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

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 $\epsilon_i$-switching in the $\epsilon_i$-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 $\{\epsilon_i\}$.

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 $A = \{1, 2, \ldots, K\}$, $K > 1$ 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 $X_{a,t} \in \mathfrak{N}$ drawn from an unknown distribution associated with $a$, whose unknown expectation and variance are $\mu_a$ and $\sigma_a^2$, respectively. We define a strategy $\pi = \{\pi_t, t = 1, 2, \ldots\}$ as a sequence of mappings such that $\pi_t$ 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 $H_a$ is an arbitrarily given nonempty subset of $\mathfrak{R}$. We denote the set of all possible strategies as $\Pi$. Given a particular sequence $h_{t-1}^a \in H_{t-1}$ of the past plays and rewards obtained by following $\pi \in \Pi$ over $t - 1$ time steps, $\pi$ selects $a \in A$ to be played by probability $\pi_t(h_{t-1}^a)(a)$ at time $t$. We assume that $\pi_t(h_{t-1}^a)$ is arbitrarily given.

Let random variable $\text{I}_t^a$ denote the arm selected by $\pi$ at time $t$ and let $P^a(t)$ be the distribution over $A$ given by $\pi$ at time $t$ so that $P^a(t) = \Pr(\text{I}_t^a = a)$, which is equal to $\sum_{h_{t-1}^a \in H_{t-1}} \pi_t(h_{t-1}^a)(a) \Pr[h_{t-1}^a]$. We assume that $X_{a, \cdot}$ and $X_{a', \cdot}$ are independent of any $a, a'$ and $t \neq t'$ and $X_{a, \cdot}$'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 $\sum_{a \in A^*} P^a(t) \rightarrow 0$ as $t \rightarrow \infty$, then we say that $\pi$ 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 $\pi$ such that, at time $t$, $\pi$ selects an arm in $\arg\max_{a \in A} \{X_{a, \cdot}^t + \lambda(T_{a, \cdot}^t(t))\}$, where $T_{a, \cdot}^t(t)$ denotes the number of times arm $a$ has been played
by π during the first \( t - 1 \) plays, and \( X_{a,m}' \), \( m \geq 1 \), 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 \( \lambda(t) \to 0 \) as \( t \to \infty \), 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 \( X_{a,T} \) and to make the effect vanish as \( T \) increases. The well-known asymptotically optimal \( ε_t \)-greedy strategy \[8\] with \( \sum_{t=1}^{∞} ε_t = \infty \) and \( \lim_{t→∞} ε_t = 0 \) follows exactly the same idea: at time \( t \), with probability \( ε_t \in (0, 1) \), it selects \( a \in A \) with probability \( 1/K \) and, with probability \( 1 - ε_t \), it selects an arm in arg max\( a \in A \{X_{a,T_\phi}(t-1)\} \) 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 \( Φ \subset Π \). 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 \( \{ε_t\} \).

### 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

\[
\text{Exp4}_t \left( h^{\text{Exp4}}_{t-1} \right) (a) = \left( 1 - \gamma \right) \sum_{\phi \in \Phi} w^\phi(t) \phi \left( h^{\text{Exp4}}_{t-1} \right) (a) + \frac{\gamma}{|\Phi|}, \quad a \in A,
\]

where \( h^{\text{Exp4}}_{t-1} \) is obtained from the predetermined sample reward sequence \( \{X_{a,T}, t = 1, \ldots, T\} \) 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 \( X_{a,T}^{\text{Exp4}}_{t-1} \) as follows:

\[
w^\phi(t) = w^\phi(t-1) e^{t\phi/|\Phi|},
\]

with \( \phi = \sum_{a \in A} \phi \left( h^{\text{Exp4}}_{t-1} \right) (a) f(a) \), where \( f(a) = X_{a,T}^{\text{Exp4}}_{t-1} \).

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 \( \max_{\phi \in \Phi} \left( \sum_{t=1}^{T} \sum_{a \in A} \phi \left( h^{\phi}_{t-1} \right) (a) X_{a,T} \right) \) 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 φ ∈ Φ is selected by the algorithm τ = ε₁-comb(Φ) at time t, the bandit is played with an arm selected by φ and a sample reward of X_{t,φ} is obtained. Let S_τ denote the random variable denoting the strategy selected by τ at time t, and (with abusing notations) let T_φ^τ(m) = \sum_{s=1}^{m} [S_τ^s = φ] denote the number of times φ has been selected by τ during the first n time steps, where |S_τ^t = φ| = 1 if S_τ^t = φ and 0 otherwise. Let t^φ\_φₜ = \min\{t | S_τ^t = φ, t ≥ 1\}, t^φ\_φ₂ = \min\{t | S_τ^t = φ, t > t^φ\_φ₁\}, ..., and t^φ\_φₙ = \min\{t | S_τ^t = φ, t > t^φ\_φₙ₋₁\} where t^φ\_φₙ is the n-th selection of the strategy φ ∈ Φ by τ. For finite m ≥ 1, we then let

\[ \overline{X}_{φ,τ} = \frac{1}{m} \sum_{i=1}^{m} X_{i,φ,τ} \phi ∈ Φ. \] (2)

We formally describe the algorithm τ = ε₁-comb(Φ) below.

