An Extension of Geiringer's Theorem for a Wide Class of Evolutionary Search Algorithms.

Boris Mitavskiy

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

B.S.Mitavskiy@cs.bham.ac.uk

Jonathan Rowe

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

J.E.Rowe@cs.bham.ac.uk

Abstract

The frequency with which various elements of the search space of a given evolutionary algorithm are sampled is affected by the family of recombination (reproduction) operators. The original Geiringer theorem tells us the limiting frequency of occurrence of a given individual under repeated application of crossover alone for the classical genetic algorithm. Recently, Geiringer's theorem has been generalized to include the case of linear GP with homologous crossover (which can also be thought of as a variable length GA). In the current paper we prove a general theorem which tells us that under rather mild conditions on a given evolutionary algorithm, call it $\mathcal{A}$, the stationary distribution of a certain Markov chain of populations in the absence of selection is unique and uniform. This theorem not only implies the already existing versions of Geiringer's theorem, but also provides a recipe of how to obtain similar facts for a rather wide class of evolutionary algorithms. The techniques which are used to prove this theorem involve a classical fact about random walks on a group and may allow us to compute and/or estimate the eigenvalues of the corresponding Markov transition matrix which is directly related to the rate of convergence towards the unique limiting distribution.

Keywords

genetic algorithms, crossover, schemata, Geiringer theorem, Markov process, stationary distribution, random walk on a group, mutation

1 Introduction

Geiringer's classical theorem (see (Geiringer, 1944)) is an important part of GA theory. It has been cited in a number of papers: see, for instance, (Poli et al., 2002), (Vose, 1999), (Spears, 2000) and (Rabani et al., 1995). The theorem has also been rediscovered and improved in various ways (see, for example, (Poli et al., 2002) and (Vose, 1999)). It deals with the limit of the sequence of probability distributions obtained by repetitive applications of the crossover operator. To state the theorem formally, let $\Omega = \prod_{i=1}^n A_i$ denote the search space of a given genetic algorithm. (Intuitively, n is the number of loci and A_i is the set of alleles corresponding to the i^{th} gene.) Denote by Λ the collection of all probability distributions on Ω . Now fix a probability distribution $\lambda \in \Lambda$ and consider the sequence of probability distributions $\lambda, \mathcal{C}(\lambda), \mathcal{C}^2(\lambda), \dots$ where $\mathcal{C}(p)(k) = \sum_{i,j} p(i)p(j)r_{(i,j \rightarrow k)}$ and $r_{(i,j \rightarrow k)}$ denotes the probability of obtaining the individual k from the parents i and j after crossover. Here crossover can be thought of as an operator which takes a pair of elements of the search space (the parents) and produces another element of the search space (the child) by mingling the

alleles of the parents. This will be discussed in more detail later in the paper. Denote by λ_i the marginal distribution of λ on A_i . The classical Geiringer theorem says that $\lim_{t \rightarrow \infty} C^t(\lambda) \rightarrow \prod_{i=1}^n \lambda_i$ (meaning that the frequency of occurrence of an individual $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \Omega$ under the limiting distribution is just the product of the frequencies of x_i under the distributions λ_i). In (Poli et al., 2002) this theorem has been generalized to cover the case of variable-length GAs (or linear GP) under homologous 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 theorem 9 and section 4.2.1 of (Poli et al., 2002)). The proofs of these results provided in (Poli et al., 2002) are modifications of Geiringer's original arguments. In the current paper a new version of this result will be proved for *finite* population models in the absence of the sampling stage, addressing the limiting distribution of the associated Markov chain. It turns out that in the known classical cases the formula for the limiting distribution coincides with the one for the infinite population. To see what the formula for the finite case is saying, recall that $\Omega = \prod_{i=1}^n A_i$ denotes the search space of a given genetic algorithm. 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 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 i^{th} column of P are members of A_i . Continuing with the notation used in (Poli et al., 2002), denote by $\Phi(h, P, i)$ where $h \in A_i$, the proportion of rows of P for which $a_{ji} = h$. In other words, let $R_h(i) = \{j \mid (a_{j1}, a_{j2}, \dots, a_{jn}) \text{ is a row of } P \text{ and } a_{ji} = h\}$. Now simply let $\Phi(h, P, i) = \frac{|R_h(i)|}{m}$. The classical Geiringer theorem (see (Geiringer, 1944) or (Poli et al., 2002) for modern notation) says that if one starts with a population P of individuals and runs a genetic algorithm in the absence of selection and mutation (2 parent $\rightarrow$ 2 children crossover as described in example 5.1 being the only operator involved) then, in the "long run", the frequency of occurrence of the individual (row) $(h_1, h_2, \dots, h_n)$ before time t , call it $\Phi(\{(h_1, h_2, \dots, h_n)\}, P, t)$ (a formal definition will be given in section 6; see definition 6.2), approaches independence:

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

Thereby, the finite population version of 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. It turns out that in case of a finite population a general version of Geiringer's theorem follows rather easily from the basic facts about Markov chains and random walks on groups. This is mainly a matter of formulating the statement of the theorem in a slightly different language. In particular, it extends some of the already existing versions of Geiringer's theorem to the case of finite populations.¹ It is worth mentioning, however, that theorem 10.7 of (Vose, 1999)

¹The assumptions of the orbit-description lemmas in the current paper are slightly stronger than the notion of allele-transferring recombination given in (Poli et al., 2002). Moreover, the proof of Geiringer's theo-

provides stronger results for the case of binary GAs than theorem 5.5 of the current paper. The primary advantage of theorem 5.5 is that it applies to a wider class of EAs. In a sequel paper we shall use the generalized Geiringer theorem together with the recipe provided in the current paper to derive a schema-based version of Geiringer's theorem for nonlinear GP with homologous crossover which is based on Poli's schemata (see, for instance, (Poli, 2000) for the notion of Poli's schemata). The extended Geiringer theorem also applies to many algorithms in the presence of mutation (see section 7 for a detailed exposition). Moreover, the elegant techniques developed by Diaconis and others (see (Rosenthal, 1995) for a detailed exposition) may allow one to estimate the rate of convergence to the limiting distribution in some cases. This topic is left for future investigation.

It must be pointed out, however, that in chapter 10 of (Vose, 1999) a rather deeper investigation of the properties of crossover and mutation operators for the case of a classical genetic algorithm (GA) has been carried out. Strong versions of the classical Geiringer theorem have been established (see theorems 10.7 and 10.8 of (Vose, 1999)). Furthermore, in (Rabani et al., 1995) the rate of convergence towards the limiting distribution for the case of classical GA's has been estimated in a very clever manner. The primary virtue of Geiringer's theorem presented in the current paper is that it applies to a wide class of evolutionary algorithms and to finite populations.

In the next section we introduce some algebraic notions and facts which allow us to extend Geiringer's theorem. Incidentally, these algebraic concepts have already been exploited in (Rowe et al., 2002) and in (Rowe et al., 2004) to study representations of evolutionary algorithms. In sections 3 and 4 a mathematical framework which rigorously describes a wide class of evolutionary algorithms is introduced. A similar description has already appeared in (Mitavskiy, 2004a) and in (Mitavskiy, 2003). The generalized Geiringer theorem (theorem 5.5) itself appears in section 5. Section 6 is devoted to the demonstration of how the classical and already existing versions of Geiringer's theorem follow from theorem 5.5. The generalized Geiringer theorem applies in the presence of mutation as well. This is the subject of section 7.

2 Mathematics Behind the Generalized Geiringer Theorem

The generalized Geiringer theorem presented in the current paper (theorem 5.5) is based on a few basic facts about group actions. To make the paper self-contained, all of the necessary material is presented below. For a detailed introduction to the theory of groups and group actions see, for instance, (Dummit and Foote, 1991).

Definition 2.1. A group is an ordered pair $(G, *)$ where G is a set and $*$ is a binary operation on G (i.e. a function $*$: $G \times G \rightarrow G$; it is commonly accepted to write $g * h$ instead of $*(g, h)$) which enjoys the following properties:

1. $\forall g, h, k \in G$ we have $g * (h * k) = (g * h) * k$ (this property is known as *associativity*).
2. $\exists$ a special element $e \in G$ such that $\forall g \in G$ we have $g * e = e * g = g$.
3. $\forall g \in G$ there exists a special element, called $g^{-1} \in G$, such that $g * g^{-1} = g^{-1} * g = e$.

Throughout the remainder of this section we shall write gh instead of $g * h$.

Example 2.1. A typical example of a group is the so-called *symmetric group* of all the permutations of a given set $\mathcal{X}$. This group is denoted by $\mathcal{S}_{\mathcal{X}}$ and it will play a significant

rem given in (Poli et al., 2002) is interesting in its own right since it attacks the problem from a completely different point of view.

role in generalizing Geiringer's theorem. It is easy to see that $\mathcal{S}_{\mathcal{X}}$ is a group under function composition.

Definition 2.2. Let G denote a finite group and $H \subseteq G$. We say that H generates G , or that G is generated by H if every element of G is a product of elements in H . In other words $\forall g \in G, \exists h_1, h_2, \dots, h_n \in H$ such that $g = h_1 h_2 \dots h_n$.

It is well known (see, for instance, (Dummit and Foote, 1991)) that every permutation (i.e., a bijection whose domain and codomain is the same set) on a finite set $\mathcal{X}$ is a composition of transpositions. A transposition is a permutation which swaps a couple of elements and leaves everything else the same. Formally, a transposition is defined as follows:

$$\sigma_{ij}(k) = \begin{cases} i & \text{if } k = j \\ j & \text{if } k = i \\ k & \text{otherwise} \end{cases}.$$

According to definition 2.2 this means that the set $H = \{\sigma_{ij} \mid i, j \in \mathcal{X}\}$ of all transpositions generates the group $\mathcal{S}_{\mathcal{X}}$ (see example 2.1).

Definition 2.3. An action of a group G on a set $\mathcal{X}$ is a function $\cdot : G \times \mathcal{X} \rightarrow \mathcal{X}$ (again, we write $g \cdot \mathbf{x}$ instead of $\cdot(g, \mathbf{x})$) satisfying the following properties:

1. $\forall h, g \in G$ and $\mathbf{x} \in \mathcal{X}$ we have $(hg) \cdot \mathbf{x} = h \cdot (g \cdot \mathbf{x})$.
2. $\forall \mathbf{x} \in \mathcal{X}$ we have $\mathbf{e} \cdot \mathbf{x} = \mathbf{x}$.

When $\cdot$ is an action of a group G on the set $\mathcal{X}$, we also say that G acts on $\mathcal{X}$. We say that G acts *transitively* on $\mathcal{X}$ if for all $\mathbf{x}$ and $\mathbf{y}$ in $\mathbf{X}$ there exists $g \in G$ such that $g \cdot \mathbf{x} = \mathbf{y}$.

The group action defined in the example below will be of special importance in the derivation of the generalized Geiringer theorem (theorem 5.5).

Example 2.2. Fix a set $\mathcal{X}$. Let G be a subgroup² of the symmetric group $\mathcal{S}_{\mathcal{X}}$. Then G acts on $\mathcal{X}$ as follows: $\pi \cdot x = \pi(x)$. It is easily seen that $\cdot$ is a group action in the sense of definition 2.3.

We now proceed to define a *random walk associated with a group action*:

Definition 2.4. Let $\mathcal{X}$ be a finite set and let G be a group acting on $\mathcal{X}$. Let μ denote a probability distribution on G . We define a random walk associated with the action of G on $\mathcal{X}$ to be the Markov process with transition probabilities given as follows: for every $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ we let $p_{\mathbf{x}\mathbf{y}} = \mu(\{g \mid g \cdot \mathbf{x} = \mathbf{y}\})$.

In the heart of the proof of the generalized Geiringer theorem lies the following fact (a variation of theorem 2.1 appears in (Rosenthal, 1995)):

Theorem 2.1. Let $\mathcal{X}$ be a set, let G be a finite group acting transitively on the set $\mathcal{X}$ and let $H \subseteq G$ be a subset of G which satisfies the following two properties:

1. H generates G .
2. $\mathbf{e} \in H$ where $\mathbf{e}$ denotes the identity of the group G .

Let μ denote a probability distribution on G which is concentrated on H (meaning that $\mu(g) > 0 \iff g \in H$). Then the random walk associated with the action of G on $\mathcal{X}$ has a unique stationary distribution and this stationary distribution is the uniform distribution.

²A subgroup of a group is a subset of elements that is a group in its own right.

Proof. Denote by $p_{\mathbf{x}\mathbf{y}}^n$ the probability that a state $\mathbf{y} \in \mathcal{X}$ is reached from the state $\mathbf{x}$ in exactly n steps. Notice that, since the group G is finite, and since H generates G , there exists a large enough n so that for every g in G we have $p_{\mathbf{x}(g \cdot \mathbf{x})}^n > 0$. Indeed, since H generates G , for every g in G there exists a positive integer n_g and elements $h_1^g, h_2^g, \dots, h_{n_g}^g$ in H such that $g = h_1^g h_2^g \dots h_{n_g}^g$. Now simply let $n = \max\{n_g \mid g \in G\}$

$$\begin{aligned}
 \text{and notice that } p_{\mathbf{x}(g \cdot \mathbf{x})}^n &= p_{\mathbf{x}((h_1^g h_2^g \dots h_{n_g}^g \underbrace{\mathbf{e} \mathbf{e} \dots \mathbf{e}}_{n - n_g \text{ times}}) \cdot \mathbf{x})}^n \stackrel{\text{by definition of a group action}}{=} \\
 &= p_{\mathbf{x}, (h_1^g \cdot (h_2^g \cdot (\dots (h_{n_g}^g \cdot (\mathbf{e} \cdot (\mathbf{e} \cdot (\dots (\mathbf{e} \cdot \mathbf{x}) \dots)) \dots)))}^{\substack{n - n_g \text{ times}}})}^n \geq \\
 &\geq \underbrace{p_{\mathbf{x}(\mathbf{e} \cdot \mathbf{x})} p_{(\mathbf{e} \cdot \mathbf{x})(\mathbf{e} \cdot (\mathbf{e} \cdot \mathbf{x}))} \dots p_{\mathbf{e} \cdot (\mathbf{e} \cdot (\dots \mathbf{e} \cdot \mathbf{x}) \dots)}^{\substack{n - n_g - 1 \text{ times}}} \underbrace{p_{\mathbf{e} \cdot (\mathbf{e} \cdot (\dots \mathbf{e} \cdot \mathbf{x}) \dots)}^{\substack{n - n_g \text{ times}}}}_{n - n_g \text{ times}} \times \\
 &\times p_{\mathbf{x}(h_{n_g}^g \cdot \mathbf{x})} p_{(h_{n_g}^g \cdot \mathbf{x})(h_{n_g-1}^g \cdot (h_{n_g}^g \cdot \mathbf{x}))} \dots p_{(h_{n_2}^g h_{n_3}^g \dots h_{n_g}^g) \cdot \mathbf{x}} p_{h_{n_1}^g \cdot ((h_{n_2}^g h_{n_3}^g \dots h_{n_g}^g) \cdot \mathbf{x})}^n \geq \\
 &\stackrel{\text{according to definition 2.4}}{\geq} (\mu(\mathbf{e}))^{(n - n_g)} \prod_{i=1}^{n_g} \mu(h_{n_i}^g) \stackrel{\text{by assumption}}{>} 0.
 \end{aligned}$$

This shows that the random walk associated to the action of G on $\mathcal{X}$ is an irreducible Markov chain with a finite state space ($\mathcal{X}$ is finite because G is finite and because G acts transitively on $\mathcal{X}$) and, therefore, it has a unique stationary distribution. To prove the desired assertion, now it suffices to check that the uniform distribution on $\mathcal{X}$ is stationary. This boils down to a direct computation: Denote by π the uniform distribution on $\mathcal{X}$. We then have $\pi(\mathbf{x}) = \frac{1}{|\mathcal{X}|}$. Now, if π is taken to be the initial distribution, then the distribution in the next generation, say ρ , is given as

$$\rho(\mathbf{x}) = \sum_{h \in H} \pi(h^{-1} \cdot \mathbf{x}) \mu(h) = \sum_{h \in H} \frac{1}{|\mathcal{X}|} \mu(h) = \frac{1}{|\mathcal{X}|} \sum_{h \in H} \mu(h) = \frac{1}{|\mathcal{X}|} = \pi(\mathbf{x})$$

($\sum_{h \in H} \mu(h) = 1$ since μ is concentrated on H). This shows that the uniform distribution is, indeed, stationary and completes the proof. $\square$

To extend Geiringer's theorem we need to recall the following basic fact which follows immediately from the definition of a group action:

Proposition 2.2. *Let G be a group acting on a set $\mathcal{X}$. Consider the relation $\sim$ on $\mathcal{X}$ defined as $\mathbf{x} \sim \mathbf{y}$ if and only if there exists an element g in G such that $g \cdot \mathbf{x} = \mathbf{y}$. Then $\sim$ is an equivalence relation on $\mathcal{X}$.*

Proof. $\mathbf{x} = \mathbf{e} \cdot \mathbf{x}$, so $\mathbf{x} \sim \mathbf{x}$ which shows that $\sim$ is reflexive. If $\mathbf{x} \sim \mathbf{y}$ then there exists a group element g in G such that $g \cdot \mathbf{x} = \mathbf{y}$. But then $\mathbf{x} = g^{-1} \cdot \mathbf{y}$ so that $\mathbf{y} \sim \mathbf{x}$. This shows that $\sim$ is symmetric. Finally, if $\mathbf{x} \sim \mathbf{y}$ and $\mathbf{y} \sim \mathbf{z}$ then there exist group elements g and h in G such that $g \cdot \mathbf{x} = \mathbf{y}$ and $h \cdot \mathbf{y} = \mathbf{z}$. Now simply observe that $(hg) \cdot \mathbf{x} = h \cdot (g \cdot \mathbf{x}) = h \cdot \mathbf{y} = \mathbf{z}$. Thereby, $\sim$ is transitive. We now conclude that $\sim$ is an equivalence relation. $\square$

The equivalence classes under $\sim$ are known as *orbits* (see (Dummit and Foote, 1991) for a more detailed treatment).

