Average Convergence Rate of Evolutionary Algorithms

He, Jun; Lin, Guangming

Published in:
IEEE Transactions on Evolutionary Computation

DOI:
10.1109/TEVC.2015.2444793

Publication date:
2015

Citation for published version (APA):
He, J., & Lin, G. (2015). Average Convergence Rate of Evolutionary Algorithms. *IEEE Transactions on Evolutionary Computation*, *20*(2), 316-321. https://doi.org/10.1109/TEVC.2015.2444793

Average Convergence Rate of Evolutionary Algorithms

Jun He and Guangming Lin

***Abstract*—In evolutionary optimization, it is important to understand how fast evolutionary algorithms converge to the optimum per generation, or their convergence rates. This paper proposes a new measure of the convergence rate, called the average convergence rate. It is a normalized geometric mean of the reduction ratio of the fitness difference per generation. The calculation of the average convergence rate is very simple and it is applicable for most evolutionary algorithms on both continuous and discrete optimization. A theoretical study of the average convergence rate is conducted for discrete optimization. Lower bounds on the average convergence rate are derived. The limit of the average convergence rate is analyzed and then the asymptotic average convergence rate is proposed.**

***Index Terms*—evolutionary algorithms, evolutionary optimization, convergence rate, Markov chain, matrix analysis**

I. Introduction

Evolutionary algorithms (EAs) belong to iterative methods. As iterative methods, a fundamental question is the convergence rate: how fast does an EA converge to the optimum per generation? According to [1], existing results on the convergence rate of genetic algorithms can be classified into two categories. The first category is related to the eigenvalues of the transition matrix associated with an EA. A lower bound of convergence rate is derived in [2] for simple genetic algorithms through analyzing eigenvalues of the transition matrix. Then the work is extended in [3] and it is found that the convergence rate is determined by the second largest eigenvalue of the transition matrix. The other category is based on Doeblin's condition. The upper bound on the convergence rate is derived using Deoblin's condition in [4]. As to continuous optimization, the local convergence rate of EAs on the sphere function, quadratic convex functions and convex objective functions are discussed in [5]–[7]. The research of the convergence rate covers various types of EAs such as isotropic algorithms [8], gene expression programming [9], multiobjective optimization EAs [10]. The relationship between the convergence rate and population size is investigated in [11], [12].

The convergence rate in previous studies [1]–[4] is based on Markov chain theory. Suppose that an EA is modeled by a finite Markov chain with a transition matrix $\mathbf{P}$, in which a state is a population [13]. Let $\mathbf{p}_t$ be the probability distribution of the tth generation population on a population space, $\boldsymbol{\pi}$ an invariant probability distribution of $\mathbf{P}$. Then $\mathbf{p}_t$ is called *convergent* to $\boldsymbol{\pi}$ if $\lim_{t\to\infty} \parallel \mathbf{p}_t - \boldsymbol{\pi} \parallel = 0$ where $\parallel \cdot \parallel$ is a norm; and the *convergence rate* refers to the order of how fast $\mathbf{p}_t$ converges to $\boldsymbol{\pi}$ [4]. The goal is to obtain a bound $\beta(t)$ such that $\parallel \mathbf{p}_t - \boldsymbol{\pi} \parallel \leq \beta(t)$. But to obtain a closed form of $\beta(t)$ often is difficult in both theory and practice.

The current paper aims to seek a convergence rate satisfying two requirements: it is easy to calculate the convergence rate in practice while it is possible to make a rigorous analysis in theory. Inspired from conventional iterative methods [14], a new measure of the convergence rate, called the average convergence rate, is presented. The paper is organized as follows: Section II defines the average convergence rate. Section III establishes lower bounds on the average convergence rate. Section IV discusses the connections between the average convergence rate and other performance measures. Section V introduces an alternative definition of the average convergence rate if the optimal fitness value is unknown. Section VI concludes the paper.

II. Definition and Calculation

Consider the problem of minimizing (or maximizing) a function $f(x)$. An EA for solving the problem is regarded as an iterative procedure (Algorithm 1): initially construct a population of solutions Φ_0; then generate a sequence of populations Φ_1, Φ_2, Φ_3 and so on. This procedure is repeated until a stopping criterion is satisfied. An archive is used for recording the best found solution.

Algorithm 1 An EA with an archive

1: initialize a population of solutions Φ_0 and set $t \leftarrow 0$;
2: an archive records the best solution in Φ_0;
3: **while** the archive doesn't include an optimal solution **do**
4: generate a new population of solutions Φ_{t+1};
5: update the archive if a better solution is generated;
6: $t \leftarrow t + 1$;
7: **end while**

The fitness of population Φ_t is defined by the best fitness value among its individuals, denoted by $f(\Phi_t)$. Since $f(\Phi_t)$ is a random variable, we consider its expected value

Manuscript received xx xx xxxx

This work was supported by EPSRC under Grant No. EP/I009809/1 (He), National Science Foundation of Guangdong Province under Grant No. S2013010014108 and Shenzhen Scientific Research Project under Grant No. JCYJ 20130401095559825 (Lin).

Jun He is with Department of Computer Science, Aberystwyth University, Aberystwyth SY23 3DB, U.K. e-mail: jqh@aber.ac.uk.

Guangming Li is with Shenzhen Institute of Information Technology, Shenzhen 518172, China. e-mail: lingm@sziit.com.cn.

This is the author's version of an article that has been published in IEEE Transactions on Evolutionary Computation 20(2): 316 - 321, 2016. Changes were made to this version by the publisher prior to publication. The final version of record is available at http://dx.doi.org/10.1109/TEVC.2015.2444793

$f_t := \mathrm{E}[f(\Phi_t)]$. Let f_{opt} denote the optimal fitness. The *fitness difference* between f_{opt} and f_t is $|f_{\mathrm{opt}} - f_t|$. The *convergence rate* for one generation is

$$\left| \frac{f_{\mathrm{opt}} - f_t}{f_{\mathrm{opt}} - f_{t-1}} \right|. \tag{1}$$

Since $|f_{\mathrm{opt}} - f_t| \approx |f_{\mathrm{opt}} - f_{t-1}|$, calculating the above ratio is unstable in practice. Thus a new average convergence rate for EAs is proposed in the current paper.

