Particle Thompson Sampling with Static Particles

Zeyu Zhou
Department of Radiology
Mayo Clinic
Rochester, MN, USA
zeyuzhou91@gmail.com

Bruce Hajek
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Champaign, IL, USA
b-hajek@illinois.edu

***Abstract*—Particle Thompson sampling (PTS) is a simple and flexible approximation of Thompson sampling for solving stochastic bandit problems. PTS circumvents the intractability of maintaining a continuous posterior distribution in Thompson sampling by replacing the continuous distribution with a discrete distribution supported at a set of weighted static particles. We analyze the dynamics of particles' weights in PTS for general stochastic bandits without assuming that the set of particles contains the unknown system parameter. It is shown that fit particles survive and unfit particles decay, with the fitness measured in KL-divergence. For Bernoulli bandit problems, all but a few fit particles decay.**

***Index Terms*—stochastic bandit, Thompson sampling, particles**

I. Introduction

A bandit problem is a sequential decision problem that elegantly captures the fundamental trade-off between exploitation and exploration. *Thompson sampling (TS)* is a Bayesian heuristic for solving bandit problems with an assumption that the rewards are generated according to a given distribution with a fixed unknown parameter. TS maintains a posterior distribution on the parameter and selects an action according to the posterior probability that the action is optimal. The biggest advantage of TS is its ability to automatically handle setups with a complex information structure, where knowing the performance of one action may inform properties about other actions. Also, it has strong empirical performance [1]. Theoretical performance guarantees of TS have been established for some bandit problems [2]–[5]. However, efficient updating, storing, and sampling from the posterior distribution in TS are only feasible for some special cases (e.g. conjugate distributions). For general bandit problems, one has to resort to various approximations, most of which are complicated and have restrictive assumptions.

Particle Thompson sampling (PTS) is an approximation of TS in which the continuous posterior distribution is replaced by a discrete distribution supported at a set of weighted static particles. Updating the posterior distribution then becomes updating the particles' weights by Bayes formula, followed by normalization. PTS applies to very general bandit setups and is easy to implement. The regret of PTS is analyzed in [5], with the assumption that the finite support set of the prior includes the unknown true system parameter (see Section II for more discussion). However, for PTS when the true parameter exists in a continuum, that assumption is unreasonable. In fact, without that assumption, PTS may be inconsistent, i.e., the running average regret may not converge to zero.

The main contribution of this paper is an analysis of the dynamics of the particles' weights in PTS for general stochastic bandit problems without assuming that the set of particles contains the unknown system parameter. The main result is a drift-based sample-path necessary condition on the surviving particles, illuminating the phenomenon that fit particles survive and unfit particles decay (Proposition 1). A consequent result applied to Bernoulli bandit problems shows that not many particles can survive in PTS with randomly generated particles (Corollary 7). The results shed light on potential improvements of PTS.

The paper is organized as follows. Section II lists some related work. Section III introduces the general setup and notation of stochastic bandit problems and PTS. Section IV provides a sample-path analysis of PTS for general stochastic bandit problems. Section V draws a corollary of PTS for Bernoulli bandit problems. Section VI concludes the paper.

II. Related Work

See [6] and [7] for a survey and recent developments in bandit problems. Upper-confidence-bound (UCB) algorithms have certain theoretical guarantees for some simple bandit models [8], [9]. KL-UCB [9] even meets a lower bound on regret established in [10]. Empirically, UCB algorithms are not very competitive in the non-asymptotic regime due to their inefficient exploration and inability to take advantage of the problem structure for complex bandit problems.

Thompson sampling (TS) [11] has strong empirical performance [1] and can handle rather general and complex stochastic bandit problems [5], [12]. TS can be implemented efficiently in setups where a conjugate prior exists for the reward distribution. In cases where a conjugate prior is not available, one need to resort to approximations of TS, such as Gibbs sampling [13], Laplace approximation [1], Langevin Monte Carlo [14], [15], and boostrapping [16]. These approximations are either complicated, or rely on restrictive assumptions. For example, Laplace approximation requires strict concavity of the log of the posterior distribution and the calculation of its Hessian. The version of boostrapping studied

This work was supported in part by NSF Grant Grant CCF 19-00636. It was done while the first author was a Ph.D. student at the Electrical and Computer Engineering Department of UIUC.

in [16] only applies to Bernoulli bandits and does not naturally generalize to more complex problems. See [12] (Chapter 5) for a detailed discussion of these approximations.

To the best of our knowledge, the term *particle Thompson sampling* first appeared in [17], where the authors apply PTS as an efficient approximation of TS to solve a matrix-factorization recommendation problem. In their work, the particles are not static, but are incrementally re-sampled at each step through an MCMC-kernel. The re-sampling method relies heavily on the specific problem structure. It is not clear how it can be generalized to other bandit problems.

Ensemble sampling [18] is similar to the idea of PTS because it aims to maintain a set of particles (called "models" in the paper) independently and identically sampled from the posterior distribution in order to approximate TS. Particles in ensemble sampling are unweighted. A major restriction of the algorithm is that it requires Gaussian noise in the observation. Also, except in special setups, updating the particles in ensemble sampling requires solving an optimization problem that accounts for all the data from the start to the current time.

