

# Модели линейной регрессии

Владимир Анатольевич Судаков

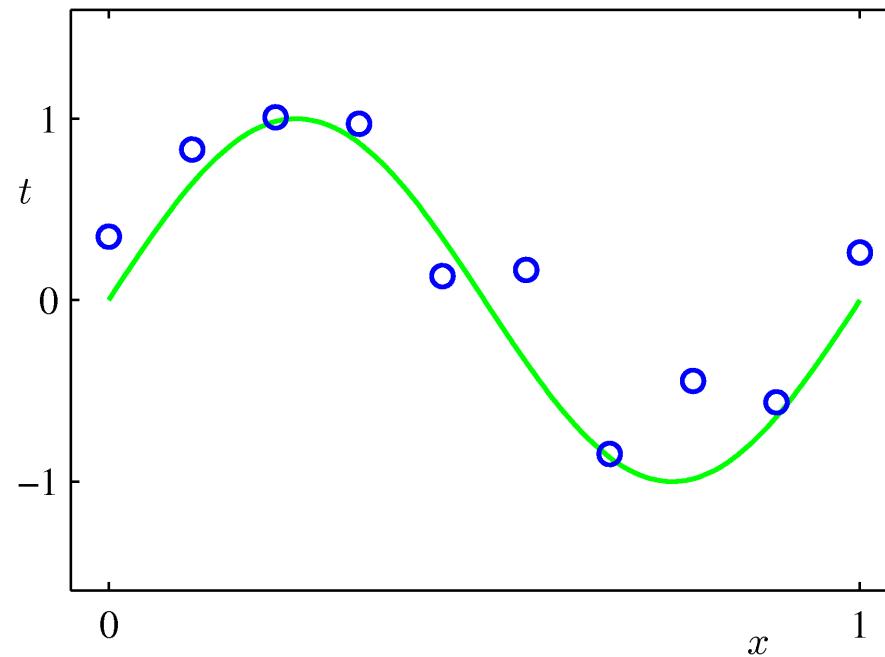
2025

---

# Модели с базисными функциями (1)

---

Пример: полиномиальная аппроксимация кривой



$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \dots + w_M x^M = \sum_{j=0}^M w_j x^j$$

---

## Модели с базисными функциями (2)

---

В формуле:

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})$$

$\phi_j(\mathbf{x})$  называются *базисными функциями*.

Обычно  $\phi_0(\mathbf{x}) = 1$ , так что  $w_0$  действует как смещение.

В простейшем случае мы используем линейные базисные функции:  $\phi_d(\mathbf{x}) = x_d$ .

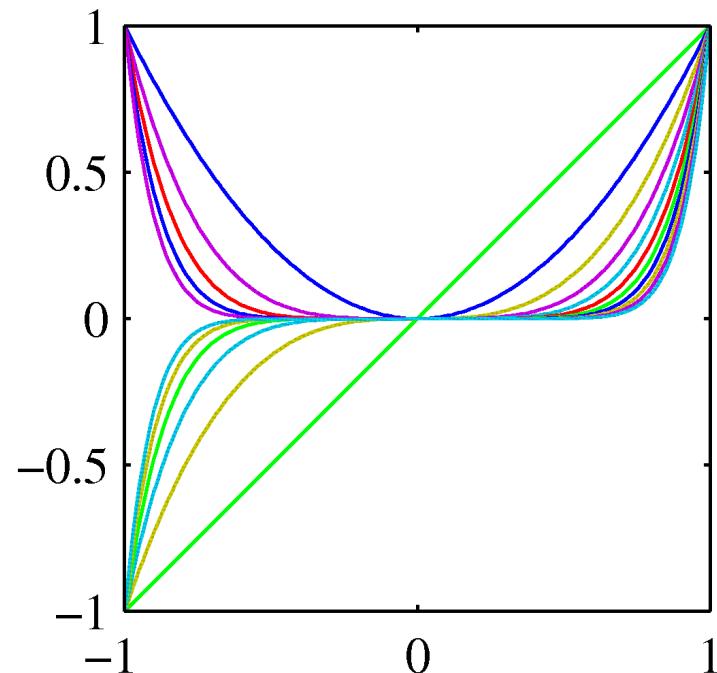
---

# Модели с базисными функциями (3)

Полиномиальные базисные  
функции:

$$\phi_j(x) = x^j.$$

Они глобальны: небольшое  
изменение  $x$  влияет на все  
базисные функции.

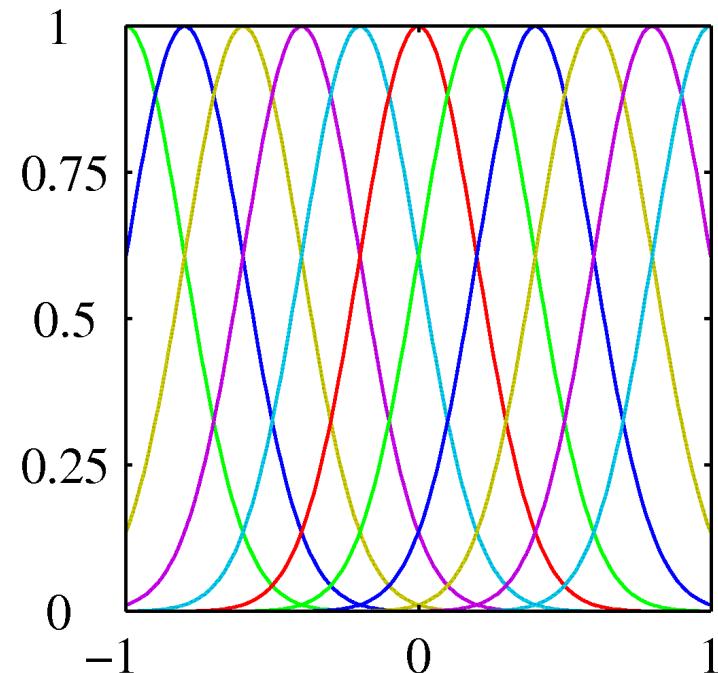


# Модели с базисными функциями (4)

Базисные функции Гаусса:

$$\phi_j(x) = \exp \left\{ -\frac{(x - \mu_j)^2}{2s^2} \right\}$$

Они локальны: небольшое изменение  $x$  влияет только на близлежащие базисные функции.  $\mu_j$  и  $s$  управляют местоположением и масштабом (шириной).



# Модели с базисными функциями (5)

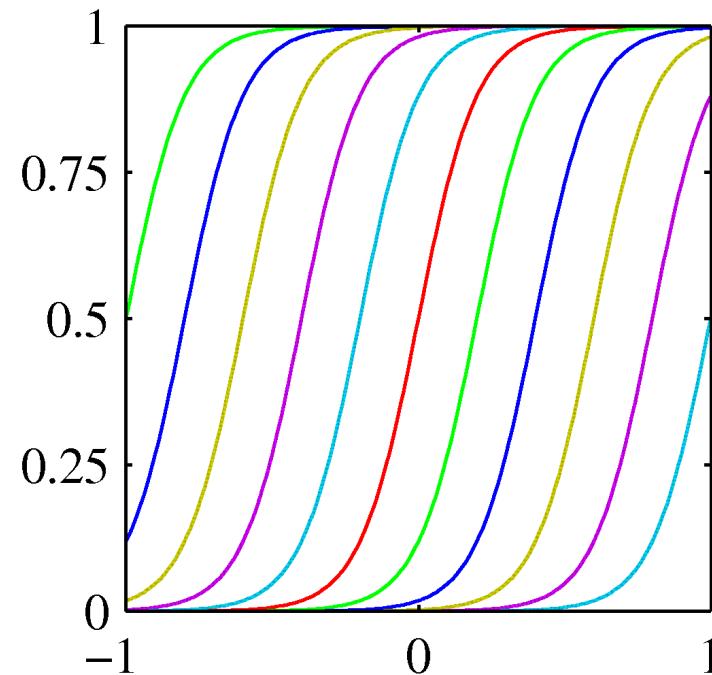
Сигмоидальные базисные функции:

$$\phi_j(x) = \sigma\left(\frac{x - \mu_j}{s}\right)$$

где

$$\sigma(a) = \frac{1}{1 + \exp(-a)}.$$

они локальны: небольшое изменение  $x$  влияет только на близлежащие базисные функции.  $\mu_j$  и  $s$  управляют местоположением и масштабом (наклоном).



## Максимальное правдоподобие и наименьшие квадраты (1)

---

Предположим, что наблюдения производятся с помощью детерминированной функции с добавлением гауссовского шума:

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon \quad \text{где} \quad p(\epsilon|\beta) = \mathcal{N}(\epsilon|0, \beta^{-1})$$

Что то же самое, что сказать:

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

$$\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$$

Учитывая наблюдаемые входные данные, и цели  
 $\mathbf{t} = [t_1, \dots, t_N]^T$ , получаем функцию правдоподобия

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^N \mathcal{N}(t_n | \mathbf{w}^T \phi(\mathbf{x}_n), \beta^{-1}).$$

---

## Максимальное правдоподобие и наименьшие квадраты (2)

---

Логарифмируя, получаем

$$\begin{aligned}\ln p(\mathbf{t}|\mathbf{w}, \beta) &= \sum_{n=1}^N \ln \mathcal{N}(t_n | \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}) \\ &= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(\mathbf{w})\end{aligned}$$

где

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{t_n - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n)\}^2$$

— ошибка суммы квадратов.

---

## Максимальное правдоподобие и наименьшие квадраты (3)

---

Вычисление градиента и установка его в ноль дает

$$\nabla_{\mathbf{w}} \ln p(\mathbf{t}|\mathbf{w}, \beta) = \beta \sum_{n=1}^N \{t_n - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n)\} \boldsymbol{\phi}(\mathbf{x}_n)^T = \mathbf{0}.$$

Решая относительно  $\mathbf{w}$ , получаем

$$\mathbf{w}_{\text{ML}} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \mathbf{t}$$

Псевдообратная  
матрица Мура-  
Пенроуза.  $\boldsymbol{\Phi}^\dagger$

где

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}.$$

# Геометрия наименьших квадратов

Учитывать

$$\mathbf{y} = \Phi \mathbf{w}_{\text{ML}} = [\varphi_1, \dots, \varphi_M] \mathbf{w}_{\text{ML}}.$$

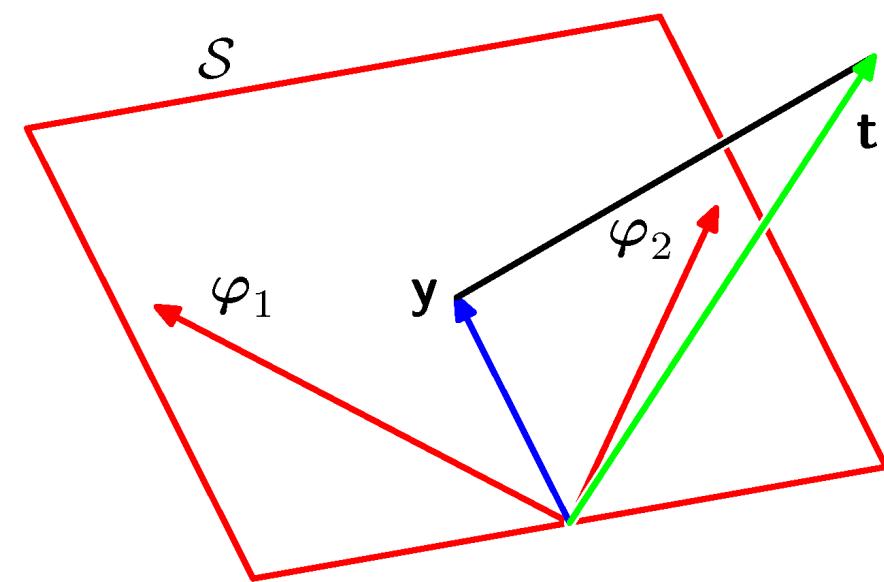
$$\mathbf{y} \in \mathcal{S} \subseteq \mathcal{T}$$

↑  
M-мерный  
↑ N-мерный

$$\mathbf{t} \in \mathcal{T}$$

$\mathcal{S}$  охватывается  $\varphi_1, \dots, \varphi_M$

$\mathbf{w}_{\text{ML}}$  минимизирует  
расстояние между  $\mathbf{t}$  и его  
ортогональной проекцией на  
 $\mathcal{S}$ , т.е.  $\mathbf{y}$ .



# Последовательное обучение

---

Элементы данных рассматриваются по одному за раз (так называемое онлайн-обучение); используется стохастический (последовательный) градиентный спуск:

$$\begin{aligned}\mathbf{w}^{(\tau+1)} &= \mathbf{w}^{(\tau)} - \eta \nabla E_n \\ &= \mathbf{w}^{(\tau)} + \eta(t_n - \mathbf{w}^{(\tau)\mathrm{T}} \phi(\mathbf{x}_n)) \phi(\mathbf{x}_n).\end{aligned}$$

Этот метод называется *алгоритм наименьших средних квадратов (LMS)*.  
Вопрос: как выбрать  $\eta$ ?

