# Introduction to Global Optimization

Rodolphe Le Riche[1] and Charlie Sire[2,1]

[1] CNRS LIMOS at Mines Saint-Etienne and UCA, France
[2] IRSN, France

as part of the Data science Major and Master "Maths in Action"
Mines Saint-Etienne, France

January 2021

# Course outline

These slides constitute a 12h introductory course on global optimization. The course starts with basic concepts specific to global optimization and different from those underlying local optimization algorithms. A selection of 6 algorithms is then presented: random search, randomly restarted local searches, simulated annealing, CMA-ES and Bayesian Optimization. This selection is meant to cover the main mechanisms behind global searches.

Pre-requisites are: linear algebra, basic probabilities and local optimization (gradient methods, necessary optimality conditions).

# Content

# Optimization = a quantitative formulation of decision

Optimization is a[1] way of mathematically modeling decision.

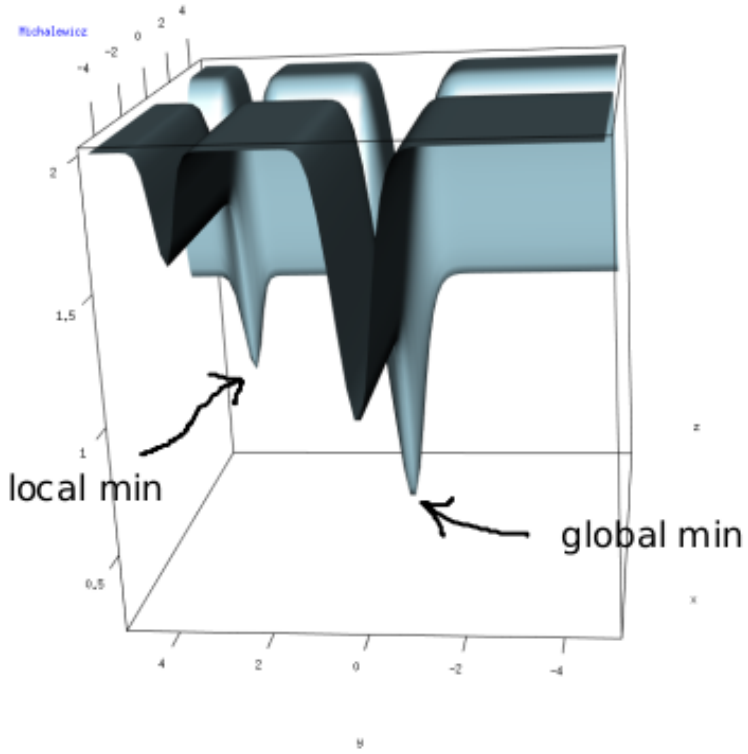
$$\min_{x \in \mathcal{S}} f(x)$$

- $x$ vector of decision parameters (variables) : dimensions, investment, tuning of a machine / program, ...
- $f(x)$ : decision cost $x$
- $\mathcal{S}$ : set of possible values for $x$, search space

[1] non unique, incomplete when considering human beings or life

# Local versus global optimum

$$\min_{x \in \mathcal{S} \subset \mathbb{R}^d} f(x)$$




$x^l \in \mathcal{S}$ is a local minimum of $f$ over $\mathcal{S}$ if there exists $\varepsilon > 0$ such that for any $x \in \mathcal{S}$ with $\|x - x^l\| \leq \varepsilon$, $f(x) \geq f(x^l)$
(it is a strict local minimum if $f(x) > f(x^l)$).

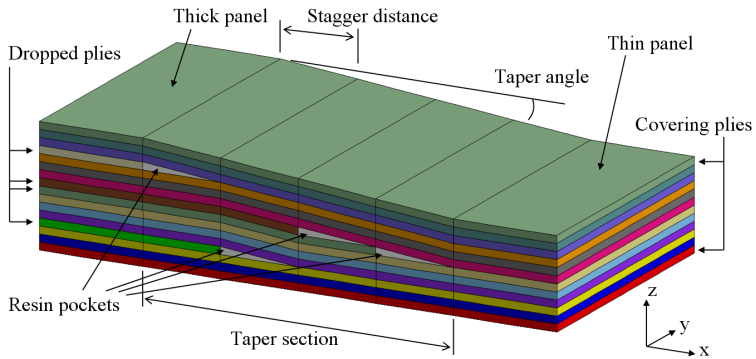
Goal of the class: learn strategies to increase the chances to locate a global optimum.

# Examples where global optimization is necessary I

In redundant structures, one of the component ($x$ composant) can take the role of another.

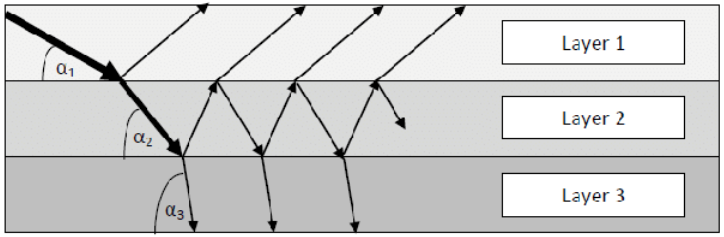
## Composite structures

$x$ is the orientation of the fibers within the plies of a composite laminate and the location where the plies are dropped. Many arrangements of the $x$'s have almost equivalent performances, leading to local optima (from [Irisarri et al., 2014]).



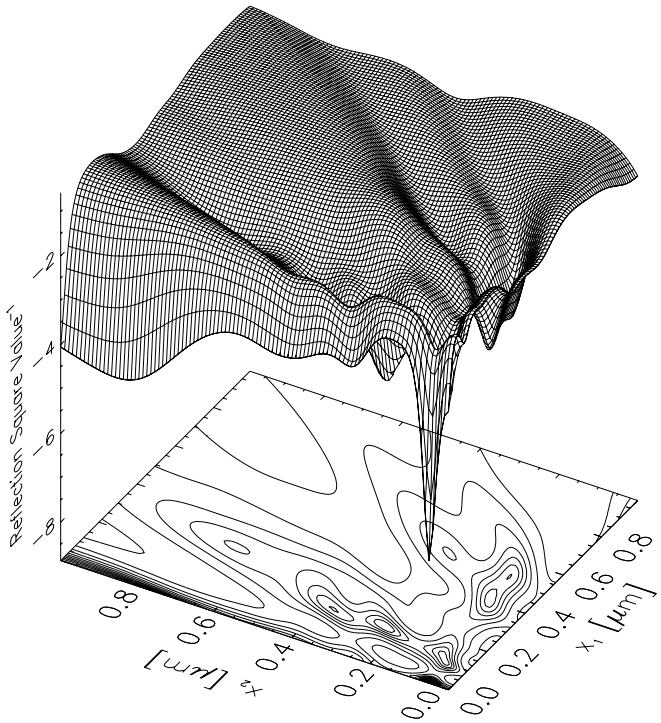
# Examples where global optimization is necessary II

## Optical filters




Root mean square distance to a target spectrum (from [Bäck et al., 1995])
Cross-section of the function w.r.t. 2 layer thicknesses. This function is multimodal, the higher dimensional function probably is.

Quiz 1: why $f(x_1, x_2)$, which is multimodal in terms of $x_1$ when $x_2$ is fixed, may not be multimodal w.r.t. $(x_1, x_2)$?

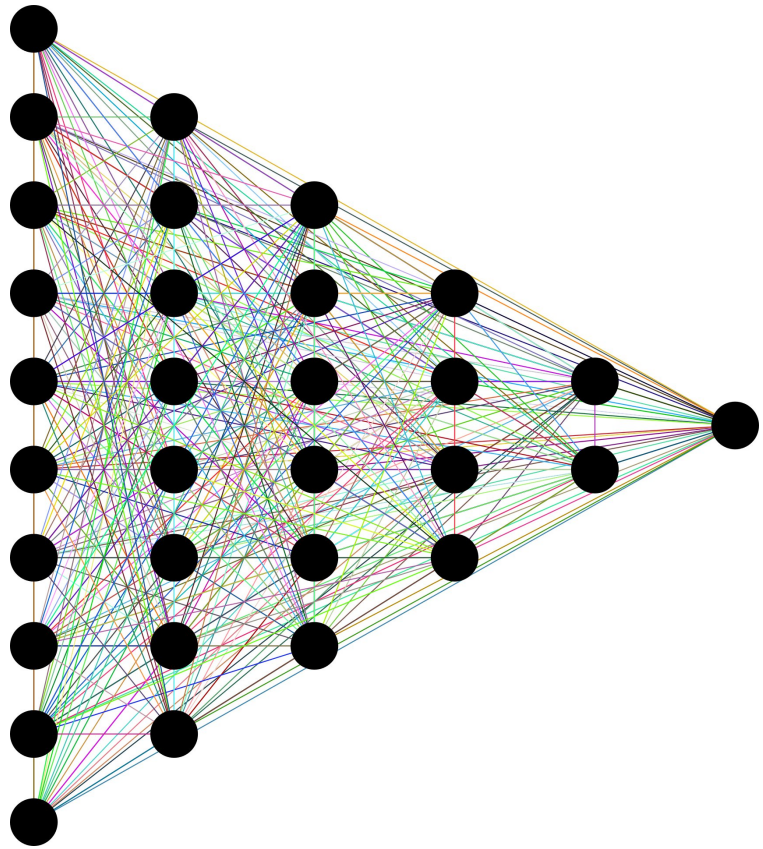


# Examples where global optimization is necessary III

## Neural networks

Very redundant structures with lots of symmetries, hence local optima.

Surprisingly, approximate convergence towards local optima works. An effect of high dimension and network flexibility?

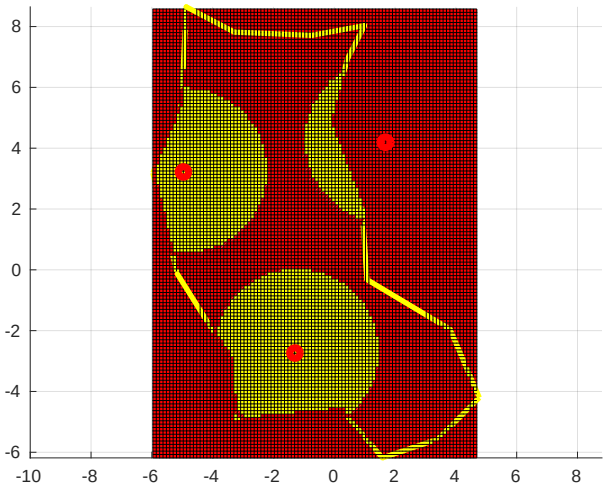


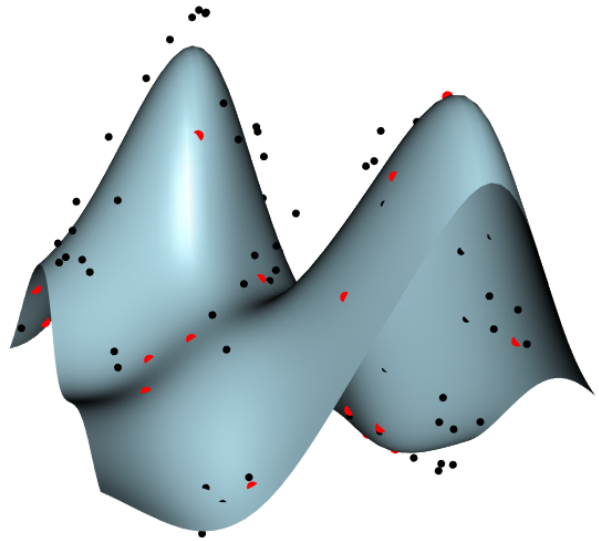
from techxplore.com/news/2020-11-neural-network.html, CC0 public domain

# Examples where global optimization is necessary IV

## Antennas positioning

Find the optimal location of a given number of antennas to cover a territory (the Loire Dept)



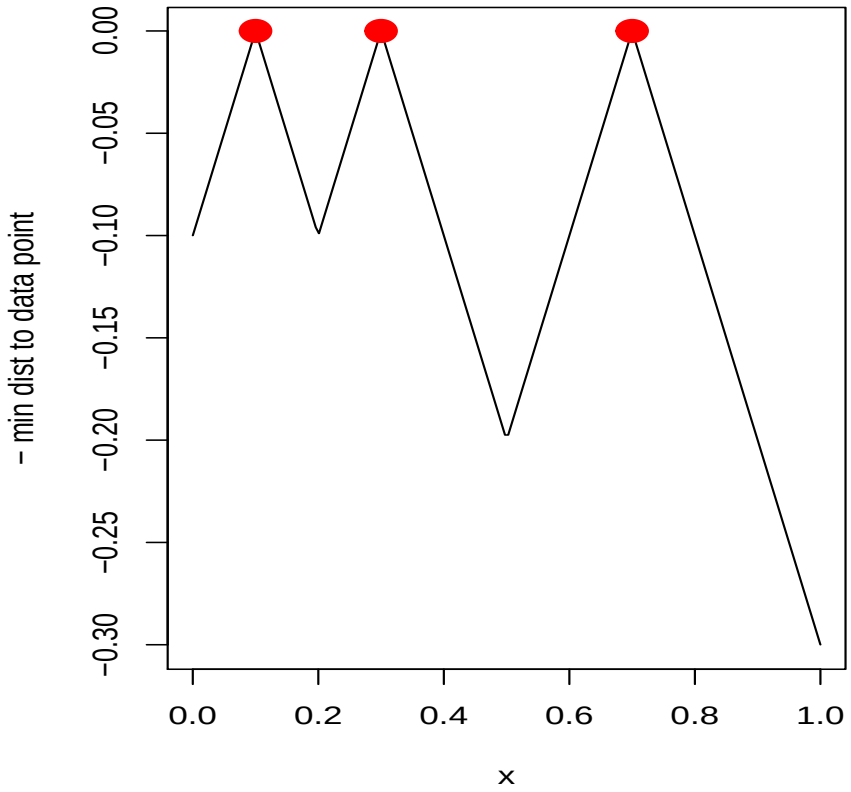
Approximation of the Loire surface covered by one antenna for varying locations, $S(x_1, x_2)$. The approximation (silver surface) is the mean of a kriging model learned from 20 examples (the red bullets). The black bullets are test points (not learned), providing some insights into the imperfections of the approximation.

# Examples where global optimization is necessary V

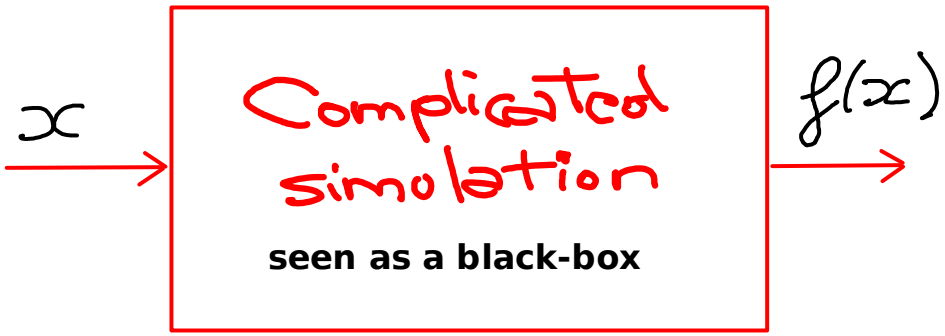
## Design of Experiments

As an archetypal example, find a point the farthest to a set of pre-existing points

$$\min_{x \in \mathcal{S}} \left( - \min_{x^i \in \text{DoE}} \|x - x^i\| \right)$$



# Examples where global optimization is necessary VI




Theoretically, global optimization is safer than local optimization for black-box functions.

