Nat Comput (2016) 15:335–354
DOI 10.1007/s11047-015-9499-0

On the model updating operators in univariate estimation of distribution algorithms

Andrey G. Bronevich[1] · **José Valente de Oliveira**[2]

Published online: 9 May 2015

Abstract The role of the selection operation—that stochastically discriminate between individuals based on their merit—on the updating of the probability model in univariate estimation of distribution algorithms is investigated. Necessary conditions for an operator to model selection in such a way that it can be used directly for updating the probability model are postulated. A family of such operators that generalize current model updating mechanisms is proposed. A thorough theoretical analysis of these operators is presented, including a study on operator equivalence. A comprehensive set of examples is provided aiming at illustrating key concepts, main results, and their relevance.

Keywords Compact genetic algorithm · Estimation of distribution algorithms · Selection operator · Theoretical analysis

1 Introduction

Estimation of distribution algorithms (EDAs) are playing a significant role as optimization tools (Larrañaga and Lozano 2001; Hauschild and Pelikan 2011; Ceberio et al. 2012). EDAs combine machine learning and evolutionary computation techniques as they are stochastic population based optimizers whose central idea is to estimate a (probabilistic) model of the population from which new individuals are sampled, evaluated, and used to update the model; the whole process being repeated until convergence. Individuals encode a candidate solution of the optimization problem. In a sense, EDAs evolve probability distributions aiming at converging them to one in which an optimal solution can be sampled from with high probability.

In recent years a considerable amount of research has been devoted to the development and applications of EDAs, see e.g., Larrañaga and Lozano (2001), Hauschild and Pelikan (2011) and Ceberio et al. (2012). Depending on the probabilistic model used, EDAs can be classified as univariate, bivariate or multivariate. In the former case it is assumed that variables of a solution are independent from each other. In this category the following algorithms are worth noting. The population based incremental learning (PBIL) algorithm (Baluja 1994); the univariate marginal distribution algorithm (UMDA) (Mühlenbein 1997) where the entire population is used for estimating probabilities, and a more parsimonious algorithm known as the compact genetic algorithm (cGA) (Harik et al. 1999), where in each step only two candidate solutions are sampled from the model and undergo a tournament-like procedure whose result is used to update the probability distribution.

Bivariate EDAs are characterized by a 2-variable modeling of interaction between variables. Bivariate models include the chain model (MMIC (De Bonet et al. 1997)), the tree (COMIT (Baluja and Davies 1997)) and forest model (BMDA (Pelikan and Mühlenbein 1999)).

In multivariate EDAs various models can be used for estimating the probability distribution of individuals. The factorized distribution algorithm (Mühlenbein et al. 1999;

✉ José Valente de Oliveira
jvo@ualg.pt

Andrey G. Bronevich
brone@mail.ru

[1] National Research University Higher School of Economics, Myasnitskaya 20, 101000 Moscow, Russia

[2] CEOT and FCT, Universidade do Algarve, Campus de Gambelas, 8005-139 Faro, Portugal

Mühlenbein and Mahnig 1998) was the first of such EDAs. An extension of the cGA, named extended compact GA (ECGA), was proposed in Harik (1999). ECGA handles multivariate interactions by searching over all possible disjoint partitions. ECGA generates all possible disjoint partitions and selects the best one in each iteration. The selection is based on information and complexity theory principles, such as the minimum description length principle. Another classical instances of multivariate EDA are the Bayesian optimization algorithm (BOA) (Pelikan et al. 2000; Pelikan and Goldberg 2001) and the estimation of Bayesian network algorithm (EBNA) (Bengoetxea et al. 2002) where the probability distribution is approximated by a Bayesian network. Another interesting approach to multivariate interactions resorts to clustering (Pelikan and Goldberg 2000; Peña et al. 2005; Emmendorfer and Pozo 2009). In this case, the population is represented by a set of clusters with a reduced complexity probability distribution being estimate for each cluster. After clustering it is possible to use probability models of lower order, such as univariate ones.

Relatively less efforts have been paid to the theoretical analysis of EDAs, cf. Hauschild (2011). Some theoretical results for univariate EDAs are available though. The work of Johnson and Shapiro (2002) was the first to stress the importance of model updating operators in EDAs showing that, for some selected problems, these operators have an higher impact on EDA's performance than model complexity. More specifically, it was shown that a univariate EDA equipped with a suitable operator outperforms a more complex bivariate EDA for such problems. In González et al. (2001) the PBIL algorithm is analyzed in the framework of discrete dynamical systems. The analysis reveals that (local) optima are stable fixed points of the correspondent discrete dynamic system; all other search space points being unstable. Convergence follows as a consequence. In Mühlenbein and Mahnig (2002) it is proposed two extensions of UMDA resorting to the Boltzmann distribution. A stochastic analysis of the proposed algorithms (BEDA and FDA) is also provided. Zhang (2004a, b) and Zhang and Mühlenbein (2004) thoroughly studied the global convergence of univariate EDAs, specially in the scope of UMDA and its variants. In Droste (2006) an assumption-free runtime analysis of cGA for linear pseudo-Boolean functions is presented. More recently (Lozada-Chang and Santana 2011) investigate the variations in the UMDA dynamics within a certain class of fitness functions. In a related subject (Echegoyen et al. 2013) established a first taxonomy of equivalent problems in which univariate EDAs have the same behavior.

In this study we are interested in further pursuing this analytical line of work for binary problems and univariate EDAs. In particular, the role of the selection operation on the updating of the probability distribution is thoroughly investigated for the univariate case. Necessary conditions for an operator to model selection in such a way that it can be used directly for updating the probability model are postulated. A family of such operators generalizing current selection effects is proposed. A comprehensive theoretical analysis of these operators is presented, including a study on operator equivalence. Examples are provided all through the text aiming at illustrating key concepts, main results, and their relevance. These include a generalization of the well-known cGA as an illustration of how the proposed theoretical results can be applied to algorithm design. While the necessary conditions must be taken into account while searching for an operator for a given optimization problem, the merit of an operator depends on that problem. To further illustrate this, an operator is derived that outperforms cGA in a parametric fitness function.

The remaining of this paper is organized as follows. In Sect. 2 a general optimization problem is formulated. The necessary notation and terminology is briefly presented resorting to the general computational scheme of an EDA. Section 3 postulates the necessary conditions a model updating operator should have and several examples of such operators are provided and their properties evaluated. In Sect. 4 several fundamental theoretical results are derived for the univariate case. A section on conclusions ends the paper.

2 Background

2.1 Problem formulation

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a basic finite set and let $\mu : 2^X \to \mathbb{R}$ be an arbitrary set function, where 2^X stands for the power set of X, the set of all subsets of X. In this paper, the following optimization problem is considered:

$$B = \arg\max_{A \in 2^X} \mu(A).$$

In plain word, the problem consist in finding a subset of X that maximizes the function μ. This problem has an unique solution under the condition $\mu(A) \neq \mu(B)$ if $A \neq B$ for any $A, B \in 2^X$. Of course, any subset of X can be described by a binary string, i.e., if $A \subseteq X$ then A can be described by the binary string $b(A) = (b_A(x_1), \ldots, b_A(x_n))$, where $b_A(x_i) = 1$ if $x_i \in A$ and $b(x_i) = 0$ otherwise. For example, let $X = \{x_1, x_2, x_3\}$ and $A = \{x_2, x_3\}$, then $b(\emptyset) = (0, 0, 0)$, $b(A) = (0, 1, 1)$, and $b(X) = (1, 1, 1)$.

Example 1 Let $X = \{x_1, x_2\}$ be the set of all possible individuals as given in Table 1. The table also includes an arbitrary set function μ viewed as a fitness function. The maximum of $\mu(A)$ is achieved for $A = \{x_2\}$.

Table 1 The set of possible individuals for $X = \{x_1, x_2\}$ together with an arbitrary set function μ viewed as a fitness function

Id	$b\ (x_1)$	$b\ (x_2$	μ
A_0	0	0	0.2
A_1	0	1	0.6
A_2	1	0	0.0
A_3	1	1	0.4

A remark is in order here. At first, the problem statement may seem trivial. However, for a sufficiently large cardinality of X, as found in many real-world applications, the problem of calculating all the values of the set function μ can be computationally intractable, thus it is reasonable to use the type of algorithms that we are interested in this study, among others metaheuristics.

A binary sequence $(b(x_1), \ldots, b(x_n))$ is called individual and consists of variables $b(x_i)$. At position x_i two values are possible, $b(x_i) = 0$ and $b(x_i) = 1$. Therefore, the above formulation is a sufficient framework for modeling a population of individuals.

2.2 The general computational scheme for EDA

EDAs rely on population modeling. This can be accomplished by resorting to probability theory. For this purpose, we can assign to a given population a probability distribution p over the power set of X, 2^X such that $\sum_{A \in 2^X} p(A) = 1$. Of course, the value $p(A)$ can be interpreted as the frequency of individual A in the population and is related to the frequency of variables in the population. For instance, under the assumption of variable independence, the probability distribution p is uniquely defined by the probabilities $p(x_i) = \Pr(b_A(x_i) = 1); i = 1, \ldots, n$, where n is the length of the individual, and obviously $p(A) = \prod_{x_i \in A} p(x_i) \prod_{x_i \notin A} (1 - p(x_i))$.

We can describe an EDA by a sequence of probability distributions $(p_1, p_2, \ldots, p_i, \ldots)$ where the index i refers to the iteration (or generation) number, and p_i is defined over 2^X. We can think about a general computation scheme for an EDA such as that presented in Algorithm 1. In the step identified by (1) in this algorithm we view the ranking procedure as a recovering from the sample D of the cumulative distribution function (cdf) $F : 2^X \to [0, 1]$ defined by

$$F(B) = \sum_{\mu(A) \leqslant \mu(B)} p(A); \quad B \in 2^X \tag{1}$$

Algorithm 1: A general computational scheme for EDAs (Larrañaga and Lozano, 2001; Ceberio et al., 2012)

Define the fitness function μ or a linear order among individuals (whatever is available);

Set the size of generated samples (progenitors) N ;

Initialize the iteration counter i ($i := 0$) ;

Initialize the probability distribution over individuals $p_{i=0}$;

$D_{i=0} :=$ Sample N individuals from $p_{i=0}$;

output: The solution A for which $p_i(A) \approx 1$

repeat

1 $D_i :=$ Rank individuals in D_i by their fitness or using linear order information (whatever is available);

$D_i^s :=$ Select $M \leq N$ individuals from D_i according to a selection method;

$p_i(x) := p(x|D_i^s)$ Update the probability distribution based on the selected individuals D_i^s;

$D_{i+1} :=$ Sample M individuals (the new population) from p_i;

$i := i + 1$;

until *A convergence condition is met*;

Table 2 shows an example of the both a probability distribution p and the corresponding cdf F under the conditions of Example 1.

3 On model updating operators

Roughly speaking, for maximization problems the selection operation prefers high fitness individuals to low fitness ones. It turns out that the selection operation plays a crucial role in EDAs, cf. Johnson and Shapiro (2002). To illustrate this, consider one of the simplest EDA, the cGA. cGA uses

Table 2 An hypothetical probability distribution p, and the corresponding cdf F, under the conditions of Example 1

Id	$b\ (x_1)$	$b\ (x_2)$	μ	p	F
A_0	0	0	0.2	1/2	2/3
A_1	0	1	0.6	0	1
A_2	1	0	0.0	1/6	1/6
A_3	1	1	0.4	1/3	1

the results of a selection operation for estimating its probability distribution model of the population. cGA adopts the so-called Block selection which is a steady-state binary tournament equivalent selection. Block selection takes two individuals and duplicates the winner while discards the looser. Based on this, cGA uses a simple updating rule for probability distribution, i.e., $p(x_i) = \Pr(b_A(x_i) = 1); i = 1, \ldots, n$, where n is the length of the individual, is updated by the amount $\pm 1/S$, S being the user-defined size of the modeled population.

Notice that by applying this rule, the proportion of variables in the population equal to 1 in a given position will increase by $1/S$ if the winning individual has a 1 in that position; it will decrease by the same amount if the winning individual has a 0 in the same position; for the general case where winner is different from the looser. Curiously enough, it was experimentally shown that when equipped with this updating rule cGA runs with the same convergence rate and reliability as the simple genetic algorithm under uniform crossover (Harik et al. 1999; Harik 1999).