---

# Регуляризованный метод наименьших квадратов (1)

---

Рассмотрим функцию ошибки:

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

Ошибка данных + Регуляризация

С функцией ошибки суммы квадратов и квадратичным регуляризатором мы получаем

$$\frac{1}{2} \sum_{n=1}^N \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w}$$

который сводится к минимуму

$$\mathbf{w} = (\lambda \mathbf{I} + \Phi^T \Phi)^{-1} \Phi^T \mathbf{t}.$$

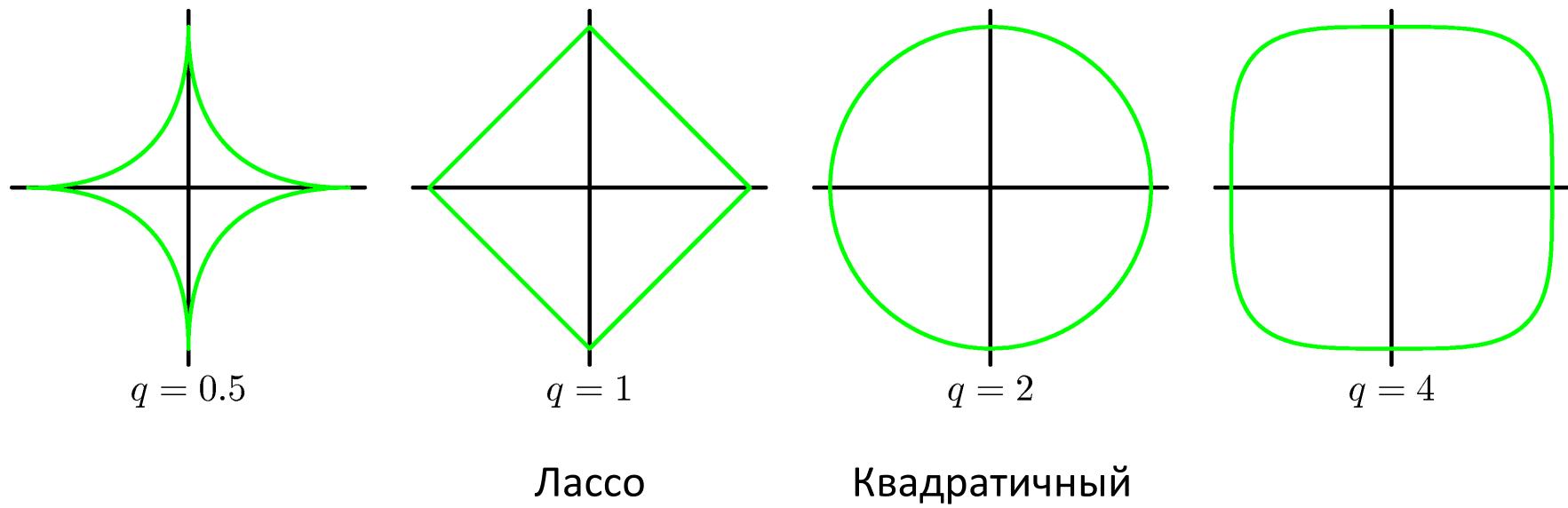
λ называется  
коэффициентом  
регуляризации.

## Регуляризованный метод наименьших квадратов (2)

---

С более общим регуляризатором мы имеем

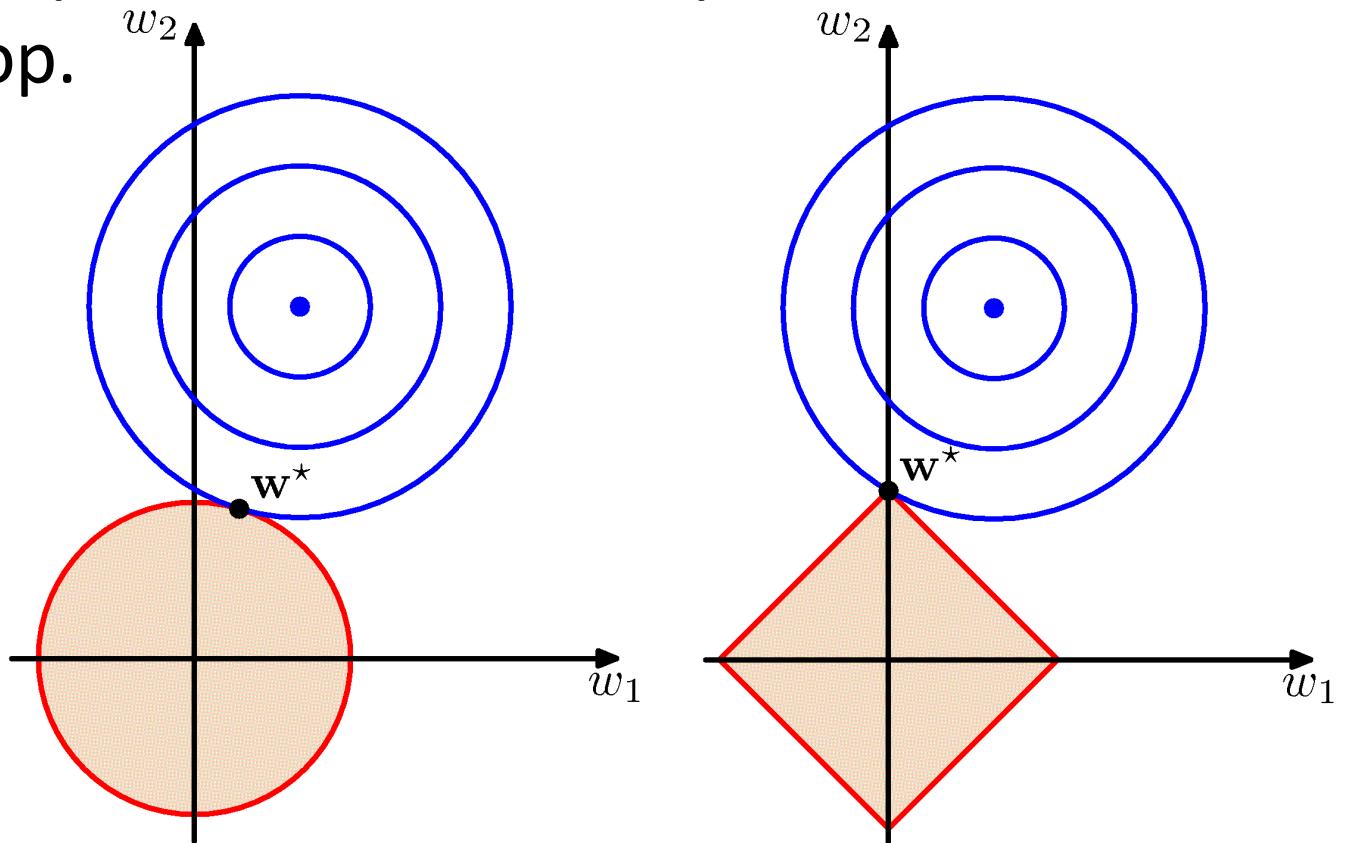
$$\frac{1}{2} \sum_{n=1}^N \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2} \sum_{j=1}^M |w_j|^q$$



## Регуляризованный метод наименьших квадратов (3)

---

Лассо имеет тенденцию генерировать более разреженные решения, чем квадратичный регуляризатор.



# Несколько выходов (1)

---

Аналогично случаю с одним выходом имеем:

$$\begin{aligned} p(\mathbf{t}|\mathbf{x}, \mathbf{W}, \beta) &= \mathcal{N}(\mathbf{t}|\mathbf{y}(\mathbf{W}, \mathbf{x}), \beta^{-1}\mathbf{I}) \\ &= \mathcal{N}(\mathbf{t}|\mathbf{W}^T\boldsymbol{\phi}(\mathbf{x}), \beta^{-1}\mathbf{I}). \end{aligned}$$

Учитывая наблюдаемые входные данные  $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$  и цели,  $\mathbf{T} = [\mathbf{t}_1, \dots, \mathbf{t}_N]^T$  мы получаем логарифмическую функцию правдоподобия

$$\begin{aligned} \ln p(\mathbf{T}|\mathbf{X}, \mathbf{W}, \beta) &= \sum_{n=1}^N \ln \mathcal{N}(\mathbf{t}_n|\mathbf{W}^T\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\mathbf{I}) \\ &= \frac{NK}{2} \ln \left( \frac{\beta}{2\pi} \right) - \frac{\beta}{2} \sum_{n=1}^N \|\mathbf{t}_n - \mathbf{W}^T\boldsymbol{\phi}(\mathbf{x}_n)\|^2. \end{aligned}$$

---

## Несколько выходов (2)

---

Максимизируя по  $\mathbf{W}$ , получаем

$$\mathbf{W}_{\text{ML}} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{T}.$$

Если мы рассмотрим одну целевую переменную,  $t_k$ , мы увидим, что

$$\mathbf{w}_k = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{t}_k = \Phi^\dagger \mathbf{t}_k$$

где  $\mathbf{t}_k = [t_{1k}, \dots, t_{Nk}]^T$ , что идентично случаю с одним выходом.

---

# Разложение смещения и дисперсии (1)

---

Напомним, что мат.ожидание ошибки

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - h(\mathbf{x})\}^2 p(\mathbf{x}) d\mathbf{x} + \underbrace{\iint \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) d\mathbf{x} dt}_{\text{где}}$$

$$h(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}] = \int tp(t|\mathbf{x}) dt.$$

Второй член  $\mathbb{E} [ L ]$  соответствует шуму, присущему случайной величине  $t$ .

А как насчет первого слагаемого?

---

## Разложение смещения-дисперсии (2)

---

Предположим, нам даны несколько наборов данных, каждый размером  $N$ . Любой конкретный набор данных  $\mathcal{D}$  даст определённую функцию  $y(\mathbf{x}; \mathcal{D})$ . Тогда мы имеем

$$\begin{aligned} & \{y(\mathbf{x}; \mathcal{D}) - h(\mathbf{x})\}^2 \\ &= \{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] + \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}^2 \\ &= \{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]\}^2 + \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}^2 \\ &\quad + 2\{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]\}\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}. \end{aligned}$$

---

## Разложение смещения-дисперсии (3)

---

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}} [\{y(\mathbf{x}; \mathcal{D}) - h(\mathbf{x})\}^2] \\ = & \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}^2}_{\text{(bias)}^2} + \underbrace{\mathbb{E}_{\mathcal{D}} [\{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]\}^2]}_{\text{variance}}. \end{aligned}$$

## Разложение смещения и дисперсии (4)

---

Таким образом, мы можем написать

$$\text{expected loss} = (\text{bias})^2 + \text{variance} + \text{noise}$$

где

$$(\text{bias})^2 = \int \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}^2 p(\mathbf{x}) d\mathbf{x}$$

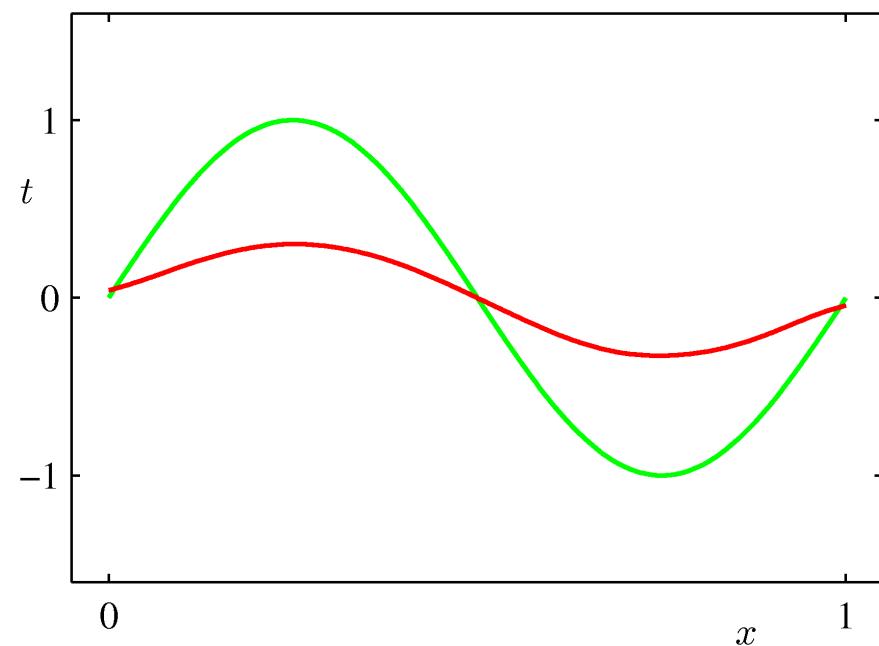
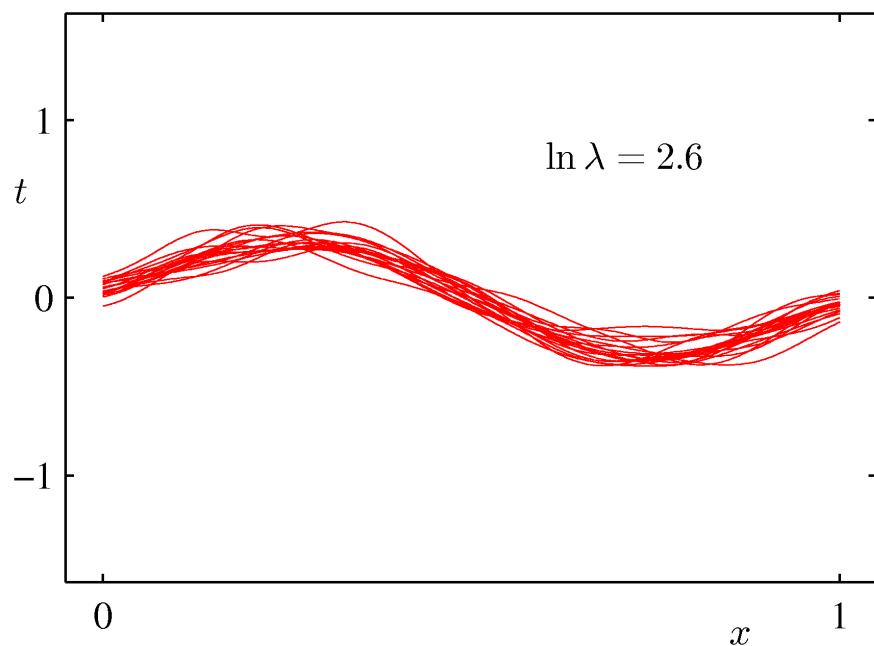
$$\text{variance} = \int \mathbb{E}_{\mathcal{D}} [\{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]\}^2] p(\mathbf{x}) d\mathbf{x}$$

$$\text{noise} = \iint \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) d\mathbf{x} dt$$

