On the Optimal Convergence Probability of Univariate Estimation of Distribution Algorithms

Reza Rastegar rastegar@iastate.edu
Department of Mathematics, Iowa State University, Ames, IA 50011, US

Abstract
In this paper, bounds on the probability of convergence to the optimal solution have been obtained for the compact Genetic Algorithm (cGA) and the Population Based Incremental Learning (PBIL). The sufficient condition for convergence of these algorithms to the optimal solution and a range of possible values of the algorithm's parameters for which they converge to the optimal solution with a confidence level are determined.

Keywords
Markov Process, Submartingale, Subregular Functions, Optimal Convergence Probability.

1 Introduction

Although univariate Estimation of Distribution Algorithms (EDAs) have low efficiency in solving difficult problems, it is still important to study them for two reasons. First, due to their simplicity in terms of memory usage and computational complexity they may be quite useful in memory-constrained applications, especially for implementing evolvable hardware. Second, it is advised to begin with a simple EDA to develop methods needed for the analysis of more complicated EDAs (Droste, 2005). Three of the simplest univariate EDAs (UEDAs) are the cGA(Harik et al., 1999b), the PBIL(Baluja and Caruana, 1995), and the UMDA (Mühlenbein, 1997) which is a special case of the PBIL. These algorithms initialize a probability vector (PV), in which each component of the PV follows a Bernoulli distribution with the parameter of 0.5, thereby randomly generating solutions by employing this PV. Some of the generated solutions are selected based on their fitness values and a selection scheme. Next, the PV is updated using learning algorithms. The process of adaptation continues until some criteria are satisfied, for example, the PV converges.

A few people have studied different theoretical aspects of these simple algorithms including their convergence and time complexity. The first theoretical study of the convergence of the PBIL with an arbitrary learning rate in $(0, 1)$ is carried out by Hohfeld and Rudolph (1997). It is argued that the PBIL converges almost surely to the maximum point of linear functions. We will return to this result later in this paper. Having a sufficiently small learning rate, Gonzalez et al. (2000) model the PBIL using a discrete dynamical system and demonstrate that the local optimum of an injective function with respect to Hamming distance are stable fixed points of the PBIL. They also study the strong dependency of the PBIL on initial values of the PV and the learning rate (Gonzalez et al., 2001). In an interesting paper, Zhang (2004) studies the stability of fixed points of limit models of the UMDA while using two-tournament selection scheme and shows that the local optima with respect to Hamming distance are asymptotically stable. In (Rastegar and Meybodi, 2005), the PBIL is studied for the case that the population size is sufficiently large and as a result dynamical properties of the algorithm are derived for different selection schema. Also, in (Rastegar and Hariri, 2006b,a), it is proven that the PBIL and the cGA with sufficiently small learning rates do not show any cyclic or chaotic behavior but instead converge weakly to the local maxima with respect to Hamming distance when they optimize an injective function. Time complexity is another aspect of these algorithms studied by a few researchers. Droste (2005) carries out the first rigorous study on the time complexity of the cGA for linear pseudo-boolean functions. He shows that not all linear functions have the same asymptotical runtime. Chen et al. (2007) study the time complexity of the PBIL and the UMDA. They extend the concept of convergence to convergence time and estimate the upper bound of the mean first hitting times of the UMDA and the PBIL on a simple pseudo-modular function and analyze the mean first hitting time of the PBIL on a hard problem. The result shows that the PBIL may spend exponential time to find the global optimum.

Another important topic is the effect of initial parameters, such as the initial PV, the learning rate, and the population size, on the probability that the cGA and the PBIL converge to optimal solutions, called optimal convergence probability, which to the best knowledge of the author is not studied deeply. The importance of this topic appears when one notice for example when the learning rate is not small enough, it is not likely that the cGA converges to a good solution for the problem and therefore, it may appear reasonable that to find solutions of high quality, the learning rate be small as much as possible. However, if the learning rate is too small, the cGA will waste time processing unnecessary individuals, and this may result in unacceptably slow performance. The problem is to find a learning rate which is small enough to permit a correct exploration of the search space without wasting computational resources (Harik et al., 1999b).

A common approach to compute the optimal convergence probability of an evolutionary algorithm (EA) with finite search sets is to model the algorithm using finite state Markov chains. However, it is barely possible to obtain analytical expressions since the probability matrices of these Markov chains are intractable even for simple optimization problems. Sometimes assumptions regarding the population size, the operators, and the optimization problem help to estimate the optimal convergence probability. These assumptions usually reduce the state space and, therefore, the size of the probability matrices, even turning these matrices into matrices with special properties.

The idea of how to approach a population-based EA with recombination and selection but without mutation is introduced in (Harik et al., 1999b). It is argued that the dynamics of such EAs are similar to the dynamics of specific random walks. The obtained results are based on many approximations without giving any estimation of possible errors. Rudolph (2005) proposes a more solid theoretical foundation upon this argument. His work gives a mathematical model to lower bound the optimal convergence probability of a variation of non-generational EAs, while optimizing the OneMax problem. The approach is still based on modeling the EA using random walks on finite space, yet he employs some estimations which make the argument not completely mathematical sound. Since the cGA mimics the behavior of a binary non-generational EA, then one can use Rudolph's idea to bound the optimal convergence probability of the cGA. However, even if we build a completely rigorous mathematical foundation upon (Harik et al., 1999b; Rudolph, 2005), we cannot study the optimal convergence probability of the PBIL by the same approach since the PBIL cannot be modeled by a finite Markov chain. This motivates us to find a more general approach covering a wider range of EAs.

A broad mathematical framework is considered in (Norman, 1972) that includes stochastic learning models with distance diminishing operators in metric spaces for experiments with finite numbers of responses and simple reinforcement. One main result of this framework is to define superregular and subregular functions and then use them to bound the convergence probability of a learning algorithm to different possible desired actions. In (Lakshmivarahan and Thathachar, 1976), it is shown the distance-diminishing property is not necessary and this method can be used in a wider range of application. This method is applied successfully to many different adaptive systems. See for example (Thathachar and Arvind, 1998) and references herein.

In this paper, we will use this method to lower bound the optimal convergence probability of the cGA and the PBIL. Then we will show that for a specific class of functions the cGA with sufficiently small learning rate and the PBIL with sufficiently small learning rate or large population size converge almost surely to the maximum. Further, using the lower bounds, we will derive some upper bounds on the learning rates and a lower bound on the population size to make sure that algorithms will converge to the optimum with a predefined confidence level. As one will see, the advantage of the approach used in this paper is that it helps us to study several properties of the cGA, the PBIL, and possibly other types of EAs under the same umbrella.

This paper is organized as follows: Section 2 describes the cGA and the PBIL precisely. Section 3 reviews basic mathematical background relevant for this paper. In Section 4 bounds on the optimal convergence probability are computed for the cGA and the PBIL. Lastly, in Section 5, computation is conducted for linear functions and several simulations are given. The paper concludes with insights toward future research.

2 Algorithms

Let $\Omega = \{0,1\}^n$ and $f : \Omega \to \mathbb{R}$ be a pseudo-boolean function. The goal is to maximize f. Assume an EDA represents the probability distribution of the population of individuals by a PV $p(k) = (p_1(k), ..., p_n(k))$ where $p_i(k)$ refers to the probability of obtaining a value of 1 in the ith component of the population of individuals in the kth generation. Let define the initial PV as $p(1) = p^0$ where $p^0 = (0.5, ..., 0.5)$.

A simple EDA is the PBIL introduced by (Baluja and Caruana, 1995). At iteration k, drawing the PV, $p(k)$, N individuals are obtained and λ of these individuals are selected using a selection scheme and named $w^{(1)}(k)$, $w^{(2)}(k)$,..., $w^{(\lambda)}(k)$. These selected individuals are then used to modify the PV according to a Hebbian-inspired rule in the form of

$$p(k+1) = (1-\alpha)\, p(k) + \alpha \frac{1}{\lambda} \sum_{t=1}^{\lambda} w^{(t)}(k) \tag{1}$$

where $\alpha \in (0,1)$ is a learning parameter. In this paper, we use two-tournament selection λ times to find $w^{(t)}(k)$s ($1 \leq t \leq \lambda$) as follows. For each $1 \leq t \leq \lambda$, two random individuals $a(k)$ and $b(k)$ are generated on the basis of $p(k)$ and then compete with each other and $w^{(t)}(k) = a(k), l^{(t)}(k) = b(k)$(resp. $w^{(t)}(k) = b(k), l^{(t)}(k) = b(k)$) when $f(a(k)) \geq f(b(k))$(resp. $f(b(k)) \geq f(a(k))$). Clearly in our case, $\lambda = \frac{N}{2}$.

