Лекция 8. Гауссовские процессы. Байесовская оптимизация

Петр Мостовский

МКН СПбГУ

7 апреля 2022

Факультет математики и компьютерных наук СПбГУ

Черный ящик

- Черный ящик $f(x)$ — сложновычислимая функция. Градиенты f, как правило, недоступны

Черный ящик

- Черный ящик $f(x)$ — сложновычислимая функция. Градиенты f, как правило, недоступны
- Хотим найти

$$x_* = \text{argmax}_{x \in \mathcal{X}} f(x)$$

Черный ящик

- Черный ящик $f(x)$ — сложновычислимая функция. Градиенты f, как правило, недоступны
- Хотим найти
$$x_* = \operatorname{argmax}_{x \in \mathcal{X}} f(x)$$
- Некоторые стратегии

Черный ящик

- Черный ящик $f(x)$ — сложновычислимая функция. Градиенты f, как правило, недоступны
- Хотим найти
$$x_* = \operatorname{argmax}_{x \in \mathcal{X}} f(x)$$
- Некоторые стратегии
  - Grid search. Перестает работать с ростом размерности $\mathcal{X}$

Черный ящик

- Черный ящик $f(x)$ — сложновычислимая функция. Градиенты f, как правило, недоступны
- Хотим найти
$$x_* = \operatorname{argmax}_{x \in \mathcal{X}} f(x)$$
- Некоторые стратегии
  - Grid search. Перестает работать с ростом размерности $\mathcal{X}$
  - Random search. Никак не использует доступную информацию об f

Черный ящик

- Черный ящик $f(x)$ — сложновычислимая функция. Градиенты f, как правило, недоступны
- Хотим найти
$$x_* = \operatorname{argmax}_{x \in \mathcal{X}} f(x)$$
- Некоторые стратегии
  - Grid search. Перестает работать с ростом размерности $\mathcal{X}$
  - Random search. Никак не использует доступную информацию об f
  - Численные градиенты. Вычисление градиента так же дорого, как вычисление самой f. Использует только последние несколько значений f

Примеры черных ящиков

- Качество обученной нейросети на валидационном датасете

Примеры черных ящиков

- Качество обученной нейросети на валидационном датасете (параметры: learning rate, momentum, batch size и т.д.)

Примеры черных ящиков

- Качество обученной нейросети на валидационном датасете (параметры: learning rate, momentum, batch size и т.д.)
- Дизайн экспериментов

Примеры черных ящиков

- Качество обученной нейросети на валидационном датасете (параметры: learning rate, momentum, batch size и т.д.)
- Дизайн экспериментов
- Дизайн машин (например, крыла самолета)

Примеры черных ящиков

- Качество обученной нейросети на валидационном датасете (параметры: learning rate, momentum, batch size и т.д.)
- Дизайн экспериментов
- Дизайн машин (например, крыла самолета)
- et cetera...

Предположения о черном ящике

- f сложновычислима

Предположения о черном ящике

- f сложновычислима
- f непрерывная, но у нее нет особой структуры (например, выпуклости)

Предположения о черном ящике

- f сложновычислима
- f непрерывная, но у нее нет особой структуры (например, выпуклости)
- $\mathcal{X}$ компактно
- градиенты f недоступны

Предположения о черном ящике

- f сложновычислима
- f непрерывная, но у нее нет особой структуры (например, выпуклости)
- $\mathcal{X}$ компактно
- градиенты f недоступны
- значения f могут быть шумными

Байесовская оптимизация

- Поскольку f — черный ящик, будем воспринимать f как случайную (введем prior $p(f)$)

Байесовская оптимизация

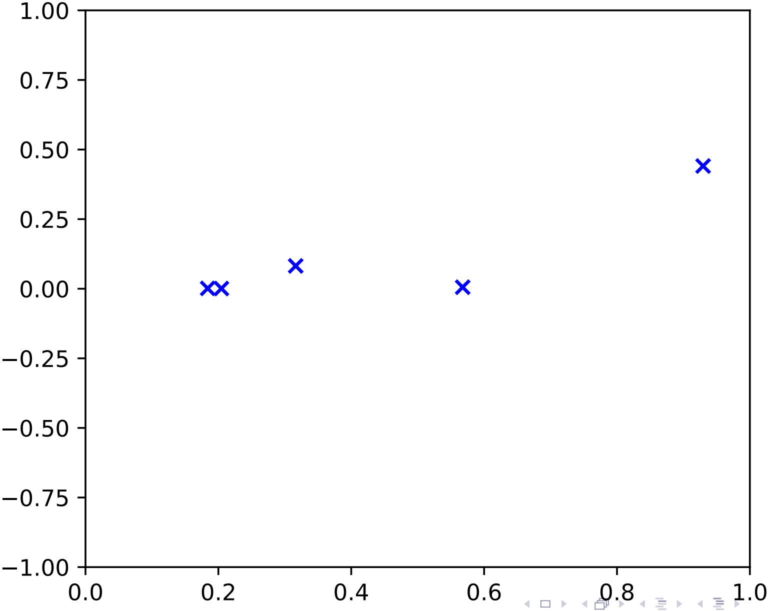
- Поскольку f — черный ящик, будем воспринимать f как случайную (введем prior $p(f)$)
- Будем обновлять prior с помощью теоремы Байеса, учитывая наблюдаемые значения $f(x_1), f(x_2), \ldots$

Байесовская оптимизация

- Поскольку f — черный ящик, будем воспринимать f как случайную (введем prior $p(f)$)
- Будем обновлять prior с помощью теоремы Байеса, учитывая наблюдаемые значения $f(x_1), f(x_2), \ldots$
- Будем искать нужный нам аргмаксимум x_* исходя из posterior

$$p(f(x_*)|f(x_1), f(x_2), \ldots)$$

Иллюстративный пример



Гауссовские процессы

- Нам нужно ввести какое-то априорное распределение на f

Гауссовские процессы

- Нам нужно ввести какое-то априорное распределение на f
- Какой подходящий prior выбрать?

Гауссовские процессы

- Нам нужно ввести какое-то априорное распределение на f
- Какой подходящий prior выбрать? Гауссовский процесс!

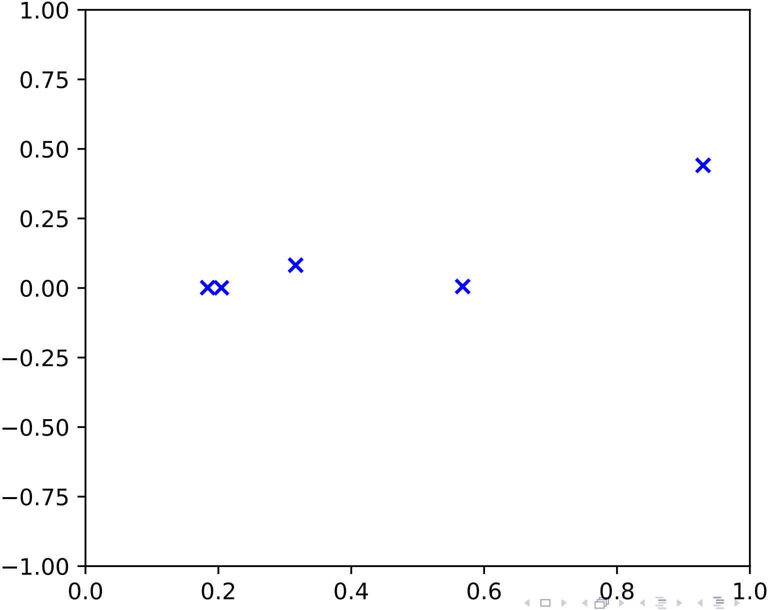
Гауссовские процессы

- Нам нужно ввести какое-то априорное распределение на f
- Какой подходящий prior выбрать? Гауссовский процесс!
- Используем информацию о всех имеющихся значениях f

