A linear convergence proof of the isotropic ES via a continuous time approximation

Youhei Akimoto Anne Auger Nikolaus Hansen

TAO - INRIA Saclay

Oct. 19, 2012

Motivation

- Adaptive Evolution Strategies sample from a Gaussian distribution and adapt the mean vector and the covariance matrix
- Adaptive Evolution Strategies including CMA-ES are successfully applied in practice and there are lots of empirical evidence of linear convergence on a wide class of functions.
- However, their linear convergence to a local optimum is so far only proven for simple algorithms compared to CMA-ES [Auger 05, Jagerskupper 06, 07].

Motivation

- In this work, we seek a new methodology to show linear convergence.
  - *theory of stochastic approximation (ODE method)*
- An isotropic variant of adaptive-ES derived from IGO is studied.
- We derive
  - global linear convergence in expectation on monotonic convex-quadratic-composite functions
  - local linear convergence w.p. $1 - \varepsilon$ on monotonic $\mathcal{C}^2$-composite functions

Information-Geometric Optimization (IGO) Algorithm

Choose a probabilistic model $\{P_\theta,\ \theta \in \Theta\}$ on $\mathbb{X}$ and initialize θ_0.

1. Draw samples x_i, $i = [\![1, \lambda]\!]$, independently from P_{θ_n}.

2. Evaluate $f(x_i)$ and compute ranking
$\mathrm{rk}(x_i) = \#\{j \in [\![1, \lambda]\!] : f(x_j) \leqslant f(x_i)\}$.

3. Update parameters as

$$\theta_{n+1} = \theta_n + \eta \sum_{i=1}^{\lambda} \underbrace{w_{\mathrm{rk}(x_i)}}_{\text{nonincreasing weights}} \overbrace{\tilde{\nabla}_\theta \ln P_{\theta_n}(x_i)}^{\text{natural gradient of log-likelihood}} .$$

4. Go back to step 1 unless a termination criterion is satisfied.

$\Longrightarrow$ PBIL if P_θ is Bernoulli on $\mathbb{X} = \{0,1\}^d$

$\Longrightarrow$ Pure rank-μ CMA-ES if P_θ is Gaussian $\mathcal{N}(m, C)$ on $\mathbb{X} = \mathbb{R}^d$

Isotropic ES

= IGO algorithm with $P_\theta = \mathcal{N}(m, vI_d)$ on $\mathbb{X} = \mathbb{R}^d$.

Initialize $\theta_0 = (m_0, v_0)$, $m_0 \in \mathbb{R}^d$ and $v_0 > 0$.

1. Draw samples x_i, $i = [\![1, \lambda]\!]$, independently from $\mathcal{N}(m_n, v_n I_d)$.

2. Evaluate $f(x_i)$ and compute ranking
$\mathrm{rk}(x_i) = \#\{j \in [\![1, \lambda]\!] : f(x_j) \leqslant f(x_i)\}$.

3. Update parameters as

$$m_{n+1} = m_n + \eta_m \sum_{i=1}^{\lambda} w_{\mathrm{rk}(x_i)} (x_i - m_n)$$

$$v_{n+1} = v_n + \eta_v \sum_{i=1}^{\lambda} w_{\mathrm{rk}(x_i)} \left(\frac{\|x_i - m_n\|^2}{d} - v_n \right) .$$

4. Go back to step 1 unless a termination criterion is satisfied.

Invariance of Isotropic ES

Invariance under monotone transformation of f

$f(x) \mapsto g(f(x))$, $g : \mathbb{R} \to \mathbb{R}$ is a strictly increasing function.

Invariance under translation of $\mathbb{X}$

$x \mapsto a + x$, $a \in \mathbb{R}^d$.

Invariance under rotation of $\mathbb{X}$

$x \mapsto Rx$, $R \in \mathbb{R}^{d \times d}$ is an orthogonal matrix.

Monotonic Convex Quadratic Composite Functions

Objective function $f(x) = g\left((x - \mathbf{x}^*)^{\mathrm{T}} A (x - \mathbf{x}^*)\right)$

- $g : \mathbb{R} \to \mathbb{R}$ is a strictly increasing function.
- $A \in \mathbb{R}^{d \times d}$ is a positive definite symmetric matrix.
- $\mathbf{x}^* \in \mathbb{R}^d$ is the global minimum.

W.l.o.g. (thanks to the invariance properties), we assume

- g is the identity function
- $A = \mathrm{diag}(A_1, \ldots, A_d)$
- $\mathbf{x}^* = 0$.

Monotonic $\mathcal{C}^2$-Composite Functions

Objective function $f(x) = g(h(x))$

- $g : \mathbb{R} \to \mathbb{R}$ is a strictly increasing function.
- $h : \mathbb{R}^d \to \mathbb{R}$ is a $\mathcal{C}^2$ function.
- $\mathbf{x}^* \in \mathbb{R}^d$ is a local minimum of h.
- $A \in \mathbb{R}^{d \times d}$ is a positive definite symmetric Hessian of h at $\mathbf{x}^*$.

W.l.o.g. (thanks to the invariance properties), we assume

- g is the identity function
- $A = \mathrm{diag}(A_1, \ldots, A_d)$
- $\mathbf{x}^* = 0$.