Harik et al. (1999a) present the cGA belonging to the EDA family. In this algorithm two-tournament selection is used just one time. At the kth iteration of the optimization process, two individuals $a(k)$ and $b(k)$ are generated on the basis of $p(k)$. Then $w(k) =$

$w^{(1)}(k)$ and $l(k) = l^{(1)}(k)$. Thus $p(k)$ is updated as follows:

$$p(k+1) = p(k) + \alpha(w(k) - l(k)) \tag{2}$$

To prevent p_is from getting smaller than 0 or larger than 1, we let α be equal to $1/m$, where m is an even positive integer. The next lemma is useful for our analysis (Hohfeld and Rudolph, 1997; Rastegar and Hariri, 2006b).

Lemma 1. *In 2-tournament selection method, let $P(w^{(t)}(k) = y)$ (resp. $P(l^{(t)}(k) = y)$) be the probability of obtaining y as the winner (resp. loser) individual at the kth iteration. Then*

$$P\left(w^{(t)}(k) = y\right) = P_k(y)\left\{\sum_{f(z)<f(y)} P_k(z) + \sum_{f(z)\leq f(y)} P_k(z)\right\} \tag{3}$$

$$P\left(l^{(t)}(k) = y\right) = P_k(y)\left\{\sum_{f(z)>f(y)} P_k(z) + \sum_{f(z)\geq f(y)} P_k(z)\right\} \tag{4}$$

where $P_k(y)$ denotes the probability of sampling the individual y at iteration k.

It is clear that for a given k, $w^{(i)}$s are independent and identically distributed (i.i.d.) random vectors and therefore $P\left(w^{(i)}(k) = y\right) = P\left(w^{(j)}(k) = y\right)$ for $1 \leq i, j \leq \lambda$.

3 Mathematical Preliminary

In this section, we define (sub,super)regular functions and mention their connection to the convergence probability of a stochastic process to an absorbing state by proving some results similar to those of (Norman, 1972; Lakshmivarahan and Thathachar, 1976) for time-homogeneous Markov processes.

Suppose $\{\xi(k)\}_{k=1}^{\infty}$ is a Markov process with stationary transition probabilities defined on the compact state space S converging almost surely to some points in $A = \{s_0, ..., s_{N-1}\} \subset S$. The topology of S is defined differently for each different S, e.g. the topology of $S = [0,1]^n$ is the usual topology inherited from $\mathbb{R}^n$ and the topology of $S = \{0, \alpha, 2\alpha, ..., 1\}^n$ is the discrete topology in which any point of S is considered to be an open set. Clearly continuity of a function on S is defined based on the topology of S, thus, in $S = [0,1]^n$ with usual topology the continuity of a function is defined as it is defined in elementary mathematics, however, in $S = \{0, \alpha, 2\alpha, ..., 1\}^n$ with discrete topology all functions are continuous. Let $C(S)$ be the space of all continuous functions from S to $\mathbb{R}$. Since S is compact, every function in $C(S)$ is bounded. Let $A_1, A_2, ..., A_r$ be a partition of A. Given an $1 \leq i \leq r$, define $\Gamma_{A_i}(s) = P\left(\lim_{k\to\infty} \xi(k) \in A_i | \xi(1) = s\right)$ as the probability that $\xi(k)$ converges to some element in A_i provided that the initial value of $\xi(1)$ is s.

If $\psi(.) \in C(S)$, the operator U is defined by

$$U\psi(s) = E\left\{\psi(\xi(k+1)) | \xi(k) = s\right\}$$

for $k \geq 1$. It may noted that U is linear and preserves non-negative function. Further

$$U^k\psi(s) = UU^{k-1}\psi(s) = E\left\{\psi(\xi(k)) | \xi(1) = s\right\}$$

for all $k > 1$ and $U^1\psi(s) = U\psi(s)$. The following lemma shows that $\Gamma_{A_i}(.)$ $(i = 1, ..., r)$ satisfies a functional equation with appropriate boundary conditions.

Lemma 2. *$\Gamma_{A_i}(.)$ is the only continuous solution of the functional equation $U\Gamma_{A_i} = \Gamma_{A_i}$ with the boundary conditions $\Gamma_{A_i}(s) = 1$ if $s \in A_i$ and $\Gamma_{A_i}(s) = 0$ otherwise.*

Proof. It is clear that $\Gamma_{A_i}(.)$ is a continuous solution of $U\Gamma_{A_i} = \Gamma_{A_i}$ with above boundary conditions. Let $h \in C(S)$ be another continuous solution where $h(s) = 1$ for $s \in A_i$ and $h(s) = 0$ otherwise. Since h is a bounded function, then for a given $s \in S$, $\left\{U^k h(s)\right\}_{k=1}^{\infty}$ is a sequence of bounded real numbers. Thus by Bolzano-Weierstrass Theorem there is a convergent subsequence $\left\{U^{k_j} h(s)\right\}_{j=1}^{\infty}$. Now an application of Bounded Convergence Theorem (Chung, 2000) in (5) gives

$$
\begin{aligned}
h(s) &= Uh(s) = \ldots = U^{k_1}h(s) = \ldots = \lim_{j\to\infty} U^{k_j}h(s) \\
&= \lim_{j\to\infty} E\left\{h(\xi(k_j))|\xi(1) = s\right\} \\
&= E\left\{\lim_{j\to\infty} h(\xi(k_j))|\xi(1) = s\right\} \qquad (5) \\
&= E\left\{h(\lim_{j\to\infty} \xi(k_j))|\xi(1) = s\right\} \\
&= E\left\{h(\lim_{k\to\infty} \xi(k))|\xi(1) = s\right\} \qquad (6) \\
&= \Sigma_{s'\in A} h(s')P(\lim_{k\to\infty} \xi(k) = s'|\xi(1) = s) \\
&= \Sigma_{s'\in A_i} P(\lim_{k\to\infty} \xi(k) = s'|\xi(1) = s) = \Gamma_{A_i}(s),
\end{aligned}
$$

which (6) comes from the fact that the each subsequence of a almost surely convergent sequence converges almost surely to the same limit random variable. □

Since solving such an equation is a difficult task, an attempt is made to determine bounds on $\Gamma_{A_i}(s)$ $(i = 1, ..., r)$ which satisfy functional inequalities. In this context subregular and superregular functions are defined. The function $\psi(.) : S \to \mathbb{R}$ is a subregular (resp. superregular) function if and only if $U\psi(s) \geq \psi(s)$ (resp. $U\psi(s) \leq \psi(s)$) for all $s \in S$.

Lemma 3. *If $\psi \in C(S)$ is subregular (resp. superregular) with $\psi(s) = 1$ when $s \in A_i$ and $\psi(s) = 0$ when $s \notin A_i$, then $\psi(s) \leq \Gamma_{A_i}(s)$ (resp. $\psi(s) \geq \Gamma_{A_i}(s)$) for all $s \in S$.*

Proof. Let ψ is subregular. Then like before there is a convergent subsequence $\left\{U^{k_j} h(s)\right\}_{j=1}^{\infty}$ and hence

$$
\begin{aligned}
\psi(s) &\leq U\psi(s) \leq \cdots \leq \lim_{j\to\infty} U^{k_j}\psi(s) \\
&= \lim_{j\to\infty} E\left\{\psi\left(\xi(k_j)\right)|\xi(1) = s\right\} \\
&= E\left\{h(\lim_{j\to\infty} \xi(k_j))|\xi(1) = s\right\} \\
&= E\left\{h(\lim_{k\to\infty} \xi(k))|\xi(1) = s\right\} \\
&= \Sigma_{s'\in A} h(s')P(\lim_{k\to\infty} \xi(k) = s'|\xi(1) = s) \\
&= \Sigma_{s'\in A_i} P(\lim_{k\to\infty} \xi(k) = s'|\xi(1) = s) = \Gamma_{A_i}(s).
\end{aligned}
$$

The same argument works when ψ is superregular. □

Lemma 3 reduces the problem of obtaining bounds on $\Gamma_{A_i}(s)$ to finding subregular and superregular functions with appropriate boundary conditions. No general method of identifying superregular and subregular functions is known. One has to start with a promising functional form and evaluate the parameters of the function so that the required inequality is satisfied. Finding a promising functional form and the best values for its the parameters is the most difficult part of the procedure. The following lemma can be usefull to simplify this procedure.

