DERIVATIVE FREE OPTIMIZATION 2025/2026

CLASS 3

Exercise on adaptive step-size

Conclusion of question 4: we need to adapt σ.

4. Explain the three phases observed on the figure.

   To accelerate the convergence, we will implement a step-size adaptive algorithm, i.e. σ is not fixed once for all. The method to adapt the step-size is called one-fifth success rule. The pseudo-code of the $(1+1)$-ES with one-fifth success rule is given by:

   Initialize $\boldsymbol{x} \in \mathbb{R}^n$ and $\sigma > 0$
   while not terminate
   $\quad \boldsymbol{x}' = \boldsymbol{x} + \sigma \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
   $\quad$ if $f(\boldsymbol{x}') \leq f(\boldsymbol{x})$
   $\quad\quad \boldsymbol{x} = \boldsymbol{x}'$
   $\quad\quad \sigma = 1.5\,\sigma$
   $\quad$ else
   $\quad\quad \sigma = (1.5)^{-1/4}\sigma$

5. Implement the (1+1)-ES with one-fifth success rule and test the algorithm on the sphere function $f_{\text{sphere}}(x)$ in dimension 5 ($n = 5$) using $\mathbf{x}^0 = (1, \ldots, 1)$, $\sigma_0 = 10^{-3}$ and as stopping criterion a maximum number of function evaluations equal to 6×10^2. Plot the evolution of the square root of the best function value at each iteration versus the number of iterations. Use a logarithmic scale for the y-axis. Compare to the plot obtained on Question 3. Plot also on the same graph the evolution of the step-size.

We observe that the step-size increase in the beginning (it was too small compared to distance to optimum).

Then both $(m_t)_{t\in\mathbb{N}}$ and $(\sigma_t)_{t\in\mathbb{N}}$ "decrease" (not strictly) linearly. We do not observe any more phase III.

Here we can prove on class of function that include convex-quadratic functions.

$$\frac{1}{t} \ln \|m_t - x^*\| \xrightarrow[t\to+\infty]{} -CR \qquad \forall\, m_0, \sigma_0$$

on the sphere

$$\frac{1}{t} \ln \sigma_t \longrightarrow -CR$$

This corresponds to linear convergence.

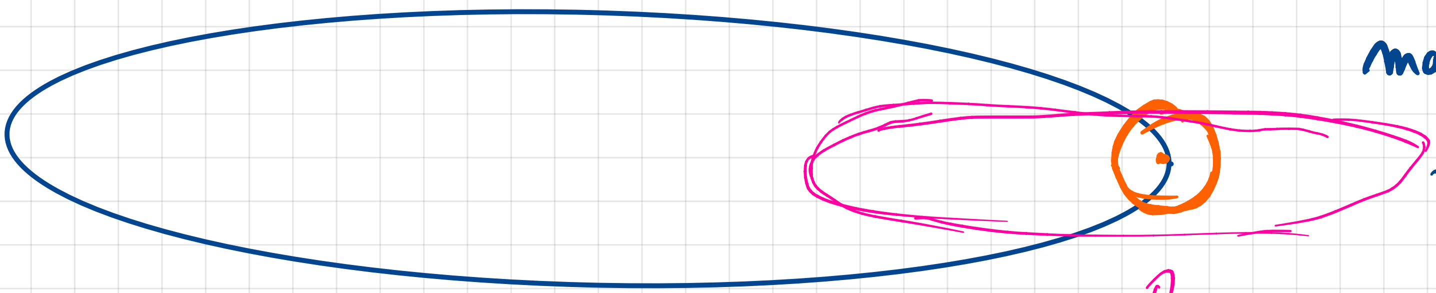
of the step size.

6. Use the algorithm to minimize the function f_{elli} in dimension $n = 5$. Plot the evolution of the objective function value of the best solution versus the number of iterations. Why is the $(1+1)$-ES with one-fifth success much slower on f_{elli} than on f_{sphere} ?

7. Same question with the function

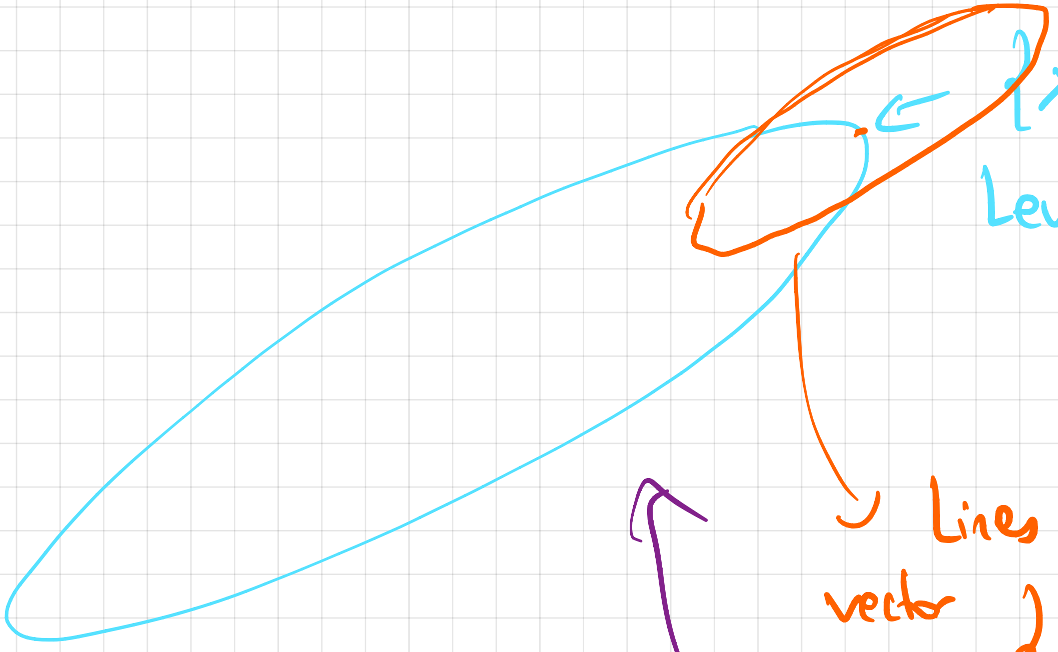
The algorithm still converges linearly but the convergence rate is slower.

The function is ill-conditionned, so sampling with a covariance matrix proportional to the identity is not well adapted.



would be better

Ideally we would like $C_t \propto H^{-1}$



$\{x \mid (x-x^*)^T H (x-x^*) = cste\}$

Level sets.

Lines of equal density of Gaussian vector $\{x \mid (x-m)^T C^{-1}(x-m) = cste\}$

$\mathcal{N}(m, C)$: density : $\dfrac{1}{\sqrt{2\pi}\,|C|^{1/2}} \exp\left(-\frac{1}{2}(x-m)^T C^{-1}(x-m)\right)$

