

# Глава 15

## Вероятностные суррогатные модели

---

---

В предыдущей главе рассказывалось, как построить суррогатные модели по расчетным точкам. Используя суррогатные модели для оптимизации, иногда возникает необходимость количественно оценить уверенность в предсказаниях этих моделей. Один из способов — использовать вероятностный подход к суррогатному моделированию. Распространенной вероятностной суррогатной моделью является гауссовский процесс, который представляет распределение вероятностей по функциям. В этой главе будет объяснено, как использовать *гауссовские процессы*, чтобы вывести распределение по значениям различных проектных точек, учитывая значения ранее вычисленных расчетных точек. Мы обсудим, как включить информацию о градиенте, а также зашумленные измерения целевой функции. Поскольку предсказания, сделанные гауссовским процессом, регулируются набором параметров, мы покажем, как вывести эти параметры непосредственно из данных.

### 15.1. Нормальное распределение

Прежде чем вводить гауссовские процессы, рассмотрим некоторые соответствующие свойства многомерного нормального распределения.<sup>1</sup> Многомерное нормальное распределение имеет параметры — математическое ожидание  $\boldsymbol{\mu}$  и ковариационную матрицу  $\boldsymbol{\Sigma}$ . Плотность вероятности в точке  $\mathbf{x}$  задается формулой

$$N(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right). \quad (15.1)$$

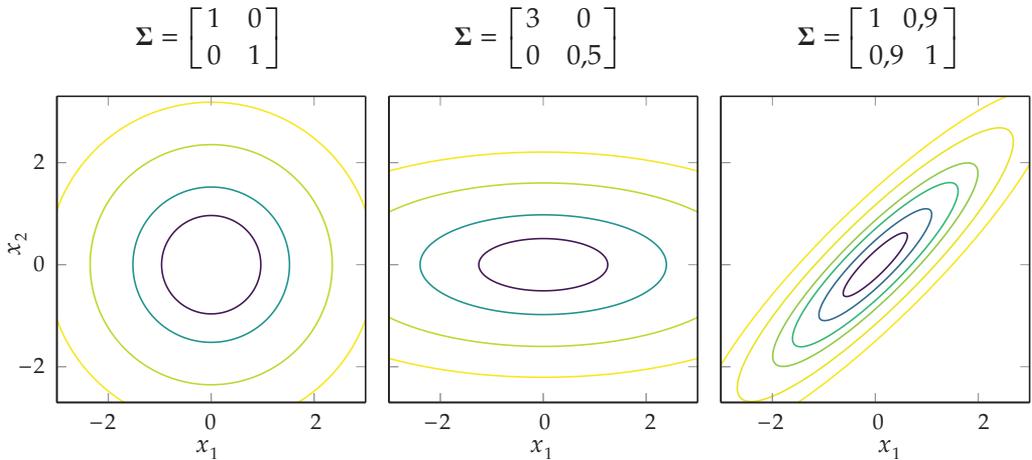
На рис. 15.1 показаны линии уровня функций плотности с различной ковариационной матрицей. Ковариационные матрицы всегда являются положительно полуопределенными.

Значение, выбранное из нормального распределения, записывается в виде

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \quad (15.2)$$

---

<sup>1</sup> Одномерное нормальное распределение обсуждается в приложении В.7.



**Рис. 15.1.** Многомерные нормальные распределения с разными ковариационными матрицами

Две случайные величины  $\mathbf{a}$  и  $\mathbf{b}$ , имеющие совместное нормальное распределение, можно записать в виде

$$\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B} \end{bmatrix}\right). \tag{15.3}$$

Маргинальное распределение<sup>2</sup> для вектора случайных величин задается соответствующими математическим ожиданием и ковариацией:

$$\mathbf{a} \sim \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{A}), \mathbf{b} \sim \mathcal{N}(\boldsymbol{\mu}_b, \mathbf{B}). \tag{15.4}$$

Условное распределение многомерного нормального распределения также можно представить в замкнутом виде:

$$\mathbf{a} | \mathbf{b} = \mathcal{N}(\boldsymbol{\mu}_{a|\mathbf{b}}, \boldsymbol{\Sigma}_{a|\mathbf{b}}), \tag{15.5}$$

$$\boldsymbol{\mu}_{a|\mathbf{b}} = \boldsymbol{\mu}_a + \mathbf{CB}^{-1}(\mathbf{b} - \boldsymbol{\mu}_b), \tag{15.6}$$

$$\boldsymbol{\Sigma}_{a|\mathbf{b}} = \mathbf{A} - \mathbf{CB}^{-1}\mathbf{C}^T. \tag{15.7}$$

Пример 15.1 иллюстрирует, как определить маргинальное и условное распределения многомерного нормального распределения.

**Пример 15.1.** Маргинальное и условное распределения многомерного нормального распределения

<sup>2</sup> Маргинальное распределение — это распределение подмножества переменных при условии, что остальные исключены путем интегрирования, т.е. маргинализированы. Для распределения по двум переменным  $a$  и  $b$  маргинальное распределение по  $a$  имеет вид:

$$p(a) = \int p(a, b)db.$$

Например, рассмотрим

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}\right).$$

Маргинальное распределение по  $x_1$  равно  $\mathcal{N}(0, 3)$ , а маргинальное распределение по  $x_2$  —  $\mathcal{N}(1, 2)$ .

