Computational Intelligence

Master in Artificial Intelligence

2023-24

Introduction to Evolution Strategies

Lluís A. Belanche

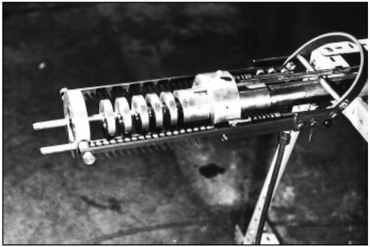
Soft Computing Research Group

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
School of Professional & Executive Development

The nozzle experiment (I)

collection of conical nozzle parts

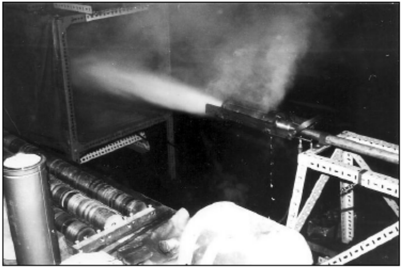
device for clamping nozzle parts



The nozzle experiment (II)

Hans-Paul Schwefel
while changing nozzle parts

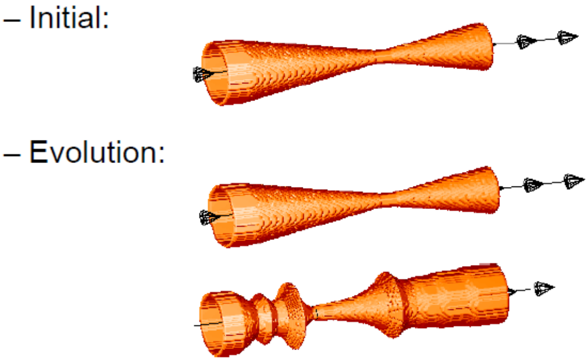
The nozzle experiment (III)



the nozzle in operation …

… while measuring degree of efficiency

The nozzle experiment (IV)



32% of increase in efficiency!

J. Klockgether and H.-P. Schwefel, "*Two-phase nozzle and hollow core jet experiments*". Proceedings of the 11th Symposium on Engineering Aspects of Magneto-Hydrodynamics, Caltech, Pasadena, California, USA, 1970.

The Gaussian Distribution

A continuous d-variate random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$ is **normally distributed**, written $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, when its joint *pdf* is:

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right\}$$

where $\boldsymbol{\mu}$ is the <u>mean vector</u> and $\Sigma_{d\times d} = (\sigma_{ij}^2)$ is the (real symmetric and p.d.) <u>covariance matrix</u>.

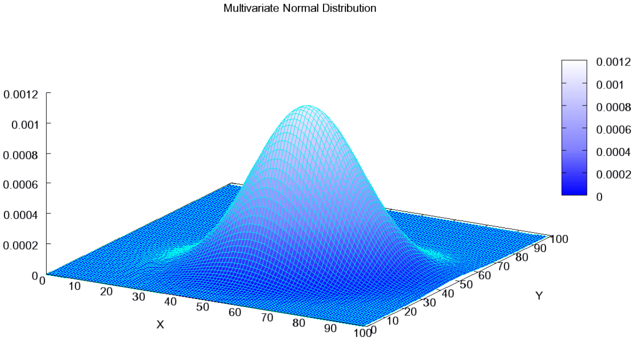
- $\mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu}$ and $\mathbb{E}[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^{\mathrm{T}}] = \Sigma$.
- $CoVar[X_i, X_j] = \sigma_{ij}^2$ and $Var[X_i] = \sigma_{ii}^2$

---

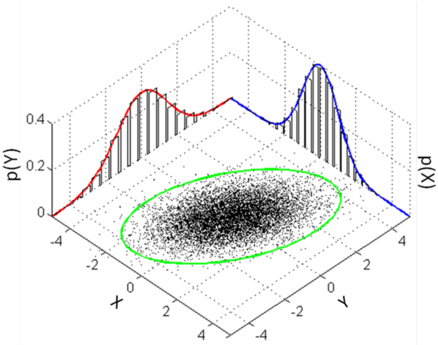
if $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$, then X_i, X_j are independent $\Longleftrightarrow$ $CoVar[X_i, X_j] = 0$