Definition 2.5. Let G be a group acting on a set $\mathcal{X}$. Fix an element $\mathbf{x} \in \mathcal{X}$. The orbit of x under the action of G , denoted by $\mathcal{O}_{\mathbf{x}}$, is the subset $\mathcal{O}_{\mathbf{x}} = \{\mathbf{y} \mid \exists g \in G \text{ such that } g \cdot \mathbf{x} = \mathbf{y}\} \subseteq \mathcal{X}$.

Notice now that the transitivity assumption in theorem 2.1 is not restrictive at all: if G is any finite group acting on a set $\mathcal{X}$ then G acts transitively on the set $\mathcal{O}_{\mathbf{x}}$ (see definition 2.5). Now simply apply theorem 2.1 to the set $\mathcal{O}_{\mathbf{x}}$ to obtain the following:

Corollary 2.3. Let $\mathcal{X}$ be a set, let G be a finite group acting on the set $\mathcal{X}$ and let $H \subseteq G$ be a subset of G which satisfies the following two properties:

1. H generates G .
2. $\mathbf{e} \in H$ where $\mathbf{e}$ denotes the identity of the group G .

Let μ denote a probability distribution on G which is concentrated on H (meaning that $\mu(g) > 0 \iff g \in H$). Fix an element $\mathbf{x} \in \mathcal{X}$. Then the random walk associated with the action of G on $\mathcal{X}$ which starts initially at $\mathbf{x}$, in the long run visits every element of $\mathcal{O}_{\mathbf{x}}$ equally likely.

Remark 2.1. Observe that in the statement of corollary 2.3, $\mathcal{O}_{\mathbf{x}}$ can be alternatively described as follows: $\mathcal{O}_{\mathbf{x}} = \{\mathbf{y} \mid p_{\mathbf{x}\mathbf{y}}^t > 0 \text{ for some } t \in \mathbb{N}\}$. This can be easily seen from the proofs of theorem 2.1 and corollary 2.3.

The description of the evolutionary computation techniques which appears below will allow us to establish Geiringer-like results in a rather general setting by means of corollary 2.3. A significant part of the framework presented below is also available in (Mitavskiy, 2004a) and in (Mitavskiy, 2003).

3 Notation

Ω is a finite set, called a *search space*. $\mathcal{F}_q$ 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 4 to describe the general evolutionary search algorithm. This approach makes it easy to state Geiringer's theorem.

Definition 3.1. 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 3.1. 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, (Schmitt, 2001) and (Schmitt, 2004)). According to definition 3.1, in the current paper we continue the development of analysis based upon this presentation. The following example illustrates an aspect of the representation which the reader should keep in mind:

Example 3.1. Let $\Omega = \{0, 1\}^3$. Consider the populations

$$\begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

According to definition 3.1 these should be considered *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. In the current paper we shall analyze only those elementary steps which model recombination. Normally, in a practical situation, an elementary step of selection is also involved. Some preliminary results about the algorithms involving selection appear in (Mitavskiy and Rowe, 2005)

In order to describe an elementary step modelling reproduction in a general setting we introduce the following preliminary definitions:

Definition 3.2. Fix an ordered k -tuple of integers $\mathbf{q} = (q_1, q_2, \dots, q_k)$. Let K denote a partition of the set $\{1, 2, \dots, m\}$ for some $m \in \mathbb{N}$. We say that partition K is $\mathbf{q}$ -fit if for every element P of the partition K we can find an i (depending on P) such that P consists of exactly q_i elements. In logical symbols this means that if $K = \{P_1, P_2, \dots, P_l\}$ then K is $\mathbf{q}$ -fit if $\forall j$ with $1 \leq j \leq l \exists i$ with $1 \leq i \leq k$ such that $|P_j| = q_i$. Denote by $\mathcal{E}_{\mathbf{q}}^m$ the family of all $\mathbf{q}$ -fit partitions of $\{1, 2, \dots, m\}$.

Definition 3.3. Let Ω be a set, $\mathcal{F}_{q_1}, \mathcal{F}_{q_2}, \dots, \mathcal{F}_{q_k}$ 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 $p_1, p_2, \dots, p_k$ be probability distributions on $(\mathcal{F}_{q_1})^{q_1}, (\mathcal{F}_{q_2})^{q_2}, \dots, (\mathcal{F}_{q_k})^{q_k}$ respectively. If $(T_1, T_2, \dots, T_{q_i}) \in (\mathcal{F}_{q_i})^{q_i}$ we shall write $p_i(T_1, T_2, \dots, T_{q_i})$ instead of $p_i(\{(T_1, T_2, \dots, T_{q_i})\})$ to denote the probability that the tuple of transformations $(T_1, T_2, \dots, T_{q_i})$ is chosen under the distribution p_i . Let $\mathbf{q} = (q_1, q_2, \dots, q_k)$. Finally, let $\wp_m$ be a probability distribution on the collection $\mathcal{E}_{\mathbf{q}}^m$ of partitions of $\{1, 2, \dots, m\}$ (see definition 3.2 above). We then say that the ordered $2(k+1)$ -tuple $(\Omega, \mathcal{F}_{q_1}, \mathcal{F}_{q_2}, \dots, \mathcal{F}_{q_k}, p_1, p_2, \dots, p_k, \wp_m)$ is a reproduction k -tuple of arity $(q_1, q_2, \dots, q_k)$.

The following definition of reproduction covers many types of q parents $\rightarrow q$ children crossovers. In particular, it covers mutation (1 parent $\rightarrow$ 1 child crossover)

Remark 3.2. Although 2 parents $\rightarrow$ 1 child crossover is often exploited in practice, the results of the current paper apply only in the case when the number of children produced is the same as the number of parents taken as an input during the recombination process. This is why definition 3.4 stated below models only q parents $\rightarrow q$ children crossovers.

Definition 3.5 (see also remark 3.3) will make it possible to combine different reproduction operators in a simple and natural way.

Definition 3.4. An elementary step of type *reproduction* associated with a given reproduction k -tuple $(\Omega, \mathcal{F}_{q_1}, \mathcal{F}_{q_2}, \dots, \mathcal{F}_{q_k}, p_1, p_2, \dots, p_k, \wp_m)$ takes a given population

$$P = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} \text{ with } x_i \in \Omega \text{ as an input.}$$

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 $K = \{(i_1^1, i_2^1, \dots, i_{q_1}^1), (i_1^2, i_2^2, \dots, i_{q_2}^2), \dots, (i_1^j, i_2^j, \dots, i_{q_j}^j) \dots\}$ the corresponding tuples are

$$Q_1 = \begin{pmatrix} x_{i_1^1} \\ x_{i_2^1} \\ \vdots \\ x_{i_{q_1}^1} \end{pmatrix} \quad Q_2 = \begin{pmatrix} x_{i_1^2} \\ x_{i_2^2} \\ \vdots \\ x_{i_{q_2}^2} \end{pmatrix} \quad \dots \quad Q_j = \begin{pmatrix} x_{i_1^j} \\ x_{i_2^j} \\ \vdots \\ x_{i_{q_j}^j} \end{pmatrix} \quad \dots$$

Having selected the partition, replace every one of the selected q_j -tuples $Q_j = \begin{pmatrix} x_{i_1^j} \\ x_{i_2^j} \\ \vdots \\ x_{i_{q_j}^j} \end{pmatrix}$

with the q_j -tuples $Q'_j = \begin{pmatrix} T_1(x_{i_1^j}, x_{i_2^j}, \dots, x_{i_{q_j}^j}) \\ T_2(x_{i_1^j}, x_{i_2^j}, \dots, x_{i_{q_j}^j}) \\ \vdots \\ T_{q_j}(x_{i_1^j}, x_{i_2^j}, \dots, x_{i_{q_j}^j}) \end{pmatrix}$ 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 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 3.5. A cycle is a finite sequence of elementary steps, $\{s_\tau\}_{\tau=1}^j$, which have the same underlying search space and the same arity of input/output.

Remark 3.3. Intuitively, the steps in a cycle are linked together in such a way that the output of step s_i is the input of 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 class of evolutionary heuristic search algorithms about which we can prove Geiringer-like results for finite populations. This class of algorithms is more general than those classes for which Geiringer-like results have been previously established, although at the same time that class does not permit as strong a collection of Geiringer-like results as have been previously established (see, for instance, theorem 10.7 of (Vose, 1999)).

4 How Does Our Heuristic Search Algorithm Work?

Our evolutionary search algorithm works as follows. Fix a *cycle* of elementary steps C . Now start the algorithm with an initial population P . This population 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 , simply apply each of the elementary steps in turn, starting with input P . The output from the first step becomes the input to the second step, and so on. We repeatedly apply C for a finite number of times.

Definition 4.1. A sub-algorithm of a given evolutionary search algorithm defined by a cycle C is simply an evolutionary search algorithm defined by a subsequence of C .

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

5 The Extended Geiringer Theorem

To understand the intuitive meaning of the definitions below, see sections 3 and 4.

Definition 5.1. Given a set Ω and a family of transformations $\mathcal{F}_q$ from Ω^q into Ω , fix a q -tuple of transformations $(T_1, T_2, \dots, T_q) \in (\mathcal{F}_q)^q$. Now consider the transformation

$\langle T_1, T_2, \dots, T_q \rangle : \Omega^q \rightarrow \Omega^q$ sending any given element

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

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

The next definition emphasizes an important property of recombination which many evolutionary search algorithms have and which is sufficient to extend the Geiringer's theorem. To alleviate the level of abstraction we try to motivate the definition by pointing out this important property in case of a classical genetic algorithm. Suppose the search space is $\Omega = \{0, 1\}^5$. For the sake of concreteness, let's assume that the individuals $\mathbf{a} = (0, 1, 0, 1, 0)$ and $\mathbf{b} = (1, 0, 1, 0, 1)$ are selected for crossover. It is possible to replace the pair $(\mathbf{a}, \mathbf{b})$ with the pair $(\mathbf{c}, \mathbf{d})$ where $\mathbf{c} = (0, 0, 0, 1, 1)$ and $\mathbf{d} = (1, 1, 0, 0, 0)$ by applying the masked crossover transformations $L_{\{1,3,4\}} : \Omega^2 \rightarrow \Omega$ sending a given pair $((a, b, c, d, e), (f, g, h, k, l))$ to the individual (a, g, c, d, l) to obtain the first child and $L_{\{2,3,5\}} : \Omega^2 \rightarrow \Omega$ sending $((a, b, c, d, e), (f, g, h, k, l))$ to the individual (f, b, c, k, e) . Traditionally, when using 2 parent $\rightarrow$ 2 children crossover, such a thing is rarely done in practice, though. The point is that by doing so we may lose one of the alleles. In the above example the allele 1 does not appear in the third position of any one of the children while it is present in the third position of the second parent. One reason why 2 parent $\rightarrow$ 2 children crossover is often exploited in practice is so that the alleles are *swapped* and, thereby, neither one of them is lost during the process. In order to perform the swap one would have to use either the transformation $L_{\{2,5\}}$ in place of $L_{\{2,3,5\}}$ to obtain the second child or the transformation $L_{\{1,4\}}$ in place of $L_{\{1,3,4\}}$ to obtain the first child. Interestingly enough, this also ensures that the pair of parents can always be obtained from the pair of children by performing crossover alone (no mutation), or, in other words, that the transformation $\langle L_{\{1,3,4\}}, L_{\{2,5\}} \rangle : \Omega^2 \rightarrow \Omega^2$ is bijective.

Example 5.1. More generally, for the classical genetic algorithm the search space $\Omega = \prod_{i=1}^n A_i$ where $|A_i| \geq 2$. Let $\mathbf{a} = (a_1, a_2, \dots, a_n)$ and $\mathbf{b} = (b_1, \dots, b_n) \in \Omega$ denote typical elements. Let $M \subseteq \{1, 2, \dots, n\}$ and consider a reproduction (a masked crossover) transformation $L_M : \Omega^2 \rightarrow \Omega$ defined as follows:

$$L_M(\mathbf{a}, \mathbf{b}) = (x_1, x_2, \dots, x_i, \dots, x_n)$$

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

Now let $\mathcal{F}_M = \{L_M \mid M \subseteq \{1, 2, \dots, n\}\}$. Swapping the alleles when performing crossover corresponds to selecting the pair of transformations $(L_M, L_{\bar{M}})$ where $\bar{M}$ denotes the complement of M in $\{1, 2, \dots, n\}$. At the same time the following is also true (for a more general version of proposition 5.1 which uses the ideas from (Mitavskiy, 2003) see theorem 3.3.7 of (Mitavskiy, 2004b)).

Proposition 5.1. A given transformation $\langle L_M, L_K \rangle : \Omega^2 \rightarrow \Omega^2$ with $M, K \subseteq \{1, 2, \dots, n\}$, is bijective (one-to-one and onto) if and only if $K = \bar{M}$.

Proof. The “if” part is immediate since the transformation $\langle L_M, L_{\bar{M}} \rangle : \Omega^2 \rightarrow \Omega^2$ is an involution (it is its own inverse) i. e. applying the same crossover transformation to the pair of children gives back the original parents.

For the “only if” part, suppose to the contrary that $\langle L_M, L_K \rangle$ is bijective and yet $K \neq \bar{M}$. Then either $K \cap M \neq \emptyset$ or $\bar{K} \cap \bar{M} \neq \emptyset$ which means that either there exists an $i \in K \cap M$ or there exists an $i \in \bar{K} \cap \bar{M}$. In either case, according to the way in which the transformations in the family $\mathcal{F}_M$ are defined (see the discussion above), there is an i such that either one of the following is true:

1. For all $\mathbf{a} = (a_1, a_2, \dots, a_n)$ and $\mathbf{b} = (b_1, b_2, \dots, b_n)$, whenever $\langle L_M, L_K \rangle(\mathbf{a}, \mathbf{b}) = (x_1, x_2, \dots, x_n)$ we have $x_i = y_i = a_i$ where $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)$.
2. For all $\mathbf{a} = (a_1, a_2, \dots, a_n)$ and $\mathbf{b} = (b_1, b_2, \dots, b_n)$, whenever $\langle L_M, L_K \rangle(\mathbf{a}, \mathbf{b}) = (x_1, x_2, \dots, x_n)$ we have $x_i = y_i = b_i$ where $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)$.

Fix u and $v \in A_i$ with $u \neq v$ (recall that $|A_i| \geq 2$ by assumption). Fix arbitrary $a_i \in A_i$ and consider the individuals $\mathbf{u} = (a_1, a_2, \dots, a_{i-1}, u_i, a_{i+1}, \dots, a_n)$ and $\mathbf{v} = (a_1, a_2, \dots, a_{i-1}, v_i, a_{i+1}, \dots, a_n)$. Without loss of generality assume that condition 1 above holds (otherwise 2 holds and the argument below can be repeated verbatim replacing $\langle L_M, L_K \rangle(\mathbf{u}, \mathbf{v})$ with $\langle L_M, L_K \rangle(\mathbf{v}, \mathbf{u})$ throughout). Notice that $\langle L_M, L_K \rangle(\mathbf{u}, \mathbf{v}) \in (S_H)^2$ where $S_H = (\prod_{j=1}^{i-1} A_j) \times \{u\} \times (\prod_{j=i+1}^n A_j)$ is the subset of the search space determined by the Holland schema $H = (*, *, \dots, *, u, *, \dots, *)$ having u in the i^{th} position and the $*$ symbol everywhere else. Since the subsets of the search space determined by Holland schemata are well known to be invariant under crossover (see, for instance, corollary 3.4 of (Mitavskiy, 2004a)) we deduce that $\langle L_M, L_K \rangle$ maps $(S_H)^2$ into itself. We also assumed that $\langle L_M, L_K \rangle$ is one-to-one. Since $(S_H)^2$ is a finite set, the restriction $\langle L_M, L_K \rangle|_{(S_H)^2} : (S_H)^2 \rightarrow (S_H)^2$ is onto as well. But then there is a pair $(\mathbf{z}, \mathbf{w}) \in (S_H)^2$ with $\langle L_M, L_K \rangle(\mathbf{z}, \mathbf{w}) = \langle L_M, L_K \rangle(\mathbf{u}, \mathbf{v})$ since $\langle L_M, L_K \rangle(\mathbf{u}, \mathbf{v}) \in (S_H)^2$. But $(\mathbf{u}, \mathbf{v}) \notin (S_H)^2$ since $u \neq v$ and so $(\mathbf{u}, \mathbf{v}) \neq (\mathbf{z}, \mathbf{w})$ contradicting the injectivity of $\langle L_M, L_K \rangle$. The desired conclusion that $K = \bar{M}$ now follows. $\square$

Definition 5.2. A given elementary step of type reproduction associated with the reproduction k -tuple $(\Omega, \mathcal{F}_{q_1}, \mathcal{F}_{q_2}, \dots, \mathcal{F}_{q_k}, p_1, p_2, \dots, p_k, \wp_m)$ is said to be bijective if it satisfies condition 1 stated below:

1. For every i with $1 \leq i \leq k$ we have $p_i(T_1, T_2, \dots, T_{q_i}) > 0 \implies \langle T_1, T_2, \dots, T_{q_i} \rangle$ is a bijection (a one-to-one and onto map of Ω^{q_i} onto itself).

