Globally Convergent Evolution Strategies and CMA-ES

Y. Diouane[1] S. Gratton[1,2]

[1] CERFACS, Toulouse

[2] ENSEEIHT-IRIT, Toulouse

Journée des Doctorants, September 7, 2012

Outline

1. Derivative free optimization

2. Direct-search and evolution strategy
   - Algorithmic changes
   - Convergence
   - Numerical results

Motivations

- Minimizing $f(x)$, f is assumed to be differentiable
- The derivatives of f are not available (expensive in CPU, legacy code, not coded)
- Basically two kinds of methods are available in the num. opt. community : model based (MB), direct search methods (DS)
- Important note : algorithm for few hundreds of **true** degrees of freedom at most (forms of model reduction should be considered)

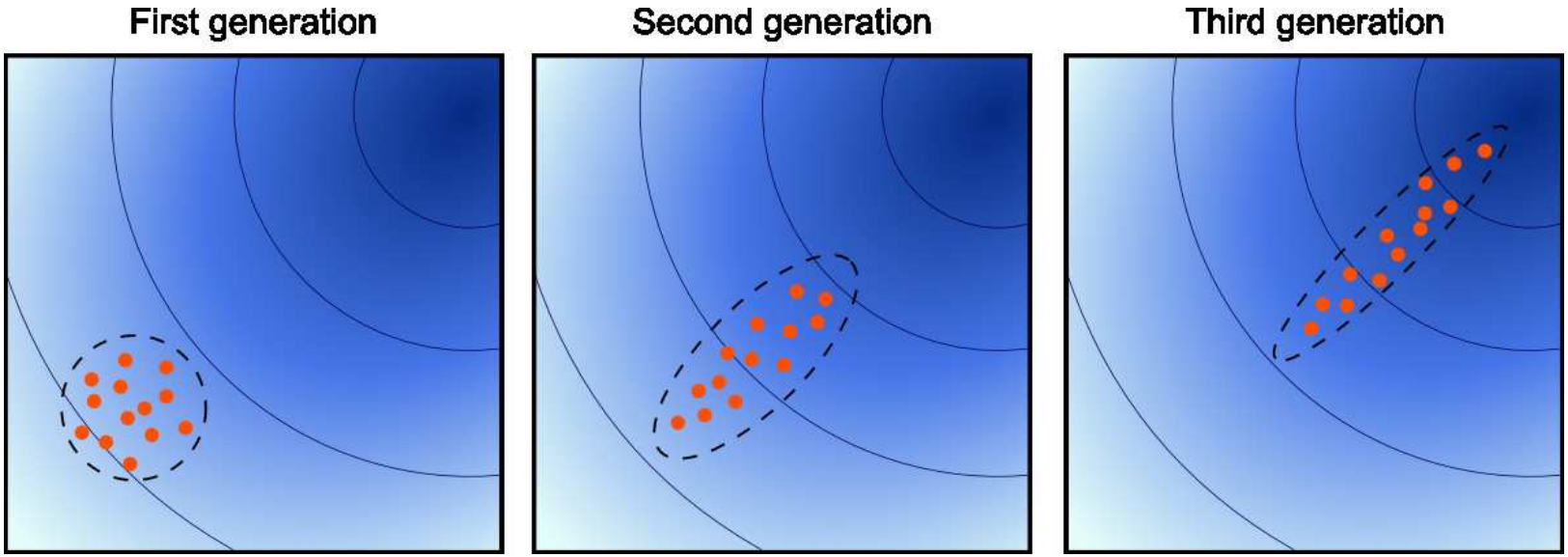
Model based algorithms

- Rely on interpolation, regression models at the current point
- Minimize the model inside a trust-region
- Accept the step depending on the ratio of achieved versus predicted reduction
- Possibly reduce the TR radius if model is good enough and the ratio is bad
- Convergence can be proved to first order critical points for differentiable functions. Quick convergence observed
- Codes BOBYQA [Powell 2009], BC-DFO [Gratton, Toint, Troeltzsch 2011]

Main objectives

- Achieve convergence from any starting point to a stationary point for methods based on sampling, such as evolution algorithms
- Obtain good performance on practical problems, in terms of function evalutations, supposed to be expensive
- Just consider the Holy Grail: approaching the global minimum of a black-box function

Evolution Strategies



- A large Class of evolution strategies (ES) [see Hansen and Auger (INRIA)]
- Main features (for our applications)
  - "Good" convergence to global minima at low cost
  - Considered as one of the best evolution strategies
  - Evolution of a population with scaling σ and distribution $\mathscr{C}$ adaptation

Evolution Strategies

Algorithm

1. Offspring Generation: Compute new sample points $Y_{k+1} = \{y_{k+1}^1, ..., y_{k+1}^{m_\lambda}\}$such that

$$y_{k+1}^i = x_k + \sigma_k^{ES} d_k^i$$

where $d_k^i$is drawn from a distribution $\mathscr{C}_k$, $i = 1, ..., m_\lambda$.

