Available online at www.sciencedirect.com
ScienceDirect

Theoretical Computer Science 361 (2006) 72–110

Theoretical Computer Science

www.elsevier.com/locate/tcs

Some results about the Markov chains associated to GPs and general EAs

Boris Mitavskiy*, Jonathan Rowe

School of Computer Science, University of Birmingham, Edgbaston, Birmingham, B15, 2TT, UK

Abstract

Geiringer's theorem is a statement which tells us something about the limiting frequency of occurrence of a certain individual when a classical genetic algorithm is executed in the absence of selection and mutation. Recently Poli, Stephens, Wright and Rowe extended the original theorem of Geiringer to include the case of variable-length genetic algorithms and linear genetic programming. In the current paper a rather powerful finite population version of Geiringer's theorem which has been established recently by Mitavskiy is used to derive a schema-based version of the theorem for nonlinear genetic programming with homologous crossover. The theorem also applies in the presence of "node mutation". The corresponding formula in case when "node mutation" is present has been established.

The limitation of the finite population Geiringer result is that it applies only in the absence of selection. In the current paper we also observe some general inequalities concerning the stationary distribution of the Markov chain associated to an evolutionary algorithm in which selection is the last (output) stage of a cycle. Moreover we prove an "anti-communism" theorem which applies to a wide class of EAs and says that for small enough mutation rate, the stationary distribution of the Markov chain modelling the EA cannot be uniform.

Keywords: Evolutionary algorithms; Markov chain; Geiringer theorem; Stationary distribution; Mutation; Crossover; Fitness-proportional selection

1. Introduction

Geiringer's classical theorem (see [3]) is an important part of GA theory. It has been cited in a number of papers: see, for instance, [7,8,12,13]. It deals with the limit of the sequence of population vectors obtained by repeatedly applying the crossover operator $\mathcal{C}(p)_k = \sum_{i,j} p_i p_j r_{(i,j \to k)}$ where $r_{(i,j \to k)}$ denotes the probability of obtaining the individual k from the parents i and j after crossover. In other words, it speaks to the limit of repeated crossover in the case of an infinite population. In [4], a new version of this result was proved for *finite* populations, addressing the limiting distribution of the associated Markov chain, as follows. Let $\Omega = \prod_{i=1}^{n} A_i$ denote the search space of a given genetic algorithm (GA) (intuitively A_i is the set of alleles corresponding to the ith gene and n is the chromosome length). Fix a population P consisting of m individuals with m being an even number. P can be thought of as an m by n matrix whose

* Corresponding author.
E-mail address: B.S.Mitavskiy@cs.bham.ac.uk (B. Mitavskiy).

0304-3975/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2006.04.006

rows are the individuals of the population P. Write

$$P = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}.$$

Notice that the elements of the ith column of P are members of A_i. Continuing with the notation used in [7], denote by $\Phi(h, P, i)$ where $h \in A_i$ the proportion of rows, say j, of P for which $a_{ji} = h$. In other words, let $R_h = \{j \,|\, 1 \leqslant j \leqslant m$ and $a_{ji} = h\}$. Now simply let $\Phi(h, P, i) = \frac{|R_h|}{m}$. The classical Geiringer theorem (see [3,7] for modern notation) says that if one starts with a population P of individuals and runs a GA in the absence of selection and mutation (crossover being the only operator involved) then, in the "long run", the frequency of occurrence of the individual $(h_1, h_2, \dots, h_n)$ before time t, call it $\Phi(h_1, h_2, \dots, h_n, t)$, approaches independence:

$$\lim_{t \to \infty} \Phi(h_1, h_2, \dots, h_n, t) = \prod_{i=1}^{n} \Phi(h, P, i).$$

Thereby, Geiringer's theorem tells us something about the limiting frequency with which certain elements of the search space are sampled in the long run, provided one uses crossover alone. In [7] this theorem has been generalized to cover the cases of variable-length GAs and homologous linear genetic programming (GP) crossover. The limiting distributions of the frequency of occurrence of individuals belonging to a certain schema under these algorithms have been computed. The special conditions under which such a limiting distribution exists for linear GP under homologous crossover have been established (see [7, Theorem 9 and Section 4.2.1]). In [4] a rather powerful extension of the finite population version of Geiringer's theorem has been established. In the current paper we shall use the recipe described in [4] to derive a version of Geiringer's theorem for nonlinear GP with homologous crossover (see Section 6 or [5] for a detailed description of how nonlinear GP with homologous crossover works) which is based on Poli hyperschemata (see Section 6 or [5]). The first step in this procedure is to describe the search space and the appropriate family of reproduction transformations so that the resulting GP algorithm is bijective and self-transient in the sense of Definition 5.2 of [4]. Then the generalized Geiringer theorem ([4, Theorem 5.2]) as well as Corollaries 6.1 and 6.2 of [4] apply. The necessary details are summarized in the next few sections. A schema based version of Geiringer's theorem for nonlinear GP applies even in the presence of "node-mutation" (see Section 9).

The finite population Geiringer theorem established in [4] may completely describe the stationary distribution of the Markov chain associated to an evolutionary algorithm only in the absence of selection. In Section 10 we introduce a pre-order relation on the states of a Markov chain associated to an evolutionary algorithm which is defined in terms of selection alone, and establish some general inequalities about the stationary distribution of this Markov chain when selection is the "last stage" in the cycle. In Section 12 we demonstrate that the stationary distribution of the Markov chain associated to most evolutionary algorithms in the presence of selection can never be uniform when mutation rate is small enough, even if the fitness function is constant.

The material in Sections 10, 11 and 12 is independent of the results in Sections 5–9. Thus, the reader has an option of jumping to read Section 10 right after Section 4.

2. Notation

Ω is a finite set, called a *search space*.

$f : \Omega \to (0, \infty)$ is a function, called a *fitness function*. The goal is to find a maximum of the function f.

$\mathcal{F}_q$ is a collection of q-ary operations on Ω. Intuitively $\mathcal{F}_q$ can be thought of as the collection of reproduction operators: some q parents produce one offspring. In nature often $q = 2$, for every child has two parents, but in the artificial setting there seems to be no special reason to assume that every child has no more than two parents. When $q = 1$, the family $\mathcal{F}_1$ can be thought of as asexual reproductions or mutations. The following definitions will be used in Section 3 to describe the general evolutionary search algorithm. This approach makes it easy to state the Geiringer Theorem.

Definition 1. A population P of size m is simply an element of Ω^m. (Intuitively it is convenient to think of a population as a "column vector".)

Remark 2. There are 2 primary methods for representing populations: multi-sets and ordered multi-sets. Each has advantages, depending upon the particular analytical goals. Lothar Schmitt has published a number of papers which use the ordered multi-set representation to advantage (see, for instance, [10,11]). According to Definition 1, in the current paper we continue the development of analysis based upon the presentation pioneered by Lothar Schmitt. The following example illustrates an aspect of the representation which the reader would do well to keep in mind:

Example 3. Let $\Omega = \{0, 1\}^3$. Consider the populations

$$\begin{pmatrix} 0\ 0\ 0 \\ 1\ 1\ 1 \\ 1\ 1\ 1 \end{pmatrix}, \quad \begin{pmatrix} 1\ 1\ 1 \\ 0\ 0\ 0 \\ 1\ 1\ 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1\ 1\ 1 \\ 1\ 1\ 1 \\ 0\ 0\ 0 \end{pmatrix}.$$

According to Definition 1 (the ordered multi-set model which is exploited in the current paper) these are distinct populations despite the fact that they represent the same population under the multi-set model.

An *elementary step* is a probabilistic rule which takes one population as an input and produces another population of the same size as an output. For example, the following elementary step corresponds to the fitness-proportional selection which has been studied in detail by Wright and Fisher (see [14,2]).

Definition 4. An elementary step of type 1 (alternatively, of type *selection*) takes a given population

$$P = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}$$

with $x_i \in \Omega$ as an input. The individuals of P are evaluated:

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} \begin{matrix} \to & f(x_1) \\ \to & f(x_2) \\ \vdots & \vdots \\ \to & f(x_m). \end{matrix}$$

A new population

$$P' = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

is obtained where y_i's are chosen independently m times form the individuals of P and $y_i = x_j$ with probability $\frac{f(x_j)}{\Sigma_{l=1}^{m} f(x_l)}$.

In other words, all of the individuals of P' are among those of P, and the expectation of the number of occurrences of any individual of P in P' is proportional to the number of occurrences of that individual in P times the individual's fitness value. In particular, the fitter the individual is, the more copies of that individual are likely to be present in P'. On the other hand, the individuals having relatively small fitness value are not likely to enter into P' at all. This is designed to imitate the natural survival of the fittest principle.

Population P' is the output of this elementary step.

In order to define an elementary step of type 2 (reproduction) in a general setting which uses the ordered multi-set representation (see Remark 2 and Example 3) one needs to introduce the following definitions:

Definition 5. Fix an ordered k-tuple of integers $\mathbf{q} = (q_1, q_2, \ldots, q_k)$. Let K denote a partition of the set $\{1, 2, \ldots, m\}$ for some $m \in \mathbb{N}$. We say that partition K is $\mathbf{q}$-fit if every element of K consists of exactly q_i elements for some i. In logical symbols this means that if $K = \{P_1, P_2, \ldots, P_l\}$ then K is $\mathbf{q}$-fit if $\forall 1 \leqslant j \leqslant l \, \exists 1 \leqslant i \leqslant k$ such that $|P_j| = q_i$. Denote by $\mathcal{E}^m_{\mathbf{q}}$ the family of all $\mathbf{q}$-fit partitions of $\{1, 2, \ldots, m\}$ (i.e. $\mathcal{E}^m_{\mathbf{q}} = \{K | K$ is a $\mathbf{q}$-fit partition of $\{1, 2, \ldots, m\}\}$).

Definition 6. Let Ω be a set, $\mathcal{F}_{q_1}, \mathcal{F}_{q_2}, \ldots, \mathcal{F}_{q_k}$ be some fixed families of q_j-ary operations on Ω ($\mathcal{F}_{q_j}$ is simply a family of functions from Ω^{q_j} into Ω), and $p_1, p_2, \ldots, p_k$ be probability distributions on $(\mathcal{F}_{q_1})^{q_1}, (\mathcal{F}_{q_2})^{q_2}, \ldots, (\mathcal{F}_{q_k})^{q_k}$, respectively. Let $\mathbf{q} = (q_1, q_2, \ldots, q_k)$. Finally, let $\wp_m$ be a probability distribution on the collection $\mathcal{E}^m_{\mathbf{q}}$ of partitions of $\{1, 2, \ldots, m\}$ (see Definition 5). We then say that the ordered $2(k+1)$-tuple $(\Omega, \mathcal{F}_{q_1}, \mathcal{F}_{q_2}, \ldots, \mathcal{F}_{q_k}, p_1, p_2, \ldots, p_k, \wp_m)$ is a reproduction k-tuple of arity $(q_1, q_2, \ldots, q_k)$.

The following definition of reproduction covers both, crossover and mutation. Definition 8 (see also Remark 9) will make it possible to combine different reproduction operators in a simple and natural way.

Definition 7. An elementary step of type 2 (alternatively, of type *reproduction*) associated to a given reproduction k-tuple $(\Omega, \mathcal{F}_{q_1}, \mathcal{F}_{q_2}, \ldots, \mathcal{F}_{q_k}, p_1, p_2, \ldots, p_k, \wp_m)$ takes a given population

$$P = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}$$

with $x_i \in \Omega$ as an input.

The individuals of P are partitioned into pairwise disjoint tuples for mating according to the probability distribution $\wp_m$. For instance, if the partition selected according to $\wp_m$ is $K = \{(i^1_1, i^1_2, \ldots, i^1_{q_1}), (i^2_1, i^2_2, \ldots, i^2_{q_2}), \ldots, (i^j_1, i^j_2, \ldots, i^j_{q_j}), \ldots\}$ the corresponding tuples are

$$Q_1 = \begin{pmatrix} x_{i^1_1} \\ x_{i^1_2} \\ \vdots \\ x_{i^1_{q_1}} \end{pmatrix} Q_2 = \begin{pmatrix} x_{i^2_1} \\ x_{i^2_2} \\ \vdots \\ x_{i^2_{q_2}} \end{pmatrix} \ldots Q_j = \begin{pmatrix} x_{i^j_1} \\ x_{i^j_2} \\ \vdots \\ x_{i^j_{q_j}} \end{pmatrix} \ldots .$$

Having selected the partition, replace every one of the selected q_j-tuples

$$Q_j = \begin{pmatrix} x_{i^j_1} \\ x_{i^j_2} \\ \vdots \\ x_{i^j_{q_j}} \end{pmatrix}$$

with the q_j-tuples

$$Q' = \begin{pmatrix} T_1(x_{i^j_1}, x_{i^j_2}, \ldots, x_{i^j_{q_j}}) \\ T_2(x_{i^j_1}, x_{i^j_2}, \ldots, x_{i^j_{q_j}}) \\ \vdots \\ T_{q_j}(x_{i^j_1}, x_{i^j_2}, \ldots, x_{i^j_{q_j}}) \end{pmatrix}$$

for a q_j-tuple of transformations $(T_1, T_2, \dots, T_{q_j}) \in (\mathcal{F}_{q_j})^{q_j}$ selected randomly according to the probability distribution p_j on $(\mathcal{F}_{q_j})^{q_j}$. This gives us a new population

$$P' = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

which serves as the output of this elementary step.

Notice that a single child does not have to be produced by exactly two parents. It is possible that a child has more than two parents. Asexual reproduction (mutation) is also allowed.

Definition 8. A cycle is a finite sequence of elementary steps, say $\{s_n\}_{n=1}^{j}$, which are either of type 1 or of type 2 and such that all of the steps in the sequence $\{s_n\}_{n=1}^{j}$ have the same underlying search space and the same arity of input/output.

Remark 9. Intuitively, these steps are linked together in such a way that the output of the step s_i is the input of the step s_{i+1}. This is why all of the steps in the same cycle must have the same underlying search space and the same arity of input/output (otherwise the input/output relationship does not make sense).

We are finally ready to describe a rather wide class of evolutionary heuristic search algorithms.

3. How does a heuristic search algorithm work?

A general evolutionary search algorithm works as follows: Fix a *cycle*, say $C = \{s_n\}_{n=1}^{j}$ (see Definition 8). Now start the algorithm with an initial population

$$P = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}.$$

The initial population P may be selected completely randomly, or it may also be predetermined depending on the circumstances. The actual method of selecting the initial population P is irrelevant for the purposes of the current paper. To run the algorithm with cycle $C = \{s_n\}$, simply input P into s_1, run s_1, then input the output of s_1 into $s_2 \dots$ input the output of s_{j-1} into s_j and produce the new output, say P'. Now use P' as an initial population and run the cycle C again. Continue this loop finitely many times depending on the circumstances.

Definition 10. A sub-algorithm of a given evolutionary search algorithm defined by a cycle $C = \{s_n\}_{n=1}^{j}$ is simply an evolutionary search algorithm defined by a subsequence $\{s_{n_q}\}_{q=1}^{l}$ of the sequence C of elementary steps.

A *recombination* sub-algorithm is sub-algorithm defined by a sequence of elementary steps of type 2 (reproduction) only.

4. The Markov chain associated to an evolutionary algorithm

In [13] it has been pointed out that heuristic search algorithms give rise to the following Markov process [1] (see also [1], for instance): The state space of this Markov process is the set of all populations of a fixed size m. This set, in our notation, is simply Ω^m. The transition probability $p_{\mathbf{xy}}$ is simply the probability that the population $\mathbf{y} \in \Omega^m$ is obtained from the population $\mathbf{x}$ by going through the cycle once (where the notion of a cycle is described in Section 3: see

[1] In the current paper the state space of this process is slightly modified for technical reasons which will be seen later.

Definition 8 and Remark 9). The aim of the current paper is to establish a few rather general properties of this Markov chain. In case when there are several algorithms present in our discussion we shall write $\{p^{\mathcal{A}}_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\Omega^m}$ to denote the Markov transition matrix associated to the algorithm $\mathcal{A}$ while $\{p^{\mathcal{B}}_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\Omega^m}$ would denote the Markov transition matrix associated to the algorithm $\mathcal{B}$.

Definition 11. Fix an evolutionary search algorithm $\mathcal{A}$. Denote by $p^n_{\mathbf{x},\mathbf{y}}$ the probability that a population $\mathbf{y}$ is obtained from the population $\mathbf{x}$ upon the completion of n complete cycles (in the sense of Definition 8 and Remark 9) of the algorithm. We say that a population $\mathbf{x}$ leads to a population $\mathbf{y}$ under $\mathcal{A}$ if and only if $p^n_{\mathbf{x},\mathbf{y}} > 0$ for some n. We also write $\mathbf{x} \xrightarrow{\mathcal{A}} \mathbf{y}$ as a shorthand notation for $\mathbf{x}$ leads to $\mathbf{y}$. (This terminology is adopted from [1].)

5. A special kind of reproduction steps and the extended Geiringer theorem

To understand the intuitive meaning of the definition below, see Sections 2 and 3.

Definition 12. Given a set Ω and a family of transformations $\mathcal{F}_q$ from Ω^q into Ω, fix a q-tuple of transformations $(T_1, T_2, \ldots, T_q) \in (\mathcal{F}_q)^q$. Now consider the transformation $\langle T_1, T_2, \ldots, T_q\rangle : \Omega^q \to \Omega^q$ sending any given element

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_q \end{pmatrix} \in \Omega^q \text{ into } \begin{pmatrix} T_1(x_1, x_2, \ldots, x_q) \\ T_2(x_1, x_2, \ldots, x_q) \\ \vdots \\ T_q(x_1, x_2, \ldots, x_q) \end{pmatrix} \in \Omega^q.$$

We say that the transformation $\langle T_1, T_2, \ldots, T_q\rangle$ is the tupling of the ordered q-tuple $(T_1, T_2, \ldots, T_q)$.

Definition 13. Given an elementary step of type 2 (reproduction) associated to the reproduction k-tuple $\Omega = (\Omega, \mathcal{F}_{q_1}, \mathcal{F}_{q_2}, \ldots, \mathcal{F}_{q_k}, p_1, p_2, \ldots, p_k, \wp_m)$, fix some index i with $1 \leqslant i \leqslant k$ and denote by

$$\mathcal{G}(\Omega, q_i) = \{\langle T_1, T_2, \ldots, T_q\rangle : \Omega^{q_i} \to \Omega^{q_i} | T_j \in \mathcal{F}_{q_i}, p_i(T_1, T_2, \ldots, T_{q_i}) > 0\}$$

the family of all tuplings which have a positive probability of being selected.

Remark 14. The family of tupling transformations $\mathcal{G}(\Omega, q_i)$ described in Definition 13 represents the family of q parents → q children crossover transformations while the family $\mathcal{F}_q$ represents the family of q parents → 1 child crossovers. Depending on the circumstances it may be more convenient to specify the family of q parents → q children crossover transformations directly rather than specifying the families $\mathcal{F}_q$ individually. We shall see an example of this situation in Section 6. The family $\mathcal{F}_q$ of q parents → 1 child crossovers can then be recovered from the family of q parents → q children crossover transformations by using coordinate projections.

As mentioned in Section 2, in nature often the arity of the reproduction transformations is 2 meaning that every child has 2 parents.

It turns out that quite many evolutionary algorithms, including the classical GA and nonlinear (as well as linear) GP are equipped with the reproduction steps having the following nice property which has been introduced and investigated in [4].

Definition 15. A given elementary step of type 2 (reproduction) associated to the reproduction k-tuple $(\Omega, \mathcal{F}_{q_1}, \mathcal{F}_{q_2}, \ldots, \mathcal{F}_{q_k}, p_1, p_2, \ldots, p_k, \wp_m)$ is said to be bijective (and self-transient) if it satisfies conditions 1 (and 2) stated below:

1. $\forall 1 \leqslant i \leqslant k$ we have $p_i(T_1, T_2, \ldots, T_{q_i}) > 0 \Longrightarrow \langle T_1, T_2, \ldots, T_{q_i}\rangle$ (see Definition 12 for the meaning of $\langle T_1, T_2, \ldots, T_{q_i}\rangle$) is a bijection (a one-to-one and onto map of Ω^{q_i} onto itself).
2. $\forall 1 \leqslant i \leqslant k \exists (T_1, T_2, \ldots, T_{q_i}) \in (\mathcal{F}_{q_i})^{q_i}$ such that $p_i(T_1, T_2, \ldots, T_{q_i}) > 0$ and $\langle T_1, T_2, \ldots, T_{q_i}\rangle = \mathbf{1}$ where $\mathbf{1} : \Omega^{q_i} \to \Omega^{q_i}$ denotes the identity map (i.e. $\forall \mathbf{x} \in \Omega^{q_i}$ we have $\langle T_1, T_2, \ldots, T_{q_i}\rangle(\mathbf{x}) = \mathbf{x}$). We say that a recombination sub-algorithm (see Definition 10) of a given evolutionary search algorithm is bijective (and self-transient) if every given term of the subsequence, s_{n_k} by which the sub-algorithm is defined is bijective (and self-transient).

Remark 16. Notice that conditions 1 and 2 of Definition 15 can be restated in terms of the family $\mathcal{G}(\Omega, q_i)$ as follows:
1. Every transformation in the family of tuplings, $\mathcal{G}(\Omega, q_i)$ is a bijection.
2. $\mathbf{1} \in \mathcal{G}(\Omega, q_i)$ where $\mathbf{1} : \Omega^{q_i} \to \Omega^{q_i}$ denotes the identity map.

In [4] the following nice facts have been established:

Proposition 17. *Let $\mathcal{A}$ denote a bijective and self-transient algorithm* (*see Definition* 15). *Then $\xrightarrow{\mathcal{A}}$ is an equivalence relation.*

Proposition 17 motivates the following definition:

Definition 18. Given a bijective and self-transient algorithm $\mathcal{A}$ and a population $P \in \Omega^m$, denote by $[P]_{\mathcal{A}}$ the equivalence class of the population P under the equivalence relation $\xrightarrow{\mathcal{A}}$.

To alleviate the level of abstraction we illustrate Proposition 17 and Definition 18 with a couple of examples.

Example 19. Consider a binary GA over the search space $\Omega = \{0, 1\}^n$ under the action of crossover alone. Let the population size be some even number m. Consider the following family of masked crossover transformations: $\mathcal{F} = \{F_M | M \subseteq \{1, 2, \ldots, n\}\}$ where each F_M is a binary operation (i.e. a function from Ω^2 into Ω) defined as follows: For every $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ and $\mathbf{b} = (b_1, b_2, \ldots, b_n) \in \Omega^n$, $F_M(\mathbf{a}, \mathbf{b}) = \mathbf{x} = (x_1, x_2, \ldots, x_n) \in \Omega^n$ where

$$x_i = \begin{cases} a_i & \text{if } i \in M, \\ b_i & \text{otherwise.} \end{cases}$$

Let $\mathcal{A}$ denote the evolutionary algorithm determined by a single elementary step of type 2 (crossover) which is associated to the reproduction $\frac{m}{2}$-tuple

$$\Omega = (\Omega, \mathcal{F}, \mathcal{F}, \ldots, \mathcal{F}, p_1, p_2, \ldots, p_{\frac{m}{2}}, \wp_m)$$

(see Definitions 6 and 7) where the probability distributions p_i have the property that $p_i(F_M, F_K) \neq 0$ only if $K = M^c$ (here M^c denotes the complement of M in $\{1, 2, \ldots, n\}$). This assumption on the distributions p_i ensures that the elementary step of crossover associated to the reproduction $\frac{m}{2}$-tuple Ω is bijective. Depending on the further properties of the distributions p_i and the distribution $\wp_m$, different types of equivalence relations $\xrightarrow{\mathcal{A}}$ would be induced. Typically, in case of a classical GA crossover, the distributions p_i are all identical (i.e. $p_1 = p_2 = \cdots = p_{\frac{m}{2}} = p$) where p is the uniform distribution on $\{1, 2, \ldots, n\}$ and the distribution $\wp_m$ is uniform over all partitions of $\{1, 2, \ldots, n\}$ into 2-element subsets. In such a case the equivalence relation $\xrightarrow{\mathcal{A}}$ is determined by the numbers of 0's in the columns (or, equivalently, by the numbers of 1's in the columns). The reason this is so is that a population Q can be reached from a population P in by performing a sequence of crossover elementary steps only if it has the same amount of "genetic material" in every column since alleles are neither lost nor created during homologous crossover. Using the fact that every permutation can be obtained by performing enough transpositions, one can show the converse of this fact. This fact is a particular case of Lemma 47 of [4]. For instance, if $n = 5$ and $m = 4$ we have

$$\begin{pmatrix} 0&1&0&1&1 \\ 0&1&0&1&0 \\ 1&0&1&0&1 \\ 0&0&1&0&1 \end{pmatrix} \xrightarrow{\mathcal{A}} \begin{pmatrix} 0&1&0&0&1 \\ 1&0&1&0&1 \\ 0&0&0&1&0 \\ 0&1&1&1&1 \end{pmatrix}.$$

Indeed, the number of 0's in both populations in the first column is 3, in the 2nd, 3rd and 4th columns is 2 and in the last column is 1. Thus the equivalence class corresponding to a given population P can be described by an ordered n-tuple $[c]_P = (c_1, c_2, \ldots, c_n)$ of numbers between 0 and m where c_i is the number of 0s in the ith column of P. For example, if P is either one of the equivalent populations above then $[c]_P = (3, 2, 2, 2, 1)$.[2]

[2] One point crossover under reasonable assumptions will produce the same equivalence relation.

