

Open Problems in the Spectral Analysis of Evolutionary Dynamics

Lee Altenberg

Information and Computer Sciences

University of Hawai'i at Manoa, Honolulu, Hawai'i U.S.A.*

altenber@hawaii.edu

Abstract

For broad classes of selection and genetic operators, the dynamics of evolution can be completely characterized by the spectra of the operators that define the dynamics, in both infinite and finite populations. These classes include generalized mutation, frequency-independent selection, uniparental inheritance. Several open questions exist regarding these spectra:

1. For a given fitness function, what genetic operators and operator intensities are optimal for finding the fittest genotype? The concept of rapid first hitting time, an analog of Sinclair's "rapidly mixing" Markov chains, is examined.
2. What is the relationship between the spectra of deterministic infinite population models, and the spectra of the Markov processes derived from them in the case of finite populations?
3. Karlin proved a fundamental relationship between selection, rates of transformation under genetic operators, and the consequent asymptotic mean fitness of the population. Developed to analyze the stability of polymorphisms in subdivided populations, the theorem has been applied to unify the reduction principle for self-adaptation, and has other applications as well. Many other problems could be solved if it were generalized to account for the interaction of different genetic operators. Can Karlin's theorem on operator intensity be extended to account for mixed genetic operators?

Introduction

A general theory for the performance and design of evolutionary algorithms has proven difficult to achieve. This difficulty sets in even before we delve into search spaces with great complexity, or search operators with great complexity. We find it in the simplest

*Copyright ©2004 by Lee Altenberg. Chapter 4 in *Frontiers of Evolutionary Computation*, ed. Anil Menon, pp. 73-102. Genetic Algorithms And Evolutionary Computation Series, Vol. 11, Kluwer Academic Publishers, Boston, MA, 2004. Last corrigenda March 27, 2008.

“canonical” models of evolutionary algorithms owing to their nonlinear structure and stochastic dynamics.

Nonlinearity and stochasticity can be eliminated by making a variety of simplifying assumptions—in essence, exploring a subspace on the boundaries of the general problem. Linearity is produced by assuming constant selection and uniparental transmission (i.e. where the offspring type is determined by the type of its one parent). Determinism can be produced by assuming an infinite population size. These assumptions produce a linear dynamical system whose trajectory and attractors can be described in closed form, and decomposed in terms of its spectrum of eigenvalues and eigenvectors.

Actual evolutionary algorithms depart from this boundary in two important ways: finite populations, and recombination between two (or more) parents.

Recombination, a central innovation of genetic algorithms, is aimed at allowing combinations of partial solutions to be assembled. Recombination between two parents changes the dynamics of the infinite population model from linear to quadratic. In a quadratic system, we can no longer obtain a spectrum of eigenvalues and eigenvectors; the methods of nonlinear analysis must be employed, such as characterization of fixed points and their stability, domains of attraction, and Lyapunov functions.

A great deal of work has been on the dynamics of recombination and selection for models at various points on the boundaries of the general problem. A recent compendium can be found in Christiansen (2000). For more on quadratic dynamical systems see Rabinovich *et al.* (1992) and Arora *et al.* (1994). Progress has been made in the dynamics of recombination in the absence of selection, in both infinite and finite population models, by Rabani *et al.* (1995), and for simple selection, by Rabinovich and Wigderson (1999). Numerous analyses for other models on the boundary of the general problem can be found in the evolutionary computation and population genetics literature.

Evolutionary algorithms employ finite populations of a size considerably less than the cardinality of the search space, since a primary goal of the algorithms is to locate desired elements of the search space without exhaustive search.

Finite population algorithms typically use Bernoulli sampling to generate new samples of the search space. This changes the model of the algorithm from deterministic to stochastic, a Markov chain which has a linear state transition matrix, but whose dimensions are exponentially increased beyond the number of elements in the search space. The first model of finite population dynamics was developed based on Bernoulli sampling by Wright (1931) and Fisher (1930). In the Wright-Fisher model, the number of states in the Markov chain for the finite population model is $\mathcal{O}(N^{|\mathcal{S}|})$, compared to a dimension of $|\mathcal{S}|$ for the infinite population model, where $|\mathcal{S}|$ is the number of different genotypes, and N is the population size. Hence, the dimensionality of the state space is vastly increased in the finite population model over the infinite population model.

This comparison can be made more concrete by describing the difference in terms of points in the $|\mathcal{S}| - 1$ dimensional simplex. In the infinite population model, the system state is represented as a single point in the simplex which moves deterministically one generation to the next. In the finite population model, the state is represented as a probability distribution over a cloud of points in the simplex, restricted to the lattice of coordinates $\{\mathbf{x} : N x_i \in \{0, 1, \dots, N\}, \sum_{i=1}^n x_i = 1\}$. The distribution of the cloud of points is what changes every generation.

Because the uniparental, infinite population model has a complete solution, in terms of the spectrum of the linear operators, it presents the logical starting point to try to understand a number of unanswered questions in the design and dynamics of evolutionary algorithms. So I begin with the uniparental, infinite population model. There are three primary open questions I want to discuss:

1. What are the optimal transmission matrices for finding global optima of a search space?
2. What is the relationship between the spectrum of the infinite population model and the spectrum of the finite population model?
3. Can a key theorem of Karlin on the effects of operator intensity be generalized?

The Canonical Model

The ‘canonical’ model I shall be referring to throughout is the model of an infinite population evolving with discrete, non-overlapping generations, under constant fitness coefficients and generalized uniparental transmission. Let \mathbf{x} be the n -dimensional vector of frequencies of different types in the population, so $x_i \geq 0$, and $\sum_{i=1}^n x_i = 1$, which is to say that $\mathbf{x} \in \Delta_n$, the $n - 1$ -dimensional simplex. Then the recursion on \mathbf{x} is:

$$\mathbf{x}' = \frac{1}{\bar{w}} \mathbf{T} \mathbf{W} \mathbf{x}, \quad (1)$$

where \mathbf{x}' is the vector of frequencies in the next time step; \mathbf{W} is the diagonal matrix of fitness coefficients, $w_i \geq 0$;

$$\bar{w} = \sum_{i=1}^n w_i x_i$$

is the mean fitness of the population, used as a normalizer to maintain the system state as frequencies; and

$$\mathbf{T} = [T_{ij}]_{i,j=1}^n$$

is the n -by- n matrix of transmission probabilities, T_{ij} , the probability that type j produces an offspring of type i , so

$$\sum_{i=1}^n T_{ij} = 1 \quad \forall j, \quad T_{ij} \geq 0.$$

In vector form, these identities are:

$$\mathbf{1}^\top \mathbf{x} = 1, \quad \mathbf{1}^\top \mathbf{T} = \mathbf{1}, \quad \text{and} \quad \bar{w} = \mathbf{1}^\top \mathbf{W} \mathbf{x},$$

where

$$\mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

The trajectory of the system is:

$$\mathbf{x}(t) = \frac{1}{\nu(t)}(\mathbf{TW})^t \mathbf{x}(0), \quad (2)$$

where $\nu(t) = \mathbf{1}^\top (\mathbf{TW})^t \mathbf{x}(0)$ is the normalizer.

1 Optimal Evolutionary Dynamics for Optimization

For an optimization problem, we assume that an objective function $f : S \rightarrow \mathbb{R}_+$ is defined on each element of the search space; here, I assume that the goal is to find the element with maximum objective function value. Exhaustive search or random search of such a space will require on the average $n/2$ samples to have sampled an optimum if it is unique (which will be by assumption throughout unless specified otherwise). If an algorithm can find the optimum in an average of $\epsilon n/2$ samples, for some small constant $\epsilon \ll 1$, then it is clearly doing better than “blind search”.

However, evolutionary algorithms can perform much better than $\mathcal{O}(n)$. The canonical example for an “evolutionary algorithm-easy” problem is the ONEMAX problem, where the fitness increases with the number of loci that have 1 as their allelic value (Ackley, 1987). The number of samples required by a simple mutation-selection algorithm to find the global optimum in the ONEMAX problem is $\mathcal{O}(L) = \mathcal{O}(\log(n))$, where L the number of loci, $n = |\mathcal{A}|^L$ is the size of the search space, \mathcal{A} is the set of alleles for each locus, $|\mathcal{A}|$ the cardinality of \mathcal{A} (for binary strings, $|\mathcal{A}| = 2$).

So, as a performance goal, we would like the time complexity our evolutionary search to be on the order of the ONEMAX problem, taking $\mathcal{O}(\log(n))$ samples in order to find the global optimum. To be a little more lenient with the performance requirements, we can relax the condition for “EA-easy” to polylogarithmic time, meaning that it takes $\mathcal{O}(P(\log(n)))$ samples to find the optimum, where $P(\log(n))$ is a polynomial in $\log(n)$.

So, we wish to know what conditions on an evolutionary algorithm will allow it to find the global optimum in $\mathcal{O}(P(\log(n)))$ samples.

Evolutionary algorithms often have multiple domains of attraction (at least in the metastable sense (van Nimwegen *et al.*, 1999)), which imposes a secondary search problem: finding the initial conditions that are in the domain of attraction containing the global optimum. The multiple-attractor problem is usually described as “multimodality” of the fitness function, but it must be understood that the fitness function by itself does not determine whether the EA has multiple domains of attraction—it is only the relationship of the fitness function to the variation-producing operators that produces multiple-attractors (Altenberg, 1995).

