Convergence proof of CMA-ES
Analysis of underlying Markov chains

Dagstuhl seminar
Theory of Randomized Optimization Heuristics

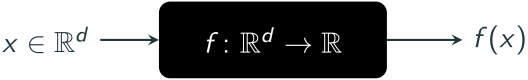
Armand Gissler

RandOpt team, Inria & École polytechnique
Advisors: Anne Auger & Nikolaus Hansen

Tuesday 2nd July, 2024

$$\text{Find } x^* \in \operatorname*{Arg\,min}_{x \in \mathbb{R}^d} f(x) \tag{P}$$

$$\text{Find } x^* \in \operatorname*{Arg\,min}_{x \in \mathbb{R}^d} f(x) \tag{P}$$



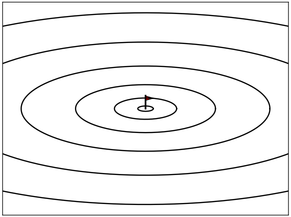
$$\text{Find } x^* \in \operatorname*{Arg\,min}_{x \in \mathbb{R}^d} f(x) \tag{P}$$

$$x \in \mathbb{R}^d \longrightarrow \boxed{f : \mathbb{R}^d \to \mathbb{R}} \longrightarrow f(x)$$

$$\xcancel{\nabla f(x)} \qquad \xcancel{\partial f(x)}$$

Algorithm 1 CMA-ES [HO01], [HMK03]

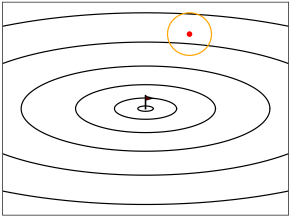
Goal: $\min_{x \in \mathbb{R}^d} f(x)$



Algorithm 1 CMA-ES [HO01], [HMK03]

Goal: $\min_{x \in \mathbb{R}^d} f(x)$

Repeat $(m_t \in \mathbb{R}^d,\ \sigma_t > 0,\ C_t \succ 0)$

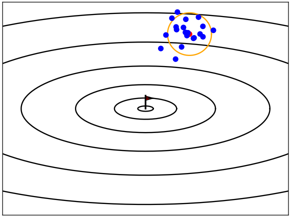


Algorithm 1 CMA-ES [HO01], [HMK03]

Goal: $\min_{x \in \mathbb{R}^d} f(x)$

Repeat $(m_t \in \mathbb{R}^d,\ \sigma_t > 0,\ C_t \succ 0)$

1. $x_{t+1}^1, \ldots, x_{t+1}^\lambda \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$



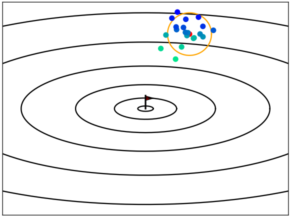
$\lambda =$ population size

Algorithm 1 CMA-ES [HO01], [HMK03]

Goal: $\min_{x \in \mathbb{R}^d} f(x)$

Repeat $(m_t \in \mathbb{R}^d,\ \sigma_t > 0,\ C_t \succ 0)$

1. $x_{t+1}^1, \ldots, x_{t+1}^\lambda \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$
2. sort $f(x_{t+1}^i)$:



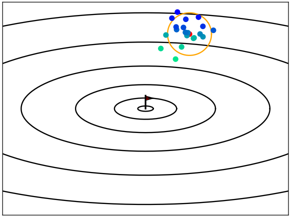
$\lambda =$ population size

Algorithm 1 CMA-ES [HO01], [HMK03]

Goal: $\min_{x \in \mathbb{R}^d} f(x)$

Repeat $(m_t \in \mathbb{R}^d,\ \sigma_t > 0,\ C_t \succ 0)$

1. $x_{t+1}^1, \ldots, x_{t+1}^\lambda \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$
2. sort $f(x_{t+1}^i)$:
   $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$



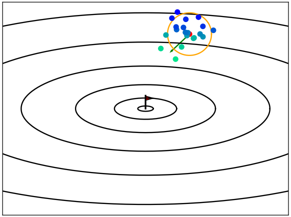
$\lambda =$ population size

Algorithm 1 CMA-ES [HO01], [HMK03]

Goal: $\min_{x \in \mathbb{R}^d} f(x)$

Repeat $(m_t \in \mathbb{R}^d,\ \sigma_t > 0,\ C_t \succ 0)$

1. $x_{t+1}^1, \ldots, x_{t+1}^\lambda \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$
2. sort $f(x_{t+1}^i)$:
   $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$
3. $m_{t+1} = \text{Average}(x_{t+1}^{1:\lambda}, \ldots, x_{t+1}^{\mu:\lambda})$



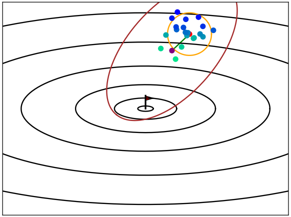
$\lambda =$ population size

$\mu =$ parent number

Algorithm 1 CMA-ES [HO01], [HMK03]

Goal: $\min_{x \in \mathbb{R}^d} f(x)$

Repeat $(m_t \in \mathbb{R}^d,\ \sigma_t > 0,\ C_t \succ 0)$



1. $x_{t+1}^1, \dots, x_{t+1}^\lambda \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$
2. sort $f(x_{t+1}^i)$:
   $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$
3. $m_{t+1} = \text{Average}(x_{t+1}^{1:\lambda}, \dots, x_{t+1}^{\mu:\lambda})$
4. $\sigma_{t+1} = \sigma_t \times$ increasing function $(\|\text{path}\|)$

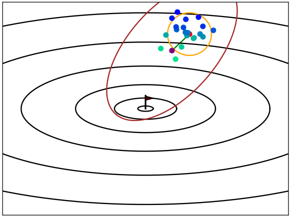
$\lambda =$ population size

$\mu =$ parent number

Algorithm 1 CMA-ES [HO01], [HMK03]

Goal: $\min_{x \in \mathbb{R}^d} f(x)$

Repeat $(m_t \in \mathbb{R}^d,\ \sigma_t > 0,\ C_t \succ 0)$



1. $x_{t+1}^1, \ldots, x_{t+1}^\lambda \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$
2. sort $f(x_{t+1}^i)$:
   $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$
3. $m_{t+1} = \text{Average}(x_{t+1}^{1:\lambda}, \ldots, x_{t+1}^{\mu:\lambda})$
4. $\sigma_{t+1} = \sigma_t \times \text{increasing function}\,(\|\mathrm{path}\|)$
5. $C_{t+1} = \text{Positive combination}\left(C_t, \overleftrightarrow{\mathrm{path}}, \text{Average}\left[\overleftrightarrow{(x_{t+1}^{i:\lambda} - m_t)}\right]\right)$

$\lambda =$ population size

$\mu =$ parent number

Mean update:

$$m_{t+1} = \text{Average}(x_{t+1}^{1:\lambda}, \ldots, x_{t+1}^{\mu:\lambda})$$

Mean update:

$$m_{t+1} = \text{Average}(x_{t+1}^{1:\lambda}, \ldots, x_{t+1}^{\mu:\lambda}) = \sum_{i=1}^{\mu} \underbrace{\text{weight}_i}_{w_i} x_{t+1}^{i:\lambda}$$

Step-size adaptation:

$$\sigma_{t+1} = \sigma_t \times \text{increasing function}\left(\|\text{path}\|\right)$$

Step-size adaptation:

$$\sigma_{t+1} = \sigma_t \times \text{increasing function}\left(\|\mathrm{path}\|\right)$$

where

$$\mathrm{path} = p^{\sigma}_{t+1} = \underbrace{(1 - c_\sigma)}_{\text{decay rate}} p^{\sigma}_t$$