The version of PTS in this paper is first considered in [5], which analyzes TS for general stochastic bandit problems. For technical tractability, [5] assumes the prior distribution of the parameter is supported over a finite (possibly huge) set instead of a continuum. Therefore, TS in [5] is tantamount to PTS, with the finite prior support set equivalent to a set of particles. The main result in [5] is that with high probability the number of plays of non-optimal actions is upper bounded by $B + C \log T$, where B, C are problem-dependent constants and T is the time horizon. This result relies on a realizability assumption (called "grain of truth" in the paper): the finite support set of the prior includes the true system parameter. However, for PTS when the true parameter exists in a continuum, the realizability assumption is unreasonable. In our paper, PTS is analyzed without the realizability assumption. The analysis is inspired by [5] on how KL divergence comes into play in the measurement of the fitness of the particles.

III. Setup and Preliminaries

A *stochastic bandit problem* contains the following elements: an action set $\mathcal{A}$, an observation space $\mathcal{Y}$, a parameter space Θ, a known observation model $P_\theta(\cdot|a)$ and a reward function $R : \mathcal{Y} \to \mathbb{R}$. Consider a player who acts at steps $t = 1, 2, \cdots$. At step t, the player takes an action $A_t \in \mathcal{A}$, then observes $Y_t \in \mathcal{Y}$ according to the observation model $P_{\theta^*}(\cdot|A_t)$ for some fixed and unknown $\theta^* \in \Theta$, independent of past observations. The observation Y_t then incurs a reward $R_t = R(Y_t)$. The goal of the player is to maximize the cumulative reward. [1] For notational convenience, we denote an instance of the stochastic bandit problem by $\mathrm{StochasticBandit}(\mathcal{A}, \mathcal{Y}, \Theta, P_\theta(\cdot|a), R, \theta^*)$. Let $\mathcal{H}_t = (A_1, \cdots, A_t, Y_1, \cdots, Y_t)$ denote the history of actions and observations up to time t. An algorithm is a (possibly randomized) mapping from $\mathcal{H}_{t-1}$ to $\mathcal{A}$, for each step t.

Thompson sampling (TS) is an algorithm for solving stochastic bandit problems, shown in Algorithm 1. TS is often difficult to implement in practice because π_t may not have a closed form. Even if a closed form can be obtained, it is not clear how it can be efficiently stored and be sampled from.

Algorithm 1 Thompson sampling

Inputs: $\mathcal{A}, \mathcal{Y}, \Theta, P_\theta(\cdot|a), R, \theta^*$
Initialize: prior π_0 over Θ

1: **for** $t = 1, 2, \cdots$ **do**
2: Sample $\theta_t \sim \pi_{t-1}$
3: Play $A_t \leftarrow \arg\max_{a \in \mathcal{A}} \mathbb{E}_{\theta_t}[R(Y)|A_t = a]$
4: Observe $Y_t \sim P_{\theta^*}(\cdot|A_t)$
5: Update π_t: $\pi_t(\theta) = \frac{P_\theta(Y_t|A_t)\pi_{t-1}(\theta)}{\int_\Theta P_\theta(Y_t|A_t)\pi_{t-1}(\theta)\,\mathrm{d}\theta} \quad \forall \theta \in \Theta.$
6: **end for**

The idea of particle Thompson sampling (PTS) (Algorithm 2) is to approximate π_t by a discrete distribution $w_t = (w_{t,1}, \cdots, w_{t,N})$ supported on a finite set of fixed particles $\mathcal{P}_N = \{\theta^{(1)}, \cdots, \theta^{(N)}\} \subset \Theta$, where N is the number of particles.

Algorithm 2 Particle Thompson sampling

Inputs: $\mathcal{A}, \mathcal{Y}, \Theta, P_\theta(\cdot|a), R, \theta^*, \mathcal{P}_N$
Initialize: $w_0 \leftarrow (\frac{1}{N}, \cdots, \frac{1}{N})$

1: **for** $t = 1, 2, \cdots$ **do**
2: Generate θ_t from $\mathcal{P}_N$ according to weights w_{t-1}
3: Play $A_t \leftarrow \arg\max_{a \in \mathcal{A}} \mathbb{E}_{\theta_t}[R(Y)|A_t = a]$
4: Observe $Y_t \sim P_{\theta^*}(\cdot|A_t)$
5: **for** $i \in \{1, 2, \cdots, N\}$ **do**
6: $\widetilde{w}_{t,i} = w_{t-1,i}\, P_{\theta^{(i)}}(Y_t|A_t)$
7: **end for**
8: $w_t \leftarrow$ normalize $\widetilde{w}_t$
9: **end for**

In practice, one can use a pre-determined set of points $\mathcal{P}_N$ in Θ, or randomly generate some points from Θ. Θ may not contain θ^*. $\widetilde{w}_{t,i}$ is the unnormalized weight of particle i at time t. Step 6 can be alternatively implemented by $\widetilde{w}_{t,i} = \widetilde{w}_{t-1,i} P_{\theta^{(i)}}(Y_t|A_t)$, with the initialization $\widetilde{w}_0 = w_0$, because it yields the same normalized vectors w_t. PTS is very flexible because it does not require any structure on the observation model $P_\theta(\cdot|a)$, as long as the model is given. Steps 5-7 in Algorithm 2 are easy to implement: they require only multiplication and normalization. For notational convenience, we denote an instance of particle Thompson sampling with particle set $\mathcal{P}_N$ by $\mathrm{PTS}(\mathcal{P}_N)$.

