Convergence Analysis of Canonical Genetic Algorithms

GÜNTER RUDOLPH

Abstract

This paper analyzes the convergence properties of the canonical genetic algorithm (CGA) with mutation, crossover and proportional reproduction applied to static optimization problems. It is proved by means of homogeneous finite Markov chain analysis that a CGA will never converge to the global optimum regardless of the initialization, crossover operator and objective function. But variants of CGAs that always maintain the best solution in the population, either before or after selection, are shown to converge to the global optimum due to the irreducibility property of the underlying original nonconvergent CGA. These results are discussed with respect to the schema theorem.

Keywords: canonical genetic algorithm, global convergence, Markov chains, schema theorem

1 Introduction

Canonical genetic algorithms (CGA) as introduced in [1] are often used to tackle static optimization problems of the type

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

assuming that $0 < f(b) < \infty$ for all $b \in \mathbb{B}^l = \{0,1\}^l$ and $f(b) \neq const.$ Although empirical evidence indicates that CGAs can sometimes find good solutions to complex problems there are few theoretical results concerning their convergence properties. 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. A probabilistic version of this definition is used in this paper.

Markov chains offer an appropriate model to analyze GAs and they have been used in [2] and [3] to prove probabilistic convergence of the best solution within a population to the global optimum under *elitist selection* (the best individual survives with probability one).

This paper analyzes the global convergence properties of the original CGA and modified versions that simply save the best discovered solution either before or after the reproduction operation. It is proved by means of homogeneous finite Markov chains that the CGA never converges to the global optimum, but modified versions do. To make these results rigorous, a precise definition of the term *convergence to the global optimum* is offered, which may be linked to the Markov chain model of the CGA. Therefore, after a brief description of the CGA given in section 2 some basic definitions and results from finite Markov chain theory are introduced in section 3. Section 4 is devoted to the Markov chain analysis of CGAs as well as their variants and contains the essential results. Finally, the results are discussed with respect to the schema theorem [1].

---

[1]The author is with the Department of Computer Science, LS XI, University of Dortmund, D–44221 Dortmund, Germany. This work was supported under project no. 107 004 91 by the Research Initiative *Parallel Computing* of the Ministry of Science and Research of Northrhine–Westfalia, Germany.

2 Canonical Genetic Algorithms

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*. Following the terminology of organic evolution the operations performed on the population are called *mutation*, *crossover* and *selection* (differential reproduction). Each individual b_i represents a feasible solution of problem (1) and its objective function value $f(b_i)$ is said to be its *fitness* which is to be maximized. The algorithm can be 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

When using so–called *proportional selection* the population of the next generation is determined by n independent random experiments. The probability that individual b_i is selected from tuple $(b_1, b_2, \ldots, b_n)$ to be a member of the next generation at each experiment is given by

$$P\{b_i \text{ is selected }\} = \frac{f(b_i)}{\sum\limits_{j=1}^{n} f(b_j)} > 0 \ . \tag{2}$$

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 $p_m \cdot (1-p_m) \cdot p_m \cdot p_m \cdot (1-p_m) = p_m^k (1-p_m)^{l-k}$ 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

$$P\{b_i \to b_i'\} = p_m^{H(b_i,b_i')} (1-p_m)^{l-H(b_i,b_i')} > 0 \ . \tag{3}$$

Usually, the 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 specific crossover operator does not effect the subsequent analysis.
This algorithm and its potential implementation is described in [4, pp. 59–70] in more detail.

3 Finite Markov Chains

A finite Markov chain describes a probabilistic trajectory over a finite state space $\mathcal{S}$ of cardinality $|\,\mathcal{S}\,| = n$, where the states may be numbered from 1 to n. The probability $p_{ij}(t)$ of transitioning from state $i \in \mathcal{S}$ to state $j \in \mathcal{S}$ 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., $p_{ij}(t) = p_{ij}(s)$ for all $i, j \in \mathcal{S}$ and for all $s, t \in \mathbb{N}$, the Markov chain is said to be *homogeneous*.