If, in addition to condition 1, condition 2 stated below is satisfied as well, then we say that the algorithm is bijective and self-transient.

2. For every i with $1 \leq i \leq k$ there exists a tuple $(T_1, T_2, \dots, T_{q_i}) \in (\mathcal{F}_{q_i})^{q_i}$ such that $p_i(T_1, T_2, \dots, T_{q_i}) > 0$ and $\langle T_1, T_2, \dots, T_{q_i} \rangle = \mathbf{1}$ where $\mathbf{1} : \Omega^{q_i} \rightarrow \Omega^{q_i}$ denotes the identity map (i. e. for every $\mathbf{x} \in \Omega^{q_i}$ we have $\langle T_1, T_2, \dots, T_{q_i} \rangle(\mathbf{x}) = \mathbf{x}$).

We say that a recombination sub-algorithm (see definition 4.1) of a given evolutionary search algorithm is bijective (and self-transient) if every given term of the sub-sequence by which the sub-algorithm is defined is bijective (and self-transient).

Remark 5.1. Notice that no explicit assumption is made about the probability distribution on the partitioning, $\wp_m$.

Remark 5.2. It is possible to weaken property 2 so that the generalized Geiringer theorem still holds by using the so-called “postage stamp problem” (see exercise 6 on page

277 of (Dummit and Foote, 1991)), but this does not introduce any new ideas of particular interest to the current paper, even more so since most reasonable evolutionary search algorithms satisfy condition 2.

The properties of definition 5.2 allow us to apply corollary 2.3 from section 2 by means of lemma 5.3 to extend the Geiringer's theorem. To alleviate notation complexity at least to some extent, we shall prove the following intermediate lemma first:

Lemma 5.2. *Let s denote a bijective elementary step and let*

$$\Omega_s = (\Omega, \mathcal{F}_{q_1}, \mathcal{F}_{q_2}, \dots, \mathcal{F}_{q_k}, p_1, p_2, \dots, p_k, \wp_m)$$

denote the reproduction k -tuple associated with the step s . Then there is a family of permutations $H \subseteq \mathcal{S}_{\Omega^m}$ (recall that $\mathcal{S}_{\Omega^m}$ denotes the group of permutations on Ω^m) and a probability distribution μ on H which depends only on $p_1, p_2, \dots, p_k$ and on $\wp_m$ such that the elementary step s can be described as follows: given an input $\mathbf{x} = (x_1, x_2, \dots, x_m) \in \Omega^m$, the output of s is obtained by selecting a permutation $\pi \in H$ according to the probability distribution μ and applying it to $\mathbf{x}$. In other words, the output of s on the input $\mathbf{x}$ is $\pi(\mathbf{x})$ with probability $\mu(\pi)$. Moreover, if the elementary step s is self-transient as well, then $\mathbf{1}_{\Omega^m} \in H$ (here $\mathbf{1}_{\Omega^m}$ denotes the identity map on Ω^m) and $\mu(\mathbf{1}_{\Omega^m}) > 0$.

Proof. Fix a partition $K = (P_1, P_2, \dots, P_l) \in \mathcal{E}_{\mathbf{q}}^m$ where $\mathbf{q} = (q_1, q_2, \dots, q_k)$ (see definition 3.2). Now for every i with $1 \leq i \leq l$ fix a particular choice of tuples of transformations $\tau_i = \langle T_1^i, T_2^i, \dots, T_{|P_i|}^i \rangle : \Omega^{|P_i|} \rightarrow \Omega^{|P_i|}$ such that $p^{P_i}(T_1^i, T_2^i, \dots, T_{|P_i|}^i) > 0$ where $p^{P_i} = p_j$ for some j (see definitions 5.1 and 3.3). First notice that we can identify Ω^m with the set $\Omega^{|P_1|} \times \Omega^{|P_2|} \times \dots \times \Omega^{|P_l|}$ via partition $K = (P_1, P_2, \dots, P_l)$ as follows: given $\mathbf{x} = (x_1, x_2, \dots, x_m) \in \Omega^m$, identify $\mathbf{x} \in \Omega^m$ with the element $\bar{\mathbf{u}}_{\mathbf{x}} = (\mathbf{u}_1^{\mathbf{x}}, \mathbf{u}_2^{\mathbf{x}}, \dots, \mathbf{u}_l^{\mathbf{x}})$ where $\mathbf{u}_i^{\mathbf{x}} = (x_{a_1}, x_{a_2}, \dots, x_{a_{|P_i|}})$, $a_1, a_2, \dots, a_{|P_i|} \in P_i$ and $a_1 < a_2 < \dots < a_{|P_i|}$. Now define a transformation $T_{\tau_1, \tau_2, \dots, \tau_l}^K : \Omega^m \rightarrow \Omega^m$ as follows: Let $T_{\tau_1, \tau_2, \dots, \tau_l}^K(\mathbf{x}) = \mathbf{y}$ where $\mathbf{y} \in \Omega^m$ corresponds to $\bar{\mathbf{u}}_{\mathbf{y}} = (\tau_1(\mathbf{u}_1^{\mathbf{x}}), \tau_2(\mathbf{u}_2^{\mathbf{x}}), \dots, \tau_l(\mathbf{u}_l^{\mathbf{x}}))$. Notice that $T_{\tau_1, \tau_2, \dots, \tau_l}^K(\mathbf{x})$ is the output of the elementary step s when the partition K and the transformations $\tau_1, \tau_2, \dots, \tau_l$ are chosen for reproduction (see definition 3.4). Observe also that the transformation $T_{\tau_1, \tau_2, \dots, \tau_l}^K$ is a bijection since by definition 5.2 every τ_i is. Indeed, the two-sided inverse of $T_{\tau_1, \tau_2, \dots, \tau_l}^K$ is the transformation $(T_{\tau_1, \tau_2, \dots, \tau_l}^K)^{-1}$ which sends a given $\mathbf{x} \in \Omega^m$ into $\mathbf{y} \in \Omega^m$ corresponding to the element

$$(\tau_1^{-1}(\mathbf{u}_1^{\mathbf{x}}), \tau_2^{-1}(\mathbf{u}_2^{\mathbf{x}}), \dots, \tau_l^{-1}(\mathbf{u}_l^{\mathbf{x}})) \in \Omega^{|P_1|} \times \Omega^{|P_2|} \times \dots \times \Omega^{|P_l|}.$$

Now let $H = \{T_{\tau_1, \tau_2, \dots, \tau_l}^K \mid K \text{ is a partition in } \mathcal{E}_{\mathbf{q}}^m \text{ and } \tau_1, \tau_2, \dots, \tau_l \text{ are chosen for reproduction as in definition 3.4}\}$ denote the set of all such transformations. By construction H consists of bijections and so $H \subseteq \mathcal{S}_{\Omega^m}$. Moreover, according to the construction above, the probability, say μ , with which a transformation in H is selected to obtain an output of s on a given input depends exclusively on the choice of a $\mathbf{q}$ -fit partition K which is made with respect to the probability distribution $\wp_m$ and on the choices of tuple-transformations $\tau_1, \tau_2, \dots, \tau_l$ which depends on K and on the probability distributions $p_1, p_2, \dots, p_k$. This establishes the first assertion of the lemma. The second assertion is almost immediate: if s is self-transient then $\tau_i = \mathbf{1}_{\Omega^{|P_i|}}$ has a positive chance of being selected. Doing so for every i with $1 \leq i \leq l$ results in $T_{\tau_1, \tau_2, \dots, \tau_l}^K$ being the identity map. $\square$

Lemma 5.3. *Let $\mathcal{A}$ denote a given bijective algorithm (see definition 5.2) defined by a cycle $C = \{s_{\tau}\}_{\tau=1}^j$. Then there is a family of permutations $H \subseteq \mathcal{S}_{\Omega^m}$ and a probability distribution*

μ on H which depends intrinsically on the cycle C such that C can be described as follows: given an input $\mathbf{x} = (x_1, x_2, \dots, x_m) \in \Omega^m$, the output of C is obtained by selecting a permutation $\pi \in H$ according to the probability distribution μ and applying it to $\mathbf{x}$. In other words, the output of C on the input $\mathbf{x}$ is $\pi(\mathbf{x})$ with probability $\mu(\pi)$ and μ does not depend on $\mathbf{x}$. Moreover, if the algorithm $\mathcal{A}$ is self-transient as well, then $\mathbf{1}_{\Omega^m} \in H$ (here $\mathbf{1}_{\Omega^m}$ denotes the identity map on Ω^m) and $\mu(\mathbf{1}_{\Omega^m}) > 0$.

Proof. According to lemma 5.2 for every τ with $1 \leq \tau \leq j$ there is a subset $H_\tau \subseteq \mathcal{S}_{\Omega^m}$ and a probability distribution μ_τ on H_τ such that the output of the elementary step s_n on any given input $\mathbf{x}$ is $\pi_\tau(\mathbf{x})$ with probability $\mu_\tau(\pi_\tau)$. Now simply define $H = \{\pi \mid \pi = \pi_1 \circ \pi_2 \circ \dots \circ \pi_\tau\}$ and let μ be the probability distribution on H which selects a given permutation $\pi = \pi_1 \circ \pi_2 \circ \dots \circ \pi_\tau \in H$ by choosing every π_k with probability μ_k independently. It is easy to see that this choice of H and μ establishes the lemma. $\square$

In (Vose, 1999) it has been pointed out that heuristic search algorithms give rise to the following Markov process³ (see also (Coffey, 1999), 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 $p_{\mathbf{x}\mathbf{y}}$ 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 4: see definition 3.5 and remark 3.3). These transition probabilities have been computed in (Vose, 1999), but the formulas are rather involved and the Markov chain obtained in such a way is, in general (without any extra assumptions about the algorithm), difficult to analyze. Lemma 5.3 shows that a Markov chain associated with a bijective and self-transient sub-algorithm is a random walk associated with a group action (see definition 2.4 and example 2.2). Indeed, if $\mathcal{A}$ is a given bijective and self-transient algorithm, let H be a set of permutations constructed in Lemma 5.3. Let G be the subgroup of $\mathcal{S}_{\Omega^m}$ generated by H (the set of all products of elements of H is a subgroup of $\mathcal{S}_{\Omega^m}$ since $\mathcal{S}_{\Omega^m}$ is a finite group). According to remark 2.1 definition 5.3, proposition 5.4 and definition 5.4 describe the orbit of a given population $\mathbf{x}$ under the action of the group G on the set of populations Ω^m (this is the action described in example 2.2):

Definition 5.3. Fix an evolutionary search algorithm $\mathcal{A}$. Denote by $p_{\mathbf{x},\mathbf{y}}^t$ the probability that a population $\mathbf{y}$ is obtained from the population $\mathbf{x}$ upon the completion of t complete cycles (in the sense of definition 3.5 and remark 3.3) of the algorithm. We say that a population $\mathbf{x}$ leads to a population $\mathbf{y}$ under $\mathcal{A}$ if and only if $p_{\mathbf{x},\mathbf{y}}^t > 0$ for some t . 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 (Coffey, 1999).)

The following proposition is then simply a restatement of proposition 2.2 in our specific case:

Proposition 5.4. Let $\mathcal{A}$ denote a bijective and self-transient algorithm (see definition 5.2). Then $\xrightarrow{\mathcal{A}}$ is an equivalence relation.

In view of proposition 5.4 (compare with definition 2.5) we can introduce the following notation:

Definition 5.4. 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 $\xrightarrow{\mathcal{A}}$.

³In the current paper the state space of this process is slightly modified for technical reasons which will be seen later.

When a given bijective and self-transient algorithm $\mathcal{A}$ starts running with the initial population P thanks to proposition 5.4, only the populations in $[P]_{\mathcal{A}}$ may occur with nonzero probability. It makes sense, therefore, to restrict the state space of our Markov chain to include only the elements of the equivalence class $[P]_{\mathcal{A}}$. We shall call such a Markov process “the Markov chain initiated at P ” (compare with the discussion following definition 2.5). The generalized Geiringer theorem is now simply a special case of the corollary 2.3:

Theorem 5.5. *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 $[P]_{\mathcal{A}}$).*

Remark 5.3. The statement of theorem 5.5 appears to be very different from the way in which the classical result has been stated. The similarity between the two versions will become more clear in section 6 (see also remark 5.5). It is also worth mentioning that the classical version of the Geiringer theorem for *infinite* population does not require the reproduction transformations to be bijective.

Remark 5.4. Property 2 of definition 5.2 ensures that $p_{x,y}^1 > 0$ which prevents the Markov chain initiated at some population P from being periodic. However, for small self-transition probability there may be rather long periodic transients. Sinclair (Sinclair, 1992) set the lower bound on the self-transition probability to $\frac{1}{2}$ to ensure that all eigenvalues of the Markov chain are nonnegative, hence preventing any transient periodicities. Such a lower bound is unreasonable in our circumstances as the self-transition probability is usually quite small. However, there may be other mechanisms which ensure reasonably fast convergence to the stationary distribution. We leave this subject for future investigation.

Remark 5.5. Notice that the framework above seems to apply to the algorithms with a *fixed* population size (which is denoted by m throughout the paper). It is worth pointing out that there is actually no restriction of this sort in the following sense. One can simply think of a population P as a probability distribution on the search space Ω defined as follows: Given a population $P \in \Omega^m$ denote by λ_P the probability distribution on Ω defined by $\lambda_P(x) = (\text{the number of times } x \text{ occurs in } P)/m$. Thereby we have a function $\phi : \Omega^m \rightarrow \Lambda_m$ between the set of all populations of size m and the collection of all probability distributions on Ω which assign to a given singleton subset $\{x\} \subseteq \Omega$ a rational probability whose denominator (when the fraction is written in the reduced form) divides m . The Markov transition matrix associated with an evolutionary algorithm with a fixed population size defines an operator, $\mathcal{M} : \mathcal{Z} \rightarrow \mathcal{Z}$, on the collection $\mathcal{Z}$ of probability distributions on $[P]_{\mathcal{A}}$. Theorem 5.5 says that for any $z \in \mathcal{Z}$ the sequence $\mathcal{M}(z), \mathcal{M}^2(z), \dots, \mathcal{M}^n(z), \dots$ converges to the uniform distribution on $[P]_{\mathcal{A}}$. Consider now the map $\alpha : \mathcal{Z} \rightarrow \Lambda_m$ defined as follows: $\alpha(z)(x) = \sum_{Q \in [P]_{\mathcal{A}}} z(Q) \lambda_Q(x)$. Evidently, the map α is continuous and so we deduce that the sequence of distributions $\alpha(\mathcal{M}(z)), \alpha(\mathcal{M}^2(z)), \dots, \alpha(\mathcal{M}^n(z)), \dots$ converges to the distribution $\alpha(\varrho)$ where ϱ denotes the uniform distribution on $[P]_{\mathcal{A}}$. This interpretation of theorem 5.5 stays more in line with the way in which the classical Geiringer theorem is stated. The primary difference between the two versions lies in the fact that the sequence of probability distributions obtained from the Markov chain is actually very different from the sequence obtained applying the “crossover operator” in the introduction. Interestingly, it turns out that the two sequences do converge to the same limit distribution when we use 2 parent $\rightarrow$ 2 children crossover in the finite population model for the classical GA and variable length GP as we shall see in later sections. Although the general circumstances

under which this happens remain an open question, some progress in this direction has already been made and it will be presented in a sequel paper.

