An Introduction to GA Theory

Jonathan E. Rowe
School of Computer Science
University of Birmingham

There are several different kinds of research that all get labelled as “theory”.

- Mathematics — describing what is true
- Science — describing what is observed
- Engineering — designing new things

The mathematical theory of GAs can be used to provide a firm foundation for the models of the scientific approach and the intuitions of the engineers.The approach we will take is:

- take a vague intuitive statement or idea
- try to formalize it mathematically
- try to prove something about the formal concept
- relate the result back to the practical case

There are many intuitive concepts used in designing GAs and describing their behaviour.

We are going to look at just one:

The idea of "convergence" in Genetic Algorithms.

"Genetic algorithms suffer from premature convergence."

Such statements refer to the idea that when a GA reaches some stopping criterion (e.g. average fitness is not increasing, the population starts to look uniform, a certain number of generations has passed), the end population does not contain the optimum.

But the notion of "convergence" in a GA is tricky. . . .

GAs are random processes, mapping populations to populations.

The probability of getting a particular population depends only on the previous generation.

This kind of random process is called a **Markov Chain**.

A Markov Chain is described by its **transition matrix**.

$Q_{i,j}$ is the probability of going from population j to population i in one generation.

Q is a large matrix!

A Markov Chain might have **absorbing states**. Once you arrive at such a state, you can't escape.

Example: a GA with selection and crossover (but no mutation). Absorbing states are the **uniform populations**.

"Convergence" might mean that the GA has arrived at an absorbing state.

Write the transition matrix in the form:

$$Q = \left[\begin{array}{c|c} I & R \\ \hline 0 & S \end{array}\right]$$

The expected time to absorption is given by the vector

$$a = (I - S^T)^{-1}\mathbf{1}$$

where a_k is the time to absorption starting from state k.

Some Markov Chains are **ergodic** and do not have absorbing states. They visit every possible state infinitely often!

Example: a GA with mutation. There is always a non-zero probability that anything could happen at the next generation.

Some states may be more likely to be visited than others. In fact there is a limiting distribution over all possible states, which the process will tend towards.

We can write down a formula for the limiting distribution, in terms of the transition matrix — but it is impractical to calculate it.

“Convergence” might mean that the behaviour of the GA is conforming to the limiting distribution.

Suppose we add elitism to our GA and watch what happens to the best item in the population. It's fitness surely increases to the optimum.

Be careful! In a random process, it could happen that the optimum is never seen! Fortunately the probability that this happens is zero.

We say that the maximum population fitness converges **almost surely** to the optimum fitness.

To make progress with the theory, we have to be able to write down equations describing the Markov process.

We can describe a population by a vector

$$p = (p_0, p_1, \ldots, p_{n-1})$$

where p_k is the proportion of the population occupied by item k.

Let

$$q = (q_0, q_1, \ldots, q_{n-1})$$

contain the probabilities that each item is generated in the next population.

The next population can be thought of as being N independent samples of the search space, using q as a probability distribution.

We can think of the action of a GA in terms of a map from vectors to vectors:

$$q = \mathcal{G}(p)$$

That is, $\mathcal{G} : \Lambda \to \Lambda$, where

$$\Lambda = \{x \in \mathbb{R}^n : x_k \geq 0, \sum x_k = 1\}$$

The transition matrix for the finite population case can be constructed from this map.

The probability that population q follows population p is

$$N! \prod \frac{(\mathcal{G}(p)_j)^{Nq_j}}{(Nq_j)!}$$

This is called a **multinomial** distribution.

To define $\mathcal{G}$ it is helpful to split it into a **selection** phase and a **mixing** phase.

$$\mathcal{G} = \mathcal{M} \circ \mathcal{F}$$

These operators can then be defined for particular selection and mixing (crossover and mutation) schemes.

The fitness-proportional selection scheme is defined by:

$$\mathcal{F}(p) = \frac{\text{diag}(f)p}{f^T p}$$

where

- f is the vector containing the fitness values
- diag(f) is the diagonal matrix with f on the diagonal.

Crossover may be defined by the application of a mask. A mask b has a probability χ_b of being used.

The mixing operator for this kind of crossover is:

$$\mathcal{M}(p)_k = \sum_b \chi_b \sum_{i,j} p_i p_j [(i \otimes b) \oplus (j \otimes \bar{b}) = k]$$

Iterating $\mathcal{G}$ will produce a sequence of points. This sequence might converge to a **fixed-point**.

What has this sequence got to do with finite populations?

- $\mathcal{G}(p)$ is the **expected** next population.
- It describes the **limiting** behaviour as the population size grow large.
- It qualitatively characterises the transient behaviour of finite populations.

Theorem Suppose we are given:

- an initial population p
- a number of generations t
- a small error $\epsilon > 0$
- a large probability $1 - \delta$

Let q be the actual population observed after t generations. Then there exists a number N such that if the population size is bigger than N, then the probability that

$$\|\mathcal{G}^t(p) - q\| < \epsilon$$

is greater than $1 - \delta$.

Populations which are close to fixed-points correspond to **metastable states**. The GA seems to spend most of its time in such states.

Often, when we say a GA has "converged", we really mean that it has arrived at a metastable state and has stayed there for a long time.

This happens because the map $\mathcal{G}$ is **continuous**.

A population that is close to a fixed-point has a high probability of staying in that region at the next generation...

...as long as the population is not too small! Small populations create more variance.

Finding the fixed-points of $\mathcal{G}$ helps us characterise the behaviour of the GA.

N.B. There may be fixed-points outside the set Λ, but which still influence the GAs behaviour. Some of these fixed-points may be complex!!

We therefore need to consider the extension of $\mathcal{G}$ to complex space.

There can also be large regions in which the GA seems to wander randomly. These regions are called "neutral" regions.

There is as yet no complete theory describing the relationship between the dynamics of $\mathcal{G}$, the population size, and the charactisation of metastable states and neutral regions.

This is an important open research problem.

"This GA is stuck at a local optimum."

This statement is often based on the confusion that a GA is a kind of local search algorithm. It is not! It is a population algorithm, in which items interact in a non-linear, stochastic way.

The idea of a local optimum in a GA might make sense if we are talking about a selection-crossover algorithm (i.e. no mutation).

We expect this GA to end up with a uniform population. The item it contains might be a local optimum (in Hamming space). Or it might not. . . .

Actually we have the following theorem.

Theorem Suppose $\mathcal{G}$ corresponds to a selection-crossover GA. A population vector corresponding to a uniform population is an **asymptotically stable** fixed-point of $\mathcal{G}$ only if the item it contains is a local optimum with respect to its Hamming neighbours.

It is also conjectured that uniform populations are the only asymptotically stable fixed-points of $\mathcal{G}$ in this case.

If this conjecture is true, it means that, generically, iterates of $\mathcal{G}$ will converge to uniform populations containing a local optimum.

Conclusion

- The term "convergence" is often used in a loose sense when describing GAs.

- We can make this concept more formal using GA theory.

- Formalizing things enables us to find out what is true.

- This helps us to sharpen up our intuitions.

Finally. . .

There has been a lot of progress made in recent years, but much remains to be done. Some highlights include:

- The algebraic characterisation of mixing.
- Generalisation to arbitrary finite search spaces.
- Differential fixed-point theory.
- Equivalence and coarse-graining.
- Generalisation to Genetic Programming.
- Statistical Mechanics models.
- Computational complexity results