The Choice of Noninformative Priors for Thompson Sampling in Multiparameter Bandit Models

Jongyeong Lee[1, 2*], **Chao-Kai Chiang**[1], **Masashi Sugiyama**[2, 1]

[1]The University of Tokyo
[2]RIKEN AIP
lee@ms.k.u-tokyo.ac.jp, chaokai@edu.k.u-tokyo.ac.jp, sugi@k.u-tokyo.ac.jp

Abstract

Thompson sampling (TS) has been known for its outstanding empirical performance supported by theoretical guarantees across various reward models in the classical stochastic multi-armed bandit problems. Nonetheless, its optimality is often restricted to specific priors due to the common observation that TS is fairly insensitive to the choice of the prior when it comes to asymptotic regret bounds. However, when the model contains multiple parameters, the optimality of TS highly depends on the choice of priors, which casts doubt on the generalizability of previous findings to other models. To address this gap, this study explores the impact of selecting noninformative priors, offering insights into the performance of TS when dealing with new models that lack theoretical understanding. We first extend the regret analysis of TS to the model of uniform distributions with unknown supports, which would be the simplest non-regular model. Our findings reveal that changing noninformative priors can significantly affect the expected regret, aligning with previously known results in other multiparameter bandit models. Although the uniform prior is shown to be optimal, we highlight the inherent limitation of its optimality, which is limited to specific parameterizations and emphasizes the significance of the invariance property of priors. In light of this limitation, we propose a slightly modified TS-based policy, called TS with Truncation (TS-T), which can achieve the asymptotic optimality for the Gaussian models and the uniform models by using the reference prior and the Jeffreys prior that are invariant under one-to-one reparameterizations. This policy provides an alternative approach to achieving optimality by employing fine-tuned truncation, which would be much easier than hunting for optimal priors in practice.

Introduction

In the classical parametric stochastic multi-armed bandit (MAB) problems, an agent plays an arm at every round. In each round, the agent observes a reward generated from the distribution associated with the played arm, whose functional form is known, but the specific values of parameters are unknown. Since the agent observes a reward only from the played arm and is not aware of the true parameters, they have to choose an arm carefully to maximize rewards based on the history of their choices and corresponding rewards. Therefore, the MAB problem is one of the elementary models that exemplify the tradeoff between the exploration to learn parameters and the exploitation of knowledge to accumulate rewards.

For this problem, we can evaluate the performance of an agent's policy by the *regret* defined as the difference between maximum rewards and the rewards obtained from the policy since minimizing the expected regret is equivalent to maximizing expected rewards. Lai and Robbins (1985) provided an asymptotic problem-dependent lower bound on the expected regret that captures the optimal problem-dependent performance, which was generalized by Burnetas and Katehakis (1996). Note that their regret bounds are on the frequentist's view, where the parameters are regarded as fixed quantities, and we say a policy matching this lower bound to be asymptotically optimal.

Out of the various policies in the bandit literature, this paper focuses on the asymptotic optimality of Thompson sampling (TS) due to its outstanding empirical performance (Chapelle and Li 2011). TS is a randomized Bayesian policy that maintains a posterior distribution over the unknown parameters (Thompson 1933). Therefore, the choice of the priors would be important since TS plays an arm according to the posterior probability of being the best arm. When there is no prior knowledge of the parameters, it is reasonable to utilize noninformative priors based on the interpretation initially proposed by Kass and Wasserman (1996) and subsequently discussed by Robert (2007, Section 3.5):

> Noninformative priors should be taken as default priors, upon which everyone could fall back when the prior information is missing.

In this study, we translate this description to the usefulness of TS with noninformative priors as a *starting point* for bandit problems where no prior knowledge is available. One naive choice would be the uniform prior that assigns equal probability to all possible values over the parameter space (Laplace 1820), which obviously represents the ignorance of the parameters and can be defined for any model. However, as pointed out in literature (Datta and Ghosh 1996), uniform priors can vary depending on the parameterization of the distribution, which means that when the same distribution is modeled by different parameters, the resulting posterior distributions may also be different. Robert (2007)

*JL is now affiliated with Seoul National University

also emphasized the importance of invariance properties, especially when one makes inferences on multiple parameters.

Nevertheless, when it comes to the problem-dependent regret bounds of TS, it is often reported that TS is not too sensitive to the choice of the prior for the model of single-parameter distributions. For example, both the uniform prior (Kaufmann, Korda, and Munos 2012) and the Jeffreys prior (Korda, Kaufmann, and Munos 2013) are found to be optimal for the Bernoulli models. Note that the reference prior also leads to the optimal regret bound for the Bernoulli bandit models since the Jeffreys prior coincides with the reference prior for the regular single-parameter models (Ghosh 2011). This would be due to the fact that in MAB problems, the focus is solely on inferring the mean of the reward model, which differs from other pure inference tasks that involve multiple parameters of interest.

However, it has been shown that the choice of noninformative priors can significantly impact the performance of TS for noncompact multiparameter bandit models, such as the Gaussian models (Honda and Takemura 2014) and the Pareto models (Lee et al. 2023). These results indicate that the choice of noninformative priors becomes more challenging in multiparameter models than that in single-parameter models. In this paper, we first show that the prior sensitivity of TS occurs not only in the noncompact multiparameter models but also in the uniform model with unknown supports, which is a compact non-regular multiparameter model. Specifically, we show that TS with the uniform prior with location-scale (LS) parameterization is asymptotically optimal, while TS with the reference prior and the Jeffreys prior are suboptimal. The implication of this discovery is twofold. Firstly, the bounds show the importance of selecting priors in multiparameter models, extending the understanding provided by Honda and Takemura (2014) and Lee et al. (2023). Moreover, the invariance problems of the uniform priors mentioned above make the optimal regret bound less informative. This is demonstrated in the O-T column of Gaussian and uniform models in Table 1, where we showed that some uniform priors are optimal while others are not.

Moreover, recent findings have demonstrated that selecting the uniform prior with scale-shape parameterization is suboptimal for Pareto bandits (Lee et al. 2023). These results raise concerns about the reliability of the uniform prior as a fallback option, as it becomes evident that the choice of parameterization in statistical models requires meticulous consideration. This brings us to the central question that serves as the driving force behind this paper:

> Is there a *universally* applicable prior in general bandit models that *consistently* leads to high-performance outcomes when employed in posterior sampling?

