

Evolution strategies

Chapter 4

ES quick overview

- Developed: Germany in the 1970's
- Early names: I. Rechenberg, H.-P. Schwefel
- Typically applied to:
 - numerical optimisation
- Attributed features:
 - fast
 - good optimizer for real-valued optimisation
 - relatively much theory
- Special:
 - self-adaptation of (mutation) parameters standard

ES technical summary tableau

Representation	Real-valued vectors
Recombination	Discrete or intermediary
Mutation	Gaussian perturbation
Parent selection	Uniform random
Survivor selection	(μ, λ) or $(\mu + \lambda)$
Specialty	Self-adaptation of mutation step sizes

Introductory example

- Task: minimise $f : \mathbb{R}^n \rightarrow \mathbb{R}$
- Algorithm: “two-membered ES” using
 - Vectors from \mathbb{R}^n directly as chromosomes
 - Population size 1
 - Only mutation creating one child
 - Greedy selection

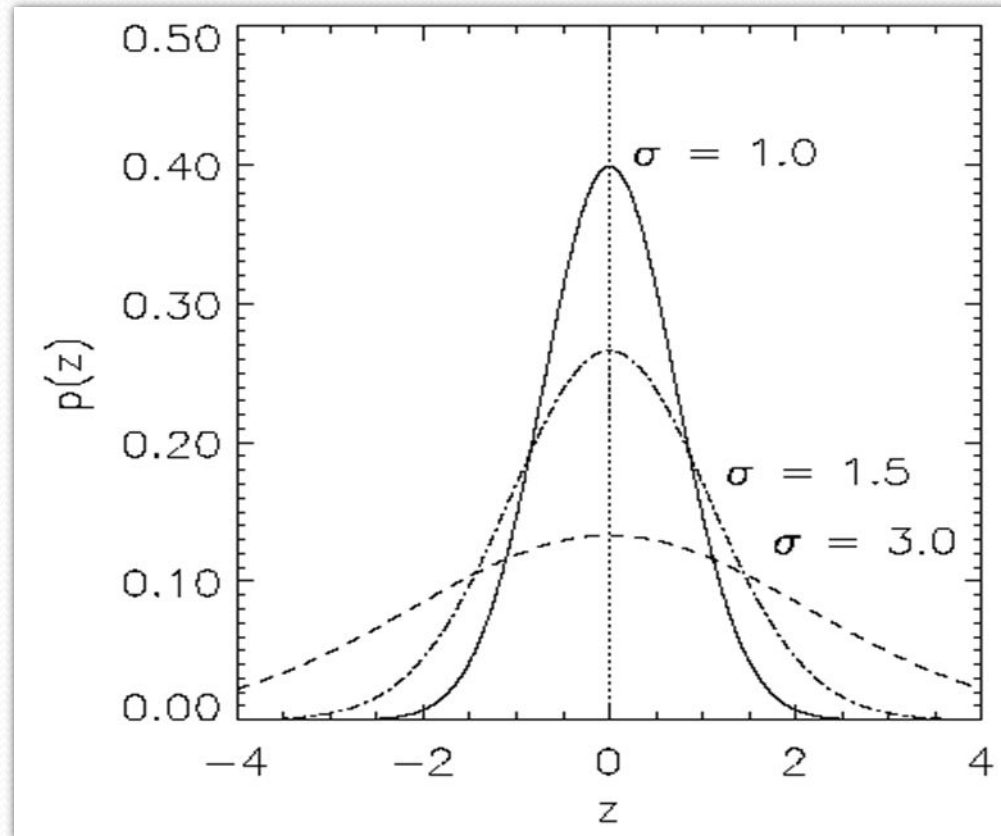
Introductory example: pseudocode

- Set $t = 0$
- Create initial point $x^t = \langle x_1^t, \dots, x_n^t \rangle$
- REPEAT UNTIL (*TERMIN.COND* satisfied) DO
 - Draw z_i from a normal distr. for all $i = 1, \dots, n$
 - $y_i^t = x_i^t + z_i$
 - IF $f(x^t) < f(y^t)$ THEN $x^{t+1} = x^t$
 - ELSE $x^{t+1} = y^t$
 - FI
 - Set $t = t+1$
- OD

