PARAMETER-FREE AND ADAPTIVE OPTIMIZATION UNDER MINIMAL ASSUMPTIONS

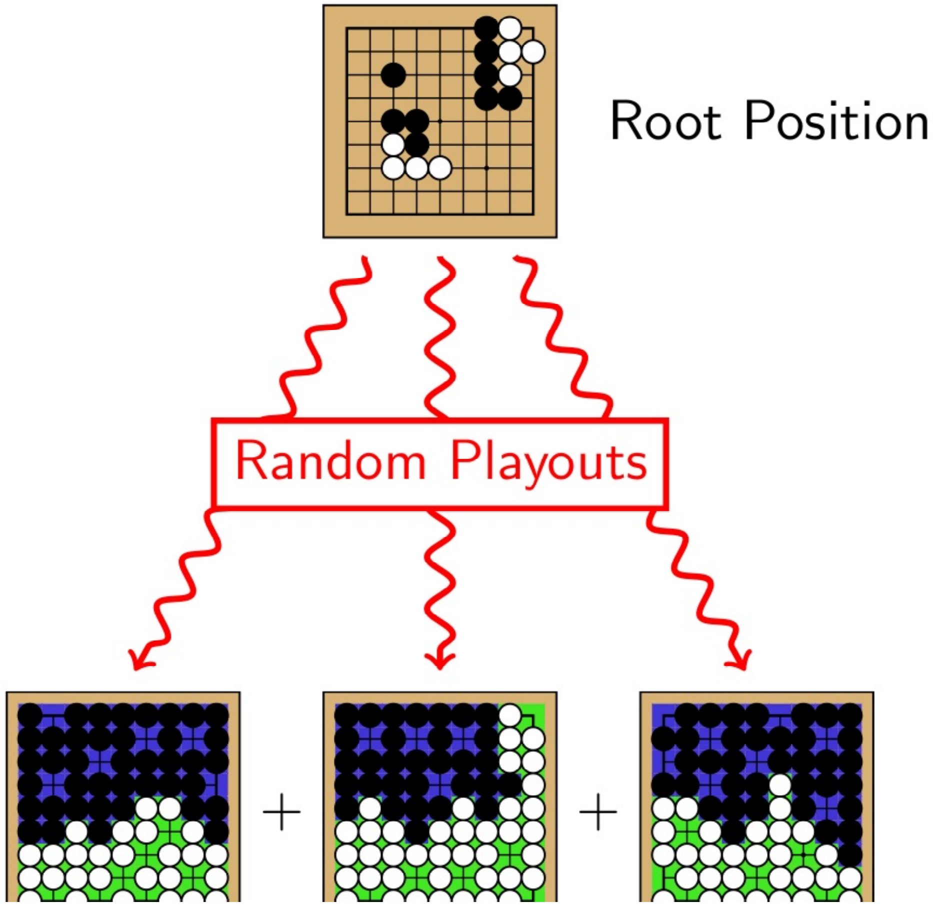
Michal Valko

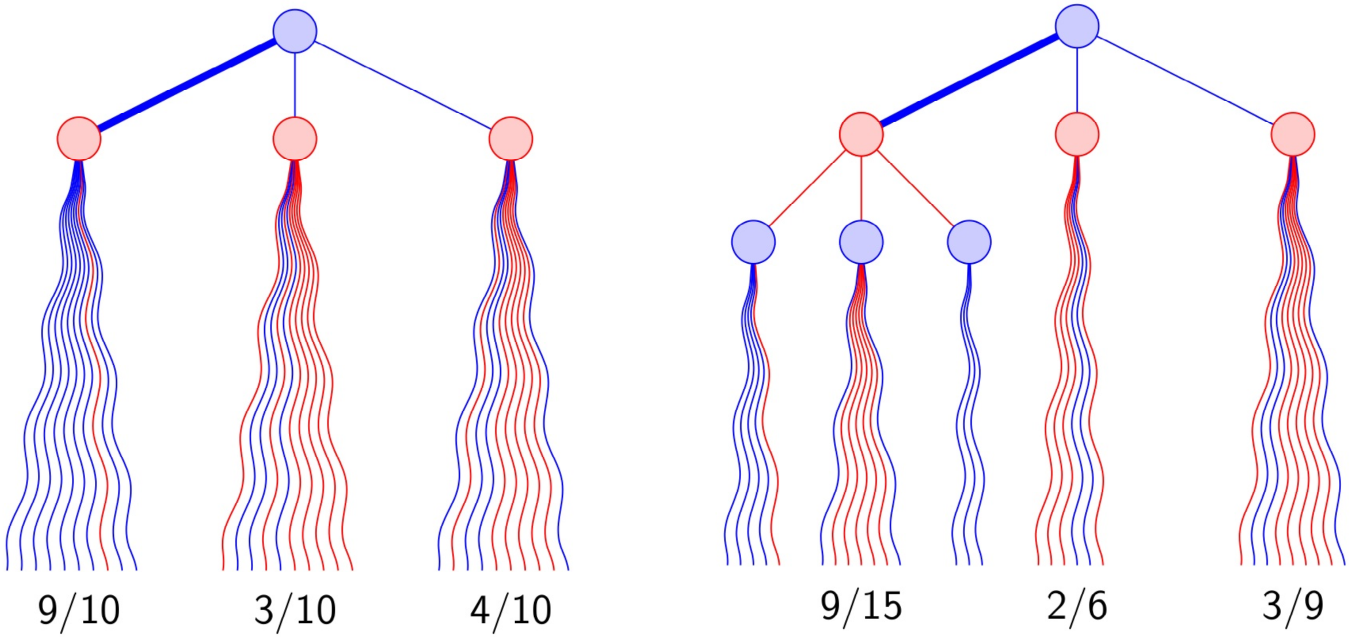
with Peter Bartlett and Victor Gabillon

SequeL @ Inria Lille — Nord Europe

inventors for the digital world

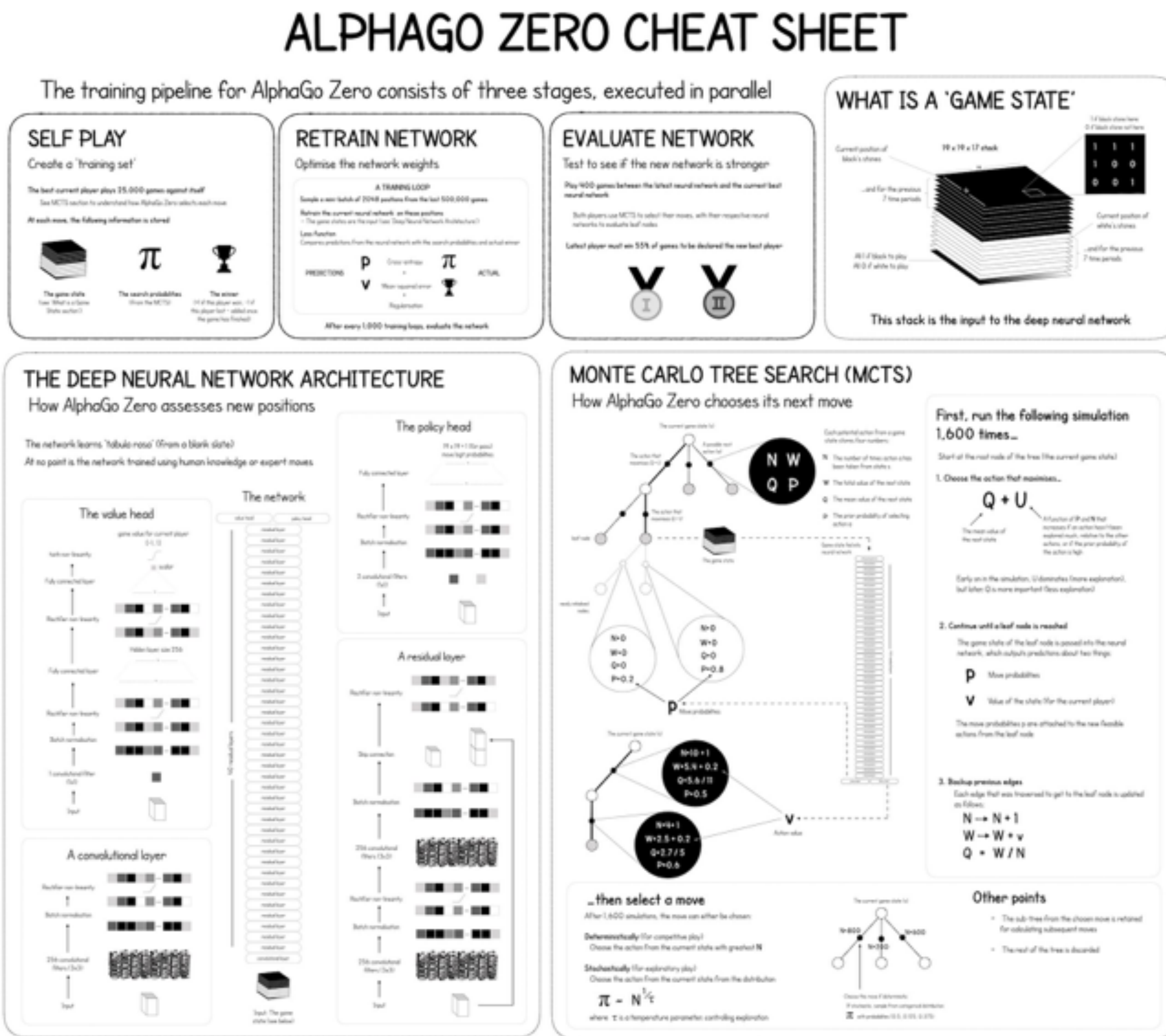
MCTS IN COMPUTER GO





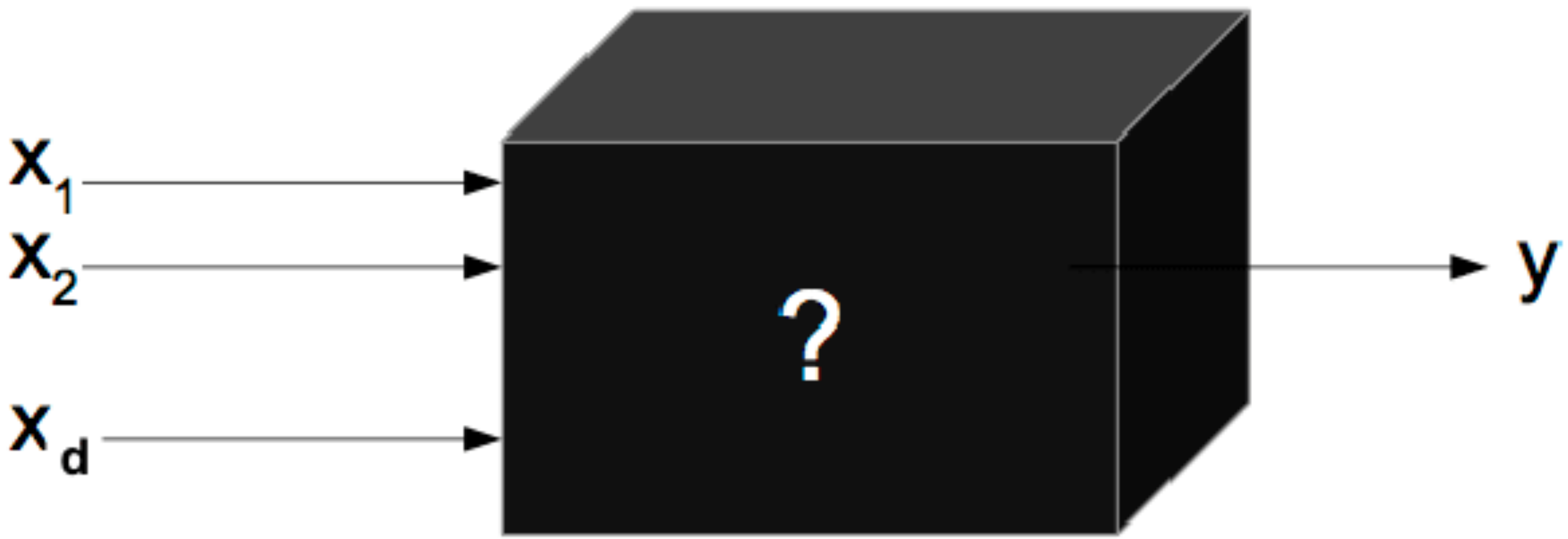
Munos: From bandits to Monte-Carlo Tree Search: The optimistic principle applied to optimization and planning, 2014

MOGO — CRAZY STONE — ALPHAGO(0)



https://medium.com/applied-data-science/alphago-zero-explained-in-one-diagram-365f5abf67e0

OPTIMIZE THIS!



