The challenges for stochastic optimization and a variable metric approach

Nikolaus Hansen, INRIA Saclay

Microsoft Research–INRIA Joint Centre, INRIA Saclay

April 6, 2009

Content

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

Problem Statement

Continuous Domain Search/Optimization

- Task: **minimize** a **objective function** (*fitness* function, *loss* function) in continuous domain

$$f : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}, \qquad \underline{x} \mapsto f(\underline{x})$$

- **Black Box** scenario (direct search scenario)

  - gradients are not available or not useful
  - problem domain specific knowledge is used only within the black box, e.g. within an appropriate encoding
- Search costs: number of function evaluations

Problem Statement

Continuous Domain Search/Optimization

- Goal
  - solution x with **small function value** with **least search cost**
    there are two conflicting objectives
  - fast convergence to the global optimum
    . . . or to a robust solution x
- Typical Examples
  - shape optimization (e.g. using CFD) — curve fitting, airfoils
  - model calibration — biological, physical
  - parameter calibration — controller, plants, images

Approach: stochastic search, Evolutionary Algorithms

. . . metaphores

Metaphors

Evolutionary Computation		Optimization
genome	$\longleftrightarrow$	decision variables
		design variables
		object variables
individual, offspring, parent	$\longleftrightarrow$	candidate solution
population	$\longleftrightarrow$	set of candidate solutions
fitness function	$\longleftrightarrow$	objective function
		loss function
		cost function
generation	$\longleftrightarrow$	iteration

. . . properties

Objective Function Properties

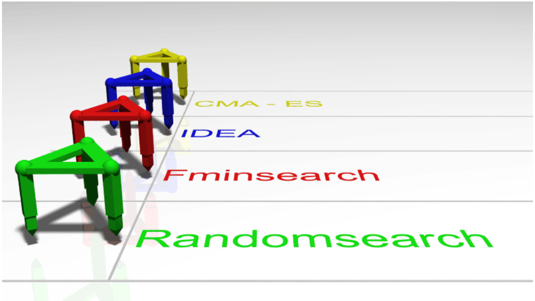
We assume $f : \mathcal{X} \subset \mathbb{R}^n \to \mathbb{R}$ to be *non-linear*, *non-separable* and to have at least moderate dimensionality, say $n \not\ll 10$.
Additionally, f can be

- non-convex
- non-smooth — derivatives do not exist
- discontinuous
- ill-conditioned
- multimodal — there are eventually many local optima
- noisy
- ...

Goal : cope with any of these function properties
they are related to real-world problems

. . . Tripeds

Comparison of CMA-ES, IDEA and Simplex-Downhill



CMA-ES: Covariance Matrix Adaptation Evolution Strategy
IDEA: Iterated Density-Estimation Evolutionary Algorithm[1]

Fminsearch: Nelder-Mead simplex downhill method[2]

see...
`http://www.icos.ethz.ch/cse/research/highlights/Race.gif`

...function properties

[1] Bosman (2003) Design and Application of Iterated Density-Estimation Evolutionary Algorithms. PhD thesis.
[2]

What Makes a Function Difficult to Solve?

Why stochastic search?

- ruggedness

  non-smooth, discontinuous, multimodal, and/or noisy function

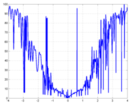
- non-separability

  dependencies between the objective variables

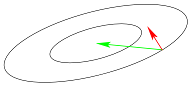
- dimensionality

  (considerably) larger than three

- ill-conditioning



cut from 5-D solvable example



gradient direction $-f'(\underline{x})^{\mathrm{T}}$

Newton direction $-\underline{\underline{H}}^{-1} f'(\underline{x})^{\mathrm{T}}$

Separable Problems

Definition (Separable Problem)

A function f is separable if

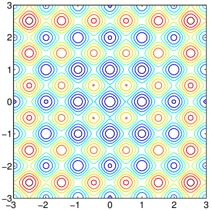
$$\arg\min_{(x_1,\dots,x_n)} f(x_1,\dots,x_n) = \left(\arg\min_{x_1} f(x_1,\dots),\dots,\arg\min_{x_n} f(\dots,x_n)\right)$$

$\Rightarrow$ it follows that f can be optimized in a sequence of n independent 1-D optimization processes

Example: Additively decomposable functions

$$f(x_1,\dots,x_n) = \sum_{i=1}^{n} f_i(x_i)$$

Rastrigin function



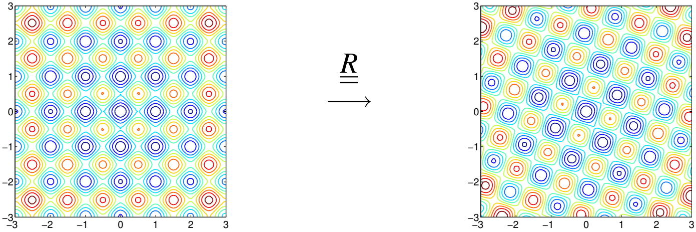
Non-Separable Problems

Building a non-separable problem from a separable one

Rotating the coordinate system

- $f : \underline{x} \mapsto f(\underline{x})$ separable
- $f : \underline{x} \mapsto f(\underline{\underline{R}}\underline{x})$ **non-separable**