Introductory example: mutation mechanism

- z values drawn from normal distribution $N(\xi, \sigma)$
 - mean ξ is set to 0
 - variation σ is called mutation step size
- σ is varied on the fly by the “1/5 success rule”:
- This rule resets σ after every k iterations by
 - $\sigma = \sigma / c$ if $p_s > 1/5$
 - $\sigma = \sigma \cdot c$ if $p_s < 1/5$
 - $\sigma = \sigma$ if $p_s = 1/5$
- where p_s is the % of successful mutations, $0.8 \leq c \leq 1$

Illustration of normal distribution



Another historical example: the jet nozzle experiment

Task: to optimize the shape of a jet nozzle

Approach: random mutations to shape + selection

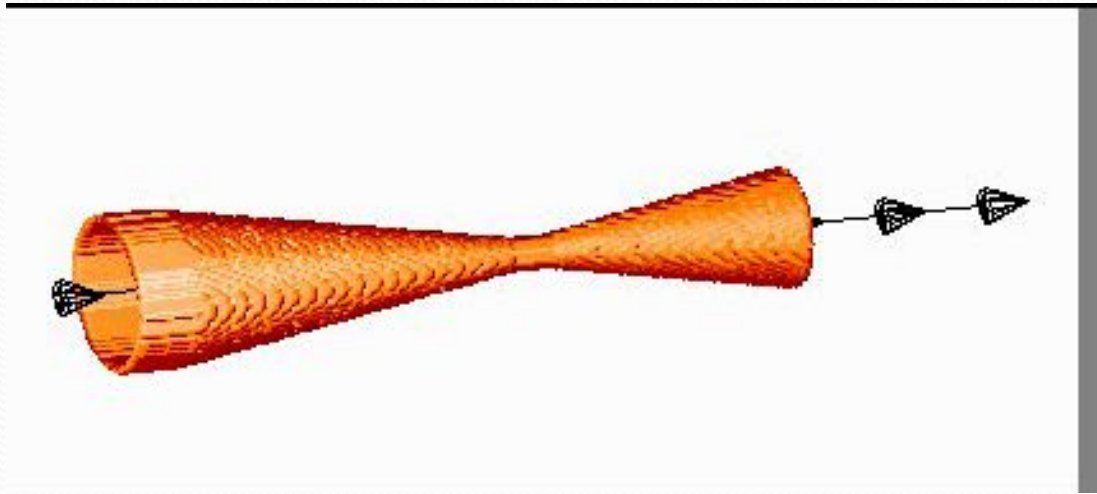


Initial shape



Final shape

The famous jet nozzle experiment (movie)



Representation

- Chromosomes consist of three parts:
 - Object variables: x_1, \dots, x_n
 - Strategy parameters:
 - Mutation step sizes: $\sigma_1, \dots, \sigma_{n\sigma}$
 - Rotation angles: $\alpha_1, \dots, \alpha_{n\alpha}$
- Not every component is always present
- Full size: $\langle x_1, \dots, x_n, \sigma_1, \dots, \sigma_n, \alpha_1, \dots, \alpha_k \rangle$
where $k = n(n-1)/2$ (no. of i, j pairs)

Mutation

- Main mechanism: changing value by adding random noise drawn from normal distribution
- $x'_i = x_i + N(0, \sigma)$
- Key idea:
 - σ is part of the chromosome $\langle x_1, \dots, x_n, \sigma \rangle$
 - σ is also mutated into σ' (see later how)
- Thus: mutation step size σ is coevolving with the solution x

