On the Convergence of a Class of Estimation of Distribution Algorithms

Qingfu Zhang, *Member, IEEE,* and Heinz Mühlenbein

***Abstract*—We investigate the global convergence of estimation of distribution algorithms (EDAs). In EDAs, the distribution is estimated from a set of selected elements, i.e., the parent set, and then the estimated distribution model is used to generate new elements. In this paper, we prove that: 1) if the distribution of the new elements matches that of the parent set exactly, the algorithms will converge to the global optimum under three widely used selection schemes and 2) a factorized distribution algorithm converges globally under proportional selection.**

***Index Terms*—Convergence, estimation of distribution algorithms (EDAs), factorized distribution algorithms (FDA).**

I. Introduction

Evolutionary algorithms (EAs) are population-based algorithms for optimization and search problems. EAs maintain and successively improve a collection of potential solutions until some stopping condition is met. Let $\mathrm{Pop}(t)$ be the population at generation t. An EA works in the following recursive way.

Step 1) **Selection:** Choose some elements from $\mathrm{Pop}(t)$ to form the parent set $\mathrm{Pop}^s(t)$, using a selection scheme.

Step 2) **Variation:** Perform variation operations on elements of $\mathrm{Pop}^s(t)$ and generate new elements to form the new population.

The most popular implementations of EAs are genetic algorithms (GAs), evolution strategies (ES), and evolutionary programming (EP). EAs employ crossover and mutation as variation operators to create the elements of the next generation. Numerous successful applications of EAs have been reported in the literature. Many search and optimization problems have also been encountered, where EAs perform badly. The behavior of EAs is often studied theoretically by using the following approaches: schema-based approach [1], Markov chain models [40], and infinite population models [21] among others (e.g., [2] and [3]). The schema (i.e., building blocks) theory for GAs, originally proposed by Holland [1], [36], [37], aims to predict the expected numbers of solutions in a given schema (a subset of the search space) at the next generation, in terms of quantities measured at the current generation. The schema theorem of Holland only gives the lower bounds for the expected quantities and does not say anything precise about schema reconstruction. Recently, the exact schema theorems for GA and genetic programming (GP) have been derived for exactly predicting the expected characteristics of the population at the next generation [38], [39]. However, much effort has to be made in utilizing these theorems to study the convergence of a GA. Markov chain-based approaches characterize an EA as a Markov model with the current population being the state variables, and then study the convergence of the population in the sense of probability [40]–[43]. A general framework has been developed for analyzing the first hitting time of an EAs Markov model [25]–[27]. Under this framework, some open questions on the roles of population and crossover have been successfully solved for several typical EAs on different problems. Infinite population models approximate the behavior of an EA as its population size tends to infinity [21]–[23]. These models are often represented by deterministic dynamical systems and, therefore, the mathematical analysis becomes easier. However, an upper bound of the error between the actual EA and its model is not easily estimated.

One recent principled alternative to traditional EAs is provided by population-based algorithms using estimation of distribution, which are often called estimation of distribution algorithms (EDAs) [6], [8]. Instances of EDAs include population-based incremental learning (PBIL) [7], univariate marginal distribution algorithm (UMDA) [5], mutual information maximization for input clustering (MIMIC) [9], combining optimizers with mutual information trees (COMIT) [10], factorized distribution algorithm (FDA) [11], Bayesian optimization algorithm (BOA) [12], Bayesian evolutionary algorithm (BEA) [13], iterated density estimation evolutionary algorithm (IDEA) [14], and global search based on reinforcement learning agents (GSBRL) [15], to name a few. There is no traditional crossover or mutation in EDAs. Instead, they explicitly extract global statistical information from the parent set $\mathrm{Pop}^s(t)$ and build a posterior probability distribution model of promising solutions, based on the extracted information. New solutions are sampled from the model thus built and fully or in part replace $\mathrm{Pop}(t)$ to form the new population $\mathrm{Pop}(t+1)$. The idea of EDAs can be easily adapted for many optimization problems. Since the dependence relationships in the distribution of the promising solutions are highly relevant to the variable interactions in the problem, EDAs are promising methods for capturing the structure of variable interactions, identifying and manipulating crucial building blocks and, hence, efficiently solving hard optimization and search problems with interactions among the variables.

Manuscript received October 29, 2002; revised August 4, 2003. The work of Q. Zhang was supported in part by the Engineering and Physical Sciences Research Council (EPSRC) under Grant GR/R64742/01 and in part by the GMD Research Fellowship.

Q. Zhang is with the Department of Computer Science, University of Essex, Colchester CO4 3SQ, U.K. (-mail: qzhang@essex.ac.uk).

H. Mühlenbein is with the Fraunhofer Institute for Autonomous Intelligent Systems, St. Augustin D-53757, Germany (e-mail: heinz.muehlenbein@ais.fraunhofer.de).

Digital Object Identifier 10.1109/TEVC.2003.820663

1089-778X/04$20.00 © 2004 IEEE

Relatively little effort has been devoted to studying the working mechanics of EDAs. Mühlenbein [5], González *et al.* [16], and Höhfeld and Rudolph [17] have studied the behaviors of UMDA and PBIL (the simplest versions of the EDA, which ignore all the variable interactions). Their results show that these algorithms can locate the optimum of a linear function but cannot solve problems with nonlinear variable interactions. In [18], Mühlenbein and Mahnig discussed the convergence of the FDA for separable additively decomposable functions (ADFs). Since there are no overlaps in their objective functions, their FDA is equivalent to the UMDA. Therefore, their work does not deal with the ability of FDA to solve problems with variable interactions.

The purpose of this paper is to study the convergence of EDAs. It is extremely difficult to analyze the behavior of EDAs with a finite population. However, the distributions of infinite populations can be studied mathematically and can be regarded as the limit distributions of large populations. For this reason, this paper focuses on the dynamics of the distributions of infinite population in EDAs. In designing practical EDA-like algorithms, researchers often explicitly or implicitly attempt to make the distribution of the population approximate that of their parents as closely as possible. We prove in this paper that EDA converges to the global optimum under three widely used selection schemes if the distribution of the next generation exactly matches that of their parents. This result shows something of the utility of this approach. To avoid exponential explosion, most existing EDAs can only use low-order dependence relationships in modeling the posterior probability of promising solutions. Obviously, the distribution of the next generation does not approximate that of their parents in these algorithms. FDA is an algorithm of this kind. We show that FDA converges to the global optimum under proportional selection for maximizing real-valued ADFs. Our analysis implies that the use of some selected low-order dependence relationships in EDAs can guarantee convergence.

The paper is organized as follows. Modeling of EDAs with infinite populations is introduced in Section II. Section III contains the convergence results for EDAs with $p(x,t+1) = p^s(x,t)$. Convergence results for FDAs with proportional selection are established in Section IV. Section V summarizes the paper.

