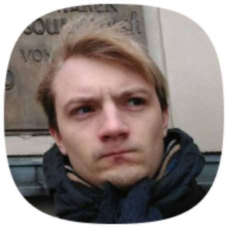
Vladislav Belavin, Maxim Borisyak



# Bayesian Optimization

combining GP and BO

2021

MLHEP 2021

HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY

LAMBDA • HSE

Yandex

SCHOOL OF DATA ANALYSIS

EPFL

Schaffhausen Institute of Technology

# Bayesian Optimization refresher

# Prerequisites

1. model: $\mathcal{M} = \{f_\theta : \mathcal{X} \to \mathbb{R} \mid \theta \in \Theta\}$;
2. prior: $P(\theta \mid \theta \in \Theta)$;
3. probability data model: $P(y \mid \theta, x)$;
4. gain/acquisition function.
5. the objective/target function $t$.

# Main loop

`for` $i = 1$ to $N$ do
    compute $P(\theta \mid X, Y)$
    search for $x_i$ with the most expected gain
    evaluate $y_i = t(x_i)$
    extend $X$ and $Y$ with $x_i$ and $y_i$
`end for`

# Posterior computation

Bayesian inference is difficult:

$$P(\theta \mid X, Y) = \frac{P(Y \mid X, \theta) P(\theta)}{P(Y \mid X)} = \frac{P(Y \mid X, \theta) P(\theta)}{\int P(Y \mid X, \theta) P(\theta) \, d\theta} = \frac{P(Y \mid X, \theta) P(\theta)}{Z}$$

$$P(y \mid X, Y) = \int P(y \mid \theta, X, Y) P(\theta \mid X, Y) \, d\theta$$

# Gaussian processes

# Role in Bayesian Optimization

```
for i = 1 to N do
    compute P(θ | X, Y)
    search for x_i with the most expected gain
    evaluate y_i = t(x_i)
    extend X and Y with x_i and y_i
end for
```

# Gaussian processes refresher

Linear Gaussian process:

- $f_w(x) = w \cdot x$;
- $w \sim \mathcal{N}(0, \Sigma)$, $\Sigma = \mathrm{diag}(\sigma_w^2)$;
- $y \mid x, w \sim \mathcal{N}(w \cdot x, \sigma_y^2)$.

# Bayesian inference on a linear model

$$P(w \mid Y, X) \propto P(Y \mid w, X)P(w) \propto$$

$$\exp\left[-\frac{1}{2\sigma_y^2}(y - X^T w)^T (y - X^T w)\right] \cdot \exp\left[-\frac{1}{2} w^T \Sigma_w^{-1} w\right] =$$

$$\exp\left[-\frac{1}{2}(w - w^*)^T A_w (w - w^*)\right]$$

where:

- $A_w = \frac{1}{\sigma_y^2} X X^T + \Sigma^{-1}$;
- $w^* = \frac{1}{\sigma_y^2} A_w^{-1} X y$.

# Bayesian inference on a linear model

To make prediction $y$ in point $x$:

$$P(y \mid Y, X, x) = \int P(y \mid w, x) P(w \mid X, y) = \mathcal{N}\left(\frac{x^T A^{-1} X Y}{\sigma_y^2}, \; x^T A^{-1} x\right)$$

- posterior distribution of model parameters is Gaussian;
- (posterior) joint distribution of any number of $y(x)$ is a Gaussian distribution.

# Basis expansion

Basis expansion:

- $x \to \phi(x)$;
- polynomial: $x \to (1, x_1, x_2, \dots, x_n, x_1 x_2, x_1, x_3, \dots, x_1^2, x_2^2, \dots, x_n^2, \dots, )$;
- Fourier: $x \to (1, \cos(2\pi x_1), \sin(2\pi x_1), \cos(2\pi x_2), \sin(2\pi x_2), \dots)$;

# Kernels

$P(y, Y, X, x)$ can be rewritten via scalar products:

$$k(x_i, x_j) = \phi^T(x_i) \cdot \phi(x_j)$$

Popular kernels:

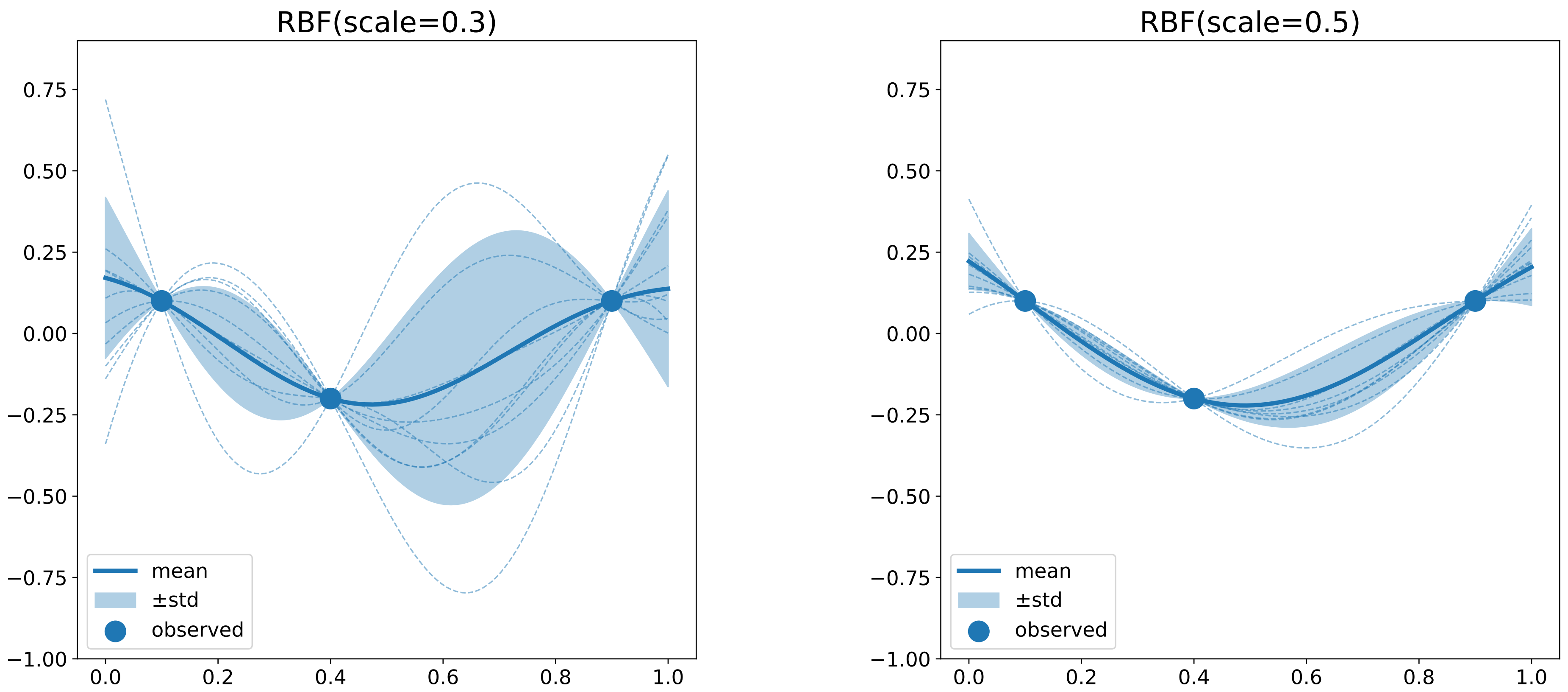
- polynomial;
- RBF:

$$\mathrm{RBF}(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right)$$

- Matern:

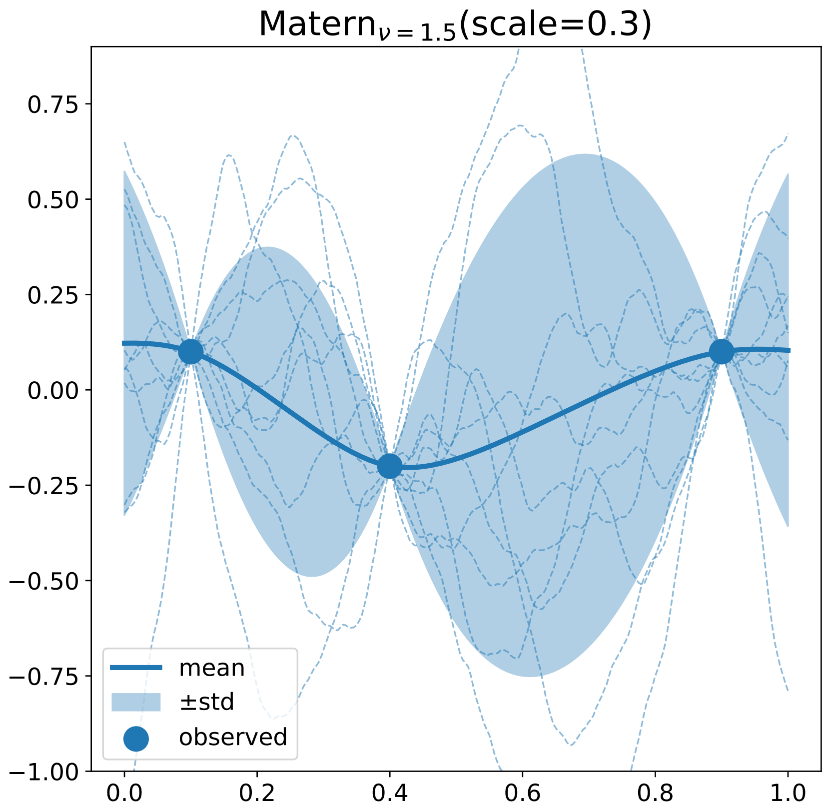
$$\mathrm{Matern}(x_i, x_j) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left( \frac{\sqrt{2\nu}}{l} \|x_i - x_j\| \right)^{\nu} K_\nu \left( \frac{\sqrt{2\nu}}{l} \|x_i - x_j\| \right)$$