Linear Convergence

aka log-linear convergence or geometric convergence

Definition (Linear convergence of x to $\mathbf{x}^*$ in $(\mathbb{X}, \rho_{\mathbb{X}})$)

Let $\{x_n\}$ be a sequence in $\mathbb{X}$, $\mathbf{x}^* \in \mathbb{X}$ and $\rho_{\mathbb{X}}$ a metric on $\mathbb{X}$. If there exists $0 < \gamma_1 < 1$ such that

$$\limsup_{n \to \infty} \frac{\rho_{\mathbb{X}}(x_{n+1}, \mathbf{x}^*)}{\rho_{\mathbb{X}}(x_n, \mathbf{x}^*)} = \gamma_1$$

or if there exists $\gamma_2 < 0$ such that

$$\limsup_{n \to \infty} \frac{1}{n} \ln \rho_{\mathbb{X}}(x_n, \mathbf{x}^*) = \gamma_2 \ ,$$

$\{x_n\}$ is said to converge linearly (aka log-linearly or geometrically) to $\mathbf{x}^*$ in $(\mathbb{X}, \rho_{\mathbb{X}})$.

Our Objective

- Expected distance of $x \sim P_{\theta_n}$ from the optimum $\mathbf{x}^*$ given θ_n

$$\mathbb{E}_{x \sim P_{\theta_n}} \left[\| x - \mathbf{x}^* \| \right] \leqslant \left(\| m_n - \mathbf{x}^* \|^2 + d \cdot v_n \right)^{1/2}$$

- Optimum parameter $\theta^* \in \overline{\Theta}$

$$\theta^* = (\mathbf{x}^*, 0) \in \partial\Theta$$

- Metric on parameter space $\overline{\Theta}$

$$\rho(\theta, \theta') := \left[\| m - m' \|^2 + d \left(\sqrt{v} - \sqrt{v'} \right)^2 \right]^{1/2}$$

$$\implies \rho(\theta, \theta^*) = \left(\| m_n - \mathbf{x}^* \|^2 + d \cdot v_n \right)^{1/2}$$

Linear convergence of θ_n to θ^* in (Θ, ρ)

$$\limsup_{n \to \infty} \frac{1}{n} \ln \mathbb{E}_{x \sim P_{\theta_n}} \left[\| x - \mathbf{x}^* \| \right] \leqslant \limsup_{n \to \infty} \frac{1}{n} \ln \rho(\theta_n, \theta^*) = \gamma < 0$$

w.p.1, in expectation, in probability, w.p. $1 - \varepsilon$, or etc.

Main Results

Under an assumption on w, (introduced later)

Global Linear Convergence on Monotonic Convex-Quadratic-Composite

For sufficiently small $\delta t > 0$, there exists $\gamma < 0$ such that

$$\limsup_{n\to\infty} \frac{1}{n} \ln \mathbb{E}\left[\rho(\theta_n, \theta^*)\right] = \gamma \ .$$

Local Linear Convergence on Monotonic $\mathcal{C}^2$-Composite

For $\theta_0 \in U$ and any $0 < \varepsilon < 1$, there exists $\overline{\delta t} > 0$ such that for $0 < \forall \delta t \leqslant \overline{\delta t}$

$$\Pr\left[\limsup_{n\to\infty} \frac{1}{n} \ln \rho(\theta_n, \theta^*) \leqslant \gamma\right] \geqslant 1 - \varepsilon \ ,$$

where the speed of convergence $\gamma < 0$ depends on δt.

Notation

Let $\theta_n = [m_n^{\mathrm{T}}, v_n]^{\mathrm{T}}$ and

$$\hat{F}_n = \begin{bmatrix} \hat{F}_n^m \\ \hat{F}_n^v \end{bmatrix} = \frac{1}{\lambda} \sum_{i=1}^{\lambda} w_{\mathrm{rk}(\mathrm{X}_i)} \begin{bmatrix} x_i - m_n \\ \|x_i - m_n\|^2 / d - v_n \end{bmatrix} .$$

Then, the isotropic ES is $\theta_{n+1} = \theta_n + \delta t \hat{F}_n$ $(\eta_m = \eta_v = \delta t)$.

Decompose

$$\hat{F}_n = F(\theta_n) + M_n \ ,$$

where

- $F(\theta) = \mathbb{E}[\hat{F}_n \mid \theta_n = \theta]$ is the *mean field*
- $M_n = \hat{F}_n - F(\theta_n)$ is the *martingale difference noise*

Then, the isotropic ES is $\theta_{n+1} = \theta_n + \delta t [F(\theta_n) + M_n]$.

Idea 1: Potential function

i.e., Drift function

Potential function

For some $\eta > 0$, define

$$\Psi_\eta(\theta) = \left(\eta \vee \frac{\|m\|}{\sqrt{v}}\right)^{1/2} \cdot \underbrace{(\|m\|^2 + d \cdot v)^{1/2}}_{=\rho(\theta,\theta^*)}$$

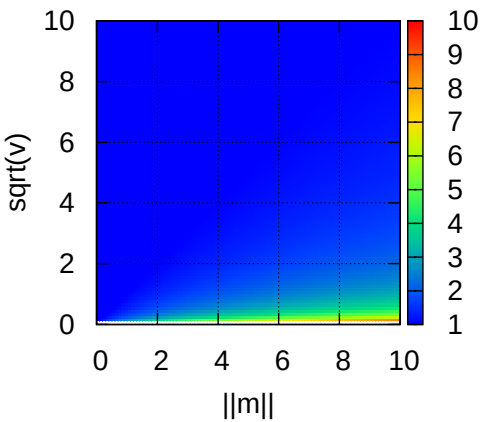
and show for some $N \in \mathbb{N}$ and $0 < \gamma < 1$

$$\mathbb{E}_n\left[\Psi_\eta(\theta_{N+n})\right] \leqslant \gamma \Psi_\eta(\theta_n) \ .$$