To be in the above scenario I need

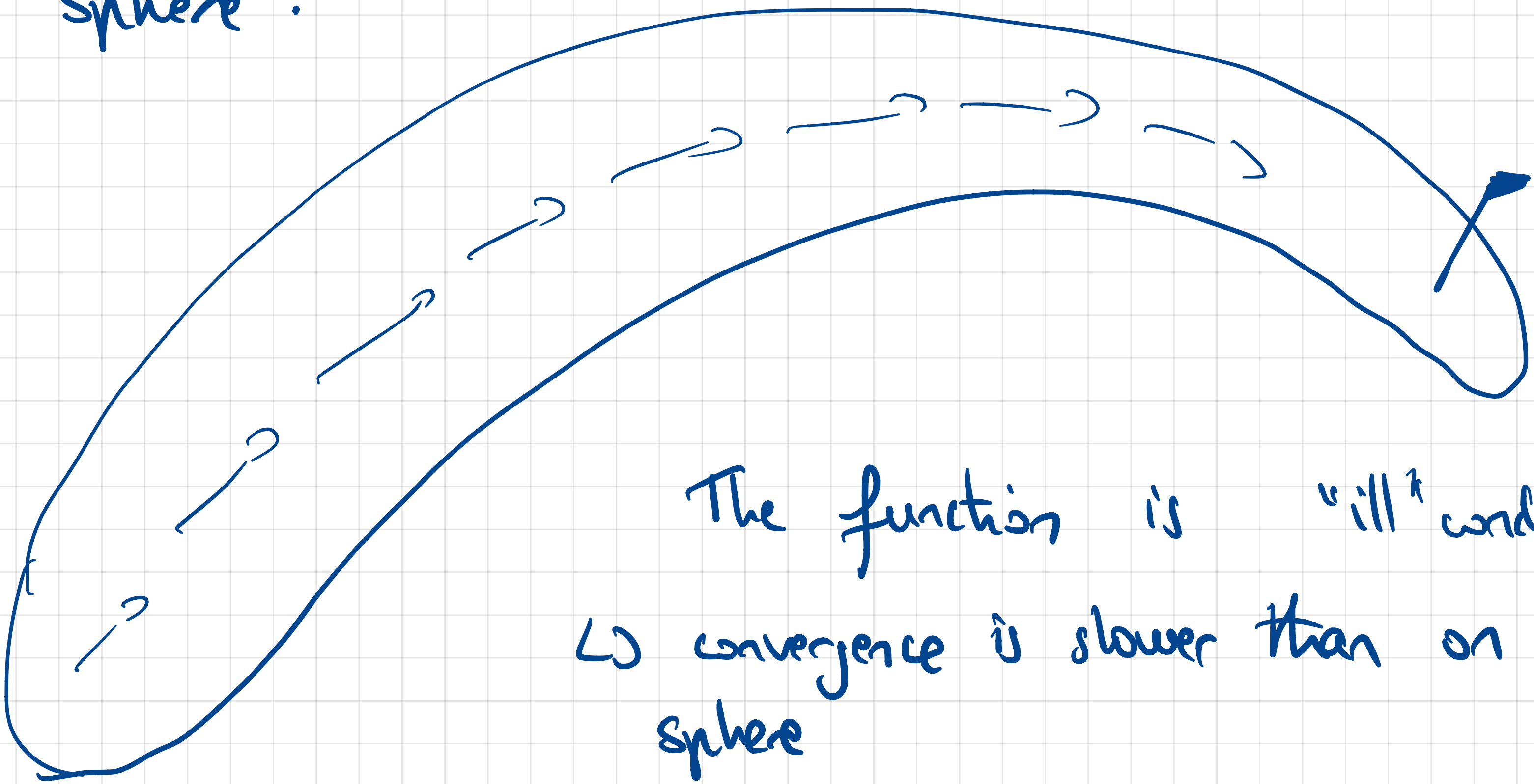
$$\boxed{C \propto H^{-1}}$$

7. Same question with the function

$$f_{\text{Rosenbrock}}(x) = \sum_{i=1}^{n-1} \left(100(x_i^2 - x_{i+1})^2 + (x_i - 1)^2\right) .$$

Starting in $m_0 \neq \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$, we observe that it converges slower than on the sphere.

Banana function.



The function is "ill" conditioned

$\hookrightarrow$ convergence is slower than on the sphere

$i{=}1$

. We now consider the functions, $g(f_{\text{sphere}})$ and $g(f_{\text{elli}})$ where $g : \mathbb{R} \to \mathbb{R}, y \mapsto y^{1/4}$. Modify your implementation in Questions 5 and 6 so as to save at each iteration the distance between $\mathbf{x}$ and the optimum. Plot the evolution of the distance to the optimum versus the number of function evaluations on the functions f_{sphere} and $g(f_{\text{sphere}})$ as well as on the functions f_{elli} and $g(f_{\text{elli}})$. What do you observe? Explain.

Observations: On fsphere versus g(fsphere), the graphs are closed to each other, sometimes one looks above and sometimes the other one is above.

On felli or g(felli), typically one is above, but from one trial to the next one sometims felli is above, someting g(felli) is above.

Note: $g: \mathbb{R}_{\geq 0} \to \mathbb{R}$, $y \mapsto y^{1/4}$ is stricly increasing

1/ If the same sequence of random vectors ($\mathcal{N}(0, Id)$) are used when optimizing $\left\{ \begin{array}{l} f_{sphere} \\ f_{elli} \end{array} \right.$ or $\begin{array}{l} g(f_{sphere}) \\ g(f_{elli}) \end{array}$ we will generate the same sequence $\left\{ \begin{array}{l} (m_t)_{t \in \mathbb{N}} \\ (s_t)_{t \in \mathbb{N}} \end{array} \right.$

Therefore the differences observed are due to stochasticity (the fact that we chose different random number sequences).

⚠ If we display $f_{elli}(m_t)$ and $g(f_{elli}(m_t))$ even with the same random numbers, we will observe something different since $f_{elli}(x) \neq g(f_{elli}(x))$

To fix the random sequence, we can fix the seed.

Why Step-size Adaptation?

Assume a (1+1)-ES algorithm with fixed step-size σ (and $C = I_d$) optimizing the function $f(x) = \sum_{i=1}^{n} x_i^2 = \|x\|^2$.

Initialize $\mathbf{m}, \sigma$

While *(stopping criterion not met)*

sample new solution:

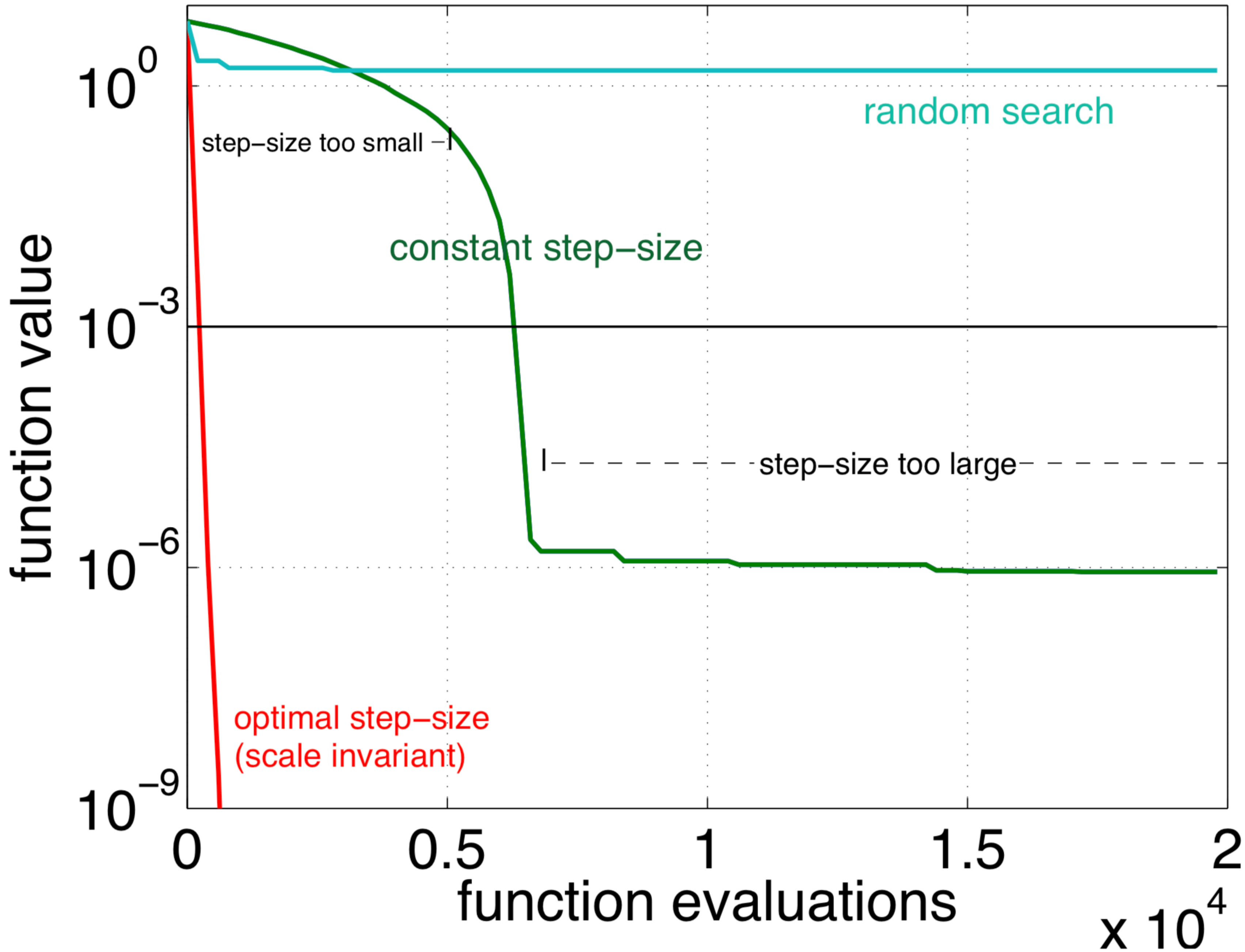
$$\mathbf{x} \leftarrow \mathbf{m} + \sigma \mathcal{N}(0, I_d)$$

if $f(\mathbf{x}) \leq f(\mathbf{m})$

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

What will happen if you look at the convergence of $f(m)$?

