Diffusion Approximations for Thompson Sampling

Lin Fan

Department of Management Science and Engineering, Stanford University, Stanford, CA 94305, linfan@stanford.edu

Peter W. Glynn

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

We study the behavior of Thompson sampling from the perspective of weak convergence. In the regime where the gaps between arm means scale as $1/\sqrt{n}$ with the time horizon n, we show that the dynamics of Thompson sampling evolve according to discrete versions of SDEs and stochastic ODEs. As $n \to \infty$, we show that the dynamics converge weakly to solutions of the corresponding SDEs and stochastic ODEs. Our weak convergence theory, which covers both multi-armed and linear bandit settings, is developed from first principles using the Continuous Mapping Theorem and can be directly adapted to analyze other sampling-based bandit algorithms, for example, algorithms using the bootstrap for exploration. We also establish an invariance principle for multi-armed bandits with gaps scaling as $1/\sqrt{n}$—for Thompson sampling and related algorithms involving posterior approximation or the bootstrap, the weak diffusion limit is in general the same regardless of the specifics of the reward distributions or the choice of prior. In particular, as suggested by the classical Bernstein-von Mises normal approximation for posterior distributions, the weak diffusion limits generally coincide with the limit for normally-distributed rewards and priors.

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

History: Manuscript version – August 31, 2023

Contents

1. Introduction

The multi-armed bandit (MAB) problem has developed into an extremely fruitful area of both research and practice in the past two decades. Modern applications are now numerous and span diverse areas ranging from personalized online advertising and news article recommendation (Li et al. 2010, Chapelle and Li 2011), to dynamic pricing and portfolio management (Shen et al. 2015, Ferreira et al. 2018, Misra et al. 2019), to mobile health and personalized medicine (Tewari and Murphy 2017, Bastani and Bayati 2020), and the list is constantly growing. Along with the widespread deployment of bandit algorithms, there has recently been a dramatic increase in theoretical bandit research, which has almost exclusively focused on optimizing finite-horizon regret bounds, either in expectation or with high probability. While such regret bounds, especially ones in expectation, are accompanied by strong (asymptotic) optimality guarantees (Lai and Robbins 1985, Burnetas and Katehakis 1996), it is well known that practical performance often deviates from what is suggested by the regret bounds. For instance, although regret bounds for Thompson sampling (originally due to Thompson (1933), popularized by Chapelle and Li (2011), and then analyzed by Agrawal and Goyal (2012) and Kaufmann et al. (2012)) are generally somewhat worse compared to those for upper-confidence bound (UCB) algorithms (originally due to Lai and Robbins (1985) and Agrawal (1995), and then further developed and popularized by Auer et al. (2002)), it has been empirically observed that Thompson sampling often significantly outperforms UCB algorithms on many adaptive decision-making tasks (Chapelle and Li 2011, Scott 2010); see also Russo et al. (2019) for a recent overview of Thompson sampling.

Moreover, regret bounds collapse all of the information in the regret distribution into a single performance measure, thus offering a reduced view of performance. In settings where bandit algorithms are deployed with only a limited number of independent runs (so that the law of large numbers does not “kick in”), the expected regret may not be the best performance measure to optimize for. One may instead be interested in controlling some other attribute of the regret distribution such as a measure of overall spread or tail decay. However, even for the most fundamental models such as stochastic MABs, it is difficult to accurately characterize such attributes of the regret distribution. Indeed, work in such directions has been very sparse in the literature, being essentially limited to that of Audibert et al. (2009) and Salomon and Audibert (2011).

As a first step towards providing distributional characterizations of the regret of bandit algorithms, we study the weak convergence (i.e., convergence in distribution) of the dynamics of Thompson sampling to diffusion processes, specifically, to solutions of stochastic differential equations (SDEs) and stochastic ordinary differential equations (ODEs). As detailed in our derivations, for a time horizon of n, when the gap between the arm means scales as $\Delta/\sqrt{n}$ for some fixed Δ, the dynamics of Thompson sampling are governed by equations that are discrete versions of SDEs and stochastic ODEs, which converge weakly to their continuous versions as $n \to \infty$. (Recently, Kuang and Wager (Kuang and Wager 2021) independently proposed this diffusion scaling regime and developed similar weak convergence results for MABs via a different proof approach. See the detailed discussion in Section 1.2 below.) This $1/\sqrt{n}$ scaling is fundamental in the theory of statistical efficiency (see van der Vaart (1998) chapters 6-9 and 25, Le Cam and Yang (2000) or Bickel et al. (1998)). Indeed, the intuition is that given enough data, any (fixed) model parameter can be learned with unlimited precision, so a better way of measuring the performance of an algorithm is to make the learning task more challenging simultaneously as the time horizon increases. Directly related to all of this is the minimax (also called problem-independent or worst-case) bandit setting. In such settings, no gap assumptions are made, but information-theoretic lower bounds on expected regret generally involve the construction of alternative bandit environments with $1/\sqrt{n}$-scale gaps (see, for instance, chapters 15 and 24 of (Lattimore and Szepesvári 2020)). Thus, our weak convergence theory allows for the characterization of the regret distribution under minimax gap regimes.

1.1. Contributions

One of our main contributions is the development of weak convergence theory to approximate the dynamics, and in particular the regret, of Thompson sampling by the solutions of SDEs and stochastic ODEs (see Section 2 and Theorems 1, 2 and 3), which (along with the work of Kuang and Wager (Kuang and Wager 2021)) is new to the literature. We also extend the weak convergence theory to linear bandits with both finite and infinite action sets, where the action sets are allowed to be stochastic/time-varying (see Section 5 and Theorems 4 and 5). On the technical side, we provide a transparent theory that clarifies from first principles using the Continuous Mapping

Theorem exactly why SDE and stochastic ODE weak limits should be obtained, and the ideas and tools that we discuss are also potentially of interest for analyzing other stochastic systems. Although our theory focuses on Thompson sampling, it can be readily adapted to yield diffusion approximations for other bandit algorithms.

One of our particularly striking results is an invariance principle for Thompson sampling in the diffusion regime (see Section 6.1 and Theorem 6). We show that for MABs, the weak diffusion limits for Thompson sampling with general reward distributions are all the same, and moreover the choice of prior does not matter as long as it allocates positive density to neighborhoods of the actual arm means. This result is intuitive in light of the classical Bernstein-von Mises Theorem, which establishes the asymptotic normality of posterior distributions in great generality, along with the fact that (reasonably specified) priors are eventually overwhelmed by data. We develop a slightly stronger locally-uniform almost sure version of the Bernstein-von Mises Theorem (see Proposition 2), which allows us to conclude that the weak diffusion limits for Thompson sampling with general reward distributions and priors coincide with the limit for Thompson sampling with normally-distributed rewards and priors. This result also indicates that in this $1/\sqrt{n}$-scale gap regime, sampling from a Laplace approximation of a posterior results in the same Thompson sampling behavior as sampling from an exact posterior. Furthermore, we develop weak diffusion limits for algorithms that use bootstrap samples for exploration (see Section 6.3 and Theorem 8), which are valid for general reward distributions (without the possibility of mis-specification) and also coincide with the limit for Thompson sampling with normally-distributed rewards and priors (due to the general asymptotic normality associated with bootstrapping the mean of a distribution). Altogether, these results highlight the central role that normality plays in determining the behavior of sampling-based bandit algorithms in minimax gap regimes.

1.2. Related Work

In the process of completing our paper, we became aware of the independent work of Kuang and Wager (WX) Kuang and Wager (2021), which was posted on arXiv prior to our manuscript. Here we discuss in detail the similarities and differences between their work and ours. The focus of WX is very broad—they provide a general framework within which the dynamics of bandit algorithms can be studied under diffusion scaling. Central to their framework is the concept of a *sampling function*, which encodes the specific characteristics of an algorithm, and at any time specifies the probabilities of playing different arms. WX show that the dynamics of any algorithm within their framework converge weakly to solutions of SDEs and stochastic ODEs involving the corresponding sampling function, with Thompson sampling being a case of special interest. On the other hand, our focus is on Thompson sampling and related algorithms. The main overlap between their work and ours is that both obtain similar SDE and stochastic ODE approximations for the dynamics of Thompson sampling in the MAB setting assuming $1/\sqrt{n}$-scale gaps and normally-distributed rewards (cf. Theorems 1 and 2; Theorems 7-10 in Kuang and Wager (2021)). However, our approach to developing the weak convergence theory is different from theirs. Furthermore, we consider extensions that are unrelated to theirs. The differences can be summarized as follows.

1) WX represent sequential algorithms as Markov chains, and use the martingale framework of Stroock and Varadhan (Stroock and Varadhan 1979) to show weak convergence of the Markov chains to diffusion processes by establishing the corresponding convergence of infinitesimal generators. On the other hand, we use representations in terms of discrete versions of SDEs and stochastic ODEs, and we argue directly via the Continuous Mapping Theorem that the discrete systems converge weakly to their continuous counterparts.

2) WX go beyond the general weak convergence theory by analyzing the behavior of one-armed and two-armed Thompson sampling in more depth, for instance, by examining the effect of different prior variance scalings, the effect of increasing the diffusion-scale gap between the mean of the known and unknown arm, as well as the evolution of sampling probabilities over time. On the other hand, our extensions focus on different directions, such as the generalization from MABs to linear bandits, as well as the development of a general invariance principle for Thompson sampling and related algorithms, for instance, involving posterior approximation or bootstrap-based exploration. We also study the effects of model mis-specification, variance estimation and batched updates on algorithm behavior.

Recently, Kalvit and Zeevi (2021) has also studied the behavior of the well-known UCB1 algorithm (from Auer et al. (2002)) in worst-case gap regimes. When the gaps between arm means scale as $\sqrt{\log(n)/n}$, they obtain weak diffusion limits for UCB1. Additionally, they provide sharp distinctions between the behavior of Thompson sampling and UCB algorithms when the gap sizes are zero or near zero.

Also related to our work, Araman and Caldentey (2022) consider a sequential binary testing environment with experiments that arrive according to an exogenous Poisson process. They obtain a diffusion limit as the intensity of arrivals increases and the informativeness of experiments decreases. They then obtain a closed-form solution for optimal experimentation and stopping for the diffusion limit, which provides nice insights and heuristics for the pre-limit problem.

1.3. Notation

For functions $F : \mathbb{R}^d \to \mathbb{R}^d$ and $G : \mathbb{R} \to \mathbb{R}^d$, with component functions $F = (F_1, \ldots, F_d)$ and $G = (G_1, \ldots, G_d)$, we denote their component-wise function composition by $F \circ G(x) =$

$(F_1(G_1(x)),\ldots,F_d(G_d(x)))$, and $G(x)=(G_1(x),\ldots,G_d(x))$, for $x\in\mathbb{R}$. We use $D^d[0,1]$ to denote $\mathbb{R}^d$-valued Skorohod space, the space of functions mapping $[0,1]\to\mathbb{R}^d$, that are continuous from the right and have limits from the left. For a function f with domain in $\mathbb{R}$, we write $f(x-)$ to denote the limit from the left at x.

2. Diffusion Approximations for Multi-armed Bandits

We sketch the development of diffusion approximations for MABs, eventually arriving at SDE and stochastic ODE approximations to the discrete system evolving according to the Thompson sampling algorithm. For concreteness and simplicity, we focus here on the two-armed setting with normal arm rewards, but our approach can be directly generalized to accommodate more than two arms, which we carry out in Sections 3 and 4. Our results can also be easily generalized to other reward distributions.

For a time horizon n, we consider an MAB model, where at time $i=1,\ldots,n$, the reward for playing arm k, $k=1,2$, is:

$$X_k(i)\overset{\text{iid}}{\sim}N(\mu_k^n,\sigma_k^2). \tag{1}$$

To obtain a diffusion approximation, without loss of generality, we assume for arm k that $\mu_k^n=\frac{\mu_k}{\sqrt{n}}$ for some fixed μ_k, and that $\mu_1>\mu_2$. We assume that the variances $\sigma_k^2>0$ are known. (We can also allow the variance of each arm k to change in the asymptotics as $n\to\infty$, for example: $(\sigma_k^n)^2\to\sigma_k^2$ for some $\sigma_k^2>0$.)

For the Thompson sampling algorithm, we choose some $b^2>0$ and put an independent $N(0,(b^2n)^{-1})$ prior on each μ_k^n. As we will see, the n^{-1} scaling of the prior variance is convenient for developing diffusion approximations, as the resulting SDEs will have unique solutions due to the drift and dispersion functions possessing Lipschitz-continuity. But this is by no means the only possible choice of scaling for the prior variance. Through a more careful treatment of the initial phase of the system evolution, it should be interesting to study other choices of scaling for the prior variance that lead to analytically tractable diffusions, and see the effect of the choice on system behavior.

Remark 1. We caution that using such $(b^2n)^{-1}$ scaling of the prior variance can result in catastrophically poor performance if the prior means are set inappropriately. For instance, suppose that the prior for μ_k^n, $k=1,2$, is $N(\mu_k',(b^2n)^{-1})$. With $X_k(1),\ldots,X_k(m_k)$ sampled according to (1), the posterior means have the form:

$$\frac{1}{1+b^2\sigma_k^2}\left(b^2\sigma_k^2\mu_k'+\frac{1}{n}\sum_{i=1}^{m_k}X_k(i)\right),$$

which suggests that if we set $\mu_2' - \mu_1' > 0$ to be large enough, then (incorrectly) the posterior mean estimate for μ_2^n will, with high probability, always be greater than that of μ_1^n when $m_k \leq n$. We will always set the prior means to be zero, thus avoiding this issue. However, Kuang and Wager (2021) show that for one-armed Thompson sampling, even with a prior mean of zero, using a n^{-1} scaling of the prior variance can result in similarly undesirable performance as $\mu_1 - \mu_2$ becomes large (see their Theorem 9).

2.1. Derivation of the SDE Approximation

We first derive the discrete approximations for the SDE. For a time horizon n, consider the setup where for each arm $k = 1, 2$ and at each time $i = 1, \ldots, n$, there is a reward $X_k(i) \overset{\text{iid}}{\sim} N(\mu_k^n, \sigma_k^2)$ that is exogenously generated. The algorithm decides which arm to play, which is reflected by the status of the indicator variables $I_k(i)$, equal to either 0 or 1, reflecting the decision to not play or play, respectively. The algorithm then receives as feedback the possibly censored rewards $I_k(i)X_k(i)$. For a time horizon n, we use the filtration

$$\mathcal{G}_j^n = \sigma\left(I_k(i), I_k(i)X_k(i) : \ k = 1, 2, \ 1 \leq i \leq j\right) \tag{2}$$

to capture the history of arm plays and rewards accumulated up to and including time j. Denote $t_j = j/n$, $0 \leq j \leq n$. After rescaling (in accordance with typical scalings for diffusion approximations), the dynamics of Thompson sampling are completely captured by the evolution of two processes: $R^n = (R_1^n, R_2^n)$ and $Y^n = (Y_1^n, Y_2^n)$, defined via

$$R_k^n(t_j) = \frac{1}{n}\sum_{i=1}^{j} I_k(i) \tag{3}$$

$$Y_k^n(t_j) = \frac{1}{\sqrt{n}}\sum_{i=1}^{j} I_k(i)\frac{X_k(i) - \mu_k^n}{\sigma_k}. \tag{4}$$

To see this, note that, at time $j+1$, having collected history $\mathcal{G}_j^n$, Thompson sampling draws from the posterior distributions:

$$\widetilde{\mu}_k^n(j+1) \sim N\left(\frac{\sum_{i=1}^{j} I_k(i)X_k(i)}{(R_k^n(t_j) + b^2\sigma_k^2)n}, \frac{\sigma_k^2}{(R_k^n(t_j) + b^2\sigma_k^2)n}\right). \tag{5}$$

So the probability of playing arm 1 can be expressed as:

$$\begin{aligned}
&\mathbb{P}\left(\widetilde{\mu}_1^n(j+1) > \widetilde{\mu}_2^n(j+1) \,\middle|\, \mathcal{G}_j^n\right) && (6)\\
&= \mathbb{P}\Bigg(N_1\left(\frac{Y_1^n(t_j)\sigma_1 + R_1^n(t_j)\mu_1}{R_1^n(t_j) + b^2\sigma_1^2}, \frac{\sigma_1^2}{R_1^n(t_j) + b^2\sigma_1^2}\right) > \\
&\qquad N_2\left(\frac{Y_2^n(t_j)\sigma_2 + R_2^n(t_j)\mu_2}{R_2^n(t_j) + b^2\sigma_2^2}, \frac{\sigma_2^2}{R_2^n(t_j) + b^2\sigma_2^2}\right) \,\Bigg|\, \mathcal{G}_j^n\Bigg) && (7)\\
&= p_1(R^n(t_j), Y^n(t_j)), && (8)
\end{aligned}$$

where the N_k are independent normal random variables with their specified means and variances, and we define

$$p_1(u,v) = \Phi\left(\frac{\frac{v_1\sigma_1+u_1\mu_1}{u_1+b^2\sigma_1^2} - \frac{v_2\sigma_2+u_2\mu_2}{u_2+b^2\sigma_2^2}}{\sqrt{\frac{\sigma_1^2}{u_1+b^2\sigma_1^2}+\frac{\sigma_2^2}{u_2+b^2\sigma_2^2}}}\right) \tag{9}$$

$$p_2(u,v) = 1 - p_1(u,v), \tag{10}$$

for any $u = (u_1,u_2) \in [0,1]^2$ and $v = (v_1,v_2) \in \mathbb{R}^2$, with Φ denoting the standard normal CDF. So at time $j+1$, the probability of playing arm k, $k = 1,2$, is $p_k(R^n(t_j), Y^n(t_j))$. Thus, it is sufficient to keep track of R^n and Y^n.