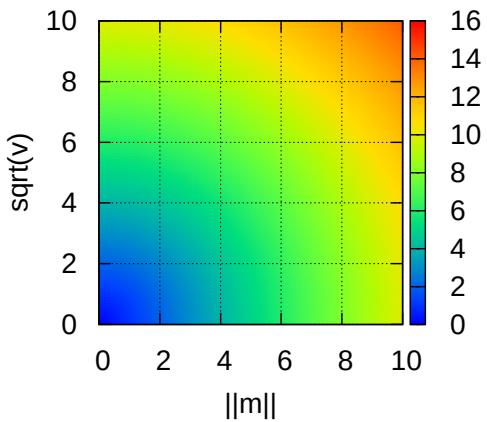
Note. Since $\rho(\theta, \theta^*) \leqslant \Psi_\eta(\theta_n)/\eta^{1/2}$,

$$\limsup_{n\to\infty} \frac{1}{n} \ln \mathbb{E}_{\theta_n}\left[\rho(\theta, \theta^*)\right] \leqslant \limsup_{n\to\infty} \frac{1}{n} \ln \mathbb{E}_{\theta_n}\left[\Psi_\eta(\theta_n)\right] \ .$$

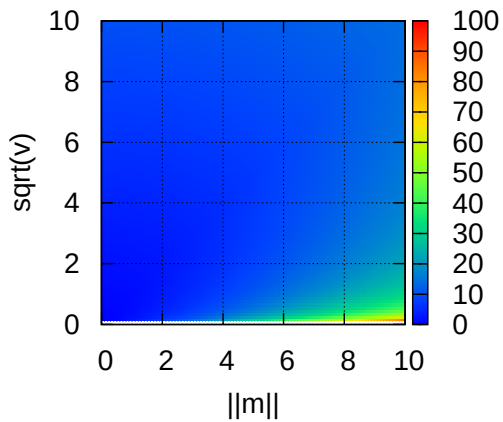
$$\eta \vee \frac{\|m\|}{\sqrt{v}}$$



$$\rho(\theta, \theta^*)$$



$$\Psi_\eta(\theta)$$



Idea 2: Stochastic Approximation

The isotropic ES as a noisy approximation of an ODE solution

Stochastic Approximation

Consider the isotropic ES

$$\theta_{n+1} = \theta_n + \delta t[F(\theta_n) + M_n]$$

as a noisy approximation of the solution to the ODE

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = F(\theta) \ .$$

In other words, θ_{N+n} is approximated by the solution $\varphi(N\delta t, \theta_n)$ of the ODE starting from θ_n.

Continuous time trajectory

Mean Field

The mean field $F(\theta)$ is expressed as

$$F(\theta) = \int w(P_\theta[y : f(y) \leqslant f(x)]) \begin{bmatrix} \alpha \tilde{\nabla}_{\theta_m} \ln p_\theta(x) \\ \tilde{\nabla}_{\theta_v} \ln p_\theta(x) \end{bmatrix} P_\theta(\mathrm{d}x) \ ,$$

where $w : [0,1] \to \mathbb{R}$ is a function defined by

$$w(p) = \sum_{i=1}^{\lambda} w_i \binom{\lambda - 1}{i - 1} p^{i-1} (1-p)^{\lambda - i} \ .$$

Global Existence and Uniqueness

The autonomous system $\frac{\mathrm{d}\theta}{\mathrm{d}t} = F(\theta)$ with initial value $\theta(0) = \theta_0$ has a unique solution on $[0, \infty)$ for each $\theta_0 \in \Theta$, i.e. there is only one solution $\varphi(\cdot, \theta_0) : t \to \theta(t)$ to the system for any $\theta_0 \in \Theta$.

Outline of the Proof

1. Bound the progress of $\varphi(N\delta t, \theta_n)$ over $\theta_n = \varphi(0, \theta_n)$.

2. Bound N-step cumulative error between θ_{n+N} and $\varphi(N\delta t, \theta_n)$.

3. Combine them and show $\mathbb{E}_n \left[\Psi_\eta(\theta_{N+n})\right] \leqslant \gamma \Psi_\eta(\theta_n)$.

Step 1: Study of Continuous-time Trajectory

Assumption on w

$$\int w(P_1[y : y \leqslant z])(z^2 - 1)P_1(\mathrm{d}z) = \alpha > 0 \ ,$$

equivalently, $F^v(\theta)/v = \alpha/d$ for $f(x) = e^{\mathrm{T}}x$, $\forall e \in \mathbb{R}^d \setminus \{0\}$.