In practice, global optimization is not always affordable.

# Global optimization as a metaphor

Let $f$ be the depth of a canal. To know if the canal is navigable, soundings are made to

find the minimal depth

$$f^\star = \min_{x \in \mathcal{S}} f(x)$$

find the location(s) of minimal depth

$$x^\star \in \arg\min_{x \in \mathcal{S}} f(x)$$

The motor vessel Sparna lists to its port side after taking on water in void spaces after running aground while transiting the Columbia River, 2016. U.S. Coast Guard Photo

Each sounding takes time. What is your search strategy?
Don't forget to use auxiliary information, e.g. former groundings.

# The exploration-intensification tradeoff

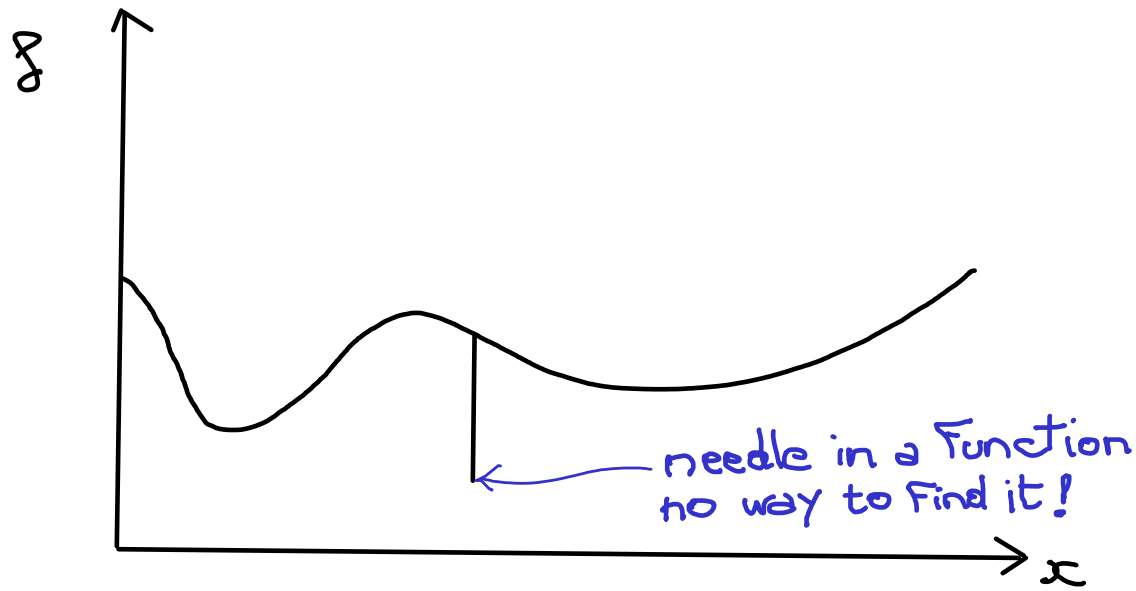
All global optimization algorithms strike a compromise between exploration and intensification.

- Intensification: use already calculated points to concentrate the search in high-return areas of the search space.
- Exploration: try the most unknown parts of the search space.

There is an infinite number of ways to handle this compromise.

# Essential global optima

Without restrictions, global optimization is a utopia: only look for essential global optima




$$x^\star \in \arg\min_{x \in \mathcal{S}} f(x)$$

$$\text{such that } \forall \varepsilon > 0 \ , \ \text{volume}[x' \in \mathcal{S} \mid f(x') < f(x) + \varepsilon] > 0$$

# Theoretically solvable problems : deterministic point of view

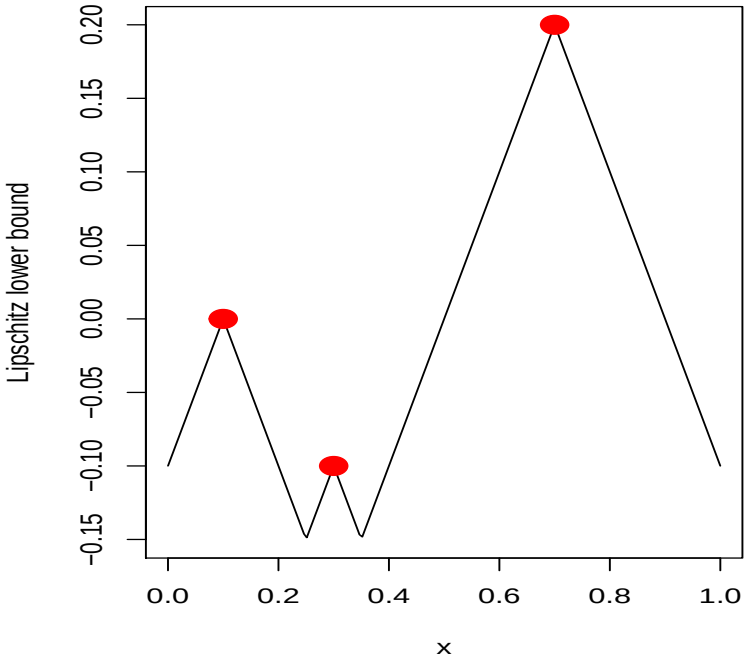
If $f$ has a known Lipschitz constant $L$,

$$\exists L \text{ such that } \forall x, x' \quad , \quad |f(x) - f(x')| \leq L\|x - x'\|$$

we can know how far the true solution is from already calculated points which tells us where to search and when to stop :

$$x^{t+1} = \arg\min_{x \in \mathcal{S}} \max_{i=1,\dots,t} f(x^i) - L\|x - x^i\|$$

Quiz 2: knowing $(x^i, f(x^i))$ , $i = 1, \dots, t$ and $L$, what is the accuracy with which the problem has been solved ? In other terms, give an upper-bound to $[\min_{i=1,\dots,t} f(x^i) - f(x^\star)]$.

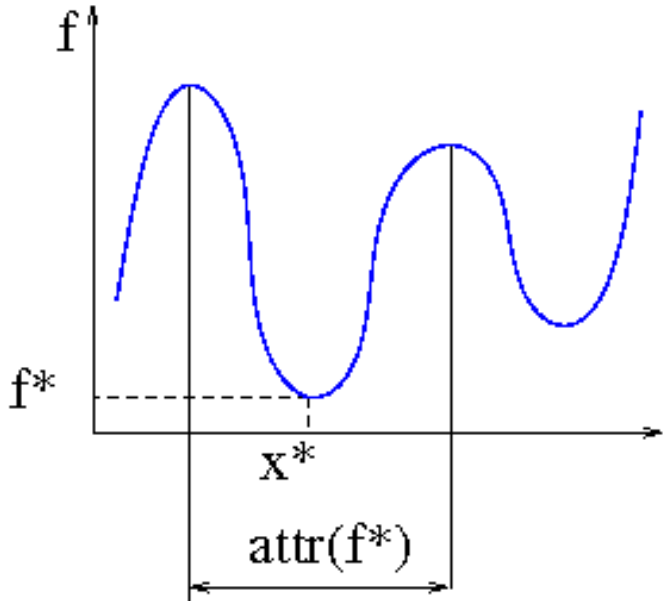


# Problems solvable in probability I

Let there be $k$ distinct local optima $x_1^\star, \ldots, x_k^\star$ such that $f^\star := f(x_1^\star) := f_1^\star \leq \ldots \leq f(x_k^\star) = f_k^\star$.

Gradient flow: set of points generated as solutions of $\dot{x} = -\nabla f(x)$ starting from a given $x$, i.e., set of iterates of an infinitesimal step gradient descent. Assume that $f$ and $\mathcal{S}$ are such that it always converges to a local optimum.

Basin of attraction, $\text{attr}(f_i^\star)$: set of points $x \in \mathcal{S}$ such that a gradient flow converges to $f(x) = f_i^\star$.



# Problems solvable in probability II

$p_i$, the probability that the gradient flow with random starting point converges to $f_i^\star$, $p_i = \text{vol}\left[\text{attr}\left(f_i^\star\right)\right] / \text{vol}\left[\mathcal{S}\right]$.

The difficulty of a problem (assuming the local search is perfect) can be measured with $p_1$:

- $p_1 = 1$: unimodal function.
- $p_1 \approx 1$: easy global problem, most local searches will solve it.
- $p_1 \geq \delta > 0$: problem solvable in probability because the probability of finding $f_1^\star \to 1$ as the number of restarts increases. The probability to have at least one search converge to $\text{attr}\left(f^\star\right)$ in $\mu$ repeats is $= 1 - (1 - p_1)^\mu$.
- $p_1 \approx 0$: unstable optimum, non essential. In practice, often avoided because it is not robust to a small error in $x^\star$.

# Content

# First, simple, algorithms I

We will now review a few basic global optimization algorithms by increasing order of complexity.

Words of caution:

- Although there are better optimization algorithms than others, there will never be an algorithm better over all possible problems, cf. No Free Lunch theorem [Wolpert and Macready, 1997]: when an algorithm improves on a subset of problems, it regresses on other subsets.
- There is an adequation between the problem and the algorithm. Auxiliary information may be used to select the algorithm (dimension, nature of the search space, budget of calls to $f$, noise affecting $f$, regularity of $f$).

# First, simple, algorithms II

- The problem formulation is a degree of freedom to change the problem and make it more amenable to optimization. Examples with shapes and PCA in [Gaudrie et al., 2020].
- A theoretically solvable problem may not be solvable in practice because of the cost of the search, in particular in high dimension ($d \gg 1$). In this case, "solving" the problem is finding the best possible solution.
- For a better understanding, look for the elements of the algorithms that contribute to the exploration or to the intensification of the search.

**Algorithm 1** Random search

**Require:** $x^{\text{LB}}$, $x^{\text{UB}}$, $t^{\max}$

$t \leftarrow 0$, $\widehat{f}^\star \leftarrow +\infty$

**while** $t < t^{\max}$ **do**

  $x' \leftarrow \mathcal{U}[x^{\text{LB}}, x^{\text{UB}}]$ {uniform law}

  calculate $f(x')$, $t \leftarrow t + 1$

  **if** $f(x') < \widehat{f}^\star$ **then**

    $\widehat{x}^\star \leftarrow x'$ , $\widehat{f}^\star \leftarrow f(x')$

  **end if**

**end while**

**return** $\widehat{x}^\star, \widehat{f}^\star$

Quiz 3: if $\mathcal{S} = [0,1]^5$ (hypercube in $d = 5$), and $x^\star = (0.2, \ldots, 0.2)$ what is the expected time until a point is sampled inside the cube of side 0.1 centered at $x^\star$ (i.e., $x^\star$ found with a 0.05 accuracy on each of its components)? What is the standard deviation of this time?

**Algorithm 2** Restarted local search

**Require:** $x^{\mathrm{LB}}$, $x^{\mathrm{UB}}$, $t^{\max}$, a `LOCAL_SEARCH` algorithm

$t \leftarrow 0$, $\widehat{f^\star} \leftarrow +\infty$

1: **while** $t < t^{\max}$ **do**
2: $x^{\mathrm{init}} \leftarrow \mathcal{U}[x^{\mathrm{LB}}, x^{\mathrm{UB}}]$ {uniform law}
3: $[x', f(x'), t'] \leftarrow \mathtt{LOCAL\_SEARCH}(x^{\mathrm{init}})$ {start from $x^{\mathrm{init}}$, $x'$ candidate local optimum, $f(x')$ its obj. function, $t'$ number of calls to $f$ done during the local search}
4: $t \leftarrow t + t'$
5: **if** $f(x') < \widehat{f^\star}$ **then**
6: $\widehat{x^\star} \leftarrow x'$ , $\widehat{f^\star} \leftarrow f(x')$
7: **end if**
8: **end while**
9: **return** $\widehat{x^\star}, \widehat{f^\star}$

It is often interesting to additionally save the candidate local optima $x'$.

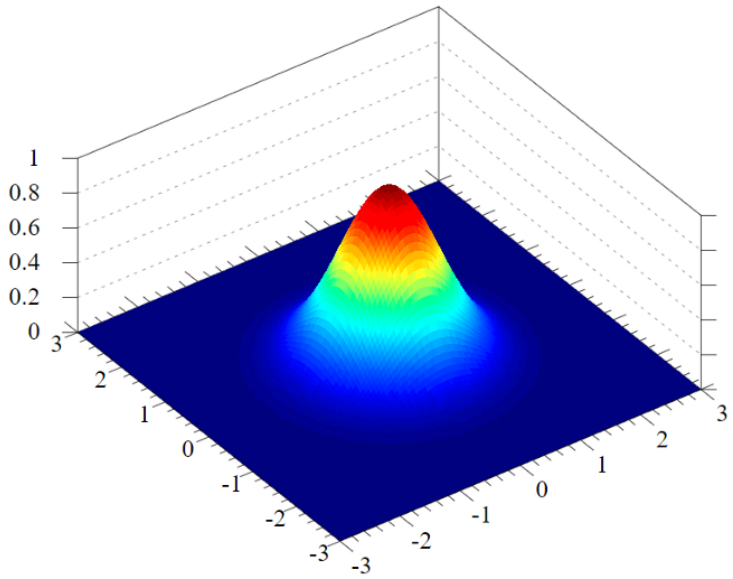
Quiz 4: Let $p_1$ be the probability to start a local search in the basin of attraction of the optimum $x^\star$, and $r$ be the number of restarts done. By how much does the probability of locating $x^\star$ increase if a $r+1$ restart is done? Why is it a case of diminishing returns?