6 Applications

In order to obtain the classical versions of the Geiringer theorem for the known examples (examples 5.1, 6.2 and 6.3) from the generalized version (theorem 5.5), one has to introduce a few definitions and to prove a few corollaries first:

Definition 6.1. We define the characteristic function $\mathcal{X} : \Omega^m \times \mathcal{P}(\Omega) \rightarrow \mathbb{N} \cup \{0\}$ as follows: $\mathcal{X}(P, S) =$ the number of individuals of P which are 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 6.1. Let

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

Suppose $\Omega = \{0, 1\}^5$, and $S \subseteq \Omega = \{0, 1\}^5$ is determined by the Holland schema $(*, 1, *, 1, *)$. Then the 1st, the 2nd, and the 5th rows of P are elements of S . The rest of the rows of P are not in S . Thereby, exactly 3 rows of P are elements of S , so that $\mathcal{X}(P, S) = 3$.

Definition 6.2. 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 cycles of the algorithm $\mathcal{A}$. 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 $\mathcal{X}(\square, S) : \Omega^m \rightarrow \mathbb{N}$ 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).

Corollary 6.1. Fix a 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 5.4). Fix a set S of individuals in Ω ($S \subseteq \Omega$). Then $\lim_{t \rightarrow \infty} \Phi(S, P, t) = \frac{1}{m} E_{\varrho_{[P]_{\mathcal{A}}}}(\mathcal{X}(\square, S))$ (here $E_{\varrho_{[P]_{\mathcal{A}}}}(f)$ denotes the expectation of the random variable f with respect to the uniform distribution on the set $[P]_{\mathcal{A}}$).⁴

Proof. This is a natural consequence of the generalized Geiringer theorem (theorem 5.5). The details are as follows: Fix an initial population P . For any given population $P' \in [P]_{\mathcal{A}}$ denote by $\#(P', t)$ the number of times that the population P' has been encountered during the run before time t . With this notation we have

$$\Phi(S, P, t) = \frac{\sum_{s=1}^t \mathcal{X}(P(s), S)}{tm} = \sum_{P' \in [P]_{\mathcal{A}}} \mathcal{X}(P', S) \frac{\#(P', t)}{tm}.$$

⁴Throughout the paper, whenever a limit is involved, the almost everywhere type of convergence is meant.

But, according to theorem 5.5, $\lim_{t \rightarrow \infty} \frac{\#(P', t)}{t} = \frac{1}{|[P]_{\mathcal{A}}|}$ so that we finally obtain

$$\begin{aligned} \lim_{t \rightarrow \infty} \Phi(S, P, t) &= \lim_{t \rightarrow \infty} \sum_{P' \in [P]_{\mathcal{A}}} \mathcal{X}(P', S) \frac{\#(P', t)}{tm} = \frac{1}{m} \sum_{P' \in [P]_{\mathcal{A}}} \mathcal{X}(P', S) \lim_{t \rightarrow \infty} \frac{\#(P', t)}{t} = \\ &= \frac{1}{m} \sum_{P' \in [P]_{\mathcal{A}}} \mathcal{X}(P', S) \frac{1}{|[P]_{\mathcal{A}}|} = \frac{1}{m} E_{\varrho_{[P]_{\mathcal{A}}}}(\mathcal{X}(\square, S)) \end{aligned}$$

which is exactly what we were after. $\square$

To establish 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 6.3. We say that a given algorithm $\mathcal{A}$ is regular if the following is true: for every population $P = (x_1, x_2, \dots, x_m) \in \Omega^m$ and for every permutation $\pi \in \mathcal{S}_m$, the population obtained by permuting the elements of P by π , namely $\pi(P) = (x_{\pi(1)}, x_{\pi(2)}, \dots, x_{\pi(m)}) \in [P]_{\mathcal{A}}$. In words this says that the equivalence classes $[P]_{\mathcal{A}}$ are permutation invariant.

Remark 6.1. Definition 6.3 is only needed because our description of an evolutionary search algorithm uses the ordered multi-set model. This makes the generalized Geiringer theorem (theorem 5.5) 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 (Vose, 1999) 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 6.2. 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 5.4). Fix a set S of individuals in Ω ($S \subseteq \Omega$). Then $\lim_{t \rightarrow \infty} \Phi(S, P, t) = \varrho_{[P]_{\mathcal{A}}}(\mathcal{V}_S)$ where

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

Proof. Denote by $\mathcal{X}_i$ the characteristic function of the set

$$\mathcal{V}_S^i = \{P \mid P \in [P]_{\mathcal{A}}, \text{ and the } i^{\text{th}} \text{ individual of } P \text{ is an element of } S\}$$

(Notice that $\mathcal{V}_S^1 = \mathcal{V}_S$. Saying that $\mathcal{X}_i$ is the characteristic function of the set $\mathcal{V}_S^i$ means that $\mathcal{X}_i(P) = \begin{cases} 1 & \text{if } P \in \mathcal{V}_S^i \\ 0 & \text{otherwise} \end{cases}$.) First, observe that $E_{\varrho_{[P]_{\mathcal{A}}}}(\mathcal{X}_i(P)) = \varrho_{[P]_{\mathcal{A}}}(\mathcal{V}_S^i)$. Moreover, since $\varrho_{[P]_{\mathcal{A}}}$ is a uniform distribution on $[P]_{\mathcal{A}}$, we have

$$\forall i, j \quad E_{\varrho_{[P]_{\mathcal{A}}}}(\mathcal{X}_i(P)) = \varrho_{[P]_{\mathcal{A}}}(\mathcal{V}_S^i) = \varrho_{[P]_{\mathcal{A}}}(\mathcal{V}_S^j) = E_{\varrho_{[P]_{\mathcal{A}}}}(\mathcal{X}_j(P))$$

(Indeed, since $\varrho_{[P]_{\mathcal{A}}}$ is the uniform distribution, any bijection on $[P]_{\mathcal{A}}$ is an isomorphism of probability spaces. Thanks to the regularity assumption (see definition 6.3) the map $t_{i \leftrightarrow j}$ which swaps i^{th} and j^{th} individuals is a well-defined bijection on $[P]_{\mathcal{A}}$. Therefore,

we have $\varrho_{[P]\mathcal{A}}(\mathcal{V}_S^i) = \varrho_{[P]\mathcal{A}}(t_{i \leftrightarrow j}(\mathcal{V}_S^i)) = \varrho_{[P]\mathcal{A}}(\mathcal{V}_S^j)$. It follows, therefore, that for every i we have $E_{\varrho_{[P]\mathcal{A}}}(\mathcal{X}_i(P)) = \varrho_{[P]\mathcal{A}}(\mathcal{V}_S)$. Now simply observe that $\mathcal{X}(\square, S) = \sum_{i=1}^m \mathcal{X}_i$ so that, according to corollary 6.1, we have

$$\begin{aligned} \lim_{t \rightarrow \infty} \Phi(S, P, t) &= \frac{1}{m} E_{\varrho_{[P]\mathcal{A}}}(\mathcal{X}(\square, S)) = \frac{1}{m} E_{\varrho_{[P]\mathcal{A}}}(\sum_{i=1}^m \mathcal{X}_i) \\ &= \frac{1}{m} \sum_{i=1}^m E_{\varrho_{[P]\mathcal{A}}}(\mathcal{X}_i) = \frac{1}{m} (m \times \varrho_{[P]\mathcal{A}}(\mathcal{V}_S)) = \varrho_{[P]\mathcal{A}}(\mathcal{V}_S). \end{aligned}$$

□

Remark 6.2. Notice that corollary 6.2 says that $\lim_{t \rightarrow \infty} \Phi(S, P, t)$ is simply the proportion of all populations whose 1st element is in S among all populations in $[P]_{\mathcal{A}}$.

The classical and the other already existing versions of the Geiringer theorem are fairly routine consequences of corollaries 6.1 and 6.2. The only relatively tedious step is to describe the equivalence class $[P]_{\mathcal{A}}$ explicitly. We shall informally refer to this step as proving the “orbit-description lemma”.

Example 6.2 (Classical Genetic algorithm). Continuing with example 5.1 let s_2 denote the elementary step of type reproduction associated with the reproduction 1-tuple $\Omega_{\text{crossover}} = (\Omega, \mathcal{F}_M, p, \wp_m)$ (see definition 3.3). Here p is a probability distribution on the set $(\mathcal{F}_M)^2$ which is concentrated on the pairs of the form $(L_M, L_{\bar{M}})$ where $\bar{M}$ denotes the complement of M in $\{1, 2, \dots, n\}$. Notice that if $M = \{1, 2, \dots, l\}$ or $M = \{l+1, \dots, n\}$ then L_M is simply the one-point crossover transformation. Often p is also concentrated on the pairs of the form $(L_M, L_{\bar{M}})$ where $M = \{1, 2, \dots, l\}$ or $M = \{l+1, \dots, n\}$ so that one uses only the one-point crossover transformations. $\wp_m$ is then a probability distribution on the collection of partitions of the set of m elements into 2-element subsets (in the classical case, usually m is even). Usually $\wp_m$ is taken to be the uniform distribution.⁵ Now we simply let s_2 be the elementary step of type reproduction associated with the reproduction 1-tuple $\Omega_{\text{crossover}}$. Notice that the classical genetic algorithm described here is bijective and self-transient regardless of whether one uses only the one-point crossover transformations or the entire family of masked crossover transformations. Indeed, p is concentrated on the transformations of the form $(L_M, L_{\bar{M}})$ and every transformation of this form is an involution on Ω^2 , hence it is bijective (see proposition 5.1). This shows that condition 1 of definition 5.2 is satisfied. Condition 2 of definition 5.2 follows from the observation that if we let $M = \{1, 2, \dots, n\}$ then $(L_M, L_{\bar{M}}) = 1$. The next step is to establish an appropriate “orbit-description” lemma. We prove Lemma 6.3 (as well as lemmas 6.5, 6.6, and 7.3) under the following two conditions on the algorithm $\mathcal{A}$:

Definition 6.4. We say that a given genetic algorithm $\mathcal{A}$ is fully shuffling if the following two conditions are satisfied:

1. For every pair $\{i, j\}$ with $i \neq j$ there exists a partition $P \in \mathcal{E}_{(2)}^m$ (see definition 3.2) such that $\{i, j\} \in P$ and $\wp_m(P) > 0$.
2. Every one-point crossover transformation has a positive probability of being chosen.

⁵The classical Geiringer theorem holds under weaker assumptions, though. The generalized Geiringer theorem holds under even weaker requirements, but its conclusion may need to be modified accordingly. This subject will be discussed in section 6.

Lemma 6.3. Let the initial population P be

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

Suppose the algorithm $\mathcal{A}$ is fully shuffling in the sense of definition 6.4. Then, independently of whether one exploits the family of one-point crossover transformations or the family of all masked crossovers, the corresponding equivalence class $[P]_{\mathcal{A}}$ is

$$\left\{ \begin{pmatrix} a_{\pi_1(1)1} & a_{\pi_2(1)2} & \dots & a_{\pi_n(1)n} \\ a_{\pi_1(2)1} & a_{\pi_2(2)2} & \dots & a_{\pi_n(2)n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{\pi_1(m)1} & a_{\pi_2(m)2} & \dots & a_{\pi_n(m)n} \end{pmatrix} \mid \pi_1, \pi_2, \dots, \pi_n \in \mathcal{S}_m \right\}$$

where $\mathcal{S}_m$ denotes the group of all permutations of the set $\{1, 2, \dots, m\}$.

Proof. This follows from the fact that every permutation on a finite set can be written as a product of transpositions (see for example (Dummit and Foote, 1991)) together with the observation that it is possible to apply masked crossover transformations which swap (transposes) two given alleles in a given column and fixes the rest of the alleles so that the group generated by all of the permutations constructed in the proofs of lemmas 5.2 and 5.3 (lemma 5.2 is sufficient since we deal with crossover alone) is $(\mathcal{S}_m)^n$ where $\mathcal{S}_m$ is the group of permutations on the set of m elements. The desired assertion for the case when we use the entire family of masked crossovers now follows at once. The observation that every masked crossover is a composition of one-point crossovers completes the remainder of the argument. $\square$

Corollary 6.2 tells us that $\lim_{t \rightarrow \infty} \Phi(S, P, t) = \varrho_{[P]_{\mathcal{A}}}(\mathcal{V}_S)$. This equation holds, in particular, for the singleton set $S = \{\mathbf{h} = (h_1, h_2, \dots, h_n)\}$. Recall that $\varrho_{[P]_{\mathcal{A}}}$ is the uniform probability distribution on the set $[P]_{\mathcal{A}}$ which is explicitly described in Lemma 6.3. The probability space $[P]_{\mathcal{A}}$ equipped with the uniform distribution $\varrho_{[P]_{\mathcal{A}}}$ is easily seen to be isomorphic to the product of the following probability spaces: Let $I_m = \{1, 2, \dots, m\}$ and let $f_i : I_m \rightarrow A_i$ be defined as follows: $f_i(j) = a_{ji}$. In words, f_i is simply the function which assigns to j the i^{th} allele of the j^{th} individual of P . Let $\Theta_i = \{f_i \circ \pi \mid \pi \in \mathcal{S}_m\}$. Now equip every Θ_i with the uniform distribution ρ_i and let $\Theta = \prod_{i=1}^n \Theta_i$. Evidently the distribution $\rho = \prod_{i=1}^n \rho_i$ is the uniform distribution on Θ (since it is the product of uniform distributions) so that the natural bijective identification $\phi : \Theta \rightarrow [P]_{\mathcal{A}}$ sending a given element $\theta = (f_1 \circ \pi_1, f_2 \circ \pi_2, \dots, f_n \circ \pi_n) \in \Theta$ to the population

$$\phi(\theta) = \begin{pmatrix} f_1(\pi_1(1)) & f_2(\pi_2(1)) & \dots & f_n(\pi_n(1)) \\ f_1(\pi_1(2)) & f_2(\pi_2(2)) & \dots & f_n(\pi_n(2)) \\ \vdots & \vdots & \ddots & \vdots \\ f_1(\pi_1(m)) & f_2(\pi_2(m)) & \dots & f_n(\pi_n(m)) \end{pmatrix}$$

is an isomorphism of probability spaces. Now, according to corollary 6.2 together with the above observation, we have $\lim_{t \rightarrow \infty} \Phi(\{\mathbf{h}\}, P, t) = \rho(K)$ where $K = \{\theta = (f_1 \circ \pi_1, f_2 \circ \pi_2, \dots, f_n \circ \pi_n) \mid \forall 1 \leq i \leq n \ f_i(\pi_i(1)) = h_i\} = \prod_{i=1}^n K_i$ where $K_i = \{f_i \circ \pi \mid \pi \in \mathcal{S}_m, \pi_i(1) \in f_i^{-1}(h_i)\}$. Since every ρ_i is a uniform distribution on K_i ,

we have $\rho_i(K_i) = \frac{|f_i^{-1}(h_i)|}{m} = \Phi(h_i, P, i)$ (recall from the introduction that $\Phi(h_i, P, i)$ traditionally denotes the proportion of the individuals in the population P belonging to the schema $(*, \dots, *, h_i, *, \dots, *)$ which has h_i in the i^{th} position and has no other fixed positions). Combining all of the above, we finally obtain:

$$\lim_{t \rightarrow \infty} \Phi(\{\mathbf{h}\}, P, t) = \rho(K) = \rho\left(\prod_{i=1}^n K_i\right) = \prod_{i=1}^n \rho_i(K_i) = \prod_{i=1}^n \Phi(h_i, P, i)$$

which is precisely what the classical Geiringer theorem asserts.

When deriving schema-based versions of Geiringer theorem for more complicated algorithms (such as linear genetic programming described in the next example) the following strategy may be implemented: Continuing with the notation in corollaries 6.1 and 6.2, suppose we are given a nested sequence of subsets of the search space: $S_1 \supseteq S_2 \supseteq \dots \supseteq S_n$. According to corollary 6.2,

$$\begin{aligned} \lim_{t \rightarrow \infty} \Phi(S_n, P, t) &= \varrho_{[P]_{\mathcal{A}}}(\mathcal{V}_{S_n}) = \frac{|\mathcal{V}_{S_n}|}{|[P]_{\mathcal{A}}|} = \frac{|\mathcal{V}_{S_n}|}{|\mathcal{V}_{S_{n-1}}|} \cdot \frac{|\mathcal{V}_{S_{n-1}}|}{|[P]_{\mathcal{A}}|} = \\ &= \frac{|\mathcal{V}_{S_n}|}{|\mathcal{V}_{S_{n-1}}|} \cdot \frac{|\mathcal{V}_{S_{n-1}}|}{|\mathcal{V}_{S_{n-2}}|} \cdot \dots \cdot \frac{|\mathcal{V}_{S_2}|}{|\mathcal{V}_{S_1}|} \cdot \frac{|\mathcal{V}_{S_1}|}{|[P]_{\mathcal{A}}|} = \\ &= \varrho_{[P]_{\mathcal{A}}}(\mathcal{V}_{S_1}) \cdot \prod_{j=0}^{n-2} \frac{|\mathcal{V}_{S_{n-j}}|}{|\mathcal{V}_{S_{n-j-1}}|} = \frac{1}{m} E_{\varrho_{[P]_{\mathcal{A}}}}(\mathcal{X}(\square, S_1)) \cdot \prod_{j=0}^{n-2} \frac{|\mathcal{V}_{S_{n-j}}|}{|\mathcal{V}_{S_{n-j-1}}|} \end{aligned}$$

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} . We summarize this observation in the following corollary:

Corollary 6.4. *Fix a regular bijective and self-transient algorithm $\mathcal{A}$ and an initial population $P \in \Omega^m$. Fix a nested sequence of subsets $S_1 \supseteq S_2 \supseteq \dots \supseteq S_n$ of individuals in Ω ($S_1 \subseteq \Omega$). Then*

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

where, as before, $\mathcal{V}_S$ denotes the set of all populations whose first individual is a member of S for a given subset $S \subseteq \Omega$.

