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 process-level dynamics of Thompson sampling in the "small gap" regime. The small gap regime is one in which the gaps between the arm means are of order $\sqrt{\gamma}$ or smaller and the time horizon is of order $1/\gamma$, where γ is small. As $\gamma \downarrow 0$, we show that the process-level dynamics of Thompson sampling converge weakly to the solutions to certain stochastic differential equations and stochastic ordinary differential equations. Our weak convergence theory, which is developed from first principles using the Continuous Mapping Theorem, can handle stationary, weakly dependent reward processes, and can also be adapted to analyze a variety of sampling-based bandit algorithms. Indeed, we show that the process-level dynamics of many sampling-based bandit algorithms—including Thompson sampling designed for any single-parameter exponential family of rewards, as well as non-parametric bandit algorithms based on bootstrap re-sampling—satisfy an invariance principle. Namely, their weak limits coincide with that of Gaussian parametric Thompson sampling with Gaussian priors. Moreover, in the small gap regime, the regret performance of these algorithms is generally insensitive to model mis-specification, changing continuously with increasing degrees of mis-specification.

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

History: Manuscript version – December 30, 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 has not "kicked 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" regime in which the sub-optimality gaps Δ_k are of order $\sqrt{\gamma}$ (or smaller), with $\gamma \downarrow 0$, and in which the total number n of arm plays is large and of order $1/\gamma$. 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 small gap regime corresponds to so-called minimax or worst-case settings in the bandit literature, and is one of the 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 close to 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. For versions of the *Gaussian Thompson sampler*, which is the TS principle implemented using the posterior updating mechanics of Gaussian priors and likelihoods, Kuang and Wager (2024) independently developed similar SDE and stochastic ODE characterizations as part of a general framework for analyzing sampling-based bandit algorithms in the small gap regime. However, directly compared to our two main contributions below, 1) their weak convergence theory is developed using a different theoretical approach, which restricts their study to iid reward processes and does not provide a transparent view of how diffusion limits arise, and 2) their study of Thompson sampling is restricted to the Gaussian Thompson sampler. See Section 1.1 for a detailed comparison of their work to ours.

1) In the small gap regime, we develop distributional approximations for the process-level dynamics of the Gaussian Thompson sampler, which have an SDE representation (Theorems 1 and 4) and also a stochastic ordinary differential equation (ODE) representation (Theorems 2 and 4). These diffusion approximations essentially only require that the centered and suitably re-scaled reward processes converge weakly to Brownian motion. They do not require the rewards themselves to be Gaussian or even iid; our Theorem 2 is developed for stationary, weakly dependent reward processes. Also, notably, our proof approach for these theorems shows explicitly how the SDE and stochastic ODE weak limits arise. In particular, 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 (CMT) and elementary arguments to obtain the SDEs and stochastic ODEs. Intuitive sketches of our proof approach are provided in Sections 3.1 and 3.2. Additionally, we fully establish the distributional equivalence between the SDE and stochastic ODE representations by showing how an SDE solution can be expressed as a stochastic ODE solution (Proposition 1), and vice versa (Theorem 3).

2) We 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 a general single-parameter exponential family likelihood and a general (bounded support) prior distribution (Theorem 5). 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 (Theorem 6). In the small gap regime, 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, the Gaussian Thompson sampler provides general insight about the behavior of the many versions of TS and related sampling-based bandit algorithms studied in the literature. Additionally, in the small gap regime, the regret performance of these algorithms is insensitive to mis-specification of reward distributions, changing continuously with increasing degrees of mis-specification (Proposition 2). This contrasts with the bandit setting of Lai and Robbins (1985), corresponding to time scales that are large relative to $1/\gamma$, over which optimized bandit algorithms achieve expected regret growing logarithmically with time. In that setting, regret is sensitive to small probability mis-identification of the optimal arm, and the algorithms become highly sensitive to even slight degrees of model mis-specification (see Fan and Glynn (2024)).

The rest of the paper is structured as follows. Related work is further discussed in Section 1.1. We then introduce 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 in Theorem 1 for the Gaussian Thompson sampler and iid reward processes, with the proofs given in Section 5.1. Similarly, in Section 3.2, we provide an intuitive derivation leading to the stochastic ODE convergence result in Theorem 2 for the Gaussian Thompson sampler and general stationary reward processes, with the proofs given in Section 5.2. Along the way, we develop Theorem 3 and Proposition 1, which establish distributional equivalences between the solution to the SDE and solutions to the corresponding stochastic ODE. We develop extensions in Corollary 1 and Theorem 4 to the diffusion approximations in Section 3.3. In Section 4.1, we show via Theorem 5 that the EF Thompson sampler has the same weak limit in the small gap regime as the Gaussian Thompson sampler. In Section 4.2, the same is shown via Theorem 6 for the bootstrap sampler. In Section 4.3, we study the insensitivity (in the small gap regime) of these sampling-based bandit algorithms to mis-specification of the reward distribution in Proposition 2. 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 and concurrent work of Kuang and Wager (2024) (abbreviated KW in the discussion below), which was posted on arXiv (Wager and Xu 2021) prior to our manuscript (Fan and Glynn 2021). Our current paper is based on an updated and expanded version of that initial arXiv posting, and also forms Chapter 4 of the first author's PhD dissertation (Fan 2023).

The overlap between our work and KW is that both obtain similar SDE and stochastic ODE approximations for the dynamics of the Gaussian Thompson sampler in the small gap regime with $\sqrt{\gamma}$-scale sub-optimality gaps over time horizons of order $1/\gamma$. In terms of formal theoretical results, the overlap is essentially between our Theorem 1 and KW Theorem 1 (applied to the Gaussian Thompson sampler), and also our Proposition 1 and KW Theorem 3. Below, we discuss in detail the key differences between these results.

First of all, as mentioned in the Introduction, the theoretical approach taken in our paper to establish these results differs from the approach taken by KW. KW expresses the dynamics of sampling-based algorithms belonging to their *Sequentially Randomized Markov Experiments* framework, which includes TS, as Markov chains, and uses 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 CMT that the discrete systems converge weakly to their continuous counterparts.

One advantage of our CMT approach is that it directly accommodates stationary, weakly dependent reward processes; see Theorem 2, proved under Assumption 2. In the rested bandit formulation, our analysis separates (i) weak convergence of the per-arm reward processes to Brownian motions from (ii) the algorithm's sampling behavior. By contrast, the generator-based analysis of KW is developed for sequentially randomized Markov experiments with iid reward draws. Extending that approach to weakly dependent rewards would require augmenting the state so that the adaptive system becomes Markovian, and then verifying generator convergence for the enlarged process—a technically challenging development not undertaken in KW.

Another advantage of our CMT approach, relative to the generator approach of KW, is that it offers a transparent and intuitive view of how the diffusion approximations arise. We restrict the scope of our paper to TS and related sampling-based algorithms, but our CMT approach can be directly applied to obtain diffusion approximations for the Sequentially Randomized Markov Experiments algorithm class of KW, with the transparent and intuitive view extending to the analysis of all algorithms within that class.

Furthermore, as mentioned previously, we fully establish the distributional equivalence between the SDE and stochastic ODE representations by showing how an SDE solution can be expressed as a stochastic ODE solution (Proposition 1), and vice versa (Theorem 3). On the other hand, KW establishes only one half of the connection (via KW Theorem 3), that the solution to the SDE in KW Theorem 1 is a solution to a corresponding stochastic ODE. (The proof of our Proposition 1 also differs from that of KW Theorem 3.)

Additionally, related to both our work and also KW, Kalvit and Zeevi (2021) has recently studied the behavior of the UCB1 algorithm of Auer et al. (2002) in worst-case/minimax gap regimes. When the gaps between arm means scale as $\sqrt{\log(n)/n}$ with the horizon n, they obtain diffusion approximations for UCB1. Furthermore, they highlight sharp distinctions between the behavior of TS and UCB algorithms when the gap sizes are $o(\sqrt{\log(n)/n})$ (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. The actual data presented to the algorithm will be assumed iid or weakly dependent. (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 $j=1,2,\dots$ 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 ways of generating reward feedback. For each arm $k\in[K]$, let $X_k(i)$, $i=1,2,\dots$ be the sequence of rewards for the arm. The first way is called the *random table model*, where at time j, $Y(j)=X_k(j)$ for the selected arm $k\in[K]$ $(I_k(j)=1)$. The second way is called 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.)

When the reward sequence for each arm is iid with independence across arms, the random table model and the reward stack model generate reward feedback in distributionally equivalent ways. We will also consider the setting where for each arm $k \in [K]$, the rewards $X_k(i)$, $i = 1, 2, \ldots$ are a *stationary sequence* (which allows for serial dependence), i.e., for any fixed integers $1 \leq i_1 \leq i_2 \leq \cdots \leq i_l < \infty$, the finite-dimensional distributions $(X_k(i_1+j), X_k(i_2+j), \ldots, X_k(i_l+j))$ are the same for all $j \geq 0$. In the stationary rewards setting, we will use the reward stack model. This leads to a *rested bandit*, where the reward process for each arm is in steady state and evolves according to a stochastic process only when the arm is played, otherwise staying "frozen".

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, both for iid and stationary reward processes.

Function Spaces and Weak Convergence

Throughout this paper, $D_m[a, \infty)$ denotes the space of functions with domain $[a, \infty)$ and range $\mathbb{R}^m$ (writing $D[a, \infty)$ when $m = 1$), that are right-continuous and have limits from the left. We use the Skorohod metric on this space, as defined in Chapter 3, Section 5 of Ethier and Kurtz (1986). 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 Chapter 3 of Ethier and Kurtz (1986).

Small Gap 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 any reward feedback mechanism, we will use the discrete-time filtration $\mathcal{H}^\gamma = (\mathcal{H}_j^\gamma, j \geq 0)$, with

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

to keep track of the algorithm's information. Below, we state and discuss two setups for the small gap regime (Assumptions 1 and 2) that we will use to develop our limit theory throughout the paper.

In the iid rewards setting, we will work under Assumption 1, given as follows.

ASSUMPTION 1 (Small Gap Regime with IID Rewards). *For each* γ *and each arm* $k \in [K]$*, the rewards* $X_k^\gamma(i)$*,* $i = 1, 2, \ldots$ *are independent across both* i *and* k*, and are distributed according to* Q_k^γ*, with mean* μ_k^γ *and variance* $(\sigma_k^\gamma)^2$*. 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*

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

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

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

REMARK 1. For the iid rewards setting, it suffices for our analysis to have finite $2 + \alpha$ (with arbitrarily small $\alpha > 0$) moments for the rewards suffices (as in (7)), while the theoretical approach of Kuang and Wager (2024) requires finite fourth moments.

In the small gap regime setup 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 small gap regime 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 fewer times results in their means being essentially indistinguishable. The conditions in Assumption 1 enable the reward processes to be well-approximated by Brownian motions.

In the stationary rewards setting, we will work under Assumption 2, given as follows.

ASSUMPTION 2 (Small Gap Regime with Stationary Rewards). *For each* γ *and each arm* $k \in [K]$*, the rewards* $X_k^\gamma(i)$*,* $i = 1, 2, \ldots$ *are a stationary sequence with mean* μ_k^γ *(which satisfies the scaling in (5) from Assumption 1), with independence across different arms* k*. For each arm* k*, there exists* $\sigma_k > 0$*, such that the family of processes (indexed by* γ*)*

$$Z_k^\gamma(t) = \sqrt{\gamma} \frac{1}{\sigma_k} \sum_{i=1}^{\lfloor t/\gamma \rfloor} (X_k^\gamma(i) - \mu_k^\gamma) \tag{8}$$

is tight in $D[0, \infty)$*, and for any continuous function* $f : \mathbb{R} \to \mathbb{R}$ *vanishing at infinity, and any* $u \in [0, \infty)^K$*,* $v > 0$*,*

$$\lim_{\gamma \downarrow 0} \mathbb{E}\left[\left| \mathbb{E}\left[f(Z_k^\gamma(u_k + v)) \mid \mathcal{F}_u^\gamma \right] - \mathbb{E}[f(Z_k^\gamma(u_k) + \sqrt{v}\mathcal{N})] \right| \right] = 0, \tag{9}$$

where $\mathcal{N}$ *is a standard Gaussian random variable independent of* Z_k^γ*, and the filtration* $\mathcal{F}^\gamma = (\mathcal{F}_u^\gamma, u \in [0, \infty)^K)$ *is defined as*

$$\mathcal{F}_u^\gamma = \sigma(X_k^\gamma(i_k), i_k \le \lfloor u_k/\gamma \rfloor, k \in [K]). \tag{10}$$

Remark 2. The condition in (9) implies that the finite-dimensional distributions of the Z_k^γ converge to the corresponding multivariate normal distributions. The condition in (9) also ensures weak convergence to "non-anticipative" (see Definition 2) solutions to the limit stochastic ODEs. Together, the tightness assumption and (9) ensure that $Z_k^\gamma \Rightarrow B_k$ in $D[0,\infty)$ as $\gamma \downarrow 0$, where B_k is standard Brownian motion. Here, because the limit processes are Brownian motions with continuous sample paths, it suffices to verify tightness in $D[0,\infty)$ of the family Z_k^γ in (8) separately for each $k \in [K]$, instead of verifying tightness in $D_K[0,\infty)$ (using, for example, Lemma 8). In (8), σ_k^2 is the long-run variance constant corresponding to the time-average of the stationary reward process; normalizing by σ_k in (8) ensures weak convergence to standard Brownian motion. If the arm rewards are not just stationary but also independent, then the conditions of Assumption 2 follow from those of Assumption 1.

Key Processes for Describing Dynamics

Here, we record the key processes that will be used throughout the paper for describing the dynamics of TS and related algorithms in the small gap regime. To keep track of the amount of sampling effort allocated to each of the K arms, we use $U^\gamma = (U_k^\gamma, k \in [K]) \in D_K[0,\infty)$, defined as:

$$U_k^\gamma(t) = \gamma \sum_{i=1}^{\lfloor t/\gamma \rfloor} I_k^\gamma(i), \tag{11}$$

which is a re-scaling of (2). To keep track of the rewards received for each arm under the random table model, we will use $S^\gamma = (S_k^\gamma, k \in [K]) \in D_K[0,\infty)$, defined as:

$$S_k^\gamma(t) = \sqrt{\gamma}\frac{1}{\sigma_k} \sum_{i=1}^{\lfloor t/\gamma \rfloor} I_k^\gamma(i)(X_k^\gamma(i) - \mu_k^\gamma), \tag{12}$$

which is a centering and re-scaling of (3). To keep track of the rewards received for each arm under the reward stack model, we will use $Z^\gamma \circ U^\gamma = (Z_k^\gamma(U_k^\gamma), k \in [K]) \in D_K[0,\infty)$, where

$$Z_k^\gamma(U_k^\gamma(t)) = \sqrt{\gamma}\frac{1}{\sigma_k} \sum_{i=1}^{U_k^\gamma(t)/\gamma} (X_k^\gamma(i) - \mu_k^\gamma), \tag{13}$$

with $Z^\gamma = (Z_k^\gamma, k \in [K])$ and each Z_k^γ as defined in (8), and with U_k^γ as defined in (11). The processes in (13), like those in (12), are also a centering and re-scaling of (3). (For vector-valued functions f and g, we use $f \circ g$ to denote component-wise composition of f and g.) In (12) and (13), the μ_k^γ and σ_k are the means and scaling factors from either Assumption 1 or Assumption 2.

