

# 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  $\gamma$  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 is developed from first principles using the Continuous Mapping Theorem, and can 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 and 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 a Gaussian prior. Moreover, in this 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 – September 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  $\text{Reg}(n) := \sum_k N_k(n) \Delta_k$  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  $\mu_j$  is the mean reward of arm  $j$ . While expected regret  $\mathbb{E}[\text{Reg}(n)]$  is the most fundamental performance measure, the probabilistic behavior of  $\text{Reg}(n)$  can depend on other aspects of its distribution, which may be crucial to understand in some applications. For example, in settings where bandit algorithms are deployed with only a limited number of runs so that the law of large numbers does not “kick in”, or in settings where risk sensitivity is a key

concern, the spread or variance of  $\text{Reg}(n)$  can be as important for designing effective algorithms as  $\mathbb{E}[\text{Reg}(n)]$ .

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  $\Delta_k$  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 two key settings which guide the design of optimal bandit algorithms; see Chapters 15-16 of Lattimore and Szepesvári (2020). Indeed, for TS, which is known to be nearly minimax-optimal, the “statistically hardest” bandit environments have sub-optimality gaps  $\Delta_k$  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, 1) their weak convergence theory is based on analysis of infinitesimal generators and 2) they do not develop the general invariance principles for algorithms besides the Gaussian Thompson sampler (nor the accompanying insights) that we do. We provide a detailed comparison of our work to theirs in Section 1.1.

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 is transparent and shows explicitly how the SDE and stochastic ODE weak limits arise. 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 general prior distributions and any single-parameter exponential family likelihood (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 instance-dependent bandit setting of Lai and Robbins (1985), where algorithms can be highly sensitive to even small degrees of mis-specification, as recently shown in Fan and Glynn (2024).

The rest of the paper is structured as follows. Related work is further discussed in Section 1.1. We then 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  $O(1/\gamma)$ . 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, relative to the generator approach, 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.

Another 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 appear to require augmenting the state so that the adaptive system becomes Markovian, and then verifying generator convergence for the enlarged process—a development not undertaken in KW.

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 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

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

corresponding to the data collected through some time  $j$ , where at each time  $i$  and for each arm  $k \in [K] := \{1, \dots, K\}$ ,  $I_k(i) = 1$  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  $k \in [K]$ ,

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

is the number of plays and

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

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  $\Delta^K$  denotes the  $K$ -dimensional probability simplex and  $\pi : \mathbb{N}^K \times \mathbb{R}^K \rightarrow \Delta^K$ . Given  $I(j + 1)$ , a reward  $Y(j + 1)$  is received for the selected arm, and the sufficient statistics  $(N(j + 1), G(j + 1))$  are updated accordingly.

TS is an important example of a sampling-based bandit algorithm and our primary focus throughout the paper. When studying TS, we will restrict attention to TS designed for parametric reward models parameterized by mean. (As mentioned in the Introduction, we will begin with the Gaussian Thompson sampler in Sections 3.1 and 3.2 before generalizing to EF Thompson samplers in Section 4.1.) As a Bayesian algorithm, TS maintains a posterior distribution for the mean reward of each arm, and in each time period, it samples a mean from each posterior and plays the arm corresponding to the highest sampled mean, after which a corresponding reward is received and the posterior is updated with the new information. More precisely, for each arm  $k$ , we start with an independent prior distribution  $\nu_k^0$  for the unknown mean  $\mu_k$ . 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  $\tilde{\mu}_k(j) \sim \nu_k(N_k(j), G_k(j))$  for each arm  $k$ , and we play the arm  $\arg \max_k \tilde{\mu}_k(j)$ . So, for TS,  $\pi_k(N(j), G(j)) := \mathbb{P}(k = \arg \max_l \tilde{\mu}_l(j))$ , 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 a sequence of rewards. We refer to the first way as the *random table model*, where at time  $j$ ,  $Y(j) = X_k(j)$  for the selected arm  $k \in [K]$  ( $I_k(j) = 1$ ). We refer to the second way as the *reward stack model*, where at time  $j$ ,  $Y(j) = X_k(N_k(j - 1) + 1)$  for the selected arm  $k \in [K]$  ( $I_k(j) = 1$ ), where  $N_k(j - 1)$  is the number of plays of arm  $k$  through time  $j - 1$ , as defined in (2) above. (The

random table and reward stack terminology is taken from [Lattimore and Szepesvári \(2020\)](#); see Chapter 4.6, page 53.)

When the rewards for each arm are iid, the random table model and the reward stack model generate reward feedback in distributionally equivalent ways. We will also consider the setting where the rewards  $X_k(i)$ ,  $i = 1, 2, \dots$  for each arm  $k \in [K]$  are *stationary* (which allows for serial dependence), i.e., for any fixed integers  $1 \leq i_1 \leq i_2 \leq \dots \leq i_l < \infty$ , the finite-dimensional distributions  $(X_k(i_1 + j), X_k(i_2 + j), \dots, X_k(i_l + j))$  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  $\mathbb{R}^m$ , that are right-continuous and have limits from the left. For this space, we use the Skorohod metric. Weak convergence is always denoted using  $\Rightarrow$ , both for stochastic processes taking values in  $D^m[a, \infty)$  and for random variables taking values in  $\mathbb{R}^m$ . Complete mathematical details for the spaces  $D^m[a, \infty)$  equipped with the Skorohod metric, as well as the theory of weak convergence in such spaces, can be found in standard references such as [Billingsley \(1999\)](#), [Ethier and Kurtz \(1986\)](#) and [Whitt \(2002\)](#).

### 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  $\gamma$ , 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  $\gamma$ -scale system, we will write a  $\gamma$  superscript on all objects defined previously to indicate we are working with the same object defined appropriately in the  $\gamma$ -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

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

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  $\gamma$  and each arm  $k \in [K]$ , we have a reward distribution  $Q_k^\gamma$  with mean  $\mu_k^\gamma$ , variance  $(\sigma_k^\gamma)^2$ , and rewards  $X_k^\gamma(i) \stackrel{\text{iid}}{\sim} Q_k^\gamma$  for  $i = 1, 2, \dots$ . 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*

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

$$\lim_{\gamma \downarrow 0} \sigma_k^\gamma = \sigma_k \quad (6)$$

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

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  $\gamma \downarrow 0$ ,  $\Delta_k^\gamma \rightarrow \Delta_k := \max_l d_l - d_k$ . The essential idea behind the small gap regime is that the arm means  $\mu_k^\gamma$  are all clustered near some fixed  $\mu_* \in \mathbb{R}$ , with small differences/gaps between the means on the scale of  $\sqrt{\gamma}$ . In order to begin distinguishing between arms, one must play each arm on the scale of  $1/\gamma$  times, so that the standard errors for estimating the means are on the scale of  $\sqrt{\gamma}$ , comparable in size to the gaps between the arm means. Playing the arms significantly less times results in their means essentially being indistinguishable. 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  $\gamma$  and each arm  $k \in [K]$ , the rewards  $X_k^\gamma(i)$ ,  $i = 1, 2, \dots$  are stationary (with independence across different arms) with mean  $\mu_k^\gamma$ , which satisfies the scaling in (5) from Assumption 1. Additionally, for each arm  $k$ , there exists  $\sigma_k > 0$  such that the process*

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

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

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

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

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



REMARK 2. The condition in (9) implies that the finite-dimensional distributions of the  $Z_k^\gamma$  converge to the corresponding multivariate normal distributions. We assume (9) to ensure 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. In (8),  $\sigma_k^2$  is the long-run variance constant corresponding to the time-average of the stationary reward process; normalizing by  $\sigma_k$  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:

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

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:

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

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

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

with  $Z^\gamma = (Z_k^\gamma, k \in [K])$  and each  $Z_k^\gamma$  as defined in (8), and with  $U_k^\gamma$  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  $\mu_k^\gamma$  and  $\sigma_k$  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  $\sigma_k^2$ , 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 Lipschitz continuity of the SDEs.

### 1) Smoothing via Concentrated Priors

One way to smooth out the initial behavior of TS is to use a concentrated prior. We use this approach in Sections 3.1 and 3.2. From Assumption 1, the arm means  $\mu_k^\gamma$  are concentrated around  $\mu_*$  with sub-optimality gaps  $\sqrt{\gamma}\Delta_k^\gamma$ , where the  $\Delta_k^\gamma$  are unknown. We assume that  $\mu_*$  and  $\sqrt{\gamma}$  are known, and we use an independent  $N(\mu_*, \gamma/b)$  prior for each arm in the Gaussian Thompson sampler, with fixed  $b > 0$ . Translated into practice, this means that the experimenter knows that the arm means are in a “ $\sqrt{\gamma}$ -scale neighborhood” of  $\mu_*$  (so that the sub-optimality gaps, i.e., effect sizes, are on the scale of  $\sqrt{\gamma}$ ), perhaps from similar experiments run in the past. (To keep the algebra simple, we assume without loss of generality that  $\mu_* = 0$ .) Then, the experimenter will run a bandit experiment over time horizons scaling as  $1/\gamma$  to learn about the sub-optimality gaps and maximize cumulative reward.

