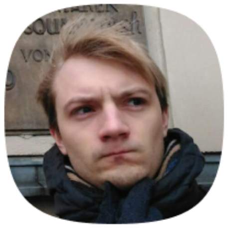
Vladislav Belavin, Maxim Borisyak



Bayesian Optimization

Introduction

2021

NATIONAL RESEARCH UNIVERSITY

LAMBDA • HSE

SCHOOL OF DATA ANALYSIS

Schaffhausen Institute of Technology

Surrogate Optimization

SHiP shield optimization

$$\text{background}(\theta) = \mathop{\mathbb{E}}_{\text{event}} \mathbb{I}\big[\text{muons} > 0 \mid \text{event}, \theta\big] \to \min$$

- computationally expensive;
- no gradient information;
- noisy.

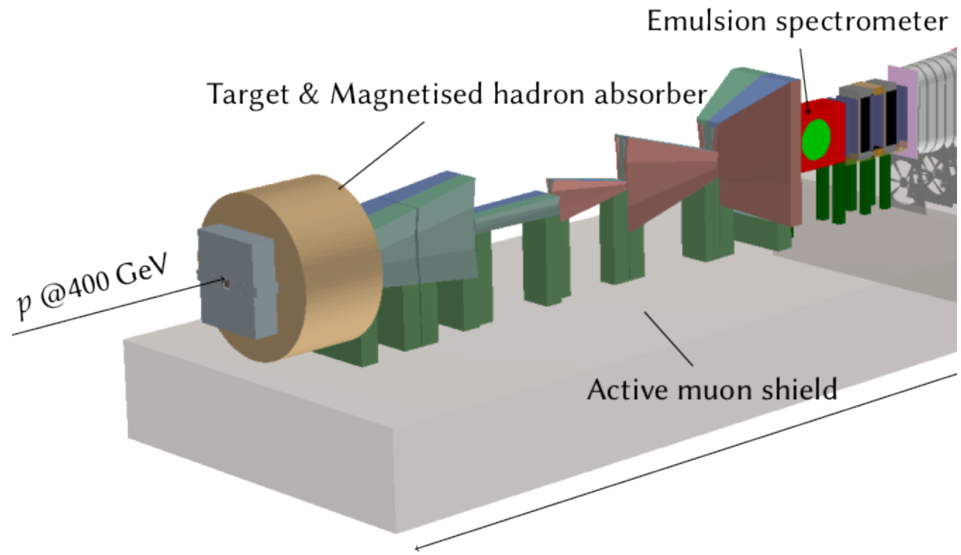


Image source: Oliver Lantwin, Bayesian optimisation of the SHiP muon shield.

Surrogates

Substitute objective function with a surrogate.

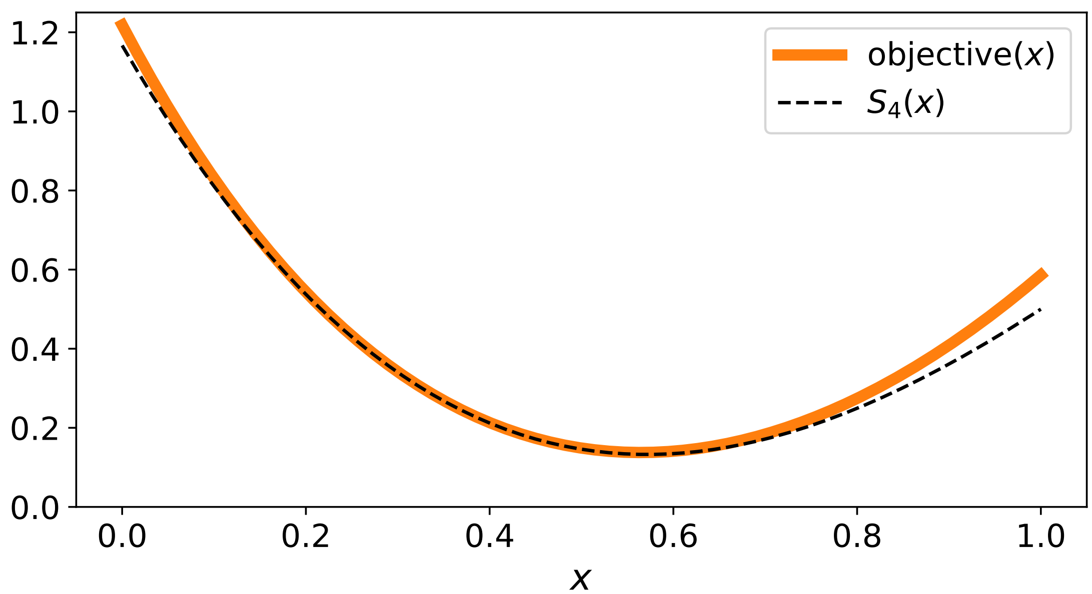
$$\begin{array}{ccc} f(x) & \to & \min; \\ & \downarrow & \\ g(x) & \to & \min; \end{array}$$

where:

- $g_\psi(x) \approx f(x)$;
- g — cheap to evaluate.

