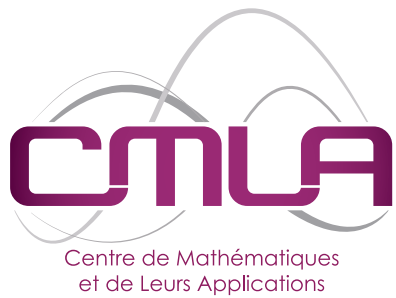
Optimisation Séquentielle par Processus Gaussiens

Emile Contal — CMLA, ENS Cachan



Formulation du Problème

Modèle

- $f : \mathcal{X} \to \mathbb{R}$, inconnue, non convexe, multimodale
- on cherche $f^\star = \sup_{x \in \mathcal{X}} f(x)$ via $f(x_1), f(x_2), \ldots$
- $\mathcal{X}$ peut être $\subset \mathbb{R}^D$ ou non paramétrique
- on observe $y_n = f(x_n) + \varepsilon$ où $\varepsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, \eta^2)$

Objectifs

- regret simple : $S_T = \min_{n \leq T} \left\{ f^\star - f(x_n) \right\}$
- regret cumulé : $R_T = \sum_{n \leq T} \left(f^\star - f(x_n) \right)$

Formulation du Problème

Modèle

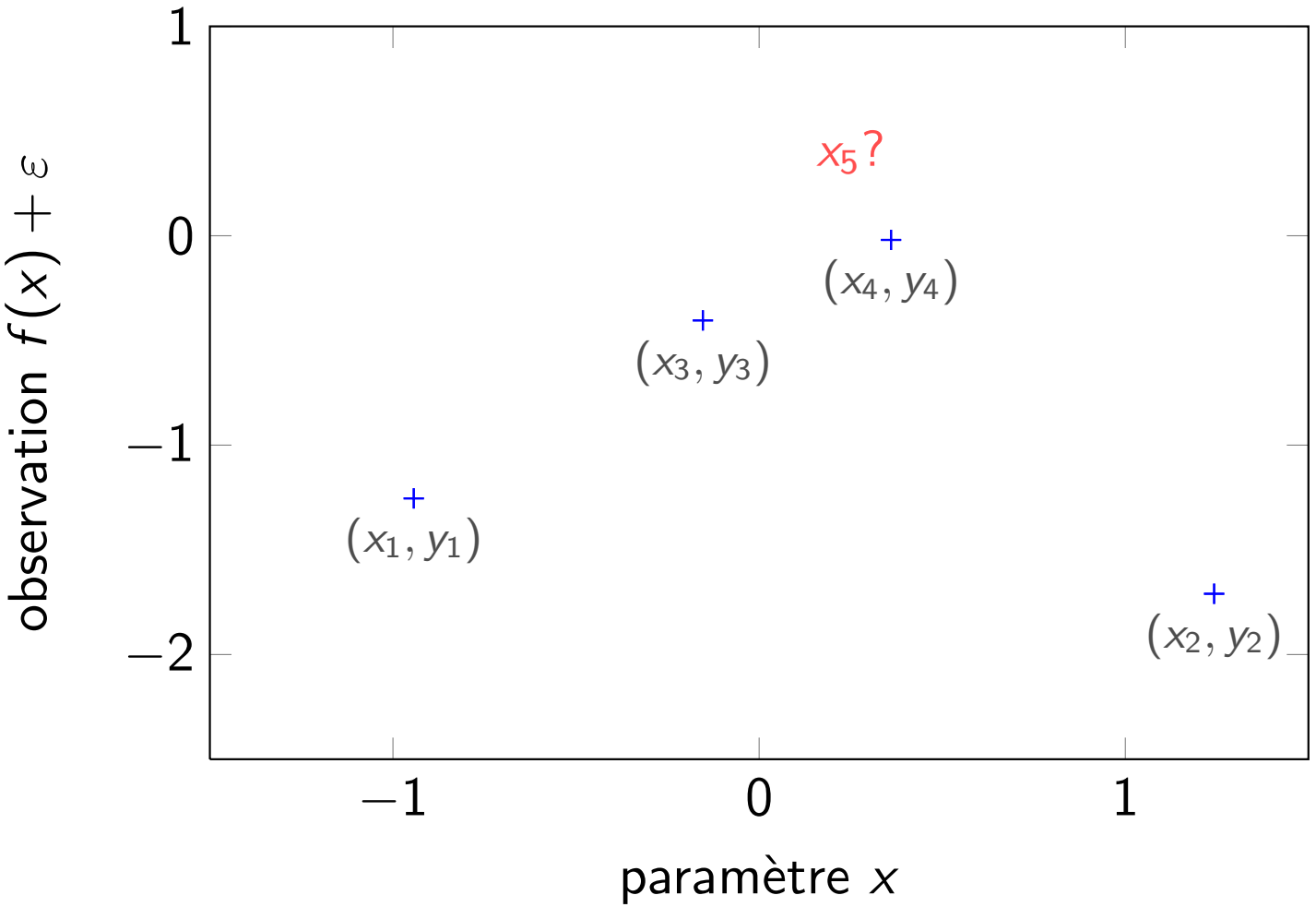
- $f : \mathcal{X} \to \mathbb{R}$, inconnue, non convexe, multimodale
- on cherche $f^\star = \sup_{x \in \mathcal{X}} f(x)$ via $f(x_1), f(x_2), \ldots$
- $\mathcal{X}$ peut être $\subset \mathbb{R}^D$ ou non paramétrique
- on observe $y_n = f(x_n) + \varepsilon$ où $\varepsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, \eta^2)$

Objectifs

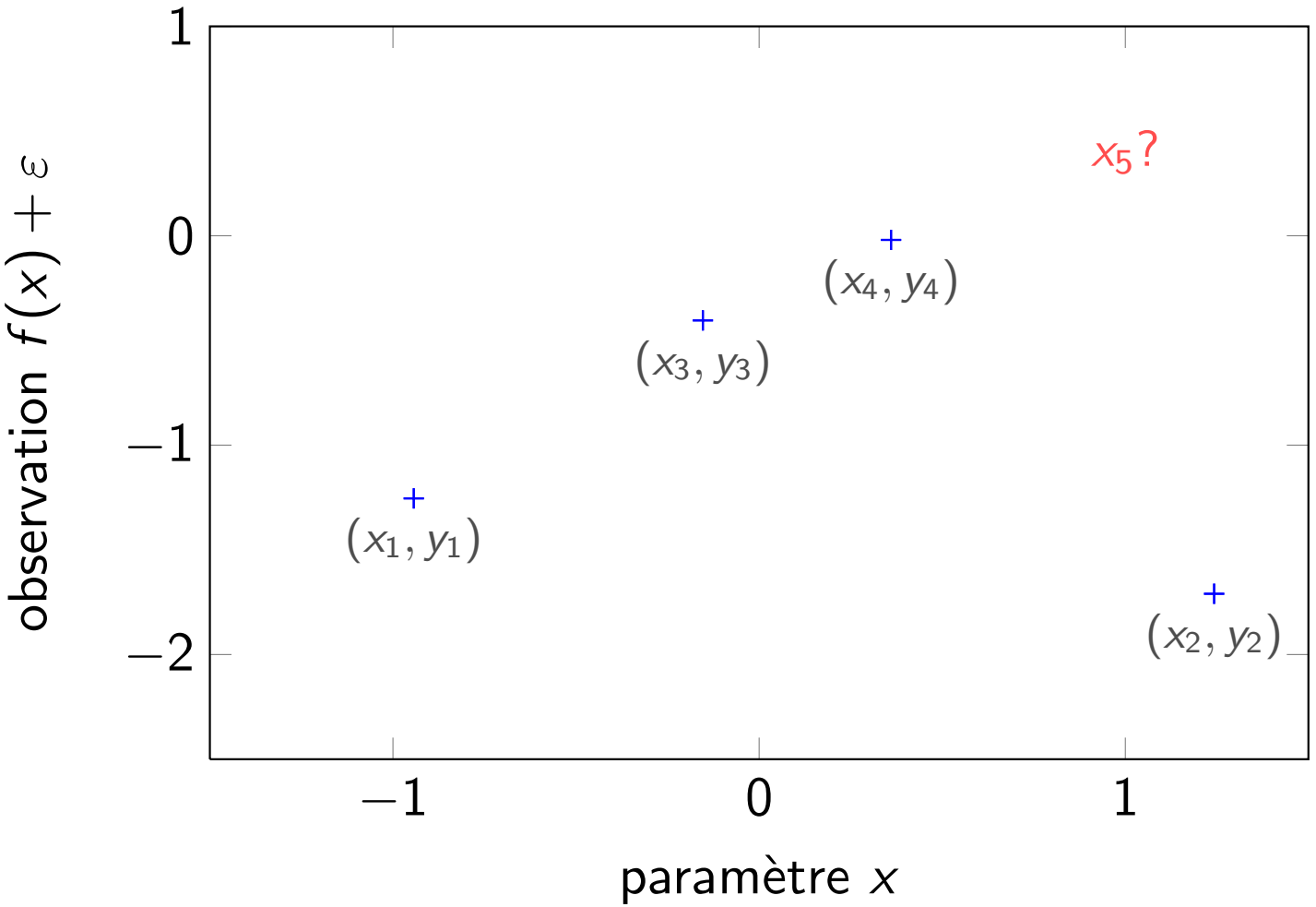
- regret simple : $S_T = \min_{n \leq T} \left\{ f^\star - f(x_n) \right\}$
- regret cumulé : $R_T = \sum_{n \leq T} \left(f^\star - f(x_n) \right)$