The τ = ε₁-comb(Φ) Algorithm

1. **Initialization:** select ε₁ ∈ (0, 1), for t = 1, 2, ..., Set t = 1 and T_φ^τ(0) = 0, \overline{X}_{φ,0} = 0, for all φ ∈ Φ.

2. **Loop:**

   2.1) Obtain S_τ = arg max_{φ∈Φ} \overline{X}_{φ,τ} to be other end of the strategy t = 1, 2, ..., Set the selected strategy φ to be S_τ. Select an arm according to φT_φ^τ(t-1) (ties broken arbitrarily);

   2.2) With probability 1 - ε₁, select S_τ and, with probability ε₁, select uniformly a strategy in Φ. Then (6) can be rewritten as

3. **T_{φ,τ}(t) = T_{φ,τ}(t-1) + 1 and t ← t + 1.

Note that ε₁-comb(Φ) as above is involved with general schedule of ε₁. Note that ε₁-comb(Φ) properly, ε₁ subsumes the schedules used in ε-greedy, ε-first, and ε-decreasing strategies [16, 17]. In particular, as a special case, if Φ = {φ_a, a = 1, ..., K} and ε₁ is given such that T_φ^τ(t) = 1 for all t, then the strategy t = 1, 2, ..., Set the selected strategy φ to be S_τ. Select an arm according to φT_φ^τ(t-1) (ties broken arbitrarily); and (4) becomes zero because T_φ^τ(t-1) → 0 from \sum_{φ=1}^{∞} ε₁ = 0 and T_φ^τ(t) < ∞. For the second term in the right-hand side of (4), we rewrite it as

\[ \frac{T_φ^τ(t-1) - T_φ^τ(t)}{T_φ^τ(t-1) - T_φ^τ(t)} \]

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

Because (T_φ^τ(t-1) - T_φ^τ(t)) goes to zero because T_φ^τ(t-1) → 0 from \sum_{φ=1}^{∞} ε₁ = 0 and T_φ^τ(t) < ∞. For the second term in the right-hand side of (4), we rewrite it as

\[ \frac{T_φ^τ(t-1) - T_φ^τ(t)}{T_φ^τ(t-1) - T_φ^τ(t)} \]

Because (T_φ^τ(t-1) - T_φ^τ(t)) goes to one.

Let N_a(T_φ^τ(t)) = \sum_{i=1}^{T_φ^τ(t)} [φ_i = a], a ∈ A. Then (6) can be rewritten such that

\[ \sum_{i=1}^{T_φ^τ(t)} X_iφ_{,φ,τ} = \sum_{i=1}^{T_φ^τ(t)} [φ_i = a] \]

Now because (T_φ^τ(t-1) - T_φ^τ(t)) goes to one.

Theorem 1. Given a finite nonempty Φ ⊂ Π, consider τ = ε₁-comb(Φ). Suppose that \lim_{t→∞} ε₁ = 0, that \sum_{t=0}^{∞} ε₁ = ∞, that there exists \alpha > 0 such that P_φ^τ(t) ≥ α for all φ ∈ Φ, a ∈ A, and t ≥ 1, and that \lim_{t→∞} P_φ^τ(t) = d_φ for all φ ∈ Φ. Then

\[ \lim_{t→∞} \Pr \left\{ S_τ^t = \arg \max_{φ∈Φ} \sum_{a∈A} d_φ^a μ_a \right\} = 1. \] (3)

Proof. Because \lim_{t→∞} P_φ^τ(t) = d_φ, for all φ ∈ Φ from the assumption, there exists \epsilon_ε < ∞ such that, for all t ≥ t_ε, max_{φ∈A} P_φ^τ(t) - d_φ < \epsilon, for all φ ∈ Φ. Fix any φ ∈ Φ. Then decomposing the sample average such that

\[ \frac{1}{1} \sum_{i=1}^{T_φ^τ(t)-1} X_iφ_{,φ,τ} = \frac{1}{1} \sum_{i=1}^{T_φ^τ(t)-1} X_iφ_{,φ,τ} \]

we see that as t → ∞, the first term in the right-hand side of (4) goes to zero because T_φ^τ(t-1) → ∞ from \sum_{φ=1}^{∞} ε₁ = 0 and T_φ^τ(t) < ∞. For the second term in the right-hand side of (4), we rewrite it as

\[ \frac{T_φ^τ(t-1) - T_φ^τ(t)}{T_φ^τ(t-1) - T_φ^τ(t)} \]