BIG QUESTIONS?

How black-box is black-box?

What can black-box optimization guarantee?

What are the minimal assumptions?

What are the absolutely minimal assumptions?

SETTING

- **Goal:** Maximize $f : \mathcal{X} \to \mathbb{R}$ given a budget of n evaluations.
- **Challenges:** f is ***stochastic*** and has ***unknown smoothness***
- **Protocol:** At round t, select state x_t, observe r_t such that

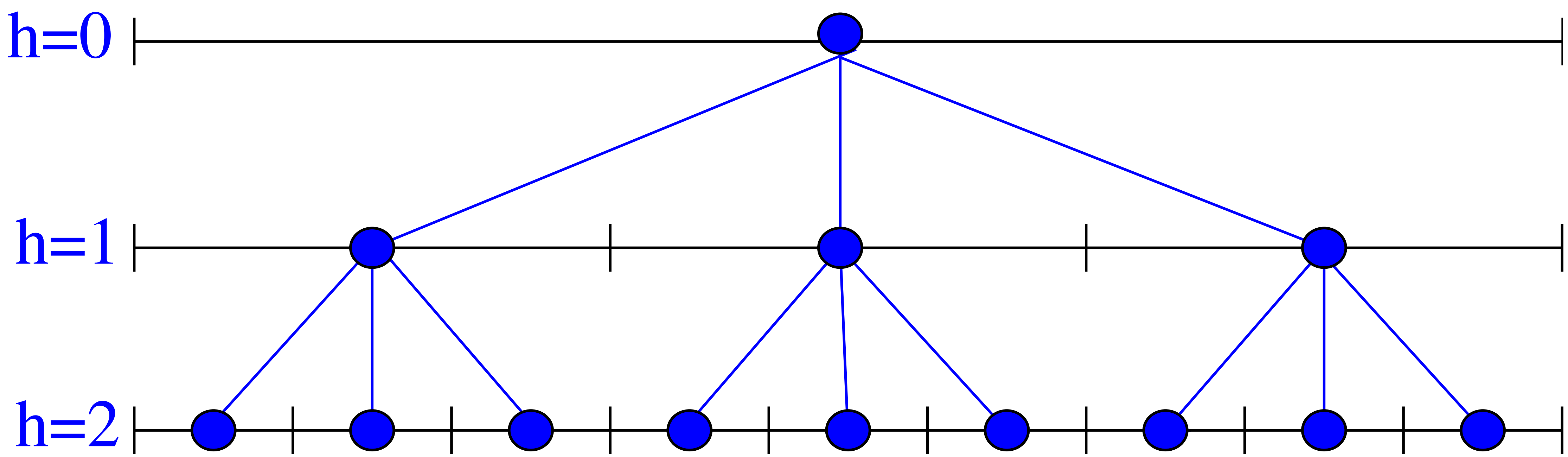
$$\mathbb{E}[r_t | x_t] = f(x_t).$$

After n rounds, return a state $x(n)$.

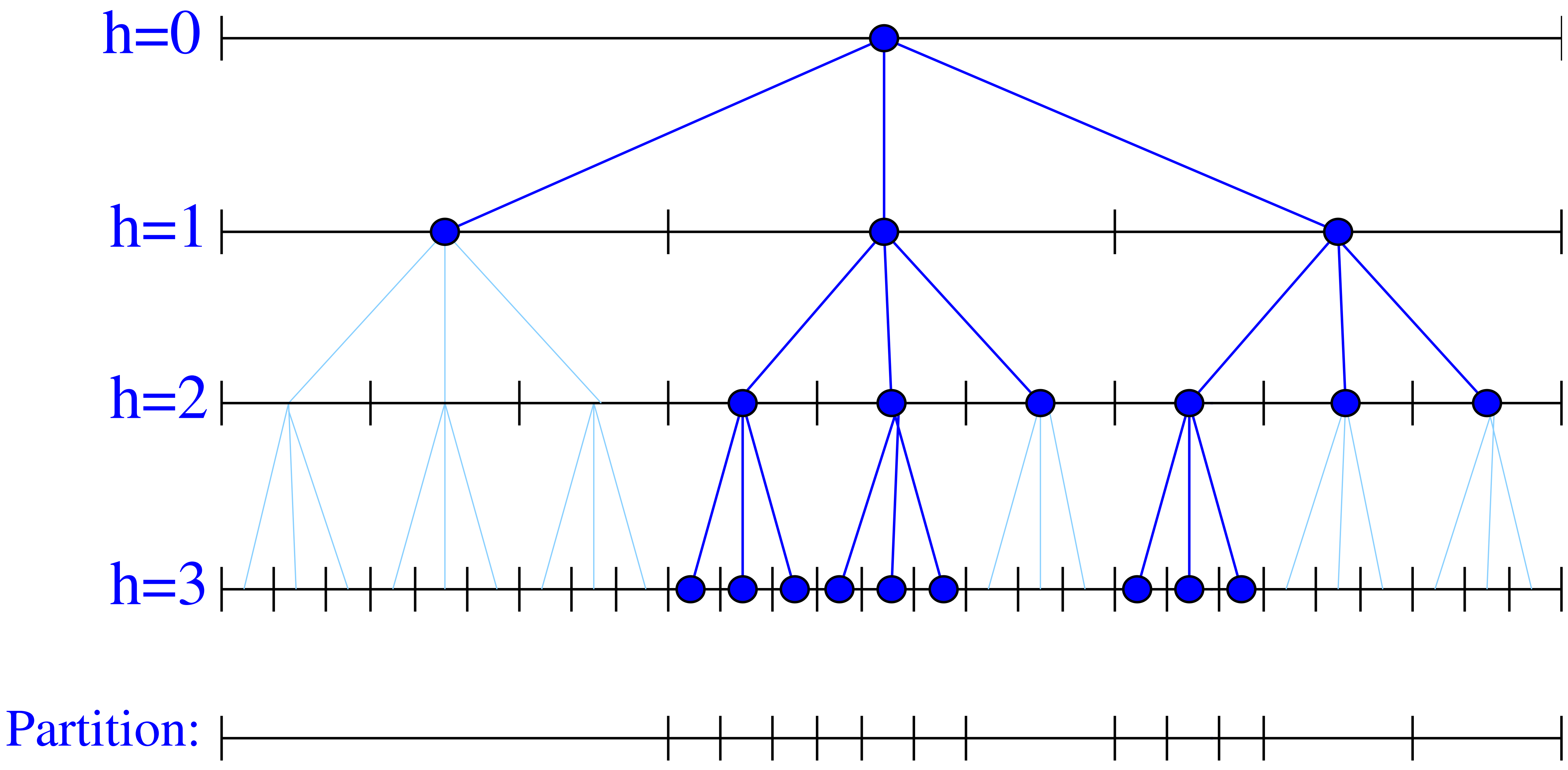
- **Loss:** $R_n = \sup_{x \in \mathcal{X}} f(x) - f(x(n))$

PARTITIONING: 1D

- For any h, $\mathcal{X}$ is partitioned in K^h cells $(X_{h,i})_{0 \leq i \leq K^h - 1}$.
- K-ary tree $\mathcal{T}_\infty$ where depth $h = 0$ is the whole $\mathcal{X}$.

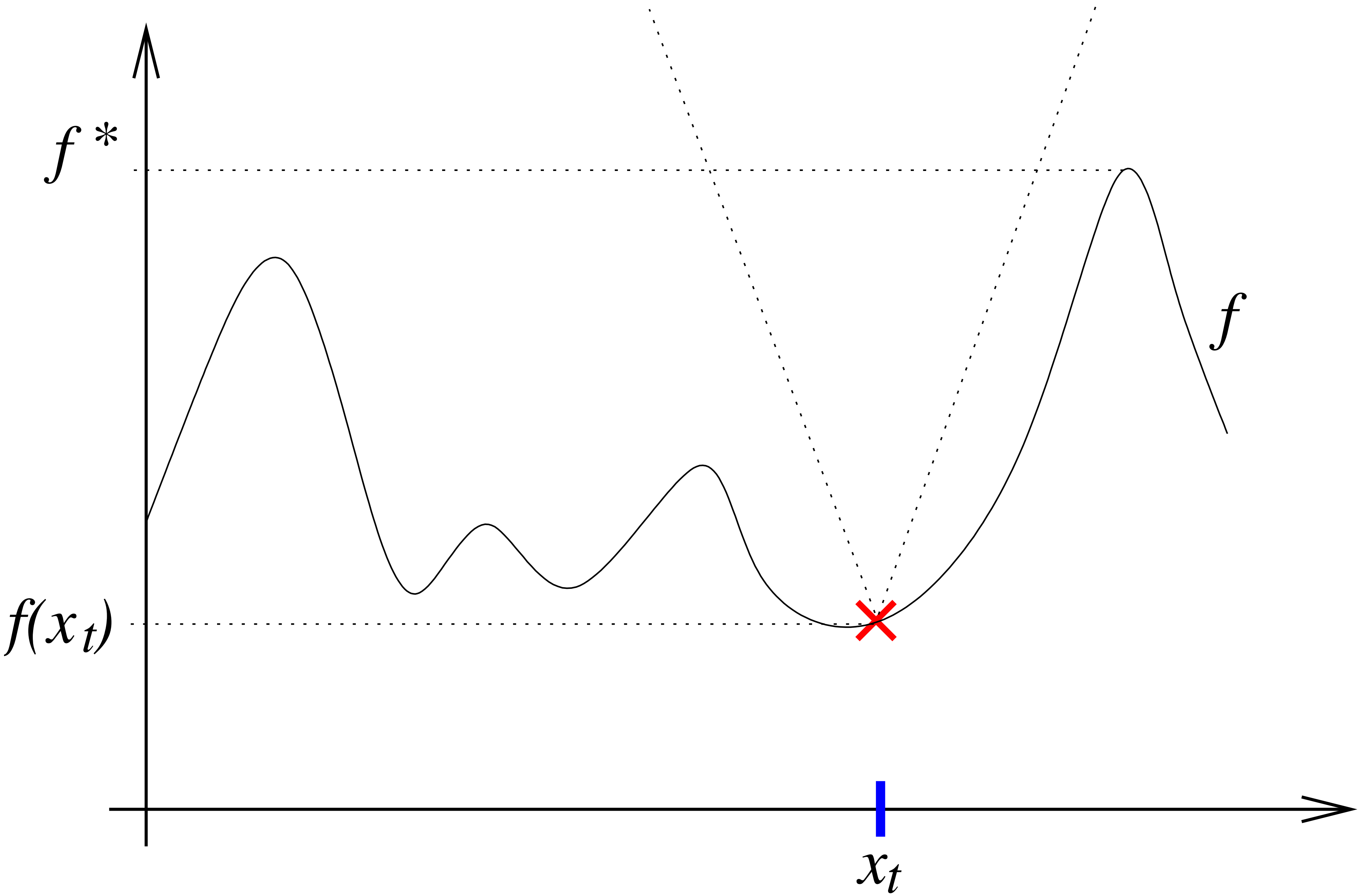


HOW IT WORKS?



