Гауссовские процессы в машинном обучении

Вячеслав Боровицкий (СПБГУ, ПОМИ РАН)

24 сентября 2020 г.

Структура презентации

Содержание

Гауссовские процессы: много миров

Теоретический мир, например

L_2-small ball asymptotics for a family of finite-dimensional perturbations of Gaussian functions

Petrova Yu.P.*

Прикладной мир, например

Efficiently Sampling Functions from Gaussian Process Posteriors

---

James T. Wilson [*1] **Viacheslav Borovitskiy** [*2 3] **Alexander Terenin** [*1]
Peter Mostowsky [*2] **Marc Peter Deisenroth** [4]

Гауссовские процессы в ML: одно из приложений

Байесовская оптимизация — алгоритм оптимизации, основанный на гауссовских процессах.

Bayesian Optimization in AlphaGo

Yutian Chen, Aja Huang, Ziyu Wang, Ioannis Antonoglou, Julian Schrittwieser, David Silver & Nando de Freitas

DeepMind, London, UK
yutianc@google.com

Была использована для подбора гиперпараметров алгоритма, который в марте 2016 со счётом 4:1 победил в игру Го профессионала 9-го (высшего) дана Ли Седоля.

Содержание

Теорема Байеса

Пусть Θ и $\mathcal{D}$ — случайные величины. Тогда

$$\underbrace{\mathbb{P}(\Theta = \theta \mid \mathcal{D} = d)}_{\text{Апостериорное распр.}} = \frac{\overbrace{\mathbb{P}(\mathcal{D} = d \mid \Theta = \theta)}^{\text{Правдоподобие}} \overbrace{\mathbb{P}(\Theta = \theta)}^{\text{Априорное распр.}}}{\underbrace{\mathbb{P}(\mathcal{D} = d)}_{\text{Нормализующая константа}}}.$$

Наиболее распространенная в байесовской статистике запись формулы:

$$p(\theta \mid d) = \frac{p(d \mid \theta)p(\theta)}{p(d)}.$$

Модель определяется выбором

- априорного распределения $p(\theta)$,
- правдоподобия $p(d \mid \theta)$.

Теорема Байеса при данной модели позволяет вычислить по данным апостериорное распределение на параметры.

Содержание

Байесовский вывод для монетки: теория

Задача: хотим оценить параметры нечестной монетки.

Пусть X — с.в., моделирующая нашу нечестную монетку.
Она принимает два значения: 1 (Орел) и 0 (Решка):

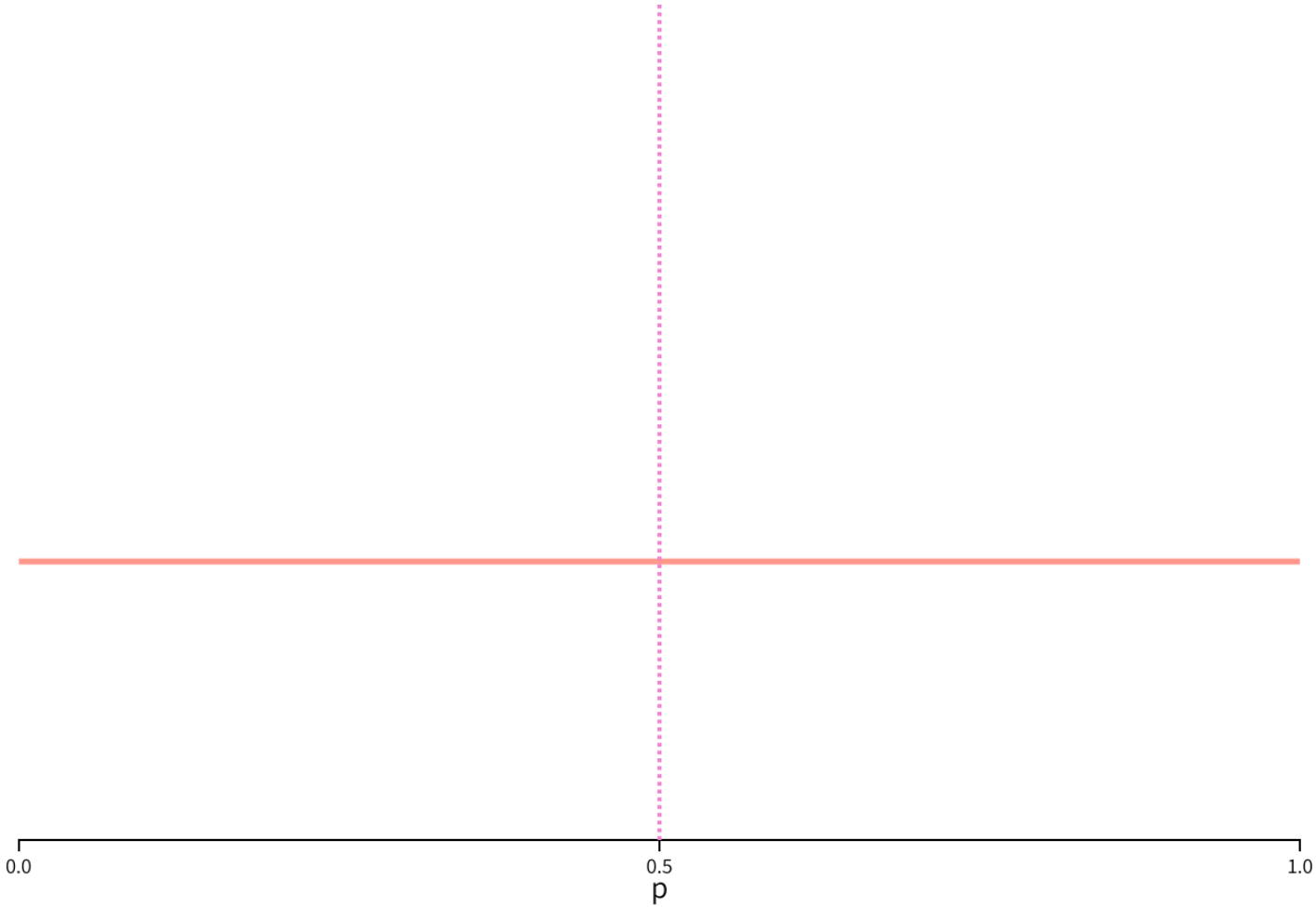
$$\mathbb{P}(X = 1) = p, \qquad \mathbb{P}(X = 0) = 1 - p.$$

Хотим оценить параметр p (вероятность выпадения Орла).

Frequentist approach	Bayesian approach
Ответ: конкретное число $\hat{p}$.	Ответ: распределение $\hat{\rho}(p)$.
Инструмент: оценка макс. правдоподобия.	Инструмент: теорема Байеса.
Дополнительно: —.	Дополнительно: априорное распределение $\rho(p)$.
$\hat{p} = \arg\max p^{\#1}(1-p)^{\#0}$	$\hat{\rho}(p) \propto p^{\#1}(1-p)^{\#0}\rho(p)$

Байесовский вывод для монетки: пример

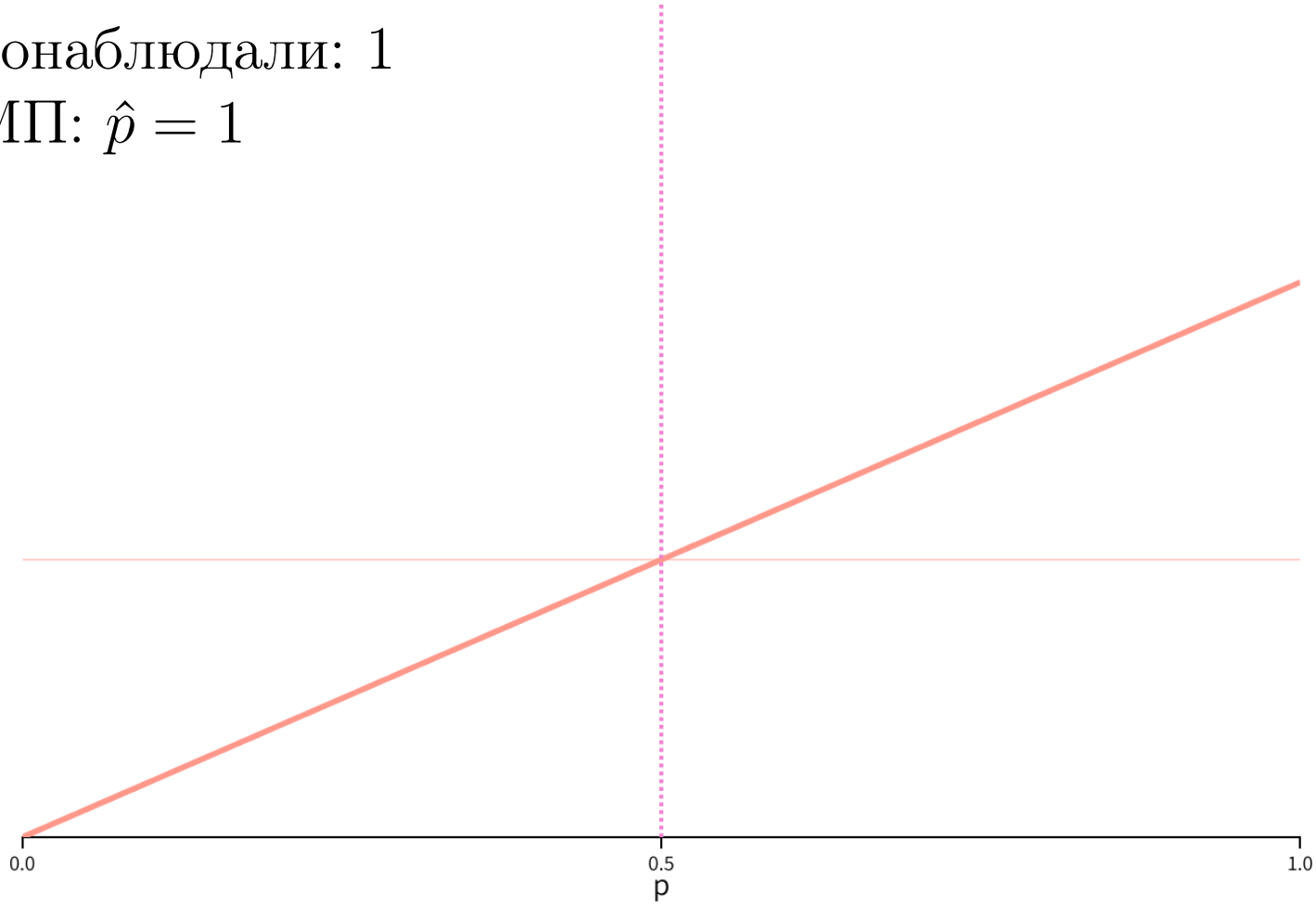
Пусть $\rho(p) := \mathbb{1}_{[0,1]}(p)$ — равномерное распределение на $[0, 1]$, то есть априори мы ничего не знаем о параметре p.
Пусть истинное значение p равно 0.5, т.е. монетка честная.



Байесовский вывод для монетки: пример

Пусть $\rho(p) := \mathbb{1}_{[0,1]}(p)$ — равномерное распределение на $[0, 1]$, то есть априори мы ничего не знаем о параметре p.
Пусть истинное значение p равно 0.5, т.е. монетка честная.

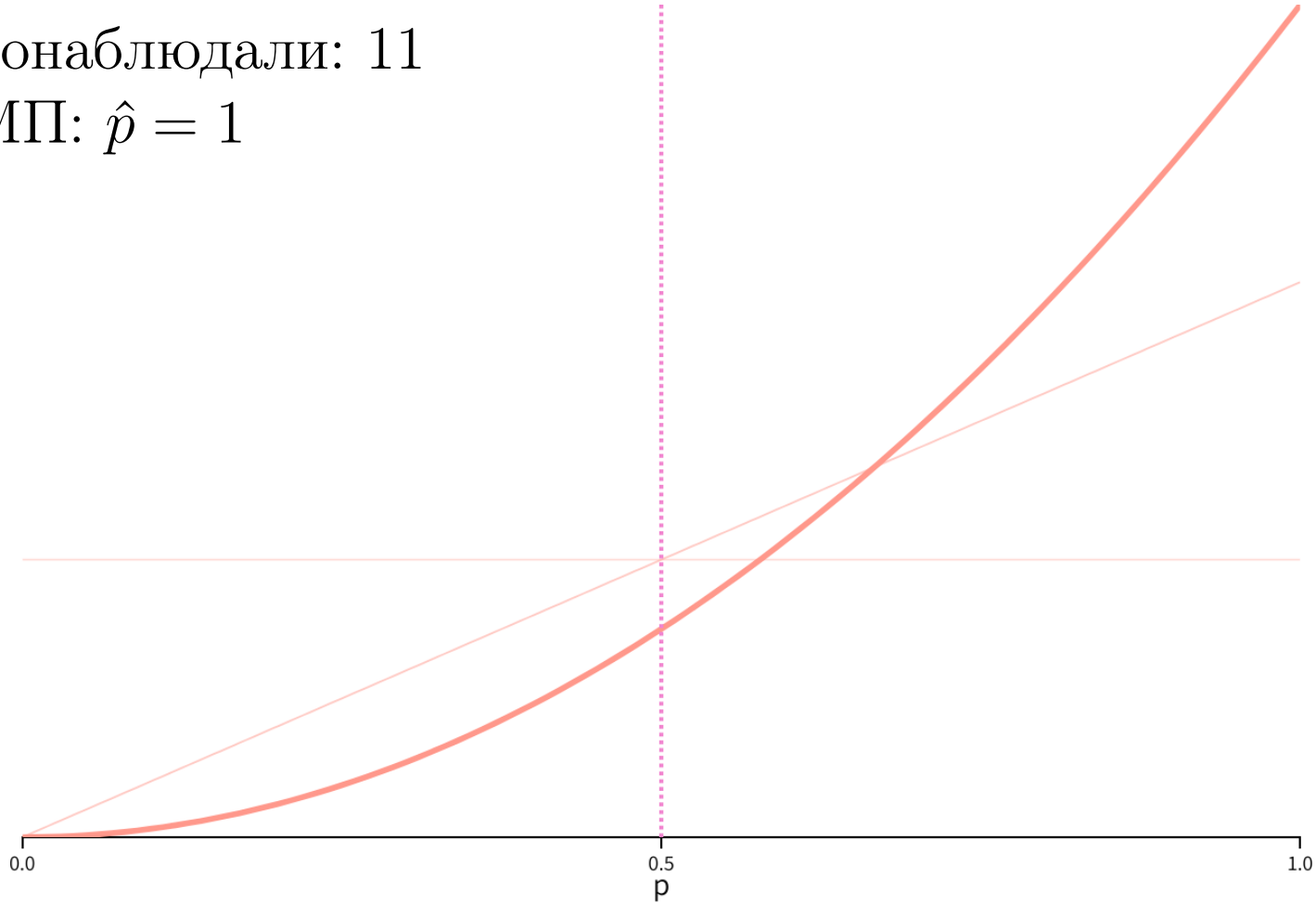
Пронаблюдали: 1
ОМП: $\hat{p} = 1$



Байесовский вывод для монетки: пример

Пусть $\rho(p) := \mathbb{1}_{[0,1]}(p)$ — равномерное распределение на $[0, 1]$, то есть априори мы ничего не знаем о параметре p.
Пусть истинное значение p равно 0.5, т.е. монетка честная.

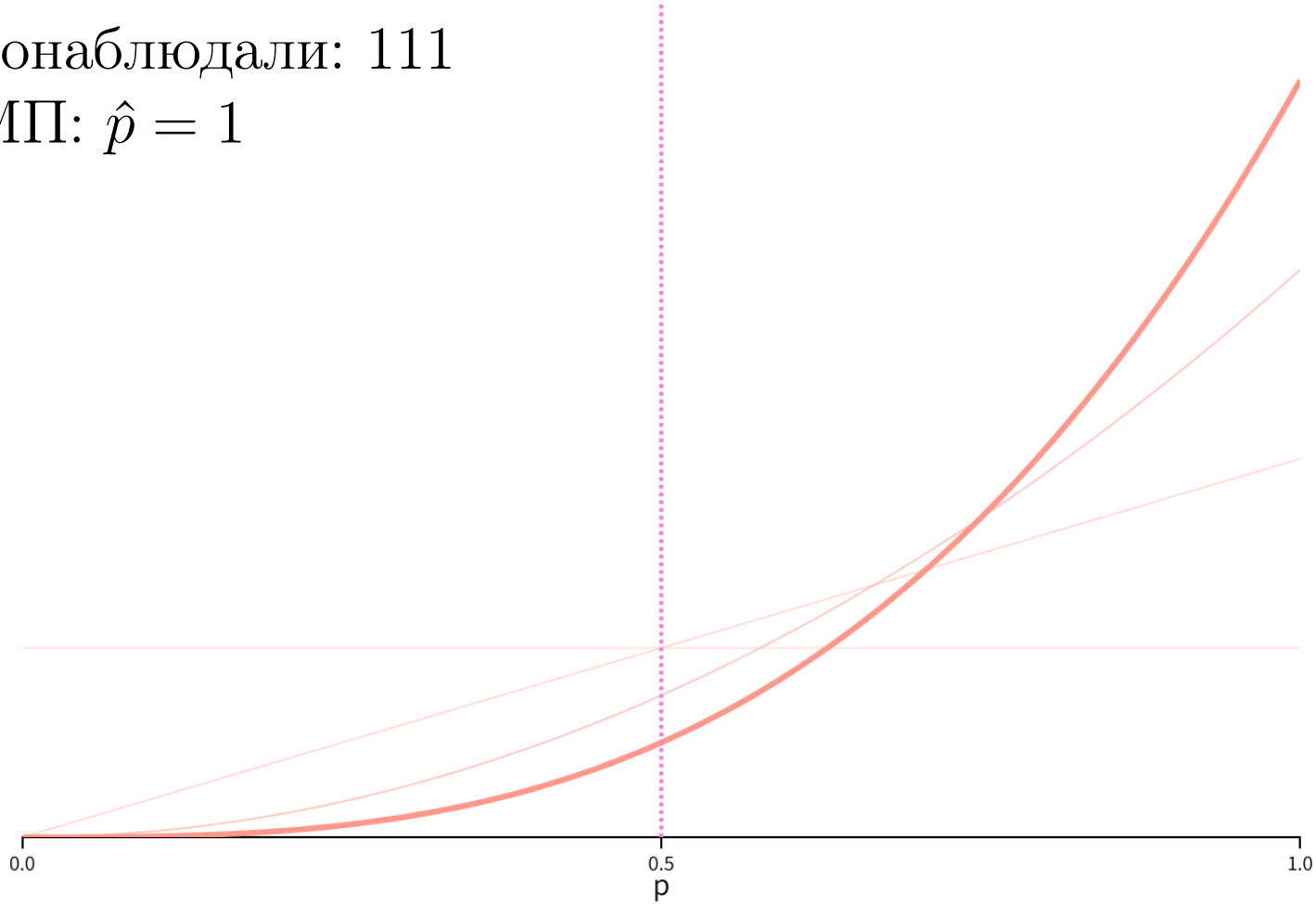
Пронаблюдали: 11
ОМП: $\hat{p} = 1$



Байесовский вывод для монетки: пример

Пусть $\rho(p) := \mathbb{1}_{[0,1]}(p)$ — равномерное распределение на $[0, 1]$, то есть априори мы ничего не знаем о параметре p.
Пусть истинное значение p равно 0.5, т.е. монетка честная.

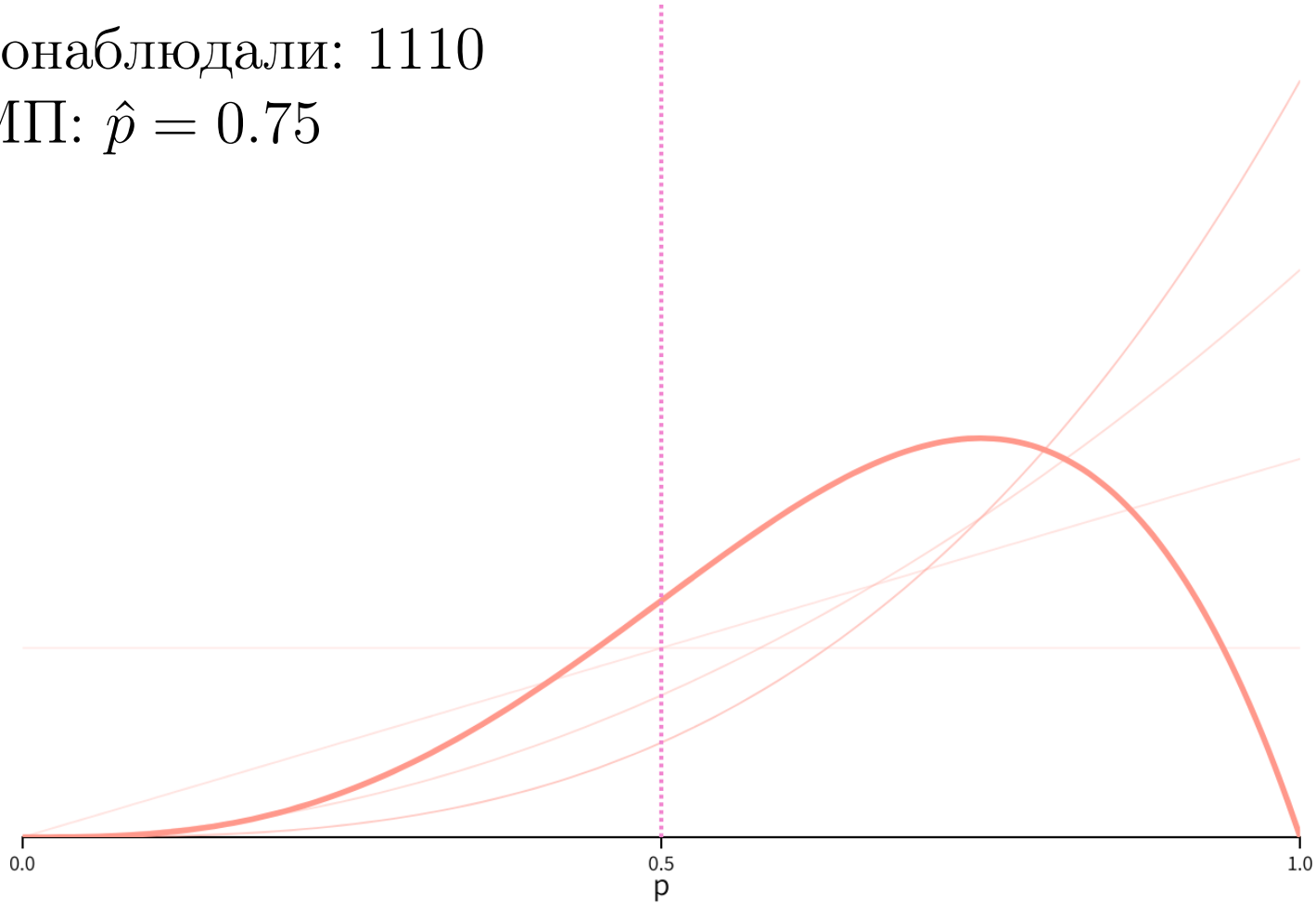
Пронаблюдали: 111
ОМП: $\hat{p} = 1$



Байесовский вывод для монетки: пример

Пусть $\rho(p) := \mathbb{1}_{[0,1]}(p)$ — равномерное распределение на $[0, 1]$, то есть априори мы ничего не знаем о параметре p.
Пусть истинное значение p равно 0.5, т.е. монетка честная.

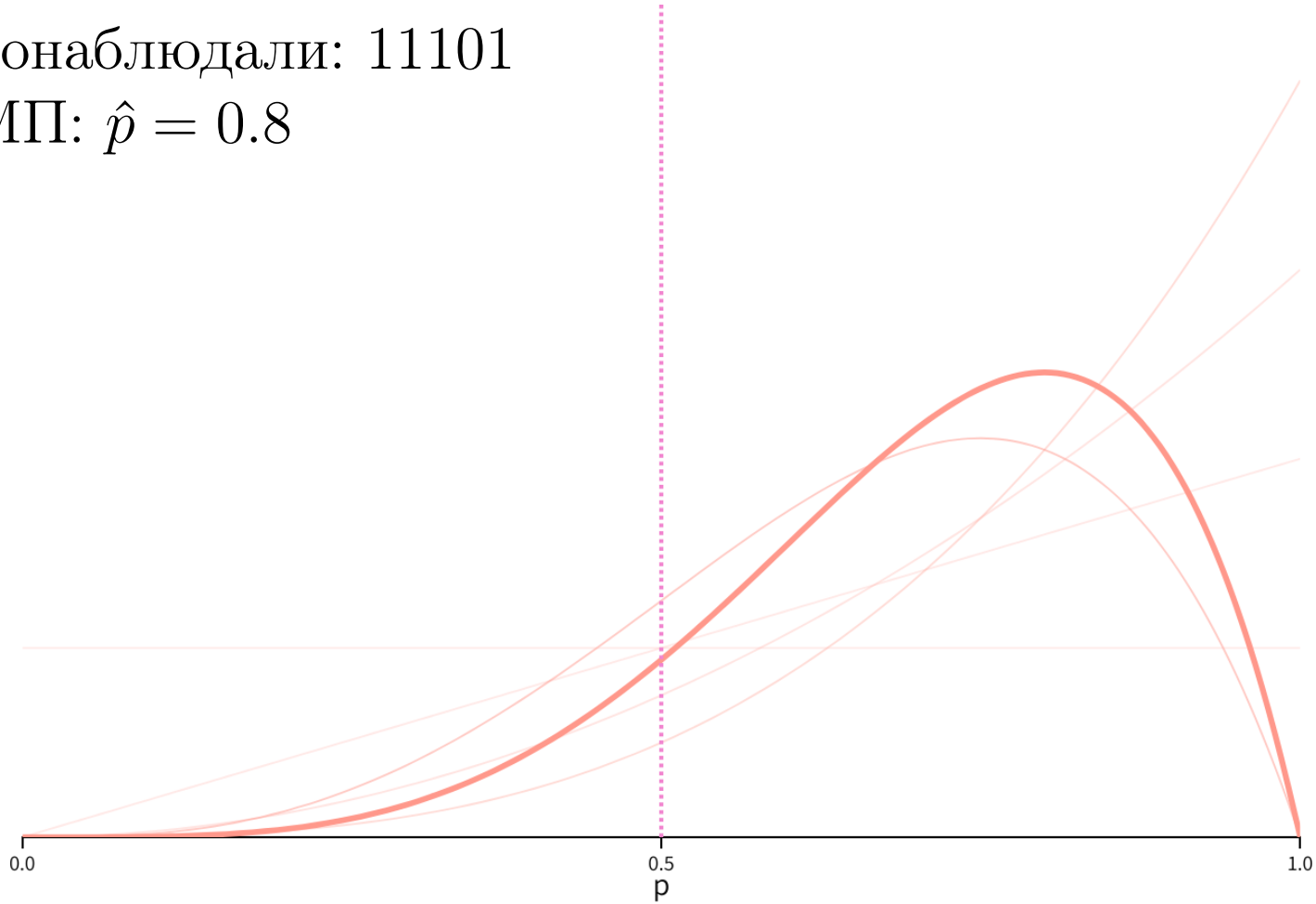
Пронаблюдали: 1110
ОМП: $\hat{p} = 0.75$



Байесовский вывод для монетки: пример

Пусть $\rho(p) := \mathbb{1}_{[0,1]}(p)$ — равномерное распределение на $[0, 1]$, то есть априори мы ничего не знаем о параметре p.
Пусть истинное значение p равно 0.5, т.е. монетка честная.