Why Step-size Adaptation?



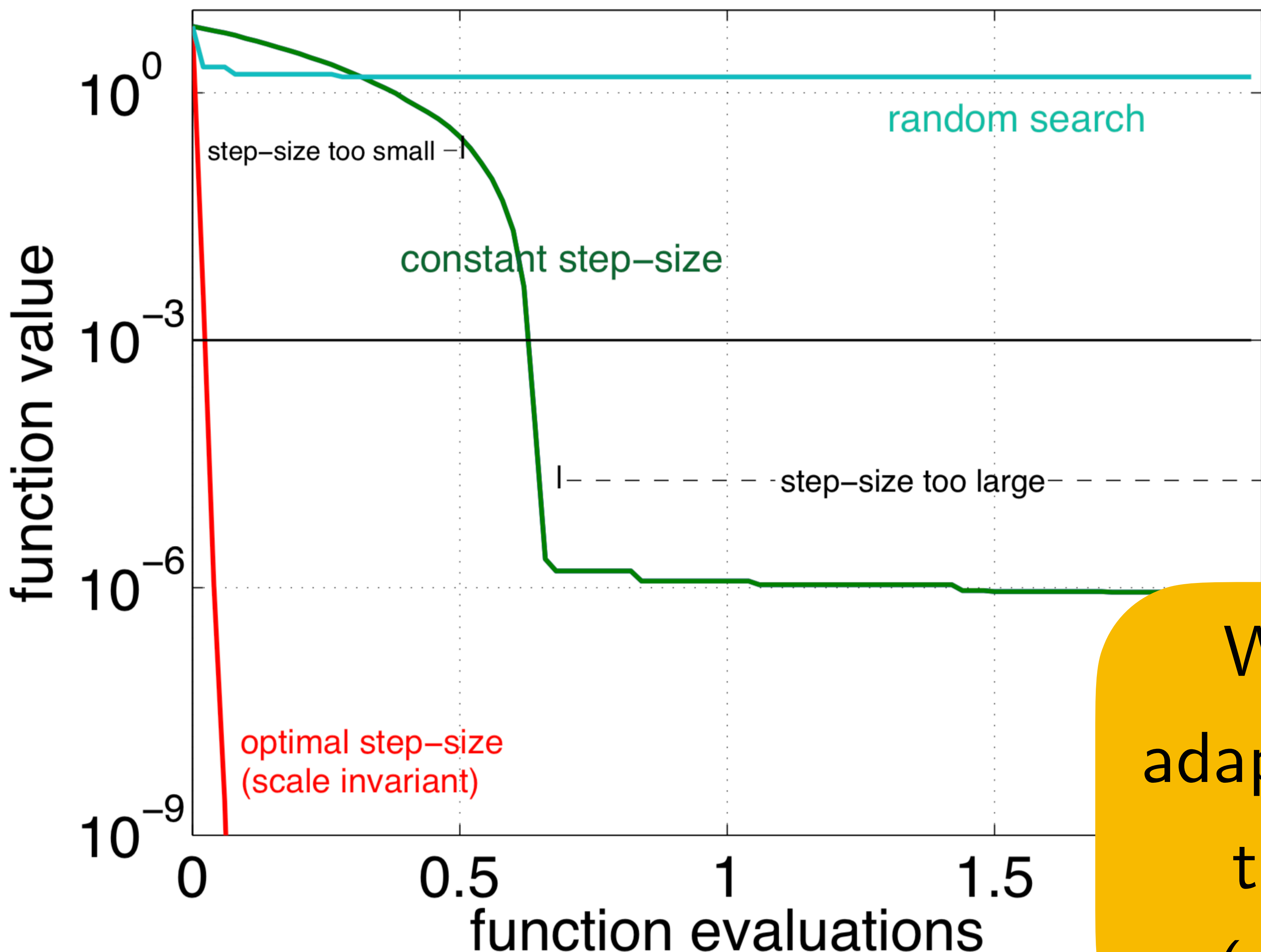
(1+1)-ES
(red & green)

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} x_i^2$$

in $[-2.2, 0.8]^n$
for $n = 10$

red curve: (1+1)-ES with optimal step-size (see later)
green curve: (1+1)-ES with constant step-size ($\sigma = 10^{-3}$)

Why Step-size Adaptation?



(1+1)-ES
(red & green)

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} x_i^2$$

We need step-size adaptation to approach the optimum fast (converge linearly)

red curve: (1+1)-ES with optimal step-size (see later)

green curve: (1+1)-ES with constant step-size ($\sigma = 10^{-3}$)

Methods for Step-size Adaptation

1/5th success rule, typically applied with "$+$" selection

[Rechenberg, 73][Schumer and Steiglitz, 78][Devroye, 72]

σ-self adaptation, applied with "," selection [Schwefel, 81]

random variation is applied to the step-size and the better one, according to the objective function value, is selected

path-length control or Cumulative step-size adaptation (CSA), applied with "," selection

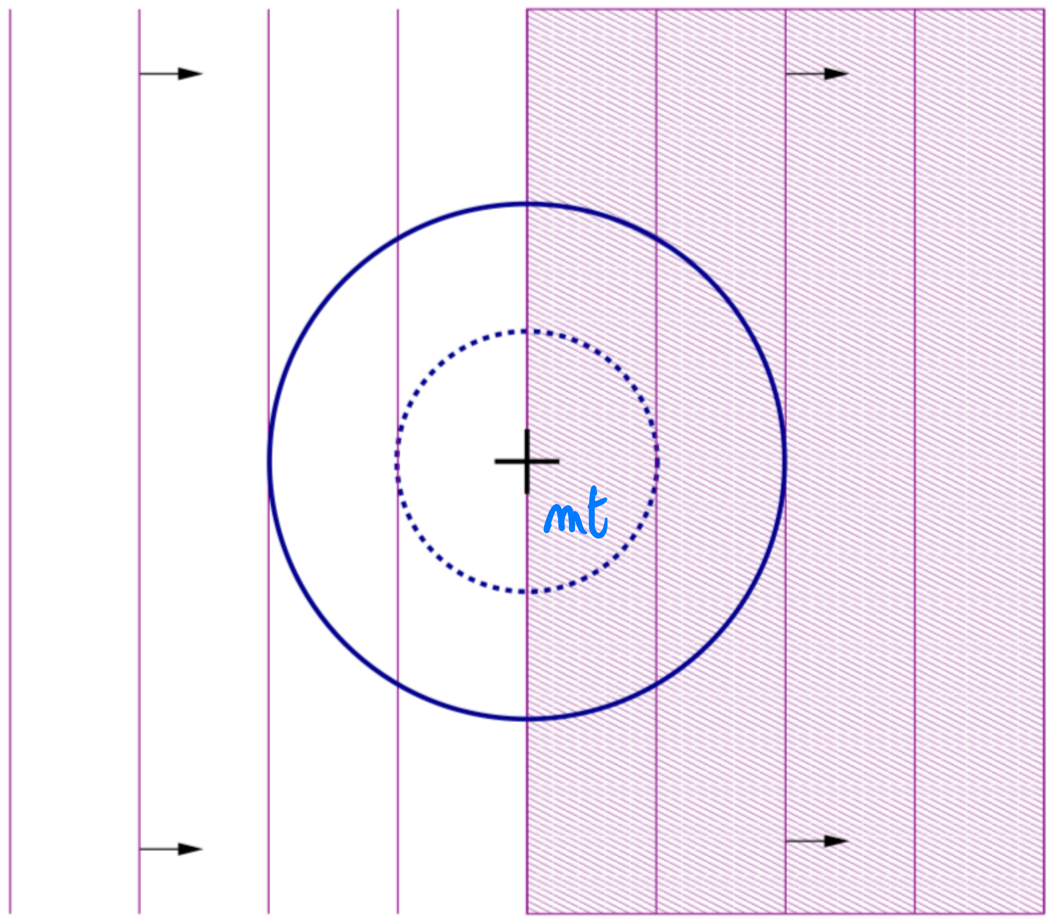
[Ostermeier et al. 84][Hansen, Ostermeier, 2001]