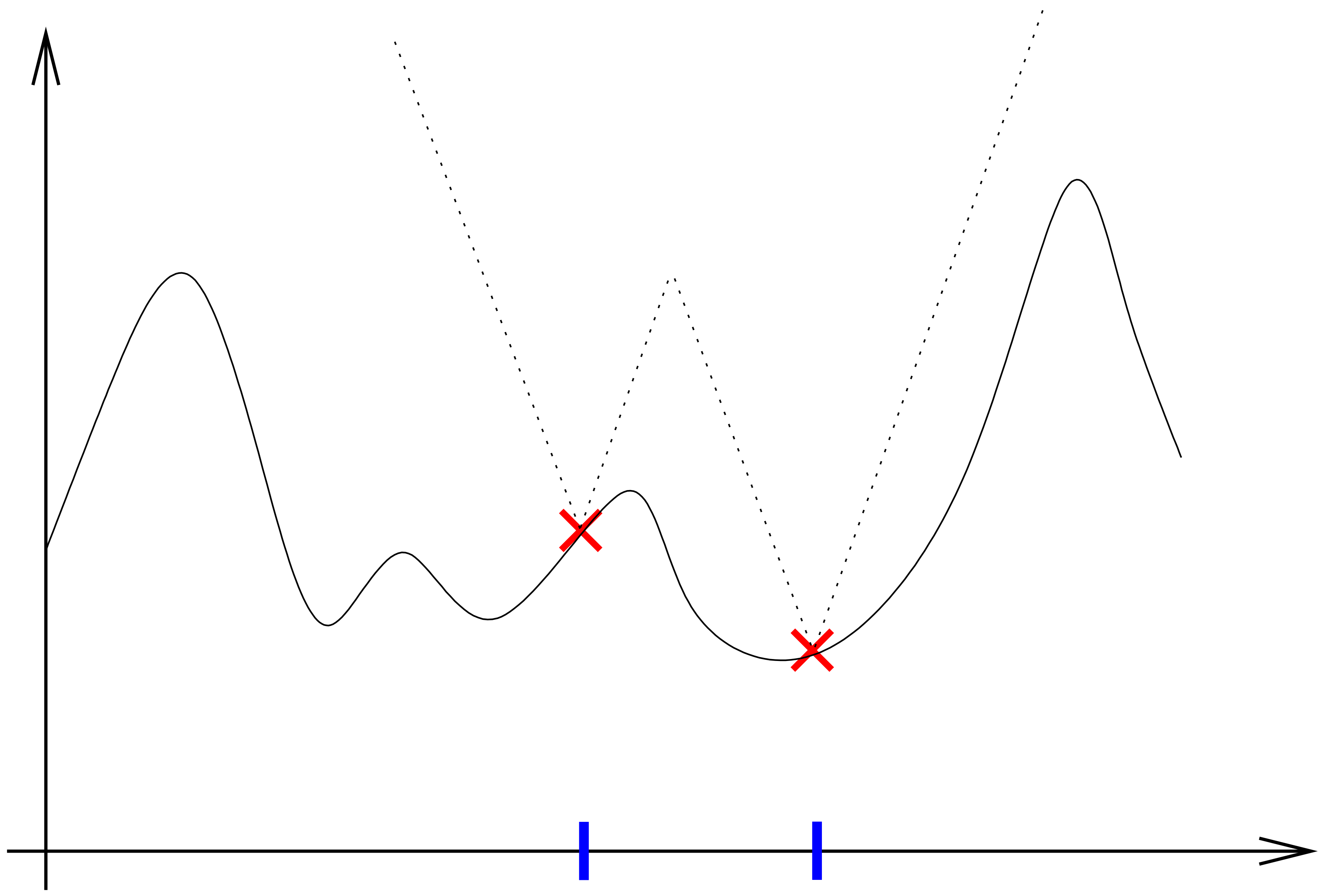
PARTITIONING: 2D

EXAMPLE: 1D



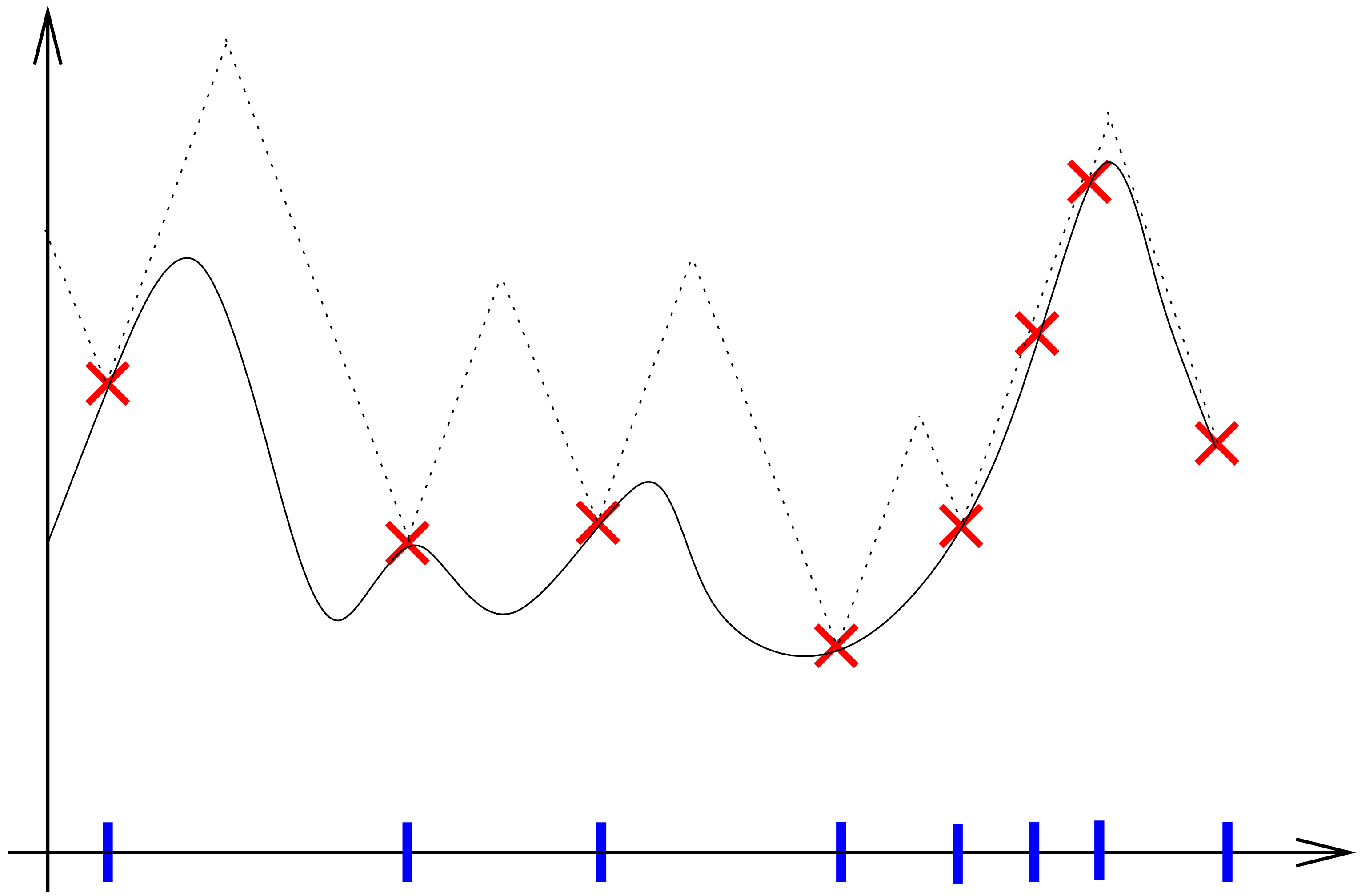
Lipschitz property → the evaluation of f at x_t provides a first upper-bound on f

EXAMPLE: 1D



New point → refined upper-bound on f

EXAMPLE: 1D



Question: where should one sample the next point?

Answer: select the point with highest upper bound!

GLOBAL OPTIMIZERS

a ZOO of possibilities

very few guarantee a global optimality

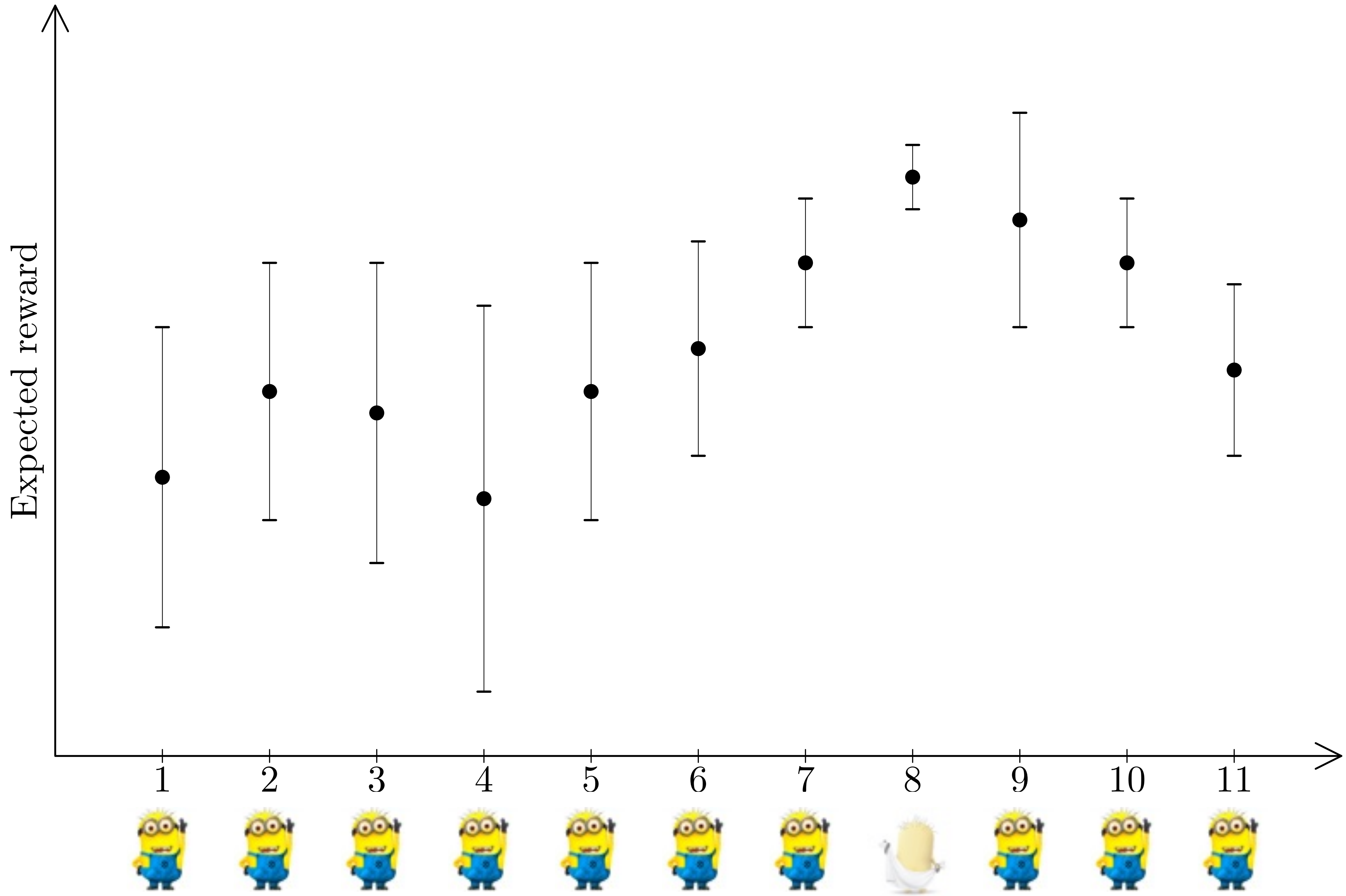
smoothness	**deterministic**	**stochastic**
known	DOO	Zooming, HOO
unknown	DiRect, SOO, SequOOL	StoSOO, POO, StroquOOL

Which functions are difficult to optimize?