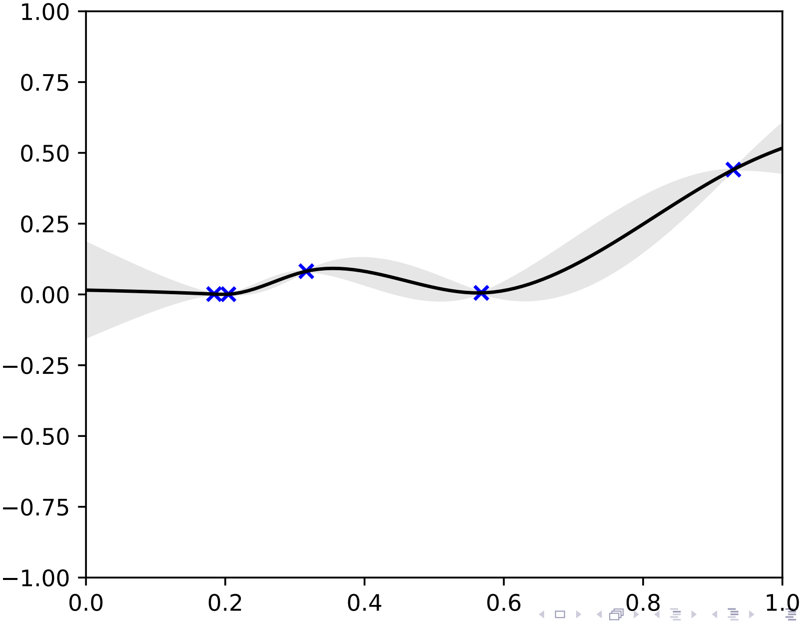
Гауссовские процессы

- Нам нужно ввести какое-то априорное распределение на f
- Какой подходящий prior выбрать? Гауссовский процесс!
- Используем информацию о всех имеющихся значениях f
- Используем апостериорное распределение, чтобы выбрать кандидата на максимум

Иллюстративный пример



Иллюстративный пример



Acquisition function

- Как выбрать кандидата на максимум?

Acquisition function

- Как выбрать кандидата на максимум?
- У нас есть целое апостериорное распределение, давайте используем его

Acquisition function

- Как выбрать кандидата на максимум?
- У нас есть целое апостериорное распределение, давайте используем его
- Например, выберем x_* такой, что

$$x_* = \operatorname{argmax}_{x \in \mathcal{X}} P(f(x) \geq f^+ + \varepsilon),$$

где f^+ – текущее наилучшее значение

Acquisition function

- Как выбрать кандидата на максимум?
- У нас есть целое апостериорное распределение, давайте используем его
- Например, выберем x_* такой, что

$$x_* = \text{argmax}_{x \in \mathcal{X}} P(f(x) \geq f^+ + \varepsilon),$$

  где f^+ – текущее наилучшее значение
- Функция

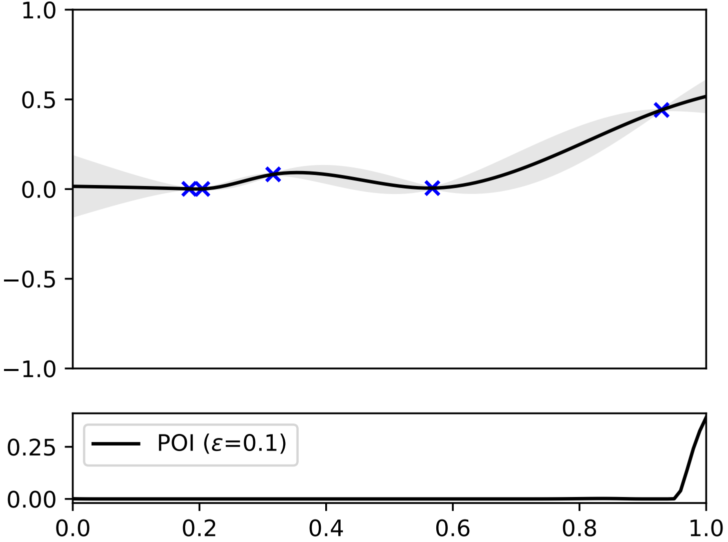
$$\alpha(x; f^+) := P(h(x) \geq f^+ + \varepsilon)$$

  называется Probability of Improvement

Acquisition function

- Как выбрать кандидата на максимум?
- У нас есть целое апостериорное распределение, давайте используем его
- Например, выберем x_* такой, что
$$x_* = \text{argmax}_{x \in \mathcal{X}} P(f(x) \geq f^+ + \varepsilon),$$
где f^+ – текущее наилучшее значение
- Функция
$$\alpha(x; f^+) := P(h(x) \geq f^+ + \varepsilon)$$
называется Probability of Improvement
- Для GP prior, Probability of Improvement считается аналитически

Иллюстративный пример



Acquisition function

- Более общо, вводится acqusition function $\alpha(x)$, которая

Acquisition function

- Более общо, вводится acqusition function $\alpha(x)$, которая
  - просто вычислима

Acquisition function

- Более общо, вводится acqusition function $\alpha(x)$, которая
  - просто вычислима
  - отражает наши представления о том, где находится кандидат на максимум, исходя из апостериорного распределения

Acquisition function

- Более общо, вводится acqusition function $\alpha(x)$, которая
  - просто вычислима
  - отражает наши представления о том, где находится кандидат на максимум, исходя из апостериорного распределения
- Примеры acquisition function

Acquisition function

- Более общо, вводится acqusition function $\alpha(x)$, которая
  - просто вычислима
  - отражает наши представления о том, где находится кандидат на максимум, исходя из апостериорного распределения
- Примеры acquisition function :
  - Expected Improvement

$$\alpha_{EI}(x) = \mathbb{E}(\max(0, h(x) - f^{+})|\mathcal{D})$$

Acquisition function

- Более общо, вводится acqusition function $\alpha(x)$, которая
  - просто вычислима
  - отражает наши представления о том, где находится кандидат на максимум, исходя из апостериорного распределения
- Примеры acquisition function :
  - Expected Improvement

$$\alpha_{EI}(x) = \mathbb{E}(\max(0, h(x) - f^{+})|\mathcal{D})$$

  - Upper Confidence Bound

$$\alpha_{UCB}(x) = \mu(x) + \lambda\sigma(x)$$

Acquisition function

- Более общо, вводится acqusition function $\alpha(x)$, которая
  - просто вычислима
  - отражает наши представления о том, где находится кандидат на максимум, исходя из апостериорного распределения
- Примеры acquisition function :
  - Expected Improvement

