Optimization with Gaussian processes via chaining

Emile Contal Cédric Malherbe Nicolas Vayatis
CMLA, ENS Cachan & CNRS UMR 8536, France

Abstract

We consider the problem of sequential optimization of Gaussian processes. We generalize the UCB algorithm to **arbitrary kernels and search spaces**. We provide a novel optimization scheme based on **covering numbers** to automatically calibrate the exploration-exploitation tradeoff. We show how to build efficiently the algorithm and demonstrate that it is empirically more efficient than its natural competitors on simple and **complex input spaces**.

The Chaining Trick

1. Related Work

Examples of Optimization Problems

- Minimization of energy consumption in engineering
- Maximization of benefits in marketing
- Minimization of validation error in machine learning

Related Work

- Bayesian optimization — Jones & al.(1998), Hennig & al.(2012)
- UCB — Freitas & al.(2012), Srinivas & al.(2012), Djolonga & al.(2013)
- Chaining — Grunewalder & al.(2010), Gaillard & Gerchinovitz.(2015)

2. Problem Statement

Sequential Optimization

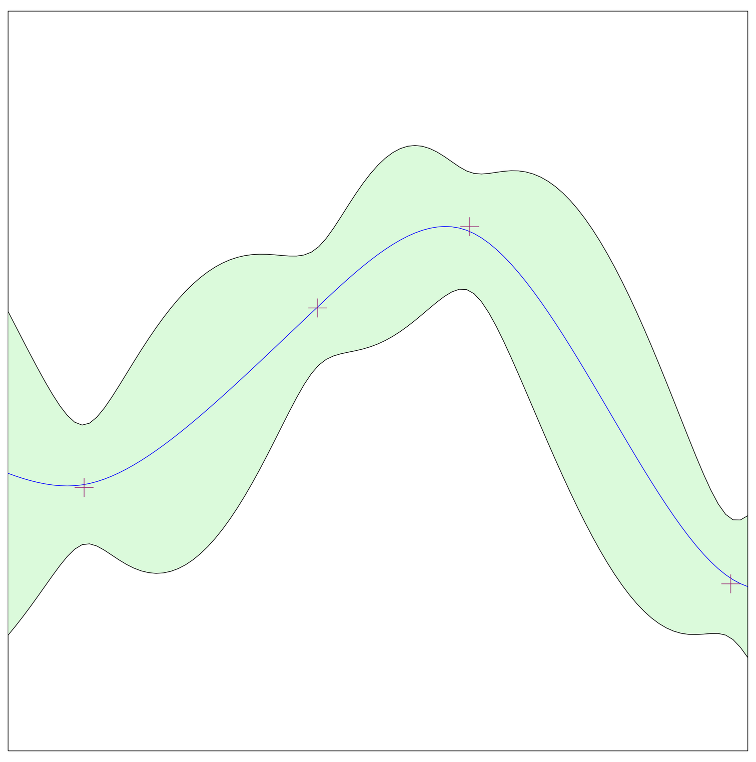
Let $f : \mathcal{X} \to \mathbb{R}$. We consider the problem of finding the maximum of f denoted by: $f(x^\star) = \sup_{x \in \mathcal{X}} f(x)$, via sequential queries $f(x_1), f(x_2), \dots$
At iteration T we choose x_{T+1} using the previous noisy observations $Y_T = \{y_1, \dots, y_T\}$, where $\forall t \leq T : y_t = f(x_t) + \epsilon_t$ and $\epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \eta^2)$.

Objective

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

$$S_T = \min_{t<T} f(x^\star) - f(x_t) \text{ and } R_T = \sum_{t<T} f(x^\star) - f(x_t)$$

3. Gaussian Processes



Definition

$f \sim \mathcal{GP}(m, k)$ with mean $m : \mathcal{X} \to \mathbb{R}$ and covariance (kernel) $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$, when for all $x_1, \dots, x_n$ we have that $(f(x_1), \dots, f(x_n))$ is a multivariate Gaussian $\mathcal{N}([m(x_i)]_i, [k(x_i, x_j)]_{i,j})$.