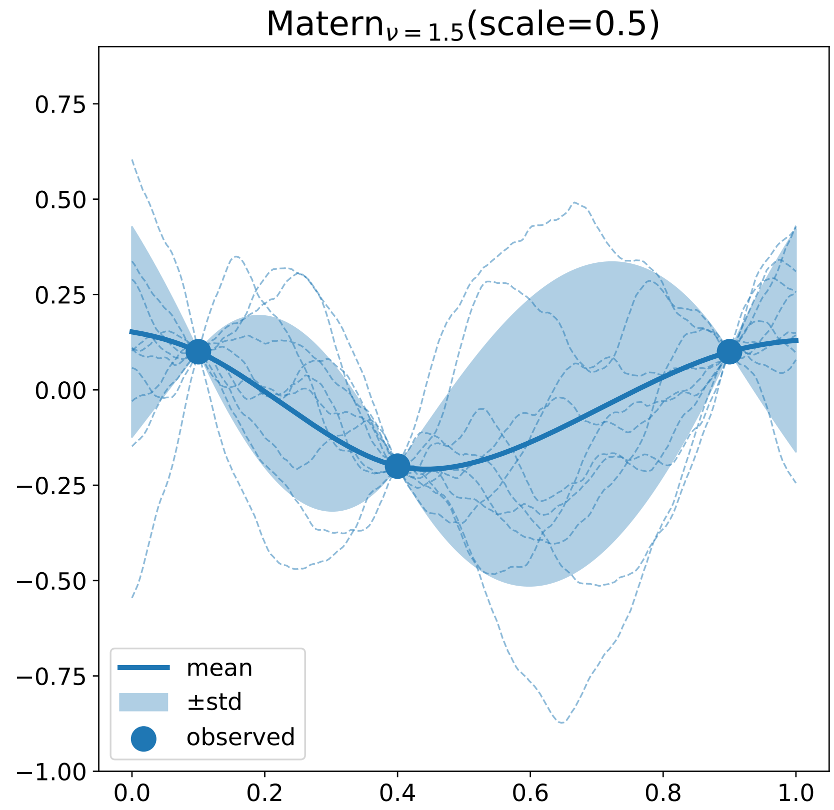
# RBF kernel



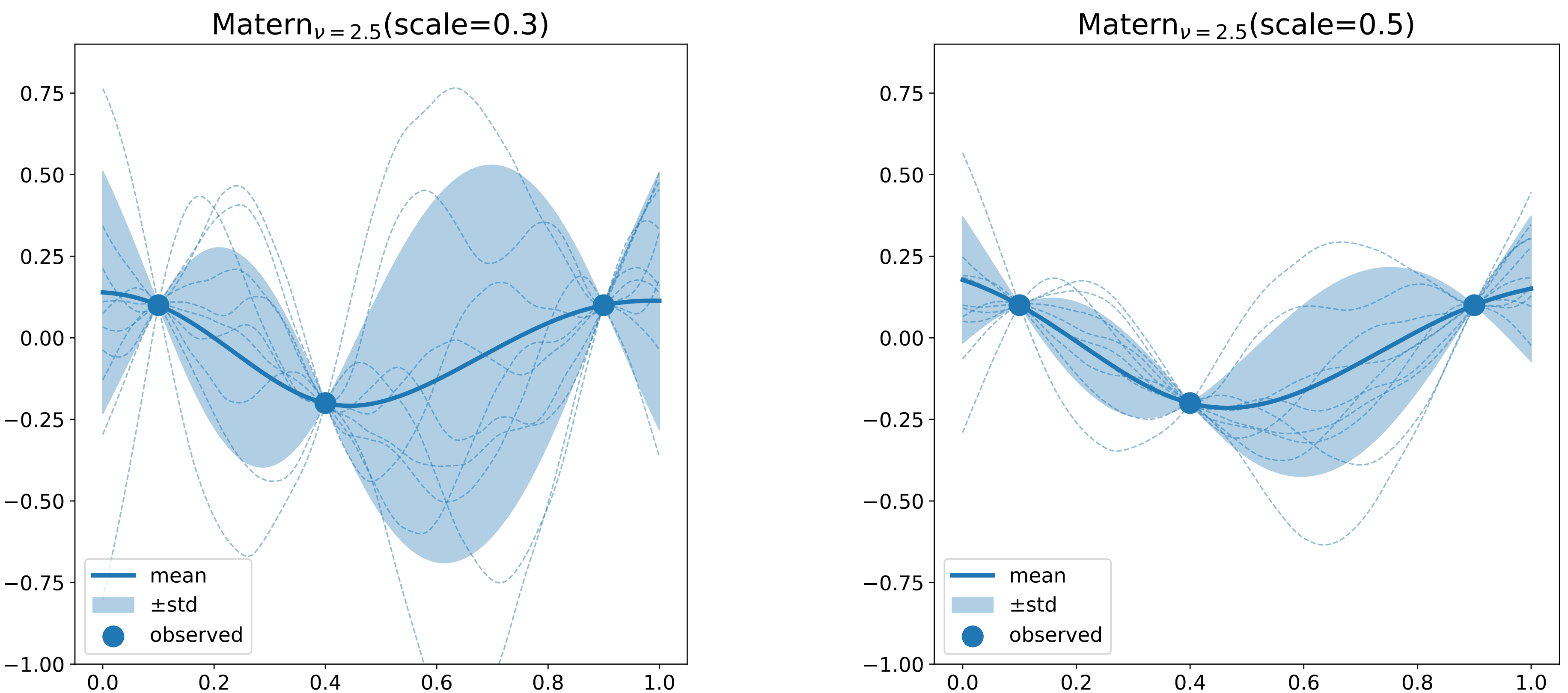
RBF(scale=0.3)
RBF(scale=0.5)
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0.0
0.2
0.4
0.6
0.8
1.0
mean
±std
observed