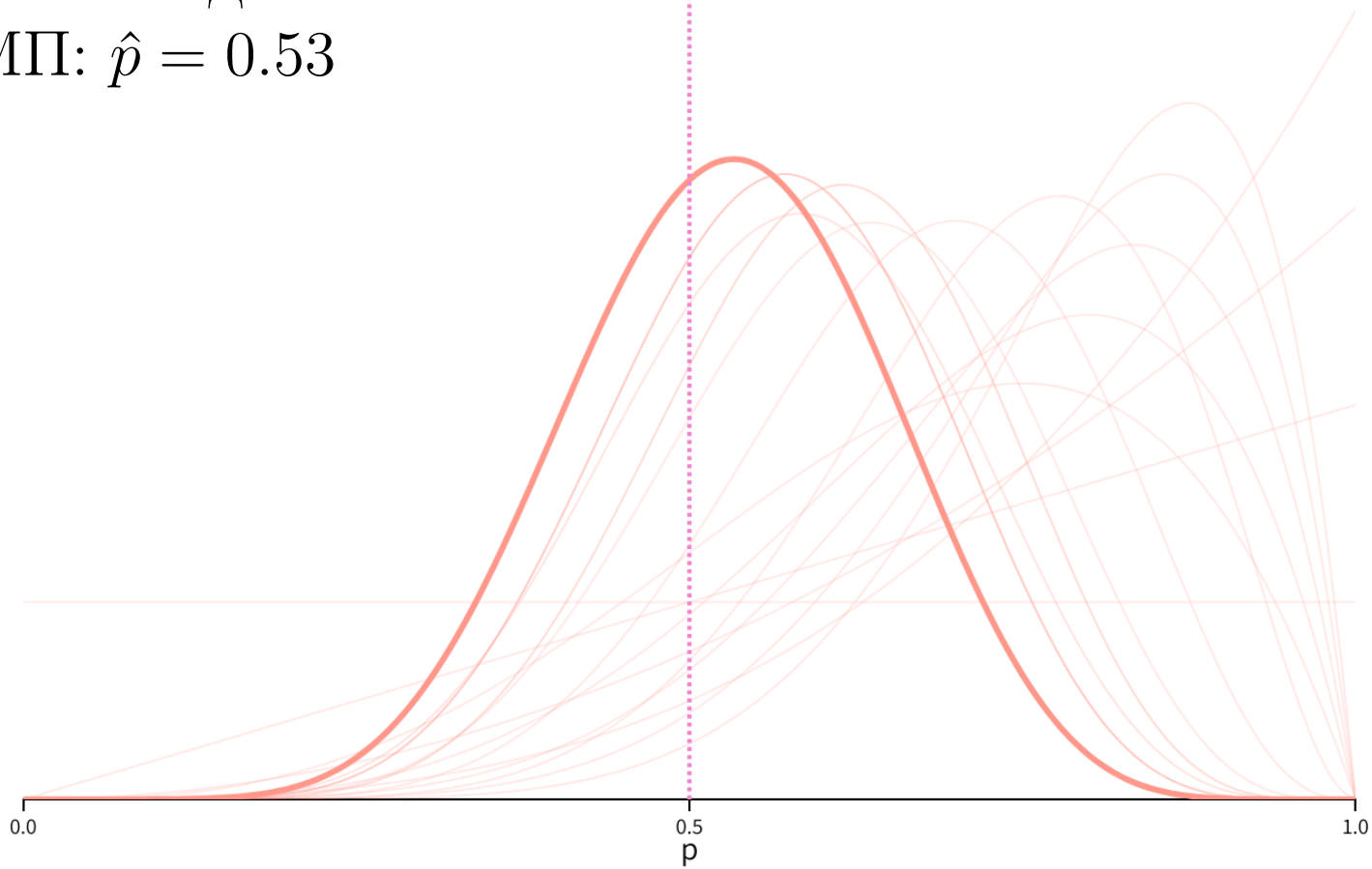
Пронаблюдали: 11101
ОМП: $\hat{p} = 0.8$



Байесовский вывод для монетки: пример

Пусть $\rho(p) := \mathbb{1}_{[0,1]}(p)$ — равномерное распределение на $[0, 1]$, то есть априори мы ничего не знаем о параметре p.
Пусть истинное значение p равно 0.5, т.е. монетка честная.

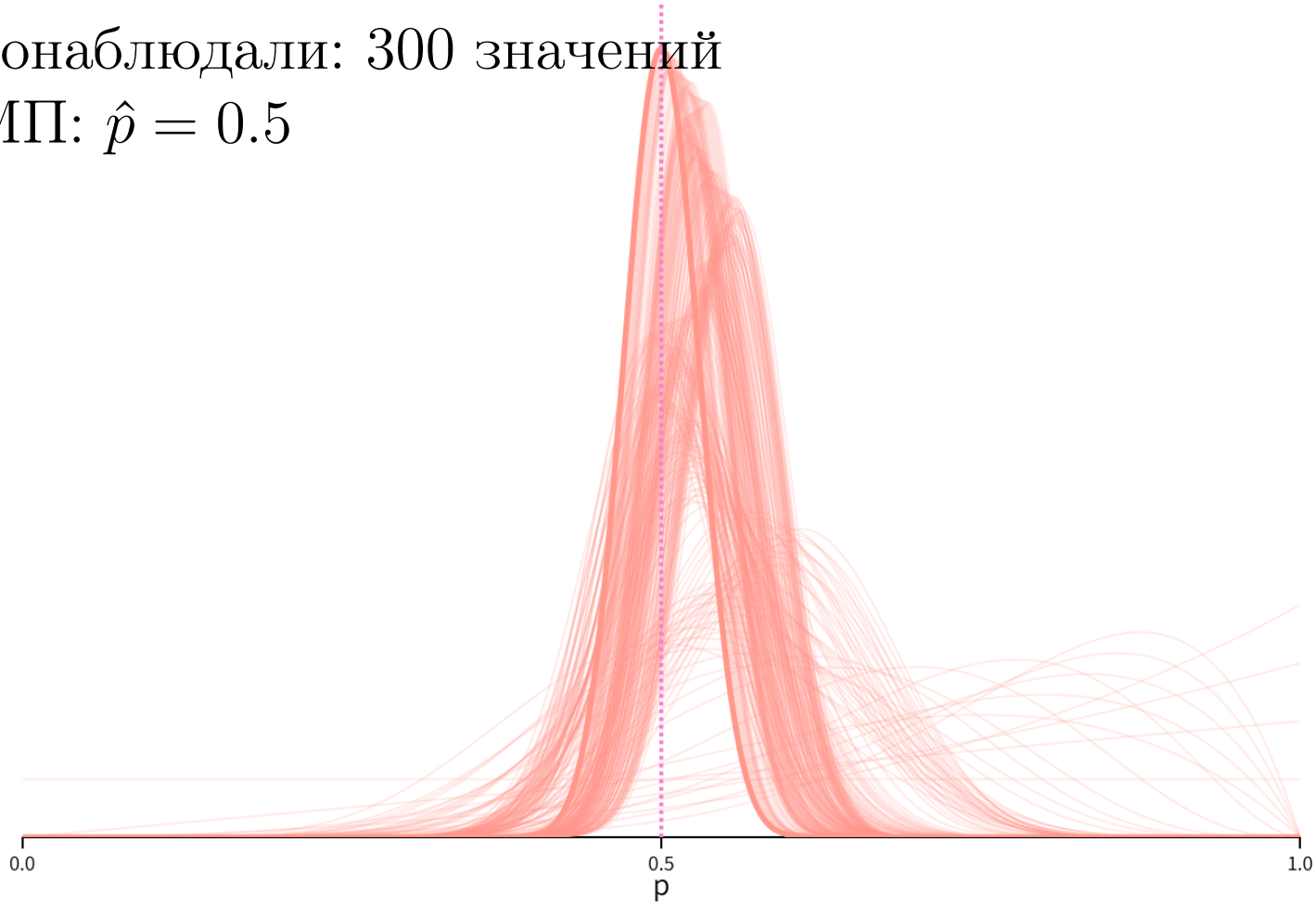
Пронаблюдали: 11101 10001 10100
ОМП: $\hat{p} = 0.53$



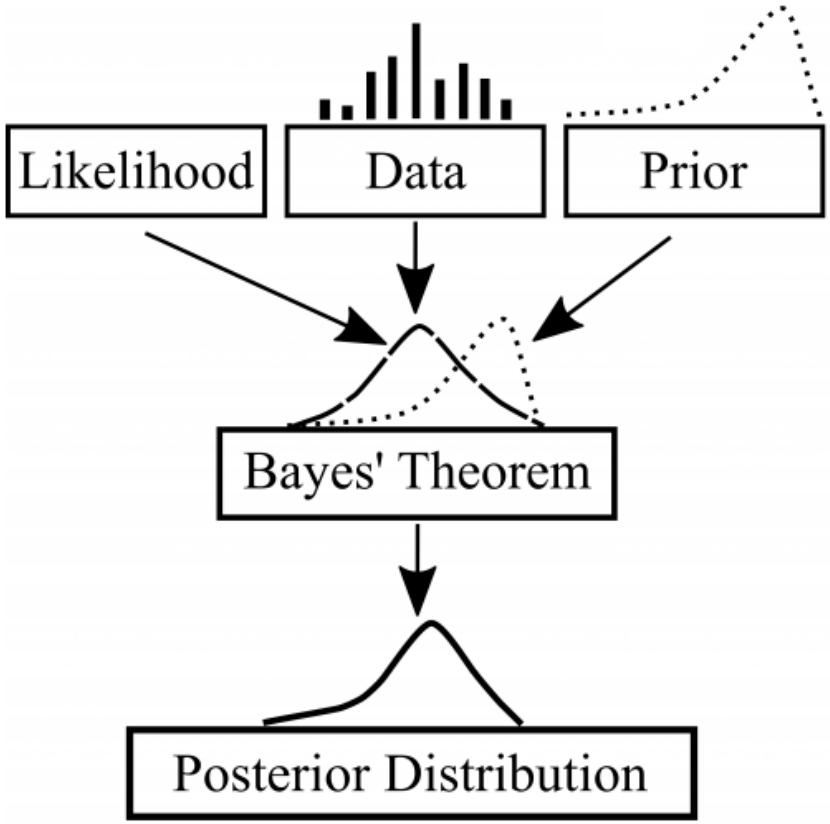
Байесовский вывод для монетки: пример

Пусть $\rho(p) := \mathbb{1}_{[0,1]}(p)$ — равномерное распределение на $[0, 1]$, то есть априори мы ничего не знаем о параметре p.
Пусть истинное значение p равно 0.5, т.е. монетка честная.

Пронаблюдали: 300 значений
ОМП: $\hat{p} = 0.5$



Байесовский вывод для монетки: итог



Важнейшим преимуществом байесовского подхода является оценка неопределенности. Меньше чуши с уверенным лицом.

Содержание

Байесовский вывод для ГП: суть

Гауссовский процесс — распределение на функциях.

Модель:

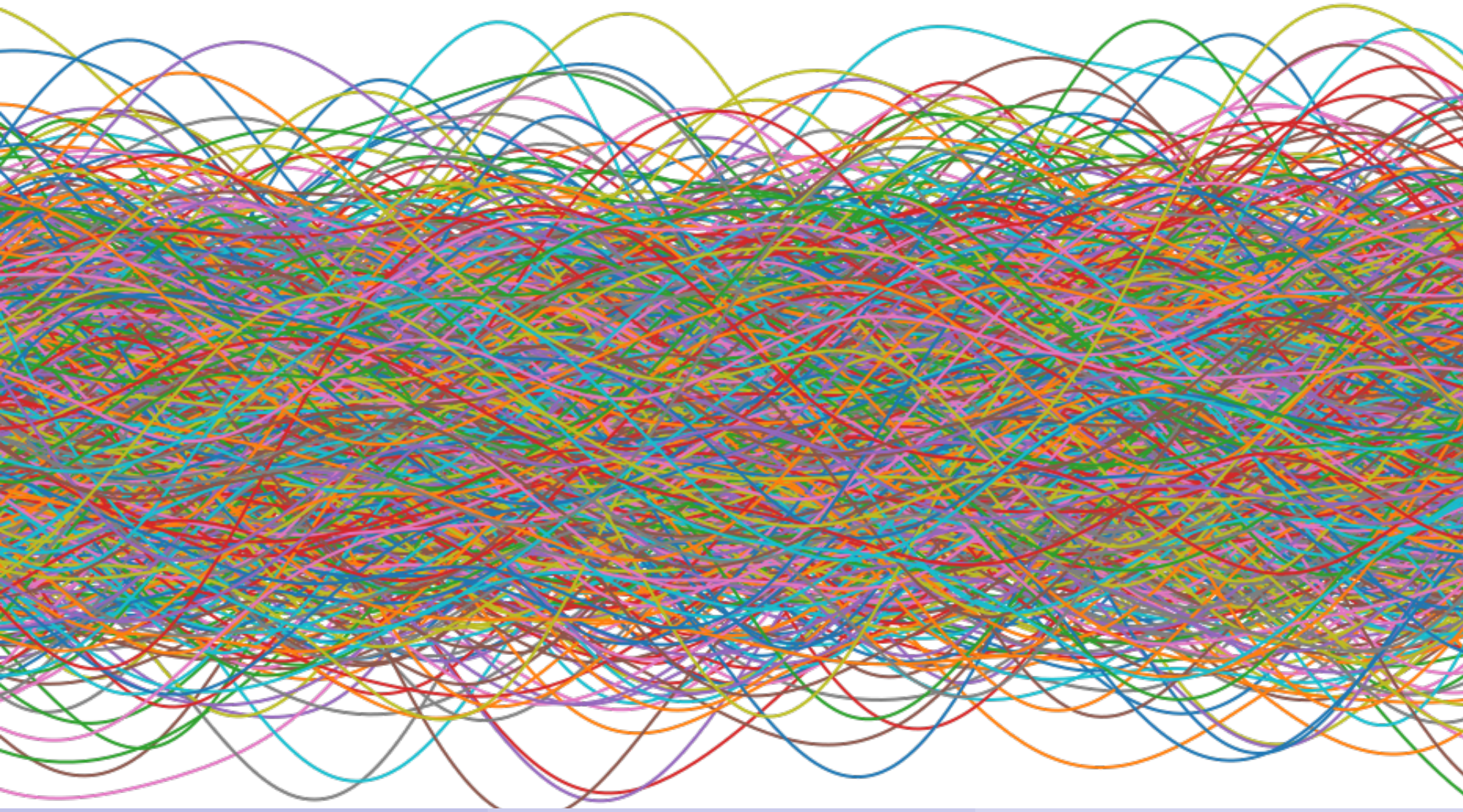
- Априорное распределение: гауссовский процесс.
- Правдоподобие задается нормальным распределением.

Данные: значения функции в каких-то точках.

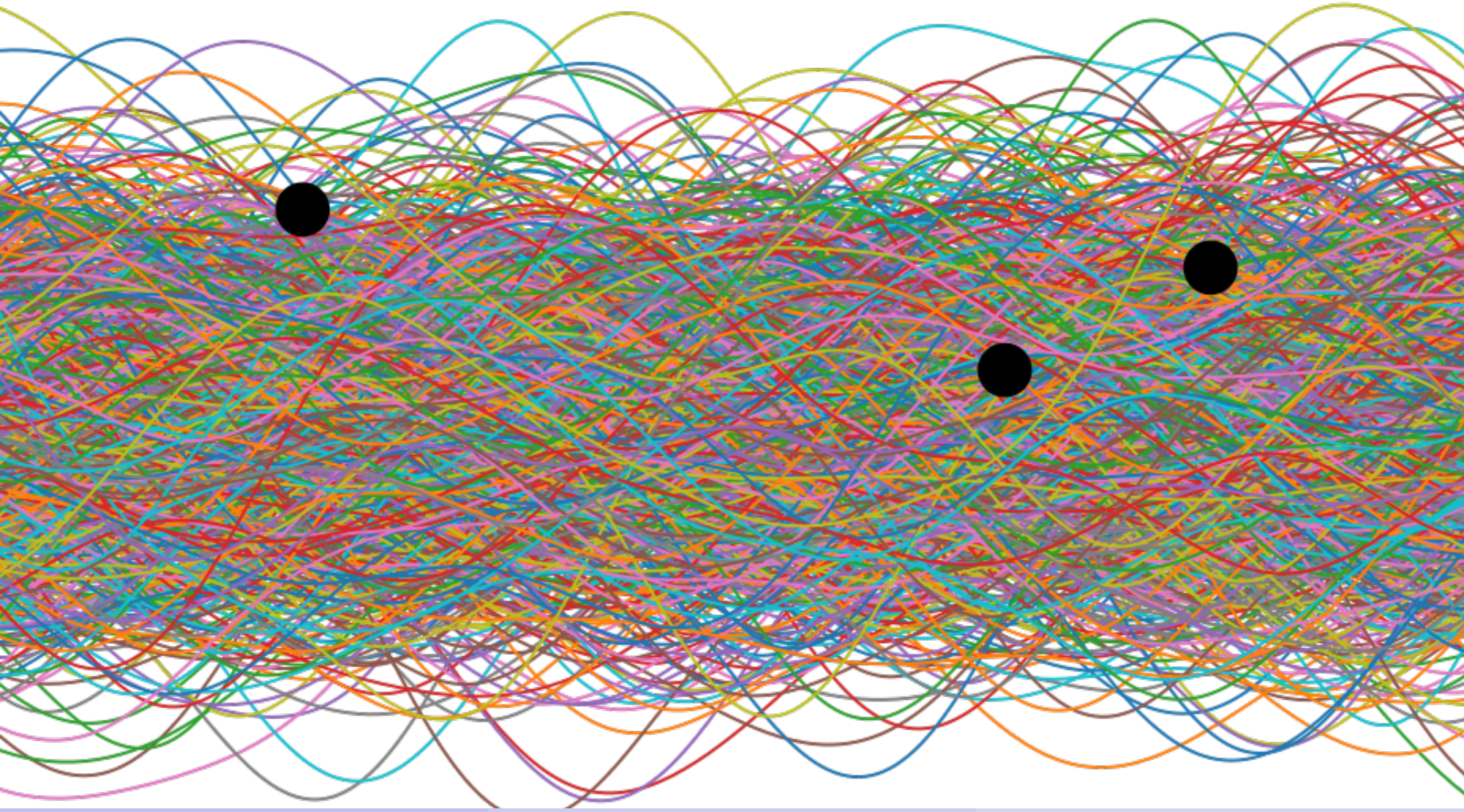
Результат (апостериорное распределение): тоже гауссовский процесс, но другой.

Байесовский вывод для ГП: иллюстрация

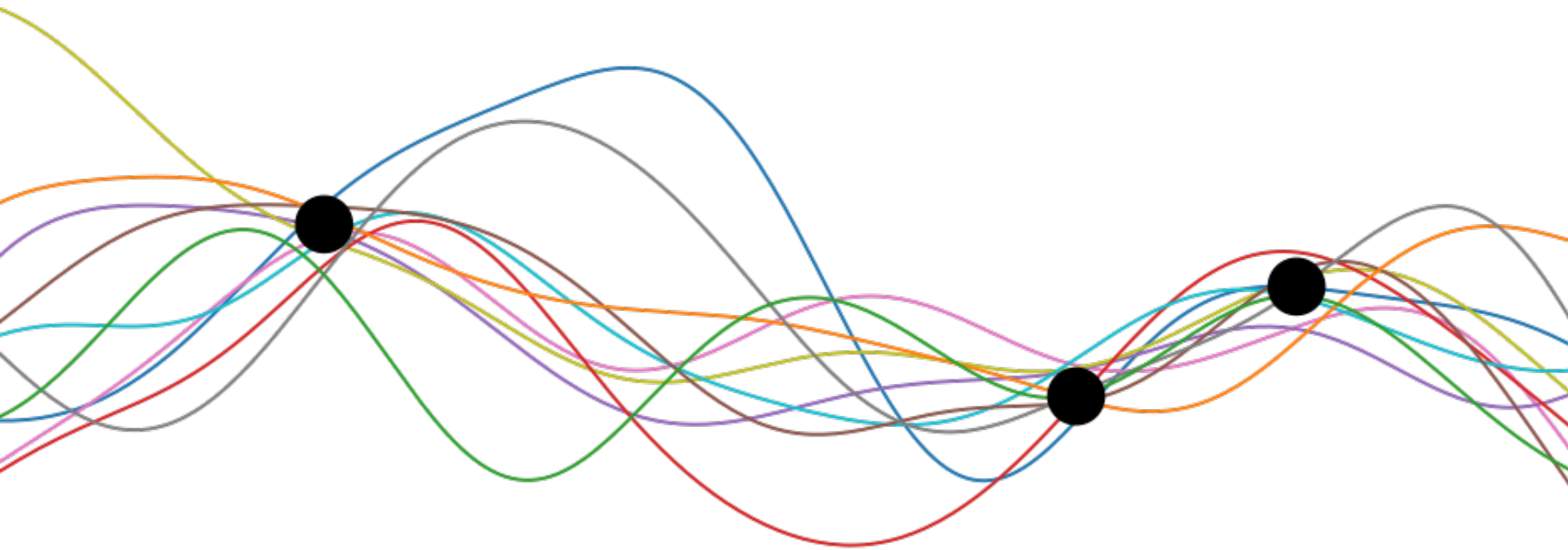
Байесовский вывод для ГП: иллюстрация



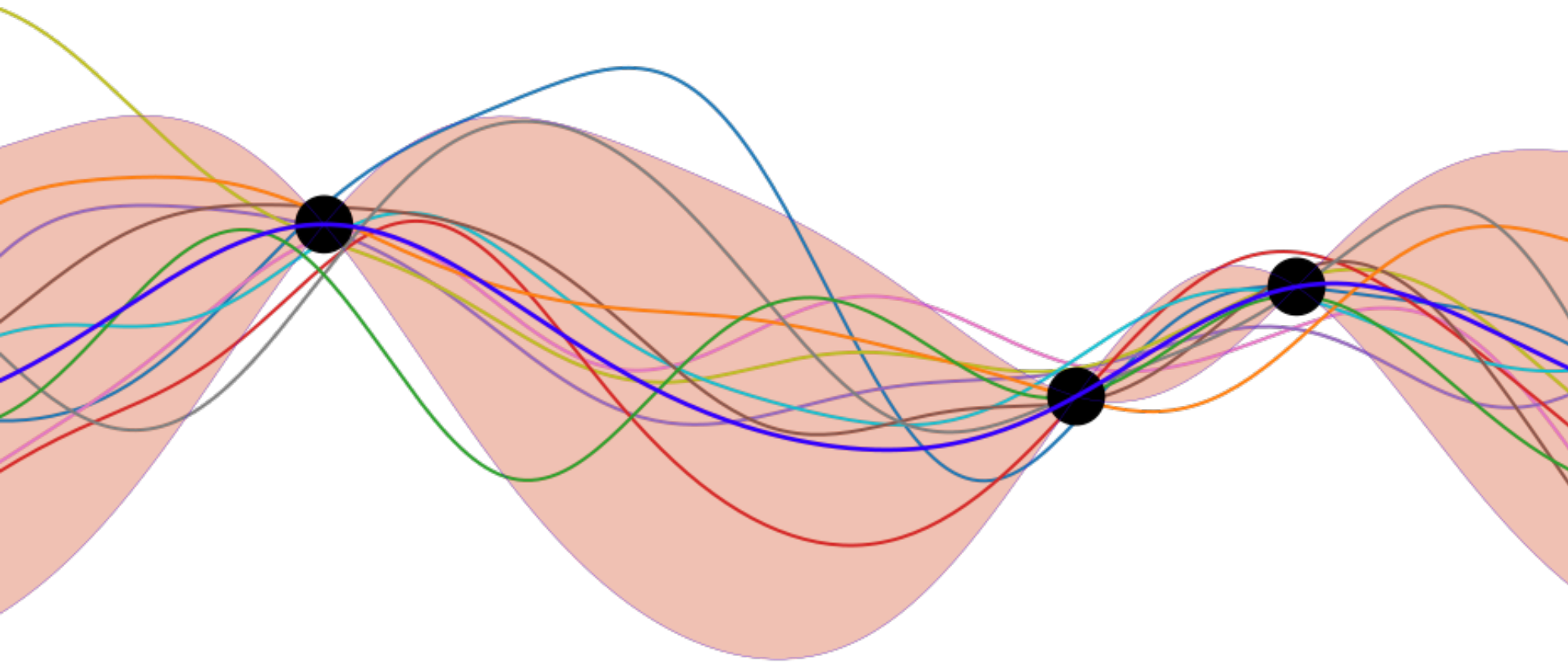
Байесовский вывод для ГП: иллюстрация



Байесовский вывод для ГП: иллюстрация



Байесовский вывод для ГП: иллюстрация



Байесовский вывод для ГП: что такое ГП?

Гауссовское распределение

- распределение на $\mathbb{R}$, обозначается $\mathrm{N}(\mu, \sigma^2)$,
- определяется двумя числами: средним μ и дисперсией σ^2.

Многомерное гауссовское распределение

- распределение на $\mathbb{R}^d$, обозначается $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,
- определяется вектором средних $\boldsymbol{\mu}$ и матрицей ковариаций $\boldsymbol{\Sigma}$.

Гауссовский процесс

- распределение функциях из X в $\mathbb{R}$, обозначается $\mathrm{GP}(m, k)$,
- определяется функциями $m : X \to \mathbb{R}$ (средним) и $k : X \times X \to \mathbb{R}$ (ковариацией).

Гауссовский процесс — одно из многих распределений на функциях (случайных процессов).

Оно выделяется простотой вычислений.

Байесовский вывод для ГП: главное

Байесовский вывод для ГП по

- априорному распределению на функции вида $GP(m, k)$,
- значениям $y_1, .., y_n$ моделируемой функции в точках $x_1, .., x_n$, "измеренным" с точн. до шума с дисперсией σ^2.

