A genetic system based on simulated crossover of sequences of two-bit genes

Marco Carpentieri

Dipartimento di Fisica E. R. Caianiello (DF), Università degli Studi di Salerno (Sa), Italy

Received 24 February 2005; received in revised form 10 August 2006; accepted 13 September 2006

Communicated by Michael David Vose

Abstract

We introduce a genetic model based on simulated crossover of fixed sequences of two-bit genes. Results are

(1) a lower bound on population size is exhibited such that a transition takes the stochastic finite population genetic system near the next state of the deterministic infinite population genetic system (provided both begin in the same state);
(2) states and dynamics of the deterministic infinite population genetic system are derived for arbitrary (finite) fitness functions (expressed in terms of multivariate polynomials);
(3) in the case of quadratic fitness defined by weight matrices with \( m \) nonnull entries it is shown that each state transition can be implemented in time \( O(m + l) \), where \( l \) is the chromosome length;
(4) the genetic algorithm (implementing the proposed infinite population system) is experimentally compared with the infinite population genetic algorithm with bit-based simulated crossover for the max-cut problem; the results show that the extension to sequences of genes with four alleles is useful to improve performances.

© 2006 Elsevier B.V. All rights reserved.

Keywords: Marginal distribution genetic algorithms; Max-cut problem

1. Introduction

Genetic algorithms are probabilistic search algorithms inspired by mechanisms of natural selection and genetics, introduced by John Holland in the 1970s. They have received considerable attention because of their many applications to several research fields such as optimization, adaptive control and others [10,14,15,18]. By simplifying natural laws, genetic algorithms simulate reproductive processes over a population of individuals in an arbitrary environment. Natural selection gives each individual a reproductive chance that depends on his fitness, where the fitness measures the capability of adaptation to the environment and it is usually evaluated by a fitness function. Heritable characters are coded in sequences of special carriers (genes) called chromosomes, typically represented by binary strings of length \( l \).
During the reproduction phase, the chromosomes of the selected individuals (parents) are duplicated and the copies are modified by swapping or reversal of genetic material and/or through (random) mutations. The offspring (new chromosomes) are evaluated to determine their fitness and replace other individuals in the population. The genetic cycle consisting of selection, reproduction, evaluation and replacement of individuals is depicted in Fig. 1.

A classic way to describe the behavior of genetic algorithms is obtained by means of homogeneous Markov chains [8,11,24] whose states encode populations and are multi-sets of binary strings. General theoretical results were introduced for infinite populations by Vose [29,30] who showed how to use them to perform the qualitative analysis of the behavior of the finite populations models. In particular, in [29,30], a dynamical system model is introduced for which simulation is computationally difficult. It is worth noting that the original formalism presented by Vose was intended to model situations in which recombination of genetic material is obtained through the crossover of the chromosomes of two mating parents selected with probability in proportion to their fitness. Thus, the intractability of simulating any general system (such as Vose’s) is due to the fact that it keeps track of every chromosome [31]. To avoid exponential complexity one may consider approximate models, or, alternatively, restrict the type of fitness functions (see for example [25]). A different approach consists in changing one’s mind about what is being modeled, thus, representing some other related genetic system. This last alternative is followed by the present paper (along with restriction of the fitness to some classes of polynomial functions) following the guidelines of Refs. [2,3]; in such papers a model has been proposed in which the recombination of the genetic material is obtained by means of the bit-based simulated crossover operator [26]. This rule, as well as the gene pool recombination introduced by Mühlenbein and Voight [28,23] maintains an infinite population in linkage equilibrium: the genotype frequencies are the product of marginal frequencies. In this context, a frequency vector with exponentially many components can be reconstructed by the vectors of marginal frequencies that in the univariate marginal distribution algorithms [2,3,22,23,28,31] are l-component vectors. Other models based on the marginal frequencies are presented in [1,12,19]. Related work can be also found in [31] in which it is analyzed a recombination-mutation-selection genetic algorithm that uses gene pool recombination. In particular, Wright et al. show that in case of linear fitness functions there is a single stable fixed point for their univariate marginal distribution genetic algorithm. Moreover, readers interested in exact mathematical analysis of simple genetic algorithms and their use as an alternative approach to combinatorial optimization are referred to [25].