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. For the Gaussian likelihood, we use a fixed variance $c_*^2 > 0$, which may or may not correspond to the σ_k^2, the limit variances (in (6)) or long-run variances (in (8)), of the arm reward processes. Later in the paper, we will complement the theory of this section by studying EF Thompson samplers in Section 4.1, the bootstrap sampler in Section 4.2, and then model mis-specification issues in Section 4.3.

Before continuing on to the derivation of diffusion approximations, we first discuss a technical issue that can arise. The sampling behavior of TS can be highly erratic at the very beginning of a bandit experiment in the small gap regime (as in Assumptions 1 and 2) when little data has been collected and the algorithm is only performing exploration. 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 (and stochastic ODEs) have unique solutions. Below, we discuss two ways of "smoothing" the initial behavior of TS to restore suitable Lipschitz continuity of the SDEs.

1) Smoothing via Concentrated Priors

One way to smooth out the initial behavior of TS is to use concentrated priors on the arm means. We use this approach in Sections 3.1 and 3.2. In Assumptions 1 and 2, 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. (Recall that $\Delta_k^\gamma \to \Delta_k$ for some $\Delta_k > 0$ as $\gamma \downarrow 0$.) We assume that the centering parameter μ_* and scale parameter γ are known, and we use an independent $N(\mu_*, \gamma/b)$ prior for each arm in the Gaussian Thompson sampler, with a chosen fixed $b > 0$. Translated into practice, this means that the experimenter knows, perhaps from related experiments run in the past, that the arm means are in a "small ($\sqrt{\gamma}$-scale) neighborhood" of some μ_*. So, the experimenter chooses priors with probability mass concentrated in such a neighborhood of μ_*. Then, over time horizons scaling as $1/\gamma$, the bandit algorithm begins to identify 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 suitable 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.

Remark 3. The Gaussian Thompson sampler with concentrated priors on the arm means is not translation-invariant. To avoid writing the centering parameter μ_* repeatedly in the derivations for Theorems 1 and 2 only, there we simply set $\mu_* = 0$. For all other results, there is translation invariance, and μ_* always gets canceled out.

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 suitable 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 centered around μ_* with variance scaling as γ, after which TS using the learned prior is deployed.

3.1. SDE Approximation

In this section, we work under Assumption 1 with iid rewards for each arm, and we use the random table model of reward feedback, as introduced in Section 2. This leads to the SDE approximation for the Gaussian Thompson sampler in Theorem 1 below.

To begin, we show that the dynamics in this setting can be described by the evolution of the processes (U^γ, S^γ) as defined in (11) and (12). 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:

$$\widetilde{\mu}_k^\gamma(j+1) \sim N\left(\frac{\gamma \sum_{i=1}^{j} I_k^\gamma(i) X_k^\gamma(i)}{U_k^\gamma(j\gamma) + bc_*^2}, \frac{c_*^2 \gamma}{U_k^\gamma(j\gamma) + bc_*^2}\right). \tag{14}$$

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

$$\mathbb{P}\left(k = \underset{l \in [K]}{\arg\max}\, \widetilde{\mu}_l^\gamma(j+1) \,\Big|\, \mathcal{H}_j^\gamma\right) \tag{15}$$

$$= \mathbb{P}\left(k = \underset{l \in [K]}{\arg\max} \left\{\frac{S_l^\gamma(j\gamma)\sigma_l + U_l^\gamma(j\gamma) d_l^\gamma}{U_l^\gamma(j\gamma) + bc_*^2} + \frac{c_*}{\sqrt{U_l^\gamma(j\gamma) + bc_*^2}} \mathcal{N}_l\right\} \,\Bigg|\, U^\gamma(j\gamma), S^\gamma(j\gamma)\right) \tag{16}$$

$$= p_k^\gamma(U^\gamma(j\gamma), S^\gamma(j\gamma)), \tag{17}$$

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

$$p_k^\gamma(u, s) = \mathbb{P}\left(k = \underset{l \in [K]}{\arg\max} \left\{\frac{s_l \sigma_l + u_l d_l^\gamma}{u_l + bc_*^2} + \frac{c_*}{\sqrt{u_l + bc_*^2}} \mathcal{N}_l\right\}\right). \tag{18}$$

We can then re-express $U_k^\gamma(t)$ and $S_k^\gamma(t)$ from (11)-(12) as

$$U_k^\gamma(t) = \gamma \sum_{i=0}^{\lfloor t/\gamma \rfloor - 1} p_k^\gamma(U^\gamma(i\gamma), S^\gamma(i\gamma)) + M_k^\gamma(t) \tag{19}$$

$$S_k^\gamma(t) = \sum_{i=0}^{\lfloor t/\gamma \rfloor - 1} \sqrt{p_k^\gamma(U^\gamma(i\gamma), S^\gamma(i\gamma))}\left(B_k^\gamma((i+1)\gamma) - B_k^\gamma(i\gamma)\right), \tag{20}$$

where $M^\gamma = (M_k^\gamma, k \in [K]) \in D_K[0,\infty)$ and $B^\gamma = (B_k^\gamma, k \in [K]) \in D_K[0,\infty)$ are defined as:

$$M_k^\gamma(t) = \gamma \sum_{i=0}^{\lfloor t/\gamma \rfloor - 1} \left(I_k^\gamma(i+1) - p_k^\gamma(U^\gamma(i\gamma), S^\gamma(i\gamma))\right) \tag{21}$$

$$B_k^\gamma(t) = \sqrt{\gamma}\frac{1}{\sigma_k} \sum_{i=0}^{\lfloor t/\gamma \rfloor - 1} \frac{I_k^\gamma(i+1)(X_k^\gamma(i+1) - \mu_k^\gamma)}{\sqrt{p_k^\gamma(U^\gamma(i\gamma), S^\gamma(i\gamma))}}, \tag{22}$$

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 converges weakly to the $D_K[0,\infty)$ zero process, and B^γ converges weakly to standard K-dimensional Brownian motion. Additionally, since $d_k^\gamma \to d_k$ from (5), we have

$$p_k^\gamma(u,s) \to p_k(u,s) \tag{23}$$

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

$$p_k(u,s) = \mathbb{P}\left(k = \arg\max_{l \in [K]} \left\{ \frac{s_l \sigma_l + u_l d_l}{u_l + bc_*^2} + \frac{c_*}{\sqrt{u_l + bc_*^2}} \mathcal{N}_l \right\}\right). \tag{24}$$

Thus, we expect (19)-(20) to be a discrete approximation to the SDE in integral form:

$$U_k(t) = \int_0^t p_k(U(v), S(v)) dv \tag{25}$$

$$S_k(t) = \int_0^t \sqrt{p_k(U(v), S(v))} dB_k(v), \quad k \in [K] \tag{26}$$

with standard K-dimensional Brownian motion B.

To conclude the above derivation, the formal 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 CMT, 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 (19)-(20) to the limit in (25)-(26).

As mentioned previously, for each function p_k in (24), both p_k and $\sqrt{p_k}$ are *locally Lipschitz continuous*, which means that they are Lipschitz continuous on any compact subset of the domain

$[0,\infty)^K \times \mathbb{R}^K$. In particular, in (24), we have $b > 0$ resulting from the use of concentrated priors in the pre-limit, which ensures local Lipschitz continuity. These functions are also bounded, which together with local Lipschitz continuity, ensures that the SDEs in (25)-(26) have a (global) unique strong solution, as summarized in Remark 4 below.

Remark 4. For the definition of a strong solution to an SDE and uniqueness of strong solutions, see Definitions 2.1 and 2.3 from Chapter 5.2 of Karatzas and Shreve (1998). The existence of a strong solution under boundedness (and thus (sub-)linear growth) and local Lipschitz continuity of the p_k and $\sqrt{p_k}$ functions follows from Theorem 3.11 from Chapter 5 of Ethier and Kurtz (1986). Uniqueness follows from Theorem 2.5 from Chapter 5.2 of Karatzas and Shreve (1998).

Before stating Theorem 1, in Remark 5 below, we note an expression for regret that will be used throughout the rest of the paper.

Remark 5. 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 by:

$$\mathrm{Reg}^\gamma(n) = \frac{1}{\sqrt{\gamma}} \sum_{k\in[K]} U_k^\gamma(n\gamma)\Delta_k^\gamma. \tag{27}$$

Theorem 1. *Consider a K-armed bandit in the small gap regime of Assumption 1 (with iid rewards for each arm) and the random table model of reward feedback. For the Gaussian Thompson sampler with concentrated priors on the arm means, we have*

$$(U^\gamma, S^\gamma) \Rightarrow (U, S) \tag{28}$$

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

$$dU_k(t) = p_k(U(t), S(t))dt \tag{29}$$

$$dS_k(t) = \sqrt{p_k(U(t), S(t))}dB_k(t) \tag{30}$$

$$U_k(0) = S_k(0) = 0, \quad k \in [K], \tag{31}$$

with standard K-dimensional Brownian motion B, and functions p_k as expressed in (24).

Moreover, for regret,

$$\sqrt{\gamma}\,\mathrm{Reg}^\gamma(\lfloor \cdot/\gamma \rfloor) \Rightarrow \sum_{k\in[K]} U_k(\cdot)\Delta_k \tag{32}$$

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

3.2. Stochastic ODE Approximation

In this section, we work under more general conditions than in Section 3.1, where we used Assumption 1 with iid rewards for each arm. Here, we work under Assumption 2 with general stationary sequences of rewards for each arm, and we use the reward stack model of reward feedback, as introduced in Section 2. As discussed in Section 2, we can think of this setup as a rested bandit, where the rewards for each arm evolve according to a stochastic process when the arm is played and stays frozen otherwise. This leads to the stochastic ODE approximation for the Gaussian Thompson sampler in Theorem 2. To conclude this section, via Theorem 3 and Proposition 1, we establish the distributional equivalence between general SDE and stochastic ODE limit representations.

Similar to the derivation of the SDE approximation, we first show that the dynamics can be described by the evolution of the processes $(U^\gamma, B^\gamma \circ U^\gamma)$ as defined in (11) and (13). 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:

$$\widetilde{\mu}_k^\gamma(j+1) \sim N\left(\frac{\gamma \sum_{i=1}^{U_k^\gamma(j\gamma)/\gamma} X_k^\gamma(i)}{U_k^\gamma(j\gamma)+bc_*^2}, \frac{c_*^2\gamma}{U_k^\gamma(j\gamma)+bc_*^2}\right). \tag{33}$$

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

$$\begin{aligned}
&\mathbb{P}\left(k = \arg\max_{l\in[K]} \widetilde{\mu}_l^\gamma(j+1) \,\middle|\, \mathcal{H}_j^\gamma\right) &(34)\\
&= \mathbb{P}\left(k = \arg\max_{l\in[K]}\left\{\frac{Z_l^\gamma(U_l^\gamma(j\gamma))\sigma_l + U_l^\gamma(j\gamma)d_l^\gamma}{U_l^\gamma(j\gamma)+bc_*^2} + \frac{c_*}{\sqrt{U_l^\gamma(j\gamma)+bc_*^2}}\mathcal{N}_l\right\} \,\middle|\, U^\gamma(j\gamma), Z^\gamma\circ U^\gamma(j\gamma)\right) &(35)\\
&= p_k^\gamma(U^\gamma(j\gamma), Z^\gamma\circ U^\gamma(j\gamma)), &(36)
\end{aligned}$$

where the $\mathcal{N}_l$ are independent standard Gaussian random variables, and functions p_k^γ are given by (24).

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

$$U_k^\gamma(t) = \gamma \sum_{i=0}^{\lfloor t/\gamma\rfloor-1} p_k^\gamma(U^\gamma(i\gamma), Z^\gamma\circ U^\gamma(i\gamma)) + M_k^\gamma(t), \quad k\in[K], \tag{37}$$

where $M^\gamma = (M_k^\gamma,\, k\in[K]) \in D_K[0,\infty)$ is defined as:

$$M_k^\gamma(t) = \gamma \sum_{i=0}^{\lfloor t/\gamma\rfloor-1} \left(I_k^\gamma(i+1) - p_k^\gamma(U^\gamma(i\gamma), Z^\gamma\circ U^\gamma(i\gamma))\right), \tag{38}$$

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 converges weakly to the $D_K[0,\infty)$ zero process. Moreover, as discussed in Remark 2, Z^γ converges weakly to standard K-dimensional Brownian motion. As in the previous section, the convergence in (23) holds. Thus, we expect (37) to be a discrete approximation to the stochastic ODE in integral form:

$$U_k(t) = \int_0^t p_k(U(v), B \circ U(v))dv, \quad k \in [K], \tag{39}$$

with standard K-dimensional Brownian motion B, and functions p_k as expressed in (24).

To conclude the above derivation, the formal 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 CMT, 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 (37) to the limit in (39).

The uniqueness of non-anticipative weak (in distribution) solutions to the stochastic ODE in Theorem 2 follows from Theorem 3 below (which establishes a general connection between SDE and stochastic ODE solutions), together with standard SDE strong uniqueness theory (as discussed previously in Remark 4). Before stating Theorem 2, we specify below in Definition 2 what it means to be a non-anticipative solution to the stochastic ODE. In Definition 3 below, we specify what it means to be a unique (in distribution) weak solution to the stochastic ODE. Definitions 2 and 3 can be found in Chapter 6, Section 2 of Ethier and Kurtz (1986). Also, one may wonder if the non-anticipative property automatically holds for all solutions to stochastic ODEs of the form in (39)—in general, the answer is no; see Remark 8 below.

Definition 2. Let B be a K-dimensional standard Brownian motion on a probability space $(\Omega, \mathbb{F}, \mathbb{P})$, and consider the augmented filtration $\mathcal{F} = (\mathcal{F}_u,\, u \in [0,\infty)^K)$, with

$$\mathcal{F}_u = \sigma\big(B_k(t_k),\, t_k \leq u_k,\, k \in [K]\big) \vee \sigma(\mathcal{L}), \tag{40}$$

where $\mathcal{L} \subset \mathbb{F}$ is the collection of all $\mathbb{P}$-probability zero sets. We say that a solution $U = (U_k,\, k \in [K])$ satisfying a.s. (almost surely) the stochastic ODE:

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

is *non-anticipative* if there exists a filtration $\mathcal{G} = (\mathcal{G}_u,\, u \in [0,\infty)^K)$ such that, denoting $\xi^u(\cdot) = (B_k(u_k + \cdot),\, k \in [K])$ for $u \in [0,\infty)^K$, the following hold.

