Bayesian optimization

Black-box optimization

$$f(x) \to \max_x$$

Gradient is known:

- Gradient descent with restarts

Gradient is unknown:

- Numerically estimate gradient
- Grid search / random search

Black-box optimization

$$f(x) \to \max_x$$

- x geographic coordinates, $f(x)$ — amount of oil,
  1 sample = $1,000,000
- x hyperparameters of NN, $f(x)$ — objective function,
  1 sample = 10 hours
- x drug, $f(x)$ — effectiveness against disease,
  1 sample = 2 months, $10,000, life of a rat

Black-box optimization

$$f(x) \to \max_x$$

Goal: Optimize with minimum number of trials

Black-box optimization

$$f(x) \to \max_x$$

Goal: Optimize with minimum number of trials

Surrogate model: $\widehat{f} \approx f$

- Approximates true function
- Cheap to evaluate

Black-box optimization

$$f(x) \to \max_x$$

Goal: Optimize with minimum number of trials

Surrogate model: $\widehat{f} \approx f$

- Approximates true function
- Cheap to evaluate

Acquisition function: $\mu(x)$

- Estimates profit for optimization
- Uses surrogate model

Surrogate model

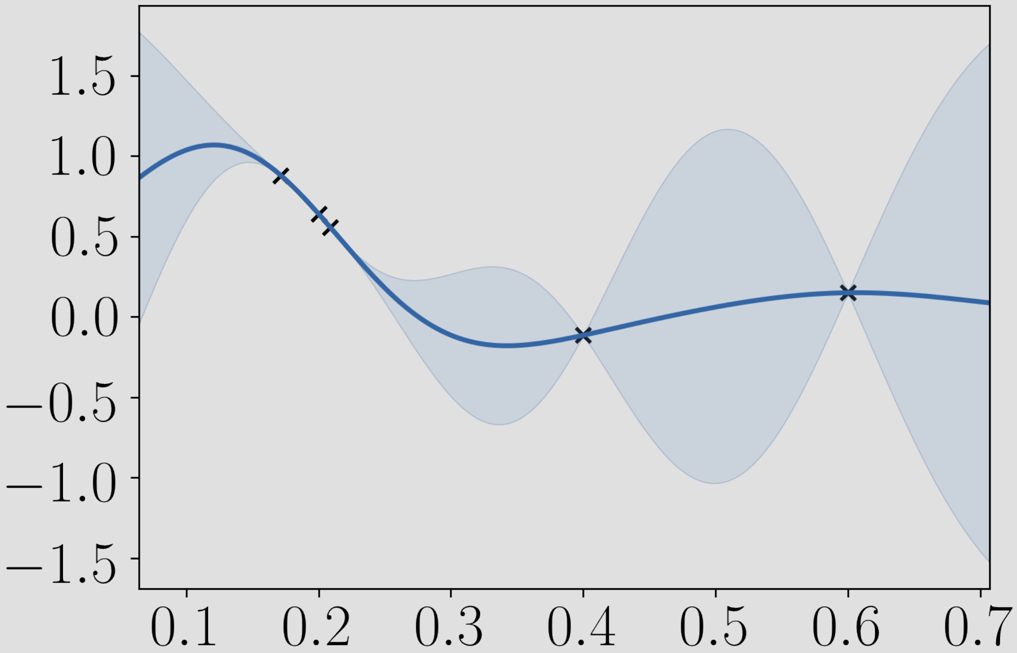
$$\widehat{f} \approx f$$

Should model arbitrary complex functions

$\Rightarrow$ Nonparametric method

Profitable to estimate uncertainty

$\Rightarrow$ Gaussian process



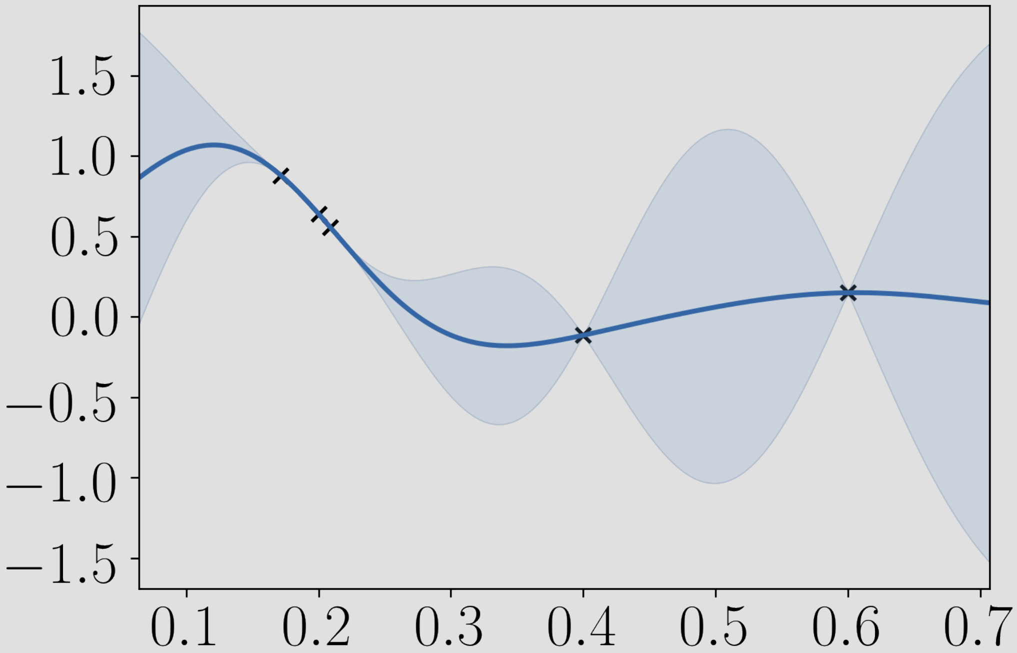
Acquisition function

Exploration:

Search in regions with high uncertainty

Exploitation:

Search in regions with high estimated value

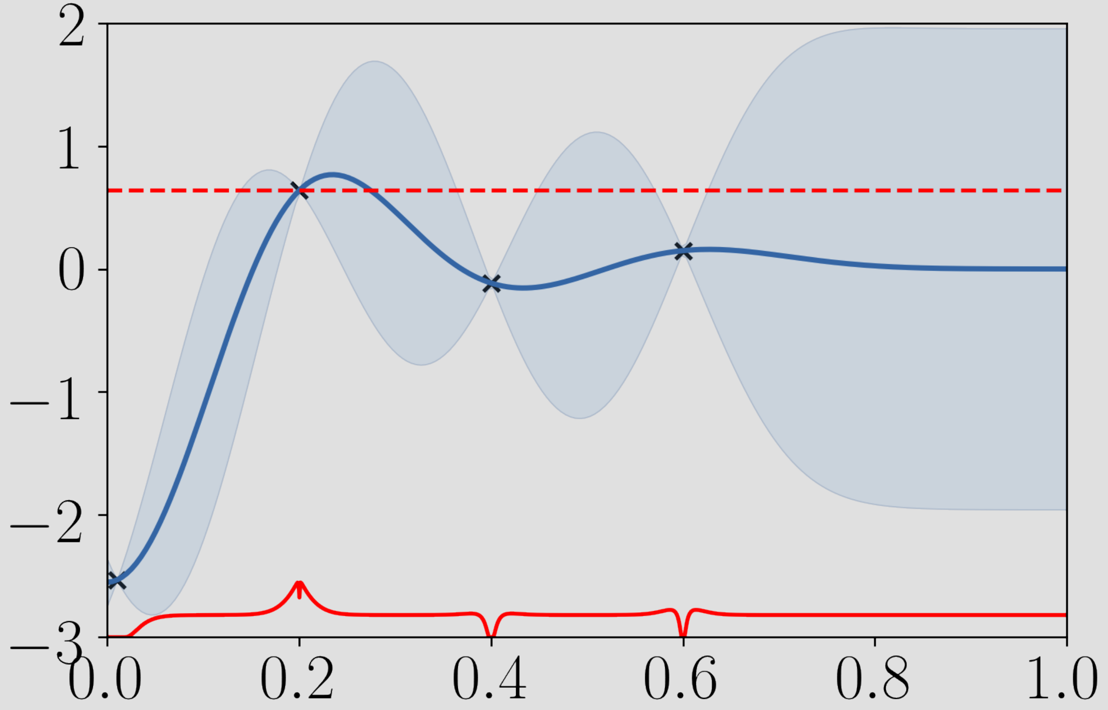


Maximum probability of improvement (MPI)

