Research Article

Combining Multiple Strategies for Multiarmed Bandit Problems and Asymptotic Optimality

Hyeong Soo Chang and Sanghee Choe

Department of Computer Science and Engineering, Sogang University, Seoul 121-742, Republic of Korea

Correspondence should be addressed to Hyeong Soo Chang; hschang@sogang.ac.kr

Received 13 January 2015; Accepted 11 March 2015

Academic Editor: Shengwei Mei

Copyright © 2015 H. S. Chang and S. Choe. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This brief paper provides a simple algorithm that selects a strategy at each time in a given set of multiple strategies for stochastic multiarmed bandit problems, thereby playing the arm by the chosen strategy at each time. The algorithm follows the idea of the probabilistic ϵ_t -switching in the ϵ_t -greedy strategy and is asymptotically optimal in the sense that the selected strategy converges to the best in the set under some conditions on the strategies in the set and the sequence of $\{\epsilon_t\}$.

1. Introduction

This paper considers the problem of stochastic non-Bayesian multiarmed bandit (MAB) in which a player with a bandit has to decide which arm to play at each time among available arms to maximize the sum of rewards earned through a sequence of playing arms. When played, each arm provides a random reward from an unknown distribution specific to that arm. The problem models the well-known trade-off between “exploration” and “exploitation” in sequential learning. The player needs to obtain new knowledge (exploration) and at the same time optimize her decisions based on existing knowledge (exploitation). The player attempts to balance these competing tasks in order to achieve the goal. Many practical problems, for example, in networking [1, 2], in games [3], and in prediction [4], and problems such as clinical trials and ad placement on the Internet (see, e.g., [1, 5, 6] and the references therein) have been studied with a (properly extended) model of the MAB problems.

Specifically, we consider a stochastic K -armed bandit problem where there is a finite set of arms $A = \{1, 2, \dots, K\}$, $K > 1$ and one arm in A needs to be sequentially played. When an arm $a \in A$ is played at time $t \geq 1$, the player obtains a sample bounded reward $X_{a,t} \in \mathfrak{R}$ drawn from an unknown distribution associated with a , whose unknown expectation and variance are μ_a and σ_a^2 , respectively. We define a strategy $\pi = \{\pi_t, t = 1, 2, \dots\}$ as a sequence of mappings such that

π_t maps from the set of past plays and rewards, $H_{t-1} := (A \times \mathfrak{R})^{t-1}$, to the set of all possible distributions over A where H_0 is an arbitrarily given nonempty subset of $\mathfrak{R}$. We denote the set of all possible strategies as Π . Given a particular sequence $h_{t-1}^\pi \in H_{t-1}$ of the past plays and rewards obtained by following $\pi \in \Pi$ over $t-1$ time steps, π selects $a \in A$ to be played by probability $\pi_t(h_{t-1}^\pi)(a)$ at time t . We assume that $\pi_1(h_0^\pi)$ is arbitrarily given.

Let random variable I_t^π denote the arm selected by π at time t and let $P_a^\pi(t)$ be the distribution over A given by π at time t so that $P_a^\pi(t) = \Pr\{I_t^\pi = a\}$, which is equal to $\sum_{h_{t-1}^\pi \in H_{t-1}} \pi_t(h_{t-1}^\pi)(a) \Pr\{h_{t-1}^\pi\}$. We assume that $X_{a,t}$ and $X_{a',t'}$ are independent of any a, a' and $t \neq t'$ and $X_{a,t}$'s for $t \geq 1$ are identically distributed for any fixed a in A .

Let $\mu^* = \max_{a \in A} \mu_a$ and $A^* = \{a \in A \mid \mu_a = \mu^*\}$. For a given $\pi \in \Pi$, if $\sum_{a \in A \setminus A^*} P_a^\pi(t) \rightarrow 0$ as $t \rightarrow \infty$, then we say that π is an *asymptotically optimal* strategy. The notion of the asymptotic optimality was introduced by Robbins [5]. He presented a strategy which achieves the optimality for the $K = 2$ case where in a single play, each $a \in A$ produces a reward of 1 or 0 with unknown probabilities p_a and $1 - p_a$, respectively. Bather [7] considered the same Bernoulli problem with $K \geq 2$ and established an asymptotically optimal index-based strategy π such that, at time t , π selects an arm in $\arg \max_{a \in A} \{\bar{X}_{a, T_a^\pi(t-1)}^\pi + \lambda(T_a^\pi(t-1))Z_a(t)\}$, where $T_a^\pi(t-1)$ denotes the number of times arm a has been played

