Diffusion Approximations for Thompson Sampling in the Small Gap Regime

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 in the "small gap" regime from the perspective of weak convergence. The small gap regime is one in which the gaps between the arm means are of order $\sqrt{\gamma}$, where γ is small. When $\sqrt{\gamma}$ is small and the number of arm plays n is large and of order $1/\gamma$ or smaller, we show that the process-level dynamics of Thompson sampling can be approximated by solutions to appropriately defined stochastic differential equations (SDEs) and stochastic ordinary differential equations (ODEs). 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. In this regime, we also show that the weak limits of the dynamics of many sampling-based bandit algorithms—including Thompson sampling designed for single-parameter exponential family rewards, and non-parametric bandit algorithms based on bootstrap re-sampling—coincide with those of Gaussian parametric Thompson sampling with Gaussian priors. Moreover, in this regime, these algorithms are generally robust to model mis-specification.

Key words: Multi-armed bandits, regret distribution, weak convergence, Gaussian approximations, model mis-specification

History: Manuscript version – August 11, 2025

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 overwhelmingly 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 $\mathrm{Reg}(n) := \sum_k N_k(n)\Delta_k$ is the sum over each arm k of the number of times $N_k(n)$ it is played over horizon n, weighted by its mean reward sub-optimality gap $\Delta_k := \max_j \mu_j - \mu_k$, where μ_j is 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). The TS principle specifies that at any given time, an arm is played with probability equal to the posterior probability that its mean reward is the highest among all arms; a precise description of TS is provided in Section 2. Our specific interest is in studying the algorithm's behavior in the challenging "small gap" environment in which the sub-optimality gaps Δ_k are of order $\sqrt{\gamma}$, with $\gamma \downarrow 0$, and in which the total number n of arm plays is large and of order $1/\gamma$ (or smaller). Thus, this analysis provides insight into the algorithm's behavior when the number of arm plays n is not yet large enough to have confidently identified the optimal arm. Sending $\gamma \downarrow 0$, 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 settings which guide the design of 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 Δ_k scaling as $1/\sqrt{n}$ for time horizon n (Agrawal and Goyal 2013, 2017). In such settings, there is not enough reward information for bandit algorithms to fully distinguish 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, as mentioned above, 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.

Our main contributions in this paper are described in the two points below. In independent work, Kuang and Wager (2023) derived similar SDE and stochastic ODE characterizations for versions of TS based on posterior updating with Gaussian priors and likelihoods within a general framework for analyzing sampling-based bandit algorithms under diffusion scaling. However, directly compared to our two main contributions, 1) their weak convergence theory invokes rather abstract theory based on infinitesimal generators and 2) they do not develop the general invariance principles and accompanying insights that we do. We provide a more detailed comparison of our work to theirs in Section 1.1.

1) Under diffusion scaling, we develop distributional approximations for the process-level dynamics of the *Gaussian Thompson sampler*, which is the TS principle implemented using the posterior updating mechanics of Gaussian priors and likelihoods. The limit dynamics of the Gaussian Thompson sampler have an SDE representation and also an equivalent stochastic ordinary differential

equation (ODE) representation; see Theorems 1, 2 and 3. These diffusion approximations only require that the centered and suitably re-scaled reward processes converge weakly to Brownian motion, and do not require the rewards themselves to be Gaussian (or even necessarily iid). Crucially, our proof approach for these theorems is transparent and explicitly shows how the SDE and stochastic ODE weak limits arise. Specifically, we start with discrete-time equations describing the evolution of the Gaussian Thompson sampler, and then pass to the limit using the Continuous Mapping Theorem and elementary arguments to obtain the SDEs and stochastic ODEs. We provide intuitive sketches of our proof approach in Sections 3.1-3.2.

2) We also develop diffusion approximations for other versions of TS and related sampling-based bandit algorithms. Notably, we develop such approximations for *exponential family (EF) Thompson samplers*, which is the TS principle implemented using the posterior updating mechanics of any prior distribution (satisfying modest regularity conditions) and any single-parameter exponential family likelihood. We further develop such approximations for the *bootstrap sampler*, which is similar to the TS principle, but involves (non-parametric) bootstrap re-sampling instead of posterior sampling. Under diffusion scaling, our theory indicates that all of these algorithms satisfy an invariance principle—namely, in the limit, their sampling behaviors and thus also their SDEs and stochastic ODEs all coincide with that of the Gaussian Thompson sampler. Thus, in minimax or worst-case settings, this positions the Gaussian Thompson sampler as the central/canonical bandit algorithm among the many versions of TS and related sampling-based bandit algorithms studied in the literature. Additionally, under diffusion scaling, the regret performance of these algorithms is insensitive to mis-specification of reward distributions, as shown in Proposition 3. This contrasts with the instance-dependent bandit setting of Lai and Robbins (1985), where algorithms can be highly sensitive to model mis-specification, as recently shown in Fan and Glynn (2024).

The rest of the paper is structured as follows. Related work is further discussed in Section 1.1. We then describe the model and setup used throughout the paper in Section 2. In Section 3.1, we provide an intuitive derivation leading to the SDE convergence result for the Gaussian Thompson sampler in Theorem 1 (with the proof given in Section 5.1). Similarly, in Section 3.2, we provide an intuitive derivation leading to the stochastic ODE convergence result for Gaussian Thompson sampler in Theorem 2 (with the proof given in Section 5.2). We provide extensions (Corollary 1 and Theorem 3) of diffusion approximations in Section 3.3. In Section 4.1, we show that the EF Thompson sampler has the same weak limit under diffusion scaling as the Gaussian Thompson sampler. The same is shown for the bootstrap sampler in Section 4.2. In Section 4.3, under diffusion scaling, we discuss the insensitivity of these sampling-based bandit algorithms to mis-specification of the reward distribution. We then conclude the paper with a quick study of batched updating 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. As mentioned in the Introduction, the overlap between our work and theirs is that both obtain similar SDE and stochastic ODE approximations for the dynamics of the Gaussian Thompson sampler under diffusion scaling with $\sqrt{\gamma}$-scale sub-optimality gaps over time horizons of $O(1/\gamma)$; see our Theorems 1 and 2, and KW's Theorems 1 and 3 (applied to the Gaussian Thompson sampler).

However, the theoretical approach taken in our paper to establish these results differs from that of KW in the following way. KW represent sampling-based algorithms, including TS, as Markov chains, and use the martingale framework of Stroock and Varadhan (Stroock and Varadhan 1979) to establish weak convergence of the Markov chains to diffusion processes by showing the convergence of the corresponding infinitesimal generators. On the other hand, as discussed in the Introduction, we use direct representations in terms of discrete versions of SDEs and stochastic ODEs, and we show from first principles using the Continuous Mapping Theorem that the discrete systems converge weakly to their continuous counterparts. Our approach has the advantage that it offers a transparent and intuitive view of how the diffusion approximations arise. Furthermore, our approach can be used to obtain diffusion approximations for sampling-based algorithms belonging to the *Sequentially Randomized Markov Experiment Framework* of KW.

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 diffusion approximations 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.

2. Model and Preliminaries

Bandit Problems and Thompson Sampling

A general sampling-based bandit algorithm operates as follows. We have a filtration $\mathcal{H} = (\mathcal{H}_j, j \geq 0)$ that the bandit process is adapted to, with

$$\mathcal{H}_j = \sigma(I(1), Y(1), \ldots, I(j), Y(j)) \tag{1}$$

corresponding to the data collected through some time j, where at each time i and for each arm $k \in [K] := \{1, \ldots, K\}$, $I_k(i) = 1$ if arm k is selected and otherwise $I_k(i) = 0$ (so that $\sum_k I_k(i) = 1$), and $Y(i)$ is the reward received for the selected arm. For the settings in this paper, the data

can be summarized by sufficient statistics $(N(j), G(j)) = ((N_k(j), G_k(j)), k \in [K])$ measurable with respect to $\mathcal{H}_j$, where for each arm $k \in [K]$,

$$N_k(j) = \sum_{i=1}^{j} I_k(i) \tag{2}$$

is the number of plays and

$$G_k(j) = \sum_{i=1}^{j} I_k(i) Y(i) \tag{3}$$

is the cumulative reward.

The algorithm selects an arm in the time period $j+1$ by generating $I(j+1)$ as an independent K-dimensional multinomial random variable with a single trial and success probability vector $\pi(N(j), G(j)) \in \Delta^K$, where Δ^K denotes the K-dimensional probability simplex and $\pi : \mathbb{N}^K \times \mathbb{R}^K \to \Delta^K$. Given $I(j+1)$, a reward $Y(j+1)$ is received for the selected arm, and the sufficient statistics $(N(j+1), G(j+1))$ are updated accordingly.

TS is an important example of a sampling-based bandit algorithm and our primary focus throughout the paper. When studying TS, we will restrict attention to TS designed for parametric reward models parameterized by mean. (As mentioned in the Introduction, we will begin with the Gaussian Thompson sampler in Sections 3.1 and 3.2 before generalizing to EF Thompson samplers in Section 4.1.) As a Bayesian algorithm, TS maintains a posterior distribution for the mean reward of each arm, and in each time period, it samples a mean from each posterior and plays the arm corresponding to the highest sampled mean, after which a corresponding reward is received and the posterior is updated with the new information. More precisely, for each arm k, we start with an independent prior distribution ν_k^0 for the unknown mean μ_k. From posterior updating, at each time $j = 1, 2, \ldots$ and for each arm k, we have a posterior distribution $\nu_k(N_k(j), G_k(j))$, which depends on the sufficient statistics $(N_k(j), G_k(j))$ for that arm. At time j, we draw an independent sample $\widetilde{\mu}_k(j) \sim \nu_k(N_k(j), G_k(j))$ for each arm k, and we play the arm $\arg\max_k \widetilde{\mu}_k(j)$. So, for TS, $\pi_k(N(j), G(j)) := \mathbb{P}(k = \arg\max_l \widetilde{\mu}_l(j))$, i.e., each arm is played according to the posterior probability that it has the highest mean reward.

Reward Feedback Mechanisms

We consider two distributionally equivalent ways of generating reward feedback. For each arm $k \in [K]$, let Q_k be the reward distribution, and on a common probability space, let $X_k(i) \overset{\text{iid}}{\sim} Q_k$ for $i = 1, 2, \ldots$. We refer to the first way as the *random table model*, where at time j, $Y(j) = X_k(j)$ for the selected arm $k \in [K]$ $(I_k(j) = 1)$. We refer to the second way as the *reward stack model*, where at time j, $Y(j) = X_k(N_k(j-1)+1)$ for the selected arm $k \in [K]$ $(I_k(j) = 1)$, where $N_k(j-1)$ is

the number of plays of arm k through time $j-1$, as defined in (2) above. (The random table and reward stack terminology is taken from Lattimore and Szepesvári (2020); see Chapter 4.6, page 53.) In Section 3.1, we will see how the random table model leads to an SDE characterization of TS dynamics. In Section 3.2, we will see how the reward stack model leads to a stochastic ODE characterization.

Diffusion Scaling Asymptotic Regime

As mentioned in the Introduction, throughout the paper, we consider a sequence of bandit models indexed by a positive, real-valued parameter γ, with $\gamma \downarrow 0$. We will consider bandit instances with arm mean separation on the scale of $\sqrt{\gamma}$, over time horizons on the scale of $1/\gamma$. When working within the corresponding γ-scale system, we will write a γ superscript on all objects defined previously to indicate we are working with the same object defined appropriately in the γ-scale system.

For each γ and each arm $k \in [K]$, we have a reward distribution Q_k^γ, with rewards $X_k^\gamma(i) \overset{\text{iid}}{\sim} Q_k^\gamma$ for $i = 1, 2, \ldots$. Regardless of the reward feedback mechanism (with reward $Y^\gamma(j)$ at time j given according to the random table model or the reward stack model introduced previously), the algorithm's information is captured by the filtration $\mathcal{H}^\gamma = (\mathcal{H}_j^\gamma, j \geq 0)$, with

$$\mathcal{H}_j^\gamma = \sigma(I^\gamma(1), Y^\gamma(1), \ldots, I^\gamma(j), Y^\gamma(j)). \tag{4}$$

The diffusion scaling asymptotic regime is defined in Assumption 1 below.

