Diffusion Approximations for Thompson Sampling

Lin Fan
Kellogg School of Management, Northwestern University, Evanston, IL 60208, lin.fan@kellogg.northwestern.edu

Peter W. Glynn
Department of Management Science and Engineering, Stanford University, Stanford, CA 94305, glynn@stanford.edu

We study the behavior of Thompson sampling from the perspective of weak convergence. In the regime where the gaps between arm means scale as $1/\sqrt{n}$ with the time horizon n, we show that the dynamics of Thompson sampling evolve according to discrete versions of SDE's and stochastic ODE's. As $n \to \infty$, we show that the dynamics converge weakly to solutions of the corresponding SDE's and stochastic ODE's. Our weak convergence theory is developed from first principles using the Continuous Mapping Theorem, and can be easily adapted to analyze other sampling-based bandit algorithms. We further show that in this regime, the weak limits of the dynamics of many sampling-based algorithms—including Thompson sampling designed for any exponential family of rewards, and algorithms involving bootstrap-based sampling—coincide with that of Gaussian Thompson sampling.

Key words: Multi-armed Bandits, Regret Distribution, Limit Theorems, Mis-specification

History: Manuscript version – August 18, 2024

---

1. Introduction

The multi-armed bandit problem is a widely studied model that is both useful in practical applications and is a valuable theoretical paradigm exhibiting the trade-off between exploration and exploitation in sequential decision-making under uncertainty. Theoretical research in this area has focused overwhelming on studying the performance of algorithms through establishing upper and lower bounds on the expected (pseudo-)regret; see Lattimore and Szepesvári (2020) for a recent detailed account of bandit theory. The regret $\mathrm{Reg}(n) := \sum_k T_k(n)\Delta_k$ is the sum over each (sub-optimal) arm k of the number of times $T_k(n)$ it is played over horizon n, weighted by its mean reward sub-optimality gap $\Delta_k := \max_j \mu_j - \mu_k$, with μ_j as the mean reward of arm j. While expected regret $\mathbb{E}[\mathrm{Reg}(n)]$ is the most fundamental performance measure, the probabilistic behavior of $\mathrm{Reg}(n)$ can depend on other aspects of its distribution, which may be crucial to understand in some applications. For example, in settings where bandit algorithms are deployed with only a limited number of runs so that the law of large numbers does not "kick in", or in settings where risk sensitivity is a key concern, the spread or variance of $\mathrm{Reg}(n)$ can be as important for designing effective algorithms as $\mathbb{E}[\mathrm{Reg(n)}]$.

In this paper, we focus on Thompson sampling (TS) (Thompson 1933), which is a Bayesian approach for balancing exploration and exploitation that has recently become one of the most popular bandit algorithms (Chapelle and Li 2011, Agrawal and Goyal 2012, Kaufmann et al. 2012, Russo and Van Roy 2014, 2016, Russo et al. 2019); a precise description of TS is provided in Section 2.1. For TS, we derive a diffusion approximation for the distribution of $\mathrm{Reg}(n)$. To do so, we consider a "triangular array" asymptotic regime consisting of a sequence of bandit problems indexed by the horizon n, such that for a particular n, the sub-optimality gaps are on the scale of $1/\sqrt{n}$. Sending $n \to \infty$, we show that the dynamics of TS, viewed as a stochastic process, converges weakly (in distribution) to a diffusion process characterized by a stochastic differential equation (SDE).

This asymptotic regime, which we will refer to as "diffusion scaling", corresponds to so-called minimax or worst-case settings in the bandit literature, and is one of the two key paradigms in which to develop optimal bandit algorithms; see Chapters 15-16 of Lattimore and Szepesvári (2020). Indeed, for TS, which is known to be nearly minimax-optimal, the "statistically hardest" bandit environments have sub-optimality gaps scaling as $1/\sqrt{n}$ with the time horizon n (Agrawal and Goyal 2013, 2017). In such settings, there is not enough reward information for bandit algorithms to distinguish with a high degree of confidence between sub-optimal and optimal arms, and so essentially all arms are played $O_{\mathbb{P}}(n)$ times over a horizon of n, resulting in $O_{\mathbb{P}}(\sqrt{n})$ regret. Moreover, the analysis of such settings provides insight about the early stages of bandit experiments in general, when algorithms are just starting to be able to distinguish between arms.

In this paper, our three main contributions are the following.

1) Under diffusion scaling, we provide distributional approximations for the limiting dynamics of TS in terms of an SDE representation and also an equivalent stochastic ordinary differential equation (ODE) representation. See Theorems 1, 2 and 3.

2) Our proof approach for these theorems shows explicitly why the SDE and stochastic ODE weak limits arise. In particular, we start with discrete-time equations describing the evolution of TS, and then directly pass to the limit using the Continuous Mapping Theorem and elementary arguments to obtain the SDE's and stochastic ODE's. We provide intuitive derivations of these results in Sections 2.2-2.3.

3) We also deduce further insights using our limit theory. For example, under diffusion scaling, the weak limits of the dynamics of many sampling-based algorithms, including TS for exponential families and bootstrap-based sampling algorithms, coincide with that of Gaussian TS, as shown in Propositions 3 and 4. Additionally, under diffusion scaling, the regret performance of TS is generally insensitive to mis-specification of reward distributions, as shown in Proposition 2. This contrasts with the instance-dependent bandit paradigm of Lai and Robbins (1985), where algorithms can be highly sensitive to model mis-specification, as shown in Fan and Glynn (2021).

The rest of the paper is structured as follows. Related work is discussed in the following Section 1.1. We then describe the model and setup used throughout the paper in Section 2.1. In Section 2.2, we provide an intuitive derivation leading to the SDE convergence result for Gaussian TS in Theorem 1, which is proved rigorously in Section 3.1. Similarly, in Section 2.3, we provide an intuitive derivation leading to the stochastic ODE convergence result for Gaussian TS in Theorem 2, which is proved rigorously in Section 3.2. We provide extensions (Theorem 3 and Proposition 1) of diffusion approximations in Section 2.4. In Section 4.1, under diffusion scaling, we discuss the insensitivity of Gaussian TS to mis-specification of the reward distribution. We show that TS designed for exponential family reward distributions can be approximated by Gaussian TS under diffusion scaling in Section 4.2. The same is shown for bootstrap-based sampling in Section 4.3. We then conclude the paper with a quick study of batched updates for TS in Section 4.4. Additional proofs and technical results can be found in Appendices A, B, C and D.

1.1. Related Work

In the process of completing our paper, we became aware of the independent work of Kuang and Wager (2023) (abbreviated KW in the discussion below), which was posted online prior to our manuscript. The main overlap between our work and theirs is that both obtain similar SDE and stochastic ODE approximations for the dynamics of TS under diffusion scaling with $(1/\sqrt{n})$-scale sub-optimality gaps for time horizon n; see our Theorems 1 and 2, and KW's Theorems 1 and 3 (applied to the TS algorithm). However, our approach for developing the weak convergence theory (the second point in our main contributions) is different from theirs. Furthermore, we consider extensions (the third point in our main contributions) that are unrelated to theirs.

In more detail, the theoretical approach taken in our paper differs from that of KW in the following way. KW represent sequential sampling-based algorithms, including TS, as Markov chains, and use the martingale framework of Stroock and Varadhan (Stroock and Varadhan 1979) to show weak convergence of the Markov chains to diffusion processes by establishing the corresponding convergence of infinitesimal generators. On the other hand, we use representations in terms of discrete versions of SDE's and stochastic ODE's, and we show from first principles using the Continuous Mapping Theorem that the discrete systems converge weakly to their continuous counterparts.

Also related to our work, Kalvit and Zeevi (2021) has recently studied the behavior of the UCB1 algorithm of Auer et al. (2002) in worst-case gap regimes. When the gaps between arm means scale as $\sqrt{\log(n)/n}$ with the horizon n, they obtain weak diffusion limits for UCB1. Additionally, they provide distinctions between the behavior of TS and UCB algorithms when the sub-optimality gap sizes are effectively zero relative to the length of the horizon n.

Recently, Araman and Caldentey (2022) has also studied sequential binary testing for exogenous Poisson arrival processes. They obtain a diffusion limit as the intensity of arrivals increases and the informativeness of experiments decreases. They then obtain a closed-form solution for optimal experimentation and stopping for the diffusion limit, from which they deduce insights for the pre-limit problem.

2. Derivations of Diffusion Approximations

2.1. Preliminaries

As discussed in the Introduction, we consider a triangular array setup throughout the paper. The rows of the triangular array correspond to the sequence of time horizons $n = 1, 2, \ldots$. We write superscript n to index this sequence. For each time horizon n and each arm $k \in [K] := \{1, \ldots, K\}$, we have a reward distribution Q_k^n, with rewards $X_k^n(i) \overset{\text{iid}}{\sim} Q_k^n$ for different time periods i. We will use the following assumption throughout the paper.

Assumption 1 (Diffusion Scaling). *For the distributions Q_k^n, with means μ_k^n and variances $(\sigma_k^n)^2$, the following hold. There exist some $\alpha > 0$, some $\mu_* \in \mathbb{R}$, and for each arm k, some fixed $d_k \in \mathbb{R}$ and $\sigma_k > 0$ such that*

$$\mu_k^n = \mu_* + \frac{d_k^n}{\sqrt{n}}, \ \lim_{n\to\infty} d_k^n = d_k \tag{1}$$

$$\lim_{n\to\infty} \sigma_k^n = \sigma_k \tag{2}$$

$$\sup_n \mathbb{E}\left[|X_k^n(i)|^{2+\alpha}\right] < \infty. \tag{3}$$

Remark 1. For our analysis, finite $2+\alpha$ (for arbitrarily small $\alpha > 0$) moments for the rewards suffices (as in (3)), while the theory of Kuang and Wager (2023) requires finite 4^{th} moments.

Bandit Problems and Thompson Sampling

We first describe how a generic, sampling-based bandit algorithm operates over a time horizon of n. For the discussion in this section, let $\mathcal{F}_j^n$ denote the sigma-algebra capturing all information, consisting of arm selection decisions and resulting rewards, collected through time period j, which induces a filtration $(\mathcal{F}_j^n : j = 0, \ldots, n-1)$. At time $j+1$, given $\mathcal{F}_j^n$, the algorithm selects one of the K arms to play, possibly using exogenous randomization. In particular, the arm selection decision is determined by sampling according to a probability kernel $\pi^n : \mathcal{F}_j^n \to \Delta^K$, where Δ^K is the $(K-1)$-probability simplex, and component π_k^n determines the probability of playing arm $k \in [K]$. Upon playing an arm k, an iid reward is received from the Q_k^n distribution. Afterwards, $\mathcal{F}_{j+1}^n$ is updated from $\mathcal{F}_j^n$ by augmenting $\mathcal{F}_j^n$ with the newly collected information for time $j+1$, specifically the arm played and the reward received.

TS is a particular case of the above. For each arm k, we start at time $j = 0$ with an independent prior distribution $\nu_{0,k}^n$ (possibly dependent on the time horizon n) for the unknown mean μ_k^n. As a result of posterior updating, at each time $j = 0, \ldots, n-1$ and for each arm k, we have a posterior distribution $\nu_{j,k}^n$ (which is $\mathcal{F}_j^n$-measurable). Then, at time $j+1$, we draw an independent sample $\widetilde{\mu}_k^n(j+1) \sim \nu_{j,k}^n$ for each arm k, and we play the arm $\arg\max_k \widetilde{\mu}_k^n(j+1)$. So, in the case of TS, for each arm k, we have $\pi_k^n(\mathcal{F}_j^n) = \mathbb{P}(k = \arg\max_{l\in[K]} \widetilde{\mu}_l^n(j+1) \mid \mathcal{F}_j^n)$. After receiving a reward for the arm played, its posterior distribution is updated (all other posterior distributions remain the same), and $\mathcal{F}_{j+1}^n$ is also updated from $\mathcal{F}_j^n$ with the new information.

Throughout the rest of the paper, we will use two particular choices of filtration, denoted $\mathcal{G}_j^n$ (as defined in (4)) and $\mathcal{H}_j^n$ (as defined in (5)). As discussed below, the filtration $\mathcal{G}_j^n$ will lead to an SDE approximation, and the filtration $\mathcal{H}_j^n$ will lead to a stochastic ODE approximation. We will further show that these approximations are equivalent.

SDE Filtration

At time i, the algorithm decides which arm to play, which is reflected by the status of the indicator variables $I_k^n(i)$, equal to either 0 or 1, reflecting the decision to not play or play, respectively. The algorithm then receives as feedback the possibly censored rewards $I_k^n(i)X_k^n(i)$. For time horizon n, the filtration is:

$$\mathcal{G}_j^n = \sigma\left(I_k^n(i), I_k^n(i)X_k^n(i) : k \in [K];\ i = 1, \ldots, j\right). \tag{4}$$

We will use this information structure to obtain an SDE approximation in Section 2.2.