$\underline{\underline{R}}$ rotation matrix



[3,4]

[3] Hansen, Ostermeier, Gawelczyk (1995). On the adaptation of arbitrary normal mutation distributions in evolution strategies: The generating set adaptation. Sixth ICGA, pp. 57-64, Morgan Kaufmann

[4] Salomon (1996). "Reevaluating Genetic Algorithm Performance under Coordinate Rotation of Benchmark Functions; A survey of some theoretical and practical aspects of genetic algorithms." BioSystems, 39(3):263-278

Curse of Dimensionality

The term *Curse of dimensionality* (Richard Bellman) refers to problems caused by the **rapid increase in volume** associated with adding extra dimensions to a (mathematical) space.

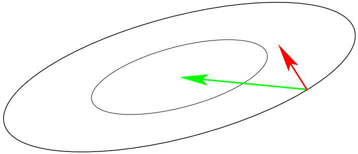
Example: Consider placing 100 points onto a real interval, say $[-1, 1]$. To get **similar coverage**, in terms of distance between adjacent points, of the 10-dimensional space $[-1, 1]^{10}$ would require $100^{10} = 10^{20}$ points. A 100 points appear now as isolated points in a vast empty space.

Consequently, a **search policy** (e.g. exhaustive search) that is valuable in small dimensions **might be useless** in moderate or large dimensional search spaces.

Ill-Conditioned Problems

Curvature of level sets

Consider the convex-quadratic function
$f(\underline{x}) = \frac{1}{2}(\underline{x} - \underline{x}^*)^T \underline{\underline{H}}(\underline{x} - \underline{x}^*)$



gradient direction $-f'(\underline{x})^{\mathrm{T}}$

Newton direction $-\underline{\underline{H}}^{-1} f'(\underline{x})^{\mathrm{T}}$

Condition number equals nine here. Condition numbers between 100 and even 10^{10} can often be observed in real world problems.

If $\underline{\underline{H}} \approx \underline{\underline{I}}$ (small condition number of $\underline{\underline{H}}$) first order information (e.g. the gradient) is sufficient. Otherwise **second order information** (estimation of $\underline{\underline{H}}^{-1}$) **is required**.

What Makes a Function Difficult to Solve?

... and what can be done

Challenge	Approach in Evolutionary Computation
Ruggedness	**non-local** policy, large sampling width (step-size) as large as possible while preserving a reasonable convergence speed
	stochastic, non-elitistic, **population-based** method
	recombination operator serves as repair mechanism
Dimensionality, Non-Separability	exploiting the problem structure locality, neighborhood, encoding
Ill-conditioning	second order approach changes the neighborhood metric

Stochastic Search

A black box search template to minimize $f : \mathbb{R}^n \to \mathbb{R}$

Initialize distribution parameters $\underline{\theta}$, set population size $\lambda \in \mathbb{N}$

While not terminate

1. **Sample distribution** $P(\underline{x}|\underline{\theta}) \to \underline{x}_1, \dots, \underline{x}_\lambda \in \mathbb{R}^n$
2. **Evaluate** $\underline{x}_1, \dots, \underline{x}_\lambda$ **on** f
3. **Update parameters** $\underline{\theta} \leftarrow F_\theta(\underline{\theta}, \underline{x}_1, \dots, \underline{x}_\lambda, f(\underline{x}_1), \dots, f(\underline{x}_\lambda))$

Everything depends on the definition of P and F_θ

deterministic algorithms are covered as well

In Evolutionary Algorithms the distribution P is often implicitly defined via **operators on a population**, in particular, selection, recombination and mutation

Natural template for *Estimation of Distribution Algorithms*

Stochastic Search

A black box search template to minimize $f : \mathbb{R}^n \to \mathbb{R}$

Initialize distribution parameters $\underline{\theta}$, set population size $\lambda \in \mathbb{N}$

While not terminate

1. **Sample distribution** $P(\underline{x}|\underline{\theta}) \to \underline{x}_1, \ldots, \underline{x}_\lambda \in \mathbb{R}^n$
2. **Evaluate** $\underline{x}_1, \ldots, \underline{x}_\lambda$ **on** f
3. **Update parameters** $\underline{\theta} \leftarrow F_\theta(\underline{\theta}, \underline{x}_1, \ldots, \underline{x}_\lambda, f(\underline{x}_1), \ldots, f(\underline{x}_\lambda))$

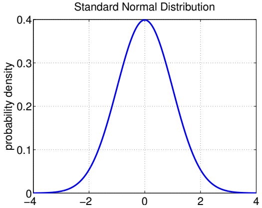
In the following

- P is a **multi-variate normal** distribution $\mathcal{N}\left(\underline{m}, \sigma^2 \underline{\underline{C}}\right) \sim \underline{m} + \sigma \mathcal{N}\left(\underline{0}, \underline{\underline{C}}\right)$

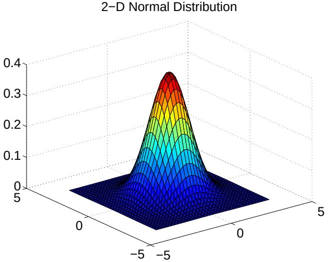
  $\underline{\theta} = \{\underline{m}, \underline{\underline{C}}, \sigma\} \in \mathbb{R}^n \times \mathbb{R}^{n \times n} \times \mathbb{R}_+$
- $F_\theta = F_\theta(\underline{\theta}, \underline{x}_{1:\lambda}, \ldots, \underline{x}_{\mu:\lambda})$, where $\mu \leq \lambda$ and $\underline{x}_{i:\lambda}$ is the i-th best of the λ points

. . . why?

