CS4: Optimization using Covariance Matrix Adaptation

Lecturer: Georgios Arampatzis

CSElab,
Chair of Computational Science

2017.03.10

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

based on “**The CMA Evolution Strategy: A Tutorial**”

by Nikolaus Hansen

GOAL

- Given and objective function in continuous domain

$$f : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}$$

- find

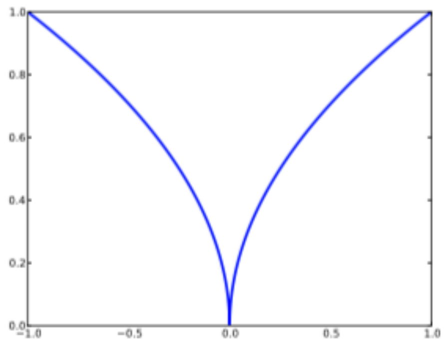
$$x^\star = \underset{x \in \mathcal{X}}{\operatorname{argmax}} \ f(x)$$

Black Box Scenario

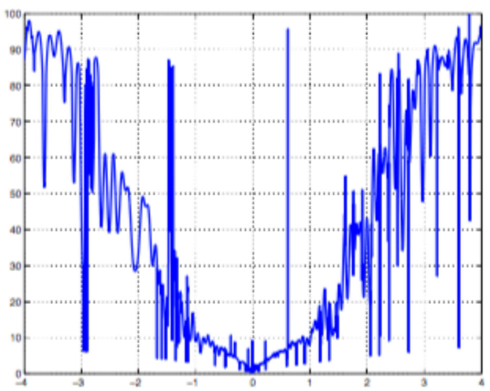
- gradients are not available
- non-convex
- non-smooth
- multimodal
- high dimensional
- noisy
- ...

Difficult Functions

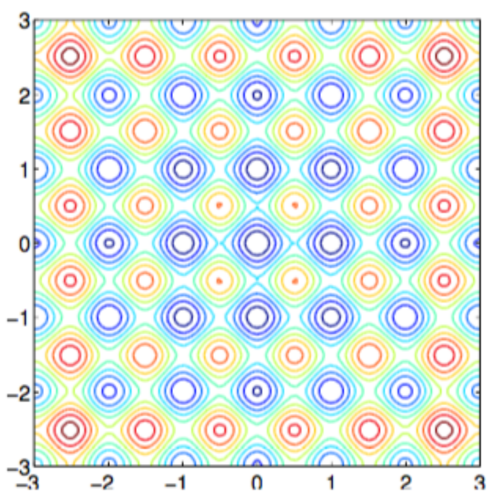
https://www.lri.fr/~hansen/gecco2013-CMA-ES-tutorial.pdf



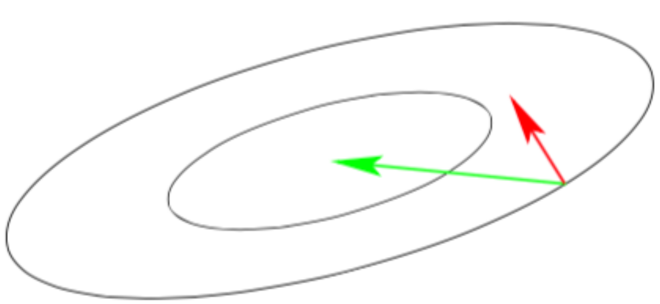
- non-linear, non-quadratic, non-convex



- non-smooth, multimodal, noisy



- non-separability



- ill conditioning

Examples

- shape optimization
    - curve fitting
    - airfoils
- model calibration
    - biological
    - physical
- parameter calibration
    - controller

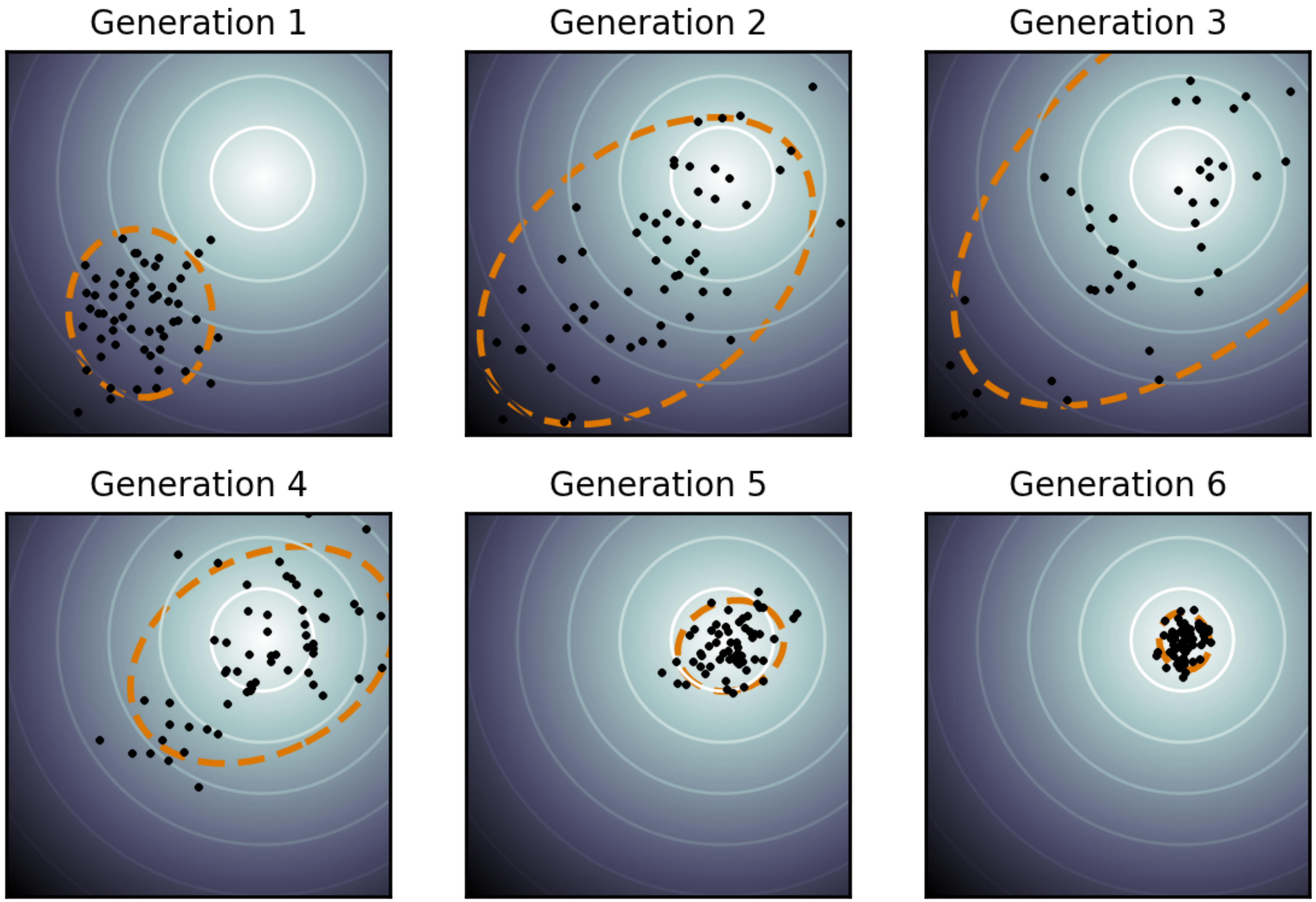
Randomized Black Box Search

- ➡initialize
  - ➡population size $\lambda \in \mathbb{N}$
  - ➡distribution parameters $\vartheta^{(0)}$

- ➡until happy
  - ➡sample $\boldsymbol{x}_i \sim P(\boldsymbol{x}|\vartheta^{(k)}), \quad i = 1, \ldots, \lambda$
  - ➡evaluate $f(\boldsymbol{x}_i), \quad i = 1, \ldots, \lambda$
  - ➡update parameters

$$\vartheta^{(k+1)} = F\big(\vartheta, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_\lambda, f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_\lambda)\big)$$

Randomized Black Box Search



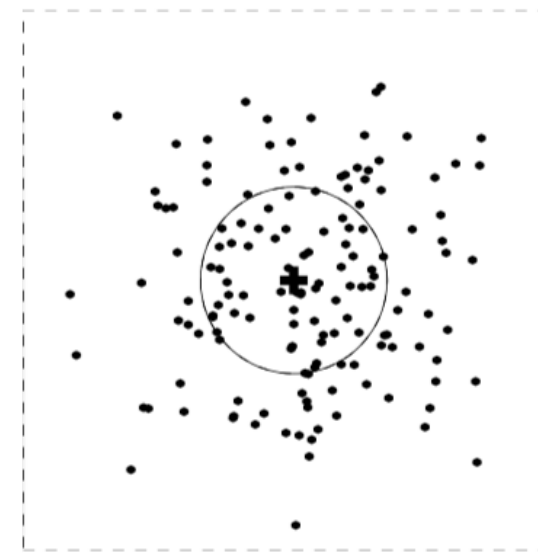
https://upload.wikimedia.org/wikipedia/commons/d/d8/Concept_of_directional_optimization_in_CMA-ES_algorithm.png