(in general, only the left-to-right implication holds)

The Gaussian Distribution ($d = 2$)

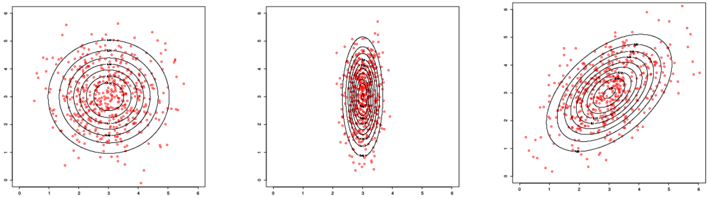


The Gaussian Distribution ($d = 2$)



Observations from a bivariate normal distribution, a contour ellipsoid, the two marginal distributions, and their histograms (all images from the Wikipedia)

Linear algebra point of view ($d = 2$)



$$\boldsymbol{\mu} = \begin{bmatrix} 3 \\ 3 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 0.1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

- The principal directions (a.k.a. PCs) of the hyperellipsoids are given by the eigenvectors $\boldsymbol{u}_i$ of Σ, which satisfy $\Sigma \boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i$.
- The lengths of the hyperellipsoids along these axes are proportional to $\sqrt{\lambda_i}$ (note $\lambda_i > 0$), where λ_i are the eigenvalues associated with $\boldsymbol{u}_i$.

Conceptual view

- What is behind the choice of a **multivariate Gaussian**?

  Examples from a class are noisy versions of an ideal class member (a prototype):

  - Prototype: modeled by the mean vector
  - Noise: modeled by the covariance matrix

- The quantity

$$d(\boldsymbol{x}) := \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

  is called the **Mahalanobis distance** for $\boldsymbol{x}$

- Very important! the number of parameters is $\frac{d(d+1)}{2} + d$

Mathematical view

Positive definiteness

For a Gaussian distribution to be well-defined, Σ has to be real symmetric and positive definite (p.d.):

- for all non-null vectors $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{x}^{\mathrm{T}} \Sigma \boldsymbol{x} > 0$ must hold true.
- alt., all eigenvalues must be positive (note they are real)

Examples: are these matrices p.d.?

$$a. \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad b. \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}$$

$$c. \begin{pmatrix} 3 & -1 \\ -1 & 2 \end{pmatrix} \qquad d. \begin{pmatrix} 1 & 4 \\ \frac{1}{2} & 1 \end{pmatrix}$$

a. YES; b. YES
c. YES; d. NO

Mathematical view

Positive definiteness

For a Gaussian distribution to be well-defined, Σ has to be real symmetric and positive definite (p.d.):

- for all non-null vectors $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{x}^{\mathrm{T}}\Sigma\boldsymbol{x} > 0$ must hold true.
- alt., all eigenvalues must be positive (note they are real)

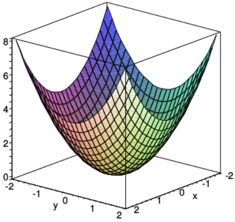
Examples: are these matrices p.d.?

$$a. \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad b. \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}$$

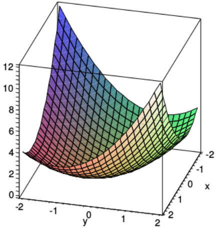
$$c. \begin{pmatrix} 3 & -1 \\ -1 & 2 \end{pmatrix} \qquad d. \begin{pmatrix} 1 & 4 \\ \frac{1}{2} & 1 \end{pmatrix}$$

a. YES; b. YES
c. YES; d. NO

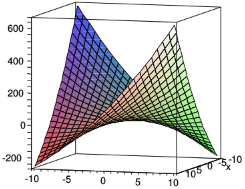
Mathematical view



a. $z_1^2 + z_2^2$;



b. $z_1^2 + z_1 z_2 + z_2^2$;



d. $z_1^2 + \dfrac{9}{2} z_1 z_2 + z_2^2$

Evolution Strategies: main characteristics

- Continuous search space $\mathbb{R}^n$ (*n* **objective** parameters)
- Various *ad hoc* recombination operators
- Deterministic (μ,λ)-replacement
- Generation of an offspring *surplus*: $\lambda \gg \mu$
- Emphasis on mutation: *n*-dimensional Gaussian
- Self-adaptation of mutation parameters
  (first self-adaptive EA!)

