

Convergence of Estimation of Distribution Algorithms in Optimization of Additively Noisy Fitness Functions

Yi Hong¹, Qingsheng Ren¹, Jin Zeng², Yuchou Chang³

¹Department of Computer Science and Engineering, Shanghai Jiao tong University, P.R.China

²Department of Mathematics, Shanghai Jiaotong University, P.R.China

³Institute of ImageProcessing and Pattern Recognition

goodji@sjtu.edu.cn

Abstract

Noise is a common phenomenon in many real-world optimizations. It has long been argued that evolutionary algorithm (EA) should be relatively robust against it. As a novel computing model in evolutionary computations, Estimation of Distribution Algorithm (EDA) is also encountered with it. This paper initially presents three dynamic models of EDA under the additively noisy environment with three different selection methods (proportional selection method, truncation selection method and tournament selection method). We verify that when the population size is infinite, EDA can converge to the global optimal point. This concept establishes the theoretic foundation for optimization of noisy fitness functions with EDA.

1. Introduction

Genetic algorithm (GA) [1] is a population based search method inspired from the notion of Darwinian evolution: Survival of the Fittest. According to this notion, GA employs such three simple operators as selection, crossover and mutation to generate the new populations. GA is simple to execute, and can search the global optimal point not depending on the gradient of target functions. Thus, it has been applied to a very wide range of problems. But Darwinian-type genetic algorithm is just semi-blind: positions of crossover and mutation are random. This randomness has greatly delayed the convergence speed of GA. Low convergence efficacy has become a major obstacle for its generalizations in real-world applications.

To tackle the problem of low efficacy of GA, many effective methods have been proposed. Among them is Estimation of Distribution Algorithm (EDA). EDA was firstly proposed by Mühlenbein in 1996 [2], and has become a novel research area in evolutionary computation(EC)[3]. Completely abandoning crossover

operator and mutation operator, EDA turns to analyze the selected subpopulations to estimate their distribution, and then uses this estimated probability distribution to produce the population in the next generation. The kernel of EDA is how to estimate their distribution. To realize it, different methods have been proposed. Univariate Marginal Distribution Algorithm (UMDA) is perhaps the simplest EDAs. It supposes that all variables are independent, and uses their joint estimated normal distribution to generate the new populations. A more complex model is MIMIC, which supposes that variables are pairwise dependent. BOA and BEA are perhaps the most commonly-used EDAs. Both of them are based on Bayesian network, and their performances are very satisfactory. The frame of EDA is shown in figure 1. EDA has greatly improved the speed of genetic algorithm, and sometimes is more effective to solve some hard deceptive problems. But EDA is just a fresh research area, where there are so many works to be done. In this paper, we analyze the convergent performance of EDA under the additively noisy conditions.

Arrangement of this paper is as follows. In section 2, we briefly introduce the existed models for EDA in noiseless conditions. Section 3 presents the dynamic models for EDA with three different selection methods under additively noisy conditions. In section 4, we verify that EDA can converge to the global optimal point under additively noisy conditions.

2. Overview of Mathematic Models of EDA

To better understand the behavior of EDA, some mathematic models have been established. The most commonly used two models are Markov chains and dynamic systems. A status in Markov chain only depends on its previous one, so if we know the current status, the probability of its subsequence can also be