Remarque : $S_T \leq T^{-1} R_T$

Compromis Exploration/Exploitation



Compromis Exploration/Exploitation



Revue de la Littérature

Bandits Stochastiques

[Lai1985, Bubeck2012]

$$\mathcal{X} = (1, \dots, K), \quad \mathbb{E}[R_T] \approx 2\eta^2 H \log T \text{ avec } H = \sum_{x \in \mathcal{X} : f(x) < f^\star} \big(f^\star - f(x)\big)^{-1}$$

Revue de la Littérature

Bandits Stochastiques

[Lai1985, Bubeck2012]

$$\mathcal{X} = (1, \dots, K), \quad \mathbb{E}[R_T] \approx 2\eta^2 H \log T \text{ avec } H = \sum_{x \in \mathcal{X} : f(x) < f^\star} \big(f^\star - f(x)\big)^{-1}$$

Algorithmes Evolutionnaires

[Garnier2001, Eiben2003]

garanties de convergence mais peu de résultats sur la vitesse, gourmand en évaluations de f

Revue de la Littérature

Bandits Stochastiques

[Lai1985, Bubeck2012]

$\mathcal{X} = (1, \dots, K), \quad \mathbb{E}[R_T] \approx 2\eta^2 H \log T$ avec $H = \sum_{x \in \mathcal{X} : f(x) < f^\star} \big(f^\star - f(x)\big)^{-1}$

Algorithmes Evolutionnaires

[Garnier2001, Eiben2003]

garanties de convergence mais peu de résultats sur la vitesse, gourmand en évaluations de f

Optimisation Lipschitzienne

[Bull2015, Grill2015]

$\mathcal{X} \subset \mathbb{R}^D, \|f\|_{\text{Lip}} < B, \qquad \mathbb{E}[R_T] \approx T^{\frac{D+1}{D+2}}$

Revue de la Littérature

Bandits Stochastiques

[Lai1985, Bubeck2012]

$\mathcal{X} = (1, \dots, K), \quad \mathbb{E}[R_T] \approx 2\eta^2 H \log T$ avec $H = \sum_{x \in \mathcal{X} : f(x) < f^\star} \big(f^\star - f(x)\big)^{-1}$

Algorithmes Evolutionnaires

[Garnier2001, Eiben2003]

garanties de convergence mais peu de résultats sur la vitesse, gourmand en évaluations de f

Optimisation Lipschitzienne

[Bull2015, Grill2015]

$\mathcal{X} \subset \mathbb{R}^D, \|f\|_{\text{Lip}} < B, \qquad \mathbb{E}[R_T] \approx T^{\frac{D+1}{D+2}}$

Optimisation Bayésienne

[deFreitas2012, Srinivas2012]

$f \sim \mathcal{GP}(0, k), \qquad R_T \lesssim \sqrt{T \gamma_T \log(T|\mathcal{X}|)}$ avec γ_T défini plus tard

Bornes de Confiance et UCB

Intervalles de Confiance

Après n itérations, connaissant $y_1, \ldots, y_n$ les observations pour $x_1, \ldots, x_n$, imaginons que l'on peut calculer $L_n : \mathcal{X} \to \mathbb{R}$ et $U_n : \mathcal{X} \to \mathbb{R}$ tels que :

$$\forall x \in \mathcal{X}, \qquad f(x) \in \big(L_n(x), U_n(x)\big) \qquad \text{avec forte probabilité}$$

Bornes de Confiance et UCB

Intervalles de Confiance

Après n itérations, connaissant $y_1, \ldots, y_n$ les observations pour $x_1, \ldots, x_n$, imaginons que l'on peut calculer $L_n : \mathcal{X} \to \mathbb{R}$ et $U_n : \mathcal{X} \to \mathbb{R}$ tels que :

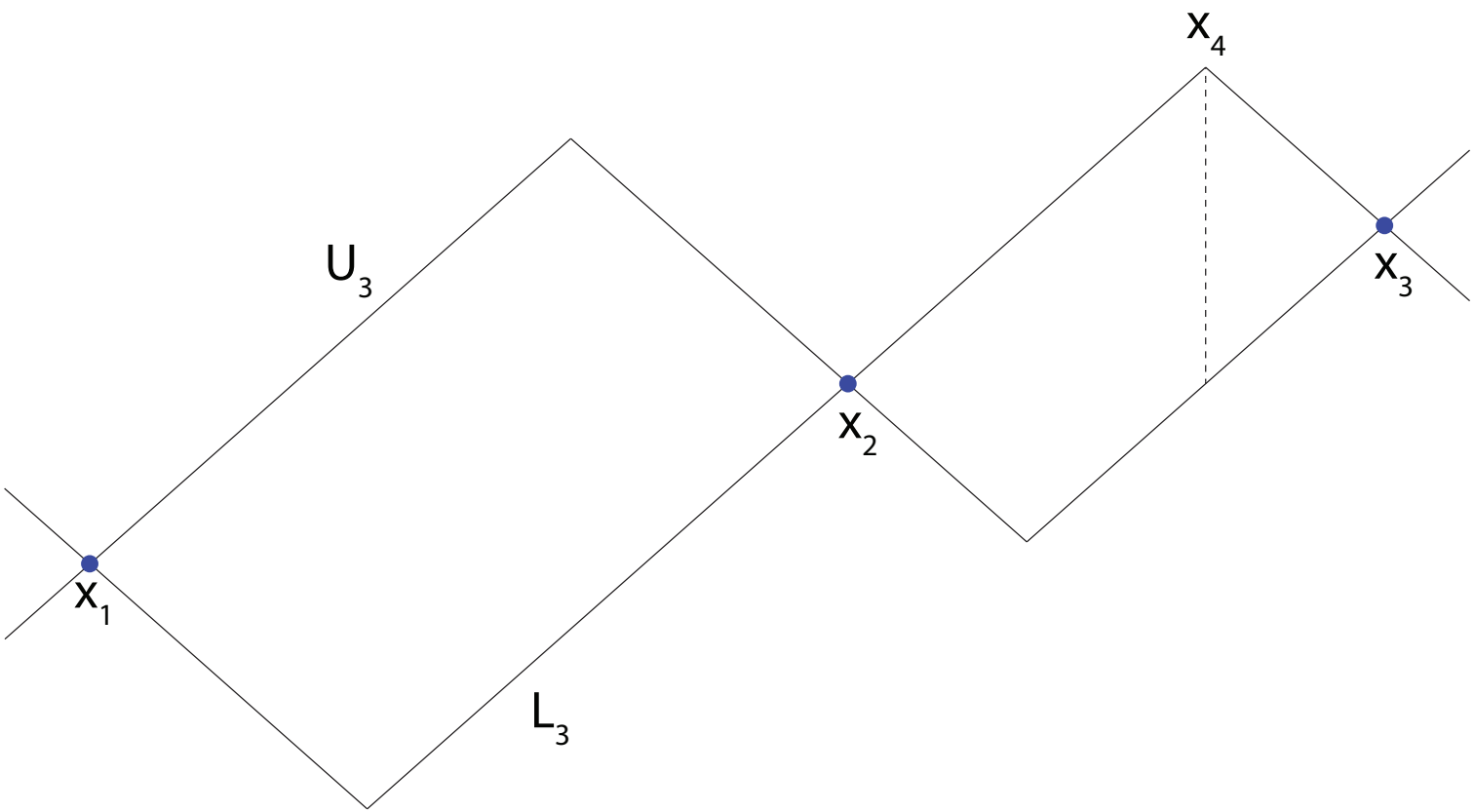
$$\forall x \in \mathcal{X}, \qquad f(x) \in \big(L_n(x), U_n(x)\big) \qquad \text{avec forte probabilité}$$

