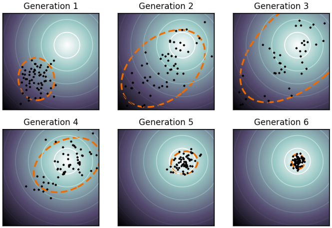
Optimization in Machine Learning

CMA-ES Algorithm



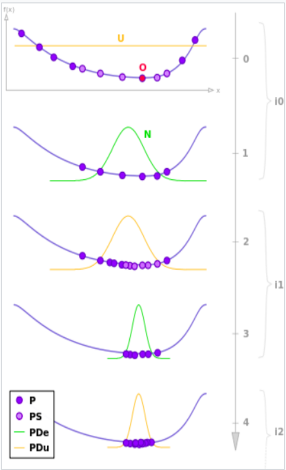
Learning goals

- CMA-ES strategy
- Estimation of distribution
- Step size control

ESTIMATION OF DISTRIBUTION ALGORITHM

- General algorithmic template
- Instead of population we maintain parameterized distribution to sample offspring from

1. Draw λ offsping $\mathbf{x}^{(i)}$ from $p(\boldsymbol{x}|\boldsymbol{\theta}^{[t]})$
2. Evaluate fitness $f(\mathbf{x}^{(i)})$
3. Update $\boldsymbol{\theta}^{[t+1]}$ with μ best offspring

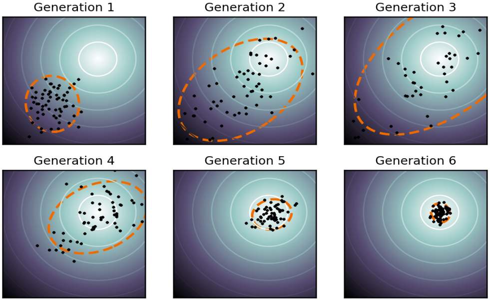


Estimation of distribution algorithm. For each iteration *i*, a random draw is performed for a population *P* in a distribution *PDu*. The distribution parameters *PDe* are then estimated using the selected points *PS*. The illustrated example optimizes a continuous objective function *f(X)* with a unique optimum *O*. The sampling (following a normal distribution *N*) concentrates around the optimum as one goes along unwinding algorithm.

COVARIANCE MATRIX ADAPTATION ES

- Sample distribution is multivariate Gaussian

$$\mathbf{x}^{[t+1](i)} \sim \boldsymbol{m}^{[t]} + \sigma^{[t]}\mathcal{N}(\mathbf{0}, \boldsymbol{C}^{[t]}) \quad \text{for } i = 1, \dots, \lambda.$$

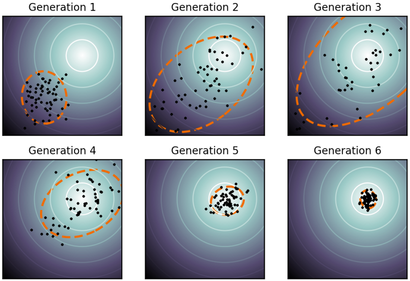


COVARIANCE MATRIX ADAPTATION ES

Sample distribution is multivariate Gaussian

$$\mathbf{x}^{[t+1](i)} \sim \boldsymbol{m}^{[t]} + \sigma^{[t]} \mathcal{N}(\mathbf{0}, \boldsymbol{C}^{[t]}) \quad \text{for } i = 1, \ldots, \lambda$$

- $\mathbf{x}^{[t+1](i)} \in \mathbb{R}^d$ i-th offspring; $\lambda \geq 2$ number of offspring
- $\boldsymbol{m}^{[t]} \in \mathbb{R}^d$ mean value and $\boldsymbol{C}^{[t]} \in \mathbb{R}^{d \times d}$ covar matrix
- $\sigma^{[t]} \in \mathbb{R}_+$ "overall" standard deviation/step size



$\rightarrow$ *How to calculate* $\mathbf{m}^{[t+1]}$, $\boldsymbol{C}^{[t+1]}$, $\sigma^{[t+1]}$ *for next generation* $t+1$ *?*

CMA-ES: BASIC METHOD - ITERATION 1

0. Initialize $\boldsymbol{m}^{[0]}, \sigma^{[0]}$ problem-dependent and $\boldsymbol{C}^{[0]} = \mathbb{I}_d$
1. **Sample** from distribution
   $\mathbf{x}^{[1](i)} = \boldsymbol{m}^{[0]} + \sigma^{[0]}\mathcal{N}(\mathbf{0}, \boldsymbol{C}^{[0]})$ multivariate Gaussian

Initial distribution $\mathcal{N}(\boldsymbol{m}^{[0]}, (\sigma^{[0]})^2\mathbb{I}_2)$ of generation $t = 0$.