Example 6.3. (Variable-Length GA or Linear GP:) This type of algorithm is considered in (Poli et al., 2002). The search space is $\Omega = \bigcup_{n=1}^N \prod_{i=1}^n A_i$ where $|A_i| \geq 2$. To define the family of masked crossover transformations we proceed as follows: Denote by $I_n = \{1, 2, \dots, n\}$ the set of all integers between 1 and n . Then $\mathcal{P}(I_n)$ denotes the power set (i. e. the set of all subsets) of I_n . The elements of $\mathcal{P}(I_n)$ (i. e. the subsets of I_n) serve as binary masks of length n . For every element $\mathbf{M} = (M_1, M_2, \dots, M_N) \in \prod_{n=1}^N \mathcal{P}(I_n)$ define a transformation $T_{\mathbf{M}} : \Omega^2 \rightarrow \Omega$ which can be described in words as follows: Fix any $\mathbf{a} = (a_1, a_2, \dots, a_l)$ and $\mathbf{b} = (b_1, b_2, \dots, b_k) \in \Omega$. Without loss of generality assume that $l < k$. Then the common region for the crossover is $I_l = \{1, 2, \dots, l\}$ and, if $i \in M_l$, the i^{th} allele of the offspring is the i^{th} allele of the first parent, while if $i \notin M_l$, the i^{th} allele of the offspring is the i^{th} allele of the second parent. The largest element of M_l has a similar meaning, except this time if the allele to be transferred to form the offspring belongs to the longer parent, then also the alleles to its right

are transferred. Formally we can write $T_M(\mathbf{a}, \mathbf{b}) = \mathbf{x} = (x_1, x_2, \dots, x_{B(l,k)})$ where $B(l, k) = \begin{cases} l & \text{if } l \in M_l \\ k & \text{if } l \notin M_l \end{cases}$ and, for $1 \leq i \leq l$ we have $x_i = \begin{cases} a_i & \text{if } i \in M_l \\ b_i & \text{if } i \notin M_l \end{cases}$ while for $l < i \leq B(l, k)$ we have $x_i = b_i$ (notice that if $B(l, k) = l$ then the last statement is vacuous). Likewise, $T_M(\mathbf{b}, \mathbf{a}) = \mathbf{x} = (x_1, x_2, \dots, x_{B(l,k)})$ where $B(l, k) = \begin{cases} l & \text{if } l \notin M_l \\ k & \text{if } l \in M_l \end{cases}$

and, for $1 \leq i \leq l$ we have $x_i = \begin{cases} a_i & \text{if } i \notin M_l \\ b_i & \text{if } i \in M_l \end{cases}$ while for $l < i \leq B(l, k)$ we

have $x_i = b_i$ (if $B(l, k) = l$ then the last statement is vacuous). For example, say $A_i = \{0, 1, 2, \dots, i, i+1\}$ and $N = 10$. Let's say we fix the individuals $\mathbf{a} = (1, 2, 3, 4, 5)$ and $\mathbf{b} = (2, 3, 4, 5, 6, 7, 8)$. Let's say we fix a sequence $\mathbf{M} = \{M_1, M_2, \dots, M_{10}\}$ of crossover masks such that $M_5 = \{1, 3, 5\}$. $T_M(\mathbf{a}, \mathbf{b}) = (1, 3, 3, 5, 5)$. Notice that the result of crossover is completely determined by M_5 . This is simply because the shorter parent has length 5. The offspring has the length of the shorter parent because it is the 1st parent and also because $5 \in M$. If the 1st parent is the longer one (namely $\mathbf{b}$) then the result of crossover is $T_M(\mathbf{b}, \mathbf{a}) = (2, 2, 4, 4, 6, 7, 8)$. We let $\mathcal{F} = \{T_M \mid \mathbf{M} = (M_1, M_2, \dots, M_N) \in \prod_{n=1}^N \mathcal{P}(I_n)\}$. Once again there are many different ways to choose a probability distribution on the set $\mathcal{F}^2$. For the elementary step (crossover) to be bijective (see definition 5.2), it is sufficient and necessary that the probability distribution is concentrated on the pairs of the form $(T_M, T_{\bar{\mathbf{M}}})$ where $\bar{\mathbf{M}} = (\bar{M}_1, \bar{M}_2, \dots, \bar{M}_N)$ and $\bar{M}_n$ denotes the complement of M_n in I_n . (In other words, $\bar{M} \neq K \Rightarrow p(T_M, T_K) = 0$.) Indeed, just like in the case of a classical genetic algorithm, every transformation of the form $(T_M, T_{\bar{\mathbf{M}}})$ is an involution, hence bijective. Moreover, if we let $\mathbf{M} = (I_1, I_2, \dots, I_N)$, then $(T_M, T_{\bar{\mathbf{M}}}) = \mathbf{1}$ (recall that $I_k = \{1, 2, \dots, k\}$) so that both conditions of definition 5.2 are satisfied. Thus, for the elementary step to be self-transient as well it is sufficient to require $p(T_M, T_{\bar{\mathbf{M}}}) > 0$ where $\mathbf{M} = (I_1, I_2, \dots, I_N)$. All we have to do now is to select a probability distribution on the set $\prod_{n=1}^N \mathcal{P}(I_n)$ (if a given element $\mathbf{M} \in \prod_{n=1}^N \mathcal{P}(I_n)$ is chosen, then this corresponds to selecting the pair $(T_M, T_{\bar{\mathbf{M}}})$). To obtain a description which is probabilistically equivalent to the one given in (Poli et al., 2002), for every i with $1 \leq i \leq N$ fix a probability distribution p_i on $\mathcal{P}(M_i)$. Now let $p = \prod_{n=1}^N p_n$ be the probability distribution on $\prod_{n=1}^N \mathcal{P}(I_n)$. To obtain the classical case where each mask is equally likely to be chosen, simply let every p_n be the uniform distribution on $\mathcal{P}(I_n)$. The probability distribution $\wp_m$ on the collection of all partitions into 2-element pieces of a set of cardinality m can be chosen in the same way as in example 6.2 above. Usually it is chosen to be uniform, but Geiringer's theorem applies under more general circumstances. We let s_2 be the elementary step of type reproduction associated with the reproduction 1-tuple $\Omega_{\text{crossover}} = (\Omega, \mathcal{F}, p, \wp_m)$. Let $\mathcal{A}$ denote the algorithm defined by a cycle $C = \{s_2\}$ ($\mathcal{A}$ simply performs crossover repeatedly). Here it is convenient to state the orbit description lemma in terms of schemata. We adopt the notation from (Poli et al., 2002): Recall that our search space $\Omega = \bigcup_{n=1}^N \prod_{i=1}^n A_i$ where $|A_i| \geq 2$. The following definitions describe explicitly which subsets of the search space are represented by which types of schemata:

Definition 6.5. The schema $H = (*^{i-1}, h_i, \#)$ represents the subset $S_H = \{\mathbf{x} = (x_1, x_2, \dots, x_l) \mid l > i \text{ and } x_i = h_i\}$. In words, S_H is simply the set of all individuals whose length is at least $i + 1$ and whose i^{th} allele is h_i .

Definition 6.6. The schema $H = (*^i, \#)$ represents the subset $S_H = \{\mathbf{x} =$

$(x_1, x_2, \dots, x_l) \mid l > i\}$. In words, S_H is simply the subset of all individuals whose length is at least $i + 1$.

Definition 6.7. The schema $H = (*^i, h_{i+1}, h_{i+2}, \dots, h_{i+j})$ represents the subset $S_H = \{\mathbf{x} = (x_1, x_2, \dots, x_{i+j}) \mid x_{i+k} = h_{i+k} \forall k \text{ with } 1 \leq k \leq j\}$ of the search space which is simply the set of all individuals of length exactly equal to $i + j$ whose last j alleles are $h_{i+1}, h_{i+2}, \dots, h_{i+j}$.

Fix a population P . Recall, that for a given subset $S \subseteq \Omega$, $\mathcal{X}(P, S)$ denotes the number of individuals of P which are elements of S (see definition 6.1), while $\mathcal{V}_S$ denotes the set of populations in $[P]_{\mathcal{A}}$ whose first individual is a member of S (see corollary 6.2). In case when $S = S_H$ for some schema H , we shall write $\mathcal{X}(P, H)$ instead of $\mathcal{X}(P, S_H)$ and $\mathcal{V}_H$ instead of $\mathcal{V}_{S_H}$ respectively. Before proving the orbit-description lemma, let's recall a well-known fact about crossover in genetic algorithms:

Remark 6.3. Notice that due to our choice of the probability distribution on $\mathcal{F}^2$, the alleles are neither lost nor created during crossover. In terms of schemata this means that $\mathcal{X}(Q, H) = \mathcal{X}(P, H)$ for every population $Q \in [P]_{\mathcal{A}}$ where $H = (*^{i-1}, h_i, \#)$ or $H = (*^{i-1}, h_i)$. (Crossover is defined in such a way that an interior allele can not become a terminal allele and vice versa: a terminal allele can not become an interior one after crossover.) In particular, the function $\mathcal{X}(\square, S_H)$ is constant on the equivalence class $[P]_{\mathcal{A}}$ and its value is $\mathcal{X}(P, H) = \mathcal{X}(P, (*^{l-1}, h_l))$. This tells us that the expectation (with respect to an arbitrary probability distribution) of $\mathcal{X}(\square, S_{(*^{l-1}, h_l)})$ is exactly $\mathcal{X}(P, (*^{l-1}, h_l))$ which is simply the number of individuals of length l whose last allele is h_l which occur in P (we shall need this fact to apply corollary 6.4).

It turns out that it is a lot more convenient to establish the orbit-description lemma in a special case when all of the alleles corresponding to the same gene occurring in the initial population P are distinct. In mathematical symbols this assumption means that if $H = (*^{i-1}, x_i, \#)$ or $H = (*^{i-1}, x_i)$ then $\mathcal{X}(P, H) = 1$ or $\mathcal{X}(P, H) = 0$. We shall then apply corollary 6.4 to derive a schema-based version of Geiringer theorem for this special case. The general case will then follow by assigning enough extra labels to the alleles involved and then looking at the projection obtained by deleting the extra labels. A completely analogous strategy will be used in a forthcoming paper when we derive a schema-based version of Geiringer theorem for non-linear genetic programming.

Before stating the next orbit-description lemma let's notice that definition 6.4 can be restated for the case of a variable-length genetic algorithm just as easily. The only thing that needs to be formally mentioned is that one-point crossover transformations for a linear GP are just these which are built from the one-point crossovers for the classical GA:

Definition 6.8. Consider the family $\mathcal{F} = \{T_{\mathbf{M}} \mid \mathbf{M} = (M_1, M_2, \dots, M_N) \in \prod_{n=1}^N \mathcal{P}(I_n)\}$ of the masked crossover transformations for the linear GP (variable-length GA) as described in the preceding discussion. We say that a transformation $F \in \mathcal{F}$ is a one-point crossover transformation if for every $\mathbf{a} = (a_1, a_2, \dots, a_l)$ and $\mathbf{b} = (b_1, b_2, \dots, b_k) \in \bigcup_{n=1}^N \prod_{i=1}^n A_i$ there is an integer q with $0 \leq q \leq \min\{l, k\}$ such that either

$$F(\mathbf{a}, \mathbf{b}) = (a_1, a_2, \dots, a_q, b_{q+1}, \dots, b_k) \text{ or } F(\mathbf{a}, \mathbf{b}) = (b_1, b_2, \dots, b_q, a_{q+1}, \dots, a_l).$$

We also say that a given variable-length genetic algorithm is fully shuffling if conditions (1) and (2) below are satisfied:

1. For every pair $\{i, j\}$ with $i \neq j$ there exists a partition $P \in \mathcal{E}_{(2)}^m$ (see definition 3.2) such that $\{i, j\} \in P$ and $\wp_m(P) > 0$.

2. Every one-point crossover transformation has a positive probability of being chosen.

Lemma 6.5. Let $\mathcal{A}$ denote a fully shuffling variable-length genetic algorithm (see definition 6.8). Fix an initial population P such that for any schema H of the form $H = (*^{i-1}, x_i, \#)$ or $H = (*^{i-1}, x_i)$ we have $\mathcal{X}(P, H) = 1$ or $\mathcal{X}(P, H) = 0$. Fix an individual $\mathbf{h} = (h_1, h_2, \dots, h_l) \in \Omega$. Consider the following sequence of subsets of the search space Ω determined by schemata: $S_{(*^{l-1}, h_l)} \supseteq S_{(*^{l-2}, h_{l-1}, h_l)} \supseteq \dots \supseteq S_{(h_1, h_2, \dots, h_l)} = \{(h_1, h_2, \dots, h_l)\}$. Suppose $\mathcal{X}(P, (*^{l-j}, h_{l-j+1}, \#)) > 0$ for every j with $2 \leq j \leq l$ and $\mathcal{X}(P, (*^{l-1}, h_l)) > 0$. Then for every j with $2 \leq j \leq l$ we have

$$\frac{|\mathcal{V}_{(*^{l-j}, h_{l-j+1}, h_{l-j+2}, \dots, h_l)}|}{|\mathcal{V}_{(*^{l-j+1}, h_{l-j+2}, h_{l-j+3}, \dots, h_l)}|} = \frac{1}{\mathcal{X}(P, (*^{l-j+1}, \#))}.$$

Proof. The key idea of the proof is to establish the following fact:

Claim: Fix i with $1 \leq i \leq l-1$. Fix $a_1, a_2 \in A_i$ with $\mathcal{X}(P, (*^{i-1}, a_1, \#)) = \mathcal{X}(P, (*^{i-1}, a_2, \#))$. Then

$$|\mathcal{V}_{(*^{i-1}, a_1, h_{i+1}, h_{i+2}, \dots, h_l)}| = |\mathcal{V}_{(*^{i-1}, a_2, h_{i+1}, h_{i+2}, \dots, h_l)}|.$$

Moreover, if $\mathcal{X}(P, (*^{i-1}, a, \#)) = 0$ then $|\mathcal{V}_{(*^{i-1}, a, h_{i+1}, h_{i+2}, \dots, h_l)}| = 0$.