In order to preclude this secondary search problem, we desire that the algorithm exhibit a single, global attractor that contains the global optimum.

So, we wish to find what spectral properties give rise to the following characteristics of an evolutionary algorithm:

1. **Rapid First Hitting Time:** It finds the global optimum using a number of samples that are $\mathcal{O}(P(\log(n)))$ where n , is the cardinality of the search space. I will call this the *rapid first hitting time* property.

2. **Global Attraction:** It finds the global optimum regardless of the initial samples taken, i.e. the simplex must have one global attractor containing the optimum.

Search problem that present obstacles to 1. include *long path problems*, and the *needle-in-a-haystack*. Search problem that present obstacles to 2. include *deception*, *rugged adaptive landscapes*, and *multimodal objective functions*.

1.1 Spectral Conditions for Global Attraction

For the canonical model Eq. (1), the *global attraction* condition, 2. above, can be stated precisely as:

$$\lim_{t \rightarrow \infty} \frac{1}{\nu(t)} (\mathbf{T}\mathbf{W})^t \mathbf{x}(0) = \pi, \text{ and } \pi_1 > 0, \forall \mathbf{x}(0) \in \Delta_n, \quad (3)$$

where we index the global optimum type as 1, so π_1 is its stationary frequency.

Condition (3) is guaranteed if and only if \mathbf{T} is primitive (irreducible and acyclic), i.e. there is some $v \geq 0$ such that $\mathbf{T}^v > 0$. From the Perron-Frobenius theorem (Gantmacher, 1959), primitiveness guarantees that there be a strictly positive eigenvector π corresponding to the leading eigenvalue of $\mathbf{T}\mathbf{W}$. This eigenvector π , normalized so $\langle \mathbf{1}, \pi \rangle = \sum_i \pi_i = 1$, is the global attractor, since the composition of the population converges to it regardless of the initial composition $\mathbf{x}(0)$.

Primitiveness in the transmission matrix corresponds to the property of ergodicity.

It should be noted that when some types have a fitness of 0, then their frequency becomes irrelevant to the dynamics, so the transmission probabilities $\{T_{ij} : j \in \mathcal{N}\}$, where $\mathcal{N} = \{i : w_i = 0\}$, are also irrelevant. Hence, primitiveness is required only for the restriction of \mathbf{T} to \mathbf{T}^+ , where

$$\mathbf{T}^+ = [T_{ij}]_{i,j \notin \mathcal{N}}.$$

For simplicity, I will henceforth assume all fitnesses are positive.

It should be noted that ergodicity in the infinite population model gives us little guarantee that the system in the finite population model will exhibit a global attractor, due to the phenomenon of metastability or *broken ergodicity* (Palmer, 1982). While ergodicity in the infinite population model is necessary for ergodicity in the finite population model, it is not sufficient. The Markov chain for the finite population model must in addition be *rapidly mixing* (Sinclair, 1992) to avoid broken ergodicity, as will be discussed later.

1.2 Spectral Conditions for Rapid First Hitting Times

What properties of \mathbf{T} and \mathbf{W} —which here completely define the canonical evolutionary algorithm—lead to rapid first hitting times? \mathbf{W} incorporates the map between the objective function and the fitness values, w_i , and we could certainly focus on the properties of this map. I can pose the following (without belaboring its precise details):

Open Question 1.1. *For a given transmission matrix, \mathbf{T} , what is the optimum selection scheme to find the global optimum with a rapid first hitting time?*

Here, however, since the canonical model assumes that \mathbf{W} is fixed, we wish to consider the problem for arbitrary \mathbf{W} . This leaves only \mathbf{T} , the transmission matrix, to be explored.

We can, without loss of generality, label the unique optimal point in the search space with $i = 1$, so

$$w_1 = \max_{i=1}^n w_i.$$

We can trivially guarantee a hitting time of 1 by simply constructing a transmission matrix that produces the optimum by mutation:

$$\mathbf{T} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}.$$

Transmission in this case is biased to find the optimum without any help from selection. Clearly, such a *a priori* knowledge does not capture the nature of the implicit knowledge that an evolutionary algorithm must contain to have rapid first hitting times (Altenberg, 1995). The essence of evolutionary search is that *transmission in the absence of selection is unable to produce adaptation or optimization*. Only when selection and transmission are combined does adaptation occur. The translation of this principle into a condition on \mathbf{T} would require that all types evolve to equal frequency in the absence of selection, i.e.

$$\lim_{t \rightarrow \infty} (\mathbf{T})^t \mathbf{x}(0) = \frac{1}{n} \mathbf{1}, \quad \forall \mathbf{x}(0) \in \Delta_n. \quad (4)$$

Condition (4) for “fair” transmission implies that

1. The transmission matrix is doubly stochastic, i.e. $\mathbf{T} \mathbf{1} = \mathbf{1}$;
2. The transmission matrix is primitive, i.e. irreducible and acyclic.

So, our question about the optimal characteristics of \mathbf{T} can be posed thus:

Open Question 1.2. *Given a fitness function on a points in a search space, what “fair” transmission matrix is optimal for finding the global optimum with rapid first hitting time?*

A rapid first hitting time refers to the number of samples that need to be taken before finding the global optimum. But in an infinite population, an infinite number of samples are taken each generation. So clearly, to adapt the infinite population model to the problem of rapid first hitting time, we need a proper translation.

In a finite population, with discrete, non-overlapping generations, the number of samples, s^* , until the optimum is found is:

$$s^* = N \tau \mu,$$

where N is the population size, τ is the first hitting time (in generations), and μ is the fraction of the population each generation that comprise new samples. Hence, to achieve rapid first hitting times, the population size and the first hitting time itself must each be polylogarithmic in $n = |S|$, the size of the search space, since $\mathcal{O}(P(\log(n))) * \mathcal{O}(P(\log(n))) = \mathcal{O}(P(\log(n)))$.

1.3 Rapid Mixing and Rapid First Hitting Times

Vitanyi (2000) has investigated the problem of rapid first hitting time in the finite population model, and proposes two criteria that will ensure rapid first hitting time:

1. the second-largest eigenvalue of the matrix representing the Markov process is bounded away far enough from 1 so that the Markov chain is rapidly mixing, as defined by Sinclair (1992).
2. the stationary distribution π gives probability greater than $1/P(\log(n))$ to the set of states that contain the global optima, where $P(\log(n))$ is a polynomial in the log of the size of the search space.

The identification of the second-largest eigenvalue as a measure of the speed of convergence of the Markov chain in evolutionary dynamics goes all the way back to Wright (1931) and Fisher (1930), who solved the second-largest eigenvalue for the Markov process representing the finite population model. This eigenvalue is $\lambda_3 = 1 - 1/N$ (since $\lambda_1 = \lambda_2 = 1$), where N is haplotype population size. It gives the rate of convergence to fixation on a single haplotype due to genetic drift, and is also the rate of decrease in the frequency of heterozygotes in the population. See Ewens (1979, pp. 17, 76, 79, 82, 85–90, 105–107, and Appendix B).

Other more recent work investigating the second-largest eigenvalue includes Suzuki (1995), Rudolph (1997), and Schmitt and Rothlauf (2001a,b)

The condition defined by Sinclair (1992) to produce what he calls *rapid mixing* in a Markov chain is as follows. Sinclair lays out his concept of rapid mixing by first defining the relative pointwise distance (r.p.d.) on a Markov process with transition matrix \mathbf{P} as:

$$d(t, n) = \max_{i, j \in \{1, \dots, n\}} \frac{\left| \left[\mathbf{P}^t \right]_{ij} - \pi_i \right|}{\pi_i},$$

where n is the cardinality of the state space for the chain. Additionally, one defines

$$\tau(\epsilon) = \min\{t \in \mathcal{Z}^+ : d(t', n) \leq \epsilon, \forall t' \geq t\}.$$

The Markov chain is said to be rapidly mixing if there exists a polynomial $P(\log(n), \log(1/\epsilon))$ such that:

$$\max_{\epsilon \in (0, 1]} \tau(\epsilon) \leq P(\log(n), \log(1/\epsilon))$$

(Sinclair, 1992, p. 56).

Rapid mixing concerns the rate of convergence of a Markov chain to its limiting probability distribution. The second-largest eigenvalue determines the rate at which the components of the probability distribution that are orthogonal to the limiting distribution die away. The definition of fast optimization which depends on rapid mixing I call *rapid first hitting time* by analogy.