Условное распределение по  $x_1$  при  $x_2 = 2$  определяется следующим образом.

$$\mu_{x_1|x_2=2} = 0 + 1 \cdot 2^{-1} \cdot (2 - 1) = 0,5;$$

$$\Sigma_{x_1|x_2=2} = 3 - 1 \cdot 2^{-1} \cdot 1 = 2,5;$$

$$x_1|(x_2 = 2) \sim \mathcal{N}(0,5; 2,5).$$

## 15.2. Гауссовские процессы

В предыдущей главе мы аппроксимировали целевую функцию  $f$ , используя суррогатную модельную функцию  $\hat{f}$ , приближенную по ранее найденным расчетным точкам. Специальный тип суррогатной модели, известный как *гауссовский процесс*, позволяет не только прогнозировать  $f$ , но и количественно определять неопределенность в этом прогнозе, используя распределение вероятностей.<sup>3</sup>

Гауссовский процесс — это распределение по функциям. Для любого конечного множества точек  $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}\}$  ассоциированные с ними значения функций  $\{y_1, \dots, y_m\}$  распределяются по формуле

$$\begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \sim N\left(\begin{bmatrix} m(\mathbf{x}^{(1)}) \\ \vdots \\ m(\mathbf{x}^{(m)}) \end{bmatrix}, \begin{bmatrix} k(\mathbf{x}^{(1)}, \mathbf{x}^{(1)}) & \dots & k(\mathbf{x}^{(1)}, \mathbf{x}^{(m)}) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}^{(m)}, \mathbf{x}^{(1)}) & \dots & k(\mathbf{x}^{(m)}, \mathbf{x}^{(m)}) \end{bmatrix}\right), \quad (15.8)$$

где  $m(\mathbf{x})$  — *функция математического ожидания*, а  $k(\mathbf{x}, \mathbf{x}')$  — *ковариационная функция*, или *ядро*.<sup>4</sup> Функция математического ожидания может представлять предварительные знания о функции. Ядро контролирует гладкость функций. Способы построения вектора математических ожиданий и ковариационной матрицы с использованием функции математического ожидания и ковариационной функции приведены в алгоритме 15.1.

<sup>3</sup> Более подробное введение в гауссовские процессы изложено в [126].

<sup>4</sup> Значением функции математического ожидания является математическое ожидание:  $m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$ , а значением ковариационной функции является ковариация:

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))].$$

**Алгоритм 15.1.** Функция  $\mu$  для построения вектора математических ожиданий при заданном списке расчетных точек и функция математического ожидания  $m$ , а также функция  $\Sigma$  для построения ковариационной матрицы при заданных одном или двух списках расчетных точек и ковариационной функции  $k$

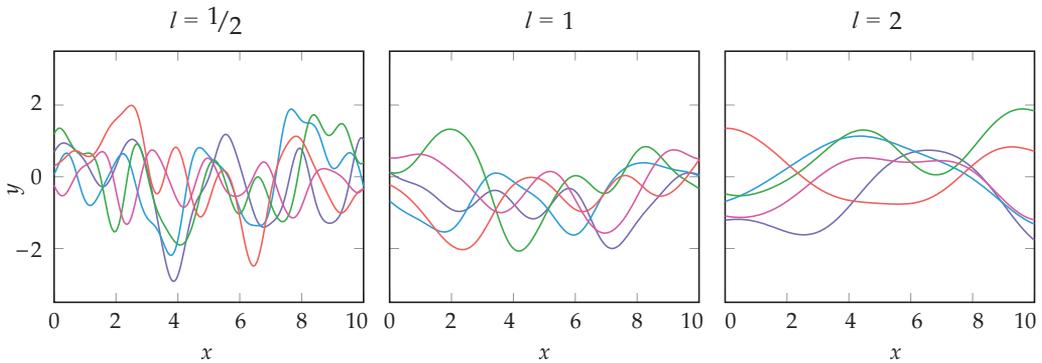
```

 $\mu(X, m) = [m(x) \text{ for } x \text{ in } X]$ 
 $\Sigma(X, k) = [k(x, x') \text{ for } x \text{ in } X, x' \text{ in } X]$ 
 $K(X, X', k) = [k(x, x') \text{ for } x \text{ in } X, x' \text{ in } X']$ 
    
```

Широко распространенной функцией ядра является *квадрат экспоненциального ядра* (squared exponential kernel), где

$$k(x, x') = \exp\left(-\frac{(x - x')^2}{2l^2}\right). \tag{15.9}$$

Параметр  $l$  соответствует так называемому *характеристическому масштабу длины*, рассматриваемой как расстояние, которое мы должны пройти в пространстве проектирования до тех пор, пока значение целевой функции существенно не изменится. Следовательно, большие значения  $l$  приводят к более гладким функциям. На рис. 15.2 показаны функции, выбранные из процесса Гаусса с функцией с нулевым математическим ожиданием и квадратом экспоненциального ядра с различными характеристическими масштабами длины.<sup>5</sup>

