Optimization with Gaussian Processes via Chaining

Emile Contal — CMLA, ENS Cachan

Joint work with Cédric Malherbe and Nicolas Vayatis

Problem Statement

Sequential Optimization

Let $f : \mathcal{X} \to \mathbb{R}$ where the input space $\mathcal{X}$ is not necessarily a subset of $\mathbb{R}^D$. We consider the problem of finding the maximum of f denoted by:

$$f^\star = \sup_{x \in \mathcal{X}} f(x)\,,$$

via sequential queries $f(x_1), f(x_2), \ldots$

Noisy Observations

At iteration T we choose x_{T+1} using the previous noisy observations $Y_T = \{y_1, \ldots, y_T\}$, where $\forall t \leq T$:

$$y_t = f(x_t) + \epsilon_t \;\text{ and }\; \epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \eta^2)\,.$$

Gaussian Processes

Definition

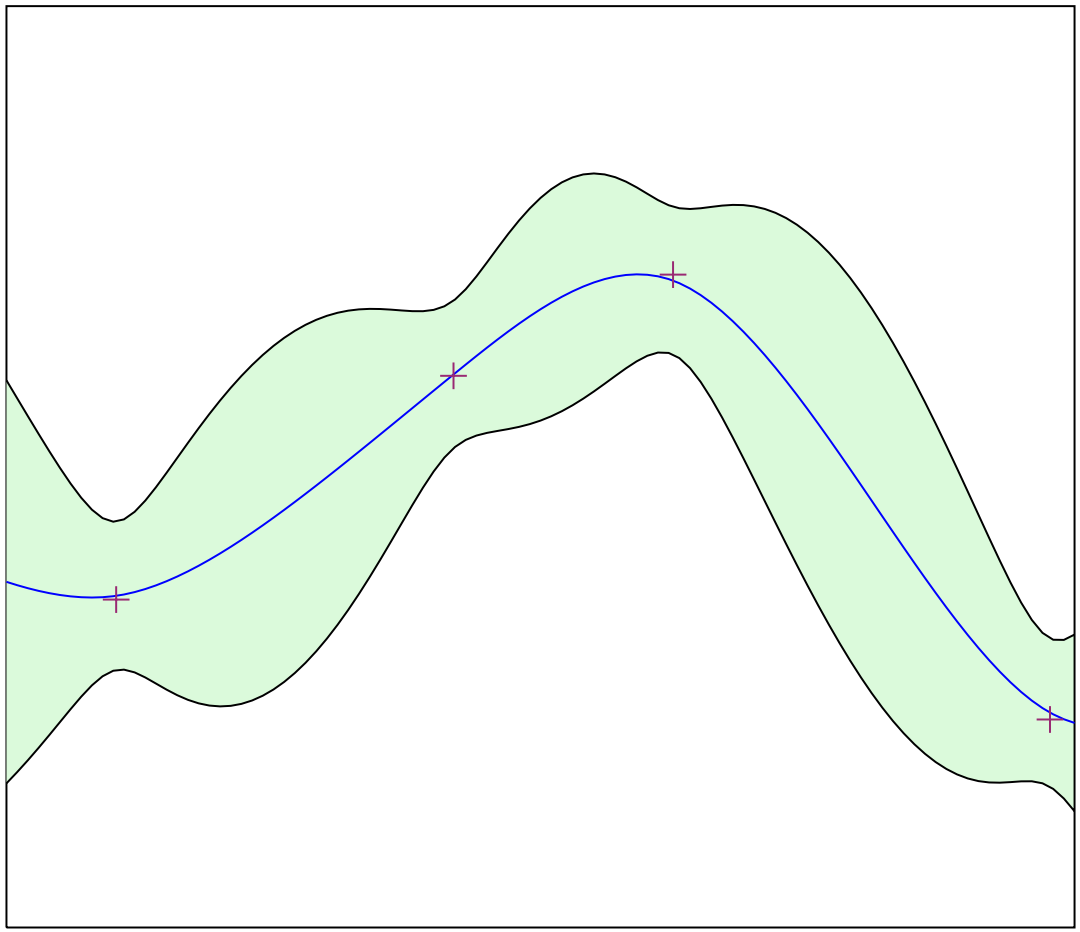
$f \sim \mathcal{GP}(m, k)$ with mean function $m : \mathcal{X} \to \mathbb{R}$ and kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$, when for all $x_1, \ldots, x_n \in \mathcal{X}$ we have:

$$\Big(f(x_1), \ldots, f(x_n)\Big) \sim \mathcal{N}\Big(\big[m(x_i)\big]_i, \big[k(x_i, x_j)\big]_{i,j}\Big) .$$

Bayesian Inference

Given Y_T at X_T, the posterior distribution $\Pr[f \mid X_T, Y_T]$ is a GP with mean μ_{T+1} (prediction) and variance σ^2_{T+1} (uncertainty) computed by Bayesian inference.

Gaussian Processes Confidence Bounds



Objective

Regret

The efficiency of a policy is measured via the simple or cumulative regret:

$$S_T = \min_{t \leq T} \left\{ f^\star - f(x_t) \right\} \text{ and } R_T = \sum_{t \leq T} \left(f^\star - f(x_t) \right).$$

Goal

The regrets are unknown in practice. Our aim is to obtain upper bounds on S_T an R_T with high probability.

Strategy: Upper Confidence Bounds (UCB)

If we have $f^\star - f(x_t) \leq UCB_t(x_t)$, then we can control the regret.

Related Work

Bayesian Optimization

- Bull (2011): Expected Improvement Algorithm
- Hennig et al. (2012): Entropy Search Algorithm

Upper Confidence Bounds

- Freitas et al. (2012): Deterministic GP
- Srinivas et al. (2012): GP-UCB
- Djolonga et al. (2013): High-dimensional GP

Chaining

- Grunewalder et al. (2010): Known horizon bandits
- Gaillard and Gerchinovitz. (2015): Online regression

Canonical Pseudo-Distance

A First UCB

$$d_t^2(x^\star, x_t) = \mathbb{V}[f(x^\star) - f(x_t) \mid X_t, Y_t]$$
$$= \sigma_t^2(x^\star) - 2k_t(x^\star, x_t) + \sigma_t^2(x_t)\,.$$

For all $\delta > 0$, it exists β_δ such that with probability at least $1 - \delta$:

$$f(x^\star) - f(x_t) \leq \mu_t(x^\star) - \mu_t(x_t) + \beta_\delta d_t(x^\star, x_t)\,.$$

Union Bound over $\mathcal{X}$

With $|\mathcal{X}| < \infty$, we have with probability at least $1 - |\mathcal{X}|\,\delta$:

$$\sup_{x^\star \in \mathcal{X}} f(x^\star) - f(x_t) \leq \sup_{x^\star \in \mathcal{X}} \mu_t(x^\star) - \mu_t(x_t) + \beta_\delta d_t(x^\star, x_t)\,.$$

But what if $|\mathcal{X}| = \infty$?

