Spectral Analysis of Evolutionary Dynamics

Lee Altenberg

Information and Computer Sciences

University of Hawai‘i at Manoa

altenber@hawaii.edu

http://dynamics.org/Altenberg/PAPERS/OPSAED/

Challenges in the analysis of evolutionary dynamics

1. General theory for performance and design of evolutionary algorithms has proven difficult to achieve.
2. Difficulty sets in with the simplest “canonical” models of evolutionary algorithms owing to
   (a) their nonlinear structure and
   (b) stochastic dynamics.

Simplifying assumptions to

eliminate nonlinearity and stochasticity:

1. **Linearity** is produced by assuming **constant selection** and **uniparental transmission**
2. **Determinism** can be produced by assuming an **infinite population size**.
3. → Linear dynamical system whose trajectory and attractors
   (a) can be described in closed form, and
   (b) decomposed into **spectrum of eigenvalues and eigenvectors**.

More realistic models:

Actual evolutionary algorithms depart from this boundary in two important ways:

1. **recombination** between two (or more) parents.
2. **finite populations**.

Departure 1: Recombination

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

1. 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.
2. A recent compendium can be found in Christiansen:2000.
3. For more on quadratic dynamical systems see Rabinovich:Sinclair:and:Wigderson:1992 and Arora:Rabani:and:Vazirani:1994.
4. Progress has been made in the dynamics of recombination in the absence of selection, in both infinite and finite population models, by Rabani:Rabinovich:and:Sinclair:1995, and for simple selection, by Rabinovich:and:Wigderson:1999.
5. Numerous analyses for other models on the boundary of the general problem can be found in the evolutionary computation and population genetics literature.

Departure 2: Finite populations

1. 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.
2. Finite population algorithms typically use Bernoulli sampling to generate new samples of the search space.
   (a) This changes the model of the algorithm from deterministic to stochastic,
   (b) a Markov chain which has a linear state transition matrix, but
   (c) whose dimensions are exponentially increased beyond the number of elements in the search space.

1. The first model of finite population dynamics was developed based on Bernoulli sampling by Wright:1931 and Fisher:1930.
2. 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,
   (a) where $|\mathcal{S}|$ is the number of different genotypes, and N is the population size.
3. Hence, the dimensionality of the state space is vastly increased in the finite population model over the infinite population model.

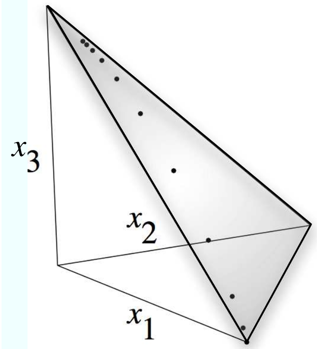
Illustrate using the frequency simplex

The difference between infinite and finite population models can be illustrated showing points in the $|\mathcal{S}| - 1$ dimensional simplex.

- $|\mathcal{S}| = 3$ alleles,
- Allele frequencies $x_1 + x_2 + x_3 = 1$.

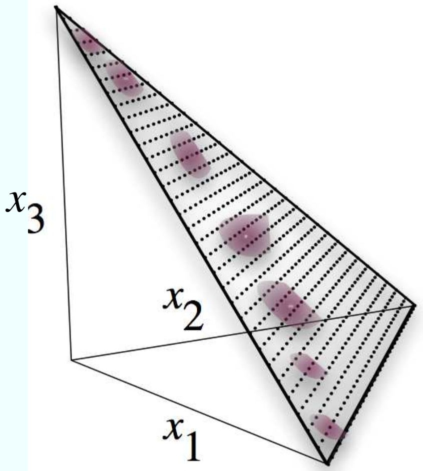
Infinite population:

System state is a single point in the simplex which moves deterministically from one generation to the next.