Normal Distribution



probability density of 1-D standard normal distribution



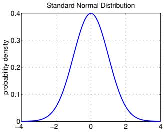
probability density of 2-D normal distribution

The Multi-Variate (n-Dimensional) Normal Distribution

Any multi-variate normal distribution $\mathcal{N}\left(\underline{m}, \underline{\underline{C}}\right)$ is uniquely determined by its mean value $\underline{m} \in \mathbb{R}^n$ and its symmetric positive definite $n \times n$ covariance matrix $\underline{\underline{C}}$.

The **mean** value $\underline{m}$

- determines the displacement (translation)
- is the value with the largest density (modal value)
- the distribution is symmetric about the distribution mean

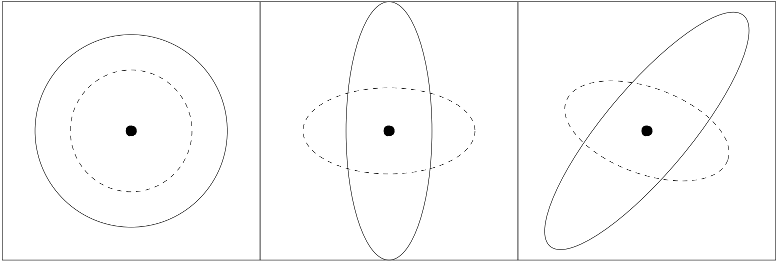


The **covariance matrix** $\underline{\underline{C}}$

- determines the shape
- has a valuable **geometrical interpretation**: any covariance matrix can be uniquely identified with the iso-density ellipsoid $\{\underline{x} \in \mathbb{R}^n \mid \underline{x}^{\mathrm{T}} \underline{\underline{C}}^{-1} \underline{x} = 1\}$

. . . any **covariance matrix** can be uniquely identified with the iso-density ellipsoid $\{\underline{x} \in \mathbb{R}^n \mid \underline{x}^{\mathrm{T}} \underline{\underline{C}}^{-1} \underline{x} = 1\}$

Lines of Equal Density



$\mathcal{N}\left(\underline{m}, \sigma^2 \underline{\underline{I}}\right) \sim \underline{m} + \sigma \mathcal{N}\left(\underline{0}, \underline{\underline{I}}\right)$

one degree of freedom σ

components are independent standard normally distributed

$\mathcal{N}\left(\underline{m}, \underline{\underline{D}}^2\right) \sim \underline{m} + \underline{\underline{D}} \mathcal{N}\left(\underline{0}, \underline{\underline{I}}\right)$

n **degrees of freedom**

components are independent, scaled

$\mathcal{N}\left(\underline{m}, \underline{\underline{C}}\right) \sim \underline{m} + \underline{\underline{C}}^{\frac{1}{2}} \mathcal{N}\left(\underline{0}, \underline{\underline{I}}\right)$

$(n^2 + n)/2$ **degrees of freedom**

components are correlated

where $\underline{\underline{I}}$ is the identity matrix (isotropic case) and $\underline{\underline{D}}$ is a diagonal matrix (reasonable for separable problems) and $\underline{\underline{A}} \times \mathcal{N}\left(\underline{0}, \underline{\underline{I}}\right) \sim \mathcal{N}\left(\underline{0}, \underline{\underline{A}}\underline{\underline{A}}^{\mathrm{T}}\right)$ holds for all $\underline{\underline{A}}$.

. . . CMA

Stochastic Search

A black box search template to minimize $f : \mathbb{R}^n \to \mathbb{R}$

Initialize distribution parameters $\underline{\theta}$, set population size $\lambda \in \mathbb{N}$

While not terminate

1. **Sample distribution** $P(\underline{x}|\underline{\theta}) \to \underline{x}_1, \dots, \underline{x}_\lambda \in \mathbb{R}^n$
2. **Evaluate** $\underline{x}_1, \dots, \underline{x}_\lambda$ **on** f
3. **Update parameters** $\underline{\theta} \leftarrow F_\theta(\underline{\theta}, \underline{x}_1, \dots, \underline{x}_\lambda, f(\underline{x}_1), \dots, f(\underline{x}_\lambda))$

P is a multi-variate normal distribution

$$\mathcal{N}\left(\underline{m}, \sigma^2 \underline{\underline{C}}\right) \sim \underline{m} + \sigma \mathcal{N}\left(\underline{0}, \underline{\underline{C}}\right)$$

