

Toward a Convergence Proof for CMA-ES—and Beyond

Anne Auger & Nikolaus Hansen

TAO Team, INRIA Saclay, France
<http://tao.lri.fr>
{nikolaus.hansen@inria.fr}

Dagstuhl 2008

Continuous minimization of a class of functions

Task: approach the global minimum of

$$\begin{aligned} f : \mathbb{R}^d &\rightarrow \mathbb{R} \\ \mathbf{x} &\mapsto g((\mathbf{x} - \mathbf{x}^*)^\top \mathbf{H}(\mathbf{x} - \mathbf{x}^*)) \end{aligned}$$

$g : \mathbb{R} \rightarrow \mathbb{R}$ is strictly monotonously increasing

\mathbf{H} is a symmetric positive definite Hessian matrix

Lower bound and convergence rate for scale-adaptive ES

$$\inf_{n \in \mathbb{N}} \frac{1}{n} E \ln \frac{\|X_n - \mathbf{x}^*\|}{\|X_0 - \mathbf{x}^*\|} \geq -\rho \quad \text{for all } f$$

set w.l.o.g. $\mathbf{x}^* = \vec{0}$

... some previous works based on continuous Markov chains

- Rudolph, CEC 1994, 1997, ES
- Bienvenüe & François, TCS 2003, self-adaptive $(1, \lambda)$ -ES
- Auger, TCS 2005, self-adaptive $(1, \lambda)$ -ES
- Andrieu & Moulines, AAP 2005, MCMC with covariance matrix update

Content

- 1 CMA-ES
- 2 Stability
- 3 The Definitions
- 4 The Markov Chain
- 5 The Results
- 6 Summary

*Einstein once spoke of the “unreasonable effectiveness of mathematics” in describing how the natural world works. Whether one is talking about basic physics, about the increasingly important environmental sciences, or the transmission of disease, **mathematics is never any more, or any less, than a way of thinking clearly.** As such, it always has been and always will be a valuable tool, but only valuable when it is part of a larger arsenal embracing analytic experiments and, above all, wide-ranging imagination.*

Lord Kay

CAVEAT: these results are very ongoing

Covariance Matrix Adaptation ES in a Nutshell

- 1 Multivariate normal distribution to generate new search points

follows the maximum entropy principle

Covariance Matrix Adaptation ES in a Nutshell

- 1 Multivariate normal distribution to generate new search points
follows the maximum entropy principle
- 2 Selection only based on the ranking of the f -values, weighted recombination
using only the ranking of f -values in CMA preserves invariance

Covariance Matrix Adaptation ES in a Nutshell

- 1 Multivariate normal distribution to generate new search points
follows the maximum entropy principle
- 2 Selection only based on the ranking of the f -values, weighted recombination
using only the ranking of f -values in CMA preserves invariance
- 3 *Covariance matrix adaptation (CMA)* increases the probability to **repeat successful steps**
learning all pairwise dependencies
⇒ conducts a sequential PCA
⇒ rotated problem representation

Covariance Matrix Adaptation ES in a Nutshell

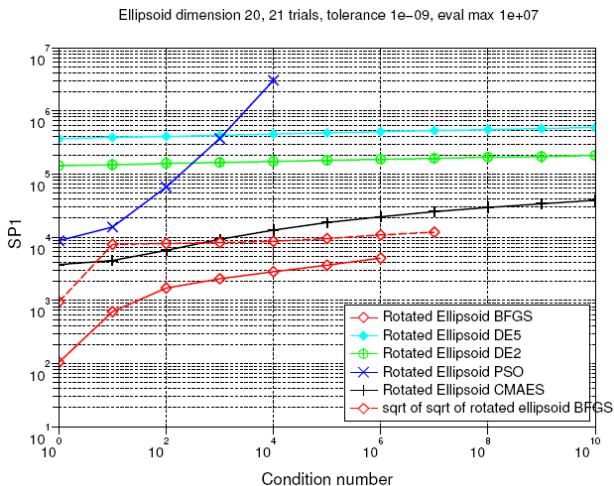
- 1 Multivariate normal distribution to generate new search points
follows the maximum entropy principle
- 2 Selection only based on the ranking of the f -values, weighted recombination
using only the ranking of f -values in CMA preserves invariance
- 3 *Covariance matrix adaptation (CMA)* increases the probability to **repeat successful steps**
learning all pairwise dependencies
⇒ conducts a sequential PCA
⇒ rotated problem representation
- 4 *Path length control* **controls the step length**
uses the evolution path,
aims at conjugate perpendicularity,
non-local criterion

