Convergence analysis of evolution strategies with covariance matrix adaptation

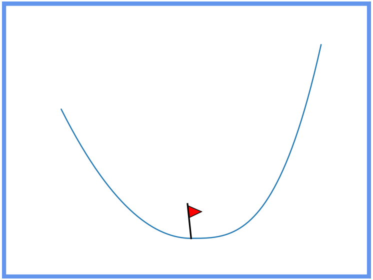
PhD students seminar

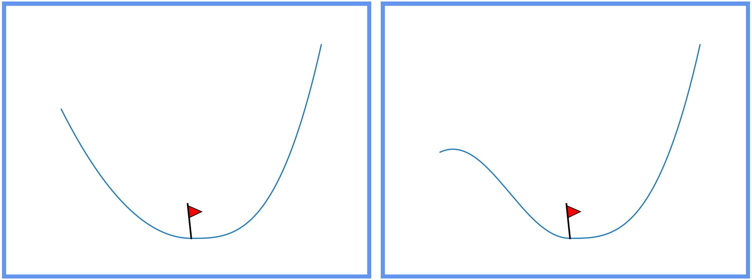
Armand Gissler

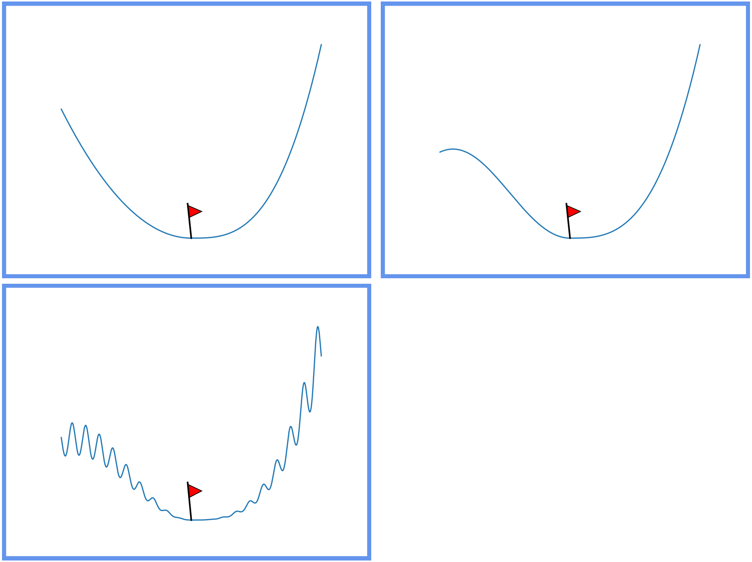
RandOpt team, Inria & École polytechnique

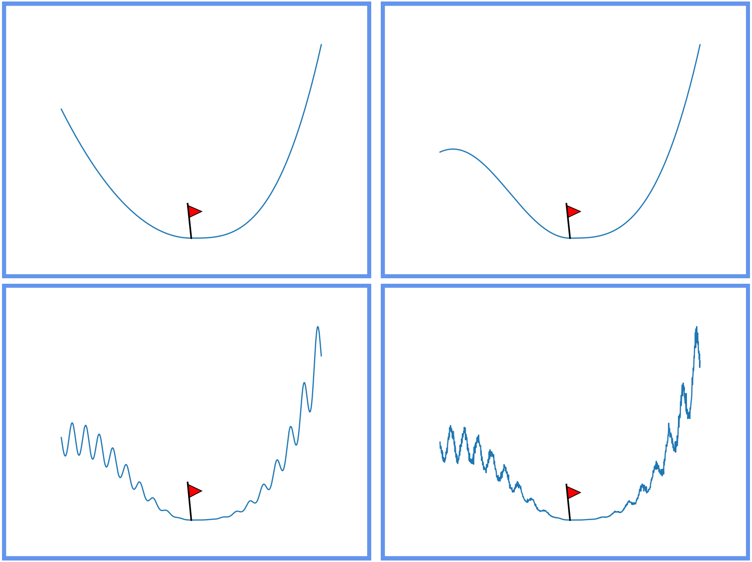
Advisors: Anne Auger & Nikolaus Hansen

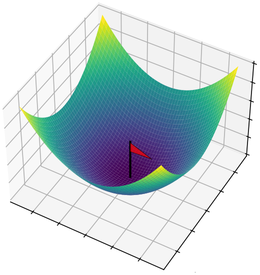
Wednesday 10th April, 2024

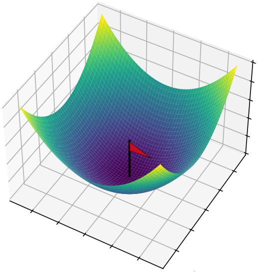


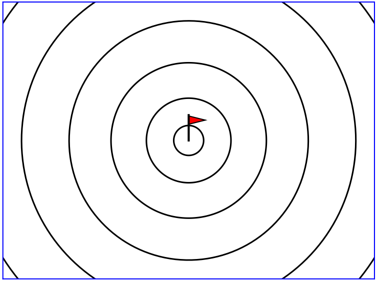


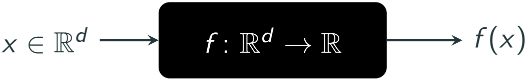


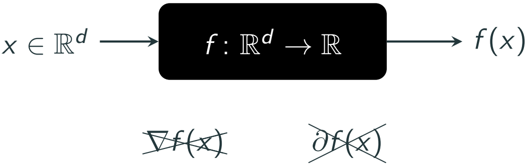


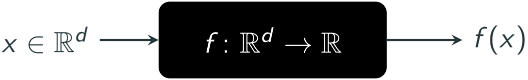






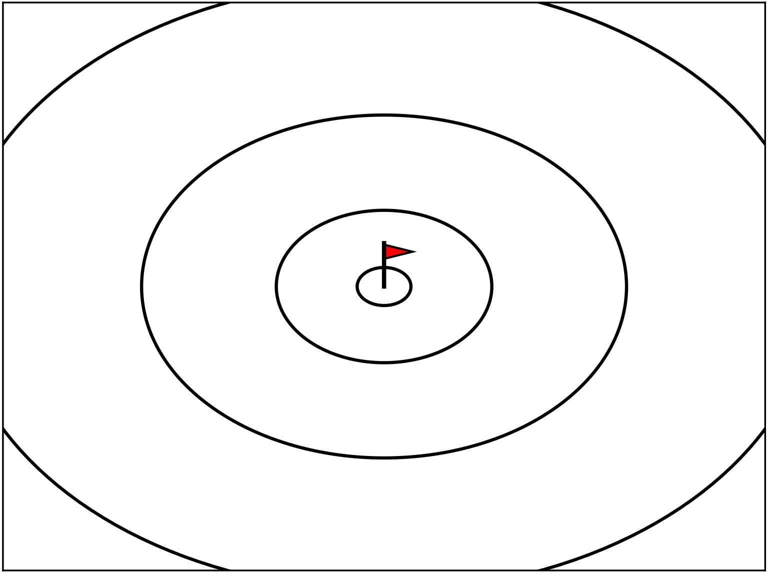

$x \in \mathbb{R}^d$
$f : \mathbb{R}^d \to \mathbb{R}$
$f(x)$

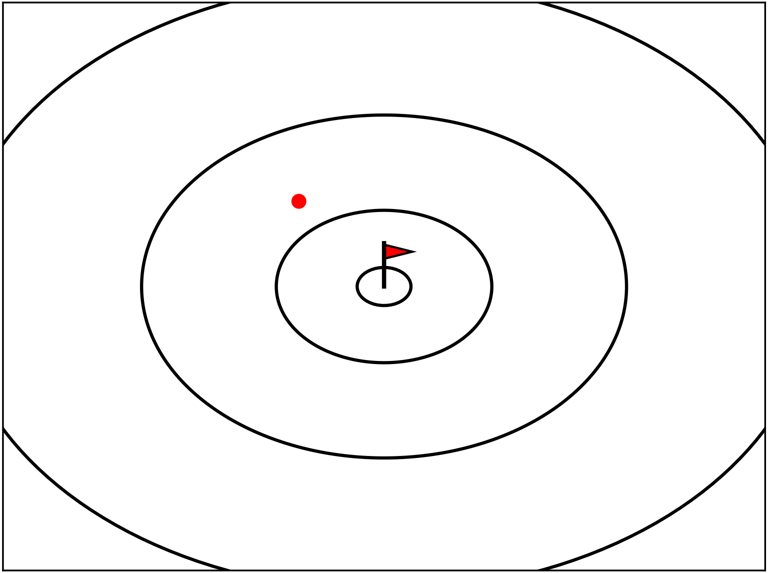

x ∈ ℝ^d
f : ℝ^d → ℝ
f(x)
∇f(x)
∂f(x)

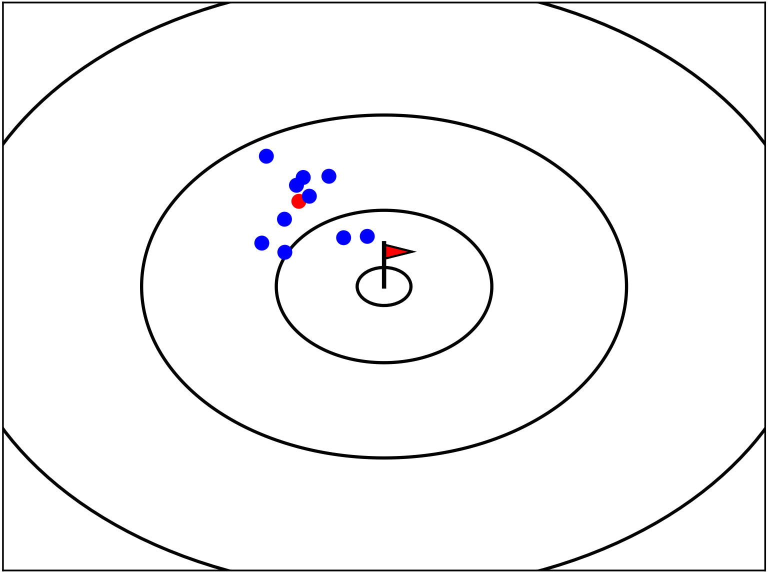


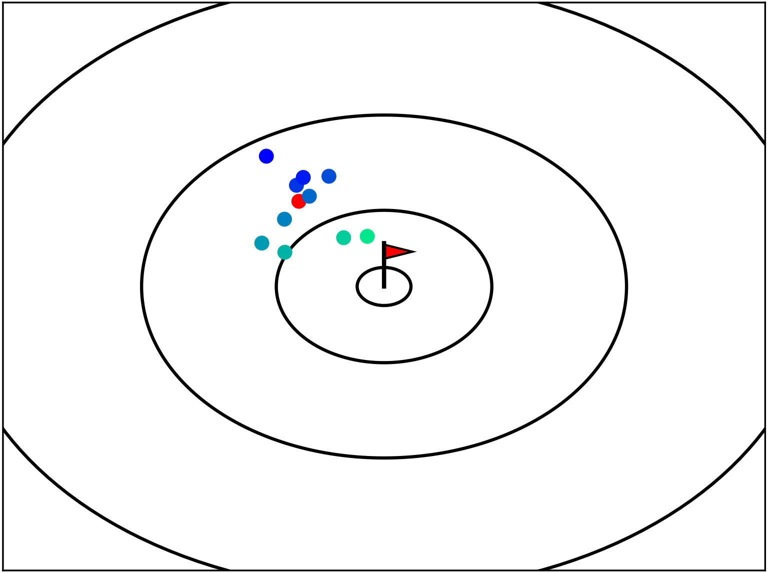
~~$\nabla f(x)$~~ ~~$\partial f(x)$~~

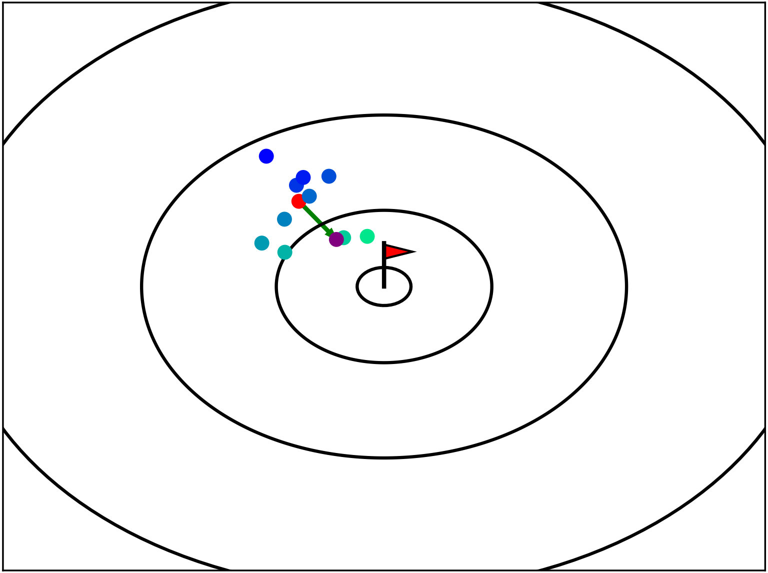
$$\text{Find } x^* \in \operatorname*{Arg\,min}_{x \in \mathbb{R}^d} f(x) \tag{P}$$





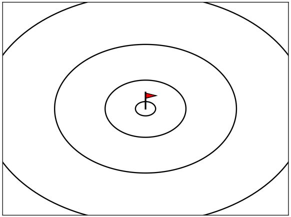






Algorithm 1 Our first ES

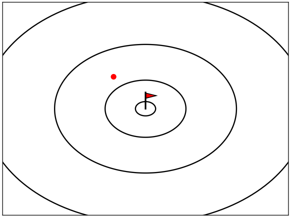
Goal: $\min_{x \in \mathbb{R}^d} f(x)$



Algorithm 1 Our first ES

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

Repeat: (Given $m_t \in \mathbb{R}^d$)

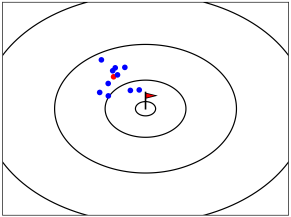


Algorithm 1 Our first ES

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

Repeat: (Given $m_t \in \mathbb{R}^d$)

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



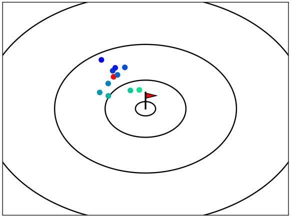
$\lambda =$ population size

Algorithm 1 Our first ES

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

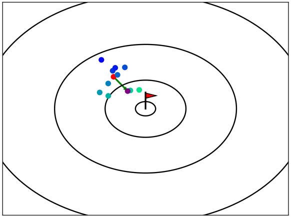
Repeat: (Given $m_t \in \mathbb{R}^d$)

1. $x_{t+1}^1, \ldots, x_{t+1}^\lambda \sim \mathcal{N}(m_t, I_d)$
2. Rank population:
   $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$



$\lambda =$ population size

Algorithm 1 Our first ES



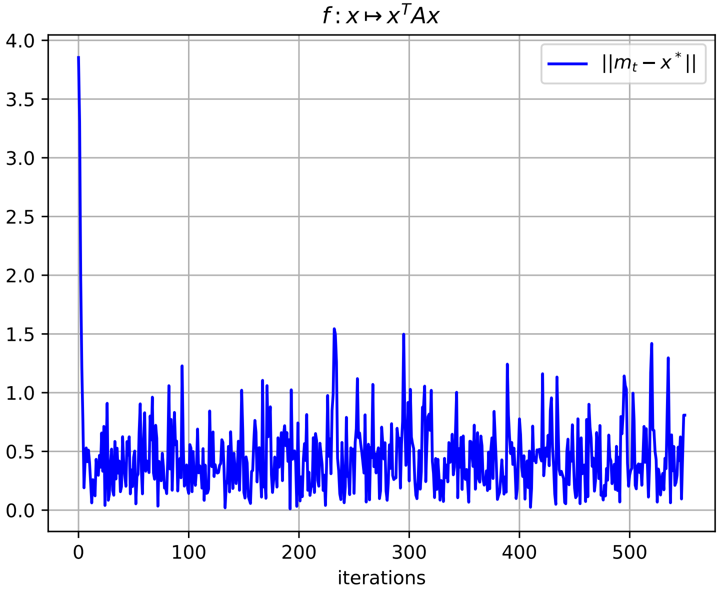
Goal: $\min_{x \in \mathbb{R}^d} f(x)$

Repeat: (Given $m_t \in \mathbb{R}^d$)

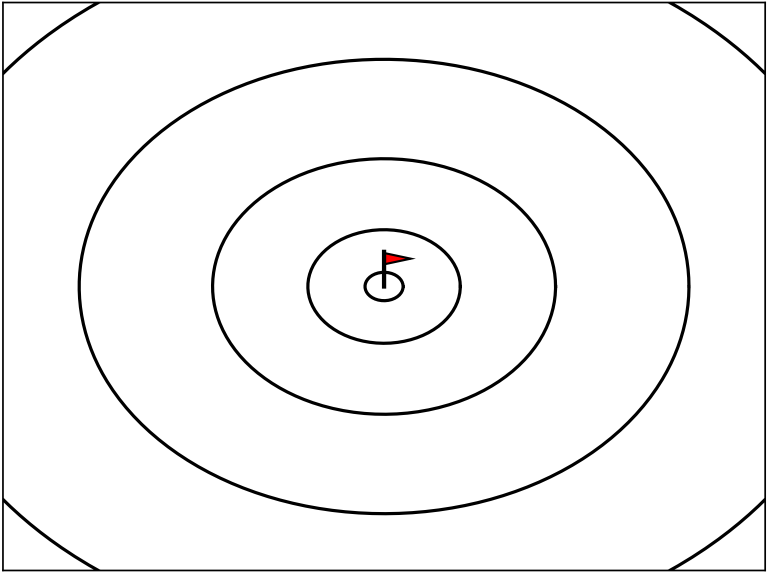
1. $x_{t+1}^1, \ldots, x_{t+1}^\lambda \sim \mathcal{N}(m_t, I_d)$
2. Rank population:
   $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$
3. Update mean: $m_{t+1} = \text{Average}(x_{t+1}^{1:\lambda}, \ldots, x_{t+1}^{\mu:\lambda})$