дает апостериорное распределение на функции, которое имеет вид

$$GP(\hat{m}, \hat{k}).$$

Значения функций $\hat{m}$ и $\hat{K}$ пересчитываются из значений функций m и k за конечное время. Более точно:

$$\tilde{m}(u) = m(u) + \mathbf{K}_{f(u)f(\boldsymbol{x})}\Big(\mathbf{K}_{f(\boldsymbol{x})f(\boldsymbol{x})} + \sigma^2 I\Big)^{-1}(\boldsymbol{y} - m(\boldsymbol{x}))$$

$$\tilde{k}(u, v) = k(u, v) - \underbrace{\mathbf{K}_{f(u)f(\boldsymbol{x})}}_{\text{вектор } 1\times n}\underbrace{\Big(\mathbf{K}_{f(\boldsymbol{x})f(\boldsymbol{x})} + \sigma^2 I\Big)^{-1}}_{\text{матрица } n\times n}\underbrace{\mathbf{K}_{f(\boldsymbol{x})f(v)}}_{\text{вектор } n\times 1}.$$

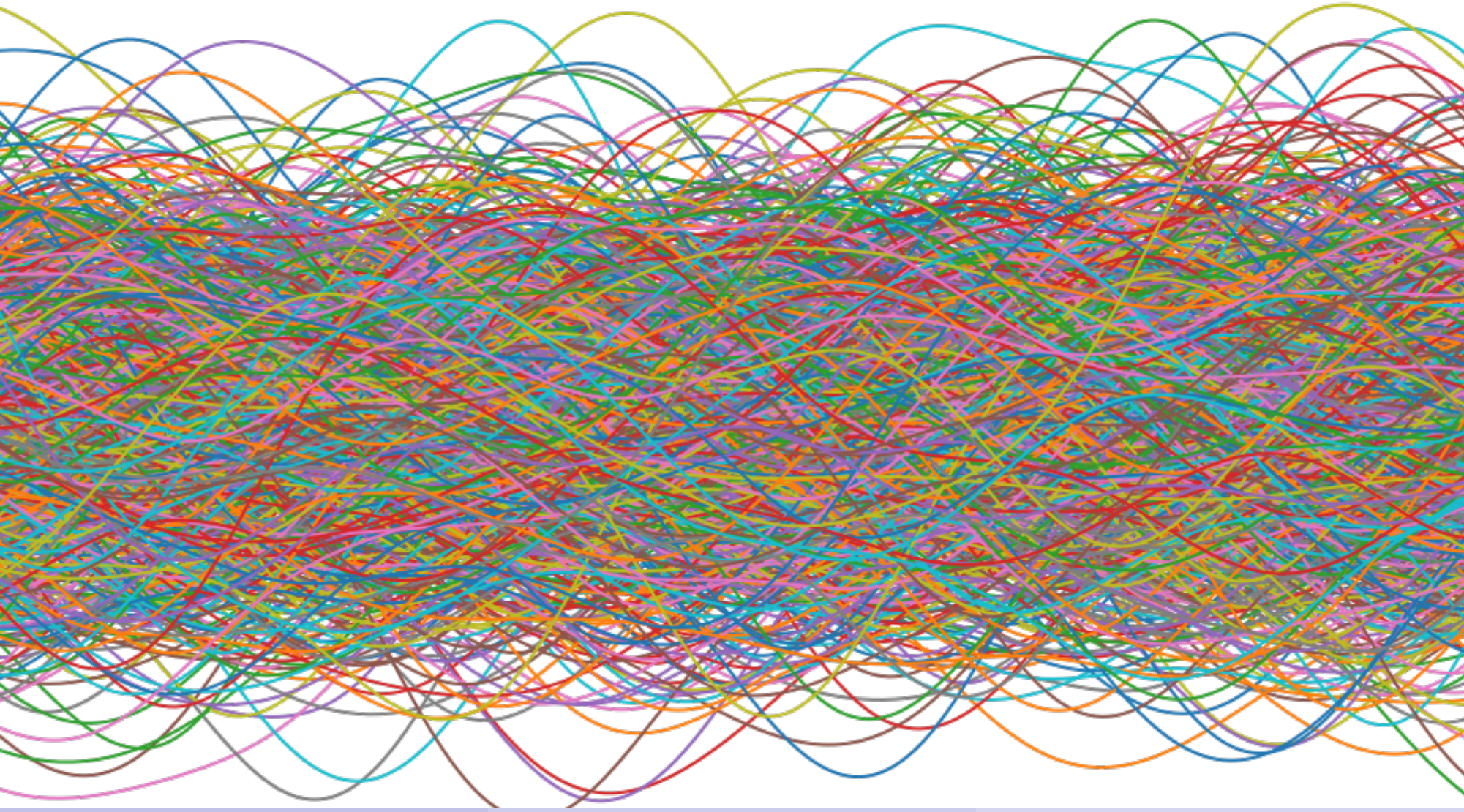
Байесовский вывод для ГП: алгоритм

Как получить по данным $(x_1, y_1), .., (x_n, y_n)$ разумную стохастическую модель, их интерполирующую?

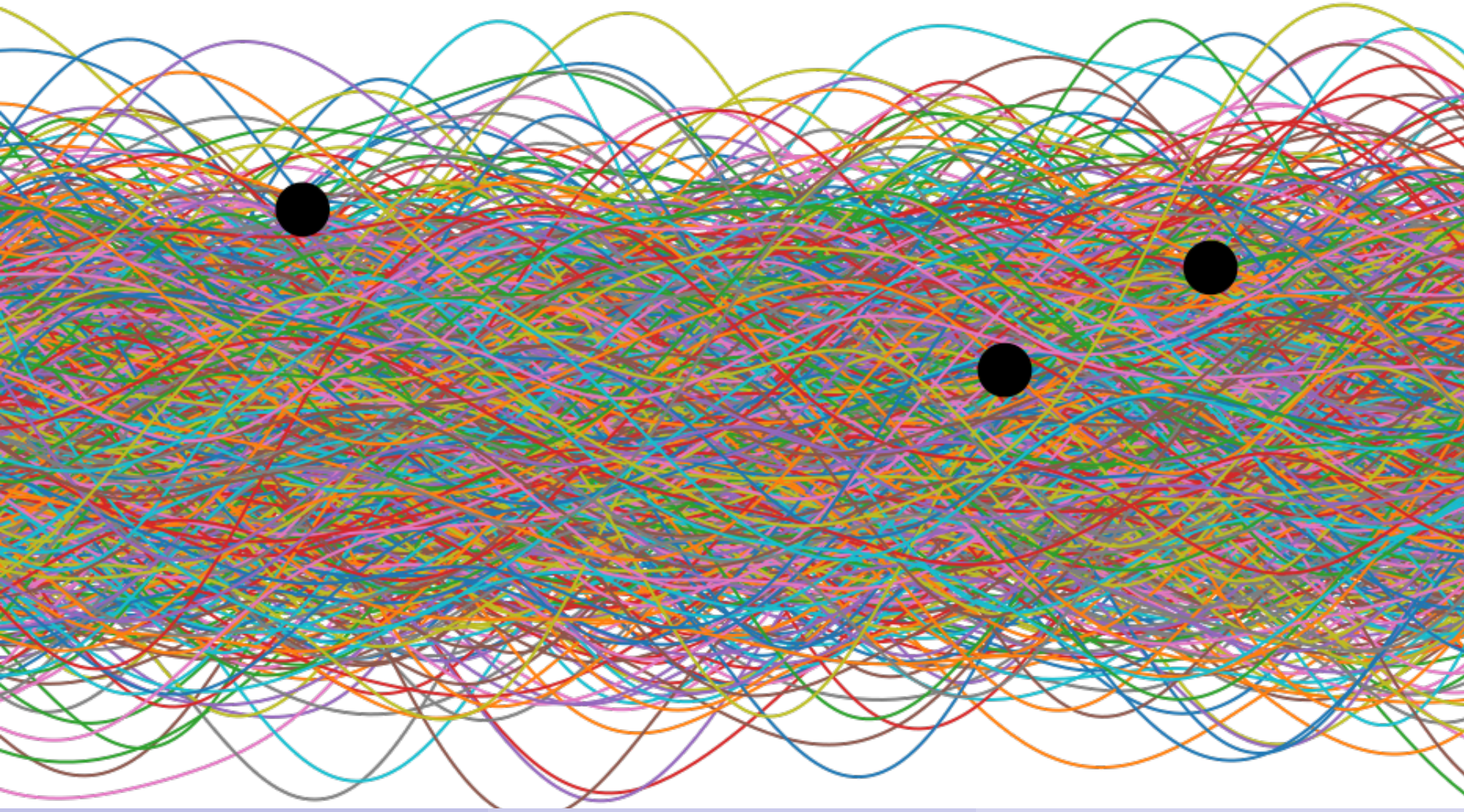
1. Взять с неба априорное среднее m и параметрическое семейство $k_{\boldsymbol{\theta}}$ априорных ковариационных функций.
2. По данным $(x_1, y_1), .., (x_n, y_n)$ выбрать оптимальное значение параметров $\boldsymbol{\theta}$ и шума σ^2.
3. Байесовский вывод с априорными m и $k_{\boldsymbol{\theta}}$ по данным $(x_1, y_1), .., (x_n, y_n)$ с шумом σ^2.
   В результате имеем апостериорные $\tilde{m}$ и $\tilde{k}$.
4. Использовать
   - $\mathrm{N}(\tilde{m}(u), \tilde{k}(u, u))$ как стохастический прогноз в точке u.
   - сэмплы апостериорного процесса $GP(\tilde{m}, \tilde{k})$ как ансамбль возможных детерминистических моделей.

Байесовский вывод для ГП: иллюстрация

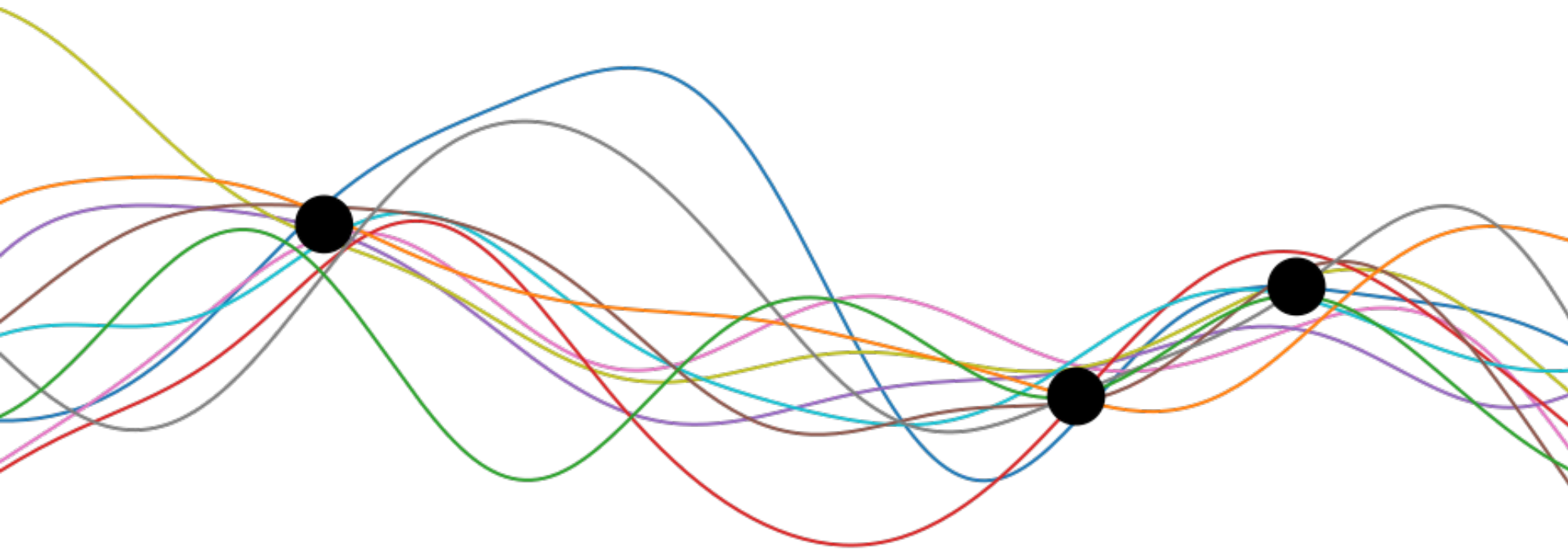
Байесовский вывод для ГП: иллюстрация



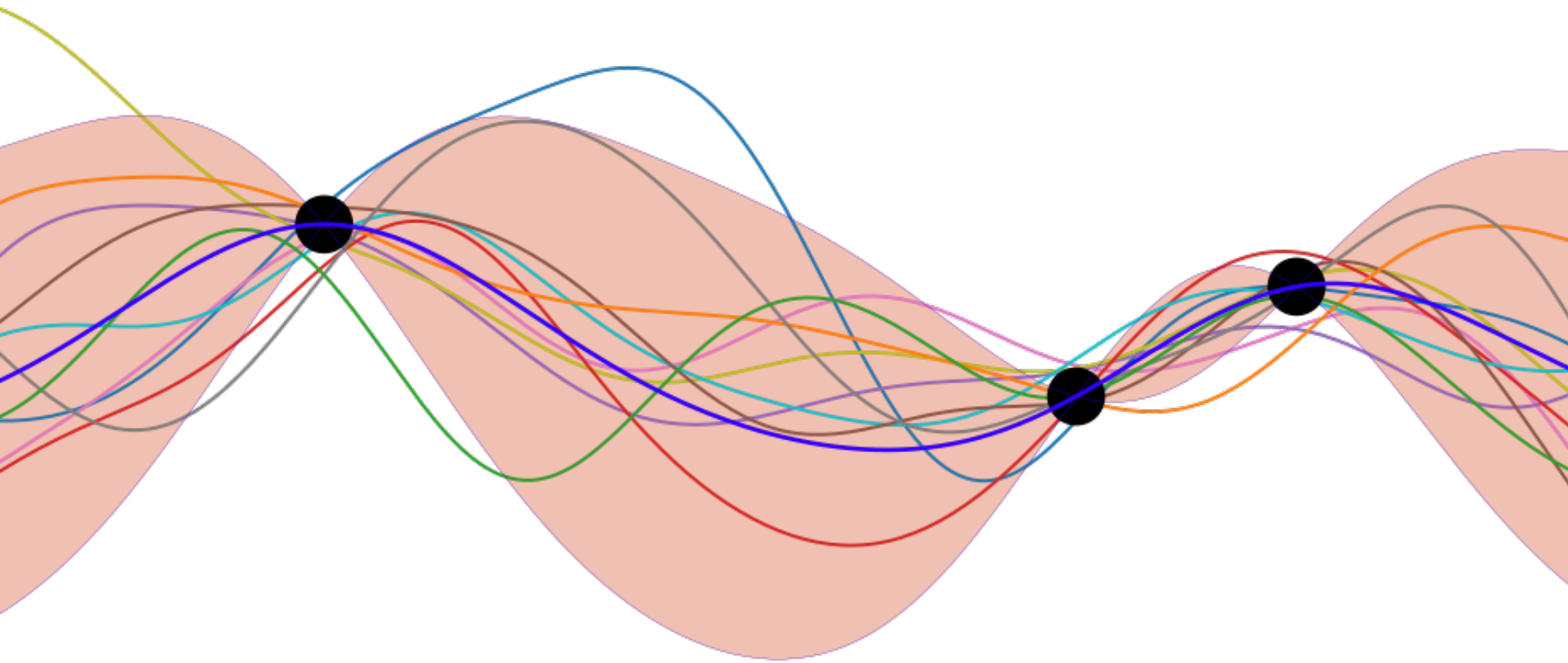
Байесовский вывод для ГП: иллюстрация



Байесовский вывод для ГП: иллюстрация



Байесовский вывод для ГП: иллюстрация

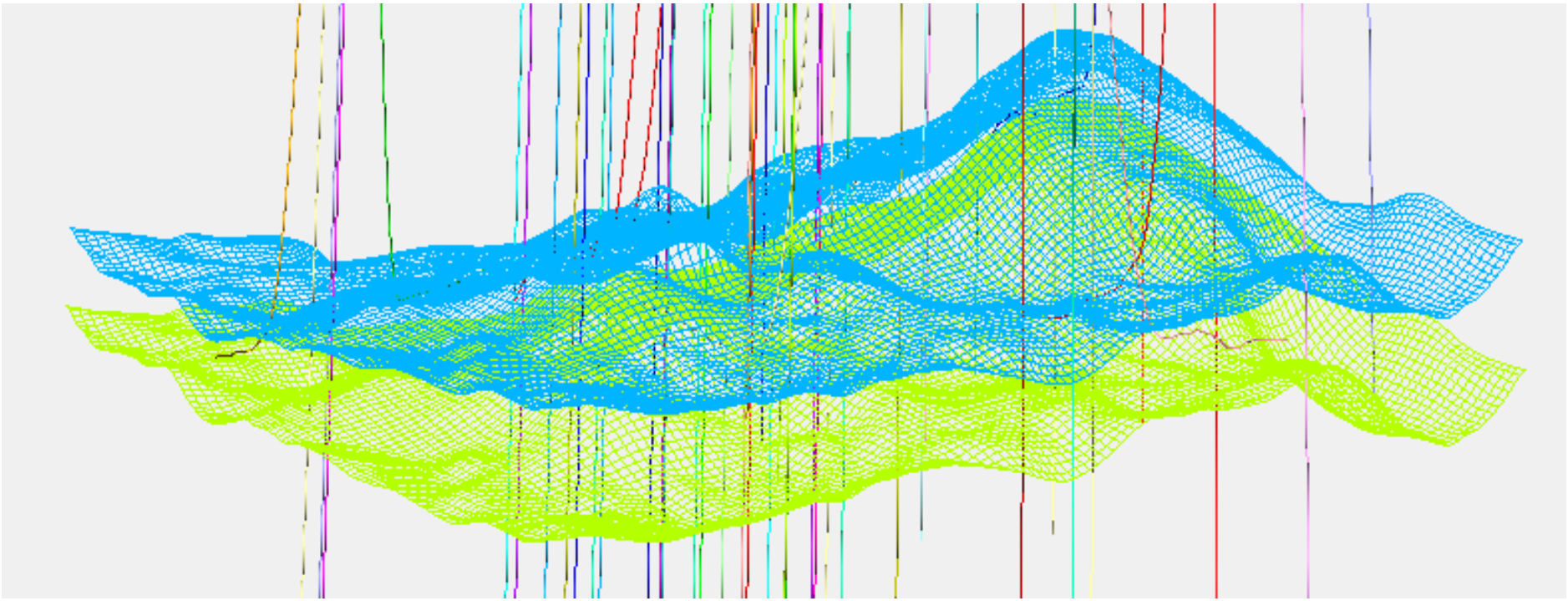


Содержание

Геостатистическое моделирование

Задача: интерполяция данных со скважин в межскважинное пространство.

Данных очень мало, поэтому детерминистический прогноз – плохо, а стохастический – хорошо.

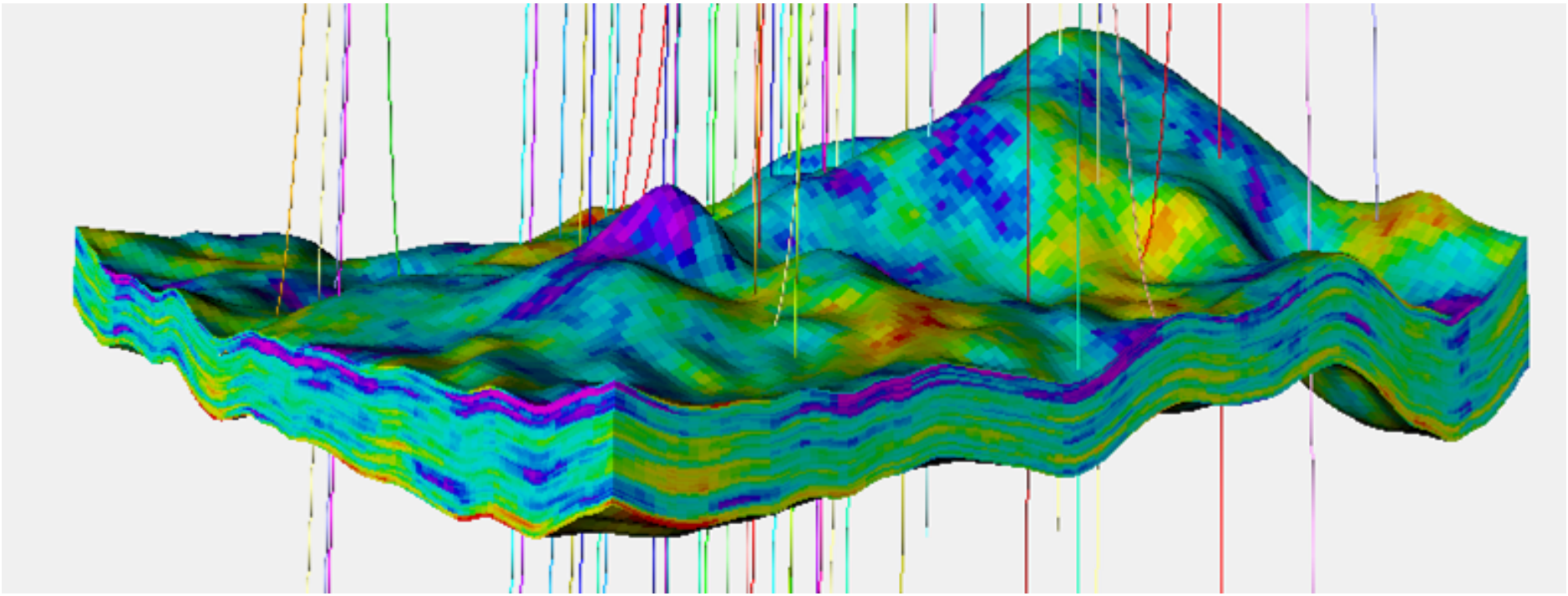


Структура пласта, расположение скважин.

Геостатистическое моделирование

Задача: интерполяция данных со скважин в межскважинное пространство.

Данных очень мало, поэтому детерминистический прогноз – плохо, а стохастический – хорошо.



Сэмпл гауссовского процесса в межскважинном пространстве

Содержание

Байесовская оптимизация: суть

Задача: минимизация сложно-вычислимой black-box function.

Пусть на шаге n оптимизируемая функция $\phi : \mathbb{R}^d \to \mathbb{R}$ вычислена в точках $x_1, .., x_n$. Как выбрать x_{n+1}?

Строим условный гауссовский процесс f по данным

$$x_1, .., x_n, \qquad \phi(x_1), .., \phi(x_n).$$

Упорядочим $x_1, .., x_n$ так, чтобы $\phi(x_1) \geq \phi(x_2) \geq .. \geq \phi(x_n)$.

Выбираем

$$x_{n+1} = \arg\max_{x \in \mathbb{R}^d} \mathbb{P}(f(x) < \phi(x_n)). \qquad (MPI)$$

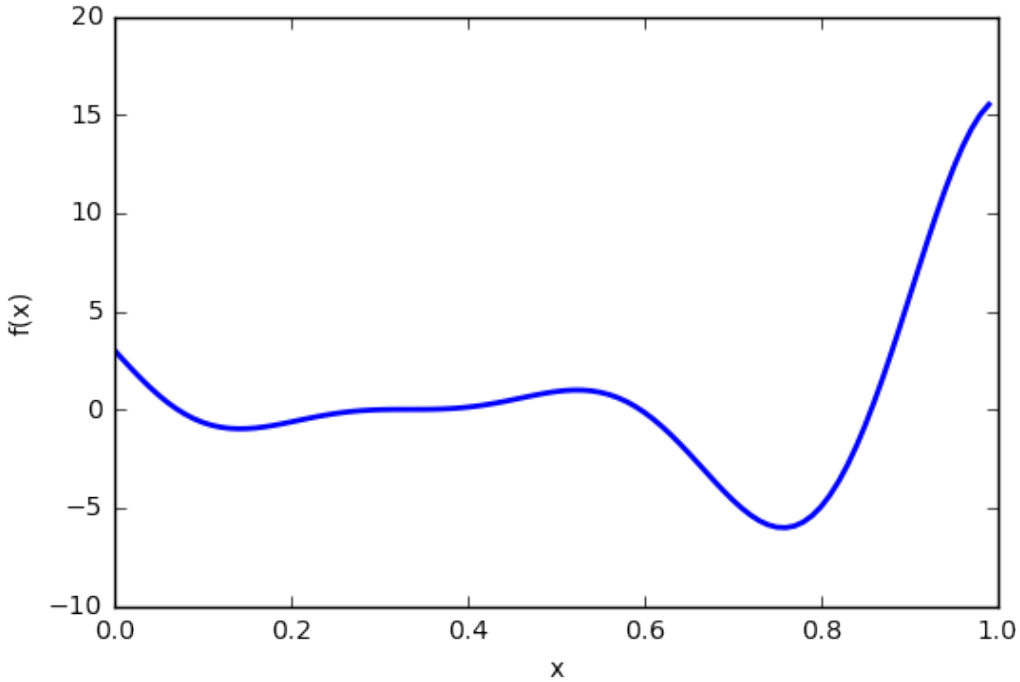
или

$$x_{n+1} = \arg\max_{x \in \mathbb{R}^d} \mathbb{E} \max(\phi(x_n) - f(x), 0). \qquad (EI)$$

Автоматический Exploration/exploitation trade-off.

Байесовская оптимизация: пример

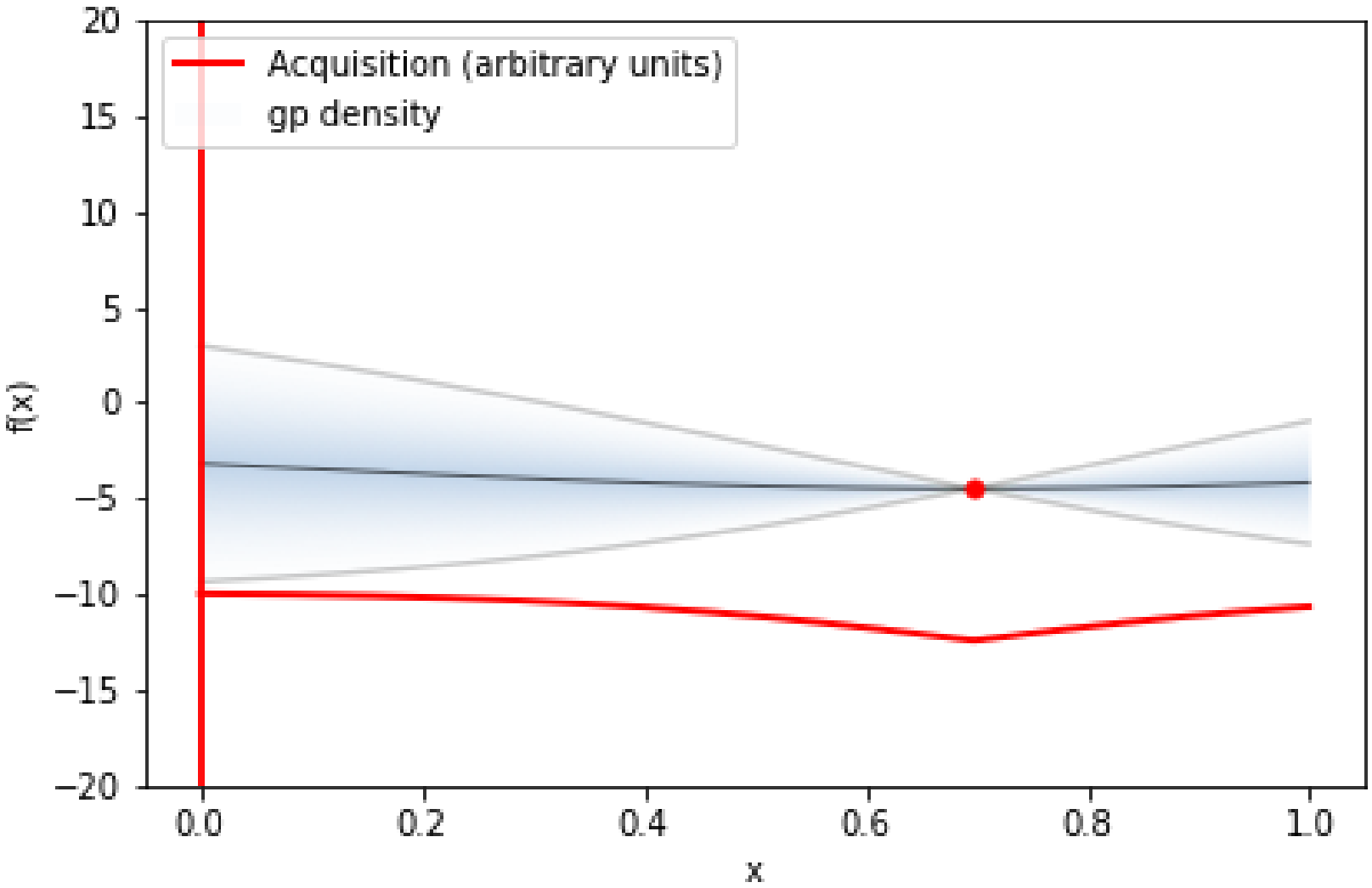
Попробуем найти минимум функции Форрестера
$f(x) = (6x - 2)^2 \sin(12x - 4)$.



