Convergence analysis of CMA-ES

ISMP 2024
Stochastic DFO 1

Armand Gissler
Thursday 25th July, 2024

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

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

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

Evolution Strategies (ES)

Given $\theta_t \in \Theta$:

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

Evolution Strategies (ES)

Given $\theta_t \in \Theta$:

1. Sample a population $x_{t+1}^1, \ldots, x_{t+1}^\lambda \sim p_{\theta_t}$

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

Evolution Strategies (ES)

Given $\theta_t \in \Theta$:

1. Sample a population $x_{t+1}^1, \ldots, x_{t+1}^\lambda \sim p_{\theta_t}$
2. Rank
$$f(x_{t+1}^{1:\lambda}) \leqslant \cdots \leqslant f(x_{t+1}^{\lambda:\lambda})$$

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

Evolution Strategies (ES)

Given $\theta_t \in \Theta$:

1. Sample a population $x_{t+1}^1, \ldots, x_{t+1}^\lambda \sim p_{\theta_t}$
2. Rank
$$f(x_{t+1}^{1:\lambda}) \leqslant \cdots \leqslant f(x_{t+1}^{\lambda:\lambda})$$
3. Update θ_{t+1}.

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

Covariance Matrix Adaptation-ES (CMA-ES)

Given $\theta_t = (m_t, \sigma_t, \mathbf{C}_t) \in \mathbb{R}^d \times \mathbb{R}_{>0} \times \mathcal{S}^d_{++}$:

1. Sample a population $x^1_{t+1}, \ldots, x^\lambda_{t+1} \sim p_{\theta_t}$
2. Rank
$$f(x^{1:\lambda}_{t+1}) \leqslant \cdots \leqslant f(x^{\lambda:\lambda}_{t+1})$$
3. Update θ_{t+1}.

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

Covariance Matrix Adaptation-ES (CMA-ES)

Given $\theta_t = (m_t, \sigma_t, \mathbf{C}_t) \in \mathbb{R}^d \times \mathbb{R}_{>0} \times \mathcal{S}^d_{++}$:

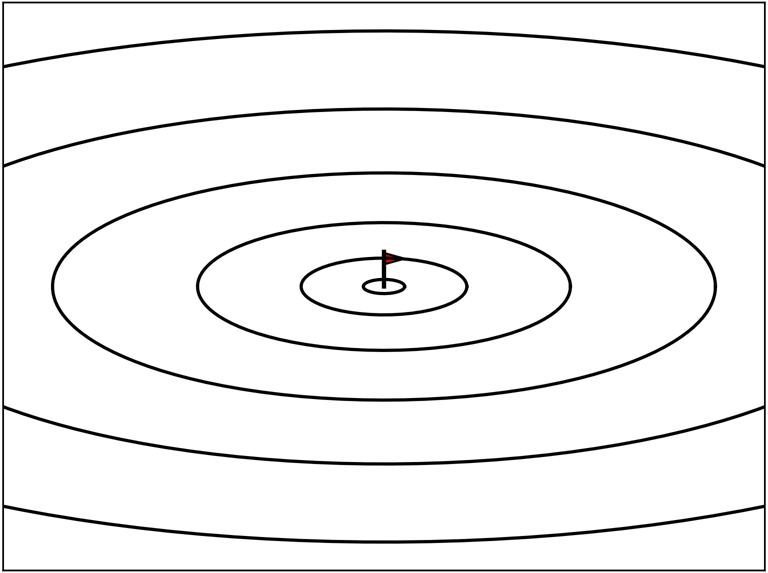
1. Sample a population $x^1_{t+1}, \ldots, x^\lambda_{t+1} \sim p_{\theta_t} = \mathcal{N}(m_t, \sigma_t^2 \mathbf{C}_t)$
2. Rank
$$f(x^{1:\lambda}_{t+1}) \leqslant \cdots \leqslant f(x^{\lambda:\lambda}_{t+1})$$
3. Update θ_{t+1}.

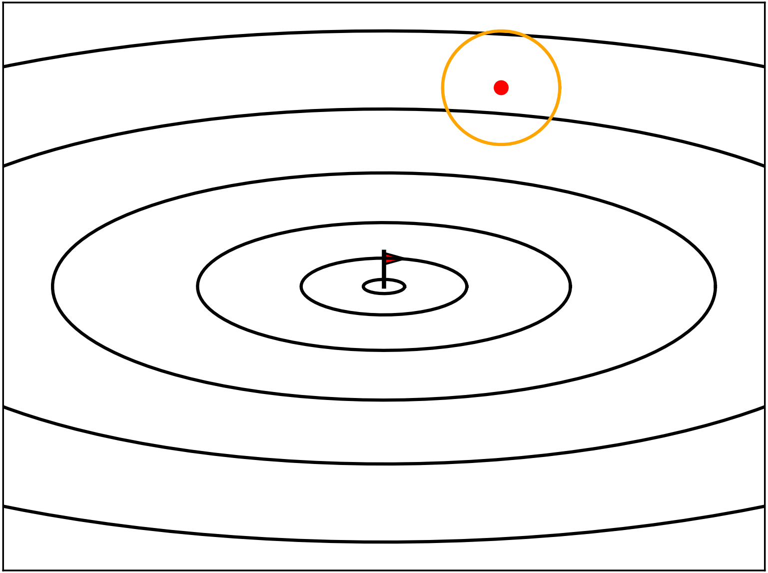
$$\text{Find } x^* \in \operatorname*{Arg\,min}_{x \in \mathbb{R}^d} f(x) \tag{P}$$

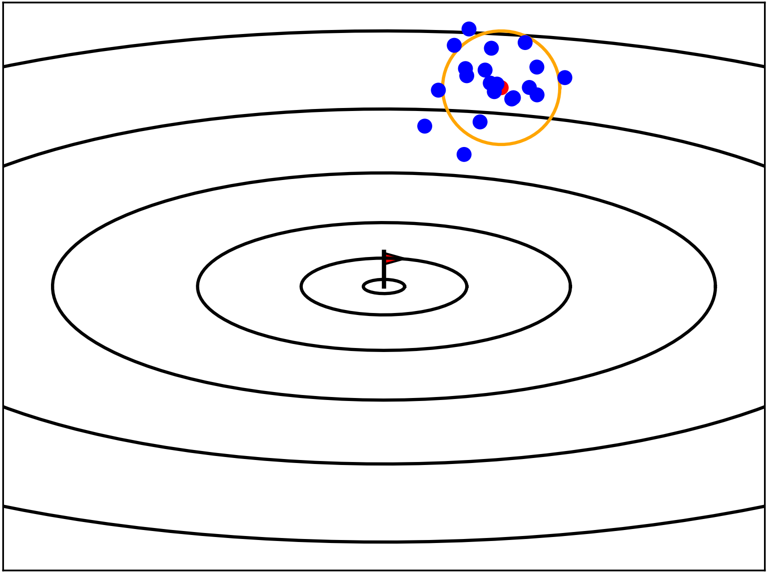
Covariance Matrix Adaptation-ES (CMA-ES)

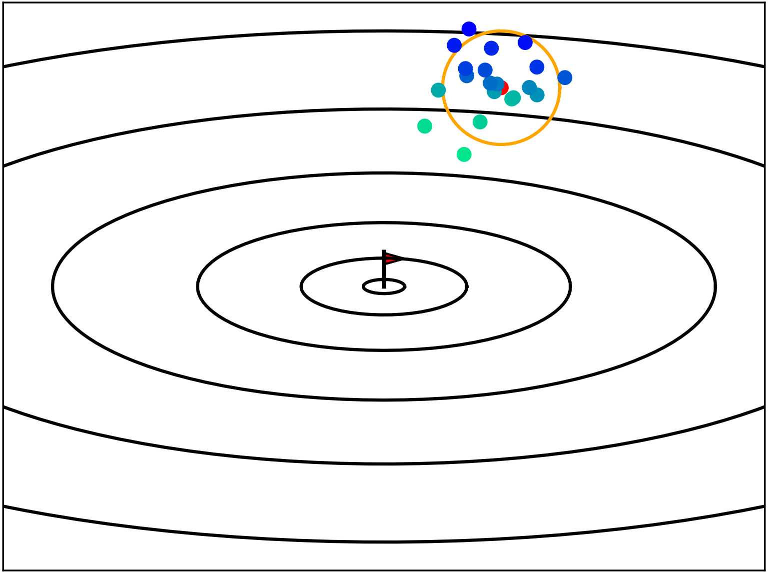
Given $\theta_t = (m_t, \sigma_t, \mathbf{C}_t) \in \mathbb{R}^d \times \mathbb{R}_{>0} \times \mathcal{S}^d_{++}$:

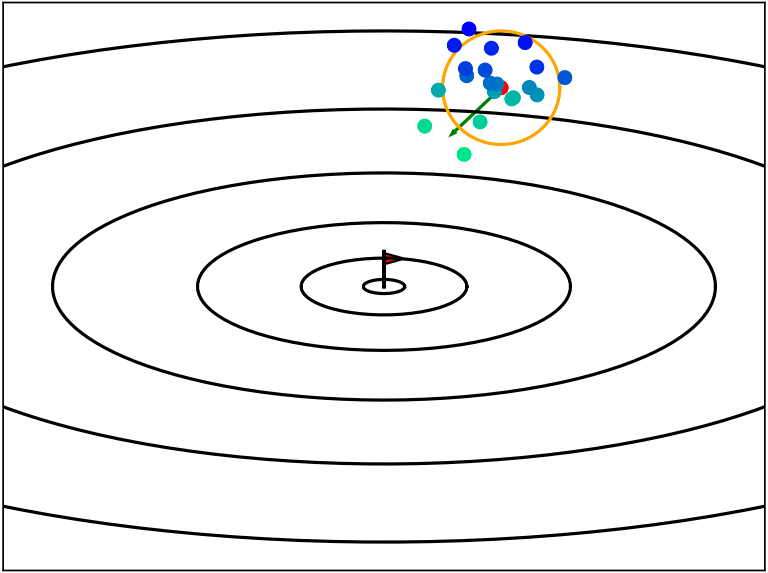
1. Sample a population $x^1_{t+1}, \ldots, x^\lambda_{t+1} \sim p_{\theta_t} = \mathcal{N}(m_t, \sigma_t^2 \mathbf{C}_t)$
2. Rank
$$f(x^{1:\lambda}_{t+1}) \leqslant \cdots \leqslant f(x^{\lambda:\lambda}_{t+1})$$
3. Update $\theta_{t+1} = (m_{t+1}, \sigma_{t+1}, \mathbf{C}_{t+1})$.