Definition 1: Given an initial population Φ_0, the *average (geometric) convergence rate of an EA for t generations* is

$$\begin{aligned} R(t \mid \Phi_0) &:= 1 - \left(\left| \frac{f_{\mathrm{opt}} - f_1}{f_{\mathrm{opt}} - f_0} \right| \cdots \left| \frac{f_{\mathrm{opt}} - f_t}{f_{\mathrm{opt}} - f_{t-1}} \right| \right)^{1/t} \\ &\equiv 1 - \left(\left| \frac{f_{\mathrm{opt}} - f_t}{f_{\mathrm{opt}} - f_0} \right| \right)^{1/t}. \end{aligned} \tag{2}$$

If $f_0 = f_{\mathrm{opt}}$, let $R(t \mid \Phi_0) = 1$. For the sake of simplicity, $R(t)$ is short for $R(t \mid \Phi_0)$.

The rate represents a normalized geometric mean of the reduction ratio of the fitness difference per generation. The larger the convergence rate, the faster the convergence. The rate takes the maximal value of 1 at $f_t = f_{\mathrm{opt}}$.

Inspired from conventional iterative methods [14, Definition 3.1], the *average (logarithmic) convergence rate* is defined as follows:

$$R^{\dagger}(t) := -\frac{1}{t} \log \left| \frac{f_{\mathrm{opt}} - f_t}{f_{\mathrm{opt}} - f_0} \right|. \tag{3}$$

Formula (3) is not adopted since its value is $+\infty$ at $f_t = f_{\mathrm{opt}}$. But in most cases, average geometric and logarithmic convergence rates are almost the same. Since usually $|f_{\mathrm{opt}} - f_t|/|f_{\mathrm{opt}} - f_{t-1}| \approx 1$ and then $(|f_{\mathrm{opt}} - f_t|/|f_{\mathrm{opt}} - f_0|)^{1/t} \approx 1$, so $R^{\dagger}(t) \div R(t) \approx 1$.

In practice, the average convergence rate is calculated as follows: given $f(x)$ with f_{opt} known in advance,

1) Run an EA for T times ($T \gg 1$).
2) Then calculate the mean fitness value f_t as follows,

$$\frac{1}{T} \left(f(\Phi_t^{[1]}) + \cdots + f(\Phi_t^{[T]}) \right), \tag{4}$$

   where $f(\Phi_t^{[k]})$ denotes the fitness $f(\Phi_t)$ at the kth run. The law of large numbers guarantees (4) approximating to the expected fitness value $f_t = \mathrm{E}(f(\Phi_t))$ when T tends towards $+\infty$.
3) Finally, calculate $R(t)$ according to formula (2).

The calculation is applicable for most EAs on both continuous and discrete optimization. We take an example to illustrate the average convergence rate. Consider the problem of minimizing Ackley's function:

$$\begin{aligned} f(x) = &-20 \exp\{-0.2[\sum_{i=1}^{n} (x_i + e)^2/n]^{\frac{1}{2}}\} \\ &- \exp[\sum_{i=1}^{n} \cos(2\pi x_i + 2\pi e)/n] + 20 + e, \end{aligned} \tag{5}$$

where $x_i \in [-32 - e, 32 - e], i = 1, \cdots, n$. The optimum is $(-e, -e, \cdots)$ and $f_{\mathrm{opt}} = 0$. We compare the Multi-grid EA (MEA) [15] with the Fast Evolutionary Programming (FEP) [16] under the same experimental setting (where n is 30 and population size is 100). Run the two EAs for 1500 generations and 100 times. Calculate f_t according to (4) and then $R(t)$ according to (2). Fig. 1 illustrates the convergence rates of MEA and FEP.

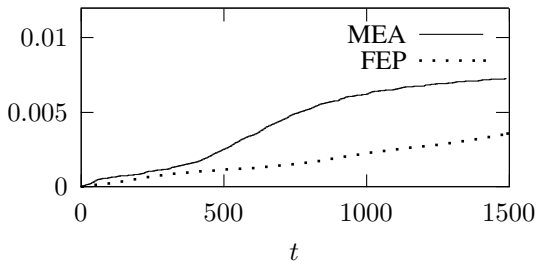

Fig. 1. A comparison of the average convergence rates of MEA and FEP on Ackley's function.

The average convergence rate is different from the progress rate such as the fitness error $|f_t - f_{\mathrm{opt}}|$ used in [17], [18] or logarithmic rate $\log|f_t - f_{\mathrm{opt}}|$ used in [19], [20]. The progress rate measures the fitness change; but the convergence rate measures the rate of the fitness change. We demonstrate this difference by an example. Let $g(x) = 100f(x)$. In terms of $|f_t - f_{\mathrm{opt}}|$, the progress rate on $g(x)$ is 100 times that on $f(x)$. In terms of $\log|f_t - f_{\mathrm{opt}}|$, the progress rate on $g(x)$ is $1 + 2/\log|f_t - f_{\mathrm{opt}}|$ times that on $f(x)$. However, the average convergence rate is the same on both $f(x)$ and $g(x)$.

III. Analysis for Discrete Optimization

Looking at Fig. 1 again, two questions may be raised: what is the lower bound or upper bound on $R(t)$? Does $R(t)$ converge or not? For discrete optimization, a theoretical answer is provided to these questions in this section. For continuous optimization, its analysis is left for future research.

In the rest of the paper, we analyze EAs for discrete optimization and assume that their genetic operators do not change with time. Such an EA can be modeled by a homogeneous Markov chain [13] with transition probabilities $\Pr(X, Y) := \Pr(\Phi_{t+1} = Y \mid \Phi_t = X), X, Y \in S$, where populations X, Y denote states of Φ_t and S denotes the set of populations (called the *population space*). Let $\mathbf{P}$ denote the transition matrix with entries $\Pr(X, Y)$.