THE SAMPLING DISTRIBUTION

- choose sampling distribution

- isotropic
- maximum entropy

$$\boldsymbol{x}_i \sim P(\boldsymbol{x}|\vartheta^{(k)}) = \boldsymbol{m}^{(k)} + \boldsymbol{\sigma}^{(k)}\mathcal{N}(\mathbf{0}, \boldsymbol{C}^{(k)})$$



https://www.lri.fr/~hansen/gecco2013-CMA-ES-tutorial.pdf

- choose how to update the parameters

$$\boldsymbol{m}^{(k)} \quad - \quad \boldsymbol{\sigma}^{(k)} \quad - \quad \boldsymbol{C}^{(k)}$$

represents the favorite solution

controls the step-size

determines the shape of the distribution

THE INGREDIENTS

$\boldsymbol{m}^{(k)}$ Evolution Strategy

$\boldsymbol{C}^{(k)}$ Covariance Matrix Adaptation

$\sigma^{(k)}$ Step Size Control

EVOLUTION STRATEGIES

A. SELECT AND RECOMBINE

B. COMPUTE THE MEAN

EVOLUTION STRATEGIES

- # parents: μ
- # children: λ

- elitist selection: $(\mu + \lambda)$-ES
- non-elitist selection: (μ, λ)-ES

$(1+1)$-ES

- sample one child from parent $\mathbf{m}$

$$x \sim \mathbf{m} + \sigma \mathcal{N}(0, \mathbf{C})$$

- if x is better than $\mathbf{m}$ select

$$\mathbf{m} \leftarrow x$$

Evolution Strategies

Selection and weighted Recombination

$$\boldsymbol{m}^{(k+1)} = \sum_{i=1}^{\mu} w_i \, \boldsymbol{x}_{i:\lambda}^{(k+1)}$$

$$(\mu \backslash \mu_W, \lambda)\text{-ES}$$

parents

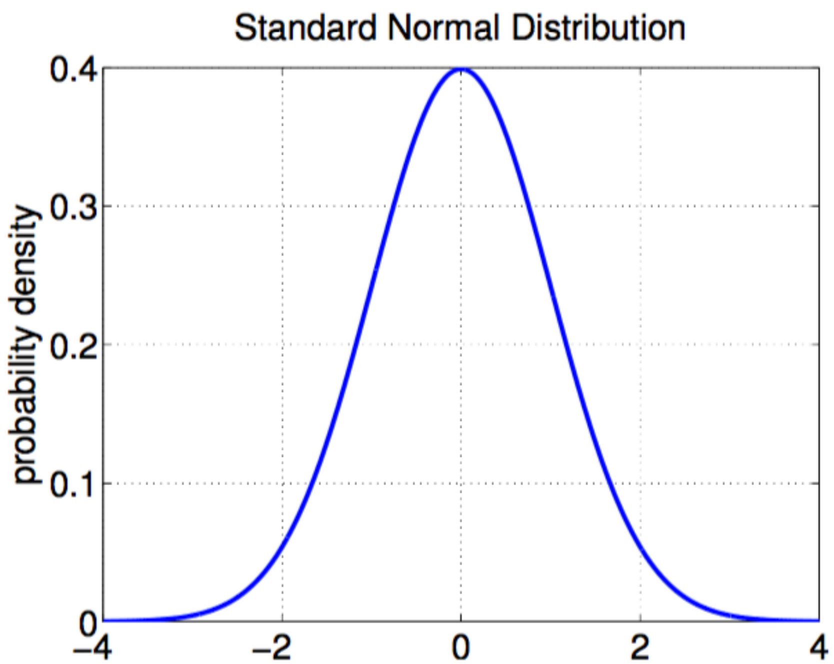
weight
+
recombine

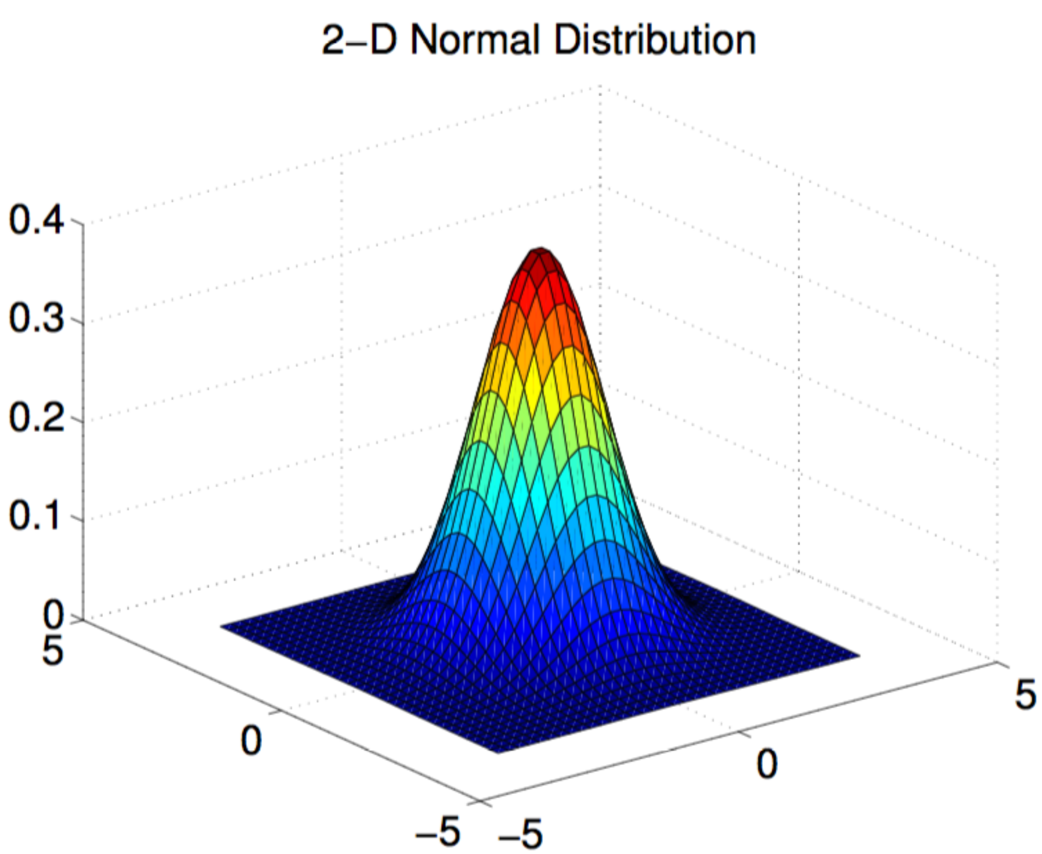
children

$w_i \propto \mu - i + 1$

- $w_1 \geq w_2 \geq \ldots \geq w_\mu > 0$
- $\sum_i w_i = 1$
- $f(\boldsymbol{x}_{1:\lambda}) \leq f(\boldsymbol{x}_{2:\lambda}) \leq \ldots \leq f(\boldsymbol{x}_{\mu:\lambda})$

