THE INSTITUTE FOR SYSTEMS RESEARCH

ISR TECHNICAL REPORT 2009-16

On Convergence of Evolutionary Computation for Stochastic Combinatorial Optimization

Hyeong Soo Chang

On Convergence of Evolutionary Computation for Stochastic Combinatorial Optimization

Hyeong Soo Chang

Abstract

Extending Rudolph's works on the convergence analysis of evolutionary computation (EC) for deterministic combinatorial optimization problems (COPs), this brief paper establishes a probability one convergence of some variants of explicit-averaging EC to an optimal solution and the optimal value for solving stochastic COPs.

Keywords: Evolutionary computation, genetic algorithm, stochastic combinatorial optimization, convergence, multi-armed bandit

I. INTRODUCTION

Consider a stochastic combinatorial optimization (SCO) problem Ψ of $\max_{i \in A}(\mu_i := E_w[r(i,w)])$, where $A = \{1, 2, ..., n\}$ is a ***finite*** set of solutions, w is a random vector supported on a set $\Omega \subset \mathbb{R}^d$, $r : A \times \Omega \to \mathbb{R}$ is a reward function, and the expectation is taken with respect to a fixed unknown distribution P of the random vector w. Solving Ψ is to obtain an optimal solution in $\arg\max_{i \in A} \mu_i$ and the optimal value of $\max_{i \in A} \mu_i$. We assume that $\mu_i, i \in A$, is finite and that n is very large so that it is impractical to apply an enumeration method (or its variant). We further assume that a closed form expression for $E_w[r(i,w)]$ cannot be found but by sampling from P, samples $w^j, j = 1, 2, 3, ...$, of independent realizations of w can be generated and $r(i, w^j)$ can be evaluated explicitly for any $i \in A$ and sample $w^j \in \Omega$.

As reported by Bianchi *et al.* [2] (see also [10] for a literature survey on applying evolutionary computation (EC) to optimization problems under "uncertainty"), theoretical convergence

H. S. Chang is with the Department of Computer Science and Engineering at Sogang University, Seoul, Korea and can be reached by e-mail at hschang@sogang.ac.kr. This work was done while he was a visiting associate professor at ECE and ISR, University of Maryland, College Park.

analysis is *still missing* in the category of "EC with sampling estimated objective function" *for SCO*. Some properties on convergence *behavior* of EC for discrete optimization under noisy function evaluation have been studied by Miller and Goldberg [14] but the analysis is restricted to the "onemax bit-counting" function with no consideration for the general problem setup of Ψ. Even though several theoretical studies on the convergence of EC under noiseless function evaluation are available (see, e.g., [20] [18] [17] [19]), those results do not carry over to Ψ with our assumptions. Recently, Nakama [16] provided a theoretical convergence result of EC under noisy function evaluation based on a Markov chain model as in Rudolph [18] (see also [5]) for EC under noiseless function evaluation. Nakama assumes that for $i \in A$, the objective function with additive noise is given as $f(i)+X$, where $f : A \to \mathbb{R}$ and X is a random variable that takes only finite number of values $x_1 > x_2 \cdots > x_N$ and shows that an elitist-based genetic algorithm (GA) finds a global optimal solution in $\arg\max_{i\in A} f(i)$ by proving that a Markov chain model of the GA algorithm visits $(i^*, x_1), i^* \in \arg\max_{i\in A} f(i)$ with probability one. However, this is essentially same to the proof of the convergence of an elitist-based GA by Rudolph in [18] under noiseless function evaluation by considering the solution set $\{(i, x_j), i \in A, j = 1, ..., N\}$. The problem under consideration in [16] is fundamentally different from Ψ. Furthermore, the GA considered by Nakama does not consider the noise compensation technique of "explicit averaging" employed here, i.e., a Monte Carlo sampling for estimating $\mu_i, i \in A,$ by a sample average.

This paper presents two sampling-based EC algorithms that use the technique of explicit averaging in the category of EC with sampling estimated objective function for solving Ψ and establishes a convergence of the algorithms, extending Rudolph's works [17] [18] on the convergence analysis of EC (including GAs) under noiseless function evaluation or for deterministic combinatorial optimization problems (COPs). It is this paper's goal that is to fill the missing analysis of a theoretical convergence of EC for stochastic COPs but not to provide a novel competitive EC with other preexisting algorithms.