A population is called *optimal* if it includes an optimal solution; otherwise called *non-optimal*. Let S_{opt} denote the set of optimal populations, and $S_{\mathrm{non}} = S \setminus S_{\mathrm{opt}}$. Because of the stopping criterion, the optimal set is always absorbing,

$$\Pr(\Phi_{t+1} \in S_{\mathrm{non}} \mid \Phi_t \in S_{\mathrm{opt}}) = 0. \tag{6}$$

Transition matrix $\mathbf{P}$ can be split into four parts:

$$\mathbf{P} = \begin{array}{c} \\ S_{\mathrm{opt}} \\ S_{\mathrm{non}} \end{array} \begin{array}{c} \begin{array}{cc} S_{\mathrm{opt}} & S_{\mathrm{non}} \end{array} \\ \begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{B} & \mathbf{Q} \end{pmatrix} \end{array} \tag{7}$$

where $\mathbf{A}$ is a submatrix representing probability transitions among optimal states; $\mathbf{O}$ a submatrix for probability transitions

from optimal states to non-optimal ones, of which all entries take the value of zero; $\mathbf{B}$ a submatrix denoting probability transitions from non-optimal states to optimal ones; and $\mathbf{Q}$ a submatrix for probability transitions among non-optimal states.

Since Φ_t is a random variable, we investigate the probability distribution of Φ_t instead of Φ_t itself. Let $q_t(X)$ denote the probability of Φ_t at a non-optimal state X, $q_t(X) := \Pr(\Phi_t = X)$. Let vector $(X_1, X_2, \cdots)$ represent all non-optimal states and vector $\mathbf{q}_t^T$ denote the probability distribution of Φ_t in the non-optimal set, where $\mathbf{q}_t := (q_t(X_1), q_t(X_2), \cdots)^T$. Here notation $\mathbf{q}$ is a column vector and $\mathbf{q}^T$ the row column with the transpose operation. For the initial probability distribution, $\mathbf{q}_0 \geq \mathbf{0}$ where $\mathbf{0} = (0, 0, \cdots)^T$. Only when the initial population is chosen from the optimal set, $\mathbf{q}_0 = \mathbf{0}$.

Consider probability transitions among non-optimal states only, which can be represented by matrix iteration

$$\mathbf{q}_t^T = \mathbf{q}_{t-1}^T \mathbf{Q} = \mathbf{q}_0^T \mathbf{Q}^t. \tag{8}$$

Definition 2: An EA is called *convergent* if for any $\mathbf{q}_0$, $\lim_{t\to+\infty} \mathbf{q}_t = \mathbf{0}$ or $\lim_{t\to+\infty} \mathbf{Q}^t = \mathbf{O}$. It is equivalent to saying that the probability of finding an optimal solution is 1 as t tends towards $+\infty$.

The expected fitness value f_t is given as follows:

$$f_t := \mathrm{E}[f(\Phi_t)] = \sum_{X\in S} f(X) \Pr(\Phi_t = X). \tag{9}$$

Then it follows

$$f_{\mathrm{opt}} - f_t = \sum_{X\in S_{\mathrm{non}}} (f(X) - f_{\mathrm{opt}}) q_t(X). \tag{10}$$

Let vector $\mathbf{f} := (f(X_1), f(X_2), \cdots)^T$ denote the fitness values of all non-optimal populations $(X_1, X_2, \cdots)$. Then (10) can be rewritten in a vector form

$$f_{\mathrm{opt}} - f_t = \mathbf{q}_t^T \cdot (f_{\mathrm{opt}} \mathbf{1} - \mathbf{f}), \tag{11}$$

where $\cdot$ denotes the vector product and $\mathbf{1} = (1, 1, \cdots)^T$.

For a vector $\mathbf{v}$, denote

$$\| \mathbf{v}^T \| := |\mathbf{v}^T \cdot (f_{\mathrm{opt}} \mathbf{1} - \mathbf{f})|. \tag{12}$$

Since $\| \mathbf{v} \| = 0$ iff $\mathbf{v} = \mathbf{0}$; $\| a\mathbf{v} \| = |a| \, \| \mathbf{v} \|$ and $\| \mathbf{v}_1 + \mathbf{v}_2 \| \geq \| \mathbf{v}_1 \| + \| \mathbf{v}_2 \|$, thus $\| \mathbf{v} \|$ is a vector norm. For a matrix $\mathbf{M}$, let $\| \mathbf{M} \|$ be the induced matrix norm, given by

$$\| \mathbf{M} \| = \sup \left\{ \frac{\| \mathbf{v}^T \mathbf{M} \|}{\| \mathbf{v}^T \|} : \mathbf{v} \neq \mathbf{0} \right\}. \tag{13}$$

Using the above Markov chain model, we are able to estimate lower bounds on the average convergence rate.

Theorem 1: Let $\mathbf{Q}$ be the transition submatrix associated with a convergent EA. For any $\mathbf{q}_0 \neq \mathbf{0}$,

1) The average convergence rate for t iterations is lower-bounded by
$$R(t) \geq 1 - \| \mathbf{Q}^t \|^{1/t}. \tag{14}$$
2) The limit of the average convergence rate for t generations is lower-bounded by
$$\lim_{t\to+\infty} R(t) \geq 1 - \rho(\mathbf{Q}), \tag{15}$$
where $\rho(\mathbf{Q})$ is the spectral radius (i.e., the supremum among the absolute values of all eigenvalues of $\mathbf{Q}$).
3) Under random initialization (that is, $\Pr(\Phi_0 = X) > 0$ for any $X \in S_{\mathrm{non}}$ or $\mathbf{q}_0 > \mathbf{0}$), it holds
$$\lim_{t\to+\infty} R(t) = 1 - \rho(\mathbf{Q}). \tag{16}$$
4) Under particular initialization (that is, set[1] $\mathbf{q}_0 = \mathbf{v}/|\mathbf{v}|$ where $\mathbf{v}$ is an eigenvector corresponding to the eigenvalue $\rho(\mathbf{Q})$ with $\mathbf{v} \geq \mathbf{0}$ but $\mathbf{v} \neq \mathbf{0}$. The existence of such a $\mathbf{v}$ is given in the proof), it holds for all $t \geq 1$,
$$R(t) = 1 - \rho(\mathbf{Q}). \tag{17}$$

Proof: 1) From (8): $\mathbf{q}_t^T = \mathbf{q}_0^T \mathbf{Q}^t$, we have

$$\frac{|f_{\mathrm{opt}} - f_t|}{|f_{\mathrm{opt}} - f_0|} = \frac{\| \mathbf{q}_t^T \|}{\| \mathbf{q}_0^T \|} = \frac{\| \mathbf{q}_0^T \mathbf{Q}^t \|}{\| \mathbf{q}_0^T \|} \leq \frac{\| \mathbf{q}_0^T \| \| \mathbf{Q}^t \|}{\| \mathbf{q}_0^T \|} = \| \mathbf{Q}^t \|. \tag{18}$$

Hence

$$1 - \left| \frac{f_{\mathrm{opt}} - f_t}{f_{\mathrm{opt}} - f_0} \right|^{1/t} \geq 1 - \| \mathbf{Q}^t \|^{1/t}, \tag{19}$$

which proves the first conclusion.