**Algorithm 3** Isotropic ES-(1+1)

**Require:** $t^{\max}$, $x$ , $f(x)$ , $\sigma^2$

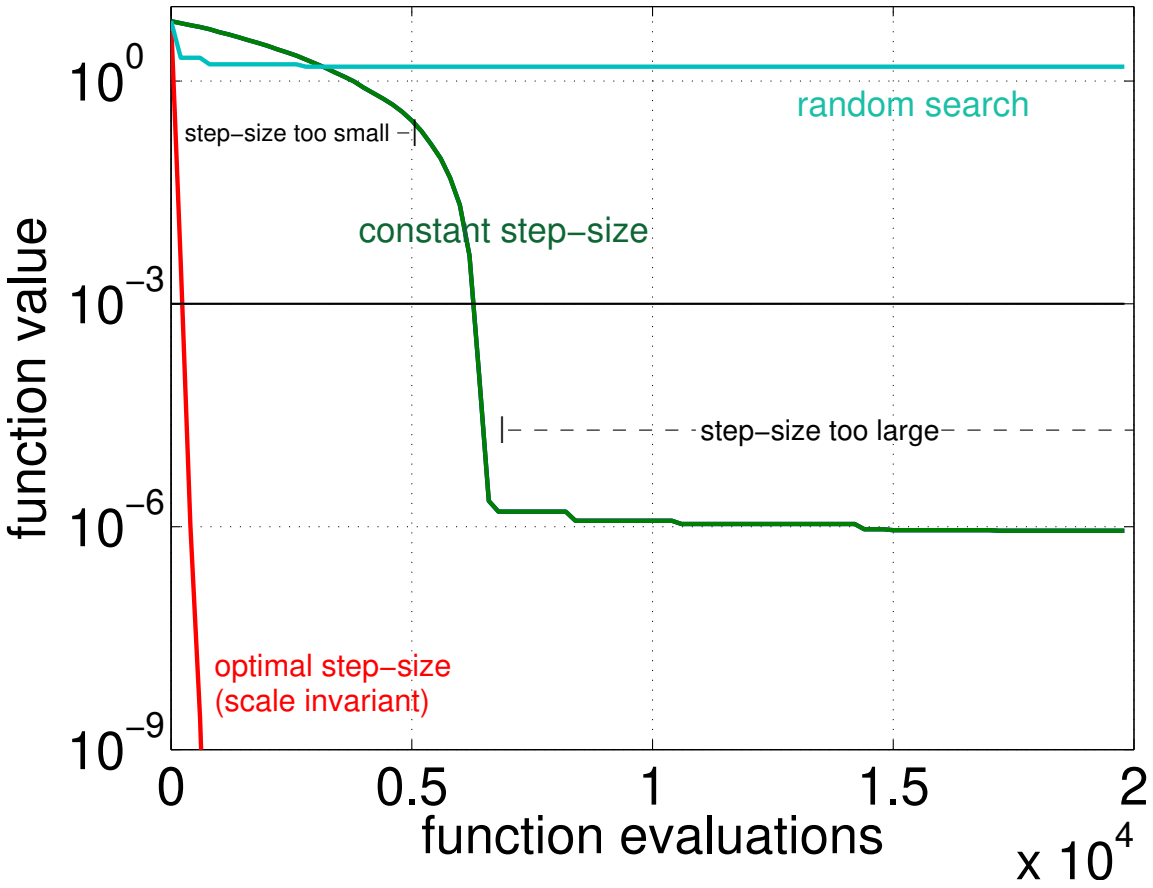
$t \leftarrow 1,\ m \leftarrow x,\ C \leftarrow \sigma^2 I$

1: **while** $t < t^{\max}$ **do**
2: $x' \leftarrow \mathcal{N}(m, C)$ {Gaussian law}
3: Calculate $f(x')$, $t \leftarrow t + 1$
4: **if** $f(x') < f(x)$ **then**
5: $x \leftarrow x'$ , $f(x) \leftarrow f(x')$
6: **end if**
7: $m \leftarrow x$ {update proposal pdf}
8: **end while**
9: **return** $x$ , $f(x)$




Simplified : no adaptation of $C$, neither the step size $\sigma$, nor the shape which remains isotropic (no priviledged search direction).

# Step adaptation is important




(1+1)-ES
(red & green)

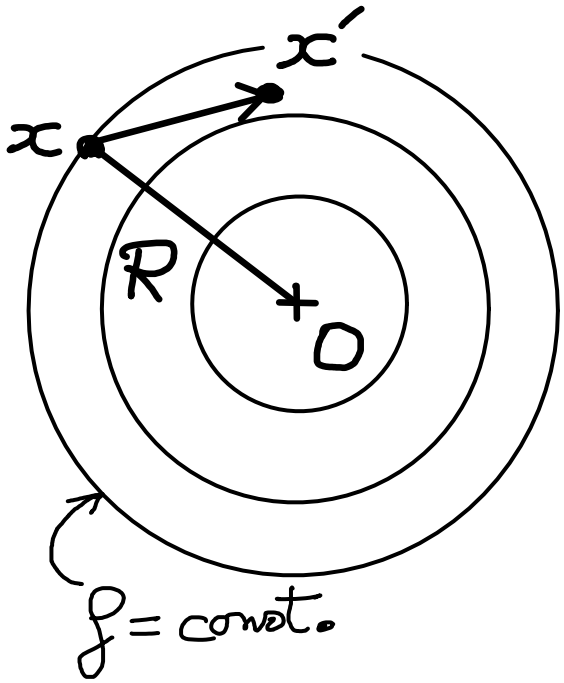
$$f(\boldsymbol{x}) = \sum_{i=1}^{n} x_i^2$$

in $[-2.2, 0.8]^n$
for $n = 10$

from [Auger and Dumas, 2017]

# Theoretical step size in ES-(1+1) I

Look at template function, $\min_{x \in \mathbb{R}^d} f(x)$ where $f(x) = \|x\|^2$.
Trivial problem but useful to understand phenomena occuring when $d$ changes.




$$X_i' = x_i + \sigma U_i \quad \text{where} \quad U_i \sim \mathcal{N}(0,1)$$

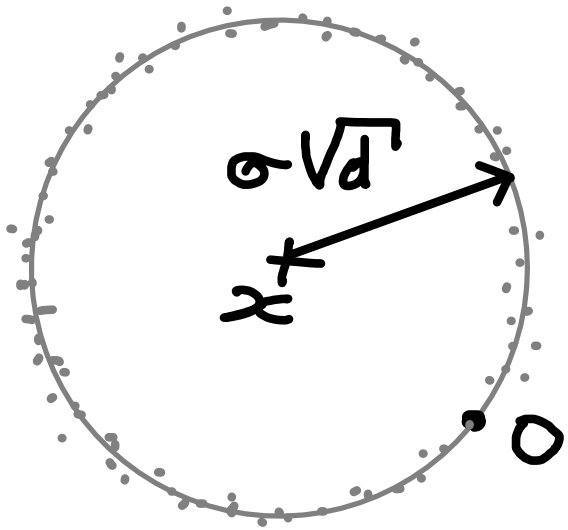
$$\|X' - x\|^2 = \sigma^2 \sum_{i=1}^{d} U_i^2$$

The last term follows a $\chi^2_d$.
When $d \geq 5$, it is approximated by $\mathcal{N}(d, 2d)$.

# Theoretical step size in ES-(1+1) II

$$\|X' - x\| = \sigma \sqrt{\sum_{i=1}^{d} U_i^2}$$

$$\overset{d \geq 5}{\rightsquigarrow} \sigma \mathcal{N}(\sqrt{d}, 1/2)$$



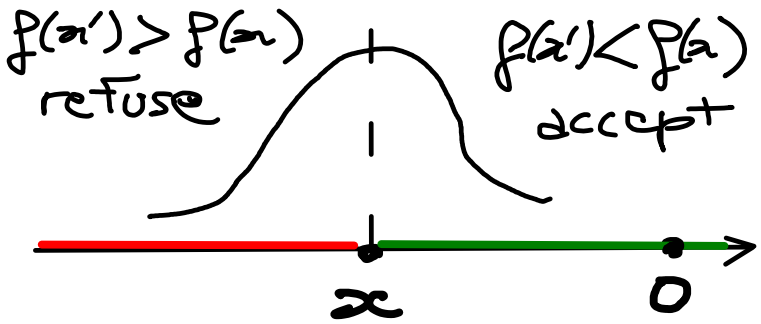

i.e., samples concentrate on a sphere centered on $x$ and of radius $\sigma\sqrt{d}$. Useful to choose $\sigma$? Sure, set $\sigma\sqrt{d} = R$, i.e., $\boxed{\sigma = R/\sqrt{d}}$.

But this step size does not account for the acceptance / refusal in the ES-(1+1) algorithm, and the acceptance rate at given $\sigma$ decreases with $d$ ↗.

# Theoretical step size in ES-(1+1) III

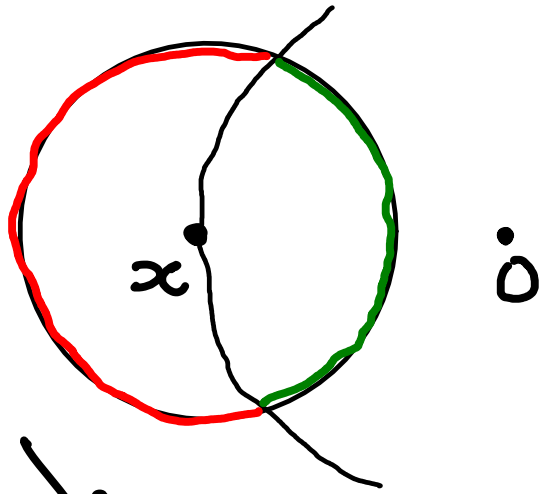
$d=1$

$P_{accept} = 1/2$




$d=2$

$P_{accept} < 1/2$




... $d \nearrow$, $P_{accept} \searrow$

# Theoretical step size in ES-(1+1) IV

## Optimal $\sigma$

Change in distance to 0 (random variable): $D_2 = \|x\|^2 - \|X'\|^2$

$$\begin{aligned} D_2 =& \|x\|^2 - \|x + \sigma U\|^2 \\ =& \|x\|^2 - (x + \sigma U)^\top (x + \sigma U) \\ =& -2\sigma x^\top U - \sigma^2 U^\top U \\ \overset{d\geq 5}{\approx}& -2\sigma x^\top U - \sigma^2 d \\ \Rightarrow D_2 \sim& \mathcal{N}(-\sigma^2 d, 4\sigma^2 R^2) \end{aligned}$$

ES-(1+1) has a test that prevents $D_2$ from being negative, the

# Theoretical step size in ES-(1+1) V

progress is $I_2 = \max(0, D_2)$. The mean progress is

$$\mathbb{E}(I_2) = \mathbb{E}(\max(0, D_2)) = \int_0^{+\infty} i_2 p(i_2) di_2$$

Change of variable : $j = \frac{i_2 + \sigma^2 d}{2\sigma R}$ , $j \sim \mathcal{N}(0,1)$

$$\mathbb{E}(I_2) = \int_{\frac{\sigma d}{2R}}^{+\infty} (2\sigma R j - \sigma^2 d)\ \mathrm{pdf}_{\mathcal{N}}(j) dj$$

$$= 2\sigma R\ \mathrm{pdf}_{\mathcal{N}}(\frac{\sigma d}{2R}) - \sigma^2 d \left[1 - \mathrm{cdf}_{\mathcal{N}}(\frac{\sigma d}{2R})\right]$$

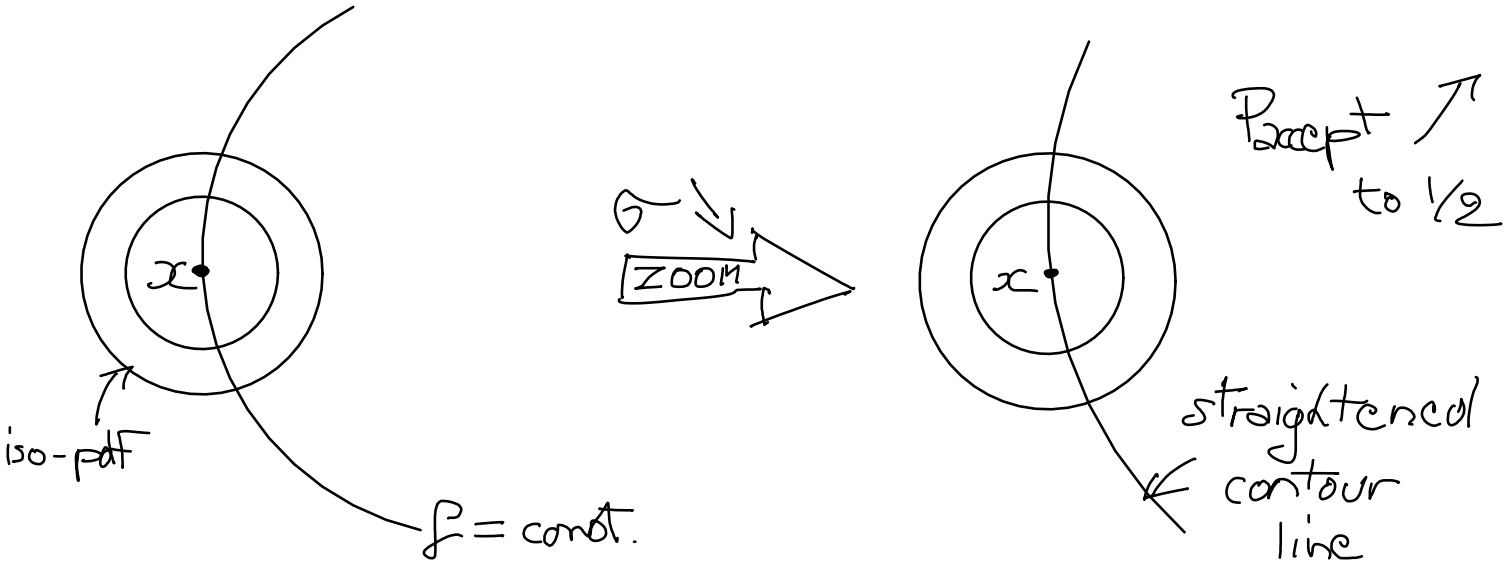
$$\frac{d}{R^2}\mathbb{E}(I_2) = 2\frac{\sigma d}{R}\ \mathrm{pdf}_{\mathcal{N}}(\frac{\sigma d}{2R}) - \frac{\sigma^2 d^2}{R^2}\left[1 - \mathrm{cdf}_{\mathcal{N}}(\frac{\sigma d}{2R})\right]$$

# Theoretical step size in ES-(1+1) VI

$\frac{d}{R^2}\mathbb{E}(I_2)$ is a function of $\tilde{\sigma} = \frac{\sigma d}{R}$.
Its maximizer is $\tilde{\sigma}^\star = \arg\max_{\tilde{\sigma}} \frac{d}{R^2}\mathbb{E}(I_2) \approx 1.22 = \frac{\sigma^\star d}{R}$

$$\Rightarrow \boxed{\sigma^\star = 1.22\frac{R}{d}}$$

which is more cautious than the $\mathcal{O}(1/\sqrt{d})$ first estimate.



# Theoretical step size in ES-(1+1) VII

## Extension to quadratic functions

$$f(x) = \frac{1}{2} x^\top H x$$

$H$ Hessian strictly positive definite. The previous optimal step size calculation can be extended.
Eigendecomposition of $H = PD^2P^\top$, $P$ matrix of eigenvectors as columns, $P^\top P = PP^\top = I$, $D^2$ diagonal matrix of eigenvalues.

$$f(x) = \frac{1}{2} x^\top PDDP^\top x = \frac{1}{2} y^\top y = f(y)$$
$$\text{where} \quad y = DP^\top x \iff x = PD^{-1} y$$

# Theoretical step size in ES-(1+1) VIII

$f(y)$ is a sphere, therefore an optimal[2] step distribution in the $y$-space is $Y' - y \sim \mathcal{N}(0, \sigma^{\star 2} I)$ where

$$\sigma^\star = 1.22 \frac{\|y\|}{d} = 1.22 \frac{\sqrt{x^\top P D^2 P^\top x}}{d} = 1.22 \frac{\sqrt{x^\top H x}}{d}$$

Translating back into the $x$-space by multiplying by $PD^{-1}$,

$$X' - x = PD^{-1}(Y' - y) \sim \mathcal{N}(0, \sigma^{\star 2} P D^{-2} P^\top) \equiv \mathcal{N}(0, \sigma^{\star 2} H^{-1})$$

Note that the shape of the covariance matrix is given by the inverse Hessian.