2. Parent Selection: Evaluate $f(y_{k+1}^i), i = 1, ..., m_\lambda$,and reorder $Y_{k+1} = \{\tilde{y}_{k+1}^1, ..., \tilde{y}_{k+1}^{m_\lambda}\}$: $f(\tilde{y}_{k+1}^1) \leq ... \leq f(\tilde{y}_{k+1}^{m_\lambda})$.

$$x_{k+1} = \sum_{i=1}^{m_\mu} \omega_k^i \tilde{y}_{k+1}^i \quad m_\lambda \geq m_\mu ..$$

3. Updates: σ_{k+1}^{ES}, $\mathscr{C}_k$ and $(\omega_0^1, ..., \omega_0^{m_\mu}) \in \mathbb{S}$. Return to step (1).

Making the convergence global

- The method resembles a direct search method, where the function is successively minimized on a dense set of directions [Audet, Denis, 2006]
- The method does not rely on integer lattices as do other methods (MADS); we have to develop an adapted convergence theory
- For that purpose, we introduced two changes in the algorithm (least possible change)
  - As for other (derivative free) optimization method, we will ask more than the simple decrease along the iterates
  - We also require a control of the scaling parameter (i.e. steplength), and possibly scale too long steps
- Idea: give a general framework into which a slight modification of CMA can fit, hopefully not jeopardizing performance