Lemma 1: Exponential Convergence

There exist constants $\eta > 0$ and $\kappa_e > 0$ such that

$$\Psi_\eta(\varphi(t, \theta_0)) \leqslant \Psi_\eta(\theta_0)e^{-\kappa_e t}, \quad \forall t \geqslant 0, \ \forall \theta_0 \in \Theta.$$

For any $\theta_n \in \Theta$ and $t = N\delta t$, we have

$$\Psi_\eta(\varphi(N\delta t, \theta_n)) \leqslant \Psi_\eta(\theta_n)e^{-\kappa_e N\delta t} \ .$$

Proof of Lemma 1

Rewrite $\Psi_\eta(\theta) = \sqrt{\eta \vee \zeta(\theta)}\sqrt{V(\theta)}$, where

- $\zeta(\theta) = \|m\| / \sqrt{v}$
- $V(\theta) = \|m\|^2 + d \cdot v = \rho(\theta, \theta^*)^2$

To show exponential convergence, we need to show

$$\frac{\mathrm{d}\Psi_\eta(\theta)}{\mathrm{d}t} = \nabla\Psi_\eta(\theta)^{\mathrm{T}} F(\theta) \leqslant -\kappa_e \Psi_\eta(\theta) \ .$$

Proposition 1

There exists a constant $\kappa_V > 0$ such that

$$\nabla V(\theta)^{\mathrm{T}} F(\theta) \leqslant -\kappa_V \sqrt{v \cdot V(\theta)}, \quad \forall \theta \in \Theta.$$

- If $\zeta(\theta) \leqslant \eta$, then $V(\theta) \leqslant \sqrt{\eta^{-2} + d} \cdot \sqrt{v}$. Therefore, letting $\alpha = \kappa_V / \sqrt{\eta^{-2} + d}$ we have $\nabla V(\theta)^{\mathrm{T}} F(\theta) \leqslant -\alpha V(\theta)$.

Proposition 2

There exist constants $\eta > 0$ and $\beta > 0$ such that

$$\zeta(\theta) \geqslant \eta \Longrightarrow \langle \nabla \zeta(\theta),\ F(\theta) \rangle \leqslant -\beta \cdot \zeta(\theta)\ .$$

If $\zeta(\theta) < \eta$, then $\Psi_\eta(\theta) = \eta V(\theta)$ and by Proposition 1

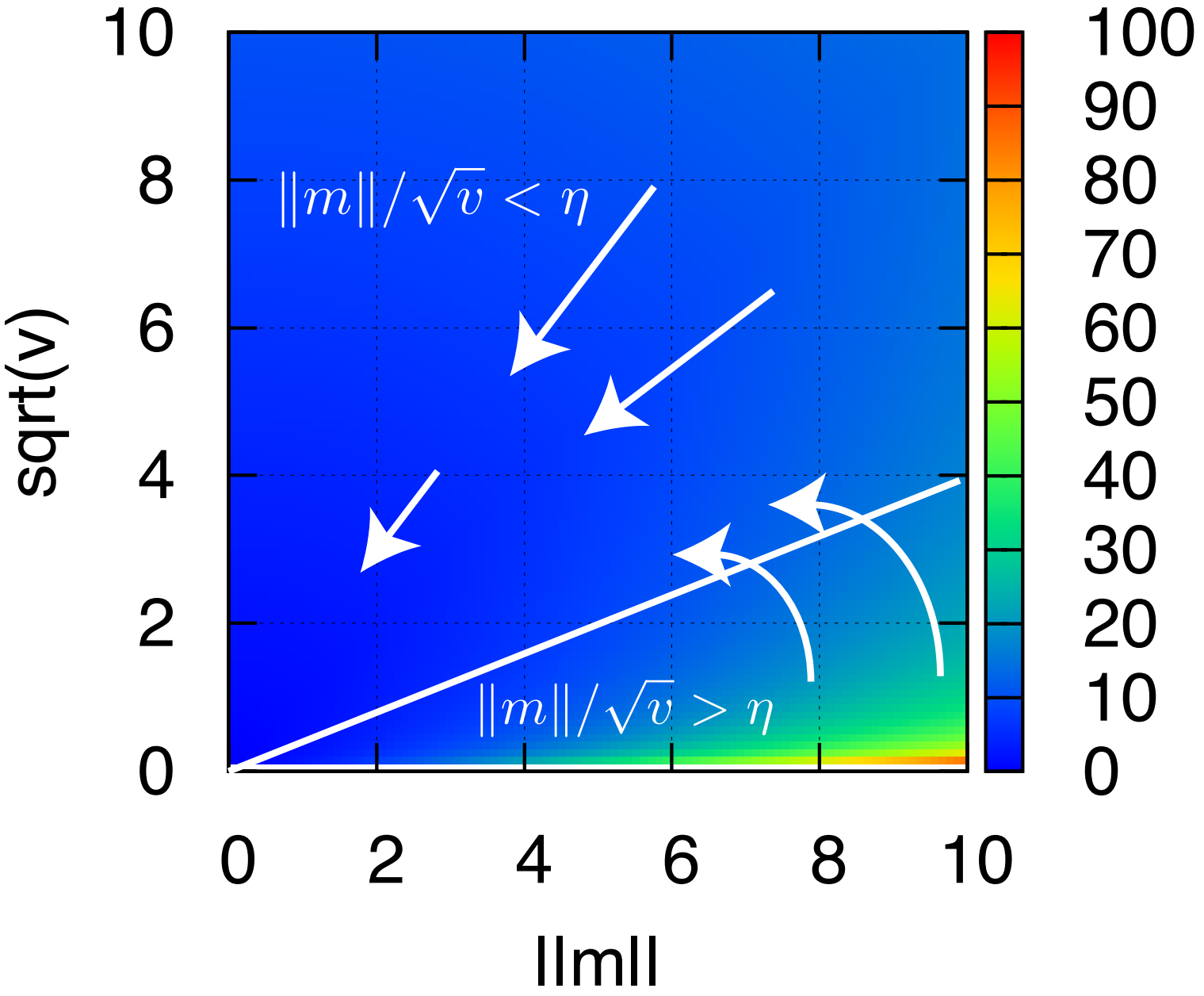
$$\begin{aligned}\nabla\Psi_\eta^2(\theta)^{\mathrm{T}} F(\theta) &= \eta \nabla V(\theta)^{\mathrm{T}} F(\theta) \\ &\leqslant -\alpha\eta V(\theta) = -\alpha\Psi_\eta^2(\theta) \ .\end{aligned}$$

