Review Convergence of Canonical Genetic Algorithm

BY TINGYU ZHANG

2023.10.16

1 Introduction

The target is to use Canonical genetic algorithm(CGA) to tackle the static optimization problems of the type:

$$\max \{ f(b) | b \in \mathbb{B}^l \} \tag{1}$$

which $\mathbb{B}^l := \{0, 1\}^l$, and has the condition $0 < f(b) < \infty$ for all $b \in \mathbb{B}^l$ and $f(b) \neq \text{const}$.

In optimization theory an algorithm is said to converge to the global optimum if it generates a sequence of solutions or function values in which the global optimum is a limit value. Especially,

it is proved by means of homogeneous finite Markov chains that the CGA neverconverges to the global optimum, but modi¿ed versions do.Precisely, probabilistic convergence of the best solution within a population to the global optimum under elitist selection, which is the theorem the rest of the paper will prove.

2 Basic Tools

We will introduce CGA and finite Markov Chain as the background:

2.1 CGA

A genetic algorithm consists of an n-tuple of binary strings b_i of length l, where the bits of each string are considered to be the genes of an individual chromosome and where the n-tuple of individual chromosomes is said to be a population.

Example 1. For a GA problem, suppose we have a genetic algorithm in which each chromosome is a binary string of length 5, and our population size is 3. A possible representation of the population could be:

Chromosome 1: 10101

Chromosome 2: 11010

Chromosome 3: 00111

The tuple consisting of these three chromosomes, $(10101, 11010, 00111)$, represents our population. b_i represents a certain chromosome of population, $l = 5, n = 3, i \in [\![1, n]\!]$

Following the terminology of organic evolution the operations performed on the population are called mutation, crossover and selection (differential reproduction,差异化繁殖). Each individual bi represents a feasible solution of problem (1) and its ob jective function value $f(b_i)$ is said to be its fitness which is to be maximized.

Remark 1. Fitness function is a judgement criterion that is positive and increase more approach to the better solution. For example, $\max g(x) = -x^2 + 4x, x \in [1, 5]$, the corresponding fitness function can be $f(x) = g(x) + 100$. And $\min g(x) = x^2, x \in [-10, 10]$, the corresponding fitness function can be

$f(x)=\frac{1}{x^2+0.1}$, where 0.1 is for avoiding the denominator to be 0.

The algorithm is sketched as follows:

```
choose an initial population
determine the fitness of each individual
perform selection
repeat
perform crossover
perform mutation
determine the fitness of each individual
perform selection
until some stopping criterion applies
```

Mutation operates independently on each individual by probabilistically perturbing each bit string. The event that the j-th bit of the i-th individual is flipped is stochastically independent and occurs with probability $p_m \in (0,1)$. For example, the probability that string $b = 00000$ transitions to string $b' = 10110$ by mutation is $p_m^k (1-p_m)^{l-k}$ with $k=3$ and $l=5$. Clearly, k is just the Hamming distance $H(b,b')$ between strings b and b'. Therefore the probability that string b_i resembles string b_i' after mutation can be aggregated to

$$P\{b_i \to b_i'\} = p_m^{H(b,b')}(1-p_m)^{l-H(b,b')} > 0 \tag{2}$$

Crossover operator is applied with some probability $p_c \in [0,1]$ in order to construct a bit string from at least two other bit strings chosen at random. Although many crossover operators have been proposed a description can be omitted because the choice of a speciac crossover operator does not effect the subsequent analysis.(Check out)

For Proportional selection, the population of the next generation is determined by n independent random experiments. The probability that individual bi is selected from tuple $(b_1, \ldots, b_n)$ to be a member of the next generation at each experiment is given by:

$$P\{b_i \,\text{is selected}\} = \frac{f(b_i)}{\sum\limits_{j=1}^{n} f(b_j)} > 0 \tag{3}$$

2.2 Finite Markov Chains

A finite Markov chain describes a probabilistic trajectory over a finite state space S of cardinality $|S| = n$, where the states may be numbered from 1 to n.