Example 20. Continuing with Example 19, consider the following family of mutation transformations $\mathcal{M} = \{T_{\mathbf{u}} | \mathbf{u} \in \Omega\}$ where each transformation $T_{\mathbf{u}}$ is defined as follows: Denote by $+_2$ the addition modulo 2 ($0 +_2 0 = 0$, $1 +_2 0 = 1$, $0 +_2 1 = 1$, $1 +_2 1 = 0$). We then define $T_{\mathbf{u}}$ to be the function from Ω into itself which sends every $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ to $T_{\mathbf{u}}(\mathbf{a}) = \mathbf{a} \oplus \mathbf{u}$ where $\oplus$ is componentwise addition modulo 2, i.e. given $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n) \in \Omega^n$, the $\oplus$ operation is defined as follows: $\mathbf{x} \oplus \mathbf{y} = \mathbf{z}$ where $\mathbf{z} = (z_1, z_2, \ldots, z_n)$ with $z_i = x_i +_2 y_i$. Notice first that every transformation $T_{\mathbf{u}}$ is bijective (in fact $T_{\mathbf{u}} \circ T_{\mathbf{u}} = \mathbf{1}$ where $\mathbf{1}$ is the identity map on Ω). Since every mutation transformation $T_{\mathbf{u}}$ is uniquely determined by the element $\mathbf{u} \in \Omega$, defining a probability distribution on the family $\mathcal{M}$ amounts to defining a probability distribution on $\Omega = \{0, 1\}^n$. To achieve a situation equivalent to the classical case where every bit is mutated independently with a small probability $\varepsilon > 0$ and remains unchanged with probability $1 - \varepsilon$, we choose 1 with probability ε and 0 with probability $1 - \varepsilon$ independently n times. Given a population of size m we let mutation be the elementary step associated to the reproduction m-tuple

$$\Omega_{\text{mutation}} = (\Omega, \mathcal{M}, \mathcal{M}, \ldots, \mathcal{M}, p, p, \ldots, p, \wp_m),$$

where p is the probability distribution on $\mathcal{M}$ described above and $\wp_m$ is the unique trivial probability distribution on the one-element set (since there is exactly one way to partition a given set into singleton subsets). Now let $\mathcal{B}$ denote the algorithm determined by the elementary step of crossover as described in Example 19 followed by the elementary step of mutation as described above. Then the algorithm $\mathcal{B}$ is ergodic in the sense of Definition 58 of [4] which means that the equivalence relation $\xrightarrow{\mathcal{B}}$ is trivial, i.e. there is only one equivalence class or, in other words, for any two populations P and Q we have $P \xrightarrow{\mathcal{B}} Q$. Indeed, thanks to the availability of mutation, any given population can be reached from any other given population in a single step with a small but a positive probability which means that any two given populations are equivalent under $\xrightarrow{\mathcal{B}}$.

The main result of [4] is the following fact:

Theorem 21. *Let $\mathcal{A}$ denote a bijective and self-transient algorithm. Then the Markov chain initiated at some population $P \in \Omega^m$ is irreducible and its unique stationary distribution is the uniform distribution* (*on* $[P]_{\mathcal{A}}$).

The classical versions of Geiringer theorem, such as the ones established in [3] and in [7] are stated in terms of the "limiting frequency of occurrence" of a certain element of the search space. The following definitions, which also appear in [4], make these notions precise in the finite population setting:

Definition 22. We define the characteristic function $\mathcal{X} : \Omega^m \times \mathcal{P}(\Omega) \to \mathbb{N} \cup \{0\}$ as follows: $\mathcal{X}(P, S) =$ the number of individuals of P which are the elements of S. (Recall that $P \in \Omega^m$ is a population consisting of m individuals and $S \in \mathcal{P}(\Omega)$ simply means that $S \subseteq \Omega$.)

Example 23. For instance, suppose $\Omega = \{0, 1\}^n$,

$$P = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

and $S \subseteq \Omega = \{0, 1\}^n$ is determined by the Holland schema $(*, 1, *, 1, *)$. Then $\mathcal{X}(P, S) = 3$ because exactly three rows of P, the 1st, the 2nd, and the 5th are in S.

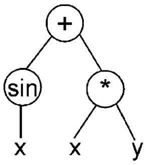
Definition 24. Fix an evolutionary algorithm $\mathcal{A}$ and an initial population $P \in \Omega^m$. Let $P(t)$ denote the population obtained upon the completion of t reproduction steps of the algorithm $\mathcal{A}$ in the absence of selection and mutation. For instance, $P(0) = P$. Denote by $\Phi(S, P, t)$ the proportion of individuals from the set S which occur before time t. That is, $\Phi(S, P, t) = \frac{\sum_{s=1}^{t} \mathcal{X}(P(s), S)}{tm}$. (Notice that tm is simply the total number of individuals encountered before time t. The same individual may be repeated more than once and the multiplicity contributes to Φ.) Denote by $\mathcal{X}(\Box, S) : \Omega^m \to \mathbb{N}$ the restriction of the function $\mathcal{X}$ when the set S is fixed (the notation suggests that one plugs a population P into the box).

Intuitively, $\Phi(S, P, t)$ is the frequency of encountering the individuals in S before time t when we run the algorithm starting with the initial population P.

6. Nonlinear genetic programming (GP) with homologous crossover

In GP, the search space, Ω, consists of the parse trees which usually represent various computer programs.

Example 25. A typical parse tree representing the program $(+(\sin(x), *(x, y)))$ is drawn below:



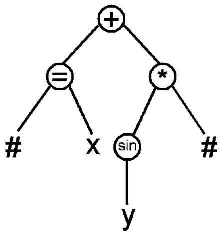
Since computers have only a finite amount of memory, it is reasonable to assume that there are finitely many basic operations which can be used to construct programs and that every program tree has depth less than or equal to some integer L. Under these assumptions Ω is a finite set. We may then define the search space as follows:

Definition 26. Fix a signature $\Sigma = (\Sigma_0, \Sigma_1, \Sigma_2, \ldots, \Sigma_N)$ where Σ_i's are finite sets.[3] We assume that $\Sigma_0 \neq \emptyset$ and $|\Sigma_j| \neq 1 \forall j$.[4] The search space Ω consists of all parse trees having depth at most L. Interior nodes having i children are labelled by the elements of Σ_i. The leaf nodes are labelled by the elements of Σ_0.

In order to study the appropriate family of reproduction (crossover) transformations with the aim of applying the generalized Geiringer theorem, it is most convenient to exploit Poli hyperschemata ([5] for a more detailed description).

Definition 27. A Poli hyperschema is a rooted parse tree which may have two additional labels for the nodes, namely # and = signs (it is assumed, of course, that neither one of these denotes an operation). The = sign may label any interior node v of the tree. Since v does occur in the tree, we must have $|\Sigma_i| > 0$. The # sign can only label a leaf node. A given Poli hyperschema represents the set of all programs whose parse tree can be obtained by replacing the = signs with any operation of the appropriate arities and attaching any program trees in place of the # signs. Different occurrences of # or = may be replaced differently. We shall denote by S_t the set of programs represented by a hyperschema t.

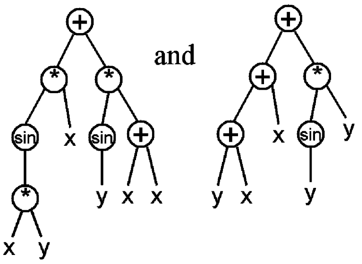
Consider, for instance, the hyperschema t defined as $(+(=(\#, x), *(\sin(y), \#))$ which is pictured below:



[3] Intuitively Σ_i is the set consisting of i-ary operations and Σ_0 consists of the input variables. Formally this does not have to be the case though.

[4] The assumption that $|\Sigma_j| \neq 1 \forall j$ does not cause any problems since we are free to select any elements from the search space that we want. On the other hand, this assumption helps us to avoid unnecessary complications when dealing with the poset of Poli hyperschemata later.

A couple of programs fitting the hyperschema t are shown below:



In order to model the family of reproduction (crossover) transformation in a way which makes it obvious that GP is a bijective and self-transient algorithm, we shall introduce a partial order on the set of all Poli hyperschema so that every two elements have the least upper bound. The notion of the least upper bound will be also used to define the *common region* (see [6] for an alternative description of the notion of a common region).

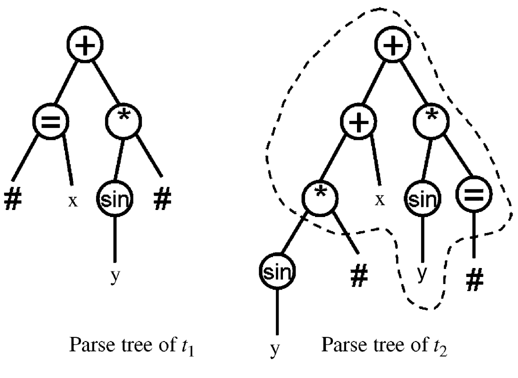
Definition 28. Denote by $\mathcal{O}$ the set of all basic operations which can be used to construct the programs (i.e. $\mathcal{O} = \Sigma_1 \cup, \ldots, \cup \Sigma_N$) and by $\mathcal{V}$ the set of all variables (i.e. $\mathcal{V} = \Sigma_0$). Put the following partial order, $\preccurlyeq$, on the set $\mathcal{O} \cup \mathcal{V} \cup \{=, \#\}$:
1. $\forall a, b \in \mathcal{O} \cup \mathcal{V}$ we have $a \preccurlyeq b \Longleftrightarrow a = b$.
2. $\forall a \in \mathcal{O}$ we have $a \preccurlyeq\, =$.
3. $\forall a \in \mathcal{O} \cup \mathcal{V}$ we have $a \preccurlyeq \#$.
4. $= \preccurlyeq =, \# \preccurlyeq \#$ and $= \preccurlyeq \#$.

We shall also write $a \succcurlyeq b$ to mean $b \preccurlyeq a$.

It is easy to see that $\preccurlyeq$ is, indeed a partial order. Moreover, every collection of elements of $\mathcal{O} \cup \mathcal{V} \cup \{=, \#\}$ has the least upper bound under $\preccurlyeq$. We are now ready to define the partial order relation on the set of all Poli hyperschemata:

Definition 29. Let t_1 and t_2 denote two Poli hyperschemata. We say that $t_1 \geqslant t_2$ if and only if the following two conditions are satisfied:
1. The tree corresponding to t_1 when all of the labels are deleted is a subtree of the tree corresponding to t_2 with all of the labels deleted.
2. Every one of the labels (which represents an operation or a variable) of t_1 is $\succcurlyeq$ the label of the node in the corresponding position of t_2.

Example 30. For instance, the hyperschema $t_1 = (+(= (\#, x)), *(\sin(y), \#)) \geqslant t_2 = (+(+(*(\sin(x), y), x)), *(\sin(y), = (\#)))$. Indeed, the parse trees of t_1 and t_2 appear on the picture below:



Parse tree of t_1

Parse tree of t_2

When all the labels in the dashed subtree of the parse tree of t_2 are deleted one gets the tree isomorphic to that obtained from t_1 by deleting all the labels. Thus condition 1 of Definition 29 is satisfied. To see that condition 2 is fulfilled as well, we notice that the labels of t_1 are $\preccurlyeq$ to the corresponding labels of the dashed subtree of t_2: Indeed, we have $+ \succcurlyeq +$, $= \succcurlyeq +$, $* \succcurlyeq *$, $\# \succcurlyeq *$, $x \succcurlyeq x$, $\sin \succcurlyeq \sin$, $\# \succcurlyeq =$ and $y \succcurlyeq y$.

Again it is easy to check that $\geqslant$ is, indeed, a partial order relation on the collection of Poli hyperschemata. Proposition 31, tells us even more:

Proposition 31. *Any given collection of Poli hyperschemata has the least upper bound under* $\geqslant$.

Proof. Denote by $\mathcal{S}$ a given collection of Poli hyperschemata. We provide an algorithm to construct the least upper bound of $\mathcal{S}$ as follows: Copies of all the trees in $\mathcal{S}$ are recursively jointly traversed starting from the root nodes to identify the parts with the same shape, i.e. the same arity in the nodes visited. Recursion is stopped as soon as an arity mismatch between corresponding nodes in some two trees from $\mathcal{S}$ is present. All the nodes and links encountered are stored. This way we obtain a tree. It remains to stick in the labels. Each one of the interior nodes is labelled by the least upper bound of the corresponding labels of the trees in $\mathcal{S}$. The label of a leaf node is a variable, say x, if all the labels of the corresponding nodes of the trees in $\mathcal{S}$ are x (which implies that they are leaf nodes themselves). In all other cases the label of the leaf node is the # sign. It is not hard to see that this produces the least upper bound of the collection $\mathcal{S}$ of parse trees. □

It was pointed out before, that programs themselves are Poli hyperschemata. The following fact is almost immediate from the explicit construction of the least upper bound carried out in the proof of Proposition 31:

Proposition 32. *A given Poli hyperschema t is the least upper bound of the set S_t of programs determined by t.*

From Proposition 32 it follows easily that $\geqslant$ is order isomorphic to the collection of subsets determined by the Poli hyperschemata:

Proposition 33. *Let t and s denote Poli hyperschemata. Denote by S_t and S_s the subsets of the search space determined by the hyperschemata t and s, respectively. Then $t \geqslant s \iff S_t \supseteq S_s$.*

There is another type of schemata which is useful to introduce in order to define the family of reproduction (crossover) transformations:

Definition 34. A shape schema is just a rooted ordered tree. If $\tilde{t}$ is a given shape schema then $S_{\tilde{t}}$ is just the set of all programs whose underlying tree when all the labels are deleted is precisely $\tilde{t}$. Given a Poli hyperschema s, we shall denote by $\tilde{s}$ the underlying shape schema of s, i.e. the tree obtained by deleting all the labels in s.

The notion of a common region which is equivalent to the one defined below also appears in [6]:

Definition 35. Given two Poli hyperschemata t and s we define their common region to be the underlying shape schema of the least upper bound of t and s.

Definition 36. Fix a shape schema $\tilde{t}$. We shall say that the set $C_{\tilde{t}} = \{(a, b) | a, b$ are program trees and $\tilde{t}$ is the common region of a and $b\}$ is a component corresponding to the shape $\tilde{t}$.

Notice that sets determined by the shape schemata partition the search space:

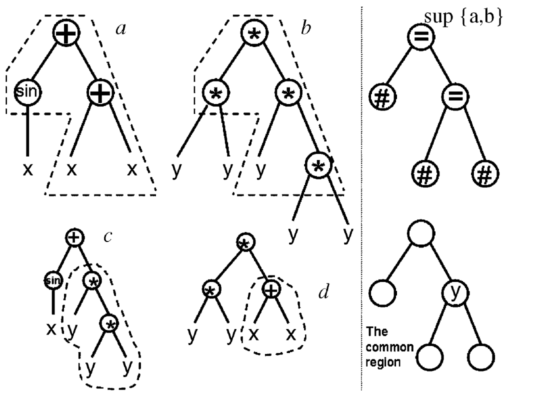
Remark 37. Notice that $\Omega^2 = \bigcup_{\tilde{t} \text{ is a shape}} C_{\tilde{t}}$. Moreover, $C_{\tilde{t}} \cap C_{\tilde{s}} = \emptyset$ for $t \neq s$. (This is so because least upper bounds in a poset are uniquely determined and so the function sending $(a, b) \to \sup(a, b) \to$ the underlying shape of $\sup(a, b)$ is well defined. But then the sets $C_{\tilde{t}}$ are simply the pre-images under this function of the individual shape schemata and, hence, form a partition of Ω^2.)

We now proceed to define the family of reproduction transformations. Our goal is to introduce a family of functions on Ω^2 in such a way that each one of them is easily seen to be bijective (see Theorem 21, Definitions 13, 15 and Remark 16). The idea is to define these transformations on each of the components first:

Definition 38. Fix a shape schema $\tilde{t}$. Fix a node, v of $\tilde{t}$. A one-point partial homologous crossover transformation $T_v : C_{\tilde{t}} \to C_{\tilde{t}}$ is defined as follows: For given $(a, b) \in C_{\tilde{t}}$ let $T_v(a, b) = (c, d)$ where c and d are obtained from the program trees of a and b as follows: First identify the node v in the parse trees of a and b, respectively. Now obtain the pair (c, d) by swapping the subtrees of a and b rooted at v. (This procedure is described in detail in [6] and it is also illustrated in the example below). Let $\mathcal{G}_{\tilde{t}} = \{T_v \mid v \text{ is a node of } \tilde{t}\}$ denote the family of all partial homologous one-point crossover transformations associated to the shape $\tilde{t}$.

The following example illustrates the concepts in Definitions 34, 35 and 38:

Example 39. In the upper left part of the picture parse trees of the two sample programs a and b are shown. Then on the upper right one can see the least upper bound of a and b. On the lower right the underlying tree of the least upper bound of a and b is drawn. According to Definition 35, this tree is precisely the common region of the programs a and b. The isomorphic subtrees inside both, a and b, are emphasized inside the dashed areas:



A node v is selected inside the common region. The pair of children $(c, d) = T_v(a, b)$ appears on the lower left of the picture above. The subtrees rooted at v which are swapped during crossover are emphasized inside the dashed area.

Remark 40. One does need to show that for $(a, b) \in C_{\tilde{t}}$ we have $T_v(a, b) \in C_{\tilde{t}}$. A rigorous argument can be given as follows: Clearly $T_v : C_{\tilde{t}} \to \bigcup_{\tilde{t} \text{ is a shape}} C_{\tilde{t}}$ is a well-defined map. Moreover, since v is a node of the least upper bound of a and b and the pair (c, d) is obtained simply by swapping the corresponding subtrees rooted at v, we get $s = \sup\{c, d\} \leqslant \sup\{a, b\}$. Now consider the transformation $F_v : C_{\tilde{s}} \to \bigcup_{\tilde{t} \text{ is a shape}} C_{\tilde{t}}$ and notice that, by definition, we have $F_v(c, d) = (a, b)$. But then, according to the reasoning above, we have $\sup\{c, d\} \leqslant \sup\{a, b\}$. Thereby, we get $\sup\{c, d\} \leqslant \sup\{a, b\} \leqslant \sup\{c, d\} \Longrightarrow \sup\{c, d\} = \sup\{a, b\} \Longrightarrow \tilde{t} = \tilde{s}$. This shows that T_v does, indeed, map into $C_{\tilde{v}}$. Moreover, in the process, we have also observed a couple of very important facts:

1. $T_v \circ T_v = \mathbf{1}_{C_{\tilde{t}}}$ where $\mathbf{1}_{C_{\tilde{t}}}$ denotes the identity map on $C_{\tilde{t}}$. This shows, in particular, that T_v is a bijection.
2. T_v preserves the least upper bounds: $\sup\{a, b\} = \sup T_v(a, b)$.

We are finally ready to define the family of reproduction transformations on the search space Ω of all programs:

Definition 41. For every shape schema $\tilde{t}$ fix a node $v_{\tilde{t}}$ of $\tilde{t}$. Define a one point crossover transformation $T_{\{v_{\tilde{t}}\}_{\tilde{t} \text{ is a shape schema}}} : \Omega^2 \to \Omega^2$ to be the set-theoretic union of all partial crossover transformations of the form $T_{v_{\tilde{t}}}$. More explicitly, this means that whenever a given pair $(a, b) \in \Omega^2$ we must have $(a, b) \in C_{\tilde{s}}$ for a unique shape schema $\tilde{s}$ (since, according to Remark 37, Ω^2 is a disjoint union of components corresponding to various shapes). But then $T_{\{v_{\tilde{t}}\}_{\tilde{t} \text{ is a shape schema}}}(a, b) = T_{v_{\tilde{s}}}(a, b)$. Denote by $\mathcal{G}$ the family of all crossover transformations together with the identity map on Ω^2. For simplicity of notation we shall denote the transformations in $\mathcal{G}$ by plain English letters: T, F etc., keeping in mind that every such transformation is determined by making choices of partial crossover transformations on every one of the components.

Remark 42. Thanks to Remark 40, every one of the crossover transformations in the family $\mathcal{G}$ is bijective (since it is a union of bijections on the pieces of a partition). It follows now that the generalized Geiringer theorem (Theorem 21) applies to the case of homologous GP.

Remark 43. It is also possible to model uniform GP crossover (this type of crossover is examined in detail in [6]) in the analogous manner. All of the results established in the current paper apply to this case without any modification.

7. The statement of the schema-based version of Geiringer's theorem for non-linear GP under homologous crossover

As mentioned before, the schema-based version of Geiringer's theorem for non-linear GP is stated in terms of Poli hyperschemata.

Definition 44. A Poli hyperschema of order i is a Poli hyperschema which has exactly i nodes whose label is not a # or an $=$ sign.

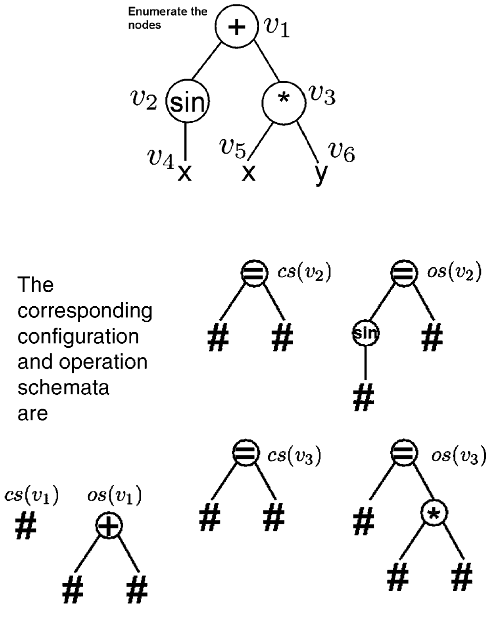
A configuration schema is a 0-order Poli hyperschema (i.e. a hyperschema which has only the equal signs in the interior nodes and # signs in the leaf nodes.)

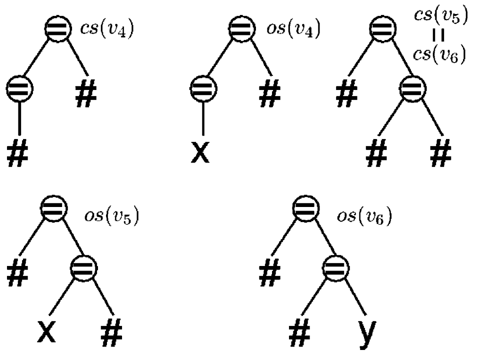
An operation schema is a Poli hyperschema of order 1 (i.e. a hyperschema which has exactly one node whose label is not a # or an $=$ sign).

Fix an individual (a parse tree) $\mathbf{u} \in \Omega$. Let v denote any node of $\mathbf{u}$. Let $B(v)$ denote the branch of the shape schema of $\mathbf{u}$ from the root down to the node v. Let $B^+(v) = B(v) \cup \{w | w \text{ is a child of some node } z \text{ of } B \text{ with } z \neq v\}$. Now define $cs(v)$ to be the configuration schema whose underlying shape schema is $B^+(v)$. Let o denote an operation or a variable (an element of Σ_i for some i between 0 and N). Now obtain the operation schema os_o from $cs(v)$ by attaching the node labelled by o in place of the # sign at the node corresponding to v of $cs(v)$. Unless v is the leaf node of $\mathbf{u}$, all the children of this new node are the leaf nodes of os_o labelled by the # sign. When o is the operation (or the variable) labelling the node v of $\mathbf{u}$, we shall write $os(v)$ instead of os_o.

Notice that if v is a root node then $cs(v)$ is just the schema which determines the entire search space, i.e. the parse tree consisting of a single node labelled by the # sign. Example 45 illustrates Definition 44.

Example 45. Below we list all of the configuration schemata and operation schemata for the individual of Example 25:





Recall from Definition 22 that $\mathcal{X}(P, S)$ denotes the number of individuals in the population P which are the elements of $S \subseteq \Omega$. The following definition makes it more convenient to state the schema-based version of Geiringer's theorem:

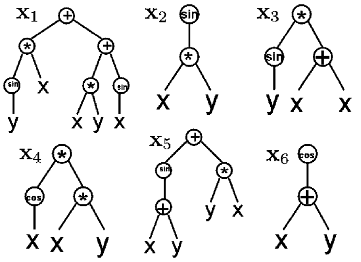
Definition 46. Given a Poli hyperschema H, we shall write $|H(P)|$ in place of $\mathcal{X}(P, S_H)$ (see Definition 27) to denote the number of individuals (counting repetitions) in the population P fitting the hyperschema H.