If $\zeta(\theta) > \eta$, then $\Psi_\eta(\theta) = \zeta(\theta)V(\theta)$ by Proposition 2

$$\begin{aligned}\nabla\Psi_\eta^2(\theta)^{\mathrm{T}} F(\theta) &= V(\theta) \cdot \underbrace{\nabla\zeta(\theta)^{\mathrm{T}} F(\theta)}_{\leqslant -\beta\cdot\zeta(\theta)} + \zeta(\theta) \cdot \underbrace{\nabla V(\theta)^{\mathrm{T}} F(\theta)}_{<0} \\ &\leqslant -\beta \cdot \zeta(\theta) V(\theta) = -\beta\Psi_\eta^2(\theta) \ .\end{aligned}$$

Let $\kappa_e = \alpha \wedge \beta/2$, then

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\Psi_\eta(\varphi(t,\theta)) = \nabla\Psi_\eta(\varphi(t,\theta))^{\mathrm{T}} F(\varphi(t,\theta)) &\leqslant -\kappa_e \Psi_\eta(\varphi(t,\theta)) \\ \Longrightarrow \Psi_\eta(\varphi(t,\theta)) &\leqslant \Psi_\eta(\theta) e^{-\kappa_e t} \ . \quad \square\end{aligned}$$


10
8
6
4
2
0
sqrt(v)
0
2
4
6
8
10
||m||
100
90
80
70
60
50
40
30
20
10
0
$\|m\|/\sqrt{v} < \eta$
$\|m\|/\sqrt{v} > \eta$

Proof of Proposition 1

Negative Correlation

Proposition 1

There exists a constant $\kappa_V > 0$ such that

$$\nabla V(\theta)^{\mathrm{T}} F(\theta) \leqslant -\kappa_V \sqrt{v \cdot V(\theta)}, \quad \forall \theta \in \Theta.$$

$$\nabla V(\theta)^{\mathrm{T}} F(\theta) \leqslant \frac{1}{\max_i A_{i,i}} \mathbb{E}_{x \sim P_\theta} \Big[w\big(P_\theta[y : f(y) \leqslant f(x)]\big) \big(f(x) - \mathbb{E}_{x \sim P_\theta}[f(x)]\big) \Big]$$

letting P_f be the distribution of $f(x)$

$$\leqslant \frac{1}{\max_i A_{i,i}} \mathbb{E}_f \Big[\underbrace{w\big(P_f[y : y \leqslant f]\big)}_{\text{nonincreasing w.r.t. } f} \overbrace{\big(f - \mathbb{E}_{f \sim P_f}[f]\big)}^{\text{nondecreasing w.r.t. } f} \Big]$$

Proof of Proposition 1

Sub propositions

Proposition 3 (negative correlation)

Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be non-decreasing and non-increasing functions, respectively. Let μ be a measure on $\mathbb{R}$ and suppose that $\int |f(x)| \, \mu(\mathrm{d}x) < \infty$ and $\int |g(x)| \, \mu(\mathrm{d}x) < \infty$. Let $M_f = \int f(x)\mu(\mathrm{d}x)$ and $M_g = \int g(x)\mu(\mathrm{d}x)$. Then,

$$\int f(x)g(x)\mu(\mathrm{d}x) \leqslant M_f M_g - \frac{1}{4} \iint |f(x) - f(y)| \, \mu(\mathrm{d}x)\mu(\mathrm{d}y) \iint |g(x) - g(y)| \, \mu(\mathrm{d}x)\mu(\mathrm{d}y) \ .$$

Proposition 4 (fourth moment method)

Let Y and $\tilde{Y}$ be i.i.d. random variable on $\mathbb{R}$ such that $\mathbb{E}[Y^2] > 0$ and $\mathbb{E}[Y^4] < \infty$. Then, the following inequality holds

$$\mathbb{E}[|Y - \tilde{Y}|] \geqslant \frac{\mathbb{E}[(Y - \mathbb{E}[Y])^2]^{3/2}}{\mathbb{E}[(Y - \mathbb{E}[Y])^4]^{1/2}} \ .$$

Proof of Proposition 1

$$\nabla V(\theta)^{\mathrm{T}} F(\theta) \leqslant \frac{1}{\max_i A_{i,i}} \mathbb{E}_f \Big[w(\underbrace{P_f[y : y \leqslant f]}_{=:U \sim U(0,1)}) \underbrace{(f - \mathbb{E}_{f \sim P_f}[f])}_{=:F} \Big]$$

by the propositions

$$\leqslant -\frac{1}{4} \frac{1}{\max_i A_{i,i}} \frac{\mathbb{E}[(w(U) - \mathbb{E}[w(U)])^2]^{3/2}}{\mathbb{E}[(w(U) - \mathbb{E}[w(U)])^4]^{1/2}} \frac{\mathbb{E}[(F - \mathbb{E}[F])^2]^{3/2}}{\mathbb{E}[(F - \mathbb{E}[F])^4]^{1/2}}$$

since F is generalized χ^2 distributed, we can evaluate the expectations

$$= -\frac{1}{2\sqrt{30}\,\mathrm{Cond}(A)} \frac{M_{w,2}^{3/2}}{M_{w,4}^{1/2}} \sqrt{v} \sqrt{2\|m\|^2 + dv} \leqslant -\kappa_V \sqrt{vV(\theta)}$$

where $M_{w,1} := \int_0^1 w(u)\mathrm{d}u$, $M_{w,2} := \int_0^1 (w(u) - M_{w,1})^2 \mathrm{d}u$ and $M_{w,4} := \int_0^1 (w(u) - M_{w,1})^4 \mathrm{d}u$. □

Proof of Proposition 2

Proposition 2

There exist constants $\eta > 0$ and $\beta > 0$ such that

$$\zeta(\theta) \geqslant \eta \Longrightarrow \nabla\zeta(\theta)^{\mathrm{T}} F(\theta) \leqslant -\beta \cdot \zeta(\theta) \ .$$

Proof idea.