Covering Numbers

ε-Cover

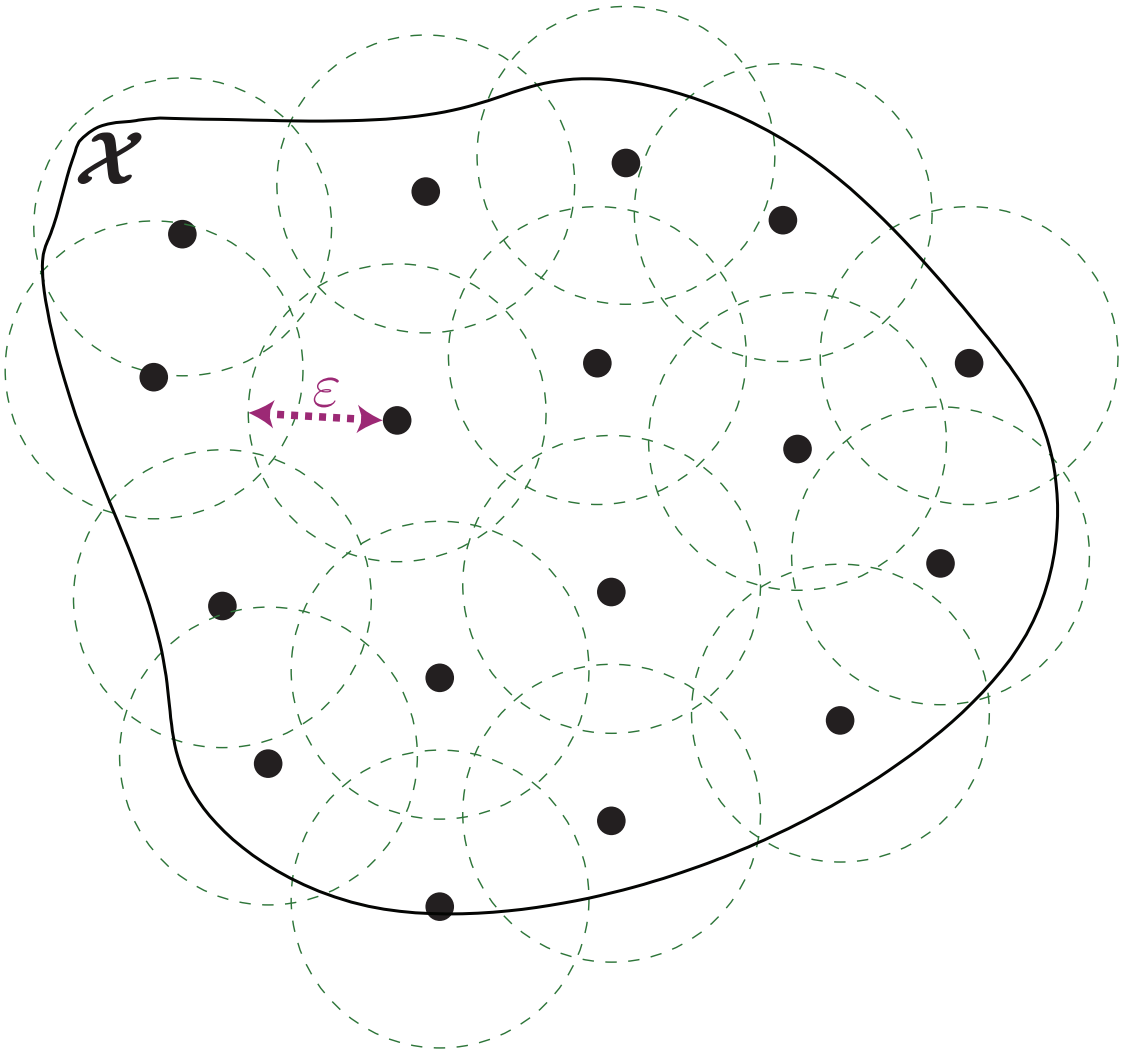
$T \subset \mathcal{X}$ is an ε-cover of $\mathcal{X}$ for d_t iff:

$$\forall x \in \mathcal{X},\ \exists x' \in T \text{ s.t. } d_t(x, x') \leq \varepsilon$$

Covering Numbers

The covering number $N(\mathcal{X}, d_t, \varepsilon)$ is the size of the smallest ε-cover.