Априорный процесс выберем как $f_0 \sim \mathrm{GP}(0, \mathrm{Matern}_{5/2})$

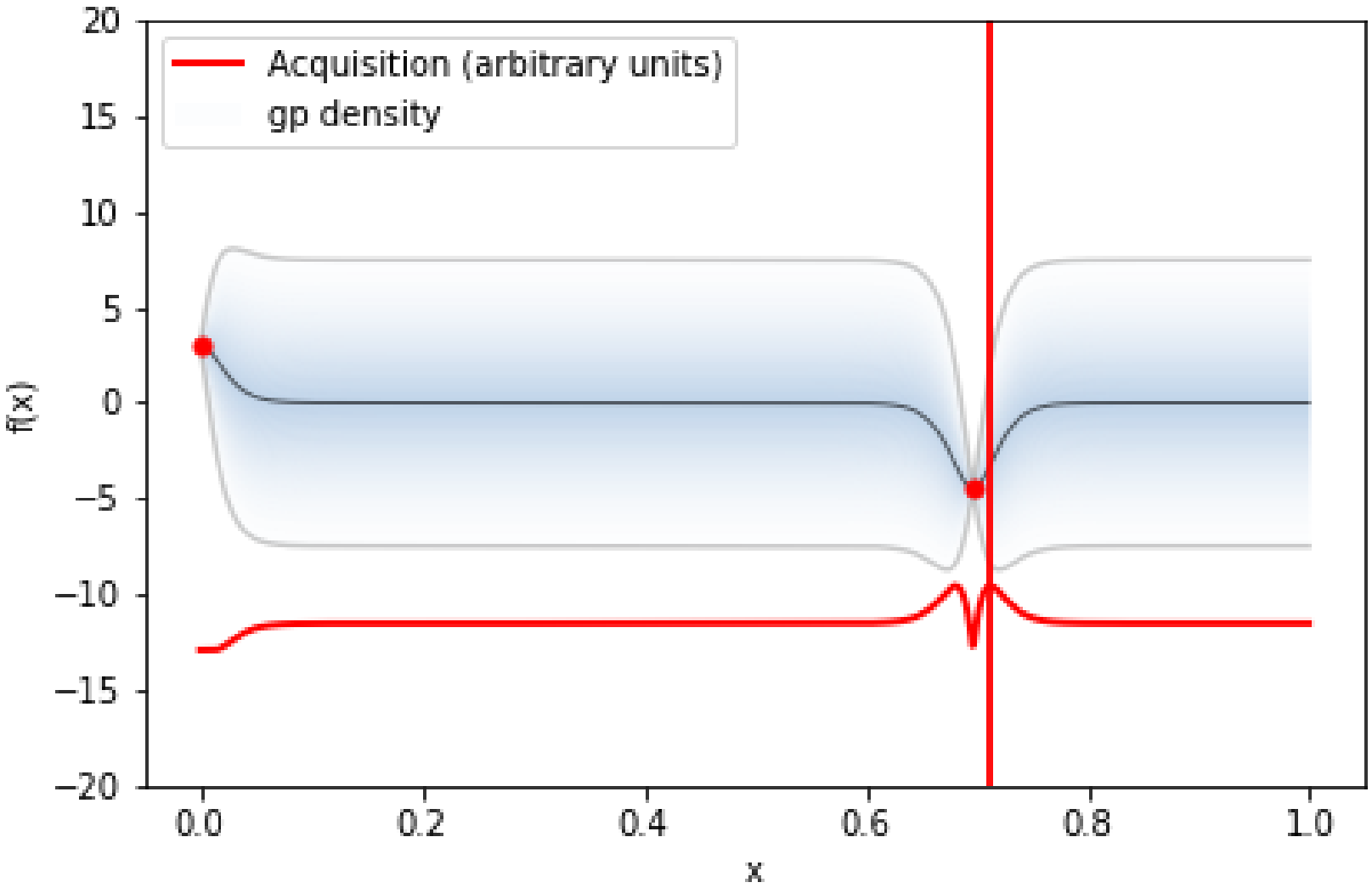
Байесовская оптимизация: пример

Итерация 1.



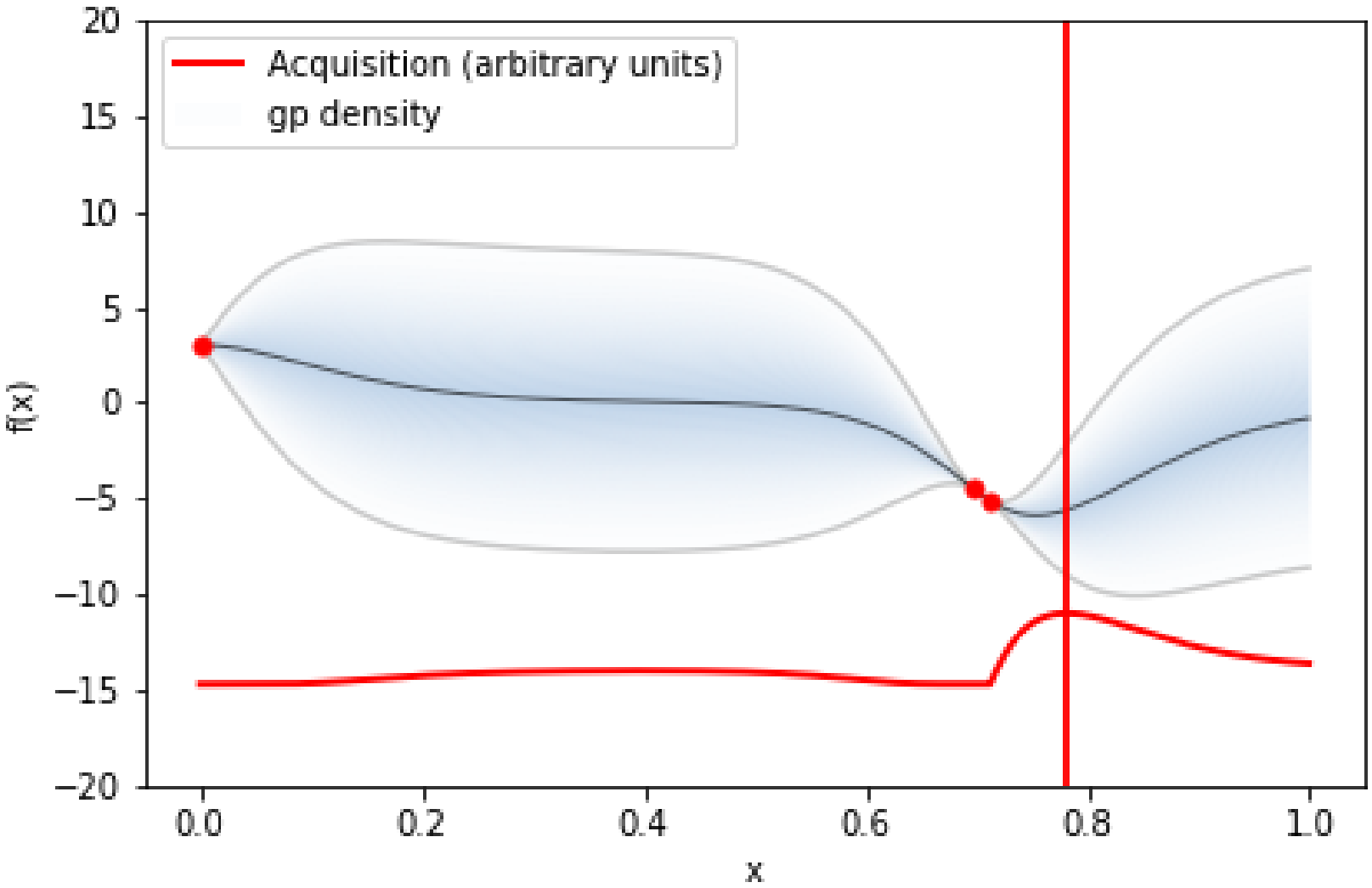
Байесовская оптимизация: пример

Итерация 2.



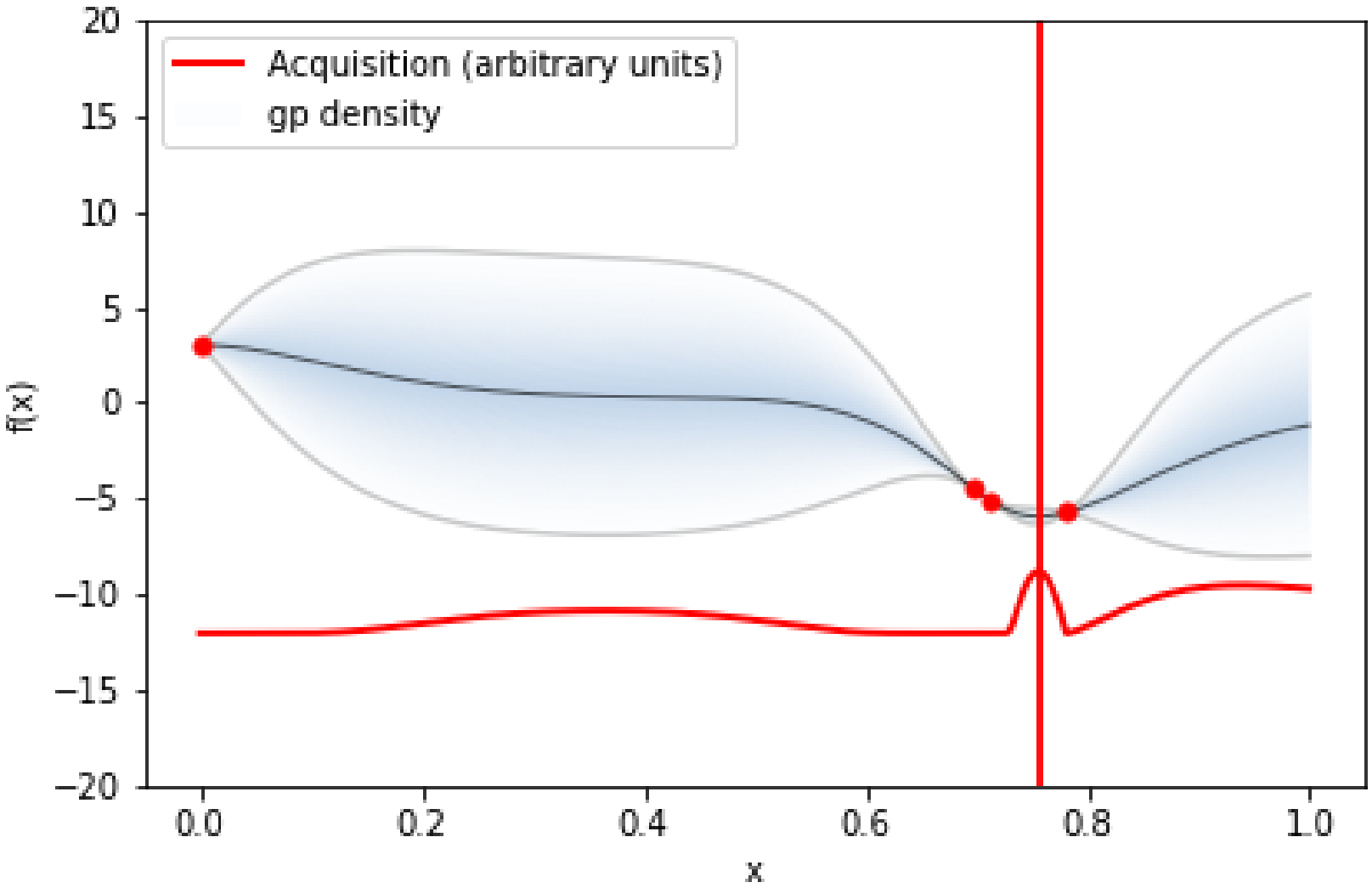
Байесовская оптимизация: пример

Итерация 3.



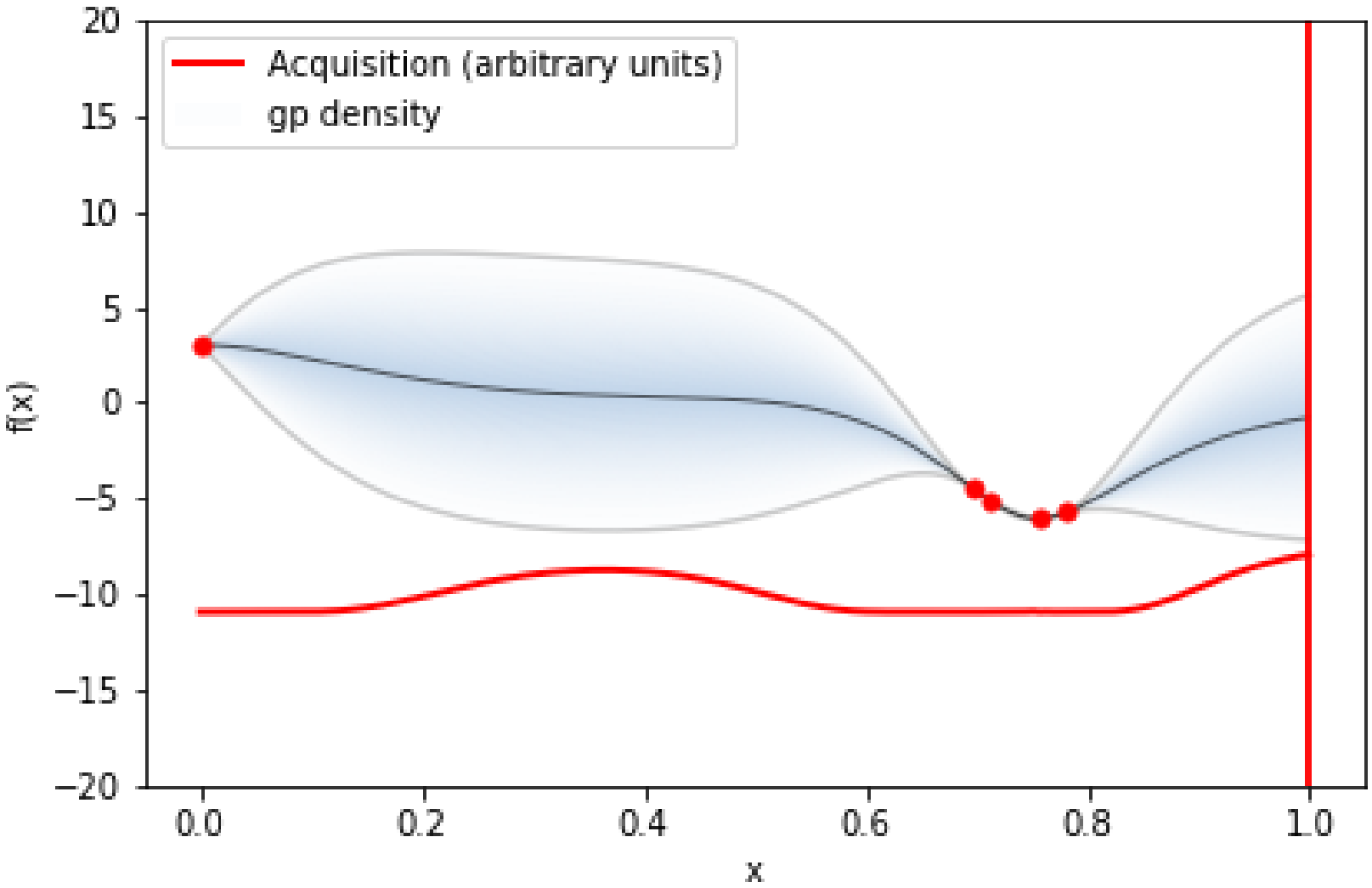
Байесовская оптимизация: пример

Итерация 4.



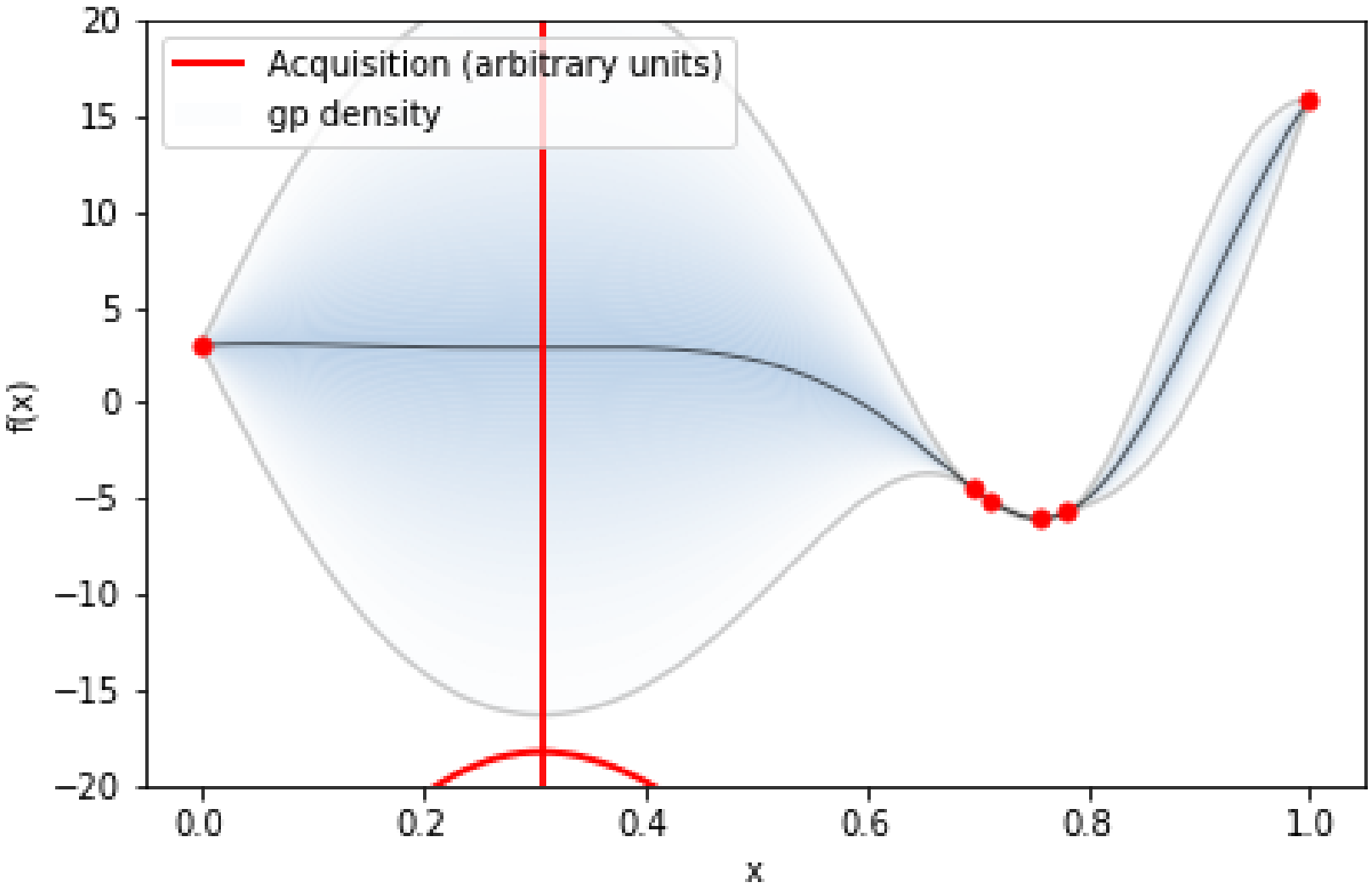
Байесовская оптимизация: пример

Итерация 5.



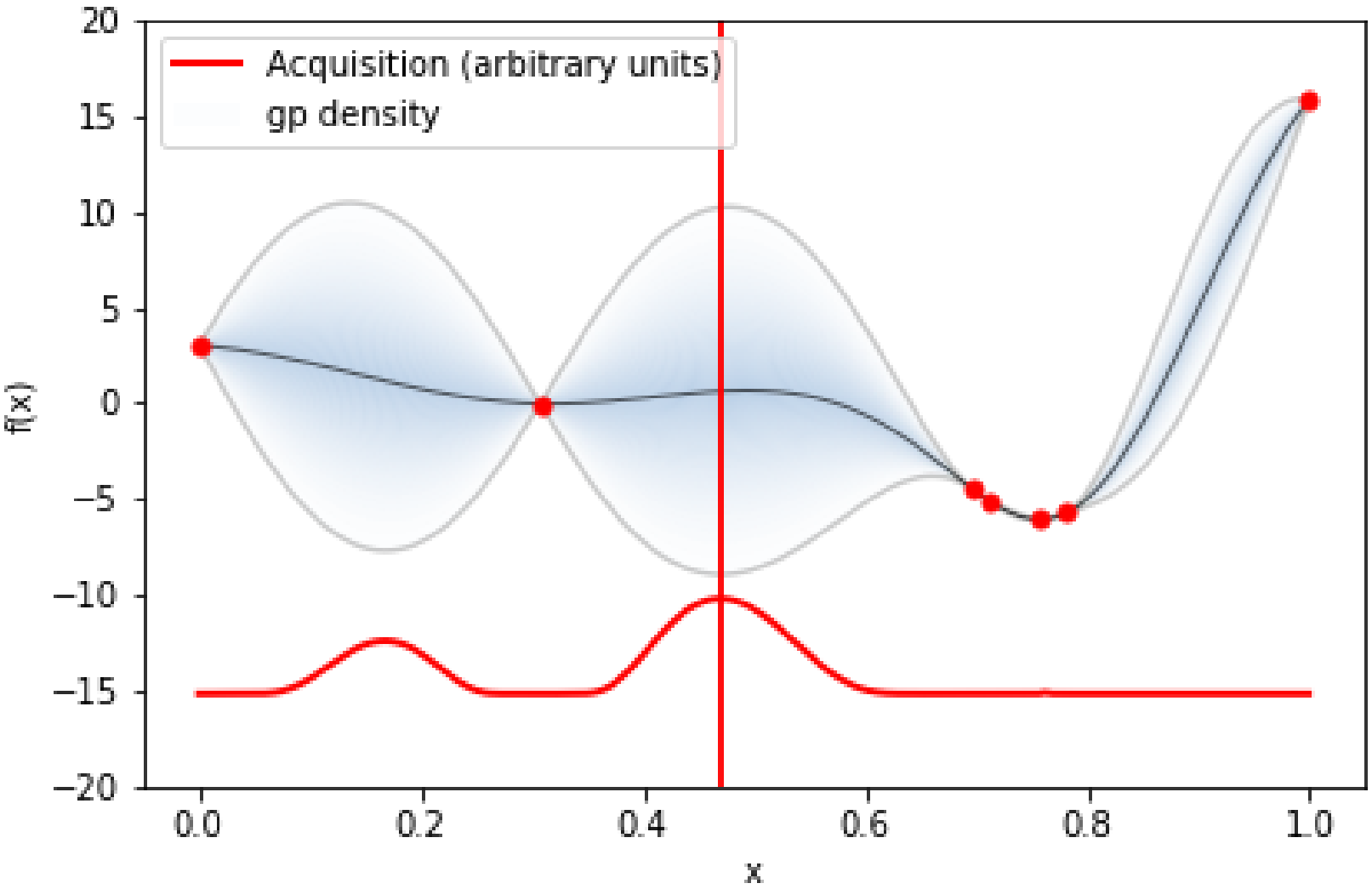
Байесовская оптимизация: пример

Итерация 6.



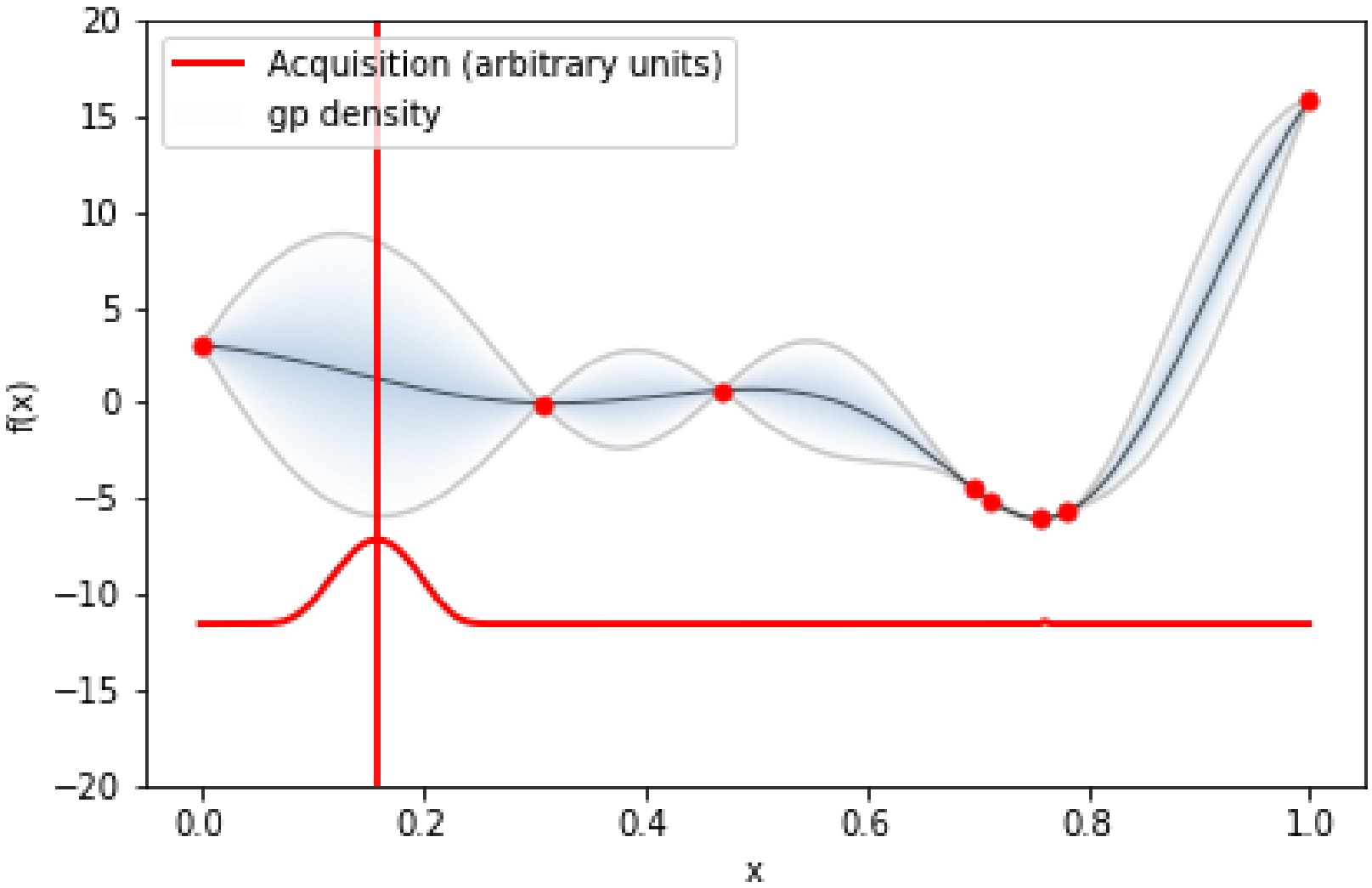
Байесовская оптимизация: пример

Итерация 7.



