Sufficient Conditions for Coarse-Graining Evolutionary Dynamics

Keki Burjorjee
DEMO Lab,
Computer Science Department,
Brandeis University, Waltham, MA 02454.
kekib@cs.brandeis.edu

Abstract

Previous theoretical results in the Evolutionary Computation literature only permit analyses of evolutionary dynamics in the *immediate term* — i.e. over a single generation — or in the asymptote of time. There are currently no theoretical results that permit a *principled* analysis of any non-trivial aspect of evolutionary dynamics in the *short term*, i.e. over a small number of generations. In the absence of such analyses we believe that accurate theories of evolutionary adaptation will continue to evade discovery. We describe a technique called coarse-graining which has been widely used in other scientific disciplines to study the emergent phenomena of complex systems. This technique is a promising approach towards the formulation of more principled theories of evolutionary adaptation because, if successfully applied, it permits a principled analysis of evolutionary dynamics across multiple generations. We develop a simple yet powerful abstract framework for studying the dynamics of an infinite population evolutionary algorithm (IPEA). Using this framework we show that the short term dynamics of an IPEA can be coarse-grained if it satisfies certain abstract conditions. We then use this result to argue that the dynamics of an infinite population *genetic* algorithm with uniform crossover and fitness proportional selection can be coarse-grained for at least a small number of generations, provided that the initial population belongs to a particular class of distributions (which includes the uniform distribution), and the fitness function satisfies a relatively *weak* constraint.

1 Introduction

Simple Genetic Algorithms (GAs) have successfully been used to adapt solutions for a wide range of search problems. One of the most important open questions, one might argue *the* most important question, in GA research today is the question of how these algorithms perform adaptation. Complete theories of how the iterated effect of selection and variation on an initially random population drives adaptation have been scarce in the thirty odd years since GAs were first proposed. We know of just one work in which a theory of adaptation has been completely laid out — the seminal work of Holland [9], in which a theory of adaptation, that later came to called the Building Block Hypothesis [8, 11] was proposed.

The Building Block Hypothesis, though well-formulated, is not necessarily accurate. It has been sharply criticized for lacking theoretical justification and experimental results have been published that draw its veracity into question. On the theoretical side, for example, Wright et. al. state in [20], "The various claims about GAs that are traditionally made under the name of the *building block hypothesis* have, to date, no basis in theory, and in some cases, are simply incoherent.". On the experimental side Syswerda has reported in [16] that uniform crossover often outperforms one-point and two-point crossover on the fitness functions that he has studied. Summing up Syswerda's results Fogel remarks [7, p.140] "Generally, uniform crossover yielded better performance than two-point crossover, which in turn yielded better performance than one-point crossover". These results contradict the building block hypothesis because uniform crossover is extremely disruptive of short schemata whereas one and two-point crossover are certainly more likely to conserve short schemata and combine their defining bits in children produced during recombination.

1.1 The Absence of Principled Theories of Adaptation

A very general way of understanding the absence of principled theories of adaptation in the Evolutionary Computation literature is to note firstly that the set of GAs comprise a class of *complex systems*, secondly that adaptation is an *emergent phenomenon* of many (but not all) of the systems in this class, and finally that it is in general very difficult to formulate accurate theories of how or why particular emergent phenomena arise (or do not arise) from the dynamics of the complex systems in some class [2].

The state of any system in a class of complex systems is typically modeled as a tuple of N state variables. The dynamics of the systems in this class is modeled by a system of N parameterized coupled difference or differential equations. A set of parameter values determines a particular dynamical system, and when these values are 'plugged-in' to the equations they describe how the state variables of that system change over time. One (dead-end) approach to understanding how some target phenomenon arises through the dynamics of a class of complex systems is to attempt to solve the set of equations, i.e. to attempt to obtain a closed form formula which, given some parameter values that determine some system and the state of that system at time step $t = 0$ (the initial condition), gives the state of the system at some time-step t. In principle, one can then attempt to understand the target phenomenon by studying the closed-form solution. The equations of a complex system however are always non-linear, and typically unsolvable, so this approach has a low likelihood of success. Another approach is to attempt to glean an understanding of the target phenomenon by *numerically simulating* the dynamics of a well chosen subset of the systems in the class under a well chosen subset of initial conditions. This approach becomes infeasible as N, the number of state variables, becomes large.

The dynamics of an EA can be modeled by a system of coupled nonlinear difference equations called an infinite population model (IPEAs). The time-dependent state variables in such a system is the expected frequencies of individual genomes[1]. The simulation of one generation of an IPEA with N state variables has time complexity $O(N^3)$. An infinite population model of a GA (IPGA) with bitstring genomes of length ℓ has $N = 2^\ell$ state variables. Hence, the time complexity for a 'naive' numeric simulation of an IPGA is $O(8^\ell)$. (See [19, p.36] for a description of how the Fast Walsh Transform can be used to bring this bound down to $O(3^\ell)$.) Even when the Fast Walsh Transform is used, computation time still increases exponentially with ℓ. Vose reported in 1999 that computational concerns force numeric simulation to be limited to cases where $\ell \leq 20$.

Given this discission, we are now in a position to give a more specific reason for the absence of principled theories of adaptation in Evolutionary Computation. Current results in the field only allow one to analyze aspects of IPEA dynamics at the asymptote of time (e.g. [19]) or in the *immediate term*, i.e. over the course of a single generation (e.g. [10, 14]). For IPEAs with large genome sizes and non-trivial fitness functions, there are currently no theoretical results that permit a *principled* analysis of *any* non-trivial aspect of IPEA dynamics in the *short term* (i.e. over a small number of generations). We believe however that a principled analysis of the short term is necessary in order to formulate accurate theories of adaptation. In the absence of results that permit such analysis we believe that accurate theories of adaptation will continue to evade discovery.

[1] Evolutionary Biologists use the word genome to refer to the total hereditary information that is encoded in the DNA of an individual, and we will as well. This same concept is called a genotype in the evolutionary computation literature. Unfortunately this usage creates a terminological inconsistency. The word genotype is used in evolutionary biology to refer to a different concept, one that is similar to what is called a schema in evolutionary computation

1.2 The Promise of Coarse-Graining

Coarse-graining is a technique that has often been used in theoretical physics for studying some target property (e.g. temperature) of many-body systems with very large numbers of state variables (e.g. gases). This technique allows one to reformulate some system of equations with many state variables (called the fine-grained system) as a new system of equations that describes the time-evolution of a smaller set of state variables (the coarse-grained system). This reformulation is done using a surjective (typically non-injective) function, called the partition function, between the fine-grained state space and the coarse-grained state space. States in the fine-grained state space that share some key property (e.g. energy) are projected to a single state in the coarse-grained state space. Metaphorically speaking, just as a stationary light bulb projects the shadow of some moving 3D object onto a flat 2D wall, the partition function projects the changing state of the fine-grained system onto states in the state space of the coarse-grained system.

Three conditions are necessary for a coarse-graining to be successful. Firstly, the dimensionality of the coarse-grained state space should be smaller than the dimensionality of the fine-grained state space (information about the original system will hence be lost). Secondly the dynamics described by the coarse-grained system of equations must 'shadow' the dynamics described by the original system of equations in the sense that if the projected state of the original system at time $t = 0$ is equal to the state of the coarse-grained system at time $t = 0$ then at any other time t, the projected state of the original system should be closely approximated by the state of the coarse-grained system. Thirdly, the coarse-grained system of equations must *not* depend on any of the original state variables.

In the second condition given above, if the approximation is instead an equality then the coarse-graining is said to be *exact*.

Suppose $\mathbf{x}^{(t)}$ and $\mathbf{y}^{(t)}$ are the time-dependent state vectors of some system and a coarse-graining of that system respectively. Now, if the partition function projects $\mathbf{x}^{(0)}$ to $\mathbf{y}^{(0)}$, then, since none of the state variables of the original system are needed to express the dynamics of the coarse-grained system, one can determine how the state of the coarse-grained system $\mathbf{y}^{(t)}$ (the shadow state) changes over time without needing to determine how the state in the fine-grained system $\mathbf{x}^{(t)}$ (the shadowed state) changes. Thus, even though for any t, one might not be able to determine $\mathbf{x}^{(t)}$, one can always be confident that $\mathbf{y}^{(t)}$ is its projection. Therefore, if the number of state variables of the coarse-grained space is small enough, one can numerically determine changes to the shadow state without first needing to determine changes to the shadowed state.