As noted in Berger and Bernardo (1992), the three most important criteria for noninformative priors would be simplicity, generality, and trustworthiness. Although several well-known noninformative priors have been studied for multiparameter models, none of them simultaneously satisfy all three criteria in the context of MAB problems. In general, there is no silver bullet that can optimally address all problems. However, it might be possible to discover a "bronze bullet", a solution that achieves optimal performance in certain scenarios while still maintaining reasonable effectiveness in others, which can serve as a valuable *baseline*.

Model	R	C	T	Parameter θ	Priors	O-T	O-TT
Uniform	✗	✓	**L**	location and scale $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$	$\pi_{\mathrm{u}}^{\mu,\sigma}$	✓	✓
					π_{j}	✗	✓
					π_{r}	✗	✓
				location and rate $(\mu, \sigma^{-1}) \in \mathbb{R} \times \mathbb{R}_+$	$\pi_{\mathrm{u}}^{\mu,\frac{1}{\sigma}}$	✗	✓
Gaussian	✓	✗	**L**	location and scale $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$	$\pi_{\mathrm{u}}^{\mu,\sigma}$	$✓_H$	✓
					π_{j}	$✗_H$	✓
					π_{r}	$✗_H$	✓
				location and rate $(\mu, \sigma^{-1}) \in \mathbb{R} \times \mathbb{R}_+$	$\pi_{\mathrm{u}}^{\mu,\frac{1}{\sigma}}$	✗	✓
Pareto	✗	✗	**H**	scale and shape $(\sigma, \alpha) \in \mathbb{R}_+ \times \mathbb{R}_{\geq 1}$	$\pi_{\mathrm{u}}^{\sigma,\alpha}$	$✗_L$	$✓_L$
					π_{j}	$✗_L$	$✓_L$
					π_{r}	$✗_L$	$✓_L$
				rate and shape $(\sigma^{-1}, \alpha) \in \mathbb{R}_+ \times \mathbb{R}_{\geq 1}$	$\pi_{\mathrm{u}}^{\frac{1}{\sigma},\alpha}$	✗	**?**

Table 1: Asymptotic optimality with different noninformative priors for multiparameter models. R, C, and T denote whether the model satisfies the Fisher regularity or not, whether it is compact or non-compact, and whether its function is light-tailed (L) or heavy-tailed (H). O-T and O-TT indicate the optimality of TS and TS with truncation (TS-T), respectively, in terms of whether they can achieve the asymptotic regret lower bound for the corresponding model or not. Notice that $_H$ and $_L$ indicate that the results are derived by Honda and Takemura (2014) and by Lee et al. (2023), respectively. π_{u}, π_{j}, and π_{r} denote the uniform prior, the Jeffreys prior, and the reference priors, respectively. For the uniform priors, we specify the parameterization in the superscript. **?** denotes unknown results.

On the other hand, one might be looking forward to an alternative approach with renowned (invariant) priors that can provide practical and optimal solutions rather than hunting for good priors. In this regard, we propose a variant of TS, called TS with Truncation (TS-T), for the uniform models and the Gaussian models. We provide a finite-time regret analysis of TS-T, which demonstrates its asymptotic optimality under the reference prior and the Jeffreys prior for both models. Our approach builds upon the basic strategy of TS, but with key modifications that improve the performance and address the limitations of TS. In particular, we devise an adaptive truncation procedure on the parameter space of the posterior distribution to control the problems in the early stage of learning, hence the name truncation in TS-T. The proposed policy is inspired by the policies proposed in Jin et al. (2021) and Lee et al. (2023), extending and generalizing their approaches. We further provide a high-level design idea that can be generalized to other reward models easily.

The main results of this paper and related works are summarized in Table 1, and our contributions are summarized as follows:

- We prove the asymptotic optimality/suboptimality of TS with noninformative priors for the uniform bandits. This extends the understanding of TS in the multiparameter models, which have not been well studied so far, emphasizing the significance of selecting noninformative priors.

- We show that some uniform priors with different parameterizations are suboptimal. This makes the optimality of TS with the uniform prior less attractive in general, as it inherently involves the non-trivial task of selecting appropriate parameterizations.
- We propose a variant of TS that is asymptotically optimal for the uniform models and the Gaussian models under the reference prior and the Jeffreys prior, where the vanilla TS is found to be suboptimal. This provides optimal results that remain consistent regardless of the way of parameterizing the models, which addresses the limitations of the vanilla TS.

Preliminaries

In this section, we formulate K-armed bandit problems and the asymptotic regret lower bound for the uniform models and Gaussian models.

Problem Formulation

Suppose that there are finite K arms associated with a reward distribution ν_θ belonging to the LS family, whose density function is denoted by $f_{l,\sigma}(x)$ with location $l \in \mathbb{R}$ and scale $\sigma \in \mathbb{R}_+$. Here, the parameters $\theta = (l, \sigma) \in \mathbb{R} \times \mathbb{R}_+$ are unknown to the agent. Note that we consider MAB problems where every arm is modeled by the *same* type of distribution but with possibly different parameters.

If a random variable X with the density function $f_\theta(x)$ belongs to the LS family, then $f_{l,\sigma}$ can be written using a probability density function $f_{0,1}(\cdot)$ as

$$f_{l,\sigma}(x) = \frac{1}{\sigma} f_{0,1}\left(\frac{x-l}{\sigma}\right). \tag{1}$$

Although location l is not necessarily equivalent to the expectation $\mu(\theta) = \mathbb{E}_{\nu_\theta}[X]$ in general, we use them interchangeably in this paper since they coincide for both the Gaussian and uniform models. One can retrieve the density function of the Gaussian distribution $\mathrm{Gaussian}(\mu, \sigma)$ with location (mean) μ and scale σ, $f^{\mathrm{G}}_{\mu,\sigma}(x)$, by substituting the standard normal density for $f_{0,1}$. The uniform distribution can be obtained by letting $f_{0,1}(x) = \mathbf{1}[0 \le x \le 1]$ for the indicator function $\mathbf{1}[\cdot]$. If X follows the uniform distribution $\mathrm{Uni}_{\mu\sigma}(\mu, \sigma)$ under the LS parameterization, then it has the density of the form with location (mean) μ and scale σ,

$$f^{\mathrm{U}_{\mu\sigma}}_{\mu,\sigma}(x) = \frac{1}{\sigma}\mathbf{1}\left[\mu - \frac{\sigma}{2} \le x \le \mu + \frac{\sigma}{2}\right].$$