The transition probabilities of a homogeneous finite Markov chain can be gathered in a *transition matrix* $\mathbf{P} = (p_{ij})$. For each entry, $p_{ij} \in [0,1]$ and $\sum_{j=1}^{|\mathcal{S}|} p_{ij} = 1$ for all $i \in \mathcal{S}$. Matrices with the above properties are called *stochastic*. Given an initial distribution $\mathbf{p}^0$ as a row vector, the distribution of the chain after the $t-$th step is determined by $\mathbf{p}^t = \mathbf{p}^0\,\mathbf{P}^t$. Therefore, a homogeneous finite Markov chain is completely determined by the pair $(\mathbf{p}^0, \mathbf{P})$. Since the limit behavior of the Markov chain depends on the structure of the transition matrix the following classification is useful [5]:

Definition 1

A square matrix $\mathbf{A} : n \times n$ is said to be

(a) *nonnegative* ($\mathbf{A} \geq \mathbf{0}$), if $a_{ij} \geq 0$ for all $i, j \in \{1, ..., n\}$,

(b) *positive* ($\mathbf{A} > \mathbf{0}$), if $a_{ij} > 0$ for all $i, j \in \{1, ..., n\}$.

A nonnegative matrix $\mathbf{A} : n \times n$ is said to be

(c) *primitive*, if there exists a $k \in \mathbb{N}$ such that $\mathbf{A}^k$ is positive,

(d) *reducible*, if $\mathbf{A}$ can be brought into the form (with square matrices $\mathbf{C}$ and $\mathbf{T}$)

$$\begin{pmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{R} & \mathbf{T} \end{pmatrix}$$

by applying the same permutations to rows and columns,

(e) *irreducible*, if it is not reducible,

(f) *stochastic*, if $\sum_{j=1}^{n} a_{ij} = 1$ for all $i \in \{1, ..., n\}$.

A stochastic matrix $\mathbf{A} : n \times n$ is said to be

(g) *stable*, if it has identical rows,

(h) *column allowable*, if it has at least one positive entry in each column. $\square$

Note that the product of stochastic matrices is again a stochastic matrix and that every positive matrix is also primitive. The following results set up the foundation of the convergence analysis for GAs given in the next section.

Lemma 1

Let $\mathbf{C}, \mathbf{M}$ and $\mathbf{S}$ be stochastic matrices, where $\mathbf{M}$ is positive and $\mathbf{S}$ is column allowable. Then the product $\mathbf{CMS}$ is positive.

Proof:

Let $\mathbf{A} = \mathbf{CM}$ and $\mathbf{B} = \mathbf{AS}$. Since $\mathbf{C}$ is stochastic there exists at least one positive entry in each row of $\mathbf{C}$. Therefore, by matrix multiplication $a_{ij} = \sum_{k=1}^{n} c_{ik} \cdot m_{kj} > 0$ for all $i, j \in \{1, \ldots, n\}$, i.e., $\mathbf{A} > \mathbf{0}$. Similarly, $b_{ij} = \sum_{k=1}^{n} a_{ik} \cdot s_{kj} > 0$ for all $i, j \in \{1, \ldots, n\}$, because $\mathbf{S}$ is column allowable. $\square$

Theorem 1 ([6, p. 123])

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

Theorem 2 ([6, p. 126])
Let $\mathbf{P}$ be a reducible stochastic matrix, where $\mathbf{C} : m \times m$ is a primitive stochastic matrix and $\mathbf{R}, \mathbf{T} \neq \mathbf{0}$. Then

$$\mathbf{P}^{\infty} = \lim_{k \to \infty} \mathbf{P}^k = \lim_{k \to \infty} \begin{pmatrix} \mathbf{C}^k & \mathbf{0} \\ \sum\limits_{i=0}^{k-1} \mathbf{T}^i \mathbf{R} \mathbf{C}^{k-i} & \mathbf{T}^k \end{pmatrix} = \begin{pmatrix} \mathbf{C}^{\infty} & \mathbf{0} \\ \mathbf{R}_{\infty} & \mathbf{0} \end{pmatrix}$$

