PSO facing non-separable and ill-conditioned problems

Anne Auger Nikolaus Hansen Nikolas Mauny
Marc Schoenauer

TAO Team, INRIA Futurs, FRANCE
http://tao.lri.fr
{Anne.Auger@inria.fr}

Séminaire OEP, Avril 2007

1. Continuous optimization and stochastic search
   - Non-separable problems
   - Ill-Conditioned Problems
   - PSO and CMA-ES
2. PSO facing ill-conditioned and non-separable problems
   - Test problems
   - Results
3. Summary and conclusion

Problem Statement
Continuous Domain Search/Optimization

- Minimize a fitness function (*objective* function, *loss* function) in continuous domain

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

- Black Box scenario (direct search)

Problem Statement
Continuous Domain Search/Optimization

- Minimize a fitness function (*objective* function, *loss* function) in continuous domain

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

- Black Box scenario (direct search)

Problem Statement
Continuous Domain Search/Optimization

- Minimize a fitness function (*objective* function, *loss* function) in continuous domain

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

- Black Box scenario (direct search)

  - gradients are not available or not useful
  - problem domain specific knowledge is used only within the black box
  - search costs: number of function evaluations
  - goal (1): find a solution $\boldsymbol{x}$ with a function value as small as possible with as small as possible search costs
  - goal (2): convergence to an optimum
  - goal (3): convergence to the global optimum

Problem Statement
Continuous Domain Search/Optimization

- Minimize a fitness function (*objective* function, *loss* function) in continuous domain

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

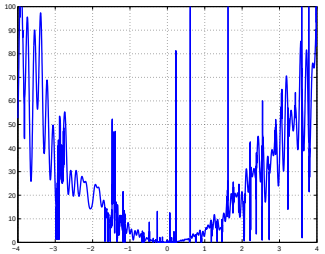
- Black Box scenario (direct search)

- Typical Examples:
  - shape optimization (e.g. using CFD) — curve fitting, airfoils
  - model calibration — biological, physical
  - parameter calibration — controller, plants, images

What makes a problem hard?
Fitness function properties - Why stochastic search?

- ruggedness

  non-smooth, discontinuous, multimodal, and/or noisy function
- dimensionality (considerably) larger than three
- non-separability
- ill-conditioning

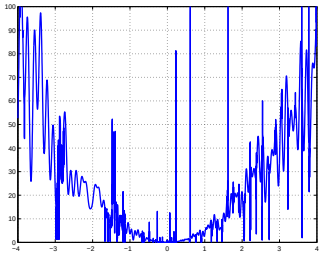


cut from 3-D example solvable with an ES

What makes a problem hard?
Fitness function properties - Why stochastic search?

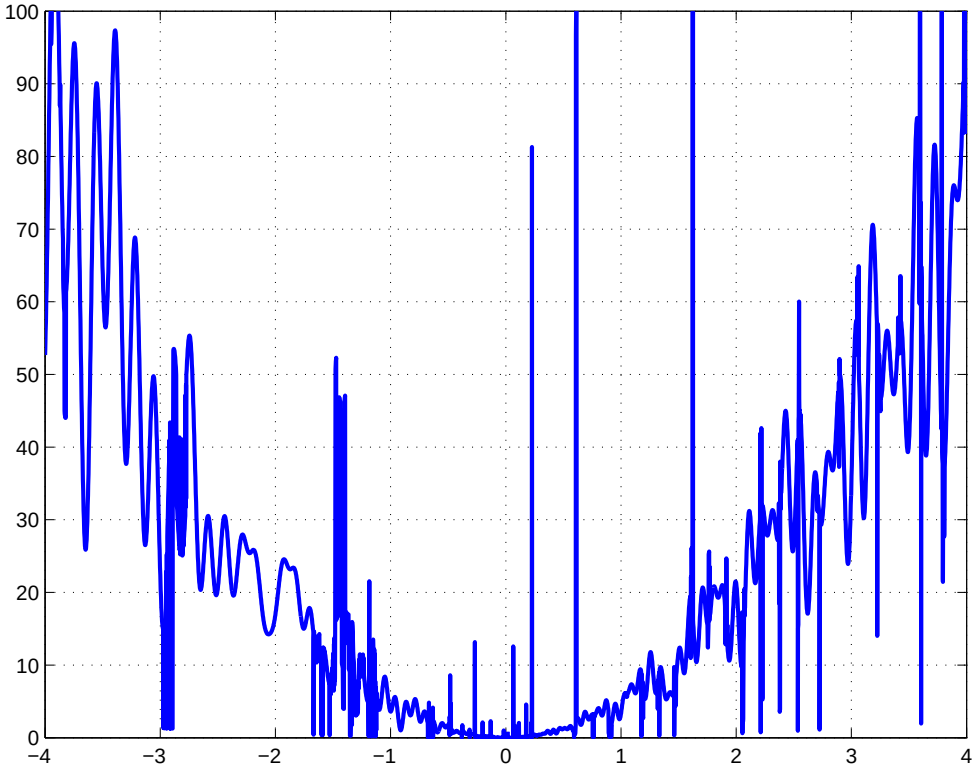
- ruggedness

  non-smooth, discontinuous, multimodal, and/or noisy function
- dimensionality (considerably) larger than three
- non-separability
- ill-conditioning



cut from 3-D example solvable with an ES

What makes a problem hard?
Fitness function properties - Why stochastic search?



cut from 3-D example solvable with an ES

Separable Problems

Definition (Separable Problem)

A function f is separable if

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

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

Separable Problems

Definition (Separable Problem)

A function f is separable if

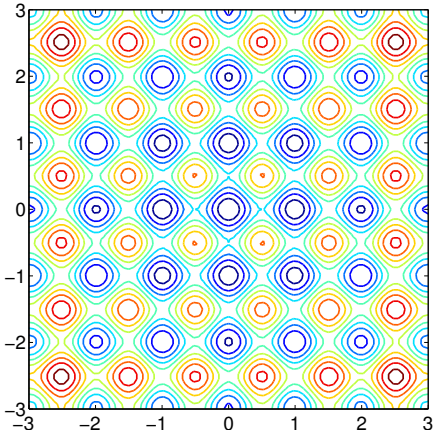
$$\left(\arg\min_{x_1} f(x_1, \ldots), \ldots, \arg\min_{x_n} f(\ldots, x_n)\right) = \arg\min_{(x_1,\ldots,x_n)} f(x_1, \ldots, x_n)$$

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

Example: Additively decomposable functions

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

Rastrigin function

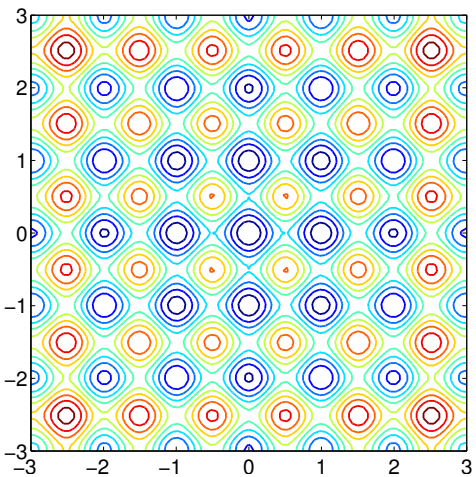


Non-Separable Problems
Building a non-separable problem from a separable one

Rotating the coordinate system

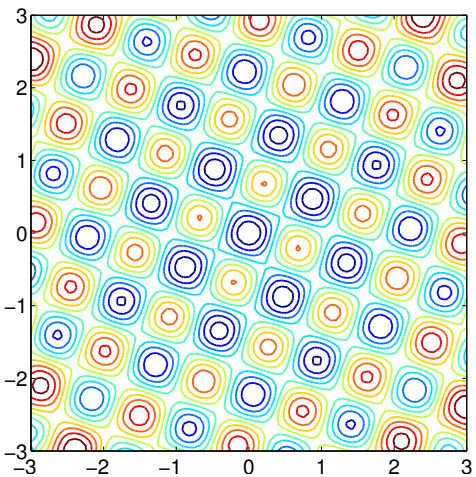
- $f : \boldsymbol{x} \mapsto f(\boldsymbol{x})$ separable
- $f : \boldsymbol{x} \mapsto f(\boldsymbol{R}\boldsymbol{x})$ non-separable

$\boldsymbol{R}$ rotation matrix



$\boldsymbol{R}$

$\longrightarrow$



[1] 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

[2] 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

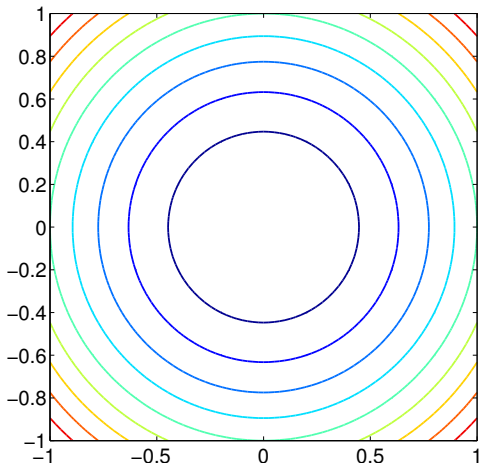
Ill-Conditioned Problems

If f is quadratic, $f : \boldsymbol{x} \mapsto \boldsymbol{x}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{x}$, ill-conditioned means a high condition number of Hessian Matrix $\boldsymbol{H}$