# Разложение смещения и дисперсии (5)

---

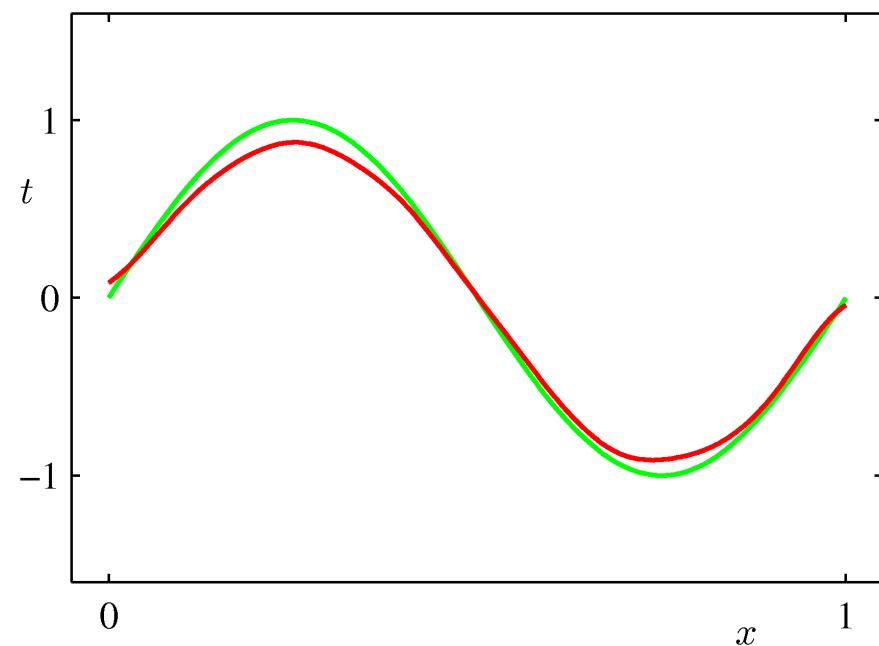
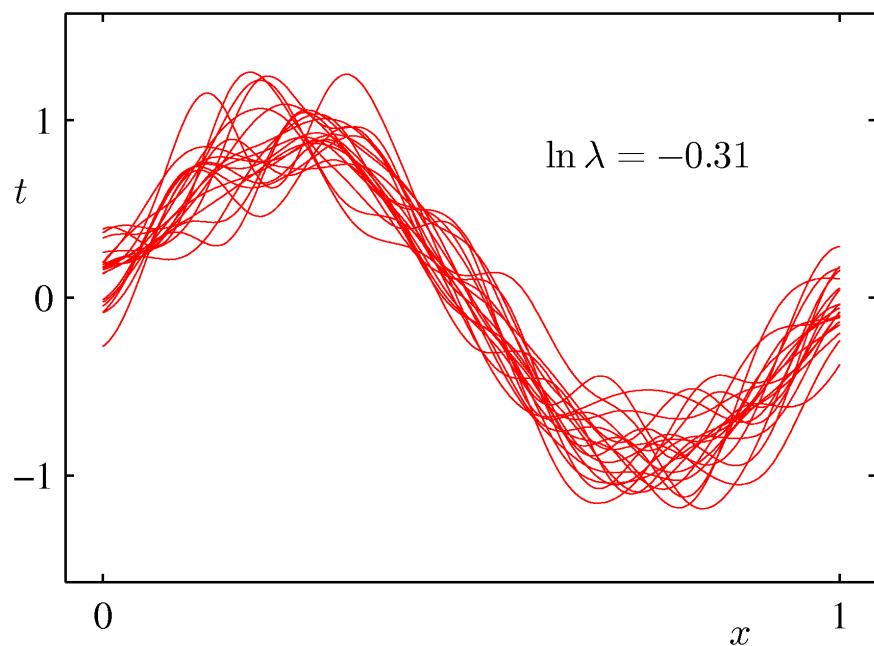
Пример: 25 наборов данных из синусоиды,  
различающихся степенью регуляризации,  $\lambda$ .



# Разложение смещения и дисперсии (6)

---

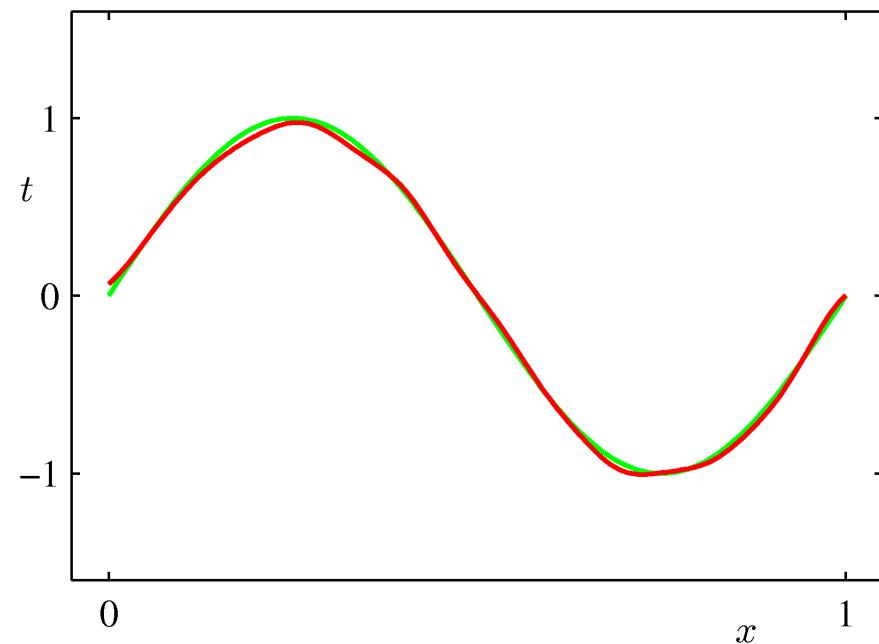
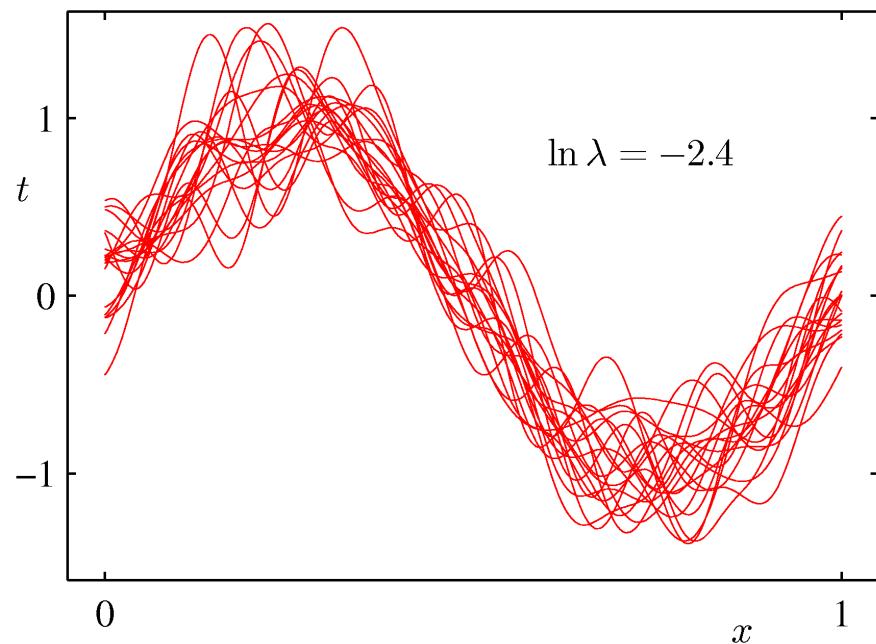
Пример: 25 наборов данных из синусоиды,  
различающихся степенью регуляризации,  $\lambda$ .



# Разложение смещения и дисперсии (7)

---

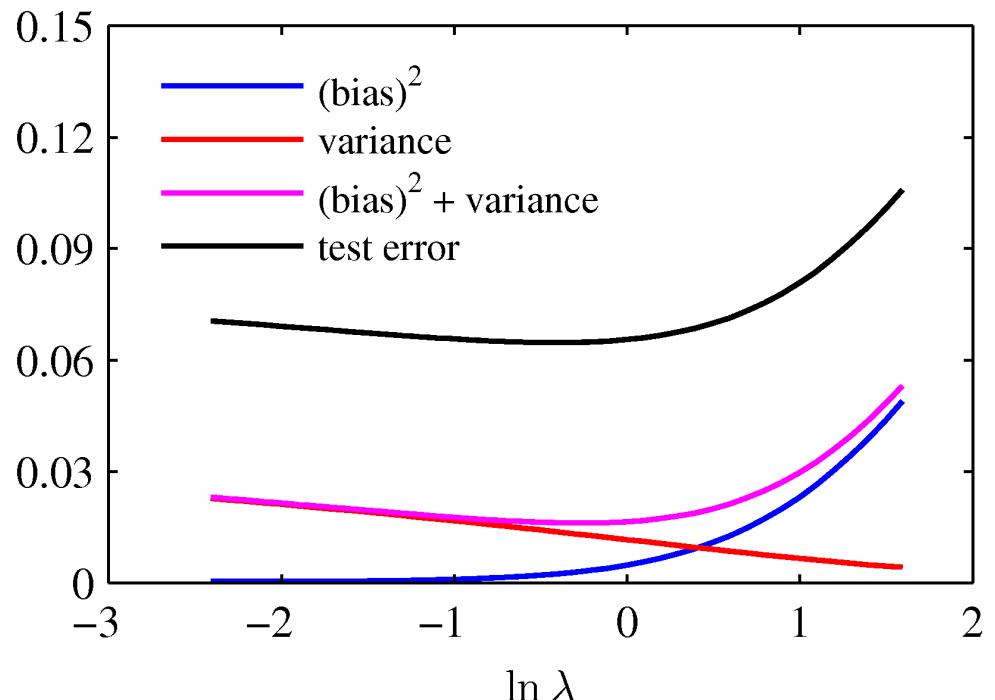
Пример: 25 наборов данных из синусоиды,  
различающихся степенью регуляризации,  $\lambda$ .



# Компромисс между смещением и дисперсией

---

Из этих графиков мы видим, что чрезмерно регуляризованная модель (большое  $\lambda$ ) будет иметь высокое смещение, тогда как недостаточно регуляризованная модель (маленькое  $\lambda$ ) будет иметь высокую дисперсию.