The algorithms analyzed in this paper are based on the on-line learning algorithms developed by Kleinberg *et al.* [12] for sleeping-experts and sleeping-bandits problems. Here a sequence $\{A_t, t = 1, 2, ..., T\}$ of nonempty subsets of A is given and each $i \in A$ corresponds to an expert or a bandit (or arm) so that all experts/bandits in $A \setminus A_t$ at t are "sleeping". At each round t, a play algorithm chooses $i_t \in A_t$ and a sample of reward $r(i, w)$ is available for all $i \in A_t$ in the

best-expert setting (the full-information setting) and a sample of $r(i_t, w)$ is available only for the chosen i_t in the multi-armed bandit setting (the partial-information setting). The goal is to devise a play algorithm that minimizes the regret given by $\sum_{t=1}^{T} \max_{i \in A_t} \mu_i - E[\sum_{t=1}^{T} \mu_{i_t}]$. Kleinberg *et al.* adapt the UCB1 algorithm in [1] for the multi-armed setting and the follow-the-leader algorithm in [11] for the best-expert setting, where each algorithm works under $A_t = A$ for all $t \geq 1$, into algorithms for the sleeping experts and bandits problems and provide some lower and upper finite-time bounds on the regret for both settings.

We modify the on-line learning algorithms by Kleinberg *et al.* into some variants of EC for solving Ψ with incorporation of the ϵ_t-greedy exploration and exploitation rule in reinforcement learning (RL) [21] for the choice of an elitist at time step t. Unlike the elitist concept under noiseless function evaluation [3], the elitist at time step t in our case is a "best currently-estimated solution" in A_t (an estimate of the optimal solutions in $\arg\max_{i \in A} \mu_i$) or is an expert uniformly selected over A in the best-expert setting (or over A_t in the multi-armed setting), where this choice is determined probabilistically with ϵ_t. We show that under some conditions on $\{\epsilon_t, t \geq 1\}$ and some assumptions on EC dynamics, the elitist becomes among the best solutions in A in the limit with probability one and this comes with the convergence to the optimal value.

II. Elitist-based Evolutionary Computation

The following description of EC model and the assumptions imposed on the model below are based on the model and the assumptions by Rudolph (see [17] for details). Let $S_t = (x_t^1, x_t^2, ..., x_t^m) \in A^m$ denote the population at time step t where $m \ll n$. Given the parent population S_t at time step t, an offspring is produced as follows. First, ρ parents in S_t are selected to serve as mates for recombination process by **mat** operation: **mat**: $A^m \rightarrow A^\rho$, $2 \leq \rho \leq m$. These solutions are recombined by **reco** operation: **reco**: $A^\rho \rightarrow A$ and further mutated by **mut** operation: **mut**: $A \rightarrow A$, yielding an offspring. After all r offspring have been produced in this way, selection procedure **sel**: $A^k \rightarrow A^m, k = q + r \geq m, 1 \leq q \leq m$, decides which offspring and possibly parents are in the next population S_{t+1}. In summary, for a given $S_t = (x_t^1, x_t^2, ..., x_t^m) \in A^m$ at each time step t, we have that for all $i = 1, ..., r$, $o_t^i = \textbf{mut}(\textbf{reco}(\textbf{mat}(x_t^1, ..., x_t^m)))$ and $S_{t+1} = (x_{t+1}^1, x_{t+1}^2, ..., x_{t+1}^m) = \textbf{sel}(x_t^1, ..., x_t^q, o_t^1, ..., o_t^r)$.

There exist many papers and books in the literature that discuss crossover, mutation, selection, etc. of the EC model in a detail (see, e.g., [4] [15] [6]). We omit a review on those processes

but impose the following assumptions A1 $\sim$ A3 for any $t \geq 1$ on the evolutionary process. (For the notational simplicity, an element in A^q also refers to a multiset of q possibly same solutions in A. That is, $(x_t^1, ..., x_t^q) \in A^q$ refers to the multiset $\{x_t^1, ..., x_t^q\}$ so that we use the membership notation with $(x_t^1, ..., x_t^q)$ to say, e.g, that $x_t^1 \in (x_t^1, ..., x_t^q)$.)

A1. There exists $\delta_r \in (0, 1]$ such that for all $x \in S_t$, $\Pr\{x \in \textbf{reco}(\textbf{mat}(x_t^1, ..., x_t^m))\} \geq \delta_r$.

