Benchmarking IPOP-CMA-ES-TPA and IPOP-CMA-ES-MSR on the BBOB Noiseless Testbed

Asma Atamna
TAO Team
Inria - LRI - University of Paris-Sud

Black Box Optimization Benchmarking 2015

GECCO 2015, Madrid

July 23, 2015

Outline

1. Motivation
2. Testbed Overview
3. Algorithms
   - The $(\mu/\mu, \lambda)$-ES
   - CMA-ES-TPA
   - CMA-ES-MSR
   - CMA-ES-CSA
4. Results
   - Experimental Setting
   - ERT vs. Dimension
   - ECDFs
5. Discussion

Motivation

- Benchmark two relatively new step-size adaptation mechanisms on the BBOB noiseless testbed
  - Covariance Matrix Adaptation Evolution Strategy with Two-Point step-size Adaptation (CMA-ES-TPA)
  - Covariance Matrix Adaptation Evolution Strategy with Median Success Rule (CMA-ES-MSR)
- Restarts
  - IPOP: Increase POPulation size by a factor of 2
- Compare CMA-ES-TPA and CMA-ES-MSR to state-of-the-art CMA-ES-CSA (Covariance Matrix Adaptation Evolution Strategy with Cumulative Step-size Adaptation)

Testbed Overview

24 functions, 5 categories

- Separable: sphere, separable ellipsoid, separable Rastrigin, Bueche-Ratrigin, linear slope
- Moderate: attractive sector, step-ellipsoid, Rosenbrock, rotated Rosenbrock
- Ill-conditioned: ellipsoid, discus, bent cigar, sharp ridge, sum of different powers
- Multi-modal: Rastrigin, Weierstrass, Schaffer F7 with condition 10, Schaffer F7 with condition 1000, Griewank-Rosenbrock F8F2
- Weakly structured multi-modal: Schwefel, Gallagher 101 peaks, Gallagher 21 peaks, Katsuuras, Lunacek bi-Rastrigin

The $(\mu/\mu, \lambda)$-ES

λ: population size
μ: number of parents
',': non-elitist selection
At time step t

- Sample λ offspring, $\mathbf{X}_t^1, \cdots, \mathbf{X}_t^\lambda$ according to

$$\mathbf{X}_t^i = \mathbf{X}_t + \sigma_t \mathcal{N}_t^i(0, \mathbf{C}_t)$$

  $i = 1, \cdots, \lambda$, $\mathcal{N}_t^i(0, \mathbf{C}_t)$ is the multivariate normal distribution with mean 0 and covariance matrix $\mathbf{C}_t$, σ_t is the step-size

- Select the μ best offspring fitness-wise and recombine them according to

$$\mathbf{X}_{t+1} = \sum_{i=1}^{\mu} w_i \mathbf{X}_t^{i:\lambda}$$

  $\mathbf{X}_t^{i:\lambda}$ is the ith best offspring, w_i are positive weights

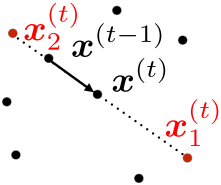
- Update σ_t, $\mathbf{C}_t$

CMA-ES-TPA

- Sample the first two offspring, $\mathbf{X}_t^1$ and $\mathbf{X}_t^2$, along the shift vector from $\mathbf{X}_{t-1}$ to $\mathbf{X}_t$, as a mirrored pair, symmetric to $\mathbf{X}_t$, according to

$$\mathbf{X}_t^{1,2} = \mathbf{X}_t \pm \sigma_t \|\mathcal{N}_t(0, \mathbf{I})\| \frac{\mathbf{X}_t - \mathbf{X}_{t-1}}{\|\mathbf{X}_t - \mathbf{X}_{t-1}\|}$$

$\mathbf{I}$: identity matrix



CMA-ES-TPA

- If $\mathbf{X}_t^1$ is better than $\mathbf{X}_t^2$, increase σ_t. Otherwise, decrease σ_t

$$s_t = (1 - c_\sigma)s_{t-1} + c_\sigma \frac{\text{rank}(\mathbf{X}_t^2) - \text{rank}(\mathbf{X}_t^1)}{\lambda - 1}$$

$$\sigma_{t+1} = \sigma_t \exp\left(\frac{s_t}{d_\sigma}\right)$$

$s_0 = 0$, $c_\sigma = 0.3$, $d_\sigma = \sqrt{D}$, D is the dimension of the search space

