Evolution Strategies

KhangTD - KHTN2018

List of content

I. Problem Statement
II. Overview of ES
    A. Basic Ideal and Algorithm
    B. Recombinations
    C. Parameters Control
    D. Survivor Selection
III. CEM (Cross Entropy Method)

List of content

IV. CMA-ES

A. Sampling
B. Selection and Recombination
C. Adapting the Covariance Matrix
D. Step-Size Control

I. Problem Statement

Continuous Domain Search/Optimization

- Task: Minimize/Maximize an objective function (fitness function, loss function) in continuous domains.

$$f \colon \chi \subseteq \mathbb{R}^n \to \mathbb{R}$$
$$x \mapsto f(x)$$

- Black box scenario (direct search scenario)
    - gradients are not available or not useful
    - problem domain specific knowledge is used only within the black box

- Search costs: number of function evaluations

Continuous Domain Search/Optimization

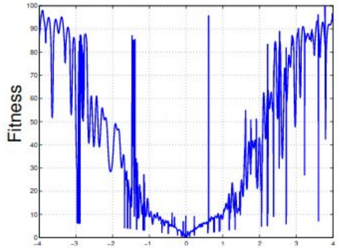
- Goal:
  - Fast convergence to the global optimum or to a robust solution x
  - Solution x with small function value f(x) with least search cost

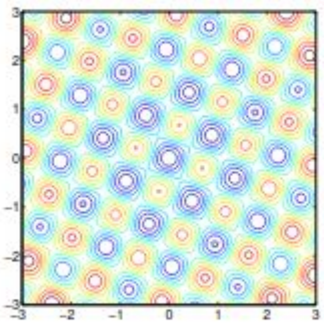
- Problems:
  - Exhaustive search is infeasible
  - Naive random search takes too long
  - Deterministic search is not successful / takes too long

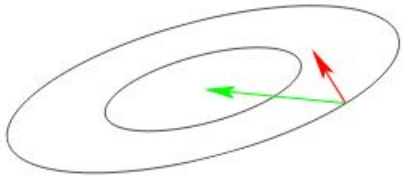
- Approach: Stochastic Search, Evolutionary Algorithms

Some objective function of real-word problem?

- f can be:
  - non-lin
  - Rugged
  - Non-se
  - Ill-cond
  - dimens
  - . . .







Nikolaus Hansen

Senior researcher (directeur de recherche) at Inria

The CMA evolution strategy: a comparing review (2006)

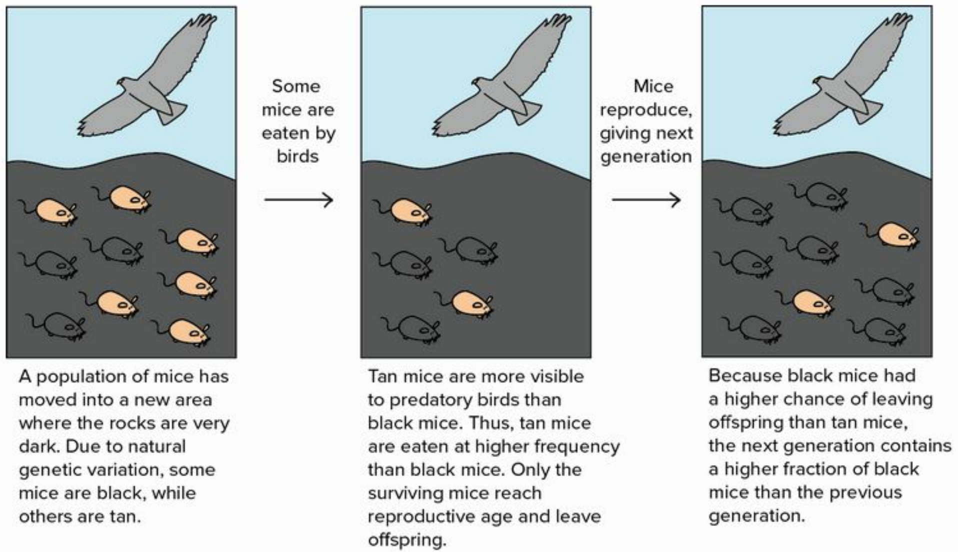
The CMA Evolution Strategy: A tutorial (2016)

http://www.cmap.polytechnique.fr/~nikolaus.hansen/

II. Overview of ES

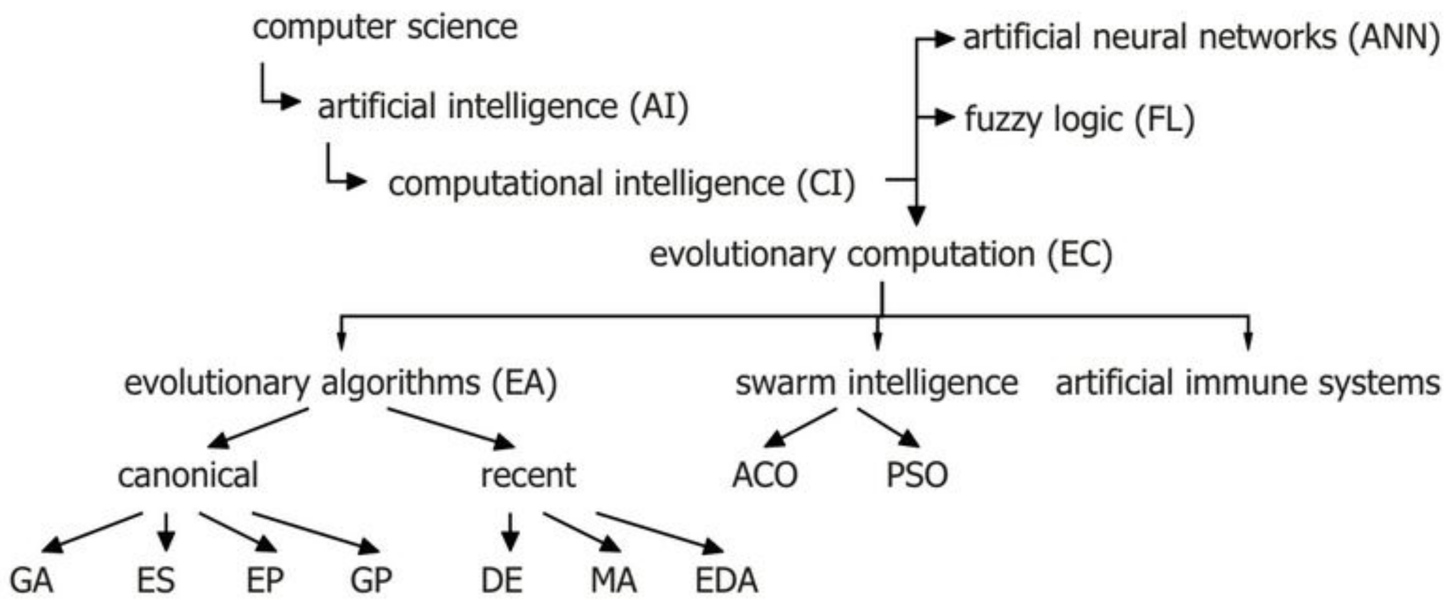
What are Evolution Strategies?

- Evolution Strategies (ES): techniques used in solving continuous domain.
- Evolution Strategies : inverted in early 1960s by Rechenberg and Schewefel.
- Inspired by *natural selection.*

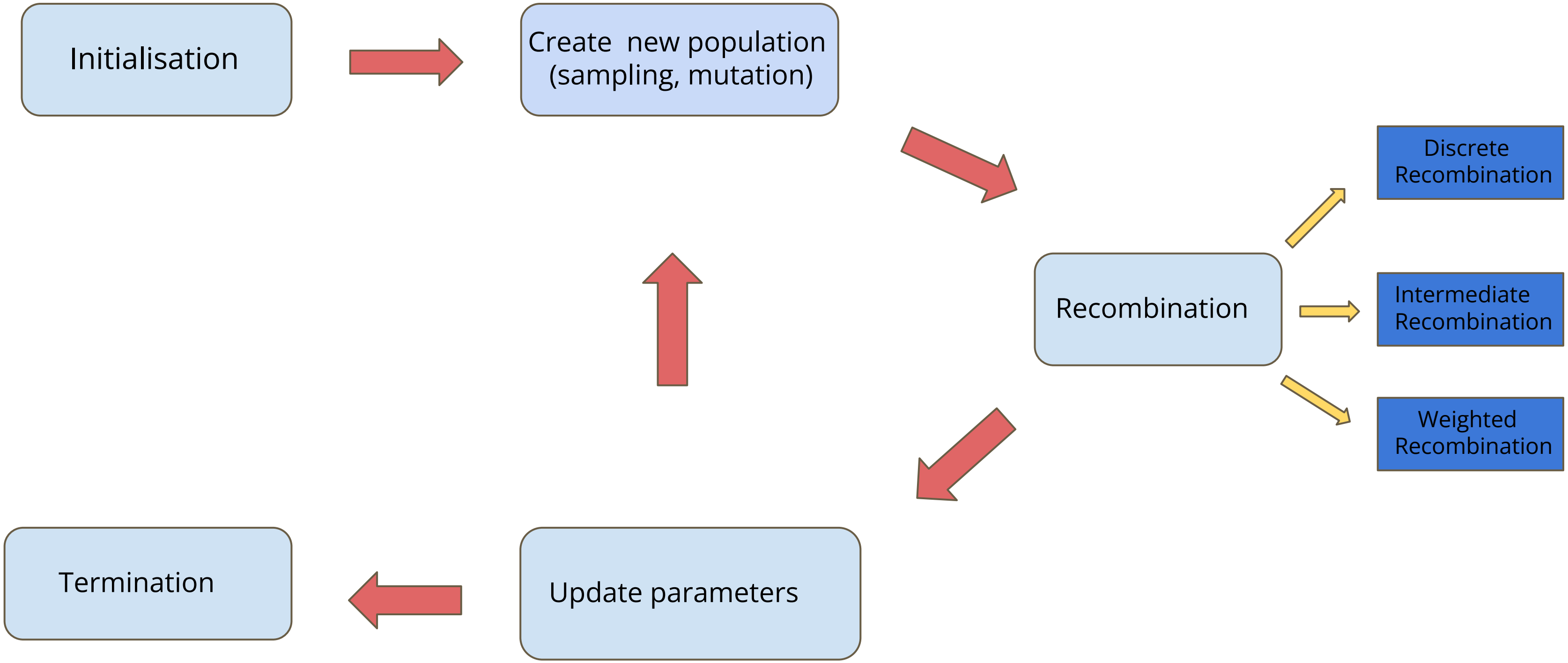


What are Evolution Strategies?

Evolution Strategies (ES) belong to the big family of Evolutionary Algorithms (EA)

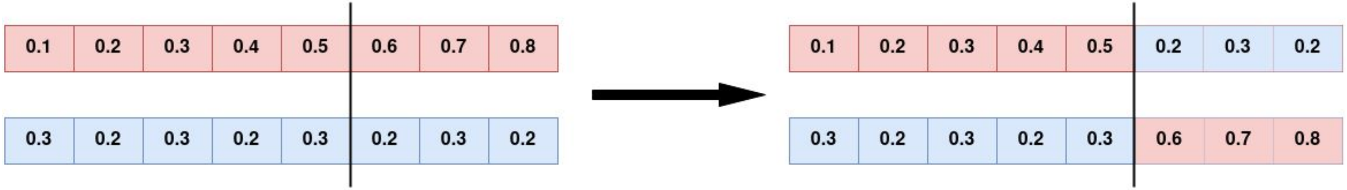


Basic Idea

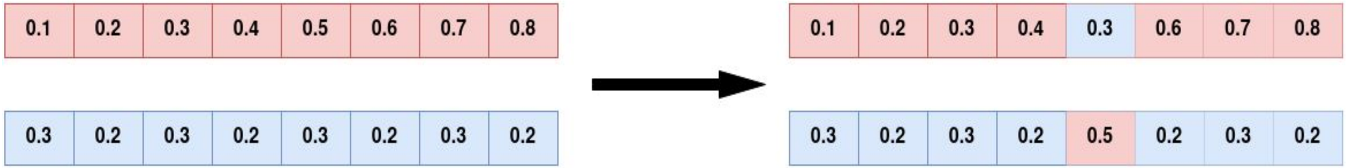

Initialisation
Create new population (sampling, mutation)
Recombination
Discrete Recombination
Intermediate Recombination
Weighted Recombination
Update parameters
Termination

Discrete Recombination

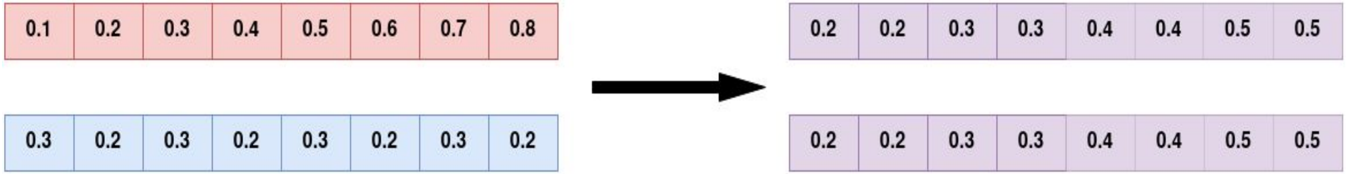
- Simple arithmetic recombination:



- Single arithmetic recombination:

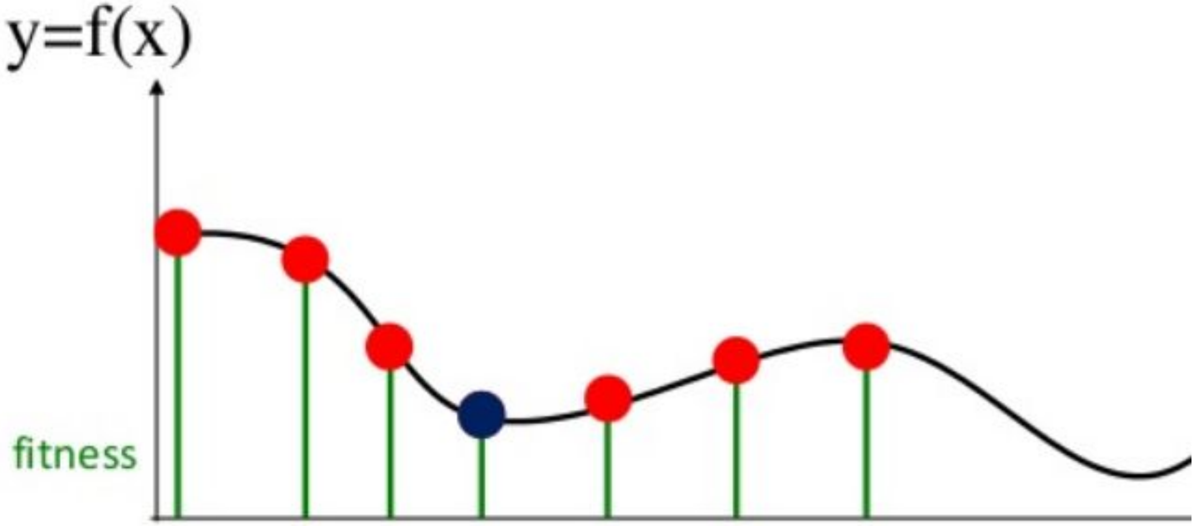


- Whole arithmetic recombination:



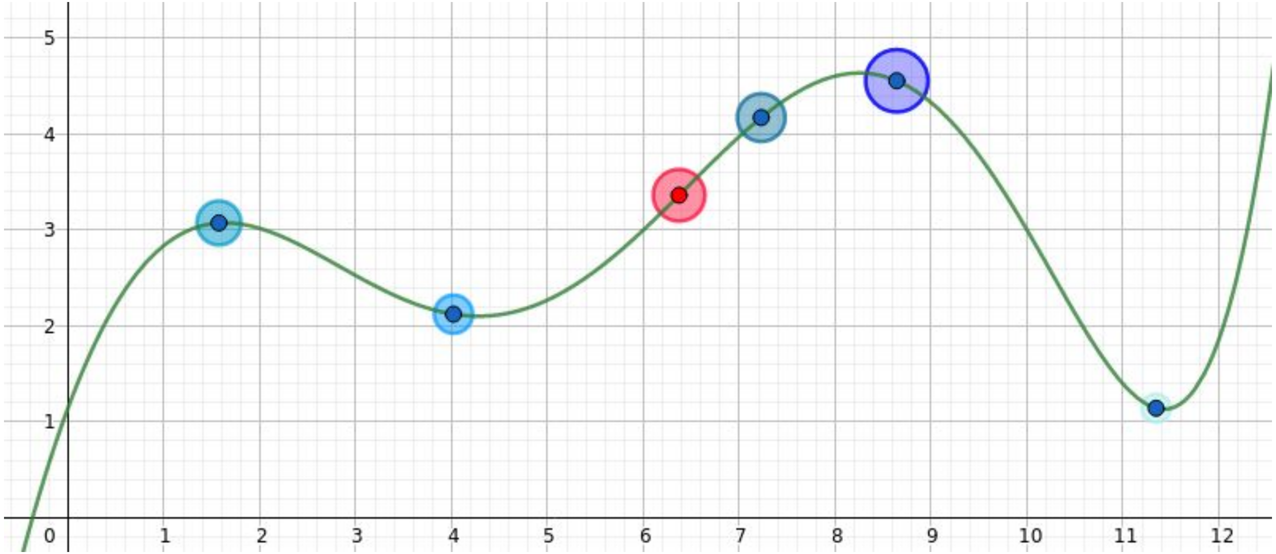
Intermediate Recombination

Takes the average value of all ρ parents (computes the center of mass, the centroid).



Weighted Recombination

- Weighted Recombination is a generalization of intermediate recombination. It takes a weighted average of ρ parents.
- The weight values depend on the fitness ranking, in that better parents never get smaller weights than inferior ones.



Survivor Selection

- Applied create children λ from μ parents by mutation and recombination.
- Two mechanism: Plus (elitist) and comma (non-elitist) selection
  - $(\mu + \lambda)$: selection μ new parents in {parent,offspring}
  - (μ, λ) : selection μ new parents in {offspring}

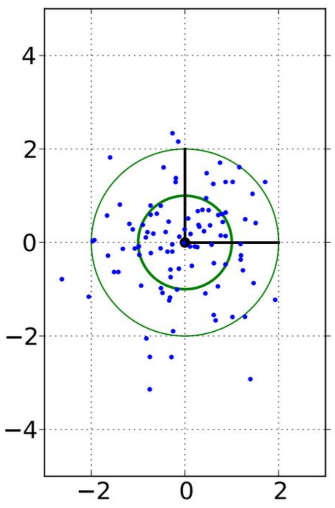
Survivor Selection cont'd

- $(\mu + \lambda)$ is an elitist strategy
- (μ, λ) is truncation selection
- Often (μ, λ) is preferred for:
  - Better in leaving local optima
  - Better in following moving optima
  - Using "plus" selection, bad strategy parameter can survive in population too long, if an individual has relatively good objective variables.

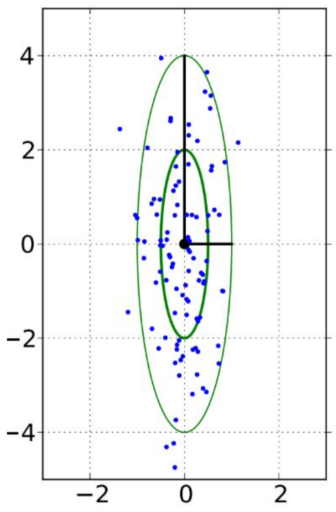
Mutation

The mutation operator introduces ("small") variations by adding a point symmetric perturbation to the result of recombination.

