Diffusion Approximations for Thompson Sampling

Kintan Saha Vipul Tejwani

December 28, 2025

Outline

Goal of the Paper

- The paper analyzes the behavior of **Thompson Sampling (TS)** during the phase where the algorithm has not yet clearly distinguished between the arms.
- This is done by studying discrete-time settings where

$$\Delta_k = O(\sqrt{\gamma}) \quad \text{up to time } O\left(\frac{1}{\gamma}\right).$$

- The reasoning is that we need at least

$$O\left(\frac{1}{\Delta_k^2}\right)$$

time steps to reliably distinguish between arms.

- The paper then analyzes the **limiting behavior** of such an algorithm as $\gamma \to 0$, leading to a diffusion approximation that captures TS dynamics in this "small-gap" regime.

Setup: Random Table Reward Model

- Let $X_k(1), X_k(2), \ldots$ denote the sequence of rewards for arm k
- If the arm chosen at time t is k, then the observed reward is:

$$Y(t) = X_k(t)$$

Time t	**Arm 1**	**Arm 2**	**Arm 3**	$\cdots$
1	$X_1(1)$	$X_2(1)$	$X_3(1)$	$\cdots$
2	$X_1(2)$	$X_2(2)$	$X_3(2)$	$\cdots$
3	$X_1(3)$	$X_2(3)$	$X_3(3)$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

- Each row corresponds to a round t.
- Each column corresponds to an arm k.
- The highlighted cell shows the reward observed at time $t = 2$ when arm 3 was chosen.

Setup: Assumption 1

Assumption 1 (Small Gap with IID Rewards)

For each γ and each arm $k \in [K]$, we have a reward distribution Q_k^γ with mean μ_k^γ, variance $(\sigma_k^\gamma)^2$, and rewards $X_k^\gamma(i) \overset{\text{iid}}{\sim} Q_k^\gamma$ for $i = 1, 2, \ldots$

There exist some $\alpha > 0$, some $\mu_* \in \mathbb{R}$, and for each arm k, some fixed $d_k \in \mathbb{R}$, $\sigma_k > 0$ such that

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

$$\lim_{\gamma \downarrow 0} \sigma_k^\gamma = \sigma_k,$$

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

For simplicity, assume $\mu^* = 0$

Interpretation: Arm means differ by order $\sqrt{\gamma}$, variances stabilize, and reward distributions have uniformly bounded $(2 + \alpha)$ moments.

Notations

- Let $I_k(t)$ denote the indicator that arm k is pulled at time t.
- Define the cumulative counts and cumulative rewards:

$$N_k(t) = \sum_{i=1}^{t} I_k(i), \qquad G_k(t) = \sum_{i=1}^{t} I_k(i) Y(i),$$

where:

  - $N_k(t)$: number of times arm k has been pulled up to time t;
  - $G_k(t)$: cumulative reward obtained from arm k.

- Define the **rescaled processes** (used for the diffusion limit):

$$U_k^{\gamma}(t) = \gamma\, N_k\left(\left\lfloor \frac{t}{\gamma} \right\rfloor\right), \qquad S_k^{\gamma}(t) = \sqrt{\gamma} \sum_{i=1}^{\lfloor t/\gamma \rfloor} I_k(i)\, \frac{(Y^{\gamma}(i) - \mu_k^{\gamma})}{\sigma_k^{\gamma}},$$

where:

  - $U_k^{\gamma}(t)$: scaled fraction of plays of arm k;
  - $S_k^{\gamma}(t)$: centered and scaled cumulative reward noise.

Posterior with concentrated Gaussian prior.

Assume for arm k the prior (Concentrated priors)

$$\mu_k \sim \mathcal{N}(\mu^*, \gamma/b) \text{ where } b > 0 \text{ is fixed,}$$

and conditional likelihoods

$$Y_k(i) \mid \mu_k \overset{\text{iid}}{\sim} \mathcal{N}(\mu_k, (c^*)^2).$$

Since N_k is the number of observations for arm k and $G_k = \sum_{i=1}^{N_k} Y_k(i)$.