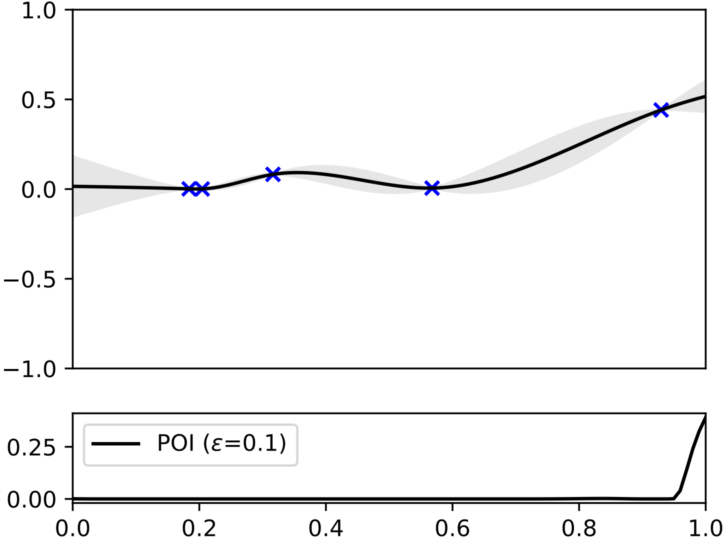
$$\alpha_{EI}(x) = \mathbb{E}(\max(0, h(x) - f^{+})|\mathcal{D})$$

  - Upper Confidence Bound

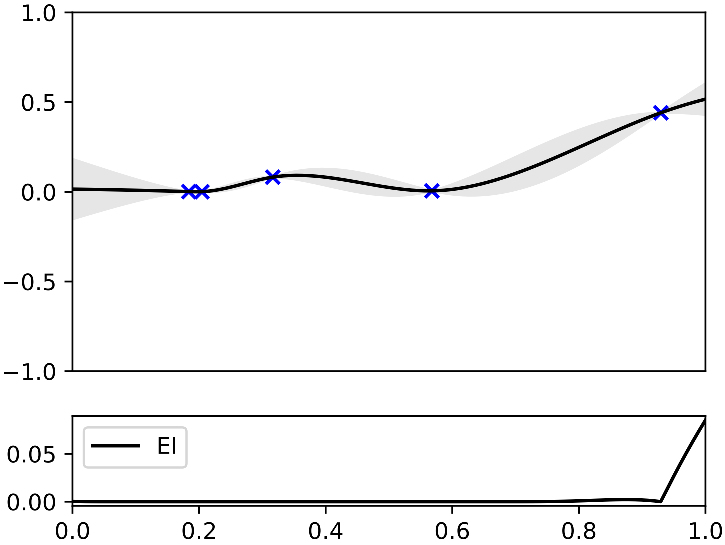
$$\alpha_{UCB}(x) = \mu(x) + \lambda\sigma(x)$$

  - И другие...

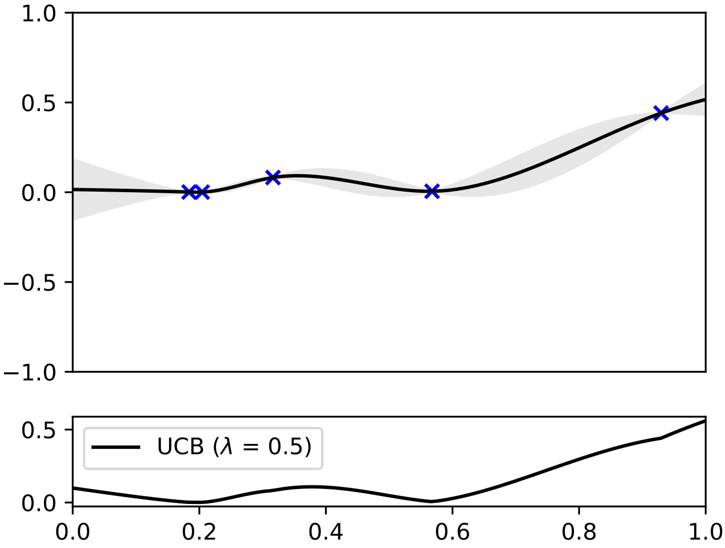
Иллюстративный пример



Иллюстративный пример



Иллюстративный пример



Формулы

Probability of Improvement

$$\alpha_{POI}(x; f^+) := P(h(x) \geq f^+ + \varepsilon) =$$

Формулы

Probability of Improvement

$$
\begin{aligned}
\alpha_{POI}(x; f^+) := P(h(x) \geq f^+ + \varepsilon) = \\
P(\mu(x) + \sigma(x) u \geq f^+ + \varepsilon) =
\end{aligned}
$$

Формулы

Probability of Improvement

$$
\begin{aligned}
\alpha_{POI}(x; f^+) := P(h(x) \geq f^+ + \varepsilon) = \\
P(\mu(x) + \sigma(x)u \geq f^+ + \varepsilon) = \\
P(u \geq \frac{f^+ + \varepsilon - \mu(x)}{\sigma(x)}) =
\end{aligned}
$$

Формулы

Probability of Improvement

$$\begin{aligned}\alpha_{POI}(x; f^+) := P(h(x) \geq f^+ + \varepsilon) =\\ P(\mu(x) + \sigma(x)u \geq f^+ + \varepsilon) =\\ P(u \geq \frac{f^+ + \varepsilon - \mu(x)}{\sigma(x)}) =\\ 1 - P\left(u \leq \frac{f^+ + \varepsilon - \mu(x)}{\sigma(x)}\right) =\end{aligned}$$

Формулы

Probability of Improvement

$$
\begin{aligned}
\alpha_{POI}(x; f^+) := P(h(x) \geq f^+ + \varepsilon) = \\
P(\mu(x) + \sigma(x)u \geq f^+ + \varepsilon) = \\
P(u \geq \frac{f^+ + \varepsilon - \mu(x)}{\sigma(x)}) = \\
1 - P\left(u \leq \frac{f^+ + \varepsilon - \mu(x)}{\sigma(x)}\right) = \\
\Phi\left(-\frac{f^+ + \varepsilon - \mu(x)}{\sigma(x)}\right)
\end{aligned}
$$

Формулы

Expected Improvement

$$\alpha_{EI}(x) = \mathbb{E}(\max(0, h(x) - f^{+}) =$$

Формулы

Expected Improvement

$$\alpha_{EI}(x) = \mathbb{E}(\max(0, h(x) - f^{+}) =$$
$$\int_{-\infty}^{\infty} \max(0, \mu(x) + \sigma(x)u - f^{+})\phi(u)du =$$

Формулы

Expected Improvement

$$\alpha_{EI}(x) = \mathbb{E}(\max(0, h(x) - f^+) =$$

$$\int_{-\infty}^{\infty} \max(0, \mu(x) + \sigma(x)u - f^+)\phi(u)du =$$

$$\left[z := \frac{f^+ - \mu(x)}{\sigma(x)}\right]$$

Формулы

Expected Improvement

$$\alpha_{EI}(x) = \mathbb{E}(\max(0, h(x) - f^{+}) =$$

$$\int_{-\infty}^{\infty} \max(0, \mu(x) + \sigma(x)u - f^{+})\phi(u)du =$$

$$\left[z := \frac{f^{+} - \mu(x)}{\sigma(x)}\right]$$

$$\int_{z}^{\infty} (\mu(x) - f^{+} + \sigma(x)u)\phi(u)du =$$

Формулы

Expected Improvement

$$\alpha_{EI}(x) = \mathbb{E}(\max(0, h(x) - f^{+}) =$$

$$\int_{-\infty}^{\infty} \max(0, \mu(x) + \sigma(x)u - f^{+})\phi(u)du =$$

$$\left[z := \frac{f^{+} - \mu(x)}{\sigma(x)} \right]$$

$$\int_{z}^{\infty} (\mu(x) - f^{+} + \sigma(x)u)\phi(u)du =$$

$$(\mu(x) - f^{+})\Phi(-z) + \sigma(x) \int_{z}^{\infty} u\phi(u)du =$$

Формулы

Expected Improvement

$$\alpha_{EI}(x) = \mathbb{E}(\max(0, h(x) - f^+) =$$

$$\int_{-\infty}^{\infty} \max(0, \mu(x) + \sigma(x)u - f^+)\phi(u)du =$$

$$\left[z := \frac{f^+ - \mu(x)}{\sigma(x)} \right]$$

$$\int_{z}^{\infty} (\mu(x) - f^+ + \sigma(x)u)\phi(u)du =$$

$$(\mu(x) - f^+)\Phi(-z) + \sigma(x) \int_{z}^{\infty} u\phi(u)du =$$

$$(\mu(x) - f^+)\Phi(-z) + \sigma(x)\phi(z)$$

Exploration vs Exploitation

Два режима оптимизации:

- Exploration. “Исследование” областей с высокой неопределенностью.

