ELSEVIER

Theoretical Computer Science 259 (2001) 1–61

Theoretical Computer Science

www.elsevier.com/locate/tcs

Fundamental Study

Theory of genetic algorithms

Lothar M. Schmitt

School of Computer Science and Engineering, The University of Aizu, Aizu-Wakamatsu City, Fukushima Prefecture, 965-8580, Japan

Received April 1999; revised September 2000; accepted October 2000
Communicated by M. Ito

This work is dedicated to Axel T. Schreiner in Osnabrück

Abstract

(i) We investigate spectral and geometric properties of the mutation-crossover operator in a genetic algorithm with general-size alphabet. By computing spectral estimates, we show how the crossover operator enhances the averaging procedure of the mutation operator in the random generator phase of the genetic algorithm. By mapping our model to the multi-set model often investigated in the literature, we compute corresponding spectral estimates for mutation-crossover in the multi-set model.

(ii) Various types of unscaled or scaled fitness selection mechanisms are considered such as proportional fitness selection, rank selection, and tournament fitness selection. We allow fitness selection mechanisms where the fitness of an individual or creature depends upon the population it resides in. We investigate contracting properties of these fitness selection mechanisms and combine them with the crossover operator to obtain a model for genetic drift. This has applications to the study of genetic algorithms with zero or extremely low mutation rate.

(iii) We discuss a variety of convergent simulated-annealing-type algorithms with mutation-crossover as generator matrix.

(iv) The theory includes proof of strong ergodicity for various types of scaled genetic algorithms using common fitness selection methods. If the mutation rate converges to a positive value, and the other operators of the genetic algorithm converge, then the limit probability distribution over populations is fully positive at uniform populations whose members have not necessarily optimal fitness.

(v) In what follows, suppose the mutation rate converges to zero sufficiently slow to assure weak ergodicity of the inhomogeneous Markov chain describing the genetic algorithm, unbounded power-law scaling for the fitness selection is used, mutation and crossover commute, and the fitness function is injective which is a minor restriction in regard to function optimization.

(v_a) If a certain integrable convergence condition is satisfied such that the selection pressure increases fast, then there is essentially no other restriction on the crossover operation, and the algorithm asymptotically behaves as the following take-the-best search algorithm: (1) mutate

E-mail address: lothar@u-aizu.ac.jp (L.M. Schmitt).

PII: S0304-3975(00)00406-0

in every step with rate decreasing to zero, and (2) map any population to the uniform population with the best creature. The take-the-best search algorithm is investigated, and its convergence is shown. Depending upon population-size, the take-the-best search algorithm does or does not necessarily converge to the optimal solution.

(v_b) If population size is larger than length of genome, and a certain logarithmic convergence condition is satisfied such that the selection pressure increases slowly but sufficiently fast, then the algorithm asymptotically converges to the optimal solution.

Keywords: Non-binary genetic algorithms; Spectral analysis of mutation-crossover; Convergent simulated annealing-type genetic algorithm; Genetic drift; Scaled proportional fitness selection; Convergence in the zero-mutation limit case

Contents

1. Introduction

The purpose of this exposition is to give a comprehensive theory of genetic algorithms that use a general-size alphabet with the following major goals:

- Establish mutation as the main thriving force in the random generator phase of a genetic algorithm which assures ergodicity of the (inhomogeneous) Markov chain describing the algorithm. This point of view is, in particular, in accordance with experimental results as in [7]. See Theorems 4.1, 4.2, Lemmas 4.3, 4.4, the discussion in Section 5.1, Theorems 6.1 and 8.3.
- Put the crossover operation in proper perspective by showing:
  - The crossover operation *assists* mutation in the random generator phase of the algorithm by adding to the contraction process over the state space S. See Theorem 6.1.
  - The crossover operation combined with fitness selection under no mutation shows a convergence effect for the algorithm, called genetic drift, which is *not* ergodic in nature, i.e., depends upon the initial population. See Section 7.5.
  - The crossover operation does *not* play a significant role in the asymptotic behavior of genetic algorithms, if the mutation rate stays positive and fitness scalings such as unbounded power-law scaling are used. See Theorem 8.3.
- Discuss a variety of *convergent* simulated annealing-type algorithms with mutation-crossover as generator matrix, including algorithms that allow dynamic bursts of mutation and crossover rate. See Section 6.2.
- Show that genetic algorithms using the most widely used fitness evaluation methods do *not* necessarily converge to populations that contain only optimal creatures. See Theorems 8.1, 8.2, 8.5 and Section 8.3.
- Show convergence to optima for genetic algorithms using unbounded power-law scaling. (The main result in [65, 66] is not correct, cf., Section 8.3.) See Theorems 8.5 and 8.6.

Genetic algorithms, introduced by Holland in [33], have been used by many researchers as a tool for search and optimization. A given optimization task is encoded in such a way that instances such as a path in a weighted graph are understood as elements in a finite collection $\mathscr{C}$ of creatures (candidate solutions) in

a model "world", and a fitness function $f:\mathscr{C}\to\mathbf{R}^+$ exists, which has to be maximized. Usually, the number of elements in $\mathscr{C}$ is very large prohibiting a complete search of $\mathscr{C}$. Genetic algorithms provide a probabilistic way to conduct a search in $\mathscr{C}$ for arbitrary f given a suitable encoding of creatures or instances into strings of symbols.

A genetic algorithm comprises three phases (operations): mutation, crossover and fitness selection. These are applied cyclically and iteratively to fixed-size, finite populations consisting of elements of $\mathscr{C}$ until a saturation condition, or another boundary condition is satisfied. A genetic algorithm is called simple, if all three operations stay constant over the course of the algorithm. Mutation models random change in the genetic information of creatures, and is inspired by random change of genetic information in living organisms, e.g., through the effects of radiation or chemical mismatch. Crossover models the exchange of genetic information of creatures, and is inspired by exchange of genetic information in living organisms, e.g., during the process of sexual reproduction. Fitness selection models reproductive success of adapted organisms in their environment.

Theoretical investigations of genetic algorithms have previously been undertaken by many authors (see, e.g., [14–16, 25, 29, 34, 50, 54, 57, 61, 62, 65, 66, 68–70]). The model most commonly investigated is the genetic algorithm with a binary alphabet, multiple-bit mutation, one-point crossover, and proportional fitness selection. These genetic algorithms are modeled almost exclusively – with the exception of work by Rudolph [57], or Fujii, Nehaniv and the author [61, 62] – by (inhomogeneous) Markov chains acting on probability distributions over populations which are understood as unordered multi-sets. This may have some advantages such as a smaller state space, but the price to pay is a non-negligible combinatorial burden. In the work of Davis and Principe [14–16], the main point of consideration is whether or not annealing the mutation rate to zero in the simple genetic algorithm implies convergence to global optima. It does not. However, Davis and Principe established strong ergodicity of the resulting inhomogeneous Markov chain describing the algorithm. The results established below, in particular Theorem 8.5 and the examples in Section 8.3, generalize the results by Davis and Principe [14–16] to genetic algorithms using scaled fitness selection. Notably another comprehensive model for the simple genetic algorithm based upon Markov chain analysis can be found in the work of Vose, Liepins, and Nix [50, 68–70], which is the basis for many subsequent investigations.

A number of authors have investigated non-binary genetic algorithms theoretically (see, e.g. [3, 9, 37, 42, 71, 72]) which have been used successfully in applications [13]. Bhattacharyya and Koehler [9] as well as Leung et al. [42] investigate non-binary genetic algorithms with cardinality 2^v. However, it is clear that this is essentially the binary model with a different crossover operator that fits the definition of a generalized crossover operator in [61]. Thus, this situation is for the most part covered by results in [61] except for Theorems 8.5 and 8.6 presented here. Koehler et al. [37] present a generalization of the multi-set model for arbitrary non-binary alphabets, and show that a theory analogous to the binary case holds to a very large extent.

The starting point in [61, 62] and in this exposition is to represent populations as strings of characters in the underlying alphabet $\mathscr{A}$. This corresponds to what actually happens in computer memory in an application of genetic algorithms. The state space is consequently the set of probability distributions over fixed-length words over $\mathscr{A}$. This approach frees the initial description from clumsy combinatorial coefficients, which may hinder a subsequent detailed analysis. It also has the advantage of allowing to model spatial structure on an evolving population. The study of evolution of spatially structured populations is an area of active interest for both computer scientists (see Mitchell's book [48]) and evolutionary ecologists and population geneticists (see Roughgarden's book [55], and the publication by Peck et al. [53]). This represents a substantial advantage of the model presented here over the multi-set-based models, which do not capture any spatial structure. The models for genetic algorithms using unordered multi-sets can be obtained as projections of the model presented here by simply by forgetting the order of creatures within populations (see Section 2.9). As an example of an application for this projection mechanism, we compute spectral estimates for the non-binary mutation-crossover matrix in the multi-set model in Theorem 6.2. This yields a canonical correspondence to similar bounds for eigenvalues implied by the Vose–Liepins conjecture (see Koehler's proof [36, p. 421], and [70]).

The next major step in the approach presented in [61, 62] and here is to *separately* describe and analyze the three phases mutation, crossover, and fitness selection of the genetic algorithm. This allows to apply spectral theory and other techniques to the matrices that arise. In particular, mutation and crossover can be analyzed conveniently in our model using the fact that the underlying vector space can be obtained by a suitable tensor product construction (see Section 2.6). The linear model for binary genetic algorithms, which was developed in [62], was found independently by Rudolph [57], who proves a part of Theorem 8.2 in the binary case, and analyses a convergent variant of the simple genetic algorithm by extending all linear operators to keep track of a best-so-far individual seen by the algorithm. The analysis [61, 62] and also the analysis presented here does not treat recording the best-so-far individual, but otherwise extends Rudolph's findings for simple genetic algorithms.

The model presented here for genetic algorithms is a Markov chain model. In this model, the basis vectors of the underlying vector space $\mathscr{V}_{\wp}$ correspond to the possible populations in the genetic algorithm which are represented as strings of letters in $\mathscr{A}$. The state of the algorithm after t steps or generations is described by a probability distribution $v = \sum v_p\, p$, $(0 \leqslant v_p \leqslant 1,\ \sum v_p = 1)$, over all possible populations p in the set of populations $\wp_s$. For each of the genetic operators, mutation $M_{\mu(t)}$ with mutation rate $\mu(t)$, crossover $C_{\chi(t)}$ with crossover rate $\chi(t)$, and fitness selection F_t, we have a corresponding stochastic matrix acting on the underlying vector space $\mathscr{V}_{\wp}$. A single iteration G_t through one step t of a genetic algorithm is given by the product linear operator $G_t = F_t C^K_{\chi(t)} M_{\mu(t)}$. The operators $C_{\chi(t)}$ and $M_{\mu(t)}$ commute for the standard methods, i.e., regular pairwise one-point crossover and multiple-spot mutation, but for a larger part of the theory this is not a requirement.

In Section 3, we determine the spectra of single- and multiple-spot mutation. In addition, we investigate contracting properties of mutation with respect to various norms. These estimates can be used to give upper bounds for the combined limit probability of the genetic algorithm over uniform populations, i.e., populations that consist of multiple copies of a single creature. We also compute estimates for the coefficient of ergodicity of mutation which subsequently can be used to obtain criteria for weak ergodicity of the inhomogeneous Markov chain underlying the genetic algorithm.

In Section 4, we determine criteria for weak ergodicity of genetic algorithms. In addition, we investigate the interplay of mutation and the fitness selection operator in regard to the change in non-uniformity of populations (or more precisely probability distributions over populations) over the course of the algorithm.

In Section 5, we give a summary of the results on crossover in [62] with a few additions. The results in [62] carry over almost verbatim, since crossover does not change letters in $\mathscr{A}$ but *positions* of letters. For the purpose of the analysis presented here, we have to introduce an extended definition of the so-called Hardy–Weinberg spaces, which are the span over populations with constant allele frequencies, and which are the invariant subspaces for the crossover operator.

It is important to note that iterated crossover converges on Hardy–Weinberg spaces to uniform distributions over the base vectors (populations), i.e., given a fixed allele frequency, every population with that frequency has equal probability in the limit. This shows that any linkage [20, p. 25] between genes is resolved, where gene means a letter of the underlying alphabet $\mathscr{A}$ in a certain position in the genome of a creature. This is related to work of Geiringer [20], and Vose and Wright [72]. Geiringer [20] considers sexual reproduction with equal distribution for diploid male and female under complete panmixia, while in the model presented here any two haploid individuals or creatures can be mated under crossover. We have included a short simple proof of the analogue of [20, p. 42, Theorem III] showing that the mean distribution of creatures in a Hardy–Weinberg space under iterated crossover in a finite population is the canonical product distribution (see Theorem 5.4). See also [72, p. 287, Theorem 3.9].

In addition to the above, we introduce the notion of *generalized crossover operation*, which covers regular crossover used in most applications and discussed in many models, as well as unrestricted crossover as discussed in [62]. In regard to ergodic-type theorems, the notion of generalized crossover operation is actually sufficient. We note that for some genetic algorithms, the limit probability distribution over populations is independent even of the crossover *method* (see Theorems 8.3 and 8.5).

In Section 6, we show in Theorem 6.1. how, i.e., on which subspaces of $\mathscr{V}_{\mathscr{P}}$, the crossover operation enhances the averaging effect of mutation in the random generator phase of the genetic algorithm. This result is – in the author's opinion – the key result in understanding mutation-crossover vs. mutation as probabilistic generator operators in the algorithm. It characterizes mutation-crossover as procedure with geometric rates of convergence towards the uniformly probability distributions over all populations. In addition to the above, we introduce several *convergent* simulated annealing-type optimization methods (see the publication by Aarts and van Laarhoven [1] for an in-

troduction), which use mutation-crossover as generator matrix. This extends discussions by Mahfoud and Goldberg [44, p. 302, 305], and [62, p. 124, Remark on simulated annealing]. The method presented here allows for dynamic bursts of mutation and crossover rates during the course of the algorithm.

In Section 7, we discuss a variety of standard fitness selection schemes used in applications of genetic algorithms. But the analysis also treats types of fitness selection mechanisms where the fitness of the individual depends upon the ambient population, including, e.g., cases in which the fitness of a genotype depends on the frequencies of the various types of individuals in the population (see publications by Axelrod and Hamilton [4], and Sigmund [64]). In particular, we analyze how fitness selection schemes contract towards the space of probability distributions over uniform populations. This is later combined with the analysis of the interplay fitness selection vs. mutation of Section 4.2 mentioned above to obtain estimates for the proportion of the limit probability distribution over non-uniform populations for a possibly scaled genetic algorithm. Our presentation illuminates the discussion of "punctuated equilibrium" by Vose in [69]. In addition to the above, we discuss the effect of genetic drift (see, e.g., [12, 19, 29, 45, 55]) as modeled in [61] in regard to the extended definitions given in this exposition. Essentially, the results of [61] stay valid: it is not hard to show that such a genetic algorithm without mutation converges pointwise on distributions over populations to a distribution over uniform populations. This analysis extends findings by Fogel in [19], and it yields insight into some applications of genetic algorithms using crossover without mutation.

A number of authors have used genetic algorithms with very low mutation rate or zero mutation rate obtaining seemingly good results. In the author's opinion, these implementations of genetic algorithms may see "convergence" of the algorithm as an instance of having virtually implemented genetic drift. We point out that this process is non-ergodic in nature, i.e., the limit probability distribution over populations depends upon the initial population. Note that Banzhaf et al. [7] report enhanced performance of genetic algorithms using higher mutation rates.

In Section 8, we show strong ergodicity of a large variety of genetic algorithms. These algorithms can be scaled in regard to the mutation rates for single- or multiple-spot mutation, and in regard to the crossover rates for a very general setting for crossover. We impose as main condition, that the mutation rates $\mu(t)$ for multiple-bit mutation converge to a strictly positive value. The algorithm can use tournament fitness selection, rank selection or scaled fitness selection but also fitness selection mechanisms where the fitness of the individual depends also on the population as discussed in Section 7. Our analysis applies to most standard methods of fitness scaling such as linear fitness scaling [25, p. 77], sigma-truncation [25, p. 124], and power-law scaling (see [25, p. 124], and Section 7.1). However, if single-spot mutation is considered, then such processes as unbounded power-law scaling are excluded. We show that the limit probability distribution of such processes is fully positive at every population of uniform fitness, which implies that the algorithm does *not* necessarily converge to populations containing only optimal solutions.

If an injective, population-independent fitness function is used,[1] and if a so-called strong fitness scaling such as unbounded power-law scaling for proportional fitness selection is used, then the limit distribution depends only on the pre-order or rank induced by the fitness function on the set of creatures, and not the particular values of the fitness function. This is shown in Theorem 8.3. This is in agreement with the viewpoint of Baker [6], who proposes to use the fitness ranking of individuals in determining fitness selection probabilities. If, in addition, the mutation and crossover operators commute, which is satisfied in all standard applications, then what actually is done as crossover operation (i.e., *method and scaling* of crossover) are irrelevant for the asymptotic behavior of the genetic algorithm. This means that the limit probability distribution over populations does not depend upon method and/or scaling of crossover.

We also give explicit bounds on the combined probability for non-uniform populations in the limit distribution. In fact, Theorems 8.1 and 8.2 show that for small mutation rates the genetic algorithm asymptotically spends most of its time in uniform populations.

The final part of our analysis contains the most striking new results under the following hypotheses: (a) the mutation rates $\mu(t)$, $t \in \mathbf{N}$, are monotonously decreasing to 0, slow enough to assure weak ergodicity of the inhomogeneous Markov chain describing the algorithm (e.g., $\mu(t) = t^{-1/L}$, $t \in \mathbf{N}$, where L is the length of a population as a word over $\mathscr{A}$); (b) mutation and crossover commute, which is usually satisfied; (c) an injective, population-independent fitness function is used, or rank based upon such a function.

If power-law scaling is used with an "integrable" exponentiation g (e.g., g is linear; see identity (22) in Section 7.1, and identity (31) in Theorem 8.2), then we have the following situation: The Markov chain describing the scaled genetic algorithm is strongly ergodic. The genetic algorithm behaves asymptotically like the so-called *take-the-best search algorithm* which consists of the two steps: (1) mutate with rate $\mu(t)$, and (2) map every population p to the uniform population containing solely the best creature in p. This means in particular, that crossover method, and fitness selection schedule are irrelevant (asymptotically) for the outcome of the genetic algorithm procedure. We investigate a particular example for a take-the-best search algorithm, and we show that it does not converge to the best element (see Section 8.3). These results contradict the main result in [65, 66]. In general, the limit probability distribution over populations, i.e., the outcome of the algorithm, depends upon the GA-landscape. However, if the size s of populations is larger than the length ℓ of the genome of creatures, then the take-the-best search algorithm *does converge* to the best element.

If power-law scaling is used with a "logarithmic" exponentiation g (see identity (22) in Section 7.1, and identities (34) and (35) in Theorem 8.3) and $s > \ell$, then

[1] Actually, we need only a weaker condition that every population contains only one creature of maximal fitness which allows, e.g., to consider population-dependent rank. The discussion in [62, p. 120] shows that an injective, population-independent fitness function on the set of creatures is a minor restriction, if a genetic algorithm is considered for the purpose of function optimization.

the Markov chain describing the scaled genetic algorithm is strongly ergodic, and the algorithm *does converge* to the best element.

The work of Davies and Principe [15, 16] showed that annealing the mutation rate to zero does not imply convergence of an otherwise constant genetic algorithm to global optima with probability 1. Rudolph [57] raised the question whether using a proper fitness scaling would imply convergence to global optima. This was answered to the negative in 1996 in the preprint version of [62] for strong fitness scalings and scaled mutation rate with positive limit (see [62, Theorem 17]). Theorems 8.2 and 8.3 give a positive answer to Rudolph's question in [57] in the remaining case, i.e., annealing the mutation rate to zero and using a strong fitness scaling under certain conditions as outlined above. The examples in Section 8.3 show that some of the conditions are absolutely necessary.

Finally in Section 9, we discuss some aspects of the theory presented here in regard to practical applications of genetic algorithms.

2. Notation and preliminaries

The notation used in this exposition is an extension of the notation used in [62]. We shall give a complete listing here for the convenience of the reader.

2.1. Scalars

Let $\mathbf{N}$, $\mathbf{Z}$, $\mathbf{R}$, $\mathbf{R}^+$, and $\mathbf{C}$ denote the strictly positive integers, the integers, the real numbers, the non-negative real numbers, and the complex numbers respectively. Let $\mathbf{N}_0 = \mathbf{N} \cup \{0\}$. Let $\mathbf{R}_*^+ = \mathbf{R}^+ / \{0\}$. For elements n, m of a set, let $\delta_{n,m} = 1$, if $n = m$, and $\delta_{n,m} = 0$ otherwise, i.e., δ is the Kronecker delta.

2.2. Vectors

The standard base of unit vectors in $\mathbf{C}^n$, $n \in \mathbf{N}$, is denoted as

$$\boldsymbol{b}_v = (\delta_{v,\kappa})_{\kappa=1}^n \in \mathbf{C}^n, \quad 1 \leqslant v \leqslant n.$$

For vectors $v = (v_v)_{v=1}^n, w = (w_v)_{v=1}^n \in \mathbf{C}^n$, $n \in \mathbf{N}$, and $r \in [1, \infty)$, we shall denote the canonical inner product of v and w, and the ℓ^r-norm [59, p. 4] of v by

$$\langle w, v \rangle = \sum_{v=1}^n \bar{w}_v v_v \quad \text{and} \quad ||v||_r = \left(\sum_{v=1}^n |v_v|^r \right)^{1/r}.$$

The ℓ^1-norm shall also be called the Hamming norm. The ℓ^∞-norm or max-norm of v is defined as

$$||v||_\infty = \max\{|v_v| : v = 1 \ldots n\}.$$

2.3. Matrices

Let $\mathbf{M}_n(\Omega)$, $n \in \mathbf{N}$, denote the $n \times n$ matrices with entries in a set Ω. Let $\mathbf{M}_n = \mathbf{M}_n(\mathbf{C})$. A matrix in $\mathbf{M}_n$ will operate by matrix multiplication from the left on vectors in $\mathbf{C}^n$. A matrix in $\mathbf{M}_n(\mathbf{R}^+)$ is called *column stochastic* or (just) *stochastic*, if every of its columns sums to 1. $\mathbf{1}$ shall always denote the identity matrix, i.e., the stochastic matrix whose diagonal entries equal 1. A matrix is called *row stochastic*, if its transpose is column stochastic. A matrix in $\mathbf{M}_n(\mathbf{R}_*^+)$ will be called *fully positive*. A matrix in $X \in \mathbf{M}_n$ shall be called *irreducible* or *indecomposable* [59, p. 19], if there exists no permutation matrix $\pi \in \mathbf{M}_n(\{0,1\})$ such that for some $X_1 \in \mathbf{M}_m$, $1 \leqslant m < n$:

$$X = \pi \cdot \begin{pmatrix} X_1 & X_2 \\ 0 & X_3 \end{pmatrix} \cdot \pi^{-1}. \tag{1}$$

The ℓ^r-norm on $\mathbf{C}^n$ induces an operator norm $||\cdot||_r$ on $\mathbf{M}_n$ [59, p. 5]. In fact, for $X \in \mathbf{M}_n$ we have

$$||X||_r = \sup\{||Xv||_r : ||v||_r = 1\}. \tag{2}$$

Note that if $X \in \mathbf{M}_n$, then – avoiding large numbers of subscripts – the coefficients of X can conveniently be denoted as

$$\langle \boldsymbol{b}_v, X\boldsymbol{b}_{v'} \rangle = X_{v,v'}, \quad 1 \leqslant v, v' \leqslant n.$$

If $X \in \mathbf{M}_n$ is column stochastic, then let the *coefficient of ergodicity with respect to the ℓ^r-norm*, be given by

$$\tau_r(X) = \max\left\{ \|Xv\|_r : v = (v_v)_{v=1}^n \in \mathbf{R}^n, \ \sum_{v=1}^n v_v = 0 \text{ and } \|v\|_r = 1 \right\}. \tag{3}$$

A useful fact from Seneta's book [63, p. 137, line 20, and formula (4.6)] gives

$$\begin{aligned} \tau_1(X) &= 1 - \min\left\{ \sum_{\kappa=1}^n \min(X_{\kappa,v_1}, X_{\kappa,v_2}) : 1 \leqslant v_1, v_2 \leqslant n \right\} \\ &\leqslant 1 - \sum_{\kappa=1}^n \min\{X_{\kappa,v} : 1 \leqslant v \leqslant n\}. \end{aligned} \tag{4}$$

Observe that Seneta's book [63] uses row-stochastic matrices. Thus, the formulas used here are transposed versions of the formulas in [63].

2.4. The alphabet

We shall denote the letters in the alphabet $\mathscr{A}$ of size a, $1 < a \in \mathbf{N}$ underlying the model for genetic algorithm described here by $\hat{a}(0), \hat{a}(1), \ldots, \hat{a}(a-1)$. Sometimes when explicitly stated, we shall identify $\mathscr{A} = \{\hat{a}(\imath) \mid 0 \leqslant \imath \leqslant a-1\}$ with $\mathbf{Z}_a = \mathbf{Z}/a\mathbf{Z}$ such that under this identification $\hat{a}(\imath) \equiv \imath$.

2.5. The basic vector space

We shall denote the free vector space over $\mathscr{A}$ by $\mathscr{V}_1$. We shall identify $\mathscr{V}_1$ with $\mathbf{C}^a$ such that the base vector $\hat{a}(\imath) \in \mathscr{V}_1$ corresponds to base vector $\boldsymbol{b}_{\imath+1} \in \mathbf{C}^a$, $0 \leqslant \imath \leqslant a-1$.

Usually, we shall identify $\mathscr{A}$ with base vectors in $\mathscr{V}_1$. Consequently, – if not otherwise stated – an expression such as $\hat{a}(0)+\hat{a}(1)$ denotes an element in $\mathscr{V}_1$, and not in $\mathbf{Z}_a$.

2.6. The vector space underlying the model

We shall consider *creatures* living in the model world to which the genetic algorithm is applied as words of fixed length $\ell \geqslant 2$ over $\mathscr{A}$. Thus, every creature can be identified with an ℓ-tuple of base vectors in $\mathscr{V}_1$.

