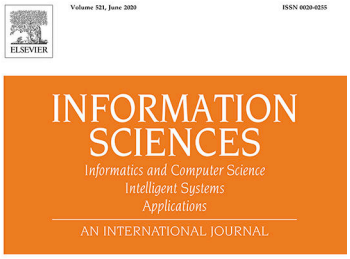
Journal Pre-proofs



Average Convergence Rate of Evolutionary Algorithms in Continuous Optimization

Yu Chen, Jun He

PII: S0020-0255(20)31242-1
DOI: https://doi.org/10.1016/j.ins.2020.12.076
Reference: INS 16112

To appear in: *Information Sciences*

Accepted Date: 27 December 2020

Please cite this article as: Y. Chen, J. He, Average Convergence Rate of Evolutionary Algorithms in Continuous Optimization, *Information Sciences* (2021), doi: https://doi.org/10.1016/j.ins.2020.12.076

Average Convergence Rate of Evolutionary Algorithms in Continuous Optimization ⋆ ,⋆⋆

Yu Chen[a], Jun He[b,*]

[a]School of Science, Wuhan University of Technology, Wuhan, 430070, China.
[b]School of Science and Technology, Nottingham Trent University, Nottingham NG11 8NS, UK

Abstract

The average convergence rate (ACR) measures how fast the approximation error of an evolutionary algorithm converges to zero per generation. It is defined as the geometric average of the reduction rate of the approximation error over consecutive generations. This paper makes a theoretical analysis of the ACR in continuous optimization. The obtained results are summarized as follows. According to the limit property, the ACR is classified into two categories: (1) linear ACR whose limit inferior value is larger than a positive and (2) sublinear ACR whose value converges to zero. Then, it is proven that the ACR is linear for evolutionary programming using positive landscape-adaptive mutation, but sublinear for that using landscape-invariant or zero landscape-adaptive mutation. The relationship between the ACR and the decision space dimension is also classified into two categories: (1) polynomial ACR whose value is larger than the reciprocal of a polynomial function of the dimension for any generation, and (2) exponential ACR whose value is less than the reciprocal of an exponential function of the dimension for an exponential long period. It is proven that for easy problems such as linear functions, the ACR of the (1+1) adaptive random univariate search is polynomial. But for hard functions such as the deceptive function, the ACR of both the (1+1) adaptive random univariate search and evolutionary programming is exponential.

Keywords: Evolutionary algorithm, Continuous optimization, Convergence rate, Markov chain, Approximation error

1. Introduction

In both empirical and theoretical studies of evolutionary algorithms (EAs), a fundamental question is: how fast does an EA converge to the optimal solution of an optimization problem? In discrete optimization, this can be measured by computational time, by either the hitting time (the number of generations) or running time (the number of fitness evaluations) when an EA first finds an optimal solution [1]. In continuous optimization, however, computational time often is infinite because for many optimization problems, the number of optimal solutions is finite. So, computational time has to be modified to the time when EAs reach an ϵ-neighbor around the optimal solutions [2, 3, 4].

The convergence rate is an alternative way to evaluate the performance of EAs in continuous optimization. It quantifies how fast an EA converges to the optimal solution set per generation in the decision space. So far, numerous theoretical work has been reported to discuss the convergence rate from different perspectives [5, 6, 7, 8, 9, 10, 11].

⋆The first author was partly supported by the National Science Foundation of China (61303028) and the Fundamental Research Funds for the Central Universities (WUT: 2020IB006). The second author was supported by EPSRC under Grant No. EP/I009809/1.

⋆⋆The paper is an extension of our conference paper: Chen, Y. and He, J. (2019). Average Convergence Rate of Evolutionary Algorithms II. In *Int. Symposium on Intelligence Computation and Applications*. Springer.

*Corresponding author
Email addresses: ychen@whut.edu.cn (Yu Chen), jun.he@ntu.ac.uk (Jun He)

Because of the equivalence of convergence in decision space and that in the objective space, it is rational to investigate how fast the approximation error converges to zero in the objective space. The convergence rate discussed in this paper is defined in the objective space in terms of the approximation error. Denote the expected fitness value of the t^{th} generation population as f_t and the approximation error as $e_t = |f_t - f^*|$ where f^* represents the optimal fitness. The geometric convergence rate $e_t \leq e_0 c^t$ can be obtained from the one-step convergence rate (CR): e_t/e_{t-1} under the condition $e_t/e_{t-1} < c$ [6, 12]. But unfortunately, randomness in EAs results in the oscillation of e_t/e_{t-1}, which in turn hinders its practical application in computer experiments. Instead, the geometric average of e_t/e_{t-1} over consecutive t generations is proposed as the average convergence rate (ACR) [13]: $ACR_t = 1 - (e_t/e_0)^{1/t}$. A major advantage of ACR_t is that it is more stable than e_t/e_{t-1} in computer simulation.

The ACR has been adopted as a practical metric of the convergence speed of EAs in continuous optimization [14, 15, 16, 17]. Although some theoretical results have been obtained for the ACR in discrete optimization [13], there is no analysis of the ACR in continuous optimization. The current paper aims to extend the study from discrete optimization to continuous optimization and to answer the following research questions: When does the ACR converge to zero? When not? What is the relationship between the ACR and the decision space dimension?

The paper is organized as follows: Section 2 reviews the related work. Section 3 presents an empirical study of the ACR. Section 4 provides a general theoretical study of the ACR. Section 5 analyses the ACR of evolutionary programming. Section 6 investigates the relation between the ACR and decision space dimension. Finally, Section 7 concludes the paper.

2. Brief Literature Review of Convergence Rate

The convergence rate of EAs has been investigated from different perspectives and in varied terms. He, Kang and Ding [7, 18] studied the convergence in distribution by considering sequence $\{\| \mu_t - \pi \|, t = 1, 2, \dots\}$ where μ_t is the probability distribution of the t^{th} generation population X_t and π a stationary probability distribution. Based on the Doeblin condition, they obtained an upper bound $(1-\delta)^{t-1}$ on $\| \mu_t - \pi \|$ for some $\delta \in (0, 1)$. He and Yu [8] derived lower and upper bounds on $1 - \mu_t(X^*_\delta)$ where $\mu_t(\mathcal{S}^*_\delta)$ denotes the probability of X_t entering in a δ-neighbor of X^* where X^* denotes the set of optimal solutions.

Rudolph [5] compared Gaussian and Cauchy mutation for minimization of the sphere function in terms of the rate of local convergence, $\mathbb{E}[\min\{\| X_{t+1} \|^2 / \| X_t \|^2, 1\} \mid X_t]$, where $\| \cdot \|$ denotes the Euclidean norm. He proved that the rate is identical for Gaussian and spherical Cauchy distributions, whereas nonspherical Cauchy mutations lead to slower local convergence. Rudolph [6] also proved under the condition $e_t/e_{t-1} \leq c < 1$, the sequence $\{e_t\}$ converges in mean geometrically fast to 0, that is, $q^t e_t = o(1)$ for some $q > 1$. For a superset of the class of quadratic functions, sharp bounds on the convergence rate are obtained.

Semenov and Terkel [19] studied the convergence velocity of a simple EA with self-adaptation using a stochastic Lyapunov function and martingale theory. They proved that the velocity is asymptotically exponential $|x_t| \leq \exp(-at)$ on the class of unimodal functions with the aid of Monte Carlo simulation.

Beyer [20] developed a systematic theory of evolutionary strategies (ES) based on the progress rate and quality gain. The progress rate measures the distance change to the optimal solution in one generation, $\mathbb{E}[\| X_t - X^* \| - \| X_{t-1} - X^* \|]$. The quality gain is the fitness change in one generation, $\mathbb{E}[\bar{f}(X_t) - \bar{f}(X_{t-1})]$, where $\bar{f}(X)$ is the fitness mean of individuals in population X. Beyer et al. [21, 22] analyzed dynamics of ES with cumulative step size adaption and ES with self-adaption and multi-recombination on the ellipsoid model and derived the quadratic progress rate. Akimoto et al.[23] investigated ES with weighted recombination on general convex quadratic functions and derived the asymptotic quality gain. However, Auger and Hansen [24] argued the limit of the predictions using the progress rate.

Auger and Hansen [25] developed the theory of ES from a new perspective using the stability of Markov chains. Auger [9] investigated the $(1, \lambda)$-SA-EA on the sphere function and proved the convergence of $(\ln \| X_t \|)/t$ based on Foster-Lyapunov drift conditions. Jebalia et al. [26] investigated convergence rate of the scale-invariant (1+1)-ES in minimizing the noisy sphere function and proved a log-linear convergence rate in the sense that: $(\ln \| X_t \|)/t \to \gamma$ for some γ as $t \to +\infty$. Auger and Hansen [10] further investigated the

comparison-based step-size adaptive randomized search on scaling-invariant objective functions and proved as $t \to +\infty$, $\ln(\| X_t \| / \| X_0 \|_t)/t \to -CR$ for some positive CR. This log-linear convergence can be regarded as an extension of the classical average rate of convergence in deterministic iterative methods [27].

The above convergence rates are 'evaluated by Markov chain analysis, however, process is complicated from theoretical and practical point of view' [16]. Unlike them, the ACR has an attention to a close link with the practice. Although optional to define convergence rate in various spaces, it is preferred in this paper to investigate it in the objective space. With respect to many applications of EAs, their performance is often evaluated by the approximation error of obtained solutions. Thus, the convergence rate of EAs is defined as the average reduction rate of the approximation error over consecutive generations in the objective space [13]. For discrete optimization, it has been proved [13] that under particular initialization, the ACR is equal to the spectral radius of a matrix corresponding to transition probabilities within non-optimal solutions and under random initialization (all solutions can be chosen into the initial population with a positive probability), the ACR converges to this spectral radius. However, there is no similar analysis for the ACR in continuous optimization.

3. Empirical Study of Average Convergence Rate

3.1. Definition of Average Convergence Rate

Consider a minimization problem:

$$\min f(\mathbf{x}), \quad \mathbf{x} = (x_1, \cdots, x_d) \in \mathcal{D} \subset \mathbb{R}^d, \tag{1}$$

where $f(\mathbf{x})$ is a continuous function and $\mathcal{D}$ is the definition domain (called the decision space) and is bounded. d is the dimension. Denote the minimal function value as $f^* := \min\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{D}\}$ and the optimal solution set as $X^* := \{\mathbf{x} \in \mathcal{D} \mid f(\mathbf{x}) = f^*\}$.

Algorithm 1 describe a general framework of EAs. In EAs, an individual is a single point (solution) $\mathbf{x}$. A population X is a union of finite individuals $X = (\mathbf{x}_1, \cdots, \mathbf{x}_\mu)$ where μ is the population size. An optimal population X satisfies $X \cap X^* \neq \emptyset$ and a non-optimal population X satisfies $X \cap X^* = \emptyset$. Let $\mathcal{S}$ denote the set of all populations and $\mathcal{S}^*$ denote the set of optimal populations.

Algorithm 1 Evolutionary Algorithm

1: generation counter $t \leftarrow 0$;
2: $X_0 \leftarrow$ initialize a population of individuals;
3: **while** the stopping criterion is not satisfied **do**
4: $\quad X_{t+1} \leftarrow$ generate a population of individuals from X_t subject to a conditional transition probability $\Pr(X_{t+1} \mid X_0, \cdots, X_t)$;
5: $\quad t \leftarrow t + 1$;
6: **end while**

Given an initial population X_0, the *fitness* of the population X_t is $f(X_t \mid X_0) := \min\{f(\mathbf{x}) \mid \mathbf{x} \in X_t\}$, and its *approximation error* is $e(X_t \mid X_0) := |f(X_t) - f^*|$. Thereafter, $f(X_t \mid X_0)$ and $e(X_t \mid X_0)$ are denoted by $f(X_t)$ and $e(X_t)$ in short respectively. An EA using *elitist selection* always keeps the best found individual, that is, $e(X_{t+1}) \leq e(X_t)$. Let $f_t := \mathbb{E}[\mathbb{E}[f(X_t)] \mid X_0]$ and $e_t := \mathbb{E}[\mathbb{E}[e(X_t) \mid X_0]]$ denote the expected values of the fitness and approximation error respectively. In computer simulation, f_t is calculated as the the average value over a number of runs.