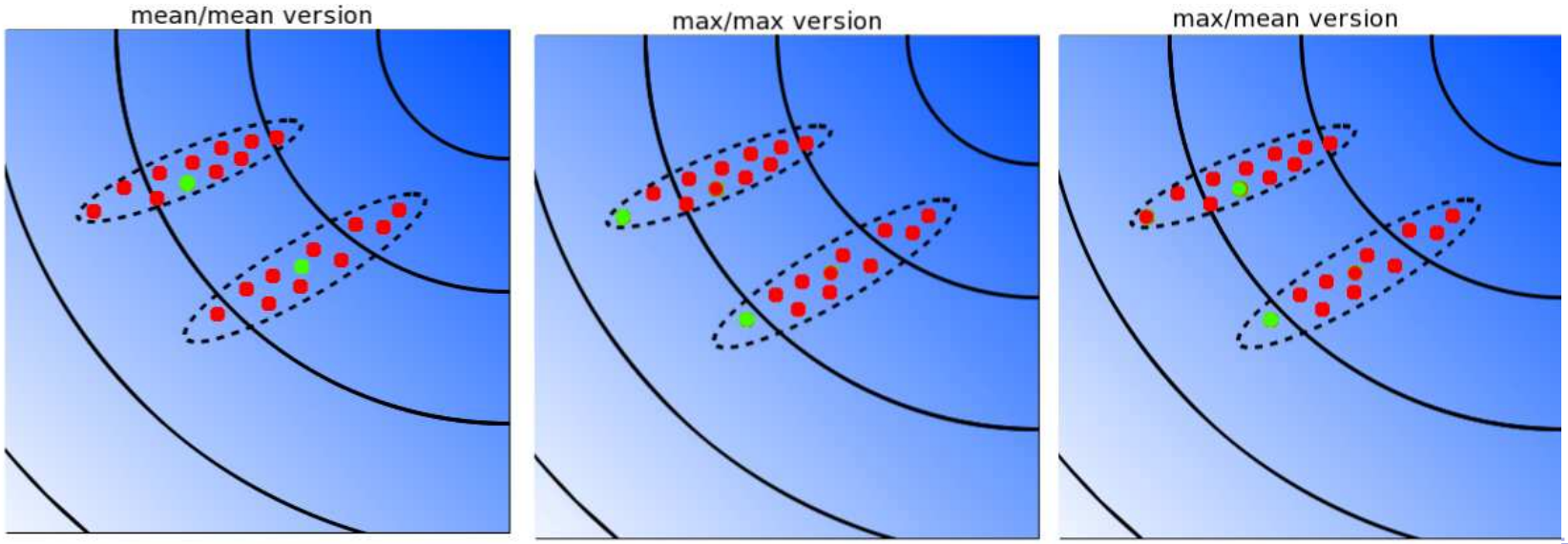
Outline

1 Derivative free optimization

2 Direct-search and evolution strategy
- Algorithmic changes
- Convergence
- Numerical results

Modified Evolution Strategies

- New point will be accepted if new population is good : whenever the decrease with the new generation is larger than a quantity $\rho(\sigma)$
- $t \mapsto \rho(t)$ taken positive, nondecreasing, and $\lim_{t \to 0} \frac{\rho(t)}{t} = 0$.
- How to measure decrease between 2 populations. Three variants (for which we can prove convergence): mean/mean, max/max, and mean/max



Modified Evolution Strategies

Algorithm

- if (version mean/mean)

$$f(x_{k+1}) \leq f(x_k) - \rho(\sigma_k),$$

- or (version max/max)

$$f(\tilde{y}_{k+1}^{m_\lambda}) \leq f(\tilde{y}_k^{m_\lambda}) - \rho(\sigma_k),$$

- or (version max/mean)

$$f(\tilde{y}_{k+1}^{m_\lambda}) \leq f(x_k) - \rho(\sigma_k),$$

- then the iteration is successfull ($\sigma_{k+1} = \max(\sigma_k, \sigma_k^{ES})$)
- Otherwise, $\sigma_{k+1} = \beta\sigma_k$ $(where\ \beta < 1)$.

Outline

1 Derivative free optimization

2 Direct-search and evolution strategy
- Algorithmic changes
- Convergence
- Numerical results

Convergence

Definition

For a function f, the Clarke generalized derivative

$$f^{\circ}(x_*; u) = \limsup_{x \to x_*, t \downarrow 0} \frac{f(x + t\,u) - f(x)}{t},$$

the point $x_*$is then Clarke stationary if $f^{\circ}(x_*; d) \geq 0, \forall d \in \mathbb{R}^n$.

- A Lipschitz continuous function is Clarke differentiable
- For Fréchet continously differentiable functions, the Clarke generalized derivative along d is the derivative along d

Convergence: Existence of a refining sequence

Theorem

Consider a sequence of iteration generated by our Algorithm without any stopping criterion. Let f be bounded below.
There exists a subsequence K of unsuccessful iterates for which $\lim_{k \in K} \sigma_k = 0$.
If $\{x_k\}$ is bounded, then there exists x_* and a K of unsuccessful iterates for which $\lim_{k \in K} \sigma_k = 0$ *and* $\lim_{k \in K} x_k = x_*$.

Convergence: Existence of a refining sequence

- Consider a sequence generated by the algorithm. Suppose $\exists \sigma > 0,\ \forall k, \sigma_k > \sigma$.
- There exists an infinite number of successful iterations (otherwise $\sigma_k \to 0$). In the mean/mean variant (idem for the max/max) $f(x_{k+1}) \leq f(x_k) - \rho(\sigma_k) \leq f(x_k) - \rho(\sigma)$.
- An infinite number of such steps cannot occur if f is bounded from below
- Take the unsuccessful iterations "just before" in the sequence

Convergence: Existence of a refining sequence

- Let $a_k = \sum_{i=1}^{m_\mu} \omega_k^i d_k^i$, $u_k = a_k / \|a_k\|$, u a limit point of (u_k). For $t_k = \sigma_k \|a_k\|$ we have $x_{k+1} = x_k + t_k u_k$.
- Then for unsuccessful iterations, using Lispschitz continuity of f

$$\begin{aligned}
\frac{f(x_k + t_k u) - f(x_k)}{t_k} &= \frac{f(x_k + t_k u_k) - f(x_k)}{t_k} \\
&- \frac{f(x_k + t_k u_k) - f(x_k + t_k u)}{t_k} \\
&\geq \frac{\rho(\sigma_k)}{\sigma_k \|a_k\|} - \gamma \frac{t_k \|u_k - u\|}{t_k}
\end{aligned}$$

- Taking the limsup yields a positive Clarke derivative along u.

Convergence

Theorem

Let x_* be a limit point of unsuccessful subsequence of iterates $\{x_k\}$ for which $\lim_{k\in K} \sigma_k = 0$. Assume that f is Lipschitz continuous near x_*.
If u is a limit point of $\{a_k/\|a_k\|\}_K$, then $f^\circ(x_*; u) \geq 0$.
if the set of limit points $\{a_k/\|a_k\|\}_K$ is dense on the unit sphere, then x_* is a Clarke stationary point.