IV. PTS for General Stochastic Bandits

This section contains an analysis of PTS for general stochastic bandits. The main result is a sample-path necessary

[1] The problem can be made more general by adding contexts. Let $\mathcal{C}$ be a context set. The observation model becomes $P_\theta(\cdot|a, c)$. At each step of the game, the game player receives an arbitrary context $c_t \in \mathcal{C}$ before taking action A_t. The observation Y_t follows distribution $P_{\theta^*}(\cdot|A_t, c_t)$. This is known as the contextual stochastic bandit model, for which PTS still works. The reason we do not use this more general model here is that we want to emphasize the key word *stochastic*, not contextual.

condition for surviving particles based on drift information.

Let $I_t \in [N]$ be the index of the particle chosen at time t. Thus, $I_t \sim w_{t-1}$. Let $A_t \in \mathcal{A}$ be the arm chosen at time t. Let $A : \Theta \to \mathcal{A}$ be the function mapping from a particle to the corresponding optimal arm, defined by $A(\theta) = \arg\max_{a\in\mathcal{A}} \mathbb{E}_\theta[R(Y)|a]$. If there are multiple maximizers, let $A(\theta)$ be one of them selected deterministically. With a slight abuse of notation, we sometimes abbreviate $A(\theta^{(i)})$ by $A(i)$. So $A_t = A(I_t)$. For any $x \in \mathbb{R}^N$, define $\mathrm{supp}(x) \triangleq \{i \in [N] : x_i \neq 0\}$ and $\arg\max x \triangleq \left\{i \in [N] : x_i = \max_{j\in[N]} x_j\right\}$. Recall from Algorithm 2 that the unnormalized weights of the particles evolve by the equation $\widetilde{w}_{t,i} = \widetilde{w}_{t-1,i} P_{\theta^{(i)}}(Y_t|A_t)$, where $Y_t \sim P_{\theta^*}(\cdot|A_t)$. Let $L_{t,i} \triangleq \ln \widetilde{w}_{t,i} - \ln \widetilde{w}_{t-1,i}$ and $L_t = (L_{t,1}, \cdots, L_{t,N})$.

Definition 1. (Drift matrix) For a given $\mathrm{StochasticBandit}(\mathcal{A}, \Theta, \mathcal{Y}, P_\theta(\cdot|a), R, \theta^*)$ problem and a set of particles $\mathcal{P}_N \subset \Theta$, the *drift matrix* D is an $N \times N$ matrix given by:

$$\begin{aligned} D_{ij} &\triangleq \mathbb{E}\left[\ln \widetilde{w}_{t,j} - \ln \widetilde{w}_{t-1,j} | I_t = i\right] \\ &= \mathbb{E}[\ln P_{\theta^{(j)}}(Y_t|A_t)|I_t = i] \\ &= \mathbb{E}_{Y\sim P_{\theta^*}(\cdot|A(i))}\left[\ln P_{\theta^{(j)}}(Y|A(i))\right] \end{aligned}$$

for $i, j \in [N]$. In words, D_{ij} is the (exponential) drift of particle j when particle i is chosen.

The following properties of D are readily verified: 1) Entries in D are non-positive; 2) D is independent of time, fundamentally because $\{\widetilde{w}_t\}$ is a time-homogeneous Markov process; 3) Row i_1 and row i_2 of D are the same if $A(i_1) = A(i_2)$. Therefore D can have at most $|\mathcal{A}|$ distinct rows. In what follows we consider drift matrices D and D' to be equivalent if each row in D' is equal to the corresponding row of D up to an additive constant. Therefore, D remains in the same equivalence class if for each i the constant $-\mathbb{E}\left[\ln P_{\theta^*}(Y|A(i))\right]$ is added to row i. Therefore, a representative choice of D is the following:

$$\begin{aligned} D_{ij} &\overset{\text{equivalent}}{=} -\mathbb{E}_{Y\sim P_{\theta^*}(\cdot|A(i))}\left[\ln \frac{P_{\theta^*}(Y|A(i))}{P_{\theta^{(j)}}(Y|A(i))}\right] \\ &= -\mathrm{KL}\left(P_{\theta^*}(\cdot|A(i)) \,\middle|\middle|\, P_{\theta^{(j)}}(\cdot|A(i))\right). \end{aligned}$$

Here D_{ij} is the negative of KL divergence between distributions $P_{\theta^*}(\cdot|A(i))$ and $P_{\theta^{(j)}}(\cdot|A(i))$. In this sense, the ith row of D gives the relative fitness of the particles for action $A(i)$, and the j^{th} column of D gives the fitness of particle j for action $A(i)$ varying over all i.

We need the following two assumptions due to technical tractability for the main result.

Assumption 1 (Sample path assumptions)**.** Consider the problem $\mathrm{StochasticBandit}(\mathcal{A}, \Theta, \mathcal{Y}, P_\theta(\cdot|a), R, \theta^*)$ and suppose $\mathrm{PTS}(\mathcal{P}_N)$ is run for a set of N particles $\mathcal{P}_N \subset \Theta$. Assume that the sample path satisfies the following: there exists a non-empty set $S \subset [N]$ that satisfies

(a) (Non-zero decaying rate gap) For any $i \notin S$ and $j \in S$, $\limsup_{t\to\infty} \frac{1}{t}\left(\ln \widetilde{w}_{t,i} - \ln \widetilde{w}_{t,j}\right) < 0$, and

(b) (Existence of survivor limiting distribution) $G_t = (\ln \widetilde{w}_{t,i} - \ln \widetilde{w}_{t,j} : i, j \in S) \in \mathbb{R}^{|S|\times|S|}$ has a limiting empirical distribution μ_G. In other words, for any bounded continuous function h on $\mathbb{R}^{|S|\times|S|}$, $\frac{1}{t}\sum_{\tau=0}^{t} h(G_\tau) \to \mathbb{E}_{\mu_G}[h]$.

