# Recent Advances in Continuous Randomized Black-Box Optimization: an Overview

Anne Auger

Optimization and Machine Learning Team (TAO)
INRIA Saclay-Ile-de-France

PGMO/COPI'14 28 - 31 October 2014 Paris Saclay
(Ecole Polytechnique)

# Black-Box Optimization - Zero

- Optimize $f : \mathbb{R}^n \mapsto \mathbb{R}$
- Zero$^{th}$ order method + Black-Box setting


- Cost = # calls to the black-box (f-calls)

Derivative-Free Optimization (DFO) setting

# Function-Value-Free (FVF) / Comparison-based / Ranked-based Optimization

- Optimize $f : \mathbb{R}^n \mapsto \mathbb{R}$
- Zero[th] order method + Black-Box setting


- Cost = # calls to the black-box (f-calls)
- Optimization algorithm only allowed to use f-comparisons

$$\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n \rightarrow f(\mathbf{x}_1) < f(\mathbf{x}_2) \ ?$$

$$\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n \rightarrow f(\mathbf{x}_1) \geq f(\mathbf{x}_2) \ ?$$

# Function-Value-Free (FVF) / Comparison-based / Ranked-based Optimization

- Optimize $f : \mathbb{R}^n \mapsto \mathbb{R}$
- Zero$^{th}$ order method + Black-Box setting


- Cost = # calls to the black-box (f-calls)
- Optimization algorithm only allowed to use f-comparisons

Well-known comparison-based algorithms:
Nelder-Mead
Hooke and Jeeves / pattern search
Evolution Strategies and many Evolutionary Algorithms

# Why Comparison-based?

- Robustness:
  - error on f-value (due to noise, ...) has an impact only if it changes the result of a comparison
  - very small or very large f-values have only a limited impact
- Generalization:
  - same result on $f$ or $g \circ f$ if $g : \mathbb{R} \mapsto \mathbb{R}$ strictly increasing

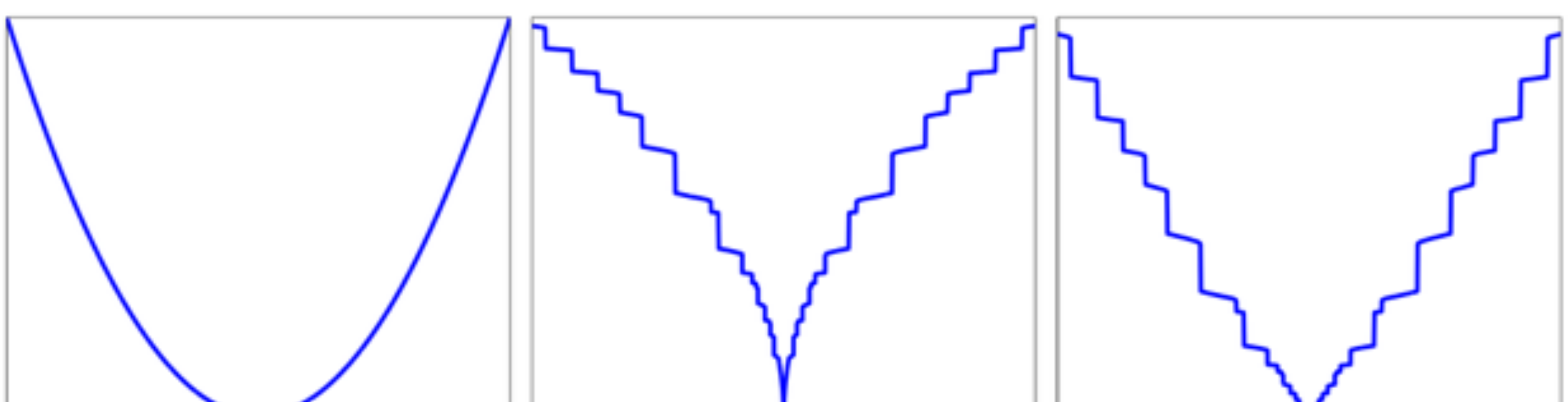
$$f(\mathbf{x}_1) \leq f(\mathbf{x}_2) \leq \ldots \leq f(\mathbf{x}_\lambda)$$

$$g \circ f(\mathbf{x}_1) \leq g \circ f(\mathbf{x}_2) \leq \ldots \leq g \circ f(\mathbf{x}_\lambda)$$