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

4 Markov Chain Analysis of Genetic Algorithms

The CGA can be described as a Markov chain: The state of the CGA depends only on the genes of the individuals so that the state space is $\mathcal{S} = \mathbb{B}^N = \mathbb{B}^{l \cdot n}$, where n denotes the population size and l is the number of genes. Each element of the state space can be regarded as an integer number in binary representation. Since this mapping is isomorphic an element $i \in S$ is considered to be in binary or integer representation just as required. 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.
The probabilistic changes of the genes within the population caused by the genetic operators are captured by the transition matrix $\mathbf{P}$, which can be decomposed in a natural way into a product of stochastic matrices $\mathbf{P} = \mathbf{C} \cdot \mathbf{M} \cdot \mathbf{S}$, where $\mathbf{C}$, $\mathbf{M}$ and $\mathbf{S}$ describe the intermediate transitions caused by crossover, mutation and selection, respectively. This leads to:

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.
Proof:
The crossover operator may be regarded as a random total function whose domain and range are $\mathcal{S}$, i.e., each state of $\mathcal{S}$ is mapped probabilistically to another state. Therefore, $\mathbf{C}$ is stochastic. The same holds for the other operators and their transition matrices.
Because the mutation operator is applied independently to each gene/bit in the population, the probability that state i becomes state j after mutation can be aggregated to $m_{ij} = p_m^{H_{ij}} (1 - p_m)^{N - H_{ij}} > 0$ for all $i, j \in \mathcal{S}$, where H_{ij} denotes the Hamming distance between the binary representations of state i and state j. Thus, $\mathbf{M}$ is positive.
The probability that selection does not alter the state generated by mutation can be bounded by

$$s_{ii} \geq \frac{\prod\limits_{k=1}^{n} f(\pi_k(i))}{\left(\sum\limits_{k=1}^{n} f(\pi_k(i)) \right)^n} > 0$$

for all $i \in \mathcal{S}$, so that $\mathbf{S}$ is column allowable.
It follows by Lemma 1 that $\mathbf{P} = \mathbf{C} \cdot \mathbf{M} \cdot \mathbf{S}$ is positive. Since every positive matrix is primitive the proof is completed. □

This result has an alternative representation, also reported in [7, pp. 177–178], albeit derived within a different Markov chain model:

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.
Proof:
Immediate consequence of Theorems 1 and 3. □

Note that the initial distribution $\mathbf{p}^0$ has no effect on the limit behavior of the Markov chain. Therefore, the initialization of the algorithm can be done arbitrarily and the selection operation performed before entering the loop can be omitted from a theoretical point of view: The limit distribution remains the same.
The ergodicity property has consequences for the convergence behavior of the CGA. To avoid confusion, a precise definition of the term *convergence of a GA* is required:

Definition 2
Let $Z_t = \max\{f(\pi_k^{(t)}(i)) \,|\, k = 1, \ldots, n\}$ 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

$$\lim_{t \to \infty} P\{Z_t = f^*\} = 1 \quad , \tag{4}$$

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

This leads to:

Theorem 4
The CGA with parameter ranges as in Theorem 3 does not converge to the global optimum.
Proof:
Let $i \in \mathcal{S}$ be any state with $\max\{f(\pi_k(i)) \,|\, k = 1, \ldots, n\} < f^*$ and p_i^t the probability that the GA is in such a state i at step t. Clearly, $P\{Z_t \neq f^*\} \geq p_i^t \Leftrightarrow P\{Z_t = f^*\} \leq 1 - p_i^t$. From Theorem 1 the probability that the GA is in state i converges to $p_i^\infty > 0$. Consequently,

$$\lim_{t \to \infty} P\{Z_t = f^*\} \leq 1 - p_i^\infty < 1$$

so that condition (4) is not fulfilled. □

One might argue that the above Markov chain does not represent a practical GA because in real–world applications the best solution found over time is always maintained. In fact, after a finite number of transitions the global solution will be visited and copied. This is an immediate consequence of the following:

Theorem 5 ([6, p. 133])
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. □

In order to recognize that this result is in conformity with Definition 2, the Markov chain description has to be adapted by enlarging the population by an additional, say, super individual which does not take part in the evolutionary process. The cardinality of the state space grows from $2^{n \cdot l}$ to $2^{(n+1) \cdot l}$. For notational convenience let the super individual be placed at the leftmost position in the $(n+1)$–tuple and let it be accessible by $\pi_0(i)$ from a population at state i. The transition probabilities of those states containing the same super individual string are assumed to be listed one below the other in the transition matrix and the better the super individual's fitness the higher the position of the corresponding state in the matrix. Since the super individual's string is not affected by the genetic operators, the extended transition matrices for crossover $\mathbf{C}^+$, mutation $\mathbf{M}^+$ and selection $\mathbf{S}^+$ can be written as block diagonal matrices

$$\mathbf{C}^+ = \begin{pmatrix} \mathbf{C} & & & \\ & \mathbf{C} & & \\ & & \ddots & \\ & & & \mathbf{C} \end{pmatrix}, \quad \mathbf{M}^+ = \begin{pmatrix} \mathbf{M} & & & \\ & \mathbf{M} & & \\ & & \ddots & \\ & & & \mathbf{M} \end{pmatrix}, \quad \mathbf{S}^+ = \begin{pmatrix} \mathbf{S} & & & \\ & \mathbf{S} & & \\ & & \ddots & \\ & & & \mathbf{S} \end{pmatrix}$$

with 2^l square matrices $\mathbf{C}$, $\mathbf{M}$ and $\mathbf{S}$ of size $2^{nl} \times 2^{nl}$ possessing the same structure as in the ergodic case so that

$$\mathbf{C}^+ \mathbf{M}^+ \mathbf{S}^+ = \begin{pmatrix} \mathbf{CMS} & & & \\ & \mathbf{CMS} & & \\ & & \ddots & \\ & & & \mathbf{CMS} \end{pmatrix}$$

with $\mathbf{CMS} > \mathbf{0}$. The copy operation is represented by an *upgrade* matrix $\mathbf{U}$ which upgrades an intermediate state containing an individual better than its super individual to a state where the super individual equals the better individual. In particular, let $b = \arg\max\{f(\pi_k(i)) \mid k = 1, \ldots, n\} \in \mathbb{B}^l$ denote the best individual of the population at any state i excluding the super individual. Then $u_{ij} = 1$ if $f(\pi_0(i)) < f(b)$ with $j \stackrel{def}{=} (b, \pi_1(i), \pi_2(i), \ldots, \pi_n(i)) \in \mathcal{S}$, otherwise $u_{ii} = 1$. Thus, there is exactly one entry in each row, which does not hold for the columns because for every state $j \in \mathcal{S}$ with $f(\pi_0(j)) < \max\{f(\pi_k(j) \mid k = 1, \ldots, n\}$ one gets $u_{ij} = 0$ for all $i \in \mathcal{S}$. In other words, a state either becomes upgraded or remains unaltered. Therefore, the structure of the upgrade matrix can be written as

$$\mathbf{U} = \begin{pmatrix} \mathbf{U}_{11} & & & \\ \mathbf{U}_{21} & \mathbf{U}_{22} & & \\ \vdots & \vdots & \ddots & \\ \mathbf{U}_{2^l,1} & \mathbf{U}_{2^l,2} & \ldots & \mathbf{U}_{2^l,2^l} \end{pmatrix}$$