II. Models of EDAs With Infinite Populations

We consider the following optimization problem:

$$\max f(x) \quad x \in D \tag{1}$$

where $x = (x_1, x_2, \ldots, x_n) \in R^n, D \subset R^n$ is a bounded and closed set with nonempty interior, and $f(x) : D \to R$ is a positive and continuous objective function. Therefore, there exists a point $x^* \in D$ such that $f(x) \le f(x^*)$ for all $x \in D$. x^* is called a globally optimal solution and $G^* = f(x^*)$ is the global maximum.[1] Throughout this paper, we assume that the Borel measure of $H = \{x \mid x \in D, f(x) > C\}$ is positive if $C < G^*$.

A. EDAs With Infinite Population

As denoted in Section I, $\mathrm{Pop}(t)$ and $\mathrm{Pop}^s(t)$, respectively, are the population and the parent set at time t in an EDA for the problem (1). Let the underlying probability distribution functions for the points in $\mathrm{Pop}(t)$ and $\mathrm{Pop}^s(t)$ be $p(x,t)$ and $p^s(x,t)$, respectively. By the famous Glivenko–Canteli theorem [45], the empirical probability density functions induced by points in $\mathrm{Pop}(t)$ and $\mathrm{Pop}^s(t)$ will converge to $p(x,t)$ and $p^s(x,t)$, respectively, as the sizes of $\mathrm{Pop}(t)$ and $\mathrm{Pop}^s(t)$ tend to infinity. Therefore, $p(x,t)$ and $p^s(x,t)$ can be thought of as the population and the parent set at time t in the EDA with infinite populations. Correspondingly, EDAs can be modeled as the following iteration of probability functions.

Step 1) **Selection:** $p(x,t) \to p^s(x,t)$.

Step 2) **Variation:** $p^s(x,t) \to p(x,t+1)$.

We define

$$E(t) = \int_D f(x) p(x,t)\, dx. \tag{2}$$

$E(t)$ is the average objective function value of the population at time t. We say that the EDA converges globally if

$$\lim_{t \to \infty} E(t) = G^*. \tag{3}$$

Intuitively, almost all the individual points in the population will move to globally optimal points as the time tends to infinity if an EDA converges globally.

B. Proportional Selection Model

Proportional selection is the most basic selection scheme in EAs. In the case of a finite population, the probability of a point being selected as a parent is proportional to its fitness (i.e., objective function value). Therefore, in the case of an infinite population, this selection scheme should be modeled as

$$p^s(x,t) = \frac{f(x)p(x,t)}{E(t)}. \tag{4}$$

C. Truncation Selection Model

Truncation selection ranks all the points in the current population according to their objective function values and selects the best ones as parents. In truncation selection with threshold $\alpha(t)$ only the $100\alpha(t)\%$ best points are selected to become the parents for the next generation. This selection has been used in the breeder GA and some implementations of EDAs [5]. It can be modeled as

$$p^s(x,t) = \begin{cases} \frac{p(x,t)}{\alpha(t)}, & \text{if } f(x) \ge \beta(t) \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $\beta(t)$ is a real number such that

$$\alpha(t) = \int_{f(x) \ge \beta(t)} p(x,t)\, dx. \tag{6}$$

Therefore, a point is selected if and only if its objective function is not less than $\beta(t)$.

[1] $f(x)$ may have many distinct globally optimal solutions.

D. Tournament Selection Model

In K-tournament selection [35], K points are randomly chosen from the current population and the best point is selected to be a parent. As pointed out in [35], $K = 2$ is a typical value used in many applications. Therefore, we consider the case $K = 2$ and assume that $p(x,t)$ is continuous. In this case, the tournament selection can be modeled as [5]

$$p^s(x,t) = 2p(x,t) \int\limits_{f(y)\le f(x)} p(y,t)\,dy. \tag{7}$$

Blickle and Thiele have studied the expected fitness distribution after repeating tournament selection several times [19], [20]. The model (7) can also be derived from their work.[2]

The mathematical models of natural and artificial selection have been extensively studied in the area of population genetics [46]–[48]. The effect of biological selection on the quantitative characteristics of the offspring can be inferred from the observed regression of offsring on parent. Mühlenbein has applied this technique to study the behaviors of GAs [5].

III. EDAs With $p(x,t+1) = p^s(x,t)$

EDAs build a probability model based on the extracted statistical information from $\mathrm{Pop}^s(t)$ and sample from the model, thus, built to generate new points for the next generation. One principle in existing EDAs is to make the probability model approximate the actual probability distribution of points in $\mathrm{Pop}^s(t)$ as closely as possible within a reasonable computational time. In the case of infinite populations, this principle is equivalent to approximating $p(x,t+1)$ to $p^s(x,t)$. Therefore, it is worthwhile studying the convergence of EDAs with $p(x,t+1) = p^s(x,t)$.

Theorem 1: If $p(x,0)$ is positive and continuous in D and $p(x,t+1) = p^s(x,t)$, then

a) the EDA with proportional selection converges;
b) the EDA with truncation selection converges if $\alpha(t) \le \theta$ for all $t > 0$, where $\theta < 1$ is a positive constant;
c) the EDA with 2-tournament selection converges.

Proof: **Case A:** Proportional Selection.
By (4), the algorithm can be described as

$$p(x,t+1) = \frac{p(x,t)f(x)}{E(t)}. \tag{8}$$

From the fact that $f(x), p(x,0) > 0$ for all $x \in D$, and $f(x)$ and $p(x)$ are continuous, it follows that $p(x,t)$ is a positive continuous probability function, and $E(t) < G^*$ for all $t \ge 0$. By (2) and (8), we have

$$E(t+1) = \frac{\int\limits_D [f(x)]^2 p(x,t)\,dx}{E(t)}. \tag{9}$$

Then

$$E(t+1) - E(t) = \frac{\int\limits_D [f(x) - E(t)]^2 p(x,t)\,dx}{E(t)} \tag{10}$$

equation (10) and (55) are equivalent to Fisher's fundamental theorem of natural selection [46], which has been discussed in the context of GAs in [5], which implies that

$$E(t+1) \ge E(t) \tag{11}$$

for all $t > 0$. Therefore, $\lim_{t\to\infty} E(t)$ exists. Let

$$G \overset{\triangle}{=} \lim_{t\to\infty} E(t). \tag{12}$$

We next prove by contradiction that $G = G^*$. Assume that

$$G < G^*. \tag{13}$$

By (8)

$$p(x,t) = p(x,0)\frac{f(x)}{E(t-1)}\frac{f(x)}{E(t-2)}\cdots\frac{f(x)}{E(0)}. \tag{14}$$

Since

$$\lim_{t\to\infty}\frac{f(x)}{E(t)} = \frac{f(x)}{G} > 1 \tag{15}$$

whenever $f(x) > G$. Noting that $p(x) > 0$ for all $x \in D$, we have

$$\lim_{t\to\infty} p(x,t) = +\infty \tag{16}$$

for all x with $f(x) > G$.

