Convergence Analysis of Evolution Strategies with Covariance Matrix Adaptation (CMA-ES) via Markov Chain Stability Analysis

Blackbox Optimization and Derivative-Free Algorithms

Armand Gissler

Friday 2nd June, 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

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

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 $\mathcal{N}(m, \sigma^2 C)$

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 $\mathcal{N}(m, \sigma^2 C)$ **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

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

$\Rightarrow$ 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 $\mathcal{N}(m, \sigma^2 C)$ **by adapting** the mean $m \in \mathbb{R}^d$, **the stepsize** $\sigma > 0$

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 $\mathcal{N}(m, \sigma^2 C)$ **by adapting** the mean $m \in \mathbb{R}^d$, the stepsize $\sigma > 0$ **and the covariance matrix** $C \in \mathcal{S}^d_{++}$.

- **converges linearly to the minimum**

- converges linearly to the minimum
- **learns the inverse Hessian** of a convex-quadratic function

- converges linearly to the minimum
- learns the inverse Hessian of a convex-quadratic function
- state-of-the-art method for **difficult (derivative-free) optimization problems** such as minimization of **ill-conditionned**, **non-separable**, **discontinuous**, **multimodal** and/or **noisy** functions

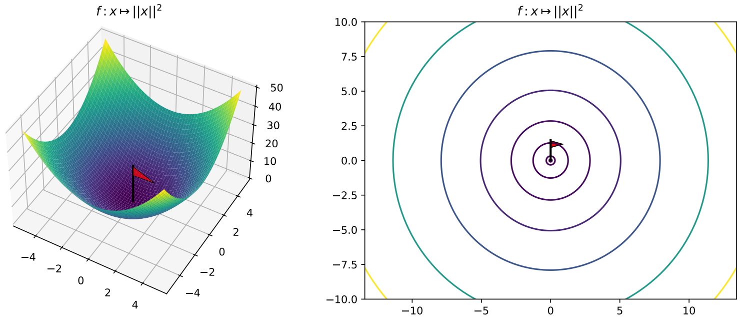
- converges linearly to the minimum
- learns the inverse Hessian of a convex-quadratic function
- state-of-the-art method for difficult (derivative-free) optimization problems such as minimization of ill-conditionned, non-separable, discontinuous, multimodal and/or noisy functions
- >40 **millions of downloads of two Python modules**

- converges linearly to the minimum
- learns the inverse Hessian of a convex-quadratic function
- state-of-the-art method for difficult (derivative-free) optimization problems such as minimization of ill-conditionned, non-separable, discontinuous, multimodal and/or noisy functions
- $>$ 40 millions of downloads of two Python modules
- **proofs of convergence require additional assumptions so far**

- Without covariance matrix adaptation: Touré et al, *Global linear convergence of Evolution Strategies with recombination on scaling-invariant functions* (2021)
- With a sufficient decrease condition: Diouane et al, *Globally convergent evolution strategies* (2015)
- Assuming that the covariance matrix is bounded: Akimoto et al, *Global linear convergence of evolution strategies on more than smooth strongly convex functions* (2022)
- Using a different update for the covariance matrix: Glasmachers et al, *Convergence analysis of the Hessian estimation evolution strategy* (2022)

CMA-ES: algorithm presentation

Level sets representation



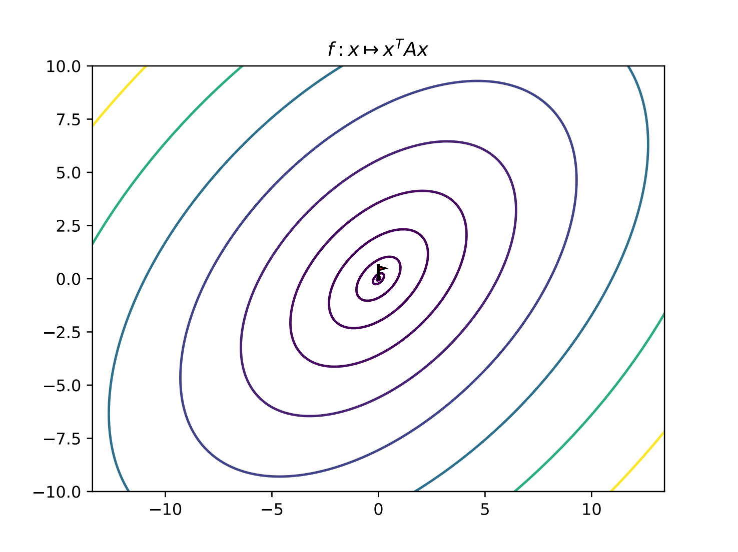
CMA-ES : presentation

Principle of Evolution Strategies (ES) : approximate the minimum of the function by a distribution $\mathcal{N}(m, \sigma^2 C)$.

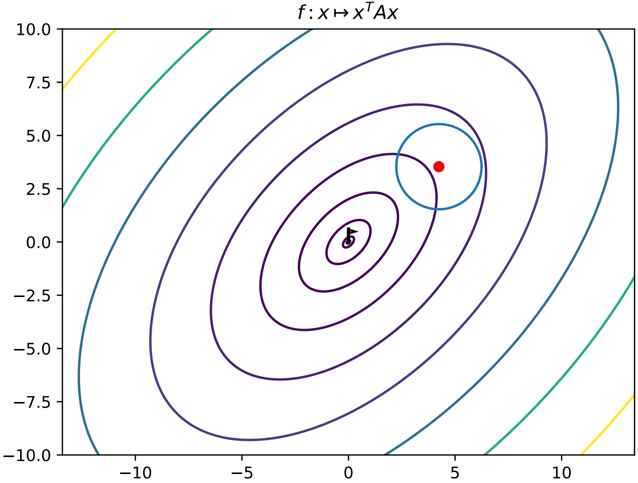
CMA-ES : presentation

At each iteration $t \in \mathbb{N}$, given a **mean** m_t, a **stepsize** σ_t and a **covariance matrix** C_t :

Start from a distribution $\mathcal{N}(m_t, \sigma_t^2 C_t)$



Start from a distribution $\mathcal{N}(m_t, \sigma_t^2 C_t)$



CMA-ES : presentation

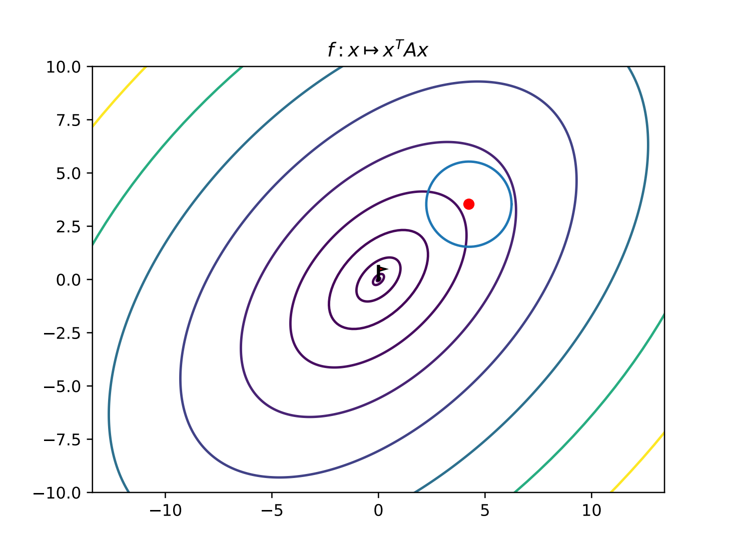
At each iteration $t \in \mathbb{N}$, given a **mean** m_t, a **stepsize** σ_t and a **covariance matrix** C_t :

CMA-ES : presentation

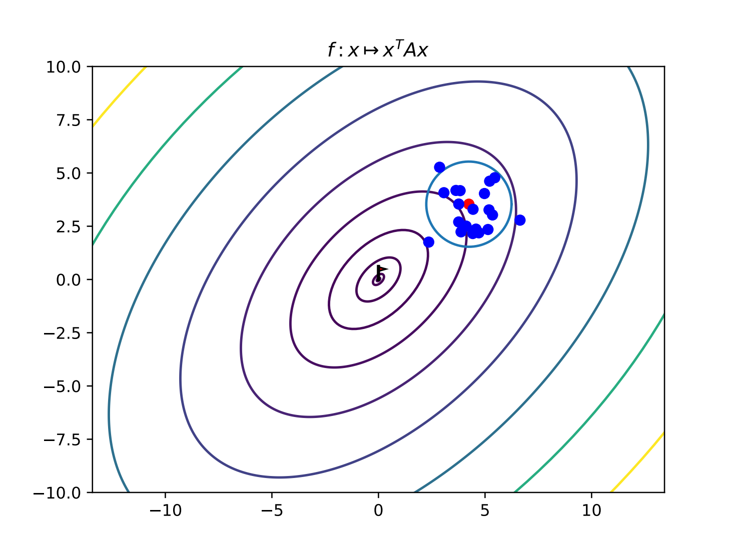
At each iteration $t \in \mathbb{N}$, given a **mean** m_t, a **stepsize** σ_t and a **covariance matrix** C_t :

1. **Generate** λ offspring $x_{t+1}^i \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$ *independently*

Generate λ offspring $x_{t+1}^i \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$ *independently*



Generate λ offspring $x_{t+1}^i \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$ *independently*



CMA-ES : presentation

At each iteration $t \in \mathbb{N}$, given a **mean** m_t, a **stepsize** σ_t and a **covariance matrix** C_t :

1. **Generate** λ offspring $x_{t+1}^i \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$ *independently*

CMA-ES : presentation

At each iteration $t \in \mathbb{N}$, given a **mean** m_t, a **stepsize** σ_t and a **covariance matrix** C_t :

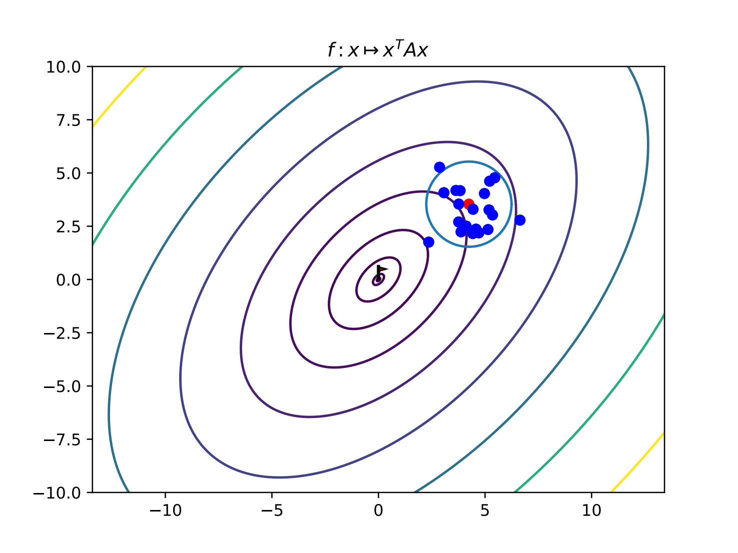
1. Generate λ offspring $x_{t+1}^i \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$ *independently*
2. **Rank** the x_{t+1}^i w.r.t. their f-values

CMA-ES : presentation

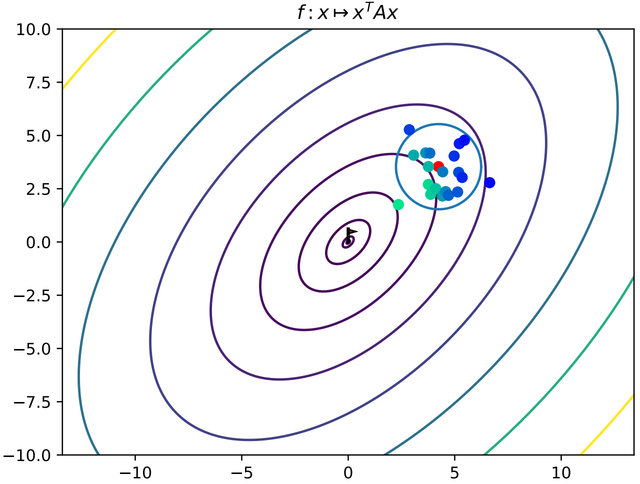
At each iteration $t \in \mathbb{N}$, given a **mean** m_t, a **stepsize** σ_t and a **covariance matrix** C_t :

1. Generate λ offspring $x_{t+1}^i \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$ *independently*
2. **Rank** the x_{t+1}^i w.r.t. their f-values: $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$

Rank the x_{t+1}^i w.r.t. their f-values $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$



Rank the x_{t+1}^{i} w.r.t. their f-values $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$



CMA-ES : presentation

At each iteration $t \in \mathbb{N}$, given a **mean** m_t, a **stepsize** σ_t and a **covariance matrix** C_t :

1. Generate λ offspring $x^i_{t+1} \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$ *independently*
2. **Rank** the x^i_{t+1} w.r.t. their f-values: $f\left(x^{1:\lambda}_{t+1}\right) \leqslant \cdots \leqslant f\left(x^{\lambda:\lambda}_{t+1}\right)$

CMA-ES : presentation

At each iteration $t \in \mathbb{N}$, given a **mean** m_t, a **stepsize** σ_t and a **covariance matrix** C_t :

1. Generate λ offspring $x_{t+1}^i \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$ *independently*
2. Rank the x_{t+1}^i w.r.t. their f-values: $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$
3. **Update the mean**

CMA-ES : presentation

At each iteration $t \in \mathbb{N}$, given a **mean** m_t, a **stepsize** σ_t and a **covariance matrix** C_t :

1. Generate λ offspring $x^i_{t+1} \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$ *independently*
2. Rank the x^i_{t+1} w.r.t. their f-values: $f\left(x^{1:\lambda}_{t+1}\right) \leqslant \cdots \leqslant f\left(x^{\lambda:\lambda}_{t+1}\right)$
3. **Update the mean**: $m_{t+1} = \text{Weighted average(population)}$

CMA-ES : presentation

At each iteration $t \in \mathbb{N}$, given a **mean** m_t, a **stepsize** σ_t and a **covariance matrix** C_t :

1. Generate λ offspring $x_{t+1}^i \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$ *independently*
2. Rank the x_{t+1}^i w.r.t. their f-values: $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$
3. **Update the mean**: $m_{t+1} = \sum w_i x_{t+1}^{i:\lambda}$

