How to update the different parameters $m, \sigma, \mathbf{C}$?

1. Adapting the mean m
2. **Adapting the step-size σ**
3. Adapting the covariance matrix C

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:

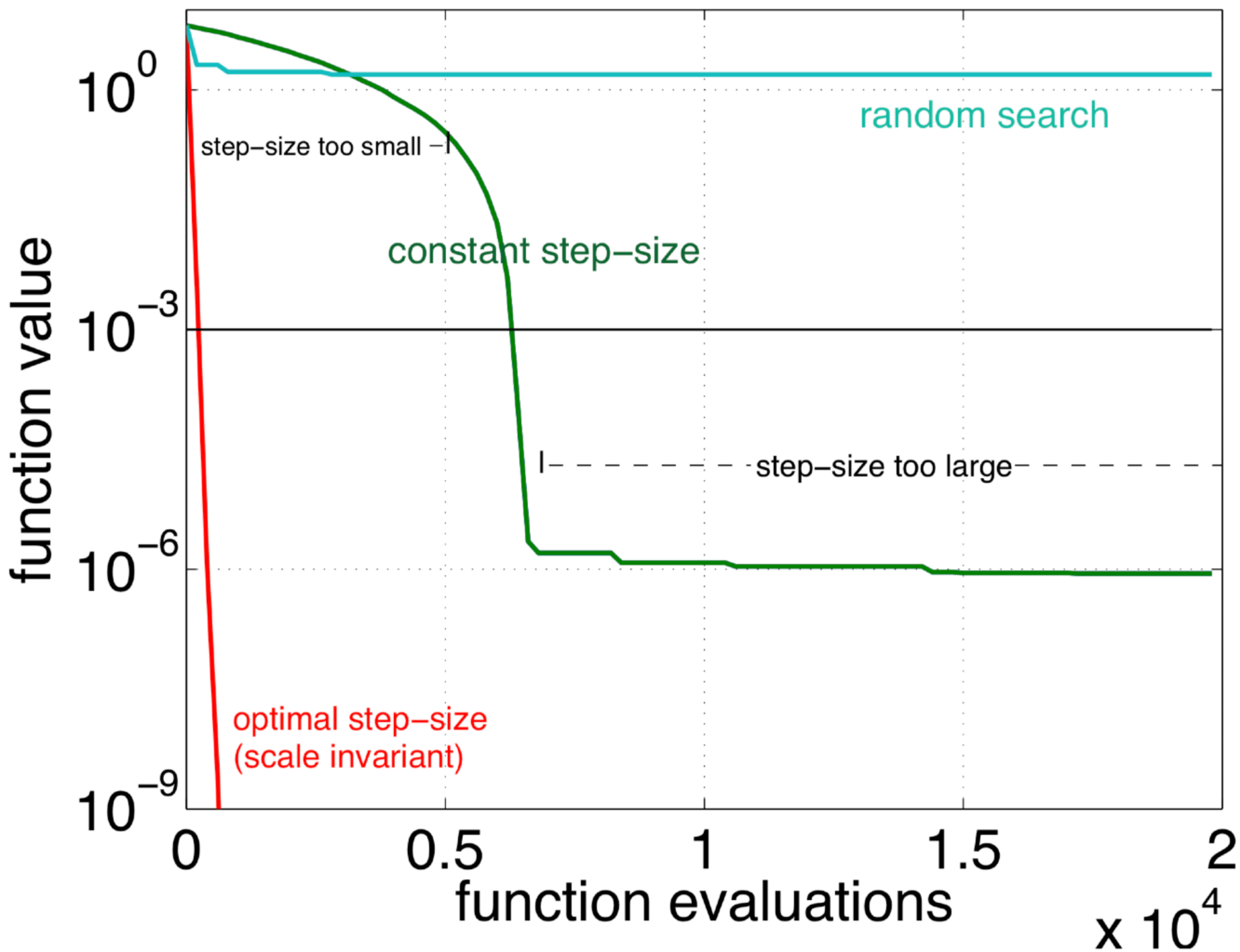
$$\mathbf{x} \leftarrow \mathbf{m} + \sigma \mathcal{N}(0, I_d)$$

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?