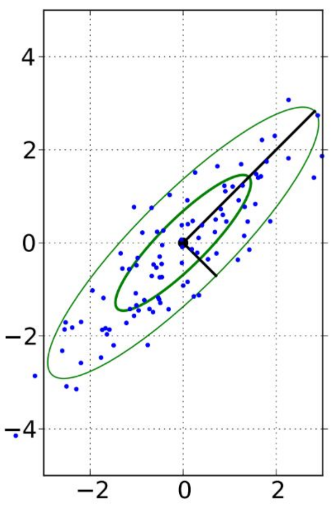
$$m' = m + \sigma\mathcal{N}(0, \mathbf{C})$$



Spherical/Isotropic: **C** is identity matrix



Axis-parallel: **C** is diagonal (positive) matrix



General: **C** is symmetric and PSD matrix

Parameters control

Controlling the parameters of the mutation operator is key to the design of evolution strategies and affects convergence speed.

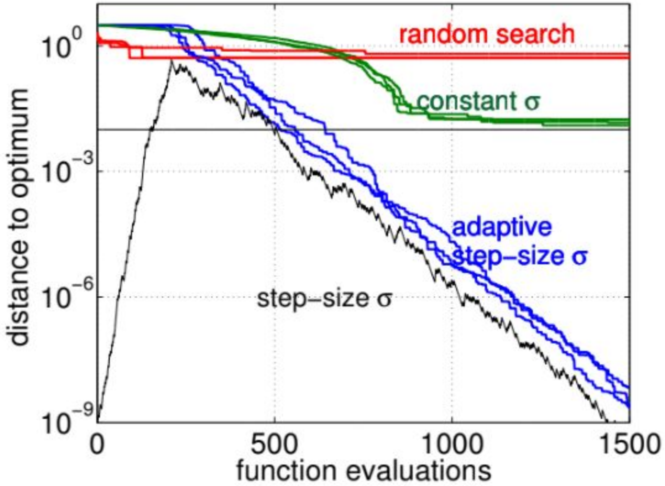


Fig: Step-size affects convergence[1]

[1]Evolution Strategies, Nikolaus Hansen, Dirk V. Arnold and Anne Auger (2015)

Parameters control

- 1/5-th success rule, often applied with Plus - selection
  - increase step-size if more than 20% of the new solutions are successful,decrease otherwise
- σ-self-adaptation, applied with Comma - selection
  - mutation is applied to the step-size and the better, according to the objective function value, is selected
- path length control (Cumulative Step-size Adaptation, CSA)
  - self-adaptation derandomized and non-localized

(1 + 1)ES

Algorithm 1 (1 + 1)ES

Hyperparameters: $c_{inc} > 0$, $c_{dec} > 0$
Input: vector $m^{(0)} \in \mathbb{R}^d$, step-size $\sigma^{(0)} \in \mathbb{R}_{>0}$

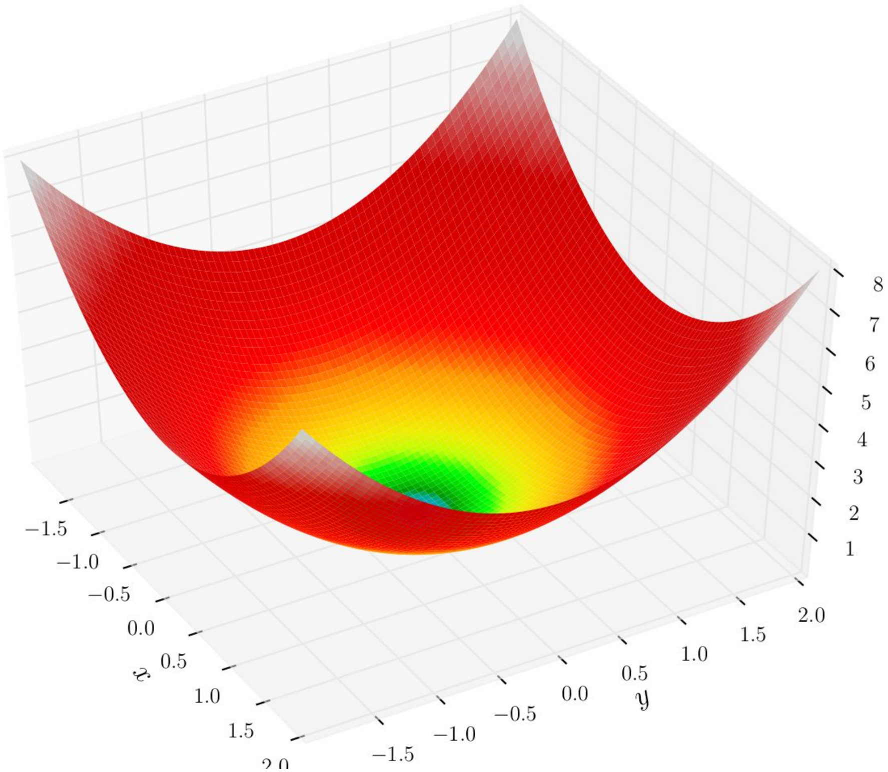
for $t = 0, \ldots, T-1$ **do**
    Create 1 offspring by adding a point symmetric perturbation to $m^{(t)}$

$$(\epsilon^{(t)}) \sim \mathcal{N}(0, \mathbf{I}_d)$$

$$x^{(t)} \leftarrow m^{(t)} + \sigma^{(t)} \epsilon^{(t)}$$

    Survival selection (1 + 1) and update step-size
    if $\mathrm{F}(x^{(t)}) \leq \mathrm{F}(m^{(t)})$ **then**
        $m^{(t+1)} \leftarrow x^{(t)}$
        $\sigma^{(t+1)} \leftarrow \sigma^{(t)} c_{inc}$
    else
        $m^{(t+1)} \leftarrow m^{(t)}$
        $\sigma^{(t+1)} \leftarrow \sigma^{(t)} c_{dec}$
    end if
end for

Test Function



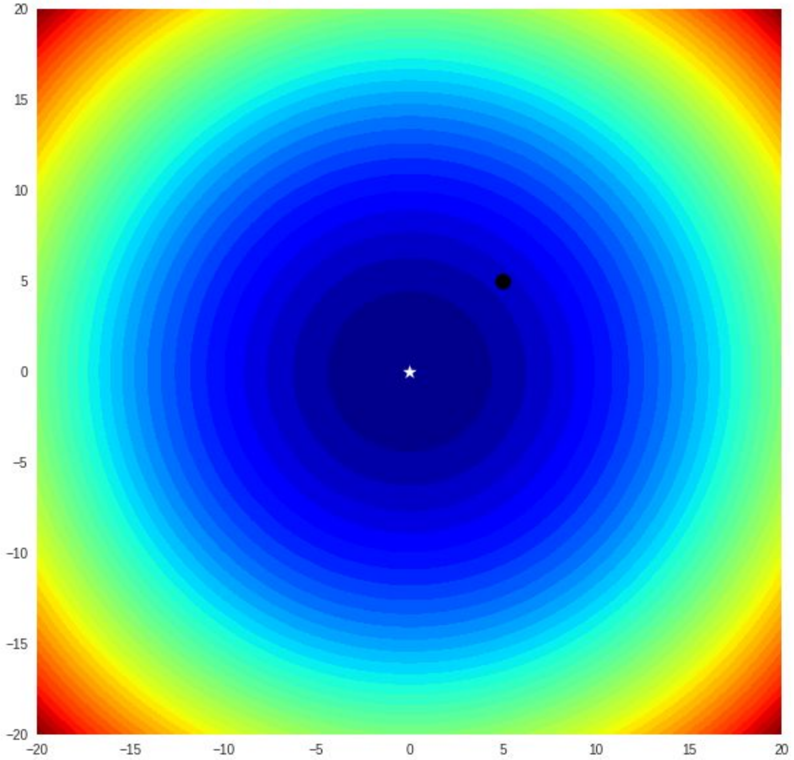
Sphere Function in 3D

Formula: $f(x) = x_1^2 + x_2^2$

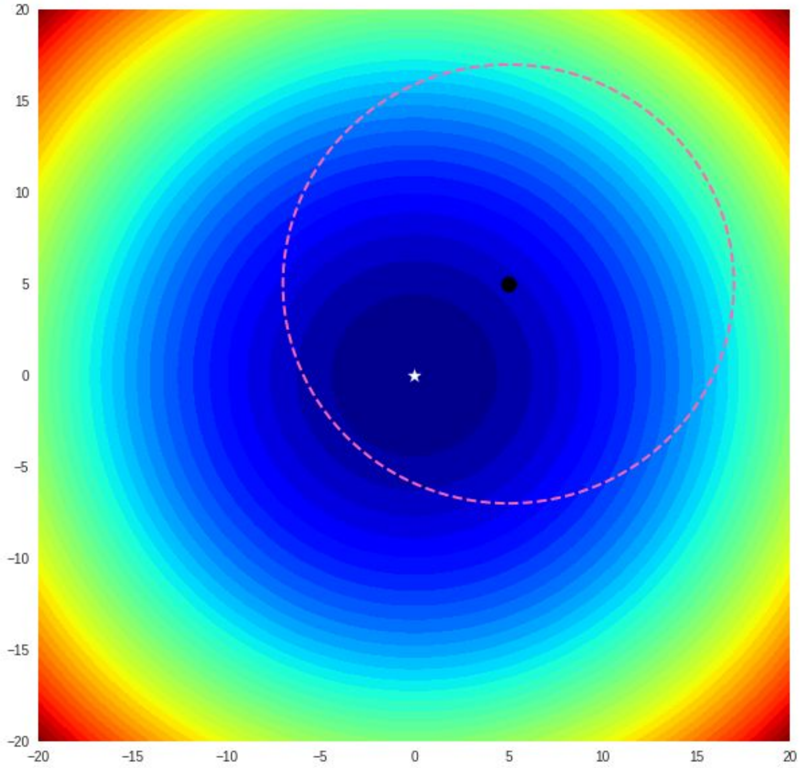
Search domain: $-\infty \leq x_i \leq \infty,\ 1 \leq i \leq 2$

Global minimum: $f(x) = f(0) = 0$

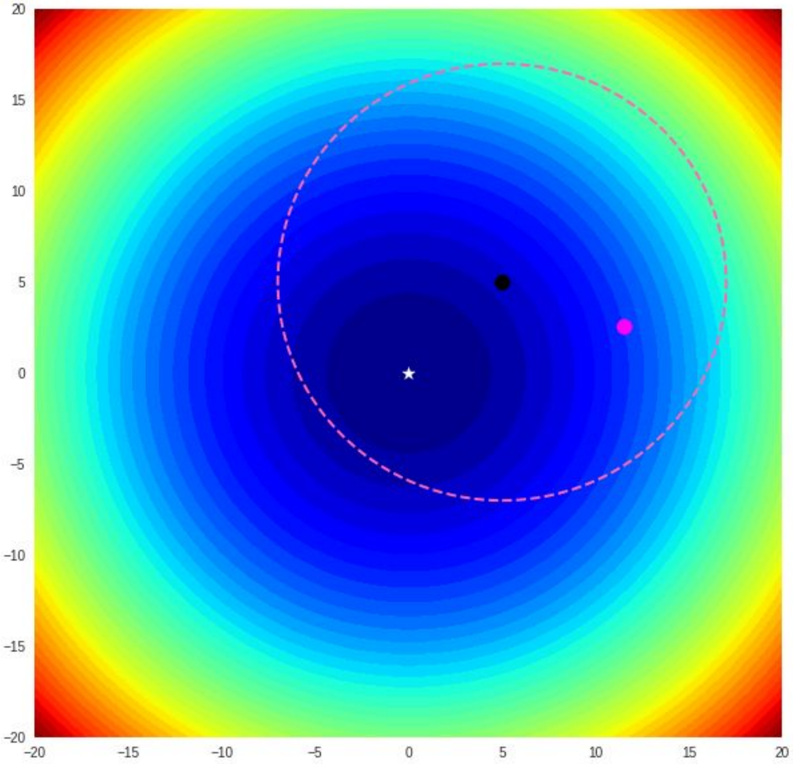
steps 0:
<Figure size 576x396 with 0 Axes>



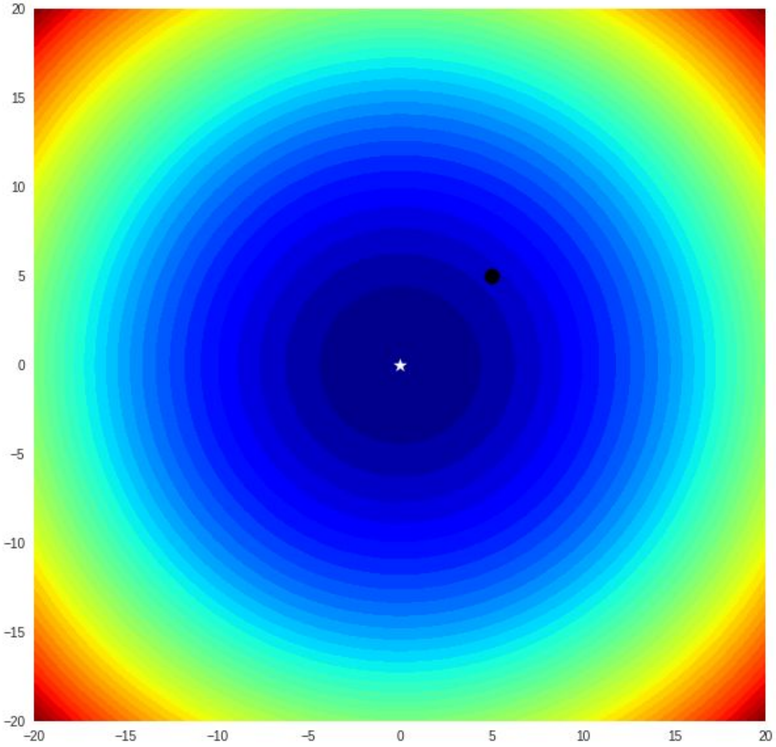
steps 1:



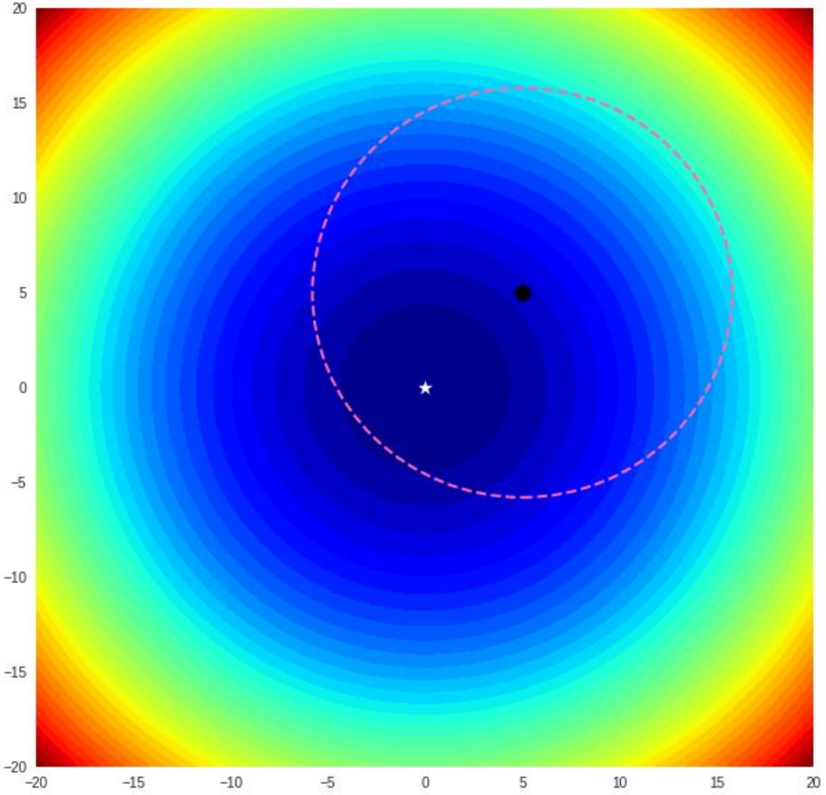
steps 2:



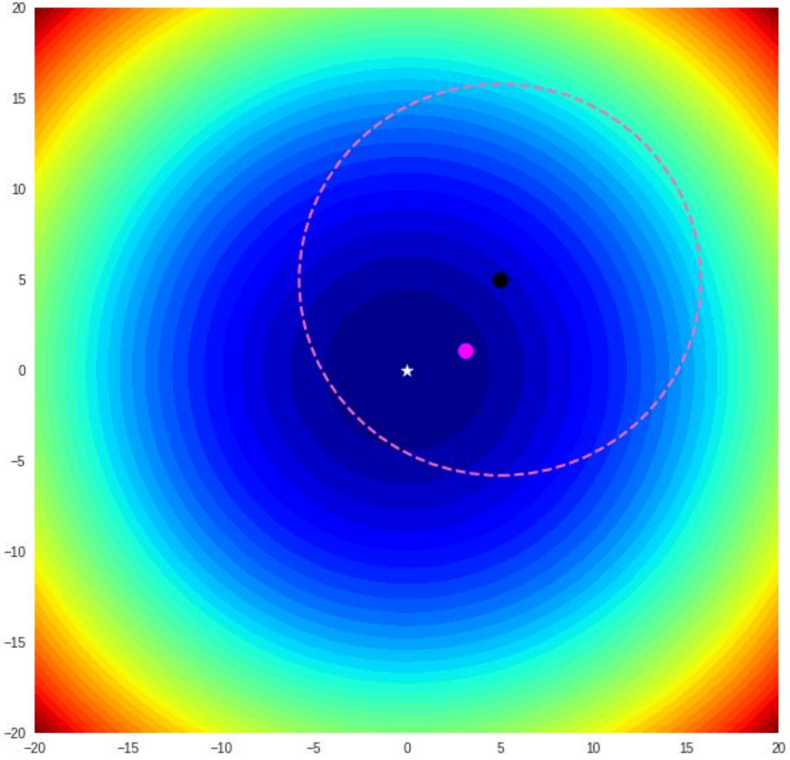
steps 3:



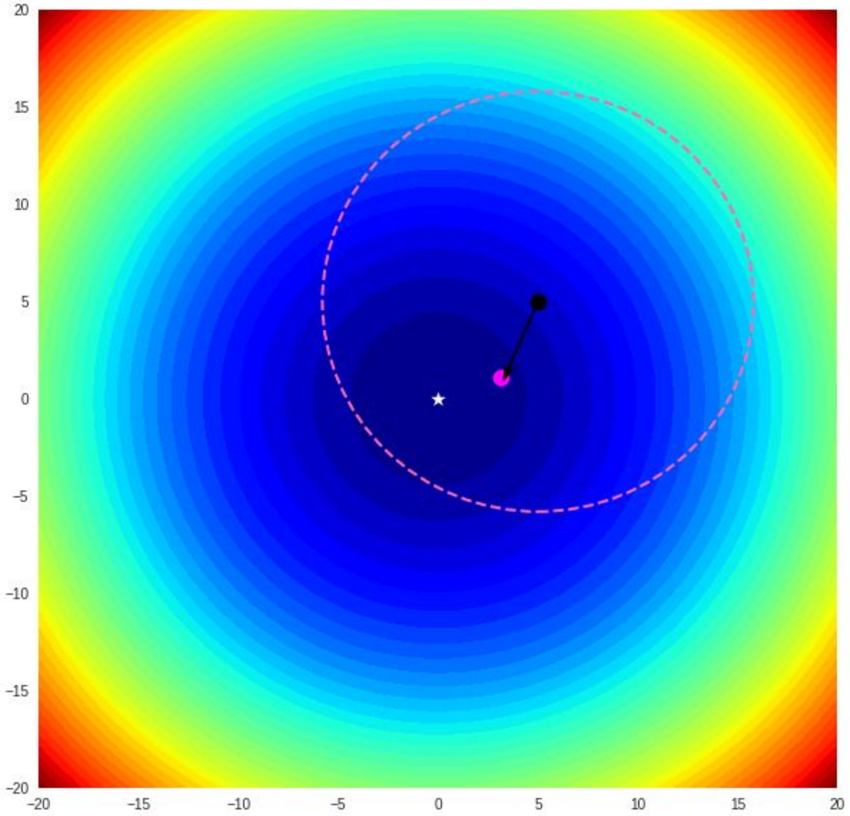
steps 4:



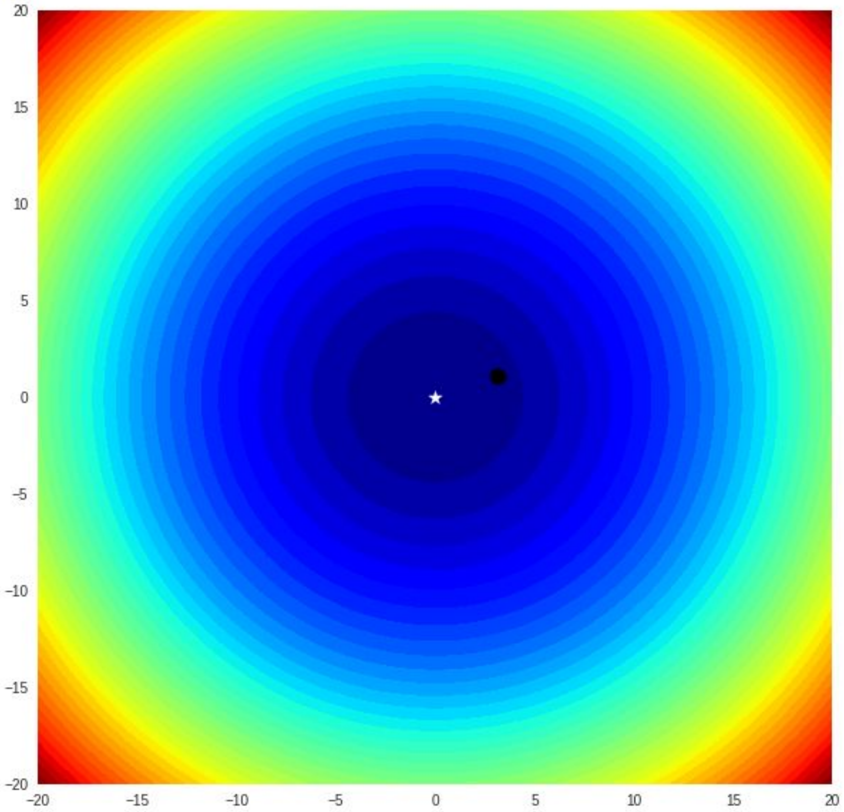
steps 5:

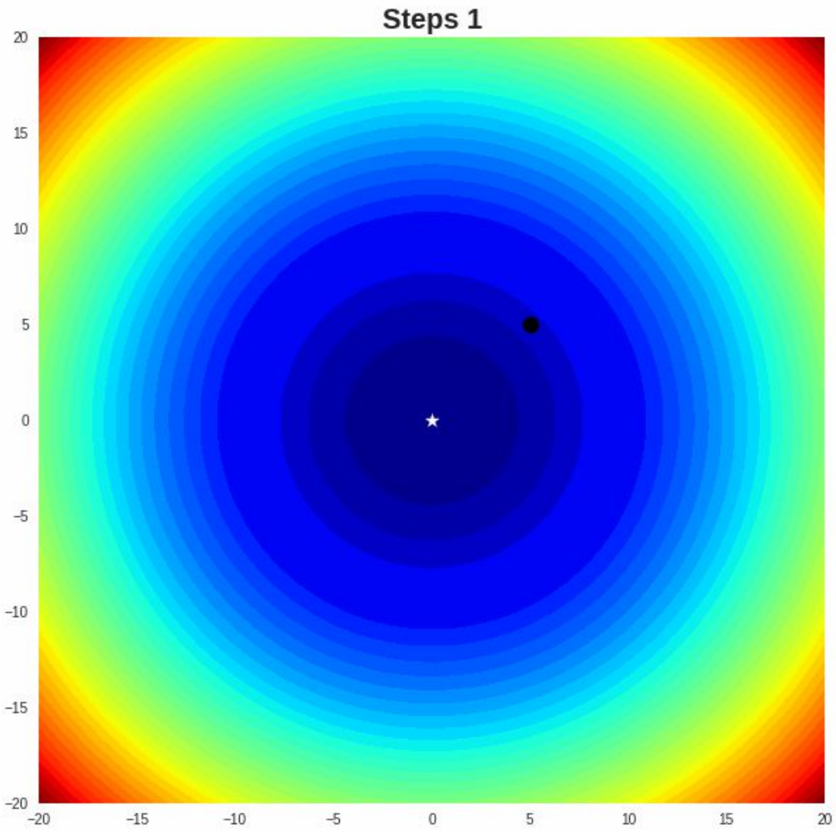


steps 6:



steps 7:




Steps 1
20
15
10
5
0
−5
−10
−15
−20
−20
−15
−10
−5
0
5
10
15
20

(1 + lambda) ES

Algorithm 2 $(1+\lambda)$ES

Hyperparameters: $c_{inc} > 0$, $c_{dec} > 0$
Input: vector $m^{(0)} \in \mathbb{R}^d$, step-size $\sigma^{(0)} \in \mathbb{R}_{>0}$, number of offspring $\lambda > 0$

for $t = 0, \ldots, T-1$ **do**
  Create λ offspring by adding a point symmetric perturbation to $m^{(t)}$

$$(\epsilon_i^{(t)})_{i=1,\ldots,\lambda} \sim \mathcal{N}(0, \mathbf{I}_d)$$
$$x_i^{(t)} \leftarrow m^{(t)} + \sigma^{(t)} \epsilon_i^{(t)}, \quad i = 1, \ldots, \lambda$$

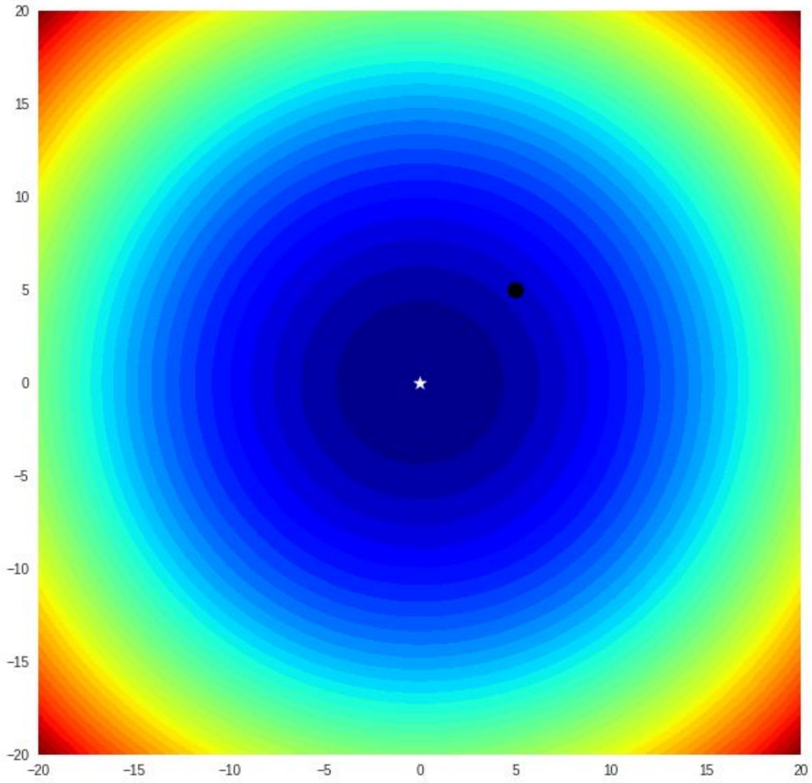
  Find the best solution in the offspring

$$x_{best}^{(t)} = \underset{x_i^{(t)} \in \text{ offspring}}{\operatorname{argmin}} \; F(x_i^{(t)})$$

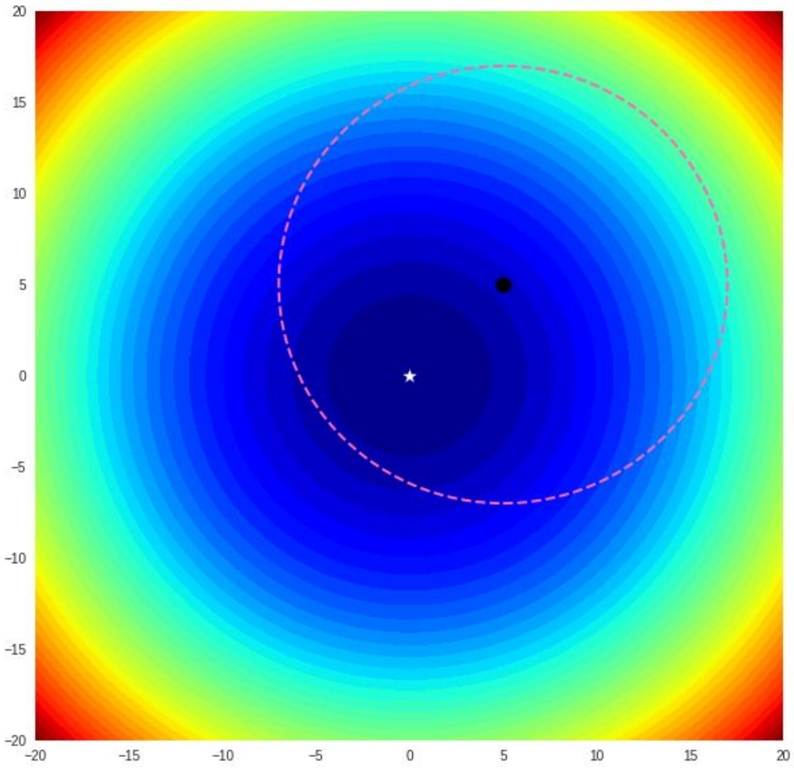
  $x_{best}^{(t)}$ compete with $m^{(t)}$ and update step-size

  if $\mathrm{F}(x_{best}^{(t)}) \leq \mathrm{F}(m^{(t)})$ **then**
    $m^{(t+1)} \leftarrow x_{best}^{(t)}$
    $\sigma^{(t+1)} \leftarrow \sigma^{(t)} c_{inc}$
  else
    $m^{(t+1)} \leftarrow m^{(t)}$
    $\sigma^{(t+1)} \leftarrow \sigma^{(t)} c_{dec}$
  end if
end for

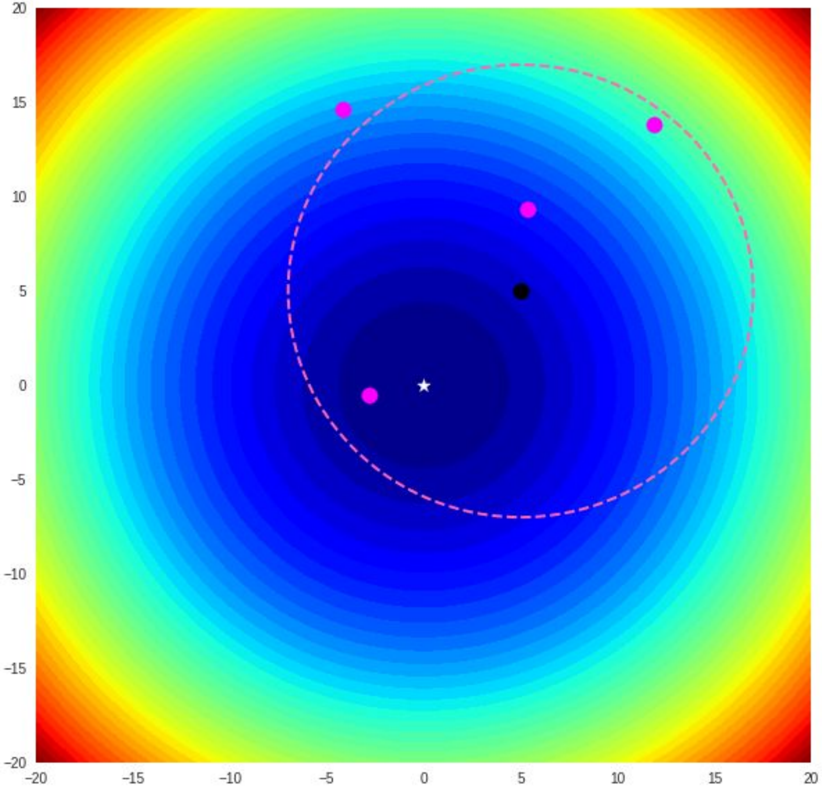
steps 0:
<Figure size 576x396 with 0 Axes>



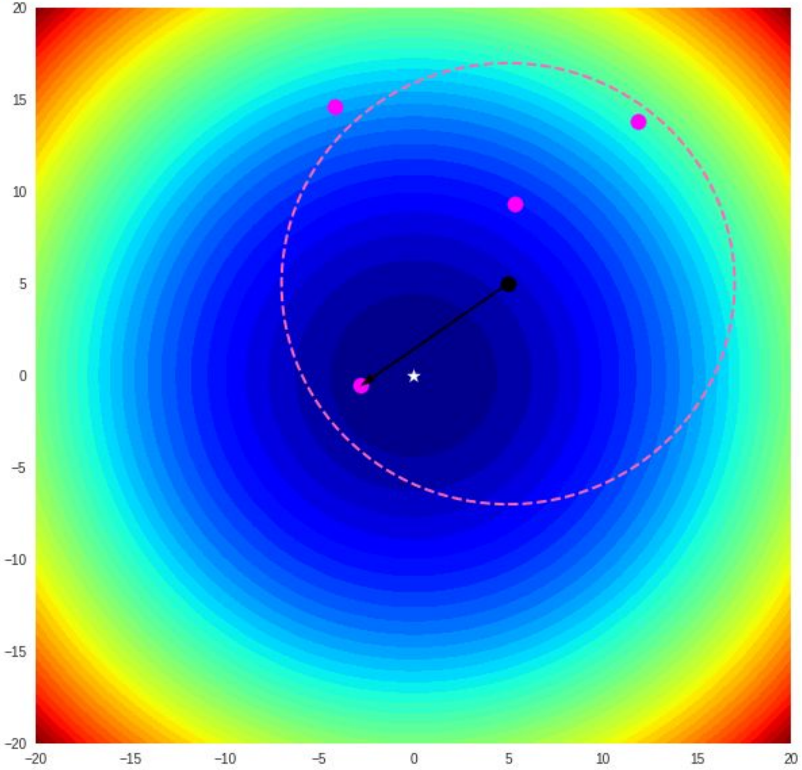
steps 1:



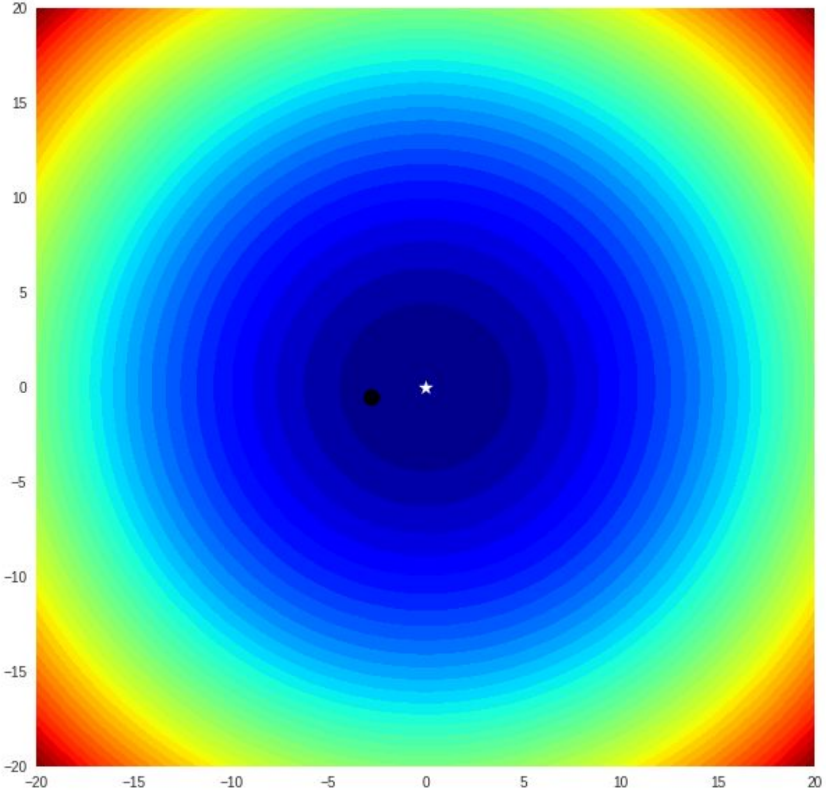
steps 2:



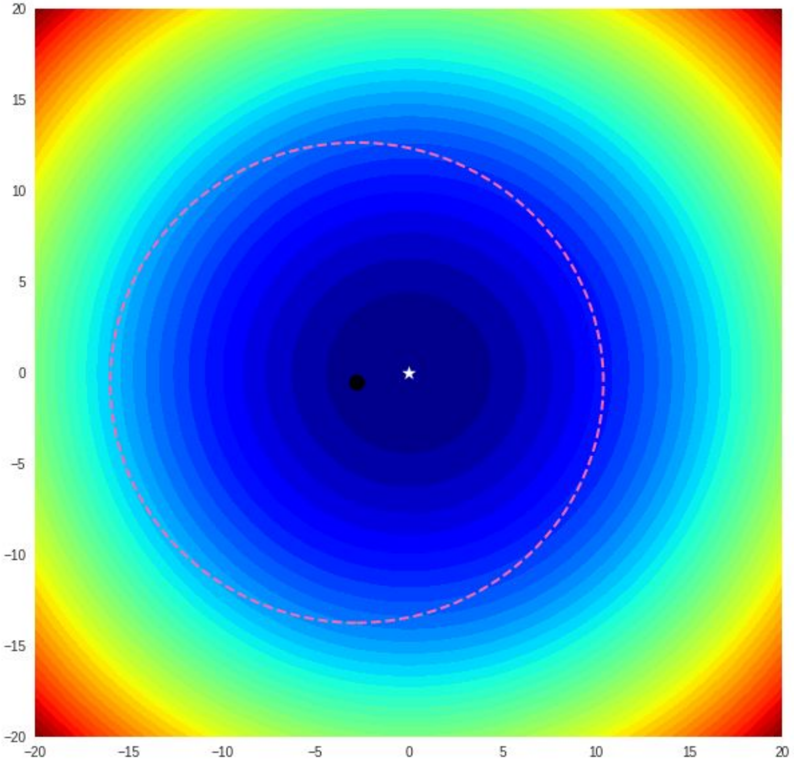
steps 3:



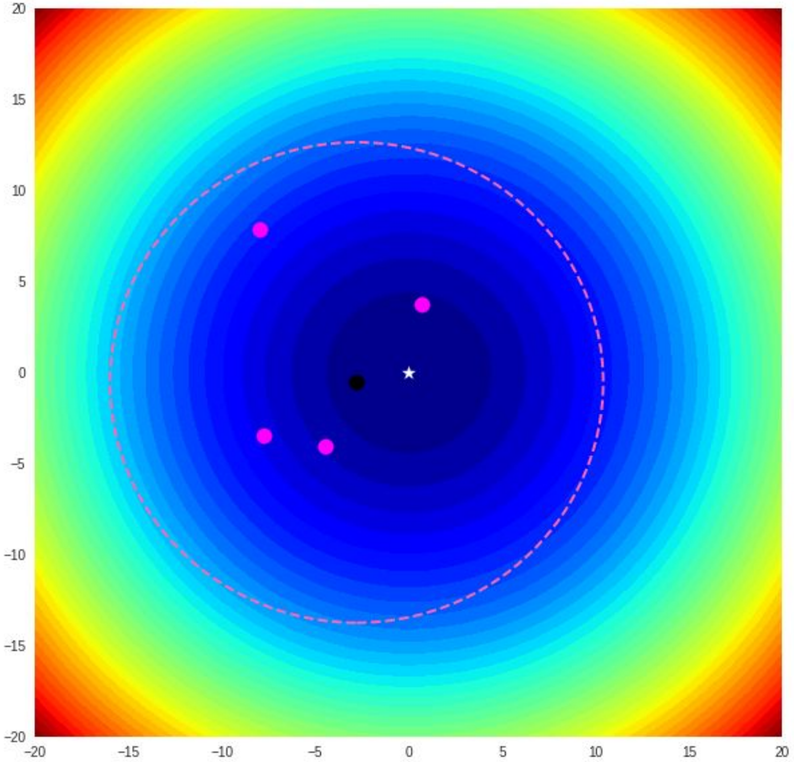
steps 4:



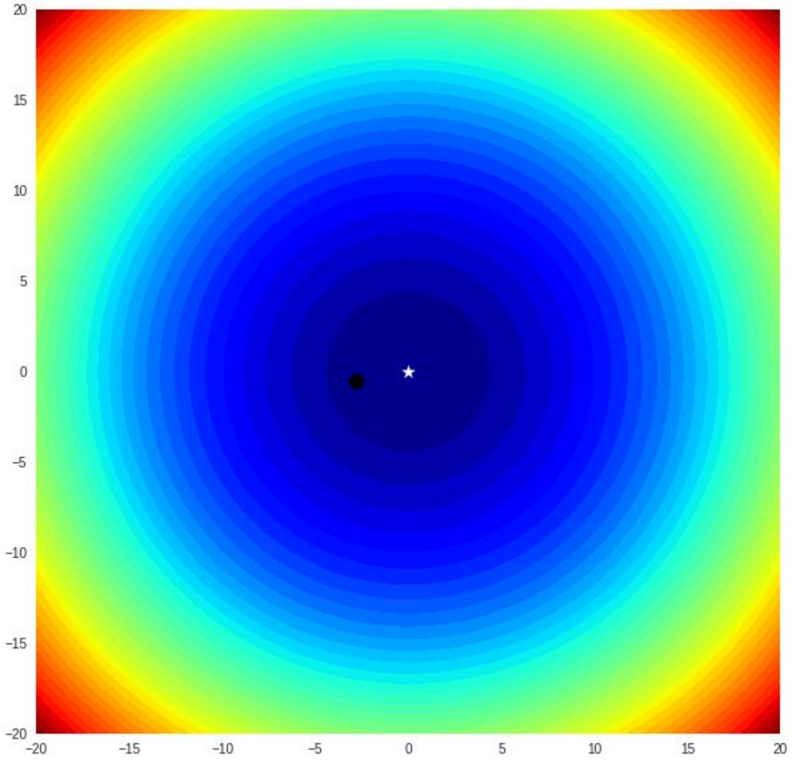
steps 5:

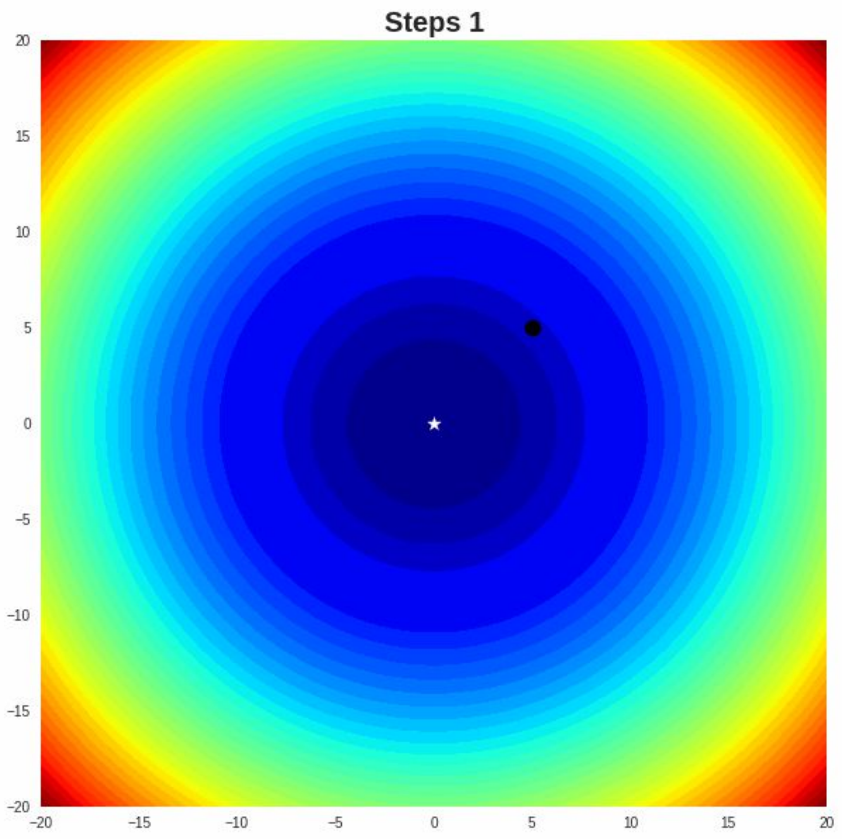


steps 6:



steps 7:




Steps 1
20
15
10
5
0
−5
−10
−15
−20
−20
−15
−10
−5
0
5
10
15
20

ES Algorithm

Algorithm 1 $(\mu/\rho,^{+}\lambda)$-Self-Adaptation ES

Input: $\rho, \lambda, \mu \in \mathbb{N}_+$

Initialize $\left(\mathcal{P}_\mu^0 \leftarrow \left\{\left(\mathbf{x}_i^0, \mathbf{s}_i^0, F(\mathbf{x}_i^0)\right), \quad i = 1, 2, \ldots, \mu\right\}\right)$

$g \leftarrow 0$

while *Termination_Condition is not satisfied* **do**

  $\widetilde{\mathcal{P}}_\lambda^g \leftarrow \emptyset$

  for $i \leftarrow 1$ **to** μ **do**

    $(x_i, s_i) \leftarrow$ **recombine** (**select_mates** $(\mathcal{P}_\mu^g, \rho)$)

    $\widetilde{s}_i \leftarrow$ **s_mutation**(s_i)

    $\widetilde{x}_i \leftarrow$ **x_mutation**(x_i)

    $\widetilde{F}_i \leftarrow F(\widetilde{x}_i)$

  end

  $\widetilde{\mathcal{P}}_\lambda^g \leftarrow \left\{\left(\widetilde{x}_i, \widetilde{s}_i, \widetilde{F}_i\right), \quad i = 1, 2, \ldots, \lambda\right\}$

  switch *selection_type* **do**

    case (μ, λ) **do**

      $\mathcal{P}_\mu^{g+1} \leftarrow$ **selection**$(\widetilde{\mathcal{P}}_\lambda^g, \mu)$

    end

    case $(\mu + \lambda)$ **do**

      $\mathcal{P}_\mu^{g+1} \leftarrow$ **selection**$(\widetilde{\mathcal{P}}_\lambda^g, \mathcal{P}_\mu^g, \mu)$

    end

  end

  $g \leftarrow g + 1$

end

Summary

Encoding	Real vectors
Recombination	Discrete or intermediate
Mutation	Random additive perturbation (uniform, Gaussian, Cauchy)
Parents selection	Uniformly random
Survivors selection	(μ,λ) or $(\mu+\lambda)$
Particularity	Self-adaptive mutation parameters

III. CEM

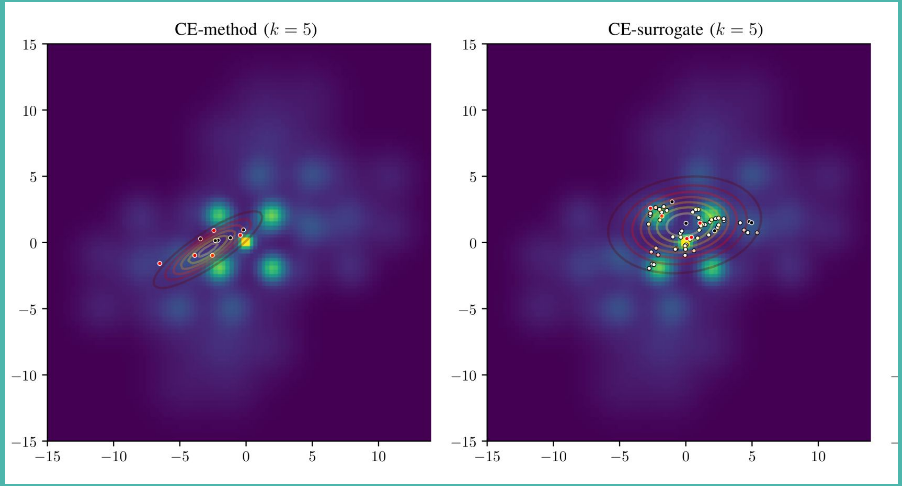


Fig:

Algorithm 3 CEM

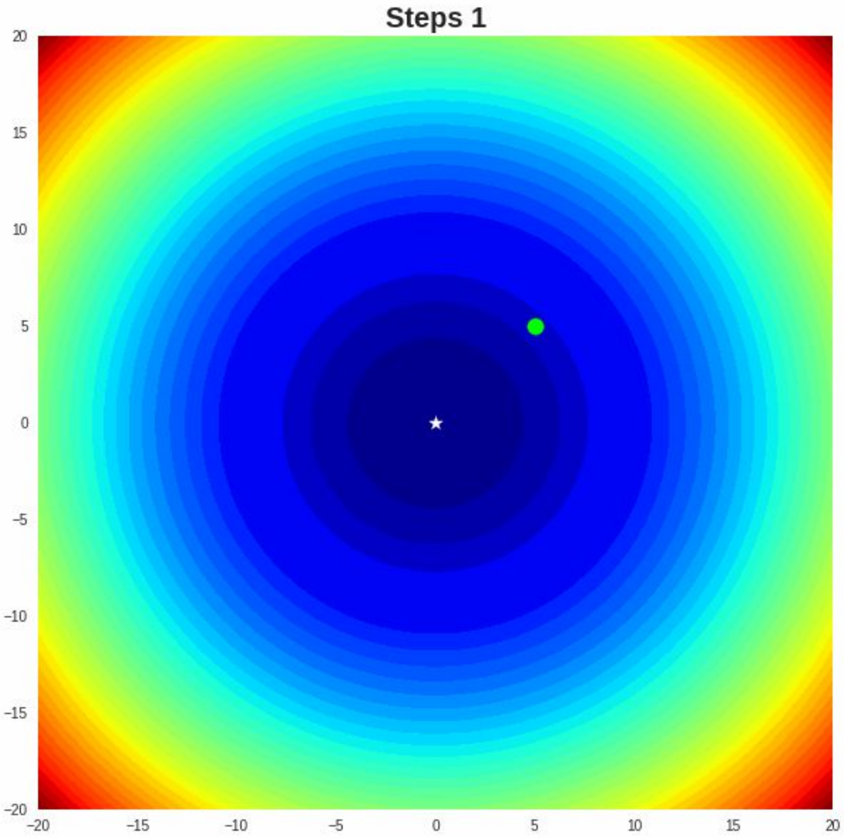
Hyperparameters: $\sigma_{init} \in R_{>0}$
Input: function F, vector $\mu_0 \in \mathbb{R}^d$,
number of sample N, number of elite set N_e ,
Initialize: $\Sigma_0 = \sigma_{init}\mathbf{I}_d$

for $t = 0, \ldots, T-1$ **do**
  Sample N search points $x_1, \ldots, x_N$ from $\mathcal{N}(\mu_t, \Sigma_t)$
  Evaluate the samples $x_1, \ldots, x_N$ on F
  Select top N_e search points $(z_i)_{i=1,\ldots,N_e}$
  Update the parameters of the distribution

$$\mu_{t+1} = \frac{1}{N_e}\sum_{i=1}^{N_e} z_i$$

$$\Sigma_{t+1} = \frac{1}{N_e}\sum_{i=1}^{N_e}(z_i - \mu_{t+1})(z_i - \mu_{t+1})^T$$

end for


Steps 1
20
15
10
5
0
−5
−10
−15
−20
−20
−15
−10
−5
0
5
10
15
20

Some modify to prevent premature convergence

Algorithm 4 CEM

Hyperparameters: extra variance ϵ, $\sigma_{init} \in R_{>0}$
Input: function F, vector $\mu_0 \in \mathbb{R}^d$,
number of sample N, number of elite set N_e ,
Initialize: $\Sigma_0 = \sigma_{init}\mathbf{I}_d$
$(w_i)_{i=1,\dots,N_e}$, where $w_i = \frac{1}{N_e}$
or $w_i = \frac{\log(N_e+1)-\log(i)}{\sum_{i=1}^{N_e}\log(N_e+1)-\log(i)}$

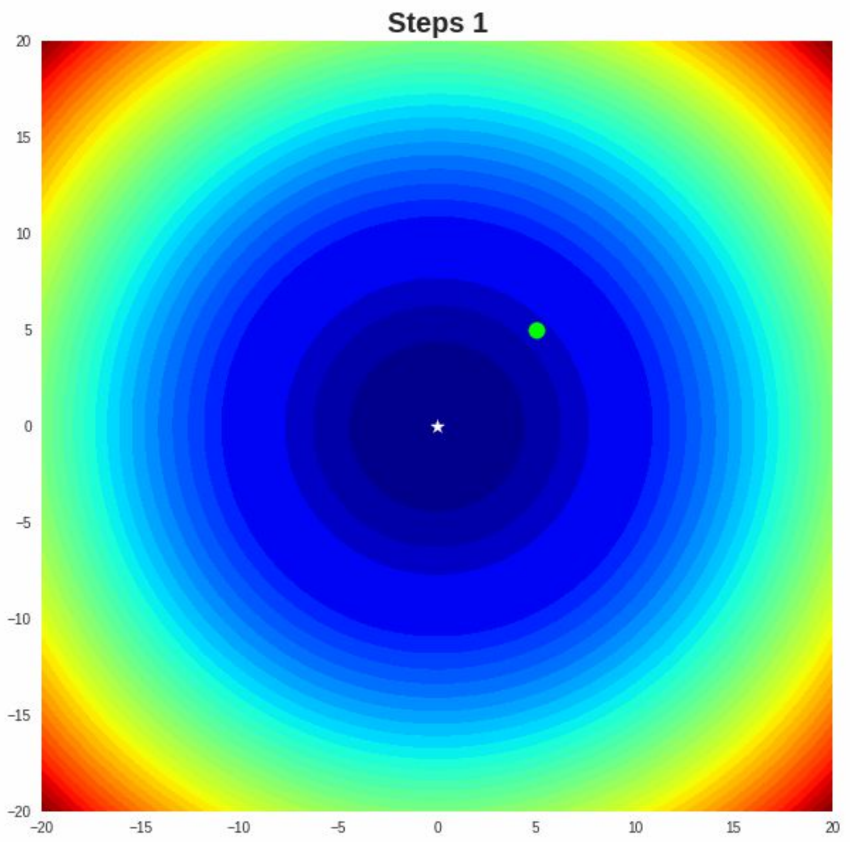
for $t = 0, \dots, T-1$ **do**
  Sample N search points $x_1, \dots, x_N$ from $\mathcal{N}(\mu_t, \Sigma_t)$
  Evaluate the samples $x_1, \dots, x_N$ on F
  Select top N_e search points $(z_i)_{i=1,\dots,N_e}$
  Update the parameters of the distribution

$$\mu_{t+1} = \sum_{i=1}^{N_e} w_i z_i$$

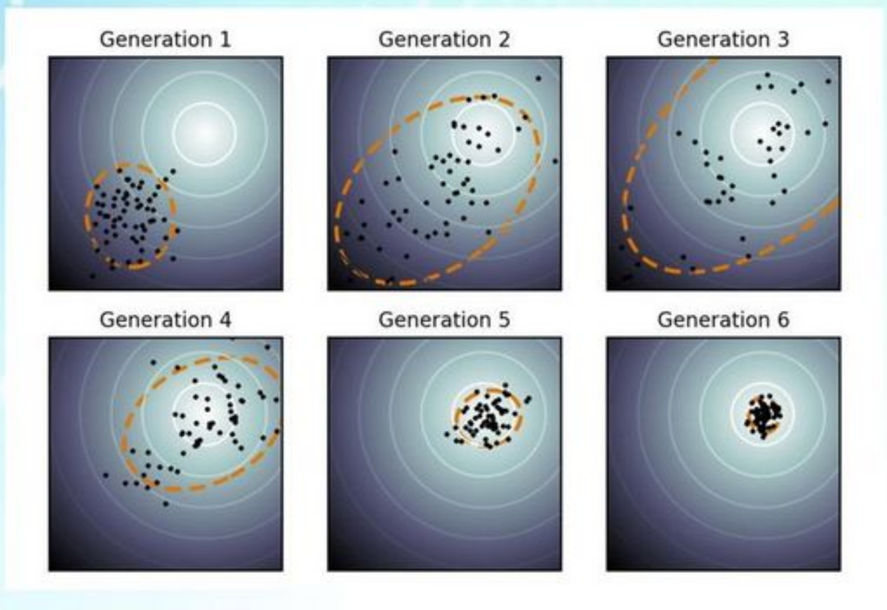
$$\boxed{\Sigma_{t+1} = \sum_{i=1}^{N_e} w_i (z_i - \mu_t)(z_i - \mu_t)^T + \epsilon \mathbf{I}_d}$$

end for

add to some noise
and use for μ_t updating Σ_{t+1}

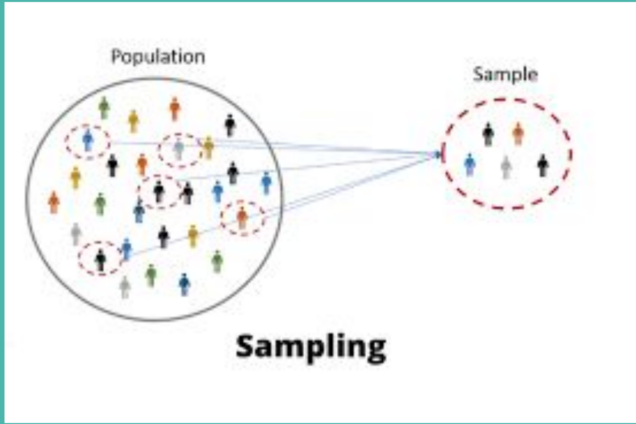

Steps 1
20
15
10
5
0
−5
−10
−15
−20
−20
−15
−10
−5
0
5
10
15
20

IV. CMA-ES



Generation 1
Generation 2
Generation 3
Generation 4
Generation 5
Generation 6

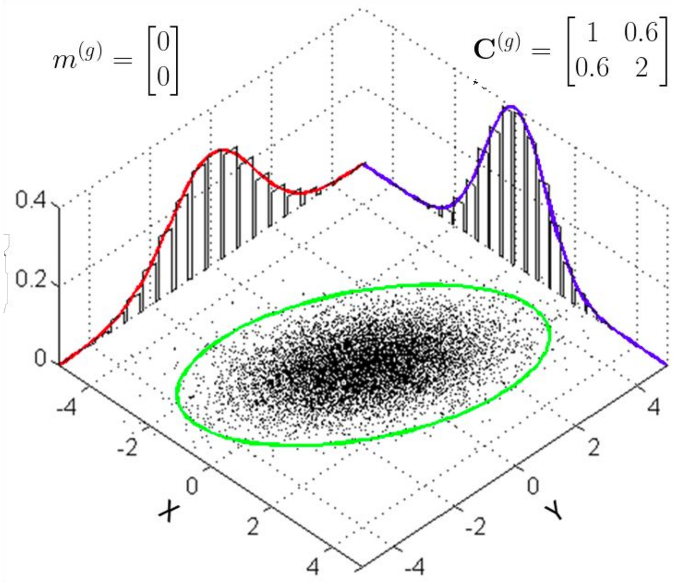
A. Sampling



Sampling

New search points is generated by sampling a multivariate normal distribution:

$$\mathbf{x}_k^{(g+1)} \sim \mathbf{m}^{(g)} + \sigma^{(g)} \mathcal{N}(\mathbf{0}, \mathbf{C}^{(g)}),$$
$$\forall k = 1, \ldots, \lambda$$



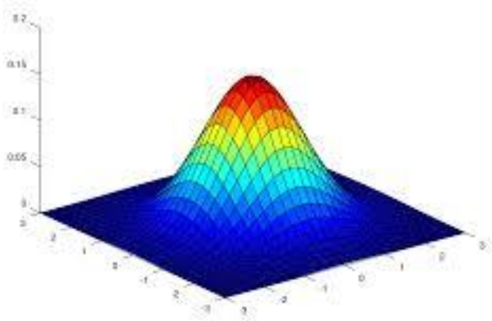
Where,

Step size $\sigma \in \mathbb{R}_+$ control the step length

The covariance matrix $\mathbf{C} \in \mathbb{R}^{n \times n}$ determines the shape of the distribution ellipsoid

Why Normal Distribution ?