Proof. In case $\mathcal{X}(P, (*^{i-1}, a_1, \#)) = \mathcal{X}(P, (*^{i-1}, a_2, \#)) = 0$, according to remark 6.3, for every population $Q \in [P]_{\mathcal{A}}$ we have $\mathcal{X}(Q, (*^{i-1}, a_2, \#)) = 0$. But then $|\mathcal{V}_{(*^{i-1}, a_1, \#)}| = |\mathcal{V}_{(*^{i-1}, a_2, \#)}| = 0$. Thanks to our extra assumption, the only other possibility is that $\mathcal{X}(P, (*^{i-1}, a_1, \#)) = \mathcal{X}(P, (*^{i-1}, a_2, \#)) = 1$. Consider the map $F : [P]_{\mathcal{A}} \rightarrow [P]_{\mathcal{A}}$ defined as follows: Given a population $Q \in [P]_{\mathcal{A}}$, let $\mathbf{x}$ and $\mathbf{y}$ denote the unique individuals of Q fitting the schemata $(*^{i-1}, a_1, \#)$ and $(*^{i-1}, a_2, \#)$ respectively (remember that for every population $Q \in [P]_{\mathcal{A}}$ we have $\mathcal{X}(Q, (*^{i-1}, a_1, \#)) = \mathcal{X}(Q, (*^{i-1}, a_2, \#)) = 1$). Let now $F(Q)$ denote the population obtained from Q by replacing the individuals $\mathbf{x}$ and $\mathbf{y}$ with the corresponding pair of children obtained by swapping the i^{th} alleles of $\mathbf{x}$ and $\mathbf{y}$ and leaving the rest of the individuals of Q as they were. Notice that $F \circ F = \mathbf{1}_{[P]_{\mathcal{A}}}$ where $\mathbf{1}_{[P]_{\mathcal{A}}} : [P]_{\mathcal{A}} \rightarrow [P]_{\mathcal{A}}$ denotes the identity map. In particular, this tells us that F is bijective. Moreover, due to the way F is defined, we immediately get

$$F(\mathcal{V}_{(*^{i-1}, a_1, h_{i+1}, h_{i+2}, \dots, h_l)}) \subseteq \mathcal{V}_{(*^{i-1}, a_2, h_{i+1}, h_{i+2}, \dots, h_l)}$$

and, likewise,

$$F(\mathcal{V}_{(*^{i-1}, a_2, h_{i+1}, h_{i+2}, \dots, h_l)}) \subseteq \mathcal{V}_{(*^{i-1}, a_1, h_{i+1}, h_{i+2}, \dots, h_l)}$$

which implies the desired assertion. $\square$

Now observe that $S_{(*^{l-j+1}, h_{l-j+2}, h_{l-j+3}, \dots, h_l)} = \bigcup_{a \in A_{l-j+1}} S_{(*^{l-j}, a, h_{l-j+2}, \dots, h_l)}$ so that $\mathcal{V}_{(*^{l-j+1}, h_{l-j+2}, h_{l-j+3}, \dots, h_l)} = \bigcup_{a \in A_{l-j+1}} \mathcal{V}_{(*^{l-j}, a, h_{l-j+2}, \dots, h_l)}$. Moreover, according to the claim above, the only alleles $a \in A_{l-j+1}$ which contribute to the disjoint union above are such that $\mathcal{X}(P, (*^{l-j}, a, \#)) = 1$. Furthermore, according to the claim, all the constituents corresponding to such alleles have the same number of elements. Evidently, $|\{a \mid \mathcal{X}(P, (*^{l-j}, a, \#)) = 1\}| = \mathcal{X}(P, (*^{l-j+1}, \#))$ (remember that we assumed

that all of the alleles corresponding to the same gene in P are distinct). Combining the deductions above together with the assumptions of the lemma we obtain

$$|\mathcal{V}_{(*^{l-j+1}, h_{l-j+2}, h_{l-j+3}, \dots, h_l)}| = \mathcal{X}(P, (*^{l-j+1}, \#)) \cdot |\mathcal{V}_{(*^{l-j}, h_{l-j+1}, h_{l-j+2}, \dots, h_l)}|.$$

Notice that $|\mathcal{V}_{(*^{l-j}, h_{l-j+1}, h_{l-j+2}, \dots, h_l)}| \neq 0$. (If not, let j denote the minimal positive integer between 1 and l such that $|\mathcal{V}_{(*^{l-j}, h_{l-j+1}, h_{l-j+2}, \dots, h_l)}| = 0$. Notice that $j \neq 1$ since by our assumption $\mathcal{X}(P, (*^{l-1}, h_l)) > 0$ which means that there exists an individual of P fitting the schema $(*^{l-1}, h_l)$. Even if this individual is not the first one in the population, since our algorithm, $\mathcal{A}$, is regular (see definition 6.3) it follows that $\mathcal{V}_{(*^{l-1}, h_l)} > 0$. So $j \neq 1$. But now simply observe that we get

$$\begin{aligned} |\mathcal{V}_{(*^{l-j+1}, h_{l-j+2}, h_{l-j+3}, \dots, h_l)}| &= \mathcal{X}(P, (*^{l-j+1}, \#)) \cdot |\mathcal{V}_{(*^{l-j}, h_{l-j+1}, h_{l-j+2}, \dots, h_l)}| = \\ &= \mathcal{X}(P, (*^{l-j+1}, \#)) \cdot 0 = 0 \end{aligned}$$

which contradicts the minimality of j .) We now immediately obtain the equation

$$\frac{|\mathcal{V}_{(*^{l-j}, h_{l-j+1}, h_{l-j+2}, \dots, h_l)}|}{|\mathcal{V}_{(*^{l-j+1}, h_{l-j+2}, h_{l-j+3}, \dots, h_l)}|} = \frac{1}{\mathcal{X}(P, (*^{l-j+1}, \#))}$$

which establishes the lemma. $\square$

We now combine corollary 6.4 with lemma 6.5 to obtain the following

Lemma 6.6. Let $\mathcal{A}$ denote a fully shuffling variable length genetic algorithm. Fix an initial population P such that for any schema H of the form $H = (^{i-1}, x_i, \#)$ or $H = (*^{i-1}, x_i)$ we have $\mathcal{X}(P, H) = 1$ or $\mathcal{X}(P, H) = 0$. Fix an individual $\mathbf{h} = (h_1, h_2, \dots, h_l) \in \Omega$. Then we have*

$$\lim_{t \rightarrow \infty} \Phi(h_1, h_2, \dots, h_l) = \frac{\mathcal{X}(P, (*^{l-1}, h_l))}{m} \cdot \prod_{i=1}^{l-1} \frac{1}{\mathcal{X}(P, (*^i, \#))} = \frac{1}{m} \cdot \prod_{i=1}^{l-1} \frac{1}{\mathcal{X}(P, (*^i, \#))}.$$

To obtain the general case, suppose we are given an initial population P (in which 2 different individuals may have the same allele in the i^{th} position). For every i between 1 and N consider the (relabelling) function $f_i : \{1, 2, \dots, m\} \rightarrow A_i \cup \{0\}$ sending $j \in \{1, 2, \dots, m\}$ into the i^{th} allele of the j^{th} individual of P if the length of the j^{th} individual of P is at least i and to 0 if the length of the j^{th} individual of P is less than i . Consider a new variable length genetic algorithm with the search space $\Theta = \bigcup_{n=1}^N \prod_{i=1}^n D_i$ where $D_i = f_i^{-1}(A_i \cup \{0\})$ (recall that m is the number of individuals in the population P). Notice that the relabelling functions induce a “projection” function $\varphi : \Theta^m \rightarrow \Omega^m$ which sends a given population Q consisting of the individuals from Θ to the population $\varphi(Q) \in \Omega^m$ such that the i^{th} allele of the j^{th} individual of $\varphi(Q)$ is $f_i(j)$ as long as $f_i(j) \neq 0$ (0 does not correspond to any allele at all and serves as the terminating symbol) and the length of the j^{th} individual of Q is the largest i such that $f_k(j) \neq 0$ for every $k \leq i$. Notice that the projection map φ commutes with the elementary step of crossover (partition and the choice of crossover transformations) on a population $Q \in \Theta^m$ and applying the projection afterwards produces the same population as applying the projection first and performing the elementary step of crossover with the same parameters afterwards. This tells us, in particular, that $\varphi([P']_{\mathcal{B}}) = [P]_{\mathcal{A}}$ where $\mathcal{B}$ is the algorithm defined in the same way as $\mathcal{A}$ with the exception that the search space is Θ

instead of Ω and $P' \in \Theta^m$ is the population obtained from P in the following way: The j^{th} individual of P' has the same length as the j^{th} individual of P and every allele of the j^{th} individual of P' is equal to j (note that $\varphi(P') = P$). In view of the observations made above, running the algorithm $\mathcal{A}$ starting with the initial population P is probabilistically equivalent to running the algorithm $\mathcal{B}$ with the initial population P' and applying "projection" φ to every output population to see what the corresponding population is for the algorithm $\mathcal{A}$. Seeing things this way tells us that a given individual $\mathbf{h} = (h_1, h_2, \dots, h_l) \in \Omega$ is encountered in the projected population $\varphi(Q)$ (here Q denotes some population in $[P']_{\mathcal{B}}$ which is encountered during a run starting at P') whenever an individual $\mathbf{x} = (x_1, x_2, \dots, x_l) \in \prod_{i=1}^l f_i^{-1}(\{h_i\})$ is encountered in Q . Thereby, the limiting frequency of occurrence of the individual $\mathbf{h} = (h_1, h_2, \dots, h_l)$,

$$\begin{aligned} \lim_{t \rightarrow \infty} \Phi(h_1, h_2, \dots, h_l) &= \lim_{t \rightarrow \infty} \sum_{x_i \in f_i^{-1}(h_i) \text{ where } 1 \leq i \leq l} \Phi(x_1, x_2, \dots, x_l) = \\ &= \sum_{x_i \in f_i^{-1}(h_i) \text{ where } 1 \leq i \leq l} \lim_{t \rightarrow \infty} \Phi(x_1, x_2, \dots, x_l). \end{aligned}$$

Notice that lemma 6.6 applies to the algorithm $\mathcal{B}$ and we immediately obtain

$$\lim_{t \rightarrow \infty} \Phi(x_1, x_2, \dots, x_l) = \frac{1}{m} \cdot \prod_{i=1}^{l-1} \frac{1}{\mathcal{X}(P', (*^i, \#))} = \frac{1}{m} \cdot \prod_{i=1}^{l-1} \frac{1}{\mathcal{X}(P, (*^{l-1}, h_l))}.$$

Plugging the right hand side of the last equation into the preceding one and observing that $|f_i^{-1}(h_i)| = \mathcal{X}(P, (*^{i-1}, h_i, \#))$ when $1 \leq i < l$ while $|f_l^{-1}(h_l)| = \mathcal{X}(P, (*^{l-1}, h_l))$, we finally obtain

$$\begin{aligned} \lim_{t \rightarrow \infty} \Phi(h_1, h_2, \dots, h_l) &= \sum_{x_i \in f_i^{-1}(h_i) \text{ where } 1 \leq i \leq l} \frac{1}{m} \cdot \prod_{i=1}^{l-1} \frac{1}{\mathcal{X}(P, (*^i, \#))} = \\ &= \sum_{x_l \in f_l^{-1}(h_l)} \sum_{x_{l-1} \in f_{l-1}^{-1}(h_{l-1})} \dots \sum_{x_1 \in f_1^{-1}(h_1)} \frac{1}{m} \cdot \prod_{i=1}^{l-1} \frac{1}{\mathcal{X}(P, (*^i, \#))} = \\ &= \sum_{x_l \in f_l^{-1}(h_l)} \frac{1}{m} \sum_{x_{l-1} \in f_{l-1}^{-1}(h_{l-1})} \frac{1}{\mathcal{X}(P, (*^{l-1}, \#))} \dots \sum_{x_1 \in f_1^{-1}(h_1)} \frac{1}{\mathcal{X}(P, (*^1, \#))} = \\ &= \frac{\mathcal{X}(P, (*^{l-1}, h_l))}{m} \cdot \prod_{i=1}^{l-1} \frac{\mathcal{X}(P, (*^{i-1}, h_i, \#))}{\mathcal{X}(P, (*^i, \#))} = \frac{\mathcal{X}(P, (*^{l-1}, h_l))}{m} \cdot \prod_{i=1}^{l-1} \frac{\frac{\mathcal{X}(P, (*^{i-1}, h_i, \#))}{m}}{\frac{\mathcal{X}(P, (*^i, \#))}{m}}. \end{aligned}$$

Recall from definition 6.2 that the ratios $\frac{\mathcal{X}(P, (*^{i-1}, h_i, \#))}{m}$ and $\frac{\mathcal{X}(P, (*^i, \#))}{m}$ are denoted by $\Phi(S_{(*^{i-1}, h_i, \#)}, P, 0)$ and $\Phi(S_{(*^i, \#)}, P, 0)$ respectively so that we can write

$$\lim_{t \rightarrow \infty} \Phi(h_1, h_2, \dots, h_l) = \Phi(S_{(*^{l-1}, h_l)}, P, 1) \cdot \prod_{i=1}^{l-1} \frac{\Phi(S_{(*^{i-1}, h_i, \#)}, P, 1)}{\Phi(S_{(*^i, \#)}, P, 1)}$$

which is precisely the version of Geiringer's theorem obtained in (Poli et al., 2002).

In a forthcoming paper we shall use corollaries 6.4 and 6.1 to derive a similar schema-based version of Geiringer theorem for non-linear GP with homologous crossover based on Poli's schemata (see, for instance, (Poli, 2000) for the notion of Poli's schemata).

7 What Does the Generalized Geiringer Theorem Say When Mutation Is Present?

It turns out, that whenever the mutation rate is positive (regardless of how small it is), the stationary distribution of the irreducible Markov chain for most evolutionary search algorithms in the absence of selection is uniform. This is, actually, almost an immediate corollary of the generalized Geiringer theorem (theorem 5.5). A less general Geiringer-like theorem dealing with positive mutation was obtained in (Spears, 2000).

Definition 7.1. We define mutation to be the elementary step of type 2 reproduction which is associated with the evolutionary 3-tuple of the form $(\Omega, \mathcal{M}, p, \wp_m)$ (see definition 3.4) where $\mathcal{M}$ is a family of unary operations, i. e. simply a family of functions on Ω . $\wp_m$ is the trivial probability distribution which assigns probability one to the partition of Ω consisting of one-element subsets. We shall, therefore, omit $\wp_m$ from the description and write $(\Omega, \mathcal{M}, p)$ instead of $(\Omega, \mathcal{M}, p, \wp_m)$.

In case when $\mathcal{M}$ consists entirely of bijections, $1_\Omega \in \mathcal{M}$ and $p(1_\Omega) > 0$ we say that mutation is invertible.

An invertible mutation step for which there is an $\mathbf{x}$ such that for every $\mathbf{y} \in \Omega$ there exists a transformation $T_{\mathbf{y}} \in \mathcal{M}$ with $T_{\mathbf{y}}(\mathbf{x}) = \mathbf{y}$ and $p(T_{\mathbf{y}}) > 0$ is said to be mixing.

Not every evolutionary search algorithm is equipped with a mixing and invertible mutation step, however, the main purpose of the mutation step is to make the algorithm ergodic in the sense of the following definition:

Definition 7.2. We say that a given bijective and self-transient (see definition 5.2) sub-algorithm $\mathcal{A}$ is ergodic if for some $P \in \Omega^m$ we have $[P]_{\mathcal{A}} = \Omega^m$ (notice that, according to proposition 5.4, if this happens for some population P then the same is true for all populations P).

Corollary 7.1. Let $\mathcal{A}$ denote an ergodic bijective and self-transient sub-algorithm. Then the Markov chain associated with the sub-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 $\lim_{t \rightarrow \infty} \Phi(\{\mathbf{x}\}, P, t) = \frac{1}{|\Omega|}$ (see definition 6.2 for the meaning of $\Phi(\{\mathbf{x}\}, P, t)$).

Proof. Since $[P]_{\mathcal{A}} = \Omega^m$ for some $P \in \Omega^m$, and collection $\{[Q]_{\mathcal{A}} \mid Q \in \Omega^m\}$ forms a partition of Ω^m (see proposition 5.4) we conclude that for every population $P \in \Omega^m$ we have $[P]_{\mathcal{A}} = \Omega^m$ and the fact that the Markov chain associated with the sub-algorithm $\mathcal{A}$ is irreducible and that its stationary distribution is uniform follows immediately from theorem 5.5. Moreover, according to corollary 6.2, $\lim_{t \rightarrow \infty} \Phi(\{\mathbf{x}\}, P, t)$ is simply the proportion of the total number of populations in Ω^m whose 1st individual is $\mathbf{x}$ out of the total number of populations in Ω^m . There are totally $|\Omega|^m$ populations in Ω^m (since there are $|\Omega|$ independent ways to choose an individual out of Ω and we do it m times), and, among these, $|\Omega|^{m-1}$ populations have their 1st individual equal to $\mathbf{x}$ (similar reason: the first individual is fixed, while the rest of the individuals are selected independently from Ω $m-1$ times). Thereby, we finally obtain $\lim_{t \rightarrow \infty} \Phi(\{\mathbf{x}\}, P, t) = \frac{|\Omega|^{m-1}}{|\Omega|^m} = \frac{1}{|\Omega|}$. $\square$

A mutation elementary step is particularly nice when it is ergodic:

Corollary 7.2. Suppose a given bijective and self-transient evolutionary sub-algorithm $\mathcal{A}$ involves a mixing mutation step (see definition 7.1). Then $\mathcal{A}$ is itself an ergodic bijective and self-transient algorithm in the sense of definition 7.2.