by π during the first $t - 1$ plays, and $\bar{X}_{a,m}^\pi, m \geq 1$, denotes the average of the m reward samples obtained by playing a when π is followed, and $\{\lambda(n), n \geq 1\}$ is a sequence of strictly positive constants such that $\lambda(t) \rightarrow 0$ as $t \rightarrow \infty$, and $Z_a(t), a \in A, t \geq K$, are i.i.d. positive and unbounded random variables whose common distribution function F satisfies that $F(0) = 0$ and $F(z) < 1$, for all $z > 0$. The idea is to ensure that each arm is played infinitely often by adding small perturbations to $\bar{X}_{a,T_a^\pi(t-1)}^\pi$ and to make the effect vanish as $T_a^\pi(t-1)$ increases. The well-known asymptotically optimal ϵ_t -greedy strategy [8] with $\sum_{t=1}^\infty \epsilon_t = \infty$ and $\lim_{t \rightarrow \infty} \epsilon_t = 0$ follows exactly the same idea: at time t , with probability $\epsilon_t \in (0, 1]$, it selects $a \in A$ with probability $1/K$ and, with probability $1 - \epsilon_t$, it selects an arm in $\arg \max_{a \in A} \{\bar{X}_{a,T_a^\pi(t-1)}^\pi\}$ where π refers to the ϵ_t -greedy strategy.

This brief paper provides a randomized algorithm ϵ_t -comb(Φ) which follows the spirit of the ϵ_t -greedy strategy for combining multiple strategies in a given finite nonempty $\Phi \subset \Pi$. At each time t , we use the probabilistic ϵ_t -switching to select uniformly a strategy in Φ or to select a strategy with the highest sample average of the rewards obtained so far by playing the bandit. Once a strategy ϕ is selected, the arm chosen by ϕ is played for the bandit. Analogous to the case of the ϵ_t -greedy strategy, it is asymptotically optimal in the sense that the selected strategy converges to the “best” in the set under some conditions on the strategies in Φ and $\{\epsilon_t\}$.

2. Related Work

In the following, we briefly summarize the most relevant works in the literature with the results of the present paper. A seminal work by Gittins and Jones [9] provides an optimal policy (or allocation index rule) to maximize the discounted reward over an infinite horizon when the rewards are given by Markov chains whose statistics are perfectly known in advance. Note that our model does not consider discounting in the rewards and assumes that the relevant statistics are unknown.

Auer et al. [10] presented an algorithm, called Exp4, which combines multiple strategies in a *nonstochastic* bandit setting. In the nonstochastic MAB, it is assumed that each arm is *initially assigned* an arbitrary and unknown sequence of rewards, one for each time step. In other words, the rewards obtained by playing a specific arm are predetermined. In Exp4, the “uniform expert,” which always selects an action uniformly over A , needs to be always included in Φ . At time t , Exp4 computes

$$\begin{aligned} \text{Exp4}_t(h_{t-1}^{\text{Exp4}})(a) &= (1 - \gamma) \sum_{\phi \in \Phi} w^\phi(t) \phi_t(h_{t-1}^{\text{Exp4}})(a) \\ &+ \frac{\gamma}{|\Phi|}, \quad a \in A, \end{aligned} \quad (1)$$

where h_{t-1}^{Exp4} is obtained from the predetermined sample reward sequence $\{X_{a,t}, t = 1, \dots, T\}$ of length $T < \infty$, $\gamma \in (0, 1]$, and $w^\phi(t)$ is a weight for ϕ such that $\sum_{\phi \in \Phi} w^\phi(t) = 1$. $w^\phi(t)$ is updated from the observed sample

reward $X_{t-1}^{\text{Exp4},t-1}$ as follows: $w^\phi(t) = w^\phi(t-1)e^{\gamma y^\phi/K}$ with $y^\phi = \sum_{a \in A} \phi_{t-1}(h_{t-2}^{\text{Exp4}})(a) f(a)$, where $f(a) = X_{a,t-1} / \text{Exp4}_{t-1}(h_{t-2}^{\text{Exp4}})(a)[a = I_{t-1}^{\text{Exp4}}]$ with $[\cdot]$ being Iverson bracket. An arm is then played according to the distribution $\text{Exp4}_t(h_{t-1}^{\text{Exp4}})$. A finite-time upper bound (but no lower bound) of $O(\sqrt{KT \ln |\Phi|})$ on the expected “regret” was analyzed against a best strategy that achieves the optimal total reward of $\max_{\phi \in \Phi} \{\sum_{t=1}^T \sum_{a \in A} \phi_t(h_{t-1}^\phi)(a) X_{a,t}\}$ with respect to the fixed sample reward sequences for the arms (see [11] for a probability version within a contextual bandit setting and [12] for a simplified derivation).

