Convergence of Evolution Strategies with Covariance Matrix Adaptation (CMA-ES)

Armand Gissler

Tuesday 26th September, 2023

CMAP, École polytechnique & Inria
(with Anne Auger & Nikolaus Hansen)

Black-box optimisation and Evolution strategies

Consider the optimisation problem

$$\min_{x \in \mathbb{R}^d} f(x) \tag{P}$$

Black-box optimisation and Evolution strategies

Consider the optimisation problem

$$\min_{x \in \mathbb{R}^d} f(x) \tag{P}$$

with

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

Black-box optimisation and Evolution strategies

Consider the optimisation problem

$$\min_{x \in \mathbb{R}^d} f(x) \tag{P}$$

with

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

$\Rightarrow$ **we only have access to a minimum amount of informations on** f

Black-box optimisation and Evolution strategies

Consider the optimisation problem

$$\min_{x \in \mathbb{R}^d} f(x) \tag{P}$$

with

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

$\Rightarrow$ we only have access to a minimum amount of informations on f
(in particular no information on the derivatives of f)

Black-box optimisation and Evolution strategies

Consider the optimisation problem

$$\min_{x \in \mathbb{R}^d} f(x) \tag{P}$$

with

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

$\Rightarrow$ we only have access to a minimum amount of informations on f (in particular no information on the derivatives of f)

CMA-ES approximates the minimum x^* of f by a multivariate normal distribution $\mathcal{N}(m, C)$

Black-box optimisation and Evolution strategies

Consider the optimisation problem

$$\min_{x \in \mathbb{R}^d} f(x) \tag{P}$$

with

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

$\Rightarrow$ we only have access to a minimum amount of informations on f (in particular no information on the derivatives of f)

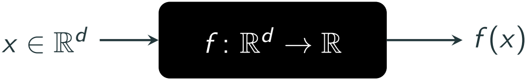
CMA-ES approximates the minimum x^* of f by a multivariate normal distribution $\mathcal{N}(m, C)$ **by adapting the mean** $m \in \mathbb{R}^d$

Black-box optimisation and Evolution strategies

Consider the optimisation problem

$$\min_{x \in \mathbb{R}^d} f(x) \tag{P}$$

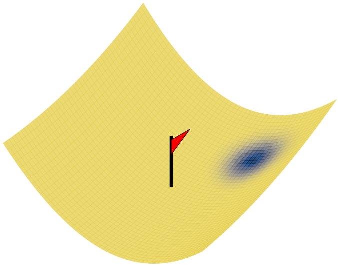
with



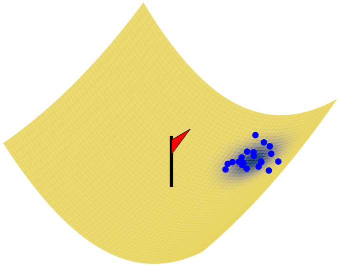
$\Rightarrow$ we only have access to a minimum amount of informations on f (in particular no information on the derivatives of f)

CMA-ES approximates the minimum x^* of f by a multivariate normal distribution $\mathcal{N}(m, C)$ **by adapting** the mean $m \in \mathbb{R}^d$ **and the covariance matrix** $C \in \mathcal{S}^d_{++}$.

