On the Optimal Convergence Probability of Univariate Estimation of Distribution Algorithms

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

Abstract

In this paper we obtain bounds on the probability of convergence to the optimal solution for the compact genetic algorithm (cGA) and the population based incremental learning (PBIL). Moreover, we give a sufficient condition for convergence of these algorithms to the optimal solution and compute a range of possible values for algorithm parameters at which there is convergence to the optimal solution with a predefined confidence level.

Keywords

EDA, PBIL, cGA, 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, Lobo, et al., 1999), the PBIL (Baluja and Caruana, 1995), and the UMDA (Mühlenbein, 1997), the latter being a special case of the PBIL.

Questions regarding convergence and time complexity of EDAs have been a topic of recent interest. In the first theoretical study of the convergence of the PBIL with an arbitrary learning rate in $(0, 1)$, Hohfeld and Rudolph (1997) argue that the PBIL converges almost surely to the maximum point of linear functions. Having a sufficiently small learning rate, Gonzalez et al. (2000) model the PBIL using a discrete dynamic system and demonstrate that the local optima of an injective function with respect to Hamming distance are stable fixed points of the PBIL. Following this research, Gonzalez et al. (2001) investigated strong dependency of the PBIL on initial values of the PV and the learning rate. In addition to the work of Hohfeld and Rudolph (1997), and Gonzales et al. (2001), Zhang (2004) studied the stability of fixed points of limit models of the UMDA while using a two-tournament selection scheme and showed that the local optima with respect to Hamming distance are asymptotically stable. Additionally, with regard specifically to the PBIL, Rastegar and Meybodi (2005) considered the case

when population size is sufficiently large, resulting in the derivation of dynamic properties for different selection schema. Following that work, Rastegar and Hariri (2006a,b) showed that the PBIL and the cGA, with sufficiently small learning rates, do not show any cyclic or chaotic behavior and moreover converge weakly to the local maxima with respect to Hamming distance.

Previous investigations concerning EDA time complexity include the first rigorous study on the time complexity of the cGA for linear pseudo-Boolean functions (Droste, 2005), giving the result that not all linear functions have the same asymptotical runtime. Chen et al. (2007) extended 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. In addition, their study includes the analysis of 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.

Of similar importance to convergence and time complexity are questions regarding effects of initial parameters, such as the initial PV, the learning rate, and the effects of population size, on the probability that the cGA and the PBIL converge to optimal solutions (called optimal convergence probability). The significance of these topics is self evident when one observes, for example, that when the learning rate is insufficiently small, it is not likely that the cGA converges to a good solution for the problem. Therefore, it may appear reasonable that the learning rate must be as small as possible in order to obtain high quality solutions. However, if the learning rate is too small, the cGA will be time inefficient, processing unnecessary individuals, resulting in unacceptably slow performance. The obvious objective is to determine a learning rate that is computationally inexpensive, yet small enough to permit a correct exploration of the search space (Harik, Lobo, et al., 1999).

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 difficult to obtain analytical expressions since the probability transient matrices of these Markov chains are intractable even for simple optimization problems. In some situations, assumptions regarding the population size, the operators, and the optimization problem aid in the estimation of the optimal convergence probability. These assumptions usually reduce the state space and, therefore, the size of the probability matrices, in some instances reducing these matrices into matrices with special properties.

Harik, Lobo, et al. (1999) argued that the dynamics of population-based EAs with recombination and selection but without mutation, are similar to the dynamics of specific random walks. The obtained results are based on multiple approximations, lacking error estimation. Rudolph (2005) proposes an improved argument, giving a mathematical model to lower bound the optimal convergence probability of a variation of nongenerational EAs, while optimizing the OneMax problem. The approach is still based on modeling the EA using random walks on finite space, yet he employs estimations which weaken the argument's mathematical integrity. Since the cGA mimics the behavior of a binary nongenerational EA, then one can use Rudolph's idea to bound the optimal convergence probability of the cGA. However, even if one builds a completely rigorous mathematical foundation upon previous work (Harik, Lobo, et al., 1999; Rudolph, 2005), one 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 considered in Norman (1972) includes stochastic learning models with distance diminishing operators in metric spaces for experiments with finite numbers of responses and simple reinforcement. A primary result of this framework is the following method of defining superregular and subregular functions, then using them to bound the convergence probability of a learning algorithm to different possible desired actions. In Lakshmivarahan and Thathachar (1976), it is shown that 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 therein.

In this paper, we will employ the method utilized in Norman (1972) to lower bound the optimal convergence probability of the cGA and the PBIL in the following manner. (1) Prove these algorithms converge to a point in Ω (i.e., Lemma 5). (2) Decompose the problem into tractable subproblems and compute bounds on the optimal convergence probability for each subproblem by bounding the interaction among the subproblems (using Lemma 3). (3) Integrate the partial bounds (by Lemma 4). Also, using the lower bounds, 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, we will derive some upper bounds on the learning rates and a lower bound on the population size to guarantee that algorithms will converge to the global optima with a predefined confidence level. As will become clear the advantage of this approach is that it facilitates the study of several properties of the cGA, the PBIL, and possibly other types of EAs under the same umbrella.

This paper is structured 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 for the optimal convergence probability are computed for the cGA and the PBIL using the methodology outlined above. 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 $p(k) = (p_1(k), \ldots, p_n(k))$ 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. Define the initial PV as $p(1) = p^0$ where $p^0 = (0.5, \ldots, 0.5)$.

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 $w^{(1)}(k)$, $w^{(2)}(k), \ldots, w^{(\lambda)}(k)$. 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 $(1 \le t \le \lambda)$ as follows. For each $1 \le t \le \lambda$, two random

individuals $c^{(1)}(k)$ and $c^{(2)}(k)$ are generated on the basis of $p(k)$ and then compete with each other and $w^{(t)}(k) = c^{(1)}(k), l^{(t)}(k) = c^{(2)}(k)$ (resp. $w^{(t)}(k) = c^{(2)}(k), l^{(t)}(k) = c^{(1)}(k)$) when $f(c^{(1)}(k)) \geq f(c^{(2)}(k))$ (resp. $f(c^{(2)}(k)) > f(c^{(1)}(k))$). Clearly, in our case, $\lambda = \frac{N}{2}$.