2) According to Gelfand's spectral radius formula [21, p.619], we get

$$\lim_{t\to+\infty} \| \mathbf{Q}^t \|^{1/t} = \rho(\mathbf{Q}). \tag{20}$$

The second conclusion follows by combining (20) with (14).

3) Since $\mathbf{Q} \geq 0$, according to the extension of Perron-Frobenius' theorems to non-negative matrices [21, pp. 670], $\rho(\mathbf{Q})$ is an eigenvalue of $\mathbf{Q}$. There exists an eigenvector $\mathbf{v}$ corresponding to $\rho(\mathbf{Q})$ such that $\mathbf{v} \geq \mathbf{0}$ but $\mathbf{v} \neq \mathbf{0}$. In particular,

$$\rho(\mathbf{Q}) \mathbf{v}^T = \mathbf{v}^T \mathbf{Q}. \tag{21}$$

Let $\max(\mathbf{v})$ denote the maximum value of the entries in vector $\mathbf{v}$. Due to random initialization, $\mathbf{q}_0 > 0$. Let $\min(\mathbf{q}_0)$ denote the minimum value of the entries in vector $\mathbf{q}_0$. Set

$$\mathbf{u} = \frac{\min(\mathbf{q}_0)}{\max(\mathbf{v})} \mathbf{v}. \tag{22}$$

From (21), we get

$$\rho(\mathbf{Q}) \mathbf{u}^T = \mathbf{u}^T \mathbf{Q}. \tag{23}$$

Thus vector $\mathbf{u}$ is an eigenvector of $\rho(\mathbf{Q})$.

Let $\mathbf{w} = \mathbf{q}_0 - \mathbf{u}$. Then from (22), we know $\mathbf{w} \geq 0$. Since $\mathbf{q}_0 = \mathbf{u} + \mathbf{w}$, $\mathbf{w} \geq 0$ and $\mathbf{Q} \geq 0$, we deduce that

$$\mathbf{q}_t^T = \mathbf{q}_0^T \mathbf{Q}^t = (\mathbf{u} + \mathbf{w})^T \mathbf{Q}^t \geq \mathbf{u}^T \mathbf{Q}^t = \rho(\mathbf{Q})^t \mathbf{u}^T. \tag{24}$$

It follows that

$$\left| \frac{f_{\mathrm{opt}} - f_t}{f_{\mathrm{opt}} - f_0} \right| = \left| \frac{\mathbf{q}_t^T \cdot (f_{\mathrm{opt}} \mathbf{1} - \mathbf{f})}{f_{\mathrm{opt}} - f_0} \right| \geq \left| \frac{\rho(\mathbf{Q})^t \mathbf{u}^T \cdot (f_{\mathrm{opt}} \mathbf{1} - \mathbf{f})}{\mathbf{q}_0^T \cdot (f_{\mathrm{opt}} \mathbf{1} - \mathbf{f})} \right|. \tag{25}$$

[1]For vector $\mathbf{v} = (v_1, v_2, \cdots)$, denote $|\mathbf{v}| := \sum_i |v_i|$.

$$\left|\frac{f_{\text{opt}}-f_t}{f_{\text{opt}}-f_0}\right|^{1/t} \geq \rho(\mathbf{Q})\left|\frac{\mathbf{u}^T\cdot(f_{\text{opt}}\mathbf{1}-\mathbf{f})}{\mathbf{q}_0^T\cdot(f_{\text{opt}}\mathbf{1}-\mathbf{f})}\right|^{1/t}. \quad (26)$$

Since both $|\mathbf{u}^T\cdot(f_{\text{opt}}\mathbf{1}-\mathbf{f})|$ and $|\mathbf{q}_0^T\cdot(f_{\text{opt}}\mathbf{1}-\mathbf{f})|$ are independent of t, we let $t\to+\infty$ and get

$$\lim_{t\to+\infty}\left|\frac{\mathbf{u}^T\cdot(f_{\text{opt}}\mathbf{1}-\mathbf{f})}{\mathbf{q}_0^T\cdot(f_{\text{opt}}\mathbf{1}-\mathbf{f})}\right|^{1/t}=1, \quad (27)$$

then we get

$$\lim_{t\to+\infty}\left|\frac{f_{\text{opt}}-f_t}{f_{\text{opt}}-f_0}\right|^{1/t}\geq\rho(\mathbf{Q}). \quad (28)$$

$$\lim_{t\to+\infty}R(t)=1-\lim_{t\to+\infty}\left|\frac{f_{\text{opt}}-f_t}{f_{\text{opt}}-f_0}\right|^{1/t}\leq 1-\rho(\mathbf{Q}). \quad (29)$$

The third conclusion follows by combining (29) with (15).