# Байесовская линейная регрессия (1)

---

Определим сопряженную априорную функцию по  $\mathbf{w}$

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

Объединяя это с функцией правдоподобия и используя результаты для маргинальных и условных гауссовых распределений, получаем апостериорную

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

где

$$\begin{aligned}\mathbf{m}_N &= \mathbf{S}_N \left( \mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \boldsymbol{\Phi}^T \mathbf{t} \right) \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi}.\end{aligned}$$

---

# Байесовская линейная регрессия (2)

---

Обычный выбор для априора — это

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

для которого

$$\begin{aligned}\mathbf{m}_N &= \beta \mathbf{S}_N \Phi^T \mathbf{t} \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \Phi^T \Phi.\end{aligned}$$

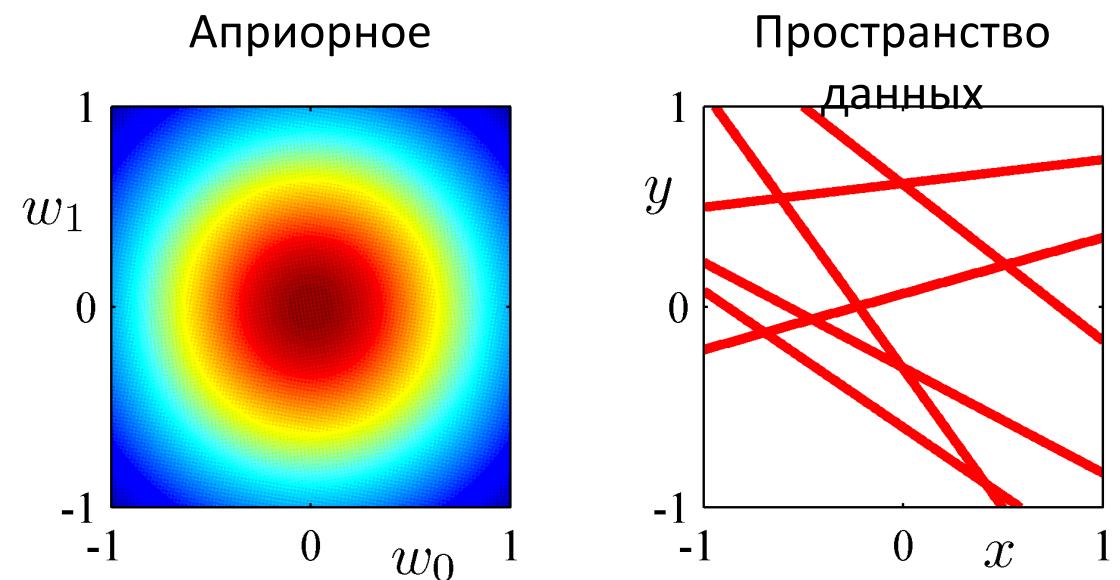
Далее рассмотрим пример...

---

# Байесовская линейная регрессия (3)

---

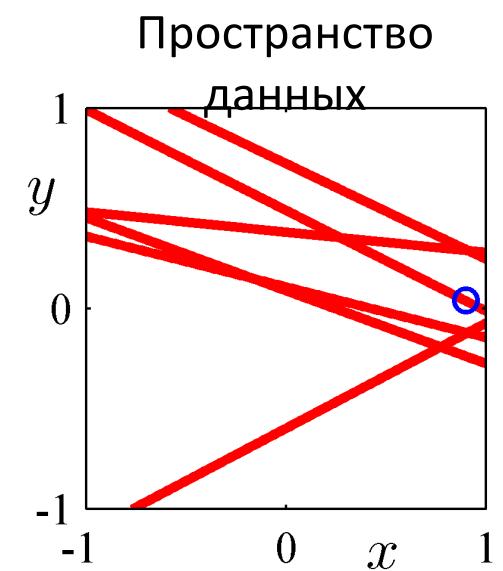
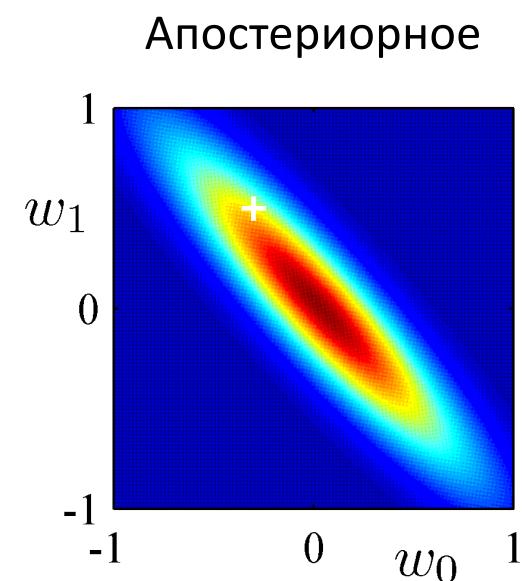
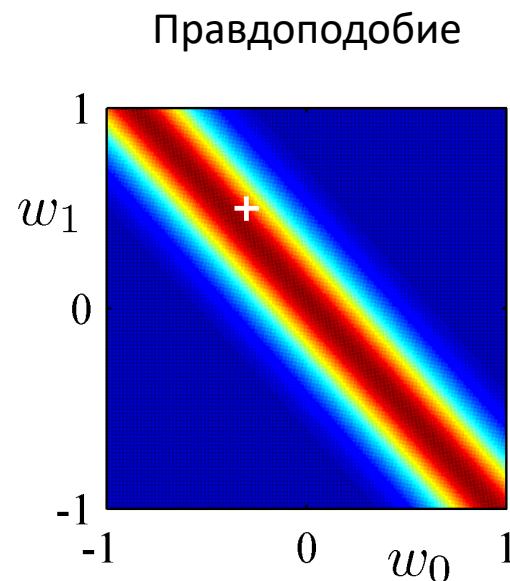
0 наблюдаемых точек данных



# Байесовская линейная регрессия (4)

---

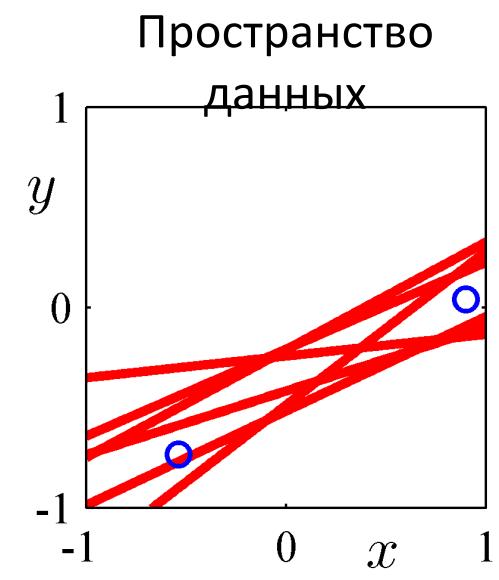
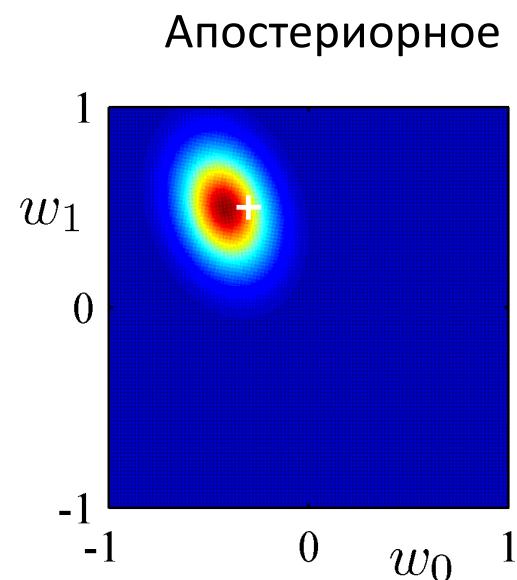
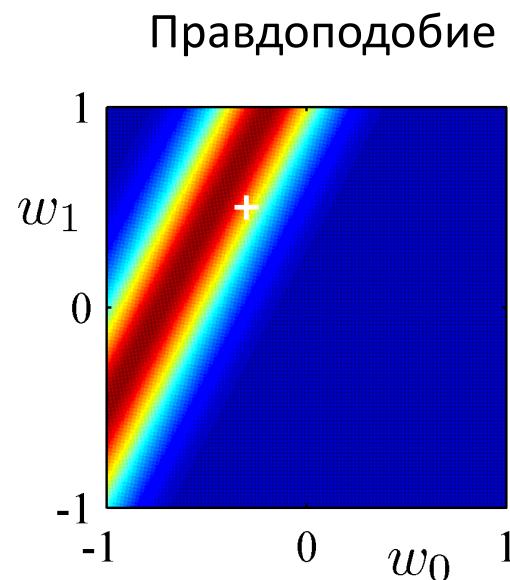
Наблюдается 1 точка данных



# Байесовская линейная регрессия (5)

---

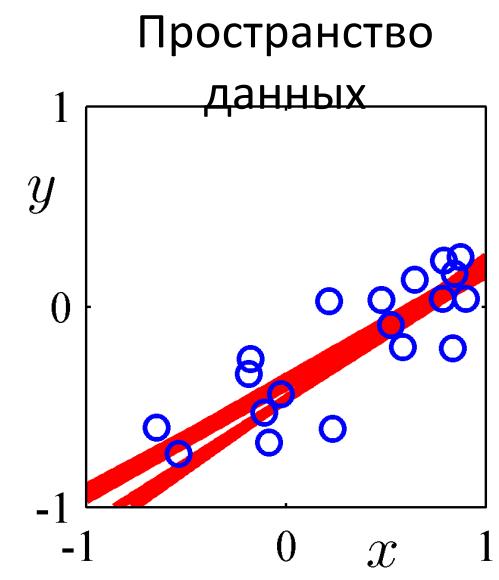
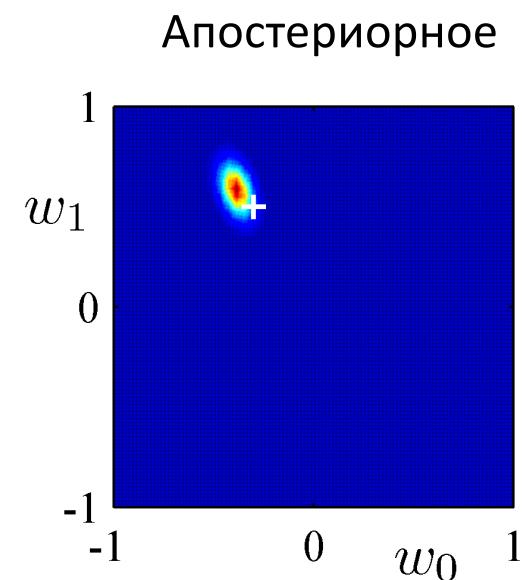
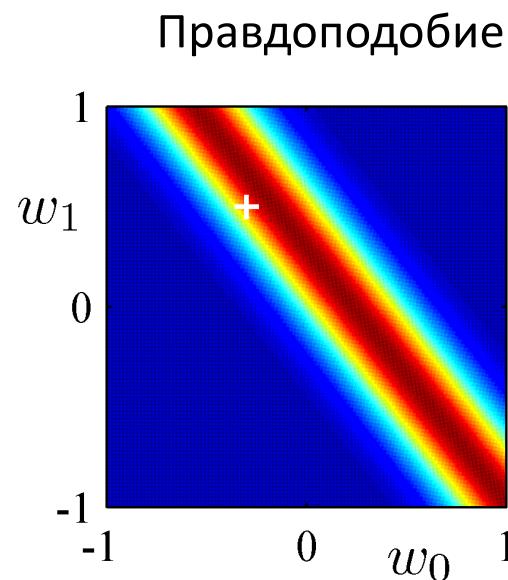
Наблюдалось 2 точки данных



# Байесовская линейная регрессия (6)

---

20 наблюдаемых точек данных



# Предиктивное распределение (1)

Предсказать  $t$  для новых значений  $\mathbf{x}$  путем интегрирования по  $\mathbf{w}$  :

$$\begin{aligned} p(t|\mathbf{t}, \alpha, \beta) &= \int p(t|\mathbf{w}, \beta)p(\mathbf{w}|\mathbf{t}, \alpha, \beta) d\mathbf{w} \\ &= \mathcal{N}(t|\mathbf{m}_N^T \boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x})) \end{aligned}$$