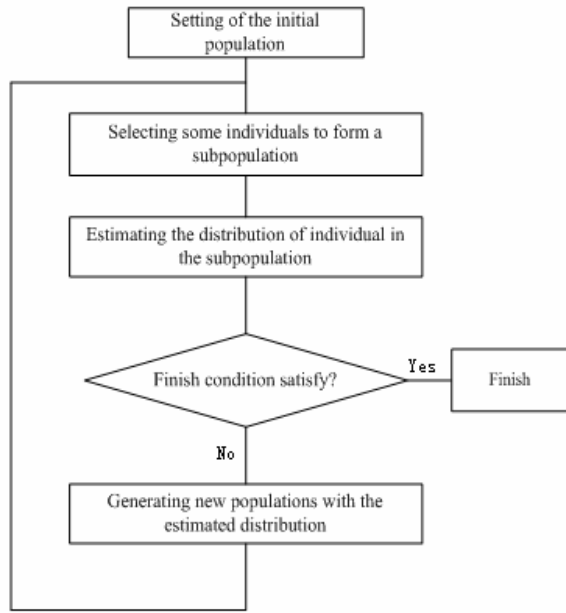


Fig. 1. Frame of EDA

calculated. EDA is in fact a Markov chain. The reason is that the population in step l only depends on the population at the step $l-1$. Moreover, neither operator used for the calculation of the transition probabilities depends on the parameter at the step l . Gonzalez and Larranaga have expatiated Markov model for EDA. He also uses this model to analyze the convergence of many kinds of EDAs. Compared with Markov chains, the dynamic model is much simpler. Mahnig and Mühlenbein drew a dynamic expression of variable probability in for UMDA, and similar result can also be seen in [3]. Berny shows that Reinforced Learning and PBIL algorithms can be derived from gradient dynamical systems acting on the probability vectors. A generally used result was given by Zhang in [4]. Zhang verified that when the population size is infinite, EDA can converge to the global optimal point under proportional selecting method, truncation selecting method and tournament selecting method. This paper just expands Zhang's result into noisy environment. In noisy environment, the behavior of UMDA has been studied by Hong in [5], where he also shows that compared with GA, EDA is more suitable to tackle noisy optimizations. His analysis is based on a certain kind of simple functions called Womax problems. He argues that UMDA can converge to the global optimal point for Womax problems even there are noise in target functions. In this study, we verified that EDA can converge to the global optimal point under the additively noisy environment.

3. Dynamic Model of EDA under Additively Noisy Environment

All existed models for EDAs are drawn under the noiseless conditions. But since noise will affect the selection operator: a good individual probably loses, while a bad individual probably wins. The existed models must be adjusted under noisy conditions. Optimization of noisy fitness functions can be defined as:

$$\max_x F(x) = f(x) + \delta \quad x \in D \quad (1)$$

Here we suppose that δ is an additive normal distributed noise. $\delta \sim N(0, \sigma^2)$. This assumption will let the modeling of noise in the target function more simply, and it also satisfies many real-world applications. In addition, we suppose that $f(x)$ is continuous and has only one global maximal point, and its value is positive and limited. That is:

$$0 < f(x) < +\infty$$

The last but also the most important premise is that the distribution of the population approximates that of their parents as closely as possible.

Define the following parameters:

$pop(x, t)$: The estimated distribution of individuals in the population in the t^{th} generation;

$pop^s(x, t)$: The estimated distribution of individuals in the selected subpopulation in the t^{th} generation;

$p(x, t)$: The ideal probability distribution of individuals in the population in the t^{th} generation;

$p^s(x, t)$: The ideal probability distribution of individuals in the selected subpopulation in the t^{th} generation.

G^* : The global optimal values of $f(x)$, x^* is the one and only optimal point.

When the population size is infinite, the following equations satisfy:

$$p(x, t) = Pop(x, t) \quad (2)$$

$$P^s(x, t) = Pop^s(x, t) \quad (3)$$

Consider our suppositions listed above, we get:

$$p(x, t+1) = Pop^s(x, t) \quad (4)$$

Calculating the expected value of population fitness in the t^{th} generation,

$$\begin{aligned} H(t) &= \int_D F(x) p(x, t) dx \\ &= \int_D (f(x) + \delta) p(x, t) dx \\ &= \int_D f(x) p(x, t) dx \end{aligned} \quad (5)$$

3.1. Dynamic Model of EDA under the Proportional Selection

Proportion selection operator selects the individuals according to the fitness of an individual. The probability of an individual x in the next generation can be calculated,

$$\begin{aligned} Pop^s(x,t) &= \frac{E(F(x))p(x,t)}{H(t)} \\ &= \frac{f(x)p(x)}{H(t)} \end{aligned} \quad (6)$$

Where E is expectation operator for a random variable. This operator will be applied throughout of this paper. According to (6), we can model the proportional selection as:

$$p(x,t+1) = \frac{f(x)p(x,t)}{H(t)} \quad (7)$$

(7) coincides with the model of EDA under the noiseless conditions.