Harik, Cantu-Paz, et al. (1999) 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 $c^{(1)}(k)$ and $c^{(2)}(k)$ 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:

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

In order 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 a two-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*

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

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

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 $P(w^{(i)}(k) = y) = P(w^{(j)}(k) = y)$ for $1 \leq i, j \leq \lambda$.

3 Mathematical Preliminary

In this section, we define (sub,super) regular functions[1] and mention their connection to the convergence probability of a stochastic process to an absorbing state by stating some results similar to those of Norman (1972) and Lakshmivarahan and Thathachar (1976) for time-homogeneous Markov processes.

Suppose $\{\xi(k)\}_{k=1}^{\infty}$ is a Markov process with stationary transition kernel K defined on the compact set $S \subset \mathbb{R}^n$, where $K : S \times \sigma(S) \to \mathbb{R}$ where $\sigma(S)$ is the Borel-σ algebra generated by S. Suppose that $\{\xi(k)\}_{k=1}^{\infty}$ converges almost surely to some points in $A = \{s_0, \ldots, s_{N\text{-}1}\} \subset S$. 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 $A_1, A_2, \ldots, A_r$ be a partition of A where for $i \neq j$, A_i and A_j are noncommunicating classes, meaning that the probability of going from a point in A_i to a point in A_j is zero.

[1] Another commonly used name in the probability theory is (sub-super) harmonic functions.

Given $1 \leq i \leq r$, 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 $\psi(.) : S \to \mathbb{R}$, the operator U is defined by

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

for $k \geq 1$. Note that U is linear and preserves nonnegative 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 $\Gamma_{A_i}(.)$ $(i = 1, \ldots, r)$ satisfies a functional equation with appropriate boundary conditions.

LEMMA 2: *$\Gamma_{A_i}(.)$ is a solution of the functional equation $U\psi = \psi$ with the boundary conditions $\psi(s) = 1$ if $s \in A_i$ and $\psi(s) = 0$ if $s \in A_j$, $j \neq i$. Also, if $h \in C(S)$ is another solution of the equation, then $h = \Gamma_{A_i}$.*

REMARK: This result holds without the assumption that h is a continuous function. Please refer to Durrett (1995), Section 5.2, Excercise 2.6 for more information.

PROOF: Clearly Γ_{A_i} satisfies the boundary conditions. Also,

$$\begin{aligned} U\Gamma_{A_i}(s) &= \int_S \Gamma_{A_i}(y)K(s, dy) = \int_S P\left(\lim_{k\to\infty} \xi(k) \in A_i | \xi(1) = y\right) K(s, dy) \\ &= \lim_{k\to\infty} \int_S P\left(\xi(k) \in A_i | \xi(1) = y\right) K(s, dy) = \lim_{k\to\infty} \int_S K^k(y, A_i)K(s, dy) \\ &= \lim_{k\to\infty} K^{k+1}(s, A_i) = \Gamma_{A_i}(s). \end{aligned}$$

Suppose $h \in C(S)$ is another solution of the equation. 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 the bounded convergence theorem (Durrett, 1995) gives

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

where Equation (6) comes from the fact that each subsequence of an 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 $\Gamma_{A_i}(s)$ $(i = 1, \ldots, r)$ 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 $s \in A_j$, $j \neq i$, then $\psi(s) \leq \Gamma_{A_i}(s)$ (resp. $\psi(s) \geq \Gamma_{A_i}(s)$) for all $s \in S$.*

PROOF: The proof is similar to that of Lemma 2. □

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 parameters is the most difficult part of the procedure. The following lemma can be useful to simplify this procedure.

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

PROOF: The application of the Chebyshev integral inequality (Tong, 1997) implies

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

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

Using Lemma 4 in finding the subregular function leads us to a more conservative result; however, it reduces the difficulty of the 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 defined in the following.

DEFINITION (Property 1): *A function* $f : \Omega \to \mathbb{R}$ *satisfies Property 1 if* $f(x \vee e_i) \geq f(x \wedge \overline{e}_i)$ *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 componentwise "AND" and "OR," respectively.*

This property essentially states that setting one bit to 0 does not increase the function value. All linear functions $f(x) = \Sigma_{i=1}^{n} \gamma_i x_i$ with $\gamma_i > 0$ have Property 1. There are also some nonlinear functions such as $f(x) = 2\Sigma_{i=1}^{n} \gamma_i x_i + \prod_{i=1}^{n} x_i$ having this property. From this point forward, we assume that f satisfies Property 1.

4.1 Lower Bound for the cGA

The cGA shows a complicated nonlinear behavior. In order to analyze the optimal convergence probability of this algorithm we approach the problem as follows. We first prove the algorithm will converge to a point in Ω. Then, we decompose the problem into tractable subproblems and we compute some bounds on the optimal convergence probability for each subproblem by bounding the interaction among the subproblems. Finally, we integrate the partial bounds.

Let the random sequence $\{p(k)\}_{k=1}^{\infty}$ be generated by the cGA while optimizing function f. It is clear that this sequence is a time-homogeneous Markov chain on $S = \{0, \alpha, 2\alpha, \ldots, 1\}^n$ with $A = \Omega$ as the absorbing points and $S - A$ as the transient states, thus the a.s. convergence of the cGA to a point in Ω is guaranteed. However, we will prove this fact using a second approach developed in Hohfeld and Rudolph (1997), since the latter can be easily used to show the convergence of the PBIL while the first approach does not work for the PBIL, and also, the second approach gives some insights about the behavior of each $\{p_d(k)\}_{k=1}^{\infty}$.