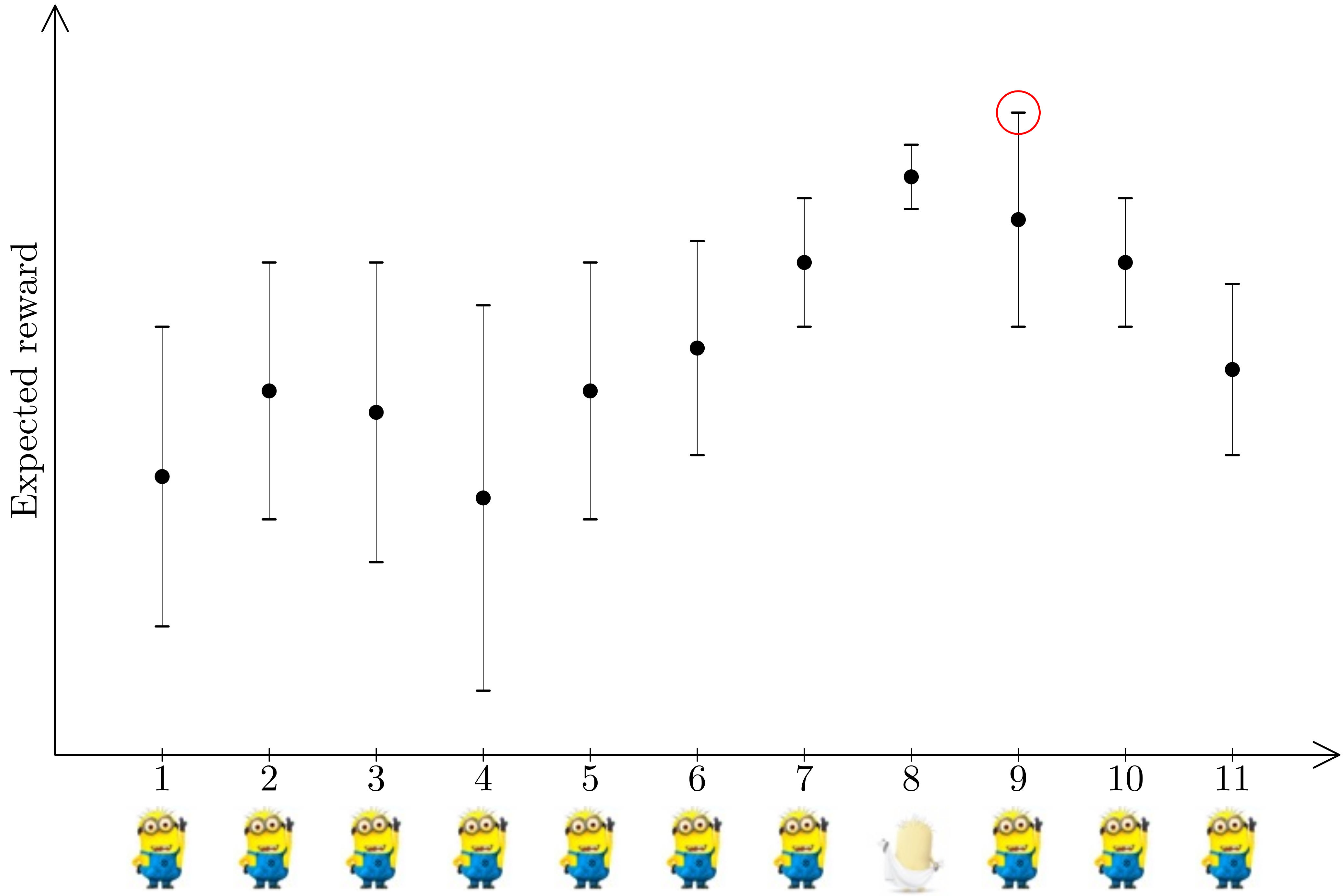
What is the right characterization of the problem?