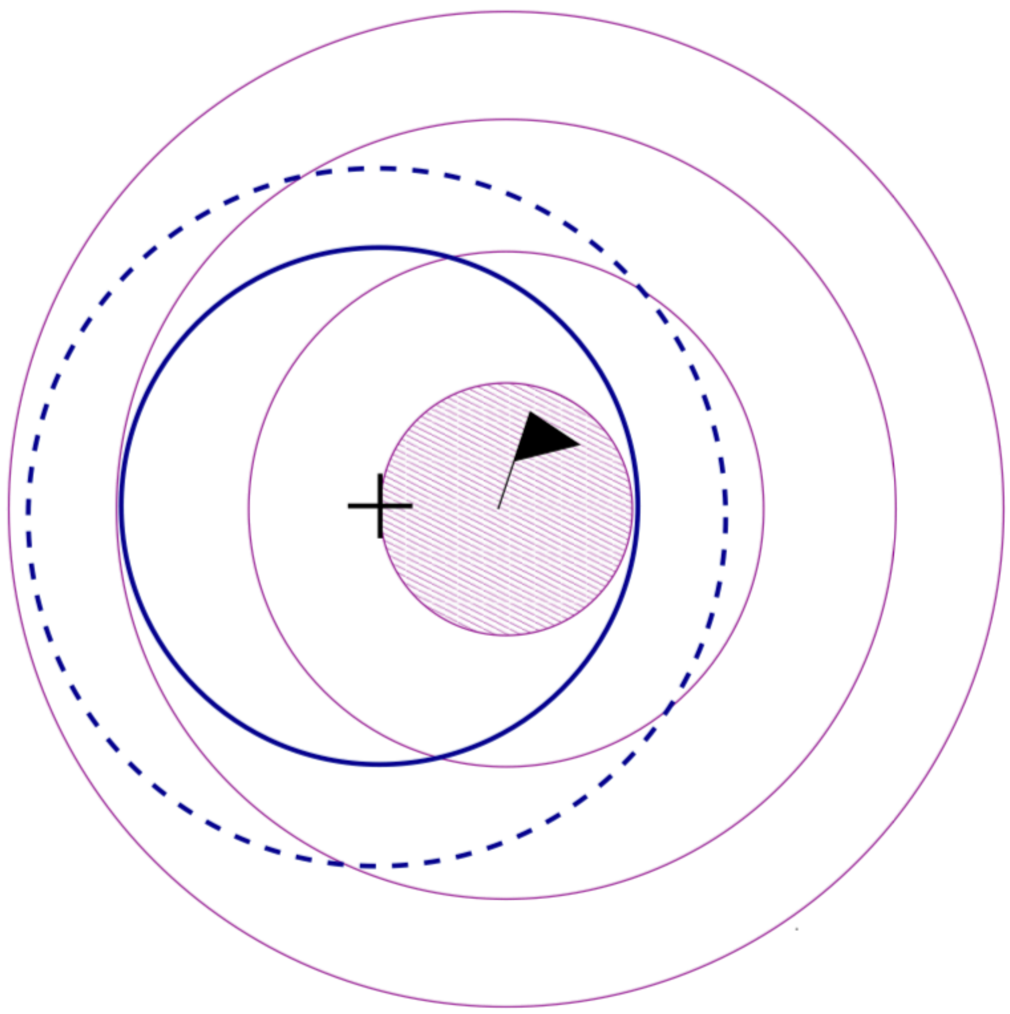
two-point adaptation (TPA), applied with "," selection [Hansen 2008]

test two solutions in the direction of the mean shift, increase or decrease accordingly the step-size