Stratégie UCB

Alors, avec $x_{n+1} = \text{argmax}_{x \in \mathcal{X}}\, U_n(x)$, on a :

$$f^\star - f(x_{n+1}) \leq U_n(x^\star) - L_n(x_{n+1}) \leq U_n(x_{n+1}) - L_n(x_{n+1})$$

Stratégie UCB pour une Fonction Lipschitzienne sans Bruit



Optimisation Bayésienne

Hypothèse de Régularité

- $f \sim \mathcal{GP}(0, k)$ avec k connu en théorie
- en pratique : famille SE ou Matérn et choix empirique des paramètres

Optimisation Bayésienne

Hypothèse de Régularité

- $f \sim \mathcal{GP}(0, k)$ avec k connu en théorie
- en pratique : famille SE ou Matérn et choix empirique des paramètres

Inférence Bayésienne

Sachant les observations $\mathbf{Y}_n = [y_1, \ldots, y_n]$ pour $X_n = (x_1, \ldots, x_n)$,

- $\mu_n(x) = \mathbb{E}[f(x) \mid X_n, \mathbf{Y}_n]$
- $\sigma_n^2(x) = \mathbb{V}[f(x) \mid X_n, \mathbf{Y}_n]$

Optimisation Bayésienne

Hypothèse de Régularité

- $f \sim \mathcal{GP}(0, k)$ avec k connu en théorie
- en pratique : famille SE ou Matérn et choix empirique des paramètres

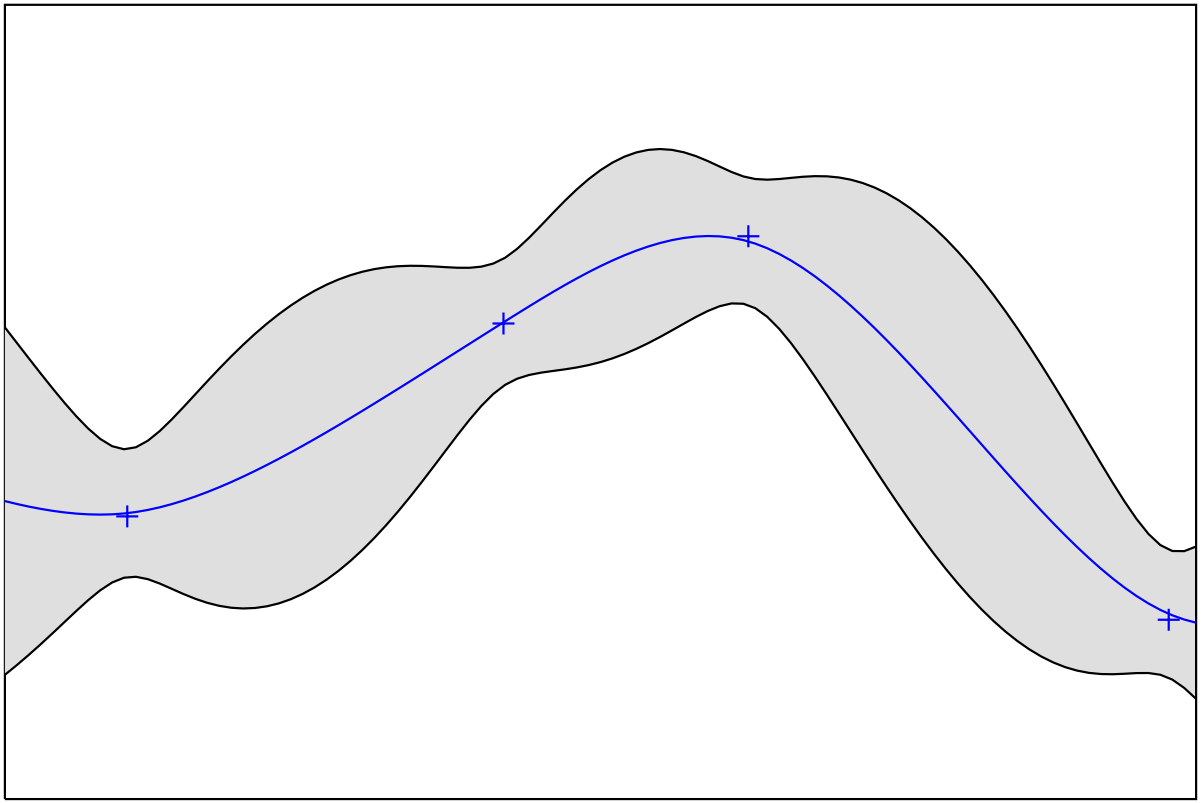
Inférence Bayésienne

Sachant les observations $\mathbf{Y}_n = [y_1, \ldots, y_n]$ pour $X_n = (x_1, \ldots, x_n)$,

- $\mu_n(x) = \mathbb{E}[f(x) \mid X_n, \mathbf{Y}_n] = \mathbf{k}_n(x)^\top \mathbf{C}_n^{-1} \mathbf{Y}_n$
- $\sigma_n^2(x) = \mathbb{V}[f(x) \mid X_n, \mathbf{Y}_n] = k(x,x) - \mathbf{k}_n(x)^\top \mathbf{C}_n^{-1} \mathbf{k}_n(x)$

où $\mathbf{C}_n = \mathbf{K}_n + \eta^{-2}\mathbf{I}$ et $\mathbf{K}_n = [k(x_i, x_j)]_{x_i, x_j \in X_n}$.