LEMMA 5: *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: Equation (2) implies $E[p_d(k+1) \,|\, p(k)] = p_d(k) + \alpha E[w_d(k) - l_d(k) | p(k)]$ for all $1 \leq d \leq n$. Since f satisfies Property 1, for a given $x \in \Omega$, $f(x \vee e_d) \geq f(x \wedge \overline{e}_d)$. Hence

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

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

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

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

Then based on Lemma 1 we have

$$\begin{aligned}\frac{P(w(k)=x\vee e_d)}{P_k(x\vee e_d)} &= \sum_{f(z)<f(x\vee e_d)} P_k(z) + \sum_{f(z)\le f(x\vee e_d)} P_k(z)\\ &\ge \sum_{f(z)<f(x\wedge \overline{e}_d)} P_k(z) + \sum_{f(z)\le f(x\wedge \overline{e}_d)} P_k(z)\\ &= \frac{P(w(k)=x\wedge \overline{e}_d)}{P_k(x\wedge \overline{e}_d)}. \end{aligned} \tag{11}$$

and in a similar way

$$\frac{P(l(k)=x\vee e_d)}{P_k(x\vee e_d)} \le \frac{P(l(k)=x\wedge \overline{e}_d)}{P_k(x\wedge \overline{e}_d)}. \tag{12}$$

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 $P_k(x\wedge\overline{e}_d)=(1-p_d(k))q_d(x,k)$ and $P_k(x\vee e_d)=p_d(k)q_d(x,k)$. Insertion of these identities into the inequalities in Equations (11) and (12) and some simplification show that

$$\begin{aligned} P\left(w(k)=x\vee e_d\right) &\ge p_d(k)\left(P(w(k)=x\wedge\overline{e}_d)+P(w(k)=x\vee e_d)\right)\\ P\left(l(k)=x\vee e_d\right) &\le p_d(k)\left(P(l(k)=x\wedge\overline{e}_d)+P(l(k)=x\vee e_d)\right). \end{aligned}$$

Thus, the above inequalities give

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

This shows that $\{p_d(k)\}_{k=1}^{\infty}$ is a sub-Martingale which is positive and uniformly bounded by one. Thus Martingale theorem (Durrett, 1995) asserts that $\lim_{k\to\infty} p_d(k)=p^*$ almost surely exists. If $p_d^*\notin\{0,1\}$, then $p_d^*(k)\neq p_d^*(k+1)$ with a nonzero 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 $\{p_d(k)\}$. This completes the proof. □

We are now in a position to apply the results of Section 3 to find a bound on the optimal convergence probability of the cGA. Without loss of generality, we assume that $x^* = (1, \ldots, 1)$ is the only maximum point of function f. Partition A to two sets of the optimal point, $A_1 = \{(1, \ldots, 1)\}$, and nonoptimal points, $A_2 = \Omega - A_1$, then the optimal convergence probability of the cGA will be $\Gamma_{A_1}((0.5, \ldots, 0.5))$, the probability that $\{p(k)\}$ converges to x^*.

The 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 $\psi(p) = 0$ for $p \in A_2$. 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, that is, $\psi(.)$ is a subregular function. In order to compute the largest value of b, we need to have the 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 find another functional form. One way is to first decompose the PV, $p(k) = (p_1(k), \ldots, p_n(k))$ to some sub-PVs. Then for a given sub-PV, we introduce a subregular function depending only on this sub-PV by bounding its interaction with other sub-PVs. The larger sub-PVs' sizes are, the sharper result we get, but at the same time, the complexity of the approach increases. Finally, we find our subregular function by multiplying the sub-PVs subregular functions. For the sake of simplicity in the notation and computation, we will consider the sub-PVs with size one, that is, we look at subregular function

$$\psi(.) = \prod_{d=1}^{n} \psi_d(p), \tag{13}$$

with

$$\psi_d(p) = \frac{1 - e^{-b_d p_d}}{1 - e^{-b_d}} \tag{14}$$

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.

In the following, we define $H_d(k)$ which is the quantity that models the interactions of $p_d(k)$ with other PV components at iteration k.

DEFINITION:

$$\begin{aligned} H_d(k) = {} & P\left(f(c^{(1)}(k)) \geq f(c^{(2)}(k)) | c_d^{(1)}(k) = 1, c_d^{(2)}(k) = 0\right) \\ & + P\left(f(c^{(2)}(k)) > f(c^{(1)}(k)) | c_d^{(2)}(k) = 1, c_d^{(1)}(k) = 0\right). \end{aligned} \tag{15}$$

In order to exclude that factor of time from the interaction among the sub-PVs, we define

$$H_d = \min_k H_d(k)$$

and use it to find the subregular functions.

Using this new notation, we have

$$\begin{aligned}
P(w_d(k)-l_d(k)=1|p(k)) &= P\left\{f(c^{(1)}(k))\geq f(c^{(2)}(k)), c_d^{(1)}(k)=1, c_d^{(2)}(k)=0|p(k)\right\} \\
&\quad +P\left\{f(c^{(2)}(k))> f(c^{(1)}(k)), c_d^{(1)}(k)=0, c_d^{(2)}(k)=1|p(k)\right\} \\
&= 2H_d(k)p_d(k)\left(1-p_d(k)\right). \qquad (16)\\
P(w_d(k)-l_d(k)=0|p(k)) &= P\left(w_d(k)=1, l_d(k)=1|p(k)\right) \\
&\quad +P\left(w_d(k)=0, l_d(k)=0|p(k)\right) \\
&= 1-2p_d(k)\left(1-p_d(k)\right). \qquad (17)\\
P(w_d(k)-l_d(k)=-1|p(k)) &= 1-P(w_d(k)-l_d(k)=0|p(k)) \\
&\quad -P(w_d(k)-l_d(k)=1|p(k)) \\
&= 2(1-H_d(k))p_d(k)\left(1-p_d(k)\right). \qquad (18)
\end{aligned}$$

Note that

$$\begin{aligned}
E\left\{p_d(k+1)-p_d(k)|p(k)\right\} &= \alpha P\left(w_d(k)-l_d(k)=1|p(k)\right) \\
&\quad -\alpha P\left(w_d(k)-l_d(k)=-1|p(k)\right) \\
&= \alpha\left(2H_d(k)-1\right)p_d(k)\left(1-p_d(k)\right). \qquad (19)
\end{aligned}$$

By lemma 5, the left-hand side of Equation (19) is always nonnegative, therefore $1 \leq 2H_d(k)$.