An EA is called *convergent in mean* [6] if starting from any X_0, $\lim_{t\to+\infty} e_t = 0$. An EA is called *convergent almost surely* [6] if starting from any X_0, the probability $\Pr(\lim_{t\to+\infty} e(X_t) = 0) = 1$. Given an approximation error sequence $\{e_t; t = 0, 1, \cdots\}$, its *one-step convergence rate* (CR) is the reduction rate of the approximation error at at the t^{th} generation.

$$CR_t := \frac{e_t}{e_{t-1}}, \quad t \in \mathbb{Z}^+, \tag{2}$$

where $\mathbb{Z}^+$ denotes the set of positive integers. The *average convergence rate for t generations* (ACR) [13] is defined by

$$ACR_t := 1 - \left(\frac{e_t}{e_0}\right)^{1/t} = 1 - \left(\prod_{k=1}^{t} CR_k\right)^{1/t}, \quad t \in \mathbb{Z}^+. \tag{3}$$

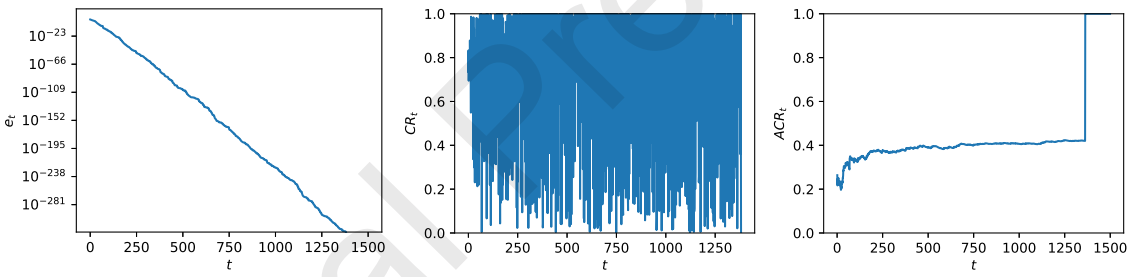
In (3), the term $(e_t/e_0)^{1/t}$ represents a geometric average of the CR over t consecutive generations. $1 - (e_t/e_0)^{1/t}$ normalizes the average between $(-\infty, 1]$. This rate can be regarded as the convergence speed because the larger ACR, the faster convergence. A negative value of the ACR means the EA moves away from the optimal point. $ACR_t = 1$ when $e_t = 0$ or equivalently the optimal solution is generated.

Similar to the convergence rate in deterministic iterative methods [27, Definition 3.1], the average convergence rate of EAs can be defined in the logarithmic form [13].

$$ACR'_t := -\frac{1}{t} \ln \frac{e_t}{e_0}. \tag{4}$$

However, this rate can be not adopted in computer simulation because its value is $+\infty$ if $e_t = 0$.

Figure 1 compares the approximation error, CR and ACR through an example. Figure 1(a) shows e_t decreases as t, but it does not quantify the convergence speed. Figure 1(b) shows CR_t oscillates significantly as t. Figure 1(c) depicts that ACR_t is more stable because it averages the CR values over consecutive generations. The ACR increases from 0.2 to 04, then jumps to 1 when the optimal solution is found.



(a) (b) (c)

Figure 1: (a) e_t converges to 0. (b) CR_t oscillates between 0 and 1. (c) ACR_t is stable at a constant 0.4.

Note: A jump to 1 is observed in Figure 1(c). It means that the EA has found the optimal solution. This phenomenon could happen in computer simulation because the number of runs is finite and the EA may find the optimum in all runs. However, in theory, it never happens. e_t is the expected value of $e(X_t)$. For a randomized search algorithm, its error e_t can be infinitely close to 0 but not equal to 0, thus, no jump exists. This claim can be validated through increasing the number of runs.

There is an essential difference between the ACR and CR. Some sequence $\{CR_t; t \in \mathbb{Z}^+\}$ oscillates, but the sequence $\{ACR_t; t \in \mathbb{Z}^+\}$ is still stable. A classical method in literature is to bound CR_t such as $\alpha \le CR_t \le \beta$ for any t, and derive the geometric convergence rate as $e_0\alpha^t \le e_t \le e_0\beta^t$ [6]. But the lower and upper bounds only provide a range of the convergence rate and are not sufficient for quantifying the convergence rate. The ACR is the geometric average of the CR and its value is between α and β.

$$\alpha \le \min_{k \le t} CR_k \le ACR_t \le \max_{k \le t} CR_k \le \beta. \tag{5}$$

In practice, the calculation of ACR_t is easy and it results in an exact expression $e(t) = e_0(ACR_t)^t$. But using CR_t only is difficult to obtain the same expression.

In the above ACR definition, the optimal fitness value f^* is required. In case of f^* unknown, the *alternative average convergence rate* is also introduced in [13].

$$ACR_t^\dagger := 1 - \left|\frac{f_{t+\tau} - f_t}{f_t - f_{t-\tau}}\right|^{1/\tau}, \quad \text{if } t \geq \tau, \tag{6}$$

where τ is an appropriate and user-defined time interval. In computer simulation, it is necessary to set τ to a large value because $ACR_t^\dagger$ with a small τ suffers big noise. A drawback of the value of the alternative ACR is not available for $t < \tau$.

3.2. Empirical Study of ACR and Alternative ACR

Using the ACR or alternative ACR, it is convenient to quantify and visualize the convergence speed of EAs. Let us show this claim through computer simulation.

For the purpose of illustration, consider (1+1) evolutionary programming (Algorithm 2) for minimizing the 2-d sphere function as an example. The sphere function is a unimodal function which is often used as a benchmark in EAs [28].

$$\min f_S(\mathbf{x}) = x_1^2 + x_2^2, \quad \mathbf{x} \in \mathbb{R}^2. \tag{7}$$

The minimal point to this function is $\mathbf{x}^* = (0, 0)$ with $f^* = 0$.

Evolutionary programming (EP) is a type of EAs which employs mutation and selection but without recombination. (1+1) EP is equivalent to (1+1) evolutionary strategies (ES) without crossover. However, a population-based ES employs a recommendation operator [29].

Algorithm 2 (1+1) Evolutionary Programming

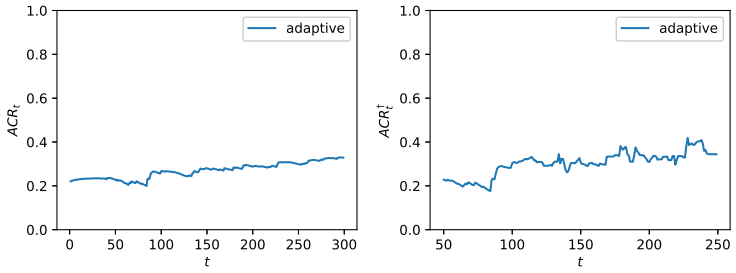
1: generation counter $t \leftarrow 0$;
2: initialize an individual $\mathbf{x}_0$;
3: **while** the maximal number of generations is not reached **do**
4: generate a new individual $\mathbf{y}_t = \mathbf{x}_t + \mathbf{z}_t$ where $\mathbf{z}_t$ obeys a probability distribution (such as Gaussian, Cauchy or Lévy) distribution;
5: select the best one from $\mathbf{y}_t$ and $\mathbf{x}_t$ as $\mathbf{x}_{t+1}$;
6: $t \leftarrow t + 1$;
7: **end while**

A child $\mathbf{y}$ is generated by Gaussian mutation $\mathbf{y} = \mathbf{x} + \mathbf{z}$, where $\mathbf{z} = (z_1, \cdots, z_d)$ obeys the Gaussian probability distribution $z_i \sim \mathcal{N}(0, \sigma_i)$. Two variants of Gaussian mutation are chosen with different settings of $\boldsymbol{\sigma} = (\sigma_1, \cdots, \sigma_d)$.

- **Invariant Gaussian mutation:** $\boldsymbol{\sigma}$ is set to constants, that is, for any i, σ_i is a constant. The (1+1) EP using invariant Gaussian mutation is called the *(1+1) invariant EP*.
- **Adaptive Gaussian mutation:** $\boldsymbol{\sigma}$ varies as $\mathbf{x}$, that is, $\boldsymbol{\sigma}$ is a function of $\mathbf{x}$. The (1+1) EP using adaptive Gaussian mutation is called the *(1+1) adaptive EP*.

In computer simulation, set $\sigma_i = 1$ in the (1+1) invariant EP and $\sigma_i = |x|_i$ in the (1+1) adaptive EP. The adaption is based on commonsense in the design of EAs: as $\mathbf{x}$ is close to the optimal solution 0, $\boldsymbol{\sigma}$ is set to a small value. This is equivalent to the practical strategy: as t increases, $\mathbf{x}_t$ is close to the optimal solution, then $\boldsymbol{\sigma}$ is reduced. $\mathbf{x}_0$ is randomly generated in $[-20, 20]^2$. Each algorithm runs $1,000$ times independently and f_t is the average over the 1,000 runs. The maximum number of generations is 300. The time interval τ for calculating the alternative ACR is chosen to be 50.

The first experiment is to compare the ACR and alternative ACR of the adaptive EP on the sphere function. Trend plots of the ACR and alternative ACR are illustrated in Figure 2. Figure 2(a) shows that

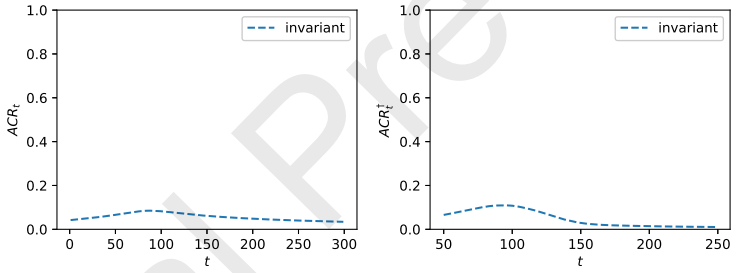


(a) (b)

Figure 2: (a) ACR_t of the (1+1) adaptive EP on the sphere function.(b) $ACR_t^\dagger$

the ACR of the adaptive EP increases to a positive, while Figure 2(b) depicts similar tends of the alternative ACR. But $ACR_t^\dagger$ suffers a little bigger noise than ACR_t due to the introduction of the τ-th order difference.

The second experiment is to compare the ACR and alternative ACR of the invariant EP on the sphere function. Trend plots of the ACR and alternative ACR are illustrated in figure 3. Figure 3(a) shows that the ACR decreases as time, while Figure 3(b) depicts a similar tend for the alternative ACR.



(a) (b)

Figure 3: (a) ACR_t of the (1+1) invariant EP on the sphere function. (b) alternative $ACR_t^\dagger$.

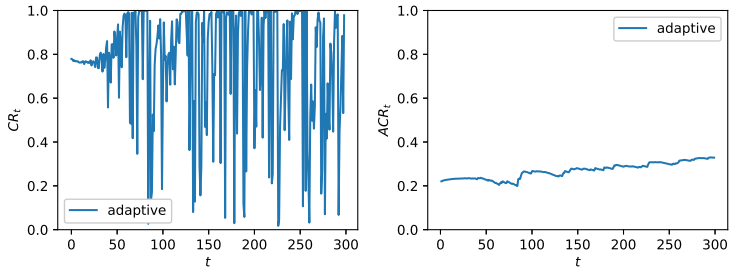
The results reveal that the adaptive EP converges faster than the invariant EP. For the adaptive EP, its ACR and alternative ACR both tend to a positive constant around 0.38 to 0.4. In general, an ACR is called *linear* if it tends towards a positive. But for the invariant EP, its ACR and alternative ACR decreases to a smaller constant and eventually towards 0. In general, an ACR is called *sublinear* if it converges to 0.

3.3. Comparison between ACR and CR

The ACR is more stable than the CR in numerical calculation. Let us show the claim through computer simulation.

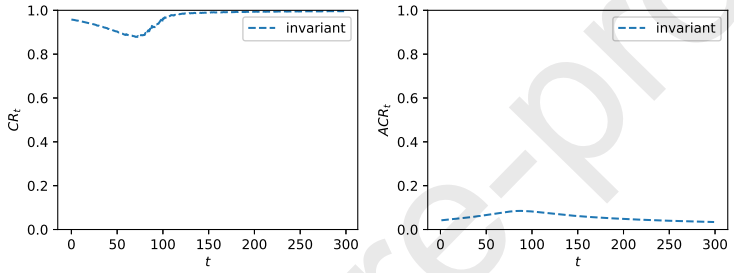
The first experiment is to compare the ACR and CR of the (1+1) adaptive EP on the sphere function. Experimental setting is the same as that in the previous subsection. Figure 4 illustrates trend plots of the ACR and CR. The CR fluctuates greatly between 0 and 1. It is impossible to quantify the convergence speed using the CR. But the ACR clearly converges to a positive constant around 0.38.