A2. There exists $\delta_m \in (0, 1]$ such that for every pair $x, y \in A$, there exists a finite path $x^1, x^2, ..., x^l$ of pairwise distinct solutions with $x^1 = x$ and $x^l = y$ such that $\Pr\{x^{i+1} = \textbf{mut}(x^i)\} \geq \delta_m$ for all $i = 1, ..., l-1$.

A3. There exists $\delta_s \in (0, 1]$ such that for all $x \in (x_t^1, ..., x_t^k)$, $\Pr\{x \in \textbf{sel}(x_t^1, ..., x_t^k)\} \geq \delta_s$.

A. Best expert setting

The ϵ-greedy elitist-based EC algorithm in Figure 1 is based on the "follow the awake leader" (FTAL) algorithm in [12] for the sleeping-experts problems where at each time step, the awake expert that has the highest sample-average reward is chosen. The average is taken over the time steps when the expert was awake.

We modify the FTAL algorithm by incorporating the ϵ_t-greedy rule used in RL and the process of $\{S_t\}$, yielding a variant of elitist-based EC where the goal in our case is finding an optimal solution and obtaining the optimal value unlike minimizing the regret with respect to the best-expert ordering. Overall, the algorithm follows Rudolph's EC model for the population sequence $\{S_t, t \geq 1, S_t \in A^m\}$ generation process under the assumptions A1 $\sim$ A3, and it generates a sequence $\{A_t, t \geq 1, A_t \subseteq A\}$ of nonempty subsets of A where the sequence $\{A_t\}$ controls the sampling process of the algorithm in the best-expert setting. Every sampling is done independently from the past samples. It is the elitist that is probabilistically determined by the ϵ_t-greedy rule: at time step t, the elitist $x(t)$ corresponds to the awake expert that has the highest sample-average reward in A_t with probability $1 - \epsilon_t$ and to the expert in A selected by uniform distribution over A with probability ϵ_t. The elitist $x(t)$ is added then into A_{t+1} of awake experts at time step $t + 1$ by doing $A_{t+1} \leftarrow \{x(t)\} \cup \Lambda_t$, where Λ_t is the set which contains all *distinct* elements in the multiset S_t. In the algorithm, $|S_t| = m$ for all $t \geq 1$ but the size of A_t is varying. A_1 is set to be a singleton set with a uniformly sampled solution in A.

Roughly, the basic idea is that each expert in A is *infinitely often* woken up to estimate its objective function value and our choice for the elitist becomes greedy in the limit with respect

to the estimated objective function value among the currently awake experts. We ensure the estimated objective function value goes to the true value and the list of awake experts contains an optimal solution in the limit, which yields the elitist being an optimal solution in the limit. In fact, the convergence result of the pure random search adapted for SCO [9] also relies on the fact that every solution in A is sampled infinitely often to estimate μ_i for all $i \in A$.

The ϵ-greedy elitist-based EC for SCO

- Initialize $z_i = 0$ and $n_i = 0$ for all $i \in A$. Sample $i \in A$ with uniform distribution over A and set $A_1 = \{i\}$. Set $S_1 \in A^m$ arbitrarily.
- **For** $t = 1$ to T do
  - Observe $r(i, w^t)$ for all $i \in A_t$ by sampling w^t from P .
  - $z_i \leftarrow z_i + r(i, w^t)$ for all $i \in A_t$.
  - $n_i \leftarrow n_i + 1$ for all $i \in A_t$.
  - With probability ϵ_t, sample $x(t) \in A$ with uniform distribution over A and with probability $1 - \epsilon_t$, $x(t) \in \arg\max_{i \in A_t} \left(\frac{z_i}{n_i}\right)$ (ties broken arbitrarily).
  - Generate S_{t+1} from S_t and set $A_{t+1} \leftarrow \{x(t)\} \cup \Lambda_t$, where Λ_t is the set which contains all distinct elements in S_t.

end

Fig. 1. The ϵ-greedy elitist-based EC algorithm for SCO description

We let $z_{i,t} = \sum_{k=1}^{t} r(i, w^k) \cdot I\{i \in A_k\}$ and $n_{i,t} = \sum_{k=1}^{t} I\{i \in A_k\}$ for $t \geq 1$ and $i \in A$ from the parameters of the ϵ-greedy elitist-based EC, where $I\{i \in A_k\} = 1$ if $i \in A_k$ at time step k and 0 otherwise. (We will use the same notations in the next subsection and the function $I\{\cdot\}$ is also used in a similar fashion at other places.) Simply, $z_{i,t}$ corresponds to the value of z_i at time step t in the algorithm and similarly $n_{i,t}$ to n_i at time step t. We further let $\hat{\mu}_{i,t} = z_{i,t}/n_{i,t}$, for $t \geq 1$ and $i \in A$.