CMA-ES : presentation

At each iteration $t \in \mathbb{N}$, given a **mean** m_t, a **stepsize** σ_t and a **covariance matrix** C_t :

1. Generate λ offspring $x_{t+1}^i \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$ *independently*
2. Rank the x_{t+1}^i w.r.t. their f-values: $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$
3. **Update the mean**: $m_{t+1} = \sum w_i x_{t+1}^{i:\lambda}$

   The best offspring have the largest weights

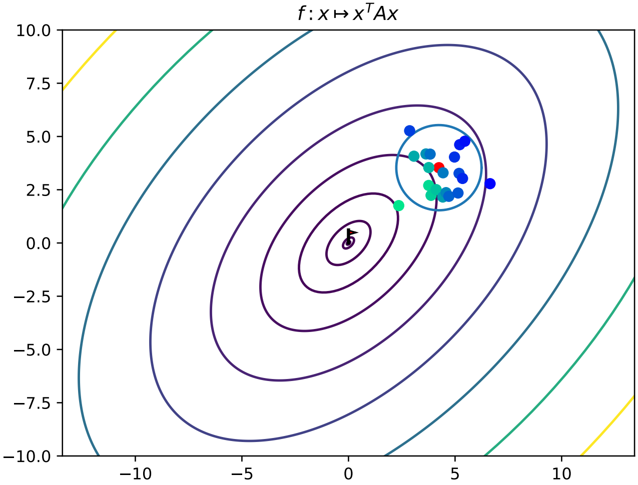
CMA-ES : presentation

At each iteration $t \in \mathbb{N}$, given a **mean** m_t, a **stepsize** σ_t and a **covariance matrix** C_t :

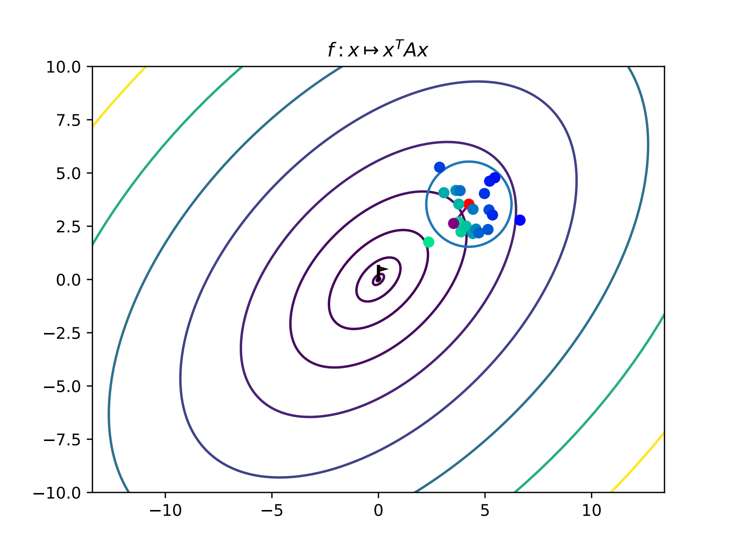
1. Generate λ offspring $x^i_{t+1} \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$ *independently*
2. Rank the x^i_{t+1} w.r.t. their f-values: $f\left(x^{1:\lambda}_{t+1}\right) \leqslant \cdots \leqslant f\left(x^{\lambda:\lambda}_{t+1}\right)$
3. **Update the mean**: $m_{t+1} = \sum w_i x^{i:\lambda}_{t+1}$

   The best offspring have the largest weights: $w_1 \geqslant w_2 \ldots$

Update the mean $m_{t+1} = \sum w_i x_{t+1}^{i:\lambda}$



Update the mean $m_{t+1} = \sum w_i x_{t+1}^{i:\lambda}$



CMA-ES : presentation

At each iteration $t \in \mathbb{N}$, given a **mean** m_t, a **stepsize** σ_t and a **covariance matrix** C_t :

1. Generate λ offspring $x_{t+1}^i \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$ *independently*
2. Rank the x_{t+1}^i w.r.t. their f-values: $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$
3. **Update the mean**: $m_{t+1} = \sum w_i x_{t+1}^{i:\lambda}$

   The best offspring have the largest weights: $w_1 \geqslant w_2 \ldots$

CMA-ES : presentation

At each iteration $t \in \mathbb{N}$, given a **mean** m_t, a **stepsize** σ_t and a **covariance matrix** C_t :

1. Generate λ offspring $x^i_{t+1} \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$ *independently*
2. Rank the x^i_{t+1} w.r.t. their f-values: $f\left(x^{1:\lambda}_{t+1}\right) \leqslant \cdots \leqslant f\left(x^{\lambda:\lambda}_{t+1}\right)$
3. Update the mean: $m_{t+1} = \sum w_i x^{i:\lambda}_{t+1}$

   The best offspring have the largest weights: $w_1 \geqslant w_2 \ldots$
4. **Update the stepsize**

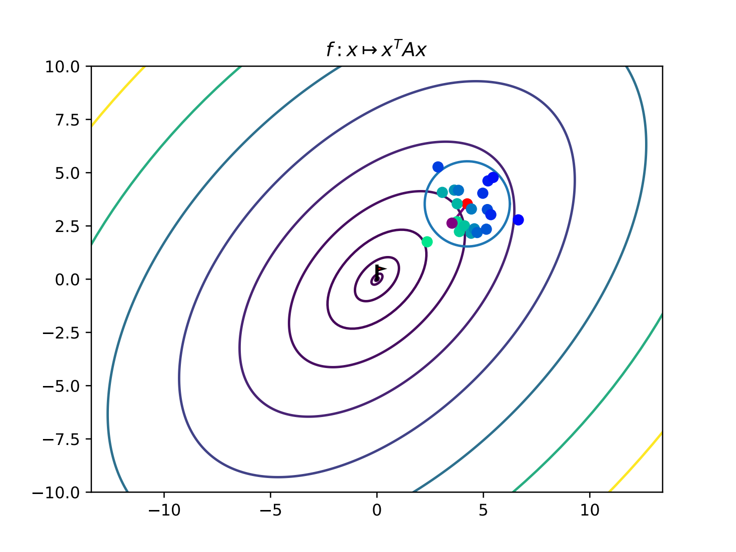
CMA-ES : presentation

At each iteration $t \in \mathbb{N}$, given a **mean** m_t, a **stepsize** σ_t and a **covariance matrix** C_t :

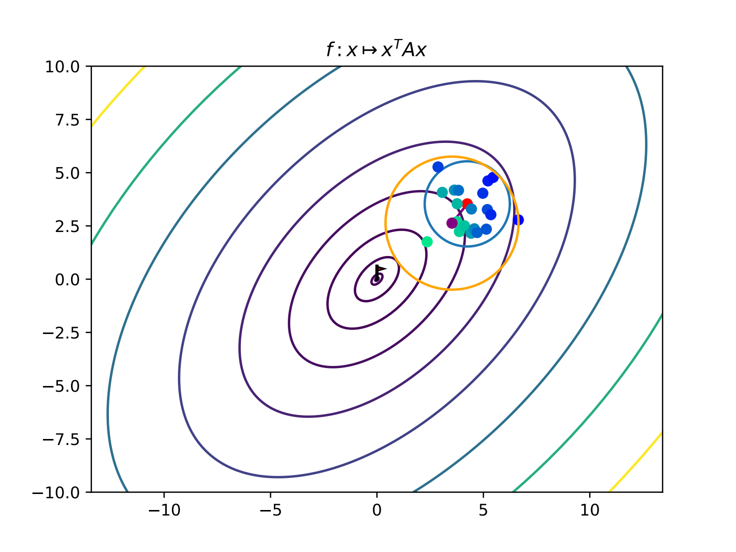
1. Generate λ offspring $x_{t+1}^i \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$ *independently*
2. Rank the x_{t+1}^i w.r.t. their f-values: $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$
3. Update the mean: $m_{t+1} = \sum w_i x_{t+1}^{i:\lambda}$

   The best offspring have the largest weights: $w_1 \geqslant w_2 \ldots$
4. **Update the stepsize**:
   Increase the stepsize if the path taken by the mean is **larger than expected** (assuming no selection)

Adapt the stepsize



Adapt the stepsize



CMA-ES : presentation

At each iteration $t \in \mathbb{N}$, given a **mean** m_t, a **stepsize** σ_t and a **covariance matrix** C_t :

1. Generate λ offspring $x^i_{t+1} \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$ *independently*
2. Rank the x^i_{t+1} w.r.t. their f-values: $f\left(x^{1:\lambda}_{t+1}\right) \leqslant \cdots \leqslant f\left(x^{\lambda:\lambda}_{t+1}\right)$
3. Update the mean: $m_{t+1} = \sum w_i x^{i:\lambda}_{t+1}$

   The best offspring have the largest weights: $w_1 \geqslant w_2 \ldots$
4. **Update the stepsize**:
   Increase the stepsize if the path taken by the mean is **larger than expected** (assuming no selection)

CMA-ES : presentation

At each iteration $t \in \mathbb{N}$, given a **mean** m_t, a **stepsize** σ_t and a **covariance matrix** C_t :

1. Generate λ offspring $x^i_{t+1} \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$ *independently*
2. Rank the x^i_{t+1} w.r.t. their f-values: $f\left(x^{1:\lambda}_{t+1}\right) \leqslant \dots \leqslant f\left(x^{\lambda:\lambda}_{t+1}\right)$
3. Update the mean: $m_{t+1} = \sum w_i x^{i:\lambda}_{t+1}$

   The best offspring have the largest weights: $w_1 \geqslant w_2 \dots$
4. Update the stepsize:
   Increase the stepsize if the path taken by the mean is larger than expected (assuming no selection)
5. **Update the covariance matrix**

CMA-ES : presentation

At each iteration $t \in \mathbb{N}$, given a **mean** m_t, a **stepsize** σ_t and a **covariance matrix** C_t :

1. Generate λ offspring $x_{t+1}^i \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$ *independently*
2. Rank the x_{t+1}^i w.r.t. their f-values: $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$
3. Update the mean: $m_{t+1} = \sum w_i x_{t+1}^{i:\lambda}$

   The best offspring have the largest weights: $w_1 \geqslant w_2 \ldots$
4. Update the stepsize:
   Increase the stepsize if the path taken by the mean is larger than expected (assuming no selection)
5. **Update the covariance matrix**

$$C_{t+1} = (1-c)C_t + c' \frac{(m_{t+1}-m_t)(m_{t+1}-m_t)^T}{\sigma_t^2}$$

CMA-ES : presentation

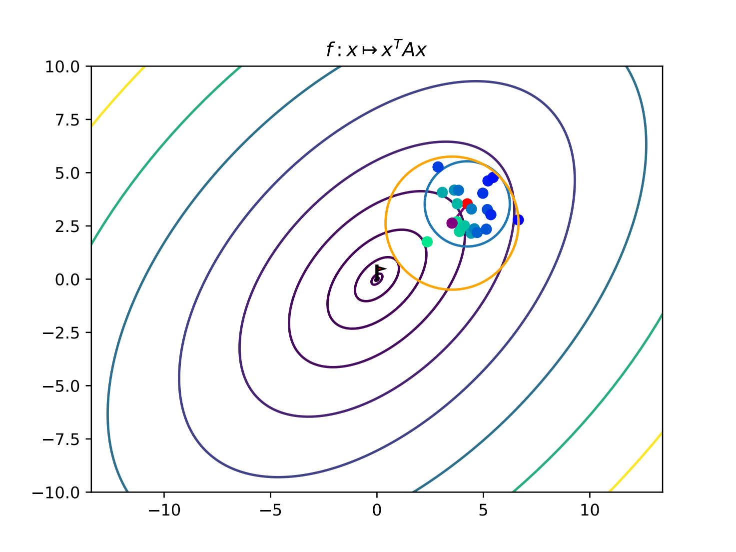
At each iteration $t \in \mathbb{N}$, given a **mean** m_t, a **stepsize** σ_t and a **covariance matrix** C_t :

1. Generate λ offspring $x^i_{t+1} \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$ *independently*
2. Rank the x^i_{t+1} w.r.t. their f-values: $f\left(x^{1:\lambda}_{t+1}\right) \leqslant \cdots \leqslant f\left(x^{\lambda:\lambda}_{t+1}\right)$
3. Update the mean: $m_{t+1} = \sum w_i x^{i:\lambda}_{t+1}$

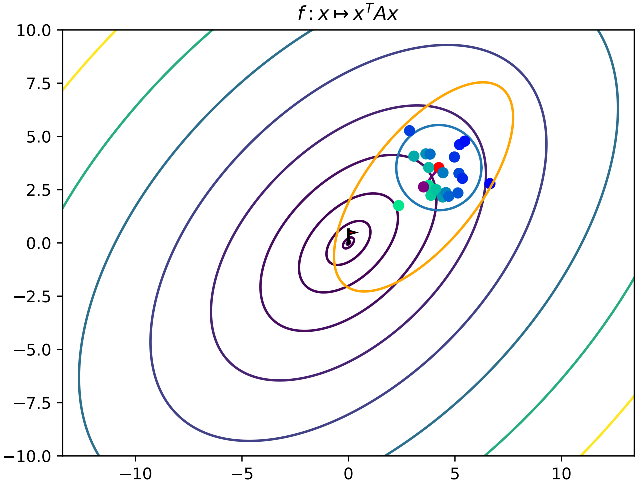
   The best offspring have the largest weights: $w_1 \geqslant w_2 \ldots$
4. Update the stepsize:
   Increase the stepsize if the path taken by the mean is larger than expected (assuming no selection)
5. **Update the covariance matrix**

$$C_{t+1} = (1-c)C_t + c \sum w_i \frac{(x^{i:\lambda}_{t+1} - m_t)(x^{i:\lambda}_{t+1} - m_t)^T}{\sigma_t^2}$$

Adapt the covariance matrix



Adapt the covariance matrix



CMA-ES : presentation

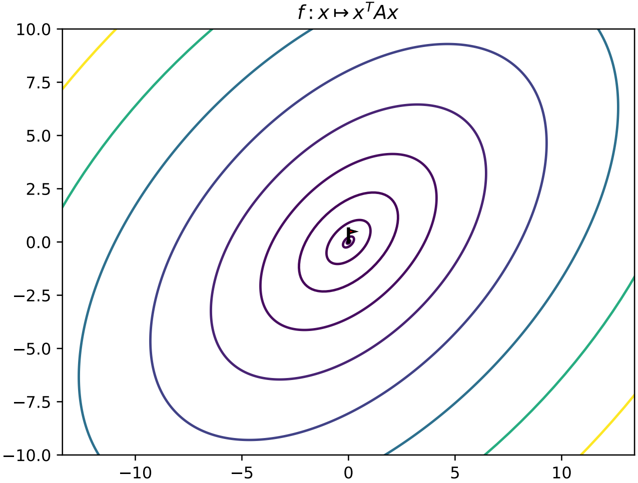
At each iteration $t \in \mathbb{N}$, given a **mean** m_t, a **stepsize** σ_t and a **covariance matrix** C_t :