[2] optimal w.r.t. $\sigma$ and position invariant on the contour $f$ =const. See [Rudolph, 1992]

# Content

# Definition

- Let's consider $\{X^{(t)}\}_{t=0,1..}$ a sequence of random variables in a state space $\mathcal{S}$
- $\{X^{(t)}\}_{t=0,1..}$ is a Markov Chain if
  $\forall t, \forall x^{(0)}, .., x^{(t-1)} : p(x^{(t)} \mid x^{(0)}, .., x^{(t-1)}) = p(x^{(t)} \mid x^{(t-1)})$
- Then we have a simpler expression of the joint density of
  $(X^{(0)}, X^{(1)}, .., X^{(t)}) : \; p(x^{(t)}, .., x^{(0)}) = p(x^0) \prod_{t'=1}^{t} p(x^{(t')} \mid x^{(t'-1)})$

The idea is that $\mathcal{S}$ is also our search space in which we will investigate the argmin of the function $f$ with Markov Chains

# Discrete states space

- Now we consider $card(\mathcal{S}) < \infty : \mathcal{S} = \{1, ..., s\}$
- $X^{(t)} = j$ for $j \in \mathcal{S}$ means the process is in state $j$ at time $t$
- Let's define the one-step transition probability from state i at time $t$ to state j at time $t+1$ : $p_{ij}^{(t)} = \mathbb{P}(X^{(t+1)} = j \mid X^{(t)} = i)$
- If $\forall t, p_{ij}^{(t)} = p_{ij}^{(0)} = p_{ij}$ : the Markov Chain is called time-homogeneous

In the following we will consider time homogeneous Markov Chains

# Transition probability matrix

For a time-homogeneous Markov Chain, we can define a transition probability matrix $\mathbf{P} = (p_{ij})_{i,j \in 1,..,s}$
Each row of the matrix must sum to one : $\sum_{j=1}^{s} p_{ij} = 1$

# Marginal probability

- We denote $\pi_i^{(t)}$ the marginal probability that $X^{(t)} = i$
- Then we can define $\boldsymbol{\pi}^{(t)} = (\pi_i^{(t)})_{i \in \mathcal{S}}$ the vector probability (the sum is 1)
- $\forall j \in \mathcal{S}, \pi_j^{(t+1)} = \mathbb{P}(X^{(t+1)} = j) = \sum_{i=1}^{s} \mathbb{P}(X^{(t+1)} = j \mid X^{(t)} = i)\mathbb{P}(X^{(t)} = i) = \sum_{i=1}^{s} p_{ij}\pi_i^{(t)} = [\boldsymbol{\pi}^{(t)^T}\mathbf{P}]_j$

$\Longrightarrow \boldsymbol{\pi}^{(t+1)} = \boldsymbol{\pi}^{(t)^T}\mathbf{P}$

$\Longrightarrow \boldsymbol{\pi}^{(t+n)} = \boldsymbol{\pi}^{(t)^T}(\mathbf{P})^n$

A time-homogeneous Markov Chain is then completely defined by :

- The state space $\mathcal{S}$
- The starting distribution $\boldsymbol{\pi}^{(0)}$
- The transition probability matrix $\mathbf{P}$

# Stationary distribution

A stationary distribution $\boldsymbol{\pi}$ for $\mathbf{P}$ is a vector that satisfies the following properties :

- $0 \leq \pi_i \leq 1$
- $\sum_i \pi_i = 1$
- $\boldsymbol{\pi}^T \mathbf{P} = \boldsymbol{\pi}^T$

If $\{X^{(t)}\}$ follows a stationary distribution, then the marginal distributions of $X^{(t)}$ and $X^{(t+1)}$ are identical

# Irreducibility and Periodicity

- A Markov Chain is irreducible if any state j can be reached from any state i in a finite number of steps, ie
  $\forall i, j \in \mathcal{S}^2, \forall t, \exists m > 0, \mathbb{P}(X^{(t+m)} = j | X^{(t)} = i) > 0$
- A state $i$ is said to have period $k$ if any return to state i must occur in multiples of k time steps. Formally we have
  $k = gcd\{n > 0 | \mathbb{P}(X^{(n)} = i \mid X^0 = i) > 0\}$
  A state i is aperiodic if $k = 1$ and the Markov Chain is aperiodic if all states are aperiodic.

gcd : greatest common divisor

# Reversibility

A Markov chain is reversible if there exists a distribution $\boldsymbol{\pi}$ which satisfies the detailed balance conditions: $\forall i, j, \pi_i p_{ij} = \pi_j p_{ji}$
Any distribution satisfying detailed balance is a stationary distribution :

# Reversibility

A Markov chain is reversible if there exists a distribution $\boldsymbol{\pi}$ which satisfies the detailed balance conditions: $\forall i, j, \pi_i p_{ij} = \pi_j p_{ji}$
Any distribution satisfying detailed balance is a stationary distribution :

$$\begin{aligned}
\forall i, j, \pi_i p_{ij} = \pi_j p_{ji} &\Rightarrow \forall j \sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} \\
&\Rightarrow \forall j, \sum_i \pi_i p_{ij} = \pi_j \\
&\Rightarrow \forall j, [\boldsymbol{\pi}^T \mathbf{P}]_j = \pi_j
\end{aligned}$$

# Convergence

If $\{X^{(t)}\}$ is an irreducible and aperiodic Markov Chain in a finite state space $\mathcal{S}$ with stationary distribution $\boldsymbol{\pi}$, then :

- $\boldsymbol{\pi}$ is the unique stationary distribution
- $\forall i, j \in \mathcal{S}, \mathbb{P}(X^{(t)} = i \mid X^0 = j) \underset{t \to \infty}{\longrightarrow} \pi_i$

As $\forall i \in \mathcal{S}, \mathbb{P}(X^{(t)} = i) = \sum_{j \in \mathcal{S}} \mathbb{P}(X^{(t)} = i \mid X^{(0)} = j)\mathbb{P}(X^{(0)} = j)$,
We have the following convergence : $\forall i \in \mathcal{S}, \mathbb{P}(X^{(t)} = i) \underset{t \to \infty}{\longrightarrow} \pi_i$

# Importance of aperiodicity

Let's show an example where a Markov chain is irreducible with stationary distribution $\pi$ but is not aperiodic, and doesn't converge.
Let's define $(X^{(t)})_{t=0,1..}$ as follows :

- $\mathcal{S} = 1, 2, 3$
- $\pi^{(0)} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T$
- $p_{1,2} = p_{3,1} = p_{2,3} = 1$

Then $\pi = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}^T$ is stationary here, the chain is irreducible but all the states have period 3.
And we have $X^{(t)} = 1$ whenever $t$ is a multiple of 3, so $\mathbb{P}(X^{(t)} = 1)$ oscillates between 0 and 1, and there $X^{(t)}$ doesn't converge.
Quiz 5: We just showed that irreducibility and a stationary distribution does not guarantee convergence. However, is aperiodicity a necessary condition for convergence ?

# Continuous states space

Comparable results hold for a continuous state space

- In the continuous case, a time-homogeneous Markov Chain is defined by the transition kernel $p(x, x') = p_{X^{(t+1)}|X^{(t)}}(x' \mid x)$
- The density $\pi$ is stationary for the Markov Chain with kernel $p(x, x')$ if : $\forall x, x' \in \mathcal{S}, \pi(x') = \int p(x, x')\pi(x)dx$
- Under analogous conditions, convergence is expressed as follows : $\forall x \in supp(\pi)$, for all measurable $A \subset \mathcal{S}$,

$$\mathbb{P}(X^{(t)} \in A \mid X^{(0)} = x) \underset{t \to \infty}{\longrightarrow} \pi(A)$$

# Content

# Objective

Idea : General method to construct a Markov Chain that converges to a target distribution $\pi$ defined here as follows :

$$\pi(x) = \frac{1}{K} exp(-\frac{f(x)}{T})$$

where $K = \int_{\mathcal{S}} exp(-\frac{f(x)}{T})dx$ is an unknown normalizing constant and $T$ is a positive "temperature" scalar

- Why this target distribution ? As $T$ tends to 0, $\pi$ becomes a Dirac with all the density mass at $\mathcal{S}_{opt} = \underset{x \in \mathcal{S}}{argmin}\, f(x)$
- Why not sampling directly $\pi$ ? Constant $K$ very hard or impossible to compute

Remark : Markov Chain Monte Carlo (MCMC) can be related to other objectives with other hard to compute target distributions

# Proof of convergence when $T \to 0$

Let $x^\star \in \mathcal{S}_{opt}$,
$\forall T, \forall x \in \mathcal{S}, \pi(x) > 0$ and $\frac{\pi(x)}{\pi(x^\star)} = exp(-\frac{f(x)-f(x^\star)}{T})$
Then

$$\forall x \in \mathcal{S} \setminus \mathcal{S}_{opt}, \frac{\pi(x)}{\pi(x^\star)} \underset{T \to 0}{\to} 0$$

$$\forall x \in \mathcal{S}_{opt}, \frac{\pi(x)}{\pi(x^\star)} \underset{T \to 0}{\to} 1$$

The density mass is concentrated at $\mathcal{S}_{opt}$

# Details with $\mathcal{S}$ finite

More precisely, with $\mathcal{S} = \{1, ..., s\}$,

$$\sum_{i=1}^{s} \pi_i = 1 \Rightarrow \sum_{i=1}^{s} \frac{\pi_i}{\pi_{i^\star}} = \frac{1}{\pi_{i^\star}}$$

$$\Rightarrow \sum_{i \in \mathcal{S}_{opt}} \frac{\pi_i}{\pi_{i^\star}} + \sum_{i \in \mathcal{S} \setminus \mathcal{S}_{opt}} \frac{\pi_i}{\pi_{i^\star}} = \frac{1}{\pi_{i^\star}}$$

$$\Rightarrow card(\mathcal{S}_{opt}) + \sum_{i \in \mathcal{S} \setminus \mathcal{S}_{opt}} \frac{\pi_i}{\pi_{i^\star}} = \frac{1}{\pi_{i^\star}}$$

$$\Rightarrow \begin{cases} \pi_{i^\star} \underset{T \to 0}{\to} \frac{1}{card(\mathcal{S}_{opt})} \\ \pi_i \underset{T \to 0}{\to} 0 \text{ if } i \notin \mathcal{S}_{opt} \end{cases}$$

# Algorithm (General Case)

**Algorithm 4** Metropolis Hastings

**Require:** $\pi$ (a target distribution), $h$ (a proposal distribution)

$x^0$ with $\pi(x^0) > 0$

$t \leftarrow 0$

1: Sample a candidate value $\tilde{x}$ from the proposal distribution $h(. \mid x^{(t)})$

2: Compute the Metropolis Hastings ratio $R(x^{(t)}, \tilde{x}) = \frac{\pi(\tilde{x})h(x^{(t)}|\tilde{x})}{\pi(x^{(t)})h(\tilde{x}|x^{(t)})}$

3: Generate $x^{(t+1)}$ as follows :

$$x^{(t+1)} = \begin{cases} \tilde{x} & \text{with probability } min(1, R(x^{(t)}, \tilde{x})) \\ x^{(t)} & \text{with probability } 1 - min(1, R(x^{(t)}, \tilde{x})) \end{cases}$$

Remark : Don't need to compute Normalizing constant $K$ in $R(x^{(t)}, \tilde{x})$

# Properties

- Clearly, a chain constructed via the Metropolis Hasting algorithm is Markov since $X^{(t+1)}$ is only dependent on $X^{(t)}$
- Whether the chain is irreducible and aperiodic depends on the choice of proposal distribution; the user must check these conditions for any implementation.
- If this check confirms irreducibility and aperiodicity, then the chain generated by the Metropolis Hastings algorithm has a unique limiting stationary distribution, which is the target distribution $\pi$ .

# Proof : Discrete states space

We denote here the target distribution $p = \pi$ And we introduce the matrix $\mathbf{H}$ with $\forall i, j, H_{ij}$ the probability of proposing the candidate value $j$ for $X^{(t+1)}$ knowing $X^{(t)} = i$
Let's show that $\pi$ satisfies the detailed balanced condition

# Proof : Discrete states space

We denote here the target distribution $p = \boldsymbol{\pi}$ And we introduce the matrix $\mathbf{H}$ with $\forall i, j, H_{ij}$ the probability of proposing the candidate value $j$ for $X^{(t+1)}$ knowing $X^{(t)} = i$
Let's show that $\boldsymbol{\pi}$ satisfies the detailed balanced condition
The transition probability matrix is $\mathbf{P}$ with $p_{ij} = min(1, \frac{\pi_j H_{ji}}{\pi_i H_{ij}}) H_{ij}$
We assume $\pi_j H_{ji} \leq \pi_i H_{ij}$ (without loss of generality) :

$$p_{ij} = \frac{\pi_j}{\pi_i} H_{ji}$$
$$p_{ji} = H_{ji}$$

Then $\pi_j p_{ji} = \pi_i p_{ij}$

# Proof : Continuous states space (1/2)

$X^{(t)} \sim \pi(x)$

The transition kernel is $p(x, x') = min(1, \frac{\pi(x')h(x|x')}{\pi(x)h(x'|x)})h(x' \mid x)$

We assume $\pi(x')h(x \mid x') \leq \pi(x)h(x \mid x')$ (without loss of generality)

:

$$p(x, x') = \frac{\pi(x')}{\pi(x)} h(x' \mid x)$$

$$p(x', x) = h(x' \mid x)$$

Then $p(x', x)\pi(x') = p(x, x')\pi(x)$ (reversibility in continuous case)

# Proof : Continuous states space (2/2)

Then

$$\begin{aligned}\int p(x', x)\pi(x')dx' &= \int p(x, x')\pi(x)dx' \\ &= \pi(x)\int p(x, x')dx' \\ &= \pi(x)\end{aligned}$$

# Proposal distribution

A well-chosen proposal distribution produces candidate values that cover the support of the stationary distribution in a reasonable number of iterations and produces candidate values that are not accepted or rejected too frequently:

- If the proposal distribution is too diffuse relative to the target distribution, the candidate values will be rejected frequently and thus the chain will require many iterations to adequately explore the space of the target distribution.
- If the proposal distribution is too focused (e.g., has too small a variance), then the chain will remain in one small region of the target distribution for many iterations while other regions of the target distribution will not be adequately explored.