1.3 Inconsistent Use of the Phrase 'Coarse-Graining'

The phrase 'coarse-graining' has, till now, been used in the Evolutionary Computation literature to describe two ways of rewriting the system of equations of an IPGA, neither of which satisfy the three conditions for a successful coarse-graining that we listed above. For instance in [14] the phrase coarse-graining is used to describe how the N equations of an IPGA can be rewritten as a system of $3^{log_2 N} (\geq N)$ equations where the state variables are the frequencies of schemata. Such a rewriting does not qualify as a coarse-graining because the state space of the new system of equations is *bigger* than the state space of the original system. This way of rewriting the system of equations was later called an *embedding* in [15].

However in [15] the phrase coarse-graining was used to refer to a reformulation of the original equations such that each equation in the new system has a form that is similar to the form of the equations in the original system. The new system of equations however is *not* independent of the state variables in the original system; the fitness function used in the new system of equations depends directly on the values of these variables. The authors do note that the "coarse-graining gives rise to a *time dependent* coarse-grained fitness". But this simply obscures the fact that the new system of equations is dependent on the (time dependent) state variables of the original system. A principled analysis of the dynamics of the new state variables *cannot* tractably be 'carried forward' over multiple generations because at each generation the determination of these values relies on the calculation of the values of the state variables of the fine-grained system.

Elsewhere the phrase coarse-graining has been defined as "a collection of subsets of the search space that covers the search space"[6], and more egregiously as "just a function from a genotype set to some other set"[5].

Coarse-graining has a precise meaning in the scientific literature. It is important that the names of useful ideas from other fields be used consistently.

1.4 Formal Approaches to Coarse-Graining

When the state variables in the fine-grained and coarse-grained systems are the frequencies of distributions, i.e. when their values always sum to 1, then the formal concept of compatibility [19, p. 188] is one way of capturing the idea of coarse-graining. Wright et al. show in [20] that there exist conditions under which variation in an IPGA is compatible. They then argue that the same cannot be said for fitness proportional selection "except in the trivial case where fitness is a constant for each schema in a schema [partition]". In other words, except in the case where the constraint on the fitness function is so severe that it renders any coarse-graining result essentially useless. Their argument suggests that it may not *in principle* be possible to show that evolution is compatible for any weaker constraint on the fitness function. This negative claim is cause for concern because it casts doubt on the possibility of obtaining a useful coarse-graining of evolutionary dynamics.

Compatibility however formalizes a very strong notion of coarse-graining. If an operator is compatible then the dynamics induced by its iterated application to some initial distribution can be exactly coarse-grained regardless of the choice of initial distribution. We introduce a concept called *concordance* which formalizes a slightly weaker notion of coarse-graining. If an operator is concordant on some subset U of the set of all distributions then the dynamics induced by the iterated application of the operator to some initial distribution can be exactly coarse-grained *if* the initial distribution is in U.

We use the concept of concordance to show that if variation and the fitness function of an IPEA satisfy certain abstract conditions then, *provided that* the initial population is in U, evolutionary dynamics over at least a small number of generations *can* be coarse-grained.

This abstract result is of no use if it cannot be used in practice. To show that this is *not* the case, we argue that these conditions will be satisfied by an IPGA with fitness proportional selection and uniform crossover provided that the initial population belongs to a particular class of distributions (which includes the uniform distribution), and the fitness function satisfies a relatively weak (statistical) constraint.

1.5 Coarse-Graining and Population Genetics

The practice of assigning fitness values to *genotypes* is so ubiquitous in Population Genetics that it can be considered to be part of the foundations of that field. Without this practice mathematical models of changes to the genotype frequencies of an evolving population would not be tractable. This practice however is not without its detractors — most notably Mayr, who has labeled it “Bean Bag Genetics”. The practice is dissatisfying to many because it reflects a willingness on the part of Population Geneticists to accept (without a principled argument) what amounts to a statement that the evolutionary dynamics over the state space of genome frequencies can be coarse-grained such that a) the new state space is a set of genotype frequencies, b) the coarse-grained dynamics over this state space is also evolutionary in form, c) the ‘fitness’ of each genotype is simply the average fitness of all genomes that belong to the genotype.

The work in this paper demonstrates that such a coarse-graining *is indeed possible* provided that certain conditions are met. We only prove, in this paper, that these conditions are *sufficient*. However, we also conjecture that they (or trivially different variants) are *necessary*. If this conjecture is true then the widespread practice of assigning fitness values to genotypes within Population Genetics amounts to an implicit assumption that these conditions are true. Thus, if our hypothesis is true, then this paper makes an important contribution to Population Genetics by unearthing implicit assumptions that are inherent within the foundations of that field.

1.6 Structure of this Paper

The rest of this paper is organized as follows: In the next section we define the basic mathematical objects and notation used in this paper. In section 3 we define the concepts of semi-concordance, concordance and global concordance that are useful for formalizing the idea of a coarse-graining. In section 4 and section 5 we prove some stepping-stone results about selection and variation. We use these results in section 6 where we prove that an IPEA that satisfies certain abstract conditions can be coarse-grained. The proofs in sections 5 and 6 rely on lemmas which have been relegated to and proved in the appendix. In section 7 we describe a class of IPGAs with non-trivial fitness functions. We argue that these IPGAs approximately satisfy the abstract conditions of section 6 and can hence be coarse-grained for a small number of generations. We conclude in section 8 with a summary of our work and a discussion of future work.

2 Mathematical Preliminaries

Let X, Y be sets and let $\xi : X \to Y$ be some function. We use the notation $\langle y \rangle_\xi$ to denote the pre-image of y, i.e. the set $\{x \in X \,|\, \beta(x) = y\}$. For any subset $A \subset X$ we use the notation $\xi(A)$ to denote the set $\{y \in Y \,|\, \xi(a) = y \text{ and } a \in A\}$

As in [17], for any set X we use the notation Λ^X to denote the set of all distributions over X, i.e. Λ^X denotes set $\{f : X \to [0,1] \mid \sum_{x \in X} f(x) = 1\}$. For any set X, let $0^X : X \to \{0\}$ be the constant zero function over X. For any set X, an m-parent transmission function [13, 1, 18] over X is an element of the set

$$\left\{ T : \prod_1^{m+1} X \to [0,1] \;\middle|\; \forall x_1, \ldots, x_m \in X, \sum_{x \in X} T(x, x_1', \ldots, x_m') = 1 \right\}$$

Extending the notation introduced above, we denote this set by Λ_m^X. Following [17], we use conditional probability notation in our denotation of transmission functions. Thus an m-parent transmission function $T(x, x_1, \ldots, x_m)$ is denoted $T(x|x_1, \ldots, x_m)$.

A transmission function can be used to model the individual-level effect of mutation, which operates on one parent and produces one child, and indeed the individual-level effect of any variation operation which operates on any numbers of parents and produces one child.

Our scheme for modeling EA dynamics is based on the one used in [17]. We model the genomic populations of an EA as distributions over the genome set. The population-level effect of the evolutionary operations of an EA is modeled by mathematical operators whose inputs and outputs are such distributions.

The expectation operator, defined below, is used in the definition of the selection operator, which follows thereafter.

Definition 1. (EXPECTATION OPERATOR) *Let X be some finite set, and let $f : X \to \mathbb{R}^+$ be some function. We define the expectation operator $\mathcal{E}_f : \Lambda^X \cup 0^X \to \mathbb{R}^+ \cup \{0\}$ as follows:*

$$\mathcal{E}_f(p) = \sum_{x \in X} f(x)p(x)$$

The selection operator is parameterized by a fitness function. It models the effect of fitness proportional selection on a population of genomes.