Finite population:

- System state is a **probability distribution over a cloud points** in the simplex, restricted to the lattice of coordinates $\{\mathbf{x} : N\,x_i \in \{0, 1, \ldots, N\}, \sum_{i=1}^{n} x_i = 1\}$.
- The distribution of the cloud of points is what changes every generation.



Strategy for analysis:

1. Characterize features of the dynamics that count as “good performance”
   (a) Global attraction
   (b) Rapid first hitting time
2. Solve the infinite population model.
3. Find properties of the infinite population model that produce good performance.
4. See what we can then say about the finite population model.

Infinite Population Models

The Canonical Model

The ‘canonical’ model:

1. an infinite population — the state space is frequencies of types i
2. discrete, non-overlapping generations,
3. constant fitness coefficients, w_i
4. generalized single-parent transmission, $\mathbf{T}$.

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.

The recursion on $\mathbf{x}$

$$x_i{}' = \sum_{j=1}^{n} T(i \leftarrow j)\, w_j\, x_j / \overline{w}, \quad \text{or, in vector form:} \quad \mathbf{x}' = \frac{1}{\overline{w}} \mathbf{T}\mathbf{W}\mathbf{x},$$

where

- $\mathbf{x}'$ is the vector of frequencies in the next time step;
- $\mathbf{W}$ is the diagonal matrix of fitness coefficients, $w_i \geq 0$;
- $\overline{w} = \sum_{i=1}^{n} w_i x_i$ is the mean fitness of the population, used as a normalizer to maintain the system state as frequencies; and
- $\mathbf{T} = \left[T_{ij}\right]_{i,j=1}^{n}$ is the n-by-n matrix of transmission probabilities, T_{ij}, the probability that type j produces an offspring of type i, so
$$\sum_{i=1}^{n} T_{ij} = 1 \; \forall j, \;\; T_{ij} \geq 0.$$

Compare: the bi-parental (sexual reproduction) model:

$$x_i{}' = \sum_{j,k=1}^{n} T(i \leftarrow j,k)\, \frac{w_j\, x_j\, w_k\, x_k}{\overline{w}},$$

or in vector form:

$$\overline{w}\, \mathbf{x}' = \mathbf{T}(\mathbf{W}\ \otimes\ \mathbf{W})(\mathbf{x} \otimes \mathbf{x}).$$

Thus, it is a quadratic system— $\mathbf{x}(t)$ not tractable in general.

$\otimes$ is the tensor product (see below);

w_{ij} is the fitness of the parental pair (i, j)

$\mathbf{W}$ is the n^2 by n^2 diagonal matrix of coefficients $w_{i_1 i_2}$;

$T(i \leftarrow j, k)$ = probability that genetic operators produce offspring i from parents j and k, $\sum_{i=1}^{n} T(i \leftarrow j,k) = 1$; $\mathbf{T}$ in this case is the n by n^2 matrix of transmission probabilities.

The trajectory of the system is:

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

where $\nu(t) = \mathbf{1}^\top (\mathbf{TW})^t\, \mathbf{x}(0)$ is the normalizer to maintain the state vector as frequencies.

Optimal Evolutionary Dynamics for Optimization

- For an optimization problem, we assume that an objective function $f : \mathcal{S} \rightarrow \Re+$ is defined on each element of the search space;
- The goal is to find an element with maximum objective function value. Assume the element is unique.
- Exhaustive search or random search will require on the average $n/2$ samples to have sampled the optimum.
- If an algorithm can find the optimum in an average of $\epsilon n/2$ samples, for some small constant $\epsilon \ll 1$, it is clearly doing better than "blind search".
- However, evolutionary algorithms can perform much better than $\mathcal{O}(n)$, namely $\mathcal{O}(\log(n))$.

As good as it gets: the ONEMAX problem

- The canonical example for an "evolutionary algorithm-easy" problem is the ONEMAX problem, where the objective function 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
  1. L the number of loci,
  2. $n = |\mathcal{A}|^L$ is the size of the search space,
  3. $\mathcal{A}$ is the set of alleles for each locus,
  4. $|\mathcal{A}|$ the cardinality of $\mathcal{A}$ (for binary strings, $|\mathcal{A}| = 2$) .