minimax-optimal sample complexity

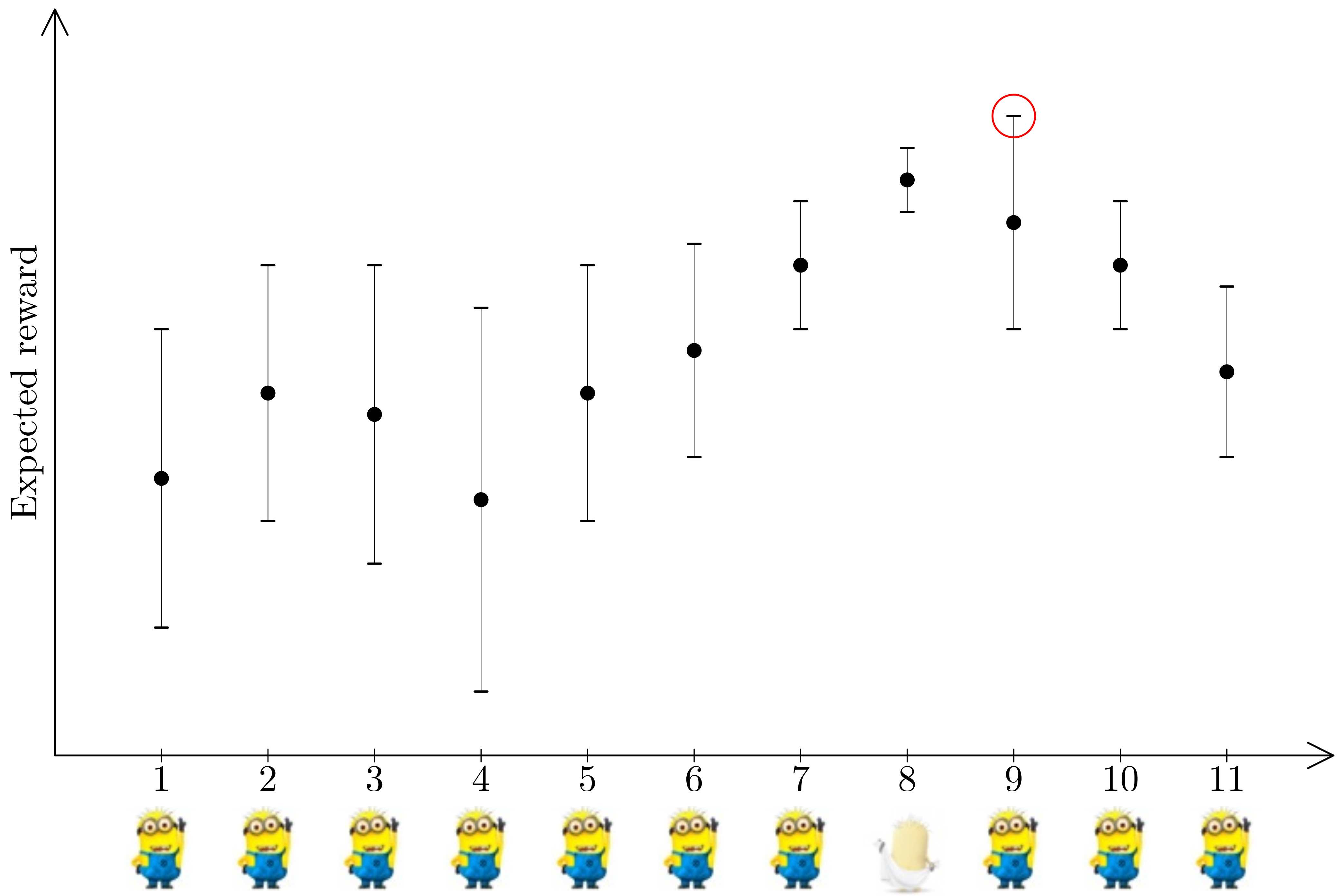
UPPER CONFIDENCE BOUND BASED ALGOS



UPPER CONFIDENCE BOUND BASED ALGOS

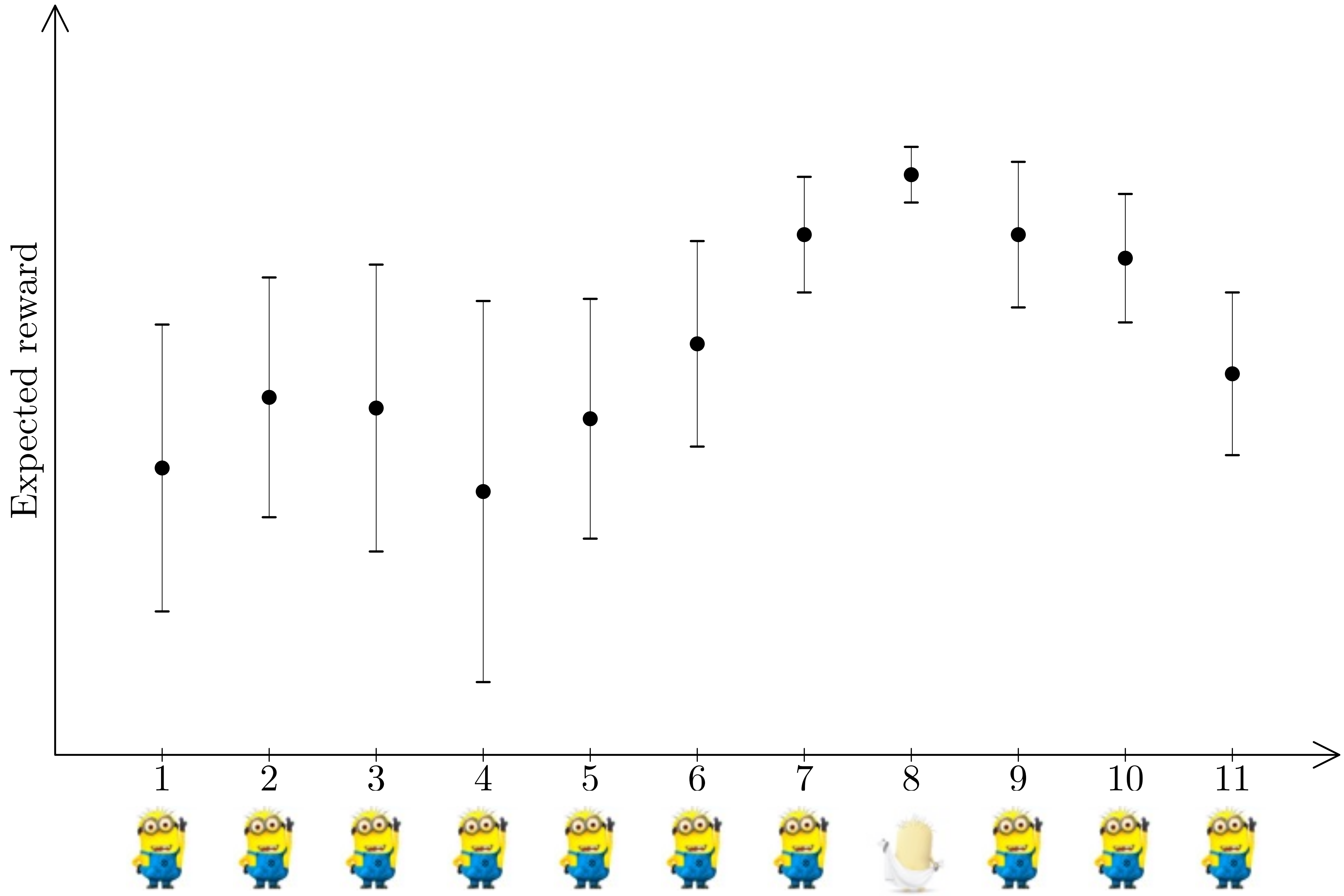


UPPER CONFIDENCE BOUND BASED ALGOS



UPPER CONFIDENCE BOUND BASED ALGOS

VIDEO EXAMPLES FOR THE CONTINUOUS FUNCTION OPTIMIZATION



COMPLICATED HISTORY

weak Lipschitz | local Lipschitz only

adapted (semi-metric) | no metric needed

needs scale of the noise as an input | adapts to noise

HOO 2010

DOO 2011

SOO 2011

StoSOO 2013

HCT 2014

POO 2015

adapts to smoothness

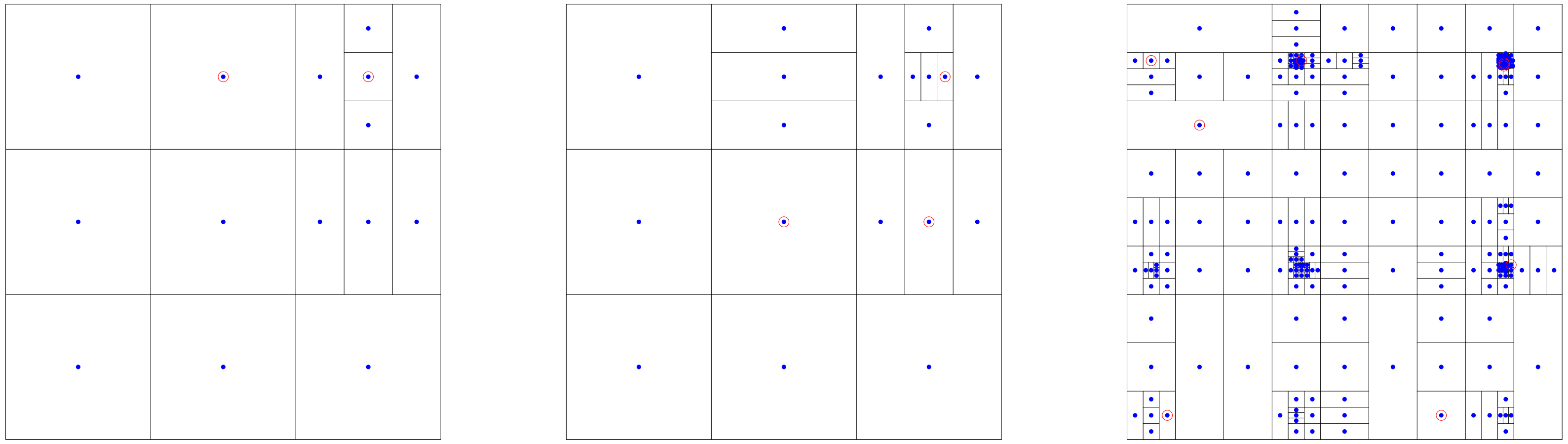
SequooL 2019

StroquOOL 2019

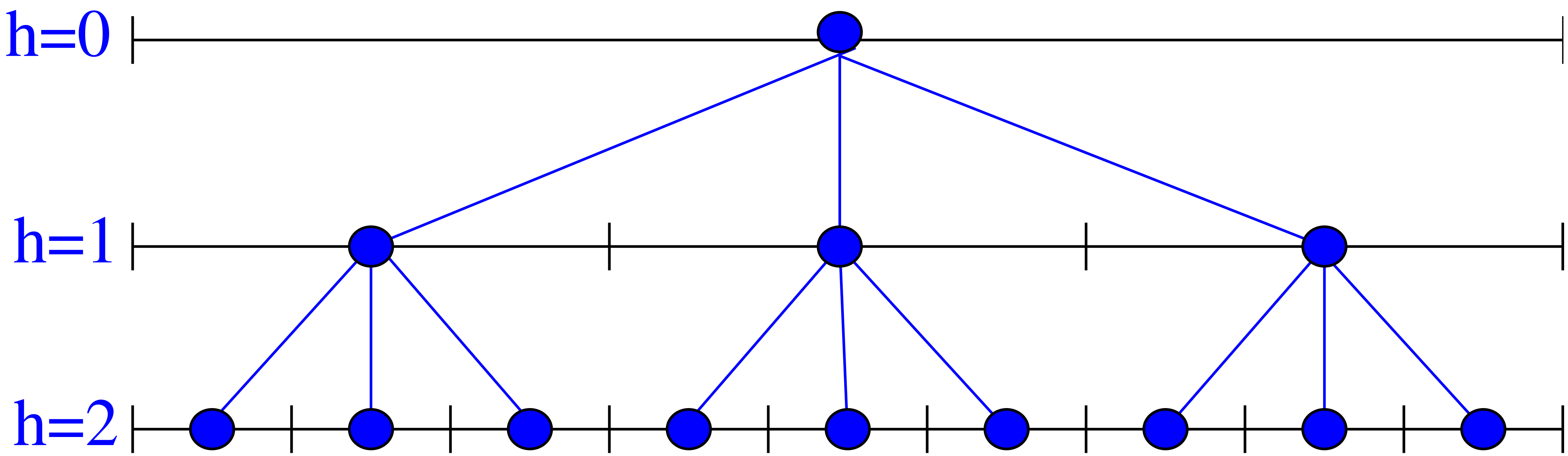
WHAT DOES OUR ALGORITHM BRING?

- **Current state-of-the-art needs noise scale as input**
  - If the noise is actually smaller, we find the optimum slower than we could
  - if the input happens to be deterministic, **we miss learning exponentially fast**
- **Current state-of-the-art is are complicated META-ALGORITHM**
  - explicitly running several algorithms that know the smoothness
  - VERY complicated analysis, high computational complexity

What is the price to pay for all this adaptivity and minimal assumptions?



hyper-parameter optimization!



Assumption 1 *For any global optimum* $x^\star$*, there exists* $\nu > 0$ *and* $\rho \in (0,1)$ *such that* $\forall h \in \mathbb{N}$, $\forall x \in \mathcal{P}_{h,i_h^\star}, f(x) \geq f(x^\star) - \nu\rho^h$.

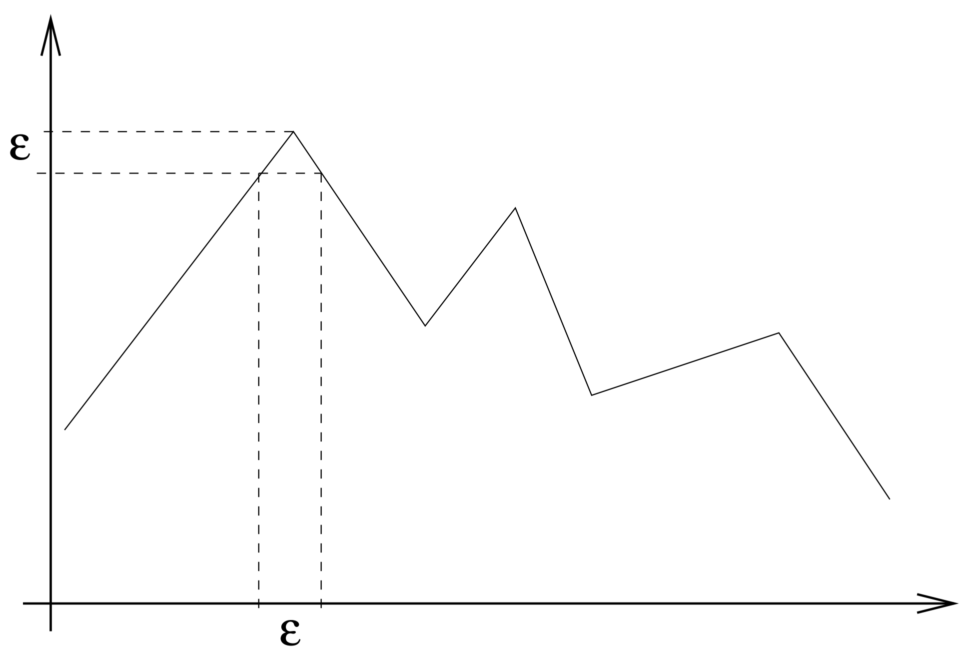
Definition 1 *For any* $\nu > 0$ *and* $\rho \in (0,1)$*, the* ***near-optimality dimension***[3] $d(\nu, \rho)$ *of* f *with respect to the partitioning* $\mathcal{P}$ *and with associated constant* C*, is*

$$d(\nu, \rho) \triangleq \inf\left\{d' \in \mathbb{R}^+ : \exists C > 1, \forall h \geq 0, \mathcal{N}_h(3\nu\rho^h) \leq C\rho^{-dh}\right\},$$

where $\mathcal{N}_h(\varepsilon)$ *is the number of cells* $\mathcal{P}_{h,i}$ *of depth* h *such that* $\sup_{x \in \mathcal{P}_{h,i}} f(x) \geq f(x^\star) - \varepsilon$.

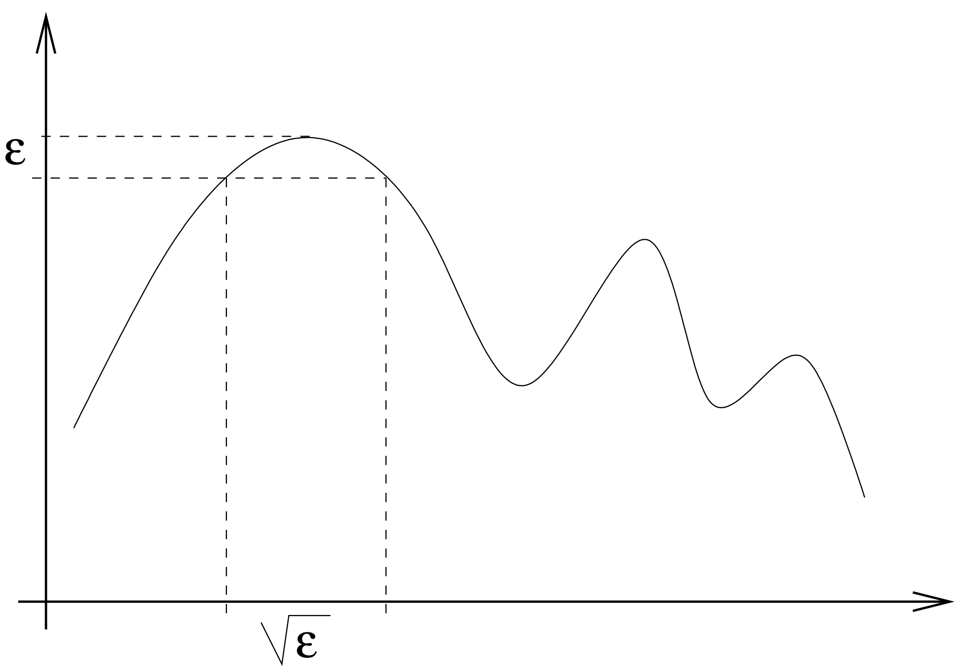
D=0

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



$$\ell(x,y) = ||x-y|| \rightarrow d = 0$$

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

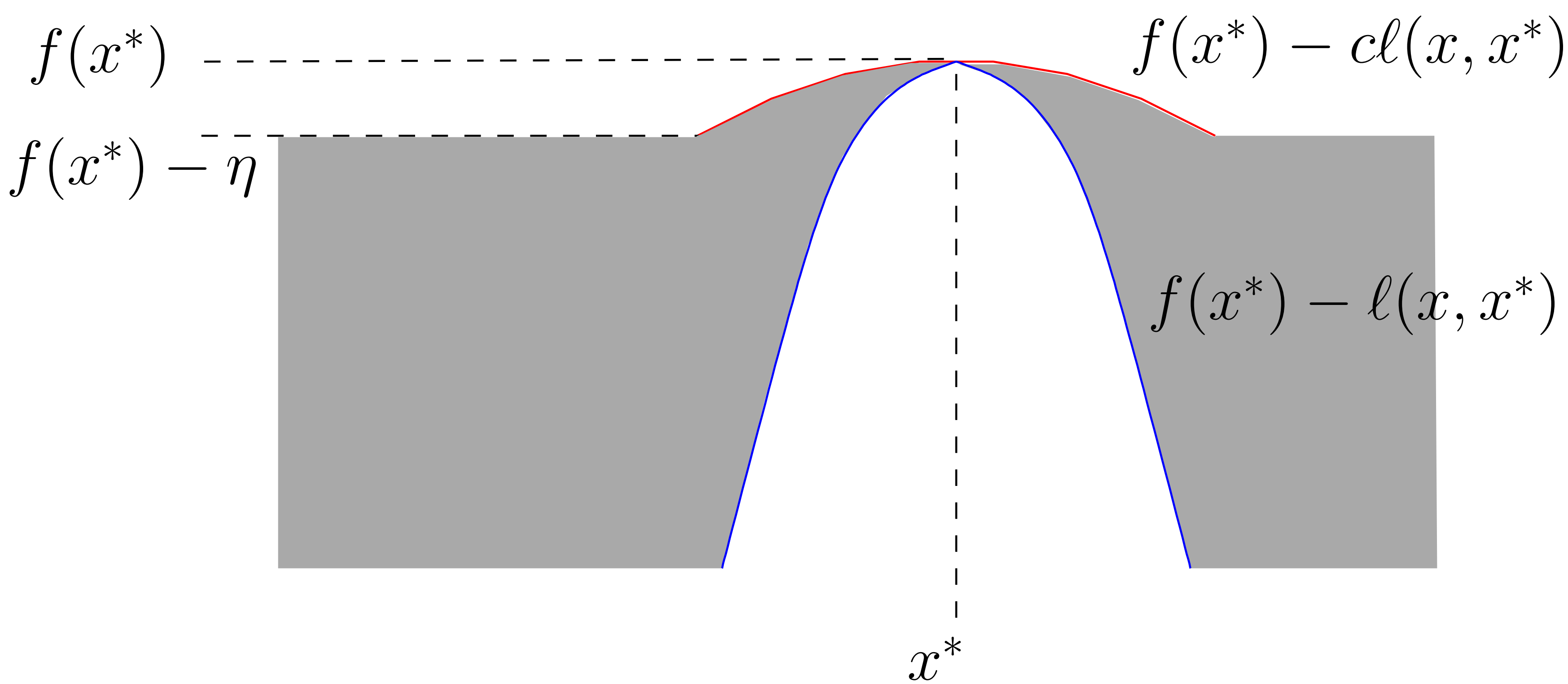


$$\ell(x,y) = ||x-y|| \rightarrow d = D/2$$
$$\ell(x,y) = ||x-y||^2 \rightarrow d = 0$$