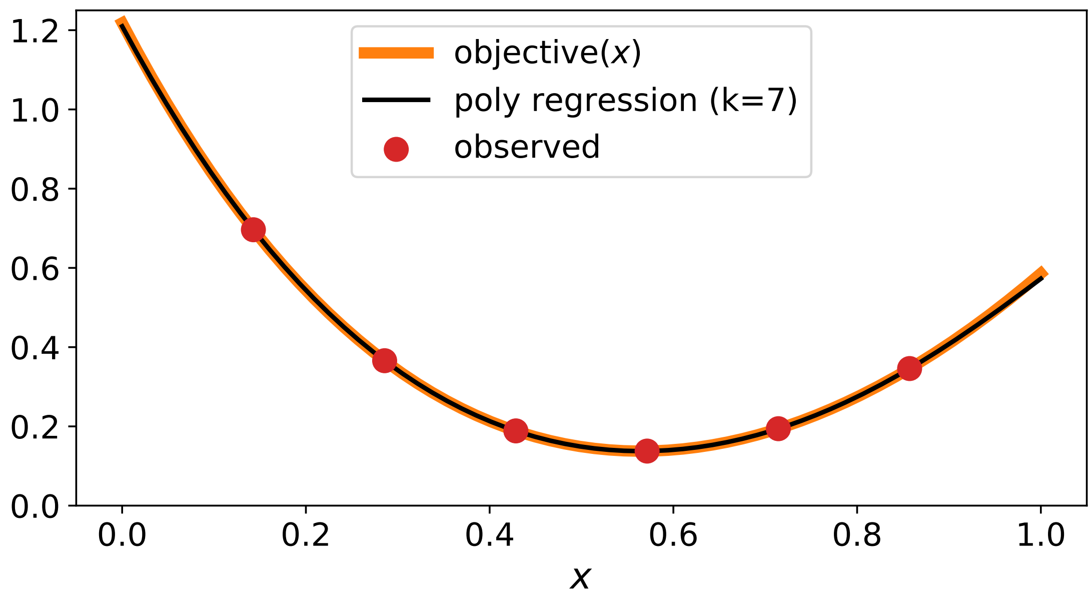
An example of a good surrogate

$$\begin{aligned} \text{objective}(x) &= \exp(-2x+1) + \exp(x) - 2.5 \approx S_k(-2x+1) + S_k(x) - 2.5 \\ S_k(x) &= \sum_{n=0}^{k} x^n / n! \end{aligned}$$

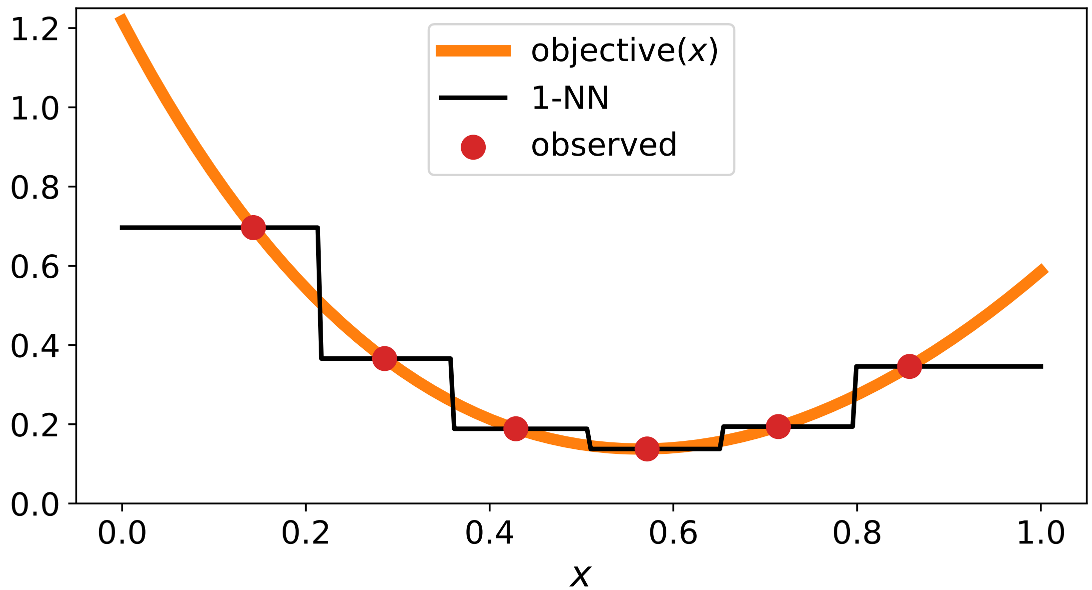


Surrogate models

Train a surrogate model.