The probability $p_{\mathrm{ij}}(t)$ of transitioning from state $i \in S$ to state $j \in S$ at step t is called the transition probability from i to j at step t.

If the transition probabilities are independent from t, i.e., $p_{\mathrm{ij}}(t) = p_{\mathrm{ij}}(s)$ for all $i, j \in S$ and for all $s; t \in \mathbb{N}$, the Markov chain is said to be homogeneous.

Remark 2. Homogeneity is a really crucial property, since only homogeneous Markov Chain can determain every step by transitioning matrix and initial state in the formula $p^t = p^0 P^t$.

Definition 1. *A square matrix $A \in \mathcal{M}_n(K)$ is said to be*

(a) nonnegative ($A \succcurlyeq 0$), if $a_{\mathrm{ij}} \geqslant 0$ for all $i, j \in \{1, \ldots, n\}$

(b) positive ($A \succ 0$), if $a_{\mathrm{ij}} > 0$ for all $i, j \in \{1, \ldots, n\}$

A nonnegative matrix $A \in \mathcal{M}_n(K)$ is said to be *(c) primitive, if there exists a* $k \in \mathbb{N}$ *such that* A^k *is positive,*

(d) reducible, if A *can be brought into the form (with square matrices* C *and* T*)*$\left(\begin{smallmatrix} C & \mathbf{0} \\ R & T \end{smallmatrix}\right)$ *by applying the same permutations to rows and columns,*

(e) irreducible, if it is not reducible,

(f) stochastic, if $\sum_{j=1}^{n} a_{\mathrm{ij}} = 1\,(\text{summation of the row of matrix})$ *for all* $i \in \{1, \ldots, n\}$

A stochastic matrix $A \in \mathcal{M}_n(K)$ *is said to be(g) stable, if it has identical rows,*

(h) column al lowable, if it has at least one positive entry in each column.

Quite a lot definitions are made, here comes what they can do:

Lemma 1. *Let* $C,$ M *and* S *be stochastic matrices, where* M *is positive and* S *is column allowable. Then the product* CMS *is positive.*

By the lamma 5, 2 theorem can be proved:

Theorem 1. *Let* P *be a primitive stochastic matrix. Then* P^k *converges as* $k \rightarrow \infty$ *to a positive stable stochastic matrix* $P^\infty = 1' p^\infty$*, where* $p^\infty = p^0 \lim_{k \rightarrow \infty} P^k = p^0 P^\infty$ *has nonzero entries andis unique regardless of the initial distribution.*

proof:

1. Positive stable stochastic get directly from definition.

2. By induction, all element in P^k *is smaller than 1. And by contradiction all eigen-valun of* P^k *must be smaller than 1. Since by* $|I - P| = 0, P\ \text{has 1 as eigen} - \text{value},$ *hence* P^k*has 1 as eigen-value. Hence convergence exist and not to 0.*

Theorem 2. *Let* P *be a reducible stochastic matrix, where* $C \in \mathcal{M}_m(K)$ *is a primitive stochastic matrix and* $R, T \neq 0$*. Then*

$$P^\infty = \lim_{k \rightarrow \infty} P^k = \lim_{k \rightarrow \infty} \left(\begin{array}{cc} C^k & 0 \\ \sum_{i=0}^{k-1} T^i \mathrm{RC}^{k-i-1} & T^k \end{array}\right) = \left(\begin{array}{c} C^k\, 0 \\ R_\infty\, 0 \end{array}\right) \tag{4}$$

is a stable stochastic matrix with $P^\infty = 1' p^\infty$, where $p^\infty = p^0 P^\infty$ is unique regardless of the initial distribution, and p^∞ satisfies: $p_i^\infty > 0$ for $1 \leqslant i \leqslant m$ and $p_i^\infty = 0$ for $m < i \leqslant n$.