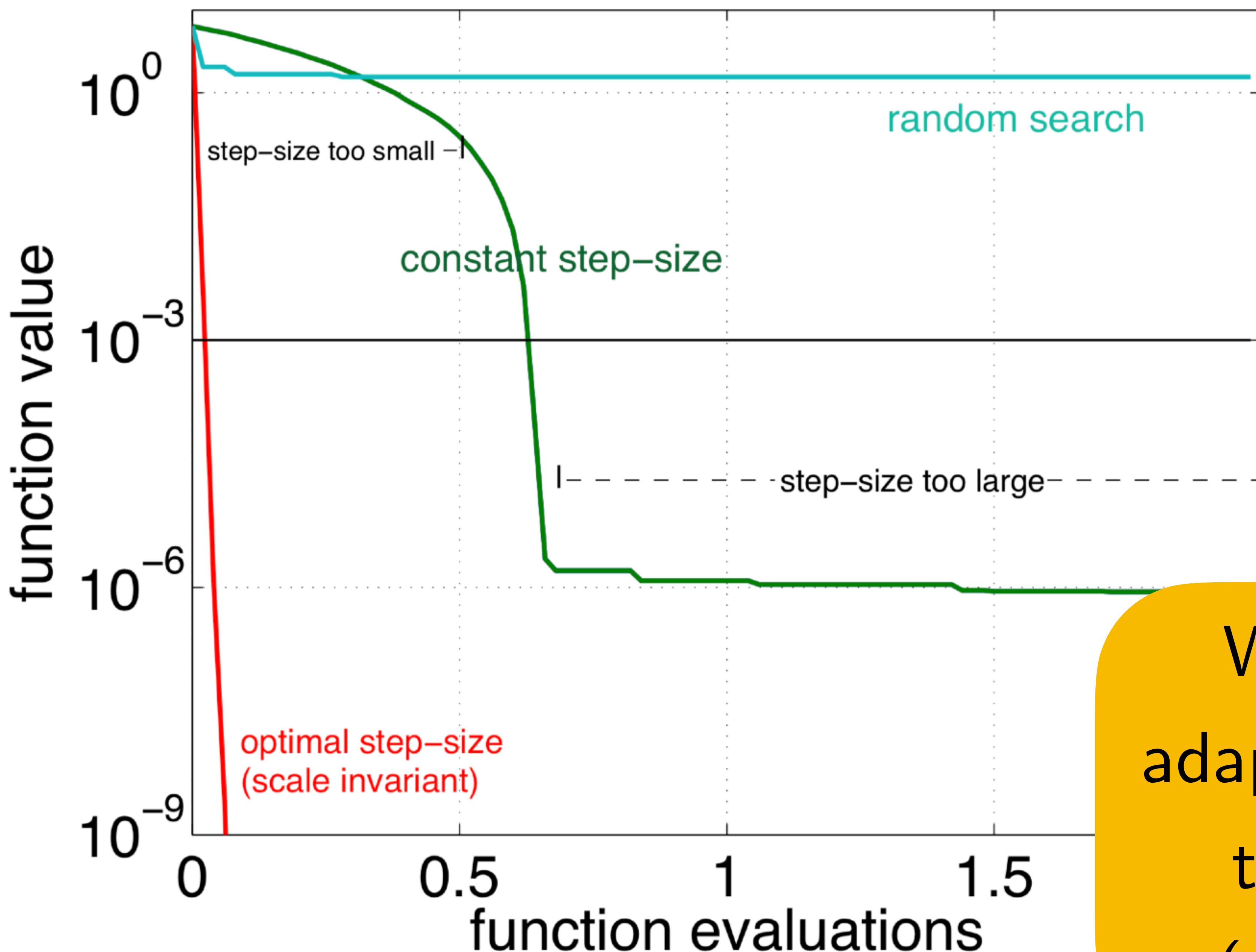
(1+1)-ES
(red & green)

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

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)

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

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

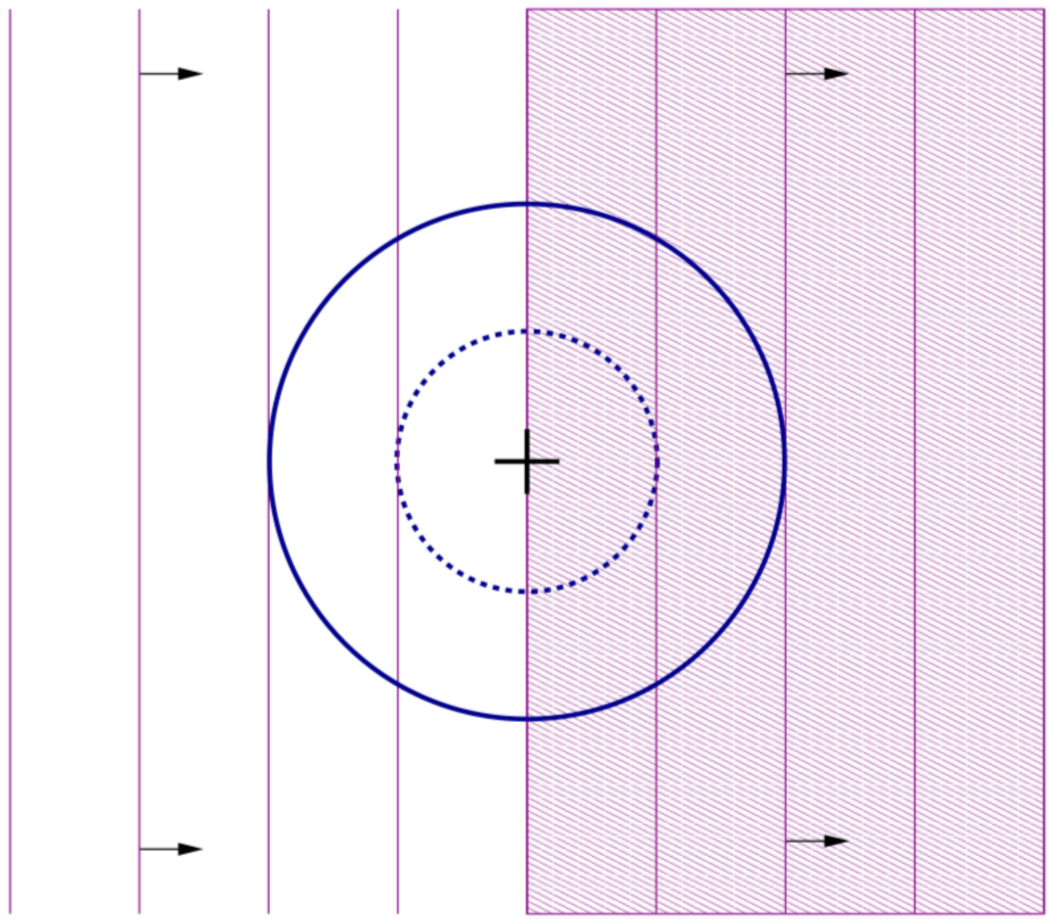
[Ostermeier et al. 84][Hansen, Ostermeier, 2001]

two-point adaptation (TPA), applied with "," selection

[Hansen 2008]