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

$$\mu_k^\gamma = \mu_* + \sqrt{\gamma} d_k^\gamma, \quad \lim_{\gamma \downarrow 0} d_k^\gamma = d_k \tag{5}$$

$$\lim_{\gamma \downarrow 0} \sigma_k^\gamma = \sigma_k \tag{6}$$

$$\sup_{\gamma > 0} \mathbb{E}\left[|X_k^\gamma(i)|^{2+\alpha} \right] < \infty. \tag{7}$$

In the diffusion scaling setting of Assumption 1, for each arm k, we will use the notation $\Delta_k^\gamma := \max_l d_l^\gamma - d_k^\gamma$. As $\gamma \downarrow 0$, $\Delta_k^\gamma \to \Delta_k := \max_l d_l - d_k$. The essential idea behind the diffusion scaling is that the arm means μ_k^γ are all clustered near some fixed $\mu_* \in \mathbb{R}$, with small differences/gaps between the means on the scale of $\sqrt{\gamma}$. In order to begin distinguishing between arms, one must play each arm on the scale of $1/\gamma$ times, so that the standard errors for estimating the means are on the scale of $\sqrt{\gamma}$, comparable in size to the gaps between the arm means. Playing the arms significantly less times results in their means essentially being indistinguishable. As we will see, because all arms are played so many times, collectively their reward processes are well-approximated by Brownian motions.

REMARK 1. For our analysis throughout the paper, finite $2+\alpha$ (with arbitrarily small $\alpha > 0$) moments for the rewards suffices (as in (7)), while the theoretical approach of Kuang and Wager (2023) requires finite fourth moments.

Function Spaces and Weak Convergence

Throughout this paper, $D^m[a,\infty)$ to denotes the space of functions with domain $[a,\infty)$ and range $\mathbb{R}^m$, that are right-continuous and have limits from the left. For this space, we use the Skorohod metric. Weak convergence is always denoted using $\Rightarrow$, both for stochastic processes taking values in $D^m[a,\infty)$ and for random variables taking values in $\mathbb{R}^m$. Complete mathematical details for the spaces $D^m[a,\infty)$ equipped with the Skorohod metric, as well as the theory of weak convergence in such spaces, can be found in standard references such as Billingsley (1999), Ethier and Kurtz (1986) and Whitt (2002).

3. Derivations of Diffusion Approximations

In the following sections, we derive an SDE approximation (Section 3.1) and a stochastic ODE approximation (Section 3.2) for the Gaussian Thompson sampler, i.e., TS implemented using posterior updating based on Gaussian priors and likelihoods. We assume that the rewards are from general distributions (not necessarily Gaussian) and satisfy the conditions of the setup in Assumption 1. For the Gaussian likelihood, we use a fixed variance $c_*^2 > 0$, which may or may not correspond to the limit variances σ_k^2 of the rewards (as in (6)). Later in the paper, we will complement the theory of this section by studying EF Thompson samplers in Section 4.1 and then model misspecification issues in Section 4.3.

Before continuing on to the derivation of diffusion approximations, we first discuss a technical issue that can arise. The behavior of TS can be highly erratic at the very beginning of a bandit experiment under diffusion scaling (Assumption 1) when little data has been collected and the algorithm is performing a lot of exploration by randomly sampling arms. This can create mathematical difficulties such as the breakdown of Lipschitz continuity in SDE approximations in an arbitrarily small initial time interval (in continuous time), which in turn makes it challenging to establish that the SDEs have unique solutions. Below, we discuss two ways of "smoothing" the initial behavior of TS to restore Lipschitz continuity of the SDEs.

1) Smoothing via Concentrated Priors

One way to smooth out the initial behavior of TS is to use a concentrated prior. We use this approach in Sections 3.1 and 3.2. From Assumption 1, the arm means μ_k^γ are concentrated around μ_* with sub-optimality gaps $\sqrt{\gamma}\Delta_k^\gamma$, where the Δ_k^γ are unknown. We assume that μ_* and $\sqrt{\gamma}$ are known, and we use an independent $N(\mu_*, \gamma/b)$ prior for each arm in the Gaussian Thompson

sampler, with fixed $b > 0$. Translated into practice, this means that the experimenter knows that the arm means are in a "$\sqrt{\gamma}$-scale neighborhood" of μ_* (so that the sub-optimality gaps, i.e., effect sizes, are on the scale of $\sqrt{\gamma}$), perhaps from similar experiments run in the past. (To keep the algebra simple, we assume without loss of generality that $\mu_* = 0$.) Then, the experimenter will run a bandit experiment over time horizons scaling as $1/\gamma$ to learn about the sub-optimality gaps and maximize cumulative reward.

Importantly, the use of γ-scale variance priors together with $(1/\gamma)$-scale time horizons ensures the SDE approximations have desirable Lipschitz continuity properties and thus a unique strong solution. The use of γ-scale variance priors together with data collected over $(1/\gamma)$-scale time horizons naturally enable Bayesian inference about the $\sqrt{\gamma}$-scale sub-optimality gaps. If the prior is less concentrated with variance scaling as $\omega(\gamma)$ as $\gamma \downarrow 0$, then it will be asymptotically dominated by the data collected over $(1/\gamma)$-scale time horizons. And if the prior is more concentrated with variance scaling as $o(\gamma)$ as $\gamma \downarrow 0$, then it will asymptotically dominate the data collected.

2) Smoothing via ϵ-warm-start

A second way to smooth out the initial erratic behavior of TS is to sample all arms with fixed, positive probabilities for an arbitrarily small initial time interval in continuous time, and then run TS afterwards. We refer to this initialization procedure as *ϵ-warm-start* (defined below), and we will use it in Section 3.3 and in Section 4.

DEFINITION 1 (ϵ-WARM-START). Fix some positive probabilities $q_1, \ldots, q_K$ (with $\sum_k q_k = 1$). For the initial $\lfloor \epsilon/\gamma \rfloor$ time periods, sample each arm k with probability q_k. Then, run TS from time $\lfloor \epsilon/\gamma \rfloor + 1$ onward.

Using ϵ-warm-start, we can ensure Lipschitz continuity of the SDE approximation, and thus a unique strong solution. Moreover, the prior used in TS can be general and need not be concentrated in any way. We can also think of ϵ-warm-start as an empirical Bayes approach, where a tiny fraction of data is collected initially to learn a prior with the centering around μ_* and the variance scale of γ, after which TS using the learned prior is deployed.

3.1. SDE Approximation

To derive the SDE approximation for the Gaussian Thompson sampler, we use the random table model of reward feedback introduced in Section 2. We will first show that the dynamics in this setting are described by the evolution of two processes: $U^\gamma = (U_k^\gamma,\, k \in [K])$ and $S^\gamma = (S_k^\gamma,\, k \in [K])$, defined via:

$$U_k^\gamma(t) = \gamma \sum_{i=1}^{\lfloor t/\gamma \rfloor} I_k^\gamma(i) \tag{8}$$

$$S_k^\gamma(t) = \sqrt{\gamma} \sum_{i=1}^{\lfloor t/\gamma \rfloor} I_k^\gamma(i) \frac{X_k^\gamma(i) - \mu_k^\gamma}{\sigma_k^\gamma}, \tag{9}$$

which are re-scaled and centered versions of (2) and (3), respectively.

REMARK 2. Under the setup of Assumption 1, for a particular γ value, the overall regret $\mathrm{Reg}^\gamma(n)$ at time n is related to the U_k^γ processes via:

$$\mathrm{Reg}^\gamma(n) = \frac{1}{\sqrt{\gamma}} \sum_{k\in[K]} U_k^\gamma(n\gamma)\Delta_k^\gamma. \tag{10}$$

At time $j+1$, conditional on $\mathcal{H}_j^\gamma$ (defined in (4)), the Gaussian Thompson sampler draws a sample from the posterior distribution of each arm k:

$$\widetilde{\mu}_k^\gamma(j+1) \sim N\left(\frac{\gamma \sum_{i=1}^j I_k^\gamma(i) X_k^\gamma(i)}{U_k^\gamma(j\gamma) + bc_*^2}, \frac{c_*^2\gamma}{U_k^\gamma(j\gamma) + bc_*^2} \right). \tag{11}$$

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

$$\mathbb{P}\left(k = \arg\max_{l\in[K]} \widetilde{\mu}_l^\gamma(j+1) \,\middle|\, \mathcal{H}_j^\gamma \right) \tag{12}$$

$$= \mathbb{P}\left(k = \arg\max_{l\in[K]} \left\{ \frac{S_l^\gamma(j\gamma)\sigma_l^\gamma + U_l^\gamma(j\gamma)d_l^\gamma}{U_l^\gamma(j\gamma) + bc_*^2} + \frac{c_*}{\sqrt{U_l^\gamma(j\gamma) + bc_*^2}} \mathcal{N}_l \right\} \,\middle|\, U^\gamma(j\gamma), S^\gamma(j\gamma) \right) \tag{13}$$

$$= p_k^\gamma(U^\gamma(j\gamma), S^\gamma(j\gamma)), \tag{14}$$

where the probability is taken over the independent standard Gaussian variables $\mathcal{N}_l$, and for $u = (u_k, k\in[K]) \in [0,\infty)^K$ and $s = (s_k, k\in[K]) \in \mathbb{R}^K$,

$$p_k^\gamma(u,s) = \mathbb{P}\left(k = \arg\max_{l\in[K]} \left\{ \frac{s_l\sigma_l^\gamma + u_l d_l^\gamma}{u_l + bc_*^2} + \frac{c_*}{\sqrt{u_l + bc_*^2}} \mathcal{N}_l \right\} \right). \tag{15}$$

We can now re-express $U_k^\gamma(t)$ and $S_k^\gamma(t)$ from (8)-(9) as

$$U_k^\gamma(t) = \gamma \sum_{i=0}^{\lfloor t/\gamma \rfloor - 1} p_k^\gamma(U^\gamma(i\gamma), S^\gamma(i\gamma)) + M_k^\gamma(t) \tag{16}$$

$$S_k^\gamma(t) = \sum_{i=0}^{\lfloor t/\gamma \rfloor - 1} \sqrt{p_k^\gamma(U^\gamma(i\gamma), S^\gamma(i\gamma))} \left(B_k^\gamma((i+1)\gamma) - B_k^\gamma(i\gamma) \right), \tag{17}$$

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

$$M_k^\gamma(t) = \gamma \sum_{i=0}^{\lfloor t/\gamma \rfloor - 1} \left(I_k^\gamma(i+1) - p_k^\gamma(U^\gamma(i\gamma), S^\gamma(i\gamma)) \right) \tag{18}$$

$$B_k^\gamma(t) = \sqrt{\gamma} \sum_{i=0}^{\lfloor t/\gamma \rfloor - 1} \frac{I_k^\gamma(i+1)(X_k^\gamma(i+1) - \mu_k^\gamma)}{\sqrt{p_k^\gamma(U^\gamma(i\gamma), S^\gamma(i\gamma))} \cdot \sigma_k^\gamma}, \tag{19}$$

and $(I_k^\gamma(i+1), k \in [K])$ is a multinomial random variable with a single trial and success probabilities $p_k^\gamma(U^\gamma(i\gamma), S^\gamma(i\gamma))$.

As $\gamma \downarrow 0$, we show that M^γ and B^γ converge weakly to the $D^K[0,\infty)$ zero process and standard K-dimensional Brownian motion, respectively. Additionally, since $d_k^\gamma \to d_k$ and $\sigma_k^\gamma \to \sigma_k$ (from (5)-(6)), we have

$$p_k^\gamma(u,s) \to p_k(u,s) \tag{20}$$

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

$$p_k(u,s) = \mathbb{P}\left(k = \arg\max_{l\in[K]} \left\{ \frac{s_l\sigma_l + u_l d_l}{u_l + bc_*^2} + \frac{c_*}{\sqrt{u_l + bc_*^2}}\mathcal{N}_l \right\}\right). \tag{21}$$

Thus, we expect (16)-(17) to be a discrete approximation to the SDE in integral form:

$$U_k(t) = \int_0^t p_k(U(v), S(v))dv \tag{22}$$

$$S_k(t) = \int_0^t \sqrt{p_k(U(v), S(v))}dB_k(v), \quad k \in [K] \tag{23}$$

with standard K-dimensional Brownian motion B. As mentioned earlier in this section, the functions p_k in (21) are Lipschitz continuous, which ensures that the SDEs in (22)-(23) 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 SDE characterization is stated in Theorem 1 below. The proof of Theorem 1 can be found in Section 5.1, along with the development of the supporting results for the proof. The rigorous argument closely follows the derivation above. The main technical tool is the Continuous Mapping Theorem, together with the property that stochastic integration is a continuous mapping of the integrand and integrator processes, which allows us to pass from the pre-limit in (16)-(17) to the limit in (22)-(23).

