

# Biologically Inspired Methods: Cheatsheet

Vishnu

Friday 26<sup>th</sup> April, 2019

# 1 Exploitation vs. Exploration

All guided searching algorithms try to strike a balance between these. *Exploitation* is taking known factors and using them to intensify their search: for example, on finding a good potential solution, search the area around it for better solutions. This allows us to follow trends in the search space, but often leads to being stuck in local minima.

*Exploration* involves diversifying the search by visiting previously unknown areas and trying different values to discover potentially better solutions. Doing this too much leads to an effectively random search, but in moderation this allows us to identify and break out of local minima.

## 2 Genetic Algorithms

Genetic Algorithms work by evolving a pool of candidate solutions, optimising them in the process till it eventually converges to an optimal enough solution. Optimising *usually* means minimising, unless otherwise specified, and this is the assumption throughout these notes. To convert from a maximisation to a minimisation problem, just multiply the cost function by -1. Each solution is represented as a **chromosome**, which contains values for the variables of the given problem (each variable is known as a **gene**). For example, if we're optimising a function of the form  $f(x, y)$ , then each chromosome contains two genes. **Binary Genetic Algorithms** represent chromosomes as a binary string, **Continuous Genetic Algorithms** represent them as a tuple (or list or array) of genes.

The general form of a genetic algorithm has 4 main parts:

1. Natural Selection
2. Pair Selection
3. Crossover
4. Mutation

Combining all of these together, the overall flow of the algorithm is:

- A Create a random initial population
- B While stopping criteria are unmet:
  - 1 Filter population with Natural Selection
  - 2 Create pairs via Pair Selection
  - 3 Make new chromosomes via Crossover
  - 4 Apply mutation to the population
- C Select the best performing chromosome as the solution

For problems that have constraints or multiple objectives to be optimised, having the cost function be a weighted sum of each individual requirement is sufficient. Note that all must be maximisation/minimisation: these can be converted with negative weights if required. Constraints should degrade the cost if not met (e.g add a large integer with absolute value higher than the lowest possible cost), and should be designed such that a solution violating a constraint cannot have a lower cost than a function that satisfies it. There is no such restriction for preferences.

## 2.1 Stopping Criteria

Any/all of the below can be used: it depends on the situation:

1. An acceptable (low enough cost) solution has been found
2. The mean/standard deviation/range of the chromosomes is within a target value
3. No change in chromosomes/best/cost/worst cost after X iterations
4. N iterations have elapsed

## 2.2 Binary Strings

For the **Binary Genetic Algorithm**, we also have to determine how to convert to and from binary strings. Using binary strings is mostly historical, back when the memory used by each chromosome was important. It's also fairly simple to implement.

### 2.2.1 Binary String Length

The overall length of the binary string is dependent on the number of possible values our gene can take. For example: for an equation  $f(x, y)$ , if  $x$  has 100 possible values and  $y$  has 30, the overall length of the binary string representation will be  $7 + 5 = 12$ .

Normally though, we're given a range and precision that our genes must adhere to. Rephrasing our earlier example:  $0 \leq x \leq 1$  with a precision of 2 decimal places,  $5 \leq y \leq 8$  with a precision of 1 decimal place. To convert from these to the above, we use the following equation:

$$Number\_of\_Values = \frac{x_{hi} - x_{lo}}{10^{-d}} \quad (1)$$

This equation is the same as saying  $integer\_range * 10^{decimal\_places}$ . Now that we have the number of possible values a gene can take, we need to find the minimum length of a binary string that can represent that many distinct values. A binary string of length  $n$  can represent  $2^n - 1$  distinct values.

For example: using the above equation, we see there are 30 distinct values for  $y$ .  $2^4 - 1$  is 15,  $2^5 - 1$  is 31. Therefore we choose a binary string of length 5 to represent  $y$ . We could have also calculated this as  $ceil(log_2(30))$ .

The above can be also be summarised as:

$$\frac{x_{hi} - x_{lo}}{10^{-d}} \leq 2^m - 1 \quad (2)$$

Where  $m$  is the required length. To get the full string, do this for each variable and add them together.

### 2.2.2 Binary Encoding and Decoding

Our binary representation roughly contains one unique string for each possible value. It's not going to be a perfect match though, as binary can't perfectly split decimal. To ease in the conversion, we calculate the **Binary Value**:

$$\frac{x_{hi} - x_{lo}}{2^m - 1}$$

This gives us the decimal value of a 1 in the binary string, and we'll represent it as  $b$ . Representing the binary string as  $BS$  and the decimal value as  $D$ , the conversions are:

$$Encoding : BS = binary(round(\frac{D - x_{lo}}{b})) \quad (3)$$

$$Decoding : D = x_{lo} + decimal(BS) * b \quad (4)$$

With these,  $x_{lo}$  is represented as  $0 \dots 0$  and  $x_{hi}$  as  $1 \dots 1$  for every gene. The complete binary string for a chromosome is just the concatenation of the strings for each gene.

### 2.2.3 Grey Codes

One problem with binary strings is adjacent numbers can have very different representations. For example: 3 and 4 have a difference of 1 in decimal, but their binary representations have every bit different (011 and 100). We can quantify this as the **Hamming Distance**, which is the number of bits that are different between two binary numbers (e.g 011 and 100 have a h.d of 3, 001 and 101 have a h.d of 1).

Grey codes are an alternate representation that guarantee consecutive numbers to have a Hamming Distance of 1. Representing the binary string as  $b$  and the grey code as  $g$ , both with length  $n$ , the conversions are as follows:

Binary to Grey:

$$g[i] = \begin{cases} b[i], & \text{if } i = n \\ b[i] \text{ XOR } g[i + 1], & \text{otherwise} \end{cases} \quad (5)$$

Grey to Binary:

$$b[i] = \begin{cases} g[i], & \text{if } i = n \\ b[i + 1] \text{ XOR } g[i], & \text{otherwise} \end{cases} \quad (6)$$

## 2.3 Natural Selection

This part of the algorithm determines which chromosomes are carried forward to the next generation.