Corollary

When f is continuously differentiable at x_*, we conclude that $\nabla f(x_*) = 0$.

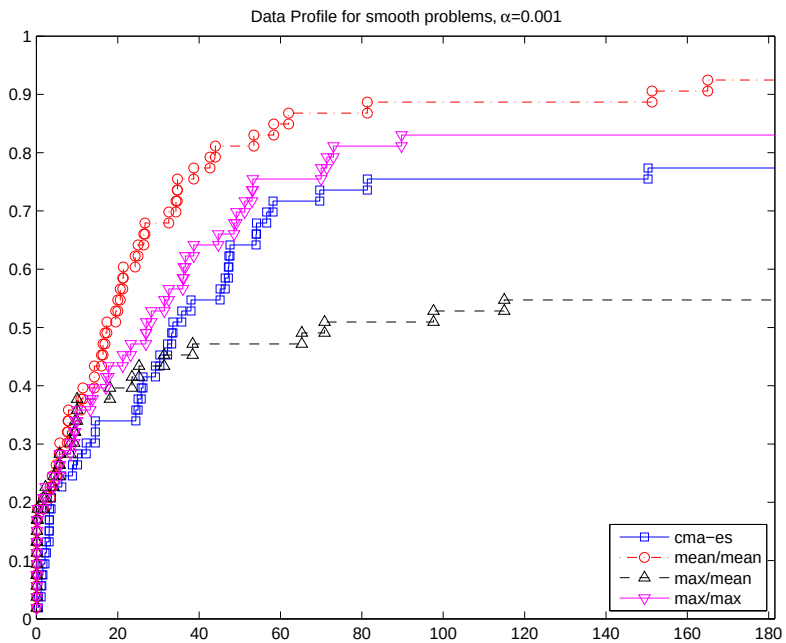
Outline

1. Derivative free optimization
2. Direct-search and evolution strategy
   - Algorithmic changes
   - Convergence
   - Numerical results

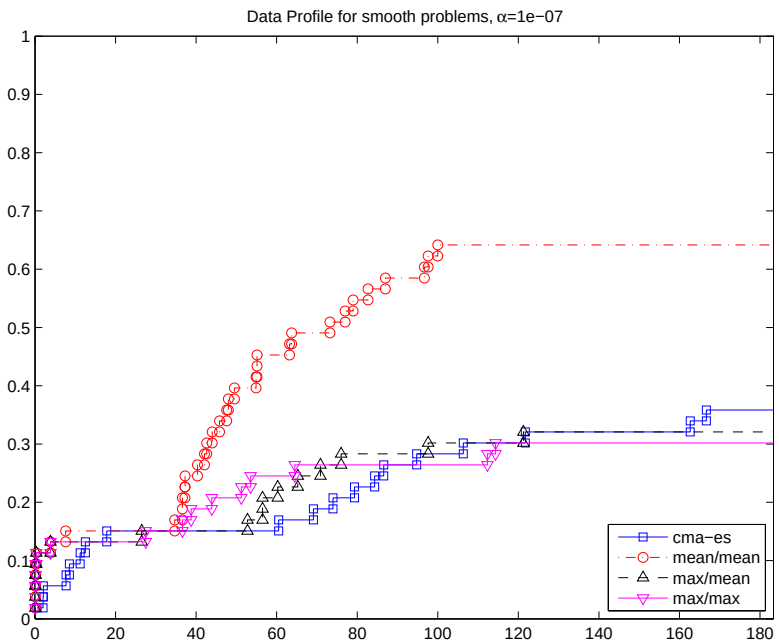
Test set

- 53 different problems taken form [Moré, Wild, 2009]
- Functions of the type "sum of squares", possibly with noise
- performance is compared using
  - data profiles: percentage of pb solved within a given function accuracy versus budget
  - performance profiles: percentage of pb solved within a given factor of the best solver
- Convergence test : $f(x_0) - f(x) \geq (1-\alpha)(f(x_0) - f^*)$