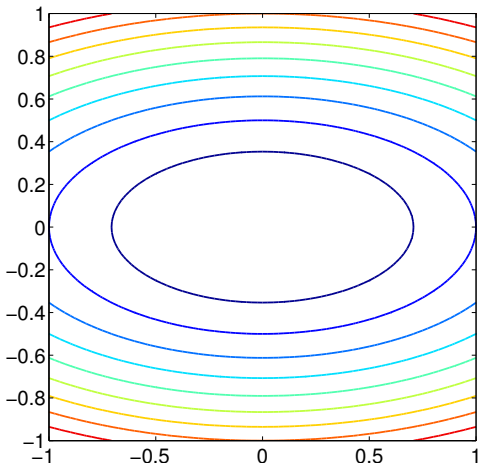
Ill-Conditioned Problems

If f is quadratic, $f : \boldsymbol{x} \mapsto \boldsymbol{x}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{x}$, ill-conditioned means a high condition number of Hessian Matrix $\boldsymbol{H}$

ill-conditioned means “squeezed” lines of equal function value



Increased
$\longrightarrow$
condition
number



consider the curvature of iso-fitness lines

Black Box Stochastic Optimization Algorithms

Template to minimize $f : \mathbb{R}^n \to \mathbb{R}$

Initialize distribution parameters $\boldsymbol{\theta}$, set population size $\lambda \in \mathbb{N}$
While not terminate

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

Examples

- Particle Swarm Optimization (PSO)
  Kennedy, Eberhart, 1995
- Covariance Matrix Adaptation Evolution Strategy (CMA-ES)
  Hansen, Ostermeier, 2001

Black Box Stochastic Optimization Algorithms

Template to minimize $f : \mathbb{R}^n \to \mathbb{R}$

Initialize distribution parameters $\boldsymbol{\theta}$, set population size $\lambda \in \mathbb{N}$
While not terminate

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

Examples

- Particle Swarm Optimization (PSO)
  Kennedy, Eberhart, 1995
- Covariance Matrix Adaptation Evolution Strategy (CMA-ES)
  Hansen, Ostermeier, 2001

Particle Swarm Optimization (PSO)

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\lambda \in \mathbb{R}^n$ be a set of particles

1. Sample new position

   For every particle index $i \in \{1, \ldots, \lambda\}$ and every coordinate $j \in \{1, \ldots, n\}$

$$x_i^j(t+1) = x_i^j(t) + w \times \underbrace{\left(x_i^j(t) - x_i^j(t-1)\right)}_{\text{previous step, momentum}} + \underbrace{c_1\, \mathcal{U}_i^j(0,1)(p_i^j - x_i^j(t))}_{\text{approach the "previous" best}} + \underbrace{c_2\, \tilde{\mathcal{U}}_i^j(0,1)(g_i^j - x_i^j(t))}_{\text{approach the "global" best}}$$

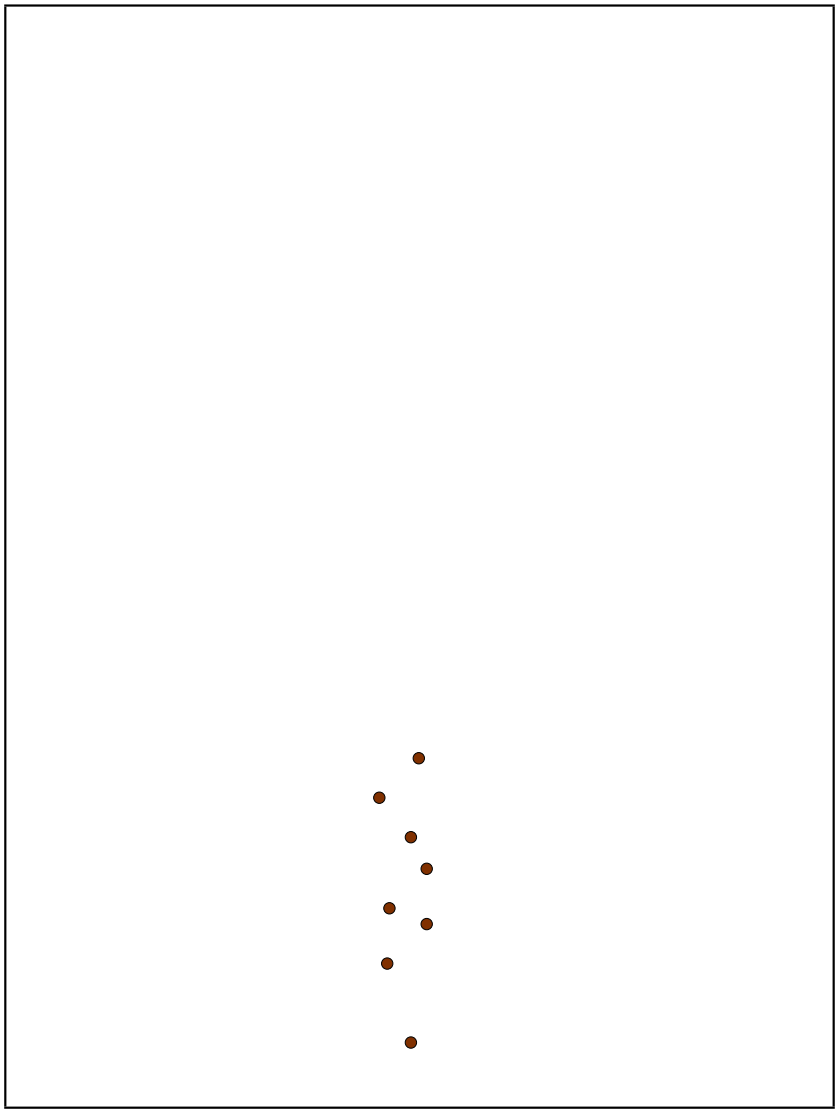
   inertia weight $w \approx 0.7$, and coefficients $c_1 = c_2 \approx 1.2$

2. Evaluate $\boldsymbol{x}_1(t+1), \ldots, \boldsymbol{x}_\lambda(t+1)$
3. Update parameter of distribution

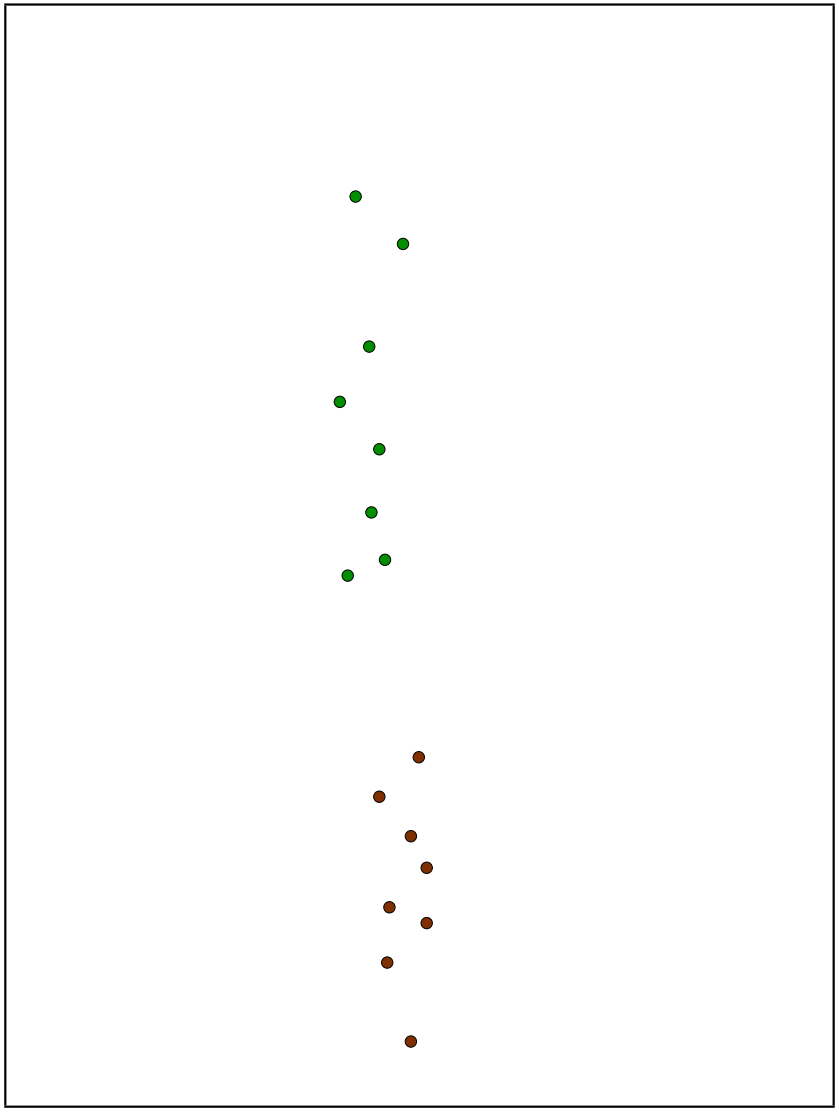
$\boldsymbol{p}_i = \boldsymbol{x}_i(t+1)$ if $f(\boldsymbol{x}_i(t+1)) < f(\boldsymbol{p}_i)$

$\boldsymbol{g}_i = \boldsymbol{x}_*(t+1)$ where $\mathrm{f}(\boldsymbol{x}_*(\mathrm{t}+1)) = \min\{\mathrm{f}(\boldsymbol{x}_\mathrm{k}(\mathrm{t}+1)), \mathrm{k} \in$ neighborhood of $i\}$