Let $S = \{x \,|\, x \in D, f(x) > G\}$. By the assumption stated earlier in Section II, the Borel measure of S is nonzero. Then, by Fatou's lemma [44]

$$\lim_{t\to\infty}\int\limits_S p(x,t)\,dx = +\infty \tag{17}$$

which contradicts the fact that $p(x,t)$ is a probability density function. This completes the proof of the theorem for Case A.

Case B: Truncation Selection.
In this case, the algorithm is as follows:

$$p(x,t+1) = \begin{cases} \frac{p(x,t)}{\alpha(t)}, & \text{if f (x)} \ge \beta(t) \\ 0, & \text{otherwise} \end{cases}. \tag{18}$$

Thus, we have $p(x,t+1) = 0$ whenever $f(x) < \beta(t)$. It follows that:

$$\int\limits_{f(x)\ge\beta(t)} p(x,t+1)\,dx = \int\limits_D p(x,t+1)\,dx = 1. \tag{19}$$

By (6)

$$\int\limits_{f(x)\ge\beta(t+1)} p(x,t+1)\,dx = \alpha(t+1) < 1. \tag{20}$$

Comparing (19) and (20) gives

$$\beta(t) < \beta(t+1), \quad t = 1,2,\ldots. \tag{21}$$

It follows that $\lim_{t\to\infty}\beta(t)$ exists. Denote

$$\beta = \lim_{t\to\infty}\beta(t). \tag{22}$$

We next prove by contradiction that $\beta = G^*$. Assume that

$$\beta < G^*. \tag{23}$$

[2] In fact, taking $t = 2$ and $N = 1$ in (8) and [19] (t is the tournament size and N is the number of times the selection repeats), we obtain a model which is mathematically equivalent to our 2-tournament model.

Therefore

$$p(x,t) = p(x,0)\prod_{i=0}^{t-1}[\alpha(i)]^{-1} \geq \theta^{-t}p(x,0) \qquad (24)$$

whenever $f(x) > \beta$. Noting that $p(x,0) > 0$ for any $x \in D$, we have

$$\lim_{t\to\infty} p(x,t) = +\infty \qquad (25)$$

for all x with $f(x) > \beta$. Let $S = \{x \,|\, x \in D, f(x) > \beta\}$. Since the Borel measure of S is positive, by Fatou's lemma we obtain

$$\lim_{t\to\infty} \int\limits_{S} p(x,t)\,dx = +\infty \qquad (26)$$

which contradicts the fact that $p(x,t)$ is a probability density function. Therefore, we obtain

$$\lim_{t\to\infty} \beta(t) = G^{*}. \qquad (27)$$

Note that

$$E(t) \geq \beta(t). \qquad (28)$$

Then, we have

$$\lim_{t\to\infty} E(t) = G^{*}. \qquad (29)$$

This completes the proof of the theorem for Case B.

Case C: Tournament Selection.

In this case, the algorithm can be modeled as

$$p(x,t+1) = 2p(x,t)\int\limits_{f(y)\leq f(x)} p(y,t)\,dy. \qquad (30)$$

For any given positive ϵ, let $N_\epsilon = \{x \,|\, x \in D, f(x) \geq G^{*} - \epsilon\}$, and denote

$$r(\epsilon,t) = \int\limits_{N_\epsilon} p(x,t)\,dx \qquad (31)$$

and

$$q(\epsilon,t) = \int\limits_{D\backslash N_\epsilon} p(x,t)\,dx. \qquad (32)$$

By (30), we obtain

$$r(\epsilon,t+1) = \int\limits_{x\in N_\epsilon}\left[2p(x,t)\int\limits_{f(y)\leq f(x)} p(y,t)\,dy\right]dx. \qquad (33)$$

Obviously

$$\int\limits_{f(y)\leq f(x)} p(y,t)\,dy = q(\epsilon,t) + \int\limits_{G^{*}-\epsilon\leq f(y)\leq f(x)} p(y,t)\,dt \qquad (34)$$

for $x \in N_\epsilon$. Inserting (34) into (33) and rearranging them leads to

$$r(\epsilon,t+1) = 2q(\epsilon,t)r(\epsilon,t) + I \qquad (35)$$

where

$$I = 2\int\limits_{x\in N_\epsilon}\int\limits_{G^{*}-\epsilon\leq f(y)\leq f(x)} p(x,t)p(y,t)\,dx\,dy. \qquad (36)$$

By the symmetry of $p(x,t)p(y,t)$

$$\begin{aligned}\int\limits_{x\in N_\epsilon}\int\limits_{G^{*}-\epsilon\leq f(y)\leq f(x)} &p(x,t)p(y,t)\,dx\,dy\\ &= \int\limits_{y\in N_\epsilon}\int\limits_{G^{*}-\epsilon\leq f(x)\leq f(y)} p(x,t)p(y,t)\,dx\,dy. \qquad (37)\end{aligned}$$

Then

$$\begin{aligned} I &= \int\limits_{x\in N_\epsilon}\int\limits_{G^{*}-\epsilon\leq f(y)\leq f(x)} p(y,t)p(x,t)\,dx\,dy\\ &\quad + \int\limits_{y\in N_\epsilon}\int\limits_{G^{*}-\epsilon\leq f(x)\leq f(y)} p(x,t)p(y,t)\,dx\,dy\\ &= \int\limits_{x\in N_\epsilon}\int\limits_{y\in N_\epsilon} p(x,t)p(y,t)\,dx\,dy\\ &= \left[\int\limits_{x\in N_\epsilon} p(x,t)\,dx\right]^{2} = [r(\epsilon,t)]^{2}. \qquad (38)\end{aligned}$$

Therefore

$$r(\epsilon,t+1) = 2q(\epsilon,t)r(\epsilon,t) + [r(\epsilon,t)]^{2}. \qquad (39)$$

Noting that $q(\epsilon,t) + r(\epsilon,t) = 1$, we have

$$r(\epsilon,t+1) = 1 - [q(\epsilon,t)]^{2}. \qquad (40)$$

Thus

$$q(\epsilon,t+1) = [q(\epsilon,t)]^{2} \qquad (41)$$

which implies

$$q(\epsilon,t+1) = [q(\epsilon,0)]^{2(t+1)}. \qquad (42)$$

Since $q(\epsilon,0) < 1$, we obtain

$$\lim_{t\to\infty} q(\epsilon,t) = 0. \qquad (43)$$

Then

$$\lim_{t\to\infty} r(\epsilon,t) = 1 \qquad (44)$$

which implies

$$\lim_{t\to\infty} E(t) = G^{*}. \qquad (45)$$

This completes the proof of the theorem for Case C. ■

Since selection schemes have a crucial impact on the performance of EAs and it is very difficult to analyze the joint effect of selection schemes and recombination operators, much effort has been devoted to analyzing the behaviors of EAs using selection schemes only. Very recently, He and Yao have made one of the first attempts toward analyzing EAs with recombination and with a population size greater than one [26]. The EAs with selection only can be regarded as implementations of EDAs with $p(x,t+1)=p^s(x,t)$. Qi and Palmieri have studied the effect of the proportional selection scheme and obtained a very similar result [21].[3] Goldberg and Deb have introduced the concept of takeover time as a measure of selective pressure for algorithms for a finite search space [24]. Several selection schemes have been studied and compared with respect to their takeover time [28]. He and Yao have used the first hitting time in the study of the time complexity of EAs with finite population for combinatorial optimization problems [25]–[27]. Selection intensity and fitness distribution have also been analyzed extensively in finite space [20], [29], [30]. We plan, in the future, to extend these concepts to EDAs for continuous optimization problems.