(i) $\mathcal{F}_u \subset \mathcal{G}_u \subset \mathbb{F} \quad \forall\, u \in [0,\infty)^K$

(ii) $\mathbb{P}(\xi^u \in \cdot \mid \mathcal{G}_u) = \mathbb{P}(\xi^u \in \cdot \mid \mathcal{F}_u) \quad \forall\, u \in [0,\infty)^K$

(iii) for each $t \geq 0$, $U(t)$ is a $\mathcal{G}_u$-stopping time

REMARK 6. In (40), each Brownian motion B_k has its own separate clock/time index u_k, and the filtration $\mathcal{F}$ is indexed by the directed set $[0,\infty)^K$. For condition (iii) in Definition 2, $U(t)$ is a $\mathcal{G}_u$-stopping time means that $\bigcap_{k\in[K]}\{U_k(t)\le u_k\}\in\mathcal{G}_u$ for any $u\in[0,\infty)^K$. For details on filtrations and martingales indexed by directed sets like $[0,\infty)^K$, and their associated stopping times and optional stopping theorem, see Kurtz (1980b) or Chapter 2, Section 8 of Ethier and Kurtz (1986).

DEFINITION 3. Given a K-dimensional standard Brownian motion B on a probability space $(\Omega,\mathbb{F},\mathbb{P})$, we say that the stochastic ODE (in integral form) in (39) has a *weak solution* if there exists a probability space $(\widetilde{\Omega},\widetilde{\mathbb{F}},\widetilde{P})$ on which is defined a K-dimensional standard Brownian motion $\widetilde{B}$ and a stochastic process $\widetilde{U}$ such that $\widetilde{P}$-a.s. for all $t\ge 0$,

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

We say that a (non-anticipative) weak solution is *unique in distribution* if any two (non-anticipative) weak solutions always have the same finite-dimensional distributions.

REMARK 7. As mentioned earlier, we will use the connection between stochastic ODEs and SDEs (in Theorem 3 below) to establish uniqueness (in distribution) of non-anticipative weak solutions to stochastic ODEs. However, tightness alone is sufficient to guarantee *existence* of non-anticipative weak solutions to stochastic ODEs of the form in (39), as long as the functions p_k are continuous. See Corollary 3.6 from Chapter 6, Section 3 of Ethier and Kurtz (1986).

REMARK 8. Given a K-dimensional standard Brownian motion B on a probability space $(\Omega,\mathbb{F},\mathbb{P})$, if a solution U to the stochastic ODE in (39) is not unique for $\mathbb{P}$-almost all sample paths of B, then it may not be non-anticipative. Thus, if we have uniqueness in distribution but not $\mathbb{P}$-a.s., then the non-anticipative property requires separate treatment. See Remark 2.3 and Problem 1 from Chapter 6 of Ethier and Kurtz (1986) for an example.

THEOREM 2. *Consider a K-armed bandit in the small gap regime of Assumption 2 (with stationary rewards for each arm) and the reward stack model of reward feedback. For the Gaussian Thompson sampler with concentrated priors on the arm means, we have*

$$(U^\gamma,Z^\gamma\circ U^\gamma)\Rightarrow(U,B\circ U) \tag{41}$$

as $\gamma\downarrow 0$ in $D_{2K}[0,\infty)$, where U is the unique (in distribution) non-anticipative weak solution to the stochastic ODE:

$$dU_k(t)=p_k(U(t),B\circ U(t))dt \tag{42}$$

$$U_k(0)=0,\quad k\in[K], \tag{43}$$

with standard K-dimensional Brownian motion B, and functions p_k as expressed in (24).

Moreover, for regret, (32) holds in this stochastic ODE setting.

REMARK 9. In the special case that the rewards for each arm are iid (not just stationary), then the current setup (Assumption 2 with the reward stack model) leading to the stochastic ODE representation in Theorem 2 is probabilistically equivalent to the setup used in Section 3.1 (Assumption 1 with the random table model) leading to the SDE representation in Theorem 1. In particular, under iid rewards, the processes $S_k^\gamma(t)$ (defined in (12)) and the processes $Z_k^\gamma(U_k^\gamma(t))$ (defined in (13)) have the same distribution. The processes $U_k^\gamma(t)$ are also defined in exactly the same way in both cases. Thus, under iid rewards, the weak limits, i.e., the unique strong solution to the SDE in Theorem 1 and the unique (in distribution) weak solution to the stochastic ODE in Theorem 2, must also have the same distribution.

In Theorem 3 below, we work with the limit processes and establish in general that the uniqueness of the strong solution to an SDE implies the uniqueness (in distribution) of non-anticipative weak solutions to the corresponding stochastic ODE. As mentioned earlier, the solution uniqueness to the stochastic ODE in Theorem 2 is obtained from Theorem 3. The proof of Theorem 3 is provided in Section 5.2.

THEOREM 3. *Suppose U is a non-anticipative solution to the stochastic ODE:*

$$dU_k(t) = p_k(U(t), B \circ U(t))dt \tag{44}$$

$$U_k(0) = 0, \quad k \in [K], \tag{45}$$

for functions $p_k : [0,\infty)^K \times \mathbb{R}^K \to (0,1)$ and a standard K-dimensional Brownian motion B. Then, there exists a standard K-dimensional Brownian motion $\widetilde{B}$ such that (U,S), with $S = B \circ U$, is a solution to the SDE:

$$dU_k(t) = p_k(U(t), S(t))dt \tag{46}$$

$$dS_k(t) = \sqrt{p_k(U(t), S(t))}\,d\widetilde{B}_k(t) \tag{47}$$

$$U_k(0) = S_k(0) = 0, \quad k \in [K]. \tag{48}$$

Therefore, if the SDE in (46)-(48) has a unique strong solution (U^*, S^*), then for any non-anticipative solution V to the stochastic ODE in (44)-(45), we have $(V, B \circ V) \stackrel{d}{=} (U^*, S^*)$.

Conversely, in Proposition 1, we show that we can always convert from the SDE representation to the stochastic ODE representation. This follows directly from a multivariate version (due to F.B. Knight) of the well-known result that a continuous local martingale such as a stochastic integral can be represented as a Brownian motion with a random time change. The proof of Proposition 1 is provided in Section 5.2.

PROPOSITION 1. *Let (U,S) be a solution to the SDE in (46)-(48), with independent standard K-dimensional Brownian motion $\widetilde{B}$ and functions $p_k : [0,\infty)^K \times \mathbb{R}^K \to (0,1)$. Then, there exists a standard K-dimensional Brownian motion B such that we have the representation $(U,S) = (U, B \circ U)$, which solves the stochastic ODE in (44)-(45), with U as a non-anticipative solution.*

3.3. Approximations Without Concentrated Priors

From the development of Theorems 1 and 2, it is important for each function p_k in (24) that both p_k and $\sqrt{p_k}$ are locally Lipschitz continuous, which together with boundedness, ensures that the limit SDE has a unique strong solution (as discussed in Remark 4) and the limit stochastic ODE has a unique (in distribution) non-anticipative weak solution. In Corollary 1, we state a result for general sampling-based bandit algorithms that does not involve locally Lipschitz continuous p_k and $\sqrt{p_k}$ limit functions. In such settings, the uniqueness theory we previously invoked no longer applies. 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 distribution, so we simply characterize their evolution equations. The justification for Corollary 1 follows directly from the proof of Theorem 2.

Corollary 1. *Consider a K-armed bandit in the small gap regime of Assumption 2 (with stationary rewards for each arm) and the reward stack model of reward feedback. For 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, 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 $(U, B \circ U)$ and satisfy the stochastic ODE:*

$$\begin{aligned} dU_k(t) &= p_k(U(t), B \circ U(t))dt \\ U_k(0) &= 0, \quad k \in [K], \end{aligned}$$

with standard K-dimensional Brownian motion B.

Next, we develop a diffusion approximation for the Gaussian Thompson sampler without use of concentrated priors with variance scaling as γ. Unlike in Sections 3.1-3.2, here the decision-maker can use any fixed Gaussian prior with no γ-dependence. (More generally, any prior can be used as long as it puts positive probability mass in a neighborhood of the centering parameter μ_* from Assumptions 1 and 2.) Then, the functions p_k in (24) become:

$$p_k(u,s) = \mathbb{P}\left(k = \underset{l \in [K]}{\arg\max}\left\{\frac{s_l \sigma_l}{u_l} + d_l + \frac{c_*}{\sqrt{u_l}}\mathcal{N}_l\right\}\right), \tag{49}$$

where the $\mathcal{N}_l$ are independent standard Gaussian random variables.

However, as discussed at the beginning of Section 3, the function $(u,s) \mapsto p_k(u,s)$ from (49), as well as $(u,s) \mapsto \sqrt{p_k(u,s)}$, are no longer Lipschitz continuous for points near $u_l = 0$, $l \in [K]$. Nevertheless, the problem with these p_k and $\sqrt{p_k}$ functions only exists for an infinitesimally small initial time interval. Whenever all inputs $U_l(t)$, $l \in [K]$ to the u_l components in (49) become strictly positive, then from that time onward, there is local Lipschitz continuity for the p_k and $\sqrt{p_k}$ functions, which together with boundedness, ensures that there is a unique strong solution to the SDE. In Theorem 4, we use ϵ-warm-start (recall Definition 1) to ensure local Lipschitz continuity holds. The proof of Theorem 4 is a direct modification of those of Theorems 1 and 2, and is thus omitted.

THEOREM 4. *Consider the Gaussian Thompson sampler with any fixed Gaussian prior (no γ-dependence), under ϵ-warm-start (with initial sampling probabilities $q_k > 0$, $\sum_k q_k = 1$). Then, Theorems 1 and 2 hold with the functions $p_k : [0,\infty)^K \times \mathbb{R}^K \to (0,1)$ defined by:*

$$p_k(u,s) = \begin{cases} q_k & \sum_l u_l \le \epsilon \\ \mathbb{P}\left(k = \arg\max_{l\in[K]} \left\{ \frac{s_l \sigma_l}{u_l} + d_l + \frac{c_*}{\sqrt{u_l}} \mathcal{N}_l \right\}\right) & \sum_l u_l > \epsilon, \end{cases} \tag{50}$$

where the $\mathcal{N}_l$ are independent standard Gaussian random variables.

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 Theorem 5 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 a general single-parameter exponential family likelihood and a general (bounded support) prior distribution.)

In the literature, minimax or worst-case regret analysis (which is essentially the small gap regime, 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 study the Gaussian Thompson sampler, which has minimax optimal dependence of expected regret on the time horizon (Agrawal and Goyal 2013, 2017).

For Theorem 5, the main step is to establish in the small gap regime 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, with the form:

$$P^\mu(dx) = \exp(\theta(\mu)\cdot x - \Lambda(\mu))P(dx). \tag{51}$$

In (51), 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}, with finite values of the tilting parameter $\theta(\mu) \in \mathbb{R}$.

Suppose Assumption 1 holds, and that for the distributions Q_k^{γ} with means μ_k^{γ} from Assumption 1, we have $Q_k^{\gamma} = P^{\mu_k^{\gamma}}$. 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). For simplicity, suppose we know that the mean reward for all arms belong to a bounded, open interval $\mathcal{I}$, with $\underline{\mu} < \inf \mathcal{I} < \sup \mathcal{I} < \overline{\mu}$. So, $\mu_* \in \mathcal{I}$ and $\mu_k^{\gamma} \in \mathcal{I}$ for all $k \in [K]$ and sufficiently small $\gamma > 0$.

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 (fixed, with no γ-dependence) prior distribution (for the mean) with density that is bounded, supported on $\mathcal{I}$, and continuous and positive on a neighborhood of μ_*. For simplicity, we use the same prior for every arm, with independence across arms.

The above setup leads to Theorem 5 below for EF Thompson samplers in the small gap regime. The proof of Theorem 5 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 3 in Appendix B.

THEOREM 5. *Consider the above setup, with a K-armed bandit in the small gap regime of Assumption 1 (with iid rewards for each arm) and the random table model of reward feedback. Suppose the arm reward distributions belong to an exponential family of the form in (51), and that the corresponding EF Thompson sampler uses a prior with bounded density that is continuous and positive in a neighborhood of μ_*.*

Then, for the EF Thompson sampler under ϵ-warm-start (with initial sampling probabilities $q_k > 0$, $\sum_k q_k = 1$), we have processes (U^{γ}, S^{γ}) from (11)-(12):

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

as $\gamma \downarrow 0$ in $D_{2K}[0,\infty)$, where (U,S) is the unique strong solution to the SDE expressed in (29)-(31) of Theorem 1, with the functions $p_k : [0,\infty)^K \times \mathbb{R}^K \to (0,1)$ defined by:

$$p_k(u,s) = \begin{cases} q_k & \sum_l u_l \le \epsilon \\ \mathbb{P}\left(k = \arg\max_{l \in [K]} \left\{ \dfrac{s_l \sigma_*}{u_l} + d_l + \dfrac{\sigma_*}{\sqrt{u_l}} \mathcal{N}_l \right\}\right) & \sum_l u_l > \epsilon, \end{cases} \tag{52}$$

where the $\mathcal{N}_l$ are independent standard Gaussian random variables.

Furthermore, for regret, (32) continues to hold.

The conclusion of Theorem 5 (with the $p_k(u,s)$ in (52)) matches that of Theorem 4 (with the $p_k(u,s)$ in (50)) when (in the context of Theorem 4) the limit variances σ_k^2 from (6) match the variance c_*^2 used in the Gaussian likelihood of the Gaussian Thompson sampler.

Our results here also suggest that in the small gap regime, 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 in the small gap regime. 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 discussed in the Introduction, 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 largest one is played.

In Theorem 6 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 Theorem 5. But unlike Theorem 5, here the reward distributions do not need to belong to any exponential family. Here, we allow for an arbitrary family of reward distributions P^μ parameterized by mean $\mu \in \mathcal{I}$, where $\mathcal{I} \subset \mathbb{R}$ is a bounded, open interval. The only requirement on the P^μ is that the conditions in (53) and (54) are satisfied, where $(\sigma^\mu)^2$ is the variance of P^μ. Given the previously discussed optimality or near optimality of the Gaussian Thompson sampler, our results here suggest that the bootstrap sampler can be an effective means of balancing exploration and exploitation in the small gap regime, with the added benefit of not needing to make distributional assumptions.

The proof of Theorem 6 is the same as that of Theorem 5, except we use a Gaussian approximation for the bootstrapped sample mean, which is developed in Proposition 4 in Appendix C.