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

\[ \frac{T_φ^τ(t-1) - T_φ^τ(t)}{T_φ^τ(t-1) - T_φ^τ(t)} \]

because (T_φ^τ(t-1) - T_φ^τ(t)) goes to one.

Let N_a(T_φ^τ(t)) = \sum_{i=1}^{T_φ^τ(t)} [φ_i = a], a ∈ A. Then (6) can be rewritten such that

\[ \sum_{i=1}^{T_φ^τ(t)-1} X_iφ_{,φ,τ} = \sum_{i=1}^{T_φ^τ(t)-1} X_iφ_{,φ,τ} \]

Now because (T_φ^τ(t-1) - T_φ^τ(t)) goes to one.

\[ \sum_{i=1}^{T_φ^τ(t)-1} X_iφ_{,φ,τ} \]

Because P_φ^τ(t) ≥ α for all t, we have that, as t → ∞, N_a(T_φ^τ(t)) = N_a(T_φ^τ(t)) → ∞. Therefore, the second product term inside of the summation in the right-hand side of (7),

\[ \left( \sum_{i=1}^{T_φ^τ(t)-1} X_iφ_{,φ,τ} \right) \]

(8)
goes to $\mu_\alpha$ 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), $(N_\alpha(T^*_\phi(t) - 1) - N_\alpha(T^*_\phi(t_c)))/(T^*_\phi(t - 1) - T^*_\phi(t_c))$, goes to $d_\alpha^\phi$ as $t \to \infty$. Let

$$
\tilde{P}_a^\phi = \frac{P_a^\phi(T^*_\phi(t) + 1) + P_a^\phi(T^*_\phi(t) + 2) + \cdots + P_a^\phi(T^*_\phi(t - 1))}{T^*_\phi(t - 1) - T^*_\phi(t_c)}.
$$

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

$$
\Pr \left\{ \left| \frac{N_\alpha(T^*_\phi(t) - 1) - N_\alpha(T^*_\phi(t_c))}{T^*_\phi(t - 1) - T^*_\phi(t_c)} - \tilde{P}_a^\phi \right| \leq \beta \right\} \geq 1 - \frac{1}{4(T^*_\phi(t) - 1) - T^*_\phi(t_c)} \beta^2.
$$

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

$$
\Pr \left\{ \left| \frac{N_\alpha(T^*_\phi(t) - 1) - N_\alpha(T^*_\phi(t_c))}{T^*_\phi(t) - 1) - T^*_\phi(t_c)} - d_\alpha^\phi \right| \leq 2\epsilon \right\} \geq 1 - \frac{1}{4(T^*_\phi(t) - 1) - T^*_\phi(t_c)} \epsilon^2.
$$

It follows that with probability 1, as $t \to \infty$, $\overline{X^\phi}_{\phi,T^*_\phi(t)} \to \sum_{a \in A^*} d_\alpha^\phi \mu_\alpha$. Finally, because $\epsilon_i \to 0$ as $t \to \infty$, the result of the theorem follows. \qed

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 $A_j, i = 1, \ldots, m$, such that $A_i \cap A_j = \emptyset, i \neq j$ and $\bigcup_{i=1}^m A_j = A$. Choose any asymptotically optimal strategy $\phi$ and associate $\phi$ with a strategy $\phi_i$ such that $\sum_{a \in A \setminus A^*} P_a^\phi(t) \to 0$ as $t \to \infty$, where $A_j^* = \{a \in A_i \mid \mu_\alpha = \max_{a' \in A_j, a' \neq a} \mu_\alpha \}$ and $P_a^\phi(t) = 0$ for all $a \in A \setminus A_j$. Then by employing $\tau = \epsilon_i$, comb(\Phi) where $\Phi = \{\phi_1, \ldots, \phi_m\}$, we have that $\sum_{a \in A \setminus A^*} P_a^\phi(t) \to 0$ as $t \to \infty$.