3.2. Dynamic Model of EDA under the Tournament Selection

Tournament selection chooses K individuals from the population randomly, and compares their fitness. The best one is selected into the subpopulation. K is called the tournament size. Commonly $K = 2$. Here we also set $K = 2$.

$$\begin{aligned} \therefore \delta &\sim N(0, \sigma^2) \\ \therefore F(x) &\sim N(f(x), \sigma^2) \end{aligned}$$

Consider any two different individuals x and y from the population,

$$F(x) - F(y) \sim N(f(x) - f(y), 2\sigma^2)$$

Then

$$p(F(x) - F(y) \geq 0) = \Phi\left(\frac{f(x) - f(y)}{\sqrt{2}\sigma}\right)$$

Where $\Phi(t)$ is the standard normal cdf.

$$\Phi(t) = \int_{-\infty}^t \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

$Pop^s(x,t)$ can be calculated as follows:

$$\begin{aligned} Pop^s(x,t) &= 2p(x,t) \int_{F(y) \leq F(x)} p(y,t) dy \\ &= 2p(x,t) \left(\int_D p(y,x) \Phi\left(\frac{f(x) - f(y)}{\sqrt{2}\sigma}\right) dy \right) \end{aligned} \quad (8)$$

According to (8), we can model the tournament selection as:

$$p(x,t) = 2p(x,t) \left(\int_D p(y,x) \Phi\left(\frac{f(x) - f(y)}{\sqrt{2}\sigma}\right) dy \right) \quad (9)$$

3.3. Dynamic Model of EDA under the Truncation Selection

Truncation selection operator first ranks all the individuals according to their fitness, then select N best

individuals to form the subpopulation, N is the subpopulation size.

$$p(x,t+1) = \frac{p(x,t)}{C} \times \Phi\left(\frac{f(x) - \beta(t)}{\sigma}\right) \quad (10)$$

Where

$$C = \frac{N}{Popsize} = \int_{F(x) \geq \beta(t)} p(x,t) dx$$

C is a constant parameter. $Popsize$ is population size.

4. Convergence of EDA under Additively Noisy Condition

lemma 1. If $f(x)$ is a continuous function in the variable domain D , then the probability distribution of individuals $p(x,t)$ is also continuous in the domain D .

Proof.

Since the initial population are generated randomly, then,

$$p(x,0) = \frac{1}{\int_D 1 ds}$$

is a continuous function in domain D . Consider the three equals (7)(9)(10), when $p(x,t)$ is continuous, we can draw the conclusion that $p(x,t+1)$ is continuous. So $p(x,t)$ is continuous function during the evolution.

lemma 2. If the following conditions satisfy,

$$\lim_{t \rightarrow \infty} H(t) = G^* \quad (11)$$

The population have converged to the global optimal point.

Proof.

According (5), $H(t) = \int_D f(x)p(x,t) dx$

since the global maximal value of $f(x)$ is G^* , that is $f(x) \leq G^*$, then

$$H(t) = \int_D f(x)p(x,t) dx \leq G^* \int_D p(x,t) dx \leq G^* \quad (12)$$

In (12), when $H(t) = G^*$, $f(x) \equiv G(x^*)$. Since $f(x)$ has a single global optima, thus, all the individuals in the current generation converge to the single optimal point.

Theorem 1. With the proportional selection, EDAs can converge to the global optimal point under the additive noisy environment.

Proof.

According the proportional selection method (7),

$$\begin{aligned} H(t+1) &= E\left(\int_D \frac{f^2(x)p(x,t)}{H(t)} dx\right) + E(\delta) \times E\left(\int_D \frac{p(x,t) \times f(x)}{H(t)} dx\right) \\ &= \int_D \frac{f^2(x)p(x,t)}{H(t)} d(x) \end{aligned}$$

Then,

$$H(t+1) - H(t) = \frac{\int_D [f(x) - H(t)]^2 p(x,t) dx}{H(t)} \geq 0 \quad (13)$$

From (13), $H(t+1) \geq H(t)$. Since $H(t) < +\infty$, $\lim_{t \rightarrow \infty} H(t)$ exists. Suppose that $H(t)$ converges to G ($G < G^*$), define the set Q :

$$Q = \{x | f(x) > G\}$$

For all the individuals in Q

$$p(x,t) = E \left(p(x,0) \frac{f(x)}{E(t-1)} \frac{f(x)}{E(t-2)} \dots \frac{f(x)}{E(0)} \right) > p(x,0) \left(\frac{f(x)}{G} \right)^t$$

Since $f(x) > G$, then,

$$\lim_{t \rightarrow \infty} p(x,t) = +\infty \quad (14)$$

From (14) and $p(x,t)$ is a continuous probability distribution function,

$$\lim_{t \rightarrow \infty} \int_Q p(x,t) = +\infty \quad (15)$$

(15) contradicts the performance of pdf. So the suppose that $H(t)$ converges to G ($G < G^*$) doesn't hold. Then

$$\lim_{t \rightarrow \infty} H(t) = G^*$$

EDA converges to the global optimal point.

Theorem 2. With the tournament selection, EDAs can converge to the global optimal point under the additive noisy environment.