THEOREM 1. *Under the diffusion scaling of Assumption 1 and the random table model of reward feedback, for a K-armed bandit and the Gaussian Thompson sampler with prior variance scaling as γ,*

$$(U^\gamma, S^\gamma) \Rightarrow (U, S) \tag{24}$$

as $\gamma \downarrow 0$ in $D^{2K}[0,\infty)$, where (U,S) is the unique strong solution to the SDE:

$$dU_k(t) = p_k(U(t), S(t))dt \tag{25}$$

$$dS_k(t) = \sqrt{p_k(U(t), S(t))}dB_k(t) \tag{26}$$

$$U_k(0) = S_k(0) = 0, \quad k \in [K], \tag{27}$$

with standard K-dimensional Brownian motion B, and functions p_k as expressed in (21).

Moreover, for regret,

$$\sqrt{\gamma}\,\mathrm{Reg}^{\gamma}(\lfloor \cdot/\gamma \rfloor) \Rightarrow \sum_{k\in[K]} U_k(\cdot)\Delta_k \tag{28}$$

as $\gamma \downarrow 0$ in $D[0,\infty)$.

3.2. Stochastic ODE Approximation

To derive the stochastic ODE approximation for the Gaussian Thompson sampler, we use the reward stack model of reward feedback introduced in Section 2. Similar to the derivation of the SDE approximation, we first show that the dynamics are described by the evolution of two processes: $U^{\gamma} = (U_k^{\gamma},\, k \in [K])$ with the same expression as in (8), and $Z^{\gamma} \circ U^{\gamma} = (Z_k^{\gamma}(U_k^{\gamma}),\, k \in [K])$ defined via:

$$Z_k^{\gamma}(U_k^{\gamma}(t)) = \sqrt{\gamma} \sum_{i=1}^{U_k^{\gamma}(t)/\gamma} \frac{X_k^{\gamma}(i) - \mu_k^{\gamma}}{\sigma_k^{\gamma}}, \tag{29}$$

where $Z^{\gamma} = (Z_k^{\gamma},\, k \in [K])$ has the expression:

$$Z_k^{\gamma}(t) = \sqrt{\gamma} \sum_{i=1}^{\lfloor t/\gamma \rfloor} \frac{X_k^{\gamma}(i) - \mu_k^{\gamma}}{\sigma_k^{\gamma}}, \tag{30}$$

which is a re-scaled and centered version of (3). (For vector-valued functions f and g, we use $f \circ g$ to denote component-wise composition of f and g.) For the stochastic ODE approximation, since U^{γ} has the same expression as in (8), the relationship to regret $\mathrm{Reg}^{\gamma}(n)$ is the same as in Remark 2.

Remark 3. We point out that the distribution of the process $S_k^{\gamma}(t)$ (as defined in (9)) and of the process $Z_k^{\gamma}(U_k^{\gamma}(t))$ (as defined in (29)) are the same. As can be seen 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, and so they are distributionally equivalent, as we will make clear below in the discussion leading up to and in the statement of Theorem 2.

At time $j+1$, conditional on $\mathcal{H}_j^{\gamma}$ (defined in (4)), the Gaussian Thompson sampler draws a sample from the posterior distribution of each arm k:

$$\widetilde{\mu}_k^{\gamma}(j+1) \sim N\left(\frac{\gamma \sum_{i=1}^{U_k^{\gamma}(j\gamma)/\gamma} X_k^{\gamma}(i)}{U_k^{\gamma}(j\gamma) + bc_*^2}, \frac{c_*^2 \gamma}{U_k^{\gamma}(j\gamma) + bc_*^2} \right). \tag{31}$$

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

$$\mathbb{P}\left(k = \arg\max_{l\in[K]} \widetilde{\mu}_l^\gamma(j+1) \,\middle|\, \mathcal{H}_j^\gamma\right) \tag{32}$$

$$= \mathbb{P}\left(k = \arg\max_{l\in[K]} \left\{ \frac{Z_l^\gamma(U_l^\gamma(j\gamma))\sigma_l^\gamma + U_l^\gamma(j\gamma)d_l^\gamma}{U_l^\gamma(j\gamma) + bc_*^2} + \frac{c_*}{\sqrt{U_l^\gamma(j\gamma) + bc_*^2}}\mathcal{N}_l \right\} \,\middle|\, U^\gamma(j\gamma), Z^\gamma \circ U^\gamma(j\gamma)\right) \tag{33}$$

$$= p_k^\gamma(U^\gamma(j\gamma), Z^\gamma \circ U^\gamma(j\gamma)), \tag{34}$$

where the probability is taken over the independent standard Gaussian variables $\mathcal{N}_l$, and functions p_k^γ are given by (21).

We can now re-express $U_k^\gamma(t)$ as

$$U_k^\gamma(t) = \gamma \sum_{i=0}^{\lfloor t/\gamma \rfloor - 1} p_k^\gamma(U^\gamma(i\gamma), Z^\gamma \circ U^\gamma(i\gamma)) + M_k^\gamma(t), \quad k \in [K], \tag{35}$$

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

$$M_k^\gamma(t) = \gamma \sum_{i=0}^{\lfloor t/\gamma \rfloor - 1} \left(I_k^\gamma(i+1) - p_k^\gamma(U^\gamma(i\gamma), Z^\gamma \circ U^\gamma(i\gamma))\right), \tag{36}$$

and $(I_k^\gamma(i+1),\, k \in [K])$ is a multinomial random variable with a single trial and success probabilities $p_k^\gamma(U^\gamma(i\gamma), Z^\gamma \circ U^\gamma(i\gamma))$.

As $\gamma \downarrow 0$, we show that M^γ and Z^γ converge weakly to the $D^K[0,\infty)$ zero process and standard K-dimensional Brownian motion, respectively. As in previous section, the convergence in (20) holds. Thus, we expect (35) to be a discrete approximation to the stochastic ODE in integral form:

$$U_k(t) = \int_0^t p_k(U(v), B \circ U(v))dv, \quad k \in [K], \tag{37}$$

with standard K-dimensional Brownian motion B, and functions p_k as expressed in (21).

To conclude the above derivation, the rigorous stochastic ODE characterization is stated in Theorem 2 below. The proof of Theorem 2 can be found in Section 5.2. The rigorous argument closely follows the derivation above, using the Continuous Mapping Theorem, together with the property that Riemann integration is a continuous mapping of the integrand and integrator processes, which allows us to pass from the pre-limit in (35) to the limit in (37). As can be seen from the proof, the stochastic ODE representation in Theorem 2 can be recovered from the SDE representation in Theorem 1 using the fact that any continuous martingale can be represented as a Brownian motion evolving according to a random clock/time-change. While we generally cannot claim that the stochastic ODE has a unique adapted (pathwise) solution (because the sample paths of Brownian motion are not Lipschitz continuous), the limit U in Theorem 2 is a particular adapted (pathwise) solution of the stochastic ODE, and is distributionally equivalent to its counterpart in the unique strong solution of the SDE from Theorem 1.

THEOREM 2. *Under the diffusion scaling of Assumption 1 and the reward stack model of reward feedback, for a K-armed bandit and the Gaussian Thompson sampler with prior variance scaling as γ,*

$$(U^\gamma, Z^\gamma \circ U^\gamma) \Rightarrow (U, B \circ U) \tag{38}$$

as $\gamma \downarrow 0$ in $D^{2K}[0,\infty)$, where U is an adapted (pathwise) solution to the stochastic ODE:

$$dU_k(t) = p_k(U(t), B \circ U(t))dt \tag{39}$$

$$U_k(0) = 0, \quad k \in [K], \tag{40}$$

with standard K-dimensional Brownian motion B, and functions p_k as expressed in (21).

The limit stochastic ODE and one of its adapted (pathwise) solutions can be derived from the limit SDE (in Theorem 1) and its unique strong solution. In particular, with (U, S) as the unique strong solution to the SDE in (25)-(27), we have $S(t) = \widetilde{B} \circ U(t)$ for some standard K-dimensional Brownian motion $\widetilde{B}$. So, the adapted (pathwise) solution to the stochastic ODE referenced above is distributionally equivalent to the unique strong solution of the SDE.

Moreover, for regret, (28) holds in the stochastic ODE setting.

3.3. Additional Approximations

Recall from the development of Theorems 1 and 2, with the functions p_k as defined in (21), that it is important for $(u, s) \mapsto p_k(u, s)$ to be Lipschitz continuous, which ensures that the limit SDEs and stochastic ODEs have unique solutions. In Corollary 1, we state a result for general sampling-based bandit algorithms that does not involve Lipschitz continuous limit sampling probabilities p_k. In such settings, there may not be a unique solution to the limit SDE or stochastic ODE. Nevertheless, the rescaled pre-limit processes, for example, $(U^\gamma, Z^\gamma \circ U^\gamma)$ in the stochastic ODE setting, will still be tight. So, every subsequence as $\gamma \downarrow 0$ of pre-limit processes will have a further subsequence that converges weakly to a limit process that satisfies the stochastic ODE. However, these weak limit processes may be distinct in general, so we simply characterize their evolution equations. The justification for Corollary 1 follows directly from the proof of Theorem 2.

COROLLARY 1. *Under the diffusion scaling of Assumption 1, for a K-armed bandit and a sampling-based algorithm, suppose that as $\gamma \downarrow 0$, the sampling probabilities $p_k^\gamma(u, s) \to p_k(u, s)$ uniformly for (u, s) in compact subsets of $[0,\infty)^K \times \mathbb{R}^K$, where p_k is a continuous function. Then, under the reward stack model of reward feedback, the weak limit points of $(U^\gamma, Z^\gamma \circ U^\gamma)$ in $D^{2K}[0,\infty)$ as $\gamma \downarrow 0$ are of the form $(U, B \circ U)$, where U is an adapted (pathwise) solution of the stochastic ODE:*

$$dU_k(t) = p_k(U(t), B \circ U(t))dt \tag{41}$$

$$U_k(0) = 0, \quad k \in [K], \tag{42}$$

with standard K-dimensional Brownian motion B.

Next, we develop a diffusion approximation for the Gaussian Thompson sampler without assuming a concentrated prior with variance scaling as γ. Specifically, we consider any fixed Gaussian prior (with constant variance) in the asymptotics as $\gamma \downarrow 0$. Unlike in Sections 3.1-3.2, here we can take $\mu_* \in \mathbb{R}$ to be unknown, since we are not using concentrated priors and do not need to center such priors on μ_*. Then, the functions p_k in (21) become:

$$p_k(u,s) = \mathbb{P}\left(k = \arg\max_{l \in [K]} \left\{ \frac{s_l \sigma_l}{u_l} + d_l + \frac{c_*}{\sqrt{u_l}} \mathcal{N}_l \right\}\right), \tag{43}$$

where the probability is taken over the independent standard Gaussian variables $\mathcal{N}_l$.

However, as discussed at the beginning of Section 3, the function $(u,s) \mapsto p_k(u,s)$ in (43) is no longer Lipschitz continuous for points near $u_l = 0$, $l \in [K]$. Nevertheless, the problem with the p_k in (43) only exists for an infinitesimally small initial interval. Whenever all inputs $U_l(t)$, $l \in [K]$ to the u_l components in (43) become strictly positive, then from that time onward, there is Lipschitz continuity of the p_k. It then follows that there is a unique strong solution to the SDE and an adapted (pathwise) solution to the corresponding stochastic ODE, which are distributionally equivalent representations of the weak limit. In Theorem 3, we use ϵ-warm-start (recall Definition 1) to ensure Lipschitz continuity. The proof of Theorem 3 is a simple modification of those of Theorems 1 and 2, and is thus omitted.

THEOREM 3. *Under the diffusion scaling of Assumption 1, consider a K-armed bandit and the Gaussian Thompson sampler with a fixed prior variance (no γ-dependence) and ϵ-warm-start. Then, under the random table model of reward feedback,*

$$(U^\gamma, S^\gamma) \Rightarrow (U, S) \tag{44}$$

as $\gamma \downarrow 0$ in $D^{2K}[\epsilon, \infty)$, where (U,S) is the unique strong solution to the SDE:

$$dU_k(t) = p_k(U(t), S(t))dt \tag{45}$$

$$dS_k(t) = \sqrt{p_k(U(t), S(t))} dB_k(t) \tag{46}$$

$$U_k(\epsilon) = q_k \epsilon \tag{47}$$

$$S_k(\epsilon) = \sqrt{q_k} B_k(\epsilon), \quad k \in [K], \tag{48}$$

with standard K-dimensional Brownian motion B, and functions p_k as expressed in (43).