IV. Convergence of Factorized Distribution Algorithm

Theorem 1 shows that approximation of $p(x,t+1)$ to $p^s(x,t)$ will drive the population to the global optimum. However, it is often very difficult to do so numerically in practical algorithms, particularly for large-scale problems. $p(x,t+1)$ has to be built within a reasonable computational time. This task will become tractable if $p(x,t+1)$ is expressed by a graphical model [31]. For this reason, most current EDAs use graphical models to represent $p(x,t+1)$ (e.g., [9]–[11] and [14]). These algorithms select some dependence relationships [i.e., multivariable marginal probabilities of $p^s(x,t)$] to construct $p(x,t+1)$. Obviously, there exists a gap between $p(x,t+1)$ and $p^s(x,t)$. To our best knowledge, no work on the convergence of these algorithms has been carried out so far. In this section, we will study the convergence of FDA [11]. FDA chooses a graphical model for building $p(x,t+1)$ based on the prior knowledge of the structure of $f(x)$. We first introduce the following definition.

Definition 2: Let $x=(x_1,x_2,\ldots,x_n)\in D, d_1,\ldots,d_m$ be nonempty subsets of $I=\{1,2,\ldots,n\}$, and x_{d_i} be the subvector of x containing x_j for $j\in d_i$. Then

$$f(x)=\sum_{i=1}^{m} f_i\left(x_{d_i}\right) \tag{46}$$

is called a canonical form of $f(x)$, if :

a) $\cup_{j=1}^m d_j = I$;

b) for each $1\le i<m, d_i$ and $\cup_{j=i+1}^m d_j$ cannot properly contain each other;

c) for each $i=1,2,\ldots,m-1$ and $i<l\le m$, let $a_i=d_i\cap(\cup_{j=i+1}^m d_j)$, then $a_i\cap d_l$ is empty or $a_i\subset d_l$.

The above requirements for d_i are the same as that in the definition of the triangulation structure in the graphical model for multivariate statistics [31].

A function can have several different canonical forms. For example, if

$$f(x)=h_1(x_1,x_2)+h_2(x_2,x_3)+h_3(x_3,x_4)+\ h_4(x_4,x_1)$$

then the above form is not canonical. But $f(x)$ can be written as the following canonical form:

$$f(x)=g_1(x_1,x_2,x_4)+g_2(x_2,x_3,x_4)$$

where $g_1(x_1,x_2,x_4)=h_1(x_1,x_2)+h_4(x_4,x_1)$ and $g_2(x_2,x_3,x_4)=h_2(x_2,x_3)+h_3(x_3,x_4)$. $f(x)$ can also be expressed as the following canonical form:

$$(x)=l_1(x_1,x_2,x_3)+l_2(x_1,x_3,x_4)$$

where $l_1(x_1,x_2,x_3)=h_1(x_1,x_2)+h_2(x_2,x_3)$ and $l_2(x_1,x_3,x_4)=h_3(x_3,x_4)+h_4(x_4,x_1)$.

In the following, we always assume that $f(x)$ is in the form (46) and that D is a closed hypercube of R^n. FDAs for continuous optimization problems have first been investigated in [32]. A FDA with finite populations for the problem (1) can be described as follows.

Step 1) **Selection:** $\text{Pop}(t)\to Pop^s(\mathbf{t})$.

Step 2) **Variation:** $\text{Pop}^s(t)\to Pop(t+1)$.

Step 2.1) Estimate the marginal probabilities $p^s(x_{d_i},t)$, $p^s(x_{a_j},t)$ of the points in $\text{Pop}^s(t)$. Denote the estimated probabilities as $\tilde{p}^s(x_{d_i},t),\tilde{p}^s(x_{a_j},t)$.[4]

Step 2.2) Sample points from the probability

$$\frac{\prod_{i=1}^m \tilde{p}^s\left(x_{d_i},t\right)}{\prod_{j=1}^{m-1}\tilde{p}^s\left(x_{a_j},t\right)} \tag{47}$$

to form $\text{Pop}(t+1)$.

In Step 2.1), only $p^s(x_{d_i},t)$ and $p^s(x_{a_j},t)$ need to be estimated. If all the d_i are very short, the computational overhead involved in Step 2.1) should be acceptable. On the other hand, a good estimation of marginal probability $p^s(x_{d_i},t)$ for short d_i could be obtained from $\text{Pop}^s(t)$ if it has a reasonable size (see [45, Ch. 12]). Therefore, Step 2.1) should be able to be implemented without much difficulty. Sampling in Step 2.2) can be easily made, based on the graphical model determined by d_i [11].

Obviously, the FDA with infinite populations should be modeled as follows.

Step 1) $p(x,t)\to p^s(x,t)$.

Step 2)

$$p(x,t+1)=\frac{\prod_{i=1}^m p^s\left(x_{d_i},t\right)}{\prod_{j=1}^{m-1}p^s\left(x_{a_j},t\right)}. \tag{48}$$

In the following, we only study the FDA with infinite populations. Using the properties of the canonical form [31], we can prove.

Lemma 3: For the above FDA, we have the following:

a) $p(x,t)\ge 0$ for all $x\in D$ and $\int_D p(x,t)\,dx=1$;

b) $p(x_{d_i},t+1)=p^s(x_{d_i},t), i=1,\ldots,m$;

c) $p(x_{\cup_{j=2}^m d_j}\,|\,x_{d_1},t)=p(x_{\cup_{j=2}^m d_j}\,|\,x_{a_1},t)$.

where $p(x_a\,|\,x_b,t)$ stands for the conditional probability, i.e.,

$$p(x_a\,|\,x_b,t)=\frac{p(x_{a\cup b},t)}{p(x_b,t)}. \tag{49}$$

In Lemma 3, a) says that $p(x,t)$ is a probability density function of x defined on D, i.e., the FDA is well-defined, b) indicates

[3] The theorem of Qi and Palmieri on proportional selection (in [21, Th. 2]) assumes that the objective function has a unique global maximum and that the global maximum has a connected neighborhood.

[4] In this paper, if $a\subset I$ is empty, the marginal probability of x_a is defined to be one.

that for each subvector x_{d_i}, FDA samples points for the next generation exactly according to $p^s(x_{d_i}, t)$, and (c) and (d) will be useful in the proof of the global convergence of the FDA.

