Average Convergence Rate of Evolutionary Algorithms II: Continuous Optimization

Yu Chen, Jun He

Abstract—A good convergence metric must satisfy two requirements: feasible in calculation and rigorous in analysis. The average convergence rate is proposed as a new measurement for evaluating the convergence speed of evolutionary algorithms over consecutive generations. Its calculation is simple in practice and it is applicable to both continuous and discrete optimization. Previously a theoretical study of the average convergence rate was conducted for discrete optimization. This paper makes a further analysis for continuous optimization. First, the strategies of generating new solutions are classified into two categories: landscape-invariant and landscape-adaptive. Then, it is proven that the average convergence rate of evolutionary algorithms using landscape-invariant generators converges to zero, while the rate of algorithms using positive-adaptive generators has a positive limit. Finally, two case studies, the minimization problems of the two-dimensional sphere function and Rastrigin function, are made for demonstrating the applicability of the theory.

Index Terms—evolutionary algorithms, continuous optimization, convergence rate, Markov chains, approximation error, genetic operators

I. INTRODUCTION

IN the theoretical study of EAs, a fundamental question is how fast can an EA find an optimal solution to a problem? In discrete optimization, this can be measured by the number of generations (hitting time) or the number of fitness evaluations (running time) when an EA finds an optimal solution [1], [2]. However, computation time is seldom applied to continuous optimization. Unlike discrete optimization, computation time is normally infinite in continuous optimization because the optimal solution set of a continuous optimization problem is usually a zero-measure set. In order to apply computation time into continuous optimization, the optimal solution must be replaced by a ε-neighbour of the optimal solution set [3], [4], [5] which forms a positive-measure set.

In continuous optimization, the performance of EAs is often evaluated by the convergence rate. Informally, the convergence rate question is how fast $\| X_t - X^* \|$ converges to 0? where $\| X_t - X^* \|$ is a distance between the tth generation population X_t and the optimal solution(s) X^*. A lot of theoretical work discussed this topic from different perspectives [6], [7], [8], [9], [10], [11], however convergence metrics studied in theory are seldom adopted in practice. This motivates us to design a practical convergence metric satisfying two requirements: feasible in calculation and rigours in theory.

Our work emphasizes the convergence rate in terms of the approximation error. The approximation error is to evaluate the solution quality of EAs. Let $f(X_t)$ denote the fitness of the best individual in population X_t, its expected value $f_t = \mathbb{E}[f(X_t)]$, and f^* the fitness of the optimal solution. The approximate error [12] is $e_t = |f_t - f^*|$. In the context of e_t, the convergence rate question is how fast e_t converges to 0? It is straightforward to derive the geometric convergence $e_t \leq e_0 c^t$ from the condition $e_t/e_{t-1} \leq c < 1$ [6].

An alternative convergence metric is the error ratio between two generations (or one-generation convergence rate): e_t/e_{t-1}. This ratio works well in deterministic iterative algorithms. But unfortunately, it is not appropriate to EAs because the calculation of e_t/e_{t-1} is numerically unstable.

A remedy to the deficiency of the two-generation error ratio e_t/e_{t-1} is to consider its average over consecutive t generations. Then the geometric average convergence rate (ACR) is proposed by He and Lin [13], which is

$$R_t = 1 - \left(\frac{e_t}{e_0} \right)^{1/t}. \tag{1}$$

From the ACR, it is straightforward to draw an exact expression of the approximation error: $e_t = (1 - R_t)^t e_0$. More importantly, the calculation of R_t is more stable than e_t/e_{t-1} in computer simulation.

For discrete optimization, it has been proven [13] under random initialization, R_t converges to a positive; and under particular initialization, R_t always equals to this positive.

The current paper extends the analysis of the ACR from discrete optimization to continuous optimization. However, the extension is not trivial due to completely different probability measures in discrete and continuous spaces. There are two essential changes in the extension.

The analyses are different. In continuous optimization, an EA is modeled by a Markov chain in a continuous state space, rather than a Markov chain in a finite state space. Thus the matrix analysis used in [13] cannot be applied to continuous optimization.

The results are different. For continuous optimization, Theorem 1 in this paper claims that given a convergent EA modelled by an homogeneous Markov chain, its ACR converges to 0 if its generator is invariant or converges to a positive if its generator is positive-adaptive. But for discrete optimization, Theorem 1 in [13] states that for all convergent EAs modelled by homogeneous Markov chains, their ACR converges to a positive.

The paper is organized as follows: Section II introduces the related work. Section III defines the ACR. Section IV provides a general analysis of the ACR. Section V provides

two case studies on the sphere function and Rastrigin function. Section VI concludes the paper.

II. RELATED WORK

The convergence rate of EAs has been investigated from different perspectives and in varied terms.

Rudolph [6] 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 is obtained.

Rudolph [7] compared Gaussian and Cauchy mutation on minimizing 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 the rate is identical for Gaussian and spherical Cauchy distributions, whereas nonspherical Cauchy mutations lead to slower local convergence.

Beyer [14] 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. Recently Beyer et al. [15], [16] 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.[17] investigated evolution strategies with weighted recombination on general convex quadratic functions and derived the asymptotic quality gain. However, Auger and Hansen [18] argued the limits of the predictions based on the progress rate.

Auger and Hansen [19] developed the theory of ES from a new perspective using stability of Markov chains. Auger [10] 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. [20] 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 [11] 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 CR. This log-linear convergence is an extension of the average rate of convergence in deterministic iterative methods [21].

He, Kang and Ding [8], [22] studied the convergence in distribution $\| \mu_t - \pi \|$ where μ_t is the probability distribution of X_t and π a stationary probability distribution. Based on the Doeblin condition, they obtained bounds on $\| \mu_t - \pi \| \leq (1 - \delta)^{t-1}$ for some $\delta \in (0, 1)$. He and Yu [9] also derived lower and upper bounds on $1 - \mu_t(X_\delta^*)$ where $\mu_t(X_\delta^*)$ denotes the probability of X_t entering in a δ-neighbour of X^*.

This paper develops Rudolph's early work [6] which showed the geometrical convergence of e_t but didn't provide a method to quantify the convergence rate. We take $1 - (e_t/e_{t-1})^{1/t}$ as a practical metric to measure the geometric convergence and make a rigorous analysis.

III. DEFINITIONS AND PRACTICAL USAGE

A. Definitions

A continuous minimization problem is to

$$\min f(\vec{x}), \quad \vec{x} = (x_1, \cdots, x_d) \in \mathcal{D} \subset \mathbb{R}^d, \quad (2)$$

where $f(\vec{x})$ is a continuous function $\mathbb{R}^d \to \mathbb{R}$ defined on a closed set $\mathcal{D}$. Denote $f^* = \min f(\vec{x})$. We assume the optimal solution set to the above problem is a finite set.

An individual $\vec{x}$ is a vector in $\mathbb{R}^d$ and a population $X = (\vec{x}_1, \cdots, \vec{x}_N)$ is a vector in $\mathbb{R}^{d \times N}$. A general framework of elitist EAs for solving optimization problems is described in Algorithm 1. Two types of genetic operators are employed in the algorithm. One is the generation operator to generate new individuals from a population such as mutation or crossover. The other is the selection operator to select individuals from a population. Any non-elitist EA can be modified into an equivalent elitist EA through adding an archive individual which preserves the best found solution but does not get involved in evolution. Thereafter we only consider elitist EAs.

