计算智能

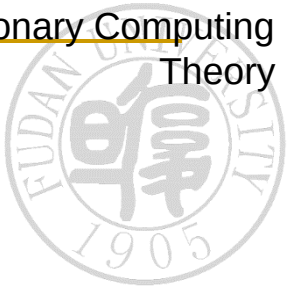
第5讲: Part I – 进化算法的理论基础

周水庚

计算机科学技术学院

2017-4-11

The Theory of EA



Overview

- Motivations and problems
- Holland's Schema Theorem
  - Derivation, Implications, Refinements
- Dynamical Systems & Markov Chain Models
- Statistical Mechanics
- Reductionist Techniques
- Techniques for Continuous Spaces
- No Free Lunch ?

Why Bother with Theory?

- Might provide performance guarantees
  - Convergence to the global optimum can be guaranteed providing certain conditions hold
- Might aid better algorithm design
  - Increased understanding can be gained about operator interplay etc.
- Mathematical Models of EAs also inform theoretical biologists

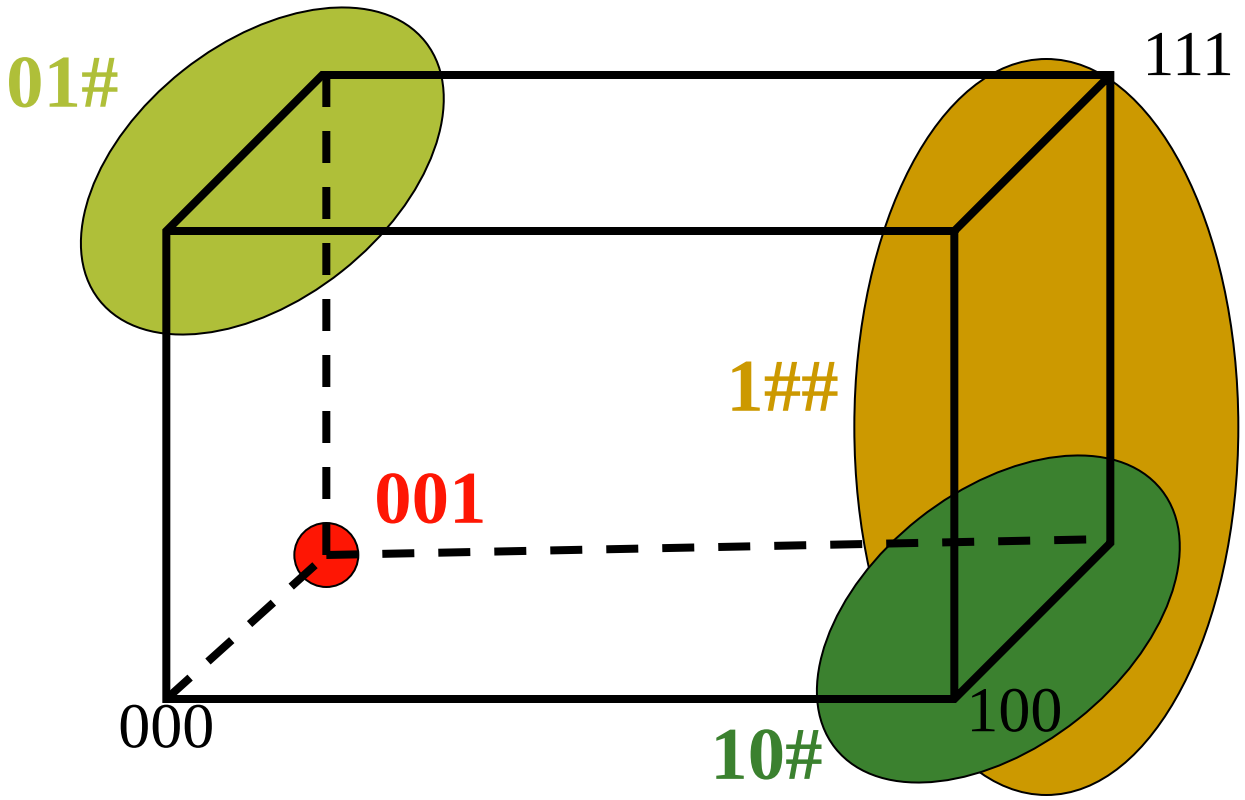
Problems with Theory ?