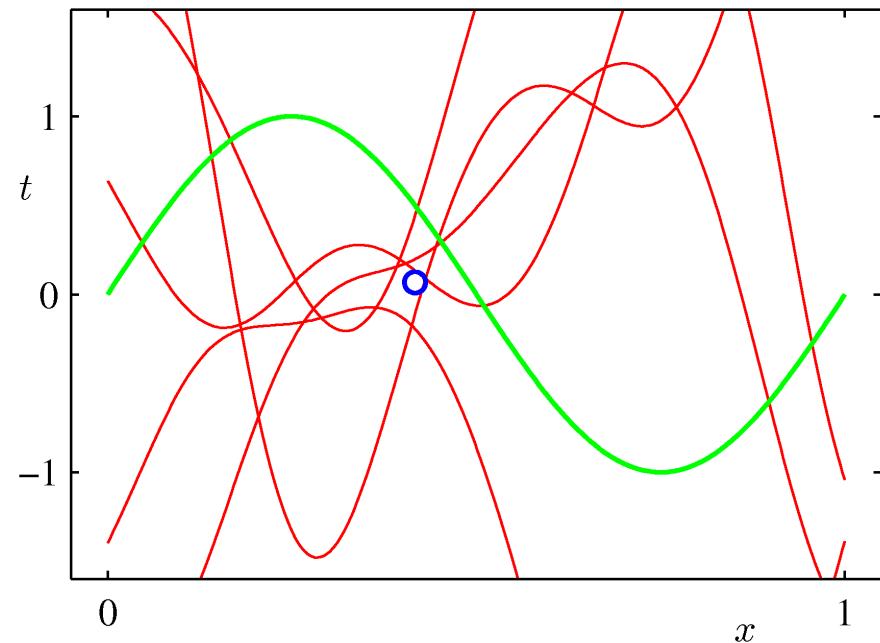
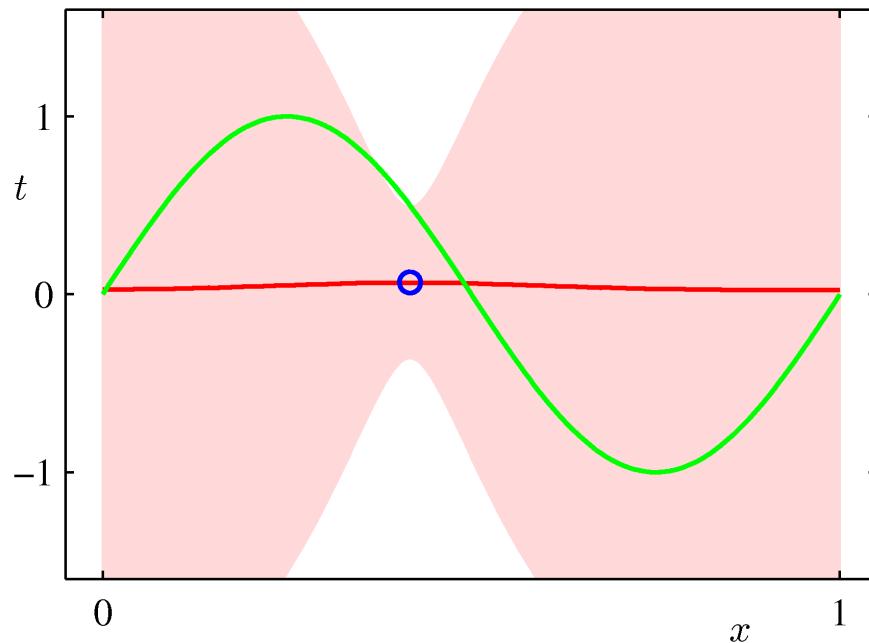
где

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^T \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}).$$

# Предиктивное распределение (2)

---

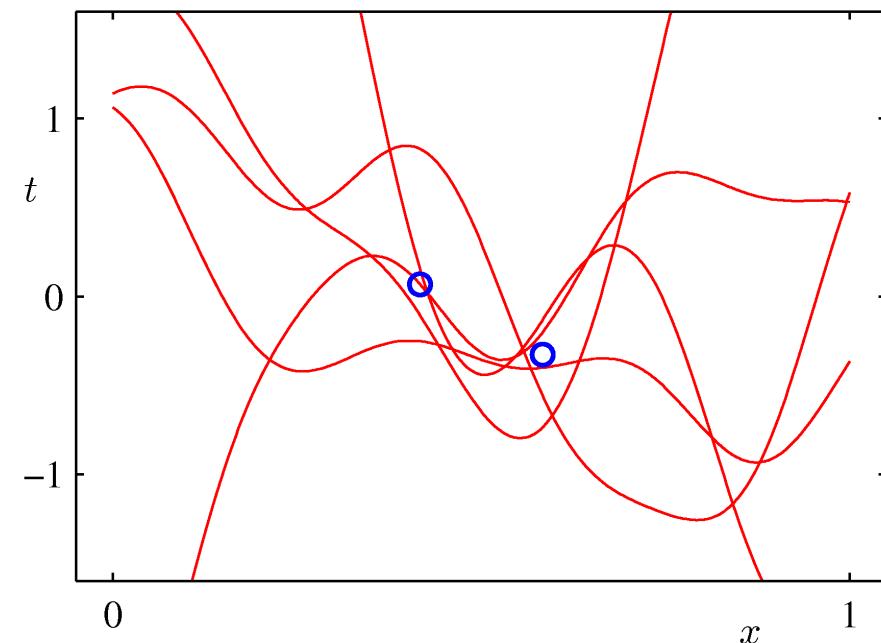
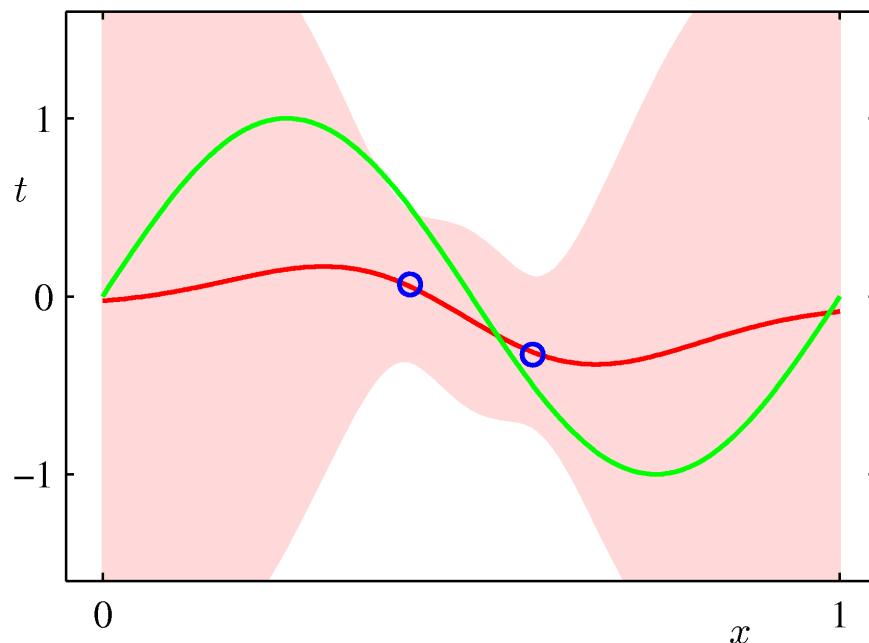
Пример: синусоидальные данные, 9 базисных функций Гаусса, 1 точка данных



# Предиктивное распределение (3)

---

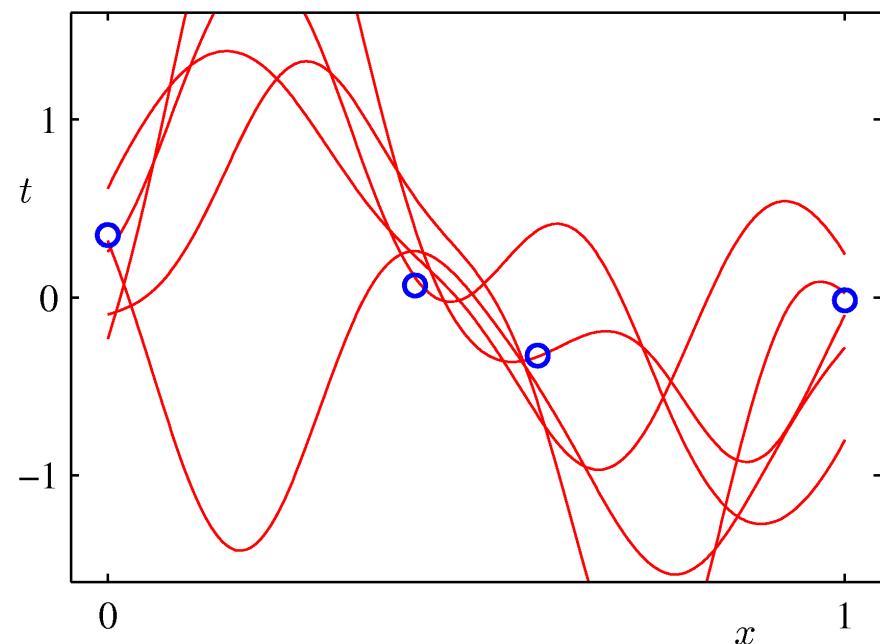
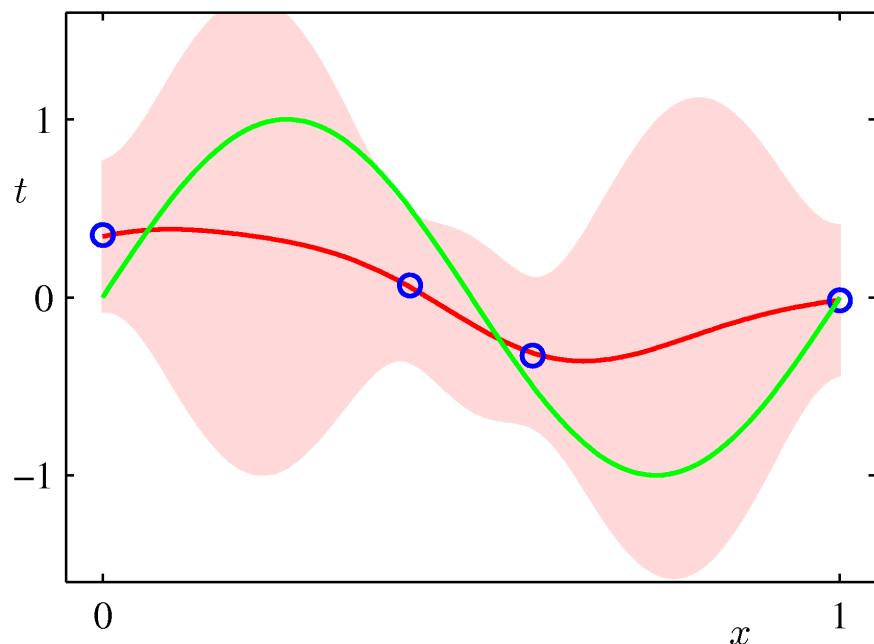
Пример: синусоидальные данные, 9 базисных функций Гаусса, 2 точки данных



# Предиктивное распределение (4)

---

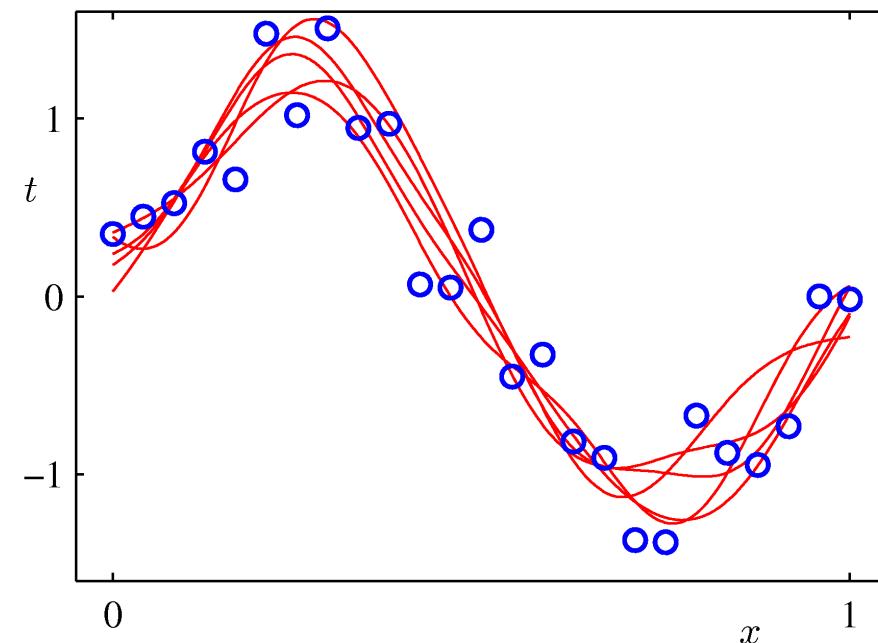
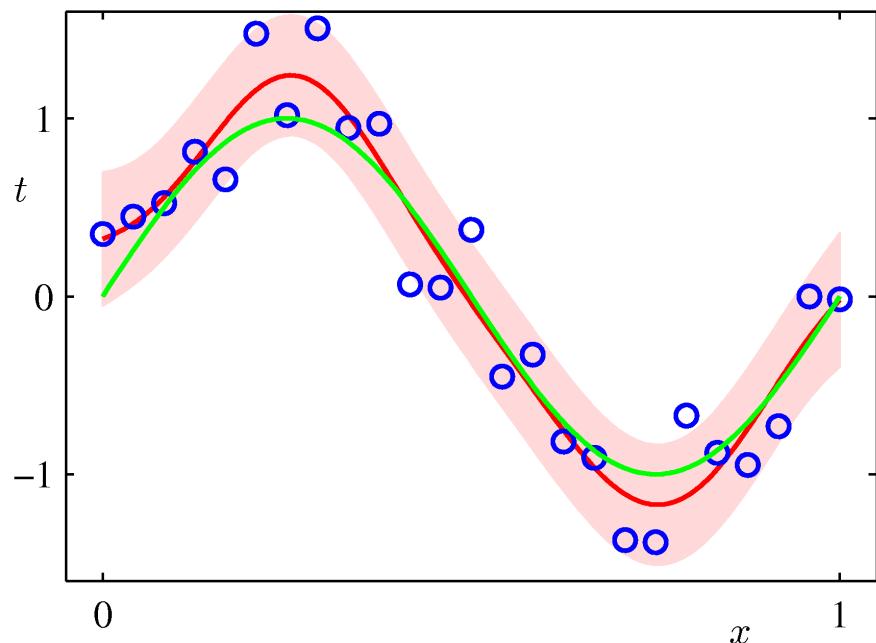
Пример: синусоидальные данные, 9 базисных функций Гаусса, 4 точки данных