Algorithm 1 A framework of elitist EAs

1: counter $t \leftarrow 0$;
2: population $X_0 \leftarrow$ initialize N individuals in the definition domain $\mathcal{D}$ at random;
3: **while** the stopping criterion is not satisfied **do**
4:    population $Y_t \leftarrow$ generate N individuals from X_t;
5:    population $X_{t+1} \leftarrow$ select N individuals from $X_t \cup Y_t$ while at least one of the best individual(s) is selected (called *elitist*) and any individual out of the domain $\mathcal{D}$ is rejected;
6:    counter $t \leftarrow t + 1$;
7: **end while**

Since population Y_t in Algorithm 1 only depends on X_t and then X_{t+1} only depends on X_t, the population sequence $\{X_t; t = 0, 1, \cdots\}$ is a Markov chain [8], [9].

Definition 1: The *fitness* of population X is $f(X) = \min\{f(\vec{x}) \mid \vec{x} \in X\}$ and the *approximation error* of X is $e(X) = |f(X) - f^*|$. The sequence $\{e(X_0), e(X_1), \cdots\}$ is called *convergent in mean* if $\lim_{t \to +\infty} \mathbb{E}[e(X_t)] = 0$ and *convergent almost sure* if $\Pr(\lim_{t \to +\infty} e(X_t) = 0) = 1$.

Thanks to elitist selection, $e(X_t) \leq e(X_{t-1})$. Then the sequence $\{e(X_t); t = 0, 1, \cdots\}$ is a supermartingale. According to Doob's convergence theorem [23], for elitist EAs, convergence in mean implies almost sure convergence [6].

Lemma 1: For elitist EAs, if the sequence $\{e(X_t); t = 0, 1, \cdots\}$ converges in mean, then it converges almost sure.

The ACR is to evaluate the average convergence speed of EAs for consecutive t generations [13]. The following definition is applicable to both elitist and non-elitist EAs.

Definition 2: Let $f_t = \mathbb{E}[f(X_t)]$ and $e_t = \mathbb{E}[e(X_t)]$. The geometric average convergence rate (ACR) of an EA for t generations is

$$R_t = 1 - \left(\frac{e_t}{e_0}\right)^{1/t} = 1 - \left(\prod_{k=1}^{t} \frac{e_k}{e_{k-1}}\right)^{1/t}. \quad (3)$$

If $e_k = 0$ for some k, let $R_t = 1$ for any $t \geq k$.

In (3), the term $(e_t/e_0)^{1/t}$ represents a geometric average of the reduction factor over t generations. $1 - (e_t/e_0)^{1/t}$ normalizes the average in the interval $(-\infty, 1]$. The ACR can be regarded as the speed of convergence while the error as the distance from the optimal set. If $R_t > 0$, then the speed is positive and $e_t < e_0$; if $R_t = 0$, then the speed is zero and $e_t = e_0$; if $R_t < 0$ (never happens in elitist EAs), then the speed is negative and $e_t > e_0$. Like the speed of light, the speed of convergence has an upper limit, that is, $R_t \leq 1$.

B. Practical Usage of Average Convergence Rate

The ACR provides a simple method to numerically measure how fast an EA converges. This is the main purpose of the ACR. In practice, the expected value f_t is replaced by a sample mean of f_t^T over T runs of the EA. The ACR R_t is calculated in four steps [13]:

1) run an EA for T times;
2) calculate the fitness sample mean f_t^T:

$$f_t^T = \frac{1}{T}\left(f(X_t^{[1]}) + \cdots + f(X_t^{[T]})\right), \quad (4)$$

where $f(X_t^{[k]})$ denotes the fitness $f(X_t)$ at the k-th run;
3) calculate the approximate error: $e_t^T = |\, f_t^T - f^* \,|$;
4) finally, calculate the ACR: $R_t^T = 1 - (e_t^T/e_0^T)^{1/t}$.

According to the Law of Large Numbers, it holds $e_t^T \to e_t$ and $R_t^T \to R_t$ as $T \to +\infty$.

An example is given to show the usage of the ACR in computer simulation. The aim is a comparison of two EAs on two benchmark functions in terms of the ACR. The benchmarks are the 2-dimensional sphere and Rastrigin functions:

$$\min f_S(\vec{x}) = x_1^2 + x_2^2, \quad \vec{x} \in \mathbb{R}^2, \quad (5)$$

$$\min f_R(\vec{x}) = 20 + \sum_{k=1}^{2}(x_k^2 - 10\cos 2\pi x_k), \quad (6)$$

The minimal point to both functions is $\vec{x}^* = (0,0)$ with $f(\vec{x}^*) = 0$. Two EAs are variants of (1+1) elitist EAs (Algorithm 2) which adopt Gaussian mutation:

$$\vec{y} = \vec{x} + \vec{z}, \quad (7)$$

where $\vec{x}$ is the parent, $\vec{y}$ the child and $\vec{z} = (z_1, \cdots, z_d)$ a Gaussian random vector obeying the probability distribution

$$z_i \sim \mathcal{N}(0, \sigma_i). \quad (8)$$

There are two ways to set the variance $\vec{\sigma} = (\sigma_1, \cdots, \sigma_N)$.

- **Invariant-σ:** $\vec{\sigma}$ is set to a constant for all $\vec{x}$. In computer simulation, set $\sigma_i = 1$.
- **Adaptive-σ:** $\vec{\sigma}$ takes varied values on different $\vec{x}$. In computer simulation, set $\sigma_i = \|\, \vec{x} \,\|_2$.

Algorithm 2 A framework of (1+1) elitist EAs

1: counter $t \leftarrow 0$;
2: individual $\vec{x}_0 \leftarrow$ initialize a solution;
3: **while** the stopping criterion is not satisfied **do**
4:     population $\vec{y}_t \leftarrow$ generate a new solution from $\vec{x}_t$ by Gaussian mutation;
5:     individual $\vec{x}_{t+1} \leftarrow$ select the best of individuals in Y_t and $\vec{x}_t$;
6:     $t \leftarrow t + 1$;
7: **end while**

For the sake of terms, the EA using invariant-σ mutation is called an *invariant EA* and the EA using adaptive-σ mutation is called an *adaptive EA*.

In the experiment, the initial solution $\vec{x}_0 = (10, 10)$. The times of running an EA is 100. The maximum number of generations is 500.

The ACR R_t quantifies the speed of convergence. Table I shows that the ACR value of the adaptive EA is much larger than that of the invariant EA on both f_R and f_S.

TABLE I: R_t: adaptive vs invariant EAs on f_S and f_R.

	generation	1	101	201	301	401
f_S	adaptive	0.28	0.42	0.45	0.46	0.51
	invariant	0.10	0.11	0.07	0.05	0.04
f_R	adaptive	0.23	0.20	1.00	1.00	1.00
	invariant	0.04	0.05	0.04	0.03	0.02

Fig. 1 illustrates the trend of R_t. The ACR of the adaptive EA tends to stabilize at some positive value, while the ACR of the invariant EA is in a decreasing tendency. This phenomenon will be strictly analyzed later.
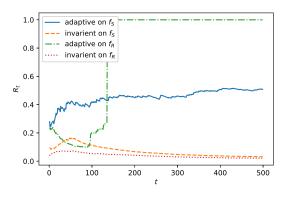

Fig. 1: R_t: adaptive vs invariant EAs on f_S and f_R.