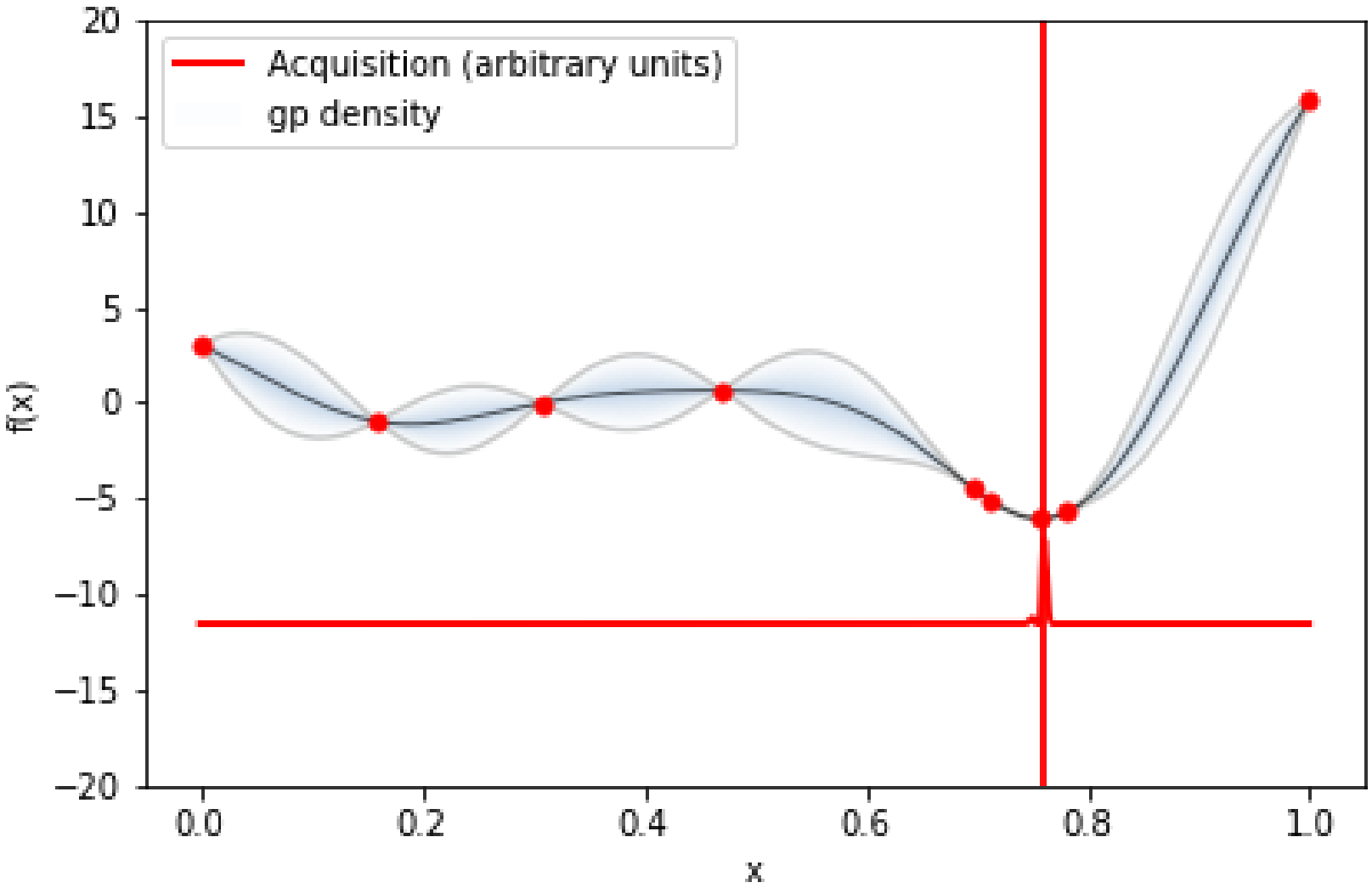
Байесовская оптимизация: пример

Итерация 8.



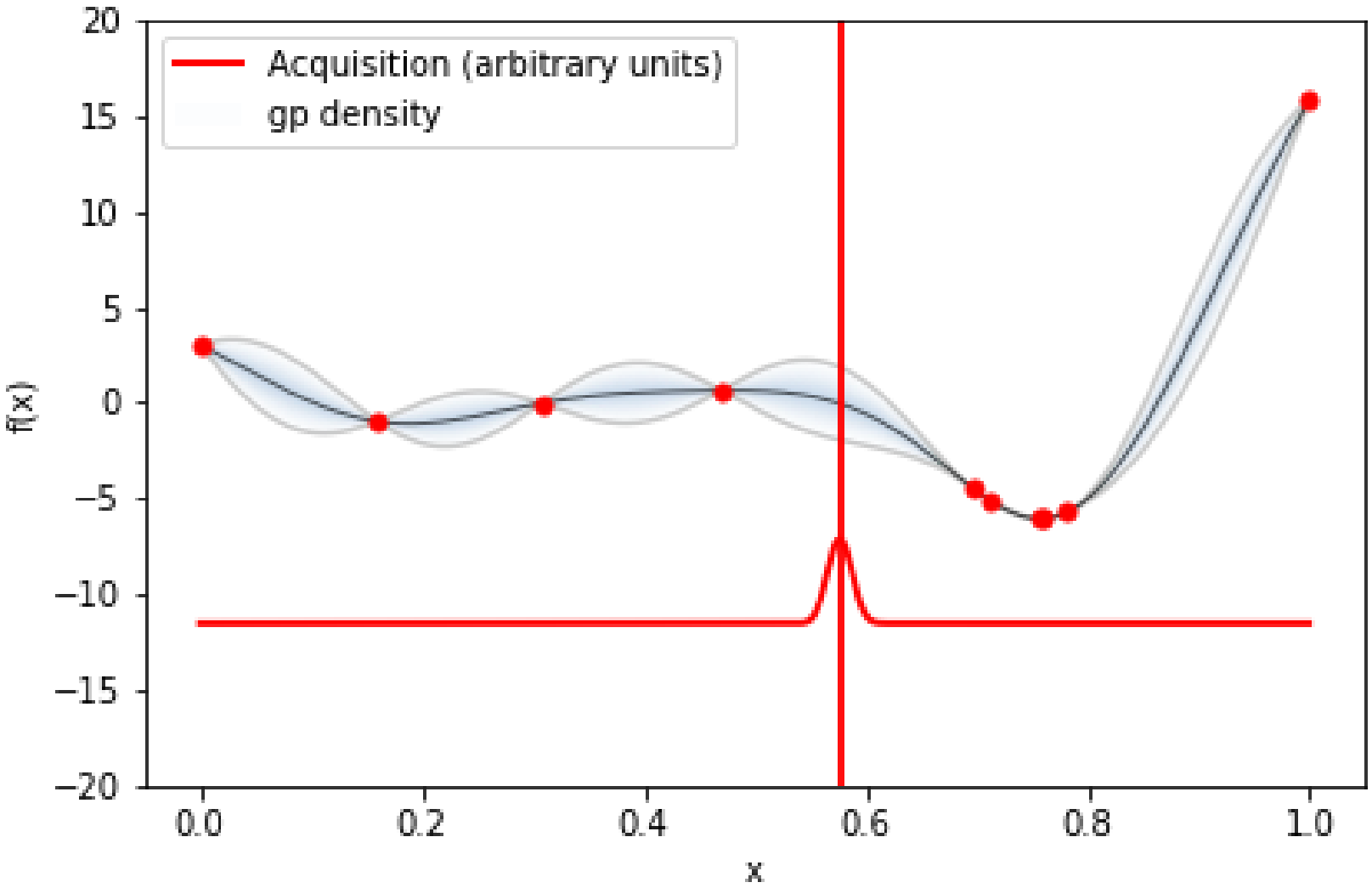
Байесовская оптимизация: пример

Итерация 9.



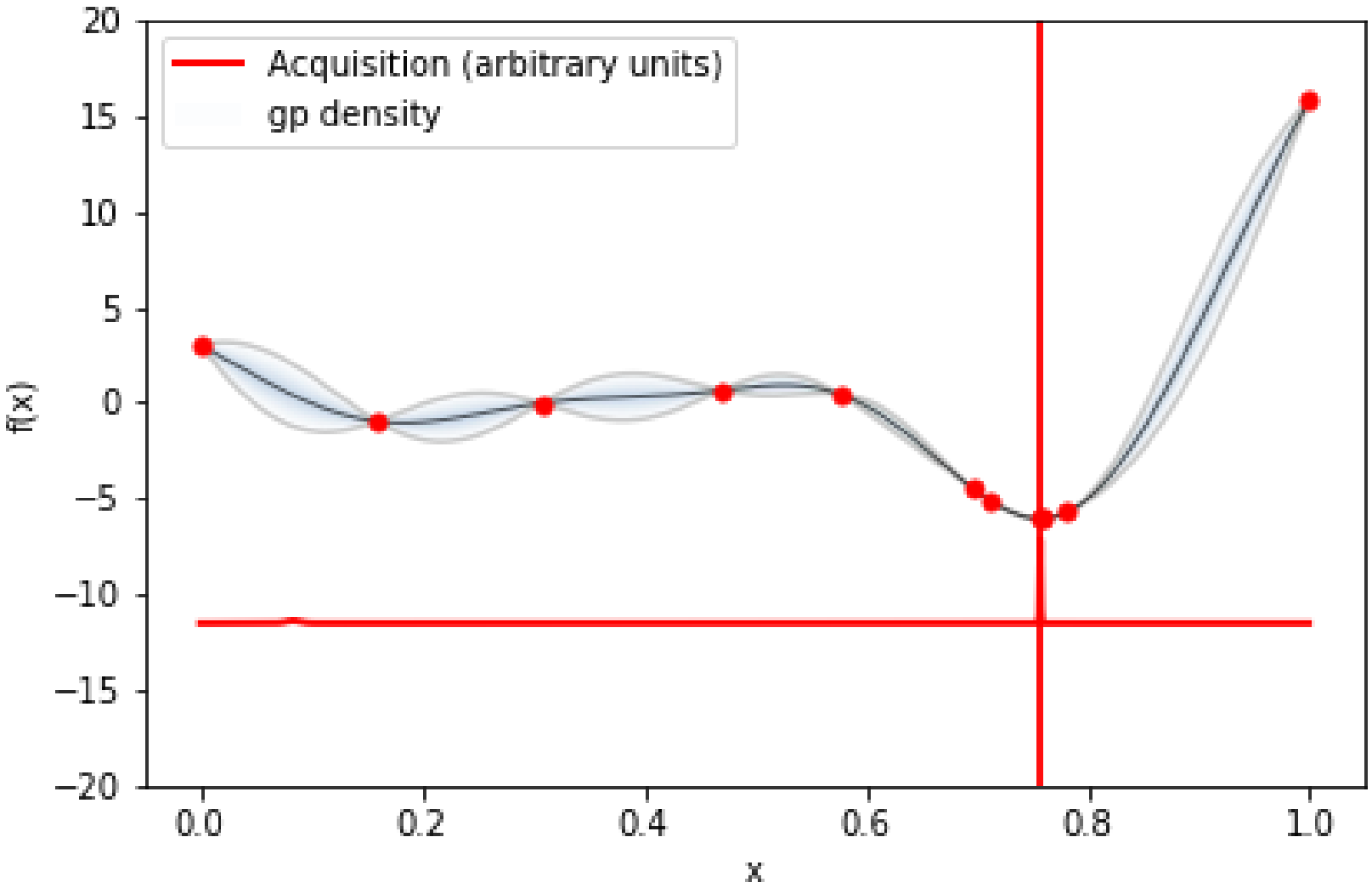
Байесовская оптимизация: пример

Итерация 10.



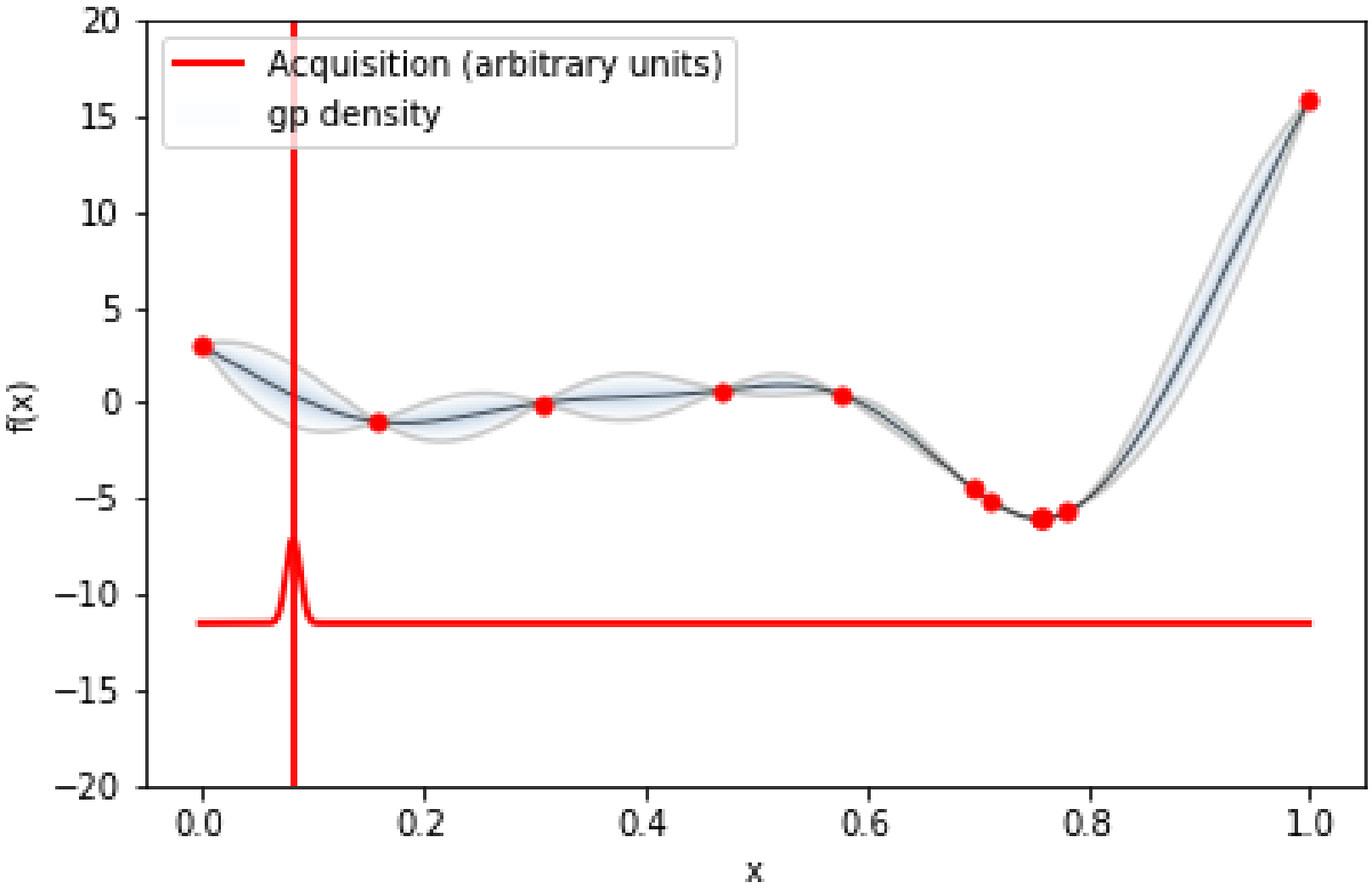
Байесовская оптимизация: пример

Итерация 11.



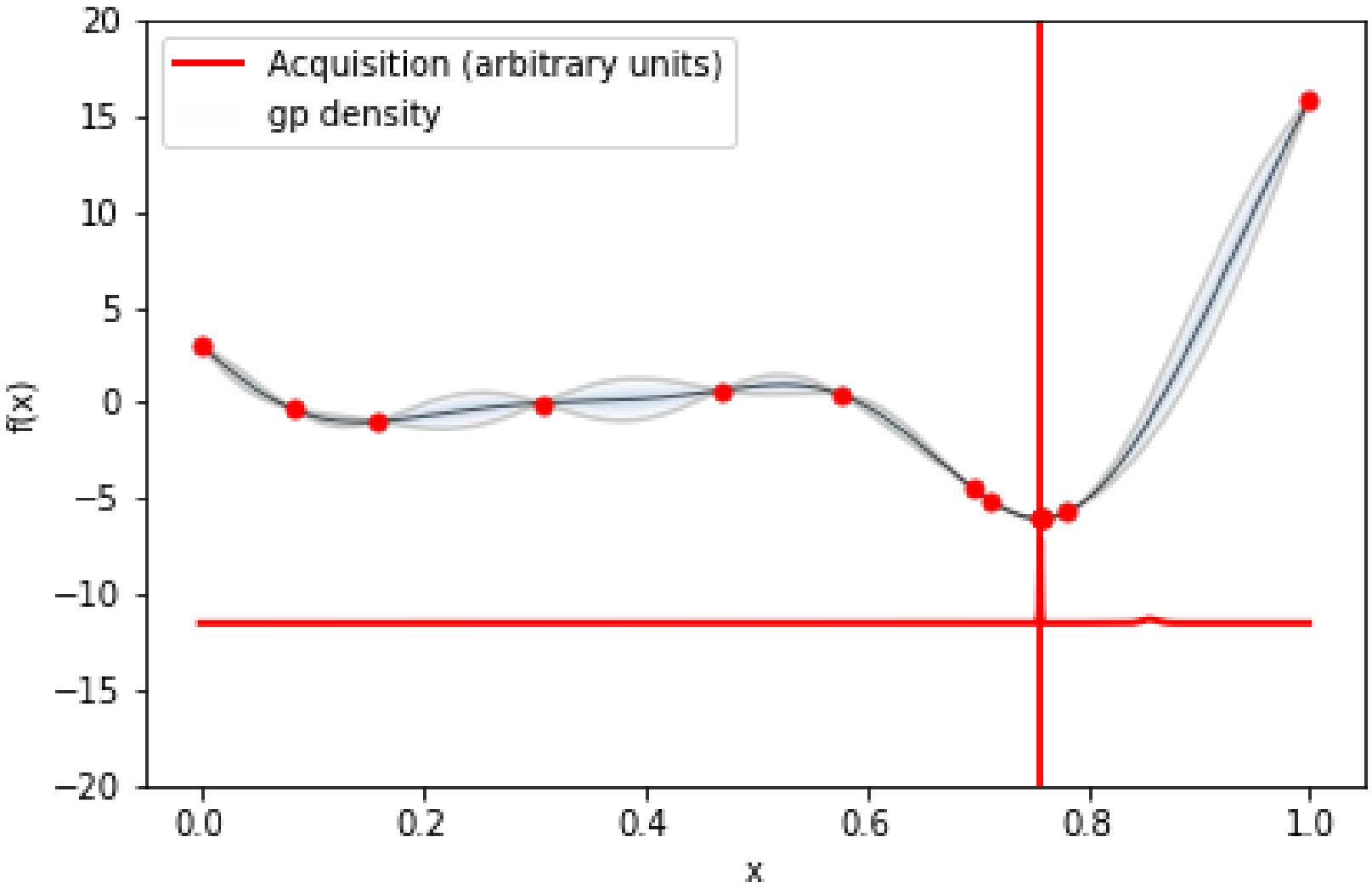
Байесовская оптимизация: пример

Итерация 12.



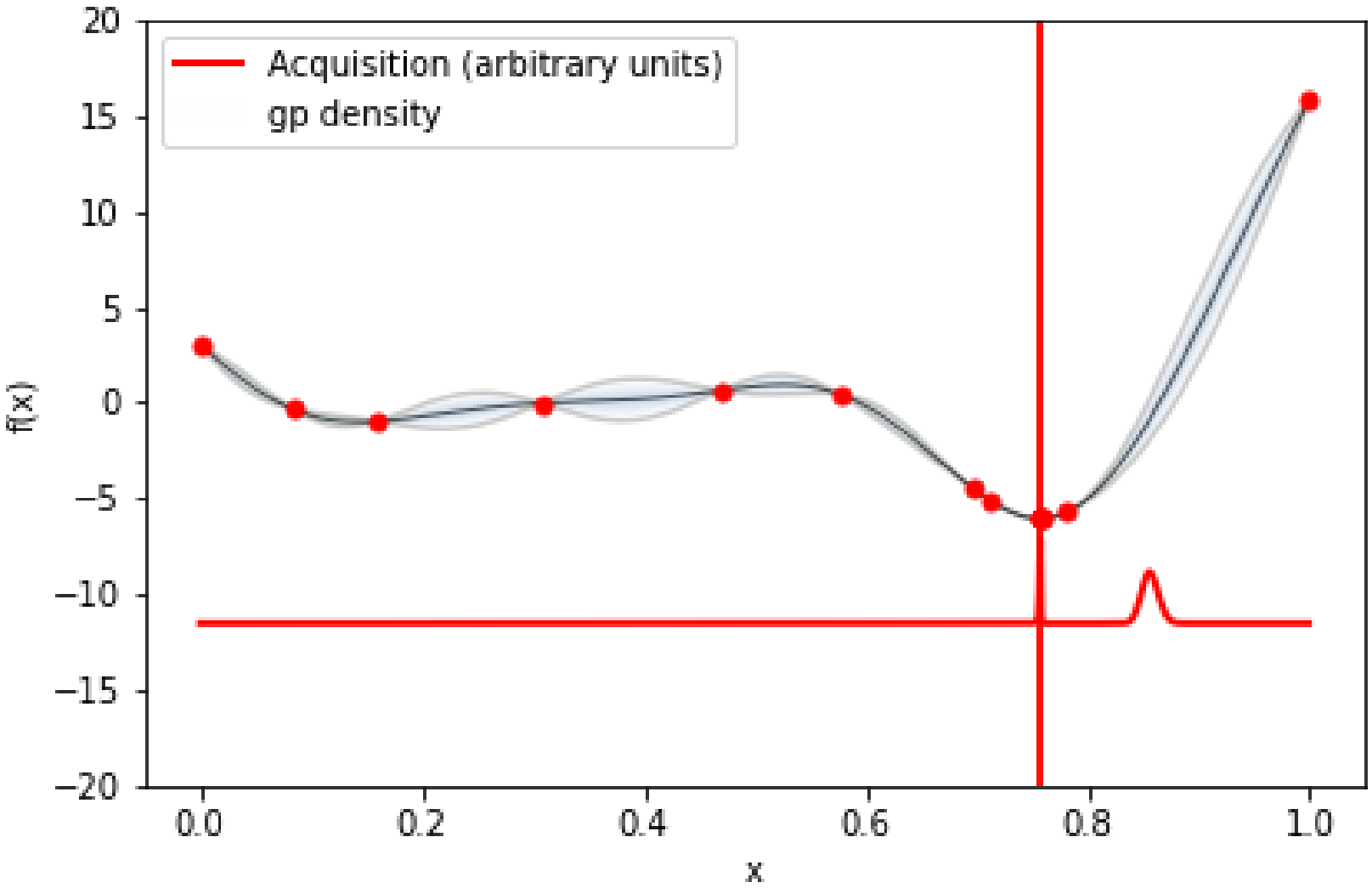
Байесовская оптимизация: пример

Итерация 13.



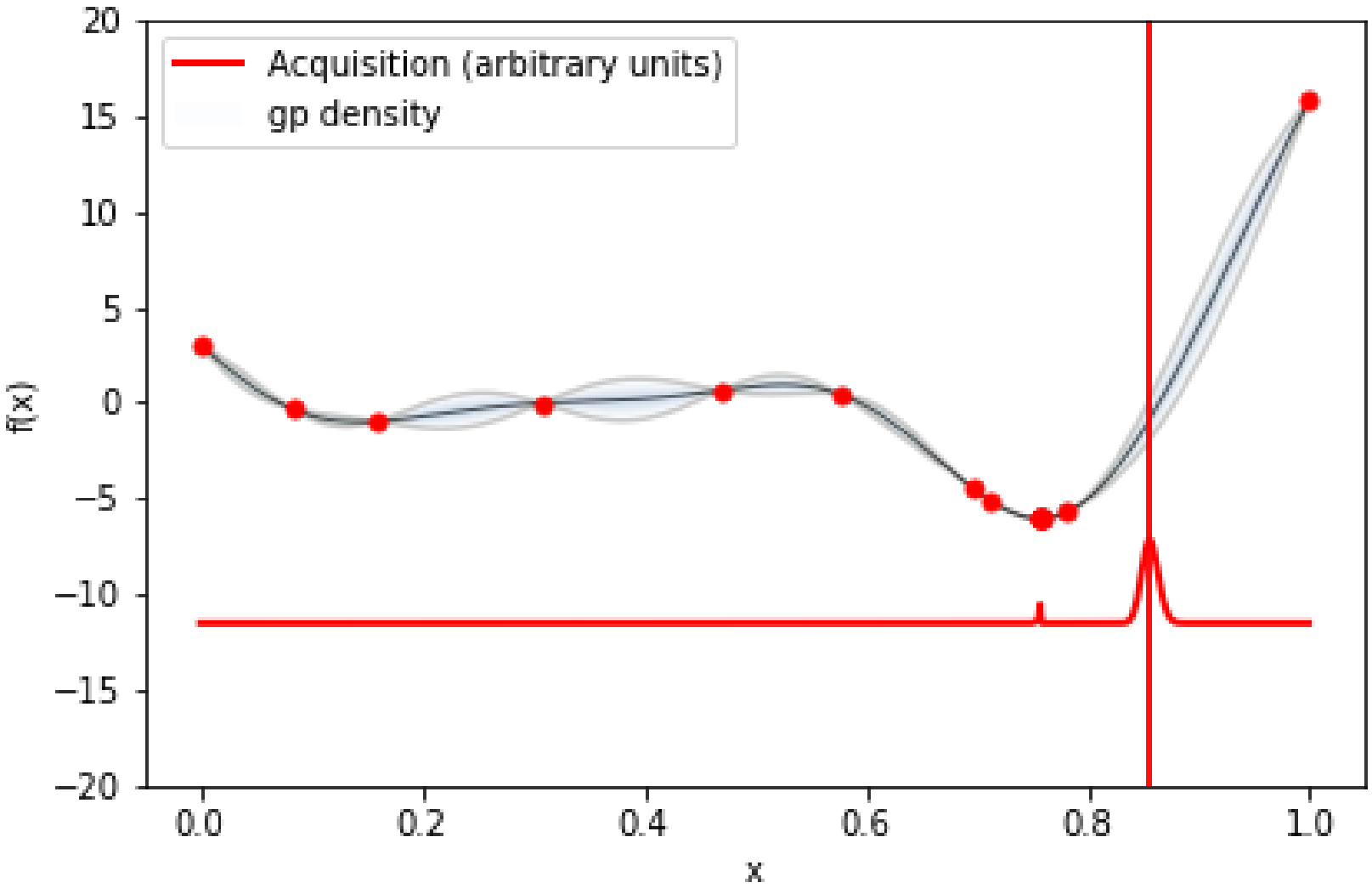
Байесовская оптимизация: пример

Итерация 14.