Stochastic ODE Filtration

Instead of a reward being generated for each arm in each time period, an alternative is for a reward to be generated for an arm only when that arm is played. So, for arm k, at time j, if $I_k^n(j) = 1$ (the algorithm decides to play arm k), then having collected $\sum_{i=1}^{j-1} I_k^n(i)$ previous rewards for arm k, the algorithm receives as feedback the reward $X_k^n(\sum_{i=1}^{j-1} I_k^n(i)+1) \overset{\text{iid}}{\sim} N(\mu_k^n, (\sigma_k^n)^2)$. For time horizon n, the filtration is

$$\mathcal{H}_j^n = \sigma\left(I_k^n(i), X_k^n(l) : k \in [K];\ i = 1, \ldots, j;\ l = 1, \ldots, \textstyle\sum_{i=1}^{j} I_k^n(i)\right). \tag{5}$$

We will use this information structure to obtain a stochastic ODE approximation in Section 2.3, which will turn out to be equivalent to the SDE approximation previously mentioned; see Remark 3 in Section 2.3.

Function Spaces and Weak Convergence

Throughout this paper, we will use $D^m[a, a']$ to denote the space of functions with domain $[a, a']$ and range $\mathbb{R}^m$, that are right-continuous and have limits from the left. We will use the Skorohod metric on $D^m[a, a']$, and we use $\Rightarrow$ to denote weak convergence in this space. These mathematical details can be found in, for example, Billingsley (1999), Pollard (1984), Ethier and Kurtz (1986), Whitt (2002).

2.2. SDE Approximation

In this section, we derive an SDE approximation for Gaussian TS, i.e., TS using a Gaussian prior and a Gaussian likelihood. We assume that the rewards are from general distributions (not necessarily Gaussian) and satisfy the conditions of the setup in Assumption 1.

In the current Section 2.2 and the following Section 2.3, we consider the case where μ_* is known. To keep algebraic derivations simple, we also assume $\mu_* = 0$, without loss of generality. Note however that from Assumption 1, for a time horizon of n, the arm means μ_k^n are concentrated around μ_* with *unknown gaps* between them on the scale of $1/\sqrt{n}$, which results in an interesting bandit problem.

In Sections 2.2-2.3, to utilize knowledge of how the arm means are clustered, for each instance of time horizon n, we use an independent $N(0,(bn)^{-1})$ prior for each arm in TS, with fixed $b > 0$. The use of concentrated priors with $(1/n)$-scale variance corresponds naturally to the level of precision needed for inference about the $(1/\sqrt{n})$-scale gaps between arm means. On a technical level, when the prior variance is scaled in this way, the resulting SDE has desirable Lipschitz continuity properties, which ensures it has a unique solution. We will further discuss this detail and other choices of prior for TS in Section 2.4.

For the Gaussian likelihood, we use a fixed variance $c_*^2 > 0$, which may or may not correspond to the limiting variances σ_k^2 of the rewards from (2). We will discuss model mis-specification in Section 4.1 and non-Gaussian likelihoods in Section 4.2.

To derive the SDE approximation, we use the reward structure associated with the filtration $\mathcal{G}_j^n$, as defined in (4) in Section 2.1. We will first show that the dynamics of TS are completely captured by the evolution of two processes: $R^n = (R_k^n : k \in [K])$ and $S^n = (S_k^n : k \in [K])$, defined via:

$$R_k^n(t) = \frac{1}{n} \sum_{i=1}^{\lfloor nt \rfloor} I_k^n(i) \tag{6}$$

$$S_k^n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} I_k^n(i) \frac{X_k^n(i) - \mu_k^n}{\sigma_k^n}. \tag{7}$$

REMARK 2. Fixing any horizon n under the setup of Assumption 1, the overall regret $\mathrm{Reg}(n)$ is related to the R_k^n processes via:

$$\mathrm{Reg}(n) = n \sum_{k \in [K]} R_k^n(1) \Delta_k^n,$$

where $\Delta_k^n := \max_l d_l^n - d_k^n$.

At time $j+1$, having collected history $\mathcal{G}_j^n$, TS draws a sample from the posterior distribution of each arm k:

$$\widetilde{\mu}_k^n(j+1) \sim N\left(\frac{\sum_{i=1}^j I_k^n(i) X_k^n(i)}{(R_k^n(\frac{j}{n}) + bc_*^2)n}, \frac{c_*^2}{(R_k^n(\frac{j}{n}) + bc_*^2)n} \right). \tag{8}$$

So, the probability of playing arm k can be expressed as:

$$\mathbb{P}\left(k = \arg\max_{l\in[K]} \widetilde{\mu}_l^n(j+1) \,\middle|\, \mathcal{G}_j^n\right) \tag{9}$$

$$= \mathbb{P}\left(k = \arg\max_{l\in[K]} \left\{ \frac{S_l^n(\frac{j}{n})\sigma_l^n + R_l^n(\frac{j}{n})d_l^n}{R_l^n(\frac{j}{n}) + bc_*^2} + \frac{c_*}{\sqrt{R_l^n(\frac{j}{n}) + bc_*^2}} N_l \right\} \,\middle|\, R^n(\tfrac{j}{n}), S^n(\tfrac{j}{n})\right) \tag{10}$$

$$= p_k^n(R^n(\tfrac{j}{n}), S^n(\tfrac{j}{n})), \tag{11}$$

where the probability is taken over the independent standard Gaussian variables N_l, and for $r = (r_k : k \in [K]) \in [0,1]^K$ and $s = (s_k : k \in [K]) \in \mathbb{R}^K$,

$$p_k^n(r,s) = \mathbb{P}\left(k = \arg\max_{l\in[K]} \left\{ \frac{s_l\sigma_l^n + r_l d_l^n}{r_l + bc_*^2} + \frac{c_*}{\sqrt{r_l + bc_*^2}} N_l \right\}\right). \tag{12}$$

For the probability kernel π^n introduced in Section 2.1, here we have the explicit expression: $\pi_k^n(\mathcal{G}_j^n) = p_k^n(R^n(\frac{j}{n}), S^n(\frac{j}{n}))$, which is in terms of the processes R^n and S^n.

We can now re-express $R_k^n(t)$ and $S_k^n(t)$ from (6)-(7) as

$$R_k^n(t) = \frac{1}{n}\sum_{i=1}^{\lfloor nt\rfloor} p_k^n(R^n(\tfrac{i-1}{n}), S^n(\tfrac{i-1}{n})) + M_k^n(t) \tag{13}$$

$$S_k^n(t) = \sum_{i=1}^{\lfloor nt\rfloor} \sqrt{p_k^n(R^n(\tfrac{i-1}{n}), S^n(\tfrac{i-1}{n}))}\left(B_k^n(\tfrac{i}{n}) - B_k^n(\tfrac{i-1}{n})\right), \tag{14}$$

where $M^n = (M_k^n : k \in [K])$ and $B^n = (B_k^n : k \in [K])$ are defined via:

$$M_k^n(t) = \frac{1}{n}\sum_{i=1}^{\lfloor nt\rfloor} \left(I_k^n(i) - p_k^n(R^n(\tfrac{i-1}{n}), S^n(\tfrac{i-1}{n}))\right) \tag{15}$$

$$B_k^n(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nt\rfloor} \frac{I_k^n(i)(X_k^n(i) - \mu_k^n)}{\sqrt{p_k^n(R^n(\tfrac{i-1}{n}), S^n(\tfrac{i-1}{n}))}\cdot\sigma_k^n}, \tag{16}$$

and $(I_k^n(i) : k \in [K])$ is a multinomial random variable with a single trial and success probabilities $p_k^n(R^n(\frac{i-1}{n}), S^n(\frac{i-1}{n}))$.

As $n \to \infty$, we show that M^n and B^n converge weakly to the $D^K[0,1]$ zero process and standard K-dimensional Brownian motion, respectively. Additionally, since $d_k^n \to d_k$ and $\sigma_k^n \to \sigma_k$, we have

$$p_k^n(r,s) \to p_k(r,s) \tag{17}$$

uniformly for (r,s) in compact subsets of $[0,1]^K \times \mathbb{R}^K$, where

$$p_k(r,s) = \mathbb{P}\left(k = \arg\max_{l\in[K]} \left\{ \frac{s_l\sigma_l + r_l d_l}{r_l + bc_*^2} + \frac{c_*}{\sqrt{r_l + bc_*^2}} N_l \right\}\right). \tag{18}$$

Thus, we expect (13)-(14) to be a discrete approximation to the SDE in integral form:

$$R_k(t) = \int_0^t p_k(R(u), S(u))du \tag{19}$$

$$S_k(t) = \int_0^t \sqrt{p_k(R(u), S(u))}dB_k(u), \quad k \in [K] \tag{20}$$

with standard K-dimensional Brownian motion B. As mentioned earlier in this section, the functions p_k in (18) are Lipschitz continuous, which ensures that the SDE's in (19)-(20) have a unique (strong) solution; the mathematical details can be found in Chapter 5.2 of Karatzas and Shreve (1998).

To conclude the above derivation, the rigorous stochastic ODE approximation is stated in Theorem 1 below. The proof of Theorem 1 can be found in Section 3.1, along with the development of the supporting results for the proof.

Theorem 1. *Under the diffusion scaling of Assumption 1, for a K-armed bandit and Gaussian TS with prior variance scaling as $1/n$ with time horizon n,*

$$(R^n, S^n) \Rightarrow (R, S) \tag{21}$$

as $n \to \infty$ in $D^{2K}[0,1]$, where (R, S) is the unique strong solution to the SDE:

$$dR_k(t) = p_k(R(t), S(t))dt \tag{22}$$

$$dS_k(t) = \sqrt{p_k(R(t), S(t))}dB_k(t) \tag{23}$$

$$R_k(0) = S_k(0) = 0, \quad k \in [K], \tag{24}$$

with standard K-dimensional Brownian motion B, and functions p_k as expressed in (18).

2.3. Stochastic ODE Approximation

In this section, we derive the corresponding stochastic ODE approximation for Gaussian TS. Here, we use the same setup previously used to derive the SDE approximation, as described in the initial four paragraphs of Section 2.2.

To derive the stochastic ODE approximation, we use the reward structure associated with the filtration $\mathcal{H}_j^n$, as defined in (5) in Section 2.1. Similar to the derivation of the SDE approximation, we first show that the dynamics of TS are completely captured by the evolution of two processes: $R^n = (R_k^n : k \in [K])$ with the same expression as in (6), and $Z^n \circ R^n = (Z_k^n(R_k^n) : k \in [K])$ defined via:

$$Z_k^n(R_k^n(t)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n \cdot R_k^n(t)} \frac{X_k^n(i) - \mu_k^n}{\sigma_k^n}, \tag{25}$$

where $Z^n = (Z_k^n : k \in [K])$ has the expression:

$$Z_k^n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} \frac{X_k^n(i) - \mu_k^n}{\sigma_k^n}. \tag{26}$$

(For vector-valued functions U and V, we use $U \circ V$ to denote component-wise composition of U and V.) For the stochastic ODE approximation, since R^n has the same expression as in (6), the relationship to regret $\mathrm{Reg}(n)$ is the same as in Remark 2.

REMARK 3. We point out that the distribution of the process $S_k^n(t)$ (as defined in (7)) and of the process $Z_k^n(R_k^n(t))$ (as defined in (25)) are the same. As will be made clear in the proof of Theorem 2, their corresponding weak limit processes also have the same distribution. All other aspects of the SDE and stochastic ODE approximations are the same. So, the SDE and stochastic ODE approximations are equivalent in the sense that their solutions have the same unique distribution.

At time $j+1$, having collected history $\mathcal{H}_j^n$, TS draws a sample from the posterior distribution of each arm k:

$$\widetilde{\mu}_k^n(j+1) \sim N\left(\frac{\sum_{i=1}^{n \cdot R_k^n(j/n)} X_k^n(i)}{(R_k^n(\frac{j}{n}) + bc_*^2)n}, \frac{c_*^2}{(R_k^n(\frac{j}{n}) + bc_*^2)n} \right). \tag{27}$$

So, the probability of playing arm k can be expressed as:

$$\mathbb{P}\left(k = \arg\max_{l \in [K]} \widetilde{\mu}_l^n(j+1) \,\Big|\, \mathcal{H}_j^n \right) \tag{28}$$

$$= \mathbb{P}\left(k = \arg\max_{l \in [K]} \left\{ \frac{Z_l^n(R_l^n(\frac{j}{n}))\sigma_l^n + R_l^n(\frac{j}{n}) d_l^n}{R_l^n(\frac{j}{n}) + bc_*^2} + \frac{c_*}{\sqrt{R_l^n(\frac{j}{n}) + bc_*^2}} N_l \right\} \,\Bigg|\, R^n(\tfrac{j}{n}), Z^n \circ R^n(\tfrac{j}{n}) \right) \tag{29}$$

$$= p_k^n(R^n(\tfrac{j}{n}), Z^n \circ R^n(\tfrac{j}{n})), \tag{30}$$

where the probability is taken over the independent standard Gaussian variables N_l, and functions p_k^n are given by (18). For the probability kernel π^n introduced in Section 2.1, here we have the explicit expression: $\pi_k^n(\mathcal{H}_j^n) = p_k^n(R^n(\frac{j}{n}), Z^n \circ R^n(\frac{j}{n}))$, which is in terms of the processes R^n and $Z^n \circ R^n$.

