New algorithms for global optimization with Gaussian processes

Emile Contal
PhD student with Nicolas Vayatis
emile.contal@cmla.ens-cachan.fr

CMLA, ENS Cachan

June 17, 2013

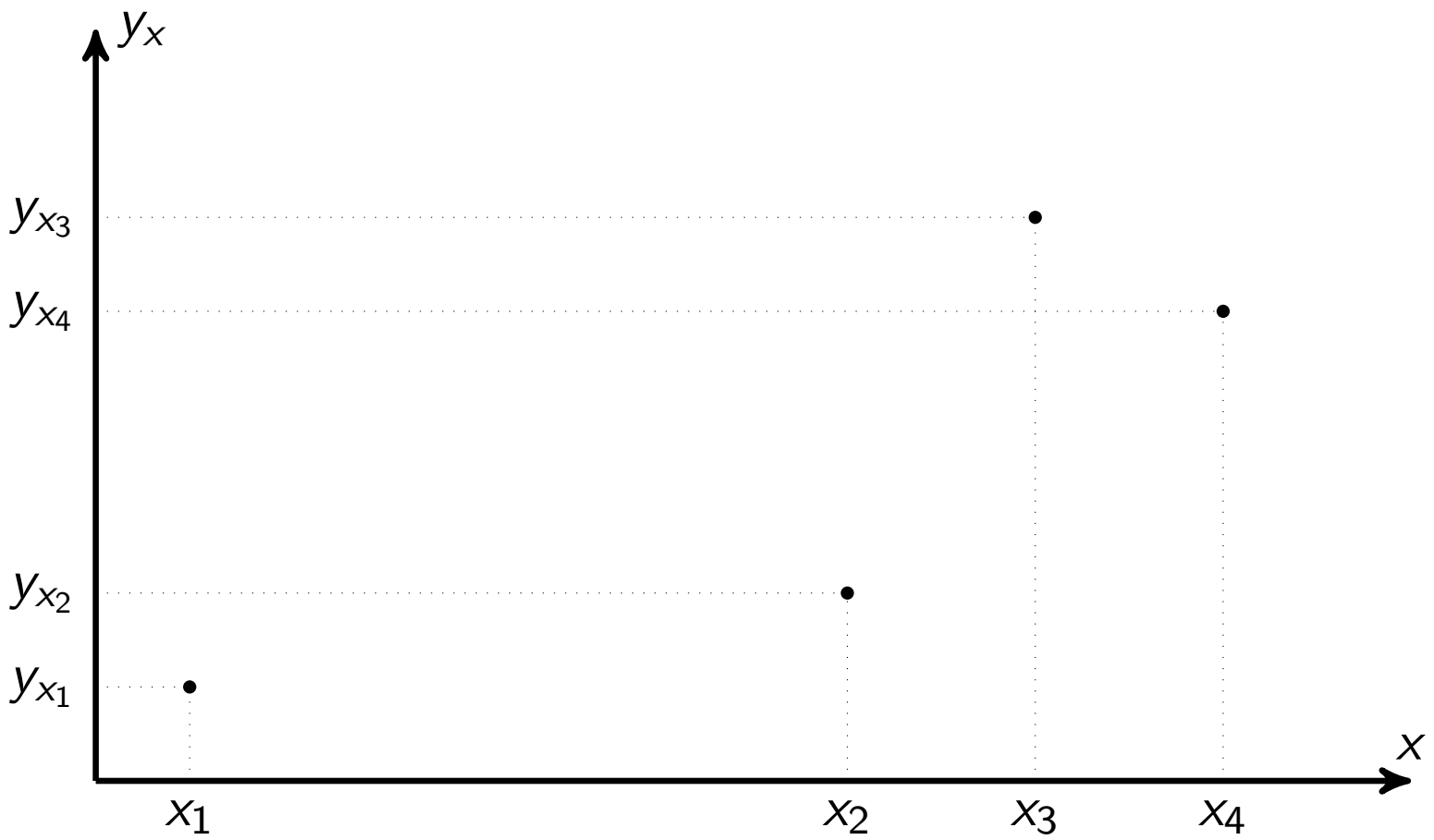
Introduction

Gaussian Processes

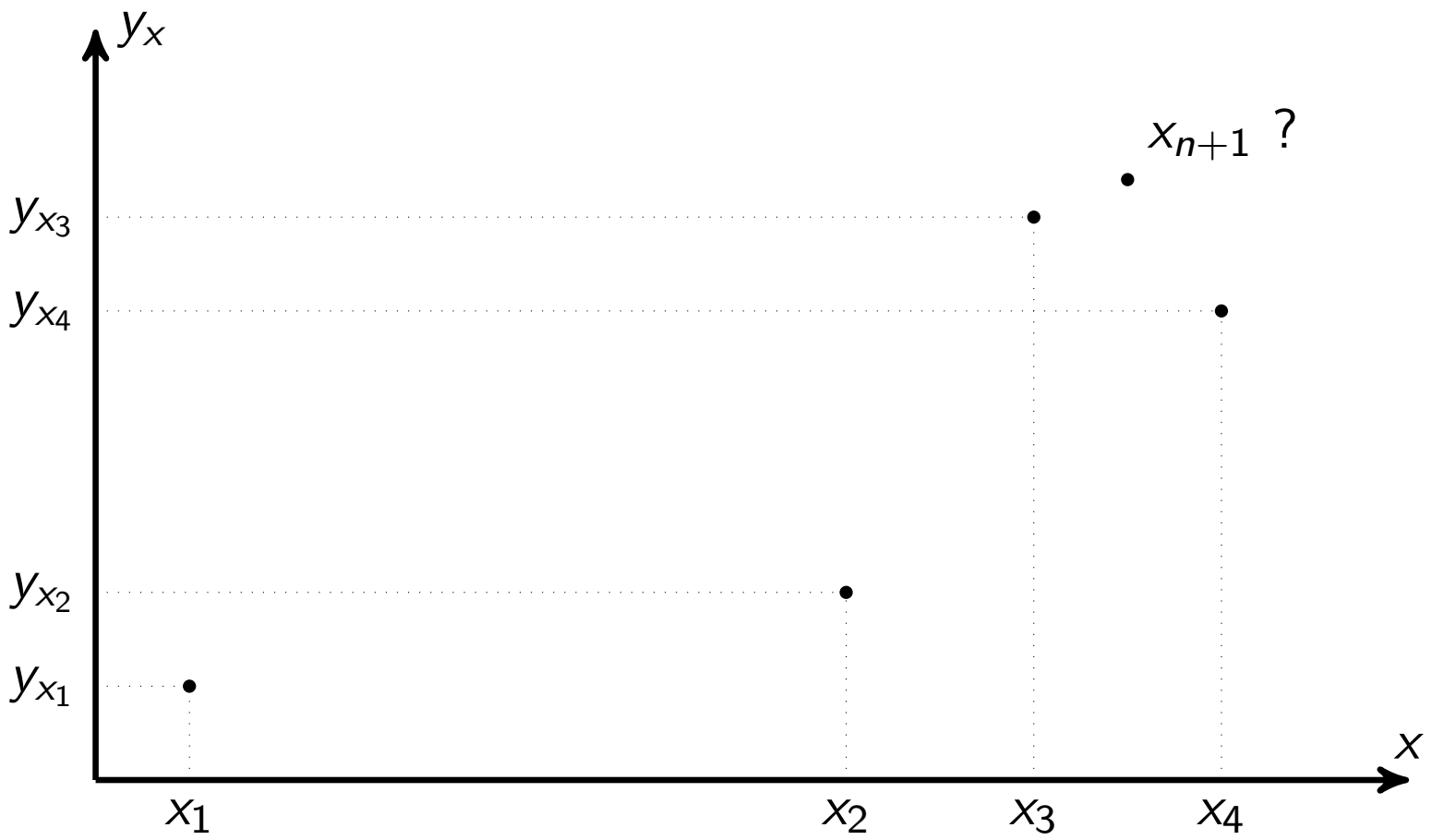
GP-UCB

Parallelism

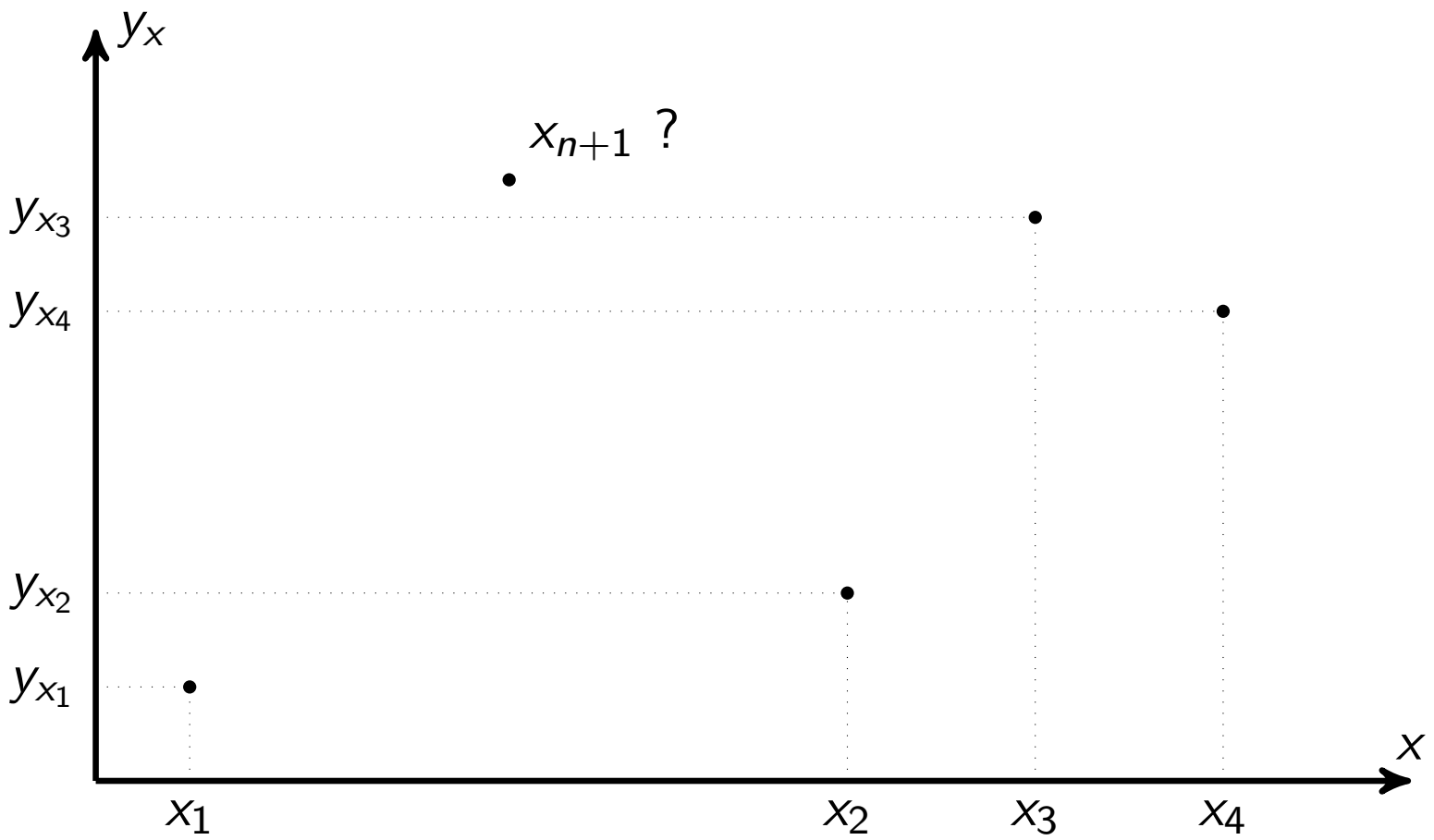
Motivating example



Motivating example



Motivating example



Sequential global optimization

Setup

- unknown $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} \in \mathbb{R}^d$
- $f(x^\star) = \max_{x \in \mathcal{X}} f(x)$
- $x_1, x_2, \cdots \in \mathcal{X}$
- $y_1, y_2, \cdots \in \mathbb{R}$, such that $y_t = f(x_t) + \epsilon_t$ where $\epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$