- EAs are vast, complex dynamical systems with many degrees of freedom
- The type of problems for which they do well, are precisely those it is hard to model
- The degree of randomness involved means
  - Stochastic analysis techniques must be used
  - Results tend to describe average behaviour
- After 100 years of work in theoretical biology, they are still using fairly crude models of very simple systems ….

Holland's Schema Theorem

- A schema (*pl.* schemata) is a string in a ternary alphabet (0,1 # = "don't care") representing a hyper plane within the solution space.
  - E.g. S1: 0001# #1# #0#
  - E.g. S2: ##1##0##
- Two values can be used to describe schemata,
  - the ***Order*** (number of defined positions) = 6,2
  - the ***Defining Length*** - length of sub-string between outmost defined positions = 9, 3

Example Schemata



H	o(H)	d(H)
001	3	2
01#	2	1
10#	2	1
1##	1	0

Schema Fitnesses

- The true "fitness" of a schema *H* is taken by averaging over all possible values in the "don't care" positions, but this is effectively sampled by the population, giving an estimated fitness *f(H)*.
- With Fitness Proportionate Selection

  $P_s(\text{instance of } H) = m(H,t) * f(H,t) / (\langle f \rangle * \mu)$

  therefore proportion in next mating pool is:

  $m'(H,t+1) = m(H,t) * f(H,t) / \langle f \rangle$

Schema Disruption I

- One Point Crossover selects a crossover point at random from the *l-1* possible points
- For a schema with defining length *d* the random point will fall inside the schema with probability = $d(H) / (l-1)$.
- If recombination is applied with probability P_c the survival probability is $1.0 - P_c * d(H)/(l-1)$

Schema Disruption II

- The probability that bit-wise mutation with probability P_m will NOT disrupt the schema is simply the probability that mutation does NOT occur in any of the defining positions,

  $P_{survive}$ (mutation) = $(1- Pm)^{o(H)}$

  $= 1 - o(H) * P_m + terms\ in\ P_m^2 + \ldots$

- For low mutation rates, this survival probability under mutation approximates to $1 - o(h)* P_m$

The Schema Theorem

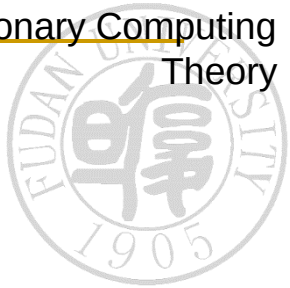
- Put together, the proportion of a schema H in successive generations varies as:

$$m(H, t+1) \geq m(H, t) \cdot \frac{f(H)}{< f >} \cdot \left[1 - \left(p_c \cdot \frac{d(H)}{l-1}\right)\right] \cdot [1 - p_m \cdot o(H)] .$$

- In words: short, low-order schemata with above average fitness will increase their representatives from generation to generation
- Inequality is due to convergence affecting crossover disruption, exact versions have been developed

Implications 1: Operator Bias

- One Point Crossover
  - less likely to disrupt schemata which have **short** defining lengths relative to their order, as it will tend to keep together adjacent genes
  - this is an example of ***Positional Bias***
- Uniform Crossover
  - No positional bias since choices independent
  - BUT is far more likely to pick 50% of the bits from each parent, less likely to pick (say) 90% from one
  - this is called ***Distributional Bias***
- Mutation
  - also shows Distributional Bias, but not Positional



Operator Biases (cont'd)

- Operator Bias has been extensively studied by Eschelman and Schaffer (empirically) and theoretically by Spears & DeJong.
- Results emphasise the importance of utilising all available problem specific knowledge when choosing a representation and operators for a new problem

Implications 2:The Building Block Hypothesis

- Closely related to the Schema Theorem is the "Building Block Hypothesis" (Goldberg 1989)
- This suggests that Genetic Algorithms work by discovering and exploiting "building blocks" (highly fit short low-order schemata) and then successively combining these (via crossover) to produce successively larger building blocks until the problem is solved.
- Has motivated study of ***Deceptive*** problems
  - Based on the notion that the lower order schemata within a partition lead the search in the opposite direction to the global optimum

Deception

- What happens if the global optimum is not an example of the short low-order schemata with the highest mean fitness?
- Example: Binary representation with 3 genes, f(x)=the number of 1's but f(000)=3.5
- Many theoretical studies on GA-deceptive functions
- No conclusive results
- Problem for all methods rather than GA only.