We can now re-express $R_k^n(t)$ as

$$R_k^n(t) = \frac{1}{n} \sum_{i=1}^{\lfloor nt \rfloor} p_k^n(R^n(\tfrac{i-1}{n}), Z^n \circ R^n(\tfrac{i-1}{n})) + M_k^n(t), \quad k \in [K], \tag{31}$$

where $M^n = (M_k^n : k \in [K])$ is defined via:

$$M_k^n(t) = \frac{1}{n} \sum_{i=1}^{\lfloor nt \rfloor} \left(I_k^n(i) - p_k^n(R^n(\tfrac{i-1}{n}), Z^n \circ R^n(\tfrac{i-1}{n})) \right), \tag{32}$$

and $(I_k^n(i) : k \in [K])$ is a multinomial random variable with a single trial and success probabilities $p_k^n(R^n(\frac{i-1}{n}), Z^n \circ R^n(\frac{i-1}{n}))$.

As $n \to \infty$, we show that M^n and Z^n converge weakly to the $D^K[0,1]$ zero process and standard K-dimensional Brownian motion, respectively. As in previous section, the convergence in (17) holds. Thus, we expect (31) to be a discrete approximation to the stochastic ODE in integral form:

$$R_k(t) = \int_0^t p_k(R(u), B \circ R(u))du, \quad k \in [K], \tag{33}$$

with standard K-dimensional Brownian motion B, and functions p_k as expressed in (18).

To conclude the above derivation, the rigorous stochastic ODE approximation is stated in Theorem 2 below. The proof of Theorem 2 can be found in Section 3.2.

Theorem 2. *Under the diffusion scaling of Assumption 1, for a K-armed bandit and Gaussian TS with prior variance scaling as $1/n$ with time horizon n,*

$$(R^n, Z^n \circ R^n) \Rightarrow (R, B \circ R) \tag{34}$$

as $n \to \infty$ in $D^{2K}[0,1]$, where R is the unique solution to the stochastic ODE:

$$dR_k(t) = p_k(R(t), B \circ R(t))dt \tag{35}$$

$$R_k(0) = 0, \quad k \in [K], \tag{36}$$

with standard K-dimensional Brownian motion B, and functions p_k as expressed in (18).

2.4. Additional Approximations

Recall from the development of Theorems 1 and 2, with the functions p_k as defined in (18), that it is important for $(r,s) \mapsto p_k(r,s)$ to be Lipschitz continuous. This property ensures that the limiting SDE's and stochastic ODE's have unique solutions. In Proposition 1, we develop a result for general sampling-based bandit algorithms (recall the description from Section 2.1 immediately preceding that of TS) that does not involve Lipschitz continuous limiting sampling probabilities p_k. Without Lipschitz continuity, there may not be a unique solution, and so in general, we can only describe the weak limit points of the algorithm's dynamics. The proof of Proposition 1 can be found in Appendix A.

Proposition 1. *Under the diffusion scaling of Assumption 1, for a K-armed bandit and a sampling-based algorithm, suppose that as $n \to \infty$, the sampling probabilities $p_k^n(r,s) \to p_k(r,s)$ uniformly for (r,s) in compact subsets of $[0,1]^K \times \mathbb{R}^K$, where p_k is a continuous function. Then, the* *weak limit points of* $(R^n, Z^n \circ R^n)$ *in* $D^{2K}[0,1]$ *as* $n \to \infty$ *are solutions* $(R, B \circ R)$ *of the stochastic ODE:*

$$dR_k(t) = p_k(R(t), B \circ R(t))dt \tag{37}$$

$$R_k(0) = 0, \quad k \in [K], \tag{38}$$

with standard K*-dimensional Brownian motion* B*.*

Next, we develop a diffusion approximation for Gaussian TS without assuming a concentrated prior with variance scaling as $1/n$ with time horizon n. Specifically, we consider any fixed Gaussian prior (with constant variance) in the asymptotics as $n \to \infty$. Unlike in Sections 2.2-2.3, here we can take $\mu_* \in \mathbb{R}$ to be unknown. Then, the functions p_k in (18) become:

$$p_k(r,s) = \mathbb{P}\left(k = \arg\max_{l \in [K]} \left\{ \frac{s_l \sigma_l}{r_l} + d_l + \frac{c_*}{\sqrt{r_l}} N_l \right\}\right), \tag{39}$$

where the probability is taken over the independent standard Gaussian variables N_l. However, in (39), $(r,s) \mapsto p_k(r,s)$ is no longer Lipschitz continuous for points near $r_l = 0$, $l \in [K]$. We saw in Proposition 1 that without Lipschitz continuity, it is challenging to ensure SDE/stochastic ODE solution uniqueness.

Nevertheless, the problem with the p_k in (39) only exists for an infinitesimally small initial interval. Whenever all inputs $R_l(t)$, $l \in [K]$ to the r_l components in (39) become strictly positive, then from that time onwards, there is Lipschitz continuity and SDE/stochastic ODE solution uniqueness. One simple way to eliminate the issue is to force-sample the arms with constant, positive probabilities $(\overline{p}_k : k \in [K])$ for some arbitrarily small $\epsilon \in (0,1)$ initial fraction of the horizon n, and then run TS. This is the approach taken in Theorem 3. The proof of Theorem 3 is a simple modification of the proofs of Theorems 1-2, and is omitted.

THEOREM 3. *Under the diffusion scaling of Assumption 1, consider a* K*-armed bandit and Gaussian TS with a fixed prior variance (not changing with* n*). For any fixed* $\epsilon \in (0,1)$*, suppose the arms are sampled with constant, positive probabilities* $(\overline{p}_k : k \in [K])$ *for the first* $\lfloor \epsilon n \rfloor$ *periods, after which TS is run. Then,*

$$(R^n, S^n) \Rightarrow (R, S) \tag{40}$$

as $n \to \infty$ *in* $D^{2K}[\epsilon, 1]$*, where* (R,S) *is the unique strong solution to the SDE:*

$$dR_k(t) = p_k(R(t), S(t))dt \tag{41}$$

$$dS_k(t) = \sqrt{p_k(R(t), S(t))}dB_k(t) \tag{42}$$

$$R_k(\epsilon) = \overline{p}_k \epsilon \tag{43}$$

$$S_k(\epsilon) = \sqrt{\overline{p}_k} B_k(\epsilon), \quad k \in [K], \tag{44}$$

with standard K-dimensional Brownian motion B, and functions p_k as expressed in (39). Moreover, an equivalent characterization of the dynamics of (R,S) in (41)-(42) is in terms of $(R, B\circ R)$, with R evolving according to the stochastic ODE:

$$dR_k(t) = p_k(R(t), B\circ R(t))dt, \quad k \in [K]. \tag{45}$$

3. Proofs of Main Theorems

3.1. Proofs for SDE Approximation

In this section, we prove the SDE approximation in Theorem 1 (from Section 2.2). We first discuss a (random) mapping due to Kurtz and Protter (1991), which allows any function in $D[0,1]$ to be approximated by a (random) step function with arbitrarily good accuracy. This mapping idea makes it simple to transition from discrete versions of Itô integrals to the Itô integrals themselves in the continuous weak limit. The mapping is defined in Definition 1, some basic properties of it are given in Remark 4, and its key properties and relevance to Itô integrals are given in Lemmas 1-2. Afterwards, the proof of Theorem 1 is given. We conclude the section with Lemmas 3-4, which establish tightness of stochastic processes and convergence to Brownian motion, and are supporting results used in the proof of Theorem 1.

Definition 1. For Lemmas 1-2 below, for any $\epsilon > 0$, we define a random mapping $\chi_\epsilon : D[0,1] \to D[0,1]$ as follows. For any $z \in D[0,1]$, define inductively the random times τ_j, starting with $\tau_0 = 0$:

$$\tau_{j+1} = \inf\{t > \tau_j : \max(|z(t) - z(\tau_j)|, |z(t-) - z(\tau_j)|) \geq \epsilon U_{j+1}\}, \tag{D1}$$

where $U_j \overset{\text{iid}}{\sim} \text{Unif}[\frac{1}{2}, 1]$. Then let

$$\chi_\epsilon(z)(t) = z(\tau_j), \quad \tau_j \leq t < \tau_{j+1}, \tag{D2}$$

and note that $\chi_\epsilon(z)$ is a step function (piecewise constant).

Remark 4. We mention here a few properties of the random mapping χ_ϵ. For justification, see the discussion preceding Lemma 6.1 of Kurtz and Protter (1991). First of all, for any $z \in D[0,1]$, with χ_ϵ as defined in (D1)-(D2), we have $\sup_{0\leq t\leq 1} |z(t) - \chi_\epsilon(z)(t)| \leq \epsilon$, so χ_ϵ yields an ϵ-uniform approximation. Note that the purpose of the uniform random variables U_j in defining the random times τ_j is to avoid pathological issues concerning, for instance, the locations of jump discontinuities of the functions $z^n, z \in D[0,1]$. (In the settings that we consider, it is not absolutely crucial that we use such randomization to avoid pathological issues. We can define deterministic step function approximations by considering sequences of partition refinements. However, this randomization idea introduced in Kurtz and Protter (1991) is both simple and adaptable to other settings, so we adopt it here.) In particular, since each U_j avoids any (fixed) countable set with probability one, for each τ_j, either z will be continuous at τ_j or we will have

$$|z(\tau_j-)-z(\tau_{j-1})|<\epsilon U_j<|z(\tau_j)-z(\tau_{j-1})|\,.$$

This ensures that if $z^n \to z$ with respect to the Skorohod metric, then $\tau_j^n \overset{\text{a.s.}}{\to} \tau_j$ and $z^n(\tau_j^n) \overset{\text{a.s.}}{\to} z(\tau_j)$ for each j, where τ_j^n is defined for z^n via (D1). These properties give rise to additional helpful properties in Lemma 1 below.

LEMMA 1 **(Continuity of ϵ-Uniform Approximation)**. *Let $\epsilon > 0$ and χ_ϵ be the random mapping defined in (D1)-(D2). For any $z \in D^m[0,1]$, with $\chi_\epsilon \circ z$ denoting the component-wise application of χ_ϵ to z, the mapping $z \mapsto (z, \chi_\epsilon \circ z)$ is continuous at z almost surely for each realization of χ_ϵ. Furthermore, let ξ^n be a sequence of processes taking values in $D^m[0,1]$ and adapted to filtrations $(\mathcal{F}_t^n : 0 \le t \le 1)$. Then $\chi_\epsilon \circ \xi^n$ is adapted to the augmented filtrations $\mathcal{G}_t^n = \sigma(\mathcal{F}_t^n \cup \mathcal{H})$, where $\mathcal{H} = \sigma(U_j : j \ge 1)$ (with the U_j from (D1)) is the sigma-algebra generated by the randomization in defining χ_ϵ, which is independent of the sequence of filtrations $\mathcal{F}_t^n$. (See Lemma 6.1 of Kurtz and Protter (1991).)*

REMARK 5. For a step function $z_1 \in D[a,b]$, with jump points $s_1 < \cdots < s_m$ (and $s_0 = a$, $s_{m+1} = b$), and a continuous function z_2 on $[a,b]$, we will always use the following definition of integration:

$$\int_a^b z_1(s)dz_2(s) = \sum_{j=0}^{m} z_1(s_j)\left(z_2(s_{j+1}) - z_2(s_j)\right). \tag{46}$$

LEMMA 2 **(Continuity of Approximate Stochastic Integration)**. *Let (ξ_1^n, ξ_2^n) and (ξ_1, ξ_2) be $D^2[0,1]$ functions such that jointly $(\xi_1^n, \xi_2^n) \to (\xi_1, \xi_2)$ with respect to the Skorohod metric, and ξ_2 is a continuous function. For any $\epsilon > 0$, define the mapping $\mathcal{I}_\epsilon : D^2[0,1] \to D[0,1]$ by*

$$\mathcal{I}_\epsilon(z_1, z_2)(t) = \int_0^t \chi_\epsilon(z_1(s))dz_2(s), \tag{47}$$

where $(z_1, z_2) \in D^2[0,1]$. (Note that the integral is defined as in (46), since $\chi_\epsilon(z_1)$ is always a step function.) Then, almost surely for each realization of χ_ϵ, we have

$$\mathcal{I}_\epsilon(\xi_1^n, \xi_2^n) \to \mathcal{I}_\epsilon(\xi_1, \xi_2)$$

with respect to the Skorohod metric.