In this paper (see also [5]) we address the problem of extending the analysis of univariate marginal distribution genetic algorithms for infinite populations to the bivariate case. We introduce a genetic model based on simulated crossover of fixed sequences of two-bit genes. Such a model represents an instance of the Random Heuristic Search (as defined in [30, Chapter 3]) and can be considered as an extension of the model presented in [26,28,2]; the main characteristic of the system that we shall consider is that the recombination of the genetic material is obtained by performing a weighted average of the alleles along each fixed two-bit locus, using such statistics to produce offspring whose alleles in distinct loci are independently generated. Note that the interest in devising marginal distribution genetic models lies not only in the fact that they consent efficient implementation for infinite populations. Indeed, in case of univariate marginal distributions, they have been used to construct approximation algorithms for hard optimization problems having error.
bounds that can be theoretically estimated. Moreover, the stability analysis of such systems has evidenced an interesting relation with neural networks (in particular with Hopfield’s networks [2,3]). We exhibit the following results:

(1) we prove an exponential lower bound on the concentration probability (viewed as function of the population size \( n \)) stating that the stochastic finite population system can be considered as an approximation of the deterministic infinite population system;

(2) states and dynamics are derived for infinite populations and arbitrary (finite) fitness functions represented in terms of multivariate polynomials;

(3) in the case of quadratic fitness defined by weight matrices with \( m \) nonnull entries it is shown that each state transition can be simulated in time \( O(m + l) \), where \( l \) is the chromosome length;

(4) the genetic algorithm based on the proposed infinite population system is experimentally compared with the infinite population genetic algorithm with bit-based simulated crossover for the max-cut problem; the results show that the extension to sequences of genes with four alleles is useful to improve performances.

2. The genetic system

2.1. The model

We describe the model on which the genetic system is based and introduce the technical formalism useful to define states and dynamics. A population \( P \) of individuals is represented by a multi-set of \( n \in \mathbb{N} \) \( l \)-length binary strings (throughout the paper we suppose \( l \) even) from the set

\[
\Omega = \{0, 1\}^l = \{\omega_1, \ldots, \omega_{2^l}\}.
\]

Each population \( P \) is associated with its frequency vector \( F = (F_{\omega_1}, \ldots, F_{\omega_{2^l}}) \) specifying the proportion of the strings in \( \Omega \) contained in \( P \), where \( F_{\omega_k} = n_k/n \) and \( n_k \) is the number of occurrences of the string \( \omega_k \) in \( P \). Let \( A_n \) denote the set of the frequency vectors that represent populations of \( n \) individuals. Each individual is evaluated by his fitness that is measured by means of a fitness function \( f : \Omega \rightarrow \mathbb{R}^+ \) that associates a positive real value to each chromosome. Throughout the paper, let \( A = \{00, 01, 10, 11\} \) and \( B = A - \{00\} \). The strings in the populations represent chromosomes and each chromosome is divided into a sequence of genes that can assume four distinct forms or alleles. For \( k = 1, \ldots, l/2 \) and \( a = a_1 \cdot a_2 \in A \), consider functions \( \chi_k[a] : \Omega \rightarrow \{0, 1\} \) defined by

\[
\chi_k[a](\omega) = \begin{cases} 
1 & \text{if } a_1, a_2 \text{ are in positions } 2k - 1, 2k, \text{ respectively, of } \omega, \\
0 & \text{otherwise.}
\end{cases}
\]

In the rest of the paper we shall use notation

\[
E_F[X] = \sum_{i=1}^{2^l} X(\omega_i) p_i
\]

to mean the expectation of function \( X : \Omega \rightarrow \mathbb{R} \) considered as a random variable along with the stochastic vector \( P = (p_1, \ldots, p_{2^l}) \).

Starting from an initial population \( P_0 \), if at time \( t \) the state of the (genetic) system is the population \( P \), represented by its frequency vector \( F \), then the population at time \( t + 1 \) is obtained as follows:

(1) for every \( k = 1, \ldots, l/2 \) and \( a \in A \) compute

\[
\phi_{k,F}[a] = \frac{E_F[\chi_k[a]]}{E_F[f]} = \frac{\sum_{i=1}^{2^l} \chi_k[a](\omega_i) f(\omega_i)n_i}{\sum_{i=1}^{2^l} f(\omega_i)n_i};
\]

(2) generate a new population \( P' \) of \( n \) \( l \)-length binary strings, denoted by

\[
P' = \{\omega_{r_1}, \ldots, \omega_{r_n}\},
\]

with probability \( \phi_{k,F}[a] \) of obtaining \( a_1, a_2 \) in positions \( 2k - 1, 2k \) independently from \( r_i \) and \( k \) for \( 1 \leq k \leq l/2 \) and \( 1 \leq i \leq n \).
Let

**Theorem 1.**

Consider the probability space $(\Omega, \Phi^{(n)}(F))$, where

$$\Phi^{(n)}(F)_{\omega_1, \ldots, \omega_n} = \prod_{j=1}^{n} \Phi(F)_{\omega_j}.$$  

Note that (2) represents the probability of generating a new population consisting of $n$ independently selected chromosomes $\omega_1, \ldots, \omega_n$ provided that the stochastic genetic system is in the state $F$. Let $(\omega_{r_1}, \ldots, \omega_{r_n})$ be $n$ chromosomes
selected according to the probability distribution \( \Phi^{(n)}(F) \) and denote by \( F' \) the associated state. We have

\[
\phi_{k,F}[a] = \frac{X_k[a](\omega_1, \ldots, \omega_n)}{X(\omega_1, \ldots, \omega_n)} \quad \text{and} \quad \phi_{k,\Phi[F]}[a] = \frac{E_{\Phi[F]}[X_k[a]]}{E_{\Phi[F]}[f]},
\]

where

\[
X_k[a](\omega_1, \ldots, \omega_n) = \frac{1}{n} \sum_{j=1}^{n} (\zeta_k[a])(\omega_{r_j}) \quad (k = 1, \ldots, q, \ a \in B)
\]

and

\[
X(\omega_1, \ldots, \omega_n) = \frac{1}{n} \sum_{j=1}^{n} f(\omega_{r_j}).
\]

