On Stability of Fixed Points of Limit Models of Univariate Marginal Distribution Algorithm and Factorized Distribution Algorithm

Qingfu Zhang, *Member, IEEE*

***Abstract*—This paper aims to study the advantages of using higher order statistics in estimation distribution of algorithms (EDAs). We study two EDAs with two-tournament selection for discrete optimization problems. One is the univariate marginal distribution algorithm (UMDA) using only first-order statistics and the other is the factorized distribution algorithm (FDA) using higher order statistics. We introduce the heuristic functions and the limit models of these two algorithms and analyze stability of these limit models. It is shown that the limit model of UMDA can be trapped at any local optimal solution for some initial probability models. However, degenerate probability density functions (pdfs) at some local optimal solutions are unstable in the limit model of FDA. In particular, the degenerate pdf at the global optimal solution is the unique asymptotically stable point in the limit model of FDA for the optimization of an additively decomposable function. Our results suggest that using higher order statistics could improve the chance of finding the global optimal solution.**

***Index Terms*—Estimation of distribution algorithms (EDAs), factorized distribution algorithm (FDA), heuristic function, stability, univariate marginal distribution algorithm (UMDA).**

I. Introduction

ESTIMATION OF distribution algorithms (EDAs) have been recently recognized as a new computing paradigm in evolutionary computation (e.g., [1]–[16]). Unlike other evolutionary algorithms, EDAs do not use crossover or mutation. Instead, they explicitly extract global statistical information from the selected solutions and build a posterior probability model of promising solutions, based on the extracted information. New solutions are sampled from the model thus built. According to the statistical information they exploit, EDAs can be classified into the following two categories.

- EDAs using only first-order statistics: This class of EDAs includes univariate marginal distribution algorithm (UMDA) [13], compact genetic algorithm (cGA) [9], and population-based incremental learning (PBIL) [3]. They employ probability models in which all the variables are independent and, hence, only need to estimate the marginal probability of each variable in the selected solutions at each iteration.
- EDAs using higher order statistics: This class of EDAs use a conditional dependence chain or network to model the probability distributions. Among them are mutual information maximization for input clustering (MIMIC) [4], combining optimizers with mutual information trees (COMIT) [12], factorized distribution algorithm (FDA) [14], estimation of Bayesian networks algorithm (EBNA) [10], and Bayesian optimization algorithm (BOA) [5], to name only a few. These algorithms need to estimate joint distributions of variable pairs or groups in the selected solutions.

The dynamic behavior of UMDA is very similar to that of genetic algorithms with uniform crossover (GAUX) [13]. Thierens [17] analyzed the ability of GAUX and concluded that GAUX may not find the global optimal solutions for a class of deceptive functions. Höhfeld and Rudolph [15] and González *et al.* [19] have studied the behavior of PBIL with elitist selection in discrete space. Based on Wright's equation, Mühlenbein and Mahning analyzed UMDA with proportional selection [20]. Berny investigated the relationship between PBIL and a gradient dynamical system [21]. These results suggest that EDAs using only first-order statistics have very limited ability to find global optimal solutions. Mühlenbein and Mahning discussed the convergence of FDA for the optimization of separable additively decomposable functions [35]. Since there are no variable overlaps in their objective functions, their FDA is mathematically equivalent to UMDA. In EDAs, the estimation of distribution is often separated into two phases: model selection and model fitting. Bosman and Thierens showed that EDAs could perform very badly if the structure of the selected model mismatches the variable-interaction structure in the objective function to be optimized [18]. Pelikan *et al.* studied the behavior of BOA and analyzed the critical population size and time complexity [16]. Zhang and Mühlenbein have proved that FDA will globally converge with proportional selection for optimization of continuous additively decomposable functions with overlaps [28]. Very recently, Zlochin *et al.* have proposed a general framework for ant colony optimization and EDAs [29]. However, the working mechanisms of EDAs are far from clear.

The commonly used approaches for studying evolutionary algorithms include the stochastic process approach [22], [23] and the heuristic function approach [31], [32] among others. The stochastic process approach often characterizes an evolutionary algorithm as a Markov chain with the current population being the state variables, and then study the convergence in the sense of probability. Such an approach exactly models the behavior of

Manuscript received June 6, 2002; revised March 7, 2003. This work was supported in part by the Engineering and Physical Sciences Research Council (EPSRC) under Grant GR/R64742/01 and in part by a Research Promotion Fund Grant from the University of Essex, Colchester, U.K.

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

Digital Object Identifier 10.1109/TEVC.2003.819431

an evolutionary algorithm (EA). It has been successfully applied to EAs with finite population for some typical examples. Several very interesting results on average time complexity have been derived [24]–[27]. Since most EDAs maintain a probability model at each generation, the transition probability is very difficult to formulate and analyze. Therefore, it is very hard in general to apply the stochastic approach to study EDAs. The heuristic function approach often assume that the size of population is infinite [31], [32]. As a result, the iterative process of an EA is modeled by a deterministic nonlinear mapping called the heuristic function. Then, analysis on the dynamics of this heuristic function is performed. The behavior of a EA with large population can be approximated by that of the deterministic dynamic system defined by the heuristic function. Therefore, analysis results on heuristic function are useful for understanding EAs with large population. However, some properties of EAs with small population cannot be observed from their heuristic functions.

The goal of this paper is to investigate the advantages of using higher order statistics in EDAs. We consider UMDA and FDA for discrete optimization problems. Both of them use the same selection scheme: two-tournament selection. We define the heuristic functions and the limit models of UMDA and FDA. We show that, in the case of a general objective function, the limit model of FDA has a better chance of obtaining a global optimal solution than that of UMDA. In the case of an additively decomposable objective function, we prove that the unique asymptotically stable fixed point of the limit model of FDA is the degenerate probability density function (pdf) at the global optimal solution, which implies that FDA can converge to the global optimal solution. We also give an example to show that an additively decomposable objective function has some local optimal solutions, where the limit model of UMDA can become stuck. The results in this paper demonstrate that the chance of converging to the global optimal solution can be increased by using higher order statistics in EDAs.

II. Preliminaries

Let $\Omega = \{0,1\}^n$, $x = (x_1, x_2, \cdots, x_n) \in \Omega$ and P be a *probability measure* on Ω. Then, P is determined by its probability density function (pdf): $p(x)$

$$p(x) = P\left\{\omega = (\omega_1, \omega_2, \cdots, \omega_n) \in \Omega : \omega_i = x_i,\ 1 \le i \le n\right\}. \tag{1}$$

Let $K = \{i_1, i_2, \cdots, i_k\}$ be a subset of $\{1, 2, \cdots, n\}$ and define $x_K = (x_{i_1}, x_{i_2}, \cdots, x_{i_k})$. The *marginal pdf* $p(x_K)$ of $p(x)$ is defined as

$$p(x_K) = P\left\{\omega = (\omega_1, \omega_2, \cdots, \omega_n) \in \Omega : \omega_i = x_i,\ i \in K\right\}. \tag{2}$$

Let J be another subset of $\{1, 2, \cdots, n\}$. A conditional probability for x_K given x_J is defined as

$$p(x_K | x_J) = \frac{p(x_{K \cup J})}{p(x_J)}.$$

To be absolutely precise, we should use $p_K(x_K)$ and $p_{K|J}(x_K|x_J)$ to denote the above marginal *pdf* and the conditional probability, respectively, but for simplicity, we use $p(x_K)$ and $p(x_K|x_J)$ as many other authors have done. The context will always make our meaning clear. To avoid confusion, the value of $p(x_K)$ at $x_K = \widetilde{x}_K$ and $p(x_K|x_J)$ at $x_K = \widetilde{x}_K$ and $x_J = \widetilde{x}_J$ will be denoted in this paper by $p(x_K = \widetilde{x}_K)$ and $p(x_K = \widetilde{x}_K | x_J = \widetilde{x}_J)$, respectively, if and when necessary.

For $i, j \in \{1, 2, \cdots, n\}$ and $i \neq j$, we call $p(x_i)$ and $p(x_i, x_j)$ a *univariate marginal pdf* (*umpdf*) and a *bivariate marginal pdf* (*bmpdf*) of $p(x)$, respectively. Clearly, there are n umpdfs and $n(n-1)/2$ bmpdfs.

We need $2^n - 1$ parameters to represent a general pdf $p(x)$ on Ω. However, some constraints on the dependence among variables with respect to $p(x)$ can simplify the representation of $p(x)$ [33]. If all the variables $x_1, x_2, \cdots, x_n$ are independent with respect to $p(x)$, then

$$p(x) = \prod_{i=1}^{n} p(x_i). \tag{3}$$

We denote the set of all the pdfs on Ω satisfying (3) by $PF(1)$. If $p(x) \in PF(1)$, we need only n parameters to represent $p(x)$. If the conditional dependence graph of variables $x_1, x_2, \cdots, x_n$ with respect to $p(x)$ is a Markov chain in which x_i is the ith node, then

$$p(x) = \frac{\prod_{i=1}^{n-1} p(x_i, x_{i+1})}{\prod_{i=2}^{n-1} p(x_i)} \tag{4}$$

where we define $0/0 = 0$. In fact, if $\prod_{i=2}^{n-1} p(x_i) = 0$, then $\prod_{i=1}^{n-1} p(x_i, x_{i+1}) = 0$ and $p(x) = 0$. We denote by $PF(2)$ the set of all the pdfs on Ω satisfying (4). If $p(x) \in PF(2)$, the representation of $p(x)$ needs $O(n)$ parameters.

There are 2^n points $x^1, x^2, \cdots, x^{2^n}$ in Ω. Let $p(x)$ be a pdf on Ω. In the following, we call the 2^n-dimensional vector $(p(x^1), \cdots, p(x^N))$ the *corresponding vector* of $p(x)$. Obviously, different pdfs have different corresponding vector. A corresponding vector can be regarded as a point in 2^n vector dimensional space.

Let $p(x)$ and $\overline{p}(x)$ be pdfs on Ω. Their respective corresponding vectors are denoted by p and $\overline{p}$. The distance between $p(x)$ and $\overline{p}(x)$ is defined as

$$d(p, \overline{p}) = \sum_{x \in \Omega} |p(x) - \overline{p}(x)| = \left|p(x^1) - \overline{p}(x^1)\right| + \left|p(x^2) - \overline{p}(x^2)\right| + \cdots + \left|p\left(x^{2^n}\right) - \overline{p}\left(x^{2^n}\right)\right|. \tag{5}$$

In other words, the distance between two pdfs is the L_1 distance between their corresponding vectors.

