18: Linear Genetic Programming 2

- Fitness, initialization, mutation, and crossover
- Evolution of programs
- Reference book chapters 3, 5
- Reference book: *Linear Genetic Programming, Brameier and Banzhaf, Springer*

An example linear genetic program

r_1, r_2 take input variables x_1, x_2

r_0, r_3 are calculation registers, and r_0 is the output register, initialized as 1

1: $r_0 = r_2 + 5$

2: $r_3 = r_1 \times 3$

3: if $r_0 > r_3$:

3: $r_3 = r_3 - 1$

4: $r_0 = r_3 \times r_0$

LGP fundamentals - Fitness

- Error-based functions on training samples
  - sum of absolute errors (SE)
  - sum of squared errors (SSE)
  - mean squared errors (MSE):

$$\mathrm{MSE}(gp) = \frac{1}{n}\sum_{k=1}^{n}(gp(\vec{i_k}) - o_k)^2$$

  - classification errors (CE):

$$\mathrm{CE}(gp) = \sum_{k=1}^{n}\{1 \mid class(gp(\vec{i_k})) \neq o_k\}$$

- Data partition: training, validation, testing

  Stochastic sampling of the training dataset

LGP fundamentals - Initialization

- Random generation of programs
  - Program size upper bound
  - Program size lower bound (minimum length - one instruction)
  - Initial program length uniformly random within the lower/upper bounds

- How to choose program size upper bound
  - Too long - reduce program variability during evolution
  - Too short - insufficient diversity of initial genetic material

LGP fundamentals - Variation

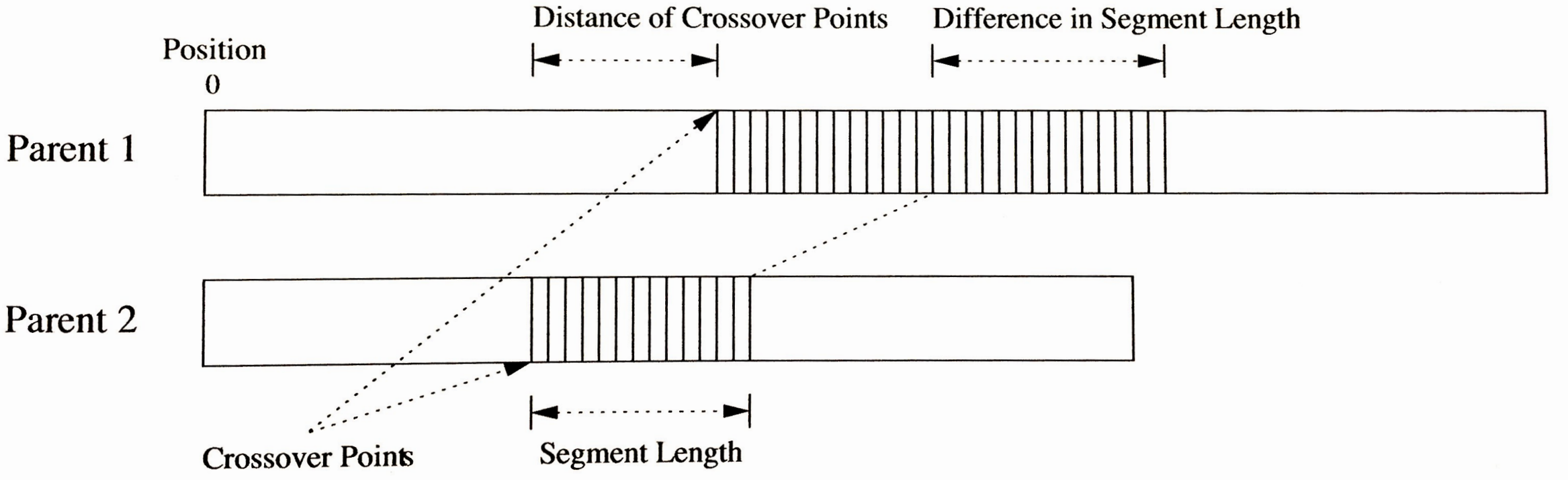
- Crossover
  - Swap segments from two parents
  - Instruction is the smallest unit for variation

- Mutation
  - Randomly replace registers, constants, operations by equivalents from predefined sets or valid ranges

- Selection of variation points
  - Tree GP, bias choosing leaf nodes (variables or constants)
  - Linear GP, each instruction is selected with the same probability

LGP fundamentals - Crossover

- Two-point crossover
  - Parameters: minimum and maximum program length
    maximum segment length
    maximum distance of crossover points
    maximum difference in segment length