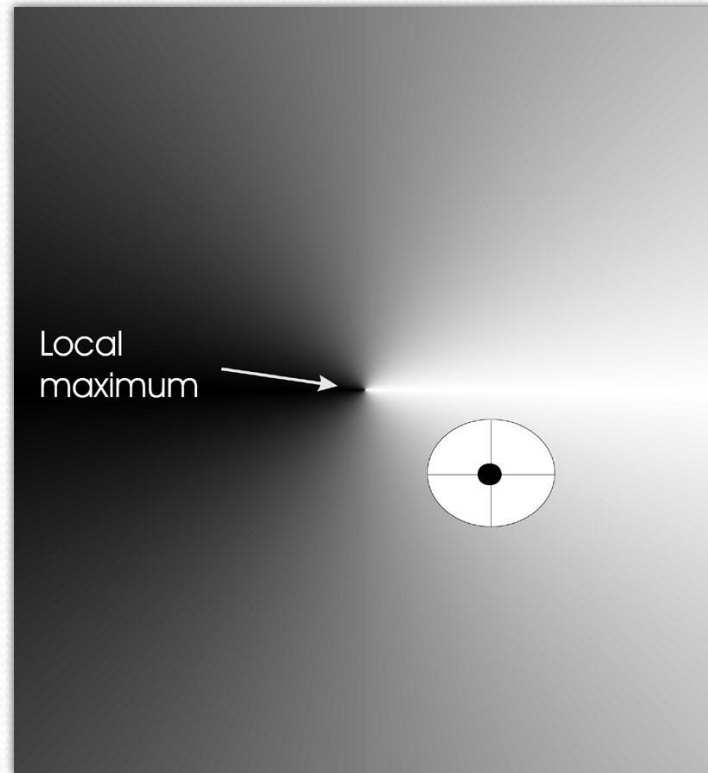
Mutate σ first

- Net mutation effect: $\langle x, \sigma \rangle \square \langle x', \sigma' \rangle$
- Order is important:
 - first $\sigma \square \sigma'$ (see later how)
 - then $x \square x' = x + N(0, \sigma')$
- Rationale: new $\langle x', \sigma' \rangle$ is evaluated twice
 - Primary: x' is good if $f(x')$ is good
 - Secondary: σ' is good if the x' it created is good
 - Step-size only survives through “hitch-hiking”
- Reversing mutation order this would not work

Mutation case 1: Uncorrelated mutation with one σ

- Chromosomes: $\langle x_1, \dots, x_n, \sigma \rangle$
 - $\sigma' = \sigma \cdot \exp(\tau \cdot N(0,1))$
 - $x'_i = x_i + \sigma' \cdot N(0,1)$
- Typically the “learning rate” $\tau \propto 1/n^{1/2}$
- And we have a boundary rule $\sigma' < \varepsilon_0 \Rightarrow \sigma' = \varepsilon_0$

Mutants with equal likelihood

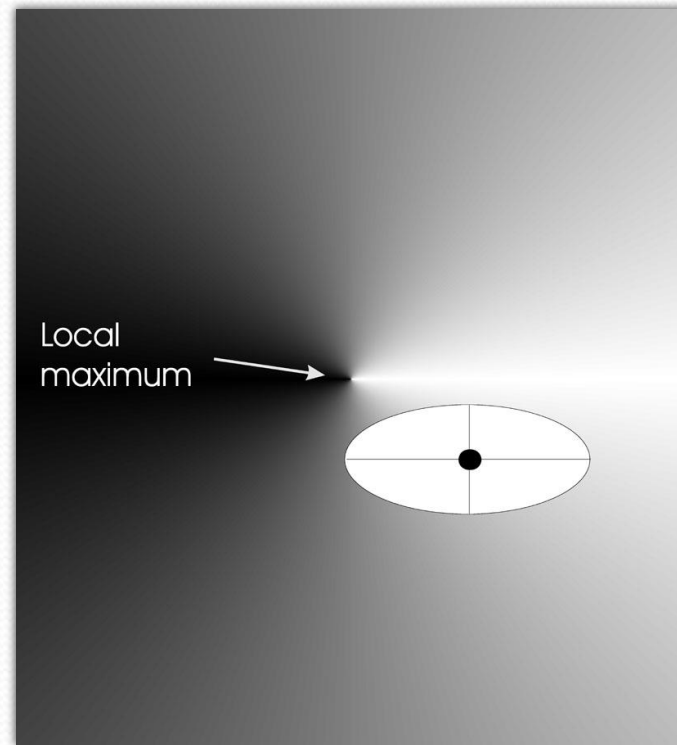


Circle: mutants having the same chance to be created

Mutation case 2: Uncorrelated mutation with n σ 's

- Chromosomes: $\langle x_1, \dots, x_n, \sigma_1, \dots, \sigma_n \rangle$
 - $\sigma'_i = \sigma_i \cdot \exp(\tau' \cdot N(0,1) + \tau \cdot N_i(0,1))$
 - $x'_i = x_i + \sigma'_i \cdot N_i(0,1)$
- Two learning rate parameters:
 - τ' overall learning rate
 - τ coordinate wise learning rate
- $\tau \propto 1/(2n)^{1/2}$ and $\tau \propto 1/(2n^{1/2})^{1/2}$
- Boundary rule: $\sigma'_i < \varepsilon_0 \Rightarrow \sigma'_i = \varepsilon_0$