Theorem 6. *Consider the above setup, with a K-armed bandit in the small gap regime of Assumption 1 (with iid rewards for each arm) and the random table model of reward feedback. Suppose that*

$$\lim_{y\to\infty}\sup_{\mu\in\mathcal{I}}\mathbb{E}[(X^\mu)^2\mathbb{I}\left((X^\mu)^2>y\right)]=0, \tag{53}$$

and

$$\inf_{\mu\in\mathcal{I}}\sigma^\mu>0, \tag{54}$$

with $X^\mu\sim P^\mu$ for each $\mu\in\mathcal{I}$.

Then, for the bootstrap sampler under ϵ-warm-start (with initial sampling probabilities $q_k>0$, $\sum_k q_k=1$), we have for the processes (U^γ,S^γ) from (11)-(12):

$$(U^\gamma,S^\gamma)\Rightarrow(U,S)$$

as $\gamma\downarrow 0$ in $D_{2K}[0,\infty)$, where (U,S) is the unique strong solution to the SDE expressed in (29)-(31) of Theorem 1, with the functions $p_k:[0,\infty)^K\times\mathbb{R}^K\to(0,1)$ defined by:

$$p_k(u,s)=\begin{cases} q_k & \sum_l u_l\le\epsilon \\ \mathbb{P}\left(k=\arg\max_{l\in[K]}\left\{\dfrac{s_l\sigma_l}{u_l}+d_l+\dfrac{\sigma_l}{\sqrt{u_l}}\mathcal{N}_l\right\}\right) & \sum_l u_l>\epsilon,\end{cases} \tag{55}$$

where the $\mathcal{N}_l$ are independent standard Gaussian random variables.

Furthermore, for regret, (32) continues to hold.

Compared to Theorem 4 (with the $p_k(u,s)$ in (50)), in Theorem 6 (with the $p_k(u,s)$ in (55)), 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 (55)), compared to the $(c_*/\sqrt{u_l})\mathcal{N}_l$ terms in (50).

4.3. Model Mis-specification

In this section, we show that in the small gap regime of Assumption 1, the regret of the Gaussian Thompson sampler, and that of other TS variants like EF Thompson samplers, are insensitive to mis-specification of the reward distributions. Asymptotically, in the small gap regime, 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 1-4, 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 2 below, we establish that in the small gap regime of Assumption 1, the regret (as expressed in (27) in Remark 5) 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 in the small gap regime. This suggests that in the small gap regime, the insensitivity of TS to model mis-specification extends to other settings as well.

Proposition 2. *Let (U,S) denote the unique strong solution to the SDE (29)-(31) in Theorem 1 with the σ dependence expressed in (24). 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 $k \in [K]$.

The same holds for the unique (in distribution) non-anticipative weak solution $(U, B \circ U)$ to the stochastic ODE (42)-(43) in Theorem 2, as well as for both (U,S) and $(U, B \circ U)$ in Theorem 4 with the σ dependence expressed in (50).

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 could also be adaptively determined). Then, the algorithm is updated all at once with the data collected during the interval. For a time horizon of $O(1/\gamma)$, suppose the batch sizes are pre-determined before the start of the experiment and are $o(1/\gamma)$. Then, in the small gap regime, we would obtain weak convergence to the same SDEs and stochastic ODEs as in the setting of ordinary non-batched TS. Indeed, a time interval of $o(1/\gamma)$ in the discrete pre-limit system corresponds to (after multiplying by γ) 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 $\gamma \downarrow 0$, and each batch is not too large ($o(1/\gamma)$ time periods), then the dynamics of TS will be approximately the same as in the non-batched setting. To make this precise, we have the following corollary, whose straightforward proof is omitted.

Corollary 2. *For the Gaussian Thompson sampler with batches of size $o(1/\gamma)$, Theorems 1, 2 and 4 hold with the same conclusions as in the non-batched setting.*

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, $O(\log\log(1/\gamma))$ number of batches is necessary and sufficient (sufficient for specially designed algorithms) to achieve the optimal order of expected regret.

We can also consider settings with batches of size $O(1/\gamma)$. Fix a sequence increasing to infinity: $0 = a_0 < a_1 < a_2 < \dots$. For each $\gamma > 0$, let a_i^γ be positive integers for $i = 1, 2, \dots$ such that $\lim_{\gamma \downarrow 0} \gamma a_i^\gamma = a_i$ for each i. For batched updating, we perform TS updates in time periods a_i^γ. For initialization, in time periods $1, \dots, a_1^\gamma - 1$, each arm k is sampled with fixed probability $q_k > 0$ ($\sum_k q_k = 1$). Then, we have the following result, where the resulting SDE and stochastic ODE limits are essentially Euler discretizations of the original processes with steps (in continuous time) of size $a_{i+1} - a_i$ for $i = 1, 2, \dots$.

Corollary 3. *For the Gaussian Thompson sampler with batches of size $O(1/\gamma)$ as described above, Theorems 1, 2 and 4 hold with the SDE limit:*

$$
\begin{aligned}
&dU_k(t) = p_k(U(a_i), S(a_i))dt \\
&dS_k(t) = \sqrt{p_k(U(a_i), S(a_i))}dB_k(t), \quad t \in [a_i, a_{i+1}),\ i = 1, 2, \dots \\
&U_k(a_1) = q_k a_1, \quad S_k(a_1) = \sqrt{q_k} B_k(a_1), \quad k \in [K],
\end{aligned}
$$

and the stochastic ODE limit:

$$
\begin{aligned}
&dU_k(t) = p_k(U(a_i), B \circ U(a_i))dt, \quad t \in [a_i, a_{i+1}),\ i = 1, 2, \dots \\
&U_k(a_1) = q_k a_1, \quad k \in [K],
\end{aligned}
$$

where p_k is given by (24) or (49).

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 construction (from Section 6 of Kurtz and Protter (1991)) that can approximate functions in $D_m[0,\infty)$ uniformly with any desired accuracy. We define this *ϵ-uniform step function approximation* in Definition 4 and define integration with step function integrands in Definition 5.

Then, we introduce Lemma 1, which establishes that the ϵ-uniform step function approximation is a continuous mapping that preserves adaptedness and martingale properties. As indicated in Lemma 1, by applying the ϵ-uniform step function approximation to pre-limit and limit integrands, the corresponding approximated pre-limit and limit stochastic integrals are mathematically well-defined, and we have weak convergence of the former to the latter. Lemma 1 also helps with verification that components of the limit process are Brownian motion.

Next, we provide the proof of Theorem 1, which uses Lemma 1 together with the CMT and standard weak convergence arguments. We conclude the section with the proof of Lemma 1, followed by Lemmas 2 and 3 and their proofs, which establish tightness of stochastic processes and convergence to Brownian motion, as used in the proof of Theorem 1.

Definition 4 (ϵ-uniform 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 $z \in D_m[0,\infty)$, inductively define the stopping times τ_j starting with $\tau_0 = 0$:

$$\tau_{j+1} = \inf\{t > \tau_j : \max(\|z(t) - z(\tau_j)\|, \|z(t-) - z(\tau_j)\|) \geq \epsilon\rho_j\}, \tag{56}$$

where $\rho_j \overset{\text{iid}}{\sim} \text{Unif}(1/2, 1)$. (Note that each z has its own corresponding sequence of stopping times τ_j.) Then, define $\chi^\epsilon(z) \in D_m[0,\infty)$ by

$$\chi^\epsilon(z)(t) = z(\tau_j), \quad t \in [\tau_j, \tau_{j+1}), \tag{57}$$

so that $\chi^\epsilon(z)$ is a step function, and a.s.,

$$\sup_{t \geq 0} \|\chi^\epsilon(z)(t) - z(t)\| \leq \epsilon. \tag{58}$$

Definition 5 (Integration with Step Functions). Let $f, g \in D[a,\infty)$, where f is a step function with jump points $t_1 < \cdots < t_j$ on the interval $[a,b]$ (set $t_0 = a$ and $t_{j+1} = b$). We will always use the following definition of integration for step function integrand f with respect to integrator g on $(a,b]$:

$$\int_a^b f(t-)dg(t) = \sum_{i=0}^{j} f(t_i)\left(g(t_{i+1}) - g(t_i)\right).$$

Lemma 1 **(Properties of ϵ-uniform Step Function Approximation)**. *Let $L^n \in D_l[0,\infty)$, $H^n \in D_m[0,\infty)$ and $W^n \in D_m[0,\infty)$ be sequences of processes adapted to a corresponding sequence of filtrations $\mathcal{F}^n = (\mathcal{F}_t^n, t \geq 0)$. For any $\epsilon > 0$, let $\chi^\epsilon : D_m[0,\infty) \to D_m[0,\infty)$ be constructed as in Definition 4, and define filtrations $\mathcal{G}^n = (\mathcal{G}_t^n, t \geq 0)$ by $\mathcal{G}_t^n = \mathcal{F}_t^n \vee \widetilde{\mathcal{G}}$, where $\widetilde{\mathcal{G}} = \sigma(\rho_j, j \geq 0)$ (with the ρ_j used to construct χ^ϵ) is independent of $\mathcal{F}^n$ for all n. If W^n is an $\mathcal{F}^n$-martingale and $(L^n, H^n, W^n) \Rightarrow (L, H, W)$ in $D_{l+2m}[0,\infty)$ as $n \to \infty$, then the following hold.*

(i) $(L^n, H^n, W^n, \chi^\epsilon(H^n)) \Rightarrow (L, H, W, \chi^\epsilon(H))$ *in* $D_{l+3m}[0,\infty)$ *as* $n \to \infty$

(ii) $(L^n, H^n, W^n, \chi^\epsilon(H^n))$ *is* $\mathcal{G}^n$*-adapted and* W^n *is a* $\mathcal{G}^n$*-martingale*

(iii) If additionally, $\sup_n \mathbb{E}[W_k^n(t)^2] < \infty$ *for each* $k \in [m]$ *and* $t > 0$*, and marginally (with respect to its own natural filtration)* W *is a standard Brownian motion, then* W *remains a standard Brownian motion with respect to the augmented natural filtration of* $(L, H, W, \chi^\epsilon(H))$*. Moreover, with* $\xi^n = (\xi_k^n, k \in [m])$ *and* $\xi = (\xi_k, k \in [m])$ *defined via*

$$\xi_k^n(t) = \int_0^t \chi_k^\epsilon(H^n)(v-)dW_k^n(v) \tag{59}$$

$$\xi_k(t) = \int_0^t \chi_k^\epsilon(H)(v-)dW_k(v), \tag{60}$$

we have

$$(L^n, H^n, W^n, \xi^n) \Rightarrow (L, H, W, \xi). \tag{61}$$

in $D_{l+3m}[0, \infty)$ *as* $n \to \infty$*.*

Proof of Theorem 1. We start with the discrete approximation (19)-(22) 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 in (29)-(30) are locally Lipschitz continuous (i.e., Lipschitz continuous on any compact subset of the domain $[0, \infty)^K \times \mathbb{R}^K$) and bounded, the SDE has a (global) unique strong solution. (For references regarding this result, see Remark 4.) 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

$$(U^\gamma, S^\gamma, B^\gamma, M^\gamma) \Rightarrow (U, S, B, M). \tag{62}$$

Because M^γ is a suitably normalized sum of bounded martingale differences, we have $M_k^\gamma(t) \xrightarrow{\mathbb{P}} 0$ for each $k \in [K]$ and any $t > 0$ as $\gamma \downarrow 0$, and thus, M is the $D_K[0, \infty)$ zero process.

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

$$A_k^\gamma(t) = p_k^\gamma(U^\gamma(t), S^\gamma(t)) \tag{63}$$

$$A_k(t) = p_k(U(t), S(t)). \tag{64}$$

Note that $p_k^\gamma(u,s) \to p_k(u,s)$ 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 $(u,s) \in [0,\infty)^K \times \mathbb{R}^K$. These properties also hold for $\sqrt{p_k^\gamma(u,s)}$ and $\sqrt{p_k(u,s)}$. So, by the Generalized CMT (Lemma 7) applied to the processes $A_k^\gamma(t)$, $A_k(t)$, $\sqrt{A_k^\gamma(t)}$ and $\sqrt{A_k(t)}$, we have from (62),

$$(U^\gamma, S^\gamma, B^\gamma, M^\gamma, A^\gamma, \sqrt{A^\gamma}) \Rightarrow (U, S, B, M, A, \sqrt{A}), \tag{65}$$

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

$$\widetilde{U}_k^\gamma(t) = \int_0^t A_k^\gamma(v)dv \tag{66}$$

$$\widetilde{U}_k(t) = \int_0^t A_k(v)dv. \tag{67}$$

Recall that

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

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

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

we have for any $T > 0$,

$$\sup_{0 \leq t \leq T} \left| U_k^\gamma(t) - \widetilde{U}_k^\gamma(t) \right| \xrightarrow{\mathbb{P}} 0. \tag{68}$$

Thus, by the continuity of integration with respect to the Skorohod metric (Theorem 11.5.1 of Whitt (2002)) and the CMT, we have from (65),

$$(U^\gamma, S^\gamma, B^\gamma, \widetilde{U}^\gamma, \sqrt{A^\gamma}) \Rightarrow (U, S, B, \widetilde{U}, \sqrt{A}). \tag{69}$$

Let $\epsilon > 0$. Let χ^ϵ be the random step function mapping defined in (56)-(57). Using (69) and Lemma 1(i), applying χ^ϵ to $\sqrt{A^\gamma}$ and $\sqrt{A}$, we have

$$(U^\gamma, S^\gamma, B^\gamma, \widetilde{U}^\gamma, \chi^\epsilon(\sqrt{A^\gamma})) \Rightarrow (U, S, B, \widetilde{U}, \chi^\epsilon(\sqrt{A})). \tag{70}$$

Let $\widetilde{\mathcal{H}}^\gamma = (\widetilde{\mathcal{H}}_t^\gamma,\, t \geq 0)$ denote the continuous, piecewise constant (and right-continuous) interpolation of the discrete-time filtration $\mathcal{H}^\gamma = (\mathcal{H}_j^\gamma,\, j \geq 0)$ defined in (4). For fixed $\gamma > 0$, the processes on the left side of (69) are adapted to $\widetilde{\mathcal{H}}^\gamma$, and B^γ is a square-integrable $\widetilde{\mathcal{H}}^\gamma$-martingale. By Lemma 1(ii), the processes on the left side of (70) are adapted to the expanded version of the filtration $\widetilde{\mathcal{H}}^\gamma$ that includes the exogenous randomization used to construct χ^ϵ, and B^γ continues to be a square-integrable martingale with respect to the expanded filtration. We can then define stochastic integration for the pre-limit processes $\chi^\epsilon(\sqrt{A^\gamma})$ and B^γ in (70) via