Let $y = (y_1, y_2, \cdots, y_n)$ be a given point in Ω. We call the following pdf (of x) on Ω:

$$\mathrm{dep}(x, y) = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

the *degenerate pdf* (*dpdf*) at y.[1] It is easy to show that $\mathrm{dep}(x, y) \in PF(1)$, and $\mathrm{dep}(x, y) \in PF(2)$ for any given y.

Let F be a metric space with the distance function d and Ψ a map from F into F. Given a point $z \in F$, Ψ defines a sequence on F

$$z, \Psi(z), \Psi^2(z), \Psi^3(z), \cdots. \tag{7}$$

[1]In $\mathrm{dep}(x, y)$, y is the function parameter vector and x is the variable vector.

A point $z \in F$ satisfying $\Phi(z) = z$ is called a *fixed point* of Ψ. A fixed point z is called *stable* if for any $\varepsilon > 0$, there exists a $\delta > 0$ such that, if $\overline{z} \in F$ satisfies $d(z, \overline{z}) < \delta$, then $d(z, \Psi^t(\overline{z})) < \varepsilon$ for all $t \geq 0$. A stable fixed point z is said to be *asymptotically stable* if there exists a $\delta > 0$ such that, if $d(\overline{z}, z) < \delta$, then $\lim_{t\to\infty} d(z, \Psi^t(\overline{z})) = 0$.

Intuitively, if a point in sequence (7) is sufficiently close to a stable fixed point, then all the points thereafter remains close to this stable fixed point. It is stable to all small perturbations. If a fixed point is not stable, no matter how close a point in (7) is to it, some points thereafter may become far from it. For a asymptotically stable fixed point, if a point in sequence (7) is sufficiently close to it then the sequence will converge to it. The convergence is stable to small perturbations.

III. UMDA AND FDA

We consider the following general combinatorial optimization problem:

$$\text{Maxmize } f(x) \tag{8}$$

where $x = (x_1, \cdots, x_n) \in \Omega$ and the objective function $f : \Omega \to R$. A point $y = (y_1, y_2, \cdots, y_n) \in \Omega$ is called a local maximal solution if $f(x) \leq f(y)$ for any $x = (x_1, x_2, \cdots, x_n) \in \Omega$ with $\|x - y\|_1 = \sum_{i=1}^{n} |x_i - y_i| = 1$. y is said to be a strict local maximal solution if $f(x) < f(y)$ for any $x \in \Omega$ with $\|x - y\|_1 = 1$. y is called a global maximal solution if $f(x) \leq f(y)$ for any $x \in \Omega$. y is said to be the strict global maximal solution if $f(x) < f(y)$ for any $x \neq y$.

EDAs for problem (8) can be regarded as a learning process in which "knowledge" about the optimal solutions is refined step by step. This knowledge is expressed by a probability function $p(x,t)$ which gives the posterior probability of each point x in Ω to be an optimal solution at iteration t. To generate $p(x, t+1)$, EDAs work as follows.[2]

Step 1) Independently sample λ points in Ω from $p(x,t)$ to form a population[3] Q.

Step 2) Select μ points from Q to form a parent population S by using a selection scheme.

Step 3) Extract statistical information from the points in S and build $p(x, t+1)$, based on this information.

Let $p(*, t)$ be the corresponding vector of $p(x,t)$, EDA generates the following sequence of points in the 2^n-dimensional vector space[4]

$$p(*, 0), p(*, 1), \cdots . p(*, t), \cdots . \tag{9}$$

Since there are random factors in EDA (e.g., a sampling point in Ω from $p(x,t)$ in Step 1 is random), the above sequence is random.

If $p(x,t)$ satisfies (3) or (4), or the conditional dependence graph of $p(x,t)$ is a Bayesian network, Step 1 will be very easily implemented. Several selection schemes can be used in Step 2. In this paper, we will focus on the widely used two-tournament selection. To select a point for S, the two-tournament selection randomly picks two points from Q and then chooses the point with larger objective function value to enter S. If two points happen to have the same objective function value, it will randomly choose one to enter S. This procedure should be repeated μ times in Step 2. Let $p^s(x,t)$ be the probability that a point x is in S. Then

$$\begin{aligned} p^s(x,t) = [p(x,t)]^2 &+ 2p(x,t) \sum_{f(z)<f(x)} p(z,t) \\ &+ p(x,t) \sum_{z\neq x,\ f(z)=f(x)} p(z,t). \end{aligned} \tag{10}$$

Suppose that S contain the following μ points:

$$\begin{aligned} y^1(t) &= \left(y_1^1(t), y_2^1(t), \cdots y_n^1(t)\right) \\ y^2(t) &= \left(y_1^2(t), y_2^2(t), \cdots y_n^2(t)\right) \\ &\vdots \\ y^\mu(t) &= \left(y_1^\mu(t), y_2^\mu(t), \cdots y_n^\mu(t)\right). \end{aligned}$$

Now, key issues in Step 3 arise.

- *Q1 Information collection*: What kinds of statistical information should we extract from the above μ points?
- *Q2 Model selection*: What kind of model should we use for $p(x, t+1)$?
- *Q3 Model fitting*: How is the extracted information used to build $p(x, t+1)$?

These highly related issues have to be tackled at a reasonable computational cost. Several different EDAs have been proposed to address these issues.

A. UMDA

UMDA [13] is the simplest version of EDAs. In UMDA, $p(x,t) \in PF(1)$ for all $t \geq 0$. In Step 3, this algorithm computes the following first-order statistics from S:

$$\begin{aligned} A_1(t) &= \frac{\sum_{j=1}^{\mu} y_1^j(t)}{\mu} \\ A_2(t) &= \frac{\sum_{j=1}^{\mu} y_2^j(t)}{\mu} \\ &\vdots \\ A_n(t) &= \frac{\sum_{j=1}^{\mu} y_n^j(t)}{\mu}. \end{aligned}$$

Obviously, $A_j(t)$ is the percentage of the points in S with the jth variable equal to 1. The computational cost for all the $A_i(t)$ is $O(n\mu)$. As to the model selection, UMDA assumes that all the variables in $p(x, t+1)$ are independent of one another. It builds $p(x, t+1)$ as follows:

$$p(x, t+1) = \prod_{i=1}^{n} A_i^{x_i}(t) \left[1 - A_i(t)\right]^{1-x_i}. \tag{11}$$

[2]Some existing EDAs use incremental learning and other techniques to improve their peformace. These algorithms are slightly different from the EDA discussed in this paper.

[3]As in genetic algorithms, a population in EDAs can have repeated elements.

[4]In practice, we can only run an EDA for finite steps. This means that we generate an finite random sequence. But for simplicity, we assume that an EDA runs for infinite steps.

Note that the expectation of $y_i^j(t)$ is

$$E\left(y_i^j(t)\right) = p^s(x_i = 1, t) \tag{12}$$

for any i and j, where $p^s(x_i, t)$ is a umpdf of $p^s(x, t)$. By the law of large numbers, $A_i(t)$ converges to $p^s(x_i = 1, t)$ almost certainly as $\lambda \to \infty$ and $\mu \to \infty$ [34]. From (11)

$$p(x, t+1) \to \prod_{i=1}^{n} p^s(x_i, t) \text{ almost certainly} \tag{13}$$

as $\lambda \to \infty$ and $\mu \to \infty$ in UMDA. Based on the observations (10) and (13), we introduce the following UMDA heuristic function $\Phi_1 : PF(1) \to PF(1)$:

$$q(x) = \Phi_1\left(p(x)\right) \tag{14}$$

where $q(x)$ is determined by

$$\boxed{\begin{aligned} p^s(x) &= [p(x)]^2 + 2p(x) \sum_{z\in\Omega, f(z)<f(x)} p(z) \\ &\quad + p(x) \sum_{z\in\Omega, z\neq x, f(z)=f(x)} p(z) \\ q(x) &= \prod_{i=1}^{n} p^s(x_i). \end{aligned}} \tag{15}$$

By (10) and (13), If $\lambda, \mu \gg 1$, a random sequence of pdfs generated by UMDA can be approximated by

$$p(x, t+1) = \Phi_1\left(p(x, t)\right), \quad (t = 0, 1, 2, \cdots). \tag{16}$$

Sequence (16) is called the *limit model* of UMDA. This limit model defines a deterministic sequence in $PF(1)$. Integrating both sides of (16) over variables $x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n$, we can prove that

$$p(x_i, t+1) = p^s(x_i, t) \tag{17}$$

holds in the limit model of UMDA for any $1 \le i \le n$.

The dynamic behavior of the limit model (16) approximates UMDA when λ and μ are large. However, it is a very challenging task to compute the approximation error of (16).