C. Discussion of Other Convergence Metrics

A good convergence metric should satisfy two requirements: feasible in calculation and rigorous in analysis. We discuss two common convergence metrics and show they don't satisfy the requirements.

The ratio e_t/e_{t-1} is a popular convergence metric used in deterministic iterative algorithms which quantifies the reduction ratio of e_t for one iteration. Fig. 2 illustrates the e_t/e_{t-1}

value for the adaptive EA on f_S. e_t/e_{t-1} fluctuates greatly. The calculation of e_t/e_{t-1} is sensitive and unstable due to $e_t, e_{t-1} \approx 0$. Therefore, it is not a practical metric to measure the convergence rate of EAs.
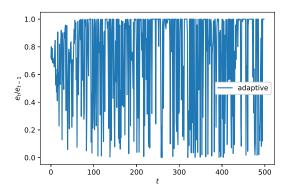

Fig. 2: e_t/e_{t-1}: the adaptive EA on f_S.

The logarithmic scale, $\log e_t$, probably is the most widely used convergence metric in comparing the convergence speed of EAs in practice. Fig. 3 displays the $\log e_t$ value of adaptive and invariant EAs on f_S. When using $\log e_t$ for comparing the speed of convergence of two EAs, it is necessary to visualize $\log e_t$ in a figure and compare the slop of $\log e_t$ via observation. Fig. 3 shows that the slop of $\log e_t$ of the adaptive EA is sharper than the invariant EA. However, an observation is an observation, not an analysis. The slop $\log e_t - \log e_{t-1}$ might be taken as a convergence metric. But like e_t/e_{t-1}, the calculation of $\log(e_t/e_{t-1})$ is sensitive and unstable in computer simulation.
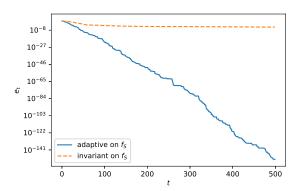

Fig. 3: e_t: adaptive vs invariant EAs on f_S.

Summarizing the above discussion, we conclude that both $\log e_t$ and e_t/e_{t-1} are not appropriate as a convergence metric.

IV. GENERAL ANALYSES

A. Transition Probabilities

An EA is determined by its operators: generator and selection. In mathematics, both can be represented by transition probabilities.

Let $\mathcal{S} = \mathcal{D}^N$ denote the set consisting of all populations. A population is represented by a capital letter such as $X = (\vec{x}_1, \cdots, \vec{x}_N)$. The tth generation population is represented by X_t which is a random vector. A population X satisfying $f(X) = f^*$ is called an optimal population, and the collection of all optimal populations is denoted as X^*.

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

$$\mathcal{S}(X, \rho) = \{Y \in \mathcal{S} | e(Y) < \rho e(X)\}, \tag{9}$$
$$\overline{\mathcal{S}}(X, \rho) = \{Y \in \mathcal{S} | e(Y) \geq \rho e(X)\}. \tag{10}$$

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

The generation of Y_t via X_t is denoted as $X_t \Rightarrow Y_t$. It can be characterized by a probability transition. Given a population $X \in \mathcal{S}$ and a population set $\mathcal{A} \subset \mathcal{S}$, the *transition probability kernel* $P_g(X; \mathcal{A})$ is defined as

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

where $p_g(X; Y)$ is a *transition probability density function* [24].

Similarly, the selection operation, $(X_t, Y_t) \Rightarrow X_{t+1}$, can be described by a probability transition too. Given any population $X, Y \in \mathcal{S}$ and a population set $\mathcal{A} \subset \mathcal{S}$, its transition probability kernel $P_s(X; Y; \mathcal{A})$ is defined as

$$P_s(X, Y; \mathcal{A}) = \int_{\mathcal{A}} p_s(X, Y; Z) dZ,$$

where $p_s(X, Y; Z)$ is a transition probability density function.

A one-generation update of population, $X_t \Rightarrow X_{t+1}$, is described by a probability transition. Given any population $X \in \mathcal{S}$ and a population set $\mathcal{A} \subset \mathcal{S}$, its transition probability kernel $P(X; \mathcal{A})$ is defined as

$$P(X; \mathcal{A}) = \int_{\mathcal{A}} p(X; Y) dY,$$

where $p(X; Y)$ is a transition probability density function.

Generally, the operators of generating new individuals may be classified into two categories.

Definition 3: Let $p_g(X; Y)$ be the probability function depicting the generation transition from X to Y.

1) **Landscape-invariant:** a generator $X \Rightarrow Y$ is called *landscape-invariant* if $Y = X + Z$ and Z is a multivariate random variable whose joint probability distribution is independent on X. Here $X + Z$ represents $(\vec{x}_1 + \vec{z}_1, \cdots, \vec{x}_N + \vec{z}_N)$.
   We assume the density function $p_z(Z)$ is continuous and bounded, such as Cauchy and Gaussian distributions.
2) **Landscape-adaptive:** otherwise, a generator $X \Rightarrow Y$ is called *landscape-adaptive*.

A landscape-invariant generator generates candidate solutions subject to the same probability distribution no matter where a parent population locates. An example is the invariant-σ Gaussian mutation described in Algorithm 2. A landscape-adaptive generator adjusts the probability distribution according to the position of a parent population. An example is the adaptive-σ Gaussian mutation in Algorithm 2.

For the landscape-invariant generator, the lemma below states that the infinum of the transition probability to the promising region equals to zero.

Lemma 2: If the number of optimal solutions is finite and the generator is landscape-invariant, then the transition probability to the promising region satisfies

$$\inf\{P_g(X, S(X, 1)); X \notin X^*\} = 0, \quad (11)$$

where inf is the abbreviation of mathematical infimum.

Proof: In order to prove (11), it is sufficient to prove

$$\lim_{e(X) \to 0} P_g(X, S(X, 1)) = 0. \quad (12)$$

That is, $\forall \varepsilon > 0, \exists \delta > 0, \forall X \in \mathcal{A}(X^*, \delta) \setminus X^*$ (where the set $\mathcal{A}(X^*, \delta) = \{X; e(X) \leq \delta\}$), it holds

$$P_g(X, \mathcal{S}(X, 1)) < \varepsilon. \quad (13)$$

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

$$\Pr(X + Z \in \mathcal{A}) = \int_{Z : X + Z \in \mathcal{A}} p_z(X + Z) dZ < \varepsilon. \quad (14)$$

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

Because f is continuous, we may choose δ sufficiently small so that $\forall X \in \mathcal{A}(X^*, \delta)$ and $Y \notin \mathcal{A}(X^*, \delta) : f(X) < f(Y)$. This implies the promising region $\mathcal{S}(X, 1) \subset \mathcal{A}(X^*, \delta)$.

According to (14) and $m(\mathcal{A}(X^*, \delta)) \leq \delta', \forall X \in \mathcal{A}(X^*, \delta) \setminus X^*$, we have

$$\Pr(X + Z \in \mathcal{A}(X^*, \delta)) < \varepsilon. \quad (15)$$

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

$$P_g(X, S(X, 1)) \leq \Pr(X + Z \in \mathcal{A}(X^*, \delta)) < \varepsilon. \quad (16)$$

The above inequality is our wanted result. ∎

B. Analysis of Landscape-invariant Generators

For elitist EAs using landscape-invariant generators, Theorem 1 below indicates that the limit of the ACR R_t is 0.