Lemma 4. *Let $\psi_i \in C(S)$ be monotonic increasing subregular functions, then $\prod \psi_i(.)$ is a subregular function.*

Proof. The application of the Chebyshev Integral Inequality (Tong, 1997) implies

$$\prod_{i=1}^{n} U\psi_i(s) \leq U \prod_{i=1}^{n} \psi_i(s) = U\psi(s).$$

Considering the subregularity of $\psi_i(.)$ shows $\psi(s) \leq U\psi(s)$. □

Using the above lemma in finding the subregular function leads us to more conservative result, however, it reduces the difficulty of problem.

4 Optimal Convergence Probability

In this section, an application of Lemma 3 provides some bounds on the optimal convergence probability of the cGA and the PBIL for a class of binary functions.

Definition (Property 1)**.** A function $f : \Omega \to \mathbb{R}$ satisfies Property 1 if $f(x \vee e_i) > f(x \wedge \overline{e}_i)$ for all $x \in \Omega$ and $1 \leq i \leq n$ where e_i is the i-th unit vector with dimension of n and $\overline{e}_i$ its binary complement and $\wedge$ and $\vee$ are component-wise "AND" and "OR", respectively.

All linear functions $f(x) = \Sigma_{i=1}^{n} w_i x_i$ with $w_i > 0$ have Property 1. There are also some nonlinear functions such as $f(x) = 2\Sigma_{i=1}^{n} w_i x_i + \prod_{i=1}^{n} x_i$ having property. From this point forward, we assume that f satisfies Property 1. Clearly $x^* = (1, ..., 1)$ is the only global maximum of f.

Let the random sequence $\{p(k)\}_{k=1}^{\infty}$ be generated by the cGA while optimizing function f. To bound the optimal convergence probability of the cGA for f, we first prove that $\{p(k)\}_{k=1}^{\infty}$ will converge to a point in Ω. Following the idea developed in (Hohfeld and Rudolph, 1997) the next lemma is given.

Lemma 5. *$\{p_d(k)\}_{k=1}^{\infty}$ is a submartingale for a given $1 \leq d \leq n$.*

Proof. Equation (2) implies $E\left[p_d(k+1)\,|p(k)\right] = p_d(k) + \alpha E\left[w_d(k) - l_d(k)|p(k)\right]$ for all $1 \leq d \leq n$. Since f satisfies Property 1, for a given $x \in \Omega$, $f(x \vee e_d) > f(x \wedge \overline{e}_d)$. Hence

$$\Sigma_{f(z)<f(x\vee e_d)} P_k(z) \quad \geq \quad \Sigma_{f(z)<f(x\wedge \overline{e}_d))} P_k(z) \tag{7}$$

$$\Sigma_{f(z)\leq f(x\vee e_d)} P_k(z) \quad \geq \quad \Sigma_{f(z)\leq f(x\wedge \overline{e}_d))} P_k(z) \tag{8}$$

$$\Sigma_{f(z)>f(x\vee e_d)} P_k(z) \quad \leq \quad \Sigma_{f(z)>f(x\wedge \overline{e}_d))} P_k(z) \tag{9}$$

$$\Sigma_{f(z)\geq f(x\vee e_d)} P_k(z) \quad \leq \quad \Sigma_{f(z)\geq f(x\wedge \overline{e}_d))} P_k(z) \tag{10}$$

Then based on Lemma 1 we have

$$\begin{aligned} P(w(k) = x \vee e_d)/P_k(x \vee e_d) &= \sum_{f(z)<f(x\vee e_d)} P_k(z) + \sum_{f(z)\leq f(x\vee e_d)} P_k(z) \\ &\geq \sum_{f(z)<f(x\wedge \overline{e}_d)} P_k(z) + \sum_{f(z)\leq f(x\wedge \overline{e}_d)} P_k(z) \\ &= P(w(k) = x \wedge \overline{e}_d)/P_k(x \wedge \overline{e}_d), \end{aligned} \tag{11}$$

and in a similar way

$$P(l(k) = x \vee e_d)/P_k(x \vee e_d) \leq P(l(k) = x \wedge \bar{e}_d)/P_k(x \wedge \bar{e}_d). \tag{12}$$

Define $q_d(x,k) = \prod_{j=1,j\neq d}^{n} p_j(k)^{x_j}(1-p_j(k))^{1-x_j}$. It is easy to see that $P_k(x \wedge \bar{e}_d) = (1-p_i(k))q_i(x,k)$ and $P_k(x \vee e_d) = p_i(k)q_i(x,k)$. Insertion of these identities into the inequalities (11) and (12) and some simplification show that

$$\begin{array}{rcl} P\left(w(k) = x \vee e_d\right) & \geq & p_d(k)\left(P(w(k) = x \wedge \bar{e}_d) + P(w(k) = x \vee e_d)\right) \\ P\left(l(k) = x \vee e_d\right) & \leq & p_d(k)\left(P(l(k) = x \wedge \bar{e}_d) + P(l(k) = x \vee e_d)\right). \end{array}$$

Thus, the above inequalities conclude

$$\begin{array}{rcl} E\left\{p_d(k+1)\left|p(k)\right.\right\} - p_d(k) & = & \alpha E\left\{w_d(k) - l_d(k)\left|p(k)\right.\right\} \\ & = & \alpha \sum_{x\in\Omega} x_d\left(P(w(k)=x) - P(l(k)=x)\right) \\ & = & \frac{\alpha}{2}\sum_{x\in\Omega}\left(P(w(k)=x\vee e_d) - P(l(k)=x\vee e_d)\right) \\ & \geq & \frac{\alpha}{2}p_d(k)\sum_{x\in\Omega}\left(P(w(k)=x\wedge\bar{e}_d) + P(w(k)=x\vee e_d)\right) \\ & - & \frac{\alpha}{2}p_d(k)\sum_{x\in\Omega}\left(P(l(k)=x\wedge\bar{e}_d) + P(l(k)=x\vee e_d)\right) \\ & = & \frac{\alpha}{2}p_d(k)\sum_{x\in\Omega}2P(w(k)=x) - \frac{\alpha}{2}p_d(k)\sum_{x\in\Omega}2P(l(k)=x) \\ & = & \alpha p_d(k) - \alpha p_d(k) = 0. \end{array}$$

This completes the proof. □

The following corollary is a direct conclusion of the above lemma.

Corollary 6. *For every $1 \leq d \leq n$, $\lim_{k\to\infty} p_d(k) = p_d^*$ exists and $p_d^* \in \{0,1\}$ almost surely.*

Proof. Lemma 5 implies that $\{p_d(k)\}$ is a submartingale. Since this submartingale is positive and uniformly bounded, then Martingale theorem (Chung, 2000) asserts that $\lim_{k\to\infty} p_d(k) = p^*$ exists almost surely. Further, if $p_d^* \notin \{0,1\}$, then $p_d^*(k) \neq p_d^*(k+1)$ with a non-zero probability for all k which is a contradiction. Hence $p_d^* \in \{0,1\}$ and $\{0,1\}$ forms the absorbing states for the Markov process $\{p_d(k)\}$. □

Thus for each $1 \leq d \leq n$, $\{p_d(k)\}_{k=1}^{\infty}$ converges almost surely to $\{0,1\}$ and consequently $\{p(k)\}_{k=1}^{\infty}$ converges almost surely to Ω. Let define

$$\begin{array}{rcl} H_d\left(k\right) & = & P\left(f\left(a\left(k\right)\right) \geq f\left(b\left(k\right)\right)\left|a_d\left(k\right)=1, b_d\left(k\right)=0\right.\right) \\ & + & P\left(f\left(b\left(k\right)\right) > f\left(a\left(k\right)\right)\left|b_d\left(k\right)=1, a_d\left(k\right)=0\right.\right), \end{array} \tag{13}$$

then

$$\begin{array}{rcl} P(w_d(k) - l_d(k) = 1|p(k)) & = & P\left\{f(a) \geq f(b), a_d = 1, b_d = 0\right\} \\ & + & P\left\{f(b) > f(a), a_d = 0, b_d = 1\right\} \\ & = & 2H_d(k)p_d(k)\left(1-p_d(k)\right). \end{array} \tag{14}$$

$$\begin{array}{rcl} P(w_d(k) - l_d(k) = 0|p(k)) & = & P\left(w_d(k)=1, l_d(k)=1|p(k)\right) \\ & + & P\left(w_d(k)=0, l_d(k)=0|p(k)\right) \\ & = & 1 - 2p_d(k)\left(1-p_d(k)\right). \end{array} \tag{15}$$

$$\begin{aligned} P(w_d(k) - l_d(k) = -1|p(k)) &= 1 - P(w_d(k) - l_d(k) = 0|p(k)) \\ &- P(w_d(k) - l_d(k) = 1|p(k)) \\ &= 2(1 - H_d(k))p_d(k)\,(1 - p_d(k))\,. \end{aligned} \tag{16}$$