Theorem 2.1: Assume that $\sum_{t=1}^{\infty} \epsilon_t = \infty$ and $\lim_{t \to \infty} \epsilon_t = 0$ and that A1 $\sim$ A3 hold. Then for any realized sequence of nonempty sets $\{S_t, t = 1, ..., T\}$ in the ϵ-greedy elitist-based EC, we have that as $T \to \infty$, $\Pr\{x(T) \in \arg\max_{i \in A} \mu_i\} \to 1$ and $\max_{i \in A_T} \hat{\mu}_{i,T} \to \max_{i \in A} \mu_i$ with probability one.

Note that the assumption on $\{\epsilon_t\}$ can be satisfied with for example, $\epsilon_t = 1/t, t \geq 1$. The proof below is partly based on the proof technique of Theorem 2.1 in [17] by Rudolph.

Proof: For any $i \in A$, $\sum_{t=1}^{\infty} \Pr\{i \in A_t\} \geq \sum_{t=1}^{\infty} \epsilon_t/|A| = \infty$ by the assumption. Therefore,

for any $i \in A$, the number of times i was included in a set in the sequence $\{A_t\}, t = 1, ..., T$ goes to infinity as $T \to \infty$ by the extended Borel-Cantelli lemma in Singh *et al.* [21, Lemma 4]. That is, $\lim_{T\to\infty} \sum_{t=1}^{T} I\{i \in A_t\} = \infty$ (a.s.). This implies that as $T \to \infty$, $n_{i,T} \to \infty$ for all $i \in A$ so that $\hat{\mu}_{i,T} \to \mu_i$ almost surely (a.s.) by the strong law of large numbers. We have that for any $\epsilon > 0$, there exists $t(\epsilon) < \infty$ such that for all $t > t(\epsilon)$ and $i \in A$,

$$\Pr\{|\hat{\mu}_{i,t} - \mu_i| < \epsilon\} > 1 - \epsilon.$$

This implies that for all $t > t(\epsilon)$,

$$\Pr\left\{\left|\max_{i\in A_t} \hat{\mu}_{i,t} - \max_{i\in A_t} \mu_i\right| < \epsilon\right\} > 1 - \epsilon,$$

which further implies that

$$\Pr\{\mu_{x(t)} + 2\epsilon \geq \max_{i\in A_t} \mu_i\} \geq (1 - \epsilon_t)(1 - \epsilon), t > t(\epsilon).$$

Now from the assumptions A1 $\sim$ A3, the probability that an optimal solution in $\arg\max_{i\in A} \mu_i$ has not been found in $\{S_t, t \geq t(\epsilon)\}$ after $l - t(\epsilon)$ steps $(l \geq t(\epsilon))$ is at most

$$\left(1 - (\delta_r \delta_m \delta_s)^{l^*-1} \delta_r \delta_m\right)^{\lfloor (l-t(\epsilon))/l^* \rfloor},$$

where

$$l^* = \max_{x \notin \arg\max_{i\in A} \mu_i} \left\{\text{the length of the shortest path between } x \text{ to the set } \arg\max_{i\in A} \mu_i\right\}.$$

This probability converges to zero as $l \to \infty$, which implies that there exists k such that $t(\epsilon) < k < \infty$ and $\Pr\{\arg\max_{i\in A} \mu_i \cap A_k \neq \emptyset\} = 1$ because $A_{t+1} \leftarrow \{x(t)\} \cup S_t$ at each t (cf., the proof of Theorem 2.1 by Rudolph in [17]). Furthermore, observe that $\Pr\{\exists k' > k$ such that $\arg\max_{i\in A} \mu_i \cap A_{k'} \neq \emptyset | \arg\max_{i\in A} \mu_i \cap A_k \neq \emptyset\} = 1$. In other words, the number of times an optimal solution is included in a set in the sequence of $\{A_t, t \geq t(\epsilon)\}$ is infinite with probability one because $\Pr\{\lim_{T\to\infty} \sum_{t=t(\epsilon)}^{T} I\{\arg\max_{i\in A} \mu_i \cap S_t \neq \emptyset\} = \infty\} = 1$.