Theorem 1: For Problem (2) and Algorithm 1, if the following conditions are true:

1) the number of optimal solutions is finite;
2) the sequence $\{e_t; t = 0, 1, \cdots\}$ converges to 0;
3) the generator is landscape-invariant;

then $\lim_{t \to +\infty} R_t = 0$.

Proof: In order to prove $\lim_{t \to +\infty} R_t = 0$, it is sufficient to prove that $\lim_{t \to +\infty} e_t / e_{t-1} = 1$, equivalently, $\lim_{t \to +\infty} (e_{t-1} - e_t) / e_{t-1} = 0$. According to the definition of limit, it is sufficient to prove that $\forall \varepsilon > 0, \exists t_0 > 0, \forall t \geq t_0$,

$$e_{t-1} - e_t < \varepsilon e_{t-1}. \quad (17)$$

From (13) in Lemma 2, we know $\forall \varepsilon > 0, \exists \delta > 0$, let $\mathcal{A}(X^*, \delta) = \{X; e(X) \leq \delta\}$, then $\forall X \in \mathcal{A}(X^*, \delta) \setminus X^*$, it holds

$$P_g(X, \mathcal{S}(X, 1)) < \varepsilon. \quad (18)$$

From Lemma 1, the sequence $\{e(X_t); t = 0, 1, \cdots\}$ converges almost surely to 0, that is, $\Pr(\lim_{t \to +\infty} e(X_t) = 0) = 1$. Denote

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

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

For the set $\mathcal{S}_2$, it holds

$$\Pr(\omega \in \mathcal{S}_2) = 0, \quad (19)$$

and for the set $\mathcal{S}_1$, we know that for the given $\delta > 0, \exists t_0 > 0$, then $\forall t > t_0$, it holds

$$e(X_{t-1}(\omega)) < \delta, \quad \forall \omega \in \mathcal{S}_1.$$

From (18) we know

$$P_g(X, \mathcal{S}(X_{t-1}(\omega), 1)) \leq \varepsilon, \quad \omega \in \mathcal{S}_1.$$

Then we obtain $\forall \omega \in \mathcal{S}_1$,

$$\mathbb{E}[e(X_{t-1}(\omega)) - e(X_t(\omega)) \mid X_{t-1}(\omega)] \leq \varepsilon e(X_{t-1}(\omega)). \quad (20)$$

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

$$\mathbb{E}[e(X_{t-1}(\omega)) - e(X_t(\omega)) \mid X_{t-1}(\omega)] \leq B. \quad (21)$$

Combining (19), (20) and (21) together, we get

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

So (17) is true. Then we complete the proof. ∎

Theorem 1 states that for EAs using landscape-invariant generators, the limit of their ACR is 0 as $t \to +\infty$. This implies that landscape-invariant generators are not appropriate for solving continuous optimization problems.

Theorem 1 may not hold if the Lebesgue measure of X^* is positive. However, for most continuous optimization problems, X^* is a zero-measure set.

C. Analysis of Landscape-adaptive Generators

Landscape-adaptive generators can be split into two types:

1) **positive-adaptive:** a landscape-adaptive generator $X \Rightarrow Y$ is called *positive-adaptive* if $\exists \rho \in (0, 1)$, the transition probability to the ρ-promising region satisfies

$$C_\rho = \inf\{P_g(X; \mathcal{S}(X, \rho)); X \notin X^*\} > 0. \quad (22)$$

2) **zero-adaptive:** a landscape-adaptive generator $X \Rightarrow Y$ is called *zero-adaptive* if the transition probability to the promising region satisfies

$$\inf\{P_g(X; \mathcal{S}(X, 1)); X \notin X^*\} = 0. \quad (23)$$

The zero-adaptive generator is bad adaptation because it causes a zero-valued ACR. (23) includes two cases:

1) $\lim_{e(X) \to 0} P_g(X, S(X, 1)) = 0$. The analysis of this case is similar to Theorem 1. Then $\lim_{t \to +\infty} R_t = 0$.
2) $\exists X \notin X^*$ such that $P_g(X; S(X, 1)) = 0$. When an EA starts from X, $f_{t+1} = f_t$ for all t and then $R_t = 0$.

However, a positive-adaptive generator is always good adaptation because it ensures that the limit of the ACR is positive.

Theorem 2: For Problem (2) and Algorithm 1, if the following conditions are true:

1) the sequence $\{e_t; t = 0, 1, \cdots\}$ converges to 0;
2) the generation operator is positive-adaptive with a contraction factor $\rho \in (0, 1)$;

then $\exists C > 0$ such that $\lim_{t \to +\infty} R_t \geq C$.

Proof: From (9), we know that for any $k - 1 \geq 0$,

$$S(X_{k-1}, \rho) = \{Y \in S \mid e(Y) \leq \rho e(X_{k-1})\}.$$

It follows that $S(X_{k-1}, \rho) \subset S(X_{k-1}, 1)$, and for any $Y \in S(X_{k-1}, \rho)$,

$$f(X_{k-1}) - f(Y) \geq (1 - \rho)(f(X_{k-1}) - f^*). \quad (24)$$

So we get

$$\mathbb{E}\left[f(X_{k-1}) - f(X_k)|X_{k-1}\right]$$
$$= \int_{S(X_{k-1}, 1)} (f(X_{k-1}) - f(Y))p_g(X_{k-1}; Y)dY$$
$$\geq \int_{S(X_{k-1}, \rho)} (f(X_{k-1}) - f(Y))p_g(X_{k-1}; Y)dY$$
$$\geq \int_{S(X_{k-1}, \rho)} (1-\rho)(f(X_{k-1})-f^*)p_g(X_{k-1};Y)dY \quad \text{(from (24))}$$
$$= (1-\rho)(f(X_{k-1})-f^*)P_g(X_{k-1}, S(X_{k-1}, \rho))$$
$$\geq (1 - \rho)C_\rho\left(f(X_{k-1}) - f^*\right). \quad \text{(from (22))} \quad (25)$$

Then

$$\frac{e_k}{e_{k-1}} = 1 - \frac{f_{k-1} - f_k}{f_{k-1} - f^*}$$
$$= 1 - \frac{\mathbb{E}[\mathbb{E}\left[f(X_{k-1}) - f(X_k)|X_{k-1}\right]]}{f_{k-1} - f^*}$$
$$\leq 1 - \frac{\mathbb{E}\left[(1-\rho)C_\rho\left(f(X_{k-1}) - f^*\right)\right]}{f_{k-1} - f^*}$$
$$\leq 1 - (1 - \rho)C_\rho.$$

Then,

$$R_t = 1 - \left(\frac{e_t}{e_0}\right)^{1/t} = 1 - \left(\prod_{k=1}^t \frac{e_k}{e_{k-1}}\right)^{1/t} \geq (1 - \rho)C_\rho.$$

Let $C = (1 - \rho)C_\rho$. It holds that $\lim_{t \to +\infty} R_t \geq C$. ∎

Theorem 2 indicates if an EA employs a positive-adaptive generator, then it converges to the optimal set with a positive ACR. How to design a generator satisfying the positive-adaptive condition (22) is important. An example is Rechenberg's 1/5th success rule for controlling the mutation strength used in evolutionary strategies [25]. From a theoretical viewpoint, Theorems 1 and 2 together confirm the necessity of using adaptive generators in continuous optimization.