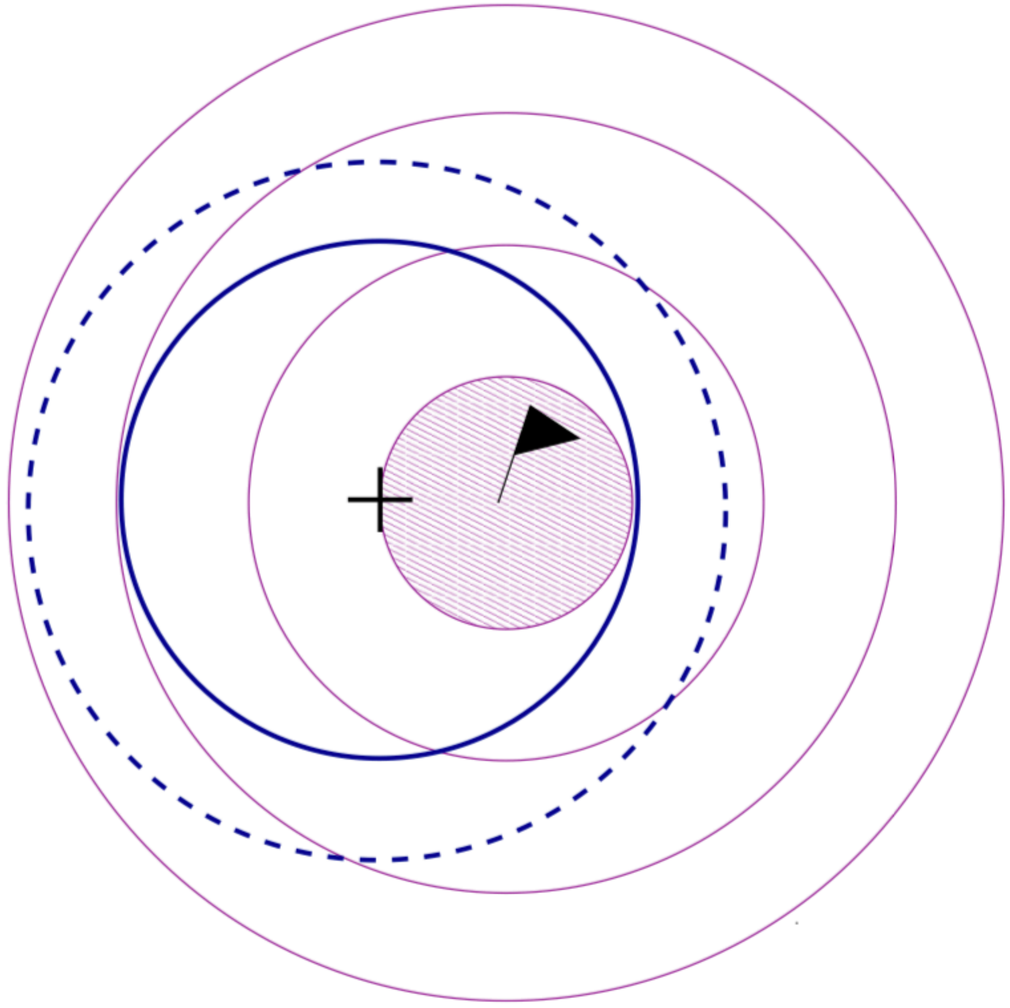
test two solutions in the direction of the mean shift, increase or decrease accordingly the step-size

