Standard GA theory

- Schema theory, schema theorem
- Two-armed bandit analogy
- Almost sure convergence
- Building block hypothesis
- Implicit parallelism
- Deceptive problems

Standard GA theory

Dominating theoretical paradigm 1975-1992:
Central notion: schema
schema analysis in many forms

Most recently:
- No Free Lunch Theorem (NFL)
- theoretical analysis with stronger predictive capabilities ($\Rightarrow$ time complexity!)
- ideas from quantitative genetics, physics, etc.

Schemata (1)

Schema (definition):
A schema H in IB^ℓ is a *partial* instantiation of a string in IB^ℓ . Usually the uninstantiated elements are denoted by '$*$', sometimes called "don't care" symbol or "wild card". A schema defines a subset of IB^ℓ :

$$H \in \{0, 1, *\}^\ell$$

Example:
H = {1, 0, 0, $*$, 1, $*$, 1, 0, $*$, $*$, $*$}

Further definitions:

- Instance of the schema H: 1 0 0 $*$ 1 $*$ 1 0 $*$ $*$ $*$ / 1 0 0 0 1 0 1 0 1 0 1

- Set of all instances of schema H = (h_1, … , h_ℓ):
  I(H) = {(a_1, … , a_ℓ) $\in$ IB^ℓ | $h_i \neq * \Rightarrow a_i = h_i$}

Schemata (2)

Further definitions continued:

- Order of the schema: Number of instantiated elements (6 in our example).

$$o(H) = |\{ i \mid h_i \in \{0,1\} \}|$$

- Defining length of the schema: length of the sub-string starting at the first and ending at the last instantiated element (7 in our example). Idea: it is the number of possible breakpoints

$$\underbrace{1\,0\,0 * 1 * 1\,0}\,* * *$$

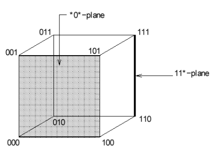
$$d(H) = \max\{ i \mid h_i \in \{0,1\} \} - \min\{ i \mid h_i \in \{0,1\} \}$$

Schemata (3)

Some numbers:
- In total there are 3^ℓ different schemata.
- Each chromosome (in IB^ℓ) is an instance of 2^ℓ different schemata.
- Thus: at most N · 2^ℓ schemata are represented in a population of size N.

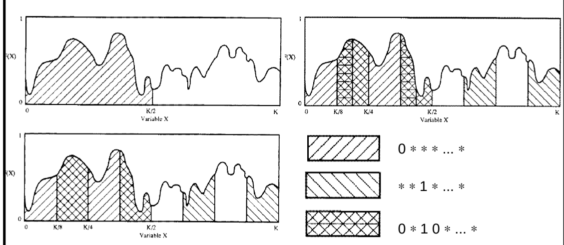
A schema can be viewed as a hyperplane of an n-dimensional space.

Examples in a 3-dimensional (hyper)cube:



Schemata (4)

A function and various (schema) partitions of hyperspace:



Schema Theorem (1)

Theorem (Holland '75):

$$m(H,t+1) \geq m(H,t) \cdot \frac{f(H)}{\bar{f}} \cdot \left(1 - p_c \frac{d(H)}{l-1}\right) \cdot (1-p_m)^{o(H)}$$

- f to be maximised, $\bar{f}$: mean fitness in population
- l: length of the string
- H: a schema
- d(H): defining length
- o(H): order
- p_m: mutation rate
- p_c: crossover rate
- f(H): (estimated) schema fitness
- m(H, t): expected number of instantiations of H in generation t

Schema Theorem (2)

- Expected number of instantiations of H selected for the gene pool:

$$m(H,t) \cdot \frac{f(H)}{\bar{f}}$$

- Probability that crossover does not occur within the defining length:

$$1 - p_c \frac{d(H)}{l-1}$$

- Probability that the schema is not mutated:

$$(1-p_m)^{o(H)}$$

Schema Theorem (3)

Critique on the schema theorem (Bäck '96):

- Most of Holland's approximations are only true for very large numbers (trials and population size).
- Within finite populations, exponentially increasing/decreasing the number of schema instances, leads to entirely filling the population and complete elimination, respectively.
- Not all schemata are represented in a typical population.
- Schemata of large defining length are likely to be destroyed by crossover (even highly fit ones).

The Goldberg example (1)

String number	Initial population	x value	$f(x)$	pselect$_i$	Expected count	Actual count
	(Randomly generated)	(Unsigned integer)	(x^2)	$\left(\frac{f_i}{\Sigma f}\right)$	$\left(\frac{f_i}{\bar{f}}\right)$	(From roulette wheel)
1	0 1 1 0 1	13	169	0.14	0.58	1
2	1 1 0 0 0	24	576	0.49	1.97	2
3	0 1 0 0 0	8	64	0.06	0.22	0
4	1 0 0 1 1	19	361	0.31	1.23	1
Sum			1170	1.00	4.00	4.0
Average			293	0.25	1.00	1.0
Max			576	0.49	1.97	2.0

Schema Processing

		Strings Representatives	Schema Average Fitness $f(H)$
H_1	1 * * * *	2,4	469
H_2	* 1 0 * *	2,3	320
H_3	1 * * * 0	2	576

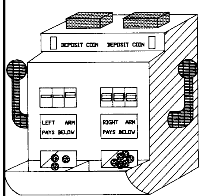
The Goldberg example (2)

Mating pool after reproduction	Mate	Crossover site	New population	x Value	$f(x)$
(Cross site shown)	(Randomly selected)	(Randomly selected)			(x^2)
0 1 1 0 | 1	2	4	0 1 1 0 0	12	144
1 1 0 0 | 0	1	4	1 1 0 0 1	25	625
1 1 | 0 0 0	4	2	1 1 0 1 1	27	729
1 0 | 0 1 1	3	2	1 0 0 0 0	16	256
					1754
					439
					729

Schema Processing

After reproduction			After all operators		
Expected Count	Actual Count	String Representatives	Expected Count	Actual Count	String Representatives
3.20	3	2,3,4	3.20	3	2,3,4
2.18	2	2,3	1.64	2	2,3
1.97	2	2	0.00	1	4

K-armed bandits and GA's (1)



Two-armed bandit problem:

- arm_1: award μ_1 with variance σ_1^2
- arm_2: award μ_2 with variance σ_2^2

- $\mu_1 > \mu_2$
- Which arm (left/right) is which (1,2)?