*Invariance to strict. increasing transformations of f*

# Why Comparison-based?

- Robustness:
  - error on f-value (due to noise, …) has an impact only if it changes the result of a comparison
  - very small or very large f-values have only a limited impact
- Generalization:
  - same result on $f$ or $g \circ f$ if $g : \mathbb{R} \mapsto \mathbb{R}$ strictly increasing

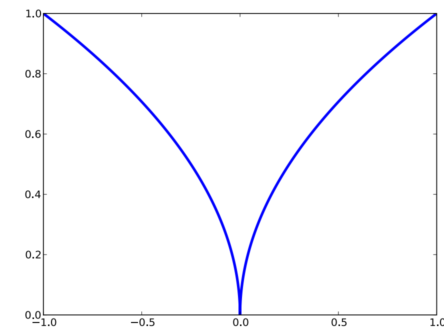


*Invariance to strict. increasing transformations of f*

# What Makes a Function Difficult to Solve

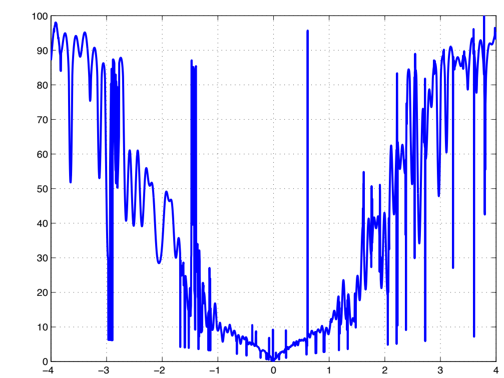
## Why Comparison-based Stochastic

- non-linear, non-quadratic, non convex



- ruggedness

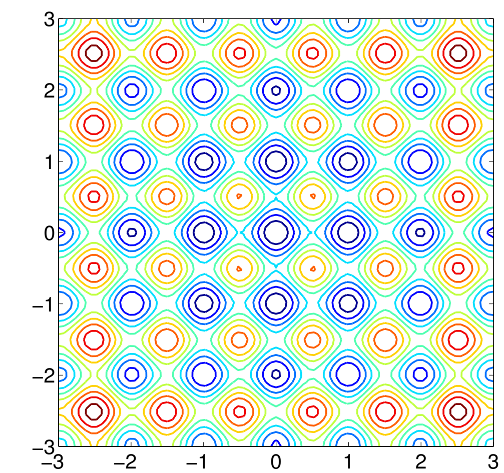
  *non-smooth, discontinuous, multi-modal, and/or noisy functions*



- dimensionality (size of the search space)