THE NORMAL DISTRIBUTION

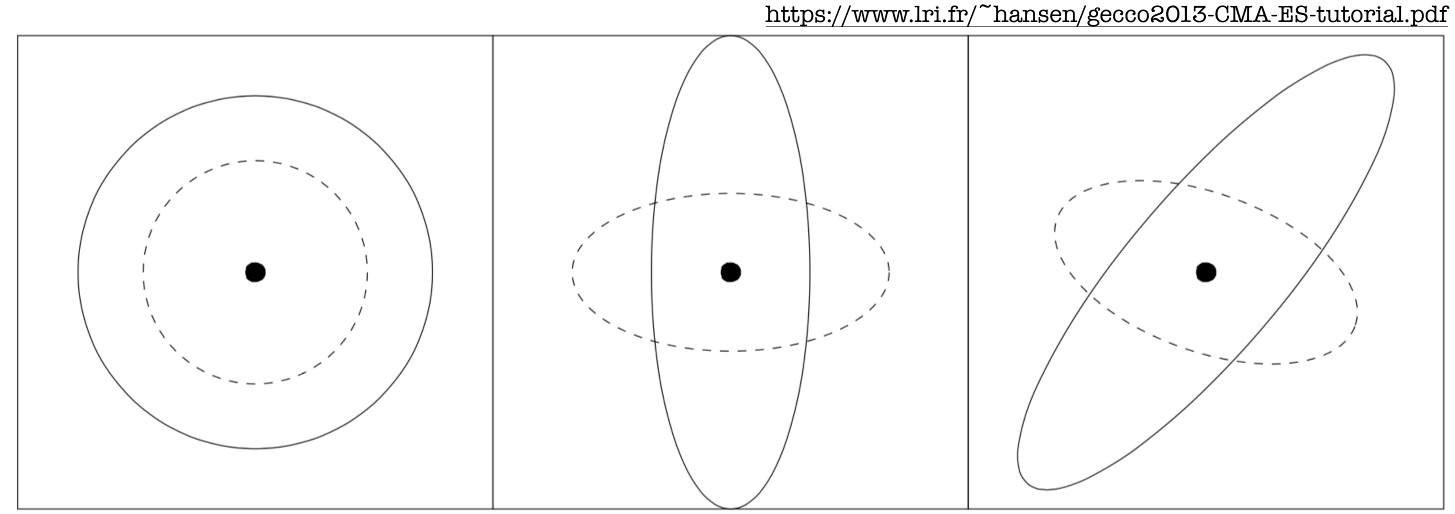




- geometrical interpretation of covariance matrix

$$(x-\mathbf{m})^\top \mathbf{C}^{-1}(x-\mathbf{m}) = \text{const.}$$

https://www.lri.fr/~hansen/gecco2013-CMA-ES-tutorial.pdf



$$\mathcal{N}(m, \sigma^2\mathbf{I}) \sim m + \sigma\mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$\mathcal{N}(m, \mathbf{D}^2) \sim m + \mathbf{D}\mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$\mathcal{N}(m, \mathbf{C}^2) \sim m + \mathbf{C}^{-\frac{1}{2}}\mathcal{N}(\mathbf{0}, \mathbf{I})$$

COVARIANCE MATRIX ADAPTATION

A. ESTIMATION FROM SCRATCH

B. RANK-MU-UPDATE

C. RANK-ONE-UPDATE

Covariance Matrix Adaptation

A. Estimation from scratch

$$\boldsymbol{x}_i \sim \boldsymbol{m}^{(k)} + \sigma^{(k)} \mathcal{N}(\mathbf{0}, \boldsymbol{C}^{(k)})$$

$$\boldsymbol{C}_{\text{emp}}^{(k+1)} = \frac{1}{\lambda - 1} \sum_{i=1}^{\lambda} (\boldsymbol{x}_{\boldsymbol{i}}^{(k+1)} - \boldsymbol{m}^{(k+1)})(\boldsymbol{x}_{\boldsymbol{i}}^{(k+1)} - \boldsymbol{m}^{(k+1)})^\top$$

$$\boldsymbol{C}_{\lambda}^{(k+1)} = \frac{1}{\lambda} \sum_{i=1}^{\lambda} (\boldsymbol{x}_i^{(k+1)} - \boldsymbol{m}^{(k)})(\boldsymbol{x}_i^{(k+1)} - \boldsymbol{m}^{(k)})^\top$$

$$\mathbb{E}\left[\boldsymbol{C}_{\text{emp}}^{(k+1)} | \boldsymbol{C}^{(k)}\right] = \boldsymbol{C}^{(k)}$$

$$\mathbb{E}\left[\boldsymbol{C}_{\lambda}^{(k+1)} | \boldsymbol{C}^{(k)}\right] = \boldsymbol{C}^{(k)}$$

unbiased estimators

Covariance Matrix Adaptation

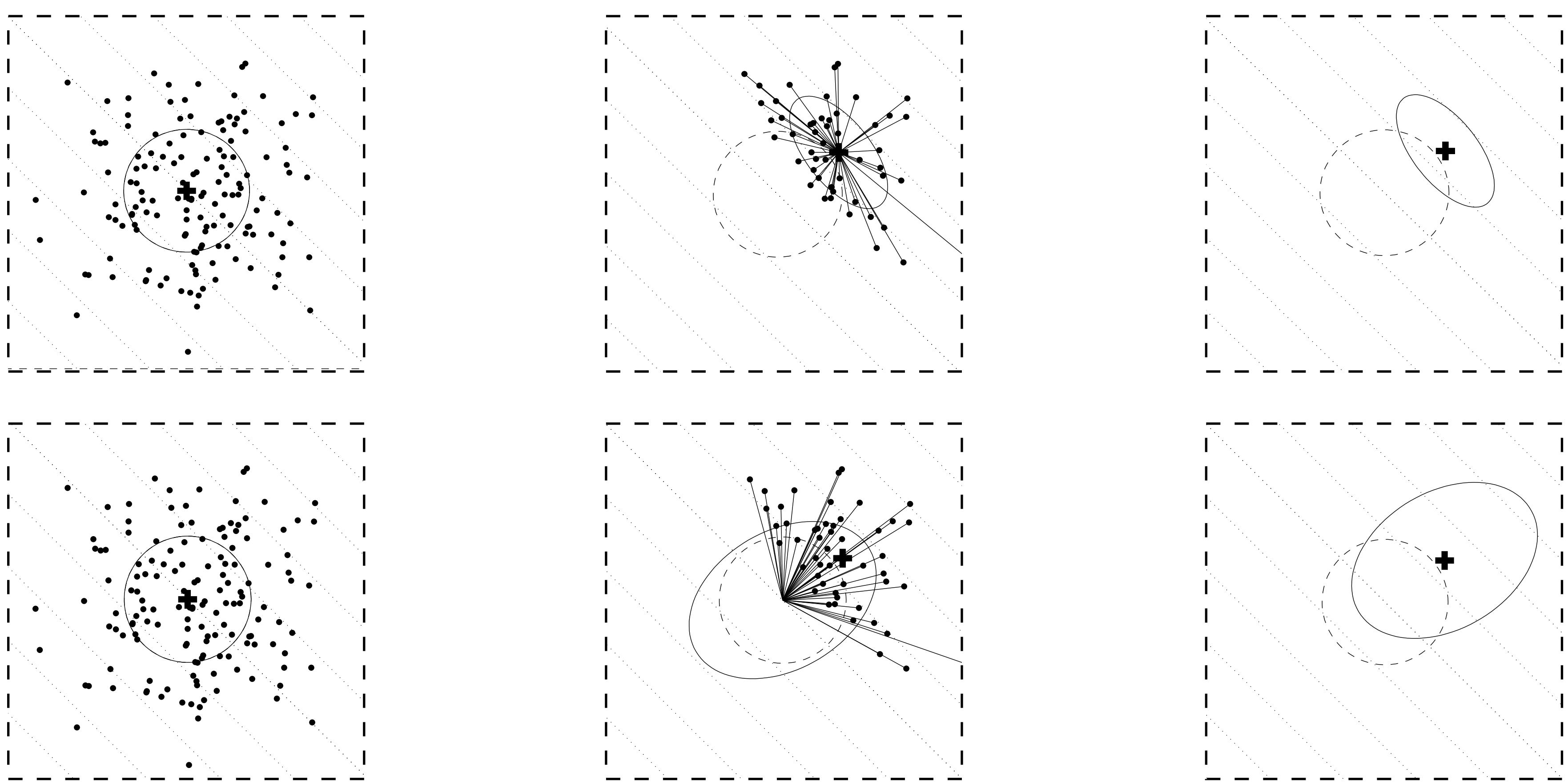
A. Estimation from scratch

w_i

$$\boldsymbol{C}_{\mathrm{EMNA}}^{(k+1)} = \frac{1}{\mu}\sum_{i=1}^{\mu}(\boldsymbol{x}_{\boldsymbol{i}}^{(k+1)} - \boldsymbol{m}^{(k+1)})(\boldsymbol{x}_{\boldsymbol{i}}^{(k+1)} - \boldsymbol{m}^{(k+1)})^\top$$

$$\boldsymbol{C}_{\mu}^{(k+1)} = \frac{1}{\mu}\sum_{i=1}^{\mu}(\boldsymbol{x}_{i}^{(k+1)} - \boldsymbol{m}^{(k)})(\boldsymbol{x}_{i}^{(k+1)} - \boldsymbol{m}^{(k)})^\top$$

https://www.lri.fr/~hansen/gecco2013-CMA-ES-tutorial.pdf



- smaller variance
- increase geometrically fast
- premature convergence

Covariance Matrix Adaptation

b. Rank-mu update

- estimation from scratch works well for large populations
- in order to be fast population must be small
- use information from past

$$C^{k+1} = \frac{1}{k+1} \sum_{i=0}^{k} \frac{1}{{\sigma^{(i)}}^2} C_{\mu}^{i+1}$$

