Numerical exploration-exploitation tradeoff for large scale function optimization

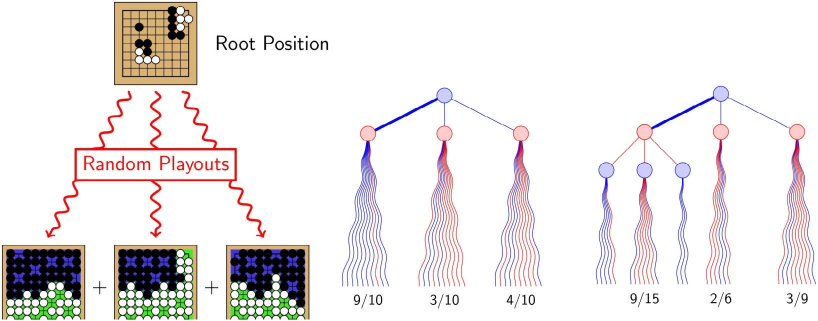
Rémi Munos

INRIA Lille - Nord Europe
Currently on leave at MSR-NE
http://researchers.lille.inria.fr/∼munos/

LSOLDM 2013, Cumberland Lodge

Initial motivation

Monte-Carlo Tree Search in computer-go



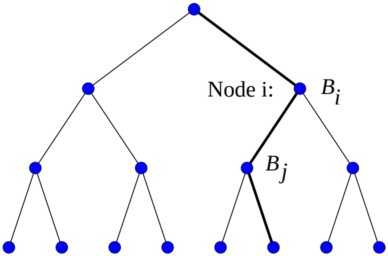
MCTS in Crazy-Stone (Rémi Coulom, 2005)

Idea: use bandits at each node of the tree search.

UCB applied to Trees

Uses Upper Confidence Bound (UCB) algorithm [Auer et al., 2002] at each node of the tree

$$B_j \stackrel{\text{def}}{=} X_{j,n_j} + \sqrt{\frac{2\log(n_i)}{n_j}}.$$



Intuition:

- Explore first the most promising branches
- Average converges to max

- Adaptive Multistage Sampling (AMS) algorithm [Chang, Fu, Hu, Marcus, 2005]
- UCB applied to Trees (UCT) [Kocsis and Szepesvári, 2006]

The MoGo program [Gelly et al., 2006]

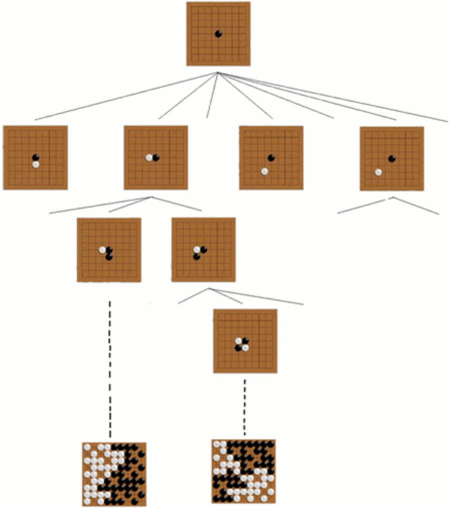
Use hierarchy of UCB bandits (UCT) [Kocsis and Szepesvári, 2006]

Features:

- Monte-Carlo evaluation
- Asymmetric tree expansion
- Anytime algo
- Use of features

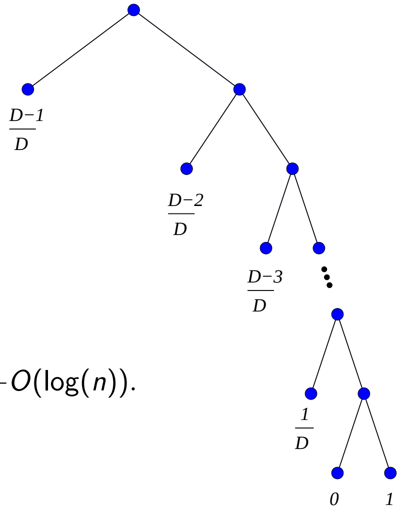
Very strong program!

MCTS and UCT very successful



No finite-time guarantee for UCT

Problem: at each node, the rewards are not i.i.d.
Consider the tree:



The left branches seem better than right branches, thus are explored for a **very** long time before the optimal leaf is eventually reached.
The regret is disastrous:

$$\mathbb{E}R_n = \Omega(\underbrace{\exp(\exp(\ldots\exp(}_{D \text{ times}}1)\ldots))) + O(\log(n)).$$

See [Coquelin and Munos, 2007]

Optimism in the face of uncertainty

"Numerical exploration-exploitation tradeoff": perform search in simulation using finite numerical resources.

Outline:

- Optimistic optimization of a deterministic Lipschitz functions
- 4 extensions:
  - Locally smooth functions,
  - Tractable algorithm
  - Unknown smoothness,
  - Noisy evaluations

Optimization of a deterministic Lipschitz function

Problem: Find online the maximum of $f : X \to \boldsymbol{R}$, assumed to be Lipschitz:

$$|f(x) - f(y)| \leq \ell(x, y).$$

Protocol:

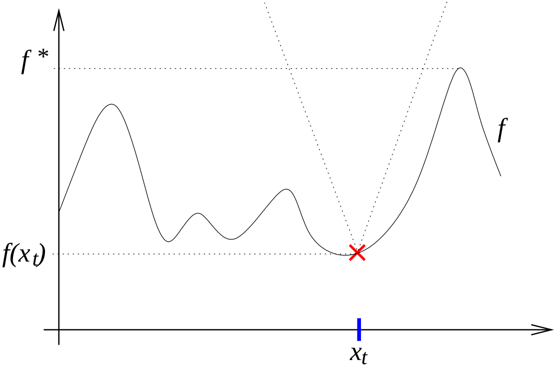
- For each time step $t = 1, 2, \ldots, n$ select a state $x_t \in X$
- Observe $f(x_t)$
- Return a state $x(n)$