Observe that

\[
E_{\Phi^{(n)}(F)}[X_k[a]] = E_{\Phi[F]}[X_k[a]] \quad \text{and} \quad E_{\Phi^{(n)}(F)}[X] = E_{\Phi[F]}[f].
\]

Moreover, for each \( k = 1, \ldots, q \) and \( a \in B \), one has that

\[
|\phi_{k,F}[a] - \phi_{k,\Phi[F]}[a]| = \left| \frac{X_k[a]}{E_{\Phi[F]}[f]} \right| - \frac{E_{\Phi[F]}[X_k[a]]}{E_{\Phi[F]}[f]}
\]

\[
= \frac{X}{E_{\Phi[F]}[f]} \left| X_k[a] - E_{\Phi[F]}[X_k[a]] \right| \leq \frac{1}{E_{\Phi[F]}[f]} \left( |X - E_{\Phi[F]}[f]| + |X_k[a] - E_{\Phi[F]}[X_k[a]]| \right)
\]

where (3) follows since

\[
0 < \frac{X_k[a]}{X} \leq 1.
\]

Therefore, for \( \varepsilon' > 0 \), we get

\[
P(|\phi_{k,F}[a] - \phi_{k,\Phi[F]}[a]| > \varepsilon') \leq P(|X - E_{\Phi[F]}[f]| + |X_k[a] - E_{\Phi[F]}[X_k[a]]| > \varepsilon' E_{\Phi[F]}[f])
\]

\[
\leq P \left( |X - E_{\Phi[F]}[f]| > \frac{\varepsilon' E_{\Phi[F]}[f]}{2} \right) + P \left( |X_k[a] - E_{\Phi[F]}[X_k[a]]| > \frac{\varepsilon' E_{\Phi[F]}[f]}{2} \right).
\]

(4)

Consequently, by Hoeffding’s inequality \(^1\) [20,27] it follows:

\[
P(|\phi_{k,F}[a] - \phi_{k,\Phi[F]}[a]| > \varepsilon') \leq 4e^{-\frac{2(\varepsilon' E_{\Phi[F]}[f]/M)^2}{2}}
\]

(5)

for each \( k = 1, \ldots, q \) and \( a \in B \). We can conclude that

\[
P(||\phi_{F^*} - \phi_{\Phi[F]}|| \leq \varepsilon) = P \left( \sqrt{\sum_{k=1}^{q} \sum_{a \in B} (\phi_{k,F^*} - \phi_{k,\Phi[F]})^2} \leq \varepsilon \right)
\]

\[
= 1 - P \left( \sqrt{\sum_{k=1}^{q} \sum_{a \in B} (\phi_{k,F^*} - \phi_{k,\Phi[F]})^2} > \varepsilon^2 \right)
\]

\[
\geq 1 - P \left( \frac{1}{3q} \sum_{k=1}^{q} \sum_{a \in B} (\phi_{k,F^*} - \phi_{k,\Phi[F]})^2 > \varepsilon^2 \right)
\]

\[
\geq 1 - \frac{q}{\sqrt{3q}} \sum_{k=1}^{q} \sum_{a \in B} P \left( |\phi_{k,F^*} - \phi_{k,\Phi[F]}| > \frac{\varepsilon}{\sqrt{3q}} \right)
\]

\(^1\) Let \( X_1, \ldots, X_n \) be independent bounded random variables such that \( X_i \) falls in the interval \([a_i, b_i]\) with probability one and let \( S_n = \sum_{i=1}^{n} X_i \) with expectation \( E[S_n] \); then, for any \( z > 0 \) we have that \( P(S_n - E[S_n] \geq z) \leq e^{-z^2/2\sum_{i=1}^{n}(b_i-a_i)^2} \) and \( P(S_n - E[S_n] \leq -z) \leq e^{-z^2/2\sum_{i=1}^{n}(b_i-a_i)^2} \).
\[ \geq 1 - 6e^{-(n/2)(\varepsilon E_{\Phi[F]}[f]/\sqrt{M})^2} \quad \text{(by (5))} \]
\[ = 1 - 6e^{-(n/3l)(\varepsilon E_{\Phi[F]}[f]/M)^2} . \]

Thus, setting
\[ \delta = 6e^{-(n/3l)(\varepsilon E_{\Phi[F]}[f]/M)^2} \]
(1) holds for every
\[ n \geq 3l \left( \frac{M}{\varepsilon E_{\Phi[F]}[f]} \right)^2 \log \left( \frac{6l}{\delta} \right) . \]