A. Global Convergence for Proportional Selection

In this section, we consider the global convergence of the FDA for the problem (1) when the following generalized proportional selection (GPS) is employed in Step 1) of the FDA:

$$p^s(x,t) = \frac{p(x,t)[F(x,t)+w(t)]}{\int\limits_D p(x,t)F(x,t)\,dx + w(t)} \tag{50}$$

where $F(x,t)$ and $w(t)$ satisfy the following conditions:

A1) $F(x,t) = \sum_{i=1}^m F_i(x_{d_i}, t)$ for all $t \geq 0$;

A2) $\lim_{t\to\infty} F_i(x_{d_i}, t) = f_i(x_{d_i})$ for $i = 1, 2 \ldots$, and for all $t \geq 0$;

A3) $0 < F_i(x_{d_i}, t) \leq F_i(x_{d_i}, t+1)$ for each $1 \leq i \leq m$ and for all $t \geq 0$;

A4) $0 \leq w(t) < M$ for all $t \geq 0$, where M is a constant.

Let $F(x,t) = f(x)$ and $w(t) = 0$ for all $t \geq 0$. Then, (50) will become (4). Therefore, proportional selection is a special case of the GPS. The reason that we use GPS (50) instead of proportional selection (4) is to make the induction in the proof of Theorem 5 much easier.

The following lemma can be regarded as an extension of Theorem 1.

Lemma 4: In the case $m = 1$, if $p(x,0)$ is positive and continuous in D, then for the FDA defined by (50) and (48), we have

$$\lim_{t\to\infty} \int\limits_D p(x,t)F(x,t)\,dx = G^*. \tag{51}$$

Proof: In this case, the algorithm is as follows:

$$p(x+1,t) = \frac{p(x,t)[F(x,t)+w(t)]}{\int\limits_D p(x,t)F(x,t)\,dx + w(t)}. \tag{52}$$

Noting A3) and A4), it is easy to show that $p(x,t)$ is a positive continuous probability function for all $t \geq 0$. Denote

$$e(t) \overset{\triangle}{=} \int\limits_D p(x,t)F(x,t)\,dx. \tag{53}$$

By A2) and A3), we know that $0 < F(x,t) \leq f(x)$ for all $t > 0$. Therefore

$$e(t) \leq G^* \tag{54}$$

for all $t > 0$.

From (52), we can derive

$$\int\limits_D p(x,t+1)F(x,t)\,dx - e(t) = \frac{\int\limits_D [F(x,t)-e(t)]^2 p(x,t)\,dx}{e(t)+w(t)} > 0. \tag{55}$$

It follows from (A3) that

$$\begin{aligned} e(t+1) &= \int\limits_D p(x,t+1)F(x,t+1)\,dx \\ &\geq \int\limits_D p(x,t+1)F(x,t)\,dx. \end{aligned} \tag{56}$$

Thus

$$e(t+1) \geq e(t) \tag{57}$$

for all $t > 0$. Therefore, $\lim_{t\to\infty} e(t)$ exists. Let

$$G \overset{\triangle}{=} \lim_{t\to\infty} e(t). \tag{58}$$

We next prove by contradiction that $G = G^*$. Assume that

$$G < G^*. \tag{59}$$

It is obvious that

$$p(x,t+1) = \frac{[F(x,t)+w(t)]p(x,t)}{e(t)+w(t)} \tag{60}$$

which establishes

$$\begin{aligned} p(x,t) = p(x,0)&\frac{F(x,t-1)+w(t-1)}{e(t-1)+w(t-1)} \\ &\times \frac{f(x,t-2)+w(t-2)}{e(t-2)+w(t-2)} \cdots \\ &\frac{F(x,0)+w(0)}{e(0)+w(0)}. \end{aligned} \tag{61}$$

Thus

$$\lim_{t\to\infty} \frac{F(x,t)+w(t)}{e(t)+w(t)} \geq \frac{f(x)+M}{G+M} > 1 \tag{62}$$

for all x with $f(x) > G$. Therefore

$$\lim_{t\to\infty} p(x,t) = +\infty \tag{63}$$

for all x with $f(x) > G$.

Let $S = \{x \mid x \in D, f(x) > G\}$. By the assumption stated earlier in Section II, the Borel measure of S is nonzero, by Fatou's lemma, we have

$$\lim_{t\to\infty} \int\limits_S p(x,t)\,dx = +\infty \tag{64}$$

which contradicts the fact that $p(x,t)$ is a probability density function. This completes the proof of the Lemma. ■

Now, we are in a position to state and prove the main result in this section.

Theorem 5: For the FDA with GPS, let $p(x,0)$ be a positive continuous probability density function on D. Then

$$\lim_{t\to\infty} \int\limits_D F(x,t)p(x,t)\,dx = G^*. \tag{65}$$

Proof: The proof is by induction on m. Lemma 4 shows that the theorem is true in the case $m = 1$. Now, we consider the case $m > 1$. We begin with some notation

$$\begin{aligned}
c_1 &\triangleq \bigcup_{i=2}^{m} d_i \\
G\left(x_{c_1}, t\right) &\triangleq \sum_{j=2}^{m} F_j\left(x_{d_j}, t\right) \\
e(t) &\triangleq \int F(x,t)p(x,t)\, dx \\
e_1(t) &\triangleq \int F_1\left(x_{d_1}, t\right) p\left(x_{d_1}, t\right) dx_{d_1} \\
e_2(t) &\triangleq \int G\left(x_{c_1}, t\right) p\left(x_{c_1}, t\right) dx_{c_1} \\
H_1\left(x_{a_1}, t\right) &\triangleq \int F_1\left(x_{d_1}, t\right) p\left(x_{d_1} \mid x_{a_1}, t\right) dx_{d_1 \setminus a_1} \\
H_2\left(x_{a_1}, t\right) &\triangleq \int G\left(x_{c_1}, t\right) p\left(x_{c_1} \mid x_{a_1}, t\right) dx_{c_1 \setminus a_1}.
\end{aligned}$$

In the above integrals, the regions of integration are the projection areas of D corresponding to the variables of integration.

Integrating both sides of (50) over $x_{c_1 \setminus a_1}$ gives

$$p^s\left(x_{d_1}, t\right) = \frac{F_1\left(x_{d_1}, t\right) p\left(x_{d_1}, t\right) + \int\limits_D G\left(x_{c_1}, t\right) p(x,t)\, dx_{c_1 \setminus a_1}}{e(t) + w(t)}. \tag{66}$$

By c) in Lemma 3

$$\begin{aligned}
&\int\limits_D G\left(x_{c_1}, t\right) p(x,t)\, dx_{c_1 \setminus a_1} \\
&\quad = \int\limits_D G\left(x_{c_1}, t\right) p\left(x_{c_1} \mid x_{d_1}, t\right) p\left(x_{d_1}, t\right) dx_{c_1 \setminus a_1} \\
&\quad = \int\limits_D G\left(x_{c_1}, t\right) p\left(x_{c_1} \mid x_{a_1}, t\right) p\left(x_{d_1}, t\right) dx_{c_1 \setminus a_1}.
\end{aligned}$$

Then

$$\int\limits_D G\left(x_{c_1}, t\right) p(x,t)\, dx_{c_1 \setminus a_1} = p\left(x_{d_1}, t\right) H_2(x_{a_1}, t). \tag{67}$$