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

### 2) Smoothing via $\epsilon$ -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  *$\epsilon$ -warm-start* (defined below), and we will use it in Section 3.3 and in Section 4.

DEFINITION 1 ( $\epsilon$ -WARM-START). Fix some positive probabilities  $q_1, \dots, q_K$  (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  $\epsilon$ -warm-start, we can ensure Lipschitz continuity of the SDE approximation, and thus a unique strong solution. Moreover, the prior used in TS can be general and need not be concentrated in any way. We can also think of  $\epsilon$ -warm-start as an empirical Bayes approach, where a tiny fraction of data is collected initially to learn a prior with the centering around  $\mu_*$  and the variance scale of  $\gamma$ , 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 Thomson sampler in Theorem 1 below.

To begin, we show that the dynamics in this setting can be described by the evolution of the processes  $(U^\gamma, S^\gamma)$  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  $k$ :

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

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

$$\mathbb{P} \left( k = \arg \max_{l \in [K]} \tilde{\mu}_l^\gamma(j+1) \mid \mathcal{H}_j^\gamma \right) \quad (15)$$

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

$$= p_k^\gamma(U^\gamma(j\gamma), S^\gamma(j\gamma)), \quad (17)$$

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

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

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

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

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

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:

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

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

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

$$p_k^\gamma(u, s) \rightarrow p_k(u, s) \quad (23)$$

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

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

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

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

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

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 earlier in this section, the functions  $p_k$  in (24) are Lipschitz continuous, which ensures that the SDEs in (25)-(26) have a unique strong solution. (For the definition and a theoretical guarantee of strong uniqueness, see Definitions 2.1 and 2.3 and also Theorem 2.5 from Chapter 5.2 of Karatzas and Shreve (1998).) Before stating Theorem 1, in Remark 3 below, we note an expression for regret that will be used throughout the rest of the paper.

REMARK 3. Under the setup of Assumption 1, for a particular  $\gamma$  value, the overall regret  $\text{Reg}^\gamma(n)$  at time  $n$  is related to the  $U_k^\gamma$  processes by:

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

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 prior variance scaling as  $\gamma$ , we have

$$(U^\gamma, S^\gamma) \Rightarrow (U, S) \quad (28)$$

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

$$dU_k(t) = p_k(U(t), S(t))dt \quad (29)$$

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

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

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

Moreover, for regret,

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

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  $k$ :

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

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

$$\mathbb{P} \left( k = \arg \max_{l \in [K]} \tilde{\mu}_l^\gamma(j+1) \mid \mathcal{H}_j^\gamma \right) \quad (34)$$

$$= \mathbb{P} \left( k = \arg \max_{l \in [K]} \left\{ \frac{Z_l^\gamma(U_l^\gamma(j\gamma))\sigma_l + U_l^\gamma(j\gamma)d_l^\gamma}{U_l^\gamma(j\gamma) + bc_*^2} + \frac{c_*}{\sqrt{U_l^\gamma(j\gamma) + bc_*^2}} \mathcal{N}_l \right\} \mid U^\gamma(j\gamma), Z^\gamma \circ U^\gamma(j\gamma) \right) \quad (35)$$

$$= p_k^\gamma(U^\gamma(j\gamma), Z^\gamma \circ U^\gamma(j\gamma)), \quad (36)$$

where the  $\mathcal{N}_l$  are independent standard Gaussian random variables, and functions  $p_k^\gamma$  are given by (24). We can then re-express  $U_k^\gamma(t)$  as

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

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

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

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  $M^\gamma$  converges weakly to the  $D^K[0, \infty)$  zero process. Moreover, as discussed in Remark 2,  $Z^\gamma$  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:

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

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 weak (in distribution) uniqueness of the stochastic ODE solution in Theorem 2 follows from Theorem 3 below (which establishes a general equivalence between SDE and stochastic ODE solutions), together with standard SDE strong uniqueness theory. (For the definition of weak uniqueness, see Definitions 3.1 and 3.4 from Chapter 5.3 of Karatzas and Shreve (1998).) Before stating Theorem 2, in Definition 2, we specify what it means to be a non-anticipative solution to the stochastic ODE.

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

$$\mathcal{F}_u = \bigcap_{\epsilon > 0} \left( \sigma(B_k(t_k), t_k \leq u_k + \epsilon, k \in [K]) \vee \sigma(\mathcal{L}) \right), \quad (40)$$

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 (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 for any  $t \geq 0$ ,  $U(t)$  is a  $\mathcal{F}_u$ -stopping time, i.e.,  $\{U_k(t) \leq u_k\} \in \mathcal{F}_u$  for any  $u \in [0, \infty)^K$ ,  $k \in [K]$ . (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).)

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 prior variance scaling as  $\gamma$ , we have

$$(U^\gamma, Z^\gamma \circ U^\gamma) \Rightarrow (U, B \circ U) \quad (41)$$

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

$$dU_k(t) = p_k(U(t), B \circ U(t)) dt \quad (42)$$

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

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 4. 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. However, in the stationary rewards setting (without independence), the random table model and the reward stack model are not probabilistically equivalent, and this line of deduction is invalid.



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. Starting with a non-anticipative solution to the stochastic ODE, we show that we can always convert from the stochastic ODE representation to the SDE representation. We can then leverage the uniqueness theory for SDEs to deduce uniqueness about solutions to the corresponding stochastic ODEs. As mentioned earlier, the solution uniqueness to the stochastic ODE in Theorem 2 follows from Theorem 3. The proof of Theorem 3 can be found in Section 5.2.

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

$$dU_k(t) = p_k(U(t), B \circ U(t))dt \quad (44)$$

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

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

$$dU_k(t) = p_k(U(t), S(t))dt \quad (46)$$

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

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

*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 can be found 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  $\tilde{B}$  and functions  $p_k : [0, \infty)^K \times \mathbb{R}^K \rightarrow (0, 1)$ . 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, with the functions  $p_k$  as defined in (24), it is important for  $(u, s) \mapsto p_k(u, s)$  to be Lipschitz continuous, which ensures that the limit SDE has a unique

strong solution 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 Lipschitz continuous limit sampling probabilities  $p_k$ . In such settings, there may not be uniqueness of solutions. 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  $p_k^\gamma(u, s) \rightarrow p_k(u, s)$  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:*

$$dU_k(t) = p_k(U(t), B \circ U(t))dt \quad (49)$$

$$U_k(0) = 0, \quad k \in [K], \quad (50)$$

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

Next, we develop a diffusion approximation for the Gaussian Thompson sampler without assuming a concentrated prior with variance scaling as  $\gamma$ . Specifically, we consider any fixed Gaussian prior (with constant variance) in the asymptotics as  $\gamma \downarrow 0$ . Unlike in Sections 3.1-3.2, here we can take  $\mu_* \in \mathbb{R}$  to be unknown, since we are not using concentrated priors and do not need to center such priors on  $\mu_*$ . Then, the functions  $p_k$  in (24) become:

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

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

However, as discussed at the beginning of Section 3, the function  $(u, s) \mapsto p_k(u, s)$  in (51) is no longer Lipschitz continuous for points near  $u_l = 0$ ,  $l \in [K]$ . Nevertheless, the problem with the  $p_k$  in (51) only exists for an infinitesimally small initial interval. Whenever all inputs  $U_l(t)$ ,  $l \in [K]$  to the  $u_l$  components in (51) become strictly positive, then from that time onward, there is Lipschitz continuity of the  $p_k$ , which ensures that there is a unique strong solution to the SDE. In Theorem 4, we use  $\epsilon$ -warm-start (recall Definition 1) to ensure Lipschitz continuity. 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 a fixed prior variance (no  $\gamma$ -dependence) and  $\epsilon$ -warm-start (with initial sampling probabilities  $q_k > 0$ ,  $\sum_k q_k = 1$ ). Then, Theorems 1 and 2 hold with the functions  $p_k : [0, \infty)^K \times \mathbb{R}^K \rightarrow (0, 1)$  defined by:

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

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 any prior distribution satisfying modest regularity conditions and any single-parameter exponential family likelihood.) In the literature, minimax or worst-case regret analysis (which is essentially 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 simply analyze 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  $P^\mu$  parameterized by mean  $\mu$ , with the form:

$$P^\mu(dx) = \exp(\theta(\mu) \cdot x - \Lambda(\mu)) P(dx). \quad (53)$$

In (53),  $P$  is a base distribution,  $\theta(\mu) \in \mathbb{R}$  is the value of the tilting parameter resulting in a mean of  $\mu$ , and  $\Lambda$  is the cumulant generating function. Let  $(\underline{\mu}, \bar{\mu})$  denote the open interval of all possible mean values achievable by the family  $P^\mu$  (for some value of the tilting parameter  $\theta(\mu) \in \mathbb{R}$ ).