Definition 2. (SELECTION OPERATOR) *Let X be some finite set and let $f : X \to \mathbb{R}^+$ be some function. We define the* Selection Operator *$\mathcal{S}_f : \Lambda^X \to \Lambda^X$ as follows:*

$$(\mathcal{S}_f p)(x) = \frac{f(x)p(x)}{\mathcal{E}_f(p)}$$

The population-level effect of variation is modeled by the variation operator. This operator is parameterized by a transmission function which models the effect of variation at the individual level.

Definition 3. (VARIATION OPERATOR[2]) *Let X be a countable set, and for any $m \in \mathbb{N}^+$, let $T \in \Lambda^X_m$ be a transmission function function over X. We define the variation operator $\mathcal{V}_T : \Lambda^X \to \Lambda^X$ as follows:*

$$(\mathcal{V}_T p)(x) = \sum_{\substack{(x_1,\ldots,x_m) \\ \in \prod_1^m X}} T(x|x_1,\ldots,x_m) \prod_{i=1}^{m} p(x_i)$$

The next definition describes the projection operator (previously used in [19] and [17]). A projection operator that is parameterized by some function β 'projects' distributions over the domain of β, to distributions over its co-domain.

[2] also called the Mixing Operator in [19] and [17]

Definition 4. (Projection Operator) *Let X be a countable set, let Y be some set, and let $\beta : X \to Y$ be a function. We define the projection operator, $\Xi_\beta : \Lambda^X \to \Lambda^Y$ as follows:*

$$(\Xi_\beta p\ \ )(y) = \sum_{x \in \langle y \rangle_\beta} p(x)$$

and call $\Xi_\beta p$ the β-projection of p.

3 A Formalization of Coarse-Graining

The following definition introduces some convenient function-related terminology.

Definition 5. (Partitioning, Theme Set, Themes, Theme Class) *Let X, K be sets and let $\beta : X \to K$ be a surjective function. We call β a* partitioning*, call the co-domain K of β the* theme set *of β, call any element in K a* theme *of β, and call the pre-image $\langle k \rangle_\beta$ of some $k \in K$, the* theme class *of k under β.*

The next definition introduces concepts which are useful for formalizing the notion of coarse-graining

Definition 6 (Semi-Concordance, Concordance, Global Concordance)**.** *Let G, K be sets, let $\mathcal{W} : \Lambda^G \to \Lambda^G$ be an operator, let $\beta : G \to K$ be a partitioning, and let $U \subseteq \Lambda^G$ such that $\Xi_\beta(U) = \Lambda^K$. We say that $\mathcal{W}$ is semi-concordant with β on U if there exists an operator $\mathcal{Q} : \Lambda^K \to \Lambda^K$ such that for all $p \in U$, $\mathcal{Q} \circ \Xi_\beta p = \Xi_\beta \circ \mathcal{W} p$, i.e. the following diagram commutes:*

$$\begin{tikzcd}
U \arrow[r, "\mathcal{W}"] \arrow[d, "\Xi_\beta"'] & \Lambda^G \arrow[d, "\Xi_\beta"] \\
\Lambda^K \arrow[r, "\mathcal{Q}"'] & \Lambda^K
\end{tikzcd}$$

Since β is surjective, if $\mathcal{Q}$ exists, it is clearly unique; we call it the quotient. We call G, K, W, and U the domain, co-domain, primary operator and turf respectively. If in addition $\mathcal{W}(U) \subseteq U$ we say that $\mathcal{W}$ is concordant with β on U. If in addition $U = \Lambda^G$ we say that $\mathcal{W}$ is globally concordant with β.

Global Concordance is a stricter condition than concordance, which in turn is a stricter condition than semi-concordance. It is easily seen that global concordance is equivalent to Vose's notion of compatibility [19, p. 188].

If some operator $\mathcal{W}$ is concordant with some function β over some turf U with some quotient $\mathcal{Q}$, then for any distribution $p_K \in \Xi_\beta(U)$, and all distributions $p_G \in \langle p_K \rangle_{\Xi_\beta}$, one can study the *projected* effect of the repeated application of $\mathcal{W}$ to p_G simply by studying the effect of the repeated application of $\mathcal{Q}$ to p_K. If the size of K is small then a computational study of the projected effect of the repeated application of $\mathcal{W}$ to distributions in $\langle p_K \rangle_{\Xi_\beta}$ becomes feasible. Therefore, if one can derive such a concordance then one has succeeded in coarse-graining the dynamics induced by $\mathcal{W}$ for any initial condition in $\langle p_K \rangle_{\Xi_\beta}$. (Note that the partition function Ξ_β of the coarse-graining is not the same as the partitioning β of the concordance)

4 Global Concordance of Variation

We show that some variation operator $\mathcal{V}_T$ is globally concordant with some partitioning if a relationship, that we call *ambivalence*, exists between the transmission function T of the variation operator and the partitioning.

To illustrate the idea of ambivalence consider a partition function β which partitions a genome set G into three subsets. Fig 1 depicts the behavior of a two-parent transmission function that is ambivalent under β. Given two parents and some child, the probability that the child will belong to some theme class depends *only* on the theme classes of the parents and *not* on the specific parent genomes. Hence the name 'ambivalent' — it captures the sense that when viewed from the coarse-grained level of the theme classes, a transmission function 'does not care' about the specific genomes of the parents or the child.

The definition of ambivalence that follows is equivalent to but more useful than the definition given in [5]

Definition 7. (Ambivalence) *Let G, K be countable sets, let $T \in \Lambda_m^G$ be a transmission function, and let $\beta : G \to K$ be a partitioning. We say that T is ambivalent under β if there exists some transmission function $D \in \Lambda_m^K$, such that for all $k, k_1, \ldots, k_m \in K$ and for any $x_1 \in \langle k_1 \rangle_\beta, \ldots, x_m \in \langle k_m \rangle_\beta$,*

$$\sum_{x \in \langle k \rangle_\beta} T(x|x_1, \ldots, x_m) = D(k|k_1, \ldots, k_m)$$

If such a D exits, it is clearly unique. We denote it by $T^{\overrightarrow{\beta}}$ and call it the theme transmission function.

Suppose $T \in \Omega_m^X$ is ambivalent under some $\beta : X \to K$, we can use the projection operator to express the projection of T under β as follows: for all $k, k_1, \ldots, k_m \in K$, and any $x_1 \in \langle k_1 \rangle_\beta, \ldots, x_m \in \langle k_m \rangle_\beta$, $T^{\overrightarrow{\beta}}(k|k_1, \ldots k_m)$ is given by $(\Xi_\beta(T(\cdot\,|x_1, \ldots, x_m)))(k)$.

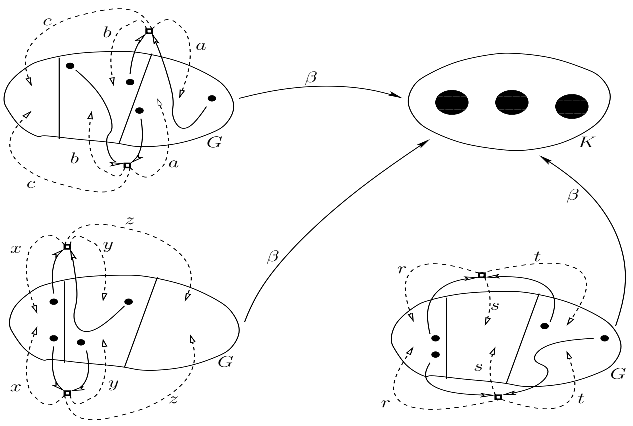

Figure 1: Let $\beta : G \rightarrow K$ be a coarse-graining which partitions the genome set G into three theme classes. This figure depicts the behavior of a two-parent variation operator that is ambivalent under β. The small dots denote specific genomes and the solid unlabeled arrows denote the recombination of these genomes. A dashed arrow denotes that a child from a recombination may be produced ‘somewhere’ within the theme class that it points to, and the label of a dashed arrow denotes the probability with which this might occur. As the diagram shows the probability that the child of a variation operation will belong to a particular theme class depends *only* on the theme classes of the parents and *not* on their specific genomes

The following theorem shows that a variation operator is globally concordant under some partitioning if it is parameterized by a transmission function which is ambivalent under that partitioning.