There are a^ℓ elements in the set $\mathscr{C}$ of possible creatures. Every creature $c \in \mathscr{C}$ can be also identified with an integer in $[0, a^\ell - 1]$ interpreting c as an integer in the a-adic number system. This induces a natural order on $\mathscr{C}$ used (up to adding 1) for indexing matrices. We set

$$\mathscr{V}_{\mathscr{C}} = \bigotimes_{\lambda=1}^{\ell} \mathscr{V}_1.$$

$\mathscr{V}_{\mathscr{C}}$ is identified with the free vector space over $\mathscr{A}^\ell$ via the map

$$\hat{a}(\imath_1) \otimes \cdots \otimes \hat{a}(\imath_\ell) \mapsto (\hat{a}(\imath_1), \ldots, \hat{a}(\imath_\ell)). \tag{5}$$

The set of populations $\wp_s$ is the set of s-tuples of creatures containing a^L elements where $L = s\ell$. We shall assume throughout this exposition, that $s \geqslant 2$, except when stated explicitly otherwise. Every population $p \in \wp_s$ can be identified with an integer in $[0, a^L - 1]$. This induces a natural order on $\wp_s$.

A *spot* in the genome is, by definition, the position of one of the letters in a word representing a creature or population. The following distance function is of use in regard to the mutation operation. For $c, d \in \mathscr{C}$, or $p, q \in \wp_s$, we define the Hamming distances $\Delta(c,d)$ resp. $\Delta(p,q)$ as the number of spots in the genome where c and d resp. p and q differ. We set

$$\mathscr{V}_{\mathscr{P}} = \bigotimes_{\sigma=1}^{s} \mathscr{V}_{\mathscr{C}} = \bigotimes_{\hat{\lambda}=1}^{L} \mathscr{V}_1.$$

where the latter identification is canonical. $\mathscr{V}_\wp$ is also identified with the free vector space over $\mathscr{A}^L$ similar as in identity (5). This tensor product construction allows a very convenient description of the mutation operator – see Section 3: Propositions 3.3 and 3.6.

Let $\mathscr{U} \subset \mathscr{V}_\wp$ be the free vector space over all populations which are uniform, i.e., which consist of s copies of a single creature. Consequently, $\mathscr{U} \cap \wp_s$ shall denote the set of uniform populations. In addition, $P_{\mathscr{U}}$ shall denote the orthogonal projection onto $\mathscr{U}$. $\mathscr{U}$ is the subspace of $\mathscr{V}_\wp$ which is element-wise invariant under the crossover and

the fitness selection operators – see Section 5.1 and Definition 7.1. Uniform populations are only affected by the mutation operation, and not by crossover or fitness selection.

Let $S \subset \mathscr{V}_{\wp}$ be the set of probability distributions over $\wp_s$, i.e., S is the positive part of the $||\cdot||_1$ unit sphere. S is the relevant state space in this investigation where the genetic operators mutation, crossover, and fitness selection act as column stochastic matrices by matrix multiplication from the left.

2.7. Matrices acting on the vector space underlying the model

If X is a linear operator on $\mathscr{V}_1$, then define a linear operator $X[\hat{\lambda}]$ on $\mathscr{V}_{\wp}$ by

$$X[\hat{\lambda}] = \mathbf{1} \otimes \mathbf{1} \otimes \cdots \otimes \mathbf{1} \otimes X \otimes \mathbf{1} \otimes \mathbf{1} \cdots \otimes \mathbf{1},$$

where X occurs at the $\hat{\lambda}$th tensor spot. This notation as well as the following identity (6) for spectra are very useful in the mathematical modeling of mutation. See Propositions 3.3 and 3.6.

For matrices $X_{\hat{\lambda}}$ acting on $\mathscr{V}_1$, we have (see, e.g., [62, p. 105])

$$\mathrm{sp}\left(\sum_{\hat{\lambda}=1}^{L} X_{\hat{\lambda}}[\hat{\lambda}]\right) = \mathrm{sp}(X_1) + \mathrm{sp}(X_2) + \cdots + \mathrm{sp}(X_L), \tag{6}$$

and

$$\mathrm{sp}\left(\prod_{\hat{\lambda}=1}^{L} X_{\hat{\lambda}}[\hat{\lambda}]\right) = \mathrm{sp}(X_1) \cdots \mathrm{sp}(X_L).$$

The group of permutations Π_s of s elements acts canonically on the set of populations $\wp_s$ rearranging creatures. In fact, for $p = (c_1, c_2, \ldots, c_s) \in \wp_s$, $c_\sigma \in \mathscr{C}$, $1 \leqslant \sigma \leqslant s$, and $\pi \in \Pi_s$, we have

$$\pi p = \pi(p) = (c_{\pi^{-1}(1)}, c_{\pi^{-1}(2)}, \ldots, c_{\pi^{-1}(s)}). \tag{7}$$

Since identity (7) establishes the action of $\pi \in \Pi_s$ on the basis of $\mathscr{V}_{\wp}$, every $\pi \in \Pi_s$ uniquely defines a linear map $\pi : \mathscr{V}_{\wp} \to \mathscr{V}_{\wp}$. With this in mind, let P_Π be given by

$$P_\Pi = s!^{-1} \sum_{\pi \in \Pi_s} \pi, \tag{8}$$

i.e., P_Π is a linear combination of elements in the space of linear maps $\mathscr{V}_{\wp} \to \mathscr{V}_{\wp}$. P_Π models arbitrary rearrangement of creatures in a population, and is used in regard to application of our theory to the multi-set model for genetic algorithms studied in [14–16, 25, 29, 34, 50, 65, 66, 68–70].

For any matrix X acting on $\mathscr{V}_{\wp}$, let $X^{[v]}$ be the matrix where the first row is replaced by v, where $v \in \mathscr{V}_{\wp}$ is seen as a row vector.

2.8. Maximal-entropy distributions

For any subset $\Omega \subseteq \wp_s$ let $e_\Omega \in \mathrm{span}_{\mathbf{C}}(\Omega)$ be the vector such that $\langle p, e_\Omega \rangle = \mathrm{card}(\Omega)^{-1}$ for every $p \in \Omega$. e_Ω is a probability distribution. Let $e = e_{\wp_s}$. Note that this notation is

slightly different from the notation introduced in [62, p. 104] (by a factor 2^{-L}), but coincides with the notation used in [61]. Let P_e be the orthogonal projection onto the one-dimensional space generated by e.

For a probability distribution $v \in S \subseteq \mathscr{V}_\wp$, the entropy [58, Section 2.1] is defined as

$$H(v) = -\sum_{p \in \mathscr{P}_s} v_p \log(v_p) \tag{9}$$

with the convention that $0\log(0) = 0$. By using induction over the dimension of the underlying vector space, it is not hard to show, that e is the maximal-entropy distribution in $S \subseteq \mathscr{V}_\wp$. In fact, by induction hypothesis, one knows the maximal values of the entropy on the boundary of S, and then can determine the maximum in the interior by differentiating identity (9). See the publication by Rudolph [58], for an investigation of genetic algorithms for integer programming on the unbounded domain $\mathbf{Z}^n$, $n \in \mathbf{N}$, using maximal-entropy distributions.

2.9. Identification with the multi-set model for populations

Many researchers have investigated genetic algorithms using the multi-set model (see [14–16, 25, 29, 34, 50, 65, 66, 68–70]). In the following paragraph, we shall establish the mathematical connection between the multi-set model and the model presented here.

Let $p = (c_1, \ldots, c_s) \in \wp_s$, $c_\sigma \in \mathscr{C}$. Suppose that c_σ occurs $v_\sigma \in \mathbf{N}$ times in p, i.e., there are v_σ copies of c_σ in p. Now, we define the set $\mathrm{mset}(p)$ by

$$\mathrm{mset}(p) = \{(c_1, v_1), \ldots, (c_s, v_s)\}.$$

Let $\mathrm{mset}(\wp_s) = \{\, \mathrm{mset}(p) : p \in \wp_s \,\}$. Next, we define a map $\Lambda : \mathrm{mset}(\wp_s) \to S$ identifying populations, which are understood as multi-sets, canonically with a probability distribution in S. In fact, we define

$$\Lambda(\mathrm{mset}(p)) = P_\Pi(p).$$

Since two populations which generate the same multi-set can be transformed into each other by a permutation, it follows that Λ is well defined.

Let $\mathscr{W}_\wp$ denote the free vector space over $\mathrm{mset}(\wp_s)$. Λ extends uniquely to a linear map $\Lambda : \mathscr{W}_\wp \to P_\Pi(\mathscr{V}_\wp) \subset \mathscr{V}_\wp$. Let $S_0 \subset \mathscr{W}_\wp$ be the set of probability distributions over $\mathrm{mset}(\wp_s)$. Clearly $\Lambda(S_0) = P_\Pi(S) \subset S$. It is easy to see that $\Lambda : \mathscr{W}_\wp \to P_\Pi(\mathscr{V}_\wp)$ is invertible. The corresponding inverse linear map $P_\Pi(\mathscr{V}_\wp) \to \mathscr{W}_\wp$ will be denoted by Λ^{-1}.

2.10. Means

$\mathrm{mean}_{\mathbf{u}}$ and $\mathrm{Mean}_{\mathbf{r}}$ as defined below, and introduced for binary genetic algorithms in [62, p. 107] measure allele/gene frequencies in populations. They are a very useful tool in analyzing the interplay mutation vs. crossover.

First, we deal with the case of the invariant for unrestricted crossover as defined in Section 5.3. If $p=(c_1,c_2,\ldots,c_s)\in\wp_s$ is a population, then for the purpose of the next definition consider every creature c_σ as an ℓ-tuple of base-vectors in $\mathscr{V}_{\hat{1}}$. Now, we set

$$\mathrm{mean}_{\mathbf{u}}(p)=s^{-1}\sum_{\sigma=1}^{s}c_\sigma\in\bigotimes_{\lambda=1}^{\ell}\mathscr{V}_{\hat{1}}.$$

For example, if $a=3$, $\ell=2$, $s=4$, and $p=((\hat{a}_0,\hat{a}_0),(\hat{a}_0,\hat{a}_1),(\hat{a}_0,\hat{a}_2),(\hat{a}_0,\hat{a}_2))$, then we have $\mathrm{mean}_{\mathbf{u}}(p)=(\hat{a}_0,\frac{1}{4}(\hat{a}_0+\hat{a}_1+2\hat{a}_2))$.

Next, we deal with the invariant $\mathrm{Mean}_{\mathbf{r}}$ related to regular crossover as defined in Section 5.2. In this case, assume that the populations size s is even. Let

$$\mathrm{Mean}_{\mathbf{r}}(p)=\tfrac{1}{2}(c_1+c_2,\ldots,c_{s-1}+c_s)\in\left(\bigoplus_{\lambda=1}^{\ell}\mathscr{V}_{\hat{1}}\right)^{s/2}.$$

For the example given above, we have $\mathrm{Mean}_{\mathbf{r}}(p)=((\hat{a}_0,\frac{1}{2}(\hat{a}_0+\hat{a}_1)),(\hat{a}_0,\hat{a}_2))$. Both $\mathrm{mean}_{\mathbf{u}}$ and $\mathrm{Mean}_{\mathbf{r}}$ extend uniquely to linear maps on $\mathscr{V}_{\wp}$.

2.11. Hardy–Weinberg spaces

Hardy–Weinberg[2] spaces as introduced in [62, p. 108] are canonically invariant subspaces of $\mathscr{V}_{\wp}$ for the standard crossover operators, i.e., regular crossover and unrestricted crossover. For a given $\xi\in\bigoplus_{\lambda=1}^{\ell}\mathscr{V}_{\hat{1}}$, let $\mathscr{V}_{\xi}$, the *Hardy–Weinberg space of populations with gene frequency* ξ, be defined as

$$\mathscr{V}_{\xi}=\mathrm{span}_{\mathbf{C}}\{p\mid p\in\mathscr{P}_s,\mathrm{mean}_{\mathbf{u}}(p)=\xi\}.$$

Define $\overline{\mathscr{V}_{\xi}}$ analogously, if $\xi=\mathrm{Mean}_{\mathbf{r}}(p)$ for a population p of even size s. Let P_{ξ}, $P_{\bar{\xi}}$ be the orthogonal projection onto $\mathscr{V}_{\xi}$, $\overline{\mathscr{V}_{\xi}}$, respectively.

2.12. The diagonal spaces

Let the diagonal space D be defined as follows: If $\xi=\mathrm{mean}_{\mathbf{u}}(p)$ for a population p, then let

$$e_{\xi}=e_{\{q\in\mathscr{P}_s\mid\xi=\mathrm{mean}_{\mathbf{u}}(q)\}}$$

and let

$$D=\mathrm{span}_{\mathbf{C}}\{e_{\xi}\mid\xi=\mathrm{mean}_{\mathbf{u}}(p),p\in\mathscr{P}_s\}.$$

Similarly to the above, if $\xi=\mathrm{Mean}_{\mathbf{r}}(p)$ for a population p of even population size s, then let

$$e_{\bar{\xi}}=e_{\{q\in\mathscr{P}_s\mid\bar{\xi}=\mathrm{Mean}_{\mathbf{r}}(q)\}},$$

[2] Assuming identical distributions for females and males, and random breeding (panmixia), Hardy showed in [32] that a state of equilibrium is reached in a population after the first generation for the probabilities of genotypes associated with one Mendelian character. In [73], Weinberg investigated the sequence of distribution of genotypes over generations for 2 Mendelian characters which converges to the canonical product distribution. See [20, p. 42, Theorem III].

and let

$$\bar{D} = \mathrm{span}_{\mathbf{C}}\{e_{\bar{\xi}} \mid \bar{\xi} = \mathrm{Mean}_{\mathbf{r}}(p),\ p \in \mathscr{P}_s\}.$$

Let P_D, and $P_{\bar{D}}$ be the orthogonal projection onto D, and $\overline{D}$, respectively. Since $e \in D$, and $e \in \bar{D}$, we have

$$P_D P_e = P_e = P_e P_D \quad \text{and} \quad P_{\bar{D}} P_e = P_e = P_e P_{\bar{D}}.$$

Observe in addition, that $\pi e_\xi = e_\xi$ for $\pi \in \Pi_s$, $\xi = \mathrm{mean}_{\mathbf{u}}(p)$, $p \in \wp_s$. Hence,

$$P_\Pi P_D = P_D = P_D P_\Pi \ \Rightarrow\ P_\Pi(\mathbf{1} - P_D) = (\mathbf{1} - P_D)P_\Pi(\mathbf{1} - P_D).$$

This shows that P_Π maps $D^\perp$ into $D^\perp$.

2.1. Lemma. *Denote $e = e_{\mathscr{A}} \in \mathscr{V}_{\hat{1}}$ in this lemma (using the above definitions for $\ell = s = 1$). Let $z \in \mathscr{V}_{\hat{1}}$ be such that $z \perp e$. Then we have*

1. $x_1 = z \otimes e \otimes e \ldots \otimes e + e \otimes e \ldots \otimes e \otimes z \otimes e \otimes e \ldots \otimes e \in \bar{D}$,
   where the second z occurs at spot $\ell + 1$, i.e., *the first spot in the second creature.*
2. $x_2 = z \otimes e \otimes e \ldots \otimes e - e \otimes e \ldots \otimes e \otimes z \otimes e \otimes e \ldots \otimes e \in \bar{D}^\perp$,
   where again the second z occurs at spot $\ell + 1$.

Proof. Let $z = \sum_{\imath=0}^{a-1} z_\imath \hat{a}(\imath)$. Then

$$\begin{aligned} z \otimes e \pm e \otimes z &= a^{-1} \sum_{\imath,\imath'} z_\imath \hat{a}(\imath) \otimes \hat{a}(\imath') \pm a^{-1} \sum_{\imath,\imath'} z_{\imath'} \hat{a}(\imath) \otimes \hat{a}(\imath') \\ &= a^{-1} \sum_{\imath,\imath'} z_\imath (\hat{a}(\imath) \otimes \hat{a}(\imath') \pm \hat{a}(\imath') \otimes \hat{a}(\imath)). \end{aligned}$$

Up to permuting tensor spots, this shows Lemma 2.1.1 Consider $\pi \in \Pi_L$ acting on populations by permuting the L letters in a population canonically, and extend π to a linear map acting on $\mathscr{V}_\wp$. Let $\pi \in \Pi_L$ be a transposition. Then $\frac{1}{2}(\mathbf{1} + \pi)$ is a self-adjoint, stochastic projection acting on $\mathscr{V}_\wp$. The orthogonal projection $P_{\bar{D}}$ onto $\bar{D}$ can be expressed as a product of commuting projections as just considered, where the associated transpositions exchange spots $(1, \ell + 1)$, $(2, \ell + 2)$, etc. Consequently, if $P_{\bar{D}}$ is applied to x_2, then this yields 0. □

3. Mutation

3.1. The basic mutation matrix

The basic mutation matrix $m^{(1)} \in \mathbf{M}_a$ models the *change* of a single spot in the genome, i.e., a change within the alphabet $\mathscr{A}$ at a given position in the genome.

In case of a binary alphabet, $m^{(1)}$ models flipping a single bit in the genome, and thus equals the flip matrix $\mathbf{f}$, i.e., we have for $a = 2$:

$$m^{(1)} = \mathbf{f} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \in \mathbf{M}_2.$$

In general, we can model mutation at the level of a single spot of the genome by any stochastic matrix $m^{(1)} \in \mathbf{M}_a$ satisfying

$$\langle \hat{a}(\imath), m^{(1)}\hat{a}(\imath)\rangle = 0, \quad \imath \in \mathbf{Z}_a.$$

If demanded by application or model, one can generalize the analysis given below to cases such as the cyclic shift, i.e., $\langle \hat{a}(\imath+1), m^{(1)}\hat{a}(\imath)\rangle = 1$, $\imath \in \mathbf{Z}_a$. Another case of interest may even be to split the alphabet at this level using a reducible matrix $m^{(1)}$ (see identity (1), or [59, p. 19]). In any case, we have $|\mathrm{sp}(m^{(1)})| \subset [0,1]$. This follows, e.g. from [59, p. 5, formula (7′)], and [56, Theorem 10.13(b)].

We remark that for application purposes, a finite discrete domain of real numbers

$$\mathscr{A} = \{x_0 + \imath \cdot \delta : x_0, \delta \in \mathbf{R}, \imath = 1 \ldots a\},$$

can be used as alphabet in a genetic algorithm.

In [58], Rudolph investigates genetic algorithms for integer programming on the unbounded domain $\mathbf{Z}^n$, $n \in \mathbf{N}$. In particular, Rudolph argues to use distributions for mutation which are "spread out as uniformly as possible" [58, p. 140], i.e., maximal-entropy distributions. For the unbounded domain, this yields geometric distributions [58, Section 2.1, Proportion 1, Section 2.2, Proportion 2 and Definition 2 ff.].

Considering the case of bounded domain $\mathscr{A}$, and following Rudolph's argument for a mechanism of *change* of letter which is uniformly spread, mutation should change the current letter of the alphabet $\mathscr{A}$ with equal probability to a different, new letter. Consequently, we define the *regular basic mutation matrix* $\hat{m}^{(1)}$ by

$$\langle \hat{a}(\imath'), \hat{m}^{(1)}\hat{a}(\imath)\rangle = (a-1)^{-1}, \quad \imath' \neq \imath \in \mathbf{Z}_a.$$

In what follows, we shall usually assume that $m^{(1)} = \hat{m}^{(1)}$, if not explicitly stated otherwise.

3.1. Lemma. *Denote* $e = e_{\mathscr{A}} \in \mathscr{V}_1$ *in this lemma* (*using the above definitions for* $\ell = s = 1$). *The spectrum of* $\hat{m}^{(1)}$ *satisfies* $\mathrm{sp}(\hat{m}^{(1)}) = \{-(a-1)^{-1}, 1\}$, *where the eigenspace to eigenvalue* 1 *is spanned by* e, *the maximal-entropy distribution over* $\mathscr{A}$.

Proof. Clearly, P_e is a self-adjoint projection with one-dimensional range. Consequently, $\mathbf{1} - P_e$ is the self-adjoint projection onto the kernel of P_e with range of dimension $a-1$. Since $\hat{m}^{(1)} = -(a-1)^{-1}(\mathbf{1} - P_e) + P_e$, we obtain Lemma 3.1. □

Let $B(\hat{m}^{(1)}) = \{e_{\mathscr{A}} \in \mathscr{V}_1, z_\imath \in \mathscr{V}_1 : \imath = 1 \ldots a-1\}$ where the $z_\imath$ form an orthonormal basis of $(\mathbf{1} - P_e)\mathscr{V}_1$.

3.2. Single-spot mutation

Let $\mu \in (0, L^{-1})$, where L is the length of populations as words over the alphabet $\mathscr{A}$. Single-spot mutation (see also [62, Section 2.1]) is defined as the following three-step procedure:

1. Decide whether a single-spot mutation should take place with probability $L\mu$.

2. If single-spot mutation takes place, then chose one of the spots in the population with probability L^{-1}.
3. If single-spot mutation takes place, then *change* the current letter in the underlying alphabet $\mathscr{A}$ in accordance with $\hat{m}^{(1)}$ in the spot selected in step (2).

Single-spot mutation corresponds closely to what in biology is called a "point mutation" in a single individual of a population. Since by design single-spot mutation allows only a limited amount of change in the genome, it fits very well into the philosophy of "small neighborhoods" used in simulated annealing [1] (see Section 6.2).

3.2. Example. Let us consider the case, $a=\ell=s=3$. Then $L=9$. Set $\mathscr{A}=\{0,1,2\}$. Let $\mu=\frac{1}{20}$. Let $p_{\hat{a}}=((\hat{a},0,0),(0,0,0),(0,0,0))\in\wp_s$, $\hat{a}\in\mathscr{A}$, and assume that p_0 is the current population. Then, $\Delta(p_0,p_1)=1$. Now, the decision to mutate is made positively in 45% of all applications of single-spot mutation. Hence, the probability of no change is $55\%=1-L\mu$. The probability that spot 1 (where p_0 and p_1 differ) is chosen for change of letter is 5%. If spot 1 is chosen for change of letter, then the first letter 0 in p_0 is changed with probability $\frac{1}{2}$ to letter 1, and otherwise to letter 2. Hence, the probability for transition from p_0 to p_1 is $2.5\%=\mu/(a-1)$.

The following proposition collects the spectral information about the stochastic matrix $M_\mu^{(1)}$ describing single-spot mutation.

3.3. Proposition. *Let $M_\mu^{(1)}$ denote the symmetric, stochastic matrix acting on $\mathscr{V}_\wp$ describing transition probabilities for entire populations under single-spot mutation. We have*

1. *Let $p,q\in\wp_s$. The coefficients of $M_\mu^{(1)}$ are as follows:*
   - $\langle p,M_\mu^{(1)}p\rangle=1-L\mu$.
   - $\langle q,M_\mu^{(1)}p\rangle=\mu(a-1)^{-1}$, *if* $\Delta(p,q)=1$.
   - $\langle q,M_\mu^{(1)}p\rangle=0$, *if* $\Delta(p,q)>1$.
2. $M_\mu^{(1)}=(1-L\mu)\mathbf{1}+L\mu M_{1/L}^{(1)}$, *where* $M_{1/L}^{(1)}=L^{-1}\sum_{\hat{\lambda}=1}^{L}\hat{m}^{(1)}[\hat{\lambda}]$.
3. $\mathrm{sp}(M_\mu^{(1)})=1-\mu a/(a-1)\cdot([0,L]\cap\mathbf{N}_0)$. *Consequently, $M_\mu^{(1)}$ is invertible and C^*-positive*[3] *if* $\mu<(a-1)/La$.

   Note that the eigenvectors for $M_\mu^{(1)}$ are all eigenvectors to corresponding eigenvalues of $M_{1/L}^{(1)}$. It is easy to see that eigenvectors of $M_{1/L}^{(1)}$ are suitable sums of orthogonal vectors of the form $\bigotimes_{\hat{\lambda}=1}^{L}z_{\hat{\lambda}}$, where every $z_{\hat{\lambda}}\in B(\hat{m}^{(1)})$ is an eigenvector of $\hat{m}^{(1)}$.
4. $\|M_\mu^{(1)}\|_r=1$ *for* $r\in[1,\infty]$, *where the operator norm* $\|\cdot\|_r$ *corresponds to the* ℓ^r*-norm on* $\mathscr{V}_\wp$ *as in* [59, *p.* 5, formula 5].
5. $M_\mu^{(1)}$ *commutes with every permutation operator* $\pi\in\Pi_s$.
6. $(M_\mu^{(1)})^L$ *is fully positive. Consequently, e is, up to scalar multiples, the only eigenvector of* $M_\mu^{(1)}$ *to eigenvalue* 1. *Thus,* P_e *is the spectral projection of* $M_\mu^{(1)}$ *to eigenvalue* 1.

[3] A C^*-positive matrix is the square of a self-adjoint matrix. Equivalently, it is self-adjoint with positive spectrum. See [62, p. 132, Corollary B3], or [56, p. 282, Definition 11.27, Theorem 11.28].

Proof. The statement Proposition 3.3.1 for the coefficients of $M_\mu^{(1)}$ follows immediately from the definition of the single-spot mutation operation given in the first paragraph of Section 3.2. Now, Proposition 3.3.2 follows directly from Proportion 3.3.1 by checking how the matrices act on populations. The formula Proportion 3.3.3 for the spectrum follows from Proportion 3.3.2 and the upper part of identity (6) which is applied to $L \cdot M_{1/L}^{(1)}$. The details in Proportion 3.3.3 about eigenvectors follow from Lemma 3.1, and the identities

$$M_{1/L}^{(1)} = L^{-1} \sum_{\hat{\lambda}=1}^{L} \hat{m}^{(1)}[\hat{\lambda}] \quad \text{and} \quad M_\mu^{(1)} = (1 - L\mu)\mathbf{1} + L\mu M_{1/L}^{(1)}.$$

$M_\mu^{(1)}$ is doubly stochastic. By [59, p. 17, Proposition 5.3], $M_\mu^{(1)}$ is a convex combination of permutation matrices. If $Q \in \mathbf{M}_{a^L}$ is a permutation matrix, then $\|Q\|_r = 1$ for $r \in [1, \infty]$, since coefficients of a vector in $\mathscr{V}_\varphi$ are only exchanged by the action of Q. Since, consequently, $M_\mu^{(1)}$ is a convex combination of matrices with operator norm equal to 1, we have $\|M_\mu^{(1)}\|_r \leqslant 1$. Since e is a fix point of $M_\mu^{(1)}$, we have $\|M_\mu^{(1)}\|_r \geqslant 1$. Proposition 3.3.5 is rather obvious. If $M_\mu^{(1)}$ has been applied L times, then every spot in any given population may have changed, i.e., $(M_\mu^{(1)})^L$ is fully positive. By [59, p. 23, Corollary 1], 1 is a simple root of the characteristic equation of $(M_\mu^{(1)})^L$. By [59, p. 9, Proposition 2.8], 1 is a simple pole of the resolvent, and thus the corresponding eigenspace is one-dimensional. This shows Proposition 3.3.6.