Covariance Matrix Adaptation Evolution Strategy

*Everything should be made as simple as possible,
but not simpler.
— Albert Einstein*

- very carefully designed algorithm
 - simplicity, invariance, parameter identification
- philosophy is based on invariance
 - invariance generalizes performance assertions
 - adaptation introduces invariance under general linear transformations
- performs well not only, but in particular, on non-separable, ill-conditioned (test-)functions and “real world problems”
 - problems with essential dependencies between the parameters

A Performance Result



SP1=average number of function evaluations to reach the target function value of 10^{-9}

Covariance Matrix Adaptation Evolution Strategy

Initialize $\mathbf{m} \in \mathbb{R}^d$,

$$\mathbf{C} = \mathbf{I}$$

set

$$c_{\text{cov}} \approx 1/d^2,$$

λ

While not terminate

$$\mathbf{x}_i = \mathbf{m} + \mathbf{z}_i, \quad \mathbf{z}_i \sim \mathcal{N}_i(\mathbf{0}, \mathbf{C}), \quad i = 1, \dots, \lambda$$

sampling

$$\mathbf{m} \leftarrow \mathbf{m} + \mathbf{z}_{\text{sel}}, \quad \text{where } \mathbf{z}_{\text{sel}} = \arg \min_{i=1}^{\mu} f(\mathbf{x}_i) - \mathbf{m}$$

update mean

$$\mathbf{C} \leftarrow (1 - c_{\text{cov}}) \mathbf{C} + c_{\text{cov}} \times \mathbf{z}_{\text{sel}} \mathbf{z}_{\text{sel}}^T$$

update \mathbf{C}

SIMPLIFIED $(1, \lambda)$ -CMA-ES

Covariance Matrix Adaptation Evolution Strategy

Initialize $\mathbf{m} \in \mathbb{R}^d$, $\sigma \in \mathbb{R}_+$, $\mathbf{C} = \mathbf{I}$, and $\mathbf{p}_c = \mathbf{0}$, $\mathbf{p}_\sigma = \mathbf{0}$

set $c_c \approx 4/d$, $c_\sigma \approx 4/d$, $c_{\text{cov}} \approx \mu_{\text{eff}}/d^2$, $\mu_{\text{cov}} = \mu_{\text{eff}}$, $d_\sigma \approx 1 + \sqrt{\frac{\mu_{\text{eff}}}{d}}$,
 λ , and $w_i, i = 1, \dots, \mu$ such that $\mu_{\text{eff}} \approx 0.3 \lambda$, where $\mu_{\text{eff}} = 1 / \sum_{i=1}^{\mu} w_i^2$

While not terminate

$\mathbf{x}_i = \mathbf{m} + \sigma \mathbf{z}_i$, $\mathbf{z}_i \sim \mathcal{N}_i(\mathbf{0}, \mathbf{C})$, $i = 1, \dots, \lambda$ sampling

$\mathbf{m} \leftarrow \mathbf{m} + \sigma \langle \mathbf{z} \rangle_{\text{sel}}$, where $\langle \mathbf{z} \rangle_{\text{sel}} = \sum_{i=1}^{\mu} w_i \mathbf{z}_{i:\lambda}$ update mean

$\mathbf{p}_c \leftarrow (1 - c_c) \mathbf{p}_c + \mathbb{1}_{\{\frac{\|\mathbf{p}_\sigma\|}{\sqrt{d}} < 1.5\}} \sqrt{1 - (1 - c_c)^2} \sqrt{\mu_{\text{eff}}} \langle \mathbf{z} \rangle_{\text{sel}}$ cumulation for \mathbf{C}

$\mathbf{C} \leftarrow (1 - c_{\text{cov}}) \mathbf{C} + c_{\text{cov}} \frac{1}{\mu_{\text{cov}}} \mathbf{p}_c \mathbf{p}_c^T$ update \mathbf{C}
 $+ c_{\text{cov}} \left(1 - \frac{1}{\mu_{\text{cov}}}\right) \sum_{i=1}^{\mu} w_i \mathbf{z}_{i:\lambda} \mathbf{z}_{i:\lambda}^T$