B. FDA

In UMDA, all the variables in the probability model $p(x, t)$ are assumed to be independent of one another for all $t \ge 0$. Only first-order statistics in S are utilized in the phase of model fitting. Several EDAs have been proposed using higher order statistics in S [4], [5], [14]. These algorithms always use a chain, a tree, or a network to model the conditional dependence graph of $p(x, t)$. Some of them involve the selection of the structure of the network according to a predefined criterion in Step 3. FDA [14] uses a fixed probability model throughout the algorithm, where the model is specified based on the knowledge of the problem to be solved. In this paper, we consider an FDA in which the conditional dependence graph of variables $x_1, x_2, \cdots, x_n$ is a Markov chain in which x_i is the ith node, i.e., $p(x, t) \in PF(2)$ for all $t \ge 0$. This FDA is very similar to MIMIC [4], but MIMIC adjusts its Markov chain at each iteration. In Step 3, FDA computes $A_{i,i+1}(g, h, t)$ $(1 \le i \le n-1, g, h = 0, 1)$, the percentage of the points in S with the ith and $(i+1)$th variables being g and h, respectively. The computational cost of FDA at Step 3 is $O(n^2\mu)$. FDA builds $p(x, t+1)$ as follows:

$$p(x, t+1) = \frac{\prod_{i=1}^{n-1} A_{i,i+1}(x_i, x_{i+1}, t)}{\prod_{i=2}^{n-1} \left[A_{i,i+1}(x_i, 0) + A_{i,i+1}(x_i, 1)\right]}. \tag{18}$$

By the law of large numbers [34], $A_{i,i+1}(g, h, t)$ converges to $p^s(x_i = g, x_{i+1} = h)$ with probability one as $\lambda \to \infty$ and $\mu \to \infty$. Consequently, the right side of (18) will converge to

$$\frac{\prod_{i=1}^{n-1} p^s(x_i, x_{i+1}, t)}{\prod_{i=2}^{n-1} p^s(x_i, t)}$$

with probability one as $\lambda \to \infty$ and $\mu \to \infty$. Therefore, we define the following FDA heuristic function $\Phi_2 : PF(2) \to PF(2)$:

$$q(x) = \Phi_2\left(p(x)\right) \tag{19}$$

where $q(x)$ is determined by

$$\boxed{\begin{aligned} p^s(x) &= [p(x)]^2 + 2p(x) \sum_{z\in\Omega, f(z)<f(x)} p(z) \\ &\quad + p(x) \sum_{z\in\Omega, z\neq x, f(z)=f(x)} p(z) \\ q(x) &= \frac{\prod_{i=1}^{n-1} p^s(x_i, x_{i+1})}{\prod_{i=2}^{n-1} p^s(x_i)}. \end{aligned}} \tag{20}$$

Therefore, if $\lambda, \mu \gg 1$, a random sequence of pdfs generated by FDA in $PF(2)$ can be approximated by

$$p(x, t+1) = \Phi_2\left(p(x, t)\right), \quad (i = 1, 2, \cdots, n). \tag{21}$$

We call the above model the limit model of FDA. Obviously, the model (21) generates a deterministic sequence. Integrating both sides of (21) over variables $x_1, \cdots, x_{i-1}, x_{i+2}, \cdots, x_n$, we can prove that

$$p(x_i, x_{i+1}, t+1) = p^s(x_i, x_{i+1}, t) \tag{22}$$

holds in the limit model of FDA for any $1 \le i \le n-1$.

IV. Fixed Points of the Limit Models of UMDA and FDA

As stated in the previous section, some EDAs (such as FDA) need to compute higher order statistics. As a result, these algorithms are computationally more expensive than UMDA at each iteration. A very natural question arises here: What are the benefits of using higher order statistics in building $p(x, t+1)$ in EDAs? In this section, we will study the stability of dynamics of the heuristic functions Φ_1 and Φ_2 in $PF(1)$ and $PF(2)$, respectively.

As shown in Section III, the limit models defined by Φ_1 and Φ_2 approximate the behaviors of UMDA and FDA with large population $(\lambda, \mu \gg 1)$, respectively. Obviously, the analysis on global dynamical behaviors of these models should be conducted. However, all our works in this paper are on the local stability of these limit models. The main reasons behind our work are the following.

- In general, it is extremely difficult to analyze the global stability of these models. Study on the local stability is prerequisite for understanding the global behavior.
- There are random mechanisms in EDA with finite population, UMDA, and FDA with finite population can be regarded as their limit models with random perturbations. Therefore, we should study the stability of these limit models.
- In some scenarios, the initial pdf $p(x,0)$ in EDA may be set as a nonuniform distribution based on problem-special knowledge or by other heuristics. Moveover, random mechanisms are involved in EDA with finite population. Therefore, it is possible that a pdf in the sequence of pdfs generated by UMDA or FDA with finite population is close to any dpdf. Investigation of the local behaviors of the limit models near a dpdf should be helpful for understanding the behaviors of these algorithms.

A. *Fixed Points of* Φ_1

Theorem 1: For any given $y=(y_1,\cdots,y_n)\in\Omega$, the *dpdf* $\mathrm{dep}(x,y)$ is a fixed point of Φ_1.

Proof: Let $p(x)=\mathrm{dep}(x,y)$, and $q(x)=\Phi_1(p(x))$. Then

$$p(x)=\begin{cases}1, & x=y\\ 0, & \text{otherwise}\end{cases}.$$

By (15)

$$p^s(y)=\underbrace{[p(y)]^2}_{=1}+2p(y)\underbrace{\sum_{f(z)<f(y)}p(z)}_{=0}+p(y)\times\underbrace{\sum_{z\neq y,f(z)=f(y)}p(z)}_{=0}=1.$$

Therefore, $p^s(x)=0$ for all $x\neq y$. The marginal probability $p^s(x_i)$ of $p^s(x)$ is

$$p^s(x_i)=\begin{cases}1, & \text{if } x_i=y_i\\ 0, & \text{otherwise}\end{cases}.$$

Since $q(x)=\prod_{i=1}^n p^s(x_i)$, we have

$$q(x)=\begin{cases}1, & \text{if } x=y\\ 0, & \text{otherwise}\end{cases}$$

which implies $q(x)=p(x)$. Therefore, $\mathrm{dep}(x,y)$ is a fixed point of Φ_1. ■

This theorem says that if $p(x,0)$ in the limit model of UMDA (16) is a dpdf at a given point y, then all the $p(x,t)$ for $t>0$ are unchanged in (16). This result is not surprising, in UMDA with finite population, if $p(x,0)$ is a dpdf at a given point y, then all the points in Q and consequently in S are y. Therefore, $p(x,t)=p(x,0)$ for all $t\geq 0$ in UMDA with finite population. In the following, we will study the stability of these dpdf fixed points.

Theorem 2: If $y=(y_1,y_2,\cdots y_n)$ is a strict local maximal solution of problem (8), then the dpdf $\mathrm{dep}(x,y)$ is an asymptotically stable fixed point of Φ_1.

Proof: Let $p(x,0)\in PF(1)$ and $p(x,t)=\Phi_1(p(x,t-1))=\Phi_1^t(p(x,0))$. Let $\mathrm{dep}(*,y)$ be the corresponding vector of $\mathrm{dep}(x,y)$. To prove the theorem, it suffices to show that there exists a $\delta>0$ such that, if $d(\mathrm{dep}(*,y),p(*,0))<\delta$ (the distance between two pdfs as defined in (5)), then $\lim_{t\to\infty}p(x,t)=\mathrm{dep}(x,y)$ for any $x\in\Omega$ and $d(p(*,t+1),\mathrm{dep}(*,y))\leq d(p(*,t),\mathrm{dep}(*,y))$ for all $t\geq 0$.

Denote $d(p(*,0),\mathrm{dep}(*,y))$ by θ. Then

$$\sum_{x\in\Omega}|p(x,0)-\mathrm{dep}(x,y)|=\theta.$$

Noting that $\sum_{x\in\Omega}p(x,0)=1$ and

$$\mathrm{dep}(x,y)=\begin{cases}1, & \text{if } x=y\\ 0, & \text{otherwise}\end{cases}$$

we have

$$p(x=y,0)=1-\frac{\theta}{2}.$$

Since $p(x,0)\in PF(1)$

$$p(x=y,0)=\prod_{i=1}^n p(x_i=y_i,0).$$

Thus

$$p(x_i=y_i,0)\geq 1-\frac{\theta}{2}.$$

Denote $p(x_i=1-y_i,0)$ by ξ_i. Since $P(x_i=1-y_i,0)=1-p(x_i=y_i,0)$, we have

$$\xi_i\leq\frac{\theta}{2}$$

for $i=1,2,\cdots,n$. Now, we calculate a lower bound for $p^s(y,0)$

$$\begin{aligned}p^s(y,0)=&\ [p(y,0)]^2+2p(y,0)\sum_{f(z)<f(y)}p(z,0)+p(y,0)\\ &\times\sum_{z\neq y,f(z)=f(y)}p(z,0)\\ &\underbrace{\geq}_{\text{since } y \text{ is a strict local maximal solution}}\left[\prod_{i=1}^n p(x_i=y_i,0)\right]^2\\ &+2\left[\prod_{i=1}^n p(x_i=y_i,0)\right]\sum_{\|z-y\|_1=1}p(z,0)\\ =&\ \underbrace{\left[\prod_{i=1}^n(1-\xi_i)\right]^2}_{p(x_i=y_i,0)=1-\xi_i}+2\sum_{j=1}^n(1-\xi_j)\xi_j\prod_{i=1,i\neq j}^n(1-\xi_i)^2\\ =&\ \prod_{i=1}^n\left(1-2\xi_i+\xi_i^2\right)+2\sum_{j=1}^n\left(\xi_j-\xi_j^2\right)\\ &\times\prod_{i=1,i\neq j}^n\left(1-2\xi_i+\xi_i^2\right)=1-O\left((\max\xi_i)^2\right).\end{aligned}$$

[5] It follows that

$$p(y,1)=\prod_{i=1}^{n}p^s(x_i=y_i,0)\underbrace{\geq}_{p^s(x_i=y_i,0)\geq p^s(x=y,0)}[p^s(x=y,0)]^n$$
$$=\left[1-O\left((\max\xi_i)^2\right)\right]^n=1-O\left((\max\xi_i)^2\right)=1-O(\theta^2).$$

Therefore

$$d\left(p(*,1),p(*,y)\right)=|1-p(y,1)|+\underbrace{\sum_{z\in\Omega,z\neq y}p(z,1)}_{=1-p(y,1)}\leq O(\theta^2).$$

Bearing in mind that $\theta=d(p(*,0),p(*,y))$, there exists a positive $\delta>0$ such that

$$d\left(p(*,1),p(*,y)\right)\leq\frac{1}{2}\theta=\frac{1}{2}d\left(p(*,0),p(*,y)\right)$$

whenever $d(p(*,0),p(*,y))\leq\delta$.

Now, we repeat the above procedure with $p(x,0)$ replaced by $p(x,i)$ $(i=1,2,\cdots,t-1)$. We obtain

$$d\left(p(*,t),p(*,y)\right)\leq\frac{1}{2}d\left(p(*,t-1),p(*,y)\right).$$

Thus

$$d\left(p(*,t),p(*,y)\right)\leq\left(\frac{1}{2}\right)^t d\left(p(*,0),p(*,y)\right)$$

which implies $\lim_{t\to\infty}p(*,t)=\mathrm{dep}(*,y)$ and $d(p(*,t+1),\mathrm{dep}(*,y))<d(p(*,t),\mathrm{dep}(*,y))$ for all $t\geq 0$ if $d(p(*,0),p(*,y))\leq\delta$. ■

By this theorem, for a pdf sequence $p(x,0),p(x,1),\cdots$, generated in the limit model of UMDA, if a pdf $p(x,t_0)$ in the sequence is sufficiently close to the dpdf at a strict local maximal solution, there all the pdfs $p(x,t)$ for $t\geq t_0$ will remains close to this dpdf even there are some small perturbations in this sequence. This results implies that for a sequence of pdfs generated by UMDA with large population $(\lambda,\mu\gg 1)$, once a pdf in this sequence is close enough to a dpdf at any strict local maximal solution of the objective function $f(x)$, then there is a very high possibility that all the dpdf thereafter will be in a small neighborhood of this dpdf.