. . . sampling

Sampling New Search Points

The Mutation Operator

New search points are sampled normally distributed

$$\underline{x}_i \sim \underline{m} + \sigma \mathcal{N}_i\left(\underline{0}, \underline{\underline{C}}\right) \qquad \text{for } i = 1, \dots, \lambda$$

as perturbations of $\underline{m}$ where $\underline{x}_i, \underline{m} \in \mathbb{R}^n$, $\sigma \in \mathbb{R}_+$, and $\underline{\underline{C}} \in \mathbb{R}^{n \times n}$

where

- the mean vector $\underline{m} \in \mathbb{R}^n$ represents the favorite solution
- the so-called step-size $\sigma \in \mathbb{R}_+$ controls the *step length*
- the covariance matrix $\underline{\underline{C}} \in \mathbb{R}^{n \times n}$ determines the **shape** of the distribution ellipsoid

The question remains how to update $\underline{m}$, $\underline{\underline{C}}$, and σ.

Update of the Distribution Mean m

Selection and Recombination

Given the i-th solution point $x_i = m + \sigma \underbrace{\mathcal{N}_i\left(0, C\right)}_{=: y_i} = m + \sigma\, y_i$

Let $x_{i:\lambda}$ the **i-th ranked** solution point, such that $f(x_{1:\lambda}) \leq \cdots \leq f(x_{\lambda:\lambda})$.
The new mean reads

$$m \leftarrow \sum_{i=1}^{\mu} w_i\, x_{i:\lambda} \;=\; m + \sigma \underbrace{\sum_{i=1}^{\mu} w_i\, y_{i:\lambda}}_{=: y_w}$$

where

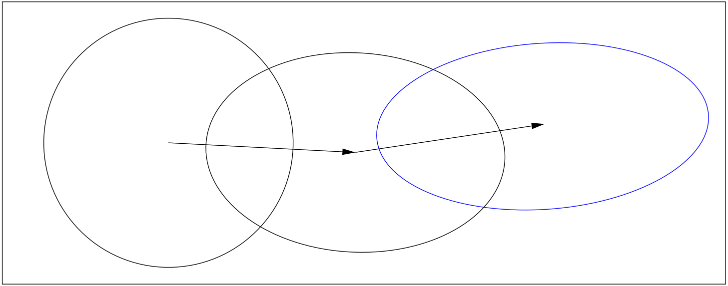
$$w_1 \geq \cdots \geq w_\mu > 0, \quad \textstyle\sum_{i=1}^{\mu} w_i = 1$$

The best μ points are selected **from the sampled solutions** (non-elitistic) and a **weighted mean** is taken.

Covariance Matrix Adaptation

Rank-One Update

$$\underline{m} \leftarrow \underline{m} + \sigma \underline{y}_w, \quad \underline{y}_w = \sum_{i=1}^{\mu} w_i \, \underline{y}_{i:\lambda}, \quad \underline{y}_i \sim \mathcal{N}_i\left(\underline{0}, \underline{\underline{C}}\right)$$



new distribution,

$\underline{\underline{C}} \leftarrow 0.8 \times \underline{\underline{C}} + 0.2 \times \underline{y}_w \underline{y}_w^{\mathrm{T}}$

the ruling principle: the adaptation **increases the likelyhood of successful steps**, $\underline{y}_w$, to appear again

. . . equations

Preliminary Set of Equations

Covariance Matrix Adaptation with Rank-One Update

Initialize $\underline{m} \in \mathbb{R}^n$, and $\underline{\underline{C}} = \underline{\underline{I}}$, set $\sigma = 1$, learning rate $c_{\text{cov}} \approx 2/n^2$
While not terminate

$$
\begin{aligned}
\underline{x}_i &= \underline{m} + \sigma\, \underline{y}_i, \qquad \underline{y}_i \sim \mathcal{N}_i\left(\underline{0}, \underline{\underline{C}}\right), \qquad i = 1, \ldots, \lambda \\
\underline{m} &\leftarrow \underline{m} + \sigma \underline{y}_w \qquad \text{where } \underline{y}_w = \sum_{i=1}^{\mu} w_i\, \underline{y}_{i:\lambda} \\
\underline{\underline{C}} &\leftarrow (1 - c_{\text{cov}})\underline{\underline{C}} + c_{\text{cov}} \mu_w \underbrace{\underline{y}_w \underline{y}_w^{\mathrm{T}}}_{\text{rank-one}} \qquad \text{where } \mu_w = \frac{1}{\sum_{i=1}^{\mu} {w_i}^2} \geq 1
\end{aligned}
$$

λ can be small

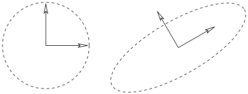
$$\underline{\underline{C}} \leftarrow (1 - c_{\mathrm{cov}})\underline{\underline{C}} + c_{\mathrm{cov}}\,\mu_w\,\underline{y}_w\underline{y}_w^{\mathrm{T}}$$

The covariance matrix adaptation

- learns all **pairwise dependencies** between variables
  off-diagonal entries in the covariance matrix reflect the dependencies

- conducts a **principle component analysis** (PCA) of steps $\underline{y}_w$, sequentially in time and space
  eigenvectors of the covariance matrix $\underline{\underline{C}}$ are the principle components
  / the principle axes of the mutation ellipsoid

