Линейная регрессия

Сергей Николенко

НИУ ВШЭ — Санкт-Петербург

15 февраля 2020 г.

Random facts:

- 15 февраля на Вануату — День Джона Фрума, центрального персонажа в культе карго, изображаемого в виде американского военнослужащего Второй мировой войны; возможно, «Джон Фрум» — это искажение «John from (America)»; последователи Джона Фрума соорудили на острове Танна символическую взлётно-посадочную полосу с макетами самолётов и символическими вышками, на которых дежурили «диспетчеры»
- 15 февраля в США — День Сьюзен Б. Энтони (Susan B. Anthony), одной из главных суфражисток XIX века; в 1863 году они с Элизабет Стэнтон организовали Women's Loyal National League, которая собрала 400 тысяч подписей в поддержку отмены рабства, а затем лоббировала равноправие всех граждан США — и женщин, и афроамериканцев (но не вылоббировала, в 14-й и 15-й поправках про женщин ничего нет)
- 15 февраля 1717 г. в Петербурге было впервые напечатано «Юности честное зерцало»
- 15 февраля 2013 г. астероид диаметром $\approx$ 17 метров и массой $\approx$ 10000 тонн вошёл в атмосферу Земли на скорости $\approx$ 18 км/с и разрушился на высоте 15—25 км в окрестностях Челябинска; снятое на видеорегистратор падение было показано в начальных титрах фильмов World War Z и Edge of Tomorrow

Сопряжённые априорные распределения

Напоминание

- Напоминаю, что основная наша задача – как обучить параметры распределения и/или предсказать следующие его точки по имеющимся данным.
- В байесовском выводе участвуют:
  - $p(x \mid \theta)$ – правдоподобие данных;
  - $p(\theta)$ – априорное распределение;
  - $p(x) = \int_\Theta p(x \mid \theta) p(\theta) d\theta$ – маргинальное правдоподобие;
  - $p(\theta \mid x) = \frac{p(x|\theta)p(\theta)}{p(x)}$ – апостериорное распределение;
  - $p(x' \mid x) = \int_\Theta p(x' \mid \theta) p(\theta \mid x) d\theta$ – предсказание нового x'.
- Задача обычно в том, чтобы найти $p(\theta \mid x)$ и/или $p(x' \mid x)$.

Априорные распределения

- Когда мы проводим байесовский вывод, у нас, кроме правдоподобия, должно быть ещё *априорное распределение* (prior distribution) по всем возможным значениям параметров.
- Мы раньше к ним специально не присматривались, но они очень важны.
- Задача байесовского вывода – как подсчитать $p(\theta \mid x)$ и/или $p(x' \mid x)$.
- Но чтобы это сделать, сначала надо выбрать $p(\theta)$. Как выбирать априорные распределения?

Сопряжённые априорные распределения

- Разумная цель: давайте будем выбирать распределения так, чтобы они оставались такими же и *a posteriori*.
- До начала вывода есть априорное распределение $p(\theta)$.
- После него есть какое-то новое апостериорное распределение $p(\theta \mid x)$.
- Я хочу, чтобы $p(\theta \mid x)$ тоже имело тот же вид, что и $p(\theta)$, просто с другими параметрами.

Сопряжённые априорные распределения

- Не слишком формальное определение: семейство распределений $p(\theta \mid \alpha)$ называется семейством *сопряжённых априорных распределений* для семейства правдоподобий $p(x \mid \theta)$, если после умножения на правдоподобие апостериорное распределение $p(\theta \mid x, \alpha)$ остаётся в том же семействе: $p(\theta \mid x, \alpha) = p(\theta \mid \alpha')$.
- α называются *гиперпараметрами* (hyperparameters), это «параметры распределения параметров».
- Тривиальный пример: семейство всех распределений будет сопряжённым чему угодно, но это не очень интересно.

Сопряжённые априорные распределения

- Разумеется, вид хорошего априорного распределения зависит от вида распределения собственно данных, $p(x \mid \theta)$.
- Сопряжённые априорные распределения подсчитаны для многих распределений, мы приведём несколько примеров.

Испытания Бернулли

- Каким будет сопряжённое априорное распределение для бросания нечестной монетки (испытаний Бернулли)?
- Ответ: это будет бета-распределение; плотность распределения нечестности монетки θ

$$p(\theta \mid \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}.$$

Испытания Бернулли

- Плотность распределения нечестности монетки θ

$$p(\theta \mid \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}.$$

- Тогда, если мы посэмплируем монетку, получив s орлов и f решек, получится

$$p(s, f \mid \theta) = \binom{s+f}{s} \theta^s (1-\theta)^f, \text{ и}$$

$$p(\theta|s,f) = \frac{\binom{s+f}{s}\theta^{s+\alpha-1}(1-\theta)^{f+\beta-1}/B(\alpha,\beta)}{\int_0^1 \binom{s+f}{s}x^{s+\alpha-1}(1-x)^{f+\beta-1}/B(\alpha,\beta)dx} =$$

$$= \frac{\theta^{s+\alpha-1}(1-\theta)^{f+\beta-1}}{B(s+\alpha, f+\beta)}.$$

Испытания Бернулли

- Итого получается, что сопряжённое априорное распределение для параметра нечестной монетки θ – это