Proof of Lemma 2. For fixed $\epsilon > 0$, let $\tau_1^n, \tau_2^n, \ldots$ denote the jump times for $\chi_\epsilon(\xi_1^n)$, and let $\tau_1, \tau_2, \ldots$ denote the jump times for $\chi_\epsilon(\xi_1)$, all according to the definitions in (D1)-(D2). Since $\xi_1 \in D[0,1]$, for some finite M (depending on the particular realization of χ_ϵ), there are only M such jump discontinuities of $\chi_\epsilon(\xi_1)$ (at $\tau_1, \ldots, \tau_M$) that are at least $\epsilon/2$ in magnitude. Note that (by Lemma 1) almost surely for each realization of χ_ϵ, we have $\chi_\epsilon(\xi_1^n) \to \chi_\epsilon(\xi_1)$ with respect to the Skorohod metric. Thus, for n sufficiently large, there are also only M such jump discontinuities of $\chi_\epsilon(\xi_1^n)$ (at $\tau_1^n, \ldots, \tau_M^n$) that are at least $\epsilon/2$ in magnitude. We denote $\tau_0^n = \tau_0 = 0$ and $\tau_{M+1}^n = \tau_{M+1} = 1$. To conclude the proof, note that

$$\begin{aligned}
&\sup_{0\le t\le 1} |\mathcal{I}_\epsilon(\xi_1^n, \xi_2^n)(t) - \mathcal{I}_\epsilon(\xi_1, \xi_2)(t)| \\
&\le \sum_{j=0}^{M} \left|\chi_\epsilon(\xi_1^n(\tau_j^n)) \left(\xi_2^n(\tau_{j+1}^n) - \xi_2^n(\tau_j^n)\right) - \chi_\epsilon(\xi_1(\tau_j)) \left(\xi_2(\tau_{j+1}) - \xi_2(\tau_j)\right)\right| \\
&\to 0
\end{aligned}$$

as $n \to \infty$, since, as discussed in Remark 4 and Lemma 1, we have $\chi_\epsilon(\xi_1^n(\tau_j^n)) \to \chi_\epsilon(\xi_1(\tau_j))$ and $\xi_2^n(\tau_j^n) \to \xi_2(\tau_j)$ for each j. $\square$

We now present the proof of Theorem 1.

Proof of Theorem 1. We start with the discrete approximation (13)-(16) from our derivation in Section 2.2. We denote the joint processes via $(R^n, S^n, B^n, M^n) = (R_k^n, S_k^n, B_k^n, M_k^n : k \in [K])$, and recall that they are processes in $D^{4K}[0,1]$.

Our proof strategy is as follows. We will show that for every subsequence of (R^n, S^n), there is a further subsequence which converges weakly to a limit that is a solution to the SDE. Because the drift and dispersion functions, p_k and $\sqrt{p_k}$, of the SDE (22)-(23) are Lipschitz-continuous and bounded on their domain of definition, Theorem 5.2.9 of Karatzas and Shreve (1998) ensures that the SDE has a unique strong solution. Thus, (R^n, S^n) must converge weakly to the unique strong solution of the SDE.

By Lemma 3 (stated and proved after the current proof), the joint processes (R^n, S^n, B^n, M^n) are tight, and thus, Prohorov's Theorem (see Chapters 1 and 3 of Billingsley (1999)) ensures that for each subsequence, there is a further subsequence which converges weakly to some limit process $(R, S, B, M) = (R_k, S_k, B_k, M_k : k \in [K])$. From now on, we work with this further subsequence, and for notational simplicity, we still index this further subsequence by n. So, we have

$$(R^n, S^n, B^n, M^n) \Rightarrow (R, S, B, M). \tag{48}$$

Since M^n consists of martingale differences, by a Chebyshev bound, we have $M_k^n(t) \xrightarrow{\mathbb{P}} 0$ for each $k \in [K]$ and $t \in (0,1]$ as $n \to \infty$, and thus, M is the $D^K[0,1]$ zero process. By Lemma 4 (stated and proved after the current proof), B is a standard Brownian motion on $\mathbb{R}^K$.

Now define the processes $G^n = (G_k^n : k \in [K])$ and $G = (G_k : k \in [K])$, where

$$G_k^n(t) = p_k^n(R^n(t), S^n(t)) \tag{49}$$

$$G_k(t) = p_k(R(t), S(t)). \tag{50}$$

Since $p_k^n(r,s) \to p_k(r,s)$ as $n \to \infty$ uniformly for (r,s) in compact subsets of $[0,1]^K \times \mathbb{R}^K$, and $p_k(r,s)$ is continuous at all $(r,s) \in [0,1]^K \times \mathbb{R}^K$, by the Generalized Continuous Mapping Theorem (see Lemma 7) applied to the processes in (49)-(50), we have from (48),

$$(R^n, S^n, B^n, M^n, G^n) \Rightarrow (R, S, B, M, G). \tag{51}$$

Additionally, define the processes $\widetilde{R}^n = (\widetilde{R}_k^n : k \in [K])$ and $\widetilde{R} = (\widetilde{R}_k : k \in [K])$, where

$$\widetilde{R}_k^n(t) = \int_0^t p_k^n(R^n(u), S^n(u))du \tag{52}$$

$$\widetilde{R}_k(t) = \int_0^t p_k(S(u), S(u))du. \tag{53}$$

Recall that

$$R_k^n(t) = \frac{1}{n}\sum_{i=1}^{\lfloor nt \rfloor} p_k^n(R^n(\tfrac{i-1}{n}), S^n(\tfrac{i-1}{n})) + M_k^n(t).$$

For each k, because M_k^n converges weakly to the $D[0,1]$ zero process, and

$$\left|\frac{1}{n}\sum_{i=1}^{\lfloor nt \rfloor} p_k(R^n(\tfrac{i-1}{n}), S^n(\tfrac{i-1}{n})) - \widetilde{R}_k^n(t)\right| \le \frac{1}{n},$$

we have

$$\sup_{0\le t\le 1}\left|R_k^n(t) - \widetilde{R}_k^n(t)\right| \xrightarrow{\mathbb{P}} 0. \tag{54}$$

Thus, by the fact that integration is a continuous functional with respect to the Skorohod metric (see Theorem 11.5.1 on page 383 of Whitt (2002)) and the Continuous Mapping Theorem, we have from (51),

$$(R^n, S^n, B^n, \widetilde{R}^n, G^n) \Rightarrow (R, S, B, \widetilde{R}, G). \tag{55}$$

For any $\epsilon > 0$, let χ_ϵ be the random mapping defined in (D1)-(D2) (see Remark 4 and Lemma 1 for some basic properties of χ_ϵ). Let $\chi_\epsilon \circ G^n$ and $\chi_\epsilon \circ G$ denote the component-wise applications of χ_ϵ to the vector-valued processes G^n and G. By Lemma 1 and the Continuous Mapping Theorem, we have from (55),

$$(R^n, S^n, B^n, \widetilde{R}^n, \chi_\epsilon \circ G^n) \Rightarrow (R, S, B, \widetilde{R}, \chi_\epsilon \circ G). \tag{56}$$

Recall from (14), (16) and (49), that for each k,

$$S_k^n(t) = \int_0^t \sqrt{G_k^n(u-)}dB_k^n(u), \tag{57}$$

and define the process $\widehat{S}^n = (\widehat{S}^n_k : k \in [K])$ by

$$\widehat{S}^n_k(t) = \int_0^t \chi_\epsilon \left(\sqrt{G^n_k(u-)} \right) dB^n_k(u). \tag{58}$$

By Lemma 2 and the Continuous Mapping Theorem (with the mapping $\mathcal{I}_\epsilon$ in (47)), we have from (56),

$$(R^n, S^n, B^n, \widetilde{R}^n, \widehat{S}^n) \Rightarrow (R, S, B, \widetilde{R}, \widehat{S}), \tag{59}$$

where the process $\widehat{S} = (\widehat{S}_k : k \in [K])$ is defined by

$$\widehat{S}_k(t) = \int_0^t \chi_\epsilon \left(\sqrt{G_k(u-)} \right) dB_k(u) \tag{60}$$

with G_k defined by (50). We also define the process $\widetilde{S} = (\widetilde{S}_k : k \in [K])$ by

$$\widetilde{S}_k(t) = \int_0^t \sqrt{G_k(u-)} dB_k(u). \tag{61}$$

Note that both of the processes in (60) and (61) are well defined as Itô integrals, since by Lemma 4, the integrands (with the G_k defined in (50)) are non-anticipative with respect to the Brownian motions B_k. (As defined in (D1)-(D2), χ_ϵ depends on exogenous randomization that is independent of the B_k.) By Lemma 1, because χ_ϵ is an ϵ-uniform approximation, for each k,

$$\mathbb{E}\left[\sup_{0\le t\le 1} \left| S^n_k(t) - \widehat{S}^n_k(t) \right| \right] \le \epsilon \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[\frac{I^n_k(i)(X^n_k(i) - \mu^n_k)^2}{p^n_k(R^n(\frac{i-1}{n}), S^n(\frac{i-1}{n})) \cdot (\sigma^n_k)^2} \,\middle|\, \mathcal{G}^n_{i-1} \right] \right]^{1/2} = \epsilon. \tag{62}$$

(Recall the expressions for S^n_k and $\widehat{S}^n_k$ in (57)-(58), as well as the definition of G^n_k in (49), of B^n_k in (16), of S^n_k in (14), and of $\mathcal{G}^n_j$ in (4).) Similarly,

$$\mathbb{E}\left[\sup_{0\le t\le 1} \left| \widehat{S}_k(t) - \widetilde{S}_k(t) \right| \right] \le \epsilon \mathbb{E}\left[\langle B_k \rangle_1 \right]^{1/2} = \epsilon, \tag{63}$$

where $t \mapsto \langle B_k \rangle_t$ denotes the quadratic variation process for B_k. Putting together (54), (59)-(63) and sending $\epsilon \downarrow 0$, we have

$$(R^n, S^n, B^n, R^n, S^n) \Rightarrow (R, S, B, \widetilde{R}, \widetilde{S}). \tag{64}$$

Recalling the definition of $\widetilde{R}$ in (53) as well as that of $\widetilde{S}$ in (61) and G_k in (50), we see from (64) that the limit process (R, S, B) satisfies the SDE:

$$R_k(t) = \int_0^t p_k(R(u), S(u)) du \tag{65}$$

$$S_k(t) = \int_0^t \sqrt{p_k(R(u), S(u))} dB_k(u), \quad k = 1, \ldots, K. \tag{66}$$

(Note that from (65)-(66), it is clear that (R, S, B) is adapted to the (augmented) filtration $\mathcal{F}_t = \sigma(\mathcal{F}^B_t \cup \mathcal{N})$, where $\mathcal{F}^B_t = \sigma(B(u) : 0 \le u \le t)$, with $\mathcal{N}$ denoting the collection of all $\mathbb{P}$-null sets.) $\square$

In the following two lemmas, we show tightness and weak convergence to Brownian motion to complete the proof of Theorem 1 above.

Lemma 3. *The processes* (R^n, S^n, B^n, M^n) *defined in (13)-(16) are tight in* $D^{4K}[0,1]$.

Proof of Lemma 3. For the convenience of the reader, we recall that the processes can be expressed as:

$$R_k^n(t) = \frac{1}{n}\sum_{i=1}^{\lfloor nt \rfloor} I_k^n(i) \tag{67}$$

$$S_k^n(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nt \rfloor} I_k^n(i)\frac{X_k^n(i)-\mu_k^n}{\sigma_k^n} \tag{68}$$

$$M_k^n(t) = \frac{1}{n}\sum_{i=1}^{\lfloor nt \rfloor} \left(I_k^n(i) - p_k^n(R^n(\tfrac{i-1}{n}), S^n(\tfrac{i-1}{n}))\right) \tag{69}$$

$$B_k^n(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nt \rfloor} \frac{I_k^n(i)(X_k^n(i)-\mu_k^n)}{\sqrt{p_k^n(R^n(\tfrac{i-1}{n}), S^n(\tfrac{i-1}{n}))}\cdot\sigma_k^n}. \tag{70}$$

In particular, (13)-(14) have the equivalent expressions in (67)-(68) above. With a slight abuse of notation, let $(\mathcal{G}_t^n : 0 \le t \le 1)$ denote the continuous, piecewise constant (and right-continuous) interpolation of the discrete-time filtration $(\mathcal{G}_j^n : 0 \le j \le n)$ defined in (4), and note that (67)-(70) are all adapted to $\mathcal{G}_t^n$. Also note that the processes in (67) are uniformly bounded and increasing, and those in (68)-(70) are square-integrable martingales.

By Lemma 8, to show tightness of the joint processes (R^n, S^n, B^n, M^n), we just need to show tightness of each component sequence of processes and each pairwise sum of component sequences of processes. We use Lemma 9 to verify tightness in each case. Condition (T1) can be directly verified using a sub-martingale maximal inequality (see, for example, Theorem 3.8(i) in Chapter 1 of Karatzas and Shreve (1998)), along with a union bound when dealing with pairwise sums of component processes. Conditions (T2)-(T3) can also be directly verified. For each individual component process, we can set $\alpha_n(\delta) = \delta$, and for each pairwise sum of component processes, we can set $\alpha_n(\delta) = 4\delta$ (by bounding via: $(x+y)^2 \le 2x^2 + 2y^2$), uniformly for all n in each case. □

Lemma 4. *Following Lemma 3, for any subsequence of* (R^n, S^n, B^n, M^n) *that converges weakly in* $D^{4K}[0,1]$ *to some limit process* (R, S, B, M)*, the component* B *is a standard Brownian motion on* $\mathbb{R}^K$*. Moreover,* R *and* S *are non-anticipative with respect to* B*, i.e.,* $B(t+u) - B(t)$ *is independent of* $(R(u'), S(u'))$ *for* $0 \le u' \le t$ *and* $u \ge 0$.