4) Set $\mathbf{q}_0=\mathbf{v}/\sum_i v_i$ where $\mathbf{v}$ is given in Step 3. Then $\mathbf{q}_0$ is an eigenvector corresponding to the eigenvalue $\rho(\mathbf{Q})$ such that $\rho(\mathbf{Q})\mathbf{q}_0^T=\mathbf{q}_0^T\mathbf{Q}$. From (8): $\mathbf{q}_t^T=\mathbf{q}_{t-1}^T\mathbf{Q}$, we get

$$\frac{f_{\text{opt}}-f_t}{f_{\text{opt}}-f_0}=\frac{\mathbf{q}_t^T\cdot(f_{\text{opt}}\mathbf{1}-\mathbf{f})}{f_{\text{opt}}-f_0}=\frac{\rho(\mathbf{Q})^t\mathbf{q}_0^T\cdot(f_{\text{opt}}\mathbf{1}-\mathbf{f})}{\mathbf{q}_0^T\cdot(f_{\text{opt}}\mathbf{1}-\mathbf{f})}.$$

Thus we have for any $t\geq 1$

$$\left|\frac{f_{\text{opt}}-f_t}{f_{\text{opt}}-f_0}\right|^{1/t}=\rho(\mathbf{Q}), \quad (30)$$

then $R(t)=1-\rho(\mathbf{Q})$ which gives the fourth conclusion. ■

The above theorem provides lower bounds on the average convergence rate. Furthermore, it reveals that $R(t)$ converges to $1-\rho(\mathbf{Q})$ under random initialization and $R(t)=1-\rho(\mathbf{Q})$ for any $t\geq 1$ under particular initialization. Similar to conventional iterative methods [14, pp. 73], we call $1-\rho(\mathbf{Q})$ the *asymptotic average convergence rate* of an EA, denoted by R_∞. According to (16), its value can be approximately calculated as follows: under random initialization, $R(t)$ approximates to $1-\rho(\mathbf{Q})$ if t is sufficiently large. Note that this definition is different from another asymptotic convergence rate, given by $-\log\rho(\mathbf{Q})$ in [22]. In most cases, the two rates are almost the same since usually $\rho(\mathbf{Q})\approx 1$ and then $-\log\rho(\mathbf{Q})\div(1-\rho(\mathbf{Q}))\approx 1$. Since $1-\rho(\mathbf{Q})$ is independent of t and initialization, hence using asymptotic average convergence rate is convenient for comparing two EAs, for example, to analyze mixed strategy EAs [22].

IV. Connections

The average convergence rate is different from other performance measures of EAs: the expected hitting time is the total number of generations for obtaining an optimal solution [13]; and fixed budget analysis focuses on the performance of EAs within fixed budget computation [23]. However, there are some interesting connections between them.

There exists a link between the asymptotic average convergence rate and the hitting time. Let $m(X)$ be the expected number of generations for a convergent EA to hit S_{opt} when starting from state X (called the *expected hitting time*). Denote $\mathbf{m}:=(m(X_1),m(X_2),\cdots)^T$ where $(X_1,X_2,\cdots)$ represent all non-optimal states.

Theorem 2: Let $\mathbf{Q}$ be the transition submatrix associated with a convergent EA. Then $1/R_\infty$ is not more than $\|\mathbf{m}\|_\infty:=\max\{m(X);X\in S_{\text{non}}\}$.

Proof: According to the fundamental matrix theorem [24, Theorem 11.5], $\mathbf{m}=(\mathbf{I}-\mathbf{Q})^{-1}\mathbf{1}$, where $\mathbf{I}$ is the unit matrix. Then

$$\begin{aligned}\|\mathbf{m}\|_\infty &=\|(\mathbf{I}-\mathbf{Q})^{-1}\mathbf{1}\|_\infty=\|(\mathbf{I}-\mathbf{Q})^{-1}\|_\infty\\ &\geq\rho((\mathbf{I}-\mathbf{Q})^{-1})=(1-\rho(\mathbf{Q}))^{-1},\end{aligned} \quad (31)$$

where the last equality takes use of a fact: $(1-\rho(\mathbf{Q}))^{-1}$ is an eigenvalue and spectral radius of $(\mathbf{I}-\mathbf{Q})^{-1}$. ■

The above theorem shows that $1/R_\infty$ is a lower bound on the expected hitting time.

Following Theorem 1, a straightforward connection can be established between the spectral radius $\rho(\mathbf{Q})$ and the progress rate $|f_{\text{opt}}-f_t|$.

Corollary 1: Let $\mathbf{Q}$ be the transition submatrix associated with a convergent EA.

1) Under random initialization (that is $\mathbf{q}_0>\mathbf{0}$), it holds

$$\lim_{t\to+\infty}\frac{|f_{\text{opt}}-f_t|^{1/t}}{\rho(\mathbf{Q})|f_{\text{opt}}-f_0|^{1/t}}=1. \quad (32)$$

2) Under particular initialization (that is, set $\mathbf{q}_0=\mathbf{v}/|\mathbf{v}|$ where $\mathbf{v}$ is an eigenvector corresponding to the eigenvalue $\rho(\mathbf{Q})$ with $\mathbf{v}\geq\mathbf{0}$ but $\mathbf{v}\neq\mathbf{0}$), it holds for all $t\geq 1$,

$$\frac{|f_{\text{opt}}-f_t|}{\rho(\mathbf{Q})^t|f_{\text{opt}}-f_0|}=1. \quad (33)$$

The corollary demonstrates that the exponential decay, $\rho(\mathbf{Q})^t|f_{\text{opt}}-f_0|$, provides a theoretical prediction for the trend of $|f_{\text{opt}}-f_t|$.

We explain the theoretical results by a simple example. Consider a (1+1) EA for maximizing the OneMax function $|\mathbf{x}|$ where $\mathbf{x}=(s_1,\cdots,s_n)\in\{0,1,\}^n$, which is the easiest fitness function to the $(1+1)$ EA [25].

Algorithm 2 A (1+1) elitits EA

Onebit Mutation: choose a bit of Φ_t (one individual) uniformly at random and flip it. Let Ψ_t denote the child.
Elitist Selection: if $f(\Psi_t)>f(\Phi_t)$, then let $\Phi_{t+1}\leftarrow\Psi_t$; otherwise $\Phi_{t+1}\leftarrow\Phi_t$.