with submatrices U_{ab} of size $2^{nl} \times 2^{nl}$. To ease the presentation assume that problem (1) has only one global optimizer. Then only $\mathbf{U}_{11}$ is a unit matrix whereas all matrices $\mathbf{U}_{aa}$ with $a \geq 2$ are unit matrices with some zero diagonal entries. With $\mathbf{P} = \mathbf{CMS}$ the transition matrix for the GA becomes

$$\begin{aligned} \mathbf{P}^+ &= \begin{pmatrix} \mathbf{P} & & & \\ & \mathbf{P} & & \\ & & \ddots & \\ & & & \mathbf{P} \end{pmatrix} \begin{pmatrix} \mathbf{U}_{11} & & & \\ \mathbf{U}_{21} & \mathbf{U}_{22} & & \\ \vdots & \vdots & \ddots & \\ \mathbf{U}_{2^l,1} & \mathbf{U}_{2^l,2} & \ldots & \mathbf{U}_{2^l,2^l} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{PU}_{11} & & & \\ \mathbf{PU}_{21} & \mathbf{PU}_{22} & & \\ \vdots & \vdots & \ddots & \\ \mathbf{PU}_{2^l,1} & \mathbf{PU}_{2^l,2} & \ldots & \mathbf{PU}_{2^l,2^l} \end{pmatrix} \end{aligned}$$

with $\mathbf{PU}_{11} = \mathbf{P} > \mathbf{0}$. The submatrices $\mathbf{PU}_{a1}$ with $a \geq 2$ may be gathered in a rectangular matrix $\mathbf{R} \neq \mathbf{0}$ so that Theorem 2 may be used to prove that the corresponding GA converges to the global optimum.

THEOREM 6
The canonical GA as in Theorem 3 maintaining the best solution found over time *after* selection converges to the global optimum.
PROOF:
Submatrix $\mathbf{PU}_{11} = \mathbf{P} > \mathbf{0}$ gathers the transition probabilities for states containing a globally optimal super individual (globally optimal states). Since $\mathbf{P}$ is a primitive stochastic matrix and $\mathbf{R} \neq \mathbf{0}$, Theorem 2 guarantees that the probability of staying in any nonglobally optimal state converges to zero. It follows that the probability of being in any globally optimal state converges to one, so that the limit of $P\{Z_t = f^*\}$ converges to one for $t \to \infty$ as well. $\square$

Of course, it is a more sophisticated strategy to copy a better solution directly after the evaluation phase because this solution could be lost after selection. But the purpose of the above analysis was to show the relationship to Theorem 5. Nevertheless, the proof for the case in which a better individual is copied before selection can be adapted straightforwardly:

THEOREM 7
The canonical GA as in Theorem 3 maintaining the best solution found over time *before* selection converges globally optimal.
PROOF:
The transition matrix is now $\mathbf{P}^+ = \mathbf{T}^+\mathbf{U}\mathbf{S}^+$ with $\mathbf{T}^+ = \mathbf{C}^+\mathbf{M}^+$. Setting $\mathbf{T} = \mathbf{CM}$ yields:

$$
\begin{aligned}
\mathbf{P}^+ &= \begin{pmatrix} \mathbf{T} & & & \\ & \mathbf{T} & & \\ & & \ddots & \\ & & & \mathbf{T} \end{pmatrix}
\begin{pmatrix} \mathbf{U}_{11} & & & \\ \mathbf{U}_{21} & \mathbf{U}_{22} & & \\ \vdots & \vdots & \ddots & \\ \mathbf{U}_{2^l,1} & \mathbf{U}_{2^l,2} & \dots & \mathbf{U}_{2^l,2^l} \end{pmatrix}
\begin{pmatrix} \mathbf{S} & & & \\ & \mathbf{S} & & \\ & & \ddots & \\ & & & \mathbf{S} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{TU}_{11}\mathbf{S} & & & \\ \mathbf{TU}_{21}\mathbf{S} & \mathbf{TU}_{22}\mathbf{S} & & \\ \vdots & \vdots & \ddots & \\ \mathbf{TU}_{2^l,1}\mathbf{S} & \mathbf{TU}_{2^l,2}\mathbf{S} & \dots & \mathbf{TU}_{2^l,2^l}\mathbf{S} \end{pmatrix}
\end{aligned}
$$