Step-size control: 1/5th Success Rule



$\downarrow$

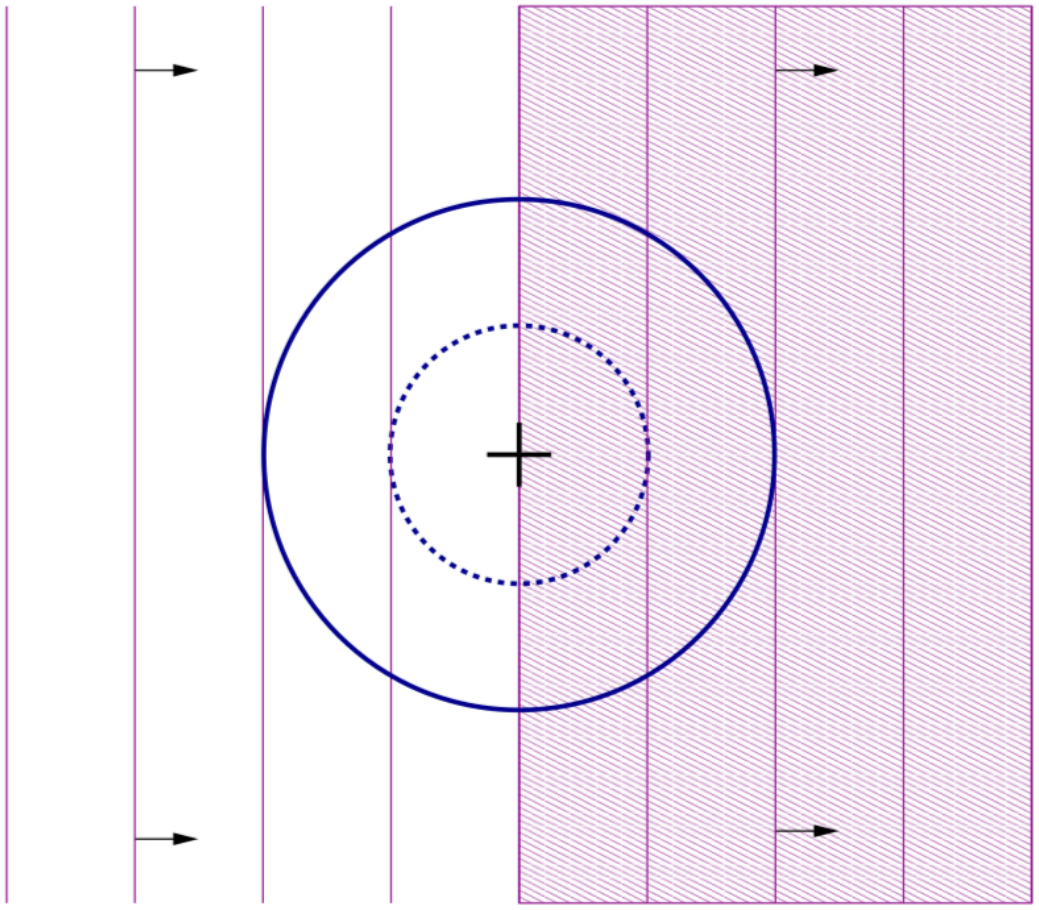
increase σ

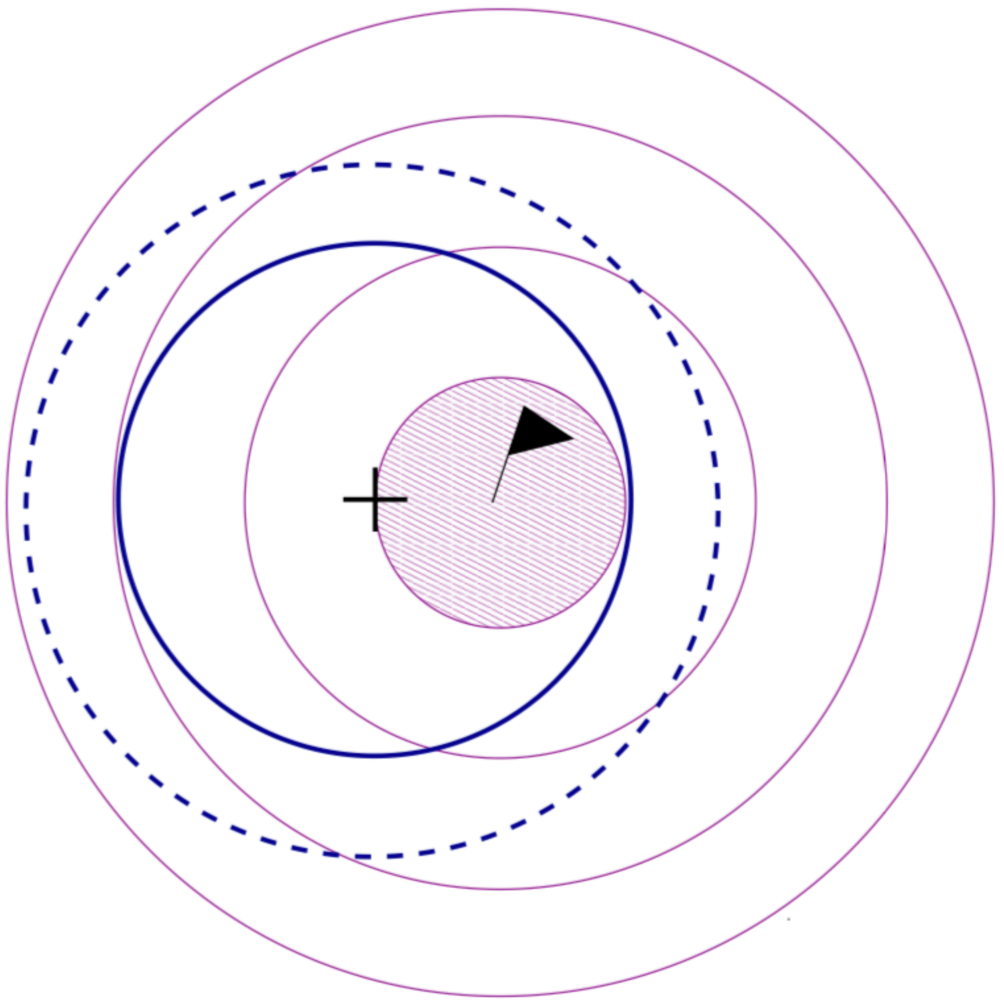


$\downarrow$

decrease σ

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 $p_s < p_{\text{target}}$

$(1+1)$-ES

$$p_{target} = 1/5$$

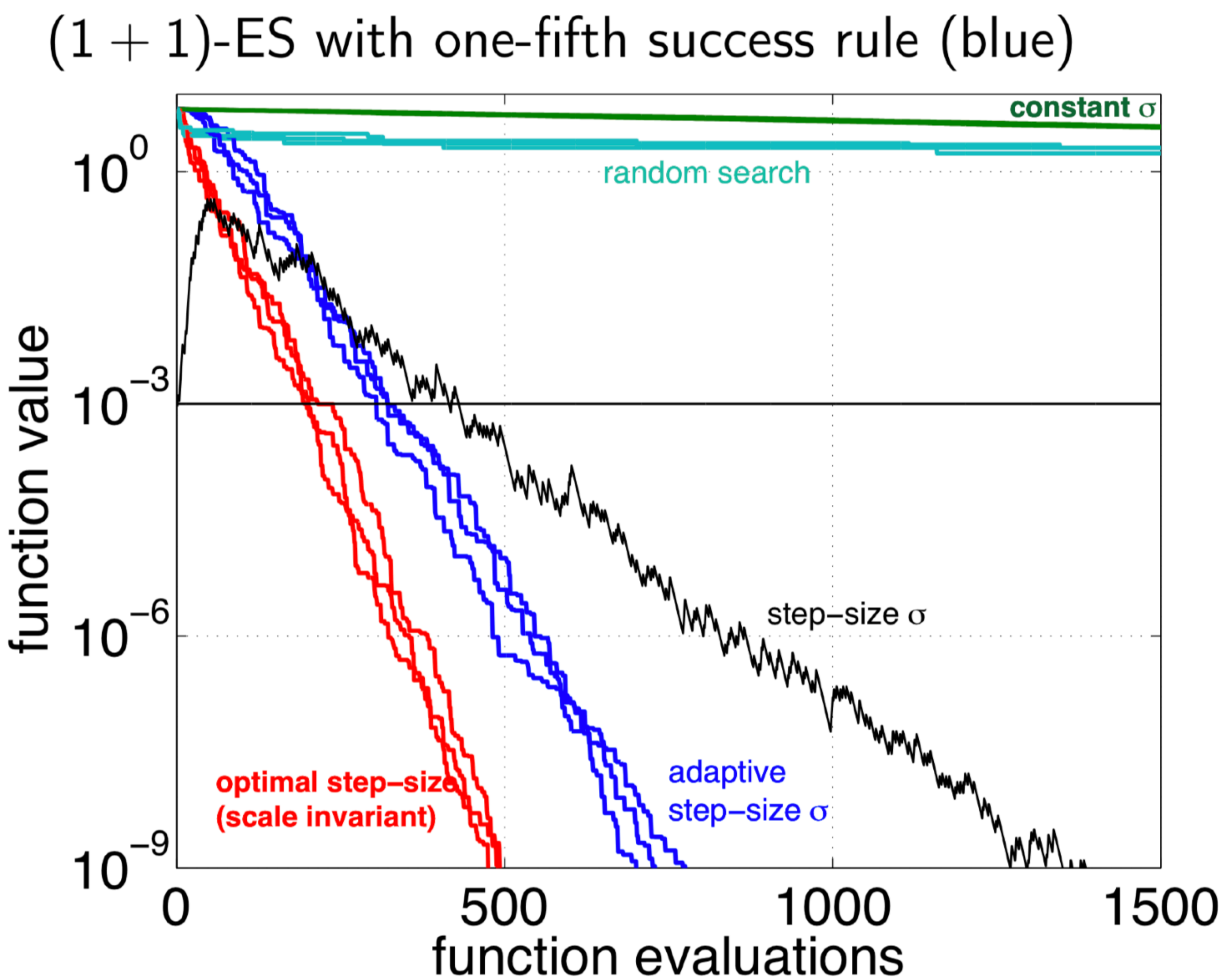
IF *offspring better parent* $[f(\mathbf{x}) \leq f(\mathbf{m})]$