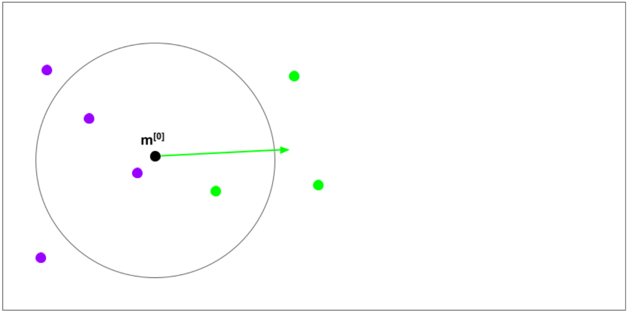
CMA-ES: BASIC METHOD - ITERATION 1

1. **Sample** from distribution
$\mathbf{x}^{[1](i)} = \boldsymbol{m}^{[0]} + \sigma^{[0]}\mathcal{N}(\mathbf{0}, \boldsymbol{C}^{[0]})$ multivariate normal distribution.

Initial distribution $\mathcal{N}(\boldsymbol{m}^{[0]}, (\sigma^{[0]})^2\mathbb{I}_2)$ of generation $t = 0$, $\lambda = 7$.

CMA-ES: BASIC METHOD - ITERATION 1

❷ **Selection and recombination** of $\mu < \lambda$ best-performing offspring using fixed weights $w_1 \geq \ldots \geq w_\mu > 0, \sum_{i=1}^{\mu} w_i = 1$.
$\mathbf{x}_{i:\lambda}$ is i-th ranked solution, ranked by $f(\mathbf{x}_{i:\lambda})$.

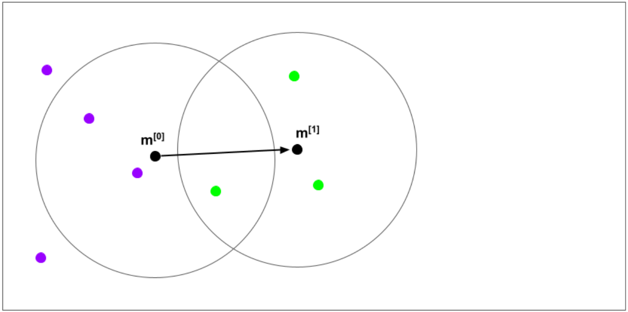


Calculation of auxiliary variables ($\mu = 3$ points)

$$\boldsymbol{y}_w^{[1]} := \sum_{i=1}^{\mu} w_i(\mathbf{x}_{i:\lambda}^{[1]} - \boldsymbol{m}^{[0]})/\sigma^{[0]} := \sum_{i=1}^{\mu} w_i \boldsymbol{y}_{i:\lambda}^{[1]}$$

CMA-ES: BASIC METHOD - ITERATION 1

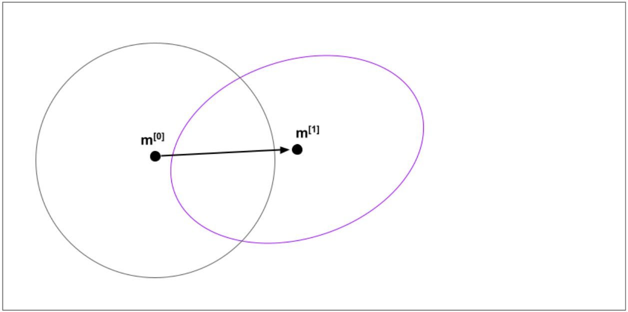
❸ **Update mean**



Movement towards the new distribution with mean
$$\boldsymbol{m}^{[1]} = \boldsymbol{m}^{[0]} + \sigma^{[0]} \boldsymbol{y}_w^{[1]}.$$

CMA-ES: BASIC METHOD - ITERATION 1

4. **Update covariance matrix**
   Roughly: elongate density ellipsoid in direction of successful steps.
   $\boldsymbol{C}^{[1]}$ reproduces successful points with higher probability than $\boldsymbol{C}^{[0]}$.

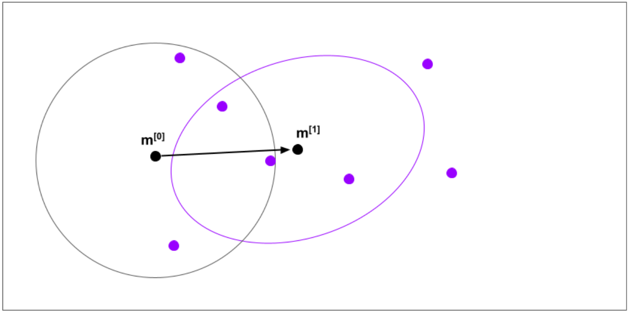


Update $\boldsymbol{C}$ using sum of outer products and learning rate c_μ (simplified):

$$\boldsymbol{C}^{[1]} = (1 - c_\mu)\boldsymbol{C}^{[0]} + c_\mu \sum_{i=1}^{\mu} w_i \boldsymbol{y}_{i:\lambda}^{[1]} (\boldsymbol{y}_{i:\lambda}^{[1]})^\top \text{ (rank-}\mu\text{ update).}$$