Notice that

$$\begin{aligned} E\left\{p_d(k+1) - p_d(k)|p(k)\right\} &= \alpha P\left(w_d(k) - l_d(k) = 1|p(k)\right) \\ &- \alpha P\left(w_d(k) - l_d(k) = -1|p(k)\right) \\ &= \alpha\left(2H_d(k) - 1\right)p_d(k)\left(1 - p_d(k)\right). \end{aligned} \tag{17}$$

By lemma 5, the left-side hand of (17) is always non-negative, therefore one has $1 \leq 2H_d(k)$. At this point, we are ready to apply the results of the section 3 to find a bound on the optimal convergence probability of the cGA. First, $\{p(k)\}_{k=1}^{\infty}$ is a time-homogeneous markov chain with the compact state set $S = \{0, \alpha, 2\alpha, ..., 1\}^n$ converging almost surely to $A = \Omega$. Lets partition A to two sets of the optimal point, $A_1 = \{(1, ..., 1)\}$, and non-optimal points, $A_2 = \Omega - A_1$, then the optimal convergence probability of the cGA will be $\Gamma_{A_1}((0.5, ..., 0.5))$, the probaility that $\{p(k)\}$ converges to $(1, ..., 1)$.

The next important step is to find an appropriate functional form, $\psi(.) : S \to \mathbb{R}$, s.t. $\psi(.)$ has the same boundary values as $\Gamma_{A_1}(.)$, that is $\psi(p) = 1$ for $p \in A_1$ and $\psi(p) = 0$ otherwise. The first candidate for such a functional form is

$$\psi(p) = \frac{1 - e^{-b\prod_{d=1}^{n} p_d}}{1 - e^{-b}},$$

where $b > 0$ is to be chosen. In this case, the best value for b giving a tight lower bound is the largest value for which $U\psi(p) \geq \psi(p)$ holds, i.e. $\psi(.)$ is a subregular function. To compute the largest value of b, we need to have transition probability matrix of the markov process $\{p(k)\}_{k=1}^{\infty}$. However, this matrix is intractable, even for simple optimization functions and accordingly we need to use another functional form such as

$$\psi(.) = \prod_{d=1}^{n} \psi_d(p), \tag{18}$$

with

$$\psi_d(p) = \frac{1 - e^{-b_d p_d}}{1 - e^{-b_d}} \tag{19}$$

where for each $1 \leq d \leq n$, p_d is the dth component of p and $b_d > 0$ is to be chosen. Since $\psi_d(.)$s are continuous, then $\psi(.) \in C(S)$. Again, the best value for b_ds are the largest values for which $\psi_d(.)$ are subregular functions. A direct computation of b_is in inequality $U\psi(p) \geq \psi(p)$ is a tedious task, however, finding the b_ds for which ψ_d are subregular is simple.

Lemma 7. *Lets define $\psi_d : S \to \mathbb{R}$ as in (19) and $H_d = \min_k H_d(k)$. If $H_d \neq 1$ then $\psi_d(.)$ is subregular provided that $b_d \leq \frac{1}{\alpha}\ln\frac{H_d}{1-H_d}$. If $H_d = 1$, then $\psi_d(.)$ is subregular for all $b_d > 0$.*

Proof. Some computations and using (14)-(16) give

$$\begin{aligned}
U\psi_d(p) - \psi_d(p) &= E\left\{\psi_d(p_d(k+1))|p(k) = p\right\} - \psi_d(p) \\
&= E\left\{\frac{1-e^{-b_d p_d(k+1)}}{1-e^{-b_d}}|p(k)\right\} - \frac{1-e^{-b_d p_d(k)}}{1-e^{-b_d}} \\
&= \frac{1}{1-e^{-b_d}}\left(e^{-b_d p_d(k)} - E\left\{e^{-b_d p_d(k+1)}|p(k)\right\}\right) \\
&= \frac{1}{1-e^{-b_d}}\left(e^{-b_d p_d(k)} - E\left\{e^{-b_d p_d(k) - b_d\alpha(w_d(k)-l_d(k))}|p(k)\right\}\right) \\
&= \frac{1}{1-e^{-b_d}}\left(e^{-b_d p_d(k)} - e^{-b_d p_d(k)} E\left\{e^{-b_d\alpha(w_d(k)-l_d(k))}|p(k)\right\}\right) \\
&= \frac{e^{-b_d p_d(k)}}{1-e^{-b_d}}\left\{1 - P\left(w_d(k) - l_d(k) = 1|p(k)\right)e^{-b_d\alpha}\right. \\
&+ \left. P\left(w_d(k) - l_d(k) = -1|p(k)\right)e^{b_d\alpha} + P\left(w_d(k) - l_d(k) = 0|p(k)\right)\right\} \\
&= 2\frac{e^{-b_d p_d(k)}}{1-e^{-b_d}} p_d(k)\left(1 - p_d(k)\right)\left(1 - H_d(k)e^{-b_d\alpha} - \left(1 - H_d(k)\right)e^{b_d\alpha}\right).
\end{aligned}$$

Hence, $\psi_d(.)$ is a subregular function if $1 \geq H_d(k)e^{-b_d\alpha} + (1 - H_d(k))\, e^{b_d\alpha}$ or equivalently

$$\left(1 - H_d(k)\right)e^{2b_d\alpha} - e^{b_d\alpha} + H_d(k) \leq 0. \tag{20}$$

If $H_d(k) = 1$, the inequality trivially holds. Suppose $H_d(k) < 1$. Since $2H_d(k) \geq 1$, solving (20) shows

$$\begin{aligned}
e^{b_d\alpha} &\leq \frac{1 + \sqrt{1 - 4H_d(k)\left(1 - H_d(k)\right)}}{2\left(1 - H_d(k)\right)} \\
&= \frac{1 + \sqrt{\left(2H_d(k) - 1\right)^2}}{2\left(1 - H_d(k)\right)} = \frac{H_d(k)}{1 - H_d(k)}.
\end{aligned} \tag{21}$$

By inequality (21), $\psi(.)$ is subregular if

$$b_d \leq \frac{1}{\alpha} \min_{1 \leq k < \infty} \ln \frac{H_d(k)}{1 - H_d(k)} = \frac{1}{\alpha} \ln \frac{H_d}{1 - H_d}$$

which completes the proof. □

The following main theorem is a direct result of the lemmas 4 and 3.

Theorem 8. *Let $p^0 = (0.5, ..., 0.5)$ be the initial PV and x^* be the optimal solution. Then*

$$\prod_{d=1}^{n}\left(1 + \left(\frac{1 - H_d}{H_d}\right)^{\frac{1}{2\alpha}}\right)^{-1} \leq \Gamma_{A_1}(p^0) = P(\lim_{k\to\infty} p(k) = x^*|p(1) = p^0) \tag{22}$$

Proof. Let $\psi(.)$ be defined as in (18). One sees that since $\psi_d(.)$s are monotonic increasing, by Lemma 4, $\psi(.)$ is subregular if each $\psi_d(.)$ is subregular. Therefore, according to lemmas 3 and 7 we have

$$\begin{aligned}
\Gamma_{A_1}(p^0) &\geq \prod_{d=1}^{n} \psi_d(p^0) = \prod_{d=1}^{n} \frac{1 - e^{-\frac{b_d}{2}}}{1 - e^{-b_d}} \\
&= \prod_{d=1}^{n} \frac{1}{1 + e^{\frac{-b_d}{2}}} = \prod_{H_d \neq 1} \frac{1}{1 + e^{\frac{-1}{2\alpha} \ln \frac{H_d}{1 - H_d}}} = \prod_{d=1}^{n} \frac{1}{1 + \left(\frac{1 - H_d}{H_d}\right)^{\frac{1}{2\alpha}}},
\end{aligned}$$

which completes the proof. □

Remark. A similar result is reported in (Rudolph, 2005) for binary non-generational evolutionary algorithm (the cGA) optimizing the OneMax problem. However, there are two questionable points in the argument. To understand these points, we review the argument. Each component $p_d(k)$ of the probability vector is modeled by a random walk on $S = \{0, 1, 2, ..., m\}$ where $m = \alpha^{-1}$. Let $P_{i,i+1}(d,k)$, $P_{i,i-1}(d,k)$, and $P_{i,i}(d,k)$ be the probabilities that $p_d(k+1) = p_d(k)+\alpha$, $p_d(k+1) = p_d(k)-\alpha$, and $p_d(k+1) = p_d(k)$ when $p_d(k) = i\alpha$. $P_{i,i+1}(d,k)$, $P_{i,i-1}(d,k)$, and $P_{i,i}(d,k)$ form transition probabilities of the dth random walk with $m+1$ states $0, 1, ..., m$. $1, ..., m-1$ are the transient states and 0 and m are the absorbing states of the random walk. Thus we have

$$\begin{aligned}
P_{i,i}(d,k) &= 1 - 2i\alpha\,(1 - i\alpha)\,, \\
P_{i,i+1}(d,k) &= 2i\alpha\,(1 - i\alpha)\,H_d(k), \\
P_{i,i-1}(d,k) &= 2i\alpha\,(1 - i\alpha)\,(1 - H_d(k)), \quad \forall 1 < i < m \\
P_{0,0}(d,k) &= 1, \\
P_{m,m}(d,k) &= 1.
\end{aligned}$$