Proof of Lemma 4. To show that $B^n \Rightarrow B$, where B is a standard Brownian motion, we apply the martingale functional central limit theorem stated in Lemma 10. Below, we verify (M1) and

(M2) to ensure Lemma 10 holds. Afterwards, the non-anticipative fact follows from a straightforward characteristic function argument, since in the pre-limit, R^n and S^n are non-anticipative with respect to B^n.

Verification of (M1)

As shorthand, denote $p^n_{k,i-1} := p^n_k(R^n(\frac{i-1}{n}), S^n(\frac{i-1}{n}))$. We have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{\lfloor nt\rfloor}\mathbb{E}\left[\frac{I^n_k(i)}{p^n_{k,i-1}}\left(\frac{X^n_k(i)-\mu^n_k}{\sigma^n_k}\right)^2 \,\middle|\, \mathcal{G}^n_{i-1}\right]\\
&=\frac{1}{n}\sum_{i=1}^{\lfloor nt\rfloor}\frac{\mathbb{E}\left[I^n_k(i)\,\middle|\,\mathcal{G}^n_{i-1}\right]}{p^n_{k,i-1}}\mathbb{E}\left[\left(\frac{X^n_k(i)-\mu^n_k}{\sigma^n_k}\right)^2 \,\middle|\, \mathcal{G}^n_{i-1}\right] \qquad (71)\\
&=\frac{\lfloor nt\rfloor}{n}\to t
\end{aligned}$$

as $n\to\infty$. Here, (71) follows from $p^n_{k,i-1} = p^n_k(R^n(\frac{i-1}{n}), S^n(\frac{i-1}{n}))$ being $\mathcal{G}^n_{i-1}$-measurable, and $I^n_k(i)$ and $(X^n_k(i)-\mu^n_k)^2/(\sigma^n_k)^2$ being independent conditional on $\mathcal{G}^n_{i-1}$.

Verification of (M2)

Denote

$$Y^n(i) = \frac{I^n_k(i)(X^n_k(i)-\mu^n_k)}{\sqrt{p^n_{k,i-1}}\cdot\sigma^n_k}.$$

By Markov's inequality, it suffices to show that for each fixed $i=1,\dots,n$,

$$\mathbb{E}\left[Y^n(i)^2\mathbb{I}\left(|Y^n(i)|>\epsilon\sqrt{n}\right)\right]\to 0 \qquad (72)$$

as $n\to\infty$.

We have the following three observations. 1) (R^n, S^n) is a tight sequence, as established in Lemma 3, which implies stochastic boundedness of each component with respect to the supremum norm. 2) $p^n_k(r,s)\to p_k(r,s)$ as $n\to\infty$ uniformly for (r,s) in compact subsets of $[0,1]^K\times\mathbb{R}^K$. 3) $p_k(r,s)$ is continuous and strictly positive for all $(r,s)\in[0,1]^K\times\mathbb{R}^K$. Given these three observations, for any $\gamma>0$, there exists $\delta\in(0,1)$ such that for n sufficiently large,

$$\mathbb{P}\left(\inf_{u\in[0,1]} p^n_k(R^n(u),S^n(u))<\delta\right)\le\gamma.$$

We then have

$$\begin{aligned}
&\mathbb{E}\left[Y^n(i)^2\mathbb{I}\left(|Y^n(i)|>\epsilon\sqrt{n}\right)\mathbb{I}\left(p^n_{k,i-1}<\delta\right)\right]\\
&\le\mathbb{E}\left[Y^n(i)^2\mathbb{I}\left(p^n_{k,i-1}<\delta\right)\right]\\
&=\mathbb{E}\left[\frac{\mathbb{I}\left(p^n_{k,i-1}<\delta\right)}{p^n_{k,i-1}}\mathbb{E}\left[I^n_k(i)\left(\frac{X^n_k(i)-\mu^n_k}{\sigma^n_k}\right)^2\,\middle|\,\mathcal{G}^n_{i-1}\right]\right] \qquad (73)
\end{aligned}$$

$$= \mathbb{P}\left(p^n_{k,i-1} < \delta\right) \tag{74}$$

$$\leq \mathbb{P}\left(\inf_{u\in[0,1]} p^n_k(R^n(u), S^n(u)) < \delta\right)$$

$$\leq \gamma, \tag{75}$$

where (73) follows from $R^n(\frac{i-1}{n})$ and $S^n(\frac{i-1}{n})$ being $\mathcal{G}^n_{i-1}$-measurable, (74) follows from conditional independence of $I^n_k(i)$ and $(X^n_k(i)-\mu^n_k)^2/(\sigma^n_k)^2$, and (75) holds for n sufficiently large, as established above.

Additionally, we have

$$\begin{aligned}
&\mathbb{E}\left[Y^n(i)^2\mathbb{I}\left(|Y^n(i)| > \epsilon\sqrt{n}\right)\mathbb{I}\left(p^n_{k,i-1} \geq \delta\right)\right] \\
&= \mathbb{E}\left[\frac{\mathbb{I}\left(p^n_{k,i-1} \geq \delta\right)}{p^n_{k,i-1}}\mathbb{E}\left[I^n_k(i)\left(\frac{X^n_k(i)-\mu^n_k}{\sigma^n_k}\right)^2\mathbb{I}\left(|Y^n(i)| > \epsilon\sqrt{n}\right)\,\middle|\,\mathcal{G}^n_{i-1}\right]\right] \\
&\leq \frac{1}{\delta}\mathbb{E}\left[\mathbb{E}\left[\left(\frac{X^n_k(i)-\mu^n_k}{\sigma^n_k}\right)^2\mathbb{I}\left(|Y^n(i)| > \epsilon\sqrt{n}\right)\,\middle|\,\mathcal{G}^n_{i-1}\right]\right] \\
&\leq \frac{1}{\delta}\mathbb{E}\left[\mathbb{E}\left[\left|\frac{X^n_k(i)-\mu^n_k}{\sigma^n_k}\right|^{2+\alpha}\right]^{2/(2+\alpha)}\mathbb{P}\left(|Y^n(i)| > \epsilon\sqrt{n}\,\middle|\,\mathcal{G}^n_{i-1}\right)^{\alpha/(2+\alpha)}\right] &(76)\\
&\leq \frac{C}{\delta}\mathbb{E}\left[\mathbb{P}\left(|Y^n(i)| > \epsilon\sqrt{n}\,\middle|\,\mathcal{G}^n_{i-1}\right)^{\alpha/(2+\alpha)}\right], &(77)
\end{aligned}$$

where (76) follows from Hölder's inequality, and (77) follows from (3) in Assumption 1, with constant $C > 0$. Furthermore, almost surely,

$$\begin{aligned}
\mathbb{P}\left(|Y^n(i)| > \epsilon\sqrt{n}\,\middle|\,\mathcal{G}^n_{i-1}\right) &\leq \frac{1}{\epsilon^2 n}\frac{1}{p^n_{k,i-1}}\mathbb{E}\left[I^n_k(i)\left(\frac{X^n_k(i)-\mu^n_k}{\sigma^n_k}\right)^2\,\middle|\,\mathcal{G}^n_{i-1}\right] \\
&= \frac{1}{\epsilon^2 n}.
\end{aligned}$$

So, by the bounded convergence theorem, the right side of (77) converges to zero as $n\to\infty$.

Therefore, from (75) and (77), we have

$$\limsup_{n\to\infty}\mathbb{E}\left[Y^n(i)^2\mathbb{I}\left(|Y^n(i)| > \epsilon\sqrt{n}\right)\right] \leq \gamma, \tag{78}$$

and sending $\gamma\downarrow 0$ yields (72). $\square$

3.2. Proofs for Stochastic ODE Approximation

In this section, we prove Theorem 2 (from Section 2.3), which is an alternative stochastic ODE representation of the SDE in Theorem 1 (from Section 2.2). While we could take a first-principles approach to prove Theorem 2, we instead take a simpler approach by leveraging Theorem 1 and the fact that in great generality, continuous local martingales, such as Itô integrals, can be represented as time-changed Brownian motion.

Proof of Theorem 2. We work with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ supporting a standard Brownian motion B on $\mathbb{R}^K$, with natural filtration $\mathcal{F}_t^B = \sigma(B(u) : 0 \le u \le t)$. We will work with the corresponding augmented filtration $\mathcal{F}_t = \sigma(\mathcal{F}_t^B \cup \mathcal{N})$, where $\mathcal{N}$ is the collection of all $\mathbb{P}$-null sets. (See Chapter 2.7 of Karatzas and Shreve (1998) for details.) By Theorem 1, there exists a solution (R, S) to the SDE (22)-(23) on this probability space with respect to the standard Brownian motion B. Writing (23) in integral form, because the p_k functions are bounded,

$$S_k(t) = \int_0^t \sqrt{p_k(R(u), S(u))} dB_k(u), \quad k \in [K]$$

are continuous $\mathcal{F}_t$-martingales with quadratic variation processes

$$\langle S_k \rangle_t = \int_0^t p_k(R(u), S(u)) du, \quad k \in [K],$$

and for $k \neq k'$, the cross-variation processes $\langle S_k, S_{k'} \rangle_t = 0$ since B_k and $B_{k'}$ are independent. Note that integrating (22) (in the Riemann sense) yields $\langle S_k \rangle_t = R_k(t)$, $k \in [K]$, which are continuous and strictly increasing processes since the p_k functions are bounded and strictly positive. Define

$$R_k^{-1}(t) = \inf\{u \ge 0 : R_k(u) \ge t\}, \quad k \in [K].$$

Now, we recall that in great generality, continuous martingales can be represented as time-changed Brownian motions. In particular, by a theorem due to F.B. Knight (see for instance Proposition 18.8 on page 355 of Kallenberg (2002) or Theorem 1.10 on page 183 of Revuz and Yor (1999)), for $k \in [K]$, we have that $\widetilde{B}_k(t) := S_k(R_k^{-1}(t))$ are independent standard Brownian motions (at least until time $t = R_k(1)$) with respect to the filtration $\mathcal{F}_t^{\widetilde{B}} = \sigma\left(\widetilde{B}(u) : 0 \le u \le t\right)$. Thus, we have $\widetilde{B}_k(R_k(t)) = S_k(t)$, and substituting this representation into the SDE (22), we obtain the stochastic ODE:

$$R_k(t) = \int_0^t p_k(R(u), \widetilde{B} \circ R(u)) du, \quad k \in [K]. \tag{79}$$

So with respect to the smaller filtration $\mathcal{F}_t^{\widetilde{B}}$, the SDE solution $R(t)$ satisfies the stochastic ODE (79), which coincides with (35). □

4. Further Insights from Diffusion Approximations

4.1. Model Mis-specification

In this section, we show that under the diffusion scaling of Assumption 1, the regret behavior of TS, as determined by the pre-limit R^n and limit R processes in Theorems 1-3, is robust to mis-specification of the reward distributions. (Recall the connection between regret and the R^n and R processes in Remark 2.) Asymptotically, under diffusion scaling, only the limiting means and variances (as in (1)-(2)) of the reward distributions affects the dynamics of TS. So, in Theorems

1-3, mis-specification corresponds to mis-match between the limiting variances σ_k^2 and the variance c_*^2 specified in the Gaussian likelihood of TS.

In Proposition 2, which is proved in Appendix A, we show that the regret behavior of Gaussian TS is continuous in $\sigma := (\sigma_k : k \in [K])$. This contrasts with the results in the instance-dependent Lai-Robbins asymptotic regime (Lai and Robbins 1985). In that setting, as studied in Fan and Glynn (2021), the slightest amount of reward distribution mis-specification (e.g., setting the variance parameter of a bandit algorithm to be just slightly less than the true variance of the rewards), can cause the regret performance to sharply deteriorate.

Proposition 2. *Let $(\widetilde{R}^\sigma, \widetilde{S}^\sigma) = (\widetilde{R}_k^\sigma, \widetilde{S}_k^\sigma : k \in [K])$ denote either the solution to (22)-(24) in Theorem 1 (equivalently, (35)-(36) in Theorem 2) corresponding to the σ in (18), or the solution to (41)-(44) in Theorem 3 corresponding to the σ in (39). Then, for any continuous function $f : D^K[0,1] \to \mathbb{R}$, the mapping $\sigma \mapsto \mathbb{E}[f(\widetilde{R}^\sigma)]$ is continuous.*

In the following Section 4.2, we will show that TS designed for exponential family rewards can be approximated by Gaussian TS under diffusion scaling. This indicates that the robustness of TS under diffusion scaling to model mis-specification extends broadly to other settings.

4.2. Approximation by Gaussian Thompson Sampling

Here, we study non-Gaussian TS in the diffusion regime. In Proposition 3 below, we show that the dynamics of TS designed for exponential family models can be approximated by those of Gaussian TS (which has been our focus until now). The main step is to show that for TS, under suitable regularity conditions, posterior distributions can be approximated by Gaussian distributions.