2.3. States and dynamics

2.3.1. Fitness functions

First of all, we briefly discuss the topic of efficient implementation in case of arbitrary finite fitness functions \( f : \Omega \to \mathbb{R}^+ \) and for infinite populations. For \( z \in B \) and \( k = 1, \ldots, q \), one has
\[ \phi_k(z) = \frac{E_{\Phi[F]}[z]f}{E_{\Phi[F]}[f]} = \frac{\sum_{j=1}^{2^l} \phi_k(z)(\omega_j)f(\omega_j)z_{2^l-1} \cdots z_1}{\sum_{j=1}^{2^l} f(\omega_j)z_{2^l-1} \cdots z_1} . \]

By (6) we observe that efficient implementation depends, not only on the dimension of the involved states, but also on the type of fitness function. In this regard, it is well known that \( f \) can be expressed in terms of multivariate polynomials and this fact can be (naturally) used to characterize classes of functions for which we are able to perform efficient implementation in the sense that state transitions can be computed in time polynomial in \( l \). Let, now,
\[ Pf(x_1, \ldots, x_l) = \sum_{y_1, \ldots, y_l \in \{0, 1\}} w_{y_1, \ldots, y_l} x_1^{y_1} \cdots x_l^{y_l} = 1 - x_i, x_i^{(0)} = x_i \quad \text{for} \quad i = 1, \ldots, l \]
be a multivariate polynomial defined on \([0, 1]^l\) and coincident with \( f \) on \( \Omega \). Notice that since \( Pf \) is a polynomial of degree at most one in each variable its global maximum is on elements in \( \Omega \).

Denote, for \( u = 1, \ldots, q \) and \( a \in A \), by \( \hat{x}_u[a] \) the product of \( x_{2u-1}^{(a_1)} x_{2u}^{(a_2)} \), that is
\[ \hat{x}_u[a] = x_{2u-1}^{(a_1)} x_{2u}^{(a_2)} . \]

The identity
\[ 1 = \sum_{a \in A} \hat{x}_u[a] \]
implies
\[ Pf(x) = \sum_{a \in A} \hat{x}_u[a] Pf(x) \quad (x = (x_1, \ldots, x_l)) \]
and therefore
\[ Pf(x) = \sum_{a \in A} \hat{x}_u[a] b_{k,a}(x) , \]
where
\[ b_{k,z}(x) = Pf(x)|_{x_{2^l-1}=z_1, x_{2^l}=z_2} \quad (z \in A) . \]

(8)
This can be used recursively to express (7) in terms of \( \hat{x}_q[a] \) (0 \( \leq u \leq q \), \( a \in A \)). Thus, the polynomial can be put in the form

\[
Pf(\hat{x}) = \sum_{a \in A} \hat{x}_k[a]b_{k,a}(\hat{x}),
\]

where \( \hat{x} = (\hat{x}_0[00], \hat{x}_1[01], \hat{x}_1[10], \hat{x}_1[11], \ldots, \hat{x}_q[00], \hat{x}_q[01], \hat{x}_q[10], \hat{x}_q[11]) \) and \( b_{k,a}(\hat{x}) \) does not depend on the variables \( x_{2k-1}, x_{2k} \) for every \( a \in A \) and \( k = 1, \ldots, q \). Calculating the expectations, being \( \phi_{k,F}[00] = 1 - \sum_{a \in B} \phi_{k,a}[a] \) for each \( k = 1, \ldots, q \), we get

\[
E_{\Phi(F)}[Pf] = \sum_{a \in A} E_{\Phi(F)}[\hat{x}_k[a]]E_{\Phi(F)}[b_{k,a}] = \sum_{a \in A} \phi_{k,F}[a]b_{k,a}(\phi_F) = Pf(\phi_F)
\]

by linearity of the mean and by independence. Moreover, since

\[
E_{\Phi(F)}[\hat{x}_k[z]\hat{x}_k[a]] = \begin{cases} \phi_{k,F}[z] & \text{if } z = a, \\ 0 & \text{otherwise} \end{cases}
\]

for \((a, z) \in A \), one has for \( k = 1, \ldots, q \) and \( z \in B \) that

\[
E_{\Phi(F)}[\hat{x}_k[z]Pf] = \sum_{a \in A} E_{\Phi(F)}[\hat{x}_k[z]\hat{x}_k[a]]E_{\Phi(F)}[b_{k,a}] = \phi_{k,F}[z]b_{k,z}(\phi_F).
\]