Time complexity goal:

- 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 $\log(n)$.

What conditions on an evolutionary algorithm will allow it to find the global optimum in $\mathcal{O}(P(\log(n)))$ samples?

Independent of initial samples:

- Evolutionary algorithms often have multiple domains of attraction (at least in the metastable sense vanNimwegen:and:Crutchfield:2000), 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 objective function,
- but the objective function by itself does not determine whether the EA has multiple domains of attraction—
- it depends on the relationship of the objective function to the genetic operators Altenberg:1995:STPT.

To preclude this secondary search problem, the algorithm should exhibit a single, global attractor that contains the global optimum.

Summary of Performance Goals for an EA:

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*.

Spectral Analysis

- Can we define properties of $\mathbf{TW}$ that correspond to the performance goals?
- Yes. The **spectrum of eigenvalues and eigenvectors** of $\mathbf{TW}$.

Spectral Conditions for Global Attraction

For the canonical model Eq. (16), the *global attraction* condition, 2 above, can be stated precisely: There exists one attractor, the frequency vector $\mathbf{x}^*$ such that:

1. $$\lim_{t\to\infty} \frac{1}{\nu(t)} (\mathbf{TW})^t \, \mathbf{x}(0) = \mathbf{x}^*, \forall \, \mathbf{x}(0) \in \Delta_n,$$

2. The frequency of the global optimum type at the attractor is positive, i.e. $x_1^* > 0$, where we index the global optimum type as 1. Thus x_1^* is its asymptotic frequency.

Conditions for Global Attraction

- Condition (1) is guaranteed if and only if $\mathbf{T}$ is primitive (irreducible and acyclic), i.e. there is some $k \geq 0$ such that $\mathbf{T}^k > 0$.
- Primitiveness in the transmission matrix corresponds to the property of ergodicity.
- From the Perron-Frobenius theorem Gantmacher:1959, primitiveness guarantees that there be a strictly positive eigenvector $\mathbf{x}^*$ corresponding to the leading eigenvalue of $\mathbf{TW}$.

$$\mathbf{TW}\mathbf{x}^* = \lambda_1 \mathbf{x}^*.$$

- This eigenvector $\mathbf{x}^*$, normalized so $\langle \mathbf{1}, \mathbf{x}^* \rangle = \sum_i x_i^* = 1$, is the global attractor. This is true since $\lambda_1 > \lambda_i \; \forall i \neq 1$, so the composition of the population converges to it regardless of the initial composition $\mathbf{x}(0)$.

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:

Open Question:

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

Assume we are given $\mathbf{W}$ as a fixed part of the problem to be solved. What transmission matrices $\mathbf{T}$ will give rapid first hitting time?

Let the unique optimum in the search space be labeled with index $i = 1$. Thus

$$w_1 = \max_{i=1}^{n} w_i > w_i \; \forall i \neq 1.$$

Trivial answer:

Guarantee a hitting time of 1 by simply constructing a transmission matrix that produces the optimum by mutation:

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

Problems with the trivial answer:

- Transmission in this case is biased to find the optimum without any help from selection.
- Clearly, such *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:STPT.
- 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.

“Fair” transmission:

The translation of these principles into a condition on $\mathbf{T}$ would require that all types evolve to equal frequency in the absence of selection, i.e.

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

Condition (30) 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:

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?

Spectral Analysis of the Convergence Rate

Assume $\mathbf{T}$ is symmetric. Then $\mathbf{TW}$ can be represented in Jordan canonical form as:

$$\mathbf{TW} = \mathbf{Q} \boldsymbol{\Lambda} \mathbf{Q}^\top,$$

- 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
- $\boldsymbol{\Lambda}$ is the diagonal matrix $\boldsymbol{\Lambda}_{ii} = \lambda_i$ of the eigenvalues of $\mathbf{TW}$.