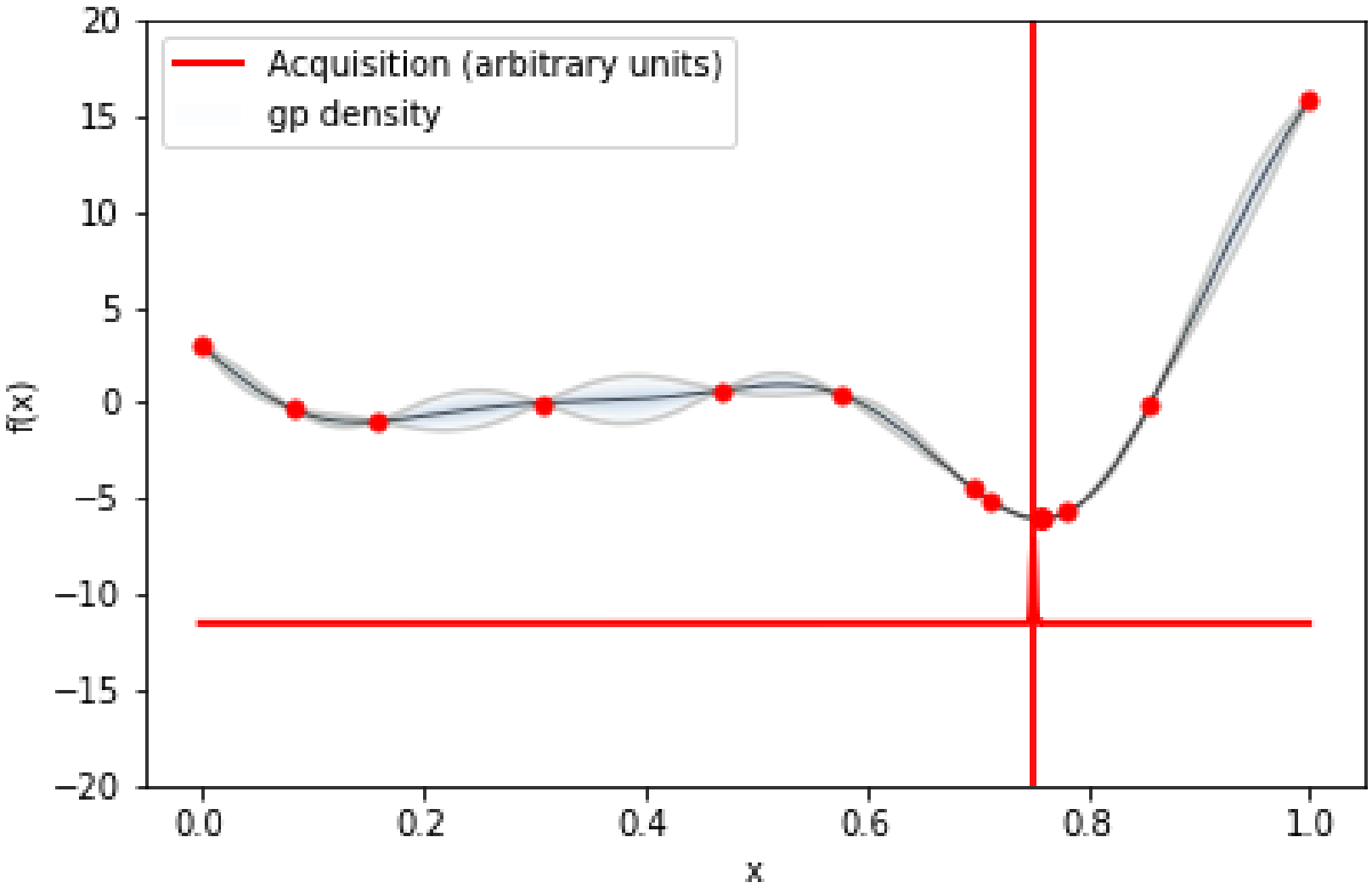
Байесовская оптимизация: пример

Итерация 15.



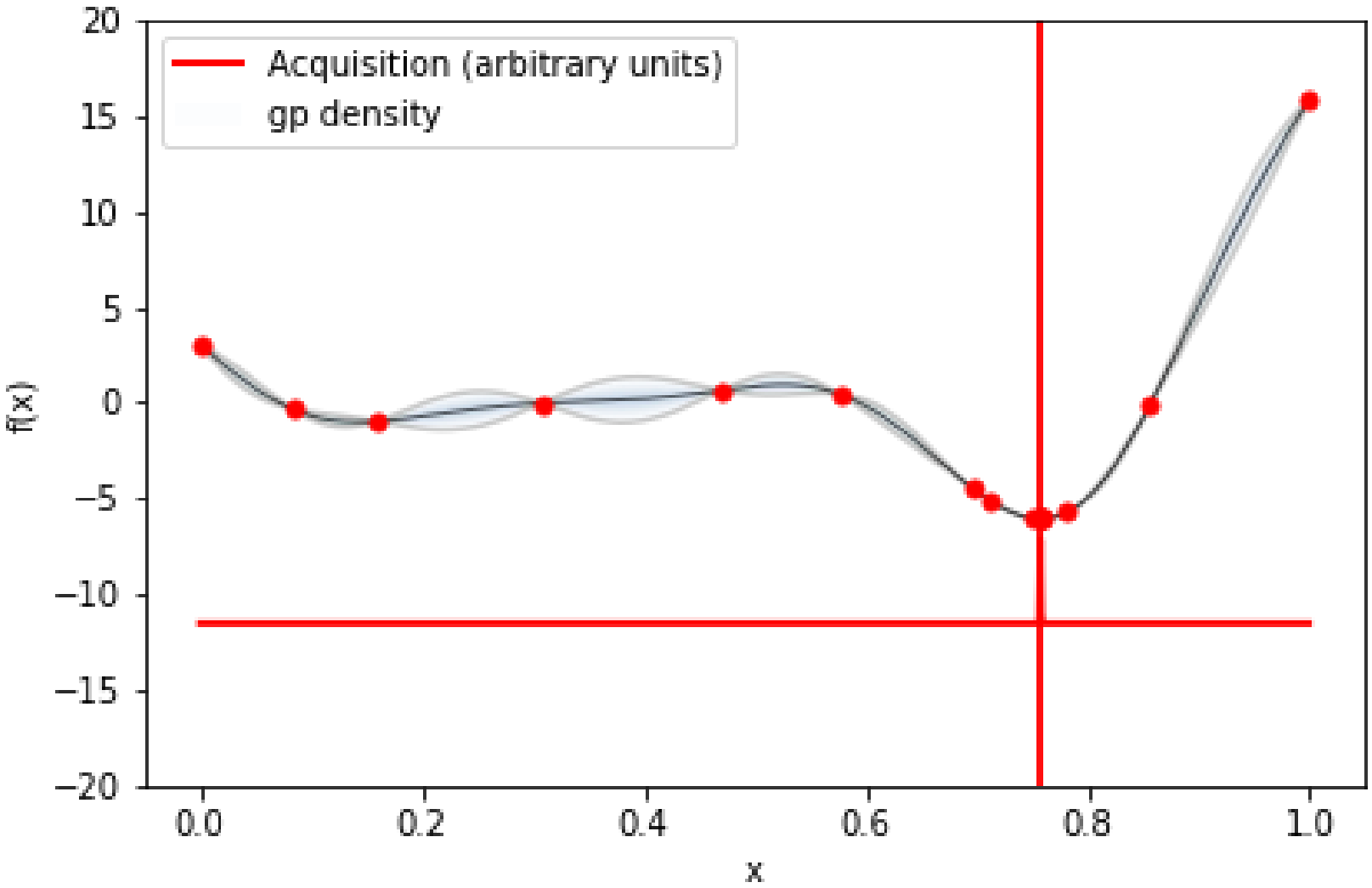
Байесовская оптимизация: пример

Итерация 16.



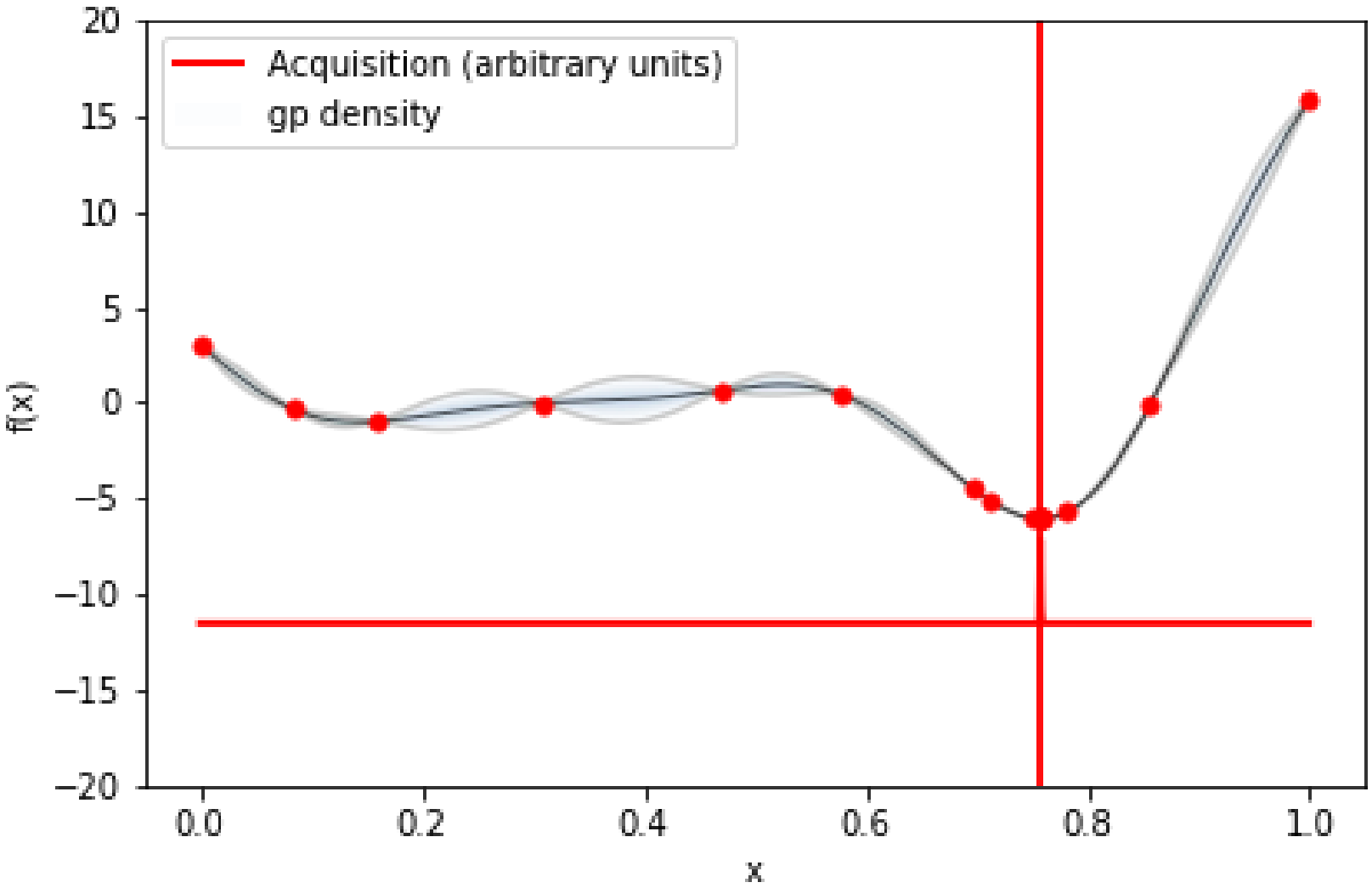
Байесовская оптимизация: пример

Итерация 17.



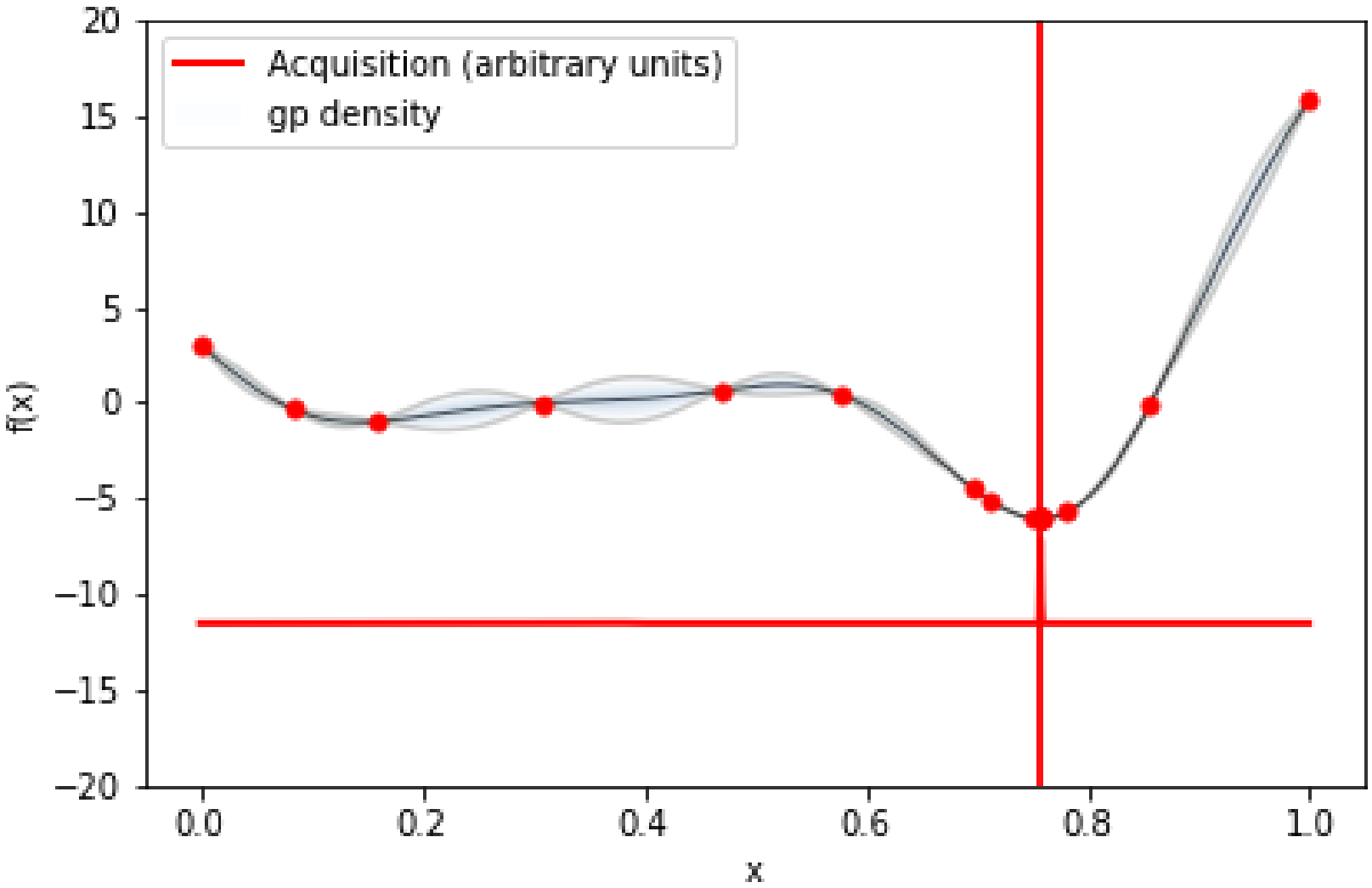
Байесовская оптимизация: пример

Итерация 18.



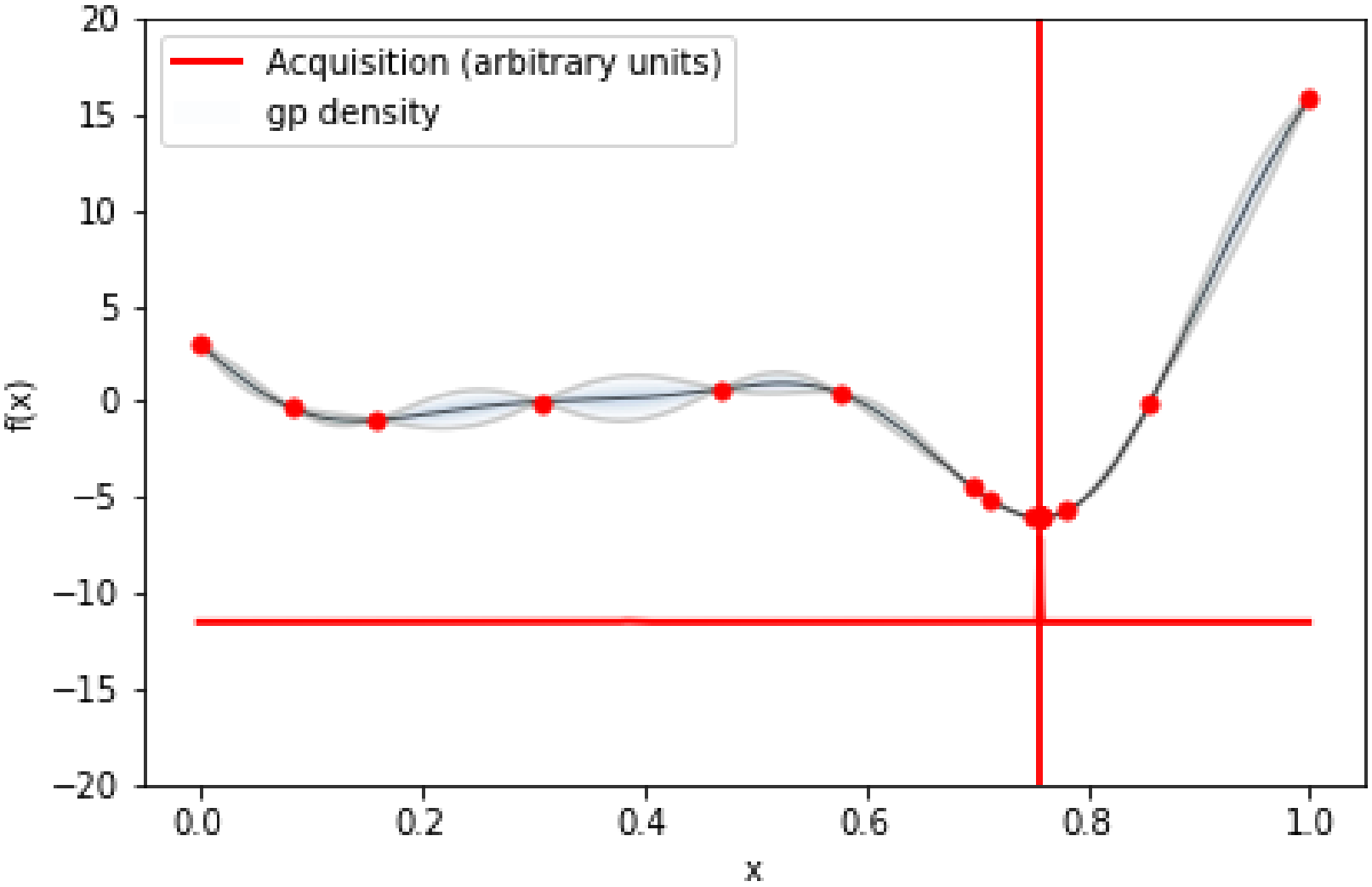
Байесовская оптимизация: пример

Итерация 19.



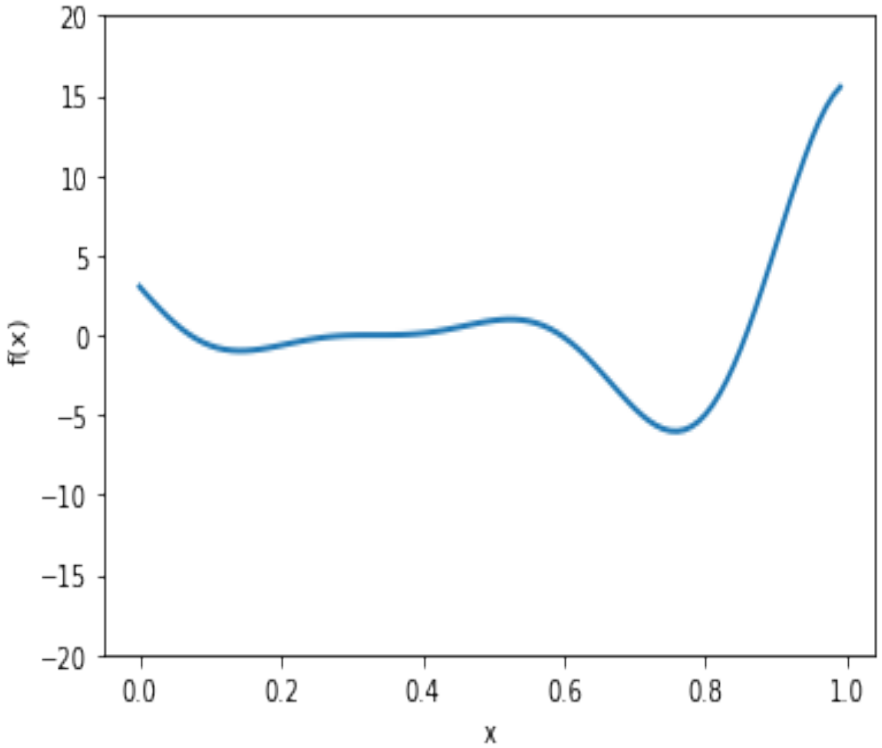
Байесовская оптимизация: пример

Итерация 20.

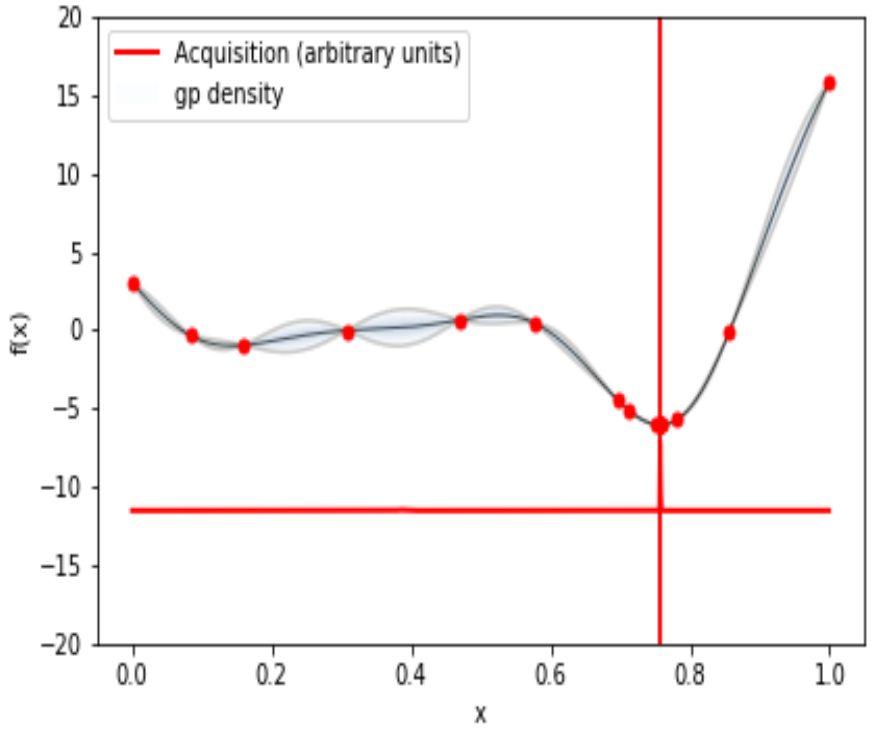


Байесовская оптимизация: пример

Сравним полученное на 20 итерации приближение с оригинальной функцией.



(a) Оригинальная функция



(b) Аппроксимирующий ГП

Содержание

Управление роботами: мотивация

Классическая постановка задачи управления: знаем физику, ищем оптимальное управление.

Современная постановка задачи управления: не знаем физику, пытаемся ее выучить и на ходу найти оптимальное управление.

Машины, для которых достаточен первый подход, дороги.

Для дешевых машин:

- мы не знаем их физику (она очень сильно отклоняется от “идеальной”),
- у нас нет возможности в ручном режиме эту физику изучать (строить подходящую модель), так как это сильно повышает цену.

Управление роботами: PILCO

PILCO (Probabilistic Inference for Learning COntrol) — подход, использующий гауссовские процессы для “изучения физики”.

PILCO: A Model-Based and Data-Efficient Approach to Policy Search

---

Marc Peter Deisenroth MARC@CS.WASHINGTON.EDU
Department of Computer Science & Engineering, University of Washington, USA

Carl Edward Rasmussen CER54@CAM.AC.UK
Department of Engineering, University of Cambridge, UK

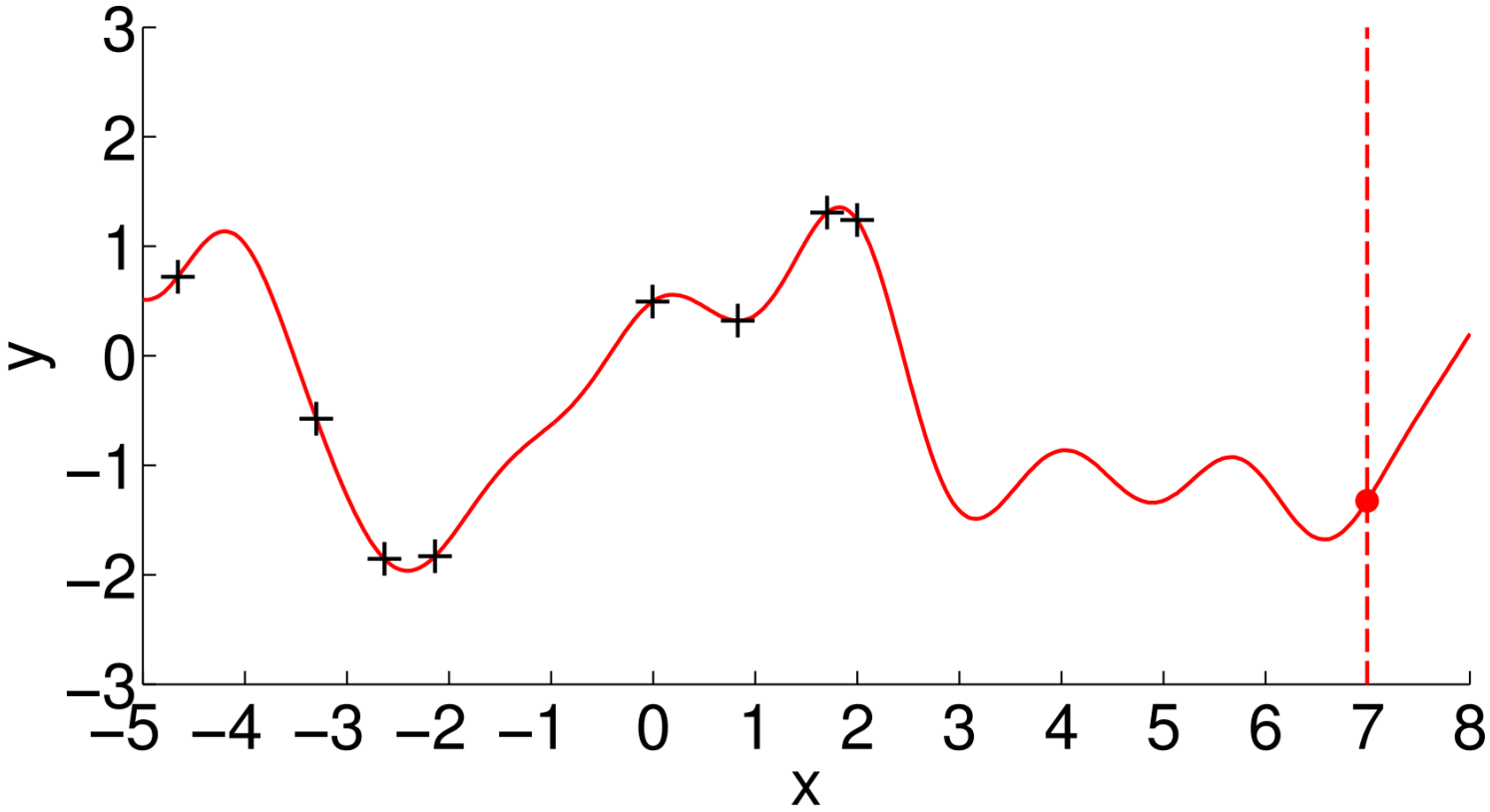
Модель системы можно записать в виде

$$x_{t+1} = f(x_t, u_t) + w, \qquad \text{где}$$

- x_t — траектория системы,
- u_t — управление,
- f моделирует физику,
- $w \sim N(0, \sigma^2)$ — случайный шум.