An ε-Cover for the Euclidean Distance



Hierarchical Covers

Assumption

$$\forall x, y \in \mathcal{X},\ k(x, y) \leq 1\,.$$

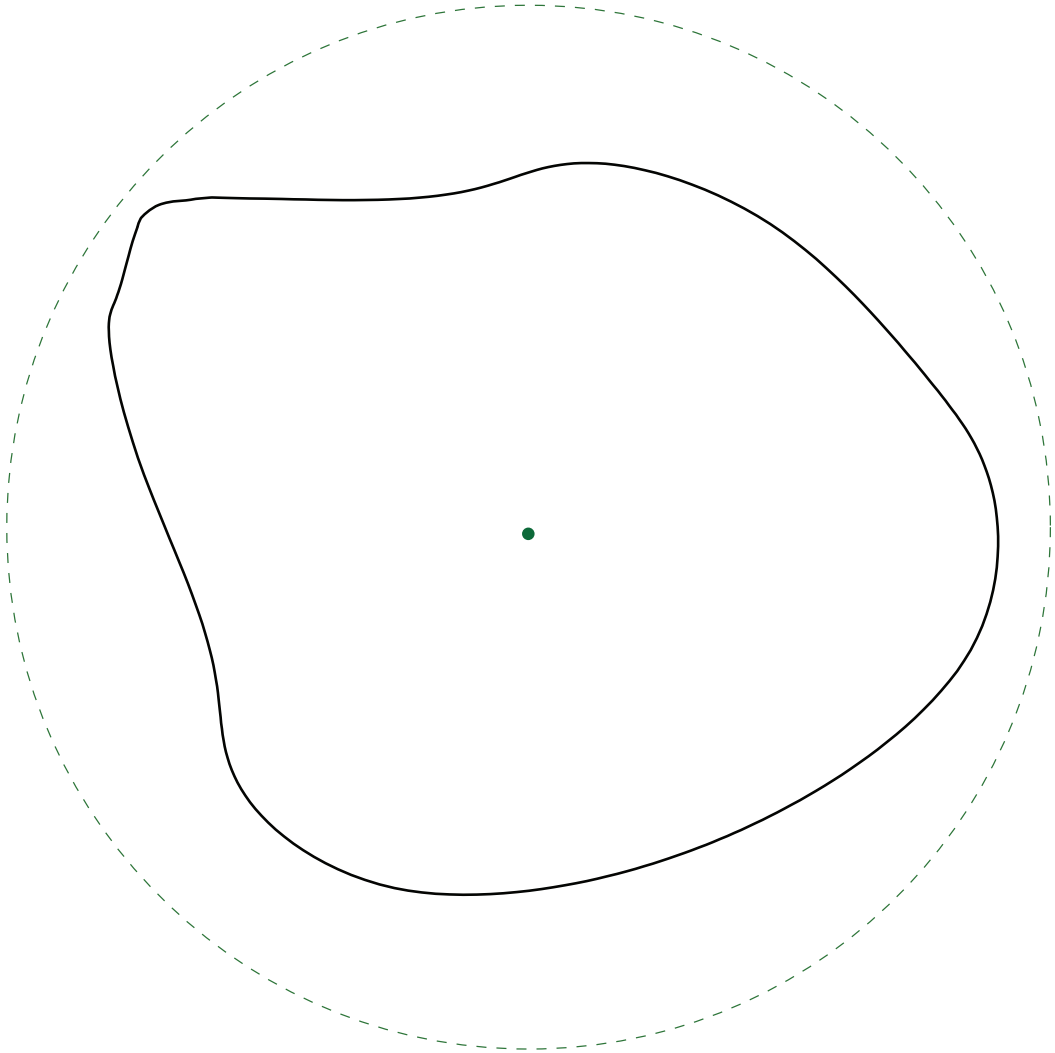
Since $d_t(x, y) \leq k(x, y)$, any point of $\mathcal{X}$ is a 1-cover of $\mathcal{X}$ for d_t.

Hierarchical Covers

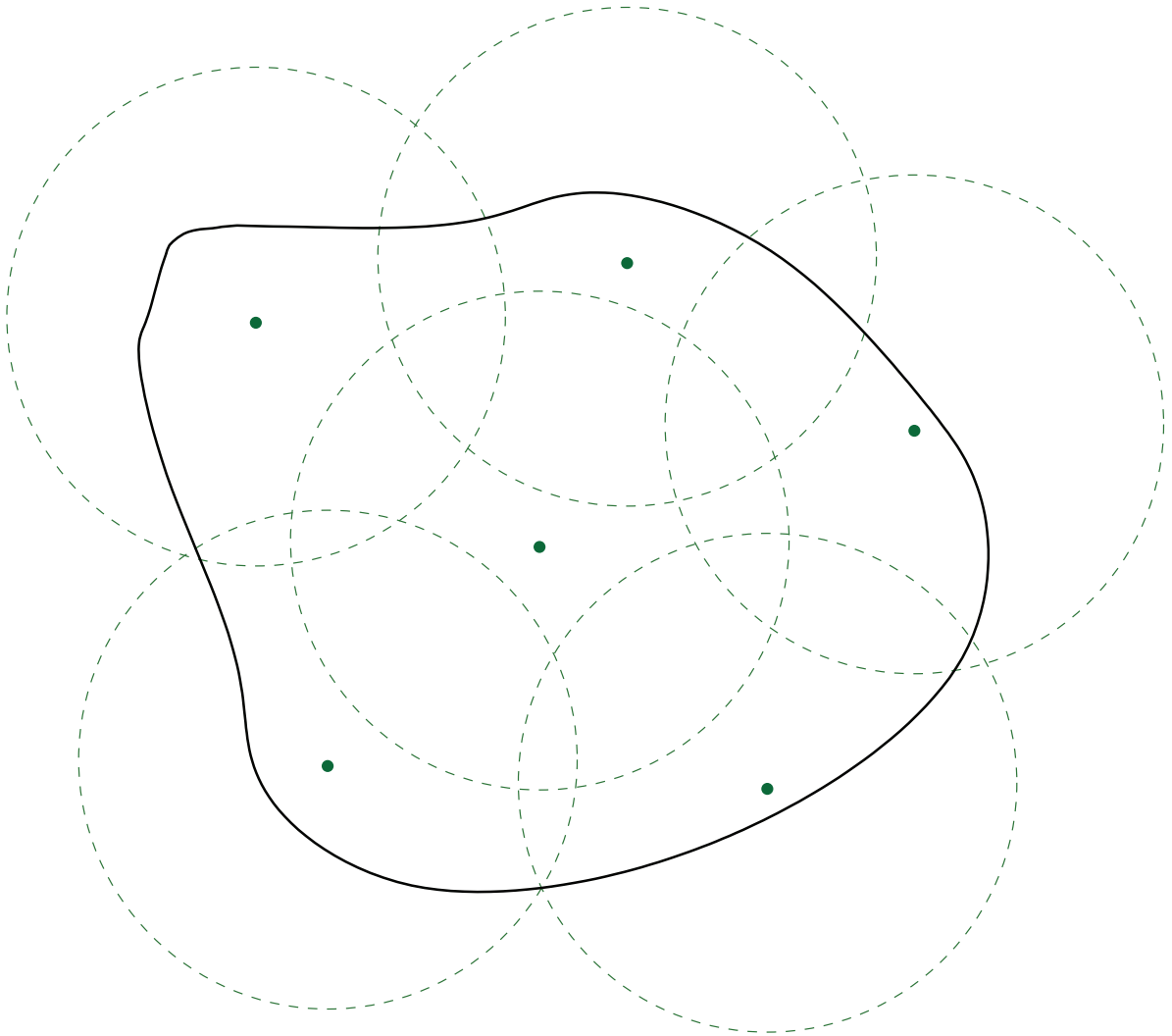
Let $\mathcal{T} = (T_i)_{i \geq 0}$ such that for all $i \geq 0$:

- T_i is an ε_i-cover with $\varepsilon_i = 2^{-i}$.
- $T_i \subset T_{i+1}$

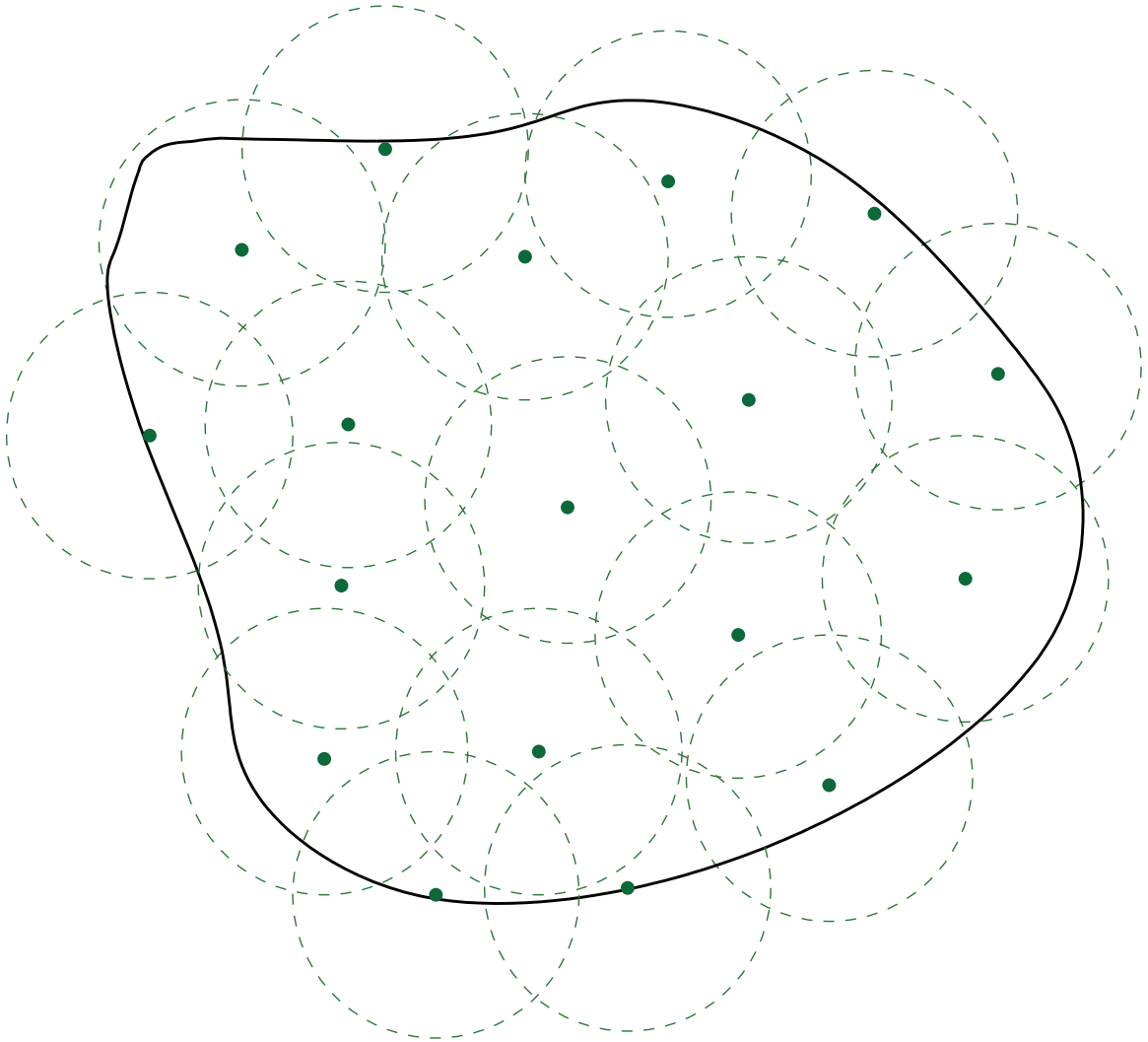
Hierarchical Covers: $\epsilon_0 = 1$



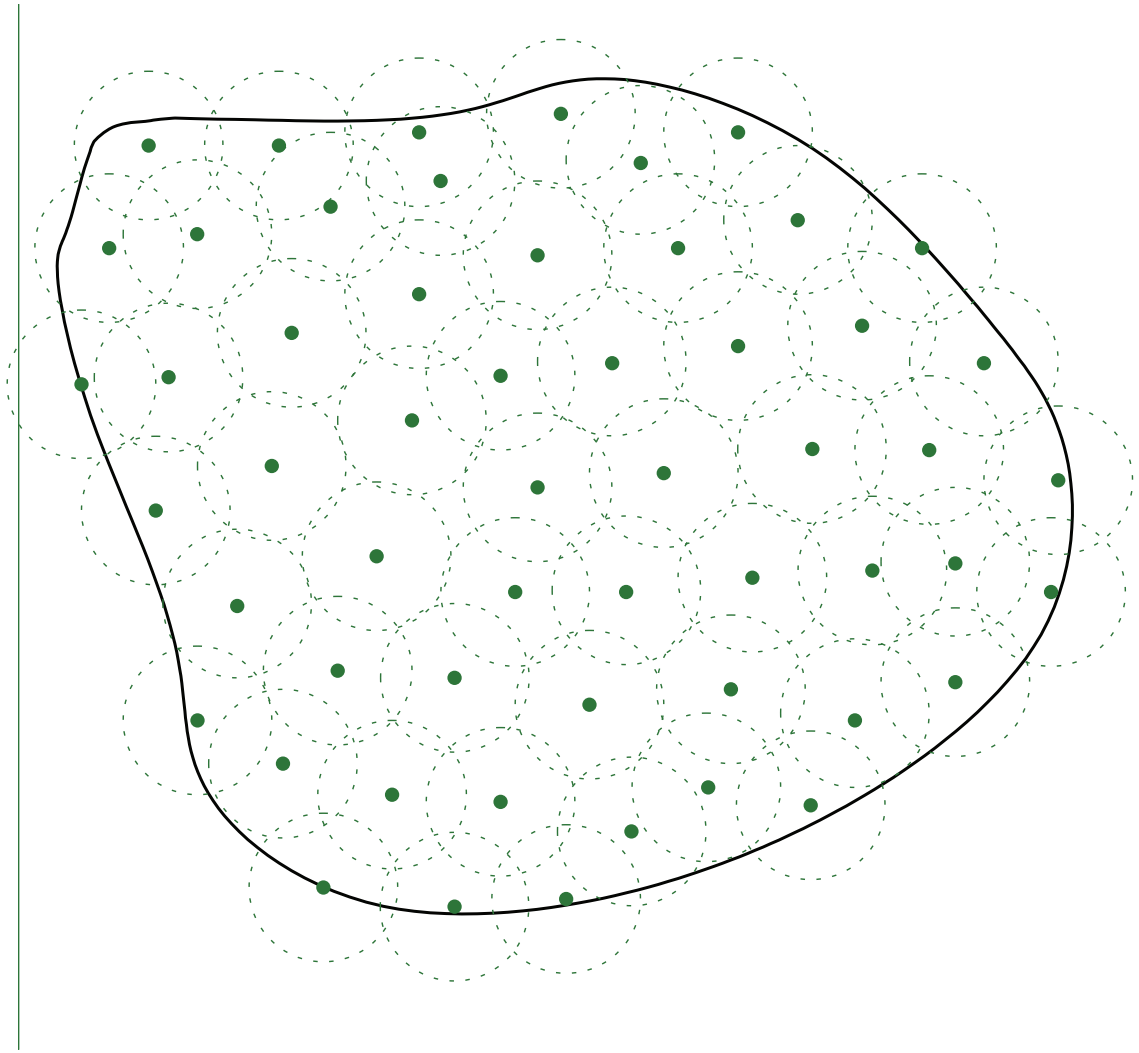
Hierarchical Covers: $\epsilon_1 = \frac{1}{2}$