Evolution Strategies: Representation

Recall the notation:

$$(\mu/\rho, \lambda) - \mathrm{ES}$$

The three parts of an individual

1. object variables $\boldsymbol{x} \in \mathbb{R}^n$ to compute fitness $F(\boldsymbol{x})$
2. standard deviations $\boldsymbol{\sigma} \in \mathbb{R}_+^{n_\sigma}$ to express variances
3. rotation angles $\boldsymbol{\alpha} \in (-\pi, \pi]^{n_\alpha}$ to express covariances (all Gaussians are zero mean)

Evolution Strategies: Mutation (I)

Simple self-adaptive Mutation

$n_\sigma = 1, n_\alpha = 0$
(one mutation parameter per individual)

$\sigma := \sigma \cdot \exp(\mathcal{N}(0, \tau_0))$
For $i \in \{1, 2, \ldots, n\}$

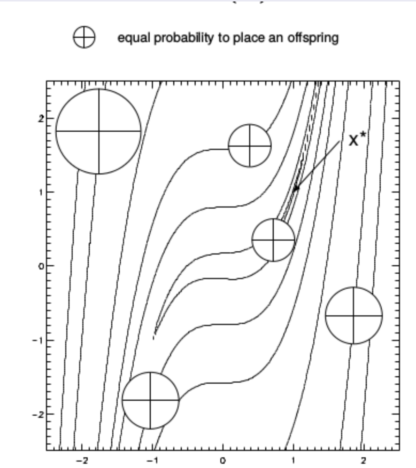
1. $x_i := x_i + \mathcal{N}_i(0, \sigma^2)$

where

$$\tau_0 \propto \frac{1}{n}$$

Evolution Strategies: Mutation (I)

Simple self-adaptive Mutation ($n = 2$)



Evolution Strategies: Mutation (II)

Diagonal self-adaptive Mutation

$n_\sigma = n, n_\alpha = 0$
(one mutation parameter per individual and variable)

For $i \in \{1, 2, \ldots, n_\sigma\}$

1. $\sigma_i := \sigma_i \cdot \exp(\mathcal{N}(0, \tau') + \mathcal{N}_i(0, \tau))$

For $i \in \{1, 2, \ldots, n\}$

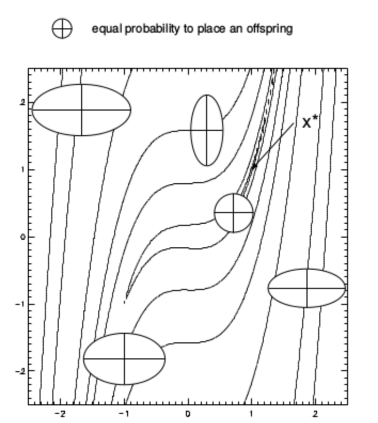
1. $x_i := x_i + \mathcal{N}_i(0, \sigma_i^2)$

where

$$\tau \propto \frac{1}{2\sqrt{n}}$$
$$\tau' \propto \frac{1}{2n}$$

Evolution Strategies: Mutation (II)

Diagonal Self-Adaptive Mutation ($n = 2$)



Evolution Strategies: Mutation (III)

Correlated self-adaptive Mutation

$n_\sigma = n, n_\alpha = \left(n - \frac{n_\sigma}{2}\right)(n_\sigma - 1)$
(one covariance matrix per individual, represented by a collection of n_α rotation angles)

For $i \in \{1, 2, \ldots, n_\sigma\}$

1. $\sigma_i := \sigma_i \cdot \exp(\mathcal{N}(0, \tau') + \mathcal{N}_i(0, \tau)),\ \tau \propto \frac{1}{2\sqrt{n}}, \tau' \propto \frac{1}{2n}$