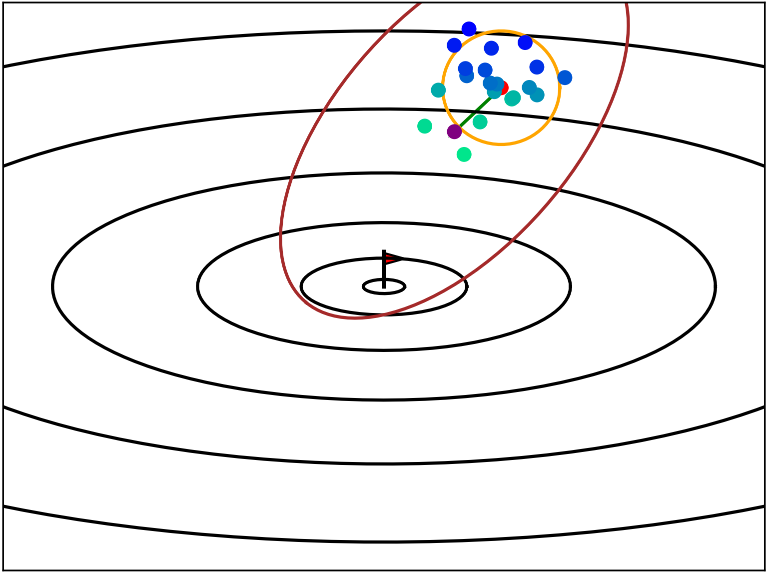












Mean update:

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

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:

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

Step-size adaptation:

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

$$= \sigma_t \times \exp\left(\frac{1}{d_\sigma}\left(\frac{\|{\sigma_t}^{-1}{\mathbf{C}_t}^{-1/2}(m_{t+1} - m_t)\|}{\|\text{weights}\|\mathbb{E}\|\mathcal{N}\|} - 1\right)\right)$$

Covariance matrix adaptation:

$$\mathbf{C}_{t+1} = \text{Positive combination}\left(\mathbf{C}_t, \text{Average}\left[\overleftrightarrow{(x_{t+1}^{i:\lambda} - m_t)}\right]\right)$$

Covariance matrix adaptation:

$$\mathbf{C}_{t+1} = \text{Positive combination}\left(\mathbf{C}_t, \text{Average}\left[\overleftrightarrow{(x_{t+1}^{i:\lambda} - m_t)}\right]\right)$$

$$= (1 - c_\mu)\mathbf{C}_t$$

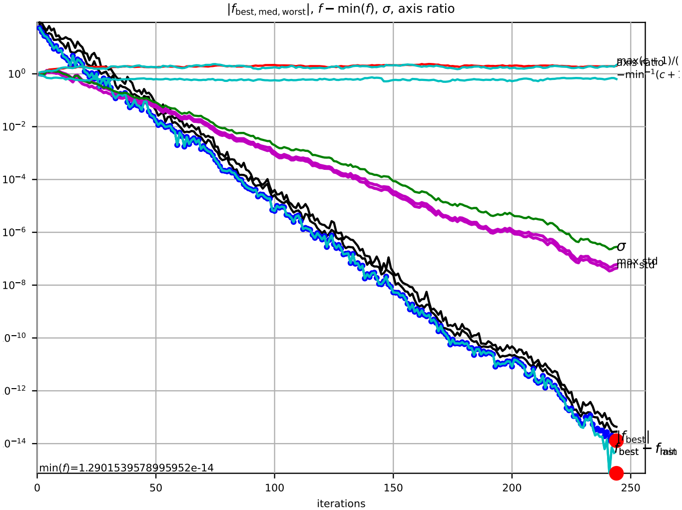
Covariance matrix adaptation:

$$\mathbf{C}_{t+1} = \text{Positive combination}\left(\mathbf{C}_t, \text{Average}\left[\overleftrightarrow{(x_{t+1}^{i:\lambda} - m_t)}\right]\right)$$

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

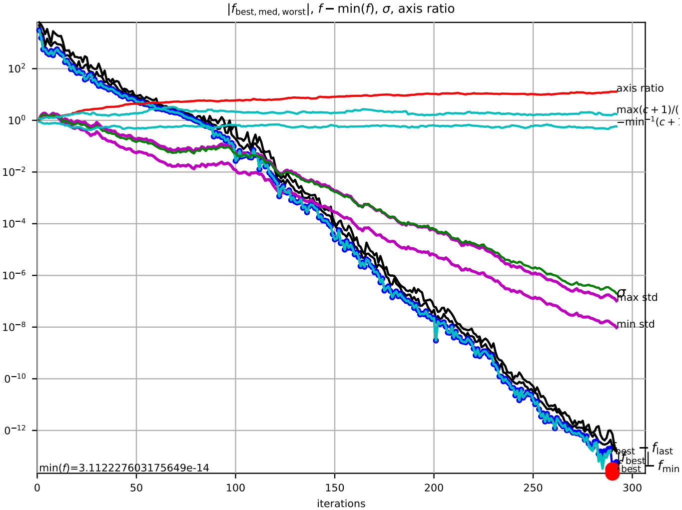
$$f(x) = \frac{1}{2} x^\top \mathbf{H} x$$

$$\mathrm{Cond}(\mathbf{H}) = 10^0$$



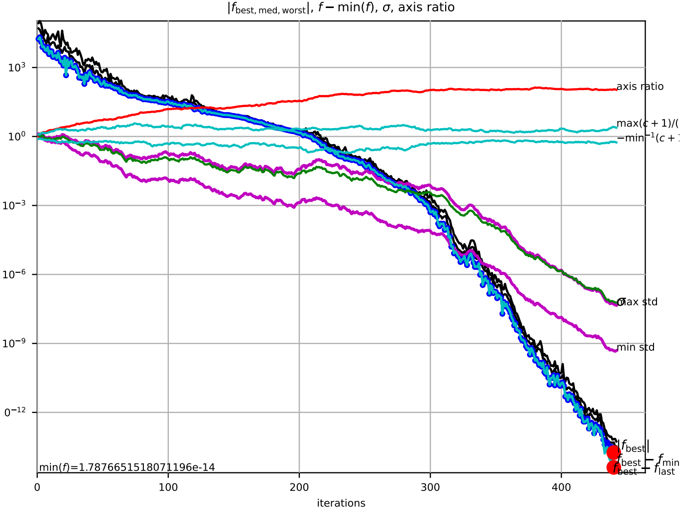
$$f(x) = \frac{1}{2} x^\top \mathbf{H} x$$

$$\mathrm{Cond}(\mathbf{H}) = 10^2$$



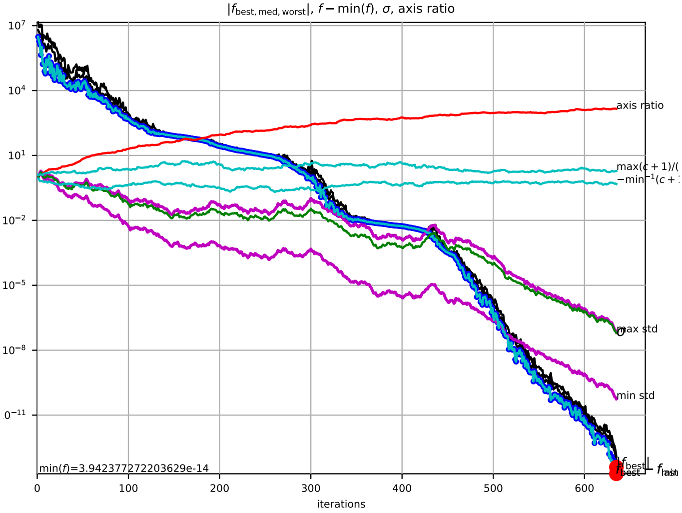
$$f(x) = \frac{1}{2} x^\top \mathbf{H} x$$