Loss:

$$r_n = f^* - f(x(n)),$$

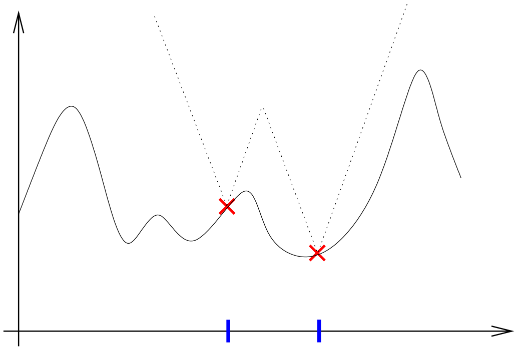
where $f^* = \sup_{x \in X} f(x)$.

Example in 1d



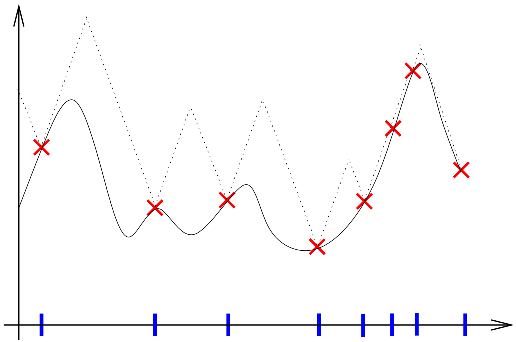
Lipschitz property $\rightarrow$ the evaluation of f at x_t provides a first upper-bound on f.

Example in 1d (continued)



New point → refined upper-bound on f.

Example in 1d (continued)



Question: where should one sample the next point?
Answer: select the point with highest upper bound!
"Optimism in the face of (partial observation) uncertainty"

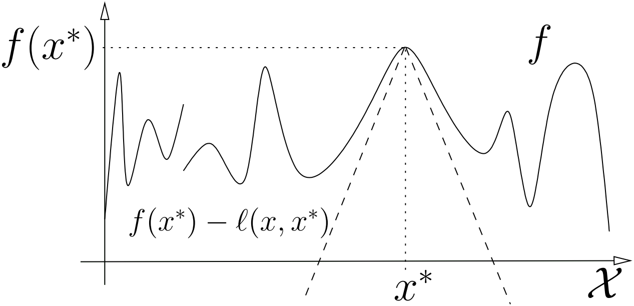
Several issues

1. Lipschitz assumption is too strong
2. Finding the optimum of the upper-bounding function may be hard!
3. What if we don't know the metric ℓ?
4. How to handle noise?

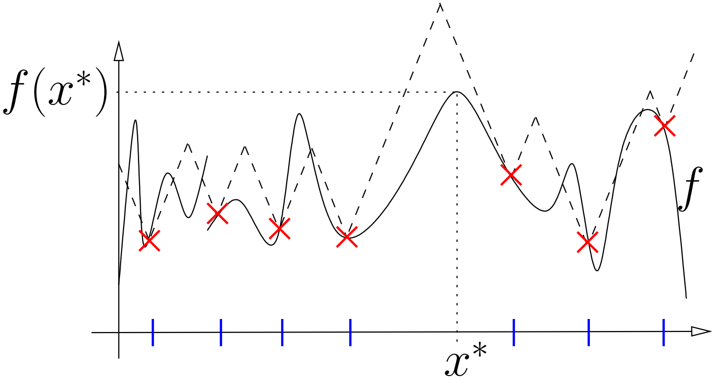
Local smoothness property

Assumption: f **is "locally smooth" around its max. w.r.t.** ℓ where ℓ is a semi-metric (symmetric, and $\ell(x, y) = 0 \Leftrightarrow x = y$):
For all $x \in \mathcal{X}$,

$$f(x^*) - f(x) \leq \ell(x, x^*).$$



Local smoothness is enough!



Optimistic principle only requires:

- a true bound at the maximum
- the bounds gets refined when adding more points

Efficient implementation

Deterministic Optimistic Optimization (DOO) builds a hierarchical partitioning of the space where cells are refined according to their upper bounds.

- For $t = 1$ to n,
  - Define an upper bound for each cell:

$$B_i = f(x_i) + \mathrm{diam}_\ell(X_i)$$

  - Select the cell with highest bound

$$I_t = \operatorname*{argmax}_i B_i.$$

  - Expand I_t: refine the grid and evaluate f in children cells
- Return $x(n) \stackrel{\mathrm{def}}{=} \mathrm{argmax}_{\{x_t\}_{1 \le t \le n}} f(x_t)$

Near-optimality dimension

Define the **near-optimality dimension** of f as the smallest $d \geq 0$ such that $\exists C$, $\forall \epsilon$, the set of ε-optimal states

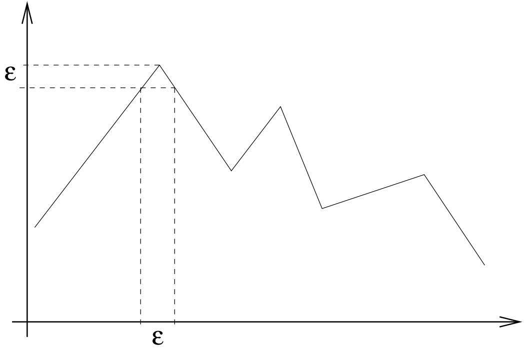
$$X_\varepsilon \stackrel{\text{def}}{=} \{x \in X, f(x) \geq f^* - \varepsilon\}$$

can be covered by $C\varepsilon^{-d}$ ℓ-balls of radius ε.