Moreover, under the reward stack model of reward feedback, a distributionally equivalent characterization of the dynamics of (U,S) in (45)-(46) is in terms of $(U, B \circ U)$, where U is an adapted (pathwise) solution to the stochastic ODE:

$$dU_k(t) = p_k(U(t), B \circ U(t))dt, \quad k \in [K]. \tag{49}$$

Furthermore, for regret, (28) holds in both the SDE and stochastic ODE settings.

4. Further Insights from Diffusion Approximations

4.1. Approximations for Exponential Family Thompson Samplers

So far, we have focused on the Gaussian Thompson sampler. In Proposition 1 below, we show that the sampling behaviors and process-level dynamics of EF Thompson samplers can be approximated by those of the Gaussian Thompson sampler. (Recall from the Introduction that EF Thompson samplers are versions of TS implemented using posterior updating with any prior distribution (satisfying modest regularity conditions) and any single-parameter exponential family likelihood.) In the literature, minimax or worst-case regret analysis (which is essentially diffusion scaling, with sub-optimality gaps scaling as $1/\sqrt{n}$ with time horizon n) is carried out on a case-by-case basis for the many variants of TS (with posterior updating based on Gaussian prior and likelihood, beta prior and Bernoulli likelihood, etc.). Our approximation of EF Thompson samplers by the Gaussian Thompson sampler suggests that for minimax regret analysis, it suffices to simply analyze the Gaussian Thompson sampler, which has minimax optimal dependence of expected regret on the time horizon (Agrawal and Goyal 2013, 2017).

For Proposition 1, the main step is to establish under diffusion scaling that the posterior distributions of EF Thompson samplers are approximately Gaussian. To develop the Gaussian approximation, in this section we assume that the arm reward distributions are from an exponential family P^μ parameterized by mean μ. The exponential family distributions have the form:

$$P^\mu(dx) = \exp(\theta(\mu) \cdot x - \Lambda(\mu))P(dx), \tag{50}$$

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. Let $(\underline{\mu}, \overline{\mu})$ denote the open interval of all possible mean values achievable by the family P^μ (for some value of the tilting parameter $\theta(\mu) \in \mathbb{R}$).

For simplicity, suppose we know that the mean reward for all arms belong to a bounded, open interval $\mathcal{I}$, with $\inf \mathcal{I} > \underline{\mu}$ and $\sup \mathcal{I} < \overline{\mu}$. (The analysis is simplified by avoiding the boundaries $\underline{\mu}$ and $\overline{\mu}$.) Suppose also that Assumption 1 holds, and that for the distributions Q_k^γ with means μ_k^γ from Assumption 1, we have $Q_k^\gamma = P^{\mu_k^\gamma}$, with all $\mu_k^\gamma \in \mathcal{I}$. For the σ_k in (6), here we have $\sigma_k = \sigma_*$ for all k, where σ_*^2 is the variance of P^{μ_*}, with the μ_* from (5).

We consider EF Thompson samplers with posterior updating based on the likelihood of the exponential family P^μ (with mean $\mu \in \mathcal{I}$), together with any prior (for the mean) with bounded density, support contained in $\mathcal{I}$, and continuous and positive density in a neighborhood of μ_*. For simplicity, we use the same prior for every arm, with independence across arms (and no γ-dependence).

The above setup leads to the following Proposition 1 for EF Thompson samplers under diffusion scaling. The proof of Proposition 1 is provided in Appendix A. It uses a version of the Bernstein-von Mises Theorem, i.e., a Gaussian approximation for the posterior distribution, which can be found in Proposition 5 in Appendix B.

PROPOSITION 1. *Consider the setup described above, with a K-armed bandit under the diffusion scaling of Assumption 1, the arm reward distributions belonging to an exponential family of the form in (50), and the corresponding EF Thompson sampler with a prior having continuous and positive density in a neighborhood of μ_*.*

Then, under ϵ-warm-start, we have weak convergence to the limits in (44)-(49) of Theorem 3, with

$$p_k(u,s) = \mathbb{P}\left(k = \arg\max_{l\in[K]}\left\{\frac{s_l\sigma_*}{u_l} + d_l + \frac{\sigma_*}{\sqrt{u_l}}\mathcal{N}_l\right\}\right), \tag{51}$$

where the probability is taken over the independent standard Gaussian variables $\mathcal{N}_l$.

Furthermore, for regret, (28) continues to hold in both the SDE and stochastic ODE settings.

The conclusion of Proposition 1 (with the $p_k(u,s)$ in (51)) matches that of Theorem 3 (with the $p_k(u,s)$ in (43)) when (in the context of Theorem 3) the limit variances σ_k^2 in (6) match the variance c_*^2 used in the Gaussian likelihood of the Gaussian Thompson sampler.

Our results here also indicate that under diffusion scaling, the Gaussian Thompson sampler is a good approximation of other variants of TS, including ones involving approximations of the posterior distribution, for example, via Laplace approximation. Since the Gaussian Thompson sampler is known to have optimal or near-optimal expected regret performance in a wide range of settings (Agrawal and Goyal 2013, Korda et al. 2013, Agrawal and Goyal 2017), 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.

4.2. Approximations for Bootstrap Sampler

The bootstrap and related ideas such as subsampling have recently been proposed as mechanisms 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 bootstrap sampler introduced earlier, which is one natural implementation of bootstrapping to induce exploration in bandit problems. For the bootstrap sampler, in each time period, a single

(non-parametric) bootstrapped sample mean is generated for each arm, and the arm with the greatest one is played.

In Proposition 2 below, we show that for general reward distributions, the sampling behavior and process-level dynamics of the bootstrap sampler can be approximated by those of the Gaussian Thompson sampler. This is similar in spirit to Proposition 1. But unlike in Proposition 1, here the reward distributions do not need to belong to any exponential family. Given the optimality or near optimality of the Gaussian Thompson sampler discussed previously, our results here suggest that the bootstrap sampler can be an effective means of balancing exploration and exploitation under diffusion scaling, with the added benefit of not needing to make distributional assumptions.

The proof of Proposition 2 is the same as that of Proposition 1, except we use a Gaussian approximation for the bootstrapped sample mean, which is developed in Proposition 6 in Appendix C.

PROPOSITION 2. *Consider a K-armed bandit under the diffusion scaling of Assumption 1, and suppose also that*

$$\lim_{y\to\infty}\sup_{\mu\in\mathcal{I}}\mathbb{E}^{\mu}[X^2\mathbb{I}\left(X^2>y\right)]=0. \tag{52}$$

Then, for the bootstrap sampler under ϵ-warm-start, we have weak convergence to the limits in (44)-(49) of Theorem 3, with

$$p_k(u,s)=\mathbb{P}\left(k=\arg\max_{l\in[K]}\left\{\frac{s_l\sigma_l}{u_l}+d_l+\frac{\sigma_l}{\sqrt{u_l}}\mathcal{N}_l\right\}\right). \tag{53}$$

Furthermore, for regret, (28) continues to hold in both the SDE and stochastic ODE settings.

Compared to Theorem 3 (with the $p_k(u,s)$ in (43)), in Proposition 2 (with the $p_k(u,s)$ in (53)), the bootstrap sampler automatically adapts to the limit variance σ_k^2 for each arm k, rather than having to specify some variance c_*^2 as in the Gaussian Thompson sampler. This is reflected in the $(\sigma_l/\sqrt{u_l})\mathcal{N}_l$ terms in (53)), compared to the $(c_*/\sqrt{u_l})\mathcal{N}_l$ terms in (43).

4.3. Model Mis-specification

In this section, we show that under the diffusion scaling of Assumption 1, the regret of the Gaussian Thompson sampler, and that of other TS variants like EF Thompson samplers, are robust to mis-specification of the reward distributions. Asymptotically, under diffusion scaling, only the limit means and variances (as in (5)-(6)) of the reward distributions influence the dynamics of the Gaussian Thompson sampler. So, in Theorems 1-3, mis-specification corresponds to mis-match between the limit variances σ_k^2 in (6) and the variance c_*^2 specified in the Gaussian likelihood.

In Proposition 3 below, we establish that under the diffusion scaling of Assumption 1, the regret (as expressed in (10) in Remark 2) of the Gaussian Thompson sampler (on the $1/\sqrt{\gamma}$ scale) is continuous with respect to the limit variances $\sigma := (\sigma_k, k \in [K])$. As mentioned in the Introduction, this contrasts with the results in the instance-dependent Lai-Robbins asymptotic regime (Lai and Robbins 1985). In that setting, as recently shown in Fan and Glynn (2024), 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 (from scaling as $\log(n)$ to polynomial in n with horizon n). Furthermore, previously in Section 4.1, we showed that EF Thompson samplers can be approximated by the Gaussian Thompson sampler under diffusion scaling. This suggests that under diffusion scaling, the robustness of TS to model mis-specification extends to other settings as well.

PROPOSITION 3. *Let $(U, S) = (U_k, S_k, k \in [K])$ denote either the solution to (25)-(27) in Theorem 1 (equivalently, (39)-(40) in Theorem 2) with σ-dependence as in (21), or the solution to (45)-(48) in Theorem 3 with σ-dependence as in (43). Then, the distribution of (U, S) is continuous with respect to σ, i.e., for any bounded continuous function $f : D^{2K}[0, \infty) \to \mathbb{R}$, the mapping $\sigma \mapsto \mathbb{E}^{\sigma}[f(U, S)]$ is continuous. Moreover, for any fixed $t > 0$,*

$$\lim_{\gamma \downarrow 0} \sqrt{\gamma}\, \mathbb{E}^{\sigma}[\mathrm{Reg}^{\gamma}(\lfloor t/\gamma \rfloor)] = \sum_{k \in [K]} \mathbb{E}^{\sigma}[U_k(t)]\Delta_k,$$

where $\sigma \mapsto \mathbb{E}^{\sigma}[U_k(t)]$ is a positive, continuous mapping for each arm $k \in [K]$.

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 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 4. *In the settings of Theorems 1, 2 and 3, the same conclusions hold for the Gaussian Thompson sampler 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.

5. Proofs for Main Results

5.1. Proofs for SDE Approximation

In this section, we prove the SDE approximation in Theorem 1 (from Section 3.1). We first discuss a (random) step function approximation (with any desired accuracy) for functions in $D^m[0,\infty)$, due to Kurtz and Protter (1991). The step function approximation is technically useful for passing from the discrete versions of Itô integrals to the Itô integrals themselves in the continuous weak limit. We describe the approximation in Definition 2, and discuss integration with the approximation applied to the integrand in Definition 3. Then, Lemma 1, which is a summary of useful technical results from Kurtz and Protter (1991) (see their Lemma 6.1 and its proof, as well as the proof of their Theorem 1), ensures that integration is a continuous mapping when the integrand is approximated in such a way. Following these, we provide proof of Theorem 1 using the continuity properties at hand together with the Continuous Mapping Theorem. We conclude the section with Lemmas 2 and 3, which establish the tightness of stochastic processes and convergence to Brownian motion used in the proof of Theorem 1.

DEFINITION 2 (STEP FUNCTION APPROXIMATION). For any $\epsilon > 0$, we construct a random step function mapping $\chi^\epsilon : D^m[0,\infty) \to D^m[0,\infty)$ as follows. We use the ℓ^1 norm, with $||w|| := \sum_{i=1}^m |w_i|$ for $w \in \mathbb{R}^m$. For any $z \in D^m[0,\infty)$, define inductively the random times $\tau_j(z)$ starting with $\tau_0(z) = 0$:

$$\tau_{j+1}(z) = \inf\{t > \tau_j(z) : \max(||z(t) - z(\tau_j(z))||, ||z(t-) - z(\tau_j(z))||) \geq \epsilon V_j\}, \tag{54}$$

where $V_j \overset{\text{iid}}{\sim} \text{Unif}(1/2, 1)$. Then, define $\chi^\epsilon(z) \in D^m[0,\infty)$ by

$$\chi^\epsilon(z)(t) = z(\tau_j(z)), \quad t \in [\tau_j(z), \tau_{j+1}(z)), \tag{55}$$

so that $\chi^\epsilon(z)$ is a step function (piecewise constant), and almost surely,

$$\sup_{t \geq 0} ||\chi^\epsilon(z)(t) - z(t)|| \leq \epsilon. \tag{56}$$

DEFINITION 3 (INTEGRATION WITH STEP FUNCTIONS). On an interval $[a,b]$, let f_1, f_2 be $\mathbb{R}$-valued right-continuous functions with left limits, where f_1 is a step function with jump points $t_1 < \cdots < t_j$ in $[a,b]$. We will always use the following definition of integration for step function integrands (setting $t_0 = a$ and $t_{j+1} = b$):

$$\int_a^b f_1(t) df_2(t) = \sum_{i=0}^{j} f_1(t_i) \left(f_2(t_{i+1}) - f_2(t_i) \right). \tag{57}$$

Then, for any $\epsilon > 0$, with the random step function mapping $\chi^\epsilon : D^m[0,\infty) \to D^m[0,\infty)$ in (55), define the integral mapping $\mathcal{I}^\epsilon : D^{2m}[0,\infty) \to D^m[0,\infty)$ component-wise for $k = 1,\ldots,m$ via

$$\mathcal{I}_k^\epsilon(g,h)(t) = \int_0^t \chi_k^\epsilon(g)(v) dh_k(v), \tag{58}$$

for $g,h \in D^m[0,\infty)$ (with the definition of integral in (57)).