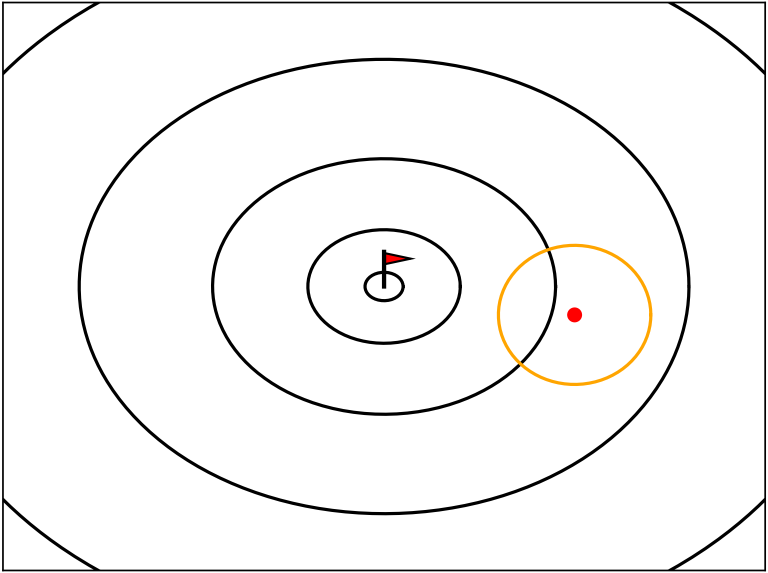
$\lambda =$ population size

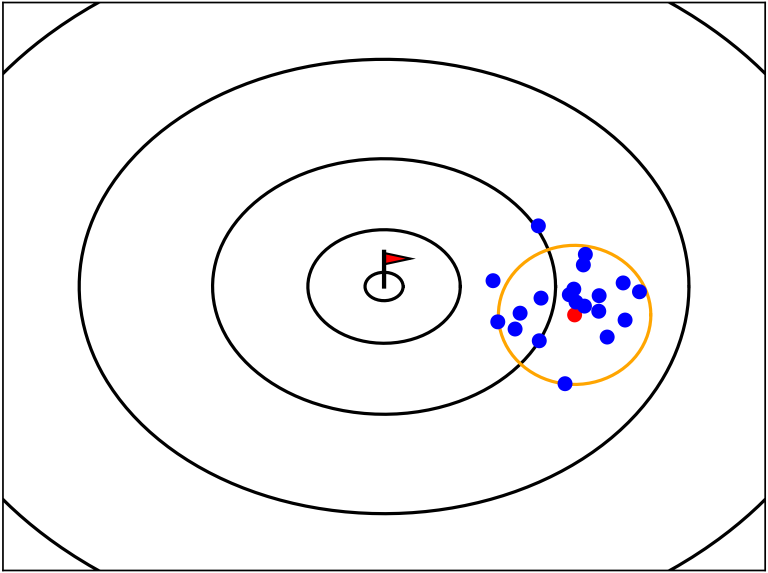
$\mu =$ parent number

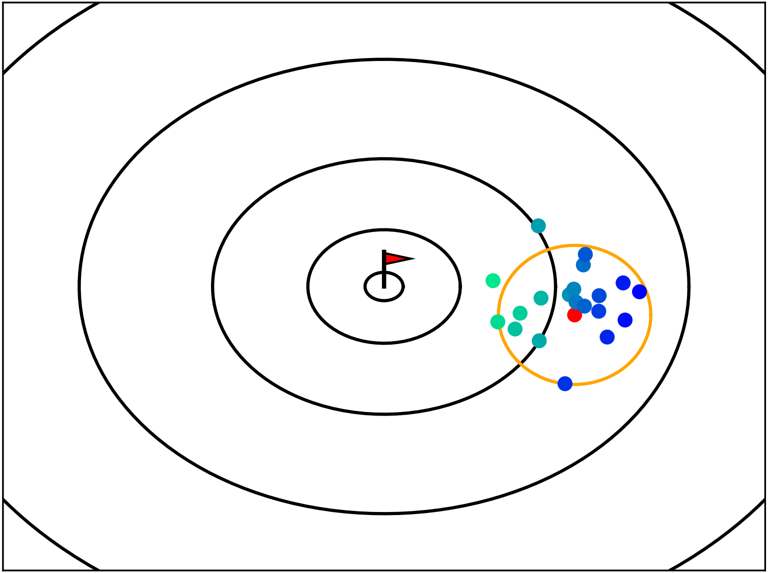


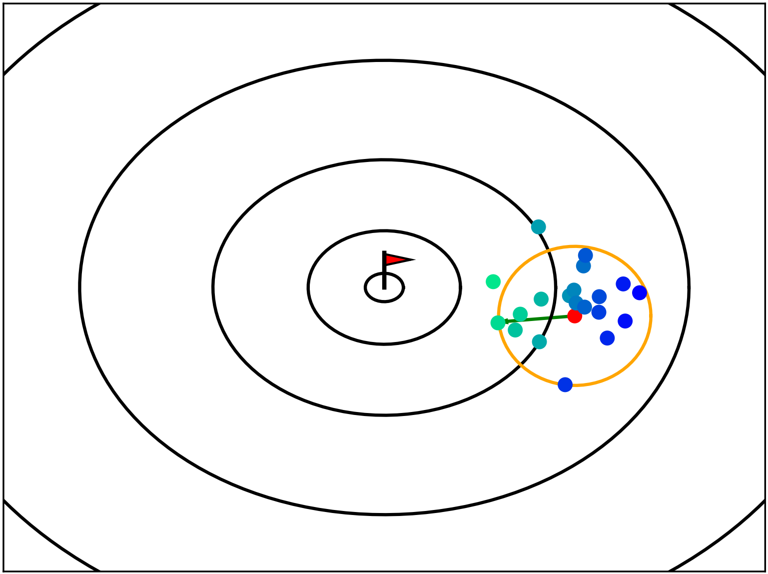
$f: x \mapsto x^T A x$
$||m_t - x^*||$
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
100
200
300
400
500
iterations

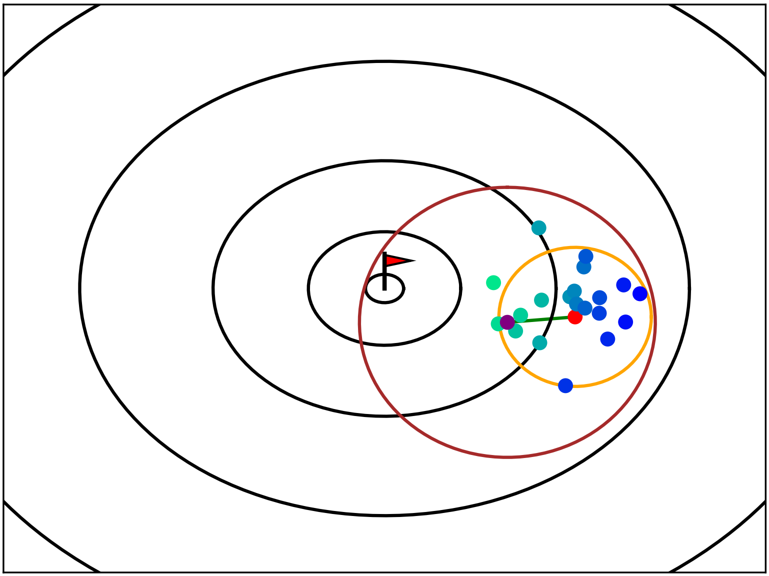


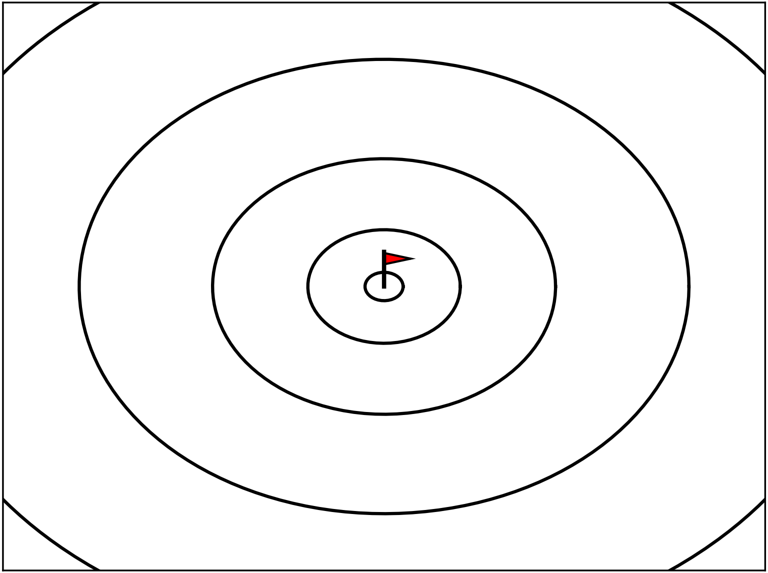


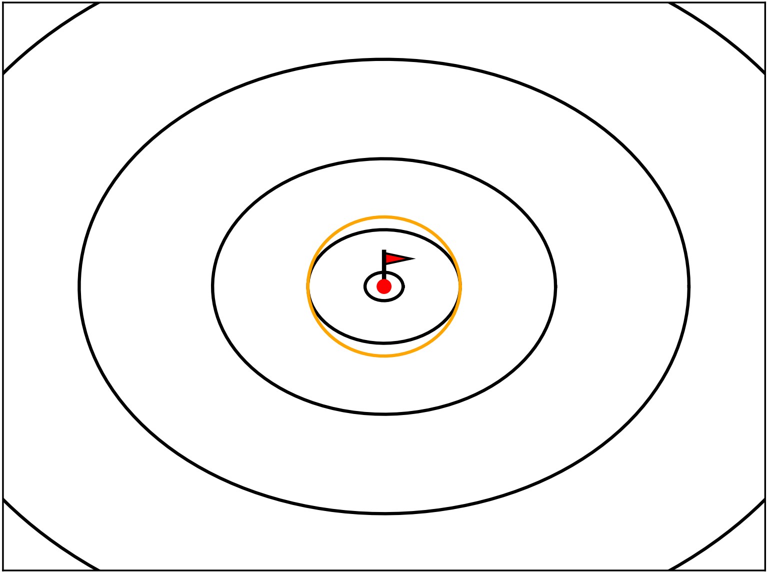


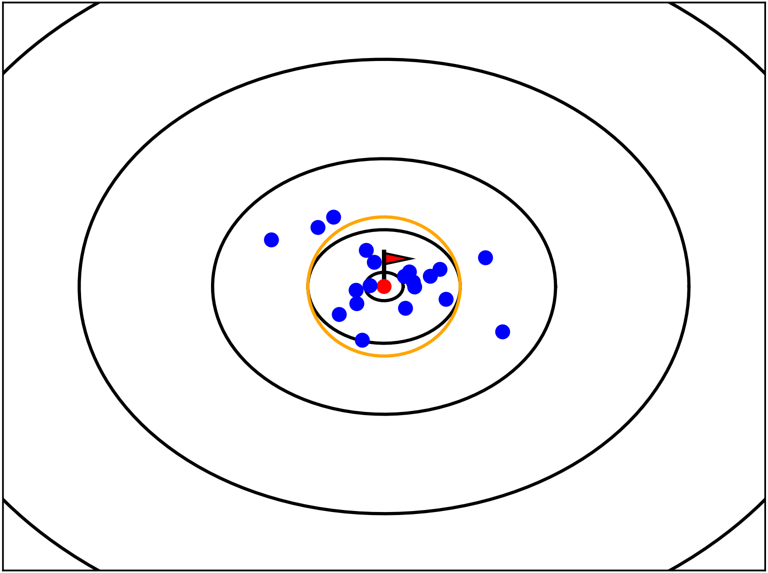


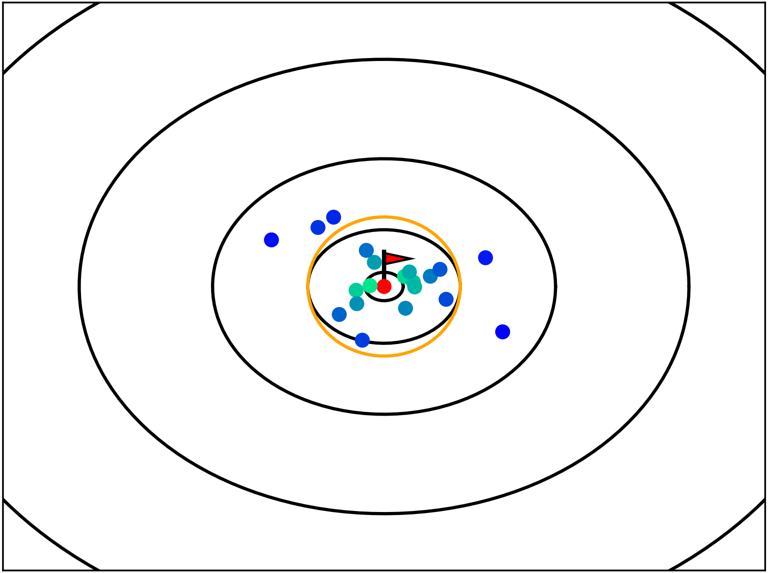


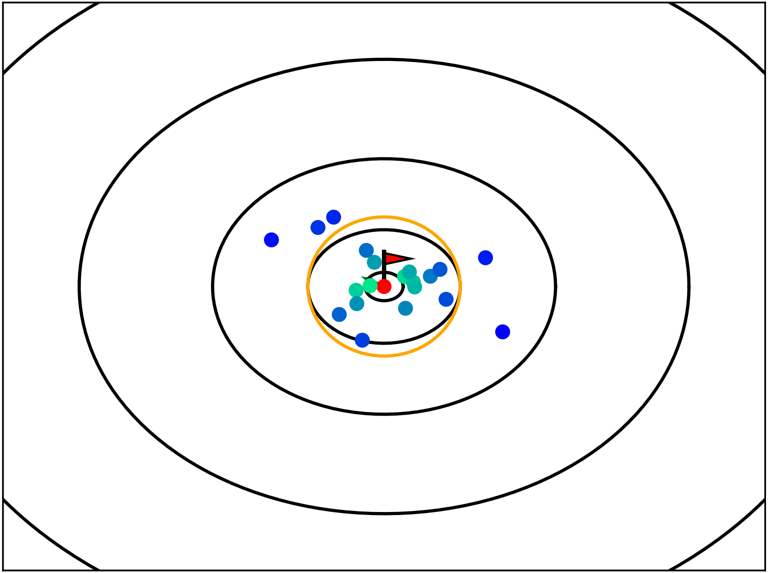


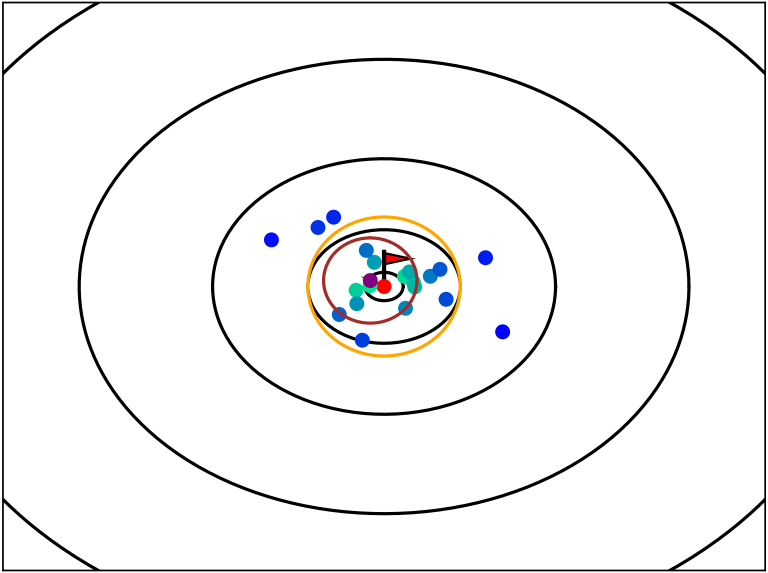






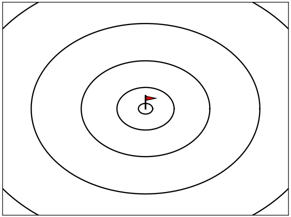






Algorithm 2 ES with step-size adaptation

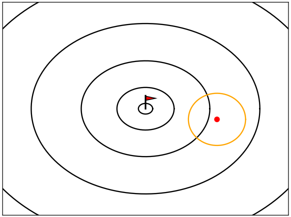
Goal: $\min_{x \in \mathbb{R}^d} f(x)$



Algorithm 2 ES with step-size adaptation

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

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

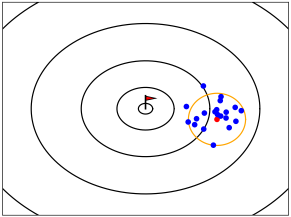


Algorithm 2 ES with step-size adaptation

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

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

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



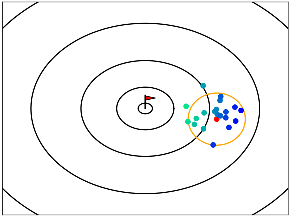
$\lambda =$ population size

Algorithm 2 ES with step-size adaptation

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

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

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



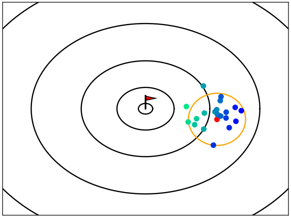
$\lambda =$ population size

Algorithm 2 ES with step-size adaptation

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

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

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



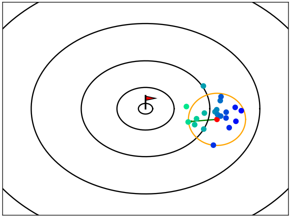
$\lambda =$ population size

Algorithm 2 ES with step-size adaptation

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

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

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



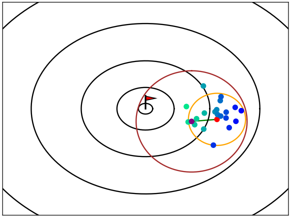
λ = population size

μ = parent number

Algorithm 2 ES with step-size adaptation

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

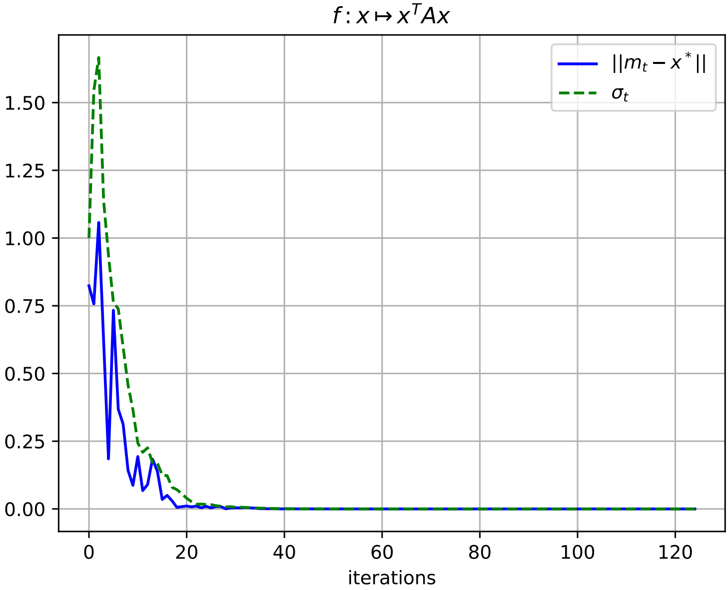
Repeat (Given $m_t \in \mathbb{R}^d$ and $\sigma_t > 0$)