Hierarchical Covers: $\epsilon_2 = \frac{1}{4}$



Hierarchical Covers: $\epsilon_3 = \frac{1}{8}$



Localized Chaining

Projection to $\mathcal{T}$ starting from x_t toward $x^\star$

Define $\pi_i(x^\star) = \text{argmin}_{x \in T_i \cup \{x_t\}} d_t(x^\star, x)$, then:

- $\pi_i(x^\star) \underset{i \to \infty}{\to} x^\star$
- $\pi_i(x^\star) = x_t$ if $d_t(x^\star, x_t) < \epsilon_i$

The Chaining Trick

$$f(x^\star) - f(x_t) = \sum_{i : \epsilon_i < d(x^\star, x_t)} f(\pi_i(x^\star)) - f(\pi_{i-1}(x^\star)) \,.$$

Localized Chaining

Projection to $\mathcal{T}$ starting from x_t toward $x^\star$

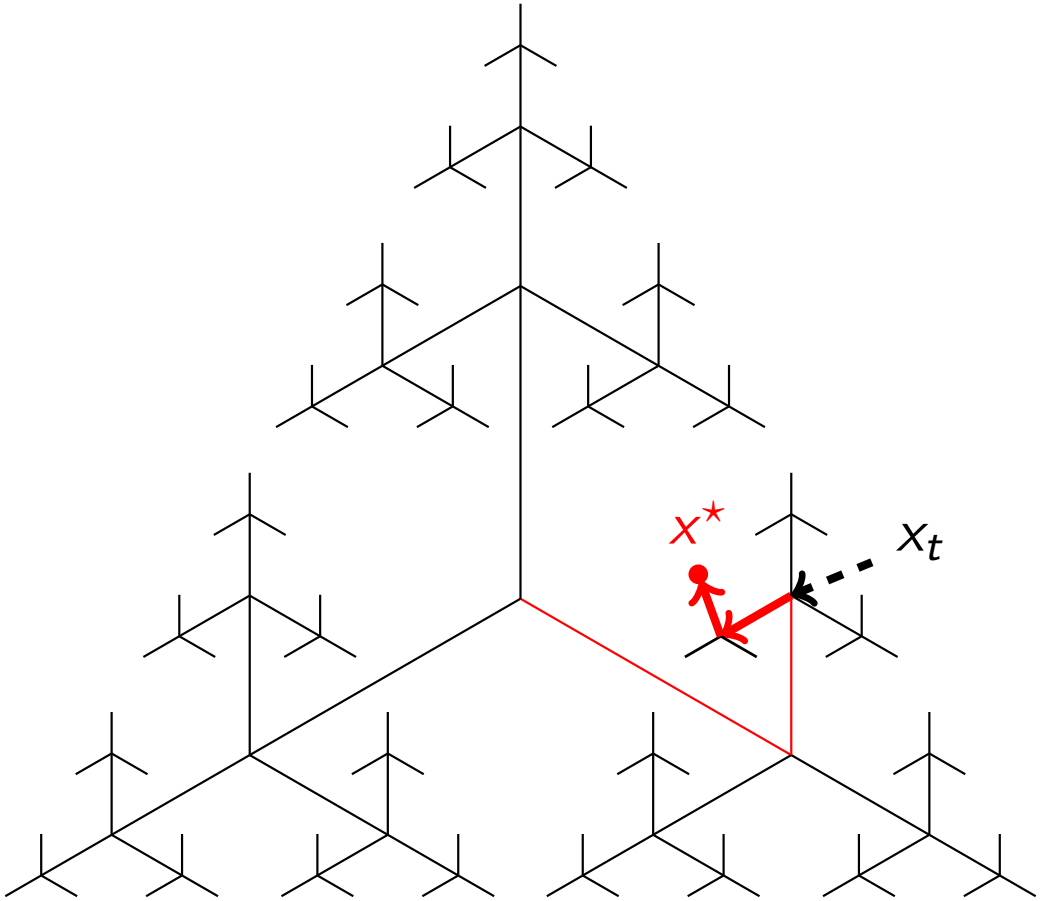
Define $\pi_i(x^\star) = \text{argmin}_{x \in T_i \cup \{x_t\}} d_t(x^\star, x)$, then:

- $\pi_i(x^\star) \underset{i \to \infty}{\to} x^\star$
- $\pi_i(x^\star) = x_t$ if $d_t(x^\star, x_t) < \epsilon_i$

The Chaining Trick

$$\sup_{x^\star \in \mathcal{X}} f(x^\star) - f(x_t) = \sup_{x^\star \in \mathcal{X}} \sum_{i : \epsilon_i < d(x^\star, x_t)} f(\pi_i(x^\star)) - f(\pi_{i-1}(x^\star)) \,.$$

Localized Chaining Illustrated



Upper Confidence Bound

Converging distances

$$d_t\big(\pi_i(x^\star), \pi_{i-1}(x^\star)\big) \leq \varepsilon_{i-1}$$

UCB at depth i: Union Bound on T_i

For any $i \geq 1$, with probability at least $1 - |T_i|\,\delta$:

$$\sup_{x^\star \in \mathcal{X}} f(\pi_i(x^\star)) - f(\pi_{i-1}(x^\star)) \leq \sup_{x^\star \in \mathcal{X}} \mu_t(\pi_i(x^\star)) - \mu_t(\pi_{i-1}(x^\star)) + \beta_\delta \varepsilon_{i-1}$$