We can now finally state the Geiringer's theorem for nonlinear GP under homologous crossover:

Theorem 47. *Fix an initial population $P \in \Omega^m$ and an individual $\mathbf{u} \in \Omega$. Suppose every pair of individuals has a positive probability to be paired up for crossover and every transformation in $\mathcal{G}$ has a positive probability of being chosen.*[5] *Then the limiting frequency of occurrence of a given individual* $\mathbf{u}$,

$$\lim_{t\to\infty} \Phi(\mathbf{u}, P, t) = \prod_{v \text{ is a node of } \mathbf{u}} \frac{|os(v)(P)|}{|cs(v)(P)|}.$$

Example 48. To illustrate how Theorem 47 can be applied in practice, suppose we are interested in computing the frequency of encountering the individual $\mathbf{u}$ from Examples 25 and 45 when the initial population of 6 individuals pictured below is chosen:



The number of individuals in P fitting the operation schema $os(v_1)$ is 2 (these are $\mathbf{x}_1$ and $\mathbf{x}_5$) while every individual fits the configuration schema $cs(v_1)$. Therefore, $\frac{|os(v_1)(P)|}{|cs(v_1)(P)|} = \frac{2}{6} = \frac{1}{3}$. Four individuals, namely $\mathbf{x}_1$, $\mathbf{x}_3$, $\mathbf{x}_4$ and $\mathbf{x}_5$ fit $cs(v_2) = cs(v_3)$, among these only two individuals, namely $\mathbf{x}_3$ and $\mathbf{x}_5$, fit $os(v_2)$ and two individuals, $\mathbf{x}_4$ and $\mathbf{x}_5$ fit $os(v_3)$ so that $\frac{|os(v_2)(P)|}{|cs(v_2)(P)|} = \frac{|os(v_3)(P)|}{|cs(v_3)(P)|} = \frac{2}{4} = \frac{1}{2}$. Individuals $\mathbf{x}_3$, $\mathbf{x}_4$ and $\mathbf{x}_5$ fit the configuration schema $cs(v_4)$ while only $\mathbf{x}_4$ fits the operation schema $os(v_4)$ so that $\frac{|os(v_4)(P)|}{|cs(v_4)(P)|} = \frac{1}{3}$. $\mathbf{x}_1$, $\mathbf{x}_3$, $\mathbf{x}_4$ and $\mathbf{x}_5$ fit $cs(v_5) = cs(v_6)$. Among these only $\mathbf{x}_3$ and $\mathbf{x}_4$ fit $os(v_5)$ while only $\mathbf{x}_4$ fits $os(v_6)$ so that $\frac{|os(v_5)(P)|}{|cs(v_5)(P)|} = \frac{2}{4} = \frac{1}{2}$ and $\frac{|os(v_6)(P)|}{|cs(v_6)(P)|} = \frac{1}{4}$. Thereby, according to Theorem 47, we obtain

$$\lim_{t\to\infty} \Phi(\mathbf{u}, P, t) = \prod_{i=1}^{6} \frac{|os(v_i)(P)|}{|cs(v_i)(P)|} = \frac{1}{3} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{288}.$$

[5] These conditions can be slightly relaxed, but we try to present the main idea only.

Roughly speaking, this means that if we run GP starting with the population P pictured above, in the absence of mutation and selection (crossover being the only step) for an infinitely long time, the individual $\mathbf{u}$ will be encountered on average 1 out of 288 times.

Example 49. Notice that linear GP (or, equivalently, variable length GA) as described in [7] is a special case of nonlinear GP when $\forall i > 1\ \Sigma_i = \emptyset$ and Σ_0 and $\Sigma_1 \neq \emptyset$. Indeed, the elements of such a search space are parse trees such that every interior node has exactly one child and the depth of the tree is bounded by some integer N. One can think of such a tree as a sequence of labels $(a_1, a_2, \ldots, a_n)$, the first label affiliated with the root node, second label with the child of the root node and so on. The label a_n is affiliated with the leaf node. This gives us a one-to-one correspondence, call it ϕ between the search space for nonlinear GP in our specific case when $\forall i > 1\ \Sigma_i = \emptyset$ while Σ_0 and $\Sigma_1 \neq \emptyset$ and the search space for linear GP which preserves crossover. The following types of schemata have been introduced in [7]:

Definition 50. The schema $H = (*^{i-1}, h_i, \#)$ represents the subset $S_H = \{\mathbf{x} = (x_1, x_2, \ldots, x_l) | l > i \text{ and } x_i = h_i\}$. In words, S_H is simply the set of all individuals whose length is at least $i + 1$ and whose ith allele is h_i.

Definition 51. The schema $H = (*^i, \#)$ represents the subset

$$S_H = \{\mathbf{x} = (x_1, x_2, \ldots, x_l) | l > i\}.$$

In words, S_H is simply the subset of all individuals whose length is at least $i + 1$.

Definition 52. The schema $H = (*^{i-1}, h_i)$ represents the subset

$$S_H = \{\mathbf{x} = (x_1, x_2, \ldots, x_i) | x_i = h_i\}$$

of the search space which is simply the set of all individuals of length exactly equal to i whose ith (last) allele is h_i.

The reader may check that under the correspondence ϕ the configuration schemata correspond to the schemata $H_i = (*^i, \#)$ for $i \geqslant 1$, operation schemata correspond to the schemata of the form $H = (*^{i-1}, h_i, \#)$ and of the form $H = (*^{i-1}, h_i)$ for $i > 1$. Finally, the hyperschema $t_{(1,1)}$ corresponds to the schema $H = (h_1, \#)$. Fix a population $P \in \Omega^m$. Recall that we denote by $|H|$ the number of individuals in P which fit the schema H counting repetitions. Also recall from Definition 24 that $\Phi(S_H, P, 1) = \frac{|H|}{m}$ denotes the fraction of the number of individuals of P which fit the schema H. To abbreviate the notation we shall write $\Phi(H, P, 1)$ instead of $\Phi(S_H, P, 1)$. Fix an individual $\mathbf{u} = (h_1, h_2, \ldots, h_n) \in \Omega$. Theorem 47 tells us that

$$\begin{aligned}
\lim_{t\to\infty} \Phi(\mathbf{u}, P, t) &= \frac{|(h_1, \#)|}{m} \cdot \left(\prod_{i=1}^{n-2} \frac{|(*^i, h_{i+1}, \#)|}{|(*^i, \#)|} \right) \cdot \frac{|(*^{n-1}, h_n)|}{|(*^{n-1}, \#)|} \\
&= \frac{|(h_1, \#)|}{m} \cdot \left(\prod_{i=1}^{n-2} \frac{\frac{|(*^i, h_{i+1}, \#)|}{m}}{\frac{|(*^i, \#)|}{m}} \right) \cdot \frac{\frac{|(*^{n-1}, h_n)|}{m}}{\frac{|(*^{n-1}, \#)|}{m}} \\
&= \Phi(h_1, \#) \cdot \left(\prod_{i=1}^{n-2} \frac{\Phi(*^i, h_{i+1}, \#)}{\Phi(*^i, \#)} \right) \cdot \frac{\Phi(*^{n-1}, h_n)}{\Phi(*^{n-1}, \#)} \\
&= \Phi(*^{n-1}, h_n) \cdot \frac{\prod_{i=n-2}^{0} \Phi(*^i, h_{i+1}, \#)}{\prod_{i=n-1}^{1} \Phi(*^i, \#)} \\
&= \Phi(*^{n-1}, h_n) \cdot \prod_{i=n-1}^{i=1} \frac{\Phi(*^{i-1}, h_i, \#)}{\Phi(*^i, \#)}
\end{aligned}$$

which is precisely the formula obtained in [7].

8. How do we obtain Theorem 47 from Theorem 21?

The following couple of corollaries from [4] are useful in obtaining the schema-based versions of Geiringer theorem for various evolutionary algorithms. Throughout, we shall denote by $\varrho_{[P]_{\mathcal{A}}}$ the uniform probability distribution on the set $[P]_{\mathcal{A}}$ (see Definition 18).

Corollary 53. *Fix a bijective and self-transient algorithm $\mathcal{A}$ and an initial population $P \in \Omega^m$. Fix a set S of individuals in Ω ($S \subseteq \Omega$). Then* $\lim_{t\to\infty} \Phi(S, P, t) = \frac{1}{m} E_{\varrho_{[P]_{\mathcal{A}}}}(\mathcal{X}(\Box, S))$ *(here $E_{\varrho_{[P]_{\mathcal{A}}}}(f)$ denotes the expectation of the random variable f with respect to the uniform distribution on the set $[P]_{\mathcal{A}}$).*[6]

To state the next corollary which brings us one step closer to deriving results similar in flavor to Geiringer's original theorem we need one more, purely formal, assumption about the algorithm:

Definition 54. We say that a given algorithm $\mathcal{A}$ is regular if the following is true: for every population $P = (x_1, x_2, \ldots, x_m) \in \Omega^m$ and for every permutation $\pi \in \mathcal{S}_m$, the population obtained by permuting the elements of P by π, namely $\pi(P) = (x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(m)}) \in [P]_{\mathcal{A}}$. In words this says that the equivalence classes $[P]_{\mathcal{A}}$ are permutation invariant.

Remark 55. Definition 54 is only needed because our description of an evolutionary search algorithm uses the ordered multi-set model. This makes the generalized Geiringer theorem (Theorem 21) look nice (the stationary distribution is uniform on $[P]_{\mathcal{A}}$). A disadvantage of the multi-set model is that it allows algorithms which are not regular. If we were to use the model of [13] where the order of elements in a population is not taken into account (a reasonable assumption since most evolutionary algorithms used in practice are, indeed, regular) then the generalized Geiringer theorem would have to be modified accordingly since the stationary distribution of the corresponding Markov process would be different from uniform (it is not difficult to compute it though since the corresponding Markov chain is just a "projection" of the one used in the current paper).

Corollary 56. *Fix a regular bijective and self-transient algorithm $\mathcal{A}$ and an initial population $P \in \Omega^m$. Denote by $\varrho_{[P]_{\mathcal{A}}}$ the uniform probability distribution on $[P]_{\mathcal{A}}$ (see Definition* 18). *Fix a set S of individuals in Ω ($S \subseteq \Omega$). Then we have* $\lim_{t\to\infty} \Phi(S, P, t) = \varrho_{[P]_{\mathcal{A}}}(\mathcal{V}_S)$ *where*

$$\mathcal{V}_S = \{P \mid P \in [P]_{\mathcal{A}} \textit{ and the } 1st \textit{ individual of } P \textit{ is an element of } S\}.$$

Corollaries 53 and 56 are proved in Section 6 of [4]. When deriving schema-based versions of Geiringer theorem for a specific algorithm the following strategy may be implemented: Continuing with the notation in Corollaries 53 and 56, suppose we are given a nested sequence of subsets of the search space: $S_1 \supseteq S_2 \supseteq \cdots \supseteq S_n$. According to Corollary 56,

$$\lim_{t\to\infty} \Phi(S_n, P, t) = \varrho_{[P]_{\mathcal{A}}}(\mathcal{V}_{S_n}) = \frac{|\mathcal{V}_{S_n}|}{|[P]_{\mathcal{A}}|} = \frac{|\mathcal{V}_{S_n}|}{|\mathcal{V}_{S_{n-1}}|} \cdot \frac{|\mathcal{V}_{S_{n-1}}|}{|[P]_{\mathcal{A}}|} = \frac{|\mathcal{V}_{S_n}|}{|\mathcal{V}_{S_{n-1}}|} \cdot \frac{|\mathcal{V}_{S_{n-1}}|}{|\mathcal{V}_{S_{n-2}}|} \cdot \ldots \cdot \frac{|\mathcal{V}_{S_2}|}{|\mathcal{V}_{S_1}|} \cdot \frac{|\mathcal{V}_{S_1}|}{|[P]_{\mathcal{A}}|}$$
$$= \varrho_{[P]_{\mathcal{A}}}(\mathcal{V}_{S_1}) \cdot \prod_{j=0}^{n-2} \frac{|\mathcal{V}_{S_{n-j}}|}{|\mathcal{V}_{S_{n-j-1}}|} = \frac{1}{m} E_{\varrho_{[P]_{\mathcal{A}}}}(\mathcal{X}(\Box, S)) \cdot \prod_{j=0}^{n-2} \frac{|\mathcal{V}_{S_{n-j}}|}{|\mathcal{V}_{S_{n-j-1}}|}.$$

Notice that $\frac{|\mathcal{V}_{S_j}|}{|\mathcal{V}_{S_{j-1}}|}$ is just the proportion of populations in $[P]_{\mathcal{A}}$ whose first individual is a member of S_j inside the set of populations in $[P]_{\mathcal{A}}$ whose first individual is a member of S_{j-1}.

Corollary 57. *Fix a regular, bijective and self-transient algorithm $\mathcal{A}$ and an initial population $P \in \Omega^m$. Fix a nested sequence of subsets $S_1 \supseteq S_2 \supseteq \cdots \supseteq S_n$ of individuals in Ω ($S_1 \subseteq \Omega$). Then* $\lim_{t\to\infty} \Phi(S_n, P, t) = \frac{1}{m} E_{\varrho_{[P]_{\mathcal{A}}}}(\mathcal{X}(\Box, S)) \cdot \prod_{j=0}^{n-2} \frac{|\mathcal{V}_{S_{n-j}}|}{|\mathcal{V}_{S_{n-j-1}}|}$ *where, as before, $\mathcal{V}_S$ denotes the set of all populations whose first individual is a member of S for a given subset $S \subseteq \Omega$.*

[6] Throughout the paper, whenever a limit is involved, the equality is meant to hold for almost every infinite sequence of trials.

Denote by $\mathcal{A}$ a given GP algorithm. Fix an individual $\mathbf{u} \in \Omega$. In order to apply Corollary 57, we may choose a descending chain of Poli hyperschemata $t_1 \geqslant t_2 \geqslant \cdots \geqslant t_n = \mathbf{u}$. Fix an initial population P. To avoid putting many subscripts, we shall write $\mathcal{V}_t$ instead of $\mathcal{V}_{S_t}$ for the set of all populations in $[P]_{\mathcal{A}}$ (see Definition 11) whose first individual is a member of S_t (the set of individuals determined by the hyperschema t). In order to construct the desired sequence of nested hyperschemata, we assign the following numerical labelling to the nodes of the parse tree of **u**: The nodes are labelled by the pairs of integer coordinates. The first coordinate shows the depth of the tree and the second coordinate shows how far to the right a given node at the depth specified by the first coordinate is located. Notice, for instance, that the root node is labelled by the coordinates $(1, 1)$. We also introduce the following lexicographic linear ordering on the set of coordinate pairs:

Definition 58. $(a, b) \leqslant (c, d)$ if and only if either $a \leqslant c$ or ($a = c$ and $b \leqslant d$).

It is well known and easy to verify that this defines a linear ordering.

Definition 59. Given a pair of coordinates (i, j), denote by $\uparrow\ (i, j)$ the immediate successor of (i, j) under the lexicographic ordering defined above. Explicitly,

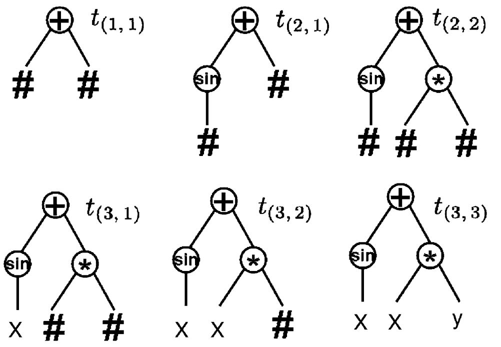
$$\uparrow (i, j) = \begin{cases} (i+1, 1) & \text{if } (i, j) \text{ labels the rightmost node of } \mathbf{u} \text{ at depth } i, \\ (i, j+1) & \text{otherwise.} \end{cases}$$

We obtain the desired nested sequence of hyperschemata for the given individual **u** recursively in the following manner:

Definition 60. Define $t_{(1,1)}$ to be the hyperschema whose root node has the same label (operation) and arity as that of the root node of **u**. All children of the root node are the leaf nodes labelled by the # sign. Once the hyperschema $t_{(i,j)}$ has been constructed, we obtain the hyperschema $t_{\uparrow(i,j)}$ by attaching the node of **u** with coordinate $\uparrow\ (i, j)$ in place of the # sign at coordinate $\uparrow\ (i, j)$ to the parse tree of $t_{(i,j)}$. Unless this node, call it v, is a leaf node of **u**, all children of this new node are the leaf nodes of $t_{\uparrow(i,j)}$ labelled by the # sign.

We illustrate the construction with an explicit example:

Example 61. Below, the nested sequence $t_{(1,1)} \geqslant t_{(2,1)} \geqslant t_{(2,2)} \geqslant t_{(3,1)}, \geqslant t_{(3,2)} \geqslant t_{(3,3)}$ corresponding to the program of Example 25 is drawn explicitly:



The formula for the limiting frequency of occurrence of a given program u in Corollary 57 involves the ratios of the form $\frac{\mathcal{V}_{t_{\uparrow(i,j)}}}{\mathcal{V}_{t_{(i,j)}}}$. It turns out that these ratios can be expressed nicely in terms of the presence of certain configuration and operation schemata in the initial population P:

Definition 62. Given a program tree $\mathbf{u}$ and the corresponding nested sequence $t_{(1,1)} \geqslant t_{(2,1)} \geqslant \cdots \geqslant t_{(i,j)} \geqslant t_{\uparrow(i,j)}, \geqslant \cdots \geqslant t_{(l,k)} = \mathbf{u}$ of hyperschemata as in Definition 60, for every $(i, j) \neq (l, k)$, denote by $cs_{(i,j)}$ $(os_{(i,j)})$ the configuration schema $cs(v)$ (operation schema $os(v)$) where v is the node of $\mathbf{u}$ with coordinate $\uparrow (i, j)$.

Example 63. Continuing with Examples 25 and 45 notice that for the individual in these examples we have $cs_{(1,1)} = cs_{(2,1)} = cs(v_2) = cs(v_3)$ while $os_{(1,1)} = os(v_2)$ and $os_{(2,1)} = os(v_3)$ (see Example 45), $cs_{(2,2)} = cs(v_4)$ while $os_{(2,2)} = os(v_4)$ and $cs_{(3,1)} = cs_{(3,2)} = cs(v_5) = cs(v_6)$ while $os_{(3,1)} = os(v_5)$ and $os_{(3,2)} = os(v_6)$.

The following “orbit description” lemma is the reason for introducing configuration and operation schemata: We prove the lemma under the following special assumption:

Definition 64. We say that a population P is special with respect to the individual $\mathbf{u}$ if for every node v of $\mathbf{u}$ and for every operation (or variable) o we have $|os_o(P)| \leqslant 1$ where os_o is obtained from $cs(v)$ by means of attaching the operation o at the leaf node of $cs(v)$ corresponding to v as described in Definition 44.

Definition 64 basically requires that no 2 operations (or variables) occurring in P at the specified location are the same. It turns out that the orbit description lemma stated below is a lot more convenient to prove under this special assumption. The general case will then follow by introducing enough extra labels for the operations and variables involved and then deleting the extra labels.

Lemma 65. *Fix an initial population P and a program $\mathbf{u} \in \Omega$. Assume that the population P is special with respect to the individual $\mathbf{u}$. Suppose every pair of individuals has a positive probability to be paired up for crossover and every transformation in $\mathcal{G}$ has a positive probability of being chosen.*[7] *Consider the sequences of hyperschemata $t_{(1,1)} \geqslant t_{(2,1)} \geqslant \cdots \geqslant t_{(i,j)} \geqslant t_{\uparrow(i,j)}, \geqslant \cdots \geqslant t_{(l,k)} = \mathbf{u}$, $\{cs_{(i,j)} | (i, j)$ is a coordinate of $\mathbf{u}$, (i, j) is not the maximal coordinate$\}$ and $\{os_{(i,j)} | (i, j)$ is a coordinate of $\mathbf{u}$, (i, j) is not the maximal coordinate$\}$ corresponding to the individual $\mathbf{u}$. For a given hyperschema t, denote by $|t(P)|$ the number of individuals in P which fit the hyperschema t counting repetitions. Suppose $\forall$ non-maximal pairs of coordinates (i, j) we have $|os_{(i,j)}(P)| \neq 0$ and $|t_{(1,1)}(P)| \neq 0$. Then it is true that $\forall (i, j) \frac{|\mathcal{V}_{t_{\uparrow(i,j)}}|}{|\mathcal{V}_{t_{(i,j)}}|} = \frac{1}{|cs_{(i,j)}(P)|}$.*

Proof. The key idea is to observe the following fact:

Claim. *Fix a coordinate (i, j). Fix any two operation schemata os_1 and os_2 which are obtained from $cs_{(i,j)}$ by attaching either a variable or an operation at the node (i, j). Suppose $\exists$ individuals in P fitting both, os_1 and os_2. Then $|\mathcal{V}_{t_{(i,j)}} \cap \mathcal{V}_{os_1}| = |\mathcal{V}_{t_{(i,j)}} \cap \mathcal{V}_{os_2}|$.*

Proof. Consider the map $F : [P]_{\mathcal{A}} \to [P]_{\mathcal{A}}$ defined as follows: Given a population, say $Q \in [P]_{\mathcal{A}}$, notice that $\exists$ an individual, say $\mathbf{x}_1$, in Q fitting the operation schema os_1 (due to the way crossover is defined, the number of individuals fitting the operation schema $os_1(Q)$ is the same in every population $Q \in [P]_{\mathcal{A}}$). Moreover, such an individual is unique since we assumed that all operations appearing in the individuals of P are distinct. Likewise, $\exists$ unique individual in Q, say $\mathbf{x}_2$ fitting the operation schema os_2. Pair up individuals $\mathbf{x}_1$ and $\mathbf{x}_2$ and pair up the rest of the individuals arbitrarily for crossover. Select the crossover transformation T_v where v is the node with coordinate (i, j) for the pair $(\mathbf{x}_1, \mathbf{x}_2)$ and choose the identity transformation for the rest of the pairs. Now let $F(Q)$ be the population obtained upon the completion of the reproduction step described above (notice that $F(Q) \in [P]_{\mathcal{A}}$ by definition of $[P]_{\mathcal{A}}$). Notice also that F is its own inverse (i.e. $F \circ F = \mathbf{1}_{[P]_{\mathcal{A}}}$). This tells us, in particular, that F is bijective. Moreover, it is clear from the definitions that $F(\mathcal{V}_{t_{(i,j)}} \cap \mathcal{V}_{os_1}) \subseteq \mathcal{V}_{t_{(i,j)}} \cap \mathcal{V}_{os_2}$ and, likewise, $F(\mathcal{V}_{t_{(i,j)}} \cap \mathcal{V}_{os_2}) \subseteq \mathcal{V}_{t_{(i,j)}} \cap \mathcal{V}_{os_1}$. The desired conclusion follows at once. □

Now observe that $t_{\uparrow(i,j)} = t_{(i,j)} \cap os_{(i,j)}$ so that $\mathcal{V}_{t_{\uparrow(i,j)}} = \mathcal{V}_{t_{(i,j)}} \cap \mathcal{V}_{os_{(i,j)}}$ and $t_{(i,j)} = \bigcup_{o \text{ is an operation or a variable}} (t_{(i,j)} \cap os_o)$ where os_o is obtained from $cs_{(i,j)}$ by attaching the operation (or variable) o at the node $\uparrow (i, j)$. Therefore,

[7] These conditions can be slightly relaxed, but we try to present the main idea only.

we also have $\mathcal{V}_{t_{(i,j)}} = \bigcup_{o \text{ is an operation or a variable}}, (\mathcal{V}_{t_{(i,j)}} \cap \mathcal{V}_{os_o})$. Since operations cannot appear or disappear from a population during crossover, $\mathcal{V}_{os_o} \neq \emptyset \Longrightarrow \exists$ an individual in P fitting the operation schema os_o. Thus the only sets of the form $\mathcal{V}_{t_{(i,j)}} \cap \mathcal{V}_{os_o}$ which may possibly contribute to the union above are these for which $\exists$ an individual in P fitting the operation schema os_o. According to the claim above, all such sets contribute exactly the same amount. Moreover, by assumption $os_{(i,j)}(P) \neq \emptyset$, and so we have $|\mathcal{V}_{t_{(i,j)}}| = n \cdot |\mathcal{V}_{t_{(i,j)}} \cap \mathcal{V}_{os_{(i,j)}}| = n \cdot |\mathcal{V}_{t_{(i,j)} \cap os_{(i,j)}}| = n \cdot |\mathcal{V}_{t_{\uparrow(i,j)}}| \Longrightarrow \frac{|\mathcal{V}_{t_{\uparrow(i,j)}}|}{|\mathcal{V}_{t_{(i,j)}}|} = \frac{1}{n}$ where n is the number of operation schemata of the form os_o which are obtained from $cs_{(i,j)}$ by attaching a variable or an operation at the node with coordinate (i, j) and for which $\exists$ an individual in P fitting the operation schema os_o and the last implication holds under the condition that $|\mathcal{V}_{t_{(i,j)}}| \neq 0$. This condition is, indeed satisfied. (Suppose not. Let (a, b) denote the smallest coordinate such that $|\mathcal{V}_{t_{(a,b)}}| = 0$. Notice that $(a, b) \neq (1, 1)$ since $|\mathcal{V}_{t_{(1,1)}}| \neq 0$. (By assumption $\exists$ an individual, say $\mathbf{x}$, in P fitting the hyperschema $t_{(1,1)}$. Even if $\mathbf{x}$ is not the first individual of P, by performing crossover of $\mathbf{x}$ with the first individual of P at the root node one gets a population $Q \in \mathcal{V}_{t_{(1,1)}}$.) But then $(a, b) =\uparrow (i, j)$ for some coordinate (i, j) and according to the equation above we have $|\mathcal{V}_{t_{(i,j)}}| = n \cdot |\mathcal{V}_{t_{\uparrow(i,j)}}| = n \cdot |\mathcal{V}_{t_{(a,b)}}| = 0$ which contradicts the minimality of the coordinate (a, b). So we conclude that $|\mathcal{V}_{t_{(i,j)}}| \neq 0$.) Thereby we have $\frac{|\mathcal{V}_{t_{\uparrow(i,j)}}|}{|\mathcal{V}_{t_{(i,j)}}|} = \frac{1}{n}$. But $cs_{(i,j)} = \bigcup_{o \text{ is an operation or a variable}} os_o \Longrightarrow cs_{(i,j)}(P) = \bigcup_{o \text{ is an operation or a variable}} os_o(P)$. Since we assumed that all of the operations and variables are distinct, $\exists$ at most one individual in P fitting the operation schema os_o and it now follows that $|cs_{(i,j)}(P)| =$ the number of operation schemata of the form os_o such that $os_o(P) \neq \emptyset$ which is precisely the number n. We finally obtain $\frac{|\mathcal{V}_{t_{\uparrow(i,j)}}|}{|\mathcal{V}_{t_{(i,j)}}|} = \frac{1}{|cs_{(i,j)}|}$ which is precisely the conclusion of the lemma. □

Remark 66. Given an individual $\mathbf{u}$ and a population P consisting of m individuals, observe that the number of individuals fitting the hyperschema $t_{(1,1)}$ is the same in every population from $[P]_{\mathcal{A}}$, i.e. $\forall Q \in [P]_{\mathcal{A}}$ we have $|t_{(1,1)}(Q)| = |t_{(1,1)}(P)| = 1$. It follows immediately now that $\frac{1}{m} E_{\varrho_{[P]_{\mathcal{A}}}}(\mathcal{X}(\square, S_{t_{(1,1)}})) = \frac{1}{m}$.