The uniform distribution can be reparameterized in terms of the boundary of the support by letting $(a, b) = \left(\mu - \frac{\sigma}{2}, \mu + \frac{\sigma}{2}\right)$, denoted by $\mathrm{Uni}_{ab}(a, b)$, whose density function is given as $f^{\mathrm{U}_{ab}}_{a,b}(x) = \frac{1}{b-a}\mathbf{1}[a \le x \le b]$. Here, we assume that the arm 1 is the unique optimal arm that has the maximum expected reward for convenience without loss of generality, i.e., $\mu_1 = \max_{i \in [K]} \mu_i$ and $\mu_1 > \mu_i$ for $i \in \{2, \dots, K\}$. This assumption is made to simplify the analysis, and it is worth noting that incorporating additional optimal arms can only decrease the expected regret of TS (see Agrawal and Goyal 2012, Appendix A).

Denote the index of the arm played at round t by $j(t)$ and the number of rounds that the arm i is played until round t by $N_i(t) = \sum_{s=1}^{t-1} \mathbf{1}[j(s) = i]$. Then, the regret at round T is defined with the sub-optimality gap $\Delta_i := \mu_1 - \mu_i$ as

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} \Delta_{j(t)} = \sum_{i=2}^{K} \Delta_i N_i(T+1).$$

When the sub-optimality gap is regarded as a fixed quantity, Burnetas and Katehakis (1996) showed that any policy, satisfying $\mathrm{Reg}(T) = o(t^\alpha)$ for all $\alpha \in (0, 1)$, must satisfy

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg}(T)]}{\log T} \ge \sum_{i=2}^{K} \frac{\Delta_i}{\inf_{\theta:\mu(\theta)>\mu_1} \mathrm{KL}(\nu_{\theta_i}; \nu_\theta)}, \tag{2}$$

where $\mathrm{KL}(\cdot;\cdot)$ denotes the Kullback-Leibler (KL) divergence. Here, an algorithm is said to be asymptotically optimal if it satisfies

$$\limsup_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg}(T)]}{\log T} \le \sum_{i=2}^{K} \frac{\Delta_i}{\inf_{\theta:\mu(\theta)>\mu_1} \mathrm{KL}(\nu_{\theta_i}; \nu_\theta)}.$$

The infimum over the KL divergence can be explicitly computed for any $i \neq 1$ under uniform models (Cowan and Katehakis 2015) as

$$\inf_{\theta:\mu(\theta)>\mu_1} \mathrm{KL}(\nu_{\theta_i}; \nu_\theta) = \log\left(1 + \frac{2\Delta_i}{\sigma_i}\right) \tag{3}$$

and under Gaussian models (Honda and Takemura 2014) as

$$\inf_{\theta:\mu(\theta)>\mu_1} \mathrm{KL}(\nu_{\theta_i}; \nu_\theta) = \frac{1}{2}\log\left(1 + \left(\frac{\Delta_i}{\sigma_i}\right)^2\right). \tag{4}$$

Thompson Sampling and the Choice of Priors

In this section, we instantiate TS and propose a variant of TS, TS-T, for the uniform model and the Gaussian model based on the noninformative priors.

Noninformative Priors in the LS Family

To develop an invariant noninformative prior, one can consider the Fisher information matrix (FIM), which does not rely on any prior information on unknown parameters. The FIM for the LS family is given as follows (Ghosh 2011):

$$I(l, \sigma) = \sigma^{-2}\begin{bmatrix} c_1 & c_2 \\ c_2 & c_3 \end{bmatrix},$$

where c_1, c_2, and c_3 are functions of f and do not involve parameters $\theta = (l, \sigma)$. Then, the FIM for the uniform model and the Gaussian model are given as follows:

$$(c_1, c_2, c_3) = \begin{cases} (0, 0, 1) & \text{if } f_{l,\sigma} = f^{\mathrm{U}_{\mu\sigma}}_{\mu,\sigma}, \\ (1, 0, 2) & \text{if } f_{l,\sigma} = f^{\mathrm{G}}_{\mu,\sigma}. \end{cases}$$

Since $c_2 = 0$, from the orthogonality, the first-order probability matching prior is of the form σ^{-k} for $k \in \mathbb{R}$ (Tibshirani 1989; Nicolaou 1993). This prior not only provides the posterior in a close form, but also encompasses various well-known noninformative priors as special cases in the LS

family such as the uniform prior $\pi_{\mathrm{u}}(l, \sigma) \propto 1$ by $k = 0$. Throughout the rest of the paper, unless otherwise stated, π_{u} denotes the uniform prior with (l, σ) parameterization.

Furthermore, when $k = 1$, it coincides with the reference prior $\pi_{\mathrm{r}}(l, \sigma) \propto \sigma^{-1}$, which is the unique second-order probability matching prior (Datta and Mukerjee 2004). On the other hand, the Jeffreys prior is not defined well for the uniform model since the determinant of the FIM is zero. Nevertheless, in this paper, we call prior with $k = 2$ as the Jeffreys prior $\pi_{\mathrm{j}}(l, \sigma) \propto \sigma^{-2}$ even for the uniform model to maintain consistency with the Gaussian model. More details on the noninformative priors are provided in the appendix for completeness.