Examples

- Heavy numerical experiment
- Laboratory experiment
- Sensor placement

Objective

Challenge

- Search space in high dimension
- Evaluations are expensive
- Exploration / Exploitation

Cumulative regret

- $r_t = f(x^\star) - f(x_t)$
- $R_T = \sum_{t=1}^{T} f(x^\star) - f(x_t)$

Introduction

Gaussian Processes

GP-UCB

Parallelism

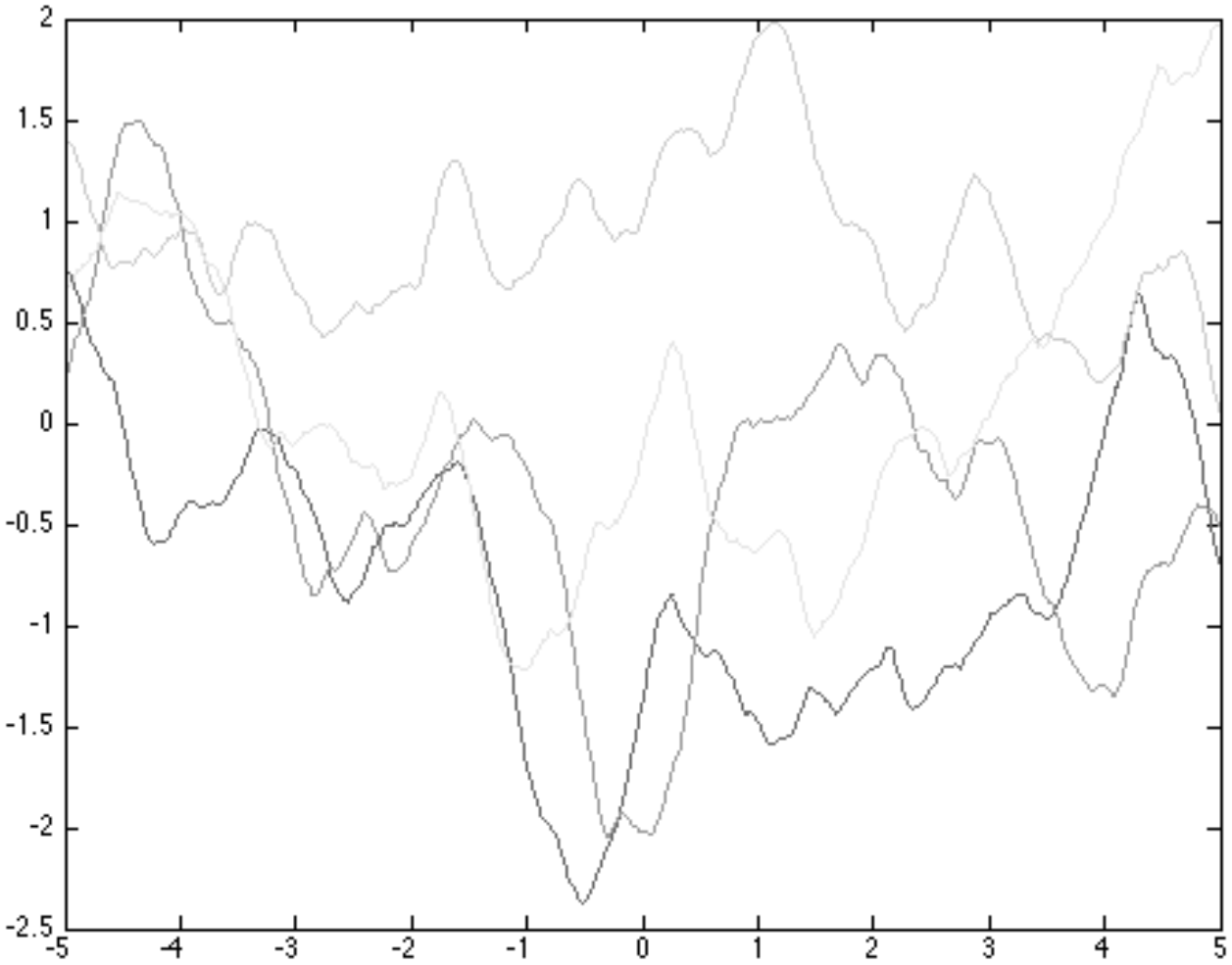
Framework

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 $x_1, \ldots, x_n$,

$$\big(f(x_1), \ldots, f(x_n)\big) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C}) \ ,$$
$$\text{with } \boldsymbol{\mu}[x_i] = m(x_i)$$
$$\text{and } \mathbf{C}[x_i, x_j] = k(x_i, x_j) \ .$$

Four 1D examples



Posterior distribution

Bayesian Inference (Rasmussen and Williams, 2005)

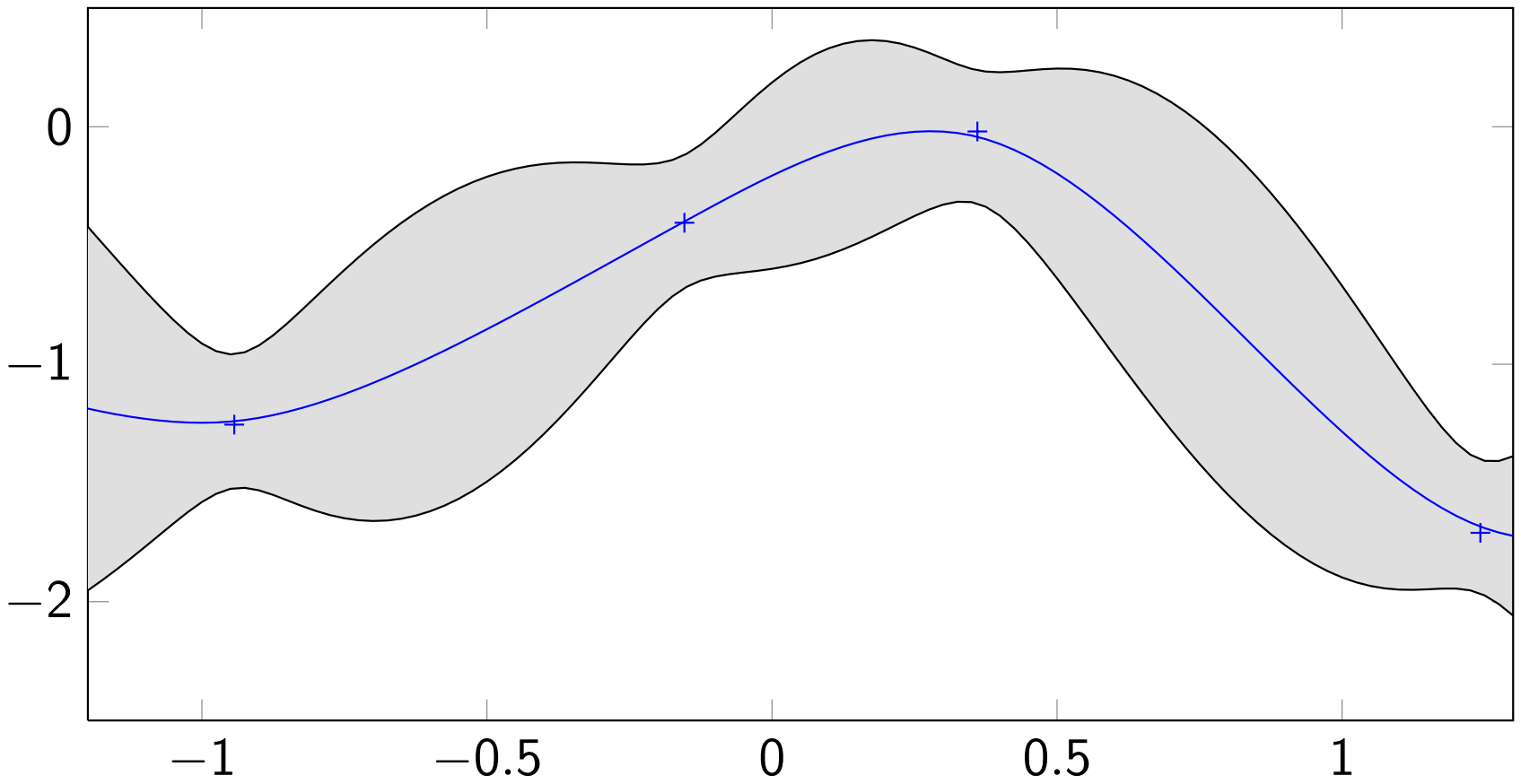
At iteration T, with observations $\mathbf{Y}_{X_T}$ at $X_T = \{x_1, \dots, x_T\}$, the posterior mean and variances are given by,

$$\mu_{T+1}(x) = \mathbf{k}_T(x)^\top \mathbf{C}_T^{-1} \mathbf{Y}_{X_T} \tag{1}$$

$$\sigma^2_{T+1}(x) = k(x,x) - \mathbf{k}_T(x)^\top \mathbf{C}_T^{-1} \mathbf{k}_T(x) \ , \tag{2}$$

where $\mathbf{C}_T = \mathbf{K}_T + \sigma^2 \mathbf{I}$ and $\mathbf{K}_T = [k(x_t, x_{t'})]_{x_t, x_{t'} \in X_T}$.