Example 1:

Assume the function is piecewise linear at its maximum:

$$f(x^*) - f(x) = \Theta(||x^* - x||).$$

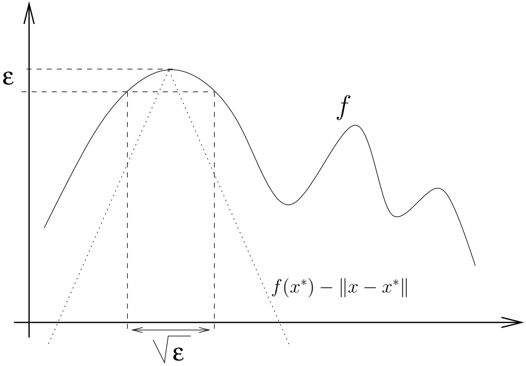


Using $\ell(x, y) = \|x - y\|$, it takes $O(\epsilon^0)$ balls of radius ϵ to cover X_ε. Thus $d = 0$.

Example 2:

Assume the function is locally quadratic around its maximum:

$$f(x^*) - f(x) = \Theta(||x^* - x||^2).$$

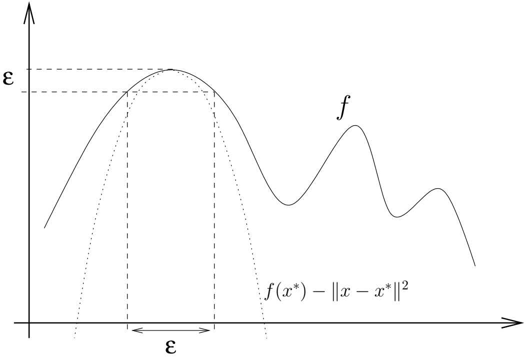


For $\ell(x, y) = ||x - y||$, it takes $O(\epsilon^{-D/2})$ balls of radius ϵ to cover X_ε (of size $O(\epsilon^{D/2})$). Thus $d = D/2$.

Example 2:

Assume the function is locally quadratic around its maximum:

$$f(x^*) - f(x) = \Theta(||x^* - x||^2)$$

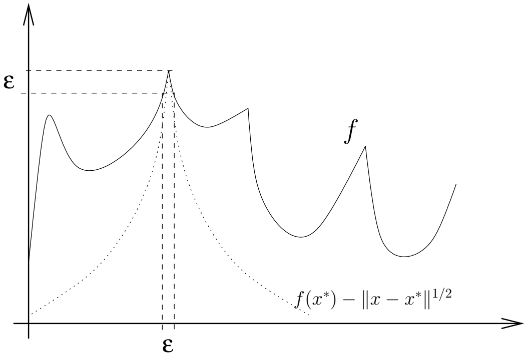


For $\ell(x, y) = ||x - y||^2$, it takes $O(\epsilon^0)$ ℓ-balls of radius ϵ to cover X_ε. Thus $d = 0$.

Example 3:

Assume the function has a square-root behavior around its maximum:

$$f(x^*) - f(x) = \Theta(||x^* - x||^{1/2})$$



For $\ell(x, y) = \|x - y\|^{1/2}$ we have $d = 0$.

Example 4:

Assume $\mathcal{X} = [0,1]^D$ and f is locally equivalent to a polynomial of degree $\alpha > 0$ around its maximum (i.e. f is α-smooth):

$$f(x^*) - f(x) = \Theta(||x^* - x||^\alpha)$$

Consider the semi-metric $\ell(x,y) = \|x - y\|^\beta$, for some $\beta > 0$.

- If $\alpha = \beta$, then $d = 0$.
- If $\alpha > \beta$, then $d = D(\frac{1}{\beta} - \frac{1}{\alpha}) > 0$.
- If $\alpha < \beta$, then the function is not locally smooth wrt ℓ.

Analysis of DOO (deterministic case)

Assume that the ℓ-diameters of the nodes of depth h decrease exponentially fast with h (i.e., $\text{diam}(h) = c\gamma^h$, for $c > 0$ $\gamma < 1$).

Example: $\mathcal{X} = [0,1]^D$ and $\ell(x,y) = \|x-y\|^\beta$ for some $\beta > 0$.

Theorem 1.
The loss of DOO is

$$r_n = \begin{cases} \left(\frac{C}{1-\gamma^d}\right)^{1/d} n^{-1/d} & \textit{for } d > 0, \\ c\gamma^{n/C-1} & \textit{for } d = 0. \end{cases}$$

(Remember that $r_n \stackrel{\text{def}}{=} f(x^*) - f(x(n))$).

About the local smoothness assumption

Assume f satisfies $f(x^*) - f(x) = \Theta(||x^* - x||^\alpha)$.

Use DOO with the semi-metric $\ell(x, y) = \|x - y\|^\beta$:

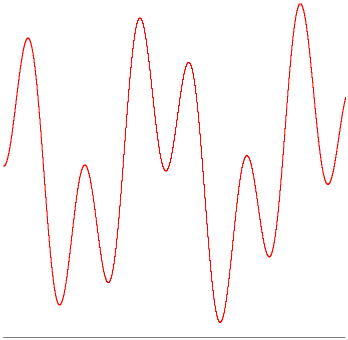
- If $\alpha = \beta$, then $d = 0$: the true "local smoothness" of the function is known, and exponential rate is achieved.
- If $\alpha > \beta$, then $d = D(\frac{1}{\beta} - \frac{1}{\alpha}) > 0$: we under-estimate the smoothness, which causes more exploration than needed.
- If $\alpha < \beta$: We over-estimate the true smoothness and DOO may fail to find the global optimum.

DOO heavilly depends on our knowledge of the true local smoothness.