Thompson Sampling

For the priors σ^{-k}, we denote the joint posterior distribution after observing n rewards from the arm i, $X_{i,n} := (x_{i,1}, \ldots, x_{i,n})$ by $\pi^k(\mu, \sigma | X_{i,n})$ or simply $\pi^k_{i,n}(\mu, \sigma)$. Let us denote the (classical) sufficient statistic $T(X_{i,n})$ for the parameter (μ_i, σ_i). Since the sufficient statistic is always Bayes-sufficient (Blackwell and Ramamoorthi 1982), one can rewrite the posterior distribution using the sufficient statistic as

$$\pi^k(\mu, \sigma | X_{i,n}) = \pi^k(\mu, \sigma | T(X_{i,n})).$$

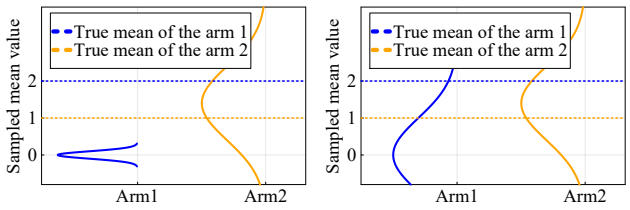
The vanilla TS observes samples $(\tilde{\mu}_i(t), \tilde{\sigma}_i(t))$ generated from the posterior $\pi^k_{i,N_i(t)}(\mu, \sigma)$ at each round. Since maximum likelihood estimators (MLEs) can be chosen as a function of sufficient statistics if any MLE exists (Moore 1971), we denote the posterior after n observations as $\pi^k(\mu, \sigma | \hat{\mu}_{i,n}, \hat{\sigma}_{i,n})$, instead of $\pi^k(\mu, \sigma | T(X_{i,n}))$, to explicitly indicate the estimates after n observations for the priors σ^{-k}. We adopt this notation as it facilitates a clear distinction between the vanilla TS and TS-T.

Thompson Sampling with Truncation

As shown in previous studies on the multiparameter bandit models (Honda and Takemura 2014; Lee et al. 2023), TS sometimes plays only suboptimal arms when the posterior of the optimal arm has a very small variance in the early stage of learning, which contributes to the suboptimality in *expectation*. To avoid such problems, TS-T samples parameters from the distributions obtained by replacing an MLE of the scale $\hat{\sigma}_n$ with a truncated estimator $\bar{\sigma}_n$ satisfying $\bar{\sigma}_n = \Omega(n^{-\beta})$ for some $\beta > 0$. Note that we choose a specific β to make regret analysis simple, but our discussion can be easily extended to any $\beta > 0$. Such truncation prevents an extreme case where $\hat{\sigma}_n \approx 0$ for small n in the regret analysis. In summary, TS-T is a policy that samples parameters from the distribution at every round, which is

$$\bar{\pi}^k_{i,n}(\mu, \sigma) = \pi^k(\mu, \sigma | \hat{\mu}_{i,n}, \bar{\sigma}_{i,n}). \tag{5}$$

Strictly speaking, TS-T is not a Bayesian policy but rather a kind of randomized probability matching policy as the distribution in (5) is not a posterior distribution anymore. However, TS-T can be seen as a pre-processed posterior probability matching policy since the truncation is applied before sampling and will behave like TS as n increases where the truncation has almost no effect.



(a) Posterior distribution. (b) "Stretched" posterior.

Figure 1: An example where the posterior distribution of each arm belongs to the Gaussian distribution. The solid lines represent the posterior probability of sampling mean values, while the blue and red dashed lines indicate the *true* expected rewards of each arm, respectively.

General Design Idea of TS-T Adaptive truncation in the parameter space of the posterior was considered in Lee et al. (2023), where they aimed to compensate for the change of the priors by replacing the MLE with a truncated one. The following design principle is a generalization of their approach to handling the problems in the first few rounds:

> Truncate the parameter space of the posterior distribution to *stretch* the distribution, which encourages a policy to *explore more* in the early stage of learning.

Here, stretching the posterior distribution can be seen as flattening the posterior distributions, which prevents them from overly concentrating on the specific value in the first few rounds. By flattening the distributions, we encourage exploration and avoid prematurely favoring a specific arm based on the small number of observations.

As an illustration, we consider a case where the posterior distribution is represented by a Gaussian distribution in Figure 1, where Figure 1a displays the posteriors of each arm. During the initial learning phase, the inherent randomness of the rewards can cause the posterior distribution of the optimal arm (arm 1) to be concentrated around a small value, such as 0, in this particular example. As a result, this concentration of the posterior may result in suboptimal behavior, where the vanilla TS is more likely to play the arm 2 that exhibits a higher expected reward according to the current posterior distribution. To address this issue, one can lift the scale parameter of the Gaussian (posterior), as depicted in Figure 1b, in order to prevent the occurrence of extreme cases during the early stage of learning. Obviously, one has to design the truncation carefully to cover the entire parameter space of the posterior as the number of samples increases.

In this paper, we truncate the parameter space by replacing sufficient statistics with truncated ones, which induces a truncated estimator instead of the MLE. Therefore, we expect that our approach can be easily applied to any model where sufficient statistics have a constant dimension, such as the (quasi-)exponential family (Robert 2007). This offers an alternative approach to achieving optimality without the need to search for an optimal or appropriate prior for each specific problem, a process we expect will be significantly more convenient in practical applications.

Comparison with Different Adaptive Approaches It is worth noting that a similar adaptive approach has been considered in the Gaussian model with known variance (Jin et al. 2021) and linear models (Hamidi and Bayati 2020). In these approaches, the posterior distribution was modeled as a Gaussian distribution and an adaptive inflation value ρ_t was introduced to the scale parameter, which effectively flattened the posterior distributions. If one extends their approaches to the LS family, it becomes a probability matching policy with the modified posterior $\pi^k(\mu, \sigma|\hat{\mu}_{i,n}, \rho_t\hat{\sigma}_{i,n})$. However, we found that this still has a similar problem to the naive TS in our analysis, which is related to Lemmas 10 and 12 in the appendix[1]. In addition, Jin et al. (2021) clipped the outputs after sampling to achieve minimax optimality, which can be seen as a post-processed posterior matching policy. While our paper does not establish the minimax optimality of TS-T, we expect that combining similar techniques with our approach could be a promising direction for the simultaneous achievement of asymptotic optimality and minimax optimality in multiparameter models, which presents an interesting problem for follow-up investigation.

Analytical Expressions of Posterior Distributions

Here, we present the formulation of the posterior for TS and TS-T in the uniform and Gaussian models. The detailed derivation for the uniform model is given in the appendix.