In this section, we generalize the updating of the probability distribution model by defining a pair of model updating operators. Assume that population is ordered by the function μ. Let $\mu(A) > \mu(B)$ then

$$p_{k+1}(x_i) = \begin{cases} p_k(x_i) & b_A(x_i) = b_B(x_i) \\ \varphi(p_k(x_i)) & b_A(x_i) < b_B(x_i) \\ \psi(p_k(x_i)) & b_A(x_i) > b_B(x_i) \end{cases}$$

if $\mu(A) < \mu(B)$ we act symmetrically. That is, we model the effects of selection in the individual probability distribution using a pair of operators (φ, ψ) where φ is called the selection operator, and ψ is its dual, and are defined as follows. The operator φ, $\varphi : [0,1] \to [0,1]$, maps the cdf F of a individual into new cdf F', i.e., $F' = \varphi \circ F$. For reasons discussed next, this mapping should satisfy the following necessary properties:

1. Boundary conditions: $\varphi(0) = 0$ and $\varphi(1) = 1$;
2. Monotonically increasing: $\varphi(x) \leqslant \varphi(y)$ if $x \leqslant y$ for $x, y \in [0,1]$;
3. Convexity: $\varphi(x + \Delta x) - \varphi(x) \leqslant \varphi(x + 2\Delta x) - \varphi(x + \Delta x)$ for $\Delta x > 0$, and $x, x + 2\Delta x \in [0,1]$.

The operator ψ is the dual of φ, i.e., $\psi : [0,1] \to [0,1]$ and $\psi(x) = 1 - \varphi(1 - x)$.

Clearly, Properties 1 and 2 are necessary and sufficient to preserve the properties of the cdf. Property 3 (convexity) allows us to specify the desirable selection effect. Again assume that population is ordered by the function μ, i.e., the population can be represented as a chain

$$(A_1, \ldots, A_{m-2}, A_{m-1}, A_m, A_{m+1}, \ldots, A_{2^n})$$

where $\mu(A_i) < \mu(A_j)$ if $i < j$, and A_{m-1}, A_m are two nearest individuals in fitness terms such that $p(A_{m-1}) - p(A_m) = \Delta z$. Then $p(A_{m-1}) = F(A_{m-1}) - F(A_{m-2})$, $p(A_m) = F(A_m) - F(A_{m-1})$, and after selection there will be new probabilities $p'(A_{m-1}) = \varphi(F(A_{m-1})) - \varphi(F(A_{m-2}))$ and $p'(A_m) = \varphi(F(A_m)) - \varphi(F(A_{m-1}))$. According to selection we should have at least $p'(A_{m-1}) \leqslant p'(A_m)$. Denoting $z = F(A_{m-1})$, we get the inequality: $\varphi(z + \Delta z) - \varphi(z) \leqslant \varphi(z + 2\Delta z) - \varphi(z + \Delta z)$. If the function φ is two times differentiable on $[0,1]$, then Properties 2 and 3 are clearly equivalent to $\varphi'(x) \geqslant 0$ and $\varphi''(x) \geqslant 0$ for any $x \in [0,1]$.

Lemma 1 *Let $\varphi : [0,1] \to [0,1]$ be a continuous function on $[0,1)$ such that $\varphi(0) = 0$, $\varphi(1) = 1$, and $\varphi(x) < x$ for all $x \in (0,1)$. Then any sequence $\{x_k\}_{k=1}^{\infty}$, defined by $x_1 \in [0,1)$, $x_2 = \varphi(x_1)$,...,$x_k = \varphi(x_{k-1})$, ..., converges to 0, i.e. $\lim_{k\to\infty} x_k = 0$.*

Proof The analyzed sequence is decreasing and bounded, therefore its limit exists, i.e., we can write $\lim_{k\to\infty} x_k = a$. Clearly, $\lim_{k\to\infty} \varphi(x_k) = \lim_{k\to\infty} x_k = a$. Because $\lim_{k\to\infty} \varphi(x_k) = \varphi(\lim_{k\to\infty} x_k) = \varphi(a)$, we get $\varphi(a) = a$ and there is only the possibility $a = 0$, because $\varphi(x) < x$ for all $x \in (0,1)$. □

Remark 1 Notice that any convex function $\varphi : [0,1] \to [0,1]$ such that $\varphi(0) = 0$, $\varphi(1) = 1$, is continuous in $[0,1)$ and $\varphi(x) \leqslant x$ for all $x \in [0,1]$, and if $\varphi(x) \neq x$ at least in one point $x \in (0,1)$, then obviously $\varphi(x) < x$ for all $x \in (0,1)$.

Remark 2 Lemma 1 gives us also sufficient and necessary conditions for convergence. Assume that ranking is known, the initial cdf being $F_0 : 2^X \to [0,1]$. Further assume that A is the best individual and its probability is not equal to zero. Then $F_0(A) = 1$ and $F_0(B) \in [0,1)$ for any $B \neq A$. Consider the sequence of cdfs $F_1 = \varphi \circ F_0$, ..., $F_n = \varphi \circ F_{n-1}$, where the function φ satisfies the conditions of Lemma 1. Then obviously, there is a limit $F = \lim_{k\to\infty} F_k$ and $F(B) = 1$ if $B = A$ and $F(B) = 0$ otherwise. If φ does not satisfies the conditions of Lemma 1, for example, $\varphi(x) = x$ for some $x \in (0,1)$, then it is possible that $F(B) \neq 0$ for some $B \neq A$, or if the probability of the individual A is equal to zero, then for the next best individual B, we also have $F(B) = 1$.

To illustrate the application of the pair (φ, ψ) we generalize one of the simplest EDA algorithm, the cGA. This algorithm constitutes a parsimonious concretization of the general computational scheme for an EDA (Algorithm 1).

Algorithm 2: A generalized compact GA (cGA) algorithm

input : the length of the individuals n; a pair (φ, ψ); fitness function μ or a linear order among individuals; a termination tolerance parameter $\epsilon > 0$

output: The individual for which $p \approx 1$

$p := [\underbrace{0.5 \quad 0.5 \quad \cdots \quad 0.5}_{n \text{ elements}}]$; // Probability distribution

repeat

  $individual1$:= generate(p) ;

  $individual2$:= generate(p) ;

  $\{winner, loser\}$:= compete ($individual1$, $individual2$) ;

  for $j := 1$ **to** n **do**

    if $winner[j]$!= $loser[j]$ **then**

      if $winner[j]==1$ **then** $p[j] := \psi(p[j])$;

      else $p[j] := \varphi(p[j])$;

    end

  end

until $p_i \notin (0+\epsilon, 1-\epsilon)$; $i = 1, \ldots, n$;

with $\psi(x) := 1 - \varphi(1-x)$;

A full theoretical justification for using the proposed pair (φ, ψ) in this algorithm is provided in Sect. 4.

It is clear that Algorithm 2 includes cGA as a special case for (φ_1, ψ_1) with

$$\varphi_1(x) = \begin{cases} 0, & x \in [0, 1/S], \\ x - 1/S, & x \in (1/S, 1), \\ 1, & x = 1, \end{cases}$$

where S is the user-defined parameter representing the population size, and $\psi_1(x) = 1 - \varphi_1(1-x)$.

In the same vein, the proposed updating operator can also be viewed as a generalization of the incremental selection operator of PBIL for (φ', ψ') with $\varphi'(x) = (1-\alpha)x + \alpha\varphi(x)$, and $\psi'(x) = 1 - \varphi'(1-x) = (1-\alpha)x + \alpha(1 - \varphi(1-x)) = (1-\alpha)x + \alpha\psi(x)$, α being the so-called learning rate.

Consider now the specification of the mapping φ. Suppose we have a sample of individuals from the probability distribution defined by a cdf, F. Suppose that we divide this sample randomly in pairs and perform tournament among pair individuals. For the sake of simplicity let us enumerate individuals A_i, $i = 1, \ldots, 2^n$, such that $\mu(A_i) < \mu(A_j)$, if $i < j$. The win probability of individual A_i is $F(A_i) - 0.5p(A_i)$. Here we subtract $0.5p(A_i)$ to model the situation where a pair consists of two identical individuals. Let us compute the cdf for the selected population. We start by simplifying first the following expression:

$$\sum_{i=1}^{j} p(A_i)(F(A_i) - 0.5p(A_i))$$

$$= 0.5\sum_{i=1}^{j} p^2(A_i) + \sum_{i=2}^{j}\sum_{k=1}^{i} p(A_i)p(A_k) = 0.5F^2(A_j)$$

Therefore, progenitors have the cdf F^2 and $\varphi(x) = x^2$. We can model also other forms of selection. For example, if we consider that a given individual has a probability a of taking part in the tournament, then the operator becomes:

$$\varphi(x) = ax^2 + (1-a)x \tag{2}$$

Example 2 In this example we apply the generalized cGA, Algorithm 2, equipped with operator (2) to the well-known and widely used onemax (or bit counting) problem. The fitness function for this problem can be given by Harik et al. (1999 and Pelikan et al. (2000): $\mu(A) = \sum_{i=1}^{n} b_A(x_i)$.

Figure 1 shows the obtained results for a sequence of a values. As the probability a of an arbitrary individual to take part in the competition increases the number of generations required to achieve convergence decreases significantly attaining his minimum at $a = 1$. Concomitantly, the reliability of the whole algorithm to attain the optimum of μ also decreases; an issue that will be further studied in the next sections.

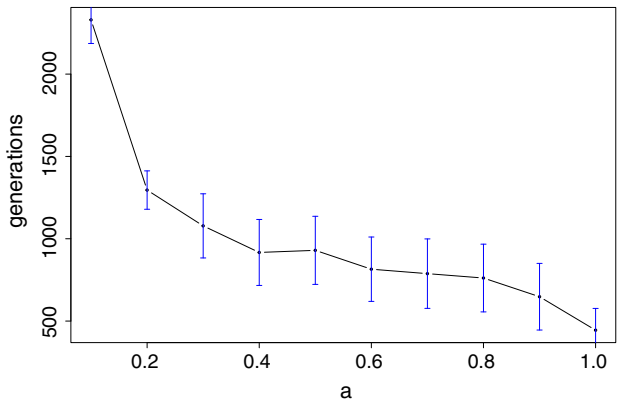

Fig. 1 Mean and standard deviation of the number of generations computed over 50 runs of the generalized cGA (Algorithm 2) versus a of the operator (2) when applied to the 100-bit onemax problem

3.1 On the comparison of operators

The merit of model updating operator φ depends on the class of fitness function (optimization problem) considered and, all other things being equal, it can be characterized by the associated (1) convergence rate, (2) convergence reliability, i.e., the ability to guide the algorithm to an extreme of the fitness function, and (3) scalability. Convergence rate is related to how φ updates probability distributions. Consider an ideal situation where the fitness function μ is linear, i.e.,

$$\mu(A) = \sum_{i=1}^{n} \alpha_i b_A(x_i) \tag{3}$$

with $\alpha_i < 0$, $i = 1, \ldots, n$, i.e., $\emptyset$ is the best individual while X is the worst. Under these conditions, probabilities $p^{(k)}(x_i)$ at generation k, are such that:

$$\begin{aligned}
&p^{(0)}(x_i) = 0.5,\\
&p^{(1)}(x_i) = \varphi(0.5),\\
&p^{(2)}(x_i) = \varphi\left(p^{(1)}(x_i)\right) = \varphi(\varphi(0.5)) = \varphi^{(2)}(0.5),\\
&\quad \ldots\ldots\\
&p^{(k)}(x_i) = \varphi\left(p^{(k-1)}(x_i)\right) = \varphi\left(\varphi^{(k-1)}(0.5)\right) = \varphi^{(k)}(0.5).
\end{aligned}$$

Here we are adopting the usual assumption, i.e., the initial population distribution $p^{(0)}$, is such that 0 and 1 are equiprobable values in any position x_i. We say that two selection operators φ_1 and φ_2 are (ε, N)-equivalent if $\varphi_1^{(N)}(0.5) = \varphi_2^{(N)}(0.5) = \varepsilon$. Therefore, we will observe the same convergence rate for two (ε, N)-equivalent selection operators after N iterations in ideal conditions. However, (ε, N)-equivalent selection operators can show different convergence rates and different reliability when applied to real optimization problems. The later can be simply evaluated by evaluating μ at the found solution.

Example 3 In the following we derive five (ε, N)-equivalent selection functions.