(c) $\left|\frac{1}{t}\sum_{\tau=1}^{t} \mathbb{1}_{\{I_\tau = i\}} - \frac{1}{t}\sum_{\tau=0}^{t-1} w_{\tau,i}\right| \to 0$ as $t \to \infty$ for any $i \in [N]$.

(d) For any $i \in [N]$ that is used infinitely many times, $\frac{1}{M}\sum_{m=1}^{M} L_{t_i(m)} \to D_i$ as $M \to \infty$, where $t_i(m)$ is the mth time particle i is chosen and D_i is the ith row of the drift matrix D.

The set S can be thought of as the set of surviving particles. Assumption 1(a) says the (unnormalized) weight decaying rate of a non-surviving particle is *strictly less* than that of a surviving particle. Consequently, the weight of a non-surviving particle converges to 0 exponentially fast. Assumption 1(b) says that the process G_t has some ergodicity property. It is similar to saying that G_t is Harris recurrent, except G_t is not Markov, because it excludes information about particles not in S. Note that knowing any row of G_t determines all the other entries of G_t. [2] In Assumption 1(c), $\mathbb{1}_{\{I_\tau=i\}}$ is a Bernoulli random variable with mean $w_{\tau-1,i}$ for each τ. Therefore Assumption 1(c) holds with probability one by the Azuma-Hoeffding inequality. Assumption 1(d) holds with probability one by the definition of D and the strong law of large numbers.

Assumption 2 (Boundedness of observation model)**.** Assume that the observation model $P_\theta(\cdot|a)$ satisfies: there exists constants $b_0, B_0 > 0$, such that for any $\theta, \theta' \in \Theta$, $b_0 \leq \frac{P_\theta(y|a)}{P_{\theta'}(y|a)} \leq B_0$ for any $y \in \mathcal{Y}, a \in \mathcal{A}$.

The assumption can be easily verified for problems in which $|\mathcal{Y}| < \infty$ and $|\mathcal{A}| < \infty$.

Define a probability vector π over $[N]$ by $\pi_i = \lim_{t\to\infty} \frac{1}{t+1}\sum_{\tau=0}^{t} w_{\tau,i}$. That is, π_i is the limiting running average weight of particle i, if it exists. Proposition 1 shows the relationship between π and the drift matrix D and provides a necessary condition for surviving particles in a sample path.

Proposition 1 (Sample-path necessary surviving condition)**.** *Let* $\mathrm{StochasticBandit}(\mathcal{A}, \Theta, \mathcal{Y}, P_\theta(\cdot|a), R, \theta^*)$ *be a given problem and* $\mathcal{P}_N \subset \Theta$ *a given set of* N *particles. Suppose* $P_\theta(\cdot|a)$ *satisfies Assumption 2. Let* D *be the drift matrix. Consider running* $\mathrm{PTS}(\mathcal{P}_N)$ *for the problem. For a sample path of the algorithm under Assumption 1,* π *is well defined and satisfies*

$$\arg\max(\pi D) = \mathrm{supp}(\pi) = S\,, \tag{1}$$

where S *is the set in Assumption 1.*

[2] Remark on Assumption 1(a-b): Roughly speaking, one may show that $\frac{1}{t}\ln \widetilde{w}_{t,j}$ converges to an I_t-dependent weighted average of the j^{th} column of D, i.e. a sample path average fitness. Assumption 1(a) holds when the sample path average fitness of each non-surviving particle i is less than that of each surviving particle j. There are setups in which the weights of the surviving particles oscillates forever and do not converge. But as long as their ratios are stochastically bounded, Assumption 1(b) holds. More intuition and evidence for Assumption 1(a-b) can be found in the analysis of PTS for the two-arm Bernoulli bandit problem ([19], Appendix B.1-B.3).

The proposition says that, if a set of particles S were to survive in a sample path, they must have a limiting average selection distribution π that satisfies (1). The jth coordinate of πD, $(\pi D)_j$, is equal to $\langle \pi, D_{\cdot j} \rangle$, where $D_{\cdot j} = (D_{1j}, \cdots, D_{Nj})$ is the jth column of D, the drifts of particle j when particles $1, 2, \cdots, N$ are chosen, which we recall can be interpreted as the fitness of particle j. Thus, $(\pi D)_j$ is the average fitness of particle j, assuming distribution π is used to select a random action $A(i)$. Therefore, (1) means that, with respect to distribution π, each surviving particle has the same average fitness, and the average fitness of each non-surviving particle is strictly smaller. In this sense, *fit particles survive, unfit particles decay*. Note the following caveat: Proposition 1 is a sample-path result. The actual set of survivors may be random. Thus, there may be more than one π that satisfies (1). If the surviving particles do not all induce the global optimal action $A(\theta^*)$, then the cumulative regret will grow linearly over time. Thus, in general PTS can be inconsistent.

The rest of this section is the proof of Proposition 1. All the lemmas in this section deal with a sample path under Assumption 1.

Lemma 2. *The probability vector π is well defined. In addition, $\mathrm{supp}(\pi) = S$. That is, if $i \notin S$, then $\pi_i = 0$; if $i \in S$, then $\pi_i > 0$.*

Proof. For $i \notin S$,

$$w_{t,i} = \frac{\widetilde{w}_{t,i}}{\sum_{j=1}^N \widetilde{w}_{t,j}} = \frac{e^{\ln \widetilde{w}_{t,i}}}{\sum_{j=1}^N e^{\ln \widetilde{w}_{t,j}}} \le \frac{e^{\ln \widetilde{w}_{t,i}}}{e^{\ln \widetilde{w}_{t,j_0}}}$$

for any $j_0 \in S$. By Assumption 1(a), $w_{t,i} \to 0$. Hence $\pi_i = \lim_{t\to\infty} \frac{1}{t+1} \sum_{\tau=0}^t w_{t,i} = 0$.