Covariance Matrix Adaptation

B. Rank-mu update

- use information from past
- assign recent generations higher weight
- exponential smoothing

$$\boldsymbol{C}^{(k+1)} = (1 - c_\mu)\boldsymbol{C}^{(k)} + c_\mu \frac{1}{{\sigma^{(k)}}^2}\boldsymbol{C}_\mu^{(k+1)}$$

$$= (1 - c_\mu)\boldsymbol{C}^{(k)} + c_\mu \sum_{i=1}^{\mu} w_i\, \boldsymbol{y}_{i:\lambda}^{(k+1)} {\boldsymbol{y}_{i:\lambda}^{(k+1)}}^\top$$

$$\boldsymbol{y}_{i:\lambda}^{(k+1)} = \frac{\boldsymbol{x}_{i:\lambda}^{(k+1)} - \boldsymbol{m}^{(k)}}{\sigma^{(k)}}$$

$\boxed{\text{rank: } \min\{\mu, n\}}$

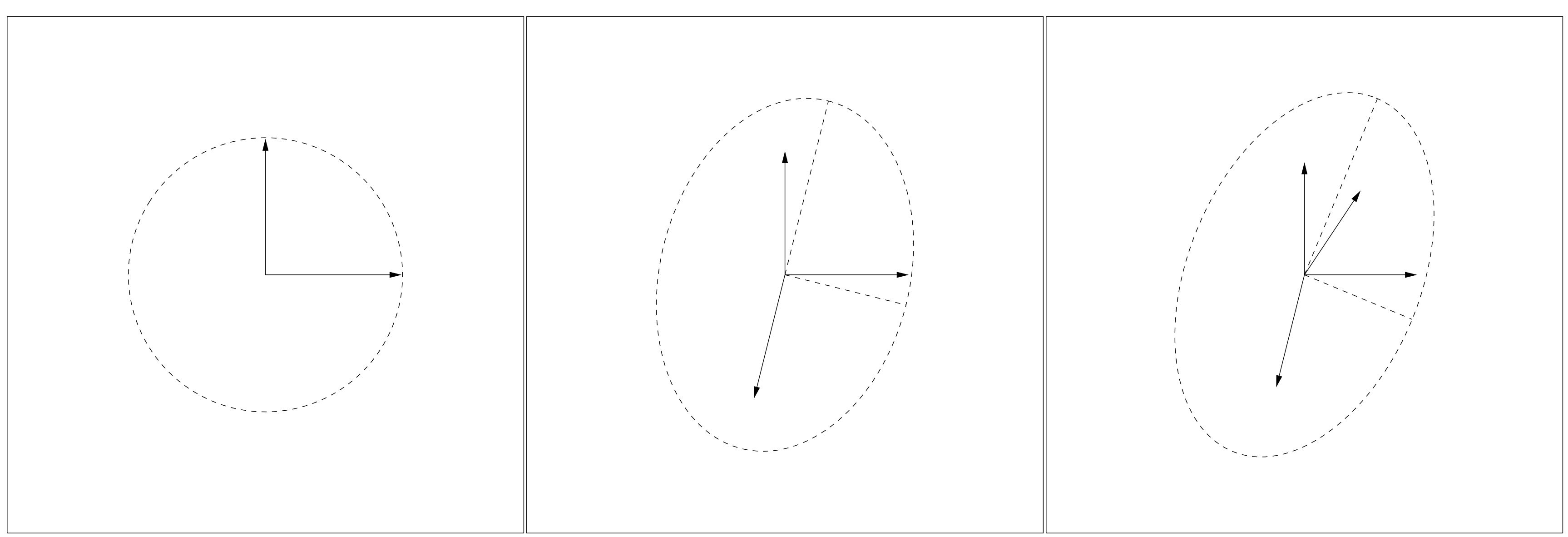
Covariance Matrix Adaptation

C. Rank-One Update

$$\mathcal{N}(0,1)\boldsymbol{y}_1 + \ldots + \mathcal{N}(0,1)\boldsymbol{y}_k \;\sim\; \mathcal{N}(\mathbf{0}, \sum_{i=1}^{k} \boldsymbol{y}_i \boldsymbol{y}_i^\top)$$

- the singular distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{y}_i \boldsymbol{y}_i^\top)$ generates the vector $\boldsymbol{y}_i$ with maximum likelihood

https://www.lri.fr/~hansen/gecco2013-CMA-ES-tutorial.pdf



$\mathcal{N}(\mathbf{0}, \boldsymbol{C}^{(0)})$ $\mathcal{N}(\mathbf{0}, \boldsymbol{C}^{(1)})$ $\mathcal{N}(\mathbf{0}, \boldsymbol{C}^{(2)})$

Covariance Matrix Adaptation

c. Rank-one update

- assume

$$\boldsymbol{y}^{(k+1)} = \frac{\boldsymbol{x}_{1:\lambda}^{(k+1)} - \boldsymbol{m}^{(k)}}{\sigma^{(k)}}$$

- then the covariance matrix

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

- increases the probability of generating $\boldsymbol{y}^{(k+1)}$ in the next generation

Covariance Matrix Adaptation

C. Rank-One Update + Cumulation

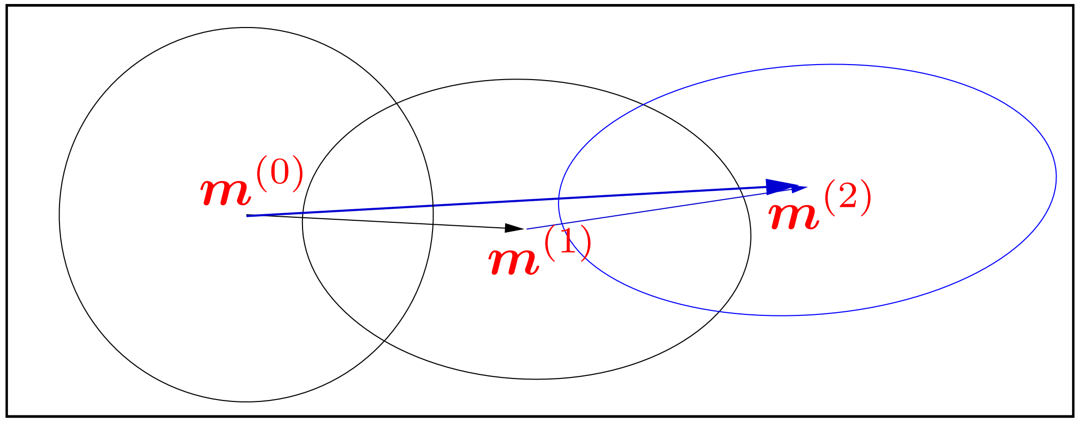
- loss of sign information $\quad \boldsymbol{y}\boldsymbol{y}^\top = -\boldsymbol{y}(-\boldsymbol{y})^\top$

- introduce the evolution path

$$\boldsymbol{p}_c^{(k+1)} = \sum_{i=1}^{k} \frac{\boldsymbol{m}^{(i+1)} - \boldsymbol{m}^{(i)}}{\sigma^{(i)}}$$

- or with exponential smoothing

$$\boldsymbol{p}_c^{(k+1)} = (1 - c_c)\boldsymbol{p}_c^{(k)} + c_c \frac{\boldsymbol{m}^{(k+1)} - \boldsymbol{m}^{(k)}}{\sigma^{(i)}}$$