For simplicity, suppose we know that the mean reward for all arms belong to a bounded, open interval  $\mathcal{I}$ , with  $\inf \mathcal{I} > \underline{\mu}$  and  $\sup \mathcal{I} < \bar{\mu}$ . (The analysis is simplified by avoiding the boundaries  $\underline{\mu}$  and  $\bar{\mu}$ .) Suppose also that Assumption 1 holds, and that for the distributions  $Q_k^\gamma$  with means  $\mu_k^\gamma$

from Assumption 1, we have  $Q_k^\gamma = P^{\mu_k^\gamma}$ , with all  $\mu_k^\gamma \in \mathcal{I}$ . For the  $\sigma_k$  in (6), here we have  $\sigma_k = \sigma_*$  for all  $k$ , where  $\sigma_*^2$  is the variance of  $P^{\mu_*}$ , with the  $\mu_*$  from (5).

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

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 4 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 (53), and that the corresponding EF Thompson sampler uses a prior having continuous and positive density in a neighborhood of  $\mu_*$ .*

*Then, for the EF Thompson sampler under  $\epsilon$ -warm-start (with initial sampling probabilities  $q_k > 0$ ,  $\sum_k q_k = 1$ ) and the processes  $(U^\gamma, S^\gamma)$  from (11)-(12), we have*

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

*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  $p_k : [0, \infty)^K \times \mathbb{R}^K \rightarrow (0, 1)$  defined by:*

$$p_k(u, s) = \begin{cases} q_k & \sum_l u_l \leq \epsilon \\ \mathbb{P} \left( k = \arg \max_{l \in [K]} \left\{ \frac{s_l \sigma_*}{u_l} + d_l + \frac{\sigma_*}{\sqrt{u_l}} \mathcal{N}_l \right\} \right) & \sum_l u_l > \epsilon, \end{cases} \quad (54)$$

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

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

For simplicity, we have only stated Theorem 5 from the SDE viewpoint. As discussed in Remark 4, we could also formulate Theorem 5 in the small gap regime of Assumption 1 (with iid rewards for each arm) and the reward stack model of reward feedback, which would naturally lead to a stochastic ODE representation. Recall that the SDE and stochastic ODE are distributionally equivalent representations of the weak limit, as established in Theorem 3 and Proposition 1.

The conclusion of Theorem 5 (with the  $p_k(u, s)$  in (54)) matches that of Theorem 4 (with the  $p_k(u, s)$  in (52)) when (in the context of Theorem 4) the limit variances  $\sigma_k^2$  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, Elmachoub et al. 2017, Vaswani et al. 2018, Kveton et al. 2019a,b, Russo et al. 2019, Kveton et al. 2020b,a, Baudry et al. 2020). In this section, we consider the bootstrap sampler introduced earlier, which is one natural implementation of bootstrapping to induce exploration in bandit problems. For the bootstrap sampler, in each time period, a single (non-parametric) bootstrapped sample mean is generated for each arm, and the arm with the greatest one is played.

In 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 in Theorem 5, here the reward distributions do not need to belong to any exponential family. Here, we allow for arbitrary reward distributions  $P^\mu$  with means  $\mu \in \mathcal{I}$ , where  $\mathcal{I} \subset \mathbb{R}$  is an open interval. The only requirement on the  $P^\mu$  is that the condition in (55) is satisfied. 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 5 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*

$$\lim_{y \rightarrow \infty} \sup_{\mu \in \mathcal{I}} \mathbb{E}[(X^\mu)^2 \mathbb{I}((X^\mu)^2 > y)] = 0, \quad (55)$$

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

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

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

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  $p_k : [0, \infty)^K \times \mathbb{R}^K \rightarrow (0, 1)$  defined by:

$$p_k(u, s) = \begin{cases} q_k & \sum_l u_l \leq \epsilon \\ \mathbb{P} \left( k = \arg \max_{l \in [K]} \left\{ \frac{s_l \sigma_l}{u_l} + d_l + \frac{\sigma_l}{\sqrt{u_l}} \mathcal{N}_l \right\} \right) & \sum_l u_l > \epsilon, \end{cases} \quad (56)$$

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

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

For simplicity, we have only stated Theorem 6 from the SDE viewpoint. As discussed in Remark 4, we could also formulate Theorem 6 in the small gap regime of Assumption 1 (with iid rewards for each arm) and the reward stack model of reward feedback, which would naturally lead to a stochastic ODE representation. Recall that the SDE and stochastic ODE representations are distributionally equivalent representations of the weak limit, as established in Theorem 3 and Proposition 1.

Compared to Theorem 4 (with the  $p_k(u, s)$  in (52)), in Theorem 6 (with the  $p_k(u, s)$  in (56)), the bootstrap sampler automatically adapts to the limit variance  $\sigma_k^2$  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 (56), compared to the  $(c_*/\sqrt{u_l})\mathcal{N}_l$  terms in (52).

### 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  $\sigma_k^2$  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 3) 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  $\sigma$  dependence expressed in (24). Then, the distribution of  $(U, S)$  is continuous with respect to  $\sigma$ , i.e., for any bounded continuous function  $f : D^{2K}[0, \infty) \rightarrow \mathbb{R}$ , the mapping  $\sigma \mapsto \mathbb{E}^\sigma[f(U, S)]$  is continuous. Moreover, for any fixed  $t > 0$ ,*

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

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  $\sigma$  dependence expressed in (52).

#### 4.4. Batched Updates

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

**PROPOSITION 3.** *In the settings of Theorems 1, 2 and 4, the same conclusions hold for the Gaussian Thompson sampler with batches of size  $o(n)$ .*



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

## 5. Proofs for Main Results

### 5.1. Proofs for SDE Approximation

In this section, we prove the SDE approximation in Theorem 1 (from Section 3.1). We first discuss a (random) step function approximation (with any desired accuracy) for functions in  $D^m[0, \infty)$ , due to [Kurtz and Protter \(1991\)](#). The step function approximation is technically useful for passing from the discrete versions of Itô integrals to the Itô integrals themselves in the continuous weak limit. We describe the approximation in Definition 3, and discuss integration with the approximation applied to the integrand in Definition 4. Then, Lemma 1, which is a summary of useful technical results from [Kurtz and Protter \(1991\)](#) (see their Lemma 6.1 and its proof, as well as the proof of their Theorem 1), ensures that integration is a continuous mapping when the integrand is approximated in such a way. Following these, we provide proof of Theorem 1 using the continuity properties at hand together with the CMT. We conclude the section with Lemmas 2 and 3, which establish the tightness of stochastic processes and convergence to Brownian motion used in the proof of Theorem 1.

**DEFINITION 3 (STEP FUNCTION APPROXIMATION).** For any  $\epsilon > 0$ , we construct a random step function mapping  $\chi^\epsilon : D^m[0, \infty) \rightarrow D^m[0, \infty)$  as follows. We use the  $\ell^1$  norm, with  $\|w\| := \sum_{i=1}^m |w_i|$  for  $w \in \mathbb{R}^m$ . For any  $z \in D^m[0, \infty)$ , define inductively the random times  $\tau_j(z)$  starting with  $\tau_0(z) = 0$ :

$$\tau_{j+1}(z) = \inf\{t > \tau_j(z) : \max(\|z(t) - z(\tau_j(z))\|, \|z(t-) - z(\tau_j(z))\|) \geq \epsilon V_j\}, \quad (57)$$

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

$$\chi^\epsilon(z)(t) = z(\tau_j(z)), \quad t \in [\tau_j(z), \tau_{j+1}(z)), \quad (58)$$

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

$$\sup_{t \geq 0} \|\chi^\epsilon(z)(t) - z(t)\| \leq \epsilon. \quad (59)$$

**DEFINITION 4 (INTEGRATION WITH STEP FUNCTIONS).** On an interval  $[a, b]$ , let  $f_1, f_2$  be  $\mathbb{R}$ -valued right-continuous functions with left limits, where  $f_1$  is a step function with jump points

$t_1 < \dots < t_j$  in  $[a, b]$ . We will always use the following definition of integration for step function integrands (setting  $t_0 = a$  and  $t_{j+1} = b$ ):

$$\int_a^b f_1(t) df_2(t) = \sum_{i=0}^j f_1(t_i) (f_2(t_{i+1}) - f_2(t_i)). \quad (60)$$

Then, for any  $\epsilon > 0$ , with the random step function mapping  $\chi^\epsilon : D^m[0, \infty) \rightarrow D^m[0, \infty)$  in (58), define the integral mapping  $\mathcal{I}^\epsilon : D^{2m}[0, \infty) \rightarrow D^m[0, \infty)$  component-wise for  $k = 1, \dots, m$  via

$$\mathcal{I}_k^\epsilon(g, h)(t) = \int_0^t \chi_k^\epsilon(g)(v) dh_k(v), \quad (61)$$

for  $g, h \in D^m[0, \infty)$  (with the definition of integral in (60)).