1. Generate λ offspring $x^i_{t+1} \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$ *independently*
2. Rank the x^i_{t+1} w.r.t. their f-values: $f\left(x^{1:\lambda}_{t+1}\right) \leqslant \cdots \leqslant f\left(x^{\lambda:\lambda}_{t+1}\right)$
3. Update the mean: $m_{t+1} = \sum w_i x^{i:\lambda}_{t+1}$

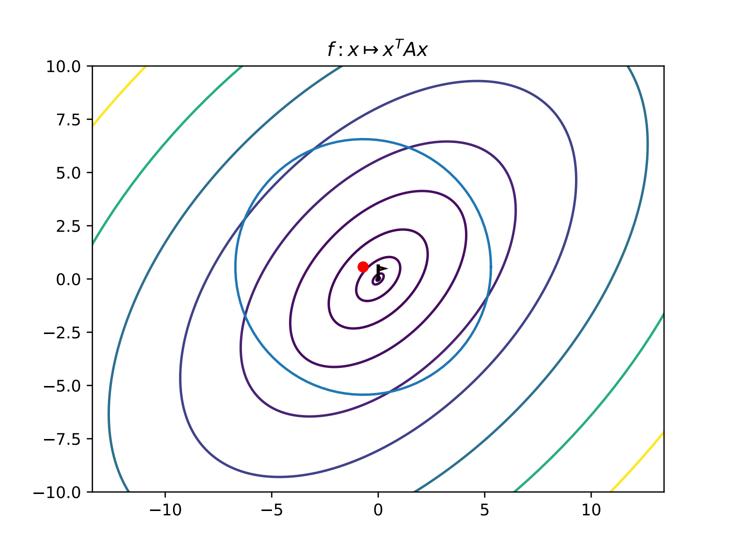
   The best offspring have the largest weights: $w_1 \geqslant w_2 \ldots$
4. Update the stepsize:
   Increase the stepsize if the path taken by the mean is larger than expected (assuming no selection)
5. Update the covariance matrix

$$C_{t+1} = (1-c)C_t + c \sum w_i \frac{(x^{i:\lambda}_{t+1} - m_t)(x^{i:\lambda}_{t+1} - m_t)^T}{\sigma_t^2}$$

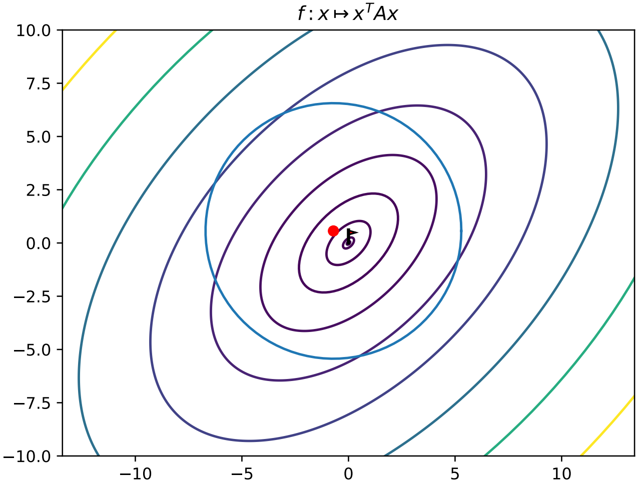
Start from a distribution $\mathcal{N}(m_t, \sigma_t^2 C_t)$



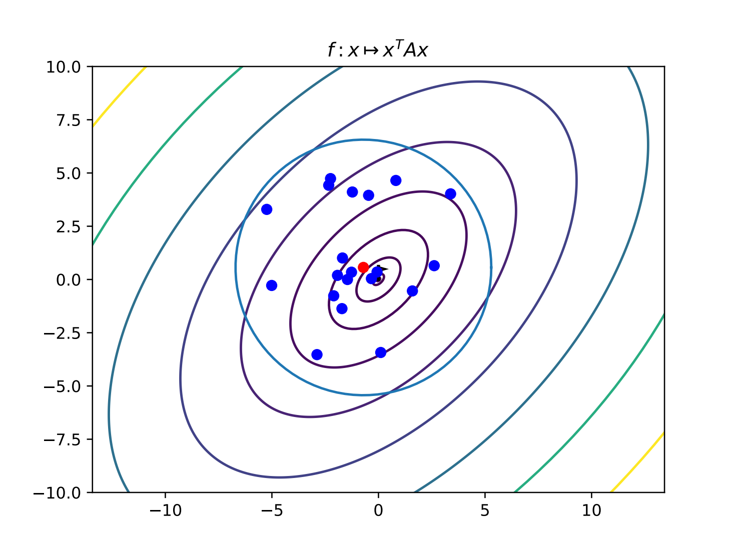
Start from a distribution $\mathcal{N}(m_t, \sigma_t^2 C_t)$



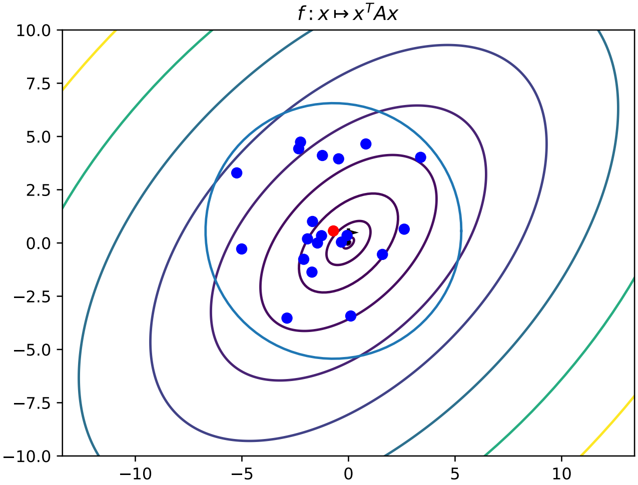
Generate λ offspring $x_{t+1}^{i} \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$ *independently*



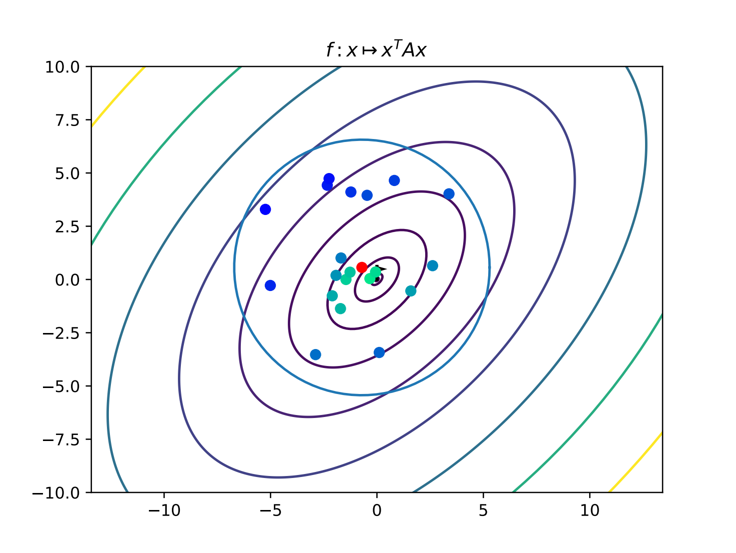
Generate λ offspring $x^i_{t+1} \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$ *independently*



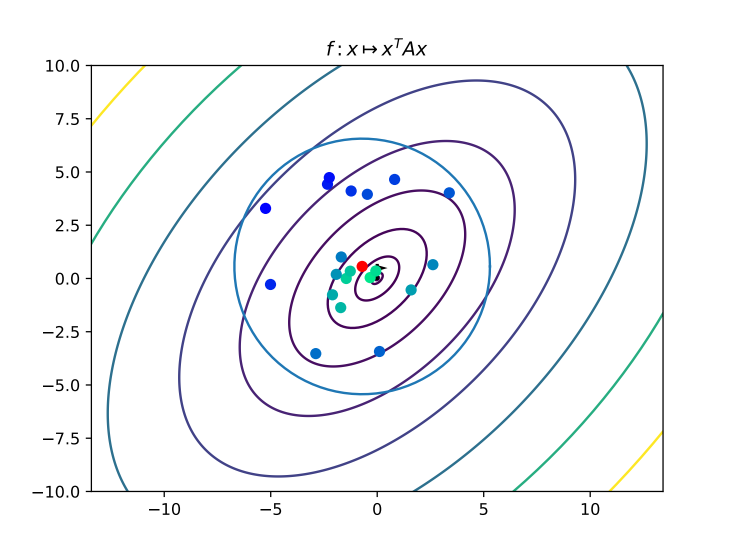
Rank the x_{t+1}^{i} w.r.t. their f-values $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$



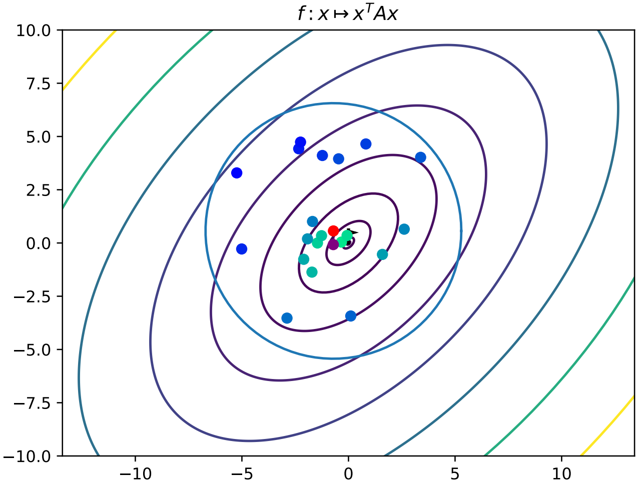
Rank the x_{t+1}^{i} w.r.t. their f-values $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$



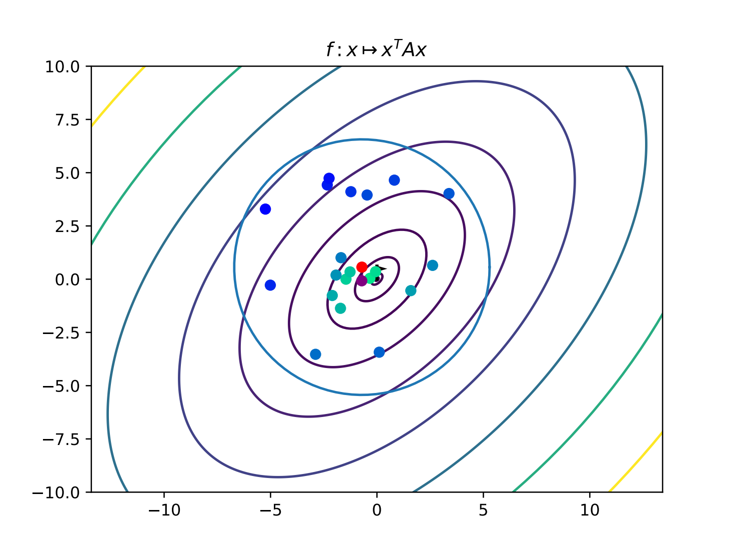
Update the mean $m_{t+1} = \sum w_i x_{t+1}^{i:\lambda}$



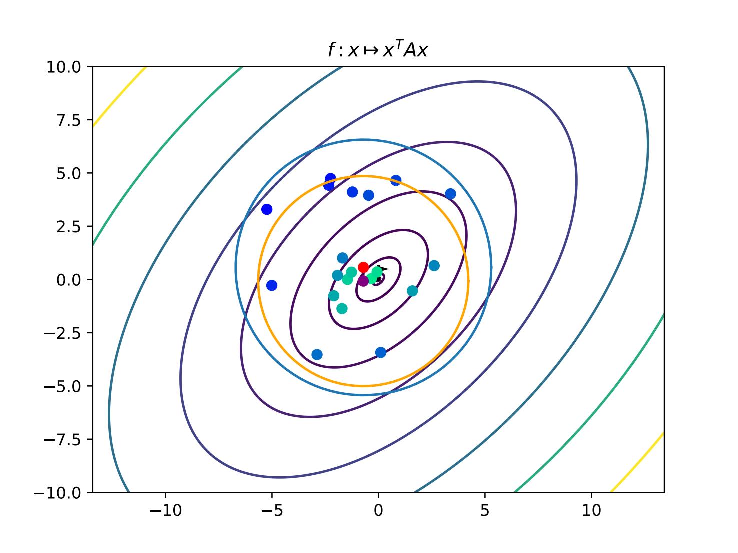
Update the mean $m_{t+1} = \sum w_i x_{t+1}^{i:\lambda}$



