# Байесовское машинное обучение

Евгений Бурнаев

Skoltech, Москва, Россия

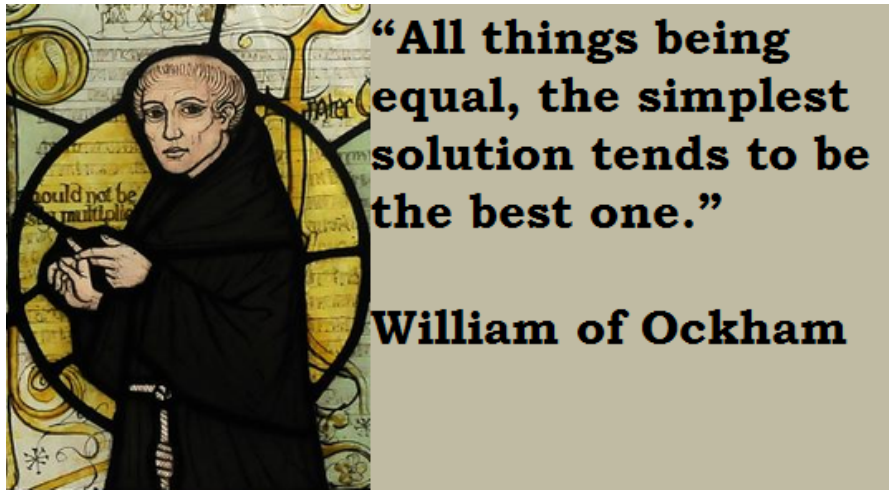
Skoltech
Skolkovo Institute of Science and Technology

# Содержание

# Основные принципы

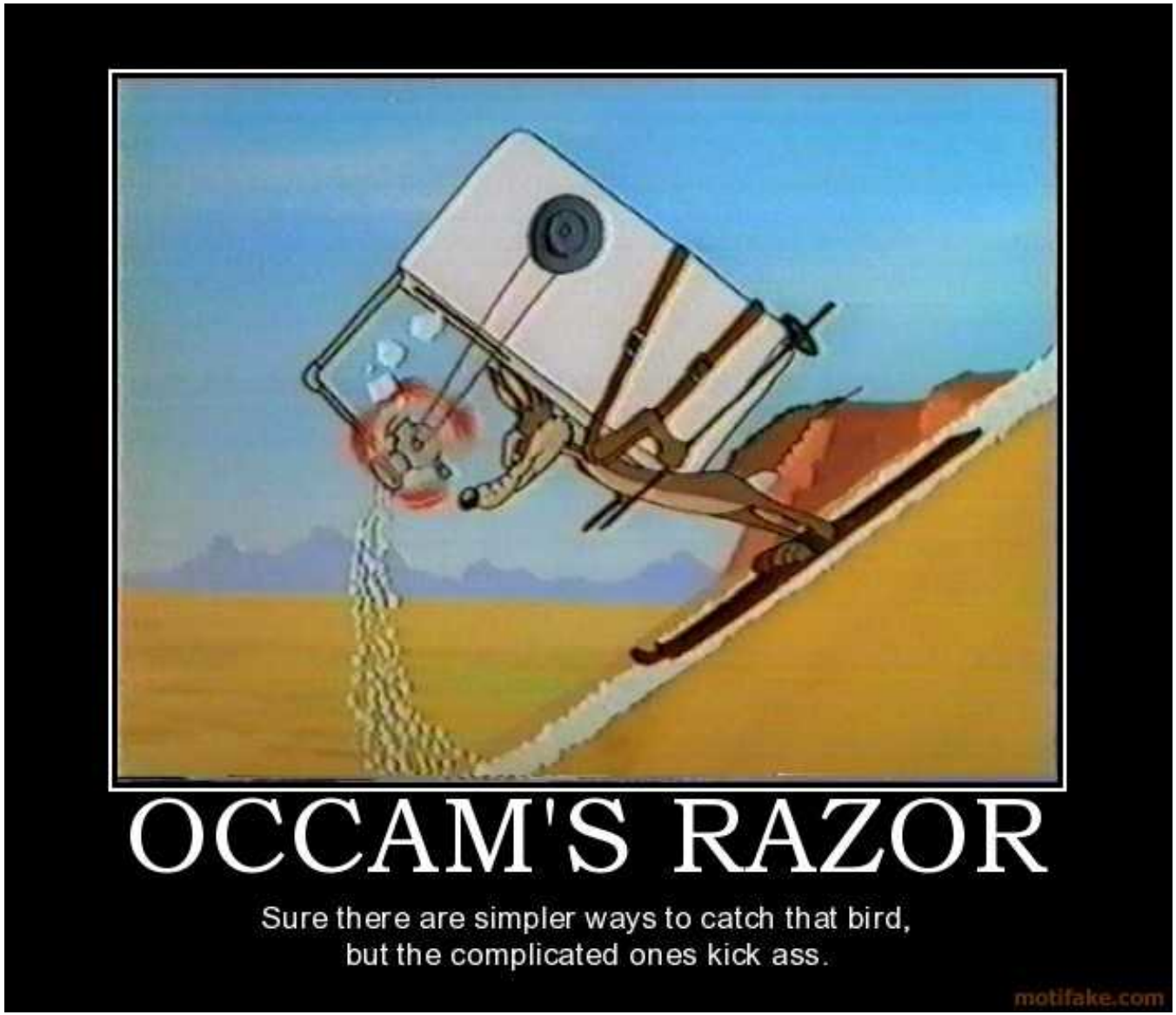
Томас Байес (с. 1701 – 7 Апреля 1761) был английским статистиком, философом и Пресвитерианским министром

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}$$

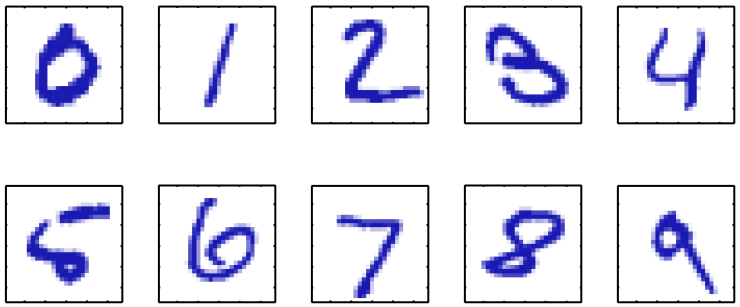



Вильям Окхам (с. 1287 – 1347) был английским был английским францисканским монахом и схоластическим философом и теологом

# Occam's razor

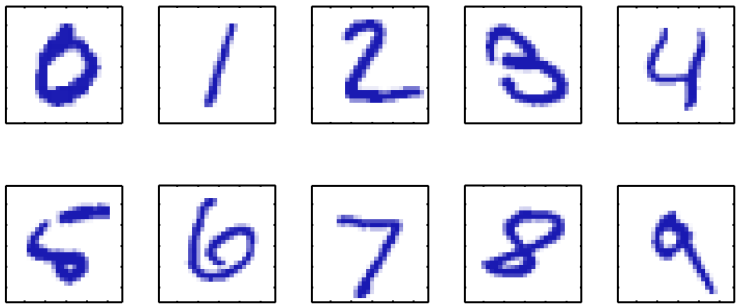


# Пример: рукописные цифры



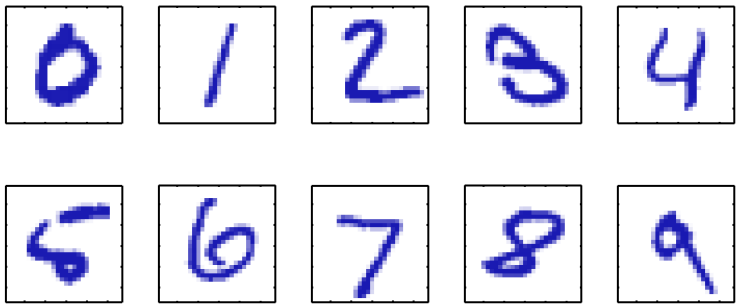
- Пример: распознавание рукописных цифр
- Каждой цифре соответствует $28 \times 28$ пиксельное изображение, представленное в виде $784$-размерного вектора $\mathbf{x}$
- $\mathcal{D}_m = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ тренировочные примеры, где $y_i$ — метка класса, соответствующая $\mathbf{x}_i$
- Используя $\mathcal{D}_m$ мы хотим восстановить $y = f(\mathbf{x})$

# Пример: рукописные цифры



- Пример: распознавание рукописных цифр
- Каждой цифре соответствует $28 \times 28$ пиксельное изображение, представленное в виде $784$-размерного вектора $\mathbf{x}$
- $\mathcal{D}_m = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ тренировочные примеры, где $y_i$ — метка класса, соответствующая $\mathbf{x}_i$
- Используя $\mathcal{D}_m$ мы хотим восстановить $y = f(\mathbf{x})$

# Пример: рукописные цифры



- Пример: распознавание рукописных цифр
- Каждой цифре соответствует $28 \times 28$ пиксельное изображение, представленное в виде $784$-размерного вектора $\mathbf{x}$
- $\mathcal{D}_m = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ тренировочные примеры, где $y_i$ — метка класса, соответствующая $\mathbf{x}_i$
- Используя $\mathcal{D}_m$ мы хотим восстановить $y = f(\mathbf{x})$

# Пример: полиномиальная регрессия

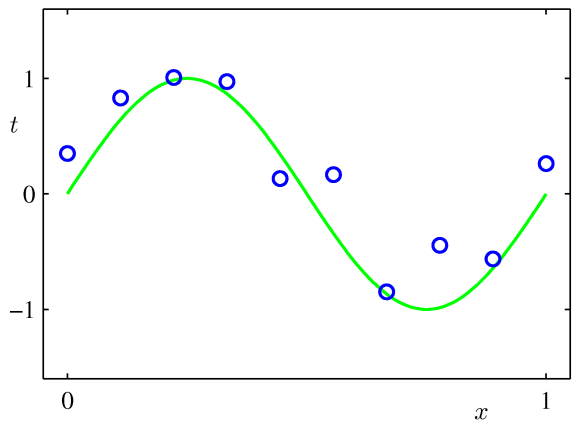



Рис. – График тренировочных данных

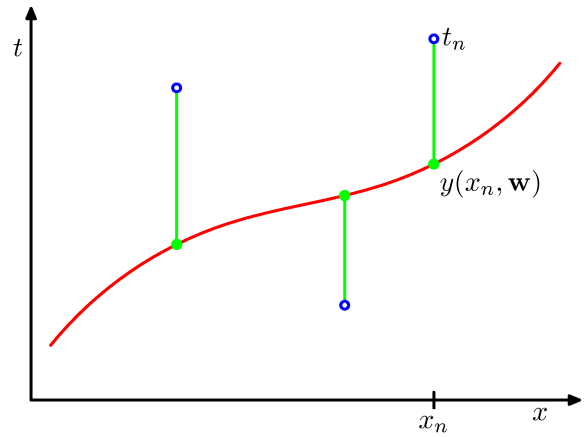



Рис. – Остатки (невязка)

- $\mathcal{D}_m = \{\mathbf{X}_m, \mathbf{Y}_m\} = \{(x_i, y_i)\}_{i=1}^m$, где $y_i = \sin(2\pi x_i) + \varepsilon_i$, $\varepsilon_i$ — гауссовский белый шум

# Пример: полиномиальная регрессия

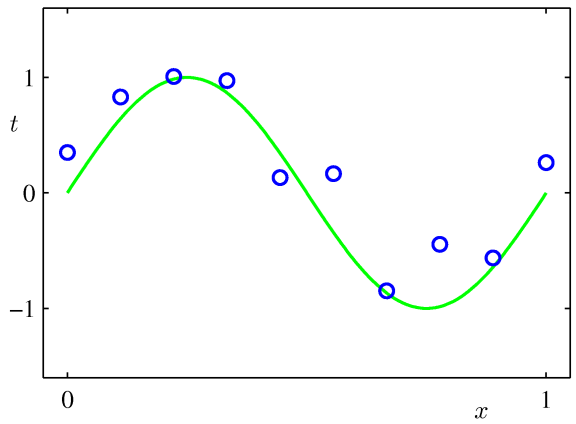



Рис. – График тренировочных данных

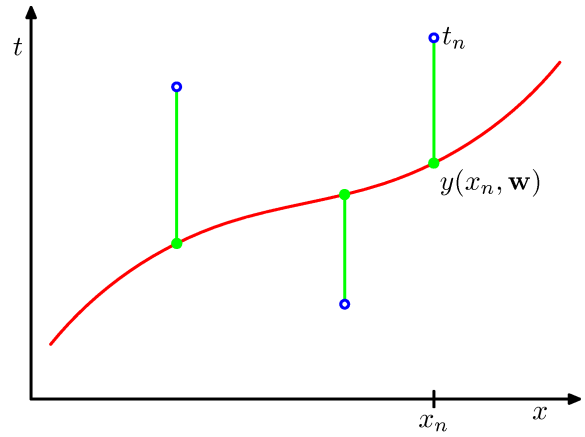



Рис. – Остатки (невязка)