Data Profiles: Smooth Problems



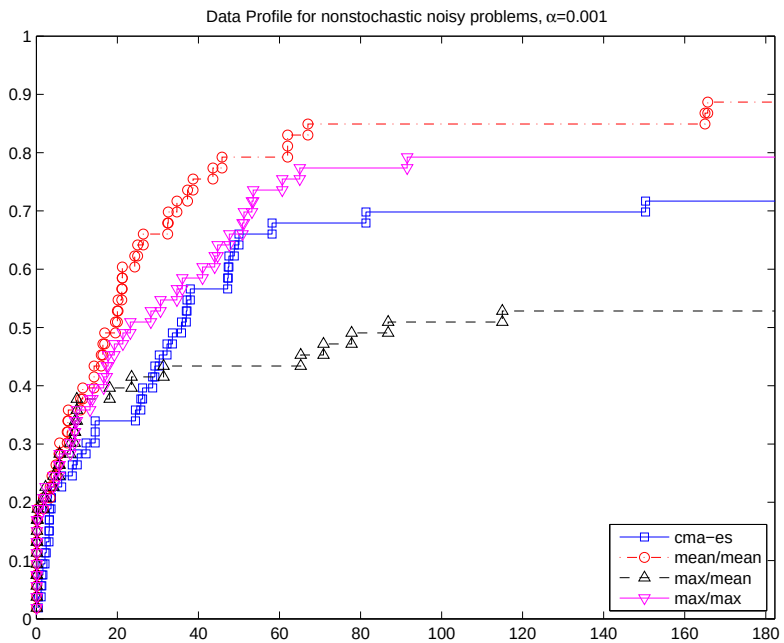
(a) Accuracy level of 10^{-3}.



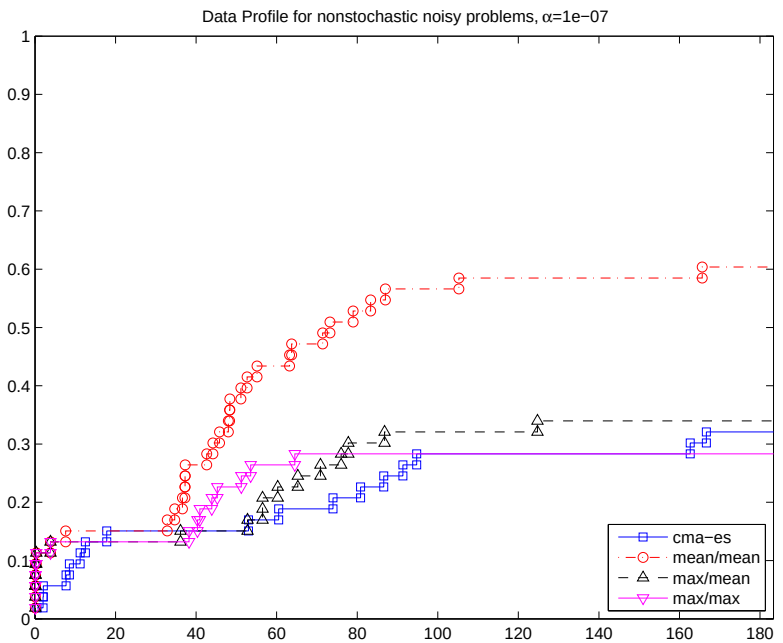
(b) Accuracy level of 10^{-7}.

Figure: Data profiles computed for the set of smooth problems, considering the two levels of accuracy, 10^{-3} and 10^{-7}.

Data Profiles: Noisy Problems



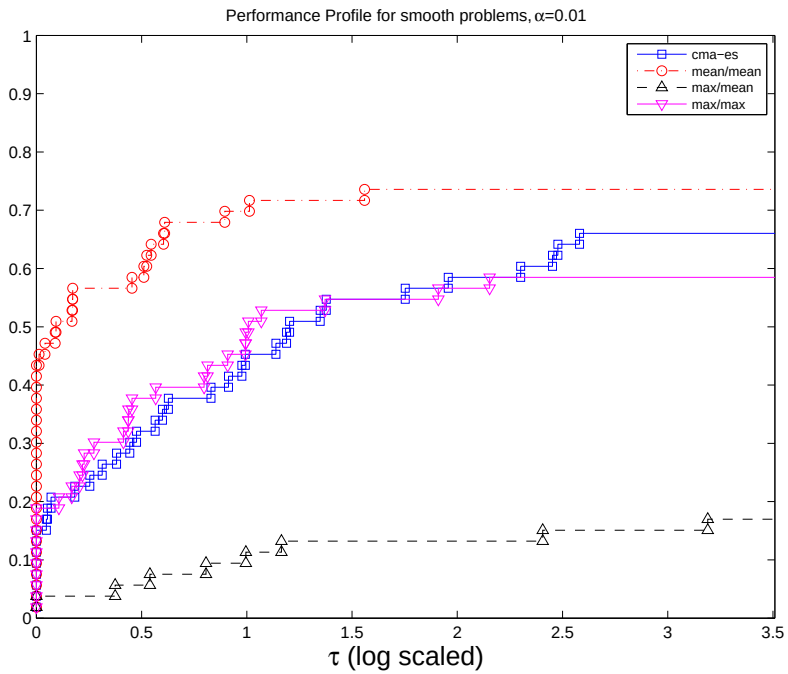
(a) Accuracy level of 10^{-3}.