Experiments [1]

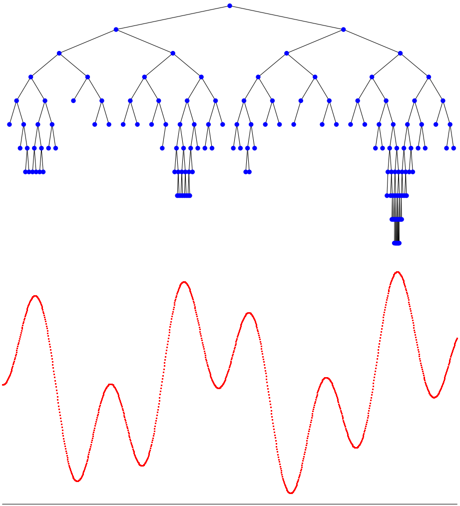
$f(x) = \frac{1}{2}(\sin(13x)\sin(27x) + 1)$ satisfies the local smoothness assumption with

- $\ell_1(x, y) = 14|x - y|$ (i.e., f is globally Lipschitz), $d = 1/2$
- $\ell_2(x, y) = 222|x - y|^2$ (i.e., f is locally quadratic), $d = 0$



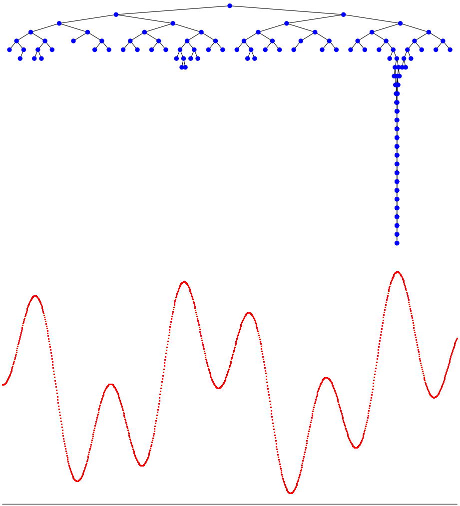
Experiments [2]

Using $\ell_1(x, y) = 14|x - y|$ (i.e., f is globally Lipschitz). $n = 150$.



Experiments [3]

Using $\ell_2(x, y) = 222|x - y|^2$ (i.e., f is locally quadratic). $n = 150$.



Experiments [4]

n	uniform grid	DOO with ℓ_1 $(d = 1/2)$	DOO with ℓ_2 $(d = 0)$
50	1.25×10^{-2}	2.53×10^{-5}	1.20×10^{-2}
100	8.31×10^{-3}	2.53×10^{-5}	1.67×10^{-7}
150	9.72×10^{-3}	4.93×10^{-6}	4.44×10^{-16}

Loss r_n for different values of n for a uniform grid and DOO with the two semi-metric ℓ_1 and ℓ_2.

What if the smoothness is unknown?

Previous algorithms heavily rely on the knowledge or the local smoothness of the function (i.e. knowledge of the best metric).

Question: When the smoothness is unknown, is it possible to implement the optimistic principle for function optimization?

DIRECT algorithm [Jones et al., 1993]

Assumes f is Lipschitz but the Lipschitz constant L is unknown.

The DIRECT algorithm expands simultaneously all nodes that may potentially contain the maximum for some value of L.

Be optimistic for all L

Illustration of DIRECT

The sin function and its upper bound for $L = 2$.

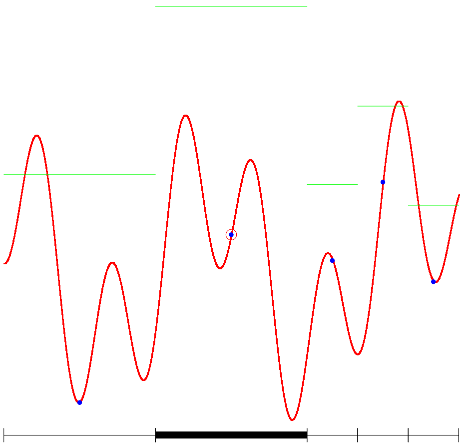
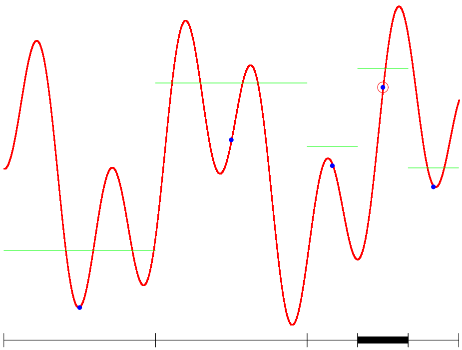


Illustration of DIRECT

The sin function and its upper bound for $L = 1/2$.



Limitations of DIRECT

- No finite-time analysis (only the consistency property $\lim_{n\to\infty} r_n = 0$ in [Finkel and Kelley, 2004])
- Global Lipschitz assumption is too strong!

We want to extend to

- any function **locally smooth** w.r.t. ℓ,
- for **any semi-metric** ℓ
- and provide performance guarantees.

