Natural Computing

Lecture 4

Michael Herrmann
mherrman@inf.ed.ac.uk
phone: 0131 6 517177
Informatics Forum 1.42

30/09/2011

The Schema Theorem and the Building Block Hypothesis

Reminder: The Canonical Genetic Algorithm

1. Old population
2. Roulette-wheel selection
3. Intermediate population
4. Single point recombination with rate p_c (per pair of individuals)
5. Mutation with rate p_m (per position in all strings)
1. New population (repeat until termination)

} one generation

- A population is a (multi-) set of individuals
- An individual (genotype, chromosome) is encoded by a string $S \in \mathcal{A}^L$ ($\mathcal{A}$: alphabet; canonical: $\mathcal{A} = \{0, 1\}$, L fixed)
- Normalised fitness represents the objective of the problem

Search Spaces as Hypercubes

Binary encoding: solution $c \in \{0,1\}^L$

$\Rightarrow$ Each Solution is a corner of the **hypercube**.

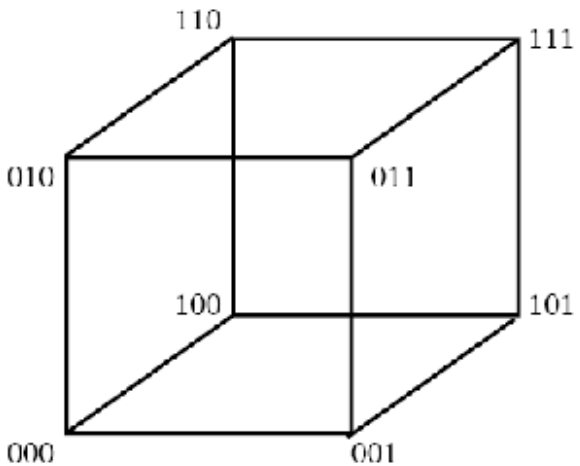
e.g. $c = (0,1,0)$ for $L = 3$ or $c = (0110)$ for $L = 4$

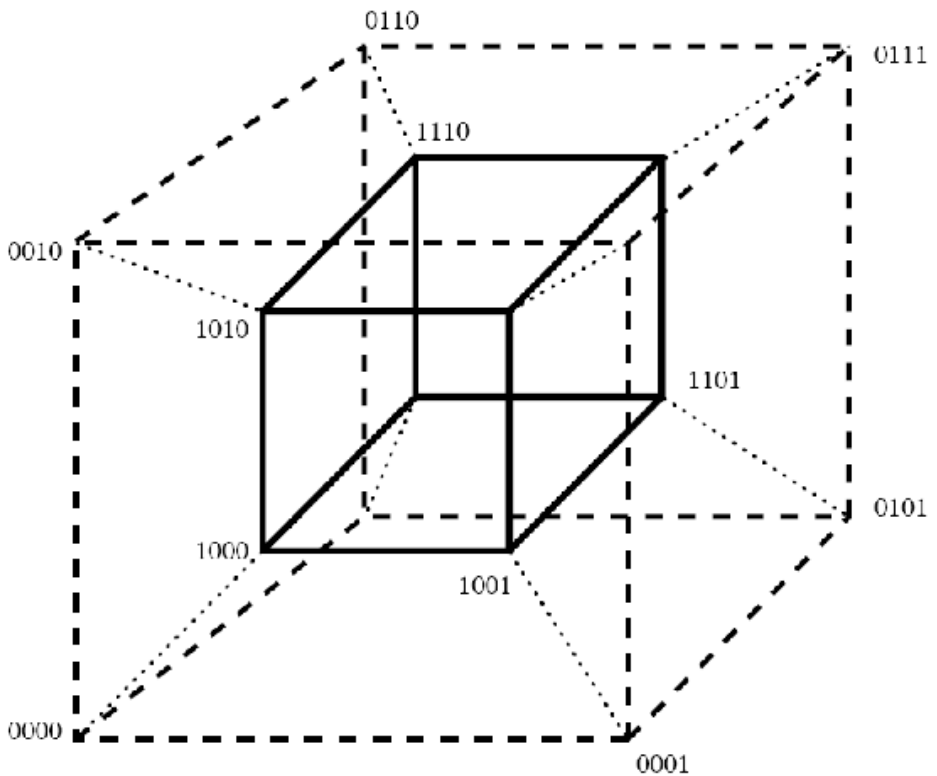
Sets of solutions:

$(0,*,0)$ denotes a line

$(*,1,*)$ denotes a plane

$(1,*,*,*)$ denotes a subcube.