LEMMA 1 **(Continuity of Integration with Step Functions)**. *Let the sequences $x_n, y_n \in D^m[0,\infty)$ and also $x,y \in D^m[0,\infty)$ such that jointly $(x^n,y^n) \to (x,y)$ in $D^{2m}[0,\infty)$ as $n \to \infty$. For $\epsilon > 0$, let $\chi^\epsilon : D^m[0,\infty) \to D^m[0,\infty)$ be the random step function mapping in (55), and let $\mathcal{I}^\epsilon : D^{2m}[0,\infty) \to D^m[0,\infty)$ be the integral mapping in (58). Then,*

$$(x^n, y^n, \mathcal{I}^\epsilon(x^n,y^n)) \overset{a.s.}{\to} (x, y, \mathcal{I}^\epsilon(x,y)) \tag{59}$$

in $D^{3m}[0,\infty)$ as $n \to \infty$.

Moreover, let z^n be a sequence of $D^m[0,\infty)$ processes, adapted to a sequence of filtrations $\mathcal{F}^n = (\mathcal{F}_t^n, t \geq 0)$. Then, $\chi^\epsilon(z^n)$ is adapted to the corresponding sequence of augmented filtrations $\mathcal{G}^n = (\mathcal{G}_t^n, t \geq 0)$, where $\mathcal{G}_t^n = \sigma(\mathcal{F}_t^n \cup \mathcal{V})$, with $\mathcal{V} = \sigma(V_j, j \geq 1)$ being the sigma-algebra generated by the extra randomization used to construct χ^ϵ (independent of the filtrations $\mathcal{F}^n$).

Proof of Theorem 1. We start with the discrete approximation (16)-(19) from our derivation in Section 3.1. We denote the joint processes via $(U^\gamma, S^\gamma, B^\gamma, M^\gamma) = (U_k^\gamma, S_k^\gamma, B_k^\gamma, M_k^\gamma, k \in [K])$, and recall that they are processes in $D^{4K}[0,\infty)$.

Our proof strategy is as follows. We will show that for every subsequence of (U^γ, S^γ), 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 (25)-(26) are Lipschitz-continuous and bounded on their domain of definition, the SDE has a unique strong solution (Theorem 5.2.9 of Karatzas and Shreve (1998)). Thus, (U^γ, S^γ) must converge weakly to the unique strong solution of the SDE.

By Lemma 2 (stated and proved after the current proof), the joint processes $(U^\gamma, S^\gamma, B^\gamma, M^\gamma)$ are tight in $D^{4K}[0,\infty)$, and thus, Prohorov's Theorem ensures that for each subsequence, there is a further subsequence which converges weakly to some limit process $(U,S,B,M) = (U_k, S_k, B_k, M_k, k \in$

$[K]$) (see Chapter 3 of Ethier and Kurtz (1986), Chapters 1 and 3 of Billingsley (1999), or Chapter 11 of Whitt (2002)). From now on, we work with this further subsequence, and for notational simplicity, we still index this further subsequence by γ. So, we have

$$(U^\gamma, S^\gamma, B^\gamma, M^\gamma) \Rightarrow (U, S, B, M). \tag{60}$$

Because M^γ consists of martingale differences, by a Chebyshev bound, we have $M_k^\gamma(t) \xrightarrow{\mathbb{P}} 0$ for each $k \in [K]$ and any $t > 0$ as $n \to \infty$, and thus, M is the $D^K[0,\infty)$ zero process. By Lemma 3 (stated and proved after the current proof), B is standard K-dimensional Brownian motion.

Now define the processes $A^\gamma = (A_k^\gamma,\, k \in [K])$ and $A = (A_k,\, k \in [K])$, where

$$A_k^\gamma(t) = p_k^\gamma(U^\gamma(t), S^\gamma(t)) \tag{61}$$

$$A_k(t) = p_k(U(t), S(t)). \tag{62}$$

Since $p_k^\gamma(u,s) \to p_k(u,s)$ as $\gamma \downarrow 0$ uniformly for (u,s) in compact subsets of $[0,\infty)^K \times \mathbb{R}^K$, and $p_k(u,s)$ is continuous at all $(u,s) \in [0,\infty)^K \times \mathbb{R}^K$, by the Generalized Continuous Mapping Theorem (Lemma 6) applied to the processes in (61)-(62), we have from (60),

$$(U^\gamma, S^\gamma, B^\gamma, M^\gamma, A^\gamma) \Rightarrow (U, S, B, M, A). \tag{63}$$

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

$$\widetilde{U}_k^\gamma(t) = \int_0^t p_k^\gamma(U^\gamma(v), S^\gamma(v))dv \tag{64}$$

$$\widetilde{U}_k(t) = \int_0^t p_k(U(v), S(v))dv. \tag{65}$$

Recall that

$$U_k^\gamma(t) = \gamma \sum_{i=0}^{\lfloor t/\gamma \rfloor - 1} p_k^\gamma(U^\gamma(i\gamma), S^\gamma(i\gamma)) + M_k^\gamma(t).$$

For each $k \in [K]$, because M_k^γ converges weakly to the $D[0,\infty)$ zero process and also

$$\sup_{t \geq 0} \left| \gamma \sum_{i=0}^{\lfloor t/\gamma \rfloor - 1} p_k^\gamma(U^\gamma(i\gamma), S^\gamma(i\gamma)) - \widetilde{U}_k^\gamma(t) \right| \leq \gamma,$$

we have for any $T > 0$,

$$\sup_{0 \leq t \leq T} \left| U_k^\gamma(t) - \widetilde{U}_k^\gamma(t) \right| \xrightarrow{\mathbb{P}} 0. \tag{66}$$

Thus, by the fact that integration is a continuous functional with respect to the Skorohod metric (Theorem 11.5.1 of Whitt (2002)) and the Continuous Mapping Theorem, we have from (63),

$$(U^\gamma, S^\gamma, B^\gamma, \widetilde{U}^\gamma, A^\gamma) \Rightarrow (U, S, B, \widetilde{U}, A). \tag{67}$$

Let $\epsilon > 0$. Let χ^ϵ be the random step function mapping defined in (54) and (55), and let $\mathcal{I}^\epsilon$ be the corresponding integral operator defined in (58). Recall from (17), (19) and (61), that for each $k \in [K]$,

$$S_k^\gamma(t) = \int_0^t \sqrt{A_k^\gamma(v-)} dB_k^\gamma(v), \tag{68}$$

and define the process $\widehat{S}^\gamma = (\widehat{S}_k^\gamma,\, k \in [K]) := \mathcal{I}^\epsilon(\sqrt{A^\gamma}, B^\gamma)$, i.e., for each $k \in [K]$,

$$\widehat{S}_k^\gamma(t) = \int_0^t \chi_k^\epsilon\left(\sqrt{A^\gamma(v-)}\right) dB_k^\gamma(v). \tag{69}$$

By Lemma 1 and the Continuous Mapping Theorem, with the continuity of the mapping $(x, y) \mapsto (x, y, \mathcal{I}^\epsilon(x, y))$ established in (59), we have from (67),

$$(U^\gamma, S^\gamma, B^\gamma, \widetilde{U}^\gamma, \widehat{S}^\gamma) \Rightarrow (U, S, B, \widetilde{U}, \widehat{S}), \tag{70}$$

where the process $\widehat{S} = (\widehat{S}_k,\, k \in [K]) := \mathcal{I}^\epsilon(\sqrt{A}, B)$, i.e., for each $k \in [K]$,

$$\widehat{S}_k(t) = \int_0^t \chi_k^\epsilon\left(\sqrt{A(v-)}\right) dB_k(v). \tag{71}$$

We also define the process $\widetilde{S} = (\widetilde{S}_k,\, k \in [K])$, where for each $k \in [K]$,

$$\widetilde{S}_k(t) = \int_0^t \sqrt{A_k(v-)} dB_k(v). \tag{72}$$

Note that both of the processes in (71) and (72) are well defined as Itô integrals, since by Lemma 3, the integrands are non-anticipative with respect to the Brownian motions B_k. (As defined in (54)-(55), χ^ϵ depends on exogenous randomization that is independent of the B_k.) By Lemma 1, because χ^ϵ is an ϵ-uniform approximation (see (56)), for each $k \in [K]$ and any $T > 0$,

$$\mathbb{E}\left[\sup_{0 \le t \le T} \left|S_k^\gamma(t) - \widehat{S}_k^\gamma(t)\right|\right] \le \epsilon \mathbb{E}\left[\gamma \sum_{i=0}^{\lfloor T/\gamma \rfloor - 1} \mathbb{E}\left[\frac{I_k^\gamma(i+1)(X_k^\gamma(i+1) - \mu_k^\gamma)^2}{p_k^\gamma(U^\gamma(i\gamma), S^\gamma(i\gamma)) \cdot (\sigma_k^\gamma)^2} \,\middle|\, \mathcal{H}_i^\gamma\right]\right]^{1/2} \le \epsilon\sqrt{T}. \tag{73}$$

Similarly, for each k and any $T > 0$,

$$\mathbb{E}\left[\sup_{0 \le t \le T} \left|\widehat{S}_k(t) - \widetilde{S}_k(t)\right|\right] \le \epsilon \mathbb{E}\left[\langle B_k \rangle_T\right]^{1/2} = \epsilon\sqrt{T}, \tag{74}$$

where $t \mapsto \langle B_k \rangle_t$ denotes the quadratic variation process for B_k. Putting together (66), (70)-(74) and sending $\epsilon \downarrow 0$, we have

$$(U^\gamma, S^\gamma, B^\gamma, U^\gamma, S^\gamma) \Rightarrow (U, S, B, \widetilde{U}, \widetilde{S}). \tag{75}$$

Recalling the definition of $\widetilde{U}$ in (65) as well as that of $\widetilde{S}$ in (72) and the A_k in (62), we see from (75) that the limit processes (U,S,B) satisfy the SDE:

$$U_k(t) = \int_0^t p_k(U(v),S(v))dv \tag{76}$$

$$S_k(t) = \int_0^t \sqrt{p_k(U(v),S(v))}dB_k(v), \quad k=1,\ldots,K. \tag{77}$$

(Note that from (76)-(77), it is clear that (U,S,B) is adapted to the (augmented) filtration $\mathcal{F}_t = \sigma(\mathcal{F}_t^B \cup \mathcal{L})$, where $\mathcal{F}_t^B = \sigma\left(B(v) : 0 \le v \le t\right)$, with $\mathcal{L}$ denoting the collection of all $\mathbb{P}$-null sets.) □

LEMMA 2. *The processes $(U^\gamma, S^\gamma, B^\gamma, M^\gamma)$ defined in (16)-(19) are tight in $D^{4K}[0,\infty)$.*

Proof of Lemma 2. We recall that the processes have the following expressions for $k=1,\ldots,K$.

$$U_k^\gamma(t) = \gamma \sum_{i=1}^{\lfloor t/\gamma \rfloor} I_k^\gamma(i) \tag{78}$$

$$S_k^\gamma(t) = \sqrt{\gamma} \sum_{i=1}^{\lfloor t/\gamma \rfloor} I_k^\gamma(i) \frac{X_k^\gamma(i) - \mu_k^\gamma}{\sigma_k^\gamma} \tag{79}$$

$$M_k^\gamma(t) = \gamma \sum_{i=0}^{\lfloor t/\gamma \rfloor - 1} \left(I_k^\gamma(i+1) - p_k^\gamma(U^\gamma(i\gamma), S^\gamma(i\gamma))\right) \tag{80}$$

$$B_k^\gamma(t) = \sqrt{\gamma} \sum_{i=0}^{\lfloor t/\gamma \rfloor - 1} \frac{I_k^\gamma(i+1)(X_k^\gamma(i+1) - \mu_k^\gamma)}{\sqrt{p_k^\gamma(U^\gamma(i\gamma), S^\gamma(i\gamma))} \cdot \sigma_k^\gamma} \tag{81}$$

Note that (78)-(79) are just different expressions of the same quantities in (16)-(17). With a slight abuse of notation, let $(\mathcal{H}_t^\gamma,\, t \ge 0)$ denote the continuous, piecewise constant (and right-continuous) interpolation of the discrete-time filtration $(\mathcal{H}_j^\gamma,\, j \ge 0)$ defined in (4), so that (78)-(81) are all adapted to $\mathcal{H}_t^\gamma$. Also, the process in (78) is uniformly bounded and increasing, and those in (79)-(81) are square-integrable martingales.