Posterior (exact). The posterior is Gaussian with precision (inverse of variance) equal to prior precision plus data precision:

$$\text{precision:} \qquad \frac{1}{v_k} = \frac{b}{\gamma} + \frac{N_k}{(c^*)^2}.$$

Hence the posterior variance and mean are

$$v_k = \left(\frac{b}{\gamma} + \frac{N_k}{(c^*)^2}\right)^{-1}, \qquad m_k = v_k\left(\frac{b}{\gamma}\mu^* + \frac{G_k}{(c^*)^2}\right).$$

Discrete dynamics and the decomposition

In Thompson Sampling, in each step we sample

$$\tilde{\mu}_k^\gamma(j+1) \sim \mathcal{N}(m_k^\gamma(j), v_k^\gamma(j)), \quad \mathbb{P}\big(I^\gamma(j+1) = k \mid \mathcal{H}_j^\gamma\big) = p_k^\gamma\big(U^\gamma(j\gamma), S^\gamma(j\gamma)\big)$$

(i.e. choose the arm with highest posterior sample).
Then the rescaled processes admit the decomposition (for each arm k):

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

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

Here the fluctuation terms are defined by

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

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

$M_k^\gamma \Rightarrow 0$ and $B_k^\gamma \Rightarrow B$

1) Vanishing martingale M_k^γ:

- M_k^γ is a zero-mean martingale
- ΔM_k^γ has mean 0 and variance $\mathrm{O}(\gamma)$
- As $\gamma \downarrow 0$, ΔM_k^γ converges weakly to 0

$$\Rightarrow \quad M_k^\gamma(t) = \sum_{i=1}^{t} \Delta M_k^\gamma(i) \text{ converges weakly to the zero process.}$$

2) Brownian limit for B_k^γ:

- B_k^γ is a sum of many small, mean-zero reward deviations scaled by $\sqrt{\gamma}$.
- Each term adds independent noise with variance $\mathrm{O}(\gamma)$.
- At time t, by CLT, $B_k^\gamma(t) - B_k^\gamma(s) \Rightarrow \mathcal{N}\big(0, (t-s)I_K\big)$ as $\gamma \downarrow 0$

$$\Rightarrow \quad B^\gamma \text{ converges weakly to a } K\text{-dimensional Brownian motion } B.$$

Intuition: M^γ represents negligible sampling noise, while B^γ captures the aggregate reward randomness that diffuses into Gaussian noise in the limit.

Diffusion Limit under Random Table Model

Theorem 1

Under the **Random Table Reward Model**, and assuming **Assumption 1** (small-gap regime) together with **concentrated Gaussian priors**, we have

$$(U^\gamma, S^\gamma) \;\Rightarrow\; (U, S) \quad \text{as } \gamma \downarrow 0 \text{ in } D_{2K}[0,\infty),$$

where (U, S) is the unique strong solution to the SDE:

$$\begin{aligned} dU_k(t) &= p_k(U(t), S(t))\, dt, \\ dS_k(t) &= \sqrt{p_k(U(t), S(t))}\, dB_k(t), \\ U_k(0) &= S_k(0) = 0, \qquad k \in [K], \end{aligned}$$

with $B(t)$ a standard K-dimensional Brownian motion, and $p_k(\cdot)$ defined as the Thompson Sampling choice probabilities.

Moreover, for (scaled) regret:

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

Discrete cumulative reward decomposition

From the definition of the scaled noise

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

we obtain the exact identity

$$\boxed{\gamma\, G_k\big(\lfloor t/\gamma \rfloor\big) = U_k^\gamma(t)\, \mu_k^\gamma \;+\; \sigma_k \sqrt{\gamma}\, S_k^\gamma(t).}$$

Multiplying by $1/\sqrt{\gamma}$ gives the $\sqrt{\gamma}$-scaled form:

$$\boxed{\sqrt{\gamma}\, G_k\big(\lfloor t/\gamma \rfloor\big) = \frac{U_k^\gamma(t)\mu_k^\gamma}{\sqrt{\gamma}} \;+\; \sigma_k\, S_k^\gamma(t).}$$

Here,

- $U_k^\gamma(t) = \gamma N_k(\lfloor t/\gamma \rfloor)$ is the scaled play count,
- $S_k^\gamma(t)$ encodes centered cumulative reward noise.

These discrete decompositions will give finite deterministic or stochastic limits depending on the scaling and centering.