McMahan and Streeter [13] proposed a variant of Exp4, called NEXP, in the same nonstochastic setting. NEXP needs to solve a linear program (LP) at each time to obtain a distribution over A that offers a “locally optimal” trade-off between exploration and exploitation. Even if some improvement over Exp4 was shown, this comes at the expense of solving an LP at every time step.

de Farias and Megiddo [14] presented another variant of Exp4, called EEE, within a “reactive” setting. In this setting, at each time a player chooses an arm and an environment chooses its state, which is unknown to the player. The reward obtained by the player depends on both the chosen arm and the current state but not necessarily determined by a distribution specific to the chosen arm and the current state. An example of this setting is playing a repeated game against another player. When an expert is selected by EEE for a phase, it is followed for multiple times during the phase, rather than picking a different expert at each time, and the average reward accumulated for that phase is kept track of. The current best strategy with respect to the estimate of the average reward or a random strategy is selected at each phase with a certain control rule of exploration and exploitation, which is similar to the ϵ_t -schedule we consider here. (See also a survey section in [15] for expert-combining algorithms in different scenarios).

Because these representative approaches combine multiple experts, we compare them with our algorithm after adapting into our setting (cf. Section 4). However, more importantly, the notion of the best strategy in nonstochastic or reactive settings does not directly apply to the stochastic setting. Establishing some kind of asymptotic optimality for Exp4 or its variants with respect to a properly defined best strategy (even after adapting Exp4 as a strategy in Π) is an open problem. In fact, to the authors’ best knowledge, there seems to be no notable work yet which studies asymptotic optimality in combining multiple strategies for *stochastic* multiarmed bandit problems.

Finally, we stress that this paper focuses on asymptotic optimality as performance measure, also termed as “instantaneous regret” [8], but not on “expected regret” which is typically considered in the (recent) existing bandit-theory literature (see, e.g., [12] for a survey). It is worthwhile to note that the instantaneous regret is a stronger measure of convergence than expected regret [8].

3. Algorithm and Convergence

Assume that a finite nonempty subset Φ of Π is given. Once $\phi \in \Phi$ is selected by the algorithm $\tau = \epsilon_t\text{-comb}(\Phi)$ at time t , the bandit is played with an arm selected by ϕ and a sample reward of $X_{I_t^\phi, t}$ is obtained. Let S_t^τ denote the random variable denoting the strategy selected by τ at time t , and (with abusing notations) let $T_\phi^\tau(n) = \sum_{t=1}^n [S_t^\tau = \phi]$ denote the number of times ϕ has been selected by τ during the first n time steps, where $[S_t^\tau = \phi] = 1$ if $S_t^\tau = \phi$ and 0 otherwise. Let $t_{\phi,1}^\tau = \min\{t \mid S_t^\tau = \phi, t \geq 1\}$, $t_{\phi,2}^\tau = \min\{t \mid S_t^\tau = \phi, t > t_{\phi,1}^\tau\}, \dots$ and $t_{\phi,s}^\tau = \min\{t \mid S_t^\tau = \phi, t > t_{\phi,s-1}^\tau\}$ where $t_{\phi,s}^\tau$ is the time for the s th selection of the strategy $\phi \in \Phi$ by τ . For finite $m \geq 1$, we then let

$$\bar{X}_{\phi,m}^\tau = \frac{1}{m} \sum_{s=1}^m X_{I_{t_{\phi,s}^\tau}^\phi, t_{\phi,s}^\tau}, \quad \phi \in \Phi. \quad (2)$$

We formally describe the algorithm $\tau = \epsilon_t\text{-comb}(\Phi)$ below.

The $\tau = \epsilon_t\text{-comb}(\Phi)$ Algorithm

(1) *Initialization*: select $\epsilon_t \in (0, 1]$, for $t = 1, 2, \dots$. Set $t = 1$ and $T_\phi^\tau(0) = 0$, $\bar{X}_{\phi,0}^\tau = 0$, for all $\phi \in \Phi$.

(2) *Loop*:

(2.1) obtain $S_t \in \arg \max_{\phi \in \Phi} \bar{X}_{\phi, T_\phi^\tau(t-1)}^\tau$ (ties broken arbitrarily);

(2.2) with probability $1 - \epsilon_t$, select S_t and, with probability ϵ_t , select uniformly a strategy in Φ . Set the selected strategy ϕ to be S_t^τ . Select an arm according to $\phi_{T_\phi^\tau(t-1)+1}^\phi(h_{T_\phi^\tau(t-1)}^\phi)$ and obtain $X_{I_t^\phi, t}$ by playing it;

(2.3) $T_{S_t^\tau}^\tau(t) \leftarrow T_{S_t^\tau}^\tau(t-1) + 1$ and $t \leftarrow t + 1$.