LGP fundamentals - Crossover

- One-point crossover
  - Programs are swapped at one point only
  - Parameters: minimum and maximum program length
    maximum distance of crossover points

- One-segment recombination
  - For parent1, *delete* a segment or *insert* a segment from parent2, and apply the same to parent2.
  - Parameters: insertion rate, deletion rate (sum to 1)
    - can adjust growth/shrink bias
    minimum and maximum program length
    maximum segment length

LGP fundamentals - Mutation

- Micro mutation
  - Modify elements of an instruction
- Macro mutation
  - Instruction mutation: Insert or delete a single instruction
  - One-segment mutation: Insert a randomly created subsequence
  - Two-segment mutation: Replace an existing segment with a randomly created subsequence
- Effective mutation

LGP fundamentals - Evolution

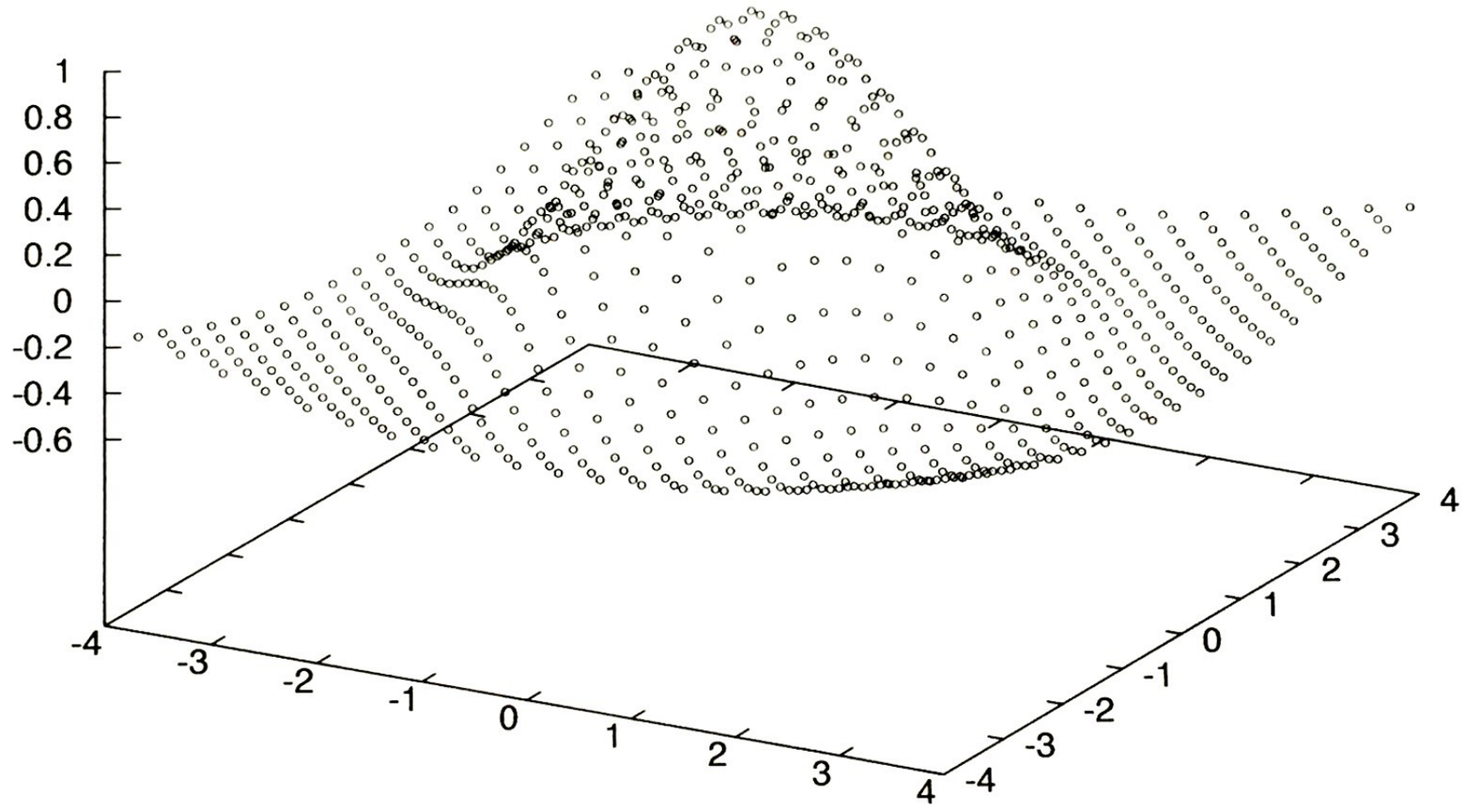
1. **Initialize** the population with random programs and calculate their fitness values
2. Randomly **select** $2*n_{ts}$ individuals from the population without replacement
3. Perform two fitness **tournaments** of size n_{ts}
4. Use copies of the two winners to **produce** two offspring by variation operators with certain probabilities
5. **Evaluate** the fitness of the two offspring
6. If any of the offspring is fitter than the current best-fit individual (saved), **validate** the new best program using validation data
7. **Replace** the two tournament losers with two offspring
8. Repeat steps 2 to 8 until the **termination** criteria is met
9. The program with the best validation accuracy is the **output** of the algorithm
10. **Test** the output program using a testing dataset and report the testing accuracy