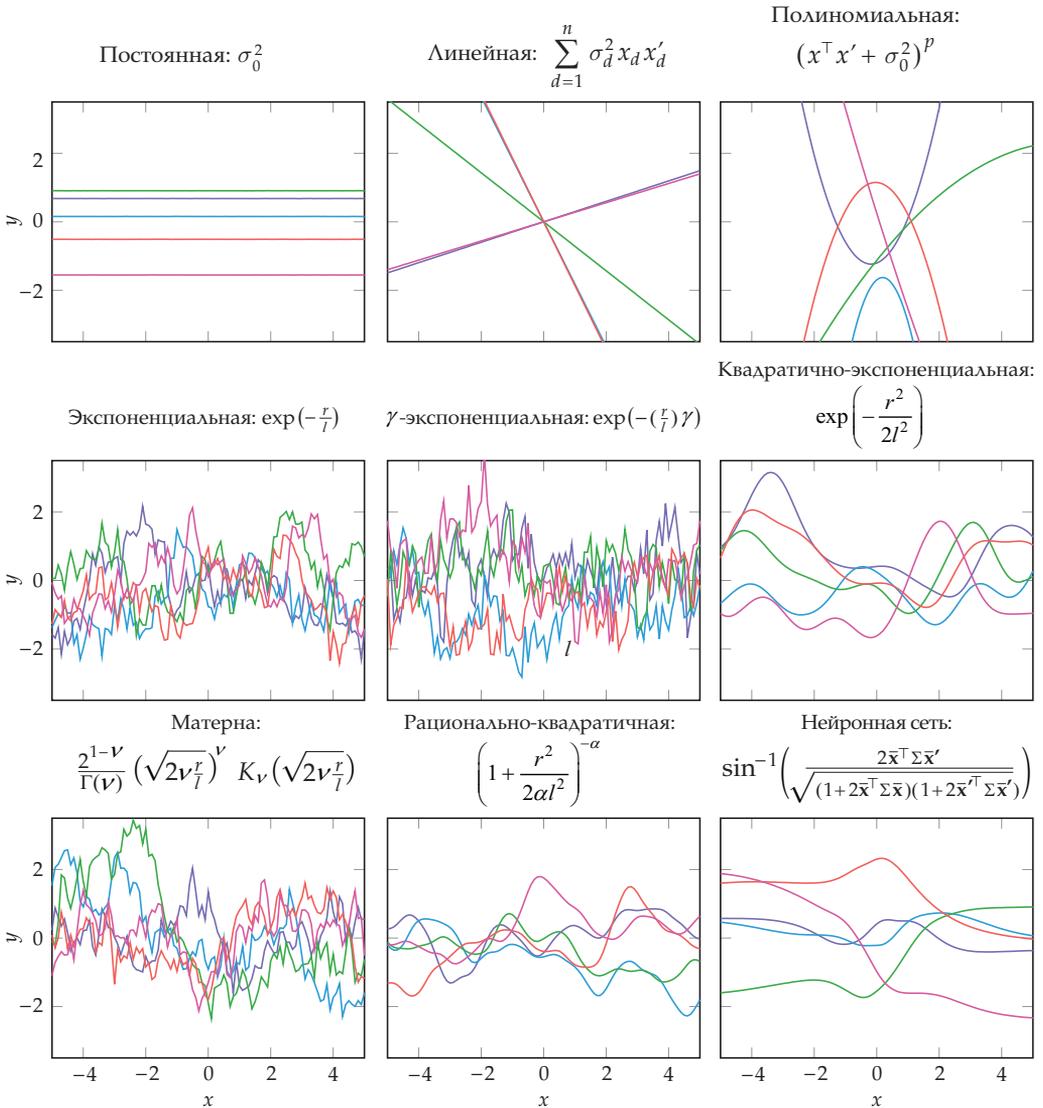


**Рис. 15.2.** Функции, извлеченные из гауссовских процессов с квадратами экспоненциальных ядер

Помимо квадрата экспоненты существует много других функций ядра. Некоторые из них показаны на рис. 15.3. Многие функции ядра используют параметр  $r$ , который является расстоянием между  $\mathbf{x}$  и  $\mathbf{x}'$ . Обычно используется евклидово расстояние. *Ядро Матерна* (Matérn kernel) применяет *гамма-функцию*  $\Gamma$ , реализо-

<sup>5</sup> Математическое определение характерной шкалы длины предоставлено в [126].

ванную функцией гамма, а  $K_\nu(x)$  — модифицированную функцию Бесселя второго рода, реализованную функцией `besselk` ( $\nu, x$ ). Ядро нейронной сети дополняет каждый расчетный вектор одним для простоты обозначения:  $\bar{\mathbf{x}} = [1, x_1, x_2, \dots]$  и  $\bar{\mathbf{x}}' = [1, x_1', x_2', \dots]$ .



**Рис. 15.3.** Функции, извлеченные из гауссовских процессов с различными функциями ядра. Показаны функции для  $\sigma_0^2 = \sigma_d^2 = l = 1, p = 2, \gamma = \nu = \alpha = 0,5$  и  $\Sigma = \mathbf{I}$

Для простоты эта глава будет сосредоточена на примерах гауссовских процессов в одномерных пространствах проектирования. Однако гауссовские процессы

могут быть определены в многомерном пространстве проектирования, как показано на рис. 15.4.

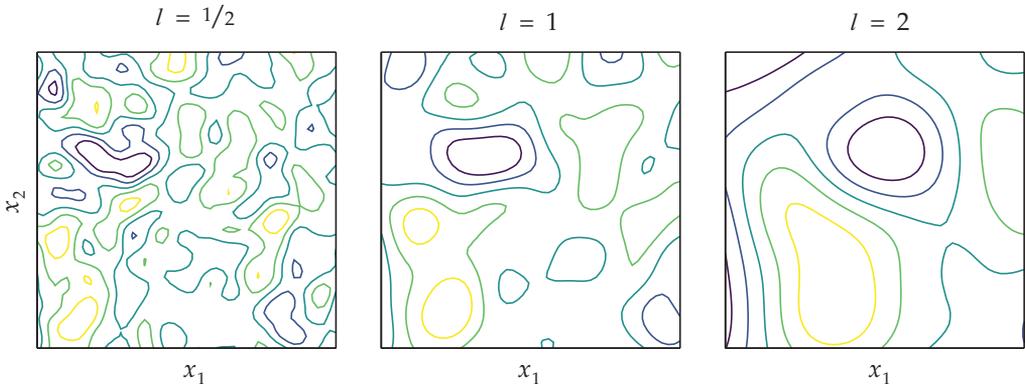


Рис. 15.4. Функции, извлеченные из гауссовских процессов

Как мы увидим в разделе 15.5, гауссовские процессы также могут включать в себя априорную зависимую дисперсию шума, обозначаемую как  $\nu$ . Таким образом, гауссовский процесс определяется функцией математического ожидания и ковариационной функцией, предыдущими расчетными точками и оценками их функций, а также дисперсией шума. Ассоциированный с ним тип приведен в алгоритме 15.2.

---

**Алгоритм 15.2.** Гауссовский процесс определяется функцией математического ожидания  $\mathbf{m}$ , ковариационной функцией  $\mathbf{k}$ , выборочными расчетными векторами  $\mathbf{X}$  и их соответствующими значениями целевой функции  $\mathbf{y}$  и дисперсией шума  $\nu$

---

```
mutable struct GaussianProcess
    m      # математическое ожидание
    k      # функция ковариации
    X      # расчетные точки
    y      # значения целевой точки
    nu     # дисперсия шума
end
```

---

### 15.3. Предсказание

Гауссовские процессы могут представлять распределения по функциям с использованием условных вероятностей. Предположим, у нас уже есть набор точек  $X$  и соответствующий вектор  $\mathbf{y}$ , но мы хотим предсказать значения  $\hat{\mathbf{y}}$  в точках  $X^*$ . Совместное распространение таково:

$$\begin{bmatrix} \hat{\mathbf{y}} \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \mathbf{m}(X^*) \\ \mathbf{m}(X) \end{bmatrix}, \begin{pmatrix} \mathbf{K}(X^*, X^*) & \mathbf{K}(X^*, X) \\ \mathbf{K}(X, X^*) & \mathbf{K}(X, X) \end{pmatrix} \right). \quad (15.10)$$