Theorem 3: If y is not a local maximal solution, then the dpdf $\mathrm{dep}(x,y)$ is not a stable fixed point of Φ_1.

Proof: If $y=(y_1,y_2,\cdots,y_n)$ is not a local maximal solution, there exists a point

$$v=(y_1,y_2,\cdots,y_{k-1},1-y_k,y_{k+1},\cdots,y_n)\in\Omega$$

such that $f(v)>f(y)$. To prove the theorem, it suffices to show that there exists a $\delta_0>0$ such that, for any given $1>\varepsilon>0$, there is a $p(x,0)\in PF(1)$ with $d(p(*,0),\mathrm{dep}(*,y))\leq\varepsilon$ and $\lim_{t\to\infty}d(p(*,t),\mathrm{dep}(*,y))\geq\delta_0$, where $p(x,t)=\Phi_1^t(p(x,0))$.

Let

$$p(x,0)=\begin{cases}1-\frac{\varepsilon}{2}, & x=y\\ \frac{\varepsilon}{2}, & x=v\\ 0, & \text{otherwise}\end{cases}.$$

[5]Let b and r be two variables, by $r=O(b)$ we mean that there exists a constant C such that $|r|\leq C|b|$.

Then, $p(x,0)\in PF(1)$ and

$$\begin{aligned}d\left(p(*,0),\mathrm{dep}(*,y)\right)&=\sum_{z\in\Omega}|p(z,0)-\mathrm{dep}(z,y)|\\&=\underbrace{|p(y,0)-\mathrm{dep}(y,y)|}_{=\frac{\varepsilon}{2}}\\&\quad+\underbrace{|p(v,0)-\mathrm{dep}(v,y)|}_{=\frac{\varepsilon}{2}}\\&\quad+\underbrace{\sum_{z\in\Omega,z\neq y,v}|p(z,0)-\mathrm{dep}(z,y)|}_{=0}=\varepsilon.\end{aligned}$$

Now, we calculate $p^s(x,0)$

$$p^s(y,0)=[p(y,0)]^2=\left(1-\frac{\varepsilon}{2}\right)^2,$$
$$p^s(v,0)=[p(v,0)]^2+2p(y,0)p(v,0)=1-\left(1-\frac{\varepsilon}{2}\right)^2$$
$$p^s(x,0)=0\text{ if }x\neq y,v.$$

Therefore

$$p^s(x_i=y_i,0)=\begin{cases}1, & i\neq k\\ \left(1-\frac{\varepsilon}{2}\right)^2, & i=k\end{cases}.$$

It follows that

$$p(x,1)=\prod_{i=1}^{n}p^s(x_i,0)=\begin{cases}\left(1-\frac{\varepsilon}{2}\right)^2, & x=y\\ 1-\left(1-\frac{\varepsilon}{2}\right)^2, & x=v\\ 0, & \text{otherwise}\end{cases}.$$

Repeating the above procedure, we obtain

$$p(x,t)=\begin{cases}\left(1-\frac{\varepsilon}{2}\right)^{2^t}, & x=y\\ 1-\left(1-\frac{\varepsilon}{2}\right)^{2^t}, & x=v\\ 0, & \text{otherwise}\end{cases}.$$

Thus

$$\begin{aligned}d\left(p(*,t),\mathrm{dep}(*,y)\right)&=|1-p(y,t)|+|p(v,t)|\\&=2\left[1-\left(1-\frac{\varepsilon}{2}\right)^{2^t}\right]\to 2,\ \text{ as }t\to\infty.\end{aligned}$$

which implies that $\mathrm{dep}(x,y)$ is not a stable fixed point. ■

This theorem says that in a sequence of pdfs generated in the limit model of UMDA, if $y\in\Omega$ is not a local maximal solution, no matter how close a pdf in this sequence is to the dpdf at y, the pdfs in the sequence may go far from this dpdf. Therefore, it is unlikely that UMDA with large population becomes trapped at a point which is not a local maximal solution.

Remark 1: Following the work of Höhfeld and Rudolph [15], González *et al.* [19] have studied the dynamical behavior of PBIL with elitist selection in the case of infinite population, which can be regarded as a variant of UMDA. They assumed that $f(x)\neq f(y)$ for any $x\neq y$. Therefore, all the local maximal solutions are strict in their case. They investigated the behavior of the probability vector, i.e. $(p(x_1=1,t),\cdots,p(x_n=1,t))$. Since the distance in $PF(1)$ defined in Section II is equivalent to the distance in the space of probability vectors, their results on stability are also true in $PF(1)$. However, we study two-tournament selection and our analyses are much simpler.

B. Fixed Point of Φ_2

Theorem 4: For any given $y = (y_1, \cdots, y_n) \in \Omega$, the dpdf $\mathrm{dep}(x, y)$ is a fixed point of Φ_2.

Proof: Let $p(x) = \mathrm{dep}(x, y)$, and $q(x) = \Phi_2(p(x))$. Then

$$p(x) = \begin{cases} 1, & x = y \\ 0, & \text{otherwise} \end{cases}.$$

As shown in Theorem 1

$$p^s(x = y) = 1.$$

We have

$$p^s(x_i) = \begin{cases} 1, & \text{if } x_i = y_i \\ 0, & \text{otherwise} \end{cases}$$

and

$$p^s(x_i, x_{i+1}) = \begin{cases} 1, & \text{if } x_i = y_i \text{ and } x_{i+1} = y_{i+1} \\ 0, & \text{otherwise} \end{cases}.$$

Since $q(x) = (\prod_{i=1}^{n-1} p^s(x_i, x_{i+1}))/(\prod_{i=2}^{n-1} p^s(x_i))$, we obtain

$$q(x) = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{otherwise} \end{cases}$$

which implies $q(x) = p(x)$. Therefore, $\mathrm{dep}(x, y)$ is a fixed point of Φ_2. ■

The following Lemma is needed in the proof of Theorem 6.

Lemma 5:

$$2\left[1 - (1 - \theta^2)^{n-1}\right] < \frac{\theta}{2}$$

for any given integer $n > 1$ and for any $\theta \in (0, (1/4(n-1)))$.

Proof: See the Appendix. ■

Now, we introduce the main results on the stability of the dpdf fixed points of Φ_2.

Theorem 6: If $y = (y_1, \cdots, y_n)$ is the strict global maximal solution, then the dpdf $\mathrm{dep}(x, y)$ is an asymptotically stable fixed point of Φ_2.

Proof: Let $p(x, 0) \in PF(2)$ and $p(x, t) = \Phi_2^t(p(x, 0))$. To prove the theorem, it suffices to show that there exists a $\delta > 0$ such that, if $d(\mathrm{dep}(*, y), p(*, 0)) < \delta$, then $\lim_{t\to\infty} p(x, t) = \mathrm{dep}(x, y)$ for all $x \in \Omega$ and $d(p(*, t+1), \mathrm{dep}(*, y)) < d(p(*, t), \mathrm{dep}(*, y))$ for all $t \geq 0$.

Denote $\theta = d(p(*, 0), \mathrm{dep}(*, y))$. Noting that

$$\begin{aligned} \theta &= \sum_{x\in\Omega} |p(x, 0) - \mathrm{dep}(x, y)| \\ &= |1 - p(y, 0)| + \sum_{x\in\Omega, x\neq y} |p(x, 0)| \end{aligned}$$

we have

$$p(y, 0) \geq 1 - \theta.$$

Therefore

$$\begin{aligned} p^s(y, 0) &= [p(y, 0)]^2 + 2p(y, 0) \underbrace{\sum_{f(z)<f(y)} p(z, 0)}_{=1-p(y,0)} \\ &\quad + p(y, 0) \underbrace{\sum_{z\neq y, f(z)=f(y)} p(z)}_{=0} \\ &= 2p(y, 0) - [p(y, 0)]^2 \geq 2(1-\theta) - (1-\theta)^2 \\ &= 1 - \theta^2. \end{aligned}$$

Then

$$\begin{aligned} p(y, 1) &= \prod_{i=1}^{n-1} p^s(x_i = y_i, x_{i+1} = y_{i+1}) / \prod_{i=2}^{n-1} p^s(x_i = y_i) \\ &\underbrace{\geq}_{p^s(x_i=y_i)\leq 1} \prod_{i=1}^{n-1} p^s(x_i = y_i, x_{i+1} = y_{i+1}) \\ &\underbrace{\geq}_{p^s(x_i=y_i, x_{i+1}=y_{i+1})\geq p^s(y,0)\geq 1-\theta^2} (1-\theta^2)^{n-1}. \end{aligned}$$

Thus

$$\begin{aligned} d(p(*, 1), \mathrm{dep}(*, y)) &= |1 - p(y, 1)| + \underbrace{\sum_{z\neq y} p(z, 1)}_{=1-p(y,1)} \\ &= 2\,|1 - p(y, 1)| \leq 2\left[1 - (1-\theta^2)^{n-1}\right]. \end{aligned}$$

By Lemma 5, if $d(p(*, 0), \mathrm{dep}(*, y)) < 1/4(n-1)$, we have

$$d(p(*, 1), \mathrm{dep}(*, y)) \leq \frac{\theta}{2} = \frac{1}{2} d(p(*, 0), \mathrm{dep}(*, y)).$$

Repeating the above procedure will lead to

$$d(p(*, t), \mathrm{dep}(*, y)) \leq \frac{1}{2^t} d(p(*, 0), \mathrm{dep}(*, y))$$

which implies that if $d(\mathrm{dep}(*, y), p(*, 0)) < (1/4(n-1))$, then $\lim_{t\to\infty} p(x, t) = \mathrm{dep}(x, y)$ and $d(p(*, t+1), \mathrm{dep}(*, y)) < d(p(*, t), \mathrm{dep}(*, y))$ for all $t \geq 0$. ■

This theorem shows that in a sequence of pdfs generated by the limit model of FDA, if a pdf in the sequence is sufficiently close to the dpdf at the strict global maximal solution, then the sequence will converge to this dpdf. A small perturbations in a pdf in the sequence does not affect the convergence.