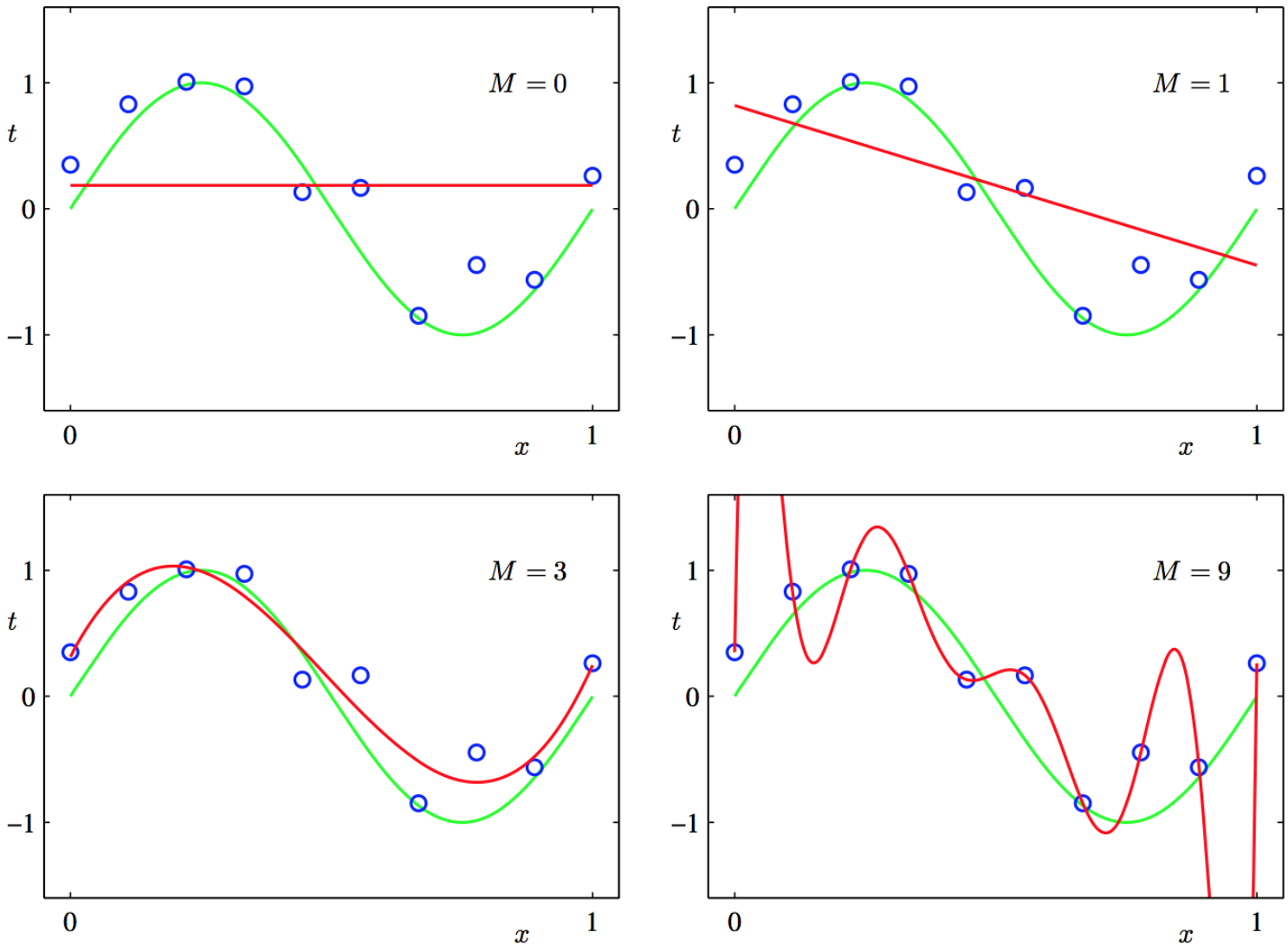
- Мы обучаем модель

$$y(x, \mathbf{w}) = \sum_{j=0}^{M} w_j x^j,$$

минимизируя ошибку

$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{m} \{y(x_i, \mathbf{w}) - y_i\}^2$$

# Графики полиномов различного порядка $M$



# Переобучение

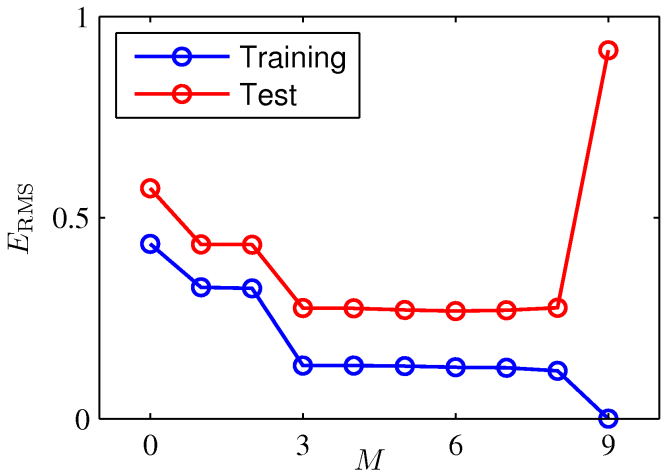



Рис. – $E_{RMS} = \sqrt{2E(\mathbf{w}^*)/n}$

| | $M=0$ | $M=1$ | $M=6$ | $M=9$ |
|---|---|---|---|---|
| $w_0^\star$ | 0.19 | 0.82 | 0.31 | 0.35 |
| $w_1^\star$ | | -1.27 | 7.99 | 232.37 |
| $w_2^\star$ | | | -25.43 | -5321.83 |
| $w_3^\star$ | | | 17.37 | 48568.31 |
| $w_4^\star$ | | | | -231639.30 |
| $w_5^\star$ | | | | 640042.26 |
| $w_6^\star$ | | | | -1061800.52 |
| $w_7^\star$ | | | | 1042400.18 |
| $w_8^\star$ | | | | -557682.99 |
| $w_9^\star$ | | | | 125201.43 |

Рис. – Коэффициенты $w^*$

# Переобучение vs. количество примеров

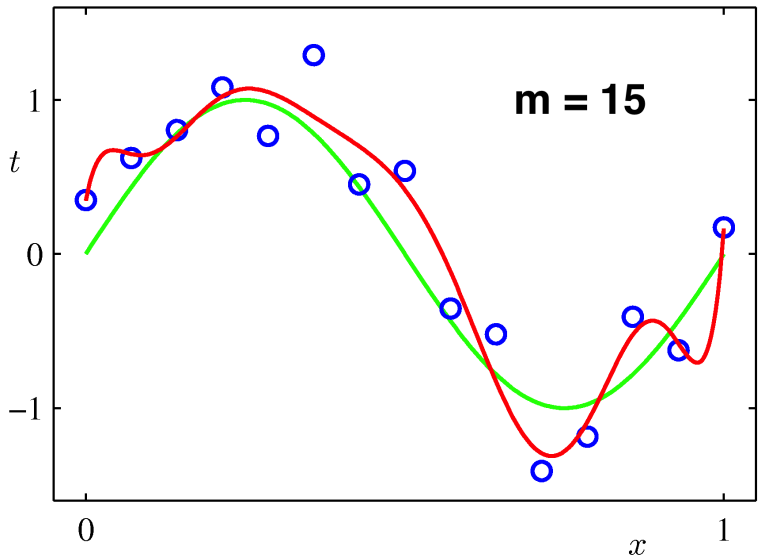



Рис. – $M = 9, m = 15$

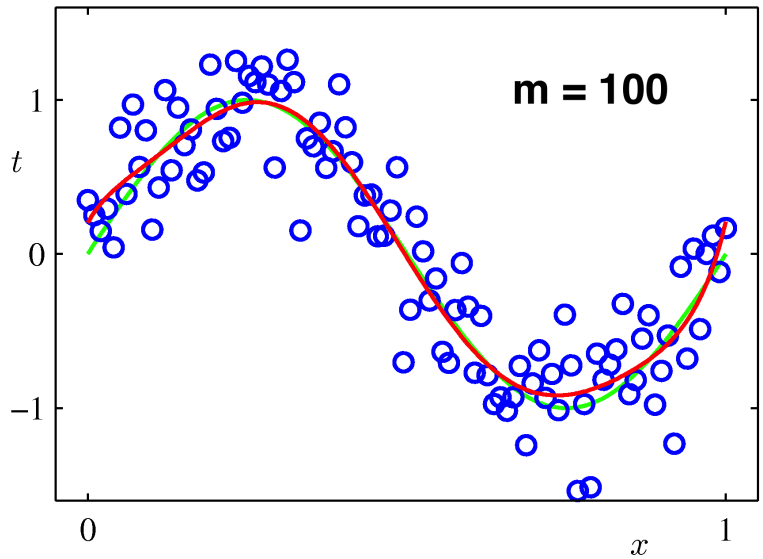



Рис. – $M = 9, m = 100$

# Переобучение vs. регуляризация

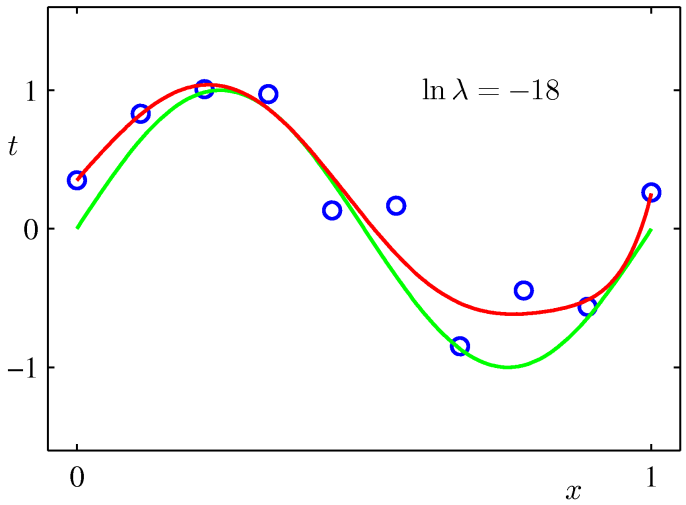



Рис. – $\lambda = e^{-18} \approx 0$

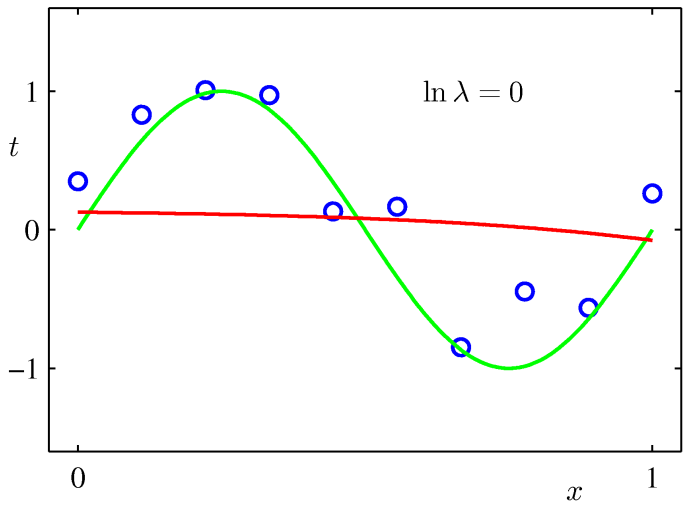



Рис. – $\lambda = 1$

- Ограничение количества параметров $M$ в соответствии с размером обучающей выборки?
- Вместо этого лучше выбрать сложность модели (количество гиперпараметров) в соответствии со сложностью проблемы!

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{m}\{y(x_i, \mathbf{w}) - y_i\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

# Переобучение vs. регуляризация

| | $\ln \lambda = -\infty$ | $\ln \lambda = -18$ | $\ln \lambda = 0$ |
|---|---|---|---|
| $w_0^\star$ | 0.35 | 0.35 | 0.13 |
| $w_1^\star$ | 232.37 | 4.74 | -0.05 |
| $w_2^\star$ | -5321.83 | -0.77 | -0.06 |
| $w_3^\star$ | 48568.31 | -31.97 | -0.05 |
| $w_4^\star$ | -231639.30 | -3.89 | -0.03 |
| $w_5^\star$ | 640042.26 | 55.28 | -0.02 |
| $w_6^\star$ | -1061800.52 | 41.32 | -0.01 |
| $w_7^\star$ | 1042400.18 | -45.95 | -0.00 |
| $w_8^\star$ | -557682.99 | -91.53 | 0.00 |
| $w_9^\star$ | 125201.43 | 72.68 | 0.01 |

Рис. – Зависимость $w^*$ от $\lambda$

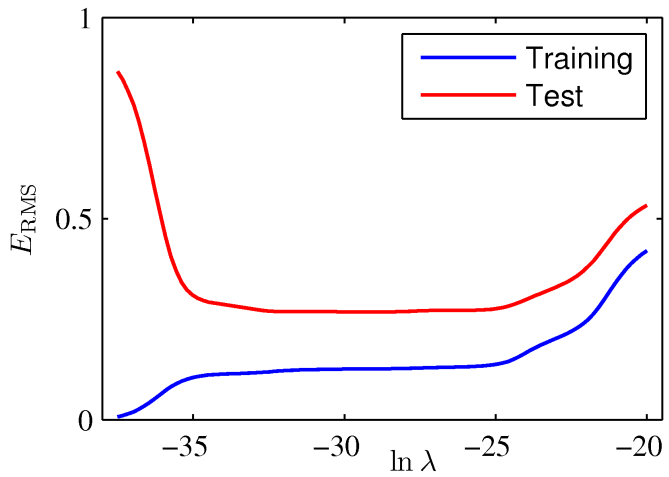



Рис. – Зависимость $E_{RMS}$ от $\lambda$

- Мы должны найти способ определить подходящее значение сложности модели!
- Отложенное множество для выбора сложности модели (либо $M$, либо $\lambda$)? Слишком ресурсозатратно $\Rightarrow$ Байесовские методы!