Simultaneous Optimistic Optimization (SOO)

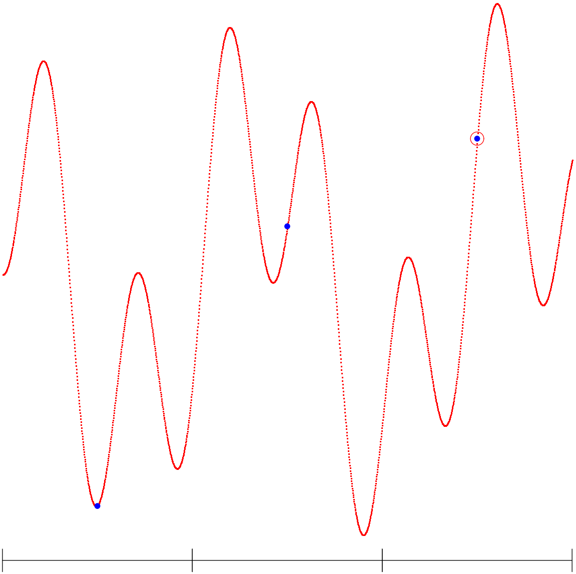
[Munos, 2011]

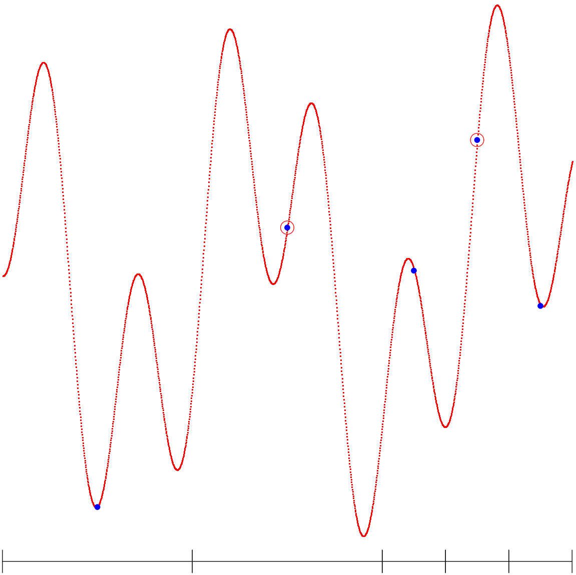
- Expand several leaves simultaneously
- SOO expands at most one leaf per depth
- SOO expands a leaf only if its value is larger that the value of all leaves of same or lower depths.
- At round t, SOO does not expand leaves with depth larger than $h_{\max}(t)$

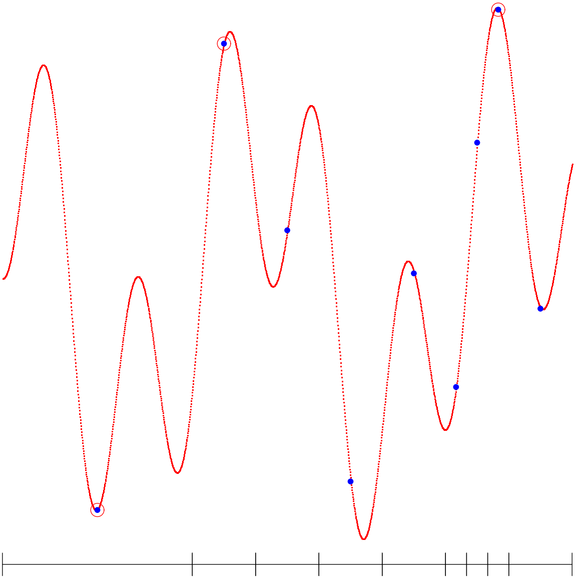
Be optimistic at all scales

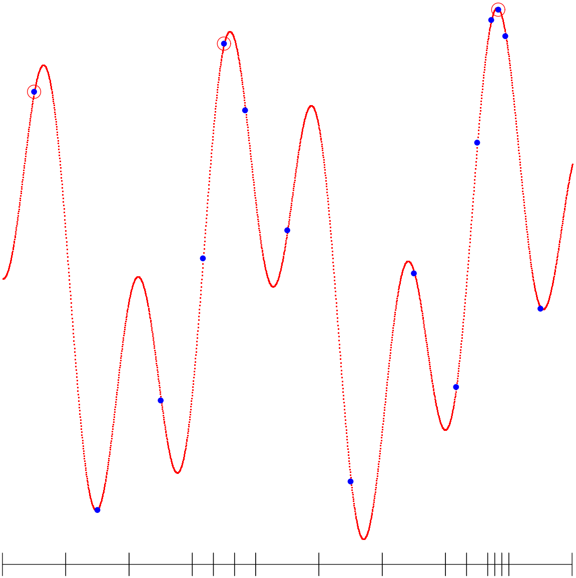
SOO algorithm

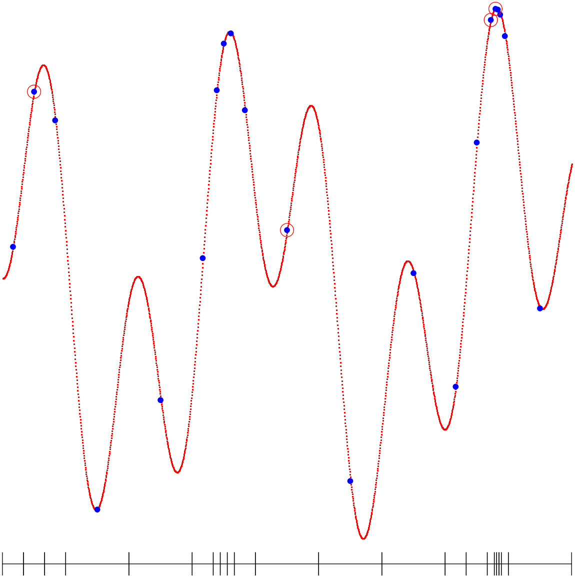
Input: the maximum depth function $t \mapsto h_{\max}(t)$
Initialization: $\mathcal{T}_1 = \{(0,0)\}$ (root node). Set $t = 1$.
while True **do**
  Set $v_{\max} = -\infty$.
  for $h = 0$ to $\min(\text{depth}(\mathcal{T}_t), h_{\max}(t))$ **do**
    Select the leaf $(h,j) \in \mathcal{L}_t$ of depth h with max $f(x_{h,j})$ value
    if $f(x_{h,i}) > v_{\max}$ **then**
      Expand the node (h,i), Set $v_{\max} = f(x_{h,i})$, Set $t = t + 1$
      if $t = n$ **then** return $x(n) = \arg\max_{(h,i)\in\mathcal{T}_n} x_{h,i}$
    end if
  end for
end while.