Uniform Bandits If rewards $(x_{i,s})$ follow $\mathrm{Uni}_{\mu\sigma}(\mu_i, \sigma_i)$, the sufficient statistic is given as $T(X_{i,n}) = (x_i^{(1)}, x_i^{(n)})$ for $x_i^{(1)} = \min_{s\in[n]} x_{i,s}$ and $x_i^{(n)} = \max_{s\in[n]} x_{i,s}$. Then, the marginal posterior of σ and the conditional posterior of μ given σ under the prior σ^{-k} are given as follows:

$$\pi^{\mathrm{U},k}(\sigma|\hat{\mu}_{i,n}, \hat{\sigma}_{i,n}) = n_k(n_k+1)\left(\hat{\sigma}_{i,n}\right)^{n_k} \frac{\sigma - \hat{\sigma}_{i,n}}{\sigma^{n_k+2}} \mathbf{1}\left[\sigma \geq \hat{\sigma}_{i,n}\right], \quad (6)$$

$$\pi^{\mathrm{U},k}(\mu|\hat{\mu}_{i,n}, \hat{\sigma}_{i,n}, \sigma = \tilde{\sigma}) = f^{\mathrm{U}_{\mu\sigma}}_{\hat{\mu}_{i,n}, \tilde{\sigma} - \hat{\sigma}_{i,n}}(\mu), \quad (7)$$

where MLEs $\hat{\mu}_{i,n} = \frac{x_i^{(n)} + x_i^{(1)}}{2}$ and $\hat{\sigma}_{i,n} = x_i^{(n)} - x_i^{(1)}$, and $n_k = n + k - 2$.

Here, following Lee et al. (2023), we employ a sequential sampling scheme to avoid the use of computationally costly approximation methods. This means that $\tilde{\sigma}$ is sampled first from the marginal posterior in (6), which can be easily implemented by using the inverse transform sampling method. Then we sample $\tilde{\mu}$ from the conditional posterior given the sampled scale parameter $\tilde{\sigma}$ in (7). This sequential sampling approach yields the same result as sampling μ from the joint posterior $\pi_{i,n}(\mu, \sigma) = \pi_{i,n}(\sigma)\pi_{i,n}(\mu|\sigma)$. Here, initial $n_0 = \max(2, 3 - \lceil k \rceil)$ plays are required to avoid improper posteriors, where $\lceil \cdot \rceil$ denotes the ceiling function.

[1]This does not necessarily imply that this approach cannot provide the optimal solution to our problem. Therefore, one might be able to show its suboptimality in a similar way to Theorem 2 or set adaptive inflation ρ_t to achieve the regret lower bounds in (3) and (4) asymptotically although it would be more difficult than our approach in the multiparameter bandit models.

As described in (5), TS-T is a sampling policy with the distribution parameterized by a truncated scale estimator. For the uniform models, we simply replace $x^{(n)}$ with a truncated statistic $\bar{x}^{(n)} = \max(x^{(1)} + n^{-1}, x^{(n)})$. In other words, we replace $\hat{\sigma}_n$ with $\bar{\sigma}_n = \bar{x}^{(n)} - x^{(1)}$, which satisfies $\bar{\sigma}_n \geq n^{-1}$. This truncation procedure is specific to the posterior sampling in (6) and (7), and is introduced to avoid the situation where parameters are sampled from a distribution whose density function is similar to the Dirac delta function. Therefore, under the TS-T policy, an agent observes samples from the following distributions:

$$\bar{\pi}^{\mathrm{U},k}_{i,n}(\sigma) = \pi^{\mathrm{U},k}(\sigma|\hat{\mu}_{i,n}, \bar{\sigma}_{i,n}) \quad (8)$$

$$\bar{\pi}^{\mathrm{U},k}_{i,n}(\mu|\sigma = \tilde{\sigma}) = f^{\mathrm{U}_{\mu\sigma}}_{\hat{\mu}_{i,n}, \tilde{\sigma} - \bar{\sigma}_{i,n}}(\mu), \quad (9)$$

where we simply replaced $\hat{\sigma}_{i,n}$ with $\bar{\sigma}_{i,n}$ in (6) and (7).

Gaussian Bandits For the Gaussian model, the sufficient statistic is given as $T(X_{i,n}) = (\hat{x}_{i,n}, S_{i,n})$ where $\hat{x}_{i,n} = \frac{1}{n}\sum_{s=1}^n x_{i,s}$, and $S_{i,n} = \sum_{s=1}^n (x_{i,s} - \hat{x}_{i,n})^2$. Then, the marginal posterior distribution of μ under the priors σ^{-k} is given as

$$\pi^{\mathrm{G},k}\left(\mu|\hat{\mu}_{i,n}, \hat{\sigma}_{i,n}\right) = f^{\mathrm{t}}_{n_k}(\mu|\hat{\mu}_{i,n}, \hat{\sigma}_{i,n}), \quad (10)$$

where $f^{\mathrm{t}}_{n_k}(\cdot|\hat{\mu}_{i,n}, \hat{\sigma}_{i,n})$ denotes the density function of the non-standardized t-distribution with the degree of freedom $n_k = n + k - 2$, location $\hat{\mu}_{i,n} = \hat{x}_{i,n}$, and scale $\hat{\sigma}_{i,n} = \sqrt{S_{i,n}/n}$. Honda and Takemura (2014) showed that TS with priors $k \geq 1$ could not achieve the lower bound with (4).

For the realization of the TS-T policy in the Gaussian models, we consider a truncated statistic and the corresponding scale estimator as follows:

$$\bar{S}_{i,n} = \max(1, S_{i,n}) \implies \bar{\sigma}_{i,n} = \sqrt{\bar{S}_{i,n} n^{-1}} \geq n^{-\frac{1}{2}}.$$

This implies that TS-T draws a sample from the distribution whose density function is given as

$$\bar{\pi}^{\mathrm{G},k}_{i,n}\left(\mu\right) = \pi^{\mathrm{G},k}\left(\mu|\hat{\mu}_{i,n}, \bar{\sigma}_{i,n}\right) = f^{t}_{n_k}(\mu|\hat{\mu}_{i,n}, \bar{\sigma}_{i,n}), \quad (11)$$

where we just replaced $\hat{\sigma}_{i,n}$ with $\bar{\sigma}_{i,n}$ in (10). In the Gaussian models, we can easily sample the location parameter directly from its marginal posterior distribution as it can be expressed by a well-known probability distribution. Note that we require n_0 initial plays to avoid improper posteriors.

Main Results

This section provides the main theoretical results of this paper, whose detailed proofs are postponed to the appendix.