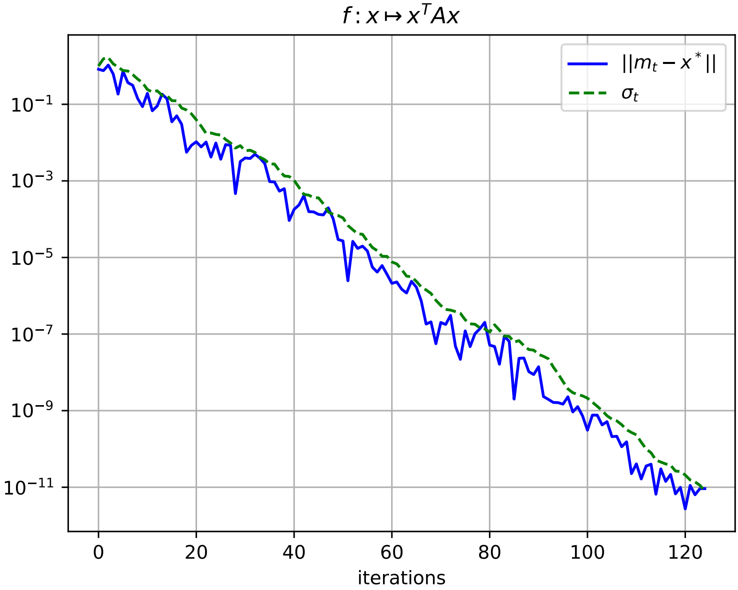


1. $x_{t+1}^1, \ldots, x_{t+1}^\lambda \sim \mathcal{N}(m_t, \sigma_t^2 I_d)$
2. sort $f(x_{t+1}^i)$:
   $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$
3. $m_{t+1} = \text{Average}(x_{t+1}^{1:\lambda}, \ldots, x_{t+1}^{\mu:\lambda})$
4. $\sigma_{t+1}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

$\lambda =$ population size

$\mu =$ parent number


$f: x \mapsto x^T Ax$
$||m_t - x^*||$
σ_t
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
20
40
60
80
100
120
iterations


$f : x \mapsto x^T A x$
$||m_t - x^*||$
σ_t
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-9}
10^{-11}
0
20
40
60
80
100
120
iterations

Prove:

$$\frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|} \approx \frac{\sigma_{t+1}}{\sigma_t} \approx \rho \in (0, 1).$$

Prove:

$$\log \frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|} \approx \log \frac{\sigma_{t+1}}{\sigma_t} \approx -\mathrm{CR}.$$

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

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

$P \colon \mathsf{X} \times \mathcal{B}(\mathsf{X}) \to \mathbb{R}$ 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:

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

$P \colon \mathsf{X} \times \mathcal{B}(\mathsf{X}) \to \mathbb{R}$ 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:

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

- When $\mathsf{X} = \{1, \ldots, n\}$ is finite, P can be represented as a $n \times n$ matrix:

$$\mathbb{P}[\theta_{t+1} = j \mid \theta_t = i] = P_{ij}$$

$P\colon \mathsf{X} \times \mathcal{B}(\mathsf{X}) \to \mathbb{R}$ 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:

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

- When $\mathsf{X} = \{1, \dots, n\}$ is finite, P can be represented as a $n \times n$ matrix:

$$\mathbb{P}[\theta_{t+1} = j \mid \theta_t = i] = P_{ij}$$

- We can define a k**-steps transition kernel** P^k which satisfies

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

If X is finite:

If $\mathsf{X} = \{1, \dots, n\}$:

$$\nu_0 = (p_1, \dots, p_n) \qquad \text{with } \sum_k p_k = 1$$

represents an initial state of the Markov chain $\{\theta_k\}_{k \in \mathbb{N}}$

If $\mathsf{X} = \{1, \ldots, n\}$:

$$\nu_0 = (p_1, \ldots, p_n) \qquad \text{with } \sum_k p_k = 1$$

represents an initial state of the Markov chain $\{\theta_k\}_{k \in \mathbb{N}}$

After k steps:

$$\nu_k = \nu_0 \times P^k$$

If $X = \{1, \ldots, n\}$:

$$\nu_0 = (p_1, \ldots, p_n) \qquad \text{with } \sum_k p_k = 1$$

represents an initial state of the Markov chain $\{\theta_k\}_{k \in \mathbb{N}}$

After k steps:

$$\nu_k = \nu_0 \times P^k$$

If

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

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

If X is **infinite**:

$$\nu_0 \textbf{ probability measure on } \mathsf{X}$$

represents an initial state of the Markov chain $\{\theta_k\}_{k\in\mathbb{N}}$

After k steps:

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

If

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

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

For an ergodic Markov chain $\Theta = \{\theta_k\}_{k \in \mathbb{N}}$:

$$\theta_k \underset{k \to \infty}{\longrightarrow} \pi$$

For an ergodic Markov chain $\Theta = \{\theta_k\}_{k \in \mathbb{N}}$:

$$\theta_k \underset{k \to \infty}{\longrightarrow} \pi$$

where π is the invariant probability measure of Θ:

$$\theta_k \sim \pi \Rightarrow \theta_{k+1} \sim \pi$$

When $\Theta = \{\theta_k\}_{k \in \mathbb{N}}$ is ergodic:

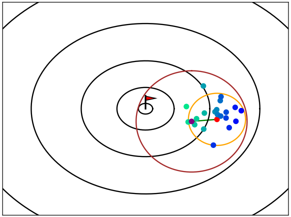
$$\lim_{T \to +\infty} \frac{1}{T} \sum_{k=0}^{T-1} g(\theta_k) = \int g(x) \mathrm{d}\pi(x)$$

Algorithm 2 ES with step-size adaptation

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

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

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



$\lambda =$ population size

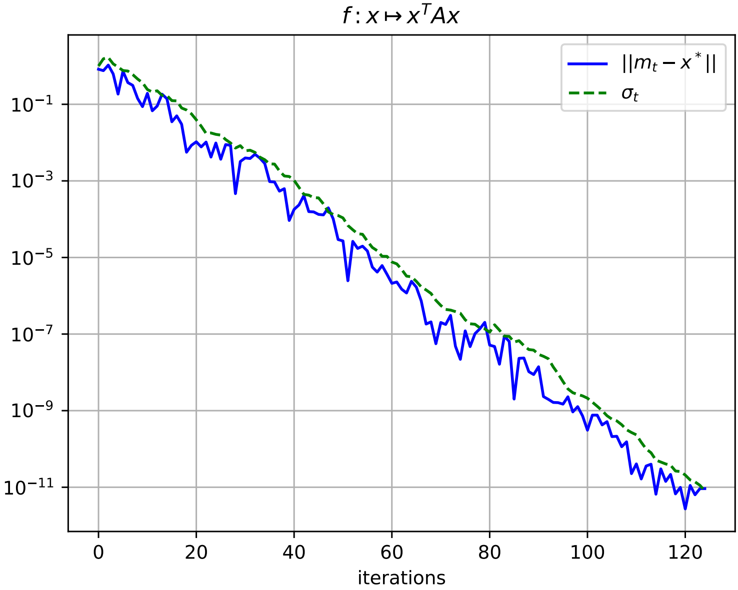
$\mu =$ parent number

$\{(m_k, \sigma_k)\}_{k \in \mathbb{N}}$ is a Markov chain valued in $X = \mathbb{R}^d \times (0, +\infty)$

$$\lim_{k\to\infty} m_k = x^* \quad \text{and} \quad \lim_{k\to\infty} \sigma_k = 0$$

$$\lim_{k\to\infty} m_k = x^* \quad \text{and} \quad \lim_{k\to\infty} \sigma_k = 0$$

$\delta_{(x^*,0)}$ is **not** a probability distribution on $\mathsf{X} = \mathbb{R}^d \times (0, +\infty)$!


$f: x \mapsto x^T A x$
$||m_t - x^*||$
σ_t
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-9}
10^{-11}
0
20
40
60
80
100
120
iterations

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

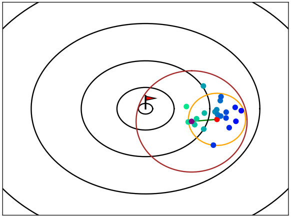
$$z_t = \frac{m_t - x^*}{\sigma_t}$$

Question: $\{z_t\}_{t\in\mathbb{N}}$ is an ergodic Markov chain?

Algorithm 3 ES with step-size adaptation

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

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



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

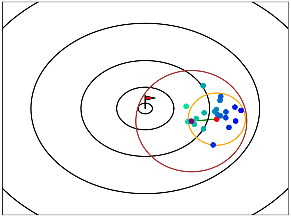
$\lambda =$ population size

$\mu =$ parent number

Algorithm 3 ES with step-size adaptation

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

Repeat (Given $z_t \in \mathbb{R}^d$) :



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

$\lambda =$ population size

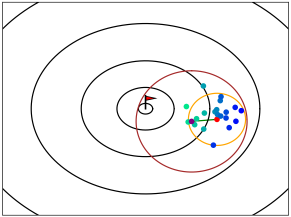
$\mu =$ parent number

Algorithm 3 ES with step-size adaptation

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

Repeat (Given $z_t \in \mathbb{R}^d$) :

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



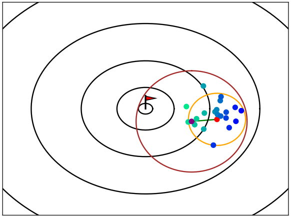
λ = population size

μ = parent number

Algorithm 3 ES with step-size adaptation

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

Repeat (Given $z_t \in \mathbb{R}^d$) :



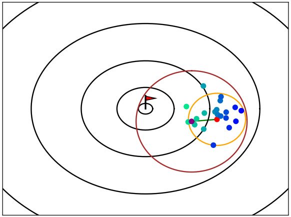
1. $z_{t+1}^1, \ldots, z_{t+1}^\lambda \sim \mathcal{N}(z_t, I_d)$
2. sort $f(x_{t+1}^i)$:
   $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$
3. $z_{t+1} = \text{Average}(z_{t+1}^{1:\lambda}, \ldots, z_{t+1}^{\mu:\lambda})$
4. $\sigma_{t+1} = \sigma_t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

$\lambda =$ population size

$\mu =$ parent number

Algorithm 3 ES with step-size adaptation

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



Repeat (Given $z_t \in \mathbb{R}^d$) :

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

$\lambda =$ population size

$\mu =$ parent number

Algorithm 3 ES with step-size adaptation

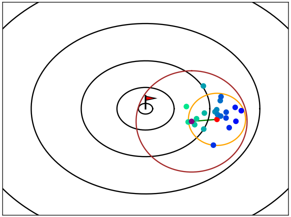
Goal: $\min_{x \in \mathbb{R}^d} f(x)$

Repeat (Given $z_t \in \mathbb{R}^d$) :

1. $z_{t+1}^1, \ldots, z_{t+1}^\lambda \sim \mathcal{N}(z_t, I_d)$
2. sort $f(x_{t+1}^i)$:

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

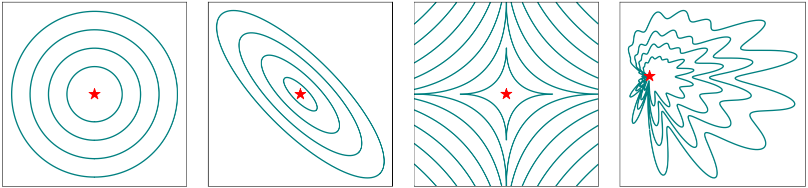
3. $z_{t+1} = \dfrac{\text{Average}(z_{t+1}^{1:\lambda}, \ldots, z_{t+1}^{\mu:\lambda})}{\text{increasing function}(\|z_{t+1} - z_t\|)}$

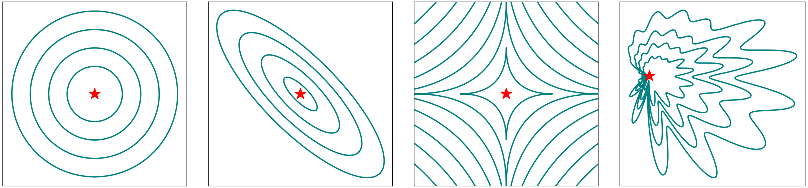


$\lambda =$ population size

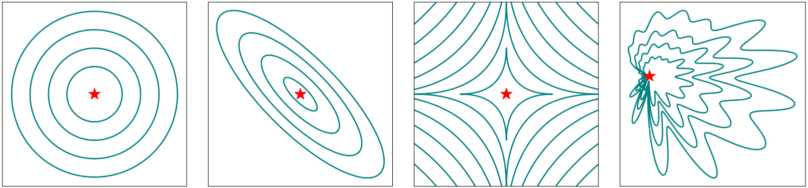
$\mu =$ parent number

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





$$f\left(m_t\right) \leqslant f\left(x_{t+1}\right) \Leftrightarrow f\left(\star + \frac{m_t - \star}{\sigma_t}\right) \leqslant f\left(\star + \frac{x_{t+1} - \star}{\sigma_t}\right)$$



$$f\left(m_t\right) \leqslant f\left(x_{t+1}\right) \Leftrightarrow f\left(\star + \frac{m_t - \star}{\sigma_t}\right) \leqslant f\left(\star + \frac{x_{t+1} - \star}{\sigma_t}\right)$$

Proposition

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

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

Question: $\{z_t\}_{t \in \mathbb{N}}$ is an ergodic Markov chain?

When X is finite:

Theorem
If $\{\theta_t\}_{t\in\mathbb{N}}$ *is an irreducible, aperiodic Markov chain, then it is ergodic.*

When X is finite:

Theorem
If $\{\theta_t\}_{t\in\mathbb{N}}$ is an $\boxed{\text{irreducible}}$, aperiodic Markov chain, then it is ergodic.

If $X = \{1, \ldots, n\}$,

If $\mathsf{X} = \{1, \dots, n\}$,

$\{\theta_t\}_{t\in\mathbb{N}}$ is irreducible if

$$\forall x, y \in \mathsf{X},\ \exists k > 0, \quad P^k_{x,y} > 0.$$

If $\mathsf{X} = \{1, \ldots, n\}$,

$\{\theta_t\}_{t \in \mathbb{N}}$ is irreducible if

$$\forall x, y \in \mathsf{X},\ \underbrace{\exists k > 0, \quad P^k_{x,y} > 0}_{x \rightsquigarrow y}.$$

If X infinite,

$\{\theta_t\}_{t\in\mathbb{N}}$ is irreducible if

$$\forall x \in \mathsf{X}, \mathsf{A} \subset \mathsf{X}, \quad \mathrm{volume}(\mathsf{A}) > 0 \Rightarrow x \rightsquigarrow \mathsf{A}$$

If X infinite,

$\{\theta_t\}_{t\in\mathbb{N}}$ is irreducible if

$$\forall x \in X, A \subset X, \quad \mathrm{volume}(A) > 0 \Rightarrow x \rightsquigarrow A$$

for some volume on X.

When X is finite:

Theorem
If $\{\theta_t\}_{t \in \mathbb{N}}$ is an irreducible, $\boxed{\text{aperiodic}}$ Markov chain, then it is ergodic.

If

$$\begin{cases} \theta_1 \in A_1 \Rightarrow \mathbb{P}[\theta_2 \in A_2] = 1 \\ \theta_2 \in A_2 \Rightarrow \mathbb{P}[\theta_3 \in A_3] = 1 \\ \vdots \\ \theta_T \in A_T \Rightarrow \mathbb{P}[\theta_{T+1} \in A_1] = 1 \end{cases}$$

Then $\text{period} = T$.

When $\text{period} = 1$,
$\{\theta_t\}$ is **aperiodic**.

When X is finite:

Theorem
If $\{\theta_t\}_{t\in\mathbb{N}}$ *is an irreducible, aperiodic Markov chain, then it is ergodic.*

When X is infinite:

Theorem
If $\{\theta_t\}_{t\in\mathbb{N}}$ *is an irreducible, aperiodic Markov chain, then it is ergodic if*

$$\mathbb{E}[V(\theta_{t+1}) \mid \theta_t] \leqslant (1-\varepsilon)V(\theta_t) \qquad \textit{if } \theta_t \notin \textit{small set}$$

for some $V \colon \mathsf{X} \to [1, +\infty]$.