By Lemma 7, to show tightness of the joint processes $(U^\gamma, S^\gamma, B^\gamma, M^\gamma)$, we just need to show tightness of each component sequence of processes and each pairwise sum of component sequences of processes. We use Lemma 8 to verify tightness in each case. Condition (T1) can be directly verified using a sub-martingale maximal inequality (for example, Theorem 3.8(i) of 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 fixed $T > 0$, $\delta > 0$ and all $\gamma > 0$, we can set $A_\delta^\gamma(T) = \delta$ for each individual component process, and we can set $A_\delta^\gamma(T) = 4\delta$ (using the bound: $(x+y)^2 \le 2x^2 + 2y^2$) for each pairwise sum of component processes. □

LEMMA 3. *Following Lemma 2, for any subsequence of $(U^\gamma, S^\gamma, B^\gamma, M^\gamma)$ that converges weakly in $D^{4K}[0,\infty)$ to some limit process (U,S,B,M), the component B is standard K-dimensional Brownian motion. Moreover, U and S are non-anticipative with respect to B, i.e., $B(t+v) - B(t)$ is independent of $(U(v'), S(v'))$ for $0 \le v' \le t$ and $v \ge 0$.*

Proof of Lemma 3. To show that $B^\gamma \Rightarrow B$, where B is standard K-dimensional Brownian motion, we apply the martingale functional central limit theorem stated in Lemma 9. Below, we verify (M1) and (M2) to ensure Lemma 9 holds. The non-anticipative property follows from the same property in the pre-limit, i.e., U^γ and S^γ are non-anticipative with respect to B^γ.

Verification of (M1)

Because $I_j^\gamma(i)I_k^\gamma(i) = 0$ for $j \neq k$ and all $i = 1, 2, \ldots$ (only one arm is played in each time period i), we have $\Sigma_{jk} = 0$ for $j \neq k$. For the diagonal elements, we have $\Sigma_{kk} = 1$ for each $k = 1, \ldots, m$, as the following argument shows. As shorthand, denote $p_k^\gamma(i) := p_k^\gamma(U^\gamma(i\gamma), S^\gamma(i\gamma))$. Then,

$$\begin{aligned}
&\gamma \sum_{i=0}^{\lfloor t/\gamma \rfloor} \mathbb{E}\left[\frac{I_k^\gamma(i+1)}{p_k^\gamma(i)} \left(\frac{X_k^\gamma(i+1) - \mu_k^\gamma}{\sigma_k^\gamma} \right)^2 \,\middle|\, \mathcal{H}_i^\gamma \right] \\
&= \gamma \sum_{i=0}^{\lfloor t/\gamma \rfloor - 1} \frac{\mathbb{E}\left[I_k^\gamma(i+1) \,\middle|\, \mathcal{H}_i^\gamma \right]}{p_k^\gamma(i)} \mathbb{E}\left[\left(\frac{X_k^\gamma(i+1) - \mu_k^\gamma}{\sigma_k^\gamma} \right)^2 \,\middle|\, \mathcal{H}_i^\gamma \right] \qquad (82) \\
&= \gamma \lfloor t/\gamma \rfloor \to t
\end{aligned}$$

as $\gamma \downarrow 0$. Here, (82) follows from $p_k^\gamma(i) = p_k^\gamma(U^\gamma(i\gamma), S^\gamma(i\gamma))$ being $\mathcal{H}_i^\gamma$-measurable, and $I_k^\gamma(i+1)$ and $(X_k^\gamma(i+1) - \mu_k^\gamma)^2/(\sigma_k^\gamma)^2$ being independent conditional on $\mathcal{H}_i^\gamma$.

Verification of (M2)

For each $k = 1, \ldots, m$, denote

$$W_k^\gamma(i+1) = \frac{I_k^\gamma(i+1)(X_k^\gamma(i+1) - \mu_k^\gamma)}{\sqrt{p_k^\gamma(i)} \cdot \sigma_k^\gamma}.$$

By Markov's inequality, it suffices to show that for each fixed $i = 0, 1, \ldots,$

$$\mathbb{E}\left[W_k^\gamma(i+1)^2 \mathbb{I}\left(|W_k^\gamma(i+1)| > \epsilon/\sqrt{\gamma} \right) \right] \to 0 \qquad (83)$$

as $\gamma \downarrow 0$.

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

$$\mathbb{P}\left(\inf_{v \in [0,t]} p_k^\gamma(U^\gamma(v), S^\gamma(v)) < \delta \right) \leq \eta.$$

We then have

$$\mathbb{E}\left[W_k^\gamma(i+1)^2 \mathbb{I}\left(|W_k^\gamma(i+1)| > \epsilon/\sqrt{\gamma} \right) \mathbb{I}\left(p_k^\gamma(i) < \delta \right) \right]$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[W_k^\gamma(i+1)^2 \mathbb{I}\left(p_k^\gamma(i) < \delta\right)\right] \\
&= \mathbb{E}\left[\frac{\mathbb{I}\left(p_k^\gamma(i) < \delta\right)}{p_k^\gamma(i)} \mathbb{E}\left[I_k^\gamma(i+1)\left(\frac{X_k^\gamma(i+1) - \mu_k^\gamma}{\sigma_k^\gamma}\right)^2 \,\middle|\, \mathcal{H}_i^\gamma\right]\right] \qquad (84) \\
&= \mathbb{P}\left(p_k^\gamma(i) < \delta\right) \qquad (85) \\
&\leq \mathbb{P}\left(\inf_{v\in[0,t]} p_k^\gamma(U^\gamma(v), S^\gamma(v)) < \delta\right) \\
&\leq \eta, \qquad (86)
\end{aligned}
$$

where (84) follows from $U^\gamma(i\gamma)$ and $S^\gamma(i\gamma)$ being $\mathcal{H}_i^\gamma$-measurable, (85) follows from conditional independence of $I_k^\gamma(i+1)$ and $(X_k^\gamma(i+1) - \mu_k^\gamma)^2/(\sigma_k^\gamma)^2$, and (86) holds for γ sufficiently close to zero, as established above.

Additionally, we have

$$
\begin{aligned}
&\mathbb{E}\left[W_k^\gamma(i+1)^2 \mathbb{I}\left(|W_k^\gamma(i+1)| > \epsilon/\sqrt{\gamma}\right) \mathbb{I}\left(p_k^\gamma(i) \geq \delta\right)\right] \\
&= \mathbb{E}\left[\frac{\mathbb{I}\left(p_k^\gamma(i) \geq \delta\right)}{p_k^\gamma(i)} \mathbb{E}\left[I_k^\gamma(i+1)\left(\frac{X_k^\gamma(i+1) - \mu_k^\gamma}{\sigma_k^\gamma}\right)^2 \mathbb{I}\left(|W_k^\gamma(i+1)| > \epsilon/\sqrt{\gamma}\right) \,\middle|\, \mathcal{H}_i^\gamma\right]\right] \\
&\leq \frac{1}{\delta}\mathbb{E}\left[\mathbb{E}\left[\left(\frac{X_k^\gamma(i+1) - \mu_k^\gamma}{\sigma_k^\gamma}\right)^2 \mathbb{I}\left(|W_k^\gamma(i+1)| > \epsilon/\sqrt{\gamma}\right) \,\middle|\, \mathcal{H}_i^\gamma\right]\right] \\
&\leq \frac{1}{\delta}\mathbb{E}\left[\mathbb{E}\left[\left|\frac{X_k^\gamma(i+1) - \mu_k^\gamma}{\sigma_k^\gamma}\right|^{2+\alpha}\right]^{2/(2+\alpha)} \mathbb{P}\left(|W_k^\gamma(i+1)| > \epsilon/\sqrt{\gamma} \,\middle|\, \mathcal{H}_i^\gamma\right)^{\alpha/(2+\alpha)}\right] \qquad (87) \\
&\leq \frac{C}{\delta}\mathbb{E}\left[\mathbb{P}\left(|W_k^\gamma(i+1)| > \epsilon/\sqrt{\gamma} \,\middle|\, \mathcal{H}_i^\gamma\right)^{\alpha/(2+\alpha)}\right], \qquad (88)
\end{aligned}
$$

where (87) follows from Hölder's inequality, and (88) follows from (7) in Assumption 1, with constant $C > 0$. Furthermore, almost surely,

$$
\begin{aligned}
\mathbb{P}\left(|W_k^\gamma(i+1)| > \epsilon/\sqrt{\gamma} \,\middle|\, \mathcal{H}_i^\gamma\right) &\leq \frac{\gamma}{\epsilon^2}\frac{1}{p_k^\gamma(i)}\mathbb{E}\left[I_k^\gamma(i+1)\left(\frac{X_k^\gamma(i+1) - \mu_k^\gamma}{\sigma_k^\gamma}\right)^2 \,\middle|\, \mathcal{H}_i^\gamma\right] \\
&= \frac{\gamma}{\epsilon^2}.
\end{aligned}
$$

So, by the bounded convergence theorem, the right side of (88) converges to zero as $\gamma \downarrow 0$.

Therefore, from (86) and (88), we have

$$
\limsup_{\gamma\downarrow 0} \mathbb{E}\left[W_k^\gamma(i+1)^2 \mathbb{I}\left(|W_k^\gamma(i+1)| > \epsilon/\sqrt{\gamma}\right)\right] \leq \eta, \qquad (89)
$$

and sending $\eta \downarrow 0$ yields (83). □

5.2. Proofs for Stochastic ODE Approximation

In this section, we prove Theorem 2 (from Section 3.2), which is an alternative stochastic ODE representation of the SDE in Theorem 1 (from Section 3.1).

Proof of Theorem 2.

Weak Convergence to the Stochastic ODE Limit

To show weak convergence to the stochastic ODE limit, we only need to slightly modify the proof of Theorem 1. We start with the discrete approximation (35)-(36) and (29)-(30) from our derivation in Section 3.2. We denote the joint processes via $(U^\gamma, Z^\gamma, M^\gamma) = (U_k^\gamma, Z_k^\gamma, M_k^\gamma, k \in [K])$, and recall that they are processes in $D^{3K}[0,\infty)$.

Consider a weakly convergent subsequence of (U^γ, Z^γ), which we will still index by γ for notational simplicity. Then, jointly $(U^\gamma, Z^\gamma, M^\gamma) \Rightarrow (U, Z, M)$, where (as in the proof of Theorem 1) M is the $D^K[0,\infty)$ zero process. By Donsker's Theorem (Chapter 3 of Billingsley (1999)), Z is standard K-dimensional Brownian motion.