Theorem 1. *Assume that the arm* 1 *is the unique optimal arm with a finite mean. Given arbitrary* $\epsilon \in \left(0, \min_{i\neq 1} \frac{\Delta_i}{2}\right)$, *the expected regret of TS with the prior* σ^{-k} *with* $k < 1$ *for the uniform models is bounded as*

$$\mathbb{E}[\mathrm{Reg}(T)] \leq \sum_{i=2}^K \Delta_i \left(\frac{\log T}{\log\left(1 + \frac{2\Delta_i - 4\epsilon}{\sigma_i}\right)} + \frac{2\sigma_i}{\epsilon} + \frac{11}{2} - \lceil k \rceil - k \right) + \Delta_{\max} C(\epsilon, k, \sigma_1),$$

where $\Delta_{\max} = \max_{i\neq 1} \Delta_i$ *and* $C(\epsilon, k, \sigma_1) = 1 + \frac{9\sigma_1}{\epsilon} + \frac{3}{16(1-k)}\frac{\sigma_1^2}{\epsilon^2}(2e^{\frac{2\epsilon}{\sigma_1}} - 1) = \mathcal{O}\left(\frac{\sigma_1^2}{(1-k)\epsilon^2}\right)$.

Since Theorem 1 holds for any $\epsilon \in \left(0, \min_{i\neq 1} \frac{\Delta_i}{2}\right)$, letting $\epsilon = \mathcal{O}((\log T)^{-1/3})$ directly implies that

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg}(T)]}{\log T} \leq \sum_{i=2}^{K} \frac{\Delta_i}{\log\left(1 + \frac{2\Delta_i}{\sigma_i}\right)},$$

which shows the asymptotic optimality of TS with $k < 1$ in terms of the regret lower bound with (3). Notice that our bound is tighter than the optimal upper-confidence bound (UCB) based policy of Cowan and Katehakis (2015), where the remaining term is $\mathcal{O}(\epsilon^{-3})$.

Theorem 1 not only establishes asymptotic optimality but also provides two additional observations: (i) A moderate choice of k can be beneficial because having a too small k induces larger regrets as it requires many initial plays, while large k increases $C(\epsilon, k, \sigma_1)$. The reduction in $C(\epsilon, k, \sigma_1)$ is preferable when ϵ is sufficiently small. (ii) We need a more delicate approach to consider the worst-case scenario where both Δ_i and σ_i are extremely large. Since σ_i is an unknown problem-dependent constant in this paper, we cannot directly apply the techniques used in the case where σ_i is assumed to be a given fixed constant (Agrawal and Goyal 2017; Jin et al. 2021).

Next, we show that the vanilla TS with $k \geq 1$ based on the posteriors in (6) and (7) cannot achieve the regret lower bound in the theorem below. To simplify the analysis, we consider two-armed bandit problems where two arms have the same left-boundary point of the support. Furthermore, we provide the full information on the arm 2 to the agent following the previous proofs (Honda and Takemura 2014; Lee et al. 2023), where the prior on the arm 2 is the Dirac measure so that $\tilde{\mu}_2(t) = \mu_2$ holds for any round $t \in \mathbb{N}$.

Theorem 2. *Assume that the arm* 1 *follows* $\mathrm{Uni}_{ab}(a_1, b_1)$ *and the arm* 2 *follows* $\mathrm{Uni}_{ab}(a_2, b_2)$ *with* $a_1 = a_2$ *and* $b_2 < b_1$*, where* $\mu_1 > \mu_2$ *holds. When* $\tilde{\sigma}_1(t)$ *and* $\tilde{\mu}_1(t)$ *are sampled from the posteriors in (6) and (7) with the priors* $k \geq 1$*, and* $\tilde{\mu}_2(t) = \mu_2$ *holds, there exists a constant* $\xi^{\mathrm{U}} > 0$ *independent of* σ_2 *satisfying*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg}(T)]}{\log T} \geq \Delta_2 \xi^{\mathrm{U}}.$$

If $k > 1$*, then there exist constants* $\xi_k^{\mathrm{U}} > 0$ *independent of* σ_2 *satisfying*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg}(T)]}{T^{\frac{k-1}{k}}} \geq \Delta_2 \xi_k^{\mathrm{U}}.$$

Theorem 2 shows that TS with $k \geq 1$ suffers at least logarithmic regrets in expectation. Although the regret lower bound with (3) approaches zero for sufficiently small $\sigma_2 = b_2 - a_2$, the regret of TS is lower-bounded by a non-zero term since the coefficient of $\log T$ converges to a non-zero constant. Therefore, TS with prior $k \geq 1$ is suboptimal, at least for sufficiently small σ_2, where the same result was found in the Gaussian models (Honda and Takemura 2014). Furthermore, one can see that priors with $k > 2$ are suboptimal even in the view of the worst-case analysis since their regret can be larger than $\sqrt{T}$ order for some instances.

From Theorem 2, we can obtain the following corollary, which shows the suboptimality of some uniform priors with different parameterizations.

Corollary 3. *For any one-to-one transformations* $g(\mu)$ *and* $h(\sigma)$*, if* $\frac{\mathrm{d}}{\mathrm{d}\mu} g^{-1}(\mu) \propto 1$ *and* $\frac{\mathrm{d}}{\mathrm{d}\sigma} h^{-1}(\sigma) \propto \sigma^{-k}$ *hold with some* $k \geq 1$*, then TS with the uniform priors with* $(g(\mu), h(\sigma))$ *parameterization,* $\pi_{\mathrm{u}}^{g(\mu),h(\sigma)}$ *is suboptimal.*

Proof. The uniform prior with $(g(\mu), h(\sigma))$ parameterization indicates that $\pi_{\mathrm{u}}^{g(\mu),h(\sigma)} \propto 1$. Let us define $f(\mu, \sigma) = (g^{-1}(\mu), h^{-1}(\sigma))$. Then, the corresponding prior with (μ, σ) parameterization can be obtained by multiplying the absolute value of the Jacobian determinant of f, which is given as $|\det \nabla f| \cdot \pi_{\mathrm{u}}^{g(\mu),h(\sigma)} = \sigma^{-k}$. Since $k \geq 1$ holds from the assumption, the proof follows from Theorem 2 in this paper for the uniform models and from Theorem 2 in Honda and Takemura (2014) for the Gaussian models. □

The result of Corollary 3 would not be surprising since one can easily expect that some arbitrary parameterizations can result in poor performance of TS with the uniform prior. However, this variability can introduce unnecessary concerns about the appropriate way to parameterize models. While the uniform prior with the LS parameterization might seem like a natural choice in the LS family, this idea cannot be generalized to other models. For instance, the uniform prior with the scale-shape parameterization in the Pareto model was shown to be suboptimal (Lee et al. 2023) and Corollary 3 further demonstrates the suboptimality of the rate-shape parameterization. Another consideration would be the use of natural parameters for exponential family models. However, the uniform prior with $\left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)$ parameterization can be seen as prior with $k = 5$ in the LS parameterization, which is suboptimal in the Gaussian bandits. Therefore, such observations emphasize the importance of the invariance property of the priors in the MAB problems, which is related to the trustworthiness of priors.