I propose a slightly different set of criteria from Vitanyi (2000) to allow rapid first hitting time to be defined in the infinite population model. We can translate the above discussion into a condition for rapid first hitting time in the deterministic model thus:

Definition: Rapid First Hitting Time. Consider a deterministic evolutionary algorithm with a unique global optimum, which we set to be type 1, so $w_1 > w_i$ for all $i \in \{2, \dots, n\}$. Let

$$\tau(\epsilon) = \max_{\mathbf{x}(0) \in \Delta_n} \min\{t \in \mathcal{Z}^+ : x_1(t) \geq \epsilon\}.$$

The evolutionary algorithm is said to possess a rapid first hitting time if there exist polynomials $P_1(\log(n))$ and $P_2(\log(n))$ in $\log(n)$, such that

$$\epsilon \geq \frac{1}{P_1(\log(n))} \quad \text{and} \quad \tau(\epsilon) \leq P_2(\log(n)). \quad (5)$$

For the canonical evolutionary algorithm, $\mathbf{x}(t) = \frac{1}{\nu(t)}(\mathbf{TW})^t \mathbf{x}(0)$, this requires that for all $\mathbf{x}(0) \in \Delta_n$, there exist polynomials $P_1(\log(n))$ and $P_2(\log(n))$ such that:

$$x_1(P_2(\log(n))) = \frac{1}{\nu(t)} [1 \ 0 \ \dots \ 0] (\mathbf{TW})^{P_2(\log(n))} \mathbf{x}(0) \geq \frac{1}{P_1(\log(n))}. \quad (6)$$

Of course, it must be emphasized that this ‘translation’ carries with it no presumption that the infinite population model adequately approximates the behavior the first hitting time in the finite population model. The first hitting time is a concept that properly belongs to stochastic processes; it is a random variable. The use of the infinite population model to approximate the first hitting time has been taken before in the ‘takeover time’ models (Goldberg and Deb, 1991), where a deterministic, infinite population model is used to approximate the time to fixation of a genotype in a finite population. *It is clear* that this approximation will be inadequate and misleading under the very circumstances in which an evolutionary algorithm is of interest, namely, when it can find the fittest elements of the search space by sampling only a fraction of the search space. These circumstances will be discussed in Section 2.1. I claim only that this use of the infinite population model may lead us to results that may be worth investigating more rigorously in the finite population model.

1.4 Some Analysis

We can assume without significant loss of generality that \mathbf{TW} permits a Jordan canonical representation as

$$\mathbf{TW} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top, \quad (7)$$

where the matrix \mathbf{Q} consists of columns that are the eigenvectors of \mathbf{TW} , $\mathbf{Q}\mathbf{Q}^\top = \mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$, and $\mathbf{\Lambda}$ is the diagonal matrix $\Lambda_{ii} = \lambda_i$ of the eigenvalues of \mathbf{TW} . This assumption will simplify the analysis.

The condition applies if we assume that transition probabilities are symmetric, i.e. $T_{ij} = T_{ji}$, which is typical of the mutation operators used on data structures in evolutionary computation. This is verified by noting that since any symmetric matrix \mathbf{S} has Jordan form $\mathbf{S} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^\top$, so we can take $\mathbf{S} = \mathbf{W}^{1/2} \mathbf{TW}^{1/2} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^\top$, hence

$$\mathbf{TW} = \left(\mathbf{W}^{-1/2}\mathbf{P}\right) \mathbf{\Lambda} \left(\mathbf{P}^\top\mathbf{W}^{1/2}\right).$$

We must assume here that all fitnesses are non-zero, $w_i > 0$.

With this assumption we can then represent the trajectory of the population as:

$$\mathbf{x}(t) = \frac{1}{\nu(t)} \mathbf{Q} \mathbf{\Lambda}^t \mathbf{Q}^\top \mathbf{x}(0).$$

We can arbitrarily permute the indices so that $\lambda_1 > \lambda_2 \geq \dots \geq \lambda_n > -\lambda_1$, and so that $w_1 > w_2 \geq \dots \geq w_n$. Then for Q_{ij} , i follows the order of the fitnesses, while j follows the order of the eigenvalues. In particular,

$$\mathbf{q}_1 = c \pi > 0$$

is the strictly positive leading eigenvector of $\mathbf{T}\mathbf{W}$, with $c = \langle \mathbf{1}, \mathbf{q}_1 \rangle$ (note that by definition $\langle \mathbf{q}_i, \mathbf{q}_i \rangle = 1$). Thus:

$$\mathbf{T}\mathbf{W}\pi = \lambda_1 \pi.$$

The trajectory of the frequency of the optimal type is:

$$\begin{aligned} x_1(t) &= \frac{1}{\nu(t)} \sum_{i=1}^n q_{1i} \lambda_i^t [\mathbf{q}_i^\top \mathbf{x}(0)] \\ &= \frac{\lambda_1^t}{\nu(t)} \left(q_{11} \langle \mathbf{q}_1, \mathbf{x}(0) \rangle + \sum_{i=2}^n q_{1i} \left(\frac{\lambda_i}{\lambda_1} \right)^t \langle \mathbf{q}_i, \mathbf{x}(0) \rangle \right). \end{aligned} \quad (8)$$

Further evaluation of $\nu(t)$ yields:

$$\begin{aligned} \nu(t) &= \sum_{i=1}^n \mathbf{1}^\top \mathbf{q}_i \lambda_i^t \mathbf{q}_i^\top \mathbf{x}(0) \\ &= \lambda_1^t \left[\sum_{i=1}^n \langle \mathbf{1}, \mathbf{q}_i \rangle \left(\frac{\lambda_i}{\lambda_1} \right)^t \langle \mathbf{q}_i, \mathbf{x}(0) \rangle \right] \\ &= \lambda_1^t \left[\langle \mathbf{1}, \mathbf{q}_1 \rangle \langle \mathbf{q}_1, \mathbf{x}(0) \rangle + \sum_{i=2}^n \langle \mathbf{1}, \mathbf{q}_i \rangle \left(\frac{\lambda_i}{\lambda_1} \right)^t \langle \mathbf{q}_i, \mathbf{x}(0) \rangle \right] \\ &= \lambda_1^t \left[c \langle \mathbf{q}_1, \mathbf{x}(0) \rangle + \sum_{i=2}^n \langle \mathbf{1}, \mathbf{q}_i \rangle \left(\frac{\lambda_i}{\lambda_1} \right)^t \langle \mathbf{q}_i, \mathbf{x}(0) \rangle \right], \end{aligned}$$

using $c = \langle \mathbf{1}, \mathbf{q}_1 \rangle$. So we obtain:

$$x_1(t) = \frac{q_{11} \langle \mathbf{q}_1, \mathbf{x}(0) \rangle + \sum_{i=2}^n q_{1i} \left(\frac{\lambda_i}{\lambda_1} \right)^t \langle \mathbf{q}_i, \mathbf{x}(0) \rangle}{c \langle \mathbf{q}_1, \mathbf{x}(0) \rangle + \sum_{i=2}^n \langle \mathbf{1}, \mathbf{q}_i \rangle \left(\frac{\lambda_i}{\lambda_1} \right)^t \langle \mathbf{q}_i, \mathbf{x}(0) \rangle}$$

Substituting the above into (6), setting $t = P_2(\log(n))$, and rearranging, we obtain the condition:

$$\begin{aligned} [P_1(\log(n)) q_{11} - c] \langle \mathbf{q}_1, \mathbf{x}(0) \rangle &\geq \\ \sum_{i=2}^n \left(\frac{\lambda_i}{\lambda_1} \right)^{P_2(\log(n))} [\langle \mathbf{1}, \mathbf{q}_i \rangle - P_1(\log(n)) q_{1i}] \langle \mathbf{q}_i, \mathbf{x}(0) \rangle &\end{aligned} \quad (9)$$

Since $\mathbf{q}_1 = c\pi$, we substitute $P_1(\log(n))q_{1i} - c = c[P_1(\log(n))\pi_1 - 1]$, and $\langle \mathbf{q}_1, \mathbf{x}(0) \rangle = c\langle \pi, \mathbf{x}(0) \rangle$, to get:

$$c^2 [P_1(\log(n))\pi_1 - 1] \langle \pi, \mathbf{x}(0) \rangle \geq \sum_{i=2}^n \left(\frac{\lambda_i}{\lambda_1} \right)^{P_2(\log(n))} (\langle \mathbf{1}, \mathbf{q}_i \rangle - P_1(\log(n))q_{1i}) \langle \mathbf{q}_i, \mathbf{x}(0) \rangle, \quad (10)$$

$\forall \mathbf{x}(0) \in \Delta_n$.

At this point, we take interest in the second-largest eigenvalue λ_2 . Let us define

$$r = \lambda_2/\lambda_1. \quad (11)$$

For any $\delta > 0$, if r is small enough, then

$$\begin{aligned} \delta &\geq \left| \sum_{i=2}^n r^{P_2(\log(n))} (\langle \mathbf{1}, \mathbf{q}_i \rangle - P_1(\log(n))q_{1i}) \langle \mathbf{q}_i, \mathbf{x}(0) \rangle \right| \\ &\geq \left| \sum_{i=2}^n \left(\frac{\lambda_i}{\lambda_1} \right)^{P_2(\log(n))} (\langle \mathbf{1}, \mathbf{q}_i \rangle - P_1(\log(n))q_{1i}) \langle \mathbf{q}_i, \mathbf{x}(0) \rangle \right| \geq 0. \end{aligned}$$

In this case, condition (10) is met provided

$$[P_1(\log(n))\pi_1 - 1] \langle \pi, \mathbf{x}(0) \rangle \geq \delta/c^2$$

or

$$\pi_1 \geq \frac{1 + \frac{\delta}{c^2 \langle \pi, \mathbf{x}(0) \rangle}}{P_1(\log(n))} > \frac{1}{P_1(\log(n))}. \quad (12)$$

Hence, for small enough r , the only condition for rapid first hitting time is that the frequency of the optimum at equilibrium be on the order of $P_1(\log(n))^{-1}$. We know that selection is required in order for $\pi_1 \geq \frac{1}{P_1(\log(n))}$ since the principle eigenvector of \mathbf{T} has $\pi_1 = \frac{1}{n}$ by the fairness assumption. Thus:

Theorem 1. *If the system $\mathbf{x}(t) = \frac{1}{\nu(t)}(\mathbf{TW})^t \mathbf{x}(0)$ exhibits rapid first hitting time, then there exists a critical value $\sigma^* \in [0, 1)$ such that the system $\mathbf{x}(t) = \frac{1}{\nu(t)}(\mathbf{TW}^\sigma)^t \mathbf{x}(0)$ no longer exhibits rapid first hitting time for all $\sigma \leq \sigma^*$.*

Characterizing the dependence of σ on \mathbf{T} and \mathbf{W} remains an open question.