**LEMMA 1 (Continuity of Integration with Step Functions).** *Let the sequences  $x_n, y_n \in D^m[0, \infty)$  and also  $x, y \in D^m[0, \infty)$  such that jointly  $(x^n, y^n) \rightarrow (x, y)$  in  $D^{2m}[0, \infty)$  as  $n \rightarrow \infty$ . For  $\epsilon > 0$ , let  $\chi^\epsilon : D^m[0, \infty) \rightarrow D^m[0, \infty)$  be the random step function mapping in (58), and let  $\mathcal{I}^\epsilon : D^{2m}[0, \infty) \rightarrow D^m[0, \infty)$  be the integral mapping in (61). Then,*

$$(x^n, y^n, \mathcal{I}^\epsilon(x^n, y^n)) \xrightarrow{a.s.} (x, y, \mathcal{I}^\epsilon(x, y)) \quad (62)$$

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

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

*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  $(U^\gamma, S^\gamma)$ , 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 (29)-(30) are Lipschitz-continuous and bounded on their domain of definition, the SDE has a unique strong solution (Theorem 5.2.9 of Karatzas and Shreve (1998)). Thus,  $(U^\gamma, S^\gamma)$  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  $\gamma$ . So, we have

$$(U^\gamma, S^\gamma, B^\gamma, M^\gamma) \Rightarrow (U, S, B, M). \quad (63)$$

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

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

$$A_k^\gamma(t) = p_k^\gamma(U^\gamma(t), S^\gamma(t)) \quad (64)$$

$$A_k(t) = p_k(U(t), S(t)). \quad (65)$$

Since  $p_k^\gamma(u, s) \rightarrow p_k(u, s)$  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 (64)-(65), we have from (63),

$$(U^\gamma, S^\gamma, B^\gamma, M^\gamma, A^\gamma) \Rightarrow (U, S, B, M, A). \quad (66)$$

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

$$\tilde{U}_k^\gamma(t) = \int_0^t p_k^\gamma(U^\gamma(v), S^\gamma(v)) dv \quad (67)$$

$$\tilde{U}_k(t) = \int_0^t p_k(U(v), S(v)) dv. \quad (68)$$

Recall that

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

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

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

we have for any  $T > 0$ ,

$$\sup_{0 \leq t \leq T} \left| U_k^\gamma(t) - \tilde{U}_k^\gamma(t) \right| \xrightarrow{\mathbb{P}} 0. \quad (69)$$

Thus, by the fact that integration is a continuous functional with respect to the Skorohod metric (Theorem 11.5.1 of Whitt (2002)) and the CMT, we have from (66),

$$(U^\gamma, S^\gamma, B^\gamma, \tilde{U}^\gamma, A^\gamma) \Rightarrow (U, S, B, \tilde{U}, A). \quad (70)$$

Let  $\epsilon > 0$ . Let  $\chi^\epsilon$  be the random step function mapping defined in (57) and (58), and let  $\mathcal{I}^\epsilon$  be the corresponding integral operator defined in (61). Recall from (20), (22) and (64), that for each  $k \in [K]$ ,

$$S_k^\gamma(t) = \int_0^t \sqrt{A_k^\gamma(v-)} dB_k^\gamma(v), \quad (71)$$

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

$$\widehat{S}_k^\gamma(t) = \int_0^t \chi_k^\epsilon \left( \sqrt{A^\gamma(v-)} \right) dB_k^\gamma(v). \quad (72)$$

By Lemma 1 and the CMT, with the continuity of the mapping  $(x, y) \mapsto (x, y, \mathcal{I}^\epsilon(x, y))$  established in (62), we have from (70),

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

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

$$\widehat{S}_k(t) = \int_0^t \chi_k^\epsilon \left( \sqrt{A(v-)} \right) dB_k(v). \quad (74)$$

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

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

Note that both of the processes in (74) and (75) are well defined as Itô integrals, since by Lemma 3, the integrands are non-anticipative with respect to the Brownian motions  $B_k$ . (As defined in (57)-(58),  $\chi^\epsilon$  depends on exogenous randomization that is independent of the  $B_k$ .) By Lemma 1, because  $\chi^\epsilon$  is an  $\epsilon$ -uniform approximation (see (59)), for each  $k \in [K]$  and any  $T > 0$ ,

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

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

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

where  $t \mapsto \langle B_k \rangle_t$  denotes the quadratic variation process for  $B_k$ . Putting together (69), (73)-(77) and sending  $\epsilon \downarrow 0$ , we have

$$(U^\gamma, S^\gamma, B^\gamma, U^\gamma, S^\gamma) \Rightarrow (U, S, B, \widetilde{U}, \widetilde{S}). \quad (78)$$

Recalling the definition of  $\tilde{U}$  in (68) as well as that of  $\tilde{S}$  in (75) and the  $A_k$  in (65), we see from (78) that the limit processes  $(U, S, B)$  satisfy the SDE:

$$U_k(t) = \int_0^t p_k(U(v), S(v)) dv \quad (79)$$

$$S_k(t) = \int_0^t \sqrt{p_k(U(v), S(v))} dB_k(v), \quad k = 1, \dots, K. \quad (80)$$

□

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  $k = 1, \dots, K$ .

$$U_k^\gamma(t) = \gamma \sum_{i=1}^{\lfloor t/\gamma \rfloor} I_k^\gamma(i) \quad (81)$$

$$S_k^\gamma(t) = \sqrt{\gamma} \frac{1}{\sigma_k} \sum_{i=1}^{\lfloor t/\gamma \rfloor} I_k^\gamma(i) (X_k^\gamma(i) - \mu_k^\gamma) \quad (82)$$

$$M_k^\gamma(t) = \gamma \sum_{i=0}^{\lfloor t/\gamma \rfloor - 1} (I_k^\gamma(i+1) - p_k^\gamma(U^\gamma(i\gamma), S^\gamma(i\gamma))) \quad (83)$$

$$B_k^\gamma(t) = \sqrt{\gamma} \frac{1}{\sigma_k} \sum_{i=0}^{\lfloor t/\gamma \rfloor - 1} \frac{I_k^\gamma(i+1) (X_k^\gamma(i+1) - \mu_k^\gamma)}{\sqrt{p_k^\gamma(U^\gamma(i\gamma), S^\gamma(i\gamma))}} \quad (84)$$

Note that (81)-(82) are just different expressions of the same quantities in (19)-(20). Let  $\tilde{\mathcal{H}}^\gamma = (\tilde{\mathcal{H}}_t^\gamma, t \geq 0)$  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), so that (81)-(84) are all adapted to  $\tilde{\mathcal{H}}_t^\gamma$ . Also, the process in (81) is uniformly bounded and increasing, and those in (82)-(84) 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. For fixed  $T > 0$ ,  $\delta > 0$  and all  $\gamma > 0$ , we can set  $A_\delta^\gamma(T) = \delta$  for each individual component process, and we can set  $A_\delta^\gamma(T) = 4\delta$  (using the bound:  $(x+y)^2 \leq 2x^2 + 2y^2$ ) for each pairwise sum of component processes. □

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

*Proof of Lemma 3.* To show that  $B^\gamma \Rightarrow B$ , where  $B$  is standard  $K$ -dimensional Brownian motion, we apply the martingale functional central limit theorem stated in Lemma 10. Below, we verify (M1) and (M2) to ensure Lemma 10 holds. The non-anticipative property follows from the same property in the pre-limit, i.e.,  $U^\gamma$  and  $S^\gamma$  are non-anticipative with respect to  $B^\gamma$ .

#### Verification of (M1)

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

$$\begin{aligned} & \gamma \left( \frac{\sigma_k^\gamma}{\sigma_k} \right)^2 \sum_{i=0}^{\lfloor t/\gamma \rfloor} \mathbb{E} \left[ \frac{I_k^\gamma(i+1)}{p_k^\gamma(i)} \left( \frac{X_k^\gamma(i+1) - \mu_k^\gamma}{\sigma_k^\gamma} \right)^2 \middle| \mathcal{H}_i^\gamma \right] \\ &= \gamma \left( \frac{\sigma_k^\gamma}{\sigma_k} \right)^2 \sum_{i=0}^{\lfloor t/\gamma \rfloor - 1} \frac{\mathbb{E} \left[ I_k^\gamma(i+1) \middle| \mathcal{H}_i^\gamma \right]}{p_k^\gamma(i)} \mathbb{E} \left[ \left( \frac{X_k^\gamma(i+1) - \mu_k^\gamma}{\sigma_k^\gamma} \right)^2 \middle| \mathcal{H}_i^\gamma \right] \end{aligned} \quad (85)$$

$$= \gamma \left( \frac{\sigma_k^\gamma}{\sigma_k} \right)^2 \lfloor t/\gamma \rfloor \rightarrow t \quad (86)$$

as  $\gamma \downarrow 0$ . Here, (85) 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 (86) follows from (6) in Assumption 1.

#### Verification of (M2)

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

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

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

$$\mathbb{E} [W_k^\gamma(i+1)^2 \mathbb{I}(|W_k^\gamma(i+1)| > \epsilon/\sqrt{\gamma})] \rightarrow 0 \quad (87)$$

as  $\gamma \downarrow 0$ .