  *(considerably) larger than three*

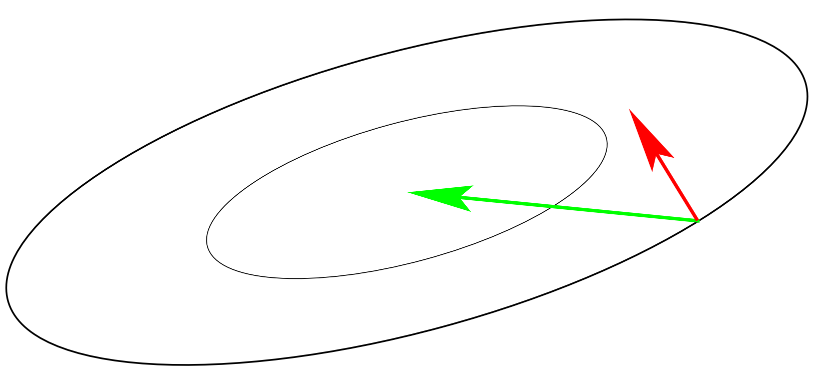
  *curse of dimensionality*




- non-separability

  *dependencies between the objective variables*

- ill-conditioning



# Adaptive Stochastic Search

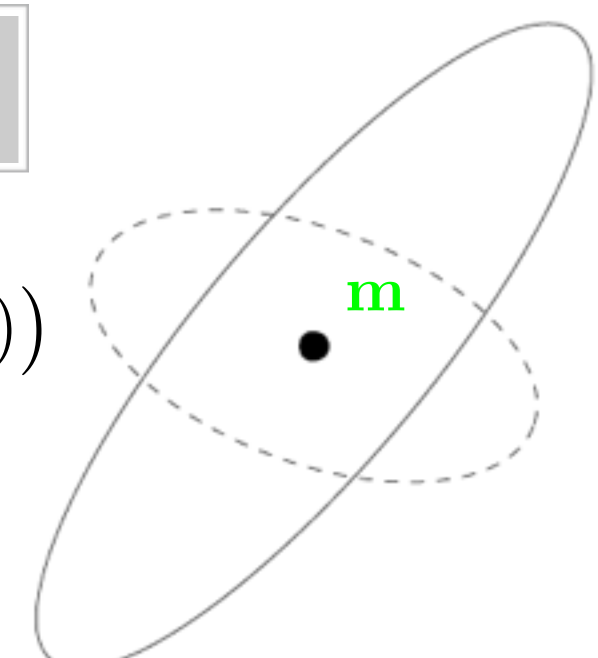
## A black-box search template to minimize $f : \mathbb{R}^n \to \mathbb{R}$

Initialize distribution parameters $\theta$, set population size $\lambda$
While not terminate

1. Sample distribution $p_\theta(x) : x_1, \ldots, x_\lambda \in \mathbb{R}^n$
2. Evaluate $x_1, \ldots, x_\lambda$ on $f$
3. Update parameters $\theta \leftarrow F(\theta, x_1, \ldots, x_\lambda, f(x_1), \ldots, f(x_\lambda))$

## Example of $p_\theta$ on $\mathbb{R}^n$

multivariate normal distribution: $\mathbf{m} + \sigma \mathcal{N}(0, \mathbf{C})$
*density* : $p_{\theta:=(\mathbf{m},\mathbf{C})}(x) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{C}|}} \exp\left(-\frac{1}{2}(x-\mathbf{m})^T \mathbf{C}^{-1}(x-\mathbf{m})\right)$




$\{x | (x-\mathbf{m})^T \mathbf{C}^{-1}(x-\mathbf{m}) = cst\}$

- ☆ Covariance Matrix Adaptation Evolution Strategies (CMA-ES) [N. Hansen et al, 2001-2013]
- ☆ Exponential Natural Evolution Strategies (xNES) [T. Glasmachers et al, 2010]