- $\frac{\nabla\zeta(\theta)^{\mathrm{T}} F(\theta)}{\zeta(\theta)} = \frac{1}{\zeta(\theta)} \underbrace{\frac{m^{\mathrm{T}}}{\|m\|} \frac{F^m(\theta)}{v^{1/2}}}_{O(1)} - \frac{F^v(\theta)}{v}$
- $\frac{F^v(\theta)}{v} \to \frac{1}{d} \int w(P_1[y : y \leqslant z])(z^2 - 1) P_1(\mathrm{d}z)$ as $\zeta(\theta) \to \infty$.
- By the assumption on w, $\int w(P_1[y : y \leqslant z])(z^2 - 1) P_1(\mathrm{d}z) = \alpha > 0$.
- $\frac{\nabla\zeta(\theta)^{\mathrm{T}} F(\theta)}{\zeta(\theta)} \to -\frac{\alpha}{d}$ as $\zeta(\theta) \to \infty$. □

Step 2: N-step cumulative error

1. Bound the progress of $\varphi(N\delta t, \theta_n)$ over $\theta_n = \varphi(0, \theta_n)$.

$$\Psi_\eta(\varphi(N\delta t, \theta_n)) \leqslant \Psi_\eta(\theta_n) e^{-\kappa_e N \delta t} \quad .$$

2. Bound N-step cumulative error between θ_{n+N} and $\varphi(N\delta t, \theta_n)$.

3. Combine them and show $\mathbb{E}_n \left[\Psi_\eta(\theta_{N+n})\right] \leqslant \gamma \Psi_\eta(\theta_n)$.

N-step cumulative error

Lemma 2: N-step cumulative error

For any $T \geqslant 0$ and $n \in \mathbb{N}$, there is a constant $\kappa_T > 0$ independent of n such that

$$\mathbb{E}_n \left[\sup_{k \in [\![0, T/\delta t]\!]} \rho(\theta_{n+k}, \varphi(k \cdot \delta t, \theta_n))^2 \right]^{1/2} \leqslant \kappa_T \sqrt{v_n} \sqrt{\delta t} \ .$$

Let $N = T/\delta t$. Then,

$$\mathbb{E}_n \left[\rho(\theta_{n+N}, \varphi(N\delta t, \theta_n))^2 \right]^{1/2} \leqslant \kappa_T \sqrt{v_n} \sqrt{\delta t} \ .$$

Proof Idea of Lemma 2

$$
\begin{aligned}
&\sup_{k\in[\![0,\lfloor T/\delta t\rfloor]\!]} \rho(\theta_{n+k}, \varphi(k\delta t, \theta_n)) \\
&\leqslant \sup_{k\in[\![0,\lfloor T/\delta t\rfloor]\!]} \left\| \sum_{i=0}^{k-1} \delta t \left(F^m(\theta_{n+i}) - F^m(\varphi(i\delta t, \theta_n))\right)\right\| \\
&\quad + \frac{\sqrt{d}}{\sqrt{v_{n+l} \vee v^n(l\delta t)}} \sup_{k\in[\![0,\lfloor T/\delta t\rfloor]\!]} \left| \sum_{i=0}^{k-1} \delta t \left(F^v(\theta_{n+i}) - F^v(\varphi(i\delta t, \theta_n))\right)\right| \\
&\quad + \sup_{k\in[\![0,\lfloor T/\delta t\rfloor]\!]} \left\| \sum_{i=0}^{k-1} \delta t M^m_{n+i}\right\| + \frac{\sqrt{d}}{\sqrt{v_{n+l} \vee v^n(l\delta t)}} \sup_{k\in[\![0,\lfloor T/\delta t\rfloor]\!]} \left| \sum_{i=0}^{k-1} \delta t M^v_{n+i}\right| \\
&\qquad\qquad \text{(martingale error)} \\
&\quad + \sup_{k\in[\![0,\lfloor T/\delta t\rfloor]\!]} \left\| \sum_{i=0}^{k-1} \int_{i\delta t}^{(i+1)\delta t} (F^m(\varphi(\tau, \theta_n)) - F^m(\varphi(i\delta t, \theta_n)))\mathrm{d}\tau\right\| \\
&\quad + \frac{\sqrt{d}}{\sqrt{v_{n+l} \vee v^n(l\delta t)}} \sup_{k\in[\![0,\lfloor T/\delta t\rfloor]\!]} \left| \sum_{i=0}^{k-1} \int_{i\delta t}^{(i+1)\delta t} (F^v(\varphi(\tau, \theta_n)) - F^v(\varphi(i\delta t, \theta_n)))\mathrm{d}\tau\right| \quad .
\end{aligned}
$$

(discretization error)