D=0

Let a function in such space have upper- and lower envelope around x^* of the same order, i.e., there exists constants $c \in (0,1)$, and $\eta > 0$, such that for all $x \in \mathcal{X}$:

$$\min(\eta, c\ell(x,x^*)) \le f(x^*) - f(x) \le \ell(x,x^*). \qquad (1)$$

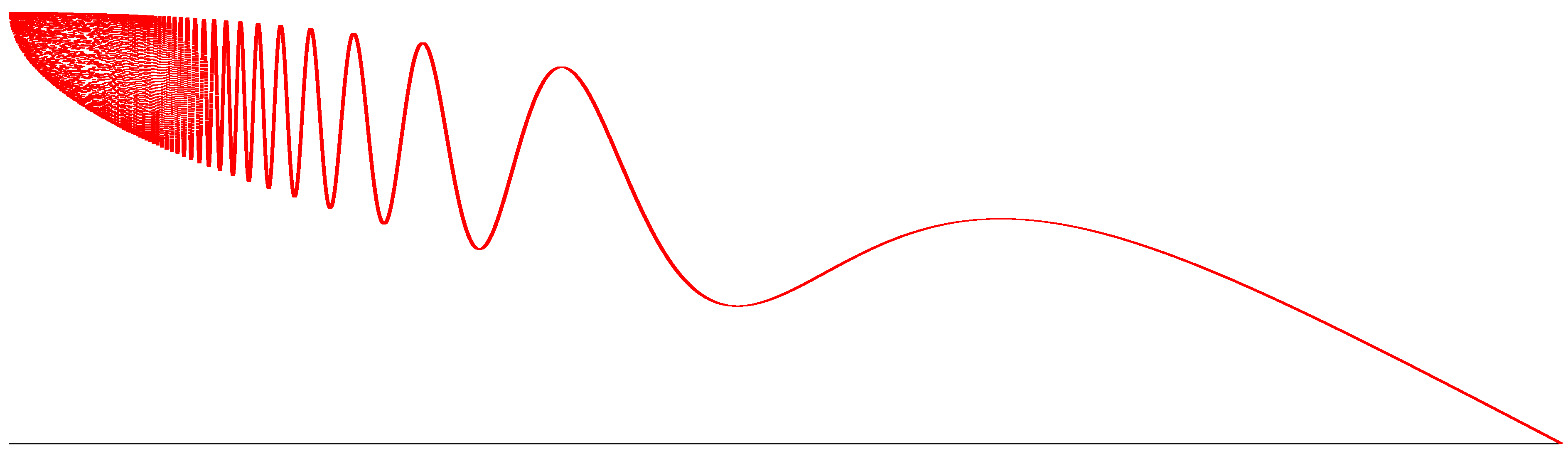


Any function satisfying (1) lies in the gray area and possesses a lower- and upper-envelopes that are of same order around x^*.

D>0

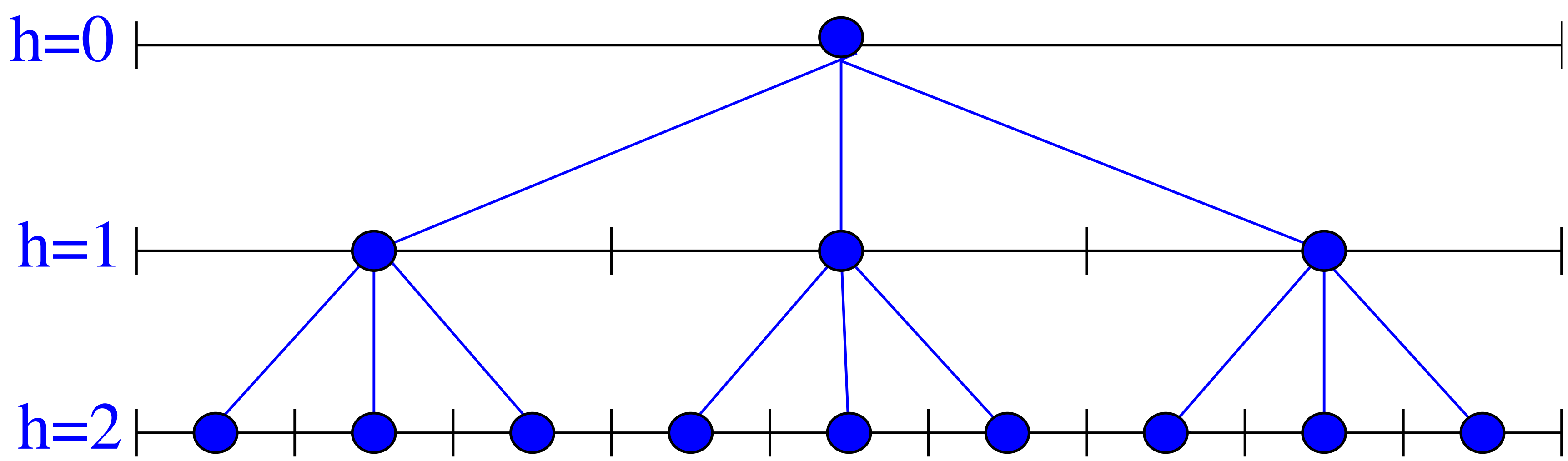
Example of a function with different order in the upper and lower envelopes, when $\ell(x, y) = |x - y|^\alpha$:

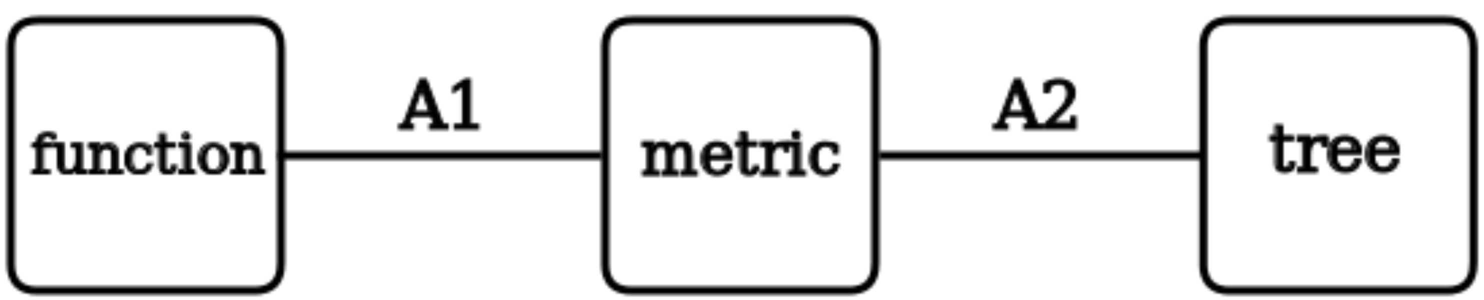
$$f(x) = 1 - \sqrt{x} + (-x^2 + \sqrt{x}) \cdot (\sin(1/x^2) + 1)/2$$

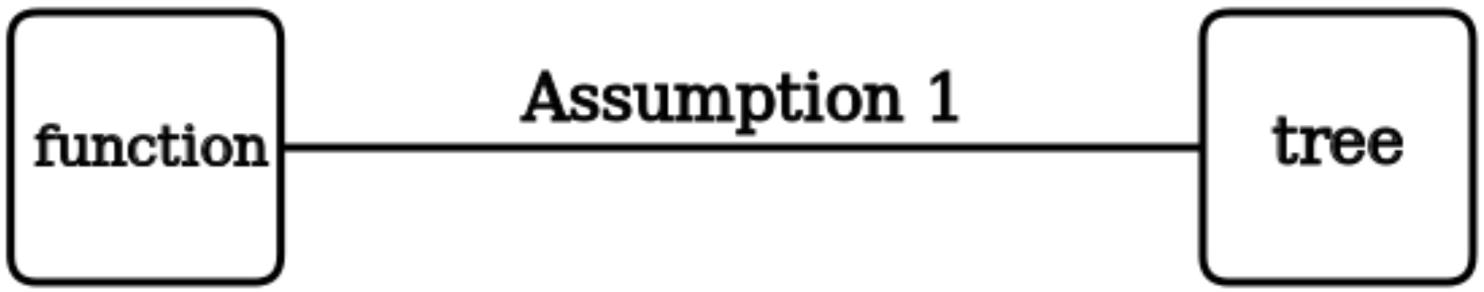


The lower-envelope behaves like a square root whereas the upper one is quadratic. There is no semi-metric of the form $|x - y|^\alpha$ for which $d < 3/2$.

GRILL, V., MUNOS, NIPS 2015







Parameters: n, $\mathcal{P} = \{\mathcal{P}_{h,i}\}$

Initialization: Open $\mathcal{P}_{0,1}$. $h_{\max} = \left\lfloor \frac{n}{\overline{\log}(n)} \right\rfloor$.

For $h = 1$ to $h_{\max}$

Open $\left\lfloor \frac{h_{\max}}{h} \right\rfloor$ cells $\mathcal{P}_{h,i}$ of depth h with largest values $f_{h,j}$.

Output $x(n) = \arg\max_{x_{h,i}:\mathcal{P}_{h,i} \in \mathcal{T}} f_{h,i}$.

as we go deeper, we open less cells per depth

Number of evaluations:

$$1 + \sum_{h=1}^{h_{\max}} \left\lfloor \frac{h_{\max}}{h} \right\rfloor \leq 1 + h_{\max} \sum_{h=1}^{h_{\max}} \frac{1}{h} = 1 + h_{\max} \overline{\log}\, h_{\max} \leq n + 1$$

SEQUOOL

OBSERVATION: The deeper we go, the better optimum we find.

Lemma 2 *For any global optimum $x^\star$ with associated (ν,ρ) as defined in Assumption 1, for any depth $h \in [h_{\max}]$, if $\frac{h_{\max}}{h} \ge C\rho^{-d(\nu,\rho)h}$, we have $\perp_h = h$, while $\perp_0 = 0$.*

SUMMARY: We go deep enough

MAIN RESULT

Theorem 3 *Let W be the standard Lambert W function (see Section 2). For any function f and one of its global optima $x^\star$ with associated (ν,ρ), and near-optimality dimension $d = d(\nu,\rho)$, we have, after n rounds, the simple regret of* `SequOOL` bounded by

- If $d=0$, $\quad r_n \le \nu\rho^{\frac{1}{C}\left\lfloor\frac{n}{\overline{\log} n}\right\rfloor}$.
- If $d>0$, $\quad r_n \le \nu e^{-\frac{1}{d}W\left(\frac{d\log(1/\rho)}{C}\left\lfloor\frac{n}{\overline{\log} n}\right\rfloor\right)}$.

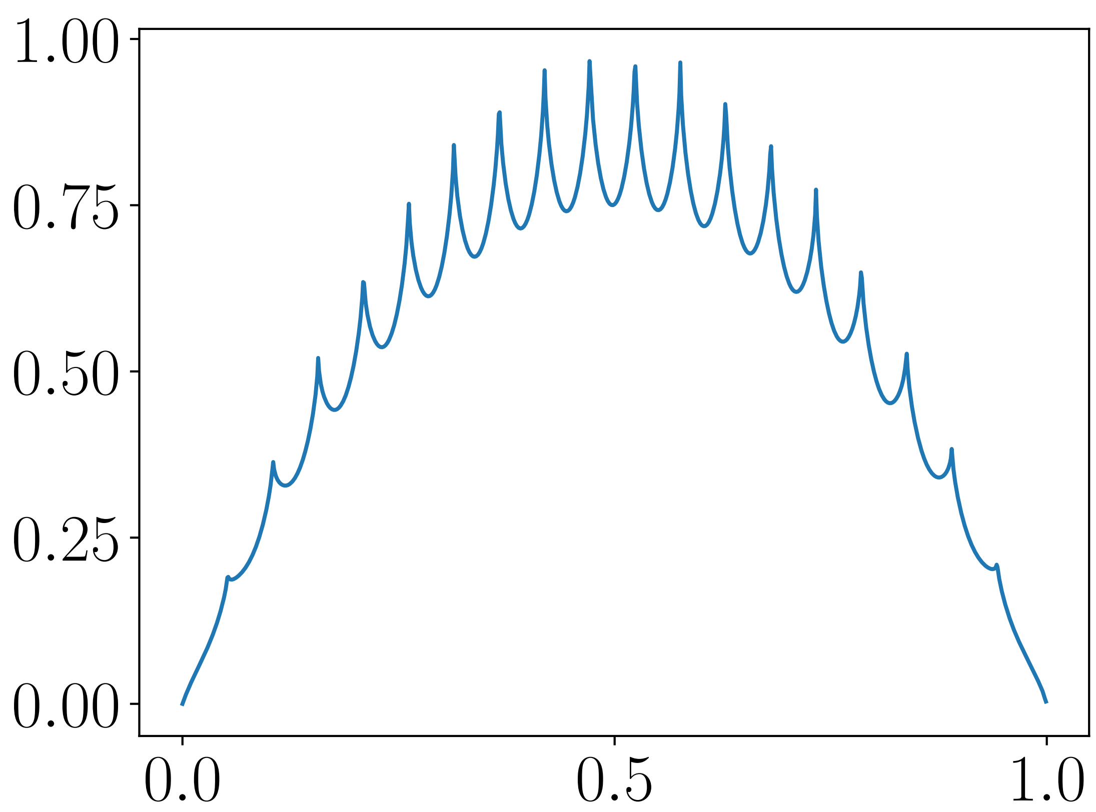
For more readability, Corollary 4 uses a lower bound on W (Hoorfar and Hassani, 2008).