Cleary, these random walks are state-dependent time-inhomogeneous Markov processes. Replacing the transition probabilities of these random walk with some new transition probabilities

$$\begin{aligned}
\widetilde{P}_{i,i+1}(d,k) &= H_d(k), \\
\widetilde{P}_{i,i-1}(d,k) &= 1 - H_d(k), \\
\widetilde{P}_{i,i}(d,k) &= 0, \quad \forall 1 < i < m \\
\widetilde{P}_{0,0}(d,k) &= 1, \\
\widetilde{P}_{m,m}(d,k) &= 1
\end{aligned}$$

gives n new random walks with the same absorption probability for state 0 and m as in the original random walks. The first fallacy arises when the author uses "Equation (1)" of the paper derived originally for absorption probability of a time homogeneous random walk to obtain the absorption probability of the new random walks, clearly not time-homogeneous. At the end, it is also concluded that a lower bound on the probability that $\{p(k)\}$ converges to $(1, ..., 1)$ is the product of lower-bounds on probabilities that random walks $\{p_d(k)\}$ converge to 1, however, since $P(\lim_{k\to\infty} p_d(k) = 1|p(1) = p^0) \geq P(\lim_{k\to\infty} p_1(k) = 1, ..., \lim_{k\to\infty} p_n(k) = 1|p(1) = p^0)$ for each $1 \leq d \leq n$ it is not clear how to lower bound $P(\lim_{k\to\infty} p(k) = x^*|p(1) = p^0)$ by lower-bounding $P(\lim_{k\to\infty} p_d(k) = 1|p(1) = p^0)$.

The bound on the optimal convergence probability can be utilized to show that for sufficient small α the cGA converges almost surely to the optimal solution of functions with Property 1. As $H_d > \frac{1}{2}$, $\frac{1-H_d}{H_d} < 1$ for all $1 \leq d \leq n$. Thus letting $\alpha \to 0$ in Theorem 8 completes the argument. Since some of the functions with Property 1, such as the OneMax, are not injective, this result can be considered a complementary result for (Rastegar and Hariri, 2006b).

Theorem 8 can further be used to determine a conservative range of possible values of the learning rate for which the cGA converges to the optimal solution with a confidence level $0 < \beta < 1$. It is clear that if

$$0 < \alpha \leq \min_{H_d < 1} \frac{\ln(1 - H_d) - \ln H_d}{2\ln(\beta^{-\frac{1}{n}} - 1)},$$

then Theorem 8 concludes $\beta \leq P(\lim_{k\to\infty} p(k) = x^*|p(1) = p^0)$. This estimate is conservative, and we underestimate the actual range of values for the learning rate.

In the remainder of this section, we obtain a lower bound for the optimal convergence probability of the PBIL. Let the random sequence $\{p(k)\}_{k=1}^{\infty}$ be generated by the PBIL while optimizing f. The state set of the time-homogeneous Markov process $\{p_d(k)\}_{k=1}^{\infty}$ is the compact set $S = [0,1]^n$. With a similar argument to that of Lemma 5, we can show for a given $1 \leq d \leq n$, $\{p_d(k)\}_{k=1}^{\infty}$ is a submartingale, $\lim_{k\to\infty} p_d(k) = p_d^*$ exists, and $p_d^* \in \{0,1\}$ almost surely. Therefore the absorbing set of $\{p(k)\}_{k=1}^{\infty}$ is Ω, i.e. $A = \Omega$. Define A_1 and A_2 as before. A promising subregular function for computing a bound on the optimal probability probability of the PBIL could be (18) where $b_d > 0$s are to be chosen. One shows

$$
\begin{aligned}
U\psi_d(p) \quad - \quad & \psi_d(p) = E\left\{\psi_d(p_d(k+1))|p(k) = p\right\} - \psi_d(p) \\
= \quad & E\left\{\frac{1-e^{-b_d p_d(k+1)}}{1-e^{-b_d}}|p(k)\right\} - \frac{1-e^{-b_d p_d(k)}}{1-e^{-b_d}} \\
= \quad & \frac{1}{1-e^{-b_d}}\left(e^{-b_d p_d(k)} - E\left\{e^{-b_d p_d(k+1)}|p(k)\right\}\right) \\
= \quad & \frac{1}{1-e^{-b_d}}\left(e^{-b_d p_d(k)} - E\left\{e^{-b_d(1-\alpha)p_d(k) - \frac{b_d\alpha}{\lambda}\Sigma_{t=1}^{\lambda} w_d^{(t)}(k)}|p(k)\right\}\right) \\
= \quad & \frac{1}{1-e^{-b_d}}\left(e^{-b_d p_d(k)} - e^{-b_d(1-\alpha)p_d(k)}\prod_{t=1}^{\lambda} E\left\{e^{-\frac{b_d\alpha}{\lambda}w_d^{(t)}(k)}|p(k)\right\}\right) \\
= \quad & \left(1 - e^{b_d\alpha p_d(k)}\prod_{t=1}^{\lambda}\left(P\left(w_d^{(t)}(k) = 1|p(k)\right)e^{\frac{-b_d\alpha}{\lambda}} + P\left(w_d^{(t)}(k) = 0|p(k)\right)\right)\right) \\
& \times\frac{e^{-b_d p_d(k)}}{1-e^{-b_d}}.
\end{aligned}
$$

Since for all i, j, k

$$
P\left(w_d^{(i)}(k) = 1|p(k)\right) = P\left(w_d^{(j)}(k) = 1|p(k)\right),
$$

we define $G_d(k) = P\left(w_d^{(1)}(k) = 1|p(k)\right)$. Therefore the most right hand side of above expression is

$$
\frac{e^{-b_d p_d(k)}}{1-e^{-b_d}}\left(1 - e^{b_d\alpha p_d(k)}\left(G_d(k)e^{\frac{-b_d\alpha}{\lambda}} + 1 - G_d(k)\right)^{\lambda}\right).
$$

For a given k, lets define

$$
u(b_d, k) = 1 - e^{b_d\alpha p_d(k)}\left(G_d(k)e^{\frac{-b_d\alpha}{\lambda}} + 1 - G_d(k)\right)^{\lambda}. \tag{23}
$$

The fact that $G_d(k) = p_d^2(k) + 2p_d(k)(1-p_d(k))H_d(k)$ shows that $G_d(k) = 1$ (resp. $G_d(k) = 0$) if and only if $p_d(k) = 1$ (resp. $p_d(k) = 0$). In these cases $u(b_d, k) = 0$ for all value b_d. Assume $0 < G_d(k) < 1$ and $0 < p_d(k) < 1$. For a given k, computing the first derivative of $u(b_d, k)$ with respect to b_d, we have

$$
\begin{aligned}
\frac{\partial u(b_d,k)}{\partial b_d} \quad = \quad & \alpha e^{b_d\alpha p_d(k)}\left(G_d(k)e^{-\frac{b_d\alpha}{\lambda}} + 1 - G_d(k)\right)^{\lambda-1} \\
\times \quad & \left(G_d(k)e^{-\frac{b_d\alpha}{\lambda}}(1-p_d(k)) - (1-G_d(k))p_d(k)\right).
\end{aligned}
$$

Solving $\frac{\partial u(b_d,k)}{\partial b_d}=0$ shows that $u(b_d,k)$ has only one critical point at

$$\begin{aligned} b_d^*(k) &= \frac{\lambda}{\alpha}\ln\frac{(1-p_d(k))\,G_d(k)}{p_d(k)\,(1-G_d(k))} \\ &= \frac{\lambda}{\alpha}\ln\frac{p_d(k)+2(1-p_d(k))H_d(k)}{1+p_d(k)-2p_d(k)H_d(k)}. \end{aligned}$$