### 2.3.1 $X_{rate}$

Only the best survive. The top (lowest cost)  $N_{keep}$  chromosomes are chosen, and the rest discarded. This method guarantees a fixed amount from each generation are kept, though this can be varied by changing  $X_{rate}$  (a user-defined percentage) at each iteration. With  $N_{pop}$  as the size of the population,  $N_{keep}$  can be calculated as:

$$N_{keep} = N_{pop} * X_{rate}$$

## 2.4 Thresholding

Only chromosomes with a score below a given threshold survive. This has the advantage of not having to sort and rank the chromosomes, but there is no guarantee on how many chromosomes survive. This can be fixed by changing the threshold at each iteration, either to an average or to a reasonable value.

## 2.5 Pair Selection

This part of the algorithm determines how chromosomes are paired up for crossovers (creating new chromosomes). This pairing up is important as it can influence how genes spread throughout the population, and can prevent good genes being overshadowed by poor ones. All of the below are independent of the cost function.i.e they don't require it to be of a certain form, such as differentiable.

### 2.5.1 Adjacency

Just pair adjacent chromosomes from top to bottom. If the population is sorted, this pairs chromosomes according to their ranks, which produces chromosomes without introducing much diversity (exploitation as opposed to exploration). It's very simple to implement, and every chromosome is used in crossovers.

### 2.5.2 Random

Randomly choose pairs. To generate the pair, a chromosome is chosen randomly, twice. This gives us unmatched diversity (exploration over exploitation), which gives a higher chance to find better quality offspring. Again very simple to implement, but for a completely random selection there's a chance we choose the same chromosome twice in a single pair.

### 2.5.3 Rank-Weighted Roulette

Choose pairs with probabilities according to their ranks. To generate the pair, two chromosomes are chosen according to the probabilities and paired together. If random can be taken as given each chromosome an equal chance to be chosen, Rank Weighted Pairing gives each chromosome a chance proportional to their rank in the population, which requires us to sort the population. This favours chromosomes with better costs. With a population of size  $N_{keep}$ , this probability for chromosome  $C_n$  of rank  $n$  is calculated as:

$$P(C_n) = \frac{N_{keep} + 1 - n}{\sum_{i=1}^{N_{keep}} i} \quad (7)$$

This gives probability  $1/sum(N_{keep})$  to the last chromosome and  $N_{keep}/sum(N_{keep})$  to the first. This gives a large difference in probability between adjacent chromosomes, regardless of their actual weights. For small populations, this gives a very skewed probability.

### 2.5.4 Cost-Weighted Roulette

Choose pairs with probabilities according to their weights. This doesn't need us to sort the population, but we do need to account that some costs may be negative. This is fixed by normalising the costs:

$$NCost(C_n) = Cost(C_n) - C_{N_{keep}+1} + 1 \quad (8)$$



$C_{N_{keep}+1}$  is the cost of the best chromosome not chosen. For  $X_{rate}$  selection this is obvious, for threshold selection or where  $N_{keep} = 0$ , instead use 2 \* the cost of the lowest chromosome. As this value is guaranteed to be  $\geq$  the lowest chromosome, subtracting it guarantees that all Ncosts are  $< 0$ . With the normalised costs, probability for each chromosome is calculated as:

$$P(C_n) = \frac{NCost(C_n)}{\sum_{i=1}^{N_{keep}} NCost(C_i)} \quad (9)$$

This gives a more accurate reflection of the relationships between the weights of the chromosomes, but is more computationally expensive.

## 2.6 Tournament Selection

A small subset of chromosomes is randomly chosen, and the best performing chromosome from this subset. This is done twice to get a pair, and the size of each subset is normally 2-3. A variation involves taking a larger subset, ranking them and then performing *rank-weighted roulette* to choose two chromosomes.

Both variations favour chromosomes with better costs. Also, since in each instance only a trivial subset has to be sorted, it reduces computation on sorting and thus works well for problems with large populations.

## 2.7 Crossover

Once the parents are selected, to generate offspring we take parts from each and combine them to make new chromosomes. Other than a few exceptions these don't generate new values: just re-arrange the existing ones. Repeatedly doing just crossovers only covers a limited section of the search space, so isn't sufficient to find an optimal solution. Crossover methods can be combined and altered: for example, a Single Point crossover followed by Blending, or only considering a subset of indices for Extrapolation.

Only the first three crossovers are applicable to the Binary Genetic Algorithm. Crossovers for ordering problems are detailed in a separate section.

### 2.7.1 Single Point

This breaks each parent into two parts, and creates offspring by swapping the second part for each. A random number  $0 \leq n < L$  is generated, and used to split the chromosomes.  $L$  is the length of the binary string / number of genes for binary and continuous genetic algorithms respectively. Using slice notation, this can be shown as:

$$\begin{aligned} Child_1 &= Parent_1[0 : n] + Parent_2[n : end] \\ Child_2 &= Parent_2[0 : n] + Parent_1[n : end] \end{aligned}$$

### 2.7.2 Double Point

This breaks each parent into 3 parts, and swaps the middle parts for each. This middle section is denoted by two crossover points, which are the start and end of the section respectively. These follow  $0 \leq n_1 < n_2 \leq L$ . Using slice notation, this can be shown as:

$$\begin{aligned} Child_1 &= Parent_1[0 : n_1] + Parent_2[n_1 : n_2] + Parent_1[n_2 : end] \\ Child_2 &= Parent_2[0 : n_1] + Parent_1[n_1 : n_2] + Parent_2[n_2 : end] \end{aligned}$$

### 2.7.3 Uniform Crossover

This uses a variable  $\mu \in (0, 1)$  to determine how much crossover occurs. For each index of the binary string/gene tuple, generate a random number  $n \in (0, 1)$ . If  $n < \mu$ , the bit/gene in that index is swapped in the children, otherwise are directly copied from the parent. The addition of  $\mu$  gives us a tuning parameter, but if it's too high or low effectively no crossover will happen (no swaps or all swaps).