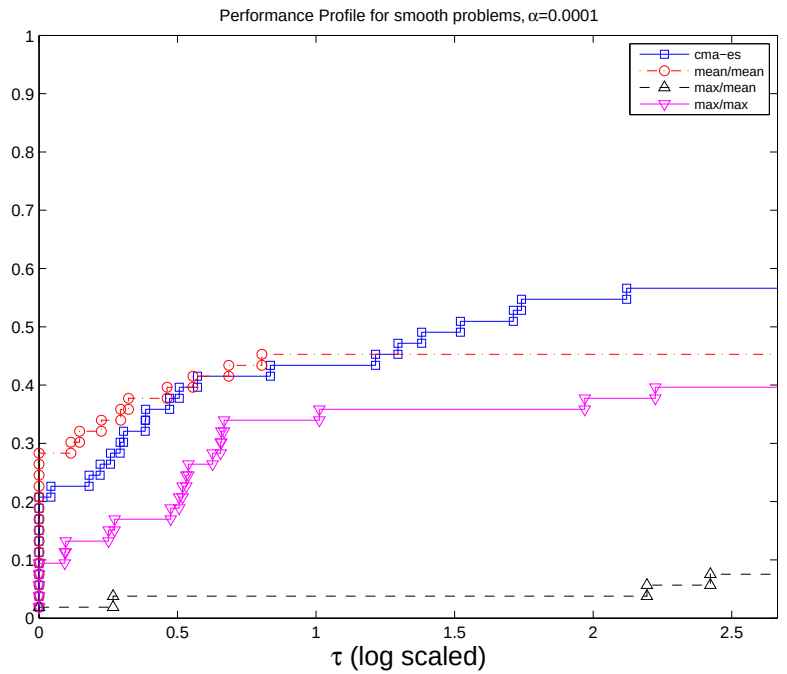
(b) Accuracy level of 10^{-7}.

Figure: Data profiles computed for the set of nonstochastic noisy problems, considering the two levels of accuracy, 10^{-3} and 10^{-7}.

Performance Profiles: Smooth Problems



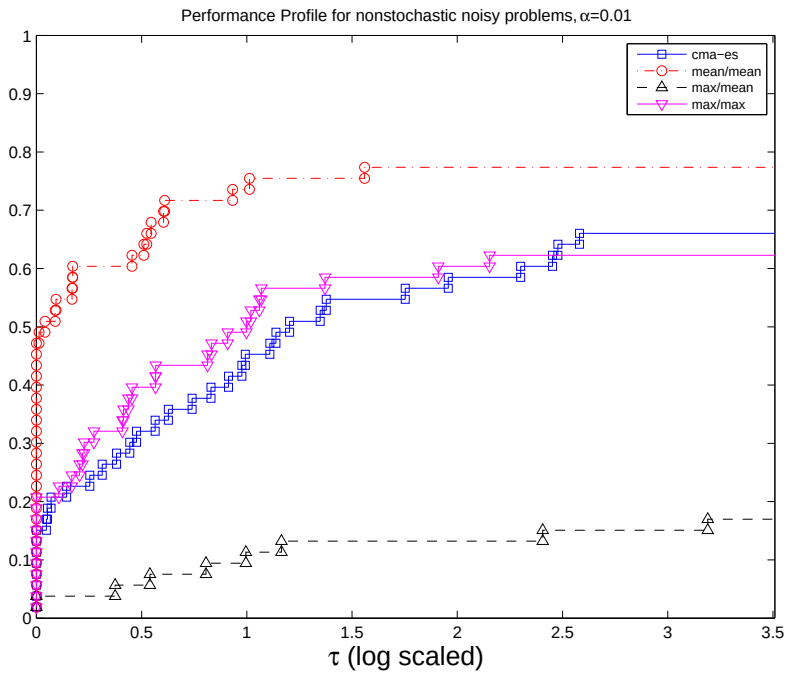
(a) Accuracy level of 10^{-2}.