For the current Section 4.2, we assume that the rewards come from an exponential family P^μ parameterized by mean $\mu \in \mathcal{I}$, where $\mathcal{I} \subset \mathbb{R}$ is a (small) bounded, open interval. The exponential family distributions have the form:

$$P^\mu(dx) = \exp(\theta(\mu) \cdot x - \Lambda(\mu))P(dx), \tag{80}$$

where P is a base distribution, $\theta(\mu) \in \mathbb{R}$ is the value of the tilting parameter resulting in a mean of μ, and Λ is the cumulant generating function. Corresponding to P^μ, the log-likelihood function (probability densities are defined with respect to Lebesgue measure on $\mathbb{R}$) is denoted by $l(\mu, x)$, and derivatives of $l(\mu, x)$ with respect to μ are denoted by $l'(\mu, x)$, $l''(\mu, x)$, etc. We use $\mathbb{E}^\mu[\cdot]$ to denote expectation with respect to distribution P^μ.

For the exponential family P^μ, we assume the following conditions C1-C2 hold. These conditions allow us to obtain a suitable Gaussian approximation for the posterior distribution of the mean, which is developed in Proposition 6 in Appendix B.

(C1) For each $\delta > 0$, there is an $\epsilon > 0$ such that for all $\mu \in \mathcal{I}$,

$$\sup_{z:|\mu - z| \geq \delta} \mathbb{E}^{\mu}[l(z, X)] \leq \mathbb{E}^{\mu}[l(\mu, X)] - \epsilon. \tag{81}$$

(C2) There exists functions η and κ such that for all x in the support of P,

$$\eta(x) \geq \sup_{\mu \in \mathcal{I}} |l'(\mu, x)| \tag{82}$$

$$\kappa(x) \geq \sup_{\mu \in \mathcal{I}} |l'''(\mu, x)|. \tag{83}$$

Moreover, for the cases: $f(x) = |x|$, $f(x) = \eta(x) + |l(\mu_0, x)|$ with some fixed $\mu_0 \in \mathcal{I}$, and $f(x) = \kappa(x)$,

$$\lim_{y \to \infty} \sup_{\mu \in \mathcal{I}} \mathbb{E}^{\mu} \left[f(X) \mathbb{I} \left(f(X) > y \right) \right] = 0. \tag{84}$$

For the current Section 4.2, we also assume that Assumption 1 holds. In particular, for the distributions Q_k^n with means μ_k^n from Assumption 1, we have $Q_k^n = P^{\mu_k^n}$, with all $\mu_k^n \in \mathcal{I}$. For the σ_k in (2), here we have $\sigma_k = \sigma_*$ for all k, where σ_*^2 is the variance of P^{μ_*}, with the μ_* from (1). For TS designed for the exponential family P^{μ}, $\mu \in \mathcal{I}$ (using the corresponding likelihood), we use any prior with positive, continuous density in a neighborhood of μ_*. For simplicity, we assume that the prior is the same and independent for every arm and does not change with n in the triangular array setting of Assumption 1. The above setup leads to the following Proposition 3 for TS in the diffusion regime with exponential family reward distributions. The proof of Proposition 3 can be found in Appendix A.

Proposition 3. *Consider a K-armed bandit under the diffusion scaling of Assumption 1, and suppose also that the reward distributions belong to an exponential family of the form in (80) and satisfy C1-C2. Consider TS designed for the likelihood of the exponential family, and that uses any fixed prior having positive, continuous density in a neighborhood of μ_*. For any fixed $\epsilon \in (0,1)$, suppose the arms are sampled with constant, positive probabilities $(\overline{p}_k : k \in [K])$ for the first $\lfloor \epsilon n \rfloor$ periods, after which TS is run. Then, for the exponential family version of TS, we have weak convergence to the limits in (40)-(45) of Theorem 3, with*

$$p_k(r, s) = \mathbb{P} \left(k = \arg\max_{l \in [K]} \left\{ \frac{s_l \sigma_*}{r_l} + d_l + \frac{\sigma_*}{\sqrt{r_l}} N_l \right\} \right), \tag{85}$$

where the probability is taken over the independent standard Gaussian variables N_l.

The conclusion of Proposition 3 (with the $p_k(r,s)$ in (85)) matches that of Theorem 3 (with the $p_k(r,s)$ in (39)) when, in the context of Theorem 3, the limiting variances σ_k^2 in (2) match the variance c_*^2 used in the Gaussian likelihood of TS. Our results here also indicate that under diffusion scaling, Gaussian TS is a good approximation for many other variants of TS, including ones involving approximations of the posterior distribution, for example, via Laplace approximation. Since Gaussian TS is known to have optimal or near-optimal expected regret performance in a wide range of settings, this suggests that bandit algorithms based on Gaussian posterior approximation can perform similarly well under diffusion scaling. See Chapelle and Li (2011) and Chapter 5 of Russo et al. (2019) for discussions of such approximations.

4.3. Bootstrap-based Exploration

The bootstrap, as well as related ideas such as subsampling, has recently been proposed for exploration in bandit problems (Baransi et al. 2014, Eckles and Kaptein 2014, Osband and Van Roy 2015, Tang et al. 2015, Elmachtoub et al. 2017, Vaswani et al. 2018, Kveton et al. 2019a,b, Russo et al. 2019, Kveton et al. 2020b,a, Baudry et al. 2020). In this section, we consider the most basic implementation of bootstrapping for exploration in bandit problems, where in each time period, a single bootstrapped mean is sampled for each arm, and the arm with the greatest bootstrapped mean is played. We will refer to this as boostrap-based exploration.

In Proposition 4 below, we show that for general reward distributions, the dynamics of bootstrap-based exploration can be approximated by those of Gaussian TS. This is similar in spirit to Proposition 3. But unlike in Proposition 3, here the reward distributions do not need to belong to any exponential family—we only need the diffusion scaling conditions of Assumption 1. The proof of Proposition 4 is the same as that of Proposition 3, except we use a Gaussian approximation for the bootstrapped mean, which is developed in Proposition 7 in Appendix C.

PROPOSITION 4. *Consider a K-armed bandit under the diffusion scaling of Assumption 1, and suppose also that*

$$\lim_{y\to\infty} \sup_{\mu\in\mathcal{I}} \mathbb{E}^{\mu}[X^2 \mathbb{I}\left(X^2 > y\right)] = 0. \tag{86}$$

For any fixed $\epsilon \in (0,1)$, suppose the arms are sampled with constant, positive probabilities $(\overline{p}_k : k \in [K])$ for the first $\lfloor \epsilon n \rfloor$ periods, after which bootstrap-based sampling is run. Then, for bootstrap-based sampling, we have weak convergence to the limits in (40)-(45) of Theorem 3, with

$$p_k(r,s) = \mathbb{P}\left(k = \arg\max_{l\in[K]} \left\{ \frac{s_l \sigma_l}{r_l} + d_l + \frac{\sigma_l}{\sqrt{r_l}} N_l \right\}\right). \tag{87}$$

Compared to Theorem 3 (with the $p_k(r,s)$ in (39)), in Proposition 4 (with the $p_k(r,s)$ in (87)), bootstrap-based sampling automatically adapts to the (limiting) variance σ_k^2 for each arm k, rather than having to specify some variance c_*^2 as in Gaussian TS. This is reflected in the $\frac{\sigma_l}{\sqrt{r_l}} N_l$ terms in (87)), compared to the $\frac{c_*}{\sqrt{r_l}} N_l$ terms in (39). Our results here suggest that bootstrap-based sampling can be an effect means of balancing exploration and exploitation, which does not require much tuning under diffusion scaling.

4.4. Batched Updates

In some settings, it may be impractical to update a bandit algorithm after each time period. Instead, updates are "batched" so that the algorithm commits to playing an (adaptively determined) arm for an interval of time (which can also be adaptively determined). Then, the algorithm is updated all at once with the data collected during the interval. For a time horizon of n, suppose the batch sizes pre-determined before the start of the experiment and are $o(n)$. Then, under diffusion scaling, we would obtain weak convergence to the same diffusion limits (SDEs and stochastic ODEs) as in the case of ordinary (non-batched) TS. Indeed, a time interval of $o(n)$ in the discrete pre-limit system corresponds to (after dividing by n) an infinitesimally small time interval in the continuous limit system. This suggests that as long as the number of batches increases to infinity (possibly at an arbitrarily slow rate) as $n \to \infty$, and each batch is not too large (at most $o(n)$ periods), then the distribution of regret will be approximately the same compared to the case in which one updates in every period (batch sizes of one). To make this precise, we have the following proposition, whose straightforward proof is omitted.

PROPOSITION 5. *In the settings of Theorems 1, 2 and 3, the same conclusions hold for TS with batches of size $o(n)$.*

The discussion and proposition above correspond nicely to results in the literature regarding optimal batching for bandits in the minimax gap regime from the perspective of expected regret. As shown in Cesa-Bianchi et al. (2013), Perchet et al. (2016) and Gao et al. (2019), in the minimax regime, a relatively tiny, $O(\log\log(n))$, number of batches is necessary and sufficient (sufficient for specially designed algorithms) to achieve the optimal order of expected regret.

Appendix A: Additional Proofs for Sections 2 and 4

Proof of Proposition 1. The proof is a simple modification of the proof of Theorem 1. We start with the discrete approximation (31)-(32) and (25)-(26) from our derivation in Section 2.3. We denote the joint processes via $(R^n, Z^n, M^n) = (R^n_k, Z^n_k, M^n_k : k \in [K])$, and recall that they are processes in $D^{3K}[0,1]$.

Consider a weakly convergent subsequence of (R^n, Z^n), which we will still index by n for notational simplicity. Then jointly, $(R^n, Z^n, M^n) \Rightarrow (R, Z, M)$, where (as in the proof of Theorem 1) M is the $D^K[0,1]$ zero process. By Donsker's Theorem (see Chapter 3 of Billingsley (1999)), Z is a standard Brownian motion on $\mathbb{R}^K$.

By the continuity of function composition (see Theorem 13.2.2 on page 430 of Whitt (2002)), since the Brownian motion limit process Z has continuous sample paths and the limit process R must have non-decreasing sample paths, we have by the Continuous Mapping Theorem,

$$(R^n, Z^n, M^n, Z^n \circ R^n) \Rightarrow (R, Z, M, Z \circ R). \tag{88}$$

Now define the processes $G^n = (G_k^n : k \in [K])$ and $G = (G_k : k \in [K])$, where

$$G_k^n(t) = p_k^n(R^n(t), Z^n \circ R^n(t)) \tag{89}$$

$$G_k(t) = p_k(R(t), Z \circ R(t)). \tag{90}$$

By assumption, $p_k^n(r,s) \to p_k(r,s)$ as $n \to \infty$ uniformly for (r,s) in compact subsets of $[0,1]^K \times \mathbb{R}^K$, and $p_k(r,s)$ is continuous at all $(r,s) \in [0,1]^K \times \mathbb{R}^K$. Therefore, by the Generalized Continuous Mapping Theorem (see Lemma 7) applied to the processes in (89)-(90), we have from (88),

$$(R^n, Z^n, M^n, G^n) \Rightarrow (R, Z, M, G). \tag{91}$$

Additionally, define the processes $\widetilde{R}^n = (\widetilde{R}_k^n : k \in [K])$ and $\widetilde{R} = (\widetilde{R}_k : k \in [K])$, where

$$\begin{aligned} \widetilde{R}_k^n(t) &= \int_0^t p_k^n(R^n(u), Z^n \circ R^n(u)) du \\ \widetilde{R}_k(t) &= \int_0^t p_k(R(u), Z \circ R(u)) du. \end{aligned} \tag{92}$$

Recall that

$$R_k^n(t) = \frac{1}{n} \sum_{i=1}^{\lfloor nt \rfloor} p_k^n(R^n(\tfrac{i-1}{n}), Z^n \circ R^n(\tfrac{i-1}{n})) + M_k^n(t).$$

For each k, because M_k^n converges weakly to the $D[0,1]$ zero process, and

$$\left| \frac{1}{n} \sum_{i=1}^{\lfloor nt \rfloor} p_k^n(R^n(\tfrac{i-1}{n}), Z^n \circ R^n(\tfrac{i-1}{n})) - \widetilde{R}_k^n(t) \right| \le \frac{1}{n},$$

we have

$$\sup_{0 \le t \le 1} \left| R_k^n(t) - \widetilde{R}_k^n(t) \right| \xrightarrow{\mathbb{P}} 0. \tag{93}$$

Thus, by the fact that integration is a continuous functional with respect to the Skorohod metric (see Theorem 11.5.1 on page 383 of Whitt (2002)) and the Continuous Mapping Theorem, we have from (91),

$$(R^n, Z^n, \widetilde{R}^n) \Rightarrow (R, Z, \widetilde{R}). \tag{94}$$

Together, (93)-(94) yield

$$(R^n, Z^n, R^n) \Rightarrow (R, Z, \widetilde{R}),$$

and recalling the definition of $\widetilde{R}$ in (92), the proof is complete. □

Proof of Proposition 3. The TS dynamics of the initial $\lfloor \epsilon n \rfloor$ periods in the pre-limit, equivalently the initial $[0, \epsilon)$ interval in the limit, are the same as in Theorem 3. To obtain the desired SDE and stochastic ODE approximations on $[\epsilon, 1]$, we verify that the TS sampling probabilities for the general exponential family setting have the desired form with $p_k(r,s)$ as in (85).