# Adaptive Comparison

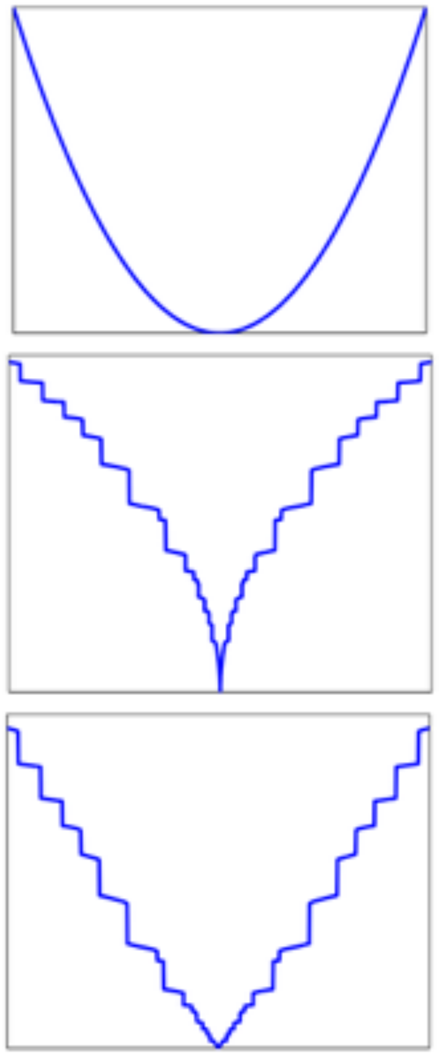
## Function-Value-Free Optimization

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

Initialize distribution parameters $\theta$, set population size $\lambda$

While not terminate

1. Sample distribution $p_\theta(x) : x_1, \ldots, x_\lambda \in \mathbb{R}^n$
2. Evaluate $x_1, \ldots, x_\lambda$ on $f$, rank solutions in $\mathcal{S}^f$
3. Update parameters $\theta \leftarrow F(\theta, x_1, \ldots, x_\lambda, \mathcal{S}^f)$



Permutation $\mathcal{S}^f$ such that:

$$f(x_{\mathcal{S}^f(1)}) \leq f(x_{\mathcal{S}^f(2)}) \leq \ldots \leq f(x_{\mathcal{S}^f(\lambda)})$$

# CMA-ES with rank-mu update

**Sample multivariate normal distribution**

$$\mathbf{x}_i = \mathbf{m}_t + \mathbf{C}_t^{1/2}\mathbf{y}_i \;\;,\;\; \mathbf{y}_i \sim \mathcal{N}(0, I_n) \;\;, i = 1, \ldots, \lambda$$

**Evaluate and rank solutions**

$$f(\mathbf{m}_t + \mathbf{C}_t^{1/2}\mathbf{y}_{1:\lambda}) \leq \ldots \leq f(\mathbf{m}_t + \mathbf{C}_t^{1/2}\mathbf{y}_{\lambda:\lambda})$$

**Update mean and covariance matrix**

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

$$\mathbf{C}_{t+1} = (1 - c_{\text{cov}})\mathbf{C}_t + c_{\text{cov}} \sum_{i=1}^{\mu} w_i \mathbf{y}_{i:\lambda}\mathbf{y}_{i:\lambda}^T$$

# Adaptive Stochastic Comparison-Based Optimization

Brooks, Pure Random Search, 1958

## Step-size adaptive algorithms

Matyas, Random optimization, 1965
Schumer, Steiglitz, Adaptive step size random search, 1968
Devroye, The compound random search, 1972
Rechenberg, Evolution Strategies (ES), One-fifth success rule, 1973

## Covariance matrix adaptive algorithms

Kjellström, Gaussian Adaptation, 1969
Hansen, Ostermeier, Covariance Matrix Adaptation ES, 2001 *State-of-the-art algorithm*
Glasmachers, Schaul, Yi, Wiestra, Schmidhuber, Exponential Natural ES, 2010

# Adaptive Stochastic Comparison-Based Optimization

Brooks, Pure Random Search, 1958

Convergence with probability one
on non-pathological functions $T(\epsilon) = \Theta(\frac{1}{\epsilon^n})$

## Step-size adaptive algorithms

Matyas, Random optimization, 1965
Schumer, Steiglitz, Adaptive step size random search, 1968
Devroye, The compound random search, 1972
Rechenberg, Evolution Strategies (ES), One-fifth success rule, 1973

Linear convergence on
wide class of functions
(ample empirical evidence)