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 $\pi$ involved with $\epsilon_i$, in this section, including $\epsilon_i$, comb(\Phi), we used $\{\epsilon_i\}$-schedule in $\epsilon_i$-greedy strategy given by $\min\{1, e(\pi)/Kd^\tau\}$ based on [8, Theorem 3], where $d = \min_{a \in A \setminus A^*}(\mu^* - \mu_a)$ and $e(\pi)$ is a constant chosen for $\pi$ to control the degree of exploration and exploitation.

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 $\epsilon_i$-greedy strategy associated with $A_i$ holding only the arms in $A_i$, corresponds to $\phi_i$ (cf. the remark given at the end of Section 3). Thus we have $\Phi = \{\phi_1, \phi_2, \phi_3\}$ and trivially the best strategy is $\phi_1$. 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 “$c$-optimality” and can be casted into a strategy for playing the bandit. (Roughly, the error probability of choosing nonoptimal solutions is bounded by $c$.)

The first PLA strategy uses the learning rate of 0.000002 which corresponds to the parameter setting in [19, Theorem 3.1] with $c = 0.1$ 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 $\Phi$ 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 of (tuned) $\epsilon_i$-comb(\Phi) and of (tuned) Exp4, NEXP, and 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 1: Percentage of selections of the best strategy of (tuned) $\epsilon_i$-comb(3) $\epsilon_i$-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).
The third case considers the case of combining two different strategies, the $\epsilon_i$-greedy strategy and UCB1-tuned [8], and was chosen to show some robustness of $\epsilon_i$-comb($\Phi$). Here the $\epsilon_i$-greedy strategy uses the value of 0.3 for $\{\epsilon_i\}$-schedule. (Tuning the $\epsilon_i$-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 $\pi$ tested over $n$ time steps for two different distributions, respectively, calculated by 

$$10^{-3} \sum_{t=1}^{T} \sum_{a \in A}(\mu^* - \mu_a)T_{a,t}^\pi(n)$$

where $T_{a,t}^\pi(n)$ denotes the number of times arm $a$ has been played by $\pi$ during the first $n$ plays in the $t$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 $\epsilon_i$-greedy strategy but for the second distribution, the $\epsilon_i$-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 $\tau = \epsilon_i$-comb(UCB1-tuned, $\epsilon_i$-greedy), we have a reasonable distribution-independent algorithm for playing the bandit. As we can see from Figure 4, the tuned $\tau$ 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 $\epsilon_i$-comb($\Phi$) for playing a given stochastic MAB when a finite nonempty strategy set $\Phi$ is available. Following the spirit of the $\epsilon_i$-greedy strategy, the algorithm combines the strategies in $\Phi$ and finds the best strategy in the limit. Specifically, at each time $t$, we use the probabilistic $\epsilon_i$-switching to select uniformly a strategy in $\Phi$ or to select a strategy with the highest sample average of the rewards obtained so far by playing the bandit. Once a strategy $\phi$ is selected, the arm chosen by $\phi$ 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 $\Phi$ and $\{\epsilon_i\}$ 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 $\mathcal{D}^\phi$ over $A$, $P^\phi(t) = \mathcal{D}^\phi$ for all $t$, then each $\phi \in \Phi$ can be viewed as an arm and playing according to $\phi$ provides a sample reward of $X_{t,\phi}$ with unknown expectation of $\sum_{a \in A} \mathcal{D}^\phi_a$. Therefore, in this case any asymptotically optimal strategy in $\Pi$ can be adapted for selecting a strategy at each time instead of $\epsilon_i$-comb($\Phi$) to achieve asymptotic optimality with respect to a strategy that achieves $\max_{\phi \in \Phi} \sum_{a \in A} \mathcal{D}^\phi_a$.

Even if we showed the convergence to the best strategy in the set, the convergence rate has not been discussed.
In the statement of Theorem 1, we assumed only that 
\[ \lim_{t \to \infty} P_\phi(t) = d_\phi \] for all \( \phi \in \Phi \). In other words, the convergence rate of each strategy in \( \Phi \) to its stationary behavior is assumed to be unknown. Because the convergence rate of \( \epsilon_t\)-comb(\( \Phi \)) needs to be given with respect to 
\[ \arg \max_{\phi \in \Phi} \sum_{a \in A} q_{a \phi}^t, \] it is crucial to know the convergence rate of each strategy in \( \Phi \). It would be a good future work to analyze the convergence rate of \( \epsilon_t\)-comb(\( \Phi \)), for example, by extending the result of Theorem 3 in [8] for \( \epsilon_t\)-greedy strategy into the case of \( \epsilon_t\)-comb(\( \Phi \)), under the conditions on the rates of \( \{P_\phi(t)\} \) to \( d_\phi \), 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**


Submit your manuscripts at http://www.hindawi.com