1. $\varphi_1(x) = \begin{cases} 0, & x \in [0, h], \\ x - h, & x \in (h, 1), \\ 1, & x = 1, \end{cases}$

   where h is a user-defined parameter. Again this operator is the equivalent to the original operator used for modeling selection in cGA with $h = 1/S$. Consequently,

   $$\varphi_1^{(N)}(x) = \begin{cases} 0, & x \in [0, hN], \\ x - Nh, & x \in (hN, 1), \\ 1, & x = 1, \end{cases}$$

   and φ_1 is in the class of (ε, N)-equivalent selection operators if $\varphi_1^{(N)}(0.5) = \varepsilon$, i.e., $0.5 - Nh = \varepsilon$ and $h = (0.5 - \varepsilon)/N$. See Fig. 2a for the characteristic of φ_1 for a range of population size parameter S.
2. $\varphi_2(x) = x^a$, where $a > 1$. Then $\varphi_2^{(N)}(x) = x^{a^N}$ and φ_2 is in the class of (ε, N)-equivalent selection operators if $\varphi_2^{(N)}(0.5) = \varepsilon$, i.e., $0.5^{a^N} = \varepsilon$ and $a = \sqrt[N]{-\lg_2(\varepsilon)}$. See Fig. 2b for the characteristic of φ_2 for the admissible range of a values.
3. $\varphi_3(x) = 1 - (1 - x)^a$, where $0 < a < 1$. Then $\varphi_3^{(N)}(x) = 1 - (1 - x)^{a^N}$ and φ_3 is in the class of (ε, N)-equivalent selection operators if $\varphi_3^{(N)}(0.5) = \varepsilon$, i.e., $1 - 0.5^{a^N} = \varepsilon$ and $a = \sqrt[N]{-\lg_2(1 - \varepsilon)}$. See Fig. 2c for the characteristic of φ_3 for the admissible range of a values.
4. $\varphi_4(x) = ax^2 + (1 - a)x$, where $0 < a \leqslant 1$. In this case it is not possible to find the analytical solution of the equation $\varphi_4^{(N)}(0.5) = \varepsilon$ but it can be solved numerically. See Fig. 2d for the characteristic of φ_4 for the admissible range of a values.
5. $\varphi_5(x) = \Phi\left(a + \Phi^{-1}(x)\right)$, where Φ is the cdf of the standard normal distribution, i.e.,

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

   Φ^{-1} is the normal inverse cdf (quantile) function, and $a < 0$. In this case $\varphi_5^{(N)}(x) = \Phi\left(aN + \Phi^{-1}(x)\right)$ and φ_5 is in the class of (ε, N)-equivalent selection operators if $\varphi_5^{(N)}(0.5) = \varepsilon$, i.e., $\Phi\left(aN + \Phi^{-1}(0.5)\right) = \varepsilon$ and $a = \Phi^{-1}(\varepsilon)/N$. See Fig. 2e for the characteristic of φ_5 for the admissible range of a values.

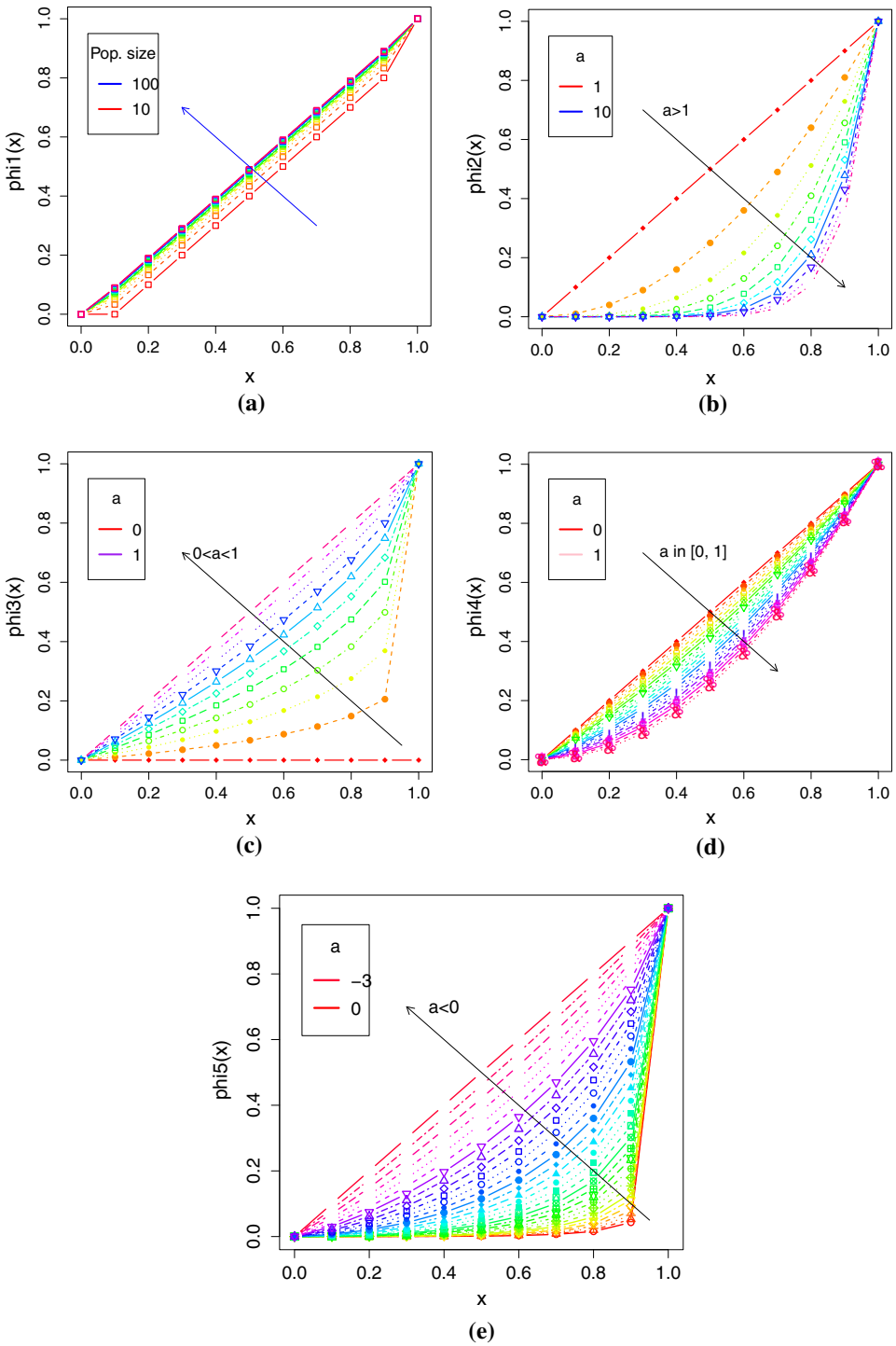

Fig. 2 Characteristics of the proposed operators as function of their parameters **a** operator $\varphi_1(x)$, i.e., cGA, **b** operator $\varphi_2(x) = x^a$, **c** operator $\varphi_3(x) = 1 - (1-x)^a$, **d** operator $\varphi_4(x) = ax^2 + (1-a)x$, **e** operator $\varphi_5(x) = \Phi\big(a + \Phi^{-1}(x)\big)$. (Color figure online)

See Sect. 4 for a theoretical justification of the above proposed operators. See Fig. 2 for the different characteristics of these operators as function of their parameters.

An alternative way of assessing the merits of a model updating operator φ on a given fitness function (optimization problem) is to establish a criterium for its convergence reliability and computes the required effort for convergence. A possibility is to compare the number of generations required by a given operator for which the maximum likelihood estimation (MLE) of the probability of success (reaching a global maximum) is greater or equal than a given confidence level. This approach is also used in

Sastry and Goldberg (2001) and followed in the example below.

3.2 An optimal updating operator based on *a prior* information

It should be clear that while conditions 1, 2, and 3 in the beginning of Sect. 3 are necessary conditions to be taken into account while searching for an operator for a given optimization problem, the merit of an operator depends on that problem. To further stress this point, assume that the fitness function μ has the following structure:

$$\mu(B) = \sum_{i=1}^{n} \xi_i b_B(x_i) \tag{4}$$

where $\xi_i = \begin{cases} 1, & i \in A \\ -1 & i \notin A \end{cases}$ and $A \subseteq \{1, 2, \ldots, n\}$. Clearly this problem has a maximum for B equals to A. Let us assume that our prior information about the optimal individual denoted as $(b_A^*(x_1), \ldots, b_A^*(x_n))$ is described by probabilities $p(x_i) = \Pr(\xi_i), i = 1, \ldots, n$, and we observe two individuals B and C such that $\mu(B) > \mu(C)$. Let us update probabilities $p(x_i)$ using the Bayes rule, i.e., compute probabilities $p^*(x_i) = \Pr(\xi_k = 1 | \mu(B) > \mu(C)), i = 1, \ldots, n$. Let us compute the probability $\Pr(\mu(B) > \mu(C))$. For this we can assume that random variables $\xi_1, \ldots, \xi_n$ are independent and analyze the random variable

$$\xi = \mu(B) - \mu(C) = \sum_{x_i \in B \setminus C} \xi_i - \sum_{x_i \in C \setminus B} \xi_i$$

then

$$\xi = \begin{cases} \eta + \xi_k, & x_k \in B \setminus C, \\ \eta - \xi_k, & x_k \in C \setminus B, \\ \eta & \text{otherwise} \end{cases}$$

with

$$\eta = \sum_{x_i \in B \setminus (C \cup \{x_k\})} \xi_i - \sum_{x_i \in C \setminus (B \cup \{x_k\})} \xi_i$$

Thus, following Bayes, the optimal updating expression for $p(x_k)$, $p^*(x_k); k = 1, \ldots, n$ is:

$$p^*(x_k) = \frac{p(x_k) \Pr(\eta + a > 0)}{p(x_k) \Pr(\eta + a > 0) + (1 - p(x_k)) \Pr(a > 0)}$$

where

$$a = \begin{cases} 1 & x_k \in B \setminus C, \\ -1 & x_k \in C \setminus B, \\ 0 & \text{otherwise} \end{cases}$$

Notice that $p^*(x_k) = p(x_k)$ if $x_k \in (B \setminus C) \cup (C \setminus B)$. Because η is the sum of independent random variables, it can be approximated by a normal distribution of mean

$$\mathbb{E}[\eta] = \mathbb{E}\left[\sum_{x_i \in B \setminus (C \cup \{x_k\})} \xi_i - \sum_{x_i \in C \setminus (B \cup \{x_k\})} \xi_i\right] = \sum_{x_i \in B \setminus (C \cup \{x_k\})} (2p(x_i) - 1) - \sum_{x_i \in C \setminus (B \cup \{x_k\})} (2p(x_i) - 1) \tag{5}$$

and variance

$$\sigma^2[\eta] = \sigma^2\left[\sum_{x_i \in B \setminus (C \cup \{x_k\})} \xi_i - \sum_{x_i \in C \setminus (B \cup \{x_k\})} \xi_i\right] = 4\left[\sum_{x_i \in B \setminus (C \cup \{x_k\})} p(x_i)(1 - p(x_i)) + \sum_{x_i \in C \setminus (B \cup \{x_k\})} p(x_i)(1 - p(x_i))\right] \tag{6}$$

Thus it is legitime to approximate probabilities as follows

$$\Pr(\eta + a > 0) = 1 - \Pr(\eta \leq -a) \approx 1 - \Phi\left(\frac{-a - \mathbb{E}[\eta]}{\sigma[\eta]}\right)$$

$$\Pr(\eta > 0) = 1 - \Pr(\eta \leq 0) \approx 1 - \Phi\left(\frac{-\mathbb{E}[\eta]}{\sigma[\eta]}\right)$$

where again Φ stands for the cdf of the standard normal distribution.[1] For the sake of computation efficiency, we can assume *a priori* $p(x_i) = 0.5$ in expressions (5) and (6) yielding $\mathbb{E}[\eta] = 0$ and $\sigma[\eta] = \sqrt{|C \setminus B| + |B \setminus C| - 1}$, respectively. Under these assumptions, the new updating operator is hereafter referred to as the Bayesian operator.

Example 4 In this example we assess the performance of the above updating operator by contrasting its performance with the performance of cGA. In particular we compare the number of iterations required to achieve the same MLE of the probability of success $\hat{p}_s$, i.e., reaching the global maximum of fitness function (4) with $n = 20$ and $A = \{1, 3, 10, 20\}$.

To begin with we examine the performance of the Algorithm 2 equipped with φ_1 above, i.e., acting as the classic cGA. From Fig. 3a we observe that at least $S = 30$ is required for achieving $\hat{p}_s = 1$ over 1000 independent runs for a stop criterium with $\epsilon = 0.01$.

Figure 3b shows the number of generations averaged over 1000 runs for the cGA versus the population size S the same conditions above. The proposed updating operator has always converged to the global maximum (i.e., it ran always with $\hat{p}_s = 1$). As can be seen in Fig. 5a the number of generations required by the proposed Bayesian operator is significantly lower than that required by cGA.

Example 5 In this example a first attempt is made to assess the scalability of the proposed operator. The

[1] Several software packages include efficient ways of computing the normal cdf.