### 2.7.4 Blending

This uses a variable  $\beta \in (0, 1)$  to combine the genes from the parents (P) into new values as below:

$$\begin{aligned} Child_1[i] &= \beta * P_1[i] + (1 - \beta) * P_2[i] \\ Child_2[i] &= \beta * P_2[i] + (1 - \beta) * P_1[i] \end{aligned}$$

These values are in the range  $(P_1[i], P_2[i])$ , so while these generate new values they are still limited to a subset of the search space.  $\beta$  can change from generation to generation, etc.

### 2.7.5 Extrapolation

This also uses a variable  $\beta$ , but there's no limit on its value. Children are generated using the below:

$$\begin{aligned} Diff_i &= abs(P_1[i] - P_2[i]) \\ Child_1[i] &= P_1[i] - \beta * Diff_i \\ Child_2[i] &= P_2[i] + \beta * Diff_i \end{aligned}$$

With some re-arrangement these are the same equations as Blending, but as there's no limit on  $\beta$  this can explore the entire search space without mutation.  $\beta$  can change from generation to generation, and if needed we can artificially limit the range of values generated.

## 2.8 Mutation

Mutation is the process that actually drives evolution: by changing the chromosomes into new values, new solutions can be evaluated and used. It is possible to have an algorithm that consists entirely of mutation: this would perform very similar to random walk or (1+1)ES. Crossovers move the genes into potentially better solutions, mutations discover new solutions.

### 2.8.1 How many Mutations?

Mutation is determined using a mutation rate,  $\mu$ . To prevent good solutions from being mutated, we can designate the top  $n$  chromosomes as *elite*, and not consider them for mutation.  $n$  can be left as 0, in which case any chromosome can be mutated.

With  $N_{pop}$  as the size of the population, the standard formula is:

$$M = (N_{pop} - n) * \mu * N_{bits} \quad (10)$$

$N_{bits}$  is the length of the binary string for a chromosome for BGA, and the number of genes in a chromosome for CGA.

### 2.8.2 Binary Mutation

Disregarding the top  $n$  chromosomes, choose  $M$  bits randomly across the population and flip them ( $0 \rightarrow 1$ ,  $1 \rightarrow 0$ ).

### 2.8.3 Numeric Mutation

This uses a variable  $\sigma$  to control how much a gene mutates. Disregarding the top  $n$  chromosomes, choose  $M$  genes randomly and apply:

$$Gene = Gene + \sigma * N_n(0, 1) \quad (11)$$

$\sigma$  is usually a constant, but for a dynamic system can be changed between generations.  $N_n$  is a random number generator, but generates numbers according to a standard distribution (i.e 0.5 is most likely to be generated, 0 and 1 are very unlikely).

## 2.9 Permutation Problems

Certain problems (such as the **Travelling Salesman Problem**) require their solution to be an ordering of a given set of variables. This means traditional mutation and crossover methods won't work, since they change the values and discover new ones respectively. Chromosomes take the form of a tuple of values, similar to CGA, but every chromosome has the same values in a different order.

### 2.9.1 Partially Matched Crossover

Perform two-point crossover, then swap duplicated genes *outside* of the two points. This can be done by going through the children, and swapping the first duplicate in Child<sub>1</sub> with the first in Child<sub>2</sub> and so on .e.g.

$$\begin{aligned} \textit{Parent1} &: (3, 4, 6, 2, 1, 5) ; \textit{Parent2} : (4, 1, 5, 3, 2, 6) \\ \textit{ChosenPoints} &: 1, 3 \\ \textit{InitialChildren} &: (3|1, 5|2, 1, 5); (4|4, 6|4, 2, 6) \\ \textit{FixedChildren} &: (3, 1, 5, 2, 4, 6); (1, 4, 6, 3, 2, 5) \end{aligned}$$

This doesn't preserve the existing ordering in the slightest: so two very good orderings could potentially produce two terrible ones.

### 2.9.2 Ordered Crossover

Similar to two-point crossover, choose two points and add the middle section to the children. Then fill in the empty spaces with the genes from the parent starting at the first crossover point, skipping the duplicates. e.g.

$$\begin{aligned} \textit{Initial : Parents} &- (3|46|215), (4|15|326) \\ \textit{Phase1 : Children} &- (x|15|xxx), (x|46|xxx) \\ \textit{Phase2 : Children} &- (4|15|623), (1|46|532) \end{aligned}$$

For Child<sub>1</sub>, the x's are filled in with 4623 (starting at crossover point 1, skipping duplicates, looping around at the end). For Child 2, 1532 is taken from the parent. An alternative is to delete the duplicates from the parents then directly fill in the children, which saves skipping the duplicates.

This method preserves the relative ordering, but not the absolute. For TSP this is fine since the start is irrelevant, and the sequence can just be rotated to get the desired start point.

### 2.9.3 Cycle Crossover

Swap the left-most genes, then keep swapping duplicates from left to right in child<sub>1</sub> till there are none left. This method has a relatively simple implementation, as only child<sub>1</sub> is checked for duplicates and the checking index only advances from left to right (looping around if necessary).e.g.

*Initial : Parents* – (|346215), (|415326)  
*Swap1 : Children* – (4|46215), (3|15326)  
*Swap2 : Children* – (41|6215), (34|5326)  
*Swap3 : Children* – (41622|5), (34531|6)  
*Swap4 : Children* – (4163|25), (3452|16)

### 2.9.4 Coding for Crossovers

Rather than a direct crossover method, the chromosomes are encoded to a numeric string, can then have any crossover method applied to them, then decoded back to a chromosome.

Encoding:

- A Make a reference list of the genes
- B For idx in range(0,len(chromosome)):
  - 1 coded[idx] = index of chromosome[idx] in reference list
  - 2 remove chromosome[idx] from the reference list

e.g.