Substituting b) in Lemma 3 and (67) into (66) yields

$$p\left(x_{d_1}, t+1\right) = \frac{p\left(x_{d_1}, t\right)\left[F_1\left(x_{d_1}, t\right) + H_2\left(x_{a_1}, t\right)\right]}{e(t) + w(t)}. \tag{68}$$

Integrating both sides of (68) over $x_{d_1 \setminus a_1}$ gives

$$p\left(x_{a_1}, t+1\right) = \frac{p\left(x_{a_1}, t\right)\left[H_1\left(x_{a_1}, t\right) + H_2\left(x_{a_1}, t\right)\right]}{e(t) + w(t)}. \tag{69}$$

By (68) and (69)

$$\begin{aligned}
&p\left(x_{d_1} \mid x_{a_1}, t+1\right) \\
&\quad = \frac{p\left(x_{d_1} \mid x_{a_1}, t\right)\left[F_1\left(x_{d_1}, t\right) + H_2\left(x_{a_1}, t\right)\right]}{H_1\left(x_{a_1}, t\right) + H_2\left(x_{a_1}, t\right)}.
\end{aligned} \tag{70}$$

For any fixed x_{a_1}, let $p(x_{d_1}, t)$ and $w(t)$ in Lemma 4 be $p(x_{d_1} \,|\, x_{a_1}, t)$ and $H_2(x_{a_1}, t)$, respectively, we obtain

$$\lim_{t \to \infty} H_1\left(x_{a_1}, t\right) = K\left(x_{a_1}\right) \tag{71}$$

where $K(x_{a_1}) = \max_{x_{d_1 \setminus a_1}} f_1(x_{d_1})$. From the proof of Lemma 4, we also know that

$$H_1\left(x_{a_1}, t+1\right) \geq H_1\left(x_{a_1}, t\right) \tag{72}$$

for all $t > 0$. Similarly, we can obtain

$$p^s\left(x_{c_1}, t\right) = \frac{p\left(x_{c_1}, t\right)\left[G\left(x_{c_1}, t\right) + H_1\left(x_{a_1}, t\right)\right]}{E(t) + w(t)}. \tag{73}$$

From b) in Definition 2, without loss of generality, we can assume that $a_1 \subset d_2$. Let

$$\begin{aligned}
L\left(x_{c_1}, t\right) &:= G\left(x_{c_1}, t\right) + H_1\left(x_{a_1}, t\right) \\
&= \left[H_1\left(x_{a_1}, t\right) + F_2\left(x_{d_2}, t\right)\right] + \sum_{j=3}^{m} F_j\left(x_{d_j}, t\right).
\end{aligned} \tag{74}$$

We can treat $[H_1(x_{a_1}, t) + f_2(x_{d_2}, t)]$ as one subfunction. Then, $L(x_{c_1}, t)$ has $m-1$ subfunctions. By the induction hypothesis, we have

$$\begin{aligned}
&\lim_{t \to \infty} \int L\left(x_{c_1}, t\right) p\left(x_{c_1}, t\right) dx_{c_1} \\
&\qquad = \max_{x_{c_1}} \left[K\left(x_{a_1}\right) + \sum_{j=2}^{m} f_j\left(x_{d_j}\right)\right].
\end{aligned} \tag{75}$$

Note that

$$E(t) = \int L\left(x_{c_1}, t\right) p\left(x_{c_1}, t\right) dx_{c_1} \tag{76}$$

and

$$G^* = \max_{x_{c_1}} \left[K\left(x_{a_1}\right) + \sum_{j=2}^{m} F_j\left(x_{d_j}\right)\right]. \tag{77}$$

Therefore, we have

$$\lim_{t \to \infty} e(t) = G^*. \tag{78}$$

This completes the proof of this theorem. ■

Note that proportional selection is a special case of GPS. We now have Theorem 6.

Theorem 6: For the FDA with proportional selection, if $p(x, 0)$ is positive and continuous, then

$$\lim_{t \to \infty} E(t) = G^*.$$

B. Discussions

1) Infinite Population Versus Finite Population: The results in this section are for the infinite population model of FDAs. However, all practical FDAs work with a finite population. Therefore, it is important to study the approximation error of the infinite population model. We have been unable to obtain an upper bound of approximation error for general FDA. In the following, we consider the behavior of UMDA (which is the simplest version of FDA) with finite and infinite population in the case when the objective function

$f(x) = (1/n)\sum_{i=1}^{n} x_i, D = [0,1]^n, p(x,0) = 1$ and the selection scheme is proportional. The UMDA with infinite population can be described as the following.

Step 1) **Proportional Selection**

$$p^s(x,t) = \frac{f(x)p(x,t)}{\int\limits_D f(x)p(x,t)\,dx}. \tag{79}$$

Step 2) **Variation**

$$p(x,t+1) = \prod_{i=1}^{n} p_i^s(x_i,t+1). \tag{80}$$

We can prove that $p(x,t)$ in the above algorithm has the form

$$p(x,t) = \prod_{i=1}^{n} g_t(x_i) \tag{81}$$

where $g_t(*)$ is a polynomial of order t

$$g_t(s) = \sum_{i=0}^{t} a_{t,i} s^t \tag{82}$$

$a_{0,0} = 1$ and $a_{t,i}$ $(t \geq 1, i = 0,1,2,\ldots,t)$ can be computed recursively

$$\begin{aligned} a_{t+1,0} &= \frac{n-1}{n} a_{t,0} \\ a_{t+1,k} &= \frac{n-1}{n} a_{t,k} + \left(n \sum_{i=0}^{t} \frac{a_{t,i}}{i+2}\right)^{-1} a_{t,k-1},\ 1 \leq k \leq t \\ a_{t+1,t+1} &= \left(n \sum_{i=0}^{t} \frac{a_{t,i}}{i+2}\right)^{-1} a_{t,t}. \end{aligned}$$

Thus, $E(t)$ (as defined in Section II-A) becomes

$$E(t) = \sum_{i=0}^{t} \frac{a_{t,i}}{i+2}.$$

There are several ways to implement UMDA with a finite population, We consider UMDA using fixed-width histogram marginal models [34], which works as follows.

Step 1) Set $t := 0$ and $\tilde{p}(x_i,0) = 1$ for all $1 \leq i \leq n$.

Step 2) Generate N points in D from $\tilde{p}(x,t) = \prod_{i=1}^{n} \tilde{p}_i(x_i,t)$ to form the population $\mathrm{Pop}(t)$.

Step 3) Use proportional selection to select M points from $\mathrm{Pop}(t)$ to form $\mathrm{Pop}^s(t)$.

Step 4) Compute $\tilde{p}_i(x_i,t+1)$, the fixed-width histogram marginal distribution on each x_i in $\mathrm{Pop}^s(t)$.

Step 5) Set $t := t+1$ and go to Step 2).