Bornes de Confiance pour un Processus Gaussien



Cas où $\mathcal{X}$ est fini

Lorsque $|\mathcal{X}| < \infty$, avec $\beta_n \approx \log(n|\mathcal{X}|)$, avec forte probabilité :

$$\forall n > 0, \forall x \in \mathcal{X}, \qquad f(x) \in \big(L_n(x), U_n(x)\big)$$

- où $L_n(x) = \mu_n(x) - \sqrt{\beta_n \sigma_n^2(x)}$
- et $U_n(x) = \mu_n(x) + \sqrt{\beta_n \sigma_n^2(x)}$

Cas où $\mathcal{X}$ est fini

Lorsque $|\mathcal{X}| < \infty$, avec $\beta_n \approx \log(n|\mathcal{X}|)$, avec forte probabilité :

$$\forall n > 0, \forall x \in \mathcal{X}, \qquad f(x) \in \big(L_n(x), U_n(x)\big)$$

- où $L_n(x) = \mu_n(x) - \sqrt{\beta_n \sigma_n^2(x)}$
- et $U_n(x) = \mu_n(x) + \sqrt{\beta_n \sigma_n^2(x)}$

pour la stratégie UCB : $R_T \leq \sum_{n=1}^{T} 2\sqrt{\beta_n \sigma_n^2(x_{n+1})}$

Cas où $\mathcal{X}$ est fini

Lorsque $|\mathcal{X}| < \infty$, avec $\beta_n \approx \log(n|\mathcal{X}|)$, avec forte probabilité :

$$\forall n > 0, \forall x \in \mathcal{X}, \qquad f(x) \in \big(L_n(x), U_n(x)\big)$$

- où $L_n(x) = \mu_n(x) - \sqrt{\beta_n \sigma_n^2(x)}$
- et $U_n(x) = \mu_n(x) + \sqrt{\beta_n \sigma_n^2(x)}$

pour la stratégie UCB : $R_T \leq \sum_{n=1}^{T} 2\sqrt{\beta_n \sigma_n^2(x_{n+1})}$

or $\sum_{n=1}^{T} \sigma_n^2(x_{n+1}) \lesssim \gamma_T$ avec $\gamma_T = \max_{\substack{S \subset \mathcal{X} \\ |S|=T}} H(f) - H(f \mid Y_S)$

Cas où $\mathcal{X}$ est fini

Lorsque $|\mathcal{X}| < \infty$, avec $\beta_n \approx \log(n|\mathcal{X}|)$, avec forte probabilité :

$$\forall n > 0, \forall x \in \mathcal{X}, \qquad f(x) \in \big(L_n(x), U_n(x)\big)$$

- où $L_n(x) = \mu_n(x) - \sqrt{\beta_n \sigma_n^2(x)}$
- et $U_n(x) = \mu_n(x) + \sqrt{\beta_n \sigma_n^2(x)}$

pour la stratégie UCB : $R_T \leq \sum_{n=1}^{T} 2\sqrt{\beta_n \sigma_n^2(x_{n+1})}$

or $\sum_{n=1}^{T} \sigma_n^2(x_{n+1}) \lesssim \gamma_T$ avec $\gamma_T = \max_{\substack{S \subset \mathcal{X} \\ |S|=T}} H(f) - H(f \mid Y_S)$

ainsi, $R_T \lesssim \sqrt{T \gamma_T \log(T|\mathcal{X}|)}$

Cas où $\mathcal{X}$ est Continu : Préliminaires

Pseudo-Métrique Canonique

$$\begin{aligned} \text{Soit } d^2(x, x') &= \mathbb{V}\big[f(x) - f(x')\big] \\ &= k(x, x) - 2k(x, x') + k(x', x') \end{aligned}$$

Cas où $\mathcal{X}$ est Continu : Préliminaires

Pseudo-Métrique Canonique

$$\begin{aligned} \text{Soit } d^2(x, x') &= \mathbb{V}\big[f(x) - f(x')\big] \\ &= k(x, x) - 2k(x, x') + k(x', x') \end{aligned}$$

Fixons $x, x' \in \mathcal{X}$, avec forte probabilité :

$$|f(x) - f(x')| \lesssim d(x, x')$$

Cas où $\mathcal{X}$ est Continu : Préliminaires

Pseudo-Métrique Canonique

$$\begin{aligned} \text{Soit } d^2(x, x') &= \mathbb{V}\big[f(x) - f(x')\big] \\ &= k(x, x) - 2k(x, x') + k(x', x') \end{aligned}$$

Fixons $x, x' \in \mathcal{X}$, avec forte probabilité :

$$|f(x) - f(x')| \lesssim d(x, x')$$

Borne pour un sous-ensemble fini $S \subset \mathcal{X}$

Pour $|S| = m$, on a avec forte probabilité :

$$\sup_{x \in S} f(x) - f(x_n) \lesssim \sqrt{\log m} \sup_{x \in S} d(x, x_n)$$

Cas où $\mathcal{X}$ est Continu : Discrétisation

ε-Recouvrement de $\mathcal{X}$

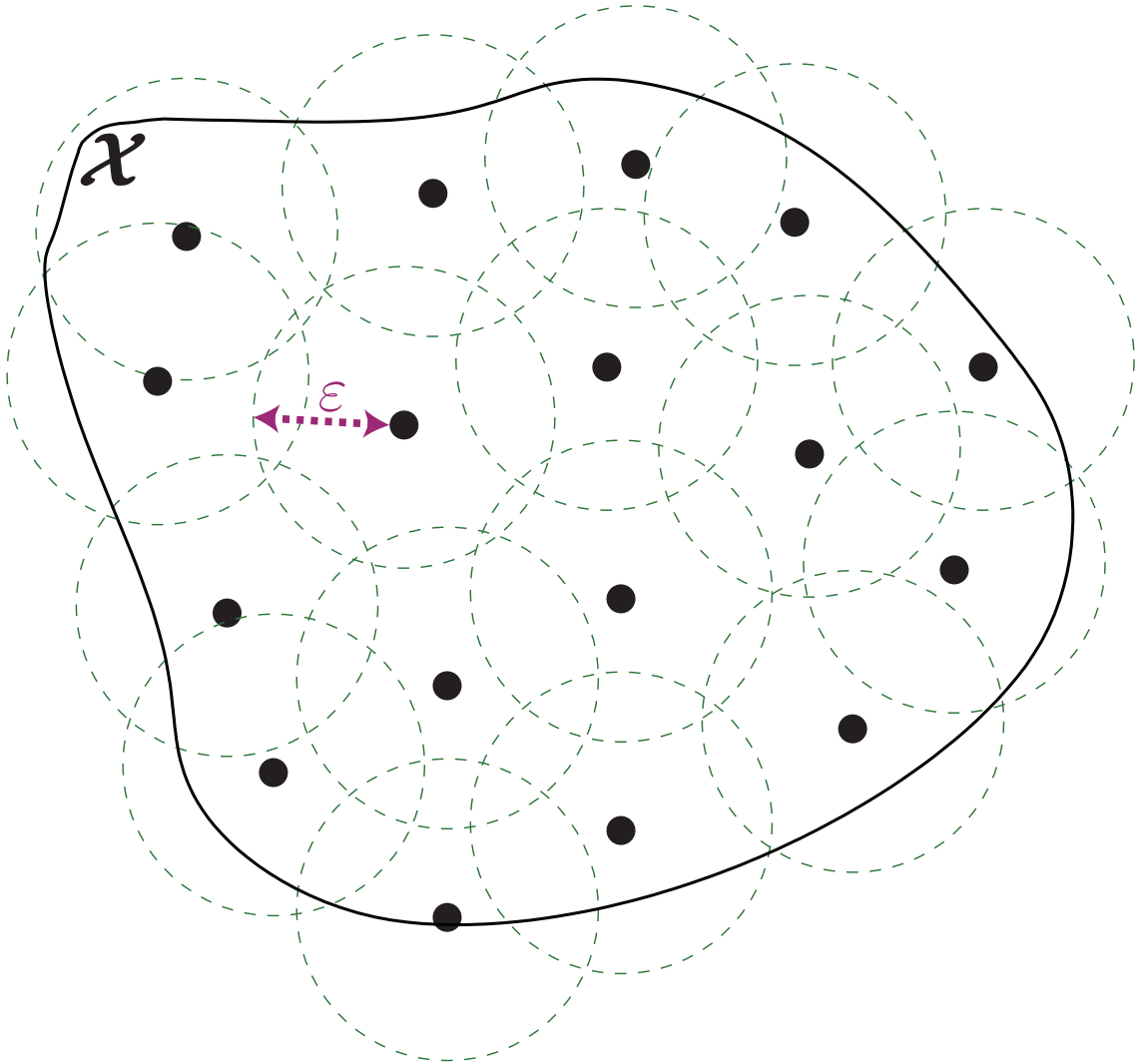
$S \subset \mathcal{X}$ est un ε-recouvrement de $\mathcal{X}$ pour d ssi :

$$\forall x \in \mathcal{X},\ \exists x' \in S \text{ t.q. } d(x, x') \leq \varepsilon$$