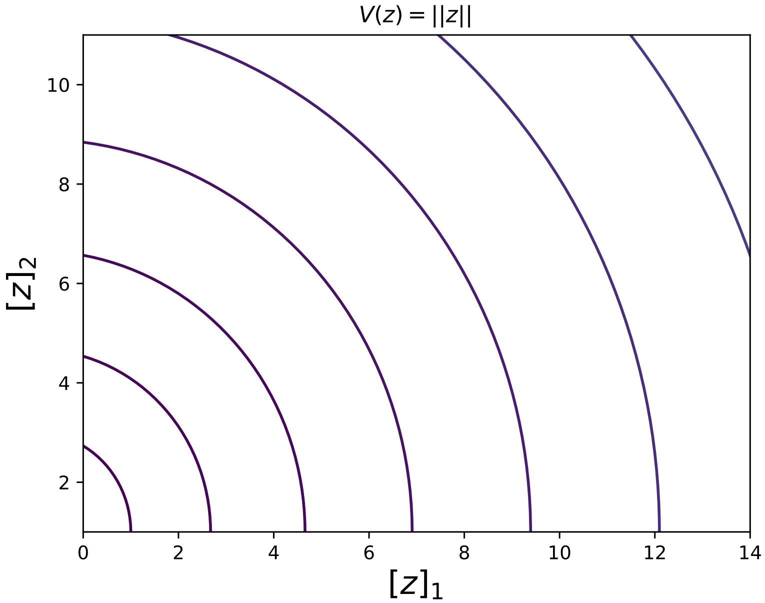
Proposition: for $\{z_t\}_{t \in \mathbb{N}}$, compact sets are small

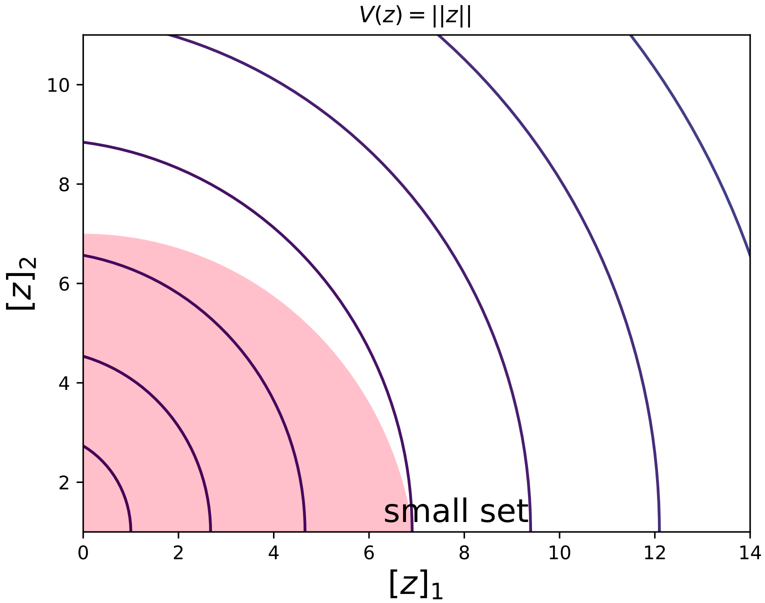
When X is infinite:

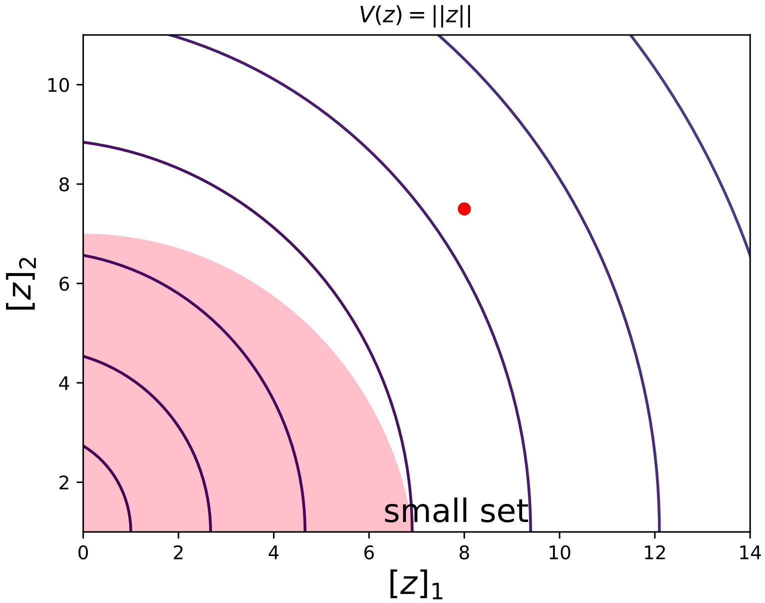
Theorem
If $\{\theta_t\}_{t\in\mathbb{N}}$ is an irreducible, aperiodic Markov chain, then it is ergodic if

$$\mathbb{E}[V(\theta_{t+1}) \mid \theta_t] \leqslant (1-\varepsilon)V(\theta_t) \quad \text{if } \theta_t \notin \text{compact set}$$

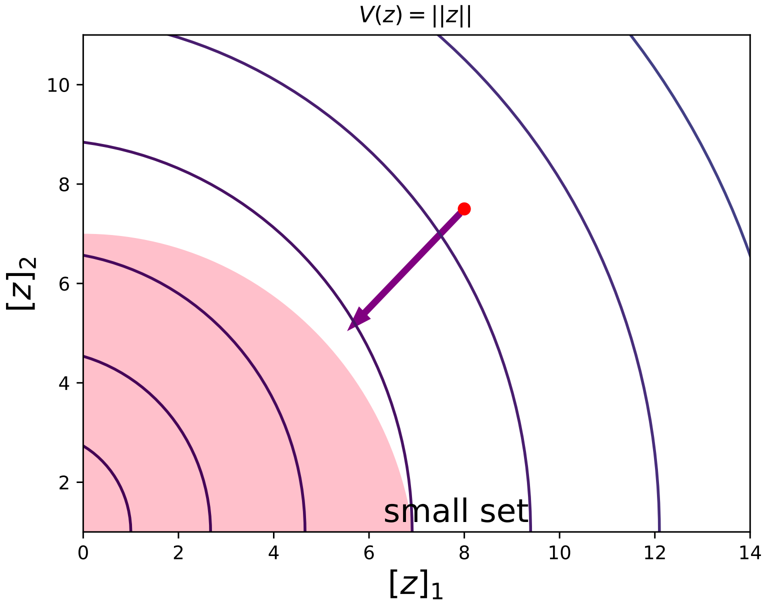
for some $V\colon \mathsf{X} \to [1,+\infty]$.


$V(z) = ||z||$
10
8
6
4
2
0
2
4
6
8
10
12
14
$[z]_2$
$[z]_1$


$V(z) = ||z||$
$[z]_2$
$[z]_1$
small set
0
2
4
6
8
10
12
14
2
4
6
8
10



$V(z) = ||z||$
small set
$[z]_1$
$[z]_2$
0
2
4
6
8
10
12
14
2
4
6
8
10



$V(z) = ||z||$
10
8
6
4
2
$[z]_2$
0
2
4
6
8
10
12
14
$[z]_1$
small set

When X is infinite:

Theorem
If $\{\theta_t\}_{t\in\mathbb{N}}$ *is an irreducible, aperiodic Markov chain, then it is ergodic if*

$$\mathbb{E}[V(\theta_{t+1}) \mid \theta_t] \leqslant (1-\varepsilon)V(\theta_t) \qquad \text{if } \theta_t \notin \text{compact set}$$

for some $V\colon \mathsf{X} \to [1,+\infty]$.

Scheme of proof:

Scheme of proof:

1. irreducibility and aperiodicity of $\{z_t\}_{t\in\mathbb{N}}$

Scheme of proof:

1. irreducibility and aperiodicity of $\{z_t\}_{t \in \mathbb{N}}$
2. drift condition: $\exists \mathsf{K} \subset \mathbb{R}^d$ compact and $V : \mathbb{R}^d \to [1, +\infty]$

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

Scheme of proof:

1. irreducibility and aperiodicity of $\{z_t\}_{t\in\mathbb{N}}$
2. drift condition: $\exists \mathsf{K} \subset \mathbb{R}^d$ compact and $V : \mathbb{R}^d \to [1, +\infty]$

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

3. deduce convergence from the ergodicity

Scheme of proof:

1. **irreducibility and aperiodicity of** $\{z_t\}_{t\in\mathbb{N}}$
2. drift condition: $\exists \mathsf{K} \subset \mathbb{R}^d$ compact and $V\colon \mathbb{R}^d \to [1, +\infty]$

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

3. deduce convergence from the ergodicity

$$\theta_{k+1} = F(\theta_k, x_{k+1}) \qquad (\mathbf{CM}(\mathbf{F}))$$

where $x_{k+1} \sim p_{\theta_k}$

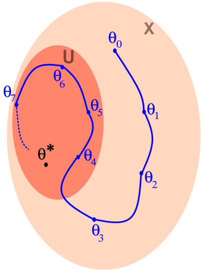
$$\begin{aligned}\theta_{k+1} &= F(\theta_k, x_{k+1}) \qquad (\mathbf{CM}(\mathbf{F}))\\ &= F_{k+1}(\theta_0, x_1, \ldots, x_{k+1})\end{aligned}$$

where $x_{k+1} \sim p_{\theta_k}$

θ^* is **attracting** when

θ^* is **attracting** when

$$\exists x_1, x_2, \ldots, \ \lim_{k\to\infty} F_k(\theta_0, x_{1..k}) = \theta^*$$



θ^* is **attracting** when

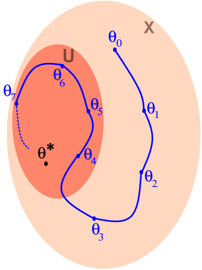
$$\exists x_1, x_2, \ldots, \lim_{k \to \infty} F_k(\theta_0, x_{1..k}) = \theta^*$$

Theorem
If

- $\exists \theta^*$ *attracting*
- $\exists x_1^*, \ldots, x_k^*$

such that $F_k(\theta^*, \cdot)$ *is a* ***submersion*** *at* $x_{1..k}^*$,

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

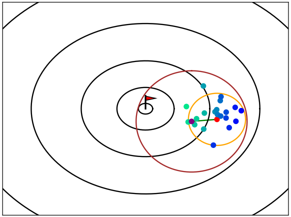


Algorithm 3 ES with step-size adaptation

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

Repeat (Given $z_t \in \mathbb{R}^d$) :

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



$\lambda =$ population size

$\mu =$ parent number

$$z_{k+1} = F(z_k, z_{k+1}^{1:\lambda}, \ldots, z_{k+1}^{\lambda:\lambda})$$

$$z_{k+1} = F(z_k, z_{k+1}^{1:\lambda}, \ldots, z_{k+1}^{\lambda:\lambda})$$

Proposition
0 *is attracting*

Proof.
Take $z_k^{i:\lambda} = 0$. Then

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

□

$$z_{k+1} = F(z_k, z_{k+1}^{1:\lambda}, \ldots, z_{k+1}^{\lambda:\lambda})$$

Proposition
0 *is attracting, and* $F(0, \cdot)$ *is submersive at* 0.

Proof.

$$F(0, h^1, \ldots, h^\lambda) = 0 + \mathrm{Average}(h^1, \ldots, h^\lambda) + o(h^1, \ldots, h^\lambda)$$

□

Corollary

If $f \in \{$ , , ,  $\}$, $\{z_t\}_{t \in \mathbb{N}}$ *is irreducible and aperiodic.*

Scheme of proof:

1. irreducibility and aperiodicity of $\{z_t\}_{t\in\mathbb{N}}$
2. **drift condition:** $\exists \mathsf{K} \subset \mathbb{R}^d$ **compact and** $V : \mathbb{R}^d \to [1, +\infty]$

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

3. deduce convergence from the ergodicity

Proposition

If $f \in$ $\Big\{$ , ,  $\Big\}$*:*

$$\mathbb{E}[\|z_{t+1}\| \mid z_t] \leqslant (1-\varepsilon) \times \|z_t\| \qquad \textit{if } \|z_t\| \gg 1$$

Proposition

If $f \in \Big\{$ , ,  $\Big\}$:

$$\mathbb{E}[\|z_{t+1}\| \mid z_t] \leqslant (1-\varepsilon) \times \|z_t\| \qquad \textit{if } \|z_t\| \gg 1$$

Theorem

If $f \in \Big\{$ , ,  $\Big\}$: $\{z_t\}_{t \in \mathbb{N}}$ *is ergodic.*

Scheme of proof:

1. irreducibility and aperiodicity of $\{z_t\}_{t\in\mathbb{N}}$
2. drift condition: $\exists \mathsf{K} \subset \mathbb{R}^d$ compact and $V\colon \mathbb{R}^d \to [1, +\infty]$

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

3. **deduce convergence from the ergodicity**

Theorem

If $f \in \Big\{$ , ,  $\Big\}$, *ES converges linearly (or geometrically):*

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

Proof.

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

$\square$

Theorem

If $f \in \Big\{$ , ,  $\Big\}$, *ES converges linearly (or geometrically):*

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

Proof.

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

□

Theorem

If $f \in$ $\Big\{$ , ,  $\Big\}$, *ES converges linearly (or geometrically):*

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

Proof.

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

□

Theorem

If $f \in \Big\{$ , ,  $\Big\}$, *ES converges linearly (or geometrically):*

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

Proof.

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

□

Theorem

If $f \in \{$ , ,  $\}$, *ES converges linearly (or geometrically):*

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

Proof.

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

□

Theorem

If $f \in \{$  ,  ,  $\}$, *ES converges linearly (or geometrically):*

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

Proof.

$$\lim_{T\to\infty} \frac{1}{T} \log \frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \mathbb{E}_{z_t\sim\pi}\left[\log \|z_{t+1}\| - \log \|z_t\| + \uparrow (\|z_{t+1} - z_t\|)\right]$$

$\square$

Theorem

If $f \in \{$ , ,  $\}$, *ES converges linearly (or geometrically):*

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

Proof.

$$\lim_{T\to\infty} \frac{1}{T} \log \frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \mathbb{E}_{z_t\sim\pi}\left[\uparrow\left(\|z_{t+1} - z_t\|\right)\right]$$

□

Theorem

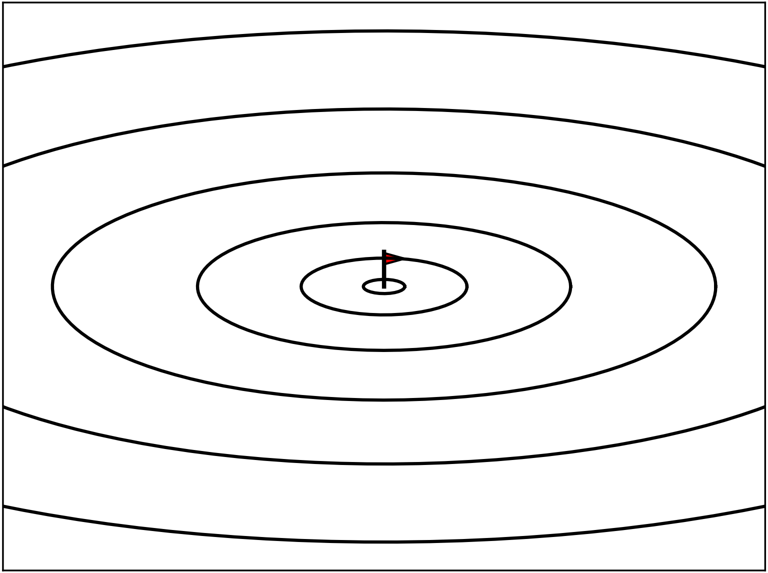
If $f \in \Big\{$ , ,  $\Big\}$, *ES converges linearly (or geometrically):*

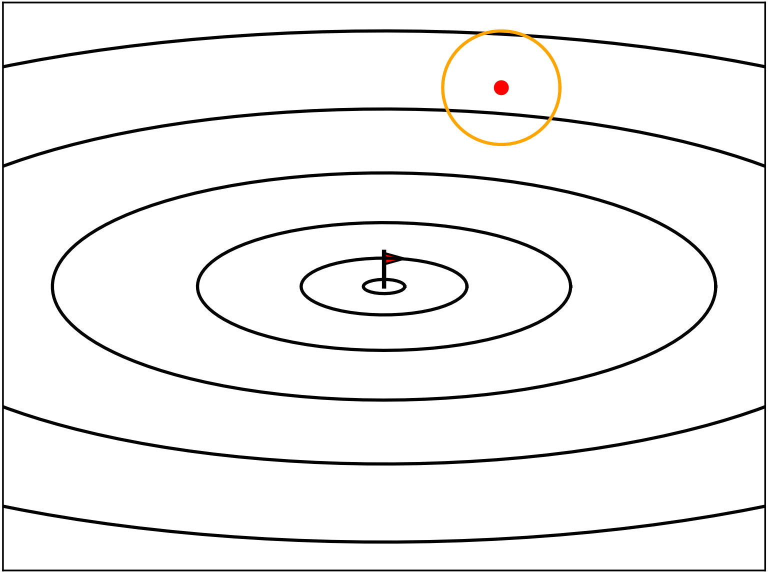
$$\lim_{t\to\infty} \frac{1}{t} \log \frac{\|m_t - x^*\|}{\|m_0 - x^*\|} = \lim_{t\to\infty} \mathbb{E}\left[\log \frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|}\right] = -\mathrm{CR}.$$

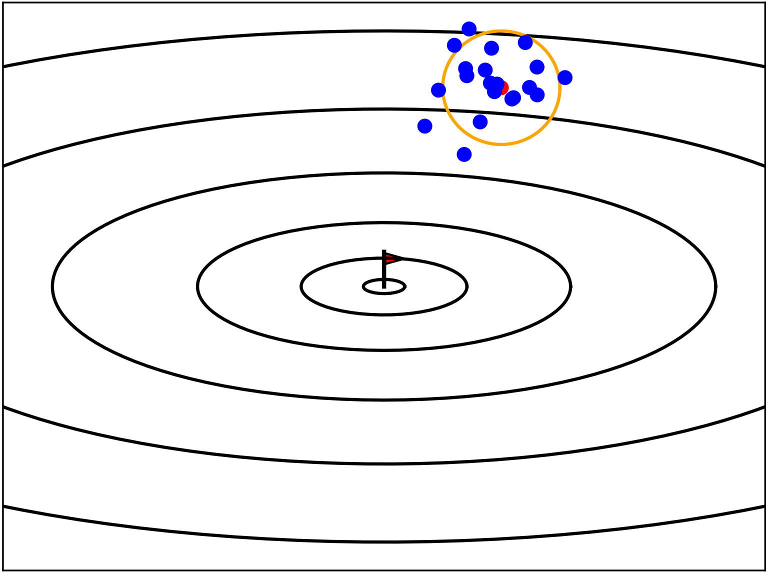
Proof.