- approximates the **inverse Hessian** on convex-quadratic functions
  overwhelming empirical evidence, proof is in progress



- learns a new, **rotated problem representation** and a **new variable metric** (Mahalanobis)
  components are independent (only) in the new representation
  rotational invariant
  equivalent with an adaptive (general) linear encoding[a]

[a] Hansen 2000, Invariance, Self-Adaptation and Correlated Mutations in Evolution Strategies, PPSN VI

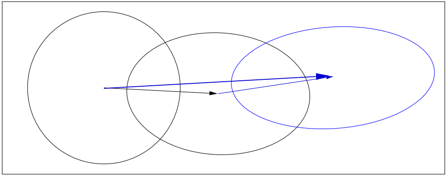
. . . cumulation, step-size control

Cumulation

The Evolution Path

Evolution Path

Conceptually, the evolution path is the path the strategy mean $\underline{m}$ takes over a number of generation steps.



An exponentially weighted sum of steps $\underline{y}_w$ is used

$$\underline{p}_c \propto \sum_{i=0}^{g} \underbrace{(1-c_c)^{g-i}}_{\text{exponentially fading weights}} \underline{y}_w^{(i)}$$

The recursive construction of the evolution path (cumulation):

$$\underline{p}_c \leftarrow \underbrace{(1-c_c)}_{\text{decay factor}} \underline{p}_c + \underbrace{\sqrt{1-(1-c_c)^2}\sqrt{\mu_w}}_{\text{normalization factor}} \underbrace{\underline{y}_w}_{\text{input, } \frac{\underline{m}-\underline{m}_{\text{old}}}{\sigma}}$$

where $\mu_w = \frac{1}{\sum w_i{}^2}$, $c_c \ll 1$. History information is accumulated in the evolution path.

"Cumulation" is a widely used technique and also know as

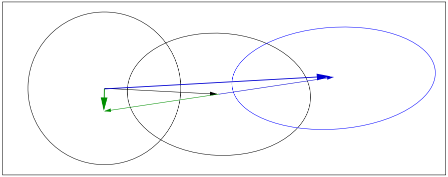
- *exponential smoothing* in time series, forecasting
- exponentially weighted *mooving average*
- *iterate averaging* in stochastic approximation
- *momentum* in the back-propagation algorithm for ANNs
- ...

. . . why?

Cumulation

Utilizing the Evolution Path

We used $\underline{y}_w \underline{y}_w^{\mathrm{T}}$ for updating $\underline{\underline{C}}$. Because $\underline{y}_w \underline{y}_w^{\mathrm{T}} = -\underline{y}_w (-\underline{y}_w)^{\mathrm{T}}$ **the sign** of $\underline{y}_w$ is neglected. The sign information is (re-)introduced by using the *evolution path*.



$$\underline{p}_{\mathrm{c}} \quad \leftarrow \quad \underbrace{(1 - c_{\mathrm{c}})}_{\text{decay factor}} \underline{p}_{\mathrm{c}} + \underbrace{\sqrt{1 - (1 - c_{\mathrm{c}})^2} \sqrt{\mu_w}}_{\text{normalization factor}} \underline{y}_w$$

where $\mu_w = \frac{1}{\sum {w_i}^2}$, $c_{\mathrm{c}} \ll 1$.

. . . equations

Preliminary Set of Equations (2)

Covariance Matrix Adaptation, Rank-One Update with Cumulation

Initialize $\underline{m} \in \mathbb{R}^n$, $\underline{\underline{C}} = \underline{\underline{I}}$, and $\underline{p}_\mathrm{c} = \underline{0} \in \mathbb{R}^n$,
set $\sigma = 1$, $c_\mathrm{c} \approx 4/n$, $c_\mathrm{cov} \approx 2/n^2$
While not terminate

$$
\begin{aligned}
\underline{x}_i &= \underline{m} + \sigma\, \underline{y}_i, \quad \underline{y}_i \sim \mathcal{N}_i\left(\underline{0}, \underline{\underline{C}}\right), \qquad i = 1, \ldots, \lambda \\
\underline{m} &\leftarrow \underline{m} + \sigma\, \underline{y}_w \quad \text{where } \underline{y}_w = \sum_{i=1}^{\mu} w_i\, \underline{y}_{i:\lambda} \\
\underline{p}_\mathrm{c} &\leftarrow (1 - c_\mathrm{c})\, \underline{p}_\mathrm{c} + \sqrt{1 - (1 - c_\mathrm{c})^2}\sqrt{\mu_w}\, \underline{y}_w \\
\underline{\underline{C}} &\leftarrow (1 - c_\mathrm{cov})\underline{\underline{C}} + c_\mathrm{cov} \underbrace{\underline{p}_\mathrm{c}\underline{p}_\mathrm{c}^{\mathrm{T}}}_{\text{rank-one}}
\end{aligned}
$$

$\ldots \mathcal{O}(n^2)$ to $\mathcal{O}(n)$

Using an **evolution path** for the **rank-one update** of the covariance matrix reduces the number of function evaluations to adapt to a straight ridge **from** $\mathcal{O}(n^2)$ **to** $\mathcal{O}(n)$.[a]