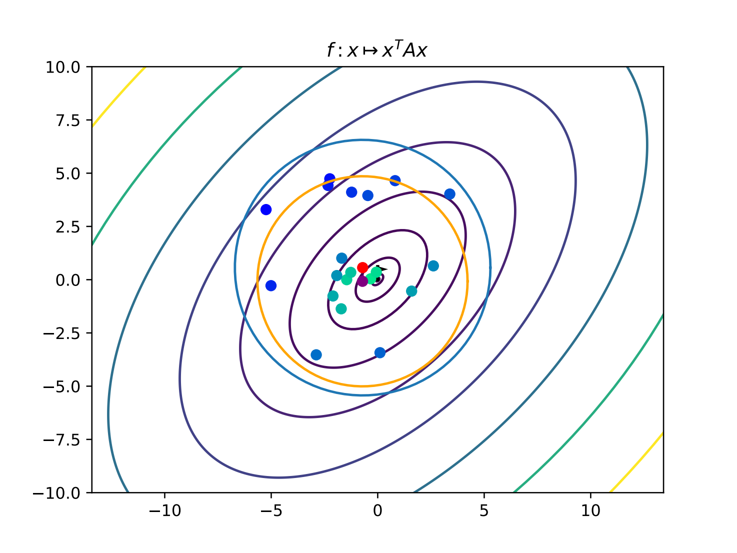
Adapt the stepsize



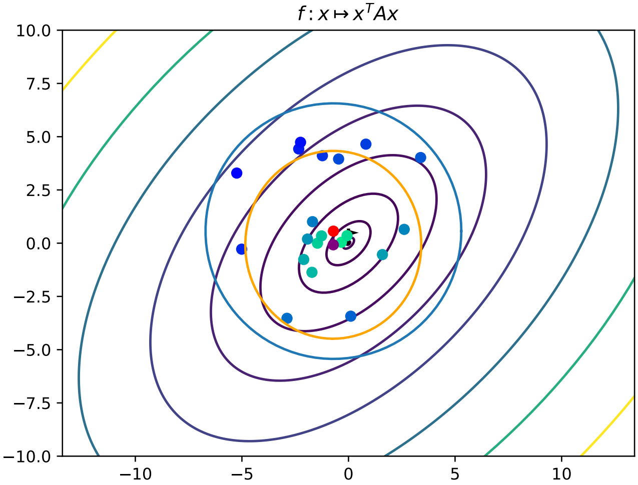
Adapt the stepsize



Adapt the covariance matrix

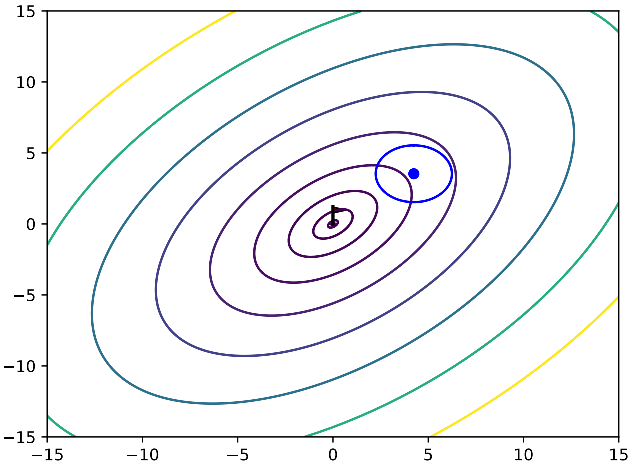


Adapt the covariance matrix

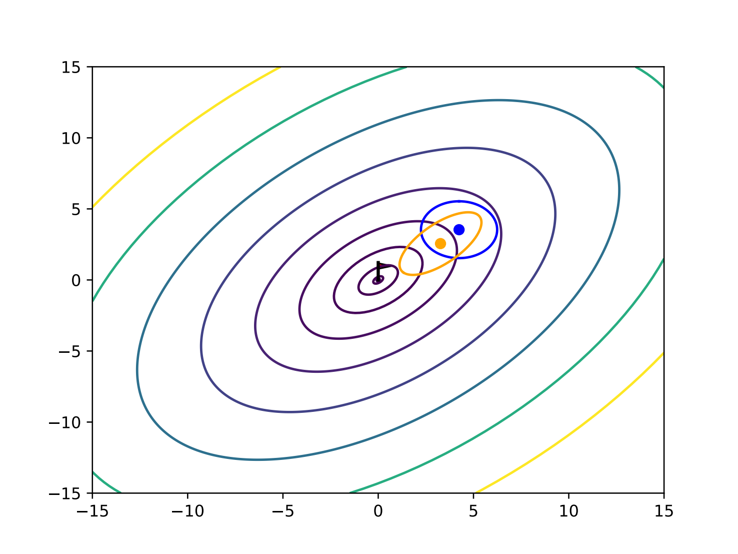


Linear convergence

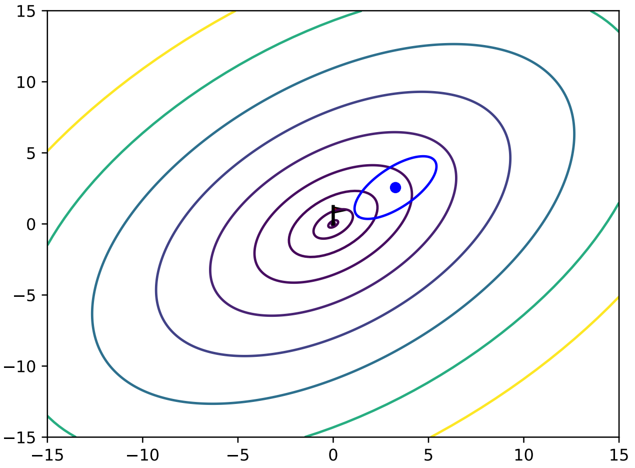
Convergence



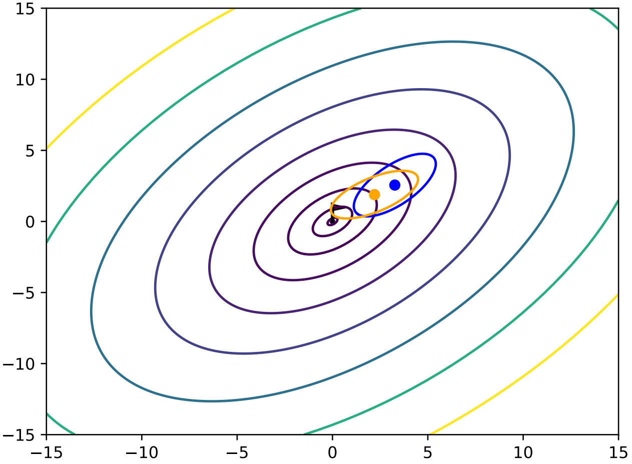
Convergence



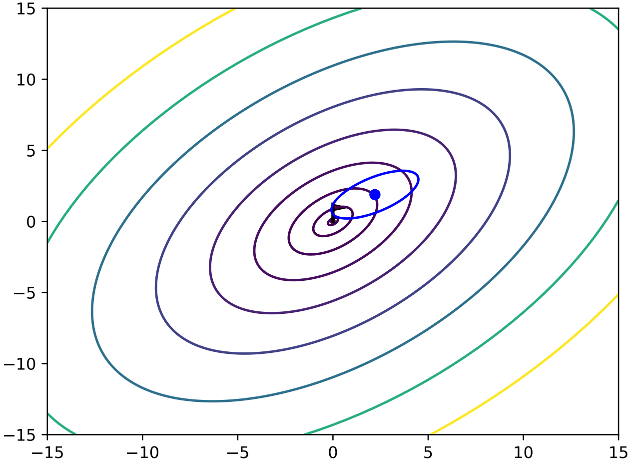
Convergence



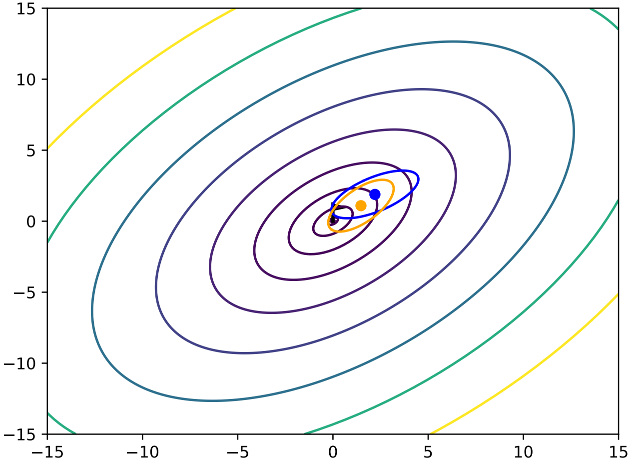
Convergence



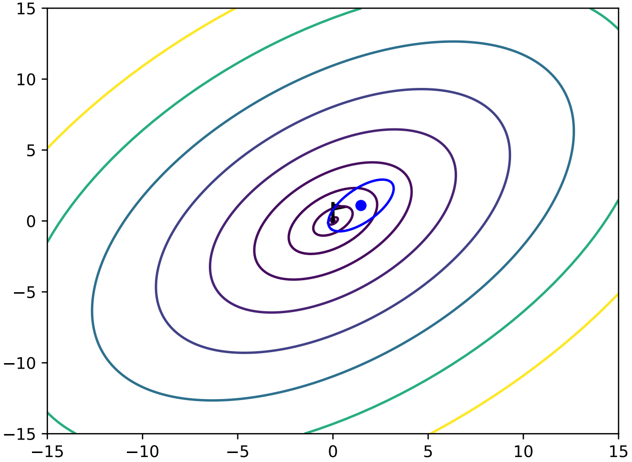
Convergence



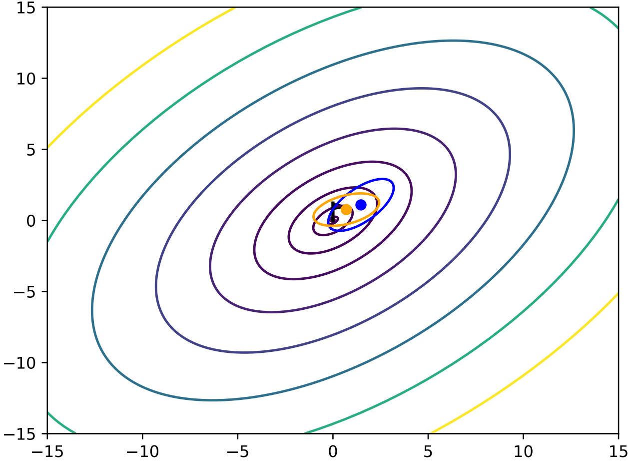
Convergence



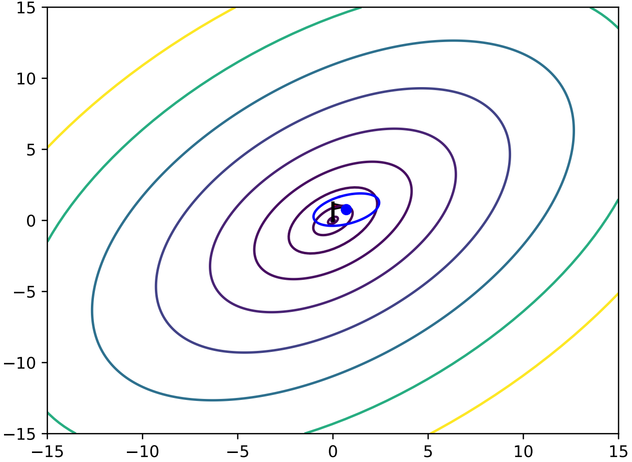
Convergence



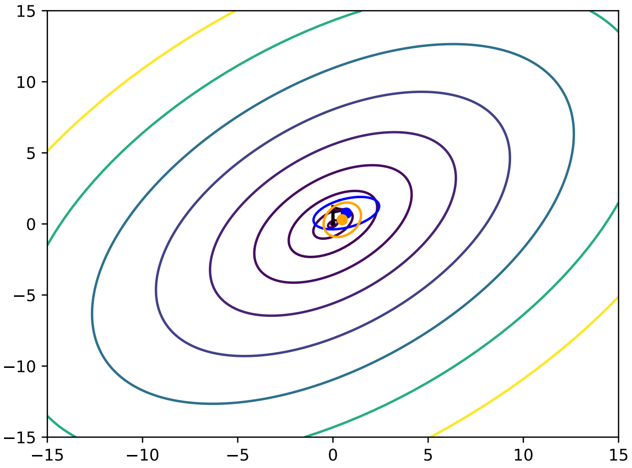
Convergence



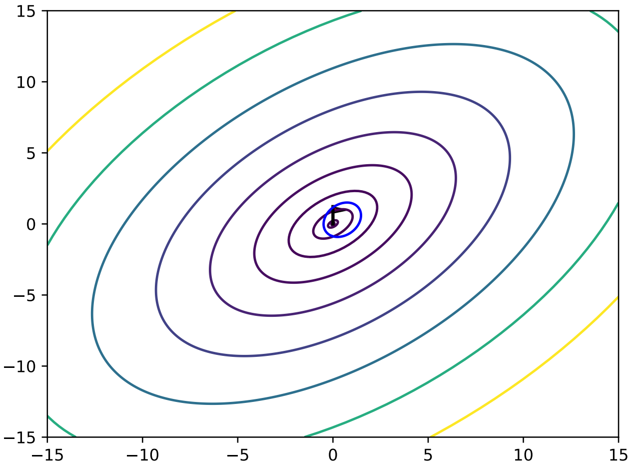
Convergence



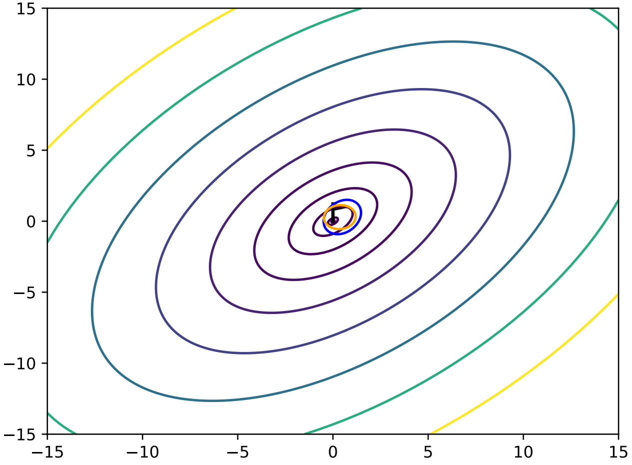
Convergence



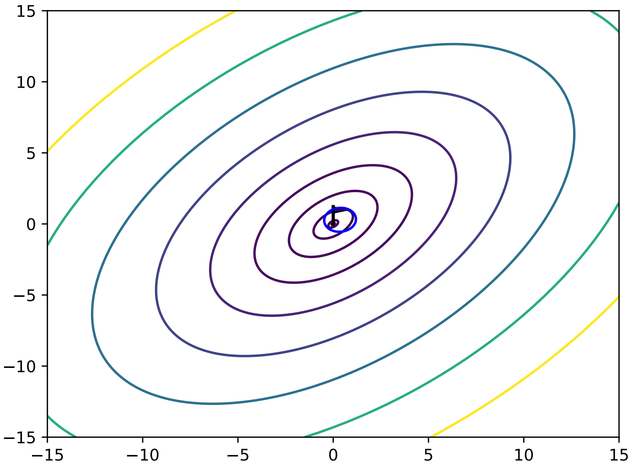
Convergence



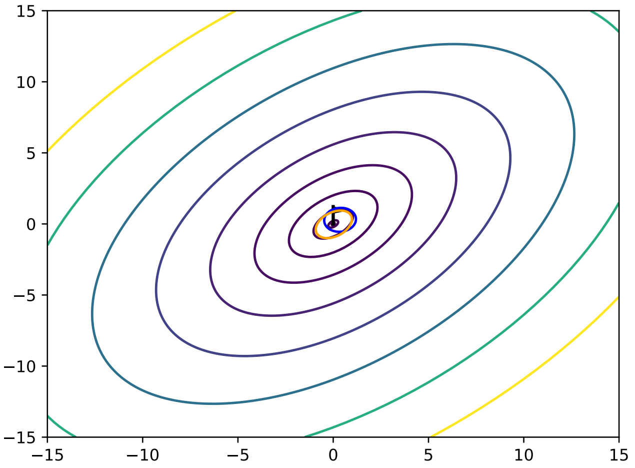
Convergence



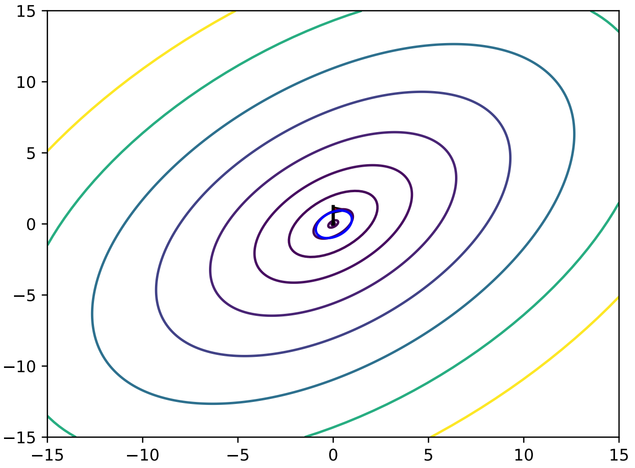
Convergence



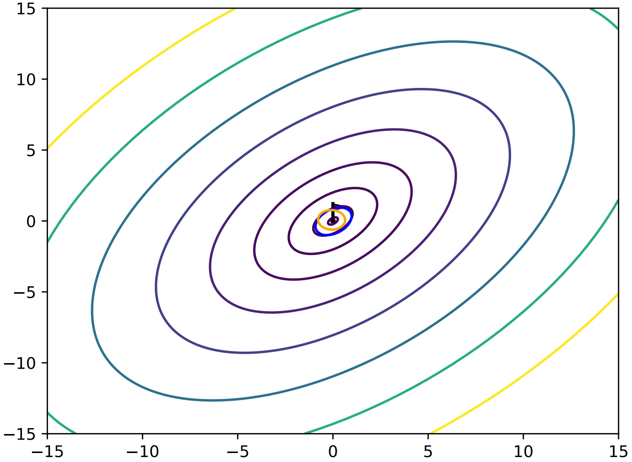
Convergence



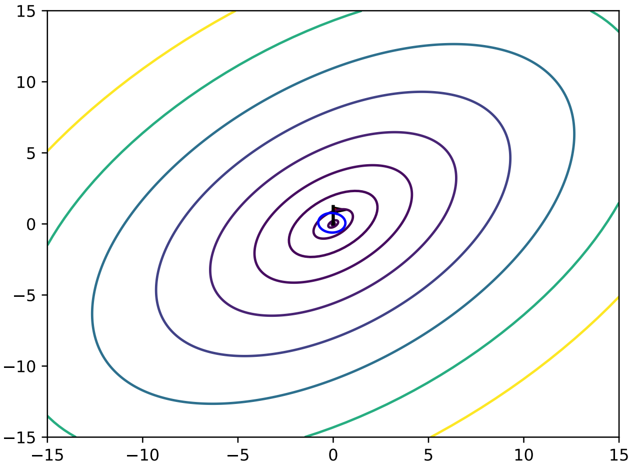
Convergence



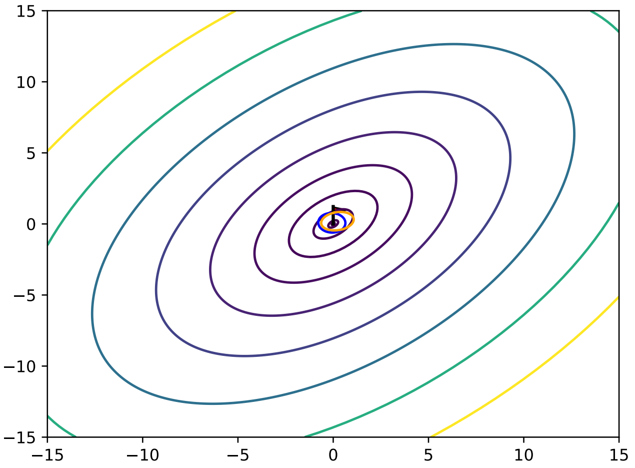
Convergence



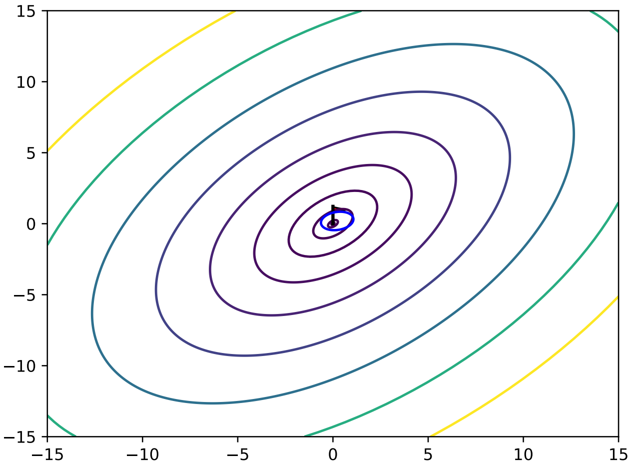
Convergence



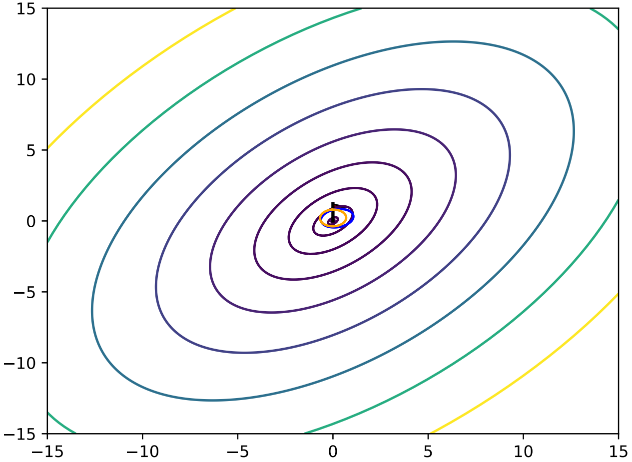
Convergence



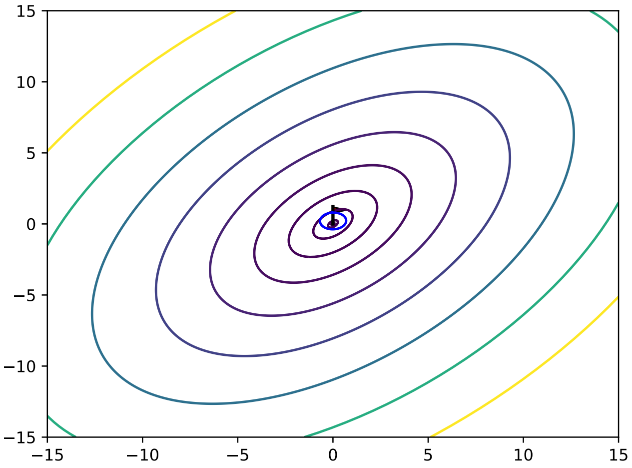
Convergence



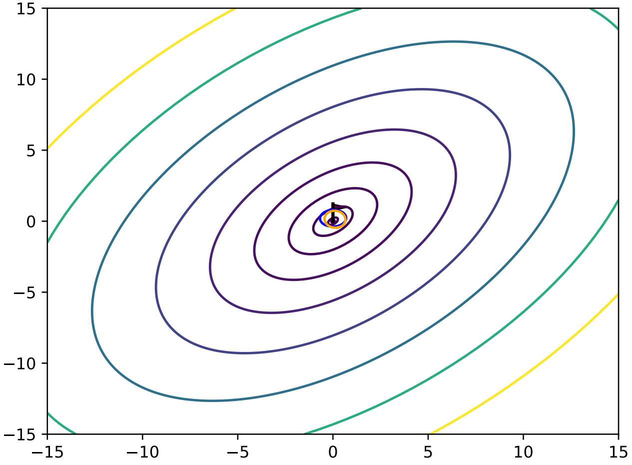
Convergence



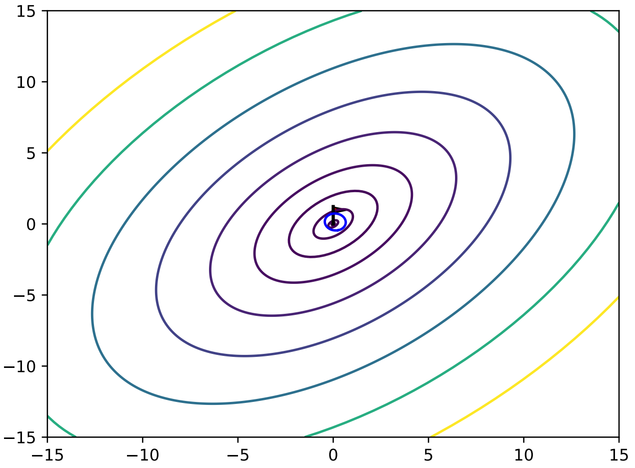
Convergence