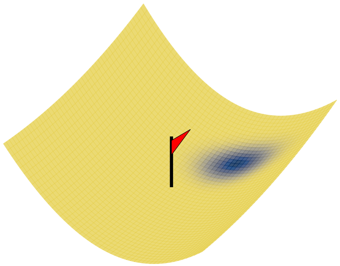
For a simple problem



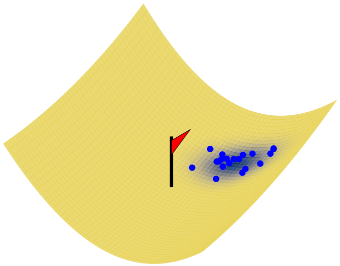
For a simple problem



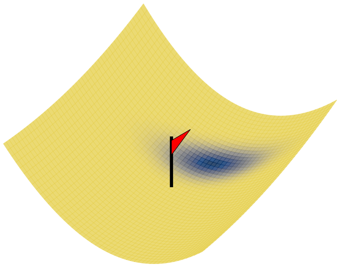
For a simple problem



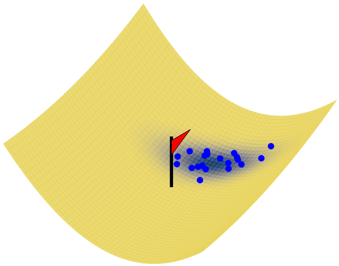
For a simple problem



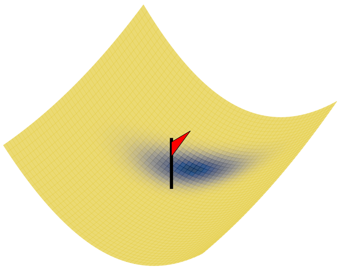
For a simple problem



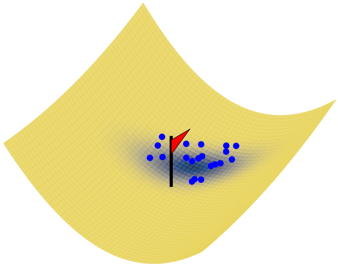
For a simple problem



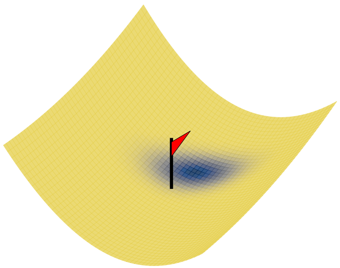
For a simple problem



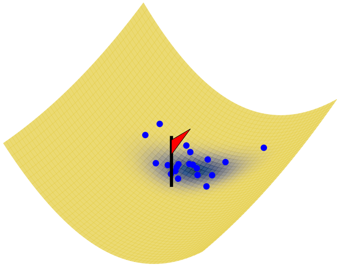
For a simple problem



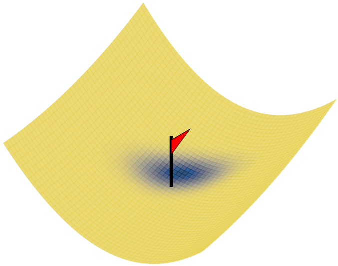
For a simple problem



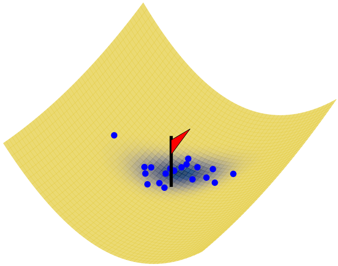
For a simple problem



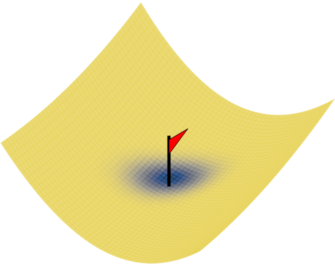
For a simple problem



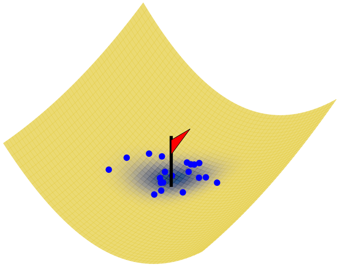
For a simple problem



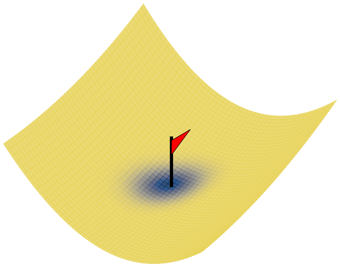
For a simple problem



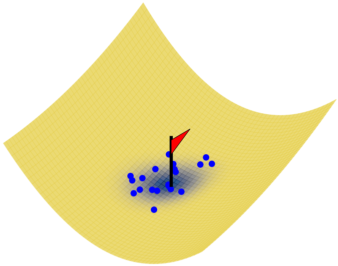
For a simple problem



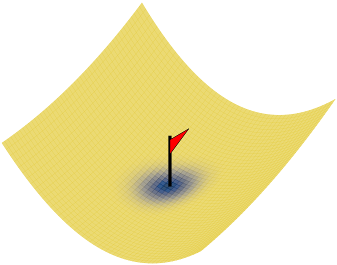
For a simple problem



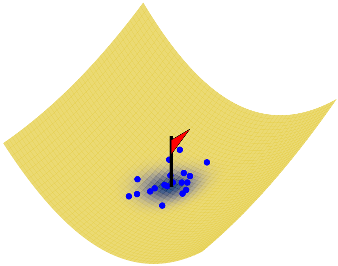
For a simple problem



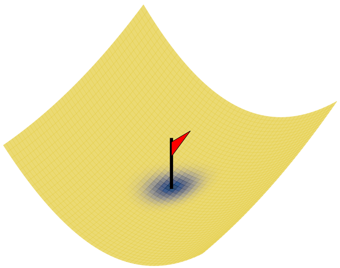
For a simple problem



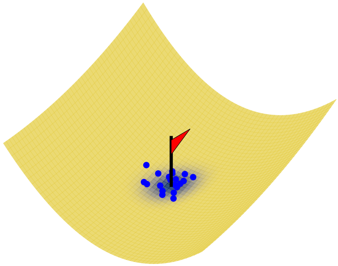
For a simple problem



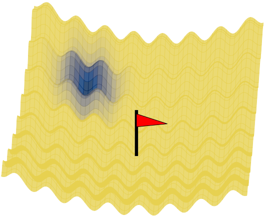
For a simple problem



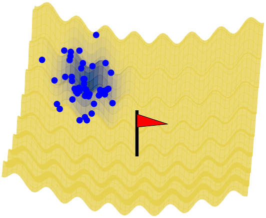
For a simple problem



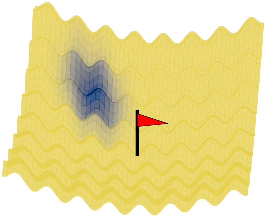
and a complicated problem



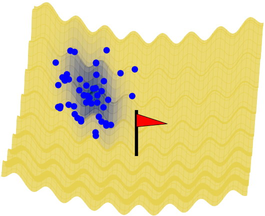
and a complicated problem



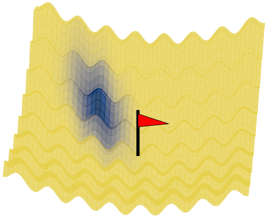
and a complicated problem



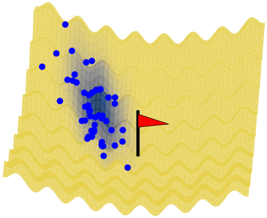
and a complicated problem



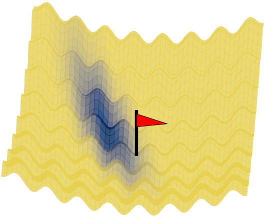
and a complicated problem



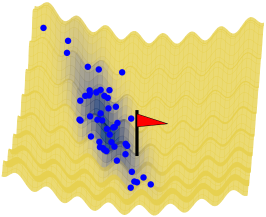
and a complicated problem



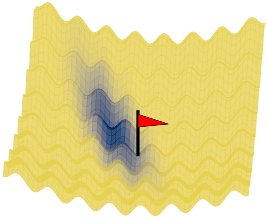
and a complicated problem



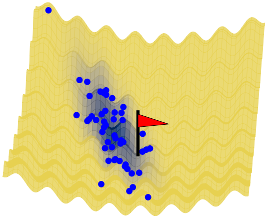
and a complicated problem



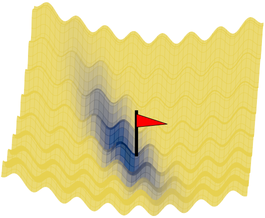
and a complicated problem



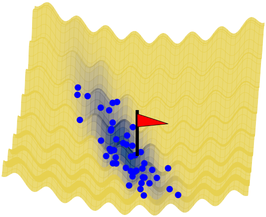
and a complicated problem



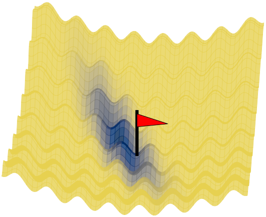
and a complicated problem



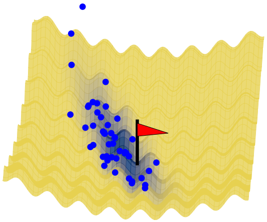
and a complicated problem



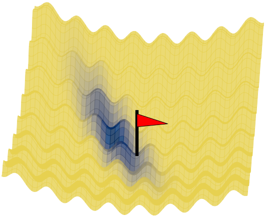
and a complicated problem



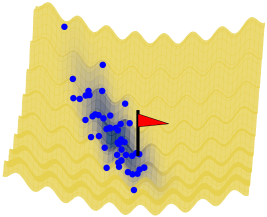
and a complicated problem



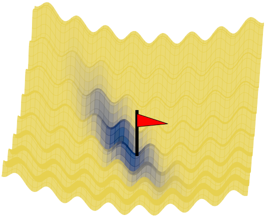
and a complicated problem



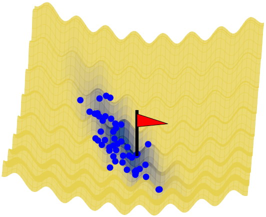
and a complicated problem



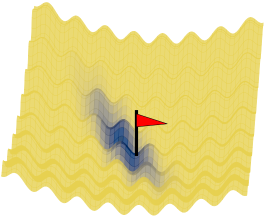
and a complicated problem



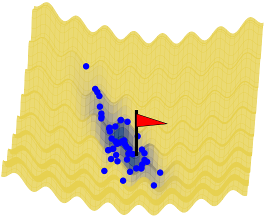
and a complicated problem



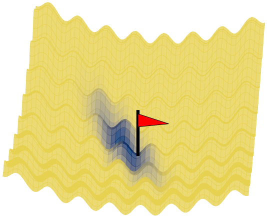
and a complicated problem



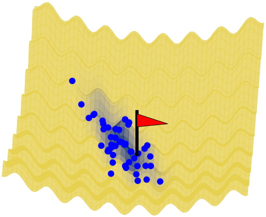
and a complicated problem



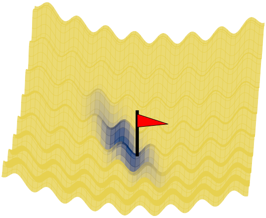
and a complicated problem



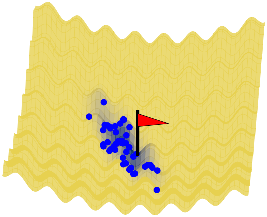
and a complicated problem



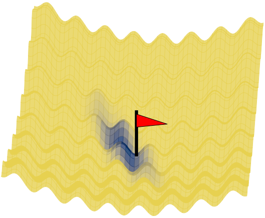
and a complicated problem



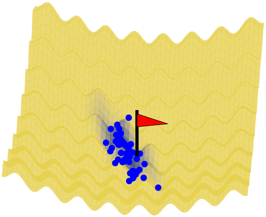
and a complicated problem



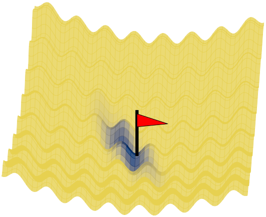
and a complicated problem



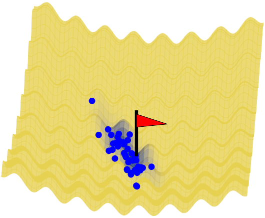
and a complicated problem



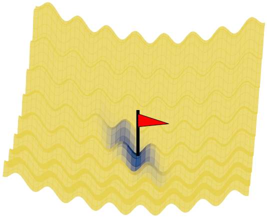
and a complicated problem



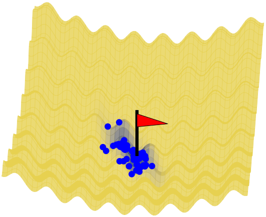
and a complicated problem



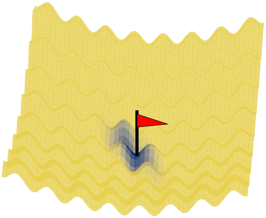
and a complicated problem



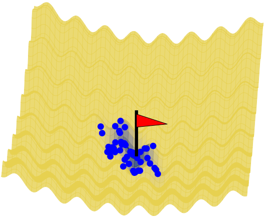
and a complicated problem



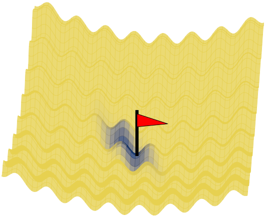
and a complicated problem



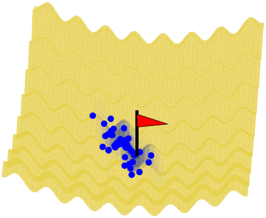
and a complicated problem



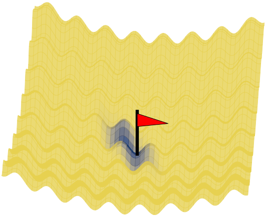
and a complicated problem



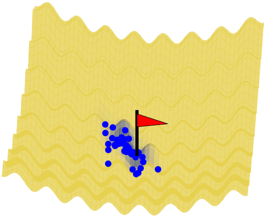
and a complicated problem



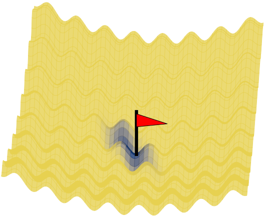
and a complicated problem



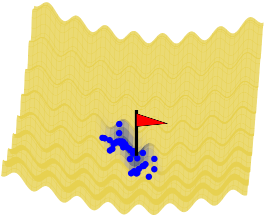
and a complicated problem



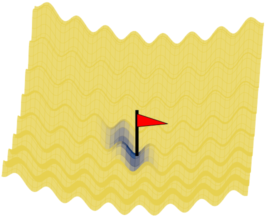
and a complicated problem



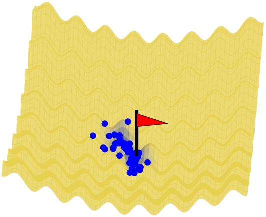
and a complicated problem



and a complicated problem

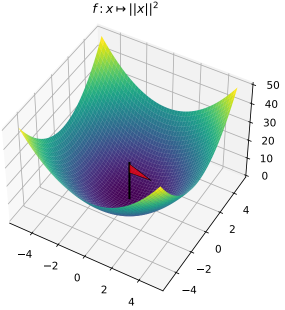


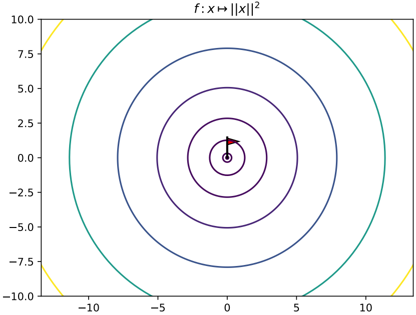
and a complicated problem



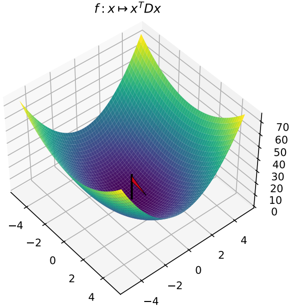
CMA-ES: algorithm presentation

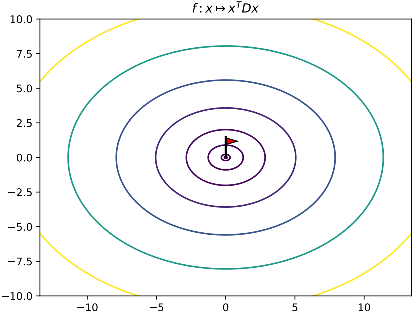
Level sets representation



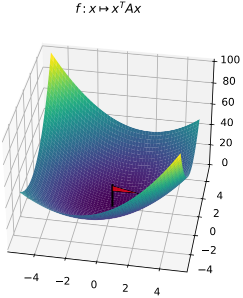


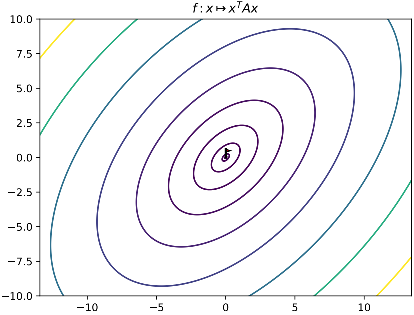
Level sets representation



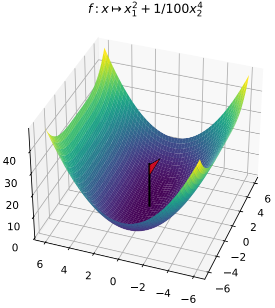


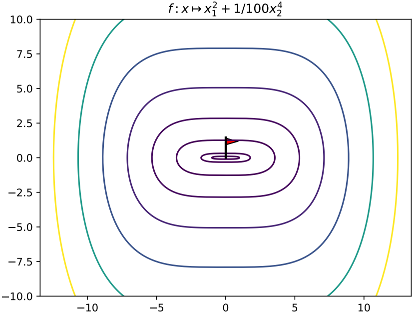
Level sets representation



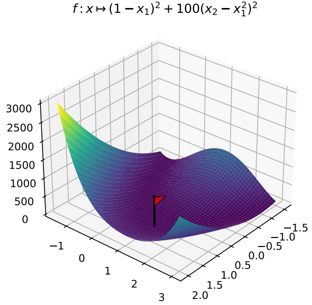


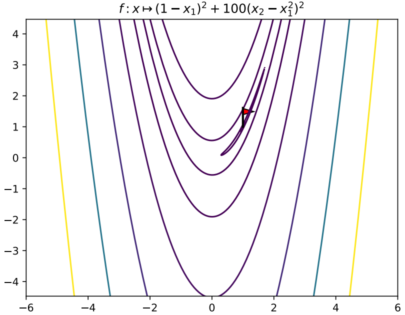
Level sets representation



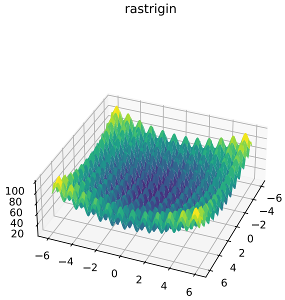


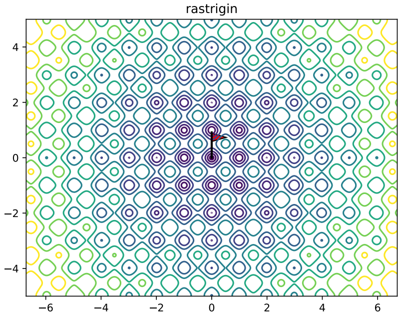
Level sets representation





Level sets representation

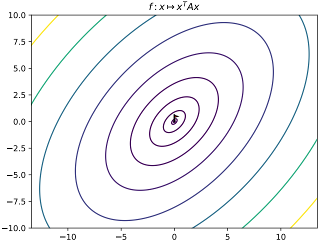




Presentation of the algorithm

Algorithm 1 CMA-ES

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

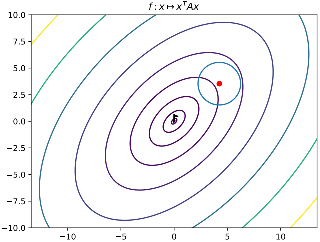


Presentation of the algorithm

Algorithm 1 CMA-ES

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

Given: $m_0 \in \mathbb{R}^d$, $C_0 \in \mathcal{S}^d_{++}$



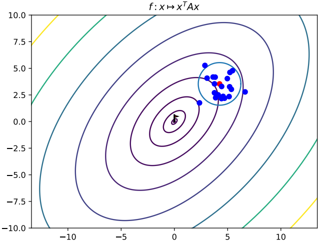
Algorithm 1 CMA-ES

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

Given: $m_0 \in \mathbb{R}^d$, $C_0 \in \mathcal{S}^d_{++}$

For $t = 0, 1, 2, \ldots$:

1. $x^1_{t+1}, \ldots, x^\lambda_{t+1} \sim \mathcal{N}(m_t, C_t)$



λ population size

Presentation of the algorithm

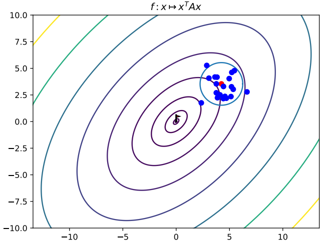
Algorithm 1 CMA-ES

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

Given: $m_0 \in \mathbb{R}^d$, $C_0 \in \mathcal{S}^d_{++}$

For $t = 0, 1, 2, \dots$:

1. $x^1_{t+1}, \dots, x^\lambda_{t+1} \sim \mathcal{N}(m_t, C_t)$
2. sort $f(x^i_{t+1})$:



λ population size

Presentation of the algorithm

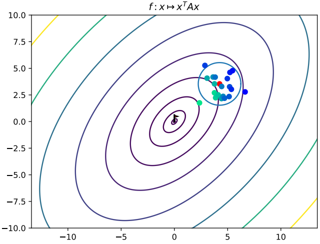
Algorithm 1 CMA-ES

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

Given: $m_0 \in \mathbb{R}^d$, $C_0 \in \mathcal{S}^d_{++}$

For $t = 0, 1, 2, \ldots$:

1. $x^1_{t+1}, \ldots, x^\lambda_{t+1} \sim \mathcal{N}(m_t, C_t)$
2. sort $f(x^i_{t+1})$:
   $f\left(x^{1:\lambda}_{t+1}\right) \leqslant \cdots \leqslant f\left(x^{\lambda:\lambda}_{t+1}\right)$



λ population size

Presentation of the algorithm

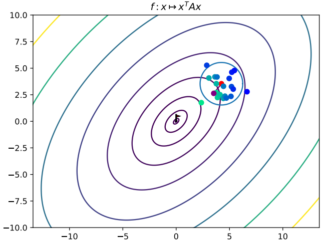
Algorithm 1 CMA-ES

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

Given: $m_0 \in \mathbb{R}^d$, $C_0 \in \mathcal{S}^d_{++}$

For $t = 0, 1, 2, \ldots$:

1. $x^1_{t+1}, \ldots, x^\lambda_{t+1} \sim \mathcal{N}(m_t, C_t)$
2. sort $f(x^i_{t+1})$:
   $f\left(x^{1:\lambda}_{t+1}\right) \leqslant \cdots \leqslant f\left(x^{\lambda:\lambda}_{t+1}\right)$
3. $m_{t+1} = \sum_{i=1}^{\mu} w_i x^{i:\lambda}_{t+1}$



λ population size

μ parent number

Presentation of the algorithm

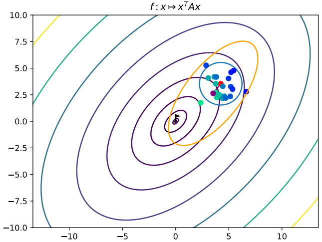
Algorithm 1 CMA-ES

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

Given: $m_0 \in \mathbb{R}^d$, $C_0 \in \mathcal{S}^d_{++}$

For $t = 0, 1, 2, \ldots$:

1. $x^1_{t+1}, \ldots, x^\lambda_{t+1} \sim \mathcal{N}(m_t, C_t)$
2. $\mathrm{sort}\ f(x^i_{t+1})$:
   $f\left(x^{1:\lambda}_{t+1}\right) \leqslant \cdots \leqslant f\left(x^{\lambda:\lambda}_{t+1}\right)$
3. $m_{t+1} = \sum_{i=1}^{\mu} w_i x^{i:\lambda}_{t+1}$
4. $C_{t+1} = (1-c)C_t + c \sum_{i=1}^{\mu} w_i \left[x^{i:\lambda}_{t+1} - m_t\right] \left[x^{i:\lambda}_{t+1} - m_t\right]^T$