- Approximates many natural phenomena so well



- Only stable distribution with finite variance
- Most convenient way to generate isotropic search points
- Maximum entropy distribution with finite variance

B. Selection and Recombination

Selection and Recombination

New mean value is computed as

$$\mathbf{m}^{(g+1)} = \mathbf{m}^{(g)} + c_m \sum_{i=1}^{\lambda} w_i \left(\mathbf{x}_{i:\lambda}^{(g+1)} - \mathbf{m}^{(g)}\right)$$

$$\sum_{i=1}^{\mu} w_i = 1, \quad w_1 \geq w_2 \geq \ldots \geq w_\mu > 0$$

Where,

$c_{\mathrm{m}} \leq 1$ is a learning rate, usually set to 1.

$w_{i=1\ldots\mu} \in \mathbb{R}_{>0}$, positive weight coefficients for recombination

$\{x_{i:\lambda} \mid i = 1 \ldots \lambda\} = \{x_i \mid i = 1 \ldots \lambda\}$ and $f(x_{1:\lambda}) \leq \cdots \leq f(x_{\mu:\lambda}) \leq f(x_{\mu+1:\lambda}) \leq \cdots,$

Selection and Recombination

- Intermediate recombination:

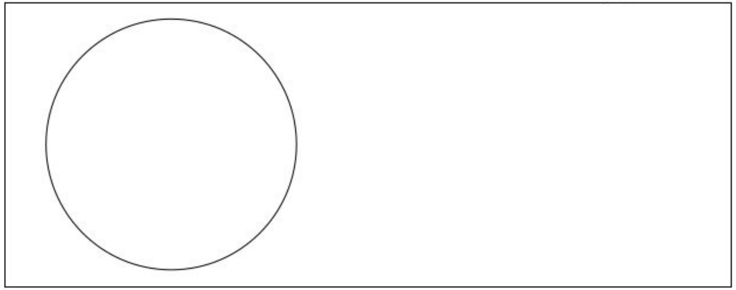
$$w_m := \begin{cases} \frac{1}{\mu}, & \text{for } 1 \leq m \leq \mu, \\ 0, & \text{otherwise,} \end{cases}$$

- Weighted recombination:

$$w_m := \begin{cases} \frac{\ln\left(\frac{\lambda+1}{2}\right) - \ln m}{\sum_{k=1}^{\mu}\left(\ln\left(\frac{\lambda+1}{2}\right) - \ln k\right)}, & \text{for } 1 \leq m \leq \mu, \\ 0, & \text{otherwise} \end{cases}$$

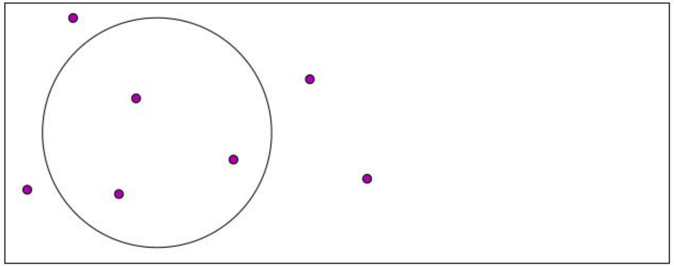
$$\mu_{\text{eff}} = \left(\frac{\|\boldsymbol{w}\|_1}{\|\boldsymbol{w}\|_2}\right)^2 = \frac{\|\boldsymbol{w}\|_1^2}{\|\boldsymbol{w}\|_2^2} = \frac{\left(\sum_{i=1}^{\mu} |w_i|\right)^2}{\sum_{i=1}^{\mu} w_i^2} = \frac{1}{\sum_{i=1}^{\mu} w_i^2}$$

$$\boldsymbol{m} \leftarrow \boldsymbol{m} + \sigma \boldsymbol{y}_w, \quad \boldsymbol{y}_w = \sum_{i=1}^{\mu} w_i \boldsymbol{y}_{i:\lambda}, \quad \boldsymbol{y}_i \sim \mathcal{N}_i(\mathbf{0}, \mathbf{C})$$



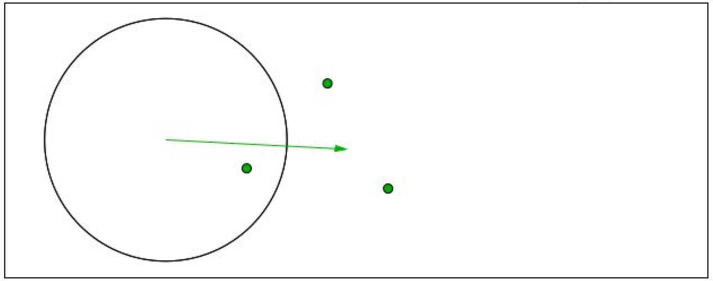
initial distribution, $\mathbf{C} = \mathbf{I}$

$$\boldsymbol{m} \leftarrow \boldsymbol{m} + \sigma \boldsymbol{y}_w, \quad \boldsymbol{y}_w = \sum_{i=1}^{\mu} w_i \boldsymbol{y}_{i:\lambda}, \quad \boldsymbol{y}_i \sim \mathcal{N}_i(\mathbf{0}, \mathbf{C})$$



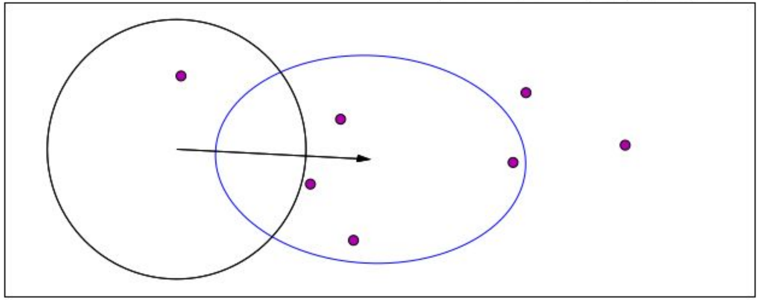
initial distribution, $\mathbf{C} = \mathbf{I}$

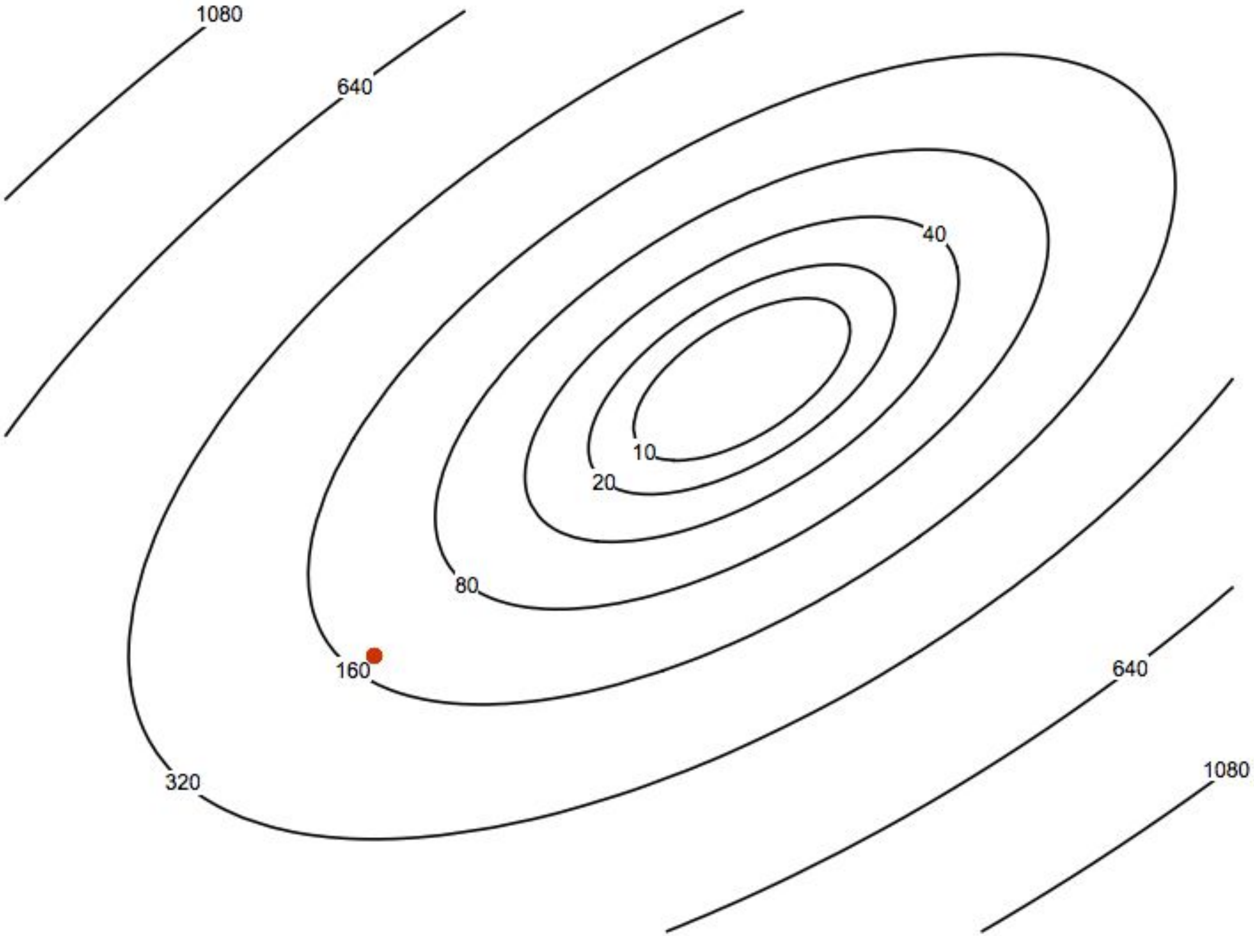
$$\boldsymbol{m} \leftarrow \boldsymbol{m} + \sigma \boldsymbol{y}_w, \quad \boldsymbol{y}_w = \sum_{i=1}^{\mu} w_i \boldsymbol{y}_{i:\lambda}, \quad \boldsymbol{y}_i \sim \mathcal{N}_i(\mathbf{0}, \mathbf{C})$$



$\boldsymbol{y}_w$, movement of the population mean $\boldsymbol{m}$ (disregarding σ)

$$\boldsymbol{m} \leftarrow \boldsymbol{m} + \sigma \boldsymbol{y}_w, \quad \boldsymbol{y}_w = \sum_{i=1}^{\mu} w_i \boldsymbol{y}_{i:\lambda}, \quad \boldsymbol{y}_i \sim \mathcal{N}_i(\mathbf{0}, \mathbf{C})$$





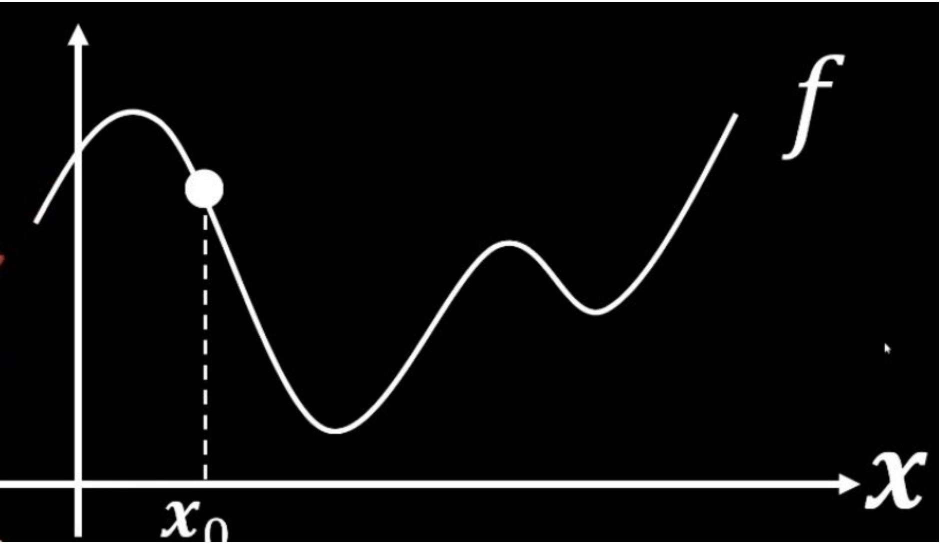
1080
640
40
10
20
80
160
320
640
1080

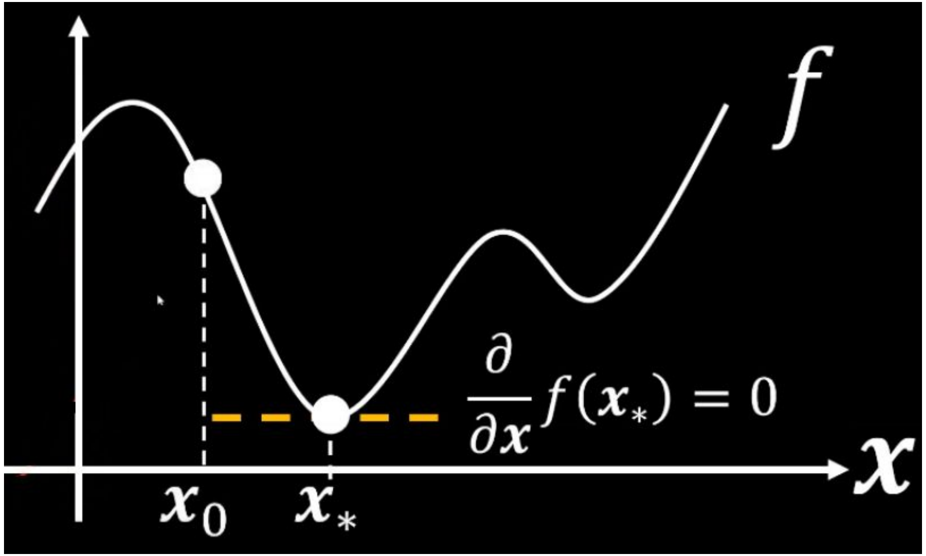
C. Adapting the Covariance Matrix

Similar to the approximation of the inverse Hessian matrix

Adaptation of the covariance matrix amounts to learning a second order model of the underlying objective function similar to the approximation of the inverse Hessian matrix in the quasi-Newton method in classical optimization.

Similar to the approximation of the inverse Hessian matrix





Similar to the approximation of the inverse Hessian matrix

$$\frac{\partial}{\partial \boldsymbol{x}} f(\boldsymbol{x}_i + \Delta \boldsymbol{x}) = 0$$

$$\underbrace{\frac{\partial}{\partial \boldsymbol{x}} f(\boldsymbol{x}_i)}_{\text{Gradient}} + \underbrace{\frac{\partial}{\partial \boldsymbol{x}} \frac{\partial}{\partial \boldsymbol{x}} f(\boldsymbol{x}_i)}_{\text{Hessian}} \Delta \boldsymbol{x} = 0$$

Similar to the approximation of the inverse Hessian matrix

$$\frac{\partial}{\partial \boldsymbol{x}} f(\boldsymbol{x}_i + \Delta \boldsymbol{x}) = 0$$

$$\boldsymbol{g} + \boldsymbol{H}\Delta \boldsymbol{x} = 0$$

$$\frac{\partial}{\partial \boldsymbol{x}} f(\boldsymbol{x}_i + \Delta \boldsymbol{x}) = 0$$

$$\Delta \boldsymbol{x} = -\boldsymbol{H}^{-1}\boldsymbol{g}$$

Estimating the Covariance Matrix From Scratch

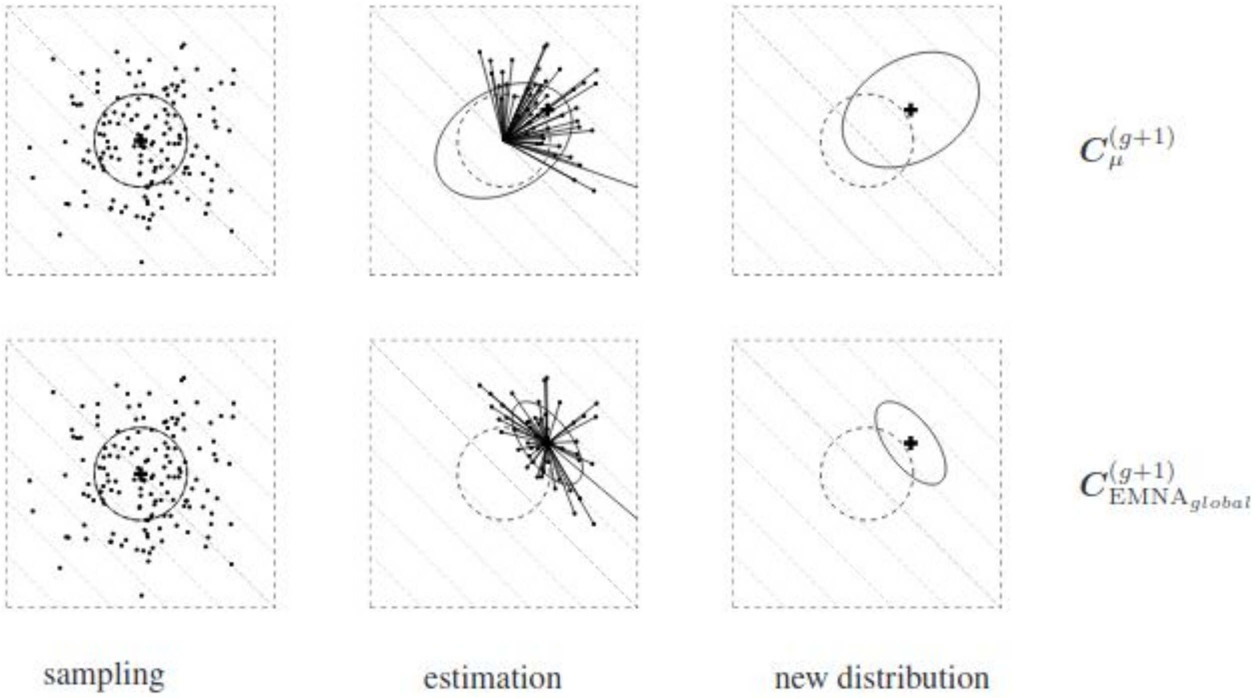
Estimate the distribution variance within the sampled points

$$\mathbf{C}_{EMNA_{global}}^{(g+1)} = \frac{1}{\sigma^{(g)2}} \sum_{i=1}^{\mu} w_i \left(\mathbf{x}_{i:\lambda}^{(g+1)} - \mathbf{m}^{(g+1)}\right) \left(\mathbf{x}_{i:\lambda}^{(g+1)} - \mathbf{m}^{(g+1)}\right)^T$$

