Mirrored Sampling and Sequental Selection for Evolution Strategies

Dimo Brockhoff[1], Anne Auger[1], Nikolaus Hansen[1], Dirk V. Arnold[2], and Tim Hohm[3]

Motivation

BBOB'2009

- (1+1)-CMA surprisingly good on some functions
- even outperforms BIPOP-CMA-ES on Gallagher functions by factor of about 10 (in 20D)
- but elitism bad for noise

Derandomization

- can improve convergence of ES
- but might introduce certain bias towards smaller step sizes

Question: How to design fast "local" non-elitist ES?
with derandomization
without bias

Conclusions

Proposed:

Mirroring and sequential selection are two independent ways to improve local non-elitist ESs

Results:

- Improved convergence on sphere function
- Implementation within CMA-ES results in better performance on some BBOB'10 functions
- $(1,4^s_m)$-ES even faster than (1+1)-ES on sphere w.r.t. the theoretical convergence rates: 0.202 (1+1)-ES ≤ 0.223 $(1,4^s_m)$-ES ≤ 0.235 $(1+1^s_m)$-ES

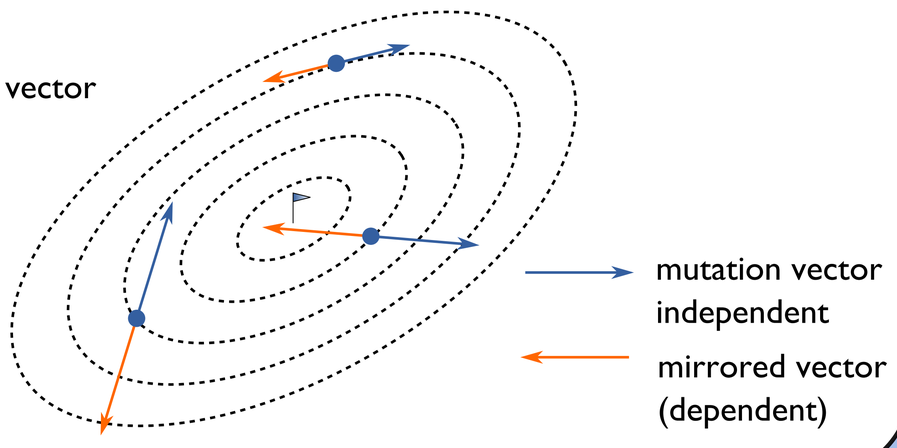
Idea 1: Mirrored Mutation

Idea

use one random vector to generate two offspring

Reasoning

often "good" and "bad" in opposite directions



Idea 2: Sequential Selection

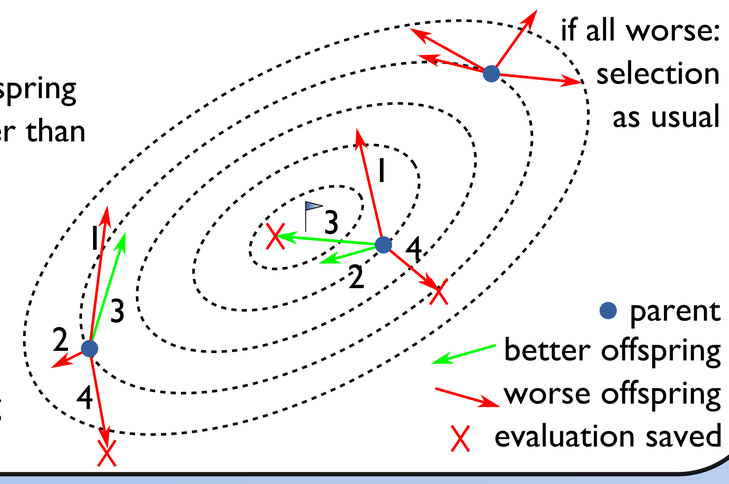
Idea

stop generation of new offspring as soon as a solution, better than the parent, is found