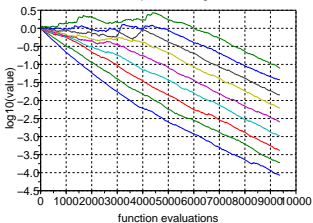
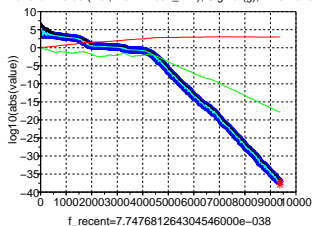
$\mathbf{p}_\sigma \leftarrow (1 - c_\sigma) \mathbf{p}_\sigma + \sqrt{1 - (1 - c_\sigma)^2} \sqrt{\mu_{\text{eff}}} \mathbf{C}^{-\frac{1}{2}} \langle \mathbf{z} \rangle_{\text{sel}}$ cumulation for σ

$\sigma \leftarrow \sigma \times \exp\left(\frac{c_\sigma}{d_\sigma} \left(\frac{\|\mathbf{p}_\sigma\|}{\mathbb{E}\|\mathcal{N}(\mathbf{0}, \mathbf{I})\|} - 1\right)\right)$ update of σ

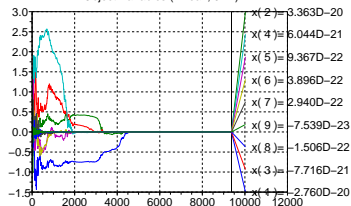
$(\mu/\mu_W, \lambda)$ -CMA-ES

Stability/Stationarity of (m_n, C_n) ?

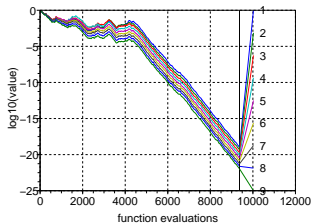
Function Value (fval, fval minus f_min), Sigma (g), Axis Ratio (r)



Object Variables (xmean, 9-D)



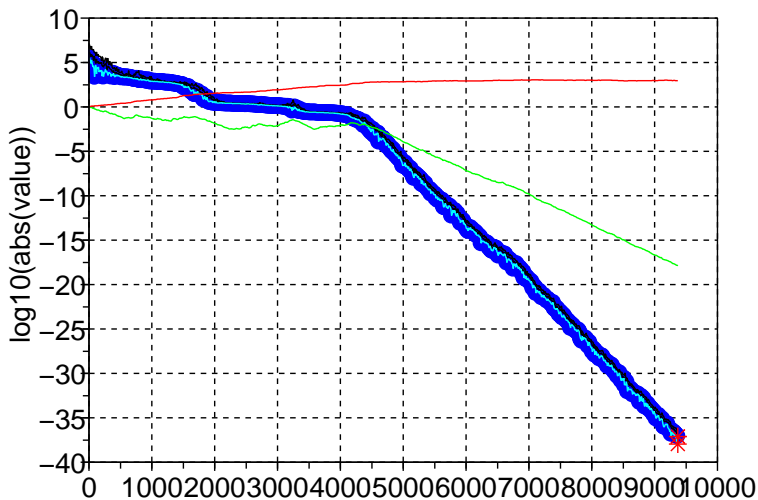
Standard Deviations



$(\mu/\mu_W, \lambda)$ -CMA-ES on $f(x) = x^T H x$

Stability/Stationarity of (m_n, C_n) ?