The second experiment is to compare the ACR and CR of the (1+1) invariant EP on the sphere function. Experimental setting is the same as that in the previous subsection. Figure 5 illustrates trend plots of the ACR and CR. The figure shows that the CR converges to 1 and the ACR to 0. The difference is caused by the normalization in the ACR to ensure the lower ACR, the slower convergence. Figure 5(b) reveals that the convergence speed of the (1+1) invariant EP eventually decreases to 0.



(a) (b)

Figure 4: (a) CR of (1+1) adaptive EP oscillates significantly on the sphere function. (b) ACR is more stable.



(a) (b)

Figure 5: (a) CR of the (1+1) invariant EP on the sphere function. (b) ACR.

The experimental results confirm that the numerical calculation of the ACR is more stable than the CR.

3.4. Relationship between ACR and Decision Space Dimension

The ACR of EAs will decrease as the decision space dimension increases. Let us verify this claim through computer simulation.

For the sake of illustration, consider a (1+1) random univariate search method (RUS), described in Algorithm 3, as an example. This algorithm can be regarded as a parallel version of random local search in discrete optimization because both make a random one-dimensional search at each generation. The RUS using adaptive Gaussian mutation is called the *adaptive RUS* in short.

Algorithm 3 (1+1) Random Univariate Search

1: $t \leftarrow 0$;
2: initialize a solution $\mathbf{x}_0 = (x_1, \cdots, x_d)$;
3: **while** the maximal number of generations is not reached **do**
4: choose one index $j \in \{1, \cdots, d\}$ at random, and generate a new solution by $\mathbf{y}_t = \mathbf{x}_t + \mathbf{z}_t$ where $\mathbf{z}_t = (z_1, \cdots, z_d)$, $z_j \sim \mathcal{N}(0, \sigma_j)$ is a Gaussian random variable and $z_i = 0$ for other $i \neq j$; if $\mathbf{y}_t$ is out of the definition domain, let $\mathbf{y}_t = \mathbf{x}_t$;
5: select the best one from $\mathbf{y}_t$ and $\mathbf{x}_t$ as $\mathbf{x}_{t+1}$;
6: $t \leftarrow t + 1$;
7: **end while**

Two test functions are used in computer simulation. The functions are inspired from the OneMax function and deceptive function in pseudo-Boolean function optimization. The OneMax function is the

easiest to a (1+1) elitist EA and the deceptive function is the hardest [30]. A variant OneMax function in continuous optimization is defined by

$$\max f_O(\mathbf{x}) := d - \sum_{i=1}^{d} x_i, \quad \mathbf{x} \in [0,1]^d. \tag{8}$$

where $\mathbf{x}^* = (0, \cdots, 0)$ and $f^* = d$.

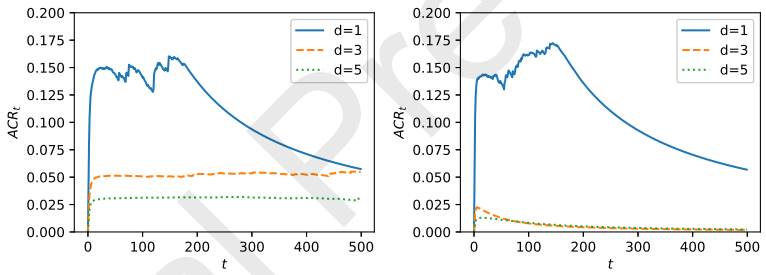
A deceptive function in continuous optimization is defined by

$$\max f_D(\mathbf{x}) = \begin{cases} \sum_{i=1}^{d} x_i, & \text{if } \sum_{i=1}^{d} x_i \geq 1/2, \\ d + 1 - \sum_{i=1}^{d} x_i, & \text{if } \sum_{i=1}^{d} x_i < 1/2, \end{cases} \quad \mathbf{x} \in [0,1]^d. \tag{9}$$

The global optimum is $\mathbf{x}^* = (0, \cdots, 0)$ and $f^* = d+1$. The basin of attraction of $\mathbf{x}^*$ is $\{\mathbf{x} \mid \sum_{i=1}^{d} x_i \leq 1/2\}$. The deceptive function (9) has a local optimum at $(1, \cdots, 1)$.

In computer simulation, set $\sigma_j = x_j$ for the selected index j in the adaptive RUS. $\mathbf{x}_0$ is chosen from $[0,1]^d$ at random. The algorithm is run 2,000 times independently on each test function. f_t is the average over the 2,000 run. The maximum number of generations is 500.

The experiments is to compare the ACR between the variant OneMax function and deceptive function. Trend plots of the ACR are illustrated in Figure 6. Figure 6(a) shows that the ACR on the variant OneMax function converges to some positive constants over generations for $d = 1, 3, 5$. But Figure 6(b) depicts the ACR on the deceptive function decreases quickly as the dimension d increases.



(a) (b)

Figure 6: (a) The ACR of adaptive RUS decays slowly as dimension $d = 1, 3, 5$ on the variant OneMax function. (b) But it decays quickly as dimension $d = 1, 3, 5$ on the deceptive function.

The experimental results demonstrate that the ACR decreases as the dimension increases. For an easy function like the OneMax function, the ACR decreases slowly as the dimension. In theory, it is expected that this ACR is larger than the reciprocal of a polynomial function of d. But for a hard function like the deceptive function, the ACR decreases quickly as the dimension increases. In theory, it is expected that this ACR is less than the reciprocal of an exponential function of d.

4. General Theoretical Study of Average Convergence Rate

4.1. Advantage of ACR over CR

Computer simulation in Section 3.3 shows an advantage of the ACR over the CR, that is, when a CR sequence oscillates, the ACR sequence is still stable. This subsection explains this difference.

The terms of linear, sub-linear or super-linear convergence has been used in describing the convergence speed of an iterative sequence. In EAs, a sequence $\{CR_t, t \in \mathbb{Z}^+\}$ converges *linearly* if $\lim_{t\to+\infty} CR_t = C < 1$ or converges *sublinearly* if $\lim_{t\to+\infty} CR_t = 1$. Similarly, a sequence $\{ACR_t, t \in \mathbb{Z}^+\}$ converges *linearly* if $\lim_{t\to+\infty} ACR_t = C < 1$ or converges *sublinearly* if $\lim_{t\to+\infty} ACR_t = 1$.

From the definition

$$ACR_t = 1 - \left(\prod_{k=1}^{t} CR_k\right)^{1/t}, \tag{10}$$

the linear convergence of the ACR can be derived from the linear convergence of the CR. But the inverse does not hold. This is the main advantage of the ACR over the CR. Let us verify the claim using an example.

Let us consider an example which is (1+1) EP (Algorithm 4) combining Cauchy and Gaussian mutation together. Because Cauchy and Gaussian mutation operators have different probability density functions (PDFs), combination of them could result in faster convergence [31, 32, 33, 34].

Algorithm 4 (1+1) Evolutionary Programming

1: generation counter $t \leftarrow 0$;
2: initialize an individual $\mathbf{x}_0$;
3: **while** the maximal number of generations is not reached **do**
4: generate a new individual by Gaussian mutation $\mathbf{y}_t = \mathbf{x}_t + \mathbf{z}_t$ where $\mathbf{z}_t$ obeys a probability distribution; if $\mathbf{y}_t$ is beyond the definition domain $\mathcal{D}$, let $\mathbf{y}_t = \mathbf{x}_t$;
5: select the best one from $\mathbf{y}_t$ and $\mathbf{x}_t$ as $\mathbf{x}_{t+1}$;
6: $t \leftarrow t + 1$;
7: **end while**

Two mutation operators are alternately used in this (1+1) EP, that is, Cauchy mutation is applied when t is an odd number and Gaussian mutation is applied when t is an even number. Thus, z_t obeys Cauchy or Gaussian probability distribution, PDFs of which are $p_c = \frac{1}{\pi(1+x^2)}$ and $p_g = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, respectively.

The (1+1) EP is used to minimize a JUMP function which is a typical multi-modal problem. This function is similar to the JUMP function in pseudo-Boolean optimization [35]. The the optimal solution set is $\{x; |x| < 1\}$.

$$f_J(x) = \begin{cases} 0 & \text{if } |x| < 1, \\ 4 - |x| & \text{if } 1 \le |x| < 2, \\ |x| & \text{otherwise.} \end{cases} \tag{11}$$

For the sake of analysis, assume that the initial point $x_0 = 2$. After $t (t \ge 1)$ iterations, x_t either jump to the flat $\{|x| < 1\}$ or stay at $|x| = 2$. Thus, the reduction rate of expected error is (without loss of generality, let $x_t = 2$ or $|x_t| < 1$)

$$\frac{e_{t+1}}{e_t} = \frac{\mathbb{E}[e(x_{t+1})|x_t = 2] \Pr(x_t = 2) + 0 \cdot \Pr(|\mathbf{x}_t| < 1)}{2 \cdot \Pr(x_t = 2) + 0 \cdot \Pr(|x_t| < 1)} = \frac{1}{2}\mathbb{E}[e(x_{t+1})|\mathbf{x}_t = 2]. \tag{12}$$

Moreover, the conditional expectation of error change is for Cauchy mutation,

$$\mathbb{E}[e(x_t) - e(x_{t+1})|x_t = 2] = \int_{-1}^{1} (2-0)\frac{1}{\pi}\frac{1}{1+(x-2)^2} dx = \frac{2}{\pi}\left(\arctan 3 - \frac{\pi}{4}\right), \tag{13}$$

and for Gaussian mutation,

$$\mathbb{E}[e(x_t) - e(\mathbf{x}_{t+1})|x_t = 2] = \int_{-1}^{1} (2-0)\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 4\left(\Phi(1) - \frac{1}{2}\right), \tag{14}$$

where $\Phi(\cdot)$ is cumulative distribution function (CDF) of the standard Gaussian distribution. From (12), (13) and (14) we know that for any integer $k \ge 1$

$$CR_{2k} = \frac{e_{2k}}{e_{2k-1}} = C_a := 2\left(\Phi(1) - \frac{1}{2}\right),$$

and

$$CR_{2k+1} = \frac{e_{2k+1}}{e_{2k}} = C_b := \frac{3}{4} - \frac{1}{\pi}\arctan 3.$$

Since $C_a \neq C_b$, the sequence $\{CR_t\}$ oscillates and does not converge.

However, from

$$ACR_t = \begin{cases} 1 - (C_aC_b)^{1/2}, & \text{if } t = 2k, \\ 1 - (C_aC_b)^{1/2}(C_b/C_a)^{1/2t}, & \text{if } t = 2k+1, \end{cases} \quad k = 1, 2, \ldots$$

we get $ACR_t \to 1 - (C_aC_b)^{1/2}$ as $t \to +\infty$, that is, the sequence $\{ACR_t\}$ is convergent.

The above example shows that the CR sequence could oscillate and not converge in some adaptive EAs. However, the ACR sequence converges thanks to the average of the CR for consecutive generations. Furthermore, for analyzing the ACR, it is necessary to consider the multi-step error change beyond the one-step error change.

4.2. Assumptions in the Theoretical Study

In order to make a theoretical analysis, EAs under investigation are assumed to satisfy several conditions.

1. (Supermartingale). The expected approximation error does not increase. For any non-optimal X_0 and any t,
$$\mathbb{E}[e(X_{t+1}) \mid X_0, \cdots, X_t] \leq e(X_t). \tag{15}$$
The sequence $\{e_t; t = 0, 1, \cdots\}$ is a monotonically decreasing function of t. This condition is different from elitism in EAs which requires $e(X_{t+1}) \leq e(X_t)$. A direct consequence from this condition is $ACR_t \in [0, 1]$.
2. (Markov chain). The state of X_{t+1} depends on X_t only. For any X_0 and any t, the transition probability
$$\Pr(X_{t+1} \mid X_0, \cdots, X_t) = \Pr(X_{t+1} \mid X_t). \tag{16}$$
3. (Stochastic algorithm). Starting from any non-optimal X_0, for any t, it holds
$$\Pr(X_t \cap X^*) < 1, \quad e_t > 0 \tag{17}$$
If $\Pr(X_t \cap X^*) = 1$, an EA reaches the optimal set at the t^{th} generation with probability 1. It degenerates to a deterministic-like algorithm which is not discussed in this paper.
4. (Normal reduction). The reduction rate of the approximation error satisfies
$$\lim_{t \to +\infty} \left(\frac{e_t}{e_{t-1}}\right)^{1/t} = 1. \tag{18}$$
The condition is mild. If $\lim_{t\to+\infty}(e_t/e_{t-1})^{1/t} < c < 1$, then for a large t, $e_t/e_{t-1} < c^t$. For example, let $c = 0.9$ and $t = 1000$, we have $e_{1000} < 1.75 \times 10^{-46} e_{999}$. This rapid reduction rate is almost impossible in normal EAs.