Example



Introduction

Gaussian Processes

GP-UCB

Parallelism

Upper and Lower bounds

Definition

$$f_T^+(x) = \mu_T(x) + \beta_T \sigma_T(x)$$
$$f_T^-(x) = \mu_T(x) - \beta_T \sigma_T(x)$$

Property

$$\forall x \in \mathcal{X}, \forall T \geq 1,$$

$$f(x) \in \left[f_T^-(x), f_T^+(x)\right] \textit{ with high probability}$$

GP-UCB (Srinivas et al., 2012)

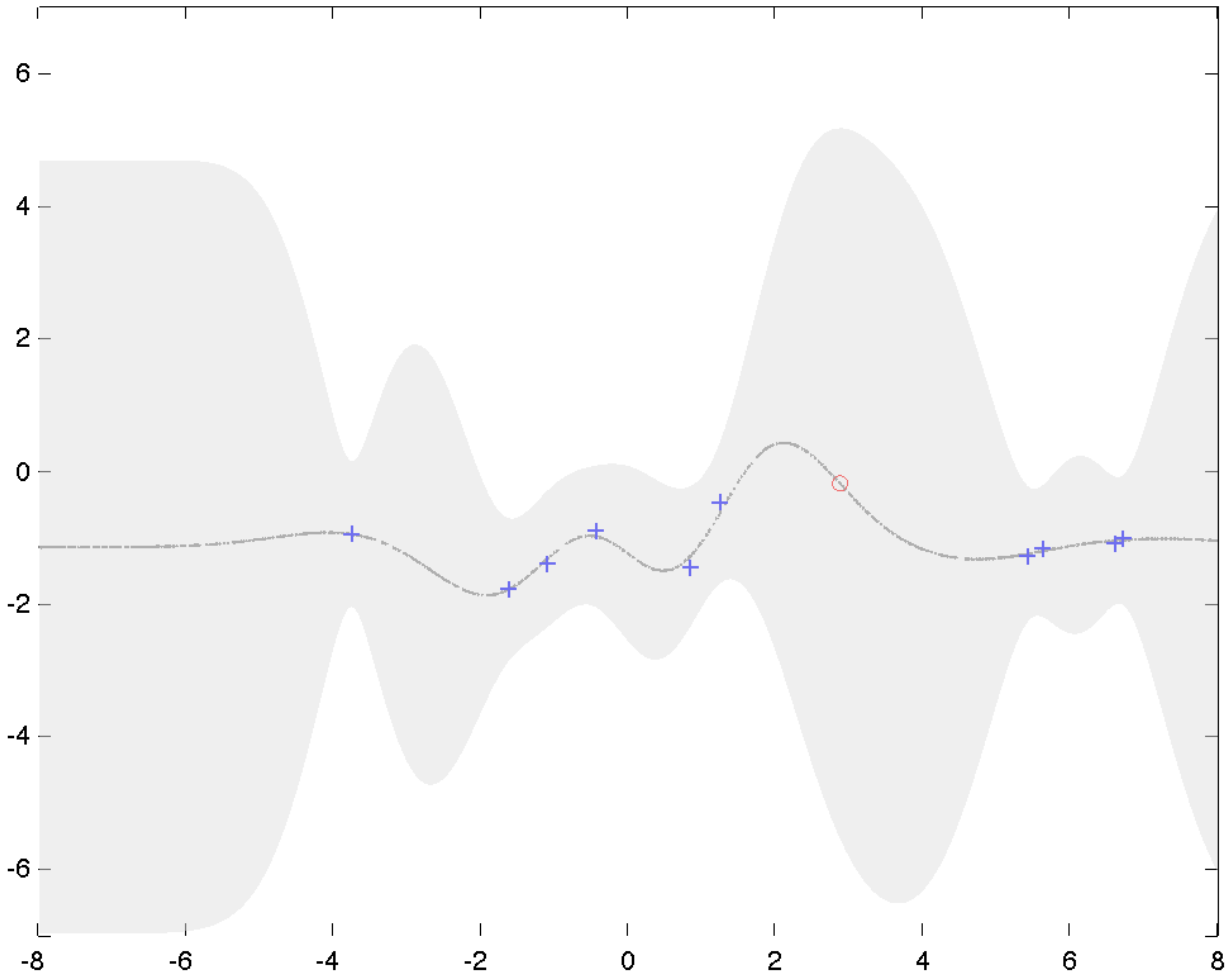
Algorithm

Algorithm 1: GP-UCB

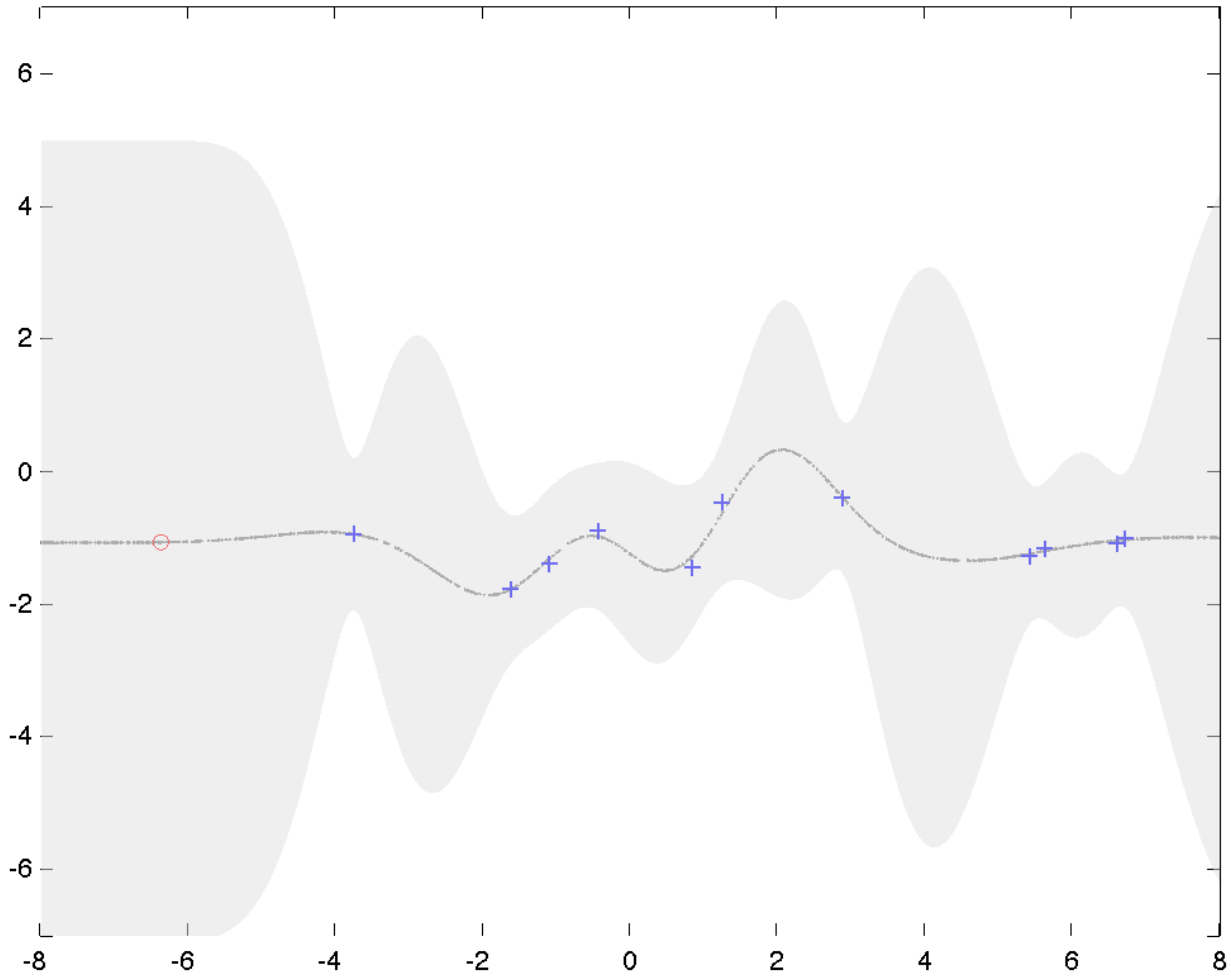
for $t = 0, 1, \ldots$ **do**

- Compute μ_t and σ_t^2 (Eq. 1, 2) with $y_1, \ldots, y_{t-1}$
- $x_t \leftarrow \text{argmax}_{x \in \mathcal{X}} f_t^+(x)$
- Query x_t and observe y_t