The next result summarizes geometric properties of $M_\mu^{(1)}$ in regard to being a contractive map.

3.4. Proposition. *Let $M_\mu^{(1)}$ be the doubly stochastic matrix describing single-spot mutation.*

1. *$M_\mu^{(1)}$ is a contracting map in the Euclidean norm on both $e^\perp$ and S with fixed points 0 and e, respectively.*
   - *If $a = 2$, then the contracting factor is given by $\max\{2\mu L - 1, 1 - 2\mu\}$. The smallest possible contracting factor $1 - 2(L+1)^{-1}$ is attained for $\mu = (L+1)^{-1}$.*
   - *If $a > 2$, then the contracting factor is given by $1 - \mu a/(a-1)$. The smallest possible contracting factor is "attained" for $\mu = L^{-1}$.*
2. *If $\mu < (a-1)/aL$, and $v \perp e$, then*

$$\left(1 - \frac{L\mu a}{a-1}\right) \|v\|_2 \leqslant \|M_\mu^{(1)} v\|_2 \leqslant \left(1 - \frac{\mu a}{a-1}\right) \|v\|_2.$$

   This identity shows that for small μ contracting by $M_\mu^{(1)}$ stays controlled in the sense that the $\|\cdot\|_2$-norm of a vector orthogonal to e cannot shrink too much.
3. *We have for the coefficients of ergodicity (see identity (3))*

$$\tau_1(M_\mu^{(1)}) = \tau_\infty(M_\mu^{(1)}) = 1.$$

4. *If $v \in S$ is a probability distribution, then*

$$L\mu + \left(1 - L\mu - \frac{\mu}{a-1}\right)\|(\mathbf{1} - P_{\mathscr{U}})v\|_1 \leqslant \|(\mathbf{1} - P_{\mathscr{U}})M_\mu^{(1)}v\|_1 \leqslant L\mu + (1 - L\mu)\|(\mathbf{1} - P_{\mathscr{U}})v\|_1.$$

This inequality describes to what degree the state of the genetic algorithm is driven away from uniform populations by single-spot mutation. Note that the upper estimate does not depend upon a, the size of the alphabet.

Proof. Proposition 3.4.1 is a direct consequence of Proposition 3.3.3, and the fact that e spans the one-dimensional eigenspace of $M_\mu^{(1)}$ to eigenvalue 1. Proposition 3.4.2 is a direct consequence of the spectral theorem for normal matrices (see, e.g., [40, p. 268], [30, p. 337], or [62, p. 131, Corollary B2]). Let p_0 and p_1 be the populations with all letters equal to $\hat{a}(0)$, and all letters equal to $\hat{a}(1)$, respectively. Since by hypothesis $L \geqslant 4$, there is no population that can reach both of these by a single-spot mutation. Therefore, $\min(\langle q, M_\mu^{(1)} p_0\rangle, \langle q, M_\mu^{(1)} p_1\rangle) = 0$ for all $q \in \wp_S$. Thus, by identity (4), we have $\tau_1(M_\mu^{(1)}) = 1$, since

$$\min_{p,p'}\left\{\sum_q \min(\langle q, M_\mu^{(1)} p\rangle, \langle M_\mu^{(1)} p'\rangle)\right\} = 0.$$

For the above p_0, define $v \in \mathscr{V}_\wp$ by $\langle v, p_0\rangle = 1$, $\langle v, p\rangle = 1$ if $\Delta(p, p_0) = 1$, and $\langle v, q\rangle = -1$ for exactly $(a-1)L + 1$ populations q with $\Delta(q, p_0) > 1$. Now $v \perp e$, $\|v\|_\infty = 1$, and

$$\langle p_0, M_\mu^{(1)} v\rangle = 1 - L\mu + \mu(a-1)^{-1}(a-1)L = 1.$$

Hence, $\tau_\infty(M_\mu^{(1)}) \geqslant 1$. Clearly, $\tau_\infty(M_\mu^{(1)}) \leqslant 1$ by [59, p. 5, formula (7)]. This shows Proposition 3.4.3. If $p \in \wp_S \cap \mathscr{U}$ is uniform, then it is mapped with probability $L\mu$ to a non-uniform population. If p is non-uniform, then changing any spot might keep it non-uniform, and it makes it uniform with probability at most $\mu(a-1)^{-1}$ (if p is uniform up to one spot). Thus, for $v \in S$

$$L\mu\|P_{\mathscr{U}} v\|_1 + \left(1 - \frac{\mu}{a-1}\right)\|(\mathbf{1} - P_{\mathscr{U}})v\|_1 \leqslant \|(\mathbf{1} - P_{\mathscr{U}})M_\mu^{(1)}v\|_1 \leqslant L\mu\|P_{\mathscr{U}} v\|_1 + \|(\mathbf{1} - P\mathscr{U})v\|_1.$$

This shows Proposition 3.4.4. □

As simple as it seems, the following result is the key observation for handling the combined mutation-crossover operation.

3.5. Proposition. *We have $M_\mu^{(1)} D \subseteq D$. Consequently, $M_\mu^{(1)} P_D = P_D M_\mu^{(1)} P_D = P_D M_\mu^{(1)}$. The same statements hold, if D is replaced by $\bar{D}$.*

Proof. Suppose $\xi=(\xi_1,\ldots,\xi_\ell)=\mathrm{mean}_{\mathbf{u}}(p)$ for $p\in\wp_s$, and $q\in\wp_s$ satisfies $\Delta(p,q)=1$. Let $\zeta=(\zeta_1,\ldots,\zeta_\ell)=\mathrm{mean}_{\mathbf{u}}(q)$. Now assume in addition, that p and q differ such that a switch from $\hat{a}(0)$ to $\hat{a}(1)$ occurred in the first spot within the creatures, i.e.,

$$\xi_1=s^{-1}((n+1)\hat{a}(0)+m\hat{a}(1)+v) \quad \text{and} \quad \zeta_1=s^{-1}(n\hat{a}(0)+(m+1)\hat{a}(1)+v),$$

where $n,m\in\mathbf{N}$, and $v\in\mathrm{span}_{\mathbf{C}}\{\hat{a}(\imath):\imath\geqslant 2\}$. Now, if we consider q fixed for a moment, then there are $m+1$ fixed first spots in creatures in q which equal $\hat{a}(1)$. Hence, there are exactly $m+1$ populations $p'\in\wp_s$ with $\xi=\mathrm{mean}_{\mathbf{u}}(p')$, which can produce q via single-spot mutation $M_\mu^{(1)}$. Thus, we have

$$\left\langle q, M_\mu^{(1)}\sum_{p\in\mathscr{V}_\xi} p\right\rangle q=\frac{(m+1)\mu}{a-1}q \;\Rightarrow\; P_\zeta M_\mu^{(1)}e_\xi=\frac{(m+1)\mu\ \dim(\mathscr{V}_\zeta)}{(a-1)\ \dim(\mathscr{V}_\xi)}e_\zeta.$$

Varying ζ yields Proposition 3.5 for D. The proof of the remainder concerning $\bar{D}$ is similar to the above, and left to the reader. □

We note that with the same argument as in [62, Proposition 1(8)], one can show the following: if $\xi=\mathrm{mean}_{\mathbf{u}}(p)$ for some $p\in\wp_s$, then $P_\xi M_\mu^{(1)}P_\xi=(1-L\mu)P_\xi$. However, this and corresponding statements for multiple-spot mutation are not needed in the sequel.

3.3. Multiple-spot mutation

Let $\mu\in(0,(a-1)/a]$. Multiple-spot mutation (see also [62, Section 2.1, p. 110]) is defined as the following procedure:

1. For $\hat{\lambda}=1,\ldots,L$ do the next two steps:
2. Decide whether or not mutation takes place at spot $\hat{\lambda}$ in the current population with probability μ.
3. If mutation takes place at spot $\hat{\lambda}$, then *change* the current letter in the underlying alphabet $\mathscr{A}$ in accordance with $\hat{m}^{(1)}$.

The following proposition collects the spectral information about the stochastic matrix $M_\mu^{(m)}$ describing multiple-spot mutation.

3.6. Proposition. *Let $M_\mu^{(m)}$ denote the symmetric, stochastic matrix acting on $\mathscr{V}_\wp$ describing transition probabilities for entire populations under multiple-spot mutation. We have*

1. *Let $p,q\in\wp_s$. The coefficients of $M_\mu^{(m)}$ are as follows*:

$$\langle q,M_\mu^{(m)}p\rangle=\left(\frac{\mu}{a-1}\right)^{\Delta(p,q)}(1-\mu)^{L-\Delta(p,q)}>0.$$

   Thus e is, up to scalar multiples, the only eigenvector of $M_\mu^{(m)}$ to eigenvalue 1. *Consequently, P_e is the spectral projection of $M_\mu^{(m)}$ to eigenvalue* 1.
2. $M_\mu^{(m)}=\prod_{\hat{\lambda}=1}^{L}((1-\mu)\mathbf{1}+\mu\hat{m}^{(1)}[\hat{\lambda}])$.
3. $\mathrm{sp}(M_\mu^{(m)})=\{(1-\mu a/(a-1))^{\hat{\lambda}}:\hat{\lambda}\in[0,L]\cap\mathbf{N}_0\}$. *Consequently, $M_\mu^{(m)}$ is C^*-positive and is invertible, except for $\mu=(a-1)/a$. If $\mu=(a-1)/a$, then $M_\mu^{(m)}=P_e$.*

Note that the eigenvectors for $M_\mu^{(m)}$ *are suitable sums of orthogonal vectors of the form* $\bigotimes_{\hat{\lambda}=1}^{L} z_{\hat{\lambda}}$, *where every* $z_{\hat{\lambda}} \in B(\hat{m}^{(1)})$ *is an eigenvector of* $\hat{m}^{(1)}$.

4. $\|M_\mu^{(m)}\|_r = 1$ *for* $r \in [1, \infty]$.
5. $M_\mu^{(m)}$ *commutes with every* $\pi \in \Pi_s$.

Proof. The coefficients of $M_\mu^{(m)}$ are immediate from the definition of multiple-spot mutation in the first paragraph of Section 3.3. The remainder of Proposition 3.6.1 follows as Proposition 3.3.6. The matrix identity Proposition 3.6.2 follows from Proposition 3.6.1 by checking the action of the matrices on populations. The description of the spectrum Proposition 3.6.3 follows from Proposition 3.6.2 and the lower part of identity (6). The statements in Proposition 3.6.3 about eigenvectors then follow from the matrix-formula in Proposition 3.6.2. Proposition 3.6.4 follows as Proposition 3.3.4. Proposition 3.6.5 is rather obvious. □

The next result summarizes geometric properties of $M_\mu^{(m)}$ in regard to being a contractive map.

3.7. Proposition. *Let* $M_\mu^{(m)}$ *denote the doubly stochastic matrix describing multiple-spot mutation.*

1. $M_\mu^{(m)}$ *is a contracting map in the Euclidean norm on both* $e^\perp$ *and* S *with fixed points* 0 *and* e, *respectively. The contracting factor is given by* $1 - \mu a/(a-1)$.
2. *If* $v \perp e$ *then*

$$\left(1 - \frac{\mu a}{a-1}\right)^L \|v\|_2 \leqslant \|M_\mu^{(m)} v\|_2 \leqslant \left(1 - \frac{\mu a}{a-1}\right) \|v\|_2.$$

Similar to Proposition 3.4.2, *this shows that convergence towards the fixed points* 0 *on* $e^\perp$, *or* e *on* S *stays controlled for small* μ *in the sense that the* $\|\cdot\|_2$*-norm of a vector orthogonal to* e *cannot shrink too much.*

3. *We have for the coefficients of ergodicity* (*see identity* (3))

$$\tau_r(M_\mu^{(m)}) \leqslant 1 - \left(\frac{a\mu}{a-1}\right)^L \quad \text{for } r = 1 \text{ or } r = \infty.$$

Consequently, $M_\mu^{(m)}$ *is a contracting map both on* $e^\perp$ *and* S *in the Hamming norm and the* ∞*-norm with contracting factor bounded above by* $1 - (a\mu/(a-1))^L$.

4. *Let* $v \in S$. *Let* $h = s/a$. *Define* $\gamma, \beta_0, \beta \in [0,1]$ *by*

$$\gamma = 1 - \left(\frac{\mu^s}{(a-1)^{s-1}} + (1-\mu)^s\right)^\ell,$$

$$\beta_0 = \left(\frac{\mu^s}{(a-1)^{s-1}} + (1-\mu)^s\right)^\ell - \left(\frac{\mu^s}{(a-1)^{s-1}} + (1-\mu)^s\right)^{(\ell-1)} \cdot \left(\frac{\mu^{s-1}(1-\mu)}{(a-1)^{s-1}} + \frac{\mu(1-\mu)^{s-1}}{a-1} + (a-2)\left(\frac{\mu}{a-1}\right)^s\right)$$

$$\beta = \left(\frac{\mu^s}{(a-1)^{s-1}} + (1-\mu)^s\right)^{\ell} - \left(\frac{\mu}{a-1}\right)^L a^{\ell} \left(\frac{(a-1)(1-\mu)}{\mu}\right)^{h\ell}.$$

Then we have the following estimates:

$$\gamma + \beta_0 \|(\mathbf{1} - P_{\mathscr{U}})v\|_1 \leqslant \|(\mathbf{1} - P_{\mathscr{U}})M_\mu^{(m)} v\|_1 \leqslant \gamma + \beta \|(\mathbf{1} - P_{\mathscr{U}})v\|_1.$$

Proof. Proposition 3.7.1 and Proposition 3.7.2 are immediate from spectral calculus with the same arguments as in the proof of Proposition 3.4.1–2. By identity (4), we have

$$\tau_1(M_\mu^{(m)}) \leqslant 1 - \sum_{p \in \wp_s} \min_{q \in \wp_s} \{\langle q, M_\mu^{(m)} p\rangle\} = 1 - \sum_{p \in \wp_s} \left(\frac{\mu}{a-1}\right)^L = 1 - \left(\frac{a\mu}{a-1}\right)^L.$$

Next, the inequality $\tau_\infty(M_\mu^{(m)}) \leqslant 1 - (a\mu/(a-1))^L$ can be seen as follows: Let $v \perp e$, such that $\|v\|_\infty = 1$, and $\tau_\infty(M_\mu^{(m)}) = \|M_\mu^{(m)} v\|_\infty$. Such a vector exists by compactness of $e^\perp \cap \mathrm{Ball}_\infty(\mathscr{V}_\wp)$. We may assume W.L.O.G., that $M_\mu^{(m)} v$ has a maximum modulus qth component, which is strictly positive. Then we have

$$\begin{aligned}
\tau_\infty(M_\mu^{(m)}) &= \langle q, M_\mu^{(m)} v\rangle = \sum_{p \in \wp_s} \langle q, M_\mu^{(m)} p\rangle \langle p, v\rangle \\
&= \sum_{p \in \wp_s} \left(\langle q, M_\mu^{(m)} p\rangle - \left(\frac{\mu}{a-1}\right)^L\right) \langle p, v\rangle \\
&\leqslant \sum_{p \in \wp_s} \left(\langle q, M_\mu^{(m)} p\rangle - \left(\frac{\mu}{a-1}\right)^L\right) = 1 - \left(\frac{a\mu}{a-1}\right)^L.
\end{aligned}$$

This shows Proposition 3.7.3. Finally, let us prove Proposition 3.7.4.

Let $p \in \wp_s$ be uniform. In order to produce a uniform population from p, one selects λ spots to be changed in the first creature of p, and then has to change $s \cdot \lambda$ corresponding spots in p. For the λ spots in the first creature, one has $(a-1)^\lambda$ combined choices for new letters. Thus, the probability of producing a uniform population from p is given by

$$\sum_{\lambda=0}^{\ell} \binom{\ell}{\lambda} (a-1)^\lambda \left(\frac{\mu}{a-1}\right)^{s\lambda} (1-\mu)^{s(\ell-\lambda)} = \left(\frac{\mu^s}{(a-1)^{s-1}} + (1-\mu)^s\right)^{\ell} \in [0,1].$$

Let $p \in P$ be non-uniform. Suppose that exactly σ_ι of the s creatures in p have $\hat{a}(\iota)$ as their first letter, $0 \leqslant \iota \leqslant a-1$. One then must change either $s - \sigma_0$, or $s - \sigma_1$, or $\ldots$ $s - \sigma_{a-1}$ spots in order to make the first letter in every creature agree. For this to happen, we obtain the probability

$$\begin{aligned}
\sum_{\iota=0}^{a-1} \left(\frac{\mu}{a-1}\right)^{s-\sigma_\iota} (1-\mu)^{\sigma_\iota} &= \left(\frac{\mu}{a-1}\right)^s \sum_{\iota=0}^{a-1} \left(\frac{(a-1)(1-\mu)}{\mu}\right)^{\sigma_\iota} \\
&\geqslant \left(\frac{\mu}{a-1}\right)^s a \left(\frac{(a-1)(1-\mu)}{\mu}\right)^h = \omega_u. \qquad (10)
\end{aligned}$$

It is an easy exercise in calculus using the gradient, and the Hessian with respect to $\sigma_1 \dots \sigma_{a-1}$ to show the last inequality in identity (10). The probability of generating a uniform population from p is bounded below by $(\omega_u)^\ell$.

Let $p \in \wp_s$ be such that $\Delta(p,q) = 1$ for some uniform $q \in \wp_s \cap \mathscr{U}$. Suppose W.L.O.G., that p and q differ at the first spot in the first creature, and the letter in question in p is $\hat{a}(0)$, the letter in q is $\hat{a}(1)$. Thus, disregarding the first spot in creatures, both p and q are uniform. If σ_ι is as above, then $\sigma_0 = 1$, $\sigma_1 = s-1$, and $\sigma_\iota = 0$ for $\iota \geqslant 2$. In this situation, the probability of producing a uniform population (not necessarily q) from p is given by

$$\left(\frac{\mu^{s-1}(1-\mu)}{(a-1)^{s-1}} + \frac{\mu(1-\mu)^{s-1}}{a-1} + (a-2)\left(\frac{\mu}{a-1}\right)^s \right) \cdot \left(\frac{\mu^s}{(a-1)^{s-1}} + (1-\mu)^s \right)^{(\ell-1)}. \tag{11}$$

For a number $t \in [1,\infty)$, and integers $n, m \in \mathbf{N}$, it is easy to show that $t^{n+1} + t^{m-1} \geqslant t^n + t^m$, if $n \geqslant m$. Using the latter fact, a discussion of the left-hand side in identity (11) – changing only two of the σ_ι by ± 1 at one step – shows that identity (10) lists the largest possible factors. Combining identities (10) and (11) yields Proposition 3.7.4 similar to the corresponding computation in the proof of Proposition 3.4.4. □

We note that the statement in [62, p. 112, Proposition 4.4] contains a misprint. In fact, the first exponent ℓ (in the left portion of the inequality shown) should be replaced by $\ell - 1$ corresponding to the exponent $\ell - 1$ in β_0. The proof given in [62] is correct showing $\ell - 1$.

Similar to Proposition 3.5, the following result is the key to handling multiple-spot mutation combined with the crossover operation.

3.8. Proposition. *We have* $M_\mu^{(m)} D \subseteq D$. *Consequently*, $M_\mu^{(m)} P_D = P_D M_\mu^{(m)} P_D = P_D M_\mu^{(m)}$. *The same statements hold, if* D *is replaced by* $\bar{D}$.

Proof. The matrix $M_{1/L}^{(1)}$ describes changing exactly one spot in a population with equal probability. Thus, the matrix describing changing exactly two spots is a linear combination of $\mathbf{1}$ (corresponding to changing a letter back), $M_{1/L}^{(1)}$ (corresponding to changing at the same spot twice, but not back), and $(M_{1/L}^{(1)})^2$. Continuing this argument by induction yields Proposition 3.8 from Proposition 3.5. □

4. Applications of contraction properties of mutation

Suppose that $T(t)$, $t \in \mathbf{N}$, is a sequence of stochastic matrices. Later, we shall set $T(t) = F_t C_{\chi(t)}^K$ where F_t is a stochastic operator modeling fitness selection, and $C_{\chi(t)}$ is an operator modeling crossover in a genetic algorithm. In this section, we are interested in contracting properties of an algorithm that can be modeled as $G_t = T(t) M_{\mu(t)}$, where $M_{\mu(t)}$ is either single- or multiple-spot mutation. First, we shall investigate weak

ergodicity of such an algorithm. Later, we shall investigate how such an algorithm contracts towards uniform populations.

4.1. Weak ergodicity of genetic algorithms

The next theorem shows that the inhomogeneous Markov chain describing a genetic algorithm using single-spot mutation is weakly ergodic under regular conditions such as: (i) the crossover rate leaves a small chance for copying of any population, and (ii) the fitness selection reproduces a given population with a fixed constant or sufficiently slowly decreasing non-zero probability.

4.1. Theorem. *Suppose that* $T(t)$, $t \in \mathbf{N}$, *is a sequence of stochastic matrices such that the diagonal of* $T(t)$ *is bounded away from* 0 *for all* $t \in \mathbf{N}$, i.e., *for* $t \in \mathbf{N}$ *there exists* $c(t) > 0$ *such that* $\langle p, T(t)p \rangle \geqslant c(t)$ *for all* $p \in \wp_S$. *Using single-spot mutation, let* $G_t = T(t)M^{(1)}_{\mu(t)}$. *Suppose that* $\mu(t) \in (0, (a-1)/(L(a-1)+1)]$, *and*

$$\sum_{t=0}^{\infty} \prod_{\hat{\lambda}=1}^{L} c(tL+\hat{\lambda})\mu(tL+\hat{\lambda}) = \infty.$$

Then the inhomogeneous Markov chain $\prod_{k=t}^{1} G_k = G_t \cdot G_{t-1} \cdots G_1$ *is weakly ergodic.*

Proof. We have $\mu/(a-1) \leqslant 1 - L\mu$. Since $(M^{(1)}_{\mu})^L$ is fully positive, the matrix

$$X_k = \prod_{k'=kL}^{L(k-1)+1} G_{k'},$$

$k \in \mathbf{N}$, is fully positive with every coefficient bounded below by

$$(a-1)^{-L} \prod_{k'=L(k-1)+1}^{kL} c(k')\mu(k').$$

As in the proof of Proposition 3.7.3, one shows using identity (4)

$$\tau_1(X_k) \leqslant 1 - a^L(a-1)^{-L} \prod_{k'=L(k-1)+1}^{kL} c(k')\mu(k').$$

Since the summation over the products is unbounded, we can use either [63, p. 137, Theorem 4.8] or [35, p. 151, Theorem V.3.2] to complete the proof. □

The following theorem generalizes a result by Suzuki [65, p. 60, Lemma 1, 66, p. 98, Lemma 1], and a result in [61, Section 3].

4.2. Theorem. *Suppose that* $T(t)$, $t \in \mathbf{N}$, *is a sequence of stochastic matrices. Using multiple-spot mutation, let* $G_t = T(t)M^{(m)}_{\mu(t)}$, $\mu(t) \in (0, (a-1)/a]$. *Suppose that*

$$\sum_{t=1}^{\infty} \mu(t)^L = \infty,$$

Then the inhomogeneous Markov chain $\prod_{k=t}^{1} G_k = G_t \cdot G_{t-1} \cdots G_1$ *is weakly ergodic.*

Proof. For the proof one combines either [63, p. 137, Theorem 4.8] or [35, p. 151, Theorem V.3.2], and Proposition 3.7.3. □

4.2. Contraction towards uniform populations

The following lemmas essentially show that the genetic algorithm asymptotically spends most time in uniform populations for small mutation rates. We shall need these technical results only for constant mutation rate $\mu = \mu(t)$. Assume in addition, that $T(t)$ is a sequence of stochastic matrices, which leave the linear span $\mathscr{U} \subset \mathscr{V}_\varphi$ over uniform populations invariant, and otherwise contract towards $\mathscr{U}$. In fact, the standard crossover operations do not change the proportions of non-uniform probabilities vs. uniform probabilities in a probability distribution over populations, i.e., an element in S. As we shall see in Section 7, the fitness operator usually contracts towards uniform populations. As outlined in detail in [61, Section 5], the latter is in most cases true with a fixed contraction coefficient $\theta < 1$. We are interested in estimates, that describe how products of the G_t as above shrink the non-uniform portion of a particular $v \in S$.

4.3. Lemma. *Let $T(t)$, $t \in \mathbf{N}$, be a sequence of stochastic matrices, and $\theta(t) \in [0,1)$ be such that*

- $T(t)P_{\mathscr{U}} = P_{\mathscr{U}}$,
- $\|(\mathbf{1} - P_{\mathscr{U}})T(t)v\|_1 \leqslant \theta(t)\|(\mathbf{1} - P_{\mathscr{U}})v\|_1$ *for every* $v \in S$.

Let $\mu \in [0, L^{-1})$, and $G_t = T(t)M_\mu^{(1)}$, i.e., *single-spot mutation is used, and by definition $M_0^{(1)} = \mathbf{1}$. Then we have*

1. $\|(\mathbf{1} - P_{\mathscr{U}})G_t v\|_1 \leqslant \theta(t)(L\mu + (1 - L\mu)\|(\mathbf{1} - P_{\mathscr{U}})v\|_1)$.
2. *In addition, we have for $t \in \mathbf{N}$:*
   - $\|(\mathbf{1} - P_{\mathscr{U}})\prod_{k=t}^{1} G_k v\|_1 \leqslant \|(\mathbf{1} - P_{\mathscr{U}})v\|_1 (1 - L\mu)^t \prod_{k=1}^{t} \theta(k) + L\mu \sum_{k=0}^{t-1} (1 - L\mu)^k \cdot \prod_{k'=t-k}^{t} \theta(k')$
   - *If $\theta(t) = \theta \in [0,1)$, then this implies* $\|(\mathbf{1} - P_{\mathscr{U}})\prod_{k=t}^{1} G_k v\|_1 \leqslant (1 - L\mu)^t \theta^t \|(\mathbf{1} - P_{\mathscr{U}})v\|_1 + L\mu\theta/(1 - (1 - L\mu)\theta)$.