# Предиктивное распределение (5)

---

Пример: синусоидальные данные, 9 базисных функций Гаусса, 25 точек данных



# Эквивалентное ядро (1)

---

Предсказательное среднее можно записать

$$\begin{aligned}y(\mathbf{x}, \mathbf{m}_N) &= \mathbf{m}_N^T \phi(\mathbf{x}) = \beta \phi(\mathbf{x})^T \mathbf{S}_N \Phi^T \mathbf{t} \\&= \sum_{n=1}^N \underbrace{\beta \phi(\mathbf{x})^T \mathbf{S}_N \phi(\mathbf{x}_n)}_{k(\mathbf{x}, \mathbf{x}_n)} t_n \\&= \sum_{n=1}^N k(\mathbf{x}, \mathbf{x}_n) t_n.\end{aligned}$$

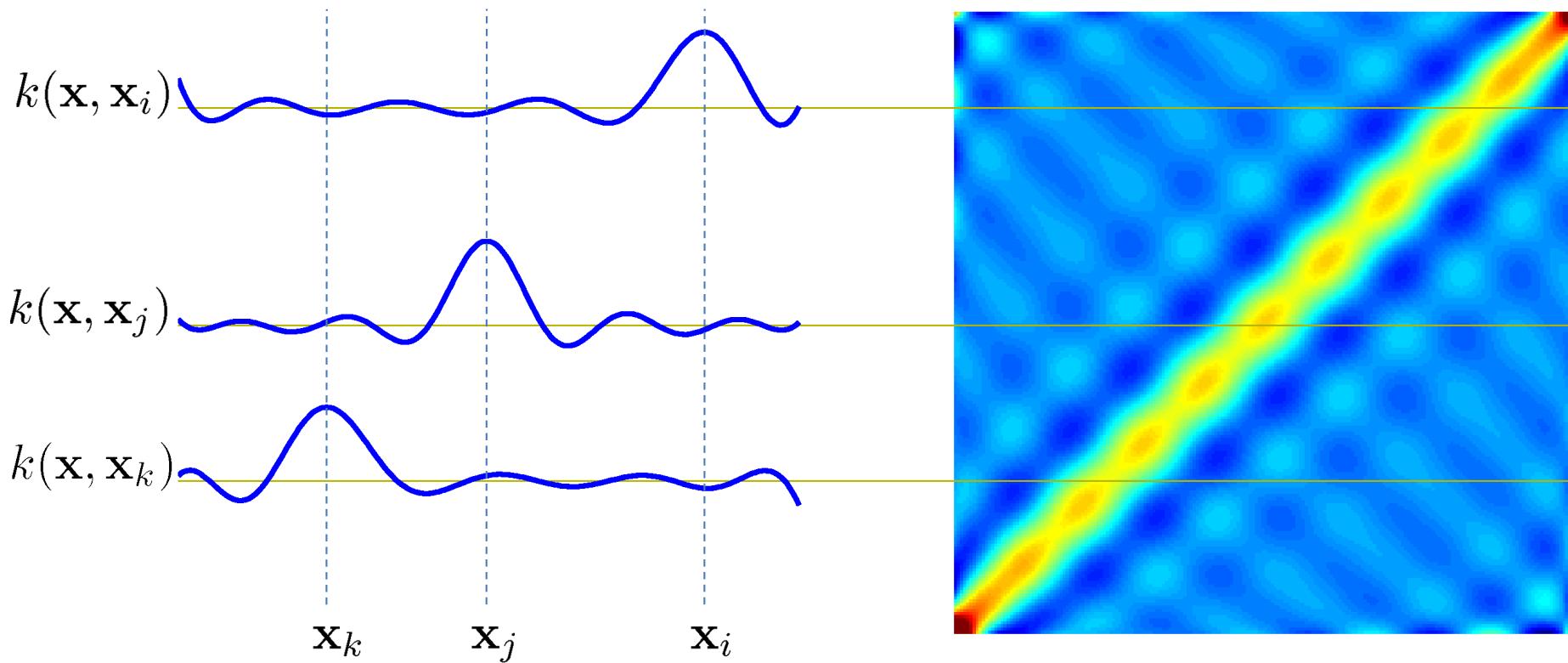
Эквивалентное  
ядро или более  
сглаженная  
матрица .

Это взвешенная сумма целевых значений  
обучающих данных,  $t_n$ .

---

## Эквивалентное ядро (2)

---



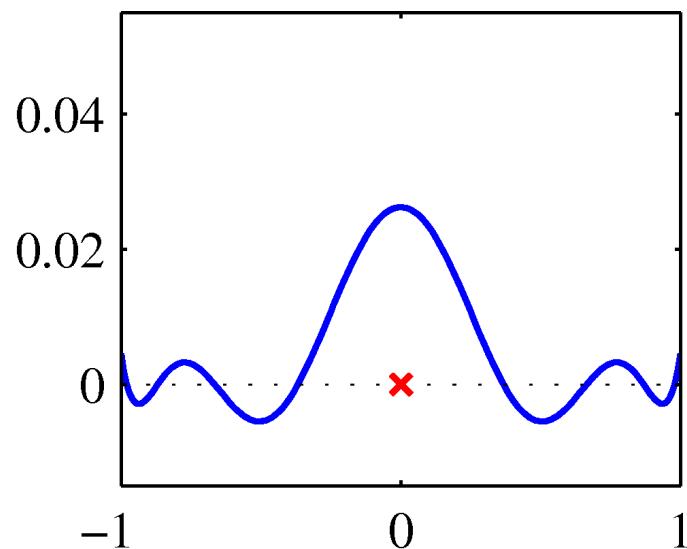
Вес  $t_n$  зависит от расстояния между  $\mathbf{x}$  и  $\mathbf{x}_n$ ;

близлежащие  $\mathbf{x}_n$  несут больший вес.

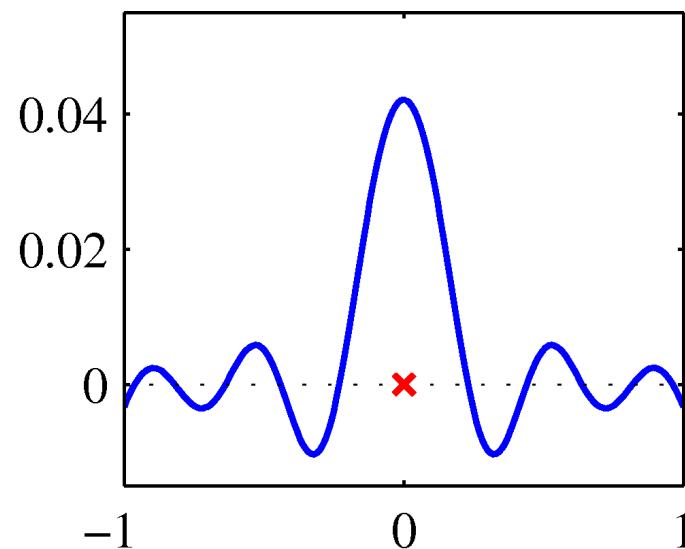
## Эквивалентное ядро (3)

---

Нелокальные базисные функции имеют локальные эквивалентные ядра:



Многочлен



Сигмоидальный

## Эквивалентное ядро (4)

---

Ядро как ковариационная функция:  
рассмотрим

$$\begin{aligned}\text{cov}[y(\mathbf{x}), y(\mathbf{x}')] &= \text{cov}[\phi(\mathbf{x})^T \mathbf{w}, \mathbf{w}^T \phi(\mathbf{x}')] \\ &= \phi(\mathbf{x})^T \mathbf{S}_N \phi(\mathbf{x}') = \beta^{-1} k(\mathbf{x}, \mathbf{x}').\end{aligned}$$

Мы можем избежать использования  
базисных функций и определить функцию  
ядра напрямую, что приведет к гауссовым  
процессам.

---

## Эквивалентное ядро (5)

---

$$\sum_{n=1}^N k(\mathbf{x}, \mathbf{x}_n) = 1$$

для всех значений  $\mathbf{x}$ ; однако эквивалентное ядро может быть отрицательным для некоторых значений  $\mathbf{x}$ .

Как и все функции ядра, эквивалентное ядро можно выразить в виде скалярного произведения:

$$k(\mathbf{x}, \mathbf{z}) = \psi(\mathbf{x})^T \psi(\mathbf{z})$$

где .  $\psi(\mathbf{x}) = \beta^{1/2} \mathbf{S}_N^{1/2} \phi(\mathbf{x})$

---

# Сравнение байесовских моделей (1)

---

Как выбрать «правильную» модель?

Предположим, мы хотим сравнить модели  $\mathcal{M}_i$ ,  
 $i=1, \dots, L$ , используя данные  $\mathcal{D}$ ; это требует  
вычислений

$$p(\mathcal{M}_i | \mathcal{D}) \propto p(\mathcal{M}_i) p(\mathcal{D} | \mathcal{M}_i).$$

Задний

Пре  
жни  
й

Модель  
доказательств  
или предельной  
вероятности

*Фактор Байеса* : соотношение доказательств для  
двух моделей

$$\frac{p(\mathcal{D} | \mathcal{M}_i)}{p(\mathcal{D} | \mathcal{M}_j)}$$

## Сравнение байесовских моделей (2)

---

Вычислив  $p(\mathcal{M}_i|\mathcal{D})$ , мы можем вычислить предсказательное (смешанное) распределение

$$p(t|\mathbf{x}, \mathcal{D}) = \sum_{i=1}^L p(t|\mathbf{x}, \mathcal{M}_i, \mathcal{D})p(\mathcal{M}_i|\mathcal{D}).$$

Более простое приближение, известное как *выбор модели*, заключается в использовании модели с наибольшим количеством доказательств.

---

# Сравнение байесовских моделей (3)

Для модели с параметрами  $\mathbf{w}$  мы получаем модельное доказательство путем маргинализации по  $\mathbf{w}$

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i)p(\mathbf{w}|\mathcal{M}_i) d\mathbf{w}.$$



Обратите внимание, что

$$p(\mathbf{w}|\mathcal{D}, \mathcal{M}_i) = \frac{p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i)p(\mathbf{w}|\mathcal{M}_i)}{p(\mathcal{D}|\mathcal{M}_i)}$$

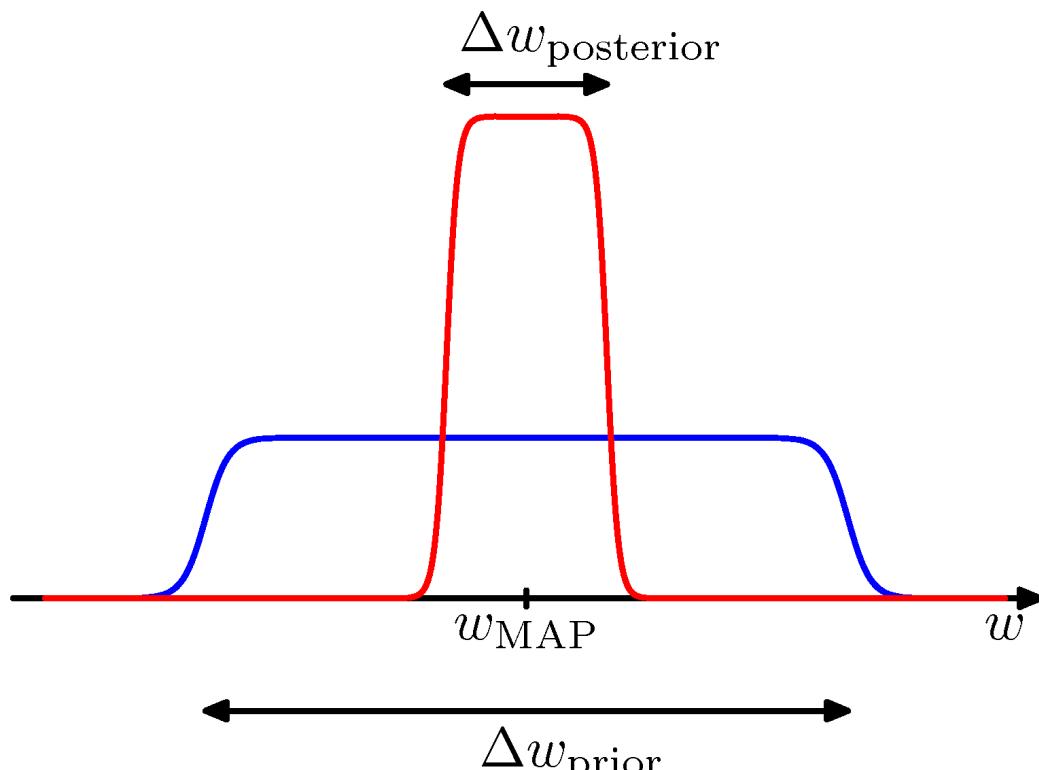


## Сравнение байесовских моделей (4)

Для данной модели с одним параметром  $w$  рассмотрим приближение

$$p(\mathcal{D}) = \int p(\mathcal{D}|w)p(w) dw$$
$$\simeq p(\mathcal{D}|w_{\text{MAP}}) \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}}$$

где предполагается, что задняя часть имеет острый пик.