CMA-ES-MSR

- Generalization of the 1/5th success rule to the case of $(\mu/\mu, \lambda)$-ES
- Success: the median offspring of the current population, $\mathbf{X}_t^{m(\lambda)}$, is better than the jth best individual of the previous population, $\mathbf{X}_{t-1}^{j:\lambda}$
- j: chosen such that the median success probability is 1/2 on the sphere with optimal step-size
- If $\mathbf{X}_t^{m(\lambda)}$ is better than $\mathbf{X}_{t-1}^{j:\lambda}$, increase σ_t. Otherwise, decrease σ_t

$$s_t = (1 - c_\sigma)s_{t-1} + c_\sigma \frac{2}{\lambda} \left(K_{\text{succ}} - \frac{\lambda}{2} \right)$$

$$\sigma_{t+1} = \sigma_t \exp\left(\frac{s_t}{d_\sigma}\right)$$

K_{succ} is the number of successful offspring, $s_0 = 0$, $c_\sigma = 0.3$, $d_\sigma = 2 - 2/D$

CMA-ES-CSA

- Record the path (consecutive steps) taken by the algorithm

$$\boldsymbol{p}_{t+1} = (1 - c_\sigma)\boldsymbol{p}_t + \sqrt{c_\sigma(2 - c_\sigma)/\sum_{k=1}^{\mu} w_k^2} \sum_{k=1}^{\mu} w_k \mathbf{Z}_t^{k:\lambda}$$

  $\mathbf{Z}_t^{k:\lambda}$: step corresponding to the kth best offspring,
  $\mathbf{X}_t^{k:\lambda} = \mathbf{X}_t + \sigma_t \mathbf{Z}_t^{k:\lambda}$

- If $\boldsymbol{p}_t$ is "too long", increase σ_t. Otherwise, decrease σ_t

$$\sigma_{t+1} = \sigma_t \exp^{c_\sigma/d}\left(\frac{\|\boldsymbol{p}_{t+1}\|}{\mathbb{E}\|\mathcal{N}(0, \boldsymbol{I})\|} - 1\right)$$

$$c_\sigma = \frac{1/\sum_{k=\mathbf{1}}^{\mu} w_k^{\mathbf{2}} + 2}{1/\sum_{k=\mathbf{1}}^{\mu} w_k^{\mathbf{2}} + D + 3}$$

Experimental Setting

- Maximum budget: $10^5 \times D$
- Tested dimensions: 2, 3, 5, 10, 20, 40
- Maximum number of restarts: 9
- $\mathbf{X}_0 \sim \mathcal{U}[-4, 4]^D$
- $\sigma_0 = 2.5$
- Source code: cma 1.1.06
- Default parameters
  - $\lambda = 4 + \lfloor 3 \ln D \rfloor$
  - $\mu = \lambda/2$

ERT vs. Dimension

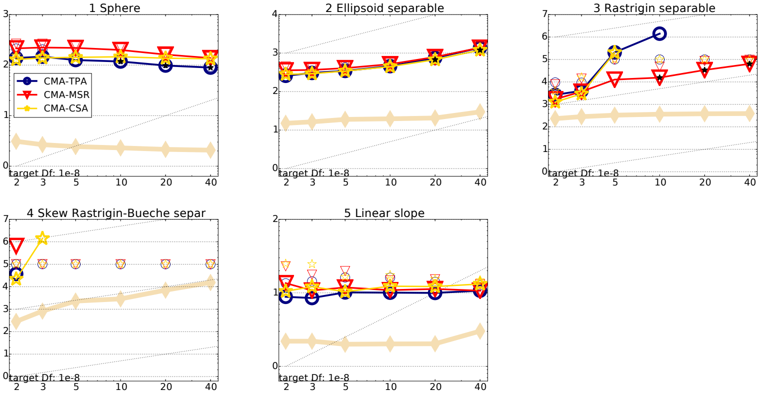
- Expected Running Time

$$\mathrm{ERT}(f_{\mathrm{target}}) = \frac{\#\mathrm{FEs}(f_{\mathrm{best}} \geq f_{\mathrm{target}})}{\#\mathrm{succ}}$$