Bayesian Inference

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

4. Covering Numbers

Canonical pseudo-distance

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

With high probability:

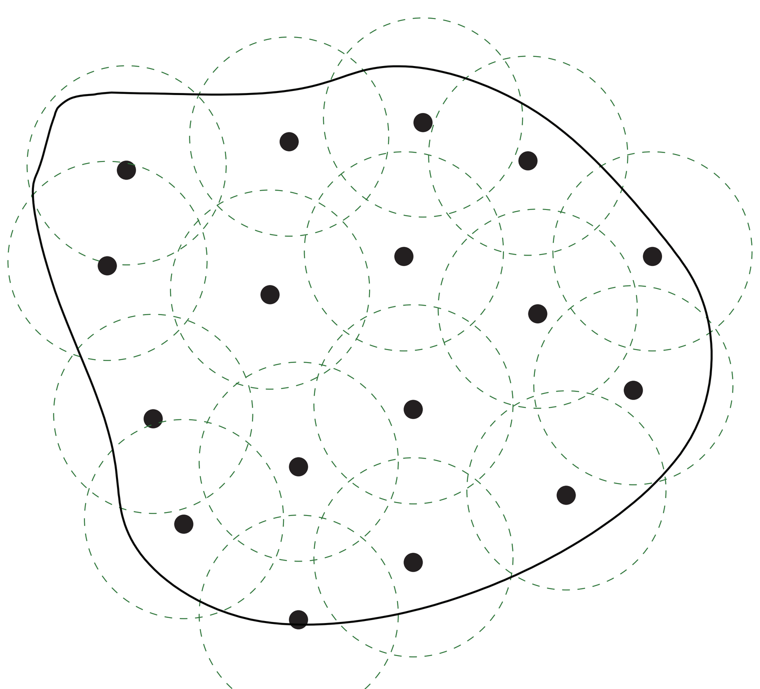
$$f(x) - f(x') \lesssim \mu_t(x) - \mu_t(x') + d_t(x, x')$$

ϵ-Cover and Covering Numbers

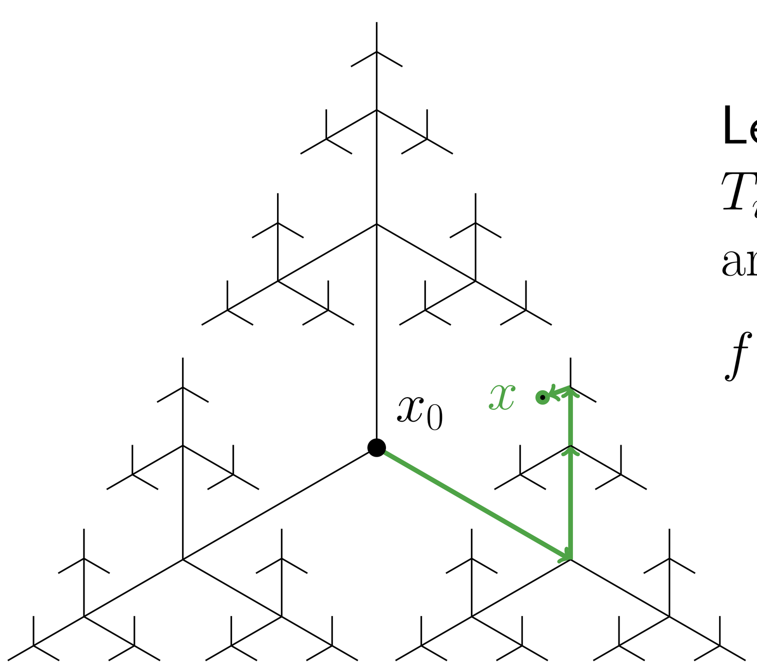
$T \subset \mathcal{X}$ is an ϵ-cover of $\mathcal{X}$ for d_t iff:

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

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



5. The Chaining Trick



Let $\epsilon_i = 2^{-i}$, and T_i be ϵ_i-covers with $T_i \subset T_{i+1}$. Define $\pi_0(\cdot) = x_0$ and $\pi_i(x) = \operatorname{argmin}_{x' \in T_i} d_t(x, x')$, then:

$$f(x) - f(x_0) = \sum_{i \geq 1} f(\pi_i(x)) - f(\pi_{i-1}(x))$$
$$\lesssim \mu_t(x) - \mu_t(x_0) + \sum_{i \geq 1} \epsilon_{i-1}$$