Epoch	Coded	Chromosome	Reference List
1		346215	1,2,3,4,5,6
2	3	46215	1,2,4,5,6
3	33	6215	1,2,5,6
4	334	215	1,2,5
5	3342	15	1,5
6	33421	5	5
7	334211		

Decoding:

- A Make a reference list of the genes
- B For idx in range(0,len(coded)):
  - 1 chromosome[idx] = index of coded[idx] in reference list
  - 2 remove coded[idx] from the reference list

e.g.

Epoch	Coded	Chromosome	Reference List
1	351311		(,2,3,4,5,6
2	51311	3	1,2,4,5,6
3	1311	3,6	1,2,4,5
4	311	3,6,1	2,4,5
5	11	3,6,1,5	2,4
6	1	3,6,1,5,2	4
7		3,6,1,2,5,4	

### 2.9.5 Mutation

As generating new values breaks the problem, the following methods can be used:

1. Invert/Reverse a randomly selected subset of the chromosome
2. Move/Swap a randomly selected subset within the chromosome
3. Randomly swap positions: this may break an encoded chromosome, so it must be decoded first

## 2.10 Genetic Programming

This application of GA allows it to build a program for us. It is provided an input and a desired output, and the algorithm builds us a program to convert between two.

Rather than strings or tuples, chromosomes are represented as trees of varying sizes and shapes. This requires extra overhead for each chromosome, since the size and shape has to be tracked as well. Leaf nodes are variables and constants, while internal tree nodes are functions that combine these.

The algorithm also has to be provided a set of possible values for variables/constants, a set of functions, and semantics on how to combine the two. e.g. the SUM function requires two numerics and outputs a numeric, the OR function requires two binary strings and outputs a binary string, etc.

### 2.10.1 Crossovers

The simplest crossover is to swap and replace subtrees between chromosomes. A single child can be generated by replacing a subtree in one parent by one from the other parent, and two children can be obtained by swapping subtrees.

### 2.10.2 Mutation

**Node Mutation** Swap a node with an appropriate alternative  
e.g replace SUM with DIFF, replace 4 with 11

**Swap Mutation** Swap the order of terminal nodes provided to a function  
e.g a DIFF b  $\rightarrow$  b DIFF a

**Grow Mutation** Replace a terminal node with a new subtree that provides the same type as the replaced node e.g. replace 5 with a SUM 7

**Gaussian Mutation** Add a value determined by a gaussian ( $N_n * \sigma$ ) to a terminal node

**Trunc Mutation** Replace a subtree with an appropriate terminal node  
e.g replace 5 MOD 2 with 11



## 2.11 Schema Theorem

This theorem is used to demonstrate that GAs are effective in discovering good solutions, and increase the quality of the solutions with more generations. The proof uses the binary genetic algorithm, but the result proves effectiveness for all GAs.

### 2.11.1 What is a Schema

A schema (or scheme) is a template that represents binary strings, similar to how regular expressions represent regular strings. They consist of '0', '1' and '\*' (wildcard).

The order of a schema is the number of fixed positions it contains ( $O(s)$ ), and the defining length ( $\delta(s)$ ) is the distance between the first and last fixed positions. e.g.

\* \* 00 \* \* 0 \* \*\*

Order : 3

Defining Length :  $7 - 3 = 4$

Examples of matching Strings : 1100010101, 0000000000, 0100100111

For this proof, we define the average cost of a schema  $s$  at generation  $t$  ( $f(s,t)$ ) as the average cost of all the chromosomes that match the schema at that generation, and  $n(s,t)$  is the number of chromosomes at  $t$  that match  $s$ .

### 2.11.2 Schema Pair Selection

Using cost-weighted roulette, the probability that a chromosome  $C$  is chosen is  $\frac{Cost(C)}{\sum_{i=1}^{N_{keep}} Cost(C_i)}$ .

Therefore, for a chromosome matching a schema at the generation, the average probability of being chosen is  $\frac{f(s,t)}{\sum_{i=1}^{N_{keep}} Cost(C_i)}$ . The probability for **any** chromosome

matching the schema being chosen is  $\left( n(s,t) * \frac{f(s,t)}{\sum_{i=1}^{N_{keep}} Cost(C_i)} \right)$ .

This selection happens  $N_{pop}$  times. If we assume that only selection takes place and no crossover/mutation occurs, we obtain:

$$n(s, t + 1) = n(s, t) * \frac{f(s, t)}{\sum_{i=1}^{N_{keep}} Cost(C_i)} * N_{pop} = n(s, t) * \frac{f(s, t)}{\text{Avg. Cost at gen } t} \quad (12)$$

With some re-arrangement, we can re-write these as the following:

$$n(s, t + 1) = n(s, t) * (1 + \varepsilon(s, t)) \quad (13)$$

$$= n(s, t - 1) * (1 + \varepsilon(s, t - 1)) * (1 + \varepsilon(t))$$

$$= n(s, 1) * \prod_{i=1}^t (1 + \varepsilon(s, t)) \quad (14)$$

$$\varepsilon(s, t) = \frac{f(s, t) - \text{Avg.cost}(t)}{\text{Avg.cost}(t)} \quad (15)$$

### 2.11.3 Schema Crossover

After being selected as a parent, a schema is said to survive a crossover if one of the child chromosomes also matches it. The only circumstance a schema breaks in single-point crossover is if the break point is inside the defined length of the schema: as the externals are wild cards it doesn't matter. Using this, and the knowledge that the last index can't be selected for a single-point crossover, the probability that a schema survives a single-point crossover is:

$$P(survival) = 1 - \frac{\delta(s)}{len(s) - 1} \quad (16)$$

### 2.11.4 Schema Mutation

Schemas are resilient to mutation as wildcards match both 0 and 1, so they only break if one of the fixed positions is mutated. If  $P_m$  is the probability for a bit to be mutated, the probability a schema survives mutation is:

$$P(survival) = (1 - P_m)^{O(s)} \quad (17)$$

### 2.11.5 Conclusion

Combining the above three sections, we obtain the schema growth equation:

$$P(survival) = n(s, t) * (1 + \varepsilon(s, t)) * \left(1 - \frac{\delta(s)}{len(s) - 1}\right) * (1 - P_m)^{O(s)} \quad (18)$$

$$= n(s, t) * \frac{\text{Avg. Cost of Matches}}{\text{Overall Avg. Cost}} * \left(1 - \frac{\text{DefiningLength}}{\text{SchemaLength}}\right) * (1 - P_m)^{Order(s)} \quad (19)$$

Therefore, we can say that above-average, short, low-order schema propagate well. The more above average the better, so with enough generations eventually the weaker solutions disappear and the more optimal ones remain.