# Matern kernel, $\nu = 1.5$



# Matern kernel, $\nu = 2.5$



# Gaussian processes, summary

- GP is a Bayesian inference over a linear model:
  - analytical form for posterior $P(y \mid X, Y, x)$;
  - functions can be sampled from GP.
- with basis expansion and kernels, GP is a powerful model:
  - kernel version is slow: $\mathcal{O}(n^3)$;
  - linear/basis expansion version: $\mathcal{O}(nd^3)$.

# Acquisition functions

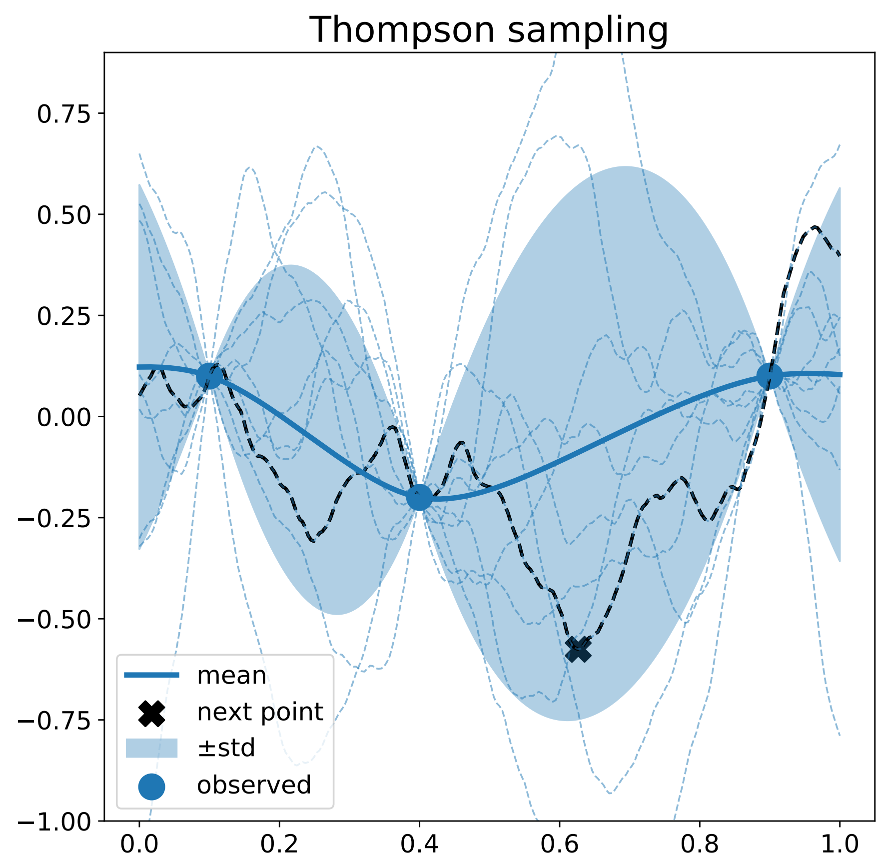
# Role in Bayesian Optimization

```
for i = 1 to N do
    compute P(θ | X, Y)
    search for x_i with the most expected gain
    evaluate y_i = t(x_i)
    extend X and Y with x_i and y_i
end for
```

# Thompson sampling

1. draw $\theta \sim P(\theta \mid X, Y)$;
2. find $x' = \arg\min f_\theta$;
3. select $x'$ as the next point.

- stochastic;
- encourages exploration;
- applicable for wide range of situations.