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

with $\widehat{S}^\gamma = (\widehat{S}_k^\gamma,\, k \in [K])$. By Lemma 3 (stated and proved after the current proof), B is marginally a standard K-dimensional Brownian motion. So, by Lemma 1(iii), B is standard K-dimensional Brownian motion with respect to the augmented natural filtration of the limit processes $(U, S, B, \widetilde{U}, \chi^\epsilon(\sqrt{A}))$ in (70). So, we can also define stochastic integration for the limit processes $\chi^\epsilon(\sqrt{A})$ and B in (70) via

$$\widehat{S}_k(t) = \int_0^t \chi_k^\epsilon(\sqrt{A})(v-)\, dB_k(v), \tag{72}$$

with $\widehat{S} = (\widehat{S}_k,\, k \in [K])$. Moreover, using Lemma 1(iii) and (70), we have weak convergence of the stochastic integrals in (71) to those in (72), jointly with the weak convergence of other components in (70):

$$(U^\gamma, S^\gamma, B^\gamma, \widetilde{U}^\gamma, \widehat{S}^\gamma) \Rightarrow (U, S, B, \widetilde{U}, \widehat{S}). \tag{73}$$

Recall from (20), (22) and (63), that for each $k \in [K]$,

$$S_k^\gamma(t) = \int_0^t \sqrt{A_k^\gamma(v-)}\, dB_k^\gamma(v). \tag{74}$$

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

$$\widetilde{S}_k(t) = \int_0^t \sqrt{A_k(v-)}\, dB_k(v). \tag{75}$$

For each $k \in [K]$, we now show that $\widehat{S}_k^\gamma$ in (71) approximates S_k^γ in (74), and $\widehat{S}_k$ in (72) approximates $\widetilde{S}_k$ in (75). Because χ^ϵ is an ϵ-uniform approximation (as in (58)), for each $k \in [K]$ and any $T > 0$,

$$\begin{aligned}
\mathbb{E}\left[\sup_{0\le t\le T}\left|S_k^\gamma(t) - \widehat{S}_k^\gamma(t)\right|\right] &\le \epsilon\sqrt{\gamma}\frac{1}{\sigma_k}\mathbb{E}\left[\sum_{i=0}^{\lfloor T/\gamma\rfloor-1}\mathbb{E}\left[\frac{I_k^\gamma(i+1)(X_k^\gamma(i+1)-\mu_k^\gamma)^2}{p_k^\gamma(U^\gamma(i\gamma), S^\gamma(i\gamma))}\,\middle|\,\mathcal{H}_i^\gamma\right]\right]^{1/2} \\
&\le \epsilon\frac{\sigma_k^\gamma}{\sigma_k}\sqrt{T}.
\end{aligned} \tag{76}$$

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

$$\mathbb{E}\left[\sup_{0\le t\le T}\left|\widehat{S}_k(t) - \widetilde{S}_k(t)\right|\right] \le \epsilon\mathbb{E}\left[\langle B_k, B_k\rangle_T\right]^{1/2} = \epsilon\sqrt{T}, \tag{77}$$

where $t \mapsto \langle B_k, B_k \rangle_t$ denotes the quadratic variation process for B_k. Putting together (68), (73), and (76)-(77), and sending $\epsilon \downarrow 0$, we obtain:

$$U = \widetilde{U} \tag{78}$$

$$S = \widetilde{S}. \tag{79}$$

Recalling the definition of A_k in (64), $\widetilde{U}$ in (67), and $\widetilde{S}$ in (75), we see from (78)-(79) that the limit processes (U, S, B) satisfy the SDE:

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

□

Proof of Lemma 1. Parts (i) and (ii) follow from Lemma 6.1 of Kurtz and Protter (1991). (See also the proof of Theorem 2.2 of Kurtz and Protter (1991).) We now establish part (iii).

The part (iii) assumption that $\sup_n \mathbb{E}[W_k^n(t)^2] < \infty$ implies that $W_k^n(t)$ is a uniformly integrable sequence for each $k \in [m]$ and $t > 0$. Using parts (i) and (ii), together with this uniform integrability, Theorem 5.3 of Whitt (2007) ensures that W remains a martingale with respect to the augmented natural filtration of $(L, H, W, \chi^\epsilon(H))$. (See also Problem 7 from Chapter 7 of Ethier and Kurtz (1986).) Since W is marginally a standard w-dimensional Brownian motion, it has quadratic variation $\langle W_k, W_k \rangle_t = t$ for all k and cross-variation $\langle W_k, W_l \rangle_t = 0$ for all $k \neq l$ almost surely. Then, using Levy's characterization of Brownian motion, W remains a standard w-dimensional Brownian motion with respect to the augmented natural filtration of $(L, H, W, \chi^\epsilon(H))$.

Lastly, we establish (61). The stochastic integrals in (59) and (60) are well-defined with predictable step function integrands and square-integrable martingale integrators. To obtain weak convergence of the stochastic integrals in (59) to those in (60), and moreover the joint weak convergence in (61), consider the following observations. Let $x^n \in D_m[0, \infty)$ and $y^n \in D_m[0, \infty)$ be sequences such that $(x^n, y^n) \to (x, y)$ in $D_{2m}[0, \infty)$ as $n \to \infty$. Suppose also that for each component $k \in [m]$, x_k^n is a step function and the number of discontinuities of x_k^n in any bounded time interval is uniformly bounded in n. Then, defining $z^n \in D_m[0, \infty)$ and $z \in D_m[0, \infty)$ with components

$$z_k^n(t) = \int_0^t x_k^n(v-)dy_k^n(v)$$
$$z_k(t) = \int_0^t x_k(v-)dy_k(v),$$

we have

$$(x^n, y^n, z^n) \to (x, y, z)$$

in $D_{3m}[0,\infty)$ as $n\to\infty$. As noted in Kurtz and Protter (1991), by the continuity of stochastic integration with step function integrands as considered above, we obtain the weak convergence of the stochastic integrals in (59) to those in (60), and moreover the desired joint weak convergence in (61). $\square$

LEMMA 2. *The processes $(U^\gamma, S^\gamma, B^\gamma, M^\gamma)$ defined in (19)-(22) are tight in $D_{4K}[0,\infty)$.*

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

$$U_k^\gamma(t)=\gamma\sum_{i=1}^{\lfloor t/\gamma\rfloor} I_k^\gamma(i) \tag{80}$$

$$S_k^\gamma(t)=\sqrt{\gamma}\frac{1}{\sigma_k}\sum_{i=1}^{\lfloor t/\gamma\rfloor} I_k^\gamma(i)(X_k^\gamma(i)-\mu_k^\gamma) \tag{81}$$

$$M_k^\gamma(t)=\gamma\sum_{i=0}^{\lfloor t/\gamma\rfloor-1}(I_k^\gamma(i+1)-p_k^\gamma(U^\gamma(i\gamma),S^\gamma(i\gamma))) \tag{82}$$

$$B_k^\gamma(t)=\sqrt{\gamma}\frac{1}{\sigma_k}\sum_{i=0}^{\lfloor t/\gamma\rfloor-1}\frac{I_k^\gamma(i+1)(X_k^\gamma(i+1)-\mu_k^\gamma)}{\sqrt{p_k^\gamma(U^\gamma(i\gamma),S^\gamma(i\gamma))}} \tag{83}$$

Note that (80)-(81) are just different expressions of the same quantities in (19)-(20). Also, the process in (80) is uniformly bounded and increasing, and those in (81)-(83) are square-integrable martingales.

By Lemma 8, 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 9 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. (Here, the corresponding sequences are ξ^γ and $A_\delta^\gamma(T)$, which are indexed by $\gamma>0$ with $\gamma\downarrow 0$.) Let $C=\max_{k\in[K]}\sup_{\gamma>0}(\sigma_k^\gamma/\sigma_k)^2$, and without loss of generality suppose that $C<\infty$. For any $T>0$, $\delta>0$ and $\gamma>0$, (T2) holds for ξ^γ equal to each individual component process of $(U^\gamma, S^\gamma, B^\gamma, M^\gamma)$ by setting $A_\delta^\gamma(T)=(C+1)\max(\delta+\gamma,(\delta+\gamma)^2)$, and (T2) holds for ξ^γ equal to each pairwise sum of component processes of $(U^\gamma, S^\gamma, B^\gamma, M^\gamma)$ by setting $A_\delta^\gamma(T)=4(C+1)\max(\delta+\gamma,(\delta+\gamma)^2)$ (using the bound: $(x+y)^2\le 2x^2+2y^2$). Then, for (T3), we have $\lim_{\delta\downarrow 0}\limsup_{\gamma\downarrow 0}A_\delta^\gamma(T)=0$ in all cases. $\square$

LEMMA 3. *Marginally, $B^\gamma\Rightarrow B$ as $\gamma\downarrow 0$, where B is standard K-dimensional Brownian motion.*

Proof of Lemma 3. We apply the martingale functional central limit theorem stated in Lemma 10. Below, we verify (M1) and (M2) to ensure Lemma 10 holds.

Verification of (M1)

Because $I_j^\gamma(i)I_k^\gamma(i)=0$ for $j\neq k$ and all $i\geq 1$ (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 $k\in[K]$, as the following argument shows. As shorthand, denote $p_k^\gamma(i):=p_k^\gamma(U^\gamma(i\gamma),S^\gamma(i\gamma))$. Then,

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

$$=\gamma\left(\frac{\sigma_k^\gamma}{\sigma_k}\right)^2\sum_{i=0}^{\lfloor t/\gamma\rfloor-1}\frac{\mathbb{E}\left[I_k^\gamma(i+1)\,\middle|\,\mathcal{H}_i^\gamma\right]}{p_k^\gamma(i)}\mathbb{E}\left[\left(\frac{X_k^\gamma(i+1)-\mu_k^\gamma}{\sigma_k^\gamma}\right)^2\,\middle|\,\mathcal{H}_i^\gamma\right] \tag{84}$$

$$=\gamma\left(\frac{\sigma_k^\gamma}{\sigma_k}\right)^2\lfloor t/\gamma\rfloor\to t \tag{85}$$

as $\gamma\downarrow 0$. Here, (84) 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 $\mathcal{H}_i^\gamma$. The convergence in (85) follows from (6) in Assumption 1.

Verification of (M2)

For each $k\in[K]$, denote

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

By Markov's inequality, it suffices to show that

$$\sup_{0\leq i\leq\lfloor t/\gamma\rfloor-1}\mathbb{E}\left[\xi_k^\gamma(i+1)^2\mathbb{I}\left(|\xi_k^\gamma(i+1)|>\epsilon/\sqrt{\gamma}\right)\right]\to 0 \tag{86}$$

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

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

$$\leq\mathbb{E}\left[\xi_k^\gamma(i+1)^2\mathbb{I}\left(p_k^\gamma(i)<\delta\right)\right]$$

$$=\mathbb{E}\left[\mathbb{I}\left(p_k^\gamma(i)<\delta\right)\frac{\mathbb{E}[I_k^\gamma(i+1)\mid\mathcal{H}_i^\gamma]}{p_k^\gamma(i)}\mathbb{E}\left[\left(\frac{X_k^\gamma(i+1)-\mu_k^\gamma}{\sigma_k^\gamma}\right)^2\,\middle|\,\mathcal{H}_i^\gamma\right]\right] \tag{87}$$

$$
\begin{aligned}
&= \mathbb{P}\left(p_k^\gamma(i) < \delta\right) && (88)\\
&\le \mathbb{P}\left(\inf_{v\in[0,t]} p_k^\gamma(U^\gamma(v), S^\gamma(v)) < \delta\right)\\
&\le \eta, && (89)
\end{aligned}
$$

where (87) follows from $U^\gamma(i\gamma)$ and $S^\gamma(i\gamma)$ being $\mathcal{H}_i^\gamma$-measurable, (88) 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 (89) holds for γ sufficiently close to zero, as established above.

Additionally, we have

$$
\begin{aligned}
&\mathbb{E}\left[\xi_k^\gamma(i+1)^2 \mathbb{I}\left(|\xi_k^\gamma(i+1)| > \epsilon/\sqrt{\gamma}\right)\mathbb{I}\left(p_k^\gamma(i)\ge\delta\right)\right]\\
&= \mathbb{E}\left[\frac{\mathbb{I}\left(p_k^\gamma(i)\ge\delta\right)}{p_k^\gamma(i)}\mathbb{E}\left[I_k^\gamma(i+1)\left(\frac{X_k^\gamma(i+1)-\mu_k^\gamma}{\sigma_k^\gamma}\right)^2 \mathbb{I}\left(|\xi_k^\gamma(i+1)| > \epsilon/\sqrt{\gamma}\right)\,\middle|\,\mathcal{H}_i^\gamma\right]\right]\\
&\le \frac{1}{\delta}\mathbb{E}\left[\mathbb{E}\left[\left(\frac{X_k^\gamma(i+1)-\mu_k^\gamma}{\sigma_k^\gamma}\right)^2 \mathbb{I}\left(|\xi_k^\gamma(i+1)| > \epsilon/\sqrt{\gamma}\right)\,\middle|\,\mathcal{H}_i^\gamma\right]\right]\\
&\le \frac{1}{\delta}\mathbb{E}\left[\mathbb{E}\left[\left|\frac{X_k^\gamma(i+1)-\mu_k^\gamma}{\sigma_k^\gamma}\right|^{2+\alpha}\right]^{2/(2+\alpha)} \mathbb{P}\left(|\xi_k^\gamma(i+1)| > \epsilon/\sqrt{\gamma}\,\middle|\,\mathcal{H}_i^\gamma\right)^{\alpha/(2+\alpha)}\right] && (90)\\
&\le \frac{C}{\delta}\mathbb{E}\left[\mathbb{P}\left(|\xi_k^\gamma(i+1)| > \epsilon/\sqrt{\gamma}\,\middle|\,\mathcal{H}_i^\gamma\right)^{\alpha/(2+\alpha)}\right], && (91)
\end{aligned}
$$

where (90) follows from Hölder's inequality, and (91) follows from (7) in Assumption 1, with constant $C>0$. Furthermore, a.s.,

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

So, by the bounded convergence theorem, the right side of (91) converges to zero as $\gamma\downarrow 0$.

Therefore, from (89) and (91), we have

$$
\limsup_{\gamma\downarrow 0}\ \sup_{0\le i\le\lfloor t/\gamma\rfloor-1}\ \mathbb{E}\left[\xi_k^\gamma(i+1)^2\mathbb{I}\left(|\xi_k^\gamma(i+1)| > \epsilon/\sqrt{\gamma}\right)\right] \le \eta,
$$

and sending $\eta\downarrow 0$ yields (86). $\square$

5.2. Proofs for Stochastic ODE Approximation

In this section, we prove Theorem 2, Theorem 3 and Proposition 1 from Section 3.2. Theorem 2 shows that in the small gap regime of Assumption 2 (with stationary rewards for each arm) and the reward stack model of reward feedback, the pre-limit processes converge weakly to the unique (in distribution) non-anticipative weak solution to the stochastic ODE, with the uniqueness ensured by Theorem 3. In Theorem 3, we show that a non-anticipative solution to the stochastic ODE is also a solution to the corresponding SDE. So, if the SDE has a unique strong solution, then the stochastic ODE must have a unique (in distribution) non-anticipative weak solution. In Proposition 1, we show the converse result that a solution to the SDE is also a non-anticipative solution to the corresponding stochastic ODE.