Estimate variances of sampled steps

$$\mathbf{C}_{\mu}^{(g+1)} = \frac{1}{\sigma^{(g)2}} \sum_{i=1}^{\mu} w_i \left(\mathbf{x}_{i:\lambda}^{(g+1)} - \mathbf{m}^{(g)}\right) \left(\mathbf{x}_{i:\lambda}^{(g+1)} - \mathbf{m}^{(g)}\right)^T$$

Estimating the Covariance Matrix From Scratch



Rank-μ-Update

To achieve fast search, the population size must be small

⟶ It is not impossible to get a reliable estimator for good covariance matrix

➡ Information from previous generations is used additionally.

$$\mathbf{C}^{(g+1)} = \left(1 - c_\mu \sum w_i\right) \mathbf{C}^{(g)} + c_\mu \sum_{i=1}^{\lambda} w_i \mathbf{y}_{i:\lambda}^{(g+1)} \mathbf{y}_{i:\lambda}^{(g+1)T}$$

Where,

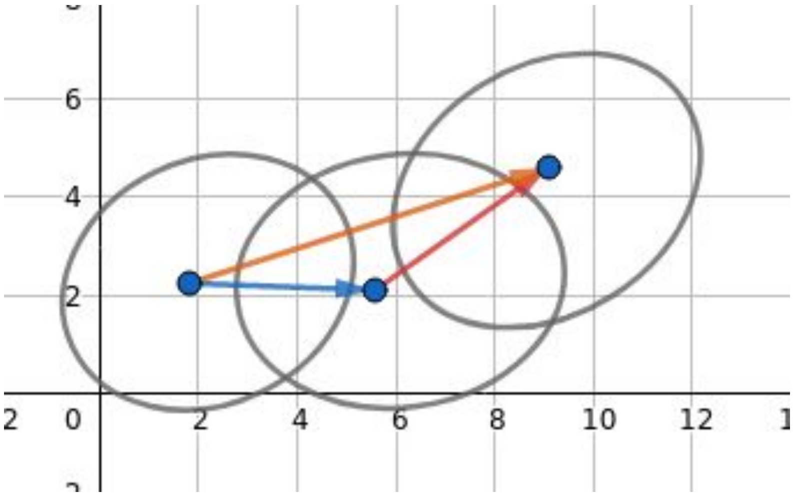
$$\mathbf{y}_{i:\lambda}^{(g+1)} = \left(\mathbf{x}_{i:\lambda}^{(g+1)} - \mathbf{m}^{(g)}\right) / \sigma^{(g)}$$

$c_\mu \leq 1$ learning rate

Cumulation The Evolution Path

Evolution Path: Conceptually, the evolution path is the search path the strategy takes over a number of generation steps. It can be expressed as a sum of consecutive steps of the mean m.

History information is accumulated in the evolution path



Cumulation The Evolution Path

- Cumulation is a widely used technique and also know as
  - **Exponential smoothing** in time series, forecasting
  - Exponentially weighted moving average
  - Iterate averaging in stochastic approximation
  - Momentum in the back-propagation algorithm for ANNs
  - . . .

The simplest form of **exponential smoothing** is given by the formulas:

$$s_0 = x_0$$
$$s_t = \alpha x_t + (1 - \alpha) s_{t-1}, \quad t > 0$$

where α is the *smoothing factor*, and $0 < \alpha < 1$.

Rank-One-Update

For any positive definite symmetric $\mathbf{C}$,

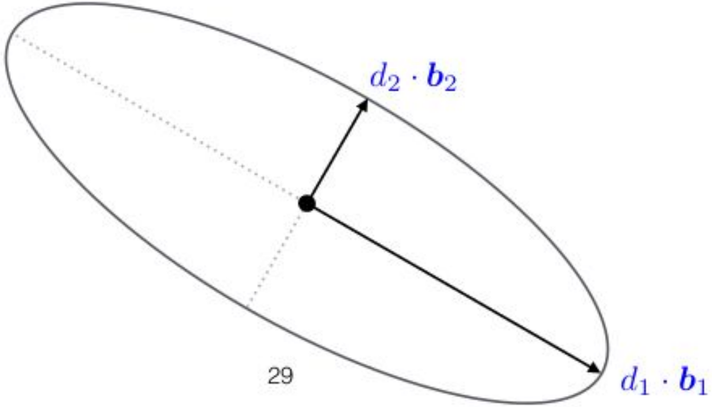
$$\mathbf{C} = d_1^2 \boldsymbol{b}_1 \boldsymbol{b}_1^{\mathrm{T}} + \cdots + d_N^2 \boldsymbol{b}_N \boldsymbol{b}_N^{\mathrm{T}}$$

d_i: square root of the eigenvalue of $\mathbf{C}$

$\boldsymbol{b}_i$: eigenvector of $\mathbf{C}$, corresponding to d_i

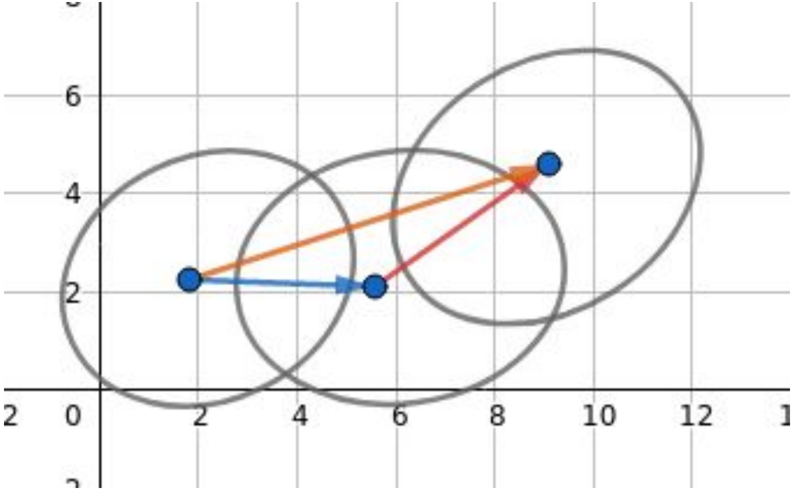
The multivariate normal distribution $\mathcal{N}(\boldsymbol{m}, \mathbf{C})$

$$\mathcal{N}(\boldsymbol{m}, \mathbf{C}) \sim \boldsymbol{m} + \mathcal{N}(0, d_1^2)\boldsymbol{b}_1 + \cdots + \mathcal{N}(0, d_N^2)\boldsymbol{b}_N$$



Rank-One-Update

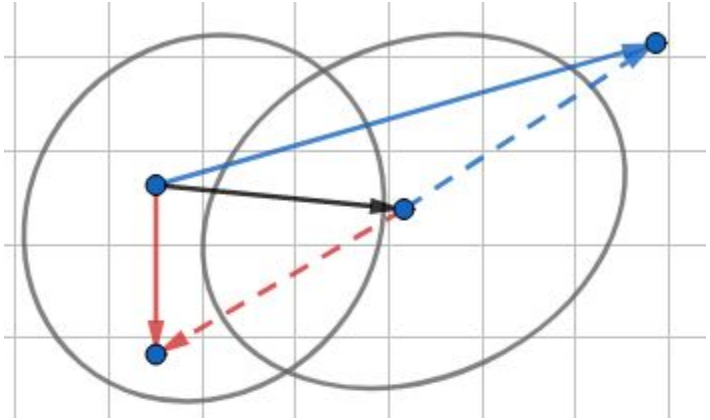
The adaptation increases the likelihood of successful steps



$$C^{(g+1)} = (1 - c_1)C^{(g)} + c_1 \, \boldsymbol{y}_{g+1} \boldsymbol{y}_{g+1}{}^{\top}$$

Rank-One-Update

Because $\mathbf{y}\mathbf{y}^T = (-\mathbf{y})(-\mathrm{y})^T$ the sign information is lost.



$$\mathbf{p}_c^{(g+1)} = (1 - c_c)\mathbf{p}_c^{(g)} + \sqrt{c_c(2 - c_c)\mu_{eff}}\,\frac{\mathbf{m}^{(g+1)} - \mathbf{m}^{(g)}}{\sigma^{(g)}}$$

$$\mathbf{C}^{(g+1)} = (1 - c_1)\mathbf{C}^{(g)} + c_1\mathbf{p}_c^{(g+1)}\mathbf{p}_c^{(g+1)T}$$

Combining Rank-μ-Update and Cumulation

CMA update of the covariance matrix combines Rank-μ-Update and Rank-One-Update

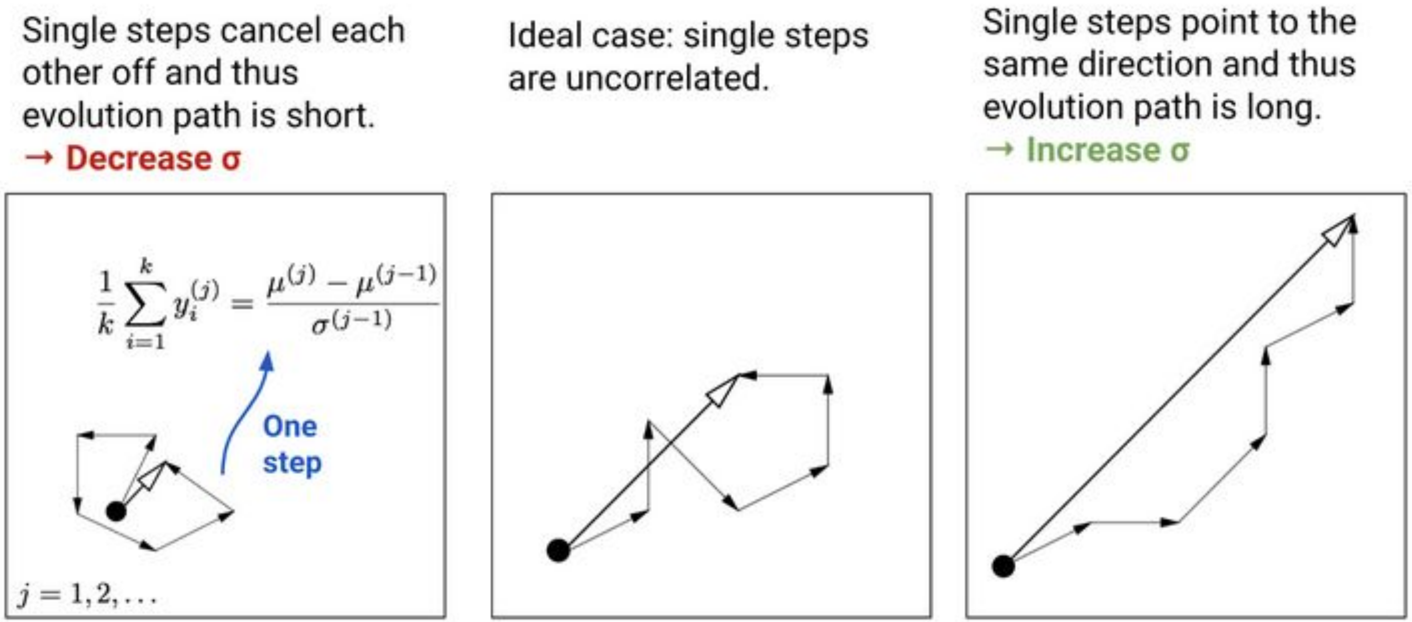
$$
\begin{aligned}
\boldsymbol{C}^{(g+1)} = \; & (1 \underbrace{- c_1 - c_\mu \textstyle\sum w_j)}_{\text{can be close or equal to 0}} \boldsymbol{C}^{(g)} \\
& + c_1 \underbrace{\boldsymbol{p}_c^{(g+1)} \boldsymbol{p}_c^{(g+1)^\mathsf{T}}}_{\text{rank-one update}} + c_\mu \underbrace{\sum_{i=1}^{\lambda} w_i \, \boldsymbol{y}_{i:\lambda}^{(g+1)} \left(\boldsymbol{y}_{i:\lambda}^{(g+1)}\right)^\mathsf{T}}_{\text{rank-}\mu\text{ update}}
\end{aligned}
$$

D. Step-Size Control

Step-Size Control

A larger step size leads to faster parameter update

→ Utilize an evolution path



Step-Size Control

Constructs an conjugate *evolution path* $\mathbf{p}_\sigma$ by summing up a consecutive sequence of moving steps

$$\mathbf{p}_\sigma^{(g+1)} = (1 - c_\sigma)\mathbf{p}_\sigma^{(g)} + \sqrt{c_\sigma(2 - c_\sigma)\mu_{eff}}\,\mathbf{C}^{(g)-\frac{1}{2}} \frac{\mathbf{m}^{(g+1)} - \mathbf{m}^{(g)}}{\sigma^{(g)}}$$

$$\ln \sigma^{(g+1)} = \ln \sigma^{(g)} + \frac{c_\sigma}{d_\sigma} \left(\frac{\|\mathbf{p}_\sigma^{(g+1)}\|}{\mathbf{E}\|\mathcal{N}(\mathbf{0}, \mathbf{I})\|} - 1 \right)$$

$$\sigma^{(g+1)} = \sigma^{(g)} \exp\left(\frac{c_\sigma}{d_\sigma} \left(\frac{\|\mathbf{p}_\sigma^{(g+1)}\|}{\mathbf{E}\|\mathcal{N}(\mathbf{0}, \mathbf{I})\|} - 1 \right) \right)$$

Algorithm

Algorithm 2 $(\mu/\mu_w, \lambda)$-CMA-ES [3]

Input: $\mathbf{m} \in \mathbb{R}^n, \lambda, \sigma \in \mathbb{R}_+$

Initialize: $\mathbf{C} = \mathbf{I}$, $\mathbf{p}_\sigma = \mathbf{0}$ và $\mathbf{p}_c = \mathbf{0}$

Set: $c_c \approx 4/n,\ c_\sigma \approx 4/n,\ c_1 \approx 2/n^2,\ c_\mu \approx \mu_w/n^2,\ c_1 + c_\mu \leq 1,\ d_\sigma \approx 1 + \sqrt{\frac{\mu_w}{n}}$ và $w_{i=1,\ldots,\lambda}$ sao cho $\mu_w = \frac{1}{\sum_{i=1}^{\mu} w_i^2} \approx 0.3\lambda$

while *Termination_Condition is not satisfied* **do**

```
/* Lấy mẫu, sinh ra các phần tử mới */
```

$\mathbf{x}_i = \mathbf{m} + \sigma \mathbf{y}_i, \qquad \mathbf{y}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \forall i = 1, \ldots \lambda$

```
/* Cập nhật giá trị trung bình */
```

$\mathbf{m} \leftarrow \mathbf{m} + \sigma \mathbf{y}_w$, trong đó $\mathbf{y}_w = \sum_{i=1}^{\mu} \mathbf{y}_{i:\lambda}$

```
/* Cập nhật ma trận hiệp phương sai */
```

$\mathbf{p}_c \leftarrow (1 - c_c)\mathbf{p}_c + \mathbb{1}_{\{\|\mathbf{p}_c\| \leq 1\}} \sqrt{c_\sigma(2-\sigma)\mu_w}\,\mathbf{y}_w$

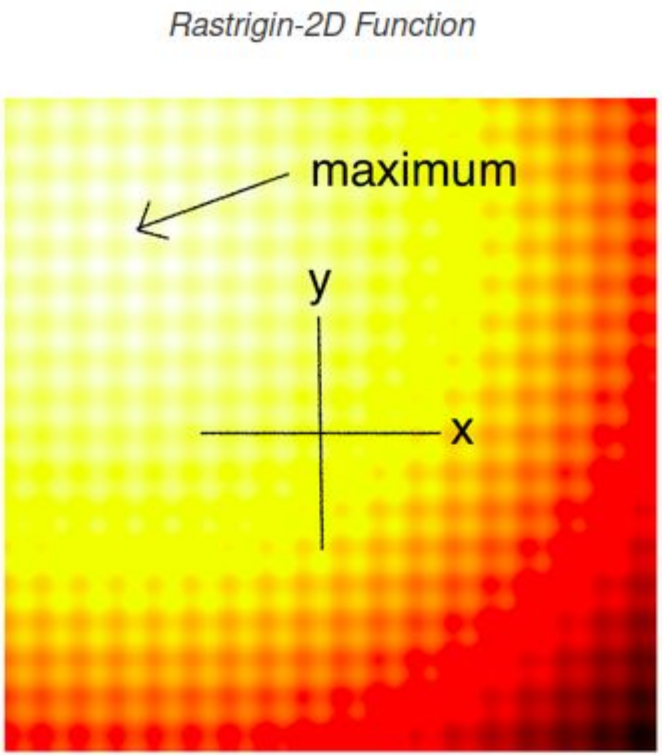
$\mathbf{C} \leftarrow (1 - c_1 - c_\mu)\mathbf{C} + c_1 \mathbf{p}_c \mathbf{p}_c^T + c_\mu \sum_{i=1}^{\mu} \mathbf{y}_{i:\lambda} \mathbf{y}_{i:\lambda}^T$

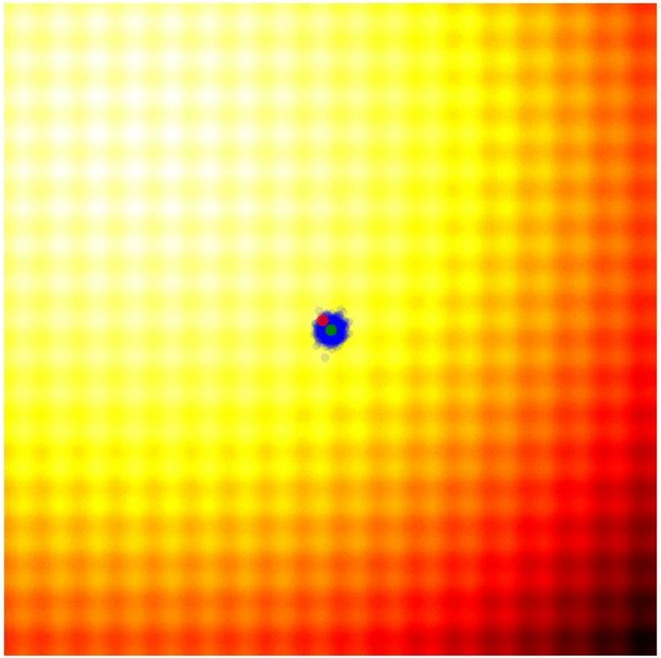
```
/* Cập nhật bước di chuyển */
```

$\mathbf{p}_\sigma \leftarrow (1 - c_\sigma)\mathbf{p}_\sigma + \sqrt{c_\sigma(2-\sigma)\mu_w}\,\mathbf{C}^{-\frac{1}{2}} \mathbf{y}_w$

$\sigma \leftarrow \sigma \times \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|\mathbf{p}_\sigma\|}{\mathbf{E}\|\mathcal{N}(\mathbf{0},\mathbf{I})\|} - 1\right)\right)$

end





Link blog: A Visual Guide to Evolution Strategies

Advantage of CMA-ES

- Non-separable problem
- The derivative of objective function is not available
- High dimension problems (n large)
- Very large search space

CMA-ES Limitations

- Partly separable problem
- The derivative of objective function is easily available
- Small dimension ($n \ll 10$)
- Small running times (number off-evaluations $< 100n$)