We now combine Corollary 57, Remark 66 and Lemma 65 to obtain the following special case of Geiringer theorem for nonlinear GP under homologous crossover in case when all of the operations appearing in the individuals of the initial population P are distinct:

$$\lim_{t \to \infty} \Phi(\mathbf{u}, P, t) = \frac{1}{m} \cdot \prod_{(i,j) \text{ is not the maximal coordinate of } \mathbf{u}} \frac{1}{|cs_{(i,j)}(P)|} = \prod_{v \text{ is a node of } \mathbf{u}} \frac{1}{|cs(v)(P)|}$$

(recall that when v is the root node of $\mathbf{u}$, $cs(v)$ determines the entire search space, and so $\frac{1}{|cs(v)(P)|} = \frac{1}{m}$). To obtain the general case, suppose we are given an initial population P. For every node v of $\mathbf{u}$ consider the set of operations $\mathcal{O}(v) = \{o \,||os_o(P)| \geqslant 1$ where os_o is obtained from $cs(v)$ as in Definition 44$\}$. Moreover, for every operation (or variable) $o \in \mathcal{O}(v)$ let $\mathbf{x}^o_1, \mathbf{x}^o_2, \ldots, \mathbf{x}^o_{|os_o(P)|}$ denote an enumeration of the individuals in P fitting the operation schema $os_o(P)$. Relabel the operation o occurring in the node v of $\mathbf{x}^o_i$ by the formally different operation (o, i) (i.e. by the ordered pair (o, i) whose first element is the operation o itself and the second element is the index telling us in which individual of P the operation o labels the node v). After all of the relabelling is complete we obtain a new population P' which is special with respect to the individual $\mathbf{u}$ in the sense of Definition 64. Formally speaking, we expand our signature $\Sigma = (\Sigma_1, \Sigma_2, \ldots, \Sigma_N)$ as in Definition 26 by adding the operations (variables) (o, i) into Σ_j where j is the arity of the operation o. This gives us a new signature $\Sigma^* = (\Sigma^*_1, \Sigma^*_2, \ldots, \Sigma^*_N)$ where

$$\Sigma^*_j = \left\{ o | o \in \Sigma_j \text{ and } o \notin \bigcup_{v \text{ is a node of } \mathbf{u}} \mathcal{O}(v) \right\} \cup \{(o, i) | o \in \mathcal{O}(v) \text{ for some } v \text{ and } 1 \leqslant i \leqslant |os_o(P)|\}.$$

Denote by Ω^* the search space induced by the signature Σ^*. The natural projection maps $p_j : \Sigma^*_j \to \Sigma_j$ sending $0 \to o$ when $o \notin \bigcup_{v \text{ is a node of } \mathbf{u}} \mathcal{O}(v)$ and $(o, i) \to o$ when $o \in \mathcal{O}(v)$ for some node v of $\mathbf{u}$, induce the natural "deletion of the extra labels" projection of the search spaces $\varphi : \Omega^* \to \Omega$ where the individual $\varphi(\mathbf{w}) \in \Omega$ is obtained from the individual $\mathbf{w} \in \Omega^*$ by replacing the label of every node w of $\mathbf{w}$ with $p_j(w)$ where j is the arity of the node w. It is easily seen that the natural projection φ commutes with the crossover transformations in the sense that for any individuals

$\mathbf{x}, \mathbf{y} \in \Omega^*$ and for any crossover transformation $T \in \mathcal{G}$ (see Definition 41) we have $\varphi(T(\mathbf{x}, \mathbf{y})) = T(\varphi(\mathbf{x}), \varphi(\mathbf{y}))$.[8] Notice also that the population P can be obtained from the population P' by applying the natural projection φ to every individual of P'. Therefore, running the algorithm with the initial population P is the same thing as running the algorithm with the initial population P' and reading the output by applying the natural projection φ. The special case does apply to the population P', as mentioned above, and so we have

$$\lim_{t\to\infty} \Phi(\mathbf{u}, P, t) = \sum_{\mathbf{w}\in\varphi^{-1}(\mathbf{u})} \lim_{t\to\infty} \Phi(\mathbf{w}, P, t) = \sum_{\mathbf{w}\in\varphi^{-1}(\mathbf{u})} \prod_{v \text{ is a node of } \mathbf{w}} \frac{1}{|cs(v)(P)|}.$$

Notice that $\mathbf{w} \in \varphi^{-1}(\mathbf{u})$ precisely when the underlying shape schema of $\mathbf{w}$ is the same as that of $\mathbf{u}$, call this shape schema $t_{\mathbf{u}}$, and the label of every node v of $\mathbf{w}$ is (o, i) where o is the label of the node v of $\mathbf{u}$. According to the way the population P' was introduced, there are precisely $|os(v)(P)|$ such labels (see also Definition 44). We can then identify the preimage $\varphi^{-1}(\mathbf{u})$ with the set $\prod_{j=1}^{K}\{i \mid 1 \leqslant i \leqslant |os(v_j)|\}$ of ordered K-tuples of integers where K is the number of nodes in the parse tree of $\mathbf{u}$ and $v_1, v_2, \ldots, v_K$ is any fixed enumeration of the nodes of $\mathbf{u}$, in the following manner: The identification map $\imath : \prod_{j=1}^{K}\{i \mid 1 \leqslant i \leqslant |os(v_j)(P)|\} \to \varphi^{-1}(\mathbf{u})$ sends a given ordered K-tuple $(i_1, i_2, \ldots, i_K)$ into the tree $\mathbf{w} = \imath((i_1, i_2, \ldots, i_K))$ whose underlying shape schema is $t_{\mathbf{u}}$ and the label of a node v_j of $\mathbf{w}$ is (o_j, i_j) where o_j is the label of the node v_j in the parse tree of $\mathbf{u}$. We finally obtain:

$$\begin{aligned}
\lim_{t\to\infty} \Phi(\mathbf{u}, P, t) &= \sum_{\mathbf{w}\in\varphi^{-1}(\mathbf{u})} \prod_{v \text{ is a node of } \mathbf{w}} \frac{1}{|cs(v)(P)|} \\
&= \sum_{(i_1, i_2, \ldots, i_K)\in\prod_{j=1}^{K}\{i \mid 1 \leqslant i \leqslant |os(v_j)|\}} \prod_{v \text{ is a node of } \mathbf{u}} \frac{1}{|cs(v)(P)|} \\
&= \sum_{i_1=1}^{|os(v_1)(P)|} \sum_{i_2=1}^{|os(v_2)(P)|} \cdots \sum_{i_K=1}^{|os(v_K)(P)|} \prod_{v \text{ is a node of } \mathbf{u}} \frac{1}{|cs(v)(P)|} \\
&= \prod_{j=1}^{K} \sum_{i_j=1}^{|os(v_j)(P)|} \frac{1}{|cs(v_j)(P)|} = \prod_{v \text{ is a node of } \mathbf{u}} \frac{|os(v)(P)|}{|cs(v)(P)|}
\end{aligned}$$

which is precisely the assertion of Theorem 47.

9. What does Theorem 21 tell us in the presence of mutation for nonlinear GP?

In general, mutation is an elementary step of type 2 (see Definition 7) which is determined by the reproduction 1-tuple of the form $(\Omega, \mathcal{M}, p, \wp_m)$ where $\mathcal{M}$ is a family of functions on Ω. Notice that the set of partitions of the set of m elements into one-element subsets consists of exactly one element—the partition into the singletons. This forces $\wp_m$ to be the trivial probability distribution. We shall, therefore, omit it from writing:

Definition 67. A mutation 1-tuple is a reproduction 1-tuple $(\Omega, \mathcal{M}, p)$ where $\mathcal{M}$ consists of functions on Ω and $1_\Omega \in \mathcal{M}$. (Here $1_\Omega : \Omega \to \Omega$ denotes the identity map.)

An ergodic mutation 1-tuple is a mutation 1-tuple $(\Omega, \mathcal{M}, p)$ such that $\forall x,$ and $y \in \Omega \exists M \in \mathcal{M}$ with $M(x) = y$ and $p(M) > 0$.

The following fact is a rather simple consequence of Theorem 21 (see Corollaries 7.1 and 7.2 of [4]):

Corollary 68. *Let $\mathcal{A}$ denote a bijective and self-transient algorithm which involves an elementary step determined by an ergodic mutation. Then the Markov chain associated to the algorithm $\mathcal{A}$ is irreducible and the unique stationary distribution of this Markov chain is uniform. In particular, the limiting frequency of occurrence of any given individual* $\mathbf{x}$ *is* $\lim_{t\to\infty} \Phi(\{\mathbf{x}\}, P, t) = \frac{1}{|\Omega|}$ (*see Definition* 24 *for the meaning of* $\Phi(\{\mathbf{x}\}, P, t)$).

[8] Of course, formally speaking, the two transformations T involved in the equation above are distinct, as they have different domains (Ω^* and Ω, respectively), but they are determined by the same set of shape schemata and the same choice of nodes for crossover so we denote them by the same symbol.

When dealing with nonlinear GP, depending on the circumstances, one may want to consider different types of mutation. Below we define one such possible mutation:

Definition 69. Let Ω denote the search space for nonlinear GP over the signature $\Sigma = (\Sigma_0, \Sigma_1, \Sigma_2, \ldots, \Sigma_N)$ where Σ_i's are finite sets (see Definition 26). Consider a configuration schema t and a node v of t. Let i denote the arity of the node v and let π denote a permutation of Σ_i. We define a node mutation transformation $M_{t,v,\pi} : \Omega \to \Omega$ to be the function which sends a given program tree $\mathbf{u}$ which fits the schema t to the program $M_{t,v,\pi}(\mathbf{u})$ obtained from $\mathbf{u}$ by replacing the label $a \in \Sigma_i$ of the node v of $\mathbf{u}$ with $\pi(a)$ whenever $\mathbf{u}$ fits the configuration schema t (if $\mathbf{u}$ does not fit the schema t then $M_{t,v,\pi}(\mathbf{u}) = \mathbf{u}$). We define the family of node mutations

$$\mathcal{M}_{\text{node}} = \{M_{t,v,\pi} : \Omega \to \Omega | \pi \in \mathcal{S}_{\Sigma_i} \text{ where } t \text{ is a configuration schema and } i \text{ is the arity of the node } v \text{ of } t\}.$$

As usual $\mathcal{S}_X$ denotes the set of all permutations of the set X. Denote by $\Omega_{\text{NodeMut}} = (\Omega, \mathcal{M}_{\text{node}}, p)$ the corresponding mutation 1-tuple.

Although node mutation described in Definition 69 is not ergodic in the sense of Definition 67, it defines a bijective elementary step (see Definition 15). Indeed, it is easy to see that the transformation $M_{t,v,\pi^{-1}}$ is a 2-sided inverse of the transformation $M_{t,v,\pi}$. Thereby Theorem 21 applies to nonlinear GP with homologous crossover and node mutation. It is also possible to derive a formula for the limiting frequency of occurrence of a given individual $\mathbf{u}$, namely $\lim_{t\to\infty} \Phi(\mathbf{u}, P, t)$ much in the same way as in Theorem 47. In order to state the corresponding result for nonlinear GP with homologous crossover and node mutation, it is convenient to introduce the following definitions first:

Definition 70. Fix an individual $\mathbf{u} \in \Omega$. Let v denote a node of $\mathbf{u}$ and consider the configuration schema $cs(\uparrow v, i)$ obtained from the configuration schema $cs(v)$ by attaching a node of degree i together with it's i children in place of the # sign at the node corresponding to v of $cs(v)$. The newly attached nodes are then labelled by the = and # signs, respectively. If the newly attached node is of arity 0 then it is a leaf node labelled by the = sign.[9] Furthermore, write $cs(\uparrow v, \mathbf{u})$ in place of $cs(\uparrow v, i)$ when i is the arity of the node v in $\mathbf{u}$. Also denote by $os(v, o)$ the operation schema obtained from the configuration schema $cs(v)$ by attaching a node labelled by the operation o together with its appropriate number of children in place of the # sign. The children of the newly attached node (if there are any) are labelled by the # signs.

Definition 71. Given a mutation 1-tuple $(\Omega, \mathcal{M}, p)$, a configuration schema t (see Definition 44), and a node v of t having arity i, denote by $G(t, v)$ the group generated by all the permutations $\pi \in \Sigma_i$ such that $p(M_{s,v,\pi}) > 0$ for some configuration schema s such that the common region of s and t contains the node v. Fix an operation $a \in \Sigma_i$. Let

$$\mathcal{O}(t, v, a) = \{o \in \Sigma_i | \exists g \in G(t, v) \text{ with } g \cdot a = o\}$$

denote the orbit of the operation a under the action of the group $G(t, v)$.

Suppose we are given a population P of size m consisting of program trees from Ω. Recall from Definition 46 that we denote by $|H(P)|$ the number of individuals in the population P fitting the schema H.

Theorem 72. *Let $\mathcal{A}$ denote an algorithm determined by 2 elementary steps of type 2 one of which is determined by the node mutation (see Definition 69) and the other one by a homologous GP crossover. Suppose every one of the transformations in the family $\mathcal{G}$ of GP homologous crossovers has a positive probability of being chosen.[10] Fix an individual (a program tree) $\mathbf{u} \in \Omega$ and an initial population P. Let $o(\mathbf{u}, v)$ denote the operation labelling the node v of the program tree $\mathbf{u}$. Denote by $\hat{\mathbf{u}}$ the configuration schema obtained from the shape schema, $\tilde{\mathbf{u}}$, of $\mathbf{u}$ (see Definition 34) by labelling all the interior nodes of $\mathbf{u}$ with the = signs and all the leaf nodes with the # signs. Suppose that the probability distribution on the collection of node mutations is such that whenever v is a node of $\mathbf{u}$ we*

[9] Formally speaking, according to Definition 44, $cs(\uparrow v, i)$ is not always a Poli hyperschema since it may contain a leaf node labelled by the = sign. However, such a schema also defines a subset of the search space Ω in much the same way as Poli hyperschemata. The only difference is that a leaf node labelled by the = sign can be replaced by a variable only. One cannot attach a non-trivial program tree to it.

[10] Again we remark that this condition can be slightly relaxed but it does not introduce any new ideas of interest.

have $p(M_{s,v,\pi}) > 0 \Longrightarrow s = cs(v)$ where as before $cs(v)$ is the configuration schema of $\mathbf{u}$ *corresponding to the node* v.[11] *Then we have*

$$\lim_{t\to\infty} \Phi(\mathbf{u}, P, t) = \prod_{v \text{ is a node of } \mathbf{u}} \frac{\sum_{o\in\mathcal{O}(\hat{\mathbf{u}},v,o(\mathbf{u},v))} |os(v,o)(P)|}{\sum_{i=0}^{N}\sum_{o\in\Sigma_i} |os(v,o)(P)| \cdot |\mathcal{O}(\hat{\mathbf{u}},v,o)|}.$$

Proof. The proof of Theorem 72 is very similar to the proof of Theorem 47 and contains essentially no new ideas. We leave most of the details for the interested reader as an exercise and provide only the rough outline: Just like Theorem 47, Theorem 75 follows from Corollary 57 by considering the nested sequence of hyperschemata $t_{(1,1)} \geqslant t_{(2,1)} \geqslant \cdots \geqslant t_{(i,j)} \geqslant t_{\uparrow(i,j)} \geqslant \cdots \geqslant t_{(l,k)} = \mathbf{u}$ corresponding to the program $\mathbf{u}$ (see Definition 60). First, we consider a special case when every set Σ_i consists of ordered pairs (l, o) where l is an integer, and mutation is allowed to modify only the operation o and is not allowed to change the integer l. We then prove Theorem 72 in the special case when all the labels contained in the initial population P have distinct first coordinates. The general case then follows by introducing the extra integer labels for the first coordinate, applying the special case and then “erasing the integer part from the labels” in exactly the same way as it was done in the proof of Theorem 47. The main difference lies in the claim proved inside Lemma 65. The corresponding claim for the proof of Theorem 72 then says the following:

Lemma 73. *Fix a node v with coordinate (i, j). Fix any two operation schemata $os(a)$ and $os(b)$ which are obtained from $cs_{(i,j)}$ by attaching either a variable or an operation at the node with coordinate (i, j). Suppose $\exists$ individuals in P fitting both, $os(c)$ and $os(d)$ where $a \in \mathcal{O}(cs(i,j), v, c)$ and $b \in \mathcal{O}(cs(i,j), v, d)$. Then $|\mathcal{V}_{t_{(i,j)}} \cap \mathcal{V}_{os(a)}| = |\mathcal{V}_{t_{(i,j)}} \cap \mathcal{V}_{os(b)}|$.*

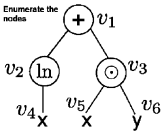
Just like the claim inside of the Lemma 65, Lemma 73 is proved by constructing an explicit bijection between the sets $\mathcal{V}_{t_{(i,j)}} \cap \mathcal{V}_{os(a)}$ and $\mathcal{V}_{t_{(i,j)}} \cap \mathcal{V}_{os(b)}$, The only difference is that these bijections make use of mutations as well as crossover. □

It may be worth mentioning that Theorem 47 is a special case of Theorem 72. Indeed, if the only mutation transformations chosen with positive probability are these which assign positive probability only to the mutations defined by the identity permutations, then every orbit $\mathcal{O}(\hat{\mathbf{u}}, v, o)$ consists of exactly one element so that $\forall t$ and v we have $|\mathcal{O}(\hat{\mathbf{u}}, v, o)| = 1$. To compress the language, we shall use $\biguplus$ to denote the union of *disjoint* sets. We then have $\biguplus_{o\in\mathcal{O}(\hat{\mathbf{u}},v,o(\mathbf{u},v))} os(v,o)(P) = os(v)(P)$ since $o(\mathbf{u}, v)$ is the only operation inside of $\mathcal{O}(\hat{\mathbf{u}}, v, o(\mathbf{u}, v))$ so that $os(v)(P)$ is the only contributor to the disjoint union above. Moreover, we also have $cs(v) = \biguplus_{i=0}^{N} \biguplus_{o\in\Sigma_i} os(v, o)$ so that we obtain

$$\sum_{i=0}^{N} \sum_{o\in\Sigma_i} |os(v,o)(P)| \cdot |\mathcal{O}(\hat{\mathbf{u}}, v, o)| = \sum_{i=0}^{N} \sum_{o\in\Sigma_i} |os(v,o)(P)| \cdot 1 = |cs(v)(P)|.$$

The formula in Theorem 72 now simplifies to the one in Theorem 47.

Example 74. Continuing with Example 48, suppose the signature Σ is defined as follows: $\Sigma = (\Sigma_0, \Sigma_1, \Sigma_2)$ where $\Sigma_0 = \{x, y, z, w\}$, $\Sigma_1 = \{\sin, \cos, \tan, \cot, \ln\}$ and $\Sigma_2 = \{+, -, *, \odot\}$ where $\odot$ is a binary operation symbol different from $+$, $*$ and $-$. (Of course, the semantics of the binary operation $\odot$ is irrelevant to the content of this example, but if the reader feels more comfortable with a concrete interpretation, they may assume, for instance, that $x \odot y = \int_x^y e^{\xi^2}\, d\xi$.) Now suppose the individual $\mathbf{u}$ is the program $(+(\ln(x), \odot(x, y))$ pictured below (with nodes being enumerated just like in Example 45) (it has the same shape schema as the individual in that example):



[11] Notice that this implies that $\mathcal{O}(\hat{\mathbf{u}}, v, o(\mathbf{u}, v)) = \mathcal{O}(cs(v), v, o(\mathbf{u}, v))$.

Notice that this individual has exactly the same set of configuration schemata as the corresponding set of configuration schemata in Example 45 (the reader may see these configuration schemata pictured in that example). Suppose that the following mutation transformations are the only ones chosen with positive probability:

$$M_{cs(v_1),v_1,(+,-)},\quad M_{cs(v_1),v_1,(*,\odot)},\quad M_{cs(v_2),v_2,(\ln,\sin,\cos)(\tan,\cot)},$$

$$M_{cs(v_2),v_2,(\tan,\cot)(\sin,\cos)},\quad M_{cs(v_2),v_2,(\ln,\sin)},\quad M_{cs(v_3),v_3,(+,\odot,*)},$$

$$M_{cs(v_3),v_3,(-,*)},\quad M_{cs(v_4),v_4,(x,w)(y,z)},\quad M_{cs(v_4),v_4,(+,*)(-,\odot)},\quad M_{cs(v_5),v_5,(x,y,z)},$$

$$M_{cs(v_5),v_5,(+,-,*,\odot)},\quad M_{cs(v_6),v_6,(x,w)(y,z)}\ \text{and}\ M_{cs(v_6),v_6,(x,w)},\quad M_{cs(v_6),v_6,(\sin,\cos,\tan,\cot)},\quad M_{cs(v_6),v_6,(+,-)}.$$

Here we represent permutations in terms of their “disjoint cycle decompositions”: for example, (ln, sin, cos)(tan, cot) represents the permutation on Σ_1 which sends ln into sin, sin into cos and cos back into ln. Likewise, it sends tan into cot and cot back into tan. If a cycle has length one (i.e. the element appearing in the cycle is a fixed point of the corresponding permutation) we omit that cycle from writing. For example, $\odot$ and $*$ are the fixed points of the permutation $(+, -)$. The corresponding permutation groups are:

$$G(cs(v_1), v_1) = \langle(+, -), (*, \odot)\rangle,$$

$$G(cs(v_2), v_2) = \langle(\ln, \sin, \cos)(\tan, \cot), (\tan, \cot)(\sin, \cos), (\ln, \sin)\rangle,$$

$$G(cs(v_3), v_3) = \langle(+, \odot, *), (-, *)\rangle,$$

$$G(cs(v_4), v_4) = \langle(x, w)(y, z), (+, *)(-, \odot)\rangle,$$

$$G(cs(v_5), v_5) = \langle(x, y, z), (+, -, *, \odot)\rangle \quad \text{and}$$

$$G(cs(v_6), v_6) = \langle(x, w)(y, z), (x, w)(\sin, \cos, \tan, \cot), (+, -)\rangle.$$

The cycle decomposition makes it easy to compute the corresponding orbits:

$$\mathcal{O}(cs(v_1), v_1, +) = \{+, -\}, \mathcal{O}(cs(v_2), v_2, \ln) = \{\ln, \sin, \cos\},$$

$$\mathcal{O}(cs(v_3), v_3, \odot) = \{+, -, *, \odot\}, \mathcal{O}(cs(v_4), v_4, x) = \{x, w\},$$

$$\mathcal{O}(cs(v_5), v_5, x) = \{x, y, z\} \text{ and } \mathcal{O}(cs(v_6), v_6, y) = \{y, z\}.$$

Now suppose the initial population is the same as in Example 48. In order to apply Theorem 72, for every node v of $\mathbf{u}$ we need to compute the number $|os(v, o)(P)|$. Recall that the schema $os(v, o)$ is obtained from the schema $os(v)$ by attaching the operation o at the node v and labelling its children nodes by the # signs. For the population P in Example 48 it was already computed that $|os(v_1, +)| = |os(v_1)| = 2$. There are exactly two individuals (namely $\mathbf{x_3}$ and $\mathbf{x_4}$) fitting the schema $os(v_1, *)$ and so $|os(v_1, *)| = 2$. Exactly one individual, namely $\mathbf{x}_2$, and one individual, namely $\mathbf{x_6}$, fit the schemata $os(v_1, \sin)$ and $os(v_2, \cos)$, respectively, and so $|os(v_1, \sin)| = |os(v_1, \cos)| = 1$. For all the other operations $o \in (\Sigma_0 \cup \Sigma_1 \cup \Sigma_2) - \{+, *, \sin, \cos\}$ there are no individuals in P fitting the schema $os(v_1, o)$ and so we have $|os(v_1, o)| = 0$. There are no individuals in P fitting the schema $os(v_2, \ln) = os(v_2)$ for the individual $\mathbf{u}$ of the current example, and no individuals fitting the schemata of the form $os(v_2, o)$ where $o \notin \{*, \sin, \cos\}$ so that for such o we have $|os(v_2, o)(P)| = 0$. Moreover, there is exactly one individual, namely $\mathbf{x}_1$ fitting the schema $os(v_2, *)$ and exactly one, namely $\mathbf{x}_4$ fitting the schema $os(v_2, \cos)$ so that $|os(v_2, *)(P)| = |os(v_2, \cos)(P)| = 1$; exactly two individuals, namely $\mathbf{x}_3$ and $\mathbf{x}_5$ fit the schema $os(v_2, \sin)$ so that $|os(v_2, \sin)(P)| = 2$. Continuing in this manner with the rest of the nodes of $\mathbf{u}$ we obtain $|os(v_3, o)(P)| = 0$ for $o \notin \{+, *\}$; $\mathbf{x}_1$ and $\mathbf{x}_3$ fit $os(v_3, +)$ while $\mathbf{x}_4$ and $\mathbf{x}_5$ fit $os(v_3, +)$ and so $|os(v_3, +)(P)| = |os(v_3, *)(P)| = 2$. $|os(v_4, o)(P)| = 0$ for $o \notin \{x, y, +\}$; $\mathbf{x_3}$ is the only individual fitting the schema $os(v_4, y)$, $\mathbf{x_4}$ is the only individual fitting the schema $os(v_4, x)$ and $\mathbf{x_5}$ is the only individual fitting the schema $os(v_4, +)$ and so we have $|os(v_4, x)(P)| = |os(v_4, y)(P)| = |os(v_4, +)(P)| = 1$. $|os(v_5, o)(P)| = 0$ for $o \notin \{*, x, y\}$; $\mathbf{x_1}$ is the only individual fitting the schema $os(v_5, *)$ and $\mathbf{x_5}$ is the only individual fitting the schema $os(v_5, y)$ while the individuals $\mathbf{x}_3$ and $\mathbf{x}_4$ are the only two which fit the schema $os(v_5, x)$ so that we have $|os(v_5, *)(P)| = |os(v_5, y)(P)| = 1$ and $|os(v_5, x)(P)| = 2$. $|os(v_6, o)(P)| = 0$ for $o \notin \{\sin, x, y\}$; moreover, $\mathbf{x_1}$ is the only individual fitting the schema

$os(v_6, \sin)$ and $\mathbf{x_4}$ is the only individual fitting the schema $os(v_6, y)$ while the individuals $\mathbf{x}_3$ and $\mathbf{x}_5$ are the only two which fit the schema $os(v_6, x)$ so that we have $|os(v_6, \sin)(P)| = |os(v_6, y)(P)| = 1$ and $|os(v_6, x)(P)| = 2$.