Exploration vs Exploitation

Два режима оптимизации:

- Exploration. “Исследование” областей с высокой неопределенностью.
- Exploitation. “Использование” знаний об области с правдоподобным максимумом.

Exploration vs Exploitation

Два режима оптимизации:

- Exploration. “Исследование” областей с высокой неопределенностью.
- Exploitation. “Использование” знаний об области с правдоподобным максимумом.
- Exploration/exploitation tradeoff — хотим исследовать области с высокой неопределенностью, но не хотим тратить слишком много ресурсов

Exploration vs Exploitation

Два режима оптимизации:

- Exploration. "Исследование" областей с высокой неопределенностью.
- Exploitation. "Использование" знаний об области с правдоподобным максимумом.
- Exploration/exploitation tradeoff — хотим исследовать области с высокой неопределенностью, но не хотим тратить слишком много ресурсов
- В UCB за tradeoff отвечает λ

$$\alpha_{UCB}(x) = \mu(x) + \lambda\sigma(x)$$

Стратегия поиска максимума

1. Warmup. Получим “начальные” значения $f(x_1), \dots, f(x_n)$.
   Датасет $\mathcal{D}_n := (x_i, y_i)$

Стратегия поиска максимума

1. Warmup. Получим “начальные” значения $f(x_1), \ldots, f(x_n)$.
   Датасет $\mathcal{D}_n := (x_i, y_i)$
2. Обучим гауссовский процесс на данных $\mathcal{D}_n$

Стратегия поиска максимума

1. Warmup. Получим “начальные” значения $f(x_1), \ldots, f(x_n)$. Датасет $\mathcal{D}_n := (x_i, y_i)$
2. Обучим гауссовский процесс на данных $\mathcal{D}_n$
3. Найдем кандидата на максимум $x_{n+1} := \operatorname{argmax}_{x \in \mathcal{X}} \alpha(x)$

Стратегия поиска максимума

1. Warmup. Получим "начальные" значения $f(x_1), \ldots, f(x_n)$. Датасет $\mathcal{D}_n := (x_i, y_i)$
2. Обучим гауссовский процесс на данных $\mathcal{D}_n$
3. Найдем кандидата на максимум $x_{n+1} := \text{argmax}_{x \in \mathcal{X}} \alpha(x)$
4. Вычислим значение y_{n+1} в x_{n+1}

Стратегия поиска максимума

1. Warmup. Получим “начальные” значения $f(x_1), \ldots, f(x_n)$.
   Датасет $\mathcal{D}_n := (x_i, y_i)$
2. Обучим гауссовский процесс на данных $\mathcal{D}_n$
3. Найдем кандидата на максимум $x_{n+1} := \mathrm{argmax}_{x \in \mathcal{X}} \alpha(x)$
4. Вычислим значение y_{n+1} в x_{n+1}
5. Добавим (x_{n+1}, y_{n+1}) в датасет $\mathcal{D}_{n+1}$

Стратегия поиска максимума

1. Warmup. Получим "начальные" значения $f(x_1), \ldots, f(x_n)$.
   Датасет $\mathcal{D}_n := (x_i, y_i)$
2. Обучим гауссовский процесс на данных $\mathcal{D}_n$
3. Найдем кандидата на максимум $x_{n+1} := \operatorname{argmax}_{x \in \mathcal{X}} \alpha(x)$
4. Вычислим значение y_{n+1} в x_{n+1}
5. Добавим (x_{n+1}, y_{n+1}) в датасет $\mathcal{D}_{n+1}$
6. Вернемся к пункту 2.

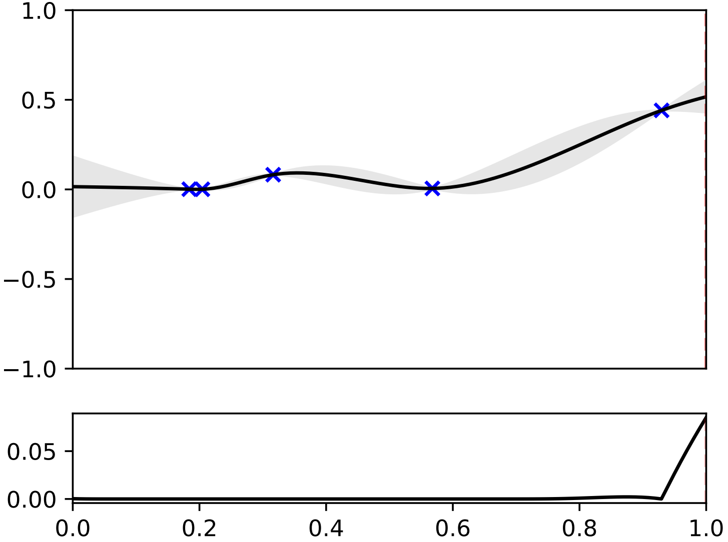
Стратегия поиска максимума

1. Warmup. Получим "начальные" значения $f(x_1), \dots, f(x_n)$. Датасет $\mathcal{D}_n := (x_i, y_i)$
2. Обучим гауссовский процесс на данных $\mathcal{D}_n$
3. Найдем кандидата на максимум $x_{n+1} := \operatorname{argmax}_{x \in \mathcal{X}} \alpha(x)$
4. Вычислим значение y_{n+1} в x_{n+1}
5. Добавим (x_{n+1}, y_{n+1}) в датасет $\mathcal{D}_{n+1}$
6. Вернемся к пункту 2.
7. ???

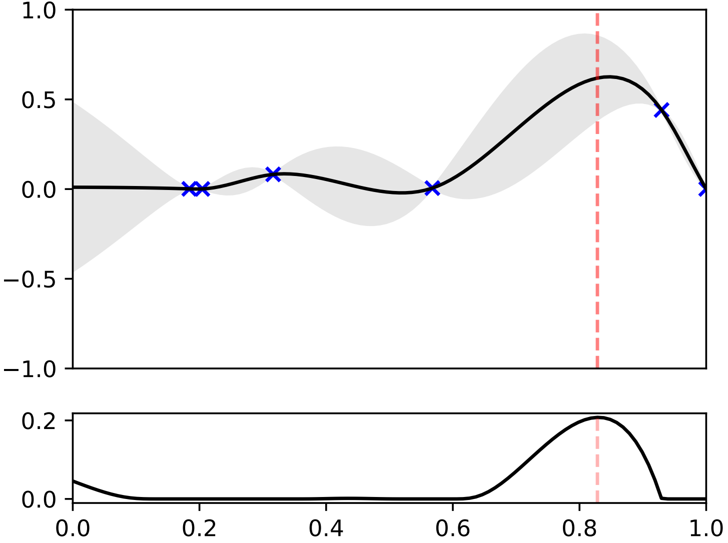
Стратегия поиска максимума

1. Warmup. Получим "начальные" значения $f(x_1), \dots, f(x_n)$. Датасет $\mathcal{D}_n := (x_i, y_i)$
2. Обучим гауссовский процесс на данных $\mathcal{D}_n$
3. Найдем кандидата на максимум $x_{n+1} := \operatorname{argmax}_{x \in \mathcal{X}} \alpha(x)$
4. Вычислим значение y_{n+1} в x_{n+1}
5. Добавим (x_{n+1}, y_{n+1}) в датасет $\mathcal{D}_{n+1}$
6. Вернемся к пункту 2.
7. ???
8. PROFIT!