Substituting $b_d^*(k)$ in (23) and simplifying, we have

$$\begin{aligned} 1-u(b_d^*(k),k) &= (p_d(k)+2\,(1-p_d(k))\,H_d(k))^{\lambda p_d(k)} \\ &\times\ (1+p_d(k)-2p_d(k)H_d(k))^{\lambda(1-p_d(k))}. \end{aligned}$$

An application of generalized A-G means inequality to the right-hand side of the above equality implies that right-hand side of above inequality is less than or equal to

$$\left(\frac{\lambda p_d(k)\,(p_d(k)+2(1-p_d(k))H_d(k))+\lambda(1-p_d(k))\,(1+p_d(k)-2p_d(k)H_d(k))}{\lambda}\right)^{\lambda}=1,$$

meaning that $u(b_d^*(k),k)\geq 0$. Suppose that there is a $b'\in(0,b_d^*(k))$ such that $u(b',k)<0$. Since $u(0,k)=0$ and $u(b_d^*(k),k)\geq 0$, by continuity of $u(.,k)$ with respect to b_d in $(0,b_d^*(k))$, there is at least a local minimum (i.e. a critical point) for $u(.,k)$ which is a contradiction since $b_d^*(k)$ is the only critical point of $u(.,k)$. Thus, $u(b',k)\geq 0$ for all $b'\in(0,b_d^*(k))$. On the other hand, for each d, ψ_d is subregular if $u(b_d,k)\geq 0$ for all k. Therefore, ψ_d is subregular if $0<b_d\leq\inf_k b_d^*(k)$. At this point, one needs to compute $\inf_k b_d^*(k)$. Some computation shows that for a given k

$$\begin{aligned} \frac{\partial b_d^*(k)}{\partial H_d(k)} &= \frac{2\lambda}{\alpha\,(1+p_d(k)-2p_d(k)H_d(k))\,(p_d(k)+2(1-p_d(k))H_d(k))}>0 \\ \text{and} & \\ \frac{\partial b_d^*(k)}{\partial p_d(k)} &= \frac{(2H_d(k)-1)^2}{(1+p_d(k)-2p_d(k)H_d(k))\,(p_d(k)+2(1-p_d(k))H_d(k))}\geq 0. \end{aligned}$$

Thus $b_d^*(k)$ is an increasing function with respect to $H_d(k)$ and $p_d(k)$, implying that $b_d^*(k)$ attains its minimum value, $\frac{\lambda}{\alpha}\ln 2H_d$, when $H_d(k)=H_d$ and $p_d(k)\to 0$. Thus, an argument similar to that of Theorem 8 shows that by selecting $b_d=\frac{\lambda}{\alpha}\ln 2H_d$, for each $1\leq d\leq n$, we have

$$\prod_{d=1}^{n}\left(1+\left(\frac{1}{2H_d}\right)^{\frac{\lambda}{2\alpha}}\right)^{-1}\leq\Gamma_{A_1}(p^0)=P(\lim_{k\to\infty}p(k)=x^*|p(1)=p^0). \tag{24}$$

Letting $\frac{\alpha}{\lambda}\to 0$ shows that for sufficiently small α or large λ, the PBIL converges almost surely to the optimal solution for functions with Property 1, a complementary result to (Gonzalez et al., 2000; Rastegar and Hariri, 2006a). Again, one computes a conservative range of possible values of the ratio of the learning rate and the population size for which the PBIL converges to the optimal solution with a confidence level $0<\beta<1$. Some computation shows that if

$$0<\frac{\alpha}{\lambda}\leq\min_{1\leq d\leq n}\frac{-\ln 2H_d}{2\ln(\beta^{-\frac{1}{n}}-1)}$$

then $\beta\leq P(\lim_{k\to\infty}p(k)=x^*|p(1)=p^0)$.

Remark. The maximum value computed for each b_d for the cGA is optimal in the sense that if $b_d > \frac{1}{\alpha}\ln\frac{H_d}{1-H_d}$, then (20) does not hold anymore, however, in the PBIL case, $b_d = \frac{\lambda}{\alpha}\ln 2H_d$ is not the optimal possible value for b_d and one can improve the bounds for the optimal convergence probability of the PBIL by finding the maximum value of b_d for which $u(b_d, k) \geq 0$.

Remark. Convergence of the PBIL is first studied in (Hohfeld and Rudolph, 1997) for a linear function with maximum point x^*. Assuming $p(1) \in (0,1)^n$ and $\alpha \in (0,1)$, it is argued that since $E\{p_d(k)\}$ is strictly monotonic when $0 < p_d(k) < 1$ for $1 \leq d \leq n$ and $E\{p_d(k)\}$ is bounded above by unity, then $p_d(k)$ converges in mean (and also almost surely) to x_d^*. However, it is proven in (Gonzalez et al., 2001) that for a 2-bit OneMax problem, $\{(p_1(k), p_2(k))\}_{k=1}^{\infty}$ converges "almost surely" to $(0,0)$ if α and $(p_1(1), p_2(1))$ are selected very close to 1 and $(0,0)$, respectively. This counterexample shows that the argument in (Hohfeld and Rudolph, 1997) is not correct for all values of $\alpha \in (0,1)$. The fallacy lies in assuming that a strictly monotonic sequence tends to x_d^* (unproven Theorem 2, same paper).

5 Computation of H_ds and Experimental Verification

In all results obtained in the last section knowing H_ds for a given function is essential. This section gives some results regarding computation of H_ds.

The following lemma shows that one computes H_ds for linear functions when the PV is uniform. Let define $A(I,k) = \Sigma_{i \notin I} w_i(a_i(k) - b_i(k))$ for a subset $I \subset \{1, ..., n\}$. For the sake of convenience in notation, we assume that w_is are natural numbers. However, with some adjustment in the notations following lemma holds for all positive real w_is.

Lemma 9. *Let $f(x) = \sum_{i=1}^{n} w_i x_i$ be a binary linear function with $w_i > 0$. Then $H_d(k)$ is minimal if $p_j = 0.5$ for all $j \neq d$.*

Proof. First note that

$$\begin{aligned} 2H_d(k) &= P\left(A\left(\{d\}, k\right) \geq -w_d\right) + P\left(A\left(\{d\}, k\right) < w_d\right) \\ &= 1 + P\left(-w_d \leq A\left(\{d\}, k\right) < w_d\right). \end{aligned}$$

Since $H_d(k)$ is a continuous function on the compact set $[0,1]^{n-1}$, it has minimum and maximum in $[0,1]^{n-1}$. Let $\widetilde{p}_i(k)$ be a vector obtained by deleting the i-th component of $p(k)$. It is easy to see that if, at iteration k_0, some components of $\widetilde{p}_d(k_0)$ are in $\{0,1\}$ and others are the same as those of $\widetilde{p}_d(k)$, then $H_d(k_0) \geq H_d(k)$. Thus, the minimum of $H_d(k)$ is a point $q \in (0,1)^{n-1}$. Suppose that $\widetilde{p}_d(k) \in (0,1)^{n-1}$. Let $z_j(k) = a_j(k) - b_j(k)$, then

$$P(z_j(k) = -1) = P(z_j(k) = 1) = \frac{1 - P(z_j(k) = 0)}{2} = p_j(k)(1 - p_j(k)).$$

Fix d such that $1 \leq j \neq d \leq n$ and $k \geq 1$. Then $H_d(k)$ can be rewritten as follows

$$\begin{aligned} 2H_d(k) - 1 &= P(-w_d \leq A(\{d\}, k) < w_d) \\ &= \sum_{i=-w_d}^{w_d - 1} P(A(\{d\}, k) = i) \\ &= \sum_{z_j = -1}^{1} \sum_{i=-w_d}^{w_d - 1} P(A(\{d, j\}, k) + z_j w_j = i) \end{aligned}$$

$$= \sum_{i=-w_d}^{w_d-1} P(A(\{d,j\},k)=i) + p_j(k)\,(1-p_j(k))\,S(j,k), \tag{25}$$

where

$$\begin{aligned} S(j,k) &= \sum_{i=-w_d}^{w_d-1} P(A(\{d,j\},k)=i-w_j) \\ &+ \sum_{i=-w_d}^{w_d-1} P(A(\{d,j\},k)=i+w_j) \\ &- 2\sum_{i=-w_d}^{w_d-1} P(A(\{d,j\},k)=i). \end{aligned}$$

Since $S(j,k)$ and $P\left(A(\{d,j\},k)=i\right)$ do not depend on p_j, the partial derivative of (25) with respect to p_j for all $j \neq d$ is

$$\frac{\partial H_d(k)}{\partial p_j} = (1-2p_j)\,S(j,k). \tag{26}$$

Obviously $\frac{\partial H_d(k)}{\partial p_j} = 0$ at q. If $S(j,k)=0$, then, by (25), p_j does not have any effect on H_d and, therefore, we let $p_j = 0.5$. If $S(j,k) \neq 0$, then, by (26), $p_j = 0.5$. This completes the proof. □

As a direct consequence of the above lemma, one derives the following lemma.