Proof Idea of Lemma 2

Proposition: bound for first two terms

For some constant $K_1 > 0$,

$$\sup_{k \in [\![0, \lfloor T/\delta t \rfloor]\!]} \left\| \sum_{i=0}^{k-1} \delta t \left(F^m(\theta_{n+i}) - F^m(\varphi(i\delta t, \theta_n)) \right) \right\|$$
$$+ \frac{\sqrt{d}}{\sqrt{v_{n+l} \vee v^n(l\delta t)}} \sup_{k \in [\![0, \lfloor T/\delta t \rfloor]\!]} \left| \sum_{i=0}^{k-1} \delta t \left(F^v(\theta_{n+i}) - F^v(\varphi(i\delta t, \theta_n)) \right) \right|$$
$$\leqslant K_1 \delta t \sum_{i=0}^{\lfloor T/\delta t \rfloor - 1} \rho(\theta_{n+i}, \varphi(i\delta t, \theta_n)) \ .$$

Proof Idea of Lemma 2

Proposition: bound for martingale error

For some constant $K_2 > 0$,

$$\mathbb{E}_n\left[\left(\sup_{k\in[\![0,\lfloor T/\delta t\rfloor]\!]}\left\|\sum_{i=0}^{k-1}\delta t M^m_{n+i}\right\|\right)^2\right]^{1/2}$$
$$+\,\mathbb{E}_n\left[\left(\frac{\sqrt{d}}{\sqrt{v_{n+l}\vee v^n(l\delta t)}}\sup_{k\in[\![0,\lfloor T/\delta t\rfloor]\!]}\left|\sum_{i=0}^{k-1}\delta t M^v_{n+i}\right|\right)^2\right]^{1/2}$$
$$\leqslant K_2\sqrt{\delta t}\sqrt{v_n}$$

Here we used Doob's L^p maximal inequality.

Proof Idea of Lemma 2

Proposition: bound for discretization error

For some constant $K_3 > 0$,

$$\sup_{k\in[\![0,\lfloor T/\delta t\rfloor]\!]}\left\|\sum_{i=0}^{k-1}\int_{i\delta t}^{(i+1)\delta t}(F^m(\varphi(\tau,\theta_n))-F^m(\varphi(i\delta t,\theta_n)))\mathrm{d}\tau\right\|$$

$$+\frac{\sqrt{d}}{\sqrt{v_{n+l}\vee v^n(l\delta t)}}\sup_{k\in[\![0,\lfloor T/\delta t\rfloor]\!]}\left|\sum_{i=0}^{k-1}\int_{i\delta t}^{(i+1)\delta t}(F^v(\varphi(\tau,\theta_n))-F^v(\varphi(i\delta t,\theta_n)))\mathrm{d}\tau\right|$$

$$\leqslant K_3\delta t\sqrt{v_n}$$

Proof Idea of Lemma 2

By the propositions,

$$\mathbb{E}_n\left[\sup_{k\in[\![0,\lfloor T/\delta t\rfloor]\!]}\rho(\theta_{n+k},\varphi(k\delta t,\theta_n))^2\right]^{1/2}$$

$$\leqslant K_1\delta t\sum_{i=0}^{\lfloor T/\delta t\rfloor-1}\mathbb{E}_n\left[\rho(\theta_{n+i},\varphi(i\delta t,\theta_n))^2\right]^{1/2}+(K_2+K_2\sqrt{\delta t})\sqrt{\delta t}\sqrt{v_n}$$

$$\leqslant K_1\delta t\sum_{i=0}^{\lfloor T/\delta t\rfloor-1}\mathbb{E}_n\left[\sup_{k\in[\![0,i]\!]}\rho(\theta_{n+k},\varphi(k\delta t,\theta_n))^2\right]^{1/2}+(K_2+K_3\sqrt{\delta t})\sqrt{\delta t}\sqrt{v_n}$$

Apply discrete Gronwall inequality:

$$x_0=0,\ x_{n+1}\leqslant C+\sum_{i=0}^{n}a_ix_i\Longrightarrow x_{n+1}\leqslant C\exp\left(\sum_{i=0}^{n}a_i\right)$$

we have

$$\mathbb{E}_n\left[\sup_{k\in[\![0,\lfloor T/\delta t\rfloor]\!]}\rho(\theta_{n+k},\varphi(k\delta t,\theta_n))^2\right]^{1/2}\leqslant(K_2+K_2\sqrt{\delta t})\sqrt{\delta t}\sqrt{v_n}\exp(K_1T)$$

$$\leqslant\kappa_T\sqrt{\delta t}\sqrt{v_n}\quad\square$$

Step 3: Completion

1. Bound the progress of $\varphi(N\delta t, \theta_n)$ over $\theta_n = \varphi(0, \theta_n)$.

$$\Psi_\eta(\varphi(N\delta t, \theta_n)) \leqslant \Psi_\eta(\theta_n) e^{-\kappa_e N \delta t} \quad .$$

2. Bound N-step cumulative error between θ_{n+N} and $\varphi(N\delta t, \theta_n)$.

$$\mathbb{E}_n \left[\rho(\theta_{n+N}, \varphi(N\delta t, \theta_n))^2 \right]^{1/2} \leqslant \kappa_T \sqrt{v_n} \sqrt{\delta t} \quad .$$