Step-size adaptation:

$$\sigma_{t+1} = \sigma_t \times \text{increasing function}\,(\|\mathrm{path}\|)$$

where

$$\mathrm{path} = p^{\sigma}_{t+1} = \underbrace{(1 - c_\sigma)}_{\text{decay rate}} p^{\sigma}_t + \sqrt{c_\sigma(2 - c_\sigma)}\frac{m_{t+1} - m_t}{\sigma_t C_t^{1/2}\|\text{weights}\|}$$

Step-size adaptation:

$$\sigma_{t+1} = \sigma_t \times \text{increasing function}\left(\|\mathrm{path}\|\right)$$
$$= \sigma_t \times \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|p^\sigma_{t+1}\|}{\mathbb{E}\|\mathcal{N}\|} - 1\right)\right)$$

where

$$\mathrm{path} = p^\sigma_{t+1} = \underbrace{(1 - c_\sigma)}_{\text{decay rate}} p^\sigma_t + \sqrt{c_\sigma(2 - c_\sigma)}\frac{m_{t+1} - m_t}{\sigma_t C_t^{1/2}\|\text{weights}\|}$$

Covariance matrix adaptation:

$$C_{t+1} = \text{Positive combination}\left(C_t, \overleftrightarrow{\text{path}}, \text{Average}\left[\overleftrightarrow{(x_{t+1}^{i:\lambda} - m_t)} \right] \right)$$

Covariance matrix adaptation:

$$C_{t+1} = \text{Positive combination}\left(C_t, \overleftrightarrow{\mathrm{path}}, \text{Average}\left[\overleftrightarrow{(x^{i:\lambda}_{t+1} - m_t)}\right]\right)$$

where

$$\mathrm{path} = p^c_{t+1} = \underbrace{(1 - c_c)}_{\text{decay rate}} p^c_t + \sqrt{c_c(2 - c_c)} \frac{m_{t+1} - m_t}{\sigma_t \|\text{weights}\|}$$

Covariance matrix adaptation:

$$C_{t+1} = \text{Positive combination}\left(C_t, \overleftrightarrow{\text{path}}, \text{Average}\left[\overleftrightarrow{(x_{t+1}^{i:\lambda} - m_t)} \right] \right)$$
$$= (1 - c_1 - c_\mu) C_t$$

where

$$\text{path} = p_{t+1}^c = \underbrace{(1 - c_c)}_{\text{decay rate}} p_t^c + \sqrt{c_c(2 - c_c)} \frac{m_{t+1} - m_t}{\sigma_t \|\text{weights}\|}$$

Covariance matrix adaptation:

$$C_{t+1} = \text{Positive combination}\left(C_t, \overleftrightarrow{\text{path}}, \text{Average}\left[\overleftrightarrow{(x_{t+1}^{i:\lambda} - m_t)} \right] \right)$$

$$= (1 - c_1 - c_\mu) C_t + c_1 \underbrace{[p_{t+1}^c][p_{t+1}^c]^\top}_{\text{rank-one update}}$$

where

$$\text{path} = p_{t+1}^c = \underbrace{(1 - c_c)}_{\text{decay rate}} p_t^c + \sqrt{c_c(2 - c_c)} \frac{m_{t+1} - m_t}{\sigma_t \|\text{weights}\|}$$

Covariance matrix adaptation:

$$C_{t+1} = \text{Positive combination}\left(C_t, \overleftrightarrow{\text{path}}, \text{Average}\left[\overleftrightarrow{(x^{i:\lambda}_{t+1} - m_t)}\right]\right)$$

$$= (1 - c_1 - c_\mu) C_t + c_1 \underbrace{[p^c_{t+1}][p^c_{t+1}]^\top}_{\text{rank-one update}}$$

$$+ \frac{c_\mu}{\sigma_t^2} \underbrace{\sum_{i=1}^{\mu} w_i (x^{i:\lambda}_{t+1} - m_t)(x^{i:\lambda}_{t+1} - m_t)^\top}_{\text{rank-mu update}}$$

where

$$\text{path} = p^c_{t+1} = \underbrace{(1 - c_c)}_{\text{decay rate}} p^c_t + \sqrt{c_c(2 - c_c)} \frac{m_{t+1} - m_t}{\sigma_t \|\text{weights}\|}$$

Goal:

Prove linear convergence of CMA-ES

Goal:

Prove linear convergence of CMA-ES:

$$\mathrm{distance}(m_t, x^*) \underset{t\to\infty}{\sim} \rho^t \times \mathrm{distance}(m_0, x^*) \quad (\rho < 1)$$

Goal:

Prove linear convergence of CMA-ES:

$$\mathrm{distance}(m_t, x^*) \underset{t\to\infty}{\sim} \rho^t \times \mathrm{distance}(m_0, x^*) \quad (\rho < 1)$$

and learning of the inverse Hessian on convex-quadratic functions $f(x) = x^\top \mathbf{H} x/2$

Goal:

Prove linear convergence of CMA-ES:

$$\mathrm{distance}(m_t, x^*) \underset{t\to\infty}{\sim} \rho^t \times \mathrm{distance}(m_0, x^*) \quad (\rho < 1)$$

and learning of the inverse Hessian on convex-quadratic functions $f(x) = x^\top \mathbf{H} x/2$:

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{C_t}{\text{normalization}}\right] \propto \mathbf{H}^{-1}$$

Without covariance matrix adaptation

We consider

$$z_t = \frac{m_t - x^*}{\sigma_t}$$

Without covariance matrix adaptation

We consider

$$z_t = \frac{m_t - x^*}{\sigma_t}$$

and we prove that $\{z_t\}_{t\in\mathbb{N}}$ is a stationary Markov chain

Without covariance matrix adaptation

We consider

$$z_t = \frac{m_t - x^*}{\sigma_t}$$

and we prove that $\{z_t\}_{t \in \mathbb{N}}$ is a stationary Markov chain
(m_t converges to x^* as fast as σ_t to 0)

Without covariance matrix adaptation

We consider

$$z_t = \frac{m_t - x^*}{\sigma_t}$$

and we prove that $\{z_t\}_{t \in \mathbb{N}}$ is a stationary Markov chain
(m_t converges to x^* as fast as σ_t to 0)

Consequence:

Without covariance matrix adaptation

We consider

$$z_t = \frac{m_t - x^*}{\sigma_t}$$

and we prove that $\{z_t\}_{t\in\mathbb{N}}$ is a stationary Markov chain
(m_t converges to x^* as fast as σ_t to 0)

Consequence:

$$\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \frac{1}{T}\sum_{t=0}^{T-1}\log\|z_{t+1}\| - \log\|z_t\| - \log\frac{\sigma_{t+1}}{\sigma_t}$$

Without covariance matrix adaptation

We consider

$$z_t = \frac{m_t - x^*}{\sigma_t}$$

and we prove that $\{z_t\}_{t \in \mathbb{N}}$ is a stationary Markov chain (m_t converges to x^* as fast as σ_t to 0)

Consequence:

$$\frac{1}{T} \log \frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \frac{1}{T} \sum_{t=0}^{T-1} \log \|z_{t+1}\| - \log \|z_t\| - \log \frac{\sigma_{t+1}}{\sigma_t}$$

$$\lim_{T \to \infty} \frac{1}{T} \log \frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \mathbb{E}_\pi[\log \|z\|] - \mathbb{E}_\pi[\log \|z\|] - \mathbb{E}_\pi \left[\log \frac{\sigma_1}{\sigma_0}\right]$$

Without covariance matrix adaptation

We consider

$$z_t = \frac{m_t - x^*}{\sigma_t}$$

and we prove that $\{z_t\}_{t\in\mathbb{N}}$ is a stationary Markov chain
(m_t converges to x^* as fast as σ_t to 0)