In the marginal fixed-width histogram distribution model, the search space for each variable x_i (which is $[0,1]$ in our case) is divided into H bins. The probability density is constant in each bin. In our experiments for the above two algorithms, we consider the case $n = 3$ and we set $N = M = H = 100$ for the algorithm with finite population. Fig. 1 shows the evolution of the average objective function value for UMDA with infinite population, the mean and standard deviation of the average objective function value for ten independent runs of UMDA with finite population. We can see that the long term behaviors of two algorithms are quite similar. The discrepancy comes from sample fluctuations.

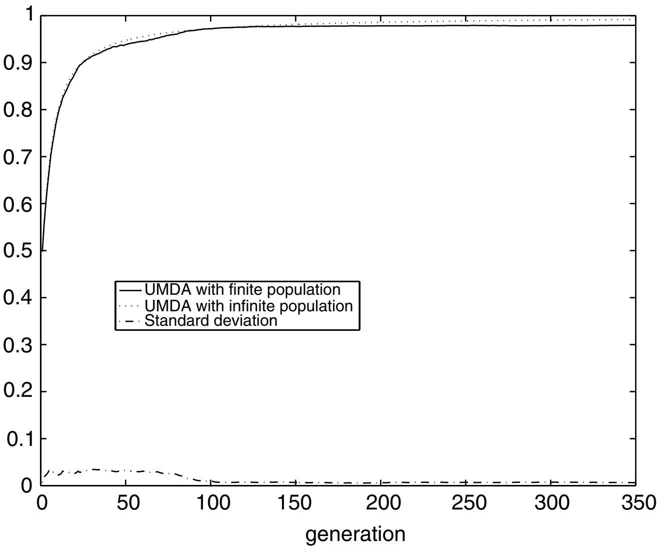

Fig. 1. Evolution of the average objective function value for UMDA with infinite population, the mean and standard deviation of the average objective function value for UMDA with finite population.

2) FDA and Building Blocks: If a point in the search space is expressed by a vector in an EA, a building block is originally defined to be a vector in which the values of several variables are fixed, while the other variables are free. The main concern about building blocks in EAs is their distributions. Mathematically, the distribution of a building block is a marginal probability. Therefore, we can identify a building block as a subvector. To optimize the objective function (46), the FDA only considers the building blocks x_{d_i} and x_{a_j} and generates the new population based on only $p^s(x_{d_i},t)$ and $p^s(x_{a_j},t)$, the marginal probabilties of x_{d_i} and x_{a_j} in the parent set. The length of each subvector should be much smaller than that of x in many applications. Therefore, it will be relatively easy to estimate these marginal probabilities. Our results imply that the utilization of some appropriately selected short building blocks is sufficient to guarantee the convergence of a population-based algorithm for the optimization of an ADF function, and other building blocks can be neglected.

V. Conclusion

EDA is a population-based optimization algorithm using estimation of distributions instead of the manipulation of strings, as is common in evolutionary algorithms. The main advantage of EDAs is that they can handle the interrelations among the variables explicitly. In this paper, the global convergence of EDAs has been shown when the distribution of the next generation is the same as that of the parent set. We have also proved the global convergence of the FDA in some cases. Our results imply that it is appropriate to approximate $p(x,t+1)$ to $p^s(x,t)$ in designing practical population-based algorithms, and it is sufficient to consider some selected crucial dependence relationships for the optimization problem of an ADF function in terms of convergence. The convergence proof is only valid for an infinite population. Several difficult questions remain: Does FDA converge globally under truncation or tournament selection? How many points does one need to estimate to obtain a reliable distribution? For discrete problems the latter question has been investigated in [33]. The answer to this question turns out to be not very informative: the more difficult the optimization problem is, the more difficult the distribution is to estimate.

Acknowledgment

The authors would like to thank X. Yao, D. Fogel, the anonymous reviewers, and the Associate Editor for their constructive comments and suggestions. They also thank A. Mackenzie, E. Tsang, J. Ford, R. Poli, H. Li, and J. Sun for helpful discussions on this topic.

References

[1] J. H. Holland, "Building blocks, cohort genetic algorithms and hyperplane-defined functions," *Evol. Comput.*, vol. 8, pp. 373–391, 2000.

[2] A. Prugel-Bennett and J. L. Shapiro, "An analysis of genetic algorithms using statistical mechanics," *Phys. Rev. Lett.*, vol. 72, pp. 1305–1309, 1994.

[3] K.-S. Leung, Q.-S. Duan, Z.-B. Xu, and C. K. Wong, "A new model of simulated evolutionary computation: Convergence analysis and specifications," *IEEE Trans. Evol. Comput.*, vol. 5, pp. 3–16, Feb. 2001.

[4] D. Whitely, "Test driving three 1995 genetic algorithms: New test functions and geometric matching," *J. Heuristics*, vol. 1, pp. 77–104, 1995.

[5] H. Mühlenbein, "The equation for response to selection and its use for prediction," *Evol. Comput.*, vol. 5, pp. 303–346, 1998.

[6] P. Larranaga and J. A. Lozano, *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation*. Norwell, MA: Kluwer, 2001.

[7] S. Baluja, "Population-Based Incremental Learning: A Method for Integrating Genetic Search Based function Optimization and Competitive Learning," Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep., 1994.

[8] H. Mühlenbein and G. Paass, "From recombination of genes to the estimation of distribution part 1, binary parameter," in *Lecture Notes in Computer Science*, Berlin, Germany: Springer-Verlag, 1996, vol. 1141, Parallel Problem Solving from Nature-PPSN IV, pp. 178–187.

[9] J. S. I. De Bonet and P. Viola, "MIMIC: Finding optima by estimating probability densities," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 1997, vol. 9, pp. 424–431.

[10] S. Baluja and S. Davies, "Fast probabilistic modeling for combinatorial optimization," in *Proc. 15th Nat. Conf. Artificial Intelligence*, 1998, pp. 469–476.

[11] H. Mühlenbein, T. Mahnig, and A. O. Rodriguez, "Schemata, distributions, and graphical models in evolutionary optimization," *J. Heuristics*, vol. 5, pp. 215–247, 1999.

[12] M. Pelikan, D. E. Goldberg, and E. Cantu-Paz, "BOA: The Bayesian optimization algorithm," in *Proc. Genetic Evolutionary Computation Conf.*, 1999, pp. 525–532.

[13] B. T. Zhang, "A Bayesian framework for evolutionary computation," in *Proc. 1999 Congr. Evolutionary Computation*, vol. 1, Washington, DC, July 1999, pp. 722–228.

[14] P. A. N. Bosman and D. Thierens, "Advancing continuous IDEA's with mixture distributions and factorization selection metrics," in *Proc. Optimization by Building and Using Probabilistic Models OBUPM Workshop at the Genetic and Evolutionary Computation Conf., GECCO-2001*. San Francisco, CA, 2001, pp. 208–212.

[15] V. V. Miagkikh and W. F. Punch, "Global search in combinatorial optimization using reinforcement learning algorithms," in *Proc. 1999 Congr. Evolutionary Computation*, vol. 1, Washington, DC, 1999, pp. 189–196.