Now, it remains to be asked, what transmission matrices \mathbf{T} minimize $r = \frac{\lambda_2}{\lambda_1}$?

1.5 Transmission Matrices Minimizing λ_2/λ_1

If we find a transmission matrix that gives $r = \lambda_2/\lambda_1 = 0$, then the only condition we require for rapid first hitting time is (12). The rank-1 matrix yields $r = 0$:

$$\mathbf{T} = \mathbf{U} = \frac{1}{n} \mathbf{1} \mathbf{1}^\top = \frac{1}{n} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & & & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}.$$

We have $\lambda_1(\mathbf{U}) = 1$, and $\lambda_2(\mathbf{U}) = \dots = \lambda_n(\mathbf{U}) = 0$. When we include selection:

$$\mathbf{U}\mathbf{W} = \frac{1}{n} \mathbf{1} \mathbf{1}^\top \mathbf{W} = \frac{1}{n} \mathbf{1} [w_1 \ w_2 \ \dots \ w_n] = \frac{1}{n} \begin{bmatrix} w_1 & w_2 & \dots & w_n \\ w_1 & w_2 & \dots & w_n \\ \vdots & & & \vdots \\ w_1 & w_2 & \dots & w_n \end{bmatrix}$$

is also a rank-1 matrix, with eigenvalues $\lambda_1(\mathbf{U}\mathbf{W}) = \frac{1}{n} \sum_{i=1}^n w_i$, and $\lambda_2(\mathbf{U}\mathbf{W}) = \dots = \lambda_n(\mathbf{U}\mathbf{W}) = 0$.

Thus, it would appear that the rank-1 matrix would be a candidate transmission matrix to achieve rapid first hitting times. However, this hope is instantly dashed by noting that for $\mathbf{U}\mathbf{W}$, $\pi_1 = 1/n$, which is not greater than $1/P_1(\log(n))$. We might ask if we can find another rank-1 matrix where $\pi_1 \geq 1/P_1(\log(n))$, but this is precluded by the condition that \mathbf{T} be ‘fair’, and thus doubly stochastic, requiring that $\pi_i = 1/n$ for all i . This result is not unexpected, when we consider that the rank-1 matrix corresponds to random search.

So, we are left with the following:

Open Question 1.3. *For a given set of fitnesses, \mathbf{W} , what classes of fair transmission matrices maximize π_1 while minimizing $r = \lambda_2/\lambda_1$ so as to satisfy the conditions for rapid first hitting time?*

One step we may take in defining the notion of classes of transmission matrices is to note that the *topology* of transmission may be separated from the *operator intensity* by the following parameterization:

$$\mathbf{T}(\mu) = (1 - \mu)\mathbf{I} + \mu \mathbf{P}, \quad (13)$$

where $\mu \in [0, 1]$ is the mutation rate, and \mathbf{P} is a transmission matrix in which at least one value $P_{ii} = 0$ (Altenberg and Feldman, 1987). For those genetic operators that can be represented as graphs, where a vertex represents a type, and an edge represents an operator transformation from one type to another, then \mathbf{P} naturally corresponds to a normalized adjacency matrix for the graph.

We can see immediately that if $\mu = 0$, the matrix $\mathbf{T}(\mu)\mathbf{W}$ becomes reducible, so if $x_1(0) = 0$, then $x_1(t) = 0$ for all t . For small μ , the following should be readily shown:

Conjecture 1. *If the system $\mathbf{x}(t) = \frac{1}{\nu(t)} (\mathbf{T} \mathbf{W})^t \mathbf{x}(0)$ exhibits rapid first hitting time, then the system*

$$\mathbf{x}(t) = \frac{1}{\nu(t)} ((1 - \mu)\mathbf{I} + \mu \mathbf{T}\mathbf{W})^t \mathbf{x}(0)$$

will exhibit rapid first hitting time for $\mu \in [\frac{1}{P(\log(n))}, 1]$, for some polynomial in $\log(n)$, $P(\log(n))$, and will not exhibit rapid first hitting time for $\mu \in [0, \frac{1}{n}]$.

Let us return to the example of the ONEMAX problem as the paradigmatic EA-easy problem. The transmission matrix for the ONEMAX problem is simple bit-flip

mutation, which produces an L -dimensional binary hypercube when represented as a graph between genotypes that mutate to one another. When fitnesses are permuted to the proper order (which Liepins and Vose (1990) prove can always be done), the problem becomes the ONEMAX problem. Hence, one can conjecture that a transmission matrix representing the binary hypercube would be a primary candidate for rapid first hitting time. However, it is clear that \mathbf{W} can be designed for which no rapid first hitting time can be achieved:

Conjecture 2. *It is possible to choose ϵ small enough so that if*

$$|\{i : \epsilon > w_1 - w_i > 0\}| \approx \mathcal{O}(n),$$

then there exists no fair transmission matrix that can produce rapid first hitting time.

With the proper constraints on \mathbf{W} , however, we may find the following:

Conjecture 3. *Consider a search space, \mathcal{S} , with $|\mathcal{S}| = n = 2^L$. Let the fitness values be $w_i = e^{-\sigma^i}$. Consider a binary encoding of the indices, $B(i)$, such that $w_i > w_j$ if and only if the Hamming distances, $H[\cdot, \cdot]$, between the binary encodings satisfies $H[B(1), B(i)] < H[B(1), B(j)]$. Let $\mathbf{T} = (1 - \mu)\mathbf{I} + \mu \mathbf{P}$, $0 \leq \mu \leq 1$, where \mathbf{P} is the normalized adjacency matrix for the L -dimensional binary hypercube Q_L under this encoding. Then for some $\sigma^* > 0$, if $\sigma \geq \sigma^*$, then there exists $\mu(\sigma)$ such that the system $\mathbf{x}(t) = (\mathbf{TW})^t \mathbf{x}(0) / \nu(t)$ has rapid first hitting time.*

Other examples of evolutionary systems that attain rapid first hitting times can be found in Vitanyi (2000).

We may also consider a class of transmission matrices which can never achieve rapid first hitting time for any set of fitnesses, namely, the “long path” (Horn *et al.*, 1994) matrices:

Conjecture 4. *Let $\mathbf{T} = (1 - \mu)\mathbf{I} + \mu \mathbf{P}$, where $P_{ij} = P_{1n} = P_{n1} = 1/2$ for $|i - j| = 1$, $P_{ij} = 0$ otherwise:*

$$\mathbf{P} = \frac{1}{2} \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 & 1 \\ 1 & 0 & 1 & & & & 0 \\ 0 & 1 & 0 & 1 & & & 0 \\ \vdots & & \ddots & \ddots & \ddots & & \vdots \\ 0 & & & 1 & 0 & 1 & 0 \\ 0 & & & & 1 & 0 & 1 \\ 1 & 0 & 0 & \cdots & 0 & 1 & 0 \end{bmatrix}$$

Then, there are no fitnesses \mathbf{W} , nor values μ , such that the system

$$\mathbf{x}(t) = \frac{1}{\nu(t)} (\mathbf{TW})^t \mathbf{x}(0)$$

has rapid first hitting time.

1.6 Rapid First Hitting Time and No Free Lunch Theorems

It should be noted that the concept of rapid first hitting times allows us to distinguish between transmission matrices in a way that the No-Free-Lunch Theorem (Wolpert and Macready, 1995, 1997) cannot.

The No-Free-Lunch Theorem, as applied to the current context, states that all transmission matrices have the same performance when averaged over all permutations of a set of fitnesses. However, Wolpert and Macready (1995) point out that search algorithms can be distinguished using minimax properties. In this case, an example of a minimax property is whether permutations of fitnesses exist for a given transmission matrix that produce rapid first hitting times.

So, while a long-path operator and a binary hypercube operator will have the same average performance in locating the global optimum over all permutations of fitnesses, they can be distinguished by their potential for rapid first hitting time. With an adequate distribution of fitness values, the binary hypercube makes possible permutations that produce ONEMAX problems having a rapid first hitting time. The long-path operator, on the other hand, allows no permutation, for any distribution of fitnesses, that can produce a rapid first hitting time. In this way, we can make a definite judgement that the binary hypercube is superior to the long-path operator for optimization.

Numerous possible directions exist for further investigation into these open questions about rapid first hitting time. I will leave these to forthcoming work.

2 Spectra for Finite Population Dynamics

One of the important open questions in evolutionary computation is the relationship between the dynamics of the infinite and the finite population models. The Wright-Fisher model of finite populations (Wright, 1931; Fisher, 1930)¹ is derived from the canonical model of an infinite population by the addition of only one free parameter—the population size. It thus provides the ideal model in which to pose this question.

2.1 Wright-Fisher Model of Finite Populations

In the Wright-Fisher model of a finite population, selection and genetic operators act on the current members of the population to produce a probability distribution from which each member of the population for the next generation is drawn independently. It is as if an infinite zygote pool was created, weighted by selection, from which only finite many can survive, each with equal probability.

The elements of the Wright-Fisher model are mostly the same as for the infinite population model. Let:

N be the population size;

¹The Markov model by Wright and Fisher is known in the genetic algorithms community as the “Nix and Vose model” (Nix and Vose, 1991). Other work on Markov chains in genetic algorithms includes Goldberg and Segrest (1987), and (Davis and Principe, 1993), since this community developed largely without awareness of prior work in mathematical population genetics.