Entropie métrique

Soit $N(\mathcal{X}, \varepsilon)$ la taille du plus petit ε-recouvrement
$H(\mathcal{X}, \varepsilon) = \log N(\mathcal{X}, \varepsilon)$ est l'entropie métrique de $\mathcal{X}$

Un ε-Recouvrement pour la Distance Euclidienne

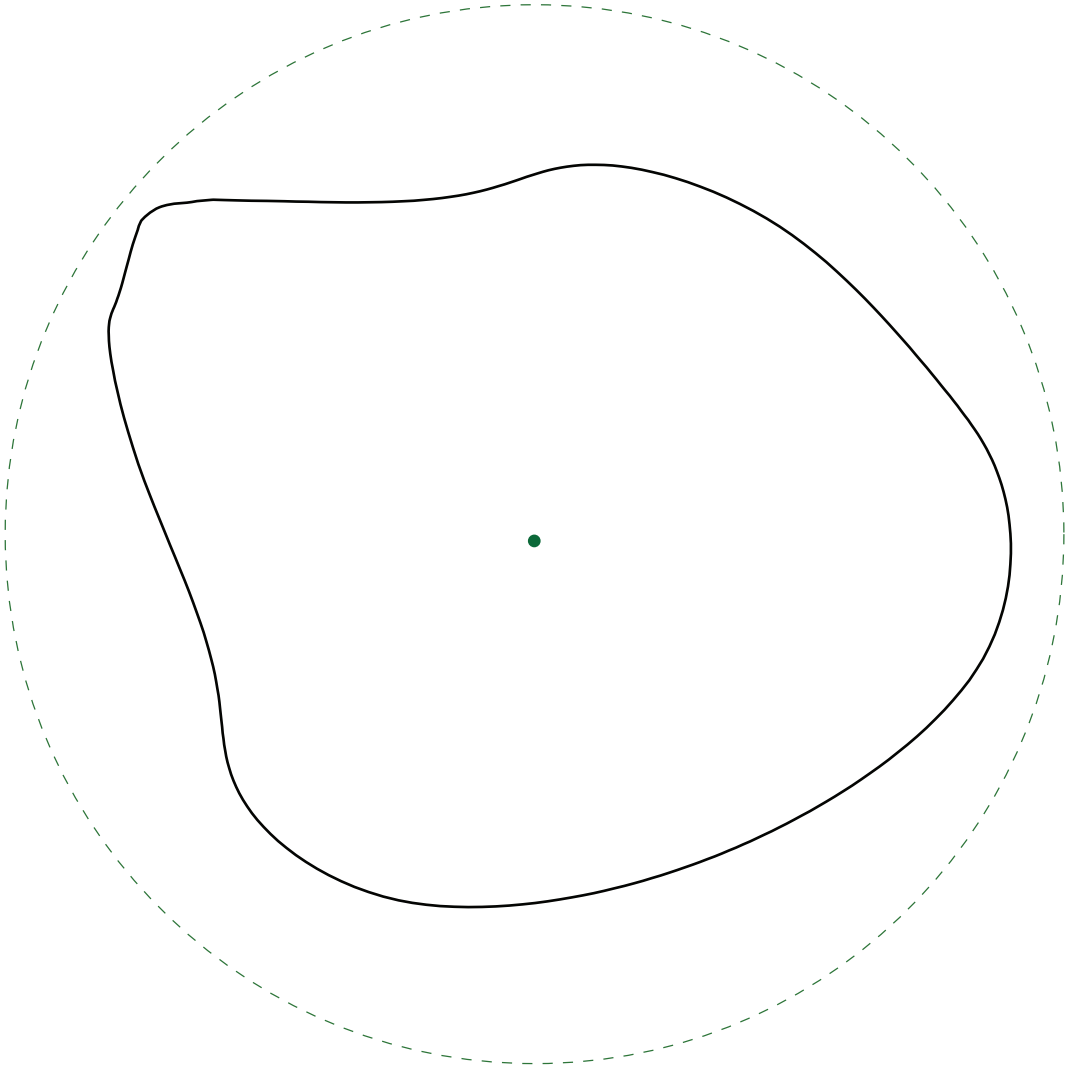


Cas où $\mathcal{X}$ est continu : Recouvrements Hiérarchiques

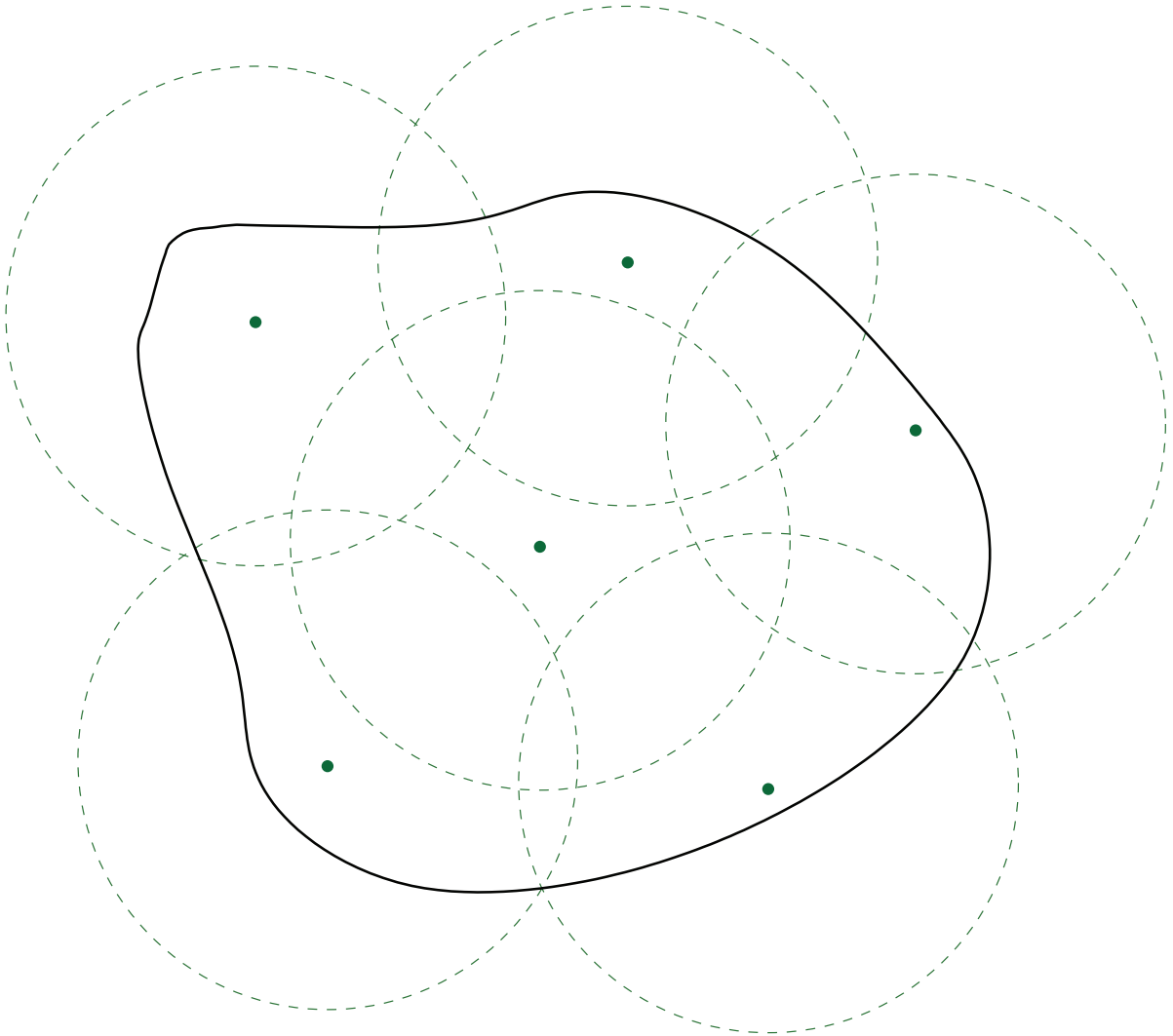
On suppose que $d(\cdot, \cdot) \leq 1$.