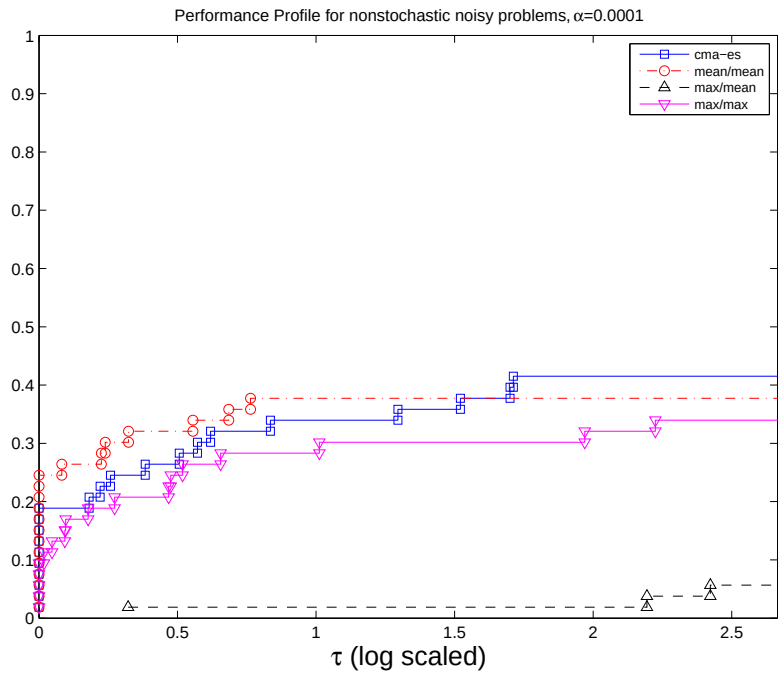
(b) Accuracy level of 10^{-4}.

Figure: Performance profiles computed for the set of smooth problems with a logarithmic scale, considering the two levels of accuracy, 10^{-2} and 10^{-4}.

Performance Profiles: Noisy Problems



(a) Accuracy level of 10^{-2}.



(b) Accuracy level of 10^{-4}.

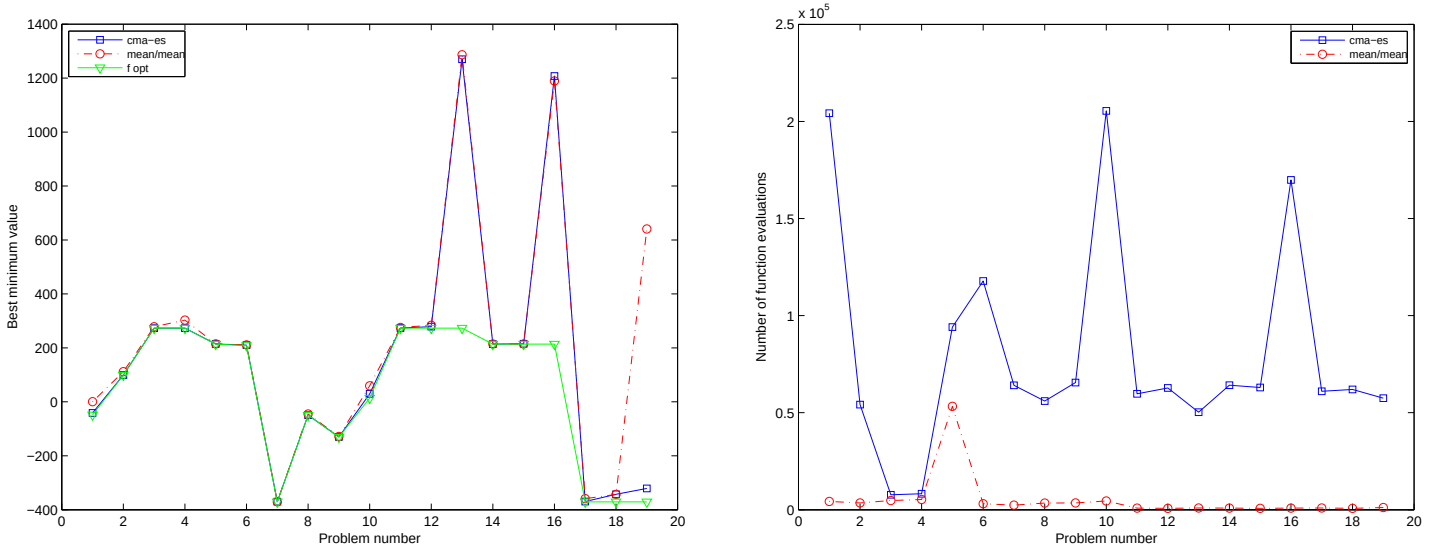
Figure: Performance profiles computed for the set of nonstochastic noisy problems with a logarithmic scale, considering the two levels of accuracy, 10^{-2} and 10^{-4}.