$$p_s = 1,\ \sigma \leftarrow \sigma \times \exp(1/3)$$

ELSE

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

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



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

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

Linear convergence

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

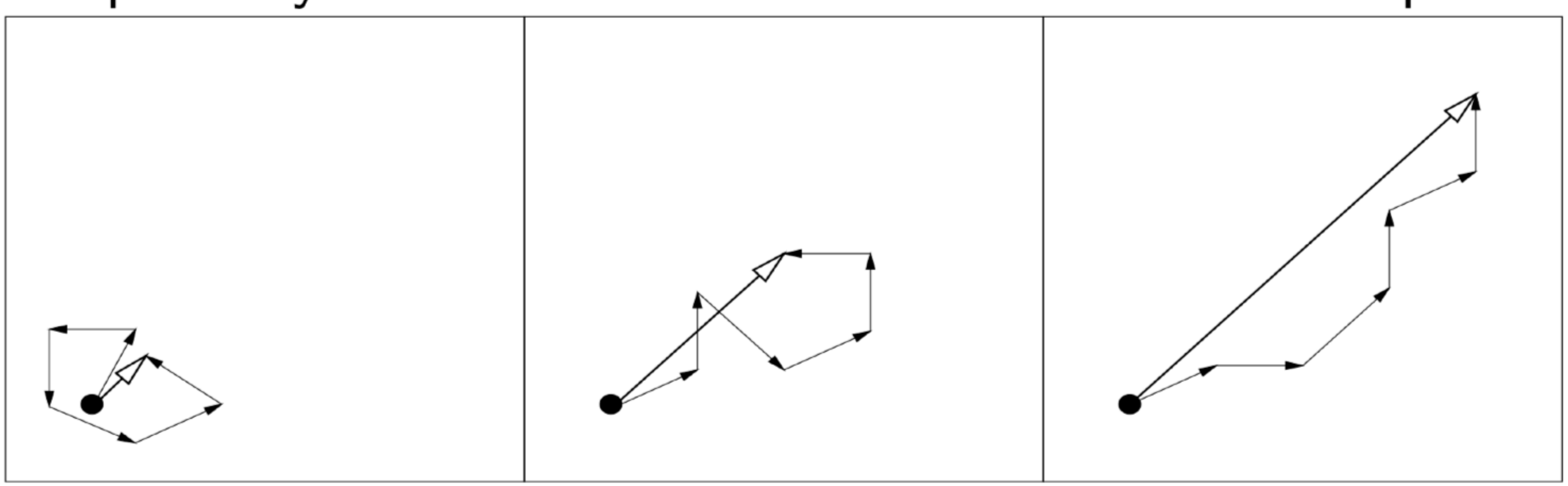
step-size adaptation used in the $(\mu/\mu_w, \lambda)$-ES algorithm framework (in CMA-ES in particular)

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



$\Downarrow$ decrease σ

$\Downarrow$ 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 $c_\sigma \approx 4/n$, $d_\sigma \approx 1$.