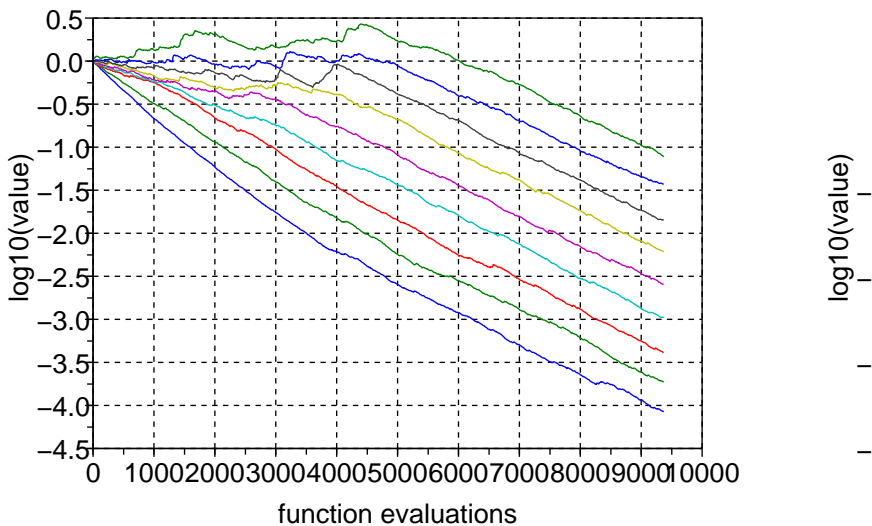
Function Value (fval, fval minus f_min), Sigma (g), Axis Ratio (r)



f_recent=7.747681264304546000e-038

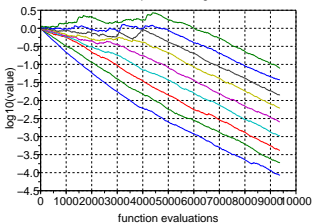
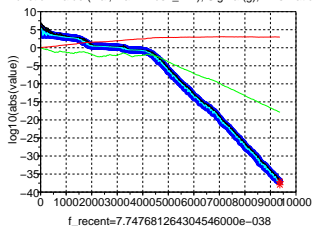
Stability/Stationarity of (m_n, C_n) ?

Principle Axis Lengths

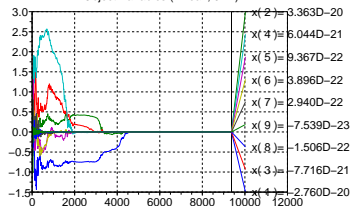


Stability/Stationarity of (m_n, C_n) ?

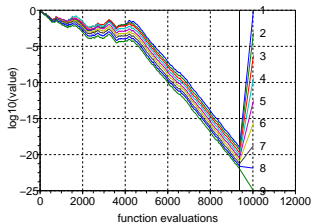
Function Value (fval, fval minus f_min), Sigma (g), Axis Ratio (r)



Object Variables (xmean, 9-D)



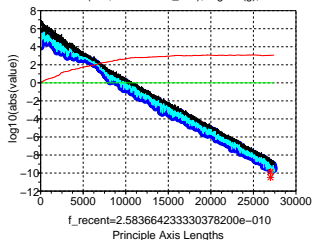
Standard Deviations



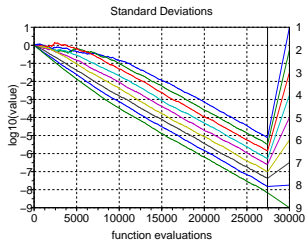
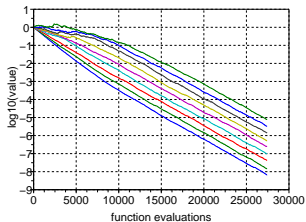
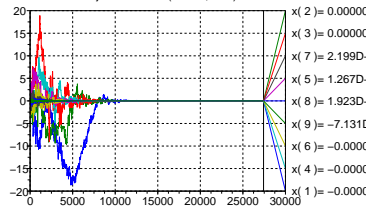
$(\mu/\mu_W, \lambda)$ -CMA-ES on $f(x) = x^T H x$

Stability/Stationarity of (m_n, C_n) ?