By the continuity of function composition (Theorem 13.2.2 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,

$$(U^\gamma, Z^\gamma, M^\gamma, Z^\gamma \circ U^\gamma) \Rightarrow (U, Z, M, Z \circ U). \tag{90}$$

Define the processes $A^\gamma = (A_k^\gamma, k \in [K])$ and $A = (A_k, k \in [K])$, where

$$A_k^\gamma(t) = p_k^\gamma(U^\gamma(t), Z^\gamma \circ U^\gamma(t)) \tag{91}$$

$$A_k(t) = p_k(U(t), Z \circ U(t)). \tag{92}$$

Since $p_k^\gamma(u,s) \to p_k(u,s)$ as $\gamma \downarrow 0$ uniformly for (u,s) in compact subsets of $[0,\infty)^K \times \mathbb{R}^K$, and $p_k(u,s)$ is continuous at all $(u,s) \in [0,\infty)^K \times \mathbb{R}^K$, by the Generalized Continuous Mapping Theorem (Lemma 6) applied to the processes in (91)-(92), we have from (90),

$$(U^\gamma, Z^\gamma, M^\gamma, A^\gamma) \Rightarrow (U, Z, M, A). \tag{93}$$

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

$$\begin{aligned} \widetilde{U}_k^\gamma(t) &= \int_0^t p_k^\gamma(U^\gamma(v), Z^\gamma \circ U^\gamma(v))dv \\ \widetilde{U}_k(t) &= \int_0^t p_k(U(v), Z \circ U(v))dv. \end{aligned} \tag{94}$$

Recall that

$$U_k^\gamma(t) = \gamma \sum_{i=0}^{\lfloor t/\gamma \rfloor - 1} p_k^\gamma(U^\gamma(i\gamma), Z^\gamma \circ U^\gamma(i\gamma)) + M_k^\gamma(t).$$

For each $k \in [K]$, because M_k^γ converges weakly to the $D[0,\infty)$ zero process and also

$$\sup_{t \geq 0} \left| \gamma \sum_{i=0}^{\lfloor t/\gamma \rfloor - 1} p_k^\gamma(U^\gamma(i\gamma), Z^\gamma \circ U^\gamma(i\gamma)) - \widetilde{U}_k^\gamma(t) \right| \leq \gamma,$$

we have for any $T > 0$,

$$\sup_{0 \le t \le T} \left| U_k^\gamma(t) - \widetilde{U}_k^\gamma(t) \right| \xrightarrow{\mathbb{P}} 0. \tag{95}$$

Thus, by the fact that integration is a continuous functional with respect to the Skorohod metric (Theorem 11.5.1 of Whitt (2002)) and the Continuous Mapping Theorem, we have from (93),

$$(U^\gamma, Z^\gamma, \widetilde{U}^\gamma) \Rightarrow (U, Z, \widetilde{U}). \tag{96}$$

Together, (95)-(96) yield

$$(U^\gamma, Z^\gamma, U^\gamma) \Rightarrow (U, Z, \widetilde{U}),$$

and recalling the definition of $\widetilde{U}$ in (94), we have established weak convergence to a weak limit point.

So, we have established that each weak limit point of the stochastic ODE pre-limit is an adapted (pathwise) solution to the stochastic ODE. (The stochastic ODE does not necessarily have a unique adapted (pathwise) solution since Brownian motion sample paths are not Lipschitz continuous, and so conventional ODE uniqueness theory does not apply.) We also know that the stochastic ODE pre-limit and the SDE pre-limit (from Theorem 1) have the same distribution (since the random table model and the reward stack model of reward feedback are distributionally equivalent), and as established in Theorem 1, the SDE pre-limit converges weakly to the unique strong solution of an SDE. Therefore, the stochastic ODE pre-limit must converge to a single weak limit, and that weak limit is an adapted (pathwise) solution to the stochastic ODE.

Random Time Change Representation

Here we show that the stochastic ODE can be recovered from the SDE through a random time change. We work with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ supporting a standard Brownian motion B on $\mathbb{R}^K$, with natural filtration $\mathcal{F}_t^B = \sigma\left(B(v) : 0 \le v \le t\right)$. We will work with the corresponding augmented filtration $\mathcal{F}_t = \sigma(\mathcal{F}_t^B \cup \mathcal{L})$, where $\mathcal{L}$ 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 (U, S) to the SDE (25)-(26) on this probability space with respect to the standard Brownian motion B. Writing (26) in integral form, because the p_k functions are bounded,

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

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

$$\langle S_k \rangle_t = \int_0^t p_k(U(v), S(v)) dv, \quad k \in [K],$$

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 (25) (in the Riemann sense) yields $\langle S_k \rangle_t = U_k(t)$, $k \in [K]$, which are continuous and strictly increasing processes since the p_k functions are bounded and strictly positive. Define

$$U_k^{-1}(t) = \inf\{v \geq 0 : U_k(v) \geq t\}, \quad k \in [K].$$

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 example, Proposition 18.8 of Kallenberg (2002) or Theorem 1.10 of Revuz and Yor (1999)), for $k \in [K]$, we have that $\widetilde{B}_k(t) := S_k(U_k^{-1}(t))$ are independent standard Brownian motions with respect to the filtration $\mathcal{F}_t^{\widetilde{B}} = \sigma\left(\widetilde{B}(u) : 0 \leq u \leq t\right)$. Thus, we have $\widetilde{B}_k(U_k(t)) = S_k(t)$, and substituting this representation into the SDE (25), we obtain the stochastic ODE:

$$U_k(t) = \int_0^t p_k(U(v), \widetilde{B} \circ U(v)) dv, \quad k \in [K]. \tag{97}$$

So with respect to the smaller filtration $\mathcal{F}_t^{\widetilde{B}}$, the SDE solution $U(t)$ satisfies the stochastic ODE (97), which coincides with (39). $\square$

Appendix A: Additional Material for Section 4

Proof of Proposition 1. Under ϵ-warm-start, we only need to establish the SDE and stochastic ODE approximations on $[\epsilon, \infty)$. We verify that the sampling probabilities for EF Thompson samplers have the desired form with $p_k(u, s)$ as in (51).

In the SDE case, as before, we use the random table model of reward feedback. At time $j+1$, conditional on $\mathcal{H}_j^\gamma$ (as defined in (4)), for each arm k, we sample a value $\widetilde{\mu}_k^\gamma(j+1)$ from the posterior distribution of μ_k^γ. Let $\widehat{\mu}_k^\gamma(j+1)$ denote the sample mean estimate at time $j+1$. (For the exponential family model, the sample mean is the maximum likelihood estimator (MLE) for the mean, and is used as the centering value for the Gaussian posterior approximation in Proposition 5.) Here, the S_k^γ and U_k^γ have the expressions from (8)-(9). The probability of playing arm k is given by:

$$\begin{aligned}
&\mathbb{P}\big(k = \arg\max_{l \in [K]} \widetilde{\mu}_l^\gamma(j+1) \,\big|\, \mathcal{H}_j^\gamma\big) \\
&= \mathbb{P}\left(k = \arg\max_{l \in [K]} \left\{ \frac{S_l^\gamma(j\gamma)\sigma_l^\gamma}{U_l^\gamma(j\gamma)} + d_l^\gamma + \frac{1}{\sqrt{\gamma}} \left(\widetilde{\mu}_l^\gamma(j+1) - \widehat{\mu}_l^\gamma(j+1)\right) \right\} \,\middle|\, U^\gamma(j\gamma), S^\gamma(j\gamma) \right) \\
&= \mathbb{P}\left(k = \arg\max_{l \in [K]} \left\{ \frac{S_l^\gamma(j\gamma)\sigma_l^\gamma}{U_l^\gamma(j\gamma)} + d_l^\gamma + \frac{\sigma_l^\gamma}{\sqrt{U_l^\gamma(j\gamma)}} \mathcal{N}_l \right\} \,\middle|\, U^\gamma(j\gamma), S^\gamma(j\gamma) \right) + o_{\mathbb{P}}(1) \qquad (98) \\
&= p_k^\gamma(U^\gamma(j\gamma), S^\gamma(j\gamma)) + o_{\mathbb{P}}(1), \qquad (99)
\end{aligned}$$

where (98) follows from Proposition 5, with the probability taken over the independent standard Gaussian variables $\mathcal{N}_l$, and

$$p_k^\gamma(u,s) = \mathbb{P}\left(k = \arg\max_{l\in[K]}\left\{\frac{s_l\sigma_l^\gamma}{u_l} + d_l^\gamma + \frac{\sigma_l^\gamma}{\sqrt{u_l}}\mathcal{N}_l\right\}\right).$$

Moreover, with $p_k(u,s)$ as in (51), $p_k^\gamma(u,s) \to p_k(u,s)$ uniformly for (u,s) on compact subsets of $[0,\infty)^K \times \mathbb{R}^K$, with the restriction that $u_k \geq \epsilon q_k > 0$ for each arm k, due to the initial sampling with constant, positive probabilities $(q_k,\, k \in [K])$ in the ϵ-warm-start procedure.

This sequence of derivations parallels what we derived in (12)-(14) in Section 3.1. From (99), the proof of Theorem 1 can be applied to yield the desired SDE approximation in (45)-(48).

The proof in the stochastic ODE case is analogous. We use the reward stack model of reward feedback, with $Z_k^\gamma(U_k^\gamma)$, as defined in (29), instead of S_k^γ. The proof of Theorem 2 can then be applied to yield the desired stochastic ODE approximation in (49). □

Appendix B: Gaussian Approximations for Posterior Distributions

In this appendix, we consider the same setup as in Section 4.1. Recall that the arm reward distributions are from an exponential family P^μ parameterized by mean μ, as expressed in (50), with all means μ known to belong to a bounded, open interval $\mathcal{I}$. Our goal here is to develop Proposition 5 below, which is a version of the Bernstein-von Mises Theorem. This version establishes weak convergence of the rescaled posterior distribution to a Gaussian distribution, almost surely as the sample size $n \to \infty$ and uniformly over the possible data-generating distributions P^μ. To make sense of the "uniform almost sure" mode of convergence, we first recall an equivalent characterization of almost sure convergence in Remark 4 below, followed by a precise definition of the mode of convergence in Definition 4 below. (For the application in Section 4.1, it suffices to establish convergence in probability, uniformly over the possible data-generating distributions, but we can directly obtain the stronger almost sure convergence using the modest technical conditions corresponding to the setup in Section 4.1.)

REMARK 4. For a sequence of random variables $Y_1, Y_2, \ldots,$

$$Y_n \overset{\text{a.s.}}{\to} 0$$

as $n \to \infty$, if and only if for any $\epsilon > 0$,

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

DEFINITION 4. 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 $\epsilon > 0$,

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

Next, we state Lemma 4, which is used in the proof of Proposition 5. This result, originally due to Chung (1951), is a strong law of large numbers that holds uniformly over a collection of underlying probability distributions.

LEMMA 4. *Let $\mathcal{P}$ be a collection of probability distributions, and for each $P \in \mathcal{P}$, let $Y_i \overset{\text{iid}}{\sim} P$. Suppose the $\mathcal{P}$-uniform integrability condition,*

$$\lim_{z\to\infty} \sup_{P\in\mathcal{P}} \mathbb{E}_P \left[|Y_1 - \mathbb{E}_P[Y_1]| \mathbb{I}\left(|Y_1 - \mathbb{E}_P[Y_1]| > z\right)\right] = 0,$$

is satisfied. Then, for every $\epsilon > 0$,

$$\lim_{m\to\infty} \sup_{P\in\mathcal{P}} \mathbb{P}_P \left(\sup_{n\geq m} \left| \frac{1}{n} \sum_{i=1}^{n} Y_i - \mathbb{E}_P[Y_1] \right| > \epsilon \right) = 0.$$

Before presenting Lemma 5 and then continuing on to Proposition 5, which is the main result of this appendix, we first formalize the (modest) technical conditions, C1 and C2 below, that are used to develop these results. It can be easily verified that the exponential family setup detailed in Section 4.1 satisfies C1 and C2. Notation-wise, corresponding to the exponential family P^μ, the log-likelihood function 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.

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

$$\sup_{z:|\mu - z|\geq\delta} \mathbb{E}^\mu[l(z, X)] \leq \mathbb{E}^\mu[l(\mu, X)] - \epsilon. \tag{100}$$

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

$$\eta(x) \geq \sup_{\mu\in\mathcal{I}} |l'(\mu, x)| \tag{101}$$

$$\kappa(x) \geq \sup_{\mu\in\mathcal{I}} |l'''(\mu, x)|. \tag{102}$$

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

$$\lim_{y\to\infty} \sup_{\mu\in\mathcal{I}} \mathbb{E}^\mu \left[f(X)\mathbb{I}\left(f(X) > y\right)\right] = 0. \tag{103}$$

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 5. *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 (101). 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 Proposition 5. The proof is adapted from the proof of Theorem 4.2 in Ghosh et al. (2006). 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.

PROPOSITION 5. *Suppose the conditions C1 and C2 hold, with bounded, open interval $\mathcal{I} \subset \mathbb{R}$. Let ν_0 be a bounded prior density with support contained in $\mathcal{I}$, that is also continuous and positive on a bounded, open sub-interval $\mathcal{I}' \subset \mathcal{I}$. Then, conditional on the data $X_i^\mu \overset{\text{iid}}{\sim} P^\mu$, the centered and scaled posterior density $y \mapsto \nu_n(y \mid X_1^\mu, \ldots, X_n^\mu)$ for $\sqrt{n}(\widetilde{m}_n^\mu - \widehat{m}_n^\mu)$ satisfies:*

$$\lim_{n\to\infty} \int_{\mathbb{R}} \left| \nu_n(y \mid X_1^\mu, \ldots, X_n^\mu) - \frac{1}{\sqrt{2\pi}\sigma^\mu} \exp\left(-\frac{1}{2(\sigma^\mu)^2} y^2\right) \right| dy = 0 \tag{104}$$

almost surely, uniformly in the underlying distribution P^μ for $\mu \in \mathcal{I}'$.