Proof. We have

$$T(t) = P_{\mathscr{U}} + T(t)(\mathbf{1} - P_{\mathscr{U}}) = P_{\mathscr{U}} + P_{\mathscr{U}}T(t)(\mathbf{1} - P_{\mathscr{U}}) + (\mathbf{1} - P_{\mathscr{U}})T(t)(\mathbf{1} - P_{\mathscr{U}}).$$

Hence, $(\mathbf{1} - P_{\mathscr{U}})T(t) = (\mathbf{1} - P_{\mathscr{U}})T(t)(\mathbf{1} - P_{\mathscr{U}})$. Set $\gamma = L\mu$, and $\beta = 1 - L\mu$. The $\|\cdot\|_1$-inequality for $T(t)$ and Proposition 3.4.4 imply

$$\|(\mathbf{1} - P_{\mathscr{U}})G_t v\|_1 \leqslant \theta(t)\|(\mathbf{1} - P_{\mathscr{U}})M_\mu^{(1)} v\|_1 \leqslant \theta(t)(\gamma + \beta\|(\mathbf{1} - P_{\mathscr{U}})v\|_1).$$

This shows Lemma 4.3.1. Now we proceed by induction to obtain Lemma 4.3.2. □

4.4. Lemma. *Let $\gamma = \gamma(\mu)$ and $\beta = \beta(\mu)$ are as in Proposition* 3.7.4 *for $\mu > 0$. Let $\gamma(0) = 0$, $\beta(0) = 1$, and $M_0^{(m)} = \mathbf{1}$. Let $T(t)$, $t \in \mathbf{N}$, be a sequence of stochastic matrices, and $\theta(t) \in [0,1)$ be such that*

- $T(t)P_{\mathscr{U}} = P_{\mathscr{U}}$,
- $\|(\mathbf{1} - P_{\mathscr{U}})T(t)v\|_1 \leqslant \theta(t)\|(\mathbf{1} - P_{\mathscr{U}})v\|_1$ *for every* $v \in S$.

Let $\mu \in [0,(a-1)/a]$, *and* $G_t = T(t)M_\mu^{(m)}$, i.e., *multiple-spot mutation is used. Then we have*

1. $\|(\mathbf{1}-P_{\mathscr{U}})G_t v\|_1 \leqslant \theta(t)(\gamma+\beta\|(\mathbf{1}-P_{\mathscr{U}})v\|_1)$.
2. *In addition, we have for* $t \in \mathbf{N}$:
   - $\|(\mathbf{1}-P_{\mathscr{U}})\prod_{k=t}^{1} G_k v\|_1 \leqslant \|(\mathbf{1}-P_{\mathscr{U}})v\|_1 \beta^t \prod_{k=1}^{t}\theta(k) + \gamma\sum_{k=0}^{t-1}\beta^k\prod_{k'=t-k}^{t}\theta(k')$
   - *If* $\theta(t)=\theta\in[0,1)$, *then this implies* $\|(\mathbf{1}-P_{\mathscr{U}})\prod_{k=t}^{1} G_k v\|_1 \leqslant \beta^t\theta^t\|(\mathbf{1}-P_{\mathscr{U}})v\|_1 + \gamma\theta/(1-\beta\theta)$.

Proof. Same as the proof of Lemma 4.3. □

5. Crossover

We shall discuss two types of specific crossover operations in this section: (i) regular crossover which is probably the crossover operation mostly used in computer implementations of genetic algorithms (see the books by Goldberg [25, p. 64], Mitchell [48, p. 8], or Holland [33, p. 97]), and (ii) unrestricted crossover which mates randomly chosen pairs in the population. Both crossover operations are discussed extensively in [62, p. 113, Section 2.2]. Most results and proofs of [62, Section 2.2] carry over verbatim, and need not be repeated here. The main reason is that regular and unrestricted crossover do not change *letters* in $\mathscr{A}$ probabilistically, but do change *positions of letters*, a procedure which is not affected by enlarging the underlying alphabet.

Note that the definition of *elementary crossover operation* given in [62, p. 113] includes exchanging parents as creatures. This is included for mathematical convenience, but is not an essential restriction. If the reader wants to exclude the exchange of creatures as an elementary crossover operation, then he or she has to change the definitions of $\mathrm{Mean}_{\mathbf{r}}$ and $\mathrm{mean}_{\mathbf{u}}$ given in Section 2.10, Hardy–Weinberg spaces given in Section 2.11, and the diagonal spaces D and $\bar{D}$ given in Section 2.12. However, the results presented here in regard to convergence of genetic algorithms (see Section 8), and the result of enhancement of mutation by crossover in Section 6.1 still remain valid, if properly reformulated.

An elementary crossover operation[4] $C(\sigma_0,\sigma_1;\lambda)$, which exchanges "heads" of the genome of two creatures at two fixed locations $\sigma_0,\sigma_1\in[1,s]\cap\mathbf{N}$ in the population with fixed cutpoint $\lambda\in[1,\ell]\cap\mathbf{N}$, is (identified with) a stochastic matrix which is its own inverse. Thus, up to rearrangement of the basis $\wp_s$ of $\mathscr{V}_\wp$, the matrix for $C(\sigma_0,\sigma_1;\lambda)$ is a direct sum of a matrix $\mathbf{1}$ of proper dimension and flip matrices $\mathbf{f}$, and has spectrum in $\{-1,1\}$. (There are only more $\mathbf{f}$'s now, and the size of $\mathbf{1}$ is larger than in the binary alphabet case, cf. [62, p. 114, Lemma 5.1].) Thus, all arguments about spectra of stochastic matrices, that model the crossover operations, remain exactly the same as in [62].

[4] We note that the version of crossover operation used here introduces an asymmetric linkage as defined in Geiringer's publication [20, p. 33] between loci (spots) in the genome. In fact, if the genes (letters) of creatures in position λ are swapped by crossover, then also the letters in the positions $\lambda'\leqslant\lambda$ are exchanged.

In addition to the above, we shall discuss *generalized crossover* as introduced in [61, Section 4], a concept that includes both regular and unrestricted crossover. For the purpose of obtaining convergence theorems, the definition of *generalized crossover* is actually sufficient in many cases.

5.1. Generalized crossover

The important property of the crossover operation in regard to convergence of genetic algorithms is that it leaves uniform populations invariant, i.e., the change from uniform populations to non-uniform populations is driven solely by mutation. The following definition leaves a lot of room for custom designed crossover operations.

An *generalized crossover operation* is a map $\chi \mapsto C_\chi$, $\chi \in [0,1]$, into the set of symmetric stochastic matrices such that $C_\chi q = q$ for every uniform population $q \in \wp_s \cap \mathscr{U}$, and $\chi \in [0,1]$. If the map $\chi \mapsto C_\chi$ is continuous, then a generalized crossover operation is called continuous.

Note that we do not require mutation and crossover to commute.

5.1.1. Summary of results

[61, Section 4.1]. If C_χ, $\chi \in [0,1]$, is the stochastic matrix associated with a generalized crossover operation, then:

1. $C_\chi P_{\mathscr{U}} = P_{\mathscr{U}} = (P_{\mathscr{U}})^* = P_{\mathscr{U}} C_\chi$ and $C_\chi(\mathbf{1} - P_{\mathscr{U}}) = C_\chi - P_{\mathscr{U}} = (\mathbf{1} - P_{\mathscr{U}})C_\chi$.
2. $\chi \mapsto C_\chi^K$, $K \in \mathbf{N}$ fixed, is also a generalized crossover operation.

5.2. Regular crossover

Regular crossover or simple crossover is defined as in [25, p. 64], [48, p. 8], or [33, p. 97] – see [62, p. 114]. The creatures in an *even-size population* are paired sequentially. For every pair, it is decided with a certain probability $\chi \in [0,1]$, the *crossover rate*, whether crossover takes place. All cutpoints have equal probability including cutpoint $\lambda = \ell$ which corresponds to exchanging the positions of the parents but leaving the parents unaltered. Denote by $C_\chi^{(r)}$ the stochastic matrix corresponding to this process.

5.2.1. Summary of results

We have the following results for $C_\chi^{(r)}$ [62, p. 115, Proposition 7.1 & 4–7]:

1. If $p \in \wp_s$, then $\langle p, C_\chi^{(r)} p \rangle \geqslant (1-\chi)^{s/2}$.
2. $C_\chi^{(r)}$ commutes with both mutation operators described in Section 3.
3. Up to rearrangement of the basis $\wp_s$ of $\mathscr{V}_\wp$, the matrix $C_\chi^{(r)}$ decomposes into a block diagonal matrix with one block $C_{\chi,\xi}^{(r)}$ for every $\xi = \mathrm{Mean}_{\mathbf{r}}(p)$, $p \in \wp_s$, i.e., $C_{\chi,\xi}^{(r)}$ acts on $\bar{\mathscr{V}}_\xi$.
4. If $\chi \in (0,1)$, then for every $\xi = \mathrm{Mean}_{\mathbf{r}}(p)$, $p \in \wp_s$, the block $(C_{\chi,\xi}^{(r)})^\ell$ is fully positive. Consequently, $e_{\bar{\xi}}$ is up to scalar multiples the only eigenvector to eigenvalue 1 of $C_{\chi,\xi}^{(r)}$.
5. $\mathrm{sp}(C_\chi^{(r)}) \subseteq [-1+\hat{\alpha}, 1-\hat{\alpha}] \cup \{1\}$, where $\hat{\alpha} = \ell^{-s/2}(1 - |1-2\chi|)$.

5.2.2. Regular crossover in the multi-set model

We note that

$$\Lambda^{-1}P_\Pi C_\chi^{(r)}P_\Pi\Lambda = \Lambda^{-1}P_\Pi C_\chi^{(r)}\Lambda$$

models the crossover operation for the multi-set model. In fact for a population $p\in\wp_s$, the expression $P_\Pi p$ models equal probability for arbitrary arrangements of creatures in p. $C_\chi^{(r)}$ models applying the crossover procedure to sequentially paired creatures, i.e., $C_\chi^{(r)}$ pairs creatures at positions $(1,2),(3,4),\ldots,(s-1,s)$ in the population. Thus, $C_\chi^{(r)}P_\Pi$ models pairing creatures at arbitrarily chosen disjoint pairs of positions for creatures in the population. Finally, $P_\Pi C_\chi^{(r)}P_\Pi$ models pairing creatures at arbitrarily chosen disjoint pairs of positions in the population, and afterwards rearranging the resulting population arbitrarily.

The relevant invariant subspaces of $P_\Pi C_\chi^{(r)}P_\Pi$ are the Hardy–Weinberg spaces $\mathscr{V}_\xi$, $\xi=\mathrm{mean}_{\mathbf{u}}(p)$, $p\in\wp_s$. Thus up to rearrangement of the basis $\wp_s$ of $\mathscr{V}_\wp$, the matrix $P_\Pi C_\chi^{(r)}P_\Pi$ decomposes into a block diagonal matrix with one block $(P_\Pi C_\chi^{(r)}P_\Pi)_\xi$ for every $\xi=\mathrm{mean}_{\mathbf{u}}(p)$, $p\in\wp_s$, i.e., $(P_\Pi C_\chi^{(r)}P_\Pi)_\xi$ acts on $\mathscr{V}_\xi$.

5.1. Lemma. *If $\chi\in(0,1)$, then for every $\xi=\mathrm{mean}_{\mathbf{u}}(p)$, $p\in\wp_s$, the block $((P_\Pi C_\chi^{(r)}P_\Pi)_\xi)^{(s-1)\ell}$ is fully positive. Consequently, e_ξ is up to scalar multiples the only eigenvector to eigenvalue 1 of $(P_\Pi C_\chi^{(r)}P_\Pi)_\xi$.*

Proof. See a population as a $\ell\times s$ matrix for a moment. Using at most $s-1$ transpositions (from left to right) for every spot $\lambda\in[1,\ell]\cap\mathbf{N}$ in creatures (from bottom to top) allow to produce any population $q\in\wp_s$ from a population $p\in\wp_s$ satisfying $\mathrm{mean}_{\mathbf{u}}(q)=\mathrm{mean}_{\mathbf{u}}(p)$. For such p and q as above, let $C(\sigma_v,\sigma_v';\lambda_v)$, $1\leqslant v\leqslant v_0\leqslant(s-1)\ell$ be a sequence of elementary crossover operations needed to transform p into q, i.e.,

$$q=\prod_{v=1}^{v_0}C(\sigma_v,\sigma_v';\lambda_v)p.$$

Any transposition $\pi=\pi_{\sigma_v,\sigma_v'}\in\Pi_s$ corresponding to $C(\sigma_v,\sigma_v';\lambda_v)$ can be written as $\pi=\tilde{\pi}\pi_{1,2}\tilde{\pi}$ for a suitable permutation $\tilde{\pi}=\tilde{\pi}^{-1}\in\Pi_s$ where $\pi_{1,2}\in\Pi_s$ denotes the transposition exchanging 1 and 2. Since $\chi<1$, a regular crossover operation may only involve swapping heads of the first two creatures in a population at cutpoint λ_v. Consequently,

$$P_\Pi C_\chi^{(r)}P_\Pi = r_v\cdot C(\sigma_v,\sigma_v';\lambda_v)+T_v,$$

where $r_v>0$, and T_v is a positive matrix for every $1\leqslant v\leqslant v_0$. Hence, $\langle q,(P_\Pi C_\chi^{(r)}P_\Pi)^{(s-1)\ell}p\rangle>0$. □

5.3. Unrestricted crossover

In unrestricted crossover, a single elementary crossover operation $C(\sigma_0,\sigma_1;\lambda)$ takes place with probability $\chi\in[0,1]$. The probability for all triples $(\sigma_0,\sigma_1;\lambda)$ is the same. Denote by $C_\chi^{(u)}$ the stochastic matrix corresponding to this process.

It is interesting to note that unrestricted crossover is related to a family of representations of Π_s on $\mathscr{V}_{\wp}$. This can be used effectively to compute bounds for the spectrum, c.f. [62, p. 118, Lemma 8].

5.3.1. Summary of results

We have the following results for $C_\chi^{(u)}$ [62, p. 118, Proposition 9.1 & 4–7]:

1. If $\chi<1$, then the diagonal of $C_\chi^{(u)}$ is strictly positive with lower bound $1-\chi$. If $s>a$, then the diagonal of $C_\chi^{(u)}$ is strictly positive with lower bound independent from the crossover rate χ.
2. $C_\chi^{(u)}$ commutes with both mutation operators described in Section 3. $C_\chi^{(u)}$ commutes with every $\pi\in\Pi_s$. Thus, it commutes with P_Π.
3. Up to rearrangement of the basis $\wp_s$ of $\mathscr{V}_{\wp}$, the matrix $C_\chi^{(u)}$ decomposes into a block diagonal matrix with one block $C_{\chi,\xi}^{(u)}$ for every $\xi=\mathrm{mean}_{\mathbf{u}}(p)$, $p\in\wp_s$, i.e., $C_{\chi,\xi}^{(u)}$ acts on $\mathscr{V}_\xi$.
4. For every $\xi=\mathrm{mean}_{\mathbf{u}}(p)$, $p\in\wp_s$, the block $(C_{\chi,\xi}^{(u)})^{(s-1)\ell}$ is fully positive. Consequently, e_ξ is up to scalar multiples the only eigenvector to eigenvalue 1 of $C_{\chi,\xi}^{(u)}$.
5. $\mathrm{sp}(C_\chi^{(u)})\subseteq[1-2\chi+\hat{\alpha},1-\hat{\alpha}]\cup\{1\}$, where $\hat{\alpha}=\ell^{-1}\chi(1-\alpha(s))$.
   Here, $\alpha(s)$ is the second largest eigenvalue of the operator

$$\Gamma_s=2(s(s-1))^{-1}\sum_{\pi\in T_s}\pi,$$

   in the left regular representation of Π_s (see, e.g., Pedersen's book [52, Chapter 7]), and $T_s\subset\Pi_s$ is the set of all transpositions in Π_s. An open conjecture is that $\alpha(s)=(s-3)(s-1)^{-1}$, c.f. [62, p. 131]. It is not hard to show $\alpha(s)\geqslant(s-3)(s-1)^{-1}$ representing Π_s on $\mathbf{C}^s$ canonically, and using that no new eigenvalues can occur in representations. (One obtains a matrix that looks like $k_1\mathbb{1}+k_2\hat{m}^{(1)}$.)

Proof. We note that if $s>a$, then at least two creatures in a population have equal first letter. Hence, a suitably chosen elementary crossover operation leaves the population invariant. This shows the second claim in Section 5.3.1.1. All other proofs are as in [62], except the proof of Section 5.3.1.4. Section 5.3.1.4 is already established in the proof of Lemma 5.1. □

5.4. On Geiringer's theorem

In this section, we shall derive an analogue of Geiringer's theorem [20, p. 42, Theorem III] for the two types of crossover considered in Sections 5.2.2 resp. in 5.3. See also [72, p. 287, Theorem 3.9] for a related result. In fact, we shall establish the result for *finite* populations using essentially only Lemma 5.1 resp. Section 5.3.1.4. The result says that, in the mean, the distribution of creatures in a population subject to iterated crossover "converges towards independence", i.e., it converges to a canonical product distribution.

Fix an initial population $p \in \wp_s$. Let $\xi = (\xi_\lambda)_{\lambda=1}^{\ell} = \mathrm{mean}_{\mathbf{u}}(p)$. For $1 \leqslant \lambda \leqslant \ell$, let

$$\xi_\lambda = \sum_{\iota=0}^{a-1} \xi_{\lambda,\iota} \cdot \hat{a}(\iota)$$

For every letter in $\mathscr{A} = \{\hat{a}(\iota) \mid 0 \leqslant \iota \leqslant a-1\}$, let

$$\mathrm{pr}(\hat{a}(\iota), \lambda, p) = \xi_{\lambda,\iota}. \tag{12}$$

denote the probability of finding $\hat{a}(\iota)$ in the λth spot in creatures of p.

5.2. Theorem. *Let $c = (\hat{a}(\iota_1), \ldots, \hat{a}(\iota_\ell)) \in \mathscr{C}$, $\hat{a}(\iota_\lambda) \in \mathscr{A}$, $1 \leqslant \lambda \leqslant \ell$. Let C denote either the crossover operation $P_{II} C_\chi^{(r)} P_{II}$ considered in Section 5.2.2, or let C denote $C_\chi^{(u)}$ considered in Section 5.3. Then the probability of finding c*

(1) *in the sequence of populations resulting from iterated application of the crossover operation C to a fixed, initial population $p \in \wp_s$, or*
(2) *in a high generation under this procedure converges in the mean over repeated applications of this procedure to*

$$\prod_{\lambda=1}^{\ell} \mathrm{pr}(\hat{a}(\iota_\lambda), \lambda, p),$$

where pr *is defined in identity* (12).

Proof. For $\ell = 1$, finding a certain creature $(\hat{a}(\iota_1))$, i.e., letter, in a population which was subject to multiple crossover operations (swaps) equals the initial probability $pr(\hat{a}(\iota_1), 1, p)$ of finding it. We shall proceed now by induction over $\ell \in \mathbf{N}$. Suppose, that the result is true for $\ell - 1$.

Let $p \in \wp_s$, and $c = (\hat{a}(\iota_1), \ldots, \hat{a}(\iota_\ell)) \in \mathscr{C}$ be fixed. We shall use that in accordance with Lemma 5.1 resp. Section 5.3.1.4 in the limit, every population $q \in \wp_s$ such that $\mathrm{mean}_{\mathbf{u}}(q) = \xi = \mathrm{mean}_{\mathbf{u}}(p)$ has equal probability of occurrence. In addition, we shall use [59, p. 10, Theorem 3.1] showing that the averaged iterates of the crossover operation converge to the projection onto the fixed point space. Let $\sigma' = s \cdot \xi_{\ell,\iota_\ell} = s \cdot \mathrm{pr}(\hat{a}(\iota_\ell), \ell, p)$, and assume W.L.O.G., $\sigma' \geqslant 1$. Define

$$\tilde{c} = (\hat{a}(\iota_1), \ldots, \hat{a}(\iota_{\ell-1})).$$

For a population $q = (d_1, d_2, \ldots, d_s) \in \wp_s$, $d_\sigma \in \mathscr{C}$, $1 \leqslant \sigma \leqslant s$, let

$$\tilde{q} = (\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_s).$$

Now define $n(\sigma)$ to be the number of populations $q \in \wp_s$ such that $\mathrm{mean}_{\mathbf{u}}(q) = \xi = \mathrm{mean}_{\mathbf{u}}(p)$, and $\tilde{c}$ occurs σ times in $\tilde{q}$. Let N_ξ be the number of populations $q \in \wp_s$ such that $\mathrm{mean}_{\mathbf{u}}(q) = \xi = \mathrm{mean}_{\mathbf{u}}(p)$. By induction hypothesis, we have for the number $N_{\tilde{c}}$ of creatures d satisfying $\tilde{d} = \tilde{c}$ in such populations

$$N_{\tilde{c}} = \sum_{\sigma=1}^{s} \sigma \cdot n(\sigma) = (sN_\xi) \cdot \prod_{\lambda=1}^{\ell-1} \mathrm{pr}(\hat{a}(\iota_\lambda), \lambda, p). \tag{13}$$

For a moment, consider the case $a=2$, i.e., binary alphabet, and $\ell=2$. In addition, consider the following population:

$$p_0 = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 & 0 & 0 & \dots & 0 \\ \hat{a}(\imath_1) & \hat{a}(\imath_2) & \hat{a}(\imath_3) & \dots & \hat{a}(\imath_\sigma) & \hat{a}(\imath_{\sigma+1}) & \hat{a}(\imath_{\sigma+2}) & \dots & \hat{a}(\imath_s) \end{pmatrix}, \tag{14}$$

where the creatures are understood as the columns in the matrix shown in identity (14). Suppose that $\mathrm{mean}_{\mathbf{u}}(p_0)=s^{-1}(\sigma,\sigma')$. We shall determine the number N_0 of creatures $c_0=\binom{1}{1}$ in this type of population. We claim that for $\sigma_-=\max\{0,\sigma'-\sigma\}$, and $\sigma_+=\min\{\sigma'-1,s-\sigma\}$

$$N_0 = \sum_{\kappa=\sigma_-}^{\sigma_+} (\sigma'-\kappa)\cdot\binom{\sigma}{\sigma'-\kappa}\cdot\binom{s-\sigma}{\kappa} = \frac{\sigma'}{s}\cdot\sigma\cdot\binom{s}{\sigma'}. \tag{15}$$

Note that the number of creatures with first component 1 in populations as in identity (14) is given by $\sigma\cdot\binom{s}{\sigma'}$. The fact, that N_0 equals the middle sum in identity (15)) is seen as follows: For every $\kappa=\sigma_-\dots\sigma_+$, one can pick $\sigma'-\kappa$ spots left of $\hat{a}(\imath_{\sigma+1})$ in p where to set $\hat{a}(\imath_{\tilde{\sigma}})=1$, and one has to pick κ spots right of $\hat{a}(\imath_\sigma)$ to maintain the desired $\mathrm{mean}_{\mathbf{u}}(p_0)=s^{-1}(\sigma,\sigma')$. In each case, this yields $\sigma'-\kappa$ copies of c_0. Using the definition of binomial coefficients, the latter equality in identity (15) is equivalent to the following:

$$\sum_{\kappa=\sigma_-}^{\sigma_+}\binom{\sigma-1}{\sigma'-1-\kappa}\cdot\binom{s-\sigma}{\kappa} = \binom{s-1}{\sigma'-1}.$$

The latter known identity is obtained by determining the coefficient of $x^{\sigma'-1}y^{s-\sigma'}$ in $(x+y)^{\sigma-1}\cdot(x+y)^{s-\sigma}$ and $(x+y)^{s-1}$.

In the case that $a>2$, i.e., non-binary alphabet, and $\ell=2$, we can use identity (15) to obtain

$$N_0 = \frac{\sigma'}{s}\cdot K\cdot\sigma\cdot\binom{s}{\sigma'}, \tag{16}$$

where K is a factor representing the possible arrangements of letters not equal to 1 in both rows of p_0. In any case, the number of creatures being identical to c_0 in populations of type p_0 is σ'/s times the number of creatures containing 1 as first letter.

Now, we only have to reinterpret identity (16) to complete the proof of Theorem 5.2. See the 1s in the first row as copies of $\tilde{c}$, and assume W.L.O.G., that $\hat{a}(\imath_\ell)=1$. Considering the possible subsets of $s-\sigma$ creatures $\tilde{d}\neq\tilde{c}$ in the first row, identity (16) says that in order to compute the number N_c of copies of c in all populations $q\in\wp_s$ such that $\mathrm{mean}_{\mathbf{u}}(q)=\xi=\mathrm{mean}_{\mathbf{u}}(p)$, one has to multiply every term in the middle of identity (13), i.e., in the sum, by σ'/s. Hence, $N_c=N_{\tilde{c}}\cdot\sigma'/s$. □

6. The mutation-crossover proposal matrix

6.1. Contraction properties of combined mutation-crossover

The following result is – in the author's opinion – the key result showing how crossover enhances the mutation operation in the random generator phase of a genetic algorithm. If $v \in S$, then $P_e v = e$. Now, let

$$v = e + d(v) + o(v),$$

where

$$d(v) = P_D(\mathbf{1} - P_e)v = (P_D - P_e)v = (\mathbf{1} - P_e)P_D v$$

and consequently $o(v) = (\mathbf{1} - P_D)v$

if unrestricted crossover is considered, and

$$d(v) = P_{\tilde{D}}(\mathbf{1} - P_e)v = (P_{\tilde{D}} - P_e)v = (\mathbf{1} - P_e)P_{\tilde{D}} v$$

and consequently $o(v) = (\mathbf{1} - P_{\tilde{D}})v$,

if regular crossover is considered.

With the notation just defined, we have the following result:

6.1. Theorem. *Let M denote either mutation operation considered in Section* 3. *If $M = M_\mu^{(1)}$, then suppose $\mu < (a-1)/La$. Let C denote either crossover operation considered in Section* 5, *and let $\hat{\alpha}$ be as in Section* 5.2.1.5 *for regular crossover, and Section* 5.3.1.5 *for unrestricted crossover. Let $v \in S$, and $v' = MC^K v$ for $K \in \mathbf{N}$. Then we have*

1. $d(v') = Md(v)$, *and* $o(v') = MC^K o(v)$.
2. $\|d(v')\|_2 \leqslant (1 - \frac{\mu a}{a-1})\|d(v)\|_2$, *and* $\|o(v')\|_2 \leqslant (1 - \frac{\mu a}{a-1})\hat{\alpha}^K \|o(v)\|_2$.