# Gaussian proposal distribution

$h(. \mid x^{(t)})$ density of $\mathcal{N}(x^{(t)}, \sigma^2)$ often used

- Irreducibility is true only on the support of $\pi$ because all $x$ with $\pi(x) = 0$ would be rejected
- But $\forall t, x^{(t)} \in supp(\pi)$
- Then, we can consider $\mathcal{S} = supp(\pi)$ and we have convergence on $\mathcal{S}$

Even if $x^{(0)} \notin supp(\pi)$, as the support of $h(. \mid x^{(t)})$ is infinite, $\mathbb{P}(\{\exists t_1, X^{(t_1)} \in supp(\pi)\}) = 1$

# Content

# Principle of Simulated Annealing

- Apply Metropolis-Hastings algorithm with $\pi(x) = \frac{1}{K} exp(-\frac{f(x)}{T})$ for a set of decreasing temperatures
- Choose a symmetric proposal distribution $h(. \mid x)$ ie $\forall x, x' \in \mathcal{S}, h(x \mid x') = h(x' \mid x)$
- Metropolis Hastings ratio is $R(x, x') = exp(\frac{f(x)-f(x')}{T})$
- Then :
  - If $f(x') \leq f(x) : R(x, x') \geq 1 \Rightarrow$ acceptance
  - Else

    Large T $\Rightarrow R(x, x') \simeq 1 \Rightarrow$ Exploration
    Small T $\Rightarrow R(x, x') << 1 \Rightarrow$ Intensification

# Algorithm

**Algorithm 5** Simulated Annealing

**Require:** $x, T_0, L_0$

$k \leftarrow 0$

**while** $T_k \not\approx 0$ **do**
  **for** $l = 0$ to $L_k$ **do**
    Sample a solution $\tilde{x}$ from the proposal distribution $h(. \mid x)$
    **if** $f(\tilde{x}) < f(x)$ **then**
      $x \leftarrow \tilde{x}$ ($\tilde{x}$ becomes the current solution)
    **else**
      $x \leftarrow \tilde{x}$ with probability $exp(\frac{f(x)-f(\tilde{x})}{T_k})$
    **end if**
  **end for**
  $k \leftarrow k + 1$
  Compute $L_k, T_k$
**end while**

# Discussion about the parameters

There are several parameter involved :

- The length of thermal equilibrum chain $L_k$
  Number of proposed transitions, Number of accepted transitions, combinations of both
- The initial temperature
  Large enough so that acceptance rate $\simeq 1$
- The cooling schedule (see next slide)
- The stopping criteria
  For instance when current solution no longer changes from one iteration to the next during a sufficiently large number of iterations

# Cooling schedule

Decreasing speed must not be too fast to avoid remaining in a local minimum

- Logarithm schedule $T_k = \frac{T_0}{1+ln(1+k)}$

Guarantee convergence but too slow in practice

- Geometric schedule : $T_k = \alpha^k T_0$

Most typical values of $\alpha$ between 0.8 and 0.99

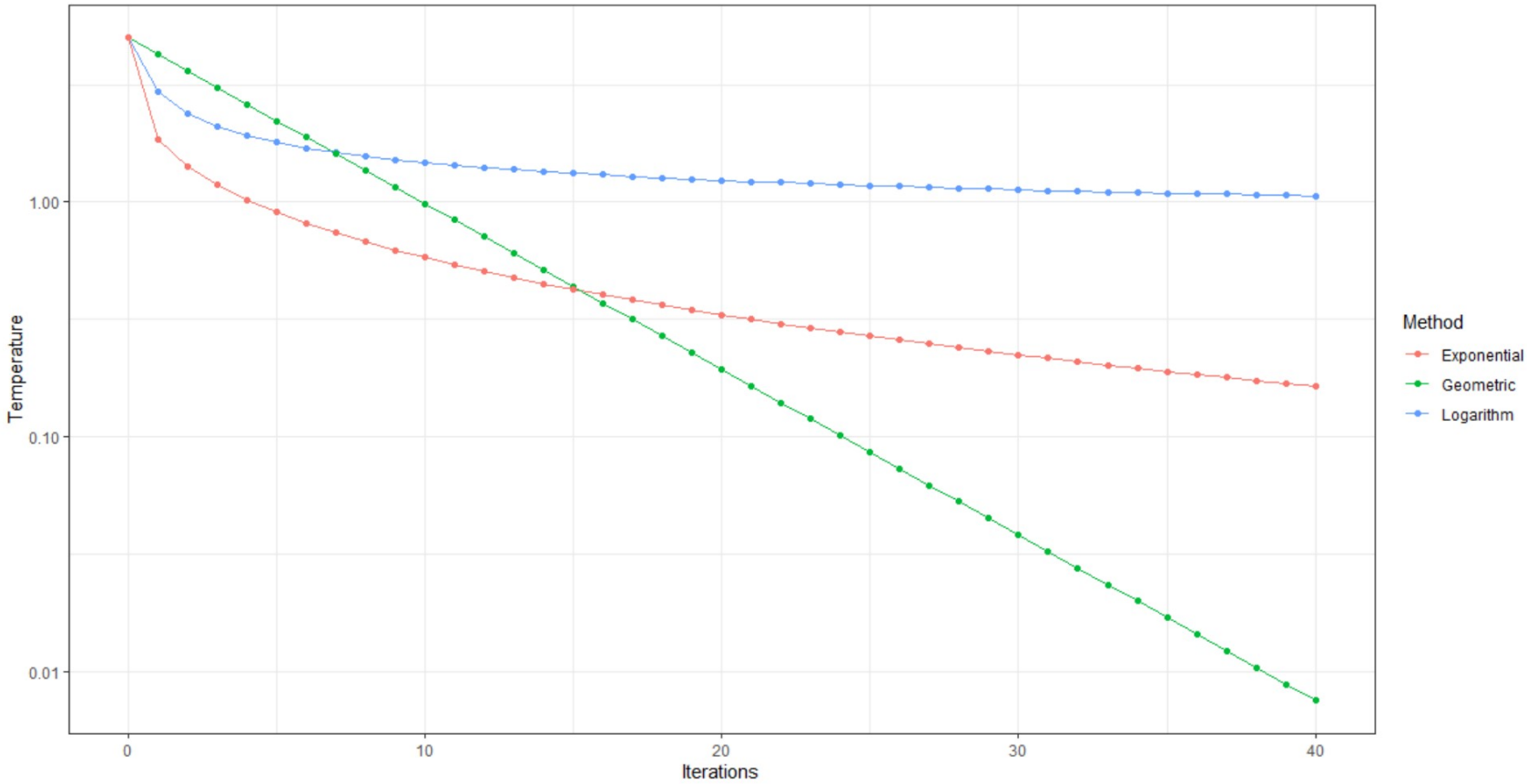
- Exponential schedule : $T_k = T_0 exp(-\alpha k^{\frac{1}{N}})$

Very fast during the first iterations

Quiz 6 As when $T$ tends to 0, $\pi(x) = \frac{1}{K} exp(-\frac{f(x)}{T})$ becomes a Dirac with all the density mass at $\mathcal{S}_{opt} = \underset{x \in \mathcal{S}}{argmin}\, f(x)$, why not starting directly with $T \simeq 0$ and build a Markov Chain converging to a random variable of density $\pi$ ?

# Cooling schedules comparison

Plot with $\alpha = 0.85$ for the geometric schedule and $\alpha = 1$, $N = 3$ for the exponential schedule



# Physical interpretation

Principle of physical annealing :

- A cristalline solid is heated : molecular structure is weaker and is more susceptible to change
- Then it's allowed to cool very slowly until it achieves its most regular possible crystal lattice configuration (its minimum lattice energy state)
- If cooling is abrupt, the solid will be found in a metastable state with non-minimal energy
- At thermal equilibrium, the probability that a system is in a macroscopic configuration $i$ with energy $E_i$ is given by the Boltzmann distribution $\pi_i = \frac{\exp\left(-\frac{E_i}{k_B T}\right)}{K}$, with K the normalizing constant and $k_B$ the Boltzmann constant

# Performance of simulated annealing

Simulated annealing not very competitive in practice compared to CMA-ES for instance.
However :

- Intuition in the concept of the algorithm is essential to understand global optimization
- Markov Chain Monte Carlo (MCMC) applied more and more in Industry (Bayesian inference)
- Simulated annealing can be combined with other algorithms (especially for the proposal distribution)

# Content

# CMA-ES

- CMA-ES = Covariance Matrix Adaptation Evolution-Strategy.
- A stochastic global optimization method,
- incrementally developed since 1996 by Hansen, Auger and others [Hansen and Ostermeier, 2001, Hansen, 2016].
- Default parameter values provided, contrarily to simulated annealing and other (stochastic) evolutionary optimization methods.

**Algorithm 6** CMA-ES structure

**Require:** $t^{\max}$, $m$, $C$, $\mu$, $\lambda$

$t \leftarrow 0$

1: **while** $t < t^{\max}$ **do**

2: $x^1, \ldots, x^\lambda \leftarrow$ i.i.d. calls to $\mathcal{N}(m, C)$

3: Calculate $f(x^1), \ldots, f(x^\lambda)$ , $t \leftarrow t + \lambda$
and rank them, $f(x^{1:\lambda}) \leq \ldots \leq f(x^{\lambda:\lambda})$

4: Update $m$ and $C$ from the $\mu$ best, $x^{1:\lambda}, \ldots, x^{\mu:\lambda}$

5: **end while**

6: **return** best observed point and $m$

It is an ES-$\mu\overset{+}{,}\lambda$ algorithm. Note the order notation $i : \lambda$ .
Metaphor: $\lambda =$ population size, $\mu =$ number of parents,
an iteration = a generation $g$, ES = Evolution Strategy.
Do not confuse this $C$ proposal density in $\mathcal{S}$ with the covariance matrix of kriging in $f$-space.
Now, let's detail the updates.

# From points to steps

Because large dimensional spaces cannot be sufficiently explored to learn accurate models, it is better to work with steps than with points: learn steps leading to progress (dynamic), and not points that were good (static).

$$\begin{aligned} \text{points}: &\quad x \sim \mathcal{N}(m^{(g)}, C^{(g)}) \\ \text{steps}: &\quad s := x - m^{(g)} \sim \mathcal{N}(0, C^{(g)}) \end{aligned}$$

# Update of the mean $m$

New mean = old mean + average of good steps

$$
\begin{aligned}
m^{(g+1)} &= m^{(g)} + \frac{1}{\mu}\sum_{i=1}^{\mu} s^{i:\lambda} \\
&= m^{(g)} + \frac{1}{\mu}\sum_{i=1}^{\mu}(x^{i:\lambda} - m^{(g)}) \\
\Rightarrow &\boxed{m^{(g+1)} = \frac{1}{\mu}\sum_{i=1}^{\mu} x^{i:\lambda}} \qquad (1)
\end{aligned}
$$

New mean is also the average of the best new points.

# Update of the covariance $C$ I

New covariance = empirical covariance of the good steps

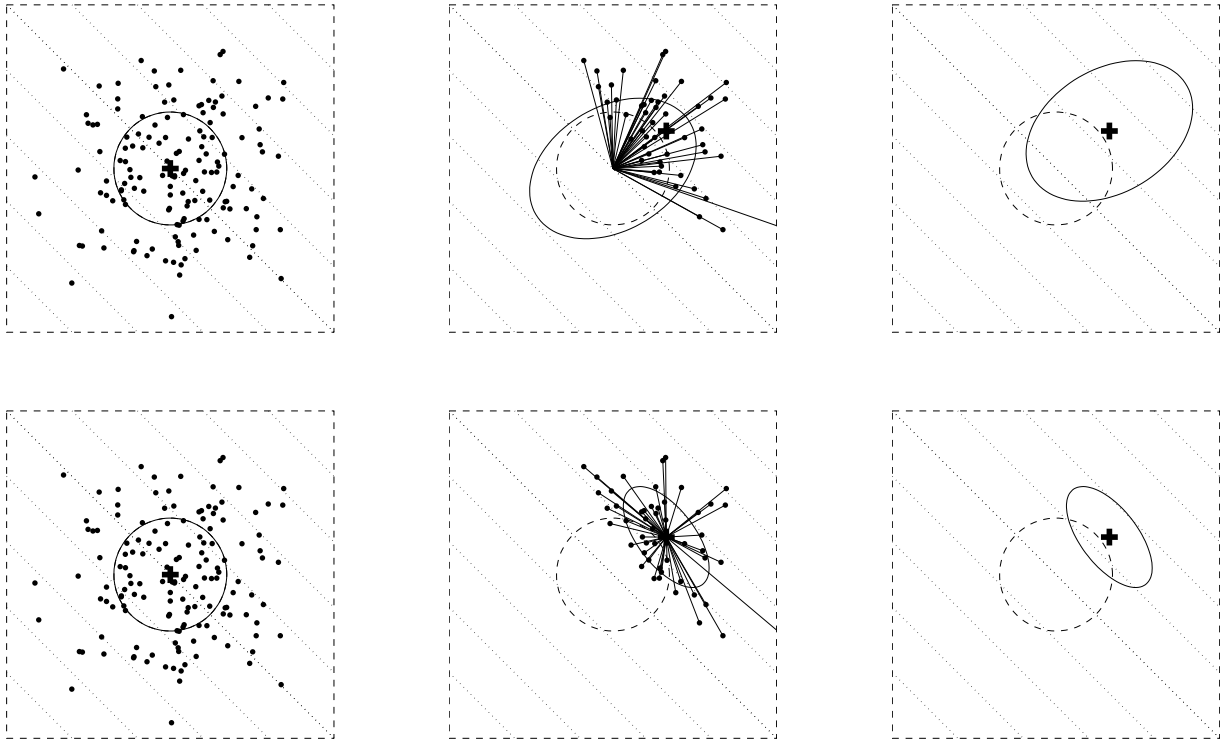
$$C^{(g+1)} = \frac{1}{\mu - 1} \sum_{i=1}^{\mu} s^{i:\lambda} s^{i:\lambda\top}$$

$$= \frac{1}{\mu - 1} \sum_{i=1}^{\mu} (x^{i:\lambda} - m^{(g)})(x^{i:\lambda} - m^{(g)})^{\top}$$

It is different from the covariance matrix of good points ([Larrañaga and Lozano, 2001])

$$C_{\text{EMNA}}^{(g+1)} = \frac{1}{\mu - 1} \sum_{i=1}^{\mu} (x^{i:\lambda} - m^{(g+1)})(x^{i:\lambda} - m^{(g+1)})^{\top}$$

# Update of the covariance $C$ II




Adaptation of $C$ (top row) versus of $C_{\text{EMNA}}$ (bottom row), from [Hansen, 2016]. The step length of $C$ correctly increases while that of $C_{\text{EMNA}}$ prematurly decreases.

# Path=cumulated steps I

To decrease the cost for estimating $C$ $(= \lambda)$,

*i)* average the steps and

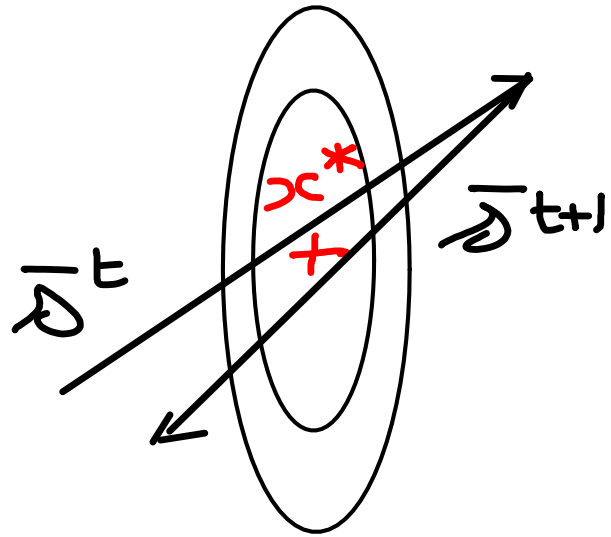
*ii)* cumulate them through time:

$$\boxed{\overline{s} = \frac{1}{\mu} \sum_{i=1}^{\mu} (x^{(i:\lambda)} - m^{(g)}) = m^{(g+1)} - m^{(g)}} \tag{2}$$

One could then have an estimation of $C$ with a time smoothing,

$$C^{(g+1)} = (1 - c_1) C^{(g)} + c_1 \overline{s}\, \overline{s}^{\top}$$

but it is insensitive to a change of direction, hence unable to converge.