Step-size control: 1/5th Success Rule

$f(x) = x_1$





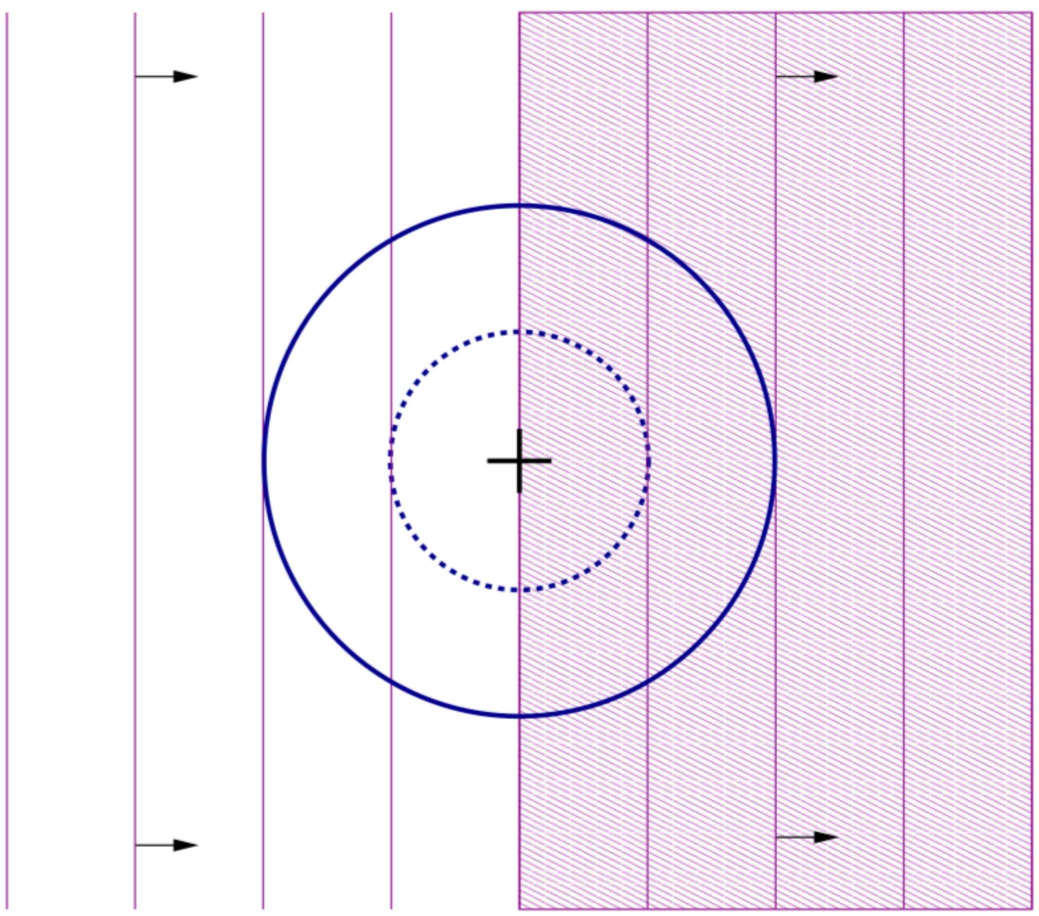
$\downarrow$ increase σ

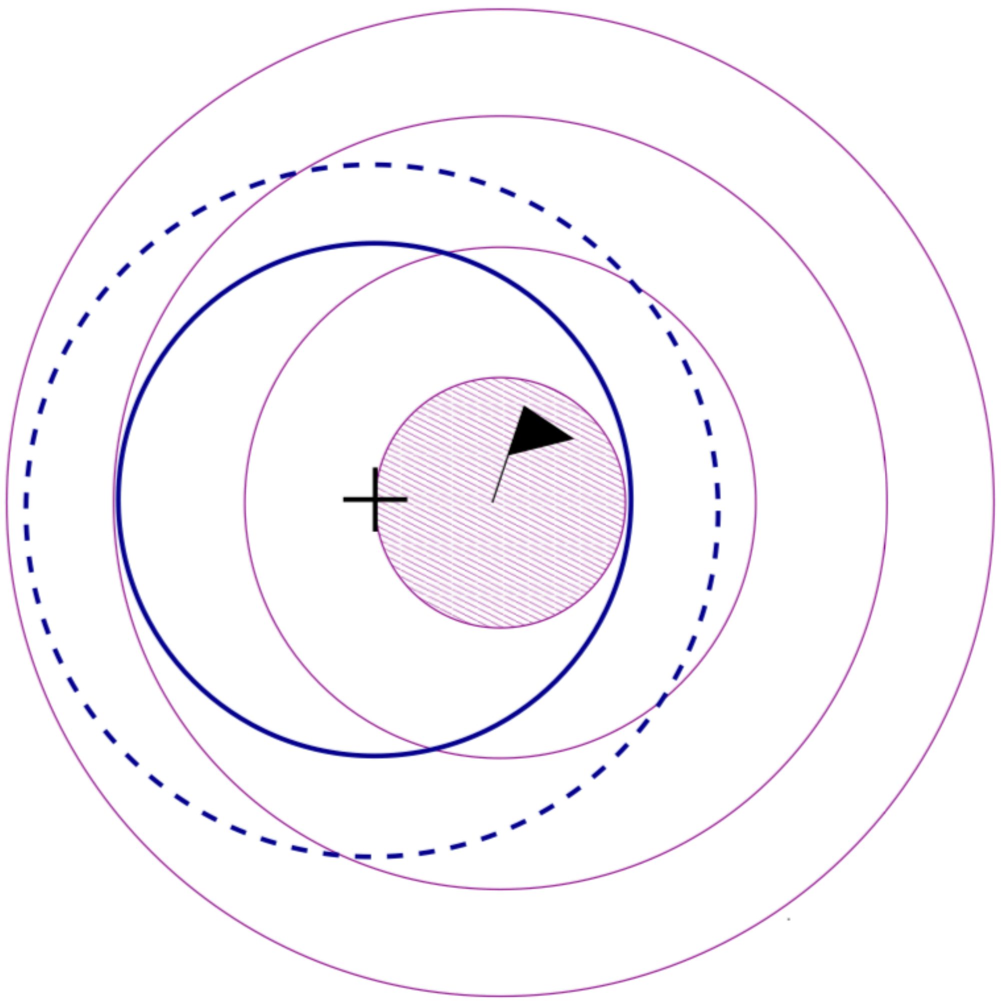
$\downarrow$ decrease σ

$p_s = \frac{1}{2}$

p_s small

Step-size control: 1/5th Success Rule





Probability of success (p_s)

1/2

1/5

Probability of success (p_s)

"too small"

Step-size control: 1/5th Success Rule

probability of success per iteration:

$$\text{ps} = \frac{\#\text{candidate solutions better than } m}{\#\text{candidate solutions}}$$

$$\sigma \leftarrow \sigma \times \exp\left(\frac{1}{3} \times \frac{p_s - p_{\text{target}}}{1 - p_{\text{target}}}\right)$$

Increase σ if $p_s > p_{\text{target}}$
Decrease σ if $p_s < p_{\text{target}}$

$\approx \frac{1}{5}$

$(1+1)$-ES

$p_{target} = 1/5$

IF *offspring better parent* $[f(\mathbf{x}) \leq f(\mathbf{m})]$

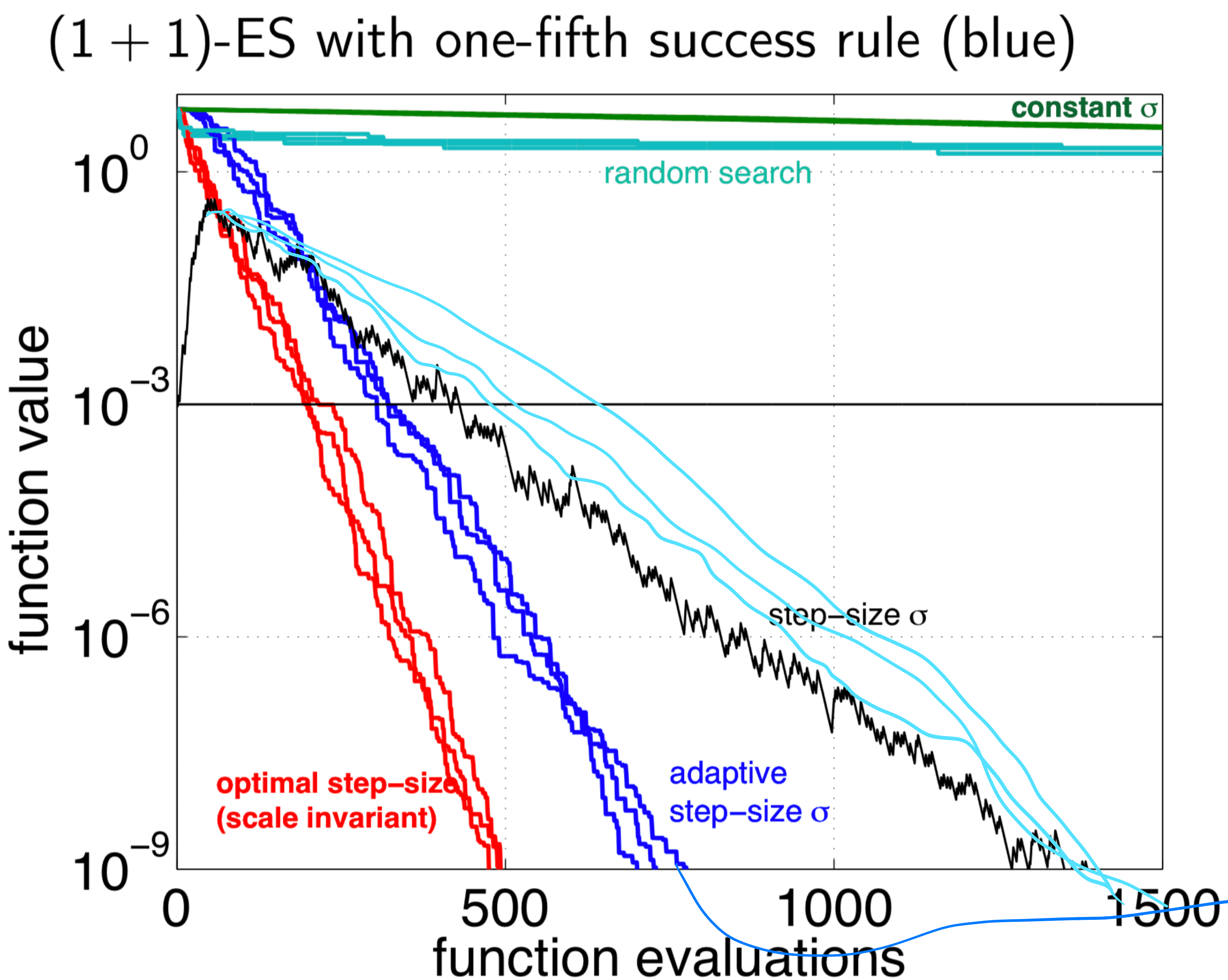
$p_s = 1,\ \sigma \leftarrow \sigma \times \underbrace{\exp(1/3)}_{1,5}$ In the exercise

ELSE

$p_s = 0,\ \sigma \leftarrow \sigma / \exp(1/3)^{1/4}$

$\sigma \leftarrow \sigma\,(1{,}5)^{-1/4}$

(1+1)-ES with One-fifth Success Rule - Convergence



$$f(x) = \sqrt{\sum_{i=1}^{n} x_i^2} = \|x\|$$

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} x_i^2$$

in $[-0.2, 0.8]^n$
for $n = 10$

$\ln \|m_t\|^2 = 2 \ln \|m_t\|$

Linear convergence

Path Length Control - Cumulative Step-size Adaptation (CSA)