Next, define

$$w'_{t,i} \triangleq \begin{cases} 0 & if \quad i \notin S \\ \frac{w_{t,i}}{\sum_{j\in S} w_{t,j}} & if \quad i \in S \end{cases} .$$

Fix $i \in S$.

$$w'_{t,i} - w_{t,i} = w_{t,i} \left(\frac{1}{\sum_{j\in S} w_{t,j}} - 1 \right) = w_{t,i} \frac{\sum_{j\notin S} w_{t,j}}{\sum_{j\in S} w_{t,j}}$$
$$= w_{t,i} \frac{\sum_{j\notin S} w_{t,j}}{1 - \sum_{j\notin S} w_{t,j}} .$$

Since the set $[N]\backslash S$ is finite, $\sum_{j\notin S} w_{t,j} \to 0$. It follows that $w'_{t,i} - w_{t,i} \to 0$. Hence

$$\frac{1}{t+1} \sum_{\tau=0}^t w'_{\tau,i} - \frac{1}{t+1} \sum_{\tau=0}^t w_{\tau,i} \to 0 . \tag{2}$$

Now, observe that $w'_{t,i}$ can be determined from $\{\ln \widetilde{w}_{t,j}\}_{j\in S}$ by $w'_{t,i} = \frac{e^{\ln \widetilde{w}_{t,i}}}{\sum_{j\in S} e^{\ln \widetilde{w}_{t,j}}}$. Therefore, $w'_{t,i}$ is a continuous and bounded function of $\{\ln \widetilde{w}_{t,j}\}_{j\in S}$, and hence of G_t. We write this as $w'_{t,i} = w'_i(G_t)$. According to Assumption 1(b),

$$\frac{1}{t+1} \sum_{\tau=0}^t w'_{\tau,i} \to \mathbb{E}_{\mu_G}[w'_i] . \tag{3}$$

Combining (2) and (3), we obtain $\pi_i = \mathbb{E}_{\mu_G}[w'_i]$. Since w'_i is a positive function and μ_G is a distribution, we conclude that $\pi_i > 0$ for $i \in S$.

Finally,

$$\sum_{i\in[N]} \pi_i = \sum_{i\in[N]} \lim_{t\to\infty} \frac{1}{t+1} \sum_{\tau=0}^t w_{\tau,i}$$
$$\overset{(i)}{=} \lim_{t\to\infty} \sum_{i\in[N]} \frac{1}{t+1} \sum_{\tau=0}^t w_{\tau,i}$$
$$= \lim_{t\to\infty} \frac{1}{t+1} \sum_{\tau=0}^t \sum_{i\in[N]} w_{\tau,i} = \lim_{t\to\infty} 1 = 1 ,$$

where in step (i) we switch the limit and summation because all summands are non-negative and N is finite. Thus π is well defined. □

Lemma 3. *$\frac{1}{t} \sum_{\tau=1}^t L_\tau \to \pi D$ as $t \to \infty$.*

Proof. Let $M_i(t)$ be the number of times particle i has been played up to time t. Let $\tau_i(m)$ be the mth time that particle i is played. Then

$$\frac{1}{t} \sum_{\tau=1}^t L_\tau = \frac{1}{t} \sum_{i=1}^N \sum_{m=1}^{M_i(t)} L_{\tau_i(m)}$$
$$= \sum_{i=1}^N \frac{M_i(t)}{t} \frac{1}{M_i(t)} \sum_{m=1}^{M_i(t)} L_{\tau_i(m)} .$$

Since $M_i(t) = \sum_{\tau=1}^t \mathbb{1}_{\{I_\tau = i\}}$, by Assumption 1(c) and the definition of π_i, $\frac{M_i(t)}{t} \to \pi_i$ for all $i \in [N]$. If particle i is played infinitely many times in the sample path, then $\frac{1}{M_i(t)} \sum_{m=1}^{M_i(t)} L_{\tau_i(m)} \to D_i$ as $t \to \infty$ by Assumption 1(d). If particle i is played finitely many times, thus $M_i(t) \le C$ for some constant C for all t, then $\frac{M_i(t)}{t} \to 0$ and $\lim_{t\to\infty} \frac{1}{M_i(t)} \sum_{m=1}^{M_i(t)} L_{\tau_i(m)} < \infty$. Either case, we have

$$\frac{M_i(t)}{t} \frac{1}{M_i(t)} \sum_{m=1}^{M_i(t)} L_{\tau_i(m)} \to \pi_i D_i \quad \text{as} \quad t \to \infty .$$

It follows that

$$\frac{1}{t} \sum_{\tau=1}^t L_\tau \to \sum_{i=1}^N \pi_i D_i = \pi D \quad \text{as} \quad t \to \infty .$$

□

Lemma 4. *If a real-valued sequence $\{x_t\}_{t\ge1}$ satisfies*

(1) *$\{x_t\}$ has a limiting distribution μ.*
(2) *$\{x_t\}$ is B-Lipschitz: there exists some constant B such that $|x_t - x_s| \le B\,|t-s|$ for all $t, s \in \mathbb{N}^+$.*

Then $\lim_{t\to} \frac{1}{t} x_t = 0$.