Therefore, we have that for any $\epsilon > 0$ arbitrarily close to zero, there exists $t'(\epsilon) < \infty$ such that $\epsilon_{t'(\epsilon)}$ is arbitrarily close to zero and $\arg\max_{i\in A} \mu_i \cap A_{t'(\epsilon)} \neq \emptyset$, which makes at $t = t'(\epsilon)$, $\Pr\{x(t) \in \arg\max_{i\in A} \mu_i\}$ arbitrarily close to one because at $t = t'(\epsilon)$, $\Pr\{\mu_{x(t)} + 2\epsilon \geq \max_{i\in A_t} \mu_i\}$ becomes arbitrarily close to $\Pr\{x(t) \in \arg\max_{i\in A_t} \mu_i\}$ and for any t,

$$\Pr\{x(t) \in \arg\max_{i\in A} \mu_i\} \geq \Pr\{\arg\max_{i\in A} \mu_i \cap A_t \neq \emptyset\} \Pr\{x(t) \in \arg\max_{i\in A_t} \mu_i\}.$$

This finally implies that there exists $t < \infty$ such that $\Pr\{x(t+k) \in \arg\max_{i \in A} \mu_i\}$ arbitrarily close to one for all $k \geq 0$. We conclude that $\Pr\{x(T) \in \arg\max_{i \in A} \mu_i\} \rightarrow 1$ as $T \rightarrow \infty$.

The convergence of $\arg\max_{i \in A_T} \hat{\mu}_{i,T}$ to $\max_{i \in A} \mu_i$ with probability one then follows from the above. ■

We remark that as stated in the theorem, the convergence here is not defined with respect to a specific optimal solution in $\arg\max_{i \in A} \mu_i$. This can be avoided if desired by a consistent selection instead of breaking ties arbitrarily, e.g., breaking ties with the smallest index solution. Moreover, we can view the above algorithm as a generalization of ordinal optimization [7] and sample average approximation [13] because by letting $A_t = A$ and $\epsilon_t = 0$ for all t and ignoring the generation process of $\{S_t, t \geq 1\}$, $x(t)$ corresponds to the estimated best solution based on sample means from t samples of w_i for each $i \in A$.

Even though our presentation is within the context of EC, a more generalized result can be stated with the sequence of $\{S_t, t \geq 1\}$: Assume that $\sum_{t=1}^{\infty} \epsilon_t = \infty$ and $\lim_{t \rightarrow \infty} \epsilon_t = 0$. If for the random sequence $\{S_t, t \geq 1, S_t \in A^m\}$ defined over a probability space, $\Pr\{\exists k < \infty$ such that $\arg\max_{i \in A} \mu_i \cap S_k \neq \emptyset\} = 1$ and $\Pr\{\exists k'$ such that $k < k' < \infty$ and $\arg\max_{i \in A} \mu_i \cap S_{k'} \neq \emptyset | \arg\max_{i \in A} \mu_i \cap S_k \neq \emptyset\} = 1$ for any integer $k \geq 1$, then we have that as $T \rightarrow \infty$, $\Pr\{x(T) \in \arg\max_{i \in A} \mu_i\} \rightarrow 1$ and $\max_{i \in A_T} \hat{\mu}_{i,T} \rightarrow \max_{i \in A} \mu_i$ with probability one.

B. Multi-armed bandit setting

For this setting, we replace the assumption A2 that for every pair $x, y \in A$, there exists a finite path $x^1, x^2, ..., x^l$ of pairwise distinct solutions with $x^1 = x$ and $x^l = y$ such that $\Pr\{x^{i+1} = \text{mut}(x^i)\} \geq \delta_m > 0$ for all $i = 1, ..., l-1$ by the assumption A2' below:

A2'. There exists $\delta_m \in (0, 1]$ such that for every pair $x, y \in A$, $\Pr\{y = \textbf{mut}(x)\} \geq \delta_m$.

Note that A2' implies A2 and that A2' corresponds to putting in essence the pure random search dynamics into EC.