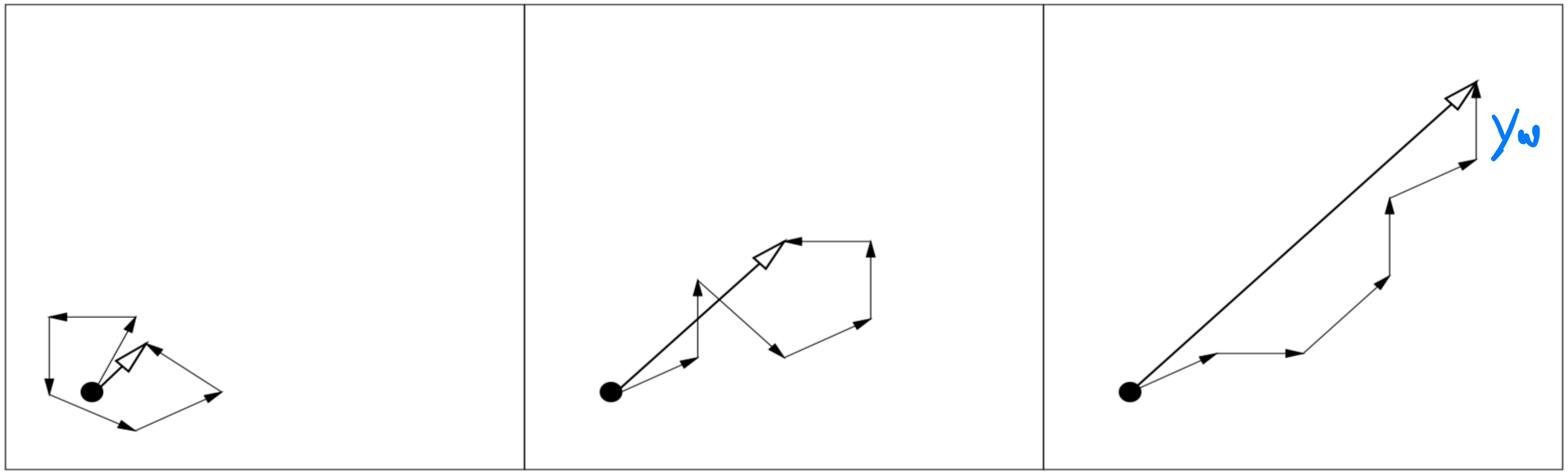
step-size adaptation used in the $(\mu/\mu_w, \lambda)$-ES algorithm framework (in CMA-ES in particular)

Main Idea:

$$\begin{aligned} \boldsymbol{x}_i &= \boldsymbol{m} + \sigma\, \boldsymbol{y}_i \\ \boldsymbol{m} &\leftarrow \boldsymbol{m} + \sigma\, \boldsymbol{y}_w \end{aligned}$$

Measure the length of the *evolution path*

the pathway of the mean vector $\boldsymbol{m}$ in the iteration sequence



$\Downarrow$		$\Downarrow$
decrease σ		increase σ

Sampling of solutions, notations as on slide "The $(\mu/\mu, \lambda)$-ES - Update of the mean vector" with $\mathbf{C}$ equal to the identity.

Initialize $\boldsymbol{m} \in \mathbb{R}^n$, $\sigma \in \mathbb{R}_+$, evolution path $\boldsymbol{p}_\sigma = \mathbf{0}$,
set $c_\sigma \approx 4/n$, $d_\sigma \approx 1$.

$x_i = m + \sigma y_i \quad , i = 1, \dots, \lambda$

$/\!/ \; \sum_{i=1}^{\mu} w_i x_{i:\lambda} = m + \sigma \sum_{i=1}^{\mu} y_{i:\lambda} \qquad f(x_{1:\lambda}) \le f(x_{2:\lambda}) \le \dots \le f(x_{\lambda:\lambda})$

$$\boldsymbol{m} \leftarrow \boldsymbol{m} + \sigma \boldsymbol{y}_w \quad \text{where } \boldsymbol{y}_w = \sum_{i=1}^{\mu} w_i \boldsymbol{y}_{i:\lambda} \qquad \text{update mean}$$

$$\boldsymbol{p}_\sigma \leftarrow (1 - c_\sigma)\boldsymbol{p}_\sigma + \underbrace{\sqrt{1-(1-c_\sigma)^2}}_{\text{accounts for } 1-c_\sigma} \; \underbrace{\sqrt{\mu_w}}_{\text{accounts for } w_i} \; \boldsymbol{y}_w$$

$$\sigma \leftarrow \sigma \times \underbrace{\exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|\boldsymbol{p}_\sigma\|}{\mathsf{E}\|\mathcal{N}(\mathbf{0},\mathbf{I})\|} - 1\right)\right)}_{>1 \iff \|\boldsymbol{p}_\sigma\| \text{ is greater than its expectation}} \qquad \text{update step-size}$$

In CSA, the scenario where we do not want to increase or decrease the step-size corresponds to a function that does not return any information, for instance

$$f(x) = \text{rand} \qquad (\text{independent of } x,$$

$$f(x^1_{t+1}), \ldots, f(x^\lambda_{t+1}))$$

$$\text{rand}^1, \ldots, \text{rand}^\lambda \quad \text{where } \text{rand}^i \text{ are } \underline{\text{iid}},$$

Now assume that the path p_σ at iteration t equals

$$p^\sigma_{t+1} = (1 - c_\sigma) p^\sigma_t + \alpha_{\Gamma} \sum_{i=1}^{\lambda} w_i y_{i:\lambda}$$

The constant α is computed such that, if f is random, if $p_t^\sigma \sim \mathcal{N}(0, I_d)$, then $p_{t+1}^\sigma \sim \mathcal{N}(0, I_d)$

Assume that

$$p_{t+1}^\sigma = (1 - c_\sigma)\, p_t^\sigma + \sqrt{1 - (1-c_\sigma)^2}\, \sqrt{\mu_w} \sum_{i=1}^{\mu} w_i\, y_{i:\lambda}$$

$$\mu_w = \frac{1}{\sum w_i^2}$$

Proposition :

If $p_t^\sigma \sim \mathcal{N}(0, I_d)$, if f = random, then $p_{t+1}^\sigma \sim \mathcal{N}(0, I_d)$.

Let us write the CSA - ES with time index notations:

$$\theta_t = (m_t, \sigma_t\ p_t^{\sigma})$$

1) Sample candidate solutions:

$$X_{t+1}^{i} = m_t + \sigma_t\ Y_{t+1}^{i}$$

$Y_{t+1}^{i} \sim \mathcal{N}(0, I_d)$

$(Y_{t+1}^{1}, \ldots, Y_{t+1}^{\lambda})$ i.i.d.

2) Evaluate on f:

$$f(X_{t+1}^{1:\lambda}) \leq \ldots \leq f(X_{t+1}^{\lambda:\lambda})$$

$\sum_{i=1}^{\mu} w_i = 1$

$\mu = \lfloor \frac{\lambda}{2} \rfloor$

$w_1 \geq w_2 \geq \ldots \geq w_\mu > 0$

3) Update θ_t:

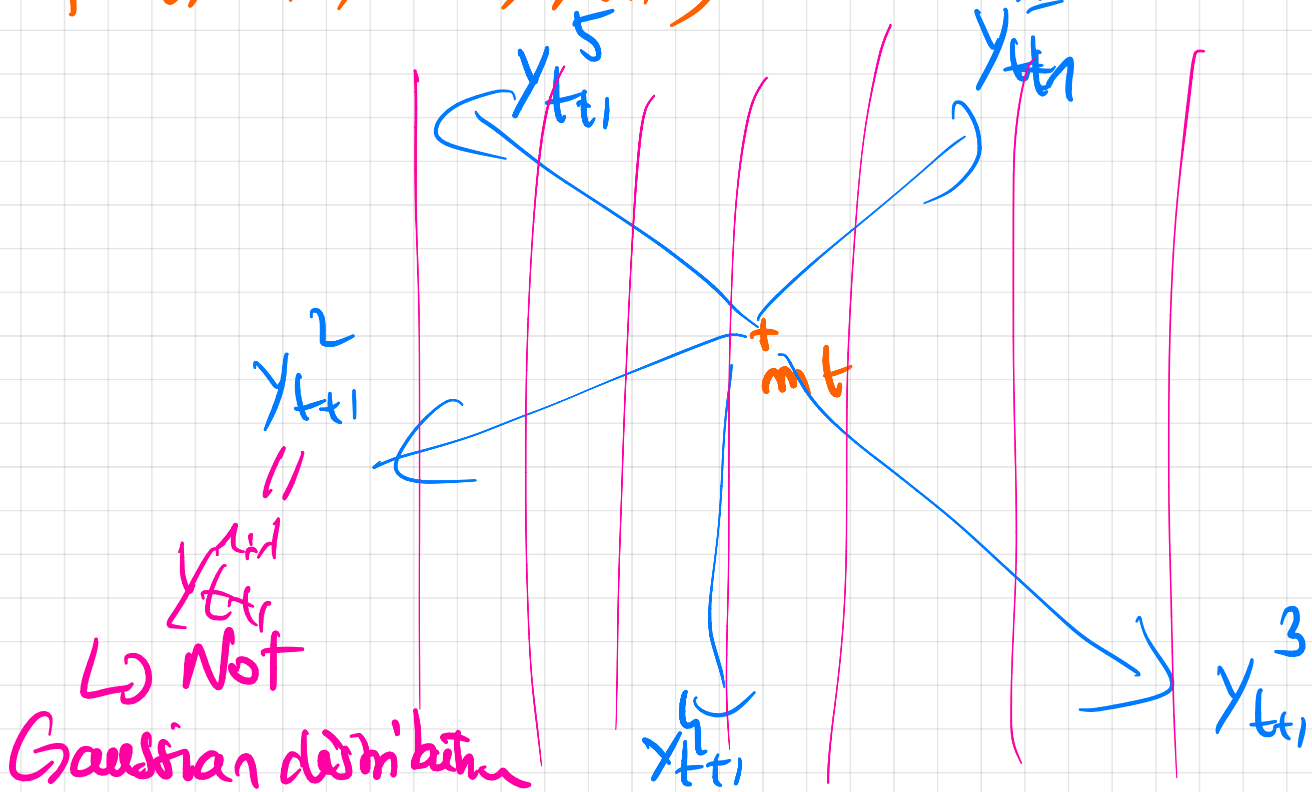
$$m_{t+1} = m_t + \sigma_t \sum_{i=1}^{\mu} w_i\ Y_{t+1}^{i:\lambda}$$