D. Analysis of Elitist EAs Not Convergent in Mean to 0

The analysis of this kind of EAs is rather simple. The theorem below states that the limit of the ACR is 0.

Theorem 3: If the sequence $\{e_t; t = 0, 1, \cdots\}$ does not converge to 0, then $\lim_{t \to +\infty} R_t = 0$.

Proof: Due to elitist selection, the sequence $\{e_t; t = 0, 1, \cdots\}$ is monotonic decreasing with $e_t \geq 0$. According to the monotone convergence theorem, $e_t \to e^\sharp$. The condition says $e^\sharp \neq 0$, thus $e^\sharp > 0$. Then $R_t \to 1 - e^\sharp/e^\sharp = 0$. ∎

V. CASE STUDIES

A. 2-D Sphere function

Consider minimization of the 2-dimensional (2-D) sphere function.

$$\min \quad f_S(\vec{x}) = x_1^2 + x_2^2, \quad \vec{x} = (x_1, x_2) \in \mathbb{R}^2. \quad (26)$$

The optimal solution is $\vec{x}^* = (0, 0)$ with $f_S(\vec{x}^*) = 0$.

The (1+1) elitist EA (Algorithm 2) is used to solve this problem. Let $\vec{x} = (x_1, x_2)$ be the individual at the t-th generation and $\vec{y} = (y_1, y_2)$ its child generated by the Gaussian mutation (7).

Since the mutation obeys the Gaussian probability distribution (8), its probability density function is

$$p_g(\vec{x}; \vec{y}) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{\frac{(y_1 - x_1)^2}{2\sigma_1^2} + \frac{(y_2 - x_2)^2}{2\sigma_2^2}\right\}. \quad (27)$$

Recalling that the sphere function is symmetric about the origin of coordinates, we set

$$\sigma_1 = \sigma_2 = \sigma.$$

Since the selection is elitist, the parent $\vec{x}$ can be replaced by a child $\vec{y}$ only if $\vec{y}$ falls in the promising region $S(\vec{x}, 1)$. For problem (26), the promising region $S(\vec{x}, 1)$ is the circle centred at $\vec{0} = (0, 0)$ with a radius $r = \| \vec{x} \|_2$. So,

$$P_g(\vec{x}; S(\vec{x}, 1))$$
$$= \frac{1}{2\pi\sigma^2} \int_{\vec{y} \in S(\vec{x}, 1)} \exp\left\{-\frac{\sum_{i=1}^2 (y_i - x_i)^2}{2\sigma^2}\right\} dy_1 dy_2$$
$$= \frac{1}{2\pi\sigma^2} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} d\theta \int_0^{2r\cos\theta} r \exp\left(-\frac{r^2}{2\sigma^2}\right) dr$$
$$= \frac{1}{2} - \frac{1}{\pi} \exp\left(-\frac{2r^2}{\sigma^2}\right) \int_0^{\frac{\pi}{2}} \exp\left(\frac{2r^2 \sin^2\theta}{\sigma^2}\right) d\theta. \quad (28)$$

If σ is a constant, then the mutation is landscape-invariant. When the (1+1) EA converges to the optimal solution, the radius r converges to 0. As a result, the value of (28) also converges to 0 since σ is a constant. This means that (12) in Lemma 2 is true. According to Theorem 1, R_t converges to 0 when $t \to +\infty$.

In order to obtain a positive ACR, the generator should be positive-adaptive, that is, $\exists C > 0$, $\rho \in (0, 1)$, $\forall \vec{x} \notin X^*$,

$$P_g(\vec{x}; S(\vec{x}, \rho)) \geq C.$$

In order to ensure a positive lower bound on $P(\vec{x}; \mathcal{S}(\vec{x}, \rho))$, we choose an adaptive σ. Denote $g(\theta) = \exp\left(\frac{2r^2 \sin^2 \theta}{\sigma^2}\right)$. From (28), we get

$$
\begin{aligned}
&P_g(\vec{x}; \mathcal{S}(\vec{x}, 1)) \\
=&\frac{1}{2} - \frac{1}{\pi} \exp\left(-\frac{2r^2}{\sigma^2}\right) \int_0^{\frac{\pi}{2}} g(\theta) d\theta \\
=&\frac{1}{2} - \frac{1}{\pi} \exp\left(-\frac{2r^2}{\sigma^2}\right) \left(\int_0^{\frac{\pi}{4}} + \int_{\frac{\pi}{4}}^{\frac{\pi}{2}}\right) g(\theta) d\theta \\
>&\frac{1}{2} - \frac{1}{\pi} \exp\left(-\frac{2r^2}{\sigma^2}\right) \left\{g\left(\frac{\pi}{4}\right) + g\left(\frac{\pi}{2}\right)\right\} \frac{\pi}{4} \\
=&\frac{1}{4} \left\{1 - \exp\left(-\frac{r^2}{\sigma^2}\right)\right\}.
\end{aligned}
$$

If r/σ is bounded below by a constant C_0, then

$$
P_g(\vec{x}; \mathcal{S}(\vec{x}, 1)) > \frac{1}{4} \left\{1 - \exp\left(-C_0^2\right)\right\}.
$$

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

$$
P_g(\vec{x}, \mathcal{S}(\vec{x}, \rho)) > P_g(\vec{x}, \mathcal{S}(\vec{x}, 1)) - \varepsilon
$$

for all ρ in $(1 - \delta, 1)$. Setting $\varepsilon = \frac{1}{8}\left\{1 - \exp\left(-C_0^2\right)\right\}$ and $\rho_0 = 1 - \frac{1}{2}\delta$, we know that the generator is positive-adaptive with the contractor factor ρ_0 for $C := \frac{1}{8}\left\{1 - \exp\left(-C_0^2\right)\right\}$.

For any σ such that $r/\sigma \geq C_0$, according to Theorem 2, the limit of R_t is a positive. A simple implementation is to let $\sigma_1 = \sigma_2 = \| \vec{x} \|_2$, which is the setting in section III-B.

This case study shows the applicability of our theory to uni-modal functions and confirms the importance of using an adaptive σ even for the sphere function. Moreover, practical EAs such as evolutionary programming and evolution strategies always adopt adaptive σ for a faster convergence speed.

B. 2-D Rastrigin Function

Consider minimization of the 2-D Rastrigin function:

$$
\min f_R(\vec{x}) = 20 + \sum_{k=1}^{2} (x_k^2 - 10 \cos 2\pi x_k), \quad (29)
$$

where $\vec{x} = (x_1, x_2) \in \mathbb{R}^2$. The optimal solution is $\vec{x}^* = (0, 0)$ with $f_R(\vec{x}^*) = 0$. The 2-D function is a sum of two 1-D Rastrigin functions as

$$
f_R(\vec{x}) = f_{R1}(x_1) + f_{R1}(x_2), \quad (30)
$$

where $f_{R1}(x) = 10 + x^2 - 10 \cos 2\pi x$.

The (1+1) elitist EA (Algorithm 2) is used to solve this minimization problem. Assume that $\vec{x} = (x_1, x_2)$ is the parent at the t-th generation at the fitness level $f_R(\vec{x}) = M$. Since the selection is elitist, the parent $\vec{x}$ is replaced by a child $\vec{y} = (y_1, y_2)$ only if $\vec{y}$ falls in the promising region $\mathcal{S}(\vec{x}, 1)$.