3. Combine them and show $\mathbb{E}_n \left[\Psi_\eta(\theta_{N+n}) \right] \leqslant \gamma \Psi_\eta(\theta_n)$.

Proof idea of Lemma 3

For simplicity, assume $\|m_n\| / \sqrt{v_n} \leqslant \eta$ for all n. Then, $\Psi_\eta(\theta) = \eta^{1/2}\rho(\theta, \theta^*)$. We have

$$\begin{aligned}
&\mathbb{E}_n[\Psi_\eta(\theta_{n+N})] \\
&\quad = \Psi_\eta(\varphi(N\cdot\delta t, \theta_n)) + \eta^{1/2}\cdot\mathbb{E}_n[\underbrace{(\rho(\theta_{n+N}, 0) - \rho(\varphi(N\cdot\delta t, \theta_n), 0))}_{\text{triangle inequality}}] \\
&\quad \leqslant \underbrace{\Psi_\eta(\varphi(N\cdot\delta t, \theta_n))}_{\text{Lemma 1}} + \eta^{1/2}\cdot\underbrace{\mathbb{E}_n[\rho(\theta_{n+N}, \varphi(N\cdot\delta t, \theta_n))]}_{\text{Lemma 2}} \\
&\quad \leqslant \Psi_\eta(\theta_n)e^{-\kappa_e N\delta t} + \kappa_T\sqrt{\delta t}\,\underbrace{\eta^{1/2}\sqrt{v_n}}_{\leqslant \Psi_\eta(\theta_n)/d^{1/2} \text{ by definition}} \\
&\quad = (e^{-\kappa_e N\delta t} + \kappa_T\sqrt{\delta t}/d^{1/2})\Psi_\eta(\theta_n)
\end{aligned}$$

taking sufficiently large $T = N\delta t$ and sufficiently small δt

$$\leqslant \gamma\Psi_\eta(\theta_n)\ .$$

$$\mathbb{E}_n[\Psi_\eta(\theta_{n+N})] \leqslant \gamma \Psi_\eta(\theta_n)$$

$$\Longrightarrow \mathbb{E}[\Psi_\eta(\theta_{N\cdot k})] \leqslant \gamma^k \Psi_\eta(\theta_0)$$

$$\Longrightarrow \frac{1}{k} \ln \mathbb{E}[\Psi_\eta(\theta_{N\cdot k})] \leqslant \ln \gamma + \frac{1}{k} \ln \Psi_\eta(\theta_0) \xrightarrow{k\to\infty} \ln \gamma < 0$$

$$\Longrightarrow \limsup_{k\to\infty} \frac{1}{k} \ln \mathbb{E}[\Psi_\eta(\theta_{N\cdot k})] < 0$$

$$\Longrightarrow \limsup_{k\to\infty} \frac{1}{k} \ln \mathbb{E}\left[\left(\|m_{N\cdot k} - \mathbf{x}^*\|^2 + d \cdot v_{N\cdot k}\right)^{1/2}\right] < 0$$

□

Global Linear Convergence in Expectation

on Monotonic Convex-Quadratic-Composite

Global Linear Convergence on Monotonic Convex-Quadratic-Composite

For sufficiently small $\delta t > 0$, there exists $\gamma < 0$ such that

$$\limsup_{n \to \infty} \frac{1}{n} \ln \mathbb{E}\left[\rho(\theta, \theta^*)\right] = \gamma \ .$$

Extension to Monotonic $\mathcal{C}^2$-Composite

1. Bound the progress of $\varphi(N\delta t, \theta_n)$ over $\theta_n = \varphi(0, \theta_n)$.

   Based on Taylor's approximation,

   $$\Psi_\eta(\varphi(N\delta t, \theta_n)) \leqslant \Psi_\eta(\theta_n) e^{-\kappa_e N\delta t}, \quad \forall \theta_n \in U \subset \Theta \ .$$

2. Bound N-step cumulative error between θ_{n+N} and $\varphi(N\delta t, \theta_n)$.

   (unchanged)

   $$\mathbb{E}_n \left[\rho(\theta_{n+N}, \varphi(N\delta t, \theta_n))^2\right]^{1/2} \leqslant \kappa_T \sqrt{v_n}\sqrt{\delta t} \ .$$

3. Combine them and we have

   $$\Pr\left[\limsup_{n\to\infty} \frac{1}{n} \ln \Psi_\eta(\theta_n) < 0\right] \geqslant 1 - \varepsilon \ .$$

   With probability at most ε, θ_n will leave the region of attraction U.

Local Linear Convergence with probability $1-\varepsilon$

on Monotonic $\mathcal{C}^2$-Composite

Local Linear Convergence on Monotonic $\mathcal{C}^2$-Composite

For $\theta_0 \in U$ and any $0 < \varepsilon < 1$, there exists $\overline{\delta t} > 0$ such that for $0 < \forall \delta t \leqslant \overline{\delta t}$

$$\Pr\left[\limsup_{n\to\infty} \frac{1}{n} \ln \Psi_\eta(\theta_n, \theta^*) \leqslant \gamma\right] \geqslant 1-\varepsilon \ ,$$

where the speed of convergence $\gamma < 0$ depends on δt.

Pros and Cons

Pros

- general weight scheme (cf $(1, \lambda)$ or $(\mu/\mu_I, \lambda)$ strategies)
- wide class of functions
- able to apply to general algorithms

Cons

- sufficiently small δt
  - the rate of convergence is loose
- rates of convergence of $\|m\|$ and $\sqrt{v}$
  - $\|m\|$ and $\sqrt{v}$ converge at the same speed in practice and this is shown by [Auger 05] for $(1, \lambda)$-σSA-ES