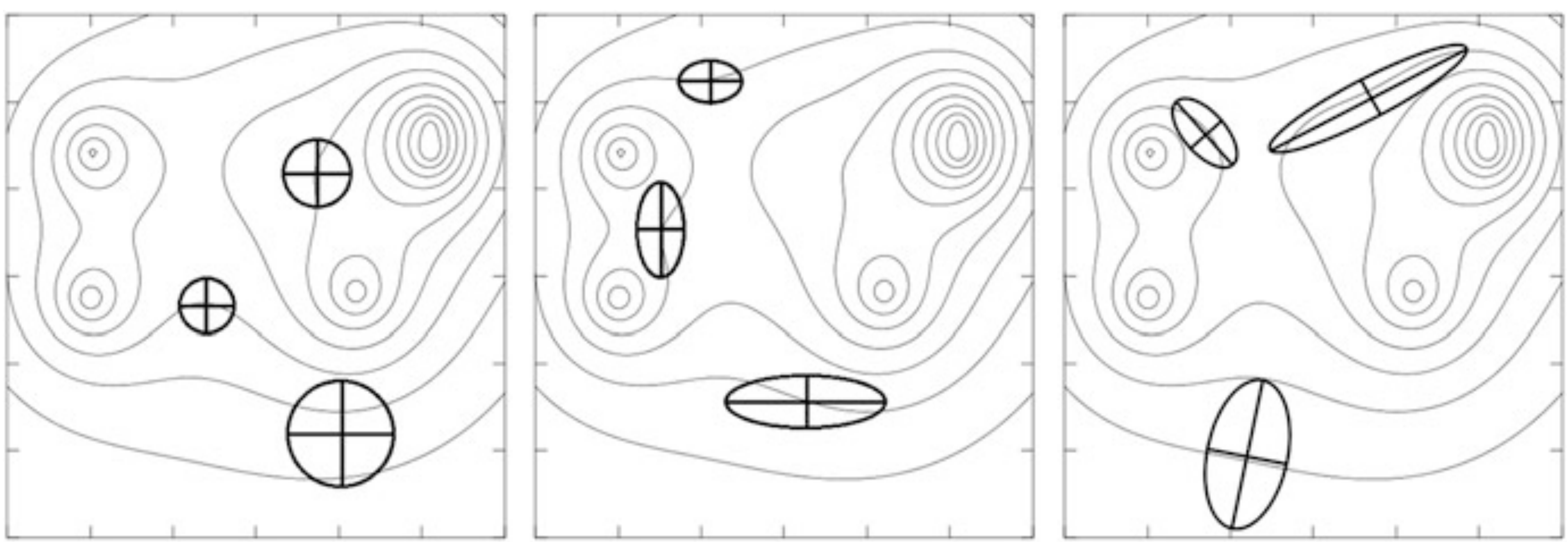
Covariance Matrix Adaptation

$$
\begin{aligned}
\boldsymbol{C}^{(k+1)} \;=\; & (1 - c_\mu - c_1)\, \boldsymbol{C}^{(k)} \\
& + c_\mu \sum_{i=1}^{\mu} w_i\, \boldsymbol{y}_{i:\lambda}^{(k+1)} \boldsymbol{y}_{i:\lambda}^{(k+1)^\top} \\
& + c_1\, \boldsymbol{p}_c^{(k+1)} \boldsymbol{p}_c^{(k+1)}
\end{aligned}
$$

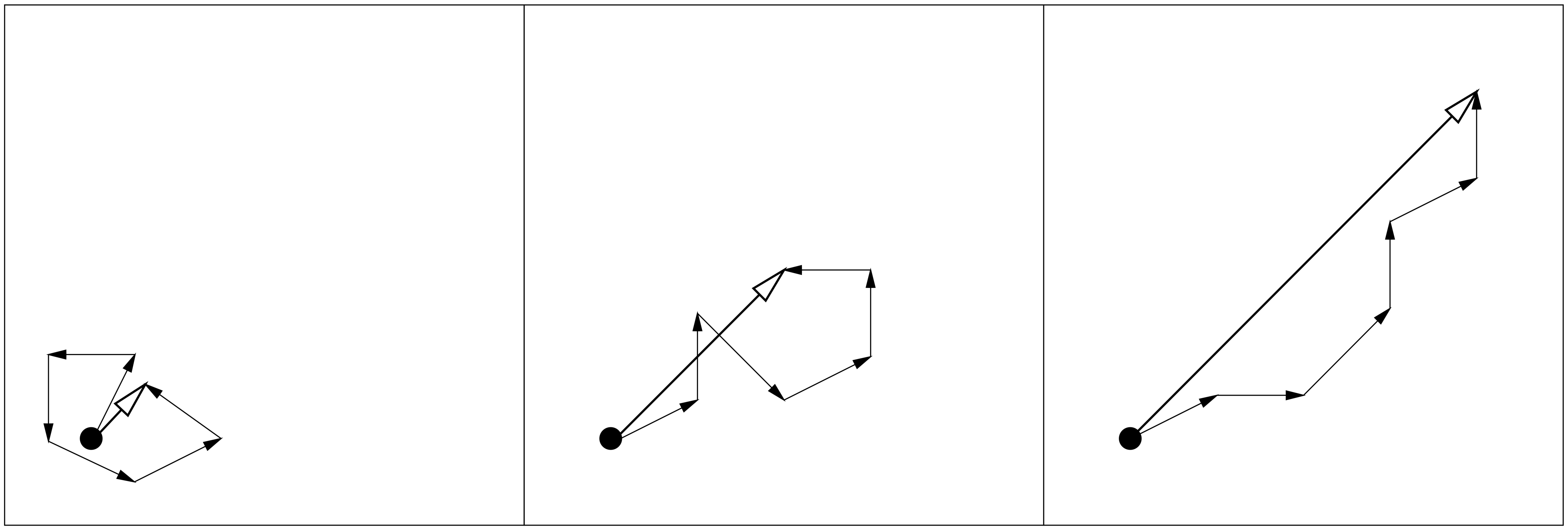
Covariance Matrix Adaptation

- ✤ learns all pairwise dependencies between variables
- ✤ learns a pre rotated problem representation
- ✤ learns a new Mahalanobis metric
- ✤ approximates the inverse Hessian on quadratic functions

Contemporary Evolution Strategies, Bäck, T., Foussette, C., Krause, P., Springer



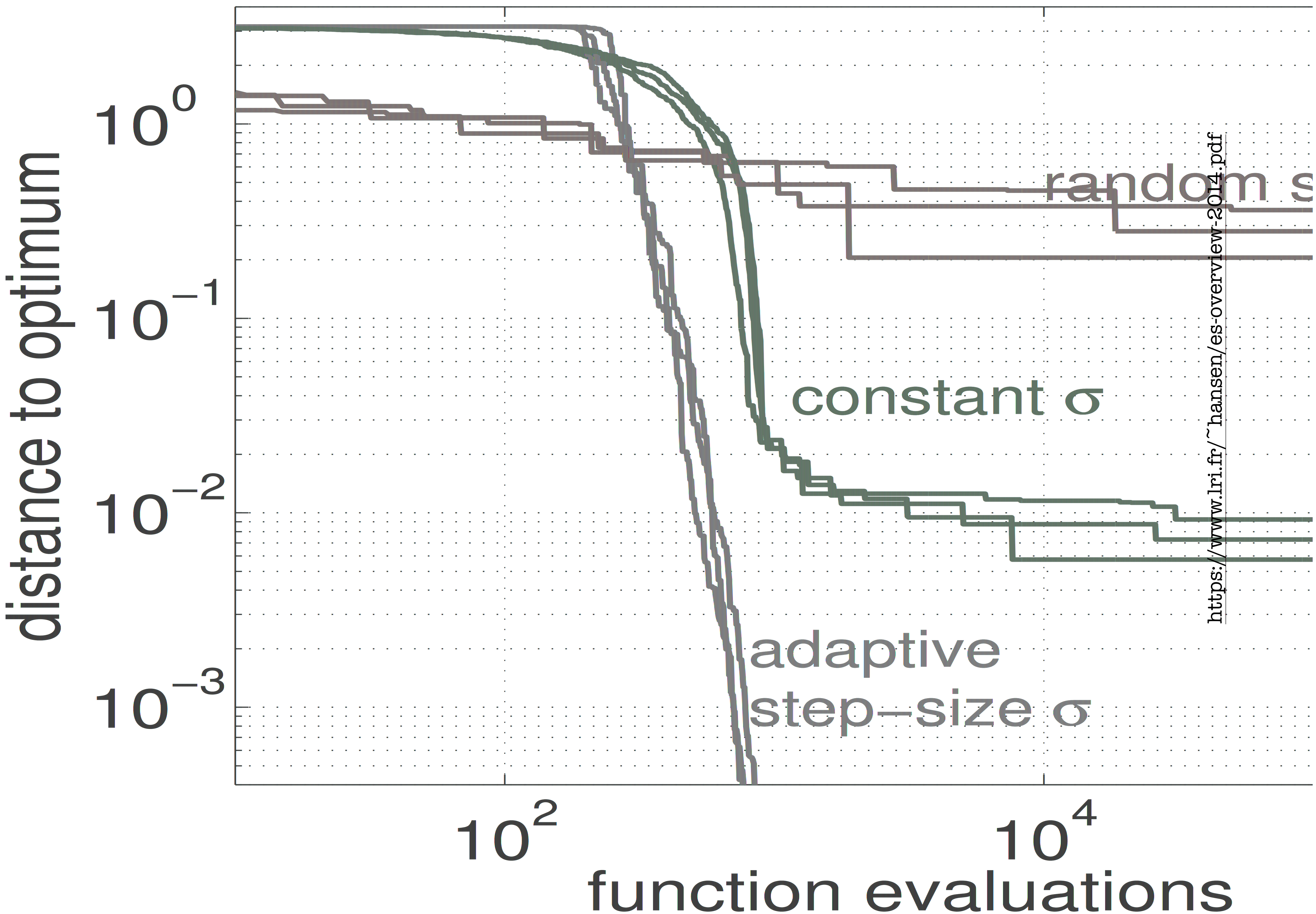
STEP-SIZE CONTROL



https://www.lri.fr/~hansen/gecco2013-CMA-ES-tutorial.pdf

- evolution path is short: single steps cancel out (anti-correlated)
- evolution path is long: steps point to the same direction (correlated)
- evolution path is OK: **uncorrelated** steps

STEP-SIZE CONTROL



THE ALGORITHM

Set parameters

Set parameters λ, μ, $w_{i=1\ldots\mu}$, c_σ, d_σ, c_c, c_1, and c_μ to their default values according to Table 1.

