

Lecture 3: Schema Theory

Suggested reading: D. E. Goldberg, *Genetic Algorithm in Search, Optimization, and Machine Learning*, Addison Wesley Publishing Company, January 1989



Schema Theorem

- Schema theorem serves as the analysis tool for the GA process
- Explain why GAs work by showing the expectation of schema survival
- Applicable to a canonical GA
 - ☐ binary representation
 - ☐ fixed length individuals
 - ☐ fitness proportional selection
 - ☐ single point crossover
 - ☐ gene-wise mutation



Schema

- A **schema** is a set of binary strings that match the template for schema *H*
- A template is made up of 1s, 0s, and *s where * is the ‘don’t care’ symbol that matches either 0 or 1



Schema Examples

- The schema $H = 10^*1^*$ represents the set of binary strings

10010, 10011, 10110, 10111

- The string '10' of length $l = 2$ belongs to $2^l = 2^2$ different schemas

** , *0 , 1* , 10



Schema: Order $o(H)$

- The order of a schema is the number of its fixed bits, i.e. the number of bits that are not ‘*’ in the schema H
- Example: if $H = 10*1*$ then $o(H) = 3$



Schema: Defining Length $\delta(H)$

- The defining length is the distance between its first and the last fixed bits
- Example: if $H = *1*01$ then $\delta(H) = 5 - 2 = 3$
- Example: if $H = 0****$ then $\delta(H) = 1 - 1 = 0$



Schema: Count

- Suppose x is an individual that belongs to the schema H , then we say that **x is an instance of H** ($x \in H$)
- $m(H, k)$ denotes the number of instances of H in the k th generation



Schema: Fitness

- $f(x)$ denotes fitness value of x
- $f(H,k)$ denotes **average fitness of H** in the k -th generation

$$f(H,k) = \frac{\sum_{x \in H} f(x)}{m(H,k)}$$



Effect of GA On A Schema

- Effect of Selection
- Effect of Crossover
- Effect of Mutation
- = Schema Theorem



Effect of Selection on Schema

- Assumption: fitness proportional selection
- Selection probability for the individual x

$$p_s(x) = \frac{f(x)}{\sum_{i=1}^N f(x_i)}$$

where the N is the total number of individuals



Net Effect of Selection

- The expected number of instances of H in the mating pool $M(H, k)$ is

$$M(H, k) = \frac{\sum_{x \in H} f(x)}{\bar{f}} = m(H, k) \frac{f(H, k)}{\bar{f}}$$

➔ **Schemas with fitness greater than the population average are likely to appear more in the next generation**



Effect of Crossover on Schema

- Assumption: single-point crossover
- Schema H survives crossover operation if
 - one of the parents is an instance of the schema H **AND**
 - one of the offspring is an instance of the schema H

Crossover Survival Examples

Consider $H = *10**$

$P_1 = 1\ 1\ 0\ 1\ 0 \in H$
 $P_2 = 1\ 0\ 1\ 1\ 1 \notin H$ \rightarrow $S_1 = 1\ 1\ 0\ 1\ 1 \in H$ Schema H
 $S_2 = 1\ 0\ 1\ 1\ 0 \notin H$ **survived**

$P_1 = 1\ 1\ 0\ 1\ 0 \in H$
 $P_2 = 1\ 0\ 1\ 1\ 1 \notin H$ \rightarrow $S_1 = 1\ 1\ 1\ 1\ 1 \notin H$ Schema H
 $S_2 = 1\ 0\ 0\ 1\ 0 \notin H$ **destroyed**



Crossover Operation

- Suppose a parent is an instance of a schema H . When the crossover is occurred within the bits of the defining length, it is destroyed unless the other parent repairs the destroyed portion
- Given a string with length l and a schema H with the defining length $\delta(H)$, the probability that the crossover occurs within the bits of the defining length is $\delta(H)/(l - 1)$



Crossover Probability Example

- Suppose $H = *1**0$
- We gave
 - $l = 5$
 - $\delta(H) = 5 - 2 = 3$
- Thus, the probability that the crossover occurs within the defining length is $3/4$



Crossover Operation

- The upper bound of the probability that the schema H being destroyed is

$$D_c(H) \leq p_c \frac{\delta(H)}{l-1}$$

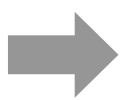
where p_c is the crossover probability



Net Effect of Crossover

- The lower bound on the probability $S_c(H)$ that H survives is

$$S_c(H) = 1 - D_c(H) \geq 1 - p_c \frac{\delta(H)}{l-1}$$



Schemas with low order are more likely to survive



Mutation Operation

- Assumption: mutation is applied gene by gene
- For a schema H to survive, all fixed bits must remain unchanged
- Probability of a gene not being changed is

$$(1 - p_m)$$

where p_m is the mutation probability of a gene



Net Effect of Mutation

- The probability a schema H survives under mutation

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

➔ **Schemas with low order are more likely to survive**



Schema Theorem

Exp. # of Schema H in Next Generation $>$

Exp. # in Mating Pool ($M(H, k) = m(H, k) \frac{f(H, k)}{\bar{f}}$)

Prob. of Surviving Crossover ($S_c(H) \geq 1 - p_c \frac{\delta(H)}{l-1}$)

Prob. of Surviving Mutation ($S_m(H) = (1 - p_m)^{o(H)}$)

Schema Theorem

■ Mathematically

$$E[m(H, k + 1)] \geq m(H, k) \frac{f(H, k)}{\bar{f}} \left(1 - p_c \frac{\delta(H)}{l - 1} \right) (1 - p_m)^{o(H)}$$

➔ **The schema theorem states that the schema with *above average fitness*, *short defining length* and *lower order* is more likely to survive**