### 3 Evolution Strategies

Evolution Strategies are similar to Genetic Algorithms, with two main differences:

1. ES can only handle real values
2. ES uses a set of values, called **Strategy Parameters** to handle mutations for each gene

The overall flow of the algorithm is:

- A Create a random initial population and strategy parameters
- B While stopping criteria are unmet:
  - 1 Create pairs via Pair Selection
  - 2 Make new chromosomes via Crossover
  - 3 Apply Mutation to the children and parameters
  - 4 Select new population
- C Select the best performing chromosome as the solution

#### 3.1 Strategy Parameters

Strategy Parameters ( $\sigma$ ) for ES are the same as the step size in gradient descent. One is created for each variable in the optimisation problem: so there will be an equal number of parameters and genes.

A larger  $\sigma$  will move further in the current population direction, exploiting a good trend, and a smaller  $\sigma$  will search locally to find a good new direction to move the population. The  $\sigma$  are mutated alongside the population, and the mutations apply even if the offspring are rejected.

#### 3.2 ES Notation

Evolution strategies are depicted with the  $(\mu, \lambda)$  or  $(\mu + \lambda)$  notation.  $\mu$  represents the size of the population at each generation, and  $\lambda$  is the number of children generated at each stage.

The  $,$  and  $+$  represent two methods of selection. With  $,$  the new population is selected only from the children, and with  $+$  the new population is selected from a combination of the parents and children. e.g.  $(1,1)$ ES will always choose the child (effectively random walk), and  $(1+1)$ ES will choose the better of the child and parent (smarter random walk).

For  $,$  notation,  $\lambda$  must be  $\geq$  than  $\mu$ : otherwise the population requirement won't be hit.  $(\mu, \lambda)$  results in faster convergence,  $(\mu + \lambda)$  gives more exploration.

### 3.3 Crossovers

Some variations don't use crossovers (the children are just the parents), but otherwise there are four main types of crossovers. They can be distinguished in the following categories:

- Source:

**Local** 2 parents, 1 child

**Global** >2 parents, 1 child

- Output:

**Discrete** Children contain elements from the parents

**Intermediate** Children contain elements derived from the parents, such as a weighted average.

To generate more children, the order of the parents is changed. Examples are shown below:

	Discrete	Intermediate
Local	Uniform Crossover	Blending
Global	Non-Uniform Crossover	Uniform Blending

#### 3.3.1 Non-Uniform Crossover

Rather than using a single  $\mu$ , a rank-based cost roulette is used amongst the parents. Otherwise this is the same as uniform crossover, selecting each index of the children from a parent. The top-ranked parent ( $P_1$ ) has  $P = 0.5$ , and the others are generated according to:

$$P(P_2) > P(P_3) > \dots > P(P_n)$$

#### 3.3.2 Uniform Blending

$$Child[i] = \frac{1}{\text{Num. of Parents}} * \sum_{k=1}^n parent_k[i]$$

### 3.4 Mutation

#### 3.4.1 Chromosomes

For a chromosome  $c$ , the mutation of each gene  $c[i]$  depends on the strategy parameter  $\sigma_i$ . This is similar to CGA, but as  $\sigma_i$  adjusts as well it allows for dynamic growth, allowing the algorithm to choose between exploring and exploiting. As before,  $N_n$  gives a random number according to a gaussian distribution.

$$c[i] = c[i] + \sigma_i * N_n(0, 1) \quad (20)$$

### 3.4.2 Strategy Parameters

These change according to how well the population is doing: if the quality of the solutions is steadily improving then the algorithm's speed is increased, and if solutions aren't improving then the speed is decreased in favour of local search. Determining an improvement could be checking if the average cost lowers, or if more children are chosen than parents.

$$\sigma_i(t+1) = \begin{cases} \sigma_i(t) * e^{\frac{1}{3}}, & \text{if } t=0 \text{ or previous mutation was successful} \\ \sigma_i(t) * e^{\frac{-1}{12}}, & \text{otherwise} \end{cases} \quad (21)$$

## 3.5 (1+1)ES

Effectively a smarter random walk, this chooses the better of the parent and child at each generation. No crossover occurs, and the offspring is simply a copy of the parent with mutation.

### 3.5.1 Parameter Mutation Variants

1. No mutation: this is similar to a GA
2. If, in the last  $X$  generations, the good mutation rate  $\geq 20\%$ , then increase  $\sigma$ . Else, decrease it.
3. With  $n_x$  being the number of genes,  $n_m$  being the number of good mutations in the last  $10n_x$  generations and  $0 < \alpha < 1$ , perform the following every  $10n_x$  generations:

$$\sigma_i(t+1) = \begin{cases} \sigma_i(t) * \alpha, & n_m < 2n_x \\ \sigma_i(t)/\alpha, & n_m > 2n_x \\ \sigma_i(t), & n_m = 2n_x \end{cases}$$

## 3.6 ( $\mu+1$ )ES

The population only potentially changes by 1 individual each time. Since there are multiple chromosomes, the offspring is obtained by crossover and thus there is a diversity in output.

## 4 Differential Evolution

Differential Evolution is similar to Evolution Strategies, but uses a guided form of mutation rather than adding a randomised value. It does this by generating difference vectors, and using these to mutate the generated offspring.