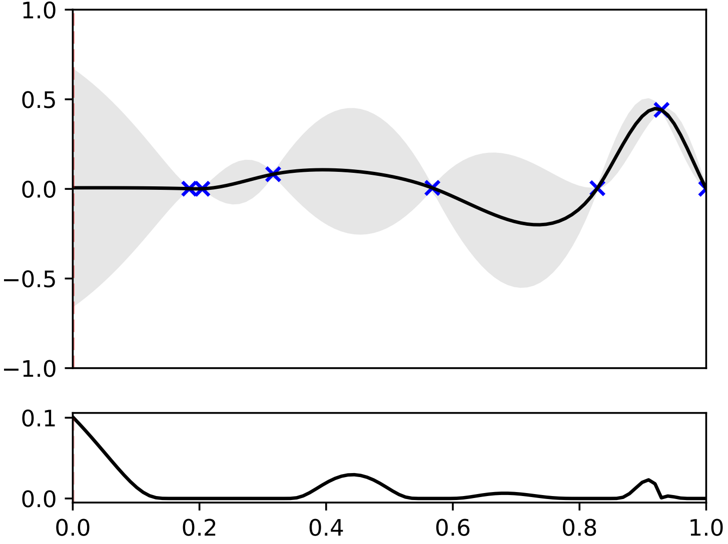
Иллюстративный пример



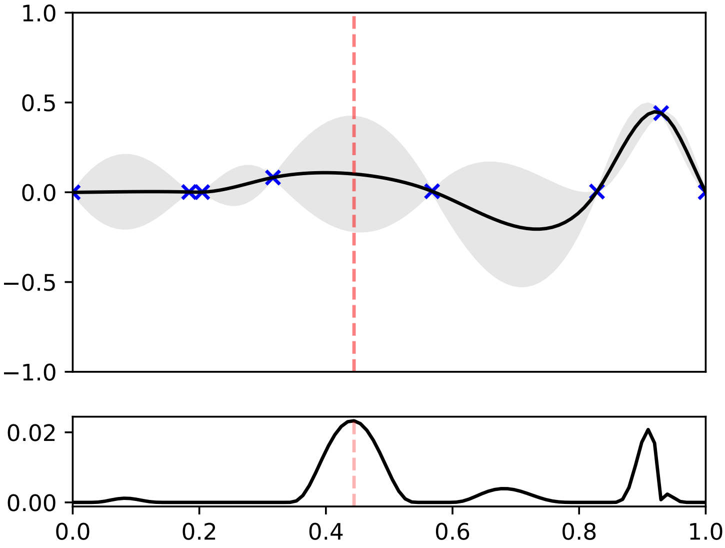
Иллюстративный пример



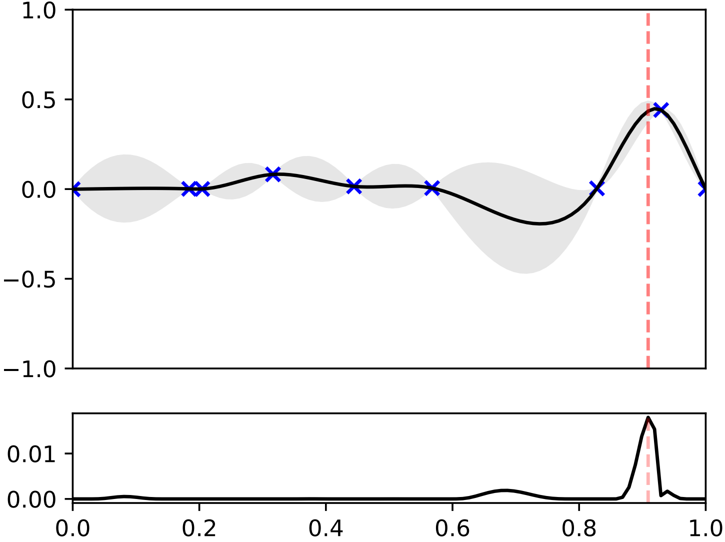
Иллюстративный пример



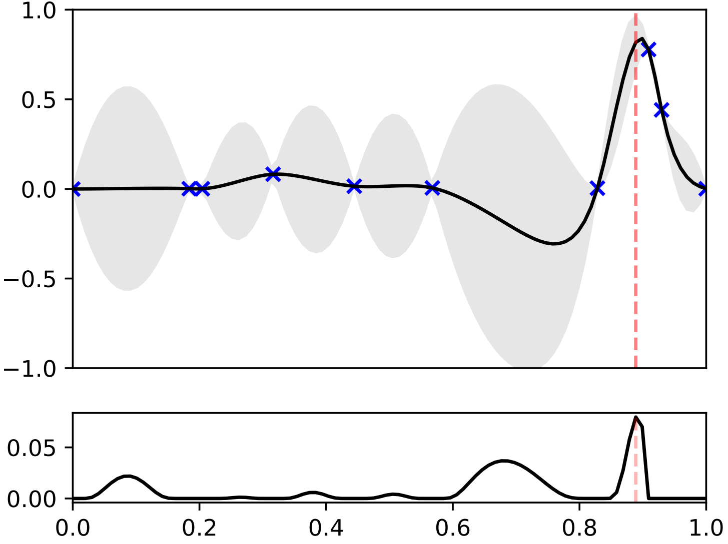
Иллюстративный пример



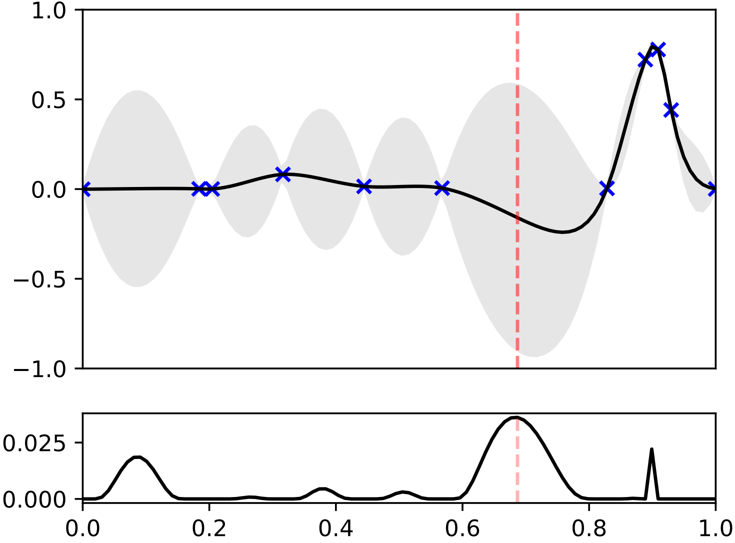
Иллюстративный пример



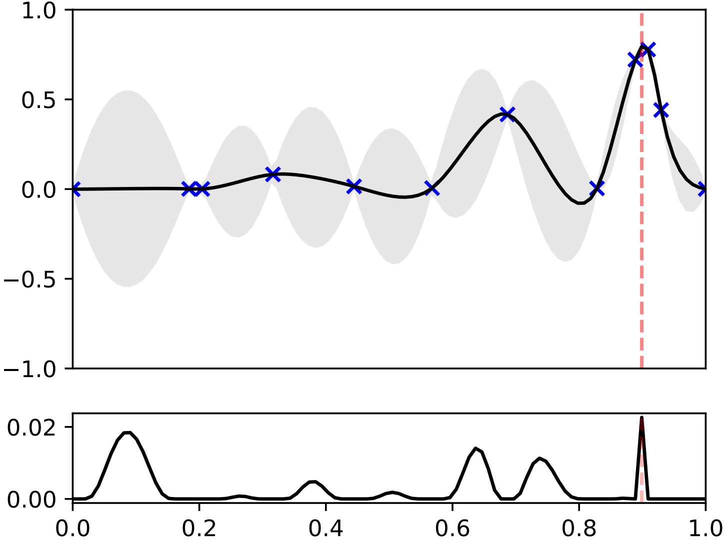
Иллюстративный пример



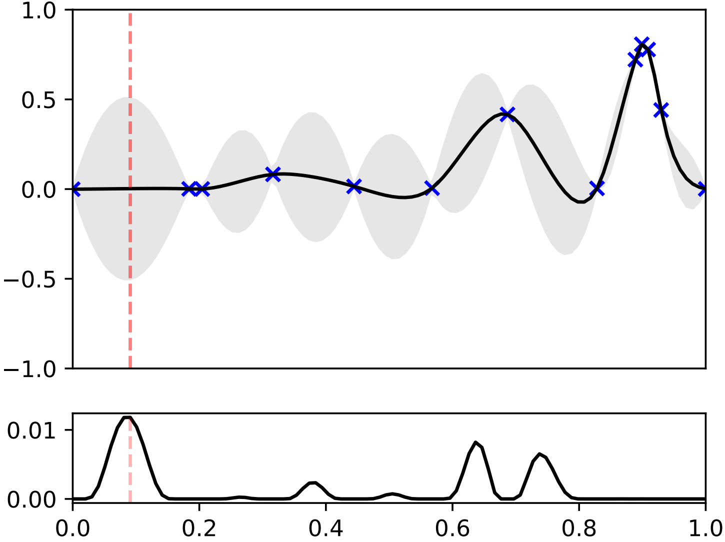
Иллюстративный пример



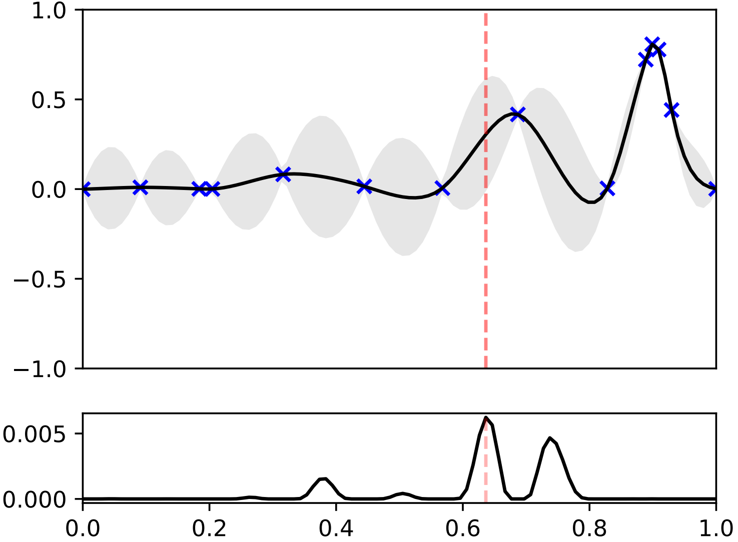
Иллюстративный пример



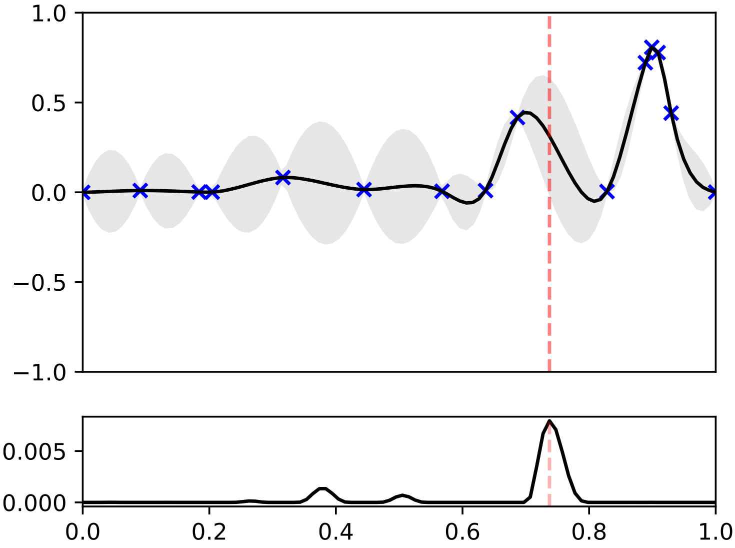
Иллюстративный пример



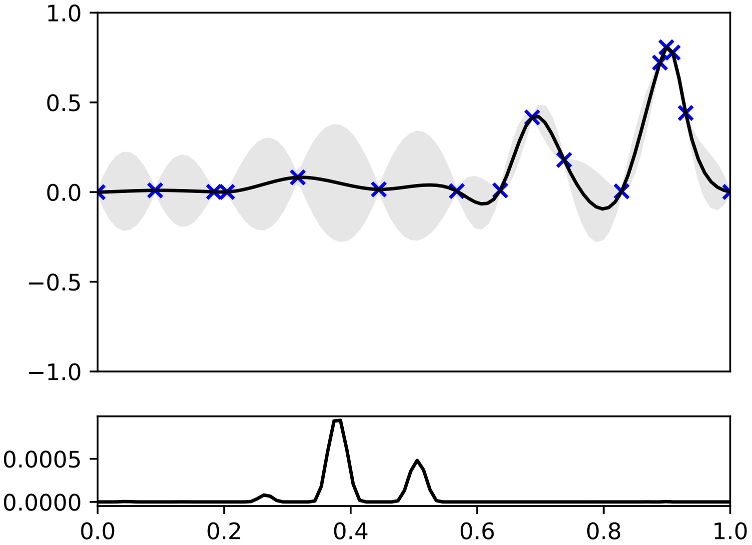
Иллюстративный пример



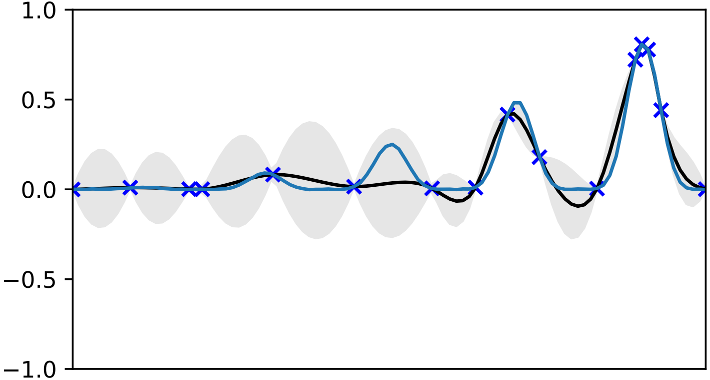
Иллюстративный пример



Иллюстративный пример



Иллюстративный пример



Thomspon Sampling

- f случайна, и ее максимум – случайная величина

$$x_* = \operatorname{argmax}_{x \in \mathcal{X}} f(x)$$
$$p(x_* | \mathcal{D}) = ?$$

где $\mathcal{D} = (x_i, y_i)_{i=1}^n$ — наблюдаемые данные

Thomspon Sampling

- f случайна, и ее максимум – случайная величина

$$x_* = \mathrm{argmax}_{x \in \mathcal{X}} f(x)$$
$$p(x_*|\mathcal{D}) = ?$$

  где $\mathcal{D} = (x_i, y_i)_{i=1}^n$ — наблюдаемые данные

- $$p(x_*|\mathcal{D}) = \int p(x_*|g, \mathcal{D}) p(g|\mathcal{D}) dg$$

  где g – гауссовский процесс, $p(x_*|g)$ — масса на аргмаксимуме траектории $g|\mathcal{D}$.