[a] Hansen, Müller and Koumoutsakos 2003. Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES). *Evolutionary Computation, 11(1)*, pp. 1-18

The overall model complexity is n^2 but important parts of the model can be learned in time of order n

. . . step-size

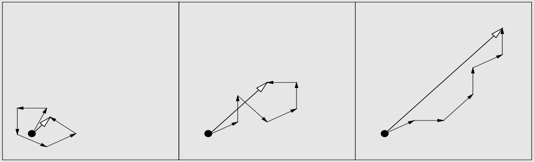
Path Length Control

The Concept

$$\underline{x}_i = \underline{m} + \sigma\, \underline{y}_i$$
$$\underline{m} \leftarrow \underline{m} + \sigma \underline{y}_w$$

Measure the length of the *evolution path*

the pathway of the mean vector $\underline{m}$ in the generation sequence



$\Downarrow$ decrease σ $\Downarrow$ increase σ

loosely speaking steps are

- perpendicular under random selection (in expectation)
- perpendicular in the desired situation (to be most efficient)

Summary

Covariance Matrix Adaptation Evolution Strategy (CMA-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

   preserves invariance

3. *Covariance matrix adaptation (CMA)* **increases the likelyhood of previously successful steps**

   learning all pairwise dependencies

   $\Longrightarrow$ adapts a variable metric

   $\Longrightarrow$ new (rotated) problem representation

4. An **evolution path** (a non-local trajectory)
   - enhances the covariance matrix (rank-one) adaptation

     yields sometimes linear time complexity
   - controls the **step-size** (step length)

     aims at conjugate perpendicularity

Summary of Equations

The Covariance Matrix Adaptation Evolution Strategy

Initialize $\underline{m} \in \mathbb{R}^n$, $\sigma \in \mathbb{R}_+$, $\underline{\underline{C}} = \underline{\underline{I}}$, and $\underline{p}_c = \underline{0}$, $\underline{p}_\sigma = \underline{0}$,
set $c_c \approx 4/n$, $c_\sigma \approx 4/n$, $c_1 \approx 2/n^2$, $c_\mu \approx \mu_w/n^2$, $c_1 + c_\mu \leq 1$,
$d_\sigma \approx 1 + \sqrt{\frac{\mu_w}{n}}$,
set λ and $w_i, i = 1, \ldots, \mu$ such that $\mu_w \approx 0.3\,\lambda$

While not terminate

$$\underline{x}_i = \underline{m} + \sigma\,\underline{y}_i, \quad \underline{y}_i \sim \mathcal{N}_i\left(\underline{0}, \underline{\underline{C}}\right), \qquad \text{sampling}$$

$$\underline{m} \leftarrow \underline{m} + \sigma\,\underline{y}_w \quad \text{where } \underline{y}_w = \textstyle\sum_{i=1}^{\mu} w_i\,\underline{y}_{i:\lambda} \qquad \text{update mean}$$

$$\underline{p}_c \leftarrow (1 - c_c)\,\underline{p}_c + \mathbb{1}_{\{\|\underline{p}_\sigma\| < 1.5\sqrt{n}\}}\sqrt{1 - (1 - c_c)^2}\sqrt{\mu_w}\,\underline{y}_w \qquad \text{cumulation for } \underline{\underline{C}}$$

$$\underline{p}_\sigma \leftarrow (1 - c_\sigma)\,\underline{p}_\sigma + \sqrt{1 - (1 - c_\sigma)^2}\sqrt{\mu_w}\,\underline{\underline{C}}^{-\frac{1}{2}}\,\underline{y}_w \qquad \text{cumulation for } \sigma$$

$$\underline{\underline{C}} \leftarrow (1 - c_1 - c_\mu)\,\underline{\underline{C}} + c_1\,\underline{p}_c\,\underline{p}_c^{\mathrm{T}} + c_\mu \textstyle\sum_{i=1}^{\mu} w_i\,\underline{y}_{i:\lambda}\underline{y}_{i:\lambda}^{\mathrm{T}} \qquad \text{update } \underline{\underline{C}}$$

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

Experimentum Crucis

What did we want to achieve?

- reduce any convex-quadratic function

$$f(\underline{x}) = \underline{x}^{\mathrm{T}} \underline{\underline{H}} \underline{x}$$

e.g. $f(\underline{x}) = \sum_{i=1}^{n} 10^{6\frac{i-1}{n-1}} x_i^2$

to the sphere model

$$f(\underline{x}) = \underline{x}^{\mathrm{T}} \underline{x}$$

without use of derivatives

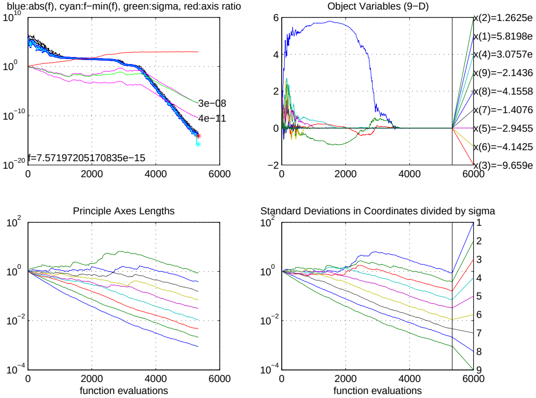
- lines of equal density align with lines of equal fitness