In the SDE case, at time $j+1$, having collected information $\mathcal{G}_j^n$ (as defined in (4)), for each arm k, we sample a value $\widetilde{\mu}_k^n(j+1)$ from the posterior distribution of μ_k^n. Let $\widehat{\mu}_k^n(j+1)$ denote the sample mean estimate at time $j+1$. (Note that the sample mean is also the statistically efficient maximum likelihood estimator for the mean, and is used as the centering value for the Gaussian posterior approximation in Proposition 6.) Here, the S_k^n and R_k^n have the expressions from (6)-(7). The probability of playing arm k is given by:

$$
\begin{aligned}
&\mathbb{P}\big(k = \arg\max_{l\in[K]} \widetilde{\mu}_l^n(j+1) \,\big|\, \mathcal{G}_j^n\big) \\
&= \mathbb{P}\left(k = \arg\max_{l\in[K]} \left\{ \frac{S_l^n(\frac{j}{n})\sigma_l^n}{R_l^n(\frac{j}{n})} + d_l^n + \sqrt{n}\,(\widetilde{\mu}_l^n(j+1) - \widehat{\mu}_l^n(j+1)) \right\} \,\middle|\, R^n(\frac{j}{n}), S^n(\frac{j}{n}) \right) \\
&= \mathbb{P}\left(k = \arg\max_{l\in[K]} \left\{ \frac{S_l^n(\frac{j}{n})\sigma_l^n}{R_l^n(\frac{j}{n})} + d_l^n + \frac{\sigma_l^n}{\sqrt{R_l^n(\frac{j}{n})}} N_l \right\} \,\middle|\, R^n(\tfrac{j}{n}), S^n(\tfrac{j}{n}) \right) + o_{\mathbb{P}}(1) &(95)\\
&= p_k^n(R^n(\tfrac{j}{n}), S^n(\tfrac{j}{n})) + o_{\mathbb{P}}(1), &(96)
\end{aligned}
$$

where (95) follows from Proposition 6, the probability is taken over the independent standard Gaussian variables N_l, and

$$p_k^n(r,s) = \mathbb{P}\left(k = \arg\max_{l\in[K]} \left\{ \frac{s_l\sigma_l^n}{r_l} + d_l^n + \frac{\sigma_l^n}{\sqrt{r_l}} N_l \right\} \right).$$

Moreover, with $p_k(r,s)$ as in (85), $p_k^n(r,s) \to p_k(r,s)$ uniformly for (r,s) on compact subsets of $[0,1]^K \times \mathbb{R}^K$ with the restriction that $r_k \geq \overline{p}_k > 0$ for all k.

This sequence of derivations parallels what we derived in (9)-(11) in Section 2.2. From (96), the proof of Theorem 1 can be applied to yield the desired SDE approximation in (41)-(44).

The proof in the stochastic ODE case only differs superficially. We would use the filtration $\mathcal{H}_j^n$, as defined in (5), instead of $\mathcal{G}_j^n$. And we would use $Z_k^n(R_k^n)$, as defined in (25), instead of S_k^n. The proof of Theorem 2 can then be applied to yield the desired stochastic ODE approximation in (45). $\square$

Appendix B: Gaussian Approximations for Posterior Distributions

In this section of the appendix, we use the same setup as in Section 4.2. Recall that the model is an exponential family P^μ parameterized by mean $\mu \in \mathcal{I}$, where $\mathcal{I} \subset \mathbb{R}$ is a (small) bounded, open interval. For each $\mu \in \mathcal{I}$, $X_i^\mu \overset{\text{iid}}{\sim} P^\mu$. The sample mean computed using $X_1^\mu, \ldots, X_n^\mu$ is denoted $\widehat{m}_n^\mu$. Given such n samples, we use $\widetilde{m}_n^\mu$ to denote a sample from the posterior of the mean μ.

Here, our goal is to develop Proposition 6, which is a Bernstein-von Mises Gaussian approximation of the posterior distribution. Although stronger than needed for the applications in Sections 4.2 and 4.1, the result holds almost surely and uniformly over data-generating distributions P^μ with $\mu \in \mathcal{I}$. To make sense of this mode of convergence, we first recall an equivalent characterization of almost sure convergence in Remark 6 below, followed by a precise definition of the mode of convergence in Definition 2 below.

Remark 6. For a sequence of random variables Y_i,

$$\mathbb{P}\left(\lim_{n\to\infty} Y_n = 0\right) = 1 \tag{97}$$

if and only if for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{j\geq n} |Y_j| > \epsilon\right) = 0. \tag{98}$$

Definition 2. Let $\mathcal{P}$ be a collection of probability distributions and Z_i be random variables defined on the probability spaces $(\Omega, \mathcal{F}, P)_{P\in\mathcal{P}}$. We say that the sequence Z_i converges almost surely to zero, uniformly in $P \in \mathcal{P}$, if for any $\epsilon > 0$,

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \mathbb{P}_P \left(\sup_{j\geq n} |Z_j| > \epsilon \right) = 0. \tag{99}$$

Next, we state Lemmas 5 and 6, which are used in the proof of Proposition 6. From Chung (1951), we have a strong law of large numbers which holds uniformly over a collection of underlying probability distributions.

Lemma 5. *Let $\mathcal{P}$ be a collection of probability distributions and Y_i be iid random variables defined on the probability spaces $(\Omega, \mathcal{F}, P)_{P\in\mathcal{P}}$ satisfying $\mathcal{P}$-uniform integrability condition:*

$$\lim_{z\to\infty} \sup_{P\in\mathcal{P}} \mathbb{E}_P \left[|Y - \mathbb{E}_P[Y]| \, \mathbb{I}\left(|Y - \mathbb{E}_P[Y]| > z \right) \right] = 0.$$

Then, for every $\epsilon > 0$,

$$\lim_{z\to\infty} \sup_{P\in\mathcal{P}} \mathbb{P}_P \left(\sup_{n\geq z} \left| \frac{1}{n} \sum_{i=1}^{n} Y_i - \mathbb{E}_P[Y] \right| > \epsilon \right) = 0.$$

Using Theorems 2.7.11 and 2.8.1 of van der Vaart and Wellner (1996) and the mean value theorem, we have the following result.

Lemma 6. *Suppose C1 holds together with C2 for the case $f(x) = \eta(x) + |l(\mu_0, x)|$ with some fixed $\mu_0 \in \mathcal{I}$ and $\eta(x)$ as defined in (82). Then, $\{l(\mu, \cdot), \mu \in \mathcal{I}\}$ is a Glivenko-Cantelli class of functions uniformly in P^μ, $\mu \in \mathcal{I}$.*

We now state and prove the main result, Proposition 6. The proof is adapted from the proof of Theorem 4.2 in Ghosh et al. (2006).

Proposition 6. *Let $\mathcal{I} \subset \mathbb{R}$ be any bounded open interval, and let ν_0 be a prior density that is positive on $\mathcal{I}$. Suppose conditions C1-C2 hold. Then, the centered and scaled posterior density $t \mapsto \nu_n(t \mid X_1^\mu, \ldots, X_n^\mu)$ for $\sqrt{n}(\widetilde{m}_n^\mu - \widehat{m}_n^\mu)$ satisfies:*

$$\lim_{n\to\infty} \int_{\mathbb{R}} \left| \nu_n(t \mid X_1^\mu, \ldots, X_n^\mu) - \frac{1}{\sqrt{2\nu}\sigma^\mu} \exp\left(-\frac{1}{2(\sigma^\mu)^2} t^2 \right) \right| dt = 0 \tag{100}$$

almost surely, uniformly in $\mu \in \mathcal{I}$.

Proof of Proposition 6. To begin, note that the posterior density can be written as

$$\nu_n(t \mid X_1^\mu, \ldots, X_n^\mu) = (C_n^\mu)^{-1} \nu_0(\widehat{m}_n^\mu + t/\sqrt{n}) \exp\left(L_n^\mu(\widehat{m}_n^\mu + t/\sqrt{n}) - L_n^\mu(\widehat{m}_n^\mu) \right), \tag{101}$$

with normalization factor $(C_n^\mu)^{-1}$ and

$$L_n^\mu(y) = \sum_{i=1}^{n} l(y, X_i^\mu).$$

Let

$$D_n^\mu(t) = \nu_0(\widehat{m}_n^\mu + t/\sqrt{n}) \exp\left(L_n^\mu(\widehat{m}_n^\mu + t/\sqrt{n}) - L_n^\mu(\widehat{m}_n^\mu) \right) - \nu_0(\mu) \exp\left(-\frac{1}{2(\sigma^\mu)^2} t^2 \right). \tag{102}$$

To show (100), it suffices to show that a.s. uniformly in μ,

$$\lim_{n\to\infty}\int_{\mathbb{R}}|D_n^\mu(t)|\,dt=0. \tag{103}$$

If (103) holds, then a.s. uniformly in μ,

$$\lim_{n\to\infty}\left(C_n^\mu-\nu_0(\mu)\sqrt{2\nu}\sigma^\mu\right)=0.$$

So we would have

$$\begin{aligned}&\int_{\mathbb{R}}\left|\nu_n(t\mid X_1^\mu,\ldots,X_n^\mu)-\frac{1}{\sqrt{2\nu}\sigma^\mu}\exp\left(-\frac{1}{2(\sigma^\mu)^2}t^2\right)\right|dt\\&\le (C_n^\mu)^{-1}\int_{\mathbb{R}}|D_n^\mu(t)|\,dt+\left|(C_n^\mu)^{-1}\nu_0(\mu)-\frac{1}{\sqrt{2\nu}\sigma^\mu}\right|\int_{\mathbb{R}}\exp\left(-\frac{1}{2(\sigma^\mu)^2}t^2\right)dt,\end{aligned}$$

and the proof of Proposition 6 would be complete.

To show (103), we consider two cases: $A_n=\{t:|t|>\gamma\sqrt{n}\}$ and $A_n^c=\{t:|t|\le\gamma\sqrt{n}\}$, where we will set $\gamma>0$ later. For the first case involving the set A_n, note that

$$\begin{aligned}\int_{A_n}|D_n^\mu(t)|\,dt\le&\int_{A_n}\nu_0(\widehat{m}_n^\mu+t/\sqrt{n})\exp\left(L_n^\mu(\widehat{m}_n^\mu+t/\sqrt{n})-L_n^\mu(\widehat{m}_n^\mu)\right)dt\\&+\int_{A_n}\nu_0(\mu)\exp\left(-\frac{1}{2(\sigma^\mu)^2}t^2\right)dt.\end{aligned} \tag{104}$$

The second integral in (104) goes to zero as $n\to\infty$, uniformly in μ. As for the first integral, note that from condition C2 with $f(x)=|x|$ and Lemma 5 that

$$\widehat{m}_n^\mu-\mu\to 0 \tag{105}$$

as $n\to\infty$, a.s. uniformly in μ. This along with Lemma 6 implies that there exists $\epsilon>0$ such that for sufficiently large n, a.s. uniformly in μ,

$$\frac{1}{n}\left(L_n^\mu(\widehat{m}_n^\mu+t/\sqrt{n})-L_n^\mu(\widehat{m}_n^\mu)\right)\le-\epsilon$$

on the set A_n. Therefore, the first integral in (104) also goes to zero as $n\to\infty$, a.s. uniformly in μ.

For the second case involving the set A_n^c, we expand L_n^μ in a Taylor series about the MLE $\widehat{m}_n^\mu$, noting that by the definition of the MLE, $(L_n^\mu)'(\widehat{m}_n^\mu)=0$. We have

$$\begin{aligned}L_n^\mu(\widehat{m}_n^\mu+t/\sqrt{n})-L_n^\mu(\widehat{m}_n^\mu)&=-\frac{1}{2}\frac{1}{n}(L_n^\mu)''(\widehat{m}_n^\mu)t^2+r_n^\mu(t)\\&=-\frac{1}{2}\left(\theta''(\widehat{m}_n^\mu)\widehat{m}_n^\mu-\Lambda''(\widehat{m}_n^\mu)\right)t^2+r_n^\mu(t),\end{aligned} \tag{106}$$

using the fact that $l''(\mu,x)=\theta''(\mu)\cdot x-\Lambda''(\mu)$, and where

$$r_n^\mu(t)=\frac{1}{6}\left(\frac{t}{\sqrt{n}}\right)^3(L_n^\mu)'''(m_{n,t}^\mu),$$

with $m_{n,t}^\mu$ being a point in between $\widehat{m}_n^\mu$ and $\widehat{m}_n^\mu+t/\sqrt{n}$. Using condition C2 with $f(x)=\kappa(x)$ and Lemma 5, for sufficiently large n, a.s. uniformly in μ,

$$|r_n^\mu(t)|\le\frac{1}{6}\frac{t^3}{\sqrt{n}}\frac{1}{n}\sum_{i=1}^n\kappa(X_i^\mu)\le\frac{1}{3}\frac{t^3}{\sqrt{n}}\mathbb{E}\left[\kappa(X_1^\mu)\right], \tag{107}$$

and so for fixed t, we have $r_n^\mu(t) \to 0$ as $n \to \infty$, a.s. uniformly in μ. On the set A_n^c, we have $t/\sqrt{n} \le \gamma$, and so (107) can be re-written as

$$|r_n^\mu(t)| \le \frac{1}{3}\gamma t^2 \mathbb{E}\left[\kappa(X_1^\mu)\right]. \tag{108}$$