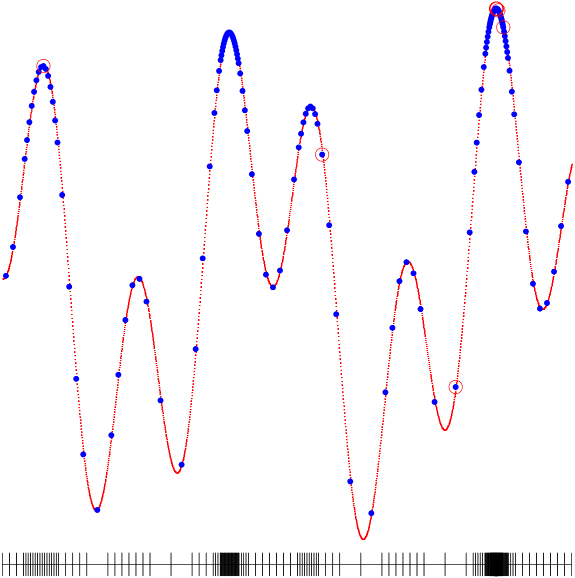


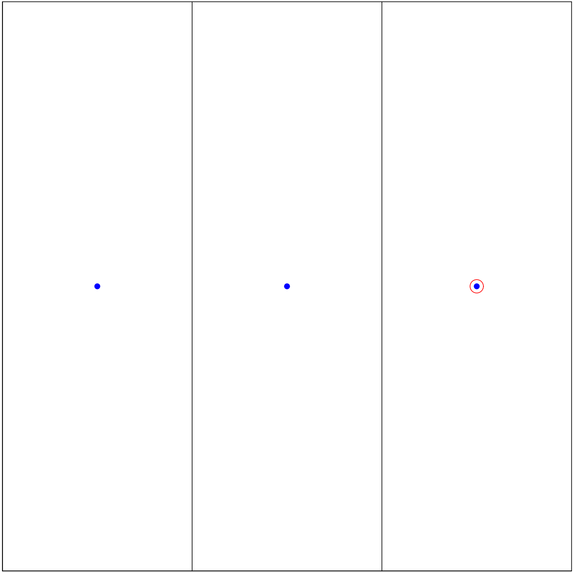


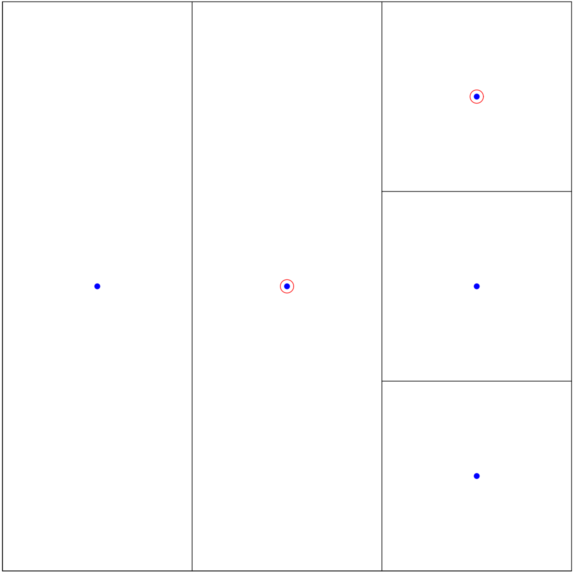


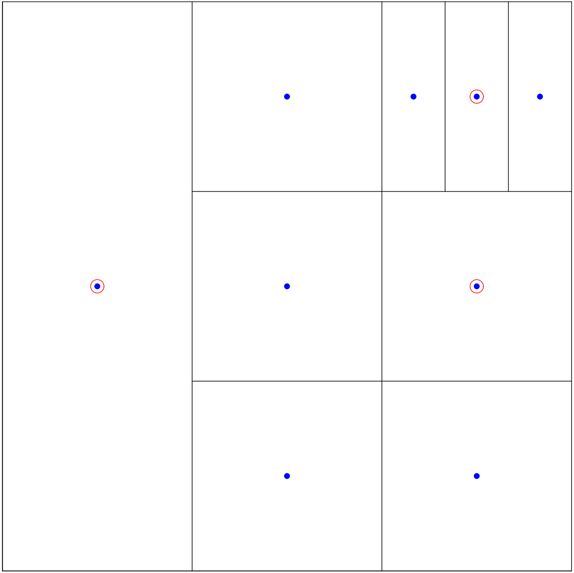


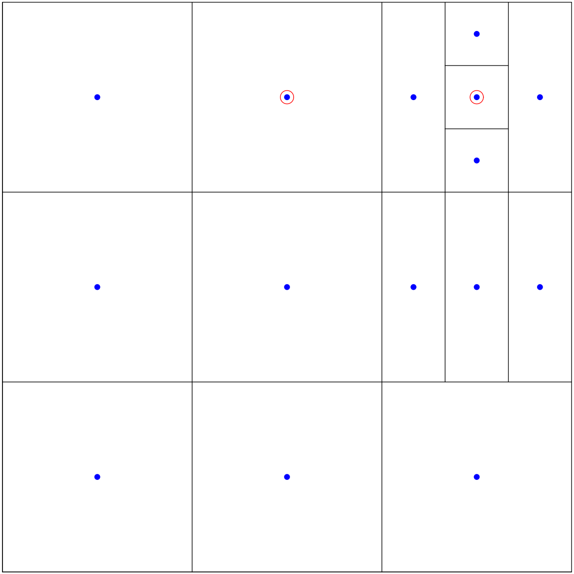


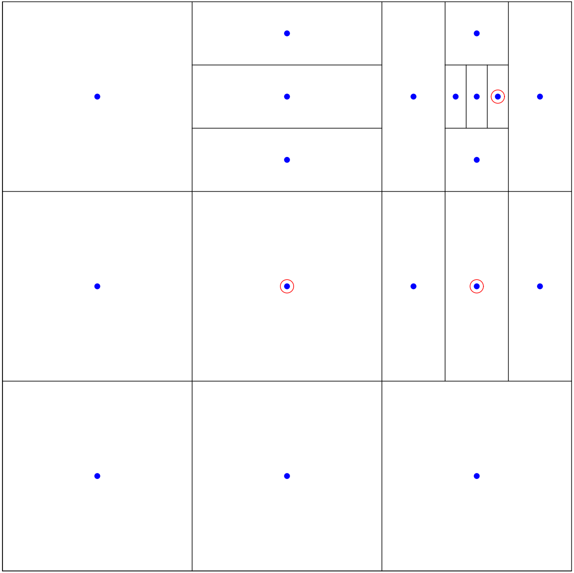


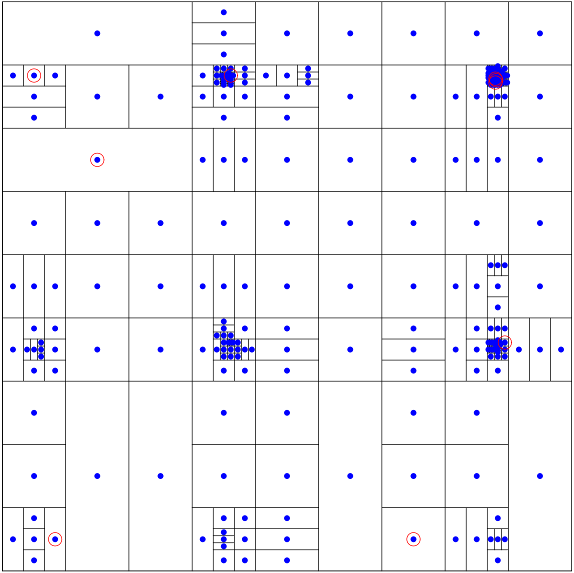












Performance of SOO

Theorem 2.
For any semi-metric ℓ such that

- *f is locally smooth w.r.t. ℓ*
- *The ℓ-diameter of cells of depth h is $c\gamma^h$*
- *The near-optimality dimension of f w.r.t. ℓ is $d = 0$,*

by choosing $h_{\max}(n) = \sqrt{n}$, the expected loss of SOO is

$$r_n \le c\gamma^{\sqrt{n}/C-1}$$

In the case $d > 0$ a similar statement holds with $\mathbb{E}r_n = \widetilde{O}(n^{-1/d})$.

Performance of SOO

Remarks:

- Since the algorithm does not depend on ℓ, the analysis holds for the best possible choice of the semi-metric ℓ satisfying the assumptions.
- **SOO does almost as well as DOO optimally fitted** (thus "adapts" to the unknown local smoothness of f).

Numerical experiments

Again for the function $f(x) = (\sin(13x)\sin(27x) + 1)/2$ we have:

n	uniform grid	DOO with ℓ_1	DOO with ℓ_2	SOO