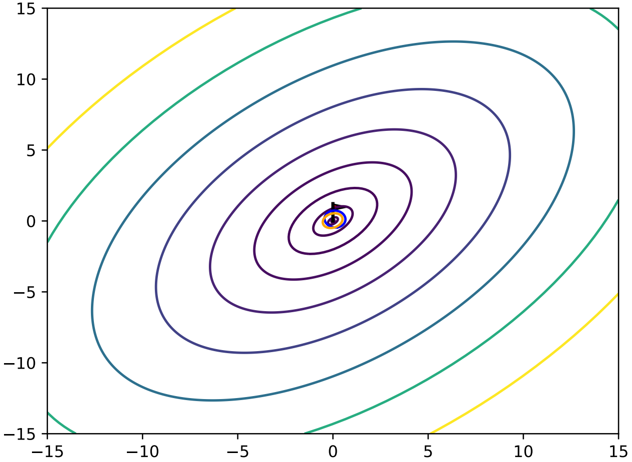
Convergence



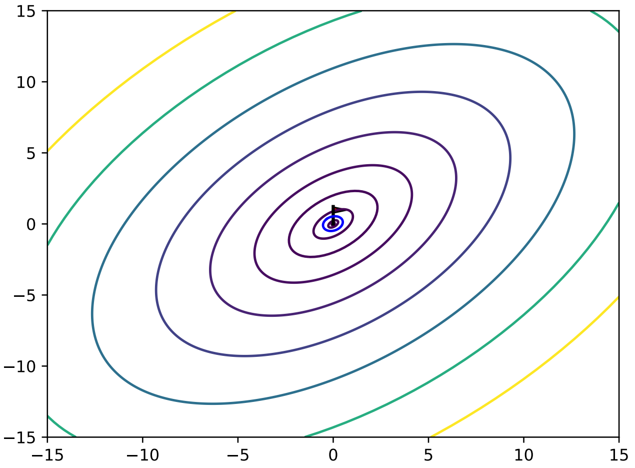
Convergence



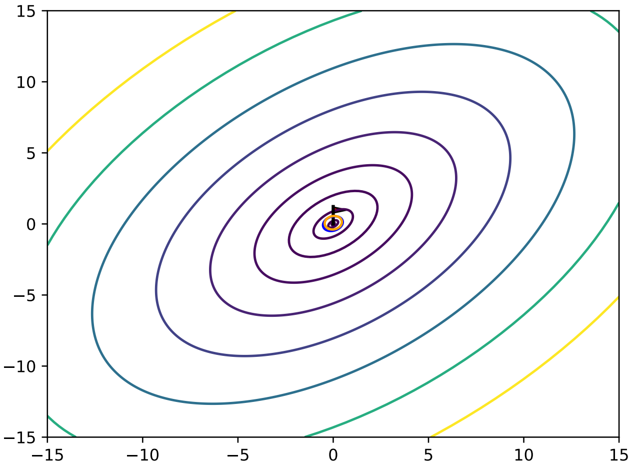
Convergence



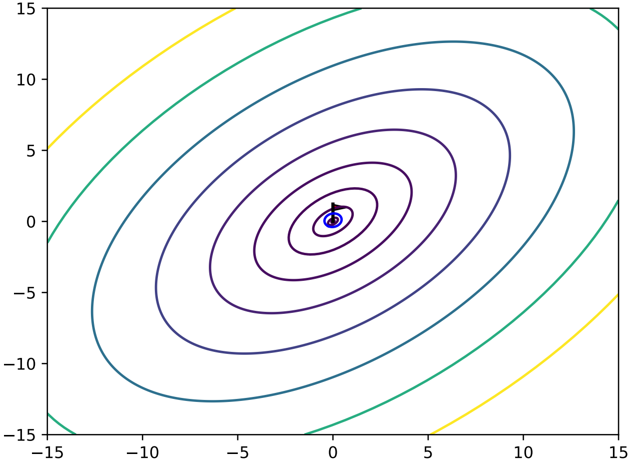
Convergence



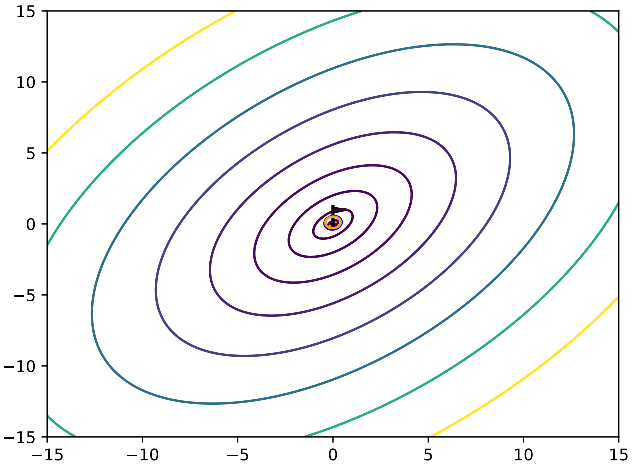
Convergence



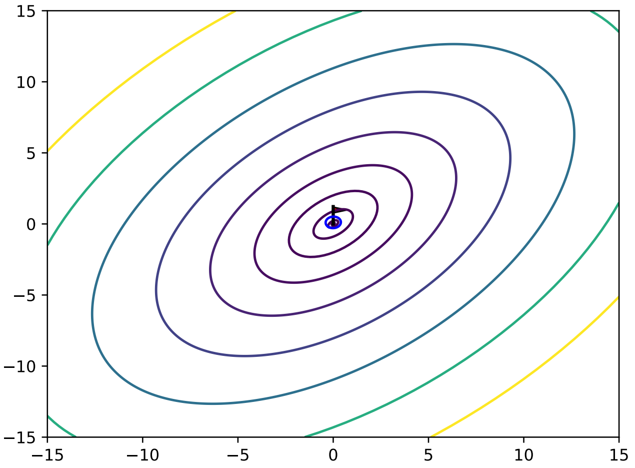
Convergence



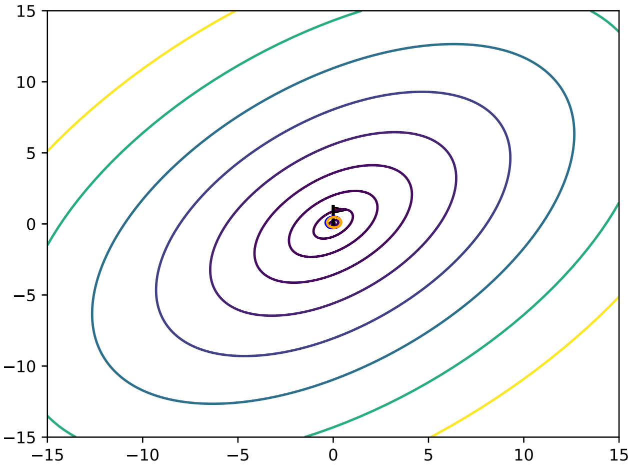
Convergence



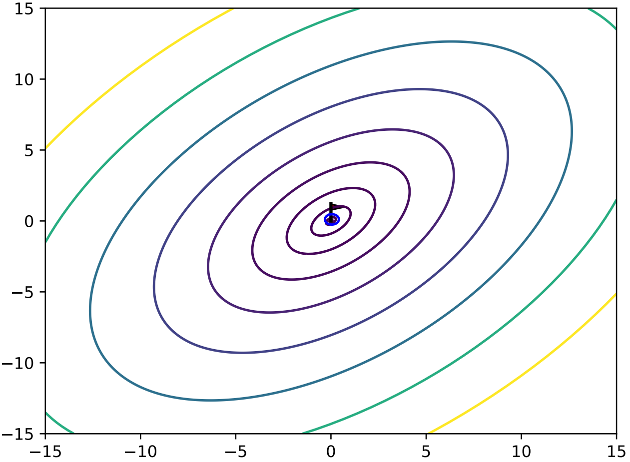
Convergence



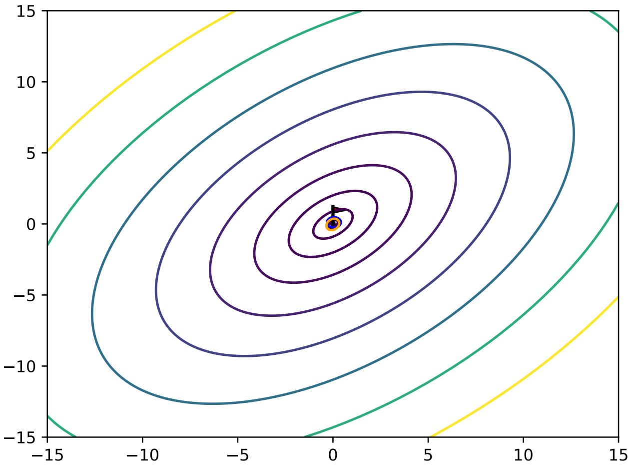
Convergence



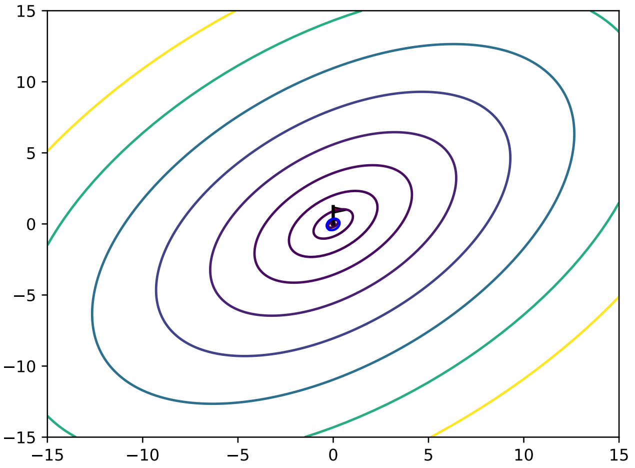
Convergence



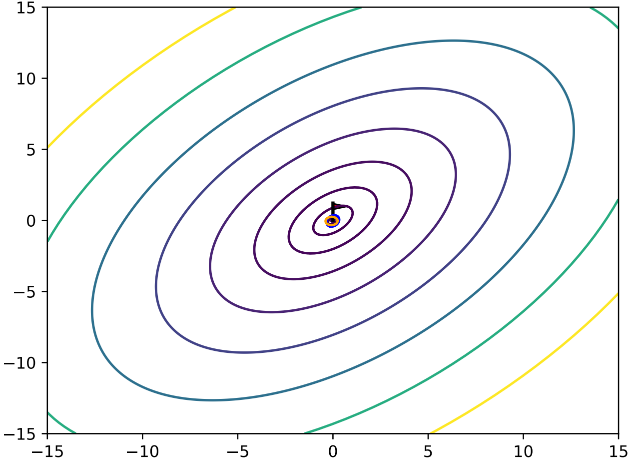
Convergence



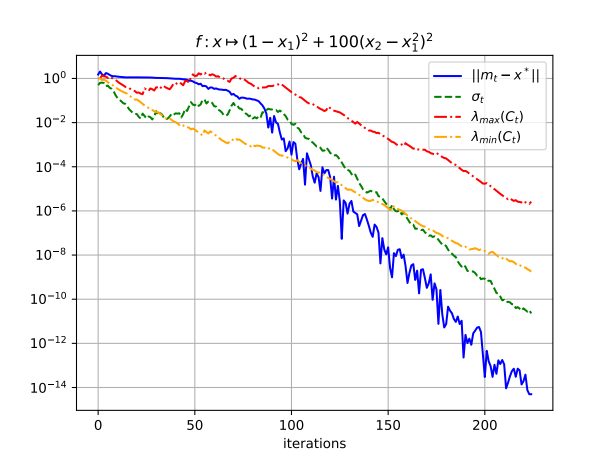
Convergence



Convergence

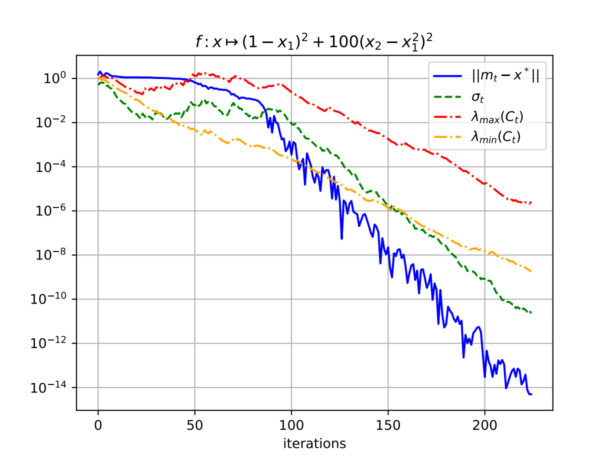


Linear convergence



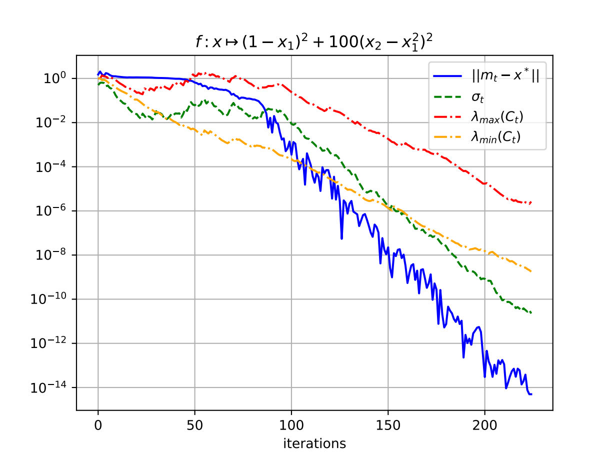
$$\frac{\|m_t - x^*\|}{\|m_0 - x^*\|}$$

Linear convergence



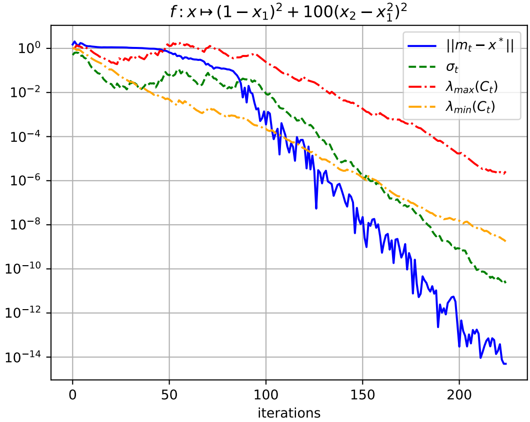
$$\log \frac{\|m_t - x^*\|}{\|m_0 - x^*\|}$$

Linear convergence



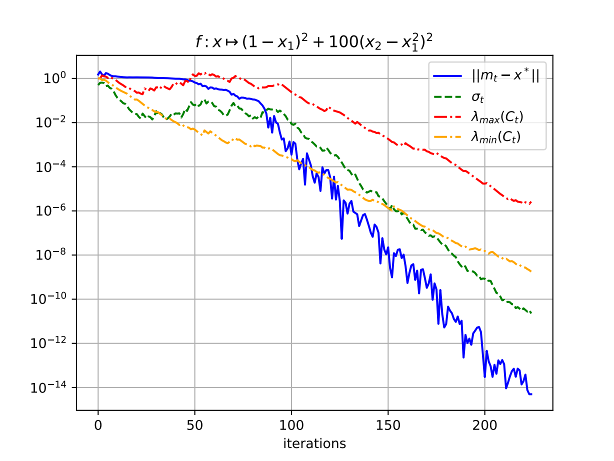
$$\log \frac{\|m_t - x^*\|}{\|m_0 - x^*\|} = -\mathrm{CR} \times t$$

Linear convergence



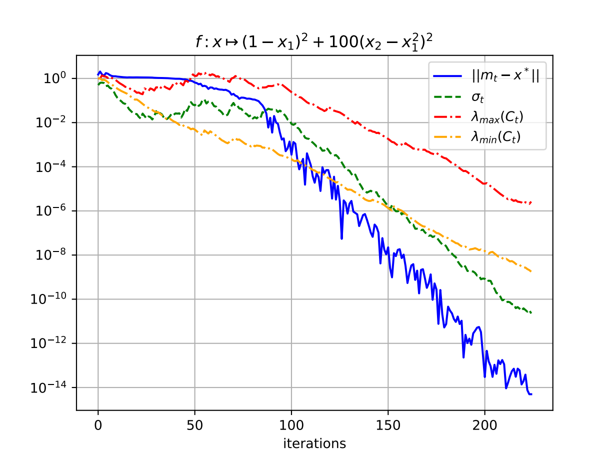
$$\frac{1}{t}\log\frac{\|m_t - x^*\|}{\|m_0 - x^*\|} = -\mathrm{CR}$$

Linear convergence



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

Linear convergence



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

Summary

Summary

- We have **invariance by increasing transformation**

Summary