LEMMA 6: *Define $\psi_d : S \to \mathbb{R}$ as in Equation (14). 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 Equations (16)–(18) give

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

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

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

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

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

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

By the inequality in Equation (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 7: *Let $p^0 = (0.5, \ldots, 0.5)$ be the initial PV and x^* be the optimal solution. Then*

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

PROOF: Let $\psi(.)$ be defined as in Equation (13). Since the $\psi_d(.)$ values are increasing, by Lemma 4, $\psi(.)$ is subregular if each $\psi_d(.)$ is subregular. Therefore, according to Lemmas 3 and 6 we have

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

which completes the proof. □

REMARK: A similar result is reported in Rudolph (2005) for binary nongenerational evolutionary algorithm (the cGA) optimizing the OneMax problem. However, there are two questionable points in the argument. In order 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

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

Clearly, these random walks are state-dependent time-inhomogeneous Markov processes. Replacing the transition probabilities of these random walks 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 Rudolph (2005) 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 $(1, \ldots, 1)$ is the product of lower bounds on probabilities that random walks $\{p_d(k)\}$ converge to 1; however, since $P(\lim_{k\to\infty} p_d(k) = 1|p(1) = p^0) \geq P(\lim_{k\to\infty} p_1(k) = 1, \ldots, \lim_{k\to\infty} p_n(k) = 1|p(1) = p^0)$ 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 sufficiently small α, the cGA converges almost surely to the optimal solution of functions with Property 1. If $H_d > \frac{1}{2}$ (this is proven at least for the linear functions in Section 5), then $\frac{1-H_d}{H_d} < 1$ for all $1 \leq d \leq n$. Thus letting $\alpha \to 0$ in Theorem 7 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 7 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 7 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.

4.2 Lower Bound for the PBIL

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 sub-Martingale, $\lim_{k\to\infty} p_d(k) = p_d^*$ exists, and $p_d^* \in \{0, 1\}$ almost surely. Therefore the absorbing set of $\{p(k)\}_{k=1}^{\infty}$ is Ω, that is, $A = \Omega$. Define A_1 and A_2 as before. A promising subregular function for computing a bound on the optimal probability of the PBIL could be Equation (13) where $b_d > 0$s are to be chosen. It can be shown that

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

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 right-most side of the 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, 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 $G_d(k) = p_d^2(k) + 2p_d(k)(1 - p_d(k))H_d(k)$, where $H_d(k)$ is defined in Section 4.1, 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 values of 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

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

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 Equation (23) and simplifying, we have

$$\begin{aligned}1 - u(b_d^*(k), k) = &\left(p_d(k) + 2\,(1 - p_d(k))\, H_d(k)\right)^{\lambda p_d(k)} \\ &\times (1 + p_d(k) - 2p_d(k)H_d(k))^{\lambda(1 - p_d(k))}.\end{aligned} \tag{24}$$

Note that a general form of arithmetic-geometric means inequality indicates that

$$\left(\frac{c_1 b_1 + c_2 b_2}{c_1 + c_2}\right)^{c_1 + c_2} \geq b_1^{c_1} b_2^{c_2}$$

where c_i and b_i are nonnegative. An application of this inequality to the right-hand side of Equation (24) implies it 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

$$\frac{\partial b_d^*(k)}{\partial H_d(k)} = \frac{2\lambda}{\alpha\,(1 + p_d(k) - 2p_d(k)H_d(k))\,(p_d(k) + 2(1 - p_d(k))H_d(k))} > 0$$

and

$$\frac{\partial b_d^*(k)}{\partial p_d(k)} = \frac{(2H_d(k) - 1)^2}{(1 + p_d(k) - 2p_d(k)H_d(k))\,(p_d(k) + 2(1 - p_d(k))H_d(k))} \geq 0.$$

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 7 shows that by selecting $b_d = \frac{\lambda}{\alpha}\ln 2H_d$, for each $1 \leq d \leq n$, we have

$$\prod_{d=1}^{n}\left(1 + \left(\frac{1}{2H_d}\right)^{\frac{\lambda}{2\alpha}}\right)^{-1} \leq \Gamma_{A_1}(p^0) = P(\lim_{k\to\infty} p(k) = x^*|p(1) = p^0). \tag{25}$$

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 $H_d > 0.5$ for all d, a complementary result to that found in Gonzalez et al. (2000) and Rastegar and Hariri (2006a). Again, this computation yields 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 Equation (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 the bounds can be improved 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 was 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 two-bit One-Max 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 counter example 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_d Values and Experimental Verification

Knowing H_d values for a given function is essential for all of our results. In this section we compute H_d values for some simple functions. Suppose $f(x) = \sum_{i=1}^{n} \gamma_i x_i$. Define $A(I,k) = \Sigma_{i \notin I} \gamma_i (c_i^{(1)}(k) - c_i^{(2)}(k))$ for a subset $I \subset \{1, \ldots, n\}$. To simplify the notation, we assume that all γ_i values are natural numbers. However, with some adjustment in the notation, the following lemma holds for all positive real γ_i.