The trajectory of the population is:

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

Details:

The assumption that transition probabilities are symmetric, i.e. $T_{ij} = T_{ji}$, is typical of the mutation operators used on data structures in evolutionary computation. Then, since any symmetric matrix $\mathbf{S}$ has Jordan form $\mathbf{S} = \mathbf{P\Lambda P}^\top$, so we can take $\mathbf{S} = \mathbf{W}^{1/2}\,\mathbf{TW}^{1/2} = \mathbf{P\Lambda P}^\top$, 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, $w_i > 0$.

Some bookkeeping:

We can arbitrarily permute the indices so that

- $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n > -\lambda_1$,
- $w_1 > w_2 \geq \cdots \geq w_n$.
- Then for Q_{ij}, i follows the order of the fitnesses, while j follows the order of the eigenvalues.
- Define the columns of $\mathbf{Q}$ so that $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n]$.
- $\mathbf{TW}\mathbf{q}_1 = \lambda_1\,\mathbf{q}_1$, hence $\mathbf{q}_1$ is the strictly positive leading eigenvector of $\mathbf{TW}$. $\mathbf{q}_1 = c\,\mathbf{x}^*$, where $c = \langle \mathbf{1}, \mathbf{q}_1\rangle$. By definition $\langle \mathbf{q}_i, \mathbf{q}_i\rangle = 1$.

Trajectory of the frequency of the optimum type:

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

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

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

using $c = \langle\mathbf{1}, \mathbf{q}_1\rangle$.

Putting them together, we obtain:

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

As $t \rightarrow \infty$, the terms $\left(\frac{\lambda_i}{\lambda_1}\right)^t$ damp out, leaving $x_1(\infty) = x_1^*$.

How fast these terms damp out is bounded by

$$\frac{\lambda_2}{\lambda_1} \geq \frac{\lambda_i}{\lambda_1}, \ \ i > 2.$$

- We are thus quite interested in λ_2/λ_1.
- This brings us to the topic of **rapidly mixing Markov chains**.

Rapidly Mixing Markov Chains

- 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.

Sinclair's Concept

Sinclair:1992 developed the concept of *rapid mixing* in a Markov chain:

- The relative pointwise distance (r.p.d.) on a Markov process with transition matrix $\mathbf{P}$ is:

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

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:

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

Rapid Mixing and Rapid First Hitting Times

Vitanyi:2000:ADoEP 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 $\mathbf{x}*$ 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.

Fisher and Wright were here

- 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 the population size.
- λ_3 gives the rate of convergence to fixation on a single type due to genetic drift, and is also the rate of decrease in the frequency of heterozygotes in the population. See Ewens:1979:MPG.
- More recent work on the second-largest eigenvalue includes Suzuki:1995, Rudolph:1997:CPoEA, and Schmitt:and:Rothlauf:2001, Schmitt:and:Rothlauf:2001

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 $n = |\mathcal{S}|$, the size of the search space, since $\mathcal{O}(P(\log(n))) * \mathcal{O}(P(\log(n))) = \mathcal{O}(P(\log(n)))$.

An infinite-population version of 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.
- Can we construct an analogous condition for the infinite population model?

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 $i \in \{2, \ldots, n\}$. Let

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

The infinite population dynamics will be 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

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

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:

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

Caveats (BIG)!

- The first hitting time is a concept that properly belongs to stochastic processes, not deterministic models; it is a random variable.
- Precedent: “Takeover time” models Goldberg:and:Deb:1991 also use a deterministic, infinite population model 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.
- 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.