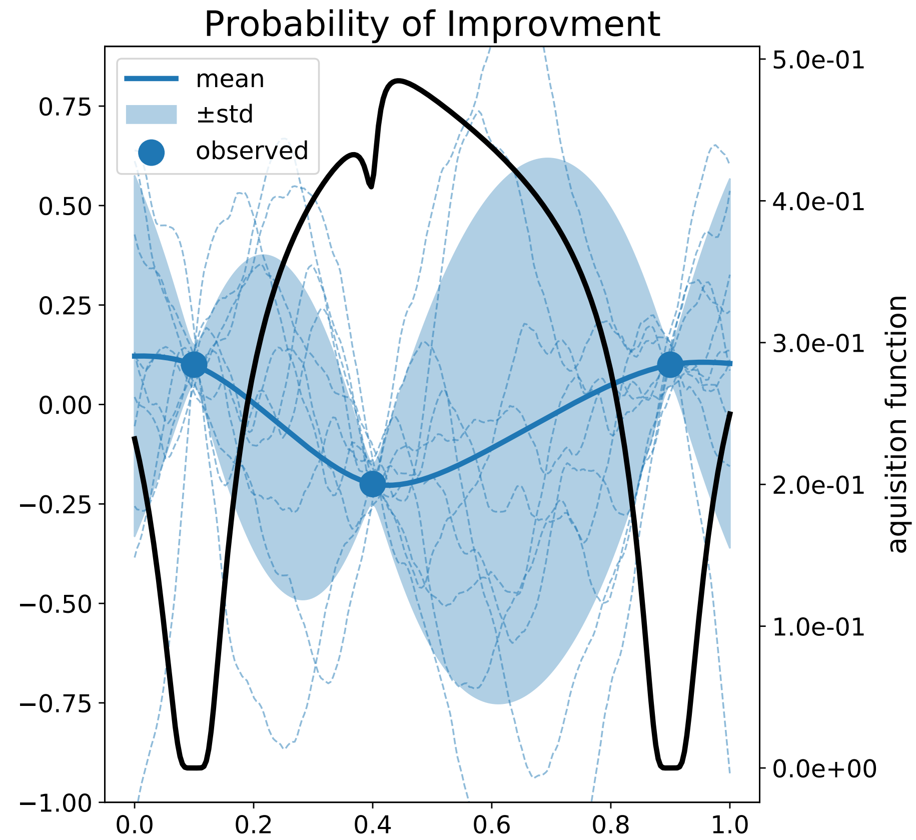


# Probability of Improvement

$$x' = \arg\max P\big(y < y^* \mid X, Y, x\big)$$

where $y^* = \min Y$.

- exploitative;
- tends to get stuck at the same points.

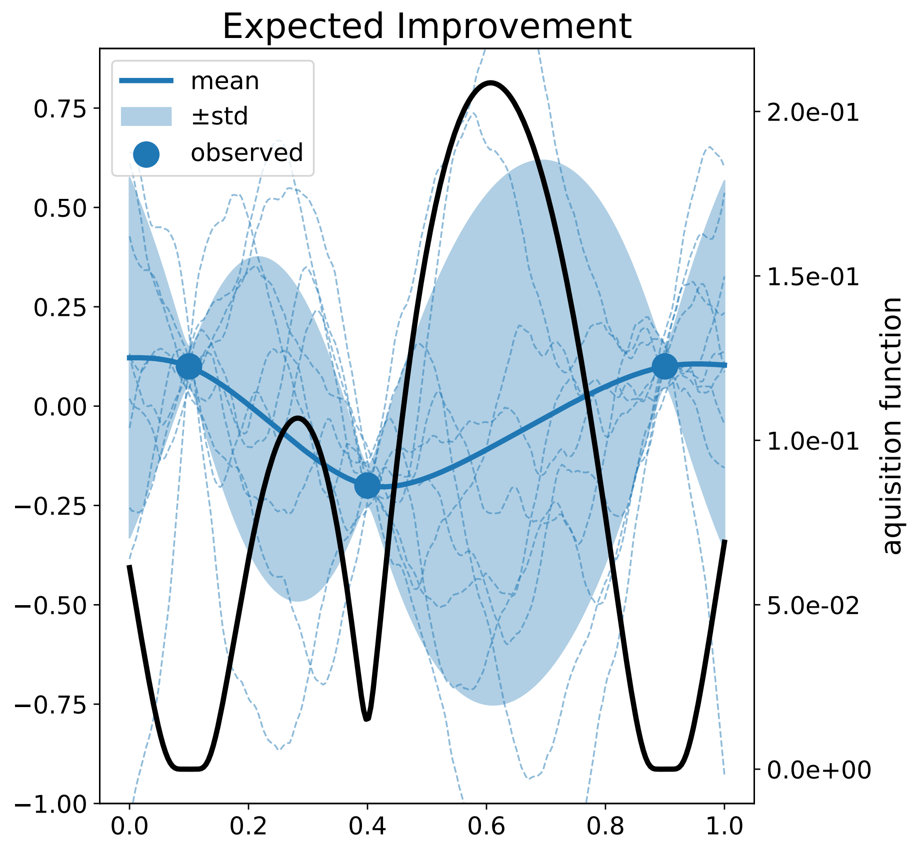


# Expected Improvement

$$x' = \arg\max \mathop{\mathbb{E}}_{y} \left[ y - y^* \mid X, Y, x \right]$$

where $y^* = \min Y$.

- explorative;
- one of the most popular choices.