Proof.

According the tournament selection method (9),

$$E(t+1) = E \left(\int_D f(x) p(x,t+1) dx \right) = 2E \left(\int_D f(x) p(x,t) \int_D p(y,t) \Phi \left(\frac{f(x) - f(y)}{\sqrt{2}\sigma} \right) dy dx \right)$$

and

$$\begin{aligned} & E \left(\int_D f(x) p(x,t) \int_D p(y,t) \Phi \left(\frac{f(x) - f(y)}{\sqrt{2}\sigma} \right) dy dx \right) + \\ & E \left(\int_D f(x) p(x,t) \int_D p(y,t) \Phi \left(\frac{f(y) - f(x)}{\sqrt{2}\sigma} \right) dy dx \right) \\ & = E \left(\int_D f(x) p(x,t) \left(\Phi \left(\frac{f(x) - f(y)}{\sqrt{2}\sigma} \right) + \Phi \left(\frac{f(y) - f(x)}{\sqrt{2}\sigma} \right) \right) dy dx \right) \\ & = E \left(\int_D f(x) p(x,t) dx \right) = H(t) \end{aligned}$$

Then

$$\begin{aligned} & E \left(\int_D f(x) p(x,t) \int_D p(y,t) \Phi \left(\frac{f(x) - f(y)}{\sqrt{2}\sigma} \right) dy dx \right) \\ & - E \left(\int_D f(x) p(x,t) \int_D p(y,t) \Phi \left(\frac{f(y) - f(x)}{\sqrt{2}\sigma} \right) dy dx \right) \\ & = E \left(\int_D p(x,t) \int_D p(y,t) (f(x) - f(y)) \Phi \left(\frac{f(x) - f(y)}{\sqrt{2}\sigma} \right) dy dx \right) \\ & = \iint_S p(x)p(y) (f(x) - f(y)) \Phi \left(\frac{f(x) - f(y)}{\sqrt{2}\sigma} \right) ds \\ & = \iint_{f(x) \geq f(y)} p(x)p(y) (f(x) - f(y)) \Phi \left(\frac{f(x) - f(y)}{\sqrt{2}\sigma} \right) ds \\ & \quad + \iint_{f(x) \leq f(y)} p(x)p(y) (f(x) - f(y)) \Phi \left(\frac{f(x) - f(y)}{\sqrt{2}\sigma} \right) ds \\ & \geq 1/2 \iint_{f(x) \geq f(y)} p(x)p(y) (f(x) - f(y)) ds \\ & \quad + 1/2 \iint_{f(x) \leq f(y)} p(x)p(y) (f(x) - f(y)) ds \\ & = 0 \end{aligned}$$

Thus

$$H(t+1) - H(t) \geq 0 \quad (16)$$

According to (16) and $H(t) < +\infty$, $\lim_{t \rightarrow \infty} H(t)$ exists.

Suppose

$$\lim_{t \rightarrow \infty} H(t) = G \quad (G < G^*)$$

For the global optimal individual x^* , $f(x^*)$ satisfies:

$$\int_D p(y,t) \Phi \left(\frac{f(x^*) - f(y)}{\sqrt{2}\sigma} \right) dy = 0.5 + \beta \quad (0 < \beta < 1)$$

Since $\int_D p(y,t) \Phi \left(\frac{f(x) - f(y)}{\sqrt{2}\sigma} \right) dy$ is a monotonous increasing continuous function, then there exists \tilde{x} satisfying:

$$\int_D p(y,t) \Phi \left(\frac{f(x^*) - f(y)}{\sqrt{2}\sigma} \right) dy = 0.5 + \kappa \quad (0 < \kappa < \beta)$$

Define the following set Q :

$$Q = \{x | f(x) > f(\tilde{x})\}$$

Thus, for all individuals in the Q ,

$$p(x,t+1) - p(x,t) > 2\kappa p(x,t) \quad (17)$$

From (17), we get:

$$\lim_{t \rightarrow \infty} \int_Q p(x,t) = +\infty$$

This equation contradicts the performance of pdf. So the suppose that $E(t)$ converges to G ($G < G^*$) doesn't hold. Then

$$\lim_{t \rightarrow \infty} H(t) = G^*$$

The population converge to the global optimal point.

Theorem 3. With the truncation selection, EDAs can converge to the global optimal point under the additive noisy environment.

Proof.