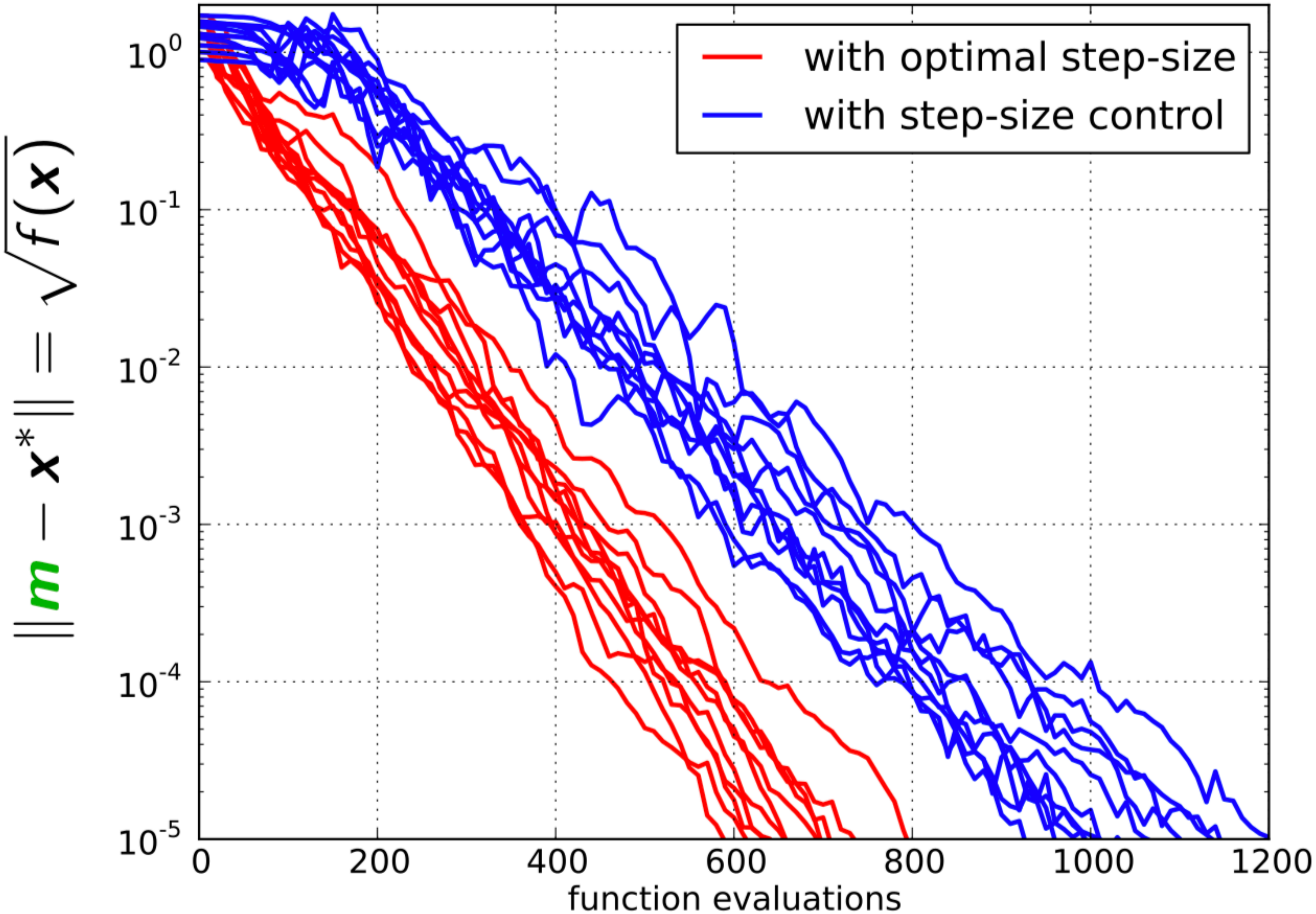
$$\boldsymbol{m} \leftarrow \boldsymbol{m} + \sigma \boldsymbol{y}_w \quad \text{where } \boldsymbol{y}_w = \sum_{i=1}^{\mu} w_i \boldsymbol{y}_{i:\lambda} \qquad \text{update mean}$$