The idea of the bandit elitist-based EC presented in Figure 2 is similar to the ϵ-greedy elitist-based EC but in this algorithm, if a sample of reward has been drawn before for all solutions in $A_t, t \geq 1$, then only one sample of reward for the chosen solution i_t at time step t is necessary. It is based on the "awake upper estimated reward" (AUER) algorithm in [12]. As before, the algorithm keeps track of the running average of rewards sampled from each arm (solution), but

also a confidence interval of width $2\sqrt{8\ln t/n_{j,t}}, j \in A$. The elitist is determined by the ϵ_t-greedy rule but here at time step t, the elitist $x(t)$ corresponds to the awake arm that has the highest "upper estimated reward" via the estimated confidence interval (instead of the highest sample-average reward) with probability $1-\epsilon_t$ and to the arm in A_t selected by uniform distribution over A_t with probability ϵ_t. The elitist $x(t)$ is added then into A_{t+1} of awake arms at time step $t+1$ by doing $A_{t+1} \leftarrow \{x(t)\} \cup \Lambda_t$. To get away with the possible problem of dividing by zero at each time step t for obtaining $x(t) \in \arg\max_{i \in A_t}(z_i/n_i + \sqrt{8\ln t/n_i})$, we have that for all $i \in A_t$ such that $n_i = 0$, we observe $r(i, w^0)$ by sampling w^0 from P and set $z_i = r(i, w^0)$ and $n_i = 1$.

We also impose a uniqueness assumption on the optimal solution set and a bounded-interval reward function assumption due to the technicality of the convergence proof unlike the ϵ-greedy elitist-based EC case.

The bandit elitist-based EC for SCO

- Initialize $z_i = 0$ and $n_i = 0$ for all $i \in A$. Sample $i \in A$ with uniform distribution over A and set $A_1 = \{i\}$. Set $S_1 \in A^m$ arbitrarily.
- **For** $t = 1$ to T do
  - For all $i \in A_t$ such that $n_i = 0$, observe $r(i, w^0)$ by sampling w^0 from P and set $z_i = r(i, w^0)$ and $n_i = 1$.
  - With probability ϵ_t, sample $x(t) \in A_t$ with uniform distribution over A_t and with probability $1-\epsilon_t$, $x(t) \in \arg\max_{i \in A_t}\left(\frac{z_i}{n_i} + \sqrt{\frac{8\ln t}{n_i}}\right)$ (ties broken arbitrarily).
  - Observe $r(x(t), w^t)$ by sampling w^t from P.
  - $z_{x(t)} \leftarrow z_{x(t)} + r(x(t), w^t)$.
  - $n_{x(t)} \leftarrow n_{x(t)} + 1$.
  - Generate S_{t+1} from S_t and set $A_{t+1} \leftarrow \{x(t)\} \cup \Lambda_t$, where Λ_t is the set which contains all distinct elements in S_t.

end

Fig. 2. The bandit elitist-based EC algorithm for SCO description

Theorem 2.2: Assume that $\sum_{t=1}^{\infty} \epsilon_t = \infty$ and $\lim_{t\to\infty} \epsilon_t = 0$ and that A1, A2', and A3 hold. Also assume that Ψ has a unique optimal solution and $r : A \times \Omega \to [0,1], i \in A$. Then for any realized sequence of nonempty sets $\{S_t, t = 1, ..., T\}$ in the bandit elitist-based EC, we have that as $T \to \infty$, $\Pr\{x(T) = \arg\max_{i \in A} \mu_i\} \to 1$ and $\max_{i \in A_T} \hat{\mu}_{i,T} \to \max_{i \in A} \mu_i$ with probability

one.

Proof: From the uniqueness assumption, let $\arg\max_{i\in A}\mu_i = \{1\}$. First, for any $t \geq 1$, we have that

$$\Pr\{x(t)\neq 1|1\in A_t\} = \sum_{j\in A_t, j\neq 1}\Pr\{x(t)=j|1\in A_t\} \tag{1}$$

$$\leq \sum_{j\in A_t,j\neq 1}\left(\epsilon_t\frac{1}{|A_t|}+(1-\epsilon_t)\Pr\left\{\hat{\mu}_{j,t}+\sqrt{\frac{8\ln t}{n_{j,t}}}\geq\hat{\mu}_{1,t}+\sqrt{\frac{8\ln t}{n_{1,t}}}\right\}\right) \tag{2}$$