- N = total number of trials
- b = $\frac{\sigma_1}{\mu_1 - \mu_2}$

K-armed bandits and GA's (2)

Theorem (Holland '75)
Expected loss is minimal if approximately:

$$[1] \quad n^* \approx b^2 \cdot \ln\left(\frac{N^2}{8\pi \cdot b^4 \cdot \ln(N^2)}\right)$$

trials are allocated to the observed worst arm.

Corollary:
Expected loss is minimal if approximately:

$$[2] \quad \mathrm{N} - n^* \approx \sqrt{8\pi \cdot b^4 \cdot \ln(N^2)} \cdot e^{\frac{n^*}{2b^2}}$$

trials are allocated to the observed best arm.

K-armed bandits and GA's (3)

We can generalize the two-armed bandit to a k-armed bandit, this gives us:

1. **Generalized corollary**:
   Optimal strategy is to allocate an exponentially increasing number of trials to the observed best arms.
2. **A link to GA's**:
   Minimizing expected losses from k-armed bandits

   ≈

   Minimizing expected losses while sampling from order $\log_2(k)$ schemata.

Thus, GA's allocate trials (near-)optimally.

Point of critique (Fogel '95): Why would this be optimal for *global* optimization? (Minimizing expected losses does not always correspond to maximizing potential gains.)

Almost sure convergence (1)

Theorem (Eiben et al. '91):

$$P[\lim_{t\to\infty} P_t \cap Optima \neq 0] = 1$$

- P_t: population at time t
- *Optima:* Set of global optima
- Prerequisites:
  - $\max_{\bar{x}\in P(t)} f(\bar{x}) \geq \max_{\bar{x}\in P(t-1)} f(\bar{x})$ (e.g. by elitist selection)
  - any point is accessible from any other point (OK if $p_m > 0$).

Almost sure convergence (2)

Critique on the theorem:

- It says nothing about convergence speed
- Theory people: I don't care if it works if only it converges
- Practice people: I don't care if it converges if only it works

Theorem later generalized by Rudolph

Implicit parallelism

- Implicit parallelism: a GA with population size N processes more than N different schemata effectively
- Reason: individuals are instantiations of more than one schema
- Effectively processing of a schema:

  Sampled at the desirable exponentially increasing rate.

- Why wouldn't a schema be processed effectively?

  Schema disruption by genetic operators!

- Holland's estimate: $O(N^3)$ schemata are processed effectively when using a population of size N

The Building Block Hypothesis

Building Block Hypothesis (Holland '75):
GA's are able to detect short, low order and highly fit schemata and combine these into highly fit individuals.

Building blocks are small and good schemata, where:

- small is:
  - short (i.e have a small defining length)
  - of low order, and
- good is: highly fit (estimated fitness in present population).

Implicit parallelism and the Building Block Hypothesis are seen as explanations for the power of GA's.

The minimal deceptive problem (1)

Goal: Define problems that mislead GA's (Goldberg '89)

- Deceptive: building blocks (small and good schemata) lead to incorrect, i.e. suboptimal, solutions
- Minimal: smallest case with such deception (two bits)
- Next we show how to construct a minimal deceptive problem

The minimal deceptive problem (2)

schema 1:	*	*	*	0	*	*	*	0	*	*		f_{00}
schema 2:	*	*	*	0	*	*	*	1	*	*		f_{01}
schema 3:	*	*	*	1	*	*	*	0	*	*		f_{10}
schema 4:	*	*	*	1	*	*	*	1	*	*		f_{11}

$|\leftarrow \quad \delta(H) \quad \rightarrow|$

- f_{xy} is the *average* fitness of the schema * * x * * * y * * (assume no variance; for expected performance this assumption can be left out).
- Assume the global optimum is an instance of schema 1 and

$$[1] \quad f_{11} > \begin{cases} f_{01} \\ f_{10} \\ f_{00} \end{cases}$$

- For the problem to be deceptive, we need:

$$[2] \quad f_{0*} > f_{1*} \quad \text{and/or} \quad [3] \quad f_{*0} > f_{*1}$$

The minimal deceptive problem (3)

- This is true iff:

$$[2] \quad \frac{f_{00}+f_{01}}{2} > \frac{f_{10}+f_{11}}{2}, \text{ or}$$

$$[3] \quad \frac{f_{00}+f_{10}}{2} > \frac{f_{01}+f_{11}}{2}$$

- Choosing (arbitrarily) for [2] we normalise with:

$$r = \frac{f_{11}}{f_{00}} \qquad c = \frac{f_{01}}{f_{00}} \qquad c' = \frac{f_{10}}{f_{00}}$$

- Using [1] we get: [4] r > c', r > 1 and r > c.
- Using [2] we get: [5] r < 1 + c - c'.
- From [4] and [5] we can conclude: [6] c'< 1 and c' < c.

The minimal deceptive problem (4)

- This leaves us with two possibilities:

Type I: $f_{01} > f_{00}$ (c > 1) Type II: $f_{01} \le f_{00}$ (c ≤ 1)