Final UCB

Therefore we can choose $\beta_{\delta,i}$ such that with probability at least $1 - \delta$:

$$\sup_{x^\star \in \mathcal{X}} f(x^\star) - f(x_t) \leq \sup_{x^\star \in \mathcal{X}} \mu_t(x^\star) - \mu_t(x_t) + \sum_{i : \varepsilon_i < d_t(x^\star, x_t)} \varepsilon_i \beta_{\delta,i}\,.$$

The Chaining-UCB Algorithm

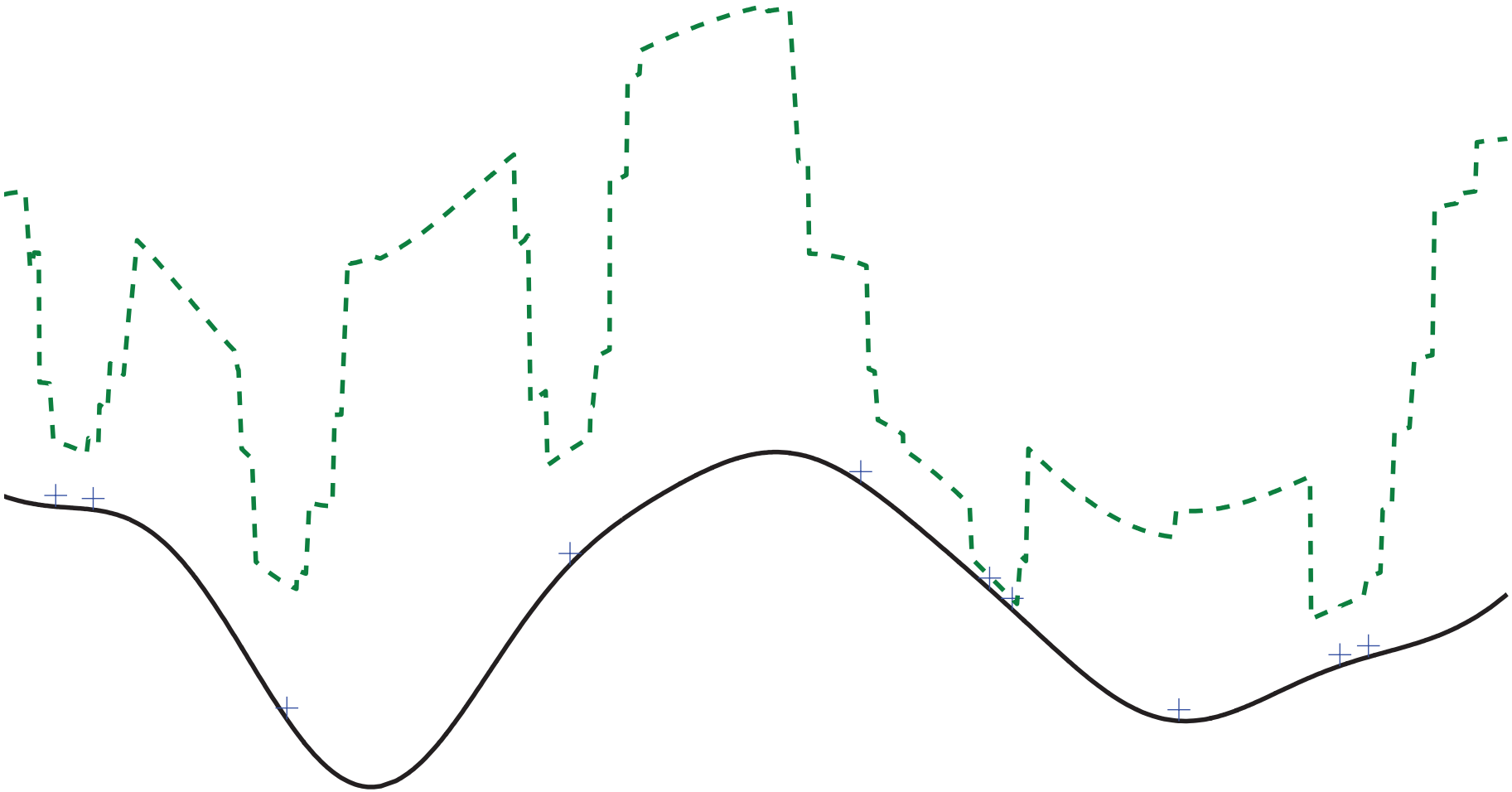
UCB policy

$$x_{t+1} \leftarrow \operatorname*{argmax}_{x \in \mathcal{X}} \mu_t(x) + \sum_{i:\varepsilon_i < \sigma_t(x)} \varepsilon_i \beta_{\delta,i} \,.$$

Practical Remark

The algorithm needs only to compute the $\beta_{\delta,i}$ for i where $\varepsilon_i > \min_{x \in \mathcal{X}} \sigma_t(x)$.

Upper Confidence Bound



Greedy Cover

NP-hardness

Computing the $\beta_{\delta,i}$ requires to compute the ε_i-cover. Finding the smallest ε_i is NP-hard.

Greedy optimal approximation

$T \leftarrow \emptyset$
$\bar{\mathcal{X}} \leftarrow \mathcal{X}$
while $\bar{\mathcal{X}} \neq \emptyset$ **do**
  $x \leftarrow \text{argmax}_{x \in \bar{\mathcal{X}}} \sum_{x' \in \bar{\mathcal{X}}} \mathbb{1}_{d(x,x') \leq \varepsilon}$
  $T \leftarrow T \cup \{x\}$
  $\bar{\mathcal{X}} \leftarrow \bar{\mathcal{X}} \setminus \{x' \in \bar{\mathcal{X}} : d(x, x') \leq \varepsilon\}$
end
return T

Theorem: Generic bounds for the Chaining-UCB algorithm

For $\delta \in (0,1)$, denoting $\sigma_t = \sigma_t(x_t)$, we have that for all $t \geq 1$, $\exists c_\delta \in \mathbb{R}$:

$$\sup_{x^\star \in \mathcal{X}} f(x^\star) - f(x_t) \leq \sigma_t\big(c_\delta - 6\log\sigma_t\big) + 9\sum_{i:\varepsilon_i < \sigma_t} \varepsilon_i \sqrt{\log N(\mathcal{X}, d_t, \varepsilon_i)}\,,$$

with probability at least $1-\delta$.