Theorem 1 (Global Concordance of Variation)**.** *Let G and K be countable sets, let $T \in \Lambda^G_m$ be a transmission function and let $\beta : G \rightarrow K$ be some partitioning such that T is ambivalent under β. Then $\mathcal{V}_T : \Lambda^G \rightarrow \Lambda^G$ is globally concordant under β with quotient $\mathcal{V}_{T^{\overrightarrow{\beta}}}$, i.e. the following diagram commutes:*

$$\begin{tikzcd}
\Lambda^G \arrow[r, "\mathcal{V}_T"] \arrow[d, "\Xi_\beta"'] & \Lambda^G \arrow[d, "\Xi_\beta"] \\
\Lambda^K \arrow[r, "\mathcal{V}_{T^{\overrightarrow{\beta}}}"'] & \Lambda^K
\end{tikzcd}$$

Proof: For any $p \in \Lambda^G$,

$(\Xi_\beta \circ \mathcal{V}_T p)(k)$

$$
\begin{aligned}
&= \sum_{x \in \langle k \rangle_\beta} \sum_{\substack{(x_1,\ldots,x_m) \\ \in \prod\limits_1^m X}} T(x|x_1,\ldots,x_m) \prod_{i=1}^m p(x_i) \\
&= \sum_{\substack{(x_1,\ldots,x_m) \\ \in \prod\limits_1^m X}} \sum_{x \in \langle k \rangle_\beta} T(x|x_1,\ldots,x_m) \prod_{i=1}^m p(x_i) \\
&= \sum_{\substack{(x_1,\ldots,x_m) \\ \in \prod\limits_1^m X}} \prod_{i=1}^m p(x_i) \sum_{x \in \langle k \rangle_\beta} T(x|x_1,\ldots,x_m) \\
&= \sum_{\substack{(k_1,\ldots,k_m) \\ \in \prod\limits_1^m K}} \sum_{\substack{(x_1,\ldots,x_m) \\ \in \prod\limits_{j=1}^m \langle k_j \rangle_\beta}} \prod_{i=1}^m p(x_i) \sum_{x \in \langle k \rangle_\beta} T(x|x_1,\ldots,x_m) \\
&= \sum_{\substack{(k_1,\ldots,k_m) \\ \in \prod\limits_1^m K}} \sum_{\substack{(x_1,\ldots,x_m) \\ \in \prod\limits_{j=1}^m \langle k_j \rangle_\beta}} \prod_{i=1}^m p(x_i) T^{\overrightarrow{\beta}}(k|k_1,\ldots,k_m) \\
&= \sum_{\substack{(k_1,\ldots,k_m) \\ \in \prod\limits_1^m K}} T^{\overrightarrow{\beta}}(k|k_1,\ldots,k_m) \sum_{\substack{(x_1,\ldots,x_m) \\ \in \prod\limits_{j=1}^m \langle k_j \rangle_\beta}} \prod_{i=1}^m p(x_i) \\
&= \sum_{\substack{(k_1,\ldots,k_m) \\ \in \prod\limits_1^m K}} T^{\overrightarrow{\beta}}(k|k_1,\ldots,k_m) \sum_{x_1 \in \langle k_1 \rangle_\beta} \ldots \sum_{x_m \in \langle k_m \rangle_\beta} p(x_1) \ldots p(x_m) \\
&= \sum_{\substack{(k_1,\ldots,k_m) \\ \in \prod\limits_1^m K}} T^{\overrightarrow{\beta}}(k|k_1,\ldots,k_m) \Bigg(\sum_{x_1 \in \langle k_1 \rangle} p(x_1) \Bigg) \ldots \Bigg(\sum_{x_m \in \langle k_m \rangle} p(x_m) \Bigg) \\
&= \sum_{\substack{(k_1,\ldots,k_m) \\ \in \prod\limits_1^m K}} T^{\overrightarrow{\beta}}(k|k_1,\ldots,k_m) \prod_{i=1}^m \Bigg((\Xi_\beta p)(k_i) \Bigg) \\
&= (\mathcal{V}_{T^{\overrightarrow{\beta}}} \circ \Xi_\beta p)(k) \quad \square
\end{aligned}
$$

The implicit parallelism theorem in [20] is similar to the theorem above. Note however that the former theorem only shows that variation is globally concordant if firstly, the genome set consists of "fixed length strings, where the size of the alphabet can vary from position to position", secondly the partition over the genome set is a schema partition, and thirdly variation is 'structural' (see [20] for details). The Global Concordance of Variation theorem has none of these specific requirements. Instead it is premised on the existence of an abstract relationship – ambivalence – between the variation operation and a partitioning. The abstract nature of this relationship makes this theorem applicable to evolutionary algorithms other than GAs. In addition this theorem illuminates the essential relationship between 'structural' variation and schemata which was used (implicitly) in the proof of the implicit parallelism theorem.

In [5] it is shown that a variation operator that models any combination of variation operations that are commonly used in GAs — i.e. any combination of mask based crossover and 'canonical' mutation, in any order — is ambivalent under any partitioning that maps bitstrings to schemata (such a partitioning was called a schema partitioning). Therefore 'common' variation in IPGAs is globally concordant with *any* schema partitioning. This is precisely the result of the implicit parallelism theorem.

5 Limitwise Semi-Concordance of Selection

For some fitness function $f : G \to \mathbb{R}^+$ and some partition function $\beta : G \to K$ let us say that f is *thematically invariant* under β if, for any schema $k \in K$, the genomes that belong to $\langle k \rangle_\beta$ all have the same fitness. Paraphrasing the discussion in [20] using the terminology developed in this paper, Wright et. al. argue (but do not prove) that if the selection operator is globally concordant under some schema partitioning $\beta : G \to K$ then the fitness function that parameterizes the selection operator is 'schematically' invariant under β. It is relatively simple to use contradiction to prove a generalization of this statement for arbitrary partitionings.

Thematic invariance is a very strict condition for a fitness function. An IPGA whose fitness function meets this condition is unlikely to yield any substantive information about the dynamics of real world GAs.

As stated above, the selection operator is not *globally* concordant unless the fitness function satisfies thematic invariance, however if the set of distributions that selection operates over (i.e. the turf) is appropriately constrained, then, as we show in this section, the selection operator is *semi*-concordant over the turf even when the fitness function only satisfies a much *weaker* condition called thematic *mean* invariance.

For any partitioning $\beta : G \to K$, any theme k, and any distribution $p \in \Lambda^G$, the theme conditional operator, defined below, returns a conditional distribution in Λ^G that is obtained by normalizing the probability mass of the elements in $\langle k \rangle_\beta$ by $(\Xi_\beta p)(k)$

Definition 8 (Theme Conditional Operator)**.** *Let G be some countable set, let K be some set, and let $\beta : G \to K$ be some function. We define the theme conditional operator $\mathcal{C}_\beta : \Lambda^G \times K \to \Lambda^G \cup 0^G$ as follow: For any $p \in \Lambda^G$, and any $k \in K$, $\mathcal{C}_\beta(p, k) \in \Lambda^G \cup 0^G$ such that for any $x \in \langle k \rangle_\beta$,*

$$(\mathcal{C}_\beta(p,k))(x) = \begin{cases} 0 & \textit{if } (\Xi_\beta p)(k) = 0 \\ \frac{p(x)}{(\Xi_\beta p)(k)} & \textit{otherwise} \end{cases}$$

A useful property of the theme conditional operator is that it can be composed with the expected fitness operator to give an operator that returns the average fitness of the genomes in some theme class. To be precise, given some finite genome set G, some partitioning $\beta : G \to K$, some fitness function $f : G \to \mathbb{R}^+$, some distribution $p \in \Lambda^G$, and some theme $k \in K$, $\mathcal{E}_f \circ \mathcal{C}_\beta(p, k)$ is the average fitness of the genomes in $\langle k \rangle_\beta$. This property proves useful in the following definition.