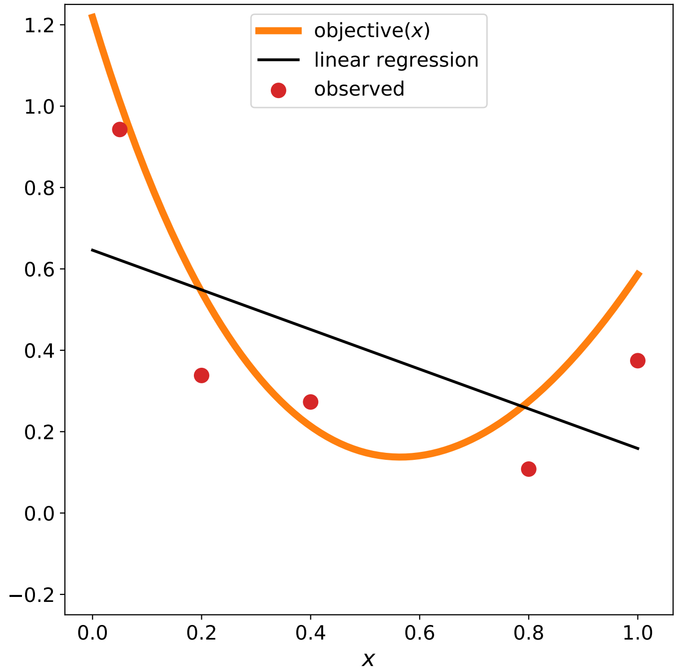


Grid search as a surrogate optimization

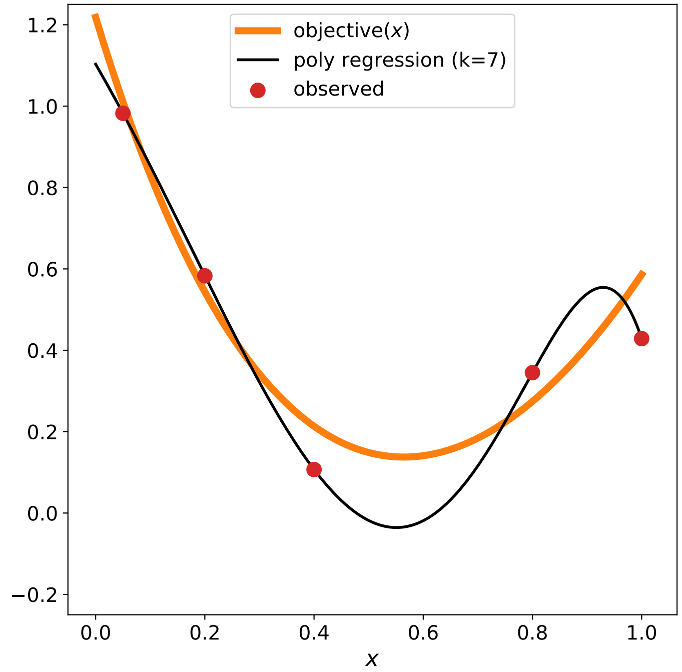


Discussion

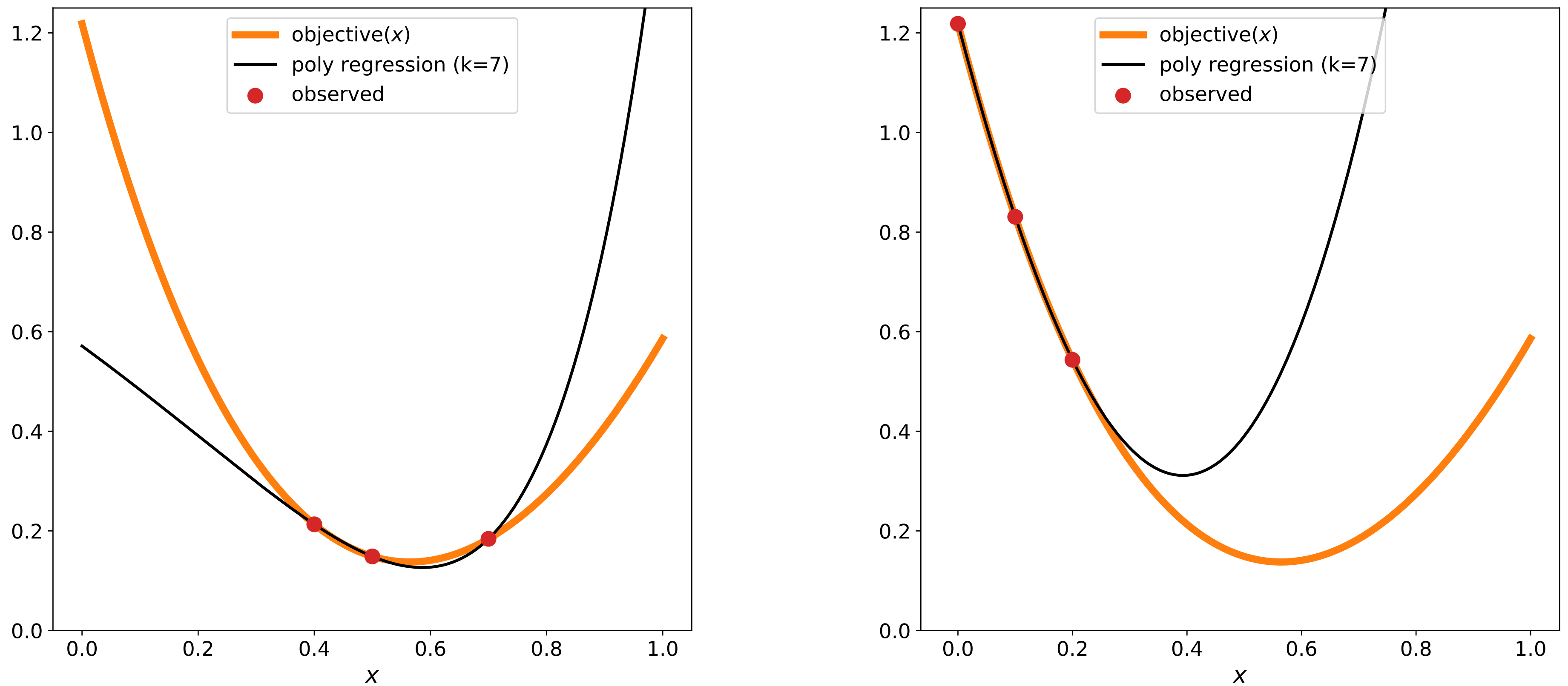
low-capacity model



high-capacity model

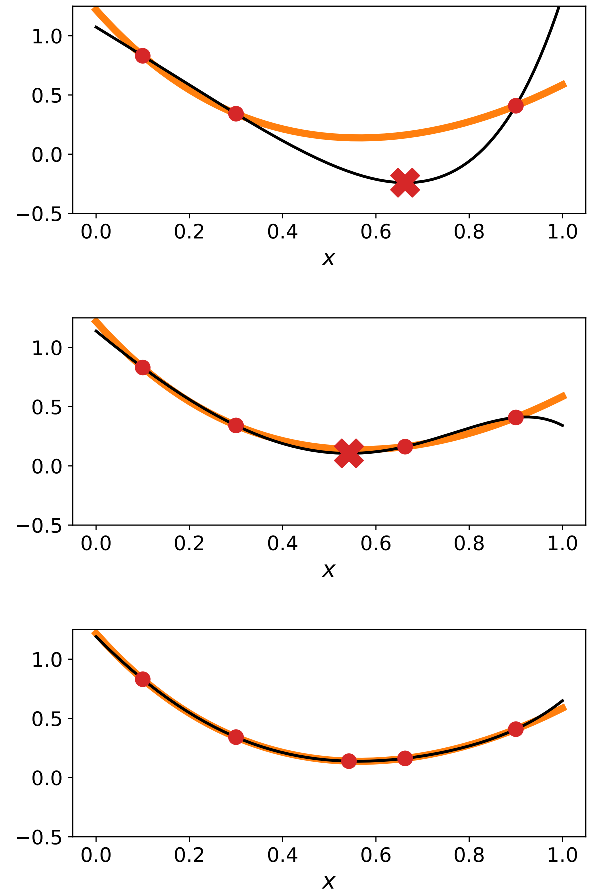


Point selection

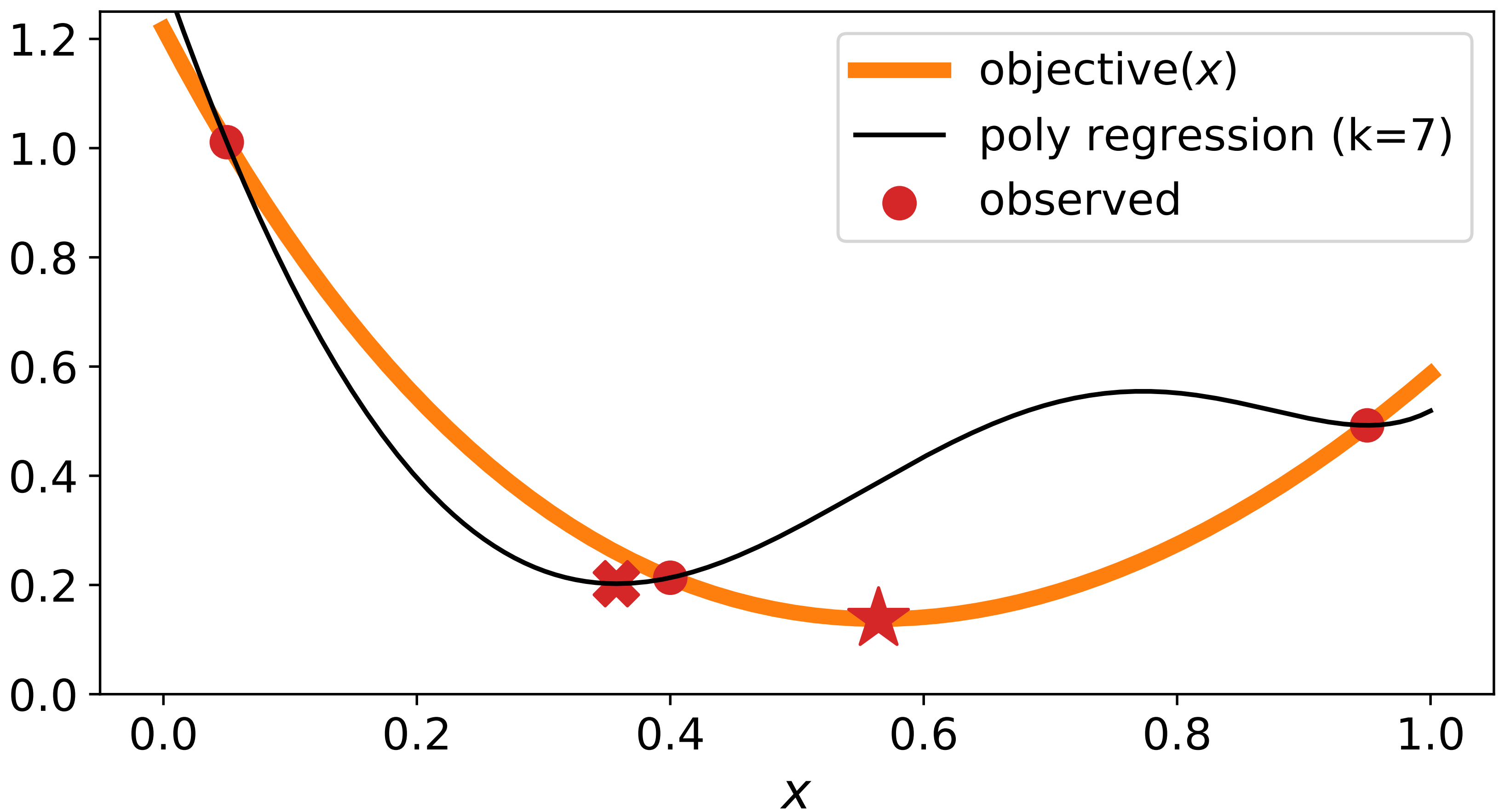