$$\leq \sum_{j\in A_t,j\neq 1}\left(\epsilon_t\frac{1}{|A_t|}+(1-\epsilon_t)\left(\Pr\left\{\hat{\mu}_{j,t}\geq\mu_j+\sqrt{\frac{8\ln t}{n_{j,t}}}\right\}+\Pr\left\{\hat{\mu}_{1,t}\leq\mu_1-\sqrt{\frac{8\ln t}{n_{1,t}}}\right\}\right.\right.$$
$$\left.\left.+\Pr\left\{\mu_1<\mu_j+2\sqrt{\frac{8\ln t}{n_{j,t}}}\right\}\right)\right) \tag{3}$$

$$\leq \sum_{j\in A_t,j\neq 1}\left(\epsilon_t\frac{1}{|A_t|}+(1-\epsilon_t)\left(\frac{2}{t^4}+\Pr\left\{\mu_1<\mu_j+2\sqrt{\frac{8\ln t}{n_{j,t}}}\right\}\right)\right), \tag{4}$$

where the last inequality of (4) comes from using Chernoff-Hoeffding bound [8] for the first two probability terms in (3).

Therefore, if $n_{j,t}\geq 32\ln t/(\mu_1-\mu_j)^2$, then

$$\Pr\{x(t)\neq 1|1\in A_t\}\leq\sum_{j\in A_t,j\neq 1}\left(\epsilon_t\frac{1}{|A_t|}+(1-\epsilon_t)\frac{2}{t^4}\right),$$

which goes to zero as $t\to\infty$ because $\epsilon_t\to 0$ as $t\to\infty$.

Second, for all $j\in A$, $n_{j,t}\to\infty$ as $t\to\infty$ because $\sum_{t=1}^{\infty}\Pr\{x(t)=j\}\geq\sum_{t=1}^{\infty}\Pr\{x(t)=j\}\Pr\{j\in A_t\}\geq\sum_{t=1}^{\infty}\epsilon_t|A_t|^{-1}\Pr\{j\in A_t\}\geq|A|^{-1}\delta_m\sum_{t=1}^{\infty}\epsilon_t=\infty$ where the last inequality follows from the assumption A2'. Therefore, the event $\{n_{j,t}\geq 32\ln t/(\mu_1-\mu_j)^2\}$ happens with probability one at some finite time step t and consequently as $t\to\infty$, $\hat{\mu}_{j,t}\to\mu_j$ (a.s.) by the strong law of large numbers.

By then using the similar arguments with the previous proof for Theorem 2.1, we can show that for any $\epsilon>0$ arbitrarily close to zero, there exists $t'(\epsilon)<\infty$ that can make $\epsilon_{t'(\epsilon)}$ arbitrarily close to zero and $\arg\max_{i\in A}\mu_i\cap A_{t'(\epsilon)}\neq\emptyset$ so that $\Pr\{x(t'(\epsilon))=\arg\max_{i\in A}\mu_i\}$ arbitrarily close to one, yielding $\Pr\{x(T)=\arg\max_{i\in A}\mu_i\}\to 1$ as $T\to\infty$. The optimal value convergence follows similarly. ■

We remark that letting $A_t=A$ for all t and replacing $x(t)\in\arg\max_{i\in A_t}\left(z_i/n_i+\sqrt{8\ln t/n_i}\right)$ with $x(t)\in\arg\max_{i\in A_t}(z_i/n_i)$ in the bandit elitist-based EC yields a similar algorithm to the

ϵ_n-GREEDY algorithm in [1]. Note that $\sum_{n=1}^{\infty} \epsilon_n = \infty$ and $\lim_{n\to\infty} \epsilon_n = 0$ in the ϵ_n-GREEDY algorithm. It turns out that doing a selection-probability analysis as in Theorem 3 in [1] for the bandit elitist-based EC replaced with $x(t) \in \arg\max_{i\in A_t} (z_i/n_i)$ seems difficult since it is difficult to obtain both lower and upper bounds on the expectation and the variance of the number of plays in which a suboptimal machine (solution) was chosen by uniform selection in the first n plays in order to apply Bernstein's inequality as used in [1] and this makes the proof of the convergence difficult. But because each solution is included in a set infinitely often in the sequence $\{A_t, t \geq 1\}$ and sampled infinitely often as an elitist in the sequence $\{x(t), t \geq 1\}$, and for $\{x(t), t \geq 1\}$, $x(t)$ becomes the optimal solution in the limit, we expect that for the bandit elitist-based EC replaced with $x(t) \in \arg\max_{i\in A_t} (z_i/n_i)$, we have the similar convergence of $\Pr\{x(T) = \arg\max_{i\in A} \mu_i\} \to 1$ as $T \to \infty$.