В приведенном выше уравнении мы используем функции  $\mathbf{m}$  и  $\mathbf{K}$ , которые определены следующим образом:

$$\mathbf{m}(X) = [m(\mathbf{x}^{(1)}), \dots, m(\mathbf{x}^{(n)})], \quad (15.11)$$

$$\mathbf{K}(X, X') = \begin{bmatrix} k(\mathbf{x}^{(1)}, \mathbf{x}'^{(1)}) & \dots & k(\mathbf{x}^{(1)}, \mathbf{x}'^{(m)}) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}^{(n)}, \mathbf{x}'^{(1)}) & \dots & k(\mathbf{x}^{(n)}, \mathbf{x}'^{(m)}) \end{bmatrix}. \quad (15.12)$$

Условное распределение дается формулой

$$\hat{\mathbf{y}}|\mathbf{y} \sim \mathcal{N} \left( \underbrace{\mathbf{m}(X^*) + \mathbf{K}(X^*, X)\mathbf{K}(X, X)^{-1}(\mathbf{y} - \mathbf{m}(X))}_{\text{Математическое ожидание}}, \underbrace{\mathbf{K}(X^*, X^*) - \mathbf{K}(X^*, X)\mathbf{K}(X, X)^{-1}\mathbf{K}(X, X^*)}_{\text{Ковариация}} \right). \quad (15.13)$$

Обратите внимание на то, что ковариация не зависит от  $\mathbf{y}$ . Это распределение часто называют апостериорным распределением.<sup>6</sup> Метод вычисления и выборки из апостериорного распределения, определенного гауссовским процессом, приведен в алгоритме 15.3.

---

**Алгоритм 15.3.** Функция `mvnrand` осуществляет выборку из многомерного нормального распределения с добавленным коэффициентом инфляции для предотвращения вычислительных проблем. Метод `rand` позволяет выбрать гауссовский процесс `GP` в заданных расчетных точках в матрице `X`

---

```
function mvnrand(μ, Σ, inflation = 1e-6)
    N = MvNormal(μ, Σ + inflation * I)
    return rand(N)
end
Base.rand(GP, X) = mvnrand(μ(X, GP.m), Σ(X, GP.k))
```

---

<sup>6</sup> На языке байесовской статистики апостериорное распределение — это распределение возможных ненаблюдаемых значений, обусловленных наблюдаемыми значениями.

Прогнозируемое математическое ожидание можно записать как функцию от  $\mathbf{x}$ :

$$\hat{\mu}(\mathbf{x}) = m(\mathbf{x}) + \mathbf{K}(\mathbf{x}, X)\mathbf{K}(X, X)^{-1}(\mathbf{y} - \mathbf{m}(X)) = \tag{15.14}$$

$$= m(\mathbf{x}) + \boldsymbol{\theta}^T \mathbf{K}(X, \mathbf{x}), \tag{15.15}$$

где  $\boldsymbol{\theta} = \mathbf{K}(X, X)^{-1}(\mathbf{y} - \mathbf{m}(X))$  можно вычислить один раз и повторно использовать для различных значений  $\mathbf{x}$ . Обратите внимание на сходство с суррогатными моделями в предыдущей главе. Ценность гауссовского процесса за пределами суррогатных моделей, обсуждавшихся ранее, состоит в том, что он также количественно определяет нашу неуверенность в предсказаниях.

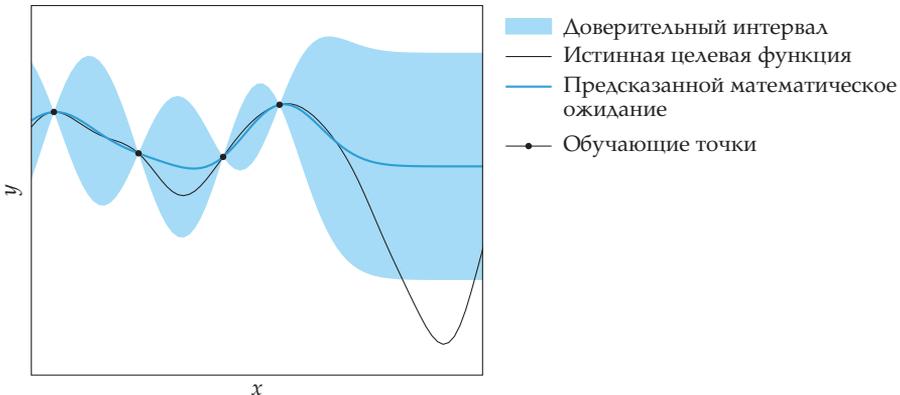
Дисперсия предсказанного математического ожидания также может быть получена как функция от  $\mathbf{x}$ :

$$\hat{v}(\mathbf{x}) = \mathbf{K}(\mathbf{x}, \mathbf{x}) - \mathbf{K}(\mathbf{x}, X)\mathbf{K}(X, X)^{-1}\mathbf{K}(X, \mathbf{x}). \tag{15.16}$$

В некоторых случаях удобнее составлять уравнения в терминах стандартного отклонения, которое является квадратным корнем из дисперсии:

$$\hat{\sigma}(\mathbf{x}) = \sqrt{\hat{v}(\mathbf{x})}. \tag{15.17}$$

Стандартное отклонение имеет те же единицы измерения, что и математическое ожидание. По стандартному отклонению мы можем вычислить 95%-ную *доверительную область*, которая представляет собой интервал, содержащий 95% вероятностной массы, связанной с распределением по  $y$  при заданном  $\mathbf{x}$ . Для конкретного  $\mathbf{x}$  95%-ная доверительная область определяется как  $\hat{\mu}(\mathbf{x}) \pm 1,96\hat{\sigma}(\mathbf{x})$ . Можно использовать доверительный уровень, отличный от 95%, но для графиков в этой главе мы будем использовать 95%. На рис. 15.5 показан график доверительной области, связанной с гауссовым процессом, подходящей для четырех оценок функций.