Initialization

Set evolution paths $\boldsymbol{p}_\sigma = \mathbf{0}$, $\boldsymbol{p}_\mathrm{c} = \mathbf{0}$, covariance matrix $\boldsymbol{C} = \mathbf{I}$, and $g = 0$.

Choose distribution mean $\boldsymbol{m} \in \mathbb{R}^n$ and step-size $\sigma \in \mathbb{R}_+$ problem dependent.[1]

Until termination criterion met, $g \leftarrow g + 1$

Sample new population of search points, for $k = 1, \ldots, \lambda$

$$\boldsymbol{z}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{35}$$

$$\boldsymbol{y}_k = \boldsymbol{B}\boldsymbol{D}\boldsymbol{z}_k \quad \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C}) \tag{36}$$

$$\boldsymbol{x}_k = \boldsymbol{m} + \sigma \boldsymbol{y}_k \quad \sim \mathcal{N}\left(\boldsymbol{m}, \sigma^2 \boldsymbol{C}\right) \tag{37}$$

Selection and recombination

$$\langle \boldsymbol{y} \rangle_\mathrm{w} = \sum_{i=1}^{\mu} w_i \, \boldsymbol{y}_{i:\lambda} \quad \text{where} \sum_{i=1}^{\mu} w_i = 1, \; w_i > 0 \tag{38}$$

$$\boldsymbol{m} \leftarrow \boldsymbol{m} + \sigma \langle \boldsymbol{y} \rangle_\mathrm{w} = \sum_{i=1}^{\mu} w_i \, \boldsymbol{x}_{i:\lambda} \tag{39}$$

Step-size control

$$\boldsymbol{p}_\sigma \leftarrow (1 - c_\sigma)\boldsymbol{p}_\sigma + \sqrt{c_\sigma(2 - c_\sigma)\mu_\mathrm{eff}} \, \boldsymbol{C}^{-\frac{1}{2}} \langle \boldsymbol{y} \rangle_\mathrm{w} \tag{40}$$

$$\sigma \leftarrow \sigma \times \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|\boldsymbol{p}_\sigma\|}{\mathsf{E}\|\mathcal{N}(\mathbf{0}, \mathbf{I})\|} - 1\right)\right) \tag{41}$$

Covariance matrix adaptation

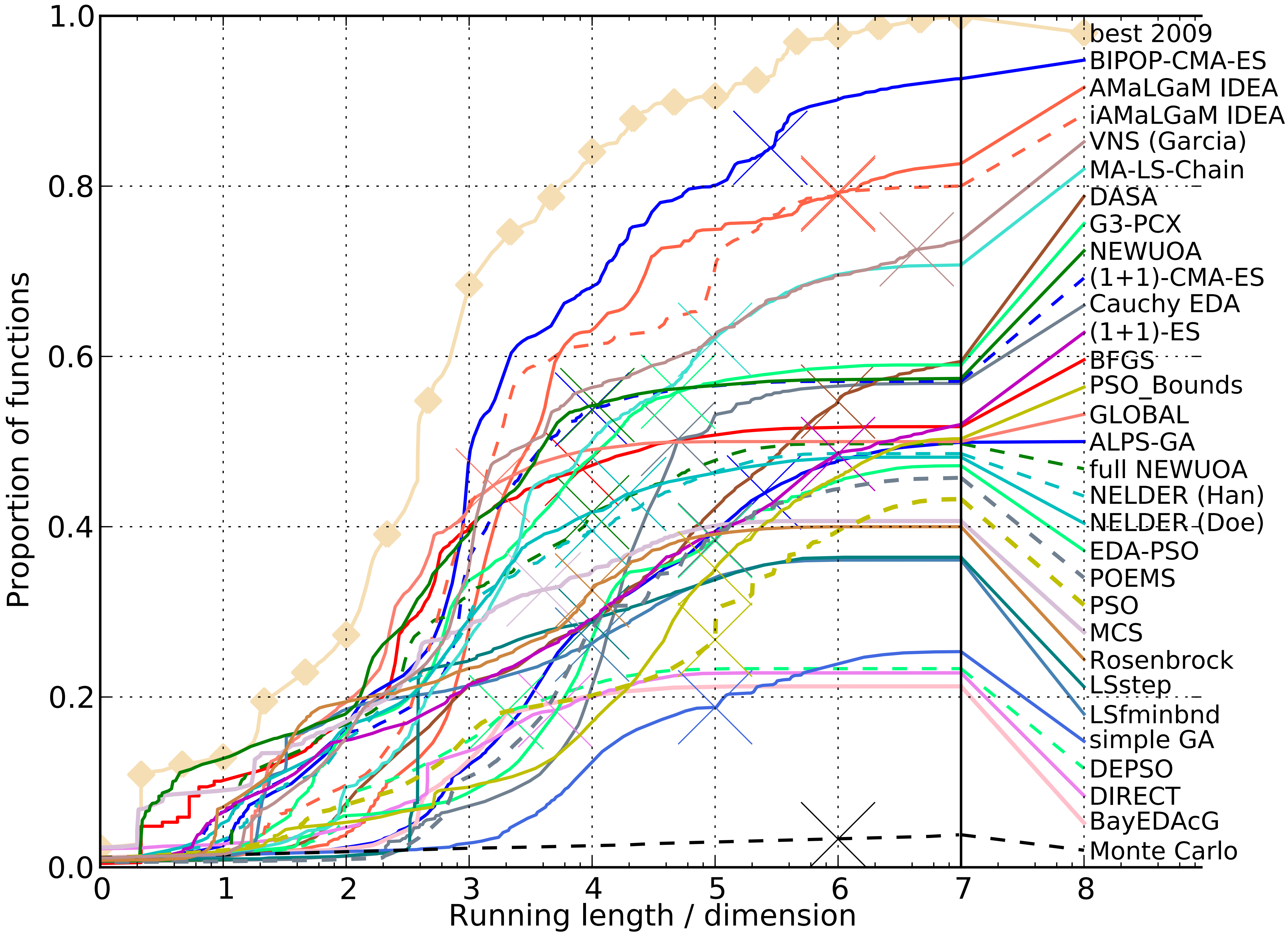
$$\boldsymbol{p}_\mathrm{c} \leftarrow (1 - c_\mathrm{c})\boldsymbol{p}_\mathrm{c} + h_\sigma \sqrt{c_\mathrm{c}(2 - c_\mathrm{c})\mu_\mathrm{eff}} \, \langle \boldsymbol{y} \rangle_\mathrm{w} \tag{42}$$

$$\boldsymbol{C} \leftarrow (1 - c_1 - c_\mu)\, \boldsymbol{C} + c_1 \left(\boldsymbol{p}_\mathrm{c}\boldsymbol{p}_\mathrm{c}^\mathrm{T} + \delta(h_\sigma)\boldsymbol{C}\right) + \; c_\mu \sum_{i=1}^{\mu} w_i \, \boldsymbol{y}_{i:\lambda}\boldsymbol{y}_{i:\lambda}^\mathrm{T} \tag{43}$$