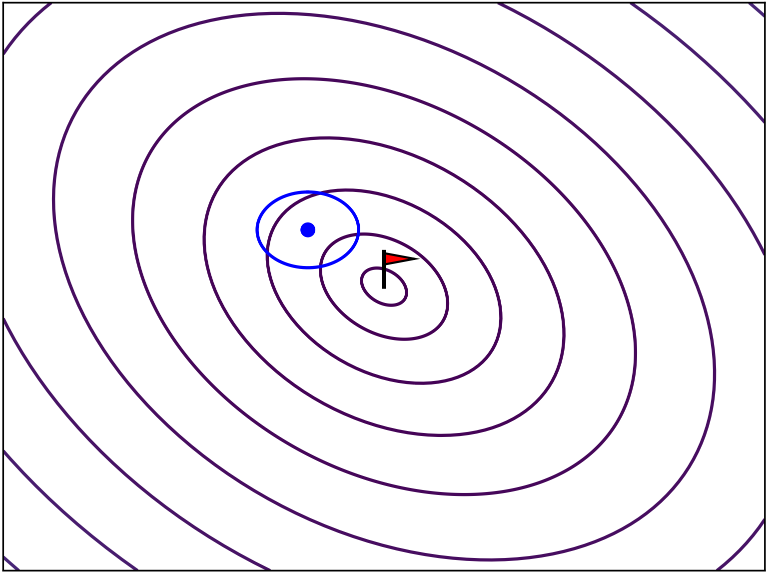
$$\mathrm{Cond}(\mathbf{H}) = 10^4$$

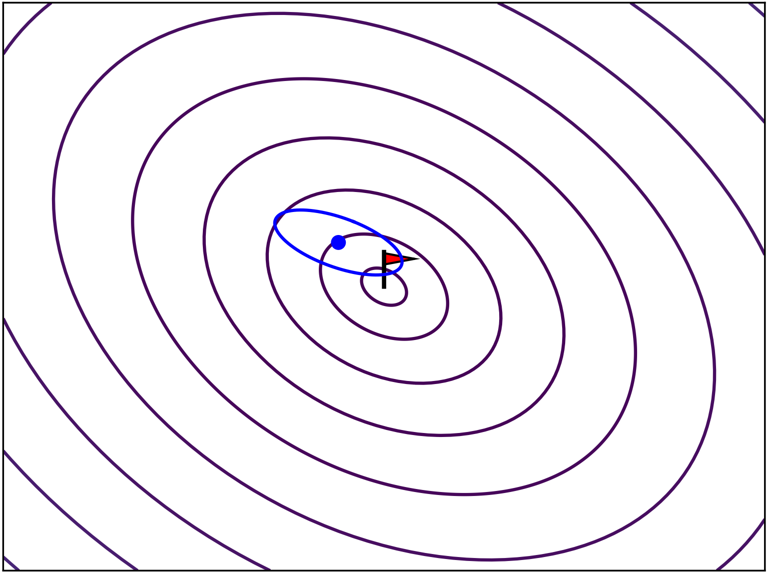


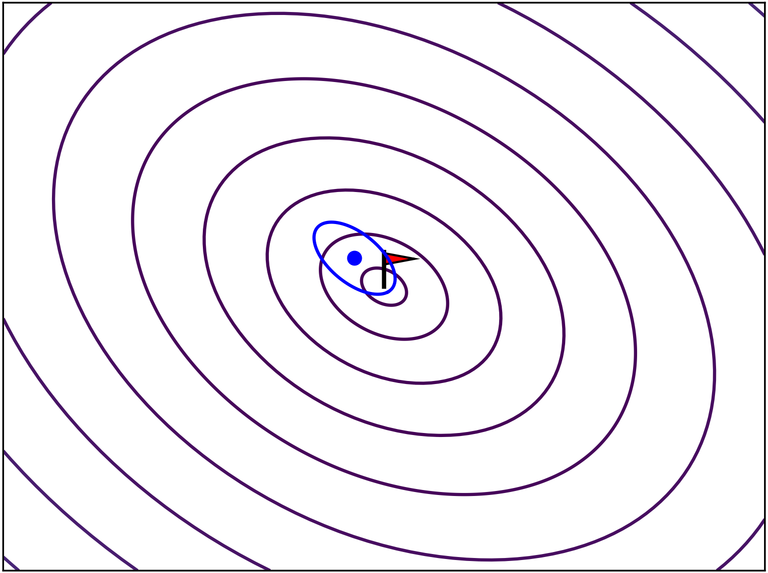
$$f(x) = \frac{1}{2} x^\top \mathbf{H} x$$