# Переобучение vs. регуляризация

| | $\ln \lambda = -\infty$ | $\ln \lambda = -18$ | $\ln \lambda = 0$ |
|---|---|---|---|
| $w_0^\star$ | 0.35 | 0.35 | 0.13 |
| $w_1^\star$ | 232.37 | 4.74 | -0.05 |
| $w_2^\star$ | -5321.83 | -0.77 | -0.06 |
| $w_3^\star$ | 48568.31 | -31.97 | -0.05 |
| $w_4^\star$ | -231639.30 | -3.89 | -0.03 |
| $w_5^\star$ | 640042.26 | 55.28 | -0.02 |
| $w_6^\star$ | -1061800.52 | 41.32 | -0.01 |
| $w_7^\star$ | 1042400.18 | -45.95 | -0.00 |
| $w_8^\star$ | -557682.99 | -91.53 | 0.00 |
| $w_9^\star$ | 125201.43 | 72.68 | 0.01 |

Рис. – Зависимость $w^*$ от $\lambda$

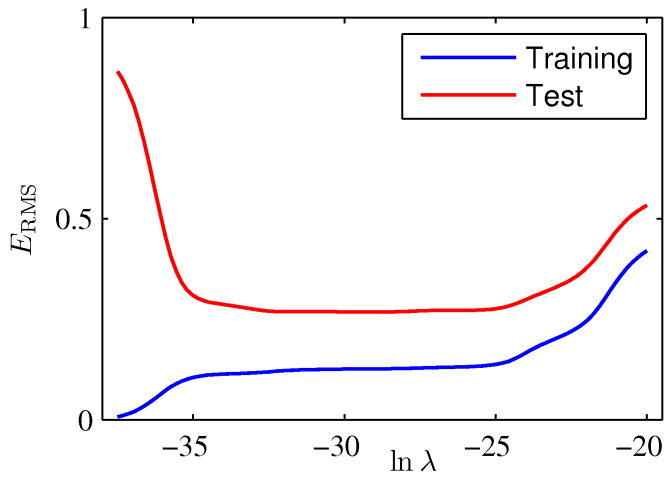



Рис. – Зависимость $E_{RMS}$ от $\lambda$

- Мы должны найти способ определить подходящее значение сложности модели!
- Отложенное множество для выбора сложности модели (либо $M$, либо $\lambda$)? Слишком ресурсозатратно $\Rightarrow$ Байесовские методы!

# Правила вероятности

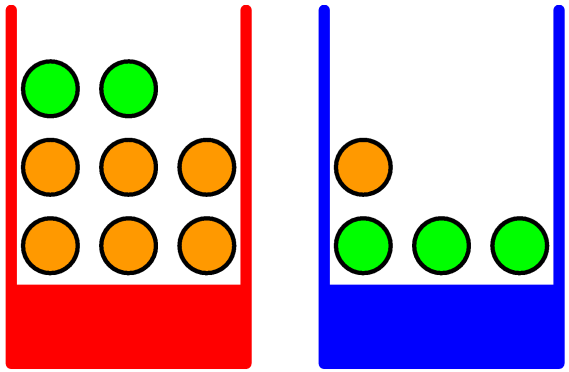


Рис. – Две коробки с фруктами (яблоки, апельсины)

- Мы случайным образов выбираем красную коробку $40\%$ процентов времени, а синюю — $60\%$
- Из этой коробки мы случайно выбираем фрукт с одинаковой вероятностью
- Определив, что это за фрукт, мы кладём его назад в ту же коробку $\Rightarrow$ повторяем эксперимент!
- $B \in \{r, b\}$ — случайно выбранная коробка, $F \in \{a, o\}$ случайно выбранный фрукт

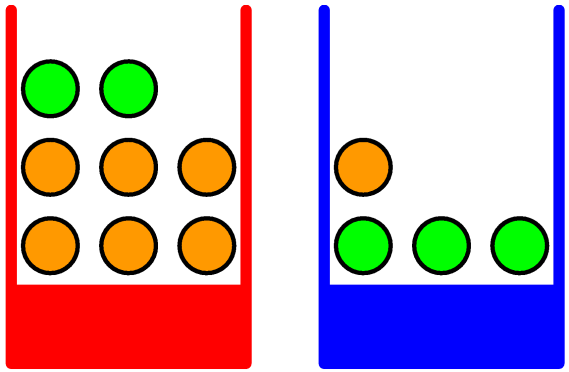


Рис. – Две коробки с фруктами (яблоки, апельсины)

- Мы случайным образов выбираем красную коробку $40\%$ процентов времени, а синюю — $60\%$
- Из этой коробки мы случайно выбираем фрукт с одинаковой вероятностью
- Определив, что это за фрукт, мы кладём его назад в ту же коробку $\Rightarrow$ повторяем эксперимент!
- $B \in \{r, b\}$ — случайно выбранная коробка, $F \in \{a, o\}$ случайно выбранный фрукт

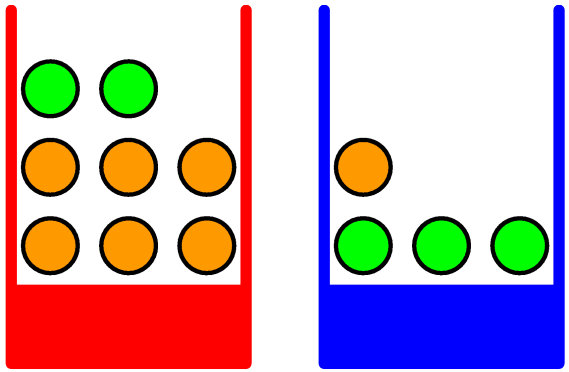


Рис. – Две коробки с фруктами (яблоки, апельсины)

- Мы случайным образов выбираем красную коробку $40\%$ процентов времени, а синюю — $60\%$
- Из этой коробки мы случайно выбираем фрукт с одинаковой вероятностью
- Определив, что это за фрукт, мы кладём его назад в ту же коробку $\Rightarrow$ повторяем эксперимент!
- $B \in \{r, b\}$ — случайно выбранная коробка, $F \in \{a, o\}$ случайно выбранный фрукт

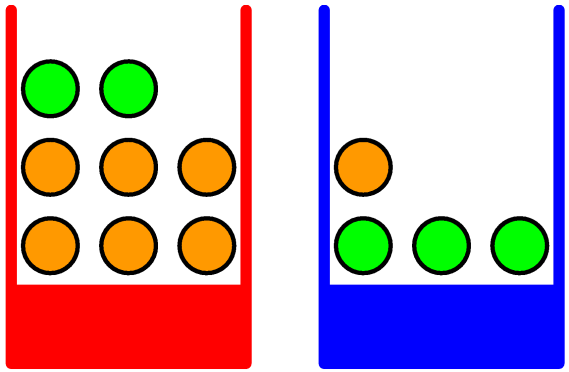


Рис. – Две коробки с фруктами (яблоки, апельсины)

- Мы случайным образов выбираем красную коробку $40\%$ процентов времени, а синюю — $60\%$
- Из этой коробки мы случайно выбираем фрукт с одинаковой вероятностью
- Определив, что это за фрукт, мы кладём его назад в ту же коробку $\Rightarrow$ повторяем эксперимент!
- $B \in \{r, b\}$ — случайно выбранная коробка, $F \in \{a, o\}$ случайно выбранный фрукт

# Правила вероятности

- $\mathbb{P}(B = r) = \frac{4}{10}$, $\mathbb{P}(B = b) = \frac{6}{10}$
- Типичные вопросы:
  - “Какая вероятность, что мы выберем яблоко?”,
  - “Какова вероятность того, что мы достали фрукт из синей коробки при условии, что это апельсин?”
- Более общий пример: 2 случайных величины $(X, Y)$, $X \in \{x_1, \ldots, x_M\}$, $Y \in \{y_1, \ldots, y_L\}$

# Правила вероятности

- $\mathbb{P}(B = r) = \frac{4}{10}$, $\mathbb{P}(B = b) = \frac{6}{10}$
- Типичные вопросы:
  - “Какая вероятность, что мы выберем яблоко?”,
  - “Какова вероятность того, что мы достали фрукт из синей коробки при условии, что это апельсин?”
- Более общий пример: 2 случайных величины $(X, Y)$, $X \in \{x_1, \ldots, x_M\}$, $Y \in \{y_1, \ldots, y_L\}$

# Правила вероятности

- $\mathbb{P}(B = r) = \frac{4}{10}$, $\mathbb{P}(B = b) = \frac{6}{10}$
- Типичные вопросы:
  - “Какая вероятность, что мы выберем яблоко?”,
  - “Какова вероятность того, что мы достали фрукт из синей коробки при условии, что это апельсин?”
- Более общий пример: 2 случайных величины $(X, Y)$, $X \in \{x_1, \ldots, x_M\}$, $Y \in \{y_1, \ldots, y_L\}$

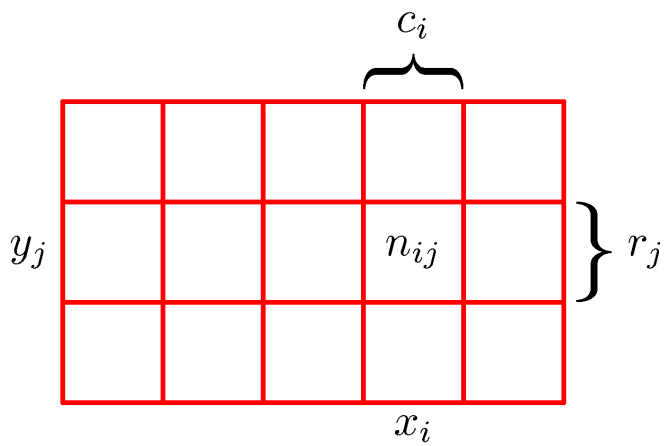



Рис. – Более общий пример

Стандартное определение:

$$\mathbb{P}(X = x_i, Y = y_j) = \frac{n_{ij}}{n},$$

где $P(X = x_i) = \frac{c_i}{n}$, $c_i = \sum_j n_{ij}$, тогда

$$\mathbb{P}(X = x_i) = \sum_{j=1}^{L} \mathbb{P}(X = x_i, Y = y_j).$$

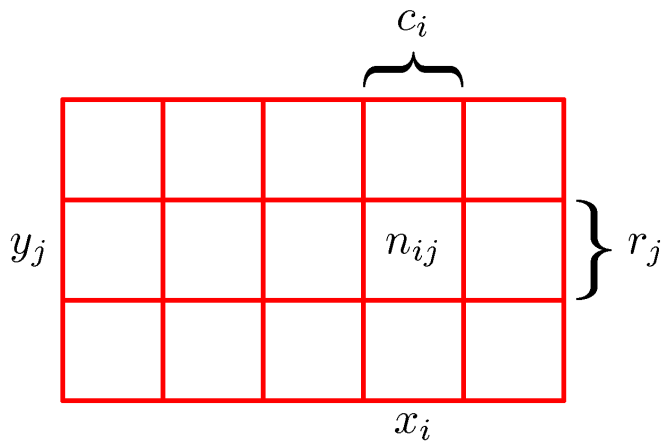



Рис. – Более общий пример

Стандартные определения: так как

$$\mathbb{P}(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i},$$

тогда

$$\begin{aligned}\mathbb{P}(X = x_i, Y = y_j) = \frac{n_{ij}}{n} &= \frac{n_{ij}}{c_i} \cdot \frac{c_i}{n}, \\ &= \mathbb{P}(Y = y_j | X = x_i)\mathbb{P}(X = x_i)\end{aligned}$$

# Правила вероятности

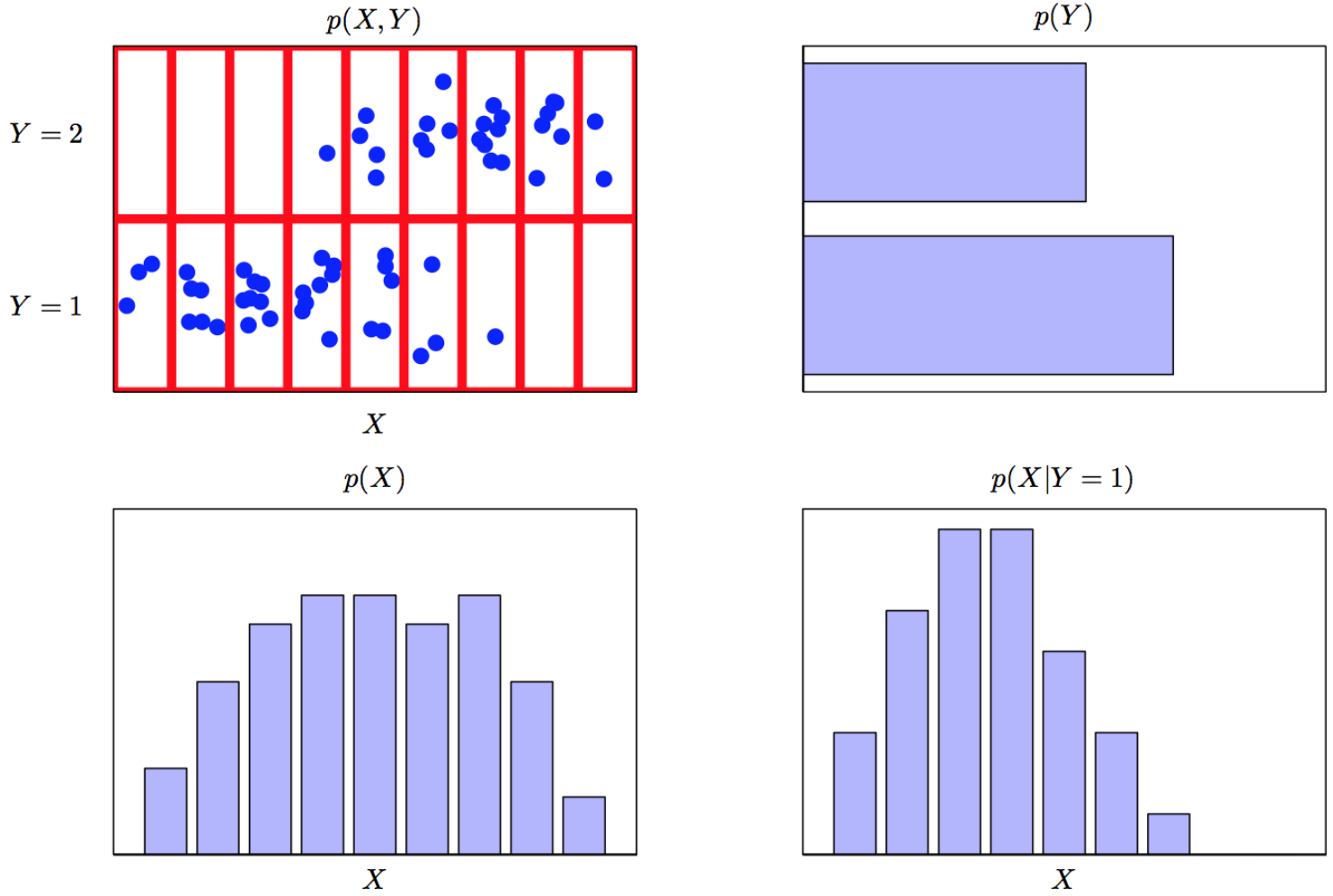



Рис. – Пример совместного распределения

- Вычисления вероятностей

$$\mathbb{P}(X) = \sum_Y \mathbb{P}(X, Y)$$

$$\mathbb{P}(X, Y) = \mathbb{P}(Y|X)\mathbb{P}(X)$$

$$\mathbb{P}(X, Y) = \mathbb{P}(Y, X)$$

$$\mathbb{P}(Y|X) = \frac{\mathbb{P}(X|Y)\mathbb{P}(Y)}{\mathbb{P}(X)}$$

$$\mathbb{P}(X) = \sum_Y \mathbb{P}(X|Y)\mathbb{P}(Y)$$

- Пример

$$\mathbb{P}(F = a) = \mathbb{P}(F = a|B = r)\mathbb{P}(B = r) + \mathbb{P}(F = a|B = b)\mathbb{P}(B = b)$$