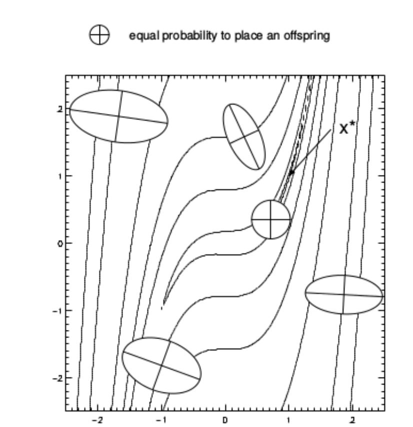
For $i \in \{1, 2, \ldots, n_\alpha\}$

1. $\alpha_i := \alpha_i + \mathcal{N}_i(0, \beta^2),\ \beta \propto 5° = \pi/36 \text{ radians}$

Build Σ using the $\boldsymbol{\sigma}$ and $\boldsymbol{\alpha}$ for individual $\boldsymbol{x}$ and then

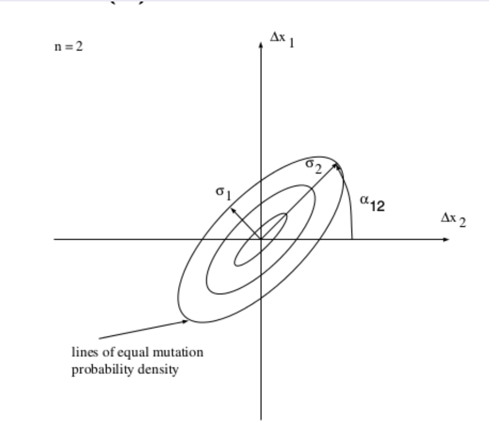
$$\boldsymbol{x} := \boldsymbol{x} + \mathcal{N}(0, \Sigma)$$

Correlated Self-Adaptive Mutation ($n = 2$)



Evolution Strategies: Mutation (III)

Illustration of the mutation ellipsoid for the case $(n = 2)$



Evolution Strategies: Mutation (III)

Theorem (Rudolph 1992)

A real symmetric matrix $\Sigma_{n\times n}$ is p.d. iff it can be decomposed as $\Sigma = (ST)^{\mathrm{T}}(ST)$, with T orthogonal and S diagonal with $s_{ii} > 0$ and:

$$T := \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} T_{ij}(\alpha_{f(i,j)})$$

- T is the product of $\frac{n(n-1)}{2}$ elementary rotation matrices T_{ij}.
- $\alpha_{f(i,j)}$ are the rotation angles (between axes i and j), represented in the chromosomic vector in position $f(i,j)$.
- $T_{ij}(\alpha_{f(i,j)})$ is built as the identity matrix and modified as:

$$r_{ii} = r_{jj} := cos(\alpha_{f(i,j)})$$
$$r_{ij} = -r_{ji} := -sin(\alpha_{f(i,j)}),\ \ i \neq j$$

Evolution Strategies: Mutation (III)

- The *dummy* function $f(i,j)$ is used to index the vector of self-adaptive parameters (angles) α, using a single index.
- As a consequence, a total of $\frac{n(n+1)}{2}$ angles and scaling parameters are sufficient to generate arbitrary correlated Normal random vectors with 0 mean and covariance matrix $\Sigma = (ST)^\top(ST)$ via:

$$\boldsymbol{x} := \boldsymbol{x} + T\mathbf{z}$$

with $\mathbf{z} \sim \mathcal{N}(0, S)$ and $S = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_{n_\sigma}^2)$.

Evolution Strategies: log-normal mutation for the σ_i

Log-normal distribution

It is a continuous probability distribution whose logarithm is normally distributed. A random variable which is log-normally distributed takes only positive real values.

$$\sigma_i := \sigma_i \cdot \exp(\mathcal{N}(0, \tau'))$$

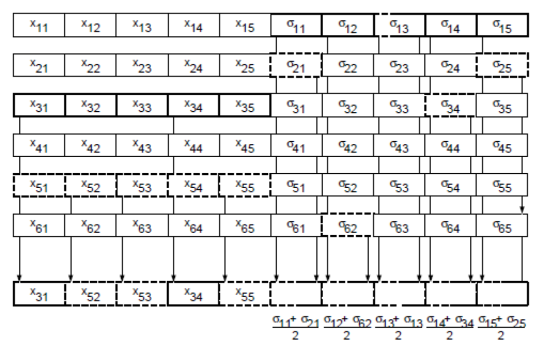
1. Multiplication by positive values preserves positivity
2. $Pr\{X = x\} = Pr\{X = \frac{1}{x}\}, x > 0$
3. Small modifications are more probable than larger ones