- $\varepsilon_0 = 1,\ \varepsilon_1 = 1/2,\ \varepsilon_2 = 1/4,\ \ldots$
- $S_0 \subset S_1 \subset \cdots \subset \mathcal{X}$
- S_i est un ε_i-recouvrement

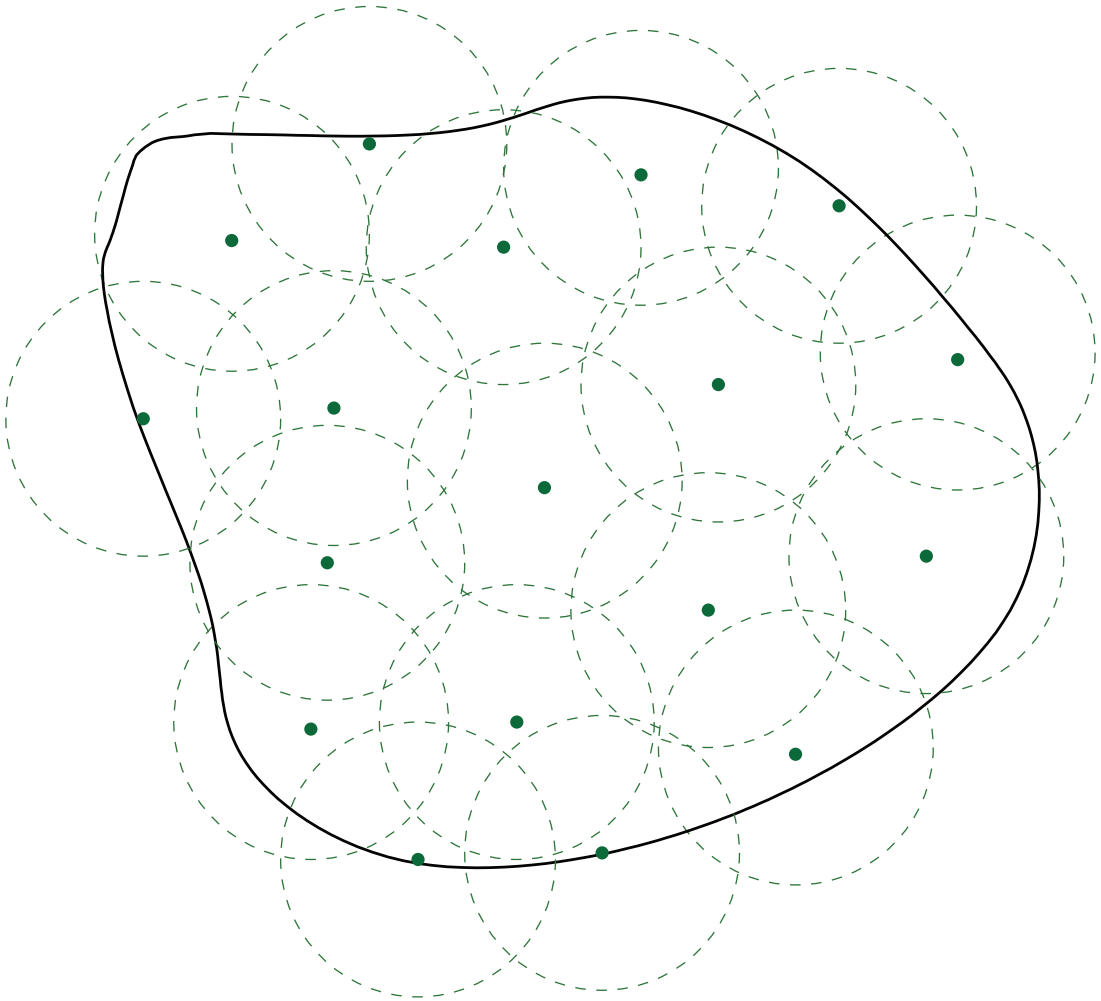
Recouvrements Hiérarchiques : $\varepsilon_0 = 1$



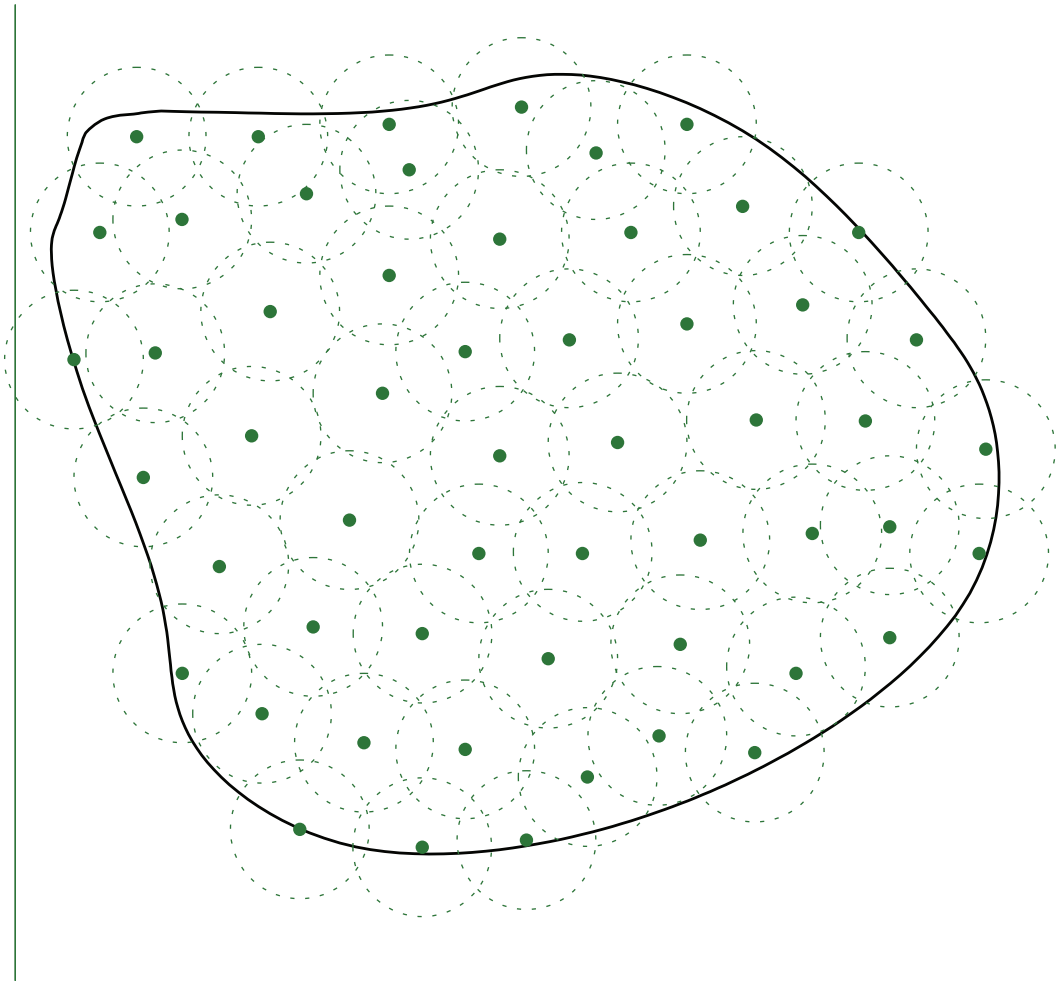
Recouvrements Hiérarchiques : $\varepsilon_1 = \frac{1}{2}$