Consequence:

$$\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \frac{1}{T}\sum_{t=0}^{T-1}\log\|z_{t+1}\| - \log\|z_t\| - \log\frac{\sigma_{t+1}}{\sigma_t}$$

$$\lim_{T\to\infty}\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \mathbb{E}_\pi[\log\|z\|] - \mathbb{E}_\pi[\log\|z\|] - \mathbb{E}_\pi\left[\log\frac{\sigma_1}{\sigma_0}\right]$$

$$\|m_T - x^*\| \underset{t\to\infty}{\sim} e^{-T\mathbb{E}_\pi\left[\log\frac{\sigma_1}{\sigma_0}\right]}\|m_0 - x^*\|$$

Without covariance matrix adaptation

We consider

$$z_t = \frac{m_t - x^*}{\sigma_t}$$

and we prove that $\{z_t\}_{t\in\mathbb{N}}$ is a stationary Markov chain
(m_t converges to x^* as fast as σ_t to 0)

Consequence:

$$\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \frac{1}{T}\sum_{t=0}^{T-1}\log\|z_{t+1}\| - \log\|z_t\| - \log\frac{\sigma_{t+1}}{\sigma_t}$$

$$\lim_{T\to\infty}\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \mathbb{E}_\pi[\log\|z\|] - \mathbb{E}_\pi[\log\|z\|] - \mathbb{E}_\pi\left[\log\frac{\sigma_1}{\sigma_0}\right]$$

$$\|m_T - x^*\| \underset{t\to\infty}{\sim} e^{-T\mathbb{E}_\pi\left[\log\frac{\sigma_1}{\sigma_0}\right]}\|m_0 - x^*\|$$

(where π is the limit distribution of $\{z_t\}$)

$$\log \frac{\sigma_1}{\sigma_0} \propto \frac{\| \sum w_i z_1^{i:\lambda} \|}{\|\text{weights}\| \mathbb{E} \|\mathcal{N}\|} - 1$$

$$\log \frac{\sigma_1}{\sigma_0} \propto \frac{\| \sum w_i z_1^{i:\lambda} \|}{\|\text{weights}\| \mathbb{E} \|\mathcal{N}\|} - 1$$

We are able to prove

$$\mathbb{E}_\pi \left[\frac{\| \sum w_i z^{i:\lambda} \|^2}{\|\text{weights}\|^2 \mathbb{E} \|\mathcal{N}\|^2} - 1 \right] > 0$$

How can we prove that $\{z_t\}_{t\in\mathbb{N}}$ is a stationary Markov chain?

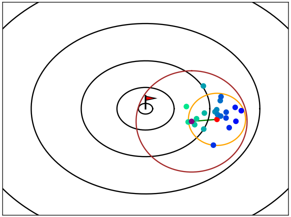
How can we prove that $\{z_t\}_{t\in\mathbb{N}}$ is a stationary Markov chain?

(and under which conditions?)

Algorithm 2 ES with step-size adaptation

Goal: $\min_{x \in \mathbb{R}^d} f(x)$

Repeat $(m_t \in \mathbb{R}^d,\ \sigma_t > 0)$



1. $x_{t+1}^1, \ldots, x_{t+1}^\lambda \sim \mathcal{N}(m_t, \sigma_t^2 I_d)$
2. $\mathrm{sort}\ f(x_{t+1}^i)$:
   $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$
3. $m_{t+1} = \text{Average}(x_{t+1}^{1:\lambda}, \ldots, x_{t+1}^{\mu:\lambda})$
4. $\sigma_{t+1} = \sigma_t \times \text{increasing function}\ (\|\mathrm{path}\|)$

$\lambda =$ population size

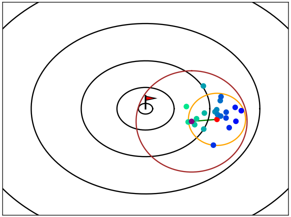
$\mu =$ parent number

Algorithm 2 ES with step-size adaptation

Goal: $\min_{x \in \mathbb{R}^d} f(x)$

Repeat $(z_t = m_t / \sigma_t)$

1. $x_{t+1}^{1}, \ldots, x_{t+1}^{\lambda} \sim \mathcal{N}(m_t, \sigma_t^2 I_d)$
2. sort $f(x_{t+1}^{i})$:
   $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$
3. $m_{t+1} = \text{Average}(x_{t+1}^{1:\lambda}, \ldots, x_{t+1}^{\mu:\lambda})$
4. $\sigma_{t+1} = \sigma_t \times$ increasing function $(\|\mathrm{path}\|)$



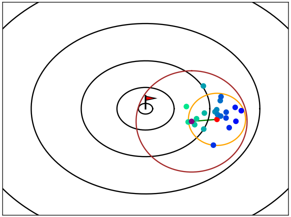
$\lambda =$ population size

$\mu =$ parent number

Algorithm 2 ES with step-size adaptation

Goal: $\min_{x \in \mathbb{R}^d} f(x)$

Repeat $(z_t = m_t / \sigma_t)$



1. $z_{t+1}^1, \ldots, z_{t+1}^\lambda \sim \mathcal{N}(z_t, I_d)$
2. sort $f(x_{t+1}^i)$:
   $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$
3. $m_{t+1} = \text{Average}(x_{t+1}^{1:\lambda}, \ldots, x_{t+1}^{\mu:\lambda})$
4. $\sigma_{t+1} = \sigma_t \times \text{increasing function}\,(\|\text{path}\|)$

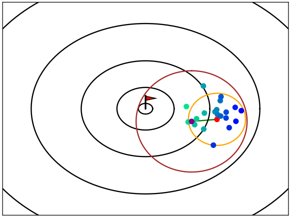
$\lambda =$ population size

$\mu =$ parent number

Algorithm 2 ES with step-size adaptation

Goal: $\min_{x \in \mathbb{R}^d} f(x)$

Repeat $(z_t = m_t / \sigma_t)$



1. $z_{t+1}^1, \ldots, z_{t+1}^\lambda \sim \mathcal{N}(z_t, I_d)$
2. sort $f(x_{t+1}^i)$:
   $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$
3. $z_{t+1} = \text{Average}(z_{t+1}^{1:\lambda}, \ldots, z_{t+1}^{\mu:\lambda})$
4. $\sigma_{t+1} = \sigma_t \times$ increasing function $(\|\text{path}\|)$

$\lambda =$ population size

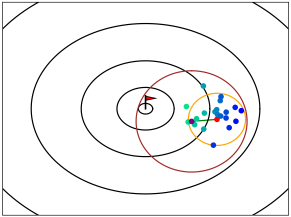
$\mu =$ parent number

Algorithm 2 ES with step-size adaptation

Goal: $\min_{x \in \mathbb{R}^d} f(x)$

Repeat $(z_t = m_t / \sigma_t)$

1. $z_{t+1}^1, \ldots, z_{t+1}^\lambda \sim \mathcal{N}(z_t, I_d)$
2. sort $f(x_{t+1}^i)$:
   $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$
3. $z_{t+1} = \frac{\text{Average}(z_{t+1}^{1:\lambda}, \ldots, z_{t+1}^{\mu:\lambda})}{\text{increasing function}(\|\text{path}\|)}$



$\lambda =$ population size

$\mu =$ parent number

Algorithm 2 ES with step-size adaptation

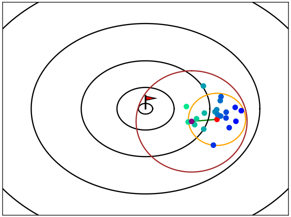
Goal: $\min_{x \in \mathbb{R}^d} f(x)$

Repeat $(z_t = m_t / \sigma_t)$

1. $z^1_{t+1}, \ldots, z^\lambda_{t+1} \sim \mathcal{N}(z_t, I_d)$
2. sort $f(x^i_{t+1})$:

$$\boxed{f\left(x^{1:\lambda}_{t+1}\right) \leqslant \cdots \leqslant f\left(x^{\lambda:\lambda}_{t+1}\right)}$$