Fig. 4a shows the fitness landscape of the 2-D Rastrigin function. Fig. 4b illustrates the projection of the landscape at a fitness level $f_R(\vec{x}) = M$ to the decision plane.

Consider the partial derivative

$$
\frac{\partial f_R(\vec{x})}{\partial x_i} = 2x_i + 20\pi \sin 2\pi x_i = 0, \quad i = 1, 2. \quad (31)
$$

Because $\sin 2\pi x$ is a periodic function with values restricted in $[-1, 1]$, all solutions to equation (31) are located in $[-10\pi, 10\pi] \times [-10\pi, 10\pi]$. So, the 2-D Rastrigin has only finite global/local optimal solutions. $\forall\, \vec{x} \in \mathbb{R}^2$, the promising region $\mathcal{S}(\vec{x}, 1)$ is decomposed into finite mutually disjoint subsets (let m denote the number of subsets):

$$
\mathcal{S}(\vec{x}, 1) = \bigcup_{k=0}^{m} \mathcal{S}_k(\vec{x}, 1).
$$

Here we denote the subset including the global optimal solution as $\mathcal{S}_0(\vec{x}, 1)$, and $\mathcal{S}_k(\vec{x}, 1)$, $k = 1, \ldots, m$ are subsets containing local optimal solutions.

Similarly, the promising region of 1-D Rastrigin function $\mathcal{S}(x, 1)$ (where $x \in \mathbb{R}$) is a union of finite mutually disjoint intervals (without causing confusion, let n denote the number of intervals here but not dimensions)

$$
\mathcal{S}(x, 1) = \bigcup_{i=0}^{n} \mathcal{S}_k(x, 1).
$$

Here $\mathcal{S}_0(x, 1)$ denotes the interval including the global optimal solution and $\mathcal{S}_k(x, 1)$ $(k = 1, \ldots, n)$ are intervals containing local optimal solutions.

Since our ultimate goal is to find the global optimal solution, we consider maximization of the probability of locating the global optimal subset $\mathcal{S}_0(\vec{x}, 1)$, which is

$$
P_g(\vec{x}; \mathcal{S}_0(\vec{x}, 1)) = \int_{\vec{y} \in \mathcal{S}_0(\vec{x}, 1)} p_g(\vec{x}; \vec{y}) d\vec{y}. \quad (32)
$$

(30) implies that

$$
\mathcal{S}_0(x_1, 1) \times \mathcal{S}_0(x_2, 1) \subset \mathcal{S}_0(\vec{x}, 1).
$$

From (32), we know that

$$
\begin{aligned}
&P_g(\vec{x}; \mathcal{S}_0(\vec{x}, 1)) \\
\geq&P_g(\vec{x}; \mathcal{S}_0(x_1, 1) \times \mathcal{S}_0(x_2, 1)) \\
=&P_g(x_1; \mathcal{S}_0(x_1, 1)) \cdot P_g(x_2; \mathcal{S}_0(x_2, 1)). \quad (33)
\end{aligned}
$$

In the following, we try to maximize

$$
P_g(x; \mathcal{S}_0(x, 1)) = \int_{y \in \mathcal{S}_0(x, 1)} p_g(x; y) dy \quad (34)
$$

for any non-optimal solution x.

Without loss of generality, we assume $x > 0$, and denote $\mathcal{S}_k(x, 1)$ as $[a_k, b_k]$, where $k = 0, 1, \ldots, n$. The interval containing the global optimal solution $x^* = 0$ of f_{R1} is

$$
\mathcal{S}_0(x, 1) = [a_0, b_0].
$$

Symmetry of 1-D Rastrigin's function indicates that

$$
\mathcal{S}_0(x, 1) = [-b_0, b_0], \quad b_0 > 0. \quad (35)
$$

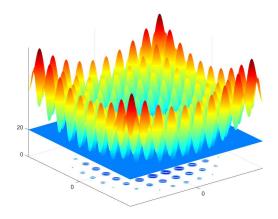To maximize $P_g(x; \mathcal{S}_0(x, 1))$, we introduce a proposition.

Proposition 1: Given interval $[a, b]$ and $x \notin [a, b]$, y is generated by 1-D Gaussian mutation $y = x + \mathcal{N}(0, \sigma)$.
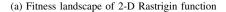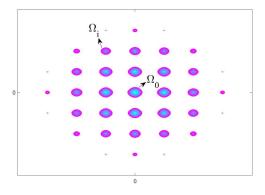
1) the transition probability from x to $[a, b]$ is

$$
P_g(x; [a, b]) = \Phi(u/\sigma) - \Phi(l/\sigma), \quad (36)
$$

where $l = \min\{|a - x|, |b - x|\}$, $u = \max\{|a - x|, |b - x|\}$; and $\Phi(\cdot)$ is the standard Gaussian distribution:

$$
\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{-\frac{y^2}{2}\right\} dy;
$$

(a) Fitness landscape of 2-D Rastrigin function

(b) Promising region with the fitness $f_R(\vec{x}) = 20$.

Fig. 4: The promising region of the 2-D Rastrigin Function when the fitness level $f_R(\vec{x}) = 20$.

2) if $l = 0$, $P_g(x, [a, b])$ is monotonously decreasing with $\sigma \in (0, +\infty)$; if $l > 0$, $P_g(x, [a, b])$ is maximized at some $\sigma_0 \in (0, +\infty)$.

Proof: The probability of $y \in [a, b]$ is

$$
\begin{aligned}
P_g(x; [a, b]) &= \Pr(y \in [a, b]) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_a^b \exp\{-\frac{(y-x)^2}{2\sigma^2}\} dy \\
&= \Phi\left(\frac{b-x}{\sigma}\right) - \Phi\left(\frac{a-x}{\sigma}\right).
\end{aligned}
$$

Then, we get

$$
P_g(x, [a, b]) = \Phi(u/\sigma) - \Phi(l/\sigma),
$$

by setting $l = \min\{|a-x|, |b-x|\}$, $u = \max\{|a-x|, |b-x|\}$.

If $l = 0$,

$$
P_g(x, [a, b]) = \Phi(u/\sigma) - \Phi(l/\sigma) = \Phi(u/\sigma) - \frac{1}{2}.
$$

Obviously, $P_g(x, [a, b])$ is monotonously decreasing with σ.

Otherwise, we have

$$
\begin{aligned}
&\frac{\partial}{\partial \sigma}\left\{\Phi\left(\frac{u}{\sigma}\right) - \Phi\left(\frac{l}{\sigma}\right)\right\} \\
&= \frac{-1}{\sqrt{2\pi}}\left\{\frac{u}{\sigma^2}\exp\left(-\frac{u^2}{2\sigma^2}\right) - \frac{l}{\sigma^2}\exp\left(-\frac{l^2}{2\sigma^2}\right)\right\}. \quad (37)
\end{aligned}
$$

Consider the partial derivative

$$
\frac{\partial}{\partial y}\left\{\frac{y}{\sigma^2}\exp\left(-\frac{y^2}{2\sigma^2}\right)\right\} = \frac{1}{\sigma^2}\exp\left(-\frac{y^2}{2\sigma^2}\right)\left(1 - \frac{y^2}{\sigma^2}\right).
$$

Its value is greater than zero when $|y| < \sigma$, smaller than zero when $|y| < \sigma$, and equal to zero when $|y| = \sigma$. So, we know that $\frac{y}{\sigma^2}\exp\left\{-\frac{y^2}{2\sigma^2}\right\}$ is

- monotonously increasing with y when $y \in (0, \sigma)$;
- locally maximized at $y = \sigma$;
- monotonously decreasing with y when $y \in (\sigma, +\infty)$.