We can now re-express (3)-(4) as

$$R_k^n(t_j) = \frac{1}{n}\sum_{i=1}^{j} p_k(R^n(t_{i-1}), Y^n(t_{i-1})) + M_k^n(t_j) \tag{11}$$

$$Y_k^n(t_j) = \sum_{i=1}^{j} \sqrt{p_k(R^n(t_{i-1}), Y^n(t_{i-1}))}\,(B_k^n(t_i) - B_k^n(t_{i-1})) \tag{12}$$

$$R_k^n(0) = Y_k^n(0) = 0, \qquad k = 1,2, \tag{13}$$

where $M^n = (M_1^n, M_2^n)$ and $B^n = (B_1^n, B_2^n)$ are defined via

$$M_k^n(t_j) = \frac{1}{n}\sum_{i=1}^{j} (I_k(i) - p_k(R^n(t_{i-1}), Y^n(t_{i-1}))) \tag{14}$$

$$B_k^n(t_j) = \frac{1}{\sqrt{n}}\sum_{i=1}^{j} \frac{I_k(i)(X_k(i) - \mu_k^n)}{\sqrt{p_k(R^n(t_{i-1}), Y^n(t_{i-1}))}\cdot\sigma_k}, \tag{15}$$

and $(I_k(i) : 1 \leq k \leq 2)$ is a multinomial random variable with a single trial and success probabilities $p_k(R^n(t_{i-1}), Y^n(t_{i-1}))$. We continuously interpolate the joint processes (R^n, Y^n, B^n, M^n) defined in (11)-(15) to be piecewise constant.

As $n \to \infty$, we show that M^n and B^n converge weakly to the $D^2[0,1]$ zero process and standard Brownian motion on $\mathbb{R}^2$, respectively. Thus, we expect (11)-(13) to be a discrete approximation to the SDE:

$$R_k(t) = \int_0^t p_k(R(s), Y(s))ds \tag{16}$$

$$Y_k(t) = \int_0^t \sqrt{p_k(R(s), Y(s))}dB_k(s) \tag{17}$$

$$R_k(0) = Y_k(0) = 0, \qquad k = 1,2, \tag{18}$$

where $B = (B_1, B_2)$ is a standard Brownian motion on $\mathbb{R}^2$. See Theorem 1 in Section 3 for a rigorous version of the above derivation for $K \geq 2$ arms.

2.2. Derivation of the Stochastic ODE Approximation

To obtain a stochastic ODE characterization, we work with a reward generation process that is equivalent in distribution to the one considered in the derivation of the SDE approximation above. Instead of considering an exogenously generated reward for each arm in each time period, here we consider the setup where an exogenous reward for an arm is generated only when that arm is played. So for arm k, at time j, if $I_k(j)=1$ (the algorithm decides to play arm k), then having collected $m_k(j-1)=\sum_{i=1}^{j-1} I_k(i)$ previous rewards for arm k, the algorithm receives as feedback the reward $X_k(m_k(j-1)+1) \overset{\text{iid}}{\sim} N(\mu_k^n, \sigma_k^2)$. For a time horizon n, we use the filtration

$$\mathcal{H}_j^n = \sigma\left(I_k(i), X_k(l): \; k=1,2, \; 1 \le i \le j, \; 1 \le l \le m_k(j)\right) \tag{19}$$

to capture the history of arm plays and rewards accumulated up to and including time j. Again, denote $t_j = j/n$, $0 \le j \le n$. And after rescaling (in accordance with typical scalings for diffusion approximations), the dynamics of Thompson sampling are completely captured by the evolution of two processes: $R^n = (R_1^n, R_2^n)$ and $Z^n \circ R^n = (Z_1^n(R_1^n), Z_2^n(R_2^n))$, defined via

$$R_k^n(t_j) = \frac{1}{n}\sum_{i=1}^{j} I_k(i) \tag{20}$$

$$Z_k^n(R_k^n(t_j)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n \cdot R_k^n(t_j)} \frac{X_k(i)-\mu_k^n}{\sigma_k}, \tag{21}$$

where $Z^n = (Z_1^n, Z_2^n)$ is defined via

$$Z_k^n(t_j) = \frac{1}{\sqrt{n}} \sum_{i=1}^{j} \frac{X_k(i)-\mu_k^n}{\sigma_k}. \tag{22}$$

To see this, note that the distributions of $Z_k^n(R_k^n(t_j))$ and $Y_k^n(t_j)$ (as defined in (4)) are the same, since the rewards $X_k(i)$ are iid and exogenous in both versions of the reward generation process. Then at time $j+1$, having collected history $\mathcal{H}_j^n$, Thompson sampling draws from the posterior distributions:

$$\widetilde{\mu}_k^n(j+1) \sim N\left(\frac{\sum_{i=1}^{n \cdot R_k^n(t_j)} X_k(i)}{(R_k^n(t_j)+b^2\sigma_k^2)n}, \frac{\sigma_k^2}{(R_k^n(t_j)+b^2\sigma_k^2)n}\right). \tag{23}$$

So the probability of playing arm 1 can be expressed as:

$$\begin{aligned}
&\mathbb{P}\left(\widetilde{\mu}_1^n(j+1) > \widetilde{\mu}_2^n(j+1) \,\middle|\, \mathcal{H}_j^n\right) && (24)\\
&= \mathbb{P}\Bigg(N_1\left(\frac{Z_1^n(R_1^n(t_j))\sigma_1 + R_1^n(t_j)\mu_1}{R_1^n(t_j)+b^2\sigma_1^2}, \frac{\sigma_1^2}{R_1^n(t_j)+b^2\sigma_1^2}\right) > \\
&\qquad N_2\left(\frac{Z_2^n(R_2^n(t_j))\sigma_2 + R_2^n(t_j)\mu_2}{R_2^n(t_j)+b^2\sigma_2^2}, \frac{\sigma_2^2}{R_2^n(t_j)+b^2\sigma_2^2}\right) \,\Bigg|\, \mathcal{H}_j^n \Bigg) && (25)\\
&= p_1(R^n(t_j), Z^n \circ R^n(t_j)), && (26)
\end{aligned}$$

where the N_k are independent normal random variables with their specified means and variances, and the p_k probabilities are as defined in (9)-(10). So at time $j+1$, the probability of playing arm k, $k=1,2$, is $p_k(R^n(t_j), Z^n \circ R^n(t_j))$. Thus, it is sufficient to keep track of R^n and $Z^n \circ R^n$.

We can now re-express (20) as

$$R_k^n(t_j) = \frac{1}{n}\sum_{i=1}^{j} p_k(R^n(t_{i-1}), Z^n \circ R^n(t_{i-1})) + M_k^n(t_j) \tag{27}$$

$$R_k^n(0) = 0, \qquad k = 1,2, \tag{28}$$

where $M^n = (M_1^n, M_2^n)$ is defined via

$$M_k^n(t_j) = \frac{1}{n}\sum_{i=1}^{j} \left(I_k(i) - p_k(R^n(t_{i-1}), Z^n \circ R^n(t_{i-1}))\right), \tag{29}$$

and $(I_k(i) : 1 \le k \le 2)$ is a multinomial random variable with a single trial and success probabilities $p_k(R^n(t_{i-1}), Z^n \circ R^n(t_{i-1}))$. We continuously interpolate the joint processes $(R^n, Z^n \circ R^n, Z^n, M^n)$ defined in (27)-(29) and (21)-(22) to be piecewise constant. (In this stochastic ODE derivation, R^n and M^n are the same as in the corresponding definitions in (3), (11) and (14) in the SDE derivation of Section 2.1, except we are working with a different but distributionally-equivalent reward generation process.)

As $n \to \infty$, we show that M^n and Z^n converge weakly to the $D^2[0,1]$ zero process and standard Brownian motion on $\mathbb{R}^2$, respectively. Thus, we expect (27)-(28) to be a discrete approximation to the stochastic ODE:

$$R_k(t) = \int_0^t p_k(R(s), B \circ R(s)) ds \tag{30}$$

$$R_k(0) = 0, \qquad k = 1,2, \tag{31}$$

where $B = (B_1, B_2)$ is a standard Brownian motion on $\mathbb{R}^2$. See Theorems 2 and 3 in Section 4 for a rigorous version of the above derivation for $K \ge 2$ arms.

2.3. Extension to $K > 2$ Arms

We conclude this section by mentioning that the above theory can be directly generalized to $K > 2$ arms. For a time horizon n, at time $j+1$, conditional on the information $\mathcal{G}_j^n$ collected up to and including time j, let $\widetilde{\mu}_k^n(j+1)$ denote the sample from the posterior distribution for arm $k = 1, \ldots, K$. By straightforward derivations, it can be shown that we can define Lipschitz-continuous functions p_k, $k = 1, \ldots, K$ (analogous to the two-armed case in (9)-(10)) such that for the SDE characterization, we have

$$p_k(R^n(t_j), Y^n(t_j)) = \mathbb{P}\left(\widetilde{\mu}_k^n(j+1) > \max_{k' \neq k} \widetilde{\mu}_{k'}^n(j+1) \mid \mathcal{G}_j^n\right), \tag{32}$$

which is the probability of playing arm k at time $j+1$. (Of course, R^n and Y^n are now $\mathbb{R}^K$-valued.) And similarly, for the stochastic ODE characterization, we have the same p_k functions, and the same probability is expressed as

$$p_k(R^n(t_j), Z^n \circ R^n(t_j)). \tag{33}$$

3. Characterization via Stochastic Differential Equations

We first discuss a (random) mapping due to Kurtz and Protter (1991), which allows any function in $D[0,1]$ to be approximated by a (random) step function with arbitrarily good accuracy. This mapping idea makes it simple to transition from discrete versions of Itô integrals to the Itô integrals themselves in the continuous weak limit. The mapping is defined in Definition 1, some basic properties of it are given in Remark 2, and its key properties and relevance to Itô integrals are given in Lemmas 1-2. The main results of this section are Theorems 1, 4 and 5, which establish that the dynamics of Thompson sampling in both the MAB and linear bandit settings (as derived in Sections 2.1, 2.3, 5.1 and 5.2) converge weakly to solutions of SDEs.

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

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

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

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

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

Remark 2. We mention here a few properties of the random mapping χ_ϵ. For justification, see the discussion preceding Lemma 6.1 of Kurtz and Protter (1991). First of all, for any $z \in D[0,1]$, with χ_ϵ as defined in (D1)-(D2), we have $\sup_{0\leq t\leq 1} |z(t) - \chi_\epsilon(z)(t)| \leq \epsilon$, so χ_ϵ yields an ϵ-uniform approximation. Note that the purpose of the uniform random variables U_j in defining the random times τ_j is to avoid pathological issues concerning, for instance, the locations of jump discontinuities of the functions $z^n, z \in D[0,1]$. (In the settings that we consider, it is not absolutely crucial that we use such randomization to avoid pathological issues. We can define deterministic step function approximations by considering sequences of partition refinements. However, this randomization idea introduced in Kurtz and Protter (1991) is both simple and adaptable to other settings, so we adopt it here.) In particular, since each U_j avoids any (fixed) countable set with probability one, for each τ_j, either z will be continuous at τ_j or we will have

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

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

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

REMARK 3. For a step function $z_1 \in D[a,b]$, with jump points $s_1 < \cdots < s_m$ (and $s_0 = a$, $s_{m+1} = b$), and a continuous function z_2 on $[a,b]$, we will always use the following definition of integration:

$$\int_a^b z_1(s)dz_2(s) = \sum_{j=0}^{m} z_1(s_j)\left(z_2(s_{j+1}) - z_2(s_j)\right). \tag{34}$$

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

$$S_\epsilon(z_1, z_2)(t) = \int_0^t \chi_\epsilon(z_1(s))dz_2(s), \tag{35}$$

where $(z_1, z_2) \in D^2[0,1]$. (Note that the integral is defined as in (34), since $\chi_\epsilon(z_1)$ is always a step function.) Then, almost surely for each realization of χ_ϵ, we have

$$S_\epsilon(\xi_1^n, \xi_2^n) \to S_\epsilon(\xi_1, \xi_2)$$

with respect to the Skorohod metric.

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

$\chi_\epsilon(\xi_1^n)$ (at $\tau_1^n,\ldots,\tau_M^n$) that are at least $\epsilon/2$ in magnitude. (See Chapter of Billingsley (1999).) We denote $\tau_0^n=\tau_0=0$ and $\tau_{M+1}^n=\tau_{M+1}=1$. To conclude the proof, note that

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

as $n\to\infty$, since, as discussed in Remark 2 and Lemma 1, we have $\chi_\epsilon(\xi_1^n(\tau_j^n))\to\chi_\epsilon(\xi_1(\tau_j))$ and $\xi_2^n(\tau_j^n)\to\xi_2(\tau_j)$ for each j. □

Remark 4. In the setting of Theorem 1, we can replace the functions p_k by functions p_k^n (possibly different for each n) in the discrete approximations (11)-(15) from our derivation in Section 2.1 (along with the generalization to $K\ge 2$ arms), as long as for each k, we have $p_k^n\to p_k$ as $n\to\infty$ uniformly on compact subsets of their domain of definition. This allows us to accommodate slight modifications of Thompson sampling. With such a modification, the proof of Theorem 1 would remain unchanged.

Theorem 1. *For the K-armed MAB, the dynamics of Thompson sampling, which are characterized by the processes R^n and Y^n (as defined in (3) and (4), except with K arms), converge weakly in $D^{2K}[0,1]$ as $n\to\infty$ to the unique strong solution of the SDE:*

$$R_k(t)=\int_0^t p_k(R(s),Y(s))ds \tag{36}$$

$$Y_k(t)=\int_0^t \sqrt{p_k(R(s),Y(s))}dB_k(s) \tag{37}$$

$$R_k(0)=Y_k(0)=0,\qquad k=1,\ldots,K, \tag{38}$$

where the B_k are independent standard Brownian motions.

Proof of Theorem 1. We start with the discrete approximation (11)-(15) from our derivation in Section 2.1, but with arms $k=1,\ldots,K$, instead of just arms $k=1,2$. We denote the joint processes via $(R^n,Y^n,B^n,M^n)=(R_k^n,Y_k^n,B_k^n,M_k^n:1\le k\le K)$, and recall that we interpolate them in a piecewise constant fashion, which results in processes in $D^{4K}[0,1]$. All of our weak convergence theory will take place in $D^d[0,1]$, for positive integer d, equipped with the Skorohod metric (see Chapter 3 of Billingsley (1999)), which makes such spaces complete, separable metric spaces.

Our proof strategy is as follows. We will show that for every subsequence of (R^n,Y^n), there is a further subsequence which converges weakly to a limit that is a solution to the SDE. Because the drift and dispersion functions, p_k and $\sqrt{p_k}$, of the SDE (36)-(38) are Lipschitz-continuous and bounded on their domain of definition, Theorem 5.2.9 of Karatzas and Shreve (1998) ensures that the SDE has a unique strong solution. Thus, (R^n, Y^n) must converge weakly to the unique strong solution of the SDE.