3. $z_{t+1} = \dfrac{\text{Average}(z^{1:\lambda}_{t+1}, \ldots, z^{\mu:\lambda}_{t+1})}{\text{increasing function}(\|\text{path}\|)}$

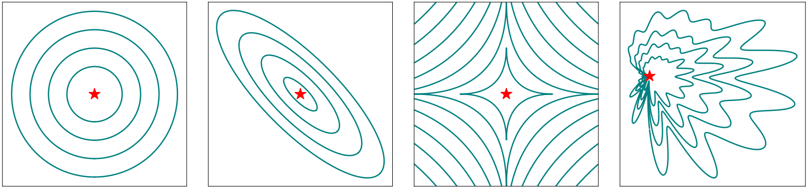


$\lambda =$ population size

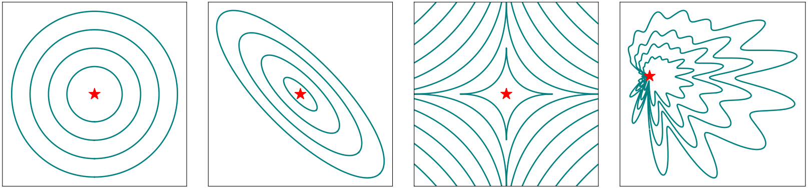
$\mu =$ parent number

$$f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right) \overset{?}{\Leftrightarrow} g\left(z_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant g\left(z_{t+1}^{\lambda:\lambda}\right)$$

Scaling-invariant functions [TGAH21]

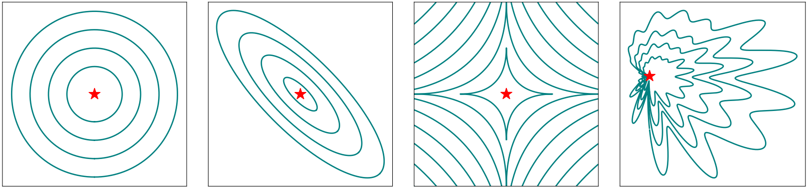


Scaling-invariant functions [TGAH21]



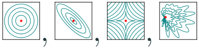
$$f\left(x^i_{t+1}\right) \leqslant f\left(x^j_{t+1}\right) \Leftrightarrow f\left(\star + \frac{x^i_{t+1} - \star}{\sigma_t}\right) \leqslant f\left(\star + \frac{x^j_{t+1} - \star}{\sigma_t}\right)$$

Scaling-invariant functions [TGAH21]



$$f\left(x^i_{t+1}\right) \leqslant f\left(x^j_{t+1}\right) \Leftrightarrow f\left(\star + \frac{x^i_{t+1} - \star}{\sigma_t}\right) \leqslant f\left(\star + \frac{x^j_{t+1} - \star}{\sigma_t}\right)$$

Proposition ([AH16])

If $f \in \{$  $\}$ *, then* $\{z_t\}_{t\in\mathbb{N}}$ *is a Markov chain.*

How to prove that $\{z_t\}_{t\in\mathbb{N}}$ is stationary

1. Irreducibility and aperiodicity of $\{z_t\}$
2. Drift condition:

$$\mathbb{E}[V(z_1)] \leqslant (1-\varepsilon)V(z_0) \qquad \forall z_0 \notin \mathsf{K}$$

How to prove that $\{z_t\}_{t\in\mathbb{N}}$ is stationary

1. Irreducibility and aperiodicity of $\{z_t\}$
2. Drift condition:

$$\mathbb{E}[V(z_1)] \leqslant (1-\varepsilon)V(z_0) \qquad \forall z_0 \notin \mathsf{K}$$

Theorem ([MT09])
If 1. and 2. hold for a small set K*, then* $\{z_t\}$ *is stationary (*V*-geometrically ergodic).*

1. Irreducibility and aperiodicity

$\{z_t\}_{t \in \mathbb{N}}$ is irreducible when

$$\forall z_{\text{start}}, z_{\text{end}} \in \mathcal{Z}, \underbrace{\exists k > 0,\ \mathbb{P}[z_k = z_{\text{end}} \mid z_0 = z_{\text{start}}] > 0}_{z_{\text{start}} \rightsquigarrow z_{\text{end}}}$$

1. Irreducibility and aperiodicity

$\{z_t\}_{t\in\mathbb{N}}$ is irreducible when

$$\forall z_{\text{start}} \in \mathcal{Z}, \forall \mathcal{Z}_{\text{end}} \subset \mathcal{Z},\ \text{Volume}(\mathcal{Z}_{\text{end}}) > 0 \Rightarrow z_{\text{start}} \rightsquigarrow \mathcal{Z}_{\text{end}}$$

1. Irreducibility and aperiodicity

Theorem ([MC91], [MT09], [CA19], [GDA24])
The Markov chain

$$z_{t+1} = F(z_t, U_{t+1})$$

is irreducible and aperiodic when

(i) *there exists a steadily attracting state* z^**;*

(ii) *there exists a path* $U_1^*, \ldots, U_k^*$ *at which* $F^k(z^*, \cdot)$ *is submersive.*

Assumptions: F is loc. Lipschitz and $U_{t+1} \sim p_{z_t}$ where $(z, u) \mapsto p_z(u)$ is l.s.c.

1. Irreducibility and aperiodicity

Theorem ([MC91], [MT09], [CA19], [GDA24])
The Markov chain

$$z_{t+1} = F(z_t, U_{t+1})$$

is irreducible and aperiodic when

(i) *there exists a steadily attracting state* z^**;*

(ii) *there exists a path* $U_1^*, \ldots, U_k^*$ *at which* $F^k(z^*, \cdot)$ *is submersive.*

For us:

$$z_{t+1} = F(z_t, z_{t+1}^{i:\lambda}) = \frac{\text{Average}(z_{t+1}^{1:\lambda}, \ldots, z_{t+1}^{\mu:\lambda})}{\text{normalization}}$$

Assumptions: F is loc. Lipschitz and $U_{t+1} \sim p_{z_t}$ where $(z, u) \mapsto p_z(u)$ is l.s.c.

(i) steadily attracting state

$$z_{t+1} = F(z_t, U_{t+1})$$

z^* is *steadily attracting* when

$$\forall z_0,\ \exists \{U_k\}_{k\in\mathbb{N}}, \quad \lim_{k\to\infty} F^k(z_0, U_1, \ldots, U_k) = z^*$$

(i) steadily attracting state

$$z_{t+1} = F(z_t, U_{t+1})$$

z^* is *steadily attracting* when

$$\forall z_0,\ \exists \{U_k\}_{k\in\mathbb{N}}, \quad \lim_{k\to\infty} F^k(z_0, U_1, \ldots, U_k) = z^*$$

Proposition
0 *is steadily attracting*

Proof.
Choose $z_{t+1}^{i:\lambda} = 0$. Then

$$z_{t+1} = \frac{\text{Average}(0, \ldots, 0)}{\text{normalization}} = 0$$

□

(ii) submersion

$F(\cdot)$ is a submersion at x when $\mathcal{D}F(x)$ is surjective.

(ii) submersion

$F(\cdot)$ is a submersion at x when $\mathcal{D}F(x)$ is surjective.