Limiting behavior and intuition

Using $\mu_k^\gamma = \mu^* - \sqrt{\gamma}\,\Delta_k$ and $(U^\gamma, S^\gamma) \Rightarrow (U, S)$ as $\gamma \downarrow 0$:

(a) γG scaling:

$$\boxed{\gamma\, G_k\big(\lfloor t/\gamma \rfloor\big) \implies \mu^*\, U_k(t),}$$

representing the dominant deterministic mean reward.

(b) Centered $\sqrt{\gamma} G$ scaling:

$$\boxed{\sqrt{\gamma}\left(G_k\big(\lfloor t/\gamma \rfloor\big) - \frac{\mu^*}{\gamma}\, U_k^\gamma(t) \right) \implies -\, U_k(t)\,\Delta_k + \sigma_k\, S_k(t).}$$

After subtracting the leading mean term $(\mu^*/\gamma)U_k^\gamma$, the remaining fluctuations capture both deterministic and stochastic corrections.

Intuition:

- The γG limit isolates the main mean reward $\mu^* U(t)$.
- Subtracting this and rescaling by $\sqrt{\gamma}$ reveals the next-order effects: the deterministic gap contribution $(-U\Delta)$ and the Gaussian noise (σS).
- Together they describe the fine-scale, stochastic evolution of cumulative reward and the diffusion-scale behavior of regret.

Setup - Reward Stack Model

Setup (reward stack model). At time j, reward for the selected arm k is

$$Y(j) = X_k\big(N_k(j-1)+1\big) \quad \text{when } I_k(j) = 1,$$

with sufficient statistics

$$N_k(j) = \sum_{i=1}^{j} I_k(i), \qquad G_k(j) = \sum_{i=1}^{j} I_k(i)\, Y(i). \tag{1}$$

This is the rested-bandit, reward–stack feedback mechanism.
Defining the rescaled and centered version of (1):

$$U_k^{\gamma}(t) = \gamma \sum_{i=1}^{\lfloor t/\gamma \rfloor} I_k^{\gamma}(i), \qquad Z_k^{\gamma}(t) = \sqrt{\gamma} \sum_{i=1}^{\lfloor t/\gamma \rfloor} \frac{X_k^{\gamma}(i) - \mu_k^{\gamma}}{\sigma_k} \tag{2}$$

Setup - Stationary rewards

Definition (Stationary Rewards)

For each arm $k \in [K]$, the rewards $\{X_k(i)\}_{i=1}^{\infty}$ are said to be *stationary* (which allows for serial dependence) if, for any fixed integers

$$1 \leq i_1 \leq i_2 \leq \cdots \leq i_\ell < \infty,$$

the finite-dimensional distributions

$$(X_k(i_1 + j), X_k(i_2 + j), \ldots, X_k(i_\ell + j))$$

are the same for all $j \geq 0$.

That is, stationary reward sequences need not be independent but however have constant mean and variance invariant over time, and hence they are a generalisation over i.i.d. rewards.

Setup - Assumptions

The following assumption can be intuitively seen through the lens of the Central Limit Theorem.

Assumption 2 (Small Gap Regime with Stationary Rewards)

For each γ and arm $k \in [K]$, the reward sequence $\{X_k^\gamma(i)\}_{i \geq 1}$ is **stationary** (with independence across arms) with mean $\mu_k^\gamma = \mu^* + \sqrt{\gamma}\, d_k$.

There exists $\sigma_k > 0$ such that the scaled, centered process

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

is tight in $D[0, \infty)$ and converges weakly to standard Brownian motion.

Methodology

At time j+1, conditioned on $\mathcal{H}_j^\gamma$, TS draws a sample from posterior of arm k as:

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

Thus, the probability of playing arm k is:

$$\mathbb{P}\left(k = \arg\max_{\ell \in [K]} \tilde{\mu}_\ell^\gamma(j+1) \,\middle|\, \mathcal{H}_j^\gamma \right) = p_k^\gamma\big(U^\gamma(j\gamma),\, Z^\gamma \circ U^\gamma(j\gamma)\big)$$

Hence,

$$U_k^\gamma(t) = \gamma \sum_{i=0}^{\lfloor t/\gamma \rfloor - 1} p_k^\gamma\big(U^\gamma(i\gamma),\, Z^\gamma \circ U^\gamma(i\gamma)\big) + M_k^\gamma(t), \quad k \in [K], \qquad (3)$$

where

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

Methodology

As $\gamma \downarrow 0$, we have $M^\gamma \Rightarrow 0$, and by Assumption 2, $Z^\gamma \Rightarrow B$. Hence, equation (3) can be interpreted as a discrete approximation of the following stochastic ODE:

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

with standard K-dimensional Brownian motion B, and functions p_k as before. Hence, we have the following result:

Setup

Theorem 2

Under Assumption 2 and concentrated priors, we have the following:

$$(U^\gamma, Z^\gamma \circ U^\gamma) \Rightarrow (U, B \circ U) \tag{4}$$

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

$$dU_k(t) = p_k(U(t), B \circ U(t))\, dt \tag{5}$$

$$U_k(0) = 0, \quad k \in [K], \tag{6}$$

with standard K-dimensional Brownian motion B, and functions p_k as before.

Moreover, for regret,

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

Connection to SDEs

It can be shown that the SDE formulation of the *Random Table model* and the SODE formulation of the *Reward Stack model* are **distributionally equivalent**.

Theorem 3

That is, any non-anticipative solution V of the SODE

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

corresponds to a solution of the SDE with a version $\tilde{B}$ of the K-dimensional Brownian motion.

$$dU_k(t) = p_k(U(t), S(t))\, dt, \qquad dS_k(t) = \sqrt{p_k(U(t), S(t))}\, d\tilde{B}_k(t),$$

such that (U, S) and $(V, B \circ V)$ are equal in distribution whenever the SDE admits a unique strong solution.

Connection to SDEs

Proposition 1

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

$$(U, S) \stackrel{d}{=} (U,\ B \circ U),$$

which solves the stochastic ODE with U as a non-anticipative solution.

Removing Concentrated Priors $\rightarrow$ ϵ-warm start

Motivation. At initialization, if some arms are never sampled, the terms $\frac{s_\ell \sigma_\ell}{u_\ell}$ in the sampling function

$$p_k(u,s) = \mathbb{P}\left(k = \arg\max_{\ell \in [K]} \left\{ \frac{s_\ell \sigma_\ell}{u_\ell} + d_\ell + \frac{c_*}{\sqrt{u_\ell}} N_\ell \right\} \right)$$

become undefined at $u_\ell = 0$. An ε-**warm start** prevents this degeneracy.

Definition. Each arm k is given a small positive initial sampling proportion:

$$q_k > 0, \qquad \sum_{k=1}^{K} q_k = 1,$$

and the initial condition for the sampling effort process is

$$U_k(0) = q_k \, \varepsilon, \quad k \in [K].$$

Effect.

- Ensures all $u_\ell > 0$ after initialization, making $p_k(u,s)$ and $\sqrt{p_k(u,s)}$ locally Lipschitz.

ϵ-warm start

Theorem 4 (Gaussian Thompson Sampling with ε-Warm Start)

Consider the Gaussian Thompson sampler with any fixed Gaussian prior (no γ-dependence), under ε-warm-start (with initial sampling probabilities $q_k > 0$ and $\sum_k q_k = 1$). Then, the conclusions of Theorems 1 and 2 (i.e., the SDE limit and SODE limit) continue to hold with the sampling probabilities defined by

$$p_k(u,s) = \begin{cases} q_k, & \text{if } \sum_\ell u_\ell \le \varepsilon, \\ \mathbb{P}\left(k = \arg\max_{\ell\in[K]} \left\{ \frac{s_\ell \sigma_\ell}{u_\ell} + d_\ell + \frac{c_*}{\sqrt{u_\ell}} N_\ell \right\}\right), & \text{if } \sum_\ell u_\ell > \varepsilon, \end{cases}$$

where the N_ℓ are independent standard Gaussian random variables.

Exponential-family distributions

A large class of distributions can be written in the **exponential-family form**:

$$p(x \mid \theta) = h(x)\, \exp\big(\theta\, T(x) - A(\theta)\big),$$

where

- θ is the **natural parameter**,
- $T(x)$ is a **sufficient statistic**,
- $A(\theta)$ ensures normalization.

Key property: The data enter the likelihood only through the average statistic

$$\bar{T}_n = \frac{1}{n} \sum_{i=1}^{n} T(x_i),$$

so all inference depends on $\bar{T}_n$.