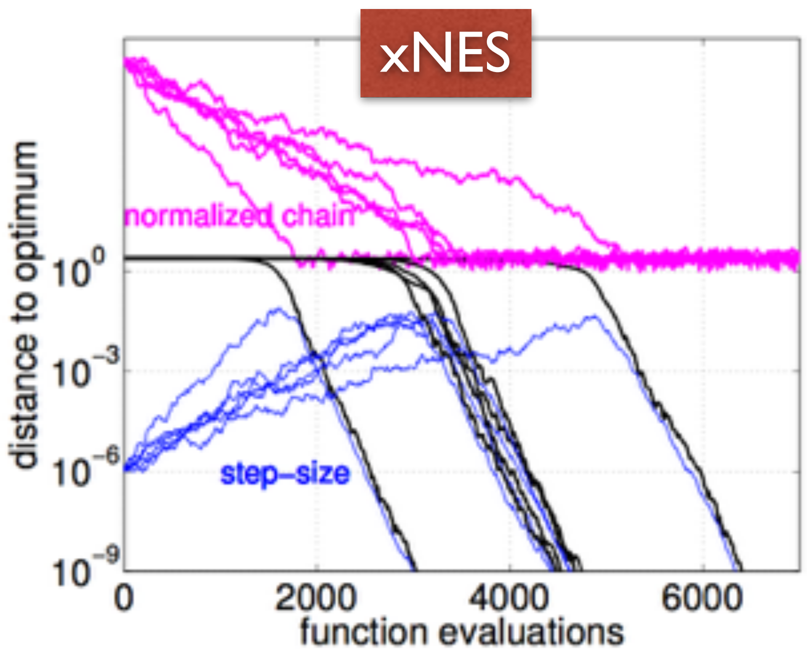
## Covariance matrix adaptive algorithms

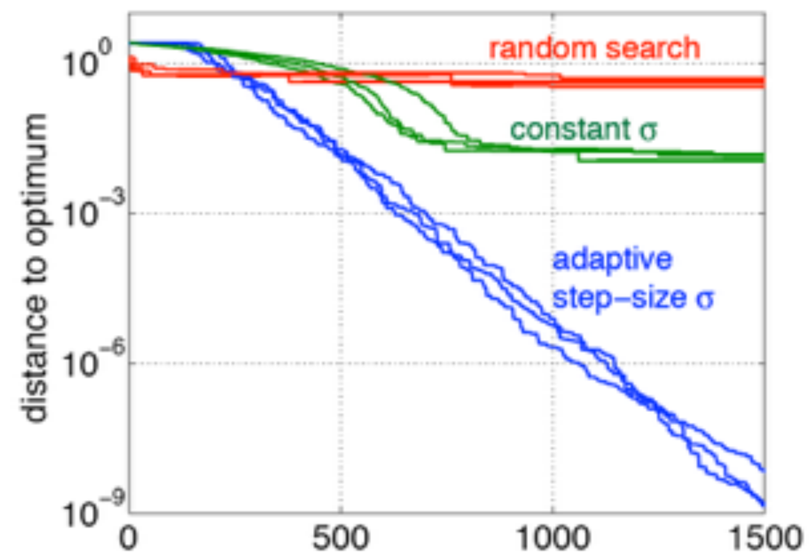
Kjellström, Gaussian Adaptation, 1969
Hansen, Ostermeier, Covariance Matrix Adaptation ES, 2001 *State-of-the-art algorithm*
Glasmachers, Schaul, Yi, Wiestra, Schmidhuber, Exponential Natural ES, 2010