Note that $\epsilon_t\text{-comb}(\Phi)$ as above is involved with general schedule of $\{\epsilon_t\}$. By setting the $\{\epsilon_t\}$ -schedule in $\epsilon_t\text{-comb}(\Phi)$ properly, $\{\epsilon_t\}$ subsumes the schedules used in ϵ -greedy, ϵ -first, and ϵ -decreasing strategies [16, 17]. In particular, as a special case, if $\Phi = \{\phi_a, a = 1, \dots, K\}$ and ϕ_a is given such that $P_a^\phi(t) = 1$ for all t , then τ degenerates to the ϵ_t -greedy strategy. As shown in the experimental results in [16, 17], the performance of *tuned* ϵ -greedy is no worse than (or very close to) those of ϵ -first, ϵ -decreasing strategies, and so forth, even with tuning the schedule (and the relevant parameters) each strategy used. However, because these schedules are usually heuristically tuned and ϵ -greedy uses a constant value of ϵ , it is not necessarily guaranteed that employing such schedules in $\epsilon_t\text{-comb}(\Phi)$ achieves asymptotic optimality. Furthermore, it is very difficult to tune the value of ϵ in advance. The theorem below establishes general conditions for asymptotic optimality of τ with respect to a properly defined best strategy in Φ . It states that if each $\phi \in \Phi$ is selected infinitely often by τ and for each ϕ , each $a \in A$ is selected infinitely often by ϕ , and each ϕ 's action-selection distribution converges to a stationary distribution, and the selection of τ goes greedy in the limit; then the selected strategy by τ converges to the best strategy in Φ .

Theorem 1. Given a finite nonempty $\Phi \subset \Pi$, consider $\tau = \epsilon_t\text{-comb}(\Phi)$. Suppose that $\lim_{t \rightarrow \infty} \epsilon_t = 0$, that $\sum_{t=0}^\infty \epsilon_t = \infty$, that there exists $\alpha > 0$ such that $P_a^\phi(t) \geq \alpha$ for all $\phi \in \Phi$, $a \in A$, and $t \geq 1$, and that $\lim_{t \rightarrow \infty} P^\phi(t) = d^\phi$ for all $\phi \in \Phi$. Then

$$\lim_{t \rightarrow \infty} \Pr \left\{ S_t^\tau \in \arg \max_{\phi \in \Phi} \sum_{a \in A} d_a^\phi \mu_a \right\} = 1. \quad (3)$$

Proof. Because $\lim_{t \rightarrow \infty} P^\phi(t) = d^\phi$, for all $\phi \in \Phi$ from the assumption, there exists $t_\epsilon < \infty$ such that, for all $t \geq t_\epsilon$, $\max_{a \in A} |P_a^\phi(t) - d_a^\phi| \leq \epsilon$, for all $\phi \in \Phi$. Fix any $\phi \in \Phi$. Then decomposing the sample average such that

$$\begin{aligned} \frac{1}{T_\phi^\tau(t-1)} \sum_{s=1}^{T_\phi^\tau(t-1)} X_{I_{t_{\phi,s}^\tau}^\phi, t_{\phi,s}^\tau} &= \frac{1}{T_\phi^\tau(t-1)} \sum_{s=1}^{T_\phi^\tau(t_\epsilon)} X_{I_{t_{\phi,s}^\tau}^\phi, t_{\phi,s}^\tau} \\ &+ \frac{1}{T_\phi^\tau(t-1)} \sum_{s=T_\phi^\tau(t_\epsilon)+1}^{T_\phi^\tau(t-1)} X_{I_{t_{\phi,s}^\tau}^\phi, t_{\phi,s}^\tau}, \end{aligned} \quad (4)$$

we see that as $t \rightarrow \infty$, the first term in the right-hand side of (4) goes to zero because $T_\phi^\tau(t-1) \rightarrow \infty$ from $\sum_{t=0}^\infty \epsilon_t = \infty$ and $T_\phi^\tau(t_\epsilon) < \infty$. For the second term in the right-hand side of (4), we rewrite it as

$$\frac{T_\phi^\tau(t-1) - T_\phi^\tau(t_\epsilon)}{T_\phi^\tau(t-1)} \cdot \frac{\sum_{s=T_\phi^\tau(t_\epsilon)+1}^{T_\phi^\tau(t-1)} X_{I_{t_{\phi,s}^\tau}^\phi, t_{\phi,s}^\tau}}{T_\phi^\tau(t-1) - T_\phi^\tau(t_\epsilon)} \quad (5)$$

and we will establish the convergence of the right product term in (5):

$$\frac{1}{T_\phi^\tau(t-1) - T_\phi^\tau(t_\epsilon)} \sum_{s=T_\phi^\tau(t_\epsilon)+1}^{T_\phi^\tau(t-1)} X_{I_{t_{\phi,s}^\tau}^\phi, t_{\phi,s}^\tau} \quad (6)$$

because $(T_\phi^\tau(t-1) - T_\phi^\tau(t_\epsilon))/T_\phi^\tau(t-1)$ goes to one.