Why the posterior becomes Gaussian

Posterior: for a smooth prior $\pi(\theta)$,

$$\pi(\theta \mid x_{1:n}) \propto \exp\Big(n[\theta\, \bar{T}_n - A(\theta)]\Big)\pi(\theta).$$

As n grows, the posterior **concentrates near the true parameter** θ^*.

Expand $A(\theta)$ around θ^* (Taylor approximation):

$$A(\theta) \approx A(\theta^*) + A'(\theta^*)(\theta - \theta^*) + \tfrac{1}{2}A''(\theta^*)(\theta - \theta^*)^2.$$

Plugging this in gives

$$\log \pi(\theta \mid x_{1:n}) \approx -\tfrac{nA''(\theta^*)}{2}(\theta - \theta^*)^2 + n(\bar{T}_n - A'(\theta^*))(\theta - \theta^*).$$

The second term is random but small ($O_p(\sqrt{n})$). So the posterior is approximately **quadratic in** $\theta \to$ **Gaussian in shape.**

$$\boxed{\theta \mid x_{1:n} \;\approx\; \mathcal{N}\Big(\hat{\theta}_n,\; \frac{1}{nA''(\theta^*)}\Big),}$$

where $A''(\theta^*)$ is the Fisher information.

Intuition and connection to Gaussian Thompson Sampling

Intuition:

- Near the true parameter θ^*, the log-likelihood is almost a **parabola** — smooth and locally quadratic.
- The randomness from the data (through $\bar{T}_n$) is approximately **Gaussian** by the Central Limit Theorem.
- Quadratic shape + Gaussian fluctuations = Gaussian posterior.

So, as the posterior concentrates:

$$\pi(\theta \mid x_{1:n}) \approx \mathcal{N}\Big(\text{mean near MLE}, \tfrac{1}{nA''(\theta^*)}\Big),$$

Connection to the paper:

- In the small-gap regime, each arm's posterior is highly concentrated.
- Exponential-family posteriors are therefore **locally Gaussian.**
- This justifies using the Gaussian posterior for all reward models

$$\tilde{\mu}_k^\gamma \sim \mathcal{N}(m_k^\gamma, v_k^\gamma)$$

The Gaussian case becomes a **canonical local approximation.**

Bootstrap Thompson Sampling (BTS)

Idea: A non-parametric alternative to Thompson Sampling that uses **bootstrap resampling** of past rewards to induce exploration without assuming any prior.

Algorithm (each round):

1. For each arm k, collect past rewards $\mathcal{D}_k = \{r_{k,1}, \ldots, r_{k,n_k}\}$.
2. Draw a **bootstrap sample** $\{r^*_{k,1}, \ldots, r^*_{k,n_k}\}$ from $\mathcal{D}_k$ *with replacement*.
3. Compute the bootstrapped mean:

$$\tilde{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} r^*_{k,i}.$$

4. Play the arm with the largest bootstrapped mean:

$$a_t = \arg\max_{k \in [K]} \tilde{\mu}_k.$$

5. Observe the reward and update $\mathcal{D}_k$.

Intuition: Random resampling of old rewards mimics posterior uncertainty.

Setup for Bootstrap Thompson Sampling (BTS)

Setting:

- K-armed bandit in the Random Table model of reward feedback under Assumption 1 with means $\mu_k \in \mathcal{I}$, where $\mathcal{I} \subset \mathbb{R}$ is an open interval.
- No parametric model is assumed; reward distributions P^μ can be arbitrary.

Moment Condition:

$$\lim_{y \to \infty} \sup_{\mu \in \mathcal{I}} \mathbb{E}\left[(X^\mu)^2 \, \mathbf{1}_{\{(X^\mu)^2 > y\}} \right] = 0,$$

ensuring finite second moments uniformly over μ i.e. X^μ is uniformly integrable.

Result: By using a Gaussian approximation for the bootstrapped sample mean, the authors show the diffusion limit process dynamics is similar to that of the Gaussian Thompson Sampling.