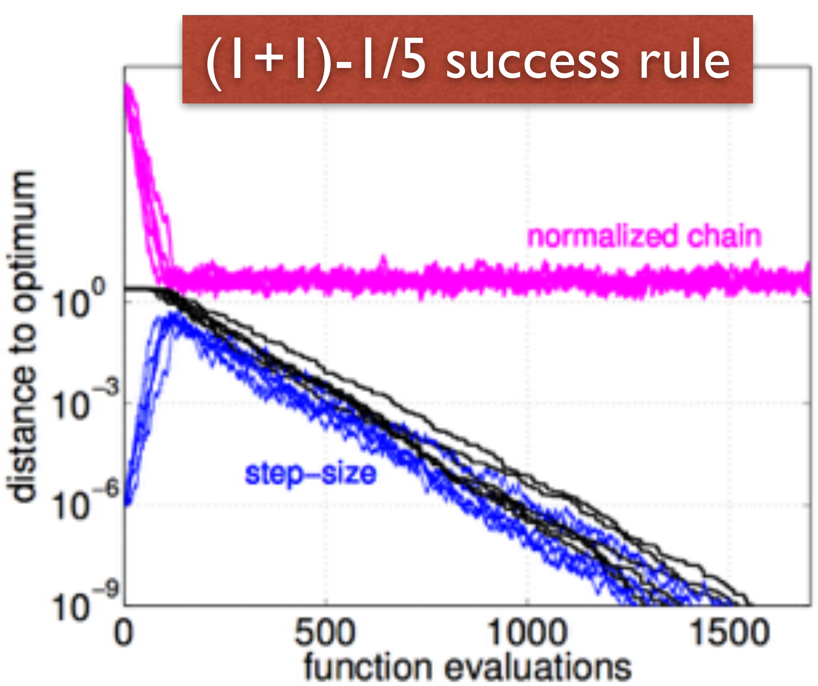
Learn second order information
*solve efficiently ill-conditioned non-separable problems*

# Linear Convergence of Step-size Adaptive Algorithms
## *Scaling-invariant Functions*

$f(\mathbf{x}) = \|\mathbf{x}\|$




**Almost sure linear convergence**

$$\frac{1}{t} \ln \frac{\|\mathbf{X}_t - \mathbf{x}^\star\|}{\|\mathbf{X}_0 - \mathbf{x}^\star\|} \xrightarrow[t\to\infty]{} -\mathrm{CR}$$

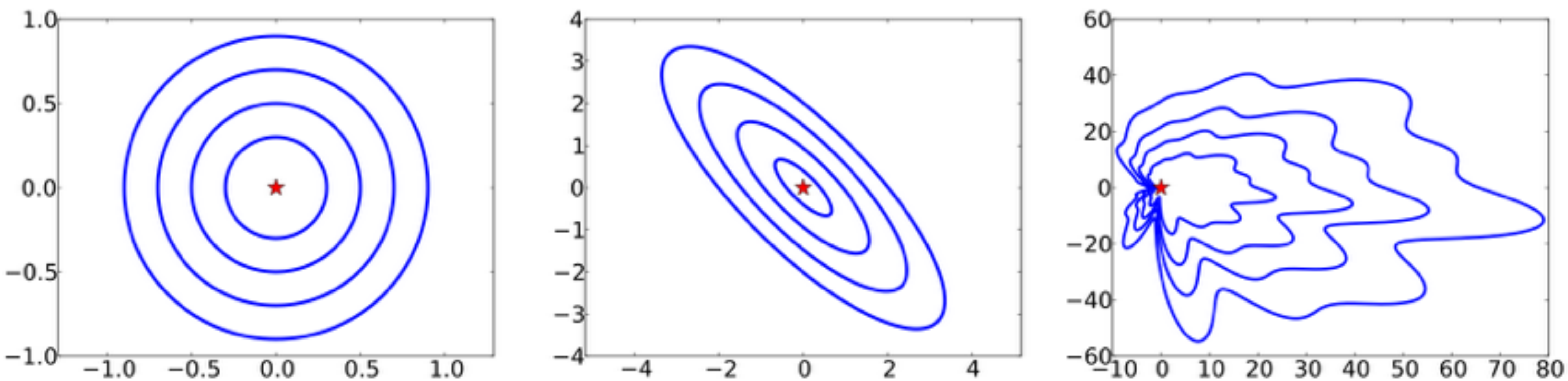
**Empirical evidences**

# Linear Convergence on Scaling-Invariant Functions

## Scaling-invariant functions

$f$ is scaling-invariant w.r.t. zero if for all $\rho > 0$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$f(\rho \mathbf{x}) \leq f(\rho \mathbf{y}) \Leftrightarrow f(\mathbf{x}) \leq f(\mathbf{y}) \ .$$