**Рис. 15.5.** Гауссовский процесс, использующий квадрат экспоненциального ядра и его 95%-ный доверительный интервал. Неопределенность возрастает по мере удаления от точки данных, и ожидаемое значение функции приближается к нулю, когда точка удаляется от данных

## 15.4. Информация о градиенте

В гауссовские процессы можно включить информацию о градиенте (см., например, [116]), расширив его определение, чтобы учесть как значение функции, так и ее градиент:

$$\begin{bmatrix} \mathbf{y} \\ \nabla \mathbf{y} \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \mathbf{m}_f \\ \mathbf{m}_\nabla \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{ff} & \mathbf{K}_{f\nabla} \\ \mathbf{K}_{\nabla f} & \mathbf{K}_{\nabla\nabla} \end{bmatrix} \right), \quad (15.18)$$

где  $\mathbf{y} \sim \mathcal{N}(\mathbf{m}_f, \mathbf{K}_{ff})$  — это традиционный гауссовский процесс,  $\mathbf{m}_\nabla$  является ковариационной матрицей между значениями функции,<sup>7</sup>  $\mathbf{K}_{\nabla f}$  — ковариационная матрица между градиентами функции и значениями функции, а  $\mathbf{K}_{\nabla\nabla}$  — ковариационная матрица между градиентами функций.

Эти ковариационные матрицы строятся с использованием ковариационных функций. Линейность нормальных распределений обеспечивает следующие зависимости между ковариационными функциями:

$$k_{ff}(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}'), \quad (15.19)$$

$$k_{\nabla f}(\mathbf{x}, \mathbf{x}') = \nabla_{\mathbf{x}} k(\mathbf{x}, \mathbf{x}'), \quad (15.20)$$

$$k_{f\nabla}(\mathbf{x}, \mathbf{x}') = \nabla_{\mathbf{x}'} k(\mathbf{x}, \mathbf{x}') \quad (15.21)$$

$$k_{\nabla\nabla}(\mathbf{x}, \mathbf{x}') = \nabla_{\mathbf{x}} \nabla_{\mathbf{x}'} k(\mathbf{x}, \mathbf{x}') \quad (15.22)$$

В примере 15.2 эти отношения используются для получения ковариационных функций высшего порядка для конкретного ядра.

**Пример 15.2.** Вывод ковариационных функций для гауссовского процесса с градиентными наблюдениями

Рассмотрим квадрат экспоненциальной ковариационной функции

$$k_{ff}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2}\|\mathbf{x} - \mathbf{x}'\|^2\right).$$

Мы можем использовать уравнения (15.19)–(15.22), чтобы получить другие ковариационные функции, необходимые для использования гауссовских процессов с градиентной информацией:

$$k_{\nabla f}(\mathbf{x}, \mathbf{x}') = -(\mathbf{x}_i - \mathbf{x}'_i) \exp\left(-\frac{1}{2}\|\mathbf{x} - \mathbf{x}'\|^2\right),$$

$$k_{\nabla\nabla}(\mathbf{x}, \mathbf{x}')_{ij} = -((i = j) - (\mathbf{x}_i - \mathbf{x}'_i)(\mathbf{x}_j - \mathbf{x}'_j)) \exp\left(-\frac{1}{2}\|\mathbf{x} - \mathbf{x}'\|^2\right).$$

<sup>7</sup> Аналогично среднему значению функции, значение  $\mathbf{m}_\nabla$  часто равно нулю.

Напоминаем, что логические выражения, такие как  $(i = j)$ , возвращают единицу, если их значение — истина, и нуль, если ложь.

Прогнозирование может быть выполнено так же, как и при традиционном гауссовском процессе. Сначала мы строим совместное распределение

$$\begin{pmatrix} \hat{\mathbf{y}} \\ \mathbf{y} \\ \nabla \mathbf{y} \end{pmatrix} \sim N \left( \begin{bmatrix} \mathbf{m}_f(X^*) \\ \mathbf{m}_f(X) \\ \mathbf{m}_\nabla(X) \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{ff}(X^*, X^*) & \mathbf{K}_{ff}(X^*, X) & \mathbf{K}_{f\nabla}(X^*, X) \\ \mathbf{K}_{ff}(X^*, X) & \mathbf{K}_{ff}(X, X) & \mathbf{K}_{f\nabla}(X, X) \\ \mathbf{K}_{\nabla f}(X, X^*) & \mathbf{K}_{\nabla f}(X, X) & \mathbf{K}_{\nabla\nabla}(X, X) \end{bmatrix} \right). \quad (15.23)$$

Для гауссовского процесса над  $n$ -мерными расчетными векторами при заданных значениях функций и градиентов и  $l$  тестовых точках блоки ковариантной матрицы имеют следующие размерности:

$$\begin{array}{ccc} l \times l & l \times m & l \times nm \\ m \times l & m \times m & m \times nm \\ nm \times l & nm \times m & nm \times nm \end{array} \quad (15.24)$$

Построение такой ковариационной матрицы описано в примере 15.3.

**Пример 15.3.** Построение ковариационной матрицы для гауссовского процесса с градиентными наблюдениями

Предположим, что мы вычислили функцию и ее градиент в двух точках,  $\mathbf{x}^{(1)}$  и  $\mathbf{x}^{(2)}$ , и хотим предсказать значение функции в точке  $\hat{\mathbf{x}}$ . Мы можем вывести совместное распределение по  $\hat{\mathbf{y}}$ ,  $\mathbf{y}$  и  $\nabla \mathbf{y}$ , используя гауссовский процесс. Ковариационная матрица имеет вид:

$$\begin{bmatrix} k_{ff}(\hat{\mathbf{x}}, \hat{\mathbf{x}}) & k_{ff}(\hat{\mathbf{x}}, \hat{\mathbf{x}}^{(1)}) & k_{ff}(\hat{\mathbf{x}}, \hat{\mathbf{x}}^{(2)}) & k_{f\nabla}(\hat{\mathbf{x}}, \hat{\mathbf{x}}^{(1)})_1 & k_{f\nabla}(\hat{\mathbf{x}}, \hat{\mathbf{x}}^{(1)})_2 & k_{f\nabla}(\hat{\mathbf{x}}, \hat{\mathbf{x}}^{(2)})_1 & k_{f\nabla}(\hat{\mathbf{x}}, \hat{\mathbf{x}}^{(2)})_2 \\ k_{ff}(\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{x}}) & k_{ff}(\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{x}}^{(1)}) & k_{ff}(\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{x}}^{(2)}) & k_{f\nabla}(\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{x}}^{(1)})_1 & k_{f\nabla}(\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{x}}^{(1)})_2 & k_{f\nabla}(\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{x}}^{(2)})_1 & k_{f\nabla}(\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{x}}^{(2)})_2 \\ k_{ff}(\hat{\mathbf{x}}^{(2)}, \hat{\mathbf{x}}) & k_{ff}(\hat{\mathbf{x}}^{(2)}, \hat{\mathbf{x}}^{(1)}) & k_{ff}(\hat{\mathbf{x}}^{(2)}, \hat{\mathbf{x}}^{(2)}) & k_{f\nabla}(\hat{\mathbf{x}}^{(2)}, \hat{\mathbf{x}}^{(1)})_1 & k_{f\nabla}(\hat{\mathbf{x}}^{(2)}, \hat{\mathbf{x}}^{(1)})_2 & k_{f\nabla}(\hat{\mathbf{x}}^{(2)}, \hat{\mathbf{x}}^{(2)})_1 & k_{f\nabla}(\hat{\mathbf{x}}^{(2)}, \hat{\mathbf{x}}^{(2)})_2 \\ k_{\nabla f}(\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{x}})_1 & k_{\nabla f}(\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{x}}^{(1)})_1 & k_{\nabla f}(\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{x}}^{(2)})_1 & k_{\nabla\nabla}(\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{x}}^{(1)})_{11} & k_{\nabla\nabla}(\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{x}}^{(1)})_{12} & k_{\nabla\nabla}(\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{x}}^{(2)})_{11} & k_{\nabla\nabla}(\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{x}}^{(2)})_{12} \\ k_{\nabla f}(\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{x}})_2 & k_{\nabla f}(\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{x}}^{(1)})_2 & k_{\nabla f}(\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{x}}^{(2)})_2 & k_{\nabla\nabla}(\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{x}}^{(1)})_{21} & k_{\nabla\nabla}(\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{x}}^{(1)})_{22} & k_{\nabla\nabla}(\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{x}}^{(2)})_{21} & k_{\nabla\nabla}(\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{x}}^{(2)})_{22} \\ k_{\nabla f}(\hat{\mathbf{x}}^{(2)}, \hat{\mathbf{x}})_1 & k_{\nabla f}(\hat{\mathbf{x}}^{(2)}, \hat{\mathbf{x}}^{(1)})_1 & k_{\nabla f}(\hat{\mathbf{x}}^{(2)}, \hat{\mathbf{x}}^{(2)})_1 & k_{\nabla\nabla}(\hat{\mathbf{x}}^{(2)}, \hat{\mathbf{x}}^{(1)})_{11} & k_{\nabla\nabla}(\hat{\mathbf{x}}^{(2)}, \hat{\mathbf{x}}^{(1)})_{12} & k_{\nabla\nabla}(\hat{\mathbf{x}}^{(2)}, \hat{\mathbf{x}}^{(2)})_{11} & k_{\nabla\nabla}(\hat{\mathbf{x}}^{(2)}, \hat{\mathbf{x}}^{(2)})_{12} \\ k_{\nabla f}(\hat{\mathbf{x}}^{(2)}, \hat{\mathbf{x}})_2 & k_{\nabla f}(\hat{\mathbf{x}}^{(2)}, \hat{\mathbf{x}}^{(1)})_2 & k_{\nabla f}(\hat{\mathbf{x}}^{(2)}, \hat{\mathbf{x}}^{(2)})_2 & k_{\nabla\nabla}(\hat{\mathbf{x}}^{(2)}, \hat{\mathbf{x}}^{(1)})_{21} & k_{\nabla\nabla}(\hat{\mathbf{x}}^{(2)}, \hat{\mathbf{x}}^{(1)})_{22} & k_{\nabla\nabla}(\hat{\mathbf{x}}^{(2)}, \hat{\mathbf{x}}^{(2)})_{21} & k_{\nabla\nabla}(\hat{\mathbf{x}}^{(2)}, \hat{\mathbf{x}}^{(2)})_{22} \end{bmatrix}$$

Условное распределение подчиняется тем же гауссовским соотношениям, что и в уравнении (15.13):

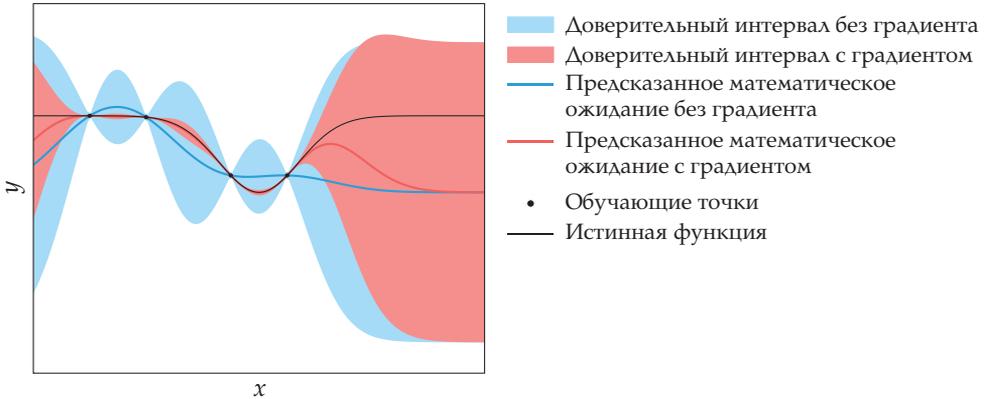
$$\hat{\mathbf{y}} | \mathbf{y}, \nabla \mathbf{y} \sim N(\boldsymbol{\mu}_\nabla, \boldsymbol{\Sigma}_\nabla), \quad (15.25)$$

