Holland's schema theorem

Nicola Gulmini

1 The theorem

Let us fix the alphabet $\Sigma = \{0, 1, \star\}$ with the special charachter $\star$. A symbol $h = \star$ is called *wildcard*, otherwise $h \neq \star$ is called *specification*. A *binary schema* is a string $H = h_1 \dots h_n \in \Sigma^{n>0}$.

A string $S = s_1 \dots s_n \in \{0, 1\}^{n>0}$ *fullfills* a schema H if and only if there is a match for each specification:

$$s_i = h_i \forall i \in \{j | h_j \neq \star\}. \tag{1}$$

If a string S fullfills the schema H, we write $S \in H$.

The number of specifications in a schema is the *order* of that schema

$$o(H) = |\{i | h_i \neq \star, h_i \in H\}|. \tag{2}$$

The *distance* between the first and the last specification in a schema is the *defining length* of that schema

$$\delta(H) = \max_{h_i \neq \star, h_i \in H} i - \min_{h_i \neq \star, h_i \in H} i \tag{3}$$

At a given iteration t of our genetic algorithm, the number of binary strings that fullfill the schema H is $r_{H,t}$.

The average observed fitness at the t-th iteration is

$$\bar{f}(t) = \frac{1}{m} \sum_{i=0}^{m} f(b_{i,t}) \tag{4}$$

where $b_{i,t}$ is the i-th individual in a population of $m > 0$ individuals (binary strings) at the t-th iteration.

The average observed fitness of the schema H at the t-th iteration is

$$\bar{f}(H, t) = \frac{1}{r_{H,t}} \sum_{i \in \{j | b_{j,t} \in H\}} f(b_{i,t}). \tag{5}$$

Teorema 1.1 (Holland) $\forall H \in \Sigma^{n>0}$, *given a genetic algorithm with:*

- *proportional selection*
- *one-point crossover with probability* p_c
- *random mutation with probability* p_m

it holds

$$\mathbb{E}[r_{H,t+1}] \geq r_{H,t} \cdot \frac{\bar{f}(H,t)}{\bar{f}(t)} \cdot \left(1 - p_c \frac{\delta(H)}{n-1}\right) \cdot (1-p_m)^{o(H)}. \tag{6}$$

2 Proof

Since the probability of an individual $b_{i,t}$ to be chosen for the $(t+1)$-th iteration is

$$\mathbb{P}[\text{choose } b_{i,t}] = \frac{f(b_i)}{\sum_{i=0}^{m} f(b_{i,t})} \tag{7}$$

the probability to choose an individual that fullfills the schema H is

$$\mathbb{P}[\text{choose a } b \in H] = \frac{\sum_{i \in \{j | b_{j,t} \in H\}} f(b_{i,t})}{\sum_{i=0}^{m} f(b_{i,t})} = p. \tag{8}$$

Since each individual is chosen independently, the expected number of individuals that fullfill the schema H at iteration $t+1$ is

$$\begin{aligned}
\mathbb{E}[r_{H,t+1}] &= m \cdot p & (9)\\
&= m \cdot \frac{\sum_{i \in \{j | b_{j,t} \in H\}} f(b_{i,t})}{\sum_{i=0}^{m} f(b_{i,t})} & (10)\\
&= m \cdot \frac{r_{H,t}}{r_{H,t}} \cdot \frac{\sum_{i \in \{j | b_{j,t} \in H\}} f(b_{i,t})}{\sum_{i=0}^{m} f(b_{i,t})} & (11)\\
&= r_{H,t} \cdot \frac{\frac{\sum_{i \in \{j | b_{j,t} \in H\}} f(b_{i,t})}{r_{H,t}}}{\frac{\sum_{i=0}^{m} f(b_{i,t})}{m}} = r_{H,t} \cdot \frac{\bar{f}(H,t)}{\bar{f}(t)}. & (12)
\end{aligned}$$

But it holds only if $p_c = 0$ and $p_m = 0$. In the case in which these probability are not null, it is necessary to take them into account.

2.1 Crossover

Property 2.1 *If two individuals fullfill the schema, also their parents fullfill the schema.*

Property 2.2 *If an indiviual fullfill the schema, and another individual does not, their children will not fullfill the same schema if and only if the cross site is between two specifications.*

Let show the last property through an example. Let $H = 01 \star 10 \star \star$ be the schema and b_i and b_j be two individuals, e.g.

$$b_i = 0111001 \in H$$
$$b_j = 1101101 \notin H.$$

If the cross site γ, i.e. the point in which the individuals are broken into two pieces in order to recombine their symbols, is before the last two symbols,

$$\text{cross}_\gamma(01110{\color{green}01}, 11011{\color{red}01}) = 01110{\color{red}01}, 11011{\color{green}01} \tag{13}$$

it is easy too see that both fullfill the schema H.

If the cross site $\bar{\gamma}$ is, for example, one position before the previous one, both the resulting individuals will not fullfill the schema H:

$$\text{cross}_{\bar{\gamma}}(0111{\color{green}001}, 1101{\color{red}101}) = 0111{\color{red}101}, 1101{\color{green}001} \tag{14}$$

Thus, since the cross site is after an index in $[1, n-1]$,

$$\mathbb{P}[\text{cross site is in the defining length}] = \frac{\delta(H)}{n-1} \tag{15}$$

and the probability to be sure that the cross site is good is $1 - \frac{\delta(H)}{n-1}$. Actually, also a cross site inside the defining length could provide a good cross, but there is not certainty. Finally, the probability that an individual that fullfills the schema leads another individual that fullfills the same schema is $p_s \geq 1 - p_c \frac{\delta(H)}{n-1}$.

Since selection and crossover are independent operations, now

$$\mathbb{E}[r_{H,t+1}] = r_{H,t} \cdot \frac{\bar{f}(H,t)}{\bar{f}(t)} \cdot p_s \tag{16}$$
$$\geq r_{H,t} \cdot \frac{\bar{f}(H,t)}{\bar{f}(t)} \cdot \left(1 - p_c \frac{\delta(H)}{n-1}\right). \tag{17}$$

2.2 Mutation

After the crossover, $r_{H,t+1}$ can only decrease if the mutation hits the wrong point. The mutation hit only one symbol and makes the logic not of the bit: $m(0) = 1$, $m(1) = 0$. The probability to choose a particular bit is $p_m = 1/n$ and each bit is independent, so the mutation is not bad if it does not hit a specification. Since the number of specification of the schema H is $o(H)$, the term $(1 - p_m)^{o(H)}$ is added to the formula, and the theorem follows.