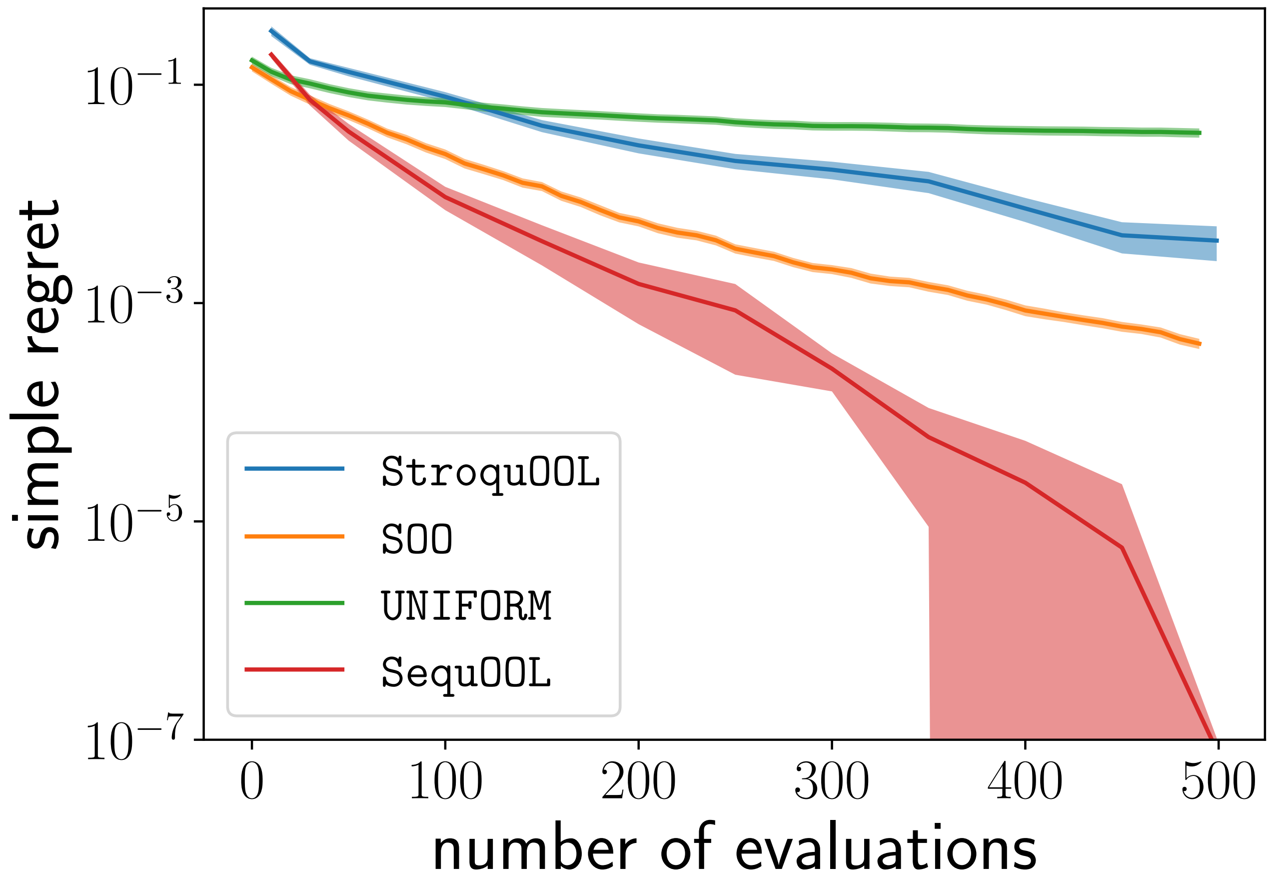
Corollary 4 *If $d>0$, assumptions in Theorem 3 hold and $\lfloor n/\overline{\log}\, n\rfloor\, d\log\frac{1}{\rho}/C > e$,*

$$r_n \le \nu\left(C/(d\log(1/\rho))\right)^{\frac{1}{d}}\left(\log\left(nd\log(1/\rho)/C\right)\right)^{\frac{1}{d}}\left\lfloor n/\overline{\log}\, n\right\rfloor^{-\frac{1}{d}}.$$

SUMMARY: d=0 → r < ρ^n AND d>0 → r < $n^{1/d}$

SEQUOOL: TRULY EXPONENTIAL RATE





STROQUOOL

Parameters: n, $\mathcal{P} = \{\mathcal{P}_{h,i}\}$

Init: Open $\mathcal{P}_{0,1}$ $h_{\max}$ times.

$$h_{\max} = \left\lfloor \frac{n}{2(\log_2 n + 1)^2} \right\rfloor, \; p_{\max} = \lfloor \log_2 (h_{\max}) \rfloor.$$

For $h = 1$ to $h_{\max}$ $\blacktriangleleft$ *Exploration* $\blacktriangleright$

For $p = \lfloor \log_2(h_{\max}/h) \rfloor$ down to 0

Open 2^p times the $\left\lfloor \frac{h_{\max}}{h2^p} \right\rfloor$ non-opened cells $\mathcal{P}_{h,i}$ with highest values $\widehat{f}_{h,i}$ and given that $T_{h,i} \geq 2^p$.

For $p \in [0 : p_{\max}]$ $\blacktriangleleft$ *Cross-validation* $\blacktriangleright$

Evaluate $h_{\max}$ times the *candidates:*

$$x(n,p) = \operatorname*{arg\,max}_{(h,i) \in \mathcal{T}, T_{h,i} \geq 2^p} \widehat{f}_{h,i}.$$

Output $x(n) = \operatorname*{arg\,max}_{\{x(n,p), p \in [0:p_{\max}]\}} \widehat{f}(x(n,p))$

Figure 2: The StroquOOL Algorithm

StroquOOL(m= 2^p) tradeoffs
- small m: quality estimates
- big m: we can go deeper

opening more promising cells more often

picking up the best point, with n samples

OBSERVATION: The deeper we go, the better optimum we find.

Lemma 5 *For any global optimum $x^\star$ with associated (ν,ρ) (see Assumption 1), with probability at least $1-\delta$, for all depths $h \in \left[\left\lfloor \frac{h_{\max}}{2^p}\right\rfloor\right]$, for all $p \in [0 : \lfloor \log_2(h_{\max}/h)\rfloor]$, if $b\sqrt{\frac{\log(4n/\delta)}{2^{p+1}}} \le \nu\rho^h$ and if $\frac{h_{\max}}{h2^p} \ge C\rho^{-d(\nu,\rho)h}$, we have $\perp_{h,p} = h$ while $\perp_{0,p} = 0$.*

SUMMARY: We go deep enough

MAIN RESULT

Theorem 6 *High-noise regime After n rounds, for any function f and one of its global optima $x^\star$ with associated (ν,ρ), and near-optimality dimension denoted for simplicity $d = d(\nu,\rho)$, if $b \ge \nu\rho^{\widetilde{h}}/\sqrt{\log(n^{3/2}/b)}$, the simple regret of* `StroquOOL` obeys

$$r_n \;\le\; \nu\rho^{\frac{1}{(d+2)\log(1/\rho)}W\left(\left\lfloor\frac{n}{2(\log_2 n+1)^2}\right\rfloor\frac{(d+2)\log(1/\rho)\nu^2}{Cb^2\log(n^{3/2}/b)}\right)} + 6b\sqrt{\log(n^{3/2}/b)\Big/\left\lfloor\frac{n}{2(\log_2 n+1)^2}\right\rfloor},$$