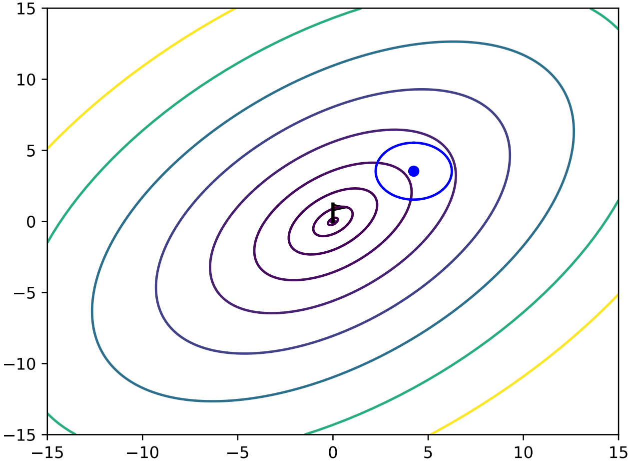


λ population size

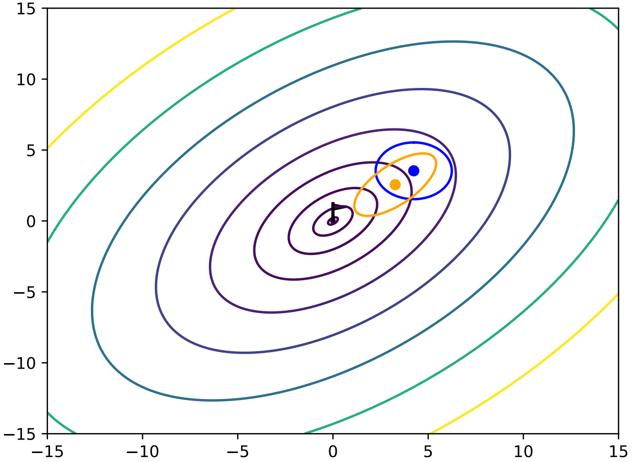
μ parent number

Linear convergence

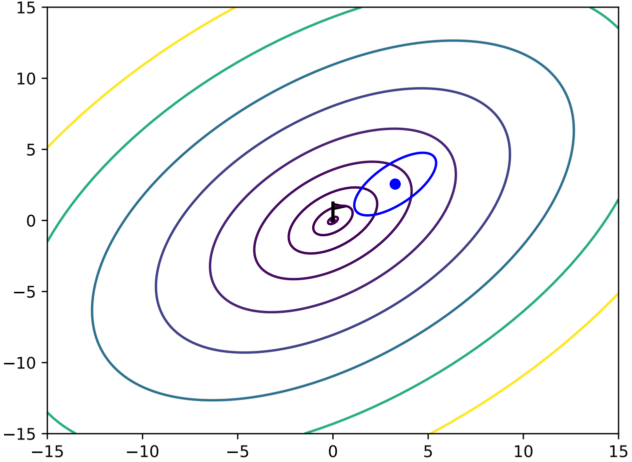
Convergence



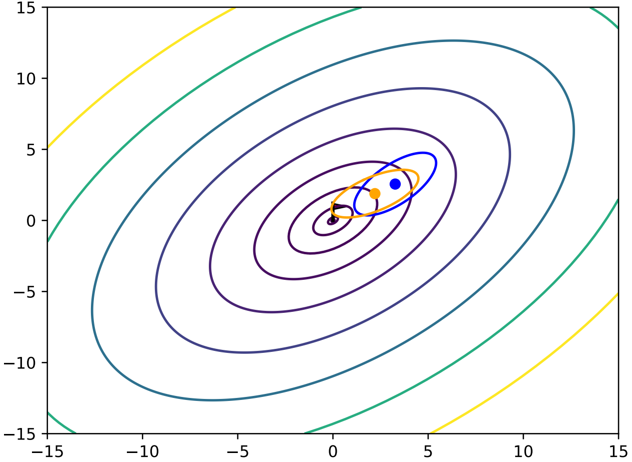
Convergence



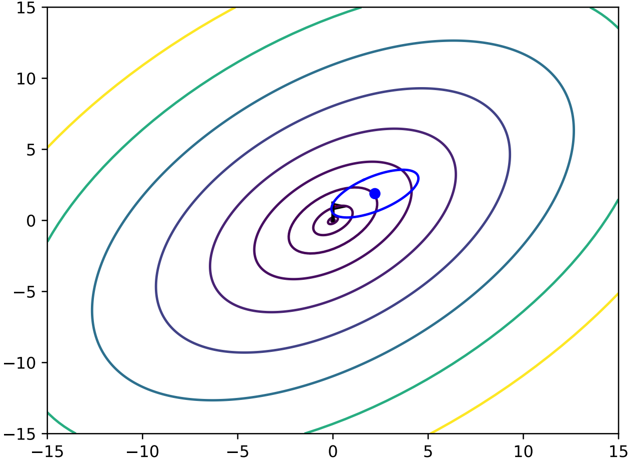
Convergence



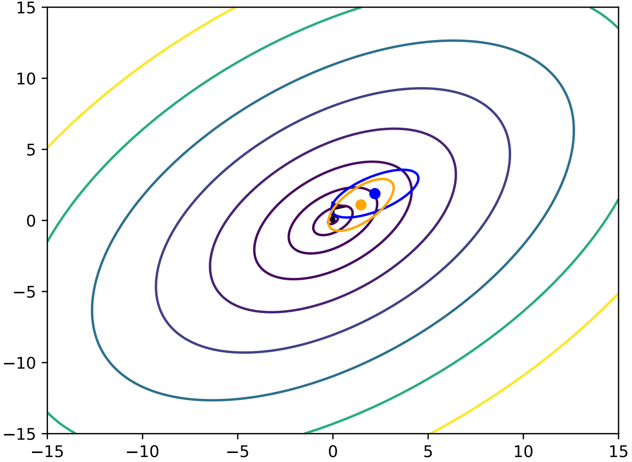
Convergence



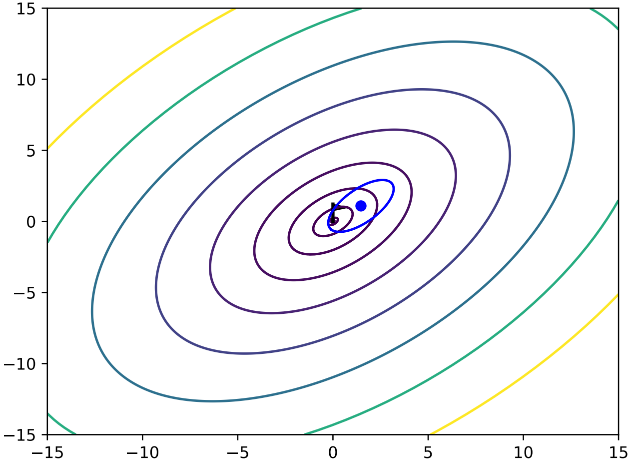
Convergence



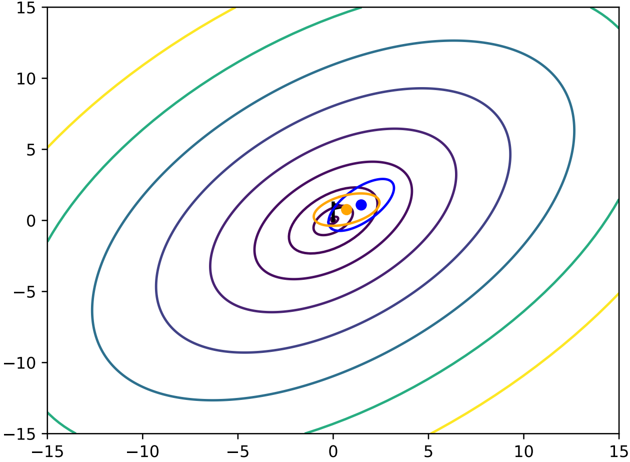
Convergence



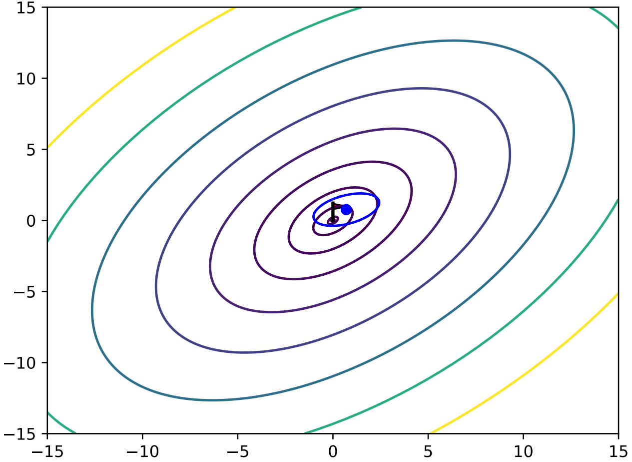
Convergence



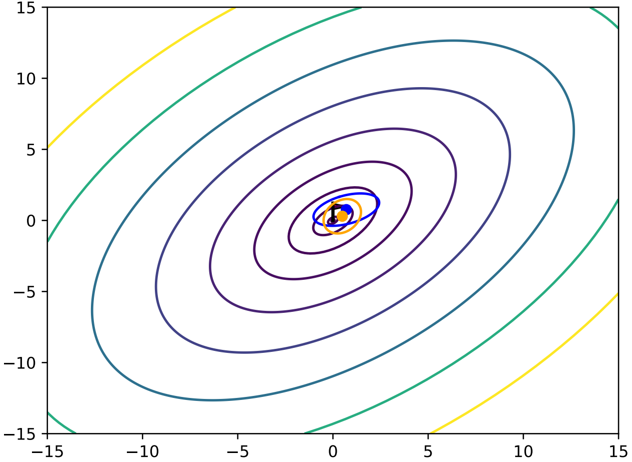
Convergence



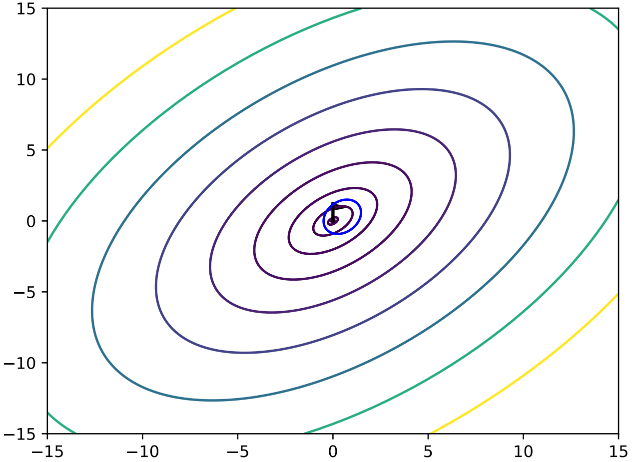
Convergence



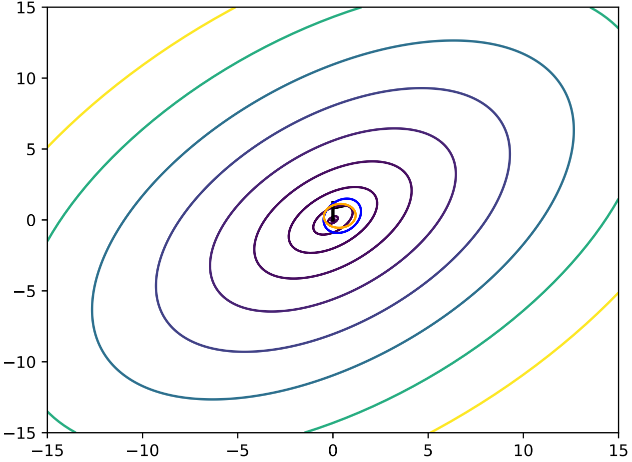
Convergence



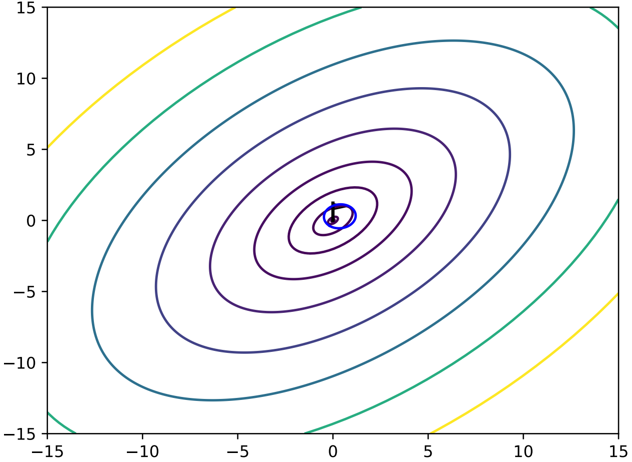
Convergence



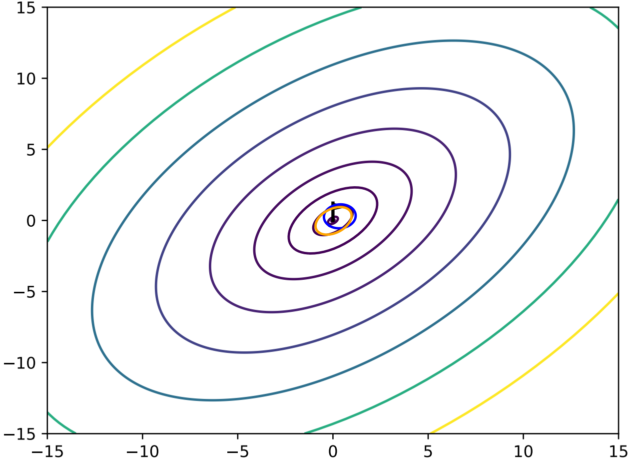
Convergence



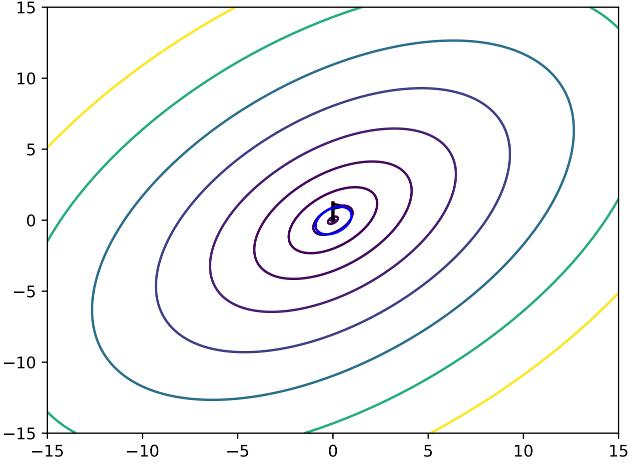
Convergence



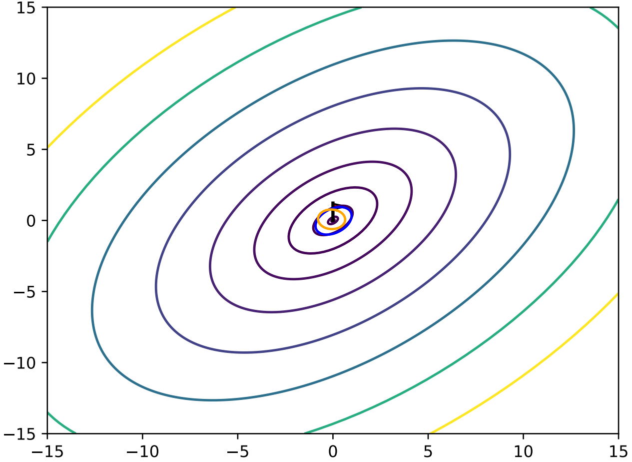
Convergence



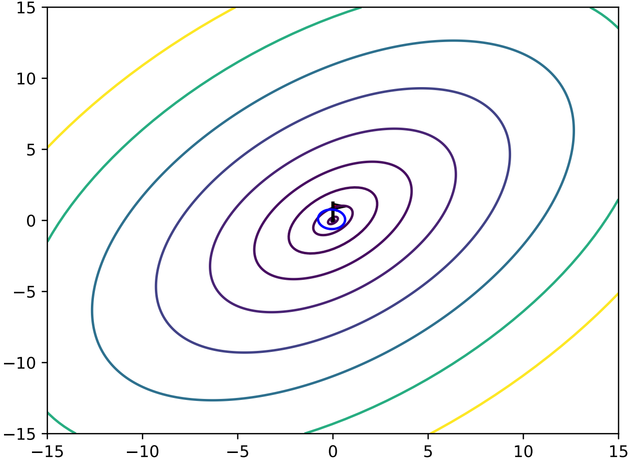
Convergence



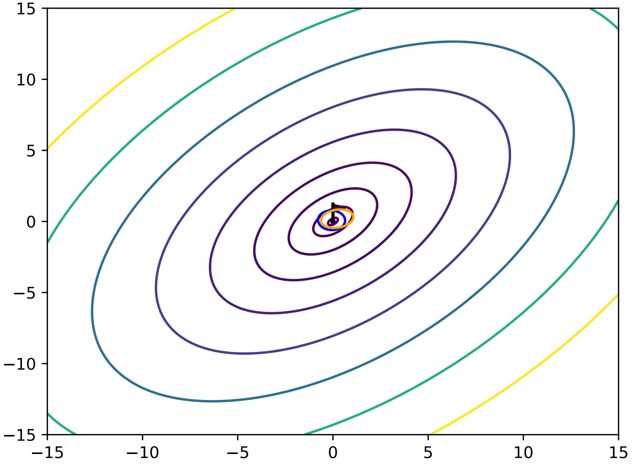
Convergence



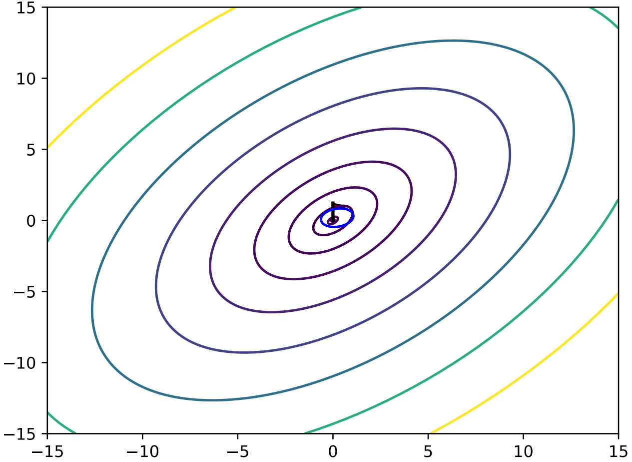
Convergence



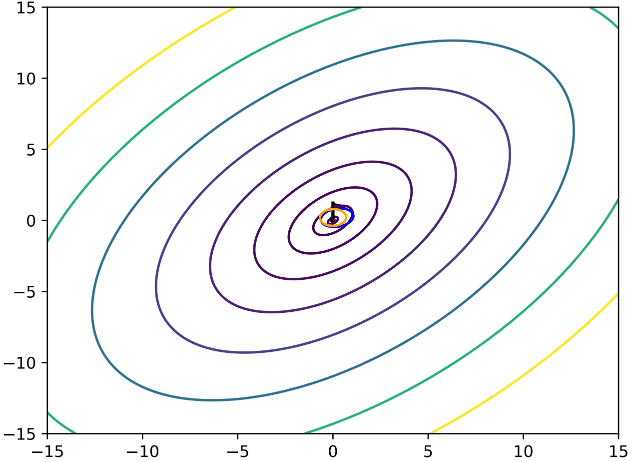
Convergence



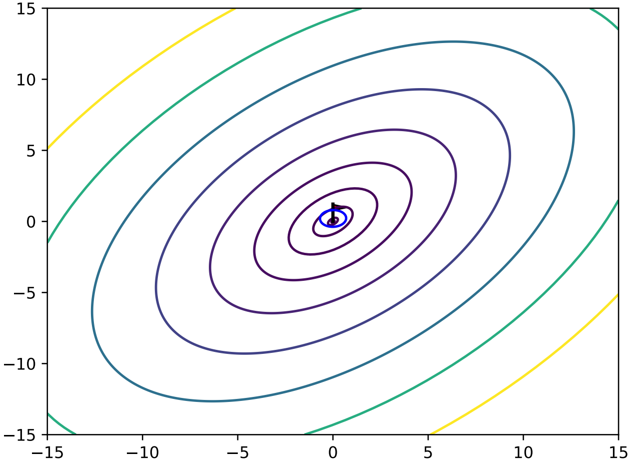
Convergence



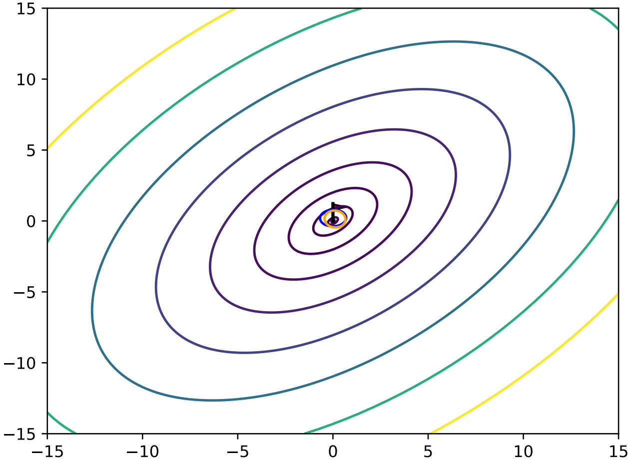
Convergence



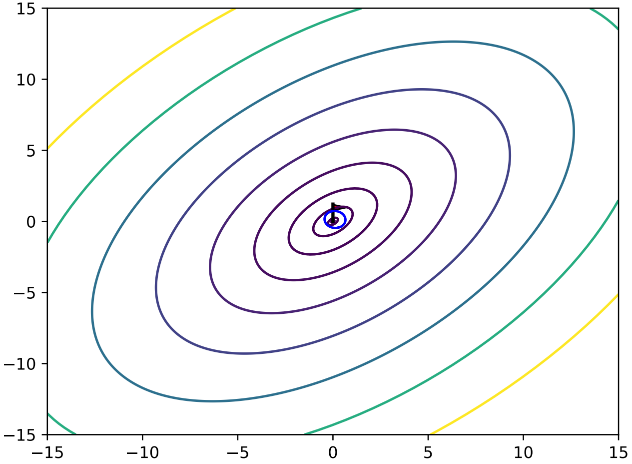
Convergence



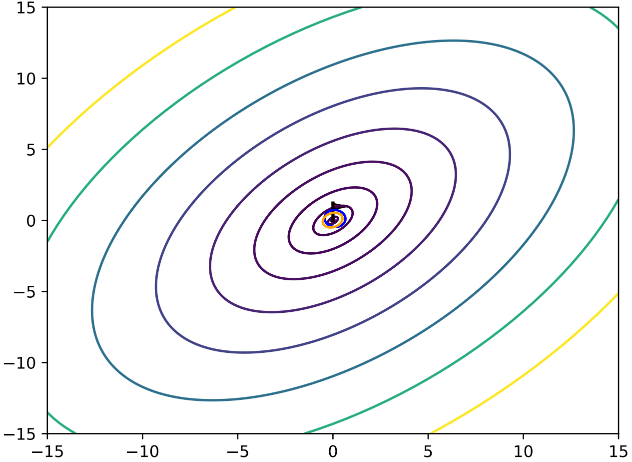
Convergence



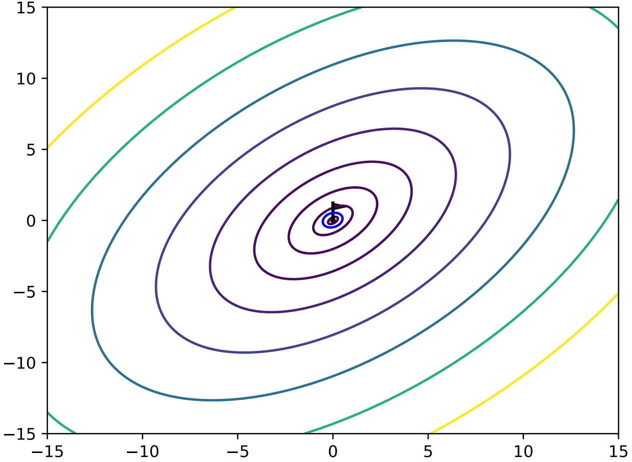
Convergence



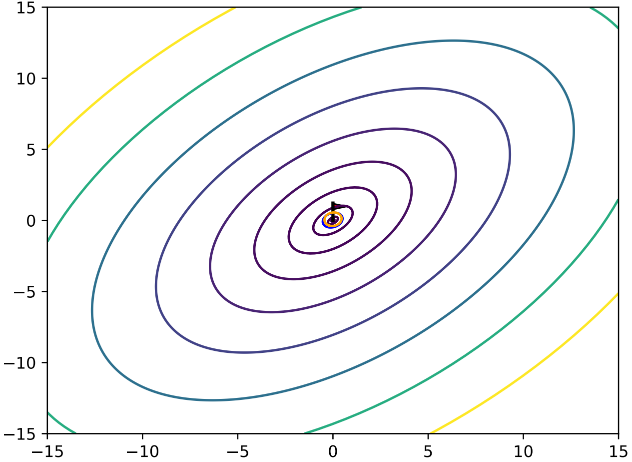