Hans-Georg Beyer

A research professor at the Research Center Business Informatics

https://homepages.fhv.at/hgb/

Simplify Your Covariance Matrix Adaptation Evolution Strategy (2017)

MA-ES

Algorithm 2 $(\mu/\mu_w, \lambda)$-CMA-ES [3]

Input: $\mathbf{m} \in \mathbb{R}^n, \lambda, \sigma \in \mathbb{R}_+$

Initialize: $\mathbf{C} = \mathbf{I}$, $\mathbf{p}_\sigma = \mathbf{0}$ và $\mathbf{p}_c = \mathbf{0}$

Set: $c_c \approx 4/n$, $c_\sigma \approx 4/n$, $c_1 \approx 2/n^2$, $c_\mu \approx \mu_w/n^2$, $c_1 + c_\mu \leq 1$, $d_\sigma \approx 1 + \sqrt{\frac{\mu_w}{n}}$ và $w_{i=1,\ldots,\lambda}$ sao cho $\mu_w = \frac{1}{\sum_{i=1}^{\mu} w_i^2} \approx 0.3\lambda$

while *Termination_ Condition is not satisfied* **do**

```
/* Lấy mẫu, sinh ra các phần tử mới */
```

$\mathbf{x}_i = \mathbf{m} + \sigma \mathbf{y}_i, \qquad \mathbf{y}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \forall i = 1, \ldots \lambda$

```
/* Cập nhật giá trị trung bình */
```

$\mathbf{m} \leftarrow \mathbf{m} + \sigma \mathbf{y}_w$, trong đó $\mathbf{y}_w = \sum_{i=1}^{\mu} \mathbf{y}_{i:\lambda}$

```
/* Cập nhật ma trận hiệp phương sai */
```

$\mathbf{p}_c \leftarrow (1 - c_c)\mathbf{p}_c + \mathbb{1}_{\{\|\mathbf{p}_c\| \leq 1\}} \sqrt{c_\sigma(2-\sigma)\mu_w} \mathbf{y}_w$

$\mathbf{C} \leftarrow (1 - c_1 - c_\mu)\mathbf{C} + c_1 \mathbf{p}_c \mathbf{p}_c^T + c_\mu \sum_{i=1}^{\mu} \mathbf{y}_{i:\lambda} \mathbf{y}_{i:\lambda}^T$

```
/* Cập nhật bước di chuyển */
```

$\mathbf{p}_\sigma \leftarrow (1 - c_\sigma)\mathbf{p}_\sigma + \sqrt{c_\sigma(2-\sigma)\mu_w} \mathbf{C}^{-\frac{1}{2}} \mathbf{y}_w$

$\sigma \leftarrow \sigma \times \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|\mathbf{p}_\sigma\|}{\mathbf{E}\|\mathcal{N}(\mathbf{0}, \mathbf{I})\|} - 1\right)\right)$

end

$(\mu/\mu_w, \lambda)$-CMA-ES

$$\textbf{Initialize}\left(\mathbf{y}^{(0)}, \sigma^{(0)}, g := 0, \mathbf{p}^{(0)} := \mathbf{0}, \mathbf{s}^{(0)} := \mathbf{0}, \mathbf{C}^{(0)} := \mathbf{I}\right) \quad \text{(C1)}$$

$$\textbf{Repeat} \quad \text{(C2)}$$

$$\mathbf{M}^{(g)} := \sqrt{\mathbf{C}^{(g)}} \quad \text{(C3)}$$

$$\textbf{For } l := 1 \textbf{ To } \lambda \quad \text{(C4)}$$

$$\tilde{\mathbf{z}}_l^{(g)} := \boldsymbol{\mathcal{N}}_l(\mathbf{0}, \mathbf{I}) \quad \text{(C5)}$$

$$\tilde{\mathbf{d}}_l^{(g)} := \mathbf{M}^{(g)} \tilde{\mathbf{z}}_l^{(g)} \quad \text{(C6)}$$

$$\tilde{\mathbf{y}}_l^{(g)} := \mathbf{y}^{(g)} + \sigma^{(g)} \tilde{\mathbf{d}}_l^{(g)} \quad \text{(C7)}$$

$$\tilde{f}_l^{(g)} := f\left(\tilde{\mathbf{y}}_l^{(g)}\right) \quad \text{(C8)}$$

$$\textbf{End} \quad \text{(C9)}$$

$$\textbf{SortOffspringPopulation} \quad \text{(C10)}$$

$$\mathbf{y}^{(g+1)} := \mathbf{y}^{(g)} + \sigma^{(g)} \left\langle \tilde{\mathbf{d}}^{(g)} \right\rangle_w \quad \text{(C11)}$$

$$\mathbf{s}^{(g+1)} := (1 - c_s)\mathbf{s}^{(g)} + \sqrt{\mu_{\text{eff}} c_s (2 - c_s)} \left\langle \tilde{\mathbf{z}}^{(g)} \right\rangle_w \quad \text{(C12)}$$

$$\mathbf{p}^{(g+1)} := (1 - c_p)\mathbf{p}^{(g)} + \sqrt{\mu_{\text{eff}} c_p (2 - c_p)} \left\langle \tilde{\mathbf{d}}^{(g)} \right\rangle_w \quad \text{(C13)}$$

$$\mathbf{C}^{(g+1)} := (1 - c_1 - c_w)\mathbf{C}^{(g)} + c_1 \mathbf{p}^{(g+1)} \left(\mathbf{p}^{(g+1)}\right)^{\mathrm{T}} + c_w \left\langle \tilde{\mathbf{d}}^{(g)} \left(\tilde{\mathbf{d}}^{(g)}\right)^{\mathrm{T}} \right\rangle_w \quad \text{(C14)}$$

$$\sigma^{(g+1)} := \sigma^{(g)} \exp\left[\frac{c_s}{d_\sigma}\left(\frac{\|\mathbf{s}^{(g+1)}\|}{\mathrm{E}\left[\|\boldsymbol{\mathcal{N}}(\mathbf{0}, \mathbf{I})\|\right]} - 1\right)\right] \quad \text{(C15)}$$

$$g := g + 1 \quad \text{(C16)}$$

$$\textbf{Until}(\text{termination condition(s) fulfilled}) \quad \text{(C17)}$$

$(\mu/\mu_w, \lambda)$-MA-ES

Initialize $\left(\mathbf{y}^{(0)}, \sigma^{(0)}, g := 0, \mathbf{s}^{(0)} := \mathbf{0}, \mathbf{M}^{(0)} := \mathbf{I}\right)$ (M1)

Repeat (M2)

For $l := 1$ **To** λ (M3)

$$\tilde{\mathbf{z}}_l^{(g)} := \boldsymbol{\mathcal{N}}_l(\mathbf{0}, \mathbf{I}) \quad \text{(M4)}$$

$$\tilde{\mathbf{d}}_l^{(g)} := \mathbf{M}^{(g)} \tilde{\mathbf{z}}_l^{(g)} \quad \text{(M5)}$$

$$\tilde{f}_l^{(g)} := f\left(\mathbf{y}^{(g)} + \sigma^{(g)} \tilde{\mathbf{d}}_l^{(g)}\right) \quad \text{(M6)}$$

End (M7)

SortOffspringPopulation (M8)

$$\mathbf{y}^{(g+1)} := \mathbf{y}^{(g)} + \sigma^{(g)} \left\langle \tilde{\mathbf{d}}^{(g)} \right\rangle_w \quad \text{(M9)}$$

$$\mathbf{s}^{(g+1)} := (1 - c_s)\mathbf{s}^{(g)} + \sqrt{\mu_{\text{eff}} c_s (2 - c_s)} \left\langle \tilde{\mathbf{z}}^{(g)} \right\rangle_w \quad \text{(M10)}$$

$$\mathbf{M}^{(g+1)} := \mathbf{M}^{(g)} \left[\mathbf{I} + \frac{c_1}{2}\left(\mathbf{s}^{(g+1)}\left(\mathbf{s}^{(g+1)}\right)^{\mathrm{T}} - \mathbf{I}\right) + \frac{c_w}{2}\left(\left\langle \tilde{\mathbf{z}}^{(g)}\left(\tilde{\mathbf{z}}^{(g)}\right)^{\mathrm{T}} \right\rangle_w - \mathbf{I}\right)\right] \quad \text{(M11)}$$

$$\sigma^{(g+1)} := \sigma^{(g)} \exp\left[\frac{c_s}{d_\sigma}\left(\frac{\left\|\mathbf{s}^{(g+1)}\right\|}{\mathrm{E}\left[\left\|\boldsymbol{\mathcal{N}}(\mathbf{0}, \mathbf{I})\right\|\right]} - 1\right)\right] \quad \text{(M12)}$$

$$g := g + 1 \quad \text{(M13)}$$

Until(termination condition(s) fulfilled) (M14)

$(\mu/\mu_w, \lambda)$-CMA-ES

Initialize $\left(\mathbf{y}^{(0)}, \sigma^{(0)}, g := 0, \mathbf{p}^{(0)} := \mathbf{0}, \mathbf{s}^{(0)} := \mathbf{0}, \mathbf{C}^{(0)} := \mathbf{I}\right)$ (C1)

Repeat (C2)

$$\mathbf{M}^{(g)} := \sqrt{\mathbf{C}^{(g)}} \quad \text{(C3)}$$

For $l := 1$ **To** λ (C4)

$$\tilde{\mathbf{z}}_l^{(g)} := \boldsymbol{\mathcal{N}}_l(\mathbf{0}, \mathbf{I}) \quad \text{(C5)}$$

$$\tilde{\mathbf{d}}_l^{(g)} := \mathbf{M}^{(g)} \tilde{\mathbf{z}}_l^{(g)} \quad \text{(C6)}$$

$$\tilde{\mathbf{y}}_l^{(g)} := \mathbf{y}^{(g)} + \sigma^{(g)} \tilde{\mathbf{d}}_l^{(g)} \quad \text{(C7)}$$

$$\tilde{f}_l^{(g)} := f\left(\tilde{\mathbf{y}}_l^{(g)}\right) \quad \text{(C8)}$$

End (C9)

SortOffspringPopulation (C10)

$$\mathbf{y}^{(g+1)} := \mathbf{y}^{(g)} + \sigma^{(g)} \left\langle \tilde{\mathbf{d}}^{(g)} \right\rangle_w \quad \text{(C11)}$$

$$\mathbf{s}^{(g+1)} := (1 - c_s)\mathbf{s}^{(g)} + \sqrt{\mu_{\text{eff}} c_s (2 - c_s)} \left\langle \tilde{\mathbf{z}}^{(g)} \right\rangle_w \quad \text{(C12)}$$

$$\mathbf{p}^{(g+1)} := (1 - c_p)\mathbf{p}^{(g)} + \sqrt{\mu_{\text{eff}} c_p (2 - c_p)} \left\langle \tilde{\mathbf{d}}^{(g)} \right\rangle_w \quad \text{(C13)}$$

$$\mathbf{C}^{(g+1)} := (1 - c_1 - c_w)\mathbf{C}^{(g)} + c_1 \mathbf{p}^{(g+1)}\left(\mathbf{p}^{(g+1)}\right)^{\mathrm{T}} + c_w \left\langle \tilde{\mathbf{d}}^{(g)}\left(\tilde{\mathbf{d}}^{(g)}\right)^{\mathrm{T}} \right\rangle_w \quad \text{(C14)}$$

$$\sigma^{(g+1)} := \sigma^{(g)} \exp\left[\frac{c_s}{d_\sigma}\left(\frac{\left\|\mathbf{s}^{(g+1)}\right\|}{\mathrm{E}\left[\left\|\boldsymbol{\mathcal{N}}(\mathbf{0}, \mathbf{I})\right\|\right]} - 1\right)\right] \quad \text{(C15)}$$

$$g := g + 1 \quad \text{(C16)}$$

Until(termination condition(s) fulfilled) (C17)

Removing the p and the C in CMA-ES

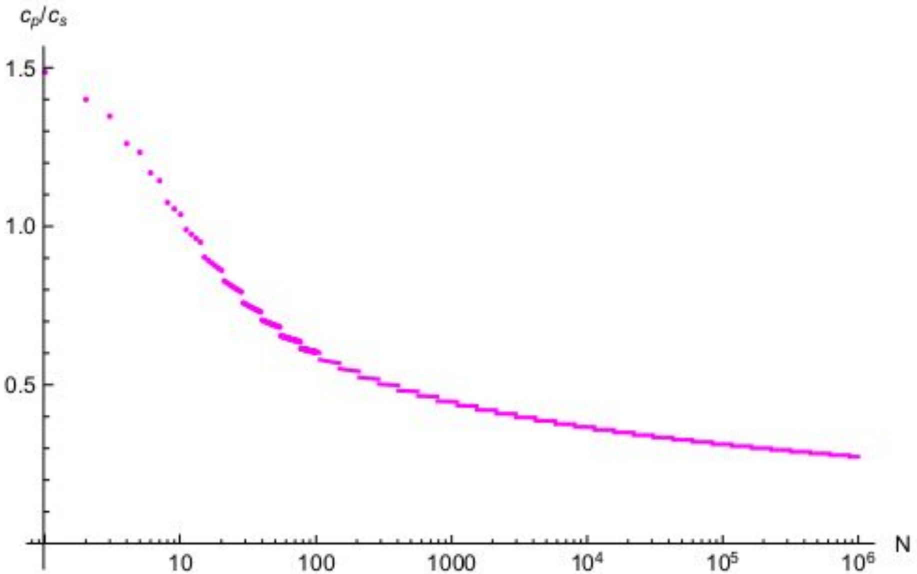
$$\begin{aligned}\mathbf{M}^{(g)}\mathbf{s}^{(g+1)} &= (1-c_s)\mathbf{M}^{(g)}\mathbf{s}^{(g)} \\ &\quad+\sqrt{\mu_{\text{eff}}c_s(2-c_s)}\left\langle \mathbf{M}^{(g)}\tilde{\mathbf{z}}^{(g)}\right\rangle_w \\ &= (1-c_s)\mathbf{M}^{(g)}\mathbf{s}^{(g)}+\sqrt{\mu_{\text{eff}}c_s(2-c_s)}\left\langle \tilde{\mathbf{d}}^{(g)}\right\rangle_w .\end{aligned}$$

Provided that $c_p = c_s$, $\Longrightarrow$ $c_p = c_s \Leftrightarrow \mathbf{M}^{(g)}\mathbf{s}^{(g)} = \mathbf{p}^{(g)} \Rightarrow \mathbf{M}^{(g)}\mathbf{s}^{(g+1)} = \mathbf{p}^{(g+1)}$

Provided that $\mathbf{M}^{(g+1)} \simeq \mathbf{M}^{(g)}$ asymptotically holds for $N \to \infty$, p can be drop

Removing the p and the C in CMA-ES

The c_p/c_s ratio is only a slightly decreasing function of N that does not deviate too much from 1. Therefore, one would not expect a much pronounced influence on the performance of the CMA-ES.



$$\lambda = 4 + \lfloor 3 \ln N \rfloor, \quad \mu = \left\lfloor \frac{\lambda}{2} \right\rfloor,$$

$$\mu_{\text{eff}} = \frac{1}{\sum_{m=1}^{\mu} w_m^2},$$

$$c_p = \frac{\mu_{\text{eff}}/N + 4}{2\mu_{\text{eff}}/N + N + 4},$$

$$c_s = \frac{\mu_{\text{eff}} + 2}{\mu_{\text{eff}} + N + 5}.$$

Removing the p and the C in CMA-ES

$$\mathbf{M}^{(g+1)}\left(\mathbf{M}^{(g+1)}\right)^{\mathrm{T}}$$
$$= \mathbf{M}^{(g)}\left[\mathbf{I} + c_1\left(\mathbf{s}^{(g+1)}\left(\mathbf{s}^{(g+1)}\right)^{\mathrm{T}} - \mathbf{I}\right)\right.$$
$$\left. + c_w\left(\left\langle \tilde{\mathbf{z}}^{(g)}\left(\tilde{\mathbf{z}}^{(g)}\right)^{\mathrm{T}}\right\rangle_w - \mathbf{I}\right)\right]\left(\mathbf{M}^{(g)}\right)^{\mathrm{T}}$$

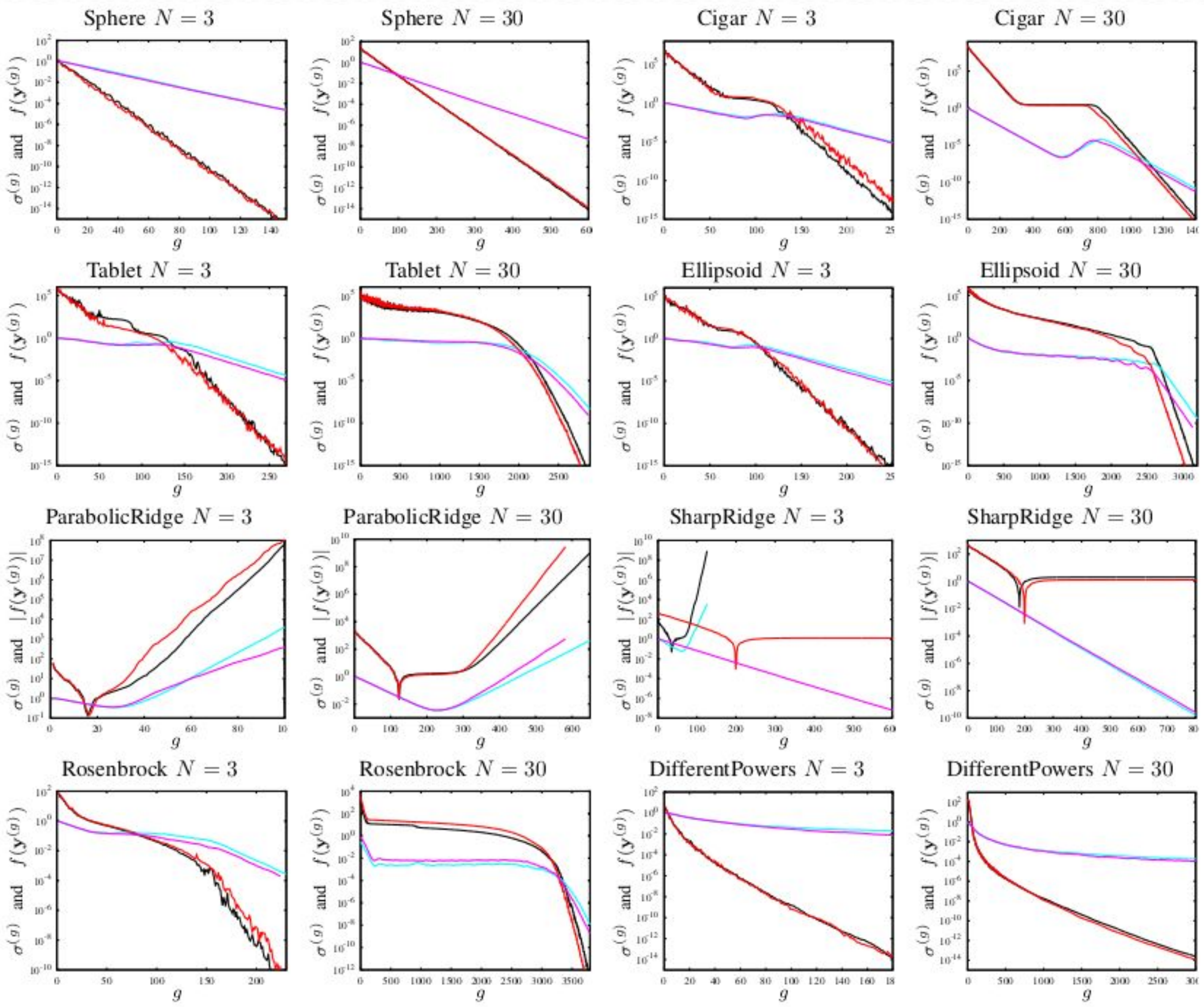
$$\mathbf{M}^{(g+1)} = \mathbf{M}^{(g)}\left[\mathbf{I} + \frac{c_1}{2}\left(\mathbf{s}^{(g+1)}\left(\mathbf{s}^{(g+1)}\right)^{\mathrm{T}} - \mathbf{I}\right)\right.$$
$$\left. + \frac{c_w}{2}\left(\left\langle \tilde{\mathbf{z}}^{(g)}\left(\tilde{\mathbf{z}}^{(g)}\right)^{\mathrm{T}}\right\rangle_w - \mathbf{I}\right) + \ldots\right]$$

and

$$c_1 = \frac{\alpha_{cov}}{(N+1.3)^2 + \mu_{\text{eff}}}$$

$$c_w = \min\left(1 - c_1,\ \alpha_{cov}\frac{\mu_{\text{eff}} + 1/\mu_{\text{eff}} - 2}{(N+2)^2 + \alpha_{cov}\mu_{\text{eff}}/2}\right)$$