Unlike previous algorithms, this allows chromosomes to work together to achieve a solution: previously each chromosome was independent, now their mutation is inter-dependent. Another crucial difference is that each chromosome is now regarded as a point/particle in an n-dimensional space (n being the number of genes). This has no effect on the structure of the chromosome (it's still a tuple of values), but allows us to use vector arithmetic on it (any point in space is a vector from the origin to that point).

The overall flow of the algorithm is:

- A Create a random initial population and control parameters
- B While stopping criteria are unmet:
  - 1 Apply mutation to each parent to generate a trial vector
  - 2 Make new chromosomes via Crossover with parent+trial vector
  - 3 Select new population from parents and offspring
- C Select the best performing chromosome as the solution

The two control parameters are the scale factor ( $\beta$ ) and the crossover parameter ( $P_r$ ). These are detailed in the following sections. DE hybrids are discussed in a later chapter.

### 4.1 Mutation

A mutated vector, known as a **trial vector**, is generated for each parent. The basic skeleton is:

$$U_i(t) = X_{i_1}(t) + \beta * (X_{i_2}(t) - X_{i_3}(t)) \quad (22)$$

Trial Vector = Target Vector + Scale Factor \* Difference Vector

Where i is the index of the parent vector, and  $i_1$ ,  $i_2$  and  $i_3$  are the indices randomly selected vectors from the population  $\neq i$ .  $\beta$  can take any positive value, but is usually in the range (0.5,2).

This mutation automatically scales between exploration and exploitation, as the difference vector shrinks as the particles get closer together. The scale factor helps with this, but early on in the search the difference vectors will be large and may dramatically spread the particles out. Another disadvantage is that the trial vector has nothing to do with the parent, though it is combined with the parent to make the offspring.

## 4.2 Crossover

Once the trial vector is generated, it is mated with the parent vector to produce an offspring. Though standard methods like Uniform Crossover, Blending or Double Point Crossover can be used, the usual method for DE is:

$$X'_{ij}(t) = \begin{cases} X_{ij}(t) & , \text{ if } j \in J \\ U_{ij}(t) & , \text{ otherwise} \end{cases} \quad (23)$$

Where  $X_{ij}(t)$  is the  $j^{th}$  gene of the parent, and  $U_{ij}(t)$  is the  $j^{th}$  gene of the trial vector.  $J$  is a list of indices, and generation methods are shown below (most of which use  $P_r$ ). This method limits the overall amount of change, as opposed to uniform crossover, and blends the entire vector together, as opposed to clearly defined subsections being from each vector. Later in the search we can expect most genes to be roughly correct, and  $P_r$  gives us a tuning parameter to reduce the amount of trial vector influence.

### 4.2.1 Random Generation

$J$  is a random list of indices  $\in (0, n_x)$ , where  $n_x$  is the number of genes. In extreme cases this may include every/no indices, and there's no tuning parameter.

### 4.2.2 Binomial Crossover (bin)

This approach is similar to uniform crossover, but uses  $P_r$  as opposed to  $\mu$ . This approach guarantees at least one value in  $J$  (so some mutation), and  $P_r$  gives us a tuning parameter: a high  $P_r$  means mostly the offspring is mostly from the parent, a low  $P_r$  means mostly from the trial vector.

1. Randomly add a value  $x \in (0, n_x)$  to  $J$
2. for  $i$  in range( $0, n_x$ ):
  - (a) if (  $i \neq x$  &&  $\text{rand}() < P_r$  )
  - Add  $i$  to  $J$

### 4.2.3 Exponential Crossover (exp)

This provides an adjacent sequence of indices (similar to Double Point Crossover), but this range is generated using  $P_r$ .

1. Randomly add a value  $i \in (0, n_x - 1)$  to  $J$
2. Do
  - (a) Add  $i+1$  to  $J$
  - (b)  $i = (i+1) \% n_x$
 while (  $\text{rand}() < P_r$  and  $\text{len}(J) < n_x$  )

This is exponential as a sequence of length 2 has probability  $(P_r)^0$ , of length 3 has  $(P_r)^1$ , and length  $n$   $(P_r)^{n-2}$ . A high  $P_r$  gives us more influence from the parent vector, a low  $P_r$  more influence from the trial vector.

### 4.3 DE Notation

Differential Evolution methods are denoted as DE x/y/z, where x is the method of selecting the target vector in mutation, y is the number of difference vectors and z is the crossover method used.

#### 4.3.1 Target Vector Selection

It was previously mentioned that the target vector for mutation was randomly selected (rand), but this isn't fixed. Another popular method is selecting the best vector ( $\hat{X}$ ) from the population (best).

Rand-to-Best sets the target vector as  $\alpha * \hat{X} + (1 - \alpha) * X_{i_1}(t)$  :  $\alpha$  then allows us to tune between exploration (the random vector) and exploitation (the best vector).  $\alpha$  can be set to dynamically increase from 0 to 1 as generations increase.

Lastly, current-to-best sets it as  $\beta * (\hat{X} - X_{i_p}(t))$ , where  $X_{i_p}(t)$  is the parent vector. This allows the parent to influence the trial vector, and minimises the divergence.

#### 4.3.2 Difference Vectors

The basic formula has a single difference vector, but any number can be chosen, denoted as  $n_v$ . Using the current-to-best target vector adds another difference, so this is denoted as  $1 + n_v$ .

#### 4.3.3 Examples

1. DE rand-to-best/2/exp

$$\textbf{Trial Vector Equation } \alpha * \hat{X} + (1 - \alpha) * X_{i_1}(t) + [\beta * (X_{i_2}(t) - X_{i_3}(t)) + \beta * (X_{i_4}(t) - X_{i_5}(t))]$$

2. DE current-to-best/1+2/bin

$$\textbf{Trial Vector Equation } \beta * (\hat{X} - X_{i_p}(t)) + [\beta * (X_{i_2}(t) - X_{i_3}(t)) + \beta * (X_{i_4}(t) - X_{i_5}(t))]$$

### 4.4 Switching DE

Studies have shown that DE rand/1/bin is good for exploring and DE current-to-best/2/bin is good for exploiting. We could swap between them after a fixed number of generations, but the usual method uses probabilities  $P_1$  and  $P_2 (= 1 - P_1)$  to decide which method to use at a generation.