где:

$$\begin{aligned} \boldsymbol{\mu}_{\nabla} = & \mathbf{m}_f(X^*) + \begin{bmatrix} \mathbf{K}_{ff}(X, X^*) \\ \mathbf{K}_{\nabla f}(X, X^*) \end{bmatrix}^T \times \\ & \times \begin{bmatrix} \mathbf{K}_{ff}(X, X) & \mathbf{K}_{f\nabla}(X, X) \\ \mathbf{K}_{\nabla f}(X, X) & \mathbf{K}_{\nabla\nabla}(X, X) \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{y} - \mathbf{m}_f(X) \\ \nabla\mathbf{y} - \mathbf{m}_{\nabla}(X) \end{bmatrix}, \end{aligned} \tag{15.26}$$

$$\begin{aligned} \boldsymbol{\Sigma}_{\nabla} = & \mathbf{K}_{ff}(X^*, X^*) - \begin{bmatrix} \mathbf{K}_{ff}(X, X^*) \\ \mathbf{K}_{\nabla f}(X, X^*) \end{bmatrix}^T \times \\ & \times \begin{bmatrix} \mathbf{K}_{ff}(X, X) & \mathbf{K}_{f\nabla}(X, X) \\ \mathbf{K}_{\nabla f}(X, X) & \mathbf{K}_{\nabla\nabla}(X, X) \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{K}_{ff}(X, X^*) \\ \mathbf{K}_{\nabla f}(X, X^*) \end{bmatrix}. \end{aligned} \tag{15.27}$$

На рис. 15.6 области, полученные при включении градиентных наблюдений, сравниваются с областями без градиентных наблюдений.



**Рис. 15.6.** Гауссовские процессы с информацией о градиенте и без нее с использованием квадратов экспоненциальных ядер. Включение информации о градиенте может значительно уменьшить доверительные интервалы

## 15.5. Информация о шуме

До сих пор мы предполагали, что целевая функция  $f$  является детерминированной. На практике, однако, оценки  $f$  могут включать в себя шум измерения, экспериментальную ошибку или числовое округление.

Мы можем смоделировать оценки шума как  $y = f(\mathbf{x}) + z$ , где  $f$  является детерминированной функцией, но  $z$  является гауссовским шумом с нулевым математи-

ческим ожиданием, т.е.  $z \sim \mathcal{N}(0, \nu)$ . Дисперсию шума можно настроить для контроля неопределенности.<sup>8</sup>

$$\begin{bmatrix} \hat{\mathbf{y}} \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \mathbf{m}(X^*) \\ \mathbf{m}(X^*) \end{bmatrix}, \begin{bmatrix} \mathbf{K}(X^*, X^*) & \mathbf{K}(X^*, X) \\ \mathbf{K}(X, X^*) & \mathbf{K}(X, X) + \nu \mathbf{I} \end{bmatrix} \right) \quad (15.28)$$

с условным распределением:

$$\hat{\mathbf{y}} | \mathbf{y}, \nu \sim \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*), \quad (15.29)$$

$$\boldsymbol{\mu}^* = \mathbf{m}(X^*) + \mathbf{K}(X^*, X)(\mathbf{K}(X, X) + \nu \mathbf{I})^{-1}(\mathbf{y} - \mathbf{m}(X)), \quad (15.30)$$

$$\boldsymbol{\Sigma}^* = \mathbf{K}(X^*, X^*) - \mathbf{K}(X^*, X)(\mathbf{K}(X, X) + \nu \mathbf{I})^{-1} \mathbf{K}(X, X^*). \quad (15.31)$$

Как показывают приведенные выше уравнения, учет гауссовского шума является простым, а последующее распределение может быть вычислено аналитически. Зашумленный гауссовский процесс показан на рис. 15.7. Прогнозирование для гауссовских процессов с зашумленными измерениями реализует алгоритм 15.4.

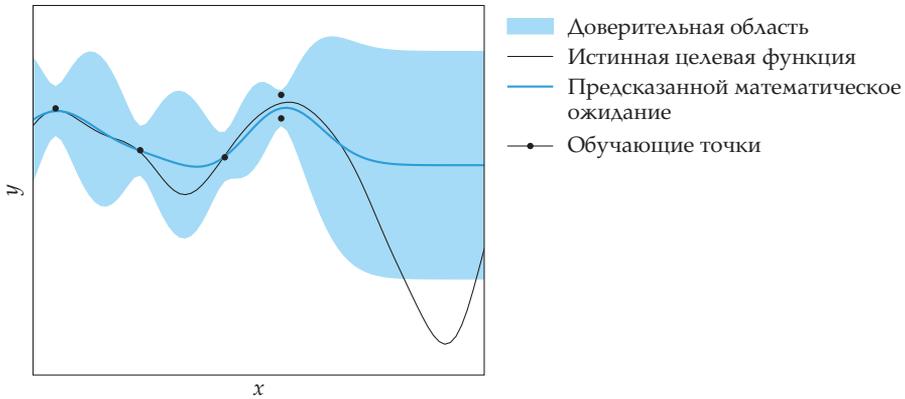


Рис. 15.7. Зашумленный гауссовский процесс с использованием квадрата экспоненциального ядра

---

**Алгоритм 15.4.** Метод получения предсказанных средних и стандартных отклонений по  $f$  при гауссовском процессе. Метод принимает гауссовский процесс GP и список точек  $\mathbf{X\_pred}$ , в которых можно оценить прогноз. Он возвращает среднее значение и дисперсию в каждой точке оценки

---

```
function predict(GP, X_pred)
    m, k, v = GP.m, GP.k, GP.v
    tmp = K(X_pred, GP.X, k) / (K(GP.X, GP.X, k) + v * I)
```

<sup>8</sup> Для настройки дисперсии шума можно использовать методы, описанные в разделе 14.5.

```

mp = mu(X_pred, m) + tmp * (GP.y - mu(GP.X, m))
S = K(X_pred, X_pred, k) - tmp * K(GP.X, X_pred, k)
vp = diag(S) .+ eps() # значение eps предотвращает вычислительные
                       # проблемы

return (mp, vp)
end

```

---

## 15.6. Подгонка гауссовских процессов

Выбор ядра и параметров оказывает большое влияние на форму гауссовского процесса между расчетными точками. Ядра и их параметры могут быть выбраны с помощью перекрестной проверки, представленной в предыдущей главе. Вместо того чтобы минимизировать квадрат ошибки на тестовых данных, мы максимизируем вероятность данных.<sup>9</sup> Иначе говоря, мы ищем параметры, которые максимизируют вероятность значений функции,  $p(\mathbf{y}|X, \theta)$ . Правдоподобность данных — это вероятность того, что наблюдаемые точки были взяты из модели. Эквивалентно мы можем максимизировать логарифмическую вероятность, которая обычно предпочтительнее, потому что умножение малых вероятностей при вычислении вероятности может привести к чрезвычайно малым значениям. При заданном наборе данных  $\mathcal{D}$  с  $n$  элементами логарифмическое правдоподобие определяется как

$$\begin{aligned} \log p(\mathbf{y}|X, \nu, \theta) = & -\frac{n}{2} \log 2\pi - \frac{1}{2} \log |K_\theta(X, X) + \nu \mathbf{I}| - \\ & - \frac{1}{2} (\mathbf{y} - \mathbf{m}_\theta(X))^T (K_\theta(X, X) + \nu \mathbf{I})^{-1} (\mathbf{y} - \mathbf{m}_\theta(X)), \end{aligned} \quad (15.32)$$

где математическое ожидание и ковариационные функции параметризованы параметром  $\theta$ .