By Lemma 3, the joint processes (R^n, Y^n, B^n, M^n) are tight, and thus, Prohorov's Theorem (see Chapters 1 and 3 of Billingsley (1999)) ensures that for each subsequence, there is a further subsequence which converges weakly to some limit process $(R, Y, B, M) = (R_k, Y_k, B_k, M_k : 1 \leq k \leq K)$. From now on, we work with this further subsequence, and for notational simplicity, we still index this further subsequence by n. So, $(R^n, Y^n, B^n, M^n) \Rightarrow (R, Y, B, M)$. Since M^n consists of martingale differences, by a Chebyshev bound, we have $M_k^n(t) \xrightarrow{\mathbb{P}} 0$ for each $k = 1, \ldots, K$ and $t \in (0, 1]$ as $n \to \infty$, and thus, M is exactly equal to the $D^K[0, 1]$ zero process. By Lemma 4, B is a standard Brownian motion on $\mathbb{R}^K$.

Now define the processes $G^n = (G_k^n : 1 \leq k \leq K)$ and $G = (G_k : 1 \leq k \leq K)$, where

$$G_k^n(t) = p_k(R^n(t), Y^n(t)) \tag{39}$$

$$G_k(t) = p_k(R(t), Y(t)). \tag{40}$$

Since the p_k functions are continuous, we have

$$(R^n, Y^n, B^n, M^n, G^n) \Rightarrow (R, Y, B, M, G). \tag{41}$$

Additionally, define the processes $\widetilde{R}^n = (\widetilde{R}_k^n : 1 \leq k \leq K)$ and $\widetilde{R} = (\widetilde{R}_k : 1 \leq k \leq K)$, where

$$\widetilde{R}_k^n(t) = \int_0^t p_k(R^n(s), Y^n(s))ds \tag{42}$$

$$\widetilde{R}_k(t) = \int_0^t p_k(R(s), Y(s))ds. \tag{43}$$

Recall that $t_j = j/n$, and note that

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

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

$$\left| \frac{1}{n}\sum_{i=1}^{\lfloor tn \rfloor} p_k(R^n(t_{i-1}), Y^n(t_{i-1})) - \widetilde{R}_k^n(t) \right| \leq \frac{1}{n},$$

we have

$$\sup_{0 \leq t \leq 1} \left| R_k^n(t) - \widetilde{R}_k^n(t) \right| \xrightarrow{\mathbb{P}} 0. \tag{44}$$

Thus, by the fact that integration is a continuous functional with respect to the Skorohod metric (see Theorem 11.5.1 on page 383 of Whitt (2002)) and the Continuous Mapping Theorem, we have from (41),

$$(R^n, Y^n, B^n, \widetilde{R}^n, G^n) \Rightarrow (R, Y, B, \widetilde{R}, G). \tag{45}$$

For any $\epsilon > 0$, let χ_ϵ be the random mapping defined in (D1)-(D2) (see Remark 2 and Lemma 1 for some basic properties of χ_ϵ). Let $\chi_\epsilon \circ G^n$ and $\chi_\epsilon \circ G$ denote the component-wise applications of χ_ϵ to the vector-valued processes G^n and G. By Lemma 1 and the Continuous Mapping Theorem, we have from (45),

$$(R^n, Y^n, B^n, \widetilde{R}^n, \chi_\epsilon \circ G^n) \Rightarrow (R, Y, B, \widetilde{R}, \chi_\epsilon \circ G). \tag{46}$$

Recall from (12), (15) and (39), that for each k,

$$Y_k^n(t) = \int_0^t \sqrt{G_k^n(s-)} dB_k^n(s), \tag{47}$$

and define the process $\widehat{Y}^n = (\widehat{Y}_k^n : 1 \leq k \leq K)$ by

$$\widehat{Y}_k^n(t) = \int_0^t \chi_\epsilon\left(\sqrt{G_k^n(s-)}\right) dB_k^n(s). \tag{48}$$

By Lemma 2 and the Continuous Mapping Theorem (with the mapping F_ϵ in (35)), we have from (46),

$$(R^n, Y^n, B^n, \widetilde{R}^n, \widehat{Y}^n) \Rightarrow (R, Y, B, \widetilde{R}, \widehat{Y}), \tag{49}$$

where the process $\widehat{Y} = (\widehat{Y}_k : 1 \leq k \leq K)$ is defined by

$$\widehat{Y}_k(t) = \int_0^t \chi_\epsilon\left(\sqrt{G_k(s-)}\right) dB_k(s) \tag{50}$$

with G_k defined by (40). We also define the process $\widetilde{Y} = (\widetilde{Y}_k : 1 \leq k \leq K)$ by

$$\widetilde{Y}_k(t) = \int_0^t \sqrt{G_k(s-)} dB_k(s). \tag{51}$$

Note that both of the processes in (50) and (51) are well defined as Itô integrals, since by Lemma 4, the integrands (with the G_k defined in (40)) are non-anticipative with respect to the Brownian motions B_k. (As defined in (D1)-(D2), χ_ϵ depends on exogenous randomization that is independent of the B_k.) By Lemma 1, because χ_ϵ is an ϵ-uniform approximation, for each k,

$$\mathbb{E}\left[\sup_{0\leq t\leq 1}\left|Y_k^n(t) - \widehat{Y}_k^n(t)\right|\right] \leq \epsilon \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\frac{I_k(i)(X_k(i)-\mu_k^n)^2}{p_k(R^n(t_{i-1}), Y^n(t_{i-1}))\cdot\sigma_k^2}\,\middle|\,\mathcal{G}_{i-1}^n\right]\right]^{1/2} = \epsilon. \tag{52}$$

(Recall the expressions for Y_k^n and $\widehat{Y}_k^n$ in (47)-(48), as well as the definition of G_k^n in (39), of B_k^n in (15), of Y_k^n in (12), and of $\mathcal{G}_j^n$ in (2).) Similarly,

$$\mathbb{E}\left[\sup_{0\le t\le 1}\left|\widehat{Y}_k(t)-\widetilde{Y}_k(t)\right|\right]\le \epsilon\mathbb{E}\left[\langle B_k\rangle_1\right]^{1/2}=\epsilon, \tag{53}$$

where $t\mapsto\langle B_k\rangle_t$ denotes the quadratic variation process for B_k. Putting together (44), (49)-(53) and sending $\epsilon\downarrow 0$, we have

$$(R^n,Y^n,B^n,R^n,Y^n)\Rightarrow(R,Y,B,\widetilde{R},\widetilde{Y}). \tag{54}$$

Recalling the definition of $\widetilde{R}$ in (43) as well as that of $\widetilde{Y}$ in (51) and G_k in (40), we see from (54) that the limit process (R,Y,B) satisfies the SDE:

$$R_k(t)=\int_0^t p_k(R(s),Y(s))ds \tag{55}$$

$$Y_k(t)=\int_0^t \sqrt{p_k(R(s),Y(s))}dB_k(s), \qquad k=1,\ldots,K. \tag{56}$$

(Note that from (55)-(56), it is clear that (R,Y,B) is adapted to the (augmented) filtration $\mathcal{F}_t=\sigma(\mathcal{F}_t^B\cup\mathcal{N})$, where $\mathcal{F}_t^B=\sigma\left(B(s):0\le s\le t\right)$, with $\mathcal{N}$ denoting the collection of all $\mathbb{P}$-null sets.) □

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

Lemma 3. *The processes (R^n,Y^n,B^n,M^n) defined in (11)-(15) (with $k=1,\ldots,K$ instead of $k=1,2$) are tight in $D^{4K}[0,1]$.*

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

$$R_k^n(t_j)=\frac{1}{n}\sum_{i=1}^{j}I_k(i) \tag{57}$$

$$Y_k^n(t_j)=\frac{1}{\sqrt{n}}\sum_{i=1}^{j}I_k(i)\frac{X_k(i)-\mu_k^n}{\sigma_k} \tag{58}$$

$$M_k^n(t_j)=\frac{1}{n}\sum_{i=1}^{j}\left(I_k(i)-p_k(R^n(t_{i-1}),Y^n(t_{i-1}))\right) \tag{59}$$

$$B_k^n(t_j)=\frac{1}{\sqrt{n}}\sum_{i=1}^{j}\frac{I_k(i)(X_k(i)-\mu_k^n)}{\sqrt{p_k(R^n(t_{i-1}),Y^n(t_{i-1}))}\cdot\sigma_k}, \tag{60}$$

where $t_j=j/n$, $0\le j\le n$, and recall that we continuously interpolate them to be piecewise constant in between the t_j. With a slight abuse of notation, let $(\mathcal{G}_t^n:0\le t\le 1)$ denote the continuous, piecewise constant (and right-continuous) interpolation of the discrete-time filtration $(\mathcal{G}_j^n:0\le j\le$

n) defined in (2), and note that (57)-(60) are all adapted to $\mathcal{G}_t^n$. Also note that the processes in (57) are uniformly bounded and increasing, and those in (58)-(60) are square-integrable martingales.

By Lemma 9, to show tightness of the joint processes (R^n, Y^n, B^n, M^n), we just need to show tightness of each component sequence of processes and each pairwise sum of component sequences of processes. We use Lemma 10 to verify tightness in each case. Condition (T1) is easily verified using a submartingale maximal inequality (for example, see page 13 of Karatzas and Shreve (1998)), along with a union bound when dealing with pairwise sums of component processes. Conditions (T2)-(T3) are also easily verified. For each individual component process, we can set $\alpha_n(\delta) = \delta$, and for each pairwise sum of component processes, we can set $\alpha_n(\delta) = 4\delta$ (by bounding via: $(a+b)^2 \leq 2a^2 + 2b^2$), uniformly for all n in each case. □

Lemma 4. *Following Lemma 3, for any subsequence of (R^n, Y^n, B^n, M^n) that converges weakly in $D^{4K}[0,1]$ to some limit process (R, Y, B, M), the component B is a standard Brownian motion on $\mathbb{R}^K$. Moreover, R and Y are non-anticipative with respect to B, i.e., $B(t+u) - B(t)$ is independent of $(R(s), Y(s))$ for $0 \leq s \leq t$ and $u \geq 0$.*

Proof of Lemma 4. We apply the martingale functional central limit theorem stated in Lemma 11. Like in the proof of Lemma 3, with a slight abuse of notation, let $(\mathcal{G}_t^n : 0 \leq t \leq 1)$ denote the continuous, piecewise constant (and right-continuous) interpolation of the discrete-time filtration $(\mathcal{G}_j^n : 0 \leq j \leq n)$ defined in (2), and note that the processes (R^n, Y^n, B^n, M^n), as defined in (11)-(15) (with $k = 1, \ldots, K$ instead of $k = 1, 2$), are adapted to the filtration $\mathcal{G}_t^n$. The condition (M1) is easily verified by noting that for a sequence of independent standard normal random variables $N_1, \ldots, N_n$, we have $\mathbb{E}\left[\max_{1 \leq i \leq n} |N_i|\right] \leq \sqrt{2\log(2n)}$ using a Chernoff bound. The condition (M2) is similarly straightforward to verify by noting that the (martingale difference) increments of B_k^n for different pairs of k are conditionally uncorrelated with respect to the filtration $\mathcal{G}_t^n$. After a straightforward calculation, we see that the C matrix is the identity matrix, so the weak limit process is standard Brownian motion. The non-anticipative fact follows from a straightforward characteristic function argument, since in the pre-limit, R^n and Y^n are non-anticipative with respect to B^n. □

4. Characterization via Stochastic Ordinary Differential Equations

In this section, we provide an alternative representation in terms of solutions of stochastic ODEs for the limiting dynamics of Thompson sampling in both the MAB and linear bandit (with finitely many fixed, non-random covariate vectors) settings, as derived in Sections 2.2 and 5.1 (in particular, Remark 7). Theorem 2 is the main result. While we could have taken a first-principles approach in its proof, we instead take a simpler approach by leveraging Theorems 1 and 4 and the fact that in great generality, continuous (local) martingales, such as Itô integrals, can be represented as time-changed Brownian motion. We go on to develop Theorem 3, which covers the setting where the p_k functions are continuous, but not necessarily Lipschitz-continuous. This situation can arise, for instance, when one uses different scalings for the prior variance.

Remark 5. Similar to Remark 4, in the setting of Theorem 2, we can replace the functions p_k by functions p_k^n (possibly different for each n) in the discrete approximations (27)-(29) from our derivation in Section 2.2 (along with the generalization to $K \geq 2$ arms), as long as for each k, we have $p_k^n \to p_k$ as $n \to \infty$ uniformly on compact subsets of their domain of definition.

Theorem 2. *For the K-armed MAB, the dynamics of Thompson sampling, which are characterized by the processes R^n and $Z^n \circ R^n$ (as defined in (20) and (21), except with K arms), converge weakly in $D^{2K}[0,1]$ as $n \to \infty$ to a solution of the stochastic ODE:*

$$R_k(t) = \int_0^t p_k(R(s), B \circ R(s)) ds \tag{61}$$

$$R_k(0) = 0, \qquad k = 1, \ldots, K, \tag{62}$$

with $Z^n \circ R^n$ converging weakly to $B \circ R$, where B is a standard Brownian motion on $\mathbb{R}^K$.

Proof of Theorem 2. The proofs for the MAB and the linear bandit with finitely many fixed, non-random covariate vectors are the same, so we only develop the proof for the MAB. We work with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ supporting a standard Brownian motion B on $\mathbb{R}^K$, with natural filtration $\mathcal{F}_t^B = \sigma(B(s) : 0 \leq s \leq t)$. We will work with the corresponding augmented filtration $\mathcal{F}_t = \sigma(\mathcal{F}_t^B \cup \mathcal{N})$, where $\mathcal{N}$ is the collection of all $\mathbb{P}$-null sets. (See Chapter 2.7 of Karatzas and Shreve (1998) for details.) By Theorem 1, there exists a solution (R, Y) to the SDE (36)-(38) on this probability space with respect to the standard Brownian motion B. Writing (37) in integral form, because the p_k functions are bounded,

$$Y_k(t) = \int_0^t \sqrt{p_k(R(s), Y(s))} dB_k(s), \qquad k = 1, \ldots, K$$

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

$$\langle Y_k \rangle_t = \int_0^t p_k(R(s), Y(s)) ds, \qquad k = 1, \ldots, K,$$

and for $k \neq k'$, the cross-variation processes $\langle Y_k, Y_{k'} \rangle_t = 0$ since B_k and $B_{k'}$ are independent. Note that integrating (36) (in the Riemann sense) yields $\langle Y_k \rangle_t = R_k(t)$, $k = 1, \ldots, K$, which are continuous and strictly increasing processes since the p_k functions are bounded and strictly positive. Define

$$R_k^{-1}(t) = \inf\{s \geq 0 : R_k(s) \geq t\}, \qquad k = 1, \ldots, K.$$

Now, we recall that in great generality, continuous martingales can be represented as time-changed Brownian motions. In particular, by a theorem due to F.B. Knight (see for instance Proposition

18.8 on page 355 of Kallenberg (2002) or Theorem 1.10 on page 183 of Revuz and Yor (1999)), for $k=1,\dots,K$, we have that $\widetilde{B}_k(t) := Y_k(R_k^{-1}(t))$ are independent standard Brownian motions (at least until time $t = R_k(1)$) with respect to the filtration $\mathcal{F}_t^{\widetilde{B}} = \sigma\left(\widetilde{B}(s) : 0 \le s \le t\right)$. Thus, we have $\widetilde{B}_k(R_k(t)) = Y_k(t)$, and substituting this representation into the SDE (36), we obtain the stochastic ODE:

$$R_k(t) = \int_0^t p_k(R(s), \widetilde{B} \circ R(s))ds, \qquad k = 1,\dots,K. \tag{63}$$

So with respect to the smaller filtration $\mathcal{F}_t^{\widetilde{B}}$, the SDE solution $R(t)$ satisfies the stochastic ODE (63), which coincides with (61). □

Theorem 3. *For the K-armed MAB, consider Thompson sampling evolving with any continuous (not Lipschitz) functions $p_k : [0,1]^K \times \mathbb{R}^K \to [0,1]$. Then, the weak limit points in $D^{2K}[0,1]$ as $n \to \infty$ of the processes R^n and $Z^n \circ R^n$ (as defined in (20) and (21), except with K arms) are solutions of the stochastic ODE:*

$$R_k(t) = \int_0^t p_k(R(s), B \circ R(s))ds \tag{64}$$

$$R_k(0) = 0, \qquad k = 1,\dots,K, \tag{65}$$

with $B \circ R$ being a weak limit of $Z^n \circ R^n$, where B is a standard Brownian motion on $\mathbb{R}^K$.