Evolution Strategies: recombination (I)

- Usually introduced as the *first* operator (before mutation)
- Generates an intermediate population size of λ by generating *one individual at a time* out of ρ parents by looping $\lambda \gg \mu$ times (generation of a **surplus**)
- Typically $\rho = 2$ (**dual**) or $\rho = \mu$ (**global** recombination):
  - dual: the two parents are chosen at random, per individual
  - global: one parent is held fixed and the other is chosen anew per each gene
- Applied to both objective and strategy parameters (and often differently)
- Two basic ways: choose randomly (**discrete**) and average (**intermediate**)

Evolution Strategies: recombination (II)

Recombination example



- $\mu = 6, n = 5, n_\sigma = n, n_\alpha = 0$ (one mutation parameter per individual and gene)
- dual discrete recombination on x_i; global intermediate on σ_i (first parent held fixed, second chosen anew)

Evolution Strategies: replacement

- Strictly deterministic, rank-based
- The μ best are treated equally
- (μ, λ) selection:
  - offspring surplus $\lambda \gg \mu$
  - important (necessary?) for self-adaptation
  - useful for moving optima, noisy F, ...

$\Rightarrow$ Very strong selective pressure

Evolution Strategies: summary

The crucial claim (Schwefel '87 '92)

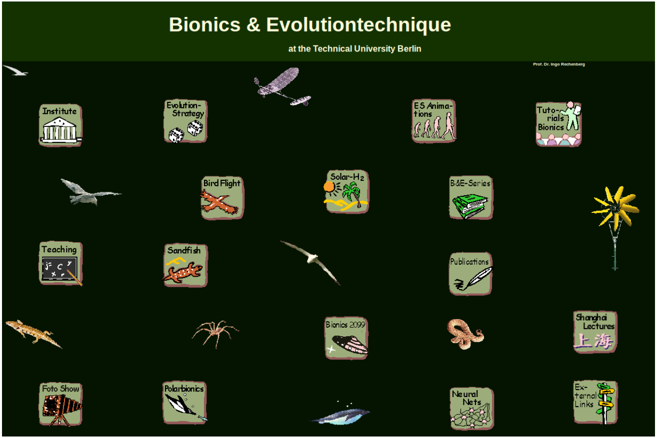
Self-adaptation of strategy parameters works!

- without exogenous or centralized control
- needs mutation of all parameters
- needs generation of a surplus and (μ, λ) replacement
- needs recombination of all parameters

default (recommended) settings:

- $\lambda \propto 7\mu$; historically $(\mu = 7, \lambda = 105)$
- dual discrete recombination on objective parameters
- global intermediate on strategy parameters

Evolution Strategies: demos



Prof. Dr. Ingo Rechenberg

https://web.archive.org/web/20180425010001/http://www.bionik.tu-berlin.de/institut/xstart.htm

Evolution Strategies: Modern developments

The CMA-ES (Covariance Matrix Adaptation Evolution Strategy, by N. Hansen) is the more recent development of ESs:

- Uses a more sophisticated method to update the covariance matrix, particularly useful if the fitness function is complex.
- Learns a second order model of the underlying function (similar to the approximation of the inverse Hessian matrix in quasi-Newton methods, used for example in neural networks).

Resources:

- `www.cmap.polytechnique.fr/~nikolaus.hansen/`
  `cma-es.github.io/index.html`
  `arxiv.org/pdf/1604.00772.pdf`
- **C, C++, Java, Matlab, Octave, Python, Scilab**
  `cma-es.github.io/cmaes_sourcecode_page.html`
- **Julia** `github.com/bionik-berlin/PURE_ES`
- **R** packages {`rCMA`, `cmaes`, `adagio`, `parma`}
- **Python** `github.com/CMA-ES/pycma`
- **Tensorflow 2** `pypi.org/project/cma-es/`