$$= \frac{1}{4}\frac{4}{10} + \frac{3}{4}\frac{6}{10} = \frac{11}{20}$$

$$\mathbb{P}(B = r|F = o) = \frac{\mathbb{P}(F = o|B = r)\mathbb{P}(B = r)}{\mathbb{P}(F = o)} = \frac{3}{4}\frac{4}{10}\frac{20}{9} = \frac{2}{3}$$

# Правила вероятности

- Вычисления вероятностей

$$\mathbb{P}(X) = \sum_Y \mathbb{P}(X, Y)$$
$$\mathbb{P}(X, Y) = \mathbb{P}(Y|X)\mathbb{P}(X)$$
$$\mathbb{P}(X, Y) = \mathbb{P}(Y, X)$$
$$\mathbb{P}(Y|X) = \frac{\mathbb{P}(X|Y)\mathbb{P}(Y)}{\mathbb{P}(X)}$$
$$\mathbb{P}(X) = \sum_Y \mathbb{P}(X|Y)\mathbb{P}(Y)$$

- Пример

$$\mathbb{P}(F = a) = \mathbb{P}(F = a|B = r)\mathbb{P}(B = r) + \mathbb{P}(F = a|B = b)\mathbb{P}(B = b)$$
$$= \frac{1}{4}\frac{4}{10} + \frac{3}{4}\frac{6}{10} = \frac{11}{20}$$
$$\mathbb{P}(B = r|F = o) = \frac{\mathbb{P}(F = o|B = r)\mathbb{P}(B = r)}{\mathbb{P}(F = o)} = \frac{3}{4}\frac{4}{10}\frac{20}{9} = \frac{2}{3}$$

# Плотность вероятности

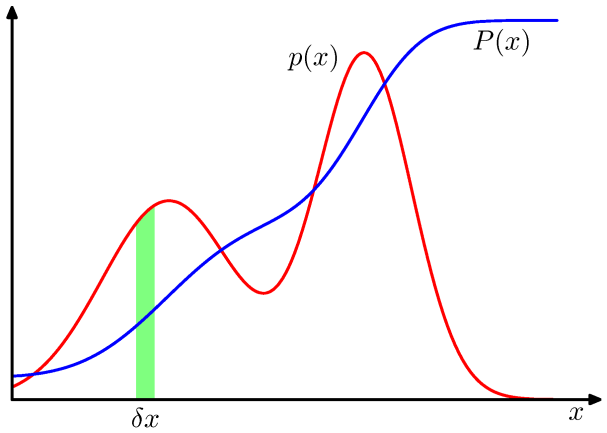



Рис. – Плотность вероятности: $\mathbb{P}(X \in (x, x + \delta x)) = p(x)\delta x$

- плотность

$$p(\mathbf{x}) \geq 0, \int_{\mathbb{R}} p(\mathbf{x}) d\mathbf{x} = 1$$

- функция распределения

$$F(\mathbf{z}) = \mathbb{P}(X_1 \leq z_1, \ldots, X_M \leq z_M)$$
$$= \int_{-\infty}^{z_1} \cdots \int_{-\infty}^{z_M} p(\mathbf{x}) d\mathbf{x}$$

- маргинальное распределение

$$p(x) = \int p(x, y) dy, \; p(x, y) = p(y|x)p(x)$$

# Плотность вероятности

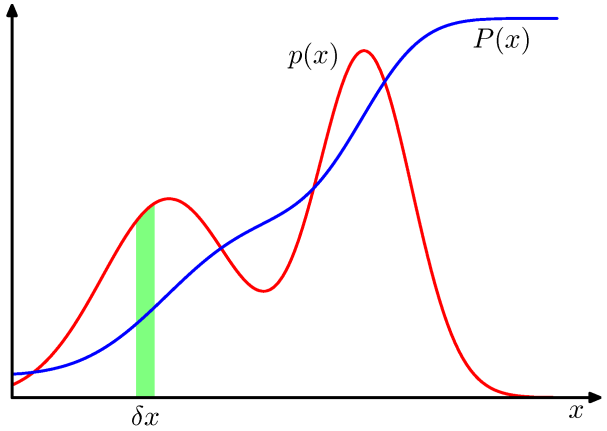



Рис. – Плотность вероятности: $\mathbb{P}(X \in (x, x + \delta x)) = p(x)\delta x$

- плотность

$$p(\mathbf{x}) \geq 0, \int_{\mathbb{R}} p(\mathbf{x})d\mathbf{x} = 1$$

- функция распределения

$$F(\mathbf{z}) = \mathbb{P}(X_1 \leq z_1, \ldots, X_M \leq z_M)$$
$$= \int_{-\infty}^{z_1} \cdots \int_{-\infty}^{z_M} p(\mathbf{x})d\mathbf{x}$$

- маргинальное распределение

$$p(x) = \int p(x, y)dy, \; p(x, y) = p(y|x)p(x)$$

# Плотность вероятности

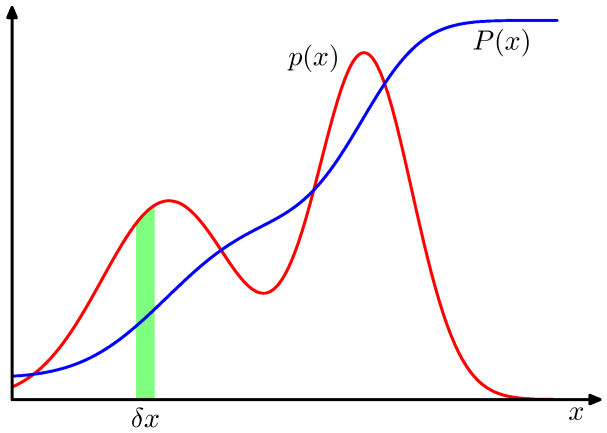



Рис. – Плотность вероятности: $\mathbb{P}(X \in (x, x + \delta x)) = p(x)\delta x$

- плотность

$$p(\mathbf{x}) \geq 0, \int_{\mathbb{R}} p(\mathbf{x}) d\mathbf{x} = 1$$

- функция распределения

$$F(\mathbf{z}) = \mathbb{P}(X_1 \leq z_1, \ldots, X_M \leq z_M)$$
$$= \int_{-\infty}^{z_1} \cdots \int_{-\infty}^{z_M} p(\mathbf{x}) d\mathbf{x}$$

- маргинальное распределение

$$p(x) = \int p(x, y) dy, \; p(x, y) = p(y|x)p(x)$$

# Математическое ожидание и ковариации

- Мат. ожидание

$$\mathbb{E}[f] = \sum_x p(x)f(x), \;\; \mathbb{E}[f] = \int_x p(x)f(x)dx$$

- Оценка мат. ожидания

$$\mathbb{E}[f] \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i), \;\; x_i \sim p(\cdot)$$

- Условное мат. ожидание

$$\mathbb{E}_x[f|y] = \sum_x f(x)p(x|y)$$

- Дисперсия

$$\mathrm{var}[f] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] = \mathbb{E}[(f(x))^2] - (\mathbb{E}[f(x)])^2$$

- Ковариация

$$\mathrm{cov}[x, y] = \mathbb{E}_{x,y}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] = \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]$$

$$\mathrm{cov}[\mathbf{x}, \mathbf{y}] = \mathbb{E}_{\mathbf{x},\mathbf{y}}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])^{\mathrm{T}}] = \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}]$$

# Математическое ожидание и ковариации

- Мат. ожидание

$$\mathbb{E}[f] = \sum_x p(x)f(x), \;\; \mathbb{E}[f] = \int_x p(x)f(x)dx$$

- Оценка мат. ожидания

$$\mathbb{E}[f] \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i), \;\; x_i \sim p(\cdot)$$

- Условное мат. ожидание

$$\mathbb{E}_x[f|y] = \sum_x f(x)p(x|y)$$

- Дисперсия

$$\mathrm{var}[f] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] = \mathbb{E}[(f(x))^2] - (\mathbb{E}[f(x)])^2$$

- Ковариация

$$\mathrm{cov}[x, y] = \mathbb{E}_{x,y}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] = \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]$$

$$\mathrm{cov}[\mathbf{x}, \mathbf{y}] = \mathbb{E}_{\mathbf{x},\mathbf{y}}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])^{\mathrm{T}}] = \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}]$$

# Математическое ожидание и ковариации

- Мат. ожидание

$$\mathbb{E}[f] = \sum_x p(x)f(x), \;\; \mathbb{E}[f] = \int_x p(x)f(x)dx$$

- Оценка мат. ожидания

$$\mathbb{E}[f] \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i), \;\; x_i \sim p(\cdot)$$

- Условное мат. ожидание

$$\mathbb{E}_x[f|y] = \sum_x f(x)p(x|y)$$

- Дисперсия

$$\mathrm{var}[f] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] = \mathbb{E}[(f(x))^2] - (\mathbb{E}[f(x)])^2$$

- Ковариация

$$\mathrm{cov}[x, y] = \mathbb{E}_{x,y}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] = \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]$$

$$\mathrm{cov}[\mathbf{x}, \mathbf{y}] = \mathbb{E}_{\mathbf{x},\mathbf{y}}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])^{\mathrm{T}}] = \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}]$$

# Математическое ожидание и ковариации

- Мат. ожидание

$$\mathbb{E}[f] = \sum_x p(x)f(x), \;\; \mathbb{E}[f] = \int_x p(x)f(x)dx$$

- Оценка мат. ожидания

$$\mathbb{E}[f] \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i), \;\; x_i \sim p(\cdot)$$

- Условное мат. ожидание

$$\mathbb{E}_x[f|y] = \sum_x f(x)p(x|y)$$

- Дисперсия

$$\mathrm{var}[f] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] = \mathbb{E}[(f(x))^2] - (\mathbb{E}[f(x)])^2$$

- Ковариация

$$\mathrm{cov}[x, y] = \mathbb{E}_{x,y}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] = \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]$$

$$\mathrm{cov}[\mathbf{x}, \mathbf{y}] = \mathbb{E}_{\mathbf{x},\mathbf{y}}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])^{\mathrm{T}}] = \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}]$$

# Математическое ожидание и ковариации

- Мат. ожидание

$$\mathbb{E}[f] = \sum_x p(x)f(x), \;\; \mathbb{E}[f] = \int_x p(x)f(x)dx$$

- Оценка мат. ожидания

$$\mathbb{E}[f] \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i), \;\; x_i \sim p(\cdot)$$

- Условное мат. ожидание

$$\mathbb{E}_x[f|y] = \sum_x f(x)p(x|y)$$

- Дисперсия

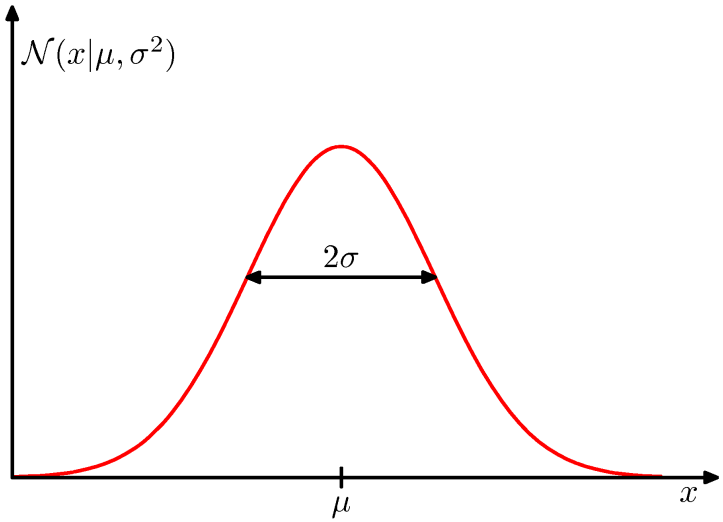
$$\mathrm{var}[f] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] = \mathbb{E}[(f(x))^2] - (\mathbb{E}[f(x)])^2$$

- Ковариация

$$\mathrm{cov}[x, y] = \mathbb{E}_{x,y}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] = \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]$$

$$\mathrm{cov}[\mathbf{x}, \mathbf{y}] = \mathbb{E}_{\mathbf{x},\mathbf{y}}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])^{\mathrm{T}}] = \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}]$$