Convergence



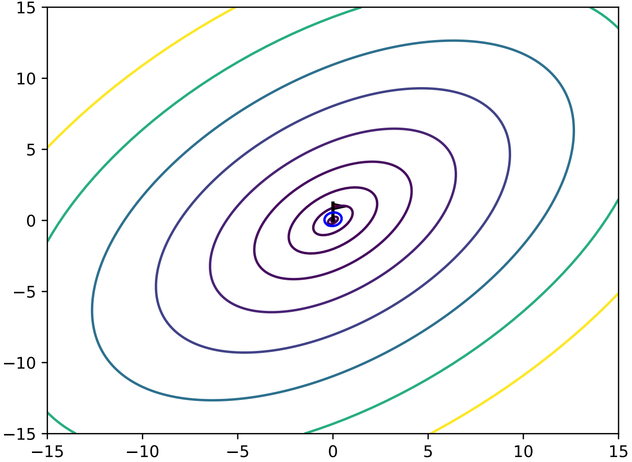
Convergence



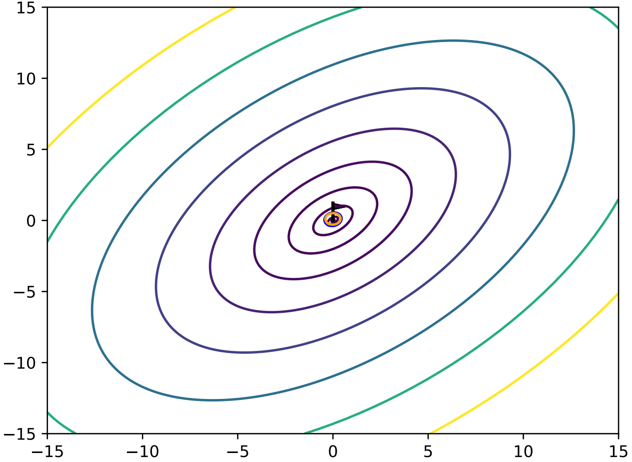
Convergence



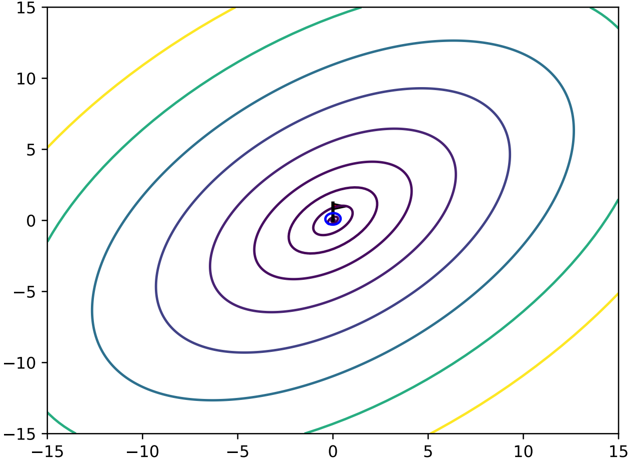
Convergence



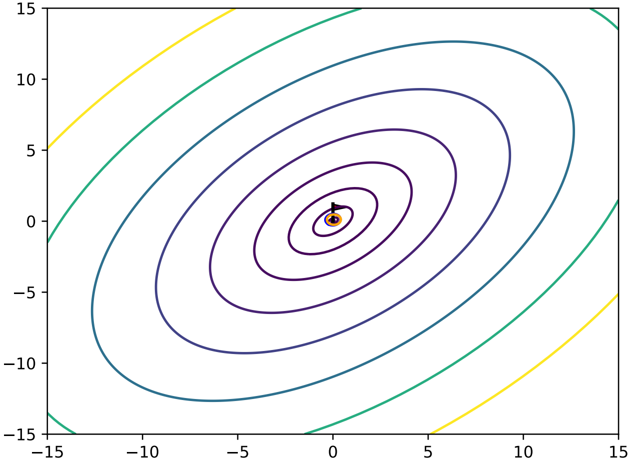
Convergence



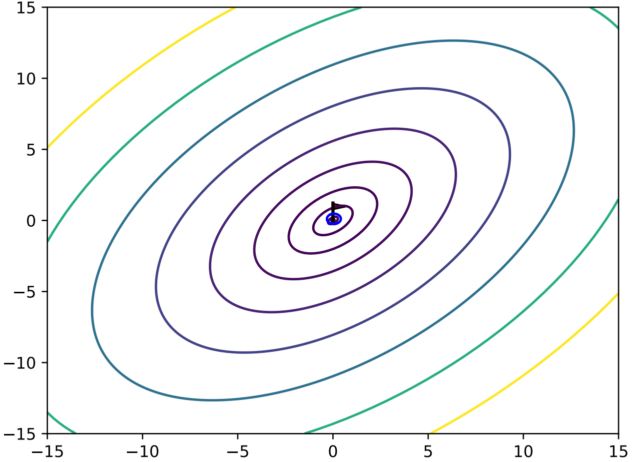
Convergence



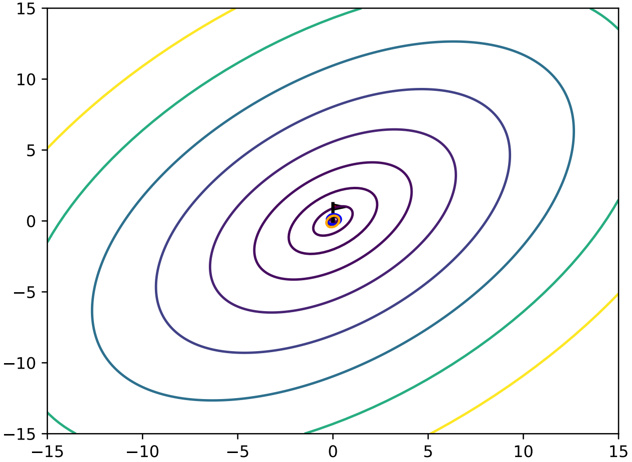
Convergence



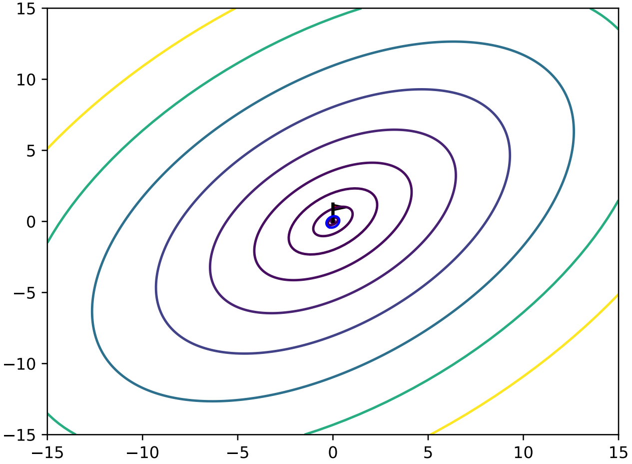
Convergence



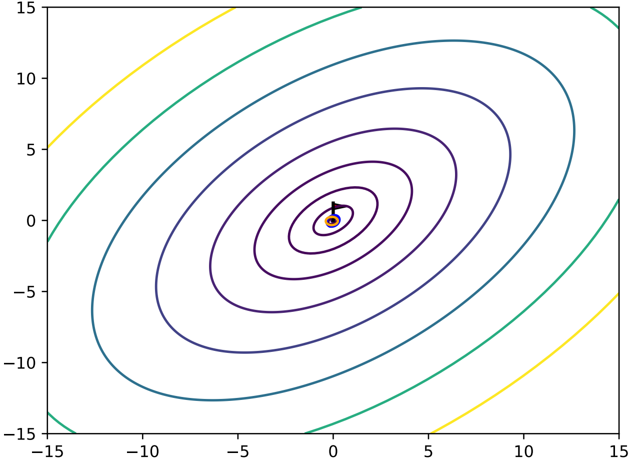
Convergence



Convergence



Convergence



Convergence

We observe

$$m_t \underset{t \to \infty}{\longrightarrow} x^* \in \arg\min f$$

and

$$C_t \underset{t \to \infty}{\longrightarrow} H^{-1}$$

Convergence

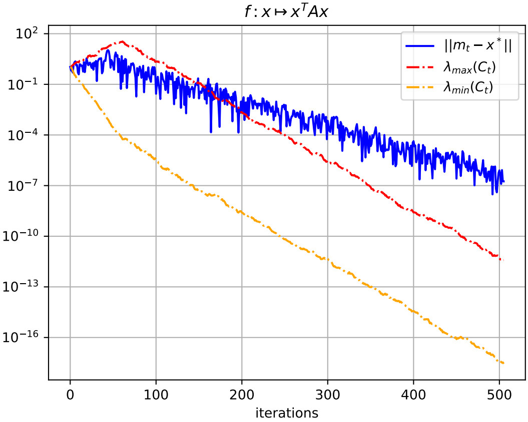
We observe

$$m_t \underset{t\to\infty}{\longrightarrow} x^* \in \arg\min f$$

and

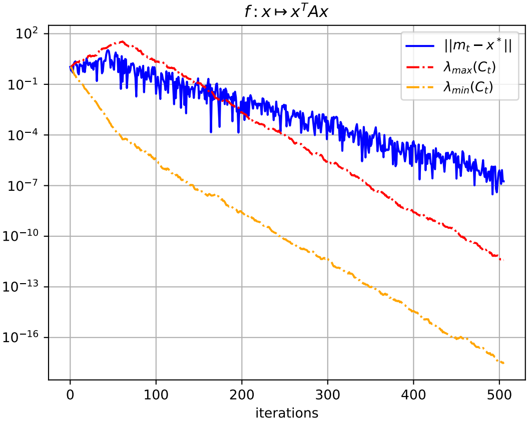
$$C_t \underset{t\to\infty}{\dashrightarrow} H^{-1}$$

Linear convergence



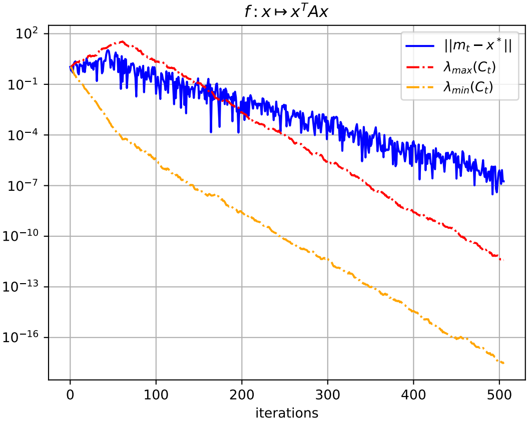
$$\frac{||m_t - x^*||}{||m_0 - x^*||}$$

Linear convergence



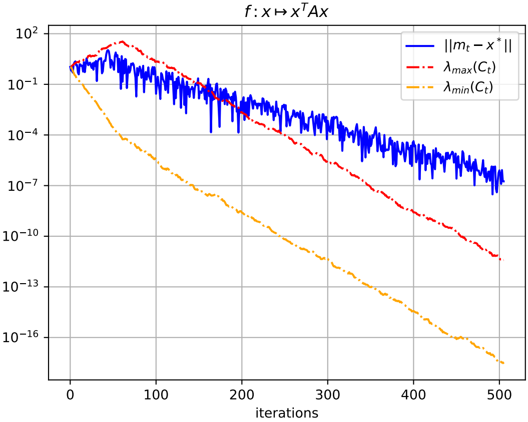
$$\log \frac{\|m_t - x^*\|}{\|m_0 - x^*\|}$$

Linear convergence



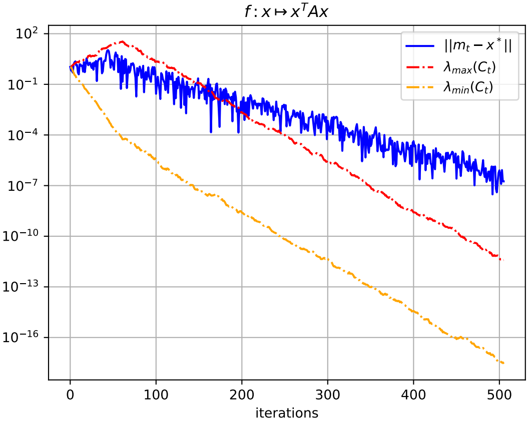
$$\log \frac{\|m_t - x^*\|}{\|m_0 - x^*\|} = -\mathrm{CR} \times t$$

Linear convergence



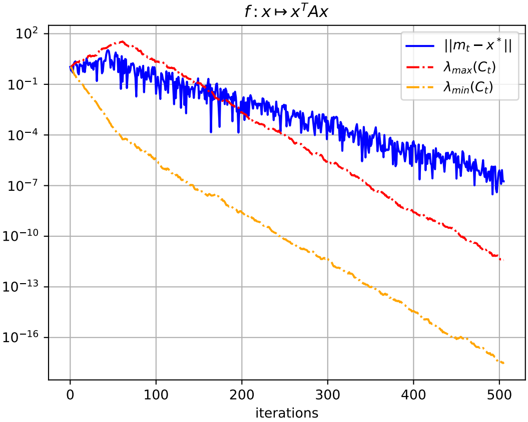
$$\frac{1}{t} \log \frac{\|m_t - x^*\|}{\|m_0 - x^*\|} = -\mathrm{CR}$$

Linear convergence



$$\lim_{t \to \infty} \frac{1}{t} \log \frac{\|m_t - x^*\|}{\|m_0 - x^*\|} = -\mathrm{CR}$$