Lemma 10. *Let $f(x) = \sum_{i=1}^{n} w_i x_i$ be a binary linear function with $w_j \geq w_i > 0$ for $1 \leq i < j \leq n$. Then $H_i \leq H_j$. Besides, we have $1 \geq H_i \geq \frac{1}{2} + \frac{1}{2^n}$ for all $n \geq i \geq 1$.*

Proof. Considering the fact that $p(1) = (0.5, ..., 0.5)$, the above lemma gives

$$\begin{aligned} 2H_i - 1 &= P\left(-w_i \leq A\left(\{i\},1\right) < w_i\right) \\ &= \frac{1}{4}P\left(-w_i \leq A\left(\{i,j\},1\right) - w_j < w_i\right) \\ &+ \frac{1}{2}P\left(-w_i \leq A\left(\{i,j\},1\right) < w_i\right) \\ &+ \frac{1}{4}P\left(-w_i \leq A\left(\{i,j\},1\right) + w_j < w_i\right). \end{aligned} \tag{27}$$

But $w_i - w_j \leq w_j - w_i$ implies

$$P(-w_i \leq w_j + A(\{i,j\},1) < w_i) \leq P(-w_j \leq w_i + A(\{i,j\},1) < w_j).$$

In the same way, one argues that

$$\begin{aligned} P\left(-w_i \leq A\left(\{i,j\}1,\right) < w_i\right) &\leq P\left(-w_j \leq A\left(\{i,j\},1\right) < w_j\right) \\ P(-w_i \leq -w_j + A(\{i,j\},1) < w_i) &\leq P(-w_j \leq w_i + A(\{i,j\},1) < w_j). \end{aligned}$$

The combination of these inequalities and (27) proves $H_i \leq H_j$. since $0 < w_1$,

$$2H_1 - 1 = P(-w_1 \leq A(\{1\},1) < w_1) \geq P(A(\{1\},1)=0) = \frac{1}{2^{n-1}},$$

and consequently $H_1 > \frac{1}{2} + \frac{1}{2^n}$. □

In the following two examples we compute the exact values of H_d for two linear problems giving us the opportunity to verify our result by conducting some simulations.

Example 5.1. The OneMax problem is a frequently used fitness function in theory of evolutionary algorithms research because of its simplicity. The fitness of an individual is equal to the number of bits set to one, i.e. $f(x) = \sum_{i=1}^{n} x_i$. This is an easy problem for UEDAs since there is no isolation or deception. Let fix d. Let $A = \Sigma_{i=1,i\neq d}^{n} a_i$ and $B = \Sigma_{i=1,i\neq d}^{n} b_i$. Theorem 9 implies that $2H_d - 1 = P(-1 \le A - B < 1)$ with $p_j = 0.5$ for all $j \neq d$. Therefore one sees that A and B have the binomial distribution $B(n-1, \frac{1}{2})$. This concludes

$$\begin{aligned} P(A-B=z) &= \sum_{i=-n+1}^{n-1-z} P(A=i)P(B=i+z) \\ &= \sum_{i=-n+1}^{n-1-z} \left(\tfrac{1}{2}\right)^{i} \left(\tfrac{1}{2}\right)^{n-1-i} \binom{n-1}{i} \left(\tfrac{1}{2}\right)^{i+z} \left(\tfrac{1}{2}\right)^{n-1-i-z} \binom{n-1}{i+z} \\ &= \left(\tfrac{1}{2}\right)^{2n-2} \binom{2n-2}{n-1+z}. \end{aligned}$$

Since $P(A-B=-1) = P(A-B=1)$, we have

$$\begin{aligned} H_d - \tfrac{1}{2} &= \tfrac{1}{2}\left(P(A-B=0) + P(A-B=1)\right) \\ &= \left(\tfrac{1}{2}\right)^{2n-1} \left(\binom{2n-2}{n-1} + \binom{2n-2}{n}\right) \\ &= \left(\tfrac{1}{2}\right)^{2n-1} \binom{2n-1}{n}. \end{aligned}$$

Example 5.2. The BinVal problem is another used fitness function in theoretical research. The fitness of an individual is equal to the integer number in decimal base represented by the individual, i.e. $f(x) = \sum_{i=1}^{n} 2^{i-1} x_i$. Let fix d. Let $A = \Sigma_{i=1,i\neq d}^{n} 2^{i-1} a_i$ and $B = \Sigma_{i=1,i\neq d}^{n} 2^{i-1} b_i$. Since $P(A = B | a_d = 1, b_d = 0) = 0$, then $P(A > B | a_d = 1, b_d = 0) = P(A \ge B | a_d = 1, b_d = 0)$. Let t be the largest index such that $a_t = 1, b_t = 0$ and $b_j = a_j$ for all $n \ge j \ge t+1$. Note that $n \ge t \ge d$ because $a_d = 1, b_d = 0$. Since, for a given j, the coefficient 2^{j-1} of x_j is larger than the sum $\Sigma_{l=1}^{j-1} 2^{l-1} = 2^j - 1$, giving two points $a, b \in \Omega$, $f(a) > f(b)$ if and only if we have $t = i$ where i is the largest index with $a_i \neq b_i$. In this case, the values of $a_{t-1}, ..., a_1, b_{t-1}, ..., b_1$ do not have any influence on the inequality $f(a) > f(b)$. Thus for $d < n$

$$\begin{aligned} H_d &= \frac{1}{2}(P(A \ge B | a_d = 1, b_d = 0) + P(A > B | a_d = 1, b_d = 0)) \\ &= P(A > B | a_d = 1, b_d = 0) = \sum_{i=d}^{n} P(t = i) \\ &= \underbrace{\prod_{j=d+1}^{n} P(a_j = b_j)}_{P(t=d)} + \sum_{i=d+1}^{n-1} \underbrace{P(a_i = 1)P(b_i = 0) \prod_{j=i+1}^{n} P(a_j = b_j)}_{P(t=i)} \\ &+ \underbrace{P(a_n = 1)P(b_n = 0)}_{P(t=n)} = \left(\frac{1}{2}\right)^{n-d} + \sum_{t=d+1}^{n} \frac{1}{4}\left(\frac{1}{2}\right)^{n-t} = \frac{1}{2} + \frac{1}{2^{n-d+1}}, \end{aligned}$$

and for $d = n$, $H_d = 1$.

In general, when n is large enough, an approximation of H_d for $f(x) = \sum_{i=1}^{n} w_i x_i$ with $w_i > 0$ can be computed as follows. Let define $F_d(x,k) = \sum_{i \neq d} w_i x_i(k)$. Central Limit Theorem (Chung, 2000) implies that $F_d(x,k)$ converges in distribution to the normal distribution $N\left(M_d\left(k\right), \Sigma_d^2\left(k\right)\right)$ where $M_d(k) = \sum_{i \neq d} p_i(k) w_i$ and $\Sigma_d^2(k) = \sum_{i \neq d} p_i(k)(1 - p_i(k)) w_i^2$. As $\Delta_F = F_d(w(k),k) - F_d(l(k),k)$ has distribution $N\left(0, 2\Sigma_d^2\left(k\right)\right)$,

$$H_d(k) \approx \frac{1}{2} + \frac{1}{2} \int_{-w_d}^{w_d} N\left(0, 2\Sigma_d^2\left(k\right)\right) d\Delta_F \tag{28}$$

Obviously, $H_d(k)$ will be minimum when Σ_d^2 is maximum. By A-G means inequality, one sees

$$\Sigma_d^2(k) = \sum_{i \neq d} p_i(k)(1 - p_i(k)) w_i^2 \leq \sum_{i \neq d} \left(w_i \frac{p_i(k) + 1 - p_i(k)}{2} \right)^2 = \frac{\sum_{i \neq d} w_i^2}{4}.$$

Thus (28) gives

$$H_d \approx \frac{1}{2} + \frac{1}{2} \left(\Phi\left(\frac{w_d}{\sqrt{\sum_{i \neq d} w_i^2}} \right) - \Phi\left(\frac{-w_d}{\sqrt{\sum_{i \neq d} w_i^2}} \right) \right),$$

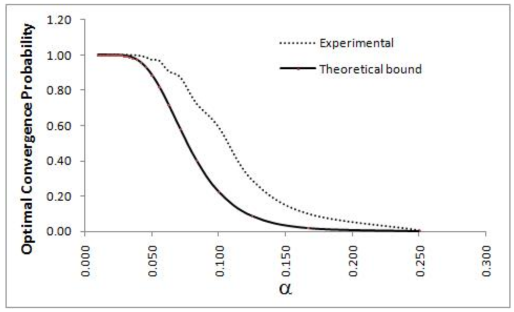
where $\Phi(.)$ is standard normal accumulation function.