The following theorem gives a sufficient condition under which a pdf is not stable fixed point of Φ_2.

Theorem 7: Let $y = (y_1, \cdots, y_n) \in \Omega$. If there exists $z = (y_1, \cdots, y_r, z_{r+1}, z_{r+2}, y_{r+3}, \cdots, y_n) \in \Omega$ such that $f(z) > f(y)$, then $\mathrm{dep}(x, y)$ is not a stable fixed point of the heuristic function Φ_2.

Proof: It suffices to show that there exists a $\delta_0 > 0$ such that for any given $1 > \varepsilon > 0$, there is a $p(x, 0) \in PF(2)$ with $d(p(*, 0), \mathrm{dep}(*, y)) \leq \varepsilon$ and $\lim_{t\to\infty} d(p(*, t), \mathrm{dep}(*, y)) \geq \delta_0$, where $p(x, t) = \Phi_2^t(p(x, 0))$.

Let

$$p(x, 0) = \begin{cases} 1 - \frac{\varepsilon}{2}, & x = y \\ \frac{\varepsilon}{2}, & x = z \\ 0, & \text{otherwise} \end{cases}.$$

Then, $p(x,0) \in PF(2)$ and

$$d\left(p(*,0), \text{dep}(*,y)\right) = \sum_{z\in\Omega} |p(z,0) - \text{dep}(z,y)| = \varepsilon.$$

Analogically to the proof of Theorem 3,

$$p^s(x,0) = \begin{cases} \left(1-\frac{\varepsilon}{2}\right)^2, & x = y \\ 1-\left(1-\frac{\varepsilon}{2}\right)^2, & x = z \\ 0, & \text{otherwise} \end{cases}.$$

It is easy to show that

$$p(x,1) = p^s(x,0).$$

Thus

$$d\left(p(*,1), \text{dep}(*,y)\right) = 2\left[1-\left(1-\frac{\varepsilon}{2}\right)^2\right].$$

Repeating the above procedure will lead to

$$d\left(p(*,t), \text{dep}(*,y)\right) = 2\left[1-\left(1-\frac{\varepsilon}{2}\right)^{2^t}\right].$$

It follows that

$$\lim_{t\to\infty} d\left(p(*,t), \text{dep}(*,y)\right) = 2.$$

■

Remark 2: As shown in the following example, some strict local maximal solutions may satisfy the condition in Theorem 7, then the dpdfs at these local maximal solutions are not stable in Φ_2. In comparison, Theorem 2 claims that a dpdf at any strict local maximal solution are stable in Φ_1. Therefore, with the same initial pdf, the limit model of UMDA may get trapped at the dpdf at a local maximal solution while the limit model of FDA will not. The following is an illustrative example.

x_1	x_2	x_3	$f(x_1,x_2,x_3)$
0	0	0	3
0	0	1	1
0	1	0	2
1	0	0	0
0	1	1	4
1	0	1	6
1	1	0	7
1	1	1	10

For this objective function, (0,0,0) is a strict local maximal solution. By Theorem 2, (0,0,0) is an asymptotically stable fixed point of the limit model of UMDA. But $f(0,1,1) > f(0,0,0)$. Then it follows from Theorem 7 that (0,0,0) is not stable fixed point of the limit model of FDA. That means that the limit model of UMDA may become stuck at the dpdf at this local maximal solution but the limit model of FDA will not for some initial pdfs.

C. Stability of Heuristic Functions for Additively Decomposable Function

Additively decomposable functions [30] have been used frequently as benchmark problems in evolutionary computation. In this subsection, we consider FDA for the maximization of the following additively decomposable function:

$$f(x) = \sum_{i=1}^{n-1} f_i(x_i, x_{i+1}). \tag{23}$$

As mentioned in Section III, an FDA chooses a conditional dependence graph for $p(x,t)$ based on the structure of the variable interaction in the objective function. The structure of variable interaction in this function is a chain in which x_i is the ith node. Therefore, in FDA for maximizing this function, the conditional dependence graph of $p(x,t)$ should also be this chain. In other words, $p(x,t) \in PF(2)$. This means that FDA considered in this subsection is the same as one introduced in Section III-B.

In the following, we always assume that $f(x) \neq f(y)$ for any $x \neq y$.[6] Under this assumption, $f(x)$ has the unique global maximal solution $y^* = (y_1^*, \cdots y_n^*)$ and $p^s(x,t)$ can be written as

$$p^s(x,t) = [p(x,t)]^2 + 2p(x,t)\sum_{f(z)<f(x)} p(z,t). \tag{24}$$

We define

$$g_k(x_k, x_{k+1:n}) = \sum_{i=k}^{n-1} f_i(x_i, x_{i+1}) \tag{25}$$

where $1 \le k < n$, $x_{k+1:n} = (x_{k+1}, x_{k+2}, \cdots, x_n)$

Lemma 8:

1) For a given $c \in \{0,1\}$, the solution to the optimization problem

$$\underset{x_{k+1:n}\in\{0,1\}^{n-k}}{\text{Maxmize}}\; g_k(c, x_{k+1:n}) \tag{26}$$

is unique; we denote it by $\pi_k(c)$.

2) Let

$$\pi_k(c) = (u_{k+1}, \cdots, u_n) = \arg\max_{x_{k+1:n}\in\{0,1\}^{n-k}} g_k(c, x_{k+1:n}). \tag{27}$$

Then for any $k < s < n$, we have

$$\pi_s(u_s) = (u_{s+1}, \cdots, u_n) = \arg\max_{x_{s+1:n}\in\{0,1\}^{n-s}} g_s(u_s, x_{s+1:n}). \tag{28}$$

Proof: See the Appendix. ■

Theorem 9: Let $z = (z_1, z_2, \cdots, z_n) \in \{0,1\}^n$, then $\text{dep}(x,z)$ is not a stable fixed point of Φ_2 if $z \neq y^*$.

Proof: We are in one of the following two cases:

Case 1: there exists $1 \le s \le n-1$ such that

$$\pi_s(z_s) \neq (z_{s+1}, \cdots, z_n). \tag{29}$$

Let r be the largest index s such that (29) holds. Denote $\pi_r(z_r) = (u_{r+1}, \cdots, u_n)$. We can prove by contradiction that $u_{r+1} = 1 - z_{r+1}$. (In fact, assume $u_{r+1} = z_{r+1}$. By Lemma 5 we have $\pi_{r+1}(u_{r+1}) = (u_{r+2}, \cdots, u_n)$. Since r is the largest index s such that (29) holds, $\pi_{r+1}(z_{r+1}) = (z_{r+2}, \cdots, z_n)$. Therefore, $(u_{r+2}, \cdots, u_n) = (z_{r+2}, \cdots, z_n)$. Consequently, $(z_{r+1}, \cdots, z_n) = (u_{r+1}, \cdots, u_n) = \pi_r(z_r)$, which gives a

[6]This assumption is not always satisfied in real-world problems. However, it makes the mathematical analysis easier and the main aspects of the dynamic system can still be illustrated under this assumption. Some other researchers also use this assumption in studying the behavior of EDA (e.g., [19]).

contradiction.) On the other hand, if there exists $m > r+1$ such that $u_m = z_m$, then $\pi_m(u_m) = \pi_m(z_m)$. Therefore, $(u_{m+1}, \cdots, u_n) = (z_{m+1}, \cdots, z_n)$, i.e., $u_i = z_i$ for all $i > m$.

Let r_0 be the largest index j such that $u_j = 1 - z_j$, then $r+1 \le r_0 \le n$. Therefore $\pi_r(z_r)$ is of the following form:

$$\pi_r(z_r) = (1 - z_{r+1}, \cdots, 1 - z_{r_0}, z_{r_0+1}, \cdots, z_n). \tag{30}$$

Let

$$\begin{aligned}\overline{z} &= (z_1, z_2, \cdots, z_r, \pi_r(z_r)) \\ &= (z_1, z_2, \cdots, z_r, 1 - z_{r+1}, \cdots, 1 - z_{r_0}, z_{r_0+1}, \cdots, z_n).\end{aligned}$$

Then

$$\begin{aligned} f(\overline{z}) &= \sum_{i=1}^{r-1} f_i(z_i, z_{i+1}) + g_r(z_r, \pi_r(z_r)) \\ &> \sum_{i=1}^{r-1} f_i(z_i, z_{i+1}) + g_r(z_r, z_{r+1:n}) \\ &= f(z). \end{aligned}$$

For a given positive $\varepsilon < 1$, let

$$p(x,0) = \begin{cases} 1 - \frac{\varepsilon}{2}, & x = z \\ \frac{\varepsilon}{2}, & x = \overline{z} \\ 0, & \text{otherwise} \end{cases}.$$

Then, $p(x,0) \in PF(2)$, and $d(p(*,0), \text{dep}(*,z)) = \varepsilon$. As shown in Theorem 3

$$p^s(x,0) = \begin{cases} \left(1 - \frac{\varepsilon}{2}\right)^2, & x = z \\ 1 - \left(1 - \frac{\varepsilon}{2}\right)^2, & x = \overline{z} \\ 0, & \text{otherwise} \end{cases}$$

in the limit model of UMDA (21). Noting (30), we can easily prove that

$$p(x,1) = p^s(x,0)$$

in (21). Then

$$d(p(*,1), \text{dep}(*,z)) = 2\left[1 - \left(1 - \frac{\varepsilon}{2}\right)^2\right].$$

Repeating the above procedure will lead to

$$d(p(*,t), \text{dep}(*,z)) = 2\left[1 - \left(1 - \frac{\varepsilon}{2}\right)^{2^t}\right].$$

Therefore

$$\lim_{t\to\infty} d(p(*,t), \text{dep}(*,z)) = 2$$

which implies that $\text{dep}(x,z)$ is not a stable fixed point of Φ_2.