Proof. Let $\mathcal{A}$ be defined by a cycle $C = \{s_n\}_{n=1}^l$ where one of the elementary steps, say s_i for some $1 \leq i \leq l$, is an ergodic mutation step. According to definition 7.1, we can find an individual $\mathbf{x} \in \Omega$ such that for every $\mathbf{y} \in \Omega$ there exists a transformation $T_{\mathbf{y}} \in \mathcal{M}$ with $T_{\mathbf{y}}(\mathbf{x}) = \mathbf{y}$ and $p(T_{\mathbf{y}}) > 0$. Let $P = (\mathbf{x}, \mathbf{x}, \dots, \mathbf{x})^T$, i.e., $P \in \Omega^m$ is simply the population all of whose individuals are the same and equal to $\mathbf{x}$. It suffices to show that any given population $Q = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m)$ can be obtained from P in finitely many steps (then any such $Q \in [P]_{\mathcal{A}}$). We show that this can be done in one cycle: Indeed, since $\mathcal{A}$ is a bijective and self-transient algorithm, for all elementary steps s_j for $j \neq i$ with some positive probability we can leave the input population unchanged. Now, for every j with $1 \leq j \leq m$ choose a transformation T_j such that $T_j(\mathbf{x}) = \mathbf{y}_j$ and $p(T_j) > 0$. Observe that if the choices above are made and one starts the algorithm with the initial population P then at the end of the 1st cycle one ends up with the population $(T_1(\mathbf{x}), T_2(\mathbf{x}), \dots, T_m(\mathbf{x}))^T = (y_1, y_2, \dots, y_m)^T = Q$. This shows that $Q \in [P]_{\mathcal{A}}$ and completes the argument. $\square$

Example 7.1. (Classical Genetic Algorithm) Continuing with example 6.2, recall that $\mathcal{S}_{A_i}$ denotes the group of permutations of A_i . (See example 2.1. Here we only need to use the fact that $\mathcal{S}_{A_i}$ consists of bijections of A_i onto itself.) Fix a subset $M \subseteq \{1, 2, \dots, n\}$ consisting of the elements $i_1 < i_2 < \dots < i_{|M|}$. Now fix an element $\vec{\pi} = (\pi_{i_1}, \pi_{i_2}, \dots, \pi_{i_{|M|}}) \in \mathcal{S}_{A_{i_1}} \times \mathcal{S}_{A_{i_2}} \times \dots \times \mathcal{S}_{A_{i_{|M|}}}$. Finally, define a mutation (asexual reproduction) transformation $T_{M\vec{\pi}} : \Omega \rightarrow \Omega$ as follows:

$$T_{M\vec{\pi}}(\mathbf{a}) = (x_1, x_2, \dots, x_i, \dots, x_n)$$

$$\text{where } x_i = \begin{cases} \pi_i(a_i) & \text{if } i \in M \\ a_i & \text{otherwise} \end{cases}$$

Denote by $\mathcal{M} = \{T_{M\vec{\pi}} \mid M \subseteq \{1, 2, \dots, n\}, \vec{\pi} \in \mathcal{S}_{A_{i_1}} \times \mathcal{S}_{A_{i_2}} \times \dots \times \mathcal{S}_{A_{i_{|M|}}}\}$. Now let s_3 be the elementary step of type reproduction (asexual reproduction or mutation) corresponding to the reproduction 1-tuple $\Omega_{\text{mutation}} = (\Omega, \mathcal{M}, \mu, \delta_m)$. Notice that for this reproduction 1-tuple, since the transformations in $\mathcal{M}$ are unary, $|\mathcal{E}_{(1)}^m| = 1$ (it consists of the unique partition of the set consisting of m elements into 1-element subsets). This forces δ_m to be the trivial probability distribution which assigns probability 1 to this unique element. The probability distribution μ on $\mathcal{M}$ can be chosen in a number of different ways. For example, the traditional choice would be to construct μ in such a way that every allele a_i in a given individual (element of Ω) has a small probability ϵ with $0 < \epsilon < 1$ of being mutated (replaced by a different allele corresponding to the same gene). With probability $1 - \epsilon$ the allele stays the same. The replacement allele is usually chosen uniformly from the set of the remaining allowable alleles $A_i - \{a_i\}$. This construction can be carried out as follows: for every i with $1 \leq i \leq n$ let $\mathcal{B}_{A_i} = \{\pi \mid \pi \in \mathcal{S}_{A_i}, \pi \text{ has no fixed points}\}$. Let ν_i denote the uniform probability distribution on the set $\mathcal{B}_{A_i}$. Now let μ_i denote the probability distribution on $\mathcal{B}_{A_i} \cup \{1\}$ ($1 \in \mathcal{S}_{A_i}$ is the identity permutation) defined as follows:

$$\mu_i(\pi) = \begin{cases} 1 - \epsilon & \text{if } \pi = 1 \\ \epsilon \nu(\pi) & \text{otherwise.} \end{cases}$$

Finally let μ be the probability distribution on $\mathcal{M}$ which selects a given transformation

$T_{M\vec{\pi}}$ with probability

$$\mu(T_{M\vec{\pi}}) = \begin{cases} \prod_{i \notin M} \mu_i(\mathbf{1}) \prod_{i \in M} \mu_i(\pi_i) = (1 - \epsilon)^{|\bar{M}|} \epsilon^{|M|} \prod_{i \in M} \nu_i(\pi_i) \\ \text{if } \vec{\pi} \in \mathcal{B}_{A_{i_1}} \times \mathcal{B}_{A_{i_2}} \times \dots \times \mathcal{B}_{A_{i_{|M|}}} \text{ for } i_1 < i_2 < \dots < i_{|M|} \in M \\ 0 \text{ otherwise.} \end{cases}$$

Fix a and $b \in A_i$ and let $\mathcal{B}_{A_i}^{a \rightarrow b} = \{\pi \mid \pi \in \mathcal{B}_{A_i} \text{ and } \pi(a) = b\}$. Notice that $|\mathcal{B}_{A_i}| = (|A_i| - 1)|\mathcal{B}_{A_i}^{a \rightarrow b}|$. (Indeed, $\mathcal{B}_{A_i} = \bigcup_{b \in A_i, b \neq a} \mathcal{B}_{A_i}^{a \rightarrow b}$. But for $b_1 \neq b_2$ we have $\mathcal{B}_{A_i}^{a \rightarrow b_1} \cap \mathcal{B}_{A_i}^{a \rightarrow b_2} = \emptyset$, and, by symmetry, for every $b_1, b_2 \in A_i$ we have $|\mathcal{B}_{A_i}^{a \rightarrow b_1}| = |\mathcal{B}_{A_i}^{a \rightarrow b_2}|$. So, there are totally $|A_i| - 1$ pairwise disjoint sets of the form $\mathcal{B}_{A_i}^{a \rightarrow b}$, all of equal size, involved in the union above, so that we finally get $|\mathcal{B}_{A_i}| = (|A_i| - 1)|\mathcal{B}_{A_i}^{a \rightarrow b}|$.) Equivalently, we can write $\frac{|\mathcal{B}_{A_i}^{a \rightarrow b}|}{|\mathcal{B}_{A_i}|} = \frac{1}{|A_i| - 1}$. Now, fix an individual $\mathbf{a} = (a_1, a_2, \dots, a_n) \in \Omega$. Fix a set $M \subseteq \{1, 2, \dots, n\}$ and for every $i \in M$ fix $x_i \in A_i$ with $x_i \neq a_i$. Let $\mathbf{x} = (x_{i_1}, x_{i_2}, \dots, x_{i_{|M|}}) \in \prod_{i \in M} A_i$ and $\mathcal{B}_M^{\mathbf{x}} = \{\vec{\pi} \mid \vec{\pi} = (\pi_1, \pi_2, \dots, \pi_n) \in \mathcal{B}_{A_{i_1}} \times \mathcal{B}_{A_{i_2}} \times \dots \times \mathcal{B}_{A_{i_{|M|}}} \text{ for } i_1 < i_2 < \dots < i_{|M|} \in M \text{ and } \forall j \pi_{i_j}(a_i) = x_i\} = \prod_{i=1}^n C_i$ where

$$C_i = \begin{cases} \mathcal{B}_{A_i}^{a_i \rightarrow x_i} & \text{if } i \in M \\ \{\mathbf{1}\} & \text{otherwise.} \end{cases}$$

Now observe that the probability that after mutation the i^{th} allele of $\mathbf{a}$ stays the same if $i \notin M$ and a_i is replaced with $x_i \neq a_i$ if $i \in M$ is $\sum_{\vec{\pi} \in \mathcal{B}_M^{\mathbf{x}}} \mu(T_{M\vec{\pi}}) = (1 - \epsilon)^{|\bar{M}|} \epsilon^{|M|} \sum_{\vec{\pi} \in \mathcal{B}_M^{\mathbf{x}}} \prod_{i \in M} \nu_i(\pi_i) = (1 - \epsilon)^{|\bar{M}|} \epsilon^{|M|} \prod_{i \in M} \sum_{\pi_i \in \mathcal{B}_{A_i}^{a_i \rightarrow x_i}} \nu_i(\pi_i) = (1 - \epsilon)^{|\bar{M}|} \epsilon^{|M|} \prod_{i \in M} \frac{|\mathcal{B}_{A_i}^{a_i \rightarrow x_i}|}{|\mathcal{B}_{A_i}|} = (1 - \epsilon)^{|\bar{M}|} \epsilon^{|M|} \prod_{i \in M} \frac{1}{|A_i| - 1}$. The elementary step of mutation described above is mixing in the sense of definition 7.1. Indeed, every one of the transformations $T_{M\vec{\pi}} \in \mathcal{M}$ is bijective: $T_{M\vec{\pi}}^{-1} = T_{M\vec{\pi}^{-1}}$ where $\forall \vec{\pi} = (\pi_1, \pi_2, \dots, \pi_{i_{|M|}}) \vec{\pi}^{-1} = (\pi_1^{-1}, \pi_2^{-1}, \dots, \pi_{i_{|M|}}^{-1})$. Moreover, by choosing $M = \emptyset$ we obtain $T_{M\vec{\pi}} = \mathbf{1}_\Omega$ so that this mutation step is, indeed, invertible in the sense of definition 7.1. To see that this mutation step is mixing, notice that given any $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\mathbf{y} = (y_1, y_2, \dots, y_n) \in \Omega^m$ we can let $M = \{i \mid x_i \neq y_i\}$, then select permutations $\pi_{i_1} \in \mathcal{S}_{A_{i_1}}$, $\pi_{i_2} \in \mathcal{S}_{A_{i_2}}, \dots, \pi_{i_{|M|}} \in \mathcal{S}_{A_{i_{|M|}}}$ such that $\pi_{i_j}(x_{i_j}) = y_{i_j}$ so that $T_{M,\vec{\pi}}(\mathbf{x}) = \mathbf{y}$. It now follows from corollary 7.2 that the Markov chain associated with the sub-algorithm $\mathcal{A}$ obtained from the classical genetic algorithm by omitting selection (meaning that $\mathcal{A}$ is defined by the cycle $C = (s_2, s_3)$) is irreducible and has uniform stationary distribution. This fact is independent of how small the mutation rate is as long as the rate is positive. A much more detailed investigation of the properties of crossover and mutation for the case of the classical genetic algorithm was carried out in (Vose, 1999) (explicit results are obtained for arbitrary distributions over arbitrary crossover and mutation masks). The primary virtue of the results such as corollary 7.2 is that they apply to a rather wide class of evolutionary algorithms.

Example 7.2. Continuing with example 6.3, to define mutation we make use of the set $\prod_{n=1}^N \mathcal{P}(I_n)$ as well (recall from example 6.3 that $I_n = \{1, 2, \dots, n\}$ and $\mathcal{P}(I_n)$ denotes the set of all subsets of I_n). There is one more complication involved here: Since we may be modelling linear GP rather than just a variable length GA, if a given individual $\mathbf{x} = (x_1, x_2, \dots, x_l)$ has length l , then it might be the case that not all of the alleles in A_l can replace x_l . We, therefore, assume that for every i $A_i = B_i \cup C_i$. (We do not assume

that B_i and C_i are disjoint. It can be the case that $B_i \cap C_i \neq \emptyset$.) The alleles in B_i are said to be internal, while these in C_i are named terminal.⁶ Recall from example 6.2 that $\mathcal{B}_K$ denotes the set of all permutations of a given set K which have no fixed points. Recall also the transformation $T_{M_n, \vec{\pi}} : (\prod_{i=1}^{n-1} B_i) \times C_n \rightarrow (\prod_{i=1}^n B_i) \times C_n$ defined as follows:

$$\forall \mathbf{a} = (a_1, a_2, \dots, a_n) \in (\prod_{i=1}^{n-1} B_i) \times C_n \text{ let } T_{M_n, \vec{\pi}}(\mathbf{a}) = (x_1, x_2, \dots, x_i, \dots, x_n)$$

$$\text{where } x_i = \begin{cases} \pi_i(a_i) & \text{if } i \in M_n \\ a_i & \text{otherwise} \end{cases}$$

(Remember that $(\prod_{i=1}^{n-1} A_i) \times C_n$ was the search space in the previous example.) Now fix an element $\mathbf{M} = (M_1, M_2, \dots, M_N) \in \prod_{n=1}^N \mathcal{P}(I_n)$ and for every n with $1 \leq n \leq N$ fix an $|M_n|$ -tuple of permutations $\vec{\pi}_{M_n} = (\pi_{i_1}, \pi_{i_2}, \dots, \pi_{i_{|M_n|}}) \in (\prod_{j=1}^{|M_n|-1} \mathcal{B}_{B_{i_j}}) \times \mathcal{B}_{C_{i_{|M_n|}}}$ where $i_j \in M_n$ for every j with $1 \leq j \leq |M_n|$. Let $\vec{\omega} = (\vec{\pi}_{M_1}, \vec{\pi}_{M_2}, \dots, \vec{\pi}_{M_N})$. Finally define the mutation transformation $U_{\mathbf{M}\vec{\omega}} : \Omega \rightarrow \Omega$ as follows:

$$\text{Whenever } \mathbf{a} = (a_1, a_2, \dots, a_n) \in (\prod_{i=1}^{n-1} B_i) \times C_n$$

we let $U_{\mathbf{M}\vec{\omega}}(\mathbf{a}) = T_{M_n, \vec{\pi}}(\mathbf{a})$. We denote by $\mathcal{M}_{\text{variable}}$ the family of all such mutation transformations. Formally,

$$\mathcal{M}_{\text{variable}} = \{U_{\mathbf{M}\vec{\omega}} \mid \mathbf{M} = (M_1, M_2, \dots, M_N) \in \prod_{n=1}^N \mathcal{P}(I_n)$$

$$\vec{\omega} = (\vec{\pi}_{M_1}, \vec{\pi}_{M_2}, \dots, \vec{\pi}_{M_N}) \text{ where } \vec{\pi}_{M_n} = (\pi_{i_1}, \pi_{i_2}, \dots, \pi_{i_{|M_n|}}) \in (\prod_{j=1}^{|M_n|-1} \mathcal{B}_{B_{i_j}}) \times \mathcal{B}_{C_{i_{|M_n|}}}\}.$$

A simple way to construct a probability distribution on the set $\mathcal{M}_{\text{variable}}$ is to construct a sequence of probability distributions μ_n on the sets $\mathcal{M}_n = \{T_{M_n, \vec{\pi}_n} \mid M_n = \{i_1, i_2, \dots, i_{|M_n|}\} \subseteq I_n \text{ and } \vec{\pi}_n \in (\prod_{j=1}^{|M_n|-1} \mathcal{B}_{B_{i_j}}) \times \mathcal{B}_{C_{i_{|M_n|}}}\}$. Now one can simply choose a transformation $U_{\mathbf{M}\vec{\omega}}$ by choosing every one of its components independently with respect to the distributions μ_n . For example, in order to obtain an algorithm which is probabilistically equivalent to the classical case, i.e., every allele is mutated with a small probability ϵ and remains unchanged with probability $1 - \epsilon$, all we have to do is to choose every μ_i in the same way as it was done in the previous example.

It turns out that the family of mutation transformations $\mathcal{M}_{\text{variable}}$ is not mixing in the sense of definition 7.1. One can still use the generalized Geiringer theorem and corollary 6.4 to compute the limiting distribution of the Markov chain associated with the sub-algorithm $\mathcal{A}$ obtained from the linear GP with the standard type of mutation (the one described first in example 7.2) by omitting selection, started with an initial population $P \in \Omega^m$. We have already seen that the crossover elementary step satisfies

⁶In view of this modification, the real search space for linear GP with homologous crossover is actually the set $\Omega = \bigcup_{n=1}^N ((\prod_{i=1}^{n-1} B_i) \times C_n)$. The restriction of any crossover transformation to this space is well-defined and does not affect the theoretical analysis presented in this paper in any way. We do need to take this into account when we consider mutation, though.

definition 5.2. To see that mutation is invertible, observe that every one of the mutation transformations from the family $\mathcal{M}_{\text{variable}}$ is simply a disjoint union of bijections (the mutation transformations of the classical genetic algorithm considered in example 7.1 everyone of which was shown to be bijective) defined on disjoint sets. As usual, one has to prove an appropriate orbit-description lemma: Again we state the orbit description lemma in the language of Holland schema in a similar style as it was done in the previous section. We, therefore, introduce one more type of schema below:

Definition 7.3. The schema $H = (*^l)$ represents the subset

$$S_H = \{\mathbf{x} = (x_1, x_2, \dots, x_l) \mid x_i \in B_i \text{ for } 1 \leq i \leq l-1 \text{ and } x_l \in C_l\} = \left(\prod_{i=1}^{l-1} B_i\right) \times C_l.$$

In words, S_H is simply the subset of all individuals whose length is exactly l .