Finally for every node v and for every operation $o \in \Sigma_0 \cup \Sigma_1 \cup \Sigma_2$ such that $|os(v, o)(P)| \neq 0$ we need to compute $|\mathcal{O}(\hat{\mathbf{u}}, v, o)|$. For the node v_1 these are $\mathcal{O}(\hat{\mathbf{u}}, v_1, +) = \mathcal{O}(cs(v_1), v_1, +) = \{+, -\}$ so that $|\mathcal{O}(cs(v_1), v_1, +)| = 2$, and, likewise, from the description of the groups $G(cs(v_i), v_i)$ given above, it is easy to compute that $\mathcal{O}(cs(v_1), v_1, *) = \{*, \odot\}$, $\mathcal{O}(cs(v_1), v_1, \sin) = \{\sin\}$ and $\mathcal{O}(cs(v_1), v_1, \cos) = \{\cos\}$ so that

$$|\mathcal{O}(cs(v_1), v_1, \sin)| = |\mathcal{O}(cs(v_1), v_1, \cos)| = 1.$$

$\mathcal{O}(cs(v_2), v_2, *) = \{*\}$, $\mathcal{O}(cs(v_2), v_2, \sin) = \mathcal{O}(cs(v_2), v_2, \cos) = \{\ln, \sin, \cos\}$ and so

$$|\mathcal{O}(cs(v_2), v_2, \sin)| = |\mathcal{O}(cs(v_2), v_2, \cos)| = 3.$$

$\mathcal{O}(cs(v_3), v_3, *) = \mathcal{O}(cs(v_3), v_3, +) = \Sigma_2$ so that $|\mathcal{O}(cs(v_3), v_3, *)| = |\mathcal{O}(cs(v_3), v_3, +)| = 4$.
$\mathcal{O}(cs(v_4), v_4, x) = \{x, w\}$ and $\mathcal{O}(cs(v_4), v_4, y) = \{y, z\}$ so that $|\mathcal{O}(cs(v_4), v_4, x)| = |\mathcal{O}(cs(v_4), v_4, y)| = 2$.
$\mathcal{O}(cs(v_4), v_4, +) = \{+, *\}$ so that $|\mathcal{O}(cs(v_4), v_4, +)| = 2$; $\mathcal{O}(cs(v_5), v_5, *) = \Sigma_2$, $\mathcal{O}(cs(v_5), v_5, x) = \mathcal{O}(cs(v_5), v_5, y) = \{x, y, z\}$ and so $|\mathcal{O}(cs(v_5), v_5, x)| = |\mathcal{O}(cs(v_5), v_5, y)| = 3$.
$\mathcal{O}(cs(v_6), v_6, \sin) = \{\sin, \cos, \tan, \cot\}$, $\mathcal{O}(cs(v_6), v_6, x) = \{x, w\}$ and, finally, $\mathcal{O}(cs(v_6), v_6, y) = \{y, z\}$ so that

$$|\mathcal{O}(cs(v_6), v_6, x)| = |\mathcal{O}(cs(v_6), v_6, y)| = 2.$$

Now we are ready to compute the ratios of the form $\frac{\sum_{o\in\mathcal{O}(\hat{\mathbf{u}},v_j,o(\mathbf{u},v_j))}|os(v_j,o)(P)|}{\sum_{i=0}^{N}\sum_{o\in\Sigma_i}|os(v_j,o)(P)|\cdot|\mathcal{O}(\hat{\mathbf{u}},v_j,o)|}$. From the data computed above we have

$$\sum_{o\in\mathcal{O}(\hat{\mathbf{u}},v_1,o(\mathbf{u},v_1))} |os(v_1, o)(P)| = |os(v_1, +)(P)| + |os(v_1, -)(P)| = 2 + 0 = 2$$

($\mathbf{x}_1$ and $\mathbf{x}_5$ are the only two individuals in P which fit the schema $os(v_1, +)$ while no individual in P fits $os(v_1, -)$);

$$\begin{aligned}\sum_{i=0}^{N}\sum_{o\in\Sigma_i} |os(v_1, o)(P)| \cdot |\mathcal{O}(\hat{\mathbf{u}}, v_1, o)| &= |os(v_1, +)(P)| \cdot |\mathcal{O}(\hat{\mathbf{u}}, v_1, +)(P)| + |os(v_1, \sin)(P)| \cdot |\mathcal{O}(\hat{\mathbf{u}}, v_1, \sin)(P)| \\ &\quad + |os(v_1, *)(P)| \cdot |\mathcal{O}(\hat{\mathbf{u}}, v_1, *)(P)| + |os(v_1, \cos)(P)| \cdot |\mathcal{O}(\hat{\mathbf{u}}, v_1, \cos)(P)| \\ &= 2 \cdot 2 + 1 \cdot 1 + 2 \cdot 2 + 1 \cdot 1 = 10\end{aligned}$$

and so

$$\frac{\sum_{o\in\mathcal{O}(\hat{\mathbf{u}},v_1,o(\mathbf{u},v_1))}|os(v_1, o)(P)|}{\sum_{i=0}^{N}\sum_{o\in\Sigma_i}|os(v_1, o)(P)| \cdot |\mathcal{O}(\hat{\mathbf{u}}, v_1, o)|} = \frac{2}{10} = \frac{1}{5}.$$

Continuing in this manner, we obtain

$$\sum_{o\in\mathcal{O}(\hat{\mathbf{u}},v_2,o(\mathbf{u},v_2))} |os(v_2, o)(P)| = |os(v_2, \ln)(P)| + |os(v_2, \sin)(P)| + |os(v_2, \cos)(P)| = 0 + 2 + 1 = 3$$

and

$$\begin{aligned}\sum_{i=0}^{N}\sum_{o\in\Sigma_i} |os(v_2, o)(P)| \cdot |\mathcal{O}(\hat{\mathbf{u}}, v_2, o)| &= |os(v_2, *)(P)|\cdot|\mathcal{O}(\hat{\mathbf{u}}, v_2, *)(P)| + |os(v_2, \sin)(P)|\cdot|\mathcal{O}(\hat{\mathbf{u}}, v_2, \sin)(P)| \\ &\quad + |os(v_2, \cos)(P)| \cdot |\mathcal{O}(\hat{\mathbf{u}}, v_2, \cos)(P)| = 1 \cdot 1 + 2 \cdot 3 + 1 \cdot 3 = 10\end{aligned}$$

so that

$$\frac{\sum_{o\in\mathcal{O}(\hat{\mathbf{u}},v_2,o(\mathbf{u},v_2))}|os(v_2, o)(P)|}{\sum_{i=0}^{N}\sum_{o\in\Sigma_i}|os(v_2, o)(P)| \cdot |\mathcal{O}(\hat{\mathbf{u}}, v_2, o)|} = \frac{3}{10},$$

$$\begin{aligned}\sum_{o\in\mathcal{O}(\hat{\mathbf{u}},v_3,o(\mathbf{u},v_3))} |os(v_3, o)(P)| &= |os(v_3, \odot)(P)| + |os(v_3, +)(P)| + |os(v_3, -)(P)| + |os(v_3, *)(P)| \\ &= 0 + 2 + 0 + 2 = 4\end{aligned}$$

and

$$\sum_{i=0}^{N}\sum_{o\in\Sigma_i}|os(v_3,o)(P)|\cdot|\mathcal{O}(\hat{\mathbf{u}},v_3,o)| = |os(v_3,+)(P)|\cdot|\mathcal{O}(\hat{\mathbf{u}},v_3,+)(P)| + |os(v_3,*)(P)|\cdot|\mathcal{O}(\hat{\mathbf{u}},v_3,*)(P)|$$
$$= 2\cdot 4 + 2\cdot 4 = 16$$

and so

$$\frac{\sum_{o\in\mathcal{O}(\hat{\mathbf{u}},v_3,o(\mathbf{u},v_3))}|os(v_3,o)(P)|}{\sum_{i=0}^{N}\sum_{o\in\Sigma_i}|os(v_3,o)(P)|\cdot|\mathcal{O}(\hat{\mathbf{u}},v_3,o)|} = \frac{4}{16} = \frac{1}{4},$$

$$\sum_{o\in\mathcal{O}(\hat{\mathbf{u}},v_4,o(\mathbf{u},v_4))}|os(v_4,o)(P)| = |os(v_4,x)(P)| + |os(v_4,w)(P)| = 1+0 = 1$$

and

$$\sum_{i=0}^{N}\sum_{o\in\Sigma_i}|os(v_4,o)(P)|\cdot|\mathcal{O}(\hat{\mathbf{u}},v_4,o)| = |os(v_4,y)(P)|\cdot|\mathcal{O}(\hat{\mathbf{u}},v_4,y)| + |os(v_4,x)(P)|\cdot|\mathcal{O}(\hat{\mathbf{u}},v_4,x)|$$
$$+ |os(v_4,+)(P)|\cdot|\mathcal{O}(\hat{\mathbf{u}},v_4,+)| = 1\cdot 2 + 1\cdot 2 + 1\cdot 2 = 6$$

and so

$$\frac{\sum_{o\in\mathcal{O}(\hat{\mathbf{u}},v_4,o(\mathbf{u},v_4))}|os(v_4,o)(P)|}{\sum_{i=0}^{N}\sum_{o\in\Sigma_i}|os(v_4,o)(P)|\cdot|\mathcal{O}(\hat{\mathbf{u}},v_4,o)|} = \frac{1}{6},$$

$$\sum_{o\in\mathcal{O}(\hat{\mathbf{u}},v_5,o(\mathbf{u},v_5))}|os(v_5,o)(P)| = |os(v_5,x)(P)| + |os(v_5,y)(P)| + |os(v_5,z)(P)| = 2+1+0 = 3$$

and

$$\sum_{i=0}^{N}\sum_{o\in\Sigma_i}|os(v_5,o)(P)|\cdot|\mathcal{O}(\hat{\mathbf{u}},v_5,o)| = |os(v_5,*)(P)|\cdot|\mathcal{O}(\hat{\mathbf{u}},v_5,*)| + |os(v_5,x)(P)|\cdot|\mathcal{O}(\hat{\mathbf{u}},v_5,x)|$$
$$+ |os(v_5,y)(P)|\cdot|\mathcal{O}(\hat{\mathbf{u}},v_5,y)| = 1\cdot 4 + 2\cdot 3 + 1\cdot 3 = 13$$

and so

$$\frac{\sum_{o\in\mathcal{O}(\hat{\mathbf{u}},v_5,o(\mathbf{u},v_5))}|os(v_5,o)(P)|}{\sum_{i=0}^{N}\sum_{o\in\Sigma_i}|os(v_5,o)(P)|\cdot|\mathcal{O}(\hat{\mathbf{u}},v_5,o)|} = \frac{3}{13}.$$

Finally,

$$\sum_{o\in\mathcal{O}(\hat{\mathbf{u}},v_6,o(\mathbf{u},v_6))}|os(v_6,o)(P)| = |os(v_6,y)(P)| + |os(v_6,z)(P)| = 1+0 = 1$$

and

$$\sum_{i=0}^{N}\sum_{o\in\Sigma_i}|os(v_6,o)(P)|\cdot|\mathcal{O}(\hat{\mathbf{u}},v_6,o)| = |os(v_6,\sin)(P)|\cdot|\mathcal{O}(\hat{\mathbf{u}},v_6,\sin)| + |os(v_6,x)(P)|\cdot|\mathcal{O}(\hat{\mathbf{u}},v_6,x)|$$
$$+ |os(v_6,y)(P)|\cdot|\mathcal{O}(\hat{\mathbf{u}},v_6,y)| = 1\cdot 4 + 2\cdot 2 + 1\cdot 2 = 10$$

so that

$$\frac{\sum_{o\in\mathcal{O}(\hat{\mathbf{u}},v_6,o(\mathbf{u},v_6))}|os(v_6,o)(P)|}{\sum_{i=0}^{N}\sum_{o\in\Sigma_i}|os(v_6,o)(P)|\cdot|\mathcal{O}(\hat{\mathbf{u}},v_6,o)|} = \frac{1}{10}.$$

Now we finally compute the product of these ratios and obtain:

$$\lim_{t\to\infty}\Phi(\mathbf{u},P,t) = \prod_{i=1}^{6}\frac{\sum_{o\in\mathcal{O}(\hat{\mathbf{u}},v_i,o(\mathbf{u},v_i))}|os(v_i,o)(P)|}{\sum_{i=0}^{N}\sum_{o\in\Sigma_i}|os(v_i,o)(P)|\cdot|\mathcal{O}(\hat{\mathbf{u}},v_i,o)|} = \frac{1}{5}\cdot\frac{3}{10}\cdot\frac{1}{4}\cdot\frac{1}{6}\cdot\frac{3}{13}\cdot\frac{1}{10} = \frac{3}{52000}.$$

At the opposite extreme is the case when every mutation transformation in the family $\mathcal{M}_{\text{node}}$ has a positive probability of being chosen. In this case $\mathcal{O}(\hat{\mathbf{u}}, v, o(\mathbf{u}, v)) = \Sigma_i$ where i is the arity of the operation o. In particular, for every operation o we have $o \in \mathcal{O}(\hat{\mathbf{u}}, v, o(\mathbf{u}, v))$ if and only if $o \in \Sigma_i$. But then we have

$$\biguplus_{o \in \mathcal{O}(\hat{\mathbf{u}}, v, o(\mathbf{u}, v))} os(v, o) = cs(\uparrow v, \mathbf{u})$$

and so

$$\sum_{o \in \mathcal{O}(\hat{\mathbf{u}}, v, o(\mathbf{u}, v))} |os(v, o)(P)| = |cs(\uparrow v, \mathbf{u})(P)|.$$

We also have $\biguplus_{o \in \Sigma_i} os(v, o) = cs(\uparrow v, i)$ so that

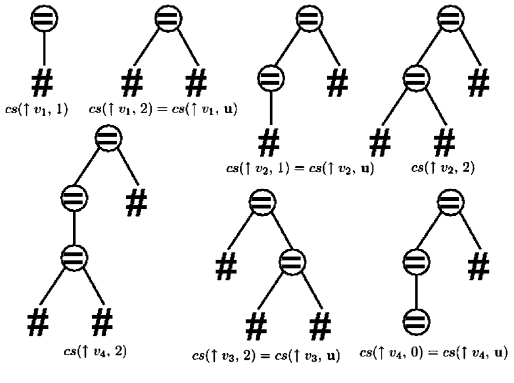
$$\sum_{o \in \Sigma_i} |os(v, o)(P)| \cdot |\mathcal{O}(\hat{\mathbf{u}}, v, o)| = |\Sigma_i| \cdot \sum_{o \in \Sigma_i} |os(v, o)(P)| = |cs(\uparrow v, i)(P)| \cdot |\Sigma_i|.$$

Combining these equations with Theorem 72 we obtain:

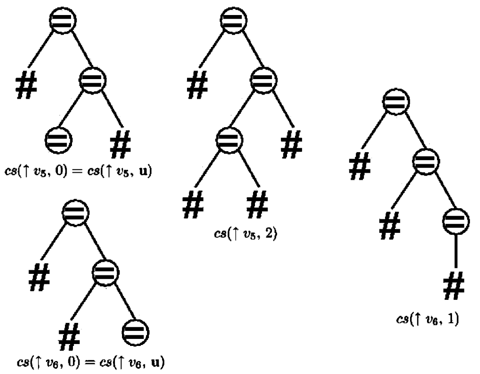
Corollary 75. *Let $\mathcal{A}$ denote an algorithm determined by* 2 *elementary steps of type* 2 *one of which is determined by the node mutation* (*see Definition* 69) *and the other one by a homologous GP crossover. Suppose every one of the transformations in the family $\mathcal{G}$ of GP homologous crossovers has a positive probability of being chosen. Suppose also that for every node v of* **u** *of arity i and for every permutation π of Σ_i we have $p(M_{\hat{\mathbf{u}}, v, \pi}) = p(M_{cs(v), v, \pi}) > 0$. Fix an individual* (*a program tree*) $\mathbf{u} \in \Omega$ *and an initial population P. Then we have*

$$\lim_{t \to \infty} \Phi(\mathbf{u}, P, t) = \prod_{v \text{ is a node of } \mathbf{u}} \frac{|cs(\uparrow v, \mathbf{u})(P)|}{\sum_{i=0}^{N} |cs(\uparrow v, i)(P)| \cdot |\Sigma_i|}.$$

Example 76. Suppose the signature $\Sigma = (\Sigma_0, \Sigma_1, \Sigma_2)$, the initial population P and the individual **u** are exactly as in Example 74. Now suppose, (unlike in Example 74, that for every permutation π of Σ_i we have $p(M_{\hat{\mathbf{u}}, v, \pi}) > 0$. Now Corollary 75 applies and we can compute the frequency of occurrence of the individual **u** according to the formula given there. To apply this formula we need to compute the numbers $|cs(\uparrow v_j, i)(P)|$ where $1 \leqslant j \leqslant 6$ and $0 \leqslant i \leqslant 2$ (the numbers $|cs(\uparrow v, \mathbf{u})(P)|$ are among these). The configuration schemata $cs(v_i)$ for the individual **u** are exactly the same as these for the individual of Example 45 (since these two individuals have the same underlying shape schema) and they are pictured in that example. Below we display only these schemata $|cs(\uparrow v_j, i)(P)|$ for which $|cs(\uparrow v_j, i)(P)| \neq 0$. According to Definition 70 they are obtained from the corresponding schemata $cs(v_i)$ by attaching a node which has i children (if $i = 0$ it has no children) in place of the # sign at the node v_i (which means that the # sign can be replaced with an arbitrary variable but not with an operation symbol):



#
$cs(\uparrow v_1, 1)$
# #
$cs(\uparrow v_1, 2) = cs(\uparrow v_1, \mathbf{u})$
# #
$cs(\uparrow v_2, 1) = cs(\uparrow v_2, \mathbf{u})$
# # #
$cs(\uparrow v_2, 2)$
# # #
$cs(\uparrow v_4, 2)$
# # #
$cs(\uparrow v_3, 2) = cs(\uparrow v_3, \mathbf{u})$
#
$cs(\uparrow v_4, 0) = cs(\uparrow v_4, \mathbf{u})$



There are exactly two individuals, namely $\mathbf{x}_2$ and $\mathbf{x}_6$ in P fitting the schema $cs(\uparrow v_1, 1)$ so that $|cs(\uparrow v_1, 1)(P)| = 2$; $\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_4$ and $\mathbf{x}_5$ are the only individuals in P which fit the schema $cs(\uparrow v_1, 2) = cs(\uparrow v_1, \mathbf{u})$ so that $|cs(\uparrow v_1, 2)(P)| = |cs(\uparrow v_1, \mathbf{u})(P)| = 4$; $\mathbf{x}_3$, $\mathbf{x}_4$ and $\mathbf{x}_5$ are the only individuals in P which fit the schema $cs(\uparrow v_2, 1) = cs(\uparrow v_2, \mathbf{u})$ so that $|cs(\uparrow v_2, 1)(P)| = |cs(\uparrow v_2, \mathbf{u})(P)| = 3$; $\mathbf{x}_1$ is the only individual fitting the schema $cs(\uparrow v_2, 2)$ so that $|cs(\uparrow v_2, 2)(P)| = 1$; $\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_4$ and $\mathbf{x}_5$ are the only individuals in P which fit the schema $cs(\uparrow v_3, 2) = cs(\uparrow v_3, \mathbf{u})$ so that $|cs(\uparrow v_3, 2)(P)| = |cs(\uparrow v_3, \mathbf{u})(P)| = 4$; $\mathbf{x}_3$ and $\mathbf{x}_4$ are the only individuals in P which fit the schema $cs(\uparrow v_4, 0) = cs(\uparrow v_4, \mathbf{u})$ so that $|cs(\uparrow v_4, 0)(P)| = |cs(\uparrow v_4, \mathbf{u})(P)| = 2$; $\mathbf{x}_5$ is the only individual fitting the schema $cs(\uparrow v_4, 2)$ and so $|cs(\uparrow v_4, 2)(P)| = 1$; $\mathbf{x}_3$, $\mathbf{x}_4$ and $\mathbf{x}_5$ are the only individuals in P fitting the schemata $cs(\uparrow v_5, 0) = cs(\uparrow v_5, \mathbf{u})$ and/or $cs(\uparrow v_6, 0) = cs(\uparrow v_6, \mathbf{u})$ and so we have $|cs(\uparrow v_5, 0)| = |cs(\uparrow v_5, \mathbf{u})| = |cs(\uparrow v_6, 0)| = |cs(\uparrow v_6, \mathbf{u})| = 3$ and $\mathbf{x}_1$ is the only individual fitting the either and both schemata $cs(\uparrow v_5, 2)$ and/or $cs(\uparrow v_6, 1)$ so that $|cs(\uparrow v_5, 2)| = |cs(\uparrow v_6, 1)| = 1$. From the definition of the signature $\Sigma = (\Sigma_0, \Sigma_1, \Sigma_2)$ in Example 74 we see that $|\Sigma_0| = 4$, $|\Sigma_1| = 5$ and $|\Sigma_2| = 4$. We are now ready to compute the ratios:

$$\frac{|cs(\uparrow v_1, \mathbf{u})(P)|}{\sum_{i=0}^{3}|cs(\uparrow v_1, i)(P)| \cdot |\Sigma_i|} = \frac{|cs(\uparrow v_1, \mathbf{u})(P)|}{|cs(\uparrow v_1, 2)(P)| \cdot |\Sigma_2| + |cs(\uparrow v_1, 1)(P)| \cdot |\Sigma_1|} = \frac{4}{4 \cdot 4 + 2 \cdot 5} = \frac{2}{13},$$

$$\frac{|cs(\uparrow v_2, \mathbf{u})(P)|}{\sum_{i=0}^{3}|cs(\uparrow v_2, i)(P)| \cdot |\Sigma_i|} = \frac{|cs(\uparrow v_2, \mathbf{u})(P)|}{|cs(\uparrow v_2, 2)(P)| \cdot |\Sigma_2| + |cs(\uparrow v_2, 1)(P)| \cdot |\Sigma_1|} = \frac{3}{1 \cdot 4 + 3 \cdot 5} = \frac{3}{19},$$

$$\frac{|cs(\uparrow v_3, \mathbf{u})(P)|}{\sum_{i=0}^{3}|cs(\uparrow v_3, i)(P)| \cdot |\Sigma_i|} = \frac{|cs(\uparrow v_3, \mathbf{u})(P)|}{|cs(\uparrow v_3, 2)(P)| \cdot |\Sigma_2|} = \frac{4}{4 \cdot 4} = \frac{1}{4},$$

$$\frac{|cs(\uparrow v_4, \mathbf{u})(P)|}{\sum_{i=0}^{3}|cs(\uparrow v_4, i)(P)| \cdot |\Sigma_i|} = \frac{|cs(\uparrow v_4, \mathbf{u})(P)|}{|cs(\uparrow v_4, 0)(P)| \cdot |\Sigma_0| + |cs(\uparrow v_4, 2)(P)| \cdot |\Sigma_2|} = \frac{2}{2 \cdot 4 + 1 \cdot 4} = \frac{1}{6},$$

$$\frac{|cs(\uparrow v_5, \mathbf{u})(P)|}{\sum_{i=0}^{3}|cs(\uparrow v_5, i)(P)| \cdot |\Sigma_i|} = \frac{|cs(\uparrow v_5, \mathbf{u})(P)|}{|cs(\uparrow v_5, 0)(P)| \cdot |\Sigma_0| + |cs(\uparrow v_5, 2)(P)| \cdot |\Sigma_2|} = \frac{3}{3 \cdot 4 + 1 \cdot 4} = \frac{3}{16},$$

$$\frac{|cs(\uparrow v_6, \mathbf{u})(P)|}{\sum_{i=0}^{3}|cs(\uparrow v_6, i)(P)| \cdot |\Sigma_i|} = \frac{|cs(\uparrow v_6, \mathbf{u})(P)|}{|cs(\uparrow v_6, 0)(P)| \cdot |\Sigma_0| + |cs(\uparrow v_6, 1)(P)| \cdot |\Sigma_1|} = \frac{3}{3 \cdot 4 + 1 \cdot 5} = \frac{3}{17}.$$

And now Corollary 75 tells us that

$$\lim_{t \to \infty} \Phi(\mathbf{u}, P, t) = \prod_{i=1}^{6} \frac{|cs(\uparrow v_i, \mathbf{u})(P)|}{\sum_{j=0}^{3}|cs(\uparrow v_i, j)(P)| \cdot |\Sigma_j|} = \frac{2}{13} \cdot \frac{3}{19} \cdot \frac{1}{4} \cdot \frac{1}{6} \cdot \frac{3}{16} \cdot \frac{3}{17} = \frac{9}{268736}.$$

It is possible to introduce mutation operators for nonlinear GP which are ergodic in the sense of Definition 67, but the easiest thing to do is probably just to define the family $\mathcal{M}_{\text{erg}}$ to be the family of all permutations of the search space Ω. The probability distribution p must then be concentrated on any subset of $\mathcal{M}$ which satisfies the ergodicity requirement of Definition 67. This would ensure that Corollary 68 applies.