50	1.25×10^{-2}	2.53×10^{-5}	1.20×10^{-2}	3.56×10^{-4}
100	8.31×10^{-3}	2.53×10^{-5}	1.67×10^{-7}	5.90×10^{-7}
150	9.72×10^{-3}	4.93×10^{-6}	4.44×10^{-16}	1.92×10^{-10}

The case $d = 0$ is non-trivial!

Example:

- f is locally α-smooth around its maximum:

$$f(x^*) - f(x) = \Theta(\|x^* - x\|^\alpha),$$

  for some $\alpha > 0$.
- SOO algorithm does not require the knowledge of ℓ,
- Using $\ell(x, y) = \|x - y\|^\alpha$ in the analysis, all assumptions are satisfied (with $\gamma = 3^{-\alpha/D}$ and $d = 0$, thus the loss of SOO is $r_n = O(3^{-\sqrt{n}\alpha/(CD)})$ (stretched-exponential loss),
- This is almost as good as DOO optimally fitted!

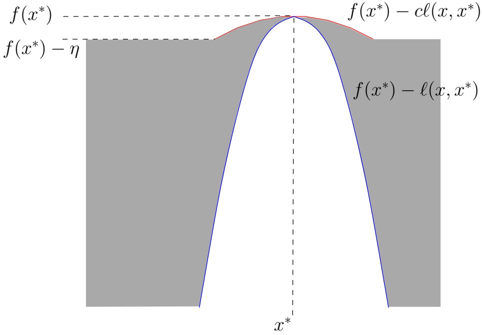
(Extends to the case $f(x^*) - f(x) \approx \sum_{i=1}^{D} c_i |x_i^* - x_i|^{\alpha_i}$)

The case $d = 0$

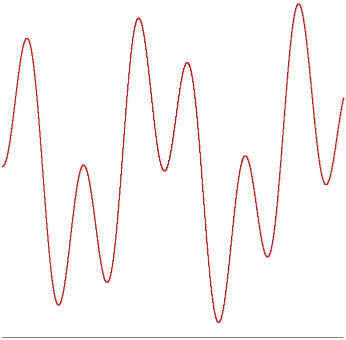
More generally, any function whose **upper- and lower envelopes around** x^* **have the same shape**: $\exists c > 0$ and $\eta > 0$, such that

$$\min(\eta, c\ell(x, x^*)) \leq f(x^*) - f(x) \leq \ell(x, x^*), \quad \text{for all } x \in \mathcal{X}.$$

has a near-optimality $d = 0$ (w.r.t. the metric ℓ).