Linear convergence proven for scale-invariant step-size adaptive algorithms on scaling-invariant functions

*stability analysis of underlying Markov chain*


Linear Convergence of Comparison-based Step-size Adaptive Randomized Search via Stability of Markov Chains, Auger, Hansen, 2014, http://arxiv.org/abs/1310.7697

Linear Convergence on Positively Homogeneous Functions of a Comparison Based Step-Size Adaptive Randomized Search: the (1+1) ES with Generalized One-fifth Success Rule, Auger, Hansen, 2014, http://arxiv.org/abs/1310.8397

# Connexion with Optimization on Manifolds

## Information Geometric Optimization

- Transform original problem into optimization problem on the statistical manifold $\Theta$ where $p_\theta$ lives

Minimize $J(\theta) = \int f(x) p_\theta(x) dx$

*not invariant to mont. transformation of f*

*Wiestra et al. Natural Evolution Strategies, CEC 2008*
*Sun et al. Efficient natural evolution strategies GECCO 2009*
*Glasmachers et al. Exponential NES GECCO 2010*

Maximize $J_{\theta_t}(\theta) = \int w(P_{\theta_t}[y : f(y) \leq f(x)]) p_\theta(x) dx$

$w : [0, 1] \mapsto \mathbb{R}$, decreasing weight function

*Ollivier et al. Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles, arXiv*

- Perform a natural gradient step on $\Theta$

gradient taken w.r.t. Fisher Information metric $I_{ij} = \int \frac{\partial \log p_\theta(x)}{\partial \theta_i} \frac{\partial \log p_\theta(x)}{\partial \theta_j} p_\theta(x) dx$

$\tilde{\nabla}_\theta = I^{-1} \frac{\partial}{\partial \theta}$

$$\theta_{t+\delta t} = \theta_t + \delta t \tilde{\nabla} J_{\theta_t}(\theta)|_{\theta=\theta_t}$$

$$= \theta_t + \delta t \int w(p_{\theta_t}[y : f(y) \leq f(x)]) \tilde{\nabla}_\theta \ln p_\theta(x) \; |_{\theta=\theta_t} p_{\theta_t}(x) dx$$

# Connexion with Optimization on Manifolds

## *Information Geometric Optimization*

☆ Monte Carlo approximation of the integral

$$\theta_{t+\delta t} = \theta_t + \delta t \int w(p_{\theta_t}[y : f(y) \leq f(x)])\tilde{\nabla}_\theta \ln p_\theta(x) \mid_{\theta=\theta_t} p_{\theta_t}(x)dx$$

Sample $X_i \sim p_{\theta_t}(x), i = 1, \ldots \lambda$

$$\theta_{t+1} = \theta_t + \delta t \frac{1}{\lambda} \sum_{i=1}^{\lambda} w_{rk(X_i)} \tilde{\nabla}_\theta \ln p_\theta(X_i)$$

For $p_\theta$ family of Gaussian distribution $\theta = (\mathbf{m}, \mathbf{C})$

CMA-ES with rank-mu update

Akimoto et al. *Bidirectional relation between CMA evolution strategies and natural evolution strategies*, 2010 PPSN XI

xNES

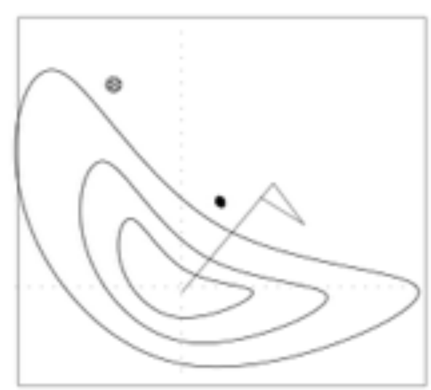
For $p_\theta$ family of Bernoulli distribution: PBIL (Baluja, 1994)