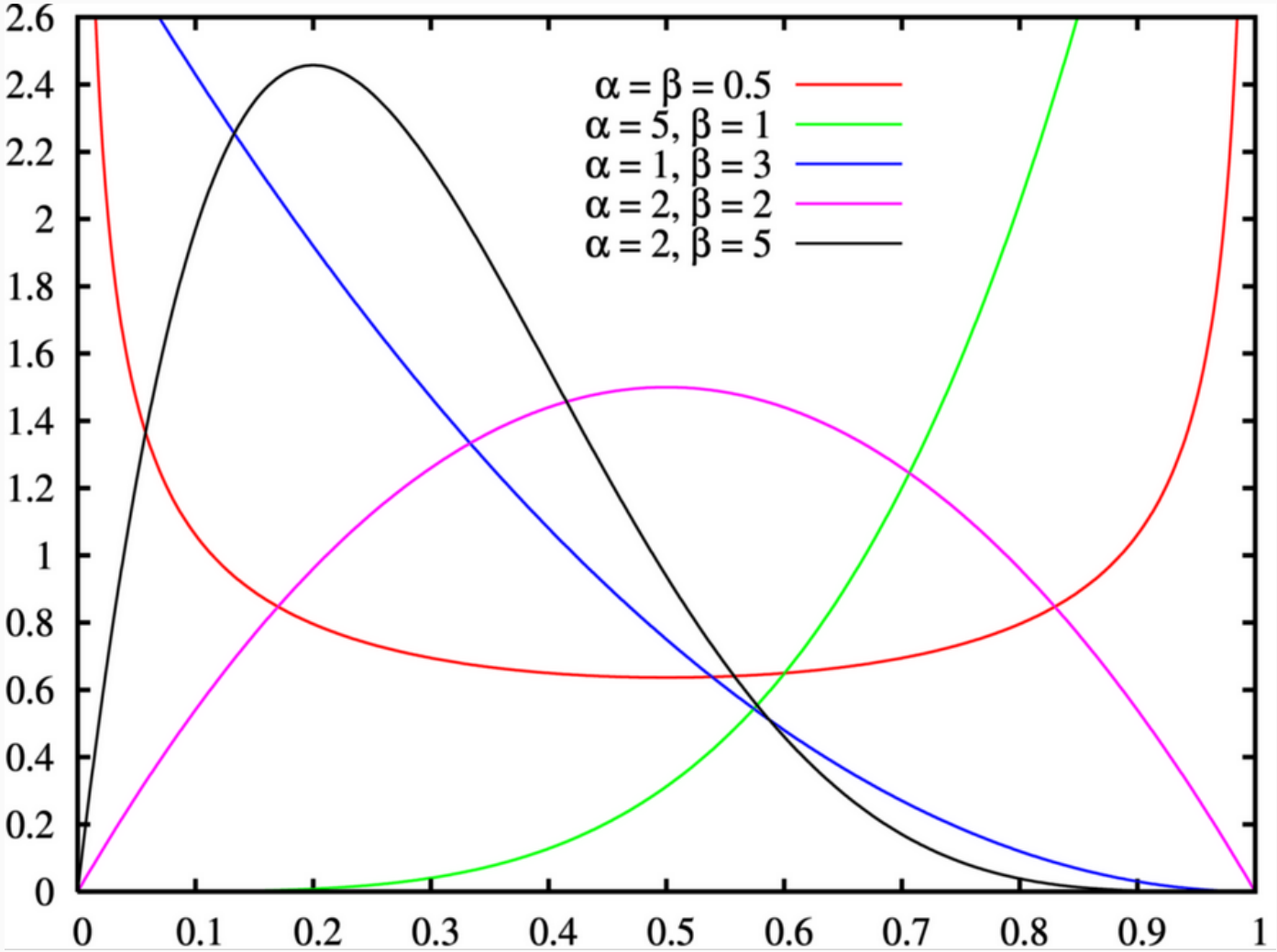
$$p(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

- После получения новых данных с s орлами и f решками гиперпараметры меняются на

$$p(\theta \mid s+\alpha, f+\beta) \propto \theta^{s+\alpha-1}(1-\theta)^{f+\beta-1}.$$

- На этом этапе можно забыть про сложные формулы и выводы, получилось очень простое правило обучения (под обучением теперь понимается изменение гиперпараметров).

Бета-распределение



Мультиномиальное распределение

- Простое обобщение: рассмотрим мультиномиальное распределение с n испытаниями, k категориями и по x_i экспериментов дали категорию i.
- Параметры θ_i показывают вероятность попасть в категорию i:

$$p(x \mid \theta) = \binom{n}{x_1, \ldots, x_n} \theta_1^{x_1} \theta_2^{x_2} \ldots \theta_k^{x_k}.$$

- Сопряжённым априорным распределением будет распределение Дирихле:

$$p(\theta \mid \alpha) \propto \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \ldots \theta_k^{\alpha_k - 1}.$$

Мультиномиальное распределение

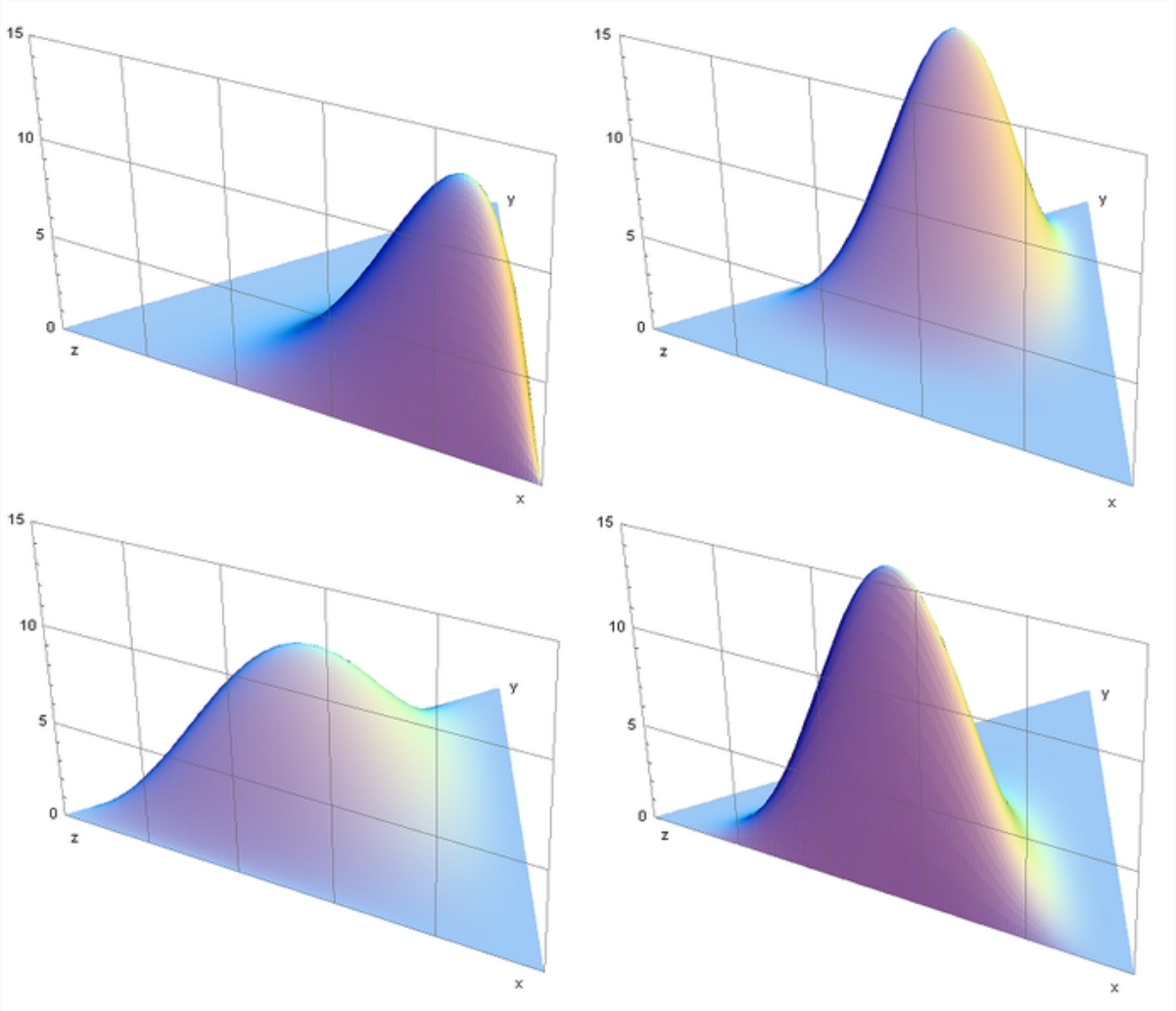
- Сопряжённым априорным распределением будет распределение Дирихле:

$$p(\theta \mid \alpha) \propto \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \dots \theta_k^{\alpha_k - 1}.$$