Let $N_a(T_\phi^\tau(n)) = \sum_{s=1}^{T_\phi^\tau(n)} [I_{t_{\phi,s}^\tau}^\phi = a]$, $a \in A$. Then (6) can be rewritten such that

$$\begin{aligned} \frac{\sum_{s=T_\phi^\tau(t_\epsilon)+1}^{T_\phi^\tau(t-1)} X_{I_{t_{\phi,s}^\tau}^\phi, t_{\phi,s}^\tau}}{T_\phi^\tau(t-1) - T_\phi^\tau(t_\epsilon)} &= \sum_{a \in A} \frac{N_a(T_\phi^\tau(t-1)) - N_a(T_\phi^\tau(t_\epsilon))}{T_\phi^\tau(t-1) - T_\phi^\tau(t_\epsilon)} \\ &\times \frac{\sum_{n=t_\epsilon+1}^{t-1} X_{a,n} [S_n^\tau = \phi \wedge I_n^\phi = a]}{N_a(T_\phi^\tau(t-1)) - N_a(T_\phi^\tau(t_\epsilon))}. \end{aligned} \quad (7)$$

Because $P_a^\phi(t) \geq \alpha$ for all t , we have that, as $t \rightarrow \infty$, $N_a(T_\phi^\tau(t-1)) - N_a(T_\phi^\tau(t_\epsilon)) \rightarrow \infty$. Therefore the second product term inside of the summation in the right-hand side of (7),

$$\frac{(\sum_{n=t_\epsilon+1}^{t-1} X_{a,n} [S_n^\tau = \phi \wedge I_n^\phi = a])}{(N_a(T_\phi^\tau(t-1)) - N_a(T_\phi^\tau(t_\epsilon)))}, \quad (8)$$

goes to μ_a by the law of large numbers. We now show that the first product term inside of the summation in the right-hand side of (7), $(N_a(T_\phi^\tau(t-1)) - N_a(T_\phi^\tau(t_\epsilon)))/(T_\phi^\tau(t-1) - T_\phi^\tau(t_\epsilon))$, goes to d_a^ϕ as $t \rightarrow \infty$. Let

$$\bar{P}_a^\epsilon = \frac{P_a^\phi(t_{\phi, T_\phi^\tau(t_\epsilon)+1}^\tau) + P_a^\phi(t_{\phi, T_\phi^\tau(t_\epsilon)+2}^\tau) + \dots + P_a^\phi(t_{\phi, T_\phi^\tau(t-1)}^\tau)}{T_\phi^\tau(t-1) - T_\phi^\tau(t_\epsilon)}. \quad (9)$$

Then by Poisson's limit theorem [18, Chapter 11] in the law of large numbers, for any $\beta > 0$,

$$\Pr \left\{ \left| \frac{N_a(T_\phi^\tau(t-1)) - N_a(T_\phi^\tau(t_\epsilon))}{T_\phi^\tau(t-1) - T_\phi^\tau(t_\epsilon)} - \bar{P}_a^\epsilon \right| \leq \beta \right\} \geq 1 - \frac{1}{4(T_\phi^\tau(t-1) - T_\phi^\tau(t_\epsilon))\beta^2}. \quad (10)$$

Because $|P_a^\phi(t) - d_a^\phi| \leq \epsilon$ for all $t \geq t_\epsilon$, $|\bar{P}_a^\epsilon - d_a^\phi| \leq \epsilon$. By setting $\beta = \epsilon$ and using triangular inequality, we have that, for any $\epsilon > 0$,

$$\Pr \left\{ \left| \frac{N_a(T_\phi^\tau(t-1)) - N_a(T_\phi^\tau(t_\epsilon))}{T_\phi^\tau(t-1) - T_\phi^\tau(t_\epsilon)} - d_a^\phi \right| \leq 2\epsilon \right\} \geq 1 - \frac{1}{4(T_\phi^\tau(t-1) - T_\phi^\tau(t_\epsilon))\epsilon^2}. \quad (11)$$