Proof. We show $\limsup_{t\to\infty} \frac{1}{t} x_t \le \delta$ for any $\delta > 0$. Suppose there exists $\delta > 0$ such that $\limsup_{t\to\infty} \frac{1}{t} x_t > \delta$. Condition (1) implies that, there exists $c \in \mathbb{R}$ such that

$$\frac{1}{t} \sum_{\tau=1}^t \mathbb{1}_{\{x_\tau \ge c\}} \le \frac{\delta}{2B} \quad \text{for all } t \text{ sufficiently large} . \tag{4}$$

Let $\{t_1, t_2, \cdots, t_n, \cdots\}$ be a sequence of positive integers such that $\lim_{n\to\infty} t_n = \infty$ and $\frac{1}{t_n} x_{t_n} \geq \delta$ for all n. Thus $x_{t_n} \geq \delta t_n$ for all n. Since $\{x_t\}$ is B-Lipschitz, for any $t \in [1, t_n]$,

$$x_t \geq x_{t_n} - B(t_n - t) \geq \delta t_n - B(t_n - t) = Bt - (B - \delta)t_n \,.$$

It follows that, if $t \geq \frac{c}{B} + \left(1 - \frac{\delta}{B}\right) t_n$, then $x_t \geq c$. Therefore, for $t_n > \frac{2c}{\delta}$,

$$\begin{aligned}
\frac{1}{t_n} \sum_{\tau=1}^{t_n} \mathbb{1}_{\{x_\tau \geq c\}} &\geq \frac{1}{t_n} \sum_{\tau=1}^{t_n} \mathbb{1}_{\left\{\tau \geq \frac{c}{L} + \left(1 - \frac{\delta}{L}\right) t_n\right\}} \\
&= \frac{1}{t_n} \left[t_n - \left(\frac{c}{B} + \left(1 - \frac{\delta}{B}\right) t_n \right) \right] \\
&= \frac{\delta}{B} - \frac{c}{B t_n} > \frac{\delta}{2B} \,,
\end{aligned}$$

which contradicts (4). Therefore, $\limsup_t \frac{1}{t} x_t \leq \delta$ for any $\delta > 0$. Similarly, we can show that $\liminf_{t\to\infty} \frac{1}{t} x_t \geq -\delta$ for any $\delta > 0$. We conclude that $\lim_{t\to\infty} \frac{1}{t} x_t = 0$. □

Lemma 5. *If $i, j \in S$, then $(\pi D)_i = (\pi D)_j$.*

Proof. Consider $i, j \in S$. Then

$$\begin{aligned}
&\frac{1}{t} \sum_{\tau=1}^{t} L_{\tau,i} - \frac{1}{t} \sum_{\tau=1}^{t} L_{\tau,j} = \frac{1}{t} \sum_{\tau=1}^{t} (L_{\tau,i} - L_{\tau,j}) \\
&= \frac{1}{t} \sum_{\tau=1}^{t} \left[(\ln \widetilde{w}_{\tau,i} - \ln \widetilde{w}_{\tau-1,i}) - (\ln \widetilde{w}_{\tau,j} - \ln \widetilde{w}_{\tau-1,j}) \right] \\
&= \frac{1}{t} \left[(\ln \widetilde{w}_{t,i} - \ln \widetilde{w}_{0,i}) - (\ln \widetilde{w}_{t,j} - \ln \widetilde{w}_{0,j}) \right] \\
&= \frac{1}{t} \left(\ln \widetilde{w}_{t,i} - \ln \widetilde{w}_{t,j} \right) = \frac{1}{t} G_t(i,j) \,.
\end{aligned}$$

The third equality above used $\ln \widetilde{w}_{0,i} = \ln \widetilde{w}_{0,j} = 0$ by initialization (although that is not important, as long as the difference is finite). By the dynamics of the weights $\{w_{t,i}\}$ and $\{w_{t,j}\}$, we have that

$$G_{t+1}(i,j) = G_t(i,j) + \ln \frac{P_{\theta^{(i)}}(Y_{t+1} | A_{t+1})}{P_{\theta^{(j)}}(Y_{t+1} | A_{t+1})} \,.$$

By Assumption 2, $|G_{t+1}(i,j) - G_t(i,j)| \leq B$, where $B = \max\{|\ln b_0|, |\ln B_0|\}$. Thus $\{G_t(i,j)\}_{t\geq 1}$ is an B-Lipschitz sequence. Therefore

$$\begin{aligned}
(\pi D)_i - (\pi D)_j &\overset{(i)}{=} \lim_{t\to\infty} \left(\frac{1}{t} \sum_{\tau=1}^{t} L_{\tau,i} - \frac{1}{t} \sum_{\tau=1}^{t} L_{\tau,j} \right) \\
&= \lim_{t\to\infty} \frac{1}{t} G_t(i,j) \overset{(ii)}{=} 0 \,,
\end{aligned}$$

where equality (i) is due to Lemma 3 and equality (ii) equality is due to Lemma 4 and Assumption 1(b). □

Lemma 6. *If $i \notin S$ and $j \in S$, then $(\pi D)_i < (\pi D)_j$.*

Proof. Similar to the proof of Lemma 5, we have

$$\frac{1}{t} \sum_{\tau=1}^{t} L_{\tau,i} - \frac{1}{t} \sum_{\tau=1}^{t} L_{t,j} = \frac{1}{t} \left(\ln \widetilde{w}_{t,i} - \ln \widetilde{w}_{t,j} \right)$$

The LHS converges to $(\pi D)_i - (\pi D)_j$ as $t \to \infty$ by Lemma 2. The RHS converges to a strictly negative value as $t \to \infty$ by Assumption 1(a). Thus $(\pi D)_i < (\pi D)_j$. □

Proof of Proposition 1. Lemma 2 shows $\mathrm{supp}(\pi) = S$. Lemma 5 and Lemma 6 show $\arg\max(\pi D) = S$. Proposition 1 is thus proved.