Having a reasonable global optimum

- Evolution algorithms are not supposed to be the best for quick convergence to local minima. They may provide more "global" solutions
- Our theory guarantees a convergence to points satisfying a first order stationarity
- Testing with a more demanding set of problems is needed. Experience on 18 multimodal problems from [Hansen, 2011], $n = 10, 20$.

Convergence Vs. Cost

Experience on 18 multimodal problems from [Hansen, 2011], $n = 10$



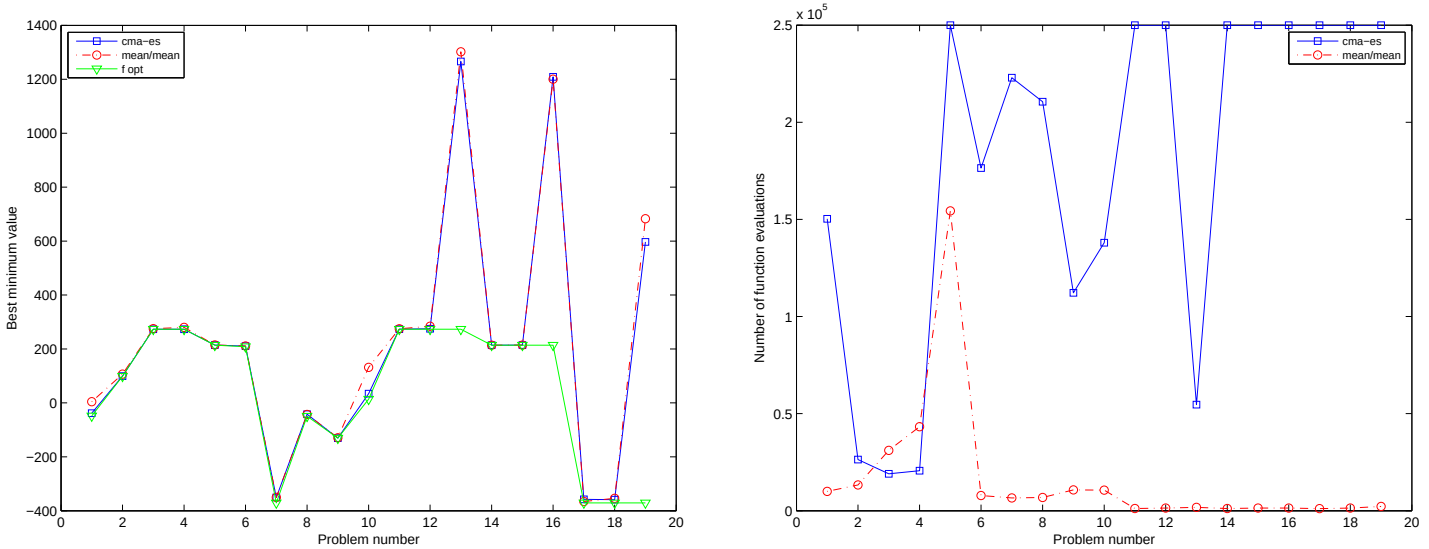
(a) Best function values (in average).

(b) Number of function evaluations taken (in average).

Figure: Results for the mean/mean version and CMA-ES on a set of multi-modal functions of dimension 10.

Convergence Vs. Cost

Experience on 18 multimodal problems from [Hansen, 2011], $n = 20$



(a) Best function values (in average).

(b) Number of function evaluations taken (in average).

Figure: Results for the mean/mean version and CMA-ES on a set of multi-modal functions of dimension 20.

Summary

- The algorithm proposed is fairly general, and incorporates decrease and step control in sampling techniques
- The most efficient decrease condition rejects points whenever the mean value of offsprings does not reduce the objective at the mean of the parents
- We are working in a generalization for constraints by minimizing the infinity barrrier function
- Our results rely dramatically on the sampling method, and the good performance of CMA-ES should be acknowledged !

Thank you for your attention !