with $\mathbf{TU}_{11}\mathbf{S} = \mathbf{TS} = \mathbf{P} > \mathbf{0}$. Again, the submatrices $\mathbf{TU}_{a1}\mathbf{S}$ with $a \geq 2$ can be gathered in a rectangular matrix $\mathbf{R} \neq \mathbf{0}$ and with the same argumentation as in Theorem 6, the limit of $P\{Z_t = f^*\}$ converges to one for $t \to \infty$. $\square$

Note that Theorems 6 and 7 do not cover the case of elitist selection. When using elitist selection the best individual is not only maintained but also used to generate new individuals. This algorithm has another transition matrix and therefore different search dynamics, which may be better in some cases and worse in other cases.
Theorem 4 indicates that the schema theorem [1] cannot imply convergence to the global optimum.

5 Discussion of results with respect to the schema theorem

A *schema* 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 $\{0, 1, \texttt{\#}\}$, where the symbol # may be instanciated to either 0 or 1. For example, schema $S = (1\texttt{\#\#}0)$ denotes the subset $S = \{1000, 1010, 1100, 1110\} \subseteq \mathbb{B}^4$. The state space $\mathcal{S} = (\mathbb{B}^l)^n$ represents all possible states of a population $X = (b_1, b_2, \dots, b_n) \in \mathcal{S}$ with n individuals of length l. In the following n–tuple X

is also regarded as a multiset, so that the multiplicity of an element in X is preserved under set operations like intersection. The utility of schema S restricted to multiset X is defined as the average objective function value over all elements contained in $S \cap X$:

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

The schema theorem states that [1, p. 102–103]

$$\mathrm{E}[\,|S \cap X_{t+1}|\,] \geq |S \cap X_t| \, \frac{u(S, X_t)}{u(\mathcal{S}, X_t)} \, (1 - c(S, X_t)) \, (1 - m(S, X_t)) \tag{5}$$

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 modified by crossover and mutation respectively, so that the resulting element is not contained in subset S. Roughly speaking, relation (5) states, 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. Although relation (5) may give some clues about the dynamics of the search, it does not imply convergence to the global optimum. If $|\,\{b^*\} \cap X_t\,| = n$, the schema theorem states that

$$\mathrm{E}[\,|\{b^*\} \cap X_{t+1}|\,] \geq n \, (1 - c(\{b^*\}, X_t)) \, (1 - m(\{b^*\}, X_t)) \quad , \tag{6}$$

which does not indicate that the expectation converges to n, because the bound in (6) is less than n. On the other hand, it does not imply nonconvergence as well. For global convergence it is neccessary and sufficient that $\lim_{t \to \infty} E[I_t] = 1$ which implies $\lim_{t \to \infty} E[\,\{b^*\} \cap X_t|\,] \geq 1$, where process

$$I_t := h(b^*, X_t) := \begin{cases} 1 & , \text{ if } b^* \in \{\pi_1(X_t), \pi_2(X_t), \ldots, \pi_n(X_t)\} \\ 0 & , \text{ otherwise} \end{cases} \tag{7}$$

indicates whether at least one optimal solution b^* is contained in population X_t. In particular:

Lemma 2

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

(b) $\lim_{t \to \infty} \mathrm{E}[I_t] = 1 \Rightarrow \lim_{t \to \infty} \mathrm{E}[\,|\{b^*\} \cap X_t|\,] \geq 1$