Предположим, что математическое значение равно нулю, т.е.  $\mathbf{m}_\theta(X) = \mathbf{0}$  и параметр  $\theta$  относится только к параметрам для ковариационной функции гауссовского процесса. Мы можем прийти к оценке *максимального правдоподобия* путем градиентного подъема. Тогда градиент определяется как

$$\frac{\partial}{\partial \theta_j} \log p(\mathbf{y}|X, \theta) = \frac{1}{2} \mathbf{y}^T \mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \theta_j} \mathbf{K}^{-1} \mathbf{y} - \frac{1}{2} \text{tr} \left( \Sigma_\theta^{-1} \frac{\partial \mathbf{K}}{\partial \theta_j} \right), \quad (15.33)$$

где  $\Sigma_\theta = \mathbf{K}_\theta(X, X) + \nu \mathbf{I}$ . Выше мы использовали матричные производные отношения

<sup>9</sup> В качестве альтернативы можно максимизировать псевдослучайность, как показано в [126].

$$\frac{\partial \mathbf{K}^{-1}}{\partial \boldsymbol{\theta}_j} = -\mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \boldsymbol{\theta}_j} \mathbf{K}^{-1}, \quad (15.34)$$

$$\frac{\partial \log |\mathbf{K}|}{\partial \boldsymbol{\theta}_j} = \text{tr} \left( \mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \boldsymbol{\theta}_j} \right), \quad (15.35)$$

где  $\text{tr}(\mathbf{A})$  обозначает *след* матрицы  $\mathbf{A}$ , определенной как сумма элементов на главной диагонали.

## 15.7. Резюме

- Гауссовские процессы — это распределения вероятностей по функциям.
- Выбор ядра влияет на гладкость функций, выбранных из гауссовского процесса.
- Многомерное нормальное распределение имеет аналитические условные и маргинальные распределения.
- Можно рассчитать математическое и стандартное отклонение прогноза целевой функции в конкретной расчетной точке с учетом ряда прошлых оценок.
- Можно включить информацию о градиенте, чтобы улучшить прогнозы значений целевой функции и ее градиента.
- Можно включить измерительный шум в гауссовский процесс.
- Можно подобрать параметры гауссовского процесса, используя максимальное правдоподобие.

## 15.8. Упражнения

**Упражнение 15.1.** Гауссовские процессы будут усложняться в процессе оптимизации по мере накопления большего количества выборок. Как это может стать преимуществом перед регрессионными моделями?

**Упражнение 15.2.** Как вычислительная сложность предсказания с гауссовским процессом увеличивается с числом точек данных  $m$ ?

**Упражнение 15.3.** Рассмотрим функцию  $f(x) = \sin x / (x^2 + 1)$  на отрезке  $[-5, 5]$ . Постройте 95%-ные доверительные границы для гауссовского процесса с информацией о производной, соответствующей оценкам в точках  $\{-5; -2,5; 0; 2,5; 5\}$ . Каково максимальное стандартное отклонение прогнозируемого распределения в диапазоне  $[-5, 5]$ ? Сколько оценок функций, равномерно распределенных по

области, необходимо для того, чтобы гауссовский процесс без информации о производной достигал одинакового максимального прогнозирующего стандартного отклонения?

Предположим, что функции имеют нулевое математическое ожидание и наблюдения без шума, а также функции ковариации:

$$k_{ff}(x, x') = \exp\left(-\frac{1}{2}\|x - x'\|_2^2\right),$$

$$k_{\nabla f}(x, x') = (x' - x) \exp\left(-\frac{1}{2}\|x - x'\|_2^2\right),$$

$$k_{\nabla\nabla}(x, x') = \left((x - x')^2 - 1\right) \exp\left(-\frac{1}{2}\|x - x'\|_2^2\right).$$

**Упражнение 15.4.** Выведите отношение

$$k_{f\nabla}(\mathbf{x}, \mathbf{x}')_i = \text{cov}\left(f(\mathbf{x}), \frac{\partial}{\partial x'_i} f(\mathbf{x}')\right) = \frac{\partial}{\partial x'_i} k_{ff}(\mathbf{x}, \mathbf{x}').$$

**Упражнение 15.5.** Предположим, что мы имеем многомерное нормальное распределение по двум переменным  $a$  и  $b$ . Покажите, что дисперсия условного распределения  $a$  при условии  $b$  не больше дисперсии маргинального распределения по  $a$ . Имеет ли это интуитивный смысл?

**Упражнение 15.6.** Допустим, что мы наблюдаем много выбросов, т.е. наблюдаем выборки, которые не попадают в доверительный интервал, заданный гауссовским процессом. Это означает, что выбранная нами вероятностная модель не подходит. Что можно сделать?

**Упражнение 15.7.** Рассмотрим выбор модели для пар оценки функций  $(x, y)$ :

$$\{(1, 0), (2, -1), (3, -2), (4, 1), (5, 0)\}$$

Используйте скользящую перекрестную проверку, чтобы выбрать ядро, которое максимизирует вероятность предсказания отложенной пары с учетом гауссовского процесса, определенного на других парах в группе. Предположим, что функция имеет нулевое математическое ожидание и не имеет шума. Выберите одно из ядер:

$$\exp(-\|x - x'\|) \quad \exp(-\|x - x'\|^2) \quad (1 + \|x - x'\|)^{-1} \quad (1 + \|x - x'\|^2)^{-1} \quad (1 + \|x - x'\|)^{-2}.$$