---

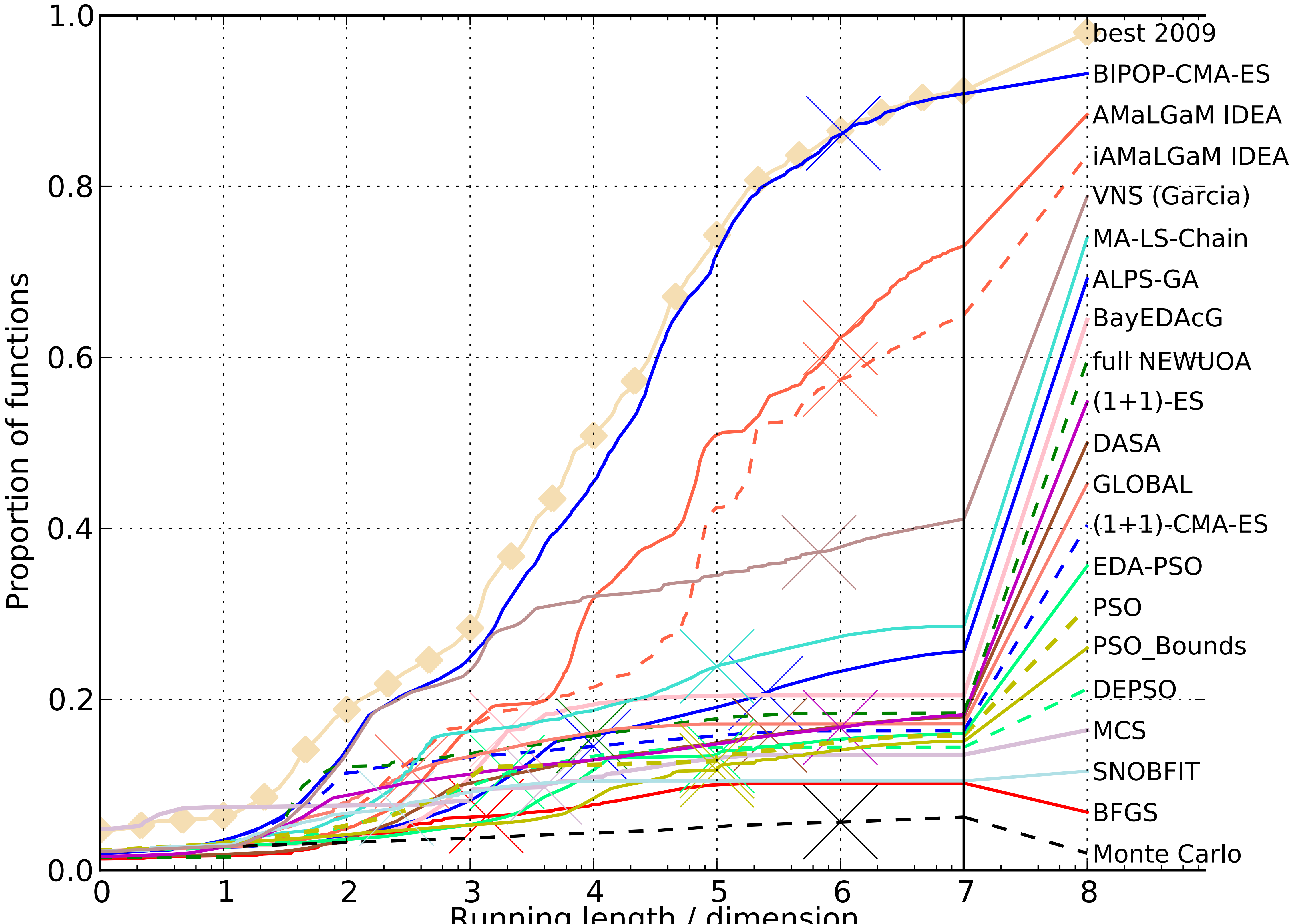
[1] The optimum should presumably be within the initial cube $\boldsymbol{m} \pm 3\sigma(1, \ldots, 1)^\mathrm{T}$. If the optimum is expected to be in the initial search interval $[a, b]^n$ we may choose the initial search point, $\boldsymbol{m}$, uniformly randomly in $[a, b]^n$, and $\sigma = 0.3(b - a)$. Different search intervals Δs_i for different variables can be reflected by a different initialization of $\boldsymbol{C}$, in that the diagonal elements of $\boldsymbol{C}$ obey $c_{ii} = (\Delta s_i)^2$. Remark that the Δs_i should not disagree by several orders of magnitude. Otherwise a scaling of the variables should be applied.

BENCHMARKS deterministic functions



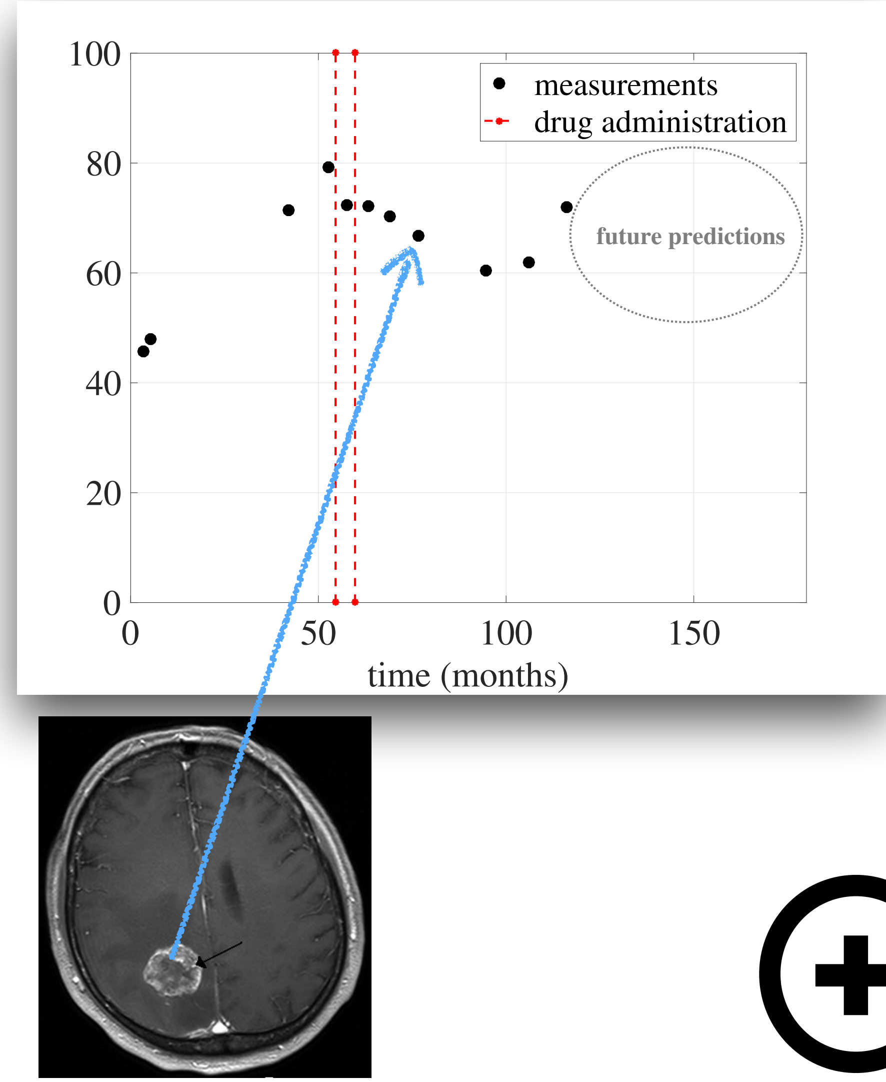
https://www.lri.fr/~hansen/gecco2013-CMA-ES-tutorial.pdf

BENCHMARKS **noisy** functions



https://www.lri.fr/~hansen/gecco2013-CMA-ES-tutorial.pdf

APPLICATION



https://en.wikipedia.org/wiki/Brain_tumor#/media/File:Hirnmetastase_MRT-T1_KM.jpg

$$\frac{dC}{dt} = -\vartheta_1 C$$

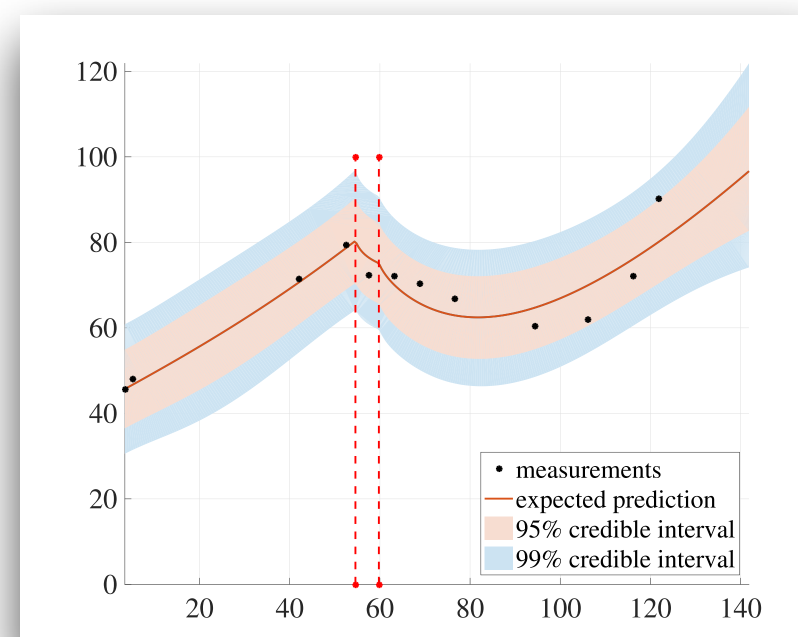
$$\frac{dP}{dt} = \vartheta_4 P(1 - \frac{P^*}{K}) + \vartheta_5 Q_P - \vartheta_3 P - \vartheta_1 \vartheta_2 C P$$

$$\frac{dQ}{dt} = \vartheta_3 P + \vartheta_1 \vartheta_2 C Q$$

$$\frac{dQ_P}{dt} = \vartheta_1 \vartheta_2 C Q - \vartheta_5 Q_P - \vartheta_6 Q_P$$

$$C(0) = 0,\ P(0) = \vartheta_7,\ Q(0) = \vartheta_8,\ Q_P(0) = 0$$

$$P^*(t) = P(t) + Q(t) + Q_P(t)$$



Application

★ data-model assumption

$$d_i \;=\; f(t_i;\vartheta) + \varepsilon, \qquad\qquad \varepsilon \sim \mathcal{N}(0,\sigma)$$