Definition 9 (Bounded Thematic Mean Divergence, Thematic Mean Invariance)**.** *Let G be some finite set, let K be some set, let $\beta : G \to K$ be a partitioning, let $f : G \to \mathbb{R}^+$ and $f^* : K \to \mathbb{R}^+$ be functions, let $U \subseteq \Lambda^G$, and let $\delta \in \mathbb{R}^+$. We say that the thematic mean divergence of f with respect to f^* on U under β is bounded by $\delta \geq 0$ if for any $p \in U$ and for any $k \in K$*

$$|\mathcal{E}_f \circ \mathcal{C}_\beta(p, k) - f^*(k)| \leq \delta$$

If $\delta = 0$ we say that f is thematically mean invariant with respect to f^* on U

The next definition gives us a means to measure a 'distance' between real valued functions over finite sets.

Definition 10 (Manhattan Distance Between Real Valued Functions)**.** *Let X be a finite set then for any functions f, h of type $X \to \mathbb{R}$ we define the manhattan distance between f and h, denoted by $d(f, h)$, as follows:*

$$d(f, h) = \sum_{x \in X} |f(x) - h(x)|$$

It is easily checked that d is a metric.

Let $f : G \to \mathbb{R}^+$, $\beta : G \to K$ and $f^* : G \to \mathbb{R}^+$ be functions with finite domains, and let $U \in \Lambda^G$. The following theorem shows that if the thematic mean divergence of f with respect to f^* on U under β is bounded by some $\delta > 0$, then in the limit as $\delta \to 0$, $\mathcal{S}_f$ is semi-concordant with β on U .

Theorem 2 (Limitwise Semi-Concordance of Selection)**.** *Let G and K be finite sets, let $\beta : G \to K$ be a partitioning, Let $U \subseteq \Lambda^G$ such that $\Xi_\beta(U) = \Lambda^K$, let $f : G \to \mathbb{R}^+$, $f^* : K \to \mathbb{R}^+$ be some functions such that the thematic mean divergence of f with respect to f^* on U under β is bounded by δ, then for any $p \in U$ and any $\epsilon > 0$ there exists a $\delta' > 0$ such that,*

$$\delta < \delta' \Rightarrow d(\Xi_\beta \circ \mathcal{S}_f p, \mathcal{S}_{f^*} \circ \Xi_\beta p) < \epsilon$$

We depict the result of this theorem as follows:

$$\begin{tikzcd}
U \arrow[r, "\mathcal{S}_f"] \arrow[d, "\Xi_\beta"'] \arrow[dr, phantom, "\lim\limits_{\delta \to 0}"] & \Lambda^G \arrow[d, "\Xi_\beta"] \\
\Lambda^K \arrow[r, "\mathcal{S}_{f^*}"'] & \Lambda^K
\end{tikzcd}$$

Proof: For any $p \in U$ and for any $k \in K$,

$$\begin{aligned}
&(\Xi_\beta \circ \mathcal{S}_f p)(k) \\
&= \sum_{g \in \langle k \rangle_\beta} (\mathcal{S}_f p)(g) \\
&= \sum_{g \in \langle k \rangle_\beta} \frac{f(g).p(g)}{\sum_{g' \in G} f(g').p(g')} \\
&= \frac{\sum\limits_{g \in \langle k \rangle_\beta} f(g).(\Xi_\beta p)(k).(\mathcal{C}_\beta(p,k))(g)}{\sum\limits_{k' \in K} \sum\limits_{g' \in \langle k' \rangle_\beta} f(g').(\Xi_\beta p)(k')(\mathcal{C}_\beta(p,k'))(g')} \\
&= \frac{(\Xi_\beta p)(k) \sum\limits_{g \in \langle k \rangle_\beta} f(g).(\mathcal{C}_\beta(p,k))(g)}{\sum\limits_{k' \in K} (\Xi_\beta p)(k') \sum\limits_{g' \in \langle k' \rangle_\beta} f(g').(\mathcal{C}_\beta(p,k'))(g')} \\
&= \frac{(\Xi_\beta p)(k).\mathcal{E}_f \circ \mathcal{C}_\beta(p,k)}{\sum\limits_{k' \in K} (\Xi_\beta p_G)(k').\mathcal{E}_f \circ \mathcal{C}_\beta(p,k')} \\
&= (\mathcal{S}_{\mathcal{E}_f \circ \mathcal{C}_\beta(p,\cdot)} \circ \Xi_\beta p)(k)
\end{aligned}$$

So we have that

$$d(\Xi_\beta \circ \mathcal{S}_f p, \mathcal{S}_{f^*} \circ \Xi_\beta p) = d(\mathcal{S}_{\mathcal{E}_f \circ \mathcal{C}_\beta(p,\cdot)} \circ \Xi_\beta p, \mathcal{S}_{f^*} \circ \Xi_\beta p)$$

By Lemma, 4 (in the appendix) for any $\epsilon > 0$ there exists a $\delta_1 > 0$ such that,

$$d(\mathcal{E}_f \circ \mathcal{C}_\beta(p,.), f^*) < \delta_1 \Rightarrow d(\mathcal{S}_{\mathcal{E}_f \circ \mathcal{C}_\beta(p,\cdot)}(\Xi_\beta p), \mathcal{S}_{f^*}(\Xi_\beta p) < \epsilon$$

Now, if $\delta < \frac{\delta'}{|K|}$, then $d(\mathcal{E}_f \circ \mathcal{C}_\beta(p,.), f^*) < \delta_1$ □

Corollary 1. *If $\delta = 0$, i.e. if f is thematically mean invariant with respect to f^* on U, then S_f is semi-concordant with β on U with quotient $\mathcal{S}_{f^*}$, i.e. the following diagram commutes:*

$$\begin{tikzcd}
U \arrow[r, "\mathcal{S}_f"] \arrow[d, "\Xi_\beta"'] & \Lambda^G \arrow[d, "\Xi_\beta"] \\
\Lambda^K \arrow[r, "\mathcal{S}_{f^*}"'] & \Lambda^K
\end{tikzcd}$$

6 Limitwise Concordance of Evolution

The two definitions below formalize the idea of an infinite population model of an EA, and its dynamics [3].

Definition 11 (Evolution Machine)**.** *An evolution machine (EM) is a tuple (G, T, f) where G is some set called the domain, $f : G \to \mathbb{R}^+$ is a function called the fitness function and $T \in \Lambda^G_m$ is called the transmission function.*

Definition 12 (Evolution Epoch Operator)**.** *Let $E = (G, T, f)$ be an evolution machine. We define the evolution epoch operator $\mathcal{G}_E : \Lambda^G \to \Lambda^G$ as follows:*

$$\mathcal{G}_E = \mathcal{V}_T \circ \mathcal{S}_f$$

For some evolution machine E, our aim is to give sufficient conditions under which, for any $t \in \mathbb{Z}^+$, $\mathcal{G}^t_E$ is approaches concordance in a limit. The following definition gives us a formal way to state one of these conditions.

Definition 13 (Non-Departure)**.** *Let $E = (G, T, f)$ be an evolution machine, let $\beta : G \to K$ be some partitioning, and let $U \subseteq \Lambda^G$. We say that E is* non-departing *over U if*

$$\mathcal{V}_T \circ \mathcal{S}_f(U) \subseteq U$$

Note that our definition does *not* require $S_f(U) \subseteq U$ in order for E to be non-departing from U.

[3] The definition of an EM given here is different from its definition in [3, 4]. The fitness function in this definition maps genomes directly to fitness values. It therefore subsumes the genotype-to-phenotype and the phenotype-to-fitness functions of the previous definition. In previous work these two functions were always composed together; their subsumption within a single function increases conceptual clarity.