Many EAs satisfy the above four conditions. By default, they are always assumed to be true in the theoretical study in this paper.

Markov chains associated with EAs can be classified into homogeneous and inhomogeneous, depending on whether genetic operators (mutation, crossover and selection) change over time [8]. This paper focuses on EAs which are modeled by homogeneous Markov chains as below.

1. The population sequence $\{X_t; t = 0, 1, \cdots\}$ is a *homogeneous Markov chain*, that is, for any t, any X and any subset $\mathcal{A} \subset \mathcal{S}$, the transition probability
$$\Pr(X_{t+1} \in \mathcal{A} \mid X_t = X) = P(X; \mathcal{A}). \tag{19}$$
Transition probabilities do not change over time.

2. A population subsequence $\{X_{\kappa t}, t = 0, 1, \cdots\}$ (where $\kappa \in \mathbb{Z}^+$) is a homogeneous Markov chain. Transition probabilities follow a periodic change over generations. Let $Y_t = X_{\kappa t}$.

$$\Pr(Y_{t+1} \in \mathcal{A} \mid Y_t = X) = P^{(\kappa)}(X; \mathcal{A}), \tag{20}$$
$$\Pr(X_{t+1} \in \mathcal{A} \mid X_t = X) \neq \Pr(X_{t+2} \in \mathcal{A} \mid X_{t+1} = X). \tag{21}$$

The original population sequence $\{X_t; t = 0, 1, \cdots\}$ is an inhomogeneous Markov chain.

The current paper will not discuss other types of EAs whose genetic operators change over time. The analysis of theses EAs needs further understanding of inhomogeneous genetic operators and advanced tools from inhomogeneous Markov chains or stochastic processes. This topic is left for future research.

4.3. Linear and Sublinear ACR

Computer simulation in Section 3.2 demonstrates different trend plots of the ACR for the (1+1) invariant EP and (1+1) adaptive EP. The ACR sequence has different limit properties. In practice, it is required that e_t converges to 0 but there is no need for ACR_t to converge as $t \to \infty$. Thus, the *limit superior* and *limit inferior* [36] of the sequence $\{ACR_t\}$, defined as (22) and (23) respectively, are introduced to describe the limit property.

$$\overline{\lim_{t\to+\infty}} ACR_t := \inf_{t\geq 0}\{\sup_{s\geq t}\{ACR_s : s \geq t\} : t \geq 0\}. \tag{22}$$
$$\underline{\lim_{t\to+\infty}} ACR_t := \sup_{t\geq 0}\{\inf_{s\geq t}\{ACR_s : s \geq t\} : t \geq 0\}, \tag{23}$$

where inf is the abbreviation of mathematical infimum and sup the abbreviation of supremum. An example of the limit superior and inferior is illustrated in Figure 7. Consider the sequence $R_t = 0.1\cos(t)\times(\exp(-0.1t)+0.1)$ where $t = 0, 1, 2, \cdots$. Although the sequence does not converge, its limit superior (-0.1) and inferior (-0.1) still exist.

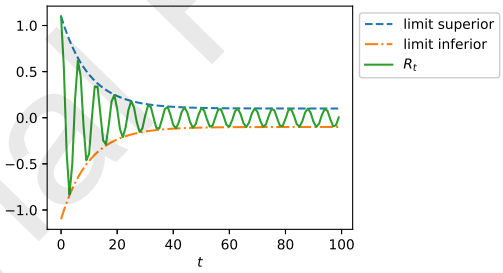


Figure 7: The limit superior and inferior of the sequence $R_t = \cos(t)(0.1 + \exp(-0.1t))$.

Existing theoretical results [37, 38, 10] show that the convergence rate of EAs in continuous optimization is up to linear. Similarly, the ACR of EAs can be classified into two categories via its limit property.

- The ACR is *linear*, if its limit inferior is a positive, that is, $\underline{\lim}_{t\to+\infty} ACR_t = C > 0$. In this case, the approximation error reduces geometrically fast to 0. For example, the ACR in Figure 2 is linear.
- The ACR is *sublinear*, if it asymptotically reduces to zero as $t \to +\infty$, that is, $\lim_{t\to+\infty} ACR_t = 0$. In this case the approximation error converges slowly to 0. For example, the ACR in Figure 3 is sublinear.

The ACR can be estimated using one-step transition or multi-step probability transition. Given an error sequence $\{e_t, t = 0, 1, \cdots\}$, the *one-step error change at the t^{th} generation* is $\Delta e_t := e_t - e_{t+1}$. It is similar to the quality gain [20] but the latter is defined on the fitness mean of individuals in a population $\bar{f}_t$ rather than the error mean e_t. The *rate of one-step error change* is $\Delta e_t/e_t$. Given a positive integer κ, the *κ-step error change at the t^{th} generation* is $\Delta^{(\kappa)} e_t := e_t - e_{t+\kappa}$. The *rate of κ-step error change* is $\Delta^{(\kappa)} e_t/e_t$.

Theorem 1. *(1) For the error sequence $\{e_t; t = 0, 1, \cdots\}$, if there exist a positive integer κ and $0 < C < 1$, the rate of κ-step error change satisfies*

$$\lim_{t\to+\infty} \frac{\Delta^{(\kappa)} e_t}{e_t} = 1 - C > 0, \tag{24}$$

then the ACR is linear, that is, $\lim_{t\to+\infty} ACR_t = 1 - C^{1/\kappa} > 0$.

(2) For the error sequence $\{e_t; t = 0, 1, \cdots\}$, if

$$\lim_{t\to+\infty} \frac{\Delta e_t}{e_t} = 0, \tag{25}$$

then the ACR is sublinear, that is, $\lim_{t\to+\infty} ACR_t = 0$.

Proof. (1) While the rate of κ-step error change converges to a positive value, there are two different cases to be discussed for the number of iteration t.

1. If $t = m\kappa$ for an integer $m > 0$, we have

$$ACR_{m\kappa} = 1 - \left(\prod_{k=1}^{m\kappa} \frac{e_k}{e_k - 1}\right)^{1/m\kappa} = 1 - \left[\prod_{l=1}^{m} \left(1 - \frac{\Delta^{(\kappa)} e_{(l-1)\kappa}}{e_{(l-1)\kappa}}\right)\right]^{1/(m\kappa)}. \tag{26}$$

Note that m tends to $+\infty$ when $t \to +\infty$. Then, from (24) and (26) we know

$$\lim_{m\to+\infty} ACR_{m\kappa} = 1 - C^{1/\kappa}. \tag{27}$$

2. If $t = m\kappa + k$ for integers m, k such that $m > 0$, $0 < k < \kappa$, we know

$$ACR_{m\kappa+k} = 1 - \left(\prod_{k=1}^{m\kappa+k} \frac{e_k}{e_{k-1}}\right)^{1/(m\kappa+k)} = 1 - \left[\prod_{l=1}^{m} \left(1 - \frac{\Delta^{(\kappa)} e_{(l-1)\kappa}}{e_{(l-1)\kappa}}\right)\left(\frac{e_{m\kappa+k}}{e_{m\kappa}}\right)\right]^{1/(m\kappa+k)} \tag{28}$$

From (24) we know that $\lim_{t\to+\infty} \frac{e_{t+\kappa}}{e_t} = C$, which implies that there exists $t_0 > 0$ such that

$$\frac{e_{t+\kappa}}{e_t} > \frac{C}{2}, \quad \forall\, t > t_0.$$

Then, monotonicity of e_t says that $\exists m_0 > 0$,

$$\frac{C}{2} < \frac{e_{m\kappa+\kappa}}{e_{m\kappa}} \le \frac{e_{m\kappa+k}}{e_{m\kappa}} \le 1, \quad \forall\, m > m_0,$$

and we conclude that

$$\lim_{m\to+\infty} \left(\frac{e_{m\kappa+k}}{e_{m\kappa}}\right)^{1/(m\kappa+k)} = 1, \quad \forall\ 0 < k < \kappa. \tag{29}$$

Combining (24), (28) and (29) we know

$$\lim_{m\to+\infty} ACR_{m\kappa+k} = \lim_{m\to+\infty} \left\{1 - \left[\prod_{l=1}^{m} \left(1 - \frac{\Delta^{(\kappa)} e_{(l-1)\kappa}}{e_{(l-1)\kappa}}\right)\right]^{1/(m\kappa+k)}\right\} = 1 - C^{1/\kappa}. \tag{30}$$

From (27) and (30), we conclude that $\lim_{t\to+\infty} ACR_t = 1 - C^{1/\kappa}$.

(2) When $\underline{\lim}_{t\to+\infty} \Delta e_t/e_t = 0$, from the ACR definition, we know $\lim_{t\to\infty} ACR_t = 0$. □

To derive the linear ACR sequence, Condition (24) requires the rate of κ-step error change larger than a positive. This condition is weaker than the rate of the one-step error change larger than zero. But to derive the sublinear convergence of the ACR sequence, Condition (25) requires the rate of one-step error change to converge to 0 too.

4.4. Link between ACR and Alternative ACR

Computation simulation in Section 3.2 shows the ACR and alternative ACR have similar behaviors. There is a link between the ACR and alternative ACR. Under some conditions, their limits are identical.

Theorem 2. *For the error sequence $\{e_t\}$, if for a positive integer κ and $0 < C < 1$,*

$$\lim_{t\to+\infty} \frac{\Delta^{(\kappa)} e_t}{e_t} = 1 - C > 0, \tag{31}$$

and in the definition of the alternative ACR, choose τ to a multiple of κ,

$$\lim_{t\to+\infty} ACR_t = \lim_{t\to+\infty} ACR_t^{\dagger}.$$

Proof. Since τ is a multiple of κ, we have $\tau = m\kappa$ for a positive integer m. Then,

$$\begin{aligned}\lim_{t\to+\infty} ACR_t^{\dagger} &= 1 - \lim_{t\to+\infty} \left(\frac{e_t - e_{t+\tau}}{e_{t-\tau} - e_t}\right)^{1/\tau} \\ &= 1 - \lim_{t\to+\infty} \left[\frac{e_t}{e_{t-\tau}} \times \left(\frac{e_t - e_{t+\tau}}{e_t}\right) \div \left(\frac{e_{t-\tau} - e_t}{e_{t-\tau}}\right)\right]^{1/\tau}.\end{aligned}$$

Since

$$\begin{aligned}\lim_{t\to+\infty} \left(\frac{e_t}{e_{t-\tau}}\right)^{1/\tau} &= \lim_{k\to+\infty} \left(\prod_{l=1}^{m} \left(1 - \frac{\Delta e_{k+(l-1)\kappa}}{e_{k+(l-1)\kappa}}\right)\right)^{1/(m\kappa)} = C^{1/\kappa}, \\ \lim_{t\to+\infty} \left(\frac{e_t - e_{t+\tau}}{e_t}\right)^{1/\tau} &= \lim_{t\to+\infty} \left(1 - \frac{e_{t+\tau}}{e_t}\right)^{1/\tau} = (1 - C^m)^{1/(m\kappa)}, \\ \lim_{t\to+\infty} \left(\frac{e_{t-\tau} - e_t}{e_{t-\tau}}\right)^{1/\tau} &= \lim_{k\to+\infty} \left(1 - \frac{e_{t-\tau}}{e_t}\right)^{1/\tau} = (1 - C^m)^{1/(m\kappa)},\end{aligned}$$

we know that

$$\lim_{t\to+\infty} ACR_t^{\dagger} = 1 - C^{1/\kappa}.$$

It has been proved in Theorem 1 that

$$\lim_{t\to+\infty} ACR_t = 1 - \lim_{t\to+\infty} \left(\prod_{k=1}^{t} \frac{e_k}{e_{k-1}}\right)^{1/t} = 1 - C^{1/\kappa},$$

and thus, we get that $\lim_{t\to+\infty} ACR_t = \lim_{t\to+\infty} ACR_t^{\dagger} = 1 - C^{1/\kappa}$. □

The above theorem proves that the limit of ACR_t is identical to the limit of $ACR_t^{\dagger}$. If the f^* value is unknown, $ACR_t^{\dagger}$ can be used as a replacement of ACR_t.