Упражнение. Докажите, что при получении данных $x_1, \dots, x_k$ гиперпараметры изменятся на

$$p(\theta \mid x, \alpha) = p(\theta \mid x + \alpha) \propto \theta_1^{x_1 + \alpha_1 - 1} \theta_2^{x_2 + \alpha_2 - 1} \dots \theta_k^{x_k + \alpha_k - 1}.$$

Распределение Дирихле



Линейная регрессия

Метод наименьших квадратов

- Линейная регрессия: рассмотрим линейную функцию

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{p} x_j w_j = \mathbf{x}^\top \mathbf{w}, \quad \mathbf{x} = (1, x_1, \ldots, x_p).$$

- Таким образом, по вектору входов $\mathbf{x}^\top = (x_1, \ldots, x_p)$ мы будем предсказывать выход y как

$$\hat{y}(\mathbf{x}) = \hat{w}_0 + \sum_{j=1}^{p} x_j \hat{w}_j = \mathbf{x}^\top \hat{\mathbf{w}}.$$

Метод наименьших квадратов

- Как найти оптимальные параметры $\hat{\mathbf{w}}$ по тренировочным данным вида $(\mathbf{x}_i, y_i)_{i=1}^N$?
- Метод наименьших квадратов: будем минимизировать

$$\mathrm{RSS}(\mathbf{w}) = \sum_{i=1}^{N} (y_i - \mathbf{x}_i^\top \mathbf{w})^2.$$

- Как минимизировать?

Метод наименьших квадратов

- Можно на самом деле решить задачу точно – записать как

$$\mathrm{RSS}(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w}),$$

  где $\mathbf{X}$ – матрица $N \times p$, продифференцировать по $\mathbf{w}$, получится

$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y},$$

  если матрица $\mathbf{X}^\top \mathbf{X}$ невырожденная.
- Замечание: $\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top$ называется *псевдообратной матрицей Мура–Пенроуза* (Moore–Penrose pseudo-inverse) матрицы $\mathbf{X}$; это обобщение понятия обратной матрицы на неквадратные матрицы.
- Много ли нужно точек, чтобы обучить такую модель?

Байесовская регрессия

- Теперь давайте поговорим о линейной регрессии по-байесовски.
- Основное наше предположение – в том, что шум (ошибка в данных) распределён нормально, т.е. переменная t, которую мы наблюдаем, получается как

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

  Иными словами,

$$p(t \mid \mathbf{x}, \mathbf{w}, \sigma^2) = \mathcal{N}(t \mid y(\mathbf{x}, \mathbf{w}), \sigma^2).$$

- Здесь пока y – любая функция.

Байесовская регрессия

- Чтобы не повторять совсем уж то же самое, мы рассмотрим не в точности линейную регрессию, а её естественное обобщение – линейную модель с базисными функциями:

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

(M параметров, $M-1$ базисная функция, $\phi_0(\mathbf{x}) = 1$).

Байесовская регрессия

- Базисные функции ϕ_i – это, например:
  - результат feature extraction;
  - расширение линейной модели на нелинейные зависимости (например, $\phi_j(x) = x^j$);
  - локальные функции, которые существенно не равны нулю только в небольшой области (например, гауссовские базисные функции $\phi_j(\mathbf{x}) = e^{-\frac{(x-\mu_j)^2}{2s^2}}$);
  - ...

Байесовская регрессия

- Рассмотрим набор данных $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ со значениями $\mathbf{t} = \{t_1, \ldots, t_N\}$.
- Будем предполагать, что данные взяты независимо по одному и тому же распределению:

$$p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(t_n \mid \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n), \sigma^2\right).$$

- Прологарифмируем (опустим $\mathbf{X}$, т.к. по нему всегда условная вероятность будет):

$$\ln p(\mathbf{t} \mid \mathbf{w}, \sigma^2) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} \left(t_n - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n)\right)^2.$$

Байесовская регрессия

- Прологарифмируем (опустим $\mathbf{X}$, т.к. по нему всегда условная вероятность будет):

$$\ln p(\mathbf{t} \mid \mathbf{w}, \sigma^2) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} \left(t_n - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n)\right)^2.$$

- И вот мы получили, что для максимизации правдоподобия по $\mathbf{w}$ нам нужно как раз минимизировать среднеквадратичную ошибку!

$$\nabla_{\mathbf{w}} \ln p(\mathbf{t} \mid \mathbf{w}, \sigma^2) = \frac{1}{\sigma^2} \sum_{n=1}^{N} \left(t_n - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n)\right) \boldsymbol{\phi}(\mathbf{x}_n).$$

Байесовская регрессия

- Решая систему уравнений $\nabla \ln p(\mathbf{t} \mid \mathbf{w}, \sigma^2) = 0$, получаем то же самое, что и раньше:

$$\mathbf{w}_{ML} = \left(\mathbf{\Phi}^\top \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^\top \mathbf{t}.$$

- Здесь $\mathbf{\Phi} = (\phi_j(\mathbf{x}_i))_{i,j}$.

Байесовская регрессия

- Теперь можно и относительно σ^2 максимизировать правдоподобие; получим

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}\left(t_n - \mathbf{w}_{ML}^\top \boldsymbol{\phi}(\mathbf{x}_n)\right)^2,$$

  т.е. как раз выборочная дисперсия имеющихся данных вокруг предсказанного значения.

Оверфиттинг
в линейной регрессии

Полиномиальная аппроксимация

- Мы говорили о регрессии с базисными функциями:

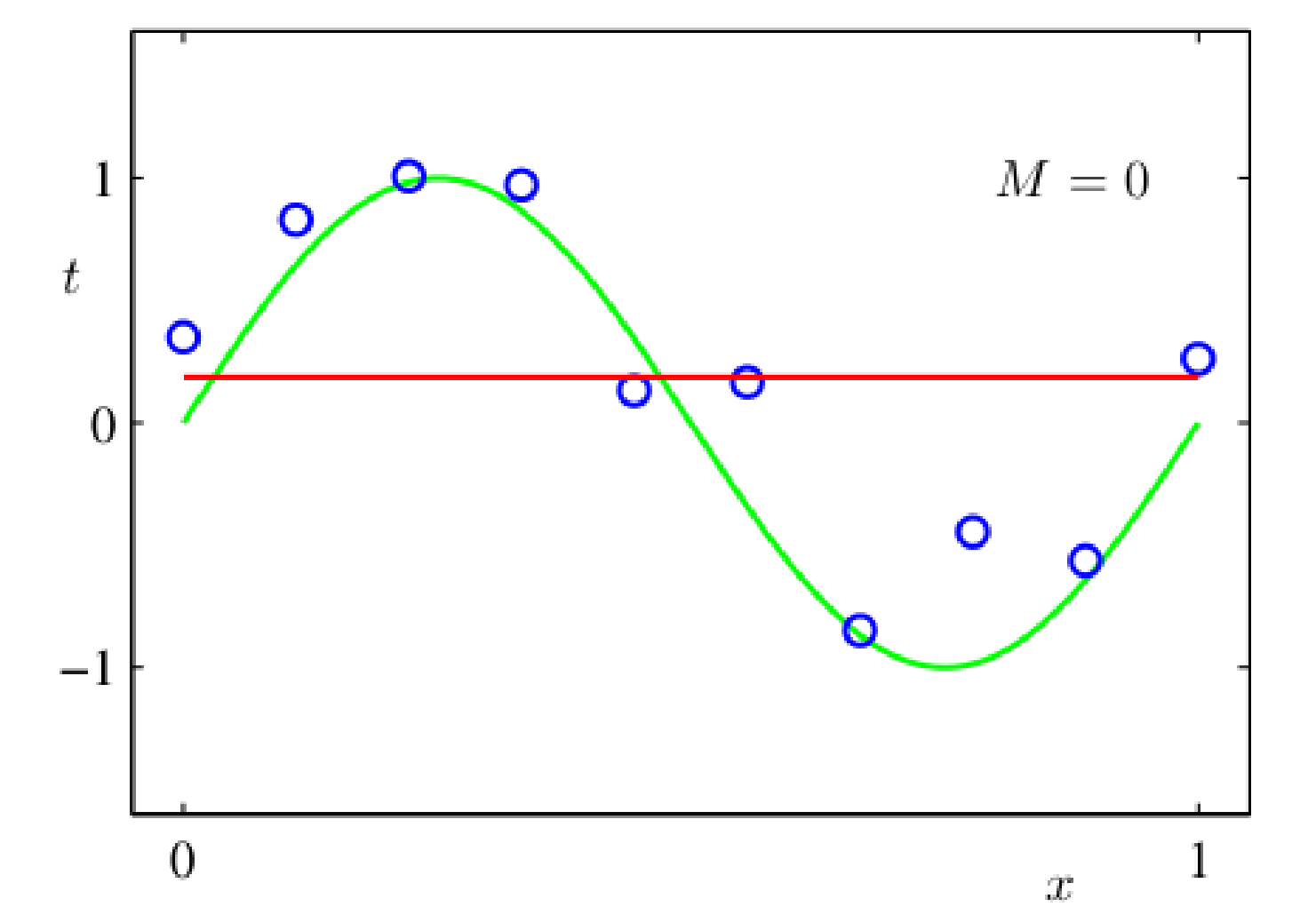
$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M} w_j \phi_j(\mathbf{x}) = \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}).$$

- Давайте для примера рассмотрим такую регрессию для $\phi_j(x) = x^j$, т.е.

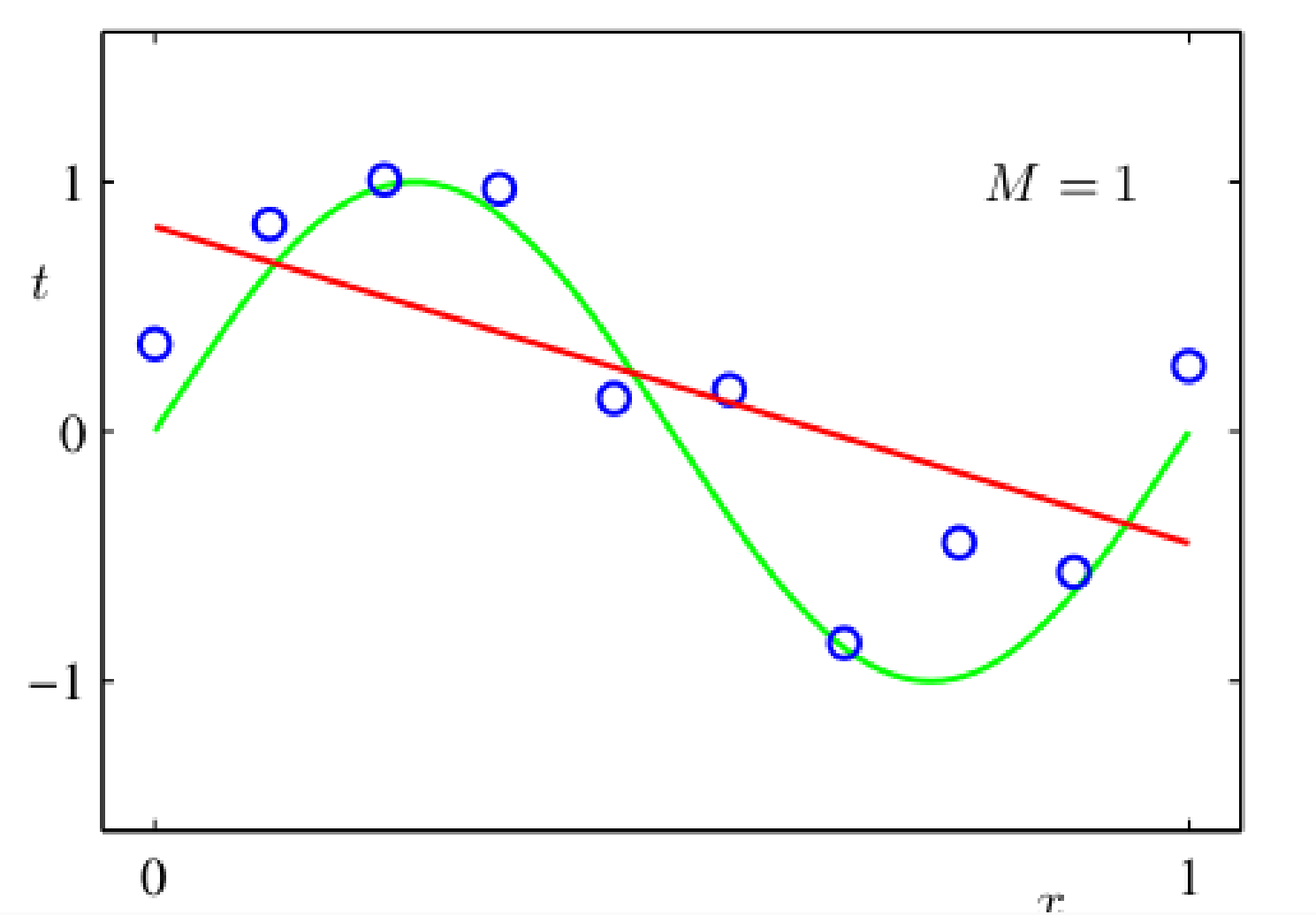
$$f(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M.$$

- И будем, как раньше, минимизировать квадратичную ошибку.
- Пример с кодом.