# Сравнение байесовских моделей (5)

---

Логарифмируя, получаем

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|w_{\text{MAP}}) + \underbrace{\ln \left( \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \right)}_{\text{Отрицательно}}.$$

При  $M$  параметрах, предполагая, что все они имеют одинаковое соотношение, мы получаем

$$\Delta w_{\text{posterior}} / \Delta w_{\text{prior}}$$

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|w_{\text{MAP}}) + M \underbrace{\ln \left( \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \right)}_{\text{Отрицательный и линейный по } M}.$$

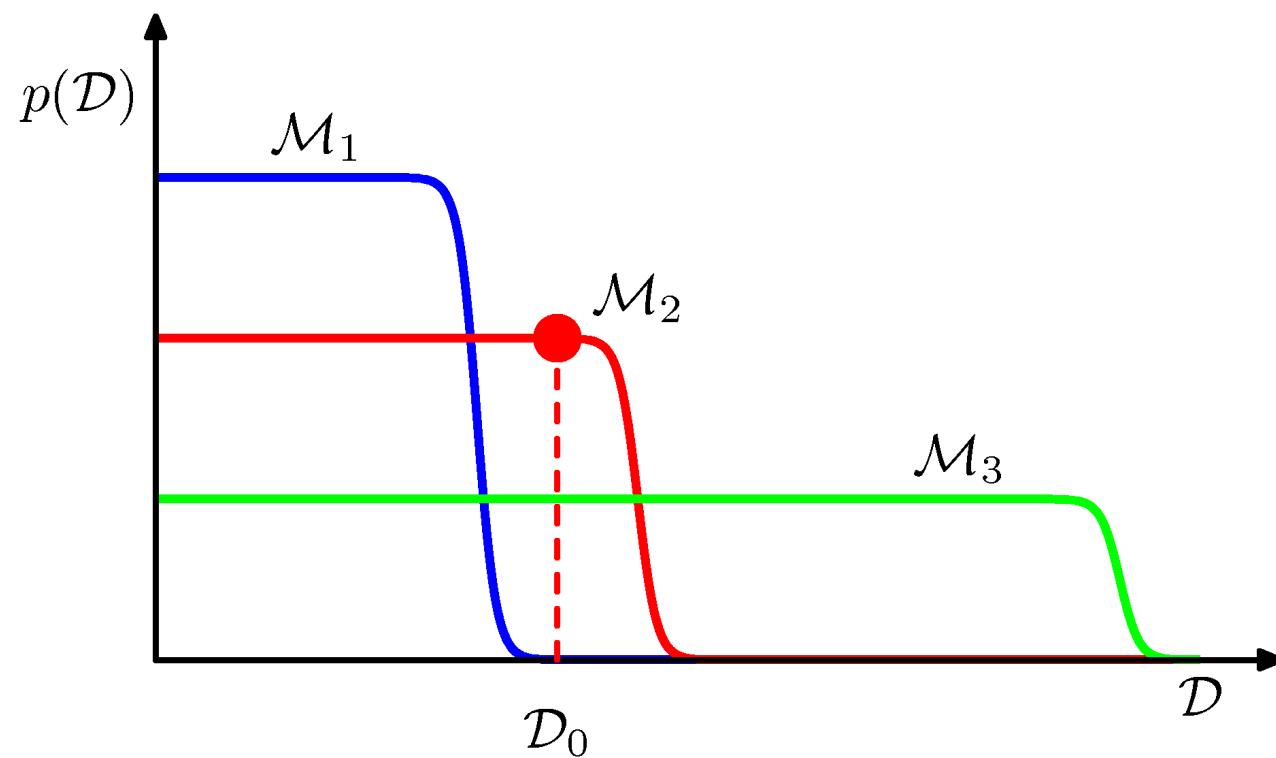
Отрицательный и линейный по  
 $M$

---

# Сравнение байесовских моделей (6)

---

Сопоставление данных и сложности модели



# Приближение доказательств (1)

---

Полностью байесовское предсказательное распределение задается формулой

$$p(t|\mathbf{t}) = \iiint p(t|\mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \alpha, \beta) p(\alpha, \beta|\mathbf{t}) d\mathbf{w} d\alpha d\beta$$

но этот интеграл труднорешаем. Приблизительно с

$$p(t|\mathbf{t}) \simeq p\left(t|\mathbf{t}, \hat{\alpha}, \hat{\beta}\right) = \int p\left(t|\mathbf{w}, \hat{\beta}\right) p\left(\mathbf{w}|\mathbf{t}, \hat{\alpha}, \hat{\beta}\right) d\mathbf{w}$$

где  $(\hat{\alpha}, \hat{\beta})$  — мода, которая, как предполагается, имеет острый пик; также известная как *эмпирический байесовский метод*,  $p(\alpha, \beta|\mathbf{t})$  *метод типа II* или *обобщенный метод максимального правдоподобия* или *приближение доказательств*.

---

## Приближение доказательств (2)

---

Из теоремы Байеса имеем

$$p(\alpha, \beta | \mathbf{t}) \propto p(\mathbf{t} | \alpha, \beta) p(\alpha, \beta)$$

и если мы предположим, что  $p(\alpha, \beta)$  плоский, мы

увидим, что

$$\begin{aligned} p(\alpha, \beta | \mathbf{t}) &\propto p(\mathbf{t} | \alpha, \beta) \\ &= \int p(\mathbf{t} | \mathbf{w}, \beta) p(\mathbf{w} | \alpha) d\mathbf{w}. \end{aligned}$$

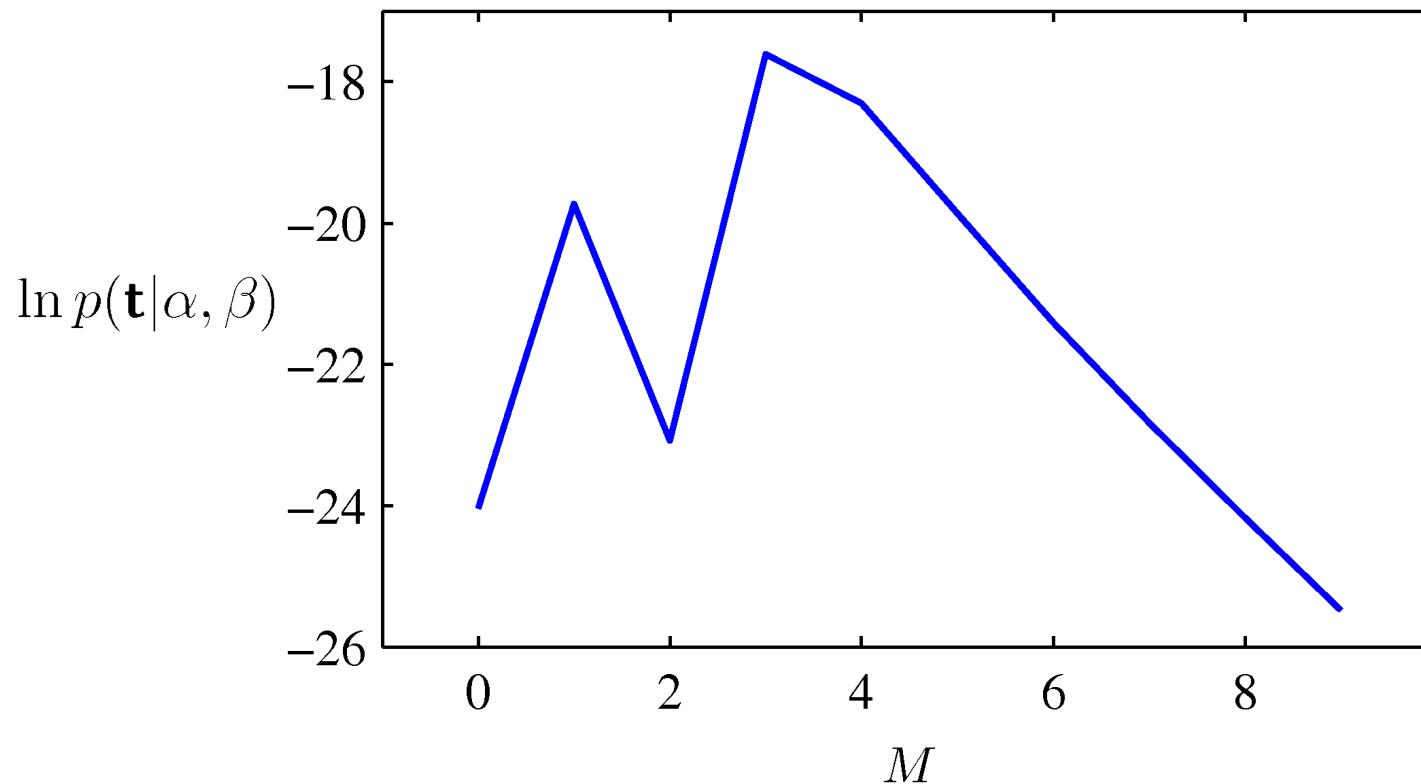
Общие результаты для гауссовых интегралов дают

$$\ln p(\mathbf{t} | \alpha, \beta) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - E(\mathbf{m}_N) + \frac{1}{2} \ln |\mathbf{S}_N| - \frac{N}{2} \ln(2\pi).$$

# Приближение доказательств (3)

---

Пример: синусоидальные данные, многочлен степени  $M$ ,  $\alpha = 5 \times 10^{-3}$



## Максимизация функции доказательства (1)

---

Чтобы максимизировать  $\ln p(\mathbf{t}|\alpha, \beta)$  по  $\alpha$  и  $\beta$ , мы определяем уравнение собственного вектора

$$\left( \beta \Phi^T \Phi \right) \mathbf{u}_i = \lambda_i \mathbf{u}_i.$$

Таким образом

$$\mathbf{A} = \mathbf{S}_N^{-1} = \alpha \mathbf{I} + \beta \Phi^T \Phi$$

имеет собственные значения  $\lambda_i + \alpha$ .

---

## Максимизация функции доказательства (2)

---

Теперь мы можем продифференцировать  $\ln p(\mathbf{t}|\alpha, \beta)$  по  $\alpha$  и  $\beta$  и приравнять результаты к нулю.