Greedy surrogate optimization

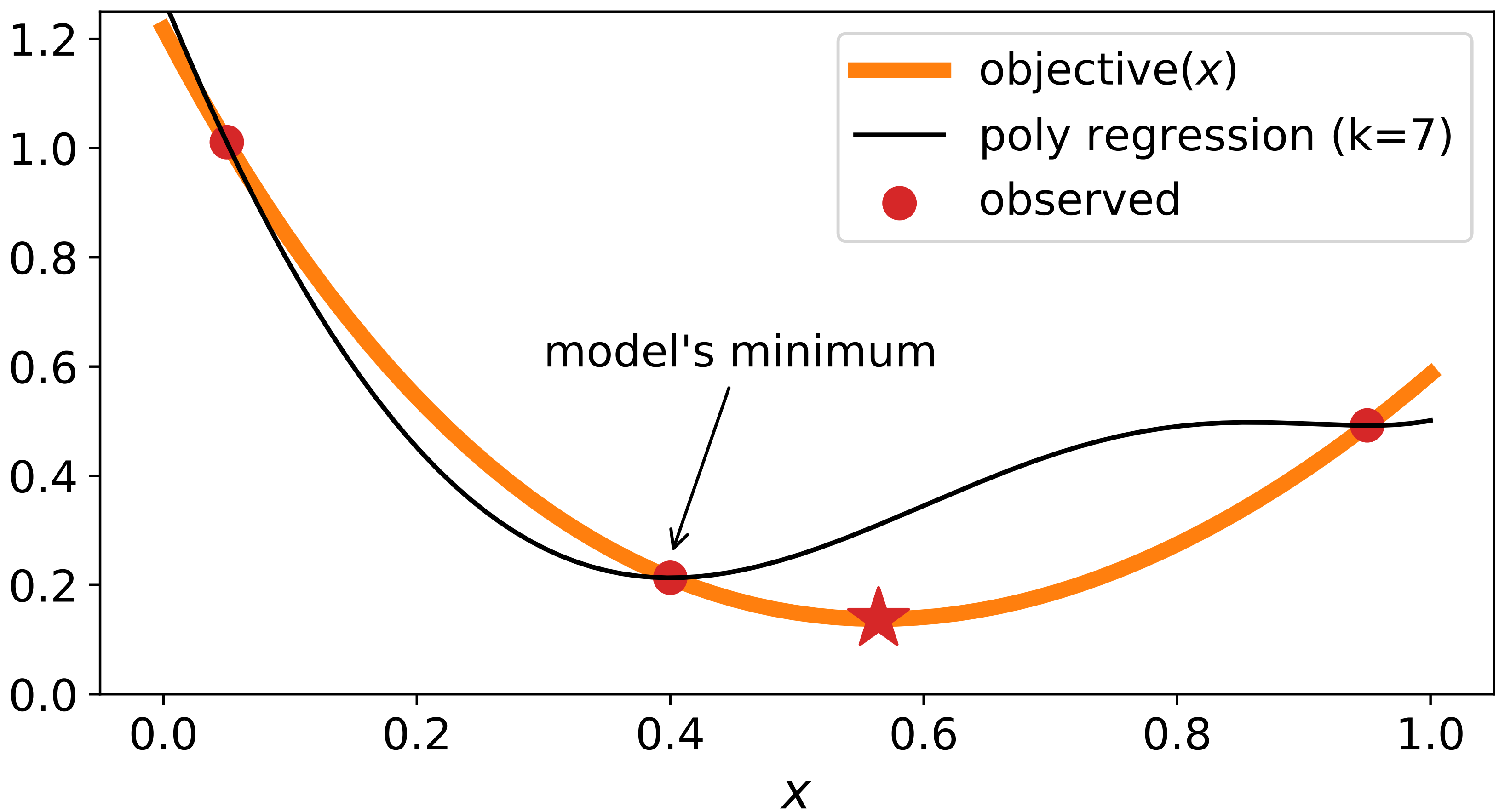
```
X ← ∅
Y ← ∅
for i = 1 to N do
    // train the model
    θ* ← argmin_θ L(X, Y, θ)
    // minimum of the trained model
    x* ← argmin_x f(x, θ*)
    y* ← objective(x*)
    X ← X ∪ {x*}
    Y ← Y ∪ {y*}
end for
```