Управление роботами: PILCO

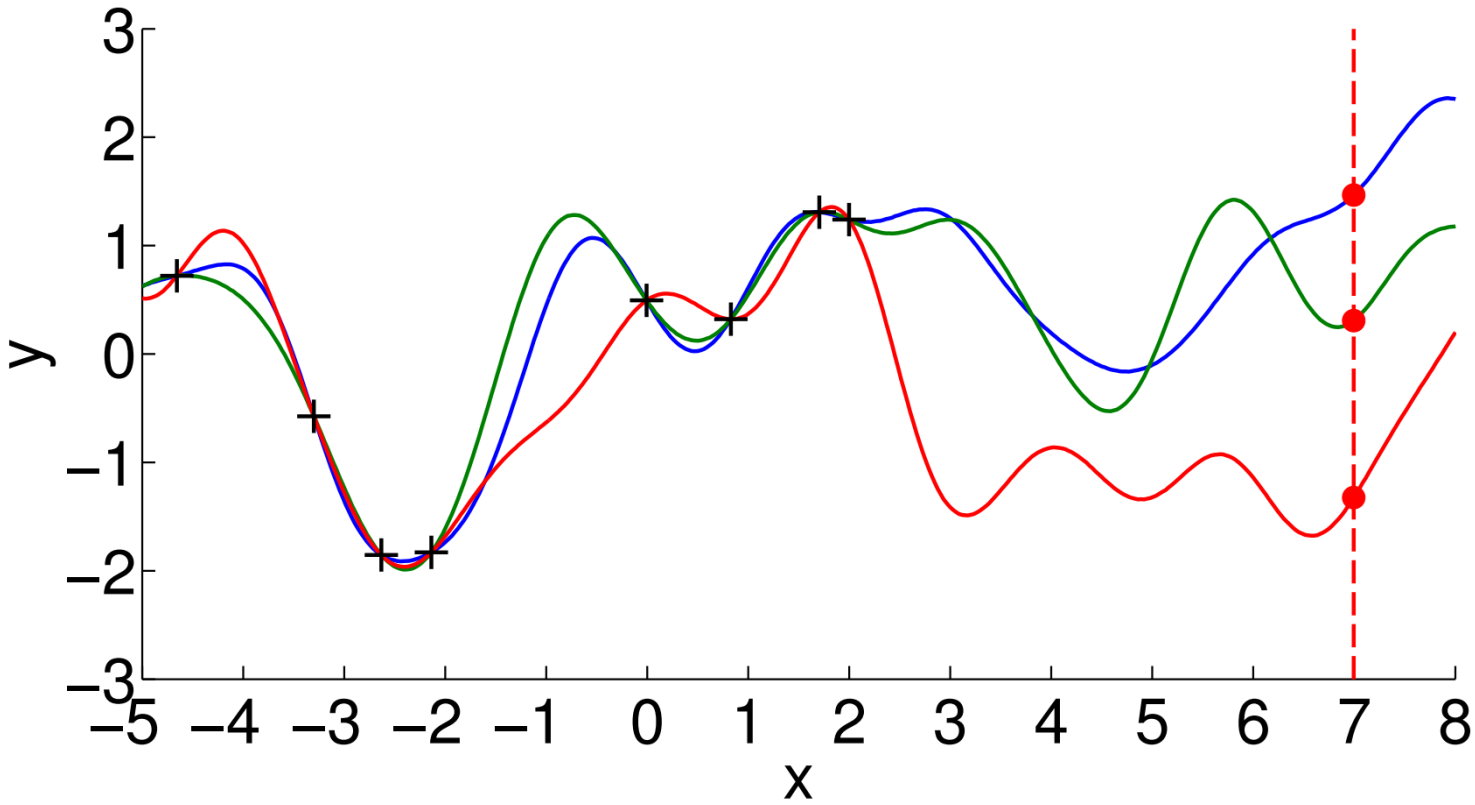
Представим, что f моделируется детерминистично.



Рассмотрим прогноз в точке $x = 7$.

Управление роботами: PILCO

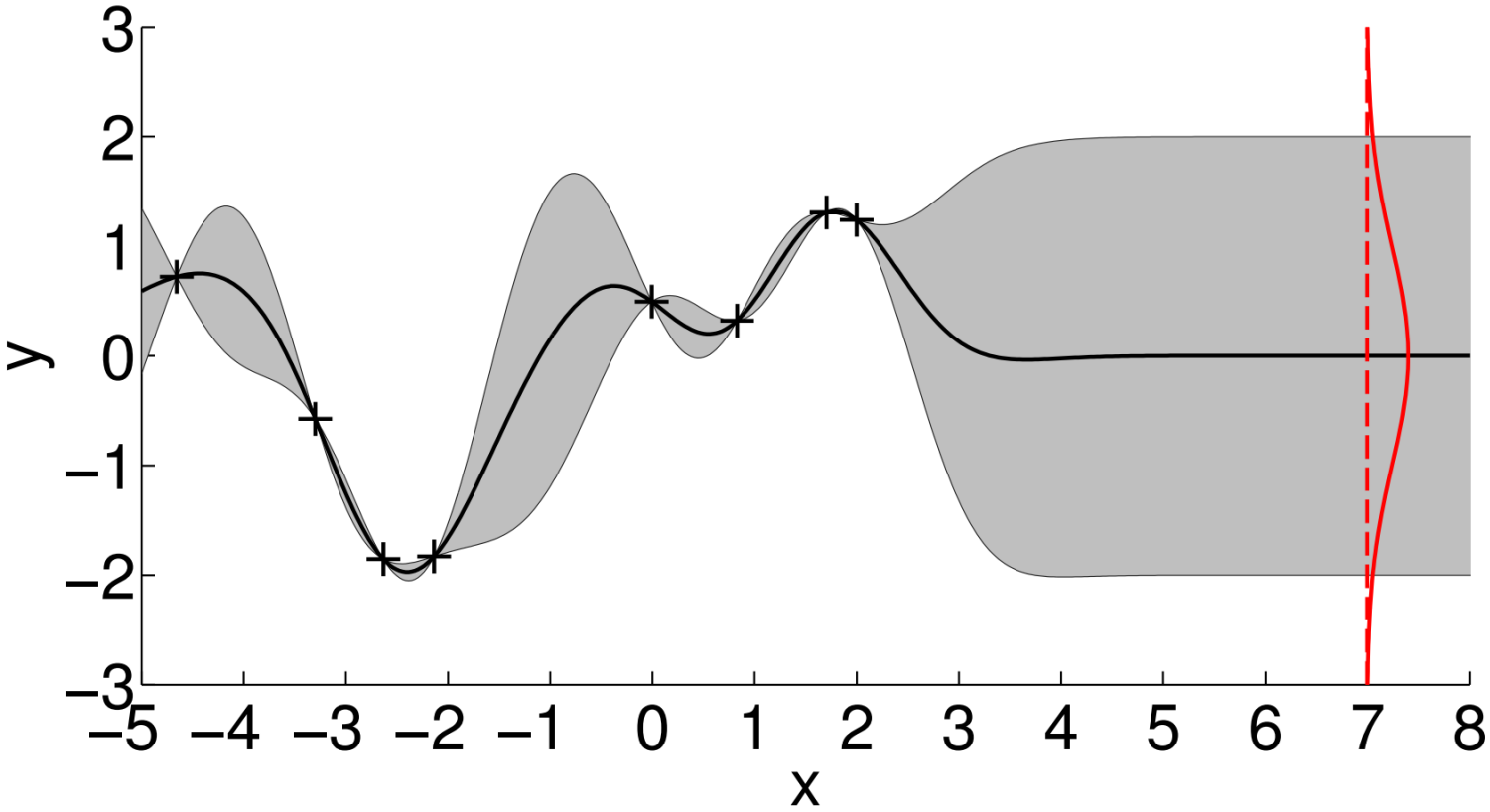
Представим, что f моделируется детерминистично.



Множество правдоподобных моделей, множество прогнозов.

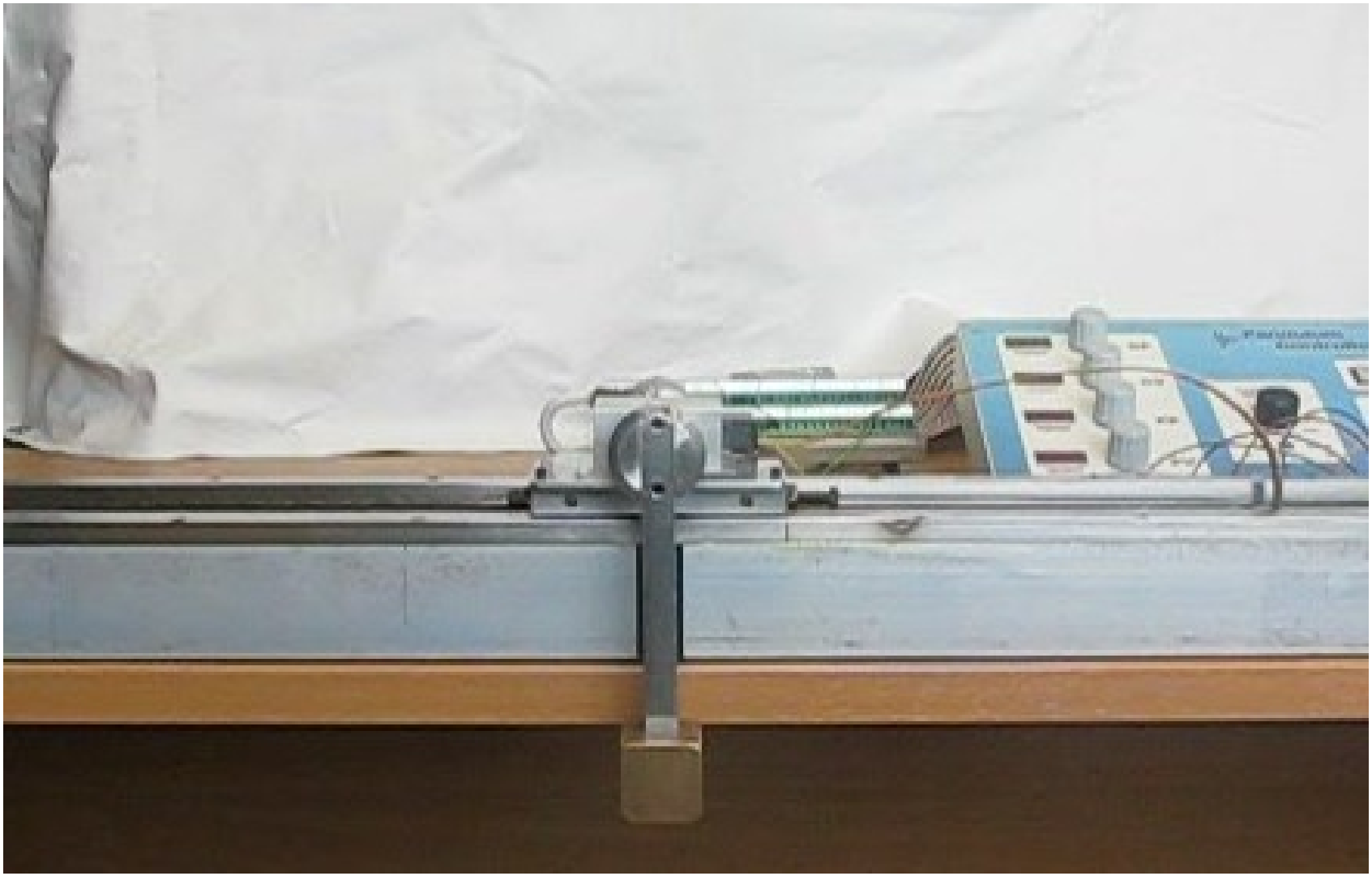
Управление роботами: PILCO

Что если моделировать f как гауссовский процесс?

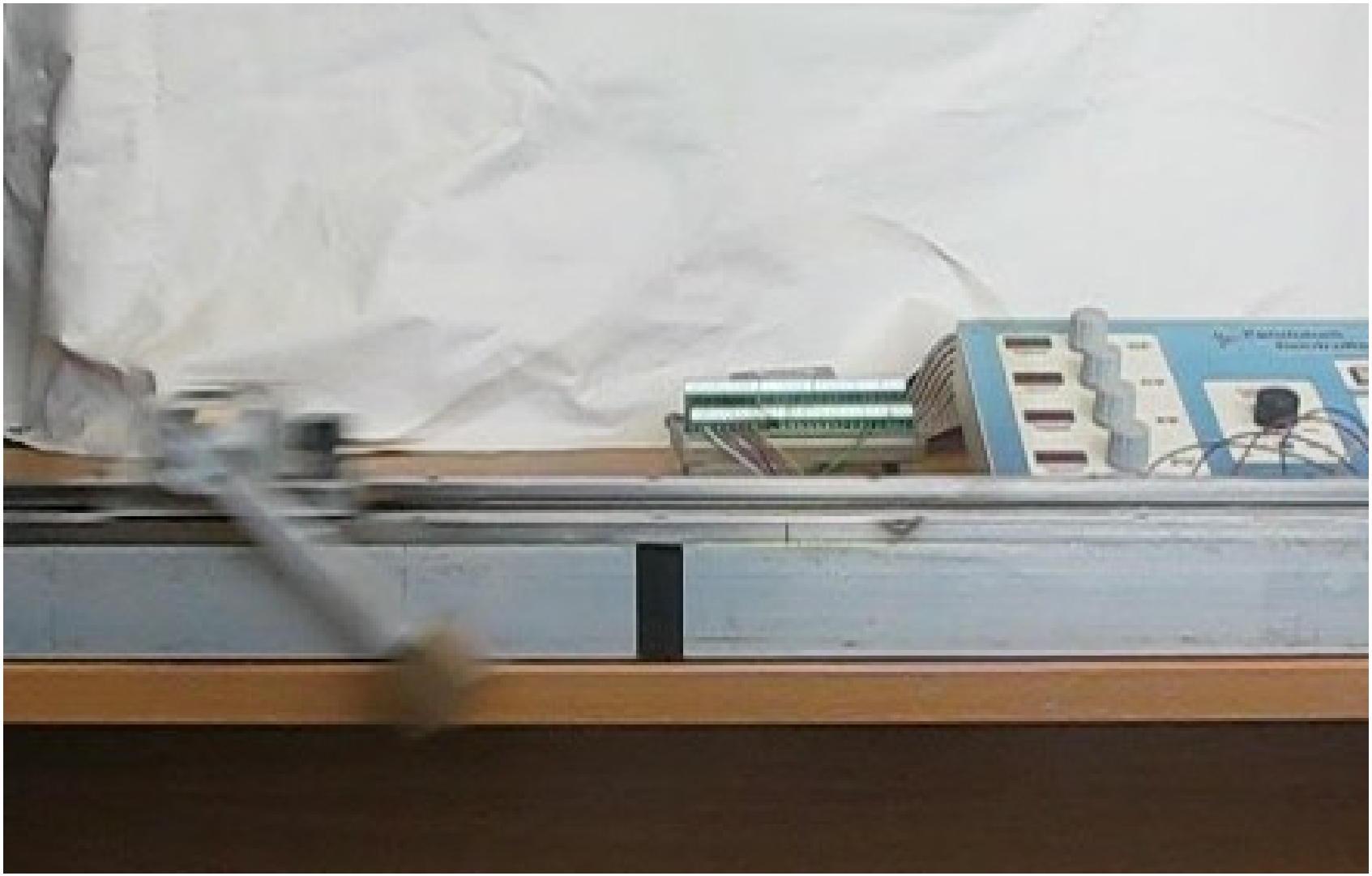


Если использовать ГП, появляется возможность учитывать неопределенность.

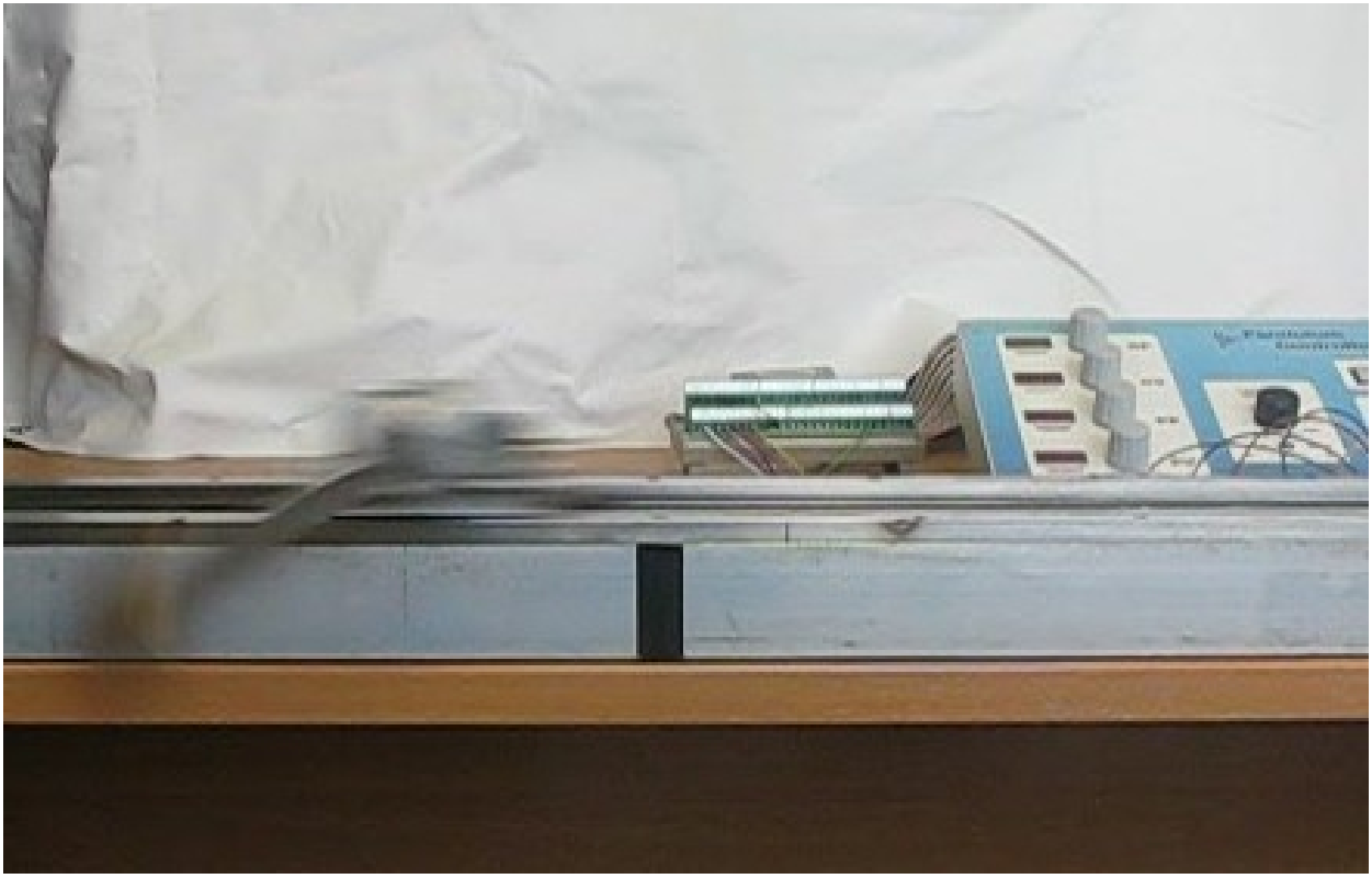
Управление роботами: пример (маятник)



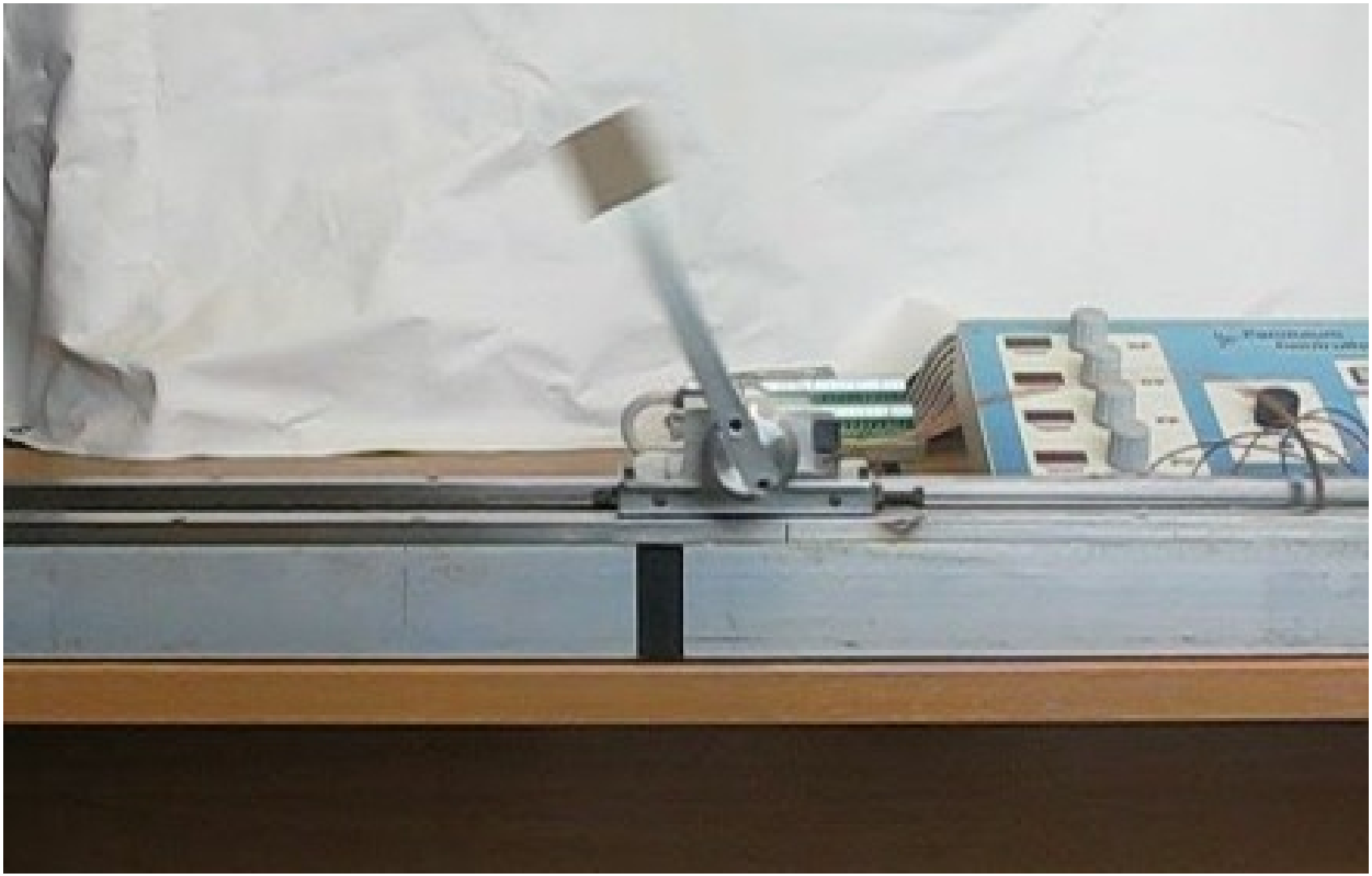
Управление роботами: пример (маятник)



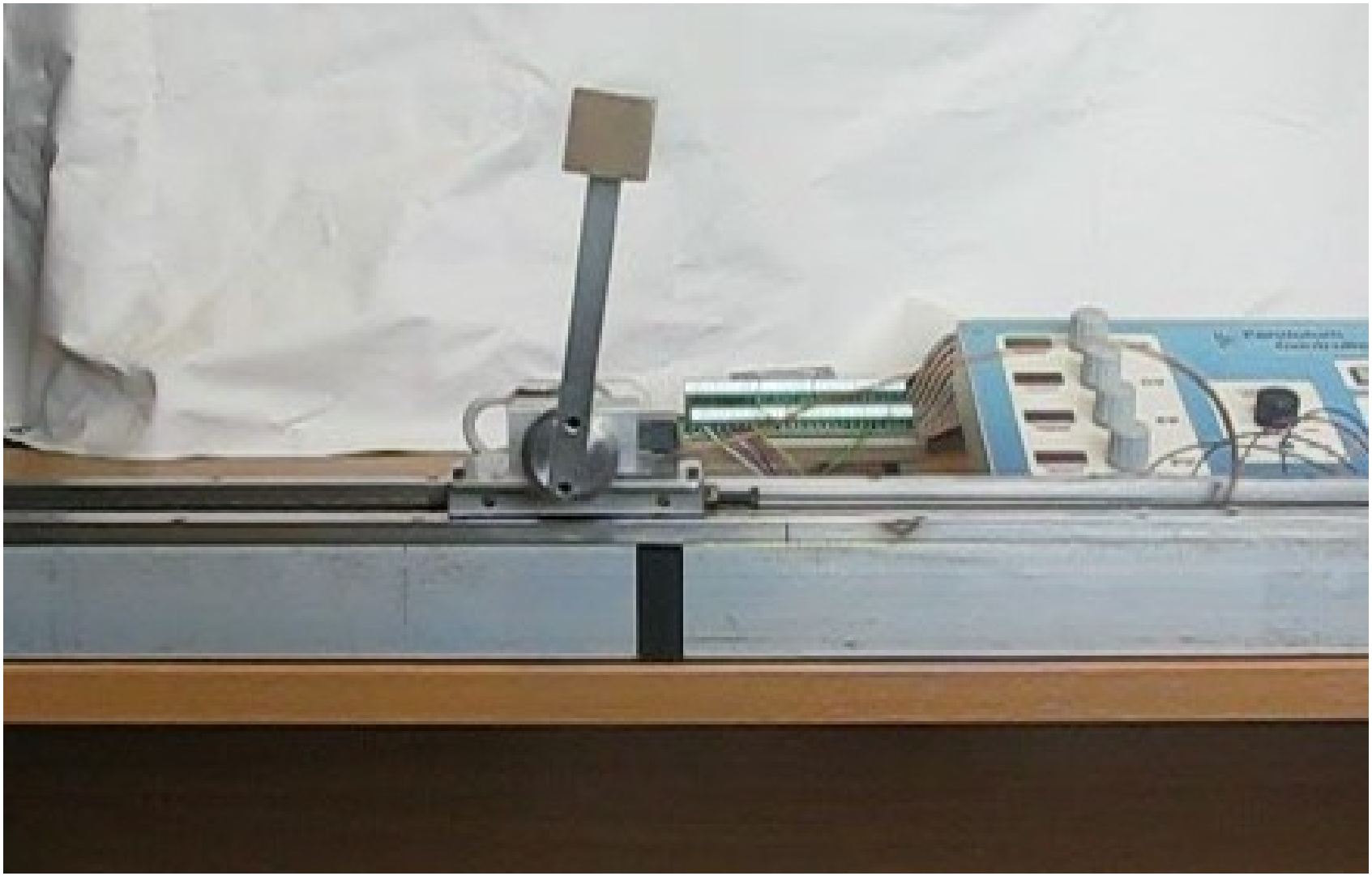
Управление роботами: пример (маятник)