From a tutorial by Erik D. Goldman GECCO09

Schemata (J. Holland, 1975)

- A schema is a string that contains wildcards ("*"), but not only asterisks, i.e. schema $H \in \{0, 1, *\}^L \setminus \{*\}^L$
- A schema defines a set of solutions (which coincide at the no-wildcard symbols)
- All (inheritable) features of the phenotype are encoded by schemata
- The order of the schema is the number of bits that are actually there, e.g. **01***1 is a schema of order 3 (and length 8)
- There are $3^L - 1$ different schemata (not counting the schema of order 0: ** ... *)
- Each solution is part of 2^L hyperplanes (or $2^L - 1$ schemata)
- Implicit parallelism: Each individual samples many hyperplanes

How do GAs work?
The schema theorem (J. Holland, 1975)

- How does selection improve fitness?
- What is the fate of the schemata in face of selection, mutation and and crossover?

Goal:

$$E\left(m\left(H,t+1\right)\right) \geq \frac{\hat{u}(H,t)}{\bar{f}(t)}\, m\left(H,t\right)\left(1-P_c\frac{d(H)}{L-1}\right)\left(1-p_m\right)^{o(H)}$$

- H is a schema
- There are m individuals carrying this schema
- t counts generations
- E is the mathematical expectation

The Schema Theorem
(before starting the proof)

Consider first an individual solution $c_i \in \mathcal{A}^L$:

$f(c_i, t)$: fitness of solution c_i in generation t

$m(c_i, t)$: number of copies of c_i in the population in generation t

$\bar{f}(t)$: average fitness of the population in generation t

$$\boxed{E(m(c_i, t+1)) = \frac{f(c_i,t)}{\bar{f}(t)} m(c_i, t)}$$

$E(\cdot)$ is the expected value

$\frac{1}{n} \frac{f(c_i,t)}{\bar{f}(t)}$ the probability of selecting c_i

$\bar{f} = \frac{1}{n} \sum_{i=1}^{n} f(c_i)$

n: population size

So above-average-fitness strings get more copies in the next generation and below average fitness strings get fewer.

Suppose c_i has above-average fitness of $(1+\delta)\,\bar{f}$ (i.e. $\delta > 0$). Then

$$E\left(m\left(c_i, t+1\right)\right) = \frac{f(c_i)}{\bar{f}} m\left(c_i, t\right) = \frac{(1+\delta)\bar{f}}{\bar{f}} m\left(c_i, t\right) = (1+\delta) m\left(c_i, t\right)$$

If δ is constant then $m\left(c_i, t\right) = (1+\delta)^t m(c_i, 0)$: Exponential growth

If $m\left(c_i\right)$ is small compared to the population size n then δ can indeed be considered constant $\Rightarrow$ Innovations that cause an increase in fitness spread quickly in the population.

Growth is self-limiting: The relative advantage shrinks because with more fit individuals also the average fitness increases$\Rightarrow$ Fit solution tend to dominate the population (crossover and mutation being ignored for the moment).

Analogously: Exponential decay for $\delta < 0$.

Fitness of Schemata

If the solutions c_i, c_j, c_k, ... all sample the same schema H their fitnesses define the (average) fitness of H at time t

$$\hat{u}(H,t) = \frac{1}{m(H,t)} \sum_{c_i \in H} m(c_i,t) f(c_i,t)$$

$m(H,t)$ is the number of instance of H in the population at time t

Note, that the sum is not taken over all possible $c_i \in H$ but only over those which are actually present in the population.

How many instances of H can be expected after selection?

$$E(m(H,t+1)) = \frac{\hat{u}(H,t)}{\bar{f}(t)} m(H,t)$$

Selection of Fit Schemata: Example

What happens when we select and duplicate strings based of fitness?

Suppose the solutions c_i, c_j, c_k sample the schema H, i.e. $c_i \in H$ etc.