Mutants with equal likelihood



Ellipse: mutants having the same chance to be created

Mutation case 3: Correlated mutations

- Chromosomes: $\langle x_1, \dots, x_n, \sigma_1, \dots, \sigma_n, \alpha_1, \dots, \alpha_k \rangle$
where $k = n \cdot (n-1)/2$
- Covariance matrix C is defined as:
 - $c_{ii} = \sigma_i^2$
 - $c_{ij} = 0$ if i and j are not correlated
 - $c_{ij} = \frac{1}{2} \cdot (\sigma_i^2 - \sigma_j^2) \cdot \tan(2 \alpha_{ij})$ if i and j are correlated
- Note the numbering / indices of the α 's

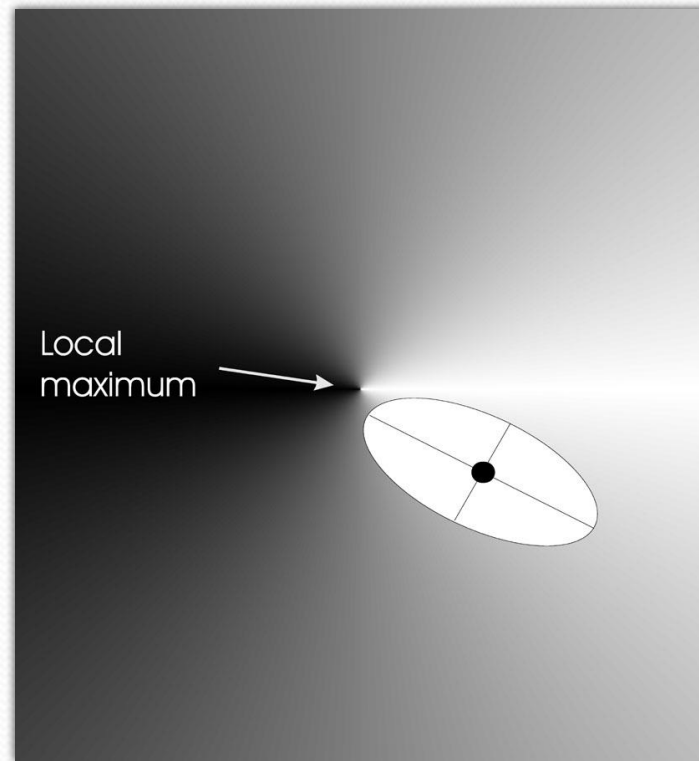
Correlated mutations cont'd

The mutation mechanism is then:

- $\sigma'_i = \sigma_i \cdot \exp(\tau' \cdot N(0,1) + \tau \cdot N_i(0,1))$
- $\alpha'_j = \alpha_j + \beta \cdot N(0,1)$
- $\mathbf{x}' = \mathbf{x} + \mathbf{N}(\mathbf{0}, \mathbf{C}')$
 - \mathbf{x} stands for the vector $\langle x_1, \dots, x_n \rangle$
 - \mathbf{C}' is the covariance matrix \mathbf{C} after mutation of the α values
- $\tau \propto 1/(2n)^{1/2}$ and $\tau' \propto 1/(2n^{1/2})^{1/2}$ and $\beta \approx 5^\circ$
- $\sigma'_i < \varepsilon_0 \Rightarrow \sigma'_i = \varepsilon_0$ and
- $|\alpha'_j| > \pi \Rightarrow \alpha'_j = \alpha'_j - 2\pi \text{sign}(\alpha'_j)$