It follows that with probability 1, as $t \rightarrow \infty$, $\bar{X}_{\phi, T_\phi^\tau(t-1)}^\tau \rightarrow \sum_{a \in A} d_a^\phi \mu_a$. Finally, because $\epsilon_t \rightarrow 0$ as $t \rightarrow \infty$, the result of the theorem follows. $\square$

We remark that this algorithm can be used for solving a bandit problem in a decomposed manner. Suppose that we partition A into m nonempty subsets $A_i, i = 1, \dots, m$, such that $A_i \cap A_j = \emptyset, i \neq j$ and $\bigcup_{i=1}^m A_i = A$. Choose any asymptotically optimal strategy ϕ and associate ϕ with a strategy ϕ_i such that $\sum_{a \in A_i \setminus A_i^*} P_a^{\phi_i}(t) \rightarrow 0$ as $t \rightarrow \infty$, where $A_i^* = \{a \in A_i \mid \mu_a = \max_{a' \in A_i} \mu_{a'}\}$ and $P_a^{\phi_i}(t) = 0$ for all $a \in A \setminus A_i$. Then by employing $\tau = \epsilon_t\text{-comb}(\Phi)$ where $\Phi = \{\phi_1, \dots, \phi_m\}$, we have that $\sum_{a \in A \setminus A^*} P_a^{S_i^\tau}(t) \rightarrow 0$ as $t \rightarrow \infty$.

4. A Numerical Example

For a proof-of-concept implementation of the approach, we consider three simple numerical examples.

We have Bernoulli reward distributions with $K = 10$ where the reward expectations of the arms 1 through 10 are given by 0.9, 0.8, 0.8, 0.8, 0.7, 0.7, 0.7, 0.6, 0.6, and 0.6, respectively. For any strategy π involved with $\{\epsilon_t\}$ in this section, including $\epsilon_t\text{-comb}(\Phi)$, we used $\{\epsilon_t\}$ -schedule in ϵ_t -greedy strategy given by $\min\{1, c(\pi)/Kd^2t\}$ based on [8, Theorem 3], where $d = \min_{a \in A \setminus A^*} (\mu^* - \mu_a)$ and $c(\pi)$ is a constant chosen for π to control the degree of exploration and exploitation.

FIGURE 1: Percentage of selections of the best strategy of (tuned) $\epsilon_t\text{-comb}$ (3 ϵ_t -greedy strategies) and of (tuned) Exp4/NEXP/EEEs for the distribution (0.9, 0.8, 0.8, 0.8, 0.7, 0.7, 0.7, 0.6, 0.6, 0.6).

For the first case, A is partitioned into $A_i, i = 1, 2, 3$, such that $A_1 = \{1, 2, 3, 4\}$, $A_2 = \{5, 6, 7\}$, and $A_3 = \{8, 9, 10\}$ and the ϵ_t -greedy strategy associated with A_i , playing only the arms in A_i , corresponds to ϕ_i (cf. the remark given at the end of Section 3). Thus we have $\Phi = \{\phi_i, i = 1, 2, 3\}$ and trivially the best strategy is ϕ_1 . The second case considers combining two pursuit learning automata (PLA) algorithms [19] with different learning rates designed for solving stochastic optimization problems. Even though PLA was not designed specifically for solving multiarmed bandit problems, PLA guarantees “ ϵ -optimality” and can be casted into a strategy for playing the bandit. (Roughly, the error probability of choosing nonoptimal solutions is bounded by ϵ .) The first PLA strategy uses the learning rate of 0.000002 which corresponds to the parameter setting in [19, Theorem 3.1] with $\epsilon = 0.1$ for a theoretical performance guarantee and the second one uses the tuned learning rate of 0.002 which achieves the best performance among the various rates we tested for the above distribution. These two PLAs are contained in Φ and the PLA with 0.002 is taken as the best strategy.

From Figures 1–3, we show the percentage of selections of the best strategy and of plays of the optimal arm of (tuned) $\epsilon_t\text{-comb}(\Phi)$ for each case along with those of (tuned) Exp4, NEXP, and EEE, respectively. (The percentage for the optimal arm for the second case is not shown due to the space constraint.) The tuned strategy corresponds to the best empirical parameter setting we obtained. The performances of all tested strategies were obtained through the average over 1000 different runs where each strategy is followed over 100,000 time steps in a single run. For the first case, that is, $\Phi = \{3\epsilon_t\text{-greedy's}\}$, we set $c(\phi) = 0.15$ for all $\phi \in \Phi$, and $c(\tau) = 0.15$ for $\tau = \epsilon_t\text{-comb}(\Phi)$, and $c(\text{EEE}) = 0.15$ for EEE. (The value of 0.15 was chosen for a reasonable allocation of exploration and exploitation.) The tuned τ uses 0.075 for $c(\tau)$ and the tuned EEE uses 0.07 for $c(\text{EEE})$. Exp4 uses 0.0116 for y

FIGURE 2: Percentage of plays of the optimal arm of (tuned) ϵ_t -comb(3 ϵ_t -greedy strategies) and of (tuned) Exp4/NEXP/EEEs for the distribution (0.9, 0.8, 0.8, 0.8, 0.7, 0.7, 0.7, 0.6, 0.6, 0.6).