The theorem below shows the asymptotic optimality of TS-T with the prior with any k, including the reference prior[2] and the Jeffreys prior that are invariant under any one-to-one transformations.

Theorem 4. *With the same notation as Theorem 1, the expected regret of TS-T with prior* $k \in \mathbb{R}$ *for the uniform models is bounded as*

$$\mathbb{E}[\mathrm{Reg}(T)] \leq \sum_{i=2}^{K} \Delta_i \left(\frac{\log T}{\log\left(1 + \frac{2\Delta_i - 4\epsilon}{\sigma_i}\right)} + \frac{2\sigma_i}{\epsilon} + \frac{1}{\sigma_i} + \max\left(\frac{7}{2}, \frac{9}{2} - \lceil k \rceil\right) \right) + \Delta_{\max} C'(\epsilon, k, \sigma_1),$$

[2] Although the reference priors are invariant under the transformation that preserves the group order of parameters in general (see Datta and Ghosh 1996, Theorem 2.1), it is invariant under any one-to-one transformation in the LS family (Ghosh 2011).

where $C'(\epsilon, k, \sigma_1) = 1 + \frac{9\sigma_1}{\epsilon} + \frac{3}{16}\frac{\sigma_1^2}{\epsilon^2}(2e^{\frac{2\epsilon}{\sigma_1}} - 1) = \mathcal{O}\left(\frac{\sigma_1^2}{\epsilon^2}\right)$ *for* $k < 1$, $C'(\epsilon, 1, \sigma_1) = \mathcal{O}\left(\frac{\sigma_1^2 \log(\sigma_1)}{\epsilon^2}\right)$, *and for* $k > 1$ $C'(\epsilon, k, \sigma_1) = \mathcal{O}\left(\frac{\sigma_1^{2k}}{\epsilon^{k+1}}\right)$.

Although Theorem 4 states that any prior σ^{-k} can achieve the regret lower bound *asymptotically*, we recommend using the priors with $k \in [0, 1]$ since small k requires many initial plays from $n_0 = \max(2, 3 - \lceil k \rceil)$, while large k will suffer from a large regret in the finite time due to large $C'(\epsilon, k, \sigma_1)$.

Not only for the uniform models, but TS-T with the reference prior and the Jeffreys prior are also asymptotically optimal for the Gaussian models, which were found to be suboptimal for TS (Honda and Takemura 2014).

Theorem 5. *Assume arm* 1 *is the unique optimal arm with a finite mean. Given arbitrary* $\epsilon \in \left(0, \min_{i\neq 1} \frac{\Delta_i}{2}\right)$*, there exists a problem-prior-dependent constant* $C''(\epsilon, k, \sigma_1)$ *such that the expected regret of TS-T with priors* σ^{-k} *for the Gaussian models is bounded for* $k \leq 2$ *as*

$$\mathbb{E}[\mathrm{Reg}(T)] \leq \sum_{i=2}^{K} \Delta_i \left(\frac{\log T}{\frac{1}{2}\log\left(1 + \frac{(\Delta_i - 2\epsilon)^2}{\sigma_i^2 + \epsilon}\right)} + \frac{1}{\sigma_i^2} + 3 - k + \frac{\sqrt{\sigma_i^2 + \epsilon}}{\Delta_i - 2\epsilon} + \frac{2\sigma_i^2 e^{\frac{\epsilon}{2\sigma_i^2}} + 2\sigma_i^4 e^{\frac{\epsilon}{\sigma_i^2}}}{\epsilon^2} \right) + \Delta_{\max} C''(\epsilon, k, \sigma_1),$$

where $C''(\epsilon, k, \sigma_1) = \mathcal{O}\left(\left(\frac{\sigma_1}{\epsilon}\right)^{4 + \lceil k \rceil \mathbf{1}[k \geq 1]}\right)$.

Letting $\epsilon = \mathcal{O}\left((\log T)^{-1/7}\right)$ provides an ϵ-free bound, which shows the asymptotic optimality of TS-T. Although the overall proofs of Theorem 5 resemble that of Honda and Takemura (2014), the introduction of the truncated estimator $\bar{\sigma}$ induces a technical challenge of integrating a product of the beta function and the incomplete gamma function, which did not occur in the previous analysis. We solve it by exploiting the modified Bessel functions of the second kind and confluent hypergeometric functions of the second kind to carefully control the effect of $\bar{\sigma}$.

Numerical Validation

This section presents simulation results to validate the theoretical analysis of TS and TS-T. To provide a baseline for comparison, we present the results of asymptotically optimal UCB-based policies, CK-UCB for the uniform bandits (Cowan and Katehakis 2015) where "CK" is the initials of the authors following the notation in the original paper.

We considered a 6-armed uniform bandit instance with parameters given as $\boldsymbol{\mu} = (5.5, 5.0, 4.5, 4.0, 4.75, 3.0)$ and $\boldsymbol{\sigma} = (4.5, 5.0, 4.5, 4, 3.75, 2.0)$, which was previously studied (Cowan and Katehakis 2015). In Figure 2, the solid lines denote the averaged regret over 10,000 independent runs of the policy that was found to be optimal in terms of the regret lower bound with (3), whereas the dashed lines denote that of the suboptimal policies. The dotted lines denote the asymptotic regret lower bound. Note that the Jeffreys prior ($k = 2$) coincides with the uniform prior with the location-rate parameterizations (μ, σ^{-1}). Validations in the Gaussian models are given in the appendix.