# Гауссовское распределение



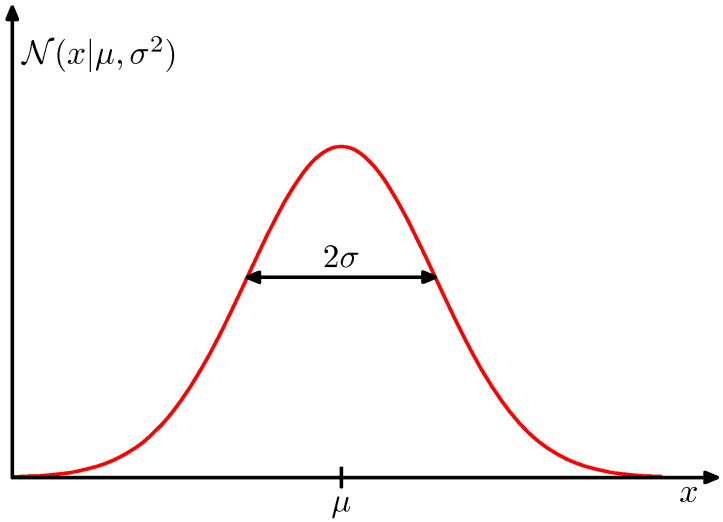

- Гауссово распределение $x \in \mathbb{R}^1$ с $\mathbb{E}[x] = \mu$, $\mathrm{var}[x] = \sigma^2$

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

- Многомерное гауссово распределение $\mathbf{x} \in \mathbb{R}^d$ с $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$, $\mathrm{cov}[\mathbf{x}] = \boldsymbol{\Sigma}$

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

# Гауссовское распределение



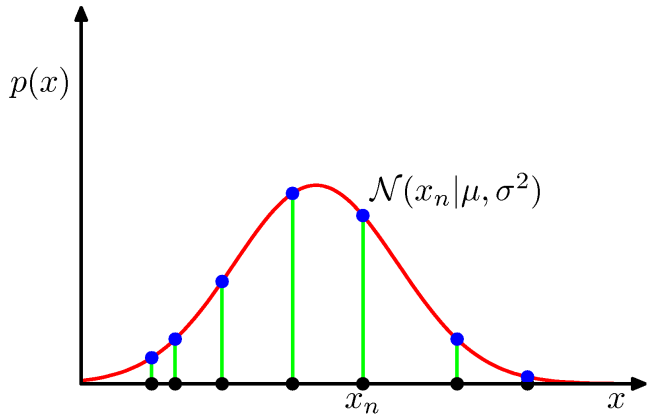

- Гауссово распределение $x \in \mathbb{R}^1$ с $\mathbb{E}[x] = \mu$, $\mathrm{var}[x] = \sigma^2$

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

- Многомерное гауссово распределение $\mathbf{x} \in \mathbb{R}^d$ с $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$, $\mathrm{cov}[\mathbf{x}] = \boldsymbol{\Sigma}$

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

# Гауссовский метод максимального правдоподобия (ММП)




- Правдоподобие i.i.d. гауссовских величин $\mathbf{X}_m = \{x_1, \ldots, x_m\}$

$$p(\mathbf{X}|\mu,\sigma^2) = \prod_{i=1}^{m} \mathcal{N}(x_i|\mu,\sigma^2)$$

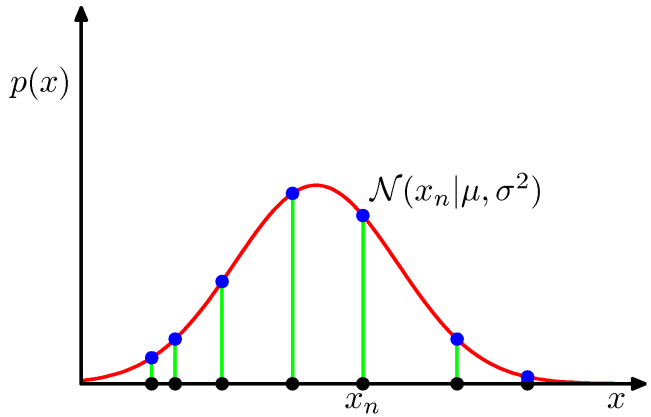
- Логарифм правдоподобия равен

$$\log p(\mathbf{X}_m|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2 - \frac{m}{2}\log\sigma^2 - \frac{m}{2}\log(2\pi) \to \max_{\mu,\sigma^2}$$

- ММП эквивалентен

$$\mu_{ML} = \frac{1}{m}\sum_{i=1}^{m} x_i,\ \sigma^2_{ML} = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{ML})^2$$

# Гауссовский метод максимального правдоподобия (ММП)




- Правдоподобие i.i.d. гауссовских величин $\mathbf{X}_m = \{x_1, \ldots, x_m\}$

$$p(\mathbf{X}|\mu,\sigma^2) = \prod_{i=1}^{m} \mathcal{N}(x_i|\mu,\sigma^2)$$

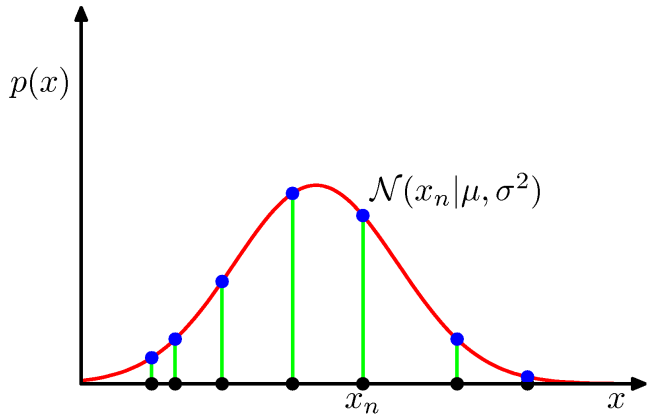
- Логарифм правдоподобия равен

$$\log p(\mathbf{X}_m|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2 - \frac{m}{2}\log\sigma^2 - \frac{m}{2}\log(2\pi) \to \max_{\mu,\sigma^2}$$

- ММП эквивалентен

$$\mu_{ML} = \frac{1}{m}\sum_{i=1}^{m} x_i,\ \sigma^2_{ML} = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{ML})^2$$

# Гауссовский метод максимального правдоподобия (ММП)




- Правдоподобие i.i.d. гауссовских величин $\mathbf{X}_m = \{x_1, \ldots, x_m\}$

$$p(\mathbf{X}|\mu,\sigma^2) = \prod_{i=1}^{m} \mathcal{N}(x_i|\mu,\sigma^2)$$

- Логарифм правдоподобия равен

$$\log p(\mathbf{X}_m|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2 - \frac{m}{2}\log\sigma^2 - \frac{m}{2}\log(2\pi) \to \max_{\mu,\sigma^2}$$

- ММП эквивалентен

$$\mu_{ML} = \frac{1}{m}\sum_{i=1}^{m} x_i,\ \sigma^2_{ML} = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{ML})^2$$

# Неопределённость

- Повторяемые события $\Rightarrow$ классическая (частотная) интерпретация вероятности
- Байесовский подход: вероятности обеспечивают количественную оценку неопределенности
- Рассмотрим неопределенное (неповторяемое) событие:
  - "арктические льды исчезнут к концу века?"
  - мы можем иметь некоторое представление о том, как быстро, по нашему мнению, тает Полярный лед
  - мы получаем свежие данные: например, со спутника наблюдения земли мы можем пересмотреть наше мнение о скорости потери льда
  - нам нужно количественно оценить наше выражение неопределенности и внести точные изменения неопределенности в свете новых данных

# Неопределённость

- Повторяемые события $\Rightarrow$ классическая (частотная) интерпретация вероятности
- Байесовский подход: вероятности обеспечивают количественную оценку неопределенности
- Рассмотрим неопределенное (неповторяемое) событие:
  - "арктические льды исчезнут к концу века?"
  - мы можем иметь некоторое представление о том, как быстро, по нашему мнению, тает Полярный лед
  - мы получаем свежие данные: например, со спутника наблюдения земли мы можем пересмотреть наше мнение о скорости потери льда
  - нам нужно количественно оценить наше выражение неопределенности и внести точные изменения неопределенности в свете новых данных

# Неопределённость

- Повторяемые события $\Rightarrow$ классическая (частотная) интерпретация вероятности
- Байесовский подход: вероятности обеспечивают количественную оценку неопределенности
- Рассмотрим неопределенное (неповторяемое) событие:
  - “арктические льды исчезнут к концу века?”
  - мы можем иметь некоторое представление о том, как быстро, по нашему мнению, тает Полярный лед
  - мы получаем свежие данные: например, со спутника наблюдения земли мы можем пересмотреть наше мнение о скорости потери льда
  - нам нужно количественно оценить наше выражение неопределенности и внести точные изменения неопределенности в свете новых данных

# Неопределённость

- Повторяемые события $\Rightarrow$ классическая (частотная) интерпретация вероятности
- Байесовский подход: вероятности обеспечивают количественную оценку неопределенности
- Рассмотрим неопределенное (неповторяемое) событие:
  - “арктические льды исчезнут к концу века?”
  - мы можем иметь некоторое представление о том, как быстро, по нашему мнению, тает Полярный лед
  - мы получаем свежие данные: например, со спутника наблюдения земли мы можем пересмотреть наше мнение о скорости потери льда
  - нам нужно количественно оценить наше выражение неопределенности и внести точные изменения неопределенности в свете новых данных

# Неопределённость

- Повторяемые события $\Rightarrow$ классическая (частотная) интерпретация вероятности
- Байесовский подход: вероятности обеспечивают количественную оценку неопределенности
- Рассмотрим неопределенное (неповторяемое) событие:
  - “арктические льды исчезнут к концу века?”
  - мы можем иметь некоторое представление о том, как быстро, по нашему мнению, тает Полярный лед
  - мы получаем свежие данные: например, со спутника наблюдения земли мы можем пересмотреть наше мнение о скорости потери льда
  - нам нужно количественно оценить наше выражение неопределенности и внести точные изменения неопределенности в свете новых данных

# Пример: оценка (подгонка) кривой

- Модель данных: $y = f(x, \mathbf{w}) + \varepsilon$, $\varepsilon$ — шум
- Количественная неопределенность параметров модели $\mathbf{w}$?
- Априорное распределение $p(\mathbf{w})$ фиксирует наши предположения о $\mathbf{w}$ перед наблюдением данных!

Вероятность vs. сложность (Колмогоров):

— Предсказать случайные редкие события практически невозможно $\Rightarrow$ их описание очень длинное $\Rightarrow$ сложное

— $\mathbf{w}$ определим "сложность" модели

— $p(\mathbf{w})$ оценивает эту сложность

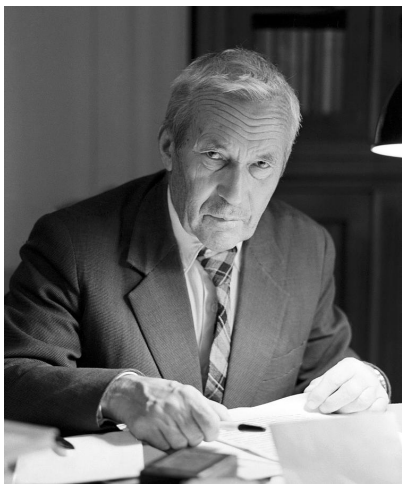


Рис. – Колмогоров А.Н. (1903-1987)

# Пример: оценка (подгонка) кривой

- Модель данных: $y = f(x, \mathbf{w}) + \varepsilon$, $\varepsilon$ — шум
- Количественная неопределенность параметров модели $\mathbf{w}$?
- Априорное распределение $p(\mathbf{w})$ фиксирует наши предположения о $\mathbf{w}$ перед наблюдением данных!

Вероятность vs. сложность (Колмогоров):

— Предсказать случайные редкие события практически невозможно $\Rightarrow$ их описание очень длинное $\Rightarrow$ сложное

— $\mathbf{w}$ определим "сложность" модели

— $p(\mathbf{w})$ оценивает эту сложность

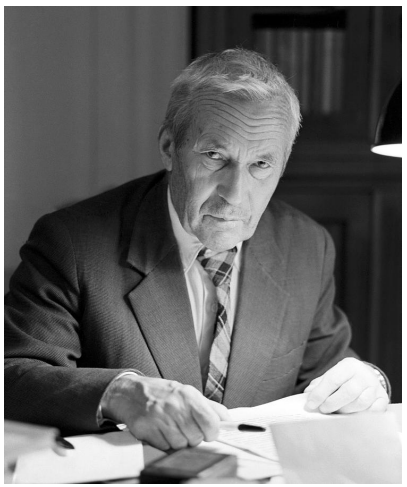


Рис. – Колмогоров А.Н. (1903-1987)

# Пример: оценка (подгонка) кривой

- Модель данных: $y = f(x, \mathbf{w}) + \varepsilon$, $\varepsilon$ — шум
- Количественная неопределенность параметров модели $\mathbf{w}$?
- Априорное распределение $p(\mathbf{w})$ фиксирует наши предположения о $\mathbf{w}$ перед наблюдением данных!