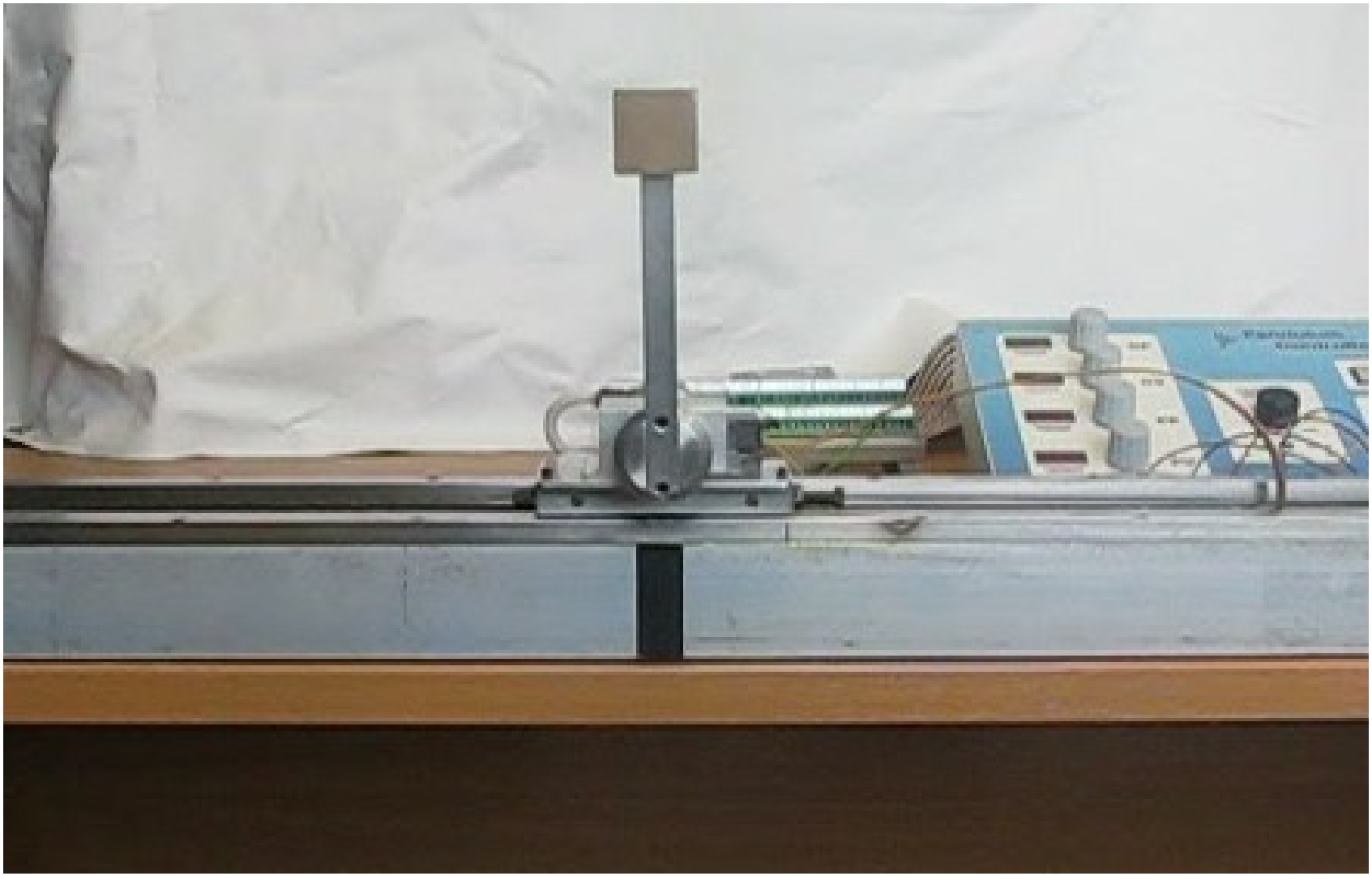
Управление роботами: пример (маятник)



Управление роботами: пример (маятник)



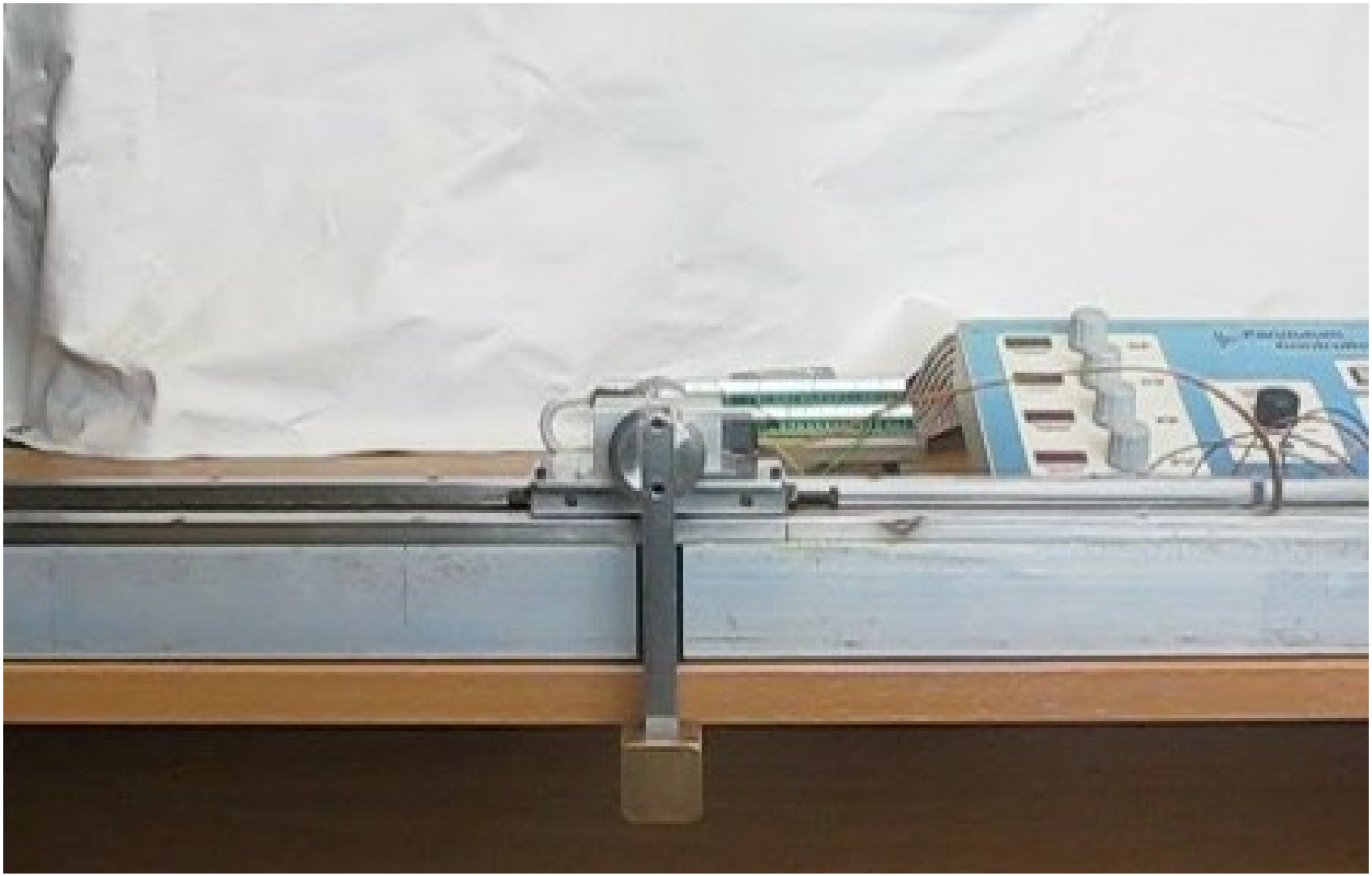
Управление роботами: пример (маятник)



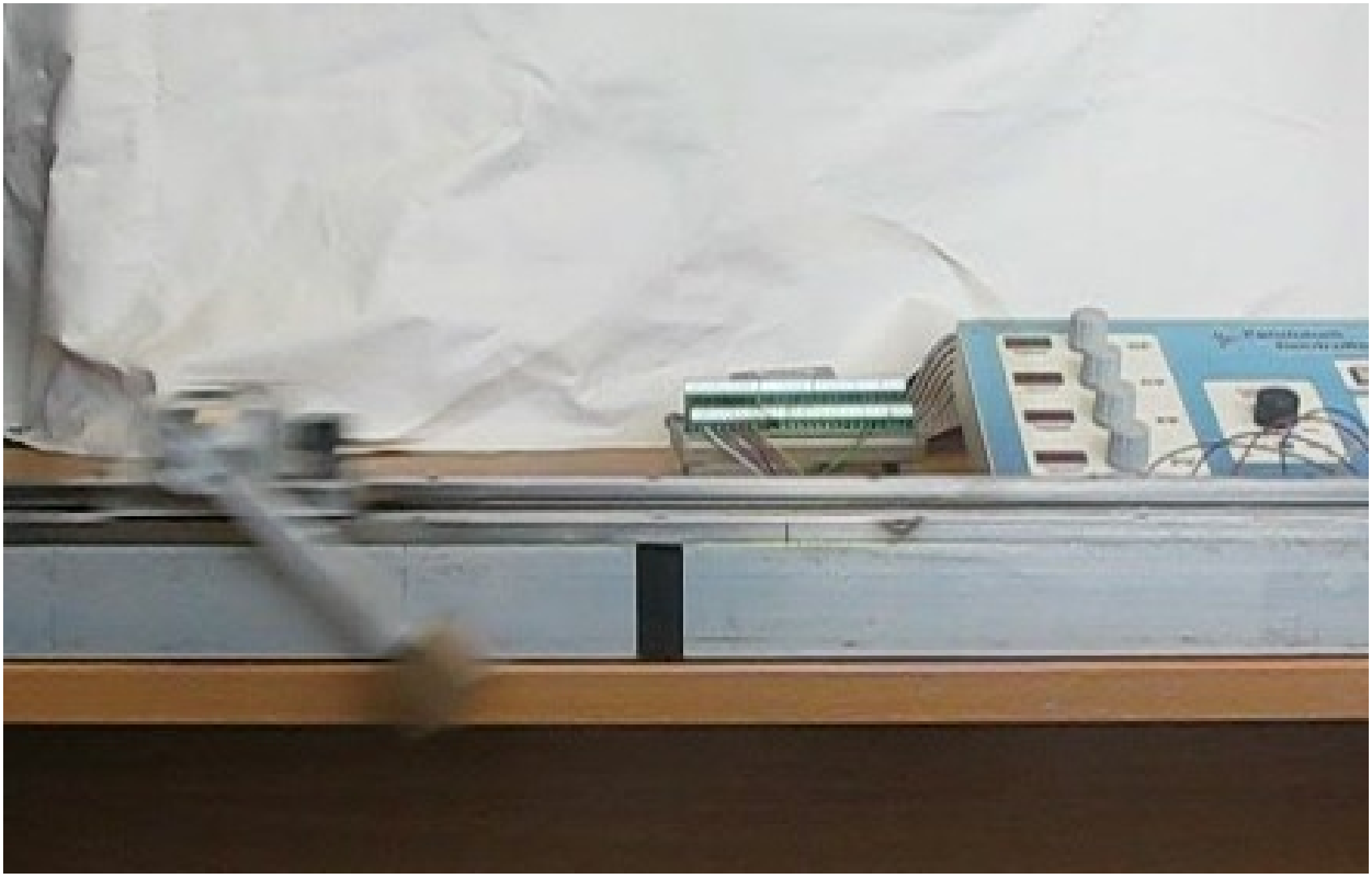
Управление роботами: пример (маятник)

Еще разок...

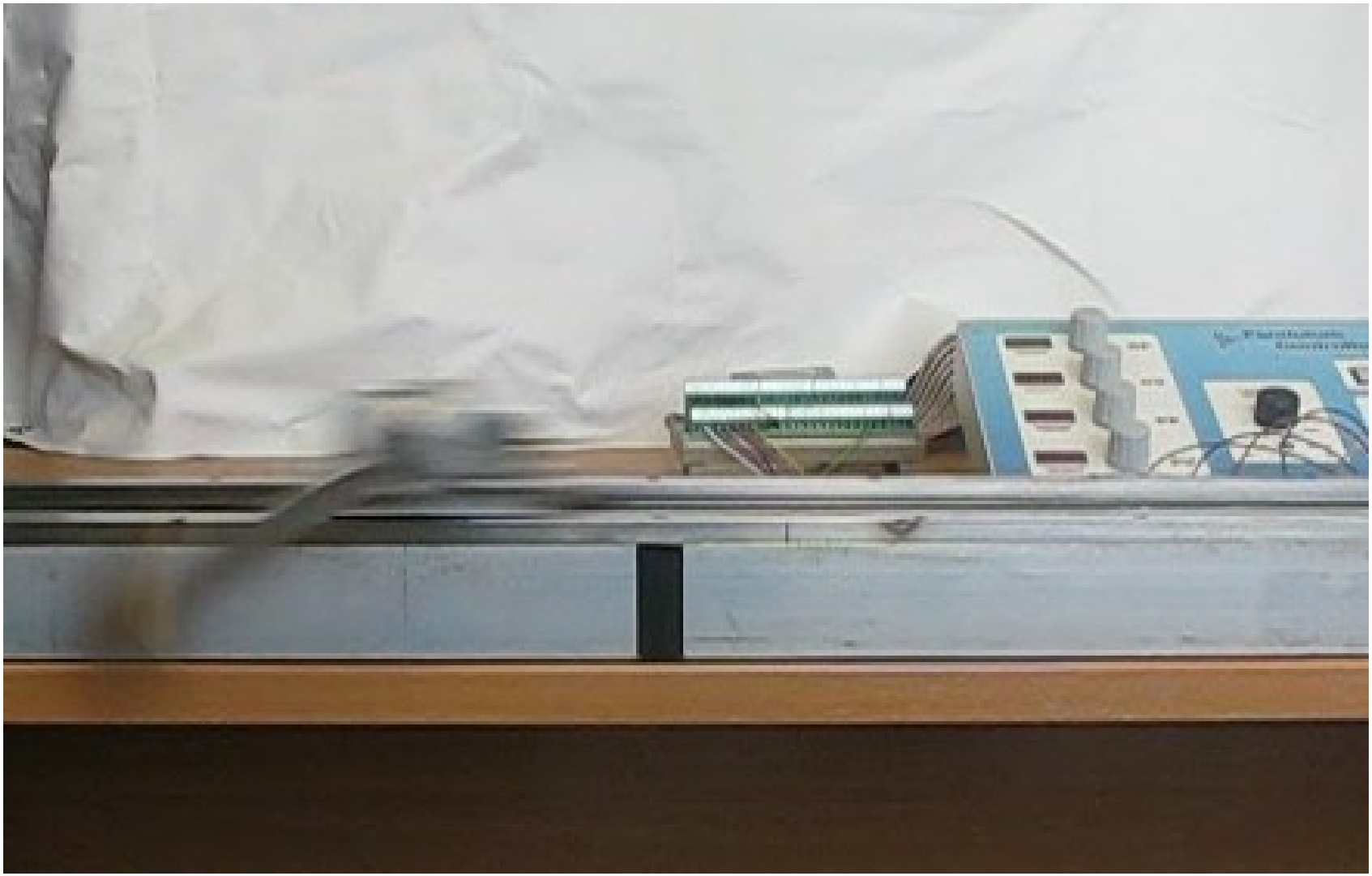
Управление роботами: пример (маятник)



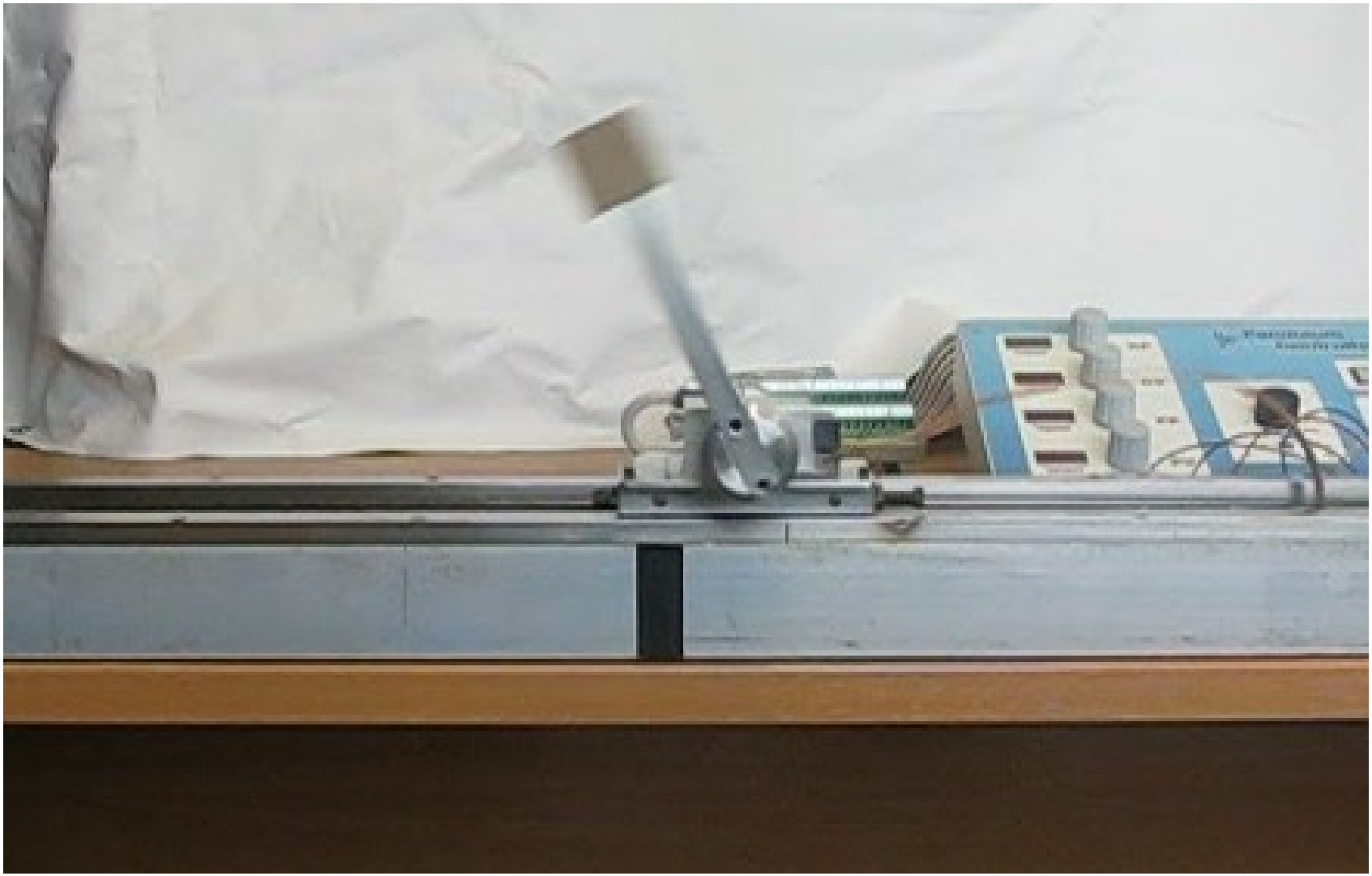
Управление роботами: пример (маятник)



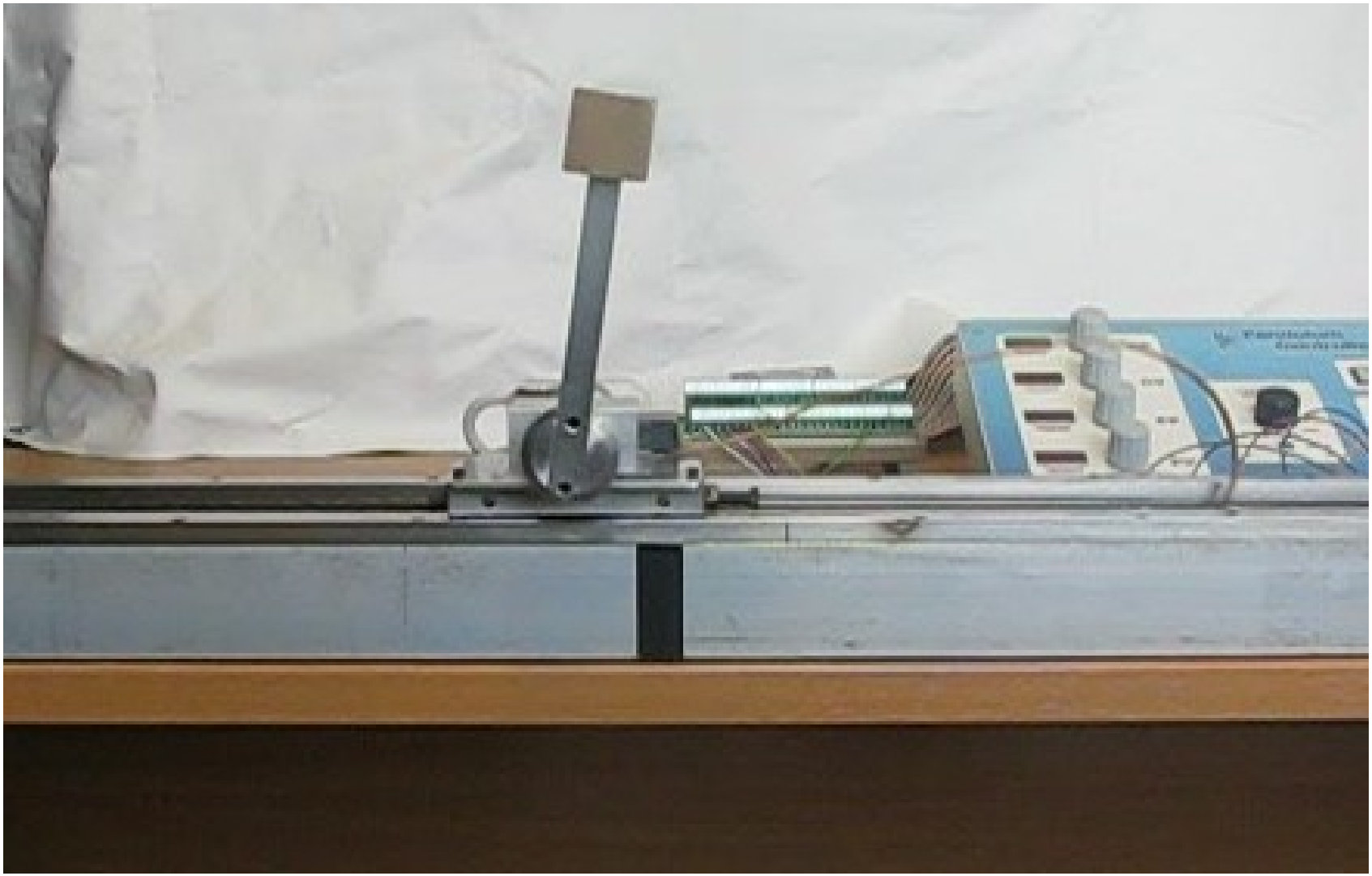
Управление роботами: пример (маятник)



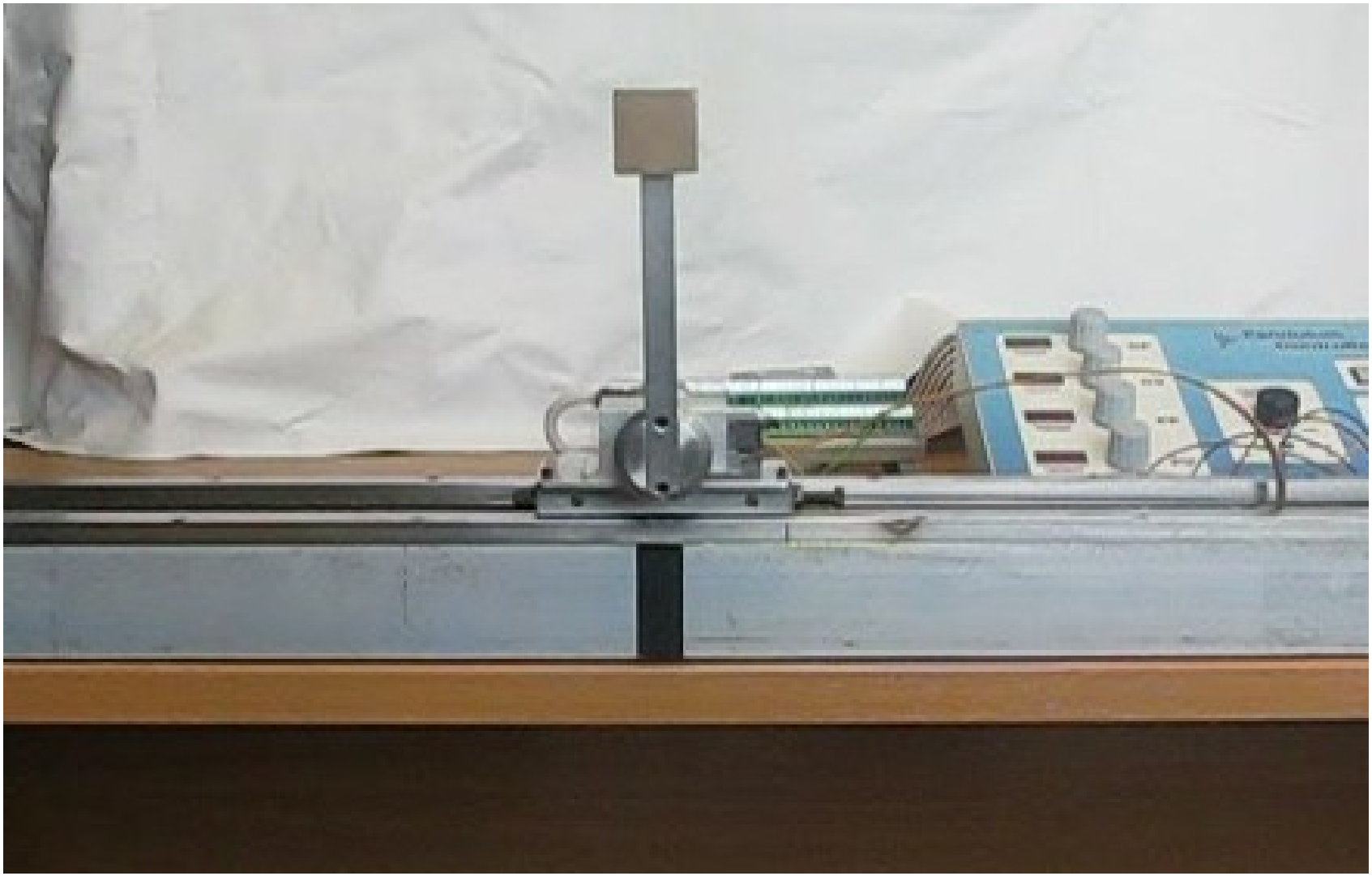
Управление роботами: пример (маятник)



Управление роботами: пример (маятник)



Управление роботами: пример (маятник)



Содержание

Место ГП в современном машинном обучении

State of the art модели для

- маленьких датасетов (small data),
- задач, где нужнна оценка неопределенности (uncertainty quantification).

Для практического применения

- библиотека (Python) https://sheffieldml.github.io/GPy/,
- туториалы nbviewer.jupyter.org/github/SheffieldML/notebook/blob/master/GPy/index.ipynb,
- книжка (в свободном доступе) C. E. Rasmussen & C. K. I. Williams, Gaussian Processes for Machine Learning,
- библитека на основе TensorFlow github.com/GPflow/GPflow.

Еще современные методы и задачи

- Что, если мы хотим делать не регрессию, а, например, классификацию?
  Ключевые слова: non-Gaussian likelihoods.
- Дополнительные приложения.
  Например, Gaussian Process Latent Variable Model — понижение размерности с помощью ГП.
- Гауссовские процессы и большие данные.
  Ключевые слова: Sparse Gaussian Processes.
- Более сложные модели.
  Например, Deep Gaussian Processes, Convolutional GPs.
- Теоретические вопросы.
  Например, сходимость байесовских нейронных сетей к ГП.

Спасибо за внимание!

viacheslav.borovitskiy@gmail.com

Материалы по теме:

- distill.pub/2019/visual-exploration-gaussian-processes
- C. E. Rasmussen & C. K. I. Williams, Gaussian Processes for Machine Learning
- cs.bath.ac.uk/~nc537/research/fonts
- mlg.eng.cam.ac.uk/pilco

В презентации были использованы иллюстрации из:
seeing-theory.brown.edu, inverseprobability.com/talks,
github.com/paintception/RandomWalk,
deisenroth.cc/talks/2020-02-13-aalto.pdf