Proposition
$F(0,\cdot)$ is submersive at 0

Proof.

$$F(0, 0 + h^i) = \frac{\text{Average}(h^1, \ldots, h^\mu)}{\text{normalization}} = \underbrace{\text{Average}(h^1, \ldots, h^\mu)}_{\text{surjective}} + o(h^i)$$

□

Consequence

$\{z_t\}$ is an irreducible aperiodic Markov chain

2. Drift condition

$V(z) = \|z\|^2$

$$\mathbb{E}[\|z_1\|^2] \leqslant (1 - \varepsilon)\|z_0\|^2$$

when $\|z_0\| \gg 1$ and $f \in \Big\{$ , ,  $\Big\}$

Theorem ([TAH23])

If $f \in \Bigg\{$ , ,  $\Bigg\}$, $\{z_t\}$ *is a stationary Markov chain.*

Theorem ([TAH23])
If $f \in \Big\{$ , ,  *$\Big\}$, $\{z_t\}$ is a stationary Markov chain.*

Conclusion:

Theorem ([TAH23])
ES with step-size adaptation converges linearly

Back to CMA-ES

$$z_t = \frac{m_t}{\sigma_t \sqrt{\lambda_{\min}(C_t)}}$$

$$\Sigma_t = \frac{C_t}{\lambda_{\min}(C_t)}$$

Proposition* ([GWAH])

If $f \in \{$ , , ,  $\}$, *then* $\{(z_t, \Sigma_t)\}_{t \in \mathbb{N}}$ *is a Markov chain.*

How to prove that $\{(z_t, \Sigma_t)\}_{t \in \mathbb{N}}$ is stationary

1. Irreducibility and aperiodicity of $\{(z_t, \Sigma_t)\}$
2. Drift condition:

$$\mathbb{E}[V(z_1, \Sigma_1)] \leqslant (1 - \varepsilon) V(z_0, \Sigma_0) \qquad \forall (z_0, \Sigma_0) \notin K$$

1. Irreducibility and aperiodicity

Theorem ([MC91], [MT09], [CA19], [GDA24])
The Markov chain

$$(z_{t+1}, \Sigma_{t+1}) = F(z_t, \Sigma_{t+1}, z_{t+1}^{i:\lambda})$$

is irreducible and aperiodic when

(i) *there exists a steadily attracting state* (z^*, Σ^*)*;*

(ii) *there exists a path* $z_1^{i:\lambda}, \ldots, z_k^{i:\lambda}$ *at which* $F^k(z^*, \Sigma^*, \cdot)$ *is submersive.*

Assumptions: F is loc. Lipschitz and $U_{t+1} \sim p_{z_t}$ where $(z, u) \mapsto p_z(u)$ is l.s.c.

Proposition* ([GWAH])
$(z^*, \Sigma^*) = (0, (1 - c_1 - c_\mu) I_d)$ *is steadily attracting and there exists* $z_1^{i:\lambda}, \ldots, z_k^{i:\lambda}$ *at which* $F^k(z^*, \Sigma^*, \cdot)$ *is submersive.*

Proof.
More complicated than before... □

Proposition* ([GWAH])
$(z^*, \Sigma^*) = (0, (1 - c_1 - c_\mu) I_d)$ *is steadily attracting and there exists* $z_1^{i:\lambda}, \ldots, z_k^{i:\lambda}$ *at which* $F^k(z^*, \Sigma^*, \cdot)$ *is submersive.*

Proof.
More complicated than before... $\square$

Consequence:

$\{(z_t, \Sigma_t)\}$ is irreducible and aperiodic.

2. Drift condition

$$V(z, \Sigma) = \alpha \|z\|^2 + \beta \|\|\Sigma\|\|$$

2. Drift condition

$$V(z, \Sigma) = \alpha \|z\|^2 + \beta |||\Sigma|||$$

(a) When $|||\Sigma_0||| \gg 1 + \|z_0\|^2$:

$$\mathbb{E}[|||\Sigma_1|||] \leqslant (1 - \varepsilon)|||\Sigma_1|||$$

2. Drift condition

$$V(z, \Sigma) = \alpha \|z\|^2 + \beta |||\Sigma|||$$

(a) When $|||\Sigma_0||| \gg 1 + \|z_0\|^2$:

$$\mathbb{E}[|||\Sigma_1|||] \leqslant (1-\varepsilon)|||\Sigma_1|||$$

(b) When $|||\Sigma_0||| \not\gg \|z_0\|^2$:

$$\mathbb{E}[\|z_1\|^2] \leqslant (1-\varepsilon)\|z_0\|^2$$

2. Drift condition

$$V(z,\Sigma) = \alpha\|z\|^2 + \beta|||\Sigma|||$$

(a) When $|||\Sigma_0||| \gg 1 + \|z_0\|^2$:

$$\mathbb{E}[|||\Sigma_1|||] \leqslant (1-\varepsilon)|||\Sigma_1|||$$

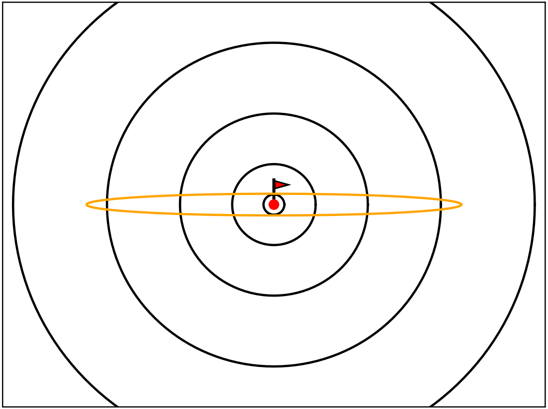
(b) When $|||\Sigma_0||| \not\gg \|z_0\|^2$:

$$\mathbb{E}[\|z_1\|^2] \leqslant (1-\varepsilon)\|z_0\|^2$$

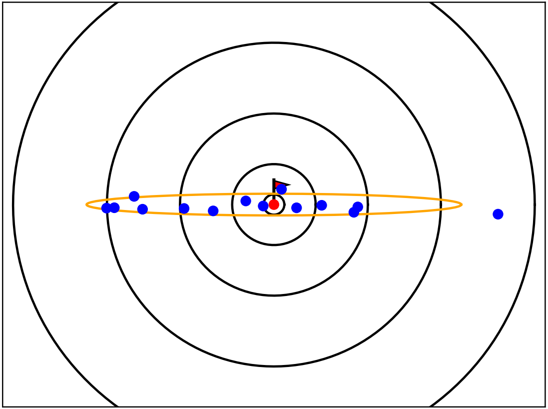
Proposition*
If (a) and (b) are true:

$$\exists \mathrm{K} \textit{ compact}, \ \mathbb{E}[V(z_1,\Sigma_1)] \leqslant (1-\varepsilon)V(z_0,\Sigma_0) \quad \forall (z_0,\Sigma_0) \notin \mathrm{K}$$

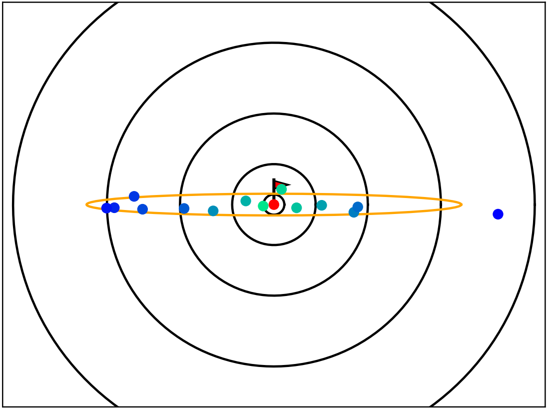
(a) When $\|\Sigma_0\| \gg 1 + \|z_0\|^2$