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

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

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

2x11 runs



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

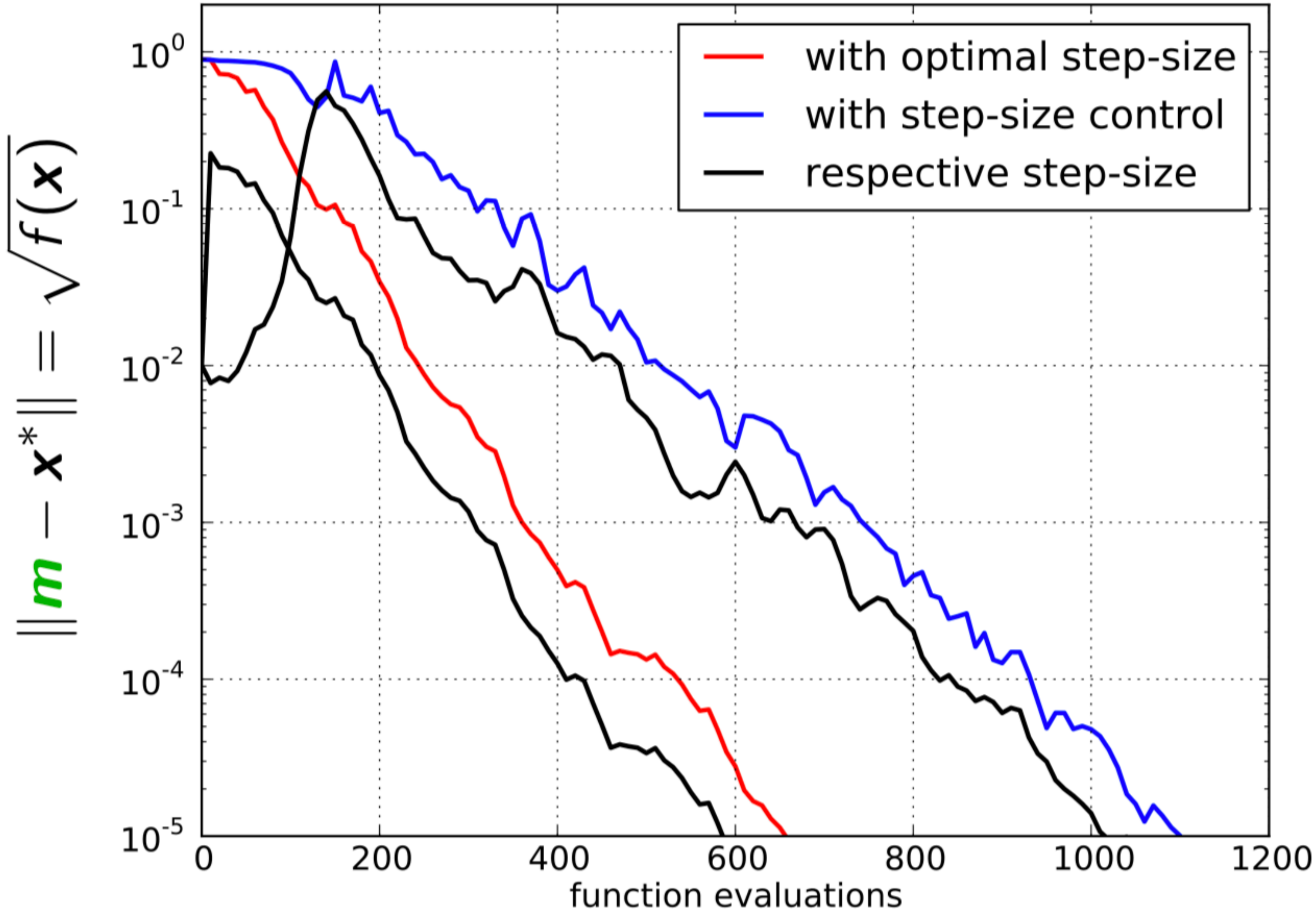
for $n = 10$

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 $n = 10$

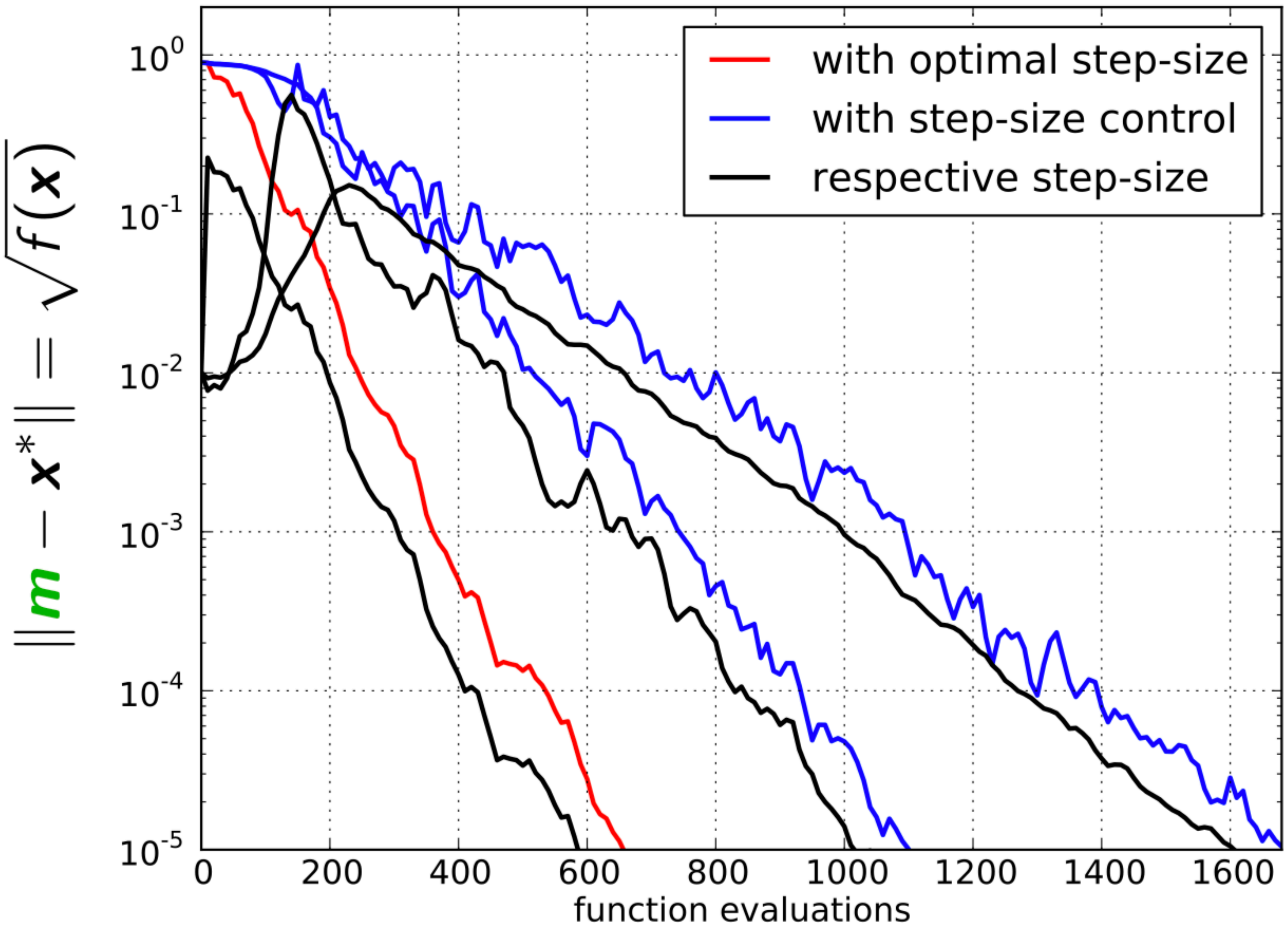
and

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

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:
$\frac{1700}{1100} \approx 1.5$

Optimal Step-size - Lower-bound for Convergence Rates

In the previous slides we have displayed some runs with "optimal" step-size.

Optimal step-size relates to step-size proportional to the distance to the optimum: $\sigma_t = \sigma \|x - x^\star\|$ where $x^\star$ is the optimum of the optimized function (with σ properly chosen).

The associated algorithm is not a real algorithm (as it needs to know the distance to the optimum) but it gives bounds on convergence rates and allows to compute many important quantities.

The goal for a step-size adaptive algorithm is to achieve convergence rates close to the one with optimal step-size

We will formalize this in the context of the (1+1)-ES. Similar results can be obtained for other algorithm frameworks.

Optimal Step-size - Bound on Convergence Rate - (1+1)-ES

Consider a (1+1)-ES algorithm with any step-size adaptation mechanism:

$$X_{t+1} = \begin{cases} X_t + \sigma_t \mathcal{N}_{t+1} & \text{if } f(X_t + \sigma_t \mathcal{N}_{t+1}) \leq f(X_t) \\ X_t & \text{otherwise} \end{cases}$$

with $\{\mathcal{N}_t, t \geq 1\}$ i.i.d. $\sim \mathcal{N}(0, I_d)$

equivalent writing:

$$X_{t+1} = X_t + \sigma_t \mathcal{N}_{t+1} 1_{\{f(X_t + \sigma_t \mathcal{N}_{t+1}) \leq f(X_t)\}}$$