The remainder of this section verifies the theoretical bounds on the optimal convergence probability of UEDAs. The experiments reported in this section are for the OneMax problem. All the results are the average over 1000 independent runs of the algorithms. For the cGA, each run was terminated when the PV had converged completely, however, for the PBIL, since the PV doesnot converge in a finite time, each run was terminated whenever for each $1 \leq i \leq n$, $p_i < 10^{-5}$ or $p_i > 1 - 10^{-5}$. We report the percentage of runs that converged to the optimal solution. The theoretical lower-bounds of the cGA and the PBIL are computed using (22) and (24), respectively.

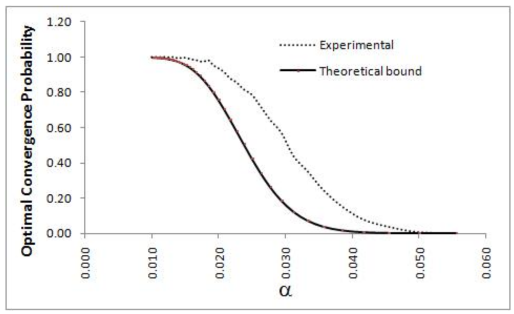
In Figures (1) and (2) the bold lines are the theoretical lowerbound and the dotted lines are the experimental results for the cGA and the PBIL, respectively, while maximizing a 20-bit and 100-bit OneMax problems.

6 Conclusion

The UEDAs are very simple and can be easily implemented in hardware. Using a small amount of memory, they may have many applications in the memory constraint problems. In addition, theoretical studying of these algorithms are very helpful to develop methods needed for the analysis of more complicated EDAs. This paper gives new theoretical results on the cGA and the PBIL, two of these kind of algorithms, which use probability distributions without dependencies between different components. The first part of the paper describes a derivation of bounds on the probability with which the cGA and the PBIL converge to the optimal solution. The approach closely follows a general approach proposed by (Norman, 1972) with several potential applications to the theory of evolutionary algorithms. Bounds are utilized to prove that the cGA and the PBIL converge almost surely to optimal solutions of functions with Property 1, as the learning rate (resp. population size) tends to zero (resp. infinity). Exact values of H_ds are computed for the OneMax and the BinVal problems, and an approximation is given for H_ds of linear functions when the size of problems is sufficiently large.



(a) 20-bit



(b) 100-bit

Figure 1: Experimental and theoretical results of the optimal convergence probability of the cGA on a 20-bit and 100-bit OneMax problems. The theoretical lower-bound is in dotted and the experimental result is in the bold line.

This paper creates the opportunity to study an aspect of EAs whose analysis is usually a difficult task. There are several natural extensions of the results here. First, we would like to compute H_ds for nonlinear functions satisfying Property 1. Since Property 1 considers only 1-bit building block, another extension would be to consider other building block sizes.

References

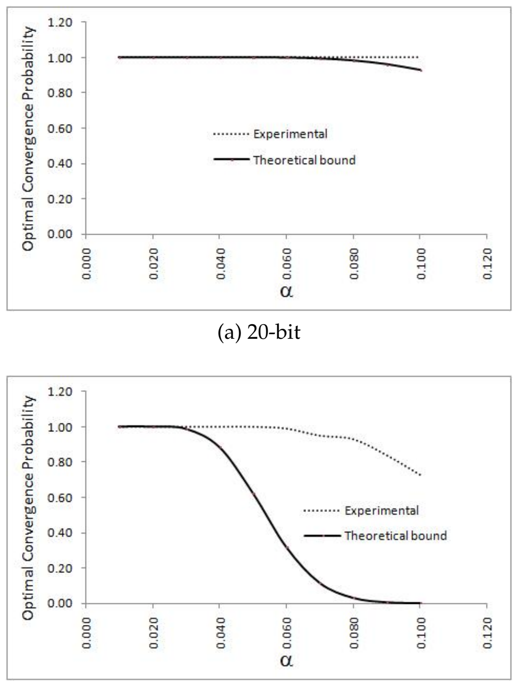
Baluja, S. and Caruana, R. (1995). Removing the genetics from the standard genetic algorithm. In Prieditis, A. and Russel, S., editors, *In the Proceedings of the Int. Conf. on Machine Learning 1995*, pages 38–46, San Mateo, CA, USA. Morgan Kaufmann Publishers.

Chen, T., Tang, K., and Yao, X. (2007). On the analysis of average time complexity of estimation of distribution algorithms. In *the Proceedings of IEEE Conference on Evolutionary Computation 2007*, pages 453–460.

Chung, K. L. (2000). *A Course in Probability Theory*. Academic Press.

Droste, S. (2005). A rigorous analysis of the compact genetic algorithm for linear functions. *Natural Computing: An International Journal*, pages 679–686.

Gonzalez, C., Lozano, J. A., and Larranaga, P. (2000). Analyzing the pbil algorithm by means of discrete dynamical systems. *Complex Systems*, 12(4):465–479.



(a) 20-bit

(b) 100-bit

Figure 2: Experimental and theoretical results of the optimal convergence probability of the PBIL with $\lambda = 5$ on a 20-bit and 100-bit OneMax problems. The theoretical lower-bound is in dotted and the experimental result is in the bold line.

Gonzalez, C., Lozano, J. A., and Larranaga, P. (2001). The convergence behavior of the pbil algorithm: A preliminary approach. In *the Proceedings 5th International Conference on Artificial Neural Networks and Genetic Algorithms*, pages 1–500. IEEE Press.

Harik, G. R., Cantu-Paz, E., Goldberg, D. E., and Miller, B. (1999a). The gamblers ruin problem, genetic algorithms, and the sizing of populations. *Evolutionary Computation*, 7(3):231–253.

Harik, G. R., Lobo, F. G., and Goldberg, D. E. (1999b). The compact genetic algorithm. *IEEE Trans. Evolutionary Computation*, 3(4):287–297.

Hohfeld, M. and Rudolph, G. (1997). Towards a theory of population based incremental learning. In *the Proceedings of the 4th IEEE Conferences on Evolutionary Computation*, pages 1–5, Indianapolis, IN, USA. IEEE Press.

Lakshmivarahan, S. and Thathachar, M. A. L. (1976). Bounds on the convergence probabilities of learning automata. *IEEE Transactions on Systems, Man, and Cybernetics, SMC*, 6:756–763.

Mühlenbein, H. (1997). The equation for response to selection and its use for prediction. *Evolutionary Computation*, 5(3):303–346.

Norman, F. (1972). *Markov Processes and Learning Models*. Academic Press.

Rastegar, R. and Hariri, A. (2006a). The population based incremental learning converges to local optima. *Neurocomputing*, 63(3):1772–1775.

Rastegar, R. and Hariri, A. (2006b). A step forward in studying the compact genetic algorithm. *Evolutionary Computation*, 14(3):277–289.

Rastegar, R. and Meybodi, M. R. (2005). A note on population based incremental learning with infinite population size. In *the Proceedings of IEEE Conference on Evolutionary Computation 2005*, pages 198–205, UK.

Rudolph, G. (2005). Analysis of a non-generational mutation less evolutionary algorithm for separable fitness functions. *International Journal of Computational Intelligence Research*, 1(1):77–84.

Thathachar, M. A. L. and Arvind, M. T. (1998). Parallel algorithms for modules of learning automata. *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, 28(1):24–33.

Tong, Y. L. (1997). Relationship between stochastic inequalities and some classical mathematical inequalities. *J. of Inequal. Appl.*, 1:85–98.

Zhang, Q. (2004). On stability of fixed points of limit models of univariate marginal distribution algorithm and factorized distribution algorithm. *IEEE Trans. Evolutionary Computation*, 8(1):80–93.