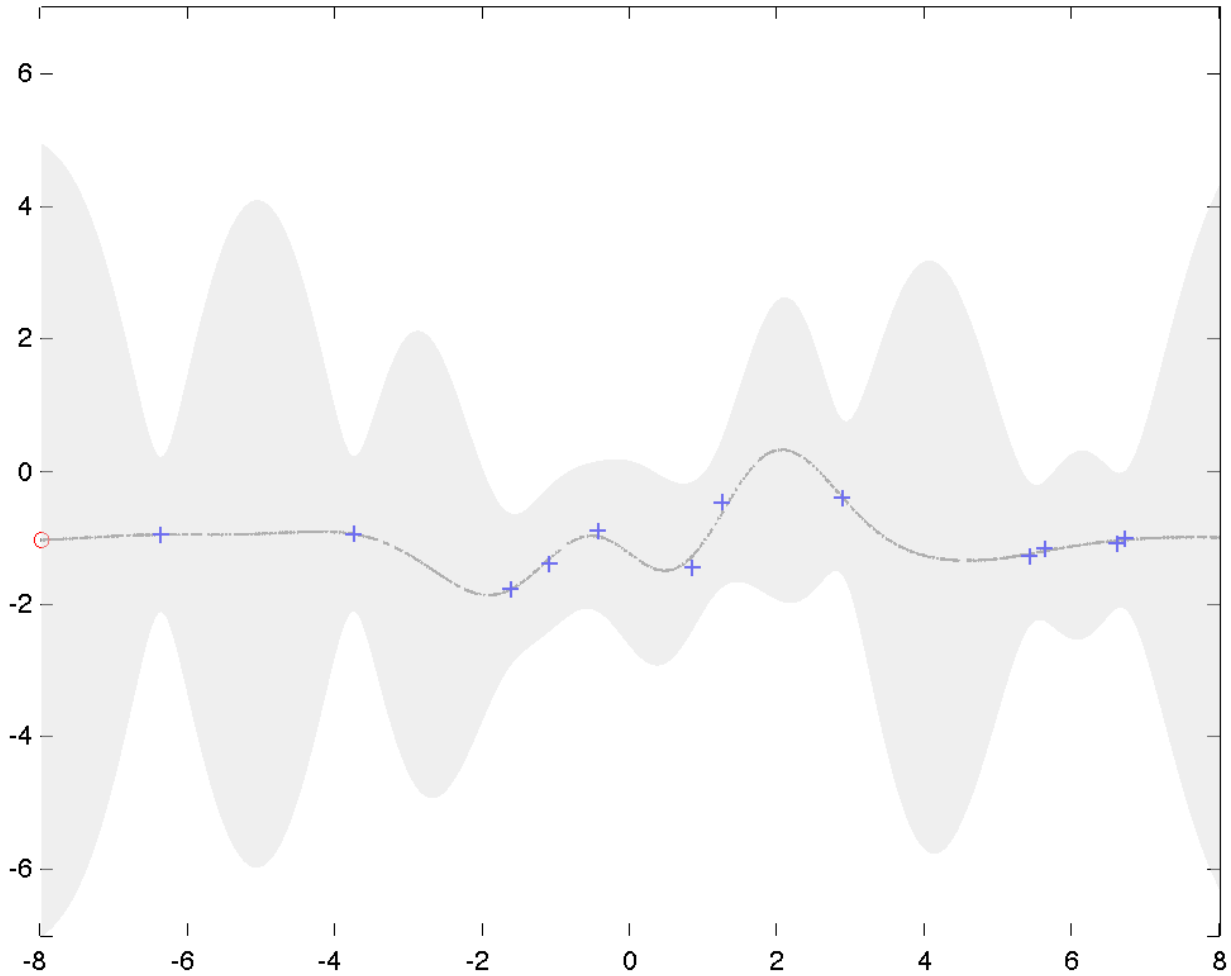
Example



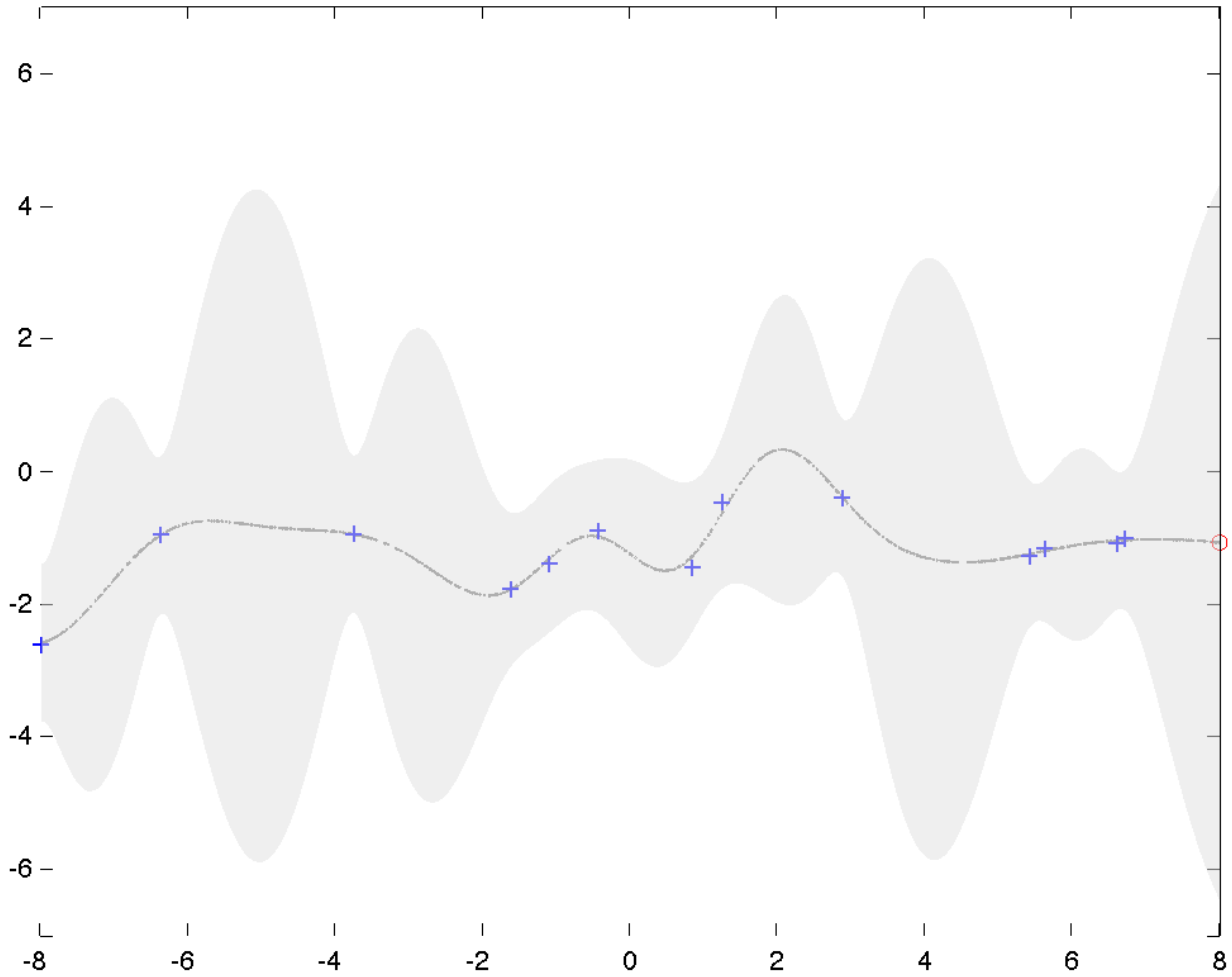
Example



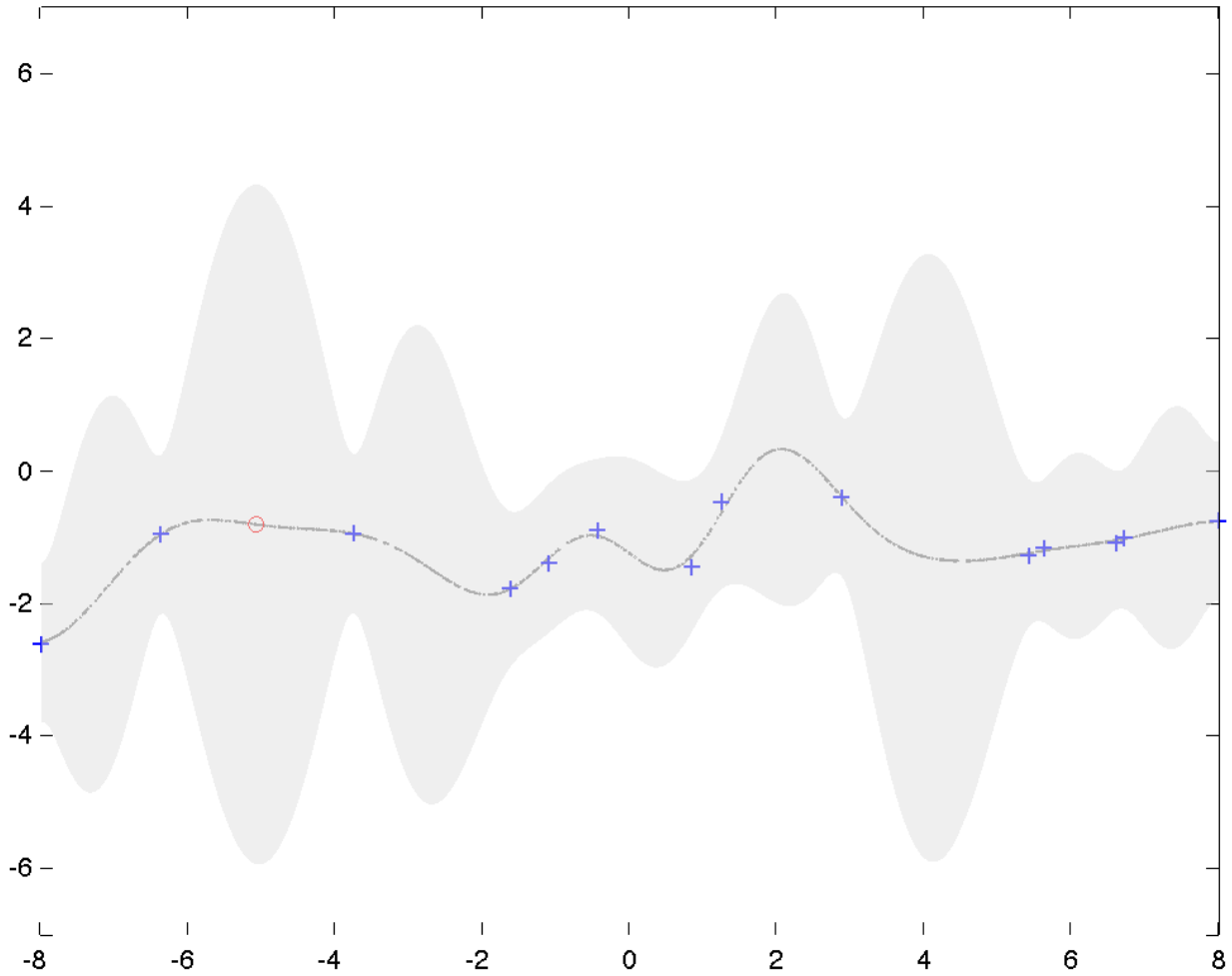
Example



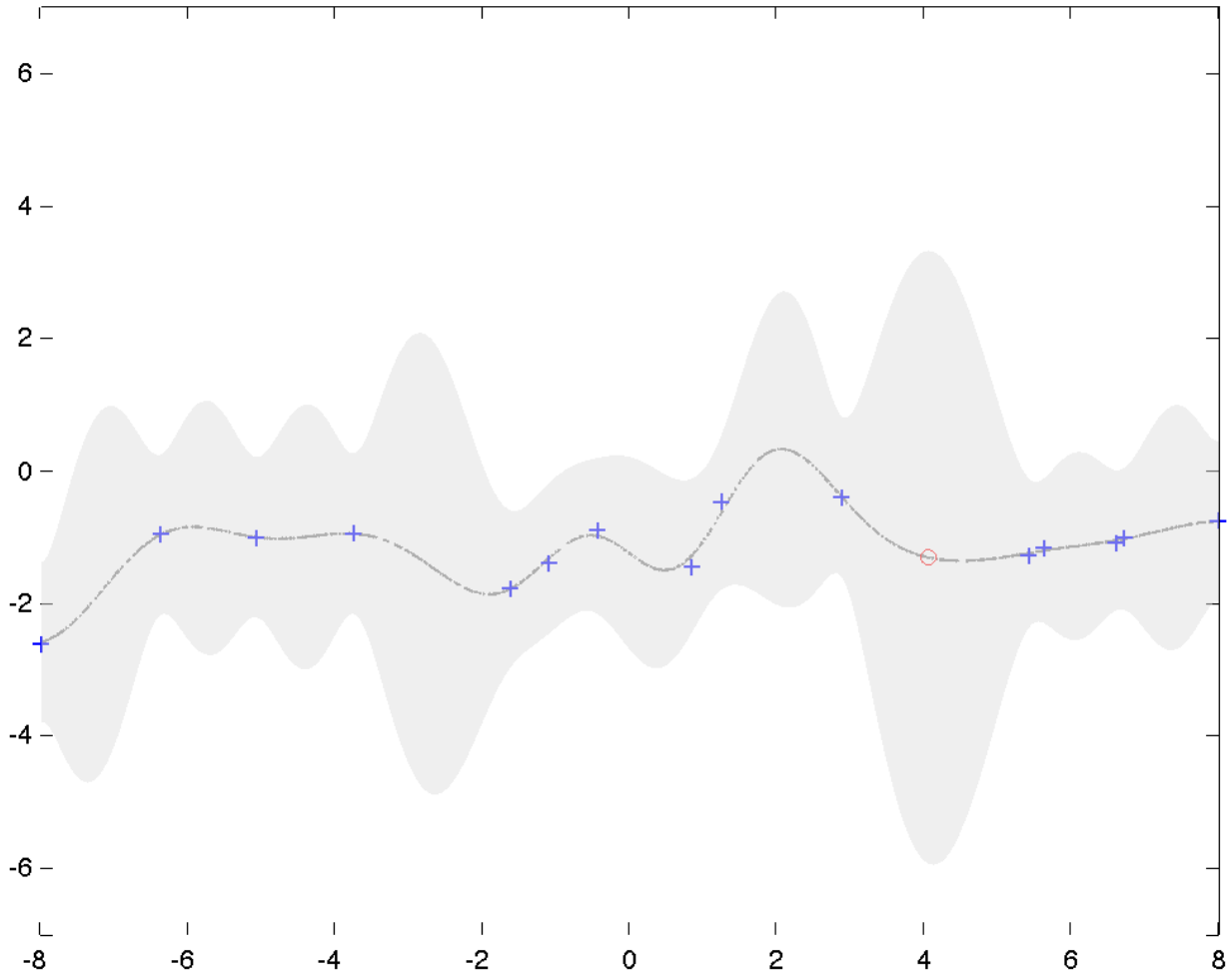
Example



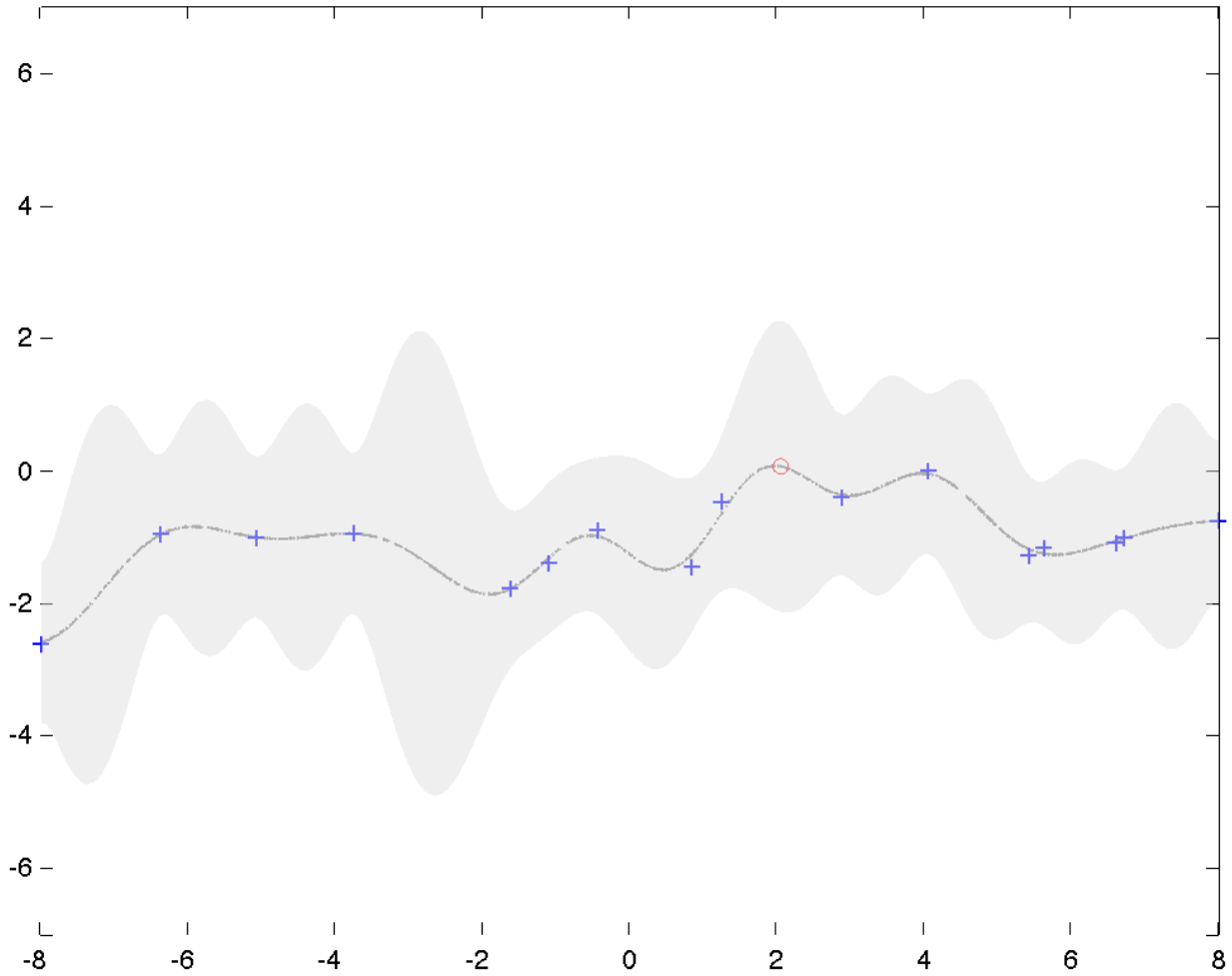
Example



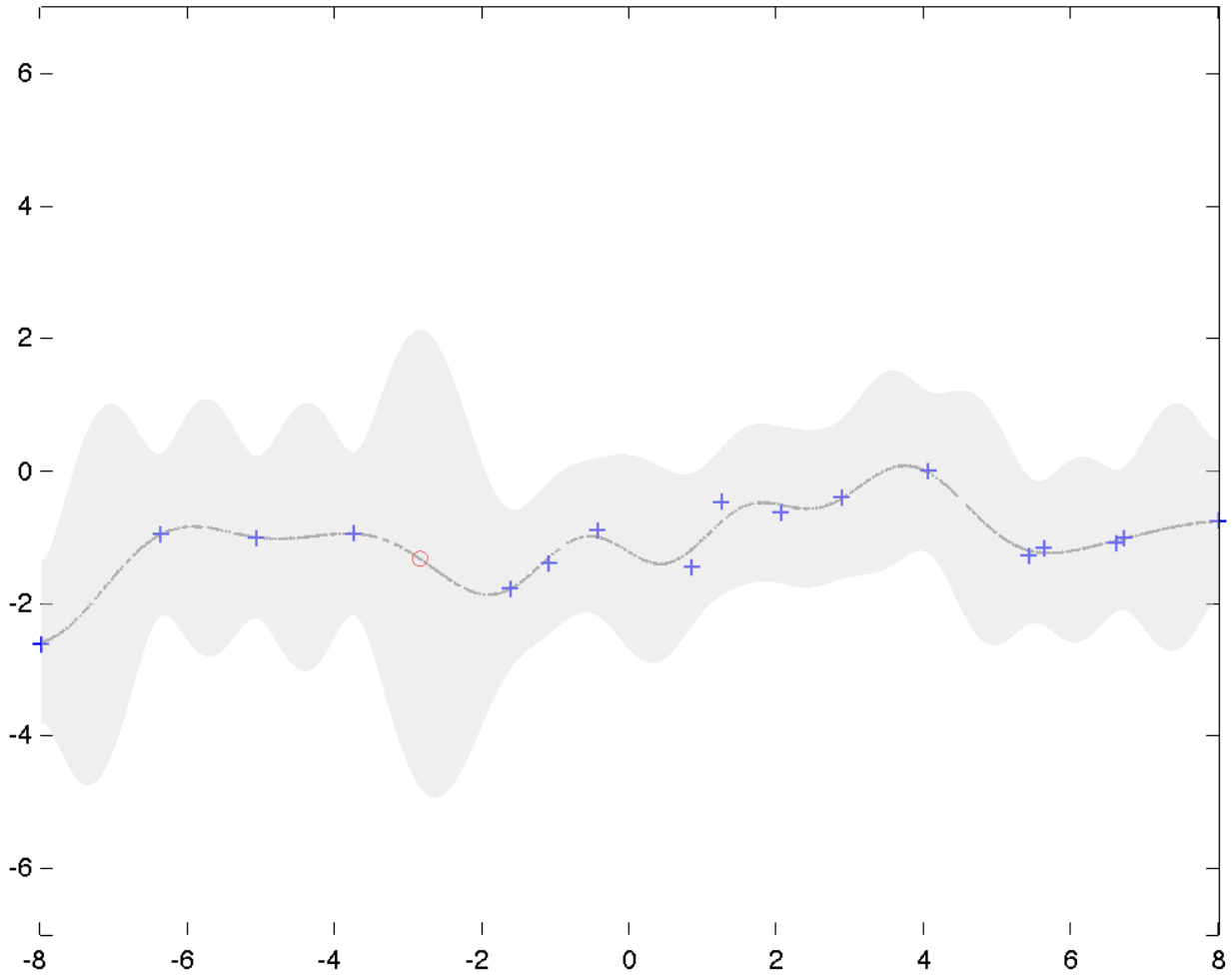
Example



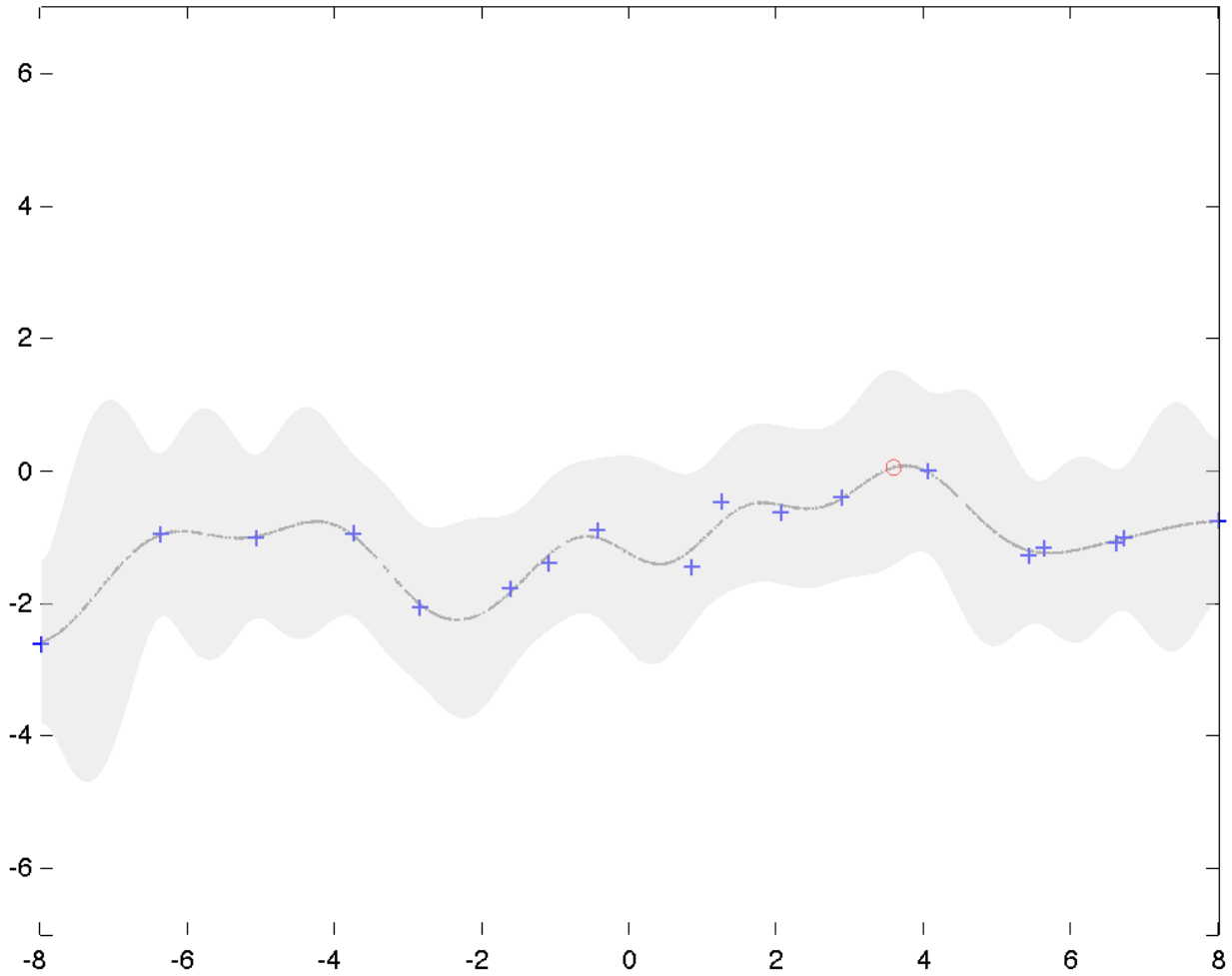
Example



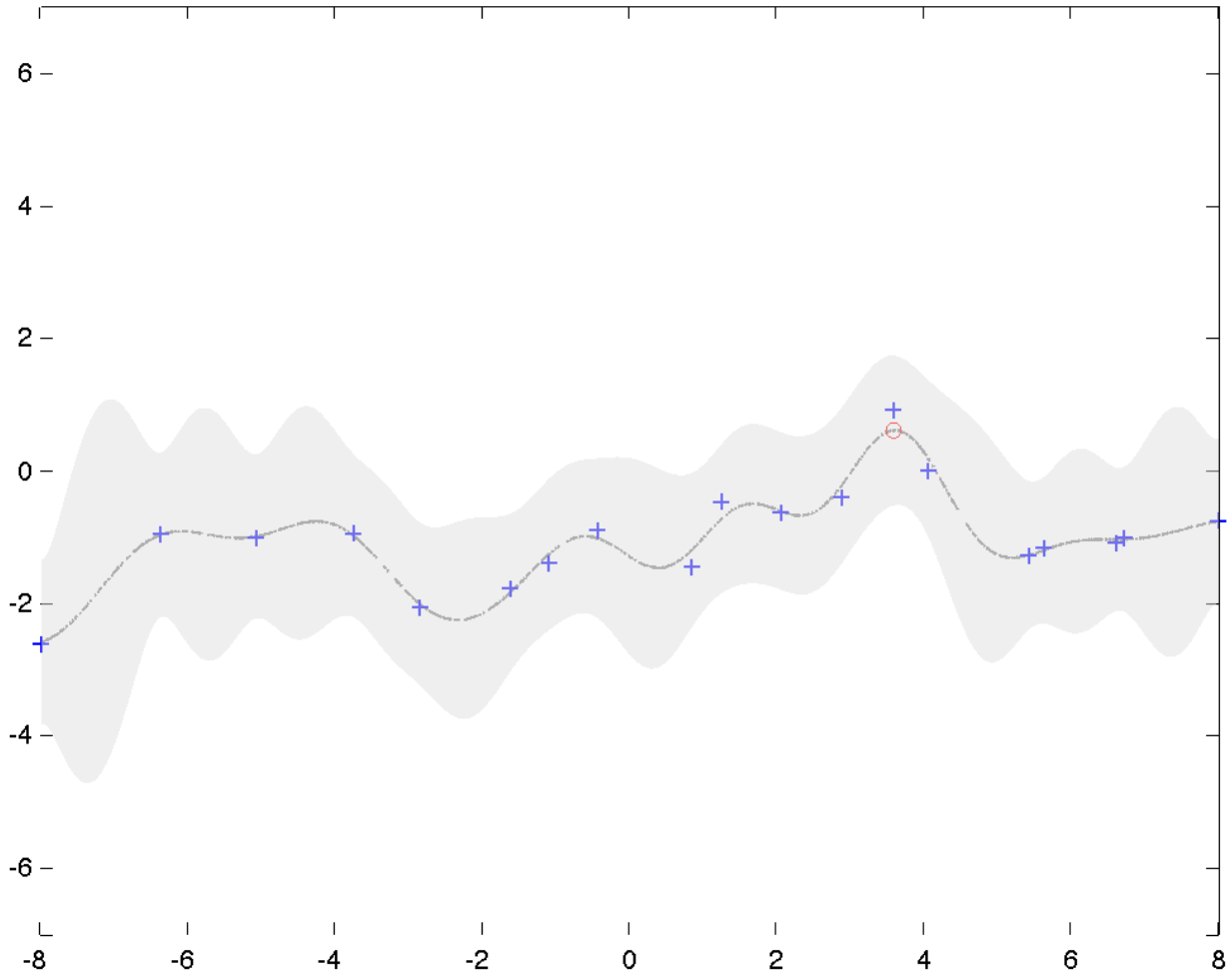
Example



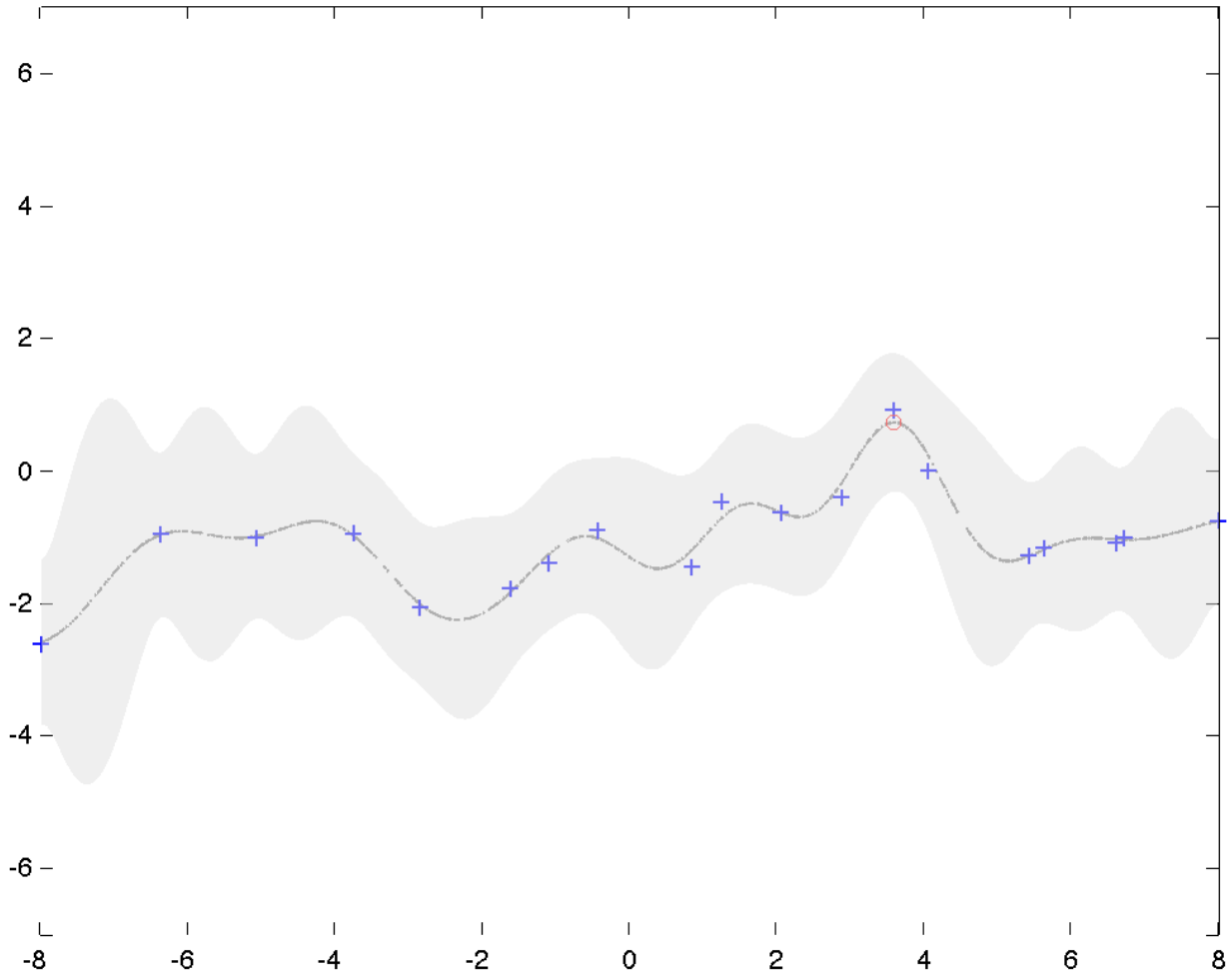
Example



Example



Example



Regret Bounds