Theorem 3 (Limitwise Concordance of Evolution)**.** *Let $E = (G, T, f)$, be an evolution machine such that G is finite, let $\beta : G \to K$ be some partitioning, and let $U \subseteq \Lambda^G$ such that $\Xi_\beta(U) = \Lambda^K$. Suppose that the following statements are true:*

1. *The thematic mean divergence of f with respect to f^* on U under β is bounded by δ*
2. *T is ambivalent under β*
3. *E is non-departing over U*

Then, letting $E^* = (K, T^{\overrightarrow{\beta}}, f^*)$ be an evolution machine, for any $t \in \mathbb{Z}^+$ and any $p \in U$,

1. *$\mathcal{G}^t_E p \ \in U$*
2. *For any $\epsilon > 0$, there exists $\delta' > 0$ such that,*

$$\delta < \delta' \Rightarrow d(\Xi_\beta \circ \mathcal{G}^t_E p \ , \ \mathcal{G}^t_{E^*} \circ \Xi_\beta p) < \epsilon$$

We depict the result of this theorem as follows:

$$\begin{tikzcd}
U \arrow[r, "\mathcal{G}^t_E"] \arrow[d, "\Xi_\beta"'] \arrow[dr, phantom, "\lim\limits_{\delta \to 0}"] & U \arrow[d, "\Xi_\beta"] \\
\Lambda^K \arrow[r, "\mathcal{G}^t_{E^*}"'] & \Lambda^K
\end{tikzcd}$$

Proof: We prove the theorem for any $t \in \mathbb{Z}^+_0$. The proof is by induction on t. The base case, when $t = 0$, is trivial. For some $n = \mathbb{Z}^+_0$, let us assume the hypothesis for $t = n$. We now show that it is true for $t = n + 1$. For any $p \in U$, by the inductive assumption $\mathcal{G}^n_E p$ is in U. Therefore, since E is non-departing over U, $\mathcal{G}^{n+1}_E p \ \in U$. This completes the proof of the first part of the hypothesis. For a proof of the second part note that,

$$\begin{aligned}
&d(\Xi_\beta \circ \mathcal{G}^{n+1}_E p \, , \, \mathcal{G}^{n+1}_{E^*} \circ \Xi_\beta p \ ) \\
&\quad = d(\Xi_\beta \circ \mathcal{V}_T \circ \mathcal{S}_f \circ \mathcal{G}^n_E p \, , \, \mathcal{V}_{T^{\overrightarrow{\beta}}} \circ \mathcal{S}_{f^*} \circ \mathcal{G}^n_{E^*} \circ \Xi_\beta p \ ) \\
&\quad = d(\mathcal{V}_{T^{\overrightarrow{\beta}}} \circ \Xi_\beta \circ \mathcal{S}_f \circ \mathcal{G}^n_E p \, , \, \mathcal{V}_{T^{\overrightarrow{\beta}}} \circ \mathcal{S}_{f^*} \circ \mathcal{G}^n_{E^*} \circ \Xi_\beta p) \quad \text{(by theorem 1)}
\end{aligned}$$

Hence, for any $\epsilon > 0$, by Lemma 2 there exists δ_1 such that

$$\begin{aligned}
&d(\Xi_\beta \circ \mathcal{S}_f \circ \mathcal{G}^n_E p \, , \, S_{f^*} \circ \mathcal{G}^n_{E^*} \circ \Xi_\beta p) < \delta_1 \Rightarrow \\
&\qquad\qquad\qquad d(\Xi_\beta \circ \mathcal{G}^{n+1}_E p \, , \, \mathcal{G}^{n+1}_{E^*} \circ \Xi_\beta p \ ) < \epsilon
\end{aligned}$$

As d is a metric it satisfies the triangle inequality. Therefore we have that

$$
\begin{aligned}
&d(\Xi_\beta \circ \mathcal{S}_f \circ \mathcal{G}_E^n p \,,\, \mathcal{S}_{f^*} \circ \mathcal{G}_{E^*}^n \circ \Xi_\beta p) \leq \\
&\qquad d(\Xi_\beta \circ \mathcal{S}_f \circ \mathcal{G}_E^n p \,,\, \mathcal{S}_{f^*} \circ \Xi_\beta \circ \mathcal{G}_E^n p) + \\
&\qquad\qquad d(\mathcal{S}_{f^*} \circ \Xi_\beta \circ \mathcal{G}_E^n p \,,\, \mathcal{S}_{f^*} \circ \mathcal{G}_{E^*}^n \circ \Xi_\beta p)
\end{aligned}
$$

By our inductive assumption $\mathcal{G}_E^n p \ \in U$. So, by theorem 2 there exists a δ_2 such that

$$
\delta < \delta_2 \Rightarrow d(\Xi_\beta \circ \mathcal{S}_f \circ \mathcal{G}_E^n p \,,\, \mathcal{S}_{f^*} \circ \Xi_\beta \circ \mathcal{G}_E^n p) < \frac{\delta_1}{2}
$$

By lemma 3 there exists a δ_3 such that

$$
\begin{aligned}
&d(\Xi_\beta \circ \mathcal{G}_E^n p \,,\, \mathcal{G}_{E^*}^n \circ \Xi_\beta p) < \delta_3 \Rightarrow \\
&\qquad\qquad d(\mathcal{S}_{f^*} \circ \Xi_\beta \circ \mathcal{G}_E^n p \,,\, \mathcal{S}_{f^*} \circ \mathcal{G}_{E^*}^n \circ \Xi_\beta p) < \frac{\delta_1}{2}
\end{aligned}
$$

By our inductive assumption, there exists a δ_4 such that

$$
\delta < \delta_4 \rightarrow d(\Xi_\beta \circ \mathcal{G}_E^n p \,,\, \mathcal{G}_{E^*}^n \circ \Xi_\beta p) < \delta_3
$$

Therefore, letting $\delta' = \min(\delta_2, \delta_4)$ we get that

$$
\delta < \delta^* \Rightarrow d(\Xi_\beta \circ \mathcal{G}_E^{n+1} p, \mathcal{G}_{E^*}^{n+1} \circ \Xi_\beta p) < \epsilon \quad \square
$$

7 Sufficient Conditions for Coarse-Graining IPGA Dynamics

We now use the result in the previous section to argue that the dynamics of an IPGA with uniform crossover and fitness proportional selection can be coarse-grained for at least a small number of generations, provided that the initial population satisfies a constraint called *approximate thematic uniformity* and the fitness function satisfies a relatively weak constraint called *low-variance schematic fitness distribution.* We stress at the outset that our argument is principled but informal, i.e. though the argument rests relatively straightforwardly on theorem 3, we do find it necessary in places to appeal to the reader's intuitive understanding of GA dynamics.

For any $n \in \mathbb{Z}^+$, let $\mathfrak{B}_n$ be the set of all bitstrings of length n. For some $\ell \gg 1$ and some $m \ll \ell$, let $\beta : \mathfrak{B}_\ell \to \mathfrak{B}_m$ be some schema partitioning. Let $f^* : \mathfrak{B}_m \to \mathbb{R}^+$ be some function. For any $k \in \mathfrak{B}_m$, let $D_k \in \Lambda^{\mathbb{R}^+}$ be some distribution over the reals with low variance such that the mean of distribution D_k is $f^*(k)$. Let $f : \mathfrak{B}_\ell \to \mathbb{R}^+$ be a fitness function such that for any $k \in \mathfrak{B}_m$, the fitness values of the elements of $\langle k\rangle_\beta$ are independently drawn from the distribution D_k. For such a fitness function we say that fitness is *schematically distributed with low-variance.*

Let U be a set of distributions such that for any $k \in \mathfrak{B}_m$ and any $p \in U$, $\mathcal{C}_\beta(p, k)$ is approximately uniform. It is easily checked that U satisfies the condition $\Xi_\beta(U) = \Lambda^{\mathfrak{B}_m}$. We say that the distributions in U are *approximately schematically uniform.*

Appealing to the weak law of large numbers, for any randomly chosen population $p \in U$, and any $\delta \in \mathbb{R}^+$, $\mathbf{P}(|\mathcal{E}_f \circ \mathcal{C}_\beta(p, k) - f^*(k)| < \delta) \to 1$ as $\ell - m \to \infty$. Because we have chosen ℓ and m such that $\ell - m$ is 'large', it is reasonable to assume that the thematic mean divergence of f on U under β is bounded by some 'small' value.