Function Value (fval, fval minus f_min), Sigma (g), Axis Ratio (r)



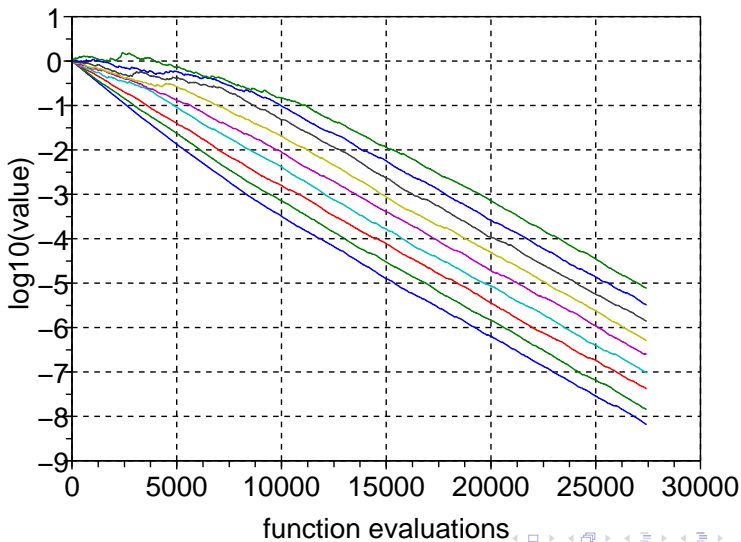
Object Variables (xmean, 9-D)



simplified $(1, \lambda)$ -CMA-ES on $f(x) = x^T H x$

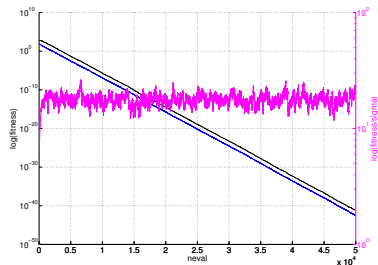
Stability/Stationarity of (m_n, C_n) ?

Principle Axis Lengths



Stability/Stationarity/Ergodicity

the “heuristic” interpretation: the “chain” returns to the “center” of the space in a recurring way



Definitions

Let (Ω, \mathcal{B}, P) be a probability space and let $Z_n \in \Omega$ be a Markov chain. Ω is the state space of the algorithm and later we will choose $\Omega = \mathbb{R}^d \times \dots$

Hitting time

For $A \in \mathcal{B}$,

$$\tau_A = \min\{n > 0 : Z_n \in A\}$$

is the first time step where $Z_n \in A$

φ -irreducibility (weak notion of stability/stationarity)

Z_n is φ -irreducible if there exists a measure φ such that for all $A \in \mathcal{B}$ with $\varphi(A) > 0$ and for all starting points $Z_0 = z \in \Omega$

$$P_z(\tau_A < \infty) > 0$$

Definitions (cont.)

Transition kernel

The transition kernel $P(., .)$ from any state $z \in \Omega$ into any $A \in \mathcal{B}$ is defined as

$$P(z, A) = P(Z_{n+1} \in A | Z_n = z)$$

Invariant measure

A measure μ is invariant under the transition kernel $P(., .)$ if

$$\mu(A) = \int_{\Omega} P(z, A) \mu(dz), \quad \forall A \in \mathcal{B}$$

If μ is a *probability* measure, Z_n is called positive (recurrent)

Definitions (final)

Ergodicity (strong notion of stability/stationarity)

We call Z_n **ergodic**, if it is φ -irreducible and positive, i.e. if there exists an unique *invariant probability* measure for Z_n

Remark A Markov chain **will converge** to its unique invariant probability measure, possibly geometrically fast

Remark Technically a *drift condition* is used to prove ergodicity

Not covered how long does it take to converge to the invariant measure? How does the invariant measure look like?

Simulations give strong hints

The Markov Chain

We consider the simplified $(1, \lambda)$ -CMA-ES on the function f .

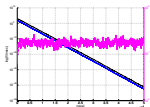
The chain $(\mathbf{m}_n, \mathbf{C}_n)_{n \in \mathbb{N}}$ obeys

$$\mathbf{m}_{n+1} = \arg \min_{i=1}^{\lambda} f(\mathbf{m}_n + \mathcal{N}_i(\mathbf{0}, \mathbf{C}))$$

$$\mathbf{C}_{n+1} = (1 - c_{\text{cov}}) \mathbf{C}_n + c_{\text{cov}} (\mathbf{m}_{n+1} - \mathbf{m}_n)(\mathbf{m}_{n+1} - \mathbf{m}_n)^{\text{T}}$$

update mean
update \mathbf{C}

We investigate the “normalized” chain



$$\mathbf{Z}_n = \frac{\mathbf{m}_n - \arg \min f(\mathbf{x})}{\sqrt{\lambda_{\max}(\mathbf{C}_n)}} \quad (1)$$

$$\mathbf{K}_n = \frac{\mathbf{C}_n}{\lambda_{\max}(\mathbf{C}_n)} \quad (2)$$

$(\mathbf{Z}_n, \mathbf{K}_n)$ is a Markov chain

Theorems

Lemma

The chain $(\mathbf{Z}_n, \mathbf{K}_n)$ is φ -irreducible

Theorem

The chain $(\mathbf{Z}_n, \mathbf{K}_n)$ admits a unique invariant probability measure