\mathbf{x} be the vector of frequencies of each type i in the population, corresponding to $N x_i$ individuals of type i ;

\mathbf{x}' be the vector of the frequencies of each type i in the population in the next generation, corresponding to $N x'_i$ individuals of type i , produced by taking N independent samples from the distribution $\mathbf{y}(\mathbf{x})$;

$\mathbf{y}(\mathbf{x}) = \frac{1}{w} \mathbf{T} \mathbf{W} \mathbf{x}$ be the vector representing the probability distribution for drawing an individual of type i to compose the population in the next generation. \mathbf{T} and \mathbf{W} again represent the transmission matrix and fitness matrix, respectively.

Since the population consists of discrete individuals, the frequency vectors are now restricted to a lattice of discrete points on the simplex Δ_n , namely

$$\Delta_n(N) = \{\mathbf{x} : N x_i \in \{0, 1, \dots, N\}, \sum_{i=1}^n x_i = 1\}.$$

The Wright-Fisher model forms a Markov chain, whose transition matrix on frequency vectors is:

$$\mathbf{M} = \left[M_{\mathbf{x}', \mathbf{x}} \right]_{\mathbf{x}, \mathbf{x}' \in \Delta_n(N)}$$

with entries

$$M_{\mathbf{x}', \mathbf{x}} = N! \prod_{i=1}^n \frac{y_i^{N x'_i}}{(N x'_i)!} = \frac{N!}{\prod_{i=1}^n (N x'_i)!} \prod_{i=1}^n \left(\frac{\mathbf{e}_i^\top \mathbf{T} \mathbf{W} \mathbf{x}}{\mathbf{1}^\top \mathbf{W} \mathbf{x}} \right)^{N x'_i}$$

where $\mathbf{e}_i^\top = [0 \ 0 \ \dots \ 1 \ \dots \ 0]$ has the 1 in the i th position.

If we make the assumption that \mathbf{T} is primitive, and $T_{ij} = T_{ji}$, then we may employ the Jordan form (7):

$$M_{\mathbf{x}', \mathbf{x}} = \frac{N!}{\prod_{i=1}^n (N x'_i)!} \prod_{i=1}^n \left(\frac{\sum_{j=1}^n q_{ij} \lambda_j \langle \mathbf{q}_j, \mathbf{x} \rangle}{\sum_{j=1}^n w_j x_j} \right)^{N x'_i}. \quad (14)$$

Wright and Fisher analyzed some simple cases for this Markov system and derived a number of their properties, including rates of convergence, probabilities of fixation, time to fixation, and stationary distributions of allele frequencies.

In the special case of $\mathbf{T} = \mathbf{W} = \mathbf{I}$, and $n = 2$, the solution for all the eigenvalues of \mathbf{M} was found by Feller (1951), and by Cannings (1974) through his method of “exchangeable processes” (see Ewens 1979, pp. 77–79). The solution is:

$$\bar{\lambda}_1 = 1 \quad \text{and} \quad \bar{\lambda}_i = \prod_{j=2}^i \left(1 - \frac{j-2}{N} \right), \quad i \in \{2, \dots, N+1\},$$

where $\bar{\lambda}_i$ refers to the eigenvalues of \mathbf{M} , not of $\mathbf{T} \mathbf{W}$.

Regrettably, the method of exchangeable processes can not be applied when different individuals have different offspring probability distributions. We are therefore left with the following:

Open Question 2.1. *What is the relationship between the eigenvalues and eigenvectors of $\mathbf{T}\mathbf{W}$ and those of \mathbf{M} ?*

Since \mathbf{M} is defined explicitly in terms of the eigenvalues and eigenvectors of $\mathbf{T}\mathbf{W}$ in (14), establishing their relationship with the eigenvalues of \mathbf{M} is simply a matter of algebra. The complexity of the algebra, however, obscures the relationship. One may be able to simplify the sums in (14) by making assumptions that cause one term to dominate the sum, for example, if $\lambda_2, \dots, \lambda_n = 0$. But the utility of such an approach has yet to be demonstrated.

One can nevertheless make the following observations. Because the state space of the system is restricted to the lattice $\Delta_n(N) \subset \Delta_n$, and the situation of interest is when $N \approx \mathcal{O}(\log(n)) \ll n$, the vast majority of the entries of any $\mathbf{x} \in \Delta_n(N)$ must be 0. Thus, $\Delta_n(N)$ has no points on the interior of Δ_n , and is in fact restricted to the low-dimensional boundaries of Δ_n .

Thus, the indices of the non-zero components of \mathbf{x} make up a sparse set. Let us define the sparse set:

$$\Psi(\mathbf{x}) = \{i : x_i > 0\} \quad (15)$$

Then we may rewrite (14) as:

$$M_{\mathbf{x}', \mathbf{x}} = \frac{N!}{\prod_{i \in \Psi(\mathbf{x}')} (Nx'_i)!} \prod_{i \in \Psi(\mathbf{x}')} \left(\frac{\sum_{j=1}^n q_{ij} \lambda_j \sum_{k \in \Psi(\mathbf{x})} q_{jk} x_k}{\sum_{j \in \Psi(\mathbf{x})} w_j x_j} \right)^{Nx'_i}. \quad (16)$$

The trajectory of points in the finite population model will be radically different from the trajectory in the infinite population model. In the finite population model, a probability distribution will move over the surface of Δ_n , while in the infinite population model, the system will enter the interior of Δ_n within v generations, for the v that gives $\mathbf{T}^v > 0$ and makes \mathbf{T} primitive. Evolution in the finite population model can be viewed as transitions between one k -dimensional ($k \leq N$) edge of Δ_n and another, with the probability of transition being highest for types i where the terms $\sum_{j=1}^n q_{ij} \lambda_j \langle \mathbf{q}_j, \mathbf{x} \rangle$ are the largest.

My earlier discussion of the transmission matrix representing the binary hypercube took place in the context of the infinite population model. I conjectured that in the infinite population dynamics it will exhibit rapid first hitting time properties. However, it seems apparent that in the finite population model, the binary hypercube mutation will be especially advantageous in traversing the low-dimensional boundaries of the simplex.

I suspect that methods which can analyze (14) as a flow along the low-dimensional boundaries of the simplex may prove to be most helpful in understanding finite population dynamics. In the work of van Nimwegen (1999) we find this approach applied to specific models of mutation and selection, with a nice harvest of analytical results. Answers to the general spectral problem, however, await discovery.

2.2 Rapid First Hitting Time in a Finite Population

For a Wright-Fisher model, we can define the criteria for rapid first hitting time in terms of the actual first hitting time for the Markov chain. Here I depart only slightly from

Vitanyi (2000).

Let us refer to the set of populations that contain the global optimum as:

$$\mathcal{B}^+ = \{\mathbf{x} \in \Delta_n(N) : x_1 > 0\},$$

and conversely, the set of populations that do not contain the global optimum as:

$$\mathcal{B}^- = \{\mathbf{x} \in \Delta_n(N) : x_1 = 0\},$$

Suppose that the population always begins fixed on one type other than the optimum, so $\mathbf{x}(0) = \mathbf{e}_i, i \in \{2, \dots, n\}$.

We cannot use \mathbf{M} itself to calculate the probability that the first transition from \mathcal{B}^- to \mathcal{B}^+ occurs at time t , because in calculating $\left[\mathbf{M}^t\right]_{\mathbf{x}^+, \mathbf{x}^-}$ where $\mathbf{x}^+ \in \mathcal{B}^+$ and $\mathbf{x}^- \in \mathcal{B}^-$, we can't know whether this transition is the first transition. However, by setting all the elements of $\mathbf{M}_{\mathbf{x}^-, \mathbf{x}^+}$ to 0, where $\mathbf{x}^- \in \mathcal{B}^-$ and $\mathbf{x}^+ \in \mathcal{B}^+$, we generate a Markov process in which \mathcal{B}^+ is an absorbing set, hence transition probabilities within \mathcal{B}^- after t iterations include the probability that the first hitting time has not yet occurred.

This modified matrix is equivalent to restriction of \mathbf{M} to \mathcal{B}^+ , which we shall call $\overline{\mathbf{M}}$. Then the probability that the global optimum first appears after generation τ is:

$$\mathbf{G}_i(t) = \sum_{\mathbf{x}^- \in \mathcal{B}^-} \left[\overline{\mathbf{M}}^t\right]_{\mathbf{x}, \mathbf{e}_i} = \langle \mathbf{1}, \overline{\mathbf{M}}^t \mathbf{e}_i \rangle$$

We can define the criteria for rapid first hitting time in terms of the speed at which $G_i(\tau)$ declines with τ .