10. What can be said in the presence of selection in the general case?

Theorem 21 established in [4] which allows us to deduce results such as Theorems 47 and 72, applies only in the absence of selection. The theme of the remainder of the current paper is to establish a few basic properties of the Markov chains associated to evolutionary algorithms in the presence of fitness-proportional selection (as described in Definition 4). Throughout the rest of the paper we shall break up our algorithm, call it $\mathcal{A}$ into sub-algorithms and then consider their composition. This idea will be made clear below:

Proposition 77. *Denote by $\mathcal{A}$ an evolutionary algorithm determined by the cycle $(s_1, s_2, \ldots, s_n)$. Fix i with $1 < i < n$ and let $\mathcal{B}$ and $\mathcal{C}$ denote the sub-algorithms determined by the cycles $(s_1, s_2, \ldots, s_i)$ and $(s_{i+1}, s_2, \ldots, s_n)$, respectively. Recall from Section 5 that $\{p^{\mathcal{A}}_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\Omega^m}$, $\{p^{\mathcal{B}}_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\Omega^m}$ and $\{p^{\mathcal{C}}_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\Omega^m}$ denote the Markov transition matrices associated to the algorithms $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$, respectively. Then we have*

$$\{p^{\mathcal{A}}_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\Omega^m} = \{p^{\mathcal{C}}_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\Omega^m} \cdot \{p^{\mathcal{B}}_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\Omega^m},$$

where $\cdot$ denotes the usual matrix multiplication.

Proof. Denote by λ a probability distribution on Ω^m. Completing a cycle of the algorithm $\mathcal{A}$ amounts to completing a cycle of $\mathcal{B}$ and then completing a cycle of $\mathcal{C}$. The next generation probability distribution upon the completion of the cycle of $\mathcal{B}$ with the input distribution λ is $\{p^{\mathcal{B}}_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\Omega^m} \cdot \lambda$. Likewise the next generation distribution obtained upon the completion of a cycle of the algorithm $\mathcal{C}$ with the input distribution $\{p^{\mathcal{B}}_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\Omega^m} \cdot \lambda$ is just

$$\{p^{\mathcal{C}}_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\Omega^m} \cdot (\{p^{\mathcal{B}}_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\Omega^m} \cdot \lambda) = (\{p^{\mathcal{C}}_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\Omega^m} \cdot \{p^{\mathcal{B}}_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\Omega^m}) \cdot \lambda$$

which means that $\{p^{\mathcal{A}}_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\Omega^m} = \{p^{\mathcal{C}}_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\Omega^m} \cdot \{p^{\mathcal{B}}_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\Omega^m}$ since the equation above holds for an arbitrary input distribution λ. □

We now proceed to study the effects of selection alone. First of all it is convenient to observe the following general fact:

Definition 78. Let $\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ denote a Markov transition matrix on a finite set $\mathcal{X}$. Fix $\mathbf{x} \in \mathcal{X}$. We define the transition support of $\mathbf{x}$ to be the set $S(\mathbf{x}) = \{\mathbf{z} | p_{\mathbf{zx}} > 0\}$ of all states $\mathbf{z}$ from which it is possible to get to $\mathbf{x}$.

Definition 79. Let $\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ denote a Markov transition matrix on a finite set $\mathcal{X}$. Fix $\mathbf{x}$ and $\mathbf{y} \in \mathcal{X}$. We say that $\mathbf{y} \trianglerighteq \mathbf{x}$ if $S(\mathbf{y}) \supseteq S(\mathbf{x})$ and $\forall \mathbf{z} \in S(\mathbf{y})$ we have $p_{\mathbf{zy}} \geqslant p_{\mathbf{zx}}$. Moreover, if either $S(\mathbf{y}) \supsetneq S(\mathbf{x})$ or $p_{\mathbf{zy}} \gneqq p_{\mathbf{zx}}$ for some $\mathbf{z} \in S(\mathbf{x})$ we write $\mathbf{y} \triangleright \mathbf{x}$.

Proposition 80 provides the reason for Definitions 78 and 79:

Proposition 80. *Let $\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ denote a Markov transition matrix on a finite set $\mathcal{X}$. Fix $\mathbf{u}$ and $\mathbf{v} \in \mathcal{X}$ with $\mathbf{u} \trianglerighteq \mathbf{v}$ and an input probability distribution λ on $\mathcal{X}$. Denote by ρ the output distribution ($\rho = \{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \lambda$). Then we have $\rho(\mathbf{u}) \geqslant \rho(\mathbf{v})$. Moreover, if $\mathbf{u} \triangleright \mathbf{v}$ and $\lambda(\mathbf{x}) > 0$ for every $\mathbf{x} \in \mathcal{X}$ then $\rho(\mathbf{u}) \gneqq \rho(\mathbf{v})$.*

Proof. This is a straightforward verification of the definitions:

$$\rho(\mathbf{u}) = \sum_{\mathbf{z}\in\mathcal{X}} \lambda(\mathbf{z}) p_{\mathbf{zu}} = \sum_{\mathbf{z}\in S(\mathbf{u})} \lambda(\mathbf{z}) p_{\mathbf{zu}} \succ \sum_{\mathbf{z}\in S(\mathbf{v})} \lambda(\mathbf{z}) p_{\mathbf{zv}} = \sum_{\mathbf{z}\in\mathcal{X}} \lambda(\mathbf{z}) p_{\mathbf{zv}} = \rho(\mathbf{v}),$$

where

$$\succ = \begin{cases} \geqslant & \text{if } \mathbf{u} \trianglerighteq \mathbf{v}, \\ \gneqq & \text{if } \mathbf{u} \triangleright \mathbf{v}. \end{cases} \qquad \square$$

The following mild technical condition on a Markov transition matrix (which is easily satisfiable by most transition matrices modelling crossover and mutation) will extend Proposition 80.

Definition 81. We call a Markov transition matrix $\{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ non-annihilating if $\forall \mathbf{y} \in \mathcal{X}\ \exists \mathbf{x} \in \mathcal{X}$ such that $q_{\mathbf{xy}} > 0$.

The main reason for introducing Definition 81 is the following fact:

Proposition 82. *A given Markov transition matrix $\{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ is non-annihilating if and only if for every input probability distribution λ on $\mathcal{X}$ with $\lambda(\mathbf{x}) > 0$ for every $\mathbf{x} \in \mathcal{X}$, the output distribution $\rho = \{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \lambda$ also satisfies the property that $\rho(\mathbf{x}) > 0$ for every $\mathbf{x} \in \mathcal{X}$.*

Proof. Given an input distribution λ with $\lambda(\mathbf{x}) > 0$ for every $\mathbf{x} \in \mathcal{X}$ and any state $\mathbf{y} \in \mathcal{X}$, we have $\rho(\mathbf{y}) = \sum_{\mathbf{x}\in\mathcal{X}} \lambda(\mathbf{x}) \cdot q_{\mathbf{xy}} > 0$ if and only if $\lambda(\mathbf{z}) \cdot q_{\mathbf{zy}} > 0$ for some $\mathbf{z} \in \mathbf{X}$ if and only if $q_{\mathbf{zy}} > 0$ for some $\mathbf{z} \in \mathbf{X}$ (since $\lambda(\mathbf{z}) > 0$ automatically by assumption) if and only if the transition matrix $\{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ is non-annihilating. □

Before proceeding any further it is worthwhile to mention that a product of non-annihilating transition matrices is non-annihilating:

Corollary 83. *Given non-annihilating Markov transition matrices $\{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ and $\{m_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$, the matrix $\{r_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} = \{m_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ is non-annihilating as well.*

Proof. Given an input distribution λ with $\lambda(\mathbf{x}) > 0$, since $\{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ is non-annihilating, the "intermediate" output distribution $\mu = \{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}(\lambda)$ also has the property that $\mu(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X}$. Now we have

$$\rho = \{r_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \lambda = \{m_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot (\{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \lambda)) = \{m_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \mu$$

also has the property that $\rho(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X}$ since it is an output of μ under the non-annihilating transition matrix $\{m_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$. By Proposition 82, the transition matrix $\{r_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ is non-annihilating as well. □

Though quite elementary, Proposition 80 readily implies subtle and rather general inequalities about the stationary distributions of the Markov chains for which the last elementary step is selection:

Corollary 84. *Let $\{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ and $\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ denote Markov transition matrices on a finite set $\mathcal{X}$. Fix $\mathbf{u}$ and $\mathbf{v} \in \mathcal{X}$ with $\mathbf{u} \trianglerighteq \mathbf{v}$ where the $\trianglerighteq$ and $\triangleright$ relations are meant with respect to the matrix $\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ and an input probability distribution λ on $\mathcal{X}$. Denote by ρ the output distribution of the composed matrix $\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{q_{\mathbf{xy}}\}$ ($\rho = \{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \lambda$). Then we have $\rho(\mathbf{u}) \geqslant \rho(\mathbf{v})$. Suppose, in addition, the matrix $\{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ is non-annihilating. Now, if $\mathbf{u} \triangleright \mathbf{v}$, and*
$\lambda(\mathbf{x}) > 0$ for every $\mathbf{x} \in \mathcal{X}$ then $\rho(\mathbf{u}) \gneqq \rho(\mathbf{v})$.

Proof. Since

$$\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \lambda = \{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot (\{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \lambda) = \{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \mu,$$

where $\mu = \{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \lambda$, the desired conclusions follow by applying Proposition 80 to the matrix $\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ and the input distribution μ. For the second conclusion we use the assumption that $\{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ is non-annihilating to deduce that $\mu(\mathbf{x}) > 0$ for every $\mathbf{x}$. □

As an almost immediate consequence we deduce the following fact:

Corollary 85. *Let $\{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$, $\{m_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ and $\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ denote Markov transition matrices on a finite set $\mathcal{X}$. Suppose that the matrices $\{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ and $\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ are non-annihilating while the matrix $\{m_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ has the property that $\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}\ m_{\mathbf{xy}} > 0$. Fix $\mathbf{u}$ and $\mathbf{v} \in \mathcal{X}$ with $\mathbf{u} \trianglerighteq \mathbf{v}$ where the $\trianglerighteq$ and $\triangleright$ relations are meant with respect to the matrix $\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$. Then the Markov chain determined by either one of the composed matrices $\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{m_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ or $\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{m_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ is irreducible. Let π denote the unique stationary distribution of the composed chain. Then we have $\pi(\mathbf{u}) \geqslant \pi(\mathbf{v})$. Suppose, in addition, the matrix $\{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ is non-annihilating. Now, if $\mathbf{u} \triangleright \mathbf{v}$, and $\lambda(\mathbf{x}) > 0$ for every $\mathbf{x} \in \mathcal{X}$ then $\pi(\mathbf{u}) \gneqq \pi(\mathbf{v})$.*

Proof. The irreducibility of the composed chain is left as an exercise for the reader. As a hint, the reader may notice that from the assumptions that $\{m_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ has all positive entries while $\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ and $\{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ are non-annihilating, it follows that every one of the composed matrices has all positive entries and, hence, determines an irreducible Markov chain. The second conclusion follows from the fact that $\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ is the leftmost matrix in the composition by applying Corollary 84 to the matrix $\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{a_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ where $\{a_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} = \{m_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ or $\{a_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} = \{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{m_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ with π being the input distribution which is then also the output distribution by stationarity. The condition of Corollary 84 is satisfied thanks to Corollary 83. □

When applying Corollary 85 we have in mind that $\{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ is the Markov transition matrix corresponding to recombination (i.e. a sub-algorithm determined by a single elementary step of type 2: see Definitions 10 and 7), $\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ is the Markov transition matrix corresponding to selection (i.e. a sub-algorithm determined by a single elementary step of type 1: see Definition 4) and $\{m_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ is the Markov transition matrix corresponding to mutation. In order to apply Proposition 80 to the case of fitness-proportional selection we need to determine the relation $\trianglerighteq$ and $\triangleright$ for this special case. Although this task is not difficult, it requires a careful "bookkeeping" analysis. This will be the subject of the next section. We end the current section with an immediate consequence (basically a restatement) of Corollary 85 [12] :

Corollary 86. *Suppose we are given an evolutionary algorithm $\mathcal{A}$ determined by the elementary steps s_1, s_2 and s_3 where s_1 are and s_2 are any elementary step (usually one of them is selection and the other is mutation) which define non-annihilating Markov transition matrices and such that one of these matrices has all positive entries, while s_3 is the elementary step of type* 1, *i.e. selection. As before, let $\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ denote the Markov transition matrix determined by the elementary step s_3 and $\triangleright$ and $\trianglerighteq$ are the defined with respect to the transition matrix $\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$. Then the Markov chain determined by the algorithm $\mathcal{A}$ with state space $\mathcal{X} = \Omega^m$ is irreducible and its unique stationary distribution π satisfies $\pi(\mathbf{x}) \geqslant \pi(\mathbf{y})$ and $\pi(\mathbf{x}) > \pi(\mathbf{y})$ whenever $\mathbf{x} \trianglerighteq \mathbf{y}$ and $\mathbf{x} \triangleright \mathbf{y}$, respectively.*

11. What are the relations $\triangleright$ and $\trianglerighteq$ for the case of fitness-proportional selection?

This section is devoted to classifying the relations $\triangleright$ and $\trianglerighteq$ for the case of fitness-proportional selection. Although this task is not difficult, it requires a careful step-by-step analysis. The reader who is interested only in the net results can read only Definitions 87 and 89, Example 90, Theorem 96 followed by Examples 97, 98, 99 and 100 and Theorem 101 which is illustrated by Example 102. It is recommended (but not essential for understanding) that the reader does not omit the discussion between Example 100 and Theorem 101. We also strongly recommend that the reader makes him/herself familiar with Lemma 88 since this fact is rather simple and reveals a very important step in the classification process.

Definition 87. Fix a population $\mathbf{x} = (x_1, x_2, \ldots, x_m) \in \Omega^m$ and denote by $I(\mathbf{x}) = \{x \mid x = x_i \text{ for some } i \text{ with } 1 \leqslant i \leqslant m\}$ the set of all individuals in the population $\mathbf{x}$.

Lemma 88. *Given populations $\mathbf{x}$ and $\mathbf{y}$, we have $S(\mathbf{x}) \supseteq S(\mathbf{x})$ if and only if $I(\mathbf{x}) \subseteq I(\mathbf{y})$. In particular, a necessary condition for $\mathbf{x} \trianglerighteq \mathbf{y}$ is that $I(\mathbf{x}) \subseteq I(\mathbf{y})$. Moreover, if $I(\mathbf{x}) \subsetneq I(\mathbf{y})$ then $\mathbf{x} \triangleright \mathbf{y}$.*

Proof. Since individuals can only disappear (and new individuals cannot appear) upon the completion of the elementary step of fitness-proportional selection (see Definition 4) it follows immediately that for any populations $\mathbf{z}$ and $\mathbf{w}$ we have $p_{\mathbf{z},\mathbf{w}} \gneq 0$ if and only if $I(\mathbf{z}) \supseteq I(\mathbf{w})$. In other words $S(\mathbf{w}) = \{\mathbf{z} \mid I(\mathbf{z}) \supseteq I(\mathbf{w})\}$. It follows immediately now that if $I(\mathbf{y}) \supseteq I(\mathbf{x})$ then $S(\mathbf{x}) \supseteq S(\mathbf{y})$. On the other hand, if $S(\mathbf{x}) \supseteq S(\mathbf{y})$, then, since $\mathbf{y} \in S(\mathbf{y}) \subseteq S(\mathbf{x})$ we also have $I(\mathbf{y}) \supseteq I(\mathbf{x})$ according to the characterization given above. We deduce now that $I(\mathbf{y}) \supseteq I(\mathbf{x})$ if and only if $S(\mathbf{x}) \supseteq S(\mathbf{y})$. In particular, it follows immediately from the previous statement that $I(\mathbf{y}) \supsetneq I(\mathbf{x})$ if and only if $S(\mathbf{x}) \supsetneq S(\mathbf{y})$. All of the remaining conclusions follow immediately from Definition 79. □

[12] It is worth mentioning that fitness-proportional selection is not the only possible type of selection. Other elementary steps of type 1 include, for instance, tournament selection and rank selection.

Definition 89. Given a population $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ and an individual $x \in I(\mathbf{x})$, denote by $n(\mathbf{x}, x) = |\{i | x = x_i\}|$ the number of times x occurs in the population $\mathbf{x}$.

Example 90. Suppose

$$\mathbf{x} = (a, a, a, b, c, a, b, b, b) \quad \text{and} \quad \mathbf{y} = (b, c, c, c, a, b, b, d, b).$$

Then $I(\mathbf{x}) = \{a, b, c\}$ and $I(\mathbf{y}) = \{a, b, c, d\}$. We also have

$$n(\mathbf{x}, a) = n(\mathbf{x}, b) = 4 \quad \text{and} \quad n(\mathbf{x}, c) = 1.$$

Likewise,

$$n(\mathbf{y}, a) = 1, \quad n(\mathbf{x}, b) = 4, \quad n(\mathbf{x}, c) = 3 \quad \text{and} \quad n(\mathbf{x}, d) = 1.$$

According to Definition 4, when performing fitness-proportional selection, the individuals are chosen independently with probability proportional to their fitness. Thus, if $\mathbf{z} = (z_1, z_2, \ldots, z_m)$ and $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ is obtained from $\mathbf{z}$ by performing fitness-proportional selection, then the probability that $x_i = z$ for a given $z \in I(\mathbf{z})$ is $\frac{n(\mathbf{z},z)\cdot f(z)}{\sum_{i=1}^m f(z_i)}$. Thus $p_{\mathbf{z},\mathbf{x}} = \prod_{j=1}^m \frac{n(\mathbf{z},x_j)\cdot f(x_j)}{\sum_{i=1}^m f(z_i)}$. Moreover, every $x \in I(\mathbf{x})$ occurs in the above product $n(\mathbf{z}, x_j)$ times while every $z \in I(\mathbf{z})$ occurs $n(\mathbf{z}, z)$ times in the denominator sum of each of the multiples and so we deduce the following:

Proposition 91. *Given populations* $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ *and* $\mathbf{z} = (z_1, z_2, \ldots, z_m)$ *we have*

$$p_{\mathbf{zx}} = \begin{cases} \left(\dfrac{1}{\sum_{z \in I(\mathbf{z})} n(\mathbf{z}, z) \cdot f(z)}\right)^m \displaystyle\prod_{x \in I(\mathbf{x})} (n(\mathbf{z}, x))^{n(\mathbf{x},x)} \cdot (f(x))^{n(\mathbf{x},x)} & \text{if } I(\mathbf{x}) \subseteq I(\mathbf{z}), \\ 0 & \text{otherwise}. \end{cases}$$

In particular, $p_{\mathbf{zx}}$ *does not depend on the way the individuals in* $\mathbf{z}$ *and in* $\mathbf{x}$ *are ordered, but only depends on the sets* $I(\mathbf{x})$ *and* $I(\mathbf{x})$ *and the numbers* $n(\mathbf{x}, x)$ *for* $x \in I(\mathbf{x})$ *and* $n(\mathbf{z}, z)$ *for* $z \in I(\mathbf{z})$. *In other words, if* σ *and* τ *demote arbitrary permutations of the set* $\{1, 2, \ldots, m\}$, *If* $\mathbf{x}_\sigma = (x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(m)})$ *and* $\mathbf{z}_\tau = (z_{\tau(1)}, z_{\tau(2)}, \ldots, z_{\tau(m)})$ *then* $p_{\mathbf{z}_\sigma \mathbf{x}_\tau} = p_{\mathbf{zx}}$.

In order to continue the investigation of the $\trianglerighteq$ relation for the case of fitness-proportional selection, it is convenient to introduce the following notions:

Definition 92. Given populations $\mathbf{x}$ and $\mathbf{y}$ of size m let

$$I(\mathbf{y}|\mathbf{x}) = \{y | y \in I(\mathbf{y}), n(\mathbf{y}, y) > n(\mathbf{x}, y)\}$$

(if $y \notin I(\mathbf{x})$ then $n(\mathbf{x}, y) = 0$). Moreover, for $y \in I(\mathbf{y}|\mathbf{x})$ let $\kappa(\mathbf{y}|\mathbf{x}, y) = n(\mathbf{y}, y) - n(\mathbf{x}, y)\}$.

Example 93. Continuing with Example 90, we have $n(\mathbf{x}, x) > n(\mathbf{y}, x)$ if and only if $x = a$ and so $I(\mathbf{x}|\mathbf{y}) = \{a\}$. Likewise, $n(\mathbf{y}, y) < n(\mathbf{y}, y)$ if and only if $y = c$ or $y = d$ (since $d \notin I(\mathbf{x})$ according to Definition 92 we have $n(\mathbf{x}, d) = 0 < 1 = n(\mathbf{x}, d)$) and so $I(\mathbf{y}|\mathbf{x}) = \{c, d\}$. Moreover, we also have $\kappa(\mathbf{x}|\mathbf{y}, a) = 4 - 1 = 3$, $\kappa(\mathbf{y}|\mathbf{x}, c) = 3 - 1 = 2$ and $\kappa(\mathbf{y}|\mathbf{x}, d) = 1 - 0 = 1$.

The sets $I(\mathbf{y}|\mathbf{x})$ play a crucial role in discovering a sufficient and necessary condition for a population $\mathbf{x} \trianglerighteq \mathbf{y}$ in view of the fact below:

Lemma 94. *Given populations* $\mathbf{z}$, $\mathbf{y}$ *and* $\mathbf{x}$ *with* $I(\mathbf{z}) \supseteq I(\mathbf{y}) \supseteq I(\mathbf{x})$, *have the following*:

$$\sum_{x \in I(\mathbf{x}|\mathbf{y})} \kappa(\mathbf{x}|\mathbf{y}, x) = \sum_{y \in I(\mathbf{y}|\mathbf{x})} \kappa(\mathbf{y}|\mathbf{x}, y),$$

$$\frac{p_{\mathbf{zx}}}{p_{\mathbf{zy}}}=\frac{\prod_{x\in I(\mathbf{x}|\mathbf{y})}(n(\mathbf{z},x))^{\kappa(\mathbf{x}|\mathbf{y},x)}\cdot(f(x))^{\kappa(\mathbf{x}|\mathbf{y},x)}}{\prod_{y\in I(\mathbf{y}|\mathbf{x})}(n(\mathbf{z},y))^{\kappa(\mathbf{y}|\mathbf{x},y)}\cdot(f(y))^{\kappa(\mathbf{y}|\mathbf{x},y)}}.$$

Proof. Given populations $\mathbf{z}$, $\mathbf{y}$ and $\mathbf{x}$, from Definition 92 it follows that for every $x\in I(\mathbf{x})$ we have

$$n(\mathbf{x},x)=\begin{cases}\min(n(\mathbf{x},x),n(\mathbf{y},x)) & \text{if } x\notin I(\mathbf{x}|\mathbf{y}),\\ \min(n(\mathbf{x},x),n(\mathbf{y},x))+\kappa(\mathbf{x}|\mathbf{y},x) & \text{if } x\in I(\mathbf{x}|\mathbf{y}).\end{cases}$$

Likewise, for every $y\in I(\mathbf{y})$ we have

$$n(\mathbf{y},y)=\begin{cases}\min(n(\mathbf{x},y),n(\mathbf{y},y)) & \text{if } y\notin I(\mathbf{y}|\mathbf{x}) \text{ and } y\in I(\mathbf{x}),\\ \min(n(\mathbf{x},x),n(\mathbf{y},x))+\kappa(\mathbf{y}|\mathbf{x},x) & \text{if } y\in I(\mathbf{y}|\mathbf{x}) \text{ and } y\in I(\mathbf{x}),\\ \kappa(\mathbf{y}|\mathbf{x},x) & \text{if } y\in I(\mathbf{y}|\mathbf{x}) \text{ and } y\notin I(\mathbf{x}).\end{cases}$$

Since there are totally m elements in every population we must have

$$\sum_{y\in I(\mathbf{y})}n(\mathbf{y},y)=\sum_{x\in I(\mathbf{y})}n(\mathbf{x},x)=m.$$

Rearranging the terms in both sides of the last equation according to the observations made above, we obtain

$$\sum_{x\in I(\mathbf{x})}\min(n(\mathbf{x},x),n(\mathbf{y},x))+\sum_{x\in I(\mathbf{x}|\mathbf{y})}\kappa(\mathbf{x}|\mathbf{y},x)=\sum_{y\in I(\mathbf{x})}\min(n(\mathbf{x},y),n(\mathbf{y},y))+\sum_{y\in I(\mathbf{y}|\mathbf{x})}\kappa(\mathbf{y}|\mathbf{x},y)$$

and the first desired equation follows by subtracting $\sum_{x\in I(\mathbf{x})}\min(n(\mathbf{x},x),n(\mathbf{y},x))$ from both sides. The second equation follows by rearranging the multiples in the formula of Proposition 91 according to the equation above and letting $k(\mathbf{z})=(\frac{1}{\sum_{z\in I(\mathbf{z})}n(\mathbf{z},z)\cdot f(z)})^m$ so that we can write

$$p_{\mathbf{zx}}=k(\mathbf{z})\cdot\prod_{x\in I(\mathbf{x})}(n(\mathbf{z},x))^{\min(n(\mathbf{x},x),n(\mathbf{y},x))}\cdot(f(x))^{\min(n(\mathbf{x},x),n(\mathbf{y},x))})\prod_{x\in I(\mathbf{x}|\mathbf{y})}(n(\mathbf{z},x))^{\kappa(\mathbf{x}|\mathbf{y},x)}\cdot(f(x))^{\kappa(\mathbf{x}|\mathbf{y},x)})$$

and, likewise,

$$p_{\mathbf{zy}}=k(\mathbf{z})\cdot\prod_{y\in I(\mathbf{x})}(n(\mathbf{z},y))^{\min(n(\mathbf{x},y),n(\mathbf{y},y))}\cdot(f(y))^{\min(n(\mathbf{x},y),n(\mathbf{y},y))})\prod_{y\in I(\mathbf{y}|\mathbf{x})}(n(\mathbf{z},y))^{\kappa(\mathbf{y}|\mathbf{x},y)}\cdot(f(y))^{\kappa(\mathbf{y}|\mathbf{x},y)}).$$