Linear convergence



$$\lim_{t\to\infty} \frac{1}{t} \log \frac{\|m_t - x^*\|}{\|m_0 - x^*\|} = -\mathrm{CR}$$

Convergence analysis via Markov chains

Markov chains

A **Markov chain** is a random sequence $(\theta_t)_{t\in\mathbb{N}}$ such that

$$\text{Distribution}\left(\theta_{t+1} \mid \theta_0, \ldots, \theta_t\right) = \text{Distribution}\left(\theta_{t+1} \mid \theta_t\right)$$

Presentation of the algorithm

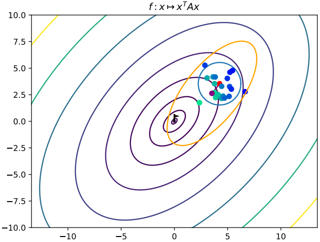
Algorithm 1 CMA-ES

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

Given: $m_0 \in \mathbb{R}^d$, $C_0 \in \mathcal{S}^d_{++}$

For $t = 0, 1, 2, \dots$:

1. $x^1_{t+1}, \dots, x^\lambda_{t+1} \sim \mathcal{N}(m_t, C_t)$
2. $\text{sort } f(x^i_{t+1})$:
   $f\left(x^{1:\lambda}_{t+1}\right) \leqslant \cdots \leqslant f\left(x^{\lambda:\lambda}_{t+1}\right)$
3. $m_{t+1} = \sum_{i=1}^{\mu} w_i x^{i:\lambda}_{t+1}$
4. $C_{t+1} = (1-c)C_t + c\sum_{i=1}^{\mu} w_i \left[x^{i:\lambda}_{t+1} - m_t\right]\left[x^{i:\lambda}_{t+1} - m_t\right]^T$



λ population size

μ parent number

Markov chains

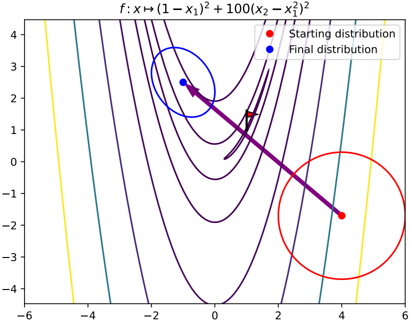
A **Markov chain** is a random sequence $(\theta_t)_{t\in\mathbb{N}}$ such that

$$\text{Distribution}\left(\theta_{t+1} \mid \theta_0, \ldots, \theta_t\right) = \text{Distribution}\left(\theta_{t+1} \mid \theta_t\right)$$

- The Markov chain $(\theta_t)_{t\in\mathbb{N}}$ is **irreducible** if any state is reachable in finite time with positive probability.