Theorem (Srinivas et al. (2012))

$\forall \delta > 0$, *set* $\forall t \leq T,\ \beta_t = 2 \log \left(|\mathcal{X}|\, t^2 \frac{\pi^2}{6\delta} \right)$,

$$\Pr \left[R_T \leq \sqrt{C T \beta_T \gamma_T} \right] \geq 1 - \delta$$

where $C = \frac{8}{1+\sigma^{-2}}$.

Information gain (1/2)

- $H(X)$ information entropy
- $I(X) = H(\mathbf{Y}_X) - H(\mathbf{Y}_X \mid f)$
- $\gamma_T = \max_{X \subset \mathcal{X}, |X| = T} I(X)$

Regret Bounds

Theorem (Srinivas et al. (2012))

$\forall \delta > 0$, *set* $\forall t \leq T,\ \beta_t = 2 \log\left(|\mathcal{X}|\, t^2 \frac{\pi^2}{6\delta} \right)$,

$$\Pr\left[R_T \leq \sqrt{C T \beta_T \gamma_T} \right] \geq 1 - \delta$$

where $C = \frac{8}{1+\sigma^{-2}}$.

Information gain (2/2)

- For linear kernel, $\gamma_T \in \mathcal{O}(d \log T)$
- For RBF kernel, $\gamma_T \in \mathcal{O}\left((\log T)^{d+1}\right)$
- For Matérn kernel, $\gamma_T \in \mathcal{O}\left(T^{\alpha} \log T\right)$, with $\alpha = \frac{d(d+1)}{2\nu + d(d+1)} \leq 1$