For the right side of (106), we have

$$\lim_{n\to\infty} \theta''(\widehat{m}_n^\mu)\widehat{m}_n^\mu - \Lambda''(\widehat{m}_n^\mu) = \frac{1}{(\sigma^u)^2} \tag{109}$$

a.s. uniformly in μ, which follows from the Cramér-Rao asymptotic lower bound for the variance of the MLE. Taking $\gamma > 0$ to be sufficiently small, and using the fact that $\inf_{\mu\in\mathcal{I}} 1/\sigma^\mu > 0$, we have from (106) that on the set A_n^c,

$$\exp\left(L_n^\mu(\widehat{m}_n^\mu + t/\sqrt{n}) - L_n^\mu(\widehat{m}_n^\mu)\right) \le \exp(-at^2) \tag{110}$$

for some fixed $a > 0$ and sufficient large n, a.s. uniformly in μ. Thus, using (105), $D_n^\mu(t)$ is dominated by an integrable function on the set A_n^c for sufficiently large n, a.s. uniformly in μ. Recall that $r_n^\mu(t) \to 0$ as $n \to \infty$, a.s. uniformly in μ, and so together with (109), we have $D_n^\mu(t) \to 0$ for each fixed t as $n \to \infty$, a.s. uniformly in μ. The dominated convergence theorem then yields

$$\int_{A_n^c} |D_n^\mu(t)|\, dt \to 0$$

as $n \to \infty$, a.s. uniformly in μ. $\square$

Appendix C: Gaussian Approximations for the Bootstrap

In Proposition 7 below, we develop a Gaussian approximation for bootstrapping the mean. Here, we allow for arbitrary reward distributions P^μ with means $\mu \in \mathcal{I}$ (not necessarily from an exponential family), where $\mathcal{I} \subset \mathbb{R}$ is a (small) bounded open interval. The only requirement on the P^μ is that the condition in (111) is satisfied. For each $\mu \in \mathcal{I}$, we use $\widehat{m}_n^{*\mu}$ to denote a bootstrap of the sample mean $\widehat{m}_n^\mu$ computed using n samples $X_i^\mu \overset{\text{iid}}{\sim} P^\mu$, $i = 1, \ldots, n$. Proposition 7 holds almost surely and uniformly over data-generating distributions P^μ with $\mu \in \mathcal{I}$. See Remark 6 and Definition 2 in Appendix C for a precise description of this mode of convergence.

Proposition 7. *Suppose that*

$$\lim_{y\to\infty} \sup_{\mu\in\mathcal{I}} \mathbb{E}^\mu[X^2 \mathbb{I}\left(X^2 > y\right)] = 0. \tag{111}$$

Then,

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left| \mathbb{P}\left(\sqrt{n}\,(\widehat{m}_n^{*\mu} - \widehat{m}_n^\mu) \le x \mid X_1^\mu, \ldots, X_n^\mu\right) - \Phi\left(\frac{x}{\sigma^\mu}\right) \right| = 0, \tag{112}$$

almost surely, uniformly in $\mu \in \mathcal{I}$.

Proof of Proposition 7. We check the conditions to be able to apply Proposition 1.3.1 part (ii) in Politis et al. (1999). First, because the class of functions $\{\mathbb{I}\,(\cdot \le x), x \in \mathbb{R}\}$ is a VC class and is uniformly bounded, it is a uniform Glivenko-Cantelli class by Theorem 2.8.1 of van der Vaart and Wellner (1996). Also, from (111) and Lemma 5, we have $\widehat{m}_n^\mu \to \mu$ and $\widehat{\sigma}_n^\mu \to \sigma^\mu$ as $n \to \infty$, almost surely, uniformly in μ. The desired result (112) then follows. $\square$

Appendix D: Weak Convergence Technical Lemmas

Lemma 7 **(Generalized Continuous Mapping Theorem)**. *Let f and f^n, $n \geq 1$, be measurable functions that map from the metric space (S_1, m_1) to the separable metric space (S_2, m_2). Let E be the set of $x \in S_1$ such that $f^n(x^n) \to f(x)$ fails for some sequence x^n, $n \geq 1$, with $x^n \to x$ in S_1. If $\xi^n \Rightarrow \xi$ in (S_1, m_1) and $P(\xi \in E) = 0$, then $f^n(\xi^n) \Rightarrow g(\xi)$ in (S_2, m_2). (See Theorem 3.4.4 on page 86 of Whitt (2002).)*

Lemma 8 **(Tightness of Multi-dimensional Processes)**. *A sequence of process $\xi^n = (\xi^n_1, \ldots, \xi^n_d)$ is tight in $D^d[0,1]$ if and only if each ξ^n_j and each $\xi^n_j + \xi^n_k$ are tight in $D[0,1]$, for all $1 \leq j, k \leq d$. (See Problem 22 of Chapter 3 on page 153 of Ethier and Kurtz (1986).)*

Lemma 9 **(Simple Sufficient Conditions for Tightness)**. *A sequence of processes ξ^n in $D[0,1]$ adapted to filtrations $\mathcal{F}^n_t$ is tight if*

$$\lim_{a\to\infty} \sup_n \mathbb{P}\left(\sup_{0\leq t\leq 1} |\xi^n(t)| > a\right) = 0, \tag{T1}$$

and for any $\delta > 0$, there exists a collection of non-negative random variables $\alpha_n(\delta)$ such that

$$\mathbb{E}\left[(\xi^n(t+u) - \xi^n(t))^2 \mid \mathcal{F}^n_t\right] \leq \mathbb{E}\left[\alpha_n(\delta) \mid \mathcal{F}^n_t\right] \tag{T2}$$

almost surely for $0 \leq t \leq 1$ and $0 \leq u \leq \delta \wedge (1-t)$, and

$$\lim_{\delta\downarrow 0} \limsup_{n\to\infty} \mathbb{E}\left[\alpha_n(\delta)\right] = 0. \tag{T3}$$

(See Lemma 3.11 from Whitt (2007), which is adapted from Ethier and Kurtz (1986).)

Lemma 10 **(Martingale Functional Central Limit Theorem for Triangular Arrays)**. *For each n, let $Y^n(i) \in \mathbb{R}$ be a martingale difference sequence adapted to the filtration $\mathcal{F}^n_i$ for $i = 1, \ldots, n$. Suppose the following hold.*

For each $t \in [0,1]$, as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{\lfloor nt\rfloor} \mathbb{E}\left[Y^n(i)^2 \mid \mathcal{F}^n_{i-1}\right] \xrightarrow{\mathbb{P}} t. \tag{M1}$$

For any $\epsilon > 0$, as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[Y^n(i)^2 \mathbb{I}\left(|Y^n(i)| > \epsilon\sqrt{n}\right) \mid \mathcal{F}^n_{i-1}\right] \xrightarrow{\mathbb{P}} 0. \tag{M2}$$

Then, in $D[0,1]$ as $n \to \infty$,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor n\cdot\rfloor} Y^n(i) \to B(\cdot),$$

where B is standard Brownian motion. (See, for example, Theorem 8.2.4 of Durrett (2019).)

References

Agrawal S, Goyal N (2012) Analysis of Thompson Sampling for the Multi-armed Bandit Problem. *Conference on Learning Theory* .

Agrawal S, Goyal N (2013) Further Optimal Regret Bounds for Thompson Sampling. *AISTATS* .

Agrawal S, Goyal N (2017) Near-optimal Regret Bounds for Thompson Sampling. *Journal of the ACM* 64(5):30:1–30:24.

Araman V, Caldentey R (2022) Diffusion Approximations for a Class of Sequential Experimentation Problems. *Management Science* 68(8):5958–5979.

Auer P, Cesa-Bianchi N, Fischer P (2002) Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning* 47:235–256.

Baransi A, Maillard O, Mannor S (2014) Sub-sampling for Multi-armed Bandits. *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* .

Baudry D, Kaufmann E, Maillard O (2020) Sub-sampling for Efficient Non-parametric Bandit Exploration. *Advances in Neural Information Processing Systems* .

Billingsley P (1999) *Convergence of Probability Measures* (Wiley).

Cesa-Bianchi N, Dekel O, Sharmi O (2013) Online Learning with Switching Costs and Other Adaptive Adversaries. *Advances in Neural Information Processing Systems* .

Chapelle O, Li L (2011) An Empirical Evaluation of Thompson Sampling. *Neural Information Processing Systems* 25.

Chung K (1951) The Strong Law of Large Numbers. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 341–353 (University of California Press).

Durrett R (2019) *Probability: Theory and Examples (Fifth Edition)* (Cambridge University Press).

Eckles D, Kaptein M (2014) Thompson Sampling with the Online Bootstrap. *arXiv:1410.4009* .

Elmachtoub A, McNellis R, Oh S, Petrik M (2017) A practical method for solving contextual bandit problems using decision trees. *Conference on Uncertainty in Artificial Intelligence* .

Ethier S, Kurtz T (1986) *Markov Processes: Characterization and Convergence* (Wiley).

Fan L, Glynn P (2021) The Fragility of Optimized Bandit Algorithms. *https://arxiv.org/abs/2109.13595* .

Gao Z, Han Y, Ren Z, Zhou Z (2019) Batched Multi-armed Bandits Problem. *Advances in Neural Information Processing Systems* .

Ghosh J, Delampady M, Samanta T (2006) *An Introduction to Bayesian Analysis: Theory and Methods* (Springer).

Kallenberg O (2002) *Foundations of Modern Probability* (Springer).

Kalvit A, Zeevi A (2021) A Closer Look at the Worst-case Behavior of Multi-armed Bandit Algorithms. *Neural Information Processing Systems* 35.

Karatzas I, Shreve S (1998) *Brownian Motion and Stochastic Calculus* (Springer).

Kaufmann E, Korda N, Munos R (2012) Thompson Sampling: An Asymptotically Optimal Finite-time Analysis. *International Conference on Algorithmic Learning Theory* 199–213.

Kuang X, Wager S (2023) Weak Signal Asymptotics for Sequentially Randomized Experiments. *Management Science* .

Kurtz T, Protter P (1991) Weak Limit Theorems for Stochastic Integrals and Stochastic Differential Equations. *The Annals of Probability* 19(3):1035–1070.

Kveton B, Manzil Z, Szepesvari C, Li L, Ghavamzadeh M, Boutilier C (2020a) Randomized Exploration in Generalized Linear Bandits. *Conference on Artificial Intelligence and Statistics* .

Kveton B, Szepesvari C, Ghavamzadeh M, Boutilier C (2019a) Perturbed History Exploration in Stochastic Multi-armed Bandits. *International Joint Conference on Artificial Intelligence* .

Kveton B, Szepesvari C, Ghavamzadeh M, Boutilier C (2020b) Perturbed History Exploration in Stochastic Linear Bandits. *Conference on Uncertainty in Artificial Intelligence* .

Kveton B, Szepesvari C, Vaswani S, Wen Z, Ghavamzadeh M, Lattimore T (2019b) Garbage In, Reward Out: Bootstrapping Exploration in Multi-armed Bandits. *International Conference on Machine Learning* .

Lai T, Robbins H (1985) Asymptotically Efficient Adaptive Allocation Rules. *Advances in Applied Mathematics* 6(1):4–22.

Lattimore T, Szepesvári C (2020) *Bandit Algorithms* (Cambridge University Press).

Osband I, Van Roy B (2015) Bootstrapped Thompson Sampling and Deep Exploration. *arXiv:1507.00300* .

Perchet V, Rigollet P, Chassang S, Snowberg E (2016) Batched Bandit Problems. *The Annals of Statistics* 44(2):660–681.

Politis D, Romano J, Wolf M (1999) *Subsampling* (Springer).

Pollard D (1984) *Convergence of Stochastic Processes* (Springer).

Revuz D, Yor M (1999) *Continuous Martingales and Brownian Motion* (Springer).

Russo D, Van Roy B (2014) Learning to Optimize via Posterior Sampling. *Mathematics of Operations Research* 39(4):1221–1243.

Russo D, Van Roy B (2016) An Information-Theoretic Analysis of Thompson Sampling. *Journal of Machine Learning Research* 17(68):1–30.

Russo D, Van Roy B, Kazerouni A, Osband I, Wen Z (2019) *A Tutorial on Thompson Sampling* (Foundations and Trends in Machine Learning).

Stroock D, Varadhan S (1979) *Multidimensional Diffusion Processes* (Springer).

Tang L, Jiang Y, Li L, Zeng C, Li T (2015) Personalized Recommendation via Parameter-free Contextual Bandits. *International ACM SIGIR Conference on Research and Development in Information Retrieval* .

Thompson W (1933) On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika* 25(3):285–294.

van der Vaart A, Wellner J (1996) *Weak Convergence and Empirical Processes* (Springer).

Vaswani S, Kveton B, Wen Z, Rao A, Schmidt M, Abbasi-Yadkori Y (2018) New Insights into Bootstrapping for Bandits. *arXiv:1805.09793* .

Whitt W (2002) *Stochastic-Process Limits* (Springer).

Whitt W (2007) Proofs of the Martingale FCLT. *Probability Surveys* 4:268–302.