3 Markov Chain Analysis of Genetic Algorithms

Determine the state space $S = \mathbb{B}^N = \mathbb{B}^{l * n}$, where l is the nunber of genes(or length of chromosome), and n denotes the population size,

Each element of the state space can be regarded as an integer number in binary representation.

$(n = \sum_{k=0}^{m} 2^k a_k)$

The projection $\pi_k(i)$ picks up the k-th bit segment of length l from the binary representation of state i and is used to identify single individuals from the population.

Example 2. For $b_i = 1111010110$, choose $l = 3$, $\pi_2(i) = 101$.

The probabilistic changes of the genes within the population caused by the genetic operators are captured by the transition matrix P, which can be decomposed in a natural way into a product of stochastic matrices $P = CMS$, where C, M and S describe the intermediate transitions caused by crossover, mutation and selection, respectively.

Theorem 3. *The transition matrix of the CGA with mutation probability* $p_m \in (0,1)$*, crossover probability* $p_c \in [0,1]$ *and proportional selection is primitive.*

proved by discussing the structure of C,G,A and lemma 1.

The result can be used to form the uniqueness limit state

Corollary 1. *The CGA with parameter ranges as in Theorem 3 is an ergodic Markov chain, i.e., there exists an unique limit distribution for the states of the chain with nonzero probability to be in any state at any time regardless of the initial distribution.*

proved by Theorom 1 and Theorom 3.

Remark 3. Ergodic Markov has another definition: the markov chain which is aperiodic, irreducible and finite recurrence.

From the result, we can apparently see that for such Markov chain the initial distribution p^0 doesn't affect the limit distribution at all !

The ergodicity property has consequences for the convergence behavior of the CGA. To avoid confusion, a precise de¸nition of the term convergence of a GA is required:

Definition 2. *Let* $Z_t = \max\left\{f\left(\pi_k^{(t)}(i) | k = 1,...,n\right)\right\}$ *be a sequence of random variables representing the best fitness within a population represented by state i at step t. A genetic algorithm converges to the global optimum, if and only if*

$$\lim_{t \to \infty} P\{Z_t = f^*\} = 1, \tag{5}$$

where$f^* = \max\{f(b) | b \in \mathbb{B}^l\}$ is is the global optimum of problem (1).

By this definition:

Theorem 4. *The CGA with parameter ranges as in Theorem 3 does not converge to the global optimum.*

Proof. *set* $\max\left\{f\left(\pi_k^{(t)}(i) | k = 1, \ldots, n\right)\right\} < f^*$, $p_i^t = \{\text{probability GA in state } i \text{ at step } t\}$;

$\underbrace{P\{Z_t \neq f^*\} \geqslant p_i^t}_{\text{see remark}} \Leftrightarrow P\{Z_t = f^*\} \leqslant 1 - p_i^t$. *Hence* $\lim_{t \to \infty} P\{Z_t = f^*\} \leqslant 1 - p_i^t < 1$. □

Remark 4. Does "$P\{Z_t \neq f^*\} \geqslant p_i^t$ " stand for the optimization may hard to get?

Can CGA be changed a little to fulfill the convergence? Actually it can be done by the theorem of ergodic Markov chain:

Theorem 5. *In an ergodic Markov chain the expected transition time between initial state i and any other state j is finite regardless of the states i and j, i.e. $E(T_{i\to j})<\infty$.*

3.1 Little Change to make convergence

To make the result in conformity with Definition 2:

Change: Enlarging the population by adding super individual which does not take part in the evolutionary process.

Cardinality: From $2^{n*l}\to 2^{(n+1)*l}$, give l bits for super inditivual leftmost and is accessible by $\pi_0(i)$ at state i.

Require :better the super individual's fitness the higher the position of the corresponding state in the matrix.

The extended transition matrices are written as diagonal matrices with 2^l square matrices C, M, S of size $2^{\text{nl}}\times 2^{\text{nl}}$.

The copy operation is represented by an upgrade matrix U which upgrades an intermediate state containing an individual better than its super individual to a state where the super individual equals the better individual.