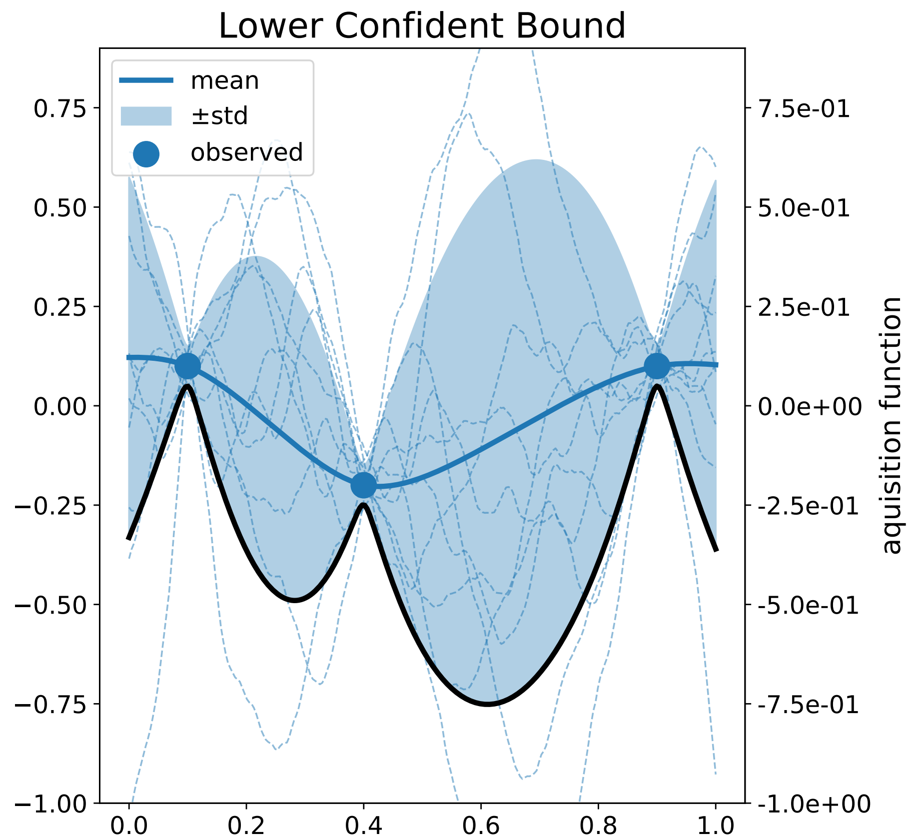


# Lower Confidence Bound

$$x' = \arg\min\big(\mu(x) - \gamma\sigma(x)\big)$$

where $\gamma \in (0, +\infty)$.

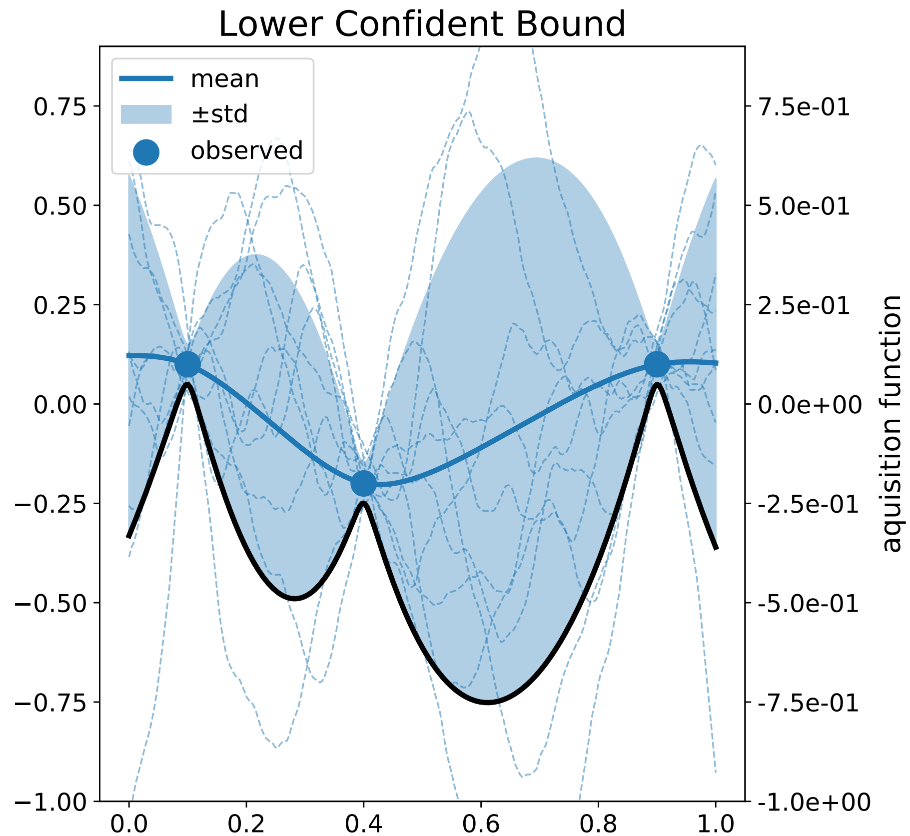
- explorative;
- one of the most popular choices.