...that being said:

Substituting the above into (45), setting $t = P_2(\log(n))$, and rearranging, we obtain the condition:

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

Since $\mathbf{q}_1 = c\,\mathbf{x}^*$, we substitute $P_1(\log(n))\,q_{11} - c = c\,[P_1(\log(n))x_1^* - 1]$, and $\langle \mathbf{q}_1, \mathbf{x}(0)\rangle = c\langle \mathbf{x}^*, \mathbf{x}(0)\rangle$, to get:

$$c^2\,[P_1(\log(n))x_1^* - 1]\langle \mathbf{x}^*, \mathbf{x}(0)\rangle \geq \sum_{i=2}^{n}\left(\frac{\lambda_i}{\lambda_1}\right)^{P_2(\log(n))}(\langle \mathbf{1}, \mathbf{q}_i\rangle - P_1(\log(n))q_{1i})\,\langle \mathbf{q}_i, \mathbf{x}(0)\rangle,$$

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

Second-largest eigenvalue λ_2

Define the ratio of the second largest eigenvalue to the largest eigenvalue as:

$$r = \lambda_2/\lambda_1.$$

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

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

In this case, condition (2) is met provided

$$[P_1(\log(n))x_1^* - 1]\langle \mathbf{x}^*, \mathbf{x}(0)\rangle \geq \delta/c^2$$

or

$$x_1^* \geq \frac{1 + \frac{\delta}{c^2\,\langle \mathbf{x}^*, \mathbf{x}(0)\rangle}}{P_1(\log(n))} > \frac{1}{P_1(\log(n))}. \tag{2}$$

The punch line:

- 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 $x_1^* = \mathcal{O}(P_1(\log(n))^{-1})$.
- We know that selection is required in order for $x_1^* \geq \frac{1}{P_1(\log(n))}$ since the principle eigenvector of $\mathbf{T}$ has $x_1^* = \frac{1}{n}$ 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.

To make a long story short:

- Rapid first hitting times require that
  1. the second largest eigenvalue of $\mathbf{TW}$ be small enough, and
  2. the selection strength $\mathbf{W}$ be large enough.

The main question:

What kind of transmission matrices $\mathbf{T}$ minimize $r = \lambda_2/\lambda_1$?

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 (2). 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 $\lambda_2(\mathbf{U}) = \cdots = \lambda_n(\mathbf{U}) = 0$.

When we include selection:

$$\mathbf{UW} = \frac{1}{n}\mathbf{1}\,\mathbf{1}^\top\mathbf{W} = \frac{1}{n}\mathbf{1}\,[w_1\; w_2 \cdots w_n] = \frac{1}{n}\begin{bmatrix} w_1 & w_2 & \cdots & w_n \\ w_1 & w_2 & \cdots & w_n \\ \vdots & & & \vdots \\ w_1 & w_2 & \cdots & w_n \end{bmatrix}$$

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

- It would appear that the rank-1 matrix would be a candidate transmission matrix to achieve rapid first hitting times.
- This hope is instantly dashed by noting that for $\mathbf{UW}$, $x_1^* = 1/n$, which is not greater than $1/P_1(\log(n))$.
- We might ask if we can find another rank-1 matrix where $x_1^* \geq 1/P_1(\log(n))$, but this is precluded by the condition that $\mathbf{T}$ be ‘fair’, and thus doubly stochastic, requiring that $x_i^* = 1/n$ for all i.
- This result is expected, when we consider that the rank-1 matrix corresponds to random search.

So, we are left with the following:

Open Question:

For a given set of fitnesses, $\mathbf{W}$, what classes of fair transmission matrices maximize x_1^* while minimizing $r = \lambda_2/\lambda_1$ so as to satisfy the conditions for rapid first hitting time?

The ONEMAX problem, reprise

- The ONEMAX problem as the paradigmatic EA-easy problem, exhibiting rapid first hitting time.
- What transmission matrix does the ONEMAX problem have? Simple **bit-flip mutation**.
- Bit-flip mutation 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), any problem becomes a ONEMAX problem.
- Hence, one can conjecture that a transmission matrix represented by the **binary hypercube** would be a primary candidate for rapid first hitting time.