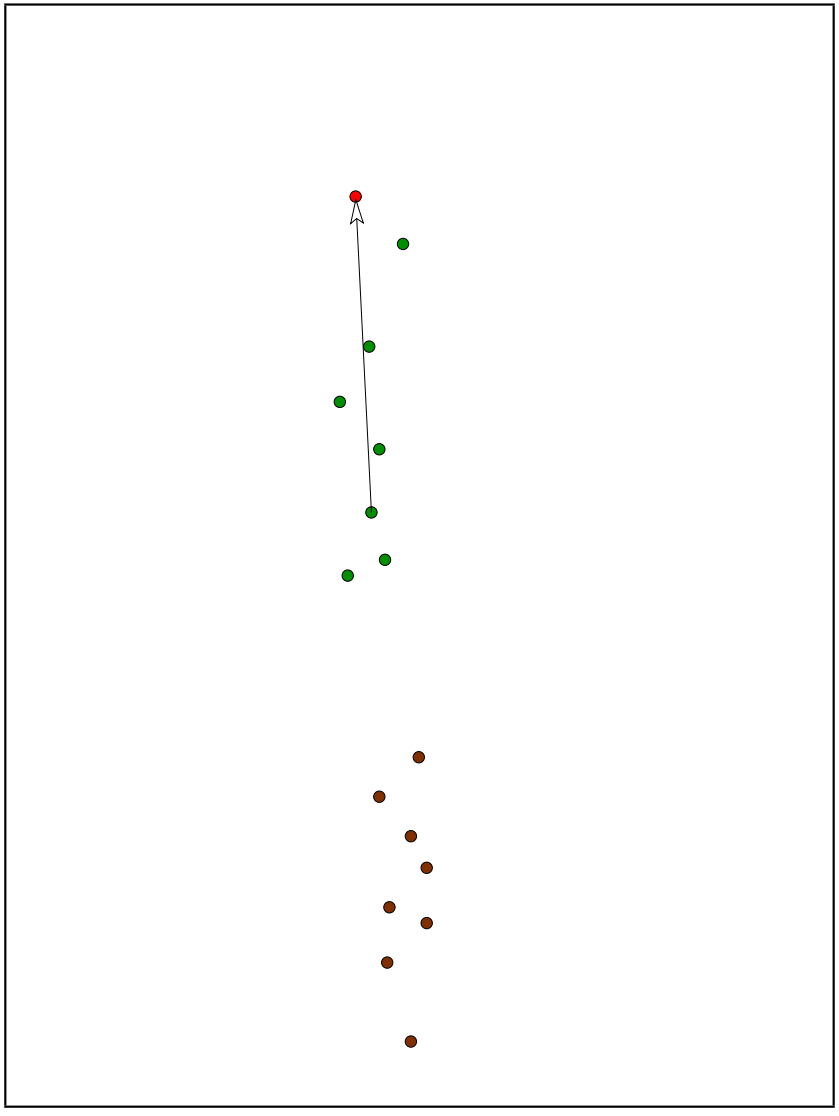
Non-invariance of PSO



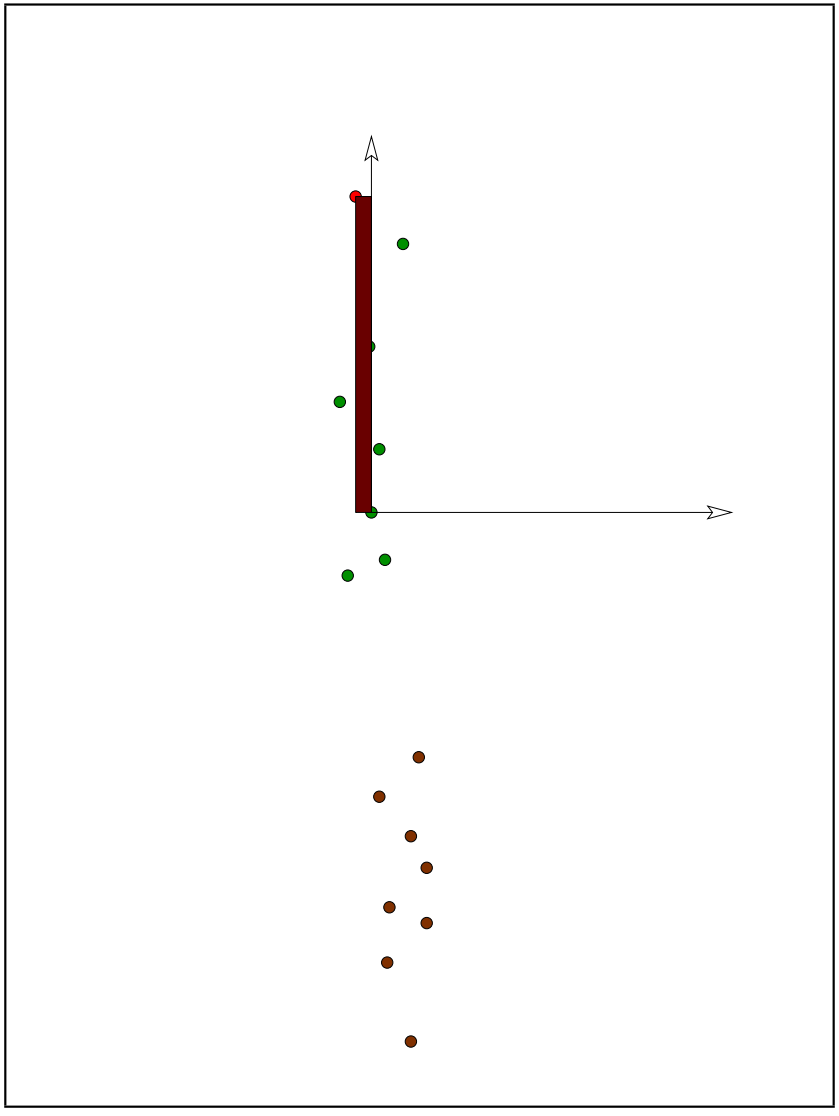
Non-invariance of PSO



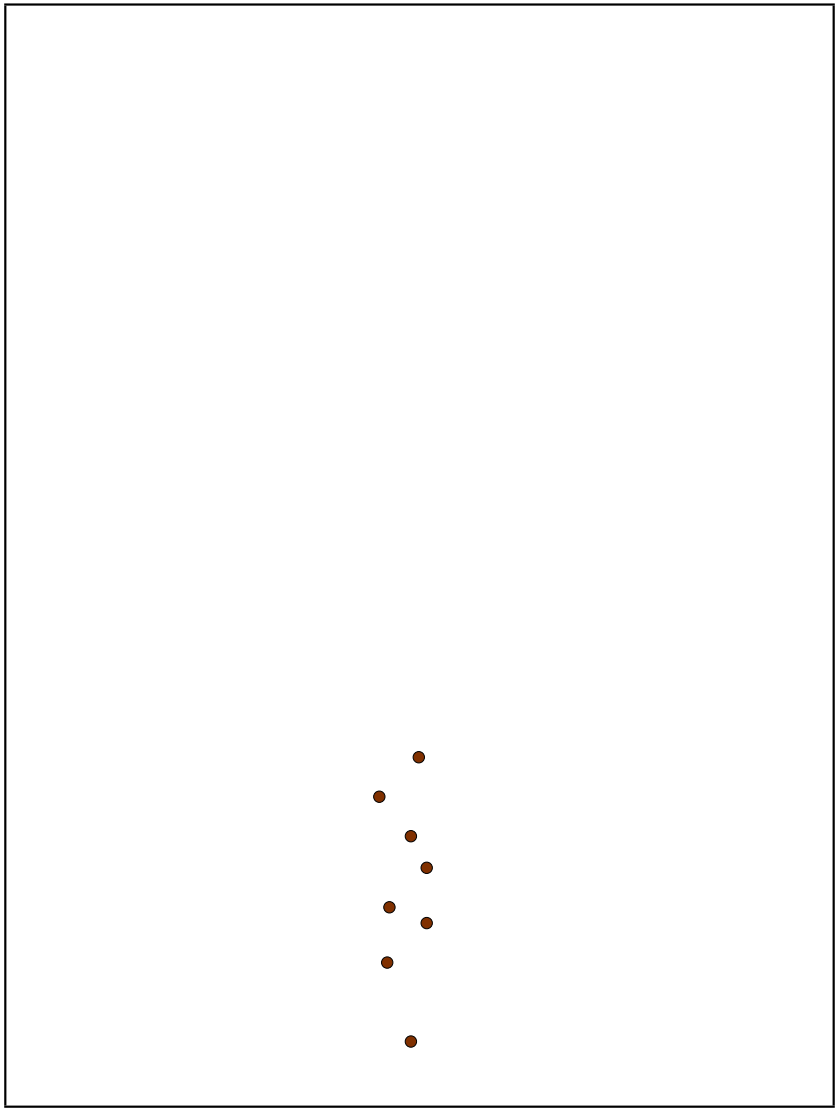
Non-invariance of PSO

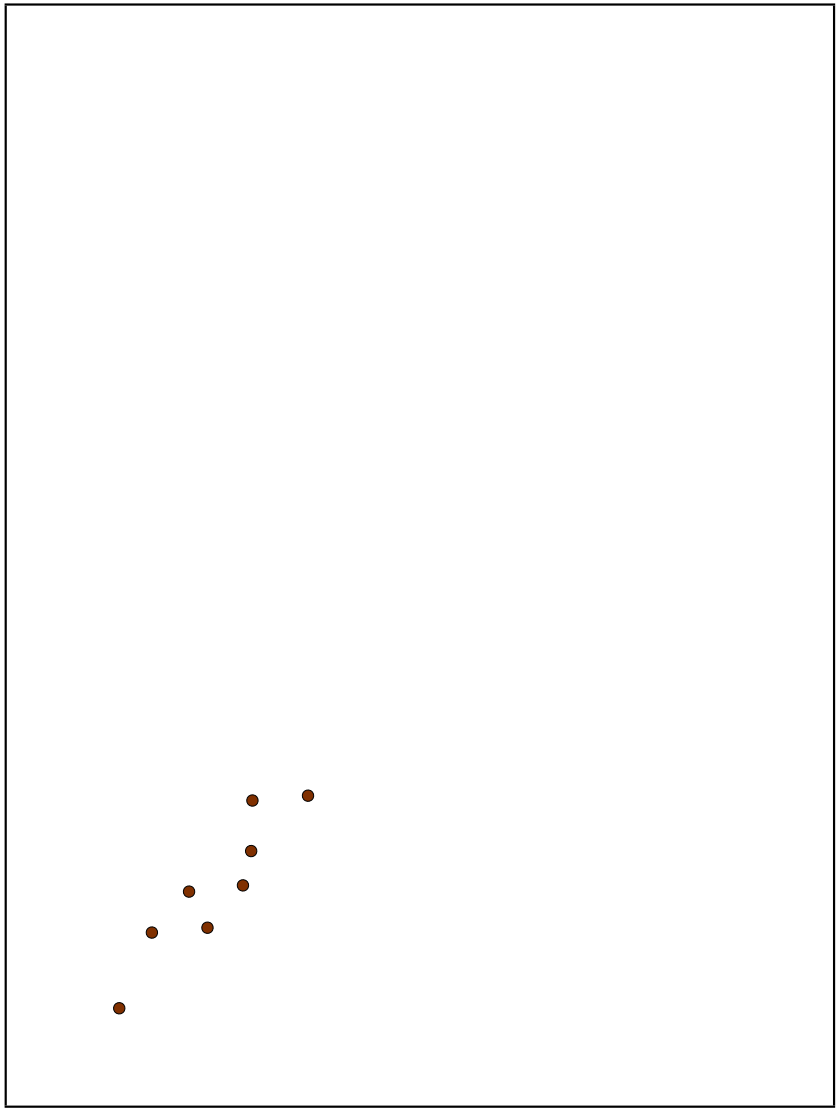


Non-invariance of PSO

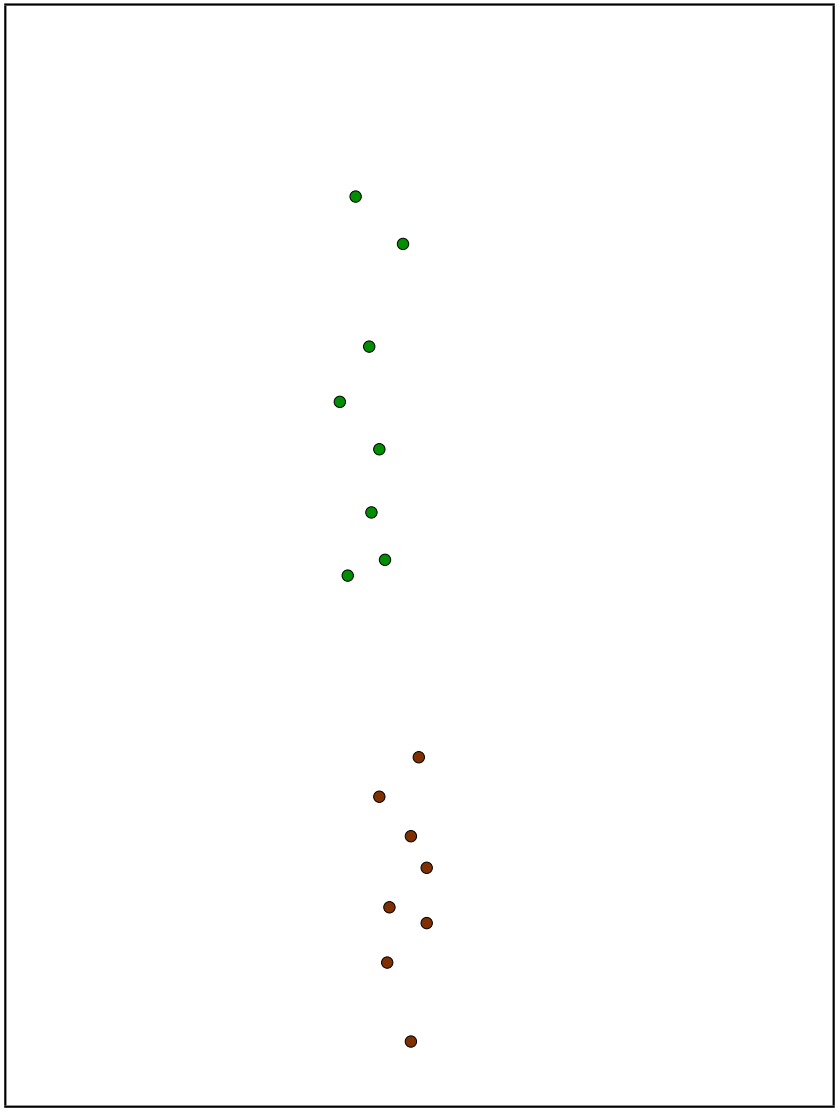


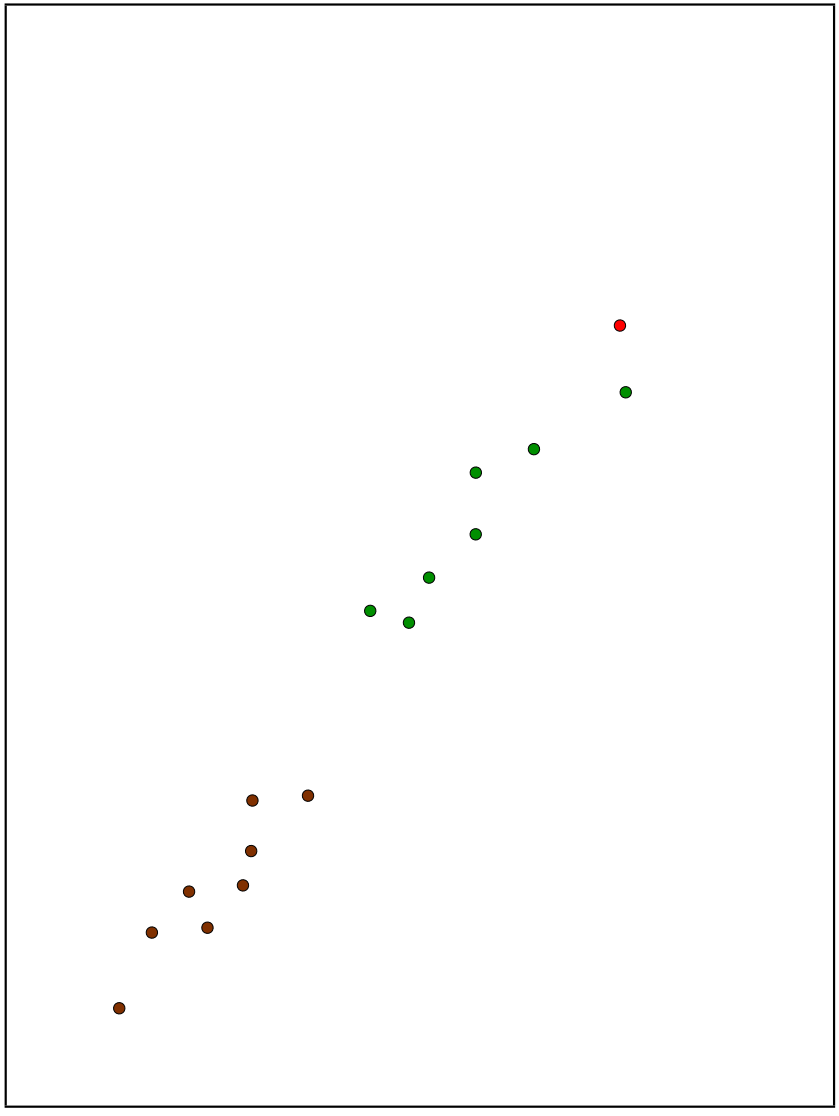
Non-invariance of PSO



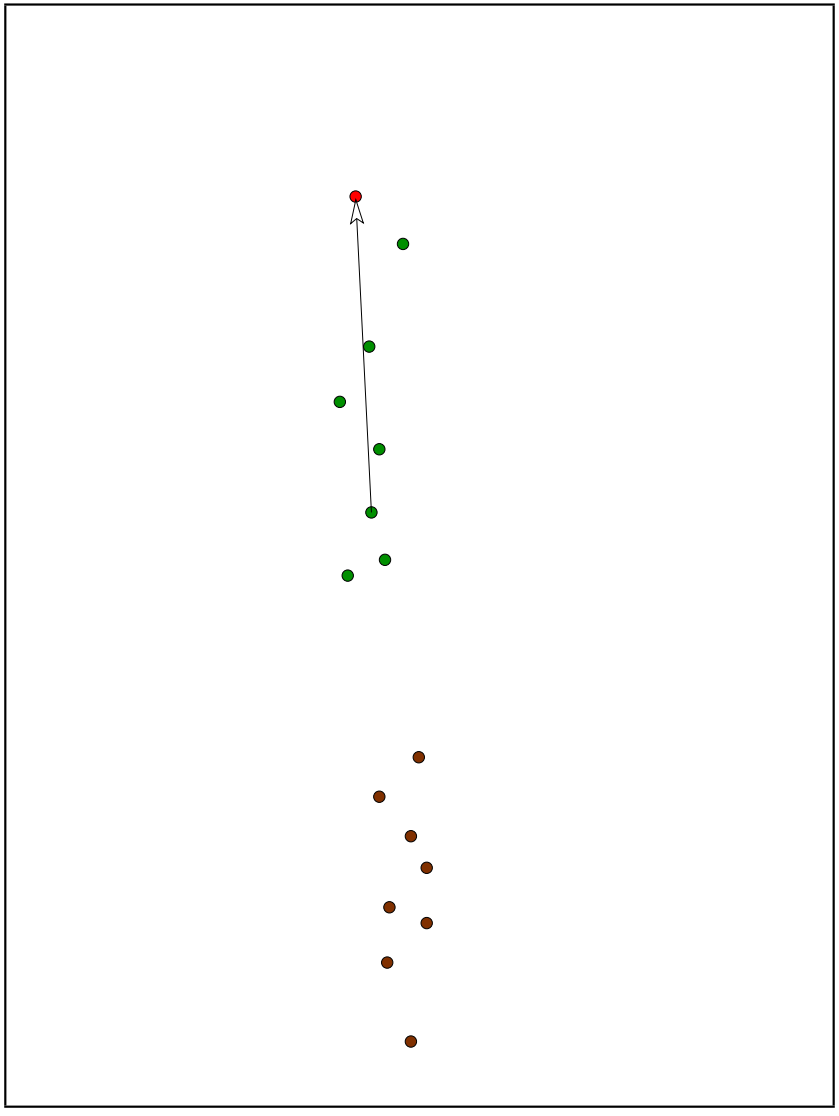


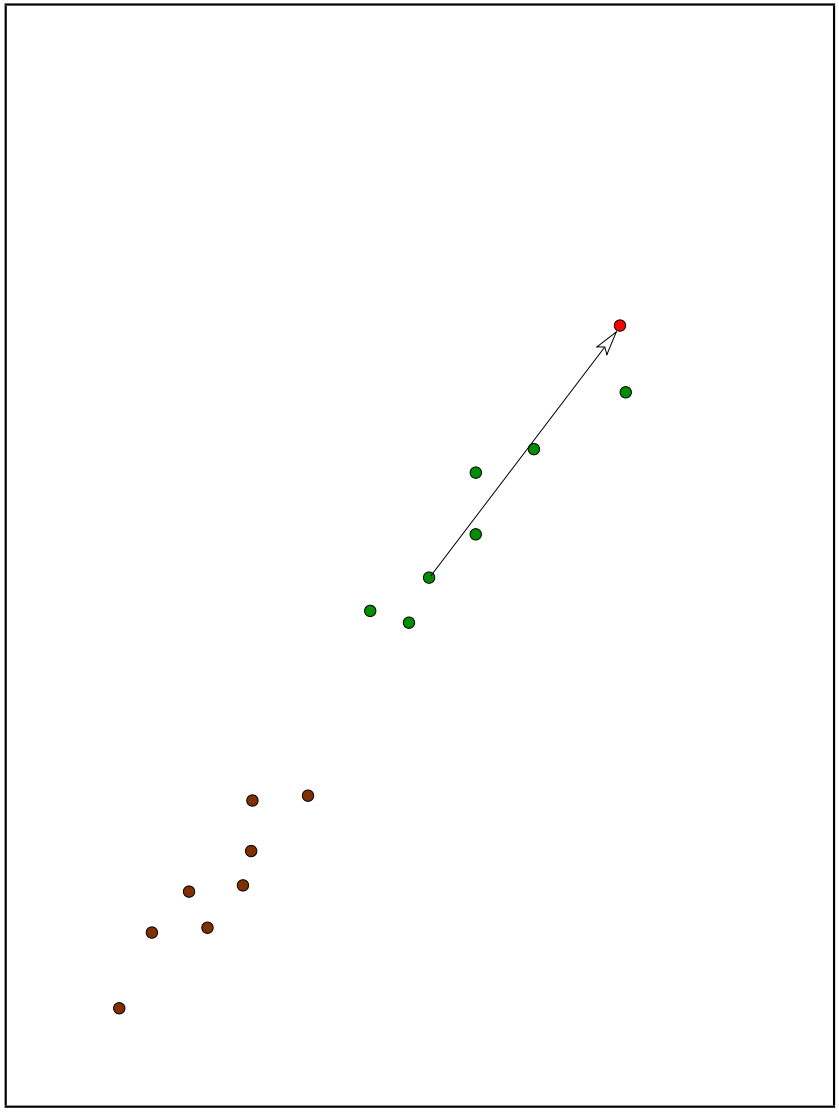
Non-invariance of PSO



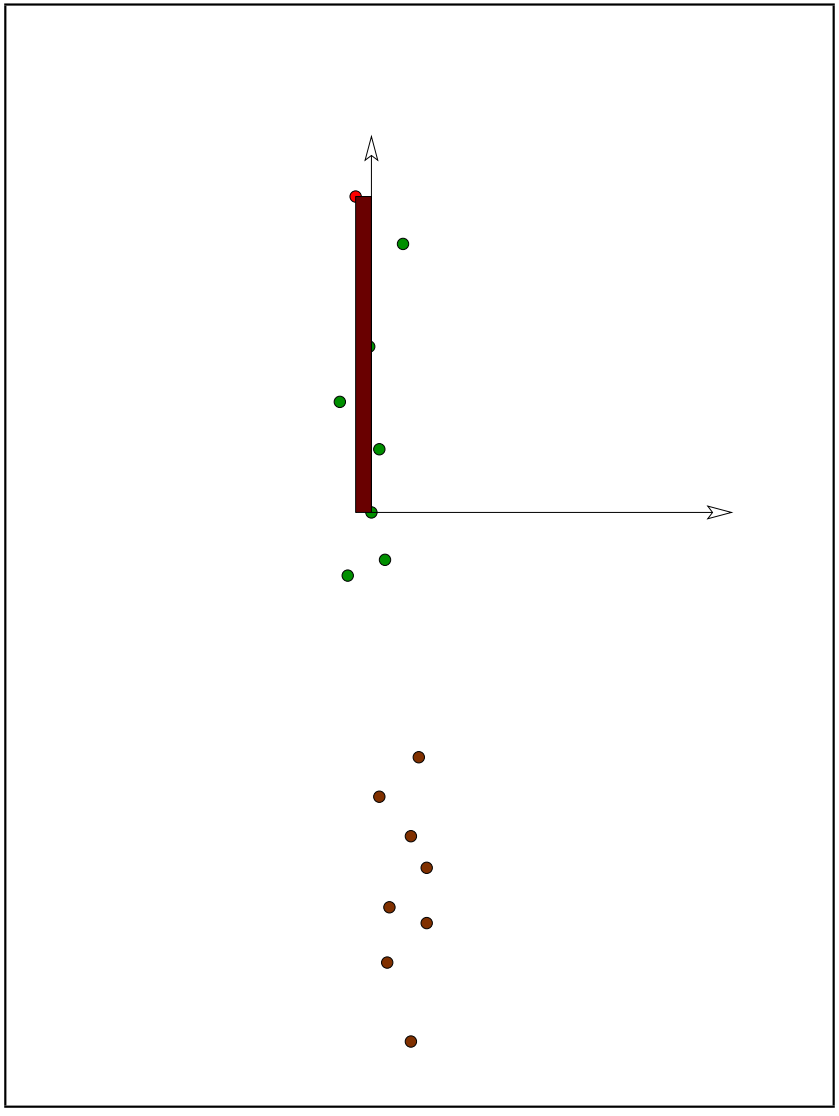


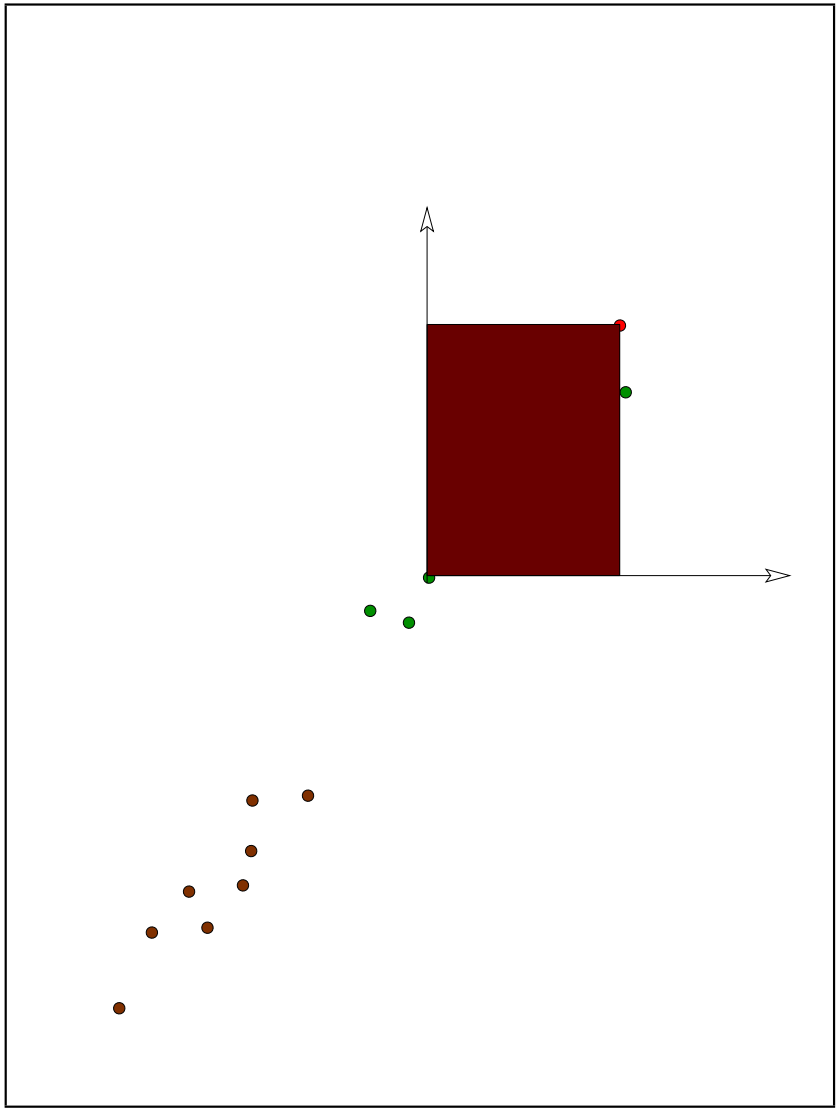
Non-invariance of PSO





Non-invariance of PSO





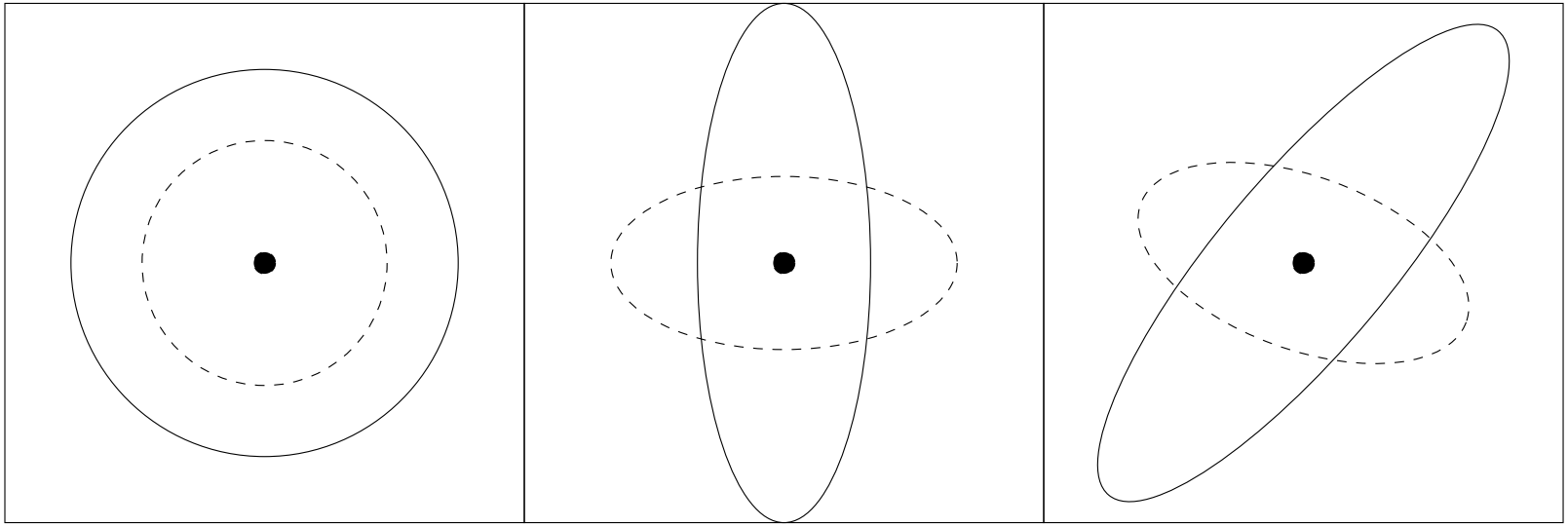
Evolution Strategies (ES)

New search points are sampled normally distributed