# Path=cumulated steps II

Path = time cumulation of the steps:

$$\boxed{p^{(g+1)} = (1 - c_c)p^{(g)} + \sqrt{c_c(2 - c_c)\mu}\ \bar{s}} \tag{3}$$

Why these coefficients? To allow convergence when everything is stable: If $p^{(g)} \sim \mathcal{N}(0, C)$ and $s^{(i:\lambda)} \sim \mathcal{N}(0, C)$, then $p^{(g+1)} \sim \mathcal{N}(0, C)$.

Proof: 1st term $\sim \mathcal{N}(0, (1 - c_c)^2 C)$, $\bar{s} \sim \mathcal{N}(0, \frac{1}{\mu^2} \mu C)$, second term $\sim \mathcal{N}(0, c_c(2 - c_c)\mu \frac{1}{\mu} C)$, then sum the covariances of the two terms $\square$

# Path=cumulated steps III

Putting it together:

$$\boxed{C^{(g+1)} = (1 - c_1)C^{(g)} + c_1 \underbrace{p^{(g+1)} p^{(g+1)^\top}}_{\text{rank 1 update}}} \tag{4}$$

Quiz 7: Why is it called "rank 1 update"?

Default values [Hansen, 2016]:

- Population size and selection pressure: $\lambda \geq 4 + \lfloor 3 \ln(d) \rfloor$, $\mu = \lambda/2$
- $c_1 \approx 2/d^2$ and $c_c \approx 3/d$.

**Algorithm 7** A simplified CMA-ES

**Require:** $t^{\max}$, $m^{(1)}$, $C^{(1)}$, $\lambda$, $\mu$, $p^{(1)}$

$t \leftarrow 0$, $g \leftarrow 1$

1: **while** $t < t^{\max}$ **do**

2: $x^1, \ldots, x^\lambda \leftarrow$ i.i.d. calls to $\mathcal{N}(m^{(1)}, C^{(1)})$

3: Calculate $f(x^1), \ldots, f(x^\lambda)$ , $t \leftarrow t + \lambda$
and rank them, $f(x^{1:\lambda}) \leq \ldots \leq f(x^{\lambda:\lambda})$

4: Update $m$ and $C$ (Eqs. (1) to (4)):

$$m^{(g+1)} = \frac{1}{\mu} \sum_{i=1}^{\mu} x^{i:\lambda}$$

$$\overline{s} = m^{(g+1)} - m^{(g)}$$

$$p^{(g+1)} = (1 - c_c) p^{(g)} + \sqrt{c_c(2 - c_c)\mu}\ \overline{s}$$

$$C^{(g+1)} = (1 - c_1) C^{(g)} + c_1 p^{(g+1)} p^{(g+1)^\top}$$

5: $g \leftarrow g + 1$

6: **end while**

7: **return** best observed point and $m^{(g)}$

# The state-of-the-art CMA-ES

The above simplified CMA-ES was given for its ease of explanations.
The following features are missing:

- Weighting of the parents: $\overline{s} = \sum_{i=1}^{\mu} w_i s^{i:\lambda}$
- Separation of the size and shape of the covariance matrix: $C = \sigma^2 R$
- Simultaneous rank-1 ($pp^\top$) and rank-$\mu$ ($\sum_{i=1}^{\mu} s^{i:\lambda} s^{i:\lambda\top}$) updates of the covariance
- Restarts with increasing population sizes [Auger and Hansen, 2005]

# Content

# Optimizing with the help of GPs

Scientific domain: Bayesian optimization.

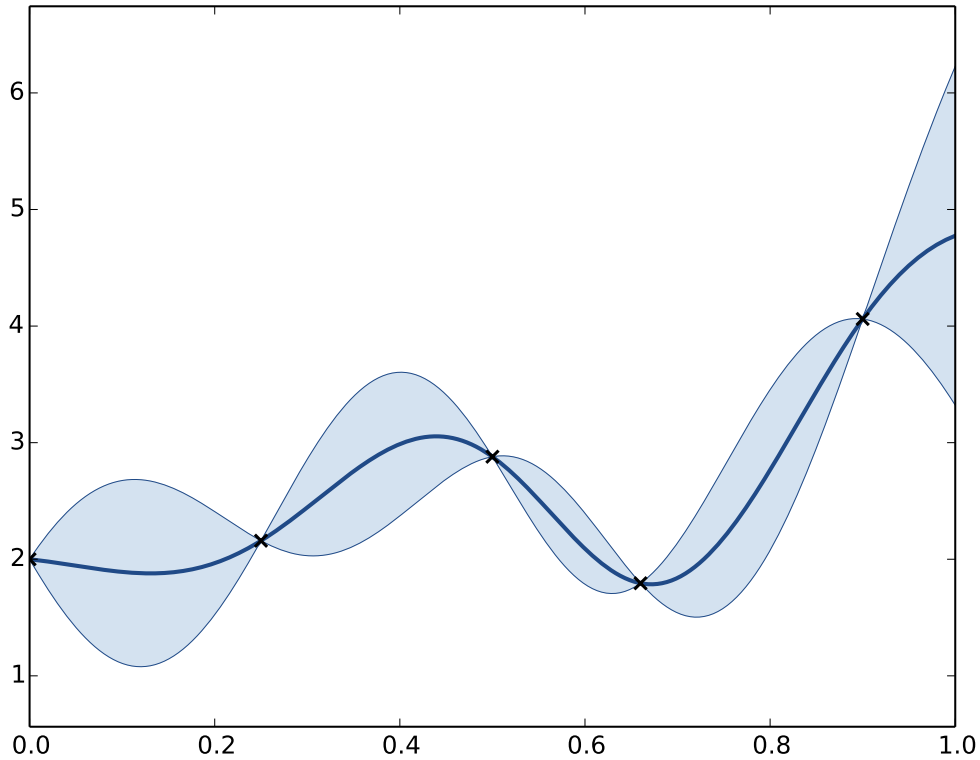
The problem:

$$\min_{x \in \mathcal{S} \subset \mathbb{R}^d} f(x) \tag{5}$$

where $f()$ is a black-box costly function. No property of $f$ that is helpful for optimization such as Lipschitz, convexity, uni-modality is known.

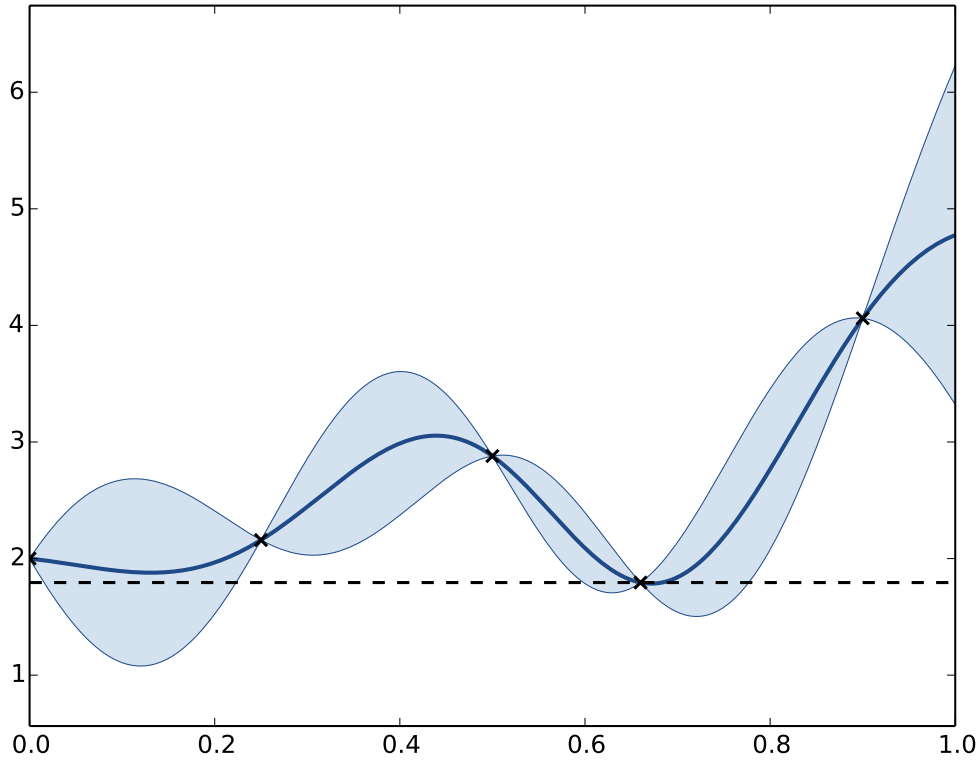
Principle: replace calls to $f()$ by the predicted density of the GP $(Y(x))$. Underlying assumption about the regularity of the function through the kernel choice, but it still works if this assumption is not satisfied.

# Given a GPR, where do you sample next?




(EGO figures from [Durrande and Le Riche, 2017])

In our example, the best observed value is 1.79



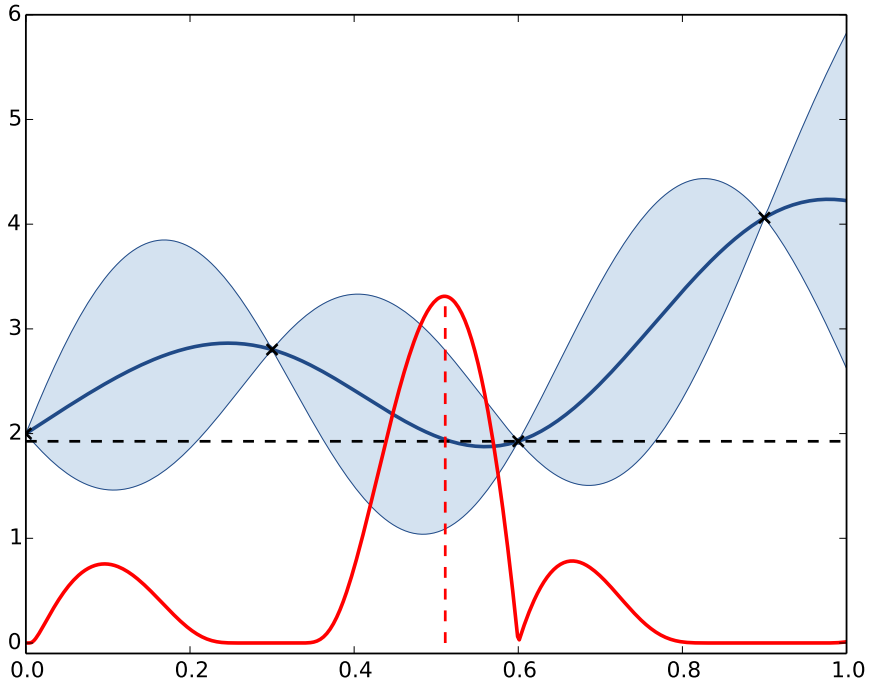

We need an acquisition criterion that uses the GP and seeks a compromise between exploration and intensification: the expected improvement ...

# The Expected Improvement

$$\mathrm{EI}(x) = \int_{-\infty}^{+\infty} \max\left(0, (\min(F) - y(x))\right)\ p(y(x))dy(x) = \cdots =$$

$$\sqrt{c(x,x)}\left[w(x)\mathrm{cdf}_{\mathcal{N}}(w(x)) + \mathrm{pdf}_{\mathcal{N}}(w(x))\right]$$

with $w(x) = \frac{\min(F) - m(x)}{\sqrt{(c(x,x))}}$.



# EI formula I

improvement as a random variable : $I(x) = \max(0, (\min(F) - Y(x)))$

$$\begin{aligned} \mathrm{EI}(x) &= \int_{-\infty}^{+\infty} \max(0, (\min(F) - y(x)))\ p(y(x))\ dy(x) \\ &= \int_{-\infty}^{\min(F)} (\min(F) - y(x))\ p(y(x))\ dy(x) + \int_{\min(F)}^{+\infty} 0\ dy(x) \end{aligned}$$

change of variable, temporarily drop some $x$'s to ease notation:
$v = v(x) = \frac{y(x) - m(x)}{s(x)}$, $s = s(x) = \sqrt{c(x,x)}$

$$\mathrm{EI}(x) = \int_{-\infty}^{\frac{\min(F)-m}{s}} (\min(F) - m - vs)\ p(v) dv$$

# EI formula II

$v$ is made to be standard, $p(v) = \text{pdf}_{\mathcal{N}}(v) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}v^2\right)$, so $\text{pdf}_{\mathcal{N}}'(v) = -v \times \text{pdf}_{\mathcal{N}}(v)$

$$\begin{aligned}\text{EI}(x) &= (\min(F) - m(x)) \times \text{cdf}_{\mathcal{N}}(w(x)) + s(x) \times \text{pdf}_{\mathcal{N}}(w(x)) \\ &= \sqrt{c(x,x)}\,[w(x) \times \text{cdf}_{\mathcal{N}}(w(x)) + \text{pdf}_{\mathcal{N}}(w(x))]\end{aligned}$$

where $w(x) = \frac{\min(F) - m(x)}{\sqrt{(c(x,x))}}$ $\square$.

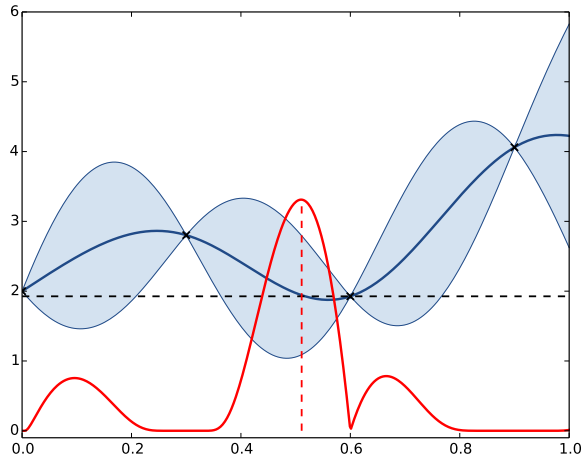
# Bayesian optimization principle

Iteratively,

$$x^{t+1} = \arg\max_{x \in \mathcal{S}} EI(x)$$

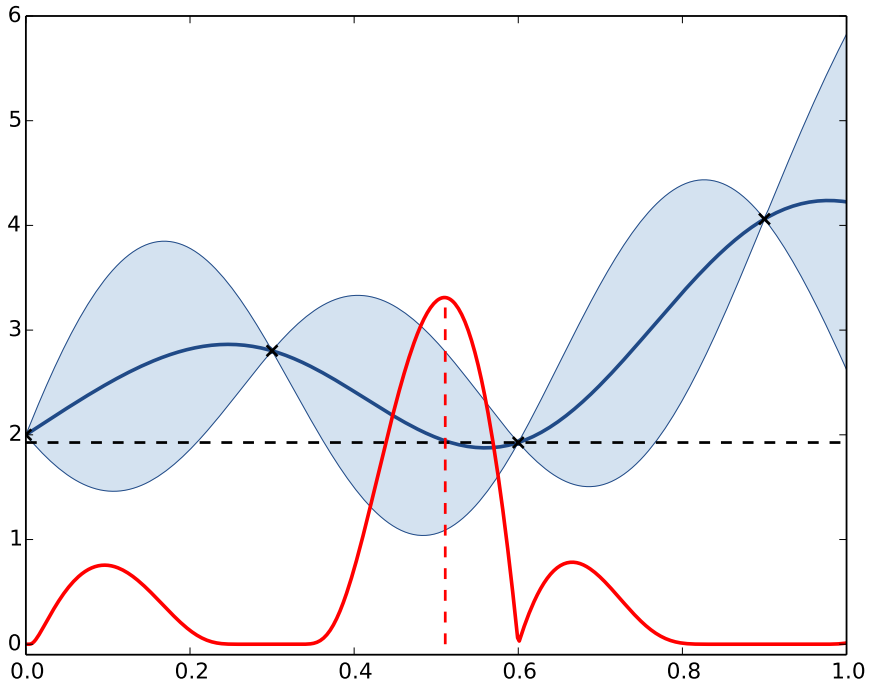
$$\text{calculate } f(x^{t+1}) \text{ , update the GP}$$

- $EI(x) \geq 0$
- Quiz 8 Prove that $EI(x^i) = 0$ when $x^i$ is already evaluated and part of the DoE of the GP.
- Hence $EI(x)$ tends to be multimodal.