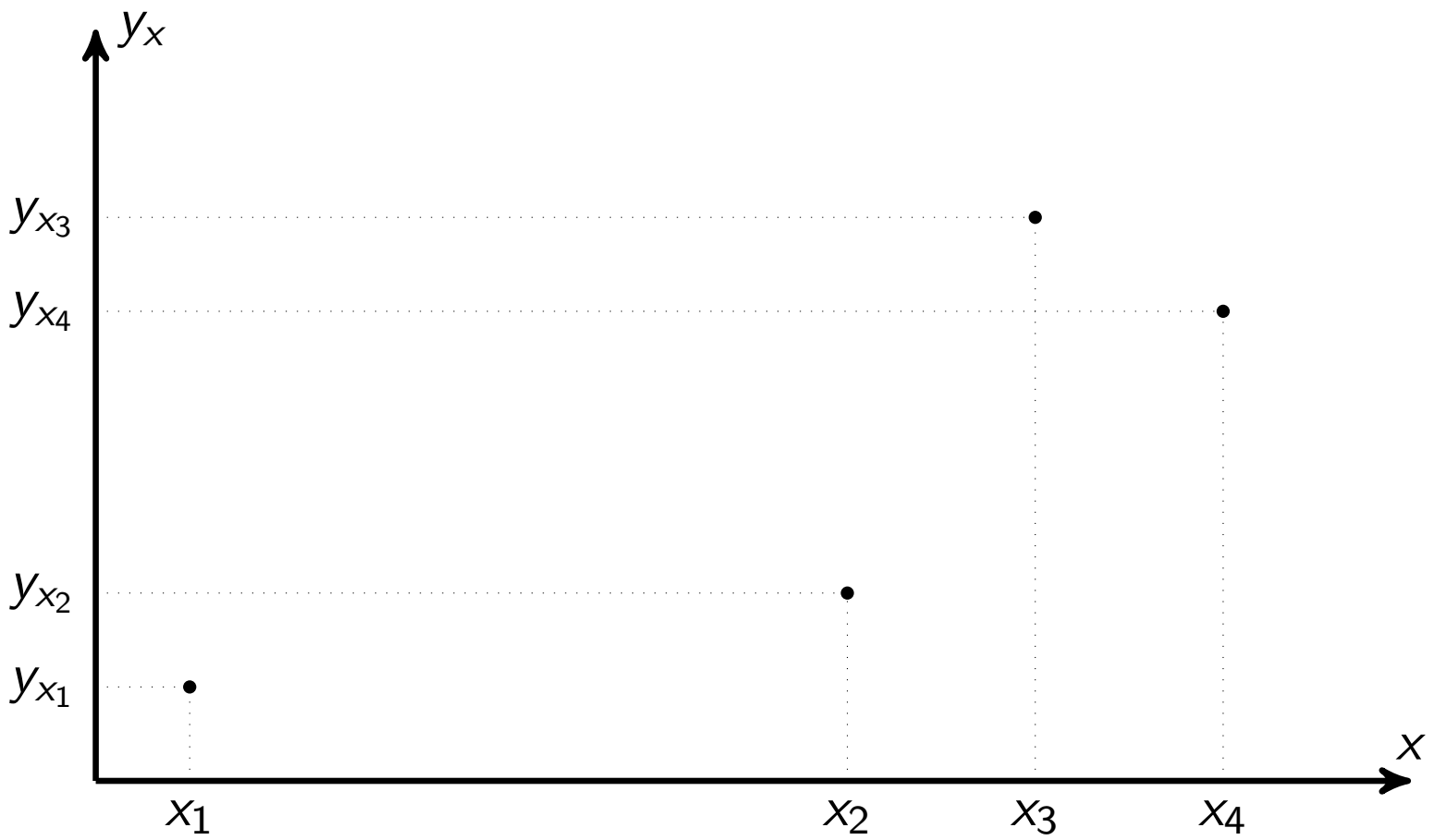
Introduction

Gaussian Processes

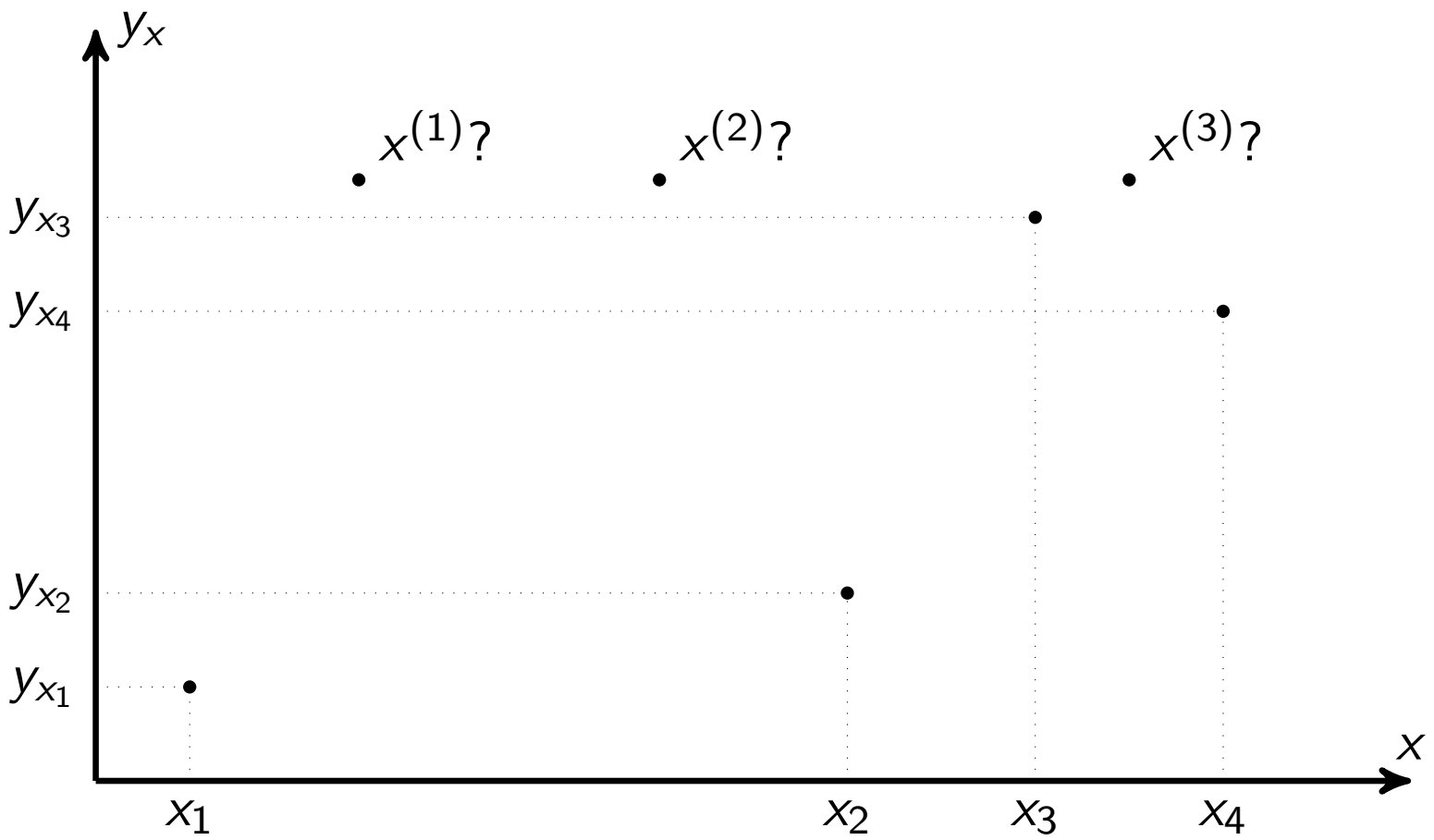
GP-UCB

Parallelism

Motivating example



Motivating example



Setup

Notation

At iteration T, select a batch of K queries $X_T^K = \left\{x_T^{(1)}, \ldots, x_T^{(K)}\right\}$

Complexity

Finding the K points X_T^K that maximizes $I(X_T^K)$ is NP-hard.

Heuristic

Due to the submodularity of I, the greedy strategy is a good approximation (Guestrin et al., 2005).

Relevant Region

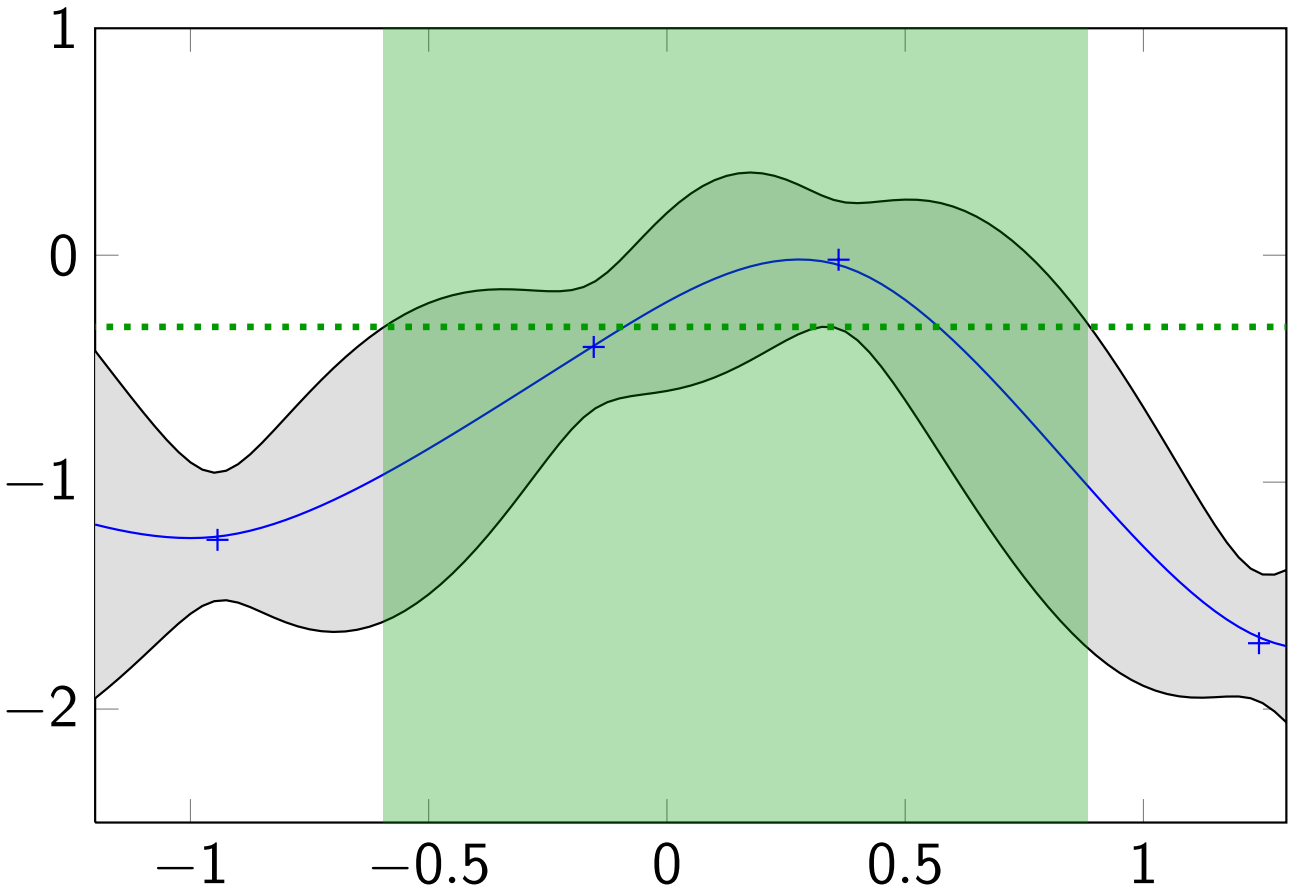
Definition

$$x_t^\bullet = \underset{x \in \mathcal{X}}{\operatorname{argmax}}\, f_t^-(x)$$

$$y_t^\bullet = f_t^-(x_t^\bullet)$$

$$\mathfrak{R}_t = \left\{ x \in \mathcal{X} \mid f_t^+(x) \geq y_t^\bullet \right\}$$

Example



GP-UCB-PE (Contal et al., 2013)