LGP applications

- Symbolic regression

  Ricker wavelet function

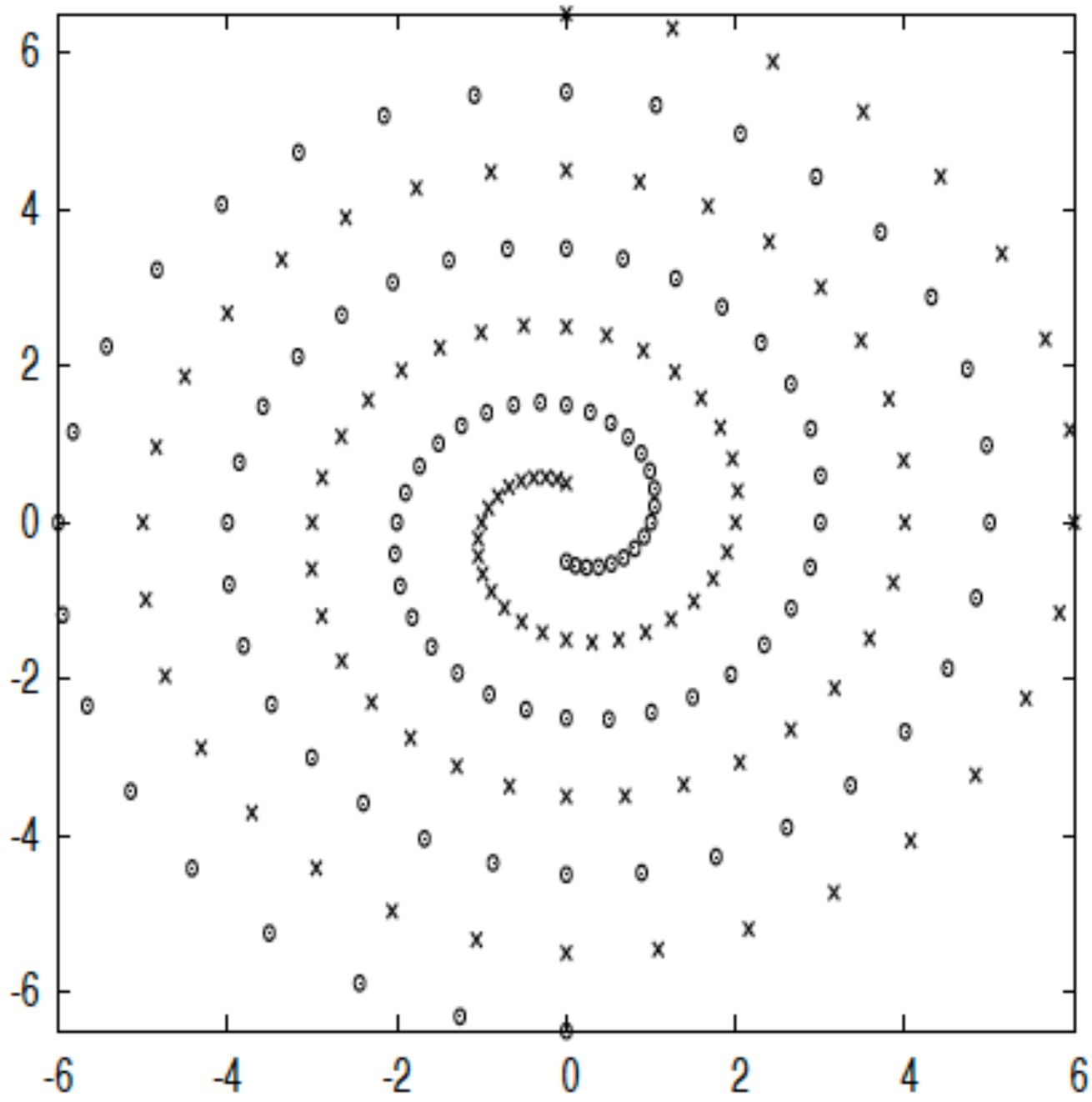
$$\left(1 - \frac{x^2}{4} - \frac{y^2}{4}\right) \times e^{\left(-\frac{x^2}{8} - \frac{y^2}{8}\right)}$$