- $f_{\mathrm{target}} = f_{\mathrm{opt}} + \Delta f$
- $\Delta f = 10^{-8}$

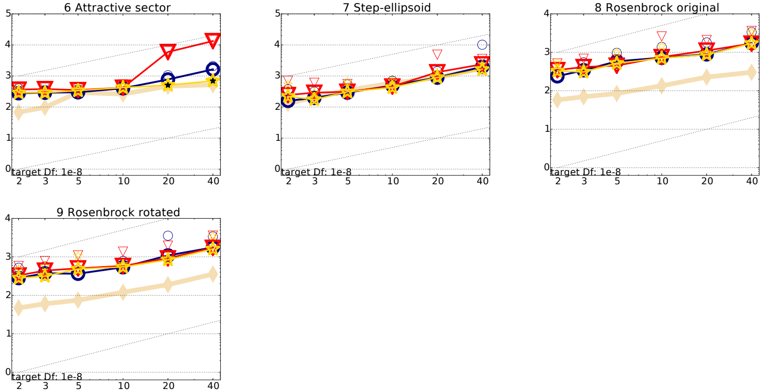
ERT vs. Dimension

Separable functions



ERT vs. Dimension

Moderate functions



ERT vs. Dimension

Single runs of CMA-MSR, CMA-TPA, and CMA-CSA on the attractive sector in 20D

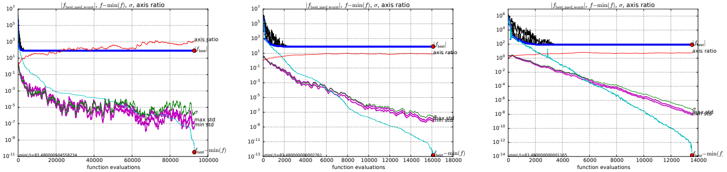
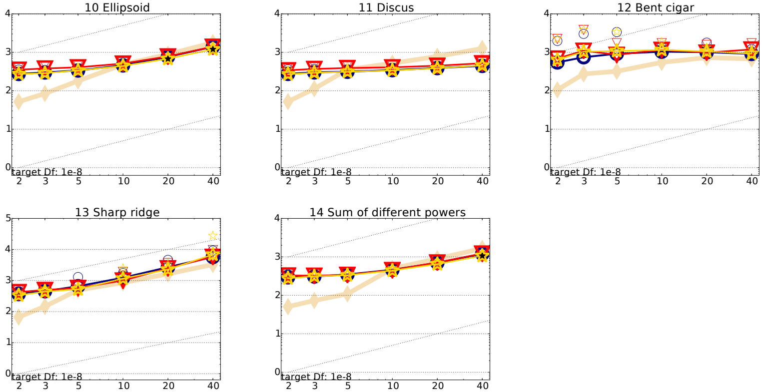