Then, we can conclude that when σ varies from 0 to $+\infty$, the value of (37) changes gradually from positive values to negative values. This means that $P_g(x, [a, b])$ reaches its maximum value at some $\sigma_0 \in (0, +\infty)$. ∎

Denoting

$$
\begin{aligned}
l_0 &= \min\{|-b_0 - x|, |b_0 - x|\}, \\
u_0 &= \max\{|-b_0 - x|, |b_0 - x|\}.
\end{aligned} \quad (38)
$$

from (35) and (36) we know that

$$
P_g(x, \mathcal{S}_0(x, 1)) = \Phi(u_0/\sigma) - \Phi(l_0/\sigma). \quad (39)
$$

Estimation of (39) can be achieved by distinguishing three different regions where x locates.

- x is located in the "Outside Region", highlighted by blue segments in Fig. 5, where the abstract value of x is sufficiently great such that

$$
f_{R1}(y) \leq f_{R1}(x), \ \forall y \in [-x, x].
$$

In this case, we have

$$
\mathcal{S}(x, 1) = \mathcal{S}_0(x, 1) = [-b_0, b_0] = [-x, x].
$$

- x is located in the "Multimodal Region", highlighted by green segments in Fig. 5. In this case, we have

$$
\mathcal{S}(x, 1) = \bigcup_{i=0}^n \mathcal{S}_k(x, 1),
$$

where $\mathcal{S}_0(x, 1) = [-b_0, b_0]$ for some $b_0 \leq x$.

- x is located in the absorbing region of $x^* = 0$, named as the "Unimodal Region" highlighted by a red segment in Fig. 5. In this case, we have

$$
\mathcal{S}(x, 1) = \mathcal{S}_0(x, 1) = [-b_0, b_0] = [-x, x].
$$

So, the value of (39) can be estimated as follows.

- When x is located in the "Outside Region" or the "Unimodal Region", we have $l_0 = 0$. Then, (38) and (39) imply that

$$
P_g(x, \mathcal{S}_0(x, 1)) = \Phi\left(\frac{u_0}{\sigma}\right) - \frac{1}{2} = \Phi\left(\frac{2x}{\sigma}\right) - \frac{1}{2}.
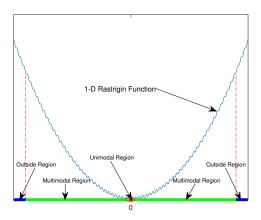$$

Fig. 5: Partition of the decision region of 1-D Rastrigin function.

By Theorems 1 and 2 we know that when $x \to 0$, a landscape-adaptive strategy should be employed to prevent $P_g(x, \mathcal{S}_0(x, 1))$ from converging to zero. A simple strategy is to set σ proportional to x. For example, letting $\sigma = x$, we have

$$P_g(x, \mathcal{S}_0(x, 1)) \mid_{\sigma=x} = \Phi(2) - \frac{1}{2}. \tag{40}$$

- When x is exploring the "Multimodal Region", we have $l_0 \geq 0$. Then Proposition 1 states that $P_g(x, \mathcal{S}_0(x, 1))$ is maximized at some $\sigma_0 \in (0, +\infty)$ which is a solution of

$$\frac{\partial}{\partial \sigma} \left(\Phi\left(\frac{u_0}{\sigma}\right) - \Phi\left(\frac{l_0}{\sigma}\right) \right) = 0.$$

That is,

$$\ln(u_0) - \ln(l_0) = \frac{u_0^2 - l_0^2}{2\sigma_0^2}. \tag{41}$$

Substituting $x_c = \frac{u_0 + l_0}{2}$ and $h = \frac{u_0 - l_0}{2}$ into (41), we know that

$$\ln(1 + \frac{h}{x_c}) - \ln(1 - \frac{h}{x_c}) = \frac{2x_c h}{\sigma_0^2}.$$

Since $0 < \frac{h}{x_c} < 1$, we can expand $\ln(1 + \frac{h}{x_c})$ and $\ln(1 - \frac{h}{x_c})$ by Taylor's series, which implies that

$$\frac{2h}{x_c} \leq \ln(1 + \frac{h}{x_c}) - \ln(1 - \frac{h}{x_c}) = \frac{2x_c h}{\sigma_0^2}.$$

So, we have $\sigma_0 \leq x_c$. Meanwhile, from (38) we know that $x_c = x$. By setting $\sigma = x$ we have

$$P_g(x, \mathcal{S}_0(x, 1)) \mid_{\sigma=x}$$
$$= \Phi\left(\frac{u_0}{x}\right) - \Phi\left(\frac{l_0}{x}\right)$$
$$= \Phi\left(1 + \frac{h}{x}\right) - \Phi\left(1 - \frac{h}{x}\right). \tag{42}$$

Then, we can take (42) as a tight lower bound of $P_g(x, \mathcal{S}_0(x, 1)) \mid_{\sigma=\sigma_0}$. Note that it is positive and continuous for any $\frac{h}{|x|}$ in the multimodal region of f_{R1}. So, $\frac{h}{|x|}$

have a positive minimum value, denoted as $\frac{h_0}{x_0}$. That is to say, $P_g(x, \mathcal{S}_0(x, 1)) \mid_{\sigma=x}$ has a general positive lower bound

$$\Phi\left(1 + \frac{h_0}{x_0}\right) - \Phi\left(1 - \frac{h_0}{x_0}\right). \tag{43}$$

For the minimization problem of 2-D Rastrigin function, a simple implementation of landscape-adaptive Gaussian mutation is to set

$$\boldsymbol{\Sigma}^2 = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} x_1^2 & 0 \\ 0 & x_2^2 \end{pmatrix}. \tag{44}$$

For the (1+1) EA with the above setting of σ, the lower bound of $P_g(\vec{x}; \mathcal{S}_0(\vec{x}, 1))$ could be estimated for one of the following cases.

1) When both x_1 and x_2 are exploring the "Outside Region", $P_g(\vec{x}; \mathcal{S}_0(\vec{x}, 1))$ is greater than $\left(\Phi(2) - \frac{1}{2}\right)^2$.
2) While one component of $\vec{x}$ is exploring the "Multimodal Region" and another is exploring the "Outside Region", $P_g(\vec{x}; \mathcal{S}_0(\vec{x}, 1))$ is bounded below by

$$\left(\Phi(2) - \frac{1}{2}\right)\left(\Phi\left(1 + \frac{h_0}{x_0}\right) - \Phi\left(1 - \frac{h_0}{x_0}\right)\right).$$

3) When both x_1 and x_2 are exploring the "Multimodal Region", an lower bound of $P_g(\vec{x}; \mathcal{S}_0(\vec{x}, 1))$ is

$$\left(\Phi\left(1 + \frac{h_0}{x_0}\right) - \Phi\left(1 - \frac{h_0}{x_0}\right)\right)^2;$$

4) While one component of $\vec{x}$ is exploring the "Multimodal Region" and another is exploiting the "Unimodal Region", $P_g(\vec{x}; \mathcal{S}_0(\vec{x}, 1))$ is bounded below by

$$\left(\Phi(2) - \frac{1}{2}\right)\left(\Phi\left(1 + \frac{h_0}{x_0}\right) - \Phi\left(1 - \frac{h_0}{x_0}\right)\right).$$

5) When both x_1 and x_2 are exploiting the "Unimodal Region", $P_g(\vec{x}; \mathcal{S}_0(\vec{x}, 1))$ is greater than $\left(\Phi(2) - \frac{1}{2}\right)^2$.