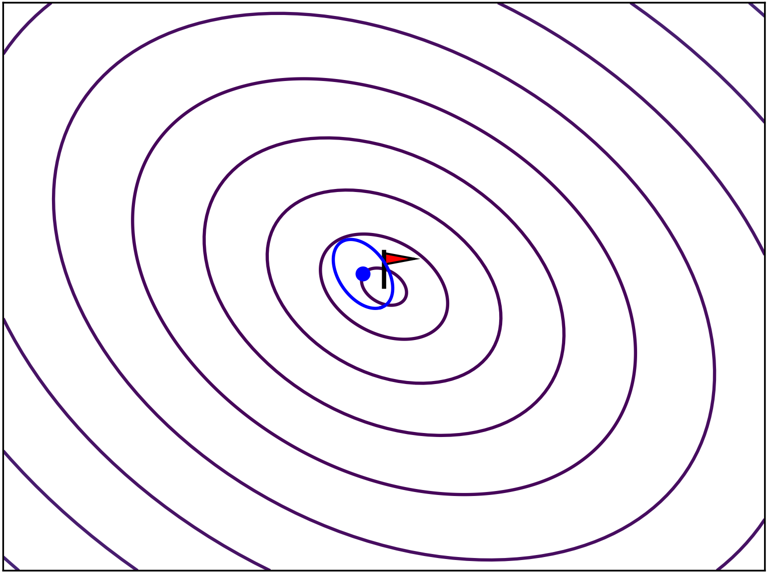
$$\mathrm{Cond}(\mathbf{H}) = 10^6$$

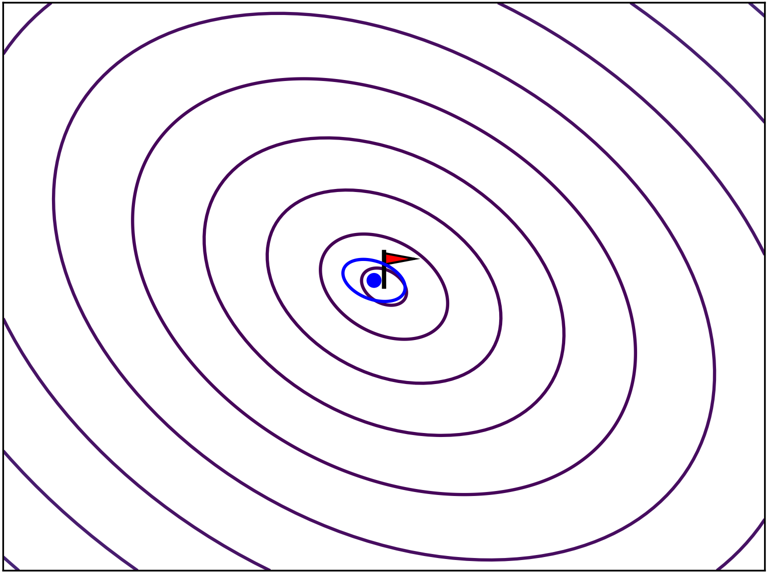


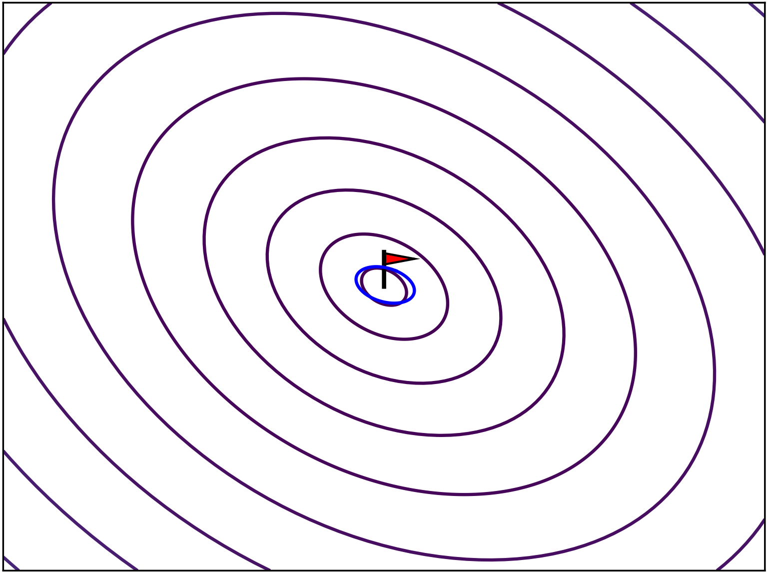


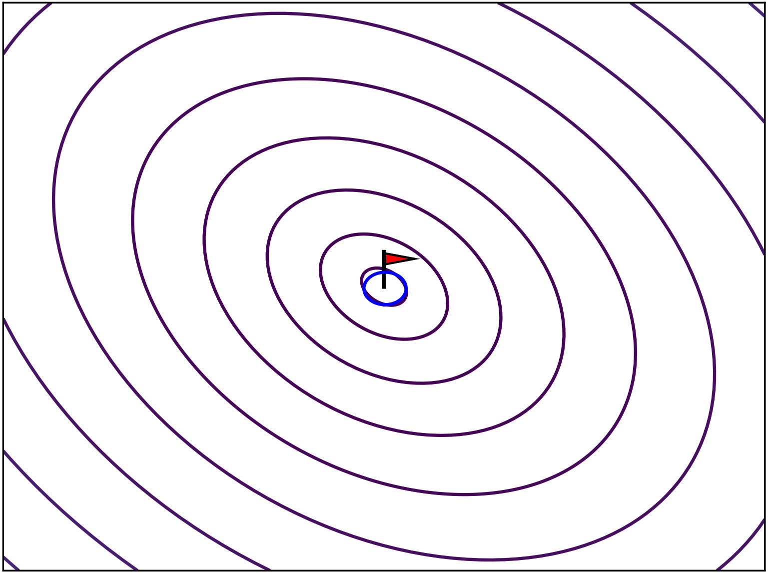


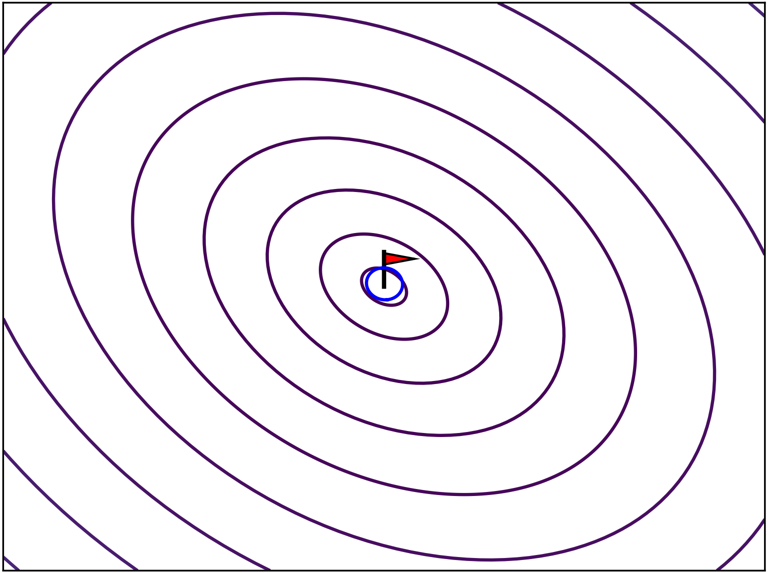


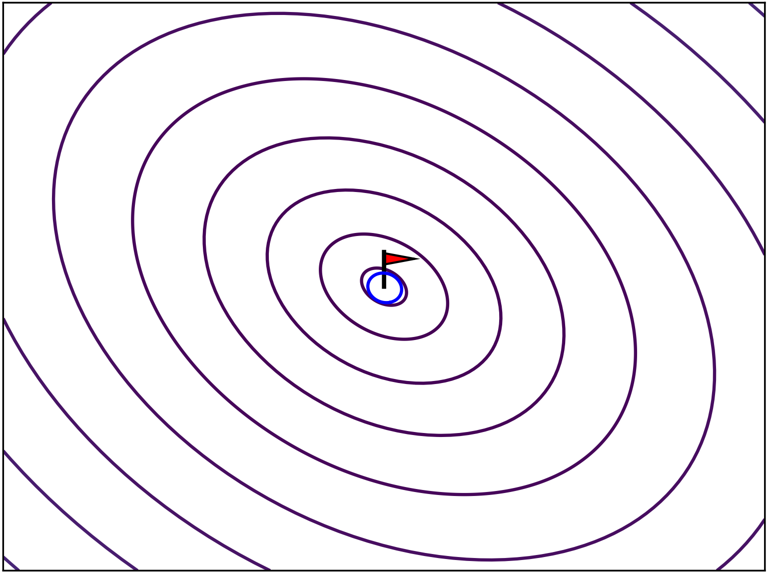


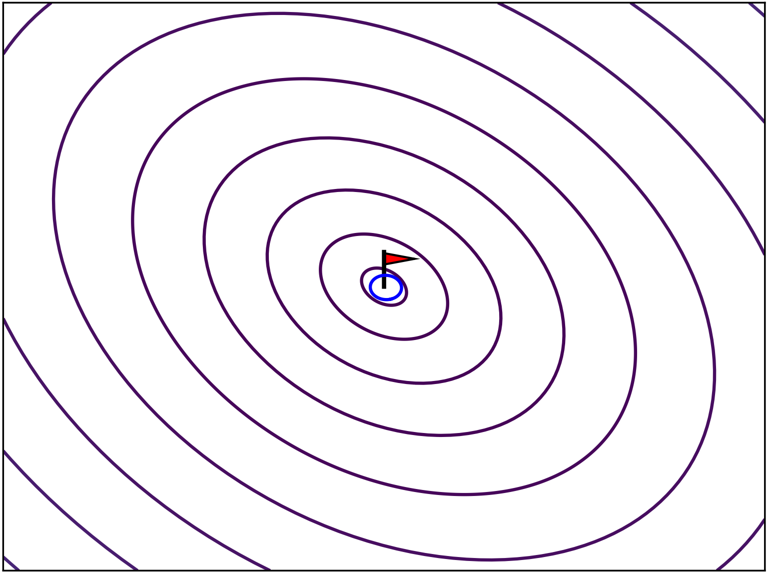


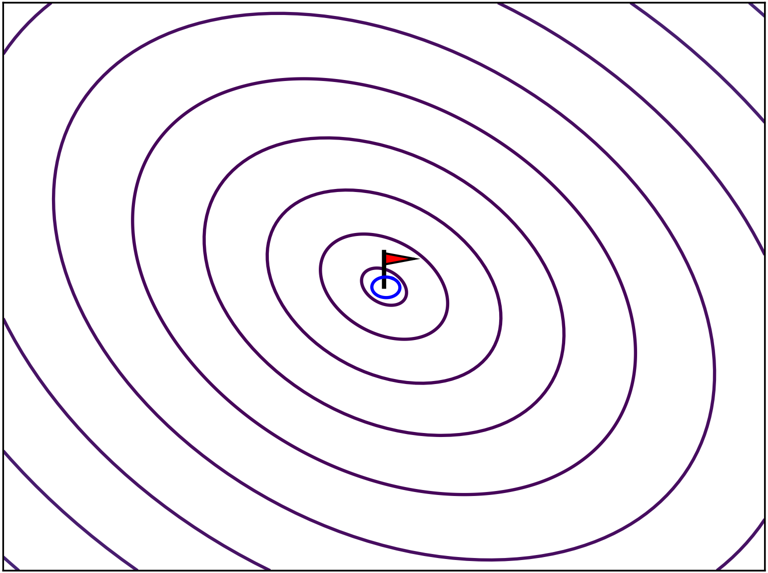


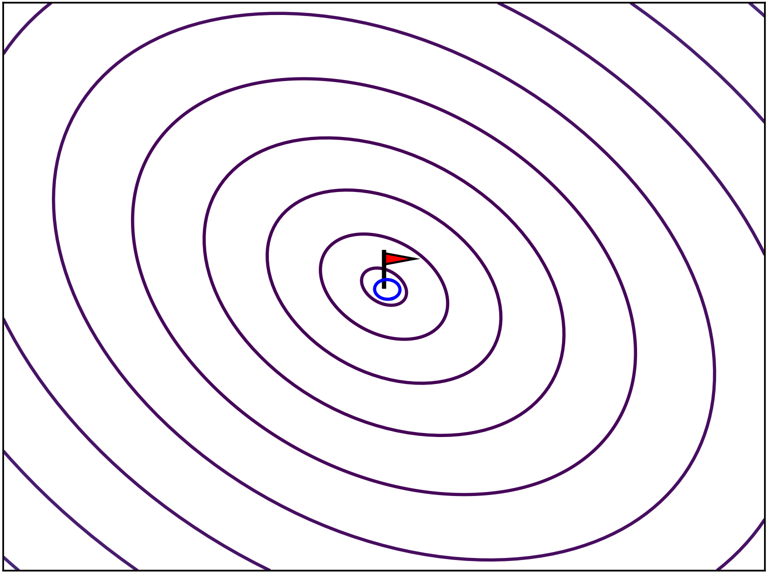


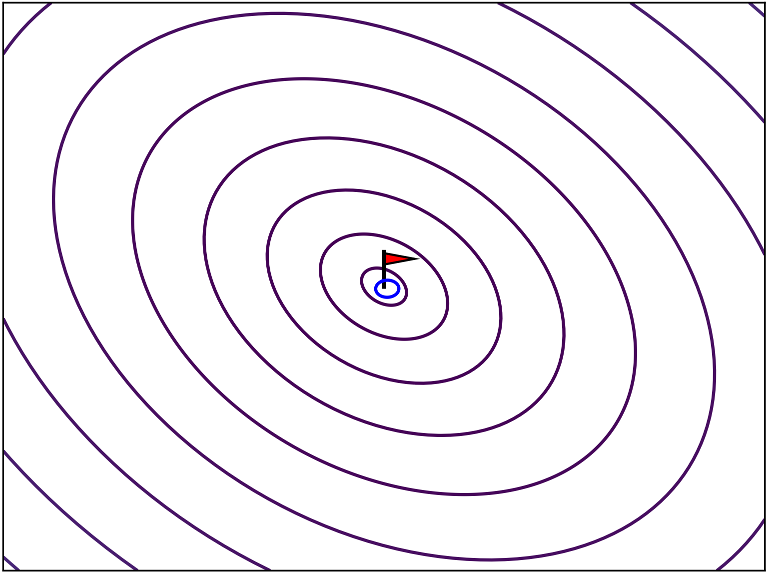


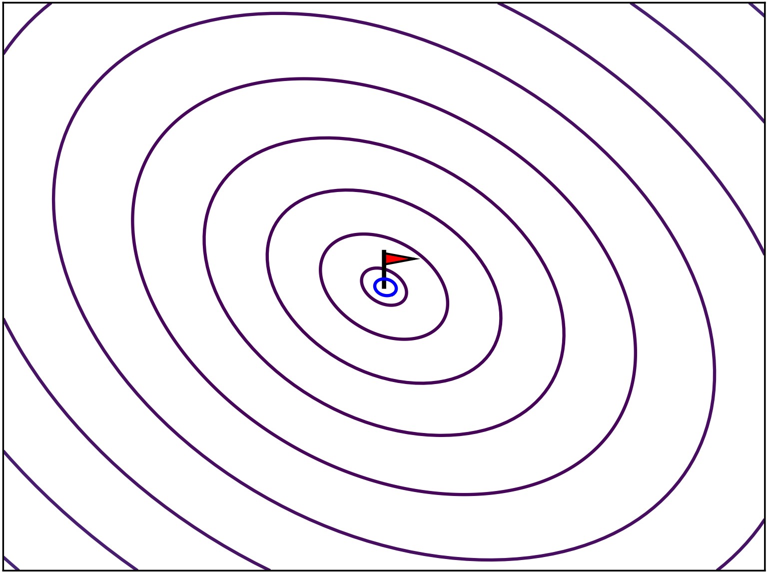


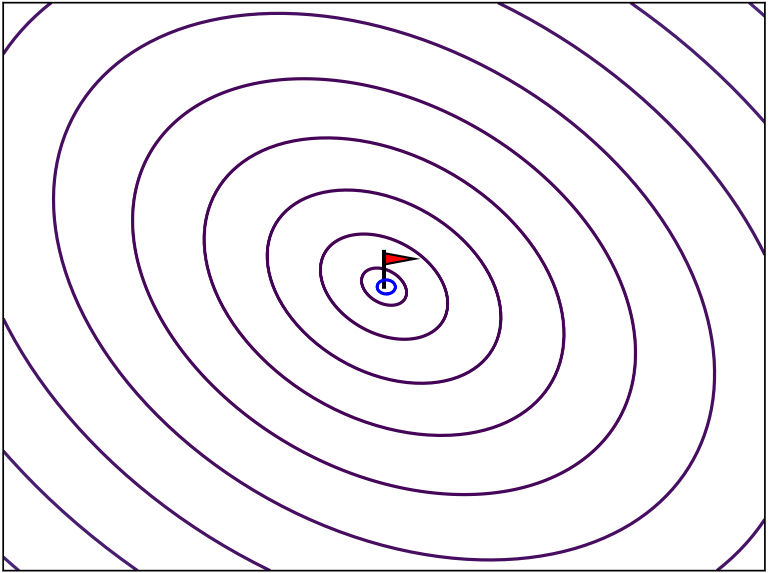


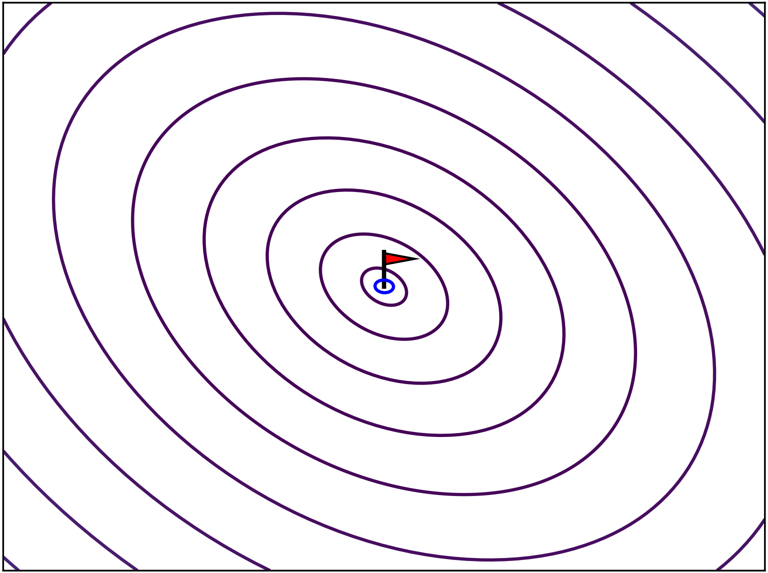


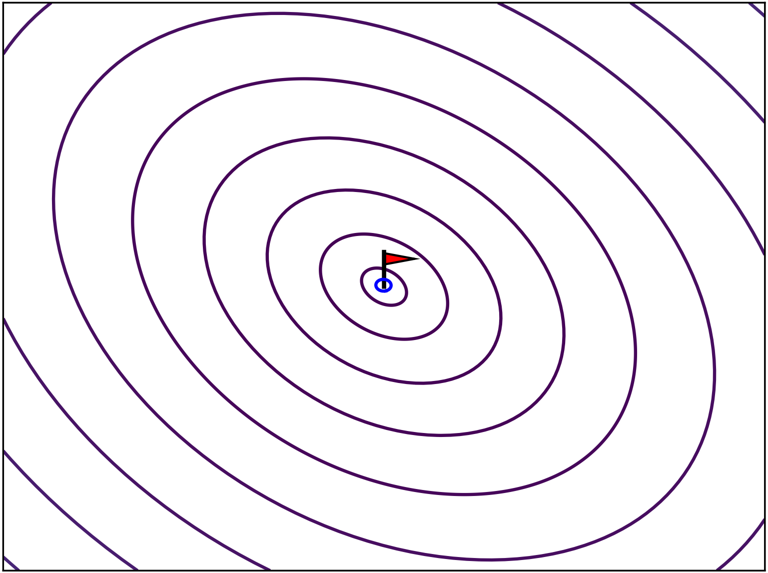


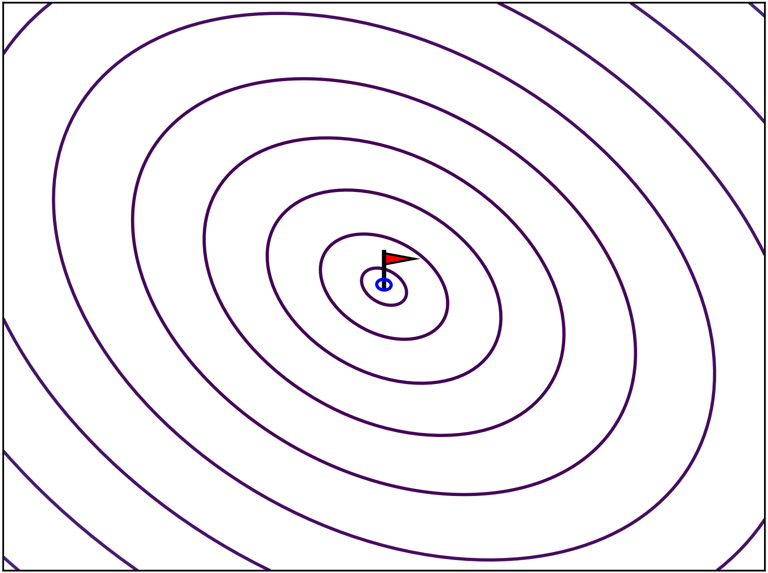






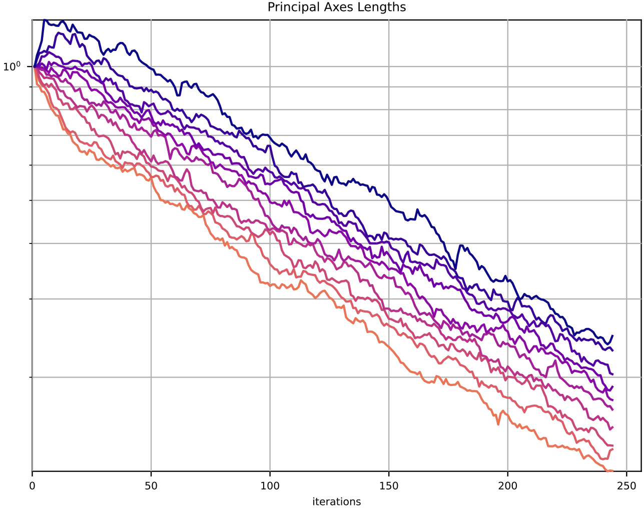






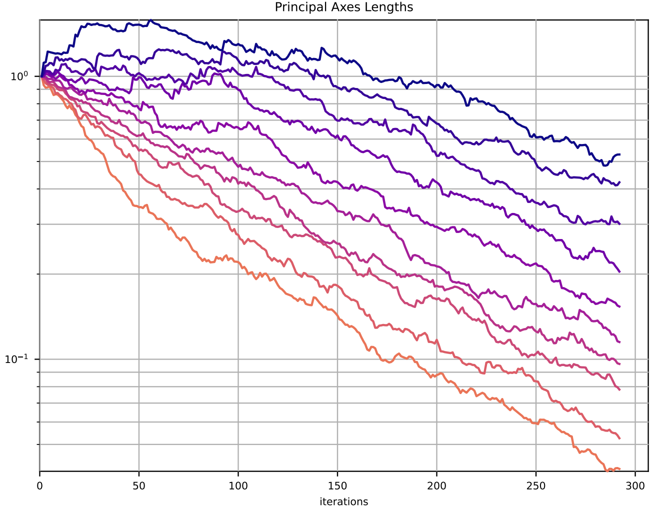
$\mathbf{C}_t$ approximates $\mathbf{H}^{-1}$.

$$\mathrm{Cond}(\mathbf{H}) = 10^0$$



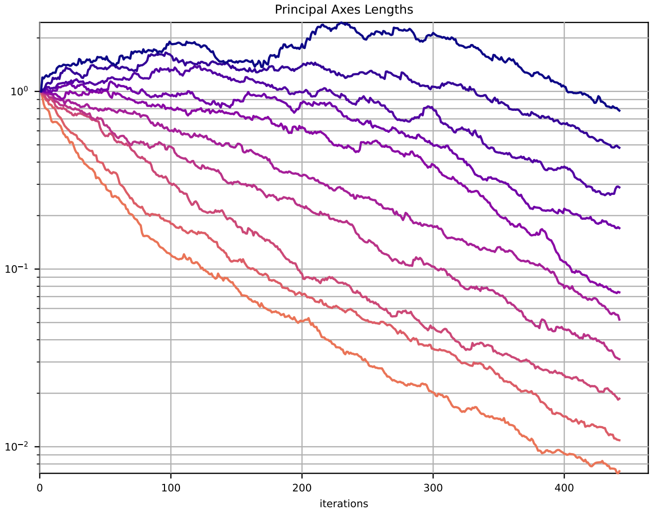
$\mathbf{C}_t$ approximates $\mathbf{H}^{-1}$.

$$\mathrm{Cond}(\mathbf{H}) = 10^2$$



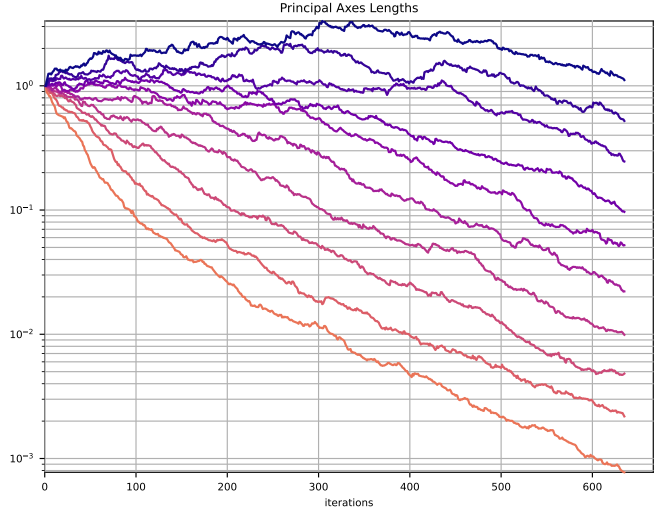
$\mathbf{C}_t$ approximates $\mathbf{H}^{-1}$.

$$\mathrm{Cond}(\mathbf{H}) = 10^4$$



$\mathbf{C}_t$ approximates $\mathbf{H}^{-1}$.

$$\mathrm{Cond}(\mathbf{H}) = 10^6$$



Goals:

Prove linear convergence of CMA-ES

Goals:

Prove linear convergence of CMA-ES:

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

Goals:

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 $f(x) = x^\top \mathbf{H} x/2$

Goals:

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 $f(x) = x^\top \mathbf{H} x/2$:

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

Markov chains and transition kernels

$P\colon \mathsf{X} \times \mathcal{B}(\mathsf{X}) \to \mathbb{R}$ is a **transition kernel** when

$$\forall x \in \mathsf{X}, \quad P(x, \cdot) \text{ is a probability measure.}$$

Markov chains and transition kernels

$P\colon \mathsf{X} \times \mathcal{B}(\mathsf{X}) \to \mathbb{R}$ is a **transition kernel** when

$$\forall x \in \mathsf{X}, \quad P(x, \cdot) \text{ is a probability measure.}$$

A **Markov chain** with transition kernel P is a random sequence $\{\theta_t\}_{t\in\mathbb{N}}$ such that:

$$\mathbb{P}[\theta_{t+1} \in \mathsf{A} \mid \theta_t = x] = P(x, \mathsf{A}).$$

Ergodic Markov chain

If $\theta_0 \sim \nu_0$

After k steps:

$$\theta_k \sim \nu_k = \nu_0 P^k = \int \nu_0(\mathrm{d}x) P^k(x, \cdot)$$

Ergodic Markov chain

If $\theta_0 \sim \nu_0$

After k steps:

$$\theta_k \sim \nu_k = \nu_0 P^k = \int \nu_0(\mathrm{d}x) P^k(x, \cdot)$$

If

$$\exists \pi, \forall \nu_0, \quad \lim_{k \to \infty} \nu_k = \pi$$

then $\{\theta_k\}_{k \in \mathbb{N}}$ is **ergodic**.