$$p_{t+1}^{\sigma} = (1 - c_\sigma)\ p_t^{\sigma} + \sqrt{1 - (1 - c_\sigma)^2}\ \sqrt{\mu_w} \sum_{i=1}^{\mu} w_i\ Y_{t+1}^{i:\lambda}$$

$$\sigma_{t+1} = \sigma_t \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|p_{t+1}^{\sigma}\|}{E(\|\mathcal{N}(0, I_d)\|)} - 1\right)\right)$$

Lemma: If $f(x) = \text{rand}$, $(y_{t+1}^{1:\lambda}, \dots, y_{t+1}^{\mu:\lambda})$ is distributed according to $(\mathcal{N}^1(0, I_d), \dots, \mathcal{N}^\mu(0, I_d))$ [μ Gaussian vectors $\mathcal{N}(0, I_d)$ that are independent]

Remark: if $f(x) = x_1$, then the selection bias the distribution of $(y_{t+1}^{1:\lambda}, \dots, y_{t+1}^{\mu:\lambda})$



In general after selection on f

$\left(Y_{t+1}^{1:\lambda}, \dots, Y_{t+1}^{\mu:\lambda}\right)$ is NOT DISTRIBUTED

According to $\left(\mathcal{N}^1(0, I_d), \dots, \mathcal{N}^\mu(0, I_d)\right)$

If $f(x)$: random $\left(Y_{t+1}^{i:\lambda}\right)_{i=1,\dots,\mu} \sim \left(\mathcal{N}(0, I_d), \dots, \mathcal{N}(0, I_d)\right)$

Therefore $\sum_{i=1}^{\mu} w_i Y_{t+1}^{i:\lambda} \sim \mathcal{N}\left(0, \sum_{i=1}^{\mu} w_i^2 I_d\right)$

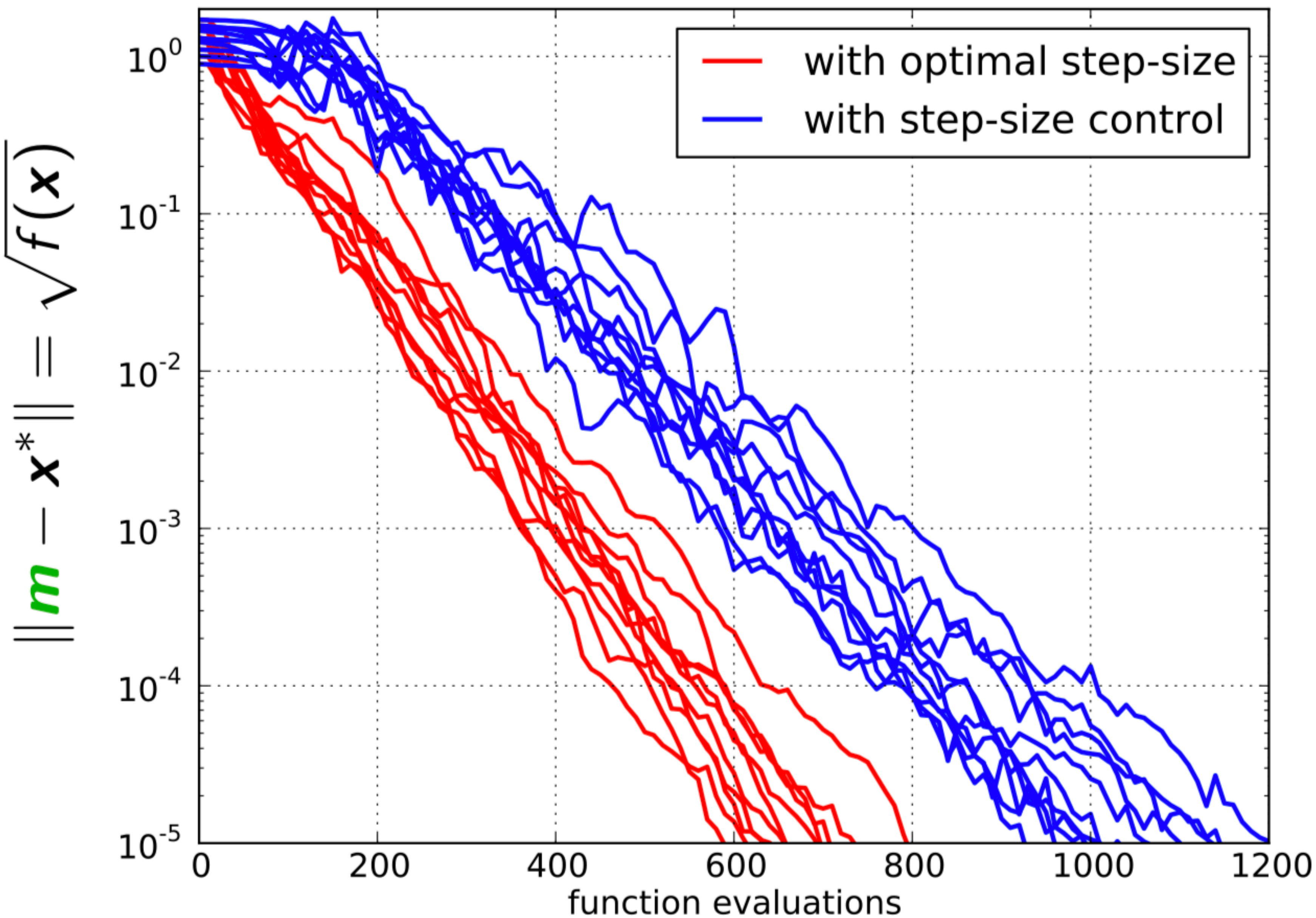
$$\sqrt{\mu_w} \sum_{i=1}^{\mu} w_i Y_{t+1}^{i:\lambda} \sim \mathcal{N}(0, I_d) \qquad \sqrt{\mu_w} = \frac{1}{\sqrt{\sum w_i^2}}$$

We have $p_{t+1}^{\sigma} = (1 - c_\sigma) p_t^\sigma + \sqrt{1-(1-c_\sigma)^2} \underbrace{\left(\sqrt{\mu_w} \sum w_i Y_{t+1}^{i:\lambda}\right)}_{\sim \mathcal{N}(0, I_d)}$

If $p_t^{\sigma} \sim \mathcal{N}(0, Id)$, then $p_{t+1}^{\sigma} \sim \underbrace{\mathcal{N}\left(0, \left[(1-c_\sigma)^2 + (1-(1-c_\sigma)^2)\right] Id\right)}_{\mathcal{N}(0,\ Id)}$