Now the common factor

$$k(\mathbf{z})\cdot\prod_{x\in I(\mathbf{x})}(n(\mathbf{z},x))^{\min(n(\mathbf{x},x),n(\mathbf{y},x))}\cdot(f(x))^{\min(n(\mathbf{x},x),n(\mathbf{y},x))})$$

on the top and the bottom of the ratio $\frac{p_{\mathbf{zx}}}{p_{\mathbf{zy}}}$ is canceled out and we obtain the desired formula. □

In fact, according to Lemma 88, we have $\mathbf{x}\trianglerighteq\mathbf{y}\Longrightarrow I(\mathbf{y})\supseteq I(\mathbf{x})$. Moreover, since new individuals cannot appear as a result of selection, whenever $\mathbf{z}\in S(\mathbf{y})$ (see Definition 78 for the meaning of $S(\mathbf{y})$) we must have $I(\mathbf{z})\supseteq I(\mathbf{y})$. Therefore, $\mathbf{x}\trianglerighteq\mathbf{y}$ if and only if $I(\mathbf{y})\supseteq I(\mathbf{x})$) and $\forall\mathbf{z}$ such that $I(\mathbf{z})\supseteq I(\mathbf{y})$ we have $p_{\mathbf{zx}}\geqslant p_{\mathbf{zy}}$. The condition $p_{\mathbf{zx}}\geqslant p_{\mathbf{zy}}$ can be restated equivalently as $\frac{p_{\mathbf{zx}}}{p_{\mathbf{zy}}}\geqslant 1$. But, thanks to Lemma 94, we have

$$\begin{aligned}\frac{p_{\mathbf{zx}}}{p_{\mathbf{zy}}}&=\frac{\prod_{x\in I(\mathbf{x}|\mathbf{y})}(n(\mathbf{z},x))^{\kappa(\mathbf{x}|\mathbf{y},x)}\cdot(f(x))^{\kappa(\mathbf{x}|\mathbf{y},x)}}{\prod_{y\in I(\mathbf{y}|\mathbf{x})}(n(\mathbf{z},y))^{\kappa(\mathbf{y}|\mathbf{x},y)}\cdot(f(y))^{\kappa(\mathbf{y}|\mathbf{x},y)}}=\frac{\prod_{x\in I(\mathbf{x}|\mathbf{y})}(n(\mathbf{z},x))^{\kappa(\mathbf{x}|\mathbf{y},x)}}{\prod_{y\in I(\mathbf{y}|\mathbf{x})}(n(\mathbf{z},y))^{\kappa(\mathbf{y}|\mathbf{x},y)}}\cdot\frac{\prod_{x\in I(\mathbf{x}|\mathbf{y})}(f(x))^{\kappa(\mathbf{x}|\mathbf{y},x)}}{\prod_{y\in I(\mathbf{y}|\mathbf{x})}(f(y))^{\kappa(\mathbf{y}|\mathbf{x},y)}}\geqslant 1\\ &\Longleftrightarrow\frac{\prod_{x\in I(\mathbf{x}|\mathbf{y})}(n(\mathbf{z},x))^{\kappa(\mathbf{x}|\mathbf{y},x)}}{\prod_{y\in I(\mathbf{y}|\mathbf{x})}(n(\mathbf{z},y))^{\kappa(\mathbf{y}|\mathbf{x},y)}}\geqslant\frac{\prod_{y\in I(\mathbf{y}|\mathbf{x})}(f(y))^{\kappa(\mathbf{y}|\mathbf{x},y)}}{\prod_{y\in I(\mathbf{x}|\mathbf{y})}(f(x))^{\kappa(\mathbf{x}|\mathbf{y},x)}}.\end{aligned}$$

Observing that

$$\frac{\prod_{y\in I(\mathbf{y}|\mathbf{x})}(f(y))^{\kappa(\mathbf{y}|\mathbf{x},y)}}{\prod_{y\in I(\mathbf{x}|\mathbf{y})}(f(x))^{\kappa(\mathbf{x}|\mathbf{y},x)}}$$

does not depend on $\mathbf{z}$ at all we deduce the following:

Lemma 95. *Given populations* $\mathbf{x}$ *and* $\mathbf{y}$ *of size* m, *we have* $\mathbf{x} \trianglerighteq \mathbf{y}$ *if and only if* $I(\mathbf{y}) \supseteq I(\mathbf{x})$ *and*

$$\min_{I(\mathbf{z})\supseteq I(\mathbf{x})} \frac{\prod_{x\in I(\mathbf{x}|\mathbf{y})}(n(\mathbf{z}, x))^{\kappa(\mathbf{x}|\mathbf{y},x)}}{\prod_{y\in I(\mathbf{y}|\mathbf{x})}(n(\mathbf{z}, y))^{\kappa(\mathbf{y}|\mathbf{x},y)}} \geqslant \frac{\prod_{y\in I(\mathbf{y}|\mathbf{x})}(f(y))^{\kappa(\mathbf{y}|\mathbf{x},y)}}{\prod_{y\in I(\mathbf{x}|\mathbf{y})}(f(x))^{\kappa(\mathbf{x}|\mathbf{y},x)}}.$$

Thanks to Lemma 95, the rest of our analysis boils down to constructing a population $\mathbf{z}$ which minimizes the ratio over $\mathbf{z}$ with $I(\mathbf{z}) \supseteq I(\mathbf{y})$. In view of Proposition 91, without loss of generality, we can assume that the first $|I(y)|$ individuals of $\mathbf{z}$ enumerate the elements of $I(y)$, i.e. $\mathbf{z} = \{y_1, y_2, \ldots, y_{|I(y)|}, z_1, z_2, \ldots, z_{m-|I(y)|}\}$. Our goal is then to select $z_1, z_2, \ldots, z_{m-|I(y)|}$ in a way which minimizes this ratio. First, it is worth pointing out, that unless $\mathbf{y} = \mathbf{x}_\sigma$ for some permutation σ of $\{1, 2, \ldots, m\}$ in the sense described in Proposition 91 (in which case we trivially have $\mathbf{x} \trianglerighteq \mathbf{y}$ thanks to Proposition 91), we can assume that $I(\mathbf{y}|\mathbf{x}) \neq \emptyset$. (Indeed, according to Lemma 94, we have $\sum_{x\in I(\mathbf{x}|\mathbf{y})}\kappa(\mathbf{x}|\mathbf{y}, x) = \sum_{y\in I(\mathbf{y}|\mathbf{x})}\kappa(\mathbf{y}|\mathbf{x}, y)$. If $I(\mathbf{y}|\mathbf{x}) \neq \emptyset$ then $\sum_{x\in I(\mathbf{x}|\mathbf{y})}\kappa(\mathbf{x}|\mathbf{y}, x) = \sum_{y\in I(\mathbf{y}|\mathbf{x})}\kappa(\mathbf{y}|\mathbf{x}, y) = 0$ which forces $I(\mathbf{x}|\mathbf{y}) = \emptyset$. But then for every $y \in I(\mathbf{y})$ we have $y \notin I(\mathbf{y}|\mathbf{x}) \Longrightarrow I(\mathbf{y}) \subseteq I(\mathbf{x})$ and $n(\mathbf{y}, y) \leqslant n(\mathbf{x}, y)$ and, since $I(\mathbf{x}|\mathbf{y}) = \emptyset$, also $n(\mathbf{y}, y) \geqslant n(\mathbf{x}, y)$. Summarizing, we obtain $n(\mathbf{y}, y) = n(\mathbf{x}, y)$ $\forall y \in I(\mathbf{y}) = I(\mathbf{y})$ which means that $\mathbf{y} = \mathbf{x}_\sigma$.) Next, we observe that $\forall i$ we have $z_i \in I(\mathbf{y}|\mathbf{x})$. (If not, then for some i we have $z_i \notin I(\mathbf{y}|\mathbf{x})$. In such a case, since replacing z_i with an element of $I(\mathbf{y}|\mathbf{x}) \neq \emptyset$ will increase the denominator and either decrease (in case if $z_i \in I(\mathbf{x}|\mathbf{y})$) or not influence the numerator in any way (since it is clear from Definition 92 that $I(\mathbf{y}|\mathbf{x}) \cap I(\mathbf{x}|\mathbf{y}) = \emptyset$) of the ratio on the L.H.S. of the inequality of Lemma 95. This in turn would only decrease this ratio so that $\mathbf{z}$ cannot minimize it.) Since z_is are chosen from the set $I(\mathbf{y}|\mathbf{x})$, and $I(\mathbf{y}|\mathbf{x}) \cap I(\mathbf{x}|\mathbf{y}) = \emptyset$, for every $x \in I(\mathbf{x}|\mathbf{y})$ $x \neq z_i$ for $1 \leqslant i \leqslant m - |I(y)|$ and $x = y_j$ for a unique j with $1 \leqslant j \leqslant |I(y)|$ (since $I(\mathbf{y}) \supseteq I(\mathbf{x})$ and $y_1, y_2, \ldots, y_{|I(y)|}$ enumerate the elements of $I(\mathbf{y})$) we have $n(\mathbf{z}, x) = 1$ for every $x \in I(\mathbf{x}|\mathbf{y})$. But then the numerator of the ratios on the L.H.S. of Lemma 95, $\prod_{x\in I(\mathbf{x}|\mathbf{y})}(n(\mathbf{z}, x))^{\kappa(\mathbf{x}|\mathbf{y},x)} = 1$ and the L.H.S. of the inequality in Lemma 94 simplifies to

$$\min_{\mathbf{z}\in Q} \frac{1}{\prod_{y\in I(\mathbf{y}|\mathbf{x})}(n(\mathbf{z}, y))^{\kappa(\mathbf{y}|\mathbf{x},y)}},$$

where $Q = \{\mathbf{z}|\mathbf{z} = (y_1, y_2, \ldots, y_{|I(y)|}, z_1, z_2, \ldots, z_{m-|I(y)|}), y_1, y_2, \ldots, y_{|I(y)|}$ enumerate $I(\mathbf{y})$ and $z_i \in I(\mathbf{y}|\mathbf{x})\}$. Moreover, notice that z_is can be chosen arbitrarily from the set $I(\mathbf{y}|\mathbf{x})$ while for every $y \in I(\mathbf{y}|\mathbf{x})$ there exists exactly one $1 \leqslant j \leqslant |I(y)|$ with $y = y_j$. We then have $n(\mathbf{z}, y) = 1 + |\{i|1 \leqslant i \leqslant m - |I(y)|, z_i = y\}|$ and $\sum_{y\in I(\mathbf{y}|\mathbf{x})}n(\mathbf{z}, y) = |I(\mathbf{y}|\mathbf{x})| + m - |I(y)|$. On the other hand, given a finite sequence of natural numbers $\{n_y\}_{y\in I(\mathbf{y}|\mathbf{x})}$ satisfying the constraint $\sum_{y\in I(\mathbf{y}|\mathbf{x})}n_y = |I(\mathbf{y}|\mathbf{x})| + m - |I(y)|$, we can construct $\mathbf{z} = \{y_1, y_2, \ldots, y_{|I(y)|}, z_1, z_2, \ldots, z_{m-|I(y)|}\}$ with $n(\mathbf{z}, y) = n_y$ by picking exactly $n_y - 1$ z_is equaling to y for every $y \in I(\mathbf{y}|\mathbf{x})$. All of this is summarized in the main theorem below:

Theorem 96. *Given populations* $\mathbf{x}$ *and* $\mathbf{y}$ *of size* m, *we have* $\mathbf{x} \trianglerighteq \mathbf{y}$ *with respect to fitness-proportional selection as described in Definition* 4, *if and only if* $I(\mathbf{y}) \supseteq I(\mathbf{x})$ *and*

$$\max_{\{n_y\}_{y\in I(\mathbf{y}|\mathbf{x})}\in Q(\mathbf{y}|\mathbf{x})} \prod_{y\in I(\mathbf{y}|\mathbf{x})} (n_y)^{\kappa(\mathbf{y}|\mathbf{x},y)} \leqslant \frac{\prod_{x\in I(\mathbf{x}|\mathbf{y})}(f(x))^{\kappa(\mathbf{x}|\mathbf{y},x)}}{\prod_{y\in I(\mathbf{y}|\mathbf{x})}(f(y))^{\kappa(\mathbf{y}|\mathbf{x},y)}},$$

where

$$Q(\mathbf{y}|\mathbf{x}) = \left\{ \{n_y\}_{y\in I(\mathbf{y}|\mathbf{x})} | \forall y \in I(\mathbf{y}|\mathbf{x}) n_y \in \mathbb{N}, \sum_{y\in I(\mathbf{y}|\mathbf{x})} n_y = |I(\mathbf{y}|\mathbf{x})| + m - |I(y)| \right\}.$$

Below we illustrate Theorem 96 with a few simple examples:

Example 97. Continuing with Examples 90 and 93, notice that we do have $I(\mathbf{y}) \supsetneq I(\mathbf{x})$ and $|I(\mathbf{y}|\mathbf{x})| + m - |I(\mathbf{y})| = 2 + 9 - 4 = 7$. so according to Theorem 96 we have $\mathbf{x} \triangleright \mathbf{y}$ if and only if

$$\frac{f(a)^3}{f(c)^2 \cdot f(d)} \geqslant \max_{n_c+n_d=7, n_c\geqslant 1 \text{ and } n_d\geqslant 1} n_c^2 \cdot n_d.$$

There are six possible pairs (n_c, n_d) over which we want to maximize the product $n_c^2 \cdot n_d$. These are (1, 6), (2, 5), (3, 4), (4, 3), (5, 2) and (6, 1). Moreover, by symmetry, since the power of the coefficient n_c in the product is bigger than that of n_d we only have to cheque 3 of these pairs: (4, 3), (5, 2) and (6, 1). The corresponding products are $4^2 \cdot 3 = 48$, $5^2 \cdot 2 = 50$ and $6^2 \cdot 1 = 36$. Then biggest one among these is 50 and so we deduce that $\mathbf{x} \rhd \mathbf{y}$ if and only if $\frac{f(a)^3}{f(c)^2 \cdot f(d)} \geqslant 50$.

Example 98. Suppose $\mathbf{y} = \{y_1, y_2, \ldots, y_m\}$ with $y_i \neq y_j$ for $i \neq j$ (i.e. $\mathbf{y}$ is a population consisting of distinct individuals). Now let $\mathbf{x} = \{x_1, x_2, \ldots, x_m\}$ with $I(\mathbf{y}) \supseteq I(\mathbf{x})$. Notice that in this example $I(\mathbf{y}|\mathbf{x}) = I(\mathbf{y}) - I(\mathbf{x})$ while $I(\mathbf{x}|\mathbf{y}) = \{x \mid \text{there is more than one } i \text{ s.t. } x = x_i\}$. Moreover, $\forall x \in I(\mathbf{x}|\mathbf{y})$ we have $\kappa(\mathbf{x}|\mathbf{y}, x) = n(\mathbf{x}, x) - 1$ (since $n(\mathbf{y}, x) = 1$) and $\forall y \in I(\mathbf{y}|\mathbf{x})$ we have $n(\mathbf{y}, y) = 1$ so that $0 < \kappa(\mathbf{y}|\mathbf{x}, y) \leqslant n(\mathbf{y}, y)$ and we have $\kappa(\mathbf{y}|\mathbf{x}, y) = 1$. Finally, observe that the set $Q(\mathbf{y}|\mathbf{x}) = \{\{1, 1, 1, \ldots, 1\}\}$ since $|I(y)| = m$ and we must have $\sum_{y \in I(\mathbf{y}|\mathbf{x})} n_y = |I(\mathbf{y}|\mathbf{x})| + m - |I(y)| = |I(\mathbf{y}|\mathbf{x})|$ and $n_y \geqslant 1$ which forces $n_y = 1$ $\forall y \in I(\mathbf{y}|\mathbf{x})$. According to Theorem 96 we have $\mathbf{x} \unrhd \mathbf{y}$ if and only if $\frac{\prod_{x \in I(\mathbf{x}|\mathbf{y})} (f(x))^{\kappa(\mathbf{x}|\mathbf{y}, x)}}{\prod_{y \in I(\mathbf{y}|\mathbf{x})} (f(y))^{\kappa(\mathbf{y}|\mathbf{x}, y)}} \geqslant 1$ if and only if

$$\prod_{x \in I(\mathbf{x}|\mathbf{y})} (f(x))^{n(\mathbf{x}, x) - 1} \geqslant \prod_{y \in I(\mathbf{y}|\mathbf{x})} f(y).$$

Example 99. Continuing with Example 98, suppose, in addition, that there is exactly one individual in $\mathbf{x}$ which occurs more than once in this population. That is, without loss of generality, let $\mathbf{x} = (y_1, y_2, \ldots, y_k, y_m, y_m, \ldots, y_m)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_k, y_{k+1}, \ldots, y_m)$ where $y_i \neq y_j$ for $i \neq j$ and $k < m$. This is a special case of Example 98 where $I(\mathbf{x}|\mathbf{y}) = \{y_m\}$ and $I(\mathbf{y}|\mathbf{x}) = \{y_{k+1}, y_{k+2}, \ldots, y_{m-1}\}$. According to the conclusion of Example 98 we have $\mathbf{x} \unrhd \mathbf{y}$ if and only if

$$(f(y_m))^{m-(k+1)} \geqslant \prod_{i=1}^{m-(k+1)} f(y_{k+i}) \quad \text{if and only if } f(y_m) \geqslant \sqrt[m-(k+1)]{\prod_{i=1}^{m-(k+1)} f(y_{k+i})}$$

which, in words, says that $\mathbf{x} \unrhd \mathbf{y}$ if and only if the fitness of the unique repeated individual of $\mathbf{x}$ is at least as large as the geometric mean of the fitness of all the individuals in $\mathbf{y}$ which do not occur in $\mathbf{x}$. It is also worth pointing out that even if the inequality above is an equality we still have $\mathbf{x} \rhd \mathbf{y}$ since $I(\mathbf{y}) \supsetneq I(\mathbf{x})$ and so $S(\mathbf{x}) \supsetneq S(\mathbf{y})$. In particular, even if the fitness function is flat, the relation $\rhd \neq \emptyset$ in case one uses fitness-proportional selection.

Example 100. Now consider an "opposite extreme" to Example 98 (in the sense of the diversity of elements in the population) where $I(\mathbf{y}) = I(\mathbf{x}) = \{x, y\}$. Let us say $n(\mathbf{x}, x) = k$ (which implies that $n(\mathbf{x}, y) = m - k$) and $n(\mathbf{y}, x) = l < k$ (hence $n(\mathbf{y}, y) = m - l$). It follows then that $I(\mathbf{x}|\mathbf{y}) = \{x\}$ and $I(\mathbf{y}|\mathbf{x}) = \{y\}$. Moreover, $\kappa(\mathbf{y}|\mathbf{x}, y) = \kappa(\mathbf{x}|\mathbf{y}, x) = k - l$. Since $|I(\mathbf{y}|\mathbf{x})| = |\{y\}|$, follows that $Q(\mathbf{y}|\mathbf{x}) = \{\{k - l\}\}$ and which makes the maximization procedure trivial. According to Theorem 96, we have $\mathbf{x} \unrhd \mathbf{y}$ if and only if $(k - l)^{k-l} \leqslant \frac{f(x)^{k-l}}{f(y)^{k-l}}$ if and only if $f(x) \geqslant (k - l) \cdot f(y)$.

Theorem 96 tells us that in order to cheque if $\mathbf{x} \unrhd \mathbf{y}$ with respect to fitness-proportional selection we need to solve an integer optimization problem subject to linear constraints. Examples 98, 99 and 100 are particularly simple mainly because the sets $Q(\mathbf{y}|\mathbf{x})$ were singletons so there was not much choice for the maximizing domain element. Although we do not intend to pursue studying this optimization problem in much detail since it is not the main subject of the current paper, it is worth mentioning that the method of Lagrange multipliers allows us to give an upper bound on the

$$\max_{\{n_y\}_{y \in I(\mathbf{y}|\mathbf{x})} \in Q(\mathbf{y}|\mathbf{x})} \prod_{y \in I(\mathbf{y}|\mathbf{x})} (n_y)^{\kappa(\mathbf{y}|\mathbf{x}, y)}$$

by letting n_y's range over positive real numbers subject to the linear constraint $\sum_{y \in I(\mathbf{y}|\mathbf{x})} n_y = |I(\mathbf{y}|\mathbf{x})| + m - |I(y)|$. Moreover, if one wants an exact solution then the method allows to narrow down the choice of suitable integer sequences significantly: Indeed, according to the method of Lagrange multipliers, if any local maximum of a differentiable function $f(\overrightarrow{n})$ on an open set $D \subseteq \mathbb{R}^n$ subject to the constraint $g(\overrightarrow{n}) = c$ where g is another differentiable function on D is achieved at a point q, then we must have $\nabla f(q) = \lambda \cdot \nabla g(q)$ where ∇ denotes the gradient (derivative of a real-valued function) operator and λ is some constant proportionality coefficient (in other words, the gradients of f and g evaluated at the point q must be collinear vectors). In our case $f, g : U \subseteq \mathbb{R}^{I(\mathbf{y}|\mathbf{x})} \to \mathbb{R}$ where $U = \{\overrightarrow{n} \mid \overrightarrow{n} \in \mathbb{R}^{I(\mathbf{y}|\mathbf{x})}$ and $n_y \geqslant 0\}$ are defined according to the following formulas: $f(\overrightarrow{n}) = \prod_{y \in I(\mathbf{y}|\mathbf{x})} (n_y)^{\kappa(\mathbf{y}|\mathbf{x}, y)}$ and $g(\overrightarrow{n}) = \sum_{y \in I(\mathbf{y}|\mathbf{x})} n_y$. Our goal is to maximize f subject to the constraint $g(\overrightarrow{n}) = |I(\mathbf{y}|\mathbf{x})| + m - |I(y)|$ where $\overrightarrow{n} = \{n_y\}_{y \in I(\mathbf{y}|\mathbf{x})}$. Clearly $\forall n_y$ we have $\frac{\partial g}{\partial n_y} = 1$ and so the condition $\nabla f(q) = \lambda \cdot \nabla g(q)$ boils down to the condition $\frac{\partial f}{\partial n_u} = \frac{\partial f}{\partial n_v}$ for every u and $v \in I(\mathbf{y}|\mathbf{x})$. For any given $w \in I(\mathbf{y}|\mathbf{x})$ we have

$$\frac{\partial f}{\partial n_w} = \kappa(\mathbf{y}|\mathbf{x}, w) \cdot (n_w)^{\kappa(\mathbf{y}|\mathbf{x}, w)-1} \cdot \prod_{y \in I(\mathbf{y}|\mathbf{x}), y \neq w} (n_y)^{\kappa(\mathbf{y}|\mathbf{x}, y)}.$$

Therefore, the equation $\frac{\partial f}{\partial n_u} = \frac{\partial f}{\partial n_v}$ holds for every u and $v \in I(\mathbf{y}|\mathbf{x})$ if and only if for every u and $v \in I(\mathbf{y}|\mathbf{x})$ we have $\kappa(\mathbf{y}|\mathbf{x}, u) \cdot n_v = \kappa(\mathbf{y}|\mathbf{x}, v) \cdot n_u$ if and only if $\frac{n_u}{\kappa(\mathbf{y}|\mathbf{x}, u)} = \frac{n_v}{\kappa(\mathbf{y}|\mathbf{x}, v)}$ for all u and $v \in I(\mathbf{y}|\mathbf{x})$. In other words, the equation $\frac{\partial f}{\partial n_u} = \frac{\partial f}{\partial n_v}$ holds for every u and $v \in I(\mathbf{y}|\mathbf{x})$ if and only if the ratio $\frac{n_y}{\kappa(\mathbf{y}|\mathbf{x}, y)} = \alpha$ is a constant independent of $y \in I(\mathbf{y}|\mathbf{x})$. Moreover, this also gives us $n_u = \frac{n_v}{\kappa(\mathbf{y}|\mathbf{x}, v)} \cdot \kappa(\mathbf{y}|\mathbf{x}, u) = \alpha \cdot \kappa(\mathbf{y}|\mathbf{x}, u)$ for every $u \in I(\mathbf{y}|\mathbf{x})$ and, according to the constraint, we also have

$$\sum_{y \in I(\mathbf{y}|\mathbf{x})} n_y = \alpha \cdot \sum_{y \in I(\mathbf{y}|\mathbf{x})} \kappa(\mathbf{y}|\mathbf{x}, y) = |I(\mathbf{y}|\mathbf{x})| + m - |I(y)|$$

which gives

$$\alpha = \frac{|I(\mathbf{y}|\mathbf{x})| + m - |I(y)|}{\sum_{y \in I(\mathbf{y}|\mathbf{x})} \kappa(\mathbf{y}|\mathbf{x}, y)}.$$