Note that h_0 is the radius of some interval $\mathcal{S}(x_0, 1)$. Then, it always hold that $\frac{h_0}{x_0} \leq 1$. That is to say, we always have

$$\Phi(2) - \frac{1}{2} \geq \Phi\left(1 + \frac{h_0}{x_0}\right) - \Phi\left(1 - \frac{h_0}{x_0}\right).$$

So the general positive lower bound of $P_g(\vec{x}; \mathcal{S}_0(\vec{x}, 1))$, obtained by applying the simple landscape-adaptive strategy (44), is

$$C_0 = \left\{ \Phi\left(1 + \frac{h_0}{x_0}\right) - \Phi\left(1 - \frac{h_0}{x_0}\right) \right\}^2.$$

Similar to the argument in section V-A, from the continuity of $P_g(\vec{x}; \mathcal{S}_0(\vec{x}, 1))$ we can know that $\forall \, \varepsilon > 0, \exists \delta \in (0, 1)$ such that

$$P_g(\vec{x}, \mathcal{S}(\vec{x}, \rho)) > P_g(\vec{x}, \mathcal{S}(\vec{x}, 1)) - \varepsilon$$

for all ρ in $(1 - \delta, 1)$. Setting $\varepsilon = \frac{1}{2}C_0$ and $\rho_0 = 1 - \frac{1}{2}\delta$, we conclude that the strategy (44) is positive-adaptive with the contractor factor $\rho = 1 - \frac{1}{2}\delta$ for the positive constant

$$C := \frac{1}{2}C_0 = \frac{1}{2}\left\{ \Phi\left(1 + \frac{h_0}{x_0}\right) - \Phi\left(1 - \frac{h_0}{x_0}\right) \right\}^2.$$

According to Theorem 2, the ACR limit of the adaptive EA is positive.

This case study demonstrates the applicability of our theory to multi-modal functions.

VI. Conclusions

The work in this paper and [13] present the average convergence rate, a new measurement for evaluating the convergence speed of EAs and other randomized search heuristics. It is shown that the ACR is a good convergence metric satisfying feasible in calculation and rigorous in analysis.

In terms of the ACR, this paper proves the necessity of using adaptive generators for solving continuous optimization problems. Theorem 1 states that for EAs using landscape-invariant generators, the limit of their ACR is 0. Therefore, landscape-invariant generators lead to a poor convergence speed. Theorem 2 indicates that for EAs using positive-adaptive generators, it converges to the optimal set with a positive ACR. How to design positive-adaptive generators is crucial in continuous optimization. Two case studies, (1+1) EAs for minimizing the 2-D sphere function and Rastrigin function, demonstrate the applicability of our theory.

An important future work is to develop methods for estimating lower and upper bounds on the ACR in both continuous and discrete optimization.

References

[1] J. He and X. Yao, "Drift analysis and average time complexity of evolutionary algorithms," *Artificial Intelligence*, vol. 127, no. 1, pp. 57–85, 2001.

[2] S. Droste, T. Jansen, and I. Wegener, "On the analysis of the (1+ 1) evolutionary algorithm," *Theoretical Computer Science*, vol. 276, no. 1-2, pp. 51–81, 2002.

[3] Y. Chen, X. Zou, and J. He, "Drift conditions for estimating the first hitting times of evolutionary algorithm," *International Journal of Computer Mathematics*, vol. 88, no. 1, pp. 37–50, 2011.

[4] A. Agapie, M. Agapie, G. Rudolph, and G. Zbaganu, "Convergence of evolutionary algorithms on the n-dimensional continuous space." *IEEE Transactions on Cybernetics*, vol. 43, no. 5, pp. 1462–1472, 2013.

[5] H. Huang, W. Xu, Y. Zhang, Z. Lin, and Z. Hao, "Runtime analysis for continuous (1+ 1) evolutionary algorithm based on average gain model," *SCIENTIA SINICA Informationis*, vol. 44, no. 6, pp. 811–824, 2014.

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

[7] G. Rudolph, "Local convergence rates of simple evolutionary algorithms with Cauchy mutations," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 4, pp. 249–258, 1997.

[8] J. He and L. Kang, "On the convergence rate of genetic algorithms," *Theoretical Computer Science*, vol. 229, no. 1-2, pp. 23–39, 1999.

[9] J. He and X. Yu, "Conditions for the convergence of evolutionary algorithms," *Journal of Systems Architecture*, vol. 47, no. 7, pp. 601–612, 2001.

[10] A. Auger, "Convergence results for the $(1, \lambda)$-sa-es using the theory of ϕ-irreducible markov chains," *Theoretical Computer Science*, vol. 334, no. 1-3, pp. 35–69, 2005.

[11] A. Auger and N. Hansen, "Linear convergence of comparison-based step-size adaptive randomized search via stability of markov chains," *SIAM Journal on Optimization*, vol. 26, no. 3, pp. 1589–1624, 2016.

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

[13] J. He and G. Lin, "Average convergence rate of evolutionary algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 2, pp. 316–321, 2016.

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

[15] H.-G. Beyer and A. Melkozerov, "The dynamics of self-adaptive multi-recombinant evolution strategies on the general ellipsoid model," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 5, pp. 764–778, 2014.

[16] H.-G. Beyer and M. Hellwig, "The dynamics of cumulative step size adaptation on the ellipsoid model," *Evolutionary computation*, vol. 24, no. 1, pp. 25–57, 2016.

[17] Y. Akimoto, A. Auger, and 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.

[18] A. Auger and 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.

[19] ——, "Theory of evolution strategies: a new perspective," in *Theory of Randomized Search Heuristics: Foundations and Recent Developments*. World Scientific, 2011, pp. 289–325.

[20] M. Jebalia, A. Auger, and N. Hansen, "Log-linear convergence and divergence of the scale-invariant (1+ 1)-es in noisy environments," *Algorithmica*, vol. 59, no. 3, pp. 425–460, 2011.

[21] R. Varga, *Matrix Iterative Analysis*. Springer, 2009.

[22] L. Ding and L. Kang, "Convergence rates for a class of evolutionary algorithms with elitist strategy," *Acta Mathematica Scientia*, vol. 21, no. 4, pp. 531–540, 2001.

[23] J. L. Doob and J. L. Doob, *Stochastic processes*. Wiley New York, 1953, vol. 7, no. 2.

[24] S. Meyn and R. Tweedie, *Markov chains and stochastic stability*. Springer Verlag, 1993.

[25] H.-G. Beyer and H.-P. Schwefel, "Evolution strategies–a comprehensive introduction," *Natural computing*, vol. 1, no. 1, pp. 3–52, 2002.