- NB Covariance Matrix Adaptation Evolution Strategy (CMA-ES) is probably the best EA for numerical optimisation, cf. CEC-2005 competition

Mutants with equal likelihood



Ellipse: mutants having the same chance to be created

Recombination

- Creates one child
- Acts per variable / position by either
 - Averaging parental values, or
 - Selecting one of the parental values
- From two or more parents by either:
 - Using two selected parents to make a child
 - Selecting two parents for each position anew

Names of recombinations

	Two fixed parents	Two parents selected for each i
$z_i = (x_i + y_i)/2$	Local intermediary	Global intermediary
z_i is x_i or y_i chosen randomly	Local discrete	Global discrete

Parent selection

- Parents are selected by uniform random distribution whenever an operator needs one/some
- Thus: ES parent selection is unbiased - every individual has the same probability to be selected
- Note that in ES “parent” means a population member (in GA’s: a population member selected to undergo variation)

Survivor selection

- Applied after creating λ children from the μ parents by mutation and recombination
- Deterministically chops off the “bad stuff”
- Two major variants, distinguished by the basis of selection:
 - (μ, λ) -selection based on the set of children only
 - $(\mu + \lambda)$ -selection based on the set of parents and children:

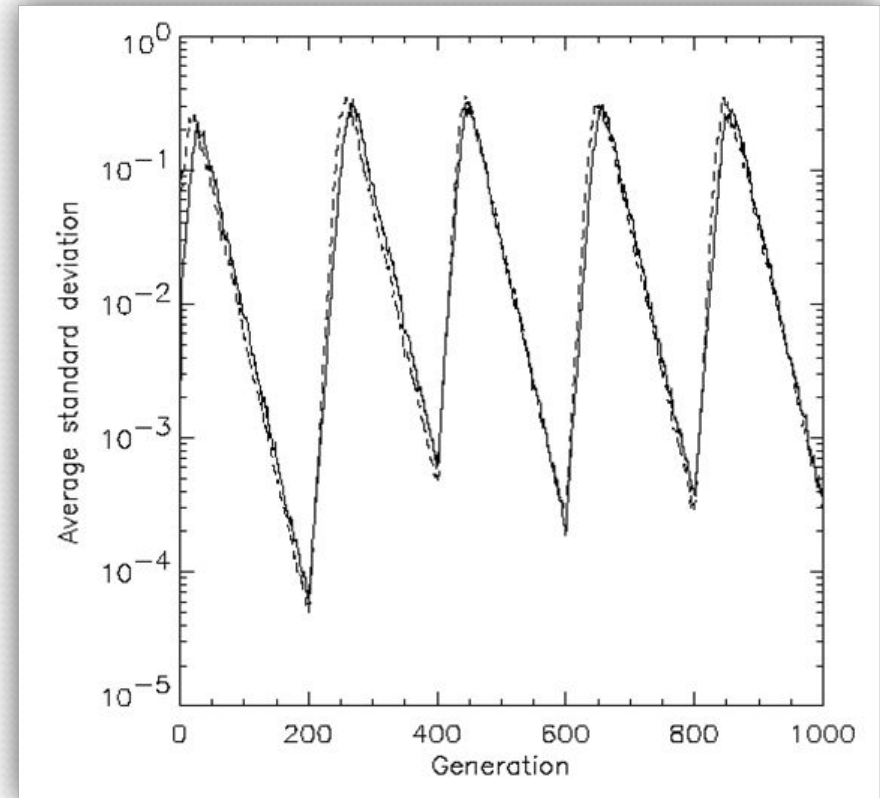
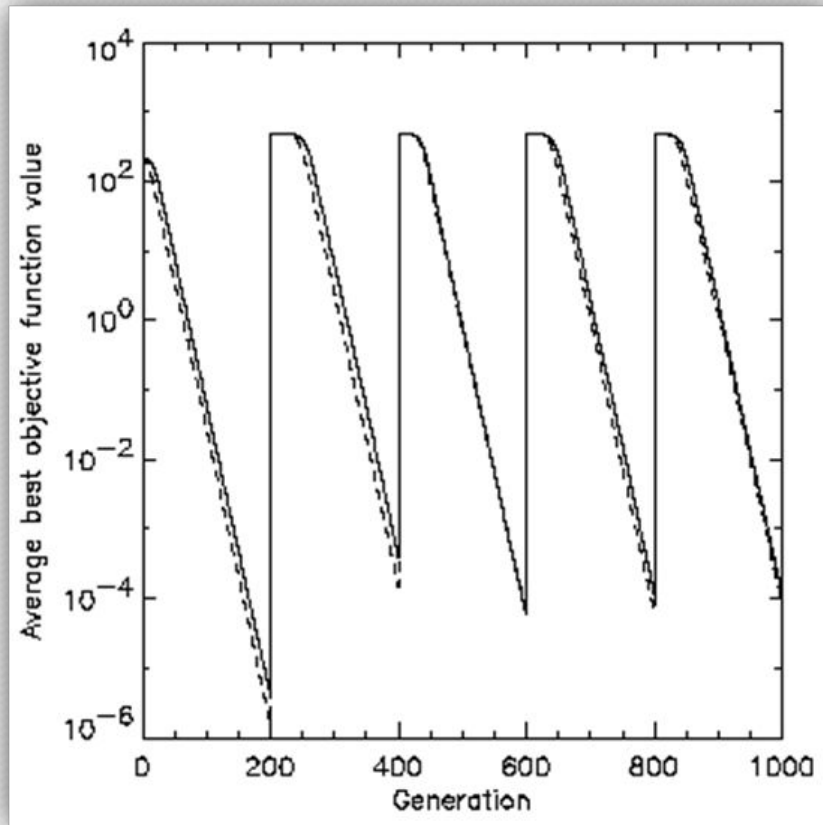
Survivor selection cont'd