# Optimization of acquisition functions

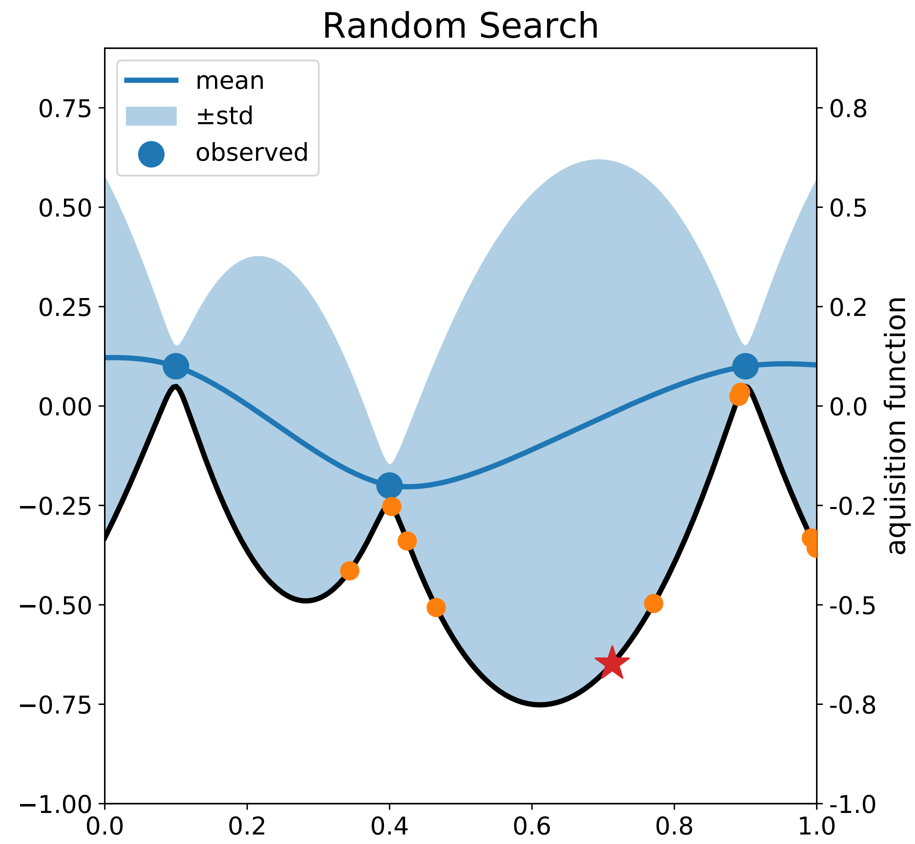
# Optimization of acquisition functions

- Gaussian Processes are differentiable;
- most acquisition functions are differentiable;
- in general, **non-convex**:
  - multiple local minima.



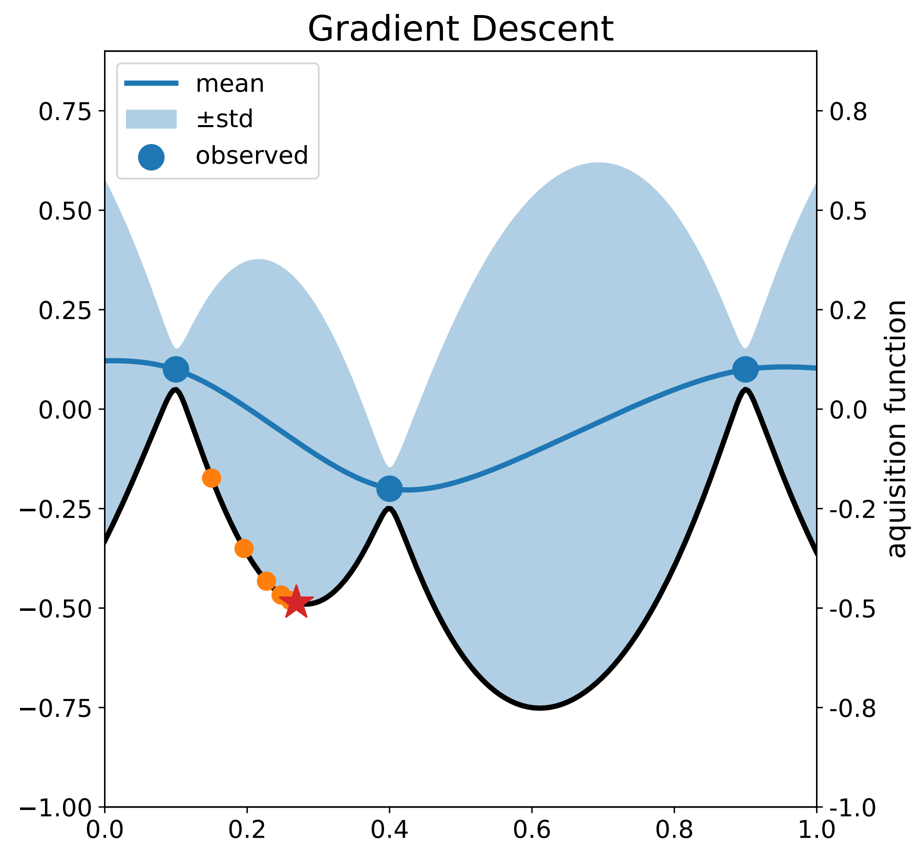
# Random Search

- **global** optimization algorithm;
- imprecise:
  - reduce convergence speed;
  - makes exploitative acquisition functions more explorative.