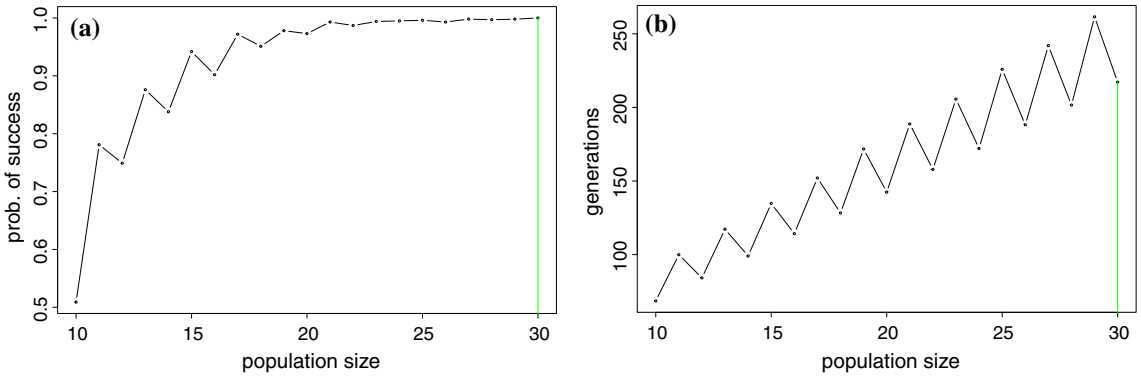

Fig. 3 **a** MLE of prob. of success $\hat{p}_s$ and **b** average number of generations over 1000 independent runs for cGA versus the population size S when applied to a 20-bit problem characterized by the fitness function (4). The *green vertical line* shows $S = 30$ as the minimum value required for achieving $\hat{p}_s = 1$. (Color figure online)

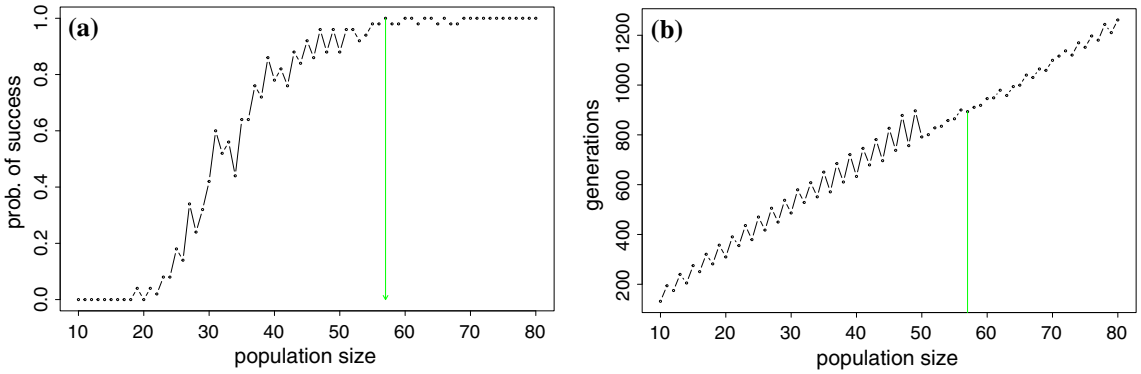

Fig. 4 **a** MLE of prob. of success $\hat{p}_s$ and **b** number of generations averaged over 50 independent runs for cGA versus the population size S when applied to a 100-bit problem characterized by the fitness function (4).The *green vertical line* shows $S = 57$ as the minimum value required for achieving $\hat{p}_s = 1$. (Color figure online)

example is essentially the same as the previous one except for the size of individuals that changed from $n = 20$ to $n = 100$ and

$$A = \{1, 3, 10, 20, 21, 22, 25, 30, 35, 38, 40, 50, 60, 70, 80, 90, 100\}$$

in function (4).

Following a process similar to the previous example we observe in Fig. 4a that at least $S = 57$ is required for achieving $\hat{p}_s = 1$ over 50 independent runs. The proposed updating operator has always converged to the global maximum. As can be seen in Fig. 5b the number of generations required by the proposed Bayesian operator is again significantly lower than that required by cGA.

Comparing Figure. 5a, b we see that while cGA has suffered an increment of 675 in the average number of generations, the proposed Bayesian operator required only more 521 generations in average when n changed from 20 to 100.

Example 6 Following the same methodology of the previous examples, here we compare the performance of the proposed operators with the performance of cGA in yet another fitness function, i.e.,

$$\mu(A) = \sum_{i=1}^{n} 2^{n-i} b_A(x_i) \tag{7}$$

for $n = 20, 30, 40$, and 50. Again we require a reliability characterized by $\hat{p}_s = 1$ over 50 independent runs for a stop criterium with $\epsilon = 0.01$.

Figures 6, 7, 8 and 9 show both (a) the MLE of prob. of success $\hat{p}_s$ and (b) the average number of generations for Algorithm 2 when equipped with each one of the proposed operators versus their respective parameters, when applied to the fitness function (7) with $n = 40$. The figures also show those parameter values that allow us to achieve the *minimum* average generations for the above defined reliability. Operator φ_2 performs very poorly in this problem; it requires more than 3000 generations in average

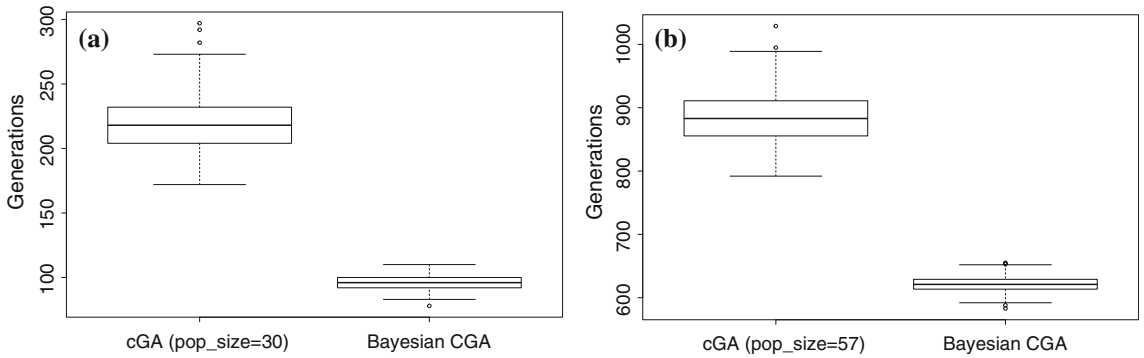

Fig. 5 Box plots exhibiting the dispersion, skewness, as well as outliers of the number of generations achieved over 100 independent runs for the proposed Bayesian operator and for cGA under the conditions required for achieving $\hat{p}_s = 1$ when applied to **a** 20-bit and **b** a 100-bit problem characterized by the fitness function (4)

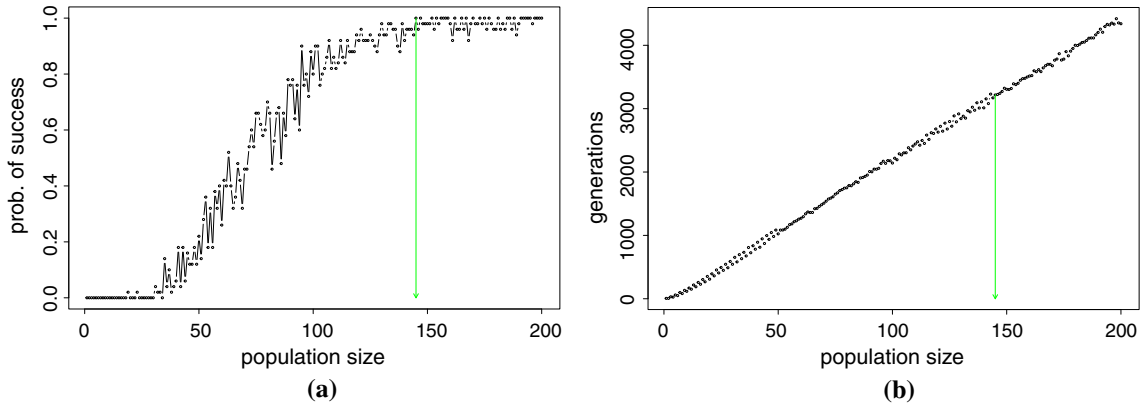

Fig. 6 Performance of operator $\varphi_1(x)$, i.e., cGA, showing (*green vertical line*) $S = 145$ as the value required for achieving $\hat{p}_s = 1$ over 50 independent runs with fitness function (7) with $n = 40$ **a** $\hat{p}_s$ versus S, **b** mean generations versus S. (Color figure online)

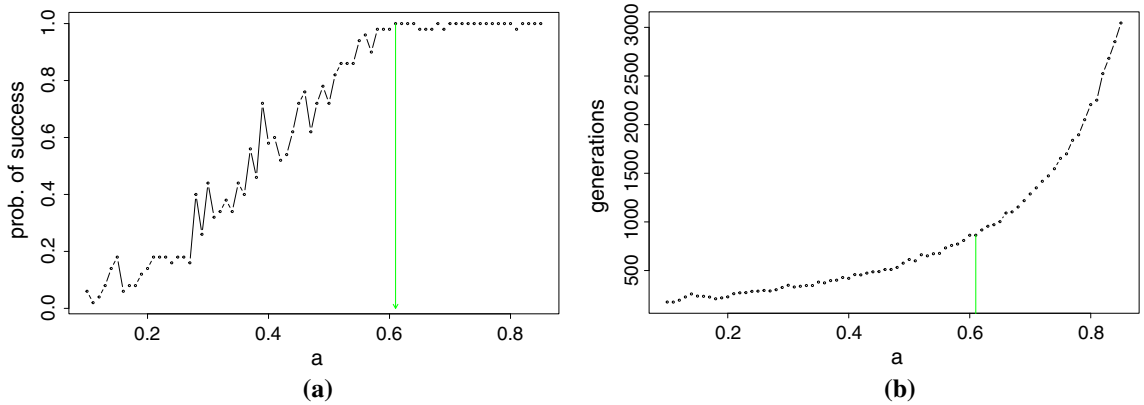

Fig. 7 Performance of operator $\varphi_3(x) = 1 - (1 - x)^a$ showing (*green vertical line*) $a = 0.61$ as the value required for achieving $\hat{p}_s = 1$ over 50 independent runs with fitness function (7) with $n = 40$ **a** $\hat{p}_s$ versus a, **b** mean generations versus a. (Color figure online)

for $n = 20$ and hence it is immediately excluded from further comparisons.

Figure 10 clearly show that, for this optimization problem, the proposed Bayesian operator has a convergence rate significantly higher than all the others, requiring an average of about 450 generations. Ranking second in this comparison is operator φ_3 with an average of 894 generations. The cGA is the last of the group with an average of about 3327 generations.

Figure 11 allows us to infer the scalability of the different operators in this problem. A first observation is

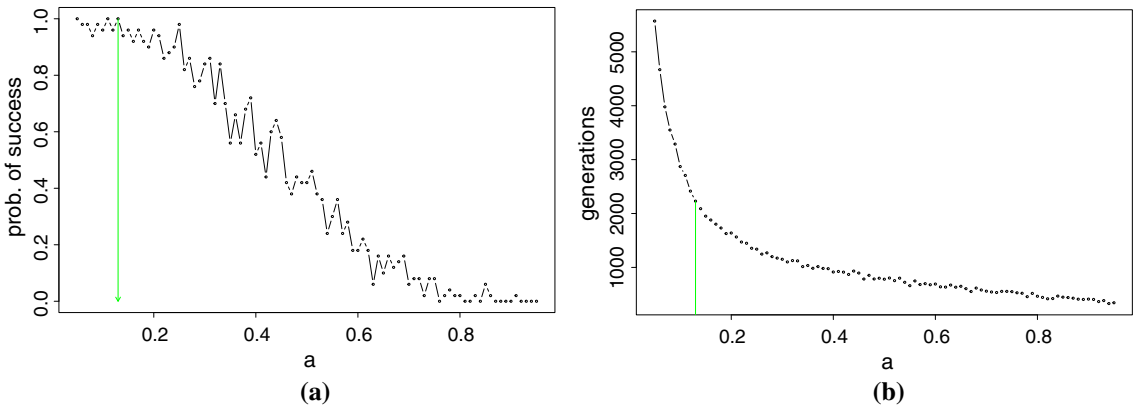

Fig. 8 Performance of operator $\varphi_4(x) = ax^2 + (1-a)x$ showing (*green vertical line*) $a = 0.13$ as the value required for achieving $\hat{p}_s = 1$ over 50 independent runs with fitness function (7) with $n = 40$, **a** $\hat{p}_s$ versus a, **b** mean generations versus a. (Color figure online)

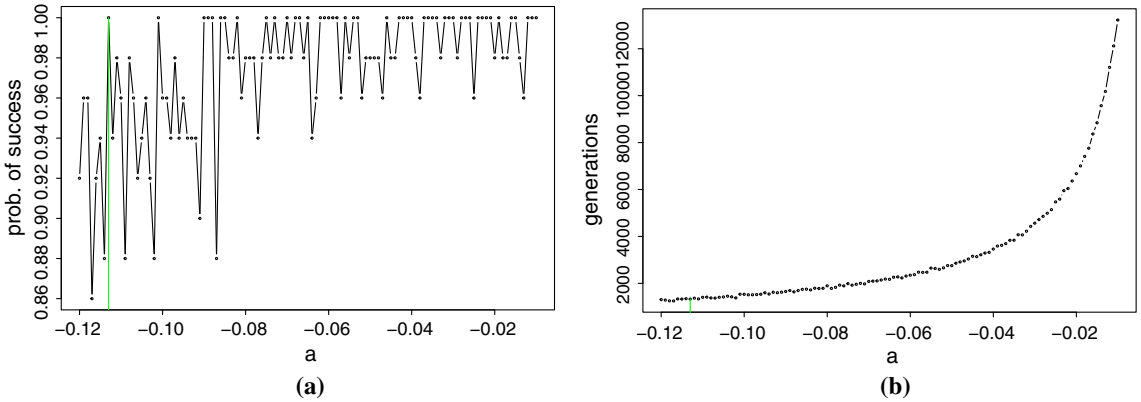

Fig. 9 Performance of operator $\varphi_5(x) = \Phi\left(a + \Phi^{-1}(x)\right)$ showing (*green vertical line*) $a = -0.11$ as the value required for achieving $\hat{p}_s = 1$ over 50 independent runs with fitness function (7) with $n = 40$ **a** $\hat{p}_s$ versus a, **b** mean generations versus a. (Color figure online)