# Invariances

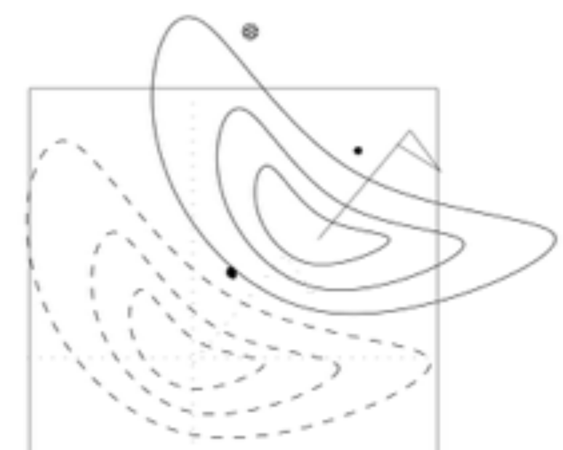
Invariance under monotonically increasing functions

*comparison-based*

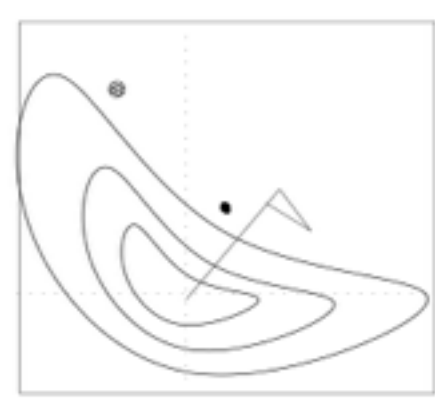
Translation invariance



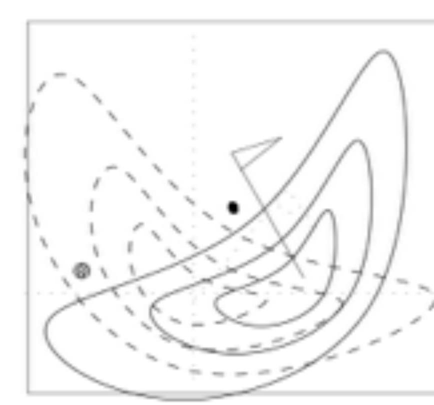
$$f(\boldsymbol{x}) \leftrightarrow f(\boldsymbol{x}-\boldsymbol{a})$$



Rotational invariance



$$f(\boldsymbol{x}) \leftrightarrow f(\mathbf{R}\boldsymbol{x})$$



Identical performance

# Why Invariance?

Empirical performance results

from test functions

from solved real world problems

are only useful if they do generalize to other problems

Invariance is a strong non-empirical statement about generalization

*generalizing performance from a single function to a whole class of functions*

# RECENT ADVANCES ON CONTINUOUS RANDOMIZED BLACK-BOX OPTIMIZATION

## Session I: Wednesday 29th October 14:30 - 16:00

A. Auger Recent Advances in Continuous Randomized Black-Box Optimization: an Overview.

N. Hansen CMA-ES: A Function Value Free Second Order Optimization Method.

I. Loshchilov LM-CMA-ES : an alternative to L-BFGS for large-scale black-box optimization.

## Session II: Thursday 30th October 11:00 - 12:30

T. Glasmachers Natural Evolution Strategies for Direct Search

D. Brockhoff Covariance Matrix Adaptation in Multiobjective Optimization

Y. Akimoto A linear time natural gradient algorithm for black-box optimization in high dimension.