$$\boldsymbol{x}_i \sim \mathcal{N}_i\left(\boldsymbol{m}, \sigma^2 \boldsymbol{C}\right) \qquad \text{for } i = 1, \ldots, \lambda$$

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

Lines of Equal Density



$\mathcal{N}\left(\boldsymbol{m}, \sigma^2 \boldsymbol{I}\right)$
one free parameter σ

$\mathcal{N}\left(\boldsymbol{m}, \boldsymbol{D}^2\right)$
n free parameters

$\mathcal{N}\left(\boldsymbol{m}, \boldsymbol{C}\right)$
$(n^2 + n)/2$ free parameters

ES, CMA-ES

Adaptation of the sampling distribution

1. one-fifth success rule

   rotational invariant isotropic distribution
   Rechenberg, 60's

2. self-adaptive ES

   adaptive search distribution along coordinate axis

   not rotational invariant
   Rechenberg, Schwefel, 70's, 80's

3. CMA-ES
   - adaptation of full covariance matrix
   - invariance to linear transformation of the search space

   adaptive *and* rotational invariant
   Hansen, Ostermeier, 90's

Performed best for CEC 05 challenge on parameter optimization

1. Continuous optimization and stochastic search

2. PSO facing ill-conditioned and non-separable problems
   - Test problems
   - Results

3. Summary and conclusion

Experimental setting

1. Test problems
   - Ellipsoid function — separable / non-separable, different condition number
   - Rosenbrock function — non-separable, different condition number
   - Rastrigin function — separable / non-separable
2. Algorithms
   - PSO (Standard PSO 2006) — code available at http://www.particleswarm.info/Standard_PSO_2006.c
   - CMA-ES — code and references available at http://www.bionik.tu-berlin.de/user/niko/

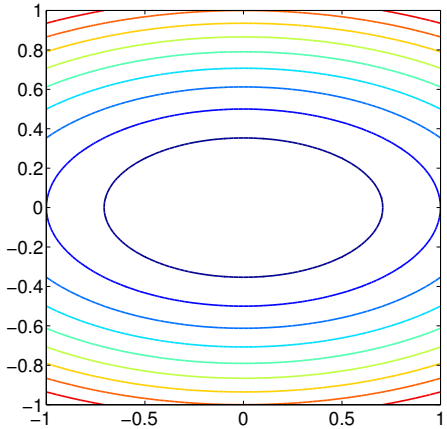