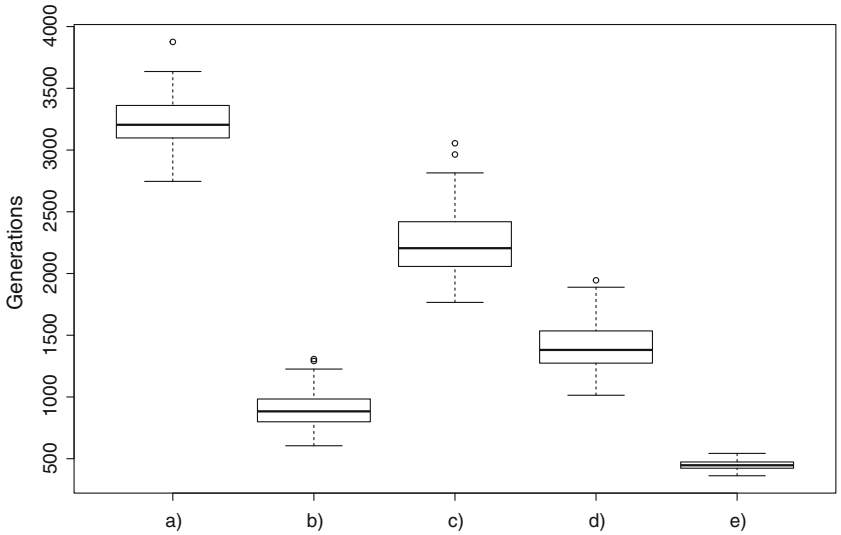

Fig. 10 Box plots exhibiting the dispersion, skewness, as well as outliers of the number of generations achieved over 100 independent runs for **a** cGA, **b** φ_3, **c** φ_4, **d** φ_5, and **e** Bayesian under the conditions required for achieving $\hat{p}_s = 1$ when applied to the fitness function (7) with $n = 40$

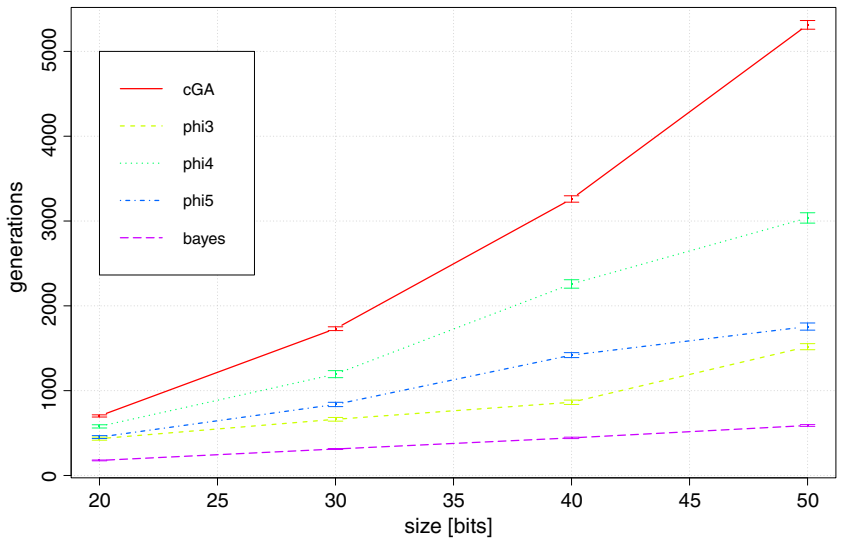

Fig. 11 Average number of generations and the corresponding confidence interval for a confidence level of 95% versus the size of the problem (7) relatively to 100 independent runs for cGA and the proposed operators. (Color figure online)

that there is a statistically significant difference between the convergence rate of the operators. The figure also shows that cGA scales worst, i.e., exhibits the higher growing rate in the average number of generations when the number of bits in (7) increases. On the hand, Algorithm 2 equipped with the proposed Bayesian operator is the fastest and has a growing rate that appear to be linear in the size of the problem. The absolute difference of the average number of generations between these two extreme cases is almost one order of magnitude for $n = 50$. The other operators are within these two cases.

In the light of reproducible research, the source code in R programming language used for generating the presented results is publicly available at w3.ualg.pt/ jvo/pubs/NACO-S-14-00107

4 Main results

Assume that the population is described by a probability distribution $p : 2^X \to [0, 1]$, where again $X = \{x_1, \ldots, x_n\}$. Then we can say that each p defines a random set $\mathcal{A}$ together with the random variables $b_{\mathcal{A}}(x_1), \ldots, b_{\mathcal{A}}(x_n)$. One of the simplest assumptions one can make is to admit independent random variables $b_{\mathcal{A}}(x_1)$,...,$b_{\mathcal{A}}(x_n)$. This is the assumption found in univariate EDAs (Baluja 1994; Mühlenbein 1997; Harik et al. 1999). In this case the probability distribution p is uniquely defined by the probabilities $p(x_i) = \Pr(b_{\mathcal{A}}(x_i) = 1)$, and obviously $p(\mathcal{A}) = \prod_{x_i \in \mathcal{A}} p(x_i) \prod_{x_i \notin \mathcal{A}} (1 - p(x_i))$.

Now suppose we know that the fitness function μ can be approximated by a linear function f:

$$f(A) = \sum_{i=1}^{n} \alpha_i b_A(x_i) \tag{8}$$

where α_i is a real number. For example, if we know the value of fitness function for two individuals B and C, and that $\mu(B) > \mu(C)$, then we can define $\alpha_i = b_B(x_i) - b_C(x_i)$ stressing which variables in B are better than in C.

We will analyze next the effect of applying the selection operator φ to the cdf $\hat{F}(B) = \sum_{f(A) \leqslant f(B)} p(A)$ that can be considered as an approximation of the cdf $F(B) = \sum_{\mu(A) \leqslant \mu(B)} p(A)$. We will show that under some assumptions on the selection operator φ, we can approximate the new probability distribution described by the cdf $\varphi \circ \hat{F}$ using probabilities $p'(x_i)$ of variables calculated by

$$p'(x_i) = \begin{cases} \varphi(p(x_i)), & \alpha_i < 0, \\ p(x_i), & \alpha_i = 0, \\ \psi(p(x_i)), & \alpha_i > 0. \end{cases}$$

Now, let us evaluate the value $\hat{F}(B)$ for an arbitrary individual B. Without restricting generality, we can assume that $\alpha_i \leqslant 0$, i.e., $\emptyset$ is the best individual while X is the worst. In this case $\mu(A) \leqslant \mu(B)$ for all $B \subseteq A$, therefore:

$$\hat{F}(B) = \sum_{f(A) \leqslant f(B)} p(A) \geqslant \sum_{B \subseteq A} p(A) = \prod_{x_i \in B} p(x_i)$$

Analogously, $f(A) > f(B)$ for all $A \subset B$, therefore, $1 - \hat{F}(B) = \sum_{f(A) > f(B)} p(A) \geqslant \sum_{A \subset B} p(A) \quad = \sum_{A \subseteq B} p(A) - p(B) = \prod_{x_i \in X \setminus B} (1 - p(x_i)) - p(B)$. Thus we have the following inequalities:

$$\prod_{x_i\in B} p(x_i) \leqslant \hat{F}(B) \leqslant 1 - \left(\prod_{x_i\in X\setminus B} (1-p(x_i))\right) \times \left(1 - \prod_{x_i\in B} p(x_i)\right).$$

Observe that the upper bound is equal to the lower bound if $\prod_{x_i\in X\setminus B} (1-p(x_i)) = 1$. Clearly, the above estimates are very rough and in the most cases $P(B)$ is considerably smaller than $\prod_{x_i\in X\setminus B} (1-p(x_i))$. Therefore, for simplicity, we assume that $\hat{F}(B) \leqslant 1 - \prod_{x_i\in X\setminus B} (1-p(x_i))$. Now, consider the following problem. Let the mapping $\varphi : [0,1] \to [0,1]$ be a selection operator, i.e., the cdf is defined as $\varphi \circ \hat{F}$ on the next generation. The question is how to approximate this probability distribution by a random set $\mathcal{A}'$ assuming that random variables $b_{\mathcal{A}'}(x_1)$,...,$b_{\mathcal{A}'}(x_n)$ are independent? Let us denote $p'(x_i) = \Pr(b_{\mathcal{A}'}(x_i) = 1)$. Clearly, if we use some approximation the algorithm has approximately, at least, the same convergence as the algorithm producing pure selection if

$$\varphi\left(\prod_{x_i\in B} p(x_i)\right) \geqslant \prod_{x_i\in B} p'(x_i), \quad \varphi\left(1 - \prod_{x_i\in X\setminus B} (1-p(x_i))\right) \leqslant 1 - \prod_{x_i\in X\setminus B} (1-p'(x_i))$$

Observe that if $\varphi(t) = t^\alpha$ then $\varphi(\prod_{x_i\in B} p(x_i)) = (\prod_{x_i\in B} p(x_i))^\alpha = \prod_{x_i\in B} p^\alpha(x_i)$, i.e., in this case at least for the lower bound the choice $p'(x_i) = \varphi(p(x_i)) = p^\alpha(x_i)$ is justifiable. This allows us to conclude that we have a good approximation at least for the lower bound if a function φ is close in some sense to the power function. Observe that the power function $\varphi(t) = t^\alpha$ obeys the following characteristic equation:

$$\varphi^{q_1}(t_1)\cdot\varphi^{q_2}(t_2) = \varphi\left(t_1^{q_1}\cdot t_2^{q_2}\right).$$

If $q_1 \in [0,1]$ and $q_2 = 1 - q_1$ the expression

$$\varphi\left(t_1^{q_1}\cdot t_2^{q_2}\right) \approx \varphi^{q_1}(t_1)\cdot\varphi^{q_2}(t_2)$$

can be conceived as the interpolation of φ in the point $t_1^{q_1}\cdot t_2^{q_2}$ by its values in points t_1 and t_2. Therefore, we can say that the function $\varphi : [0,1] \to [0,1]$ gives us the *lower interpolation* of the power function if

$$\varphi\left(t_1^q\cdot t_2^{1-q}\right) \leqslant \varphi^q(t_1)\cdot\varphi^{1-q}(t_2) \text{ for any } t_1,t_2,t_1+t_2,q\in[0,1] \tag{9}$$

We can also analyze the required conditions for the selection operator considering the upper bound. In this case the inequality can be rewritten as

$$\varphi\left(1 - \prod_{x_i\in X\setminus B} (1-p(x_i))\right) \geqslant 1 - \prod_{x_i\in X\setminus B} (1-\varphi(p(x_i))).$$

Using the auxiliary function $\psi(x) = 1 - \varphi(1-x)$ it is straightforward to rewrite the above inequality as

$$\prod_{x_i\in X\setminus B} \psi(1-p(x_i)) \geqslant \psi\left(\prod_{x_i\in X\setminus B} (1-p(x_i))\right)$$

i.e., in this case it is desirable that the function ψ should be the *upper interpolation* of power function:

$$\psi\left(t_1^q\cdot t_2^{1-q}\right) \geqslant \psi^q(t_1)\cdot\psi^{1-q}(t_2) \text{ for any } t_1,t_2,t_1+t_2,q\in[0,1] \tag{10}$$

This choice is analyzed in the following lemma.