Proof: using drift conditions

Theorem

The expected covariance matrix is proportional to the inverse Hessian matrix

$$E_{\mu}(\mathbf{K}_n) \propto \mathbf{H}^{-1}$$

Proof: using invariance properties

Implications

Linear convergence (slope of the graph)

Corollary (linear convergence)

The mean, \mathbf{m}_n , of the simplified $(1, \lambda)$ -CMA-ES converges, or diverges, log-linearly to the optimum of f , that is

$$\frac{1}{n} \ln \frac{\|\mathbf{m}_n - \mathbf{x}^*\|}{\|\mathbf{m}_0 - \mathbf{x}^*\|} \rightarrow c(d) \quad \text{for } n \rightarrow \infty$$

where $c(d)$ is a dimension dependent convergence rate

Proof: $\ln \frac{\|\mathbf{m}_n - \mathbf{x}^*\|}{\|\mathbf{m}_0 - \mathbf{x}^*\|} = \sum_{k=0}^{n-1} \ln \frac{\|\mathbf{m}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{m}_k - \mathbf{x}^*\|} = \sum_{k=0}^{n-1} \ln \frac{\|\mathbf{Z}_{k+1}\|}{\|\mathbf{Z}_k\|} \frac{\lambda_{\max}(\mathbf{C}_k)}{\lambda_{\max}(\mathbf{C}_{k+1})}$ and we can show that

$\frac{\|\mathbf{Z}_{k+1}\|}{\|\mathbf{Z}_k\|} \frac{\lambda_{\max}(\mathbf{C}_k)}{\lambda_{\max}(\mathbf{C}_{k+1})} =: \mathcal{G}(\mathbf{Z}_k, \mathbf{K}_k)$ can be written as a function of $(\mathbf{Z}_k, \mathbf{K}_k)$. Because $(\mathbf{Z}_n, \mathbf{K}_n)$ admits a stationary probability measure, μ , it satisfies the strong law of large numbers

$$\frac{1}{n} \sum_{k=1}^n \mathcal{G}(\mathbf{Z}_k, \mathbf{K}_k) \rightarrow E_{\mu}(\mathcal{G}(\mathbf{Z}, \mathbf{K})) \quad \text{for } n \rightarrow \infty$$

Progress Rate

Remark (not rigorously proved)

The log-progress

$$-E_{\mu} \ln \frac{\|m_{n+1} - \mathbf{x}^*\|}{\|m_n - \mathbf{x}^*\|}$$

equals to the negative convergence rate $-c(d)$

Remark

For $d \rightarrow \infty$ the log-progress aligns with the classical progress definition in evolution strategies

Arnold 2002

Auger & Hansen, GECCO 2006

Summary

- we can prove the existence of a unique stationary regime for a simplified $(1, \lambda)$ -CMA-ES on a class of unimodal functions, $E_{\mu}(\mathbf{C}_n) \propto \mathbf{H}^{-1}$
for a “normalized” chain on monotonous transformations of positive quadratic forms
- stationarity implies log-linear convergence, or divergence
- we exploited the theory of φ -irreducible Markov chains
Meyn & Tweedie 1993
- we exploited invariance properties of CMA-ES

... and beyond

The mathematical tools can be very likely used to also prove stability properties of

- CSA-ES (cumulative step-size adaptation)
- more complete variants of CMA-ES
- some continuous EDAs
- the cross entropy method in continuous domain

... and beyond

The mathematical tools can be very likely used to also prove stability properties of

- CSA-ES (cumulative step-size adaptation)
- more complete variants of CMA-ES
- some continuous EDAs
- the cross entropy method in continuous domain

Thank You