According to the truncation selection model,

$$p(x, t+1) = \frac{p(x, t)}{C} \times \Phi\left(\frac{f(x) - \beta(t)}{\sigma}\right) \quad (18)$$

$$C = \int_{F(x) \geq \beta(x)} p(x, t) dx \quad (C \text{ is a constant})$$

Since both $p(x, t)$ and $p(x, t+1)$ are pdfs, then

$$\int_D p(x, t) dx = 1$$

$$\int_D p(x, t+1) dx = 1$$

Combining these two equations:

$$\int_D p(x, t+1) - p(x, t) dx = 0$$

Since $p(x, t)$ is a continuous function, then there exists \tilde{x} satisfying the following equation:

$$p(\tilde{x}, t+1) - p(\tilde{x}, t) = 0$$

Viz.

$$\frac{\Phi\left(\frac{f(x) - \beta(t)}{\sigma}\right)}{C} = 1$$

Classify the population into two subsets Q_1 and Q_2 :

$$Q_1 = \{x | f(x) < f(\tilde{x})\}$$

$$Q_2 = \{x | f(x) \geq f(\tilde{x})\}$$

Since $\frac{\Phi\left(\frac{f(x) - \beta(t)}{\sigma}\right)}{C}$ is a monotonous increasing function. Thus,

$$p(x, t+1) \geq p(x, t) \quad x \in Q_1 \quad (19)$$

$$p(x, t+1) \leq p(x, t) \quad x \in Q_2 \quad (20)$$

$$\int_{Q_2} p(x, t+1) - p(x, t) dx = \int_{Q_1} p(x, t) - p(x, t+1) dx \quad (21)$$

(19)(20)(21) tell us that when the population size is static, during the evolution of the population, the individual with a high $f(x)$ will increase; while the one with a low $f(x)$ will decrease.

$$\begin{aligned} H(t+1) - H(t) &= \int_{Q_2} f(x)(p(x, t+1) - p(x, t)) dx \\ &\quad - \int_{Q_1} f(x)(p(x, t) - p(x, t+1)) dx \\ &\geq f(\tilde{x}) \int_{Q_2} p(x, t+1) - p(x, t) dx - f(\tilde{x}) \int_{Q_1} p(x, t) - p(x, t+1) dx \\ &\geq 0 \end{aligned}$$

Thus

$$H(t+1) \geq H(t)$$

$\lim_{t \rightarrow \infty} H(t)$ exists. Suppose

$$\lim_{t \rightarrow \infty} H(t) = G \quad (G < G^*)$$

Define the set Q :

$$Q = \{x | f(x) > G\}$$

For all the individuals in Q

$$p(x, t) = \frac{\Phi\left(\frac{f(x) - \beta(0)}{\sigma}\right)}{C} \frac{\Phi\left(\frac{f(x) - \beta(1)}{\sigma}\right)}{C} \dots \frac{\Phi\left(\frac{f(x) - \beta(t-1)}{\sigma}\right)}{C}$$

Where

$$\frac{\Phi\left(\frac{f(x) - \beta(i)}{\sigma}\right)}{C} > 1$$

Since $f(x) > G$, then,

$$\lim_{t \rightarrow \infty} p(x, t) = +\infty$$

$p(x, t)$ is a continuous probability distribution function,

$$\lim_{t \rightarrow \infty} \int_Q p(x, t) = +\infty$$

This equation contradicts the performance of pdf. So the suppose that $H(t)$ converges to G ($G < G^*$) doesn't hold. Then

$$\lim_{t \rightarrow \infty} H(t) = G^*$$

EDA converges to the global optimal point.

5. Acknowledgement

This project was supported by the National Natural Science Foundation of China (Nos. 60271033). Thanks for the comments and suggestions from the reviewers.

6. References

- [1] J. H. Holland, "Adaptation in Nature and Artificial System", MIT Press, 1990
- [2] H. Muhlenbein, "From recombination of genes to the estimation of distributions I. Binary parameters", LNCS1411: Parallel Problem Solving from Nature-PPSN IV, 1996, pp.178-187
- [3] P. Larrañaga, and J. A. Lozano, "Estimation of Distribution Algorithms", Kluwer Academic Publisher, 2002
- [4] Q. Zhang, and H. Mühlenbein, "On the convergent order of a class of Estimation of Distribution Algorithms", IEEE transaction on evolutionary computation. 2004, pp.127-136
- [5] Y. Hong, Q. Ren, and J. Zeng, "Optimization of Noisy Fitness functions with Univariate Marginal Distribution Algorithm", IEEE CEC2005. Edinburgh, UK. 2005, Accepted