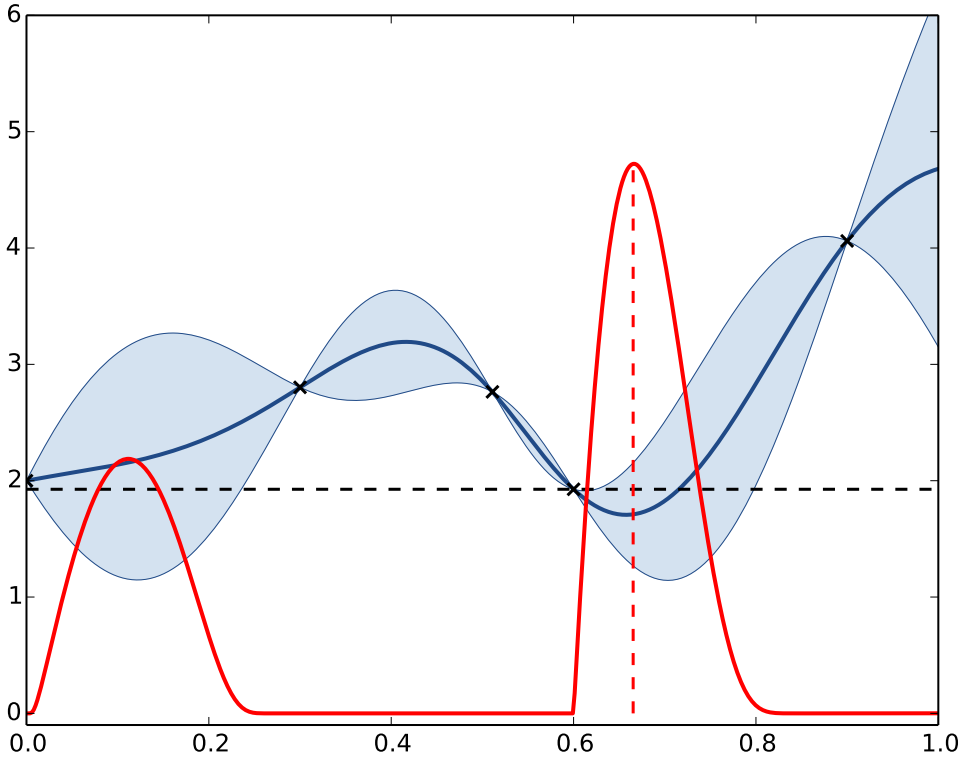
# Expected Improvement

Let's see how it works... iteration 0



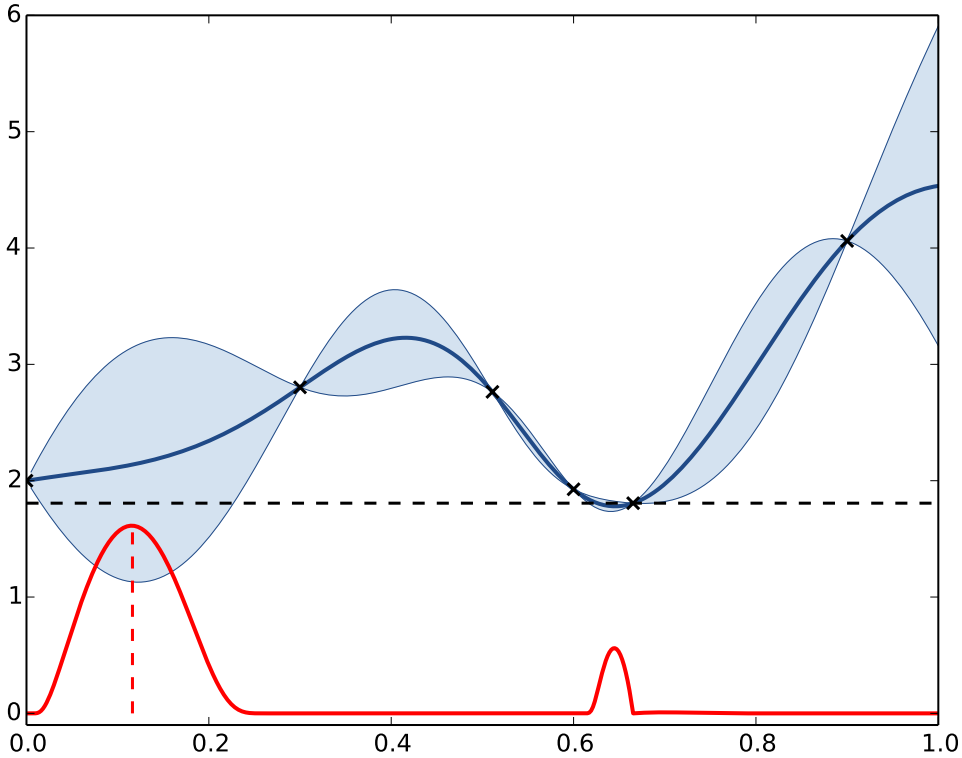
# Expected Improvement

Let's see how it works... iteration 1



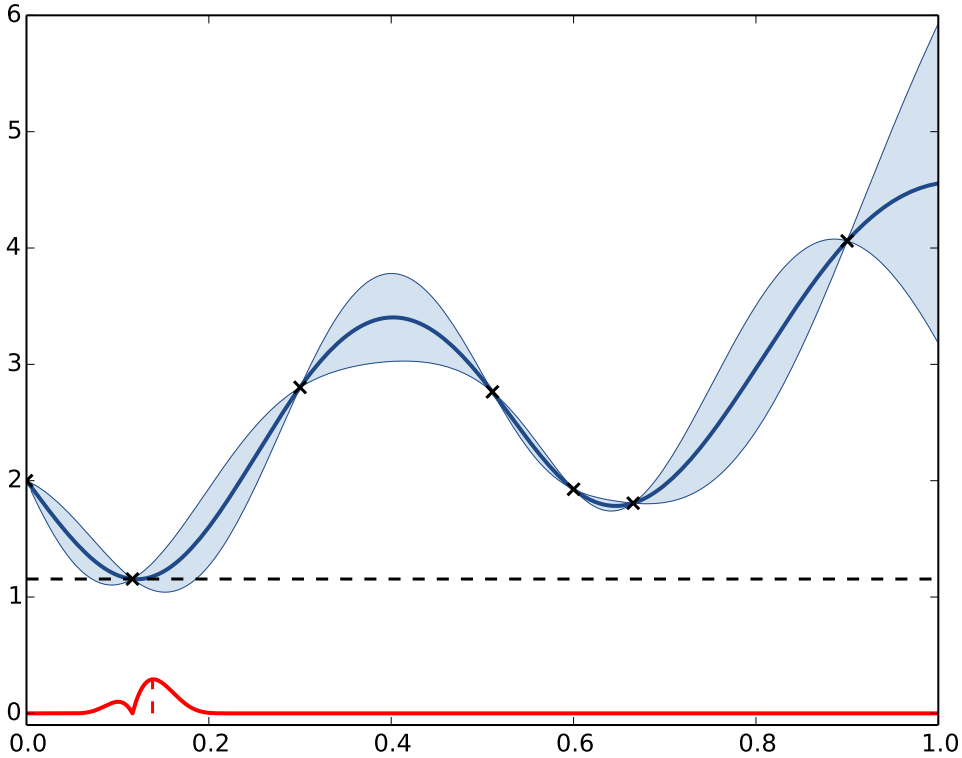
# Expected Improvement

Let's see how it works... iteration 2



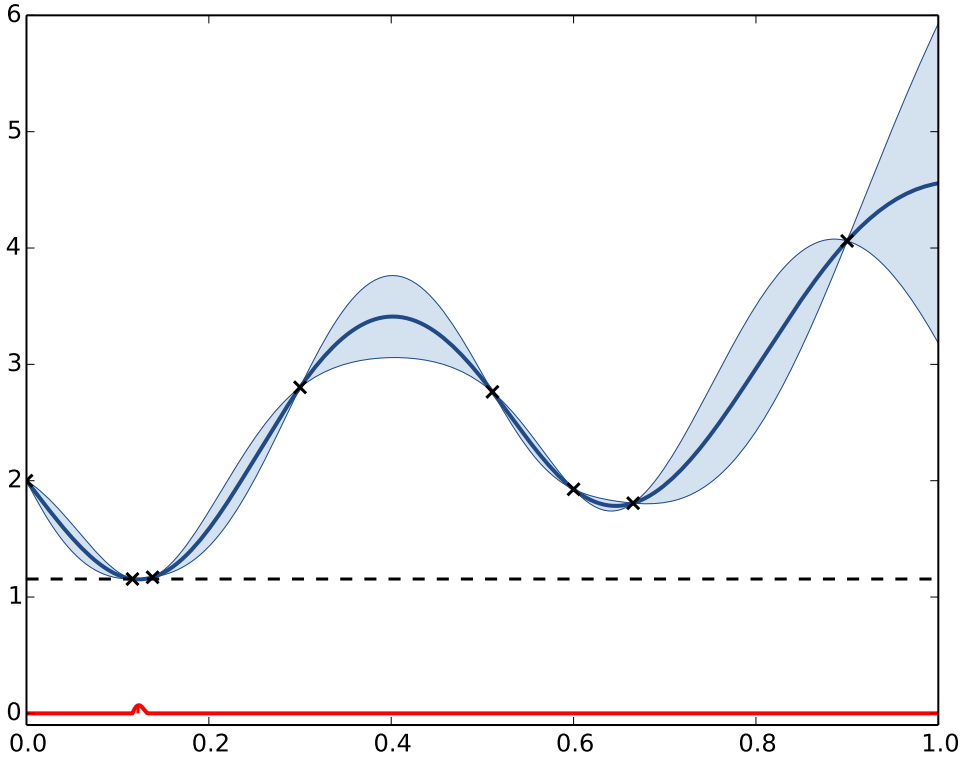
# Expected Improvement

Let's see how it works... iteration 3



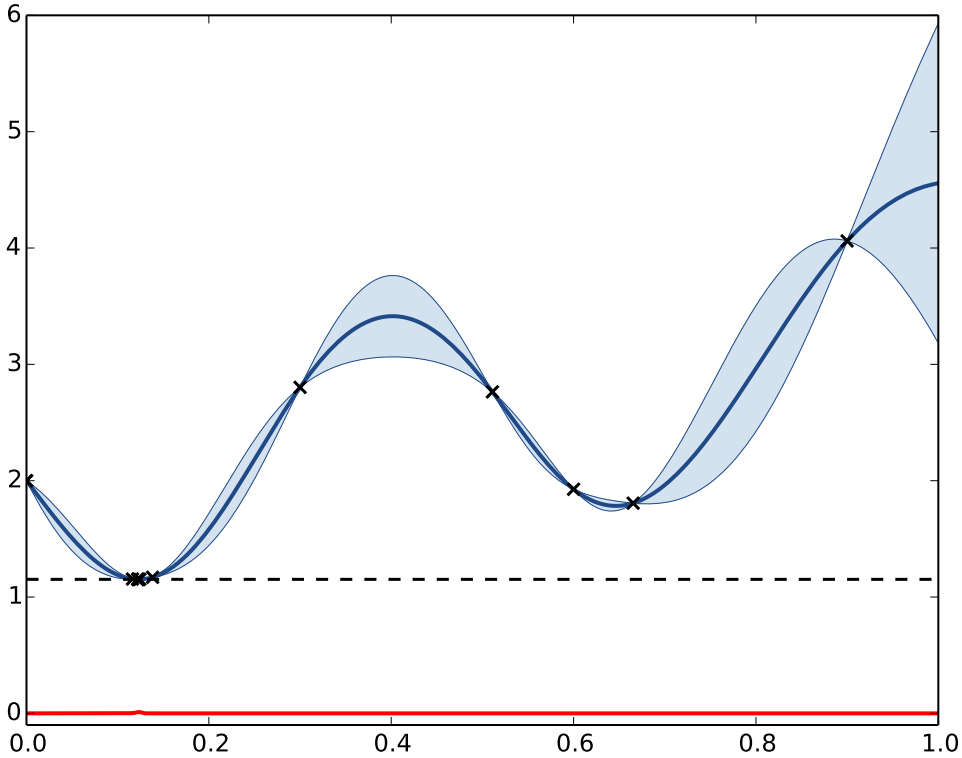
# Expected Improvement

Let's see how it works... iteration 4



# Expected Improvement

Let's see how it works... iteration 5



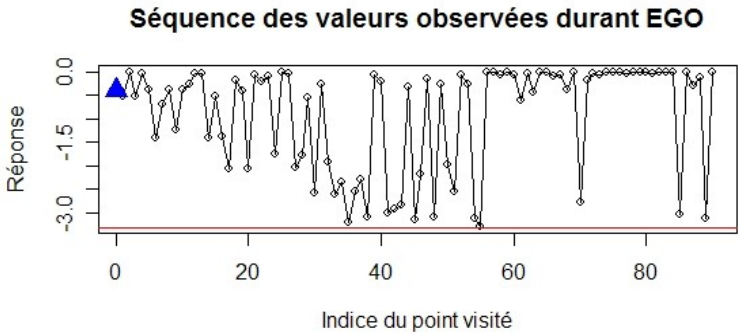
**Algorithm 8** Bayesian optimization with EI acquisition (EGO, [Jones et al., 1998])