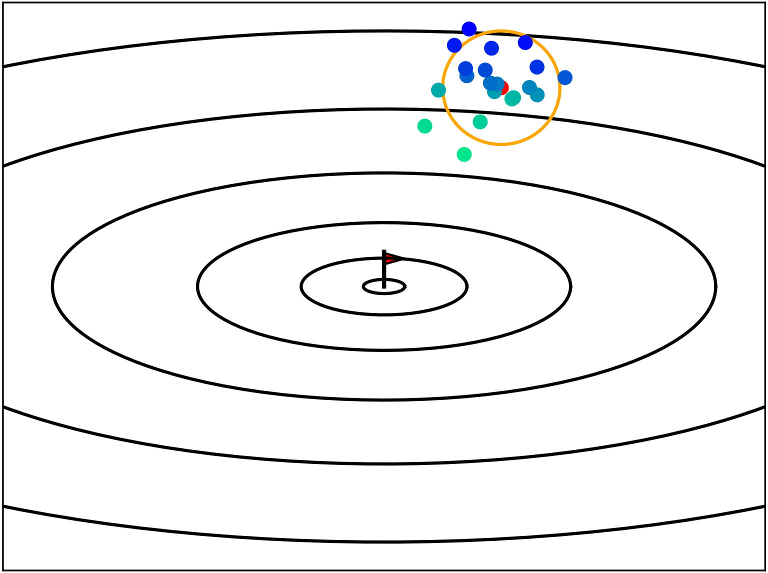
$$\lim_{T\to\infty} \frac{1}{T} \log \frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \mathbb{E}_{z_t\sim\pi}\left[\uparrow (\|z_{t+1} - z_t\|)\right]$$

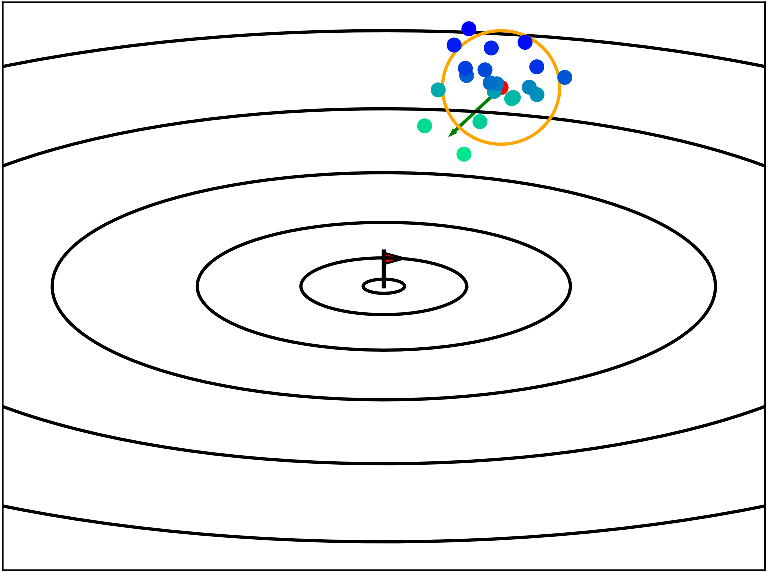
$$\mathrm{CR} = -\mathbb{E}_{z_t\sim\pi} \uparrow (\|z_{t+1} - z_t\|)$$ □

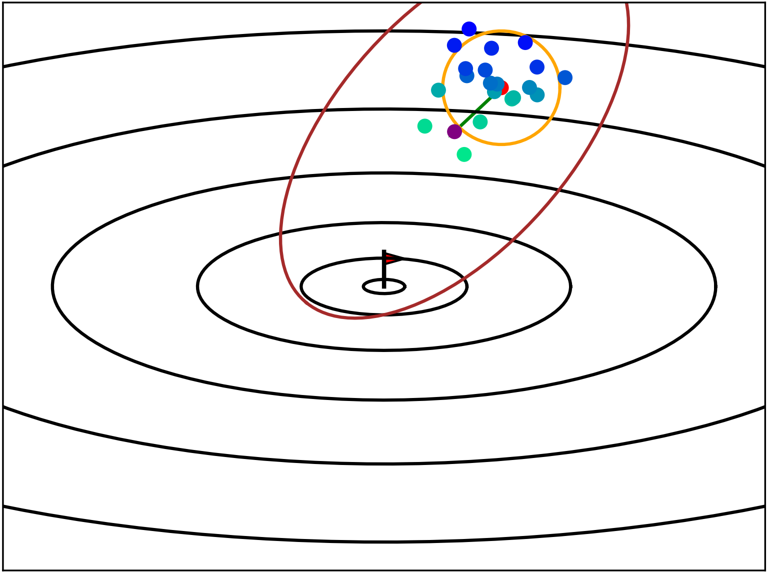












$$x_{t+1}^{1}, \ldots, x_{t+1}^{\lambda} \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$$

where C_t is the covariance matrix at iteration t

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

where C_t is the covariance matrix at iteration t

$x_{t+1}^{1:\lambda}$ is the best point among $\{x_{t+1}^{1}, \ldots, x_{t+1}^{\lambda}\}$

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

where C_t is the covariance matrix at iteration t

$x_{t+1}^{1:\lambda}$ is the best point among $\{x_{t+1}^1, \ldots, x_{t+1}^\lambda\}$

Idea: sample more in the direction $x_{t+1}^{1:\lambda} - m_t$ at iteration $t+1$

If v is a vector,

$\overleftrightarrow{v}$ is a matrix with range collinear to v

If v is a vector,

$\overleftrightarrow{v}$ is a matrix with range collinear to v:

$$\overleftrightarrow{v} = v \otimes v = v \times v^\top$$

If v is a vector,

$\overleftrightarrow{v}$ is a matrix with range collinear to v:

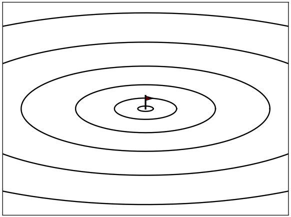
$$\overleftrightarrow{v} = v \otimes v = v \times v^\top$$

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

favors more the sampling in the direction $x_{t+1}^{1:\lambda} - m_t$ than C_t

Algorithm 4 ES with covariance matrix adaptation

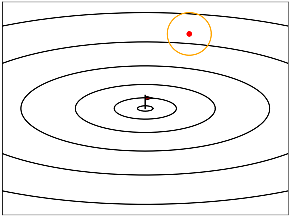
Goal: $\min_{x \in \mathbb{R}^d} f(x)$



Algorithm 4 ES with covariance matrix adaptation

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

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

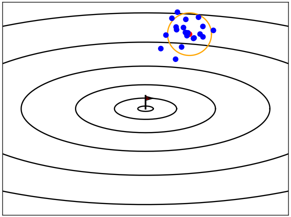


Algorithm 4 ES with covariance matrix adaptation

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

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

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



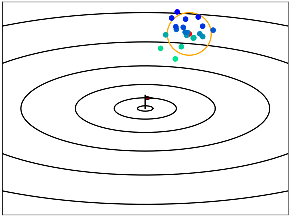
$\lambda =$ population size

Algorithm 4 ES with covariance matrix adaptation

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

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

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



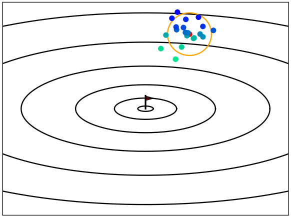
$\lambda =$ population size

Algorithm 4 ES with covariance matrix adaptation

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

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

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



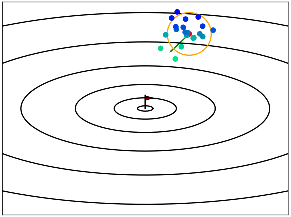
$\lambda =$ population size

Algorithm 4 ES with covariance matrix adaptation

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

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

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



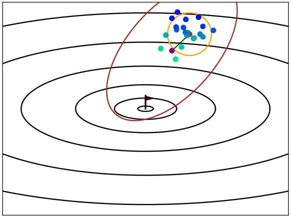
λ = population size

μ = parent number

Algorithm 4 ES with covariance matrix adaptation

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

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



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

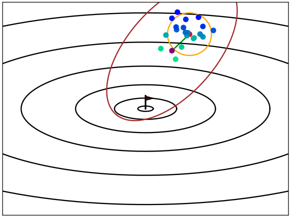
λ = population size

μ = parent number

Algorithm 4 ES with covariance matrix adaptation

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

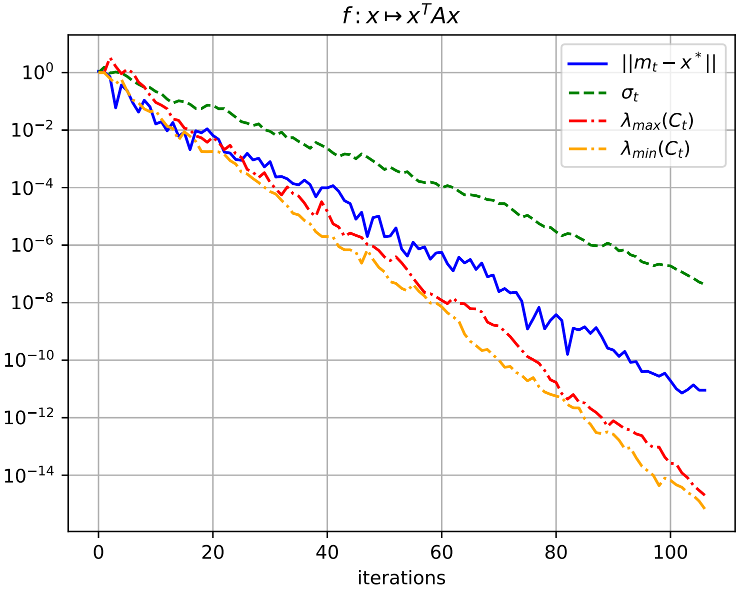
Repeat $(m_t \in \mathbb{R}^d,\ \sigma_t > 0,\ C_t \succ 0)$



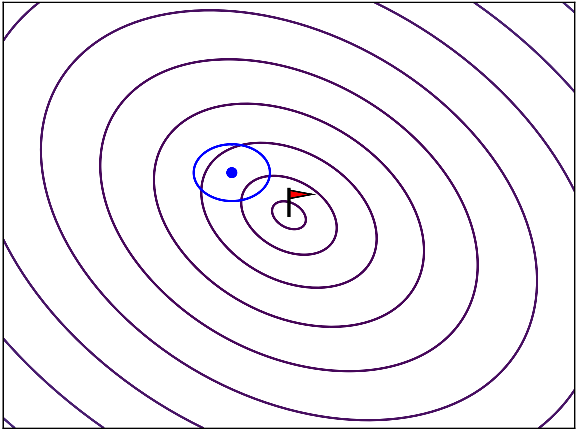
1. $x_{t+1}^1, \ldots, x_{t+1}^\lambda \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$
2. $\mathrm{sort}\ f(x_{t+1}^i)$:
   $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$
3. $m_{t+1} = \text{Average}(x_{t+1}^{1:\lambda}, \ldots, x_{t+1}^{\mu:\lambda})$
4.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

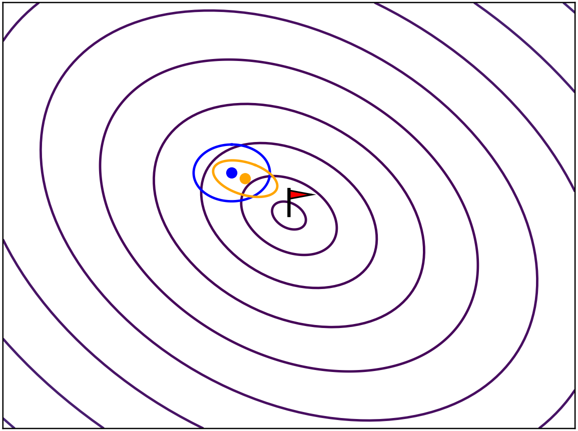
$\lambda =$ population size

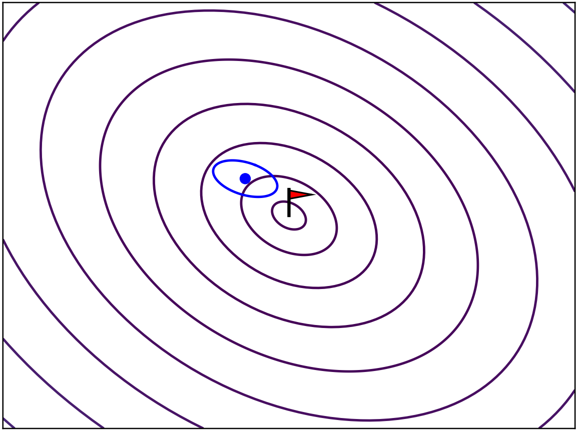
$\mu =$ parent number

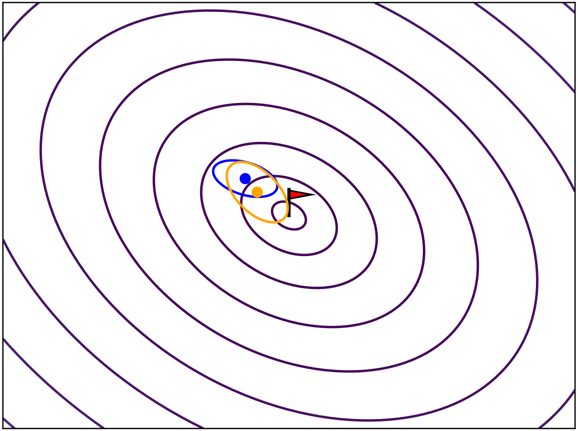


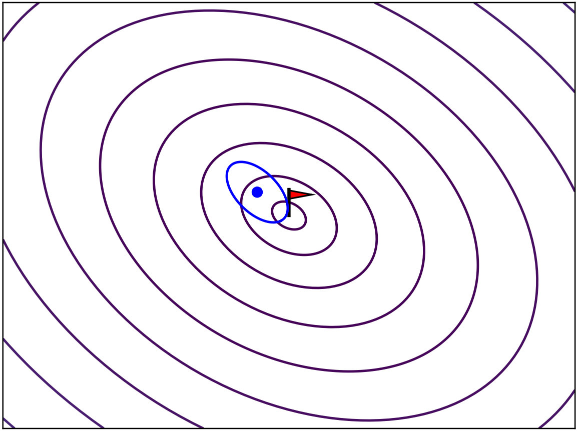
$f: x \mapsto x^T A x$
$||m_t - x^*||$
σ_t
$\lambda_{max}(C_t)$
$\lambda_{min}(C_t)$
10^{0}
10^{-2}
10^{-4}
10^{-6}
10^{-8}
10^{-10}
10^{-12}
10^{-14}
0
20
40
60
80
100
iterations