Proof of Theorem 3. Once again, the proofs for the MAB and the linear bandit with finitely many fixed, non-random covariate vectors are the same, so we only develop the proof for the MAB. We start with the discrete approximation (27)-(29) and (21)-(22) from our derivation in Section 2.2, but with arms $k=1,\dots,K$, instead of just arms $k=1,2$. We denote the joint processes via $(R^n, Z^n, M^n) = (R_k^n, Z_k^n, M_k^n : 1 \le k \le K)$, and recall that we interpolate them in a piecewise constant fashion, which results in processes in $D^{3K}[0,1]$. All of our weak convergence theory will take place in $D^d[0,1]$, for positive integer d, equipped with the Skorohod metric (see Chapter 3 of Billingsley (1999)), which makes such spaces complete, separable metric spaces.

Consider a weakly convergent subsequence of (R^n, Z^n), which we will still index by n for notational simplicity. Then jointly, $(R^n, Z^n, M^n) \Rightarrow (R, Z, M)$, where M is the $D^K[0,1]$ zero process. (Note that since M^n consists of martingale differences, by a Chebyshev bound, we have $M_k^n(t) \xrightarrow{\mathbb{P}} 0$ for each $k=1,\dots,K$ and $t \in (0,1]$ as $n \to \infty$. Also, M^n is a tight sequence of processes using what we showed in Lemma 3.) By Donsker's Theorem (see Chapter 3 of Billingsley (1999)), Z is a standard Brownian motion on $\mathbb{R}^K$.

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

$$(R^n, Z^n, M^n, Z^n \circ R^n) \Rightarrow (R, Z, M, Z \circ R). \tag{66}$$

Now define the processes $G^n = (G^n_k : 1 \le k \le K)$ and $G = (G_k : 1 \le k \le K)$, where

$$G^n_k(t) = p_k(R^n(t), Z^n \circ R^n(t))$$
$$G_k(t) = p_k(R(t), Z \circ R(t)).$$

Since the p_k functions are continuous, we have from (66),

$$(R^n, Z^n, M^n, G^n) \Rightarrow (R, Z, M, G). \tag{67}$$

Additionally, define the processes $\widetilde{R}^n = (\widetilde{R}^n_k : 1 \le k \le K)$ and $\widetilde{R} = (\widetilde{R}_k : 1 \le k \le K)$, where

$$\begin{aligned}\widetilde{R}^n_k(t) &= \int_0^t p_k(R^n(s), Z^n \circ R^n(s))ds \\ \widetilde{R}_k(t) &= \int_0^t p_k(R(s), Z \circ R(s))ds.\end{aligned} \tag{68}$$

Recall that $t_j = j/n$, and note that

$$R^n_k(t) = \frac{1}{n}\sum_{i=1}^{\lfloor tn \rfloor} p_k(R^n(t_{i-1}), Z^n \circ R^n(t_{i-1})) + M^n_k(t).$$

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

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

we have

$$\sup_{0 \le t \le 1} \left| R^n_k(t) - \widetilde{R}^n_k(t) \right| \xrightarrow{\mathbb{P}} 0. \tag{69}$$

Thus, by the fact that integration is a continuous functional with respect to the Skorohod metric (see Theorem 11.5.1 on page 383 of Whitt (2002)) and the Continuous Mapping Theorem, we have from (67),

$$(R^n, Z^n, \widetilde{R}^n) \Rightarrow (R, Z, \widetilde{R}). \tag{70}$$

Together, (69)-(70) yield

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

and recalling the definition of $\widetilde{R}$ in (68), the proof is complete. □

5. Diffusion Approximations for Linear Bandits

The linear bandit is a useful generalization of the MAB to situations where the arm means are no longer estimated separately, but information is shared so that the rewards for one arm provide information about the means of other arms. In this section, we develop weak diffusion limits for Thompson sampling when the set of actions available to the decision-maker in each time period is finite and possibly time-varying/stochastic (Section 5.1), as well as infinite (Section 5.2). (In accordance with much of the literature on linear bandits, here we choose to use the term *action* instead of *arm*.)

The results in this section also indicate that one may derive diffusion approximations for TS in more complex bandit models, going beyond the multi-armed bandit setting, by using different state processes. In the multi-armed setting, the state processes (suitably normalized/standardized) essentially track the evolution of the sample mean reward estimates for each arm over time. In the linear bandit setting, the state processes we define below essentially track the evolution of the ordinary least squares estimate of the unknown linear regression parameter vector over time.

5.1. Finitely Many Actions

We consider a d-dimensional linear bandit model, where at time $i=1,\dots,n$, the reward for playing action k, $k=1,\dots,K$, is:

$$X_k(i) = (\theta^n)^\top A_k(i) + \epsilon_k(i), \qquad \epsilon_k(i) \overset{\text{iid}}{\sim} N(0,1). \tag{71}$$

For each k, the covariate vector $A_k(i)$ is iid exogenously generated from some distribution and is revealed to the decision-maker prior to the decision for time i. The distribution is allowed to be a point mass so that $A_k(i)$ is just a (known) fixed, non-random vector that does not change with i. The parameter vector θ^n is unknown and must be learned. With a time horizon of n, we assume that $\theta^n = \frac{\theta_*}{\sqrt{n}}$, where θ_* is a fixed vector that does not change with n. This scaling assumption allows us to obtain a diffusion limit, just like the $\frac{\mu_1-\mu_2}{\sqrt{n}}$ arm mean gap assumption in Section 2.

Remark 6. In this section, the finite action set is allowed to change in time, but will always have a fixed K number of elements. We refer to them simply as actions $k=1,\dots,K$, and we refer to their associated $A_k(i)$ (which can change with the time index i) as covariate vectors.

Similar to our choice of prior for the MAB model, we put a $N(0,(b^2 n)^{-1} I)$ prior on θ^n, where I is the identity matrix. For linear bandits, the probabilities of playing various actions parallel (5)-(8)

and (23)-(26) for MABs, except that the posterior distribution of θ^n is now obtained through regularized least-squares regression. To describe the dynamics of Thompson sampling in this setting, it suffices to keep track of two processes: $V^n = (V_1^n, \ldots, V_K^n)$ and $S^n = (S_1^n, \ldots, S_K^n)$, defined via

$$V_k^n(t_j) = \frac{1}{n}\sum_{i=1}^{j} I_k(i)A_k(i)A_k(i)^\top \tag{72}$$

$$S_k^n(t_j) = \frac{1}{\sqrt{n}}\sum_{i=1}^{j} I_k(i)A_k(i)\epsilon_k(i), \qquad k = 1, \ldots, K. \tag{73}$$

Conditional on the information collected up to and including time j:

$$\mathcal{G}_j^n = \sigma\left(A_k(i),\, I_k(i)X_k(i) \,:\, k = 1, \ldots, K,\, 1 \le i \le j\right), \tag{74}$$

we sample a vector from the posterior distribution of θ^n (up to a $1/\sqrt{n}$ scaling factor):

$$\widetilde{\theta}^n(t_{j+1}) = \left(b^2 I + \sum_k V_k^n(t_j)\right)^{-1} \sum_k \left(V_k^n(t_j)\theta_* + S_k^n(t_j)\right) + \left(b^2 I + \sum_k V_k^n(t_j)\right)^{-1/2} N_{j+1} \tag{75}$$

where the $N_{j+1} \overset{\text{iid}}{\sim} N(0, I_{d\times d})$ are exogenously generated in accordance with the randomization of Thompson sampling. The conditional probability of playing action k at time $j+1$ is then

$$\begin{aligned} &p_k(A(j+1), V^n(t_j), S^n(t_j)) \\ &= \mathbb{P}\left(\widetilde{\theta}^n(t_{j+1})^\top A_k(j+1) > \max_{k' \ne k} \widetilde{\theta}^n(t_{j+1})^\top A_{k'}(j+1) \,\middle|\, A(j+1), V^n(t_j), S^n(t_j)\right). \end{aligned} \tag{76}$$

To obtain convergence of the processes V^n and S^n in (72)-(73), we define for each k:

$$\Lambda_k(V^n(t_j), S^n(t_j)) = \mathbb{E}\left[p_k(A(j+1), V^n(t_j), S^n(t_j)) \cdot A_k(j+1)A_k(j+1)^\top \,\middle|\, V^n(t_j), S^n(t_j)\right]. \tag{77}$$

This allows us to rewrite:

$$V_k^n(t_j) = \frac{1}{n}\sum_{i=1}^{j} \Lambda_k(V^n(t_{i-1}), S^n(t_{i-1})) + M_k^n(t_j) \tag{78}$$

$$S_k^n(t_j) = \sum_{i=1}^{j} \Lambda_k(V^n(t_{i-1}), S^n(t_{i-1}))^{1/2}\left(B_k^n(t_i) - B_k^n(t_{i-1})\right), \qquad k = 1, \ldots, K, \tag{79}$$

where

$$M_k^n(t_j) = \frac{1}{n}\sum_{i=1}^{j}\left(I_k(i)A_k(i)A_k(i)^\top - \Lambda_k(V^n(t_{i-1}), S^n(t_{i-1}))\right) \tag{80}$$

$$B_k^n(t_j) = \frac{1}{\sqrt{n}}\sum_{i=1}^{j} \Lambda_k(V^n(t_{i-1}), S^n(t_{i-1}))^{-1/2} I_k(i)A_k(i)\epsilon_k(i). \tag{81}$$

Like before, $(I_k(i) : 1 \leq k \leq 2)$ is a multinomial random variable with a single trial and success probabilities $p_k(A(i), V^n(t_{i-1}), S^n(t_{i-1}))$. We continuously interpolate all of these processes to be piecewise constant.

As $n \to \infty$, M_k^n and B_k^n converge weakly to the zero process in $D^{2d}[0,1]$ and a Brownian motion on $\mathbb{R}^d$ with covariance $C_k C_k^{-1}$, where $C_k = \mathbb{E}[A_k(1)A_k(1)^\top]$ and C_k^{-1} is its generalized inverse. Thus, we expect (78)-(79) to be a discrete approximation to the SDE:

$$V_k(t) = \int_0^t \Lambda_k(V(s), S(s))ds \tag{82}$$

$$S_k(t) = \int_0^t \Lambda_k(V(s), S(s))^{1/2} dB_k(s) \tag{83}$$

$$V_k(0) = 0,\ S_k(0) = 0, \qquad k = 1, \ldots, K, \tag{84}$$

where $V = (V_1, \ldots, V_K)$, $S = (S_1, \ldots, S_K)$, and each B_k is an independent standard Brownian motion on $\mathbb{R}^d$.

To keep track of the regret, we consider the processes

$$R_k^n(t_j) = \frac{1}{n} \sum_{i=1}^{j} I_k(i) A_k(i), \qquad k = 1, \ldots, K. \tag{85}$$

Similar to before, we define for each k,

$$\widetilde{\Lambda}_k(V^n(t_j), S^n(t_j)) = \mathbb{E}\Big[p_k(A(j+1), V^n(t_j), S^n(t_j)) \cdot A_k(j+1) \,\Big|\, V^n(t_j), S^n(t_j) \Big], \tag{86}$$

which allows us to re-write:

$$R_k^n(t_j) = \frac{1}{n} \sum_{i=1}^{j} \widetilde{\Lambda}_k(V^n(t_{i-1}), S^n(t_{i-1})) + \widetilde{M}_k^n(t_j), \tag{87}$$

where

$$\widetilde{M}_k^n(t_j) = \frac{1}{n} \sum_{i=1}^{j} \Big(I_k(i) A_k(i) - \widetilde{\Lambda}_k(V^n(t_{i-1}), S^n(t_{i-1})) \Big) \tag{88}$$

We continuously interpolate all of these processes to be piecewise constant. As before, when $n \to \infty$, we expect each R_k^n in (85) to converge weakly in $D^d[0,1]$ to

$$R_k(t) = \int_0^t \widetilde{\Lambda}_k(V(s), S(s))ds. \tag{89}$$

Then, the ($1/\sqrt{n}$ re-scaled) regret process in the limit system is $t\mathbb{E}[\max_k \theta_*^\top A_k(1)] - \sum_k \theta_*^\top R_k(t)$. For a rigorous statement of the above results, see Theorem 4 in Section 3.

Theorem 4 covers the finite-action linear bandit setting. Its proof follows from a direct extensions of the proof of Theorem 1, and is thus omitted.

THEOREM 4. *For the linear bandit with finitely many actions, the dynamics of Thompson sampling, which are characterized by the processes V^n and S^n (as defined in (72) and (73)), converge weakly in $D^{2Kd}[0,1]$ as $n \to \infty$ to the unique strong solution of the SDE:*

$$V_k(t) = \int_0^t \Lambda_k(V(s), S(s))ds \tag{90}$$

$$S_k(t) = \int_0^t \Lambda_k(V(s), S(s))^{1/2} dB_k(s) \tag{91}$$

$$V_k(0) = 0,\ S_k(0) = 0, \qquad k = 1, \ldots, K, \tag{92}$$

with the Λ_k defined in (77), and where each B_k is an independent standard Brownian motion on $\mathbb{R}^d$.

Furthermore, in the limit, the ($1/\sqrt{n}$ re-scaled) regret process is given by $t\mathbb{E}[\max_k \theta_*^\top A_k(1)] - \sum_k \theta_*^\top R_k(t)$, where

$$R_k(t) = \int_0^t \widetilde{\Lambda}_k(V(s), S(s))ds, \qquad k = 1, \ldots, K, \tag{93}$$

with the $\widetilde{\Lambda}_k$ defined in (86).

REMARK 7. When the covariate distribution for each action is a point mass, so that for each time i, $A_k(i)$ is equal to some fixed vector A_k for each action k, then the above theory simplifies significantly. In particular, the processes V_k^n and S_k^n defined in (72)-(73) simplify:

$$V_k^n(t_j) = \frac{1}{n}\sum_{i=1}^{j} I_k(i)A_k(i)A_k(i)^\top = A_k A_k^\top \frac{1}{n}\sum_{i=1}^{j} I_k(i) = A_k A_k^\top R_k^n(t_j) \tag{94}$$

$$S_k^n(t_j) = \frac{1}{\sqrt{n}}\sum_{i=1}^{j} I_k(i)A_k(i)\epsilon_k(i) = A_k \frac{1}{\sqrt{n}}\sum_{i=1}^{j} I_k(i)\epsilon_k(i) = A_k Y_k^n(t_j), \qquad k = 1, \ldots, K, \tag{95}$$

with the R_k^n and Y_k^n processes defined as in (3)-(4) from our derivations for the MAB. (Note that $X_k(i) - \mu_k^n$ from the MAB setting is exactly equivalent to $\epsilon_k(i)$ in this linear bandit setting.) Therefore, the processes $R^n = (R_1^n, \ldots, R_K^n)$ and $Y^n = (Y_1^n, \ldots, Y_K^n)$ completely capture the dynamics of Thompson sampling in this linear setting with finitely many fixed, non-random covariate vectors. SDE and stochastic ODE weak limits, similar to the ones we encountered earlier in (16)-(18) and (30)-(31), respectively, also hold here. The difference is that the functions p_k, as defined (in general for K arms) in (32) and (33), need to be modified to incorporate least-squares estimation of the parameter vector θ^n, which involves the covariate vectors A_k. Indeed, we can replace each p_k by Λ_k defined via (cf. (76)-(77)):

$$\Lambda_k(R^n(t_j), Y^n(t_j)) = \mathbb{P}\left(\widetilde{\theta}^n(t_{j+1})^\top A_k > \max_{k' \neq k} \widetilde{\theta}^n(t_{j+1})^\top A_{k'} \,\middle|\, R^n(t_j), Y^n(t_j)\right), \tag{96}$$

where now (cf. (75)):

$$\widetilde{\theta}^n(t_{j+1}) = \mathcal{C}^n(t_j)^{-1}\sum_k \left(A_kA_k^\top\theta_*R_k^n(t_j) + A_kY_k^n(t_j)\right) + \mathcal{C}^n(t_j)^{-1/2}N_{j+1}, \tag{97}$$

with

$$\mathcal{C}^n(t_j) = b^2I + \sum_k A_kA_k^\top R_k^n(t_j). \tag{98}$$

5.2. Infinitely Many Actions

We again consider a d-dimensional linear bandit model, where at time $i = 1, \dots, n$, the reward for playing action $A(i) \in \mathcal{A}$ is:

$$X(i) = (\theta^n)^\top A(i) + \epsilon(i), \qquad \epsilon(i) \overset{\text{iid}}{\sim} N(0,1). \tag{99}$$

But here, the set of possible actions $\mathcal{A}$ is infinite, and for illustration, we take $\mathcal{A}$ to be the unit ℓ_2-ball in $\mathbb{R}^d$, which is an important setting in its own right. Once again, the parameter vector θ^n is unknown and must be learned. And with a time horizon of n, we assume that $\theta^n = \frac{\theta_*}{\sqrt{n}}$, where θ_* is a fixed vector that does not change with n.