Denote subset $S_k:=\{\mathbf{x}:|\mathbf{x}|=n-k\}$ where $k=0,\cdots,n$. Transition probabilities satisfy that $\Pr(\Phi_{t+1}\in S_{k-1}\mid\Phi_t\in S_k)=\frac{k}{n}$ and $\Pr(\Phi_{t+1}\in S_k\mid\Phi_t\in S_k)=1-\frac{k}{n}$. Writing them in matrix $\mathbf{P}$ (where submatrix $\mathbf{Q}$ in the bold font):

$$\begin{pmatrix}1 & 0 & 0 & \cdots & 0 & 0 & 0\\ \frac{1}{n} & \mathbf{1-\frac{1}{n}} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0}\\ 0 & \mathbf{\frac{2}{n}} & \mathbf{1-\frac{2}{n}} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0}\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ 0 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{\frac{n-1}{n}} & \mathbf{1-\frac{n-1}{n}} & \mathbf{0}\\ 0 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{1} & \mathbf{0}\end{pmatrix}. \quad (34)$$

The spectral radius $\rho(\mathbf{Q})=1-\frac{1}{n}$ and the asymptotic average convergence rate $R_\infty=\frac{1}{n}$. Notice that $1/R_\infty(=n)$ is less than the expected hitting time $(=n(1+\frac{1}{2}+\cdots+\frac{1}{n}))$.

In the OneMax function, set $n = 10$, and then $\rho(\mathbf{Q}) = 0.9$ and $R_\infty = 0.1$. Choose Φ_0 uniformly at random, run the (1+1) EA for 60 generations and 2000 times, and then calculate f_t according to (4) and $R(t)$ according to formula (2). Since Φ_0 is chosen uniformly at random, $f_0 \approx 5$. Fig. 2 demonstrates that $R(t)$ approximates $0.1(= R_\infty)$. Fig. 3 shows that the theoretical exponential decay, $\rho(\mathbf{Q})^t|f_{\rm opt} - f_0|$, and the computational progress rate, $|f_{\rm opt} - f_t|$, coincide perfectly.

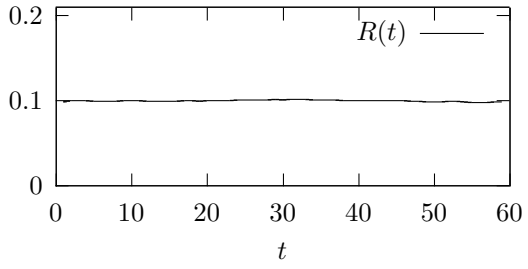

Fig. 2. $R(t)$ approximates 0.1 for the (1+1) EA on the OneMax function with $n = 10$.

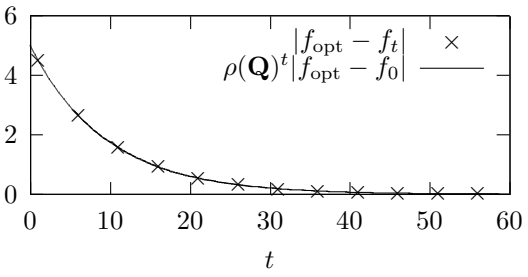

Fig. 3. A comparison of the theoretical prediction $\rho(\mathbf{Q})^t|f_{\rm opt} - f_0|$ and the computational result $|f_{\rm opt} - f_t|$ for the (1+1) EA on the OneMax function with $n = 10$, $f_0 = 5$, $f_{\rm opt} = 10$ and $\rho(\mathbf{Q}) = 0.9$.

V. Alternative Rate

So far the calculation of the average convergence rate needs the information about $f_{\rm opt}$. However this requirement is very strong. Here we introduce an alternative average convergence rate without knowing $f_{\rm opt}$, which is given as below,

$$R^\ddagger(t) := 1 - \left|\frac{f_{t+\tau} - f_t}{f_t - f_{t-\tau}}\right|^{1/\tau}, \tag{35}$$

where τ is an appropriate time interval. Its value relies on an EA and a problem.

For the (1+1) EA on the OneMax function with $n = 10$, we set $\tau = 10$. Choose Φ_0 uniformly at random, run the (1+1) EA for 60 generations and 2000 times, and then calculate f_t according to (4) and $R^\ddagger(t)$ according to formula (35). Due to $\tau = 10$, $R^\ddagger(t)$ has no value for $t < 10$ and $t > 50$ according to formula (35). Fig. 4 demonstrates that $R^\ddagger(t)$ approximates $0.1(= 1 - \rho(\mathbf{Q}))$.

The above average convergence rate converges to $1 - \rho(\mathbf{Q})$ but under stronger conditions than that in Theorem 1.

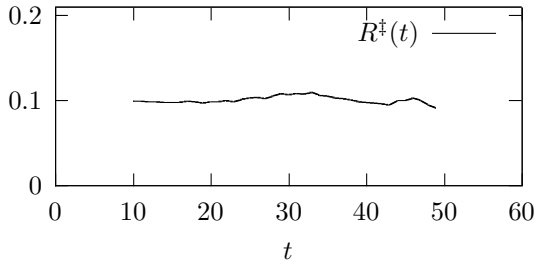

Fig. 4. $R^\ddagger(t)$ approximates 0.1 for the (1+1) EA on the OneMax function with $n = 10$.

Theorem 3: Let $\mathbf{Q}$ be the transition submatrix associated with a convergent EA.

1) Under particular initialization (that is, set $\mathbf{q}_0 = \mathbf{v}/|\mathbf{v}|$ where $\mathbf{v}$ is an eigenvector corresponding to the eigenvalue $\rho(\mathbf{Q})$ with $\mathbf{v} \geq \mathbf{0}$ but $\mathbf{v} \neq \mathbf{0}$), it holds for all $t \geq 1$,

$$R^\ddagger(t) = 1 - \rho(\mathbf{Q}). \tag{36}$$

2) Under random initialization (that is $\mathbf{q}_0 > \mathbf{0}$), choose an appropriate τ such that $\mathbf{g} := (\mathbf{I} - \mathbf{Q}^\tau)(f_{\rm opt}\mathbf{1} - \mathbf{f}) > \mathbf{0}$ for a maximization problem (or $\mathbf{g} < \mathbf{0}$ for a minimization problem)[2]. If $\mathbf{Q}$ is positive[3], then it holds

$$\lim_{t\to+\infty} R^\ddagger(t) = 1 - \rho(\mathbf{Q}). \tag{37}$$

Proof: From (8): $\mathbf{q}_t^T = \mathbf{q}_{t-1}^T\mathbf{Q}$ and (11), we get

$$\begin{aligned} f_{t+\tau} - f_t &= f_{t+\tau} - f_{\rm opt} + f_{\rm opt} - f_t \\ &= \mathbf{q}_t^T \cdot (f_{\rm opt}\mathbf{1} - \mathbf{f}) - \mathbf{q}_{t+\tau}^T \cdot (f_{\rm opt}\mathbf{1} - \mathbf{f}) \\ &= \mathbf{q}_t^T \cdot (f_{\rm opt}\mathbf{1} - \mathbf{f}) - \mathbf{q}_t^T\mathbf{Q}^\tau(f_{\rm opt}\mathbf{1} - \mathbf{f}) \\ &= \mathbf{q}_t^T \cdot \mathbf{g}. \end{aligned} \tag{38}$$