Вероятность vs. сложность (Колмогоров):

- Предсказать случайные редкие события практически невозможно $\Rightarrow$ их описание очень длинное $\Rightarrow$ сложное
- $\mathbf{w}$ определим "сложность" модели
- $p(\mathbf{w})$ оценивает эту сложность

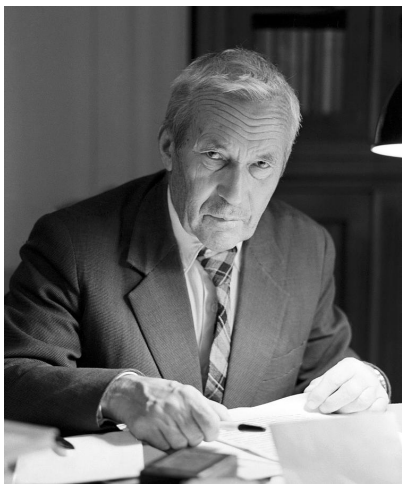


Рис. – Колмогоров А.Н. (1903-1987)

# Пример: оценка (подгонка) кривой

- Модель данных: $y = f(x, \mathbf{w}) + \varepsilon$, $\varepsilon$ — шум
- Количественная неопределенность параметров модели $\mathbf{w}$?
- Априорное распределение $p(\mathbf{w})$ фиксирует наши предположения о $\mathbf{w}$ перед наблюдением данных!

Вероятность vs. сложность (Колмогоров):

— Предсказать случайные редкие события практически невозможно $\Rightarrow$ их описание очень длинное $\Rightarrow$ сложное

— $\mathbf{w}$ определим "сложность" модели

— $p(\mathbf{w})$ оценивает эту сложность

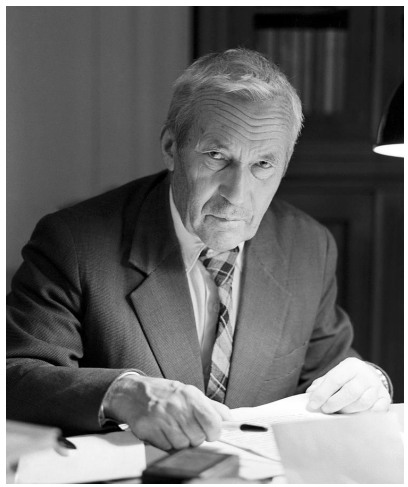


Рис. – Колмогоров А.Н. (1903-1987)

# Пример: оценка (подгонка) кривой

- Модель данных: $y = f(x, \mathbf{w}) + \varepsilon$, $\varepsilon$ — шум
- Количественная неопределенность параметров модели $\mathbf{w}$?
- Априорное распределение $p(\mathbf{w})$ фиксирует наши предположения о $\mathbf{w}$ перед наблюдением данных!

Вероятность vs. сложность (Колмогоров):

— Предсказать случайные редкие события практически невозможно $\Rightarrow$ их описание очень длинное $\Rightarrow$ сложное

— $\mathbf{w}$ определим “сложность” модели

— $p(\mathbf{w})$ оценивает эту сложность

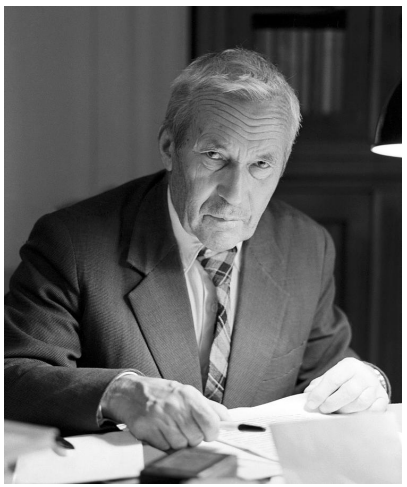


Рис. – Колмогоров А.Н. (1903-1987)

# Пример: оценка (подгонка) кривой

- Наблюдаемые данные $\mathcal{D}_m = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ влияют на условное распределение $p(\mathbf{w}|\mathcal{D}_m)$:

$$p(\mathbf{w}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D}_m)}$$

- $p(\mathcal{D}_m|\mathbf{w})$ — функция правдоподобия (насколько вероятен наблюдаемый набор данных для различных настроек вектора параметров $\mathbf{w}$)
- Константа нормализации (фактические данные)

$$p(\mathcal{D}_m) = \int p(\mathcal{D}_m|\mathbf{w})p(\mathbf{w})d\mathbf{w}$$

- Общая форма:

$$\text{posterior} \sim \text{likelihood} \times \text{prior}$$

# Пример: оценка (подгонка) кривой

- Наблюдаемые данные $\mathcal{D}_m = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ влияют на условное распределение $p(\mathbf{w}|\mathcal{D}_m)$:

$$p(\mathbf{w}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D}_m)}$$

- $p(\mathcal{D}_m|\mathbf{w})$ — функция правдоподобия (насколько вероятен наблюдаемый набор данных для различных настроек вектора параметров $\mathbf{w}$)
- Константа нормализации (фактические данные)

$$p(\mathcal{D}_m) = \int p(\mathcal{D}_m|\mathbf{w})p(\mathbf{w})d\mathbf{w}$$

- Общая форма:

$$\text{posterior} \sim \text{likelihood} \times \text{prior}$$

# Пример: оценка (подгонка) кривой

- Наблюдаемые данные $\mathcal{D}_m = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ влияют на условное распределение $p(\mathbf{w}|\mathcal{D}_m)$:

$$p(\mathbf{w}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D}_m)}$$

- $p(\mathcal{D}_m|\mathbf{w})$ — функция правдоподобия (насколько вероятен наблюдаемый набор данных для различных настроек вектора параметров $\mathbf{w}$)
- Константа нормализации (фактические данные)

$$p(\mathcal{D}_m) = \int p(\mathcal{D}_m|\mathbf{w})p(\mathbf{w})d\mathbf{w}$$

- Общая форма:

$$\text{posterior} \sim \text{likelihood} \times \text{prior}$$

- **Частотные подход**:
  - $\mathbf{w}$ — фиксированный параметр,
  - погрешности ее оценок, полученные с учетом распределения возможных наборов данных $\mathcal{D}_m$
- **Байесовский подход**:
  - существует только один (наблюдаемый) набор данных,
  - неопределенность параметров выражается через распределение вероятностей $\mathbf{w}$
- Включение априорных знаний возникает естественным образом

- **Частотные подход**:
  - $\mathbf{w}$ — фиксированный параметр,
  - погрешности ее оценок, полученные с учетом распределения возможных наборов данных $\mathcal{D}_m$
- **Байесовский подход**:
  - существует только один (наблюдаемый) набор данных,
  - неопределенность параметров выражается через распределение вероятностей $\mathbf{w}$
- Включение априорных знаний возникает естественным образом

# Байесовский vs. частотный подходы

- **Частотные подход**:
  - $\mathbf{w}$ — фиксированный параметр,
  - погрешности ее оценок, полученные с учетом распределения возможных наборов данных $\mathcal{D}_m$
- **Байесовский подход**:
  - существует только один (наблюдаемый) набор данных,
  - неопределенность параметров выражается через распределение вероятностей $\mathbf{w}$
- Включение априорных знаний возникает естественным образом

- ММП:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \log p(\mathcal{D}_m|\mathbf{w})$$

- Оценка максимума апостериорной вероятности (МАП)
  - Апостериорное распределение

$$p(\mathbf{w}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D}_m)}$$

  - МАП

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathcal{D}_m)$$

  - МАП $=$ регуляризованный ММП:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} [\log p(\mathcal{D}_m|\mathbf{w}) + \log p(\mathbf{w})]$$

- ММП:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \log p(\mathcal{D}_m|\mathbf{w})$$

- Оценка максимума апостериорной вероятности (МАП)
  - Апостериорное распределение

$$p(\mathbf{w}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D}_m)}$$

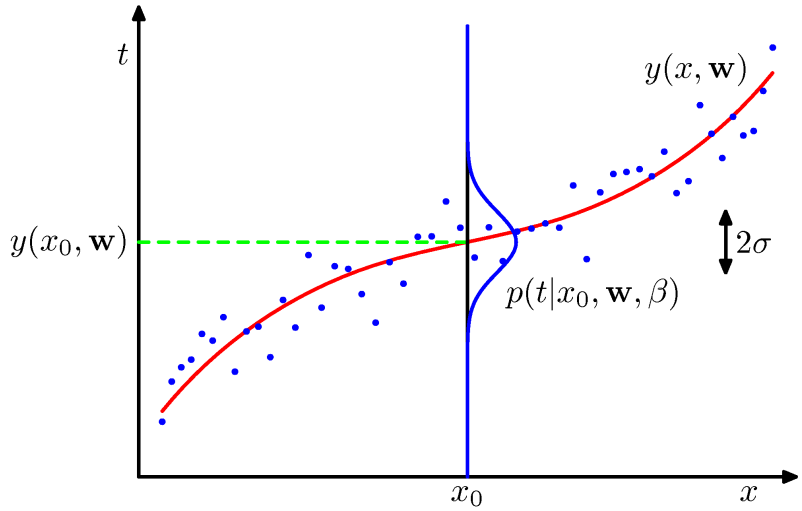
  - МАП

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathcal{D}_m)$$

  - МАП $\equiv$ регуляризованный ММП:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} [\log p(\mathcal{D}_m|\mathbf{w}) + \log p(\mathbf{w})]$$

# Приближение функции по данным




- Данные $\mathcal{D}_m = \{\mathbf{X}_m, \mathbf{Y}_m\} = \{(x_i, y_i)\}_{i=1}^m$
- Вероятностная модель

$$p(y|x, \mathbf{w}, \beta) = \mathcal{N}(y|y(x, \mathbf{w}), \beta^{-1}),$$

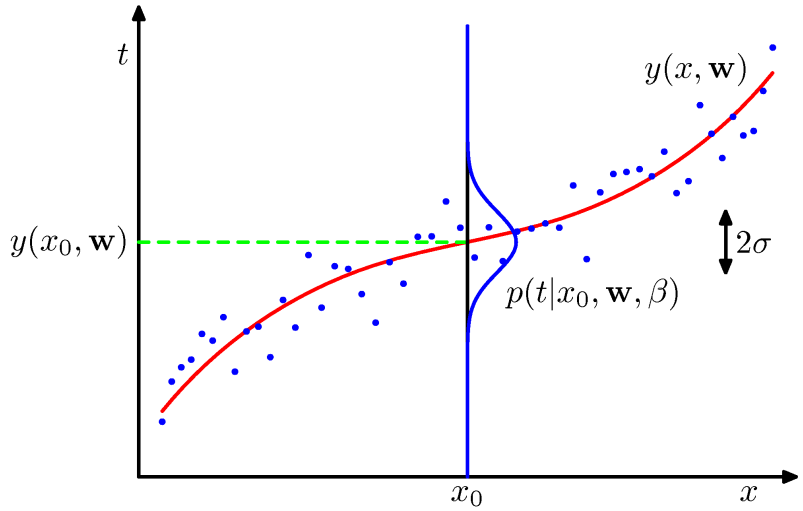
где

  - среднее значение задаётся полиномом $y(x, \mathbf{w})$
  - точность шума задается параметром $\beta^{-1} = \sigma^2$

- Правдоподобие

$$p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) = \prod_{i=1}^m \mathcal{N}(y_i|y(x_i, \mathbf{w}), \beta^{-1})$$

# Приближение функции по данным




- Данные $\mathcal{D}_m = \{\mathbf{X}_m, \mathbf{Y}_m\} = \{(x_i, y_i)\}_{i=1}^m$
- Вероятностная модель

$$p(y|x, \mathbf{w}, \beta) = \mathcal{N}(y|y(x, \mathbf{w}), \beta^{-1}),$$

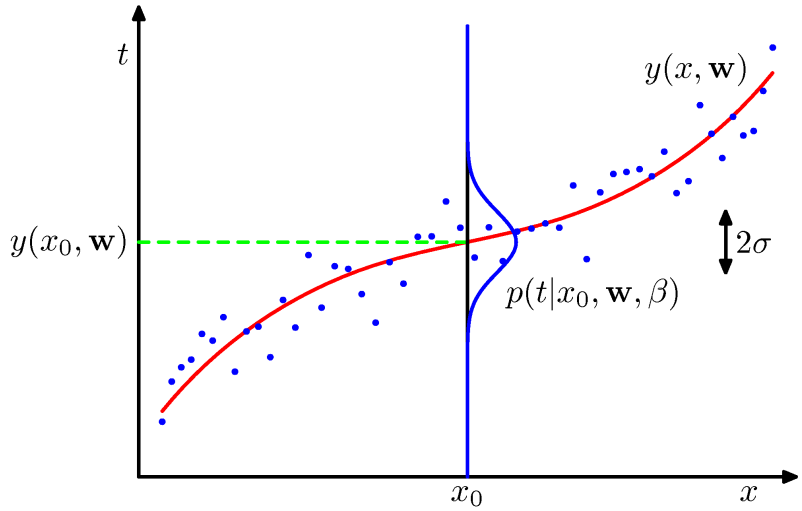
где

  - среднее значение задаётся полиномом $y(x, \mathbf{w})$
  - точность шума задается параметром $\beta^{-1} = \sigma^2$

- Правдоподобие

$$p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) = \prod_{i=1}^m \mathcal{N}(y_i|y(x_i, \mathbf{w}), \beta^{-1})$$

# Приближение функции по данным




- Данные $\mathcal{D}_m = \{\mathbf{X}_m, \mathbf{Y}_m\} = \{(x_i, y_i)\}_{i=1}^m$
- Вероятностная модель

$$p(y|x, \mathbf{w}, \beta) = \mathcal{N}(y|y(x, \mathbf{w}), \beta^{-1}),$$

где

  - среднее значение задаётся полиномом $y(x, \mathbf{w})$
  - точность шума задается параметром $\beta^{-1} = \sigma^2$

- Правдоподобие

$$p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) = \prod_{i=1}^m \mathcal{N}(y_i|y(x_i, \mathbf{w}), \beta^{-1})$$

# Приближение функции по данным

- Логарифм правдоподобия

$$\log p(\mathbf{Y}_m|\mathbf{X}_m,\, \mathbf{w}, \beta) = -\frac{\beta}{2}\sum_{i=1}^{m}(y(x_i, \mathbf{w}) - y_i)^2 + \frac{m}{2}\log\beta - \frac{m}{2}(2\pi)$$