First note that

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

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 ith component of $p(k)$. Fix $1 \leq d \leq n$. It is not hard 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

$$P(z_j(k) = -1) = P(z_j(k) = 1) = \frac{1 - P(z_j(k) = 0)}{2} = p_j(k)(1 - p_j(k)). \tag{26}$$

Fix $j \neq d$. Note that, using Equation (26), $H_d(k)$ can be rewritten as follows

$$\begin{aligned} 2H_d(k) - 1 &= P(-\gamma_d \leq A(\{d\},k) < \gamma_d) \\ &= \sum_{i=-\gamma_d}^{\gamma_d - 1} P(A(\{d\},k) = i) \\ &= \sum_{z_j=-1}^{1} \sum_{i=-\gamma_d}^{\gamma_d - 1} P(A(\{d,j\},k) + z_j\gamma_j = i) \\ &= \sum_{i=-\gamma_d}^{\gamma_d - 1} P(A(\{d,j\},k) = i) + p_j(k)\left(1 - p_j(k)\right) S(j,k), \end{aligned} \tag{27}$$

where

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

Since $S(j,k)$ and $P\left(A(\{d,\, j\},k)=i\right)$ do not depend on p_j, the partial derivative of Equation (27) with respect to p_j for all $j \neq d$ is

$$\frac{\partial H_d(k)}{\partial p_j} = \left(1-2p_j\right) S(j,k). \tag{28}$$

Obviously $\frac{\partial H_d(k)}{\partial p_j}=0$ at q. If $S(j,k)=0$, then, by Equation (27), p_j does not have any contribution to the value of H_d and, therefore, we let $p_j=0.5$. If $S(j,k)\neq 0$, then, by Equation (28), $p_j=0.5$. This argument shows that the $H_d(k)$ values are minimum at the time 1 when $p(1)=(0.5,\ldots.0.5)$, that is, $H_d=H_d(1)$.

LEMMA 8: *Let $f(x)=\sum_{i=1}^{n}\gamma_i x_i$ be a binary linear function with $\gamma_j \geq \gamma_i > 0$ 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 $p(1)=(0.5,\ldots,0.5)$, the above lemma gives

$$\begin{aligned} 2H_i-1 &= P\left(-\gamma_i \leq A\left(\{i\},1\right) < \gamma_i\right) \\ &= \frac{1}{4}P\left(-\gamma_i \leq A\left(\{i,\, j\},1\right) - \gamma_j < \gamma_i\right) \\ &\quad +\frac{1}{2}P\left(-\gamma_i \leq A\left(\{i,\, j\},1\right) < \gamma_i\right) \\ &\quad +\frac{1}{4}P\left(-\gamma_i \leq A\left(\{i,\, j\},1\right) + \gamma_j < \gamma_i\right). \end{aligned} \tag{29}$$

Since $\gamma_i-\gamma_j \leq \gamma_j-\gamma_i$,

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

In the same way,

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

The combination of these inequalities and Equation (29) proves $H_i \leq H_j$. Since $0<\gamma_1$,

$$2H_1-1 = P(-\gamma_1 \leq A(\{1\},1) < \gamma_1) \geq P(A(\{1\},1)=0) = \frac{1}{2^{n-1}},$$

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

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

EXAMPLE 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, that is, $f(x)=\sum_{i=1}^{n} x_i$. This is an easy problem

for UEDAs since there is no isolation or deception. For a fixed d, let $A = \Sigma_{i=1,i\neq d}^{n} c_i^{(1)}$ and $B = \Sigma_{i=1,i\neq d}^{n} c_i^{(2)}$, where $c^{(1)}$ and $c^{(2)}$ are defined as in Section 2. The above argument implies that $2H_d - 1 = P(-1 \leq A - B < 1)$ with $p_j = 0.5$ for all $j \neq d$. Therefore, A and B have the binomial distribution $B(n-1, \frac{1}{2})$. This yields

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

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 2: The BinVal problem is another fitness function used in theoretical research. The fitness of an individual is equal to the integer number in decimal base represented by the individual, that is, $f(x) = \sum_{i=1}^{n} 2^{i-1} x_i$. For a fixed $1 \leq d \leq n$ and $a, b \in \Omega$, let $A = \Sigma_{i=1,i\neq d}^{n} 2^{i-1} c_i^{(1)}$ and $B = \Sigma_{i=1,i\neq d}^{n} 2^{i-1} c_i^{(2)}$. Since $P(A = B | c_d^{(1)} = 1, c_d^{(2)} = 0) = 0$, then $P(A > B | c_d^{(1)} = 1, c_d^{(2)} = 0) = P(A \geq B | c_d^{(1)} = 1, c_d^{(2)} = 0)$. Let t be the largest index such that $c_t^{(1)} = 1, c_t^{(2)} = 0$ and $c_j^{(2)} = c_j^{(1)}$ for all $n \geq j \geq t+1$. Note that $n \geq t \geq d$ because $c_d^{(1)} = 1, c_d^{(2)} = 0$. Since, for a given j, the coefficient 2^{j-1} of x_j is larger than the sum $\Sigma_{l=1}^{j-1} 2^{l-1} = 2^{j-1} - 1$, $f(a) > f(b)$ if and only if we have $t = i$ where i is the largest index with $c_i^{(1)} \neq c_i^{(2)}$. In this case, the values of $c_{t-1}^{(1)}, \ldots, c_1^{(1)}, c_{t-1}^{(2)}, \ldots, c_1^{(2)}$ do not have any influence on the inequality $f(a) > f(b)$. Thus for $d < n$

$$\begin{aligned} H_d &= \frac{1}{2}\left(P(A \geq B | c_d^{(1)} = 1, c_d^{(2)} = 0) + P(A > B | c_d^{(1)} = 1, c_d^{(2)} = 0)\right) \\ &= P(A > B | c_d^{(1)} = 1, c_d^{(2)} = 0) = \sum_{i=d}^{n} P(t=i) \\ &= \underbrace{\prod_{j=d+1}^{n} P(c_j^{(1)} = c_j^{(2)})}_{P(t=d)} + \sum_{i=d+1}^{n-1} \underbrace{P(c_i^{(1)} = 1) P(c_i^{(2)} = 0) \prod_{j=i+1}^{n} P(c_j^{(1)} = c_j^{(2)})}_{P(t=i)} \\ &\quad + \underbrace{P(c_n^{(1)} = 1) P(c_n^{(2)} = 0)}_{P(t=n)} = \left(\frac{1}{2}\right)^{n-d} + \sum_{t=d+1}^{n} \frac{1}{4}\left(\frac{1}{2}\right)^{n-t} = \frac{1}{2} + \frac{1}{2^{n-d+1}}, \end{aligned}$$