Proof. In accordance with Proposition 3.3.3 and Proposition 3.6.3 the contracting factor (the second largest eigenvalue) for M is $1 - \mu a/(a-1)$. After noticing this fact, the proof continues as the proof of [60, p. 119, Proposition 10]. □

Mutation-crossover is seen in some treatise as an operation which preserves and collects schemata (see Holland's book [33, Chapter 4]). Here, it is shown that crossover accelerates the averaging processes by the mutation operation. Heuristically speaking, the enhanced averaging property of mutation-crossover "spreads" out the covered search space faster. Note that as a consequence of this proof, any vector in $e^\perp$ is shrunk by MC^K by at least a factor $1 - \mu a/(a-1)$ in its $\|\cdot\|_2$-norm.

Many researchers have investigated the mutation-crossover matrix for regular crossover at the multi-set level (see, e.g., [14–16, 25, 29, 34, 50, 65, 66, 68–70]). Let us show how one can derive spectral estimates with the methods presented here. Observe that on the multi-set level, the contraction process described in Theorem 6.1 is actually

accelerated in regard to crossover. Note that these bounds correspond up to a factor $\frac{1}{2}$ to eigenvalues associated with the Vose–Liepins conjecture [36, p. 411].

6.2. Theorem. *Let $M_0 = \Lambda^{-1} M_\mu^{(m)} \Lambda$ denote the stochastic matrix acting on $\mathscr{W}_\wp$ describing multiple-spot mutation in the multi-set model for genetic algorithms. Let $C_0 = \Lambda^{-1} P_\Pi C_\chi^{(r)} P_\Pi \Lambda$ denote the stochastic matrix acting on $\mathscr{W}_\wp$ describing the regular crossover operation in the multi-set model. Then we have,*

1. $\mathrm{sp}(M_0 C_0) \subseteq \mathrm{sp}(M_\mu^{(m)} P_\Pi C_\chi^{(r)} P_\Pi)$.
2. $|\mathrm{sp}(M_0 C_0)| \subseteq [0, (1-\mu a/(a-1))^2] \cup \{1-\mu a/(a-1), 1\}$.

Proof. We first note that by Proposition 3.6.5

$$M_\mu^{(m)} \Lambda = M_\mu^{(m)} P_\Pi \Lambda = P_\Pi M_\mu^{(m)} P_\Pi \Lambda \quad \text{and} \quad M_0 = \Lambda^{-1} P_\Pi M_\mu^{(m)} P_\Pi \Lambda = \Lambda^{-1} M_\mu^{(m)} \Lambda.$$

Let $w_0 \in \mathscr{W}_\wp$ be an eigenvector to eigenvalue t of $M_0 C_0$. Let $w = P_\Pi w = \Lambda w_0 \in \mathscr{V}_\wp$. We obtain

$$tw = P_\Pi \Lambda t w_0 = P_\Pi \Lambda M_0 C_0 w_0 = P_\Pi \Lambda M_0 \Lambda^{-1} \Lambda C_0 \Lambda^{-1} P_\Pi w = M_\mu^{(m)} P_\Pi C_\chi^{(r)} P_\Pi w.$$

This shows Theorem 6.2.1.

Suppose W.L.O.G., that $\mu < (a-1)/a$. For $\mu = (a-1)/a$, $M_0 C_0$ models the random restart of the algorithm and equals the projection $\Lambda^{-1} P_e \Lambda$, where $e = e_{\wp_\mathscr{A}}$.

As in Lemma 2.1, we denote $e = e_{\wp_\mathscr{A}} \in \mathscr{V}_1$ in the remainder of this proof (using the definitions in the notation section for $\ell = s = 1$). Let

$$x(z, \hat{\lambda}) = e \otimes e \ldots \otimes e \otimes z \otimes e \otimes e \ldots \otimes e.$$

where $z \perp e$ for $\hat{\lambda} \in [1, L] \cap \mathbf{N}$ can be chosen from $a-1$ orthonormal base vectors in $B(\hat{m}^{(1)}) \subset \mathscr{V}_1$, and z occurs at the $\hat{\lambda}$th tensor spot. It is easy to see that $x(z, \hat{\lambda})$ is a non-zero eigenvector to eigenvalue $1-\mu a/(a-1)$ of $M_\mu^{(m)}$, and that every eigenvector to this eigenvalue is a linear combination of orthogonal eigenvectors of this type. Using Lemma 2.1.1, we see that $P_\Pi x(z, \hat{\lambda}) \in \bar{D}$.

Note that $M_\mu^{(m)} P_\Pi C_\chi^{(r)} P_\Pi$ is a symmetric stochastic matrix, since $M_\mu^{(m)}$ commutes with both P_Π by Proposition 3.6.5 and $C_\chi^{(r)}$ by Section 5.2.1.2.

Suppose that $(1-\mu a/(a-1))t > (1-\mu a/(a-1))^2$ is an eigenvalue of $P_\Pi C_\chi^{(r)} P_\Pi M_\mu^{(m)}$, where $t < 1$. Let $y \in \mathscr{V}_\wp$ be an eigenvector of $P_\Pi C_\chi^{(r)} P_\Pi M_\mu^{(m)}$ to eigenvalue $(1-\mu a/(a-1))t$. Since $M_\mu^{(m)}$ and $P_\Pi C_\chi^{(r)} P_\Pi$ commute, they have a joint spectral decomposition by, e.g., [56, p. 306, Theorem 12.22], and y must be an eigenvector of both $M_\mu^{(m)}$ and $P_\Pi C_\chi^{(r)} P_\Pi$. Thus, y must be a linear combination of orthogonal vectors of the form $x(z, \hat{\lambda})$ as above. In addition, we must have

$$\begin{aligned} P_\Pi y &= P_\Pi \left(\left(1 - \frac{\mu a}{a-1} \right) t \right)^{-1} M_\mu^{(m)} P_\Pi C_\chi^{(r)} P_\Pi y \\ &= \left(\left(1 - \frac{\mu a}{a-1} \right) t \right)^{-1} M_\mu^{(m)} P_\Pi C_\chi^{(r)} P_\Pi y = y. \end{aligned}$$

Hence, $y \in \bar{D}$. Since $C_\chi^{(r)}$ acts identically on $\bar{D}$, we have $M_\mu^{(m)} P_\Pi C_\chi^{(r)} P_\Pi y = M_\mu^{(m)} y = (1 - \mu a/(a-1))y$, which yields a contradiction. The case of a negative eigenvalue with modulus larger that $(1 - \mu a/(a-1))^2$ is treated similarly. This shows that the third largest modulus of eigenvalues of $M_0 C_0$ must be bounded by $(1 - \mu a/(a-1))^2$. □

6.2. Convergent simulated annealing type algorithms

Simulated annealing is a well-established technique using a probabilistic algorithm solve, e.g., large combinatorial optimization problems. It has been extensively studied in the literature – see, e.g., [1, 10, 11, 21, 31, 46, 49]. In this section, we shall discuss several convergent simulated annealing type algorithms generalizing previously proposed algorithms by Mahfoud and Goldberg in [44, p. 304, 305], and in [62, p. 124, Remark on simulated annealing], where variations of mutation-crossover are used as *generator matrix*. What we consider here are different algorithms than Boltzmann selection as described in Michell's book [48, pp. 168–169], and Goldberg's publication [24, p. 449]. Michell [48, pp. 168–169] seems to describe Boltzmann selection as a genetic algorithm with exponentially rescaled fitness function and proportional selection. (The expectation values for the newly selected creatures only depend upon the current population.) Goldberg [24, p. 449] proposed a tournament selection mechanism (see also Section 7.2) which is shown [24, Section 3] to have the usual steady-state Boltzmann distribution over *all* possible creatures.

In general, a simulated annealing type algorithm supposes a function $E : \Upsilon \to \mathbf{R}$ defined on a set Υ of states (candidate solutions). The goal of the algorithm is to *minimize* E on the domain Υ. We shall call E the *inverse fitness function* or *energy function*. We shall set $\Upsilon = \wp_S$ in the discussion below.

For *temperature* $\theta \in \mathbf{R}_*^+$, let $A(\theta)$ be the matrix of *acceptance probabilities* describing transition from the current state $p \in \Upsilon$ in the course of the simulated annealing type algorithm to newly generated (proposed) state $q \in \Upsilon$. By definition, $A(\theta)$ is given by

$$\langle q, A(\theta) p \rangle = 1 \quad \text{if } E(q) \leqslant E(p)$$

and

$$\langle q, A(\theta) p \rangle = \exp(\theta^{-1}(E(p) - E(q))) \text{ otherwise.} \tag{17}$$

In addition to the above, a simulated annealing type algorithm requires a symmetric, stochastic *generator matrix* or *proposal matrix* G (see, e.g., [1, p. 36], [49, p. 752]), which determines the probability of proposed state $q \in \Upsilon$ being generated from the current state $p \in \Upsilon$ in the course of the algorithm. Altogether, the simulated annealing type algorithm consists of the following steps:

1. Initialize a state $p \in \Upsilon$. $(t = 1)$.
2. Generate a new (proposed) state $q \in \Upsilon$ from the current state $p \in \Upsilon$ probabilistically, where the probability of generating q from p is given by $\langle q, Gp \rangle$.

3. Accept q as the new state with probability $\langle q, A(\theta(t))p\rangle$, $t \in \mathbf{N}$, where the cooling schedule $t \mapsto \theta(t)$ must be appropriately chosen (see the publication by Aarts and van Laarhoven [1, pp. 40–41] for an overview of results on cooling schedules).
4. If a certain stopping criterion is satisfied, then halt the algorithm. Otherwise, increment $t \leftarrow t+1$, and proceed to step (2).

Suppose now, that there is a given fitness function $f : \mathscr{C} \to \mathbf{R}^+$ defined on the set of creatures $\mathscr{C}$, which is supposed to be maximized. In [44, p. 305], Mahfoud and Goldberg propose a regular simulated annealing algorithm with the following characteristics: Define the energy $E(p) = -\sum_{\sigma=1}^{s} f(c_\sigma)$ for a population $p = (c_1, c_2, \ldots, c_s) \in \wp_s$, $c_\sigma \in \mathscr{C}$, $1 \leqslant \sigma \leqslant s$. In addition, use the following method to generate the proposed state $q \in \wp_s$ for the next generation:

2(a) Randomly chose an index $\sigma \in [1, s-1] \cap \mathbf{N}$, and a cutpoint $\lambda \in [1, \ell] \cap \mathbf{N}$.
2(b) Apply the elementary crossover operation $C(\sigma, \sigma+1; \lambda)$ to the current population $p \in \wp_s$ with crossover rate $\chi \in (0,1)$.
2(c) Apply multiple spot mutation to the spots in c_σ and $c_{\sigma+1}$ with mutation rate $\mu \in (0, (a-1)/a)$ to obtain q.

Let $G_2 = M_\mu^{(m)} \cdot C_\chi^{(u)}$ be the product of multiple-spot mutation and unrestricted crossover in the case $s = 2$. Then G_2 is a symmetric matrix by Section 5.3.1.2. The generator matrix associated with the algorithm proposed by Mahfoud and Goldberg in [44, p. 305] is consequently given by

$$G = P_\Pi^k \cdot \left((s-1)^{-1} \sum_{\sigma=1}^{s-1} \mathbf{1} \otimes \mathbf{1} \ldots \otimes \mathbf{1} \otimes G_2 \otimes \mathbf{1} \otimes \mathbf{1} \ldots \otimes \mathbf{1} \right) \cdot P_\Pi^k$$

$$= P_\Pi^k \cdot \left((s-1)^{-1} \sum_{\sigma=1}^{s-1} G_2[\sigma] \right) \cdot P_\Pi^k, \tag{18}$$

where there are $\sigma - 1$ operators $\mathbf{1}$ to the left of G_2, and $s - \sigma - 1$ operators $\mathbf{1}$ to the right of G_2 in a term in the sum for $\sigma = 1, \ldots, s-1$. In addition, $k = 1$ if the population is remixed randomly in every step, or represented as multiset, $k = 0$ otherwise. Identity (18) shows that G is symmetric.

We note that Mahfoud and Goldberg [44] also propose a parallel algorithm with creature-wise simulated annealing type acceptance scheme and mutation-crossover as generator matrix on the population level. However, a proof of convergence for this algorithm is not given in [44].

Generalizing [62, p. 124, Remark on simulated annealing] and some of the above, we set for $p = (c_1, \ldots, c_s) \in \wp_s$, $c_\sigma \in \mathscr{C}$, and fixed $r \in [1, \infty]$

$$E(p) = -||(f(c_1), \ldots, f(c_s))||_r \tag{19}$$

to define an energy function on populations. Using the Hamming or ℓ^1-norm, i.e., $r = 1$, corresponds to a conservative approach where the fitness of the population is averaged, and the algorithm (slowly) seeks to improve the overall fitness in the population. Setting $r > 1$ corresponds to a more greedy strategy, where populations with one or several above-average elements are favored. For large r, this resembles the strategy of

messy genetic algorithms (see, e.g., publications by Deb, Goldberg and Korb [17, 28]): creatures with large fitness dominate the energy value of the population while creatures with low fitness can be considered as a "hidden" reservoir of promising genetic material.

If the energy of a population is defined as in identity (19), then we can use either the Mahfoud–Goldberg generator matrix as in identity (18), or any of the six operators of type $G = M \cdot C$ as a generator matrix, where M stands for single/multiple spot mutation considered in Section 3, and C denotes either regular crossover $P_{\Pi} C_{\chi}^{(r)} P_{\Pi}$, or the unrestricted crossover operators $C_{\chi}^{(u)}$ or $C_{\chi}^{(u)} P_{\Pi}$ considered in Section 5.

We note that the proof of convergence as described in [1, pp. 37–38] can be modified without much difficulty to allow for the following alternate of step (2) of the simulated annealing type algorithm with mutation-crossover as generator mechanism:

2(a) Let $\mu_0 > 0$ be fixed. If single-spot mutation is used, then chose at time t a mutation rate $\mu(t) \in [\mu_0, L^{-1} - \mu_0] \neq \emptyset$; otherwise, chose at time t a mutation rate $\mu(t) \in [\mu_0, (a-1)/a] \neq \emptyset$. Chose at time t a crossover rate $\chi(t) \in [0, \chi_0]$, $\chi_0 \leqslant 1$. Use a suitable continuous generalized crossover operation C_{χ} which has strictly positive diagonal entries for all $\chi \in [0, \chi_0]$, and commutes with mutation. The crossover operators mentioned in the last paragraph all satisfy these conditions for suitable χ_0. Let the symmetric, generator matrix be given by $G(t) = M_{\mu(t)} \cdot C_{\chi(t)}$, $t \in \mathbf{N}$.

2(b) Generate a new (proposed) state $q \in \wp_s$ from the current state $p \in \wp_s$ probabilistically, where the probability of generating q from p is given by $\langle q, G(t)p \rangle$.

One shows weak ergodicity of the Markov chain underlying the algorithm by using an appropriate monotone cooling schedule, the bounds on the coefficients of the generator matrices imposed in (2.a) above, and [63, p. 137, Theorem 4.8] or [35, p. 151, Theorem V.3.2]. Strong ergodicity then follows from [35, p. 160, Theorem V.4.3], formula (5) in [1, p. 37, Theorem 1], and the monotonity of the cooling schedule which allows to use the usual telescoping-sum argument for summability. Formula (5) in [1, p. 37, Theorem 1] shows the steady-state distribution for an individual step of the algorithm. In the limit $t \to \infty$, it also shows the convergence to optima. The algorithm introduced above allows dynamic bursts of mutation-crossover rates. Bursts of mutation-crossover rates in the generator matrix allow dynamic scaling of the "neighborhood" of the current state, and thus allow acceleration of convergence of an algorithm should the latter repeatedly return to a local minimum due to insufficient "size" of the local neighborhood of the current state.

7. Selection

Fitness selection in genetic algorithms models reproductive success of individuals or creatures in their environment. In many computer applications, the selection pressure is modeled in the following way: given a finite collection $\mathscr{C}$ of creatures in a model "world", a *fitness function* $f_0 : \mathscr{C} \to \mathbf{R}^+$ is defined which determines the chances of survival for $c \in \mathscr{C}$ in every step of the genetic algorithm. In our setting, $\mathscr{C}$ is the set

of words of length ℓ over the alphabet $\mathscr{A}$. A genetic algorithm becomes a function optimizer, if the task is then to find an element $c \in \mathscr{C}$ such that $f_0(c)$ is maximal.

Implementations of genetic algorithms mostly use the fitness selection strategies listed next. See [61, Section 5] for a more detailed discussion.

7.1. Proportional fitness selection for population-dependent fitness functions

Proportional fitness selection has probably been investigated the most as part of genetic algorithms. See, e.g., [15, 16, 29, 57, 65, 66, 69], [62, Section 2.3]. We shall list here only a few facts.

At the population level, the fitness operator $F_{\mathrm{PFS}}^{(f_0)}(t)$, $t \in \mathbf{N}$, yields a new population from the current population, based on a fitness-proportional probability for each creature to be reproduced in the next generation. Let $p = (c_1, c_2, \ldots, c_s) \in \wp_s$ be an arbitrary population, and suppose that a raw fitness value $f_0(c_\sigma, p) \in \mathbf{R}^+$ is defined for every creature c_σ in the population. Thus, we explicitly allow population-dependent fitness values of creatures.

Let $f_t(c_\sigma, p) = g(t, f_0(c_\sigma, p)) \in \mathbf{R}^+$ be the (possibly) scaled fitness value of creature c_σ in p at step t of the genetic algorithm. We shall suppose that for every population $p \in \wp_s$ and every $t \in \mathbf{N}$:

$$\sum_{\sigma=1}^{s} f_t(c_\sigma, p) > 0.$$

If this condition is violated, then redefine the fitness function to be constant 1 in a population where it was constant 0. For a population $p = (c_1, c_2, \ldots, c_s) \in \wp_s$ consisting of creatures $c_\sigma \in \mathscr{C}$, $\sigma \in [1, s] \cap \mathbf{N}$, the probability that any given new creature in step $t+1$ is produced by literally copying c_σ is

$$\frac{f_t(c_\sigma, p)}{\sum_{\sigma'=1}^{s} f_t(c_{\sigma'}, p)}. \tag{20}$$

From this, it is not hard to determine the associated stochastic matrix $F_{\mathrm{PFS}}^{(f_0)}(t)$ [62, Proposition 11.1]: for $p = (c_1, c_2, \ldots, c_s) \in \wp_s$, $c_\sigma \in \mathscr{C}$, and $q = (d_1, d_2, \ldots, d_s) \in \wp_s$, $d_\sigma \in \mathscr{C}$, $\sigma \in [1, s] \cap \mathbf{N}$, let $n(d_\sigma, p)$ denote the number of copies of d_σ in the population p. The probability that q is generated from p by fitness selection is given by

$$\langle q, F_{\mathrm{PFS}}^{(f_0)} p \rangle = \prod_{\sigma=1}^{s} \frac{n(d_\sigma, p)\, f_t(d_\sigma, p)}{\sum_{\sigma'=1}^{s} f_t(c_{\sigma'}, p)}. \tag{21}$$

As a consequence we obtain: let $c_{\max} \in \mathscr{C}$ be a creature of maximal fitness in p, and $q = (c_{\max}, \ldots, c_{\max}) \in \wp_s \cap \mathscr{U}$, then

$$\langle q, F_{\mathrm{PFS}}^{(f_0)}(t)\, p \rangle \geqslant s^{-s} =: 1 - \theta_{PFS}.$$

A common choice for a fitness scaling used with proportional fitness selection is the so-called *power-law scaling* [25, p. 124], [65, p. 65], [66, p. 100], where

$$f_t(c, p) = (f_0(c, p))^{g(t)}. \tag{22}$$

We shall say that a power-law scaling is unbounded, if $\lim_{t\to\infty} g(t)=\infty$. With the above notation, we define the stochastic matrix $F_\infty^{(f_0)}$ as follows:

$$\langle q, F_\infty^{(f_0)} p\rangle = \lim_{t\to\infty} \prod_{\sigma=1}^{s} \frac{n(d_\sigma, p)\, f_0(d_\sigma, p)^t}{\sum_{\sigma'=1}^{s} f_0(c_{\sigma'}, p)^t}. \tag{23}$$

For a fitness function f_0 where every population contains exactly one creature of maximal fitness (which can occur in multiple copies), one sees that $F_\infty^{(f_0)} p = q$, where $q=(c_{\max},\ldots,c_{\max})$, and $c_{\max}\in\mathscr{C}$ is one of the identical creatures of maximal fitness in p. Otherwise, $F_\infty^{(f_0)} p$ is a probability distribution over populations with best creatures in p depending upon the frequencies of those best creatures in p.

Mitchell [48, pp. 168–169] describes the so-called *Boltzmann selection* by a probability distribution for the current population in such a way that the method becomes an exponentially rescaled proportional fitness selection. See in particular the formula on p. 169 in [48]. This is different from Goldberg's approach in [24, p. 449], which amounts to a variation of tournament selection as described in Section 7.2. Goldberg shows [24, Section 3] that the inhomogeneous Markov chain describing his algorithm has the usual steady-state Boltzmann distribution over *all* possible creatures. Both approaches just mentioned are different from the approach by Mahfoud and Goldberg in [44, p. 304] as outlined in Section 6.2.

7.2. Tournament fitness selection

Goldberg and Deb [27, p. 78] (see also [48, p. 170][5]) propose the following tournament selection mechanism: Fix $\phi\in[0,\frac{1}{2})$. For $\sigma=1,\ldots,s$ do:

1. Select two creatures c_{σ_1} and c_{σ_2}, $\sigma_{1,2}\in[1,s]\cap\mathbf{N}$, at random from the current population $p=(c_1,\ldots,c_s)\in\wp_s$. It is assumed that all creatures $c_1,\ldots,c_s$ participate in this random selection process in both rounds, i.e., a particular creature may be selected twice.
2. If the selected creatures c_{σ_1} and c_{σ_2} have the same fitness, then select one of them with probability $\frac{1}{2}$. Otherwise, set d_σ to the creature with lower fitness value with probability ϕ, and to the creature with higher fitness value with probability $1-\phi$.

Goldberg's presentation [26] and Michalewicz's book [47, p. 59] contain variations of this procedure. Observe that this tournament fitness selection mechanisms only depends upon the rank of creatures, and not the actual fitness values.

Let $F_{\mathrm{TS}}(\phi)$ be the stochastic matrix acting on $\mathscr{V}_\wp$ associated with the tournament selection mechanism described above. Let $p\in\wp_s$, $c_{\max}\in\mathscr{C}$ be a creature of maximal fitness in p, and $q=(c_{\max},\ldots,c_{\max})\in\wp_s\cap\mathscr{U}$. It is easy to see that

$$\langle q, F_{\mathrm{TS}}(\phi)p\rangle \geqslant s^{-s} =: 1-\theta_{\mathrm{TS}}. \tag{24}$$

We shall call the tournament selection mechanism *scaled*, if ϕ is varied in accordance with a predetermined schedule $(\phi_t)_{t\in\mathbf{N}}$ during the genetic algorithm.

[5] The author thanks the anonymous referee A for pointing out that the tournament selection mechanism described here is apparently due to Goldberg and Deb, and not Mitchell as assumed in [61].

7.3. Rank selection

Baker [6] proposed to use only the fitness ranking of individuals in determining fitness selection probabilities, i.e., given a fitness function $f(c_\sigma, p) \in \mathbf{R}$ for every $p = (c_1, c_2, \ldots, c_s) \in \wp_s$, define

$$f_0(c_\sigma, p) = \mathrm{card}(\{c_{\sigma_1} : \sigma_1 \in [1,s] \cap \mathbf{N}, f(c_{\sigma_1}, p) \leqslant f(c_\sigma, p)\}). \tag{25}$$

Creatures are selected for the next step of the genetic algorithm, e.g., by scaled or unscaled proportional fitness selection as in Section 7.1 using f_0. See also Mitchell's book [48, p. 170] for genetic algorithms whose fitness selection schemes use rank.

7.4. Generalized fitness selection scaling

All the definitions given in Sections 7.1–7.3 and such procedures as *linear fitness scaling* [25, p. 77], and *sigma-truncation* [25, p. 124] can be summarized in the following definition, which extends a definition given in [61, Section 5.1].

7.1. Definition. A *generalized fitness selection scaling* is a map $t \mapsto F_t$, $t \in \mathbf{N}$, into the set of stochastic matrices, such that

1. $F_t P_{\mathscr{U}} = P_{\mathscr{U}}$ for every $t \in \mathbf{N}$, i.e., F_t leaves uniform populations invariant.
2. There exist disjoint subsets $\wp_s^{\mathrm{I}}$ and $\wp_s^{\mathrm{II}}$ of $\wp_s$ such that $\wp_s \cap \mathscr{U} \subset \wp_s^{\mathrm{I}}$, and $\wp_s^{\mathrm{I}} \cup \wp_s^{\mathrm{II}} = \wp_s$, and the conditions in the next section are satisfied where $\mathscr{W}_\kappa = \mathrm{span}_{\mathbf{C}}(\wp_s^\kappa)$, $\kappa = \mathrm{I}, \mathrm{II}$.
3. We have for some fixed $\rho \in \mathbf{N}, \theta_t \in [0,1]$, and every $v \in S$, $t \in \mathbf{N}$:
   - $P_{\mathscr{W}_{\mathrm{I}}} F_t P_{\mathscr{W}_{\mathrm{I}}} = F_t P_{\mathscr{W}_{\mathrm{I}}}$, i.e. F_t maps $\mathscr{W}_{\mathrm{I}}$ into $\mathscr{W}_{\mathrm{I}}$.
   - $P_{\mathscr{W}_{\mathrm{I}}} (\prod_{k=\rho+t}^{t} F_k) P_{\mathscr{W}_{\mathrm{II}}} v = (\prod_{k=\rho+t}^{t} F_k) P_{\mathscr{W}_{\mathrm{II}}} v$, i.e., after $\rho+1$ stages of "development" all populations end up in $\wp_s^{\mathrm{I}}$.
   - $\|(\mathbf{1} - P_{\mathscr{U}}) F_t P_{\mathscr{W}_{\mathrm{I}}} v\|_1 \leqslant \theta_t \|(\mathbf{1} - P_{\mathscr{U}}) P_{\mathscr{W}_{\mathrm{I}}} v\|_1$, i.e., $(\mathbf{1} - P_{\mathscr{U}}) F_t$ is a contracting map on $(\mathbf{1} - P_{\mathscr{U}})\mathscr{W}_{\mathrm{I}}$. We shall assume that θ_t is chosen minimal.
   - $\liminf_{t\to\infty} \theta_t =: \theta < 1$, i.e., this assures that the combined shrinking by products of all F_t transports everything into $\mathscr{U} \cap \wp_s$.