$$\underline{\underline{C}} \propto \underline{\underline{H}}^{-1}$$

in a stochastic sense

Experimentum Crucis (1)

f convex-quadratic, separable

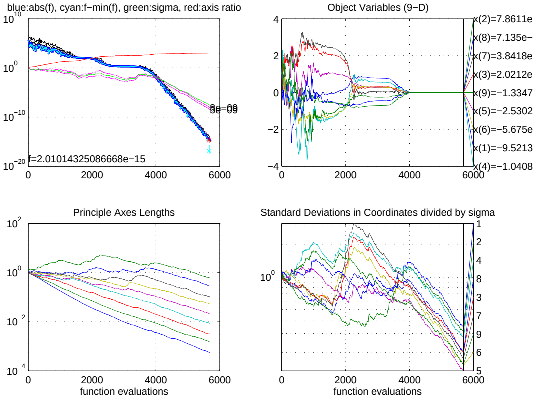


$$f(\underline{x}) = \sum_{i=1}^{n} 10^{\alpha \frac{i-1}{n-1}} x_i^2, \alpha = 6$$

. . . crucis rotated

Experimentum Crucis (2)

f convex-quadratic, as before but non-separable (rotated)



$\underline{C} \propto \underline{\underline{H}}^{-1}$ **for all** $g, \underline{\underline{H}}$

$f(\underline{x}) = g\left(\underline{x}^{\mathrm{T}}\underline{\underline{H}}\underline{x}\right)$, $g : \mathbb{R} \to \mathbb{R}$ stricly monotonic

. . . on convergence

On Global Convergence

- convergence on a very **broad class of functions**, e.g. for Monte Carlo pure random search

very slow

- convergence with practically **feasible convergence rates** on, e.g., $\|\underline{x}\|^\alpha$

Markov Chain analysis

Stability/Stationarity/Ergodicity of a markov chain

- the markov chain always **returns to "the center"** of the state space (recurrence)
- the chain exhibits an *invariant measure*, a limit probability distribution

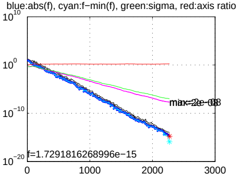
implies convergence/divergence

The Convergence Rate

Optimal convergence rate can be achieved

The converence rate for evolution strategies on $f(\underline{x}) = g(\|\underline{x} - \underline{x}^*\|)$ in iteration t reads[5]

$$\lim_{t\to\infty} \frac{1}{t} \sum_{k=1}^{t} \log \frac{\|\underline{m}_k - \underline{x}^*\|}{\|\underline{m}_{k-1} - \underline{x}^*\|} \quad \propto \quad -\frac{1}{n}$$



loosely

$$\|\underline{m}_t - \underline{x}^*\| \quad \propto \quad \exp\left(-\frac{t}{n}\right) = \left(\frac{1}{e^t}\right)^{1/n}$$

random search exhibits $\|\underline{m}_t - \underline{x}^*\| \propto \left(\frac{1}{t}\right)^{1/n}$

which is the **lower bound** for randomized direct search with isotropic sampling [6]

[5] Auger 2005

[6] Jägersküpper 2008

Convergence of the Covariance Matrix

Yet to be proven

Theorem (convergence of covariance matrix $\underline{\underline{C}}$)

Given the function

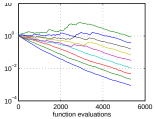
$$f(\underline{x}) = g(\underline{x}^{\mathrm{T}} \underline{\underline{H}} \underline{x})$$

where $\underline{\underline{H}}$ is positive and g is monotonic, we have

$$E(\underline{\underline{C}}) \propto \underline{\underline{H}}^{-1}$$

where the expectation is taken with respect to the invariant measure

without use of derivatives



Evaluation/Selection of Search Algorithms

Evaluation (of the performance) of a search algorithm needs

- meaningful quantitative measure on benchmark functions or real world problems
- account for meta-parameter tuning

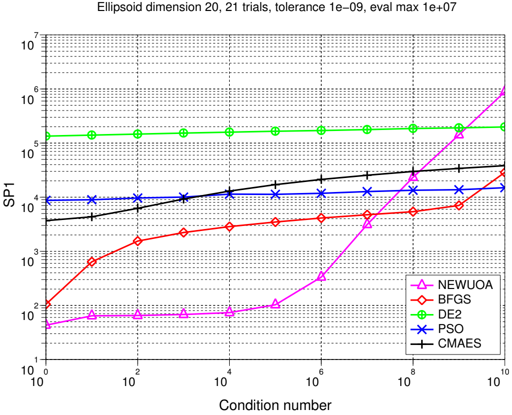
  can be quite expensive
- account for invariance properties (symmetries)

  prediction of performance is based on “similarity”, ideally equivalence classes of functions
- account for algorithm internal cost

  often negligible, depending on the objective function cost

Comparison to BFGS, NEWUOA, PSO and DE (1)

f convex-quadratic, separable with varying α



$f(\underline{x}) = g(\underline{x}^{\mathrm{T}}\underline{\underline{H}}\underline{x})$ with