As a basis for comparison, we can consider how $G_i(\tau)$ behaves for random search, i.e. $\mathbf{M} = \frac{1}{\hat{n}} \mathbf{1} \mathbf{1}^\top$, where $\hat{n} = |\Delta_n(N)|$. Then $\mathbf{G}_i(t) = \sum_{\mathbf{x}^- \in \mathcal{B}^-} \frac{1}{\hat{n}} = \left(1 - \frac{1}{\hat{n}}\right)^N$ for all i . So

$$G(\tau) = \left(1 - \frac{1}{\hat{n}}\right)^{\tau N}.$$

In order for $G(\tau)$ to be reduced to $\mathcal{O}(1)$, to be specific, say $G(\tau) \leq \frac{1}{e}$, we have:

$$\log[G(\tau)] = \tau N \log\left(1 - \frac{1}{\hat{n}}\right) \leq -1,$$

thus for large n , $\log\left(1 - \frac{1}{\hat{n}}\right) \approx -\frac{1}{\hat{n}}$, hence $\tau N \geq \hat{n}$, which is what we expect. The essential idea for rapid first hitting time is that we would like $\tau N \leq P(\log(n))$. The obvious candidate for a condition to define rapid first hitting time would be:

Definition: Rapid First Hitting Time in a Finite Population. *The evolutionary algorithm is said to possess a rapid first hitting time if there exist polynomials in $\log(n)$, $P_1(\log(n))$ and $P_2(\log(n))$, such that $N \leq P_1(\log(n))$, and*

$$\max_{i \in \mathcal{B}^-} \langle \mathbf{1}, \overline{\mathbf{M}}^\tau \mathbf{e}_i \rangle \leq \frac{1}{e}, \text{ for all } \tau \geq P_2(\log(n)), \quad (17)$$

Clearly, the smaller the spectral radius of $\overline{\mathbf{M}}$, the more easily that rapid first hitting time is achieved. We are then ready to pose the main open question regarding the spectrum of evolutionary systems:

Open Question 2.2. *What conditions on the eigenvalues and eigenvectors of $\mathbf{T}\mathbf{W}$ satisfy condition (17) for rapid first hitting time?*

3 Karlin's Spectral Theorem for Genetic Operator Intensity

Samuel Karlin derived theorem of fundamental significance for evolutionary dynamics in an article that examines the role of population subdivision in maintaining genetic diversity (Karlin, 1982). The problem at hand was to understand whether migration would enhance or inhibit the maintenance of genetic diversity. Karlin took the approach of finding the conditions that would prevent an allele from becoming extinct, i.e. which would cause it to increase when rare. To this end, he proved a general theorem on the spectral radius of the stability matrix, which solved his problem for the case of any number of demes, any migration pattern, and any selection regime—quite extraordinary in its generality. The theorem is presented below. The square matrix \mathbf{P} represents the migration pattern, ξ represents the overall migration rate, and the diagonal matrix \mathbf{W} represents the average (or 'marginal') selection coefficients of an allele when rare.

Theorem 2 (Karlin, 1982).

Let

$$\mathbf{M}(\xi) = (1 - \xi)\mathbf{I} + \xi\mathbf{P},$$

where \mathbf{P} is an irreducible Markov matrix, and let \mathbf{W} be a diagonal matrix with strictly positive diagonal elements, where $\mathbf{W} \neq a\mathbf{I}$, for any scalar a . Then the spectral radius $\rho(\mathbf{M}(\xi)\mathbf{W})$ is strictly decreasing in ξ :

$$\frac{d}{d\xi}\rho(\mathbf{M}(\xi)\mathbf{W}) < 0, \text{ for } 0 \leq \xi \leq 1.$$

The matrix $\mathbf{M}(\xi)\mathbf{W}$ represents the stability matrix for the introduction of an allele into this multi-deme system. The allele goes extinct when rare if $\rho(\mathbf{M}(\xi)\mathbf{W}) < 1$, and is protected from extinction if $\rho(\mathbf{M}(\xi)\mathbf{W}) > 1$. Since $\rho(\mathbf{M}(\xi)\mathbf{W})$ decreases with ξ , the consequence of Karlin's result is that more migration makes it more difficult to maintain genetic diversity.

Karlin's result, obtained to answer a question about migration, proves to have much deeper significance for evolutionary dynamics. It reveals a basic property of the relation between selection and genetic operators. Its first application outside of the migration question was by Altenberg (Altenberg, 1984; Altenberg and Feldman, 1987) in modifier gene theory. A number of studies in the theory of modifier genes — genes that control the genetic system, such as rates of mutation (Karlin and McGregor, 1974), recombination (Feldman, 1972; Feldman and Balkau, 1973; Teague, 1976; Feldman

et al., 1980), and migration (Teague, 1977)— all found same result: new alleles which reduced these rates could always invade a population, suggesting a general “reduction principle”. Altenberg (1984) showed that all of these models could be unified with a common representation, $\mathbf{M}(\xi)\mathbf{W}$. With this unified representation, the application of Karlin’s theorem immediately proves the reduction principle: when an allele that modifies rates of mutation, recombination, migration, or any other transformation of type is introduced into a population near equilibrium, it will increase in frequency if it uniformly reduces the rates of transformation, and go extinct if it uniformly increases the rates of transformation.

Another immediate result from Karlin’s theorem regards the mean fitness under a mutation-selection balance. The mean fitness of haploid system decreases with increasing mutation rates:

Corollary 1. *Consider an evolutionary system consisting of*

- *constant selection,*
- *asexual genetic operators, and*
- *discrete, non-overlapping generations.*

The mean fitness of the population at an attractor is a decreasing function of the probability of the genetic operator acting.

Proof. Let the asexual genetic operator be represented by the Markov matrix \mathbf{M} , and let μ is the probability of applying the operator. Then the transmission matrix for the algorithm is:

$$\mathbf{T} = (1 - \mu)\mathbf{I} + \mu\mathbf{M},$$

and the recursion for discrete, non-overlapping generations is:

$$\bar{w} \mathbf{x}' = [(1 - \mu)\mathbf{I} + \mu\mathbf{M}] \mathbf{W} \mathbf{x}$$

For the global attractor, $\hat{\mathbf{x}}$, which is the leading eigenvector of $\mathbf{T}\mathbf{W}$ (whose existence and positive value are established by the Perron-Frobenius Theorem (Gantmacher, 1959)), we have:

$$[(1 - \mu)\mathbf{I} + \mu\mathbf{M}] \mathbf{W} \hat{\mathbf{x}} = \mathbf{T} \mathbf{W} \hat{\mathbf{x}} = \hat{\mathbf{x}} \rho(\mathbf{T}\mathbf{W}) = \hat{\mathbf{x}} \bar{w}.$$

Hence the mean fitness of the global attractor,

$$\hat{\bar{w}} = \rho(\mathbf{T}(\mu)\mathbf{W}),$$

is a decreasing function of the operator probability μ . ■

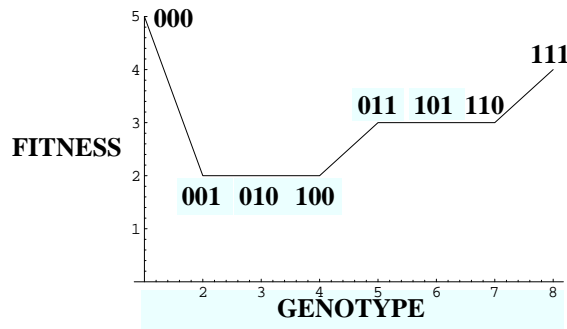


Figure 1: The Deceptive Trap fitness landscape for three loci with two alleles.

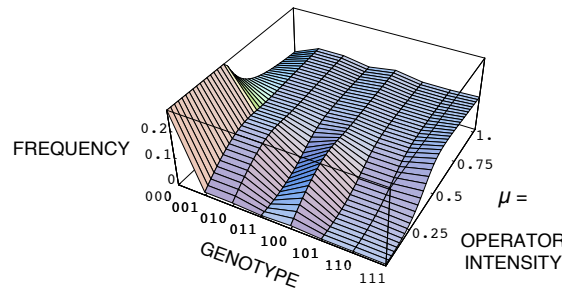


Figure 2: There is only one attractor at each value μ , but an ‘error catastrophe’ is evident for $\mu \approx 0.5$.

3.1 Karlin’s Theorem illustrated with the Deceptive Trap Function

Suppose a mutation operator is ergodic: i.e. repeated application of the operator can mutate any genotype into any other genotype. Then, under an algorithm of constant selection and mutation, the Perron-Frobenius Theorem shows that there is only one domain of attraction of the system—i.e. one ‘fitness peak’, as discussed in Section 1.1. This may seem contradictory to intuition about ‘multi-modal’ fitness landscapes, in which one would expect multiple domains of attraction. But multiple domains do not occur in haploid, infinite population models under ergodic mutation; finite populations are required to produce quasi-stability of multiple attractors. The global nature of the attractor for ergodic mutation under infinite population size is illustrated with the Deceptive Trap fitness landscape (Ackley, 1987), shown in Figure 1. In terms of the hypercube topology of the graph representing mutation, this is a bimodal fitness function. The frequency vector of the global attractor is shown as a function of the mutation rate, for a simple point mutation model, in Figure 2. The mean fitness of the attractor is seen to decrease as a function of the mutation rate, as the Karlin theorem proves. This is shown in Figure 3.

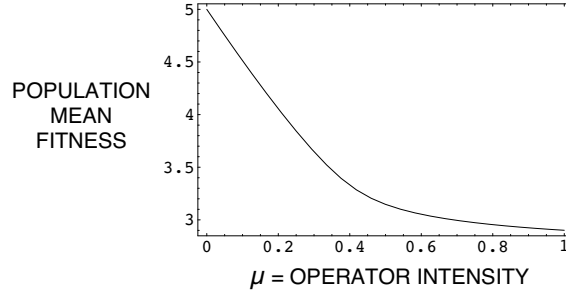


Figure 3: The mean fitness of the population at the global attractor as a function of mutation rate. It decreases in accord with Karlin’s theorem.