Algorithm & Results

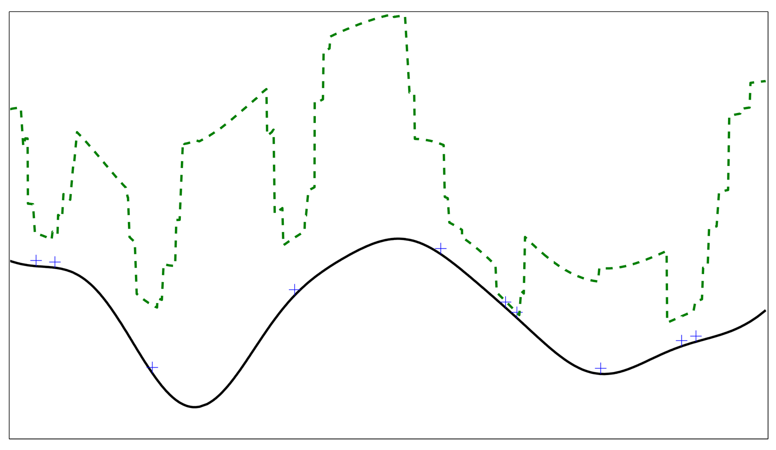
6.1 Chaining-UCB

for $t = 1, 2, \dots$ **do**
  Compute μ_t, σ_t and d_t
  $T_0 \leftarrow \emptyset$; $\sigma_t^{\text{min}} = \min_{x \in \mathcal{X}} \sigma_t(x)$
  for $i = 1 \dots \lceil 1 - \log_2(\sigma_t^{min}) \rceil$ **do**
    $\epsilon_i \leftarrow 2^{-i+1}$
    $\bar{\mathcal{X}} \leftarrow \{x \in \mathcal{X} : d_t(x, T_{i-1}) > \epsilon_i\}$
    $T_i \leftarrow T_{i-1} \cup \textsc{Cover}(\bar{\mathcal{X}}, d_t, \epsilon_i)$
    $H_i \leftarrow \epsilon_i \sqrt{2 \log \left((|T_i| + 1) i^2 t^2 \frac{\pi^4}{\delta 6^2} \right)}$
  end
  $x_t \leftarrow \operatorname*{argmax}_{x \in \mathcal{X}} \mu_t(x) + \sum_{i : \sigma_t^{\text{min}} \leq \epsilon_i < \sigma_t(x)} H_i$
  Sample x_t and observe y_t
end

6.2 Greedy-Cover

$T \leftarrow \emptyset$
$\bar{\mathcal{X}} \leftarrow \mathcal{X}$
$\forall x, x' \in \mathcal{X}, \; G[x, x'] \leftarrow \mathbb{1}_{d(x,x') \leq \epsilon}$
while $\bar{\mathcal{X}} \neq \emptyset$ **do**
  $x \leftarrow \operatorname{argmax}_{x \in \bar{\mathcal{X}}} \sum_{x' \in \bar{\mathcal{X}}} G[x, x']$
  $T \leftarrow T \cup \{x\}$
  $\bar{\mathcal{X}} \leftarrow \bar{\mathcal{X}} \setminus \{x' \in \bar{\mathcal{X}} : G[x, x'] = 1\}$
end
return T

7. Upper Confidence Bound



With probability at least $1 - \delta$ we have:

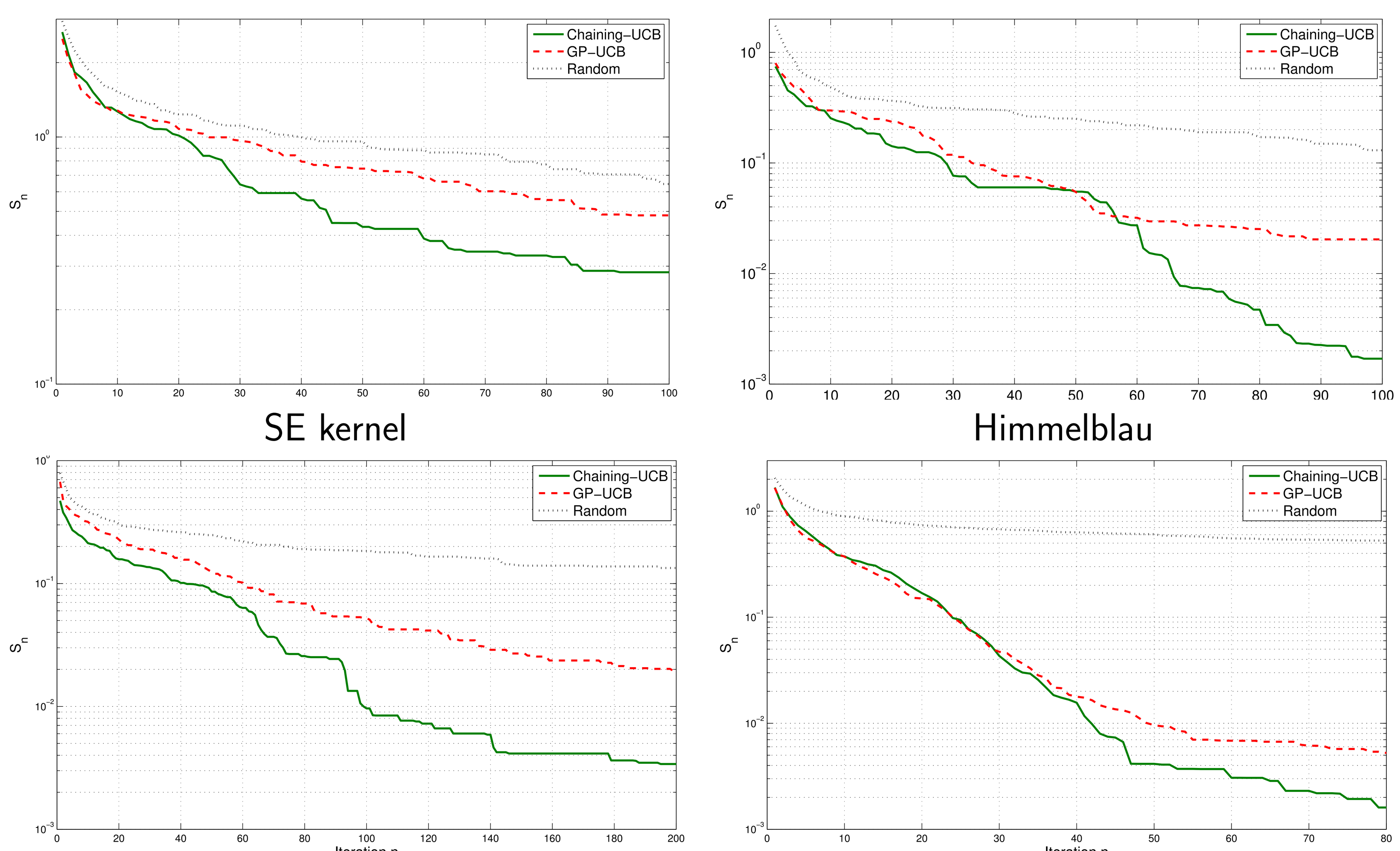
$$f(x^\star) - f(x_t) \lesssim \sum_{i : \epsilon_i < \sigma_t(x_t)} \epsilon_i \sqrt{\log N(\mathcal{X}, d_t, \epsilon_i)}$$

8. Guarantees on the regrets

For the SE kernel $k(x, x') = e^{-\frac{1}{2}\|x - x'\|_2^2}$ and a compact $\mathcal{X} \subseteq [0, R]^D$, the Chaining-UCB algorithm returns a sequence $x_1, x_2, \dots$ such that:

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

9. Experiments



SE kernel — Himmelblau — WEC — Graph kernel

References

S. Grunewalder, J-Y. Audibert, M. Opper, and J. Shawe-Taylor. Regret bounds for Gaussian process bandit problems. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 273–280. MIT Press, 2010.

C. E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

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

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