Let $T \in \Lambda^{\mathfrak{B}_\ell}$ be a transmission function that models the application of uniform crossover. In sections 6 and 7 of [5] we rigorously prove that a transmission function that models any mask based crossover operation is ambivalent under any schema partitioning. Uniform crossover is mask based, and β is a schema partitioning, therefore T is ambivalent under β.

Let $p_{\frac{1}{2}} \in \Lambda^{\mathfrak{B}_1}$ be such that $p_{\frac{1}{2}}(0) = \frac{1}{2}$ and $p_{\frac{1}{2}}(1) = \frac{1}{2}$. For any $p \in U$, $\mathcal{S}_f p$ may be 'outside' U because there may be one or more $k \in \mathfrak{B}_m$ such that $\mathcal{C}_\beta(\mathcal{S}_f p, k)$ is not quite uniform. Recall that for any $k \in \mathfrak{B}_m$ the variance of D_k is low. Therefore even though $\mathcal{S}_f p$ may be 'outside' U, the deviation from thematic uniformity is not likely to be large. Furthermore, given the low variance of D_k, the marginal distributions of $\mathcal{C}_\beta(\mathcal{S}_f p, k)$ will be very close to $p_{\frac{1}{2}}$. Given these facts and our choice of transmission function, for all $k \in K$, $\mathcal{C}_\beta(\mathcal{V}_T \circ \mathcal{S}_f p, k)$ will be more uniform than $\mathcal{C}_\beta(\mathcal{S}_f p, k)$, and we can assume that $\mathcal{V}_T \circ \mathcal{S}_f p$ is in U. In other words, we can assume that E is non-departing over U.

Let $E = (\mathfrak{B}_\ell, T, f)$ and $E^* = (\mathfrak{B}_m, T^{\overrightarrow{\beta}}, f^*)$ be evolution machines. By the discussion above and the limitwise concordance of evolution theorem one can expect that for any approximately thematically uniform distribution $p \in U$ (including of course the uniform distribution over $\mathfrak{B}_\ell$), for at least a small number of generations, the dynamics of E^* when initialized with $\Xi_\beta p$ will closely approximate the projected dynamics of E when initialized with p.

8 Conclusion

We believe that accurate theories that explain evolutionary adaptation will only be forthcoming when one can perform principled analyses of the short term dynamics of evolutionary systems. Previous theoretical results in the Evolutionary Computation literature do not permit such analyses. The technique of coarse-graining, which has widely been used in other sciences, is a promising approach towards the formulation of more accurate theories of evolutionary adaptation because, if successfully applied, it permits a principled analysis of evolutionary dynamics across multiple generations.

Previous coarse-graining results in EC were only obtained when variation was limited to mutation[12], or when selection was not applied at all [20]. The technique of obtaining a coarse-graining by showing compatibility, which was used in [20, 12], can only be successfully applied to selecto-recombinative evolution if the fitness function is schematically invariant, i.e. constant for each schema in some schema partition. Schematic invariance is unfortunately a very strong constraint to place on a fitness function. It is therefore highly unlikely that any contribution to a general theory of GA adaptation will be forthcoming from studies of IPGAs that satisfy this constraint.

In this paper we developed a simple yet powerful abstract framework for analyzing evolutionary dynamics. We then argued that the evolutionary dynamics of an IPGA with fitness proportional selection and uniform crossover can be coarse-grained for at least a small number of generations if its initial distribution satisfies a constraint called approximate schematic uniformity (a very reasonable condition), and its fitness is low-variance schematically distributed. The latter condition is much weaker than the schematic invariance constraint previously required to coarse-grain selecto-recombinative evolutionary dynamics.

Acknowledgements: I thank Jordan Pollack for supporting this work and the anonymous reviewers for their very helpful comments and suggestions.

References

[1] Lee Altenberg. The evolution of evolvability in genetic programming. In Kenneth E. Kinnear, Jr., editor, *Advances in Genetic Programming*. MIT Press, 1994.

[2] Sunny Y. Auyang. *Foundations of Complex-system Theories : In Economics, Evolutionary Biology, and Statistical Physics*. Cambridge University Press, August 1999.

[3] Keki Burjorjee and Jordan B. Pollack. Theme preservation and the evolution of representation. In *Theory of Representation Workshop, GECCO*, 2005.

[4] Keki Burjorjee and Jordan B. Pollack. Theme preservation and the evolution of representation. In *IICAI*, pages 1444–1463, 2005.

[5] Keki Burjorjee and Jordan B. Pollack. A general coarse-graining framework for studying simultaneous inter-population constraints

induced by evolutionary operations. In *GECCO 2006: Proceedings of the 8th annual conference on Genetic and evolutionary computation*. ACM Press, 2006.

[6] Andrés Aguilar Contreras, Jonathan E. Rowe, and Christopher R. Stephens. Coarse-graining in genetic algorithms: Some issues and examples. In *GECCO*, pages 874–885, 2003.

[7] David B. Fogel. *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence*. IEEE Press, NY, 1995.

[8] David E. Goldberg. *Genetic Algorithms in Search, Optimization & Machine Learning*. Addison-Wesley, Reading, MA, 1989.

[9] John H. Holland. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. University of Michigan, 1975.

[10] W. B. Langdon and Riccardo Poli. *Foundations of Genetic Programming*. Springer-Verlag, 2002.

[11] Melanie Mitchell. *An Introduction to Genetic Algorithms*. The MIT Press, Cambridge, MA, 1996.

[12] Jonathan E. Rowe, Michael D. Vose, and Alden H. Wright. Differentiable coarse graining. *Theor. Comput. Sci*, 361(1):111–129, 2006.

[13] Montgomery Slatkin. Selection and polygenic characters. *PNAS*, 66(1):87–93, 1970.

[14] Chris Stephens and Henri Waelbroeck. Schemata evolution and building blocks. *Evolutionary Computation*, 7(2):109–124, 1999.

[15] Christopher R. Stephens and Adolfo Zamora. EC theory: A unified viewpoint. In *Genetic and Evolutionary Computation – GECCO-2003*, Berlin, 2003. Springer-Verlag.

[16] G. Syswerda. Uniform crossover in genetic algorithms. In J. D. Schaffer, editor, *Proceeding of the Third International Conference on Genetic Algorithms*. Morgan Kaufmann, 1989.

[17] Marc Toussaint. *The Evolution of Genetic Representations and Modular Neural Adaptation*. PhD thesis, Institut fr Neuroinformatik, Ruhr-Universiät-Bochum, Germany, 2003.

[18] Marc Toussaint. On the evolution of phenotypic exploration distributions. In *Foundations of Genetic Algorithms 7 (FOGA VII)*. Morgan Kaufmann, 2003.

[19] Michael D. Vose. *The simple genetic algorithm: foundations and theory*. MIT Press, 1999.

[20] Alden H. Wright, Michael D. Vose, and Jonathan E. Rowe. Implicit parallelism. In *GECCO*, 2003.

Appendix

Lemma 1. *For any finite set X, and any metric space (Υ, d),let $\mathcal{A} : \Upsilon \to \Lambda^X$ and let $\mathcal{B} : X \to [\Upsilon \to [0,1]]$ be functions*[4] *such that for any $h \in \Upsilon$, and any $x \in X$, $(\mathcal{B}(x))(h) = (\mathcal{A}(h))(x)$. For any $x \in X$, and for any $h^* \in \Upsilon$, if the following statement is true*

$$\begin{aligned}&\forall x \in X, \forall \epsilon_x > 0, \exists \delta_x > 0, \forall h \in \Upsilon,\\ &\qquad\qquad d(h, h^*) < \delta_x \Rightarrow |(\mathcal{B}(x))(h) - (\mathcal{B}(x))(h^*)| < \epsilon_x\end{aligned}$$

Then we have that

$$\forall \epsilon > 0, \exists \delta > 0, \forall h \in \Upsilon, d(h, h^*) < \delta \Rightarrow d(\mathcal{A}(h), \mathcal{A}(h^*)) < \epsilon$$

This lemma says that $\mathcal{A}$ is continuous at h^* if for all $x \in X$, $\mathcal{B}(x)$ is continuous at h^*.