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

In general, when n is large enough, an approximation of H_d for $f(x) = \sum_{i=1}^{n} \gamma_i x_i$ with $\gamma_i > 0$ can be computed as follows. Define $F_d(x, k) = \sum_{i \neq d} \gamma_i x_i(k)$. The central limit theorem (Durrett, 1995) implies that $F_d(x, k)$ converges in distribution to the normal distribution $N\left(M_d(k), \Sigma_d^2(k)\right)$ where $M_d(k) = \sum_{i \neq d} p_i(k)\gamma_i$ and $\Sigma_d^2(k) = \sum_{i \neq d} p_i(k)(1 - p_i(k))\gamma_i^2$. Since $\Delta_F = F_d(w(k), k) - F_d(l(k), k)$ has distribution $N\left(0, 2\Sigma_d^2(k)\right)$,

$$H_d(k) \approx \frac{1}{2} + \frac{1}{2} \int_{-\gamma_d}^{\gamma_d} N\left(0, 2\Sigma_d^2(k)\right) d\Delta_F. \tag{30}$$

Obviously, $H_d(k)$ will be a minimum when Σ_d^2 is a maximum. By arithmetic-geometric means inequality,

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

Thus Equation (30) gives

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

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 OneMax and BinVal problems. All the results are averaged over 1 000 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 does not 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 Equations (22) and (25), respectively.

In Figures 1 and 2, the solid lines are the theoretical lower bound and the dotted lines are the experimental results for the cGA and the PBIL, respectively, while maximizing five-bit and 100-bit OneMax problems. As it is clear in the figures, in the case of the OneMax problem, the obtained lower bound for the cGA is sharper in comparison to the lower bound of the PBIL. One main reason for this difference is related to the optimality of the computed b for the cGA. Please refer to the first remark in Section 4.2 for details. Also, simulation shows that lower bounds are in general sharper for the OneMax problem in comparison to the bounds for the BinVal problem (compare Figures 1, 2, and 3, for example). The main reason for this difference is that the contribution of all bits in the OneMax problem is the same, and so considering one-bit subproblems in the process of finding the lower bound is a reasonable decision; however, for the BinVal, the contribution of different bits is very different and by dividing the problem to one-bit subproblems, we lose a lot of information about the dynamics of the algorithm. The author believes that the bounds will be considerably improved if we use two-bit subproblems.

At the end, note that the large gap between the experimental result and the theoretical bound usually happens when the learning rate is not sufficiently small. In practice,