Therefore, in case of infinite populations, Theorem 1 implies that as \( n \to \infty \) the stochastic genetic system converges to the deterministic infinite population system; the states of such a deterministic system are \( 3q \)-component vectors

\[
\psi = (\psi_{1,01}, \psi_{1,10}, \psi_{1,11}, \psi_{2,01}, \psi_{2,10}, \psi_{2,11}, \ldots, \psi_{q,01}, \psi_{q,10}, \psi_{q,11}) \in [0,1]^{3q}
\]

with \( \sum_{a \in B} \psi_{k,a} \leq 1 \) (1 \( \leq k \leq q \)) and the dynamics is described, for \( z \in B \) and \( k = 1, \ldots, q \), by the equations

\[
\psi_{k,z}(t + 1) = \frac{\psi_{k,z}(t)b_{k,z}(\psi(t))}{Pf(\psi(t))},
\]

where

\[
\psi(t) = (\psi_{1,01}(t), \psi_{1,10}(t), \psi_{1,11}(t), \ldots, \psi_{q,01}(t), \psi_{q,10}(t), \psi_{q,11}(t)).
\]

Note that \( \psi_{k,z}(t) \) in Eq. (11) represents the probability of having \( z \) as the \( k \)th allele after \( t \) transitions. The state space of the iterative deterministic genetic system for infinite populations is a subset \( A(3q) \subset [0,1]^{3q} \) of \( 3q \)-component vectors \( \psi \) such that

\[
\sum_{a \in B} \psi_{k,a}(t) \leq 1 \quad (1 \leq k \leq q).
\]

Moreover, by (11) it is clear that, to be able to perform (state transition) efficient implementation, the terms \( b_{k,z}(\psi(t)) \) for all \( k = 1, \ldots, q \) and \( z \in B \) must be computed in time polynomial in the chromosome length \( l \). In such class of functions there are important types of fitness functions such as quadratic ones that are useful to model hard optimization problems; a more general class consists of the functions than can be expressed as sums of monomials (products) of at most \( O(\log^c l) \) variables, where \( c > 0 \) is constant. Let \( N(3q) = A(3q) \cap [0,1]^{3q} \), that is, the subset of \( A(3q) \) composed by the vectors in \( A(3q) \) with \( 3q \)-components each assuming values in \([0,1]\). Intuitively, by definition of \( \psi_{k,z}(t) \), all points in \( N(3q) \) are fixed points of system (11) as stated by the following lemma.

**Lemma 1.** Let \( x \in N(3q) \) and \( \Delta \psi_{k,z}(t) = \psi_{k,z}(t + 1) - \psi_{k,z}(t) \); then it holds that

\[
\Delta \psi_{k,z}(t)_{\psi(t) = x} = 0 \quad \text{for each } k = 1, \ldots, q \text{ and } z \in B.
\]

**Proof.** By Eq. (11) it follows that

\[
\Delta \psi_{k,z}(t) = \psi_{k,z}(t) \left( \frac{b_{k,z}(\psi(t))}{Pf(\psi(t))} - 1 \right)
\]

(12)
The replacement stated by rules (14) can be easily understood suitably rewriting fitness functions (for example a second-order Taylor’s expansion). In this context, an important case we shall discuss is that of quadratic case in which it is difficult to put the fitness in the form in (9), a solution consists in using a suitable approximation and if \( \psi_{k,z}(t) = 0 \) the right-hand side of (12) is zero for each \( k = 1, \ldots, q \) and \( z \in B \). In the case \( \psi_{k,z}(t) = 1 \), since exactly one of \( \psi_{k,a}(t) | \psi(t) = x \) for \( a \in A \) is nonzero, we get \( Pf(\overline{\psi(t)}) = b_{k,z}(\overline{\psi(t)}) \) and the result follows again by (12).

### 2.3.2. Quadratic fitness

In practical applications the fitness is not, in general, expressed as in Eq. (9). Thus, there exists the problem of determining the polynomial terms \( b_{k,z}(\psi(t)) \) to implement the dynamics of the infinite population system. In the case in which it is difficult to put the fitness in the form in (9), a solution consists in using a suitable approximation (for example a second-order Taylor’s expansion). In this context, an important case we shall discuss is that of quadratic fitness functions \( f: \Omega \to \mathbb{R}^+ \) coincident with \( f \) on \( \Omega \) defined on \([0, 1]^l\) by

\[
Qf(x_1, \ldots, x_l) = C + \sum_{i=1}^{l} w_i x_i + \sum_{i,j=1 \atop i \neq j}^{l} w_{i,j} x_i x_j.
\]

Of course, such type of fitness is a special case of multivariate polynomial fitness, but it is of particular interest since it is used to model the dynamics of systems in several research fields [17] and consents to provide a mathematical formalism to deal with hard optimization problems [16]. The transition from \( Pf(x) \) to \( Pf(\overline{\psi(t)}) \) may be accomplished by replacing the indeterminants \( \hat{x}_u[a] \) in \( Pf(x) \) with \( \psi_{u,a}(t) \) (after \( \hat{x}_u[0] \) has been macro-expanded to \( 1 - \sum_{a'\in B} \hat{x}_u[a'] \)) for all \( u = 1, \ldots, q \) and \( a \in B \). Moreover, \( b_{k,z}(\psi(t)) \) can be similarly obtained from \( b_{k,z}(x) \); obviously, such conversion ignores the question of efficient computation (with respect to time or space) of \( Pf(\overline{\psi(t)}) \) and \( b_{k,z}(\overline{\psi(t)}) \) \((1 \leq u \leq q, a \in B)\).