5. Theoretical Analysis of Evolutionary Programming

5.1. Landscape-invariant and Landscape-adaptive Mutation

In Section 3.2, it is observed that the ACR of invariant EP tends to zero, while the ACR of adaptive EP tends to a positive constant. This section analyzes general EP (Algorithm 5) using landscape-invariant or landscape-adaptive mutation. We only consider genetic operators which are unchanged over time or have periodic change. So, the population sequence or a subsequence is a homogeneous Markov chain.

Algorithm 5 Evolutionary Programming

1: counter $t \leftarrow 0$;
2: initialize μ individuals $X_0 = (\mathbf{x}_1, \cdots, \mathbf{x}_\mu)$;
3: **while** the stopping criterion is not satisfied **do**
4: generate $Y_t = (\mathbf{y}_1, \cdots, \mathbf{y}_\mu)$ by mutation $Y_t = X_t + Z_t$ where $Z_t = (\mathbf{z}_1, \cdots, \mathbf{z}_\mu)$ is μ random variables; if $\mathbf{y}_i$ is out of the definition domain $\mathcal{D}$, let $\mathbf{y}_i = \mathbf{x}_i$;
5: evaluate each individual's fitness in population $X_t \cup Y_t$;
6: $X_{t+1} \leftarrow$ select μ individuals from $X_t \cup Y_t$, where the best individual is always selected;
7: counter $t \leftarrow t + 1$;
8: **end while**

Mutation $Y = X + Z$ can be characterized by probability transition. Given a population $X \in \mathcal{S}$ and a subset $\mathcal{A} \subset \mathcal{S}$, the transition probability kernel $P_g(X; \mathcal{A})$ [39] is defined as

$$P_g(X; \mathcal{A}) = \int_{\mathcal{A}} p_g(X; Y) dY,$$

where $p_g(X; Y)$ is the probability density function depicting the mutation transition from X to Y. Furthermore, the κ-step kernel is defined as

$$P_g^{(\kappa)}(X; \mathcal{A}) = \int_{Y \in \mathcal{A}} p_g^{(\kappa)}(X_{t+\kappa} = Y \mid X_t = X) dY.$$

In this paper, we assume it is continuous and bounded.

Generally, mutation operators can be classified into two categories.

- **Landscape-invariant mutation**: mutation $Y = X + Z$ is called *landscape-invariant* if Z is a random variable vector whose joint probability density function $p_g(0; Z)$ is independent on X. For simplicity, denote $p_g(0; Z)$ as $p_z(Z)$. EP using invariant mutation is named *invariant EP*. For example, the Gaussian mutation using invariant σ in Section 3.2 belongs to invariant mutation.
- **Landscape-adaptive mutation**: if the probability density function of Z varies on X, mutation $Y = X + Z$ is called *landscape-adaptive*. EP using adaptive mutation is named *adaptive EP*. For example, the Gaussian mutation using adaptive σ in Section 3.2 is adaptive mutation.

Given a contraction factor $\rho \in (0, 1]$ and a population X, the population set $\mathcal{S}$ can be divided into two disjoint subsets:

$$\mathcal{S}(X, \rho) = \{Y \in \mathcal{S} | e(Y) < \rho e(X)\}, \quad \overline{\mathcal{S}}(X, \rho) = \mathcal{S} \setminus \mathcal{S}(X, \rho). \tag{32}$$

The set $\mathcal{S}(X, \rho)$ is named as a *ρ-promising region*. When $\rho = 1$, the set $\mathcal{S}(X, 1)$ is called a *promising region*.

5.2. Analysis of Landscape-invariant EP

For EP using landscape-invariant mutation, we prove that its ACR converges to 0. First we demonstrate that the infinum of the transition probability to the promising region is zero under a mild condition.

Lemma 1. *If the number of optimal solutions is finite, then the transition probability of the landscape-invariant EP to the promising region satisfies*

$$\inf\{P_g(X, S(X, 1)); X \notin \mathcal{S}^*\} = 0. \tag{33}$$

Proof. In order to prove (33), it is sufficient to prove $\lim_{e(X) \to 0} P_g(X, S(X, 1)) = 0$. That is, $\forall \varepsilon > 0, \exists \delta > 0$, $\forall X \in \mathcal{A}(\mathcal{S}^*, \delta) \setminus \mathcal{S}^*$ (where the set $\mathcal{A}(\mathcal{S}^*, \delta) = \{X; e(X) \leq \delta\}$), it holds

$$P_g(X, \mathcal{S}(X, 1)) < \varepsilon. \tag{34}$$

For a Lebesgue-measurable set $\mathcal{A} \subset \mathcal{S}$, let $m(\mathcal{A})$ denote its Lebesgue measure. Because $p_z(Z)$ is continuous and bounded, the probability of $X+Z$ falling in a small area is small for a fixed X. That is, $\forall \varepsilon > 0$, $\exists \delta' > 0$ (set $\delta' = \varepsilon / \sup p_z(Z)$), it holds $\forall \mathcal{A} \subset \mathcal{S} : m(\mathcal{A}) \leq \delta'$ and $\forall X \in \mathcal{S}$,

$$\Pr(X + Z \in \mathcal{A}) = \int_{Z: X+Z \in \mathcal{A}} p_z(X+Z) dZ < \varepsilon. \tag{35}$$

Because the number of optimal solutions is finite (then $m(\mathcal{S}^*) = 0$) and f is continuous, for the set $\mathcal{A}(\mathcal{S}^*, \delta)$, we can choose δ sufficiently small so that $m(\mathcal{A}(\mathcal{S}^*, \delta)) \leq \delta'$.

Because f is continuous, we may set δ sufficiently small so that $\forall X \in \mathcal{A}(\mathcal{S}^*, \delta)$ and $Y \notin \mathcal{A}(\mathcal{S}^*, \delta)$: $f(X) < f(Y)$. This implies that $\mathcal{S}(X, 1) \subset \mathcal{A}(\mathcal{S}^*, \delta)$. According to (35) and $m(\mathcal{A}(\mathcal{S}^*, \delta)) \leq \delta'$, $\forall X \in \mathcal{A}(\mathcal{S}^*, \delta) \setminus \mathcal{S}^*$, we have

$$\Pr(X + Z \in \mathcal{A}(\mathcal{S}^*, \delta)) < \varepsilon.$$

Because $\mathcal{S}(X, 1) \subset \mathcal{A}(\mathcal{S}^*, \delta)$, we have

$$P_g(X, S(X, 1)) \leq \Pr(X + Z \in \mathcal{A}(\mathcal{S}^*, \delta)) < \varepsilon.$$

Then we get (34), and the proof is completed. □

The theorem below analyzes the limit property of invariant EP.

Theorem 3. *If the number of optimal solutions is finite and the invariant EP using elitist selection converges in mean, then starting from any X_0, $\lim_{t \to +\infty} ACR_t = 0$.*

Proof. According to Theorem 1, it is sufficient to prove $\lim_{t \to +\infty} \Delta e_t / e_{t-1} = 0$. That is $\forall \varepsilon > 0, \exists t_0 > 0$, $\forall t \geq t_0$,

$$\Delta e_t < \varepsilon e_t. \tag{36}$$

From (34) in Lemma 1, we know $\forall \varepsilon > 0$, $\exists \delta > 0$, let $\mathcal{A}(\mathcal{S}^*, \delta) = \{X; e(X) \leq \delta\}$. Then $\forall X \in \mathcal{A}(\mathcal{S}^*, \delta) \setminus \mathcal{S}^*$, it holds

$$P_g(X, \mathcal{S}(X, 1)) < \varepsilon. \tag{37}$$

Since the sequence $\{e_t; t = 0, 1, \cdots\}$ converges to 0, EP converges almost surely to 0, that is,

$$\Pr(\lim_{t \to +\infty} e(X_t) = 0) = 1.$$

Denote

$$\mathcal{S}_1 = \{\omega \in \mathcal{S} | \lim_{t \to +\infty} e(X_t(\omega)) = 0\},\ \mathcal{S}_2 = \{\omega \in \mathcal{S} | \lim_{t \to +\infty} e(X_t(\omega)) \neq 0\}.$$

It is obvious that

$$\Pr(\omega \in \mathcal{S}_2) = 0, \tag{38}$$

and for the given $\delta > 0$, $\exists\, t_0 > 0$, then $\forall\, t > t_0$, it holds

$$e(X_t(\omega)) < \delta, \quad \forall\, \omega \in \mathcal{S}_1.$$

From (37) we know

$$P_g(X, \mathcal{S}(X_t(\omega), 1)) \leq \varepsilon, \quad \forall \omega \in \mathcal{S}_1.$$

Then we obtain

$$\mathbb{E}[e(X_t(\omega)) - e(X_{t+1}(\omega)) \mid X_t(\omega)] \leq \varepsilon e(X_t(\omega)), \quad \forall \omega \in \mathcal{S}_1. \tag{39}$$

While $\forall \omega \in \mathcal{S}_2$, we know there exists a positive B:

$$\mathbb{E}[e(X_t(\omega)) - e(X_{t+1}(\omega)) \mid X_t(\omega)] \leq B. \tag{40}$$

Combining (38), (39) and (40) together, we get

$$\begin{aligned}\Delta e_t =& \int_{\mathcal{S}_1} \mathbb{E}[e(X_t(\omega)) - e(X_{t+1}(\omega)) \mid X_t(\omega)] \Pr(d\omega) + \int_{\mathcal{S}_2} \mathbb{E}[e(X_t(\omega)) - e(X_{t+1}(\omega)) \mid X_t(\omega)] \Pr(d\omega) \\ \leq& \varepsilon \int_{\mathcal{S}_1} e(X_t(\omega)) \Pr(d\omega) + B \cdot 0 \leq \varepsilon e_t.\end{aligned}$$

So (36) is true, and we complete the proof. □

Theorem 3 states that for the invariant elitist EP, $ACR_t \to 0$ as $t \to +\infty$. This means that landscape-invariant mutation is less efficient in continuous optimization. This phenomenon can be explained by the lazy convergence for general Markov search [40]. As the population approaches the optimum, the probability of generating a better state from one step to another goes to zero. This causes the slow convergence.

Note: Theorem 3 does not hold if the number of the optimal set X^* is not finite, for example, the Jump function (11) in Section 4.1.

5.3. Analysis of Landscape-adaptive EP

Landscape-adaptive mutation can be further split into two categories according to the probability of locating promising regions.

- **Zero landscape-adaptive mutation**: adaptive mutation $Y = X + Z$ is called *zero landscape-adaptive* if the transition probability to the promising region satisfies

$$\inf\{P_g(X; \mathcal{S}(X, 1)); X \notin \mathcal{S}^*\} = 0. \tag{41}$$

- **Positive landscape-adaptive mutation**: adaptive mutation $Y = X + Z$ is called *positive landscape-adaptive* if $\exists \rho \in (0, 1)$, the transition probability to the ρ-promising region satisfies

$$C_\rho = \inf\{P_g^{(\kappa)}(X; \mathcal{S}(X, \rho)); X \notin \mathcal{S}^*\} > 0. \tag{42}$$

  for some positive integer κ.

Thus, landscape-adaptive EP can be classified into two categories: the *zero landscape-adaptive EP* employing zero landscape-adaptive mutation and the *positive landscape-adaptive EP* with positive landscape-adaptive mutation. Theorems 4 and 5 analyze the limit property of the ACR of the two types of EP.

Theorem 4. *If the number of optimal solutions is finite and the zero landscape-adaptive EP using elitist selection converges in mean, then starting from some X_0, the ACR satisfies $\lim_{t\to+\infty} ACR_t = 0$.*

Proof. For zero landscape-adaptive mutation, (41) implies that there exists a subsequence $\{X_{t'}\}$ such that $\lim_{t'\to+\infty} P_g(X_{t'}, S(X_{t'}, 1)) = 0$. Similar to the proof of Theorem 3, we know that $\lim_{t\to+\infty} ACR_{t'} = 0$. □

Theorem 4 states that for the zero landscape-adaptive elitist EP, its ACR tends to 0 when starting from some initial population. Thus, zero landscape-adaptive mutation is not always efficient. It is different from Theorem 3 which holds for any initial non-optimal population.