- $(\mu+\lambda)$ -selection is an elitist strategy
- (μ,λ) -selection can “forget”
- Often (μ,λ) -selection is preferred for:
 - Better in leaving local optima
 - Better in following moving optima
 - Using the + strategy bad σ values can survive in $\langle x,\sigma \rangle$ too long if their host x is very fit
- Selective pressure in ES is high compared with GAs,
- $\lambda \approx 7 \cdot \mu$ is a traditionally good setting (decreasing over the last couple of years, $\lambda \approx 3 \cdot \mu$ seems more popular lately)

Self-adaptation illustrated

- Given a dynamically changing fitness landscape (optimum location shifted every 200 generations)
- Self-adaptive ES is able to
 - follow the optimum and
 - adjust the mutation step size after every shift !

Self-adaptation illustrated cont'd



Changes in the fitness values (left) and the mutation step sizes (right)

Prerequisites for self-adaptation

- $\mu > 1$ to carry different strategies
- $\lambda > \mu$ to generate offspring surplus
- Not “too” strong selection, e.g., $\lambda \approx 7 \cdot \mu$
- (μ, λ) -selection to get rid of misadapted σ 's
- Mixing strategy parameters by (intermediary) recombination on them

Example application: the cherry brandy experiment

- Task: to create a colour mix yielding a target colour (that of a well known cherry brandy)
- Ingredients: water + red, yellow, blue dye
- Representation: $\langle w, r, y, b \rangle$ no self-adaptation!
- Values scaled to give a predefined total volume (30 ml)
- Mutation: lo / med / hi σ values used with equal chance
- Selection: (1,8) strategy

Example application: cherry brandy experiment cont'd

- Fitness: students effectively making the mix and comparing it with target colour
- Termination criterion: student satisfied with mixed colour
- Solution is found mostly within 20 generations
- Accuracy is very good

Example application: the Ackley function (Bäck et al '93)

- The Ackley function (here used with $n = 30$):

$$f(x) = -20 \cdot \exp\left(-0.2 \sqrt{\frac{1}{n} \cdot \sum_{i=1}^n x_i^2}\right) - \exp\left(\frac{1}{n} \sum_{i=1}^n \cos(2\pi x_i)\right) + 20 + e$$

- Evolution strategy:
 - Representation:
 - $-30 < x_i < 30$ (coincidence of 30's!)
 - 30 step sizes
 - (30,200) selection
 - Termination : after 200000 fitness evaluations
 - Results: average best solution is $7.48 \cdot 10^{-8}$ (very good)