Lemma 2 *Let $\varphi : [0,1] \to [0,1]$ be a selection operator. Then it satisfies Properties 2 and 3 of selection operators for $\psi(x) = 1 - \varphi(1-x)$ iff*

1. *the function $\gamma_1(y) = \ln(\varphi(e^y))$ is convex on $(-\infty, 0]$;*
2. *the function $\gamma_2(y) = \ln(1 - \varphi(1-e^y))$ is concave on $(-\infty, 0]$.*

Proof Let us show that (9) is equivalent to 1. The inequality (9) can be rewritten as $\ln\left(\varphi\left(t_1^q\cdot t_2^{1-q}\right)\right) \leqslant q\ln(\varphi(t_1)) + (1-q)\ln(\varphi(t_2))$. Then after changing variables $t_1 = e^{y_1}$ and $t_2 = e^{y_2}$, we get the inequality

$$\gamma_1(qy_1 + (1-q)y_2) \leqslant q\gamma_1(y_1) + (1-q)\gamma_2(y_2)$$

for all $y_1, y_2, y_1 + y_2 \leqslant 0$ and $q \in [0,1]$. This means the convexity of γ_1 in $(-\infty, 0]$. In the same way, one can show that (10) is equivalent to 2. □

Remark 3 The conditions formulated in statements 1 and 2 of Lemma 2 are different from those that define the convexity of φ and the concavity of $f(x) = 1 - \varphi(1-x)$. To illustrate it, suppose that the function $\gamma_1(y) = \ln(\varphi(e^y))$ is convex. Then it is equivalent to

$$\gamma_1(x) + \gamma_1(y) \geqslant 2\gamma_1\left(\frac{x+y}{2}\right),$$

or

$$\varphi(e^x)\varphi(e^y) \geqslant \varphi^2\left(e^{\frac{x+y}{2}}\right).$$

Denoting $a = e^x$ and $b = e^y$, we get $\varphi(a)\varphi(b) \geqslant \varphi^2\left(\sqrt{ab}\right)$ for all $a, b \in [0,1]$. While the convexity of φ on $[0,1]$ means that for all $a, b \in [0,1]$,

$$\varphi(a)+\varphi(b)\geqslant 2\varphi\left(\frac{a+b}{2}\right)$$

Example 7 Let $\varphi(t)=at^2+(1-a)t$, where $a\in[0,1]$. Then $\gamma_1(y)=\ln(ae^{2y}+(1-a)e^y)$ and $\gamma_2(y)=\ln((1+a)e^y-ae^{2y})$. In this case, $\frac{d^2}{dy^2}\gamma_1(y)=\frac{(1-a)ae^y}{(ae^y+(1-a))^2}\geqslant 0$ for all $y\in(-\infty,0]$ and $\frac{d^2}{dy^2}\gamma_2(y)=-\frac{(1+a)ae^y}{((1+a)-ae^y)^2}\leqslant 0$ for all $y\in(-\infty,0]$, i.e., γ_1 is convex on $(-\infty,0]$ and γ_2 is concave on $(-\infty,0]$.

Example 8 Let $\varphi(t)=t^\alpha$, where $\alpha>1$. Then $\gamma_1(y)=\alpha y$, $\gamma_2(y)=\ln\left(1-e^{\alpha(1-y)}\right)$,

$$\frac{d^2}{dy^2}\gamma_2(y)=-\frac{\alpha e^y(1-e^y)^\alpha((1-e^y)^\alpha+\alpha e^y-1)}{(1-e^y)^2(1-(1-e^y)^\alpha)^2}\leqslant 0,$$

because all the involved factors are non-negative. For example, to prove that for all $y\in(-\infty,0]$, $(1-e^y)^\alpha+\alpha e^y-1\geqslant 0$, start by denoting $x=1-e^y$ then the corresponding inequality is rewritten as $f(x)=x^\alpha+\alpha(1-x)-1\geqslant 0$, where $x\in[0,1]$. The minimum of this function on $[0,1]$ is $f(1)=0$. Therefore, γ_1 is convex on $(-\infty,0]$ and γ_2 is concave on $(-\infty,0]$.

Example 9 Let

$$\varphi(t)=\begin{cases}\dfrac{bx}{a}, & x\in[0,a],\\ \dfrac{(1-b)(x-1)}{1-a}+1, & x\in(a,1],\end{cases}$$

where $0<b\leqslant a<1$. As φ is a convex function,

$$\frac{d^2}{dy^2}\gamma_1(y)=\frac{\left(1-\frac{1-b}{1-a}\right)\left(\frac{1-b}{1-a}\right)e^y}{\left(\left(\frac{1-b}{1-a}\right)(e^y-1)+1\right)^2}$$

for $y\in(\ln a,0]$ and $\frac{d^2}{dy^2}\gamma_1(y)\geqslant 0$ for $y\in(\ln a,0]$ iff $a\leqslant b$, i.e., the condition 1) of Lemma 2 is fulfilled iff $a=b$ or equivalently $\varphi(t)=t$, $t\in[0,1]$. Analogously, one can show that condition 2) of Lemma 2 is fulfilled iff $a=b$ or equivalently $\varphi(t)=t$, $t\in[0,1]$.

The following theorems give the full description of functions that satisfy conditions of Lemma 2 under certain differentiable conditions.

Theorem 1 *Let $\varphi:[0,1]\to[0,1]$ be an increasing and convex function such that*

(a) $\varphi(0)=0$, $\varphi(1)=1$;
(b) *φ is increasing, convex and differentiable on* $[0,1]$;
(c) *the function $\gamma_1(y)=\ln(\varphi(e^y))$ is convex on* $(-\infty,0]$;
(d) *the function $\gamma_2(y)=\ln(1-\varphi(1-e^y))$ is concave on* $(-\infty,0]$;
(e) *$F(x)=\frac{\varphi'(x)x}{\varphi(x)}$ is defined and differentiable on* $[0,1]$. *Here we assume that $F(0)=\lim_{x\to+0}F(x)=a$. Then for $a\geqslant 1$, F is an increasing function on* $[0,1]$ *and φ can be represented as*

$$\varphi(x)=x^a\left(\frac{e^{\int_0^x\frac{F_0(y)}{y}dy}}{e^{\int_0^1\frac{F_0(y)}{y}dy}}\right) \tag{11}$$

where $F_0(y)=F(y)-a$.

Conversely, let $F_0:[0,1]\to[0,+\infty)$ be an increasing and differentiable function in $[0,1]$ *with $F_0(0)=0$. Then any function φ defined by* (11) *satisfies the conditions* (a)–(e) *for $a\geqslant 1$.*

Proof See "Appendix". □

We will use also the following counterpart of Theorem 1.

Theorem 1* *Let $\psi:[0,1]\to[0,1]$ be an increasing and concave function such that*

(a) $\psi(0)=0$, $\psi(1)=1$;
(b) *ψ is increasing, concave and differentiable on* $[0,1]$;
(c) *the function $\gamma_1(y)=\ln(\psi(e^y))$ is concave on* $(-\infty,0]$;
(d) *the function $\gamma_2(y)=\ln(1-\psi(1-e^y))$ is convex on* $(-\infty,0]$;
(e) *$F(x)=\frac{\varphi'(x)x}{\varphi(x)}$ is defined and differentiable* on $[0,1]$. *Again, assume that $F(0)=\lim_{x\to+0}F(x)=a$.*

Then for $0<a\leqslant 1$, F is an increasing function on $[0,1]$ *and φ can be represented as*

$$\varphi(x)=x^a\left(\frac{e^{\int_0^x\frac{F_0(y)}{y}dy}}{e^{\int_0^1\frac{F_0(y)}{y}dy}}\right) \tag{12}$$

where $F_0(y)=F(y)-a$.

Conversely, let $F_0:[0,1]\to(-\infty,0]$ be an decreasing and differentiable function on $[0,1]$ *with $F_0(0)=0$ and $F(1)\geqslant -a$. Then any function φ defined by* (12) *satisfies the conditions* (a)–(e) *for $0<a\leqslant 1$.*

Proof Follows the same path as the proof of Theorem 1 and is omitted for brevity. □

Proposition 1 *The inequality*

$$\varphi\left(\prod_{x_i\in B} p(x_i)\right) \geqslant \prod_{x_i\in B} \varphi(p(x_i)) \tag{13}$$

is satisfied for a selection operator φ if the function $g(y) = \gamma_1(y)/y$ is increasing in $(-\infty, 0]$, where $\gamma_1(y) = \ln(\varphi(e^y))$.

Proof Take the logarithm from both sides of inequality (13) and let $e^{y_i} = p(x_i)$. Then the inequality (13) can be rewritten as

$$f\left(\sum_{x_i\in B} y_i\right) \geqslant \sum_{x_i\in B} f(y_i)$$

or

$$\left(\sum_{x_i\in B} y_i\right) g\left(\sum_{x_i\in B} y_i\right) = \sum_{x_i\in B} y_i g\left(\sum_{x_i\in B} y_i\right) \geqslant \sum_{x_i\in B} y_i g(y_i) \tag{14}$$

Notice that the function g is non-negative in $(-\infty, 0]$ and $g\left(\sum_{x_i\in B} y_i\right) \leqslant g(y_i)$ by our assumption, i.e., $y_i g\left(\sum_{x_i\in B} y_i\right) \geqslant y_i g(y_i)$. This implies that the inequality (14) is valid. □

Proposition 2 *The inequality*

$$\prod_{x_i\in X\setminus B} \psi(1 - p(x_i)) \leqslant \psi\left(\prod_{x_i\in X\setminus B} (1 - p(x_i))\right) \tag{15}$$

is satisfied for a function ψ if the function $g(y) = \gamma_1(y)/y$ is increasing in $(-\infty, 0]$, where $\gamma_1(y) = \ln(\psi(e^y))$.

Proof This follows the same reasoning of the proof of Proposition 1 and is omitted for brevity. □

Corollary 1 (1) *A selection operator φ obeys the inequality (13) if the function $g_1(x) = \frac{\ln \varphi(x)}{\ln x}$ is increasing, and* (2) *it obeys the inequality (15) for $\psi(x) = 1 - \varphi(1 - x)$ if the function $g_2(x) = \dfrac{\ln(1 - \varphi(x))}{\ln(1 - x)}$ is increasing.*

Proof The first statement follows directly from Proposition 1 after putting $y = \ln x$ in function g. Analogously, Proposition 2 implies that the inequality (15) is valid if the function $g_2(1 - x) = \dfrac{\ln(1 - \psi(1 - x))}{\ln x}$ is decreasing, i.e., g_2 is an increasing function. □

Corollary 2 *Let the operators φ and ψ obey the conditions of Theorem 1 and Theorem 1*, respectively. Then the inequalities (13) and (15) are also respectively satisfied.*

Proof Let a selection operator φ obey the conditions of Proposition 1. Then it can be represented as $\varphi(x) = xe^{g_1(x)}$, where g_1 is an increasing function. We can derive the analogous representation if the selection operator φ obey the conditions of Theorem 1. Actually, in this case $\varphi(x) = xe^{(a-1)\ln x + \int_0^x \frac{F_0(y)}{y} dy - A}$, where $a \geqslant 1$, F_0 is an increasing function with $F_0(0) = 0$, and $A = \int_0^1 \frac{F_0(y)}{y} dy$. We see that in this case $g_1(x) = (a - 1)\ln x + \int_0^x \frac{F_0(y)}{y} dy - A$ is an increasing function, i.e., the inequality (13) is fulfilled by Proposition 1.

Let a selection operator φ obey the conditions of Proposition 2. Then the function $\psi(x) = 1 - \varphi(1 - x)$ can be represented as $\psi(x) = xe^{g_2(x)}$, where g_2 is a decreasing function. We can derive the analogous representation if the selection operator φ obeys the conditions of Theorem 1*. Actually, in this case $\psi(x) = xe^{(a-1)\ln x + \int_0^x \frac{F_0(y)}{y} dy - A}$, where $0 < a \leqslant 1$, F_0 is a decreasing function with $F_0(0) = 0$ and $F_0(1) \leqslant a$, and $A = \int_0^1 \frac{F_0(y)}{y} dy$. We see that in this case $g_2(x) = (a - 1)\ln x + \int_0^x \frac{F_0(y)}{y} dy - A$ is a decreasing function, i.e., the inequality (15) is fulfilled by Proposition 2. □

Remark 4 It is possible to weaken the second part of Theorem 1. Namely, if $F_0 : [0, 1] \to [0, +\infty)$ is an increasing function on $[0, 1]$ with $F_0(0) = 0$. Then any function φ defined by (10) satisfies the conditions a)-d) for $a \geqslant 1$. For example, if we take

$$F_0(x) = \begin{cases} 0, & x \in [0, b], \\ c, & x \in (b, 1], \end{cases}$$

where $b \in [0, 1]$ and $c > 0$. Then

$$\varphi(x) = \begin{cases} b^c x^a, & x \in [0, b], \\ x^{a+c}, & x \in (b, 1]. \end{cases}$$

Remark 5 Theorem 1 gives us also a simple way to check conditions (b)–(d). To do this, it is sufficient to check whether function F is increasing and $F(0) \geqslant 1$.