Notice that the point $\overrightarrow{q}$ with coordinates $n_u = \frac{|I(\mathbf{y}|\mathbf{x})| + m - |I(y)|}{\sum_{y \in I(\mathbf{y}|\mathbf{x})} \kappa(\mathbf{y}|\mathbf{x}, y)} \cdot \kappa(\mathbf{y}|\mathbf{x}, y)$ is the unique point which satisfies $\nabla f(\overrightarrow{q}) = \lambda \cdot \nabla g(\overrightarrow{q})$. We argue that this point must be the global maximum of the function f on the domain $D = \{n_y \mid n_y \geqslant 0\} \cap g^{-1}(\{|I(\mathbf{y}|\mathbf{x})| + m - |I(y)|\})$ which is a closed and bounded subset of $\mathbb{R}^{I(\mathbf{y}|\mathbf{x})}$ and, hence, is compact. Clearly the function f is continuous on D and, since D is compact it must achieve a minimum and a maximum on D. The only interesting case to consider is when $|I(\mathbf{y}|\mathbf{x})| > 1$ (indeed, if $|I(\mathbf{y}|\mathbf{x})| = 1$, then D is a singleton set whose only point is $\overrightarrow{q}$ so that it is trivially a global maximum). If maximum of f was not the point $\overrightarrow{q}$ then it must be the point on the boundary of D (since it is the only interior point satisfying $\nabla f(\overrightarrow{q}) = \lambda \cdot \nabla g(\overrightarrow{q})$). But every boundary point of D has at least one zero coordinate so that $f(\overrightarrow{y}) = 0$ for every boundary point $\overrightarrow{y}$ of D. On the other hand $f(\overrightarrow{q}) > 0$. Thus we deduce that f achieves a global maximum at the point $\overrightarrow{q}$. We then have the following sufficient condition for $\mathbf{x} \trianglerighteq \mathbf{y}$:

Theorem 101. *Given populations* $\mathbf{x}$ *and* $\mathbf{y}$ *of size m, we have* $\mathbf{x} \trianglerighteq \mathbf{y}$ *with respect to fitness-proportional selection as described in Definition* 4, *if* $I(\mathbf{y}) \supseteq I(\mathbf{x})$ *and*

$$\frac{\prod_{x \in I(\mathbf{x}|\mathbf{y})} (f(x))^{\kappa(\mathbf{x}|\mathbf{y}, x)}}{\prod_{y \in I(\mathbf{y}|\mathbf{x})} (f(y))^{\kappa(\mathbf{y}|\mathbf{x}, y)}} \geqslant \left(\frac{|I(\mathbf{y}|\mathbf{x})| + m - |I(y)|}{\sum_{y \in I(\mathbf{y}|\mathbf{x})} \kappa(\mathbf{y}|\mathbf{x}, y)} \right)^{\sum_{y \in I(\mathbf{y}|\mathbf{x})} \kappa(\mathbf{y}|\mathbf{x}, y)} \cdot \prod_{y \in I(\mathbf{y}|\mathbf{x})} \kappa(\mathbf{y}|\mathbf{x}, y)^{\kappa(\mathbf{y}|\mathbf{x}, y)}$$

Example 102. Continuing with Examples 90, 93 and 97, according to Theorem 101 we have $\mathbf{x} \rhd \mathbf{y}$ (recall that we do have $I(\mathbf{y}) \supseteq I(\mathbf{y})$) if

$$\frac{f(a)^3}{f(c)^2 \cdot f(d)} \geqslant \left(\frac{|I(\mathbf{y}|\mathbf{x})| + m - |I(y)|}{\kappa(\mathbf{y}|\mathbf{x}, c) + \kappa(\mathbf{y}|\mathbf{x}, d)} \right)^{\kappa(\mathbf{y}|\mathbf{x}, c) + \kappa(\mathbf{y}|\mathbf{x}, d)} \cdot \kappa(\mathbf{y}|\mathbf{x}, c)^{\kappa(\mathbf{y}|\mathbf{x}, c)} \cdot \kappa(\mathbf{y}|\mathbf{x}, d)^{\kappa(\mathbf{y}|\mathbf{x}, d)}$$
$$= \left(\frac{7}{3} \right)^3 \cdot 2^2 \cdot 1^1 = 50.(814).$$

Notice that the bound is only slightly larger than the exact one given in Example 97. Moreover, although Theorem 101 itself only provides a numerical bound, the method of Lagrange multipliers which was used to establish Corollary 101, suggests how one can narrow down the search for the optimizing choice of coefficients by considering only the pairs

(n_c, n_d) with integer coordinates which are closest to the point with coordinates $x_u = \frac{|I(\mathbf{y}|\mathbf{x})|+m-|I(y)|}{\sum_{y\in I(\mathbf{y}|\mathbf{x})}\kappa(\mathbf{y}|\mathbf{x},y)} \cdot \kappa(\mathbf{y}|\mathbf{x}, u)$ in "every direction". In our specific example, this point is $(\frac{7}{3}\cdot 2, \frac{7}{3}\cdot 1) = (4\frac{2}{3}, 2\frac{1}{3})$ and so the only potential candidates are $(4, 3)$ and $(5, 2)$. We saw in 97 that the point $(5, 2)$ is the winner. Of course, this "narrowing down" procedure is particularly useful for the cases when $|I(\mathbf{y}|\mathbf{x})|$ is a large number.

12. What can be said when the last elementary step is mutation?

Although not nearly as much can be said when the last elementary step is mutation, the following result is a rather general "anti-communism" theorem. It should be noted that a much stronger and more informative result which depends on the assumption that crossover is "pure" in the sense of [9] (meaning that identical pair of parents produce a pair of the same identical children) shall be established in a sequel paper.

Theorem 103. *Let $\{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ and $\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ denote Markov transition matrices on a finite set $\mathcal{X}$. Suppose $\{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ is non-annihilating in the sense of Definition* 82. *Also let $\{\{m^{\delta}_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} | 0 < \delta < 1\}$ denote an indexed family of Markov transition matrices such that for every $\varepsilon > 0$ there exists $r > 0$ such that for all $\delta < r$ we have $\|\{m^{\delta}_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} - I\| < \varepsilon$ for some norm on the finite-dimensional vector space of $|\mathcal{X}| \times |\mathcal{X}|$ matrices.*[13] *Suppose also that for all $\delta > 0$ with $\delta < 1$ the composed Markov chain $\mathcal{M}(\delta) = \{m^{\delta}_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ is irreducible. Let $\rhd$ denote the relation associated with the Markov transition matrix $\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ (see Definition* 79). *Finally, let π_δ denote the unique stationary distribution of the Markov chain $\mathcal{M}(\delta)$. Then, for all small enough δ, either there exists a state $\mathbf{z} \in \mathcal{X}$ such that $\pi_\delta(\mathbf{z}) < \frac{1}{|\mathcal{X}|}$ or, whenever $\mathbf{x} \rhd \mathbf{y}$, we also have $\pi_\delta(\mathbf{x}) > \pi_\delta(\mathbf{y})$. In particular, as long as $\rhd \neq \emptyset$, the stationary distribution of the Markov chain determined by the transition matrix $\mathcal{M}(\delta)$ is never uniform for all sufficiently small "mutation rates" δ.*

Proof. Denote by $\Lambda = \{\{\lambda_{\mathbf{z}}\}_{\mathbf{z}\in\mathcal{X}} | \sum_{\mathbf{z}\in\mathcal{X}}\lambda_{\mathbf{z}} = 1\lambda_{\mathbf{z}} \geqslant 0\}$ the probability simplex and let $\Lambda_{1/|\mathcal{X}|} = \{\{\lambda_{\mathbf{z}}\}_{\mathbf{z}\in\mathcal{X}} | \sum_{\mathbf{z}\in\mathcal{X}}\lambda_{\mathbf{z}} = 1, \lambda_{\mathbf{z}} \geqslant \frac{1}{2|\mathcal{X}|}\}$. For any given $\mathbf{x} \rhd \mathbf{y} \in \mathcal{X}$ consider a function $f_{\mathbf{x},\mathbf{y}} : \Lambda_{1/|\mathcal{X}|} \to \mathbb{R}$ which sends a given $\lambda \in \Lambda_{1/|\mathcal{X}|}$ to the number

$$\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}(\lambda)(\mathbf{x}) - \{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}(\lambda)(\mathbf{y}) > 0$$

thanks to Proposition 84. From basic point-set topology we know that the set $\Lambda_{1/|\mathcal{X}|}$ is a compact topological space (it is a closed and bounded subset of $\mathbb{R}^{|\mathcal{X}|}$ with $|\mathcal{X}| < \infty$) and, moreover, the function $f_{\mathbf{x},\mathbf{y}}$ is continuous (it is a restriction of a linear map). It follows then that the function $f_{\mathbf{x},\mathbf{y}}$ achieves a minimum, $\min(f_{\mathbf{x},\mathbf{y}})$, on $\Lambda_{1/|\mathcal{X}|}$. Thanks to Proposition 84 this minimum must be a positive number since the matrix $\{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ is non-annihilating and every $\lambda \in \Lambda_{\frac{1}{|\mathcal{X}|}}$ has the property that $\lambda(\mathbf{z}) \geqslant \frac{1}{2|\mathcal{X}|} > 0$ for every $\mathbf{z} \in \mathcal{X}$. We now conclude that

$$\alpha = \min\{\min\{f_{\mathbf{x},\mathbf{y}}(\lambda) | \lambda \in \Lambda_{\frac{1}{|\mathcal{X}|}}\} | \mathbf{x} \rhd \mathbf{y} \in \mathcal{X}\} > 0.$$

Now choose $r > 0$ small enough so that whenever $0 < \delta < r$ we have

$$\|\{m^{\delta}_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} - I\|_{\mathrm{op}} < \min\left\{\frac{\alpha}{3}, \frac{1}{3|\mathcal{X}|}\right\}.$$

Choose any δ satisfying $0 < \delta < r$. Now there are exactly two mutually exclusive and exhaustive cases:

Case 1: $\exists n \in \mathbb{N}$ such that $\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \mathcal{M}(\delta)^n(\Lambda) \subseteq \Lambda_{1/|\mathcal{X}|}$.

In this case, let $\gamma_\delta = \{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \mathcal{M}(\delta)^n(\pi_\delta)$. Since π_δ is the stationary distribution of $\mathcal{M}(\delta)$ (see the statement of the theorem), it is also the stationary distribution of $\mathcal{M}(\delta)^{n+1}$ and it follows that

$$\begin{aligned}\{m^{\delta}_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}(\gamma_\delta) &= \{m^{\delta}_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot (\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \mathcal{M}(\delta)^n(\pi_\delta)) \\ &= (\{m^{\delta}_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}) \cdot \mathcal{M}(\delta)^n(\pi_\delta) = \mathcal{M}(\delta)^{n+1}(\pi_\delta) = \pi_\delta\end{aligned}$$

[13] It is a fact that all the norms on finite-dimensional vector spaces are equivalent. It is then irrelevant which norm we consider. For practical applications it is convenient to use $\|\{a_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}\|_{\max} = \max\{|a_{\mathbf{xy}}| | \mathbf{x}, \mathbf{y} \in \mathcal{X}\}$. For the purpose of proving the theorem it seems most convenient to use the operator norm defined as $\|\{a_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}\|_{\mathrm{op}} = \sup\{\|\{a_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}(\vec{v})\| | \|\vec{v}\| = 1\}$ where $\|(v_1, v_2, \ldots, v_{|\mathcal{X}|})\| = \sum_{i=1}^{|\mathcal{X}|}|v_i|$.

and so

$$\|\gamma_\delta - \pi_\delta\| = \|(I - \{m^\delta_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}})(\gamma_\delta)\| \leqslant \|I - \{m^\delta_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}\|_{\text{op}} < \frac{\alpha}{3}.$$

In particular, $\forall \mathbf{x} \rhd \mathbf{y} \in \mathcal{X}$ we have

$$|\gamma_\delta(\mathbf{x}) - \pi_\delta(\mathbf{x})| < \frac{\alpha}{3} \quad \text{and} \quad |\gamma_\delta(\mathbf{y}) - \pi_\delta(\mathbf{y})| < \frac{\alpha}{3}$$

so that

$$\pi_\delta(\mathbf{x}) > \gamma_\delta(\mathbf{x}) - \frac{\alpha}{3} \quad \text{and} \quad \pi_\delta(\mathbf{y}) < \gamma_\delta(\mathbf{y}) + \frac{\alpha}{3}$$

and, finally,

$$\pi_\delta(\mathbf{x}) - \pi_\delta(\mathbf{y}) > \gamma_\delta(\mathbf{x}) - \frac{\alpha}{3} - \left(\gamma_\delta(\mathbf{y}) + \frac{\alpha}{3}\right) = \gamma_\delta(\mathbf{x}) - \gamma_\delta(\mathbf{y}) - \frac{2\alpha}{3} > \frac{\alpha}{3} > 0$$

thanks to the choice of α, and it follows, in this case, that $\pi_\delta(\mathbf{x}) > \pi_\delta(\mathbf{y})$.

Case 2: $\forall n \in \mathbb{N}$ we have $\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \mathcal{M}(\delta)^n(\Lambda) \nsubseteq \Lambda_{1/|\mathcal{X}|}$.

In this case, first we claim that for every $n \in \mathbb{N}$ there exists a distribution $\gamma \in \mathcal{M}(\delta)^{n+1}(\Lambda)$ such that $\gamma(\mathbf{z}) < \frac{5}{6|\mathcal{X}|}$ for some $\mathbf{z} \in \mathcal{X}$. Indeed, the assumption of case 2 says that there exists a distribution $\lambda \in \{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \mathcal{M}(\delta)^n(\Lambda)$ such that $\lambda(\mathbf{z}) < \frac{1}{2\mathcal{X}}$. But then $\gamma = \{m^\delta_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}(\lambda)$ is the distribution with the desired property. Indeed, since $\lambda \in \{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \mathcal{M}(\delta)^n(\Lambda)$, it follows that $\lambda = \{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \mathcal{M}(\delta)^n(\eta)$ for some distribution $\eta \in \Lambda$. But then

$$\begin{aligned}\gamma &= \{m^\delta_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot (\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \mathcal{M}(\delta)^n(\eta)) = (\{m^\delta_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \cdot \{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}) \cdot \mathcal{M}(\delta)^n(\eta)\\ &= \mathcal{M}(\delta)^{n+1}(\eta) \in \mathcal{M}(\delta)^{n+1}(\Lambda).\end{aligned}$$

Moreover, since $\delta < r$ we have

$$|\gamma(\mathbf{z}) - \lambda(\mathbf{z})| \leqslant \|\gamma - \lambda\| = \|\{m^\delta_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}(\lambda) - \lambda\| = \|(\{m^\delta_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} - I)(\lambda)\| \leqslant \|(\{m^\delta_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}} - I)\|_{\text{op}} < \frac{1}{3|\mathcal{X}|}.$$

But then we also have $\gamma(\mathbf{z}) \leqslant \lambda(\mathbf{z}) + \frac{1}{3|\mathcal{X}|} < \frac{1}{2|\mathcal{X}|} + \frac{1}{3|\mathcal{X}|} < \frac{5}{6|\mathcal{X}|}$ as desired. So we deduce every one of the sets $\mathcal{M}(\delta)^{n+1}(\Lambda)$ contains a point γ_{n+1} with $\gamma_{n+1}(\mathbf{z}) < \frac{5}{6|\mathcal{X}|}$ for some $\mathbf{z} \in \mathcal{X}$. It is well-known from Markov chain theory that the sequence of convex compact sets $\{\mathcal{M}(\delta)^{n+1}(\Lambda)\}_{n=1}^\infty$ is nested ($\mathcal{M}(\delta)^{n+1}(\Lambda) \supseteq \mathcal{M}(\delta)^{n+2}(\Lambda)$) and $\bigcap_{n=1}^\infty \mathcal{M}(\delta)^{n+1}(\Lambda) = \{\pi_\delta\}$ where π_δ is the unique stationary distribution of the Markov chain determined by the matrix $\mathcal{M}(\delta)$. Also, all the elements of the sequence $\{\gamma_{n+1}\}_{n=1}^\infty$ inside of the compact set Λ, and, hence, the sequence $\{\gamma_{n+1}\}_{n=1}^\infty$ has a convergent subsequence $\{\gamma_{(n+1)_k}\}_{k=1}^\infty$. But then $\gamma_{(n+1)_k} \to \pi_\delta$ as $k \to \infty$ (since the limit point must lie inside of every one of the compact sets $\mathcal{M}(\delta)^{n+1}(\Lambda)$ and there intersection consists of a single point π_δ). Moreover, notice that since $\mathcal{X}$ is a finite set while $\{\gamma_{(n+1)_k}\}_{k=1}^\infty$ is an infinite sequence, according to the "pigeonhole principle" it follows that $\exists \mathbf{z} \in \mathcal{X}$ such that infinitely many elements of the subsequence $\{\gamma_{(n+1)_k}\}_{k=1}^\infty$ have the property that $\{\gamma_{(n+1)_k}\}_{k=1}^\infty(\mathbf{z}) < \frac{5}{6|\mathcal{X}|}$. In other words, $\exists \mathbf{z} \in \mathcal{X}$ for which we can extract a subsequence $\{\gamma_{(n+1)_{k_s}}\}_{s=1}^\infty$ of the convergent sequence $\{\gamma_{(n+1)_k}\}_{k=1}^\infty$ with the property that $\gamma_{(n+1)_{k_s}}(\mathbf{z}) < \frac{5}{6|\mathcal{X}|}$. In particular, $\gamma_{(n+1)_{k_s}}(\mathbf{z}) \to \pi_\delta(\mathbf{z})$ as $s \to \infty$ and it follows that $\pi_\delta(\mathbf{z}) \leqslant \frac{5}{6|\mathcal{X}|} < \frac{1}{|\mathcal{X}|}$ which is what we were after. □

When applying Theorem 103 we have in mind that $\{q_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ is the Markov transition matrix corresponding to recombination (i.e. a sub-algorithm determined by a single elementary step of type 2: see Definitions 10 and 7), $\{p_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ is the Markov transition matrix corresponding to selection (i.e. a sub-algorithm determined by a single elementary step of type 1: see Definition 4) and $\{m^\delta_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}$ is the Markov transition matrix corresponding to mutation with some "rate" δ. For the purpose of the current section, thanks to the generality of Theorem 103, it is sufficient to assume only that $m^\delta_{\mathbf{xy}} > 0 \forall \mathbf{x}, \mathbf{y}$ and that $\max(\{m^\delta_{\mathbf{xy}} | \mathbf{x} \neq \mathbf{y} \in \mathcal{X}\}) \to 0$ as $\delta \to 0$. The following proposition tells us when mutation determined by the reproduction 1-tuple $(\Omega, \mathcal{M}, p)$ satisfies conditions of Theorem 103:

Definition 104. An ergodic family of mutations is an indexed family of ergodic mutation 1-tuples (see Definition 67) of the form $\{(\Omega, \mathcal{M}, p_\delta)\}_{\delta \in (0,1)}$ where $p_\delta(1_\Omega) \geqslant 1 - \delta$.

Proposition 105. *Suppose* $\{(\Omega, \mathcal{M}, p_\delta)\}_{\delta \in (0,1)}$ *is an ergodic family of mutations as in Definition* 104. *Then* $\forall \delta \in (0, 1)$ *the Markov transition matrix* $\{m^\delta_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y} \in \mathcal{X}}$ *associated to the sub-algorithm determined by the mutation* 1-*tuple* $(\Omega, \mathcal{M}, p_\delta)$ *has the property that* $\|I - \{m^\delta_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y} \in \mathcal{X}}\| \to 0$ *as* $\delta \to 0$.

Proof. Notice that $\|I - \{m^\delta_{\mathbf{x},\mathbf{y}}\}_{\mathbf{x},\mathbf{y} \in \mathcal{X}}\| = \max\{m^\delta_{\mathbf{xy}} | \mathbf{x} \neq \mathbf{y}\}$ and so it suffices to show that for every $\mathbf{x} \neq \mathbf{y}$ we have $m^\delta_{\mathbf{x},\mathbf{y}} \to 0$ as $\delta \to 0$ (since the state space is finite). Notice also that $\forall \mathbf{x} \neq \mathbf{y}$ we have $0 < m^\delta_{\mathbf{x},\mathbf{y}} < 1 - m^\delta_{\mathbf{x},\mathbf{x}}$. It now suffices to show only that $\forall \mathbf{x} \in \mathcal{X} = \Omega^m$ we have $1 - m^\delta_{\mathbf{x},\mathbf{x}} \to 0$ as $\delta \to 0$, or, equivalently, that $\forall \mathbf{x} \in \mathcal{X} = \Omega^m$ we have $m^\delta_{\mathbf{x},\mathbf{x}} \to 1$ as $\delta \to 0$. If we write $\mathbf{x} = (x_1, x_2, \ldots, x_m) \in \Omega^m$ then, since $\mathbf{1}_\Omega(x_i) = x_i$ we see that $1 \geqslant m^\delta_{\mathbf{x},\mathbf{x}} \geqslant (p_\delta(1_\Omega))^m \geqslant (1 - \delta)^m \to 1$ as $\delta \to 0$ and the desired conclusion follows. □

Combining Theorem 103 with the conclusion of Example 99 (saying that $\rhd \neq \emptyset$ for fitness-proportional selection) we deduce the following:

Corollary 106. *Suppose for every* $0 < \delta < 1$ *we are given an evolutionary algorithm* $\mathcal{A}_\delta$ *determined by the cycle* s_1, s_2, s^δ_3 *where* s_1 *is any elementary step* (*but usually it is an elementary step of type* 2), s_2 *is the elementary step of type* 1 (*fitness-proportional selection as described in Definition* 4) *and* s^δ_3 *is an elementary step of type* 2 *determined by an ergodic mutation* 1-*tuple chosen from an ergodic family of mutations* (*see Definition* 104). *Then the Markov chain determined by the algorithm* $\mathcal{A}_\delta$ *with state space* $\mathcal{X} = \Omega^m$ *is irreducible and, for all small enough* δ, *the unique stationary distribution of this Markov chain is not uniform.*

Corollary 106 tells us, in particular, that the stationary distribution of the Markov chain associated to an algorithm $\mathcal{A}$ with the second elementary step being of type 1 (selection) is never uniform, even when the fitness function is flat. It is still reasonable to conjecture though, that in case of flat-fitness selection, under certain symmetry assumptions on recombination and mutation, everyone of the individuals in a given population is equally likely to occur "in the long run" in the sense of Definition 24. Results of this nature (and even stronger) shall be established in the upcoming paper.

13. Conclusions

In the current paper we applied the methods developed in [4] to obtain a schema-based version of Geiringer's theorem for nonlinear GP with homologous crossover. The result enables us to calculate exactly the limiting distribution of the Markov chain associated with the evolution of a finite (fixed size) population under the action of repeated crossover, or the action of the mixture of crossover and mutation. This is an extension of the results for fixed and variable-length strings given in [4] for finite populations.

The main result established in [4] applies only in the absence of selection and only when crossover and mutation are bijective (which is often, but not always the case). In the current paper we established a property of the stationary distribution of the Markov chain for a rather wide class of EAs. More specifically, we introduced a pre-order relation on the state space of a Markov chain which allows us to establish rather general inequalities concerning the stationary distribution of the Markov chain determined by an EA. This pre-order relation depends primarily on selection and not on the other stages determining an EA. In Section 11 this partial order is completely classified for the case of fitness-proportional selection in Section 11. More results on this issue, as well as some connection between the infinite and the finite population Geiringer theorems will appear in a forthcoming paper.

Acknowledgements

The authors are grateful to Professor Riccardo Poli and to the other anonymous reviewers for helpful suggestions on improvements and corrections. Boris Mitavskiy is very grateful to Professor Andreas Blass for his professional guidance and for the numerous educative advisor meetings during his (author's) years in graduate school.

References

[1] S. Coffey, An applied probabilist's guide to genetic algorithms, Master's Thesis, The University of Dublin, 1999.

[2] R.A. Fisher, The Genetical Theory of Natural Selection, Clarendon Press, Oxford, 1930.

[3] H. Geiringer, On the probability of linkage in mendelian heredity, Annu. Math. Statist. 15 (1944) 25–57.

[4] B. Mitavskiy, J. Rowe, An extension of Geiringer theorem for a wide class of evolutionary algorithms, Evolutionary Computation 14 (2006) 87–108.

[5] R. Poli, Hyperschema theory for gp with one-point crossover, building blocks, and some new results in ga theory, in: R. Poli, W. Banzhaf et al. (Eds.), Genetic Programming, Proc. EuroGP'2000, Springer, Berlin, 2000, pp. 163–180.

[6] R. Poli, W. Langdon, On the search properties of different crossover operators in genetic programming, in: Proc. Third Annu. Genetic Programming Conf., 1998, pp. 293–301.

[7] R. Poli, C. Stephens, A. Wright, J. Rowe, A schema-theory-based extension of Geiringer's theorem for linear gp and variable-length gas under homologous crossover, in: K. De Jong, R. Poli, J.E. Rowe (Eds.), Foundations of Genetic Algorithms, Vol. 7, 2002, pp. 45–62.

[8] Y. Rabani, Y. Rabinovich, A. Sinclair, A computational view of population genetics, Annu. ACM Symp. Theory of Comput., 1995, pp. 83–92.

[9] N. Radcliffe, The algebra of genetic algorithms, Annu. Math. Artif. Intell. 10 (1994) 339–384.

[10] L. Schmitt, Theory of genetic algorithms, Theoret. Comput. Sci. 259 (2001) 1–61.

[11] L. Schmitt, Theory of genetic algorithms ii: models for genetic operators over the string-tensor representation of populations and convergence to global optima for arbitrary fitness function under scaling, Theoret. Comput. Sci. 310 (2004) 181–231.

[12] W. Spears, The equilibrium and transient behavior of mutation and recombination, in: W. Martin, W. Spears (Eds.), Foundations of Genetic Algorithms, Vol. 6, 2000, pp. 241–260.

[13] M. Vose, The Simple Genetic Algorithm: Foundations and Theory, MIT Press, 1999.

[14] S. Wright, Evolution in mendelian populations, Genetics 16 (1931) 97–159.