Proof of Proposition 5. The posterior density can be expressed as

$$\nu_n(y \mid X_1^\mu, \ldots, X_n^\mu) = (C_n^\mu)^{-1} \nu_0(\widehat{m}_n^\mu + y/\sqrt{n}) \exp\left(L_n^\mu(\widehat{m}_n^\mu + y/\sqrt{n}) - L_n^\mu(\widehat{m}_n^\mu)\right), \tag{105}$$

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

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

Consider the following difference between unnormalized densities.

$$D_n^\mu(y) = \nu_0(\widehat{m}_n^\mu + y/\sqrt{n}) \exp\left(L_n^\mu(\widehat{m}_n^\mu + y/\sqrt{n}) - L_n^\mu(\widehat{m}_n^\mu)\right) - \nu_0(\mu) \exp\left(-\frac{1}{2(\sigma^\mu)^2} y^2\right) \tag{106}$$

To show (104), it suffices to show that a.s. uniformly in $\mu \in \mathcal{I}'$,

$$\lim_{n\to\infty} \int_{\mathbb{R}} |D_n^\mu(y)|\, dy = 0. \tag{107}$$

Indeed, if (107) holds, we must also have

$$\lim_{n\to\infty} C_n^\mu = \nu_0(\mu)\sqrt{2\pi}\sigma^\mu, \tag{108}$$

a.s. uniformly in $\mu \in \mathcal{I}'$. So, we would have

$$\begin{aligned} &\int_{\mathbb{R}} \left| \nu_n(y \mid X_1^\mu, \ldots, X_n^\mu) - \frac{1}{\sqrt{2\pi}\sigma^\mu} \exp\left(-\frac{1}{2(\sigma^\mu)^2} y^2\right) \right| dy \\ &\le (C_n^\mu)^{-1} \int_{\mathbb{R}} |D_n^\mu(y)|\, dy + \left| (C_n^\mu)^{-1} \nu_0(\mu) - \frac{1}{\sqrt{2\pi}\sigma^\mu} \right| \int_{\mathbb{R}} \exp\left(-\frac{1}{2(\sigma^\mu)^2} y^2\right) dy \end{aligned} \tag{109}$$

Applying (107) and (108) to (109) would then lead to the desired conclusion in (104).

To show (107), we split the integral into two pieces on $A_n = \{y : |y| > \beta\sqrt{n}\}$ and $A_n^c = \{y : |y| \leq \beta\sqrt{n}\}$, with $\beta > 0$ to be specified later in the proof. In the first case on A_n,

$$\int_{A_n} |D_n^\mu(y)|\, dy \leq \int_{A_n} \nu_0(\widehat{m}_n^\mu + y/\sqrt{n}) \exp\left(L_n^\mu(\widehat{m}_n^\mu + y/\sqrt{n}) - L_n^\mu(\widehat{m}_n^\mu)\right) dy + \int_{A_n} \nu_0(\mu) \exp\left(-\frac{1}{2(\sigma^\mu)^2} y^2\right) dy. \tag{110}$$

Clearly the second integral on the right side of (110) goes to zero as $n \to \infty$, uniformly in $\mu \in \mathcal{I}'$. For the first integral on the right side of (110), from condition C2 with $f(x) = |x|$ and Lemma 4, it follows that

$$\widehat{m}_n^\mu - \mu \to 0 \tag{111}$$

as $n \to \infty$, a.s. uniformly in $\mu \in \mathcal{I}'$. This along with Lemma 5 implies that there exists $\epsilon > 0$ such that

$$\sup_{y \in A_n,\, \widehat{m}_n^\mu + y/\sqrt{n} \in \mathcal{I}} \frac{1}{n}\left(L_n^\mu(\widehat{m}_n^\mu + y/\sqrt{n}) - L_n^\mu(\widehat{m}_n^\mu)\right) \leq -\epsilon, \tag{112}$$

for sufficiently large n, a.s. uniformly in $\mu \in \mathcal{I}'$. (For the first integral on the right side of (110), we only need to consider y such that $\widehat{m}_n^\mu + y/\sqrt{n} \in \mathcal{I}$, since the prior density ν_0 has support contained in $\mathcal{I}$.) Therefore, using (112), the first integral in (110) also goes to zero as $n \to \infty$, a.s. uniformly in $\mu \in \mathcal{I}'$.

For the second case on A_n^c, we analyze

$$\int_{A_n^c} |D_n^\mu(y)|\, dy. \tag{113}$$

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

$$\begin{aligned} L_n^\mu(\widehat{m}_n^\mu + y/\sqrt{n}) - L_n^\mu(\widehat{m}_n^\mu) &= -\frac{1}{2}\frac{1}{n}(L_n^\mu)''(\widehat{m}_n^\mu) y^2 + r_n^\mu(y) \\ &= -\frac{1}{2}\left(\theta''(\widehat{m}_n^\mu)\widehat{m}_n^\mu - \Lambda''(\widehat{m}_n^\mu)\right) y^2 + r_n^\mu(y), \end{aligned} \tag{114}$$

using the fact that $l''(z, x) = \theta''(z) \cdot x - \Lambda''(z)$ (recall the definitions of $\theta(z)$ and $\Lambda(z)$ from (50)), with

$$r_n^\mu(y) = \frac{1}{6}\left(\frac{y}{\sqrt{n}}\right)^3 (L_n^\mu)'''(m_{n,y}^\mu),$$

where $m_{n,y}^\mu$ is a point in between $\widehat{m}_n^\mu$ and $\widehat{m}_n^\mu + y/\sqrt{n}$. Using condition C2 with $f(x) = \kappa(x)$ and Lemma 4, there exists $\delta > 0$ such that for sufficiently large n, a.s. uniformly in $\mu \in \mathcal{I}'$,

$$|r_n^\mu(y)| \leq \frac{1}{6}\frac{y^3}{\sqrt{n}}\frac{1}{n}\sum_{i=1}^n \kappa(X_i^\mu) \leq \frac{1}{6}\frac{y^3}{\sqrt{n}}\left(\mathbb{E}\left[\kappa(X_1^\mu)\right] + \delta\right). \tag{115}$$

For $y \in A_n^c$, (115) can be re-written as

$$|r_n^\mu(y)| \leq \frac{1}{6}\beta y^2 \left(\mathbb{E}\left[\kappa(X_1^\mu)\right] + \delta\right). \tag{116}$$

On the right side of (114), we have

$$\lim_{n\to\infty} \theta''(\widehat{m}_n^\mu)\widehat{m}_n^\mu - \Lambda''(\widehat{m}_n^\mu) = \frac{1}{(\sigma^\mu)^2} \tag{117}$$

a.s. uniformly in $\mu \in \mathcal{I}'$, which follows from the Cramér-Rao asymptotic lower bound for the variance of the MLE. Now, defining $c_0 := \inf_{w\in\mathcal{I}'} 1/(\sigma^w)^2$ and recognizing that $c_0 > 0$, we can choose $\beta > 0$ to satisfy:

$$-\frac{1}{2}c_0 + \frac{1}{6}\beta\left(\mathbb{E}\left[\kappa(X_1^\mu)\right] + \delta\right) = -\frac{1}{4}c_0.$$

Then, using (116) and (117), we have from (114) that

$$\sup_{y\in A_n^c} \frac{\exp\left(L_n^\mu(\widehat{m}_n^\mu + y/\sqrt{n}) - L_n^\mu(\widehat{m}_n^\mu)\right)}{\exp(-c_0 y^2/4)} \leq 1 \tag{118}$$

for sufficient large n, a.s. uniformly in $\mu \in \mathcal{I}'$. Thus, on A_n^c, $D_n^\mu(y)$ (as defined in (106)) is dominated by an integrable function for sufficiently large n, a.s. uniformly in $\mu \in \mathcal{I}'$. Furthermore, from (115) we have for any fixed y that $r_n^\mu(y) \to 0$ as $n \to \infty$, a.s. uniformly in $\mu \in \mathcal{I}'$. This, together with (111), (114) and (117), we have for any fixed y that $D_n^\mu(y) \to 0$ as $n \to \infty$, a.s. uniformly in $\mu \in \mathcal{I}'$. Then, by the dominated convergence theorem, the quantity in (113) converges to zero as $n \to \infty$, a.s. uniformly in $\mu \in \mathcal{I}'$. □

Appendix C: Gaussian Approximations for the Bootstrap

In Proposition 6 below, we develop a Gaussian approximation for bootstrapping the sample 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 an open interval. The only requirement on the P^μ is that the condition in (119) 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 6 holds almost surely and uniformly over data-generating distributions P^μ with $\mu \in \mathcal{I}$. See Remark 4 and Definition 4 in Appendix C for a precise description of this mode of convergence.

Proposition 6. *Suppose that*

$$\lim_{y\to\infty} \sup_{\mu\in\mathcal{I}} \mathbb{E}^\mu[X^2 \mathbb{I}\left(X^2 > y\right)] = 0. \tag{119}$$

Then,

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left| \mathbb{P}\left(\sqrt{n}\left(\widehat{m}_n^{*\mu} - \widehat{m}_n^\mu\right) \leq x \mid X_1^\mu, \ldots, X_n^\mu\right) - \Phi\left(\frac{x}{\sigma^\mu}\right) \right| = 0, \tag{120}$$

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

Proof of Proposition 6. 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 $\{\mathbb{I}(\cdot \leq x), x \in \mathbb{R}\}$ 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 (119) and Lemma 4, 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 (120) then follows. □

Appendix D: Weak Convergence Technical Lemmas

LEMMA 6 **(Generalized Continuous Mapping Theorem)**. *Let f and f^n, $n \geq 1$, be measurable functions that map from the metric space $(\mathcal{S}_1, r_1)$ to the separable metric space $(\mathcal{S}_2, r_2)$. Let E be the set of $x \in \mathcal{S}_1$ such that $f^n(x^n) \to f(x)$ fails for some sequence x^n, $n \geq 1$, with $x^n \to x$ in $\mathcal{S}_1$. If $\xi^n \Rightarrow \xi$ in $(\mathcal{S}_1, r_1)$ and $P(\xi \in E) = 0$, then $f^n(\xi^n) \Rightarrow g(\xi)$ in $(\mathcal{S}_2, r_2)$. (See Theorem 3.4.4 of Whitt (2002).)*

LEMMA 7 **(Tightness of Multi-dimensional Processes)**. *A sequence of process $\xi^n = (\xi_1^n, \ldots, \xi_d^n)$ is tight in $D^d[0,\infty)$ if and only if each ξ_j^n and each $\xi_j^n + \xi_k^n$ are tight in $D[0,\infty)$, for all $1 \leq j, k \leq d$. (See Problem 22 of Chapter 3 of Ethier and Kurtz (1986).)*

LEMMA 8 **(Simple Sufficient Conditions for Tightness)**. *A sequence of processes ξ^n in $D[0,\infty)$ adapted to filtrations $(\mathcal{F}_t^n, t \geq 0)$ is tight if, for each $T > 0$,*

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

and there exists a collection of non-negative random variables $\{A_\delta^n(T), n \geq 1, \delta > 0\}$ such that

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

almost surely for $0 \leq t \leq T$ and $0 \leq u \leq \delta$, and

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

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

LEMMA 9 **(Martingale Functional Central Limit Theorem)**. *For each n, let $Y^n(i) \in \mathbb{R}^m$ be a martingale difference sequence adapted to the filtration $\mathcal{F}_i^n$ for $i = 1, 2, \ldots$. Suppose for any $t > 0$, the following conditions (M1) and (M2) hold as $n \to \infty$.*
There exists a symmetric positive-definite matrix Σ such that

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

For any $\epsilon > 0$ and each component $k = 1, \ldots, m$,

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

Then,

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

in $D[0,\infty)$, where B is m-dimensional Brownian motion with covariance matrix Σ.

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.

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).

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 (2024) The Fragility of Optimized Bandit Algorithms. *Operations Research* .

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.

Korda N, Kaufmann E, Munos R (2013) Thompson Sampling for One-dimensional Exponential Family Bandits. *NeurIPS* 26.

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).

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.