Current best value : f^*

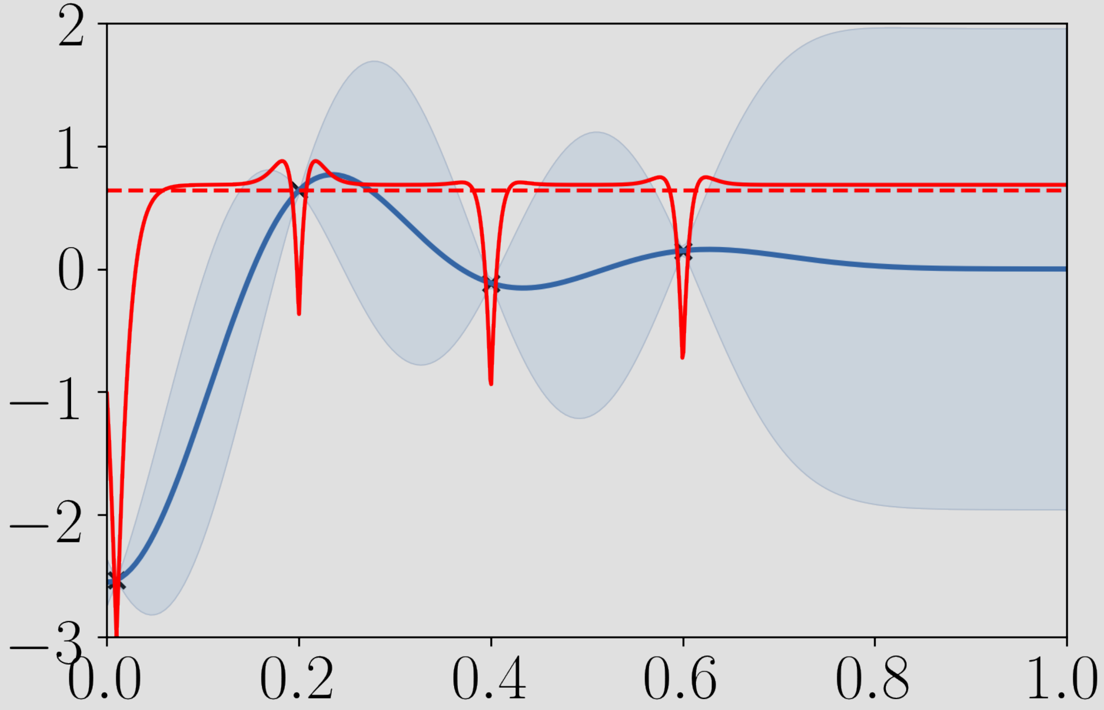
$$\mu(x) = P(\widehat{f}(x) \geq f^* + \epsilon) = \Phi\left(\frac{\mathbb{E}\widehat{f}(x) - f^* - \epsilon}{\mathrm{Var}[\widehat{f}(x)]}\right)$$

Works well if value of maximum is known



Upper confidence bound (UCB)

$$\mu(x) = \mathbb{E}\widehat{f}(x) + \eta \mathrm{Var}[\widehat{f}(x)]$$

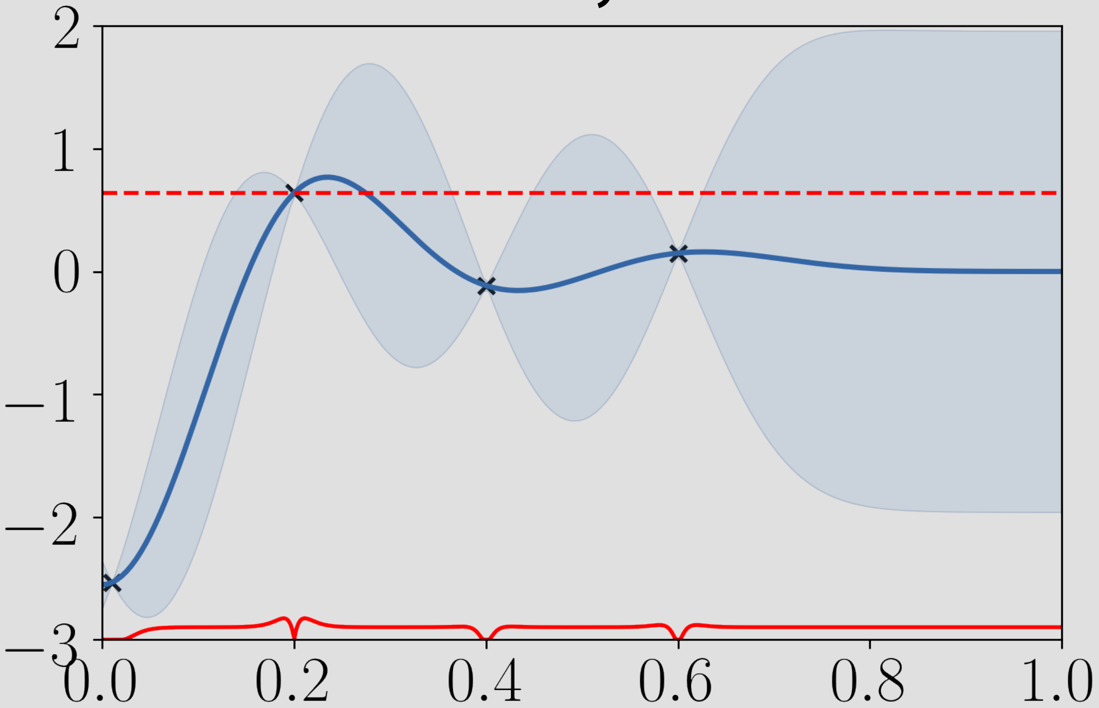


Expected improvement (EI)

$$\mu(x) = \mathbb{E}\max(f(x) - f^*, 0) = \mathrm{Var}[\widehat{f}(x)] \cdot [z\Phi(z) + \phi(z)]$$