Thomspon Sampling

- f случайна, и ее максимум – случайная величина

$$x_* = \text{argmax}_{x \in \mathcal{X}} f(x)$$
$$p(x_*|\mathcal{D}) = ?$$

  где $\mathcal{D} = (x_i, y_i)_{i=1}^n$ — наблюдаемые данные

- $$p(x_*|\mathcal{D}) = \int p(x_*|g, \mathcal{D}) p(g|\mathcal{D}) dg$$

  где g – гауссовский процесс, $p(x_*|g)$ — масса на аргмаксимуме траектории $g|\mathcal{D}$.
- Thompson sampling

Thomspon Sampling

- f случайна, и ее максимум – случайная величина
$$x_* = \text{argmax}_{x \in \mathcal{X}} f(x)$$
$$p(x_* | \mathcal{D}) = ?$$
где $\mathcal{D} = (x_i, y_i)_{i=1}^n$ — наблюдаемые данные
-
$$p(x_* | \mathcal{D}) = \int p(x_* | g, \mathcal{D}) p(g | \mathcal{D}) dg$$
где g – гауссовский процесс, $p(x_* | g)$ — масса на аргмаксимуме траектории $g | \mathcal{D}$.
- Thompson sampling:
  - Сэмплируем траекторию $g | \mathcal{D}$

Thomspon Sampling

- f случайна, и ее максимум – случайная величина

$$x_* = \text{argmax}_{x \in \mathcal{X}} f(x)$$
$$p(x_*|\mathcal{D}) = ?$$

  где $\mathcal{D} = (x_i, y_i)_{i=1}^n$ — наблюдаемые данные

- $$p(x_*|\mathcal{D}) = \int p(x_*|g, \mathcal{D}) p(g|\mathcal{D}) dg$$

  где g – гауссовский процесс, $p(x_*|g)$ — масса на аргмаксимуме траектории $g|\mathcal{D}$.

- Thompson sampling:
  - Сэмплируем траекторию $g|\mathcal{D}$ (формула Матерона)

Thomspon Sampling

- f случайна, и ее максимум – случайная величина

$$x_* = \text{argmax}_{x \in \mathcal{X}} f(x)$$
$$p(x_*|\mathcal{D}) = ?$$

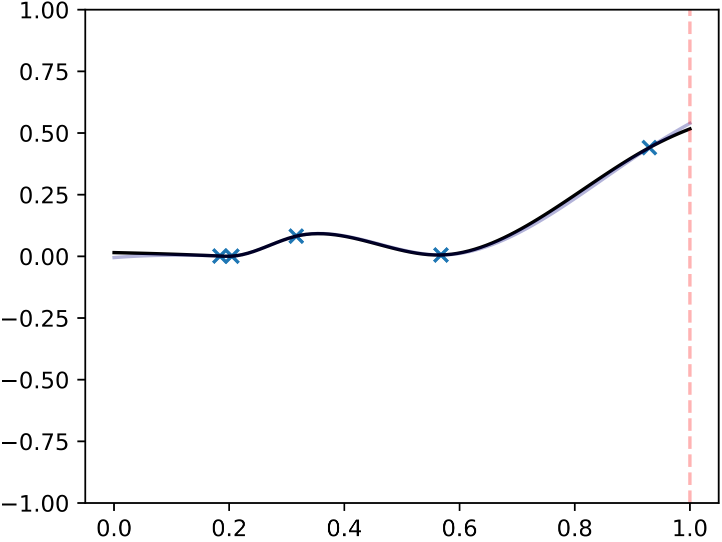
  где $\mathcal{D} = (x_i, y_i)_{i=1}^n$ — наблюдаемые данные

- $$p(x_*|\mathcal{D}) = \int p(x_*|g, \mathcal{D}) p(g|\mathcal{D}) dg$$

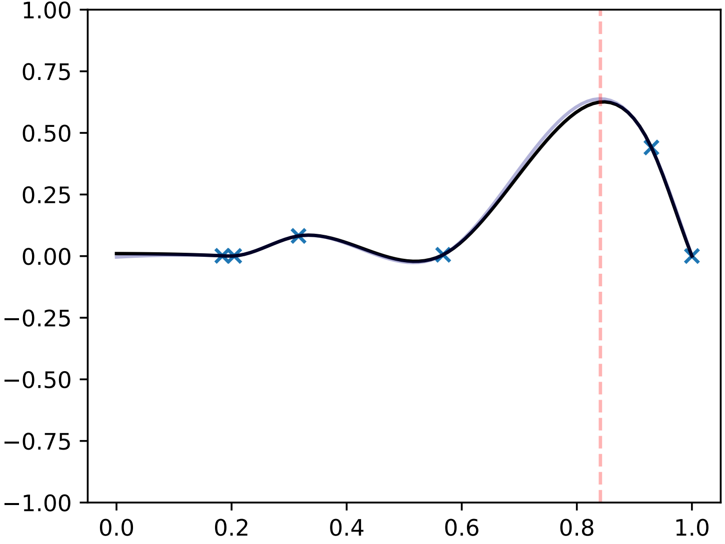
  где g – гауссовский процесс, $p(x_*|g)$ — масса на аргмаксимуме траектории $g|\mathcal{D}$.

- Thompson sampling:
  - Сэмплируем траекторию $g|\mathcal{D}$ (формула Матерона)
  - Находим ее аргмаксимум