Reasoning

if sublevel sets convex one better is enough

in particular with mirroring



Resulting Variants of the (1,λ)-ES

given: $\boldsymbol{X}_k \in \mathbb{R}^d, \sigma_k \in \mathbb{R}_{>0}, j \in \mathbb{N}, \lambda \in \mathbb{N}^+, f : \mathbb{R}^d \to \mathbb{R}$
$i \leftarrow 0$
while $i < \lambda$ **do**
  $i \leftarrow i + 1, j \leftarrow j + 1$
  if *mirrored sampling* and $j \equiv 0 \pmod 2$ **then**
    $\boldsymbol{X}_k^i = \boldsymbol{X}_k - \sigma_k \boldsymbol{\mathcal{N}}_k^{i-1}$ use previous sample
  else
    $\boldsymbol{X}_k^i = \boldsymbol{X}_k + \sigma_k \boldsymbol{\mathcal{N}}_k^i$
  if *sequential selection* and $f(\boldsymbol{X}_k^i) < f(\boldsymbol{X}_k)$ **then**
    $j \leftarrow 0$ start with a new sample in the next iteration
    break;
end while
return $\boldsymbol{X}_{k+1} = \operatorname{argmin}\{f(\boldsymbol{X}_k^1), \ldots, f(\boldsymbol{X}_k^i)\}$

mirroring and sequentialism are independent

standard: (1,λ)-ES

mirroring only: $(1,\lambda_m)$-ES

sequentialism only: $(1,\lambda^s)$-ES

mirroring and sequentialism: $(1,\lambda^s_m)$-ES

Some Algorithms' Properties

If f has convex sublevel sets:

never both mirrored offspring are better than parent

⇒ accepted parent after one iteration of $(1,2_m)$-ES and $(1,2^s_m)$-ES are the same (assuming the same random numbers are used)

⇒ $(1,2^s_m)$-ES converges faster than $(1,2_m)$-ES on those functions (and diverges faster as well)

$(1,\infty^s)$-ES = (1+1)-ES if σ_k scale-invariant or constant

[proofs submitted to FOGA]

Theoretical Convergence Rates of (1,λ)-ES Variants

Idea

explaining theoretically what can be gained by mirrored mutation and sequential selection in terms of convergence rate

Scenario

- scale-invariant (1,λ)-ES variants: $\sigma_k = \sigma \, \| X_k \|$: gives optimal convergence rate among all step size adaptive (1,λ)-ES for σ well chosen
- sphere function: f(x) = g(||x||) (g: monotonically increasing)

Linear Convergence

$$\frac{1}{T_k} \ln \frac{\|\boldsymbol{X}_k\|}{\|\boldsymbol{X}_0\|} \to c \;\; a.s.$$

T_k: number of evaluations until iteration k; c: convergence rate

Results

Theorem 4. *For a $(1, 2^s_m)$-ES with scale-invariant step-size ($\sigma_k = \sigma\|\boldsymbol{X}_k\| > 0$) on the sphere function $g(\|\boldsymbol{x}\|)$, for $g \in \mathcal{M}$, linear convergence holds and*

$$\frac{1}{T_k} \ln \frac{\|\boldsymbol{X}_k\|}{\|\boldsymbol{X}_0\|} \xrightarrow[k\to\infty]{} \frac{1}{2} \frac{1}{2 - p_s(\sigma)} \times E\left[\ln\left(1 - 2\sigma|[\mathcal{N}]_1| + \sigma^2\|\mathcal{N}\|^2\right)\right] \; a.s.$$

exemplary

where T_k is the random variable for the number of function evaluations until iteration k, $\mathcal{N}$ is a random vector following a multivariate normal distribution, and $p_s(\sigma) = \Pr([\mathcal{N}]_1 + \sigma\|\mathcal{N}\|^2 < 0)$ is the probability that the first offspring is successful.