Let $b=\text{argmax}\{f(\pi_k(i))|\,k=1,\ldots,n\}\in\mathbb{B}^l:=\{\text{best individual of the population at any state } i \text{ (exclude}$ super indivitual)$\}$.Then $u_{\text{ij}}=1$ if $f(\pi_0(i))<f(b)$ with $j\overset{\text{def}}{=}(b,\pi_1(i),\pi_2(i),\ldots,\pi_n(i))<S$, otherwise $u_{\text{ii}}=1$.(WHY?)

Thus, there is exactly one entry in each row, which does not hold for the columns because for every state $j\in s$ with $f(\pi_0(i))<\max\{f(\pi_k(i))|\,k=1,\ldots,n\}$ one gets $u_{\text{ij}}=0$ for all $i\in S$.

Theorem 6. *The canonical GA as in Theorem 3 maintaining the best solution found over time after selection converges to the global optimum.*

Theorem 7. *The canonical GA as in Theorem 3 maintaining the best solution found over time before selection converges globally optimal.*

Note 1. Theorems 6 and 7 do not cover the case of elitist selection.

4 Discussion of results with respect to the schema theorem

A schema(模式) S describes a specific type of subsets of the feasible region $\mathbb{B}^l$ of problem (1) which is again assumed to have only one global optimal point $b^*\in\mathbb{B}^l$.

Usually, these subsets are represented by a string of length l over the alphabet $\{0,1,\#\}$.

The utility of schema S restricted to multiset X is defined as the average objective function value over all elements contained in $S\cap X$:

$$u(S,X):=\frac{1}{|S\cap X|}\sum_{b\in S\cap X}f(b) \tag{6}$$

Then schema theorem is

$$E[|S\cap X|]\geqslant|S\cap X_t|\frac{u(S,X_t)}{u(\mathcal{S},X_t)}(1-c(S,X_t))(1-m(S,X_t)) \tag{7}$$

almost surely, where (X_t) is the sequence of populations generated by the CGA and c(.) and m(.) are bounds for the probability that an element of subset S is modfied by crossover and mutation respectively, so that the resulting element is not contained in subset S.

Remark 5. (7) state that the number of individuals in population $X_{t+1}$with above average fitness is expected to be no smaller than in population X_t,if the probabilities c(.) and m(.) are sufficiently small. But not fairly demonstrate the convergence to the global optimum.

Generally, the specific content of the schema theorem is: under the action of genetic operators selection, crossover, and mutation, schemas with low order, short defining length, and average fitness higher than the population average fitness will grow exponentially in the offspring.(Without proof)

Technically, it is neccessary and sufficient that $\lim_{t\to\infty} E[I_t]=1$ which implies $\lim_{t\to\infty} E[\{b^*\}\bigcap X_t]\geqslant 1$, where

$$I_t:=h(b^*,X_t):=\begin{cases}1,\text{if } b^*\in\{\pi_1(X_t),\ldots,\pi_n(X_t)\}\\ 0,\text{otherwise}\end{cases} \tag{8}$$

means when $t>N$ for certain $N\in\mathbb{N}$, b^* will be in one of the states.

In particular:

Lemma 2.

(a) $\lim_{t\to\infty} E[I_t]=1\Leftrightarrow\lim_{t\to\infty} P\{Z_t=f^*\}=1$

(b) $\lim_{t\to\infty} E[I_t]=1\Rightarrow\lim_{t\to\infty} E[\{b^*\}\bigcap X_t]\geqslant 1$

Proof:

(a) Take $I_t=1_{\{\pi_1(X_t),\ldots,\pi_n(X_t)\}}(b^*)$, and $b^*\in\{\pi_1(X_t),\ldots,\pi_n(X_t)\}\Leftrightarrow\{I_t=1\}\Leftrightarrow\{Z_t=f^*\}$