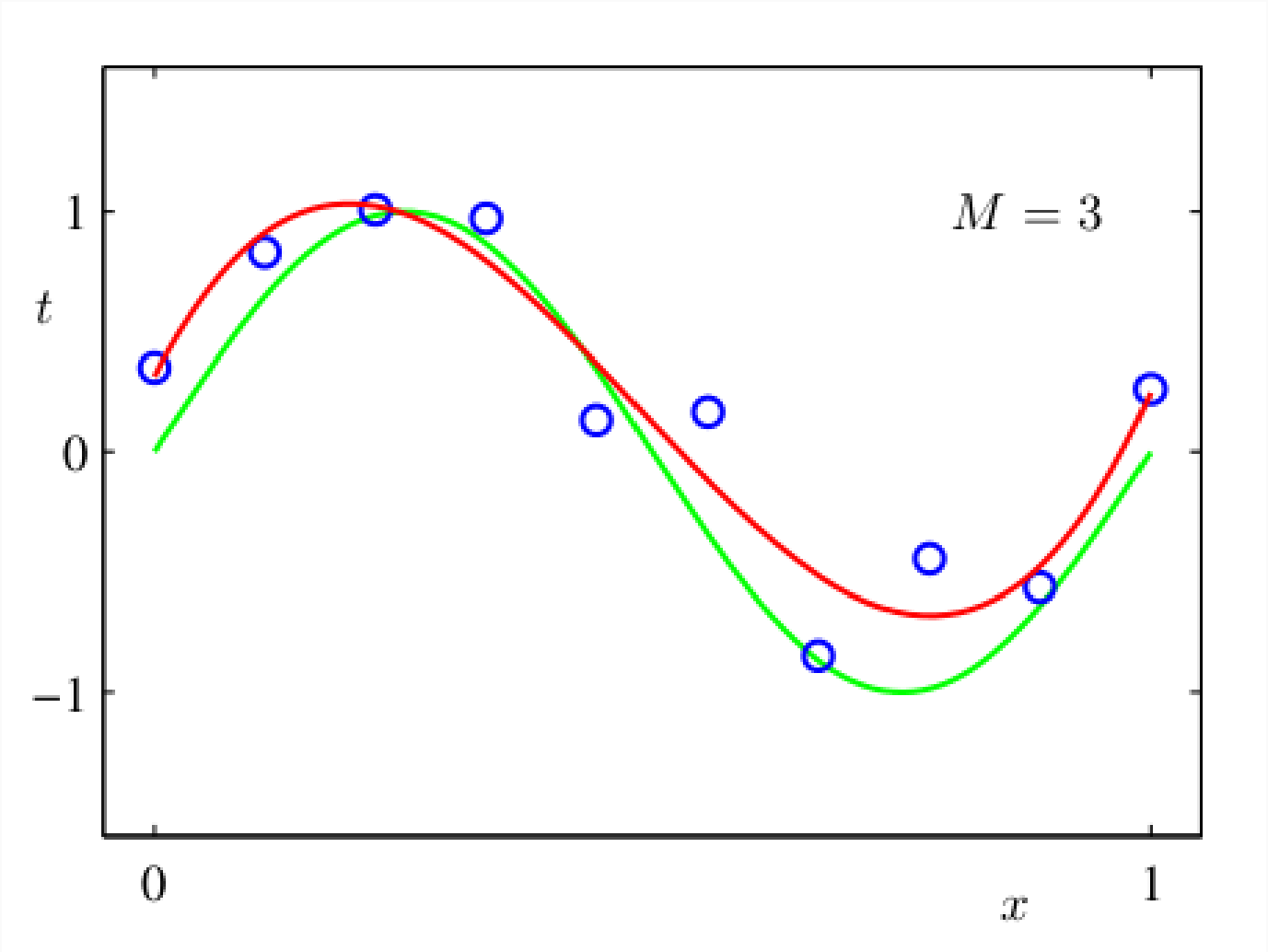
Полиномиальная аппроксимация



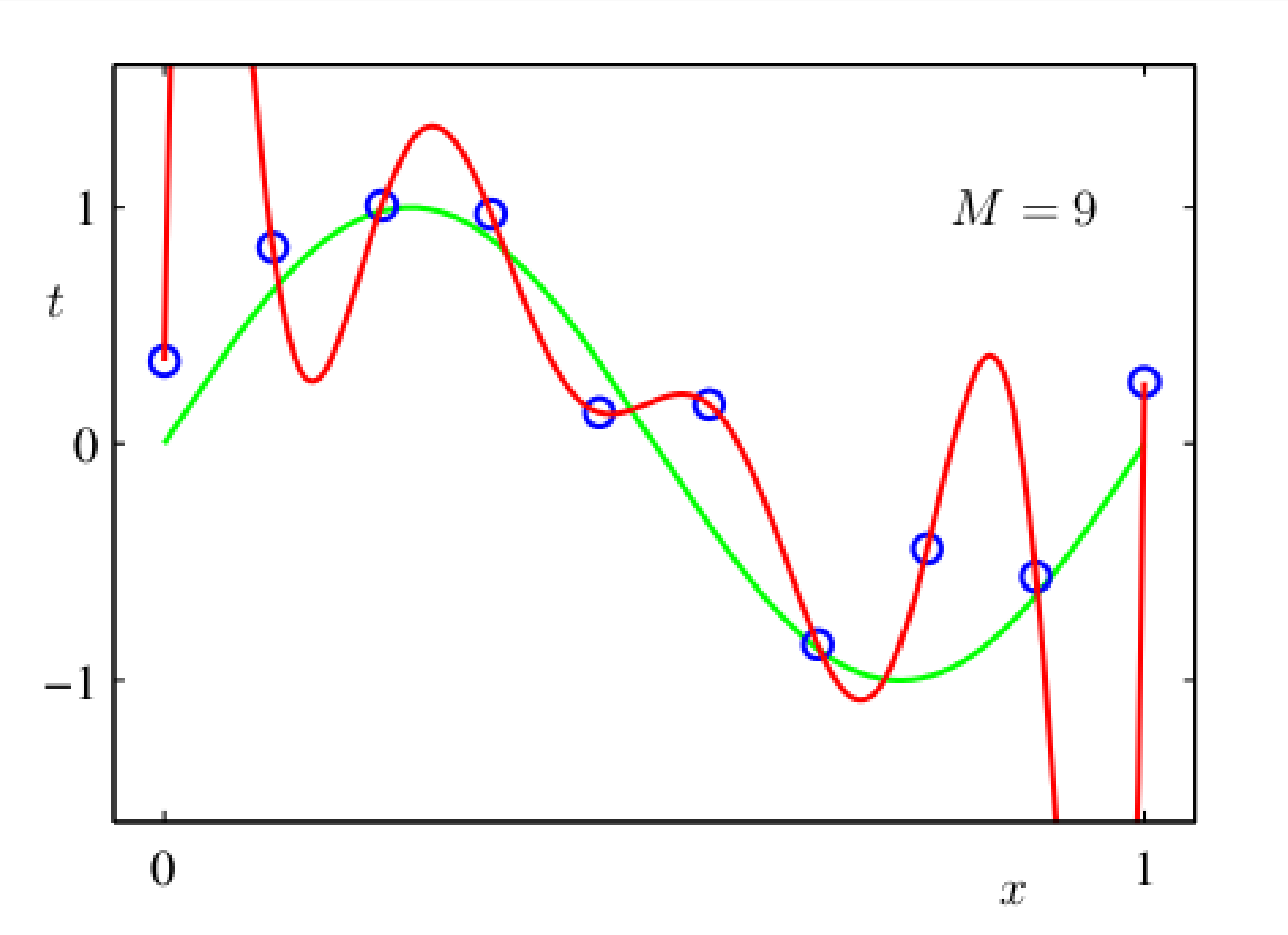
Полиномиальная аппроксимация



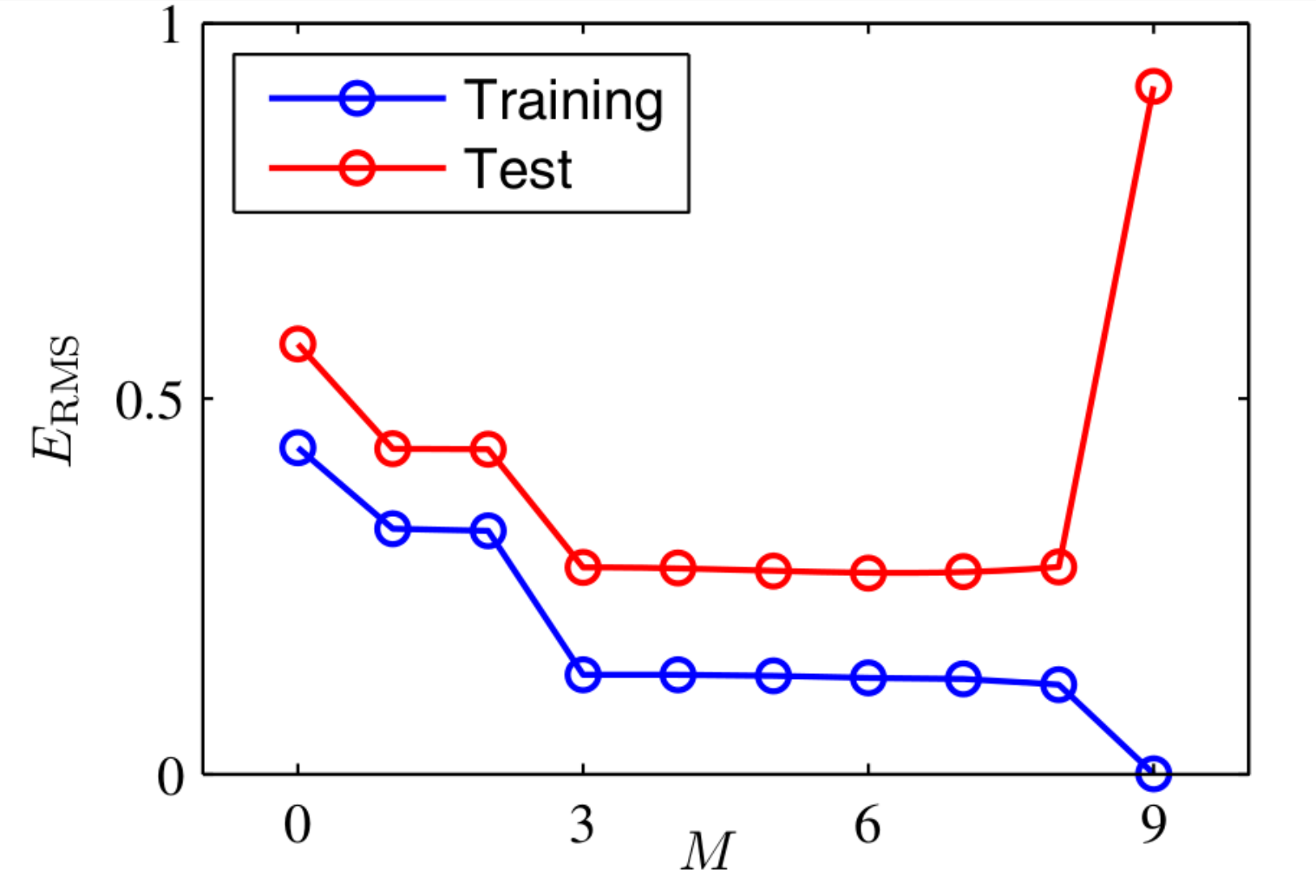
Полиномиальная аппроксимация



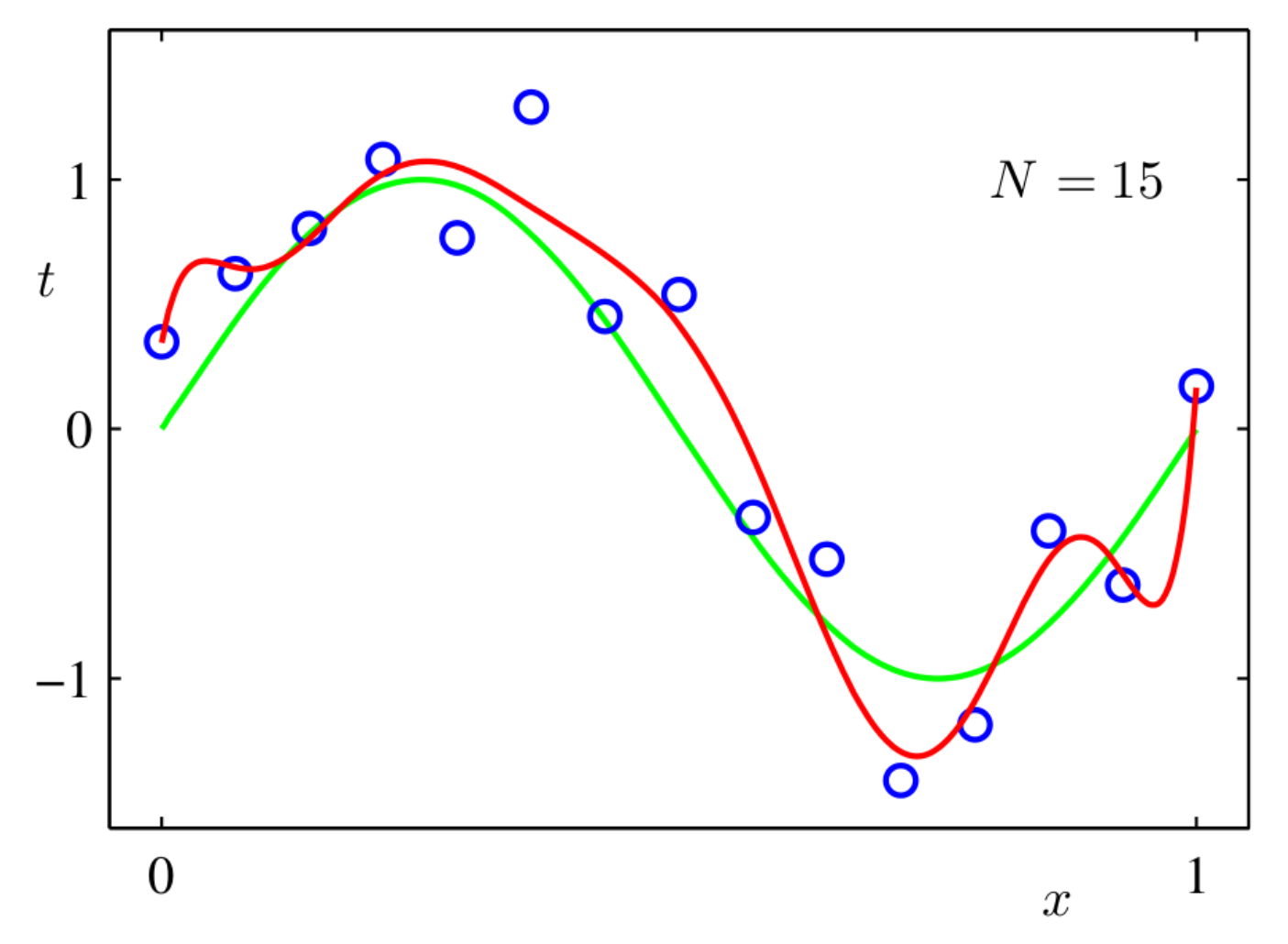
Полиномиальная аппроксимация



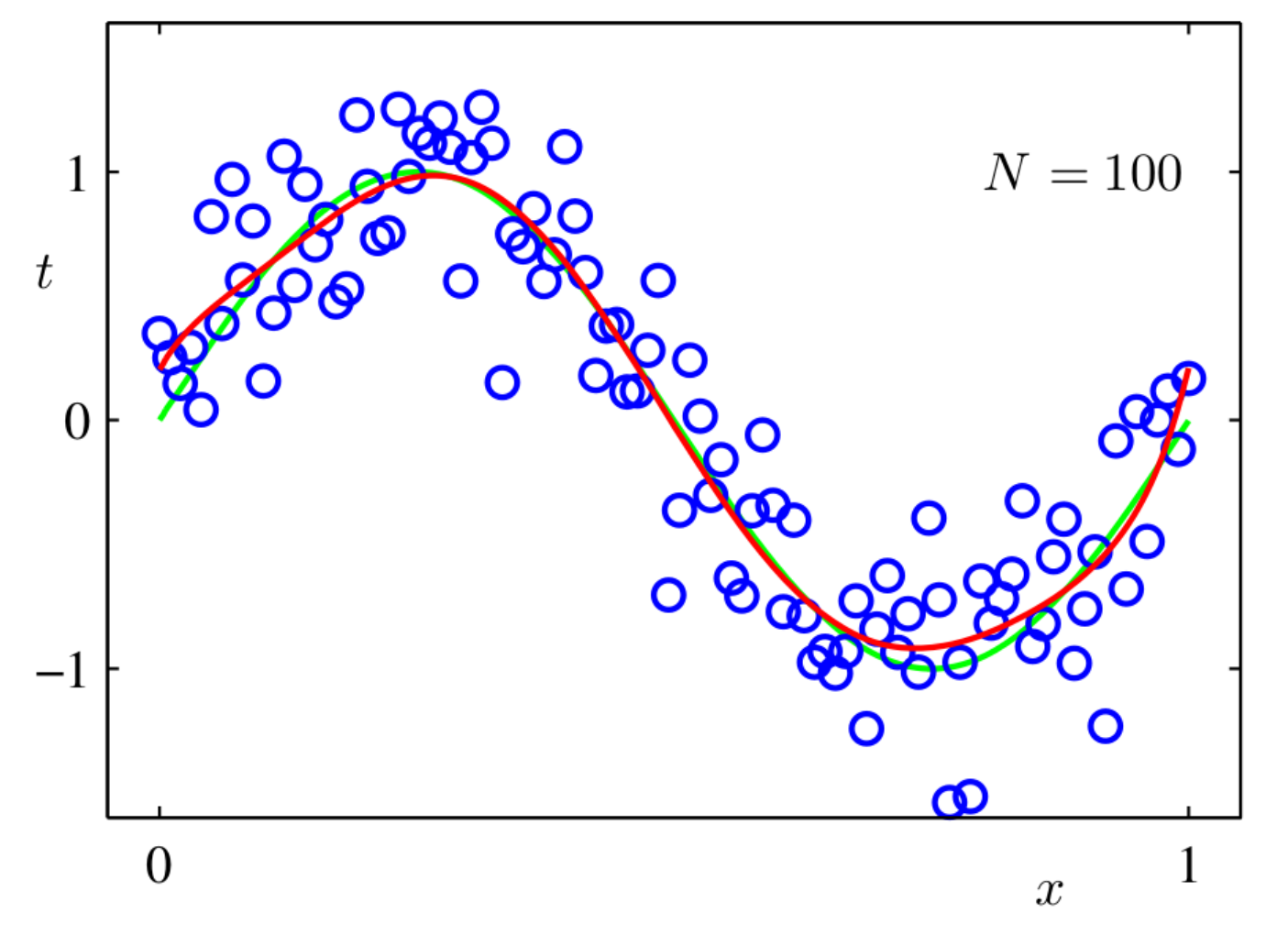
Значения RMS



Можно собрать больше данных...



Можно собрать больше данных...



Значения коэффициентов

	$M=0$	$M=1$	$M=6$	$M=9$
$w_0^\star$	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

Оверфиттинг

- Итак, мы увидели, что даже в линейной регрессии может наступить оверфиттинг.
- Что же делать?..

Регуляризация в линейной регрессии

Регуляризация

- Итак, получается, что у нас сильно растут коэффициенты.
- Давайте попробуем с этим бороться. Бороться будем прямолинейно и простодушно: возьмём и добавим размер коэффициентов в функцию ошибки.

Регуляризация

- Было (для тестовых примеров $\{(x_i, y_i)\}_{i=1}^N$):

$$\mathrm{RSS}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}(f(x_i, \mathbf{w}) - y_i)^2.$$

- Стало:

$$\mathrm{RSS}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}(f(x_i, \mathbf{w}) - y_i)^2 + \frac{\alpha}{2}\left\|\mathbf{w}\right\|^2,$$

  где α – коэффициент регуляризации (его надо будет как-нибудь выбрать).

- Как оптимизировать эту функцию ошибки?