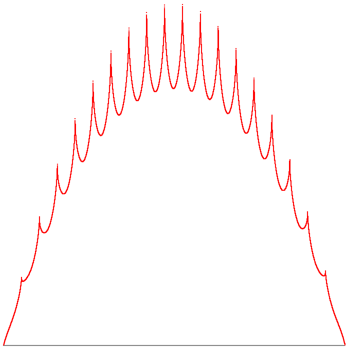


Example of functions for which $d = 0$



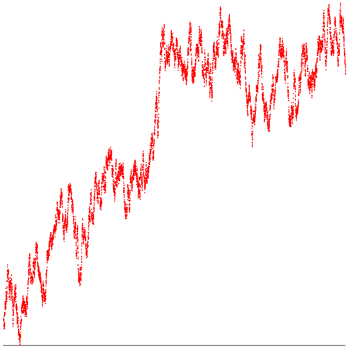
$$\ell(x, y) = c\|x - y\|^2$$

Example of functions with $d = 0$



$$\ell(x, y) = c\|x - y\|^{1/2}$$

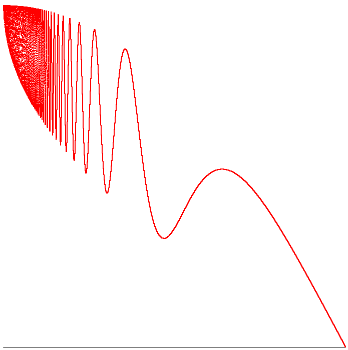
$d = 0$?



$$\ell(x, y) = c\|x - y\|^{1/2}$$

$d > 0$

$$f(x) = 1 - \sqrt{x} + (-x^2 + \sqrt{x}) * (\sin(1/x^2) + 1)/2$$



The lower-envelope is of order $1/2$ whereas the upper one is of order 2. We deduce that $d = 3/2$ and $r_n = O(n^{-2/3})$.

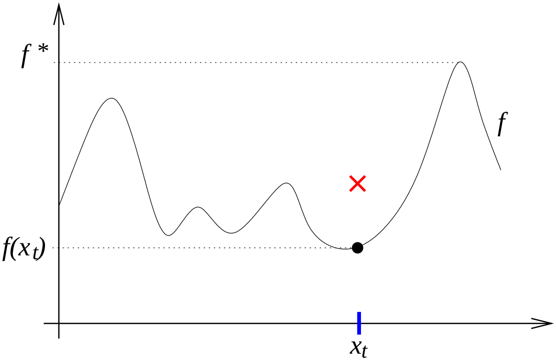
SOO versus DIRECT

- **SOO is much more general than DIRECT**: the function is only locally smooth and the space is semi-metric.
- **Finite-time analysis of SOO**
- **SOO is a rank-based algorithm**: any transformation of the values while preserving their rank will not change anything in the algorithm. Thus extends to the optimization of function givens pair-wise comparisons.

How to handle noise?

The evaluation of f at x_t is perturbed by noise:

$$y_t = f(x_t) + \epsilon_t, \text{ with } \mathbb{E}[\epsilon_t | x_t] = 0.$$



Stochastic SOO (StoSOO)

Extends SOO to stochastic evaluations:

- Select the cells X_i (at most one per depth) according to SOO based on the UCBs:

$$\widehat{\mu}_{i,t} + c\sqrt{\frac{\log n}{T_i(t)}},$$

  and get one more value $y_t = f(x_i) + \epsilon_t$ of f at x_i,
- If $T_i(t) \geq k$, then split the cell X_i.

Remark: This really looks like UCT, except that

- several cells are selected at each round,
- a cell is split only after observing k values.

Performance of StoSOO

Theorem 3 (Valko et al., 2013).

For any semi-metric ℓ such that

- *f is locally smooth w.r.t. ℓ*
- *The ℓ-diameters of the cells decrease exponentially fast with their depth,*
- *The near-optimality dimension of f w.r.t. ℓ is $d = 0$,*

by choosing $k = \frac{n}{(\log n)^3}$, $h_{\max}(n) = (\log n)^{3/2}$, the expected loss of StoSOO is

$$\mathbb{E} r_n = O\left(\frac{(\log n)^2}{\sqrt{n}}\right).$$

This is almost as good as HOO [Bubeck et al., 2011] and Zooming [Kleinberg et al., 2008] optimally fitted! Complementary to the adaptive-treed bandits of [Bull, 2013].

Range of application

All illustrations are in Euclidean spaces $[0,1]^D$ only.

But there are many other semi-metric spaces...

- Trees (games, ...)
- Graphs (social networks, ...),
- Combinatorial spaces (shortest paths problems, ...)
- Other structured spaces (policies in MDPs, ...)

We only require:

- the search space $\mathcal{X}$ to be equipped with a semi-metric ℓ,
- a nested (hierarchical) partitioning of the space,
- f to satisfy a local smoothness property w.r.t. ℓ,
- ℓ may or may not be known.

Conclusions

Provide a **measure of the complexity of optimization**.

This **multi-scale optimistic optimization**

- provides an efficient exploration of the search space by exploring the most promising areas first
- provides a natural transition from global to local search
- Performance depends on the “smoothness” of the function around the maximum w.r.t. some metric,
  - and a measure of the quantity of near-optimal solutions,
  - and our knowledge or not of this smoothness.

Thanks !!!

See the review paper

From bandits to Monte-Carlo Tree Search: The optimistic principle applied to optimization and planning.

from my web page:

`http://chercheurs.lille.inria.fr/~munos/`