In the case of fitness functions of the form (13) a basic rule to obtain \( Qf(\overline{\psi(t)}) \) from \( Qf(x) \) (or \( b_{k,z}(\overline{\psi(t)}) \) from \( b_{k,z}(x) \)) consists simply in performing the following replacement in the following (top–down) order

\[
\begin{align*}
x_{2u-1} x_{2u} &\to \psi_{u,11} \quad \text{for all } 0 < u \leq q, \\
x_{2u-1} &\to (\psi_{u,10} + \psi_{u,11}) \quad \text{for all } 0 < u \leq q, \\
x_{2u} &\to (\psi_{u,01} + \psi_{u,11}) \quad \text{for all } 0 < u \leq q.
\end{align*}
\]

The replacement stated by rules (14) can be easily understood suitably rewriting \( Qf(x) \) (for example), computing the mean \( E_{\Phi_{\{1\}}[Qf]} \) and is justified by the properties of linearity (of \( E \)) and independence. First, note that

\[
Qf(x_1, \ldots, x_l) = C + \sum_{u=1}^{q} \sum_{i=1}^{2} w_{2(u-1)+i} x_{2(u-1)+i}
\]

\[
+ \sum_{u=1}^{q} \sum_{i,j=1 \atop i \neq j}^{2} w_{2(u-1)+i,2(u-1)+j} x_{2(u-1)+i} x_{2(u-1)+j}
\]

\[
+ \sum_{u,u'=1}^{q} \sum_{i=1}^{2} w_{2(u-1)+i,2(u'-1)+j} x_{2(u-1)+i} x_{2(u'-1)+j},
\]

where the last sum is nonzero only when \( q > 2 \); consequently applying rules (14) it follows that

\[
Qf(\overline{\psi(t)}) = C + \sum_{u=1}^{q} \sum_{i=1}^{2} w_{2(u-1)+i} \sum_{a_1=1}^{2} \psi_{u,a}(t)
\]

\[
+ \sum_{u=1}^{q} \sum_{i,j=1 \atop i \neq j}^{2} w_{2(u-1)+i,2(u-1)+j} \sum_{a_1=1}^{2} \psi_{u,a}(t) \sum_{a_2=1}^{2} \psi_{u',a'}(t).
\]

(15)
Moreover, we can obtain the terms \( b_{k,z}(\psi(t)) \) for \( k = 1, \ldots, q \) rewriting \( Q_f \) as follows:

\[
Q_f(x_1, \ldots, x_l) = C + \sum_{i = 1}^{2} w_{2(k-1)+i}X_2(k-1)+i + \sum_{u = 1}^{q} \sum_{i = 1}^{2} w_{2(u-1)+i}X_2(u-1)+i
\]

\[
+ \sum_{i,j=1\atop i \neq j}^{2} w_{2(k-1)+i,2(k-1)+j}X_2(k-1)+iX_2(k-1)+j
\]

\[
+ \sum_{u = 1}^{q} \sum_{i,j=1\atop i \neq j}^{2} w_{2(u-1)+i,2(u-1)+j}X_2(u-1)+iX_2(u-1)+j
\]

\[
+ \sum_{u = 1}^{q} w_{2(k-1)+i,2(k-1)+j}X_2(k-1)+iX_2(u-1)+j
\]

\[
+ \sum_{u = 1}^{q} \sum_{i,j=1}^{2} w_{2(u-1)+i,2(u-1)+j}X_2(u-1)+iX_2(u-1)+j
\]

\[
+ \sum_{u = 1}^{q} \sum_{i,j=1}^{2} w_{2(u-1)+i,2(u-1)+j}X_2(u-1)+iX_2(u-1)+j
\]

\[
+ \sum_{u = 1}^{q} \sum_{i,j=1}^{2} w_{2(u-1)+i,2(u-1)+j}X_2(u-1)+iX_2(u-1)+j
\]

thus, by (8) and (14) for each \( k = 1, \ldots, q \) and \( z \in B \) we get

\[
b_{k,z}(\psi(t)) = C + \sum_{i = 1}^{2} w_{2(k-1)+i}z_i + \sum_{u = 1}^{q} \sum_{i = 1}^{2} w_{2(u-1)+i}z_i \sum_{a \in B : a_i = 1} \psi_{u,a}(t)
\]

\[
+ \sum_{i,j=1\atop i \neq j}^{2} w_{2(k-1)+i,2(k-1)+j}z_i z_j + \sum_{u = 1}^{q} \sum_{i,j=1\atop i \neq j}^{2} w_{2(u-1)+i,2(u-1)+j}z_i z_j \sum_{a \in B : a_i = 1} \psi_{u,a}(t)
\]

\[
+ \sum_{i = 1}^{2} z_i \sum_{u = 1}^{q} \sum_{i,j=1\atop i \neq j}^{2} \left( w_{2(k-1)+i,2(u-1)+j} + w_{2(u-1)+j,2(k-1)+i} \right) \sum_{a \in B : a_i = 1} \psi_{u,a}(t)
\]

\[
+ \sum_{u = 1}^{q} \sum_{i,j=1\atop i \neq j}^{2} w_{2(u-1)+i,2(u-1)+j} \sum_{a \in B : a_i = 1} \psi_{u,a}(t) \sum_{a' \in B : a'_j = 1} \psi_{u,a'}(t).
\]