FIGURE 3: Percentage of selections of the best strategy of (tuned) ϵ_t -comb(2 PLAs) and of (tuned) Exp4/NEXP/EEEs for the distribution (0.9, 0.8, 0.8, 0.8, 0.7, 0.7, 0.7, 0.6, 0.6, 0.6).

which was obtained by the parameter setting in [10, Corollary 3.2] and the tuned Exp4 uses 0.03. For the second case, that is, $\Phi = \{2 \text{ PLAs}\}$, we set again $c(\tau) = c(\text{EEE}) = 0.15$. The tuned τ uses 0.3333 for $c(\tau)$, and the tuned EEE uses 0.04 for $c(\text{EEE})$ and (tuned) Exp4 uses the same γ value in the first case. As we can see from Figures 1–3 (tuned) τ successfully combines the strategies in Φ achieving the asymptotic optimality with respect to the best strategy for each case. In particular, τ 's convergence rate is much faster than that of (even tuned) Exp4, NEXP, and EEE for each case. Note that by tuning $c(\tau)$, τ achieves a better performance, which shows a similar sensitivity to $\{\epsilon_t\}$ -schedule as in the ϵ_t -greedy strategy case [8].

The third case considers the case of combining two different strategies, the ϵ_t -greedy strategy and UCB1-tuned [8], and was chosen to show some robustness of ϵ_t -comb(Φ). Here the ϵ_t -greedy strategy uses the value of 0.3 for $\{\epsilon_t\}$ -schedule. (Tuning the ϵ_t -greedy strategy will make it more competitive with UCB1-tuned. This particular value was chosen for an illustration purpose only.) Figure 4 shows the average regret of each strategy π tested over n time steps for two different distributions, respectively, calculated by $10^{-3} \sum_{r=1}^{10^3} \sum_{a \in A} (\mu^* - \mu_a) T_{a,r}^\pi(n)$ where $T_{a,r}^\pi(n)$ denotes the number of times arm a has been played by π during the first n plays in the r th run. The first distribution is the same as the above one used for the first and the second cases. The second distribution is again Bernoulli reward distributions with $K = 10$ but the reward expectations are given by 0.9 for the optimal arm 1 and 0.7 for all the remaining arms. For the first distribution, UCB1-tuned's regret is much smaller than that of the ϵ_t -greedy strategy but for the second distribution, the ϵ_t -greedy strategy is slightly better than UCB1-tuned. In sum, we have the case where the better performance of the two strategies is distribution-dependent in terms of the average regret even if both achieve asymptotic optimality empirically (not shown here). By combining two strategies by $\tau = \epsilon_t$ -comb($\{\text{UCB1-tuned}, \epsilon_t$ -greedy $\}$), we have a reasonable distribution-independent algorithm for playing the bandit. As we can see from Figure 4, the tuned τ with $c(\tau) = 0.0375$ and 0.01, respectively, shows a robust performance across the two distributions. For both cases, the performances of the tuned Exp4, NEXP, and EEE are not good.

5. Concluding Remarks

In this paper, we provided a randomized algorithm ϵ_t -comb(Φ) for playing a given stochastic MAB when a finite nonempty strategy set Φ is available. Following the spirit of the ϵ_t -greedy strategy, the algorithm combines the strategies in Φ and finds the best strategy in the limit. Specifically, at each time t , we use the probabilistic ϵ_t -switching to select uniformly a strategy in Φ or to select a strategy with the highest sample average of the rewards obtained so far by playing the bandit. Once a strategy ϕ is selected, the arm chosen by ϕ is played for the bandit. We showed that it is asymptotically optimal in the sense that the selected strategy converges to the best in the set under some conditions on the strategies in Φ and $\{\epsilon_t\}$ and illustrated the result by simulation studies on some example problems.