SUMMARY: r < n$^{1/(d+2)}$ ....before it was r < n$^{1/d}$

STROQUOOL: ADAPTATION TO NOISE

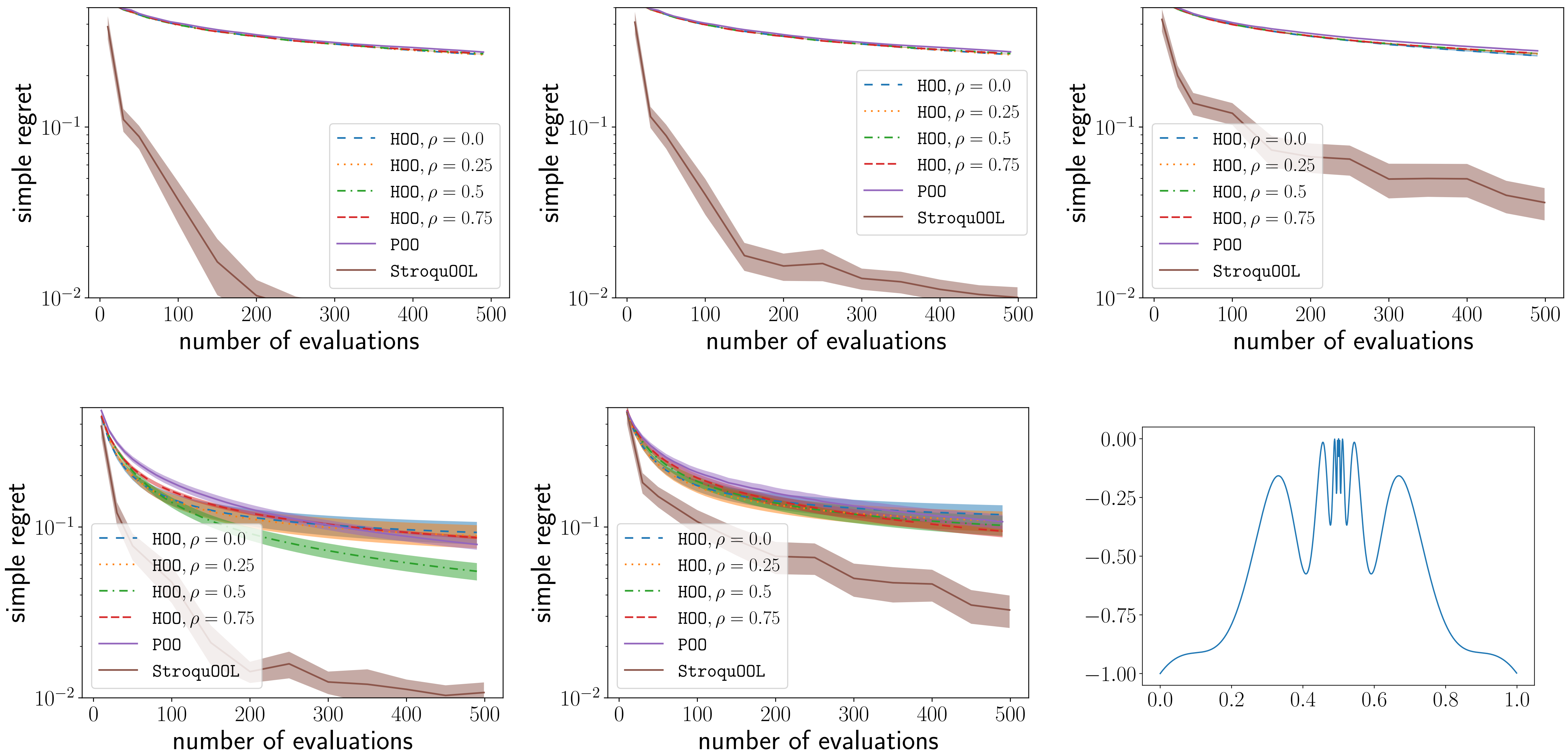
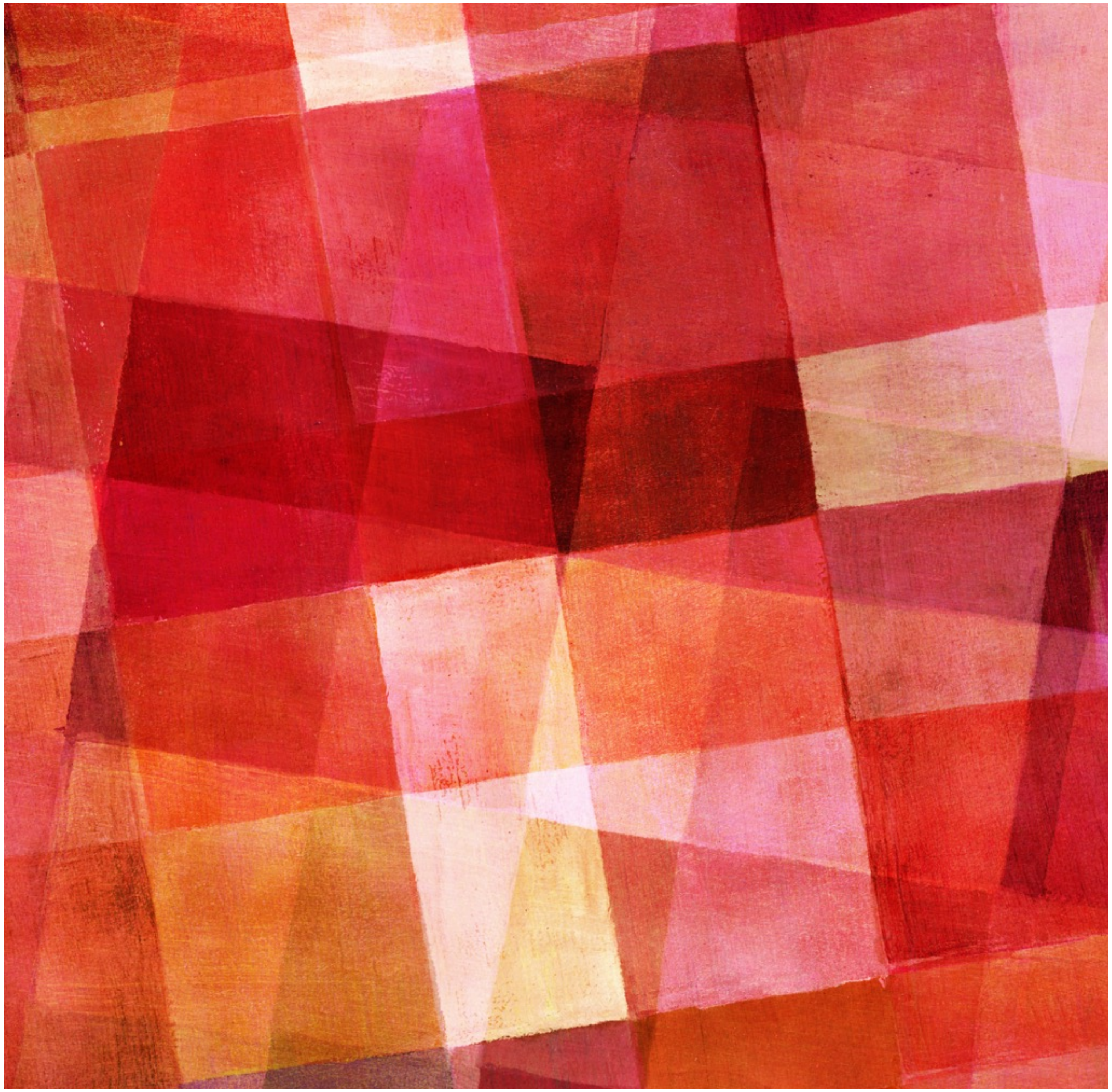
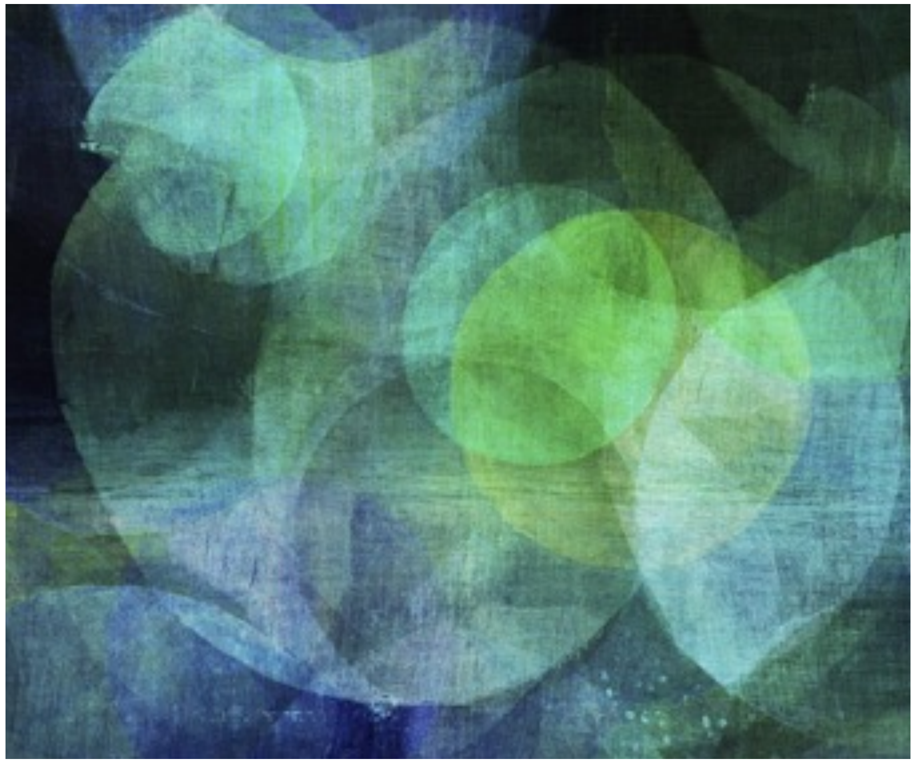


Figure 3: *Bottom right:* **Wrapped-sine** function ($d > 0$). The true range of the noise b and the range used by HOO and POO is $\widetilde{b}$. *Top:* $b = 0, \widetilde{b} = 1$ left — $b = 0.1, \widetilde{b} = 1$ middle — $b = \widetilde{b} = 1$ right. *Bottom:* $b = \widetilde{b} = 0.1$ left — $b = 1, \widetilde{b} = 0.1$ middle.

DISCUSSION AND WHAT'S NEXT

- we sample ~1/h = Zipf law (ex.: the frequency of any word is inversely proportional to its rank in the frequency table)
- **Adaption to smoothness**
- **Adaptation to noise** (no need to provide it as input)
  - even to the noise = 0 — deterministic
  - deterministic case, **exp(-n)** for d=0, first exponential rate
    - before only possible with very strong assumptions
- **Not a panacea:** price to pay for minimal assumptions and global guarantee
  - hyper-parameter optimization
- adversarial/stochastic (COLT 2018)
- **NEXT:** make MCTS for faster by adapting it to noise





Michal Valko, SequeL, Inria Lille - Nord Europe, michal.valko@inria.fr
http://researchers.lille.inria.fr/~valko/hp/

APPENDIX: PROOF SKETCH

$$f(x(n)) \overset{\textbf{(a)}}{\geq} f_{\perp_{h_{\max}}+1,i^\star} \overset{\textbf{(b)}}{\geq} f(x^\star) - \nu\rho^{\perp_{h_{\max}}+1}$$

DEFINITION: how deep we can go

$$\frac{h_{\max}}{\overline{h}} = C\rho^{-d\overline{h}}$$

PROPERTY: if we go that deep, we are near-optimal

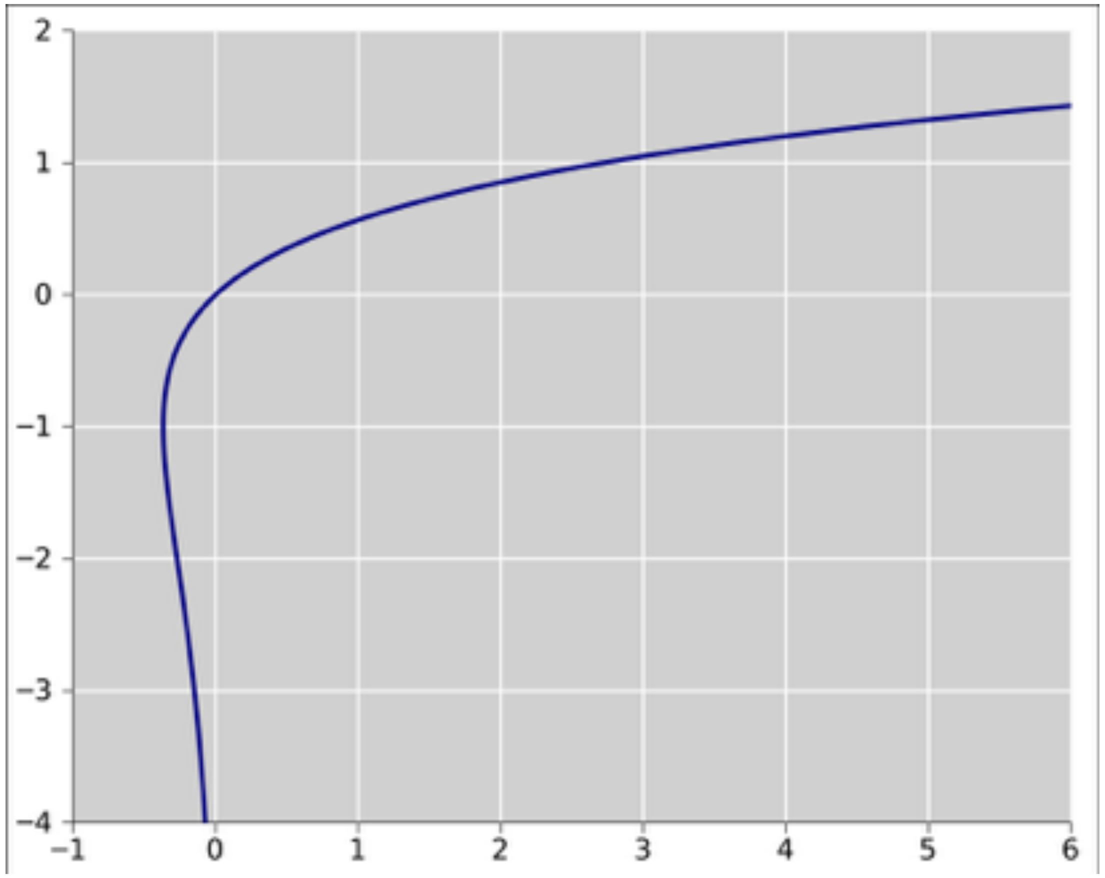
$$\frac{h_{\max}}{\lfloor \overline{h} \rfloor} \geq \frac{h_{\max}}{\overline{h}} = C\rho^{-d\overline{h}} \geq C\rho^{-d\lfloor \overline{h} \rfloor}$$

SUMMARY: We go deep enough

$$\frac{r_n}{\nu} \leq \rho^{\frac{1}{d_\rho}\left(\log\left(\frac{h_{\max} d_\rho / C}{\log(h_{\max} d_\rho / C)}\right)\right)} = e^{\frac{1}{d\log(1/\rho)}\left(\log\left(\frac{h_{\max} d_\rho / C}{\log\left(\frac{h_{\max} d_\rho}{C}\right)}\right)\right)\log(\rho)} = \left(\frac{h_{\max} d_\rho / C}{\log\left(\frac{h_{\max} d_\rho}{C}\right)}\right)^{-\frac{1}{d}}$$

APPENDIX: LAMBERT W FUNCTION

$$z = f^{-1}\left(ze^z\right) = W\left(ze^z\right)$$



The Lambert W function Our results use the **Lambert W function**. Solving for the variable z, the equation $A = ze^z$ gives $z = W(A)$. W is multivalued for $z \leq 0$. However, in this paper, we consider $z \geq 0$ and $W(z) \geq 0$, referred to as the *standard* W. W cannot be expressed in terms of elementary functions. Yet, we have $W(z) = \log(z/\log z) + o(1)$ (Hoorfar and Hassani, 2008). W has applications in physics and applied mathematics (Corless et al., 1996).