Recouvrements Hiérarchiques : $\varepsilon_2 = \frac{1}{4}$



Recouvrements Hiérarchiques : $\varepsilon_3 = \frac{1}{8}$



Cas où $\mathcal{X}$ est continu : Chaînage

Soit $\pi_i : \mathcal{X} \to S_i$ qui associe à un point de $\mathcal{X}$ son point le plus proche de S_i

$$\sup_{s \in S_i} f(s) - f(\pi_{i-1}(s)) \lesssim \sqrt{H(\mathcal{X}, \varepsilon_i)}\varepsilon_{i-1}$$

Théorème : Erreur de Discrétisation

$$\forall i \geq 0, \quad \sup_{x \in \mathcal{X}} f(x) - f(\pi_i(x)) \lesssim \omega_i$$

$$\text{où } \omega_i = \sum_{j>i} \sqrt{H(\mathcal{X}, \varepsilon_j)}\varepsilon_{j-1}$$

Cas où $\mathcal{X}$ est continu : UCB et Discrétisations Adaptatives

A l'itération n, on choisit le niveau de discrétisation $i_n = \lceil 1/2 \log_2 n \rceil$

Alors, $\omega_{i_n} \lesssim \sqrt{\frac{d \log n}{n}}$

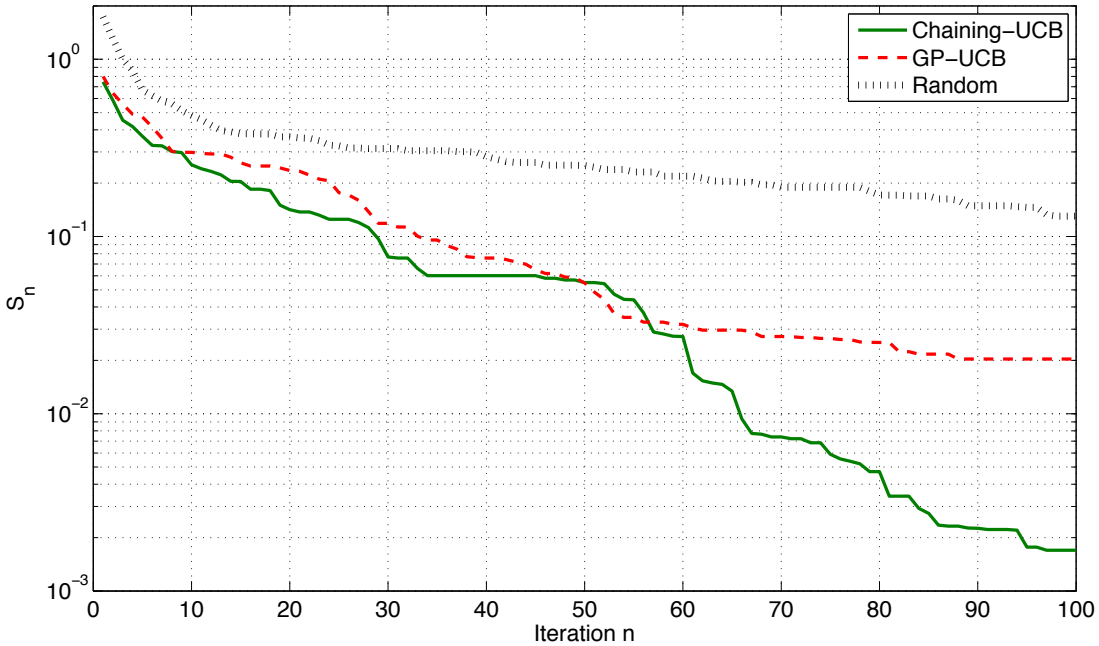
Avec $x_{n+1} = \text{argmax}_{x \in S_{i_n}} U_n(x)$ on obtient :