Получим

$$\alpha = \frac{\gamma}{\mathbf{m}_N^T \mathbf{m}_N}$$

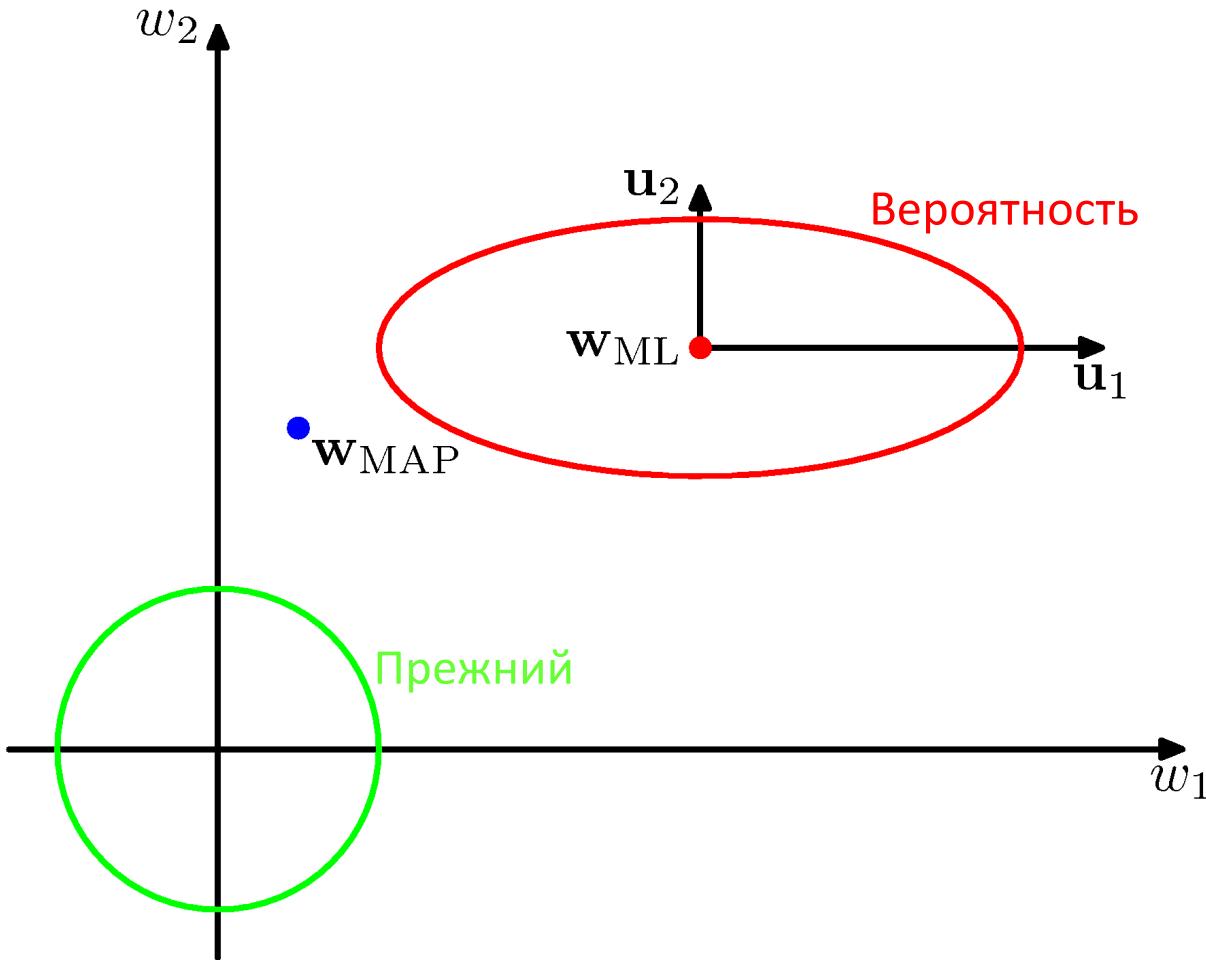
$$\frac{1}{\beta} = \frac{1}{N - \gamma} \sum_{n=1}^N \{t_n - \mathbf{m}_N^T \boldsymbol{\phi}(\mathbf{x}_n)\}^2$$

где

$$\gamma = \sum_i \frac{\lambda_i}{\alpha + \lambda_i}.$$

Примечание:  $\gamma$  зависит как от  $\alpha$  так и от  $\beta$ .

# Эффективное количество параметров (3)



$\lambda_1 \ll \alpha$   
 $w_1$  не очень хорошо  
определяется  
вероятностью

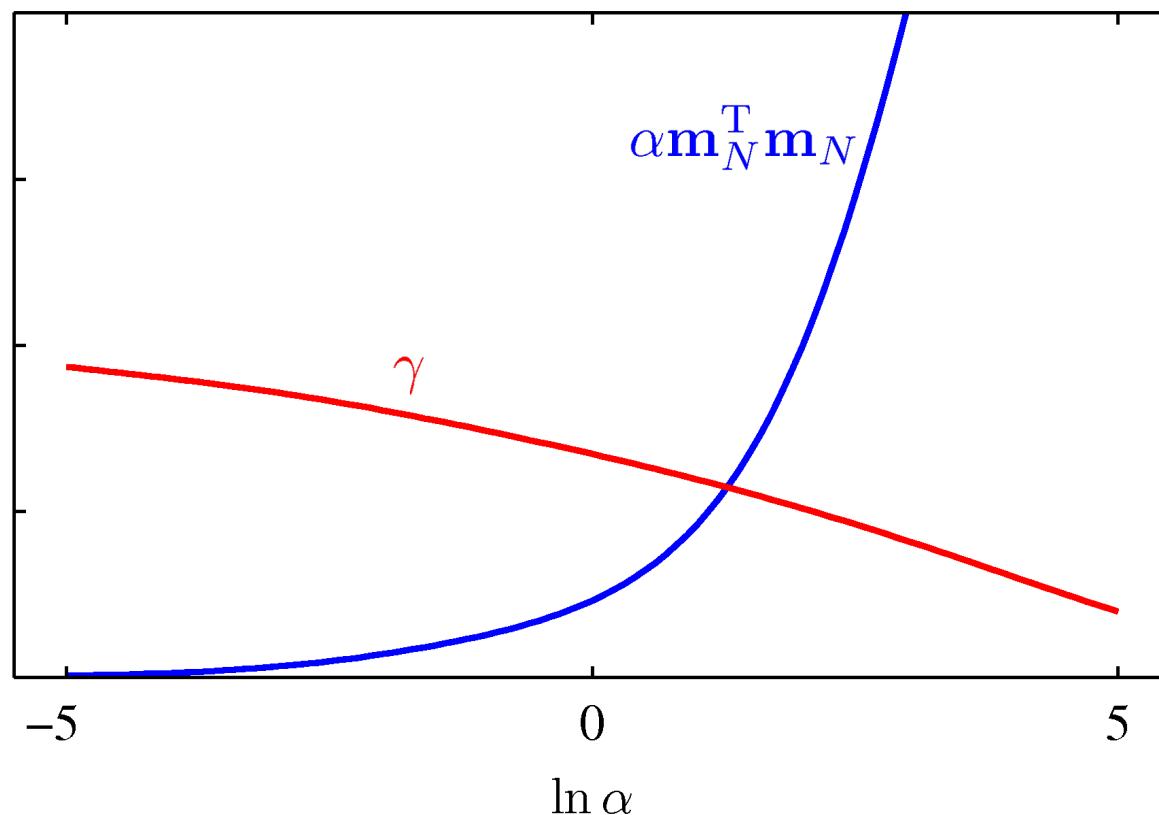
$\lambda_2 \gg \alpha$   
 $w_2$  хорошо  
определяется  
вероятностью

$\gamma$  — это количество  
хорошо определенных  
параметров

## Эффективное количество параметров (2)

---

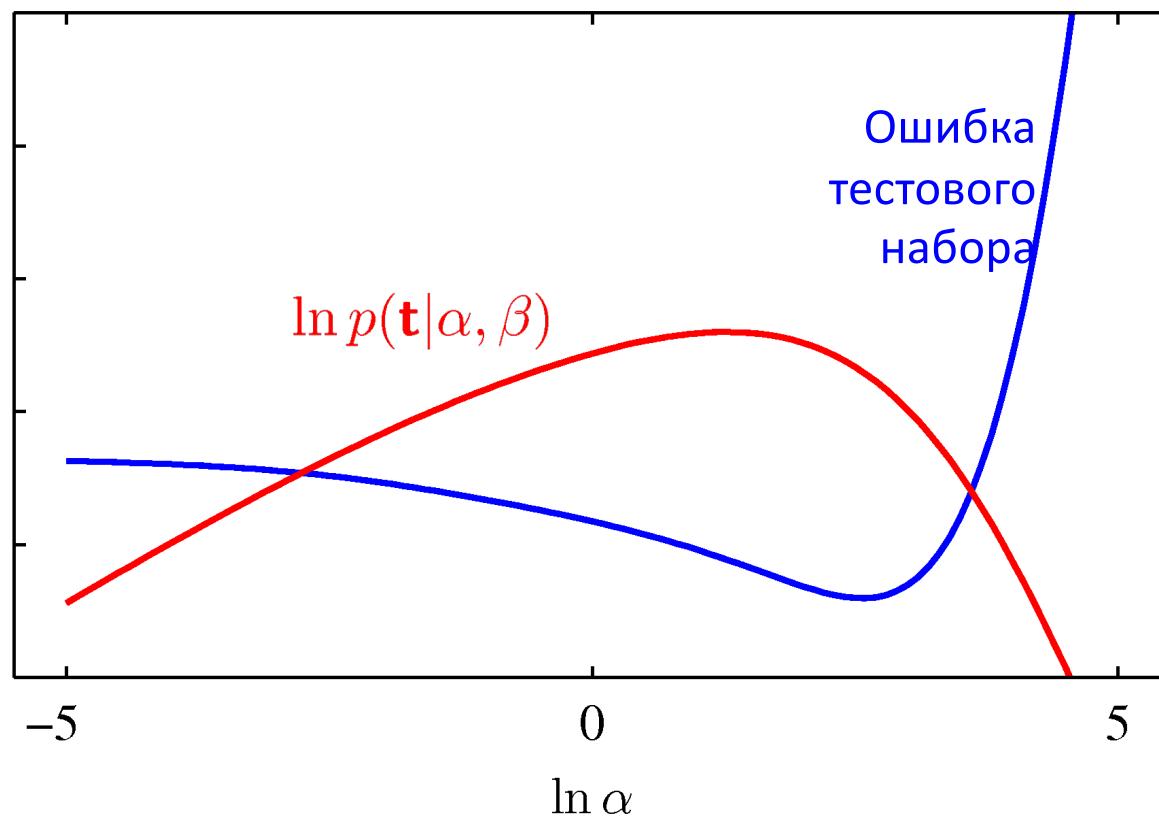
Пример: синусоидальные данные, 9 базисных функций Гаусса,  $\beta = 11.1$ .



## Эффективное количество параметров (3)

---

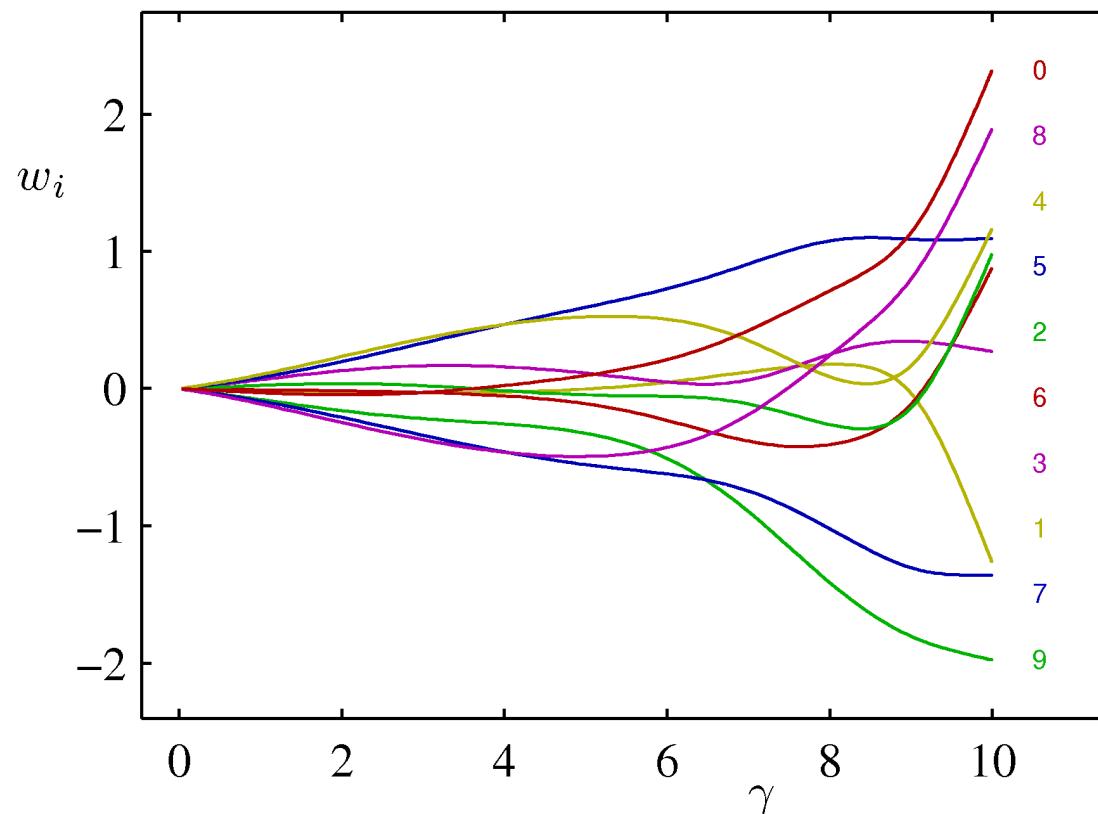
Пример: синусоидальные данные, 9 базисных функций Гаусса,  $\beta = 11.1$ .



# Эффективное количество параметров (4)

---

Пример: синусоидальные данные, 9 базисных функций Гаусса,  $\beta = 11.1$ .



## Эффективное количество параметров (5)

---

В пределе  $\gamma=M$   $N \gg M$  и мы можем  
рассмотреть возможность использования легко  
вычисляемого приближения

$$\begin{aligned}\alpha &= \frac{M}{\mathbf{m}_N^T \mathbf{m}_N} \\ \frac{1}{\beta} &= \frac{1}{N} \sum_{n=1}^N \{t_n - \mathbf{m}_N^T \boldsymbol{\phi}(\mathbf{x}_n)\}^2.\end{aligned}$$

## Ограничения функций фиксированного базиса

---

- $M$  по каждому измерению  $D$ -мерного входного пространства требует базисных функций  $M^D$  - проклятие размерности.
- Далее мы увидим, как можно обойтись меньшим количеством базисных функций, выбирая их с использованием обучающих данных.