3.2 Applications for an Extended Karlin Theorem

Several problems are encountered for which an extended Karlin theorem would allow solution, but which are currently unsolved. One of these is in modifier theory. This has been called ‘self-adaptation’ (Schwefel, 1987; Bäck, 1996) in the Evolutionary Computation literature as. In Altenberg (1984) and Altenberg and Feldman (1987) it is proven that the Reduction Principle for linear variation in transmission holds for modifiers that are tightly linked to haplotypes under viability selection (Altenberg and Feldman, 1987, Result 3, p. 565). It is conjectured that the result would also hold for looser linkage to the modifier locus. The analysis requires that we show for $r > 0$ that the spectral radius of $\mathbf{M}(\mu, r)\mathbf{W}$ decreases in μ , where:

$$\mathbf{M}(\mu, r) = (1 - \mu)[(1 - r)\mathbf{I} + r\mathbf{Q}] + \mu[(1 - r)\mathbf{S} + r\tilde{\mathbf{S}}],$$

with \mathbf{Q} , \mathbf{S} , and $\tilde{\mathbf{S}}$ being Markov matrices (see Altenberg and Feldman (1987) for details). The proof awaits an extension of Karlin’s theorem for $r > 0$.

The other context of unsolved problems occurs when several transformation processes act on types in the population, such as the simultaneous action of mutation, recombination, and migration. This can result in recursions of the form:

$$\begin{aligned} \bar{w}\mathbf{x}' &= \mathbf{M}(\mu, \beta, \gamma)\mathbf{W}\mathbf{x} \\ &= [(1 - \mu)\mathbf{I} + \mu\mathbf{A}][(1 - \beta)\mathbf{I} + \beta\mathbf{B}][(1 - \gamma)\mathbf{I} + \gamma\mathbf{C}]\mathbf{W}\mathbf{x}, \end{aligned}$$

where \mathbf{A} , \mathbf{B} , and \mathbf{C} are Markov matrices representing different transformation processes, and μ , β , and γ are the overall rates of those processes. We wish to know how the spectral radius of $\mathbf{M}(\mu, \beta, \gamma)\mathbf{W}$ changes as a function of each parameter μ , β , and γ . Such a result would allow understanding of genetic recombination can evolve in the presence of mutation under certain circumstances (Altenberg, 1984; Kondrashov, 1988). It is clear that for certain cases of \mathbf{A} , \mathbf{B} , \mathbf{C} , and \mathbf{W} , the spectral radius is not monotonically decreasing in each of μ , β , and γ . However, specifying the conditions that produce an increase in the spectral radius with respect to μ , β , γ , etc. requires an extension of Karlin’s theorem. The existence of cases of increase led to the ‘‘Principle of Partial Control’’ for the evolution of genetic modifiers:

Conjecture 5. (Altenberg, 1984, p. 149) *When a modifier gene has only partial control over the transformations occurring at selected loci, then it is possible for this part of the transformation to evolve an increase.*

3.3 Extending Karlin's Theorem

We need to extend Karlin's theorem on linear variation from products of the form

$$[(1 - \mu)\mathbf{I} + \mu\mathbf{B}]\mathbf{W}$$

to products of the more general form

$$[(1 - \mu)\mathbf{A} + \mu\mathbf{B}]\mathbf{W}.$$

Open Question 3.1. *Let*

$$\mathbf{M}(\mu) = [(1 - \mu)\mathbf{A} + \mu\mathbf{B}],$$

where \mathbf{A} and \mathbf{B} are irreducible Markov matrices, and \mathbf{W} is a diagonal matrix with strictly positive diagonal elements not similar to the identity matrix. For what conditions on \mathbf{A} , \mathbf{B} , and \mathbf{W} is the spectral radius $\rho(\mathbf{M}(\mu)\mathbf{W})$ strictly decreasing in μ :

$$\frac{d}{d\mu}\rho(\mathbf{M}(\mu)\mathbf{W}) < 0, \text{ for } 0 \leq \mu \leq 1?$$

Karlin proved that the spectral radius $\rho([(1 - \mu)\mathbf{I} + \mu\mathbf{B}]\mathbf{W})$ is decreasing in μ . Clearly each matrix pair $\{\mathbf{B}, \mathbf{W}\}$ determines a class of matrices \mathbf{A} for which the spectral radius $\rho([(1 - \mu)\mathbf{A} + \mu\mathbf{B}]\mathbf{W})$ is decreasing in μ . Explicit characterization of this class is not immediately obvious. However, one can follow Karlin's proof to produce a condition which would provide the answer if it could be solved.

Suppose that

$$\mathbf{M}(\mu, r) = [(1 - \mu)\mathbf{I} + \mu\mathbf{A}][(1 - r)\mathbf{I} + r\mathbf{B}],$$

where \mathbf{A} and \mathbf{B} are Markov matrices.

I retrace the analysis of Karlin (1982, pp. 195–196). Define

$$\phi(\mathbf{p}, \mu, r) = \sup_{\mathbf{x} > 0} \sum_i p_i \log \left(\frac{x_i}{[\mathbf{M}(\mu, r)\mathbf{x}]_i} \right). \quad (18)$$

Let $\mathbf{x}(\mu, r)$ be the vector for which the supremum is attained. The Donsker-Varadhan (1975) variational formula for the spectral radius gives:

$$\log \rho(\mathbf{M}(\mu, r)\mathbf{D}) = \sup_{\mathbf{p} > 0} [\langle \mathbf{p}, \log(\mathbf{D} \mathbf{1}) \rangle - \phi(\mathbf{p}, \mu, r)] \quad (19)$$

where $\mathbf{D} = \mathbf{diag}[w_i/\bar{w}]$, $\mathbf{1}$ is the vector of ones, $\log(\mathbf{v})$ stands for the vector of components $\log(v_i)$,

$$\log(\mathbf{D} \mathbf{1}) = [(\log w_i - \log \bar{w})],$$

and we set $\sum_i p_i = 1$. Let $\mathbf{p}(\mu, r)$ be the vector at which this supremum is attained.

Since both $\mathbf{x}(\mu, r)$ and $\mathbf{p}(\mu, r)$ as implicitly defined are unique critical points of $\phi(\mathbf{p}, \mu, r)$,

$$\frac{\partial \phi(\mathbf{p}, \mu, r)}{\partial \mathbf{x}(\mu, r)} = \frac{\partial \phi(\mathbf{p}, \mu, r)}{\partial \mathbf{p}(\mu, r)} = \mathbf{0}$$

for all i . Hence

$$\frac{\partial \rho}{\partial \mu} = -\rho \frac{\partial}{\partial \mu} \phi(\mathbf{p}, \mu, r)$$

with $\mathbf{p} = \mathbf{p}(\mu, r)$ fixed. Further evaluation paralleling Karlin (1982) yields the condition

$$\begin{aligned} \frac{\partial \rho}{\partial \mu} \leq 0 &\iff \\ 1 &\leq \sum_i p_i(\mu, r) \frac{[\mathbf{M}(0, r)\mathbf{x}(\mu, r)]_i}{[\mathbf{M}(\mu, r)\mathbf{x}(\mu, r)]_i}. \end{aligned} \quad (20)$$

For $r = 0$, Karlin uses Jensen's inequality to give:

$$\begin{aligned} \log \sum_i p_i(\mu, 0) \frac{x_i(\mu, 0)}{[\mathbf{M}(\mu, 0)\mathbf{x}(\mu, 0)]_i} \\ \geq \phi(\mathbf{p}, \mu, 0) = \sum_i p_i(\mu, 0) \log \frac{x_i(\mu, 0)}{[\mathbf{M}(\mu, 0)\mathbf{x}(\mu, 0)]_i}. \end{aligned} \quad (21)$$

By using the principal eigenvector

$$\tilde{\mathbf{x}} = \mathbf{M}(\mu, 0) \tilde{\mathbf{x}},$$

the supremum definition of ϕ gives:

$$\phi(\mathbf{p}, \mu, 0) \geq \sum_i p_i(\mu, 0) \log \frac{\tilde{x}_i}{[\mathbf{M}(\mu, 0)\tilde{\mathbf{x}}]_i} = 0.$$

Thus was it is proved that for $r = 0$, $\partial \rho / \partial \mu \leq 0$. The analysis of (20) for $r > 0$ does not allow us to use (21), and is unsolved. This leaves us with:

Open Question 3.2. *What conditions on the matrices \mathbf{A} and \mathbf{B} , and scalars r and μ , produce*

$$1 \leq \sum_i p_i(\mu, r) \frac{[(1-r)\mathbf{I} + r\mathbf{B}]\mathbf{x}(\mu, r)_i}{[(1-\mu)\mathbf{I} + \mu\mathbf{A}][(1-r)\mathbf{I} + r\mathbf{B}]\mathbf{x}(\mu, r)_i},$$

where $\mathbf{x}(\mu, r)$ and $\mathbf{p}(\mu, r)$ are the vectors producing the suprema of expressions (18) and (19)?

Another direction to extend Karlin's theorem, which would be quite relevant to the issue of rapidly mixing Markov chains and rapid first hitting times, is to say something about the second-largest eigenvalue. I would offer (without claiming undue certitude) the following:

Conjecture 6.