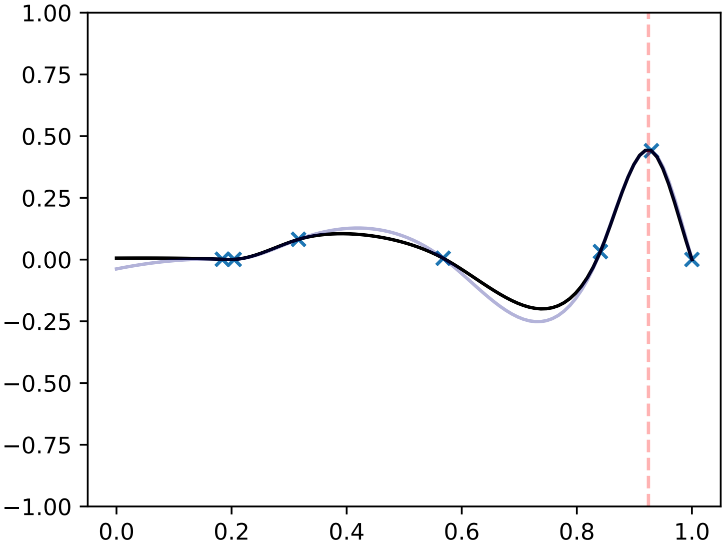
Иллюстративный пример



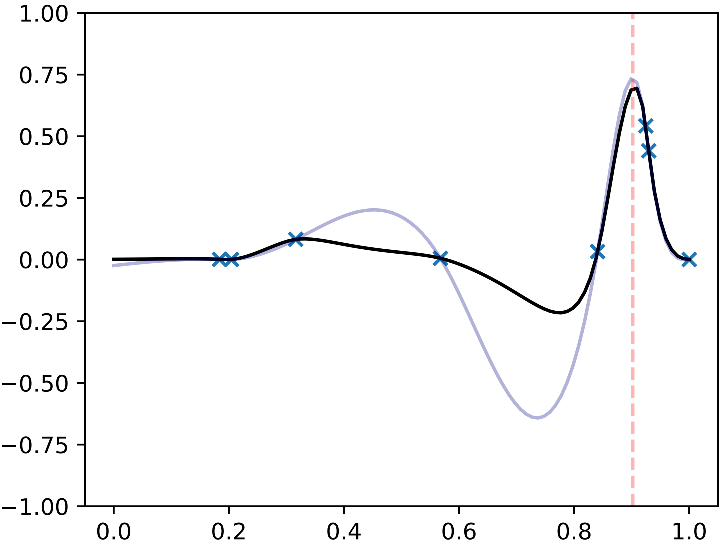
Иллюстративный пример



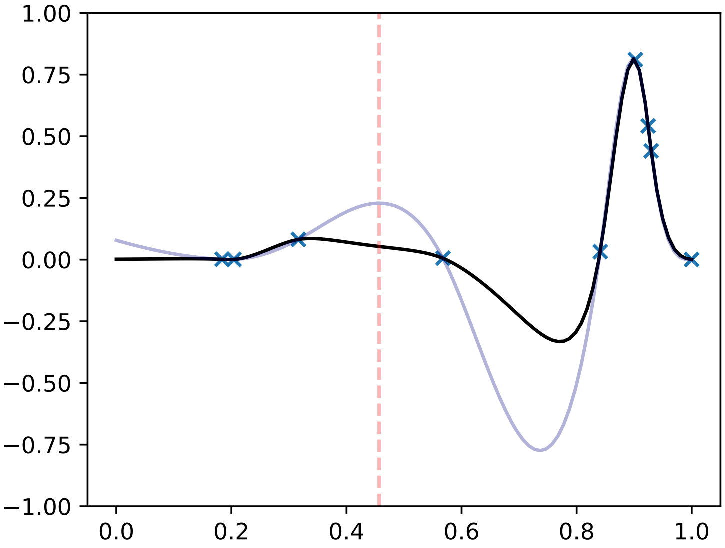
Иллюстративный пример



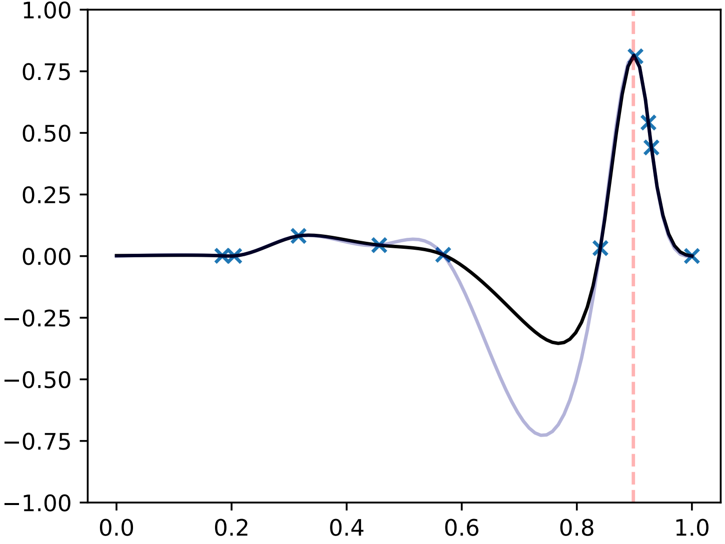
Иллюстративный пример



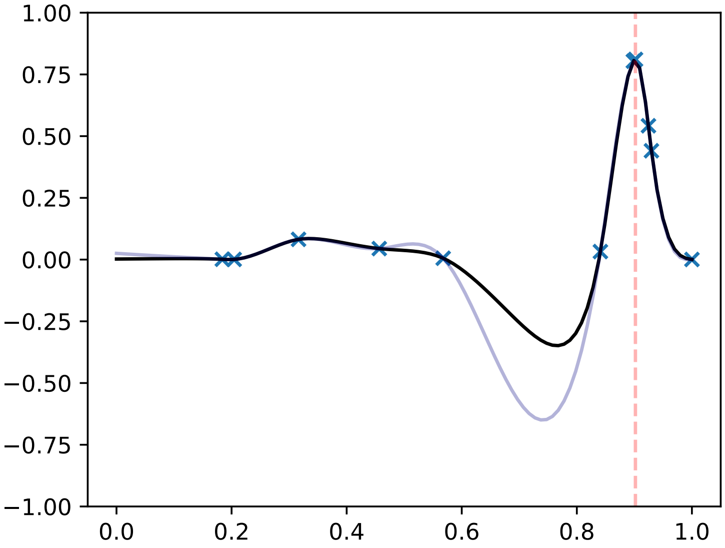
Иллюстративный пример



Иллюстративный пример



Иллюстративный пример



Regrets

- Regret
$$r_n := f(x_*) - f(x_n)$$
- Cumulative regret
$$R_n := \sum_{i=1}^{n} [f(x_*) - f(x_i)]$$
- No regret:
$$\frac{R_n}{n} \longrightarrow 0$$

No Silver Bullet

Выбор ядра для априорного гауссовского процесса

Нужно каким-то образом выбирать ядро для ГП — это вводит новые гиперпараметры, которые мы должны найти, оптимизируя MLL.

No Silver Bullet

Выбор ядра для априорного гауссовского процесса

Нужно каким-то образом выбирать ядро для ГП — это вводит новые гиперпараметры, которые мы должны найти, оптимизируя MLL.

Внутренняя оптимизация

Оптимизируем acquistion function, чтобы найти кандидата на минимум — нет гарантий сходимости к глобальному максимуму.

No Silver Bullet

Выбор ядра для априорного гауссовского процесса

Нужно каким-то образом выбирать ядро для ГП — это вводит новые гиперпараметры, которые мы должны найти, оптимизируя MLL.

Внутренняя оптимизация

Оптимизируем acquistion function, чтобы найти кандидата на минимум — нет гарантий сходимости к глобальному максимуму.

Проблема Large dimensions

Высокая размерность пространства поиска плохо дружит с байесовской оптимизацией. Rule of thumb — 20D.

ARD

- Стационарное ядро

$$k(x, x') = k(\|x - x'\|) = k\left(\sqrt{\sum_{d=1}^{D} (x_d - x'_d)^2}\right)$$

ARD

- Стационарное ядро

$$k(x, x') = k(\|x - x'\|) = k\left(\sqrt{\sum_{d=1}^{D}(x_d - x'_d)^2}\right)$$

- Параметр масштаба (lengthscale)

$$\frac{\sqrt{\sum_{d=1}^{D}(x_d - x'_d)^2}}{\ell}$$

ARD

- Стационарное ядро

$$k(x, x') = k(\|x - x'\|) = k\left(\sqrt{\sum_{d=1}^{D}(x_d - x'_d)^2}\right)$$

- Параметр масштаба (lengthscale)

$$\frac{\sqrt{\sum_{d=1}^{D}(x_d - x'_d)^2}}{\ell}$$

- Automatic Relevance Determination

$$\sqrt{\sum_{d=1}^{D}\left(\frac{x_d - x'_d}{\ell_d}\right)^2}$$

Спасибо!