Criticisms of the Schema Theorem

- It presents an inequality that does not take into account the constructive effects of crossover and mutation
  - Exact versions have been derived
  - Has links to Price's theorem in biology
- Because the mean population fitness and the estimated fitness of a schema will vary from generation to generation, it says ***nothing*** about gen. t+2, .. etc.
- "Royal Road" problems constructed to be GA-easy based on schema theorem turned out to be better solved by random mutation hill-climbers
- BUT it remains a useful conceptual tool and has historical importance

Other Landscape Metrics

- As well as epistasis and deception, several other features of search landscapes have been proposed as providing explanations as to what sort of problems will prove hard for GAs
  - fitness-distance correlation
  - number of peaks present in the landscape
  - the existence of plateaus
  - all these imply a neighbourhood structure to the search space.

Vose' Dynamical Systems Model

- Let *n* be the size of a finite search space
- Construct a population vector ***p*** with *n* elements giving the proportion of the population in each possible state.
- *n x n* ***Mixing Matrix M,*** represents operation of crossover and mutation on population
- *n x n* ***Selection Matrix F*** represents action of selection

$$\bar{p}^{t+1} = F \circ M\bar{p}^{t} = G\bar{p}^{t}$$

Dynamical Systems 2

- The existence, location and stability of fixed points or attractors for this system depend on the set of coupled equations defining G
- Note that these are infinite population models
  - extensions to finite populations are possible but computationally intractable
- Lots of interest in ways of aggregating states into equivalence classes
  - schemata are one option

Markov Chain Analysis

- A system is called a Markov Chain if
  - It can exist only in one of a finite number of states
  - So can be described by a variable X^t
  - The probability of being in any state at time $t+1$ depends only on the state at time t.
- Frequently these probabilities can be defined in a transition matrix, and the theory of stochastic processes allows us to reason using them.
- Has been used to provide convergence proofs
- Can be used with F and M to create exact probabilistic models for binary coded GAs, but these are huge

Statistical mechanics models

- Use techniques borrowed from physics
- Just as per schemata, these use some statistics to model the behaviour of a complex system
- Statistics chosen are the cumulants of the fitness distribution
  - related to mean, variance, skewness, etc. of population fitness
  - Cannot model e.g. best fitness
- Can provide more accurate predictions of short term (rather than steady state) behaviour of finite pops. than dynamical systems approaches

Reductionist Approaches

- "Engineering" type approach of studying different operators and processes in isolation
  - Analysis of Takeover times for different selection operators via Markov Chains
  - Analysis of "mixing times" for crossover to put together building blocks to create new solutions
  - Analysis of population sizes needed for different problems, under different conditions.
- Can provide useful pointers for designing EAs in practice, e.g.

  T(takeover) < T(mixing) =>premature convergence

Continuous Space models

- Theory (with ES) is more advanced than for discrete spaces, includes self-adaptation
- Most analyses models two variables:
  - Progress Rate: distance of centre of mass of pop from global optimum as a function of time
  - Quality Gain : expected improvement in fitness between generations
- Lots of theory describing behaviour on simple models (e.g. spheres, corridors)
  - These are often good descriptors of *local* properties of landscapes
  - Theory has been extended to noisy environments

No Free Lunch Theorems

- IN LAYMAN'S TERMS,
  - Averaged over all problems
  - For any performance metric related to number of distinct points seen
  - All non-revisiting black-box algorithms will display the same performance
- Implications
  - New black box algorithm is good for one problem => probably poor for another
  - Makes sense not to use "black-box algorithms"
- Lots of ongoing work showing counter-examples