- ММП $\beta$

$$\frac{1}{\beta_{ML}} = \frac{1}{m}\sum_{i=1}^{m}(y(x_i, \mathbf{w}_{ML}) - y_i)^2$$

- Предсказательное распределение

$$p(y|x, \mathbf{w}_{ML}, \beta_{ML}) = \mathcal{N}(y|y(x, \mathbf{w}_{ML}), \beta_{ML}^{-1})$$

# Приближение функции по данным

- Логарифм правдоподобия

$$\log p(\mathbf{Y}_m|\mathbf{X}_m,\ \mathbf{w}, \beta) = -\frac{\beta}{2}\sum_{i=1}^{m}(y(x_i, \mathbf{w}) - y_i)^2 + \frac{m}{2}\log\beta - \frac{m}{2}(2\pi)$$

- ММП $\beta$

$$\frac{1}{\beta_{ML}} = \frac{1}{m}\sum_{i=1}^{m}(y(x_i, \mathbf{w}_{ML}) - y_i)^2$$

- Предсказательное распределение

$$p(y|x, \mathbf{w}_{ML}, \beta_{ML}) = \mathcal{N}(y|y(x, \mathbf{w}_{ML}), \beta_{ML}^{-1})$$

# Приближение функции по данным

- Априорное распределение над коэффициентами полиномов $\mathbf{w}$

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathrm{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

- Апостериорное распределение

$$p(\mathbf{w}|\mathbf{X}_m, \mathbf{Y}_m, \alpha, \beta) \sim p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

- Максимум апостериорной вероятности

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left[\frac{\beta}{2}\sum_{i=1}^{n}(y(x_i, \mathbf{w}) - y_i)^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right]$$

- МАП $\equiv$ Гребневая регрессия с $\lambda = \frac{\alpha}{\beta}$

# Приближение функции по данным

- Априорное распределение над коэффициентами полиномов $\mathbf{w}$

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathrm{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

- Апостериорное распределение

$$p(\mathbf{w}|\mathbf{X}_m, \mathbf{Y}_m, \alpha, \beta) \sim p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

- Максимум апостериорной вероятности

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left[\frac{\beta}{2}\sum_{i=1}^{n}(y(x_i, \mathbf{w}) - y_i)^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right]$$

- МАП $\equiv$ Гребневая регрессия с $\lambda = \frac{\alpha}{\beta}$

- Используя обучающую выборку входов $\mathbf{X}_m$ и выходов $\mathbf{Y}_m$, нужно предсказать для тестового объекта $x$ значение $y$
- Мы хотим оценить распределение для прогнозирования $p(y|x, \mathbf{X}_m, \mathbf{Y}_m)$
- Распределение для прогнозирования (predictive distribution)

$$p(y|x, \mathbf{X}_m, \mathbf{Y}_m) = \int p(y|x, \mathbf{w}) p(\mathbf{w}|\mathbf{X}_m, \mathbf{Y}_m) d\mathbf{w}$$

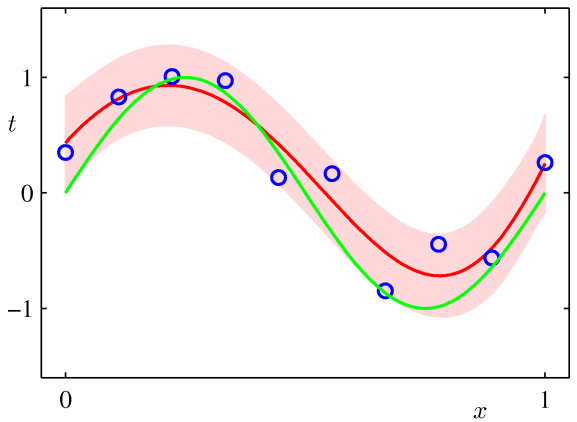



Рис. – Распределения для прогноза для полинома степени $M = 9$ и фиксированными параметрами $\alpha = 5 \times 10^{-3}$ и известной дисперсией шума $\beta = 11.1$

# Байесовский подход к задаче приближения функции по данным

- Используя обучающую выборку входов $\mathbf{X}_m$ и выходов $\mathbf{Y}_m$, нужно предсказать для тестового объекта $x$ значение $y$
- Мы хотим оценить распределение для прогнозирования $p(y|x, \mathbf{X}_m, \mathbf{Y}_m)$
- Распределение для прогнозирования (predictive distribution)

$$p(y|x, \mathbf{X}_m, \mathbf{Y}_m) = \int p(y|x, \mathbf{w}) p(\mathbf{w}|\mathbf{X}_m, \mathbf{Y}_m) d\mathbf{w}$$

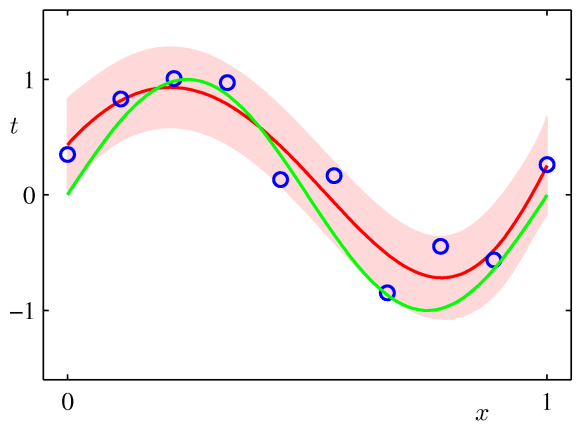



Рис. – Распределения для прогноза для полинома степени $M = 9$ и фиксированными параметрами $\alpha = 5 \times 10^{-3}$ и известной дисперсией шума $\beta = 11.1$

# Байесовский подход к задаче приближения функции по данным

- Используя обучающую выборку входов $\mathbf{X}_m$ и выходов $\mathbf{Y}_m$, нужно предсказать для тестового объекта $x$ значение $y$
- Мы хотим оценить распределение для прогнозирования $p(y|x, \mathbf{X}_m, \mathbf{Y}_m)$
- Распределение для прогнозирования (predictive distribution)

$$p(y|x, \mathbf{X}_m, \mathbf{Y}_m) = \int p(y|x, \mathbf{w}) p(\mathbf{w}|\mathbf{X}_m, \mathbf{Y}_m) d\mathbf{w}$$

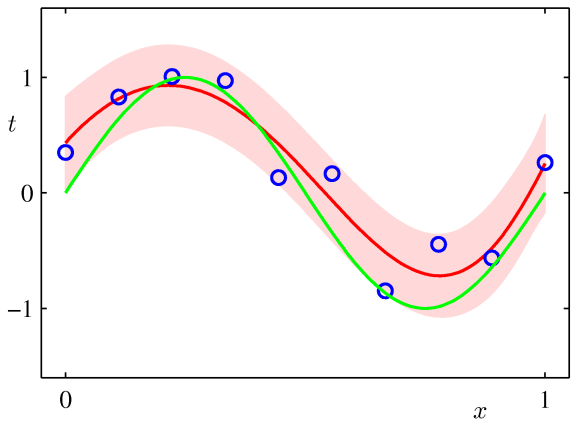



Рис. – Распределения для прогноза для полинома степени $M = 9$ и фиксированными параметрами $\alpha = 5 \times 10^{-3}$ и известной дисперсией шума $\beta = 11.1$

- Модель с линейным функциональным базисом

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

  где $\phi_j(\mathbf{x})$ — известные базисные функции

- Типичные базисные функции

$$\phi_j(\mathbf{x}) = x_{j_1}^{j_0},\ \phi_j(\mathbf{x}) = \exp\left\{-\frac{\|\mathbf{x} - \boldsymbol{\mu}_j\|^2}{2s^2}\right\},$$

$$\phi(\mathbf{x}) = \sigma\left({\boldsymbol{\mu}_{j,1}}^{\mathrm{T}} \mathbf{x} + \mu_{j,0}\right),\ \sigma(a) = \frac{1}{1 + e^{-a}}$$

- Мы предполагаем, что параметры базисных функций фиксированы

- Модель с линейным функциональным базисом

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

  где $\phi_j(\mathbf{x})$ — известные базисные функции
- Типичные базисные функции

$$\phi_j(\mathbf{x}) = x_{j_1}^{j_0},\ \phi_j(\mathbf{x}) = \exp\left\{-\frac{\|\mathbf{x} - \boldsymbol{\mu}_j\|^2}{2s^2}\right\},$$

$$\phi(\mathbf{x}) = \sigma\left({\boldsymbol{\mu}_{j,1}}^{\mathrm{T}}\mathbf{x} + \mu_{j,0}\right),\ \sigma(a) = \frac{1}{1 + e^{-a}}$$

- Мы предполагаем, что параметры базисных функций фиксированы

# Метод наименьших квадратов (МНК) = оценка максимального правдоподобия

- Оптимизируем логарифм правдоподобия:

$$\mathbf{w}_{ML} = \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}, \;\; \boldsymbol{\Phi} = \left\{\left(\boldsymbol{\phi}_i(\mathbf{x}_j)\right)_{j=0}^{M-1}\right\}_{i=1}^{N}$$

$$\frac{1}{\beta_{ML}} = \frac{1}{m}\sum_{i=1}^{m}\left\{y_i - \mathbf{w}_{ML}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_i)\right\}^2$$

- Регуляризованный МНК

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w}) \to \min_{\mathbf{w}}$$

$$\frac{1}{2}\sum_{i=1}^{m}\left\{y_i - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_i)\right\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \to \min_{\mathbf{w}}$$

$$\mathbf{w}_{LS} = \left(\lambda\mathbf{I} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{Y}_m$$

# Метод наименьших квадратов (МНК) = оценка максимального правдоподобия

- Оптимизируем логарифм правдоподобия:

$$\mathbf{w}_{ML} = \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}, \;\; \boldsymbol{\Phi} = \left\{\left(\boldsymbol{\phi}_i(\mathbf{x}_j)\right)_{j=0}^{M-1}\right\}_{i=1}^{N}$$

$$\frac{1}{\beta_{ML}} = \frac{1}{m}\sum_{i=1}^{m}\left\{y_i - \mathbf{w}_{ML}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_i)\right\}^2$$

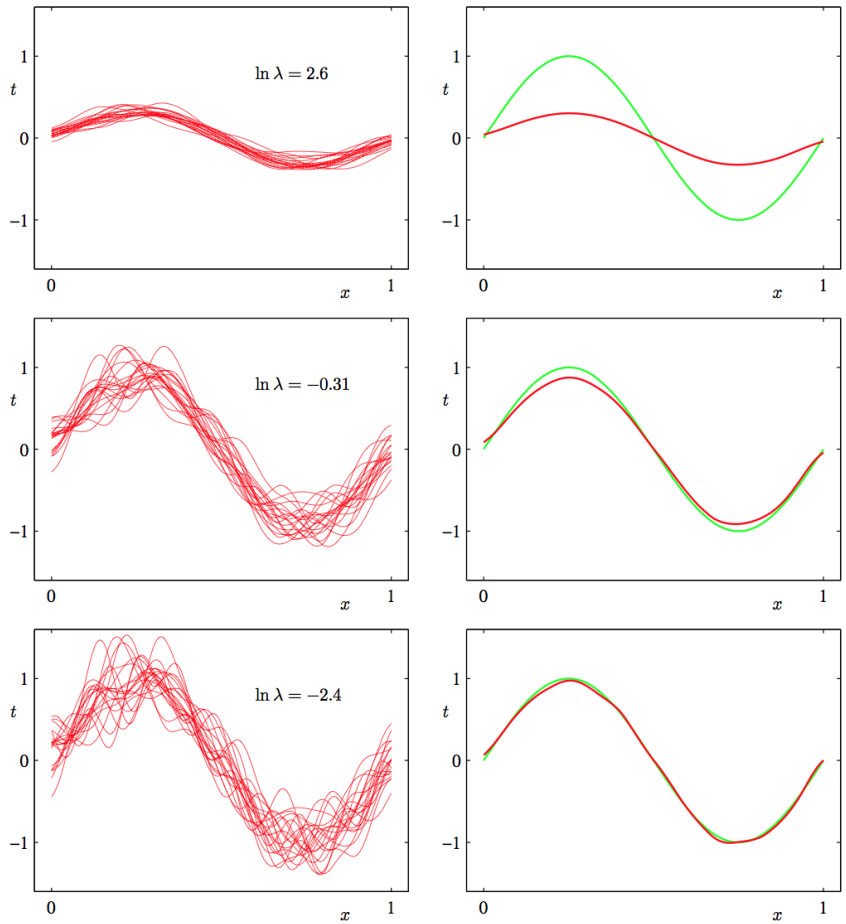
- Регуляризованный МНК

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w}) \to \min_{\mathbf{w}}$$

$$\frac{1}{2}\sum_{i=1}^{m}\left\{y_i - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_i)\right\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \to \min_{\mathbf{w}}$$

$$\mathbf{w}_{LS} = \left(\lambda\mathbf{I} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{Y}_m$$

# Разложение на смещение и разброс




Зависимость смещения/разброса от сложности модели (регуляризации): $L = 100$ наборов данных, $m = 25$ точек (объектов) в каждом, $M = 25$ гауссовских базисных функций. Правая колонка: среднее по $100$ обучением (красные кривые)

- Правдоподобие $p(\mathcal{D}_m|\mathbf{w})$ — экспонента квадрата функции от $\mathbf{w}$. Таким образом, сопряженное априорное распределение

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$$

- Апостериорное распределение на $\mathcal{D}_m = \{\mathbf{Y}_m, \mathbf{X}_m\}$

$$p(\mathbf{w}|\mathcal{D}_m) = \mathcal{N}(\mathbf{w}|\mathbf{m}_m, \mathbf{S}_m),$$

где

$$\mathbf{m}_m = \mathbf{S}_m \left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{Y}_m\right)$$
$$\mathbf{S}_m^{-1} = \mathbf{S}_0^{-1} + \beta \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$$