Case 2: there exists no $1 \le s \le n-1$ such that

$$\pi_s(z_s) \ne (z_{s+1}, \cdots, z_n).$$

In this case

$$\pi_1(z_1) = (z_2, z_3, \cdots, z_n).$$

Noting the fact that $z \ne y^*$, we have

$$z_1 = 1 - y_1^*.$$

Using arguments similar to those in the proof of Case 1, we can prove by Lemma 5 that z is of the following form:

$$z = \left(1 - y_1^*, 1 - y_2^*, \cdots, 1 - y_{r_1}^*, y_{r_1+1}^*, \cdots, y_n^*\right)$$

where $1 \le r_1 \le n$. Obviously, $f(z) < f(y^*)$. For a given positive $\varepsilon < 1$, let

$$p(x,0) = \begin{cases} 1 - \frac{\varepsilon}{2}, & x = z \\ \frac{\varepsilon}{2}, & x = y^* \\ 0, & \text{otherwise} \end{cases}.$$

As in the proof of Case 1, we can show that

$$\lim_{t\to\infty} d(p(*,t), \text{dep}(*,z)) = 2.$$

Therefore, $\text{dep}(x,z)$ is not a stable fixed point of Φ_2. ∎

Theorem 10: $\overline{p}(x) \in PF(2)$ is not a fixed point of Φ_2 if $\overline{p}(x)$ is not a dpdf.

Proof: Let $\overline{p}(x) \in PF(2)$. If $\overline{p}(x)$ is not a dpdf, then there exists $1 \le i \le n-1$ with

$$0 < \overline{p}(x_i = j, x_{i+1} = 0), \overline{p}(x_i = j, x_{i+1} = 1) < 1 \tag{31}$$

for $j = 0$ or 1 (otherwise, it is from (4) that $\overline{p}(x)$ will be a dpdf). Let r be the largest integer i in $[1, n-1]$ such that (31) holds. Without loss of generality, we assume

$$0 < \overline{p}(x_r = 0, x_{r+1} = 0), \overline{p}(x_r = 0, x_{i+1} = 1).$$

Then, there exist $u = (u_{r+2}, \cdots u_n) \in \{0,1\}^{n-r-1}$ and $v = (v_{r+2}, \cdots v_n) \in \{0,1\}^{n-r-1}$ such that

$$\overline{p}\left((x_r, x_{r+1}, \cdots, x_n) = (0,0,u)\right) = \overline{p}(x_r = 0, x_{r+1} = 0)$$

and

$$\overline{p}\left((x_r, x_{r+1}, \cdots, x_n) = (0,1,v)\right) = \overline{p}(x_r = 0, x_{r+1} = 1)$$

[otherwise, it will contradict to the fact that r is the largest integer such that (31) holds]. Therefore

$$\overline{p}\left((x_{r+2}, \cdots, x_n) = u | x_{r+1} = 0\right) = 1 \tag{32}$$

and

$$\overline{p}\left((x_{r+2}, \cdots, x_n) = v | x_{r+1} = 1\right) = 1. \tag{33}$$

Since $\overline{p}(x) \in PF(2)$, it is necessary that $v = (v_{r+2}, \cdots v_n)$ is of the following form[7]:

$$v = (1 - u_{r+2}, \cdots 1 - u_{r_1}, u_{r_1+1}, \cdots, u_n). \tag{34}$$

Noting that

$$\begin{aligned}&\overline{p}(x_{r+1} = 0 | x_r = 0) \\ &\quad = \frac{\overline{p}(x_r = 0, x_{r+1} = 0)}{\overline{p}(x_r = 0, x_{r+1} = 0) + \overline{p}(x_r = 0, x_{r+1} = 1)}\end{aligned}$$

and

$$\begin{aligned}&\overline{p}(x_{r+1} = 1 | x_r = 0) \\ &\quad = \frac{\overline{p}(x_r = 0, x_{r+1} = 1)}{\overline{p}(x_r = 0, x_{r+1} = 0) + \overline{p}(x_r = 0, x_{r+1} = 1)}\end{aligned}$$

[7]The form includes the two special cases: 1) $v = (1 - u_{r+2}, \cdots, 1 - u_n)$ and 2) $v = (u_{r+2}, \cdots, u_n)$.

we have

$$0 < \overline{p}(x_{r+1} = 0|x_r = 0), \overline{p}(x_{r+1} = 1|x_r = 0) < 1.$$

Denote

$$\beta = \overline{p}(x_{r+1} = 0|x_r = 0). \tag{35}$$

Then

$$\overline{p}(x_{r+1} = 1|x_r = 0) = 1 - \beta.$$

Without loss of generality, we assume that

$$\begin{aligned} &g_r\left(x_r = 0, (x_{r+1}, \cdots, x_n) = (0, u)\right) \\ &\quad > g_r(x_r = 0, (x_{r+1}, \cdots, x_n) = (1, v)). \end{aligned} \tag{36}$$

There are $L = 2^{r-1}$ points $w^1, \cdots, w^L$ in $\{0,1\}^{r-1}$. Denote

$$\alpha_i = \overline{p}\left((x_1, \cdots, x_{r-1}, x_r) = (w^i, 0)\right) \tag{37}$$

for $1 \leq i \leq L$. Define

$$h\left(x_{1:(r-1)}\right) = \sum_{k=1}^{r-2} f_k(x_k, x_{k+1}) + f_{r-1}(x_{r-1}, 0).$$

Since $f(x) \neq f(y)$ for any $x \neq y$, we have

$$h(w^i) \neq h(w^j)$$

for any $i \neq j$. Therefore, we can assume that

$$h(w^1) > h(w^2) > \cdots > h(w^L). \tag{38}$$

$\Omega = \{0,1\}^n$ can be divided into the following three disjoint sets:

$$\begin{aligned} S_1 &= \{x = (x_1, \cdots, x_n) | x_r = 0, x_{r+1} = 0\} \\ S_2 &= \{x = (x_1, \cdots, x_n) | x_r = 0, x_{r+1} = 1\} \end{aligned}$$

and

$$S_3 = \{x = (x_1, \cdots, x_n) | x_r = 1\}.$$

Let

$$y^i = (w^i, 0, 0, u) \quad (1 \leq i \leq L)$$

and

$$z^i = (w^i, 0, 1, v) \quad (1 \leq i \leq L).$$

Obviously, $y^i \in S_1$ and $z^i \in S_2$ for $i = 1, 2, \cdots, L$. From (32) and (33), it follows that:

$$\overline{p}(x) = 0$$

whenever $x \in S_1 \backslash \{y^1, \cdots y^L\}$ or $x \in S_2 \backslash \{z^1, \cdots z^L\}$. There are $J = 2^{n-1}$ points $x^1, \cdots, x^J$ in S_3. By (36) and (38), we have

$$f(y^i) > f(z^i)\ (1 \leq i \leq L) \tag{39}$$

$$f(y^1) > f(y^2) > \cdots > f(y^L) \tag{40}$$

and

$$f(z^1) > f(z^2) > \cdots > f(z^n). \tag{41}$$

Since $\overline{p}(x) \in PF(2)$, we have

$$\begin{aligned} \overline{p}(y^i) = &\underbrace{\overline{p}\left((x_1, \cdots, x_{r-1}, x_r) = (w^i, 0)\right)}_{=\alpha_i \text{ (by (37))}} \underbrace{\overline{p}(x_{r+1} = 0|x_r = 0)}_{=\beta \text{ (by (35))}} \\ &\times \underbrace{\overline{p}\left((x_{r+2}, \cdots x_n) = u | x_{r+1} = 0\right)}_{=1 \text{ (by (32))}} = \alpha_i \beta. \end{aligned}$$

Similarly

$$\overline{p}(z^i) = \alpha_i (1 - \beta).$$

Let $\overline{q}(x) = \Phi_2(\overline{p}(x))$. We have the following claims.

Claim 1: r is still the largest integer i in $[1, n-1]$ with

$$0 < \overline{q}(x_i = j, x_{i+1} = 0), \overline{q}(x_i = j, x_{i+1} = 1)$$

for $j = 0$ or 1, and

$$0 < \overline{q}(x_r = 0, x_{r+1} = 0), \overline{q}(x_r = 0, x_{i+1} = 1).$$

The proof of Claim 1 is trivial based on (34).

Claim 2:

$$\begin{aligned} &\overline{q}(x_{r+1} = 0|x_r = 0) - \overline{p}(x_{r+1} = 0|x_r = 0) \\ &\qquad \geq \frac{1}{2L} \overline{p}(x_r = 0, x_{r+1} = 1) \overline{p}(x_{r+1} = 0|x_r = 0). \end{aligned}$$

The proof of Claim 2 can be found in the Appendix.

Now, let $p(x, t+1) = \Phi_2(p(x,t))$ and $p(x, 0) = \overline{p}(x)$. Based on the Claims 1 and 2, we have

$$\begin{aligned} &p(x_{r+1} = 0|x_r = 0, t+1) - p(x_{r+1} = 0|x_r = 0, t) \\ &\quad \geq \frac{1}{2L} p(x_r = 0, x_{r+1} = 1, t) p(x_r = 0|x_{r+1} = 0, t). \end{aligned} \tag{42}$$

Therefore, $\{p(x_{r+1} = 0|x_r = 0, t)\}_{t=0,1,2,\cdots,}$ is increasing. Since

$$0 < p(x_{r+1} = 0|x_r = 0, t) \leq 1 \tag{43}$$

we have

$$\lim_{t \to \infty} [p(x_{r+1}=0|x_r=0, t+1) - p(x_{r+1}=0|x_r=0, t)] = 0. \tag{44}$$

By (42)–(44)

$$\lim_{t \to \infty} p(x_r = 0, x_{r+1} = 1, t) = 0$$

which implies $\overline{p}(x)$ cannot be a fixed point of Φ_2.

The main result in this section is immediately derived from Theorems 6, 9, and 10. ■

Theorem 11: $\text{dep}(x, y^*)$ is the unique asymptotically stable fixed point of Φ_2.