- We have **invariance by increasing transformation**, i.e. if $g : \mathbb{R} \to \mathbb{R}$ is **increasing**

Summary

- We have **invariance by increasing transformation**, i.e. if $g : \mathbb{R} \to \mathbb{R}$ is **increasing**, then **minimizing** f **and** $g \circ f$ **will be equivalent**

CMA-ES : presentation

At each iteration $t \in \mathbb{N}$, given a **mean** m_t, a **stepsize** σ_t and a **covariance matrix** C_t :

1. Generate λ offspring $x_{t+1}^i \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$ *independently*
2. Rank the x_{t+1}^i w.r.t. their f-values: $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$
3. Update the mean: $m_{t+1} = \sum w_i x_{t+1}^{i:\lambda}$

   The best offspring have the largest weights: $w_1 \geqslant w_2 \ldots$
4. Update the stepsize:
   Increase the stepsize if the path taken by the mean is larger than expected (assuming no selection)
5. Update the covariance matrix

$$C_{t+1} = (1-c)C_t + c\sum w_i \frac{(x_{t+1}^{i:\lambda} - m_t)(x_{t+1}^{i:\lambda} - m_t)^T}{\sigma_t^2}$$

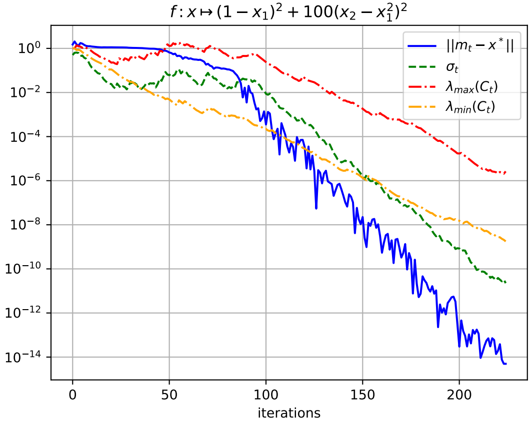
Summary

- We have **invariance by increasing transformation**, i.e. if $g : \mathbb{R} \to \mathbb{R}$ is **increasing**, then **minimizing** f **and** $g \circ f$ **will be equivalent**

Summary

- We have invariance by increasing transformation, i.e. if $g \colon \mathbb{R} \to \mathbb{R}$ is increasing, then minimizing f and $g \circ f$ will be equivalent
- We observe **linear convergence** $m_t \to x^*$