We have the following three observations. 1)  $(U^\gamma, S^\gamma)$  is a tight sequence, as established in Lemma 2, which implies stochastic boundedness of each component with respect to the supremum norm. 2)  $p_k^\gamma(u, s) \rightarrow p_k(u, s)$  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  $\gamma$  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

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

$$= \mathbb{P}(p_k^\gamma(i) < \delta) \quad (89)$$

$$\begin{aligned} & \leq \mathbb{P} \left( \inf_{v \in [0, t]} p_k^\gamma(U^\gamma(v), S^\gamma(v)) < \delta \right) \\ & \leq \eta, \end{aligned} \quad (90)$$

where (88) follows from  $U^\gamma(i\gamma)$  and  $S^\gamma(i\gamma)$  being  $\mathcal{H}_i^\gamma$ -measurable, (89) 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 (90) holds for  $\gamma$  sufficiently close to zero, as established above.

Additionally, we have

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

$$\leq \frac{C}{\delta} \mathbb{E} \left[ \mathbb{P} \left( |W_k^\gamma(i+1)| > \epsilon/\sqrt{\gamma} \middle| \mathcal{H}_i^\gamma \right)^{\alpha/(2+\alpha)} \right], \quad (92)$$

where (91) follows from Hölder's inequality, and (92) follows from (7) in Assumption 1, with constant  $C > 0$ . Furthermore, almost surely,

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

So, by the bounded convergence theorem, the right side of (92) converges to zero as  $\gamma \downarrow 0$ .

Therefore, from (90) and (92), we have

$$\limsup_{\gamma \downarrow 0} \mathbb{E} \left[ W_k^\gamma(i+1)^2 \mathbb{I}(|W_k^\gamma(i+1)| > \epsilon/\sqrt{\gamma}) \right] \leq \eta, \quad (93)$$

and sending  $\eta \downarrow 0$  yields (87).  $\square$



## 5.2. Proofs for Stochastic ODE Approximation

In this section, we prove Theorem 2 and Theorem 3 from Section 3.2. Theorem 2 shows that under Assumption 2 (with stationary rewards) and the reward stack model of reward generation, the pre-limit processes converge weakly to the unique (in distribution) weak solution to the stochastic ODE. Theorem 3 establishes the equivalence in distribution between the solution to the SDE and the solution to the stochastic ODE, and moreover, that the stochastic ODE has a unique (in distribution) weak solution whenever the SDE has a unique strong solution.

*Proof of Theorem 2.*

We start with the pre-limit processes (8), (13), (37) and (38) from our derivation in Section 3.2. We denote the joint processes via  $(U^\gamma, Z^\gamma, M^\gamma) = (U_k^\gamma, Z_k^\gamma, M_k^\gamma, k \in [K])$ , and recall that they are processes in  $D^{3K}[0, \infty)$ .

Analogous to the proof of Theorem 1, we will show that any weakly convergent subsequence of  $(U^\gamma, Z^\gamma)$  must converge weakly to the unique (in distribution) weak solution to the stochastic ODE. (Recall that the existence and uniqueness (in distribution) of the weak solution to the stochastic ODE was established by A) and B) above.) Consider such a weakly convergent subsequence of  $(U^\gamma, Z^\gamma)$ , which we will still index by  $\gamma$  for notational simplicity. Then, jointly  $(U^\gamma, Z^\gamma, M^\gamma) \Rightarrow (U, Z, M)$ , where (as in the proof of Theorem 1)  $M$  is the  $D^K[0, \infty)$  zero process. By Donsker's Theorem (Chapter 3 of Billingsley (1999)),  $Z$  is standard  $K$ -dimensional Brownian motion.

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

$$(U^\gamma, Z^\gamma, M^\gamma, Z^\gamma \circ U^\gamma) \Rightarrow (U, Z, M, Z \circ U). \quad (94)$$

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

$$A_k^\gamma(t) = p_k^\gamma(U^\gamma(t), Z^\gamma \circ U^\gamma(t)) \quad (95)$$

$$A_k(t) = p_k(U(t), Z \circ U(t)). \quad (96)$$

Since  $p_k^\gamma(u, s) \rightarrow p_k(u, s)$  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 (95)-(96), we have from (94),

$$(U^\gamma, Z^\gamma, M^\gamma, A^\gamma) \Rightarrow (U, Z, M, A). \quad (97)$$

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

$$\begin{aligned} \tilde{U}_k^\gamma(t) &= \int_0^t p_k^\gamma(U^\gamma(v), Z^\gamma \circ U^\gamma(v)) dv \\ \tilde{U}_k(t) &= \int_0^t p_k(U(v), Z \circ U(v)) dv. \end{aligned} \quad (98)$$

Recall that

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

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

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

we have for any  $T > 0$ ,

$$\sup_{0 \leq t \leq T} \left| U_k^\gamma(t) - \tilde{U}_k^\gamma(t) \right| \xrightarrow{\mathbb{P}} 0. \quad (99)$$

Thus, by the fact that integration is a continuous functional with respect to the Skorohod metric (Theorem 11.5.1 of [Whitt \(2002\)](#)) and the CMT, we have from (97),

$$(U^\gamma, Z^\gamma, \tilde{U}^\gamma) \Rightarrow (U, Z, \tilde{U}). \quad (100)$$

Together, (99)-(100) yield

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

and recalling the definition of  $\tilde{U}$  in (98), we have established the desired weak convergence.  $\square$

*Proof of Theorem 3.*

Let  $(\Omega, \mathbb{F}, \mathbb{P})$  be a probability space, and let  $\mathcal{F} = (\mathcal{F}_u, u \in [0, \infty)^K)$  be an augmented (right-continuous) filtration as defined in (40), with independent standard Brownian motions  $B_k$ ,  $k \in [K]$ . (Here, each of the Brownian motions  $B_k$  has its own separate clock/time index  $u_k$ .) 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,  $\phi_\theta(u)$  is a  $\mathcal{F}_u$ -martingale, where the  $\mathcal{F}_u$  filtration is indexed by the directed set  $[0, \infty)^K$ ; see [Kurtz \(1980b\)](#) or Chapter 2, Section 8 of [Ethier and Kurtz \(1986\)](#). Let  $U = (U_k, k \in [K])$  be a non-anticipative solution (as in Definition 2) to the stochastic ODE:

$$U_k(t) = \int_0^t p_k(U(v), B \circ U(v)) dv, \quad k \in [K], \quad (101)$$

with  $K$ -dimensional standard Brownian motion  $B$ . So, the  $U(t)$ ,  $t \in [0, \infty)$  are  $\mathcal{F}_u$ -stopping times. Define  $S = (S_k, k \in [K])$  by  $S_k(t) := B_k(U_k(t))$ . Then, the conditions of Theorem 6.3 of [Kurtz \(1980a\)](#) are satisfied. Using part (a) of Theorem 6.2, 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  $\tilde{\mathcal{F}} = (\tilde{\mathcal{F}}_t, t \geq 0)$  defined

by  $\tilde{\mathcal{F}}_t := \mathcal{F}_{U(t)}$ . Using the fact that  $U_k(t) \leq t$  almost surely, together with the Doob's Maximal Inequality (see Theorem 3.8(iv) from Chapter 1 of [Karatzas and Shreve \(1998\)](#)), we have for each  $k \in [K]$ ,

$$\mathbb{E} \left[ \sup_{v \in [0, t]} S_k^2(v) \right] \leq \mathbb{E} \left[ \left( \sup_{v \in [0, t]} |B_k(v)| \right)^2 \right] \leq 4\mathbb{E}[B_k^2(t)] = 4t. \quad (102)$$

From (102), 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  $\tilde{\mathcal{F}}_t$ . 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,

$$\langle S_k \rangle_t = U_k(t) \quad (103)$$

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

$$\langle S_k, S_l \rangle_t = U_k(t) \mathbb{I}(k = l). \quad (104)$$

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

$$\tilde{B}_k(t) := \int_0^t \frac{1}{\sqrt{p_k(U(v), S(v))}} dS_k(v), \quad k \in [K]. \quad (105)$$

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

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

Using (104), 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  $\langle \tilde{B}_k, \tilde{B}_l \rangle_t = t\mathbb{I}(k = l)$ . So, by the Lévy characterization of multi-dimensional Brownian motion (see Theorem 3.16 from Chapter 3 of [Karatzas and Shreve \(1998\)](#)),  $\tilde{B}$  is a standard  $K$ -dimensional Brownian motion with respect to the filtration  $\tilde{\mathcal{F}}_t$ . Moreover, from (105) and Corollary 2.20 from Chapter 3 of [Karatzas and Shreve \(1998\)](#), we have for each  $k \in [K]$ ,

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

Thus, we have the representation:

$$U_k(t) = \int_0^t p_k(U(v), S(v)) dv \quad (106)$$

$$S_k(t) = \int_0^t \sqrt{p_k(U(v), S(v))} d\tilde{B}_k(v), \quad k \in [K]. \quad (107)$$

So, the non-anticipative solution  $U$  to the stochastic ODE in (101) with the Brownian motion  $B$ , together with  $S = B \circ U$ , solves the SDE in (106)-(107) with the Brownian motion  $\tilde{B}$ . Furthermore, we have  $(U, B \circ U) = (U, S)$  almost surely.  $\square$

LEMMA 4. In the proof of Theorem 3, (103) holds, i.e.,  $\langle S_k \rangle_t = U_k(t)$ .