Generalizing:

- The hypercube graph may be generalized to any graph that has diameter $L \approx \mathcal{O}(\log(n))$.
- This includes "small world" graphs, where random edges have been added to a lattice.
  - But not all "small world" graphs are **navigable** Kleinberg:2000.
  - Non-navigable small world graphs may not allow rapid first hitting time.
  - ... an open question.

Guaranteed Slow First Hitting Time

- What kind of transmission matrices can never achieve rapid first hitting time for any set of fitnesses?
- Clearly, any whose graph has diameter $L \approx \mathcal{O}(n^d)$. These include:
  - Lattices, where the diameter is $\mathcal{O}(n^d)$, where d is the dimension of the lattice;
  - "Long path" Horn:Goldberg:and:Deb:1994 matrices (1-lattices):

The "Long path" Transmission Matrix

A 1-Lattice:

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:*

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

Conjecture:

Then, there are no fitnesses $\mathbf{W}$, nor values μ, such that the system with long-path mutation, $\mathbf{P}$,

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

has rapid first hitting time.

Rapid First Hitting Time and No Free Lunch Theorems

- 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,Wolpert:and:Macready: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.

Different Lunches

- A **long-path operator** and a **a binary hypercube operator** will have **the same average performance** in locating the global optimum over all permutations of fitnesses (NFL).
- But they can be distinguished by their **potential** for rapid first hitting time:
  - The binary hypercube makes possible permutations of fitnesses that produce ONEMAX problems having a rapid first hitting time, with an adequate distribution of fitnesses.
  - The long-path operator 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.

Finite Populations

- A key open question in evolutionary computation:

  the relationship between the dynamics of the **infinite** and the **finite** population models.
- The Wright-Fisher model of finite populations Wright:1931,Fisher:1930 is derived from the canonical model of an infinite population by the addition of only one free parameter
  - the population size.

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;

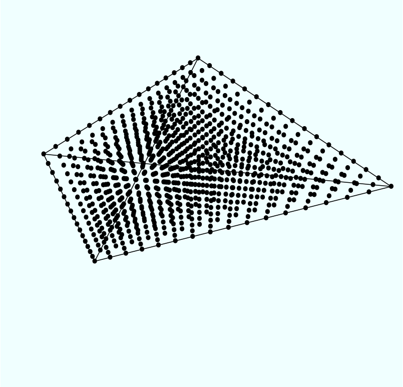
$\mathbf{x}$ be the vector of frequencies of each type i in the population, corresponding to $N\,x_i$ individuals of type i;

$\mathbf{x}'$ 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 $\mathbf{y}(\mathbf{x})$;

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

Discrete frequencies

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



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

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

with entries

$$\begin{aligned} M_{\mathbf{x}',\mathbf{x}} &= N!\prod_{i=1}^n \frac{y_i^{Nx_i'}}{(Nx_i')!} \\ &= \frac{N!}{\prod_{i=1}^n (Nx_i')!}\prod_{i=1}^n \left(\frac{\mathbf{e}_i^\top \mathbf{TWx}}{\mathbf{1}^\top \mathbf{Wx}}\right)^{Nx_i'} \end{aligned}$$

where $\mathbf{e}_i^\top = [0\;0\;\cdots 1 \cdots 0\;0]$ has the 1 in the ith position.

Jordan Form

Incorporating the Jordan canonical form from before (32):

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

Finite-Infinite Relationship

Open Question:

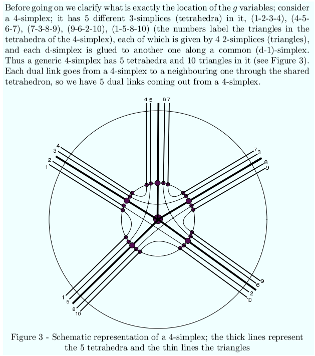
What is the relationship between the eigenvalues and eigenvectors of $\mathbf{TW}$ *and those of* $\mathbf{M}$*?*

Finite Dynamics

1. Situation of interest: $N \approx \mathcal{O}(\log(n)) \ll n$,
2. Thus the vast majority of the entries of any $\mathbf{x} \in \Delta_n(N)$ must be 0.
3. 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.
4. The trajectory of points in the finite population model will be radically different from the trajectory in the infinite population model.
5. In the infinite population model, the system will immediately enter the interior of Δ_m since $\mathbf{TW} > 0$.
6. In the finite population model, a probability distribution will move over the surface of Δ_n.