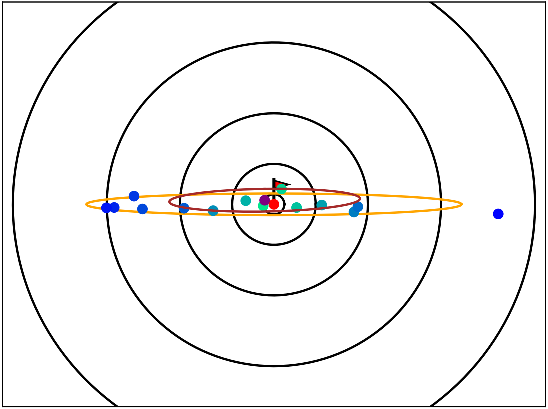
(a) When $\|\Sigma_0\| \gg 1 + \|z_0\|^2$



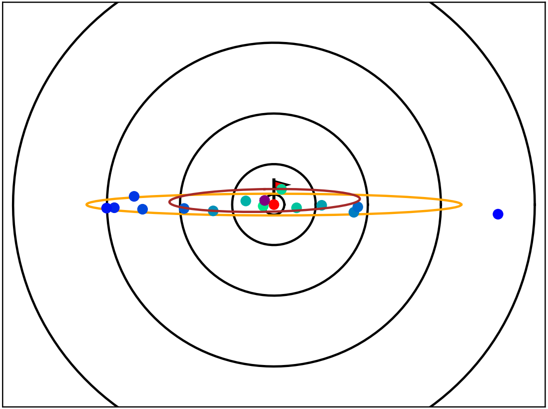
(a) When $\|\Sigma_0\| \gg 1 + \|z_0\|^2$



(a) When $\|\Sigma_0\| \gg 1 + \|z_0\|^2$



(a) When $\|\Sigma_0\| \gg 1 + \|z_0\|^2$



Proposition* ([GAH23], [GAHa])

When $f =$  *and* $\|\Sigma_0\| \gg 1 + \|z_0\|^2$*:*

$$\mathbb{E}[\|\Sigma_1\|] \leqslant (1 - \varepsilon)\|\Sigma_1\|$$

(b) When $\|\Sigma_0\| \not\gg \|z_0\|^2$

(b) When $\|\Sigma_0\| \not\gg \|z_0\|^2$

$$z_1 = \frac{\text{Average}(z_1^{1:\lambda}, \ldots, z_1^{\mu:\lambda})}{\text{normalization}}$$

(b) When $\|\Sigma_0\| \not\gg \|z_0\|^2$

$$z_1 = \frac{\text{Average}(z_1^{1:\lambda}, \ldots, z_1^{\mu:\lambda})}{\text{normalization}}$$

with

$$\text{normalization} = \text{increasing function}(\|m_{t+1} - m_t\|) \times \sqrt{\lambda_{\min}(\Sigma_1)}$$

(b) When $|||\Sigma_0||| \not\gg \|z_0\|^2$

$$z_1 = \frac{\text{Average}(z_1^{1:\lambda}, \dots, z_1^{\mu:\lambda})}{\text{normalization}}$$

with

$$\text{normalization} = \text{increasing function}(\|m_{t+1} - m_t\|) \times \sqrt{\lambda_{\min}(\Sigma_1)}$$

Proposition* ([GAHa])

When $f =$  *and* $|||\Sigma_0||| \not\gg \|z_0\|^2$*:*

$$\mathbb{E}[\|m_{t+1} - m_t\|] > \mathbb{E}\|\mathcal{N}\|$$

(b) When $\|\Sigma_0\| \ggg \|z_0\|^2$

$$z_1 = \frac{\text{Average}(z_1^{1:\lambda}, \ldots, z_1^{\mu:\lambda})}{\text{normalization}}$$

with

$$\text{normalization} = \text{increasing function}(\|m_{t+1} - m_t\|) \times \sqrt{\lambda_{\min}(\Sigma_1)}$$

Proposition* ([GAHa])

When $f =$  *and* $\|\Sigma_0\| \ggg \|z_0\|^2$*:*

$$\mathbb{E}[\|m_{t+1} - m_t\|] > \mathbb{E}\|\mathcal{N}\|$$

If we choose the hyperparameters correctly:

$$\mathbb{E}[\textit{normalization}] > 1$$

(b) When $|||\Sigma_0||| \not\gg \|z_0\|^2$

$$z_1 = \frac{\text{Average}(z_1^{1:\lambda}, \ldots, z_1^{\mu:\lambda})}{\text{normalization}}$$

with

$$\text{normalization} = \text{increasing function}(\|m_{t+1} - m_t\|) \times \sqrt{\lambda_{\min}(\Sigma_1)}$$

Proposition* ([GAHa])

When $f =$  *and $|||\Sigma_0||| \not\gg \|z_0\|^2$:*

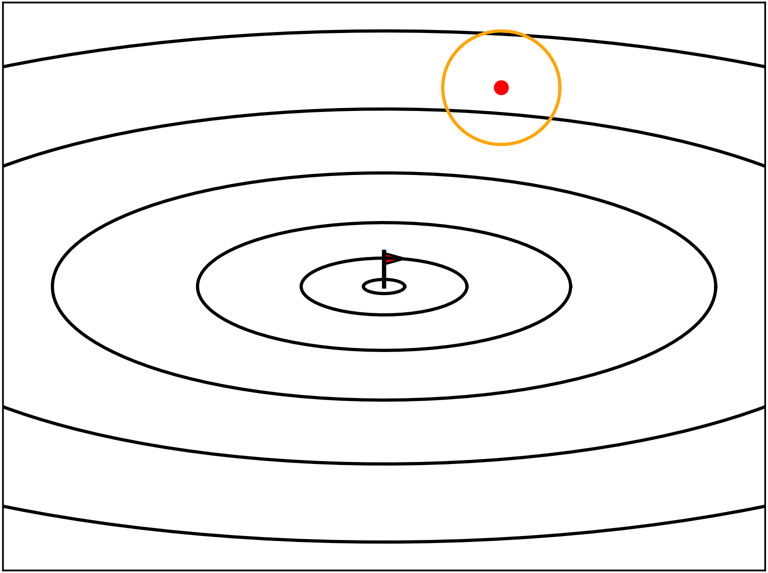
$$\mathbb{E}[\|m_{t+1} - m_t\|] > \mathbb{E}\|\mathcal{N}\|$$

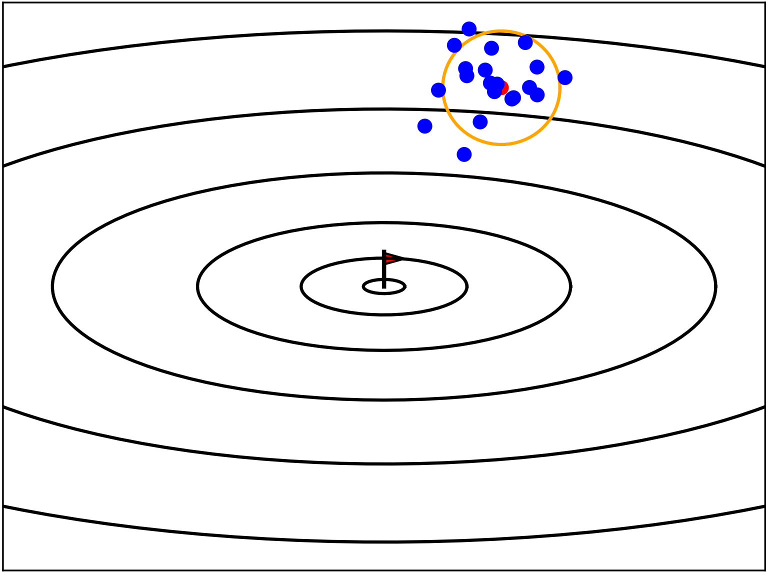
If we choose the hyperparameters correctly:

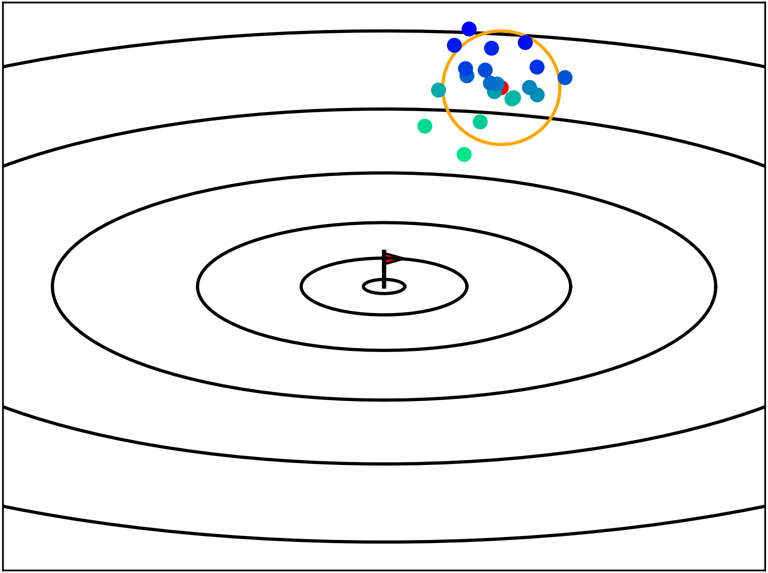
$$\mathbb{E}[normalization] > 1$$

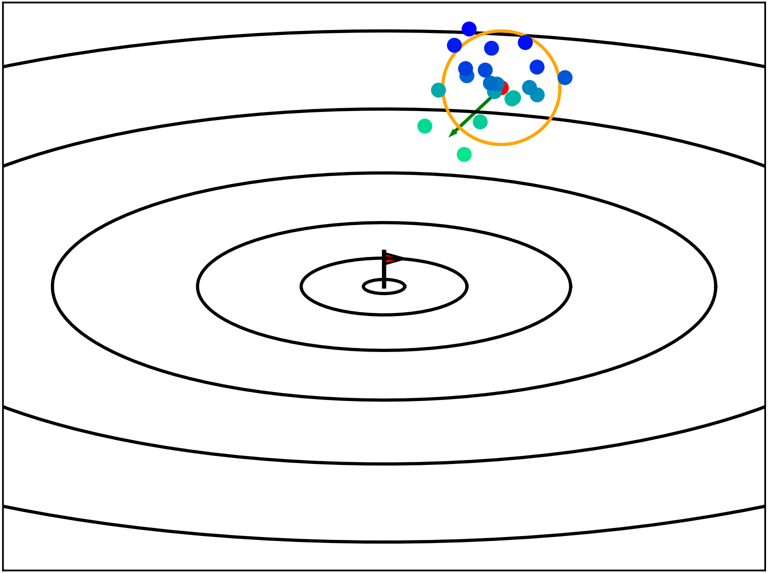
and

$$\mathbb{E}[\|z_1\|^2] \leqslant (1 - \varepsilon)\|z_0\|^2$$









Theorem* ([GAHa])

When $f =$ 

$$\exists \mathsf{K} \text{ compact}, \ \mathbb{E}[V(z_1, \Sigma_1)] \leqslant (1-\varepsilon) V(z_0, \Sigma_0) \quad \forall (z_0, \Sigma_0) \notin \mathsf{K}$$

Theorem* ([GAHb])

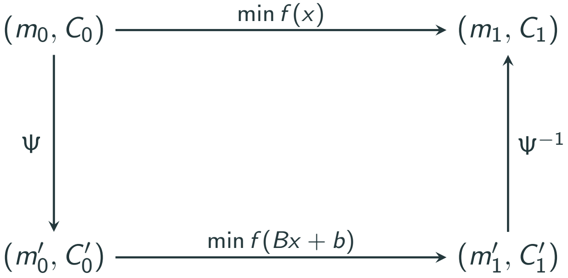
When $f =$ *, CMA-ES converges linearly.*

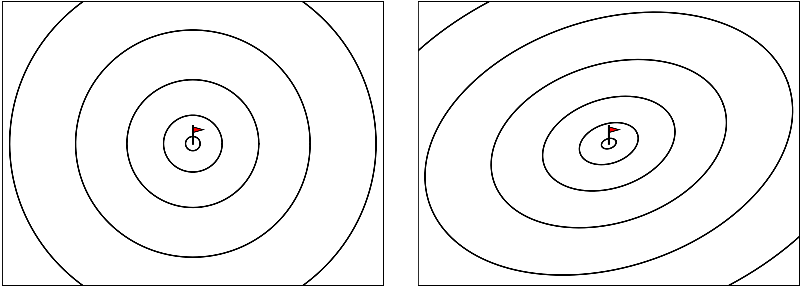
Theorem* ([GAHb])

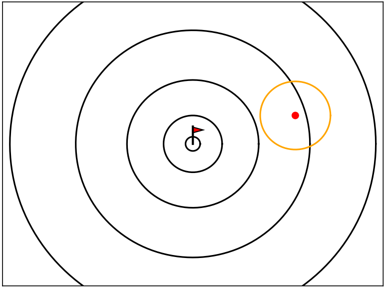
When $f =$ *, CMA-ES converges linearly.*