21 runs in each case and default paramters.

Ellipsoid

- $f_{\text{elli}}(x) = \sum_{i=1}^{n} 10^{\alpha \frac{i-1}{n-1}} x_i^2 = x^T H_{\text{elli}} x$

$$H_{\text{elli}} = \begin{pmatrix} 1 & 0 & \cdots \\ & \ddots & \\ \cdots & 0 & 10^{\alpha} \end{pmatrix}$$

separable

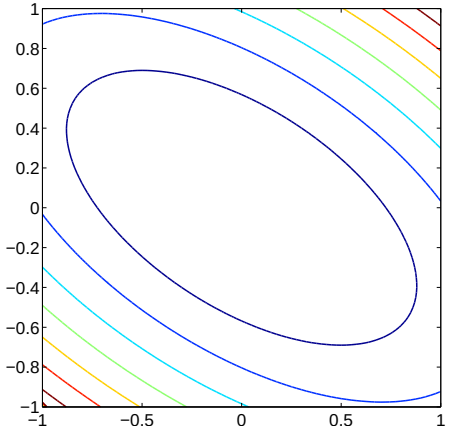


- $f_{\text{elli}}^{\text{rot}}(x) = f_{\text{elli}}(\boldsymbol{R}x) = x^T H_{\text{elli}}^{\text{rot}} x$

$\boldsymbol{R}$ random rotation

$H_{\text{elli}}^{\text{rot}} = \boldsymbol{R}^T H_{\text{elli}} \boldsymbol{R}$

non-separable



- $\text{cond}(H_{\text{elli}}) = \text{cond}(H_{\text{elli}}^{\text{rot}}) = 10^{\alpha}$

$\alpha = 1, \ldots, 10$

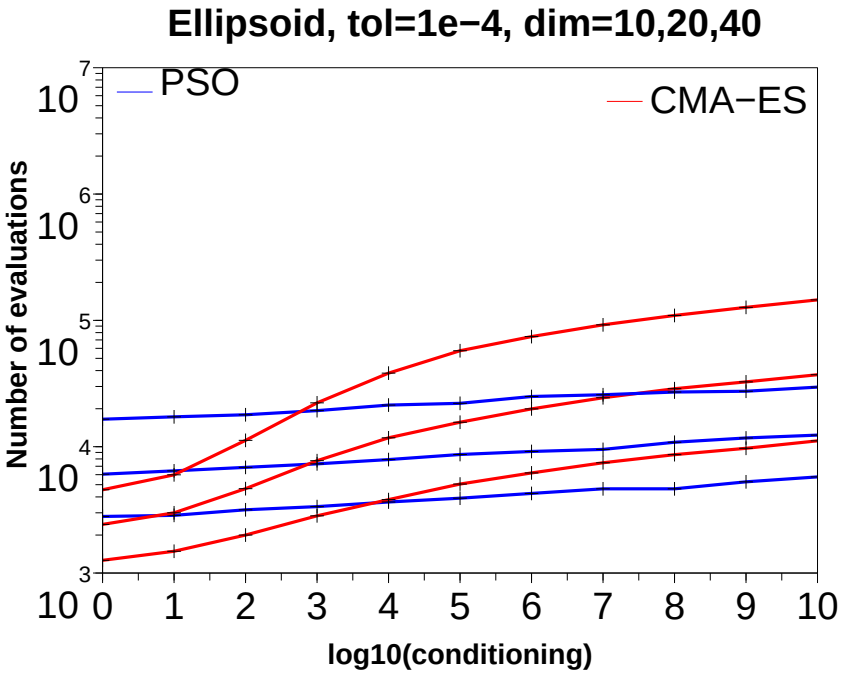
$\alpha = 6 \equiv$ axis ratio of 10^3, typical for real-world problem