*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  $\mathcal{F}_u$ -martingales (with the  $\mathcal{F}_u$  filtration as defined in (40)). Also, for  $0 \leq s \leq t$ , we have that  $U(s)$  and  $U(t)$  are bounded  $\mathcal{F}_u$ -stopping times with  $0 \leq U_k(s) \leq U_k(t) \leq t$  for all  $k \in [K]$ . Then, for each  $k \in [K]$ , we have

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

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  $k \in [K]$ ,  $B_k(U_k(t))^2 - U_k(t)$  is a  $\tilde{\mathcal{F}}_t := \mathcal{F}_{U(t)}$  martingale. The desired result then follows from the Doob-Meyer decomposition (see Definition 5.3 from Chapter 1 of Karatzas and Shreve (1998)).  $\square$

*Proof of Proposition 1.*

Let  $(\Omega, \mathbb{F}, \mathbb{P})$  be a probability space supporting a standard  $K$ -dimensional Brownian motion  $B$ , and let  $\mathcal{F}^B = (\mathcal{F}_t^B, t \geq 0)$  be the augmented (right-continuous) natural filtration corresponding to  $B$ . By Theorem 1, there exists a solution  $(U, S)$  to the SDE (29)-(30) on this probability space with respect to the standard Brownian motion  $B$ . Writing (30) in integral form, because the  $p_k$  functions are bounded,

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

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

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

For  $k \neq l$ , the quadratic co-variation processes  $\langle S_k, S_l \rangle_t = 0$  since  $B_k$  and  $B_l$  are independent. From (29), we see that  $\langle S_k \rangle_t = U_k(t)$ ,  $k \in [K]$ , which are continuous and strictly increasing processes since the  $p_k$  functions are bounded and strictly positive. Define

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

Also, we define  $U_k(\infty) := \lim_{t \rightarrow \infty} U_k(t)$ , 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:  $\tilde{B}_k(t) := S_k(U_k^{-1}(t))$ , appending on an additional independent standard Brownian motion if  $U_k(\infty) < \infty$ . Specifically, define

$$\tilde{B}_k(t) := \begin{cases} S_k(U_k^{-1}(t)), & t < U_k(\infty) \\ S_k(\infty) + W_k(t - U_k(\infty)), & t \geq U_k(\infty), \end{cases}$$

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

By construction, we have  $\tilde{B}_k(U_k(t)) = S_k(t)$ , and substituting this representation into the SDE in (29), we obtain the stochastic ODE representation:

$$U_k(t) = \int_0^t p_k(U(v), \tilde{B} \circ U(v)) dv, \quad k \in [K]. \quad (108)$$

So, with respect to the filtration  $\mathcal{F}^{\tilde{B}}$ , the SDE solution has the representation  $(U, S) = (U, \tilde{B} \circ U)$ , which satisfies the stochastic ODE in (108), with independent  $K$ -dimensional standard Brownian motion  $\tilde{B}$ .

Lastly, to establish the non-anticipative property (recall Definition 2), define the augmented (right-continuous) filtration  $\tilde{\mathcal{F}} = (\tilde{\mathcal{F}}_u, u \in [0, \infty)^K)$ , with

$$\tilde{\mathcal{F}}_u = \bigcap_{\epsilon > 0} \left( \sigma(\tilde{B}_k(t_k), t_k \leq u_k + \epsilon, k \in [K]) \vee \sigma(\mathcal{L}) \right),$$

where  $\mathcal{L} \subset \mathbb{F}$  denotes the collection of all  $\mathbb{P}$ -probability zero sets. Then, from the representation in (108), it is clear that for any  $t \geq 0$ , we have  $\{U_k(t) \leq u_k\} \in \tilde{\mathcal{F}}_u$  for any  $u \in [0, \infty)^K$ ,  $k \in [K]$ . Thus,  $U$  is non-anticipative.  $\square$

## Appendix A: Proof of Theorem 5

*Proof of Theorem 5.* Under  $\epsilon$ -warm-start, we only need to establish the SDE and stochastic ODE approximations on  $[\epsilon, \infty)$ . We verify that the sampling probabilities for EF Thompson samplers have the desired form with  $p_k(u, s)$  as in (54).

In the SDE case, as before, we use the random table model of reward feedback. At time  $j+1$ , conditional on  $\mathcal{H}_j^\gamma$  (as defined in (4)), for each arm  $k$ , we sample a value  $\tilde{\mu}_k^\gamma(j+1)$  from the posterior distribution of  $\mu_k^\gamma$ . Let  $\hat{\mu}_k^\gamma(j+1)$  denote the sample mean estimate at time  $j+1$ . (For the exponential family model, the sample mean is the maximum likelihood estimator (MLE) for the mean, and is used as the centering value for the Gaussian posterior approximation in Proposition 4.) Here, the  $S_k^\gamma$  and  $U_k^\gamma$  have the expressions from (11)-(12). The probability of playing arm  $k$  is given by:

$$\mathbb{P}(k = \arg \max_{l \in [K]} \tilde{\mu}_l^\gamma(j+1) \mid \mathcal{H}_j^\gamma)$$

$$\begin{aligned}
&= \mathbb{P} \left( k = \arg \max_{l \in [K]} \left\{ \frac{S_l^\gamma(j\gamma)\sigma_l}{U_l^\gamma(j\gamma)} + d_l^\gamma + \frac{1}{\sqrt{\gamma}} (\tilde{\mu}_l^\gamma(j+1) - \hat{\mu}_l^\gamma(j+1)) \right\} \mid U^\gamma(j\gamma), S^\gamma(j\gamma) \right) \\
&= \mathbb{P} \left( k = \arg \max_{l \in [K]} \left\{ \frac{S_l^\gamma(j\gamma)\sigma_l}{U_l^\gamma(j\gamma)} + d_l^\gamma + \frac{\sigma_l}{\sqrt{U_l^\gamma(j\gamma)}} \mathcal{N}_l \right\} \mid U^\gamma(j\gamma), S^\gamma(j\gamma) \right) + o_{\mathbb{P}}(1) \tag{109} \\
&= p_k^\gamma(U^\gamma(j\gamma), S^\gamma(j\gamma)) + o_{\mathbb{P}}(1), \tag{110}
\end{aligned}$$

where (109) follows from Proposition 4, with the probability taken over the independent standard Gaussian random variables  $\mathcal{N}_l$ , and

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

Moreover, with  $p_k(u, s)$  as in (54),  $p_k^\gamma(u, s) \rightarrow p_k(u, s)$  uniformly for  $(u, s)$  on compact subsets of  $[0, \infty)^K \times \mathbb{R}^K$ , with the restriction that  $u_k \geq \epsilon q_k > 0$  for each arm  $k$ , due to the initial sampling with constant, positive probabilities  $(q_k, k \in [K])$  in the  $\epsilon$ -warm-start procedure.

This sequence of derivations parallels what we derived in (15)-(17) in Section 3.1. From (110), the proof of Theorem 1 can be applied to yield the desired SDE approximation in (29)-(31).  $\square$

## 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  $P^\mu$  parameterized by mean  $\mu$  (as expressed in (53)), with means  $\mu$  known to belong to a bounded, open interval  $\mathcal{I}$ . Our goal here is to develop Proposition 4 below, which is a version of the Bernstein-von Mises Theorem. This version establishes weak convergence of the rescaled posterior distribution to a Gaussian distribution, almost surely as the sample size  $n \rightarrow \infty$ , and uniformly over the possible data-generating distributions  $P^\mu$ ,  $\mu \in \mathcal{I}$ . 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  $\mu_k^\gamma$  in a  $\sqrt{\gamma}$ -neighborhood of some  $\mu_*$ . As  $\gamma \downarrow 0$  (and the time horizon goes to infinity), the mean parameters  $\mu_k^\gamma$  change. So, for a given large time horizon, the Gaussian approximation to the posterior in Proposition 4 should be valid simultaneously for all distributions  $P^\mu$  corresponding to a range of mean parameters  $\mu$ ; a fixed (not depending on  $\gamma$ ) neighborhood of  $\mu_*$  suffices. Below, we first discuss the “uniform almost sure” mode of convergence, and then move on to the development of Proposition 4.

### “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 5 below, followed by a precise definition of the mode of convergence in Definition 5 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 5. For a sequence of random variables  $Y_1, Y_2, \dots$ ,

$$Y_n \xrightarrow{\text{a.s.}} 0$$

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

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

DEFINITION 5. 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 almost surely to zero, uniformly in  $Q \in \mathcal{Q}$ , if for any  $\epsilon > 0$ ,

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

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

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

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

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

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

### Development of Proposition 4

Before presenting Lemma 6 and then continuing on to Proposition 4, 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 (53) from Section 4.1 satisfies C1 and C2. Notation-wise, corresponding to the exponential family  $P^\mu$ , the log-likelihood function is denoted by  $l(\mu, x)$ , and derivatives of  $l(\mu, x)$  with respect to  $\mu$  are denoted by  $l'(\mu, x)$ ,  $l''(\mu, x)$ , etc. For each  $\mu \in \mathcal{I}$ , let  $X^\mu, X_i^\mu \stackrel{\text{iid}}{\sim} P^\mu$ .