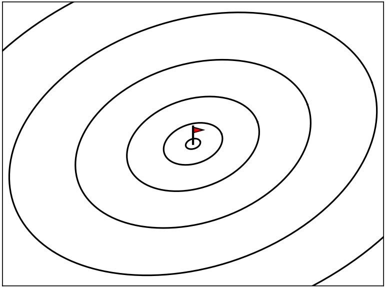
How to extend to $f =$ **?**

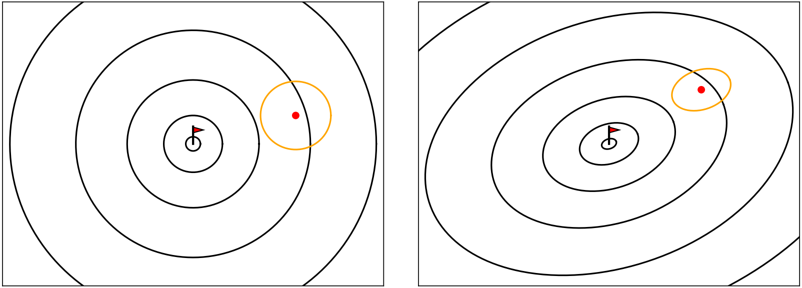
Affine-invariance

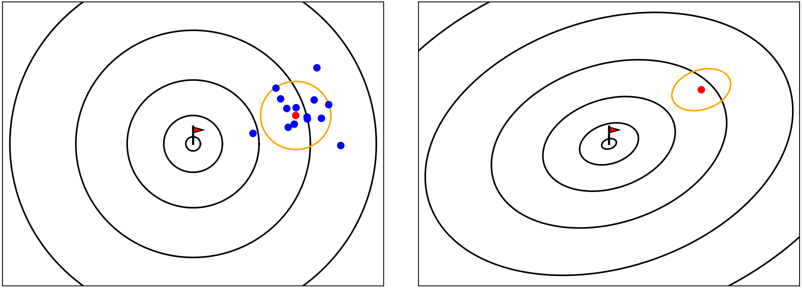


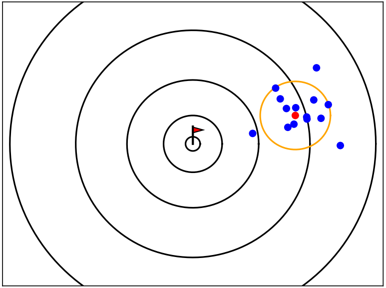


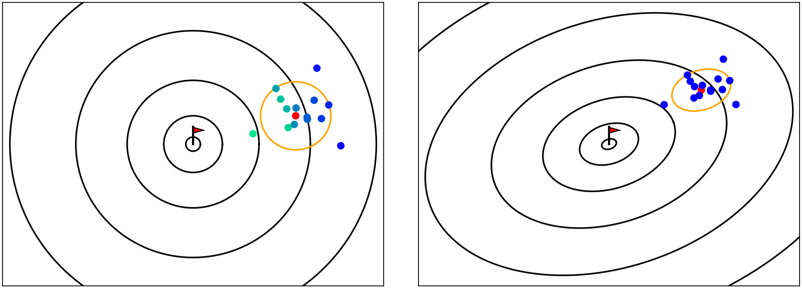


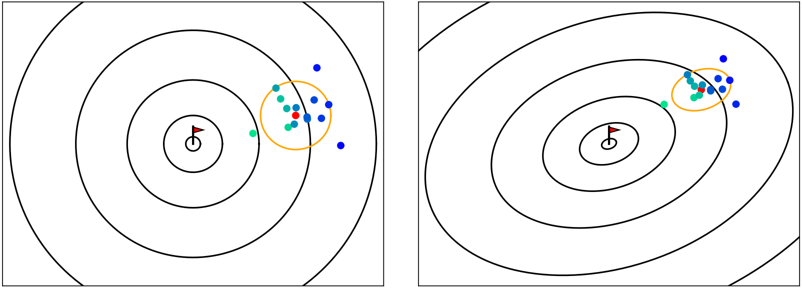


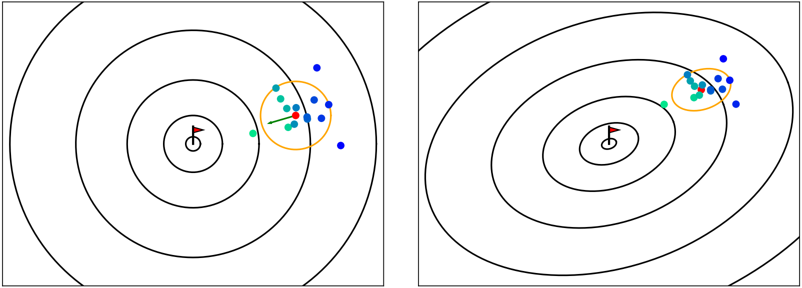


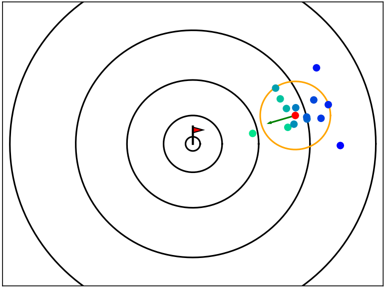


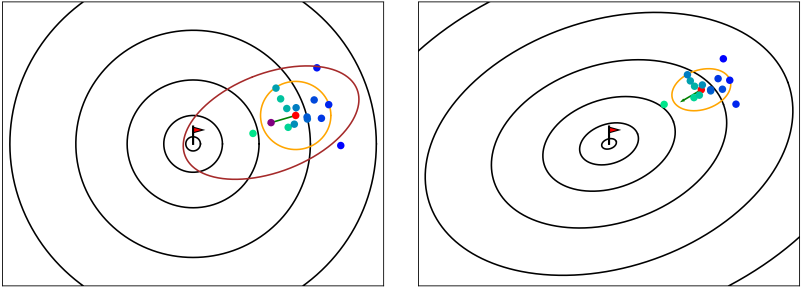


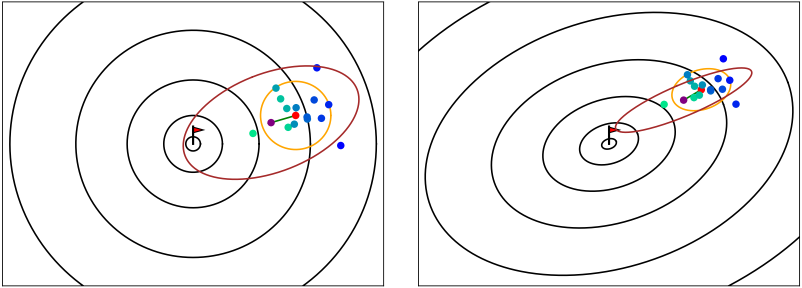












Theorem ([HA14], [A16])
CMA-ES is affine-invariant

Theorem* ([GAHb])

When f = *, CMA-ES converges linearly.*

Theorem* ([GAHb])

When f = *, CMA-ES converges linearly.*

(with the same convergence rate than *)*

Learning of the inverse Hessian

When $f =$ , we find

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{C_t}{\text{normalization}}\right] = I_d$$

Learning of the inverse Hessian

When $f =$ , we find

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{C_t}{\text{normalization}}\right] = I_d$$

Since  $=$ Hessian$^{-1/2} \times$ :

$f =$  $\Rightarrow \lim_{t\to\infty} \mathbb{E}\left[\frac{C_t}{\text{normalization}}\right] = \text{Hessian}^{-1/2} \times I_d \times \text{Hessian}^{-1/2}$

$= \text{Hessian}^{-1}$

Learning of the inverse Hessian

When $f =$ , we find

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{C_t}{\text{normalization}}\right] = I_d$$

Since  $=$ Hessian$^{-1/2} \times$ :

$f =$ 

$$\Rightarrow \lim_{t\to\infty} \mathbb{E}\left[\frac{C_t}{\text{normalization}}\right] = \text{Hessian}^{-1/2} \times I_d \times \text{Hessian}^{-1/2}$$

$$= \text{Hessian}^{-1}$$

Theorem* ([GAHb])

CMA-ES learns the inverse Hessian of .

Conclusions

- CMA-ES converges linearly when $f =$ 
- The convergence rate does not depend on 
- The covariance matrix approximates the inverse Hessian

Thank you

Bibliography (1/4)

[MC91] Meyn & Caines, 1991, Asymptotic Behavior Stochastic Systems Possessing Markov Processes

[HO01] Hansen & Ostermeier, 2001, Completely Derandomized Self-Adaptation in Evolution Strategies

[HAK03] Hansen, Müller & Koumoutsakos, 2003, Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES)

[MT09] Meyn & Tweedie, 2009, Markov Chains and Stochastic Stability

Bibliography (2/4)

[HA14] Hansen & Auger, 2014, Principled Design of Continuous Stochastic Search: From Theory to Practice

[AH16] Auger & Hansen, 2016, Linear Convergence of Comparison-based Step-size Adaptive Randomized Search via Stability of Markov Chains

[A16] Auger, 2016, Analysis of Comparison-based Stochastic Continuous Black-Box Optimization Algorithms

[CA19] Chotard & Auger, 2019, Verifiable conditions for the irreducibility and aperiodicity of Markov chains by analyzing underlying deterministic models

Bibliography (3/4)

[TGAH21] Touré, Gissler, Auger & Hansen, 2021, Scaling-invariant Functions versus Positively Homogeneous Functions

[TAH23] Touré, Auger & Hansen, 2023, Global linear convergence of evolution strategies with recombination on scaling-invariant functions

[GAH23] Gissler, Auger & Hansen, 2023, Asymptotic estimations of a perturbed symmetric eigenproblem

[GDA24] Gissler, Durmus & Auger, 2024, On the irreducibility and convergence of a class of nonsmooth nonlinear state-space models on manifolds

Bibliography (4/4)

[GWAH] Gissler, Wolfe, Auger & Hansen, soon submitted, Irreducible nonsmooth state-space models and application to CMA-ES

[GAHa] Gissler, Auger & Hansen, soon submitted, A (state-dependent) Foster-Lyapunov drift condition for CMA-ES

[GAHb] Gissler, Auger & Hansen, soon submitted, Linear convergence of CMA-ES and learning of second-order information of ellipsoid functions