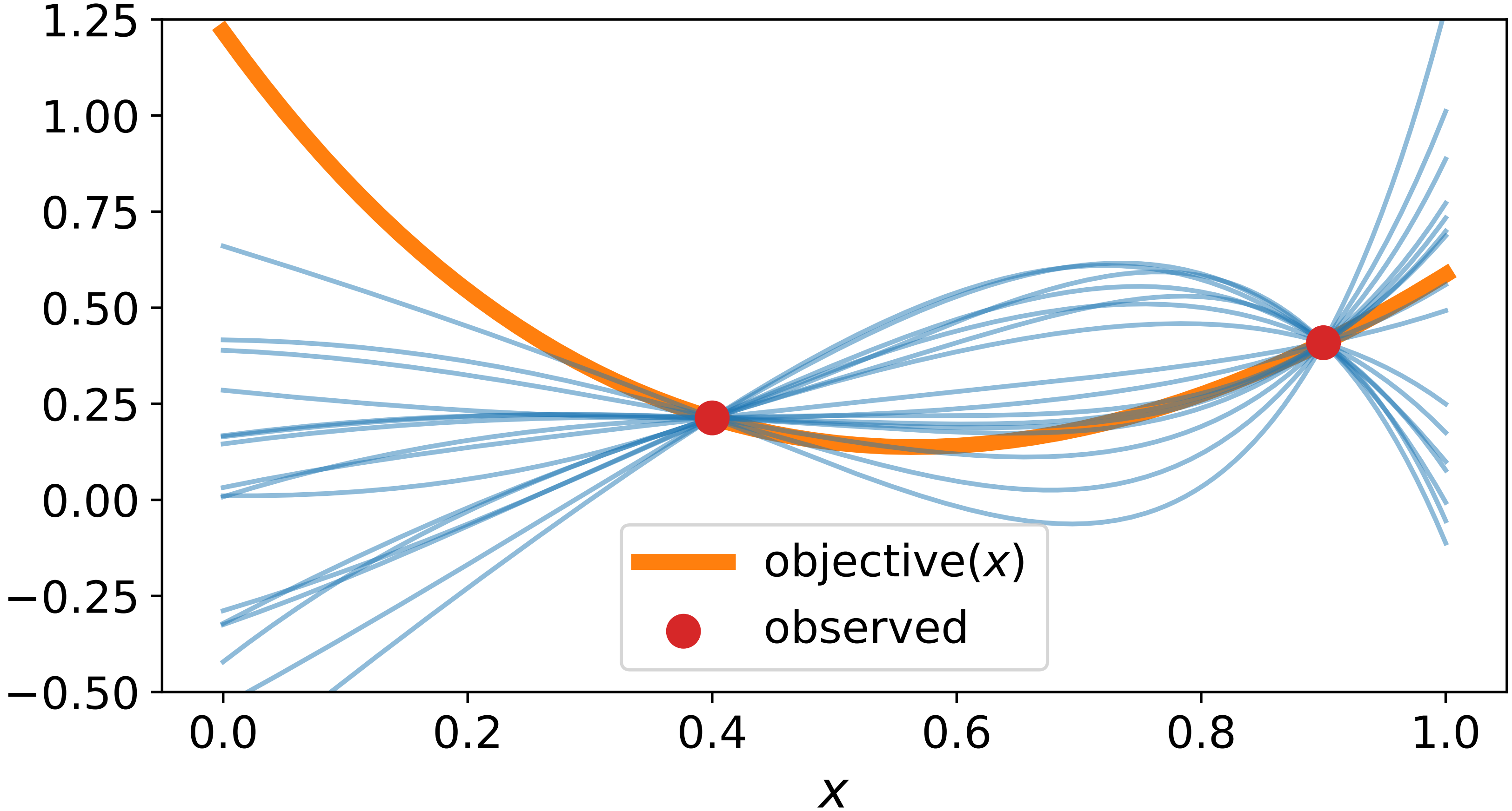
Greedy optimization: counterexample



Greedy optimization: counterexample



The source of the problem



Bayesian Inference

Bayesian inference

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

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

- $P(\theta)$ — prior;
- $P(y \mid x, \theta)$ — data model;
- $P(\theta \mid X, Y)$ — posterior.