g identity (BFGS, NEWUOA) or

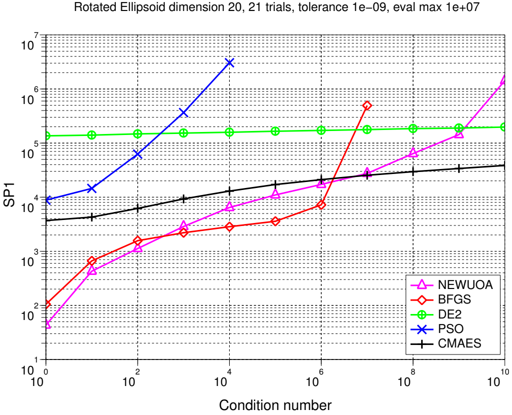
g order-preserving (strictly increasing, all other)

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

…population size, invariance

Comparison to BFGS, NEWUOA, PSO and DE (2)

f convex-quadratic, non-separable (rotated) with varying α



$f(\underline{x}) = g(\underline{x}^{\mathrm{T}} \underline{\underline{H}} \underline{x})$ with

g identity (BFGS, NEWUOA) or

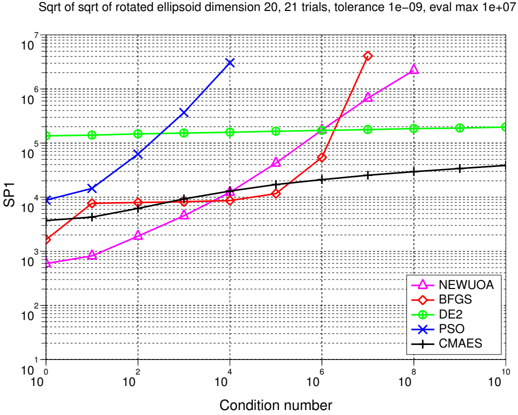
g order-preserving (strictly increasing, all other)

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

…population size, invariance

Comparison to BFGS, NEWUOA, PSO and DE (3)

f non-convex, non-separable (rotated) with varying α



$f(\underline{x}) = g(\underline{x}^{\mathrm{T}} \underline{\underline{H}} \underline{x})$ with

$g(.) = (.)^{1/4}$ (BFGS, NEWUOA) or

g any order-preserving (strictly increasing, all other)

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

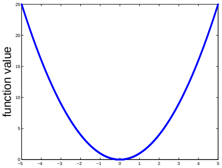
...population size, invariance

Invariance

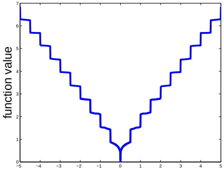
The short version

The grand aim of all science is to cover the greatest number of empirical facts by logical deduction from the smallest number of hypotheses or axioms.

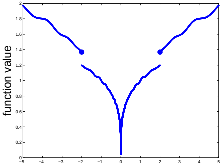
— Albert Einstein



$f(x) = x^2$



$f(x) = g_1(x^2)$



$f(x) = g_2(x^2)$

- all three functions are **equivalent** for rank-based search methods

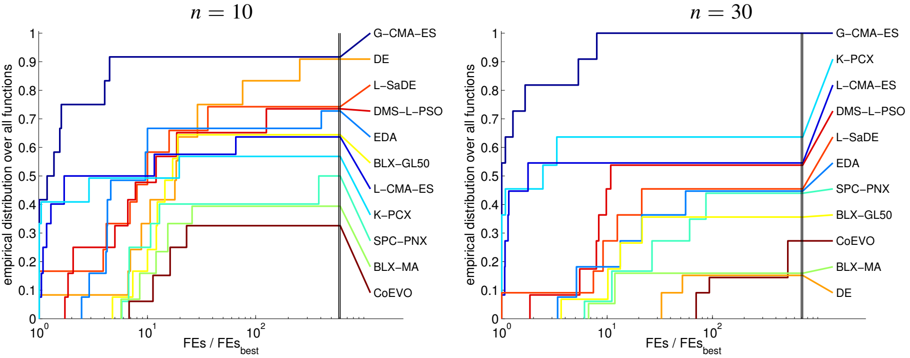
  large equivalence class

- invariance allows a save **generalization** of empirical results

  here on $f(x) = x^2$ (left) to any $f(x) = g(x^2)$, where g is monotonous

Comprehensive Comparison of 11 Algorithms

Empirical Distribution of Normalized Success Performance



$\mathtt{FEs} = \mathtt{mean}(\#\mathit{fevals}) \times \frac{\#\text{all runs (25)}}{\#\text{successful runs}}$, where $\#\mathit{fevals}$ includes only successful runs.

Shown: **empirical distribution function** of the Success Performance FEs divided by FEs of the best algorithm on the respective function.

Results of all functions are used where at least one algorithm was successful at least once, i.e. where the target function value was reached in at least one experiment (out of 11×25 experiments).

Small values for FEs and therefore large (cumulative frequency) values in the graphs are preferable.

Merci !

http://www.lri.fr/~hansen/cmaesintro.html
or google NIKOLAUS HANSEN