$$z = \frac{\mathbb{E}\widehat{f}(x) - m(x)}{\mathrm{Var}[\widehat{f}(x)]}$$

Most widely used

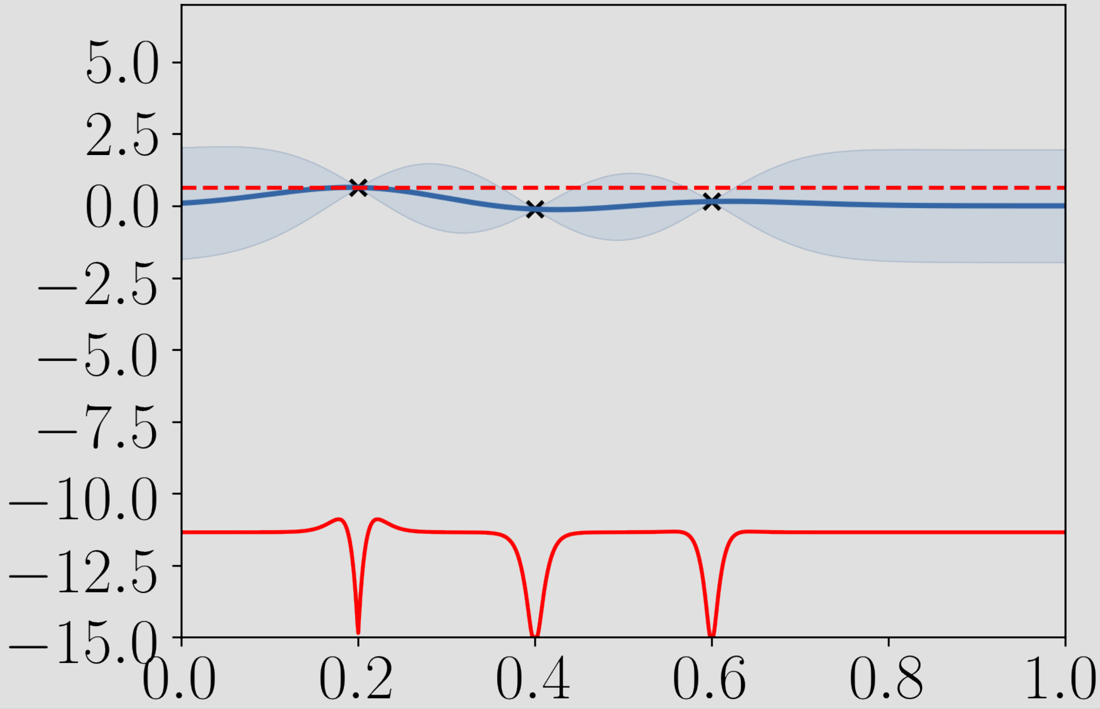


Example

Start with few points

While **not converged**:

1. Train GP
2. Find maximum of $\mu(x)$ using e.g. gradient ascent
3. Evaluate function at maximum of $\mu(x)$

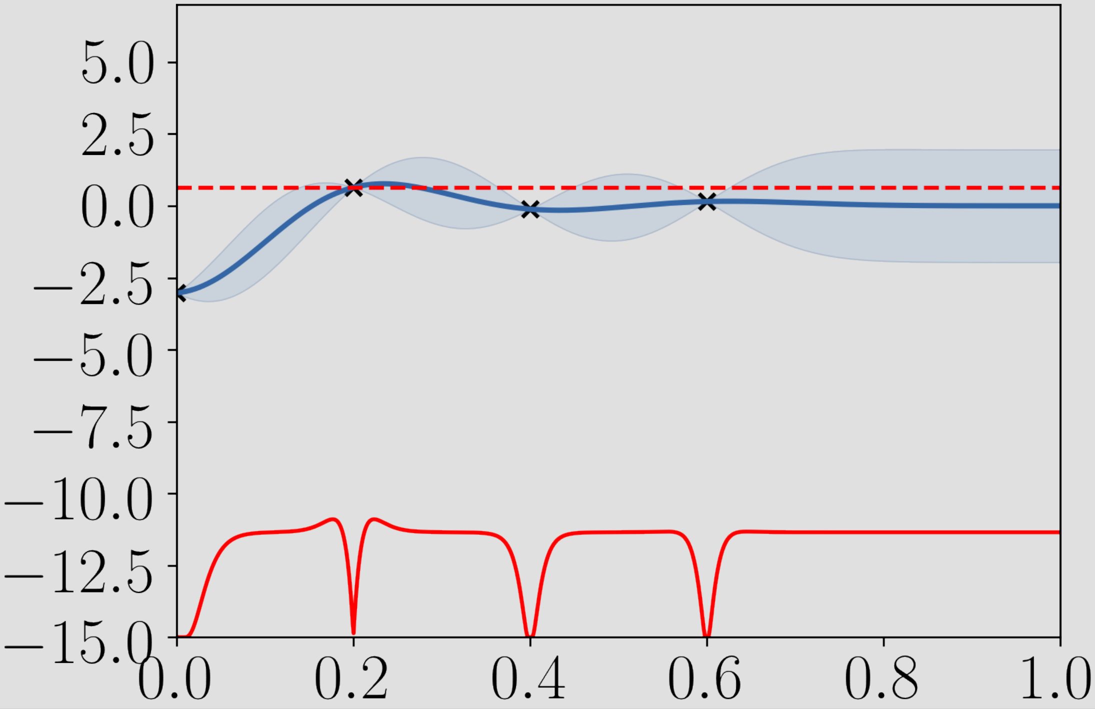


Example

Start with few points

While **not converged**:

1. Train GP
2. Find maximum of $\mu(x)$ using e.g. gradient ascent
3. Evaluate function at maximum of $\mu(x)$

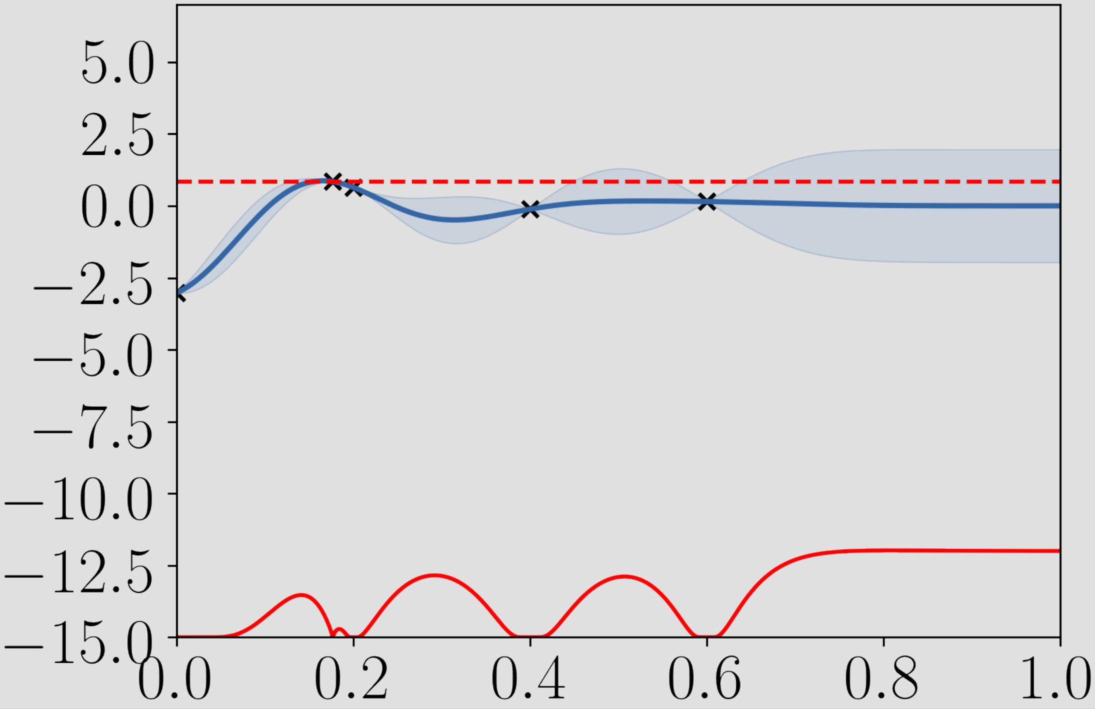


Example

Start with few points

While **not converged**:

1. Train GP
2. Find maximum of $\mu(x)$ using e.g. gradient ascent
3. Evaluate function at maximum of $\mu(x)$

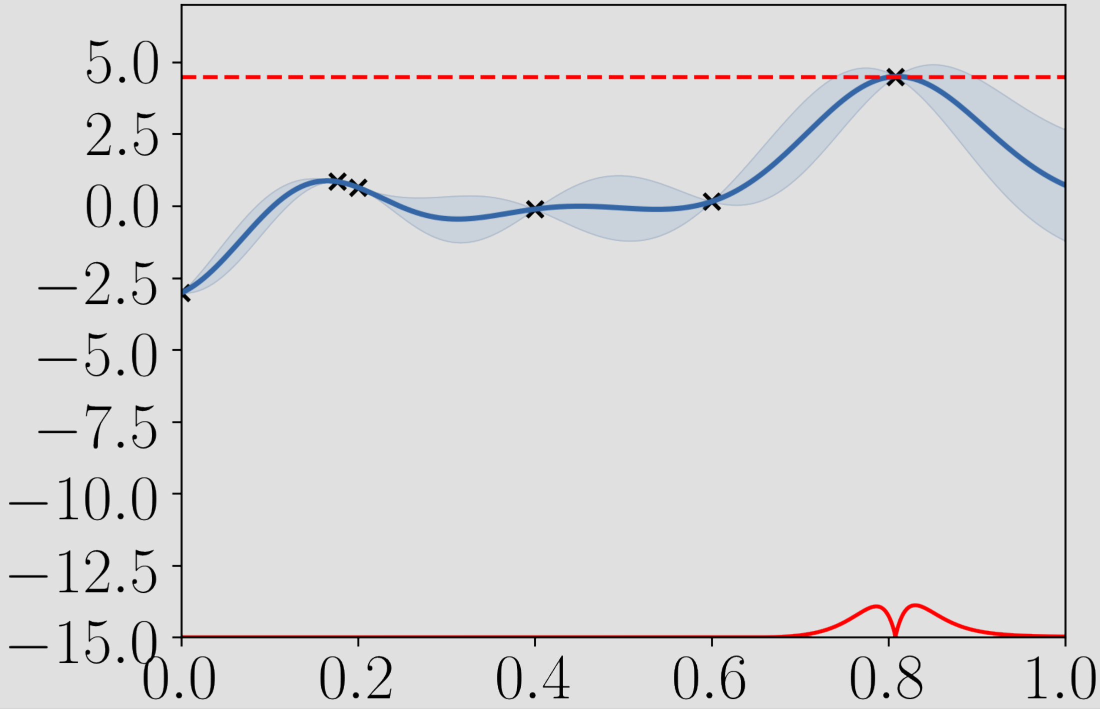


Example

Start with few points

While **not converged**:

1. Train GP
2. Find maximum of $\mu(x)$ using e.g. gradient ascent
3. Evaluate function at maximum of $\mu(x)$

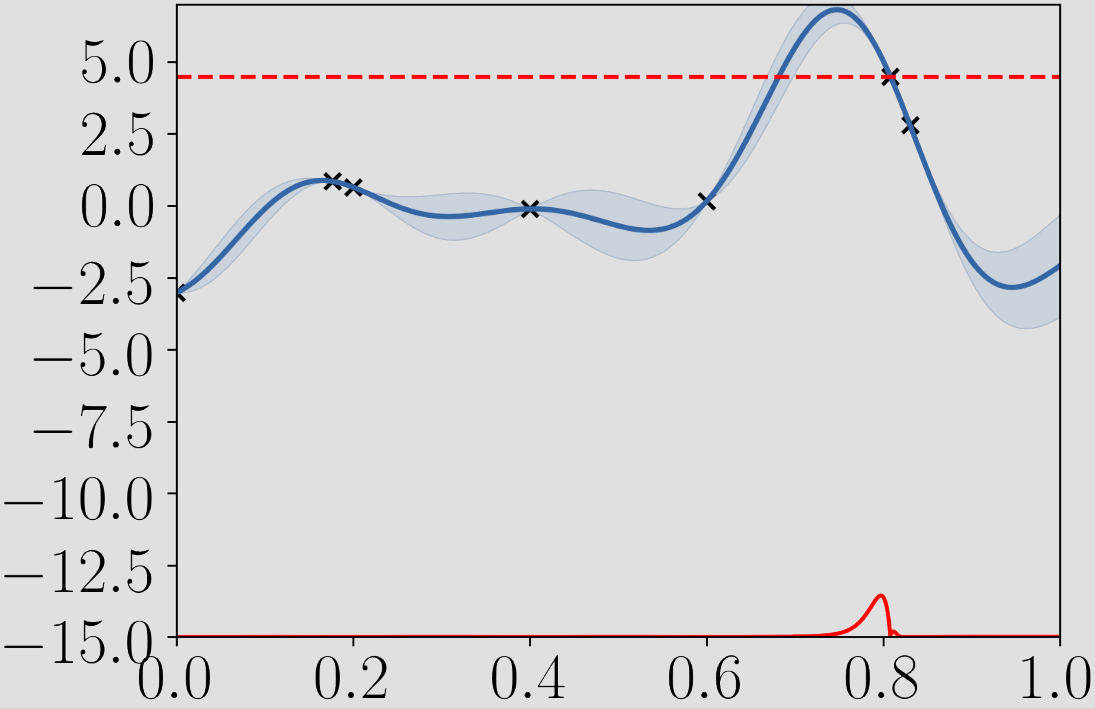


Example

Start with few points

While **not converged**:

1. Train GP
2. Find maximum of $\mu(x)$ using e.g. gradient ascent
3. Evaluate function at maximum of $\mu(x)$

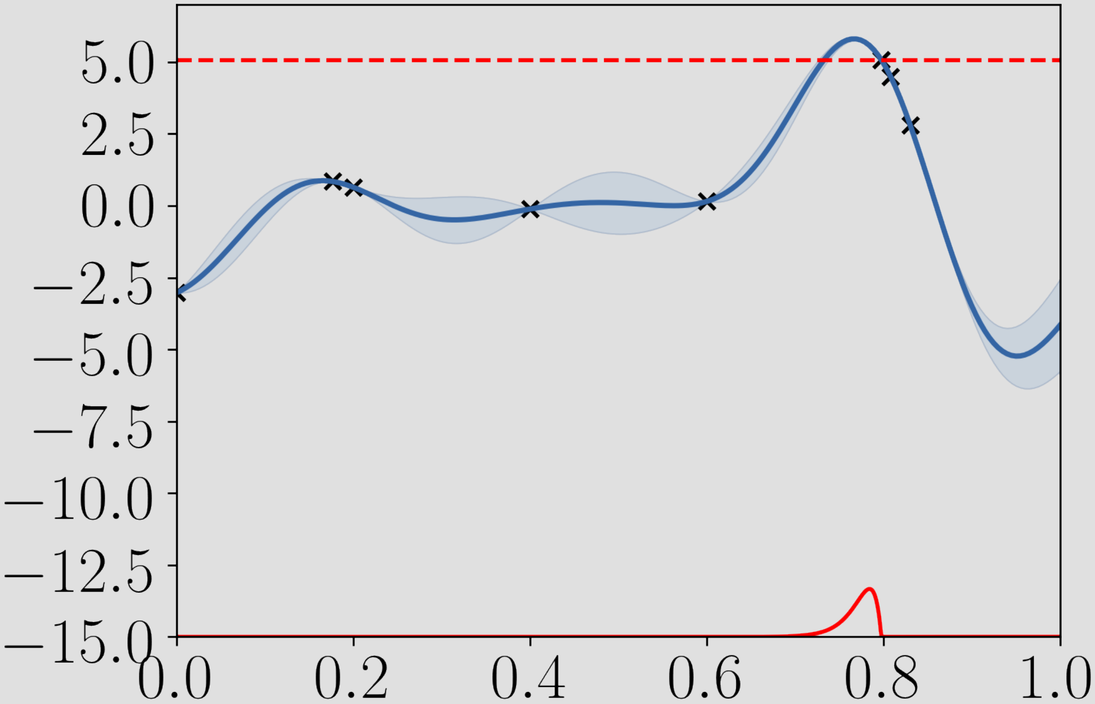


Example

Start with few points

While **not converged**:

1. Train GP
2. Find maximum of $\mu(x)$ using e.g. gradient ascent
3. Evaluate function at maximum of $\mu(x)$

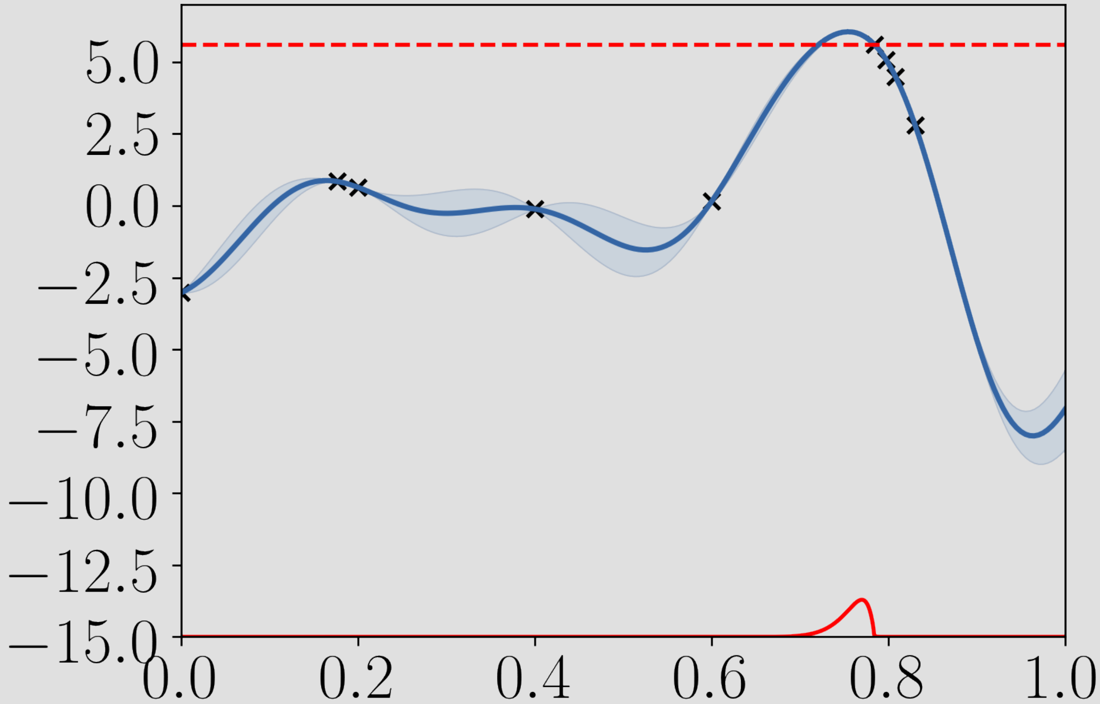


Example

Start with few points

While **not converged**:

1. Train GP
2. Find maximum of $\mu(x)$ using e.g. gradient ascent
3. Evaluate function at maximum of $\mu(x)$

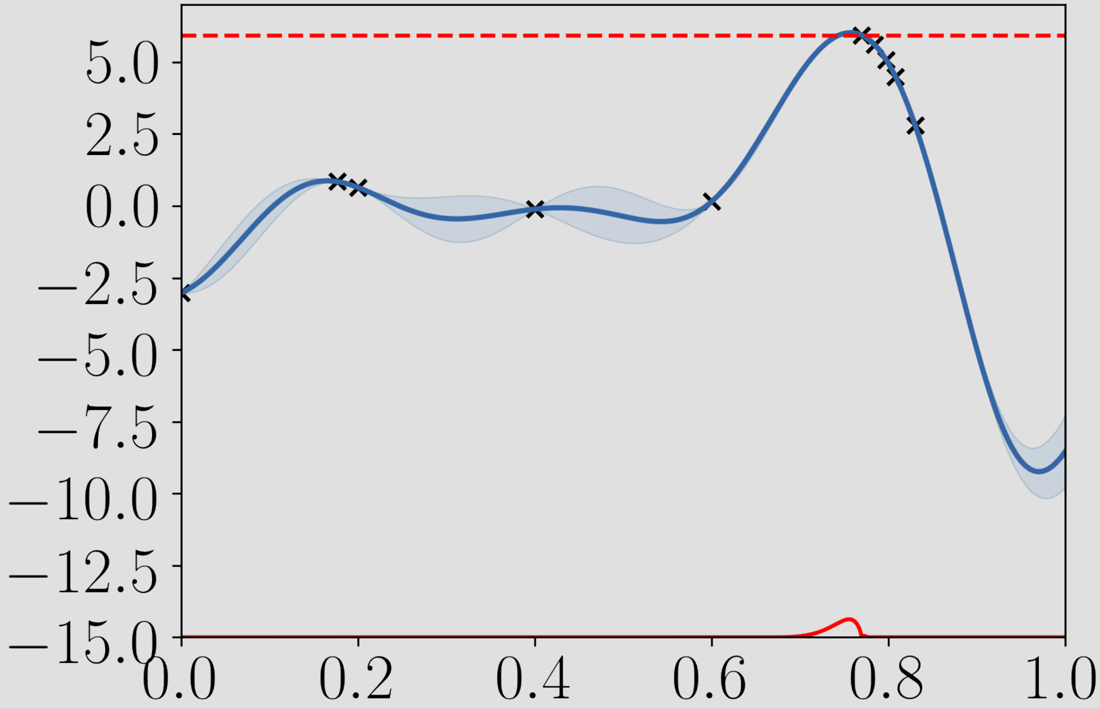


Example

Start with few points

While **not converged**:

1. Train GP
2. Find maximum of $\mu(x)$ using e.g. gradient ascent
3. Evaluate function at maximum of $\mu(x)$

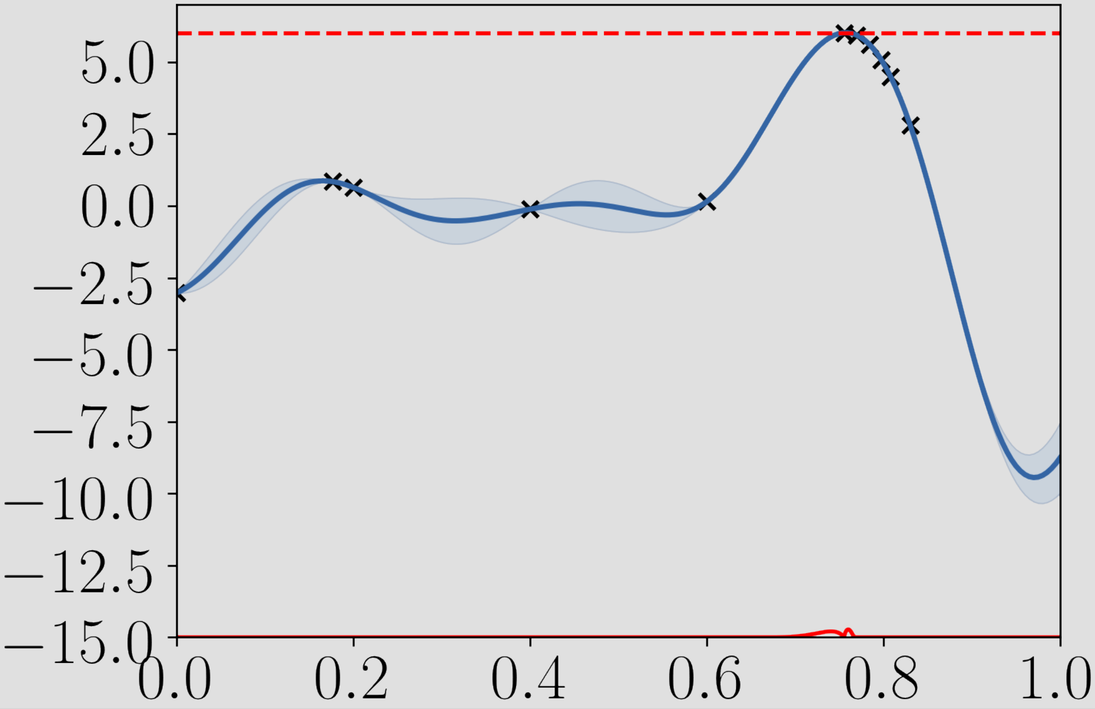


Example

Start with few points

While **not converged**:

1. Train GP
2. Find maximum of $\mu(x)$ using e.g. gradient ascent
3. Evaluate function at maximum of $\mu(x)$

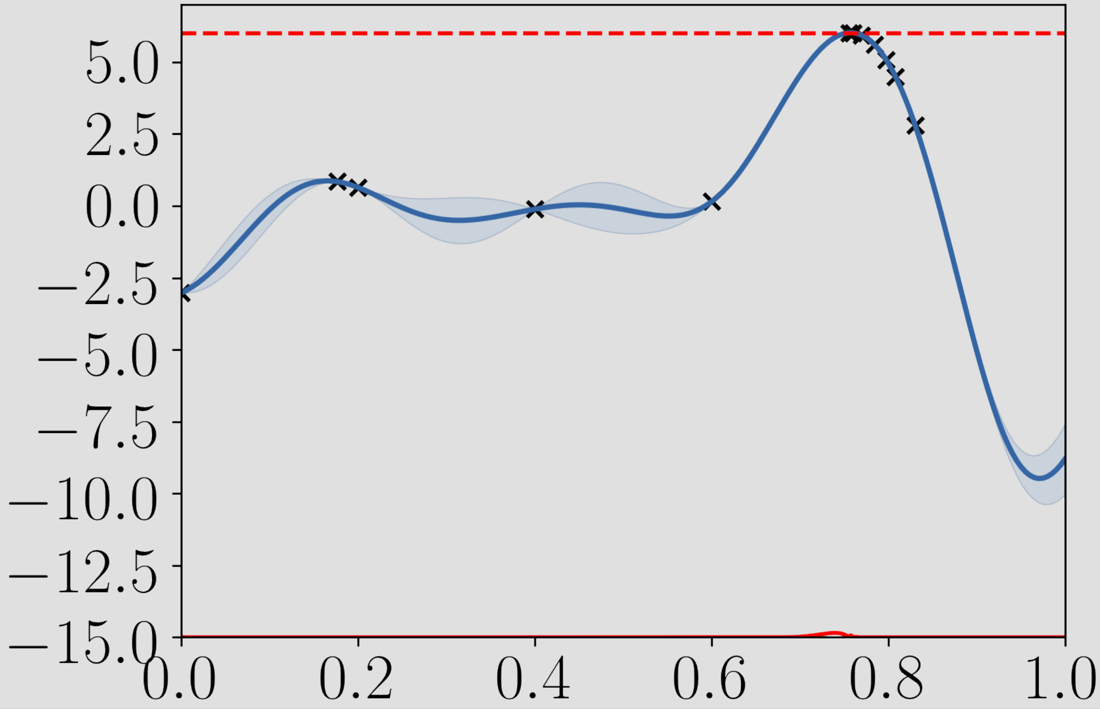


Example

Start with few points

While **not converged**:

1. Train GP
2. Find maximum of $\mu(x)$ using e.g. gradient ascent
3. Evaluate function at maximum of $\mu(x)$

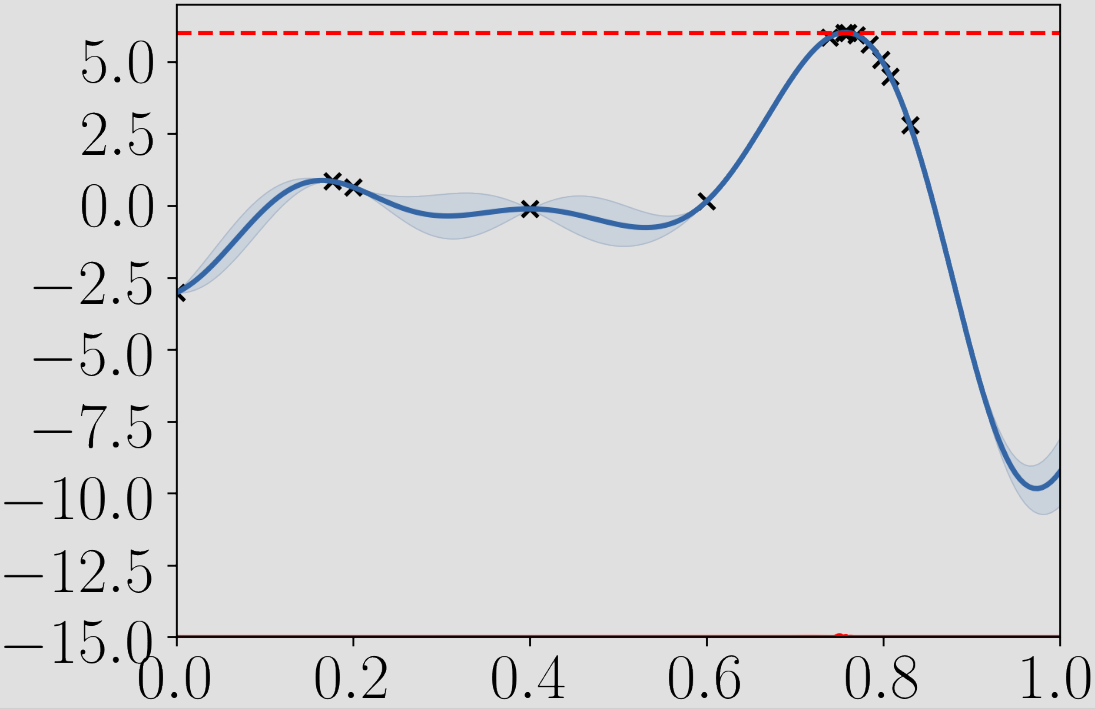


Example

Start with few points

While **not converged**:

1. Train GP
2. Find maximum of $\mu(x)$ using e.g. gradient ascent
3. Evaluate function at maximum of $\mu(x)$

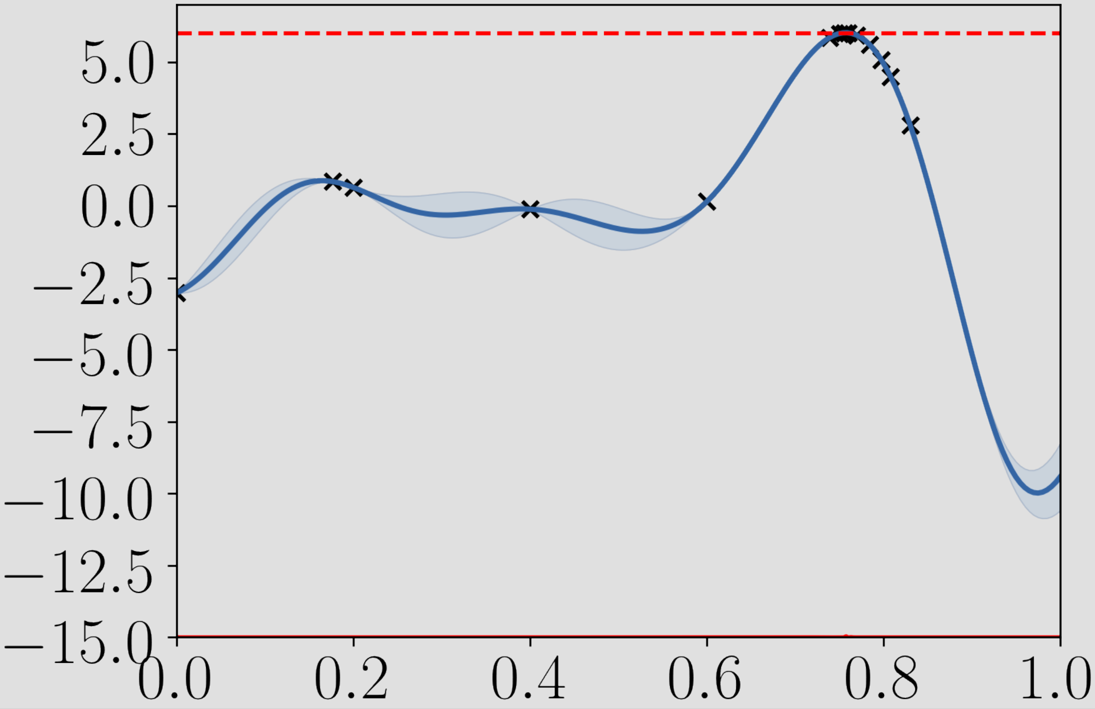


Example

Start with few points

While **not converged**:

1. Train GP
2. Find maximum of $\mu(x)$ using e.g. gradient ascent
3. Evaluate function at maximum of $\mu(x)$

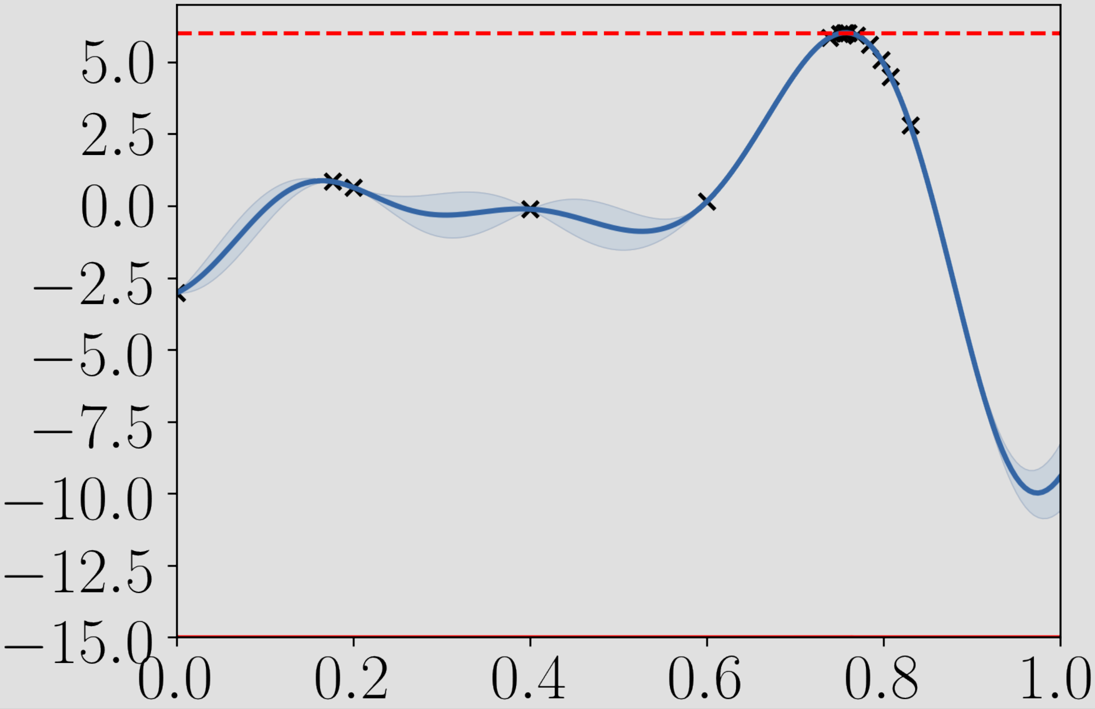


Example

Start with few points

While **not converged**:

1. Train GP
2. Find maximum of $\mu(x)$ using e.g. gradient ascent
3. Evaluate function at maximum of $\mu(x)$

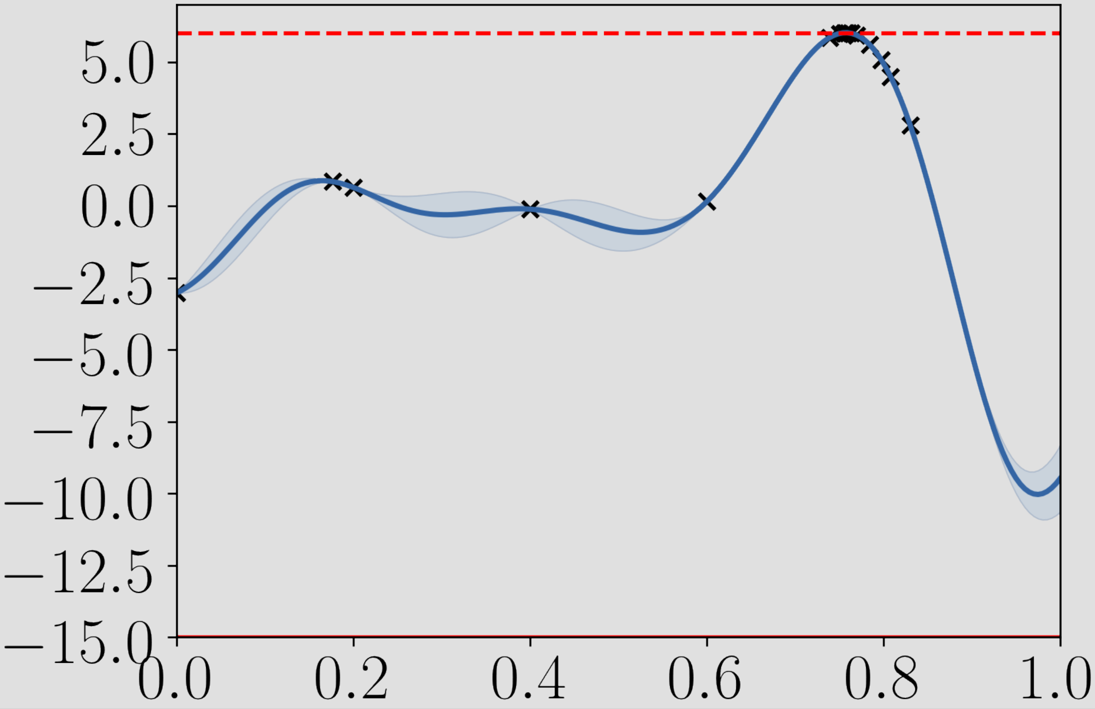


Example

Start with few points

While **not converged**:

1. Train GP
2. Find maximum of $\mu(x)$ using e.g. gradient ascent
3. Evaluate function at maximum of $\mu(x)$

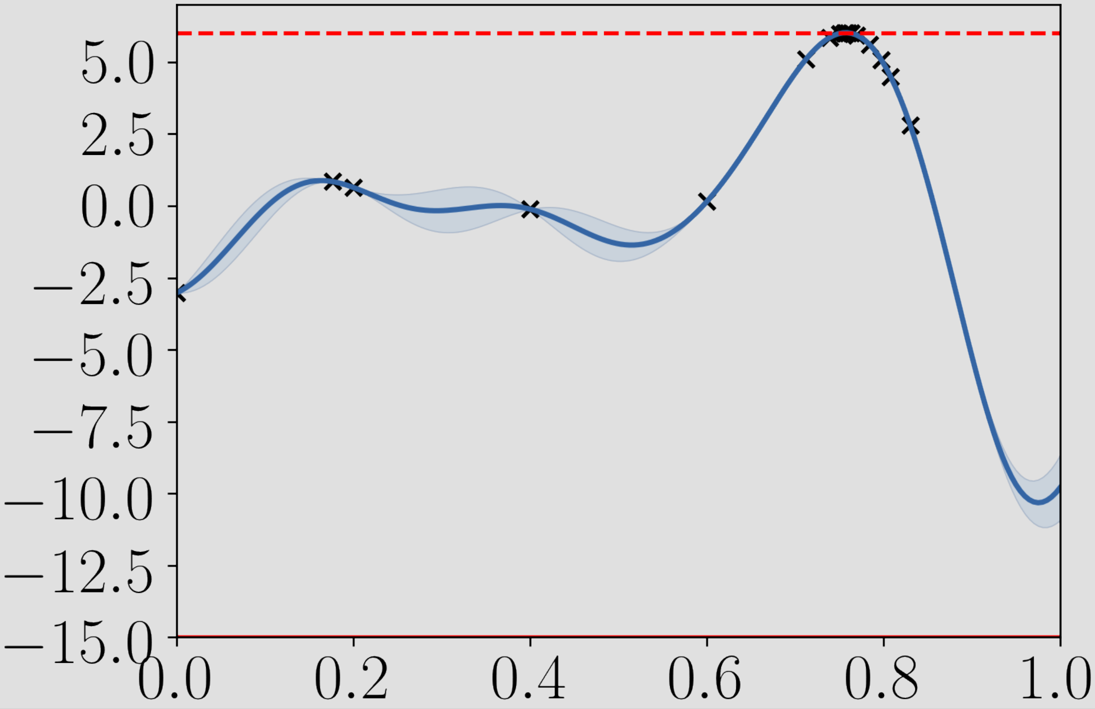


Example

Start with few points

While **not converged**:

1. Train GP
2. Find maximum of $\mu(x)$ using e.g. gradient ascent
3. Evaluate function at maximum of $\mu(x)$

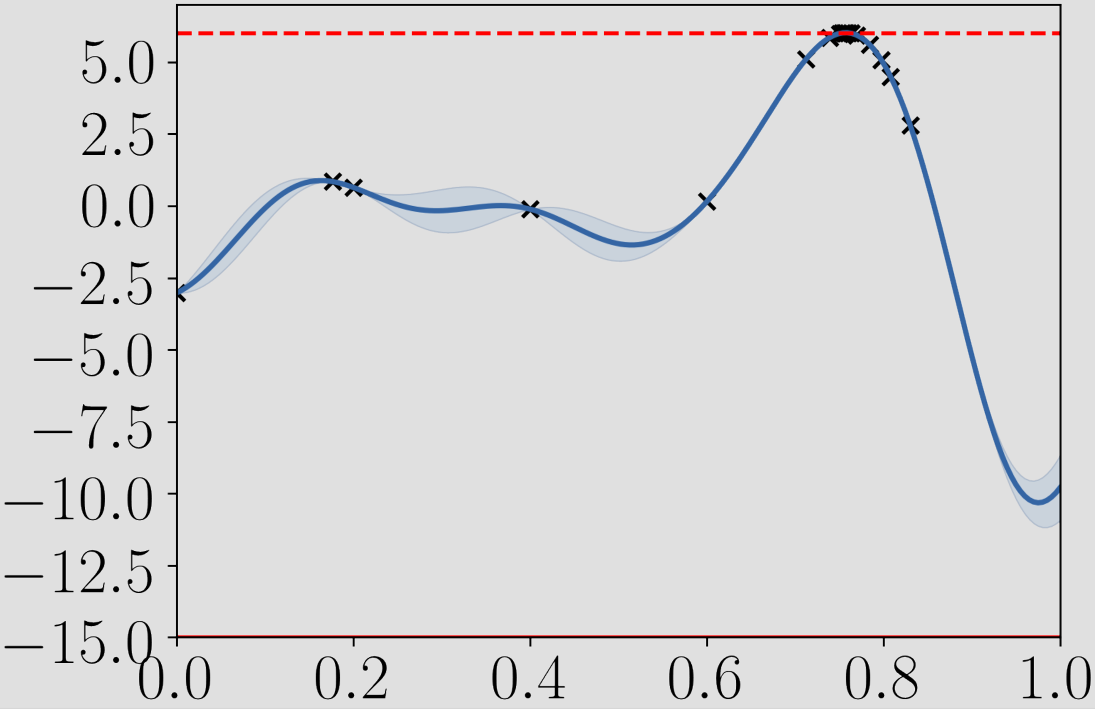


Example

Start with few points

While **not converged**:

1. Train GP
2. Find maximum of $\mu(x)$ using e.g. gradient ascent
3. Evaluate function at maximum of $\mu(x)$



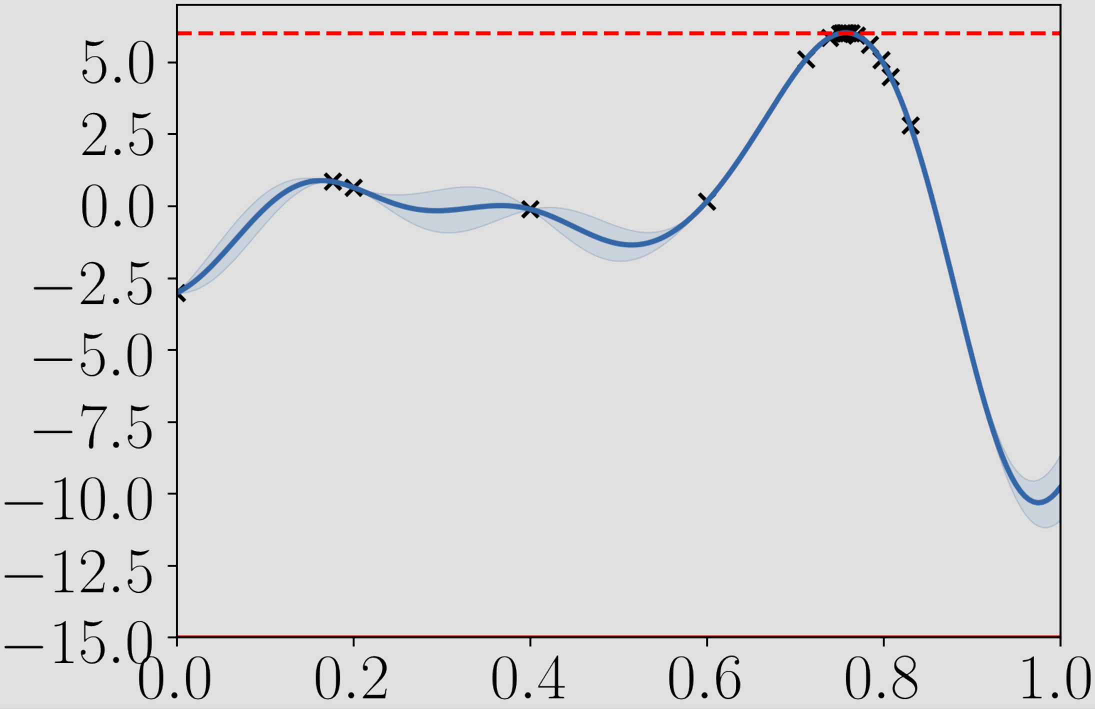
ТЕХНИЧЕСКИЙ СЛАЙД

Start with few points

While **not converged**:

1. Train GP
2. Find maximum of $\mu(x)$ using e.g. gradient ascent
3. Evaluate function at maximum of $\mu(x)$

STOP



Random search vs Gaussian processes

RS	GP
+ Parallelizable	- Hard to parallelize experiments
- Needs many more points for high dimensions	+ Requires less points on average
Any function	Each evaluation is expensive