Naive approximate Bayesian inference

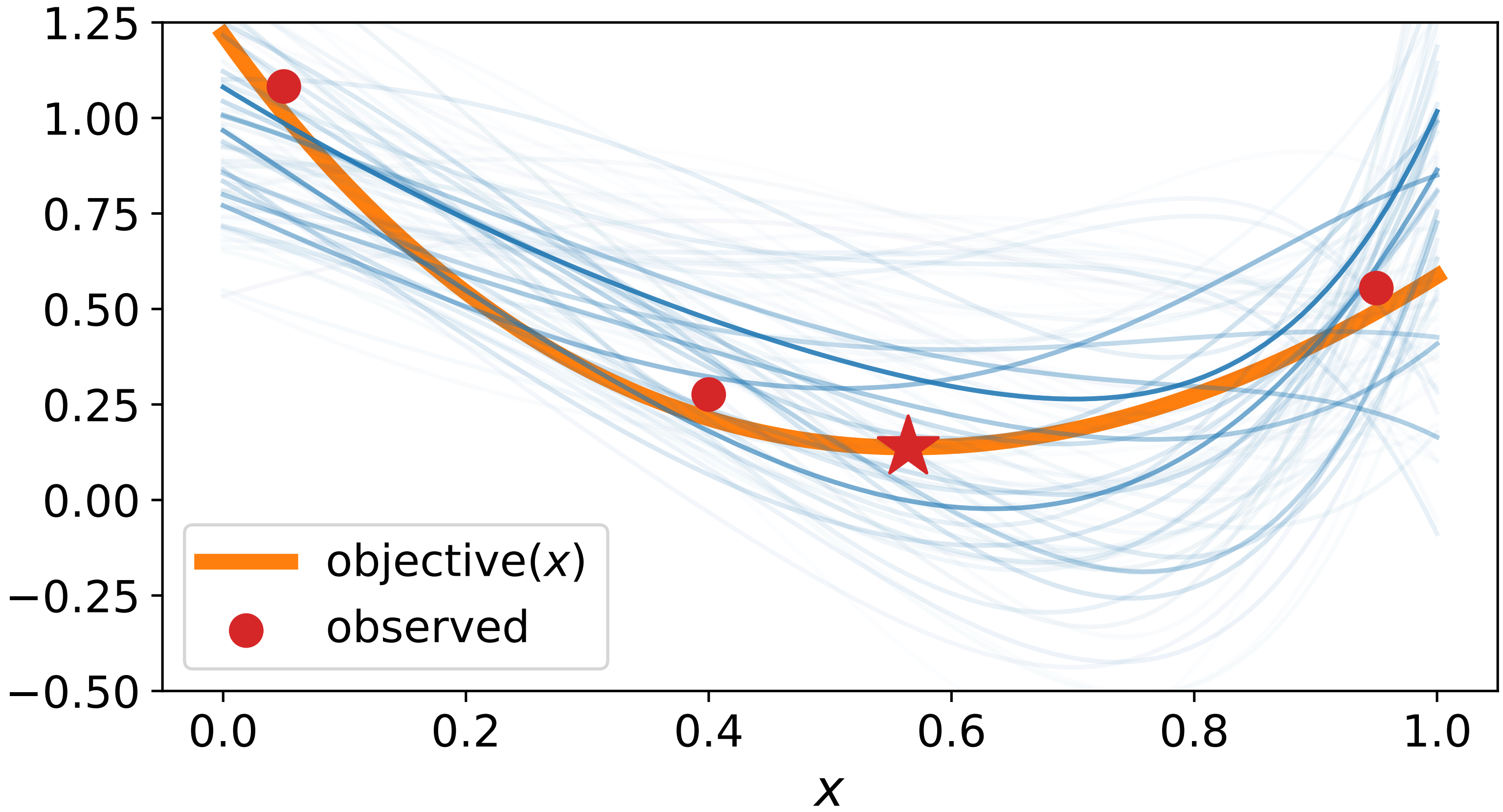
Do not try this at home...

$$P(\theta \mid X, Y) = \frac{P(Y \mid X, \theta) P(\theta)}{\int P(Y \mid X, \psi) P(\psi) \, d\psi} = \frac{1}{Z} \cdot P(Y \mid X, \theta) P(\theta)$$

$$\begin{aligned} Z &\approx \frac{1}{N} \sum_{i=1}^{N} P(Y \mid X, \theta_i); \\ \theta_i &\sim P(\theta) \end{aligned}$$

- biased and computationally inefficient.