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

$$\sup_{z: |\mu - z| \geq \delta} \mathbb{E}[l(z, X^\mu)] \leq \mathbb{E}[l(\mu, X^\mu)] - \epsilon. \quad (111)$$



(C2) There exists functions  $\eta$  and  $\kappa$  such that for all  $x$  in the support of the base distribution  $P$  (in (53)),

$$\eta(x) \geq \sup_{\mu \in \mathcal{I}} |l'(\mu, x)| \quad (112)$$

$$\kappa(x) \geq \sup_{\mu \in \mathcal{I}} |l'''(\mu, x)|. \quad (113)$$

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

$$\lim_{y \rightarrow \infty} \sup_{\mu \in \mathcal{I}} \mathbb{E}[f(X^\mu) \mathbb{I}(f(X^\mu) > y)] = 0. \quad (114)$$

Applying Theorems 2.7.11 and 2.8.1 of [van der Vaart and Wellner \(1996\)](#) (together with the mean value theorem), we have the following result.

LEMMA 6. Suppose [C1](#) holds together with [C2](#) for the case  $f(x) = \eta(x) + |l(\mu_0, x)|$  with some fixed  $\mu_0 \in \mathcal{I}$  and  $\eta(x)$  as defined in (112). Then,  $\{l(\mu, \cdot), \mu \in \mathcal{I}\}$  is a Glivenko-Cantelli class of functions uniformly in  $P^\mu$ ,  $\mu \in \mathcal{I}$ , i.e., for any  $\epsilon > 0$ ,

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

We now state and prove Proposition 4. The proof is adapted from the proof of Theorem 4.2 in [Ghosh et al. \(2006\)](#). As before, for each  $\mu \in \mathcal{I}$ , let  $X^\mu, X_i^\mu \stackrel{\text{iid}}{\sim} P^\mu$ , with mean  $\mu$  and corresponding variance  $(\sigma^\mu)^2$ . (Below, we will write all relevant quantities with superscript  $\mu$  to keep track of the distribution  $P^\mu$  that we work with.) The sample mean of  $X_1^\mu, \dots, X_n^\mu$  is denoted by  $\hat{m}_n^\mu$ . Given  $n$  such samples, we use  $\tilde{m}_n^\mu$  to denote a sample from the posterior distribution of the mean  $\mu$ .

PROPOSITION 4. Suppose the conditions [C1](#) and [C2](#) hold, with bounded, open interval  $\mathcal{I} \subset \mathbb{R}$ . Let  $\nu_0$  be a bounded prior density with support contained in  $\mathcal{I}$ , that is also continuous and positive on a bounded, open sub-interval  $\mathcal{I}' \subset \mathcal{I}$ . Then, conditional on the data  $X_i^\mu \stackrel{\text{iid}}{\sim} P^\mu$ , the centered and scaled posterior density  $y \mapsto \nu_n(y | X_1^\mu, \dots, X_n^\mu)$  for  $\sqrt{n}(\tilde{m}_n^\mu - \hat{m}_n^\mu)$  satisfies:

$$\lim_{n \rightarrow \infty} \int_{\mathbb{R}} \left| \nu_n(y | X_1^\mu, \dots, X_n^\mu) - \frac{1}{\sqrt{2\pi}\sigma^\mu} \exp \left( -\frac{1}{2(\sigma^\mu)^2} y^2 \right) \right| dy = 0 \quad (115)$$

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

*Proof of Proposition 4.* The posterior density can be expressed as

$$\nu_n(y | X_1^\mu, \dots, X_n^\mu) = (C_n^\mu)^{-1} \nu_0(\hat{m}_n^\mu + y/\sqrt{n}) \exp \left( L_n^\mu(\hat{m}_n^\mu + y/\sqrt{n}) - L_n^\mu(\hat{m}_n^\mu) \right), \quad (116)$$

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

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

Consider the following difference between unnormalized densities.

$$D_n^\mu(y) = \nu_0(\hat{m}_n^\mu + y/\sqrt{n}) \exp(L_n^\mu(\hat{m}_n^\mu + y/\sqrt{n}) - L_n^\mu(\hat{m}_n^\mu)) - \nu_0(\mu) \exp\left(-\frac{1}{2(\sigma^\mu)^2}y^2\right) \quad (117)$$

To show (115), it suffices to show that a.s. uniformly in  $\mu \in \mathcal{I}'$ ,

$$\lim_{n \rightarrow \infty} \int_{\mathbb{R}} |D_n^\mu(y)| dy = 0. \quad (118)$$

Indeed, if (118) holds, we must also have

$$\lim_{n \rightarrow \infty} C_n^\mu = \nu_0(\mu) \sqrt{2\pi} \sigma^\mu, \quad (119)$$

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

$$\begin{aligned} & \int_{\mathbb{R}} \left| \nu_n(y | X_1^\mu, \dots, X_n^\mu) - \frac{1}{\sqrt{2\pi}\sigma^\mu} \exp\left(-\frac{1}{2(\sigma^\mu)^2}y^2\right) \right| dy \\ & \leq (C_n^\mu)^{-1} \int_{\mathbb{R}} |D_n^\mu(y)| dy + \left| (C_n^\mu)^{-1} \nu_0(\mu) - \frac{1}{\sqrt{2\pi}\sigma^\mu} \right| \int_{\mathbb{R}} \exp\left(-\frac{1}{2(\sigma^\mu)^2}y^2\right) dy. \end{aligned} \quad (120)$$

Applying (118) and (119) to (120) would then lead to the desired conclusion in (115).

To show (118), we split the integral into two pieces on  $A_n = \{y : |y| > \beta\sqrt{n}\}$  and  $A_n^c = \{y : |y| \leq \beta\sqrt{n}\}$ , with  $\beta > 0$  to be specified later in the proof. In the first case on  $A_n$ ,

$$\begin{aligned} \int_{A_n} |D_n^\mu(y)| dy & \leq \int_{A_n} \nu_0(\hat{m}_n^\mu + y/\sqrt{n}) \exp(L_n^\mu(\hat{m}_n^\mu + y/\sqrt{n}) - L_n^\mu(\hat{m}_n^\mu)) dy \\ & \quad + \int_{A_n} \nu_0(\mu) \exp\left(-\frac{1}{2(\sigma^\mu)^2}y^2\right) dy. \end{aligned} \quad (121)$$

By the boundedness of  $\nu_0(\mu)$  and  $(\sigma^\mu)^2$  for  $\mu \in \mathcal{I}'$ , the second integral on the right side of (121) goes to zero as  $n \rightarrow \infty$ , uniformly in  $\mu \in \mathcal{I}'$ . For the first integral on the right side of (121), from condition C2 with  $f(x) = |x|$  and Lemma 5, it follows that

$$\hat{m}_n^\mu - \mu \rightarrow 0 \quad (122)$$

as  $n \rightarrow \infty$ , a.s. uniformly in  $\mu \in \mathcal{I}'$ . This along with Lemma 6 implies that there exists  $\epsilon > 0$  such that

$$\sup_{y \in A_n, (\hat{m}_n^\mu + y/\sqrt{n}) \in \mathcal{I}} \frac{1}{n} (L_n^\mu(\hat{m}_n^\mu + y/\sqrt{n}) - L_n^\mu(\hat{m}_n^\mu)) \leq -\epsilon, \quad (123)$$

for sufficiently large  $n$ , a.s. uniformly in  $\mu \in \mathcal{I}'$ . (For the first integral on the right side of (121), we only need to consider  $y$  such that  $(\hat{m}_n^\mu + y/\sqrt{n}) \in \mathcal{I}$ , since the prior density  $\nu_0$  has support contained in  $\mathcal{I}$ .) Therefore, using (123), the first integral in (121) also goes to zero as  $n \rightarrow \infty$ , a.s. uniformly in  $\mu \in \mathcal{I}'$ .