To illustrate

- How do we visualize a large n-simplex?
- Here is the 4-simplex (from Oriti:and:Williams:2001) :

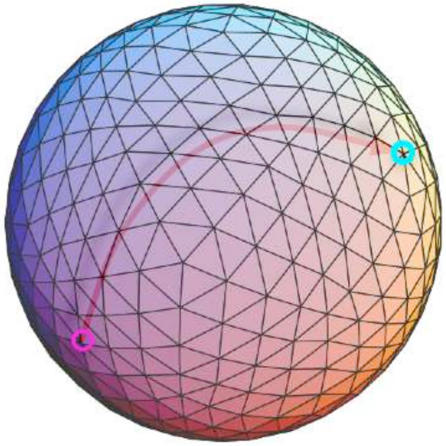


Before going on we clarify what is exactly the location of the g variables; consider a 4-simplex; it has 5 different 3-simplices (tetrahedra) in it, (1-2-3-4), (4-5-6-7), (7-3-8-9), (9-6-2-10), (1-5-8-10) (the numbers label the triangles in the tetrahedra of the 4-simplex), each of which is given by 4 2-simplices (triangles), and each d-simplex is glued to another one along a common (d-1)-simplex. Thus a generic 4-simplex has 5 tetrahedra and 10 triangles in it (see Figure 3). Each dual link goes from a 4-simplex to a neighbouring one through the shared tetrahedron, so we have 5 dual links coming out from a 4-simplex.

Figure 3 - Schematic representation of a 4-simplex; the thick lines represent the 5 tetrahedra and the thin lines the triangles

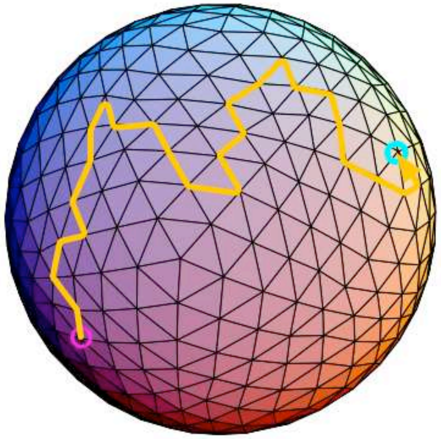
... not very promising.

INFINITE POPULATION (N >> n):



TRAJECTORY GOES THROUGH THE *INTERIOR* OF THE *n*-SIMPLEX (figurative representation)

FINITE POPULATION (N << n):



TRAJECTORY GOES OVER *O*(*N*)-DIMENSIONAL *BOUNDARIES* OF THE *n*-SIMPLEX

- Evolution in the finite population model can be views as

  transitions between one k-dimensional ($k \leq N$) boundary simplex 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.
- When the relaxation time within a low-dimensional boundary simplex is faster than the transition time between simplices, van Nimwegen and Crutchfield call this **metastability**.
- Long transition times between boundaries is known as **broken ergodicity** Palmer:1982:BE.

- Ergodicity in the infinite population model is necessary for ergodicity in the finite population model, but it is not sufficient. The Markov chain for the finite population model must in addition be *rapidly mixing* to avoid broken ergodicity.
- Methods which can analyze (69) as a flow along the low-dimensional boundaries of the simplex may prove to be most helpful in understanding finite population dynamics.
- van Nimwegen and Crutchfield (1999) apply this approach to specific models of mutation and selection, and have made good inroads to understanding the process in these specific cases.
- Answers to the general finite population dynamics problem remain **an open question**.