Proof:
(a) Note that $\{I_t = 1\} \Leftrightarrow \{Z_t = f^*\}$ and let $1_A(x)$ denote the indicator function. Using the identity $\mathrm{E}[\,1_A\,] = P\{A\}$ for some event A and taking limits on both sides gives the result.
(b) Let $g(b^*, X_t)$ count the number of optimal solutions b^* in population X_t and $h(b^*, X_t)$ as in (7). Since $g(.) \geq h(.)$ one gets

$$\lim_{t \to \infty} \mathrm{E}[\,|\{b^*\} \cap X_t|\,] = \sum_{i=1}^{|\mathcal{S}|} g(b^*, i) \cdot p_i^\infty \geq \sum_{i=1}^{|\mathcal{S}|} h(b^*, i) \cdot p_i^\infty = \lim_{t \to \infty} \mathrm{E}[\,I_t\,] = 1 \quad ,$$

where p^∞ denotes the limit distribution of the Markov chain. $\square$

Note that the converse of Lemma 2(b) is not true in general: Let $\mathcal{S} = \{00, 01, 10, 11\}$ with $g(1, \mathcal{S}) = \{0, 1, 1, 2\}$ and $p^\infty = (0.01, 0.25, 0.25, 0.49)$. Then $\lim_{t\to\infty} \mathrm{E}[\,|\{b^*\} \cap X_t|\,] = 1.48 > 1$ whereas $\lim_{t\to\infty} \mathrm{E}[\,I_t\,] = 0.99 < 1$. That means, even if the expected number of optimal solutions within the population converges to a value greater equal one, global convergence is not guaranteed. The reason is quite clear: 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 (see e.g. [8, p. 201]) that this event will occur with probability one. On the other hand, there is a minimal probability to find again a global solution if it was lost, so that this event will also occur with probability one. Summarizing, the global solution will be lost and found infinitely often, so that the sequence $(|\{b^*\} \cap X_t|)$ is an irreducible markov chain on the state space $\{0, 1, \ldots, n\}$, which does not converge although the expectation does.
The bounds for the probabilities of losing and generating the optimal solution can be derived as follows: Assume that $k' \geq 1$ optimal solutions are contained in population X_t 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. First, consider the case $k \geq 1$. The probability that at least one bit of an optimal solution is flipped is given by

$$p_F := 1 - (1 - p_m)^l > 0 \quad ,$$

so that the probability that all k optimal solutions are destroyed becomes

$$p_F^k \, (1 - p_F)^{n-k} > 0$$

which is bounded below by

$$\gamma_1 := \min \left\{ p_F^n, p_F (1 - p_F)^{n-1} \right\} > 0 \quad .$$

Now consider the case $k = 0$, i.e., all optimal solutions have been destroyed by crossover. The probability that all bits within the population remain unaltered is $(1 - p_m)^{n \cdot l} = (1 - p_F)^n =: \gamma_2 > 0$. Summarizing, the probability that the optimal solution is lost after crossover and mutation is at least

$$p_L = \min\{\gamma_1, \gamma_2\} > 0 \quad .$$

It remains to derive the bound for the probability to generate an optimal solution. Assume that the optimal solution is not contained in the population X_t. Although the crossover operator may assemble some optimal solutions, say $k \geq 1$, the worst case with $k = 0$ is assumed. The probability that mutation generates the optimal solution b^* from individual b_i is given by