Bootstrap Thompson Sampling: Limit Dynamics

Process-level convergence: For the bootstrap sampler under ε-warm-start ($q_k > 0,\ \sum_k q_k = 1$), the scaled processes satisfy

$$(U^\gamma, S^\gamma) \ \Rightarrow\ (U, S) \qquad \text{as } \gamma \downarrow 0,$$

where (U, S) solves the SDE system (unique strong solution):

$$dU_k(t) = p_k(U(t), S(t))\, dt,\ dS_k(t) = \sqrt{p_k(U(t), S(t))}\, dB_k(t),\ U_k(0) = S_k(0) = 0$$

Arm–selection function p_k: With $u \in [0, \infty)^K$, $s \in \mathbb{R}^K$ and i.i.d. $N_\ell \sim \mathcal{N}(0, 1)$,

$$p_k(u, s) = \begin{cases} q_k, & \sum_\ell u_\ell \le \varepsilon, \\ \mathbb{P}\left(k = \arg\max\limits_{\ell \in [K]} \left\{ \dfrac{s_\ell \sigma_\ell}{u_\ell} + d_\ell + \dfrac{\sigma_\ell}{\sqrt{u_\ell}} N_\ell \right\} \right), & \sum_\ell u_\ell > \varepsilon. \end{cases}$$

Regret in the limit:

$$\sqrt{\gamma}\, \mathrm{Reg}^\gamma(\lfloor \cdot/\gamma \rfloor) \ \Rightarrow\ \sum_{k \in [K]} U_k(\cdot)\, \Delta_k \quad \text{in } D[0, \infty)$$

Remark Compared to Gaussian TS, where a fixed noise scale c_*^2 must be specified, the bootstrap sampler *automatically adapts* to each arm's limit variance σ_k^2

Model Mis-specification

When dealing with **concentrated priors**, we assume the variance of the conditional likelihoods to be $(c^*)^2$.

Mis-specification corresponds to a mismatch between the true reward variances $(\sigma_k)^2$ and the assumed prior variance $(c^*)^2$.

However, the sampling probabilities $p_k(u, s)$ are **continuous in** c^*, so a small mismatch between σ_k and c^* leads only to a small perturbation in $p_k(u, s)$. Hence, the limiting behavior of the algorithm remains stable even under mild variance mis-specification.

Moreover, since

$$\sqrt{\gamma}\,\mathrm{Reg}^{\gamma}(\lfloor t/\gamma \rfloor) \;\Rightarrow\; \sum_{k\in[K]} U_k(t)\,\Delta_k \quad \text{as } \gamma \downarrow 0,$$

we also have

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

since weak convergence implies convergence in distribution, and expectations of bounded continuous functionals are preserved.

Batched Updates: Setting & Assumptions

Motivation. Updating after every period may be impractical. Instead, commit to one arm for an interval, collect data, then update in batch.

Setup.

- Time horizon n; batches are *pre-determined* (can be adaptive in length selection but fixed before the run).
- Each batch length is $o(n)$; within a batch, the chosen arm is played throughout.
- After each batch, update the algorithm using all rewards gathered in that batch.

Scaling intuition.

- A discrete interval of length $o(n)$ becomes an *infinitesimal* interval after dividing time by n in the limit.
- If the number of batches $\to \infty$ (possibly very slowly) and each batch is $o(n)$, batched and per-period updates should have the same limiting behavior.

Batched Updates: Limit Behavior

Proposition

Under the settings of the main diffusion results, all conclusions continue to hold for the Gaussian Thompson sampler when batch sizes are $o(n)$.

Weak-limit equivalence

- In the small-gap regime, batched Thompson Sampling (with batch sizes $o(n)$) has the *same* SDE / stochastic-ODE limits as ordinary (non-batched) TS.
- Consequently, the **distribution of regret** matches that of per-period updates under the same scaling.

Related Work

Concurrent work: Kuang & Wager (2024) independently developed diffusion approximations for sampling-based algorithms (including TS) in the small-gap regime.

Methodological differences:

- **KW (2024):** Uses Stroock–Varadhan martingale and infinitesimal-generator framework for *Markovian* reward models.
- **This work:** Employs the Continuous Mapping Theorem (CMT) and direct discrete SDE representations, giving an intuitive and general proof structure.

Advantages of the CMT approach:

- Accommodates *stationary, weakly dependent* (non-i.i.d.) reward processes.
- Separates reward-process convergence ($\Rightarrow$ Brownian) from algorithmic sampling dynamics.
- Establishes two-way equivalence between SDE and stochastic ODE forms (Prop. 1 $\Leftrightarrow$ Thm. 3).

Extensions: The framework also shows that EF Thompson and bootstrap samplers share the same diffusion limit and remain robust under mild model mis-specification.