Proof of Theorem 2.
We start with the joint process $(U^\gamma, Z^\gamma, M^\gamma)$ with components defined in (8), (11) and (38). From Assumption 2 and Remark 2, the individual components of Z^γ are tight in $D[0,\infty)$, and their weak limit points have continuous sample paths. Also, the individual components of U^γ and M^γ are tight in $D[0,\infty)$ (which can be argued as in Lemma 2), and their weak limit points have continuous sample paths. Thus, the joint process $(U^\gamma, Z^\gamma, M^\gamma)$ is tight in $D_{3K}[0,\infty)$. By Prohorov's Theorem, any subsequence of $(U^\gamma, Z^\gamma, M^\gamma)$ has a further subsequence that converges weakly. Consider any weakly convergent subsequence of $(U^\gamma, Z^\gamma, M^\gamma)$, which for notational simplicity we still index using γ, and let (U, Z, M) denote its weak limit. As in the proof of Theorem 1, M is the $D_K[0,\infty)$ zero process.

By the continuity of function composition (Theorem 13.2.2 of Whitt (2002)), since the individual components of Z have continuous sample paths and those of U have non-decreasing sample paths, we have by the CMT,

$$(U^\gamma, Z^\gamma, M^\gamma, Z^\gamma \circ U^\gamma) \Rightarrow (U, Z, M, Z \circ U), \tag{92}$$

in $D_{4K}[0,\infty)$, with $Z^\gamma \circ U^\gamma$ as defined in (13) and $Z \circ U = (Z_k(U_k),\, k \in [K]) \in D_K[0,\infty)$. Define the processes $A^\gamma = (A_k^\gamma,\, k \in [K])$ and $A = (A_k,\, k \in [K])$, where

$$A_k^\gamma(t) = p_k^\gamma(U^\gamma(t), Z^\gamma \circ U^\gamma(t)) \tag{93}$$

$$A_k(t) = p_k(U(t), Z \circ U(t)). \tag{94}$$

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 CMT (Lemma 7) applied to the processes in (93)-(94), we have from (92),

$$(U^\gamma, Z^\gamma, M^\gamma, A^\gamma) \Rightarrow (U, Z, M, A). \tag{95}$$

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

$$\begin{aligned} \widetilde{U}_k^\gamma(t) &= \int_0^t A_k^\gamma(v)dv \\ \widetilde{U}_k(t) &= \int_0^t A_k(v)dv. \end{aligned} \tag{96}$$

Recall that

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

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

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

we have for any $T > 0$,

$$\sup_{0\leq t\leq T} \left| U_k^\gamma(t) - \widetilde{U}_k^\gamma(t) \right| \xrightarrow{\mathbb{P}} 0. \tag{97}$$

By the continuity of integration with respect to the Skorohod metric (Theorem 11.5.1 of Whitt (2002)) and the CMT, we have from (95),

$$(U^\gamma, Z^\gamma, \widetilde{U}^\gamma) \Rightarrow (U, Z, \widetilde{U}). \tag{98}$$

Together, (97)-(98) yield

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

and recalling the definition of $\widetilde{U}$ in (96) and (94), we obtain:

$$(U^\gamma, Z^\gamma) \Rightarrow (U, Z),$$

where U and Z satisfy

$$U_k(t) = \int_0^t p_k(U(v), Z \circ U(v)) dv, \quad k \in [K]. \tag{99}$$

Moreover, by the continuity of function composition, we have

$$(U^\gamma, Z^\gamma \circ U^\gamma, Z^\gamma) \Rightarrow (U, Z \circ U, Z). \tag{100}$$

Next, we show that Z is a standard K-dimensional Brownian motion with respect to the augmented natural filtration of the limit process $(U, Z\circ U, Z)$ in (100), and that U is a non-anticipative solution (as in Definition 2) to the stochastic ODE in (99). Define the filtration $\mathcal{G}^\gamma = (\mathcal{G}_u^\gamma, u \in [0,\infty)^K)$ via

$$\mathcal{G}_u^\gamma = \mathcal{F}_u^\gamma \vee \sigma(\eta_j, j = 1, \ldots, \lfloor \Sigma_k u_k/\gamma \rfloor), \tag{101}$$

where the filtration $\mathcal{F}^\gamma = (\mathcal{F}_u^\gamma, u \in [0,\infty)^K)$ is as defined in (10), and η_j is the exogenous iid random variable used at time j by TS to randomly draw an arm (in the original discrete-time system). Then, we can see that the $U^\gamma(t)$, $t \geq 0$ are $\mathcal{G}_u^\gamma$-stopping times. Moreover, because the η_j TS randomization variables are independent of $\mathcal{F}^\gamma$, we have that (9) (from Assumption 2) holds with $\mathcal{G}_u^\gamma$ in the place of $\mathcal{F}_u^\gamma$. This, together with (97) and the weak convergence in (100), ensures that the conditions of Theorem 3.4(a) from Chapter 6 of Ethier and Kurtz (1986) are satisfied. Assumption 2 ensures that marginally (with respect to its augmented natural filtration) Z is a standard K-dimensional Brownian motion. Together with Theorem 3.4(a), we have that Z remains a standard K-dimensional Brownian motion with respect to the augmented natural filtration of the limit process $(U, Z \circ U, Z)$ in (100), and that U is a non-anticipative solution (as in Definition 2) to the stochastic ODE in (99).

Finally, we can apply Theorem 3 to conclude that all such weakly converging subsequences (U^γ, Z^γ) have the same weak limit. Specifically, by Theorem 3, with (U, Z) denoting the weak limit (as in (100), which satisfies the stochastic ODE in (99)), we have that $(U, S) := (U, Z \circ U)$ is also a solution to the corresponding SDE, which has a unique strong solution. This establishes the desired weak convergence in Theorem 2 to the unique (in distribution) non-anticipative weak solution to the stochastic ODE. $\square$

Proof of Theorem 3.
Let B be a K-dimensional standard Brownian motion on a probability space $(\Omega, \mathbb{F}, \mathbb{P})$, and let $\mathcal{F} = (\mathcal{F}_u,\, u \in [0,\infty)^K)$ be the completed filtration defined in (40). Let $U = (U_k,\, k \in [K])$ be a non-anticipative solution, as in Definition 2, to the stochastic ODE:

$$U_k(t) = \int_0^t p_k(U(v), B \circ U(v))dv, \quad k \in [K], \tag{102}$$

with K-dimensional standard Brownian motion B. So, there exists a filtration $\mathcal{G} = (\mathcal{G}_u,\, u \in [0,\infty)^K)$ for which conditions (i)-(iii) of Definition 2 hold. With $\theta = (\theta_k,\, k \in [K])$, define for $u \in [0,\infty)^K$:

$$\phi_\theta(u) = \prod_{k \in [K]} \exp\left(i\theta_k B_k(u_k) + \frac{1}{2}\theta_k^2 u_k \right).$$

Then, by condition (ii) of Definition 2, $\phi_\theta(u)$ is a $\mathcal{G}_u$-martingale. Moreover, by condition (iii) of Definition 2, for each $t \geq 0$, $U(t)$ is a $\mathcal{G}_u$-stopping time. Thus, the conditions of Theorem 6.3(a) of Kurtz (1980a) are satisfied. Define $S = (S_k,\, k \in [K])$ by $S_k(t) := B_k(U_k(t))$. Using Theorem 6.3(a), for any $k, l \in [K]$ with $k \neq l$, both $S_k(t)$ and $S_k(t)S_l(t)$ are continuous local martingales with respect to the filtration $\widetilde{\mathcal{G}} = (\widetilde{\mathcal{G}}_t,\, t \geq 0)$ defined by $\widetilde{\mathcal{G}}_t := \mathcal{G}_{U(t)}$. Using the fact that $U_k(t) \leq t$ a.s., together with the Doob's Maximal Inequality (see Theorem 3.8(iv) from Chapter 1 of Karatzas and Shreve (1998)), we have for each $k \in [K]$,

$$\mathbb{E}\left[\sup_{v \in [0,t]} S_k^2(v) \right] \leq \mathbb{E}\left[\left(\sup_{v \in [0,t]} |B_k(v)| \right)^2 \right] \leq 4\mathbb{E}[B_k^2(t)] = 4t. \tag{103}$$

From (103), we can conclude that the $S_k(t)$ are continuous square-integrable martingales and the $S_k(t)S_l(t)$ for $k \neq l$ are continuous martingales with respect to the filtration $\widetilde{\mathcal{G}}_t$. Since the $S_k(t)S_l(t)$ for $k \neq l$ are continuous martingales, by Theorem 5.13 from Chapter 1 of Karatzas and Shreve (1998), their quadratic co-variations $\langle S_k, S_l\rangle_t$ must be zero. Also,

$$\langle S_k, S_k\rangle_t = U_k(t) \tag{104}$$

by Lemma 4 below. So, in summary, for $k, l \in [K]$,

$$\langle S_k, S_l\rangle_t = U_k(t)\mathbb{I}\,(k = l)\,. \tag{105}$$

Now, define $\widetilde{B} = (\widetilde{B}_k,\, k \in [K])$ by

$$\widetilde{B}_k(t) := \int_0^t \frac{1}{\sqrt{p_k(U(v), S(v))}} dS_k(v), \quad k \in [K]. \tag{106}$$

Then, for any $k, l \in [K]$, by Proposition 2.17 from Chapter 3 of Karatzas and Shreve (1998),

$$\langle \widetilde{B}_k, \widetilde{B}_l\rangle_t = \int_0^t \frac{1}{\sqrt{p_k(U(v), S(v))}} \frac{1}{\sqrt{p_l(U(v), S(v))}} d\langle S_k, S_l\rangle_v.$$

Using (105), together with the fact that $d\langle S_k, S_k\rangle_v = dU_k(v) = p_k(U(v), S(v))dv$, we have $\langle \widetilde{B}_k, \widetilde{B}_l\rangle_t = t\mathbb{I}\,(k = l)$. So, by the Lévy characterization of multi-dimensional Brownian motion (see Theorem 3.16 from Chapter 3 of Karatzas and Shreve (1998)), $\widetilde{B}$ is a standard K-dimensional Brownian motion with respect to the filtration $\widetilde{\mathcal{G}}_t$. Moreover, from (106) and Corollary 2.20 from Chapter 3 of Karatzas and Shreve (1998), we have for each $k \in [K]$,

$$\int_0^t \sqrt{p_k(U(v), S(v))}\, d\widetilde{B}_k(v) = \int_0^t \sqrt{p_k(U(v), S(v))} \frac{1}{\sqrt{p_k(U(v), S(v))}} dS_k(v) = S_k(t).$$

Thus, we have the representation:

$$U_k(t) = \int_0^t p_k(U(v), S(v))dv \tag{107}$$

$$S_k(t) = \int_0^t \sqrt{p_k(U(v), S(v))}\, d\widetilde{B}_k(v), \quad k \in [K]. \tag{108}$$

So, the non-anticipative solution U to the stochastic ODE in (102) with the Brownian motion B, together with $S = B \circ U$, solves the SDE in (107)-(108) with the Brownian motion $\widetilde{B}$. Furthermore, we have $(U, B \circ U) = (U, S)$ a.s. $\square$

LEMMA 4. *In the proof of Theorem 3, (104) holds, i.e.,* $\langle S_k, S_k\rangle_t = U_k(t)$.

Proof of Lemma 4.

By definition, $S_k(t) = B_k(U_k(t))$. With index $u \in [0,\infty)^K$, we have that $M(u) = (M_k(u_k),\, k \in [K])$ with $M_k(u_k) := B_k(u_k)^2 - u_k,\ k \in [K]$ are $\mathcal{G}_u$-martingales. Also, for $0 \le s \le t$, we have that $U(s)$ and $U(t)$ are bounded $\mathcal{G}_u$-stopping times with $0 \le U_k(s) \le U_k(t) \le t$ for all $k \in [K]$. Then, for each k, we have

$$\mathbb{E}[B_k(U_k(t))^2 - U_k(t) \mid \mathcal{G}_{U(s)}] = B_k(U_k(s))^2 - U_k(s),$$

using Theorem 8.7 from Chapter 2 of Ethier and Kurtz (1986) (which is an optional stopping theorem for filtrations and martingales indexed by directed sets like $[0,\infty)^K$). So, for each k, $B_k(U_k(t))^2 - U_k(t)$ is a $\widetilde{\mathcal{G}}_t$ martingale (recall $\widetilde{\mathcal{G}}_t := \mathcal{G}_{U(t)}$). The desired result then follows from the Doob-Meyer decomposition (see Definition 5.3 from Chapter 1 of Karatzas and Shreve (1998)).
□

Proof of Proposition 1.

Let $\widetilde{B}$ be a K-dimensional standard Brownian motion on a probability space $(\Omega, \mathbb{F}, \mathbb{P})$, and let $\mathcal{F}^{\widetilde{B}} = (\mathcal{F}_t^{\widetilde{B}},\, t \ge 0)$ be the augmented natural filtration corresponding to $\widetilde{B}$. Let (U, S) be a solution to the SDE (46)-(48) on this probability space with respect to the standard Brownian motion $\widetilde{B}$. Writing (47) in integral form, because the p_k functions are bounded, we have that

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

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

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

For $k \ne l$, the quadratic co-variation processes $\langle S_k, S_l \rangle_t = 0$ since $\widetilde{B}_k$ and $\widetilde{B}_l$ are independent. From (46), we see that $\langle S_k, S_k \rangle_t = U_k(t)$, $k \in [K]$, which are continuous and strictly increasing processes since the p_k functions are bounded and strictly positive. Define

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

Also, we define $U_k(\infty) := \lim_{t\to\infty} U_k(t)$, and if $U_k(\infty) < \infty$, then we define $S_k(\infty) := \lim_{t \uparrow U_k(\infty)} S_k(U_k^{-1}(t))$.