Since state-transitions become increasingly expensive as the length \( l \) of the chromosomes increases, it is important to realize quick transitions in implementing the dynamics of the genetic system. In this regard, note that, suitably representing (the weighted graph [7] with adjacency matrix) \([w_{i,j}]\) by adjacency lists, we are able to compute (15) and

\[
Q_k,z = \sum_{u = 1}^{q} \sum_{i,j=1}^{2} \left( w_{2(k-1)+i,2(u-1)+j} + w_{2(u-1)+j,2(k-1)+i} \right) \sum_{a \in B : a_i = 1} \psi_{k,a}(t) \sum_{a' \in B : a'_j = 1} \psi_{k,a'}(t)
\]

for \( k = 1, \ldots, q \) and \( z \in B \) in time \( O(|E| + l) \), where \( E \) is the set

\[
E = \{(i, j) : w_{i,j} \neq 0 \text{ for } i, j = 1, \ldots, l\}.
\]

Thus, since it holds that \( b_{k,z}(\psi(t)) = (b_{k,z}(\psi(t)) - Q_f(\psi(t))) + Q_f(\psi(t)) \), where

\[
b_{k,z}(\psi(t)) - Q_f(\psi(t)) = \sum_{i = 1}^{2} w_{2(k-1)+i}z_i + \sum_{i,j=1\atop i \neq j}^{2} w_{2(k-1)+i,2(k-1)+j}z_i z_j
\]

\[
- \sum_{i = 1}^{2} w_{2(k-1)+i} \sum_{a \in B : a_i = 1} \psi_{k,a}(t) - \sum_{i,j=1\atop i \neq j}^{2} w_{2(k-1)+i,2(k-1)+j} \sum_{a \in B : a_i = a_j = 1} \psi_{k,a}(t) + Q_k,z
\]

for \( k = 1, \ldots, q \) and \( z \in B \), a state transition may be implemented in time \( O(|E| + l) \).
3. Application to the max-cut problem

The topic of designing approximation algorithms based on infinite population genetic models (with simulated crossover of one-bit genes) has been studied in [2–4] for hard problems. In summary, the available main theoretical results about such models evidence that the fitness function becomes a Lyapunov function for asynchronous variants of the corresponding iterative dynamical systems, where by asynchronous we mean that the components of the state vector are updated one at a time in a predefined order. As a consequence of this property, it is conceivable to design univariate marginal distribution genetic algorithms that can be used as approximation algorithms to solve combinatorial optimization problems. Other results connecting such genetic systems with Hopfield’s networks (well known local optimizers on which some approximation algorithms are based) can be found in [2,3]. The usual way of using the infinite population systems to get local optimization of the fitness is to initialize with (slightly perturbed) equally likely marginal probability distributions for each allele (see also [2,3] and in particular [4]). In the experiments, the synchronous variants of the considered systems have exhibited convergence and optimization properties very similar to those shown by the asynchronous systems. Following the guidelines of Refs. [2–4], system (11) has been implemented to solve (in the sense of an approximation algorithm) the max-cut problem, namely, that of partitioning the vertices of an arbitrary graph \( G \) in two other subsets \( V_1 \) and \( V_2 \) in such a way that the number \( \omega(V_1) \) of edges with one endpoint in \( V_1 \) and the other in \( V_2 \) is maximized. We remind that the decision version of the max-cut problem is NP-complete [16]. In the genetic algorithm, we have considered the quadratic fitness \( f : \Omega \rightarrow \mathbb{N}^+ \) defined by

\[
Qf(x_1, \ldots, x_l) = \sum_{\substack{i,j=1 \atop i \neq j}}^l w_{i,j} x_i (1 - x_j),
\]