Consider the case where (8) verifies $\alpha_1 < \alpha_2 < \cdots < \alpha_n < 0$ and $\sum_{k=i+1}^n \alpha_k > \alpha_i$ for any $i = 1, \ldots, n - 1$. Then $f(A) < f(B)$ iff $b_A(x_i) = b_B(x_i)$ for $i = 1, \ldots, k - 1$, but $b_A(x_i) = 1$ and $b_B(x_i) = 0$ for $i = k$. Consider an increasing sequence of sets $\{C_k\}_{k=1}^n$, such that $C_k = \{x_1, \ldots, x_k\}$. Then the above condition can be rewritten as $f(A) < f(B)$ iff there is a $k \in \{1, \ldots, n\}$ such that $A \cap C_{k-1} = B \cap C_{k-1}$ and $A \cap C_k \supset B \cap C_k$. Therefore,

$$\{B | f(A) < f(B)\} = \bigcup_{k | x_k \in A} \{B | B \cap C_{k-1} = A \cap C_{k-1}, x_k \notin B\}.$$

We see that

$$P(\{B|B\cap C_{k-1}=A\cap C_{k-1}, x_k\notin B\}) = (1-p(x_k))\prod_{x_i\in A\cap C_{k-1}} p(x_i)\prod_{x_i\in C_{k-1}\setminus A}(1-p(x_i)).$$

Since sets $\{B|B\cap C_{k-1}=A\cap C_{k-1}, x_k\notin B\}$ are disjoint for different $x_k\in A$, we get

$$P(\{B|f(A)<f(B)\}) = \sum_{k|x_k\in A}(1-p(x_k))\prod_{x_i\in A\cap C_{k-1}} p(x_i)\prod_{x_i\notin C_{k-1}\setminus A}(1-p(x_i)),$$

or

$$\hat{F}(B)=\sum_{f(A)\leqslant f(B)} p(A) = 1-\sum_{k|x_k\in B}(1-p(x_k))\prod_{x_i\in B\cap C_{k-1}} p(x_i)\prod_{x_i\in C_{k-1}\setminus B}(1-p(x_i)).$$

In particular,

$$\hat{F}(\{x_k\})=\sum_{f(A)\leqslant f(B)} p(A) = 1-\prod_{i=1}^{k}(1-p(x_i)),$$

and if $B=\{x_{k+1},\ldots,x_n\}$, then

$$\hat{F}(B)=1-\left(\prod_{x_i\in X\setminus B}(1-p(x_i))\right)\left(1-\prod_{x_i\in B}p(x_i)\right)$$

In this case the upper bound of $\hat{F}(B)$ calculated above is achieved.

For this ordering, it is also possible to get another formula for F. This is derived from:

$$\{A|f(A)<f(B)\}=\bigcup_{k|x_k\notin B}\{A|B\cap C_{k-1}=A\cap C_{k-1}, x_k\in A\}.$$

Clearly,

$$P(\{A|B\cap C_{k-1}=A\cap C_{k-1}, x_k\in A\}) = p(x_k)\prod_{x_i\in B\cap C_{k-1}} p(x_i)\prod_{x_i\in C_{k-1}\setminus B}(1-p(x_i)),$$

and $\{A|B\cap C_{k-1}=A\cap C_{k-1}, x_k\in A\}$ are disjoint for different $x_k\notin B$. Therefore,

$$P(\{A|f(A)<f(B)\}) = \sum_{k|x_k\notin B} p(x_k)\prod_{x_i\in B\cap C_{k-1}} p(x_i)\prod_{x_i\in C_{k-1}\setminus B}(1-p(x_i))$$

and

$$\hat{F}(B)=\sum_{f(A)\leqslant f(B)} p(A)=\sum_{k|x_k\notin B} p(x_k)\prod_{x_i\in B\cap C_{k-1}} p(x_i)\prod_{x_i\in C_{k-1}\setminus B}(1-p(x_i))+P(B).$$

In particular,

$$\hat{F}(X\setminus\{x_k\})=\prod_{i=1}^{k}p(x_k)+P(X\setminus\{x_k\})$$

and

$$\hat{F}(\{x_1,\ldots,x_k\})=\prod_{i=1}^{k}p(x_k),$$

i.e., in this case the lower bound of $\hat{F}(B)$ calculated above is achieved.

Assume that $\alpha_1=\alpha_2=\cdots=\alpha_n=-1$. Then $f(A)\leqslant f(B)$ if $|A|\geqslant|B|$. Therefore,

$$\hat{F}(B)=\sum_{f(A)\leqslant f(B)} p(A)=\sum_{|A|\geqslant|B|}\prod_{x_i\in A} p(x_i)\prod_{x_i\notin A}(1-p(x_i)).$$

In particular, a random value f has a binomial distribution if $p(x_i)=a$ for all $i\in\{1,\ldots,n\}$. Observe that in this case the function $\hat{F}$ may be approximated by a cdf of a normal distribution. For this case, we have

$$\hat{F}(B)=\sum_{f(A)\leqslant f(B)} p(A)\approx\Phi\left(\frac{x+na}{\sqrt{na(1-a)}}\right) \tag{16}$$

where again Φ is the cdf of the standard normal distribution. Expression (16) can be interpreted in another way. Instead of approximating $\hat{F}$, we can approximate first independent random variables $b_{\mathcal{A}}(x_1)$,...,$b_{\mathcal{A}}(x_n)$ by continuous independent random variables $\xi_1,\ldots,\xi_n$, distributed normally with means equal to a and variances equal to $\sigma^2=a(1-a)$. Then we approximate

$$f(\mathcal{A})=\sum_{i=1}^{n}\alpha_i b_{\mathcal{A}}(x_i)\approx\sum_{i=1}^{n}\alpha_i\xi_i.$$

Observe that a random variable $\xi=\sum_{i=1}^{n}\alpha_i\xi_i$ has normal distribution with a mean $\sum_{i=1}^{n}\alpha_i a=-na$ and a variance $\sum_{i=1}^{n}\alpha_i^2\sigma^2=n\sigma^2$, i.e., in this case we get (16) again. Using the same approximation, but for the general case, when α_i and $p(x_i)$ are arbitrary, we obtain

$$\hat{F}(B)\approx\Phi\left(\frac{x-\sum_{i=1}^{n}\alpha_i p(x_i)}{\sqrt{\sum_{i=1}^{n}\alpha_i^2 p(x_i)(1-p(x_i))}}\right)$$

The conditions under which this expression is justifiable can be found in the various generalizations of the central limit theorem.

In the same vein, it is of interest the problem of describing all justifiable selection operators that map any normal distribution to another normal distribution. Let us remind that any cdf of normally distributed random value ξ can be represented as $F(x)=\Phi((x-a)/\sigma)$, where a is the mean value of ξ, and σ^2 is its variance. The selection operator $\varphi:[0,1]\to[0,1]$ with $\varphi(0)=0$ and $\varphi(1)=1$ that

maps cdf $\Phi((x-a_1)/\sigma_1)$ to cdf $\Phi((x-a_2)/\sigma_2)$ can be found as a solution of the following equation:

$$\varphi(\Phi((x-a_1)/\sigma_1)) = \Phi((x-a_2)/\sigma_2).$$

Let us denote $z = \Phi((x-a_1)/\sigma_1)$. Then $x = \sigma_1\Phi^{-1}(z) + a_1$ and

$$\varphi(z) = \Phi\big(\alpha + \beta\Phi^{-1}(z)\big) \tag{17}$$

where $\alpha = (a_1 - a_2)/\sigma_2$ and $\beta = \sigma_1/\sigma_2$. Notice that by using a selection operator we are searching for a maximum of the function μ while simultaneously attempt to reduce of search region, this implies that $a_2 > a_1$ and $\sigma_1 \geqslant \sigma_2$, i.e., $\alpha < 0$ and $\beta \geqslant 1$. The next proposition establishes when such a φ has desirable properties.

Proposition 3 *Let $\varphi : [0,1] \to [0,1]$ be defined by (17). Then φ is a selection operator and it satisfies the conditions of Theorem 1 iff $\alpha < 0$ and $\beta = 1$.*

Proof See "Appendix". □

Let us analyze the consequences of Proposition 3. Let $f(\mathcal{A}) = \sum_{i=1}^{n} \alpha_i b_{\mathcal{A}}(x_i)$, with $\alpha_i < 0$, $i = 1, \ldots, n$. Assume the cdfs

$$F_i(t) = \Pr\{\alpha_i b_{\mathcal{A}}(x_i) \leqslant t\} = \begin{cases} 0, & t < \alpha_i, \\ p(x_i), & \alpha_i \leqslant t < 0, \\ 1, & t \geqslant 0. \end{cases}$$

and apply the selection operator φ to F_i to produce the cdf

$$\varphi \circ F_i(t) = \Pr\{\alpha_i b_{\mathcal{A}}(x_i) \leqslant t\} = \begin{cases} 0, & t < \alpha_i, \\ \varphi(p(x_i)), & \alpha_i \leqslant t < 0, \\ 1, & t \geqslant 0. \end{cases}$$

Notice that in Algorithm 2 we approximate the probability distribution after selection by probabilities $(\varphi(p(x_i)), 1 - \varphi(p(1-x_i)))$, i.e., we show that our approximation of applying the selection operator is produced by applying the same selection operator to random variables $\alpha_i b_{\mathcal{A}}(x_i)$. Let us assume that we approximate random variables $\alpha_1 b_{\mathcal{A}}(x_1)$, ..., $\alpha_n b_{\mathcal{A}}(x_n)$ by independent normally distributed random variables $\xi_1, \ldots, \xi_n$ and $\xi = \sum_{i=1}^{n} \xi_i$. Let a_i be a mean value of ξ_i and σ_i^2 be its variance. Then the application of the selection operator φ to the cdf of ξ_i means that we obtain a new normally distributed random variable ξ_i' with mean value $a_i' = a_i - \alpha\sigma_i$ and with variance σ_i^2. Analogously, the normally distributed random variable ξ has the mean value $a = \sum_{i=1}^{n} a_i$ and the variance $\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$ and after applying the selection operator to its cdf, we get the normally distributed random variable ξ' with the mean value $a' = a - \alpha\sigma$ and variance σ^2. Let us consider the sum $\xi'' = \sum_{i=1}^{n} \xi_i'$. This normally distributed random variable has a mean value $a'' = a - \alpha\sum_{i=1}^{n} \sigma_i$ and variance σ^2. Notice that $a'' > a'$ since $\sigma^2 = \sum_{i=1}^{n} \sigma_i^2 \leqslant \left(\sum_{i=1}^{n} \sigma_i\right)^2$ and $\alpha < 0$.

5 Conclusions

Estimation of distribution algorithms (EDAs) are stochastic population-based optimizers that combine techniques of machine learning and evolutionary computation. Currently there is a wealth of EDAs and many applications have been reported. However relatively less theoretical results are available.

In this paper, the role of the selection operation on the updating of the probability distribution was thoroughly investigated. Necessary conditions for an operator to model selection in such a way that it can be directly used for updating the probability model were postulated. A family of such operators was proposed. The properties of these operators were thoroughly analyzed, including a study on the equivalence of operators. The proposed operators were shown to generalize existing operators, namely they generalize both the operator used in the cGA and the updating operator used in the PBIL algorithm. As an application example in algorithm design, a generalization of cGA was presented. As our main theoretical results a comprehensive theoretical rationale for the proposed operators and for their application to univariate EDA design was provided.

Examples aiming at illustrating key concepts, main results, and their relevance were presented. These include simulation studies revealing some empirical evidence on the convergence rate, convergence reliability, and scalability of the proposed operators. While it should be clear that the merit of an operator depends on the optimization problem, our studies reveal that, when equipped with a selection of the proposed operators, the presented generalized cGA was able to outperform cGA in two different optimization problems under different parametrization and problem sizes.

Although the presented results are mainly for the univariate case, these can also be used for extending some multivariate EDAs, specially those relying in clustering for obtaining groups of variables where intra group independence is reasonable to assume.

Acknowledgments Andrey Bronevich is grateful to the Erasmus Mundus Triple I Consortium that supported a 10-months academic visit to the University of Algarve in 2010. This work is an outcome of a research cooperation between the authors that began with this visit. Andrey Bronevich also thanks the National Research University Higher School of Economics, Moscow, Russia for providing him with 1 month research grant for visiting University of Algarve in July 2014 facilitating the conclusion of the work. José Valente de Oliveira also thanks the National Research University Higher School of Economics, for inviting him for one week visit in November 2014.

Appendix

Proof of Theorem 1 Under conditions (a)–(e) function γ_1 is differentiable in $(-\infty, 0]$ and its convexity implies that

$$\frac{d}{dy}\gamma_1(y) = \frac{\varphi'(e^y)e^y}{\varphi(e^y)}$$

is increasing in $(-\infty, 0]$. This is obviously equivalent to have $F(x) = \frac{\varphi'(x)x}{\varphi(x)}$ as an increasing function in $(0, 1]$. Solving this differential equation w.r.t. φ we see that the function φ can be expressed as $\varphi(x) = Cx^a e^{\int_0^x \frac{F_0(y)}{y}dy}$, where an arbitrary constant C should be chosen such that $\varphi(1) = 1$, i.e., $C = e^{-\int_0^1 \frac{F_0(y)}{y}dy}$.

Let us show that the value $a \geqslant 1$(clearly $a > 0$ in order to have $\varphi(0) = 0$). Denote $\varphi_0(x) = Ce^{\int_0^x \frac{F_0(y)}{y}dy}$. Then

$$\begin{aligned}
&\varphi(x) = x^a\varphi_0(x);\\
&\varphi'(x) = x^a\varphi_0'(x) + ax^{a-1}\varphi_0(x)\\
&\quad = x^{a-1}\varphi_0(x)F_0(x) + ax^{a-1}\varphi_0(x)\\
&\quad = x^{a-1}\varphi_0(x)(F_0(x) + a);\\
&\varphi''(x) = \varphi'(x)\left(\frac{a-1}{x} + \frac{\varphi_0'(x)}{\varphi_0(x)} + \frac{F_0'(x)}{F_0(x)+a}\right)\\
&\quad = \varphi'(x)\left(\frac{a-1}{x} + \frac{F_0(x)}{x} + \frac{F_0'(x)}{F_0(x)+a}\right).
\end{aligned}$$

Therefore, if $a - 1 < 0$ then $\varphi''(x) < 0$ for some values that are close to 0, since $\lim_{x\to+0}\left(\frac{F_0(x)}{x}/\frac{a-1}{x}\right) = 0$. This means that $a \geqslant 1$.