Proof Ideas

$$\frac{1}{T_k} \ln \frac{\|\boldsymbol{X}_k\|}{\|\boldsymbol{X}_0\|} = \frac{1}{T_k} \sum_{i=0}^{k-1} \underbrace{\ln \frac{\|\boldsymbol{X}_{i+1}\|}{\|\boldsymbol{X}_i\|}}$$

independent and identically distributed:
isotropy of sphere function and normal distribution
+ scale-invariant step size rule
LLN shows that RHS converges to $E(\ln(\|X_{i+1}\|/\|X_i\|))$

Numerical Results

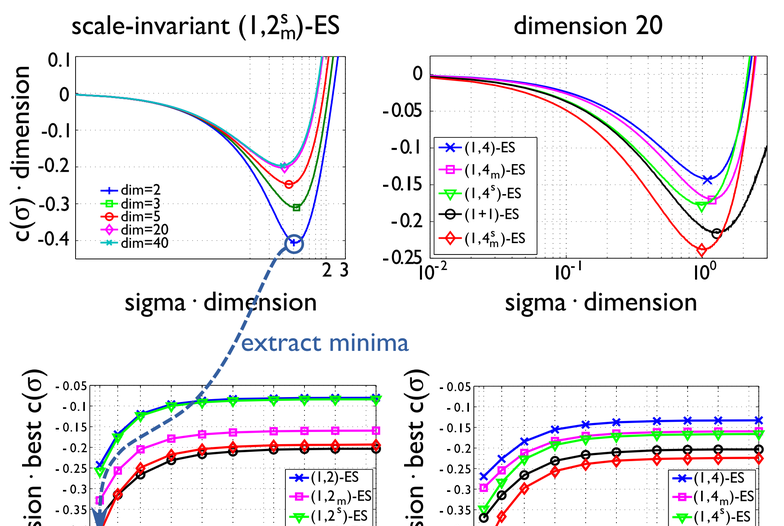
Idea

Estimate convergence rates by exploiting theoretical results via Monte Carlo sampling

Procedure

1) Exploitation of theoretical expression of convergence rate: perform Monte Carlo estimate of probability of success and expected value for different σ ($0.01 \le \sigma d \le 3$); averaged over 10^6 samples
2) extract smallest value (corresponds to optimal conv. rate though bias towards smaller values)

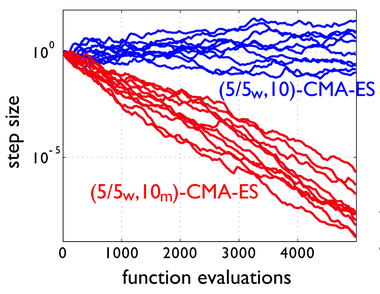
Simulated Convergence Rates

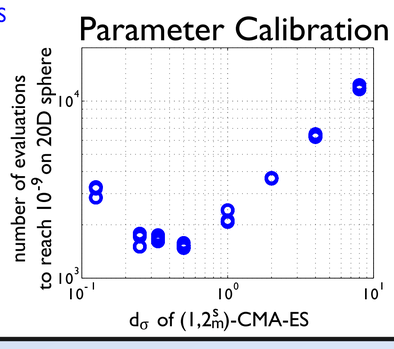


Application to the CMA-ES

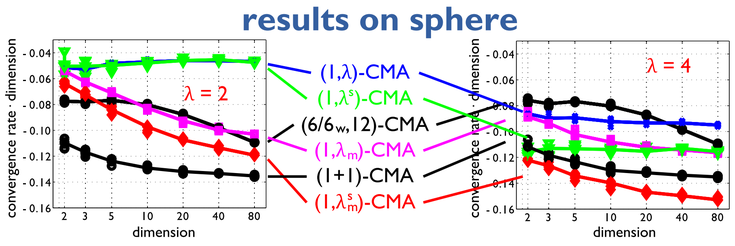
implementation straight forward,
but with bias if μ>1

⇒ only μ=1 here



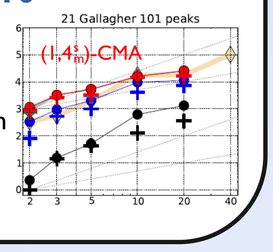


results on sphere



results on BBOB'2010

- $(1,4^s_m)$-CMA-ES turned out to be fastest local non-elitist strategy tested
- 3rd best of BBOB'2009/10 on Gallagher with 101 peaks (3x faster than (1+1)-CMA-ES)
- even more competitive on noisy functions



[1] INRIA Saclay Ile-de-France, TAO, LRI, Paris Sud University, Orsay, France, firstname.lastname@inria.fr
[2] Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada, dirk@cs.dal.ca
[3] Department of Medical Genetics, University of Lausanne, Lausanne, Switzerland tim.hohm@unil.ch