Theorem 5. *If the number of optimal solutions is finite and the positive landscape-adaptive EP using elitist selection converges in mean, then starting from any non-optimal population X_0, the ACR satisfies $\underline{\lim}_{t\to+\infty} ACR_t \geq C > 0$.*

Proof. Let us estimate the lower bound of the ACR limit. From (32), we know that for any $k \geq 0$,

$$\mathcal{S}(X_k, \rho) = \{Y \in \mathcal{S} \mid e(Y) \leq \rho e(X_k)\}.$$

It follows that $\mathcal{S}(X_k, \rho) \subset \mathcal{S}(X_k, 1)$, and for any $Y \in \mathcal{S}(X_k, \rho)$,

$$f(X_k) - f(Y) \geq (1 - \rho)(f(X_k) - f^*). \tag{43}$$

So we get for κ consecutive steps that

$$\begin{aligned}
&e(X_k) - e(X_{k+\kappa}) \\
=&\int_{\mathcal{S}(X_k,1)} (f(X_k) - f(Y)) p_g^{(\kappa)}(X_{t+\kappa} = Y \mid X_t = X) dY \\
\geq&\int_{\mathcal{S}(X_k,\rho)} (f(X_k) - f(Y)) p_g^{(\kappa)}(X_{t+\kappa} = Y \mid X_t = X) dY \\
\geq&\int_{\mathcal{S}(X_k,\rho)} (1-\rho)\left(f(X_k) - f^*\right) p_g^{(\kappa)}(X_{t+\kappa} = Y \mid X_t = X) dY \quad \text{(from (43))} \\
\geq&(1-\rho) C_\rho e(X_k). \quad \text{(from (42))}
\end{aligned} \tag{44}$$

Then

$$\frac{\Delta^{(\kappa)} e_k}{e_k} \geq \frac{(1-\rho) C_\rho \mathbb{E}[e(X_k)]}{e_k} = (1-\rho) C_\rho. \tag{45}$$

It holds that

$$\lim_{t\to+\infty} ACR_t = \lim_{t\to+\infty} \left[1 - \left(\frac{e_t}{e_0}\right)^{1/t}\right] = \lim_{t\to+\infty} \left[1 - \left(\prod_{k=1}^{t} \frac{e_k}{e_{k-1}}\right)^{1/t}\right] \geq (1-\rho) C_\rho = C.$$

Let $C = (1-\rho) C_\rho > 0$. We complete the proof. □

Theorem 5 proves that positive landscape-adaptive mutation is efficient because it results in a positive lower bound of the ACR. Theorems 3 to 5 confirms the different behaviors between the landscape-adaptive EP and landscape-invariant EP. To design an efficient EA, mutation should be positive landscape-adaptive.

5.4. Case Study of Landscape-invariant and Landscape-adaptive EP

This section makes a case study of EP using landscape-adaptive and landscape-invariant mutation. Consider the following minimization problem:

$$\min f_A(\mathbf{x}) = \|\mathbf{x}\|_\infty = \max\{|x_1|, |x_2|\}, \quad \mathbf{x} = (x_1, x_2) \in \mathbb{R}^2. \tag{46}$$

It is optimized by $(1+1)$ EP using mutation $\mathbf{y} = \mathbf{x} + \boldsymbol{z}\boldsymbol{\Gamma}$, where $\mathbf{z}$ is a random vector subject to a uniform distribution $\mathcal{U}([-1,1]\times[-1,1])$. $\boldsymbol{\Gamma} = diag\{\gamma_1, \gamma_2\}$ represents the step sizes along x_1 and x_2 axes respectively. Denote the individual at the t^{th} generation as $\mathbf{x}_t = (x_1, x_2)$. Due to symmetry of the landscape, we can postulate without loss of generality that $x_1 \geq x_2$. Then, $\|\mathbf{x}_t\|_\infty = x_1$.

Scenario 1: $\boldsymbol{\Gamma}$ is constant. In this case, the uniform mutation is landscape-invariant. Theorem 3 implies that ACR_t converge to 0 when $t \to +\infty$.

We first validate that this landscape-invariant (1+1) EP converges in probability. Let $\mathbf{x}_t = (x_1, x_2)$, set

$$\gamma_1 = |x_1|, \quad \gamma_2 = |x_2|.$$

$\forall 0 < \delta < \min\{\gamma_1, \gamma_2\}$, denote

$$\mathcal{R}_\delta = [-\delta, \delta] \times [-\delta, \delta].$$

The elitist selection confirms that

$$\Pr(\mathbf{x}_{t+1} \notin \mathcal{R}_\delta | \mathbf{x}_t \notin \mathcal{R}_\delta) \leq \Pr(\mathbf{x}_1 \notin \mathcal{R}_\delta | \mathbf{x}_0 \notin \mathcal{R}_\delta) = \frac{\delta^2}{4\gamma_1\gamma_2} \quad \forall t \in \mathbb{Z}^+.$$

Thus,

$$\begin{aligned}\Pr(\mathbf{x}_{t+1}\notin\mathcal{R}_\delta)&=\Pr(\mathbf{x}_{t+1}\notin\mathcal{R}_\delta|\mathbf{x}_t\notin\mathcal{R}_\delta)\Pr(\mathbf{x}_t\notin\mathcal{R}_\delta)\\&\leq\frac{\delta^2}{4\gamma_1\gamma_2}\Pr(\mathbf{x}_t\notin\mathcal{R}_\delta)\leq\cdots\leq\left(\frac{\delta^2}{4\gamma_1\gamma_2}\right)^{t+1}\Pr(\mathbf{x}^{(0)}\notin\mathcal{R}_\delta)\end{aligned}$$

and we know

$$\lim_{t\to\infty}\Pr(\mathbf{x}_{t+1}\notin\mathcal{R}_\delta)=0. \tag{47}$$

That is, this landscape-invariant (1+1) EP converges in probability.

From (47) we know $\forall\varepsilon_1>0$, there exists $t_0>0$ such that

$$\Pr(\mathbf{x}_t\notin\mathcal{R}_\delta)<\varepsilon_1,\quad t>t_0. \tag{48}$$

Meanwhile,

$$\begin{aligned}\mathbb{E}[e(\mathbf{x}_t)-e(\mathbf{x}_{t+1})|\mathbf{x}_t=(x_1,x_2)]&=\int_{\mathcal{S}(\mathbf{x}_t,1)}\frac{1}{4\gamma_1\gamma_2}(x_1-\max\{|y_1|,|y_2|\})dy_1dy_2\\&=\frac{2}{4\gamma_1\gamma_2}\int_{-x_1}^{x_1}\left[\int_{-x_1}^{y_1}(x_1-|y_1|)dy_2\right]dy_1=\frac{[x_1]^3}{2\gamma_1\gamma_2}.\end{aligned} \tag{49}$$

By setting $\epsilon=\frac{\delta^2}{\gamma_1\gamma_2}$, (48) and (49) imply that

$$\begin{aligned}\Delta e_t&=\mathbb{E}[\mathbb{E}[e(\mathbf{x}_t)-e(\mathbf{x}_{t+1})|\mathbf{x}_t=(x_1,x_2)]]\\&=\int_{\mathbf{x}_t\in\mathcal{R}_\delta}\frac{[x_1]^3}{2\gamma_1\gamma_2}d\mathbf{P}_{\mathbf{x}_t}+\int_{\mathbf{x}_t\notin\mathcal{R}_\delta}\frac{[x_1]^3}{2\gamma_1\gamma_2}d\mathbf{P}_{\mathbf{x}_t}\\&<\frac{\epsilon}{2}\mathbb{E}[e(\mathbf{x}_t)]+M\epsilon_1\mathbb{E}[e(\mathbf{x}_t)],\end{aligned}$$

where $M=\max\{\frac{[x_1]^2}{2\gamma_1\gamma_2}-\frac{\epsilon}{2}\}$. Hence, $\lim_{t\to\infty}\frac{\Delta e_t}{e_t}\leq\frac{\epsilon}{2}$. Since ϵ is arbitrarily small, we can conclude that

$$\lim_{t\to\infty}ACR_t=1-\lim_{t\to\infty}\left(\frac{e_t}{e_0}\right)^{1/t}=0.$$

Scenario 2: $\boldsymbol{\Gamma}$ is adaptive to the landscape. We consider an adaptive mutation strategy which dynamically adjusts the step size as below. Let $\mathbf{x}_t=(x_1,x_2)$,

$$\gamma_i(t)=\begin{cases}0.1|x_i|, & 2m\kappa\leq t<(2m+1)\kappa,\\0.2|x_i|, & 2(m+1)\kappa\leq t<(2m+2)\kappa,\end{cases}\quad m\in\mathbb{Z}^+,i=1,2, \tag{50}$$

where κ is a positive integer.

Because $\frac{\gamma_i(t)}{|x_i|}$ is bounded from below by 0.1, we have

$$\inf\{P_g(X;\mathcal{S}(X,1));X\notin\mathcal{S}^*\}>0. \tag{51}$$

Thus, the above strategy is positive landscape-adaptive. According to Thereom 5, there exists some positive C such that $\lim_{t\to+\infty}ACR_t\geq C$. Now let us deduce the lower and upper bounds on the limit of ACR_t. Set

$$\lambda=\begin{cases}0.1, & 2m\kappa\leq t<(2m+1)\kappa;\\0.2, & 2(m+1)\kappa\leq t<(2m+2)\kappa,\end{cases}\quad m\in\mathbb{Z},i=1,2.$$

The analysis is split into two different cases.

1. While $0 < (1+\lambda)x_2 < x_1$, we discuss two cases: $(1+\lambda)x_2 \le (1-\lambda)x_1$ and $(1-\lambda)x_1 < (1+\lambda)x_2$.
   (a) If $(1+\lambda)x_2 \le (1-\lambda)x_1$, we know

$$\begin{aligned}\mathbb{E}[e(\mathbf{x}_t) - e(\mathbf{x}_{t+1})|\mathbf{x}_t = (x_1, x_2)] =& \int_{\mathcal{S}(\mathbf{x}_t,1)} \frac{1}{4\gamma_1\gamma_2}(x_1 - \max\{|y_1|, |y_2|\})dy_1dy_2 \\ =& \frac{1}{4\lambda^2 x_1 x_2}\int_{(1-\lambda)x_1}^{x_1}\left[\int_{(1-\lambda)x_2}^{(1+\lambda)x_2}(x_1 - y_1)dy_2\right]dy_1 = \frac{\lambda}{4}x_1.\end{aligned}$$

   Then,

$$\frac{\Delta e_t}{e_t} = \frac{1}{4}\lambda. \tag{52}$$

   (b) If $(1-\lambda)x_1 < (1+\lambda)x_2$, we know

$$\begin{aligned}\mathbb{E}[e(\mathbf{x}_t) - e(\mathbf{x}_{t+1})|\mathbf{x}_t = (x_1, x_2)] =& \int_{\mathcal{S}(\mathbf{x}_t,1)} \frac{1}{4\gamma_1\gamma_2}(x_1 - \max\{|y_1|, |y_2|\})dy_1dy_2 \\ \le& \frac{1}{4\lambda^2 x_1 x_2}\int_{(1-\lambda)x_1}^{x_1}\left[\int_{(1-\lambda)x_2}^{(1+\lambda)x_2}(x_1^{(t)} - y_1)dy_2\right]dy_1 = \frac{\lambda}{4}x_1,\end{aligned}$$

   and

$$\begin{aligned}\mathbb{E}[e(\mathbf{x}_t - e(\mathbf{x}_{t+1})|\mathbf{x}_t] = (x_1, x_2] =& \int_{\mathcal{S}(\mathbf{x}_t,1)} \frac{1}{4\gamma_1\gamma_2}(x_1 - \max\{|y_1|, |y_2|\})dy_1dy_2 \\ \ge& \frac{1}{4\lambda^2 x_1 x_2}\int_{(1-\lambda)x_1}^{x_1}\left[\int_{(1-\lambda)x_2}^{(1-\lambda)x_1}(x_1 - y_1)dy_2\right]dy_1 = \frac{1}{8}\lambda(1-\lambda)x_1.\end{aligned}$$