A generalized fitness scaling shall be called a *strong fitness scaling*, if $\wp_s^{\mathrm{II}} = \emptyset$, and if a raw fitness function f_0 is given such that

$$\lim_{t\to\infty} F_t = F_\infty^{(f_0)}, \tag{26}$$

where $F_\infty^{(f_0)}$ is defined as in identity (23). In that case, we have

$$\theta = \lim_{t\to\infty} \theta_t = \lim_{t\to\infty} \max\{\|(\mathbf{1} - P_{\mathscr{U}}) F_t v\|_1 : v \in S, \|(\mathbf{1} - P_{\mathscr{U}}) v\|_1 = 1\}$$

$$= \max\{\|(\mathbf{1} - P_{\mathscr{U}}) F_\infty^{(f_0)} v\|_1 : v \in S, \|(\mathbf{1} - P_{\mathscr{U}}) v\|_1 = 1\},$$

since identity (26) implies/means convergence in the $\|\cdot\|_1$ operator norm (see [59, p. 5, formula 5]). The prime example for a strong fitness scaling is unbounded power-law scaling – see identity (22).

Several examples for generalized fitness selection can be found in [61, Section 5.1]. Let us include two additional examples. For various examples of frequency-dependent selection see also the publication by Sigmund [64].

7.2. Example. Given a fitness function $f(c_\sigma, p) \in \mathbf{R}$ for every $p = (c_1, c_2, \ldots, c_s) \in \wp_s$, define

$$f_0(c_\sigma, p) = f(c_\sigma, p) - \min\{f(c_{\sigma_1}, p) \colon \sigma_1 \in [1,s] \cap \mathbf{N}\}$$

and reset $f_0(c_\sigma, p) = 1$, if f is constant on the population. Let $p_t \in \wp_s$ be the current population in the genetic algorithm after step t. Let $\tilde{p}_{t+1}$ be the current population in the genetic algorithm after mutation and crossover for step $t+1$. If $\tilde{p}_{t+1} \in \wp_s \backslash \mathscr{U}$, then use scaled or unscaled proportional fitness selection based upon f_0 to determine p_{t+1} from $\tilde{p}_{t+1}$. In the setting of Example 7.2, $\wp_s^1 = \wp_s$, and $\theta_t \leqslant \theta_F = 1 - s^{-s}$.

7.3. Example. Suppose, a fitness function $f(c_\sigma, p) \in \mathbf{R}$ is defined for every $p = (c_1, c_2, \ldots, c_s) \in \wp_s$. Let $p_t \in \wp_s$ be the current population in the genetic algorithm after step t. Let $\tilde{p}_{t+1}$ be the current population in the genetic algorithm after mutation and crossover for step $t+1$.

- If $\tilde{p}_{t+1} \in \wp_s \backslash \mathscr{U}$ is of non-uniform fitness, then eliminate one of the randomly selected creatures in $\tilde{p}_{t+1}$ with lowest fitness value, and replace it by a randomly selected creature with highest fitness value.
- If $\tilde{p}_{t+1} \in \wp_s \backslash \mathscr{U}$ is of uniform fitness, then select the creatures in the new population p_{t+1} at all positions randomly from creatures in $\tilde{p}_{t+1}$.

In the setting of Example 7.3, $\wp_s^1$ is the set of populations of uniform fitness, $\theta_t \leqslant \theta_{FF} = 1 - s^{-s}$, and $\rho = s - 2$. (FF stands for "forget fast".) Similar to [61, Lemma 5.1–2], we have the following two results.

7.4. Lemma. *Let $t \mapsto F_t$, $t \in \mathbf{N}$, be a generalized fitness selection scaling. Then we have*

1. $(\mathbf{1} - P_{\mathscr{U}})F_t = (\mathbf{1} - P_{\mathscr{U}})F_t(\mathbf{1} - P_{\mathscr{U}})$.
2. *For every $k \in \mathbf{N}$, there exists $k_0 \in [0, k-1] \cap \mathbf{N}_0$ such that $\sum_{t=0}^{\infty}(1 - \theta_{tk+k_0}) = \infty$.*

Proof. We have $(\mathbf{1} - P_{\mathscr{U}})F_t(\mathbf{1} - P_{\mathscr{U}}) = (\mathbf{1} - P_{\mathscr{U}})F_t - (\mathbf{1} - P_{\mathscr{U}})F_t P_{\mathscr{U}} = (\mathbf{1} - P_{\mathscr{U}})F_t - (\mathbf{1} - P_{\mathscr{U}})P_{\mathscr{U}} = (\mathbf{1} - P_{\mathscr{U}})F_t$. By hypotheses, $\liminf_{t\to\infty} \theta_t < 1$, which yields $\sum_{t=0}^{\infty}(1-\theta_t) = \infty$. Partitioning this sum of positive terms yields Lemma 7.4.2. □

7.5. Lemma. *Let $t \mapsto F_t$, $t \in \mathbf{N}$, be a generalized fitness selection scaling with associated $\theta_t \in [0,1]$ and $\rho \in \mathbf{N}$. Then we have.*

$$\left\| (\mathbf{1} - P_{\mathscr{U}}) \prod_{k=t}^{1} F_k v \right\|_1 \leqslant \left(\prod_{k=t}^{\rho+2} \theta_k \right) \|(\mathbf{1} - P_{\mathscr{U}})v\|_1$$

for every $v \in S$, and every $t \in \mathbf{N}$ such that $t > \rho + 1$.

Proof. Using $P_{\mathscr{W}_1}F_kP_{\mathscr{W}_1}v = F_kP_{\mathscr{W}_1}v$, we have by the $\|\cdot\|_1$-inequality for F_t:

$$\left\|(\mathbf{1}-P_{\mathscr{U}})\left(\prod_{k=t}^{\rho+2}F_k\right)P_{W_1}v\right\|_1 \leqslant \left(\prod_{k=t}^{\rho+2}\theta_k\right)\|(\mathbf{1}-P_{\mathscr{U}})P_{W_1}v\|_1$$

for $t\in\mathbf{N}$. Using Lemma 7.4.1, we have for $t>\rho+1$:

$$\begin{aligned}\left\|(\mathbf{1}-P_{\mathscr{U}})\left(\prod_{k=t}^{\rho+2}F_k\right)\left(\prod_{k=\rho+1}^{1}F_k\right)v\right\|_1 &\leqslant \left(\prod_{k=\rho+2}^{t}\theta_k\right)\left\|(\mathbf{1}-P_{\mathscr{U}})P_{W_1}\prod_{k=\rho+1}^{1}F_kv\right\|_1\\ &= \left(\prod_{k=t}^{\rho+2}\theta_k\right)\left\|(\mathbf{1}-P_{\mathscr{U}})F_{\rho+1}(\mathbf{1}-P_{\mathscr{U}})\prod_{k=\rho}^{1}F_kv\right\|_1\\ &\leqslant\cdots\leqslant\left(\prod_{k=t}^{\rho+2}\theta_k\right)\|(\mathbf{1}-P_{\mathscr{U}})v\|_1.\end{aligned}$$

since $\mathbf{1}-P_{\mathscr{U}}$ is a diagonal matrix, and $\|F_k\|_1=1$ (see [59, p. 5, formula (5)]). □

7.5. A short note on the effect of genetic drift

Genetic drift, well known from treatises on evolutionary genetics (see, e.g., the books by Maynard–Smith [45], Crow and Kimura [12], Roughgarden [55]), refers to the phenomenon that populations not subject to mutation tend to become uniform, i.e., over the generations an increasing number of creatures (individuals) in the population become genetically identical. See also the paper by Goldberg and Segrest [29, p. 3].

A number of authors have used genetic algorithms with very low mutation rate or zero mutation rate obtaining seemingly good results. The convergence of such algorithms to (local) optima can be understood as an instance of genetic drift combined with the averaging property of $C_\chi^{(r)}$ and $C_\chi^{(u)}$ on Hardy–Weinberg spaces (cf., 2.11, and Sections 5.2.1.4, 5.3.1.4). We point out that this process is non-ergodic in nature, i.e., the limit population strongly depends upon the initial population in general. Note that Banzhaf et al. [7] report enhanced performance of genetic algorithms using higher mutation rates.

Since we have extended the definition of generalized fitness selection scaling, we discuss shortly genetic drift for a genetic algorithm with fitness selection but without mutation. In fact, the results in [61, Section 6] remain valid. The proofs in [61, Section 6] actually carry through almost verbatim and are only indicated here.

Consider a generalized crossover operation $\chi\mapsto C_\chi$, $\chi\in[0,1]$, and a generalized fitness selection scaling F_t, $t\in\mathbf{N}$ as defined in Sections 5.1 and 7.4 respectively. Set $H_t=\prod_{k=t}^{1}F_kC_{\chi(k)}$. Then we have

- If $\wp_s^{\mathrm{I}}=\wp_s$, and $\wp_s^{\mathrm{II}}=\emptyset$ in the definition of generalized fitness selection scaling Definition 7.1.2, then applying Lemma 4.3.2 or 4.4.2 for mutation rate $\mu=0$,

$t \in \mathbf{N}$, yields:

$$\lim_{t\to\infty} \|(\mathbf{1} - P_{\mathscr{U}})H_t v_0\|_1 = 0 \tag{27}$$

since $\prod_{t=1}^{\infty} \theta_t = 0$ by Definition 7.1.3.

- Now consider the case with $\wp_s^{\mathrm{II}} \neq \emptyset$. Suppose that $\chi_0 \in [0,1]$ and $\varepsilon \in (0,1]$ exist such that $C_\chi - \varepsilon\mathbf{1}$ is a matrix with positive entries for $\chi \in [0, \chi_0]$. This condition can be satisfied for simple crossover for every $\chi_0 \in [0,1)$ by Section 5.2.1.1, and for unrestricted crossover for every $\chi_0 \in [0,1)$ or $s > a$ and $\chi_0 = 1$ by Section 5.3.1.1. Suppose now, that $\chi(t) \in [0, \chi_0]$ for every $t \in \mathbf{N}$. Then, also identity (27) holds.

Proof. The property of a positive diagonal of the $C_{\chi(t)}$ shows that $(\mathbf{1} - P_{\mathscr{U}}) \prod_{k=t+\rho+1}^{t} F_k C_{\chi(k)}$ is a contracting map for $t \in \mathbf{N}$ similar to the proof of Lemma 7.5 with contracting factor $1 - (1 - \theta_{t+\rho+2})\varepsilon^{\rho+2}$. Now, Lemma 7.4.2 shows the claim. □

Both cases considered show the phenomenon of genetic drift in a very general setting: we impose no requirement on convergence of the crossover operators or fitness selection operators themselves. If one wants to remove the condition $\chi(t) \in [0, \chi_0]$ on the crossover rates $\chi(t)$ in the second case, then one has to analyze the interplay of the crossover operators $C_{\chi(t)}$ with the sets $\wp_s^{\mathrm{I}}$ and $\wp_s^{\mathrm{II}}$. As in [61, Section 6], it is easy to show that the average convergence time is finite for this type of process or algorithm.

8. Ergodic behavior of genetic algorithms

In the following section, we shall derive ergodic-type theorems for (scaled) genetic algorithms. Ergodic-type theorems describe the asymptotic behavior of inhomogeneous Markov chains associated with probabilistic algorithms such as scaled genetic algorithms. Such an analysis is important: the short-term behavior of a probabilistic algorithm is determined by the interplay *random* generator phase vs. selection – the long-term, i.e., asymptotic behavior, may or may not indicate that the algorithm is properly designed to, e.g., find global maxima of an optimization problem. As Aarts and van Laarhoven point out for the simulated annealing algorithm [1, p. 49, first •]: "The algorithm has a potential for finding high-quality solutions; the required amount of computation time to realize this potential is usually quite large". Theoretical analysis of genetic algorithms is in many cases devoted to the study of asymptotic behavior. In that regard, we point out the work of Aytug and Koehler [5] (approximation of the steady-state distribution), Davies [14], Davies and Principe [15, 16] (asymptotic behavior for annealing the mutation rate in a genetic algorithm to zero), Goldberg [24] (asymptotic behavior for Boltzmann tournament selection), Leung et al. [42, p. 19] (asymptotic approximation of the steady-state distribution), Mahfoud [43, p. 157] (asymptotic behavior for Boltzmann tournament selection), Mahfoud and Goldberg [44] (asymptotic behavior for parallel recombinative simulated annealing), Nix and Vose [50, Section 4] (asymptotics of the simple genetic algorithm), Poli and Langdon [54] (asymptotic behavior of

schemata), Rodolph [57] (asymptotic behavior for the simple genetic algorithm), and Vose [69] (asymptotic behavior for the simple genetic algorithm).

In the first part of this section, we shall establish strong ergodicity for (inhomogeneous) Markov chains associated with scaled genetic algorithms, if all three operators converge separately, and the mutation rate converges to a strictly positive value. This establishes such methods using, e.g., linear fitness scaling [25, p. 77], or sigma-truncation [25, p. 124] as genuine experiments with probabilistic but non-arbitrary outcome. It is also relevant in that regard that these results cover the case of the simple genetic algorithm. However, Theorems 8.1 and 8.2 show that such algorithms do not converge to a probability distribution over populations containing only optimal solutions.

Next, we deal with the case, where still the mutation rate converges to a strictly positive value, but in addition the fitness scaling is a so-called strong fitness scaling, cf., [62, p. 123] or Section 7.4 identity (26), and the raw fitness function f_0 is such that every population contains exactly one creature with maximal fitness (which may occur multiple times). This is in particular the case for a population independent raw fitness function which is an injective functions on creatures. As outlined in [62, p. 121], the latter is a minor restriction, if genetic algorithms are considered for the purpose of function optimization. The idea behind loosening the requirement that the raw fitness function is injective on creatures, is population-dependent rank f_0 as in identity (25) based upon an injective fitness function f on creatures. The prime example for a strong fitness scaling is unbounded power-law scaling – see Section 7.1 in particular identity (22). In that case, Theorem 8.3 is remarkable in regard to use of the crossover operation: the limit probability distribution over populations is independent even of the crossover method as long as it commutes with mutation. We shall also show that for such scalings the limit probability distribution is independent of the strong fitness scaling, i.e., both the method used, and any schedule used.

The second part of this section deals with the case that the mutation rate converges monotonously to 0 but slowly enough to assure weak ergodicity of the inhomogeneous Markov chain describing the genetic algorithm. In addition, we shall, essentially, assume a power-law fitness scalings which is subject to convergence conditions (see Theorem 8.5, identity (31), and Theorem 8.6, identities (34) and (35)), and a raw fitness function f_0 which is population-independent and injective on creatures, or population-dependent rank based upon an injective function on creatures, and that mutation and crossover commute. The results are quite striking.

If a certain "integrable" convergence condition is satisfied such that the selection pressure increases fast, then there is essentially no other restriction on the crossover operation, and the algorithm asymptotically behaves as the following take-the-best search algorithm: (1) mutate in every step with rate decreasing to zero, and (2) map any population to the uniform population with the best creature. The take-the-best search algorithm is investigated, and its convergence is shown. Depending upon population-size, the take-the-best search algorithm does or does not converge to the optimal solution.

If populations size s is larger than the length of the genome of creatures ℓ and a certain logarithmic convergence condition is satisfied such that the selection pressure increases slowly but sufficiently fast, then the algorithm asymptotically converges to the optimal solution.

Note that most of our results apply to types of fitness selection mechanisms where the fitness of the individual also depends on the ambient population (see Examples 7.2, 7.3 and 2 and 3 in [61, Section 5.1]. Note in addition, that we strengthen the results in [61, 62] in that single-spot mutation is incorporated in the results, and in that we show that the crossover operation is almost arbitrary in some cases.

8.1. Non-zero limit mutation rate

We treat the cases of single-spot mutation and multiple-spot mutation separately. The theorem concerning single-spot mutation requires a rather strong technical condition, which is however satisfied for the simple genetic algorithm with most fitness selection schemes including scaled tournament fitness selection (see [27, p. 78], [48, p. 170]) and discussed in Section 7.2.

8.1. Theorem. *Consider the following hypotheses*:

- *Let* $M_\mu^{(1)}, \mu \in (0, L^{-1})$, *describe single-spot mutation. Let* $\mu(t) \in (0, L^{-1}), t \in \mathbf{N}$, *be such that* $\lim_{t\to\infty} \mu(t) = \mu_\infty \in (0, L^{-1})$ *exists.*
- *Let* C_χ, $\chi \in [0,1]$, *be a continuous generalized crossover operation as defined in Section* 5.1. *Let* $\chi(t) \in [0,1]$ *be such that* $\lim_{t\to\infty} \chi(t) = \chi_\infty$ *exists. Suppose that* $C_{\chi_\infty} - \varepsilon \mathbf{1}$ *is a matrix with positive entries for some* $\varepsilon \in (0,1]$. *This condition can be satisfied for simple crossover for* $\chi_\infty \in [0,1)$ *by Section* 5.2.1.1, *and is satisfied for unrestricted crossover by Section* 5.3.1.1 *either for* $\chi_\infty \in [0,1)$ *or* $s > a$.
- *Let* F_t, $t \in \mathbf{N}$, *be a generalized fitness selection scaling as defined in Section* 7.4 *with associated* $\theta_t \in [0,1]$, $\theta \in [0,1)$ *and* $\rho \in \mathbf{N}$. *Suppose that* $\lim_{t\to\infty} F_t = F_\infty$ *exists such that* F_∞ *has a fully positive diagonal.* (*The latter implies that* $\wp_s^{\mathrm{II}} = \emptyset$.)

This technical and rather strong condition assures a strongly ergodic Markov chain with fully positive matrices, and a fully positive limit matrix. It is satisfied for most simple genetic algorithms using scaled tournament fitness selection in the sense of [27, p. 78], [48, p. 170], or bounded power-law scaling as discussed in Section 7.1. It is not satisfied for unbounded power-law scaling.

Let $G_t = F_t\, C_{\chi(t)})\, M_{\mu(t)}^{(1)}$. Then we have:

1. If $v_0 \in S$ is the initial probability distribution over populations for the scaled genetic algorithm, then

$$v_\infty = \lim_{t\to\infty} \prod_{k=t}^{1} G_k v_0 = \lim_{t\to\infty} (F_\infty C_{\chi_\infty} M_{\mu_\infty}^{(1)})^t v_0$$

exists and is independent of $v_0 \in S$. In fact, the inhomogeneous Markov chain associated with the scaled genetic algorithm is strongly ergodic.

2. The limit probability distribution v_∞ is fully positive. Hence, the algorithm does not converge to a probability distribution over populations containing only optimal solutions.
3. $\wp_s = \wp_s^1$ for $\wp_s^1$ as in the definition of generalized fitness selection scaling in Section 7.4.1, and we have:

$$\|(\mathbf{1} - P_{\mathscr{U}})v_\infty\|_1 \leqslant \theta \frac{L\mu_\infty}{1 - \theta(1 - L\mu_\infty)}.$$

Proof. For the proof of Theorem 8.1.1, we may assume W.L.O.G., that every matrix $C_{\chi(t)}$, and F_t, $t \in \mathbf{N}$ has a fully positive diagonal. Let $H_t^{t_1} = \prod_{k=t}^{t_1} F_k\, C_{\chi(k)} M_{\mu(k)}^{(1)}$, $t_1, t \in \mathbf{N}$, $t_1 \leqslant t$. Using [62, Theorem 16] or Gidas' theorem [22, Theorem 1.1], we see that $H_{tM+t_0}^{t_0+1}$ is strongly ergodic in $t \in \mathbf{N}$, for every fixed t_0, $M \in \mathbf{N}_0$, $L \leqslant M$. Now fix t_0, and a prime number $n > L$. Then $L^{n-1} \doteq 1 \,(\mathrm{mod}\, n)$ by [67, p. 131]. Hence,

$$(n-1)t_0 L^{(n-1)k} + t_0 \doteq nt_0 \doteq 0 \,(\mathrm{mod}\, n) \quad \text{and} \quad nLk \doteq 0 \,(\mathrm{mod}\, n). \tag{28}$$

Consequently, we have

$$\lim_{t\to\infty} (H_{tL+t_0}^1 v_0 - H_{tL}^1 v_0) = 0,$$

since both terms in the difference converge separately, and both sequences $(tL + t_0)_{t\in\mathbf{N}}$ and $(tL)_{t\in\mathbf{N}}$ contain a respective subsequence which is, at the same time, a subsequence of $n \cdot \mathbf{N}$ by the above identities in (28). Hence,

$$v_\infty = \lim_{t\to\infty} \prod_{k=t}^{1} G_k v_0 = \lim_{t\to\infty} (F_\infty C_{\chi_\infty} M_{\mu_\infty}^{(1)})^{Lt} v_0 = \lim_{t\to\infty} (F_\infty C_{\chi_\infty} M_{\mu_\infty}^{(1)})^{t} v_0$$

exists, and is independent of $v_0 \in S$. To prove Theorem 8.1.2, we observe that

$$(F_\infty C_{\chi_\infty} M_{\mu_\infty}^{(1)})^L$$

is a fully positive matrix. To obtain Theorem 8.1.3, we apply Lemma 4.3.2 to constant matrices $T(t) = F_\infty\, C_{\chi_\infty}$, and $\mu = \mu_\infty$. □

The result corresponding to Theorem 8.1 for multiple-spot mutation is actually much stronger since the fact that the multiple-spot mutation matrix is fully positive allows to lessen the requirements for the fitness scaling.

8.2. Theorem. *Consider the following hypotheses*:

- *Let* $M_\mu^{(m)}$, $\mu \in (0, (a-1)/a]$, *describe multiple-spot mutation. Let* $\mu(t) \in (0, (a - 1)/a]$, $t \in \mathbf{N}$, *be such that* $\lim_{t\to\infty} \mu(t) = \mu_\infty > 0$ *exists. Let* $\gamma = \gamma(\mu_\infty)$ *and* $\beta = \beta(\mu_\infty)$ *be given as in Proposition* 3.7.4 *for* $\mu = \mu_\infty$.
- *Let* C_χ, $\chi \in [0, 1]$, *be a continuous generalized crossover operation as defined in Section* 5.1 *Let* $\chi(t) \in [0, 1]$ *be such that* $\lim_{t\to\infty} \chi(t) = \chi_\infty$ *exists.*

- *Let* F_t, $t \in \mathbf{N}$, *be a generalized fitness selection scaling as defined in Section* 7.4 *with associated* $\theta_t \in [0,1]$, $\theta \in [0,1)$ *and* $\rho \in \mathbf{N}$. *Suppose that* $\lim_{t\to\infty} F_t = F_\infty$ *exists.*

Let $G_t = F_t\, C_{\chi(t)})M^{(m)}_{\mu(t)}$. Then we have:

1. If $v_0 \in S$ is the initial probability distribution over populations for the scaled genetic algorithm, then we have

$$v_\infty = \lim_{t\to\infty} \prod_{k=t}^{1} G_k v_0 = \lim_{t\to\infty} (F_\infty C_{\chi_\infty} M^{(m)}_{\mu_\infty})^t v_0$$

   exists and is independent of $v_0 \in S$. In fact, the inhomogeneous Markov chain associated with the scaled genetic algorithm is strongly ergodic.
2. If the fitness selection scaling F_t, $t \in \mathbf{N}$, stands for scaled proportional fitness selection, tournament fitness selection in the sense of [27, p. 78], [48, p. 170], [26], or [47, p. 59] as discussed in Section 7.2, or rank selection combined with (scaled) proportional fitness selection as discussed in Section 7.3, then the coefficients $\langle v_\infty, p\rangle$ of the limit probability distribution are strictly positive for every population $p \in \wp_s$ of uniform fitness. In particular, for a non-constant fitness function, the genetic algorithm G_t, $t \in \mathbf{N}$, does not converge to a population consisting solely of creatures with maximal fitness value.
3. If $\wp_s = \wp^{\mathrm{I}}_s$ for $\wp^{\mathrm{I}}_s$ as in the definition of generalized fitness selection scaling in Section 7.1, then we have

$$\|(\mathbf{1} - P_{\mathscr{U}})v_\infty\|_1 \leqslant \theta \frac{\gamma}{1-\theta\beta}.$$

4. Suppose that $C_{\chi_\infty} - \varepsilon\mathbf{1}$ is a matrix with positive entries for some $\varepsilon \in (0,1]$. This condition can be satisfied for simple crossover for $\chi_\infty \in [0,1)$ by Section 5.2.1.1, and is satisfied for unrestricted crossover by Section 5.3.1.1 either for $\chi_\infty \in [0,1)$ or $s > a$. If $\wp_s = \wp^{\mathrm{I}}_s \cup \wp^{\mathrm{II}}_s$, and $\wp^{\mathrm{II}}_s \neq \emptyset$ (see Section 7.1), then we have

$$\|(\mathbf{1} - P_{\mathscr{U}})v_\infty\|_1 \leqslant \frac{1-(1-\mu_\infty)^{L(\rho+2)}}{1-(1-\mu_\infty)^{L(\rho+2)}(1-\varepsilon^{\rho+2}(1-\theta))}.$$

Proof. The proof of Theorem 8.2.1 is omitted as it is very similar to the proof in [62, pp. 128–129, Theorem 17], or the proof of Theorem 8.1. Let us turn to the proof of Theorem 8.2.2:

Let f_0 be the given raw fitness function. In the case of scaled proportional fitness selection, let $p \in \wp_s$ be a population of uniform fitness, i.e., $p = (c_1, c_2, \ldots, c_s), c_1, \ldots, c_s \in \mathscr{C}$, and $f_0(c_1, p) = f_0(c_2, p) = \cdots = f_0(c_s, p)$. We have

$$v_\infty = F_\infty C_{\chi_\infty} M^{(m)}_{\mu_\infty} v_\infty. \tag{29}$$

Consider the p-component of $C_{\chi_\infty} M^{(m)}_{\mu_\infty} v_\infty$. Using that C_{χ_∞} is self-adjoint, we obtain

$$\langle p, C_{\chi_\infty} M^{(m)}_{\mu_\infty} v_\infty\rangle = \langle M^{(m)}_{\mu_\infty} C_{\chi_\infty} p, v_\infty\rangle > 0. \tag{30}$$

Identities (30), (29), and (20) imply $\langle p, v_\infty \rangle > 0$. This includes the case of rank selection combined with (scaled) proportional fitness selection. In the case of (scaled) tournament fitness selection following [27, 26, 47, 48], we have $\langle v_\infty, p \rangle > 0$ by combining (30), (29), and an argument similar to showing identity (24). This completes the proof of Theorem 8.2.2. To obtain Theorem 8.2.3, we apply Lemma 4.4.2. To show Theorem 8.2.4, we observe that

$$M_{\mu_\infty}^{(1)} = (1-\mu_\infty)^L \mathbf{1} + R,$$

where $\|R\|_1 = 1-(1-\mu_\infty)^L$ by [59, p. 5, formula $(7')$]. Using Lemma 7.5, we have

$$\begin{aligned}
\|(\mathbf{1}-P_{\mathscr{U}})v_\infty\|_1 &= \|(\mathbf{1}-P_{\mathscr{U}})F_\infty C_{\chi_\infty} M_{\mu_\infty}^{(1)} v_\infty\|_1 \\
&= (1-\mu_\infty)^L \|(\mathbf{1}-P_{\mathscr{U}})F_\infty C_{\chi_\infty} v_\infty\|_1 + \|(\mathbf{1}-P_{\mathscr{U}})F_\infty C_{\chi_\infty} R v_\infty\|_1 \\
&\leqslant (1-\mu_\infty)^L \|(\mathbf{1}-P_{\mathscr{U}})(F_\infty C_{\chi_\infty})^2 M_{\mu_\infty}^{(1)} v_\infty\|_1 + 1-(1-\mu_\infty)^L \\
&\leqslant (1-\mu_\infty)^{2L} \|(\mathbf{1}-P_{\mathscr{U}})(F_\infty C_{\chi_\infty})^3 M_{\mu_\infty}^{(1)} v_\infty\|_1 \\
&\quad +(1-\mu_\infty)^L(1-(1-\mu_\infty)^L) + 1-(1-\mu_\infty)^L \\
&\leqslant (1-\mu_\infty)^{(\rho+2)L} \|(\mathbf{1}-P_{\mathscr{U}})(F_\infty C_{\chi_\infty})^{\rho+2} v_\infty\|_1 + 1-(1-\mu_\infty)^{(\rho+2)L} \\
&= (1-\mu_\infty)^{(\rho+2)L} \|(\mathbf{1}-P_{\mathscr{U}})(F_\infty(\varepsilon\mathbf{1}+C^{(0)}))^{\rho+2} v_\infty\|_1 + 1-(1-\mu_\infty)^{(\rho+2)L} \\
&\leqslant (1-\mu_\infty)^{(\rho+2)L} (\|(\mathbf{1}-P_{\mathscr{U}})\varepsilon F_\infty(F_\infty(\varepsilon\mathbf{1}+C^{(0)}))^{\rho+1} v_\infty\|_1 \\
&\quad +(1-\varepsilon)\|(\mathbf{1}-P_{\mathscr{U}})(F_\infty(\varepsilon\mathbf{1}+C^{(0)}))^{\rho+1} v_\infty\|_1) + 1-(1-\mu_\infty)^{(\rho+2)L} \\
&\leqslant (1-\mu_\infty)^{(\rho+2)L} (\varepsilon^{\rho+2}\theta + (1-\varepsilon^{\rho+2}))\|(\mathbf{1}-P_{\mathscr{U}})v_\infty\|_1 + 1-(1-\mu_\infty)^{(\rho+2)L},
\end{aligned}$$

where $C^{(0)} = C_{\chi_\infty} - \varepsilon\mathbf{1}$, and $\|C^{(0)}\|_1 = 1-\varepsilon$. Now, Theorem 8.2.4 is obtained with the geometric series type argument as in the proof of Lemma 4.3.2. □

The estimates given in Theorem 8.2.4 are slightly better than the estimates given in [61, Theorem 7.4(2)]. The proof of [61, Theorem 7.4(2)] contains a misprint. In fact, the above arguments explicitly show that it is not necessary to assume that mutation and crossover commute.