Irreducibility of CMA-ES

The Markov chain $(\theta_t)_{t\in\mathbb{N}}$ is **irreducible** if any state is reachable in finite time with positive probability.



Markov chains

A **Markov chain** is a random sequence $(\theta_t)_{t\in\mathbb{N}}$ such that

$$\text{Distribution}\,(\theta_{t+1} \mid \theta_0, \ldots, \theta_t) = \text{Distribution}\,(\theta_{t+1} \mid \theta_t)$$

- The Markov chain $(\theta_t)_{t\in\mathbb{N}}$ is **irreducible** if any state is reachable in finite time with positive probability.
- Then, it admits a **period** $P \geqslant 1$. When $P = 1$, $(\theta_t)_{t\in\mathbb{N}}$ is **aperiodic**.

Markov chains

A **Markov chain** is a random sequence $(\theta_t)_{t \in \mathbb{N}}$ such that

$$\mathrm{Distribution}\left(\theta_{t+1} \mid \theta_0, \ldots, \theta_t\right) = \mathrm{Distribution}\left(\theta_{t+1} \mid \theta_t\right)$$

- The Markov chain $(\theta_t)_{t \in \mathbb{N}}$ is **positive recurrent** if there exists a unique **invariant** probability measure π, i.e.,

$$\theta_t \sim \pi \Rightarrow \theta_{t+1} \sim \pi$$

Markov chains

A **Markov chain** is a random sequence $(\theta_t)_{t\in\mathbb{N}}$ such that

$$\text{Distribution}\left(\theta_{t+1} \mid \theta_0, \ldots, \theta_t\right) = \text{Distribution}\left(\theta_{t+1} \mid \theta_t\right)$$

- The Markov chain $(\theta_t)_{t\in\mathbb{N}}$ is **positive recurrent** if there exists a unique **invariant** probability measure π, i.e.,

$$\theta_t \sim \pi \Rightarrow \theta_{t+1} \sim \pi$$

  (π is a fixed point for $(\theta_t)_{t\in\mathbb{N}}$)

Markov chains

A **Markov chain** is a random sequence $(\theta_t)_{t\in\mathbb{N}}$ such that

$$\text{Distribution}\left(\theta_{t+1} \mid \theta_0, \ldots, \theta_t\right) = \text{Distribution}\left(\theta_{t+1} \mid \theta_t\right)$$

- The Markov chain $(\theta_t)_{t\in\mathbb{N}}$ is **positive recurrent** if there exists a unique **invariant** probability measure π, i.e.,

  $$\theta_t \sim \pi \Rightarrow \theta_{t+1} \sim \pi$$

  (π is a fixed point for $(\theta_t)_{t\in\mathbb{N}}$)
- If the chain is irreducible, aperiodic, positive recurrent, then a **Law of Large Numbers** (LLN) holds

  $$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} f\left(\theta_t\right) = \int f(\theta)\, d\pi(\theta).$$

CMA-ES as a Markov chain

$$\theta_t = (\underbrace{m_t}_{\text{mean}}, \overbrace{C_t}^{\text{covariance matrix}})$$

defines a Markov chain

CMA-ES as a Markov chain

$$\theta_t = (\underbrace{m_t}_{\text{mean}}, \overbrace{C_t}^{\text{covariance matrix}})$$

defines a Markov chain

Question: Could we use the **LLN** for Markov chains to prove the **linear convergence** of CMA-ES?

Invariant measure for CMA-ES?

If π is an invariant measure of $(m_t, C_t)_{t\in\mathbb{N}}$

Invariant measure for CMA-ES?

If π is an invariant measure of $(m_t, C_t)_{t\in\mathbb{N}}$

$$(m_t, C_t) \sim \pi \Rightarrow (m_{t+1}, C_{t+1}) \sim \pi$$

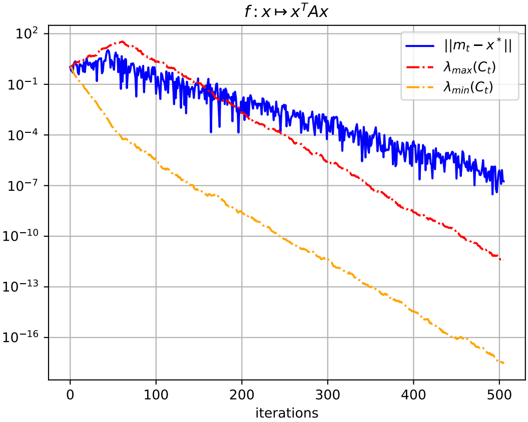
Invariant measure for CMA-ES?

If π is an invariant measure of $(m_t, C_t)_{t\in\mathbb{N}}$

$$(m_t, C_t) \sim \pi \Rightarrow (m_{t+1}, C_{t+1}) \sim \pi$$

Not possible if $m_t \to x^*$ and $C_t \to 0$.

Linear convergence



$$\lim_{t \to \infty} \frac{1}{t} \log \frac{\|m_t - x^*\|}{\|m_0 - x^*\|} = -\mathrm{CR}$$

Normalization

$\|m_t - x^*\|$ and $\lambda_{\min}(C_t) \to 0$

Normalization

$\|m_t - x^*\|$ and $\lambda_{\min}(C_t) \to 0$

$$z_t \stackrel{\text{def}}{=} \frac{m_t - x^*}{\sqrt{\lambda_{\min}(C_t)}} \qquad \Sigma_t \stackrel{\text{def}}{=} \frac{C_t}{\lambda_{\min}(C_t)}$$

Normalization

$\|m_t - x^*\|$ and $\lambda_{\min}(C_t) \to 0$

$$z_t \stackrel{\text{def}}{=} \frac{m_t - x^*}{\sqrt{\lambda_{\min}(C_t)}} \qquad \Sigma_t \stackrel{\text{def}}{=} \frac{C_t}{\lambda_{\min}(C_t)}$$

The sequence $(z_t, \Sigma_t)_{t\in\mathbb{N}}$ might eventually be **stationary**

Normalization

$\|m_t - x^*\|$ and $\lambda_{\min}(C_t) \to 0$

$$z_t \stackrel{\text{def}}{=} \frac{m_t - x^*}{\sqrt{\lambda_{\min}(C_t)}} \qquad \Sigma_t \stackrel{\text{def}}{=} \frac{C_t}{\lambda_{\min}(C_t)}$$

The sequence $(z_t, \Sigma_t)_{t \in \mathbb{N}}$ might eventually be stationary

Proposition (Normalized Markov chain)

$$(z_t, \Sigma_t)_{t \in \mathbb{N}}$$

is a Markov chain.

Normalization

$\|m_t - x^*\|$ and $\lambda_{\min}(C_t) \to 0$

$$z_t \overset{\text{def}}{=} \frac{m_t - x^*}{\sqrt{\lambda_{\min}(C_t)}} \qquad \Sigma_t \overset{\text{def}}{=} \frac{C_t}{\lambda_{\min}(C_t)}$$

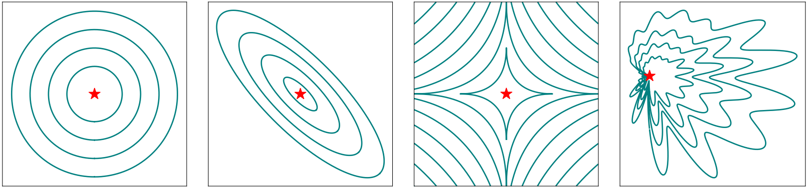
The sequence $(z_t, \Sigma_t)_{t\in\mathbb{N}}$ might eventually be stationary

Proposition (Normalized Markov chain)

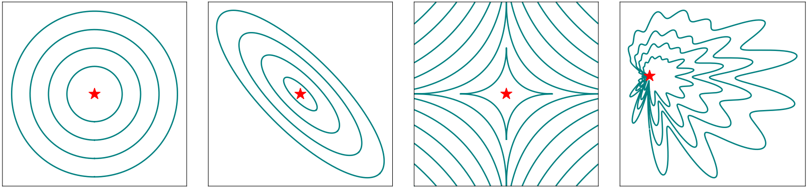
$$(z_t, \Sigma_t)_{t\in\mathbb{N}}$$

is a Markov chain. (if f is **scaling-invariant***)*

Scaling-invariant functions



Scaling-invariant functions



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

Algorithm

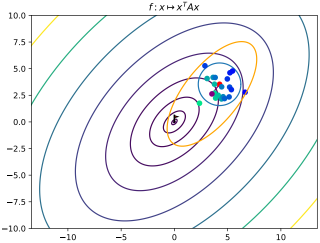
Algorithm 1 CMA-ES

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

Given: $m_0 \in \mathbb{R}^d$, $C_0 \in \mathcal{S}^d_{++}$

For $t = 0, 1, 2, \ldots$:

1. $x^1_{t+1}, \ldots, x^\lambda_{t+1} \sim \mathcal{N}(m_t, C_t)$
2. $\text{sort } f(x^i_{t+1})$:
   $f\left(x^{1:\lambda}_{t+1}\right) \leqslant \cdots \leqslant f\left(x^{\lambda:\lambda}_{t+1}\right)$
3. $m_{t+1} = \sum_{i=1}^{\mu} w_i x^{i:\lambda}_{t+1}$
4. $C_{t+1} = (1-c)C_t + c\sum_{i=1}^{\mu} w_i \left[x^{i:\lambda}_{t+1} - m_t\right]\left[m^{i:\lambda}_{t+1} - m_t\right]^T$



λ population size

μ parent number

Algorithm

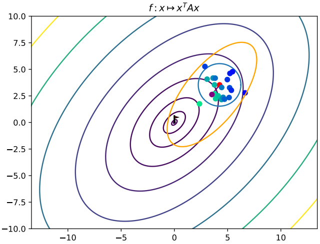
Algorithm 1 normalized CMA-ES

Goal: Converge to π

Given: $z_0 \in \mathbb{R}^d$, $\Sigma_0 \in \mathcal{S}^d_{++}$

For $t = 0, 1, 2, \ldots$:

1. $z^1_{t+1}, \ldots, z^\lambda_{t+1} \sim \mathcal{N}(z_t, \Sigma_t)$
2. sort $f(z^i_{t+1})$:
   $f\left(z^{1:\lambda}_{t+1}\right) \leqslant \cdots \leqslant f\left(z^{\lambda:\lambda}_{t+1}\right)$
3. $\tilde{z}_{t+1} = \sum_{i=1}^{\mu} w_i z^{i:\lambda}_{t+1}$
4. $\tilde{\Sigma}_{t+1} = (1-c)\Sigma_t + c\sum_{i=1}^{\mu} w_i \left[z^{i:\lambda}_{t+1} - z_t\right]\left[z^{i:\lambda}_{t+1} - z_t\right]^T$



λ population size

μ parent number

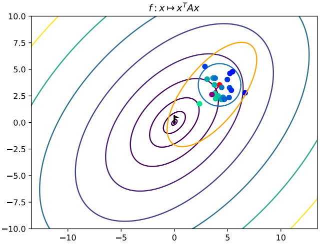
Algorithm

Algorithm 1 normalized CMA-ES

Goal: Converge to π

Given: $z_0 \in \mathbb{R}^d$, $\Sigma_0 \in \mathcal{S}^d_{++}$

For $t = 0, 1, 2, \ldots$:



1. $z^1_{t+1}, \ldots, z^\lambda_{t+1} \sim \mathcal{N}(z_t, \Sigma_t)$
2. sort $f(z^i_{t+1})$:
   $f\left(z^{1:\lambda}_{t+1}\right) \leqslant \cdots \leqslant f\left(z^{\lambda:\lambda}_{t+1}\right)$
3. $\tilde{z}_{t+1} = \sum_{i=1}^{\mu} w_i z^{i:\lambda}_{t+1}$
4. $\tilde{\Sigma}_{t+1} = (1-c)\Sigma_t + c\sum_{i=1}^{\mu} w_i \left[z^{i:\lambda}_{t+1} - z_t\right]\left[z^{i:\lambda}_{t+1} - z_t\right]^T$
5. $z_{t+1} = \tilde{z}_{t+1}/\lambda^{1/2}_{\min}(\tilde{\Sigma}_{t+1})$
   $\Sigma_{t+1} = \tilde{\Sigma}_{t+1}/\lambda_{\min}(\tilde{\Sigma}_{t+1})$

λ population size

μ parent number

Sketch of proof

1. $(\theta_t)_{t\in\mathbb{N}} = (z_t, \Sigma_t)_{t\in\mathbb{N}}$ is irreducible and aperiodic ;

Sketch of proof

1. $(\theta_t)_{t\in\mathbb{N}} = (z_t, \Sigma_t)_{t\in\mathbb{N}}$ is irreducible and aperiodic ;
2. $(\theta_t)_{t\in\mathbb{N}}$ is positive recurrent ;

Sketch of proof

1. $(\theta_t)_{t\in\mathbb{N}} = (z_t, \Sigma_t)_{t\in\mathbb{N}}$ is irreducible and aperiodic ;
2. $(\theta_t)_{t\in\mathbb{N}}$ is positive recurrent ;
3. By 1. and 2., it follows a LLN

Sketch of proof

1. $(\theta_t)_{t\in\mathbb{N}} = (z_t, \Sigma_t)_{t\in\mathbb{N}}$ is irreducible and aperiodic ;
2. $(\theta_t)_{t\in\mathbb{N}}$ is positive recurrent ;
3. By 1. and 2., it follows a LLN and

$$\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \frac{1}{T}\sum_{t=0}^{T-1}\log\frac{\|z_{t+1}\|}{\|z_t\|} - \frac{1}{2}\log\lambda_{\min}(\tilde{\Sigma}_{t+1})$$

Sufficient conditions for irreducibility and aperiodicity

Suppose

$$\theta_{t+1} = F(\theta_t, u_{t+1})$$

Sufficient conditions for irreducibility and aperiodicity

Suppose

$$\theta_{t+1} = F(\theta_t, u_{t+1})$$

where $u_{t+1} \sim p_{\theta_t}$.

Sufficient conditions for irreducibility and aperiodicity

Suppose

$$\theta_{t+1} = F(\theta_t, u_{t+1})$$

where $u_{t+1} \sim p_{\theta_t}$.

If there exists $\theta^* \in \Theta$ and $u^* \in U$

Sufficient conditions for irreducibility and aperiodicity

Suppose

$$\theta_{t+1} = F(\theta_t, u_{t+1})$$

where $u_{t+1} \sim p_{\theta_t}$.

If there exists $\theta^* \in \Theta$ and $u^* \in U$ such that

- θ^* can be reached from any starting state of Θ ;

Sufficient conditions for irreducibility and aperiodicity

Suppose

$$\theta_{t+1} = F(\theta_t, u_{t+1})$$

where $u_{t+1} \sim p_{\theta_t}$.

If there exists $\theta^* \in \Theta$ and $u^* \in U$ such that

- θ^* can be reached from any starting state of Θ ;
- $p_{\theta^*}(u^*) > 0$;

Sufficient conditions for irreducibility and aperiodicity

Suppose

$$\theta_{t+1} = F(\theta_t, u_{t+1})$$

where $u_{t+1} \sim p_{\theta_t}$.

If there exists $\theta^* \in \Theta$ and $u^* \in U$ such that

- θ^* can be reached from any starting state of Θ ;
- $p_{\theta^*}(u^*) > 0$;
- $\mathrm{rank}\, \partial_u F(\theta^*, u^*) = \dim \Theta$;

Sufficient conditions for irreducibility and aperiodicity

Suppose

$$\theta_{t+1} = F(\theta_t, u_{t+1})$$

where $u_{t+1} \sim p_{\theta_t}$.

If there exists $\theta^* \in \Theta$ and $u^* \in U$ such that

- θ^* can be reached from any starting state of Θ ;
- $p_{\theta^*}(u^*) > 0$;
- $\mathrm{rank}\, \partial_u F(\theta^*, u^*) = \dim \Theta$;

then $(\theta_t)_{t \in \mathbb{N}}$ is irreducible and aperiodic.

Proposition

Under assumptions on f, $\theta^* = (0, I_d)$ satisfies the previous conditions.

Proposition

Under assumptions on f, $\theta^* = (0, I_d)$ satisfies the previous conditions.

Corollary

Then

$$(z_t, \Sigma_t)_{t\in\mathbb{N}}$$

is **irreducible** *and* **aperiodic**.

Ergodicity of the normalized chain

$$(z_t, \Sigma_t)_{t \in \mathbb{N}}$$

is **positive recurrent** if

Ergodicity of the normalized chain

$$(z_t, \Sigma_t)_{t\in\mathbb{N}}$$

is positive recurrent if

- it is **irreducible** and **aperiodic**

Ergodicity of the normalized chain

$$(z_t, \Sigma_t)_{t \in \mathbb{N}}$$

is positive recurrent if

- it is irreducible and aperiodic
- there exists a **drift function** $V\colon \Theta \to [0, +\infty]$ such that

Ergodicity of the normalized chain

$$(z_t, \Sigma_t)_{t \in \mathbb{N}}$$

is positive recurrent if

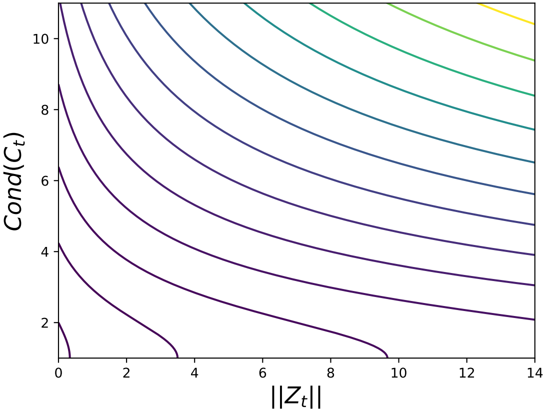
- it is irreducible and aperiodic
- there exists a drift function $V \colon \Theta \to [0, +\infty]$ such that

$$\mathbb{E}_t \left[V \left(z_{t+1}, \Sigma_{t+1} \right) \right] \leqslant (1 - \varepsilon) V \left(z_t, \Sigma_t \right)$$

outside of a compact $\mathcal{K} \subset \Theta$.

Drift condition

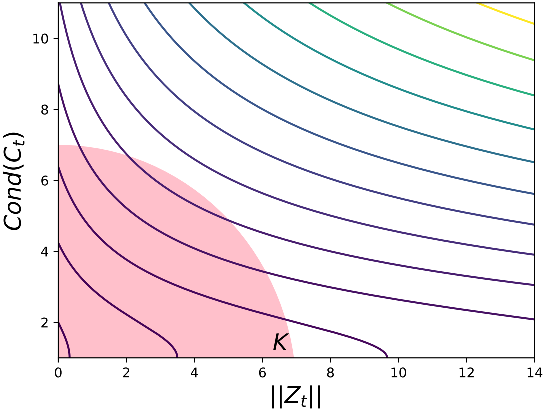
$$\mathbb{E}_t \left[V\left(z_{t+1}, \Sigma_{t+1}\right)\right] \leqslant (1-\varepsilon) V\left(z_t, \Sigma_t\right)$$



Drift condition

$$\mathbb{E}_t \left[V\left(z_{t+1}, \Sigma_{t+1}\right)\right] \leqslant (1-\varepsilon) V\left(z_t, \Sigma_t\right)$$

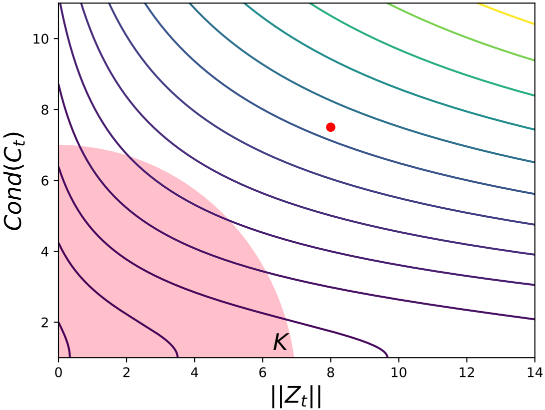
outside of a compact K



Drift condition

$$\mathbb{E}_t \left[V\left(z_{t+1}, \Sigma_{t+1}\right)\right] \leqslant (1-\varepsilon) V\left(z_t, \Sigma_t\right)$$

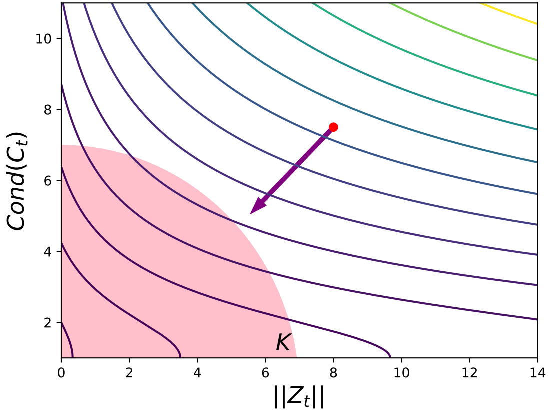
outside of a compact K



Drift condition

$$\mathbb{E}_t \left[V\left(z_{t+1}, \Sigma_{t+1}\right)\right] \leqslant (1-\varepsilon) V\left(z_t, \Sigma_t\right)$$

outside of a compact K



Theorem (Drift condition for the normalized chain)

When minimizing a **spherical** *function* $f : x \mapsto g\left(x^T x\right)$ *then* $(z_t, \Sigma_t)_{t\in\mathbb{N}}$ *satisfies a drift condition with*

$$V(z, \Sigma) = \alpha \times \frac{\|\sqrt{\Sigma} z\|^2}{\lambda_{\max}(\Sigma)} + \beta \times \|\Sigma\|$$

Theorem (Drift condition for the normalized chain)

When minimizing a **spherical** *function* $f : x \mapsto g\left(x^T x\right)$ *then* $(z_t, \Sigma_t)_{t\in\mathbb{N}}$ *satisfies a drift condition with*

$$V(z,\Sigma) = \alpha \times \frac{\|\sqrt{\Sigma} z\|^2}{\lambda_{\max}(\Sigma)} + \beta \times \|\Sigma\|$$

This can be generalized to **ellipsoid** functions $f(x) = g(x^T H x)$ using the **affine-invariance** of CMA-ES.

Affine-Invariance

$$
\begin{array}{ccc}
(m_0, C_0) & \xrightarrow{\min f(x)} & (m_1, C_1) \\
\Psi \big\downarrow & & \big\uparrow \Psi^{-1} \\
(m_0', C_0') & \xrightarrow{\min f(Bx + b)} & (m_1', C_1')
\end{array}
$$

Convergence

Theorem

When $f = g(x^T Hx)$, *then*

$$\lim_{T\to\infty} \frac{1}{T} \log \frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \lim_{t\to\infty} \mathbb{E}\left[\log \frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|}\right] = -\mathrm{CR}$$

and

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

Thank you!