   Then,

$$a_1(\lambda) \le \frac{\Delta e_t}{e_t} \le b_1(\lambda), \tag{53}$$

   where

$$a_1(\lambda) = \frac{1}{8}\lambda(1-\lambda), \qquad b_1(\lambda) = \frac{1}{4}\lambda.$$

2. While $(1+\lambda)x_2 \ge x_1 \ge x_2$, we know

$$\begin{aligned}\mathbb{E}[e(\mathbf{x}_t) - e(\mathbf{x}_{t+1})|\mathbf{x}_t] = (x_1, x_2)] =& \int_{\mathcal{S}(\mathbf{x}_t,1)} \frac{1}{4\gamma_1\gamma_2}(x_1 - \max\{|y_1|, |y_2|\})dy_1dy_2 \\ =& \frac{1}{4\lambda^2 x_1 x_2}\int_{(1-\lambda)x_1}^{x_1}\left[\int_{(1-\lambda)x_2}^{x_1}(x_1 - \max\{|y_1|, |y_2|\})dy_2\right]dy_1 \\ =& \frac{1}{4\lambda^2 x_1 x_2}\left[(\frac{1}{2} - \frac{1}{6}\lambda)(x_1)^3 - \frac{1}{2}(1-\lambda)(x_1)^2 x_2\right] \\ =& \frac{x_1}{4}\left[(\frac{1}{2} - \frac{1}{6}\lambda)\frac{x_1}{x_2} - \frac{1}{2}(1-\lambda)\right].\end{aligned}$$

   Then,

$$\frac{1}{12}\lambda x_1 \le \mathbb{E}[e(\mathbf{x}_t) - e(\mathbf{x}_{t+1})|\mathbf{x}_t] = (x_1, x_2)] \le \frac{1}{6}\lambda(5-\lambda)x_1,$$

   and it holds

$$a_2(\lambda) \le \frac{\Delta e_t}{e_t} \le b_2(\lambda), \tag{54}$$

   where

$$a_2(\lambda) = \frac{1}{12}\lambda, \qquad b_2(\lambda) = \frac{1}{6}\lambda(5-\lambda).$$

In summary, (52), (53) and (54) imply

$$\min\{\frac{1}{8}\lambda(1-\lambda), \frac{1}{12}\lambda\} \leq \frac{\Delta e_t}{e_t} \leq \frac{1}{6}\lambda(5-\lambda).$$

Since

$$ACR_t = 1 - \left(\prod_{k=0}^{t-1}(1-\frac{\Delta e_k}{e_k})\right)^{1/t},$$

The ACR is bounded by

$$\min\{\frac{1}{8}\lambda(1-\lambda), \frac{1}{12}\lambda\} \leq ACR_t \leq \frac{1}{6}\lambda(5-\lambda).$$

Set $\lambda = 0.1$ or $\lambda = 0.2$. Thus, the ACR of the landscape-adaptive EP bounded by

$$\frac{1}{120} \leq ACR_t \leq \frac{4}{25}. \tag{55}$$

6. Relationship between ACR and Decision Space Dimension

6.1. Polynomial ACR versus Exponential ACR

In Section 3.4, computer simulation demonstrates that the ACR decrease as the the decision space dimension increases. In theory, Teytaud and Selly [37] discussed the convergence by the logarithm of the distance to the optimum and proved a linear convergence with constant in $[1-O(1/d), 1]$ for the comparison-based algorithms. In general, the relationship between the ACR and decision space dimension can be studied similar to the time complexity. Recall in the theory of time complexity, the runtime of algorithms is often classified into two categories: exponential runtime and polynomial runtime as the problem input size. Similarly, the ACR of EAs can be classified into polynomial and exponential categories as follows.

- **Polynomial ACR**: starting from any X_0, for any t, ACR_t is not less than a *reciprocal of a polynomial function of* d, that is
$$\exists a > 1, \forall t \geq 0 : ACR_t = \Omega\left(d^{-a}\right), \tag{56}$$
where Ω is under Bachmann-Landau notation. (56) means ACR_t reduces slowly as d increases.

- **Exponential ACR**: starting from some X_0, for a period as long as an exponential function of d, ACR_t is not more than a *reciprocal of an exponential function of* d, that is
$$\exists a > 1, b > 1, \forall t = \Omega(b^d) : ACR_t = O\left(a^{-d}\right). \tag{57}$$
where O is under Bachmann-Landau notation. (57) means ACR_t reduces quickly as d increases.

According to the definition, a polynomial ACR is a linear ACR. However, an exponential ACR could be either a sublinear or linear ACR. For example, an exponential ACR such that $\lim_{t\to+\infty} ACR_t(d) = \exp(-d)$ is still a linear ACR.

The theorem below provides sufficient conditions of determining whether an ACR is polynomial and exponential.

Theorem 6. *(1) If starting from any X_0, for any $t \geq 0$, some $a > 1$ and $\kappa > 0$, the rate of κ-step error change satisfies*

$$\frac{\Delta^{(\kappa)} e_t}{e_t} = \Omega(d^{-a}). \tag{58}$$

then starting from any X_0, for any $t \geq 0$, $ACR_t = \Omega\left(d^{-a/\kappa}\right)$.

(2) If starting from some X_0, for some $t = O(b^d)$ where $b > 1$, some $a > 1$ and and $\kappa > 0$, the rate of error change satisfies

$$\frac{\Delta^{(\kappa)} e_t}{e_t} = O\left(a^{-d}\right). \tag{59}$$

then starting from the above X_0, for some $t = O(b^d)$, $ACR_t = O\left(a^{-d/\kappa}\right)$.

Proof. Denote $t = m\kappa + k$, where m is a positive integer, $k \in \{0, 1, \ldots, \kappa - 1\}$. Because

$$ACR_t = 1 - \left[\prod_{k=1}^{t} \frac{e_k}{e_{k-1}}\right]^{1/t} = 1 - \left[\prod_{l=0}^{m-1} \left(1 - \frac{\Delta^{(\kappa)} e_{l\kappa}}{e_{l\kappa}}\right) \frac{e_{m\kappa+k}}{e_{m\kappa}}\right]^{1/(m\kappa+k)},$$

from (58), we derive $ACR_t = \Omega(d^{-a/\kappa})$. From (59), we have for some $t = O(b^d)$, $ACR_t = O(d^{-a/\kappa})$. □

In the above theorem, two parameters a and b relies on problems and algorithms.

6.2. Polynomial ACR on Easy Problems

To exemplify a polynomial ACR on easy problems, consider the (1+1) adaptive RUS for minimizing the linear function.

$$\min f_L(\mathbf{x}) = \textstyle\sum_{i=1}^{d} c_i x_i, \quad \mathbf{x} = (x_1, \ldots, x_d) \in [0, 1]^d \subset \mathbb{R}^d. \tag{60}$$

f_L is a natural extension of linear functions from pseudo-Boolean function optimization [1] to continuous optimization. The OneMax function (8) is a special instance of linear functions. Since the OneMax function is easiest to the adaptive (1+1) EP, the ACR slowly decreases as d on the OneMax function. Let us make a rigorous analysis of this claim on linear functions.

Let $\mathbf{x}_t = \mathbf{x} = (x_1, \ldots, x_d)$ be a non-optimal solution. An offspring $\mathbf{y}$ is generated by search along a randomly selected dimension j with probability $1/d$. That is,

$$\mathbf{y}_j = (x_1, \ldots, x_{j-1}, y_j, x_{j+1}, \ldots, x_d),$$

where $y_j = x_j + \mathcal{N}(0, \sigma_j)$. Thanks to elitist selection, $\mathbf{y}$ is accepted if and only if it satisfies $y_j < x_j$. It is trivial to validate that the probability of hitting the promising region is

$$P_g(\mathbf{x}; \mathcal{S}(\mathbf{x}, 1)) = \sum_{j=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_j d} \int_0^{x_j} e^{-\frac{(y-x_j)^2}{2\sigma_j^2}} dy = \frac{1}{d} \sum_{j=1}^{d} \left(\frac{1}{2} - \Phi\left(-\frac{x_j}{\sigma_j}\right)\right). \tag{61}$$

Let σ_j such that $x_j/\sigma_j = C_0$ where $C_0 > 0$. In this case, we prove that the Gaussian mutation is positive landscape-adaptive, that is, $\exists C > 0$, $\rho \in (0, 1)$, $\forall \mathbf{x} \notin X^*$, $P_g(\mathbf{x}; \mathcal{S}(\mathbf{x}, \rho)) \geq C$.

Since $x_j/\sigma_j = C_0 > 0$, we have

$$P_g(\mathbf{x}; \mathcal{S}(\mathbf{x}, 1)) = \frac{1}{2} - \Phi(-C_0). \tag{62}$$

Take $P_g(\mathbf{x}, \mathcal{S}(\mathbf{x}, \rho))$ as a function of ρ defined in the interval $(0, 1]$. Obviously $P_g(\mathbf{x}, \mathcal{S}(\mathbf{x}, \rho))$ is continuous, that is, $\forall\, \varepsilon > 0$, $\exists \rho(\varepsilon) \in (0, 1)$ such that

$$P_g(\mathbf{x}, \mathcal{S}(\mathbf{x}, \rho(\varepsilon))) > P_g(\mathbf{x}, \mathcal{S}(\mathbf{x}, 1)) - \varepsilon. \tag{63}$$

Let $\rho = \rho(\varepsilon)$ and $C_\rho = \frac{1}{2} - \Phi(-C_0) - \varepsilon$ for any given $\frac{1}{2} - \Phi(-C_0) > \varepsilon > 0$. (63) and (42) confirm that the Gaussian mutation with $x_j/\sigma_j = C_0$ is positive landscape-adaptive. Then, Theorem 5 claims that the ACR is not less than a positive,

$$ACR_t \geq C = (1 - \rho(\varepsilon))(\frac{1}{2} - \Phi(-C_0) - \varepsilon). \tag{64}$$

However, the lower bound presented by (64) does not show how the ACR is connected to the dimension d. In the following, we demonstrate a relationship between the ACR and decision space dimension.

Given a fixed j, for the Gaussian mutation with $x_j/\sigma_j = C_0$, the error change is

$$\frac{1}{\sqrt{2\pi}\sigma_j}\int_0^{x_j} c_j(x_j - y)\exp\left\{-\frac{(y-x_j)^2}{2\sigma_j^2}\right\}dy = \frac{\sigma_j c_j}{\sqrt{2\pi}}\left(1 - e^{-\frac{x_j^2}{2\sigma_j^2}}\right) = \frac{\sigma_j c_j}{\sqrt{2\pi}}\left(1 - e^{-\frac{C_0^2}{2}}\right). \tag{65}$$

Since j is chosen at random from $\{1, \cdots, d\}$, the average error change over all $j = 1, \cdots, d$ is

$$\Delta e_t = \frac{1}{d}\sum_{j=1}^{d}\frac{\sigma_j c_j}{\sqrt{2\pi}}\left(1 - e^{-\frac{C_0^2}{2}}\right) = \frac{1}{\sqrt{2\pi}dC_0}\left(1 - e^{-\frac{C_0^2}{2}}\right)\sum_{j=1}^{d} c_j x_j.$$

Then we know that the rate of error change is

$$\frac{\Delta e_t}{e_t} = \frac{1}{d}\frac{1}{\sqrt{2\pi}C_0}\left(1 - e^{-\frac{C_0^2}{2}}\right) = \Theta(\frac{1}{d}).$$

According to Theorem 6, ACR_t is the reciprocal of a polynomial function of d for any t.

6.3. Exponential ACR on Hard Problems

To exemplify an exponential ACR on hard problems, consider the (1+1) adaptive EP for minimizing the deceptive function (9). Since the deceptive function is hard to the adaptive (1+1) EP, the ACR quickly decreases as d on the deceptive function. Let us make a rigorous analysis of this claim.

Let $\mathbf{x}_t = \mathbf{x}$ be a non-optimal solution. By Gaussian mutation, an offspring $\mathbf{y}$ is generated by $\mathbf{x}$ as $y_i = x_i + \mathcal{N}(0, \sigma_i)$ where $i = 1, \cdots, d$. Let $\sigma_i = x_i$ where $i = 1, \cdots, d$. We prove that this Gaussian mutation is positive landscape-adaptive.