Further suppose the average fitness in the population is $\bar{f} = 1$

Using the formula for solutions:

$$f(c_i, t) = 2.0,\ m(c_i, t) = 2 \quad \Rightarrow \quad E(m(c_i, t+1)) = 2 \times \frac{2.0}{1.0} = 4$$
$$f(c_j, t) = 2.5,\ m(c_j, t) = 2 \quad \Rightarrow \quad E(m(c_j, t+1)) = 2 \times \frac{2.5}{1.0} = 5$$
$$f(c_k, t) = 1.5,\ m(c_k, t) = 2 \quad \Rightarrow \quad E(m(c_k, t+1)) = 2 \times \frac{1.5}{1.0} = 3$$

All are fitter than average, all increase in their number in the population.

For the schema H (assume sampled only by c_i, c_j, c_k): $m(H, t) = 6$,

$$\hat{u}(H, t) = \frac{1}{6}(2 \times 2.0 + 2 \times 2.5 + 2 \times 1.5) = 2$$

$$\hat{u}(H, t+1) = \frac{1}{12}(4 \times 2.0 + 5 \times 2.5 + 3 \times 1.5) = 2,083$$

Number of samples in this hyperplane is expected to increase, but...

Disruption of Schemata

Crossover and mutation are both disruptive and constructive with regards to schemata. consider only disruptive effects.

Crossover:

1 1 * * * * * *	Probability of disruption by crossover?
1 * * * * * * 1	

Mutation:

1 1 0 0 1 0 0 1 1 1 0 1 * *	Many disruptive possibilities
1 1 * * * * 0 1 * * * * * *	Only 4 disruptive possibilities

Schema Jargon

Number of defined bits is the **order** $o(H)$ of the schema H

1 0 * * 1 1 0 * order 5

* 0 * * 1 1 * * order 3

Defining length is the distance $d(H)$ between the first and the last bit of the schema (i.e. number of potential cuts)

1 0 * * 1 1 0 * defining length 6

* 0 * * 1 1 * * defining length 4

i.e. bit position of last 0/1 minus bit position of first 0/1

Disruptive Effects of Crossover

- 1-point crossover with probability p_c
- $d(H)$ is the defining length of H
  $H = * * 1\ 0 * 1 * * \Rightarrow d(H) = 3$
- In a single crossover there are $L-1$ crossover points:
  1 0 1 0 0 1 0 0 7 crossover points
- Of these, $d(H)$ points will disrupt the schema

$$\Pr(\text{disruption}) = p_c \frac{d(H)}{L-1}$$

- Higher chance of survival if $d(H)$ is low

Example: Suppose $p_c = 0.8$, $d(H) = 3$, $L = 100 \Rightarrow$

$$\Pr(\text{disruption}) = 0.8 \times \frac{3}{100} = 0.024$$

Disruptive Effects of Mutation

- Single-point mutation with probability p_m (applied to each bit in turn)

  $o(H)$ is the order of H
  $H = * * 1\ 0 * 1 * * \Rightarrow o(H) = 3$
  $H = 1\ 1\ 1\ 0 * 1 * 1 \Rightarrow o(H) = 6$

- Probability that a bit survives is $1 - p_m$
- Flipping a defined bit always disrupts a schema, so the probability that the schema survives is

$$\Pr(\text{survival}) = (1 - p_m)^{o(H)}$$

- Best chances for surviving crossover and mutation when $d(H)$ and $o(H)$ are both low

Towards the Schema Theorem

First Component of the Schema Theorem

$$E\left(m\left(H,t+1\right)\right)=\frac{\hat{u}\left(H,t\right)}{\bar{f}\left(t\right)}m\left(H,t\right)$$

The other parts of the Schema Theorem

$$\Pr\left(\text{surviving crossover}\right)=1-p_c\frac{d\left(H\right)}{L-1}$$

$$\Pr\left(\text{surviving mutation}\right)=\left(1-p_m\right)^{o(H)}$$

$$E\left(m\left(H,t+1\right)\right)=\frac{\hat{u}\left(H,t\right)}{\bar{f}\left(t\right)}m\left(H,t\right)\left(1-p_c\frac{d\left(H\right)}{L-1}\right)\left(1-p_m\right)^{o(H)}\quad ???$$

The Schema Theorem

Schemata are not only being destroyed, but can also be created though crossover and mutation. So we should write an inequality

Goal:

$$E\left(m\left(H,t+1\right)\right) \geq \frac{\hat{u}(H,t)}{\bar{f}(t)} m\left(H,t\right)\left(1-P_c\frac{d(H)}{L-1}\right)\left(1-p_m\right)^{o(H)}$$

Highest when

- $\hat{u}\left(H,t\right)$ is large: fit
- $d\left(H\right)$ is small: short
- $o\left(H\right)$ is small: small number of defined bits

The Schema Theorem in words:
Short, low-order, above-average schemata receive exponentially increasing trials in subsequent generation of a genetic algorithm.