LGP applications

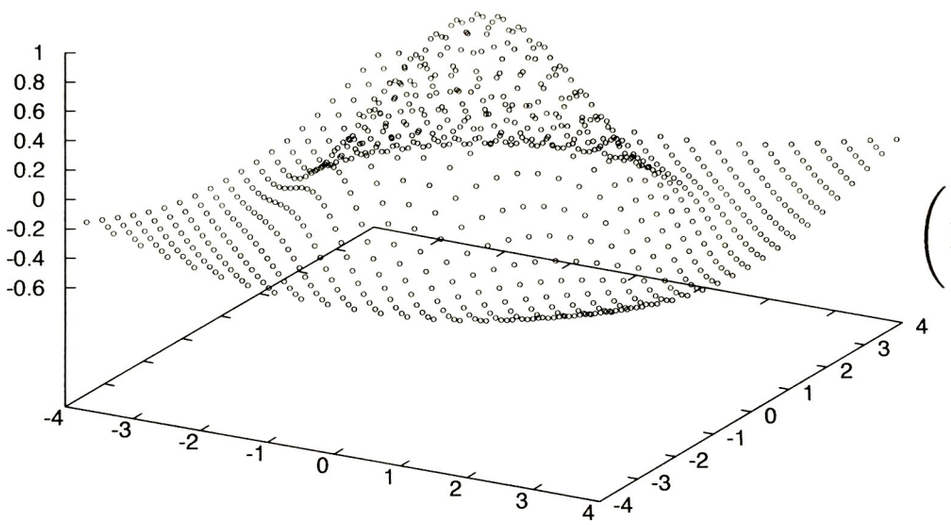
- Classification

  Spiral problem

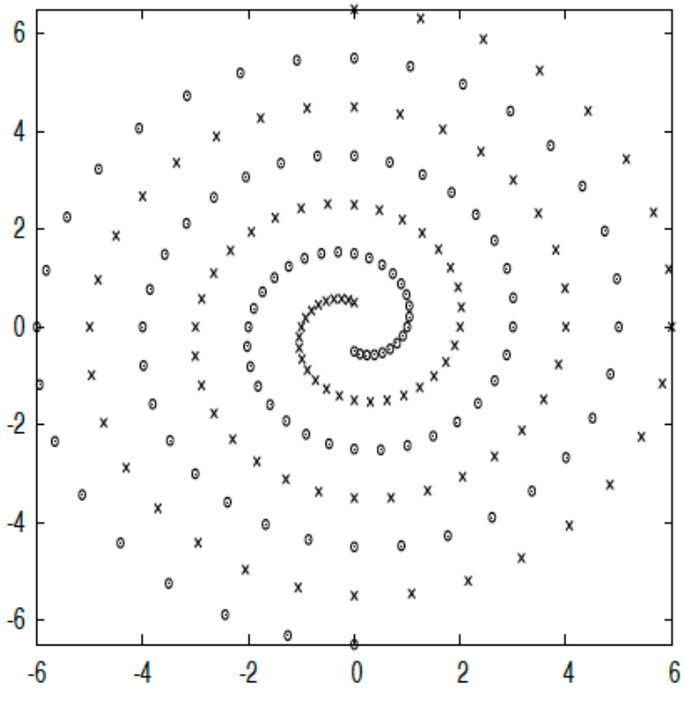


LGP applications

- LGP configuration and parameter setting



$$\left(1 - \frac{x^2}{4} - \frac{y^2}{4}\right) \times e^{\left(-\frac{x^2}{8} - \frac{y^2}{8}\right)}$$



Problem	*mexican hat*	*spiral*
Problem type	regression	classification
Input range	$[-4.0, 4.0]$	$[-2\pi, 2\pi]$
Output range	$[-1, 1]$	$\{0, 1\}$
Number of inputs	2	2
Number of outputs	1	1
Number of output classes	—	2
Number of examples	400	194
Number of registers	$2 + 4$	$2 + 4$
Fitness function	SSE	CE
Instruction set	$\{+, -, \times, /, x^y\}$	$\{+, -, \times, /, sin, cos, if >\}$
Constants	$\{1, .., 9\}$	$\{1, .., 9\}$

Parameter	Setting
Number of generations	1,000
Population size	1,000
Tournament size	2
Maximum program length	200
Initial program length	5–15
Macro variation rate	75%
Micro mutation rate	25%
Reproduction rate	100%