Also defined are  $n_1$  and  $n_2$  (the total number of offspring used after methods 1 and 2 respectively) and  $f_1$  and  $f_2$  (the total number of offspring discarded after methods 1 and 2 respectively).

The first X (normally 50) generations are set as a learning period, with  $P_1 = P_2 = 0.5$  and recording values for  $n_1, n_2, f_1$  and  $f_2$ . After generation X, the following equations are used:



$$\begin{aligned}
P_1 &= \frac{n_1 * (n_2 + f_2)}{n_1 * (n_2 + f_2) + n_2 * (n_1 + f_1)} \\
P_2 &= 1 - P_1
\end{aligned} \tag{24}$$

With some re-arrangement, this is simply  $P_1 = \frac{\text{Method 1 success \%}}{\text{Method 1 success \%} + \text{Method 2 success \%}}$

## 5 Swarm Optimisation

Swarm optimisation/intelligence takes DE a step further: rather than a group of solutions deriving from each other, a group of simple agents collectively find a single optimal solution. A highlight of this method is there is no central control or global model: each agent is a simple individual, and their local interactions cause global patterns to emerge. Some further characteristics are:

**Flexible** it can handle internal/external problems and unexpected events

**Resilient** even if some agents fail to complete, the overall system will continue

**Self-Organised** Paths to solutions aren't predefined: the system discovers the correct method

### 5.1 Self-Organisation

By using simple individual mechanisms, the interactions of the agents form structures on the global level of the system without external planning or control.

1. Positive Feedback: good solutions are highlighted and reinforced
2. Negative Feedback: introduce a time scale to the algorithm by reducing the effects of older feedback, prevents early/premature convergence on local optima
3. Exploration and diversity are amplified, by moving randomly then falling into the positive feedback loops
4. Agents communicate with each other, directly or indirectly

### 5.2 Stigmergy

Stigmergy is a method of indirect agent communication by modifying the environment the agents are in. These may trigger certain actions from the receiving agents, or simply be positive/negative feedback. There are two main types:

**Sematectonic** Communicate via changing the physical environment (the current state of the solution).e.g. well-travelled routes show more wear and tear

**Sign-Based** Communicate via signalling mechanisms.e.g laying a trail of breadcrumbs on the travelled route

### 5.3 Binary Bridge Experiment

An example of the above is the binary bridge experiment: there are two paths from the start to the goal, one longer than the other. The agents in this case are ants, which leave a pheromone behind them as they travel (Sign-Based Stigmergy). This pheromone attracts other ants, which increases the probability that future ants pick that path.

The ants reach the goal then return to the source via the same path. As the shorter path will have ants completing their journey faster, more pheromones will be deposited, which

attracts more ants until eventually every agent takes the shorter path. The pheromones dissipate after a while (negative feedback), so the longer path will slowly lose attractiveness while the shorter is constantly reinforced.

There is no guarantee that the ants will always pick the path with more pheromones though: a small random chance is added to pick a previously unexplored path, to prevent a scenario where the initial ants only traverse the longer path, and all ants only pick that path.

## 6 Ant Colony Optimisation

An type of swarm optimisation where the agents mimic ants, using sign-based stigmergy. Though the base form solves shortest path and TSP problems, a large number of problems can be reduced to these forms.

Problems are given in a graph structure, with each edge (i,j) having a cost/length and pheromone concentration ( $\tau_{ij}(t)$ ) associated with it. Also defined are  $L^k(t)$ , which is the length of the path taken by ant k at generation t, and  $N_k^i(t)$ , the list of nodes reachable from node i for ant k at gen t. The latter two will become more clear in the next section. The overall flow of the algorithm is:

- A Create random initial pheromone concentrations ( $\tau_{ij}(0)$ )
- B While stopping criteria are unmet:
  - 1 Create and evaluate a path for each ant
  - 2 Evaporate the Pheromones
  - 3 Apply feedback to the Pheromones
- C Select the path with the most ants/ the most optimal path

### 6.1 Creating an Ant Path

At each generation/iteration, a path is created for each ant according to the transition probabilities. These probabilities vary for each method (and are detailed later), but the overall method for shortest path problems is:

- I  $X^k(t) = \{\text{start\_node}\}$ ,  $i = \text{start\_node}$
- 1. While  $i \neq \text{goal\_node}$ :
  - (a) If  $N_k^i(t)$  is empty, append the predecessor node to i
  - (b) Select a node j from  $N_k^i(t)$  according to the transition probabilities
  - (c) Append (i,j) to  $X^k(t)$ ,  $i = j$
- 2. Remove all loops from  $X^k(t)$
- 3. Calculate  $L^k(t)$

Removing the loops from  $X^k(t)$  affects  $L^k(t)$ , so care must be taken: for example, there may be two conflicting loops of different lengths, so depending on which one is removed this changes the attractiveness of the path.