Fix a population P . Recall from example 6.3 that we write $\mathcal{V}_H$ instead of $\mathcal{V}_{S_H}$ to denote the set of all populations in $[P]_{\mathcal{A}}$ whose first individual is a member of S_H where S_H is the subset of the search space Ω determined by the schema H . We are now ready to state the desired orbit-description lemma which is very similar, but simpler than lemma 6.5:

Lemma 7.3. Suppose the algorithm $\mathcal{A}$ defined by the cycle (s_2, s_3) is fully shuffling in the sense of definition 6.8 and every transformation in the family $\mathcal{M}_{\text{variable}}$ has a positive probability of being chosen. Fix a population P and an individual $\mathbf{h} = \{h_1, h_2, \dots, h_l\} \in \Omega$. Then for every i with $1 \leq i \leq l-2$ and for any choice of a and b in B_{i+1} we have $|\mathcal{V}_{(*^i, a, h_{i+2}, h_{i+3}, \dots, h_l)}| = |\mathcal{V}_{(*^i, b, h_{i+2}, h_{i+3}, \dots, h_l)}|$. Likewise, for any choice of a and b in C_l we have $|\mathcal{V}_{(*^{l-1}, a)}| = |\mathcal{V}_{(*^{l-1}, b)}|$. Moreover, if the initial population, P contains an individual of length l then for every i with $1 \leq i \leq l-2$ we have $\frac{|\mathcal{V}_{(*^i, h_{i+1}, h_{i+2}, h_{i+3}, \dots, h_l)}|}{|\mathcal{V}_{(*^{i+1}, h_{i+2}, h_{i+3}, \dots, h_l)}|} = \frac{1}{|B_i|}$ while $\frac{|\mathcal{V}_{(*^{l-1}, h_l)}|}{|\mathcal{V}_{(*^l)}|} = \frac{1}{|C_l|}$.

Proof. The argument is very much analogous to the corresponding part of the proof of lemma 6.5. Fix i with $1 \leq i \leq l-1$ and let a and $b \in B_{i+1}$. Now consider the function $F_{a \rightarrow b} : \mathcal{V}_{(*^i, a, h_{i+2}, h_{i+3}, \dots, h_l)} \rightarrow \mathcal{V}_{(*^i, b, h_{i+2}, h_{i+3}, \dots, h_l)}$ defined as follows: given a population $Q \in \mathcal{V}_{(*^i, a, h_{i+2}, h_{i+3}, \dots, h_l)}$, let $F_{a \rightarrow b}(Q)$ denote the population whose last $m-1$ individuals are the same as those of Q and the first individual of $F_{a \rightarrow b}(Q)$ is obtained from the first individual of Q by performing mutation at position $i+1$ which replaces the $i+1^{\text{st}}$ allele, a with b and leaves the rest of the alleles unchanged. It is easily seen that $F_{a \rightarrow b}(Q) \in [P]_{\mathcal{A}}$, and, evidently, $F_{a \rightarrow b}(Q) \in \mathcal{V}_{(*^i, b, h_{i+2}, h_{i+3}, \dots, h_l)}$ so that the function $F_{a \rightarrow b} : \mathcal{V}_{(*^i, a, h_{i+2}, h_{i+3}, \dots, h_l)} \rightarrow \mathcal{V}_{(*^i, b, h_{i+2}, h_{i+3}, \dots, h_l)}$ is well defined. Moreover, it is easy to check that $F_{b \rightarrow a} : \mathcal{V}_{(*^i, b, h_{i+2}, h_{i+3}, \dots, h_l)} \rightarrow \mathcal{V}_{(*^i, a, h_{i+2}, h_{i+3}, \dots, h_l)}$ is a two-sided inverse of $F_{a \rightarrow b}$ so that $F_{a \rightarrow b}$ is a bijection and the desired conclusion that

$$|\mathcal{V}_{(*^i, a, h_{i+2}, h_{i+3}, \dots, h_l)}| = |\mathcal{V}_{(*^i, b, h_{i+2}, h_{i+3}, \dots, h_l)}|$$

follows at once. The second assertion of the lemma is obtained in a completely analogous way by considering the map $G_{a \rightarrow b} : \mathcal{V}_{(*^{l-1}, a)} \rightarrow \mathcal{V}_{(*^{l-1}, b)}$ which is defined in the corresponding manner. We leave the details for the reader.

Now observe that as long as mutation rate is positive (in other words, every member of the family $\mathcal{M}_{\text{variable}}$ has a positive probability of being chosen) if the initial population P contains an individual of length l then $[P]_{\mathcal{A}}$ contains a population Q_1 whose first individual has length l (the algorithm $\mathcal{A}$ is easily seen to be regular in the sense

of definition 6.3), and by applying an appropriate mutation transformation to the first individual of Q_1 we obtain a population Q whose first individual is $\mathbf{h}$. In particular, this shows that

$$\mathcal{V}_{(*)^l} \supseteq \mathcal{V}_{(*)^{l-1}, h_l} \supseteq \dots \supseteq \mathcal{V}_{(*)^i, h_{i+1}, h_{i+2}, h_{i+3}, \dots, h_l} \supseteq \dots \supseteq \mathcal{V}_{(h_1, h_2, \dots, h_l)} \neq \emptyset.$$

Now simply observe that $\mathcal{V}_{(*)^l} = \bigcup_{a \in C_l} \mathcal{V}_{(*)^{l-1}, a}$ while $\mathcal{V}_{(*)^{i+1}, h_{i+2}, h_{i+3}, \dots, h_l} = \bigcup_{a \in B_{i+1}} \mathcal{V}_{(*)^i, a, h_{i+2}, h_{i+3}, \dots, h_l}$. Since all of the constituents in the unions above are disjoint and are also of equal size (thanks to the first assertion of the lemma which we already proved), it follows that $|\mathcal{V}_{(*)^l}| = |C_l| \cdot |\mathcal{V}_{(*)^{l-1}, h_l}|$ and $|\mathcal{V}_{(*)^{i+1}, h_{i+2}, h_{i+3}, \dots, h_l}| = |B_{i+1}| \cdot |\mathcal{V}_{(*)^i, h_{i+1}, h_{i+2}, h_{i+3}, \dots, h_l}|$ which immediately implies the desired conclusion $\frac{|\mathcal{V}_{(*)^i, h_{i+1}, h_{i+2}, h_{i+3}, \dots, h_l}|}{|\mathcal{V}_{(*)^{i+1}, h_{i+2}, h_{i+3}, \dots, h_l}|} = \frac{1}{|B_i|}$ and $\frac{|\mathcal{V}_{(*)^{l-1}, h_l}|}{|\mathcal{V}_{(*)^l}|} = \frac{1}{|C_l|}$ (denominators are not 0 since the sets were shown to be nonempty). $\square$

The algorithm $\mathcal{A}$ is regular in the sense of definition 6.3 and so we are in a position to apply corollary 6.4 to the nested sequence of subsets

$$\mathcal{V}_{(*)^l} \supseteq \mathcal{V}_{(*)^{l-1}, h_l} \supseteq \dots \supseteq \mathcal{V}_{(*)^i, h_{i+1}, h_{i+2}, h_{i+3}, \dots, h_l} \supseteq \dots \supseteq \mathcal{V}_{(h_1, h_2, \dots, h_l)}.$$

Notice that the function $\mathcal{X}(\square, S_{(*)^i})$ is constant on the equivalence class $[P]_{\mathcal{A}}$ and its value is $\mathcal{X}(P, S_{(*)^i})$ since the number of individuals of a given length does not change when performing the elementary steps of crossover or mutation. This tells us that the expectation of $\mathcal{X}(\square, S_{(*)^i})$ is exactly $\mathcal{X}(P, S_{(*)^i})$ which is simply the number of individuals of length l in the population P . Combining corollary 6.4 with lemma 7.3 and the observation made above we immediately deduce that if the initial population P has an individual of length l then

$$\lim_{t \rightarrow \infty} \Phi(h_1, h_2, \dots, h_l) = \frac{\mathcal{X}(P, S_{(*)^l})}{m} \cdot \prod_{i=1}^{l-1} \frac{1}{|B_i|} \cdot \frac{1}{|C_l|}.$$

As mentioned above, the family of mutation transformations $\mathcal{M}_{\text{variable}}$ is not mixing in the sense of definition 7.1. The simplest way to repair the situation in such cases, is to extend the family $\mathcal{M}_{\text{variable}}$ to include all permutations (bijections) on the search space. Of course, one still needs to select a probability distribution on the set of all permutations $\mathcal{S}_{\Omega}$. There are many possibilities, but the simplest choice seems to be the following: select a probability distribution (α_1, α_2) on the set of two elements (simply 2 nonnegative numbers adding up to 1). Usually we select $\alpha_1 \gg \alpha_2 > 0$. Select a probability distribution, say μ on the family $\mathcal{M}_{\text{variable}}$ (this is likely to be chosen in the same way as it was done above) and a probability distribution η on the collection $\mathcal{S}_{\Omega} - \mathcal{M}_{\text{variable}}$ (usually η is selected to be the uniform distribution). It will be easily shown in example 7.1 that $\mathcal{M}_{\text{variable}} \subseteq \mathcal{S}_{\Omega}$ so that $\mathcal{S}_{\Omega} = \mathcal{M}_{\text{variable}} \cup (\mathcal{S}_{\Omega} - \mathcal{M}_{\text{variable}})$. We now define a probability distribution ν on $\mathcal{S}_{\Omega}$ as follows:

$$\nu(T) = \begin{cases} \alpha_1 \cdot \mu(T) & \text{if } T \in \mathcal{M}_{\text{variable}} \\ \alpha_2 \cdot \eta(T) & \text{if } T \in \mathcal{S}_{\Omega} - \mathcal{M}_{\text{variable}}. \end{cases}$$

One may use the entire family of permutations $\mathcal{S}_{\Omega}$ with a type of probability distribution described above in which case the corresponding elementary step of mutation is invertible and mixing. Indeed, by definition, $\mathcal{S}_{\Omega}$ consists entirely of bijections,

$\nu(\mathbf{1}_\Omega) = \alpha_1 \cdot \mu(\mathbf{1}_\Omega) > 0$ because $\alpha_1 > 0$ by assumption and above we have shown that μ is a probability distribution for another invertible mutation step so that $\mu(\mathbf{1}_\Omega) > 0$. Moreover, by definition of ν (the fact that α_1 and α_2 are positive, η is the uniform distribution on $\mathcal{S}_\Omega - \mathcal{M}_{\text{variable}}$ so that each transformation in $\mathcal{S}_\Omega - \mathcal{M}_{\text{variable}}$ has a positive chance of being chosen and μ assigns a positive probability to every transformation in the family $\mathcal{M}_{\text{variable}}$) ν assigns a positive probability to every permutation. It is a general fact that for all $\mathbf{x}$ and $\mathbf{y} \in \Omega$ there exists a permutation T on Ω with $T(\mathbf{x}) = \mathbf{y}$. We have shown now that the elementary step of mutation with the choices made above is, indeed, mixing.

8 Conclusions and Future Work

In the current paper the following contributions have been made:

1. It has been shown that the stationary distribution of the Markov chain associated with a bijective and self-transient sub-algorithm of a given evolutionary algorithm (see definition 5.2) is unique and it is always uniform on the set of populations $[P]_{\mathcal{A}}$ (see definitions 5.4, and 5.3 as well as proposition 5.4 and theorem 5.5). This fact is rather general and applies to a rather wide class of evolutionary algorithms.
2. Corollaries 6.1 and 6.2, which are stated and proved in section 6, demonstrate how theorem 5.5 generalizes the classical Geiringer theorem as well as the version established in (Poli et al., 2002). The methodology developed in section 6 shall be used in a sequel paper to derive a schema-based version of the Geiringer theorem for nonlinear genetic programming with homologous crossover. This version is based on Poli's schemata (see (Poli, 2000) for a detailed description of Poli's schemata).
3. The extension of the Geiringer theorem (theorem 5.5) for the case of a finite population applies to a wide class of evolutionary algorithms in the presence of mutation. This subject has been discussed and illustrated with examples in section 7.

Remark 8.1. Notice, however, that theorem 5.5 applies only in the absence of a sampling stage. In fact, it is shown in (Mitavskiy and Rowe, 2005), as a corollary of another theorem, that if a given algorithm is determined by a cycle whose last elementary step involves a sampling stage then the stationary distribution of the corresponding algorithm is *never* uniform⁷. Various modifications of this fact will also be established in the forthcoming paper. In particular, it will be shown that results such as corollary 7.1 still hold in the presence of flat-fitness selection despite the fact that the stationary distribution of the corresponding Markov chain is not uniform.

Remark 8.2. The formal mathematical framework in which the version of the Geiringer theorem presented in the current paper is established is essentially different from the framework in which the classical version has been proved. The framework for the classical case was briefly discussed in the introduction. The main difference between the two frameworks, lies in the fact that theorem 5.5 relies heavily on the bijective property of crossover (requirement 1 of definition 5.2). This property is not at all necessary for the infinite population model. The classical Geiringer result applies to the usual 2-parent $\rightarrow$ 1-child crossover under rather mild assumptions (which certainly do not require the crossover transformations to be bijective). Interestingly, the formulas obtained in the finite population case are very similar to those derived in the classical case. The general circumstances under which the two frameworks produce the same answer

⁷As biologists would say, sampling in a finite population induces *drift* away from the uniform distribution.

remains an open question so far, although some progress has already been made. This issue will be addressed in detail in a forthcoming paper.

The proof of Geiringer's theorem (theorem 5.5), which is given in section 2, is based on classical applications of algebra to the theory of Markov processes (see (Rosenthal, 1995) for a detailed introduction to this subject). Incidentally, algebraic notions such as groups and group actions have already been applied to the study of evolutionary algorithms (see, for instance, (Rowe et al., 2002) and (Rowe et al., 2004)). In many cases the eigenvalues of the Markov transition matrix can be computed as the values of characters on certain kinds of permutations (see section 5.2 of (Rosenthal, 1995)). This gives some hope for estimating the rate of convergence towards the uniform stationary distribution in these cases. We leave this subject for future investigation.

9 Acknowledgments

The authors would like to thank the following for their advice and help with this paper: Andreas Blass, Divakar Viswanath, John Holland, Rick Riolo, and the University of Michigan complex systems group. We are also grateful to the anonymous referees for their suggestions, improvements and hints for future work.

References

- Coffey, S. (1999). An applied probabilist's guide to genetic algorithms. Master's thesis, The University of Dublin.
- Dummit, D. and Foote, R. (1991). *Abstract Algebra*. Prentice-Hall.
- Geiringer, H. (1944). On the probability of linkage in mendelian heredity. *Annals of Mathematical Statistics*, 15:25–57.
- Mitavskiy, B. (2003). Comparing evolutionary computation techniques via their representation. In Cantu-Paz, E., et al, editors, *Proceedings of the Genetic and Evolutionary Computation (GECCO) Conference*, volume 1, pages 1196–1209. Springer-Verlag.
- Mitavskiy, B. (2004a). Crossover invariant subsets of the search space for evolutionary algorithms. *Evolutionary Computation*, 12(1):19–46.
- Mitavskiy, B. (2004b). *A mathematical model of evolutionary computation and some consequences*. PhD thesis, University of Michigan.
- Mitavskiy, B. and Rowe, J. E. (2005). Some results about the markov chains associated with GPs and general EAs. *Theoretical Computer Science*. Under submission.
- Poli, R. (2000). Hyperschema theory for GP with one-point crossover, building blocks, and some new results in GA theory. In Poli, R., et al, editors, *Genetic Programming, Proceedings of EuroGP'2000*, pages 163–180. Springer-Verlag.
- Poli, R., Stephens, C., Wright, A., and Rowe, J. (2002). A schema-theory-based extension of Geiringer's theorem for linear GP and variable-length GAs under homologous crossover. In DeJong, K., Poli, R., and Rowe, J. E., editors, *Foundations of Genetic Algorithms*, volume 7, pages 45–62.
- Rabani, Y., Rabinovich, Y., and Sinclair, A. (1995). A computational view of population genetics. In *Annual ACM Symposium on the Theory of Computing*, pages 83–92.

- Rosenthal, J. (1995). Convergence rates for markov chains. *SIAM Review*, 37(3):387–405.
- Rowe, J., Vose, M., and Wright, A. (2002). Group properties of crossover and mutation. *Evolutionary Computation*, 10(2):151–184.
- Rowe, J., Vose, M., and Wright, A. (2004). Structural search spaces and genetic operators. *Evolutionary Computation*, 12(4):461–494.
- Schmitt, L. (2001). Theory of genetic algorithms. *Theoretical Computer Science*, 259:1–61.
- Schmitt, L. (2004). 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. *Theoretical Computer Science*, 310:181–231.
- Sinclair, A. (1992). *Algorithms for Random Generation and Counting: A Markov Chain Approach*. Birkhauser, Boston.
- Spears, W. (2000). The equilibrium and transient behavior of mutation and recombination. In Martin, W. and Spears, W., editors, *Foundations of genetic Algorithms*, volume 6, pages 241–260.
- Vose, M. (1999). *The Simple Genetic Algorithm: Foundations and Theory*. MIT Press.