CMA-ES vs MA-ES



CMA-ES vs MA-ES

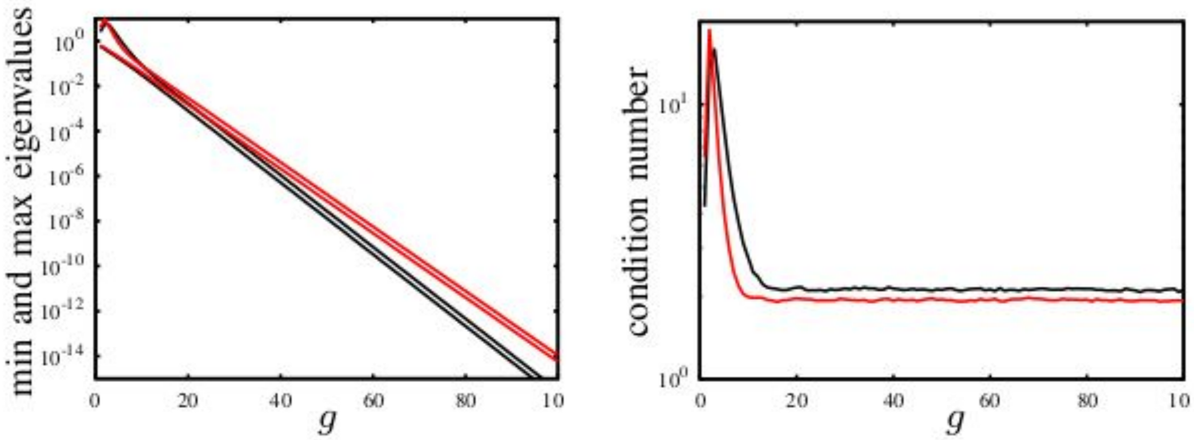


Fig. 8. Left figure: On the evolution of the minimal and the maximal eigenvalues of $\mathbf{C}$ (black curves) and $\mathbf{M}\mathbf{M}^{\mathrm{T}}$ (red curves) for a $(1800/1800_I, 3600)$-CMA-ES and $(1800/1800_I, 3600)$-MA-ES, respectively, on the $N = 30$-dimensional Sphere model. Right figure: Corresponding condition number dynamics.

Fast MA-ES (matrix x vector)

$$\mathbf{M}^{(g+1)} = \mathbf{M}^{(g)}\Big[\mathbf{I} + \frac{c_1}{2}\left(\mathbf{s}^{(g+1)}\left(\mathbf{s}^{(g+1)}\right)^{\mathrm{T}} - \mathbf{I}\right) + \frac{c_w}{2}\left(\left\langle \tilde{\mathbf{z}}^{(g)}\left(\tilde{\mathbf{z}}^{(g)}\right)^{\mathrm{T}}\right\rangle_w - \mathbf{I}\right) + \dots\Big]$$

$O(n^3)$

$$\mathbf{M}^{(t+1)} \leftarrow \left(1 - \frac{c_1}{2} - \frac{c_\mu}{2}\right)\mathbf{M}^{(t)} + \frac{c_1}{2}\boldsymbol{d}_\sigma^{(t)}(\boldsymbol{p}_\sigma^{(t)})^T + \frac{c_\mu}{2}\sum_{i=1}^{\mu} w_i \boldsymbol{d}_{i:\lambda}^{(t)}(\boldsymbol{z}_{i:\lambda}^{(t)})^T,$$

$O(n^2)$

LM MA ES

$$\mathbf{M}^{(t+1)} \leftarrow \mathbf{M}^{(t)} \left[\mathbf{I} + \frac{c_1}{2} \left(\boldsymbol{p}_\sigma^{(t+1)} (\boldsymbol{p}_\sigma^{(t+1)})^T - \mathbf{I} \right) + \frac{c_\mu}{2} \left(\sum_{i=1}^{\mu} \boldsymbol{w}_i \boldsymbol{z}_{i:\lambda}^{(t)} (\boldsymbol{z}_{i:\lambda}^{(t)})^T - \mathbf{I} \right) \right];$$

By omitting the rank-μ update for the sake of simplicity (i.e., by setting cμ = 0), we obtain:

$$\mathbf{M}^{(1)} \leftarrow \mathbf{I} + \frac{c_1}{2} \left(\boldsymbol{p}_\sigma^{(1)} (\boldsymbol{p}_\sigma^{(1)})^T - \mathbf{I} \right) = \left(1 - \frac{c_1}{2} \right) \mathbf{I} + \frac{c_1}{2} \boldsymbol{p}_\sigma^{(1)} (\boldsymbol{p}_\sigma^{(1)})^T$$

$$\Downarrow$$

$$\boldsymbol{d}_i^{(1)} = \mathbf{M}^{(1)} \boldsymbol{z}_i^{(1)} = \left(\left(1 - \frac{c_1}{2} \right) \mathbf{I} + \frac{c_1}{2} \boldsymbol{p}_\sigma^{(1)} (\boldsymbol{p}_\sigma^{(1)})^T \right) \boldsymbol{z}_i^{(1)} = \boldsymbol{z}_i^{(1)} \left(1 - \frac{c_1}{2} \right) + \frac{c_1}{2} \boldsymbol{p}_\sigma^{(1)} \left((\boldsymbol{p}_\sigma^{(1)})^T \boldsymbol{z}_i^{(1)} \right)$$

LM MA ES

$\left((\boldsymbol{p}_\sigma^{(1)})^T \boldsymbol{z}^{(1)}\right)$ is a scalar does not require $\mathbf{M}^{(1)}$ to be stored in memory.

$$\boldsymbol{d}_i^{(t)} = \mathbf{M}^{(t)}\boldsymbol{z}_i^{(t)} = \mathbf{M}^{(t-1)}\mathbf{P}^{(t)}\boldsymbol{z}_i^{(t)} = \mathbf{M}^{(t-1)} \underbrace{\left(\left(1 - \frac{c_1}{2}\right)\mathbf{I} + \frac{c_1}{2}\boldsymbol{p}_\sigma^{(t)}(\boldsymbol{p}_\sigma^{(t)})^T\right)}_{:=\mathbf{P}^{(t)}} \boldsymbol{z}_i^{(t)}$$

$$\Downarrow$$

$$\begin{aligned}\boldsymbol{d}_i^{(t)} &= \left(\left(1 - \frac{c_1}{2}\right)\mathbf{I} + \frac{c_1}{2}\boldsymbol{p}_\sigma^{(1)}(\boldsymbol{p}_\sigma^{(1)})^T\right) \cdot \ldots \\ &\quad \ldots \cdot \left(\left(1 - \frac{c_1}{2}\right)\mathbf{I} + \frac{c_1}{2}\boldsymbol{p}_\sigma^{(t-1)}(\boldsymbol{p}_\sigma^{(t-1)})^T\right) \cdot \left(\left(1 - \frac{c_1}{2}\right)\mathbf{I} + \frac{c_1}{2}\boldsymbol{p}_\sigma^{(t)}(\boldsymbol{p}_\sigma^{(t)})^T\right) \boldsymbol{z}_i^{(t)}\end{aligned}$$

Using the last m vector (direction vectors) to update matrix **M**

LM MA ES

Algorithm 1 CMA-ES , MA-ES and LM-MA-ES

given $n \in \mathbb{N}_+$, $\lambda = 4 + \lfloor 3 \ln n \rfloor$, $\mu = \lfloor \lambda/2 \rfloor$, $w_i = \frac{\ln(\mu+\frac{1}{2}) - \ln i}{\sum_{j=1}^{\mu}(\ln(\mu+\frac{1}{2}) - \ln j)}$ for $i = 1, \ldots, \mu$, $\mu_w = \frac{1}{\sum_{i=1}^{\mu} w_i^2}$, $c_\sigma = \frac{\mu_w + 2}{n + \mu_w + 5}$, $c_c = \frac{4}{n+4}$, $c_1 = \frac{2}{(n+1.3)^2 + \mu_w}$, $c_\mu = \min\left(1 - c_1, \frac{2(\mu_w - 2 + 1/\mu_w)}{(n+2)^2 + \mu_w}\right)$,
$m = 4 + \lfloor 3 \ln n \rfloor$, $c_\sigma = \frac{2\lambda}{n}$, $c_{d,i} = \frac{1}{1.5^{i-1} n}$, $c_{c,i} = \frac{\lambda}{4^{i-1} n}$ for $i = 1, \ldots, m$

initialize $t \leftarrow 0, \boldsymbol{y}^{(t=0)} \in \mathbb{R}^n, \sigma^{(t=0)} > 0, \boldsymbol{p}_\sigma^{(t=0)} = \mathbf{0}$, $\boldsymbol{p}_c^{(t=0)} = \mathbf{0}, \mathbf{C}^{(t=0)} = \mathbf{I}$, $\mathbf{M}^{(t=0)} = \mathbf{I}$,
$\boldsymbol{m}_i^{(t=0)} \in \mathbb{R}^n, \boldsymbol{m}_i^{(t=0)} = \mathbf{0}$ for $i = 1, \ldots, m$

repeat

  for $i \leftarrow 1, \ldots, \lambda$ **do**

    $\boldsymbol{z}_i^{(t)} \leftarrow \mathcal{N}(\mathbf{0}, \mathbf{I})$

    $\boldsymbol{d}_i^{(t)} \leftarrow \boldsymbol{z}_i^{(t)}$

    if $t \mod \frac{n}{\lambda} = 0$ **then** $\mathbf{M}^{(t)} \leftarrow \sqrt{\mathbf{C}^{(t)}}$ **else** $\mathbf{M}^{(t)} \leftarrow \mathbf{M}^{(t-1)}$ ▷ CMA-ES

    $\boldsymbol{d}_i^{(t)} \leftarrow \mathbf{M}^{(t)} \boldsymbol{d}_i^{(t)}$ ▷ CMA-ES and MA-ES

    for $j \leftarrow 1, \ldots, \min(t, m)$ **do** ▷ LM-MA-ES

      $\boldsymbol{d}_i^{(t)} \leftarrow (1 - c_{d,j})\boldsymbol{d}_i^{(t)} + c_{d,j}\boldsymbol{m}_j^{(t)} \left((\boldsymbol{m}_j^{(t)})^T \boldsymbol{d}_i^{(t)}\right)$ ▷ LM-MA-ES

    $f_i^{(t)} \leftarrow f(\boldsymbol{y}^{(t)} + \sigma^{(t)} \boldsymbol{d}_i^{(t)})$

  $\boldsymbol{y}^{(t+1)} \leftarrow \boldsymbol{y}^{(t)} + \sigma^{(t)} \sum_{i=1}^{\mu} w_i \boldsymbol{d}_{i:\lambda}^{(t)}$ ▷ the symbol $i:\lambda$ denotes i-th best sample on f

  $\boldsymbol{p}_\sigma^{(t+1)} \leftarrow (1 - c_\sigma)\boldsymbol{p}_\sigma^{(t)} + \sqrt{\mu_w c_\sigma (2 - c_\sigma)} \sum_{i=1}^{\mu} w_i \boldsymbol{z}_{i:\lambda}^{(t)}$

  $\boldsymbol{p}_c^{(t+1)} \leftarrow (1 - c_c)\boldsymbol{p}_c^{(t)} + \sqrt{\mu_w c_c (2 - c_c)} \sum_{i=1}^{\mu} w_i \boldsymbol{d}_{i:\lambda}^{(t)}$ ▷ CMA-ES

  $\mathbf{C}^{(t+1)} \leftarrow (1 - c_1 - c_\mu)\mathbf{C}^{(t)} + c_1 \boldsymbol{p}_c (\boldsymbol{p}_c^{(t)})^T + c_\mu \sum_{i=1}^{\mu} w_i \boldsymbol{d}_{i:\lambda}^{(t)} (\boldsymbol{d}_{i:\lambda}^{(t)})^T$ ▷ CMA-ES

  $\mathbf{M}^{(t+1)} \leftarrow \mathbf{M}^{(t)} \left[\mathbf{I} + \frac{c_1}{2}\left(\boldsymbol{p}_\sigma^{(t)} (\boldsymbol{p}_\sigma^{(t)})^T - \mathbf{I}\right) + \frac{c_\mu}{2}\left(\sum_{i=1}^{\mu} w_i \boldsymbol{z}_{i:\lambda}^{(t)} (\boldsymbol{z}_{i:\lambda}^{(t)})^T - \mathbf{I}\right)\right]$ ▷ MA-ES

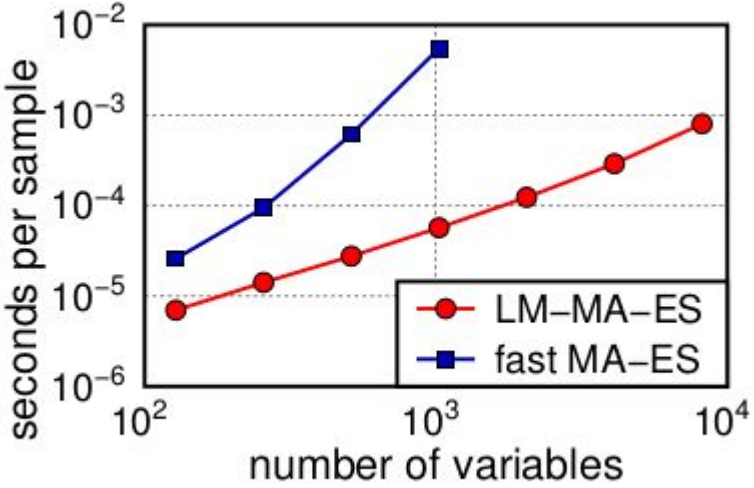
  for $i \leftarrow 1, \ldots, m$ **do** ▷ LM-MA-ES

    $\boldsymbol{m}_i^{(t+1)} \leftarrow (1 - c_{c,i})\boldsymbol{m}_i^{(t)} + \sqrt{\mu_w c_{c,i} (2 - c_{c,i})} \sum_{j=1}^{\mu} w_j \boldsymbol{z}_{j:\lambda}^{(t)}$ ▷ LM-MA-ES

  $\sigma^{(t+1)} \leftarrow \sigma^{(t)} \cdot \exp\left[\frac{c_\sigma}{2}\left(\frac{\|\boldsymbol{p}_\sigma^{(t+1)}\|^2}{n} - 1\right)\right]$

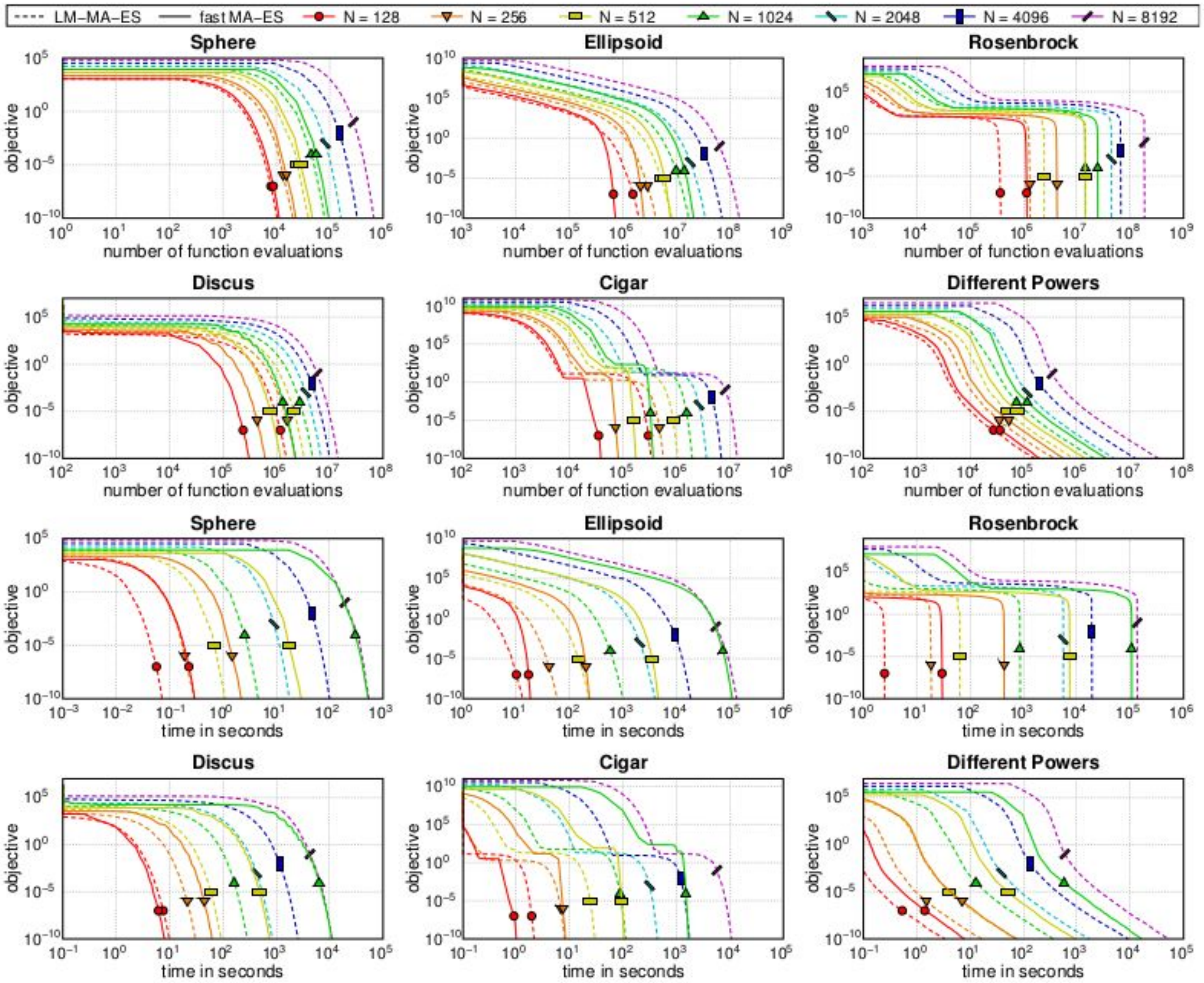
  $t \leftarrow t + 1$

LM MA ES



LM–MA–ES
fast MA–ES
seconds per sample
number of variables
10^−2
10^−3
10^−4
10^−5
10^−6
10^2
10^3
10^4

LM MA ES



LM–MA–ES
fast MA–ES
N = 128
N = 256
N = 512
N = 1024
N = 2048
N = 4096
N = 8192
Sphere
Ellipsoid
Rosenbrock
Discus
Cigar
Different Powers
objective
number of function evaluations
time in seconds


THANKS
for
WATCHING
Any Questions