where the weights \( w_{i,j} \) are set \( w_{i,j} = 1 \) if the input graph \( G \) has edge \( \{i, j\} \) and \( w_{i,j} = 0 \) otherwise. We remark that initial equally likely marginal probability distributions for each allele are fixed points, in case of undirected graphs, for the system based on recombination of sequences of one-bit genes; however, in [2,3] it has been shown that such points are not asymptotically stable. Conversely, for the system based on simulated crossover of sequences of two-bit genes, equally likely marginal distributions are not fixed points, but they initialize trajectories that do not converge towards states in \( \Omega \). Nevertheless, the initial slight perturbations \( \psi_{k,z}(0) = \frac{1}{4} \pm \gamma \), where \( \gamma \approx 10^{-2} \), for \( k = 1, \ldots, q \) and \( z \in B \) consent to solve this problem. In our experiments, the genetic algorithm with simulated crossover of two-bit genes showed convergence characteristics very similar to those held by the univariate marginal distributions genetic algorithm. In Table 1 there are the experimental results intended to compare the performances of the algorithm 2BGSC based on simulated crossover of sequences of two-bit genes with the univariate 1BGSC based on recombination of sequences of binary genes introduced in [2]. In the table it is reported the mean size of the cuts found by the two algorithms for \( p \)-random graphs with \( p = \frac{1}{4} \) and \( p = \frac{1}{3} \). In the table it is also reported the expected number of edges (row edges) of the \( p \)-random graphs. The algorithm 2BGSC consents to improve (slightly and in average) the performances of the algorithm 1BGSC that in [2], in case of \( p \)-random graphs for \( p = \frac{1}{4}, \frac{1}{3} \), exhibited performances slightly worse than those of an Hopfield’s network [17]. In Table 1 there are also the mean sizes of the cuts obtained

<table>
<thead>
<tr>
<th>Vertices</th>
<th>34</th>
<th>38</th>
<th>42</th>
<th>46</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p = \frac{1}{4} )</td>
<td>1BGSC</td>
<td>62.25</td>
<td>76.40</td>
<td>92.65</td>
<td>108.25</td>
</tr>
<tr>
<td></td>
<td>2BGSC</td>
<td>63.35</td>
<td>76.70</td>
<td>93.10</td>
<td>108.70</td>
</tr>
<tr>
<td></td>
<td>12BGSC</td>
<td>63.70</td>
<td>77.50</td>
<td>94.05</td>
<td>109.55</td>
</tr>
<tr>
<td></td>
<td>Edges</td>
<td>80.14</td>
<td>100.42</td>
<td>123.00</td>
<td>147.85</td>
</tr>
<tr>
<td>( p = \frac{1}{3} )</td>
<td>1BGSC</td>
<td>96.65</td>
<td>119.75</td>
<td>145.30</td>
<td>176.05</td>
</tr>
<tr>
<td></td>
<td>2BGSC</td>
<td>97.15</td>
<td>120.60</td>
<td>146.60</td>
<td>177.05</td>
</tr>
<tr>
<td></td>
<td>12BGSC</td>
<td>97.95</td>
<td>121.40</td>
<td>147.2</td>
<td>178.45</td>
</tr>
<tr>
<td></td>
<td>Edges</td>
<td>140.25</td>
<td>175.75</td>
<td>215.25</td>
<td>258.75</td>
</tr>
</tbody>
</table>
by choosing the best ones, for a same $p$-random graph, found by the two algorithms (rows 12BGSC). From such results, we notice that the performances were dependent on specific generated $p$-random graphs; this fact suggests that mutations (dependent on the input instances) in positions and/or lengths of genes could be helpful to improve results and speed convergence. In this regard, even if our models are merely computational, such mutations could be explained as having the precise aim of adaptation to environment modifications that, in case of evolutionary processes, are very slow. As a concluding remark note that the experimental results have to be read also by considering that the two-bit genes chromosomes structure is very simple and there is no dependence between specific properties of the input graphs and the gene codification.

4. Conclusions

In this paper, we have addressed the problem of extending the analysis of univariate marginal distribution genetic algorithms for infinite populations to the bivariate case. We have introduced a genetic model based on simulated crossover of fixed sequences of two-bit genes. States and dynamics have been derived for infinite populations and finite fitness functions. The choice of the bivariate model is for sake of conciseness and simplicity. Devising (other multivariate) models is the first step to provide more general infinite population genetic systems that can be efficiently implemented. The results exhibited in this paper can quite straightforwardly be translated into the multivariate framework. However, further analysis is required to better understand how to provide a suitable more general model for marginal distribution genetic algorithms. An important topic in this scenario seems to be constituted by the possibility of extending the relation found between (infinite population) univariate marginal distribution genetic systems and discrete Hopfield’s networks. This topic is not considered here and is referred to further research. However, based on the results obtained in this paper and on a preliminary recent analysis, we conjecture that, even if with some differences, the results obtained in [2–4] can be generalized to the multivariate framework. Such extension is considered to be of particular relevance to answer to some crucial questions such as the meaningfulness of genes codification, how to improve performances and, in particular, the design of new models that could be more convenient to try to solve hard optimization problems (in this regard the reader is also referred to the area of classic approximation algorithms with special attention to the Goemans and Williamson results [9]).

Acknowledgment

This and the papers [2–4] are based on ideas that the author developed at the beginning of the 1990s while he was at the CNR institute in Naples and that were formalized later on during his Ph.D. at the Milan University. A special thank is due to the anonymous referees, to the handling editor for their helpful suggestions and comments and to the journal manager Mick van Gijlswijk. This paper is in memory of Prof. Mario Carpentieri and is dedicated to my friends Paolo, Angelo, Donato, to Marianna and all honest people.

References