1) Since $\mathbf{q}_0$ is an eigenvector corresponding to the eigenvalue $\rho(\mathbf{Q})$ such that $\rho(\mathbf{Q})\mathbf{q}_0^T = \mathbf{q}_0^T\mathbf{Q}$. From (38) and (8): $\mathbf{q}_t^T = \mathbf{q}_{t-1}^T\mathbf{Q}$, we get

$$\begin{aligned} \left|\frac{f_{t+\tau} - f_t}{f_t - f_{t-\tau}}\right|^{1/\tau} &= \left|\frac{\mathbf{q}_t^T \cdot \mathbf{g}}{\mathbf{q}_{t-\tau}^T \cdot \mathbf{g}}\right|^{1/\tau} = \left|\frac{\mathbf{q}_0^T\mathbf{Q}^t\mathbf{g}}{\mathbf{q}_0^T\mathbf{Q}^{t-\tau}\mathbf{g}}\right|^{1/\tau} \\ = \left|\frac{\rho(\mathbf{Q})^t}{\rho(\mathbf{Q})^{t-\tau}} \times \frac{\mathbf{q}_0^T \cdot \mathbf{g}}{\mathbf{q}_0^T \cdot \mathbf{g}}\right|^{1/\tau} &= \rho(\mathbf{Q}). \end{aligned}$$

Then $R^\ddagger(t) = 1 - \rho(\mathbf{Q})$ which gives the first conclusion.

2) Without loss of the generality, consider $\mathbf{g} > \mathbf{0}$. Since

$$\frac{f_{t+\tau} - f_t}{f_t - f_{t-\tau}} = \frac{\mathbf{q}_t^T \cdot \mathbf{g}}{\mathbf{q}_{t-\tau}^T \cdot \mathbf{g}} = \frac{\mathbf{q}_{t-\tau}^T\mathbf{Q}^\tau\mathbf{g}}{\mathbf{q}_{t-\tau}^T \cdot \mathbf{g}}, \tag{39}$$

let

$$\underline{\lambda}_t = \min_i \frac{[\mathbf{q}_{t-\tau}^T\mathbf{Q}^\tau]_i}{[\mathbf{q}_{t-\tau}^T]_i}, \qquad \overline{\lambda}_t = \max_i \frac{[\mathbf{q}_{t-\tau}^T\mathbf{Q}^\tau]_i}{[\mathbf{q}_{t-\tau}^T]_i}, \tag{40}$$

[2]It is always true for a large time interval τ since $\lim_{\tau\to+\infty}(\mathbf{I} - \mathbf{Q}^\tau) = \mathbf{I}$ and $f_{\rm opt}\mathbf{1} - \mathbf{f} > \mathbf{0}$ for a maximization problem (or $f_{\rm opt}\mathbf{1} - \mathbf{f} < \mathbf{0}$ for a minimization problem).

[3]The condition of positive $\mathbf{Q}$ could be relaxed to non-negative $\mathbf{Q}$ if taking a similar argument to the extension of Perron-Frobenius' theorems to non-negative matrices [21, pp. 670].

where $[\mathbf{v}]_i$ represents the ith entry in vector $\mathbf{v}$.

According to Collatz formula [26] [27, Theorem 2],

$$\lim_{t\to+\infty} \underline{\lambda}_t = \lim_{t\to+\infty} \overline{\lambda}_t = \rho(\mathbf{Q}^\tau). \tag{41}$$

Hence for any $[\mathbf{g}]_i > 0$, it holds

$$\lim_{t\to+\infty} \min_i \frac{[\mathbf{q}_{t-\tau}^T \mathbf{Q}^\tau]_i [\mathbf{g}]_i}{[\mathbf{q}_{t-\tau}^T]_i [\mathbf{g}]_i} = \rho(\mathbf{Q}^\tau), \tag{42}$$

$$\lim_{t\to+\infty} \max_i \frac{[\mathbf{q}_{t-\tau}^T \mathbf{Q}^\tau]_i [\mathbf{g}]_i}{[\mathbf{q}_{t-\tau}^T]_i [\mathbf{g}]_i} = \rho(\mathbf{Q}^\tau). \tag{43}$$

Using $\min\{\frac{a_1}{b_1}, \frac{a_2}{b_2}\} \le \frac{a_1+a_2}{b_1+b_2} \le \max\{\frac{a_1}{b_1}, \frac{a_2}{b_2}\}$, we get

$$\lim_{t\to+\infty} \frac{\sum_i [\mathbf{q}_{t-\tau}^T \mathbf{Q}^\tau]_i [\mathbf{g}]_i}{\sum_i [\mathbf{q}_{t-\tau}^T]_i [\mathbf{g}]_i} = \rho(\mathbf{Q}^\tau). \tag{44}$$

Equivalently

$$\lim_{t\to+\infty} \frac{\mathbf{q}_{t-\tau}^T \mathbf{Q}^\tau \mathbf{g}}{\mathbf{q}_{t-\tau}^T \cdot \mathbf{g}} = \rho(\mathbf{Q}^\tau). \tag{45}$$

Then

$$\lim_{t\to+\infty} \left| \frac{f_{t+\tau} - f_t}{f_t - f_{t-\tau}} \right|^{1/\tau} = \rho(\mathbf{Q}^\tau)^{1/\tau} = \rho(\mathbf{Q}). \tag{46}$$

Finally it comes to the second conclusion. ■

The theorem shows that the average convergence rate $R^\ddagger(t)$ may play the same role as $R(t)$ does. But the calculation of $R^\ddagger(t)$ is not as stable as that of $R(t)$ in practice.