Limit theorems

If $\{\theta_t\}_{t\in\mathbb{N}}$ is ergodic with limit law π:

$$\lim_{t\to\infty} \mathbb{E}[f(\theta_t)] = \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} f(\theta_t) = \mathbb{E}_{\theta\sim\pi}[f(\theta)]$$

Methodology of convergence proof of ES

Methodology of convergence proof of ES

1. Define a normalized process $\{\phi_t\}_{t\in\mathbb{N}}$ of the algorithm $\{\theta_t\}_{t\in\mathbb{N}}$

Methodology of convergence proof of ES

1. Define a normalized process $\{\phi_t\}_{t\in\mathbb{N}}$ of the algorithm $\{\theta_t\}_{t\in\mathbb{N}}$
2. Prove: $\{\phi_t\}_{t\in\mathbb{N}}$ is an ergodic Markov chain that tends to π

Methodology of convergence proof of ES

1. Define a normalized process $\{\phi_t\}_{t\in\mathbb{N}}$ of the algorithm $\{\theta_t\}_{t\in\mathbb{N}}$
2. Prove: $\{\phi_t\}_{t\in\mathbb{N}}$ is an ergodic Markov chain that tends to π
3. Use limit theorems to prove linear convergence of $\{\theta_t\}_{t\in\mathbb{N}}$

Methodology of convergence proof of ES

1. Define a normalized process $\{\phi_t\}_{t\in\mathbb{N}}$ of the algorithm $\{\theta_t\}_{t\in\mathbb{N}}$
2. Prove: $\{\phi_t\}_{t\in\mathbb{N}}$ is an ergodic Markov chain that tends to π
3. Use limit theorems to prove linear convergence of $\{\theta_t\}_{t\in\mathbb{N}}$

This approach was successful for stepsize adaptive-ES

Definition of a normalized Markov chain for CMA-ES

In order to obtain a stationary Markov chain:

$$z_t = \frac{m_t - x^*}{\sigma_t \sqrt{\lambda_{\min}(\mathbf{C}_t)}}$$

$$\mathbf{\Sigma}_t = \frac{\mathbf{C}_t}{\lambda_{\min}(\mathbf{C}_t)}$$

Proposition

If $f \in \Big\{$ , , ,  $\Big\}$,[1] *then* $\{(z_t, \mathbf{\Sigma}_t)\}_{t\in\mathbb{N}}$ *is a Markov chain.*

[1] scaling-invariant functions

If $\{(z_t, \boldsymbol{\Sigma}_t)\}_{t \in \mathbb{N}}$ is ergodic:

$$\frac{1}{T} \log \frac{\|m_T - x^*\|}{\|m_0 - x^*\|}$$

If $\{(z_t, \mathbf{\Sigma}_t)\}_{t\in\mathbb{N}}$ is ergodic:

$$\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \frac{1}{T}\sum_{t=0}^{T-1}\log\|z_{t+1}\| - \log\|z_t\| - \log\frac{\sigma_{t+1}\lambda_{\min}(\mathbf{C}_{t+1})^{1/2}}{\sigma_t\lambda_{\min}(\mathbf{C}_t)^{1/2}}$$

If $\{(z_t, \boldsymbol{\Sigma}_t)\}_{t\in\mathbb{N}}$ is ergodic:

$$\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \frac{1}{T}\sum_{t=0}^{T-1}\log\|z_{t+1}\| - \log\|z_t\| - \log\frac{\sigma_{t+1}\lambda_{\min}(\mathbf{C}_{t+1})^{1/2}}{\sigma_t\lambda_{\min}(\mathbf{C}_t)^{1/2}}$$

$$\lim_{T\to\infty}\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \mathbb{E}_\pi[\log\|z\|] - \mathbb{E}_\pi[\log\|z\|] - \mathbb{E}_\pi\left[\log\frac{\sigma_1\lambda_{\min}(\mathbf{C}_1)^{1/2}}{\sigma_0\lambda_{\min}(\mathbf{C}_0)^{1/2}}\right]$$

If $\{(z_t, \mathbf{\Sigma}_t)\}_{t\in\mathbb{N}}$ is ergodic:

$$\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \frac{1}{T}\sum_{t=0}^{T-1}\log\|z_{t+1}\| - \log\|z_t\| - \log\frac{\sigma_{t+1}\lambda_{\min}(\mathbf{C}_{t+1})^{1/2}}{\sigma_t\lambda_{\min}(\mathbf{C}_t)^{1/2}}$$

$$\lim_{T\to\infty}\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \mathbb{E}_\pi[\log\|z\|] - \mathbb{E}_\pi[\log\|z\|] - \mathbb{E}_\pi\left[\log\frac{\sigma_1\lambda_{\min}(\mathbf{C}_1)^{1/2}}{\sigma_0\lambda_{\min}(\mathbf{C}_0)^{1/2}}\right]$$

$$\|m_T - x^*\| \underset{T\to\infty}{\sim} e^{-T\mathbb{E}_\pi\left[\log\frac{\sigma_1\lambda_{\min}(\mathbf{C}_1)^{1/2}}{\sigma_0\lambda_{\min}(\mathbf{C}_0)^{1/2}}\right]}\|m_0 - x^*\|$$

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

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

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 $\{(z_t, \boldsymbol{\Sigma}_t)\}_{t \in \mathbb{N}}$ is an ergodic Markov chain?

How can we prove that $\{(z_t, \boldsymbol{\Sigma}_t)\}_{t\in\mathbb{N}}$ is an ergodic Markov chain?

(and under which conditions?)

How to prove that $\{\phi_t\}_{t\in\mathbb{N}}$ is ergodic

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

$$\mathbb{E}[V(\phi_1)] \leqslant (1-\varepsilon)V(\phi_0) \qquad \forall \phi_0 \notin \mathsf{K}$$

How to prove that $\{\phi_t\}_{t\in\mathbb{N}}$ is ergodic

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

$$\mathbb{E}[V(\phi_1)] \leqslant (1-\varepsilon)V(\phi_0) \qquad \forall \phi_0 \notin \mathsf{K}$$

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

1. Irreducibility and aperiodicity

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

$$\forall \phi_{\text{start}}, \phi_{\text{end}} \in \Phi, \underbrace{\exists k > 0,\ \mathbb{P}[\phi_k = \phi_{\text{end}} \mid \phi_0 = \phi_{\text{start}}] > 0}_{\phi_{\text{start}} \rightsquigarrow \phi_{\text{end}}}$$

1. Irreducibility and aperiodicity

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

$$\forall \phi_{\text{start}} \in \Phi, \forall \Phi_{\text{end}} \subset \Phi,\ \text{Volume}(\Phi_{\text{end}}) > 0 \Rightarrow \phi_{\text{start}} \rightsquigarrow \Phi_{\text{end}}$$

1. Irreducibility and aperiodicity

Theorem*
The Markov chain

$$\phi_{t+1} = F(\phi_t, U_{t+1})$$

is irreducible and aperiodic when

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

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

Assumptions: F is loc. Lipschitz and $U_{t+1} \sim p_{\phi_t}$ where $(\phi, u) \mapsto p_\phi(u)$ is lsc

1. Irreducibility and aperiodicity

Theorem*
The Markov chain

$$\phi_{t+1} = F(\phi_t, U_{t+1})$$

is irreducible and aperiodic when

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

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

For us:

$$(z_{t+1}, \mathbf{\Sigma}_{t+1}) = F((z_t, \mathbf{\Sigma}_t), z_{t+1}^{i:\lambda})$$

Assumptions: F is loc. Lipschitz and $U_{t+1} \sim p_{\phi_t}$ where $(\phi, u) \mapsto p_\phi(u)$ is lsc

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

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

Consequence:

$\{(z_t, \mathbf{\Sigma}_t)\}$ is irreducible and aperiodic.

2. Drift condition

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

2. Drift condition

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

(a) When $|||\mathbf{\Sigma}_0||| \gg 1 + \|z_0\|^2$:

$$\mathbb{E}[|||\mathbf{\Sigma}_1|||] \leqslant (1 - \varepsilon) |||\mathbf{\Sigma}_1|||$$

2. Drift condition

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

(a) When $|||\mathbf{\Sigma}_0||| \gg 1 + \|z_0\|^2$:

$$\mathbb{E}[|||\mathbf{\Sigma}_1|||] \leqslant (1 - \varepsilon)|||\mathbf{\Sigma}_1|||$$

(b) When $|||\mathbf{\Sigma}_0||| \not\gg \|z_0\|^2$:

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

2. Drift condition

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

(a) When $|||\mathbf{\Sigma}_0||| \gg 1 + \|z_0\|^2$:

$$\mathbb{E}[|||\mathbf{\Sigma}_1|||] \leqslant (1 - \varepsilon)|||\mathbf{\Sigma}_1|||$$

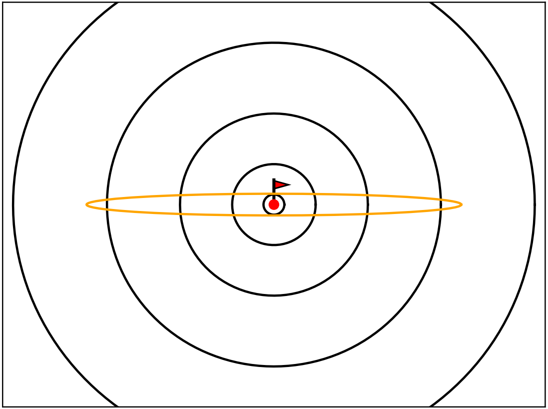
(b) When $|||\mathbf{\Sigma}_0||| \not\gg \|z_0\|^2$:

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

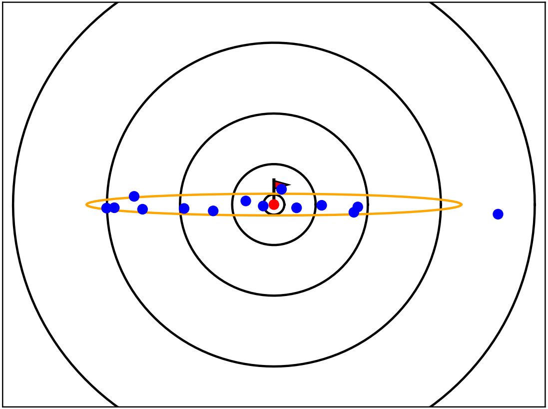
Proposition
If (a) and (b) are true:

$$\exists \mathsf{K} \text{ compact}, \; \mathbb{E}[V(z_1, \mathbf{\Sigma}_1)] \leqslant (1 - \varepsilon)V(z_0, \mathbf{\Sigma}_0) \quad \forall (z_0, \mathbf{\Sigma}_0) \notin \mathsf{K}$$

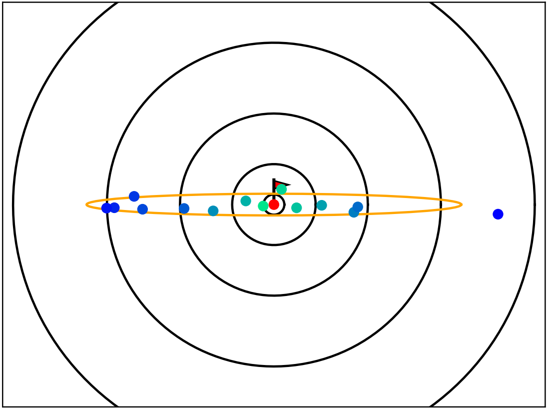
(a) When $\|\boldsymbol{\Sigma}_0\| \gg 1 + \|z_0\|^2$