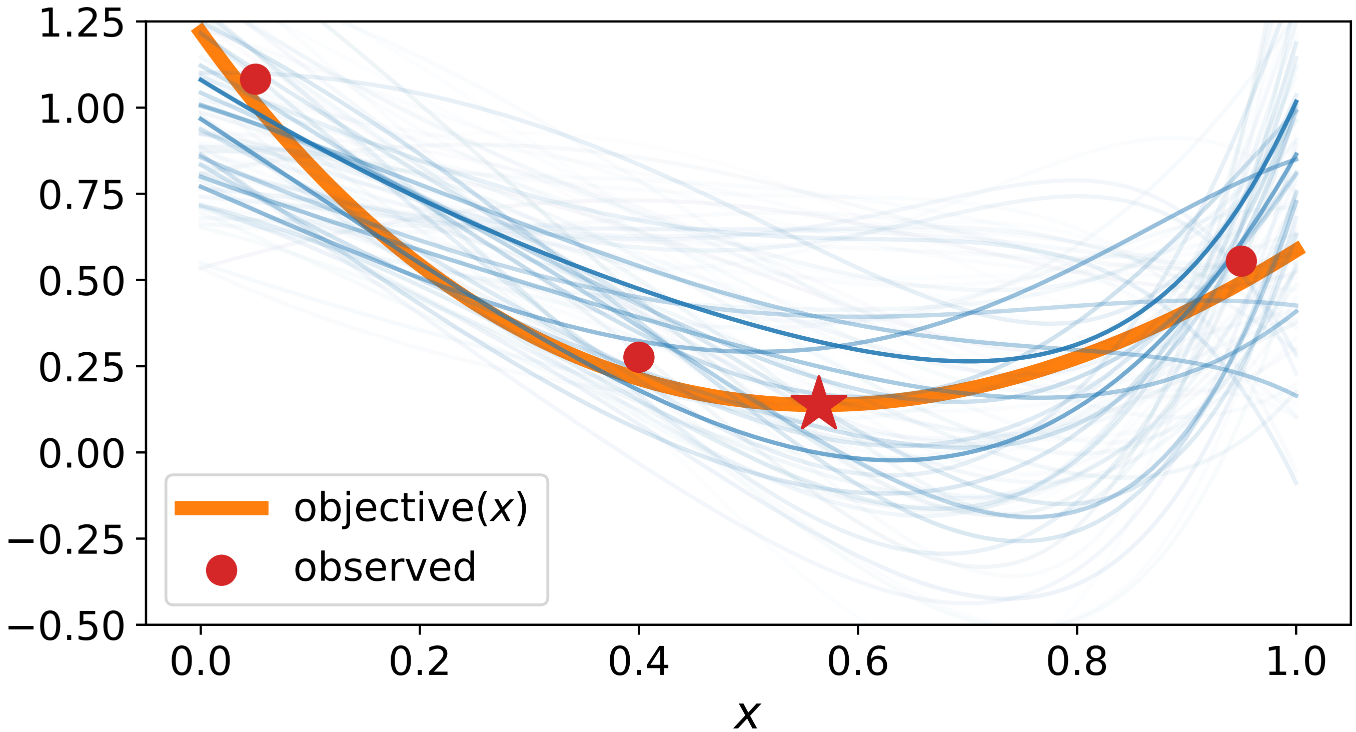
Bayesian inference



Bayesian Optimization

Bayesian Optimization

What is next?



Acquisition functions

Strategy for selecting the next point is called **acquisition function**:

- simple:

$$x_{\text{next}} = \arg\max_x J(y \mid X, Y, x);$$

- lookahead/expected gain:

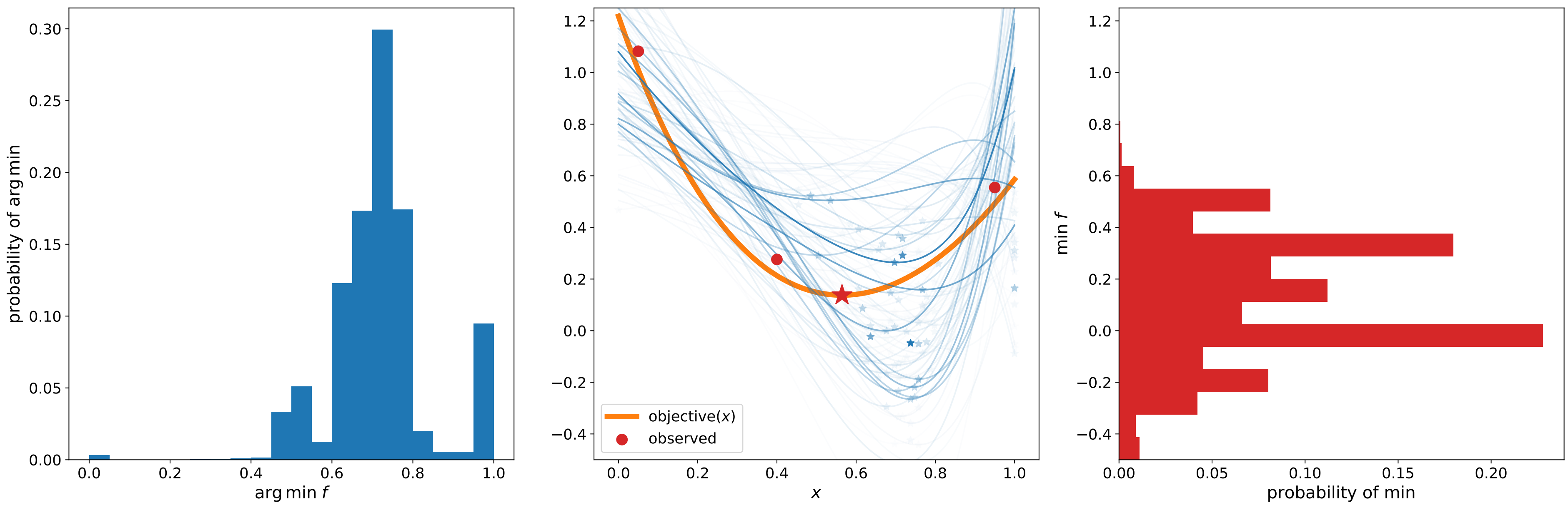
$$x_{\text{next}} = \arg\max_x \mathop{\mathbb{E}}_{y \sim P(y \mid X, Y)} J(\theta \mid X \cup \{x\}, Y \cup \{y\}).$$

Main loop

$X \leftarrow \varnothing$
$Y \leftarrow \varnothing$
for $i = 1$ to N do
    compute $P(\theta \mid X, Y)$
    search for x_i with the most expected gain
    evaluate $y_i = t(x_i)$
    $X \leftarrow X \cup \{x^*\}$
    $Y \leftarrow Y \cup \{y^*\}$
end for

Entropy search

Distribution of minima



Entropy search

- entropy of a random variable X, a measure of uncertainty:

$$\mathrm{H}(X) = -\mathbb{E} \log P(X);$$

- current uncertainty on the position of the minimum:

$$\mathrm{H}\big(\arg\min f_\theta \mid X, Y\big);$$

- uncertainty after measurements (x, y):

$$\mathrm{H}\big(\arg\min f_\theta \mid X \cup \{x\}, Y \cup \{y\}\big);$$

- **expected** uncertainty after evaluating the objective in x:

$$\boxed{x_{\mathrm{next}} = \arg\min_x \mathop{\mathbb{E}}_{y \sim P(y \mid X, Y)} \mathrm{H}\big(\arg\min f_\theta \mid X \cup \{x\}, Y \cup \{y\}\big).}$$

Entropy search

$$x_{\text{next}} = \arg\min_{x} \mathop{\mathbb{E}}_{y \sim P(y \mid X, Y)} \mathrm{H}\big(\arg\min f_\theta \mid X \cup \{x\},\, Y \cup \{y\}\big)$$

x_{next} is expected to bring the most information about $\arg\min f$:

- **hard to compute**;
- also consider:
  - $\mathrm{H}(\min f_\theta)$ - max-value entropy search;
  - $\mathrm{H}(\arg\min f_\theta, \min f_\theta)$;
  - $\mathrm{H}(\theta)$.