Remark 8. In this section, the action set $\mathcal{A}$ is uncountably infinite and does not change with time. Hence, we do not enumerate the actions like we did in the finite-action setting and simply refer to them as elements $A \in \mathcal{A}$.

Similar to the finite-action setting of previous section, to describe the dynamics of Thompson sampling in this setting, it suffices to keep track of two processes:

$$V^n(t_j) = \frac{1}{n}\sum_{i=1}^{j} A(i)A(i)^\top \tag{100}$$

$$S^n(t_j) = \frac{1}{\sqrt{n}}\sum_{i=1}^{j} A(i)\epsilon(i), \tag{101}$$

where $A(i)$ is the action chosen at time i. Conditional on the information collected up to and including time j:

$$\mathcal{G}_j^n = \sigma\left(A(i),\, X(i)\, :\, 1 \le i \le j\right), \tag{102}$$

we sample a vector from the posterior distribution of θ^n (up to a $1/\sqrt{n}$ scaling factor):

$$\widetilde{\theta}^n(t_{j+1}) = \left(b^2I + V^n(t_j)\right)^{-1}\left(V^n(t_j)\theta_* + S^n(t_j)\right) + \left(b^2I + V^n(t_j)\right)^{-1/2}N_{j+1}, \tag{103}$$

where the $N_{j+1} \overset{\text{iid}}{\sim} N(0, I_{d\times d})$ are exogenously generated in accordance with the randomization of Thompson sampling. Then, the action $A(j+1) \in \mathcal{A}$ which maximizes $\widetilde{\theta}^n(t_{j+1})^\top A(j+1)$ is

$$A(j+1) = \frac{\widetilde{\theta}^n(t_{j+1})}{\left\|\widetilde{\theta}^n(t_{j+1})\right\|}. \tag{104}$$

To obtain convergence of the processes V^n and S^n in (100)-(101), we define:

$$\Lambda(V^n(t_j), S^n(t_j)) = \mathbb{E}\left[\frac{\widetilde{\theta}^n(t_{j+1})\widetilde{\theta}^n(t_{j+1})^\top}{\left\|\widetilde{\theta}^n(t_{j+1})\right\|^2} \,\middle|\, V^n(t_j), S^n(t_j)\right]. \tag{105}$$

This allows us to rewrite:

$$V^n(t_j) = \frac{1}{n}\sum_{i=1}^{j} \Lambda(V^n(t_{i-1}), S^n(t_{i-1})) + M^n(t_j) \tag{106}$$

$$S^n(t_j) = \sum_{i=1}^{j} \Lambda(V^n(t_{i-1}), S^n(t_{i-1}))^{1/2} \left(B^n(t_i) - B^n(t_{i-1})\right), \tag{107}$$

where

$$M^n(t_j) = \frac{1}{n}\sum_{i=1}^{j}\left(\frac{\widetilde{\theta}^n(t_i)\widetilde{\theta}^n(t_i)^\top}{\left\|\widetilde{\theta}^n(t_i)\right\|^2} - \Lambda(V^n(t_{i-1}), S^n(t_{i-1}))\right) \tag{108}$$

$$B^n(t_j) = \frac{1}{\sqrt{n}}\sum_{i=1}^{j} \Lambda(V^n(t_{i-1}), S^n(t_{i-1}))^{-1/2} \frac{\widetilde{\theta}^n(t_i)}{\left\|\widetilde{\theta}^n(t_i)\right\|}\epsilon(i). \tag{109}$$

We continuously interpolate all of these processes to be piecewise constant.

As $n \to \infty$, M^n and B^n converge weakly to the zero process in $D^{2d}[0,1]$ and standard Brownian motion on $\mathbb{R}^d$. Thus, we expect (106)-(107) to be a discrete approximation to the SDE:

$$V(t) = \int_0^t \Lambda(V(s), S(s))ds \tag{110}$$

$$S(t) = \int_0^t \Lambda(V(s), S(s))^{1/2} dB(s) \tag{111}$$

$$V(0) = 0,\ S(0) = 0, \tag{112}$$

where B is a standard Brownian motion on $\mathbb{R}^d$.

To keep track of the regret, we consider the process

$$R^n(t_j) = \frac{1}{n}\sum_{i=1}^{j} \frac{\widetilde{\theta}^n(t_i)}{\left\|\widetilde{\theta}^n(t_i)\right\|}. \tag{113}$$

Similar to before, we define

$$\widetilde{\Lambda}(V^n(t_j), S^n(t_j)) = \mathbb{E}\left[\frac{\widetilde{\theta}^n(t_{j+1})}{\left\|\widetilde{\theta}^n(t_{j+1})\right\|} \,\middle|\, V^n(t_j), S^n(t_j)\right], \tag{114}$$

which allows us to re-write:

$$R^n(t_j) = \frac{1}{n}\sum_{i=1}^{j} \widetilde{\Lambda}(V^n(t_{i-1}), S^n(t_{i-1})) + \widetilde{M}^n(t_j), \tag{115}$$

where

$$\widetilde{M}^n(t_j) = \frac{1}{n}\sum_{i=1}^{j}\left(\frac{\widetilde{\theta}^n(t_i)}{\left\|\widetilde{\theta}^n(t_i)\right\|} - \widetilde{\Lambda}(V^n(t_{i-1}), S^n(t_{i-1}))\right). \tag{116}$$

We continuously interpolate all of these processes to be piecewise constant. As before, when $n \to \infty$, we expect R^n in (113) to converge weakly in $D^d[0,1]$ to

$$R(t) = \int_0^t \widetilde{\Lambda}(V(s), S(s))ds. \tag{117}$$

Then, the ($1/\sqrt{n}$ re-scaled) regret process in the limit system is $t\|\theta_*\| - \theta_*^\top R(t)$. For a rigorous statement of the above results, see Theorem 5 in Section 3.

Theorem 5 covers the infinite-action linear bandit setting. Its proof follows from a direct extensions of the proof of Theorem 1, and is thus omitted.

THEOREM 5. *For the linear bandit with the unit ℓ_2-ball as the action space, the dynamics of Thompson sampling, which are characterized by the processes V^n and S^n (as defined in (100) and (101)), converge weakly in $D^{2d}[0,1]$ as $n \to \infty$ to the unique strong solution of the SDE:*

$$V(t) = \int_0^t \Lambda(V(s), S(s))ds \tag{118}$$

$$S(t) = \int_0^t \Lambda(V(s), S(s))^{1/2} dB(s) \tag{119}$$

$$V(0) = 0,\ S(0) = 0, \tag{120}$$

with Λ defined in (105), and where B is a standard Brownian motion on $\mathbb{R}^d$.

Furthermore, in the limit, the ($1/\sqrt{n}$ re-scaled) regret process is given by $t\|\theta_*\| - \theta_*^\top R(t)$, where

$$R(t) = \int_0^t \widetilde{\Lambda}(V(s), S(s))ds, \tag{121}$$

with $\widetilde{\Lambda}$ defined in (114).

6. Applications of Diffusion Approximations

6.1. General Reward Distributions and Posterior Approximations

In this section, we study Thompson sampling with other (non-normal) reward distributions and priors in the MAB setting. We show that Thompson sampling satisfies a general invariance principle, which indicates that in $1/\sqrt{n}$-scale gap regimes, the behavior of Thompson sampling is largely independent of the particular characteristics of the arm reward distributions as well as that of prior distributions. In general, the diffusion limits all coincide with the limit for normally-distributed rewards and priors. We also point out that in minimax gap regimes, using a Laplace approximation of the posterior distribution leads to the same weak diffusion limits. This observation is in agreement with the frequentist view of Thompson sampling, in which sampling from the exact posterior distribution is not emphasized beyond its usefulness as a principled means of exploration.

For the rewards of different arms, we consider a general one-dimensional parametric family $\{F_\mu : \mu \in \Theta\}$ parameterized by mean μ, where Θ is a small open interval. For each arm k, we consider a sequence of distributions (CDFs) F_k^n from this family, with corresponding means $\mu_k^n \in \Theta$. We use σ_k^n to denote the corresponding standard deviations and Q_k^n to denote the corresponding quantile functions. To keep our derivations as concrete as possible in this section, we focus on the two-armed MAB setting, but it is easy to verify that Theorem 6 extends to multi-armed settings, following the discussion in Section 2.3. We assume there exists $\mu_0 \in \Theta$ such that for each arm k, $\mu_k^n \to \mu_0$ and $\mu_1^n - \mu_2^n = \Delta/\sqrt{n}$ for some fixed $\Delta > 0$. (Arm 1 is optimal.) We also assume there exists $\sigma_0 > 0$ such that for each arm k, $\sigma_k^n \to \sigma_0$. As before, we work with a "triangular array" setup where for each n and each arm k, the rewards $X_k(i)$, $i = 1, \dots, n$ are iid from the distribution F_k^n.

We assume the following conditions A1-A4 hold. These conditions allow us to obtain a suitable normal approximation for the posterior distribution of the mean; see Proposition 2 in Appendix A. (Proposition 2 is developed for models with general parameterization, with the models parameterized by mean in this section being a special case.) For the family F_μ, $l(\mu, x)$ denotes the log-likelihood function (observed values are plugged into the x argument), and $l^{(m)}(\mu, x)$ denotes the mth derivative with respect to μ. An expectation taken with respect to F_μ is written as $\mathbb{E}^\mu[\cdot]$. Fisher information evaluated at μ is given by $\mathcal{I}(\mu) = -\mathbb{E}^\mu[l^{(2)}(\mu, X)]$, which in this setting is equal to the variance of F_μ. Also, Q_μ denotes the quantile function of F_μ.

(A1) The set $\{x : l(\mu, x) > 0\}$ is the same for all $\mu \in \Theta$. For each x, $l(\mu, x)$ is three-times differentiable with respect to μ. Both $l(\mu, x)$ and $l^{(2)}(\mu, x)$ are jointly continuous with respect to μ and x.

(A2) At each $\mu \in \Theta$, differentiation and integration can be interchanged so that:

$$\mathbb{E}^\mu\left[l^{(1)}(\mu, X)\right] = 0 \tag{122}$$

$$\mathbb{E}^\mu\left[l^{(2)}(\mu, X)\right] = -\mathbb{E}^\mu\left[l^{(1)}(\mu, X)^2\right]. \tag{123}$$

(A3) For each $\delta > 0$, there is an $\epsilon > 0$ such that for all $\mu \in \Theta$,

$$\sup_{\mu' : |\mu - \mu'| \geq \delta} \mathbb{E}^\mu[l(\mu', X)] \leq \mathbb{E}^\mu[l(\mu, X)] - \epsilon. \tag{124}$$

Also, $\inf_{\mu \in \Theta} \mathcal{I}(\mu) > 0$.

(A4) There is a function η_1 and a continuous function η_2 such that for all x:

$$\eta_1(x) \geq \sup_{\mu\in\Theta} |l(\mu,x)| \vee \left|l^{(2)}(\mu,x)\right| \tag{125}$$

$$\eta_2(x) \geq \sup_{\mu\in\Theta} \left|l^{(3)}(\mu,x)\right| \tag{126}$$

$$\int_0^1 \sup_{\mu\in\Theta} \eta_i(Q_\mu(y))dy < \infty, \qquad i = 1,2. \tag{127}$$

The conditions A1-A3 are standard for parametric statistical estimation problems. Condition A4 is a sufficient condition that ensures the convergence to normality of the posterior distribution is suitably uniform in μ. It is not restrictive since, for example, it is easily satisfied by distributions with bounded support, given some smoothness of the log-likelihood function $l(\mu,x)$. Conditions weaker than A1-A4 suffice for location-scale families of distributions. When dealing with such families of distributions related by simple linear transformations, we can, for example, weaken condition A4—we just need a function η satisfying $\eta(x) \geq \sup_{\mu\in\Theta}|l(\mu,x)|$ for all x and $\int_0^1 \sup_{\mu\in\Theta} \eta(Q_\mu(y))dy < \infty$.

In the Thompson sampling algorithm, for each arm k, we put a prior on the arm mean such that there is positive density in a neighborhood of μ_0. For simplicity, we assume that the prior does not change with n in the triangular array setup. We emphasize that the prior need not be a conjugate prior for the reward distribution.

For simplicity, we focus on the two-armed MAB setting in our derivations, but it is easy to verify that Theorem 8 extends to multi-armed settings, as discussed in Section 2.3. As in Section 2.1, to obtain an SDE approximation, it suffices to consider the two processes $R^n = (R^n_1, R^n_2)$ and $Y^n = (Y^n_1, Y^n_2)$ defined via:

$$R^n_k(t_j) = \frac{1}{n}\sum_{i=1}^{j} I_k(i) \tag{128}$$

$$Y^n_k(t_j) = \frac{1}{\sqrt{n}}\sum_{i=1}^{j} I_k(i)\frac{X_k(i)-\mu^n_k}{\sigma^n_k}, \tag{129}$$

which we interpolate to be piecewise constant. We will describe the dynamics of Thompson sampling after time $\lfloor \epsilon n \rfloor$ for some arbitrarily small $\epsilon \in (0,1)$. At time $\lfloor \epsilon n \rfloor$, we assume that arm 1 has been played $\lfloor \epsilon n \cdot \alpha \rfloor$ times (and arm 2 the remainder), for any $\alpha \in (0,1)$. Then at a later time $j+1$, having collected history:

$$\mathcal{G}^n_j = \sigma\left(I_k(i), I_k(i)X_k(i) : \; k=1,2, \;\; 1 \leq i \leq j\right), \tag{130}$$

for each arm k, we sample a value $\widetilde{\mu}^n_k(j+1)$ from the posterior distribution of μ^n_k. The probability of playing the optimal arm 1 is

$$\mathbb{P}\big(\widetilde{\mu}^n_1(j+1) > \widetilde{\mu}^n_2(j+1) \,\big|\, \mathcal{G}^n_j\big)$$

$$= \mathbb{P}\Bigg(\sqrt{n}\left(\widetilde{\mu}_1^n(j+1) - \widehat{\mu}_1^n(j+1)\right) + \frac{Y_1^n(t_j)\sigma_1^n}{R_1^n(t_j)} + \Delta > \\ \sqrt{n}\left(\widetilde{\mu}_2^n(j+1) - \widehat{\mu}_2^n(j+1)\right) + \frac{Y_2^n(t_j)\sigma_2^n}{R_2^n(t_j)} \,\Bigg|\, \mathcal{G}_j^n \Bigg), \tag{131}$$

where for arm k, using the data collected through time j, $\widehat{\mu}_k^n(j+1)$ is the sample mean estimate. Then, assuming the conditions A1-A4 are satisfied for the family of distributions $\{F_\mu : \mu \in \Theta\}$, Proposition 2 ensures that for each arm k, as $n \to \infty$ (with $j \geq \epsilon n$):

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left(\sqrt{n}\left(\widetilde{\mu}_k^n(j+1) - \widehat{\mu}_k^n(j+1)\right) \leq x \,|\, \mathcal{G}_j^n \right) - \Phi\left(\frac{x\sqrt{R_k^n(t_j)}}{\sigma_k^n} \right) \right| \Rightarrow 0. \tag{132}$$

Therefore, with probability converging to 1 as $n \to \infty$, (131) is asymptotically equivalent to $\overline{p}_1(R^n, Y^n)$, where for $u = (u_1, u_2)$ and $v = (v_1, v_2)$,

$$\overline{p}_1(u,v) = \Phi\left(\frac{v_1\sigma_0 u_1^{-1} - v_2\sigma_0 u_2^{-1} + \Delta}{\sigma_0\sqrt{u_1^{-1} + u_2^{-1}}} \right). \tag{133}$$

And the probability of playing the sub-optimal arm 2 is asymptotically equivalent to $\overline{p}_2(R^n, Y^n) = 1 - \overline{p}_1(R^n, Y^n)$. Continuing the derivation like we did in Section 2.1 leads to the following theorem, whose proof is very similar to that of Theorem 1 and is thus omitted.