Linear convergence



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

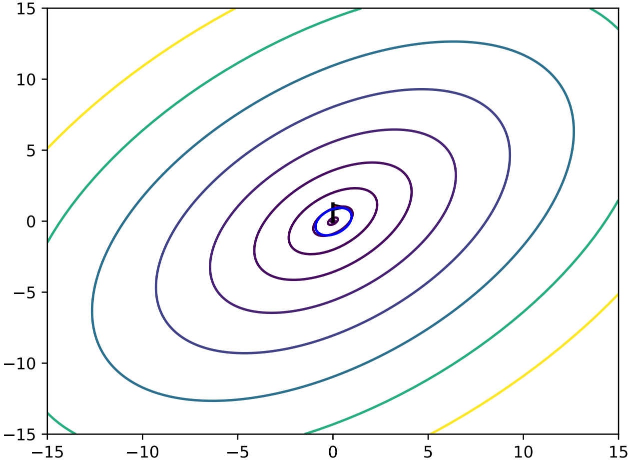
Summary

- We have invariance by increasing transformation, i.e. if $g\colon \mathbb{R} \to \mathbb{R}$ is increasing, then minimizing f and $g \circ f$ will be equivalent
- We observe **linear convergence** $m_t \to x^*$

Summary

- We have invariance by increasing transformation, i.e. if $g : \mathbb{R} \to \mathbb{R}$ is increasing, then minimizing f and $g \circ f$ will be equivalent
- We observe linear convergence $m_t \to x^*$
- The covariance matrix **learns the inverse Hessian of** f (when f is e.g. convex-quadratic)

Convergence



Summary

- We have invariance by increasing transformation, i.e. if $g\colon \mathbb{R} \to \mathbb{R}$ is increasing, then minimizing f and $g \circ f$ will be equivalent
- We observe linear convergence $m_t \to x^*$
- The covariance matrix **learns the inverse Hessian of** f (when f is e.g. convex-quadratic)

Summary

- We have invariance by increasing transformation, i.e. if $g : \mathbb{R} \to \mathbb{R}$ is increasing, then minimizing f and $g \circ f$ will be equivalent
- We observe linear convergence $m_t \to x^*$
- The covariance matrix learns the inverse Hessian of f (when f is e.g. convex-quadratic)

Goal:

Summary

- We have invariance by increasing transformation, i.e. if $g : \mathbb{R} \to \mathbb{R}$ is increasing, then minimizing f and $g \circ f$ will be equivalent
- We observe linear convergence $m_t \to x^*$
- The covariance matrix learns the inverse Hessian of f (when f is e.g. convex-quadratic)

Goal: proof of linear convergence

Summary

- We have invariance by increasing transformation, i.e. if $g : \mathbb{R} \to \mathbb{R}$ is increasing, then minimizing f and $g \circ f$ will be equivalent
- We observe linear convergence $m_t \to x^*$
- The covariance matrix learns the inverse Hessian of f (when f is e.g. convex-quadratic)

Goal: proof of linear convergence and of the learning of the inverse Hessian

Analysis via Markov chains

Markov chains

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

$$\mathrm{Distribution}\left(\phi_{t+1} \mid \phi_0, \ldots, \phi_t\right) = \mathrm{Distribution}\left(\phi_{t+1} \mid \phi_t\right)$$

Markov chains

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

$$\text{Distribution}\left(\phi_{t+1} \mid \phi_0, \ldots, \phi_t\right) = \text{Distribution}\left(\phi_{t+1} \mid \phi_t\right)$$

- The Markov chain $(\phi_t)_{t\in\mathbb{N}}$ is **irreducible** if all states[1] are reachable in finite time with positive probability.

[1] with positive measure (e.g. Lebesgue measure)

Markov chains

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

$$\text{Distribution}\left(\phi_{t+1} \mid \phi_0, \dots, \phi_t\right) = \text{Distribution}\left(\phi_{t+1} \mid \phi_t\right)$$

- The Markov chain $(\phi_t)_{t\in\mathbb{N}}$ is **irreducible** if all states[1] are reachable in finite time with positive probability.
- A probability distribution π is **invariant** for the Markov chain $(\phi_t)_{t\in\mathbb{N}}$ when

$$\phi_t \sim \pi \Rightarrow \phi_{t+1} \sim \pi$$

[1] with positive measure (e.g. Lebesgue measure)

Markov chains

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

$$\text{Distribution}\left(\phi_{t+1} \mid \phi_0, \ldots, \phi_t\right) = \text{Distribution}\left(\phi_{t+1} \mid \phi_t\right)$$

- The Markov chain $(\phi_t)_{t\in\mathbb{N}}$ is **irreducible** if all states[1] are reachable in finite time with positive probability.
- A probability distribution π is **invariant** for the Markov chain $(\phi_t)_{t\in\mathbb{N}}$ when

$$\phi_t \sim \pi \Rightarrow \phi_{t+1} \sim \pi$$

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

[1] with positive measure (e.g. Lebesgue measure)

Markov chains

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

$$\text{Distribution}\left(\phi_{t+1} \mid \phi_0, \ldots, \phi_t\right) = \text{Distribution}\left(\phi_{t+1} \mid \phi_t\right)$$

- The Markov chain $(\phi_t)_{t\in\mathbb{N}}$ is **irreducible** if all states[1] are reachable in finite time with positive probability.
- A probability distribution π is **invariant** for the Markov chain $(\phi_t)_{t\in\mathbb{N}}$ when

$$\phi_t \sim \pi \Rightarrow \phi_{t+1} \sim \pi$$

  (π is a "fixed point" for $(\phi_t)_{t\in\mathbb{N}}$)
- The Markov chain is **ergodic** when it satisfies the following LLN

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} f\left(\phi_t\right) = \int f(x)\pi(dx).$$

[1] with positive measure (e.g. Lebesgue measure)

CMA-ES as a Markov chain

Consider the random sequence

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

CMA-ES as a Markov chain

Consider the random sequence

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

This defines a Markov chain!

CMA-ES as a Markov chain

Consider the random sequence

$$\big(\underbrace{m_t}_{\text{mean}}, \overbrace{\sigma_t}^{\text{stepsize}}, \underbrace{C_t}_{\text{covariance matrix}}\big)$$

This defines a Markov chain!

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

Invariant measure for CMA-ES?

Suppose that $(m_t, \sigma_t, C_t)_{t\in\mathbb{N}}$ has an invariant measure π.

Invariant measure for CMA-ES?

Suppose that $(m_t, \sigma_t, C_t)_{t\in\mathbb{N}}$ has an invariant measure π.

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

Invariant measure for CMA-ES?

Suppose that $(m_t, \sigma_t, C_t)_{t\in\mathbb{N}}$ has an invariant measure π.

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

We do not progress towards the optimum anymore!

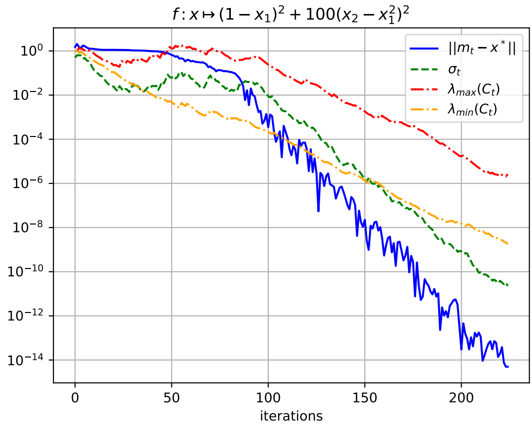
Invariant measure for CMA-ES?

Suppose that $(m_t, \sigma_t, C_t)_{t \in \mathbb{N}}$ has an invariant measure π.

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

We do not progress towards the optimum anymore! The **existence of an invariant measure** seems to be **incompatible** with the **convergence** to the optimum.

Linear convergence



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

Normalization

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

Normalization

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

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

Normalization

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

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

The sequence $(Z_t)_{t\in\mathbb{N}}$ could eventually become **stationary**

Normalization

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

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

The sequence $(Z_t)_{t \in \mathbb{N}}$ could eventually become stationary

Proposition (Normalized Markov chain)

The sequence

$$\left(Z_t, \frac{C_t}{\lambda_{\min}(C_t)} \right)_{t \in \mathbb{N}}$$

defines a Markov chain.

Normalization

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

$$Z_t \stackrel{\mathrm{def}}{=} \frac{m_t - x^*}{\sigma_t \sqrt{\lambda_{\min}(C_t)}}$$

The sequence $(Z_t)_{t\in\mathbb{N}}$ could eventually become stationary

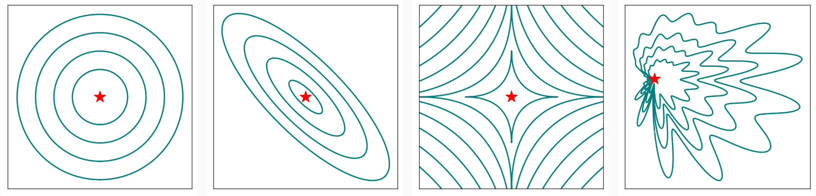
Proposition (Normalized Markov chain)

The sequence

$$\left(Z_t, \frac{C_t}{\lambda_{\min}(C_t)}\right)_{t\in\mathbb{N}}$$

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

Scaling-invariant functions

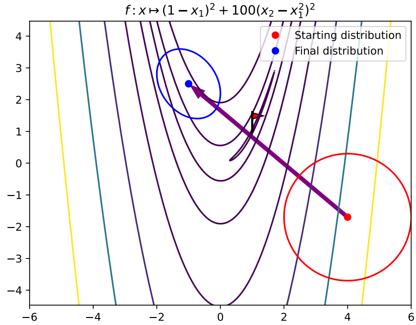


Irreducibility of CMA-ES

A Markov chain $(\phi_t)_{t\in\mathbb{N}}$ is **irreducible** if all states are reachable in finite time with positive probability.

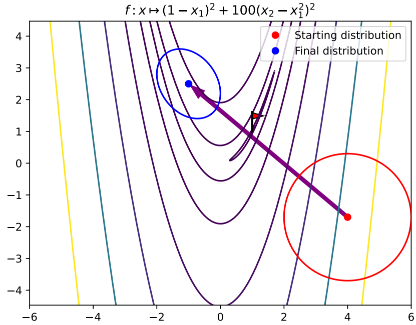
Irreducibility of CMA-ES

A Markov chain $(\phi_t)_{t\in\mathbb{N}}$ is **irreducible** if all states are reachable in finite time with positive probability.



Irreducibility of CMA-ES

A Markov chain $(\phi_t)_{t\in\mathbb{N}}$ is **irreducible** if all states are reachable in finite time with positive probability.



Given a starting and a final distributions, **can we reach the final distribution in finite time with positive probability?**

Theorem (Irreducibility of the normalized chain)

When minimizing a scaling-invariant function f with Lebesgue-negligible level sets, the sequence

$$\left(Z_t, \frac{C_t}{\lambda_{\min}(C_t)}\right)_{t\in\mathbb{N}}$$

defines a irreducible, aperiodic Markov chain, and compact sets are small sets.

Ergodicity of the normalized chain[2]

$$\left(Z_t, \frac{C_t}{\lambda_{\min}(C_t)}\right)_{t\in\mathbb{N}}$$

is **ergodic** if

[2]Sean P. Meyn and Richard L. Tweedie. **Markov Chains and Stochastic Stability.** Springer Science & Business Media, Dec. 2012. ISBN: 978-1-4471-3267-7.

Ergodicity of the normalized chain[2]

$$\left(Z_t, \frac{C_t}{\lambda_{\min}(C_t)}\right)_{t\in\mathbb{N}}$$

is **ergodic** (and satisfies a LLN) if

[2] Sean P. Meyn and Richard L. Tweedie. **Markov Chains and Stochastic Stability.** Springer Science & Business Media, Dec. 2012. ISBN: 978-1-4471-3267-7.

Ergodicity of the normalized chain[2]

$$\left(Z_t, \frac{C_t}{\lambda_{\min}(C_t)}\right)_{t\in\mathbb{N}}$$

is **ergodic** (and satisfies a LLN) if

- it is **irreducible** and aperiodic

[2]Sean P. Meyn and Richard L. Tweedie. **Markov Chains and Stochastic Stability.** Springer Science & Business Media, Dec. 2012. ISBN: 978-1-4471-3267-7.

Ergodicity of the normalized chain[2]

$$\left(Z_t, \frac{C_t}{\lambda_{\min}(C_t)}\right)_{t\in\mathbb{N}}$$

is **ergodic** (and satisfies a LLN) if

- it is irreducible and aperiodic
- it satisfies the following **drift** condition: $\exists V : X \to [0, +\infty]$

$$\mathbb{E}_t\left[V\left(Z_{t+1}, \frac{C_{t+1}}{\lambda_{\min}(C_{t+1})}\right)\right] \leqslant (1-\varepsilon) V\left(Z_t, \frac{C_t}{\lambda_{\min}(C_t)}\right)$$

[2] Sean P. Meyn and Richard L. Tweedie. **Markov Chains and Stochastic Stability.** Springer Science & Business Media, Dec. 2012. ISBN: 978-1-4471-3267-7.

Ergodicity of the normalized chain[2]

$$\left(Z_t, \frac{C_t}{\lambda_{\min}(C_t)}\right)_{t\in\mathbb{N}}$$

is **ergodic** (and satisfies a LLN) if

- it is irreducible and aperiodic
- it satisfies the following drift condition: $\exists V : X \to [0, +\infty]$

$$\mathbb{E}_t\left[V\left(Z_{t+1}, \frac{C_{t+1}}{\lambda_{\min}(C_{t+1})}\right)\right] \leqslant (1-\varepsilon)V\left(Z_t, \frac{C_t}{\lambda_{\min}(C_t)}\right)$$

for $\left(Z_t, \frac{C_t}{\lambda_{\min}(C_t)}\right)$ **outside of a compact** $\mathcal{K}$.

[2]Sean P. Meyn and Richard L. Tweedie. **Markov Chains and Stochastic Stability.** Springer Science & Business Media, Dec. 2012. ISBN: 978-1-4471-3267-7.