□

V. PTS for Bernoulli Bandits

Let K be a positive integer. A Bernoulli bandit problem depicts a player who picks an arm indexed by $a \in \{1, \cdots, K\}$ at each step, which generates a reward of either 0 or 1 according to a Bernoulli distribution parameterized by $\theta_a^* \in [0,1]$, fixed and unknown. This is a stochastic bandit problem with $\mathcal{A} = \{1, 2, \cdots, K\}$, $\mathcal{Y} = \{0, 1\}$, $\Theta = [0,1]^K$, $P_\theta(\cdot|a) \sim \mathrm{Bernoulli}(\theta_a)$, and $R(y) = y$. This is a bandit problem with separable actions – the observation model for each action is parametrized by a corresponding coordinate of θ^*.

Applying Proposition 1 to Bernoulli bandit with randomly generated particles in PTS yields the following corollary that says that not many particles can survive.

Corollary 7. *Let $\mathcal{P}_N$ be a set of N points generated independently and uniformly at random from $[0,1]^K$. Consider running $\mathrm{PTS}(\mathcal{P}_N)$ for a given Bernoulli bandit problem with K arms and with $\theta^* \in [0,1]^K$. Suppose that any sample path satisfies Assumption 1. Then with probability one, at most K particles can survive, i.e. $|\mathrm{supp}(\pi)| \leq K$.*

We suspect that something similar can be said about the fewness of survivors for other bandit problems in which the action space $\mathcal{A}$ has a finite dimension K (the total number of actions $|\mathcal{A}|$ may be much larger). But we don't have a proof.

For more evidence and intuition of Proposition 1 and Corollary 7, see [19] (Appendix B), where a thorough analysis of PTS for two-arm Bernoulli bandit is provided. The rest of this section is about the proof of Corollary 7.

Proof of Corollary 7. If $N \leq K$, then $|\mathrm{supp}(\pi)| \leq N \leq K$ trivially. Let $N > K$. The observation model of a Bernoulli bandit problem satisfies Assumption 2 trivially. By Proposition 1, with probability one, for any sample path, the probability vector π is well-defined and π and S satisfy $\arg\max(\pi D) = \mathrm{supp}(\pi) = S$, which implies the following constraints on π:

$$\begin{aligned}
&\pi_i = 0 \text{ for } i \notin S \,, \\
&(\pi D)_i = (\pi D)_j \text{ for all } i, j \in S \,,
\end{aligned} \tag{5}$$

where S is the subset of $[N]$ in Assumption 1. Suppose $|S| > K$. The remainder of the proof shows that, with probability one, any π that satisfies (5) is the all-zero vector (thus π cannot be a probability vector). This leads to a contradiction with $|S| > K$ and therefore we conclude that $|S| \leq K$.

We construct a matrix $\widetilde{D} \in \mathbb{R}^{K \times N}$ and a probability (row) vector $\widetilde{\pi} \in [0,1]^K$ from D and π, as follows.

Recall that, row i_1 and row i_2 of D are the same if $A(i_1) = A(i_2)$. Since there are K arms, there can be at most K unique

rows in D. Let $\widetilde{D}$ be D reduced to its unique K rows. That is, $\widetilde{D}_k = \mathbb{E}[L_t | A_t = k]$ (which is independent of t) for $k \in [K]$.

For $k \in [K]$, let $\widetilde{\pi}_k = \sum_{i:i\in S, A(i)=k} \pi_i$. That is, $\widetilde{\pi}_k$ is the sum of the asymptotic weights of surviving particles with the optimal arm k. If no $i \in S$ satisfies $A(i) = k$, then $\widetilde{\pi}_k = 0$. It is easy to verify that $\widetilde{\pi}_1 + \cdots + \widetilde{\pi}_K = 1$.

Now, observe that,

$$\begin{aligned}\pi D &= \sum_{i=1}^{N} \pi_i D_i = \sum_{i\in S} \pi_i D_i = \sum_{k=1}^{K} \sum_{i:i\in S, A(i)=k} \pi_i D_i \\ &= \sum_{k=1}^{K} \sum_{i:i\in S, A(i)=k} \pi_i \widetilde{D}_k = \sum_{k=1}^{K} \left(\sum_{i:i\in S, A(i)=k} \pi_i \right) \widetilde{D}_k \\ &= \sum_{k=1}^{K} \widetilde{\pi}_k \widetilde{D}_k = \widetilde{\pi} \widetilde{D} \,.\end{aligned}$$

Therefore, the constraints (5) on π imply the following constraints on $\widetilde{\pi}$:

$$(\widetilde{\pi}\widetilde{D})_i = (\widetilde{\pi}\widetilde{D})_j \text{ for all } i, j \in S \,. \tag{6}$$

Let $\widetilde{D}_i$ be the ith column of $\widetilde{D}$. Then $(\widetilde{\pi}\widetilde{D})_i = \left\langle \widetilde{\pi}, \widetilde{D}_i \right\rangle$. Constraints (6) can thus be re-written as

$$\left\langle \widetilde{\pi}, \widetilde{D}_i - \widetilde{D}_j \right\rangle = 0 \text{ for all } i, j \in S \,. \tag{7}$$