For the second part of the theorem, if we choose the function F_0 as stated by the theorem, then conditions (a)–(c), and (e) are obviously satisfied. It remains to show that (d) is also satisfied. Function $\gamma_2(y) = \ln(1 - \varphi(1 - e^y))$ is differentiable and the function

$$\frac{d}{dy}\gamma_2(y) = -\frac{\varphi'(1-e^y)e^y}{1-\varphi(1-e^y)}$$

should be decreasing. This is equivalent to

$$g(x) = \frac{\varphi'(x)(1-x)}{1-\varphi(x)}$$

be an increasing function. Substituting $\varphi(x) = x^a\varphi_0(x)$, $\varphi'(x) = x^{a-1}\varphi_0(x)(F_0(x) + a)$, we get

$$g(x) = \frac{x^{a-1}\varphi_0(x)(F_0(x)+a)(1-x)}{1-x^a\varphi_0(x)}.$$

Then

$$\begin{aligned}
g'(x) = g(x)\Bigg(&\frac{a-1}{x} + \frac{\varphi_0(x)}{\varphi_0'(x)} + \frac{F_0'(x)}{F_0(x)+a}\\
&+ \frac{x^{a-1}\varphi_0(x)(F_0(x)+a)}{1-x^a\varphi_0(x)} - \frac{1}{1-x}\Bigg).
\end{aligned}$$

Notice that

$$\begin{aligned}
&\frac{\varphi_0(x)}{\varphi_0'(x)} = \frac{F_0(x)}{x};\\
&\frac{x^{a-1}\varphi_0(x)(F_0(x)+a)}{1-x^a\varphi_0(x)} = \frac{F_0(x)+a}{x(1-x^a\varphi_0(x))} - \frac{F_0(x)+a}{x}.
\end{aligned}$$

Therefore,

$$g'(x) = g(x)\left(\frac{F_0'(x)}{F_0(x)+a} + \frac{F_0(x)+a}{x(1-x^a\varphi_0(x))} - \frac{1}{x(1-x)}\right).$$

Because $\varphi_0(x) \leqslant 1$ and $F_0(x) \geqslant 0$, $F_0'(x)/(F_0(x)+a) \geqslant 0$, we get

$$g'(x) \geqslant g(x)\left(\frac{a}{x(1-x^a)} - \frac{1}{x(1-x)}\right).$$

We see that

$$\frac{a}{x(1-x^a)} - \frac{1}{x(1-x)} = \frac{x^a + a(1-x) - 1}{x(1-x^a)(1-x)} \geqslant 0,$$

where the inequality $x^a + a(1-x) - 1 \geqslant 0$ for $a \geqslant 1$ and $x \in [0, 1]$ is proved as in Example 3, i.e., $g'(x) \geqslant 0$ and the condition d) is also fulfilled. □

Proof of Proposition 3 We will use the notation from Theorem 1. Computing the function g for this case, we get

$$g(x) = \frac{\beta\Phi'\big(\alpha + \beta\Phi^{-1}(x)\big)x}{\Phi'\big(\Phi^{-1}(x)\big)\Phi\big(\alpha + \beta\Phi^{-1}(x)\big)}$$

Let us compute the limit $b = \lim_{x\to+0} g(x)$, involving a new variable $y = \Phi^{-1}(x)$:

$$b = \lim_{y\to-\infty}\frac{\beta\Phi'(\alpha+\beta y)\Phi(y)}{\Phi'(y)\Phi(\alpha+\beta y)} = \beta\lim_{y\to-\infty}\frac{e^{-\frac{(\alpha+\beta y)^2-y^2}{2}}\Phi(y)}{\Phi(\alpha+\beta y)}$$

Obviously, $b = 1$ for $\beta = 1$. Applying the L'Hospital rule for the general case, one can obtain that $b = \beta^2$, i.e., we conclude that $\beta \geqslant 1$.

Let us check now when g is an increasing function. This condition is equivalent to the non-negativity of the derivative

$$f(y) = \frac{d}{dy}\ln\left(\frac{\beta\Phi'(\alpha+\beta y)\Phi(y)}{\Phi'(y)\Phi(\alpha+\beta y)}\right)$$

for any $y \in (-\infty, +\infty)$. Making simple calculations, we get

$$f(y) = \beta\frac{\Phi''(\alpha+\beta y)}{\Phi'(\alpha+\beta y)} + \frac{\Phi'(y)}{\Phi(y)} - \frac{\Phi'(y)}{\Phi''(y)} - \beta\frac{\Phi'(\alpha+\beta y)}{\Phi(\alpha+\beta y)}$$
$$= -\alpha\beta - \left(\beta^2-1\right)y + \frac{\Phi'(y)}{\Phi(y)} - \beta\frac{\Phi'(\alpha+\beta y)}{\Phi(\alpha+\beta y)}.$$

Notice that $\lim_{y\to+\infty} f(y) = -\infty$ for $\beta > 1$ and $\lim_{y\to+\infty} f(y) = -\alpha$ for $\beta = 1$, therefore, the function φ can satisfy the conditions of Theorem 1 if $\alpha<0$ and $\beta = 1$. For this case we can represent the function f in the form $f(y) = w(y) - w(y+\alpha)$, where $w(y) = y + \frac{\Phi'(y)}{\Phi(y)}$ and $f(y) \geqslant 0$ for all $y \in \mathbb{R}$ and $\alpha<0$, if $\frac{dw}{dy} \geqslant 0$. Let us denote $u(y) = \frac{\Phi'(y)}{\Phi(y)}$. The function u can be conceived as a partial solution of the Bernoulli differential equation $\frac{du}{dy} = -uy - u^2$. This equation and the condition $\frac{dw}{dy} = \frac{du}{dy} + 1 \geqslant 0$ imply that the function f is increasing iff $\frac{du}{dy} = -uy - u^2 \geqslant -1$ or, taking in account that the function u is non-negative, that

$$u \leqslant \frac{-y+\sqrt{y^2+4}}{2}$$

By expressing the last inequality in terms of cdf for the standard normal distribution one gets:

$$\Phi(y) \geqslant \frac{2}{\left(-y+\sqrt{y^2+4}\right)\sqrt{2\pi}} e^{-\frac{y^2}{2}} = \frac{\left(y+\sqrt{y^2+4}\right)}{2\sqrt{2\pi}} e^{-\frac{y^2}{2}} \tag{18}$$

which implies

$$\frac{d}{dy}\Phi(y) \geqslant \frac{d}{dy}\frac{\left(y+\sqrt{y^2+4}\right)}{2\sqrt{2\pi}} e^{-\frac{y^2}{2}} \text{ for all } y \in \mathbb{R},$$

or

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \geqslant \frac{1}{2\sqrt{2\pi}} e^{-\frac{y^2}{2}} \left(1 + \frac{y}{\sqrt{y^2+4}} - y^2 - y\sqrt{y^2+4}\right)$$

or

$$1 \geqslant \frac{1}{2}\left(1 + \frac{y}{\sqrt{y^2+4}} - y^2 - \frac{y(y^2+4)}{\sqrt{y^2+4}}\right)$$

Now denote $v = y/\sqrt{y^2+4}$. Then $y^2 = 4v^2/(1-v^2)$ and we get:

$$1 \geqslant \frac{1}{2}\left(1 + v - \frac{4v^2}{1-v^2} - v\left(4 + \frac{4v^2}{1-v^2}\right)\right)$$
$$= \frac{1}{2}\left(1 - 3v - \frac{4v^2}{1-v^2}(1+v)\right)$$

We proceed taking in to account that $|v|<1$:

$$1 \geqslant -3v - \frac{4v^2}{1-v}$$

or

$$v^2 + 2v + 1 \geqslant 0$$

As this inequality is satisfied for all real v we conclude that g is an increasing function and that the function φ satisfies the conditions of Theorem 1 for $\alpha<0$ and $\beta = 1$. □

References

Baluja S (1994) Population-based incremental learning: a method for integrating genetic search based function optimization and competitive learning. Technical report CMU-CS-94-13. Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

Baluja S, Davies S (1997) Using optimal dependency-trees for combinatorial optimization: learning the structure of the search space. In: Proceedings of the 14-th International Conference on Machine Learning. San Francisco, California, USA, pp 30–38

Bengoetxea E, Larrañaga P, Bloch I, Perchant A, Boeres C (2002) Inexact graph matching by means of estimation of distribution algorithms. Pattern Recognit 35(12):2867–2880

Ceberio J, Irurozki E, Mendiburu A, Lozano JA (2012) A review on estimation of distribution algorithms in permutation-based combinatorial optimization problems. Prog AI 1(1):103–117

De Bonet JS, Isbell CL, Viola P (1997) MIMIC: finding optima by estimating probability densities. In: Petsche T, Mozer MC, Jordan MI (eds) Advances in neural information processing systems. MIT Press, Cambridge, pp 424–430

Droste S (2006) A rigorous analysis of the compact genetic algorithm for linear functions. Nat Comput 5:257–283

Echegoyen C, Mendiburu A, Santana R, Lozano JA (2013) On the taxonomy of optimization problems under estimation of distribution algorithms. Evolut Comput 21(3):471–495

Emmendorfer LR, Pozo AT (2009) Effective linkage learning using low-order statistics and clustering. IEEE Trans Evolut Comput 13(6):1233–1246

González C, Lozano JA, Larrañaga P (2001) Analyzing the PBIL algorithm by means of discrete dynamical systems. Complex Syst 12(4):465–479

Harik GR (1999) Linkage learning via probabilistic modeling in the ECGA. Technical report 99010. Illinois Genetic Algorithms Laboratory, University of Illinois, Urbana, Illinois, USA

Harik GR, Lobo FG, Goldberg DE (1999) The compact genetic algorithm. IEEE Trans Evolut Comput 3(4):287–297

Hauschild M, Pelikan M (2011) An introduction and survey of estimation of distribution algorithms. Swarm Evolut Comput 1:111–128

Johnson A, Shapiro JL (2002) The importance of selection mechanisms in distribution estimation algorithms. In: Collet P, Fonlupt C, Hao JK, Lutton E, Schoenauer M (ed) Evolution artificial, vol 2310. Lecture Notes in Computer Science, Springer, pp 91–103

Larrañaga P, Lozano JA (2001) Estimation of distribution algorithms: a new tool for evolutionary computation. Kluwer Academic Publishers, Norwell

Lozada-Chang L, Santana R (2011) Univariate marginal distribution algorithm dynamics for a class of parametric functions with unitation constraints. Inf Sci 181(11):2340–2355

Mühlenbein H (1997) The equation for response to selection and its use for prediction. Evolut Comput 5:303–346

Mühlenbein H, Mahnig T (1998) Convergence theory and applications of the factorized distribution algorithm. J Comput Inf Technol 7:19–32

Mühlenbein H, Mahnig T (2002) Evolutionary algorithms and the Boltzmann distribution. In: DeJong KA, Poli R, Rowe J (eds) Foundation of genetic algorithms 7. Morgan Kaufmann, Burlington, pp 133–150

Mühlenbein H, Mahnig T, Rodriguez AO (1999) Schemata, distributions and graphical models in evolutionary optimization. J Heuristics 5:215–247

Pelikan M, Goldberg DE (2000) Genetic algorithms, clustering, and the breaking of symmetry. Lecture Notes in Computer Science 1917, pp 385–394

Pelikan M, Goldberg DE (2001) Escaping hierarchical traps with competent genetic algorithms. In: Genetic and Evolutionary Computation Conference, pp 511–518

Pelikan M, Goldberg DE, Cantu-Paz E (2000) Linkage problem, distribution estimation, and Bayesian networks. Evolut Comput 8:311–340

Pelikan M, Mühlenbein H (1999) The bivariate marginal distribution algorithm. In: Chawdhry PK, Roy R, Furuhashi T (eds) Advances in soft computing—engineering design and manufacturing. Springer, Berlin, pp 521–535

Peña J, Lozano J, Larrañaga P (2005) Globally multimodal problem optimization via an estimation of distribution algorithm based on unsupervised learning of Bayesian networks. Evolut Comput 13(1):43–66

Sastry K, Goldberg DE (2001) Modeling tournament selection with replacement using apparent added noise. In: Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001). San Francisco, California, USA, p 781

Zhang Q (2004a) On stability of fixed points of limit models of univariate marginal distribution algorithm and factorized distribution algorithm. IEEE Trans Evolut Comput 8(1):80–93

Zhang Q (2004b) On the convergence of a factorized distribution algorithm with truncation selection. Complexity 9(4):17–23

Zhang Q, Mühlenbein H (2004) On the convergence of a class of estimation of distribution algorithms. IEEE Trans Evolut Comput 8(2):127–136