The following theorem strengthens [62, Theorem 17] considerably. In fact, it shows for some regular genetic algorithm applications that the outcome of the genetic algorithm for positive limit mutation rate is, asymptotically, independent from any crossover procedure or crossover scaling schedule.

8.3. Theorem. *Consider the following hypotheses*:

- *Let* $M_\mu^{(m)}$, $\mu \in (0,(a-1)/a]$, *describe multiple-spot mutation. Let* $\mu(t) \in (0,(a-1)/a]$, $t \in \mathbf{N}$, *be such that* $\lim_{t\to\infty} \mu(t) = \mu_\infty > 0$ *exists. Let* $\gamma = \gamma(\mu_\infty)$ *and* $\beta = \beta(\mu_\infty)$ *be given as in Proposition* 3.7.4 *for* $\mu = \mu_\infty$.

- *Let* C_χ, $\chi \in [0,1]$, *be a continuous generalized crossover operation as defined in Section* 5.1. *Let* $\chi(t) \in [0,1]$ *be such that* $\lim_{t\to\infty} \chi(t) = \chi_\infty$ *exists.*
- *Let* F_t, $t \in \mathbf{N}$, *be a strong generalized fitness selection scaling as defined in Section* 7.4 *with associated* $\theta_t \in [0,1]$, *and* $\theta \in [0,1)$. *By definition,* $\lim_{t\to\infty} F_t = F_\infty^{(f_0)}$, *where* $F_\infty^{(f_0)}$ *is as in Section* 7.2 *identity* (23) *for a given possibly population-dependent raw fitness function* f_0. *This is the situation for proportional fitness selection and unbounded power-law scaling.*

Let $G_t = F_t C_{\chi(t)}) M_{\mu(t)}^{(m)}$. Then we have:

1. The limit distribution v_∞ of the inhomogeneous Markov chain associated with the genetic algorithm is independent of the method of fitness scaling. In particular for power-law scaling, the limit distribution v_∞ is independent of the scaling schedule.
2. If the raw fitness function f_0 is independent of the population and is (globally) injective on creatures, then v_∞ only depends upon rank induced by f_0, and not the particular values of f_0.
3. If the raw fitness function f_0 is such that every population contains exactly one creature with maximal fitness (which may occur multiple times), and if mutation commutes with crossover, then the conditions that the generalized crossover operation is continuous and that $\lim_{t\to\infty} \chi(t)$ exists can be dropped.
4. If the raw fitness function f_0 is such that every population contains exactly one creature with maximal fitness (which may occur multiple times), then $\|(\mathbf{1} - P_{\mathscr{U}})v_\infty\|_1 = 0$.

Proof. First observe that the limit distribution only depends upon the limit of the fitness selection matrices, i.e., $F_\infty^{(f_0)}$. This shows Theorem 8.3.1 and Theorem 8.3.2.

To show Theorem 8.3.3, we first make the following observation:

$$\begin{aligned}\lim_{t\to\infty} \|C_{\chi(t+1)}F_t - F_\infty^{(f_0)}\|_1 &= \lim_{t\to\infty} \|C_{\chi(t+1)}F_t - C_{\chi(t+1)}F_\infty^{(f_0)}\|_1 \\ &\leqslant \lim_{t\to\infty} \|F_t - F_\infty^{(f_0)}\|_1 = 0.\end{aligned}$$

Hence, $\lim_{t\to\infty} M_{\mu(t+1)}^{(m)} C_{\chi(t+1)} F_t = M_{\mu_\infty}^{(m)} F_\infty^{(f_0)}$ for any sequence $C_{\chi(t)}$. Note, that $M_{\mu_\infty}^{(m)} F_\infty^{(f_0)}$ is a fully positive matrix. Let $w \in S$ be the uniquely determined fixed-point of $M_{\mu_\infty}^{(m)} F_\infty^{(f_0)}$. Now, [62, Theorem 16] shows that the limit vector of the inhomogeneous, strongly ergodic Markov chain determined by the $M_{\mu(t+1)}^{(m)} C_{\chi(t+1)} F_t$ is w, and is thus independent from any choice of crossover operation.

Let $v_0 \in S$. Applying [62, Theorem 16] twice, we have for commuting mutation and crossover operator:

$$\begin{aligned}\lim_{n\to\infty} (F_\infty^{(f_0)} M_{\mu_\infty}^{(m)})^n v_0 = F_\infty^{(f_0)} \lim_{n\to\infty} (M_{\mu_\infty}^{(m)} F_\infty^{(f_0)})^n M_{\mu_\infty}^{(m)} v_0 &= F_\infty^{(f_0)} w \\ &= \lim_{t\to\infty} F_t \prod_{k=t-1}^{1} (M_{\mu(k+1)}^{(m)} C_{\chi(k+1)} F_k)(C_{\chi(1)} M_{\mu_1}^{(m)} v_0) \\ &= \lim_{t\to\infty} G_t v_0 = v_\infty.\end{aligned}$$

This shows Theorem 8.3.3. Theorem 8.3.4 follows from $v_\infty = F_\infty^{(f_0)} w$ shown above, if mutation and crossover commute. Otherwise, Theorem 8.3.4 follows from $\theta = 0$ and Theorem 8.2.3, if $\lim_{t\to\infty} \chi(t)$ exists, and crossover is continuous in $\chi(t)$.

□

8.2. Zero limit mutation rate and strong fitness scaling

Before we come to the final and most important results of Section 8, we need a definition.

8.4. Definition. Consider the following:

- Let $M_\mu^{(m)}$, $\mu \in (0, (a-1)/a]$, describe multiple-spot mutation. Let $\mu(t) \in (0, (a-1)/a]$, $t \in \mathbf{N}$, be a monotonously decreasing sequence such that $\lim_{t\to\infty} \mu(t) = 0$. Suppose that

$$\sum_{t=1}^{\infty} \mu(t)^L = \infty.$$

  Then we shall call this situation a *weakly ergodic multiple-spot mutation annealing*.
- If a weakly ergodic multiple-spot mutation annealing is given, and $F_\infty^{(f_0)}$ be as in Section 7.1 identity (23) for a possibly population-dependent raw fitness function f_0, then the inhomogeneous Markov chain

$$\prod_{k=t}^{1} F_\infty^{(f_0)} M_{\mu(k)}^{(m)}$$

  will be called a *take-the-best search algorithm*.

If the raw fitness function f_0 is such that every population contains exactly one creature with maximal fitness (which may occur multiple times), then a take-the-best search algorithm does the following: (i) mutate with respect to the given schedule $\mu(t)$ using multiple-spot mutation, and (ii) map a population $p \in \wp_s$ to the uniform population $(c_{\max}, \ldots, c_{\max})$ containing solely the creature $c_{\max}$ with maximal fitness in p.

8.5. Theorem. *Consider the following hypotheses*:

- *Let* $M_\mu^{(m)}$, $\mu(t) \in (0, (a-1)/a]$, $t \in \mathbf{N}$, *be a weakly ergodic multiple-spot mutation annealing.*
- *Let* C_χ, $\chi \in [0,1]$, *be a generalized crossover operation as defined in Section* 5.1. *Suppose that the crossover operation commutes with the mutation operation. This is satisfied in all regular applications by Section* 5.2.1.2 *and Section* 5.3.1.2.
- *Let* F_t, $t \in \mathbf{N}$, *be a strong generalized fitness selection scaling as defined in Section* 7.4. *Let* $F_\infty^{(f_0)}$ *be as in Section* 7.1, *identity* (23), *for a possibly population-dependent raw fitness function* f_0. *Suppose that for* $q, p \in \wp_s$ *we have*

$$\sum_{t=1}^{\infty} |\langle q, (F_t - F_\infty^{(f_0)}) p \rangle| < \infty. \tag{31}$$

Suppose that the given raw fitness function f_0 is such that every population contains exactly one creature with maximal fitness (which may occur multiple times). Condition (31) *is satisfied for proportional fitness selection, and power-law scaling for a large variety of scalings. For example, a linear growth function $g(t)$ as in Section* 7.2, *identity* (22), *satisfies condition* (31). *One has to use a suitable logarithmic growth function $g(t)$ to violate the condition.*

Let $G_t = F_t C_{\chi(t)}) M^{(m)}_{\mu(t)}$. Then we have:

1. The inhomogeneous Markov chain associated with the genetic algorithm is strongly ergodic. The limit probability distribution v_∞ of the inhomogeneous Markov chain associated with the genetic algorithm is independent of the method of fitness scaling. In particular for power-law scaling, the limit distribution v_∞ is independent of the scaling schedule as long as condition (31) holds.
2. In regard to v_∞, it is arbitrary what is used as generalized crossover operation as long as the crossover operator commutes with the mutation operator.
3. In the limit, the genetic algorithm behaves like the take-the-best search algorithm $\prod_{k=t}^{1} F^{(f_0)}_\infty M^{(m)}_{\mu(k)}$, and both inhomogeneous Markov chains are strongly ergodic. In particular, both inhomogeneous Markov chains have the same limit probability distribution over populations which is independent of the particular sequence $\mu(t)$, $t \in \mathbf{N}$.
4. Suppose that the given raw fitness function f_0 is such that there exists exactly one creature $c_+ \in \mathscr{C}$ which has strictly best fitness in all populations it occurs in. If the population size s is larger than the length of creatures ℓ, then the genetic algorithm converges to the uniform population with the optimal creature c_+.

Proof. The sequence $(\|F_t - F^{(f_0)}_\infty\|_1)_{t\in\mathbf{N}}$ is summable, i.e., an element of $\ell^1(\mathbf{N})$. In fact, since the coefficients are summable, so is

$$\left(\sum_{q,p\in\wp_s} |\langle q, (F_t - F^{(f_0)}_\infty) p\rangle| \right)_{t\in\mathbf{N}} \in \ell^1(\mathbf{N}).$$

The expression $\sum_{q,\,p\in\wp_s} |\langle q, Xp\rangle|$ is a norm for matrices X, and since all norms on a finite vector space are equivalent, we are done with this observation.

By Theorem 4.2, the inhomogeneous Markov chain underlying the genetic algorithm is weakly ergodic. We now show that this inhomogeneous Markov chain has the same limit as the corresponding take-the-best search algorithm. Consider for $t_1, t \in \mathbf{N}$, $t_1 \leqslant t$:

$$\hat{G}_t = M^{(m)}_{\mu(t+1)} F_t C_{\chi(t)}, \quad \hat{H}^{t_1}_t = \prod_{k=t}^{t_1} \hat{G}_k, \quad \tilde{G}_t = M^{(m)}_{\mu(t+1)} F^{(f_0)}_\infty \quad \text{and} \quad \tilde{H}^{t_1}_t = \prod_{k=t}^{t_1} \tilde{G}_k.$$

Note, that $\lim_{t\to\infty} (M^{(m)}_{\mu(t)})^{-1} = \mathbf{1}$, and thus for the purpose of establishing strong ergodicity of the inhomogeneous Markov chain describing the genetic algorithm, we may (with the same limit(s)) consider $\hat{H}^1_t$, and $\tilde{H}^1_t$ instead. Fix $v \in S$. Consider a sequence $(t_\kappa)_{\kappa\in\mathbf{N}}$ of integers, such that by compactness of $\mathrm{Ball}_1(\mathscr{V}_\wp)$ the vectors $\hat{H}^1_{t_\kappa} v$ converge.

Then we have, for $M \in \mathbf{N}$:

$$\lim_{\kappa\to\infty}(\hat{H}^1_{t_\kappa} - \hat{H}^{M+1}_{t_\kappa}\tilde{H}^1_M)v = 0,$$

applying weakly ergodicity to the inhomogeneous Markov chain determined by $\hat{H}^{M+1}_t$, $t \in \mathbf{N}$. Now we have

$$\begin{aligned}\|(\hat{H}^{M+1}_{t_\kappa}\tilde{H}^1_M - \tilde{H}^1_{t_\kappa})v\|_1 &= \|(\hat{H}^{M+1}_{t_\kappa}\tilde{H}^1_M - \hat{H}^{M+2}_{t_\kappa}\tilde{H}^1_{M+1} + \hat{H}^{M+2}_{t_\kappa}\tilde{H}^1_{M+1} - \tilde{H}^1_{t_\kappa})v\|_1 \\ &\leqslant \|F_{M+1} - F^{(f_0)}_\infty\|_1 + \|(\hat{H}^{M+2}_{t_\kappa}\tilde{H}^1_{M+1} - \tilde{H}^1_{t_\kappa})v\|_1 \\ &\leqslant \sum_{k=M+1}^{t_\kappa} \|F_k - F^{(f_0)}_\infty\|_1\end{aligned}$$

Here, the assumption, that mutation and crossover commute, is used. Now suppose that $w = \lim_{t\to\infty}\tilde{H}^1_t v$ exists and is independent of v, which we shall prove below. (In fact, we show that the associated inhomogeneous Markov chain is strongly ergodic.) Then we have

$$\lim_{\kappa\to\infty}\|(\hat{H}^1_{t_\kappa} - \tilde{H}^1_{t_\kappa})v\|_1 = \lim_{\kappa\to\infty}\|(\hat{H}^{M+1}_{t_\kappa}\tilde{H}^1_M - \tilde{H}^1_{t_\kappa})v\|_1 \leqslant \sum_{k=M+1}^{\infty}\|F_t - F^{(f_0)}_\infty\|_1 \to 0$$

as $M \to \infty$. Hence, if $w = \lim_{t\to\infty}\tilde{H}^1_t v$ exists, then all points of accumulation of $(\hat{H}^1_t v)_{t\in\mathbf{N}}$ are the same, i.e., the sequence converges to w.

Now, let us show that the take-the-best search algorithm determined by $\mu(t)$ is strongly ergodic. By Theorem 4.2, the inhomogeneous Markov chain underlying the take-the-best search algorithm is weakly ergodic. $\tilde{G}_t$ is a fully positive matrix which has a uniquely determined, fully positive eigenvector $w_t \in S$ to eigenvalue 1 by [59, p. 9, Proposition 2.8, p. 23, Corollary 2]. The equation determining w_t can be written as

$$(\tilde{G}_t - \mathbf{1})^{[e]}w_t = (a^{-L}, 0, 0, \ldots, 0)^{\mathrm{T}}. \tag{32}$$

Summing up the rows of $(\tilde{G}_t - \mathbf{1})$ in the first row yields $(\tilde{G}_t - \mathbf{1})^{[0]}$. The equation $(\tilde{G}_t - \mathbf{1})^{[0]}w_t = 0$ still determines the fully positive w_t up to a scalar factor. Thus, the kernel of $(\tilde{G}_t - \mathbf{1})^{[e]}$ is $\{0\}$. Hence, identity (32) is a full a^L-rank system of equations. If we apply Cramer's Rule [40, p. 182, Theorem 3], then the solution to identity (32) is a family of rational functions in $\mu(t)$ which stay bounded as $\mu(t) \to 0$. Consequently, the sequence $\|w_t - w_{t-1}\|_1$ is summable since the convergence is actually Lipschitz bounded (the rational functions determining w_t are well-defined and differentiable at $\mu = 0$). Applying [35, p. 160, Theorem V.4.3], completes the proof of Theorem 8.5.3.

Finally, let us show Theorem 8.5.4. Suppose that $\ell < s$. For $\mu > 0$, let $w_\mu \in S$ be the uniquely determined, invariant probability distribution of $M^{(m)}_\mu \cdot F^{(f_0)}_\infty$, and let $v_\mu = (M^{(m)}_\mu)^{-1}w_\mu \in S$ be the uniquely determined, invariant probability distribution of $F^{(f_0)}_\infty \cdot$

$M_\mu^{(m)}$ using [59, p. 7, Proposition 2.3]. v_μ and w_μ have the same limit as $\mu \to 0$. Let $c_+ = (\hat{a}(\imath_1^+), \ldots, \hat{a}(\imath_\ell^+)) \in \mathscr{C}$ be the creature with maximal fitness in $\mathscr{C}$. Let $p_+ = (c_+, c_+, \ldots, c_+) \in \wp_s \cap \mathscr{U}$. Let $\omega(\mu) = \langle p_+, v_\mu \rangle$. Then, $1 - \omega(\mu)$ is the combined probability for uniform populations over non-optimal creatures in v_μ.

Let $c \in \mathscr{C}$, $c \neq c_+$, and $p = (c, c, \ldots, c) \in \wp_s \cap \mathscr{U}$. In order to make a transition with $M_\mu^{(m)}$ from p to a population containing c_+, one may have to change all ℓ letters of c within one of the s copies of c in p. In order to make a transition with $M_\mu^{(m)}$ from p_+ to a population which does not contain c_+, one has to change at least one letter (arbitrarily) in every creature of p_+. Hence, we have for small μ:

$$1 - \omega(\mu) \leqslant (1 - K \cdot \mu^\ell)(1 - \omega(\mu))) + K' \cdot \mu^s \cdot \omega(\mu), \tag{33}$$

where $K > 0$, $K' > 0$ are constant. Now, identity (33) implies that

$$\lim_{\mu \to 0} (1 - \omega(\mu))/\omega(\mu) = 0,$$

which completes the proof. □

Note that the end of the proof of Theorem 8.5.3 is a simplification of a proof by Davies and Principle in [16]. We emphasize, that as outlined in [62, pp. 120–121], it is usually a minor restriction for an optimization problem to assume that a population-independent raw fitness function $f_0 : \mathscr{C} \to \mathbf{R}^+$ is injective.

8.6. Theorem. *Consider the following hypotheses*:

- *Multiple-spot mutation* $M_{\mu(t)}^{(m)}$ *with the cooling schedule* $\mu(t) = \min\{(a - 1)/2a, t^{-1/L}\}, t \in \mathbf{N}$ *is used.*
- *Let* C *be a constant generalized crossover operation as defined in Section* 5.1. *Suppose that the crossover operation commutes with the mutation operation. This is satisfied for simple crossover by Section* 5.2.1.2, *and for unrestricted crossover by Section* 5.3.1.2.
- *Let* $F_{\mathrm{PFS}}^{(f_0)}(t)$, $t \in \mathbf{N}$, *stand for power-law scaled proportional fitness selection as defined in identities* (21) *and* (22) *in Section* 7.1 *based upond a raw fitness function* f_0. *Let* $B \in \mathbf{R}_*^+$, *and*

$$g(t) = B \log(t + 1) \quad (\textit{see identity } (22)). \tag{34}$$

Let $G_t = F_{\mathrm{PFS}}^{(f_0)}(t) C M_{\mu(t)}^{(m)}$. Then we have:

1. The inhomogeneous Markov chain associated with the genetic algorithm is strongly ergodic.
2. Suppose that the given raw fitness function f_0 is such that there exists exactly one creature $c_+ \in \mathscr{C}$ which has strictly best fitness in all populations it occurs in. Suppose $s > \ell$, i.e., the population size s is larger than the length of creatures ℓ, and $B \in \mathbf{R}_*^+$, is chosen such that

$$\ell - 1 < L \cdot B \log(\rho_2), \tag{35}$$

where $\rho_2 > 1$ is the smallest value of $f_0(c_+, p)/f_0(c, p) > 1$ over all populations p containing c_+ and creatures $c \neq c_+$ in p. Then the genetic algorithm converges to the uniform population with the optimal creature c_+.

Proof. This proof is similar to the proof of Theorem 8.5. $M^{(m)}_{\mu(t)}$ with $\mu(t)$, $t \in \mathbf{N}$, as above is a weakly ergodic multiple-spot mutation annealing. First, we show strong ergodicity of the underlying inhomogeneous Markov chain for the genetic algorithm. Consider for $t \in \mathbf{N}$:

$$\hat{G}_t = M^{(m)}_{\mu(t+1)} F^{(f_0)}_{\mathrm{PFS}}(t) C, \quad \hat{H}_t = \prod_{k=t}^{1} \hat{G}_k.$$

Now, weak ergodicity of the inhomogeneous Markov chain $\hat{H}_t$, $t \in \mathbf{N}$, follows from Theorem 4.2. In order to show strong ergodicity of $\hat{H}_t$, we use [35, p. 160, Theorem V.4.3]. Thus, similar to the argument in the proof of Theorem 8.5.3, we have to show that the norms of differences $\|w_{t+1} - w_t\|_1$ of the steady-state distributions w_t of the $\hat{G}_t$ are summable.

It is enough to show summability for $|\langle p, w_{t+1} - w_t \rangle|$, $t \in \mathbf{N}$, for every population $p \in \wp_s$. To obtain this, we compute w_t with now continuous parameter $t = x^{-L} - 1$, $x = \mu(t+1) \in (0, (a-1)/a)$, as the solution of

$$(\hat{G}_t - \mathbf{1})^{[e]} w_t = (a^{-L}, 0, 0, \ldots, 0)^{\mathrm{T}}. \tag{36}$$

using Cramer's Rule [40, p. 182, Theorem 3]. Taking a look at the coefficients of F_t as in Section 7.1, identity (21), we obtain for populations $q, p \in \wp_s$: the probability that $q = (d_1, \ldots, d_s)$ is generated from $p = (c_1, \ldots, c_s)$, $d_\sigma, c_\sigma \in \mathscr{C}$, $\sigma = 1, \ldots, s$, by fitness selection is given by

$$\langle q, F^{(f_0)}_{\mathrm{PFS}}(t) p \rangle = \prod_{\sigma=1}^{s} \frac{n(d_\sigma, p) f_0(d_\sigma, p)^{g(t)}}{\sum_{\sigma'=1}^{s} f_0(c_{\sigma'}, p)^{g(t)}} = \prod_{\sigma=1}^{s} \frac{n(d_\sigma, p) x^{-L \cdot B \log(f_0(d_\sigma, p))}}{\sum_{\sigma'=1}^{s} x^{-L \cdot B \log(f_0(c'_\sigma, p))}}. \tag{37}$$

As a consequence of identity (37), we obtain that the solution $\langle p, w_{t = x^{-L} - 1} \rangle$ to identity (36) is a "rational" function in x where the powers of x in the denominator and numerator are *positive* real numbers, i.e.,

$$\phi(x) = \langle p, w_{t = x^{-L} - 1} \rangle = \frac{\sum_{v=1}^{n_1} r_{v,1} \cdot x^{r_{v,2}}}{\sum_{v=1}^{n_2} r_{v,3} \cdot x^{r_{v,4}}}, \quad r_{v,1}, r_{v,3} \in \mathbf{R}, \ r_{v,2}, r_{v,4} \in \mathbf{R}^+,$$
$$n_1, n_2 \in \mathbf{N}, \ r_{1,3} = 1, \ r_{1,4} = 0. \tag{38}$$

We can assume W.L.O.G., that $r_{1,3} = 1$ and $r_{1,4} = 0$ since $\langle p, w_{t = x^{-L} - 1} \rangle$ stays bounded as $x \to 0$. We substitute $x = z^K$ for a constant $K \geqslant 1$ in identity (38) to obtain

$$\phi(z^K) = \frac{\sum_{v=1}^{n_1} r_{v,1} \cdot z^{r'_{v,2}}}{1 + \sum_{v=2}^{n_2} r_{v,3} \cdot z^{r'_{v,4}}}, \quad r_{v,1}, r_{v,3} \in \mathbf{R}, \ r'_{v,2}, r'_{v,4} \in \{0\} \cup [1, \infty), \ n_1, n_2 \in \mathbf{N}.$$

Hence, $|(\mathrm{d}/\mathrm{d}z)\phi(z^K)|$ is bounded by some constant $K'>0$ for z in $[0,z_0]$, $z_0>0$. Consequently, we have for some $t_0\in\mathbf{N}$:

$$\sum_{t=t_0}^{\infty}|\langle p,w_{t+1}-w_t\rangle|\leqslant\sum_{t=t_0}^{\infty}K'\cdot|(t+2)^{-1/(K\cdot L)}-(t+1)^{-1/(K\cdot L)}|<\infty.$$

This completes the proof of Theorem 8.6.1.