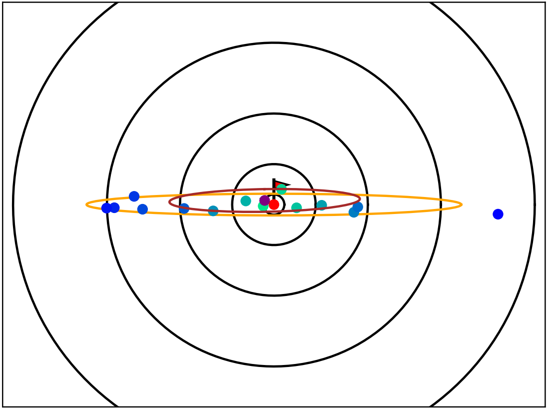
(a) When $\|\boldsymbol{\Sigma}_0\| \gg 1 + \|z_0\|^2$



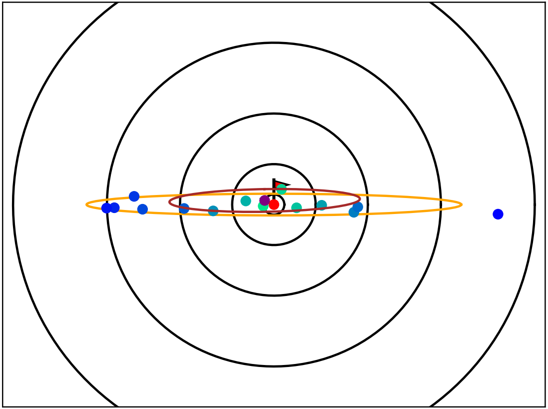
(a) When $\|\boldsymbol{\Sigma}_0\| \gg 1 + \|z_0\|^2$



(a) When $\|\boldsymbol{\Sigma}_0\| \gg 1 + \|z_0\|^2$



(a) When $|||\boldsymbol{\Sigma}_0||| \gg 1 + \|z_0\|^2$



Proposition*

When $f =$  *and* $|||\boldsymbol{\Sigma}_0||| \gg 1 + \|z_0\|^2$*:*

$$\mathbb{E}[|||\boldsymbol{\Sigma}_1|||] \leqslant (1-\varepsilon)|||\boldsymbol{\Sigma}_1|||$$

(b) When $\|\boldsymbol{\Sigma}_0\| \not\gg \|z_0\|^2$

(b) When $\|\!|\boldsymbol{\Sigma}_0|\!\| \not\gg \|z_0\|^2$

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

(b) When $\|\!|\boldsymbol{\Sigma}_0|\!\| \not\gg \|z_0\|^2$

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

with

$$\text{normalization} = \text{increasing function}(\|m_{t+1} - m_t\|) \times \sqrt{\lambda_{\min}(\boldsymbol{\Sigma}_1)}$$

(b) When $\|\|\|\boldsymbol{\Sigma}_0\|\|\| \not\gg \|z_0\|^2$

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

with

$$\text{normalization} = \text{increasing function}(\|m_{t+1} - m_t\|) \times \sqrt{\lambda_{\min}(\boldsymbol{\Sigma}_1)}$$

Proposition*

When $f =$  *and* $\|\|\|\boldsymbol{\Sigma}_0\|\|\| \not\gg \|z_0\|^2$*:*

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

(b) When $|||\boldsymbol{\Sigma}_0||| \not\gg \|z_0\|^2$

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

with

$$\text{normalization} = \text{increasing function}(\|m_{t+1} - m_t\|) \times \sqrt{\lambda_{\min}(\boldsymbol{\Sigma}_1)}$$

Proposition*

When $f =$  *and* $|||\boldsymbol{\Sigma}_0||| \not\gg \|z_0\|^2$*:*

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

If we choose the hyperparameters correctly:

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

(b) When $\|\|\|\boldsymbol{\Sigma}_0\|\|\| \not\ggg \|z_0\|^2$

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

with

$$\text{normalization} = \text{increasing function}(\|m_{t+1} - m_t\|) \times \sqrt{\lambda_{\min}(\boldsymbol{\Sigma}_1)}$$

Proposition*

When $f =$  *and* $\|\|\|\boldsymbol{\Sigma}_0\|\|\| \not\ggg \|z_0\|^2$*:*

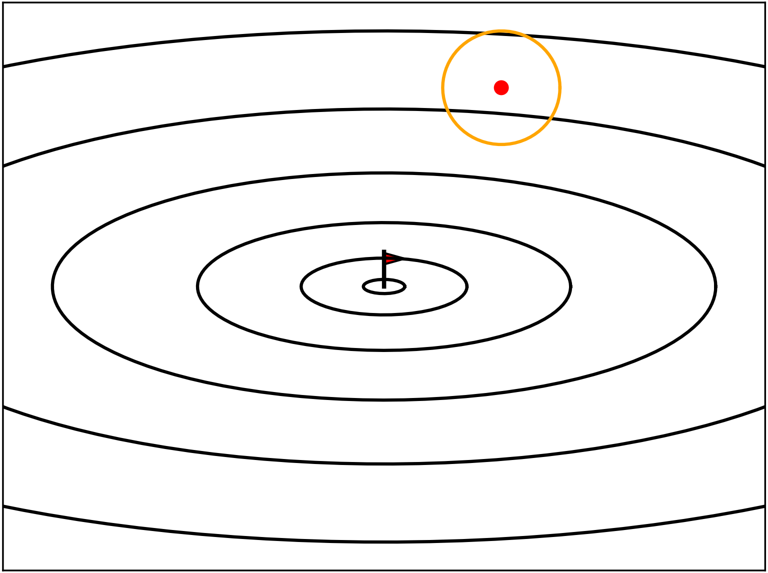
$$\mathbb{E}[\|m_{t+1} - m_t\|] > \mathbb{E}\|\mathcal{N}\|$$

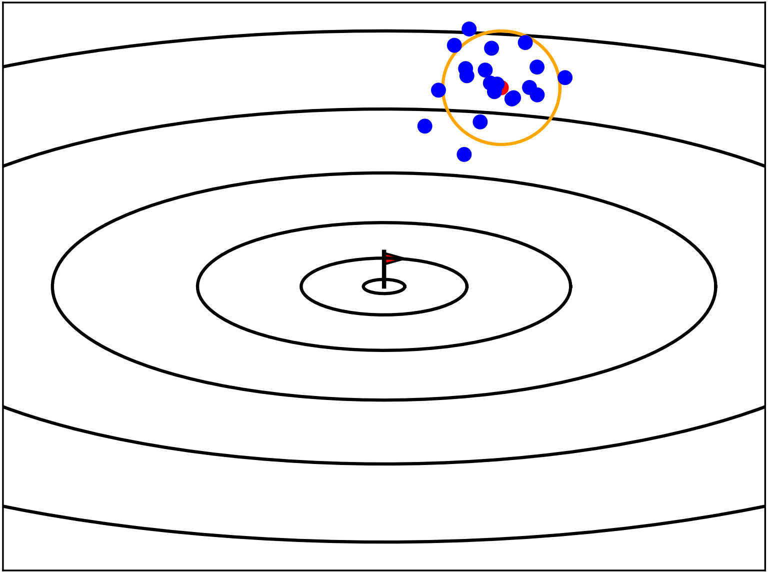
If we choose the hyperparameters correctly:

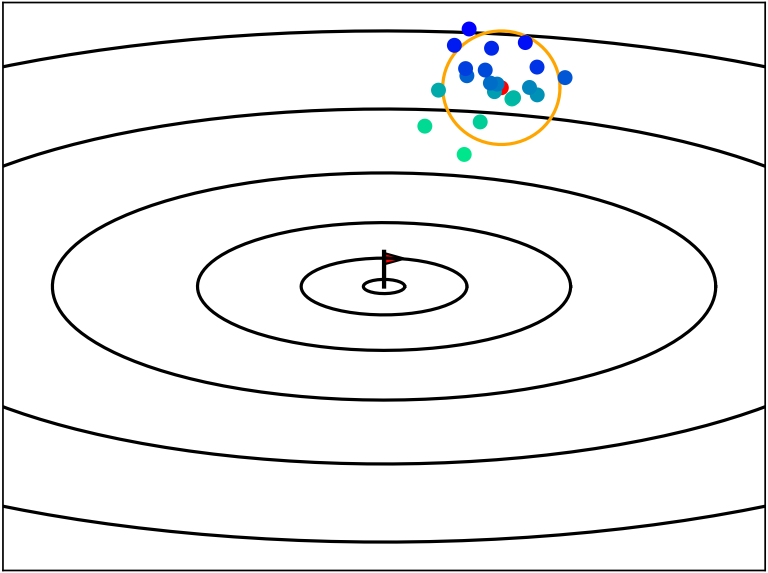
$$\mathbb{E}[\textit{normalization}] > 1$$

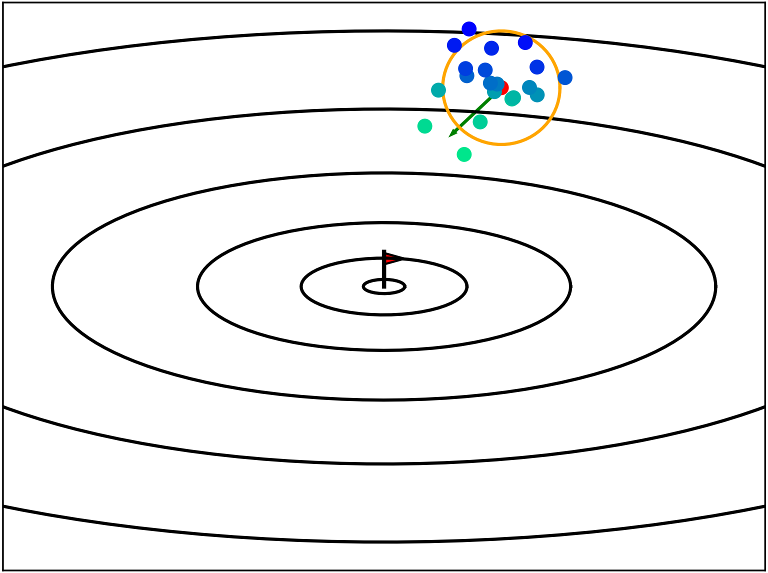
and

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









Theorem*

When $f =$ 

$$\exists \mathrm{K} \text{ compact}, \ \mathbb{E}[V(z_1, \mathbf{\Sigma}_1)] \leqslant (1-\varepsilon) V(z_0, \mathbf{\Sigma}_0) \quad \forall (z_0, \mathbf{\Sigma}_0) \notin \mathrm{K}$$

Consequence:

$\{(z_t, \mathbf{\Sigma}_t)\}_t$ is ergodic

Theorem*

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

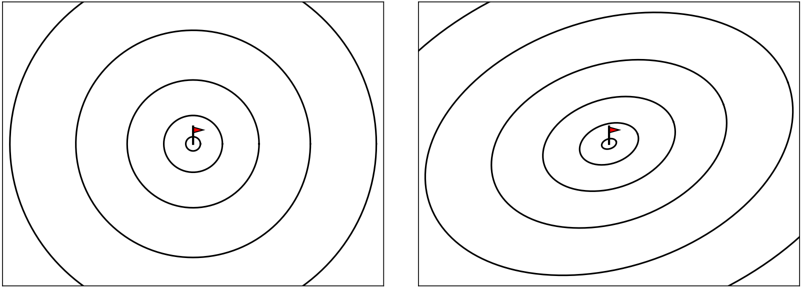
Theorem*

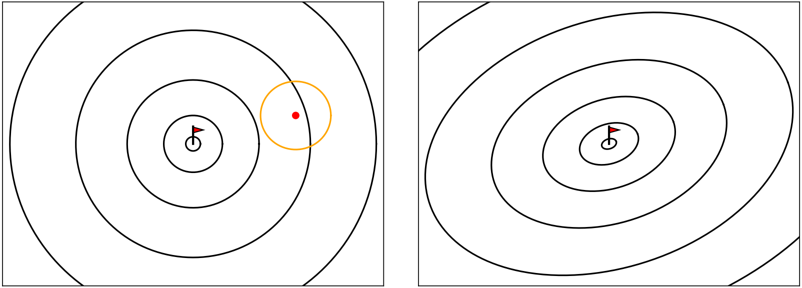
When $f =$ *, CMA-ES converges linearly.*

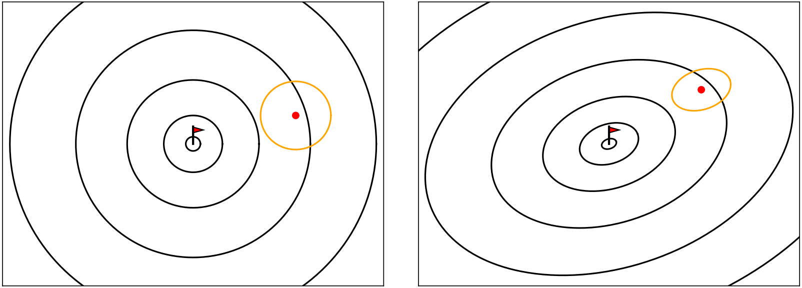
How to extend to $f =$ **?**

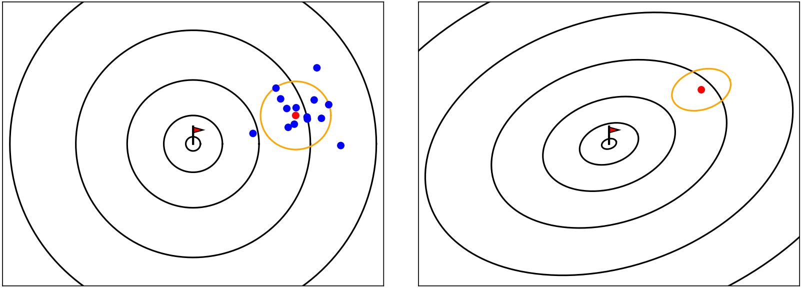
Affine-invariance

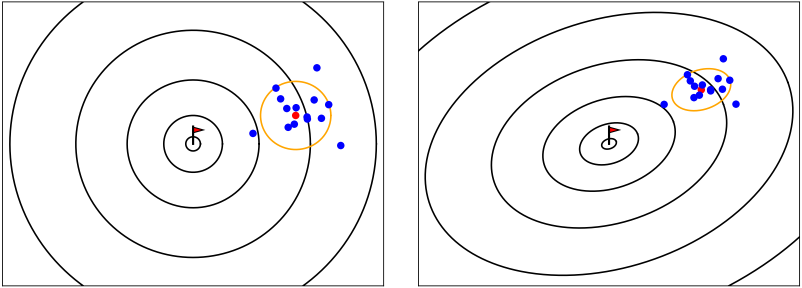
$$
\begin{array}{ccc}
(m_0, \mathbf{C}_0) & \xrightarrow{\min f(x)} & (m_1, \mathbf{C}_1) \\
\Psi \big\downarrow & & \big\uparrow \Psi^{-1} \\
(m_0', \mathbf{C}_0') & \xrightarrow{\min f(Bx+b)} & (m_1', \mathbf{C}_1')
\end{array}
$$

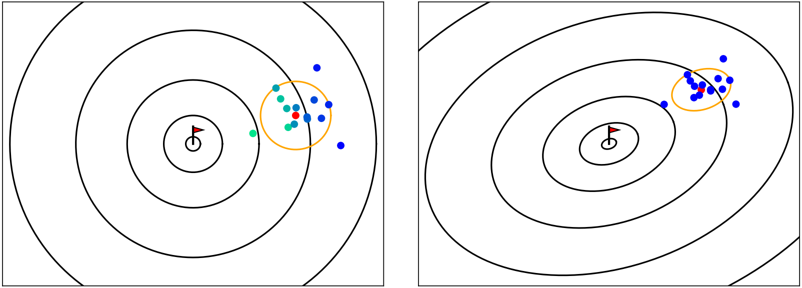


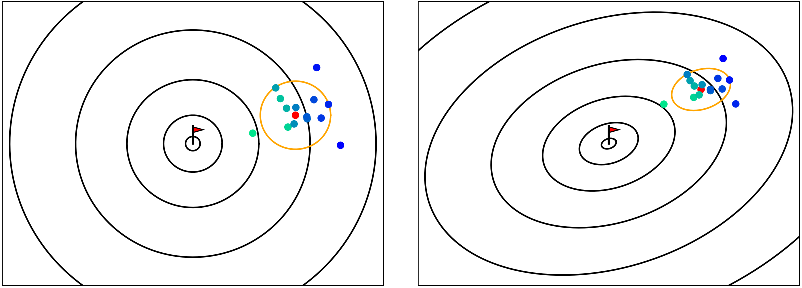


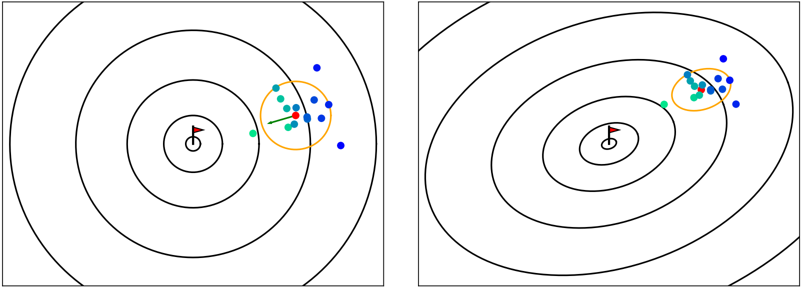


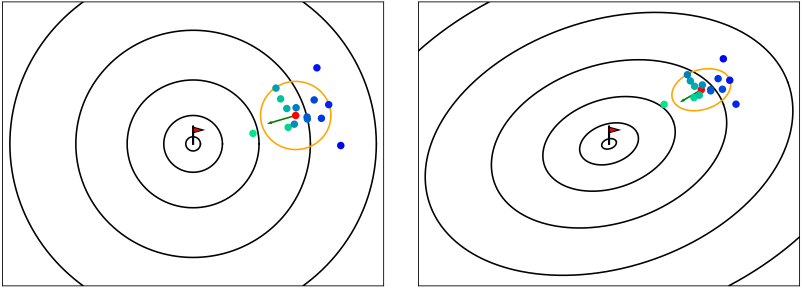


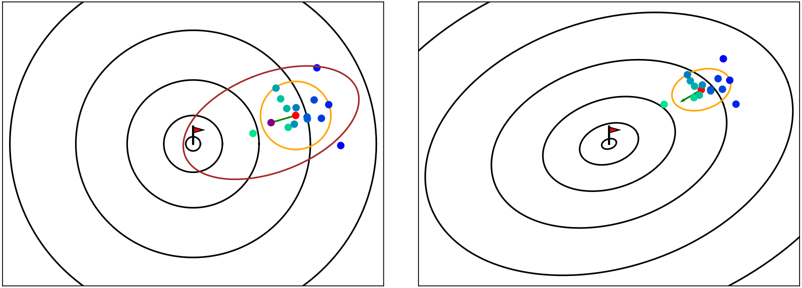


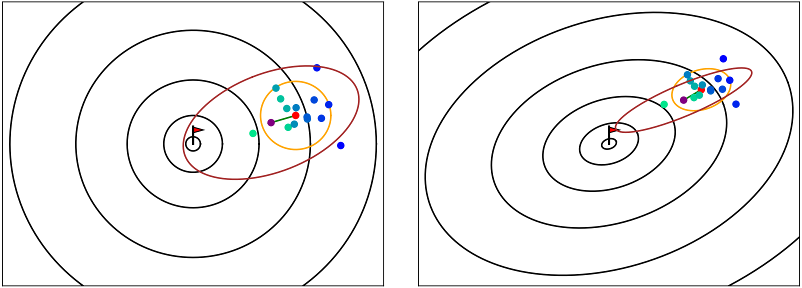












Theorem
CMA-ES is affine-invariant

Theorem
CMA-ES is affine-invariant

Consequence

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

Theorem
CMA-ES is affine-invariant

Consequence

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

(with the same convergence rate than *)*

Learning of the inverse Hessian

When $f =$ , we find

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{\mathbf{C}_t}{\text{normalization}}\right] = \mathbf{I}_d$$

Learning of the inverse Hessian

When $f =$ , we find

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{\mathbf{C}_t}{\text{normalization}}\right] = \mathbf{I}_d$$

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

$f =$  $\Rightarrow \lim_{t\to\infty} \mathbb{E}\left[\frac{\mathbf{C}_t}{\text{normalization}}\right] = \text{Hessian}^{-1/2} \times \mathbf{I}_d \times \text{Hessian}^{-1/2}$

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

Learning of the inverse Hessian

When $f =$ , we find

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{\mathbf{C}_t}{\text{normalization}}\right] = \mathbf{I}_d$$

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

$f =$  $\Rightarrow \lim_{t\to\infty} \mathbb{E}\left[\frac{\mathbf{C}_t}{\text{normalization}}\right] = \text{Hessian}^{-1/2} \times \mathbf{I}_d \times \text{Hessian}^{-1/2}$

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

Theorem*

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