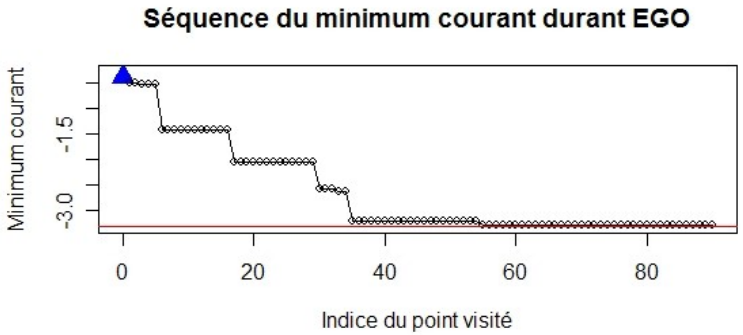
**Require:** $x^{\text{LB}}$, $x^{\text{UB}}$, $t^{\max}$

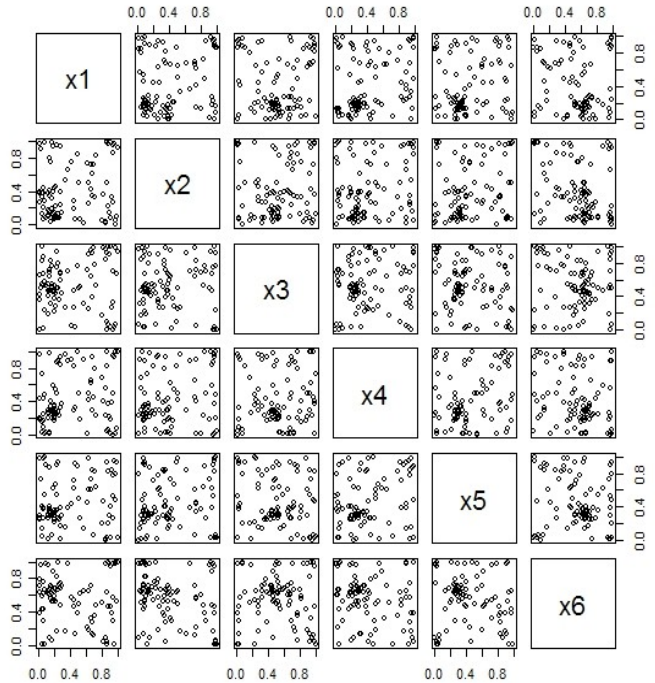
1: make an initial design of experiments $X$ and calculate the associated $F$, $t = \texttt{length}(F)$
2: build a GP from $(X, F)$ (max. log-likelihood on GP parameters – variance, length scales–)
3: **while** $t < t^{\max}$ **do**
4: $x^{t+1} = \arg\max_x EI(x)$ (with another optimizer, e.g. CMA-ES [Hansen and Ostermeier, 2001])
5: calculate $F_{t+1} = f(x^{t+1})$, $(X, F) \leftarrow (X, F) \cup (x^{t+1}, F_{t+1})$, $t \leftarrow t + 1$
6: **end while**
7: **return** best point in $(X, F)$, the GP

# EGO in higher dimension

Hartmann function, $f(x^\star) = -3.32$, 10 points in initial design of experiments




Notice the global sampling of the search space with point clusters

(from [Ginsbourger, 2009])

# EI and the intensification/exploration tradeoff

+ EI provides a good trade-off between intensification and exploration without arbitrary parameters.
  $\mathrm{EI}(x)$ increases when the variance $c(x,x)$ increases and it increases when the mean $m(x)$ decreases.

  Quiz 9: Prove that $\partial \mathrm{EI}(x)/\partial m(x) < 0$.
  Quiz 10: Prove that $\partial \mathrm{EI}(x)/\partial c(x,x) > 0$.

# Concluding comments on Bayesian Optimization

+ Bayesian optimization requires few function observations to get close to optima.
  - However its performance is not good in more than 10 dimensions [Le Riche and Picheny, 2021].
  - Special treatments required for more than 1000 points because of the covariance matrix inversion of the GP.
× EGO does not converge in the traditional sense: it creates dense samples in the volume of S. The efficiency comes from the order in which points are sampled.

# References I



Auger, A. and Dumas, L. (2017).
Derivative free optimization.


Auger, A. and Hansen, N. (2005).
A restart CMA evolution strategy with increasing population size.
In *2005 IEEE congress on evolutionary computation*, volume 2, pages 1769–1776. IEEE.


Bäck, T., Schütz, M., et al. (1995).
Evolution strategies for mixed-integer optimization of optical multilayer systems.
In *Evolutionary Programming*, pages 33–51.


Durrande, N. and Le Riche, R. (2017).
Introduction to Gaussian Process Surrogate Models.
Lecture at 4th MDIS form@ter workshop, Clermont-Fd, France.
HAL report cel-01618068.


Gaudrie, D., Le Riche, R., Picheny, V., Enaux, B., and Herbert, V. (2020).
Modeling and optimization with gaussian processes in reduced eigenbases.
*Structural and Multidisciplinary Optimization*, 61(6):2343–2361.

# References II



Ginsbourger, D. (2009).
*Multiples métamodèles pour l'approximation et l'optimisation de fonctions numériques multivariables*.
PhD thesis, Mines de Saint-Etienne.


Hansen, N. (2016).
The CMA evolution strategy: A tutorial.
*arXiv preprint arXiv:1604.00772*.


Hansen, N. and Ostermeier, A. (2001).
Completely derandomized self-adaptation in evolution strategies.
*Evol. Comput.*, 9(2):159–195.


Irisarri, F.-X., Lasseigne, A., Leroy, F.-H., and Le Riche, R. (2014).
Optimal design of laminated composite structures with ply drops using stacking sequence tables.
*Composite Structures*, 107:559–569.


Jones, D. R., Perttunen, C. D., and Stuckman, B. E. (1993).
Lipschitzian optimization without the lipschitz constant.
*Journal of optimization Theory and Applications*, 79(1):157–181.

# References III



Jones, D. R., Schonlau, M., and Welch, W. J. (1998).
Efficient Global Optimization of expensive black-box functions.
*Journal of Global optimization*, 13(4):455–492.


Larrañaga, P. and Lozano, J. A. (2001).
*Estimation of distribution algorithms: A new tool for evolutionary computation*, volume 2.
Springer Science & Business Media.


Le Riche, R. and Picheny, V. (2021).
Revisiting bayesian optimization in the light of the COCO benchmark.
*Journal of Structural and Multidisciplinary Optimization*.

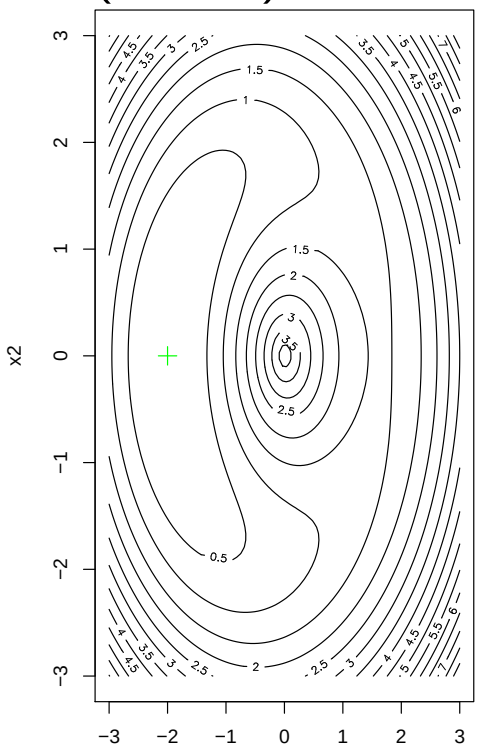

Rudolph, G. (1992).
On correlated mutations in evolution strategies.
In Männer, R. and Manderick, B., editors, *Parallel Problem Solving from Nature 2, PPSN-II, Brussels, Belgium, September 28-30, 1992*, pages 107–116. Elsevier.

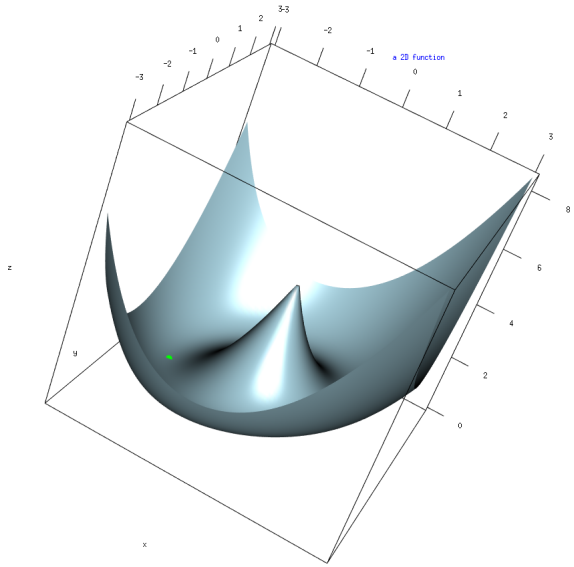

Wolpert, D. H. and Macready, W. G. (1997).
No free lunch theorems for optimization.
*IEEE transactions on evolutionary computation*, 1(1):67–82.

# Answers to the quizzes I

- Quiz 1 : Let's consider $\min_{x_1} f(x_1, x_2 = z)$ where $z$ is fixed. At the local optima, the necessary optimality condition for is $\partial f(x_1, z)/\partial x_1 = 0$ but these $(x_1, z)$ may not be optima in 2 dimensions since $\partial f(x_1, z)/\partial x_2$ may not be 0. As an example, the function $f(x_1, x_2) = (\|x\| - R)^2 + 0.1\|x - (-2, 0)^\top\|^2$ has a unique optimum at $(-2, 0)^\top$. It is multimodal for many $x_2$ fixed,

e.g., $x_2 = 0.5$ :



# Answers to the quizzes II

- Quiz 2 :

$$\forall i \in \{1, \dots, t\} \quad , \quad f(x^i) - L\|x - x^i\| \leq f(x)$$
$$\text{tightest lower bound : } \max_{i=1,\dots,t} \left(f(x^i) - L\|x - x^i\|\right) \leq f(x)$$

the min of the left-hand-side remains smaller than the min of the right-hand-side,

$$\min_{x \in \mathcal{S}} \max_{i=1,\dots,t} \left(f(x^i) - L\|x - x^i\|\right) \leq f(x^\star) = \min_{x \in \mathcal{S}} f(x)$$

multiply this inequality by $-1$ and add it to $\min_{i=1,\dots,t} f(x^i)$ to upperbound the $f$-distance to the optimum,

$$\min_{i=1,\dots,t} f(x^i) - f(x^\star) \leq \min_{i=1,\dots,t} f(x^i) - \min_{x \in \mathcal{S}} \max_{i=1,\dots,t} \left(f(x^i) - L\|x - x^i\|\right)$$

# Answers to the quizzes III

- Quiz 3: because the point is uniformly sampled, the probability to be in the neighborhood of $x^\star$ is $p =$ volume neighborhood/volume $\mathcal{S} = 10^{-5}/1 = 10^{-5}$. The number of trials before a point falls in the neighborhood, $T$, follows a geometric distribution, $\mathbb{P}(T = k) = (1-p)^{k-1}p$ whose expectation is $1/p = 10^5$ and whose standard deviation is $\sqrt{1-10^{-5}}/10^{-5} \approx 10^5$.
- Quiz 4: Let $p_r$ be the probability of locating the optimum within $r$ restarts, $p_r = 1-(1-p_1)^r$.
  $p_{r+1} - p_r = 1-(1-p_1)^{r+1} - 1 + (1-p_1)^r = p_1(1-p_1)^r$ which is positive, so there is a gain in a $(r+1)$-th search, but it is a decreasing gain with $r$.

# Answers to the quizzes IV

- Quiz 5: We will take the same Markov Chain as the one slide 44 but with a different starting distribution, considering $(X^{(t)})_{t=0,1..}$ defined as follows :
  - $\mathcal{S} = 1, 2, 3$
  - $\pi^{(0)} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}^T$
  - $p_{1,2} = p_{3,1} = p_{2,3} = 1$

  Then again $\pi = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}^T$ is stationary here, the chain is irreducible but all the states have period 3. However, as the starting distribution is equal to the stationary distribution, then the marginal distribution of $X^{(t)}$ is the same $\forall t$ and the Markov Chain converges even if all the states of a period equal to 3.

# Answers to the quizzes V

- Quiz 6: If $T \simeq 0$, $exp(\frac{f(x)-f(\tilde{x})}{T}) \simeq 0$ when $f(x) < f(\tilde{x})$ and then a candidate point is almost never accepted in this case. Then, the behaviour algorithm is almost only defined by the proposal distribution. For instance :
  - If the proposal distribution is gaussian, the algorithm is almost similar to normal search
  - If the proposal distribution is uniform, the algorithm is almost similar to random search

  The Markov Chain still converges asymptotically to a random variable of density $\pi(x) = \frac{1}{K} exp(-\frac{f(x)}{T})$ but if the starting point is not located in the zone of the "almost dirac" of $\pi$, this convergence will be very long. This is precisely the idea behind the slowly decreasing temperature, we want to bring $x$ progressively closer to this unknown zone so that for the last

# Answers to the quizzes VI

temperature $T_f \simeq 0$, the starting point of the Markov Chain is in this zone.

- Quiz 7: It is called "rank 1 update" because the correction to the covariance matrix, $p^{(g+1)} p^{(g+1)^\top}$, is a matrix of rank 1. Indeed, the rank of a matrix is the dimension of its image space, which is the number of non-zero eigenvalues. A matrix like $pp^\top$ has 1 non-zero eigenvalue, $\|p\|^2$, because $(pp^\top)\frac{p}{\|p\|} = \|p\|^2 \frac{p}{\|p\|}$, and all $d-1$ vectors perpendicular to $p$ have a 0 eigenvalue.
- Quiz 8: $x^i$ is one of the points learned by the GP which is interpolating, therefore $Y(x^i) = f(x^i)$.
$I(x^i) = \max(0, \min(F) - Y(x^i)) = \max(0, \min(F) - f(x^i)) = \max(0, \min(f(x^1), \ldots, f(x^t)) - f(x^i)) = 0$.

# Answers to the quizzes VII

- Quiz 9: switch to shorthand notation dropping all $x$'s (everything is at $x$), $s = \sqrt{c(x,x)}$, $w = (\min(F) - m)/s$, one has: $\frac{\partial w}{\partial m} = -\frac{1}{s}$,
$\frac{\partial \text{EI}}{\partial m} = s[-\frac{1}{s} \times \text{cdf}_{\mathcal{N}}(w) + w(-\frac{1}{s}) \times \text{pdf}_{\mathcal{N}}(w) - w(-\frac{1}{s}) \times \text{pdf}_{\mathcal{N}}(w)]$
$= -\text{cdf}_{\mathcal{N}}(w) < 0.$
- Quiz 10: same notation as previous quiz, $\frac{\partial w}{\partial s} = -\frac{1}{s}w$,
$\frac{\partial \text{EI}}{\partial s} = w \times \text{cdf}_{\mathcal{N}}(w) + \times \text{pdf}_{\mathcal{N}}(w) + s[-\frac{1}{s}w \times \text{cdf}_{\mathcal{N}}(w) +$
$w(-\frac{1}{s}w) \times \text{pdf}_{\mathcal{N}}(w) - w(-\frac{1}{s}w) \times \text{pdf}_{\mathcal{N}}(w)] = \text{pdf}_{\mathcal{N}}(w) > 0.$