Remark 3: This theorem implies that the limit model of FDA is able to find the global optimal solution of the additively decomposable function. We should point out that there is no guarantee that the limit model of UMDA will locate the global optimal solution of the additively decomposable function for any initial pdf $p(x, 0)$. The following is an example: Let

x_1	x_2	$f_1(x_1, x_2)$	$f_2(x_1, x_2)$
0	0	10	10
0	1	1	4
1	0	1	3
1	1	9	9

Then

x_1	x_2	x_3	$f(x_1,x_2,x_3) = f_1(x_1,x_2) + f_2(x_2,x_3)$
0	0	0	20
0	0	1	14
1	0	0	11
1	0	1	5
0	1	0	4
0	1	1	10
1	1	0	12
1	1	1	18

In this function, $f(x) \neq f(y)$ for $x \neq y$. (0,0,0) is the unique global maximal solution. (1,1,1) is a local maximal solution. Therefore, the limit model of UMDA may become stuck at the local maximal solution (1,1,1) for some initial pdfs.

Remark 4: Theorem 9 is stronger than Theorem 7 in the case of additively decomposable functions. For example, Theorem 7 cannot be used to study the stability of (1,1,1) in Φ_2 in the above example.

Remark 5: Consider the limit model of FDA: $p(x,0) \in PF(2)$ and $p(x,t+1) = \Phi_2(p(x,t))(t = 0,1,2,\cdots)$. If $p(x,0)$ is not a degenerate pdf, let r be the largest integer i in $[1, n-1]$ such that

$$0 < p(x_i = j, x_{i+1} = 0, t = 0), \overline{p}(x_i = j, x_{i+1} = 1, t = 0)$$

for $j = 0$ or 1. Let $q(x)$ be a limit point of $\{p(x,t)\}_{t=0,1,\ldots}$. We can easily prove that $q(x) \in PF(2)$. Let $\overline{r}$ be the largest integer i in $[1, n-1]$ such that

$$0 < q(x_i = j, x_{i+1} = 0), q(x_i = j, x_{i+1} = 1).$$

Therefore, by the proof of Theorem 10, $\overline{r} < r$, which means that more conditional probabilities $q(x_{i+1}|x_i)$ in $q(x)$ are degenerate than in $p(x,t)$.

V. Conclusion

Many experimental results reported in the literature suggest that using higher order statistics will improve the performance of EDAs. In this paper, we have studied the dynamic behaviors of the limit models of UMDA and FDA with tournament selection. Our results suggest that FDA has a better chance of converging to the global optimal solutions than UMDA for some initial pdfs. In particular, the limit model of FDA should be able to find the global optimal solution of an additively decomposable function but UMDA may become stuck at any local optimal solution for some initial pdfs. We should point out that all the results in this paper are on the local dynamical behavior of the limit models of these algorithms. Possible future work may include the following.

- Analysis of global behaviors of these limit models. In practice, the initial pdf $p(x,0)$ can be set as the uniform distribution (i.e., $p(x,0) = (1/2^n)$ for any $x \in \Omega$) or any other distribution on Ω. Study of global convergence and stability of a pdf sequence generated by these limit models with arbitrary initial pdf will be very important. As pointed out by a reviewer, the limit model of UMDA will converge to the global maximal solution of $f(x)$ in Remark 3 if the initial pdf is the uniform distribution. The attraction basin of the dpdf at the global maximal solution for these limit models should be investigated for some test problems.
- Study of the approximation error between these limit models and the algorithms with finite population.
- Investigation of EDAs with finite population. It is a challenging task. Recently, some results on the computational time complexity of EAs with finite populations have been obtained [24]–[27]. It is worthwhile studying if these results can be to extended to EDAs for some typical instances.
- Study of other EDAs with model selection at each generation such as MIMIC and BOA.

Appendix

Proof of Lemma 5[8]: By Bernoulli inequality [34], $(1-\tau)^n \geq 1 - n\tau$ for any $\tau \in [0,1]$, we obtain

$$(1-\theta^2)^{n-1} \geq 1 - (n-1)\theta^2$$

for any $-1 \leq \theta \leq 1$. Therefore

$$1 - (1-\theta^2)^{n-1} \leq (n-1)\theta^2 < \frac{\theta}{4}$$

whenever $0 < \theta < (1/4(n-1))$. ■

Proof of 1) in Lemma 8: Assume that $v = (v_{k+1}, v_{k+2}, \cdots, v_n)$ and $w = (w_{k+1}, w_{k+2}, \cdots, w_n)$ are the solutions to (26). Therefore

$$g_k(c,v) = g_k(c,u).$$

Let $\overline{u} = (1,1,\cdots 1, c, u_{k+1}, u_{k+2}, \cdots, u_n)$ and $\overline{v} = (1,1,\cdots 1, c, v_{k+1}, v_{k+2}, \cdots, v_n)$. Since

$$f(x) = \sum_{i=1}^{k-1} f_i(x_i, x_{i+1}) + g_k(x_k, x_{k+1:n})$$

we have

$$f(\overline{u}) = f(\overline{v}).$$

Since $f(x) \neq f(y)$ for any $x \neq y$, we have $\overline{u} = \overline{v}$. Consequently, $u = v$, which implies that (26) has a unique solution.

Proof of 2) in Lemma 8: We prove it by contradiction. Assume that 2) is not true. Then, there exists an integer r in (k,n) such that

$$(u_{r+1}, \cdots, u_n) \neq \pi_r(u_r).$$

Let

$$\pi_r(u_r) = (w_{r+1}, \cdots, w_n).$$

We have

$$g_r(u_r, w_{r+1}, \cdots, w_n) > g_r(u_r, u_{r+1}, \cdots, u_n).$$

[8]This proof was provided by the anonymous Associate Editor of this paper.

Therefore

$$\begin{aligned}g_k(c,u_{k+1},\cdots,u_n)&=f_k(c,u_{k+1})+\sum_{i=k+1}^{r-1}f_i(u_i,u_{i+1})\\&\quad+g_r(u_r,u_{r+1},\cdots,u_n)\\&<f_k(c,u_{k+1})+\sum_{i=k+1}^{r-1}f_i(u_i,u_{i+1})\\&\quad+g_r(u_r,w_{r+1},\cdots,w_n)\\&=g_k(c,u_{k+1},\cdots,u_r,w_{r+1},\cdots,w_n),\end{aligned}$$

which contradicts the fact that $\pi_k(c)=(u_{k+1},\cdots,u_n)$.

Proof of Claim 2 in Theorem 10:

$$\begin{aligned}&\overline{p}^s(x_r=0,x_{r+1}=0)\\&=\sum_{x\in S_1}\overline{p}^s(x)\\&=\sum_{x\in S_1}\left\{[\overline{p}(x)]^2+2\overline{p}(x)\sum_{f(z)<f(x)}\overline{p}(z)\right\}\\&=\sum_{x\in S_1}[\overline{p}(x)]^2\\&\quad+2\sum_{x\in S_1}\left\{\overline{p}(x)\left[\sum_{f(z)<f(x),z\in S_1}\overline{p}(z)+\sum_{f(z)<f(x),z\in S_2}\overline{p}(z)+\sum_{f(z)<f(x),z\in S_3}\overline{p}(z)\right]\right\}\\&=\left[\sum_{x\in S_1}\overline{p}(x)\right]^2+2\sum_{x\in S_1,z\in S_2,f(z)<f(x)}\overline{p}(x)\overline{p}(z)\\&\quad+2\sum_{x\in S_1,z\in S_3,f(z)<f(x)}\overline{p}(x)\overline{p}(z)\\&\geq\left[\sum_{i=1}^{L}\overline{p}(y^i)\right]^2+2\sum_{i=1}^{L}\overline{p}(y^i)\sum_{j=i}^{L}\overline{p}(z^j)\\&\quad+2\sum_{x\in S_1,z\in S_3,f(z)<f(x)}\overline{p}(x)\overline{p}(z)\\&=\beta^2\left[\sum_{i=1}^{L}\alpha_i\right]^2+2\beta(1-\beta)\sum_{i=1}^{L}\alpha_i\sum_{j=i}^{L}\alpha_j\\&\quad+2\sum_{x\in S_1,z\in S_3,f(z)<f(x)}\overline{p}(x)\overline{p}(z).\end{aligned}$$

Therefore

$$\overline{p}^s(x_r=0,x_{r+1}=0)\geq\beta\left[\sum_{i=1}^{L}\alpha_i\right]^2+\beta(1-\beta)\sum_{i=1}^{L}\alpha_i^2+2I_1$$

where $I_1=\sum_{x\in S_1,z\in S_3,f(z)<f(x)}\overline{p}(x)\overline{p}(z)$. Similarly, we have

$$\overline{p}^s(x_r=0,x_{r+1}=1)\leq(1-\beta)\left[\sum_{j=1}^{L}\alpha_j\right]^2-\beta(1-\beta)\sum_{i=1}^{L}\alpha_i^2+2I_2$$

where $I_2=\sum_{x\in S_2,z\in S_3,f(z)<f(x)}\overline{p}(x)\overline{p}(z)$.

Therefore

$$\begin{aligned}&\overline{q}(x_{r+1}=0|x_r=0)\\&\quad=\frac{\overline{q}(x_r=0,x_{r+1}=0)}{\overline{q}(x_r=0,x_{r+1}=0)+\overline{q}(x_r=0,x_{r+1}=1)}\\&\quad\underbrace{=}_{\text{by (22)}}\frac{\overline{p}^s(x_r=0,x_{r+1}=0)}{\overline{p}^s(x_r=0,x_{r+1}=0)+\overline{p}^s(x_r=0,x_{r+1}=1)}.\end{aligned}$$

Consequently

$$\begin{aligned}&\overline{q}(x_{r+1}=0|x_r=0)\\&\quad\geq\frac{\beta\left[\sum_{i=1}^{L}\alpha_i\right]^2+\beta(1-\beta)\sum_{i=1}^{L}\alpha_i^2+2I_1}{\left[\sum_{j=1}^{L}\alpha_j\right]^2+2(I_1+I_2)}.\end{aligned}\tag{45}$$