Exploration vs exploitation

Acquisition function:

- decreases uncertainty — **explorative**, e.g.:

$$x_{\text{next}} = \arg\max_x \mathop{\mathbb{E}}_{y \sim P(y \mid X, Y)} \mathrm{H}(\theta \mid X \cup \{x\},\, Y \cup \{y\});$$

- probing for minimum — **exploitative**, e.g.:

$$x_{\text{next}} = \arg\max_x P(x = \arg\min f_\theta \mid X,\, Y)/$$

Entropy search tends to be explorative.

Summary

Summary

Surrogate optimization:

- objective $\to$ surrogate;
- surrogate $\to$ regression model;
  - decrease overfitting by evaluating more point;
  - often fails/ineffective.

Bayesian Optimization:

- regression model $\to$ posterior distribution of surrogates:
  - no overfitting;
- acquisition functions.

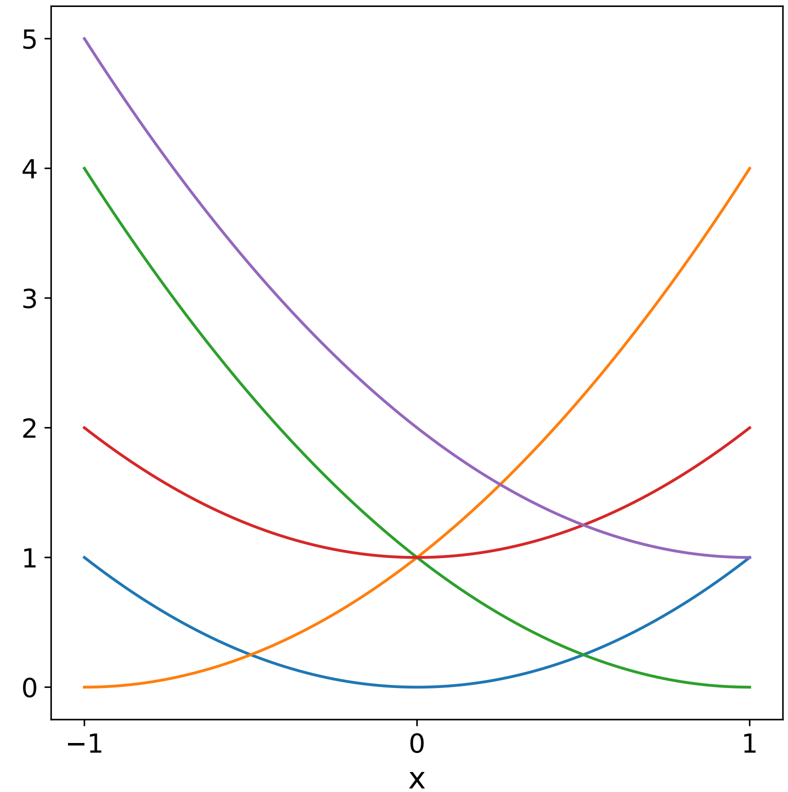
Quiz

Assuming that:

- each function has equal posterior;
- $P(y \mid f, x) = \mathcal{N}(f(x), \sigma^2), \sigma \ll 1$;

which of the following points achieves the highest entropy gain w.r.t location of $\arg\min$?

1. $x_{\text{next}} = -1$;
2. $x_{\text{next}} = 0$;
3. $x_{\text{next}} = +1$;
4. $x_{\text{next}} = 1/2$.



References

- Audet, C. and Hare, W., 2017. Derivative-free and blackbox optimization.
- Snoek, J., Larochelle, H. and Adams, R.P., 2012. Practical bayesian optimization of machine learning algorithms. In Advances in neural information processing systems (pp. 2951-2959).
- Wang Z, Jegelka S. Max-value Entropy Search for Efficient Bayesian Optimization. InInternational Conference on Machine Learning 2017 Jul 17 (pp. 3627-3635).
- Hennig, Philipp and Schuler, Christian J. Entropy search forinformation-efficient global optimization.Journal of MachineLearning Research, 13:1809–1837, 2012.
- Herńandez-Lobato, Jos´e Miguel, Hoffman, Matthew W, andGhahramani, Zoubin. Predictive entropy search for efficientglobal optimization of black-box functions. In Advances in Neural Information Processing Systems (NIPS), 2014.

Extra

Maximum Likelihood estimation

$$L(\theta) = P(Y \mid X, \theta) = \prod_i P(y_i \mid x_i, \theta) \to \max;$$

$$\mathcal{L}(\theta) = -\sum_i \log P(y_i \mid x_i, \theta) \to \min .$$

Gaussian noise:

$$P(y \mid x, \theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y - f_\theta(x))^2}{2\sigma^2}\right)$$

$$\mathcal{L}(\theta) = -\sum_i \log P(y_i \mid x_i, \theta) = \sum_i \frac{(y - f_\theta(x))^2}{2\sigma^2} + \mathrm{const} \propto \sum_i (y - f_\theta(x))^2 \to \min$$

Maximum a Posteriori estimation

$$
\begin{aligned}
P(\theta \mid X, Y) &= \frac{P(Y \mid X, \theta) P(\theta)}{P(Y \mid X)} \to \max \\
\mathcal{L}(\theta) &= -\log P(\theta) - \sum_i \log P(y_i \mid x_i, \theta) \to \min
\end{aligned}
$$

Gaussian noise and Gaussian prior:

$$
\mathcal{L}(\theta) = -\log P(\theta) - \sum_i \log P(y_i \mid x_i, \theta) \propto \sum_i (y - f_\theta(x))^2 + \alpha \|\theta\|^2 \to \min
$$

Maximum a Posteriori trap