Then, by a theorem due to F.B. Knight (see Proposition 18.8 of Kallenberg (2002) or Theorem 1.10 from Chapter V of Revuz and Yor (1999)), for $k \in [K]$, we can obtain independent standard Brownian motions via a random time change: $B_k(t) := S_k(U_k^{-1}(t))$, appending on an additional independent standard Brownian motion if $U_k(\infty) < \infty$. Specifically, define

$$B_k(t) := \begin{cases} S_k(U_k^{-1}(t)), & t < U_k(\infty) \\ S_k(\infty) + W_k(t - U_k(\infty)), & t \ge U_k(\infty), \end{cases}$$

where the W_k are independent standard Brownian motions (which we can have by enlarging the original probability space). Then, $B = (B_k,\, k \in [K])$ is an independent standard K-dimensional Brownian motion with respect to its augmented natural filtration $\mathcal{F}^B = (\mathcal{F}^B_t,\, t \geq 0)$.

By construction, we have $B_k(U_k(t)) = S_k(t)$, and substituting this representation into the SDE in (46), we obtain the stochastic ODE representation:

$$U_k(t) = \int_0^t p_k(U(v), B \circ U(v))dv, \quad k \in [K]. \tag{109}$$

So, with respect to the filtration $\mathcal{F}^B$, the SDE solution has the representation $(U, S) = (U, B \circ U)$, which satisfies the stochastic ODE in (109), with independent K-dimensional standard Brownian motion B.

Lastly, we establish that U is non-anticipative according to Definition 2. Define the (complete and right-continuous) filtration $\mathcal{G} = (\mathcal{G}_u,\, u \in [0,\infty)^K)$, with

$$\mathcal{G}_u = \mathcal{F}^B_u \vee \bigcap_{k\in[K]} \mathcal{F}^{\widetilde{B}}_{U_k^{-1}(u_k)}.$$

For condition (i) of Definition 2, we clearly have $\mathcal{F}^B_u \subset \mathcal{G}_u$ for all $u \in [0,\infty)^K$. For condition (ii) of Definition 2, set $\xi^u(\cdot) = (B_k(u_k + \cdot),\, k \in [K])$ for $u \in [0,\infty)^K$. Then, $\mathbb{P}(\xi^u \in \cdot \mid \mathcal{G}_u) = \mathbb{P}(\xi^u \in \cdot \mid \mathcal{F}^B_u)$ can be verified as in Theorem 2.8(b) (see also Theorem 5.3(b)) from Chapter 6 of Ethier and Kurtz (1986). Finally, for any $t \geq 0$, condition (iii) of Definition 2 follows from:

$$\bigcap_{k\in[K]} \{U_k(t) \leq u_k\} = \left\{\min_{k\in[K]} U_k^{-1}(u_k) \geq t\right\} \in \bigcap_{k\in[K]} \mathcal{F}^{\widetilde{B}}_{U_k^{-1}(u_k)} \subset \mathcal{G}_u.$$

□

Appendix A: Proof of Theorem 5

Proof of Theorem 5. We establish Theorem 5 for the SDE setting, but the same arguments below work in the stochastic ODE setting. Under ϵ-warm-start (from Definition 1), we only need to establish the SDE approximation on $[\epsilon, \infty)$. We verify that the sampling probabilities for EF Thompson samplers have the desired form with $p_k(u,s)$ as in (52).

In the SDE setting, we work under Assumption 1 (with iid rewards) and the random table model of reward feedback. At time $j+1 > \lfloor \epsilon/\gamma \rfloor$, conditional on $\mathcal{H}^\gamma_j$ (as defined in (4)), for each arm $k \in [K]$, we sample once from the posterior distribution of μ^γ_k and denote the sample by $\widetilde{\mu}^\gamma_k(j+1)$. For each arm $k \in [K]$, let $\widehat{\mu}^\gamma_k(j+1)$ 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 quantity for the Gaussian posterior approximation in Proposition 3.) Recall that the U^γ_k and S^γ_k processes have the expressions from (11) and (12). Then, the probability of playing arm $k \in [K]$ is given by:

$$\mathbb{P}\big(k = \arg\max_{l\in[K]} \widetilde{\mu}^\gamma_l(j+1) \,\big|\, \mathcal{H}^\gamma_j\big)$$

$$= \mathbb{P}\left(k = \arg\max_{l\in[K]}\left\{\frac{S_l^\gamma(j\gamma)\sigma_l}{U_l^\gamma(j\gamma)} + d_l^\gamma + \frac{1}{\sqrt{\gamma}}\left(\widetilde{\mu}_l^\gamma(j+1) - \widehat{\mu}_l^\gamma(j+1)\right)\right\} \,\Big|\, U^\gamma(j\gamma), S^\gamma(j\gamma)\right)$$

$$= \mathbb{P}\left(k = \arg\max_{l\in[K]}\left\{\frac{S_l^\gamma(j\gamma)\sigma_l}{U_l^\gamma(j\gamma)} + d_l^\gamma + \frac{\sigma_l}{\sqrt{U_l^\gamma(j\gamma)}}\mathcal{N}_l\right\} \,\Big|\, U^\gamma(j\gamma), S^\gamma(j\gamma)\right) + o_{\mathbb{P}}(1) \tag{110}$$

$$= p_k^\gamma(U^\gamma(j\gamma), S^\gamma(j\gamma)) + o_{\mathbb{P}}(1), \tag{111}$$

where the asymptotics are with respect to $\gamma \downarrow 0$. The result in (110) follows from Proposition 3, with the probability taken over the independent standard Gaussian random variables $\mathcal{N}_l$. In (111), we have

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

Moreover, with $p_k(u,s)$ as in (52), we have that $p_k^\gamma(u,s) \to p_k(u,s)$ uniformly for (u,s) in compact subsets of $[0,\infty)^K \times \mathbb{R}^K$, with the restriction that $u_k \geq \epsilon q_k > 0$ for each arm $k \in [K]$, 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 that of (15)-(17) in Section 3.1. Continuing from (111), the proof of Theorem 1 can be applied to yield the desired SDE approximation of Theorem 5. □

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 (51)), with means μ known to belong to a bounded, open interval $\mathcal{I}$. Our goal here is to develop Proposition 3 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, a.s. as the sample size $n \to \infty$, and uniformly over the possible data-generating distributions P^μ, $\mu \in \mathcal{I}$. The reason we develop the result uniformly over the possible data-generating distributions is the following. For a time horizon of $O(1/\gamma)$, the small gap regime of Assumption 1 involves mean parameters μ_k^γ in a $\sqrt{\gamma}$-neighborhood of some μ_*. As $\gamma \downarrow 0$ (and the time horizon goes to infinity), the mean parameters μ_k^γ change. So, for a given large time horizon, the Gaussian approximation to the posterior in Proposition 3 should be valid simultaneously for all distributions P^μ corresponding to a range of mean parameters μ; a fixed (not depending on γ) neighborhood of μ_* suffices. Below, we first discuss the "uniform almost sure" mode of convergence, and then move on to the development of Proposition 3.

"Uniform Almost Sure" Convergence

To make sense of the "uniform almost sure" mode of convergence, we first recall an equivalent characterization of almost sure convergence in Remark 10 below, followed by a precise definition of the mode of convergence in Definition 6 below. For any particular distribution Q, we use $\mathbb{E}_Q[\cdot]$ and $\mathbb{P}_Q(\cdot)$ to denote expectation and probability taken with respect to Q.

Remark 10. 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 6. Let $\mathcal{Q}$ be a collection of probability distributions and Z_i be random variables defined on the probability spaces $(\Omega, \mathbb{F}, Q)_{Q\in\mathcal{Q}}$. We say that the sequence Z_i converges to zero, a.s. uniformly in $Q \in \mathcal{Q}$, if for any $\epsilon > 0$,

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

Next, we state Lemma 5, which is used in the proof of Proposition 3. This result, originally due to Chung (1951), is a strong law of large numbers (SLLN) that holds uniformly over a collection of underlying probability distributions.

Lemma 5. *Let $\mathcal{Q}$ be a collection of probability distributions, and for each $Q \in \mathcal{Q}$, let $Y, Y_i \overset{\text{iid}}{\sim} Q$. Suppose the $\mathcal{Q}$-uniform integrability condition,*

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

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

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

Development of Proposition 3

Before presenting Lemma 6 and then continuing on to Proposition 3, 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 (51) from 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. Recall from Section 4.1 that $\mathcal{I}$ is a bounded, open interval containing μ_*, satisfying $\underline{\mu} < \inf \mathcal{I} < \sup \mathcal{I} < \overline{\mu}$. (For the exponential family P^μ, recall that $(\underline{\mu}, \overline{\mu})$ denotes the open interval of all possible mean values achievable with finite values of the tilting parameter.) For each $\mu \in \mathcal{I}$, let $X^\mu, X_i^\mu \overset{\text{iid}}{\sim} P^\mu$.

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

$$\sup_{z:|\mu - z|\geq\delta} \mathbb{E}[l(z, X^\mu)] \leq \mathbb{E}[l(\mu, X^\mu)] - \epsilon(\delta). \tag{112}$$

(C2) There exist functions η and κ such that for all x in the support of the base distribution P (in (51)),

$$\eta(x) \geq \sup_{\mu\in\mathcal{I}} |l'(\mu, x)| \tag{113}$$

$$\kappa(x) \geq \sup_{\mu\in\mathcal{I}} |l'''(\mu, x)| \, . \tag{114}$$

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

$$\lim_{y\to\infty} \sup_{\mu\in\mathcal{I}} \mathbb{E}\left[f(X^\mu) \mathbb{I}\left(f(X^\mu) > y \right) \right] = 0. \tag{115}$$

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

Lemma 6. *Suppose (115) in C2 holds 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 (113). Then, $\{l(\mu, \cdot), \mu \in \mathcal{I}\}$ is a Glivenko-Cantelli class of functions uniformly in P^μ, $\mu \in \mathcal{I}$, i.e., for any $\epsilon > 0$,*

$$\lim_{m\to\infty} \sup_{\mu\in\mathcal{I}} \mathbb{P}\left(\sup_{n\geq m} \sup_{z\in\mathcal{I}} \left| \frac{1}{n}\sum_{i=1}^{n} l(z, X_i^\mu) - \mathbb{E}[l(z, X^\mu)] \right| > \epsilon \right) = 0.$$

We now state and prove Proposition 3. The proof is adapted from the proof sketch for Theorem 4.2 in Ghosh et al. (2006). As before, for each $\mu \in \mathcal{I}$, let $X^\mu, X_i^\mu \overset{\text{iid}}{\sim} P^\mu$, with mean μ and corresponding variance $(\sigma^\mu)^2$. (Below, we will write all relevant quantities with superscript μ to keep track of the distribution P^μ that we work with.) The sample mean of $X_1^\mu, \ldots, X_n^\mu$ is denoted by $\widehat{m}_n^\mu$. Given n such samples, we use $\widetilde{m}_n^\mu$ to denote a sample from the posterior distribution of the mean μ.

Proposition 3. *For the exponential family P^μ in (51) from Section 4.1, suppose the conditions C1 and C2 hold for a bounded, open interval $\mathcal{I} \subset \mathbb{R}$. Let $\mathcal{J}$ be a compact sub-interval of $\mathcal{I}$. Let ν_0 be a bounded prior density that has support contained in $\mathcal{I}$, is continuous on an open interval containing $\mathcal{J}$, and is strictly positive on $\mathcal{J}$. 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:*

$$\lim_{n\to\infty} \int_{\mathbb{R}} \left| \nu_n(y \mid X_1^\mu, \ldots, X_n^\mu) - \frac{1}{\sqrt{2\pi}\sigma^\mu} \exp\left(-\frac{1}{2(\sigma^\mu)^2} y^2 \right) \right| dy = 0, \tag{116}$$

a.s. uniformly in the underlying distribution P^μ for $\mu \in \mathcal{J}$, in the sense of Definition 6.

Proof of Proposition 3. The posterior density can be expressed as

$$\nu_n(y \mid X_1^\mu, \ldots, X_n^\mu) = (C_n^\mu)^{-1} \nu_0(\widehat{m}_n^\mu + y/\sqrt{n}) \exp\left(L_n^\mu(\widehat{m}_n^\mu + y/\sqrt{n}) - L_n^\mu(\widehat{m}_n^\mu) \right), \tag{117}$$

with normalization factor

$$C_n^\mu = \int_{\mathbb{R}} \nu_0(\widehat{m}_n^\mu + y/\sqrt{n}) \exp\left(L_n^\mu(\widehat{m}_n^\mu + y/\sqrt{n}) - L_n^\mu(\widehat{m}_n^\mu) \right) dy,$$

and

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

Consider the following difference between unnormalized densities.

$$D_n^\mu(y) = \nu_0(\widehat{m}_n^\mu + y/\sqrt{n}) \exp\left(L_n^\mu(\widehat{m}_n^\mu + y/\sqrt{n}) - L_n^\mu(\widehat{m}_n^\mu) \right) - \nu_0(\mu) \exp\left(-\frac{1}{2(\sigma^\mu)^2} y^2 \right) \tag{118}$$

To establish (116), it suffices to show that

$$\lim_{n\to\infty} \int_{\mathbb{R}} |D_n^\mu(y)|\, dy = 0, \tag{119}$$

a.s. uniformly in $\mu \in \mathcal{J}$ (i.e., a.s. uniformly in the underlying distribution P^μ for $\mu \in \mathcal{J}$, in the sense of Definition 6). Indeed, if (119) holds, we must also have

$$\lim_{n\to\infty} C_n^\mu = \nu_0(\mu)\sqrt{2\pi}\sigma^\mu, \tag{120}$$

a.s. uniformly in $\mu \in \mathcal{J}$. Then, we would have

$$\int_{\mathbb{R}} \left| \nu_n(y \mid X_1^\mu, \ldots, X_n^\mu) - \frac{1}{\sqrt{2\pi}\sigma^\mu} \exp\left(-\frac{1}{2(\sigma^\mu)^2} y^2\right) \right| dy$$
$$\leq (C_n^\mu)^{-1} \int_{\mathbb{R}} |D_n^\mu(y)| \, dy + \left| (C_n^\mu)^{-1} \nu_0(\mu) - \frac{1}{\sqrt{2\pi}\sigma^\mu} \right| \int_{\mathbb{R}} \exp\left(-\frac{1}{2(\sigma^\mu)^2} y^2\right) dy. \tag{121}$$

Applying (119) and (120) to (121) would then lead to the desired conclusion in (116). So, it suffices to show that (119) holds. To do this, we split the integral over $\mathbb{R}$ 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, we have

$$\int_{A_n} |D_n^\mu(y)| \, dy \leq \int_{A_n} \nu_0(\widehat{m}_n^\mu + y/\sqrt{n}) \exp\left(L_n^\mu(\widehat{m}_n^\mu + y/\sqrt{n}) - L_n^\mu(\widehat{m}_n^\mu)\right) dy$$
$$+ \int_{A_n} \nu_0(\mu) \exp\left(-\frac{1}{2(\sigma^\mu)^2} y^2\right) dy. \tag{122}$$

By the boundedness of $\nu_0(\mu)$ and $(\sigma^\mu)^2$ for $\mu \in \mathcal{J}$, the second integral on the right side of (122) goes to zero as $n \to \infty$, uniformly in $\mu \in \mathcal{J}$. For the first integral on the right side of (122), using (115) in condition C2 for the case $f(x) = |x|$ together with Lemma 5, it follows that

$$\lim_{n \to \infty} \widehat{m}_n^\mu = \mu, \tag{123}$$

a.s. uniformly in $\mu \in \mathcal{J}$. This, together with condition C1 and Lemma 6, implies that there exists $\epsilon(\beta) > 0$ such that

$$\sup_{y \in A_n : (\widehat{m}_n^\mu + y/\sqrt{n}) \in \mathcal{I}} \frac{1}{n} \left(L_n^\mu(\widehat{m}_n^\mu + y/\sqrt{n}) - L_n^\mu(\widehat{m}_n^\mu)\right) \leq -\epsilon(\beta), \tag{124}$$

for sufficiently large n, a.s. uniformly in $\mu \in \mathcal{J}$. (For the first integral on the right side of (122), we only need to consider y for which $(\widehat{m}_n^\mu + y/\sqrt{n}) \in \mathcal{I}$ because the prior density ν_0 has support contained in $\mathcal{I}$.) So, using (122), (124) and the boundedness of the density ν_0, we have

$$\lim_{n \to \infty} \int_{A_n} |D_n^\mu(y)| \, dy = 0, \tag{125}$$

a.s. uniformly in $\mu \in \mathcal{J}$. (Note that this result does not depend on how the constant $\beta > 0$ will be chosen below.)