If each $\phi \in \Phi$ is stationary (as opposed to the general case studied in the present paper) in that, for a distribution d^ϕ over A , $P^\phi(t) = d^\phi$ for all t , then each $\phi \in \Phi$ can be viewed as an arm and playing according to ϕ provides a sample reward of $X_{t,\phi}$ with unknown expectation of $\sum_{a \in A} d_a^\phi \mu_a$. Therefore, in this case any asymptotically optimal strategy in Π can be adapted for selecting a strategy at each time instead of ϵ_t -comb(Φ) to achieve asymptotic optimality with respect to a strategy that achieves $\max_{\phi \in \Phi} \sum_{a \in A} d_a^\phi \mu_a$.

Even if we showed the convergence to the best strategy in the set, the convergence rate has not been discussed.

FIGURE 4: Average regret of tuned ϵ_t -comb($\{\epsilon_t$ -greedy, UCB1-tuned $\}$) and of tuned Exp4 for the distributions of (0.9,0.8,0.8,0.8,0.7,0.7,0.7, 0.6,0.6,0.6) and (0.9,0.7,0.7,0.7,0.7,0.7,0.7,0.7,0.7,0.7).

In the statement of Theorem 1, we assumed only that $\lim_{t \rightarrow \infty} P^\phi(t) = d^\phi$, for all $\phi \in \Phi$. In other words, the convergence rate of each strategy in Φ to its stationary behavior is assumed to be unknown. Because the convergence rate of ϵ_t -comb(Φ) needs to be given with respect to $\arg \max_{\phi \in \Phi} \sum_{a \in A} d_a^\phi \mu_a$, it is crucial to know the convergence rate of each strategy in Φ . It would be a good future work to analyze the convergence rate of ϵ_t -comb(Φ), for example, by extending the result of Theorem 3 in [8] for ϵ_t -greedy strategy into the case of ϵ_t -comb(Φ), under the conditions on the rates of $\{P^\phi(t)\}$ to d^ϕ , for all $\phi \in \Phi$.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

- [1] C. Tekin and M. Liu, "Online learning methods for networking," *Foundations and Trends in Networking*, vol. 8, no. 4, pp. 281–409, 2013.
- [2] A. Mahajan and D. Teneketzis, "Multi-armed bandit problems," in *Foundations and Applications of Sensor Management*, A. O. Hero III, D. A. Castanon, D. Cochran, and K. Kastella, Eds., pp. 121–151, Springer, New York, NY, USA, 2008.
- [3] C. B. Browne, E. Powley, D. Whitehouse et al., "A survey of Monte Carlo tree search methods," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 4, no. 1, pp. 1–43, 2012.
- [4] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*, Cambridge University Press, 2006.
- [5] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, vol. 58, pp. 527–535, 1952.
- [6] T. Santner and A. Tamhane, *Design of Experiments: Ranking and Selection*, CRC Press, 1984.
- [7] J. Bather, "Randomized allocation of treatments in sequential trials," *Advances in Applied Probability*, vol. 12, no. 1, pp. 174–182, 1980.
- [8] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2-3, pp. 235–256, 2002.
- [9] J. C. Gittins and D. M. Jones, "A dynamic allocation index for sequential design of experiments," in *Progress in Statistics*, vol. 1, pp. 241–266, European Meeting of Statisticians, 1972.
- [10] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, 2002/03.
- [11] A. Beygelzimer, J. Langford, L. Li, L. Reyzin, and R. E. Schapire, "Contextual bandit algorithms with supervised learning guarantees," *Journal of Machine Learning Research*, vol. 15, pp. 19–26, 2011.
- [12] 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.
- [13] H. B. McMahan and M. Streeter, "Tighter bounds for multi-armed bandits with expert advice," in *Proceedings of the 22nd Conference on Learning Theory (COLT '09)*, Montreal, Canada, 2009.
- [14] D. P. de Farias and N. Megiddo, "Combining expert advice in reactive environments," *Journal of the ACM*, vol. 53, no. 5, pp. 762–799, 2006.
- [15] A. György and G. Ottucsák, "Adaptive routing using expert advice," *The Computer Journal*, vol. 49, no. 2, pp. 180–189, 2006.
- [16] V. Kuleshov and D. Precup, "Algorithms for multi-armed bandit problems," <http://arxiv.org/abs/1402.6028>.
- [17] J. Vermorel and M. Mohri, "Multi-armed bandit algorithms and empirical evaluation," in *Machine Learning: ECML 2005*, vol. 3720 of *Lecture Notes in Computer Science*, pp. 437–448, Springer, Berlin, Germany, 2005.

- [18] J. V. Uspensky, *Introduction to Mathematical Probability*, McGraw-Hill, London, UK, 1937.
- [19] K. Rajaraman and P. S. Sastry, "Finite time analysis of the pursuit algorithm for learning automata," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 26, no. 4, pp. 590–598, 1996.