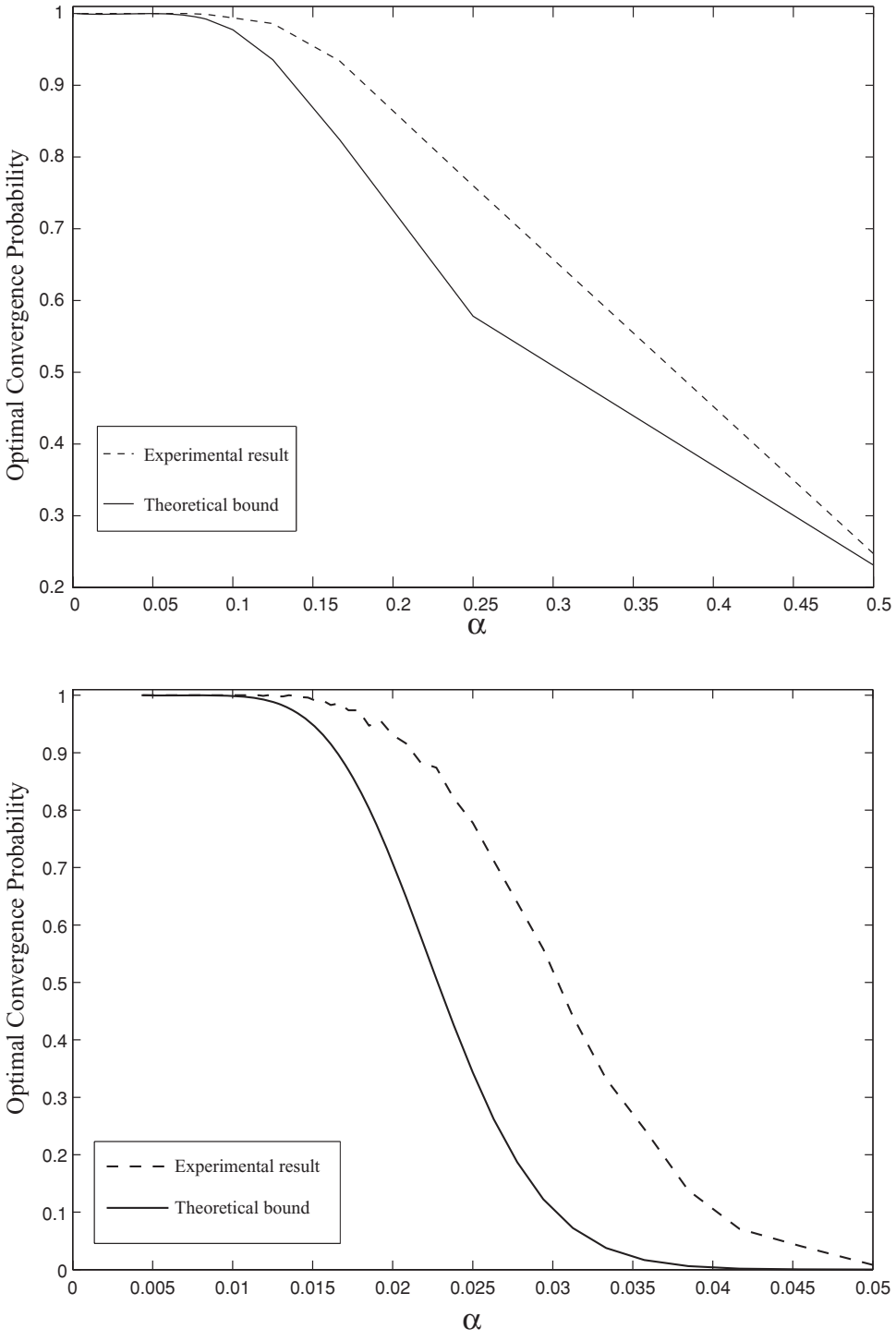

Figure 1: Experimental and theoretical results of the optimal convergence probability of the cGA on a five-bit (top plot) and 100-bit (bottom plot) OneMax problem. The theoretical lower bound is the solid line and the experimental result is the dotted line.

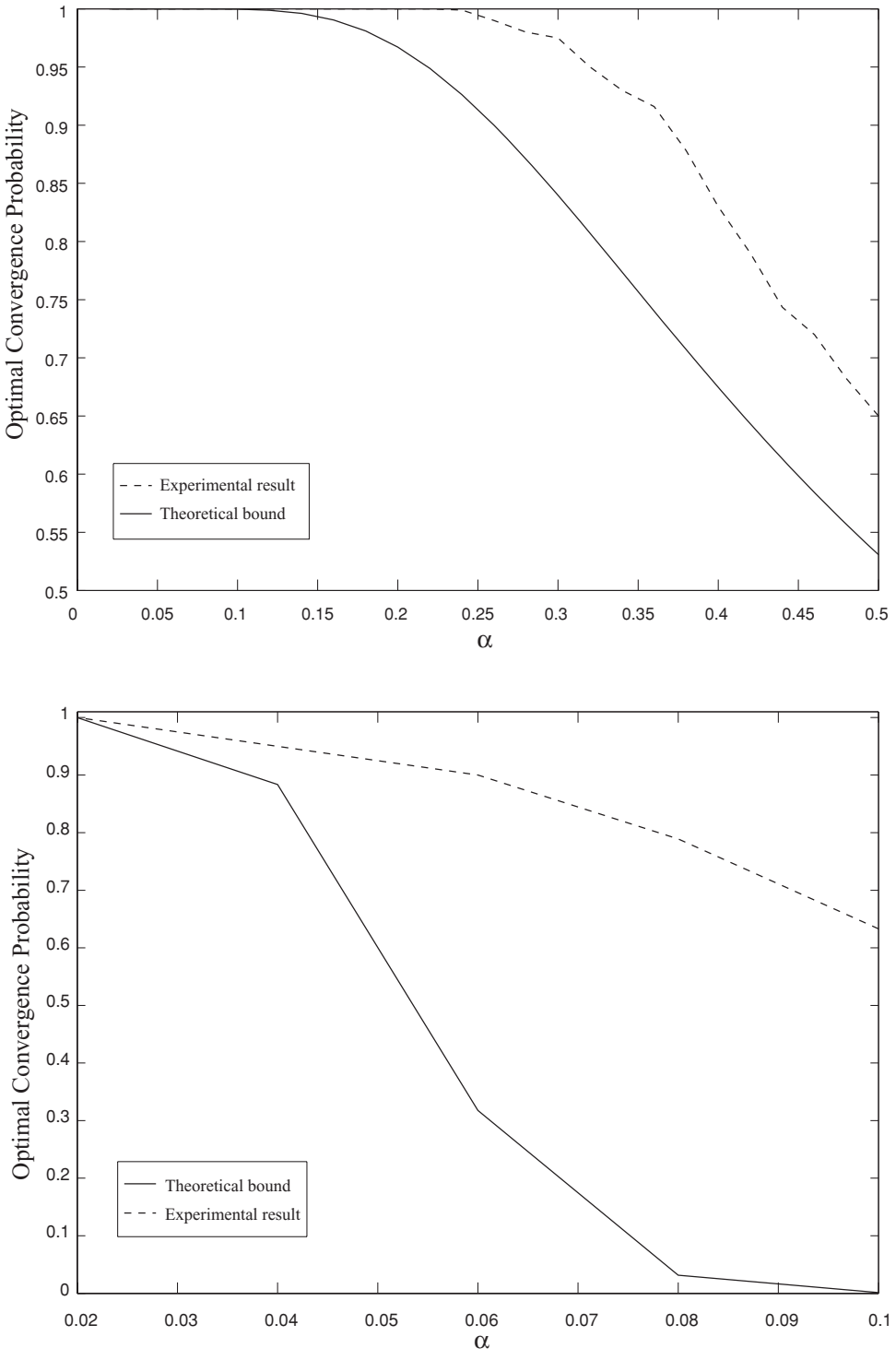

Figure 2: Experimental and theoretical results of the optimal convergence probability of the PBIL with $\lambda = 5$ on a five-bit (top plot) and 100-bit (bottom plot) OneMax problem. The theoretical lower bound is the solid line and the experimental result is the dotted line.