For the second case on A_n^c, we analyze $\int_{A_n^c} |D_n^\mu(y)| \, dy$. 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

$$L_n^\mu(\widehat{m}_n^\mu + y/\sqrt{n}) - L_n^\mu(\widehat{m}_n^\mu) = \frac{1}{2}\frac{1}{n}(L_n^\mu)''(\widehat{m}_n^\mu) y^2 + r_n^\mu(y)$$
$$= \frac{1}{2}\left(\theta''(\widehat{m}_n^\mu)\widehat{m}_n^\mu - \Lambda''(\widehat{m}_n^\mu)\right) y^2 + r_n^\mu(y), \tag{126}$$

using the fact that $l''(z, x) = \theta''(z) \cdot x - \Lambda''(z)$ (recall the definitions of $\theta(z)$ and $\Lambda(z)$ in (51)), 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 $\widehat{m}_n^\mu + y/\sqrt{n}$. Here, we require that $\beta > 0$ satisfy:

$$[\min \mathcal{J} - 2\beta, \max \mathcal{J} + 2\beta] \subset \mathcal{I}, \tag{127}$$

so that $m^\mu_{n,y} \in \mathcal{I}$ for $y \in A^c_n$, for n sufficiently large, a.s. uniformly in $\mu \in \mathcal{J}$. Then, using condition C2 with $f(x) = \kappa(x)$ and Lemma 5, there exists $\delta > 0$ such that for sufficiently large n, a.s. uniformly in $\mu \in \mathcal{J}$,

$$|r^\mu_n(y)| \leq \frac{1}{6}\frac{|y|^3}{\sqrt{n}}\frac{1}{n}\sum_{i=1}^n \kappa(X^\mu_i) \leq \frac{1}{6}\frac{|y|^3}{\sqrt{n}}\left(\mathbb{E}\left[\kappa(X^\mu)\right] + \delta\right). \tag{128}$$

For $y \in A^c_n$, (128) can be re-expressed as

$$|r^\mu_n(y)| \leq \frac{1}{6}\beta y^2 \left(\mathbb{E}\left[\kappa(X^\mu)\right] + \delta\right). \tag{129}$$

For the first term on the right side of (126), we have

$$\lim_{n\to\infty} \theta''(\widehat{m}^\mu_n)\widehat{m}^\mu_n - \Lambda''(\widehat{m}^\mu_n) = \theta''(\mu)\mu - \Lambda''(\mu) = -\frac{1}{(\sigma^\mu)^2} \tag{130}$$

a.s. uniformly in $\mu \in \mathcal{J}$. The first equality in (130) follows from the uniform continuity of θ'' and Λ'' on $\mathcal{I}$, together with the convergence result in (123). The second equality in (130) is from a standard identity relating Fisher information and the variance $(\sigma^\mu)^2$ of the X^μ_i. Defining

$$c_0 = \inf_{\mu\in\mathcal{J}} \frac{1}{(\sigma^\mu)^2}, \tag{131}$$

and recognizing that $c_0 > 0$, we can choose $\beta > 0$ to satisfy both (127) and also:

$$-\frac{1}{3}c_0 + \frac{1}{6}\beta\left(\sup_{\mu\in\mathcal{J}} \mathbb{E}\left[\kappa(X^\mu)\right] + \delta\right) \leq -\frac{1}{4}c_0. \tag{132}$$

Then, using (129) and (130), we have from (126) that

$$\sup_{y\in A^c_n} \frac{\exp\left(L^\mu_n(\widehat{m}^\mu_n + y/\sqrt{n}) - L^\mu_n(\widehat{m}^\mu_n)\right)}{\exp(-c_0 y^2/4)} \leq 1 \tag{133}$$

for sufficiently large n, a.s. uniformly in $\mu \in \mathcal{J}$. (Note that using $-(1/3)c_0$ instead of $-(1/2)c_0$ on the left side of (132) offers some "slack" to ensure compatibility with the convergence in (130).) Thus, with $D^\mu_n(y)$ as defined in (118), we have using (126), (130), (131) and (133),

$$|D^\mu_n(y)|\,\mathbb{I}\left(y \in A^c_n\right) \leq \left(\sup_{z\in\mathcal{I}} \nu_0(z)\right)\Big(\exp(-c_0y^2/4) + \exp(-c_0y^2/2)\Big), \tag{134}$$

for sufficiently large n, a.s. uniformly in $\mu \in \mathcal{J}$. Let $g(y)$ denote the function on the right side of (134). Let $\epsilon > 0$. Then, there exists $L(\epsilon) > 0$ such that

$$\int_{|y|>L(\epsilon)} g(y)dy \leq \epsilon/2. \tag{135}$$

Moreover, using the continuity of ν_0 on an open interval containing $\mathcal{J}$, (123), (126), (128) and (130), we have

$$\lim_{n\to\infty} \sup_{|z|\leq L(\epsilon)} |D^\mu_n(z)| = 0, \tag{136}$$

a.s. uniformly in $\mu \in \mathcal{J}$. From (134) and (135), we have

$$\begin{aligned}\int_{y\in A^c_n} |D^\mu_n(y)|\,dy &\leq 2L(\epsilon) \sup_{|z|\leq L(\epsilon)} |D^\mu_n(z)| + \int_{|y|>L(\epsilon)} g(y)dy \\ &\leq 2L(\epsilon) \sup_{|z|\leq L(\epsilon)} |D^\mu_n(z)| + \epsilon/2.\end{aligned} \tag{137}$$

Then, using (136) and (137), we have

$$\begin{aligned}\limsup_{m\to\infty} \sup_{\mu\in\mathcal{J}} \mathbb{P}_{P^\mu}\left(\sup_{n\geq m} \int_{y\in A^c_n} |D^\mu_n(y)|\,dy > \epsilon\right) &\leq \limsup_{m\to\infty} \sup_{\mu\in\mathcal{J}} \mathbb{P}_{P^\mu}\left(\sup_{n\geq m} 2L(\epsilon) \sup_{|z|\leq L(\epsilon)} |D^\mu_n(z)| > \epsilon/2\right) \\ &= 0,\end{aligned}$$

thus establishing that

$$\lim_{n\to\infty} \int_{y\in A^c_n} |D^\mu_n(y)|\,dy = 0, \tag{138}$$

a.s. uniformly in $\mu \in \mathcal{J}$. Together, (125) and (138) yield (119). $\square$

Appendix C: Gaussian Approximations for the Bootstrap

In Proposition 4 below, we develop a Gaussian approximation for bootstrapping the sample mean. Recall that here, we allow for an arbitrary family (not necessarily from an exponential family) of reward distributions P^μ parameterized by mean $\mu \in \mathcal{I}$, with corresponding variances $(\sigma^\mu)^2$, where $\mathcal{I} \subset \mathbb{R}$ is a bounded, open interval. The only requirement on the P^μ is that the condition in (139) is satisfied. As before, for each $\mu \in \mathcal{I}$, let $X^\mu, X_i^\mu \overset{\text{iid}}{\sim} P^\mu$. We use $\widehat{m}_n^\mu$ and $(\widehat{\sigma}_n^\mu)^2$ to denote the sample mean and variance computed using n samples $X_1^\mu, \ldots, X_n^\mu$. Also, we use $\widehat{m}_n^{*\mu}$ to denote a bootstrap of the sample mean $\widehat{m}_n^\mu$ computed using n re-samples with replacement. Proposition 4 holds a.s. and uniformly over data-generating distributions P^μ, $\mu \in \mathcal{I}$. Recall that a precise description of this mode of convergence was given in Remark 10 and Definition 6 in Appendix B.

Proposition 4. *Suppose that*

$$\lim_{y\to\infty} \sup_{\mu\in\mathcal{I}} \mathbb{E}\left[(X^\mu)^2 \mathbb{I}\left((X^\mu)^2 > y\right)\right] = 0. \tag{139}$$

and

$$\inf_{\mu\in\mathcal{I}} \sigma^\mu > 0. \tag{140}$$

Then,

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left| \mathbb{P}\left(\sqrt{n}\,(\widehat{m}_n^{*\mu} - \widehat{m}_n^\mu) \le x \mid X_1^\mu, \ldots, X_n^\mu\right) - \Phi\left(\frac{x}{\sigma^\mu}\right) \right| = 0, \tag{141}$$

a.s. uniformly in $\mu \in \mathcal{I}$.

Proof of Proposition 4. Using (139) and Lemma 5, we have

$$\lim_{n\to\infty} \widehat{m}_n^\mu = \mu \tag{142}$$

$$\lim_{n\to\infty} \widehat{\sigma}_n^\mu = \sigma^\mu, \tag{143}$$

a.s. uniformly in $\mu \in \mathcal{I}$. The rest of the proof adapts the approach for establishing boostrap consistency described Example 3.1 from Shao and Tu (1995). By the Berry-Esseen Theorem (Appendix A.9 of Shao and Tu (1995)), there exists some universal constant $c > 0$ such that for all n, a.s. for all $\mu \in \mathcal{I}$,

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}\left(\sqrt{n}\,(\widehat{m}_n^{*\mu} - \widehat{m}_n^\mu) \le x \mid X_1^\mu, \ldots, X_n^\mu\right) - \Phi\left(\frac{x}{\widehat{\sigma}_n^\mu}\right) \right| \le c(\widehat{\sigma}_n^\mu)^{-3} n^{-3/2} \sum_{i=1}^n \left|X_i^\mu - \widehat{m}_n^\mu\right|^3. \tag{144}$$

For the right side of (144), we have

$$\left|X_i^\mu - \widehat{m}_n^\mu\right|^3 \le 4\left(\left|X_i^\mu - \mu\right|^3 + \left|\widehat{m}_n^\mu - \mu\right|^3\right),$$

and moreover,

$$\lim_{n\to\infty} \frac{1}{n^{3/2}} \sum_{i=1}^n \left|X_i^\mu - \mu\right|^3 = 0 \tag{145}$$

$$\lim_{n\to\infty} \frac{1}{n^{3/2}} \sum_{i=1}^n \left|\widehat{m}_n^\mu - \mu\right|^3 = \lim_{n\to\infty} \frac{1}{\sqrt{n}} \left|\widehat{m}_n^\mu - \mu\right|^3 = 0, \tag{146}$$

a.s. uniformly in $\mu\in\mathcal{I}$, where (145) follows from (139) and a uniform version of the Marcinkiewicz-Zygmund SLLN (Theorem 1(ii) of Waudby-Smith et al. (2024) with $q=2/3$), and (146) follows from (142). Then, using (140), (143), (145) and (146), the right side of (144) converges to zero as $n\to\infty$, a.s. uniformly in $\mu\in\mathcal{I}$. Using (140), (143), and the continuity properties of $\Phi(\cdot)$ together with Polya's Theorem (pointwise convergence to a continuous limit CDF implies uniform convergence), we have

$$\lim_{n\to\infty}\sup_{x\in\mathbb{R}}\left|\Phi\left(\frac{x}{\widehat{\sigma}_n^{\mu}}\right)-\Phi\left(\frac{x}{\sigma^{\mu}}\right)\right|=0, \tag{147}$$

a.s. uniformly in $\mu\in\mathcal{I}$. Using the convergence results for (144) and (147), the desired conclusion in (141) is established. □

Appendix D: Weak Convergence Technical Lemmas

Lemma 7 **(Generalized Continuous Mapping Theorem)**. *Let f and f^n, $n\geq 1$, be measurable functions that map from the metric space $(\mathcal{S}_1,r_1)$ 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 $f^n(\xi^n)\Rightarrow f(\xi)$ in $(\mathcal{S}_2,r_2)$. (See Theorem 3.4.4 of Whitt (2002).)*

Lemma 8 **(Tightness of Multi-dimensional Processes)**. *A sequence of process $\xi^n=(\xi_1^n,\dots,\xi_d^n)$ is tight in $D_d[0,\infty)$ 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 9 **(Simple Sufficient Conditions for Tightness)**. *A sequence of processes ξ^n in $D[0,\infty)$ adapted to filtration $(\mathcal{F}_t^n, t\geq 0)$ 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

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

a.s. for $0\leq t\leq T$ and $0\leq v\leq\delta$, 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 10 **(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 $i=1,2,\dots$. 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 $k=1,\dots,m$,

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

Then,

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

in $D_m[0,\infty)$, where W 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 (2023) *Data-driven Decisions in Stochastic Systems: Reliability and Efficiency*. Ph.D. thesis, Stanford University.

Fan L, Glynn P (2021) Diffusion Approximations for Thompson Sampling. URL https://arxiv.org/abs/2105.09232v1.

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 (2024) Weak Signal Asymptotics for Sequentially Randomized Experiments. *Management Science* 70(10):7024–7041.

Kurtz T (1980a) Representations of Markov Processes as Multiparameter Time Changes. *The Annals of Probability* 8(4):682–715.

Kurtz T (1980b) The Optional Sampling Theorem for Martingales Indexed by Directed Sets. *The Annals of Probability* 8(4):675–681.

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.

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).

Shao J, Tu D (1995) *The Jackknife and the Bootstrap* (Springer).

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* .

Wager S, Xu K (2021) Diffusion Asymptotics for Sequential Experiments. URL https://arxiv.org/abs/2101.09855v1.

Waudby-Smith I, Larsson M, Ramdas A (2024) Distribution-uniform strong laws of large numbers URL https://arxiv.org/abs/2402.00713.

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

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