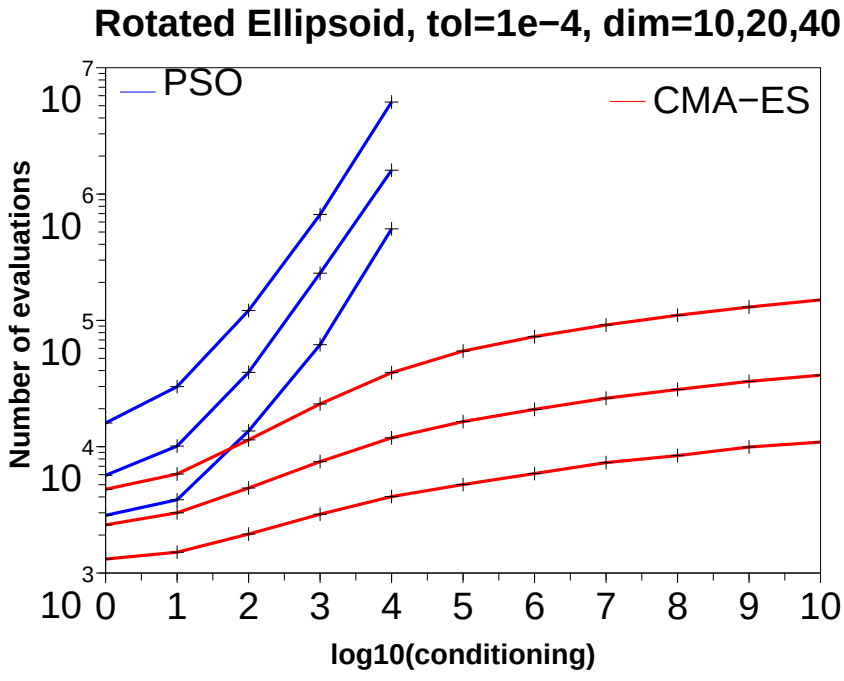
Results

Ellipsoid functions

$f_{\text{elli}}(x) = \sum_{i=1}^{n} (10^{\alpha})^{\frac{i-1}{n-1}} x_i^2$

PSO vs CMA-ES



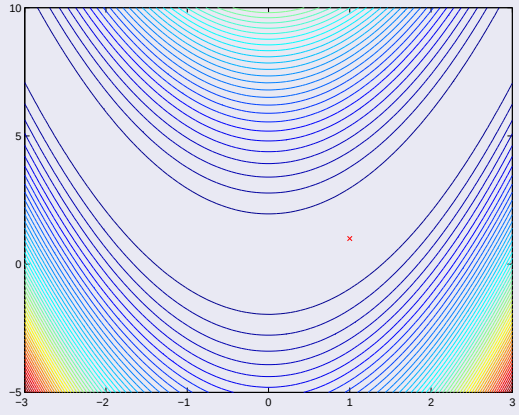


Rosenbrock (Banana) function

$$f_{\text{rosen}}(x) = \sum_{i=1}^{n-1} \left[(1 - x_i)^2 + 10^{\alpha}(x_{i+1} - x_i^2)^2\right]$$

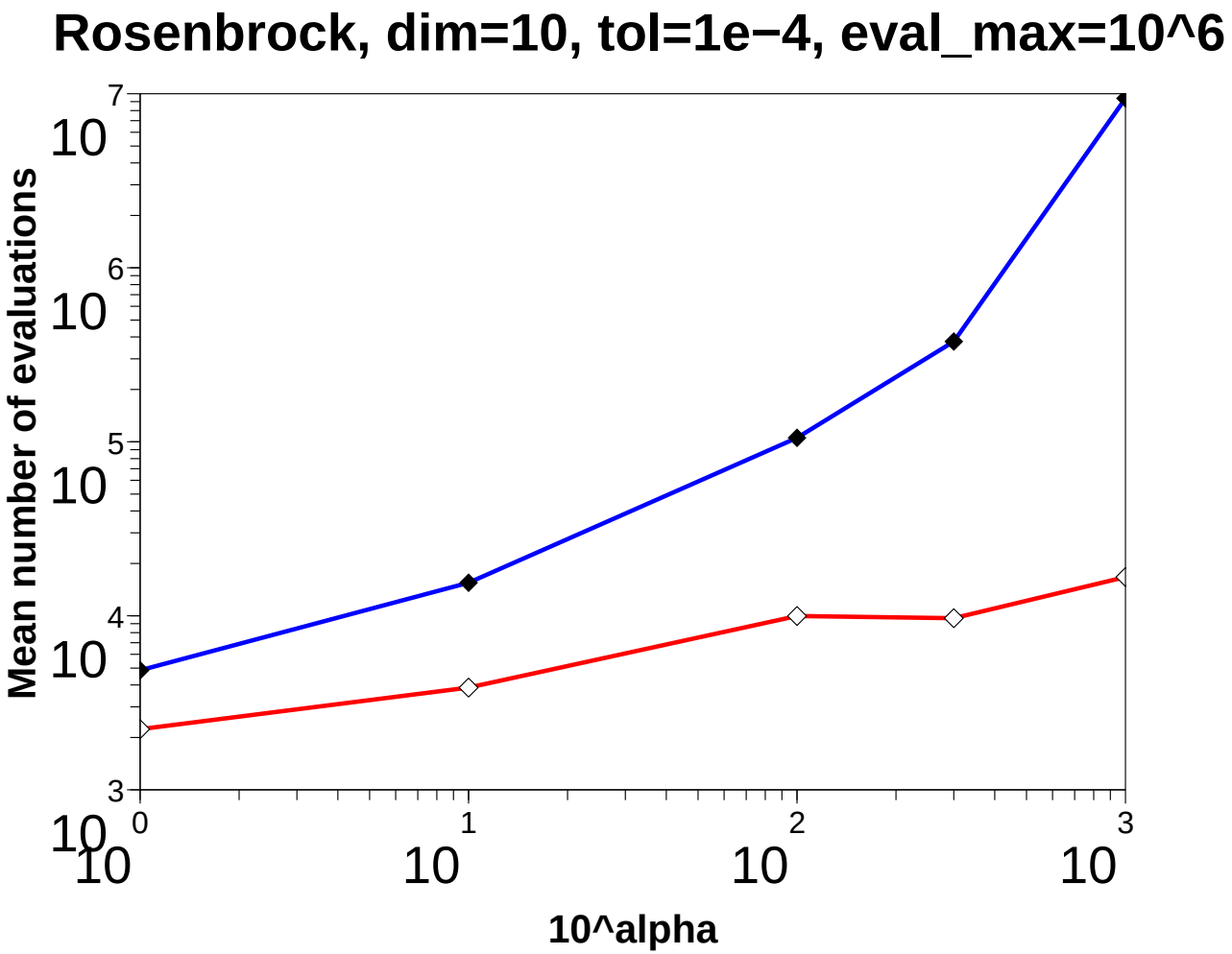
$\alpha = 1, \ldots, 4$

- non-separable
- $\alpha = 2$, classical Rosenbrock function



Rosenbrock functions

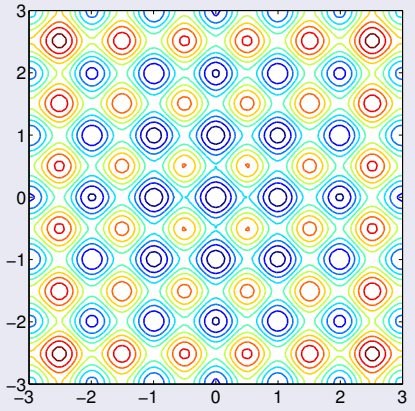
$f_{\text{rosen}}(x) = \sum_{i=1}^{n-1} \left[(1 - x_i)^2 + 10^{\alpha}(x_{i+1} - x_i^2)^2\right]$



Rastrigin function

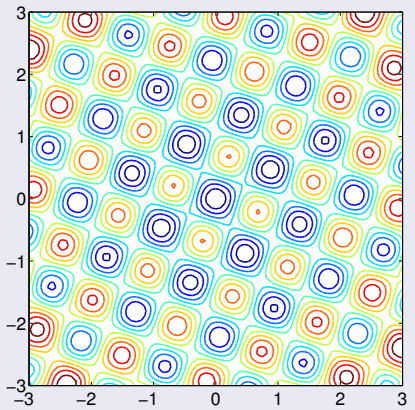
$f_{\text{rast}}(x) = 10n + \sum_{i=1}^{n} x_i^2 - 10\cos(2\pi x_i)$

- separable
- multi-modal



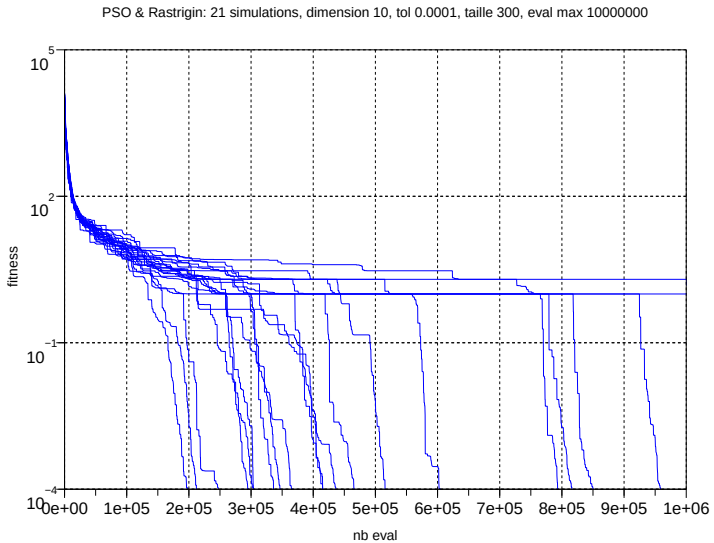
$f_{\text{rast}}^{\text{rot}}(x) = f_{\text{rast}}(\boldsymbol{R}x)$