# Gradient methods

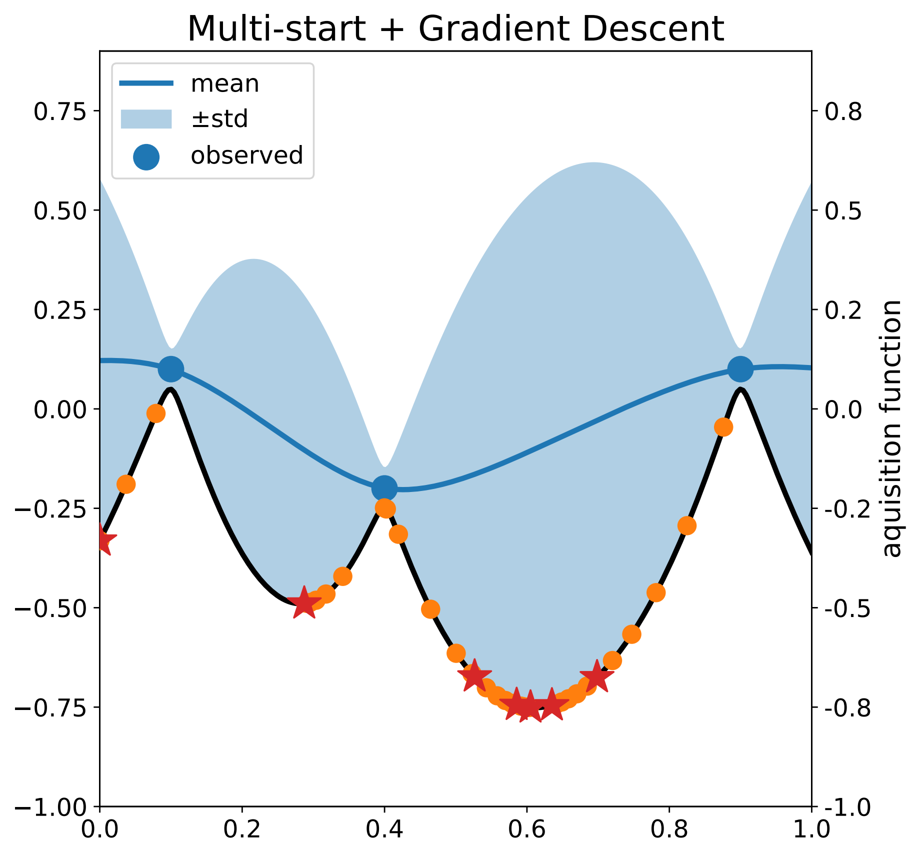
- local optimization:
  - optimization converges to a local minimum;
- **precise**.



# Multi-start

1. randomly draw initial guess;
2. descend with a gradient method;
3. repeat;

- **global** optimization;
- **precise**.



# Summary

# Gaussian Processes in practice

- basis expansion:
  - good basis is known;
- RBF kernel:
  - popular choice;
  - the objective is expected to be smooth;
- Matern kernel:
  - can be adjusted via $\nu$;
  - $\nu = 1.5$ — once differentiable functions;
  - $\nu \to +\infty$ — approaches RBF.

# Acquisition functions

- Thompson sampling:
  - stochastic;
  - only sampling from $P(\theta \mid X, Y)$;
- Probability of Improvement:
  - tends to get stuck at the same place;
  - greedy (exploitative);
- Expected improvement:
  - popular choice;
  - explorative;
- Lower Confidence Bound:
  - popular choice;
  - easy to computer for GP;
  - explorative.

# Optimization of acquisition functions

Multi-start with gradient methods:

- global;
- precise;
- de facto standard.

# Quizzz

Consider a Gaussian process with mean $\mu(x)$ and variance $\sigma^2(x)$, and the following acquisition function:

$$x' = \arg\max_x \sigma(x).$$

Which characteristics can be applied to the acquisition function:

1. explorative;
2. exploitative;
3. both;
4. neither?

*Is there any relation to Lower Confidence Bound?*

# References, BO


- Shahriari, B., Swersky, K., Wang, Z., Adams, R.P. and De Freitas, N., 2015. Taking the human out of the loop: A review of Bayesian optimization. Proceedings of the IEEE, 104(1), pp.148-175.
- Daniel James Lizotte. 2008. Practical bayesian optimization. Ph.D. Dissertation. University of Alberta, CAN. Order Number: AAINR46365.
- Frazier, P.I., 2018. A tutorial on bayesian optimization. arXiv preprint arXiv:1807.02811.