Suppose now, that the conditions of Theorem 8.6.2 are satisfied. Let $w_x\in S$ be the uniquely determined, invariant probability distribution of $M_x^{(m)}F_{\mathrm{PFS}}^{(f_0)}(x^{-L}-1)C$, and let $v_x=(M_x)^{-1}w_x\in S$ be the uniquely determined, invariant probability distribution of $F_{\mathrm{PFS}}^{(f_0)}(x^{-L}-1)CM_x^{(m)}$. Let $c_+\in\mathscr{C}$ be the creature with maximal fitness in all populations it occurs in. Let $p_+=(c_+,\ldots,c_+)\in\wp_s$. Let $\omega(x)=\langle p_+,v_x\rangle$. We first observe that for every populations $p\in\wp_s$ that *contains* c_+, one has

$$\langle p_+,F_{\mathrm{PFS}}^{(f_0)}p\rangle\geqslant(1+(s-1)x^{L\cdot B\log(\rho_2)})^{-s}=\alpha_x. \tag{39}$$

Note that for some $K_1>0$:

$$1-\alpha_x\leqslant K_1\cdot x^{L\cdot B\log(\rho_2)}. \tag{40}$$

Let $p\in\wp_s$. We determine a lower estimate for the probability $\langle p_+,F_{\mathrm{PFS}}^{(f_0)}(x^{-L}-1)C M_x^{(m)}p\rangle$. Applying crossover (first) to p generates another population $p'\in\wp_s$. In order to make a transition with $M_\mu^{(m)}$ from p' to a population $p''\in\wp_s$ containing c_+, one has at most to change all ℓ letters in the spots corresponding to one creature of p'. Making a transition from p'' to p_+ under fitness evaluation occurs now with probability bounded below by $\alpha_x\geqslant s^{-s}$. Hence, there exists a constant $K_2>0$, such that for small x:

$$\langle p_+,F_{\mathrm{PFS}}^{(f_0)}(x^{-L}-1)CM_x^{(m)}p\rangle\geqslant K_2\cdot x^{\ell}. \tag{41}$$

Now, we consider the transition probability $\sum_{p\neq p_+}\langle p,F_{\mathrm{PFS}}^{(f_0)}(x^{-L}-1)CM_x^{(m)}p_+\rangle$. Applying crossover (first) to p_+ does not affect this population. We distinguish two cases.

Case 1: In order to make a transition with multiple-spot mutation from p_+ to a population which does not contain c_+, one has to change at least one letter (arbitrarily) in every creature of p_+. Fitness selection then cannot generate p_+ again.

Case 2: If c_+ is contained in a population $p'\in\wp_s$ after changing at least one letter of p_+, then the probability to generate p_+ again via proportional fitness selection is bounded below by α_x using identity (39).

Hence using identity (40), we have for a constant $K_3>0$:

$$\sum_{p\neq p_+}\langle p,F_{\mathrm{PFS}}^{(f_0)}(x^{-L}-1)CM_x^{(m)}p_+\rangle\leqslant K_3\cdot(x^s+x\cdot x^{L\cdot B\log(\rho_2)})\cdot\omega(x). \tag{42}$$

Combining identities (41) and (42), we obtain:

$$1-\omega(x)\leqslant(1-K_2\cdot x^{\ell})\cdot(1-\omega(x))+K_3\cdot(x^s+x^{1+L\cdot B\log(\rho_2)}))\cdot\omega(x) \tag{43}$$

Thus, $\lim_{x\to 0}\omega(x)=1$. □

8.7. Remark. We note that Theorem 8.6 with the proof given above also holds in case crossover is a "reasonable" function in the mutation rate x, e.g., the coefficients are as in identity (38).

8.3. Examples for non-optimal convergence

Example 1. Consider the smallest possible example: $a=\ell=s=2$. Let the raw fitness of creatures $(00),(01),(10),(11)$ be given by $f_0((00),p)=4, f_0((01),p)=1, f_0((10),p)=2, f_0((11),p)=3$, respectively, regardless of the population p.

One computes the 16×16 matrix $(M_\mu^{(m)}F_\infty^{(f_0)}-\mathbf{1})^{[e]}$ using computer algebra (MATHEMATICA [74]), and then solves the associated system, i.e., Eq. (32) using the Gauss algorithm simplifying rational expressions at every step. The explicit MATHEMATICA-computation can be downloaded from [60]. The coefficient of the solution w_t to Eq. (32) for population $((11),(11))$ is then determined as

$$\langle((11),(11)),w_t\rangle = \frac{2-11\mu(t+1)+28\mu(t+1)^2-38\mu(t+1)^3+28\mu(t+1)^4-8\mu(t+1)^5}{6-4\mu(t+1)},$$

which converges to $\frac{1}{3}$ as $t\to\infty$.

Example 2. Consider $a=s=2$, $\ell=3$. Let the raw fitness of creatures be given by $f_0((000),p)=8$, $f_0(bin(\kappa),p)=\kappa$, $\kappa=1,\ldots,7$ where $bin(\kappa)$ means the binary string representing an integer κ. Then a MATHEMATICA-computation yields:

$$\lim_{t\to\infty}\langle((111),(111)),w_t\rangle = 1.$$

These examples show that, in general, the take-the-best search algorithm and the genetic algorithm as in Theorem 8.5 do not converge to the uniform population with the best possible creature with probability 1 contradicting the main result in [65, 66]. The proof given in [65, 66] contains gaps which are substantial and not notational in nature, and occur for arbitrary size $s>1$ of populations, and arbitrary length $\ell>1$ of the genome of a creature.

9. Discussion of applicability

One of the goals of this work is to establish scaled genetic algorithms mathematically as a probabilistic, convergent all-purpose method for optimization similar to results obtained for the simulated annealing algorithm as outlined in [1]. Essentially, this is achieved in Theorem 8.6. Obviously, finding "just one copy of one optimal creature" in the course of the algorithm is sufficient as goal for optimization. However, in order to be guaranteed just to do so with high probability, it is necessary, as outlined below, to run the algorithm for a long time, and likely "end up" with a uniform population.

In the generic situation of a *blind* search with a fitness function of largely unknown behavior, one must, first of all, make sure that the search space is sufficiently explored. In the language of simulated annealing: "the algorithm needs time for a uniform, controlled overall 'cooling' procedure". This is achieved by scaling the parameters (mutation rate $\mu(t)$, crossover rate $\chi(t)$, and exponentiation $g(t)$ of the fitness function (see identity (22)), $t \in \mathbf{N}$) in such a way that the inhomogeneous Markov chain describing the algorithm becomes strongly ergodic. As a result, the asymptotic behavior of the scaled genetic algorithm becomes, in particular, independent from any choice of initial population or population sequence. Denote, as before, the stochastic matrices describing the individual steps of the scaled genetic algorithm by G_t, $t \in \mathbf{N}$, and let w_t denote the steady-state distribution of G_t. The proof of [35, p. 160, Theorem V.4.3] shows that the asymptotic behavior of the scaled genetic algorithm is determined by (1) weak ergodicity of the sequence G_t, (2) summability of the sequence $\|w_{t+1} - w_t\|_1$, and (3) the approximation of the limit probability distribution by the w_t. This shows that even if the genetic algorithm is being executed for a finite but larger number of cycles, its behavior is bounded by the convergence behavior of the w_t, $t \in \mathbf{N}$, and weak ergodicity. A larger mutation rate at the beginning of the algorithm enhances weak ergodicity, in particular, it shrinks the distance $\| \prod_{k=t}^{1} G_k(v_0 - v_0')\|_1$ between trajectories of the algorithm for initial distributions over populations v_0 and v_0'. But in the situation of a blind search, one needs a sequence of mutation rates decreasing to zero in order to be able to estimate the probability for the current population being among (uniform) populations with optimal creatures as in Eq. (43), since in this situation, the limit probability distribution is known. The fact that for any fitness scaling and zero limit mutation rate, the limit probability distribution is non-zero only over uniform populations (see Theorem 8.2), is thereby an unavoidable consequence of the selection method.

In addition to the points discussed above, the analysis presented in this work sets boundary conditions for application of genetic algorithms some of which are listed next. First, Theorems 8.1 and 8.2 show that one must always be aware of finding suboptimal solutions for positive mutation rates. In particular, the simple genetic algorithm has positive probability for doing so. Second, as discussed in Section 7.5, crossover alone is not really suitable as random generator phase in a genetic algorithm since this leads to implementation of genetic drift which is non-ergodic in nature. Third, if a scaled genetic algorithm is used, then the selection pressure should not increase too fast. Otherwise, one can use the take-the-best algorithm as simpler, equivalent alternative, cf., Theorem 8.5. Fourth, if a scaled genetic algorithm is used, then the conditions set in Theorem 8.5 or 8.6 in particular in regard to size of the population have to be observed.

10. Conclusion

As a consequence of contributions by many researchers and the analysis in this exposition, we can present the following table of asymptotic behavior of genetic algorithms. We have established here in all cases listed below, except for [24] and the

parallel algorithm in [44], that the Markov chain describing the genetic algorithm is strongly ergodic. What we mean below with "*convergence*" is convergence of the underlying (inhomogeneous) Markov chain to a steady state probability distribution which is non-zero only over populations containing optimal creatures (individuals, candidate solutions). Observe that the notion of *generalized crossover* introduced in Section 5.1 contains the commonly used crossover operations. Note also that some of the results in this exposition have been obtained in weaker versions by R.F. Fujii, C.L. Nehaniv and the author in [61, 62].

Case 1: The mutation and crossover operations are constant (over time).

(a) *Case*: Constant proportional fitness selection is used (simple genetic algorithm). Included in (2a).
- *Non-convergence* shown for binary alphabet, regular crossover and multiple-bit mutation by Davis and Principe [15], and Rudolph [57].

(b) *Case*: Simulated annealing type selection is used with a population-dependent fitness. Largely included in (3).
- *Convergence* shown for binary alphabet, regular crossover and multiple-bit mutation as well as a special crossover-mutation operator by Mahfoud and Goldberg [44]. See also [62]. A parallel algorithm based upon a creature-dependent fitness is proposed in [44].

(c) *Case*: Boltzmann selection is used with a logistic acceptance scheme.
- *Convergence* shown for binary alphabet and regular crossover by Goldberg [24].

Case 2: The mutation rate converges to a strictly positive value.

(a) *Case*: The crossover operators and the fitness selection operators converge.
- *Non-convergence* shown for general-size alphabet, generalized crossover and single/multiple-spot mutation in Theorems 8.1 and 8.2. See also [61, 62].

(b) *Case*: Unbounded power-law scaled fitness selection is used for an injective fitness function. The crossover operators need not converge to assure ergodicity.
- *Non-convergence* shown for general-size alphabet, generalized crossover and multiple-spot mutation in Theorem 8.3.

Case 3: The mutation and crossover rates vary in an interval. Simulated annealing type selection is used for a population-dependent fitness.
- *Convergence* shown for general-size alphabet, generalized crossover and single/multiple-spot mutation as discussed in Section 6.2 similar to the proof for a constant generator matrix outlined in [1].

Case 4: The mutation rate converges to zero. Multiple-spot mutation and proportional fitness selection are used.

(a) *Case*: The crossover operator is constant. The fitness evaluation is constant.
- *Non-convergence* shown for binary alphabet, regular crossover and multiple-bit mutation by Davis and Principe [15].
- *Non-convergence* shown, in principle, for general-size alphabet and generalized crossover as in the proof of Theorem 8.5, and Section 8.3.

(b) *Case*: Unbounded power-law scaled fitness selection is used for an injective fitness function.
- *Non-convergence* is shown, in general, in Theorem 8.5, and Section 8.3.
- *Convergence* is shown under certain conditions in Theorems 8.5, 8.6, and Remark 8.7.

Acknowledgements

The author thanks the anonymous referee A for pointing out a number of interesting references in regard to this exposition, and the anonymous referee B for the challenging report. Special thanks go to M. Ito in Kyoto.

References

[1] E.H.L. Aarts, P.J.M. van Laarhoven, Simulated annealing: an introduction, Statist. Neerlandica 43 (1989) 31–52.
[2] S. Anily, A. Federgruen, Simulated annealing methods with general acceptance probabilities, J. Appl. Probab. 24 (1987) 657–667.
[3] J. Antonisse, A new interpretation of Schema notation that overturns the binary encoding constraint, in: J.D. Schaffer (Ed.), Proceedings of the Third International Conference on Genetic Algorithms, Morgan Kaufmann, Los Altos, CA, 1989, pp. 86–97.
[4] R. Axelrod, W.D. Hamilton, The evolution of cooperation, Science 211 (1981) 1390–1396.
[5] H. Aytug, G.J. Koehler, Stopping criteria for finite length genetic algorithms, Inform. J. Comput. 8 (1996) 183–191.
[6] J.E. Baker, Reducing bias and inefficiency in the selection algorithm, in: J.J. Grefenstette (Ed.), Genetic Algorithms and Their Applications: Proceedings of the Second International Conference on Genetic Algorithms, Lawrence Erlbaum, London, 1987.
[7] W. Banzhaf, F.D. Francone, P. Nordin, The effect of extensive use of the mutation operator on generalization in genetic programming using sparse data sets, in: W. Ebeling, I. Rechenberg, H.P. Schwefel, H.M. Voigt (Eds.), Proceedings of the Fourth International Conference on Parallel Problem Solving from Nature (PPSN96), Springer, Berlin, 1996, pp. 300–309.
[8] A.D. Bethke, Genetic algorithms as function optimizers, Ph.D. Dissertation, University of Michigan, Dissertation Abstracts International, 41(9), 3503B, University Microfilms No. 8106101, 1981.
[9] S. Bhattacharyya, G.J. Koehler, An analysis of non-binary genetic algorithms with cardinality 2^v, Complex Systems 8 (1994) 227–256.
[10] K. Binder, Monte Carlo Methods in Statistical Physics, Springer, Berlin, 1978.
[11] V. Cerny, Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm, J. Optim. Theory Appl. 45 (1985) 41–51.
[12] J.F. Crow, M. Kimura, An Introduction to Populations Genetics Theory, Harper & Row, New York, 1970.
[13] L. Davis, Handbook of Genetic Algorithms, Van Nostrand Reinhold, New York, 1991.
[14] T.E. Davis, Toward an extrapolation of the simulated annealing convergence theory onto the simple genetic algorithm, Ph.D. Dissertation, University of Florida, 1991.
[15] T.E. Davis, J.C. Principe, A simulated annealing-like convergence theory for the simple genetic algorithm, in: R.K. Belew, L.B. Booker (Eds.), Proceedings of the Fourth International Conference on Genetic Algorithms '91, Morgan Kaufmann, Los Atlas, CA, 1991, pp. 174–181.
[16] T.E. Davis, J.C. Principe, A Markow chain framework for the simple genetic algorithm, Evol. Comput. 1 (3) (1993) 269–288.
[17] K. Deb, D.E. Goldberg, mGA in C: A messy genetic algorithm in C, Dept. of General Engineering, University of Illinois at Urbana-Champaign IlliGAL, Report No. 91008, 1991.

[18] D.B. Fogel, Evolving artificial intelligence, Ph.D. Dissertation, The University of California, San Diego, 1992.
[19] D.B. Fogel, Asymptotic convergence properties of genetic algorithms and evolutionary programming: analysis and experiments, Cybernet. Systems 25 (3) (1994) 389–407.
[20] H. Geiringer, On the probability theory of linkage in Mendelian heredity, Ann. Math. Statist. 15 (1944) 25–57.
[21] S. Geman, D. Geman, Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images, IEEE Proc. Pattern Anal. Mach. Intell. 6 (1984) 721–741.
[22] B. Gidas, Nonstationary Markov chains and convergence of the annealing algorithm, J. Statist. Phys. 39 (1985) 73–131.
[23] A.M. Gillies, Machine learning procedures for generating image domain feature detectors, Ph.D. Dissertation, University of Michigan, 1985.
[24] D.E. Goldberg, A note on Boltzmann tournament selection for genetic algorithms and population oriented simulated annealing, Complex Systems 4 (1990) 445–460.
[25] D.E. Goldberg, Genetic Algorithms, in Search, Optimization & Machine Learning, Addison-Wesley, Reading, MA, 1989.
[26] D.E. Goldberg, Genetic Algorithms Tutorial, Genetic Programming Conference, Stanford University, July 13, 1997.
[27] D.E. Goldberg, K. Deb, A comparative analysis of selection schemes used in genetic algorithms, in: G.J.E. Rawlins (Ed.), Foundations of Genetic Algorithms, Morgan Kaufmann, Los Altos, 1991, pp. 69–93.
[28] D.E. Goldberg, K. Deb, B. Korb, Messy genetic algorithms revisited: studies in mixed size and scale, Complex Systems 4 (1990) 415–444.
[29] D.E. Goldberg, P. Segrest, Finite Markov chain analysis of genetic algorithms, in: J.J. Grefenstette (Ed.), Genetic Algorithms and their Applications: Proceedings of the Second International Conference on Genetic Algorithms '87, Lawrence Erlbaum, London, 1987, pp. 1–8.
[30] W. Greub, Linear Algebra, Springer, Berlin, 1975.
[31] B. Hajec, Cooling schedules for optimal annealing, Math. Oper. Res. 13 (1988) 311–329.
[32] G.H. Hardy, Mendelian proportions in a mixed population, Science 28 (1908) 49–50.
[33] J.H. Holland, Adaptation in Natural and Artificial Systems, University of Michigan Press, 1975; Extended new Edition, MIT Press, Cambridge, 1992.
[34] J. Horn, Finite Markov chain analysis of genetic algorithms with niching, Illinois Genetic Algorithms Laboratory Report No. 93002, Dept. of General Engineering, University of Illinois, Urbana-Champaign, 1993.
[35] D.L. Isaacson, R.W. Madsen, Markov Chains: Theory and Applications, Prentice-Hall, Englewood Cliffs, NJ, 1961.
[36] G.J. Koehler, A proof of the Vose–Liepins conjecture, Ann. Math. Artificial Intelligence 10 (1994) 408–422.
[37] G.J. Koehler, S. Bhattacharyya, M.D. Vose, General cardinality genetic algorithms, Evol. Comput. 5 (4) (1998) 439–459.
[38] J.R. Koza, Genetic Programming, MIT Press, Cambridge, MA, 1992.
[39] J.R. Koza, Genetic Programming II, MIT Press, Cambridge, MA, 1994.
[40] S. Lang, Linear Algebra, 2nd Edition, Addison-Wesley, Reading, MA, 1970.
[41] S. Lang, Complex Analysis, Addison-Wesley, Reading, MA, 1977.
[42] Y. Leung, Z.-P. Chen, Z.-B. Xu, K.-S. Leung, Convergence rate for non-binary genetic algorithms with different crossover operators, The Chinese University of Hong Kong, preprint, 1998.
[43] S.W. Mahfoud, Finite Markov chain models of an alternative selection strategy for genetic algorithms, Complex Systems 7 (1993) 155–170.
[44] S.W. Mahfoud, D.E. Goldberg, A genetic algorithm for parallel simulated annealing, in: R. Männer, B. Manderick (Eds.), Parallel Problem Solving from Nature, vol. 2, Elsevier, Amsterdam, 1992, pp. 301–310.
[45] J. Maynard Smith, Evolutionary Genetics, Oxford University Press, Oxford, 1989.
[46] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, Equations of state calculations by fast computing machines, J. Chem. Phys. 21 (1953) 1087–1091.
[47] Z. Michalewicz, Genetic Algorithms + Data Structures = Evolution Programs, 2nd, extended Edition, Springer, Berlin, 1994.

[48] M. Mitchell, An Introduction to Genetic Algorithms, MIT Press, Cambridge, MA, 1996.
[49] D. Mitra, F. Romeo, A. Sangiovanni-Vincentelli, Convergence and finite time behavior of simulated annealing, Adv. Appl. Probab. 18 (1986) 747–771.
[50] A.E. Nix, M.D. Vose, Modeling genetic algorithms with Markov chains, Ann. Math. Artificial Intelligence 5 (1992) 79–88.
[51] C.L. Nehaniv (Ed.), Mathematical and Computational Biology: Computational Morphogenesis, Hierarchical Complexity, and Digital Evolution, An International Workshop, 21–25 October 1997, Aizu, Japan, Lectures on Mathematics in the Life Sciences Series, vol. 26, American Mathematical Society, Providence, RI, 1999.
[52] G.K. Pedersen, C^*-Algebras and Their Automorphism Groups, London Mathematical Society Monographs No. 14, Academic Press, New York, 1979.
[53] J.R. Peck, J.M. Yearsley, D. Waxman, Why do asexual and self-fertilizing populations tend to occur in marginal environments?, in: C.L. Nehaniv (Ed.), Mathematical and Computational Biology: Computational Morphogenesis, Hierarchical Complexity, and Digital Evolution, An International Workshop, 21–25 October 1997, Aizu, Japan, Lectures on Mathematics in the Life Sciences Series, vol. 26, American Mathematical Society, Providence, RI, 1999, pp. 121–132.
[54] A. Poli, M. Langdon, Schema theory for genetic programming with one-point crossover and point mutation, Evol. Comput. 6 (3) (1998) 231–252.
[55] J. Roughgarden, Theory of Population Genetics and Evolutionary Ecology, MacMillan, New York, 1976 (Reprinted by Prentice-Hall, Englewood Cliffs, NJ, 1996).
[56] W. Rudin, Functional Analysis, McGraw-Hill, New York, 1973.
[57] G. Rudolph, Convergence analysis of canonical genetic algorithms, IEEE Trans. Neural Networks 5 (1994) 96–101.
[58] G. Rudolph, An evolutionary algorithm for integer programming, in: Y. Davidor, H.-P. Schwefel, R. Männer (Eds.), Proceedings of the Third International Conference on Parallel Problem Solving From Nature (PPSN III), Springer, Berlin, 1994, pp. 139–148.
[59] H.H. Schaefer, Banach Lattices and Positive Operators, Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen Band 215, Springer, Berlin, 1974.
[60] L.M. Schmitt, Mathematica Computation, `ftp://ftp.u-aizu.ac.jp/u-aizu/doc/Tech-Report/1999/99-2-004.tar.gz`
[61] L.M. Schmitt, C.L. Nehaniv, The linear geometry of genetic operators with applications to the analysis of genetic drift and genetic algorithms using tournament selection, in: C.L. Nehaniv (Ed.), Mathematical and Computational Biology: Computational Morphogenesis, Hierarchical Complexity, and Digital Evolution, An International Workshop, 21–25 October 1997, Aizu, Japan, Lectures on Mathematics in the Life Sciences Series, vol. 26, American Mathematical Society, Providence, RI, 1999, pp. 147–166.
[62] L.M. Schmitt, C.L. Nehaniv, R.H. Fujii, Linear analysis of genetic algorithms, Theoret. Comput. Sci. 200 (1998) 101–134.
[63] E. Seneta, in: Non-negative Matrices and Markov Chains, Springer Series in Statistics, Springer, Berlin, 1981.
[64] K. Sigmund, The social life of automata, in: C.L. Nehaniv (Ed.), Mathematical and Computational Biology: Computational Morphogenesis, Hierarchical Complexity, and Digital Evolution, An International Workshop, 21–25 October 1997, Aizu, Japan, Lectures on Mathematics in the Life Sciences Series, vol. 26, American Mathematical Society, Providence, RI, 1999, pp. 133–146.
[65] J. Suzuki, A further result on the Markov chain model of genetic algorithms and its application to a simulated annealing-like strategy, in: R.K. Belew, M.D. Vose (Eds.), Foundations of Genetic Algorithms, vol. 4, Morgan Kaufmann, Los Altos, CA, 1997, pp. 53–72.
[66] J. Suzuki, A further result on the Markov chain model of genetic algorithms and its application to a simulated annealing-like strategy, IEEE Trans. Systems Man, Cybernet. – Part B 28 (1998) 95–102.
[67] B.L. van der Warden, Algebra I (Achte Auflage der Modernen Algebra), Heidelberger Taschenbücher Band 12, Springer, Berlin, 1971.
[68] M.D. Vose, Formalizing genetic algorithms, in: Proceedings of the IEEE Workshop on Genetic Algorithms, Neural Networks and Simulated Annealing Applied to Problems in Signal and Image Processing, May 1990, Glasgow, UK, 1990.
[69] M.D. Vose, Modeling simple genetic algorithms, in: G. Rawlins (Ed.), Foundations of Genetic Algorithms, Morgan Kaufmann, Los Altos, CA, 1991, pp. 94–101.

[70] M.D. Vose, G.E. Liepins, Punctuated equilibria in genetic search, Complex Systems 5 (1991) 31–44.
[71] M.D. Vose, A.H. Wright, The simple genetic algorithm and the Walsh transform: Part I, theory, Evol. Comput. 6 (3) (1998) 253–273.
[72] M.D. Vose, A.H. Wright, The simple genetic algorithm and the Walsh transform: Part II, the inverse, Evol. Comput. 6 (3) (1998) 275–289.
[73] W. Weinberg, Über Vererbungsgesetze beim Menschen, Zeitschrift für induktive Abstammungs- und Vererbungslehre 1 (1909) 277–330.
[74] S. Wolfram, Mathematica – A System for Doing Mathematics by Computer, Addison Wesley, Reading, MA, 1991.
[75] S. Wright, Statistical genetics and evolution, Bull. Amer. Math. Soc. 48 (1942) 223–246.