(b) $g(b^*,X_t):=$ count the number of optimal solution b^* in population $X_{t.}$ Compare to $h(b^*,X_t)$ which only count once $\Rightarrow$ $g(b^*,X_t)\geqslant h(b^*,X_t)\Rightarrow$

$$\lim_{t\to\infty} E\big[\{b^*\}\bigcap X_t\big]=\sum_{i=1}^{|S|} g(b^*,X_t)\cdot p_i^\infty\geqslant\sum_{i=1}^{|S|} h(b^*,X_t)\cdot p_i^\infty=\lim_{t\to\infty} E[I_t]=1 \tag{9}$$

Converse for (b) is not true:

Remark 6. For a CGA there is a minimal probability bounded from zero to lose the global optimum solution at each generation. It follows from the Borel-Cantelli Lemma that this event will occur with probability one.

Borel-Cantelli Lemma: $\{E_n\}$ belongs to a certain probability space, if $\sum_{n=1}^\infty \mathbb{P}(E_n)<\infty\Rightarrow\mathbb{P}\Big(\lim_{n\to\infty}\sup(E_n)\Big)=0$

Intuition : For converge series like $\{a_n\}$, $\sum_{n=N}^\infty a_n\to 0$ as $N\to\infty$.

conclude,the global solution will be lost and found infinitely often →the sequence $(|\{b^* \bigcap X_t\}|)$ is an irreducible markov chain on the state space $\{0,\ldots,n\}$ → does not converge although the expectation does

4.1 Analysis the bounds for the probabilities of losing and generating the optimal solution

Assume that $k \geqslant 1$ optimal solutions are contained in population Xt at generation t:

The crossover operator may destroy or assemble some optimal solutions, so that there are k optimal solutions within the population after crossover.

4.1.1 The Optimal Solution Gets Lost

$1.k \geqslant 1.$

The probability that at least one bit of an optimal solution is ipped is given by:

$p_F := 1 - \underbrace{(1-p_m)^l}_{\text{all} \, l - \text{bit do not change}} > 0,$ the probability that all k optimal solutions are destroyed becomes:

$p_F^k (1-p_F)^{n-k} > 0,$ bounded below: $\gamma_1 := \min \left\{ \underbrace{p_F^n}_{p_F < 1 - p_F}, \underbrace{(1-p_F)^{n-1}}_{1-p_F < p_F} \right\} > 0$

$2.k = 0.$(all optimal solutions have been destroyed by crossover)

The probability that all bits within the population remain unaltered:

$(1-p_m)^{n \cdot l} = (1-p_F)^n := \gamma_2 > 0.$

the probability that the optimal solution is lost after crossover and mutation is at least:

$p_L = \min\,(\gamma_1, \ \gamma_2) = \min\,\{p_F^n, \ (1-p_F)^{n-1}, \ (1-p_F)^n\} = \min\,\{p_F^n, \ (1-p_F)^n\} > 0$(all n destoyed, at least one not, all not)

4.1.2 The Optimal Solution Genrating

It remains to derive the bound for the probability to generate an optimal solution:

The probability that mutation generates the optimal solution b^* from individual b_i:

$p_{B_i} := p_m^{H(b_i,b^*)} (1-p_m)^{l - H(b_i,b^*)} > 0;$ Bounded below by:$p_B := \min\{p_{B_i} | \, i \in [\![1,n]\!]\} > 0.$

The probability that this event occurs at least once is:$p_G := 1 - (1-p_B)^n > 0.$

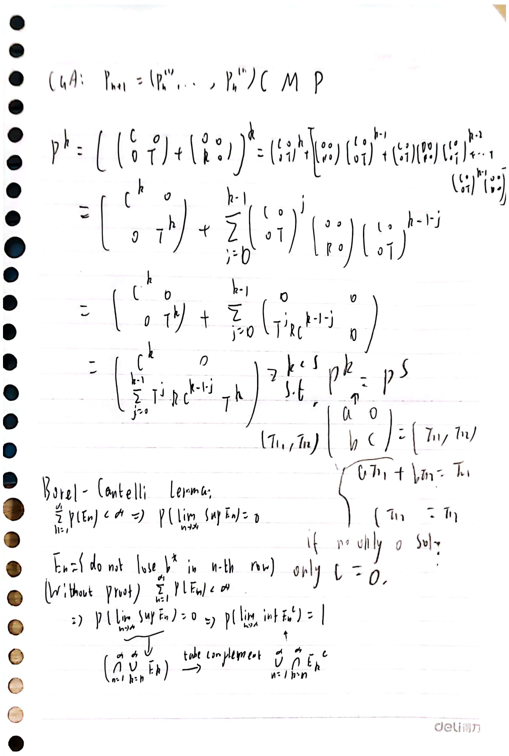
Next, consider the selection operator:

Assume that only one optimal solution has been generated by mutation with probability p_G.

The probability to select the optimal solution is given by: $p_{b^*} := \frac{f(b^*)}{\sum_{j=1}^n f(b_j)} > 0,$

The probability that this event occurs at least once becomes: $p_S := 1 - (1 - p_{b^*})^n > 0.$

The probability that a global solution is generated by mutation and that it survives the selection procedure can be bounded:$p_G \cdot p_S > 0.$



GA: $P_{n+1} = (P_n^{(1)}, \dots, P_n^{(n)})$ C M P

$$P^k = \left(\begin{pmatrix} C & 0 \\ 0 & T \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ R & 0 \end{pmatrix} \right)^k = \begin{pmatrix} C & 0 \\ 0 & T \end{pmatrix}^k + \left[\begin{pmatrix} 0 & 0 \\ R & 0 \end{pmatrix} \begin{pmatrix} C & 0 \\ 0 & T \end{pmatrix}^{k-1} + \begin{pmatrix} C & 0 \\ 0 & T \end{pmatrix} \begin{pmatrix} 0 & 0 \\ R & 0 \end{pmatrix} \begin{pmatrix} C & 0 \\ 0 & T \end{pmatrix}^{k-2} + \dots + \begin{pmatrix} C & 0 \\ 0 & T \end{pmatrix}^{k-1} \begin{pmatrix} 0 & 0 \\ R & 0 \end{pmatrix} \right]$$

$$= \begin{pmatrix} C^k & 0 \\ 0 & T^k \end{pmatrix} + \sum_{j=0}^{k-1} \begin{pmatrix} C & 0 \\ 0 & T \end{pmatrix}^j \begin{pmatrix} 0 & 0 \\ R & 0 \end{pmatrix} \begin{pmatrix} C & 0 \\ 0 & T \end{pmatrix}^{k-1-j}$$

$$= \begin{pmatrix} C^k & 0 \\ 0 & T^k \end{pmatrix} + \sum_{j=0}^{k-1} \begin{pmatrix} 0 & 0 \\ T^j R C^{k-1-j} & 0 \end{pmatrix}$$

$$= \begin{pmatrix} C^k & 0 \\ \sum_{j=0}^{k-1} T^j R C^{k-1-j} & T^k \end{pmatrix} \quad \exists\, k < s \text{ s.t. } P^k = P^s$$

$$(\pi_1, \pi_2) \begin{pmatrix} a & 0 \\ b & c \end{pmatrix} = (\pi_1, \pi_2)$$

$$\begin{cases} a\pi_1 + b\pi_2 = \pi_1 \\ c\pi_2 = \pi_2 \end{cases}$$

if no only 0 sol: only $c = 0$.

Borel-Cantelli Lemma:

$\sum_{n=1}^{\infty} P(E_n) < \infty \Rightarrow P(\lim_{n\to\infty} \sup E_n) = 0$

$E_n = \{$do not lose b^* in n-th row$\}$

(without proof) $\sum_{n=1}^{\infty} P(E_n) < \infty$

$\Rightarrow P(\underbrace{\lim_{n\to\infty} \sup E_n}_{\left(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} E_k\right)}) = 0 \Rightarrow P(\lim_{n\to\infty} \inf E_n^c) = 1$

$\left(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} E_k\right) \xrightarrow{\text{take complement}} \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} E_k^c$