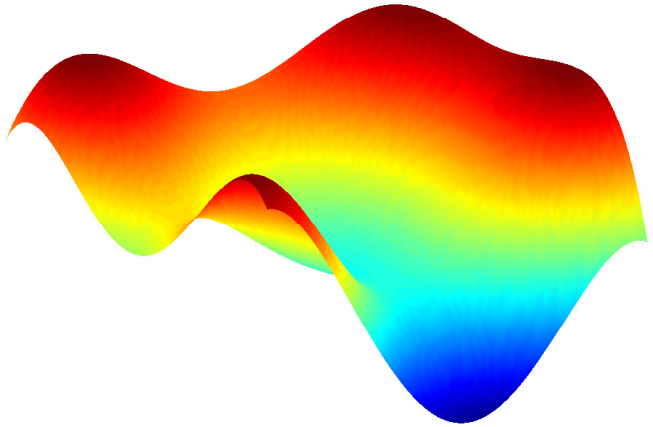
Théorème : Regret

$$R_T \lesssim \sqrt{T \gamma_T d \log^2 T}$$

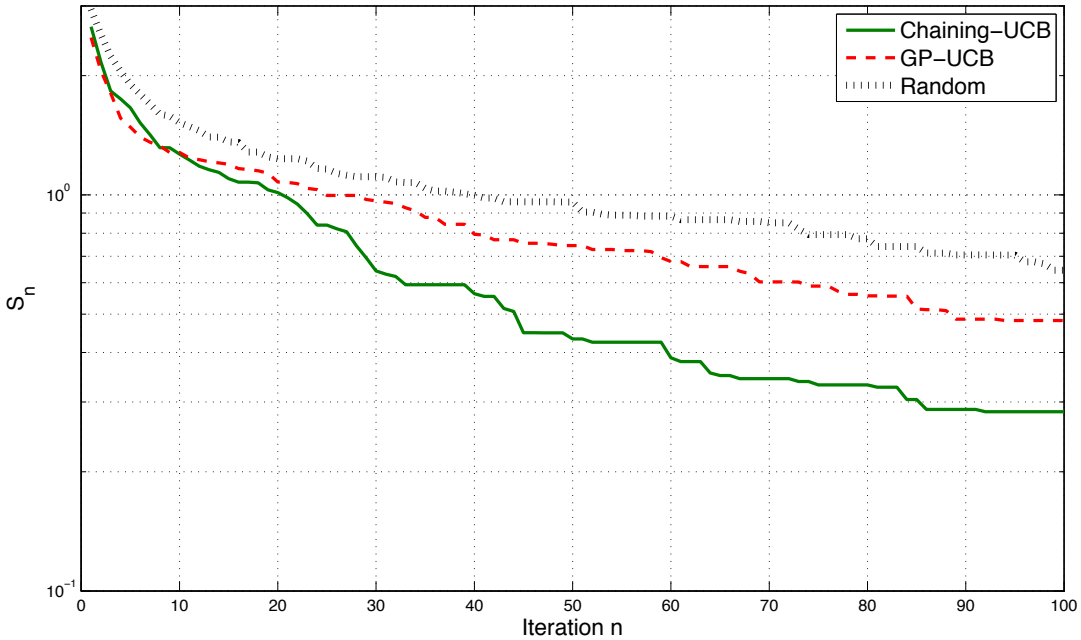
avec forte probabilité

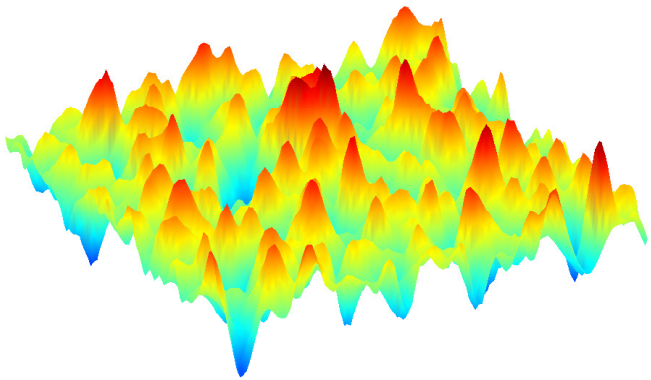
Expérience 1/4 : Fonction d'Himmelblau



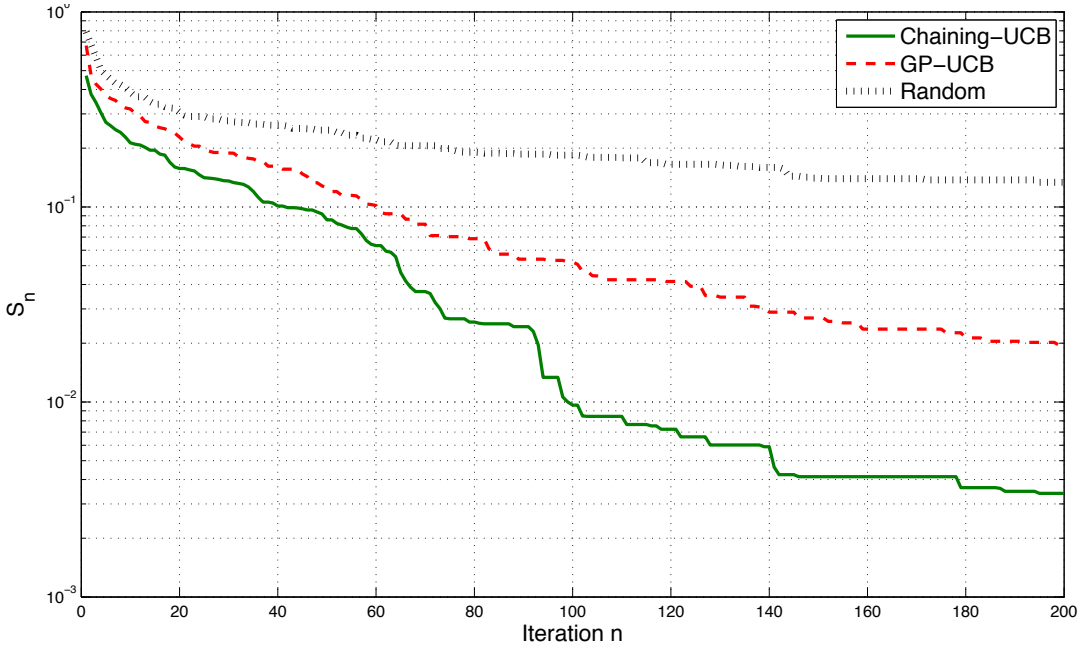


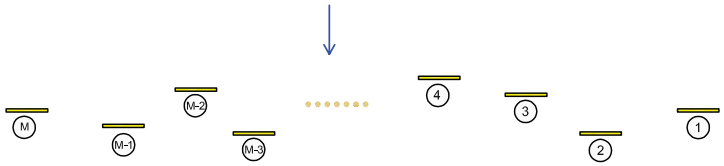
Expérience 2/4 : Noyau SE

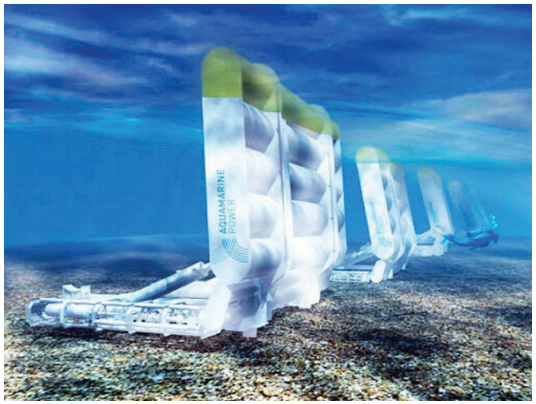




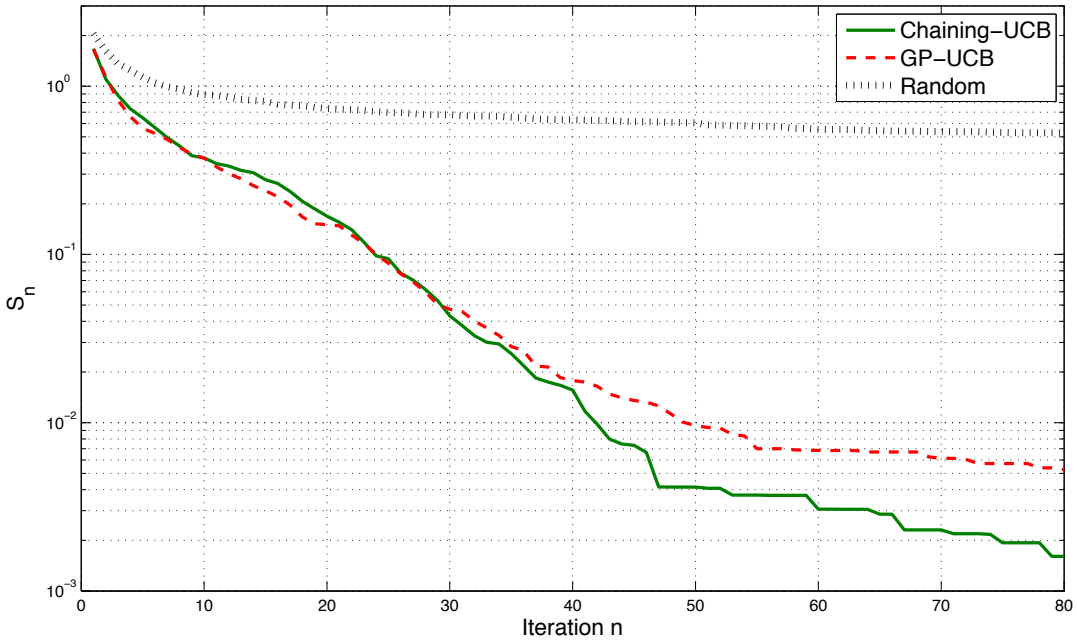
Expérience 3/4 : Wave Energy Converter

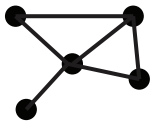


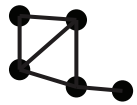


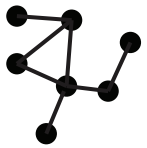


Expérience 4/4 : Noyaux de Graphes



$f($  $) = 1.3$

$f($  $) = 0.2$

$f($  $) = 2.7$

Conclusion

L'Algorithme UCB par Chaînage

- construit en suivant la théorie
- calibre automatiquement le compromis exploration/exploitation
- s'adapte à des cadres divers
- calculs raisonnables : itération en $\mathcal{O}(n^2)$

Code Matlab en Ligne (et bientôt Python)

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

Contal, E., Malherbe, C., and Vayatis, N. (2015). Optimization for gaussian processes via chaining. *NIPS Workshop on Bayesian Optimization*.

Munos, R. (2011). Optimistic optimization of deterministic functions without the knowledge of its smoothness. In *Advances in neural information processing systems 25 (NIPS)*.

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. (2012). Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265.

Talagrand, M. (2014). *Upper and Lower Bounds for Stochastic Processes: Modern Methods and Classical Problems*, volume 60. Springer-Verlag Berlin Heidelberg.