Consequences of the Schema Theorem

- How do schemata arise? Constructive role of mutation and crossover
- Which genes belong to a good schema?
  The algorithm does not easily distinguish important genes from "hitchhikers"
- How well does the expectation describe the population?
- Gradual reduction of relative fitness advantage:
  Other ways to change the fitness?

The Building Block Hypothesis

During crossover, these "building blocks" become exchanged and combined

So the Schema Theorem identifies the building blocks of a good solution although it only addresses the disruptive effects of crossover (and the constructive effects of crossover are supposed to be a large part of why GA work).
How do we address the constructive effects?

> Building block hypothesis: A genetic algorithm seeks optimal performance through the juxtaposition of short, low-order, high-performance schemata, called the building blocks.

Crossover combines short, low-order schemata into increasingly fit candidate solutions

- short low-order, high-fitness schemata
- "stepping stone" solutions which combine H_i and H_j to create even higher fitness schemata

The Building Block Hypothesis

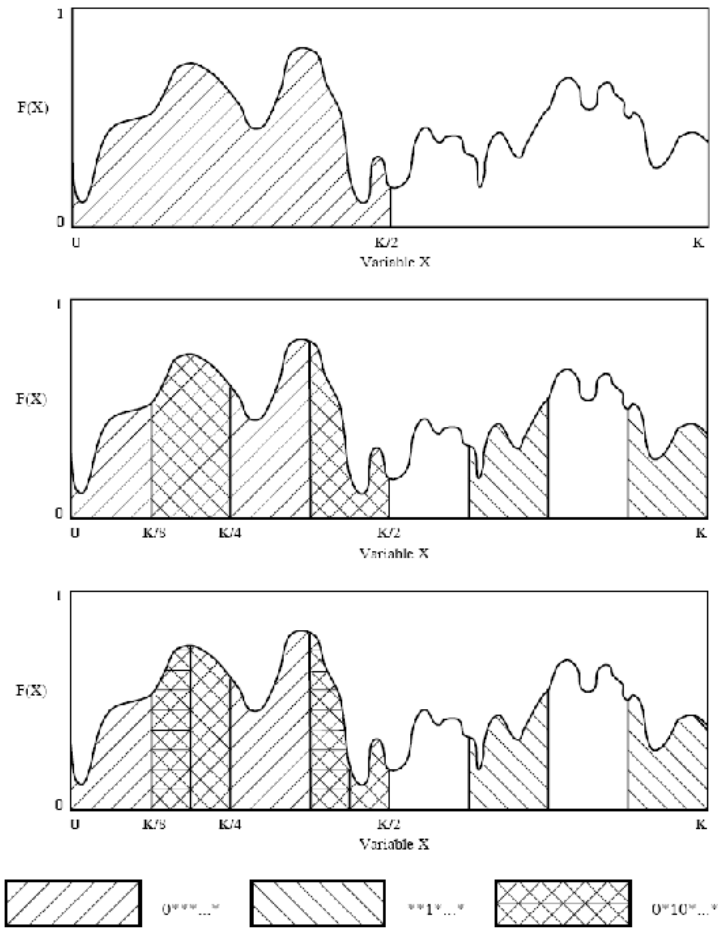
Arguments against the validity of the BBH

- **Collateral convergence:** Once the population begins to converge, even slightly, it is no longer possible to estimate the static average fitness of schemata using the information present in the current population.
- **Fitness variance within schemata:** In populations of realistic size, the observed fitness of a schema may be arbitrarily far from the static average fitness, even in the initial population.
- **Compositionality:** Superposition of fit schemata does not guarantee larger schemata that are more fit and these are less likely to survive.

Adapted from John J. Grefenstette. Deception Considered Harmful. 1992

Encoding and Local Optima

- Binary encoding of a 1-D variable
- Fitness of a schema is the average over the corresponding hyperplane (or rather the sample across the population)
- Sampling of the hyperplanes is essentially unaffected by local optima



From: Whitley, 1992

The Building Block Hypothesis
Experimental Evidence

The Building Block Hypothesis is a hypothesis – so we can do an experiment to test it.