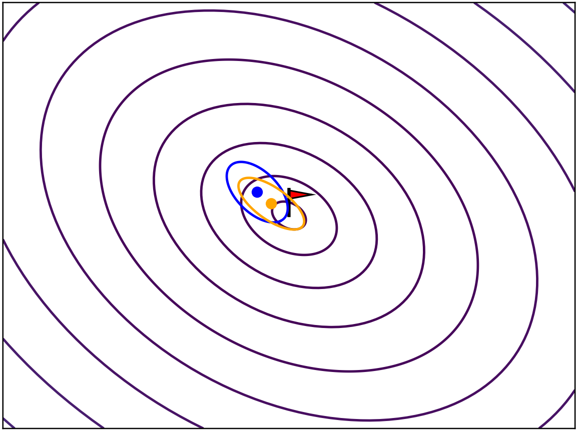


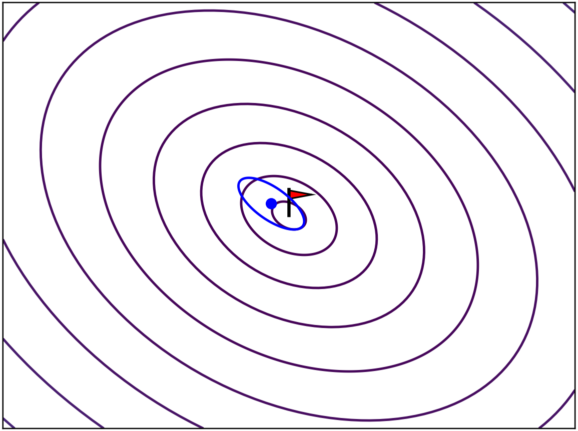


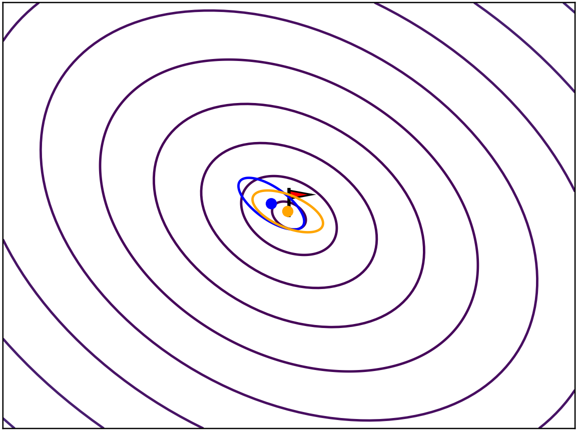


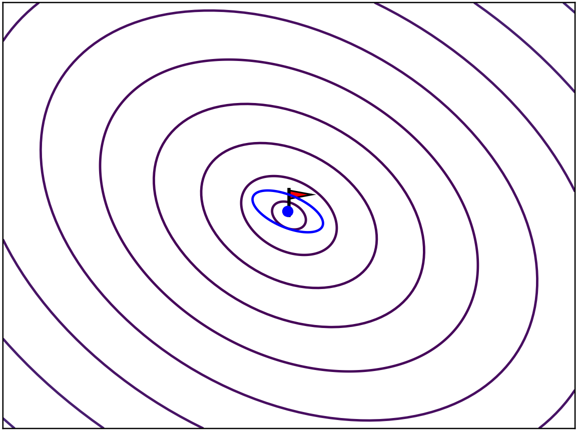


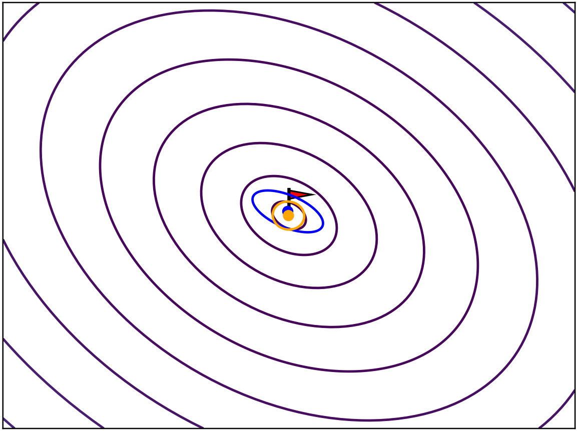


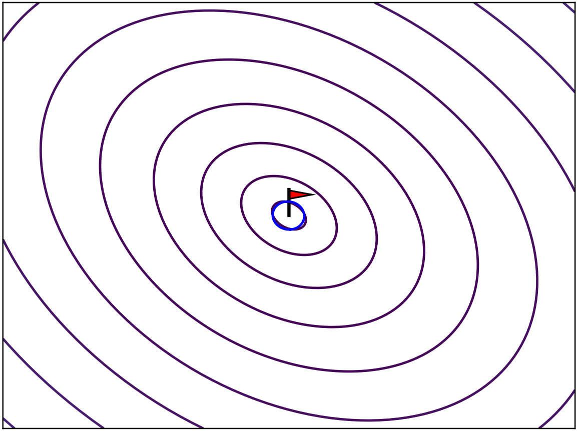


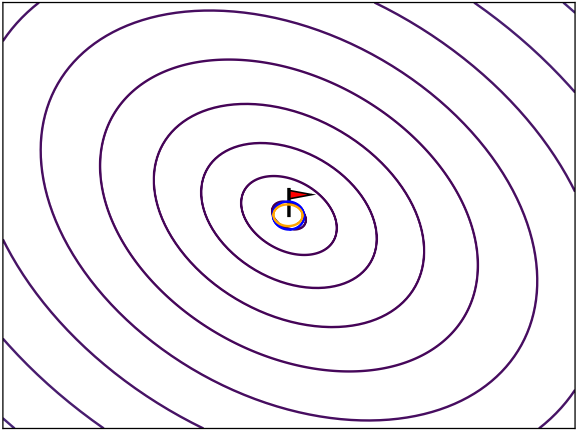


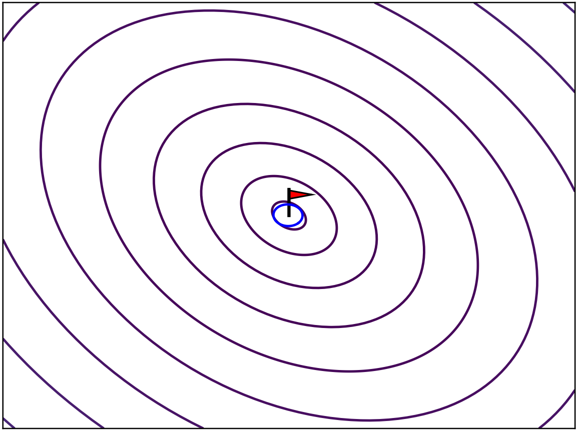


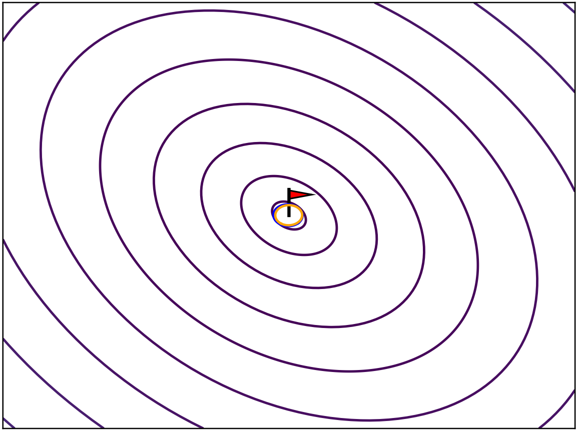


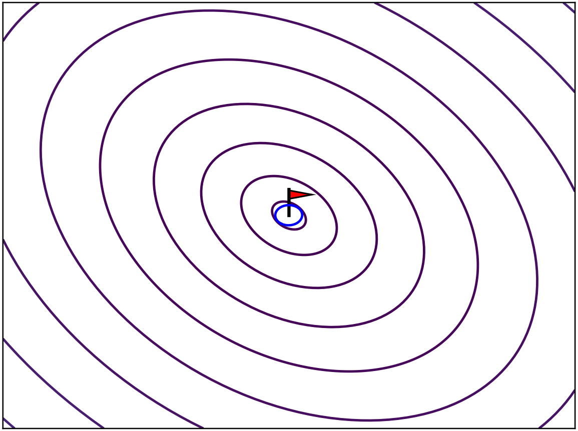


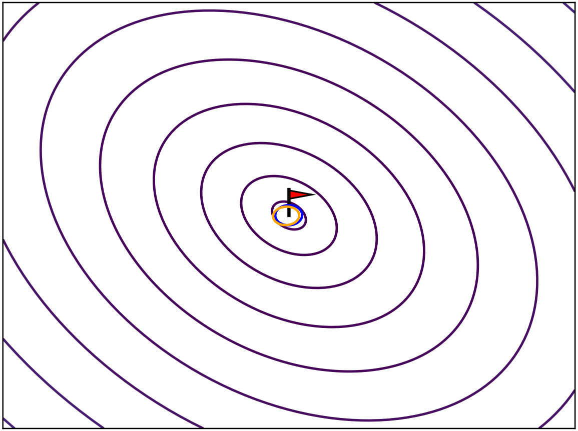


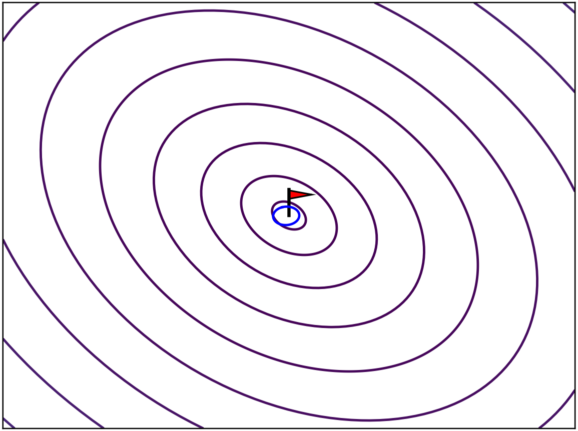


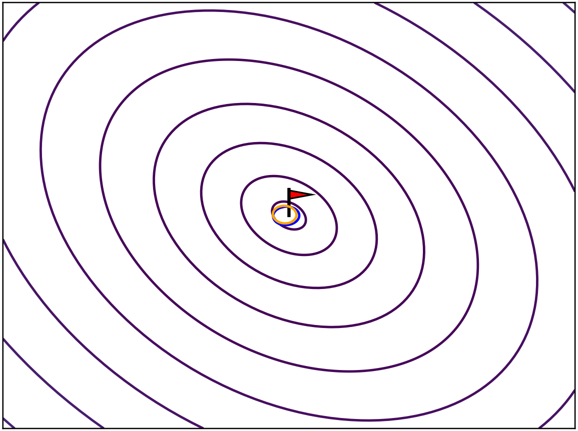


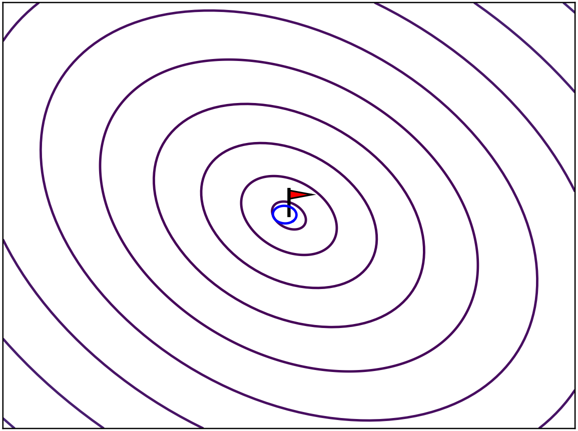


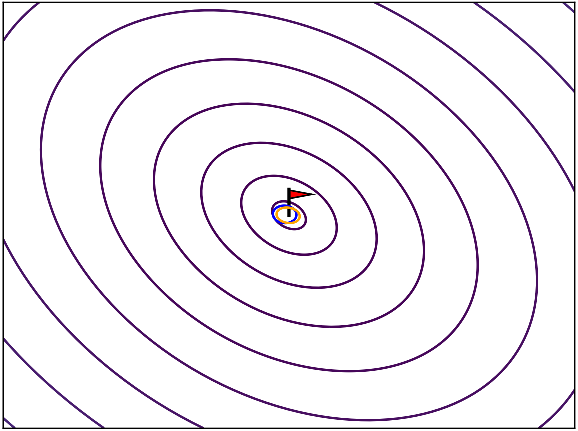


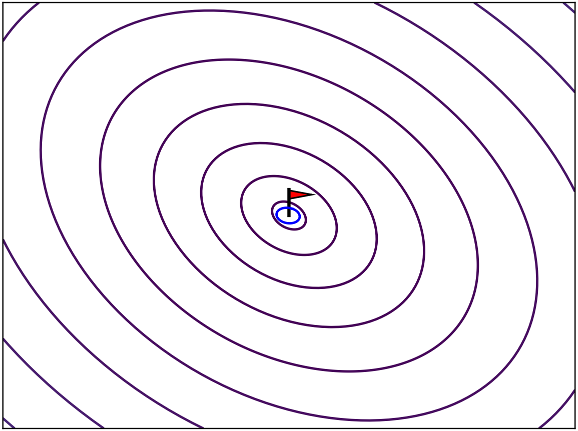


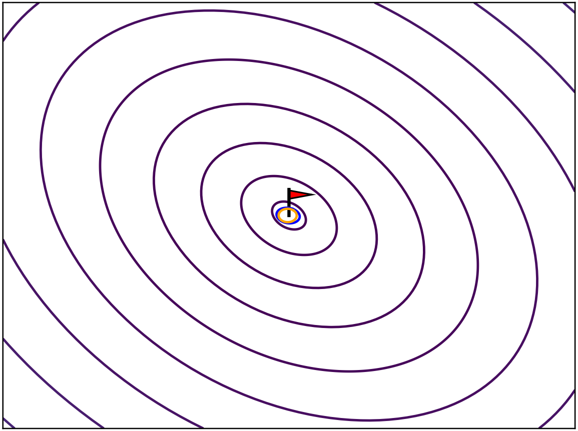


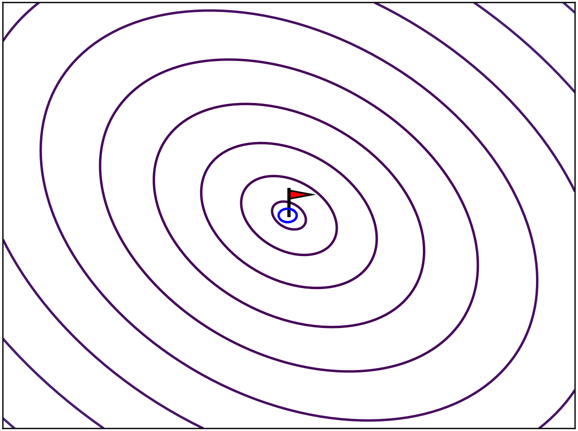


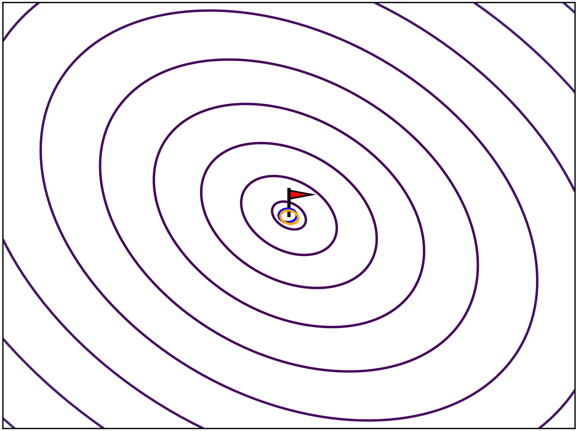


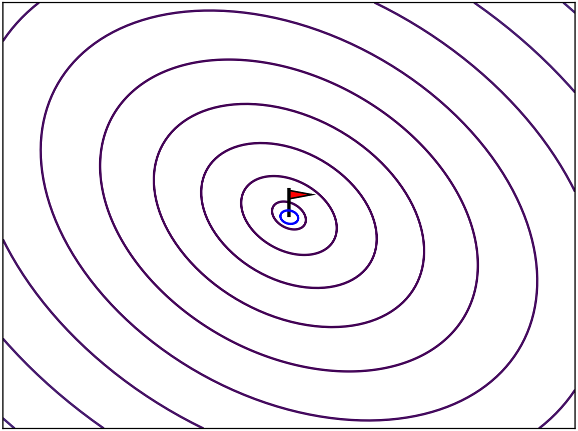


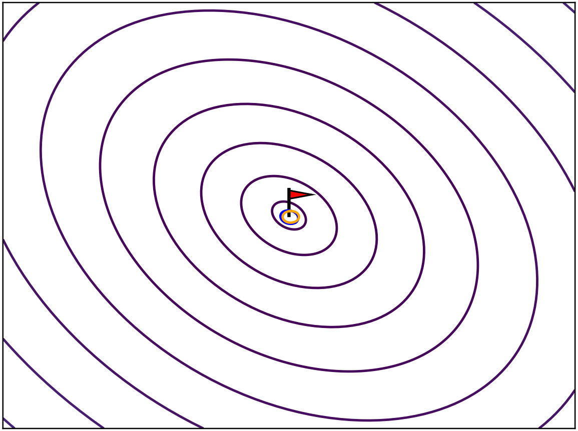












Observations:

$$\lim_{t \to \infty} m_t = x^*$$

Observations:

$$\lim_{t \to \infty} m_t = x^*$$

$$\lim_{t \to \infty} \sigma_t = \lim_{t \to \infty} C_t = 0$$

Observations:

$$\lim_{t\to\infty} m_t = x^*$$

$$\lim_{t\to\infty} \sigma_t = \lim_{t\to\infty} C_t = 0$$

and

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{C_t}{\text{normalization}}\right] \propto \text{Hessian}(f)^{-1}$$

Observations:

$$\lim_{t\to\infty} m_t = x^*$$

$$\lim_{t\to\infty} \sigma_t = \lim_{t\to\infty} C_t = 0$$

and

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{C_t}{\text{normalization}}\right] \propto \text{Hessian}(f)^{-1}$$

For the proof: we rely (again) on Markov chains

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

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

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

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

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

Proposition

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

Scheme of proof:

1. irreducibility and aperiodicity of $\{(z_t, \Sigma_t)\}_{t\in\mathbb{N}}$
2. drift condition: $\exists \mathsf{K} \subset \mathbb{R}^d \times \lambda_{\min}^{-1}(\{1\})$ compact and $V\colon \mathbb{R}^d \times \lambda_{\min}^{-1}(\{1\}) \to [1, +\infty]$

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

3. deduce convergence from the ergodicity

Scheme of proof:

1. **irreducibility and aperiodicity of** $\{(z_t, \Sigma_t)\}_{t\in\mathbb{N}}$
2. drift condition: $\exists \mathsf{K} \subset \mathbb{R}^d \times \lambda_{\min}^{-1}(\{1\})$ compact and $V\colon \mathbb{R}^d \times \lambda_{\min}^{-1}(\{1\}) \to [1, +\infty]$

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

3. deduce convergence from the ergodicity

θ^* is **attracting** when

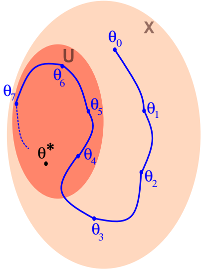
$$\exists x_1, x_2, \ldots, \lim_{k\to\infty} F_k(\theta_0, x_{1..k}) = \theta^*$$

Theorem
If

- $\exists \theta^*$ *attracting*
- $\exists x_1^*, \ldots, x_k^*$

such that $F_k(\theta^*, \cdot)$ *is a* ***submersion*** *at* $x_{1..k}^*$,

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



$$(z_{k+1}, \Sigma_{k+1}) = F(z_k, \Sigma_k, z_{k+1}^{1:\lambda}, \ldots, z_{k+1}^{\lambda:\lambda})$$

$$(z_{k+1}, \Sigma_{k+1}) = F(z_k, \Sigma_k, z_{k+1}^{1:\lambda}, \dots, z_{k+1}^{\lambda:\lambda})$$

Proposition
$(0, I_d)$ *is attracting*

$$(z_{k+1}, \Sigma_{k+1}) = F(z_k, \Sigma_k, z_{k+1}^{1:\lambda}, \ldots, z_{k+1}^{\lambda:\lambda})$$

Proposition
$(0, I_d)$ *is attracting, and* $F_k(0, I_d, \cdot)$ *is submersive somewhere.*

$$(z_{k+1}, \Sigma_{k+1}) = F(z_k, \Sigma_k, z_{k+1}^{1:\lambda}, \dots, z_{k+1}^{\lambda:\lambda})$$

Proposition
$(0, I_d)$ *is attracting, and* $F_k(0, I_d, \cdot)$ *is submersive somewhere.*

Corollary
If $f \in \Big\{$ , , ,  $\Big\}$, $\{z_t, \Sigma_t\}_{t\in\mathbb{N}}$ *is irreducible and aperiodic.*

Scheme of proof:

1. irreducibility and aperiodicity of $\{(z_t, \Sigma_t)\}_{t\in\mathbb{N}}$
2. **drift condition:** $\exists \mathsf{K} \subset \mathbb{R}^d \times \lambda_{\min}^{-1}(\{1\})$ **compact and** $V\colon \mathbb{R}^d \times \lambda_{\min}^{-1}(\{1\}) \to [1, +\infty]$

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

3. deduce convergence from the ergodicity

$$V(z, \Sigma) = \text{linear combination}(\|z\|^2, \|\Sigma\|)$$

$$V(z, \Sigma) = \text{linear combination}(\|z\|^2, \|\Sigma\|)$$

Proposition

If $f =$ , *then:*

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

when $\|z_t\| \gg 1$ *or* $\|\Sigma_t\| \gg 1$

When $\|\Sigma_t\| \gg \|z_t\|^2$:

When $\lambda_{\max}(\Sigma_t) \gg \|z_t\|^2$:

When $\lambda_{\max}(\Sigma_t) \gg \|z_t\|^2$: we want

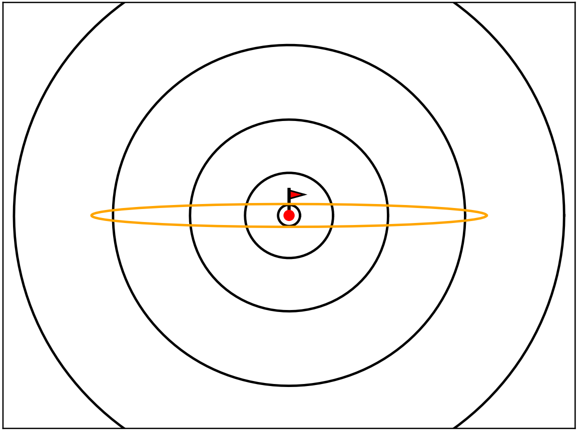
$$\mathbb{E}_t \left[\text{linear combination}(\|z_{t+1}\|^2, \|\Sigma_{t+1}\|)\right] \leqslant (1-\varepsilon) \times \text{linear combination}(\|z_t\|^2, \|\Sigma_t\|)$$

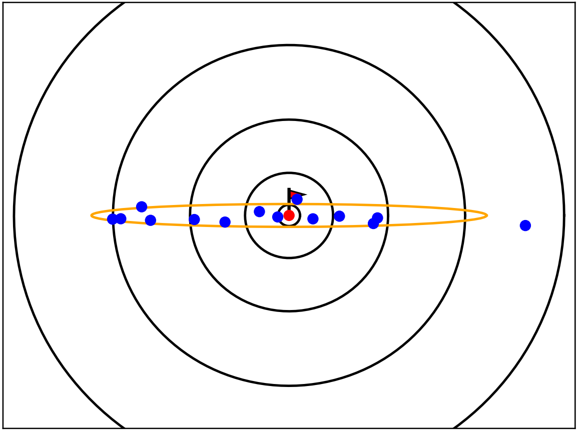
When $\lambda_{\max}(\Sigma_t) \gg \|z_t\|^2$: we want

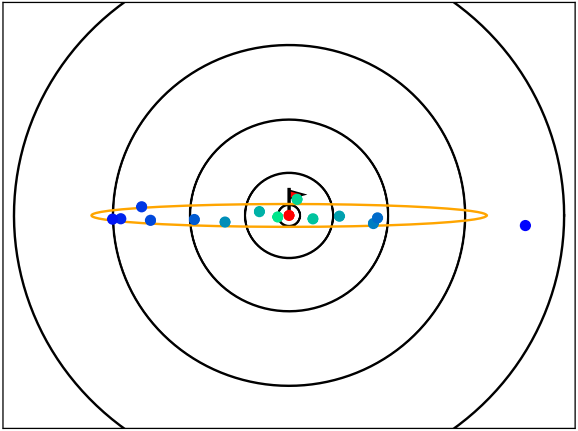
$$\mathbb{E}_t\left[\|\Sigma_{t+1}\|\right] \leqslant (1-\varepsilon) \times \|\Sigma_t\|$$

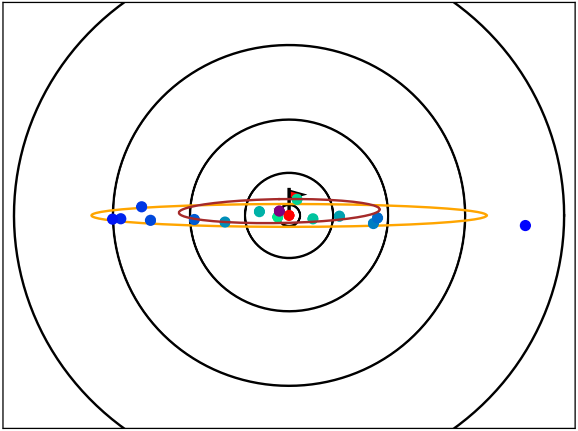
When $\lambda_{\max}(\Sigma_t) \gg \|z_t\|^2$: we want

$$\mathbb{E}_t\left[\lambda_{\max}(\Sigma_{t+1})\right] \leqslant (1-\varepsilon) \times \lambda_{\max}(\Sigma_t)$$

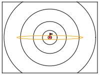








Proposition

When  *:*

$$\mathbb{E}\left[\lambda_{\max}(\Sigma_{t+1})\right] \leqslant (1-\varepsilon) \times \lambda_{\max}(\Sigma_t)$$

When $\|z_t\|^2 \gg \|\Sigma_t\|$:

When $\|z_t\|^2 \gg \|\Sigma_t\|$:   we want

$$\mathbb{E}_t \left[\text{linear combination}(\|z_{t+1}\|^2, \|\Sigma_{t+1}\|)\right] \leqslant (1-\varepsilon) \times \text{linear combination}(\|z_t\|^2, \|\Sigma_t\|)$$

When $\|z_t\|^2 \gg \|\Sigma_t\|$: we want

$$\mathbb{E}_t \left[\|z_{t+1}\|^2 \right] \leqslant (1 - \varepsilon) \times \|z_t\|^2$$

When $\|z_t\|^2 \gg \|\Sigma_t\|$: we want

$$\mathbb{E}_t\left[\|z_{t+1}\|^2\right] \leqslant (1-\varepsilon) \times \|z_t\|^2$$

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

When $\|z_t\|^2 \gg \|\Sigma_t\|$: we want

$$\mathbb{E}_t\left[\|z_{t+1}\|^2\right] \leqslant (1-\varepsilon) \times \|z_t\|^2$$

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

$$z_{t+1} = \frac{\text{Update mean}(z_t)}{\text{normalization}}$$

When $\|z_t\|^2 \gg \|\Sigma_t\|$: we want

$$\mathbb{E}_t\left[\|z_{t+1}\|^2\right] \leqslant (1-\varepsilon) \times \|z_t\|^2$$

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

$$z_{t+1} = \frac{\text{Update mean}(z_t)}{\text{normalization}}$$

$$\text{normalization} = \frac{\sigma_{t+1}}{\sigma_t}\sqrt{\frac{\lambda_{\min}(C_{t+1})}{\lambda_{\min}(C_t)}}$$

When $\|z_t\|^2 \gg \|\Sigma_t\|$: we want

$$\mathbb{E}_t\left[\|z_{t+1}\|^2\right] \leqslant (1-\varepsilon) \times \|z_t\|^2$$

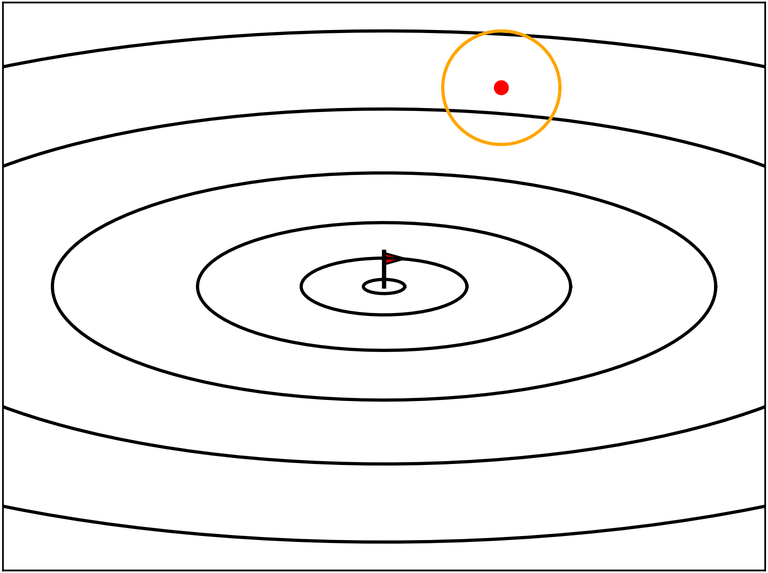
$$z_t = \frac{m_t - x^*}{\sigma_t \sqrt{\lambda_{\min}(C_t)}}$$

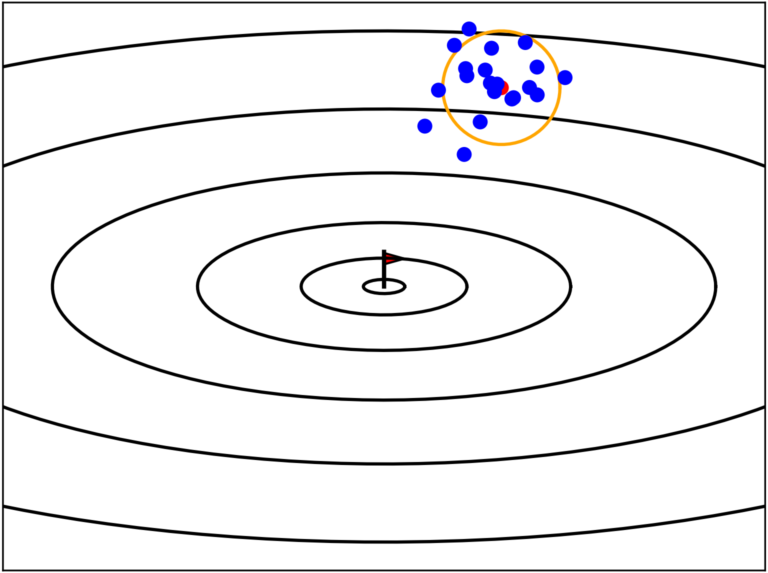
$$z_{t+1} = \frac{\text{Update mean}(z_t)}{\text{normalization}}$$

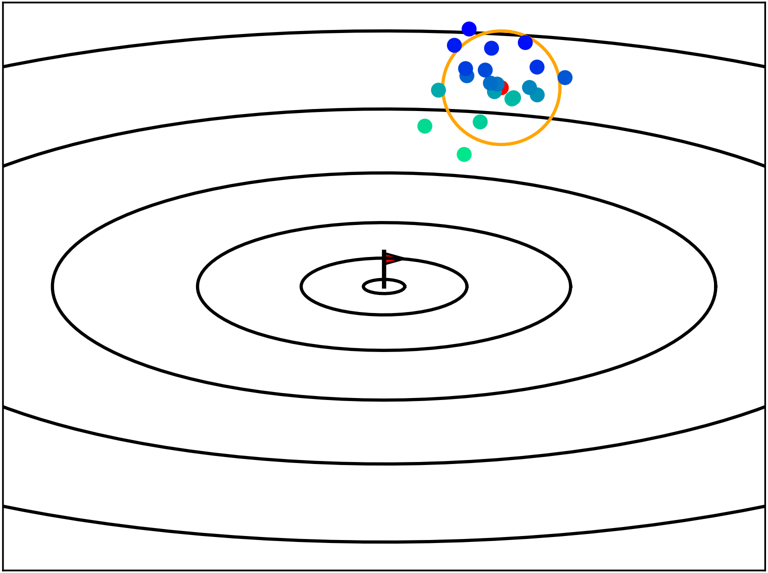
$$\text{normalization} = \underbrace{\frac{\sigma_{t+1}}{\sigma_t}} \sqrt{\frac{\lambda_{\min}(C_{t+1})}{\lambda_{\min}(C_t)}}$$

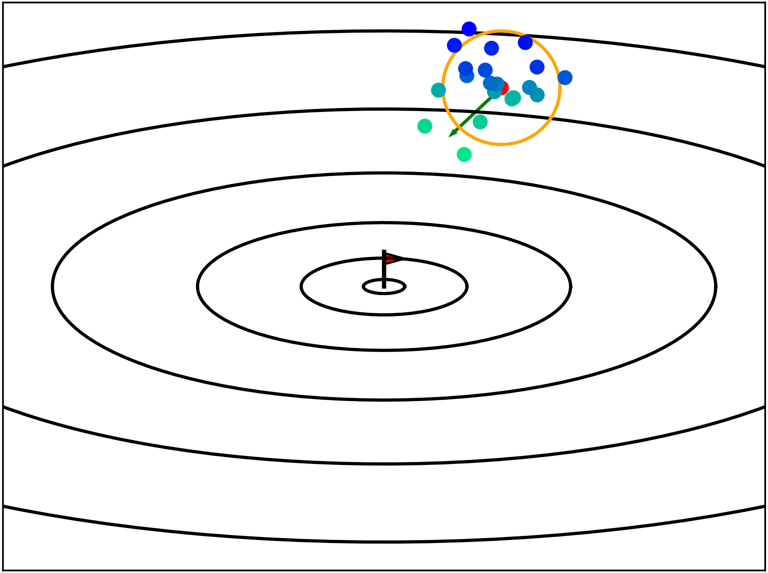
$$\sigma_{t+1} = \sigma_t \times \text{increasing function}(\|m_{t+1} - m_t\|)$$

$$\frac{\sigma_{t+1}}{\sigma_t} = \text{increasing function}(\|z_{t+1} - z_t\|)$$



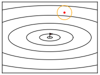






$$\frac{\sigma_{t+1}}{\sigma_t} = \text{increasing function}(\|z_{t+1} - z_t\|)$$

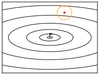
Proposition

When  *:*

$$\mathbb{E}[\|z_{t+1} - z_t\|] \geqslant \textit{constant}.$$

$$\frac{\sigma_{t+1}}{\sigma_t} = \text{increasing function}(\|z_{t+1} - z_t\|)$$

Proposition

When  *:*

$$\mathbb{E}[\|z_{t+1} - z_t\|] \geqslant \textit{constant}.$$

Corollary

$\exists$ *increasing function s.t.:*

$$\mathbb{E}\left[\|z_{t+1}\|^2\right] \leqslant (1-\varepsilon) \times \|z_t\|^2$$

$$V(z, \Sigma) = \text{linear combination}(\|z\|^2, \|\Sigma\|)$$

Proposition

If $f =$ , *then:*

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

when $\|z_t\| \gg 1$ *or* $\|\Sigma_t\| \gg 1$

Corollary

If $f =$ , $\{(z_t, \Sigma_t)\}_{t \in \mathbb{N}}$ *is ergodic.*

Scheme of proof:

1. irreducibility and aperiodicity of $\{(z_t, \Sigma_t)\}_{t\in\mathbb{N}}$
2. drift condition: $\exists \mathsf{K} \subset \mathbb{R}^d \times \lambda_{\min}^{-1}(\{1\})$ compact and $V\colon \mathbb{R}^d \times \lambda_{\min}^{-1}(\{1\}) \to [1, +\infty]$

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

3. **deduce convergence from the ergodicity**

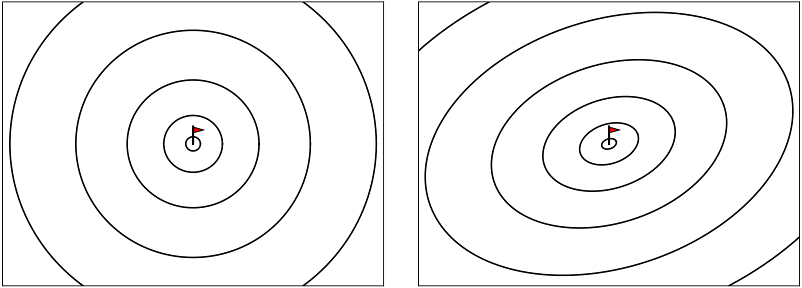
Theorem

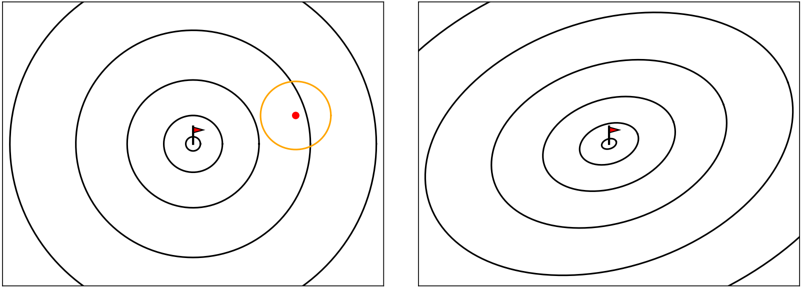
If $f =$ *, CMA-ES converges linearly (or geometrically).*

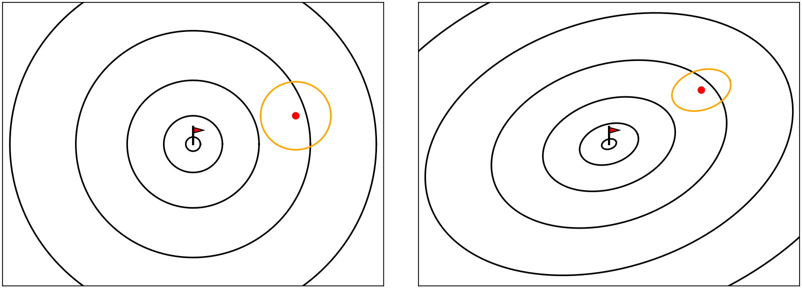
Theorem

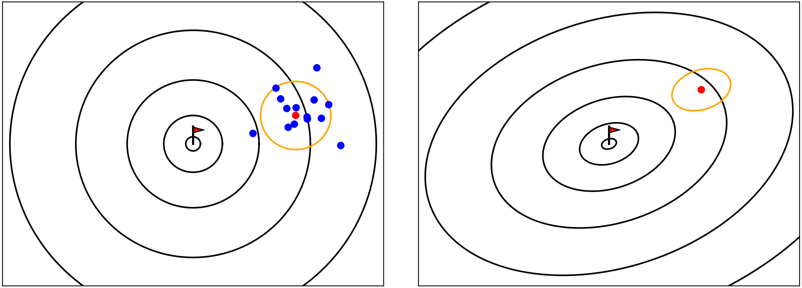
If $f =$ , CMA-ES converges linearly (or geometrically).

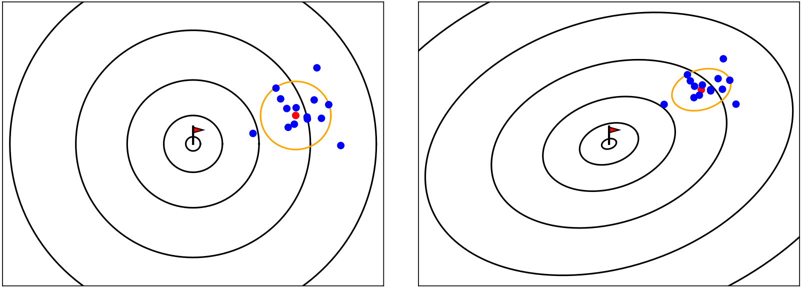
Question: how to extend to $f =$ ?