Theorem 6. *Consider a two-armed bandit where each arm's reward distribution comes from a parametric model satisfying conditions A1-A4. Then the dynamics of Thompson sampling, which are described by the processes $R^n = (R_1^n, R_2^n)$ and $Y^n = (Y_1^n, Y_2^n)$ (defined in (128)-(129)), converge weakly in $D^4[\epsilon, 1]$ as $n \to \infty$ to the processes $R = (R_1, R_2)$ and $Y = (Y_1, Y_2)$, which are unique strong solutions of the SDE:*

$$R_k(t) = R_k(\epsilon) + \int_\epsilon^t \overline{p}_k(R(s), Y(s))ds \tag{134}$$

$$Y_k(t) = Y_k(\epsilon) + \int_\epsilon^t \sqrt{\overline{p}_k(R(s), Y(s))}dB_k(s), \qquad k = 1, 2, \tag{135}$$

$$R_1(\epsilon) = \epsilon\alpha, \quad R_2(\epsilon) = \epsilon(1-\alpha), \tag{136}$$

$$Y_1(\epsilon) = B_1(\epsilon\alpha), \quad Y_2(\epsilon) = B_2(\epsilon(1-\alpha)), \tag{137}$$

where the B_k are independent standard Brownian motions.

As in Section 2.2, we can also obtain similar corresponding stochastic ODE representations of the SDEs (134)-(137).

6.2. Model Mis-specification

We use almost the same setup as in Section 6.1, with the following differences. Regarding the posterior distribution for each arm mean, we allow for the possibility that Thompson sampling may be designed for a family of (log-)likelihoods $\{l(\mu,x) : \mu \in \Theta\}$ that are mis-specified relative to the true family of reward distributions $\{F_\mu : \mu \in \Theta\}$. With this consideration in mind, we assume that the mis-specified log-likelihood $l(\mu,x)$, together with the distributional family $\{F_\mu : \mu \in \Theta\}$ (with quantile functions Q_μ), satisfies A1-A4 with the following exceptions. In A2, we assume that (122) holds, but we do not assume that (123) holds. In A3, in the place of the Fisher information $\mathcal{I}(\mu)$, we have the mis-specified version of Fisher information defined as:

$$\mathcal{I}^*(\mu) = -\mathbb{E}^\mu[l^{(2)}(\mu, X)]. \tag{138}$$

We additionally assume that $F_{\mu'} \Rightarrow F_\mu$ for each $\mu \in \Theta$ as $\mu' \to \mu$.

By slightly modifying Proposition 2, and following the derivation as in Section 6.1, we have the following. With probability converging to 1 as $n \to \infty$, the probability of playing the optimal arm 1 is asymptotically equivalent to $p_1^*(R^n, Y^n)$, where for $u = (u_1, u_2)$ and $v = (v_1, v_2)$,

$$p_1^*(u,v) = \Phi\left(\frac{v_1\sigma_0 u_1^{-1} - v_2\sigma_0 u_2^{-1} + \Delta}{\sqrt{\mathcal{I}^*(\mu_0)(u_1^{-1} + u_2^{-1})}}\right). \tag{139}$$

And the probability of playing the sub-optimal arm 2 is asymptotically equivalent to $p_2^*(R^n, Y^n) = 1 - p_1^*(R^n, Y^n)$. We then obtain the following theorem, whose proof is similar to previous ones and is thus omitted. Unlike the fragility results from Fan and Glynn (2021), in the diffusion regime, the behavior of algorithms such as Thompson sampling changes more smoothly with respect to increasing degrees of model mis-specification.

Theorem 7. *Under the assumptions described above, the dynamics of mis-specified Thompson sampling, which are described by the processes $R^n = (R_1^n, R_2^n)$ and $Y^n = (Y_1^n, Y_2^n)$ (defined in (128)-(129)), converge weakly in $D^4[\epsilon, 1]$ as $n \to \infty$ to the processes $R = (R_1, R_2)$ and $Y = (Y_1, Y_2)$, which are unique strong solutions of the SDE:*

$$R_k(t) = R_k(\epsilon) + \int_\epsilon^t p_k^*(R(s), Y(s))ds \tag{140}$$

$$Y_k(t) = Y_k(\epsilon) + \int_\epsilon^t \sqrt{p_k^*(R(s), Y(s))}dB_k(s), \qquad k = 1, 2, \tag{141}$$

$$R_1(\epsilon) = \epsilon\alpha, \quad R_2(\epsilon) = \epsilon(1-\alpha), \tag{142}$$

$$Y_1(\epsilon) = B_1(\epsilon\alpha), \quad Y_2(\epsilon) = B_2(\epsilon(1-\alpha)), \tag{143}$$

where the B_k are independent standard Brownian motions.

As in Section 2.2, we can also obtain similar corresponding stochastic ODE representations of the SDEs (140)-(143).

6.3. Bootstrap-based Exploration

The bootstrap (Efron 1979), a powerful way to approximate the sampling distributions of estimators through resampling, as well as related ideas such as subsampling, have recently been proposed for exploration in bandit problems (Baransi et al. 2014, Eckles and Kaptein 2014, Osband and Van Roy 2015, Tang et al. 2015, Elmachtoub et al. 2017, Vaswani et al. 2018, Kveton et al. 2019a,b, 2020b,a, Baudry et al. 2020). In this section, we consider the most basic implementation of bootstrapping for exploration in MABs, where in each time period, a single bootstrapped (sample) mean is sampled for each arm, and the arm with the greatest bootstrapped (sample) mean is played. We will refer to this as boostrap-based exploration. We find that the weak diffusion limits for bootstrap-based exploration with general reward distributions all coincide with the limit for Thompson sampling with normally-distributed rewards and priors (and also Thompson sampling with general reward distributions and priors, as we saw in Theorem 6). Thompson sampling with normally-distributed rewards and priors is known to be essentially optimal (up to a $\sqrt{\log(K)}$ factor, where K is the number of arms) in minimax gap regimes, from the perspective of expected regret (Agrawal and Goyal 2017). Thus, the fact that the simplest implementation of bootstrapping yields the same behavior as that of Thompson sampling in these minimax gap regimes suggests that boostrapping can indeed be a very effective means of exploration in bandit problems.

We assume the same model setup described in the second paragraph of Section 6.1. However, instead of assuming A1-A4, we assume the following conditions B1-B2. These conditions allow us to obtain a suitable normal approximation for bootstrapping the mean; see Proposition 3 in Appendix A.

(B1) For each $\mu \in \Theta$, $F_{\mu'} \Rightarrow F_\mu$ as $\mu' \to \mu$.

(B2) There exists a function η_3 such that $\eta_3(y) \geq \sup_{\mu\in\Theta} Q_\mu(y)^2$ for all $y \in (0,1)$ and $\int_0^1 \eta_3(y)dy < \infty$.

Again, we focus on the two-armed MAB setting in our derivations, but it is easy to verify that Theorem 8 extends to multi-armed settings, as discussed in Section 2.3. As in Sections 2.1 and 6.1, to obtain an SDE approximation, it suffices to consider the two processes $R^n = (R_1^n, R_2^n)$ and $Y^n = (Y_1^n, Y_2^n)$ defined in (128)-(129). Like in the previous sections, we will describe the dynamics of bootstrap-based exploration after time $\lfloor \epsilon n \rfloor$ for some arbitrarily small $\epsilon \in (0,1)$. At time $\lfloor \epsilon n \rfloor$, we assume that arm 1 has been played $\lfloor \epsilon n \cdot \alpha \rfloor$ times (and arm 2 the remainder), for any $\alpha \in (0,1)$. Then at a later time $j+1$, having collected history $\mathcal{G}_j^n$ as defined in (130), the probability of playing the optimal arm 1 is

$$\mathbb{P}\big(\widehat{\mu}_1^{*n}(j+1) > \widehat{\mu}_2^{*n}(j+1) \,\big|\, \mathcal{G}_j^n\big)$$

$$
\begin{aligned}
= \mathbb{P}\Bigg(& \sqrt{n}\left(\widehat{\mu}_1^{*n}(j+1) - \widehat{\mu}_1^{n}(j+1)\right) + \frac{Y_1^n(t_j)\sigma_1^n}{R_1^n(t_j)} + \Delta > \\
& \sqrt{n}\left(\widehat{\mu}_2^{*n}(j+1) - \widehat{\mu}_2^{n}(j+1)\right) + \frac{Y_2^n(t_j)\sigma_2^n}{R_2^n(t_j)} \,\Bigg|\, \mathcal{G}_j^n \Bigg),
\end{aligned} \tag{144}
$$

where for arm k, using the data collected through time j, $\widehat{\mu}_k^n(j+1)$ is the sample mean estimate and $\widehat{\mu}_k^{*n}(j+1)$ is a bootstrapped sample mean estimate. Then, assuming the conditions B1-B2 are satisfied for the family of distributions $\{F_\mu : \mu \in \Theta\}$, Proposition 3 ensures that for each arm k, as $n \to \infty$ (with $j \geq \epsilon n$):

$$
\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left(\sqrt{n}\left(\widehat{\mu}_k^{*n}(j+1) - \widehat{\mu}_k^{n}(j+1)\right) \leq x \mid \mathcal{G}_j^n \right) - \Phi\left(\frac{x\sqrt{R_k^n(t_j)}}{\sigma_k^n} \right) \right| \Rightarrow 0. \tag{145}
$$

Therefore, with probability converging to 1 as $n \to \infty$, (144) is asymptotically equivalent to $\overline{p}_1(R^n, Y^n)$, with $\overline{p}_1$ being exactly the probability function in (133). And the probability of playing the sub-optimal arm 2 is asymptotically equivalent to $\overline{p}_2(R^n, Y^n) = 1 - \overline{p}_1(R^n, Y^n)$. Continuing the derivation like we did in Section 2.1 leads to the following theorem, whose proof is very similar to that of Theorem 1 and is thus omitted.

Theorem 8. *Consider a two-armed bandit where each arm's reward distribution comes from a parametric model satisfying conditions B1-B2. The dynamics of bootstrap-based exploration, which are described by the processes $R^n = (R_1^n, R_2^n)$ and $Y^n = (Y_1^n, Y_2^n)$ (defined in (128)-(129)), converge weakly in $D^4[\epsilon, 1]$ as $n \to \infty$ to the processes $R = (R_1, R_2)$ and $Y = (Y_1, Y_2)$, which are unique strong solutions of the SDE:*

$$
R_k(t) = R_k(\epsilon) + \int_\epsilon^t \overline{p}_k(R(s), Y(s))ds \tag{146}
$$

$$
Y_k(t) = Y_k(\epsilon) + \int_\epsilon^t \sqrt{\overline{p}_k(R(s), Y(s))}dB_k(s), \qquad k = 1, 2, \tag{147}
$$

$$
R_1(\epsilon) = \epsilon\alpha, \quad R_2(\epsilon) = \epsilon(1-\alpha), \tag{148}
$$

$$
Y_1(\epsilon) = B_1(\epsilon\alpha), \quad Y_2(\epsilon) = B_2(\epsilon(1-\alpha)), \tag{149}
$$

where the B_k are independent standard Brownian motions.

As in Section 2.2, we can also obtain similar corresponding stochastic ODE representations of the SDEs (146)-(149).

6.4. Estimated Variances

In the bandit literature, it is almost always assumed that the variances (or variance proxies in the sub-Gaussian case) of the arm distributions are known or a bound on them is known. In particular, there have been few works on algorithmic modifications which allow arm variances to be estimated along with the means. The earliest work in this direction is that of Audibert et al. (2009), who propose the UCB-V algorithm (where "V" stands for variance), which incorporates variance estimation into its upper confidence bounds. UCB-V has been shown to significantly outperform its alternatives, which assume equal variances, in settings such as when the suboptimal arms have lower variance than the optimal arm. For Thompson sampling, Honda and Takemura (2014) showed that it is possible to incorporate variance estimation within a Bayesian framework and still achieve asymptotically optimal expected regret by using certain priors. And recently, Cowan et al. (2018) showed that asymptotic optimality is achievable using variance-adaptive UCB algorithms. One can see from these papers, novel work notwithstanding, that it can be technically complicated to bound the expected regret when there is both mean and variance estimation involved.

Fortunately, from the weak convergence perspective, it is technically straightforward to accommodate variance estimation. For simplicity, we again focus on the two-armed case and assume that the rewards for playing arms $k = 1, 2$ are:

$$X_k(i) \overset{\text{iid}}{\sim} N(\mu_k^n, \sigma_k^2),$$

where as before, $\mu_1^n - \mu_2^n = \Delta/\sqrt{n}$, but now the variances σ_k^2 (which we assume to be constant for all n) are also unknown. First, we define sample variances (recall $t_j = j/n$):

$$S_k^n(t_j) = \frac{1}{R_k^n(t_j)n} \sum_{i=1}^{j} \left(I_k(i)X_k(i) - \overline{m}_k^n(t_j)\right)^2 \tag{150}$$

$$\overline{m}_k^n(t_j) = \frac{1}{R_k^n(t_j)n} \sum_{i=1}^{j} I_k(i)X_k(i), \tag{151}$$

and S_k^n can be made into a process on $D[0,1]$ by piecewise constant interpolation. We use a simple modification of Thompson sampling to incorporate variance estimation, which is essentially what one would use when taking a fully Bayesian approach and defining a joint prior on (μ_k^n, σ_k^2) that is proportional to $(\sigma_k^2)^{-1-\alpha}$ for some $\alpha < 0$. (See Honda and Takemura (2014) and Cowan et al. (2018).)

As in Sections 2.1 and 6.1, to obtain an SDE approximation, it suffices to consider the two processes $R^n = (R_1^n, R_2^n)$ and $Y^n = (Y_1^n, Y_2^n)$ defined in (128)-(129). Again, we describe the dynamics of Thompson sampling after time $\lfloor \epsilon n \rfloor$ for some arbitrarily small $\epsilon \in (0,1)$. At time $\lfloor \epsilon n \rfloor$, we assume that arm 1 has been played $\lfloor \epsilon n \cdot \alpha \rfloor$ times (and arm 2 the remainder), for any $\alpha \in (0,1)$. Then at a later time $j+1$, having collected history $\mathcal{G}_j^n$ as defined in (130), for each arm k, we sample a value from the posterior distribution:

$$\widetilde{\mu}_k^n \sim N\left(\frac{\sum_{i=1}^{j} I_k(i)X_k(i)}{R_k^n(t_j)n}, \frac{S_k^n(t_j)}{R_k^n(t_j)n}\right). \tag{152}$$

Then, the probability of playing the optimal arm 1 is:

$$
\begin{aligned}
&\mathbb{P}\left(\widetilde{\mu}_2^n > \widetilde{\mu}_1^n \,\middle|\, \mathcal{G}_j^n\right) && (153)\\
&= \mathbb{P}\left(N_1\left(\frac{Y_1^n(t_j)\sigma_1}{R_1^n(t_j)} + \Delta, \frac{S_1^n(t_j)}{R_1^n(t_j)}\right) > N_2\left(\frac{Y_2^n(t_j)\sigma_2}{R_2^n(t_j)}, \frac{S_2^n(t_j)}{R_2^n(t_j)}\right) \,\middle|\, \mathcal{G}_j^n\right), && (154)
\end{aligned}
$$

where the N_k are independent normal random variables with their specified means and variances. It is straightforward to see that $S_k^n(t) \xrightarrow{\mathbb{P}} \sigma_k^2$ for any $t \in [\epsilon, 1]$ as $n \to \infty$. So the weak limits of the sample variance processes S_k^n are the constant $D[\epsilon, 1]$ processes taking values σ_k^2. Then, following the rest of the derivation in Section 2.1, we end up with the SDE:

$$
\begin{aligned}
&R_k(t) = R_k(\epsilon) + \int_\epsilon^t p_k^{\mathrm{v}}(R(s), Y(s))ds && (155)\\
&Y_k(t) = Y_k(\epsilon) + \int_\epsilon^t \sqrt{p_k^{\mathrm{v}}(R(s), Y(s))}dB_k(s), \qquad k = 1, 2, && (156)\\
&R_1(\epsilon) = \epsilon\alpha, \quad R_2(\epsilon) = \epsilon(1-\alpha), && (157)\\
&Y_1(\epsilon) = B_1(\epsilon\alpha), \quad Y_2(\epsilon) = B_2(\epsilon(1-\alpha)), && (158)
\end{aligned}
$$

where the B_k are independent standard Brownian motions, and for any $u = (u_1, u_2) \in [0,1]^2$ and $v = (v_1, v_2) \in \mathbb{R}^2$,

$$
\begin{aligned}
&p_1^{\mathrm{v}}(u, v) = \Phi\left(\frac{v_1\sigma_1 u_1^{-1} - v_2\sigma_2 u_2^{-1} + \Delta}{\sqrt{\sigma_1^2 u_1^{-1} + \sigma_2^2 u_2^{-1}}}\right) && (159)\\
&p_2^{\mathrm{v}}(u, v) = 1 - p_1^{\mathrm{v}}(u, v). && (160)
\end{aligned}
$$

As in Section 2.2, we can also obtain similar corresponding stochastic ODE representations of the SDEs (155)-(158).