In Figure 2a, TS with the uniform prior $\pi_{\mathrm{u}}^{\mu,\sigma}$ shows the best performance, while TS with the Jeffreys prior π_{j} and the reference prior π_{r} suffer from a large regret. Although TS with the reference prior shows a similar finite-time performance to CK-UCB, it seems to have a larger regret order compared to asymptotically optimal policies. However, as shown in Figure 2b, the performance of TS-T with the reference prior improves significantly, which highlights the effectiveness of the truncation procedure in the TS-based policy.

Conclusion

In this paper, we first demonstrated the importance of choosing noninformative priors for the vanilla TS under the uniform bandit models with unknown supports. Although the uniform prior is optimal in terms of the expected problem-dependent regret, we showed that the use of the uniform prior is problematic due to its dependency on parameterizations, which makes the optimality under the specific parameterization less informative in general. On the other hand, invariant noninformative priors, the reference prior and the Jeffreys prior, are shown to be suboptimal.

Nevertheless, in the various multiparameter models, the reference priors have been shown to be on the borderline between optimal and suboptimal in terms of prior parameter k (Honda and Takemura 2014; Lee et al. 2023). Therefore, we expect that TS with the reference prior could serve as a *baseline* for other models since the reference posterior can be derived generally (Berger and Bernardo 1992) and that an optimal policy would perform at least better than TS with the reference priors. Furthermore, by combining with TS-T, one can focus on the adaptive truncation, which provides an alternative solution to achieve optimality with renowned invariant priors. We expect that adaptively truncating parameter space would be more convenient than finding good priors for each model in practice. Our analysis was supported by the simulation results, where the invariant priors under TS-T showed a better performance than those under TS.

Acknowledgements

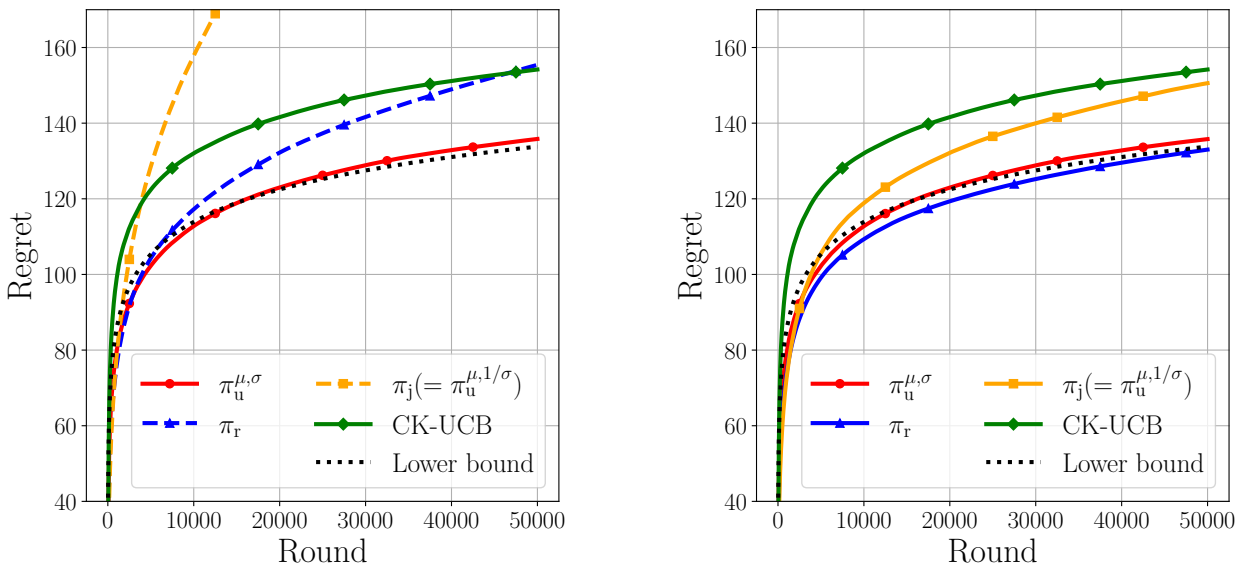
JL was supported by JST SPRING, Grant Number JPMJSP2108. CC and MS were supported by the Institute for AI and Beyond, UTokyo.

References

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

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

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



(a) Regret of TS.

(b) Regret of TS-T.

Figure 2: Cumulative regret for the 6-armed uniform bandit instance. The solid lines and the dashed lines denote the averaged values over 10,000 independent runs of the policies that can and cannot achieve the lower bound, respectively.

Blackwell, D.; and Ramamoorthi, R. 1982. A Bayes but not classically sufficient statistic. *The Annals of Statistics*, 10(3): 1025–1026.

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

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

Cowan, W.; and Katehakis, M. N. 2015. An asymptotically optimal policy for uniform bandits of unknown support. arXiv:1505.01918.

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

Datta, G. S.; and Mukerjee, R. 2004. *Probability Matching Priors: Higher Order Asymptotics: Higher Order Asymptotics*, volume 178. Springer Science & Business Media.

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

Hamidi, N.; and Bayati, M. 2020. On worst-case regret of linear Thompson sampling. arXiv:2006.06790.

Honda, J.; and Takemura, A. 2014. Optimality of Thompson sampling for Gaussian bandits depends on priors. In *International Conference on Artificial Intelligence and Statistics*.

Jin, T.; Xu, P.; Shi, J.; Xiao, X.; and Gu, Q. 2021. MOTS: Minimax optimal Thompson sampling. In *International Conference on Machine Learning*, 5074–5083.

Kass, R. E.; and Wasserman, L. 1996. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435): 1343–1370.

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

Korda, N.; Kaufmann, E.; and Munos, R. 2013. Thompson sampling for 1-Dimensional Exponential Family Bandits. In *Advances in Neural Information Processing Systems*.

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

Laplace, P. S. 1820. *Théorie analytique des probabilités*. Courcier.

Lee, J.; Honda, J.; Chiang, C.-K.; and Sugiyama, M. 2023. Optimality of Thompson Sampling with Noninformative Priors for Pareto Bandits. In *International Conference on Machine Learning*.

Moore, D. 1971. Maximum likelihood and sufficient statistics. *The American Mathematical Monthly*, 78(1): 50–52.

Nicolaou, A. 1993. Bayesian intervals with good frequentist behaviour in the presence of nuisance parameters. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(2): 377–390.

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

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

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