Algorithm 2: GP-UCB-PE

for $t = 0, 1, \ldots$ **do**

  Compute μ_t and σ_t^2 (Eq. 1, 2) with $y_1, \ldots, y_{t-1}$

  $x_t^0 \leftarrow \text{argmax}_{x \in \mathcal{X}} \widehat{f}_t^+(x)$

  Compute $\mathfrak{R}_t^+$

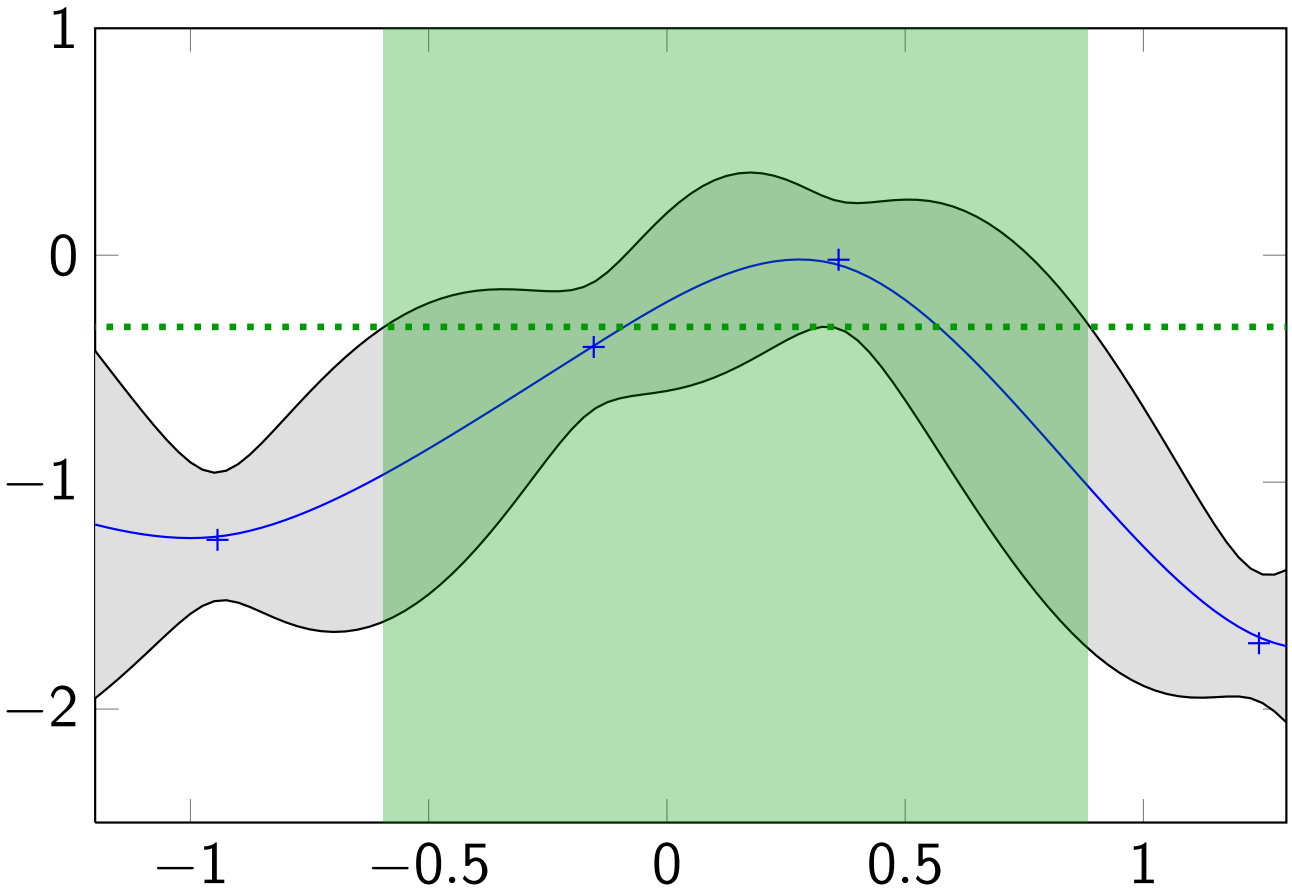
  for $k = 1, \ldots, K-1$ **do**

    Compute $\widehat{\sigma}_t^{(k)}$ (Eq.2)

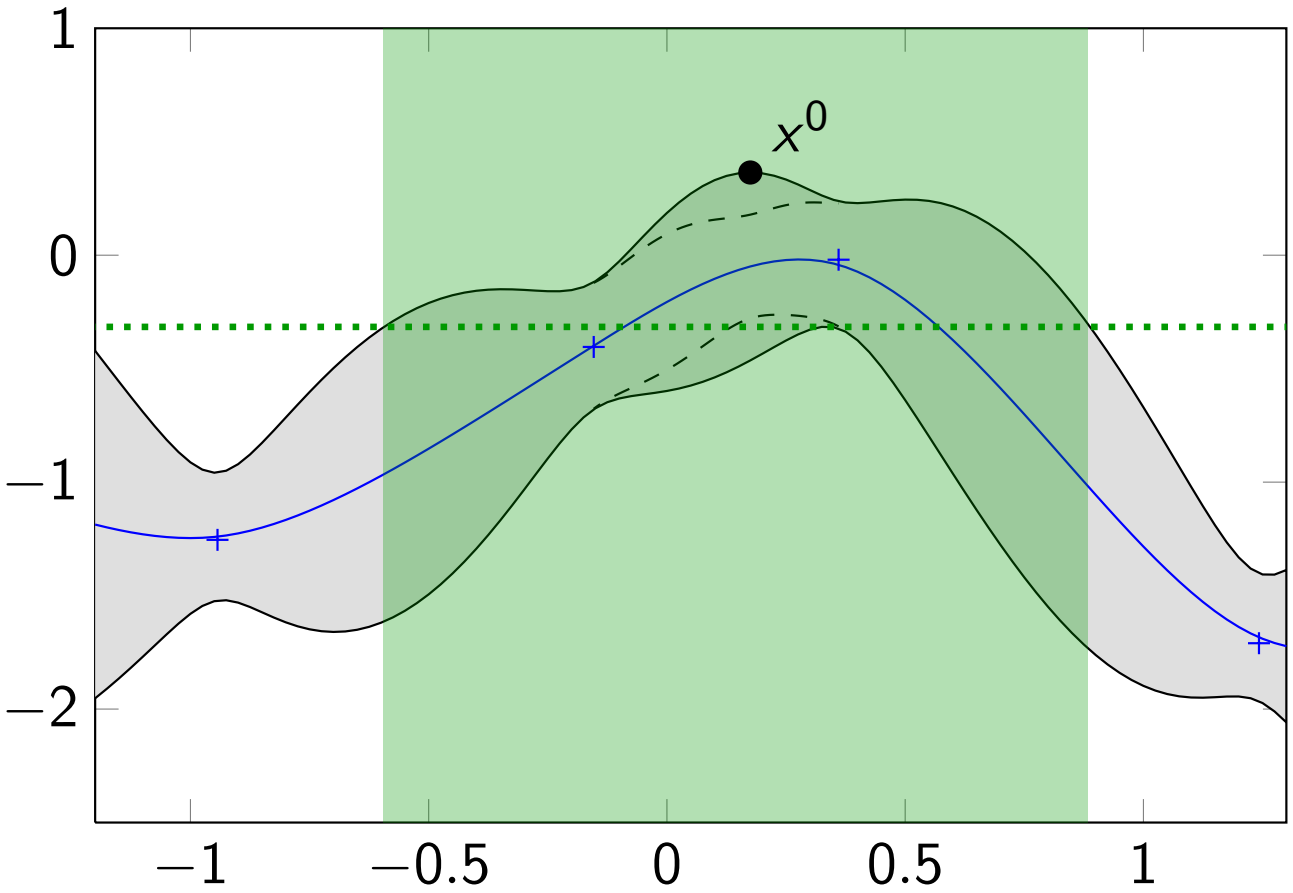
    $x_t^k \leftarrow \text{argmax}_{x \in \mathfrak{R}_t^+} \widehat{\sigma}_t^{(k)}(x)$

  Query $\{x_t^k\}_{k<K}$

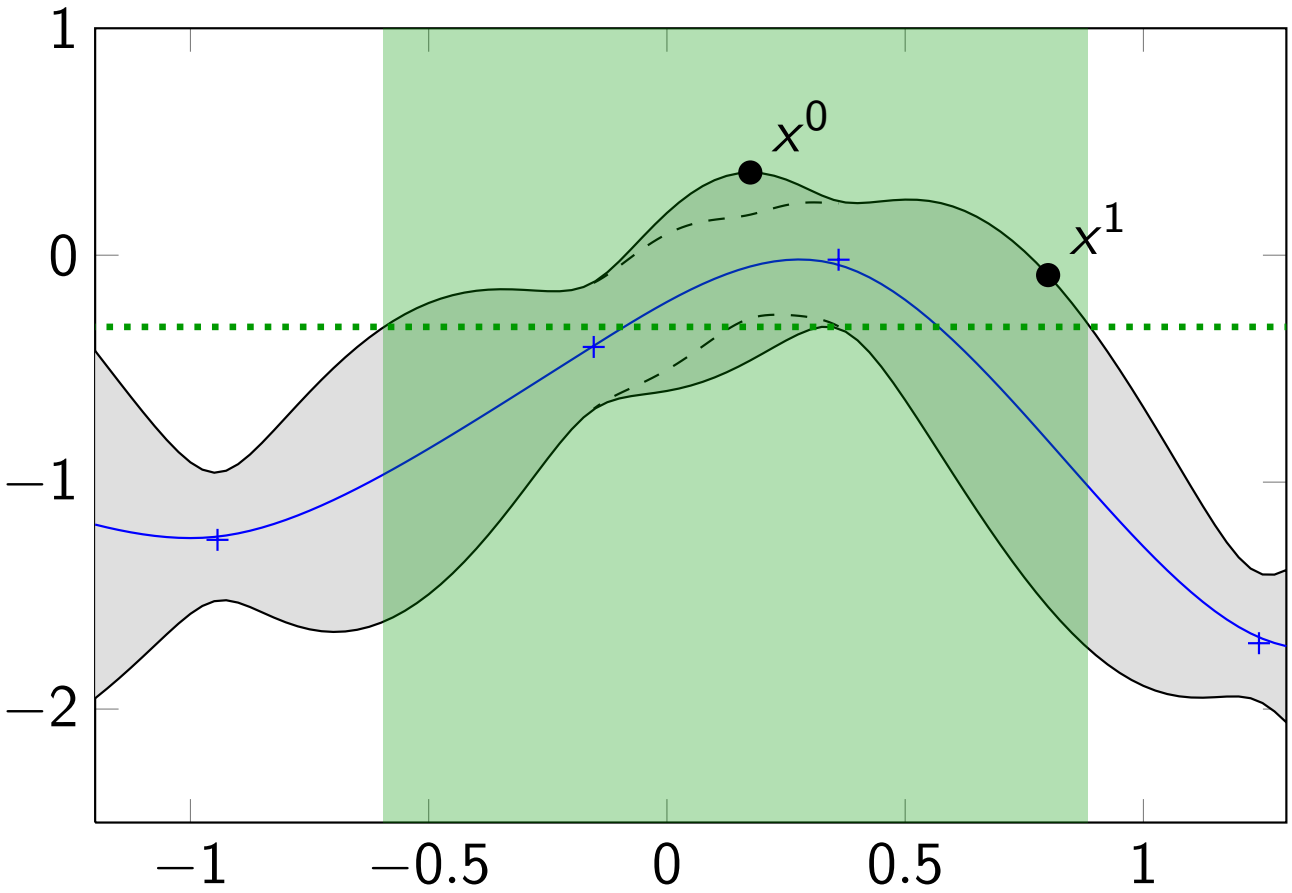
Example



Example



Example



Regret Bounds

Theorem (Contal et al. (2013))

$\forall \delta > 0$, β_t defined as above,

$$\Pr\left[R_{TK} \leq \sqrt{C_1 TK \beta_T \gamma_{TK} + C_2}\right] \geq 1 - \delta$$

where $C_1 = \frac{36}{\log(1+\sigma^{-2})}$, and $C_2 = \frac{\pi}{\sqrt{6}}$.

Corollary

When the cost for a batch is fixed and $K \ll T$, the improvement of the parallel strategy over the sequential one is $\sqrt{K}$.

Contal, E., Buffoni, D., Robicquet, A., and Vayatis, N. (2013). Parallel gaussian process optimization with pure exploration. In *Submitted to ECML (pending approval)*.

Guestrin, C., Krause, A., and Singh, A. (2005). Near-optimal sensor placements in Gaussian processes. In *Proceedings of ICML*, pages 265–272. ACM.

Rasmussen, C. E. and Williams, C. (2005). *Gaussian Processes for Machine Learning*. MIT Press.

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