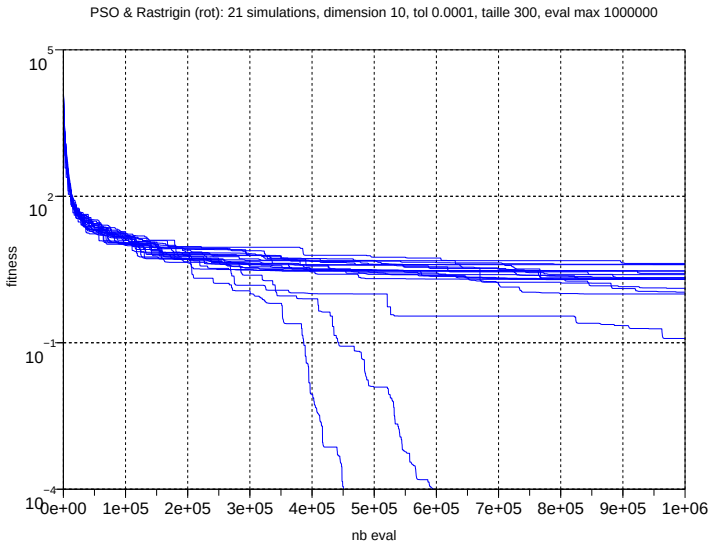
- $\boldsymbol{R}$ random rotation
- non-separable
- multimodal

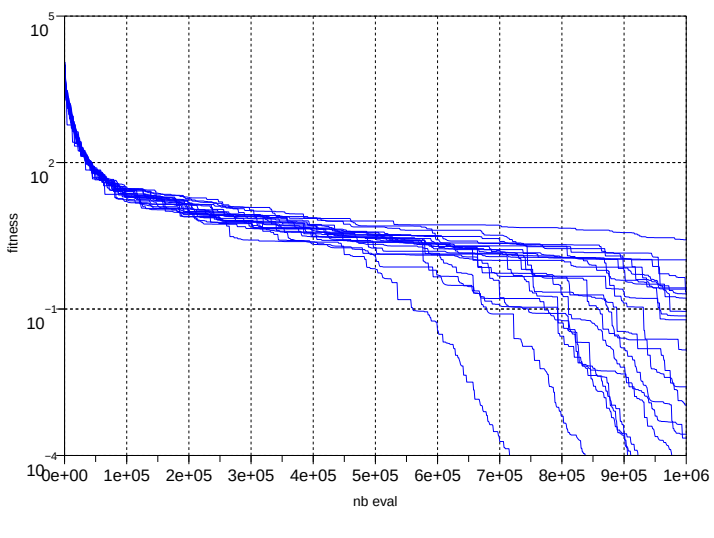


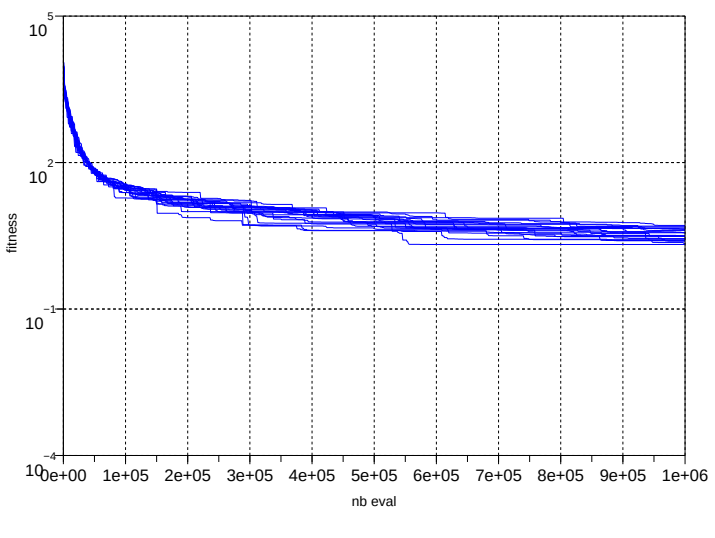
Results

Rastrigin function
PSO exploits the separability



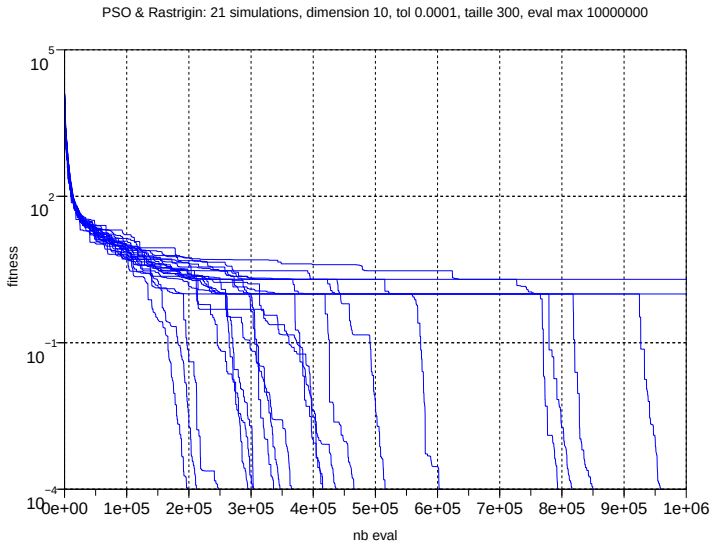


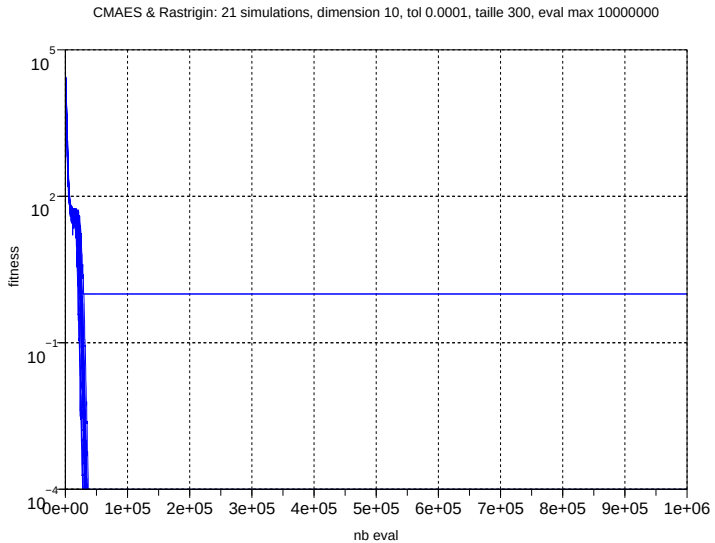


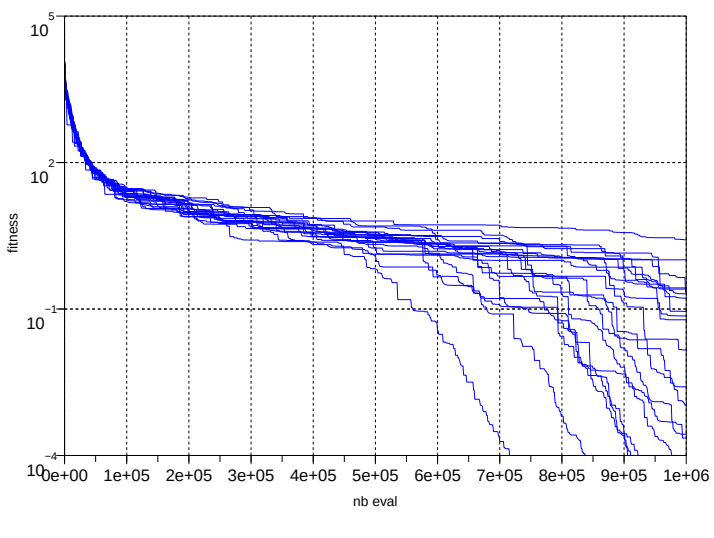


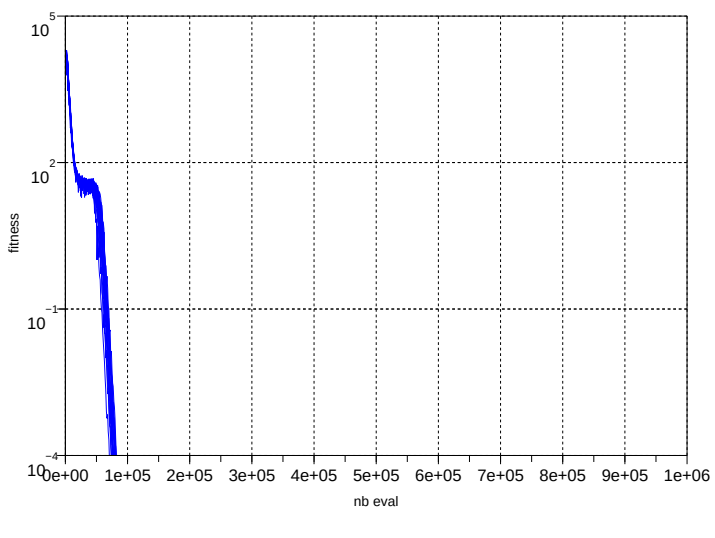
Results

Rastrigin function
PSO vs CMA-ES









Summary and conclusion

- PSO exploits the separability

  ellipsoid vs rotated ellipsoid
  rastrigin vs rotated rastrigin

  - faster than CMA-ES for separable ellipsoid and cond > 3

    up to 6 times faster

  - more than 100 times slower on rotated ellipsoid than non-rotated (cond = 10^4)
- PSO virtually unable to solve rotated ellipsoid for cond > 10^4
- PSO 70 times slower than CMA-ES on classical Rosenbrock function
- PSO non-invariant

Can we built an PSO that performs well on functions with dependencies between variables?