Experiment: Use a problem which contains explicit building blocks and observe the population. Do the building blocks combine to give a good solution in the way the BBH predicts?

Mitchel, Forrest, Holland set up such a problem, using royal Road functions. Details: Mitchel, Chapter 4, pp 127-133.

Define fitness in terms of particular schemata:
Substrings that, if present in a population ought to be combinable into the optimal solution.
They should lay out a "Royal Road" to the global optimum.

The first RR function R_1 is defined using a list of schemata s_i.
Each s_i has a fitness coefficient c_iThe fitness $R_1(x)$ of some bit string x is given by: $R_1(x) = \sum_i c_i \delta_i$, $\delta_i(x) = \begin{cases} 1 & \text{if } x \subset s_i \\ 0 & \text{otherwise} \end{cases}$

Royal Road Functions

Simple example using 16 bits. Suppose:

$s_1 = 1111************$

$s_2 = ****1111********$

$s_3 = ********1111****$

$s_4 = ************1111$

and suppose $c_1 = c_2 = c_3 = c_4 = 4$ and

$S_{\text{opt}} = 1111111111111111$

Then $R_1\left(S_{\text{opt}}\right) = \sum_{i=1}^{4} c_i \delta_i\left(S_{\text{opt}}\right) = 16$

Take the string 1111010010011111. It samples matches s_1 and s_4. So $\delta_1(x) = \delta_4(x) = 1$ and $\delta_2(x) = \delta_3(x) = 0$.

And $R_1(1111010010011111) = 8$

Royal Road Functions

Several Royal Road functions defined in terms of different combinations of schemata with building blocks at different levels, e.g. 4 contiguous 1s, 8 contiguous 1s, 18 contiguous 1s, etc.

Try to evolve the string with all 1s and compare performance of GA against a number of hill-climbing schemes

- Steepest-ascent hill climbing (SAHC)
- Next-ascent hill climbing (NAHC)
- Random mutation hill climbing (RMHC)

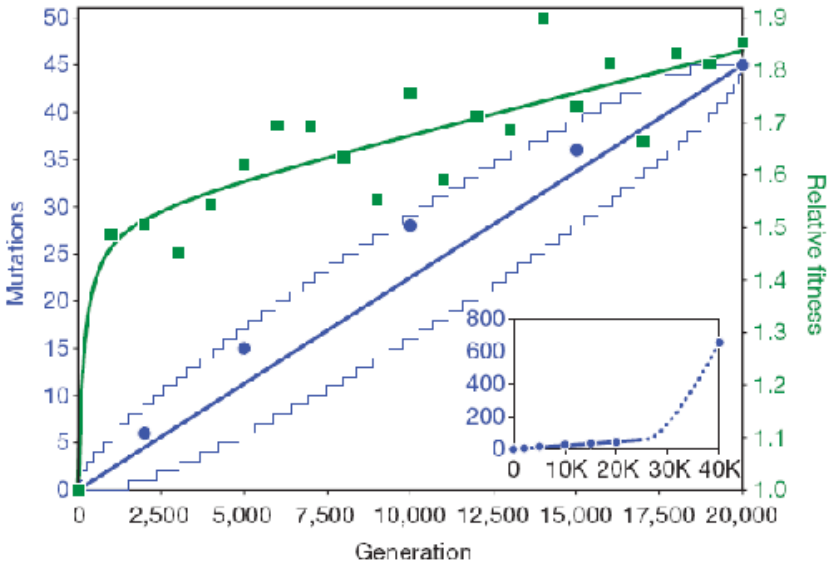
Will the GA do better?

Outlook

- More implications of the schema theorem
- More examples
- Variants and hybrid algorithms
- Genetic programming
- Continuous evolutionary algorithms

GA: Experimental Evidence for Parameter Variation

- 40000 generations 1988-2009: 21 years
- Initial population: 12 strains of Escherichia Coli
- Constant conditions: restricted glucose supply



- First half (20000 generations): 45 mutations, often related to life span and efficiency
- Second half (20000 generations): 653 mutations in some strains but without any obvious effects on fitness
- Conclusions: Relations between mutation rate and fitness are more complex than expected

J. Barrick et al. (2009) Genome evolution and adaptation in a long term experiment with Escherichia coli. Nature 08480.