[16] C. González, J. A. Lozano, and P. Larrañaga, "Analyzing the PBIL algorithm by means of discrete dynamical systems," *Complex Syst.*, vol. 12, pp. 465–479, 2000.

[17] M. Höhfeld and G. Rudolph, "Toward a theory of population-based incremental learning," in *Proc. 4th IEEE Conf. Evolutionary Computation*, Indianapolis, IN, 1997, pp. 1–5.

[18] H. Mühlenbein and T. Mahnig, "Convergence theory and application of the factorized distribution algorithm," *J. Comput. Inform. Technol.*, vol. 7, pp. 19–32, 1999.

[19] T. Blickle and L. Thiele, "A mathematical analysis of tournament selection, genetic algorithms," in *Proc. 6th Int. Conf. (ICGA95)*. San Francisco, CA, 1995, pp. 9–16.

[20] ——, "A comparison of selection schemes used in evolutionary algorithms," *Evol. Comput.*, vol. 4, pp. 361–394, 1996.

[21] X. Qi and F. Palmieri, "Theoretical analysis of evolutionary algorithms with an infinite population size in continuous space part 1: Basic properties of selection and mutation," *IEEE Trans. Neural Networks*, vol. 5, pp. 102–119, Jan. 1994.

[22] M. D. Voss, *The Simple Genetic Algorithm: Foundations and Theory*. Cambridge, MA: MIT Press, 1999.

[23] I. Karcz-Duleba, "Dynamics of infinite populations evolving in a landscape of uni- and bimodal fitness functions," *IEEE Trans. Evol. Comput.*, vol. 5, pp. 398–409, Aug. 2001.

[24] D. E. Goldberg and K. Deb, "A comparative analysis of selection schmes used in genetic algorithms," in *Foundations of Genetic Algorithms 1*. San Mateo, CA: Morgan Kaufmann, 1991, pp. 69–93.

[25] J. He and X. Yao, "Drift analysis and average time complexity of evolutionary computation," *Artif. Intell.*, vol. 127, pp. 57–85, 2001.

[26] ——, "From an individual to a population: An analysis of the frist hitting time of population-based evolutionary algorithms," *IEEE Trans. Evol. Comput.*, vol. 6, pp. 495–511, Oct. 2002.

[27] ——, "Toward an analytic framework for analyzing the computation time of evolutionary algorithms," *Artif. Intell.*, vol. 145, pp. 59–97, 2003.

[28] T. Bäck, "Selective pressure in evolutionary algorithms: A characterization of selection mechanisms," in *Proc. 1st IEEE Conf. Evolutionary Computation*, Orlando, FL, June 1994, pp. 57–62.

[29] H. Meuhlenbein and D. Schlierkamp-Voosen, "Predictive models for the breeder genetic algorithms," *Evol. Comput.*, vol. 1, pp. 25–49, 1993.

[30] D. Thierens, "Analysis and Design of Genetic Algorithms," Ph.D. dissertation, Dept. Elec. Eng., Kath. Univ., Leuven, Belgium, 1995.

[31] J. Whittaker, *Graphical Models in Applied Multivariate Statistics*, ser. Probability and Mathematics Statistics. New York: Wiley, 1991.

[32] H. Mühlenbein and T. Mahnig, "The factorized distribution algorithm for additively decomposed functions," in *Proc. 1999 Congr. Evolutionary Computation*, Washington, DC, July 1999, pp. 752–759.

[33] ——, "FDA—A scalable evolutionary algorithm for the optimization of additively decomposed functions," *Evol. Comput.*, vol. 4, pp. 353–376, 1999.

[34] S. Tsutsui, M. Pelikan, and D. E. Goldberg, "Evolutionary algorithm using marginal histogram models in continuous domain," Univ. Illinois, Chicago, IlloGAL Rep., 2001.

[35] Z. Michalewicz, *Genetic Algorithms+Data Structure=Evolution Programs*. Berlin, Germany: Springer-Verlag, 1996. 3rd version.

[36] J. H. Holland, *Adaptation in Natural and Artificial Systems*. Cambridge, MA: MIT Press, 1975.

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

[38] C. R. Stephens and H. Waelbroeck, "Schemata evolution and building blocks," *Evol. Comput.*, vol. 7, pp. 109–129, 1999.

[39] R. Poli, "Exact schema theory for GP and variable-length gas with one-point crossover," *Genetic Program. Evolvable Machines*, vol. 2, pp. 123–163, 2001.

[40] A. E. Nix and M. D. Vose, "Modeling genetic algorithms with markov chains," *Ann. Math. Artif. Intell.*, vol. 5, pp. 79–88, 1992.

[41] G. Rudolph, "Convergence analysis of canonical genetic algorithm," *IEEE Trans. Neural Networks*, vol. 5, pp. 96–101, Jan. 1994.

[42] R. Cerf, "Asymptotic convergence of genetic algorithms," *Adv. Appl. Prob.*, vol. 30, pp. 521–550, 1998.

[43] D. Greenhalgh and S. Marshall, "Convergence criteria for GAs," *SIAM J. Comput.*, vol. 30, pp. 269–282, 2000.

[44] Y. S. Chow and H. Yeicher, *Probability Theory*, 3rd ed. Berlin, Germany: Springer-Verlag, 1997.

[45] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. Berlin, Germany: Springer-Verlag, 1996.

[46] Yu. M. Svirezhev and V. P. Passekov, *Fundamental of Mathematical Evolutionary Genetics*, ser. Mathematics and Its Applications (Soviet Series). Norwell, MA: Kluwer, 1990.

[47] S. Wright, *Genetic and Biometric Foundations*. Chicago, IL: Univ. Chicago Press, 1968, vol. 1–3.

[48] M. G. Bulmer, *The Mathematical Theory of Quantitative Genetics*. Oxford, U.K.: Clarendon, 1985.

Qingfu Zhang (M'01) received the B.Sc. degree in mathematics from Shanxi University, Shanxi, China, in 1984, the M.Sc. degree in applied mathematics, and the Ph.D. degree in information engineering from Xidian University, Xi'an, China, in 1991 and 1994, respectively.

He has been a Lecturer in the Department of Computer Science, University of Essex, Colchester, U.K., since 2000. From 1994 to 2000, he worked with the Changsha Institute of Technology, China, Hong Kong Polytechnic University, Kowloon, Hong Kong, the German National Research Centre for Information Technology, Germany, and the University of Manchester Institute of Science and Technology, Manchester, U.K. His main research areas are evolutionary computation, optimization, neural networks, data analysis and their applications.

Heinz Mühlenbein received the Ph.D. degree in applied mathematics from Bonn University, Bonn, Germany, in 1971.

He is currently a Research Manager with the Institute of Autonomous Intelligent Systems, Fraunhofer Gesellschaft, St. Augustin, Germany. His research activities spanned the areas of time-sharing operating systems, computer networks, and parallel programming. Since 1985, he has been actively conducting research in soft computing including neural networks, evolutionary computation, and probabilistic logic.