Corollary: Squared Exponential Kernel

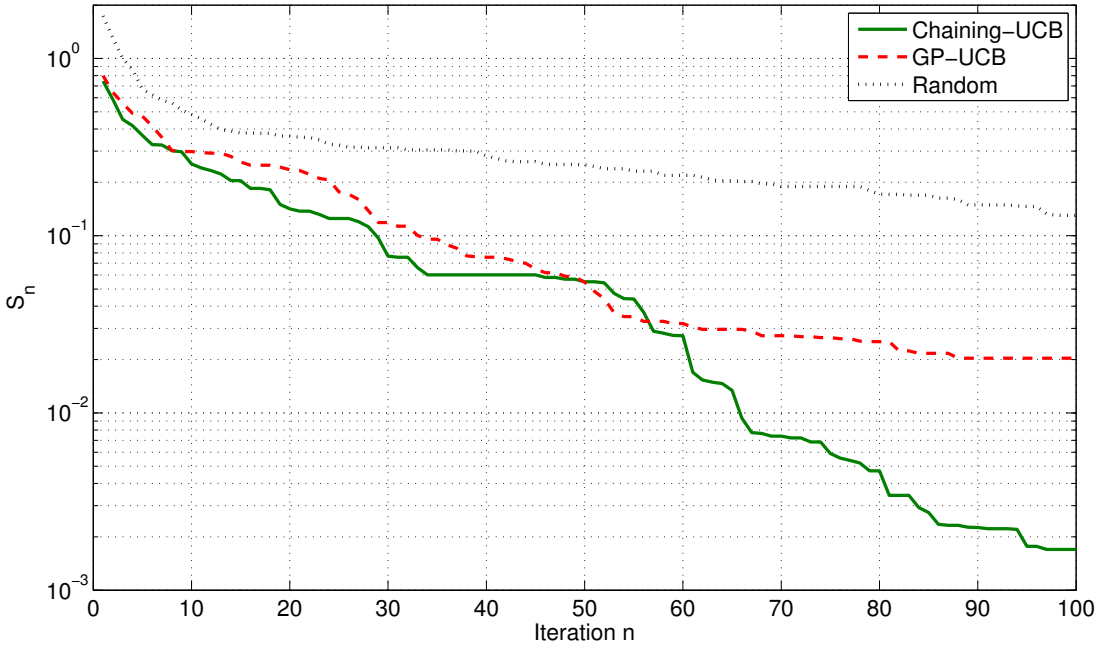
For the SE kernel $k(x, x') = e^{-\frac{1}{2}\|x-x'\|_2^2}$ and a compact $\mathcal{X} \subseteq [0, R]^D$,

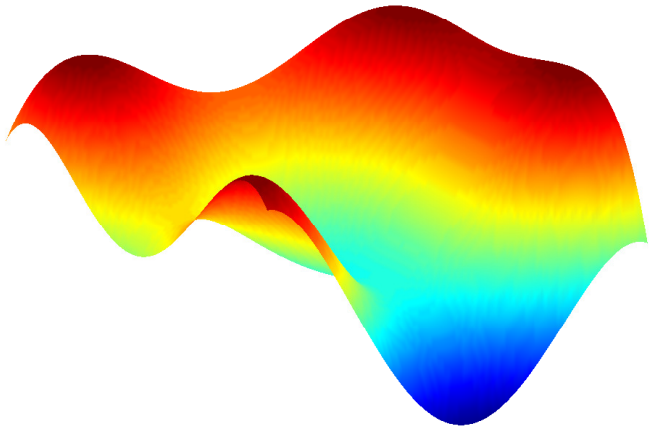
$$N(\mathcal{X}, d_t, \varepsilon) \leq \left(\frac{R}{\varepsilon}\right)^D.$$

Then the Chaining-UCB algorithm incurs regrets such that

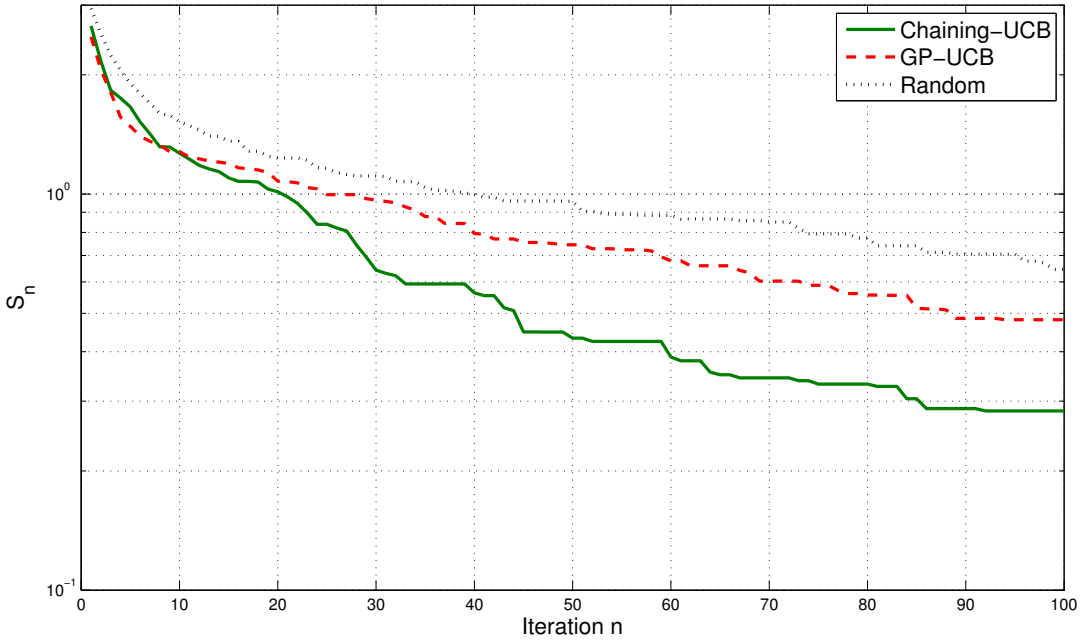
$$R_T = \mathcal{O}\left(\sqrt{T(\log T)^{D+2}}\right) \text{ and } S_T = \mathcal{O}\left(\sqrt{\frac{(\log T)^{D+2}}{T}}\right).$$