Denoting $\overline{p}(x=x^j)$ by γ_j, we have

$$\begin{aligned}I_1&=\sum_{x\in S_1,z\in S_3,f(z)<f(x)}\overline{p}(x)\overline{p}(z)\\&=\sum_{i=1}^{L}\sum_{f(x^i)<f(y^j)}\overline{p}(x=y^i)\overline{p}(x=x^j)\\&=\sum_{i=1}^{L}\alpha_i\beta\sum_{f(x^i)<f(y^j)}\gamma_j=\sum_{i=1}^{L}\alpha_i\beta A_i\end{aligned}\tag{46}$$

where $A_i=\sum_{f(x^i)<f(y^j)}\gamma_j$ and

$$\begin{aligned}I_2&=\sum_{x\in S_2,z\in S_3,f(z)<f(x)}\overline{p}(x)\overline{p}(z)\\&=\sum_{i=1}^{L}\sum_{f(x^i)<f(z^j)}\overline{p}(x=z^i)\overline{p}(x=x^j)\\&=\sum_{i=1}^{L}\alpha_i(1-\beta)\sum_{f(x^j)<f(z^i)}\gamma_j\\&=\sum_{i=1}^{L}\alpha_i(1-\beta)B_i\end{aligned}\tag{47}$$

where $B_i=\sum_{f(x^j)<f(z^i)}\gamma_j$. It follows from (39)–(41) that

$$A_i\geq B_i.$$

Substituting (46) and (47) into (45) gives

$$\begin{aligned}&\overline{q}(x_{r+1}=0|x_r=0)\\&\quad\geq\frac{\beta\left[\sum_{i=1}^{L}\alpha_i\right]^2+\beta(1-\beta)\sum_{i=1}^{L}\alpha_i^2+2\sum_{i=1}^{L}\alpha_i\beta A_i}{\left[\sum_{j=1}^{L}\alpha_j\right]^2+2\left\{\sum_{i=1}^{L}\alpha_i(1-\beta)B_i+\sum_{i=1}^{L}\alpha_i\beta A_i\right\}}\\&\quad\underbrace{\geq}_{A_i\geq B_i}\frac{\beta\left[\sum_{i=1}^{L}\alpha_i\right]^2+\beta(1-\beta)\sum_{i=1}^{L}\alpha_i^2+2\beta\sum_{i=1}^{L}\alpha_iA_i}{\left[\sum_{j=1}^{L}\alpha_j\right]^2+2\sum_{i=1}^{L}\alpha_iA_i}\\&\quad=\beta+\frac{(1-\beta)\beta\sum_{j=1}^{L}\alpha_j^2}{\left[\sum_{j=1}^{L}\alpha_j\right]^2+2\sum_{i=1}^{L}\alpha_iA_i}\\&\quad\geq\beta+\frac{(1-\beta)\beta\sum_{j=1}^{L}\alpha_j^2}{\left[\sum_{j=1}^{L}\alpha_j\right]^2+M}\end{aligned}$$

where $M = 2\sum_{i=1}^{L}\alpha_i\sum_{j=1}^{J}\gamma_j = 2\sum_{i=1}^{L}\alpha_i(1-\sum_{i=1}^{L}\alpha_i)$. (In fact, $\sum_{i=1}^{L}\alpha_i = \overline{p}(x_r=0)$, and $\sum_{j=1}^{J}\gamma_j = \overline{p}(x_r=1)$.) Therefore

$$\begin{aligned}\overline{q}(x_{r+1}&=0|x_r=0)-\beta\\ &\geq \frac{(1-\beta)\beta\sum_{j=1}^{L}\alpha_j^2}{\left[\sum_{j=1}^{L}\alpha_j\right]^2+M}\\ &=\frac{(1-\beta)\beta\sum_{j=1}^{L}\alpha_j^2}{\left[\sum_{j=1}^{L}\alpha_j\right]^2+2\sum_{i=1}^{L}\alpha_i\left(1-\sum_{i=1}^{L}\alpha_i\right)}\\ &=\frac{(1-\beta)\beta\sum_{j=1}^{L}\alpha_j^2}{2\sum_{i=1}^{L}\alpha_i-\left[\sum_{j=1}^{L}\alpha_j\right]^2}\\ &=(1-\beta)\beta\,\underbrace{\frac{1}{\left[2-\sum_{i=1}^{L}\alpha_i\right]}}_{\geq\frac{1}{2}}\sum_{j=1}^{L}\underbrace{\left[\frac{\alpha_j}{\sum_{i=1}^{L}\alpha_i}\right]^2}_{\geq\frac{1}{L}}\sum_{i=1}^{L}\alpha_i\\ &\geq\frac{\beta}{2L}\sum_{i=1}^{L}\alpha_i(1-\beta)=\frac{1}{2L}\beta\overline{p}(x_r=0,x_{r+1}=1).\end{aligned}$$

Noting that $\beta = \overline{p}(x_{r+1}=0|x_r=0)$, we obtain

$$\begin{aligned}\overline{q}(x_{r+1}=0|x_r=0)&-\overline{p}(x_{r+1}=0|x_r=0)\\ &\geq\frac{1}{2L}\overline{p}(x_r=0,x_{r+1}=1)\overline{p}(x_{r+1}=0|x_r=0).\end{aligned}$$

ACKNOWLEDGMENT

The author would like to thank X. Yao, the anonymous reviewers, and the Associate Editor for their constructive comments and suggestions. He also thanks H. Mühlenbein, E. Tsang, J. Ford, and R. Poli for helpful discussions on this topic.

REFERENCES

[1] H. Mühlenbein and G. Paass, "From recombination of genes to the estimation of distribution part 1, binary parameter," in *Lecture Notes in Computer Science*, H.-M. Voigt *et al.*, Eds. Berlin, Germany: Springer-Verlag, 1996, vol. 1141, Proc. 4th Int. Conf. Parallel Problem Solving From Nature—PPSN IV, pp. 178–187.

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

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

[4] S. De Bonet, J. S. Isbell, and P. Viola, "MIMIC: finding optima by estimating probability densities," in *Advances in Neural Information Processing Systems*, C. Mozer *et al.*, Eds. Cambridge, MA: MIT Press, 1997, vol. 9, pp. 424–431.

[5] M. Pelikan, D. E. Goldberg, and E. Cantu-Paz, "BOA: the Bayesian Optimization Algorithm," in *Proc. Genetic Evolutionary Computation Conf.*, W. Banzhaf *et al.*, Eds. Orlando, FL, 1999, pp. 525–532.

[6] ——, "Linkage problem, distribution estimation, and Bayesian networks," *Evol. Comput.*, vol. 8, pp. 311–341, 2002.

[7] B. T. Zhang, "A Bayesian framework for evolutionary computation," in *Proc. 1999 Congress Evolutionary Computation*, P. J. Angeline *et al.*, Eds., Washington, DC, 1999, pp. 722–228.

[8] V. V. Miagkikh and W. F. Punch, "Global search in combinatorial optimization using reinforcement learning algorithms," in *Proc. 1999 Congress Evolutionary Computation*, P. J. Angeline *et al.*, Eds. Washington, DC, 1999, pp. 189–196.

[9] G. Harik, F. Lobo, and D. E. Goldberg, "The compact genetic algorithm," *IEEE Trans. Evol. Comput.*, vol. 3, pp. 287–297, Nov. 1999.

[10] R. Etxeberria and P. Larrañaga, "Global optimization using Bayesian networks," in *Proc. 2nd Symp. Artificial Intelligence*, A. O. Rodriguez *et al.*, Eds., Cuba, 1999, pp. 332–339.

[11] P. A. N. Bosman and D. Thierens, "Expanding from discrete to continuous EDAs: the IEDA," in *Proc. 6th International Conf. Parallel Problem Solving From Nature—PPSN VI*, vol. 1917, Lecture Notes in Computer Science, M. Schoenauer *et al.*, Eds., 2000, pp. 767–776.

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

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

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

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

[16] M. Pelikan, K. Sastry, and D. E. Goldberg, "Evolutionary Algorithms + Graphical Models = Scalable Black-Box Optimization," Illinois Genetic Algorithms Laboratory, University of Illinois at Urbana–Champaign, IlliGAL Rep. 2 001 029, 2001.

[17] D. Thierens, "Analysis and design of genetic algorithms," Ph.D. dissertation, Katholieke Universiteit Leuven, Leuven, Belgium, 1995.

[18] P. A. N. Bosman and D. Thierens, "Linkage information processing in distribution estimation algorithms," in *Proc. 1999 Genetic Evolutionary Computation Conf.*, W. Banzhaf *et al.*, Eds., San Francisco, CA, 1999, pp. 60–67.

[19] 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.

[20] H. Mühlenbein and T. Mahnig, "Application of Wright's equation in evolutionary computation," in *Advances in Fuzzy Systems and Evolutionary Optimization*, N. Mastorakis, Ed. Singapore: World Scientific, 2001, pp. 265–270.

[21] A. Berny, "Selection and reinforcement learning for combinatorial optimization," in *Lecture Notes in Computer Science*, M. Schoenauer *et al.*, Eds. Berlin, Germany: Springer-Verlag, 2000, vol. 1917, Proc. Parallel Problem Solving from Nature, PPSN VI, pp. 601–610.

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

[23] J. Suzuki, "A Markov chain analysis on simple genetic algorithm," *IEEE Trans. Syst., Man, Cybern.*, vol. 25, pp. 655–659, Apr. 1995.

[24] J. He and X. Yao, "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, 2002.

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

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

[27] T. Jansen and L. Wegener, "The analysis of evolutionary algorithms—a proof that crossover really can help," *Algorithmica*, vol. 33, pp. 47–66, 2002.

[28] Q. Zhang and H. Mühlenbein, "On the convergence of a class of estimation of distribution algorithms, conditionally," *IEEE Trans. Evol. Compur.*, 2003, to be published.

[29] M. Zlochin *et al.*, "Model-Based Search for Combinatorial Optimization," IRIDIA, Universite Libre de Bruxelles, Tech. Rep. TR/IRIDIA/2001-15, 2001.

[30] R. K. Thompson and A. H. Wright, "Additively Decomposable Fitness Functions, Tech. Rep.," Dept. Comput. Sci., Univ. Montana, Missoula, 1997.

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

[32] A. H. Wright and J. E. Rowe, "Continuous dynamical sytem models of steady-state genetic algorithms," in *Foundations of Genetic Algorithms 6*, W. Martin *et al.*, Eds. San Mateo, CA: Morgan Kaufmann, 2001, pp. 209–226.

[33] J. Whittaker, *Graphical Models in Applied Multivariate Statistics*. New York: Wiley, 1991.

[34] Y. Chow and H. Teicher, *Probability Theory: Independence, Interchangability, Martingales*, 3rd ed. New York: Springer-Verlag, 1997.

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

Qingfu Zhang (M'01) received the B.Sc. degree in mathematics from Shanxi University, Shanxi, China, in 1984, the M.Sc. 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 in 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.