## 6.2 Simple Ant Colony Optimisation

### 6.2.1 Transition Probability

SACO just uses a pheromone-weighted roulette for its transition probabilities: these are evaluated at each node while creating a path. The probability that ant  $k$  chooses edge  $(i,j)$  from node  $i$  at generation  $t$  is:

$$P_{ij}^k(t) = \begin{cases} \frac{\tau_{ij}(t)}{\sum_{ab \in N_k^i(t)} \tau_{ab}(t)} & , if j \in N_k^i(t) \\ 0 & , otherwise \end{cases} \quad (25)$$

### 6.2.2 Pheromone Evaporation

This uses a variable,  $\rho \in (0, 1)$ , to determine how much the pheromones evaporate at each stage. The evaporation removes the effects of older pheromones on current probabilities, so for a path to be consistently picked the pheromones on it need to be reinforced.

$$\tau_{ij}(t) = (1 - \rho) * \tau_{ij}(t) \quad (26)$$

### 6.2.3 Pheromone Updates

SACO uses a simple update system: pheromones for an edge are increased based on the number of ants that included that edge in their path (path for ant  $k$  is  $X^k(t)$ ), and this increase is based on the quality of the ants' solutions. It uses a constant,  $Q$ , to moderate the increase in the pheromones.

$$\tau_{ij}(t+1) = \tau_{ij}(t) + \sum_{k \in ants} \Delta \tau_{ij}^k(t) \quad (27)$$

$$\Delta \tau_{ij}^k(t) = \begin{cases} \frac{Q}{L^k(t)} & , if (i,j) \in X^k(t) \\ 0 & , otherwise \end{cases} \quad (28)$$

## 6.3 Ant System

AS builds off SACO by adding more heuristics for the transition probability and implementing elitism. Most versions also pre-process a list of connections from each node rather than calculating the probability for every edge at every node (even the obviously unreachable ones).

### 6.3.1 Transition Probability

As the system pre-processes  $N_k^i(t)$ , this doesn't need to be checked at each stage. Instead it uses the inverse of the cost ( $\eta_{ij}(t) = 1/\text{cost}(i,j)$ ) and tuning parameters  $\alpha$  and  $\beta$ . There are two possible methods, both of which are weighted roulettes using  $\tau_{ij}(t)$  and  $\eta_{ij}(t)$ :

#### Exponent-Based

$$P_{ij}^k(t) = \frac{\tau_{ij}(t)^\alpha * \eta_{ij}(t)^\beta}{\sum_{ab \in N_k^i(t)} \tau_{ab}(t)^\alpha * \eta_{ab}(t)^\beta} \quad (29)$$

#### Fraction-Based

$$P_{ij}^k(t) = \frac{\alpha * \tau_{ij}(t) + (1 - \alpha) * \eta_{ij}(t)}{\sum_{ab \in N_k^i(t)} \alpha * \tau_{ab}(t) + (1 - \alpha) * \eta_{ab}(t)} \quad (30)$$

### 6.3.2 Pheromone Updates

AS keeps the same overall equation, but uses a different calculation for  $\Delta\tau_{ij}^k(t)$ . Again there are three different calculations, each gives slightly different results:

#### Regular (Ant-Cycle)

$$\Delta\tau_{ij}^k(t) = \begin{cases} \frac{Q}{L^k(t)} & , if(i, j) \in X^k(t) \\ 0 & , otherwise \end{cases} \quad (31)$$

#### Density-Based

$$\Delta\tau_{ij}^k(t) = \begin{cases} Q & , if(i, j) \in X^k(t) \\ 0 & , otherwise \end{cases} \quad (32)$$

#### Quantity-Based

$$\Delta\tau_{ij}^k(t) = \begin{cases} \frac{Q}{\text{cost}(i,j)} & , if(i, j) \in X^k(t) \\ 0 & , otherwise \end{cases} \quad (33)$$

### 6.3.3 Elite Pheromone Updates

On top of one of the above updates, AS also gives a special bonus to edges that were on the best path ( $X^e(t)$ ). This is achieved by adding  $n_e * \Delta_{ij}^e(t)$  to all paths, where  $n_e$  is the number of ants that took the best path.

$$\Delta_{ij}^e(t) = \begin{cases} \frac{Q}{L^{X^e(t)}} & , if(i, j) \in X^e(t) \\ 0 & , otherwise \end{cases} \quad (34)$$

## 6.4 Ant Colony System

ACS builds even further of AS, providing more tuning parameters and increasing the system's exploration. It also slightly changes the functions of Pheromone Evaporation and Pheromone Updates: they become global and local updates respectively, where global affects all pheromones and local only those on the best path.

### 6.4.1 Transition Probability

It uses a user-provided variable  $r_0 \in (0, 1)$  to balance between exploration and exploitation (higher  $r_0 \rightarrow$  more exploitation). From an ant  $k$  at node  $i$ , the edge taken is decided by:

$$\text{Next edge for ant } k = \begin{cases} \text{Lowest cost edge} \in N_k^i(t) & , \text{if } \text{rand}(0,1) < r_0 \\ \text{Use Exponent-based from Ant System} & , \text{otherwise} \end{cases} \quad (35)$$

### 6.4.2 Pheromone Evaporation (Local Update)

Unlike previous systems, this both evaporates and adds to the pheromone concentration. This local addition gives the system more chance to explore since all edges get a boost, but this addition remains constant so its effects is reduced in later generations when the system is closer to finding an optimal path. It uses the user-provided constants  $\rho_L \in (0, 1)$  and  $\tau_o > 0$ . If either were 0 then it would be regular pheromone evaporation, and if  $\rho_L$  was 1 then the pheromone intensities would be reset at every generation.

$$\tau_{ij}(t) = (1 - \rho_L) * \tau_{ij}(t) + \rho_L * \tau_o \quad (36)$$

### 6.4.3 Pheromone Updates (Global Update)

The global updates gives the edges on the global best path an extra bonus, similar to the elite pheromone updates from AS, and thus pushes the system towards exploitation. It also adds an extra layer of evaporation, which reduces the universal bonus given in the previous step. The global best path ( $X^+(t)$ ) can either be the best in this generation ( $\tilde{x}(t)$ ) or the best from the first to the current generation ( $\hat{x}(t)$ ). This update also uses a user-provided constant  $\rho_G \in (0, 1)$ , which again isn't 0 or 1 for the reasons mentioned above.

$$\tau_{ij}(t+1) = (1 - \rho_G) * \tau_{ij}(t) + \rho_G * \Delta\tau_{ij}(t) \quad (37)$$

$$\Delta\tau_{ij}(t) = \begin{cases} \frac{1}{L^{X^+(t)}} & , \text{if } (i, j) \in X^+(t) \\ 0 & , \text{otherwise} \end{cases} \quad (38)$$