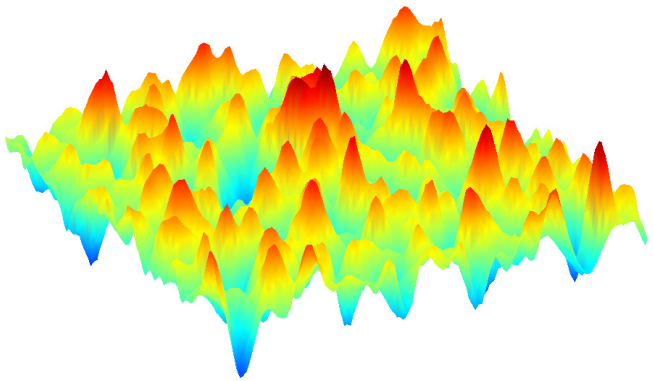
Experiments 1/4: Himmelblau's function



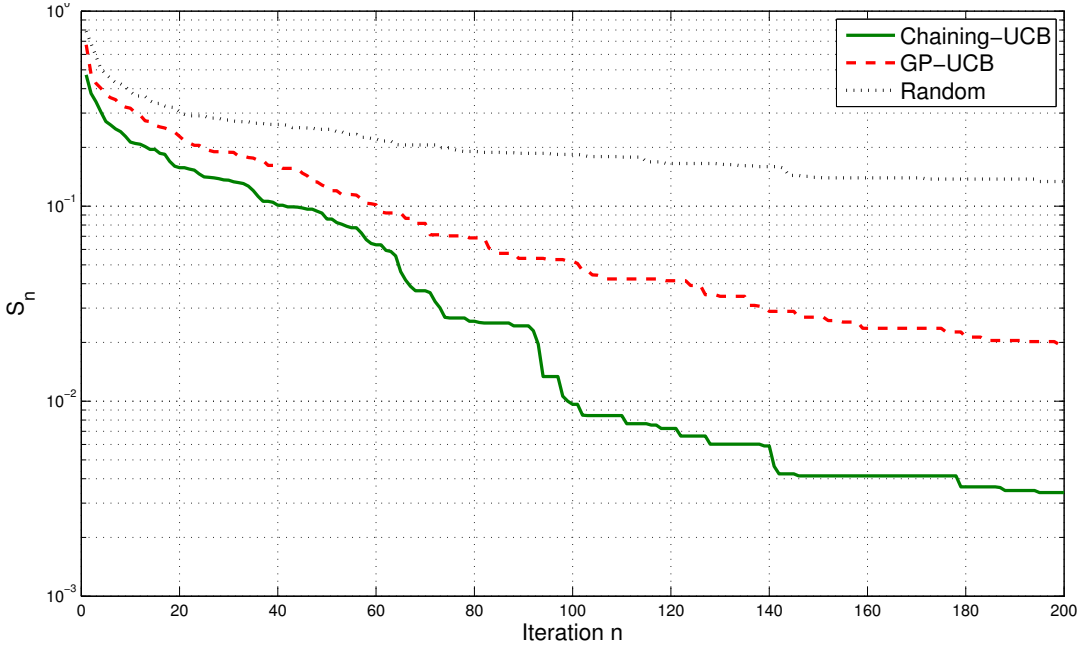


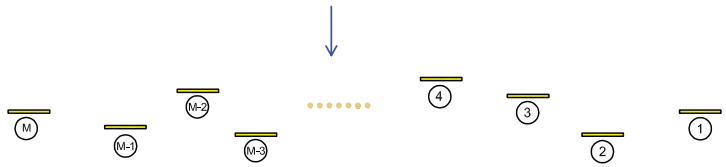
Experiments 2/4: SE kernel

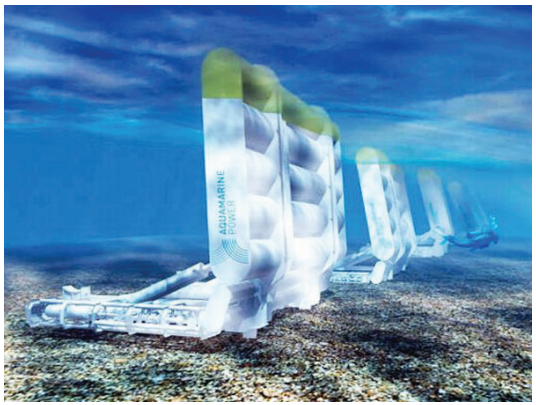




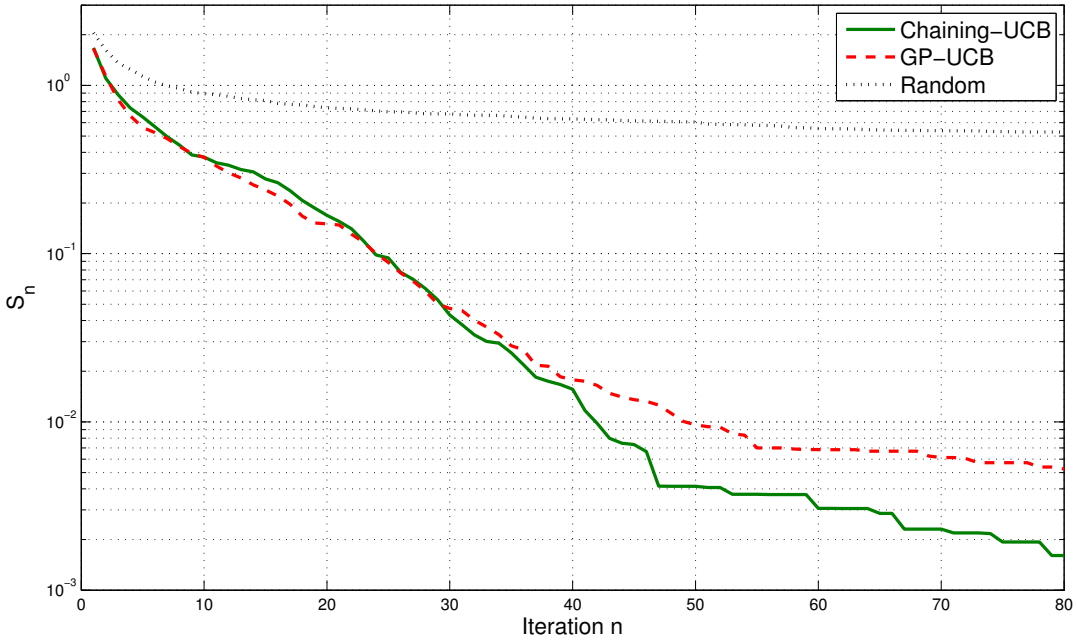
Experiments 3/4: Wave Energy Converter

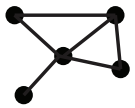


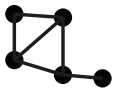


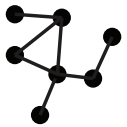


Experiments 4/4: Graphs kernel



$f($  $) = 1.3$

$f($  $) = 0.2$

$f($  $) = 2.7$

Conclusion

The Chaining-UCB Algorithm

- automatic calibration of the exploration/exploitation tradeoff
- adapts to various settings
- computationally tractable

Matlab code online!

`http://econtal.perso.math.cnrs.fr/software/`