We make some further remarks. First, although we have been considering Thompson sampling designed for normally-distributed rewards, the specifics of the reward distributions do not matter. As long as the reward distributions have finite variances, we are guaranteed to end up with weak convergence to Brownian limits by the functional central limit theorem in the martingale (leading to our SDEs) or iid (leading to our stochastic ODEs) settings. (See Whitt (2007), Chapter 7 of Ethier and Kurtz (1986) and Chapter 2 of Billingsley (1999)). Second, although both the arm means and variances are assumed to be unknown in this section, we have seen that variance estimation is basically a trivial task. This is in agreement with the intuition that the variances associated with Brownian-like quantities can generally be assumed to be known exactly. In fact, the probability measures associated with Brownian motions having different variances are mutually singular(!), and statistical inference for diffusion processes always assumes from the outset that variances and dispersion functions are known exactly (Kutoyants 2004). The non-trivial task in diffusion models is to estimate drift, which is essentially the case here—the Δ parameter (the $\sqrt{n}$-rescaled gap between arm means) appearing in all of our previous discussions plays the role of an unknown drift parameter that must be learned by Thompson sampling. This discussion suggests that variance estimation should generally be part of bandit algorithm design in such minimax settings. Indeed, assuming the time horizon is long enough to allow small gaps between arm means to be estimated with some degree of confidence, there will generally be enough data collected to accurately estimate variances.

6.5. Batched Updates

In many 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 some arm (adaptively determined) for an interval of time (also possibly adaptively determined), and then the algorithm is updated all at once with the data collected during the interval. For a time horizon of n, suppose we perform batched Thompson sampling by committing to play an arm for $o(n)$ time periods, and then updating all at once before making a new commitment for the next $o(n)$ periods. Assuming the gap sizes are of magnitude $1/\sqrt{n}$, as $n \to \infty$, we would obtain weak convergence to the same diffusion limit processes (SDEs and stochastic ODEs) as in the case of ordinary (non-batched) Thompson sampling. Indeed, a time interval of $o(n)$ in the discrete pre-limit system corresponds to (after dividing by n) an infinitesimally small time interval in the continuous limit system. This suggests that as long as the number of batches increases to infinity (possibly at an arbitrarily slow rate) as $n \to \infty$, and each batch is not too large (at most $o(n)$ periods), then the distribution of regret will be approximately the same compared to the case in which one updates in every period (batch sizes of one). To make this precise, we have the following proposition, whose straightforward proof is omitted.

Proposition 1. *For the K-armed MAB or linear bandit (with non-random covariate vectors), let $R^n = (R^n_k : 1 \le k \le K)$ and $Y^n = (Y^n_k : 1 \le k \le K)$ denote the dynamics of Thompson sampling, as in (11)-(13) from the derivation of the SDE approximation in Section 2.1. Let $\bar{R}^n = (\bar{R}^n_k : 1 \le k \le K)$ and $\bar{Y}^n = (\bar{Y}^n_k : 1 \le k \le K)$ denote the same resulting quantities (with the same arm reward sequences) when Thompson sampling is implemented with $o(n)$ batched updates. Then the processes R^n and $\bar{R}^n$, as well as Y^n and $\bar{Y}^n$, have the same weak limits as $n \to \infty$. An analogous result holds for Thompson sampling with $o(n)$ batched updates under the stochastic ODE approximation derived in Section 2.2.*

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. It should be an interesting future direction to study general batched bandit algorithms designed for the minimax regime from the perspective of weak convergence/diffusion approximation.

Appendix A: Normal Approximations for Posteriors and the Bootstrap

We consider a general one-dimensional parametric family $\{F_\theta : \theta \in \Theta\}$, where Θ is a small interval. Our version of the Bernstein-von Mises theorem in Proposition 2 is a locally-uniform almost sure version, in the sense that almost sure convergence to normality of the posterior distribution holds uniformly for all parameter values $\theta \in \Theta$. In order for almost sure convergence simultaneously for different parameter values θ to make sense, we consider an independent sequence of uniform $(0,1)$ random variables U_i, $i \geq 1$, and we obtain sequences of observations for different θ by considering $Q_\theta(U_i)$, $i \geq 1$, where Q_θ is the quantile function corresponding to CDF F_θ. Hence, the $Q_\theta(U_i)$ are iid observations from F_θ. For a sequence of parameter values θ^n, we use F^n and Q^n to denote the corresponding CDF's and quantile functions.

We assume the following conditions BvM.1-BvM.4 hold. We use $l(\theta,x)$ to denote the log-likelihood function (observed values are plugged into the x argument), and $l^{(m)}(\theta,x)$ denotes the mth derivative with respect to θ. We use $\mathcal{I}(\theta)$ to denote the Fisher information (of a single observation) evaluated at θ. Expectations taken with respect to F_θ are written using $\mathbb{E}^\theta[\cdot]$.

(BvM.1) The set $\{x : l(\theta,x) > 0\}$ is the same for all $\theta \in \Theta$. For each x, $l(\theta,x)$ is three-times differentiable with respect to θ. Both $l(\theta,x)$ and $l^{(2)}(\theta,x)$ are jointly continuous with respect to θ and x.

(BvM.2) At each $\theta \in \Theta$, differentiation and integration can be interchanged so that:

$$\mathbb{E}^\theta\left[l^{(1)}(\theta,X)\right] = 0 \tag{161}$$

$$\mathbb{E}^\theta\left[l^{(2)}(\theta,X)\right] = -\mathbb{E}^\theta\left[l^{(1)}(\theta,X)^2\right]. \tag{162}$$

(BvM.3) For each $\delta > 0$, there is an $\epsilon > 0$ such that for all $\theta \in \Theta$,

$$\sup_{\theta' : |\theta-\theta'| \geq \delta} \mathbb{E}^\theta[l(\theta',X)] \leq \mathbb{E}^\theta[l(\theta,X)] - \epsilon. \tag{163}$$

Also, $\inf_{\theta\in\Theta} \mathcal{I}(\theta) > 0$.

(BvM.4) There is a function η_1 and a continuous function η_2 such that for all x:

$$\eta_1(x) \geq \sup_{\theta\in\Theta} |l(\theta,x)| \vee \left|l^{(2)}(\theta,x)\right| \tag{164}$$

$$\eta_2(x) \geq \sup_{\theta\in\Theta} \left|l^{(3)}(\theta,x)\right| \tag{165}$$

$$\int_0^1 \sup_{\theta\in\Theta} \eta_i(Q_\theta(u))du < \infty, \qquad i = 1,2. \tag{166}$$

Proposition 2. *Suppose conditions BvM.1-BvM.4 hold for some small interval* Θ. *Consider a sequence of parameter values* $\theta^n \to \theta_0$ *for some (fixed deterministic)* $\theta_0 \in \Theta$ *as* $n \to \infty$. *Use a (fixed, non-changing with* n*) prior* π^0 *with continuous positive density in a neighborhood of* θ_0. *For a single sequence* U_i, $i \geq 1$, *of independent uniform* $(0,1)$ *random variables, for each* n, *we condition on the observations* $Q^n(U_1), \ldots, Q^n(U_n)$ *to obtain the posterior distribution* $\widetilde{\theta}^n \mid Q^n(U_1), \ldots, Q^n(U_n)$ *for* θ^n. *Let* $\widehat{\theta}^n$ *denote the maximum likelihood estimator (MLE) of* θ^n *using the observations* $Q^n(U_1), \ldots, Q^n(U_n)$. *Then, almost surely, the centered and scaled posterior density* $\pi^n(t \mid Q^n(U_1), \ldots, Q^n(U_n))$, *for* $\sqrt{n}(\widetilde{\theta}^n - \widehat{\theta}^n)$, *satisfies:*

$$\lim_{n\to\infty} \int_{\mathbb{R}} \left| \pi^n(t \mid Q^n(U_1), \ldots, Q^n(U_n)) - \sqrt{\frac{\mathcal{I}(\theta^n)}{2\pi}} \exp\left(-\frac{1}{2}t^2 \mathcal{I}(\theta^n)\right) \right| dt = 0. \tag{167}$$

Proof of Proposition 2. The proof here is adapted from Theorem 4.2 on page 104 of Ghosh et al. (2006). All of the statements to follow are almost sure statements, so generally we will not repeatedly state so. To begin, note that the posterior density can be written as

$$\pi^n(t \mid Q^n(U_1), \ldots, Q^n(U_n)) = (C^n)^{-1} \pi^0(\widehat{\theta}^n + t/\sqrt{n}) \exp\left(L_n(\widehat{\theta}^n + t/\sqrt{n}) - L_n(\widehat{\theta}^n)\right), \tag{168}$$

with normalization factor $(C^n)^{-1}$ and

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

Let

$$D^n(t) = \pi^0(\widehat{\theta}^n + t/\sqrt{n}) \exp\left(L_n(\widehat{\theta}^n + t/\sqrt{n}) - L_n(\widehat{\theta}^n)\right) - \pi^0(\theta^n) \exp\left(-\frac{1}{2}t^2 \mathcal{I}(\theta^n)\right). \tag{169}$$

To show (167), it suffices to show that

$$\lim_{n\to\infty} \int_{\mathbb{R}} |D^n(t)|\, dt = 0. \tag{170}$$

If (170) holds, then

$$\lim_{n\to\infty} \left(C^n - \pi^0(\theta^n)\sqrt{2\pi/\mathcal{I}(\theta^n)}\right) = 0.$$

So we would have

$$\begin{aligned}
&\int_{\mathbb{R}} \left| \pi^n(t \mid Q^n(U_1), \ldots, Q^n(U_n)) - \sqrt{\mathcal{I}(\theta^n)/2\pi} \exp\left(-\frac{1}{2}t^2 \mathcal{I}(\theta^n)\right) \right| dt \\
&\leq (C^n)^{-1} \int_{\mathbb{R}} |D^n(t)|\, dt + \left| (C^n)^{-1}\pi^0(\theta^n) - \sqrt{\mathcal{I}(\theta^n)/2\pi} \right| \int_{\mathbb{R}} \exp\left(-\frac{1}{2}t^2 \mathcal{I}(\theta^n)\right) dt,
\end{aligned}$$

and the proof of Proposition 2 would be complete.

To show (170), we consider two cases: $A^n = \{t : |t| > \gamma\sqrt{n}\}$ and $(A^n)^c = \{t : |t| \leq \gamma\sqrt{n}\}$, where we will set $\gamma > 0$ later. For the first case involving the set A^n, note that

$$\begin{aligned}\int_{A^n} |D^n(t)|\, dt \leq & \int_{A^n} \pi^0(\widehat{\theta}^n + t/\sqrt{n}) \exp\left(L_n(\widehat{\theta}^n + t/\sqrt{n}) - L_n(\widehat{\theta}^n)\right) dt \\ & + \int_{A^n} \pi^0(\theta^n) \exp\left(-\frac{1}{2}t^2 \mathcal{I}(\theta^n)\right) dt. \end{aligned} \tag{171}$$

Since $\theta^n \to \theta_0$ as $n \to \infty$, it is straightforward to verify that the second integral in (171) goes to zero as $n \to \infty$. As for the first integral, note from Lemma 5 that we have (locally-uniform) almost sure consistency of the MLE: $\widehat{\theta}^n - \theta^n \to 0$ as $n \to \infty$. Using this nice property, we are able to establish in Lemma 6 that there exists $\epsilon > 0$ such that almost surely, for sufficiently large n,

$$\frac{1}{n}\left(L_n(\widehat{\theta}^n + t/\sqrt{n}) - L_n(\widehat{\theta}^n)\right) \leq -\epsilon$$

on the set A^n. Therefore, the first integral in (171) also goes to zero as $n \to \infty$.

For the second case involving the set $(A^n)^c$, we expand L_n in a Taylor series about the MLE $\widehat{\theta}^n$, noting that by the definition of the MLE, $L_n^{(1)}(\widehat{\theta}^n) = 0$, where $L_n^{(m)}(y)$ is the mth derivative of L_n with respect to the argument y. For large n, we have

$$L_n(\widehat{\theta}^n + t/\sqrt{n}) - L_n(\widehat{\theta}^n) = -\frac{1}{2}t^2 \widehat{\mathcal{I}}^n(\widehat{\theta}^n) + r^n(t), \tag{172}$$

where

$$\widehat{\mathcal{I}}^n(\widehat{\theta}^n) = -\frac{1}{n} L_n^{(2)}(\widehat{\theta}^n)$$

and

$$r^n(t) = \frac{1}{6}\left(\frac{t}{\sqrt{n}}\right)^3 L_n^{(3)}(\theta_t^n),$$

where θ_t^n lies between $\widehat{\theta}^n$ and $\widehat{\theta}^n + t/\sqrt{n}$. Using condition BvM.4 and Lemma 8, for sufficiently large n,

$$|r^n(t)| \leq \frac{1}{6}\frac{t^3}{\sqrt{n}}\frac{1}{n}\sum_{i=1}^n \eta_2(Q^n(U_i)) \leq \frac{1}{3}\frac{t^3}{\sqrt{n}} \mathbb{E}\left[\sup_{\theta \in \Theta} \eta_2(Q_\theta(U_1))\right], \tag{173}$$

and so for fixed t, we have $r^n(t) \to 0$ as $n \to \infty$. On the set $(A^n)^c$, we have $t/\sqrt{n} \leq \gamma$, and so (173) can be re-written as

$$|r^n(t)| \leq \frac{1}{3}\gamma t^2 \mathbb{E}\left[\sup_{\theta \in \Theta} \eta_2(Q_\theta(U_1))\right]. \tag{174}$$

Taking $\gamma > 0$ to be sufficiently small, and using Lemma 7, along with the fact that $\inf_{\theta \in \Theta} \mathcal{I}(\theta) > 0$ from BvM.3, we have from (172) that on the set $(A^n)^c$,

$$\exp\left(L_n(\widehat{\theta}^n + t/\sqrt{n}) - L_n(\widehat{\theta}^n)\right) \leq \exp(-at^2) \tag{175}$$

for some fixed $a > 0$ and sufficient large n. Thus, using condition BvM.1 and Lemma 5, $D^n(t)$ is dominated by an integrable function on the set $(A^n)^c$ for sufficiently large n. Using (173), $r^n(t) \to 0$ for fixed t as $n \to \infty$, so together with Lemma 7, we have $D^n(t) \to 0$ for fixed t as $n \to \infty$. The dominated convergence theorem then yields

$$\int_{(A^n)^c} |D^n(t)|\, dt \to 0$$

as $n \to \infty$, thereby concluding the proof. □

Lemmas 5-8 below verify certain technical details in the proof of Proposition 2.

LEMMA 5. *Almost surely, $\widehat{\theta}^n - \theta^n \to 0$ as $n \to \infty$.*

Proof of Lemma 5. Using condition BvM.1, for any $\theta \in \Theta$ and x, $l(\theta', x) \to l(\theta, x)$ as $\theta' \to \theta$, so (by Scheffé's Lemma) $F_{\theta'} \Rightarrow F_\theta$ as $\theta' \to \theta$. Therefore, $Q_{\theta'}(u) \to Q_\theta(u)$ for all points of continuity $u \in (0,1)$ of Q_θ. (See, for instance, Proposition 3.1 on page 112 of Shorack (2000).) Note that for each $\theta' \in \Theta$, $Q_{\theta'}$ can be discontinuous at only countably many points (as a non-decreasing function), and $l(\theta, x)$ is jointly continuous in θ and x by condition BvM.1. Also, there is a function η_1 for which $\sup_{\theta' \in \Theta} \eta_1(Q_{\theta'}(u))$ is integrable and dominates $l(\theta, Q_{\theta'}(u))$ for all $\theta, \theta' \in \Theta$ by BvM.4. Therefore, the class of functions $\{u \mapsto l(\theta, Q_{\theta'}(u)) : \theta, \theta' \in \Theta\}$ has finite L^1-bracketing numbers (see Example 19.8 on page 272 of van der Vaart (1998)), and so it is a Glivenko-Cantelli class of functions:

$$\sup_{\theta, \theta' \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n l(\theta, Q_{\theta'}(U_i)) - \int_0^1 l(\theta, Q_{\theta'}(u)) du \right| \to 0 \tag{176}$$

almost surely as $n \to \infty$. And we also have

$$\frac{1}{n} \sum_{i=1}^n l(\theta^n, Q^n(U_i)) - \int_0^1 l(\theta^n, Q^n(u)) du \to 0 \tag{177}$$

almost surely as $n \to \infty$.