Figure: Left: CMA-ES-MSR, middle: CMA-ES-TPA, right: CMA-ES-CSA

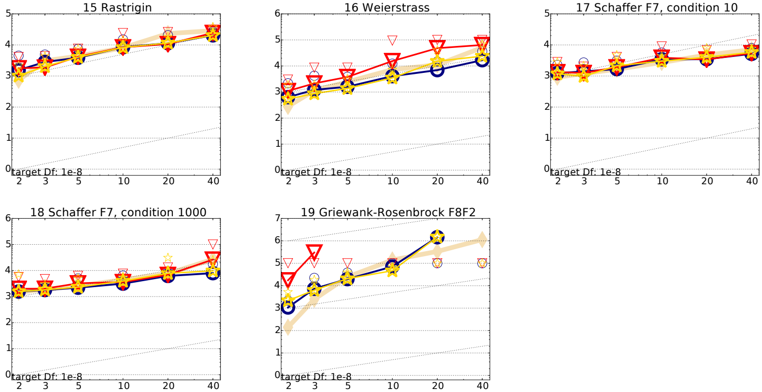
ERT vs. Dimension

Ill-conditioned functions



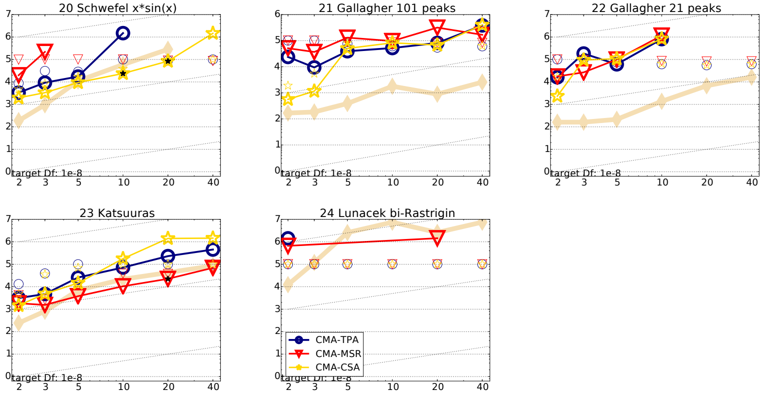
ERT vs. Dimension

Multi-modal functions



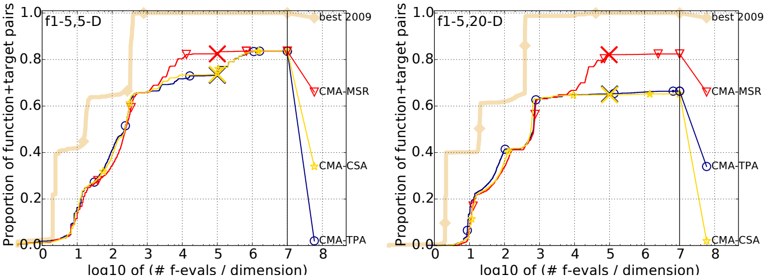
ERT vs. Dimension

Weakly structured multi-modal functions



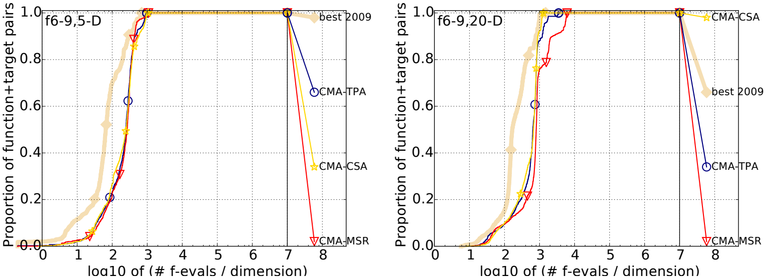
ECDFs

Separable functions



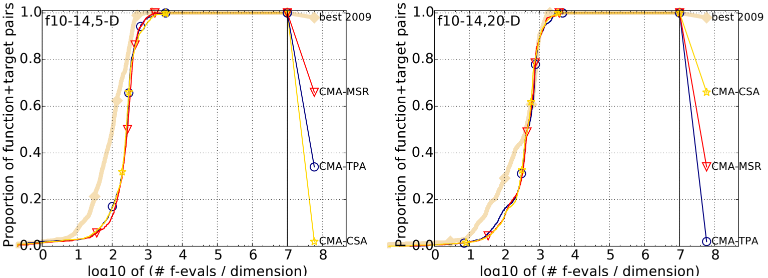
ECDFs

Moderate functions



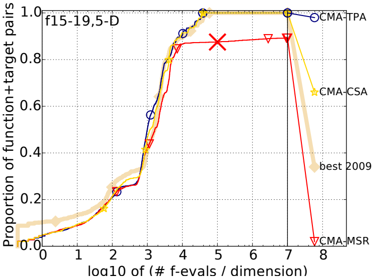
ECDFs

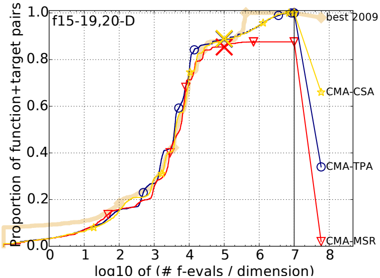
Ill-conditioned functions



ECDFs

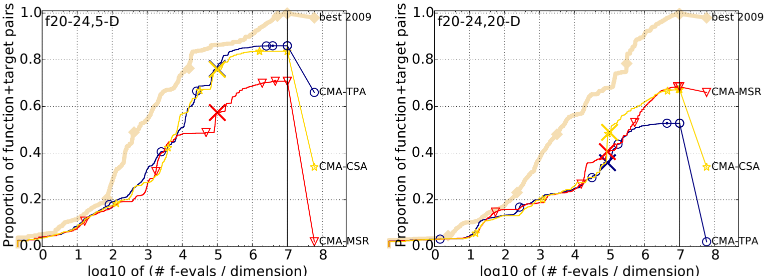
Multi-modal functions





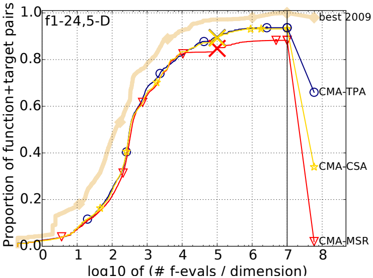
ECDFs

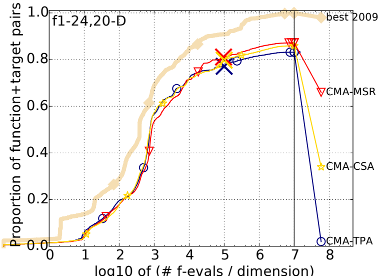
Weakly structured multi-modal functions



ECDFs

All functions





Discussion

- As expected by design, the three algorithms have comparable performance on most of the functions
- However, significant differences were observed: attractive sector, separable Rastrigin
- The influence of the step-size is more important on multi-modal and weakly structured multi-modal functions
- CMA-ES-CSA and CMA-ES-TPA have similar behavior
- Different performance on separable and rotated Rastrigin: different functions

Thank You