III. Concluding Remarks

Boltzmann exploration rule can be used instead of the ϵ_t-greedy rule in the ϵ-greedy elitist-based EC and the bandit elitist-based EC while preserving the same convergence guarantee. See Appendix B.1 in [21] for the GLIE (greedy in the limit with infinite exploration) learning-policy property.

References

[1] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, pp. 235–256, 2002.

[2] L. Bianchi, M. Dorigo, L. M. Gambardella, and W. J. Gutjahr, "A survey on metaheuristics for stochastic combinatorial optimization," *Natural Computing*, to appear.

[3] K. A. De Jong, *An Analysis of the Behavior of a Class of Genetic Adaptive Systems*, Ph.D. Thesis, Univ. of Michigan, Ann Arbor, MI, 1975.

[4] K. A. De Jong, *Evolutionary Computation*, The MIT Press, 2002.

[5] A. E. Eiben, E. H. L. Aarts, and K. M. van Hee, "Global convergence of genetic algorithms: a Markov chain analysis," *Parallel Problem Solving from Nature* (H.-P. Schwefel and R. Männer, eds.), Springer, pp. 4–12.

[6] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, 1989.

[7] Y. C. Ho, C. G. Cassandras, C. H. Chen, and L. Dai, "Ordinal optimization and simulation," *J. Oper. Res. Society*, vol. 51, pp. 490–500, 2000.

[8] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol.58, pp. 13–30, 1963.

[9] T. Homem-De-Mello, "Variable-sample methods for stochastic optimization," *ACM Trans. on Modeling and Computer Simulation*, vol. 13, no. 2, pp. 108–133, 2003.

[10] Y. Jin and J. Branke, "Evolutionary optimization in uncertain environments-a survey," *IEEE Trans. on Evolutionary Computation*, vol. 9, no. 3, pp. 303–317, 2005.

[11] A. Kalai and S. Vempala, "Efficient algorithms for on-line optimization," *J. Computer and System Sciences*," vol. 71, no. 3, pp. 291–307, 2005.

[12] R. D. Kleinberg, A. Niculescu-Mizil, and Y. Sharma, "Regret Bounds for Sleeping Experts and Bandits," in *Proc. of the 21st Annual Conference on Learning Theory* (COLT), 2008, pp. 425-436.

[13] A. J. Kleywegt, A. Shapiro, and T. Homem-De-Mello, "The sample average approximation method for stochastic discrete optimization," *SIAM J. Optim.*, vol. 12, no. 2, pp. 479–502, 2001.

[14] B. L. Miller and D. E. Goldberg, "Genetic algorithms, selection schemes, and the varying effects of noise," *Evolutionary Computation*, vol. 4, no. 2, pp. 113–131, 1997.

[15] M. Mitchell, *An Introduction to Genetic Algorithms*, the MIT press, 1998.

[16] T. Nakama, "Theoretical analysis of genetic algorithms in noisy environments based on a Markov model," in *Proc. of the Genetic and Evolutionary Computation (GECCO) Conf.*, 2008, pp. 1001–1008.

[17] G. Rudolph, "Finite Markov chain results in evolutionary computation: a tour d'horizon," *Fundamenta Informaticae*, vol. 35, no. 1-4, pp. 67–89, 1998.

[18] G. Rudolph, "Convergence analysis of canonical genetic algorithms," *IEEE Trans. on Neural Networks*, vol. 5, no. 1, pp. 96–101, 1994.

[19] L. M. Schmitt, "Theory of genetic algorithms II: models for genetic operators over the string-tensor representation of populations and convergence to global optima for arbitrary fitness function under scaling," *Theoretical Computer Science*, vol. 310, pp. 181–231, 2004.

[20] M. A. Semenov and D. A. Terkel, "Analysis of convergence of an evolutionary algorithm with self-adaptation using a stochastic Lyapunov function," *Evolutionary Computation*, vol. 11, no. 4, pp. 363–379, 2003.

[21] S. Singh, T. Jaakkola, M. Littman, and C. Szepesvari, "Convergence results for single-step on-policy reinforcement-learning algorithms," *Machine Learning*, vol. 39, pp. 287–308, 2000.