- Правдоподобие $p(\mathcal{D}_m|\mathbf{w})$ — экспонента квадрата функции от $\mathbf{w}$. Таким образом, сопряженное априорное распределение

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$$

- Апостериорное распределение на $\mathcal{D}_m = \{\mathbf{Y}_m, \mathbf{X}_m\}$

$$p(\mathbf{w}|\mathcal{D}_m) = \mathcal{N}(\mathbf{w}|\mathbf{m}_m, \mathbf{S}_m),$$

где

$$\mathbf{m}_m = \mathbf{S}_m \left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{Y}_m\right)$$
$$\mathbf{S}_m^{-1} = \mathbf{S}_0^{-1} + \beta \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$$

- Типичное априорное распределение

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

- Апостериорное распределение задаётся как

$$p(\mathbf{w}|\mathcal{D}_m) = \mathcal{N}(\mathbf{w}|\mathbf{m}_m, \mathbf{S}_m)$$
$$\mathbf{m}_m = \beta \mathbf{S}_m \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{Y}_m$$
$$\mathbf{S}_m^{-1} = \alpha^{-1}\mathbf{I} + \beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}$$

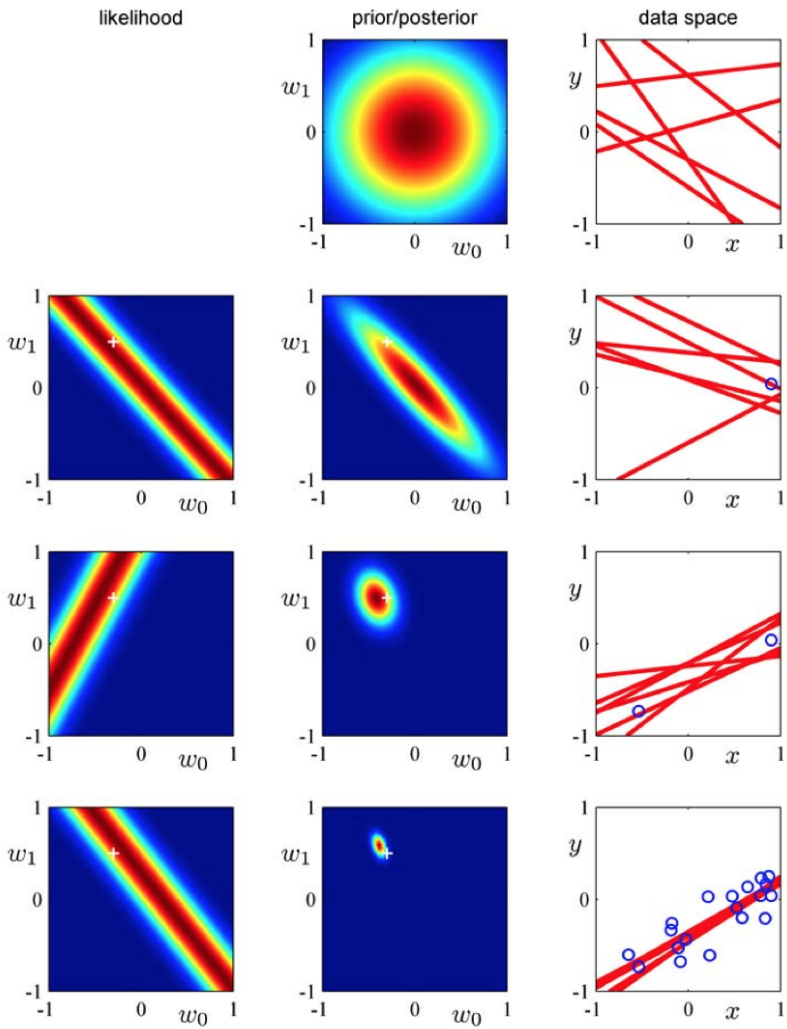
- Типичное априорное распределение

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

- Апостериорное распределение задаётся как

$$p(\mathbf{w}|\mathcal{D}_m) = \mathcal{N}(\mathbf{w}|\mathbf{m}_m, \mathbf{S}_m)$$
$$\mathbf{m}_m = \beta \mathbf{S}_m \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{Y}_m$$
$$\mathbf{S}_m^{-1} = \alpha^{-1}\mathbf{I} + \beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}$$

# Зависимость Байесовской модели от размера выборки




Модель $y(x, \mathbf{w}) = w_0 + w_1 x$

- Делаем предсказания $y$ для нового значения $\mathbf{x}$:

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \int p(y|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta) d\mathbf{w}$$

- Т.к. $p(y|\mathbf{x}, \mathbf{w}, \beta)$ — нормальное, и апостериорное распределение $p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta)$ тоже нормальное, тогда

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \mathcal{N}(y|\mathbf{m}_m^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}), \sigma_m^2(\mathbf{x})),$$

$$\sigma_m^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \mathbf{S}_m \boldsymbol{\phi}(\mathbf{x}),\ \mathbf{S}_m^{-1} = \alpha^{-1}\mathbf{I} + \beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}$$

- $p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta)$ зависит от $\alpha$ и $\beta$! Как их задать? $\Rightarrow$ Полный Байесовский вывод!

- Делаем предсказания $y$ для нового значения $\mathbf{x}$:

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \int p(y|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta) d\mathbf{w}$$

- Т.к. $p(y|\mathbf{x}, \mathbf{w}, \beta)$ — нормальное, и апостериорное распределение $p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta)$ тоже нормальное, тогда

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \mathcal{N}(y|\mathbf{m}_m^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}), \sigma_m^2(\mathbf{x})),$$

$$\sigma_m^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \mathbf{S}_m \boldsymbol{\phi}(\mathbf{x}),\ \mathbf{S}_m^{-1} = \alpha^{-1}\mathbf{I} + \beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}$$

- $p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta)$ зависит от $\alpha$ и $\beta$! Как их задать? $\Rightarrow$ Полный Байесовский вывод!

- Делаем предсказания $y$ для нового значения $\mathbf{x}$:

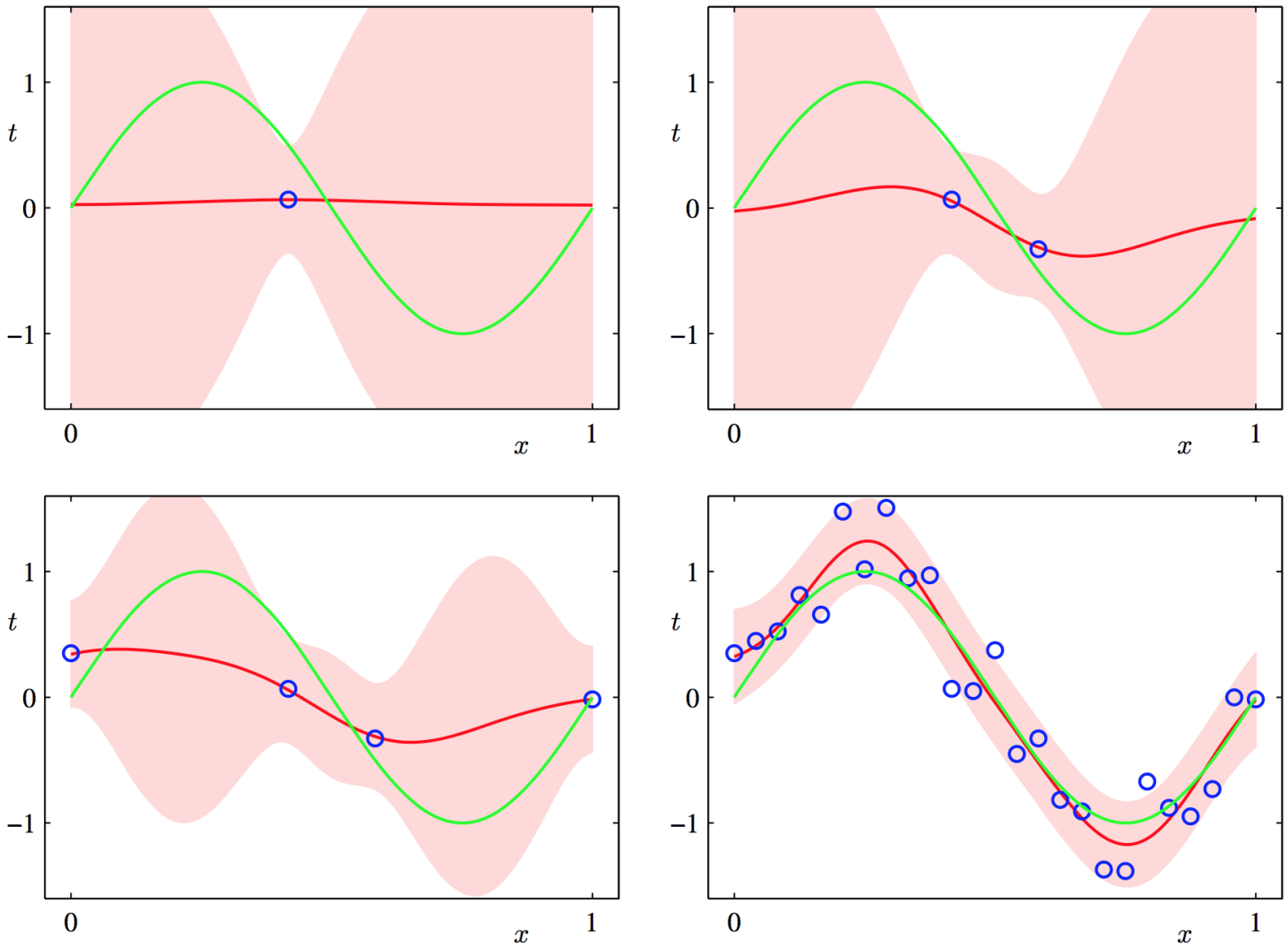
$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \int p(y|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta) d\mathbf{w}$$

- Т.к. $p(y|\mathbf{x}, \mathbf{w}, \beta)$ — нормальное, и апостериорное распределение $p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta)$ тоже нормальное, тогда

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \mathcal{N}(y|\mathbf{m}_m^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}), \sigma_m^2(\mathbf{x})),$$
$$\sigma_m^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \mathbf{S}_m \boldsymbol{\phi}(\mathbf{x}),\ \mathbf{S}_m^{-1} = \alpha^{-1}\mathbf{I} + \beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}$$

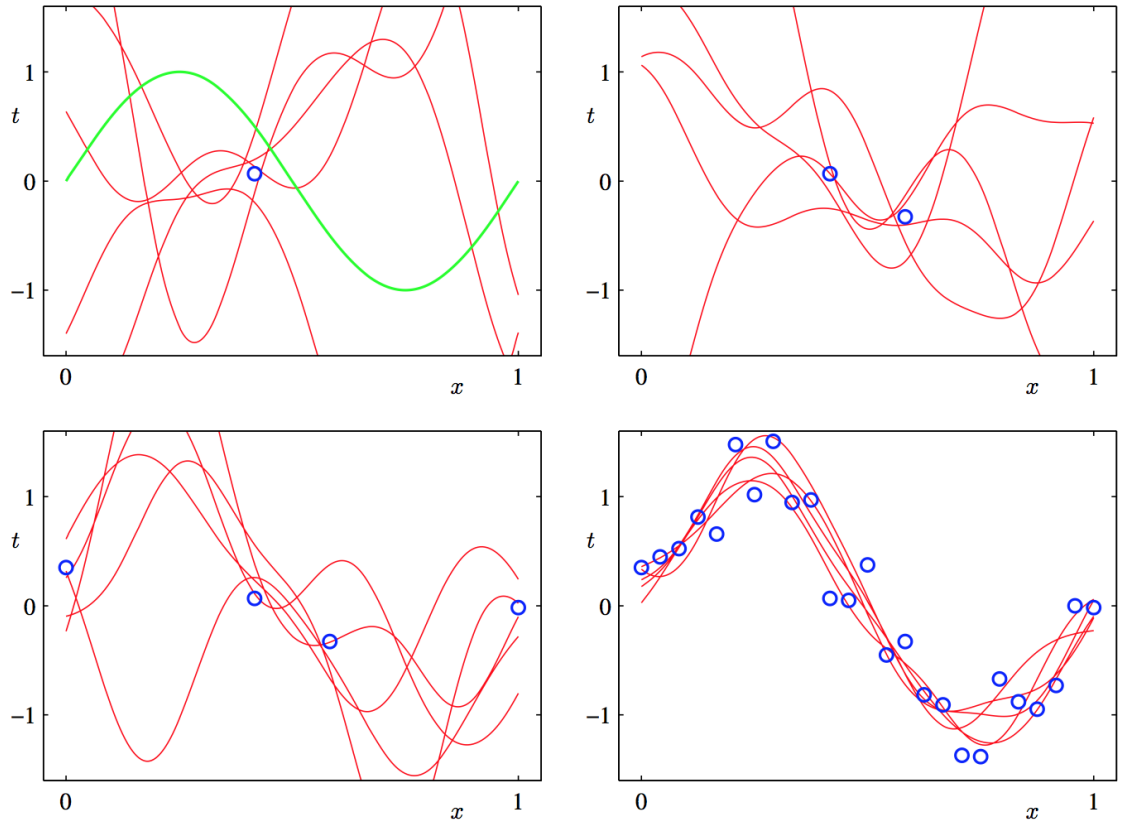
- $p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta)$ зависит от $\alpha$ и $\beta$! Как их задать? $\Rightarrow$ Полный Байесовский вывод!

# Распределение для прогноза




$M = 9$ Гауссовских функций

# Распределение для прогноза




Графики $y(\mathbf{x}, \mathbf{w})$ используют реализации апостериорного распределения над $\mathbf{w} \sim p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta)$ для некоторых $\alpha$ и $\beta$