For a Bernouli bandit problem, the entries in $\widetilde{D} = [\widetilde{D}_{kj}]_{1\le k\le K, 1\le j\le N}$ are in the form $\widetilde{D}_{kj} = -d(\theta_k^* || \theta_k^{(j)})$, where $d(x||y) = x \ln \frac{x}{y} + (1-x) \ln \frac{1-x}{1-y}$ for $x, y \in [0,1]$ and $\theta_k^{(j)}$ is uniformly distributed in $[0,1]$ and is independent across $k \in [K]$ and $j \in [N]$. Therefore, since $|S| > K$, with probability one, the set of vectors $\{\widetilde{D}_i - \widetilde{D}_j : i, j \in S\}$ spans $\mathbb{R}^K$, in which case the only $\widetilde{\pi} \in \mathbb{R}^K$ that satisfies (7) is the all-zero vector. By construction of $\widetilde{\pi}$, with probability one, the only vector $\pi \in \mathbb{R}^N$ that satisfies (5) is the all-zero vector. □

VI. Conclusion and Discussion

This paper analyzes the particle dynamics of PTS for general stochastic bandit problems. It shows a sample-path particle surviving condition, illuminating the phenomenon that fit particles survive and unfit particles decay. Applying PTS with randomly generated particles to Bernoulli bandits with K arms, it is shown that no more than K particles can survive.

The fitness of a particle is measured in terms of KL divergence with respect to the true system parameter θ^*, which in many particular problems translates to spatial closeness to θ^*. Since results in this paper suggest that not many particles can survive, continuing using these particles after their weights become negligible is a waste of computational resources. One possible improvement of PTS is to periodically delete unfit decaying particles and regenerate new particles that are spatially close to the fit surviving particles. In this way, the set of particles may get closer to θ^* and induce actions closer to the global optimal action $A(\theta^*)$. We leave this as future work. Also, the necessary survival condition in Proposition 1 may be further explored to provide insight on which particles can survive for some specific bandit problems.

References

[1] O. Chapelle and L. Li, "An empirical evaluation of Thompson sampling," in *Advances in Neural Information Processing Systems 24*. Curran Associates, Inc., 2011, pp. 2249–2257.

[2] E. Kaufmann, N. Korda, and R. Munos, "Thompson sampling: An asymptotically optimal finite-time analysis," in *Algorithmic Learning Theory*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 199–213.

[3] S. Agrawal and N. Goyal, "Analysis of Thompson sampling for the multi-armed bandit problem," in *Proceedings of the 25th Annual Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 23. Edinburgh, Scotland: PMLR, 25–27 Jun 2012, pp. 39.1–39.26.

[4] ——, "Thompson sampling for contextual bandits with linear payoffs," in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ser. ICML'13. JMLR.org, 2013, pp. III–1220–III–1228.

[5] A. Gopalan, S. Mannor, and Y. Mansour, "Thompson sampling for complex online problems," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ser. ICML'14. JMLR.org, 2014, pp. I–100–I–108.

[6] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends® in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012. [Online]. Available: http://dx.doi.org/10.1561/2200000024

[7] T. Lattimore and C. Szepesvári, *Bandit Algorithms*, 1st ed. Cambridge University Press, 2019. [Online]. Available: https://tor-lattimore.com/downloads/book/book.pdf

[8] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2–3, p. 235–256, May 2002.

[9] A. Garivier and O. Cappé, "The KL-UCB algorithm for bounded stochastic bandits and beyond," in *Proceedings of the 24th Annual Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 19. Budapest, Hungary: PMLR, 09–11 Jun 2011, pp. 359–376.

[10] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[11] W. R. Thompson, "On the theory of apportionment," *American Journal of Mathematics*, vol. 57, no. 2, pp. 450–456, 1935.

[12] D. J. Russo, B. V. Roy, A. Kazerouni, I. Osband, and Z. Wen, "A tutorial on thompson sampling," *Foundations and Trends® in Machine Learning*, vol. 11, no. 1, pp. 1–96, 2018. [Online]. Available: http://dx.doi.org/10.1561/2200000070

[13] G. Casella and E. I. George, "Explaining the gibbs sampler," *The American Statistician*, vol. 46, no. 3, pp. 167–174, 1992.

[14] G. O. Roberts and R. L. Tweedie, "Exponential convergence of langevin distributions and their discrete approximations," *Bernoulli*, vol. 2, no. 4, pp. 341–363, 1996. [Online]. Available: http://www.jstor.org/stable/3318418

[15] J. Mattingly, A. Stuart, and D. Higham, "Ergodicity for sdes and approximations: locally lipschitz vector fields and degenerate noise," *Stochastic Processes and their Applications*, vol. 101, no. 2, pp. 185–232, 2002.

[16] D. Eckles and M. Kaptein, "Thompson sampling with the online bootstrap," 2014. [Online]. Available: https://arxiv.org/abs/1410.4009

[17] J. Kawale, H. H. Bui, B. Kveton, L. Tran-Thanh, and S. Chawla, "Efficient Thompson Sampling for online matrix-factorization recommendation," in *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc., 2015, pp. 1297–1305.

[18] X. Lu and B. Van Roy, "Ensemble sampling," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. USA: Curran Associates Inc., 2017, pp. 3260–3268.

[19] Z. Zhou, B. Hajek, N. Choi, and A. Walid, "Regenerative particle thompson sampling," *arXiv preprint arXiv:2203.08082*, 2022.