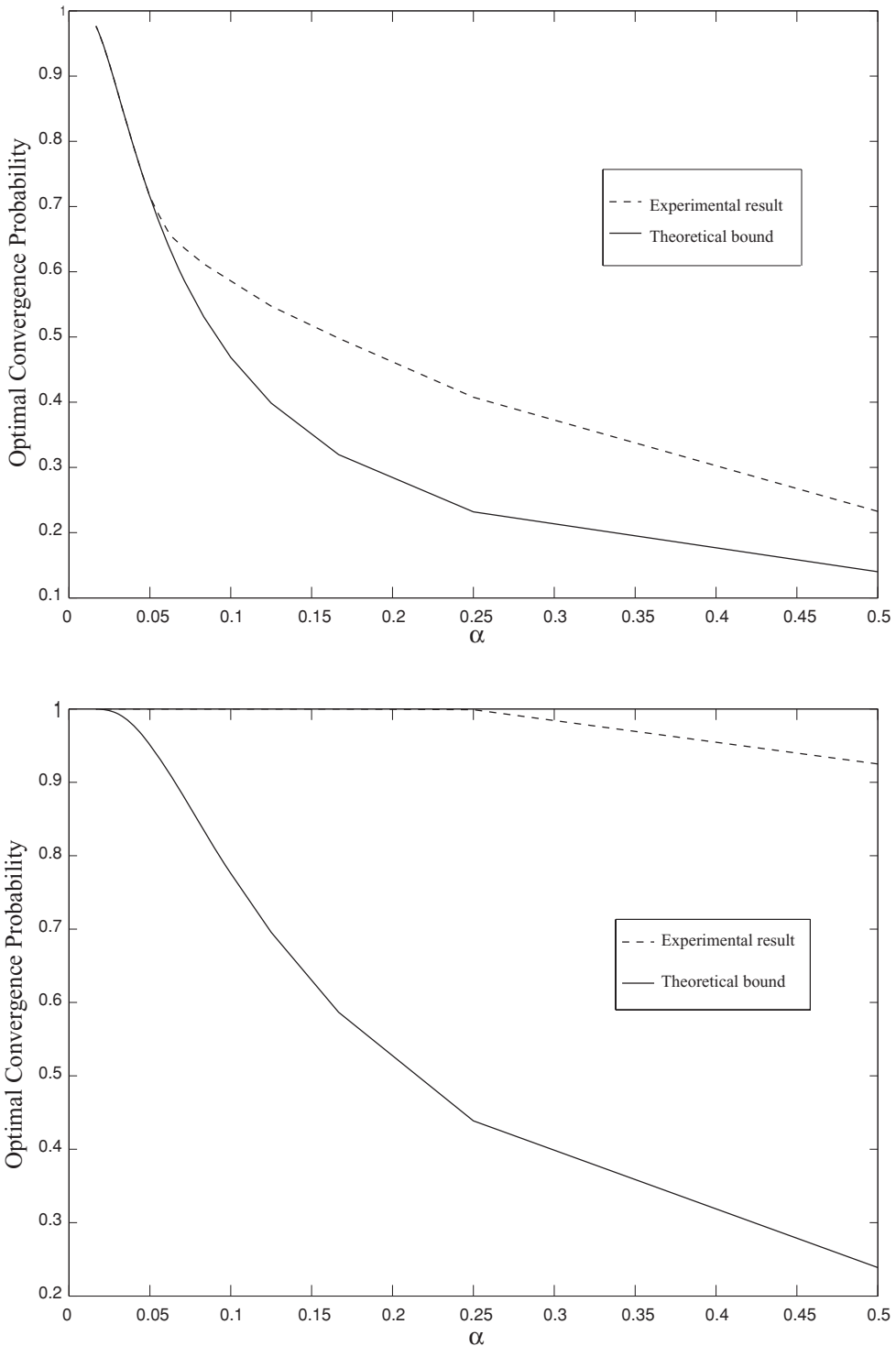

Figure 3: Experimental and theoretical results of the optimal convergence probability of the cGA (top plot) and the PBIL (bottom plot) on a five-bit BinVal problem. The theoretical lower bound is the solid line and the experimental result is the dotted line.

we usually use a reasonable small value of α, that is, ≤ 0.01, in which the theoretical results are sharp.

6 Conclusion

The UEDAs are very simple and can be easily implemented in hardware. Because they use a small amount of memory, they may have many applications in memory constraint problems. In addition, theoretical studies 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 kinds of algorithms, which use probability distributions without dependencies between different components. The first part of the paper describes a derivation of lower 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_d are computed for the OneMax and the BinVal problems, and an approximation is given for H_d values of linear functions when the size of problems is sufficiently large.

There are several natural extensions of the results here. The first extension is to compute H_d values for nonlinear functions satisfying Property 1. Since Property 1 considers only the one-bit building block, another extension would be to consider other building block sizes. This perhaps improves the bounds, especially for the BinVal. Finding an appropriate form of super regular function can also be used to find upper bounds. Having an upper bound gives us a better picture of the behavior of the algorithms and the average of the upper bounds and lower bounds could be a better estimate for the optimal convergence probability of the algorithms.

Acknowledgments

The author is very grateful to Sunder Sethuraman for useful discussion and Alexander Roitershtein and Diana Hay for the careful reading of a draft of this paper and many helpful remarks and suggestions. Thanks are also given to the anonymous referees for very valuable comments which improved both content and presentation of this paper.

References

Baluja, S., and Caruana, R. (1995). Removing the genetics from the standard genetic algorithm. In A. Prieditis and S. Russel (Eds.), *Proceedings of the International Conference on Machine Learning 1995*, pp. 38–46.

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

Droste, S. (2005). A rigorous analysis of the compact genetic algorithm for linear functions. *Natural Computing: An International Journal*, 5(3):679–686.

Durrett, R. (1995). *Probability: Theory and examples*. Belmont, CA: Duxbury Press.

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.

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

Harik, G. R., Cantu-Paz, E., Goldberg, D. E., and Miller, B. (1999). 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. (1999). The compact genetic algorithm. *IEEE Transactions on Evolutionary Computation*, 3(4):287–297.

Hohfeld, M., and Rudolph, G. (1997). Towards a theory of population based incremental learning. In *Proceedings of the 4th IEEE Conference on Evolutionary Computation*, pp. 1–5.

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*. San Diego, CA: 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 *Proceedings of the IEEE Conference on Evolutionary Computation 2005*, pp. 198–205.

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. *Journal of Inequality Applications*, 1:85–98.

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