Proof: We first prove the following two claims

Claim 1.

$$\begin{aligned}&\forall x \in X \;\; s.t. \;\; (\mathcal{B}(x))(h^*) > 0, \forall \epsilon_x > 0, \exists \delta_x > 0, \forall h \in \Upsilon,\\ &\qquad d(h, h^*) < \delta_x \Rightarrow |(\mathcal{B}(x))(h) - (\mathcal{B}(x))(h^*)| < \epsilon_x.(\mathcal{B}(x))(h^*)\end{aligned}$$

This claim follows from the continuity of $\mathcal{B}(x)$ at h^* for all $x \in X$ and the fact that $(\mathcal{B}(x))(h^*)$ is a positive constant w.r.t. h.

Claim 2. *for all $h \in \Upsilon$*

$$\sum_{\substack{x \in X s.t.\\ (\mathcal{A}(h^*))(x) >\\ (\mathcal{A}(h))(x)}} |(\mathcal{A}(h^*))(x) - (\mathcal{A}(h))(x)| = \sum_{\substack{x \in X s.t.\\ (\mathcal{A}(h))(x) >\\ (\mathcal{A}(h^*))(x)}} |(\mathcal{A}(h))(x) - (\mathcal{A}(h^*))(x)|$$

The proof of this claim is as follows: for all $h \in \Upsilon$,

$$\begin{aligned}&\sum_{x \in X} (\mathcal{A}(h^*)(x)) - (\mathcal{A}(h))(x) = 0\\ &\Rightarrow \sum_{\substack{x \in X\text{s.t.}\\ (\mathcal{A}(h^*))(x) >\\ (\mathcal{A}(h))(x)}} (\mathcal{A}(h^*))(x) - (\mathcal{A}(h))(x) - \sum_{\substack{x \in X\text{s.t.}\\ (\mathcal{A}(h))(x) >\\ (\mathcal{A}(h^*))(x)}} (\mathcal{A}(h))(x) - (\mathcal{A}(h^*))(x) = 0\\ &\Rightarrow \sum_{\substack{x \in X\text{s.t.}\\ (\mathcal{A}(h^*))(x) >\\ (\mathcal{A}(h))(x)}} (\mathcal{A}(h^*))(x) - (\mathcal{A}(h))(x) = \sum_{\substack{x \in X\text{s.t.}\\ (\mathcal{A}(h))(x) >\\ (\mathcal{A}(h^*))(x)}} (\mathcal{A}(h))(x) - (\mathcal{A}(h^*))(x)\end{aligned}$$

[4] For any sets X, Y we use the notation $[X \to Y]$ to denote the set of all functions from X to Y

$$\Rightarrow \left| \sum_{\substack{x\in X \text{s.t.} \\ (\mathcal{A}(h^*))(x)> \\ (\mathcal{A}(h))(x)}} (\mathcal{A}(h^*))(x) - (\mathcal{A}(h))(x) \right| = \left| \sum_{\substack{x\in X \text{s.t.} \\ (\mathcal{A}(h))(x)> \\ (\mathcal{A}(h^*))(x)}} (\mathcal{A}(h))(x) - (\mathcal{A}(h^*))(x) \right|$$

$$\Rightarrow \sum_{\substack{x\in X \text{s.t.} \\ (\mathcal{A}(h^*))(x)> \\ (\mathcal{A}(h))(x)}} |(\mathcal{A}(h^*))(x) - (\mathcal{A}(h))(x)| = \sum_{\substack{x\in X \text{s.t.} \\ (\mathcal{A}(h))(x)> \\ (\mathcal{A}(h^*))(x)}} |(\mathcal{A}(h))(x) - (\mathcal{A}(h^*))(x)|$$

We now prove the lemma. Using claim 1 and the fact that X is finite, we get that $\forall \epsilon > 0$, $\exists \delta > 0$, $\forall h \in [X \to \mathbb{R}]$ such that $d(h, h^*) < \delta$,

$$\sum_{\substack{x\in X \text{s.t.} \\ (\mathcal{A}(h^*))(x)> \\ (\mathcal{A}(h))(x)}} |(\mathcal{B}(x))(h^*) - (\mathcal{B}(x))(h)| < \sum_{\substack{x\in X \text{s.t.} \\ (\mathcal{A}(h^*))(x)> \\ (\mathcal{A}(h))(x)}} \frac{\epsilon}{2}.(\mathcal{B}(x))(h^*)$$

$$\Rightarrow \sum_{\substack{x\in X \text{s.t.} \\ (\mathcal{A}(h^*))(x)> \\ (\mathcal{A}(h))(x)}} |(\mathcal{A}(h^*))(x) - (\mathcal{A}(h))(x)| < \sum_{\substack{x\in X \text{s.t.} \\ (\mathcal{A}(h^*))(x)> \\ (\mathcal{A}(h))(x)}} \frac{\epsilon}{2}.(\mathcal{A}(h^*))(x)$$

$$\Rightarrow \sum_{\substack{x\in X \text{s.t.} \\ (\mathcal{A}(h^*))(x)> \\ (\mathcal{A}(h))(x)}} |(\mathcal{A}(h^*))(x) - (\mathcal{A}(h))(x)| < \frac{\epsilon}{2} \qquad \square$$

By Claim 2 and the result above, we have that $\forall \epsilon > 0$, $\exists \delta > 0$, $\forall h \in [X \to \mathbb{R}]$ such that $d(h, h^*) < \delta$,

$$\sum_{\substack{x\in X \text{s.t.} \\ (\mathcal{A}(h))(x)> \\ (\mathcal{A}(h^*))(x)}} |(\mathcal{A}(h))(x) - (\mathcal{A}(h^*))(x)| < \frac{\epsilon}{2}$$

Therefore, given the two previous results, we have that $\forall \epsilon > 0$, $\exists \delta > 0$, $\forall h \in [X \to \mathbb{R}]$ such that $d(h, h^*) < \delta$,

$$\sum_{x\in X} |(\mathcal{A}(h))(x) - (\mathcal{A}(h^*)(x))| < \epsilon \qquad \square$$

Lemma 2. *Let X be a finite set, and let $T \in \Lambda_m^X$ be a transmission function. Then for any $p' \in \Lambda^X$ and any $\epsilon > 0$, there exists a $\delta > 0$ such that for any $p \in \Lambda^X$,*

$$d(p\,,\,p') < \delta \Rightarrow d(\mathcal{V}_T p\,,\,\mathcal{V}_T p') < \epsilon$$

Sketch of Proof: Let $\mathcal{A} : \Lambda^X \to \Lambda^X$ be defined such that $(A(p))(x) = (\mathcal{V}_T p)(x)$. Let $\mathcal{B} : X \to [\Lambda^X \to [0,1]]$ be defined such that $(\mathcal{B}(x))(p) = (\mathcal{V}_T p)(x)$. The reader can check that for any $x \in X$, $\mathcal{B}(x)$ is a continuous function. The application of lemma 1 completes the proof.

By similar arguments, we obtain the following two lemmas.

Lemma 3. *Let X be a finite set, and let $f : X \to \mathbb{R}^+$ be a function. Then for any $p' \in \Lambda^X$ and any $\epsilon > 0$, there exists a $\delta > 0$ such that for any $p \in \Lambda^X$,*

$$d(p\,,\,p') < \delta \Rightarrow d(\mathcal{S}_f p\,,\,\mathcal{S}_f p') < \epsilon$$

Lemma 4. *Let X be a finite set, and let $p \in \Lambda^X$ be a distribution. Then for any $f' \in [X \to \mathbb{R}^+]$, and any $\epsilon > 0$, there exists a $\delta > 0$ such that for any $f \in [X \to \mathbb{R}^+]$,*

$$d(f\,,\,f') < \delta \Rightarrow d(\mathcal{S}_f p\,,\,\mathcal{S}_{f'} p) < \epsilon$$