Convergence of $(\mu/\mu_w, \lambda)$-CSA-ES

2x11 runs



$$f(\boldsymbol{x}) = \sum_{i=1}^{n} x_i^2$$

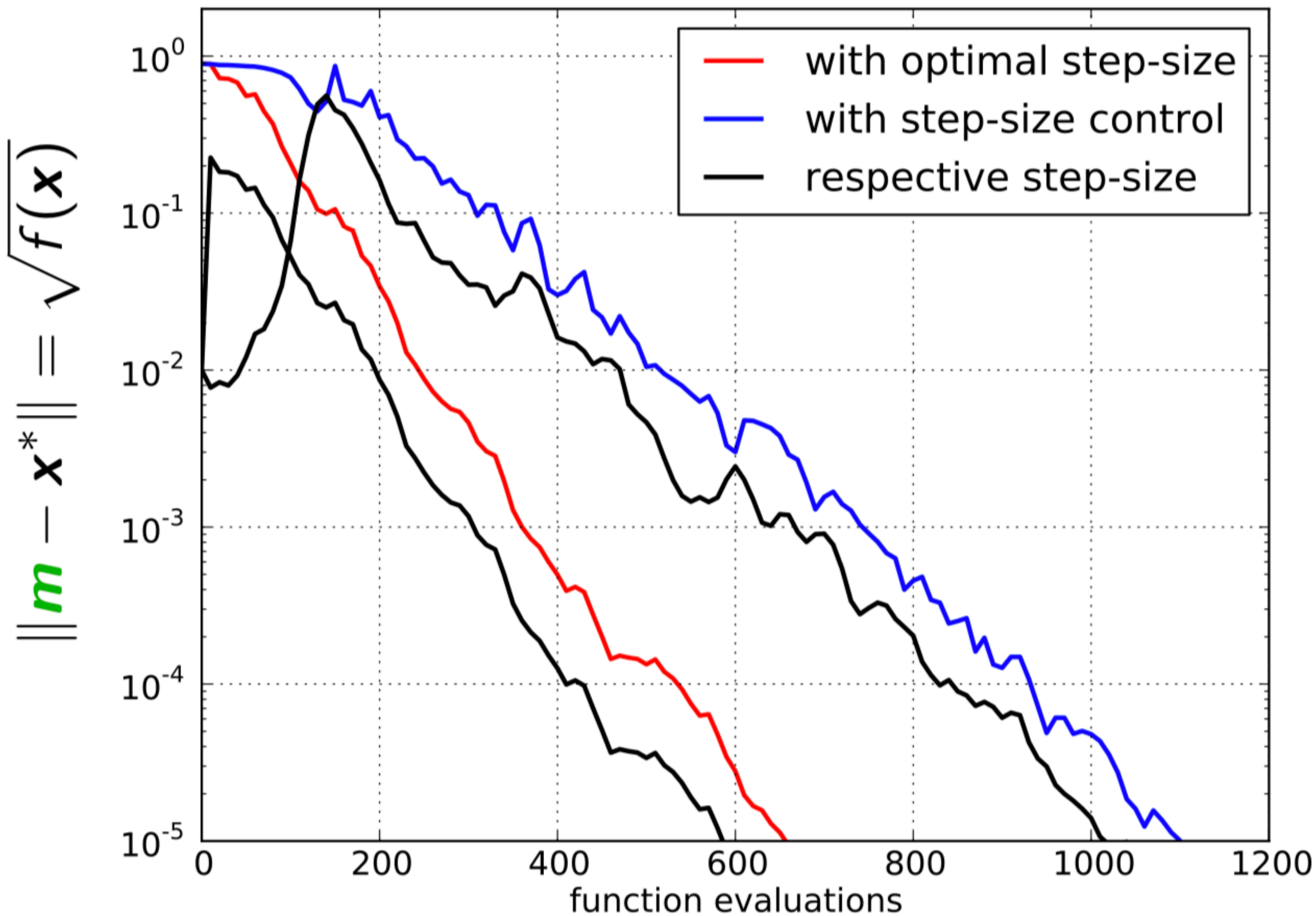
for $n = 10$

and

$\boldsymbol{x}^0 \in [-0.2, 0.8]^n$

with optimal versus adaptive step-size σ with too small initial σ

Convergence of $(\mu/\mu_w, \lambda)$-CSA-ES



$$f(\boldsymbol{x}) = \sum_{i=1}^{n} x_i^2$$

for $n = 10$

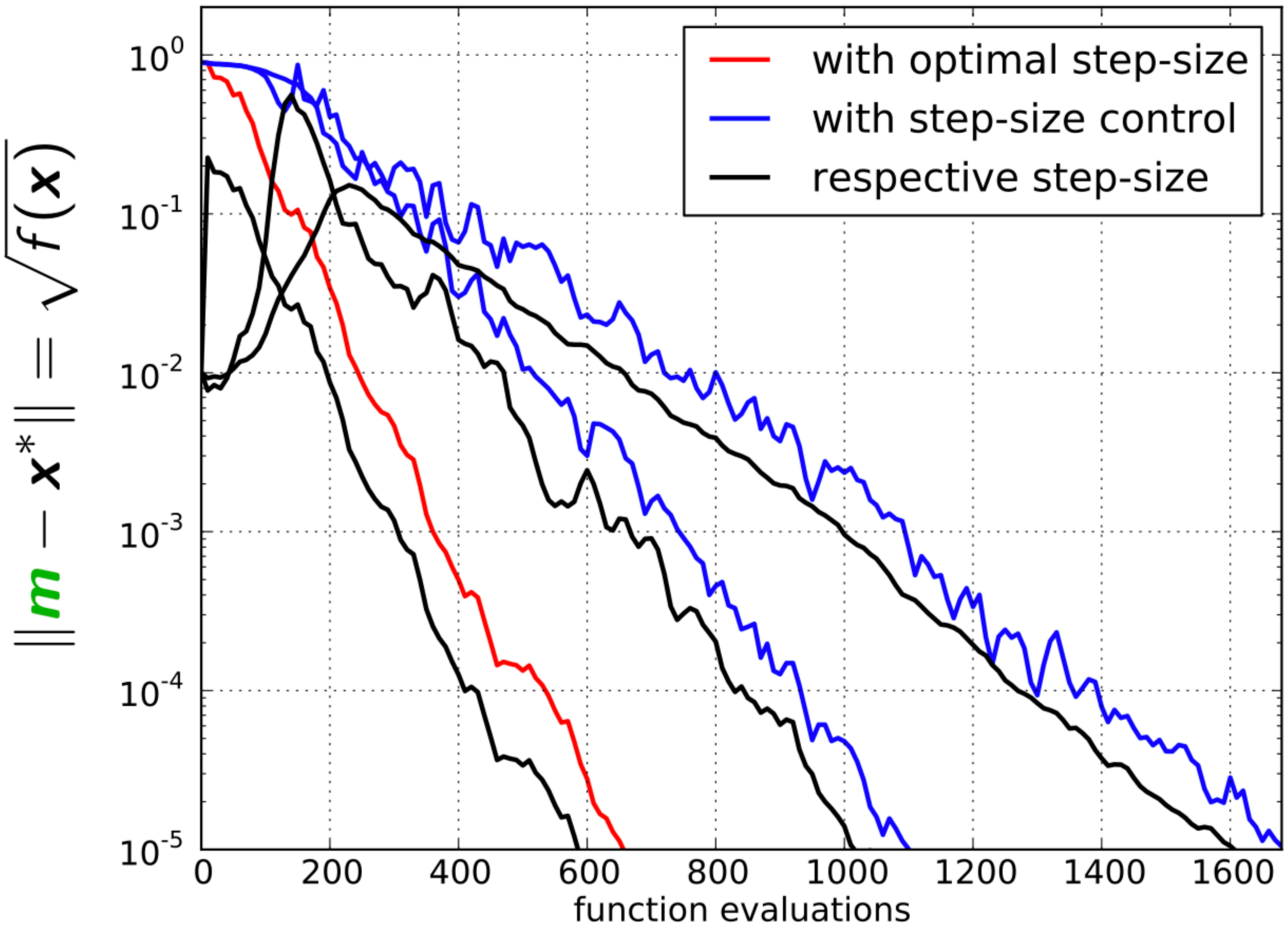
and

$\boldsymbol{x}^0 \in [-0.2, 0.8]^n$

comparing number of f-evals to reach $\|\boldsymbol{m}\| = 10^{-5}$: $\frac{1100-100}{650} \approx 1.5$

Note: initial step-size taken too small ($\sigma_0 = 10^{-2}$) to illustrate the step-size adaptation

Convergence of $(\mu/\mu_w, \lambda)$-CSA-ES



$$f(\boldsymbol{x}) = \sum_{i=1}^{n} x_i^2$$

for $n = 10$
and
$\boldsymbol{x}^0 \in [-0.2, 0.8]^n$

comparing optimal versus default damping parameter d_σ:
$\frac{1700}{1100} \approx 1.5$