VI. Conclusions

This paper proposes a new convergence rate of EAs, called the average (geometric) convergence rate. The rate represents a normalized geometric mean of the reduction ratio of the fitness difference per generation. The calculation of the average convergence rate is simple and easy to implement on most EAs in practice. Since the rate is normalized, it is convenient to compare different EAs on optimization problems.

For discrete optimization, lower bounds on the average convergence rate of EAs have been established. It is proven that under random initialization, the average convergence rate $R(t)$ for t generations converges to a limit, called the asymptotic average convergence rate; and under particular initialization, $R(t)$ equals to the asymptotic average convergence rate for any $t \ge 1$.

The analysis of EAs for continuous optimization is different from that for discrete optimization. In continuous optimization, an EA is modeled by a Markov chain on a general state space, rather than a finite Markov chain. So a different theoretical analysis is needed, rather than matrix analysis used in the current paper. This topic is left for future research.

References

[1] L. Ming, Y. Wang, and Y.-M. Cheung, "On convergence rate of a class of genetic algorithms," in *Proceedings of 2006 World Automation Congress*. IEEE, 2006, pp. 1–6.

[2] J. Suzuki, "A Markov chain analysis on simple genetic algorithms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 25, no. 4, pp. 655–659, 1995.

[3] F. Schmitt and F. Rothlauf, "On the importance of the second largest eigenvalue on the convergence rate of genetic algorithms," in *Proceedings of 2001 Genetic and Evolutionary Computation Conference*, H. Beyer, E. Cantu-Paz, D. Goldberg, Parmee, L. Spector, and D. Whitley, Eds. Morgan Kaufmann Publishers, 2001, pp. 559–564.

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

[5] 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.

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

[7] ——, "Convergence rates of evolutionary algorithms for quadratic convex functions with rank-deficient hessian," in *Adaptive and Natural Computing Algorithms*. Springer, 2013, pp. 151–160.

[8] O. Teytaud, S. Gelly, and J. Mary, "On the ultimate convergence rates for isotropic algorithms and the best choices among various forms of isotropy," in *Parallel Problem Solving from Nature (PPSN IX)*. Springer, 2006, pp. 32–41.

[9] X. Du and L. Ding, "About the convergence rates of a class of gene expression programming," *Science China Information Sciences*, vol. 53, no. 4, pp. 715–728, 2010.

[10] N. Beume, M. Laumanns, and G. Rudolph, "Convergence rates of SMS-EMOA on continuous bi-objective problem classes," in *Proceedings of the 11th Workshop on Foundations of Genetic Algorithms*. ACM, 2011, pp. 243–252.

[11] M. Jebalia and A. Auger, "Log-linear convergence of the scale-invariant $(\mu/\mu_w, \lambda)$-ES and optimal μ for intermediate recombination for large population sizes," in *Parallel Problem Solving from Nature (PPSN XI)*. Springer, 2010, pp. 52–62.

[12] F. Teytaud and O. Teytaud, "Convergence rates of evolutionary algorithms and parallel evolutionary algorithms," in *Theory and Principled Methods for the Design of Metaheuristics*. Springer, 2014, pp. 25–39.

[13] J. He and X. Yao, "Towards an analytic framework for analysing the computation time of evolutionary algorithms," *Artificial Intelligence*, vol. 145, no. 1-2, pp. 59–97, 2003.

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

[15] J. He and L. Kang, "A mixed strategy of combining evolutionary algorithms with multigrid methods," *International Journal of Computer Mathematics*, vol. 86, no. 5, pp. 837–849, 2009.

[16] X. Yao, Y. Liu, and G. Lin, "Evolutionary programming made faster," *Evolutionary Computation, IEEE Transactions on*, vol. 3, no. 2, pp. 82–102, 1999.

[17] S.-M. Guo and C.-C. Yang, "Enhancing differential evolution utilizing eigenvector-based crossover operator," *Evolutionary Computation, IEEE Transactions on*, vol. 19, no. 1, pp. 31–49, Feb 2015.

[18] L. Tang, Y. Dong, and J. Liu, "Differential evolution with an individual-dependent mechanism," *Evolutionary Computation, IEEE Transactions on*, vol. 19, no. 4, pp. 560–574, Aug 2015.

[19] R. Salomon, "Evolutionary algorithms and gradient search: similarities and differences," *IEEE Transactions on Evolutionary Computation*, vol. 2, no. 2, pp. 45–55, 1998.

[20] R. Sarker, S. Elsayed, and T. Ray, "Differential evolution with dynamic parameters selection for optimization problems," *Evolutionary Computation, IEEE Transactions on*, vol. 18, no. 5, pp. 689–707, Oct 2014.

[21] C. Meyer, *Matrix Analysis and Applied Linear Algebra*. SIAM, 2000.

[22] J. He, F. He, and H. Dong, "Pure strategy or mixed strategy?" in *Evolutionary Computation in Combinatorial Optimization*, J.-K. Hao and M. Middendorf, Eds. Springer, 2012, pp. 218–229.

[23] T. Jansen and C. Zarges, "Performance analysis of randomised search heuristics operating with a fixed budget," *Theoretical Computer Science*, vol. 545, pp. 39–58, 2014.

[24] C. Grinstead and J. Snell, *Introduction to Probability*. American Mathematical Society, 1997.

[25] J. He, T. Chen, and X. Yao, "On the easiest and hardest fitness functions," *IEEE Transactions on Evolutionary Computation*, vol. 19, no. 2, pp. 295 – 305, 2015.

[26] L. Collatz, "Einschließungssätze für charakteristische zahlen von matrizen," *Mathematische Zeitschrift*, vol. 48, no. 1, pp. 221–226, 1942.

[27] R. J. Wood and M. J. O'Neill, "Finding the spectral radius of a large sparse non-negative matrix," in *Proceedings of the 13th Biennial Computational Techniques and Applications Conference, CTAC-2006*, ser. ANZIAM J., W. Read and A. J. Roberts, Eds., vol. 48, Aug. 2007, pp. C330–C345.