Bound on Convergence Rate - (1+1)-ES

Theorem: For any objective function $f : \mathbb{R}^n \to \mathbb{R}$, for any $y^\star \in \mathbb{R}^n$

$$E[\ \ln \|X_{t+1} - y^\star\|] \geq E[\ln \|X_t - y^\star\|] - \tau$$

lower bound

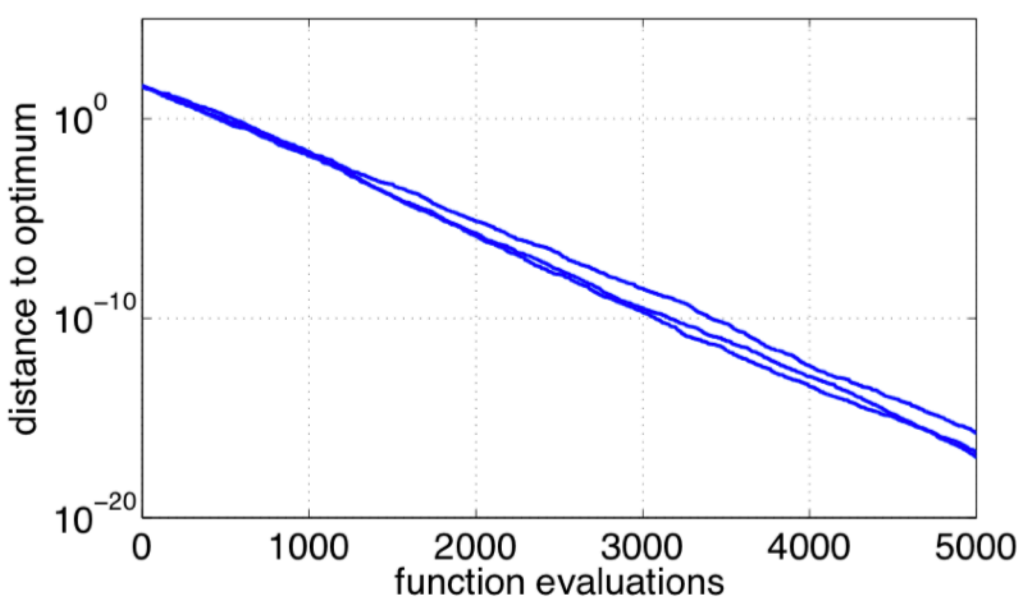
where $\tau = \max_{\sigma \in \mathbb{R}>} \underbrace{E[\ln^- \|e_1 + \sigma \mathcal{N}\|]}_{=:\varphi(\sigma)}$ with $e_1 = (1,0,\ldots,0)$

Theorem: The convergence rate lower-bound is reached on spherical functions $f(x) = g(\|x - x^\star\|)$ (with $g : \mathbb{R}_{\geq 0} \to \mathbb{R}$ strictly increasing) and step-size proportional to the distance to the optimum $\sigma_t = \sigma_{\text{opt}}\|x - x^\star\|$ with σ_{opt} such that $\varphi(\sigma_{\text{opt}}) = \tau$.

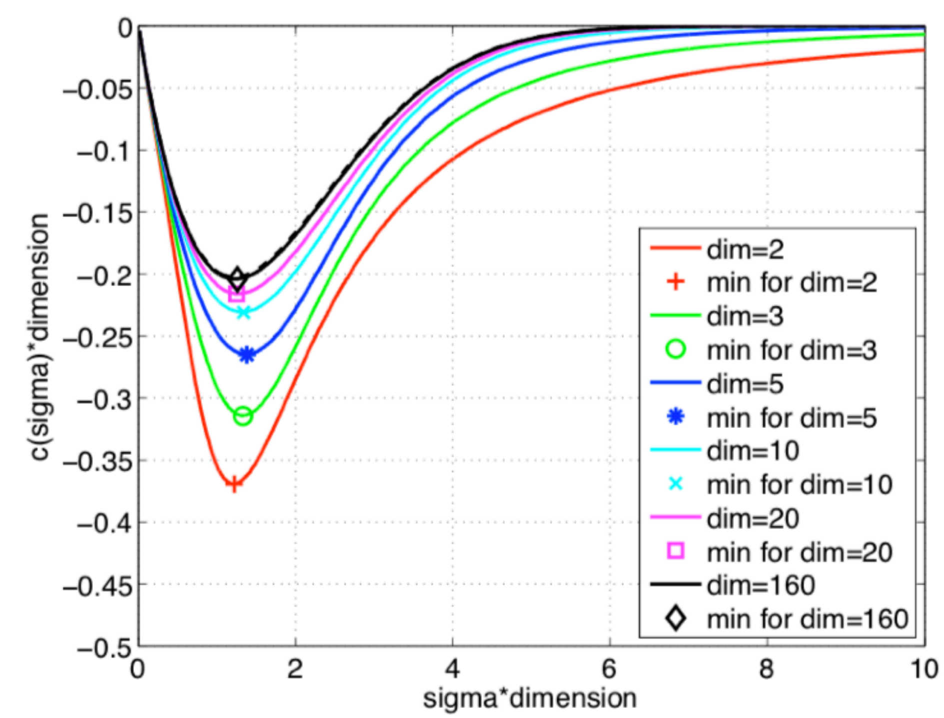
Log-Linear Convergence of scale-invariance step-size ES

Theorem: The (1+1)-ES with step-size proportional to the distance to the optimum $\sigma_t = \sigma \|x\|$ converges (log)-linearly on the sphere function $f(x) = g(\|x\|)$ almost surely:

$$\frac{1}{t} \ln \frac{\|X_t\|}{\|X_0\|} \xrightarrow[t \to \infty]{} -\varphi(\sigma) =: \mathrm{CR}_{(1+1)}(\sigma)$$



$n = 20$ and $\sigma = 0.6/n$



Asymptotic Results ($n \to \infty$)

Theorem

Let $\sigma > 0$, the convergence rate of the (1+1)-ES with scale-invariant step-size on spherical functions satisfies at the limit

$$\lim_{n\to\infty} n \times \mathrm{CR}_{(1+1)}\left(\frac{\sigma}{n}\right) = \frac{-\sigma}{\sqrt{2\pi}} \exp\left(-\frac{\sigma^2}{8}\right) + \frac{\sigma^2}{2}\Phi\left(-\frac{\sigma}{2}\right)$$

where Φ is the cumulative distribution of a normal distribution.

optimal convergence rate decreases to zero like $\frac{1}{n}$