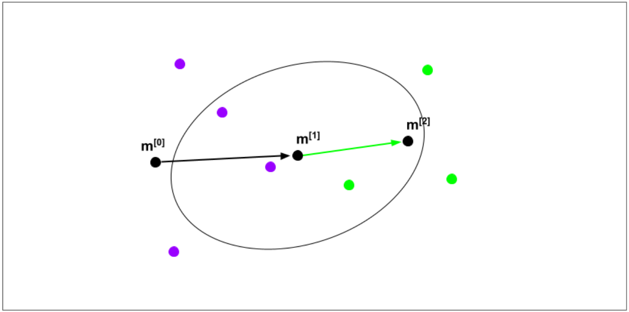
CMA-ES: BASIC METHOD - ITERATION 2

1. **Sample** from distribution for new generation



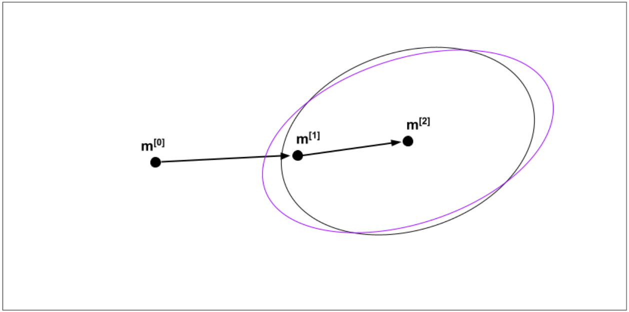
CMA-ES: BASIC METHOD - ITERATION 2

2. **Selection and recombination** of $\mu < \lambda$ best-performing offspring
3. **Update mean**



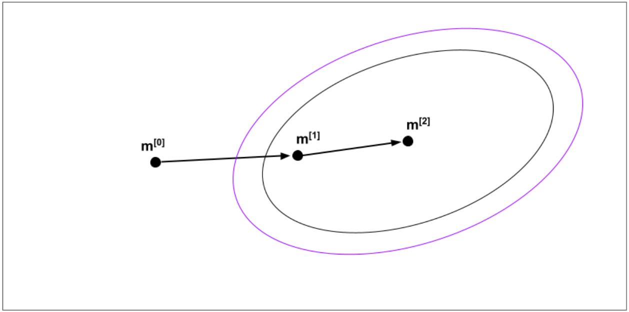
CMA-ES: BASIC METHOD - ITERATION 2

4. **Update covariance matrix**



CMA-ES: BASIC METHOD - ITERATION 2

5. **Update step-size** exploiting correlation in history of steps.
   steps point in similar direction $\Longrightarrow$ increase step-size
   steps cancel out $\Longrightarrow$ decrease step-size



UPDATING $\boldsymbol{C}$: FULL UPDATE

Full CMA update of $\boldsymbol{C}$ combines rank-μ update with a rank-1 update using exponentially smoothed evolution path $\boldsymbol{p}_c \in \mathbb{R}^d$ of successive steps and learning rate c_1:

$$\boldsymbol{p}_c^{[0]} = \mathbf{0}, \quad \boldsymbol{p}_c^{[t+1]} = (1 - c_1)\boldsymbol{p}_c^{[t]} + \sqrt{\frac{c_1(2 - c_1)}{\sum_{i=1}^{\mu} w_i^2}}\boldsymbol{y}_w$$

Final update of $\boldsymbol{C}$ is

$$\boldsymbol{C}^{[t+1]} = (1 - c_1 - c_\mu \sum w_j)\boldsymbol{C}^{[t]} + c_1 \underbrace{\boldsymbol{p}_c^{[t+1]}(\boldsymbol{p}_c^{[t+1]})^\top}_{\text{rank-1}} + c_\mu \underbrace{\sum_{i=1}^{\mu} w_i \boldsymbol{y}_{i:\lambda}^{[t+1]}(\boldsymbol{y}_{i:\lambda}^{[t+1]})^\top}_{\text{rank-}\mu}$$

- Correlation between generations used in rank-1 update
- Information from entire population is used in rank-μ update

UPDATING σ: METHODS STEP-SIZE CONTROL

- $1/5$**-th success rule**: increases the step-size if more than 20 % of the new solutions are successful, decrease otherwise
- σ**-self-adaptation**: mutation is applied to the step-size and the better - according to the objective function value - is selected
- **Path length control via cumulative step-size adaptation (CSA)**
  Intuition:
  - Short cumulative step-size $\triangleq$ steps cancel $\rightarrow$ decrease $\sigma^{[t+1]}$
  - Long cumulative step-size $\triangleq$ corr. steps $\rightarrow$ increase $\sigma^{[t+1]}$