Let

$$\mathbf{M}(\xi) = (1 - \xi)\mathbf{I} + \xi\mathbf{P},$$

where \mathbf{P} is an irreducible Markov matrix, and let \mathbf{W} be a diagonal matrix with strictly positive diagonal elements not similar to the identity matrix. Then the ratio of the second-largest eigenvalue $\lambda_2(\mathbf{M}(\xi)\mathbf{W})$, to the spectral radius $\rho(\mathbf{M}(\xi)\mathbf{W})$, is strictly increasing in ξ :

$$\frac{d}{d\xi} \frac{\lambda_2(\mathbf{M}(\xi)\mathbf{W})}{\rho(\mathbf{M}(\xi)\mathbf{W})} > 0, \text{ for } 0 \leq \xi \leq 1.$$

3.4 Discussion

Karlin's theorem, because it holds for arbitrary Markov and fitness matrices, captures a fundamental property of Darwinian dynamics, the interaction of selection and transformation caused by genetic operators. What is not generally understood is how multiple genetic operators interact with one another. Theories that depend on the interaction of recombination, mutation, migration, selection, and drift, such as Wright's Shifting Balance Theory (Wright, 1931), pose formidable analytical difficulties. Attempting to understand the interaction of multiple genetic operators brings us to the need to extend Karlin's theorem.

4 Conclusion

I hope that the reader, having followed the lines of discussion through this chapter, may come away with the conclusion that the spectra of evolutionary systems provide a useful means to pose, and occasionally to solve, problems in evolutionary dynamics. I have used the spectral representation of the generalized mutation-selection system to address the question of when an evolutionary algorithm is useful for function optimization. I have described an analog to "rapidly mixing Markov chains" (Sinclair, 1992) that is appropriate for optimization, "rapid first hitting time". The conditions needed for an evolutionary algorithm to exhibit rapid first hitting time can be described in terms of the spectra of the linear systems that, under broad circumstances, can be used to represent them.

I have also posed questions on the dynamics of finite populations in terms of the spectra of the underlying operators. Tying together the spectra of infinite population models with the spectra of the finite population models into which they are embedded remains a major open question in the theory of evolutionary dynamics. Progress may result if flows over the low-dimensional boundaries of the simplex can be modeled.

Lastly, I have reviewed an important theorem by Karlin (1982) on the spectral properties of genetic operator intensity. Extensions of this theorem would find immediate application.

Since these are spectral problems, there may indeed already be analytic techniques that could be applied to their solution. It is hoped that this chapter may bring attention to these problems and thus hasten their solution.

5 Acknowledgements

I thank Anil Menon for initiating this book as an outgrowth of the "Hilbert Challenges" for Evolutionary Algorithms session organized for the 2001 International Conference on Artificial Intelligence. The open questions on Karlin's Theorem emerged during work on my doctoral dissertation with Marcus W. Feldman and Samuel Karlin in 1983. Early versions of my ideas on rapid first hitting times were vetted in discussions with Richard Palmer and other colleagues while I was a visiting researcher at the Santa Fe Institute in the summer of 1995.

Bibliography

Ackley, D. H., 1987, *A Connectionist Machine for Genetic Hillclimbing*, volume SECS28 of *The Kluwer International Series in Engineering and Computer Science*. (Kluwer Academic Publishers, Boston).

Altenberg, L., 1984, *A Generalization of Theory on the Evolution of Modifier Genes*, Ph.D. thesis, Stanford University, searchable online and available from University Microfilms, Ann Arbor, MI.

Altenberg, L., 1995, in *Foundations of Genetic Algorithms 3*, edited by D. Whitley and M. D. Vose (Morgan Kaufmann, San Mateo, CA), pp. 23–49.

Altenberg, L., and M. W. Feldman, 1987, *Genetics* **117**, 559.

Arora, S., Y. Rabani, and U. Vazirani, 1994, in *Proceedings of the 26th Annual ACM Symposium on Theory of Computing*, pp. 459–467, URL citeseer.nj.nec.com/arora94simulating.html.

Bäck, T., 1996, *Evolutionary Algorithms in Theory and Practice: Evolutionary Strategies, Evolutionary Programming, Genetic Algorithms* (Oxford University Press, Oxford).

Cannings, C., 1974, *Advances in Applied Probability* **6**, 260.

Christiansen, F. B., 2000, *Population Genetics of Multiple Loci* (John Wiley and Sons, LTD, Chichester).

Davis, T. E., and J. C. Principe, 1993, *Evolutionary Computation* **1**(3), 269.

Donsker, M. D., and S. R. S. Varadhan, 1975, *Proceedings of the National Academy of Sciences USA* **72**, 780.

Ewens, W. J., 1979, *Mathematical Population Genetics* (Springer-Verlag, Berlin).

Feldman, M., F. Christiansen, and L. D. Brooks, 1980, *Proceedings of the National Academy of Sciences U.S.A.* **77**, 4838.

- Feldman, M. W., 1972, *Theoretical Population Biology* **3**, 324.
- Feldman, M. W., and B. Balkau, 1973, *Genetics* **74**, 713.
- Feller, W., 1951, in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, edited by J. Neyman (University of California Press, Berkeley), pp. 227–246.
- Fisher, R. A., 1930, *The Genetical Theory of Natural Selection* (Clarendon Press, Oxford).
- Gantmacher, F. R., 1959, *The Theory of Matrices*, volume 2 (Chelsea Publishing Company, New York).
- Goldberg, D. E., and K. Deb, 1991, in *Foundations of Genetic Algorithms*, edited by G. Rawlins (Morgan Kaufmann, San Mateo, CA), pp. 69–93.
- Goldberg, D. E., and P. Segrest, 1987, in *Proceedings of the Second International Conference on Genetic Algorithms*, pp. 1–8.
- Horn, J., D. E. Goldberg, and K. Deb, 1994, in *Parallel Problem Solving from Nature—PPSN III*, edited by H. P. Schwefel and R. Männer (Springer-Verlag, Berlin), volume 866.
- Karlin, S., 1982, in *Evolutionary Biology*, edited by M. K. Hecht, B. Wallace, and G. T. Prance (Plenum Publishing Corporation), volume 14, pp. 61–204.
- Karlin, S., and J. McGregor, 1974, *Theoretical Population Biology* **5**, 59.
- Kondrashov, A. S., 1988, *Nature (London)* **336**(6198), 435.
- Liepins, G., and M. Vose, 1990, *Journal of Experimental and Theoretical Artificial Intelligence* **2**(2), 101.
- Nix, A. E., and M. D. Vose, 1991, *Annals of Mathematics and Artificial Intelligence* **5**, 79.
- Palmer, R. G., 1982, *Advances in Physics* **31**, 669.
- Rabani, Y., Y. Rabinovich, and A. Sinclair, 1995, in *Annual ACM Symposium on the Theory of Computing*, pp. 83–92, ISBN 0-89791-718-9, URL citeseer.nj.nec.com/41563.html.
- Rabinovich, Y., A. Sinclair, and A. Wigderson, 1992, in *IEEE Symposium on Foundations of Computer Science*, pp. 304–313, URL citeseer.nj.nec.com/article/rabinovich92quadratic.html.
- Rabinovich, Y., and A. Wigderson, 1999, *Random Structures Algorithms* **14**, 111.
- Rudolph, G., 1997, *Convergence properties of evolutionary algorithms* (Verlag Kovač, Hamburg).

Schmitt, F., and F. Rothlauf, 2001a, in *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001)*, edited by L. Spector, E. D. Goodman, A. Wu, W. B. Langdon, H.-M. Voigt, M. Gen, S. Sen, M. Dorigo, S. Pezeshk, M. H. Garzon, and E. Burke (Morgan Kaufmann, San Francisco, California, USA), pp. 559–564, ISBN 1-55860-774-9, URL citeseer.nj.nec.com/schmitt01importance.html.

Schmitt, F., and F. Rothlauf, 2001b, *On the Mean of the Second Largest Eigenvalue on the Convergence Rate of Genetic Algorithms*, Technical Report Working Paper 1/2001, University of Bayreuth, Department of Information Systems, Universitaetsstrasse 30, D-95440 Bayreuth, Germany, working Papers in Information Systems.

Schwefel, H.-P., 1987, Preprints of the 31st Annual Meeting of the International Society for General System Research, Budapest **2**, 1025.

Sinclair, A., 1992, *Algorithms for Random Generation and Counting: A Markov Chain Approach* (Birkhäuser, Boston), ISBN 0-8176-3658-7.

Suzuki, J., 1995, IEE Transactions on Systems, Man and Cybernetics **25**(4), 655.

Teague, R., 1976, Journal of Theoretical Biology **59**, 25.

Teague, R., 1977, Theoretical Population Biology **12**, 86.

van Nimwegen, E., 1999, *The Statistical Dynamics of Epochal Evolution*, Ph.D. thesis, Universiteit Utrecht, Amsterdam.

van Nimwegen, E. J., J. P. Crutchfield, and M. Huynen, 1999, Bulletin of Mathematical Biology **62**, 799.

Vitanyi, P., 2000, Theoretical Computer Science **241**(1–2), 3, ISSN 0304-3975, URL <http://xxx.lanl.gov/abs/cs.NE/9902006>.

Wolpert, D. H., and W. G. Macready, 1995, *No Free Lunch Theorems for Search*, Technical Report SFI-TR-95-02-010, Santa Fe Institute, Santa Fe, NM, URL citeseer.nj.nec.com/wolpert95no.html.

Wolpert, D. H., and W. G. Macready, 1997, IEEE Transactions on Evolutionary Computation **1**(1), 67, URL citeseer.nj.nec.com/wolpert96no.html.

Wright, S., 1931, Genetics **16**, 97.