For the second case on  $A_n^c$ , we analyze

$$\int_{A_n^c} |D_n^\mu(y)| dy. \quad (124)$$

We expand  $L_n^\mu$  in a Taylor series about the MLE  $\hat{m}_n^\mu$ , noting that by the definition of the MLE,  $(L_n^\mu)'(\hat{m}_n^\mu) = 0$ . We have

$$\begin{aligned} L_n^\mu(\hat{m}_n^\mu + y/\sqrt{n}) - L_n^\mu(\hat{m}_n^\mu) &= -\frac{1}{2} \frac{1}{n} (L_n^\mu)''(\hat{m}_n^\mu) y^2 + r_n^\mu(y) \\ &= -\frac{1}{2} (\theta''(\hat{m}_n^\mu) \hat{m}_n^\mu - \Lambda''(\hat{m}_n^\mu)) y^2 + r_n^\mu(y), \end{aligned} \quad (125)$$

using the fact that  $l''(z, x) = \theta''(z) \cdot x - \Lambda''(z)$  (recall the definitions of  $\theta(z)$  and  $\Lambda(z)$  from (53)), 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  $\hat{m}_n^\mu$  and  $\hat{m}_n^\mu + y/\sqrt{n}$ . 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  $\mu \in \mathcal{I}'$ ,

$$|r_n^\mu(y)| \leq \frac{1}{6} \frac{y^3}{\sqrt{n}} \frac{1}{n} \sum_{i=1}^n \kappa(X_i^\mu) \leq \frac{1}{6} \frac{y^3}{\sqrt{n}} (\mathbb{E}[\kappa(X^\mu)] + \delta). \quad (126)$$

For  $y \in A_n^c$ , (126) can be re-expressed as

$$|r_n^\mu(y)| \leq \frac{1}{6} \beta y^2 (\mathbb{E}[\kappa(X^\mu)] + \delta). \quad (127)$$

For the first term on the right side of (125), we have

$$\lim_{n \rightarrow \infty} \theta''(\hat{m}_n^\mu) \hat{m}_n^\mu - \Lambda''(\hat{m}_n^\mu) = \theta''(\mu) \mu - \Lambda''(\mu) = \frac{1}{(\sigma^\mu)^2} \quad (128)$$

a.s. uniformly in  $\mu \in \mathcal{I}'$ . In (128), the first equality follows from the uniform continuity of  $\theta''$  and  $\Lambda''$  on  $\mathcal{I}$ , together with the convergence result in (122). The second equality is from a standard identity relating Fisher information and the variance  $(\sigma^\mu)^2$  of the  $X_i^\mu$ . Now, defining  $c_0 := \inf_{w \in \mathcal{I}'} 1/(\sigma^w)^2$  and recognizing that  $c_0 > 0$ , we can choose  $\beta > 0$  to satisfy:

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

Then, using (127) and (128), we have from (125) that

$$\sup_{y \in A_n^c} \frac{\exp(L_n^\mu(\hat{m}_n^\mu + y/\sqrt{n}) - L_n^\mu(\hat{m}_n^\mu))}{\exp(-c_0 y^2/4)} \leq 1 \quad (129)$$

for sufficient large  $n$ , a.s. uniformly in  $\mu \in \mathcal{I}'$ . Thus, on  $A_n^c$ ,  $D_n^\mu(y)$  (as defined in (117)) is dominated by an integrable function for sufficiently large  $n$ , a.s. uniformly in  $\mu \in \mathcal{I}'$ . Furthermore, from (126) we have for any fixed  $y$  that  $r_n^\mu(y) \rightarrow 0$  as  $n \rightarrow \infty$ , a.s. uniformly in  $\mu \in \mathcal{I}'$ . This, together with (122), (125) and (128), we have for any fixed  $y$  that  $D_n^\mu(y) \rightarrow 0$  as  $n \rightarrow \infty$ , a.s. uniformly in  $\mu \in \mathcal{I}'$ . Then, by the dominated convergence theorem, the quantity in (124) converges to zero as  $n \rightarrow \infty$ , a.s. uniformly in  $\mu \in \mathcal{I}'$ .  $\square$

## Appendix C: Gaussian Approximations for the Bootstrap

In Proposition 5 below, we develop a Gaussian approximation for bootstrapping the sample mean. Recall that here, we allow for arbitrary reward distributions  $P^\mu$  (not necessarily from an exponential family), with means  $\mu \in \mathcal{I}$  and corresponding variances  $(\sigma^\mu)^2$ , where  $\mathcal{I} \subset \mathbb{R}$  is an open interval. The only requirement on the  $P^\mu$  is that the condition in (130) is satisfied. As before, for each  $\mu \in \mathcal{I}$ , let  $X^\mu, X_i^\mu \stackrel{\text{iid}}{\sim} P^\mu$ . We use  $\hat{m}_n^\mu$  and  $(\hat{\sigma}_n^\mu)^2$  to denote the sample mean and variance computed using  $n$  samples  $X_1^\mu, \dots, X_n^\mu$ . Also, we use  $\hat{m}_n^{*\mu}$  to denote a bootstrap of the sample mean  $\hat{m}_n^\mu$  computed using  $n$  re-samples with replacement. Proposition 5 holds almost surely and uniformly over data-generating distributions  $P^\mu$ ,  $\mu \in \mathcal{I}$ . Recall that a precise description of this mode of convergence was given in Remark 5 and Definition 5 in Appendix C.

PROPOSITION 5. *Suppose that*

$$\lim_{y \rightarrow \infty} \sup_{\mu \in \mathcal{I}} \mathbb{E}[(X^\mu)^2 \mathbb{I}((X^\mu)^2 > y)] = 0. \quad (130)$$

*Then,*

$$\lim_{n \rightarrow \infty} \sup_{x \in \mathbb{R}} \left| \mathbb{P}(\sqrt{n}(\hat{m}_n^{*\mu} - \hat{m}_n^\mu) \leq x \mid X_1^\mu, \dots, X_n^\mu) - \Phi\left(\frac{x}{\sigma^\mu}\right) \right| = 0, \quad (131)$$

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

*Proof of Proposition 5.* We check the conditions to be able to apply Proposition 1.3.1 part (ii) in Politis et al. (1999). First, because the class of functions  $\{\mathbb{I}(\cdot \leq x), x \in \mathbb{R}\}$  is a VC class and is uniformly bounded, by Theorem 2.8.1 of van der Vaart and Wellner (1996), it is a Glivenko-Cantelli class of functions, uniformly in  $P^\mu$ ,  $\mu \in \mathcal{I}$ . Also, from (130) and Lemma 5, we have  $\hat{m}_n^\mu \rightarrow \mu$  and  $\hat{\sigma}_n^\mu \rightarrow \sigma^\mu$  as  $n \rightarrow \infty$ , almost surely, uniformly in  $\mu$ . The desired result (131) then follows.  $\square$

## Appendix D: Weak Convergence Technical Lemmas

LEMMA 7 (**Generalized Continuous Mapping Theorem**). *Let  $f$  and  $f^n$ ,  $n \geq 1$ , be measurable functions that map from the metric space  $(\mathcal{S}_1, r_1)$  to the separable metric space  $(\mathcal{S}_2, r_2)$ . Let  $E$  be the set of  $x \in \mathcal{S}_1$  such that  $f^n(x^n) \rightarrow f(x)$  fails for some sequence  $x^n$ ,  $n \geq 1$ , with  $x^n \rightarrow x$  in  $\mathcal{S}_1$ . If  $\xi^n \Rightarrow \xi$  in  $(\mathcal{S}_1, r_1)$  and  $P(\xi \in E) = 0$ , then  $f^n(\xi^n) \Rightarrow g(\xi)$  in  $(\mathcal{S}_2, r_2)$ . (See Theorem 3.4.4 of Whitt (2002).)*

LEMMA 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 and only if each  $\xi_j^n$  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  $\xi^n$  in  $D[0, \infty)$  adapted to filtrations  $(\mathcal{F}_t^n, t \geq 0)$  is tight if, for each  $T > 0$ ,*

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

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

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

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

$$\lim_{\delta \downarrow 0} \limsup_{n \rightarrow \infty} \mathbb{E} [A_\delta^n(T)] = 0. \quad (\text{T3})$$

(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  $n \rightarrow \infty$ .*

*There exists a symmetric positive-definite matrix  $\Sigma$  such that*

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

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

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

*Then,*

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

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

## 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.
- Politis D, Romano J, Wolf M (1999) *Subsampling* (Springer).
- Revuz D, Yor M (1999) *Continuous Martingales and Brownian Motion* (Springer).
- Russo D, Van Roy B (2014) Learning to Optimize via Posterior Sampling. *Mathematics of Operations Research* 39(4):1221–1243.
- Russo D, Van Roy B (2016) An Information-Theoretic Analysis of Thompson Sampling. *Journal of Machine Learning Research* 17(68):1–30.
- Russo D, Van Roy B, Kazerouni A, Osband I, Wen Z (2019) *A Tutorial on Thompson Sampling* (Foundations and Trends in Machine Learning).
- Stroock D, Varadhan S (1979) *Multidimensional Diffusion Processes* (Springer).
- Tang L, Jiang Y, Li L, Zeng C, Li T (2015) Personalized Recommendation via Parameter-free Contextual Bandits. *International ACM SIGIR Conference on Research and Development in Information Retrieval* .

- Thompson W (1933) On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika* 25(3):285–294.
- van der Vaart A, Wellner J (1996) *Weak Convergence and Empirical Processes* (Springer).
- Vaswani S, Kveton B, Wen Z, Rao A, Schmidt M, Abbasi-Yadkori Y (2018) New Insights into Bootstrapping for Bandits. *arXiv:1805.09793* .
- Wager S, Xu K (2021) Diffusion Asymptotics for Sequential Experiments. URL <https://arxiv.org/abs/2101.09855v1>.
- Whitt W (2002) *Stochastic-Process Limits* (Springer).
- Whitt W (2007) Proofs of the Martingale FCLT. *Probability Surveys* 4:268–302.