Ergodicity of the normalized chain[2]

$$\left(Z_t, \frac{C_t}{\lambda_{\min}(C_t)}\right)_{t\in\mathbb{N}}$$

is **ergodic** (and satisfies a LLN) if

- it is irreducible and aperiodic
- it satisfies the following drift condition: $\exists V : X \to [0, +\infty]$

$$\mathbb{E}_t\left[V\left(Z_{t+1}, \frac{C_{t+1}}{\lambda_{\min}(C_{t+1})}\right)\right] \leqslant (1-\varepsilon)V\left(Z_t, \frac{C_t}{\lambda_{\min}(C_t)}\right)$$

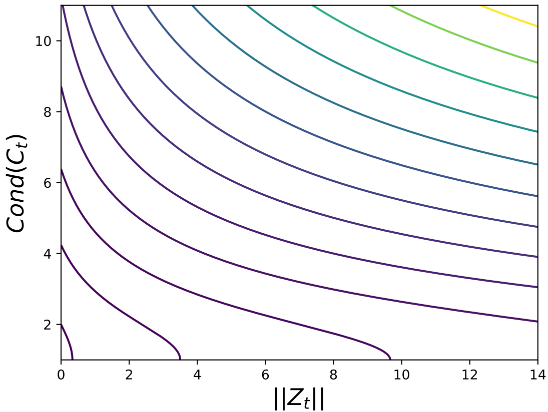
for $\left(Z_t, \frac{C_t}{\lambda_{\min}(C_t)}\right)$ outside of a compact $\mathcal{K}$.

The function V is called the **potential function** or the **drift function** or the **Lyapunov function**.

[2]Sean P. Meyn and Richard L. Tweedie. **Markov Chains and Stochastic Stability.** Springer Science & Business Media, Dec. 2012. ISBN: 978-1-4471-3267-7.

Drift condition

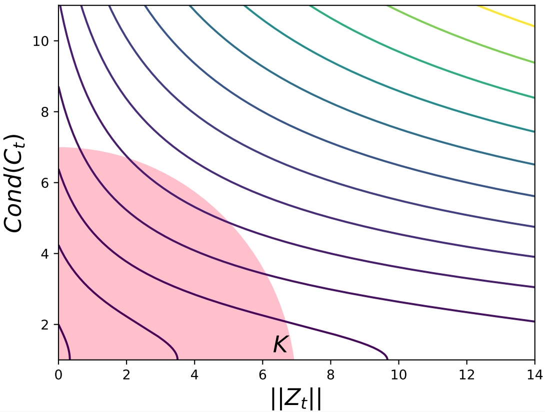
$$\mathbb{E}_t\left[V\left(Z_{t+1}, \frac{C_{t+1}}{\lambda_{\min}(C_{t+1})}\right)\right] \leqslant (1-\varepsilon)V\left(Z_t, \frac{C_t}{\lambda_{\min}(C_t)}\right)$$



Drift condition

$$\mathbb{E}_t\left[V\left(Z_{t+1}, \frac{C_{t+1}}{\lambda_{\min}(C_{t+1})}\right)\right] \leqslant (1-\varepsilon)V\left(Z_t, \frac{C_t}{\lambda_{\min}(C_t)}\right)$$

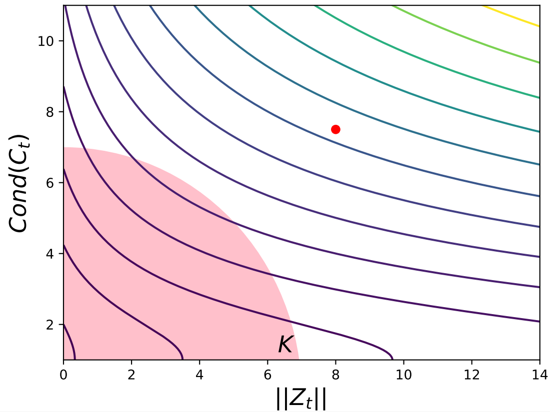
outside of a compact K



Drift condition

$$\mathbb{E}_t \left[V \left(Z_{t+1}, \frac{C_{t+1}}{\lambda_{\min}(C_{t+1})} \right) \right] \leqslant (1-\varepsilon) V \left(Z_t, \frac{C_t}{\lambda_{\min}(C_t)} \right)$$

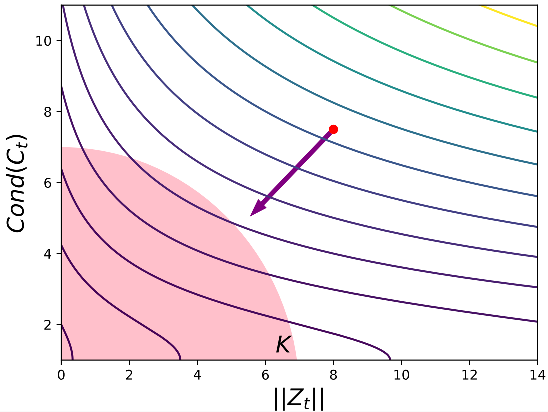
outside of a compact K



Drift condition

$$\mathbb{E}_t\left[V\left(Z_{t+1}, \frac{C_{t+1}}{\lambda_{\min}(C_{t+1})}\right)\right] \leqslant (1-\varepsilon)V\left(Z_t, \frac{C_t}{\lambda_{\min}(C_t)}\right)$$

outside of a compact K



Theorem (Drift condition for the normalized chain)

When minimizing a spherical function $f: x \mapsto g\left(x^T x\right)$ $(g: \mathbb{R} \to \mathbb{R}$ *increasing), then the irreducible, aperiodic Markov chain* $(Z_t, C_t/\lambda_{\min}(C_t))_{t\in\mathbb{N}}$ *satisfies a Foster-Lyapunov condition with the potential defined by*

$$V(Z,C) = \sum_{k=1}^{d} \left\{ \frac{\lambda_k(C)}{\lambda_1(C)} |\langle v_k(C), Z\rangle|^2 \right\} + \beta \times \mathrm{Cond}(C)$$

Theorem (Drift condition for the normalized chain)

When minimizing a spherical function $f : x \mapsto g\left(x^T x\right)$ *(*$g : \mathbb{R} \to \mathbb{R}$ *increasing), then the irreducible, aperiodic Markov chain* $(Z_t, C_t/\lambda_{\min}(C_t))_{t\in\mathbb{N}}$ *satisfies a Foster-Lyapunov condition with the potential defined by*

$$V(Z, C) = \sum_{k=1}^{d} \left\{ \frac{\lambda_k(C)}{\lambda_1(C)} |\langle v_k(C), Z\rangle|^2 \right\} + \beta \times \mathrm{Cond}(C)$$

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

Affine-Invariance

$$
\begin{array}{ccc}
(m_0, \sigma_0, C_0) & \xrightarrow{\min f(x)} & (m_1, \sigma_1, C_1) \\
\Psi \downarrow & & \uparrow \Psi^{-1} \\
(m_0', \sigma_0', C_0') & \xrightarrow{\min f(Bx + b)} & (m_1', \sigma_1', C_1')
\end{array}
$$

Proof of linear convergence

$$\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} \to -\mathrm{CR}?$$

Proof of linear convergence

$$\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \frac{1}{T}\sum_{t=0}^{T-1}\log\frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|}$$

Proof of linear convergence

$$\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \frac{1}{T}\sum_{t=0}^{T-1}\log\frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|}$$

$$=\frac{1}{T}\sum_{t=0}^{T-1}\log\|m_{t+1} - x^*\| - \log\|m_t - x^*\|$$

Proof of linear convergence

$$\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \frac{1}{T}\sum_{t=0}^{T-1}\log\frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|}$$

$$= \frac{1}{T}\sum_{t=0}^{T-1}\log\|m_{t+1} - x^*\| - \log\|m_t - x^*\|$$

Proof of linear convergence

$$\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \frac{1}{T}\sum_{t=0}^{T-1}\log\frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|}$$

$$= \frac{1}{T}\sum_{t=0}^{T-1}\log\frac{\|m_{t+1} - x^*\|}{\sigma_{t+1}\sqrt{\lambda_{\min}(C_{t+1})}} + \log\sigma_{t+1}$$

$$+ \frac{1}{2}\log\lambda_{\min}(C_{t+1}) - \log\|m_t - x^*\|$$

Proof of linear convergence

$$\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \frac{1}{T}\sum_{t=0}^{T-1}\log\frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|}$$

$$= \frac{1}{T}\sum_{t=0}^{T-1}\log\|Z_{t+1}\| + \log\sigma_{t+1}$$

$$+ \frac{1}{2}\log\lambda_{\min}(C_{t+1}) - \log\|m_t - x^*\|$$

Proof of linear convergence

$$\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \frac{1}{T}\sum_{t=0}^{T-1}\log\frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|}$$

$$= \frac{1}{T}\sum_{t=0}^{T-1}\log\|Z_{t+1}\| + \log\sigma_{t+1}$$

$$+ \frac{1}{2}\log\lambda_{\text{min}}(C_{t+1}) - \log\|m_t - x^*\|$$

Proof of linear convergence

$$\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \frac{1}{T}\sum_{t=0}^{T-1}\log\frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|}$$

$$= \frac{1}{T}\sum_{t=0}^{T-1}\log\|Z_{t+1}\| + \log\sigma_{t+1}$$

$$+ \frac{1}{2}\log\lambda_{\min}(C_{t+1}) - \log\|m_t - x^*\|$$

Proof of linear convergence

$$\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \frac{1}{T}\sum_{t=0}^{T-1}\log\frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|}$$

$$= \frac{1}{T}\sum_{t=0}^{T-1}\log\|Z_{t+1}\| - \log\|Z_t\|$$

$$+ \log\frac{\sigma_{t+1}}{\sigma_t} + \frac{1}{2}\log\frac{\lambda_{\min}(C_{t+1})}{\lambda_{\min}(C_t)}$$

Proof of linear convergence

$$
\begin{aligned}
\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} &= \frac{1}{T}\sum_{t=0}^{T-1}\log\frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|} \\
=&\frac{1}{T}\sum_{t=0}^{T-1}\log\|Z_{t+1}\| - \log\|Z_t\| \\
&+ \log\frac{\sigma_{t+1}}{\sigma_t} + \frac{1}{2}\log\frac{\lambda_{\min}(C_{t+1})}{\lambda_{\min}(C_t)}
\end{aligned}
$$

Proof of linear convergence

$$
\begin{aligned}
\lim_{T\to\infty} \frac{1}{T} \log \frac{\|m_T - x^*\|}{\|m_0 - x^*\|} &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \log \frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|} \\
&= \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \log \|Z_{t+1}\| - \log \|Z_t\| \\
&\quad + \log \frac{\sigma_{t+1}}{\sigma_t} + \frac{1}{2} \log \frac{\lambda_{\min}(C_{t+1})}{\lambda_{\min}(C_t)}
\end{aligned}
$$

Proof of linear convergence

$$\lim_{T\to\infty} \frac{1}{T} \log \frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \log \frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|}$$

$$= \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \log \|Z_{t+1}\| - \log \|Z_t\|$$

$$+ \log \frac{\sigma_{t+1}}{\sigma_t} + \frac{1}{2} \log \frac{\lambda_{\min}(C_{t+1})}{\lambda_{\min}(C_t)}$$

Proof of linear convergence

$$\lim_{T\to\infty} \frac{1}{T} \log \frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \log \frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|}$$

$$= \int \log \|z\| \, d\pi - \int \log \|z\| \, d\pi$$

$$+ \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \log \frac{\sigma_{t+1}}{\sigma_t} + \frac{1}{2} \log \frac{\lambda_{\min}(C_{t+1})}{\lambda_{\min}(C_t)}$$

Proof of linear convergence

$$\lim_{T\to\infty}\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\log\frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|}$$

$$= \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\log\frac{\sigma_{t+1}}{\sigma_t} + \frac{1}{2}\log\frac{\lambda_{\min}(C_{t+1})}{\lambda_{\min}(C_t)}$$

Proof of linear convergence

$$\lim_{T\to\infty}\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\log\frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|}$$

$$= \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\log\frac{\sigma_{t+1}}{\sigma_t} + \frac{1}{2}\log\frac{\lambda_{\min}(C_{t+1})}{\lambda_{\min}(C_t)}$$

Proof of linear convergence

$$\lim_{T\to\infty} \frac{1}{T} \log \frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \log \frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|}$$

$$= \lim_{T\to\infty} \frac{1}{T} \log \frac{\sigma_T}{\sigma_0} + \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{2} \log \frac{\lambda_{\min}(C_{t+1})}{\lambda_{\min}(C_t)}$$

Proof of linear convergence

$$\lim_{T\to\infty}\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\log\frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|}$$

$$= \lim_{T\to\infty}\frac{1}{T}\log\frac{\sigma_T}{\sigma_0} + \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{2}\log\frac{\lambda_{\min}(C_{t+1})}{\lambda_{\min}(C_t)}$$

Proof of linear convergence

$$\lim_{T\to\infty} \frac{1}{T} \log \frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \log \frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|}$$

$$= \lim_{T\to\infty} \frac{1}{T} \log \frac{\sigma_T}{\sigma_0} + \frac{1}{2T} \log \frac{\lambda_{\min}(C_T)}{\lambda_{\min}(C_0)}$$

Proof of linear convergence

$$\lim_{T\to\infty}\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \lim_{T\to\infty}\frac{1}{T}\log\frac{\sigma_T}{\sigma_0} + \frac{1}{2}\lim_{T\to\infty}\frac{1}{T}\log\frac{\lambda_{\min}(C_T)}{\lambda_{\min}(C_0)}$$

In progress

- **Proof of irreducibility of a normalized chain**

In progress

- Proof of irreducibility of a normalized chain
- **Found a potential function on which we have a drift condition for ergodicity**

In progress

- Proof of irreducibility of a normalized chain
- Found a potential function on which we have a drift condition for ergodicity[3]

[3] currently when the objective function is an increasing transformation of a convex-quadratic function

In progress

- Proof of irreducibility of a normalized chain
- Found a potential function on which we have a drift condition for ergodicity[3]
- **Proof of linear convergence**

[3]currently when the objective function is an increasing transformation of a convex-quadratic function

In progress

- Proof of irreducibility of a normalized chain
- Found a potential function on which we have a drift condition for ergodicity[3]
- Proof of linear convergence
- **When minimizing a convex-quadratic function**

$$\mathbb{E}\left[\frac{C_t}{\text{normalization}}\right] \underset{t\to\infty}{\longrightarrow} \text{constant} \times H^{-1}$$

[3] currently when the objective function is an increasing transformation of a convex-quadratic function

Thank you!