By the definition of the MLE $\widehat{\theta}^n$,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n l(\widehat{\theta}^n, Q^n(U_i)) &= \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n l(\theta, Q^n(U_i)) \\ &\geq \frac{1}{n} \sum_{i=1}^n l(\theta^n, Q^n(U_i)). \end{aligned}$$

Thus, using (177),

$$\frac{1}{n} \sum_{i=1}^n l(\widehat{\theta}^n, Q^n(U_i)) \geq \int_0^1 l(\theta^n, Q^n(u)) du + a(n),$$

for some sequence $a(n) \to 0$ almost surely as $n \to \infty$. We can then write

$$\begin{aligned}
0 &\le \int_0^1 l(\theta^n, Q^n(u))du - \int_0^1 l(\widehat{\theta}^n, Q^n(u))du \\
&\le \frac{1}{n}\sum_{i=1}^n l(\widehat{\theta}^n, Q^n(U_i)) - \int_0^1 l(\widehat{\theta}^n, Q^n(u))du - a(n) \\
&\le \sup_{\theta,\theta' \in \Theta} \left| \frac{1}{n}\sum_{i=1}^n l(\theta, Q_{\theta'}(U_i)) - \int_0^1 l(\theta, Q_{\theta'}(u))du \right| - a(n) \\
&\to 0
\end{aligned}$$

almost surely as $n \to \infty$ by (176). Finally, applying condition BvM.3, we must have $\widehat{\theta}^n - \theta^n \to 0$ almost surely as $n \to \infty$. □

Lemma 6. *For any $\gamma > 0$, there exists $\epsilon > 0$ such that almost surely, for sufficiently large n,*

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

on the set $A^n = \{t : |t| > \gamma\sqrt{n}\}$.

Proof of Lemma 6. Note that

$$\frac{1}{n}\left(L_n(\widehat{\theta}^n + t/\sqrt{n}) - L_n(\widehat{\theta}^n)\right) = \frac{1}{n}\sum_{i=1}^n l(\widehat{\theta}^n + t/\sqrt{n}, Q^n(U_i)) - \frac{1}{n}\sum_{i=1}^n l(\widehat{\theta}^n, Q^n(U_i)).$$

So using (176) from the proof of Lemma 5, we then have

$$\begin{aligned}
&\frac{1}{n}\left(L_n(\widehat{\theta}^n + t/\sqrt{n}) - L_n(\widehat{\theta}^n)\right) \\
&\le \int_0^1 l(\widehat{\theta}^n + t/\sqrt{n}, Q^n(u))du - \int_0^1 l(\widehat{\theta}^n, Q^n(u))du + a(n)
\end{aligned} \tag{178}$$

for some sequence $a(n) \to 0$ almost surely as $n \to \infty$. Recall that $\theta^n \to \theta_0$ as $n \to \infty$, $\widehat{\theta}^n - \theta^n \to 0$ almost surely as $n \to \infty$, and $l(\theta, Q_{\theta'}(u))$ is jointly continuous in θ and θ' for all but at most countably many $u \in (0,1)$ by condition BvM.1. Therefore, condition BvM.4 and the dominated convergence theorem yield

$$\int_0^1 l(\widehat{\theta}^n, Q^n(u))du - \int_0^1 l(\theta^n, Q^n(u))du \to 0 \tag{179}$$

almost surely as $n \to \infty$. For sufficiently large n, on the set A^n, we have $\left|\widehat{\theta}^n + t/\sqrt{n} - \theta^n\right| \ge \gamma/2$. The desired conclusion then follows from condition BvM.3, (178) and (179). □

Lemma 7. *Almost surely, $\widehat{\mathcal{I}}^n(\widehat{\theta}^n) - \mathcal{I}(\theta^n) \to 0$ as $n \to \infty$.*

Proof of Lemma 7. Note that

$$\begin{aligned}&\widehat{\mathcal{I}}^n(\widehat{\theta}^n) - \mathcal{I}(\widehat{\theta}^n)\\&= -\frac{1}{n}\sum_{i=1}^n l^{(2)}(\widehat{\theta}^n, Q^n(U_i)) + \int_0^1 l^{(2)}(\widehat{\theta}^n, Q^n(u))du.\end{aligned}$$

Using condition BvM.1, we have for any $\theta \in \Theta$ that $Q_{\theta'}(u) \to Q_\theta(u)$ for all points of continuity $u \in (0,1)$ of Q_θ as $\theta' \to \theta$. (See, for instance, Proposition 3.1 on page 112 of Shorack (2000).) Since at each $\theta \in \Theta$, $u \mapsto Q_\theta(u)$ can be discontinuous at only countably many $u \in (0,1)$ (as a non-decreasing function), and $l^{(2)}(\theta, x)$ is jointly continuous in θ and x by BvM.1, at each $\theta, \theta' \in \Theta$, $l(\theta, Q_{\theta'}(u))$ is jointly continuous in θ and θ' for all but at most countably many $u \in (0,1)$. By BvM.4, there exists an integrable function which dominates $l(\theta, Q_{\theta'}(u))$ for all θ and θ', and so the class of functions $\{u \mapsto l(\theta, Q_{\theta'}(u)) : \theta, \theta' \in \Theta\}$ has finite L^1-bracketing numbers (see Example 19.8 on page 272 of van der Vaart (1998)). Therefore, it is a Glivenko-Cantelli class of functions:

$$\sup_{\theta,\theta' \in \Theta}\left|\frac{1}{n}\sum_{i=1}^n l^{(2)}(\theta, Q_{\theta'}(U_i)) - \int_0^1 l^{(2)}(\theta, Q_{\theta'}(u))du\right| \to 0$$

almost surely as $n \to \infty$. Lastly, $\mathcal{I}(\widehat{\theta}^n) - \mathcal{I}(\theta^n) \to 0$ almost surely as $n \to \infty$ by Lemma 5, BvM.1 and BvM.4, and the dominated convergence theorem, thereby completing the proof. □

LEMMA 8. *Almost surely,*

$$\frac{1}{n}\sum_{i=1}^n \eta_2(Q^n(U_i)) - \mathbb{E}[\eta_2(Q^n(U_1))] \to 0$$

as $n \to \infty$.

The proof follows very similarly to the proof of Lemma 7 and is thus omitted.

In Proposition 3 below, we develop a normal approximation for bootstrapping the mean, similar in spirit to the normal approximation for posterior distributions in Proposition 2. Like before, for almost sure convergence to make sense in this setting, we consider a sequence of independent uniform $(0,1)$ random variables U_i, $i \geq 1$, and we obtain sequences of observations for different θ by considering $Q_\theta(U_i)$, $i \geq 1$ (which are thus iid from F_θ). We assume the following conditions.

(Boot.1) For each $\theta \in \Theta$, $F_{\theta'} \Rightarrow F_\theta$ as $\theta' \to \theta$.

(Boot.2) There exists function η_3 such that $\eta_3(u) \geq \sup_{\theta \in \Theta} Q_\theta(u)^2$ for all $u \in (0,1)$ and $\int_0^1 \eta_3(u)du < \infty$.

PROPOSITION 3. *Suppose conditions Boot.1-Boot.2 hold for some small interval* Θ. *Consider a sequence of parameter values* $\theta^n \to \theta_0$ *for some (fixed deterministic)* $\theta_0 \in \Theta$ *as* $n \to \infty$. *For a single* *sequence U_i, $i \geq 1$, of independent uniform $(0,1)$ random variables, for each n, let $\widehat{\mu}^n$ denote the sample mean (for estimating the mean of F^n) using the observations $Q^n(U_1), \ldots, Q^n(U_n)$. Let $\widehat{\mu}^{*n}$ denote the bootstrapped sample mean. Then, almost surely,*

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left| \mathbb{P}\left(\sqrt{n}\,(\widehat{\mu}^{*n} - \widehat{\mu}^n) \leq x \mid Q^n(U_1), \ldots, Q^n(U_n)\right) - \Phi\left(\frac{x}{\sigma_0}\right) \right| = 0, \tag{180}$$

where σ_0^2 is the variance of F_{θ_0}.

Proof of Proposition 3. We check conditions to be able to apply Proposition 1.3.1 on page 11 of Politis et al. (1999).

First of all, with $\widehat{F}^n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}(Q^n(U_i) \leq x)$, we have

$$\begin{aligned}
\sup_{x\in\mathbb{R}} \left| \widehat{F}^n(x) - F^n(x) \right| &= \sup_{x\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^n \mathbb{I}(U_i \leq F^n(x)) - F^n(x) \right| \\
&\leq \sup_{y\in(0,1)} \left| \frac{1}{n}\sum_{i=1}^n \mathbb{I}(U_i \leq y) - y \right| \\
&\to 0
\end{aligned}$$

almost surely as $n \to \infty$ by the classical Glivenko-Cantelli Theorem.

Using condition Boot.1, we have for any $\theta \in \Theta$ that $Q_{\theta'}(u) \to Q_\theta(u)$ for all points of continuity $u \in (0,1)$ of Q_θ. (See, for instance, Proposition 3.1 on page 112 of Shorack (2000).) Since at each θ, Q_θ can be discontinuous at only countably many points (as a non-decreasing function), and by condition Boot.2, there exists an integrable function η_3 which dominates Q_θ^2 for all $\theta \in \Theta$, the class of functions $\{u \mapsto Q_\theta(u)^2 : \theta \in \Theta\}$ has finite L^1-bracketing numbers (see Example 19.8 on page 272 of van der Vaart (1998)). Therefore, it is a Glivenko-Cantelli class of functions:

$$\sup_{\theta\in\Theta} \left| \frac{1}{n}\sum_{i=1}^n Q_\theta(U_i)^2 - \mathbb{E}[Q_\theta(U_1)^2] \right| \to 0$$

almost surely as $n \to \infty$. By a similar argument, we also have

$$\sup_{\theta\in\Theta} \left| \frac{1}{n}\sum_{i=1}^n Q_\theta(U_i) - \mathbb{E}[Q_\theta(U_1)] \right| \to 0$$

almost surely as $n \to \infty$.

Using conditions Boot.1-Boot.2 and the dominated convergence theorem, we have as $n \to \infty$:

$$\begin{aligned}
\int x^2 dF^n(x) = \mathbb{E}[Q^n(U_1)^2] \to \mathbb{E}[Q_{\theta_0}(U_1)^2] = \int x^2 dF_{\theta_0}(x) \\
\int x dF^n(x) = \mathbb{E}[Q^n(U_1)] \to \mathbb{E}[Q_{\theta_0}(U_1)] = \int x dF_{\theta_0}(x).
\end{aligned}$$

Thus, the conditions of Proposition 1.3.1 on page 11 of Politis et al. (1999) are verified. □

Appendix B: Weak Convergence Technical Lemmas

For the convenience of the reader, here we collect some results from the literature about processes in $D^d[0,1]$, equipped with the Skorohod metric.

Lemma 9 **(Tightness of Multi-dimensional Processes)**. *A sequence of process $\xi^n = (\xi_1^n, \ldots, \xi_d^n)$ is tight in $D^d[0,1]$ if and only if each ξ_j^n and each $\xi_j^n + \xi_k^n$ are tight in $D[0,1]$, for all $1 \le j,k \le d$. (See Problem 22 of Chapter 3 on page 153 of Ethier and Kurtz (1986).)*

Lemma 10 **(Simple Sufficient Conditions for Tightness)**. *A sequence of processes ξ^n in $D[0,1]$ adapted to filtrations $\mathcal{F}_t^n$ is tight if*

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

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

$$\mathbb{E}\left[(\xi^n(t+u) - \xi^n(t))^2 \,|\, \mathcal{F}_t^n \right] \le \mathbb{E}\left[\alpha_n(\delta) \,|\, \mathcal{F}_t^n \right] \tag{T2}$$

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

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

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

Lemma 11 **(Martingale Functional Central Limit Theorem)**. *Let $\xi^n = (\xi_1^n, \ldots, \xi_d^n)$ be a sequence of martingales in $D^d[0,1]$ (equipped with the Skorohod metric), adapted to filtrations $\mathcal{F}_t^n$ and satisfying the conditions:*

$$\lim_{n\to\infty} \mathbb{E}\left[\sup_{0\le t\le 1} \left| \xi_j^n(t) - \xi_j^n(t-) \right| \right] = 0, \qquad j = 1, \ldots, d, \tag{M1}$$

and suppose there exists a symmetric positive-definite matrix $C \in \mathbb{R}^{d\times d}$ with components $C_{j,k}$ such that the cross-variation processes satisfy

$$\langle \xi_j^n, \xi_k^n \rangle_t \xrightarrow{\mathbb{P}} C_{j,k} t. \tag{M2}$$

Then the joint process ξ^n converges weakly to a Brownian motion on $\mathbb{R}^d$ with covariance structure C (and zero drift). (See Theorem 2.1 of Whitt (2007), which is adapted from Theorem 1.4 of Chapter 7 on page 339 of Ethier and Kurtz (1986).)

References

Agrawal R (1995) Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability* 27(4):1054–1078.

Agrawal S, Goyal N (2012) Analysis of Thompson sampling for the multi-armed bandit problem. *Conference on Learning Theory* .

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

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

Audibert J, Munos R, Szepesvari C (2009) Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science* 410(19):1876–1902.

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* .

Bastani H, Bayati M (2020) Online decision-making with high-dimensional covariates. *Operations Research* 68(1):276–294.

Baudry D, Kaufmann E, Maillard O (2020) Sub-sampling for efficient non-parametric bandit exploration. *Advances in Neural Information Processing Systems* .

Bickel P, Klaassen C, Ritov Y, Wellner J (1998) *Efficient and Adaptive Estimation for Semiparametric Models* (Springer).

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

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

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.

Cowan W, Honda J, Katehakis M (2018) Normal bandits of unknown means and variances. *Journal of Machine Learning Research* 18(1):1–28.

Eckles D, Kaptein M (2014) Thompson sampling with the online bootstrap. *arXiv:1410.4009* .

Efron B (1979) Bootstrap methods: another look at the jackknife. *The Annals of Statistics* 7(1):1–26.

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

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

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

Ferreira K, Simchi-Levi D, Wang H (2018) Online network revenue management using Thompson sampling. *Operations Research* 66(6):1586–1602.

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).

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

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. *arXiv:2106.02126* .

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

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

Kuang X, Wager S (2021) Diffusion asymptotics for sequential experiments. *arXiv:2101.09855v2* .

Kurtz T, Protter P (1991) Weak limit theorems for stochastic integrals and stochastic differential equations. *The Annals of Probability* 19(3):1035–1070.

Kutoyants Y (2004) *Statistical Inference for Ergodic Diffusion Processes* (Springer).

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).

Le Cam L, Yang G (2000) *Asymptotics in Statistics: Some Basic Concepts* (Springer).

Li L, Chu W, Langford J, Schapire R (2010) A contextual-bandit approach to personalized news article recommendation. *International World Wide Web Conference* 661–670.

Misra K, Schwartz E, Abernethy J (2019) Dynamic online pricing with incomplete information using multi-armed bandit experiments. *Marketing Science* 38(2):226–252.

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

Salomon A, Audibert J (2011) Deviations of stochastic bandit regret. *International Conference on Algorithmic Learning Theory* 159–173.

Scott S (2010) A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry* 26(6):639–658.

Shen W, Wang J, Jiang Y, Zha H (2015) Portfolio choices with orthogonal bandit learning. *International Joint Conference on Artificial Intelligence* 974–980.

Shorack G (2000) *Probability for Statisticians* (Springer).

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

Tang L, Jiang Y, Li L, Zeng C, Li T (2015) Personalized recommendation via parameter-free contextual bandits. *International ACM SIGIR Conference on Research and Development in Information Retrieval* .

Tewari A, Murphy S (2017) From ads to interventions: contextual bandits in mobile health. Rehg J, Murphy S, Kumar S, eds., *Mobile Health: Sensors, Analytic Methods, and Applications*, chapter 25, 495–517 (Springer).

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 (1998) *Asymptotic Statistics* (Cambridge University Press).

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

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

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