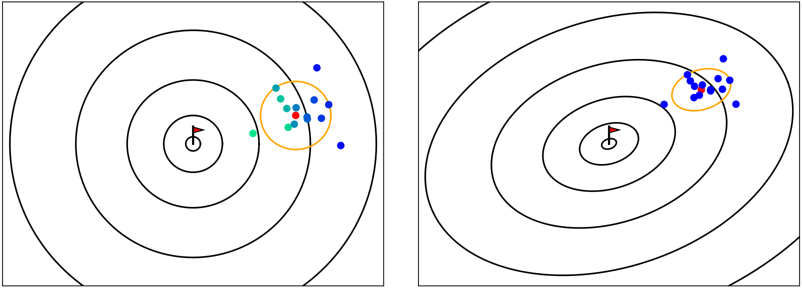


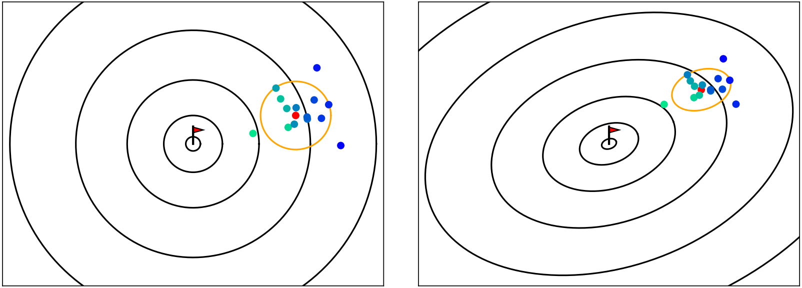


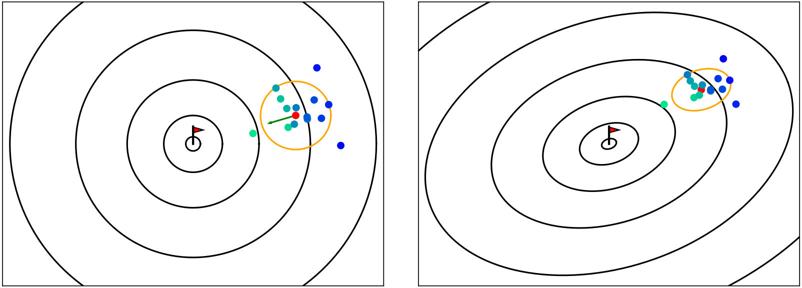


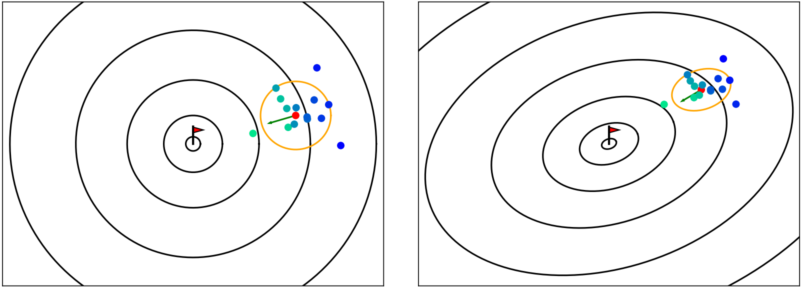


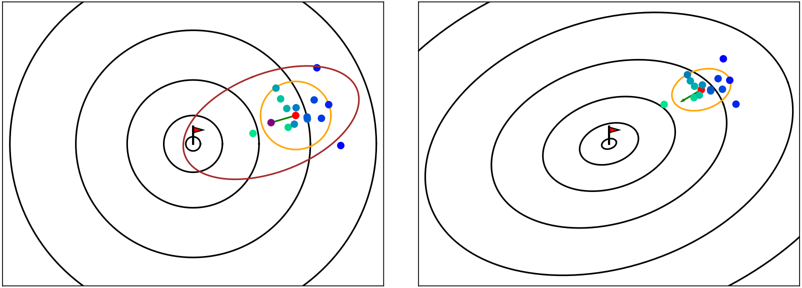


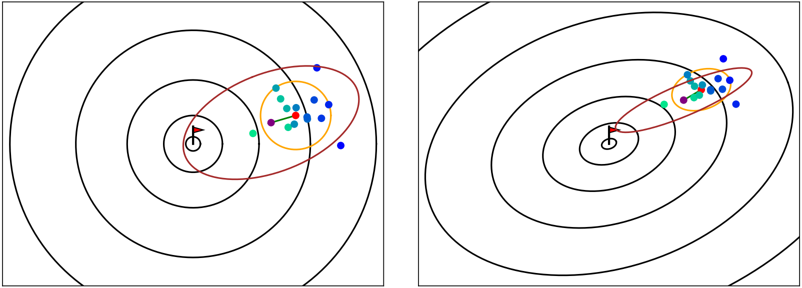


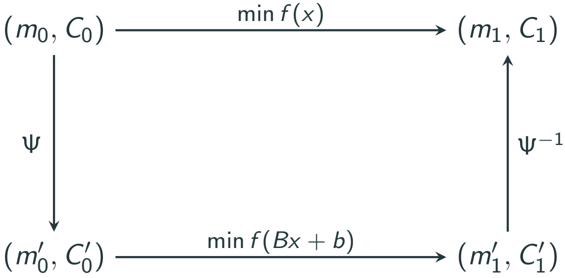








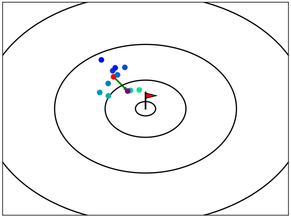



(m_0, C_0)
$\min f(x)$
(m_1, C_1)
Ψ
Ψ^{-1}
(m'_0, C'_0)
$\min f(Bx + b)$
(m'_1, C'_1)

Algorithm 1 Our first ES

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

Repeat: (Given $m_t \in \mathbb{R}^d$)



1. $x_{t+1}^1, \ldots, x_{t+1}^\lambda \sim \mathcal{N}(m_t, I_d)$
2. Rank population:
   $f\left(x_{t+1}^{1:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right)$
3. Update mean: $m_{t+1} = \text{Average}(x_{t+1}^{1:\lambda}, \ldots, x_{t+1}^{\mu:\lambda})$

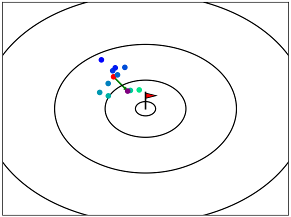
$\lambda =$ population size

$\mu =$ parent number

Algorithm 1 Our first ES

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

Repeat: (Given $m_t \in \mathbb{R}^d$, $C_t \in \mathcal{S}^d_{++}$)



1. $x^1_{t+1}, \ldots, x^\lambda_{t+1} \sim \mathcal{N}(m_t, C_t)$
2. Rank population:
   $f\left(x^{1:\lambda}_{t+1}\right) \leqslant \cdots \leqslant f\left(x^{\lambda:\lambda}_{t+1}\right)$
3. Update mean: $m_{t+1} = \text{Average}(x^{1:\lambda}_{t+1}, \ldots, x^{\mu:\lambda}_{t+1})$
4. $C_{t+1} = C_t$

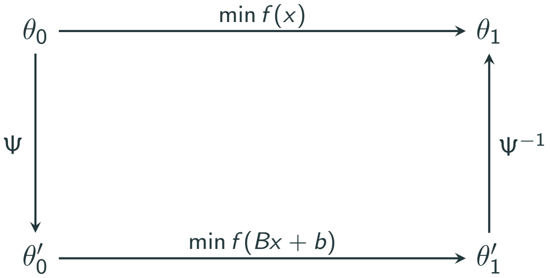
$\lambda =$ population size

$\mu =$ parent number

Definition

An algorithm $\Theta = \{\theta_t\}_{t \in \mathbb{N}}$ is affine-invariant if

(i)



(ii) From θ_0, Θ can reach a trajectory which starts at $\theta'_0 = \Psi(\theta_0)$.

Theorem
CMA-ES is affine-invariant.

Theorem
CMA-ES is affine-invariant.

Theorem
If $f \in \{$ ,  $\}$*:*

$$\lim_{t \to \infty} m_t = x^* \quad \textit{geometrically}$$

Theorem

$f \in \Big\{$ ,  $\Big\}$:

$$\lim_{t \to +\infty} \mathbb{E}\left[\frac{C_t}{normalization}\right] \propto Hessian^{-1}(f)$$

Proof.

□

Theorem

$f \in \{$ ,  $\}$:

$$\lim_{t \to +\infty} \mathbb{E}\left[\frac{C_t}{normalization}\right] \propto Hessian^{-1}(f)$$

Proof.

When $f =$ :

□

Theorem

$f \in \{$ ,  $\}$:

$$\lim_{t \to +\infty} \mathbb{E}\left[\frac{C_t}{normalization}\right] \propto Hessian^{-1}(f)$$

Proof.

When $f =$ :

$R \times$  $=$ 

for R a rotation matrix.

□

Theorem

$f \in \Big\{$ ,  $\Big\}$:

$$\lim_{t \to +\infty} \mathbb{E}\left[\frac{C_t}{normalization}\right] \propto Hessian^{-1}(f)$$

Proof.

When $f =$ :

$R \times$  $=$ 

for R a rotation matrix.

$$RC_tR^\top \quad \text{behaves like} \quad C_t$$

□

Theorem

$f \in \Big\{$ ,  $\Big\}$:

$$\lim_{t \to +\infty} \mathbb{E}\left[\frac{C_t}{\mathit{normalization}}\right] \propto \mathit{Hessian}^{-1}(f)$$

Proof.

When $f =$ :

$R \times$  $=$ 

for R a rotation matrix.

$$\lim_{t \to +\infty} \mathbb{E}\left[\frac{R C_t R^\top}{\text{normalization}}\right] = \lim_{t \to +\infty} \mathbb{E}\left[\frac{C_t}{\text{normalization}}\right]$$

□

Theorem

$f \in \Big\{$ ,  $\Big\}$:

$$\lim_{t \to +\infty} \mathbb{E}\left[\frac{C_t}{normalization}\right] \propto Hessian^{-1}(f)$$

Proof.

When $f =$ :

$R \times$  $=$ 

for R a rotation matrix.

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

□

Theorem

$f \in \Big\{$  ,  $\Big\}$:

$$\lim_{t \to +\infty} \mathbb{E}\left[\frac{C_t}{normalization}\right] \propto Hessian^{-1}(f)$$

Proof.

When $f =$ :

$R \times$  $=$ 

for R a rotation matrix.

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

□

Theorem

$f \in \Big\{$ ,  $\Big\}$:

$$\lim_{t\to+\infty} \mathbb{E}\left[\frac{C_t}{normalization}\right] \propto Hessian^{-1}(f)$$

Proof.

When $f =$ :

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

When $f =$ : □

Theorem

$f \in \{$ ,  $\}$:

$$\lim_{t \to +\infty} \mathbb{E}\left[\frac{C_t}{\textit{normalization}}\right] \propto \textit{Hessian}^{-1}(f)$$

Proof.

When $f =$ :

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

When $f =$ :

C_t behaves like $\text{Hessian}^{-1/2} C_t($  $)\text{Hessian}^{-1/2}$

□

Theorem

$f \in \{$ ,  $\}$:

$$\lim_{t \to +\infty} \mathbb{E}\left[\frac{C_t}{normalization}\right] \propto Hessian^{-1}(f)$$

Proof.

When $f =$ :

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

When $f =$ :

$$\lim_{t \to +\infty} \mathbb{E}\left[\frac{C_t}{\text{normalization}}\right] \propto \text{Hessian}^{-1}(f)$$

$\square$

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

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

Goal: remember previous iterations to update σ.

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

Goal: remember previous iterations to update σ.

$$\text{path}^{\sigma}_{t+1} = \text{linear function}\left(\text{path}^{\sigma}_{t}, m_{t+1} - m_t\right)$$

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

Goal: remember previous iterations to update σ.

$$\text{path}^{\sigma}_{t+1} = \text{linear function}\left(\text{path}^{\sigma}_{t}, m_{t+1} - m_t\right)$$

New update:

$$\boxed{\sigma_{t+1} = \sigma_t \times \text{increasing function}(\|\text{path}^{\sigma}_{t+1}\|)}$$

$$\text{path}^c_{t+1} = \text{linear function}\left(\text{path}^c_t, m_{t+1} - m_t\right)$$

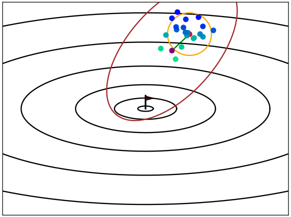
$$\text{path}^c_{t+1} = \text{linear function}\left(\text{path}^c_t, m_{t+1} - m_t\right)$$

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

Algorithm 5 CMA-ES

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

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



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

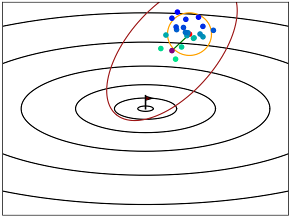
$\lambda =$ population size

$\mu =$ parent number

Algorithm 5 CMA-ES

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

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



1. $x^1_{t+1}, \dots, x^\lambda_{t+1} \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$
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. $m_{t+1} = \text{Average}(x^{1:\lambda}_{t+1}, \dots, x^{\mu:\lambda}_{t+1})$
4. $\text{path}^\sigma_{t+1} = \text{Linear}(\text{path}^\sigma_t, m_{t+1} - m_t)$
   $\sigma_{t+1} = \sigma_t \times \text{increasing function}\left(\|m_{t+1} - m_t\|\right)$
5. $C_{t+1} = \text{Linear combination}\left(C_t, \text{Average}\left[\overleftrightarrow{(x^{i:\lambda}_{t+1} - m_t)}\right]\right)$

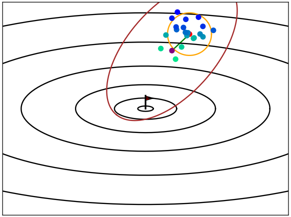
λ = population size

μ = parent number

Algorithm 5 CMA-ES

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

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



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

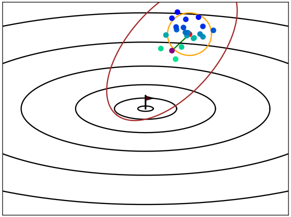
λ = population size

μ = parent number

Algorithm 5 CMA-ES

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

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



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

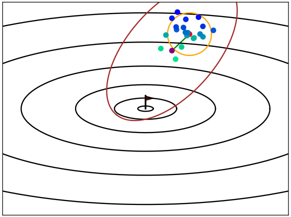
$\lambda =$ population size

$\mu =$ parent number

Algorithm 5 CMA-ES

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

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



1. $x^1_{t+1}, \ldots, x^\lambda_{t+1} \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$
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. $m_{t+1} = \text{Average}(x^{1:\lambda}_{t+1}, \ldots, x^{\mu:\lambda}_{t+1})$
4. $\text{path}^\sigma_{t+1} = \text{Linear}(\text{path}^\sigma_t, m_{t+1} - m_t)$
   $\sigma_{t+1} = \sigma_t \times \text{increasing function}\left(\|\text{path}^\sigma_{t+1}\|\right)$
5. $\text{path}^c_{t+1} = \text{Linear}(\text{path}^c_t, m_{t+1} - m_t)$
   $C_{t+1} = \text{Linear combination}\left(C_t, \text{Average}\left[\overleftrightarrow{(x^{i:\lambda}_{t+1} - m_t)}\right], \overleftrightarrow{\text{path}^c_{t+1}}\right)$

$\lambda =$ population size

$\mu =$ parent number

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

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

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

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

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

$$\text{normalized path}_t^{c,\sigma} = \frac{\text{path}_t^{c,\sigma}}{\text{normalization}}$$

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

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

$$\text{normalized path}_t^{c,\sigma} = \frac{\text{path}_t^{c,\sigma}}{\text{normalization}}$$

Proposition

If $f \in \Big\{$ , , ,  $\Big\}$, *then* $\{(z_t, \Sigma_t, n.\ path_t^c, n.\ path_t^\sigma)\}_{t\in\mathbb{N}}$ *is a Markov chain.*

When X is infinite:

Theorem
If $\{\theta_t\}_{t\in\mathbb{N}}$ *is an irreducible, aperiodic Markov chain, then it is ergodic if*

$$\mathbb{E}[V(\theta_{t+1}) \mid \theta_t] \leqslant (1-\varepsilon)V(\theta_t) \qquad \textit{if } \theta_t \notin \textit{compact set}$$

for some $V\colon \mathsf{X} \to [1, +\infty]$.

When X is infinite:

Theorem
If $\{\theta_t\}_{t\in\mathbb{N}}$ is an irreducible, aperiodic Markov chain, then it is ergodic if

$$\mathbb{E}[V(\theta_{t+n(\theta_t)}) \mid \theta_t] \leqslant (1-\varepsilon)^{n(\theta_t)} V(\theta_t) \qquad \text{if } \theta_t \notin \text{compact set}$$

for some $V\colon \mathsf{X} \to [1, +\infty]$.

$$V(z, \Sigma, \text{n. path}^c, \text{n. path}^\sigma) = \text{linear combination}(\|z\|^2, \|\Sigma\|, \|\text{n. paths}\|^2)$$

$$V(z, \Sigma, \text{n. path}^c, \text{n. path}^\sigma) = \text{linear combination}(\|z\|^2, \|\Sigma\|, \|\text{n. paths}\|^2)$$

Proposition

If $f =$ *, then:*

$$\mathbb{E}[V(\textit{iteration}_{t+k}) \mid \textit{iteration}_t] \leqslant (1-\varepsilon) \times V(\textit{iteration}_t)$$

when $\|z_t\| \gg 1$ *or* $\|\Sigma_t\| \gg 1$ *or* $\|\textit{n. path}_t^c\| \gg 1$ *or* $\|\textit{n. path}_t^\sigma\| \gg 1$.

Theorem

If $f \in \Big\{$  ,  $\Big\}$*:*

$$\lim_{t \to \infty} m_t = x^* \quad \textit{geometrically}$$

and

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

Thank you!