★ likelihood

$$p(\boldsymbol{d}|\vartheta) = \mathcal{N}(\boldsymbol{f}(\cdot\,;\vartheta),\sigma)$$

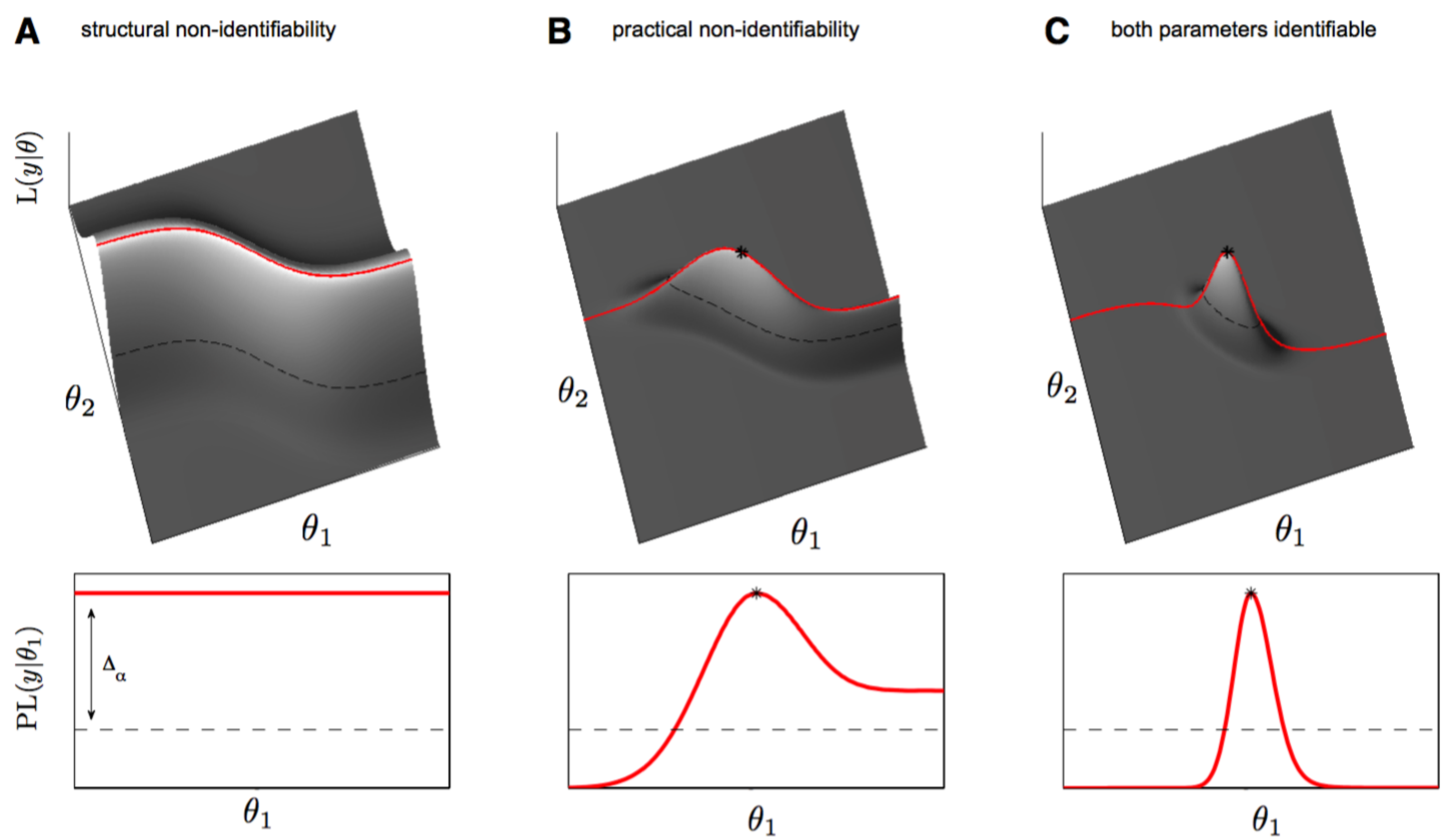
★ optimization

$$\vartheta^\star \;=\; \operatorname*{argmax}_{\vartheta} p(\boldsymbol{d}|\vartheta)$$

APPLICATION
with P. Chadjidoukas

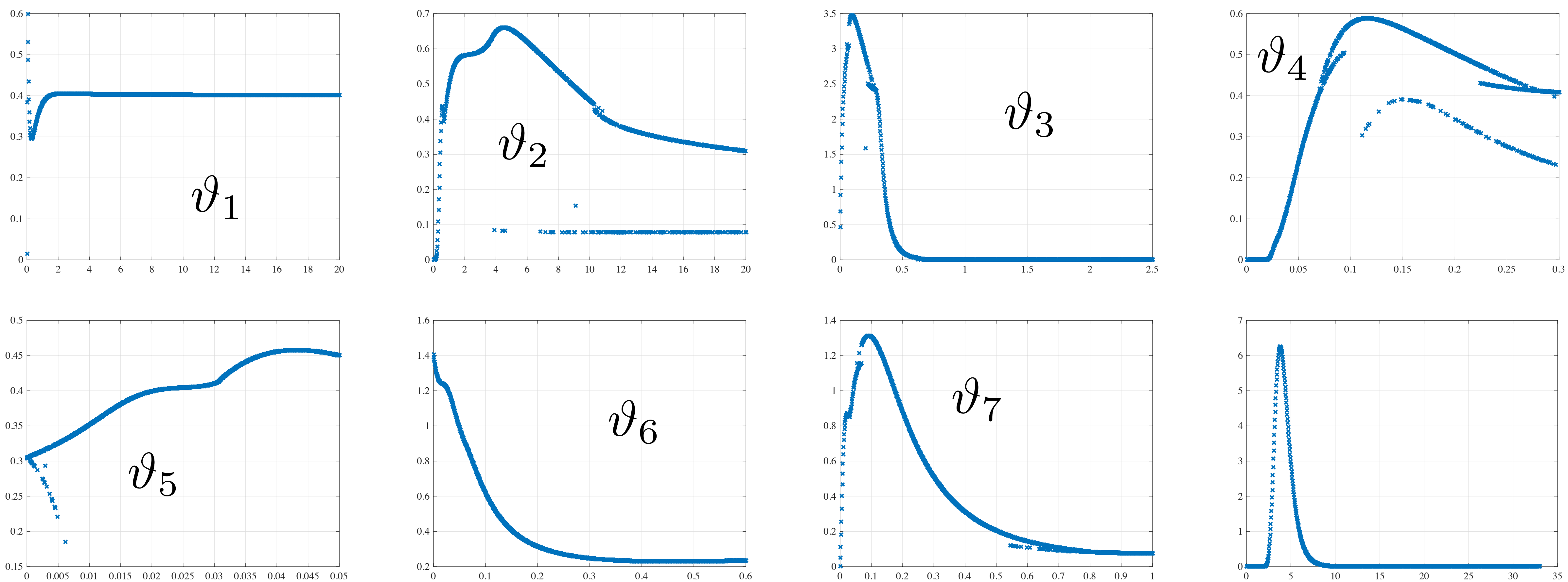
★profile likelihood function

$$q(\boldsymbol{d}|\vartheta_i) \;=\; \max_{\vartheta_{j\neq i}} p(\boldsymbol{d}|\vartheta)$$



Joining forces of Bayesian and frequentist methodology: a study for inference in the presence of non-identifiability, A. Raue et. al.,

APPLICATION
with P. Chadjidoukas



Implementation

- Π4U
- High Performance framework for Uncertainty Quantification
- See the next talk of P. Chadjidoukas

BIBLIOGRAPHY

- Hansen N. and Ostermeier A. 2001. **Completely Derandomized Self-Adaptation in Evolution Strategies**
- Hansen N., Müller S. D., Koumoutsakos P. 2003. **Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES)**
- Hansen N., **The CMA Evolution Strategy: A Tutorial**

THANK YOU