For the deceptive function, let (1+1) EP start from the local optimum $(1, \cdots, 1)$. According to the definition of the deceptive function (9), $f_D(\mathbf{y}) < f_D(\mathbf{x})$ if and only if $\sum_{i=1}^{d} y_i \leq 1/2$. Thus, only an offspring $\mathbf{y}$ with $\sum_{i=1}^{d} y_i \leq 1/2$ can be accepted, and the transition probability to the promising region is

$$\begin{aligned} P_g(\mathbf{x}, \mathcal{S}(\mathbf{x}, 1)) &= \int_{\sum_i y_i \leq 1/2} p_{\mathbf{x}}(\mathbf{y})d\mathbf{y} \\ &= \int_{\sum_i y_i \leq 1/2} \frac{1}{(\sqrt{2\pi})^d}\exp\left\{-\frac{\sum_i (y_i - 1)^2}{2}\right\}dy_1 \ldots dy_d \\ &\leq \int_{\mathbf{y}\in[0,1/2]^d} \frac{1}{(\sqrt{2\pi})^d}\exp\left\{-\frac{\sum_i (y_i - 1)^2}{2}\right\}dy_1 \ldots dy_d \leq b^{-d}, \end{aligned} \tag{66}$$

for some $b > 1$.

The expected hitting time to the promising region is at least b^d. We choose a constant $c \in (0, 1)$ such that $1 < cb < b$, then for $t \leq (cb)^d$, the event of $\mathbf{x}_t$ leaving the local optimum happens with a probability at most $1 - (1 - b^{-d})^{(cb)^d} \leq c^d$. Thus $e_t = d - O(c^d)$.

For any $\mathbf{x}_t$ satisfying $\sum_{i=1}^{d} x_i \leq 1/2$, its average error change is at most $c^d/2$. For $\mathbf{x}_t = (1, \cdots, 1)$, its

average error change is

$$\begin{aligned}
&\int_{\sum_i y_i \le 1/2} \left((d+1-\sum_i x_i) - \sum_i y_i \right) p_{\mathbf{x}}(\mathbf{y}) d\mathbf{y} \\
&= \int_{\sum_i y_i \le 1/2} \left(1 - \sum_i y_i \right) \frac{1}{(\sqrt{2\pi})^d} \exp\left\{ -\frac{\sum_i (y_i-1)^2}{2} \right\} dy_1 \dots dy_d \\
&\le \int_{\mathbf{y} \in [0,1/2]^d} \left(1 - \sum_i y_i \right) \frac{1}{(\sqrt{2\pi})^d} \exp\left\{ -\frac{\sum_i (y_i-1)^2}{2} \right\} dy_1 \dots dy_d \\
&= \int_{\mathbf{z} \in [-1/2,0]^d} \left(1 - \sum_i (1+z_i) \right) \frac{1}{(\sqrt{2\pi})^d} \exp\left\{ -\frac{\sum_i z_i^2}{2} \right\} dz_1 \dots dz_d \\
&\le \int_{\mathbf{z} \in [-1/2,0]^d} \frac{1}{(\sqrt{2\pi})^d} \exp\left\{ -\frac{\sum_i z_i^2}{2} \right\} dz_1 \dots dz_d \\
&= \left(\Phi(0) - \Phi(-1/2) \right)^d . \qquad (67)
\end{aligned}$$

Then, by setting $a = 1/(\Phi(0) - \Phi(-1/2))$, we know that $a > 1$. Then the rate of error change is not more than the reciprocal of an exponential function of d, that is

$$\frac{\Delta e_t}{e_t} \le \frac{a^{-d} + O(c^d)}{d - O(c^d)}. \qquad (68)$$

According to Theorem 6, we conclude that ACR_t is not more than the reciprocal of an exponential function of d for $t \le (cb)^d$. This theoretical result confirms ACR_t decreases quickly as d on the deceptive function.

For the (1+1) adaptive RUS on the deceptive function, it is trivial to prove that starting from the local optimum $\mathbf{x}_0 = (1, \cdots, 1)$, the algorithm cannot generate a better child $\mathbf{y}$ such that $f_D(y) < f_D(1, \cdots, 1)$. Thus, for any t, we have $\mathbf{x}_t = \mathbf{x}_0$ and then $ACR_t = 0$.

7. Conclusions

This paper conducts a theoretical analysis of the ACR of EAs in continuous optimization. According to the limit property, the ACR is classified into two categories: (1) linear ACR whose limit inferior value is larger than a positive and (2) sublinear ACR whose value converges to zero. Then, it is proven that for EP using positive landscape-adaptive mutation, its ACR is linear. But for EP using landscape-invariant or zero landscape-adaptive mutation, its ACR is sublinear.

The relation between the ACR and the decision space dimension is also classified into two categories: (1) polynomial ACR whose value is larger than the reciprocal of a polynomial function of the dimension for any generation, and (2) exponential ACR whose value is less than the reciprocal of an exponential function of the dimension for an exponential long period. It is proven that for easy functions such as linear functions, the ACR of the (1+1) adaptive RUS is polynomial. But for hard functions such as the deceptive function, the ACR of both the (1+1) adaptive EP and RUS is exponential.

This paper does not discuss EAs whose genetic operators change over time. This topic will be left for future research.

References

[1] J. He, X. Yao, Drift analysis and average time complexity of evolutionary algorithms, Artificial Intelligence 127 (1) (2001) 57–85.

[2] Y. Chen, X. Zou, J. He, Drift conditions for estimating the first hitting times of evolutionary algorithm, International Journal of Computer Mathematics 88 (1) (2011) 37–50.

[3] A. Agapie, M. Agapie, G. Rudolph, G. Zbaganu, Convergence of evolutionary algorithms on the n-dimensional continuous space., IEEE Transactions on Cybernetics 43 (5) (2013) 1462–1472.

[4] H. Huang, J. Su, Y. Zhang, Z. Hao, An experimental method to estimate running time of evolutionary algorithms for continuous optimization, IEEE Transactions on Evolutionary Computation 24 (2) (2020) 275–289.

[5] G. Rudolph, Local convergence rates of simple evolutionary algorithms with Cauchy mutations, IEEE Transactions on Evolutionary Computation 1 (4) (1997) 249–258.

[6] G. Rudolph, et al., Convergence rates of evolutionary algorithms for a class of convex objective functions, Control and Cybernetics 26 (1997) 375–390.

[7] J. He, L. Kang, On the convergence rate of genetic algorithms, Theoretical Computer Science 229 (1-2) (1999) 23–39.

[8] J. He, X. Yu, Conditions for the convergence of evolutionary algorithms, Journal of Systems Architecture 47 (7) (2001) 601–612.

[9] A. Auger, Convergence results for the (1, λ)-sa-es using the theory of ϕ-irreducible markov chains, Theoretical Computer Science 334 (1-3) (2005) 35–69.

[10] A. Auger, N. Hansen, Linear convergence of comparison-based step-size adaptive randomized search via stability of markov chains, SIAM Journal on Optimization 26 (3) (2016) 1589–1624.

[11] J. He, Y. Zhou, G. Lin, An initial error analysis for evolutionary algorithms, in: Proceedings of the Genetic and Evolutionary Computation Conference Companion, ACM, 2017, pp. 317–318.

[12] D. Tarłowski, On geometric convergence rate of markov search towards the fat target, Operations Research Letters 46 (1) (2018) 33–36.

[13] J. He, G. Lin, Average convergence rate of evolutionary algorithms, IEEE Transactions on Evolutionary Computation 20 (2) (2016) 316–321.

[14] H. Dong, T. Li, R. Ding, J. Sun, A novel hybrid genetic algorithm with granular information for feature selection and optimization, Applied Soft Computing 65 (2018) 33–46.

[15] T. Dhivyaprabha, P. Subashini, M. Krishnaveni, Synergistic fibroblast optimization: a novel nature-inspired computing algorithm, Frontiers of Information Technology & Electronic Engineering 19 (7) (2018) 815–833.

[16] D. Janiga, R. Czarnota, J. Stopa, P. Wojnarowski, Self-adapt reservoir clusterization method to enhance robustness of well placement optimization, Journal of Petroleum Science and Engineering 173 (2019) 37–52.

[17] T. Li, H. Dong, J. Sun, Binary differential evolution based on individual entropy for feature subset optimization, IEEE access 7 (2019) 24109–24121.

[18] L. Ding, L. Kang, Convergence rates for a class of evolutionary algorithms with elitist strategy, Acta Mathematica Scientia 21 (4) (2001) 531–540.

[19] M. A. Semenov, D. A. Terkel, Analysis of convergence of an evolutionary algorithm with self-adaptation using a stochastic lyapunov function, Evolutionary Computation 11 (4) (2003) 363–379.

[20] H.-G. Beyer, The theory of evolution strategies, Springer Science & Business Media, 2013.

[21] H.-G. Beyer, A. Melkozerov, The dynamics of self-adaptive multirecombinant evolution strategies on the general ellipsoid model, IEEE Transactions on Evolutionary Computation 18 (5) (2014) 764–778.

[22] H.-G. Beyer, M. Hellwig, The dynamics of cumulative step size adaptation on the ellipsoid model, Evolutionary computation 24 (1) (2016) 25–57.

[23] Y. Akimoto, A. Auger, N. Hansen, Quality gain analysis of the weighted recombination evolution strategy on general convex quadratic functions, in: Proceedings of the 14th ACM/SIGEVO Conference on Foundations of Genetic Algorithms, ACM, 2017, pp. 111–126.

[24] A. Auger, N. Hansen, Reconsidering the progress rate theory for evolution strategies in finite dimensions, in: Proceedings of the 8th annual conference on Genetic and evolutionary computation, ACM, 2006, pp. 445–452.

[25] A. Auger, N. Hansen, Theory of evolution strategies: a new perspective, in: Theory of Randomized Search Heuristics: Foundations and Recent Developments, World Scientific, 2011, pp. 289–325.

[26] M. Jebalia, A. Auger, N. Hansen, Log-linear convergence and divergence of the scale-invariant (1+ 1)-es in noisy environments, Algorithmica 59 (3) (2011) 425–460.

[27] R. Varga, Matrix Iterative Analysis, Springer, 2009.

[28] B.-Y. Qu, J. Liang, Z. Wang, Q. Chen, P. N. Suganthan, Novel benchmark functions for continuous multimodal optimization with comparative results, Swarm and Evolutionary Computation 26 (2016) 23–34.

[29] T. Bäck, G. Rudolph, H.-P. Schwefel, Evolutionary programming and evolution strategies: Similarities and differences, in: D. B. Fogel, W. Atmar (Eds.), Proceedings of the Second Annual Conference on Evolutionary Programming, Evolutionary Programming Society, San Diego, CA, 1993, pp. 11–22.

[30] J. He, T. Chen, X. Yao, On the easiest and hardest fitness functions, IEEE Transactions on Evolutionary Computation 19 (2) (2015) 295–305.

[31] K. Chellapilla, Combining mutation operators in evolutionary programming, IEEE transactions on Evolutionary Computation 2 (3) (1998) 91–96.

[32] X. Yao, Y. Liu, G. Lin, Evolutionary programming made faster, IEEE Transactions on Evolutionary computation 3 (2) (1999) 82–102.

[33] H. Dong, J. He, H. Huang, W. Hou, Evolutionary programming using a mixed mutation strategy, Information Sciences 177 (1) (2007) 312–327.

[34] R. Mallipeddi, S. Mallipeddi, P. N. Suganthan, Ensemble strategies with adaptive evolutionary programming, Information Sciences 180 (9) (2010) 1571–1581.

[35] T. Jansen, C. Zarges, Performance analysis of randomised search heuristics operating with a fixed budget, Theoretical Computer Science 545 (2014) 39–58.

[36] J. Snow, K. Weller, Exploratory Examples for Real Analysis, Mathematical Association of America, 2003.
[37] O. Teytaud, S. Gelly, General lower bounds for evolutionary algorithms, in: Parallel Problem Solving from Nature-PPSN IX, Springer, 2006, pp. 21–31.
[38] H.-G. Beyer, Convergence analysis of evolutionary algorithms that are based on the paradigm of information geometry, Evolutionary Computation 22 (4) (2014) 679–709.
[39] S. Meyn, R. Tweedie, Markov chains and stochastic stability, Springer Verlag, 1993.
[40] D. Tarłowski, On the convergence rate issues of general markov search for global minimum, Journal of Global Optimization 69 (4) (2017) 869–888.