$$p_{B_i} := p_m^{H(b_i, b^*)} (1 - p_m)^{l - H(b_i, b^*)} > 0 \quad ,$$

which is bounded below by

$$p_B := \min\{ p_{B_i} \,|\, i = 1, 2, \ldots, n \} > 0 \quad .$$

Thus, the probability that this event occurs at least once is

$$p_G := 1 - (1 - p_B)^n > 0 \quad .$$

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

$$p_{b^*} := f(b^*) / \sum_{j=1}^{n} f(b_j) > 0$$

and the probability that this event occurs at least once becomes

$$p_S := 1 - (1 - p_{b*})^n > 0 \quad .$$

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

6 Conclusions

The analysis given above reveals that convergence to the global optimum is not an inherent property of the CGA but rather is a consequence of the algorithmic trick of keeping track of the best solution found over time. In other words, the original CGA cannot be regarded as an optimization algorithm for static optimization problems (see also [9]) because it is provable that it will not converge to any subset of the set of states containing at least one global solution, even in infinite time. Static optimization, however, was not the original purpose in the design of the CGA [1]. In fact, the interest was concentrated on a strategy which performs an optimal allocation of trials so as to minimize expected losses in an uncertain environment, possibly with time varying rewards — a problem which is certainly not equivalent to a static optimization problem. The schema theorem [1] does not imply that the CGA will converge to the global optimum in static optimization problems. Moreover, due to the irreducibility of the CGA it is clear that the CGA will not converge at all.

As long as the mutation operator is applied in the usual manner, the only chance for making the CGA converge to the global optimum is to modify the selection operator, whose transition matrix must be necessarily not column allowable. Another possible route to globally optimal convergence might be the introduction of time varying mutation *and* selection probabilities, so that the corresponding Markov process becomes inhomogeneous. It has been shown in [10] that the introduction of time varying mutation probabilities alone does not help, which confirms the insight that the selection operator is the key problem of the CGA.

Although convergence to the global optimum as time approaches infinity should be the minimal requirement for a stochastic optimization algorithm, the knowledge of this property is not of practical interest because the solutions of a finite search space can be enumerated in finite time. A more practical question regards the time complexity of the algorithm to achieve the globally optimal solution. The first steps in this direction have been made in [11] and [12] both using simplified but globally optimal convergent GAs.

Acknowledgment

The author would like to thank David B. Fogel and the anonymous referees for helpful comments on the paper.

References

[1] J.H. Holland, *Adaptation in natural and artificial systems*, Ann Arbor: The University of Michigan Press, 1975.

[2] A.E. Eiben, E.H.L. Aarts, and K.M. Van Hee, "Global convergence of genetic algorithms: A markov chain analysis", in *Parallel Problem Solving from Nature*, H.-P. Schwefel and R. Männer (Eds.), Berlin and Heidelberg: Springer, 1991, pp. 4–12.

[3] D.B. Fogel, *Evolving Artificial Intelligence*, PhD dissert., San Diego: University of California, 1992.

[4] D.E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Reading/Mass.: Addison Wesley, 1989.

[5] E. Seneta, *Non–negative Matrices and Markov Chains*, 2nd edition, New York: Springer, 1981.

[6] M. Iosifescu, *Finite Markov Processes and Their Applications*, Chichester: Wiley, 1980.

[7] T.E. Davis and J.C. Principe, "A simulated annealing like convergence theory for the simple genetic algorithm", in *Proceedings of the fourth Conference on Genetic Algorithms*, R.K. Belew and L.B. Booker (Eds.), San Mateo: Morgan Kaufmann, 1991, pp. 174–181.

[8] W. Feller, *An Introduction to Probability Theory and Its Applications Vol. 1*, 3rd (revised) edition, Singapore: Wiley, 1970.

[9] K.A. De Jong, "Are genetic algorithms function optimizers ?", in [13], pp. 3–13.

[10] T.E. Davis, *Toward an extrapolation of the simulated annealing convergence theory onto the simple genetic algorithm*, PhD dissert., Gainesville: University of Florida, 1991.

[11] T. Bäck, "The interaction of mutation rate, selection, and self-adaptation within a genetic algorithm", in [13], pp. 85–94.

[12] H. Mühlenbein, "How genetic algorithms really work I: Mutation and hillclimbing", in [13], pp. 15–25.

[13] R. Männer and B. Manderick (Eds.), *Parallel Problem Solving from Nature, 2*, Amsterdam: North Holland, 1992.