Регуляризация

- Да точно так же – запишем как

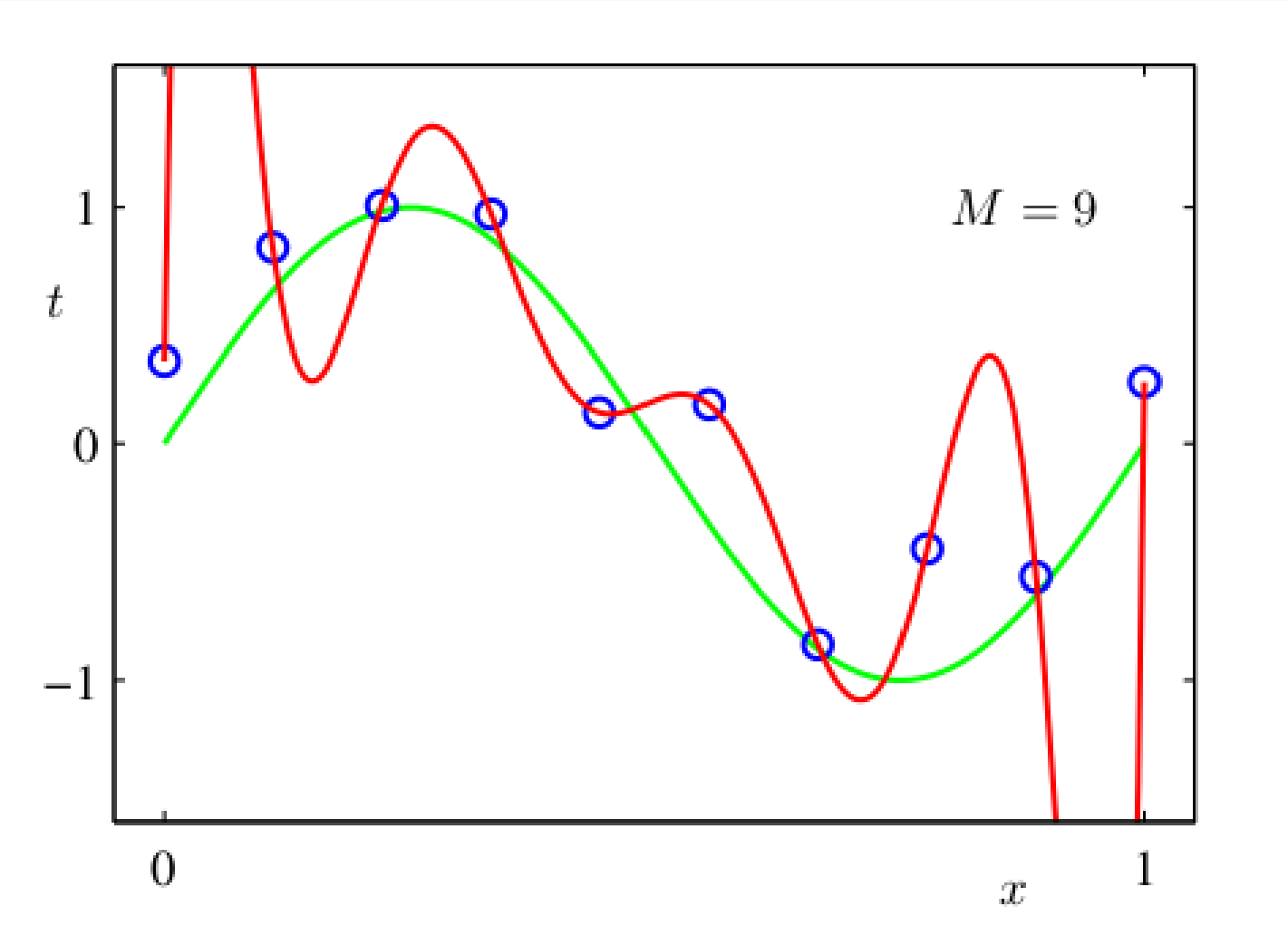
$$\mathrm{RSS}(\mathsf{w}) = \frac{1}{2}(\mathsf{y} - \mathsf{X}\mathsf{w})^\top (\mathsf{y} - \mathsf{X}\mathsf{w}) + \frac{\alpha}{2}\mathsf{w}^\top \mathsf{w}$$

  и возьмём производную; получится

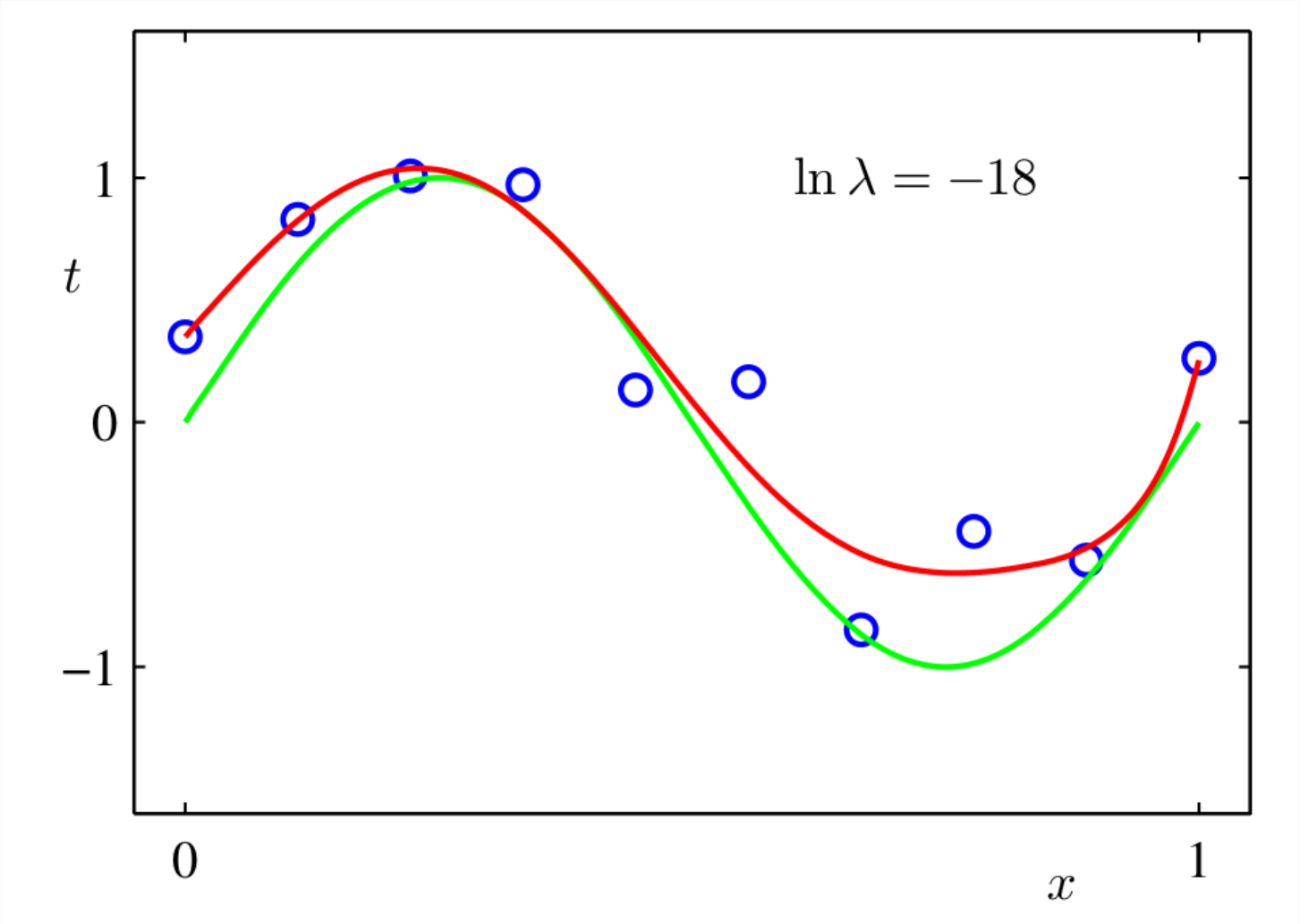
$$\mathsf{w}^* = \left(\mathsf{X}^\top \mathsf{X} + \alpha \mathsf{I}\right)^{-1} \mathsf{X}^\top \mathsf{y}.$$

- Это *гребневая регрессия* (ridge regression); кстати, добавление $\alpha\mathsf{I}$ к матрице неполного ранга делает её обратимой; это и есть *регуляризация*, и это и было исходной мотивацией для гребневой регрессии.

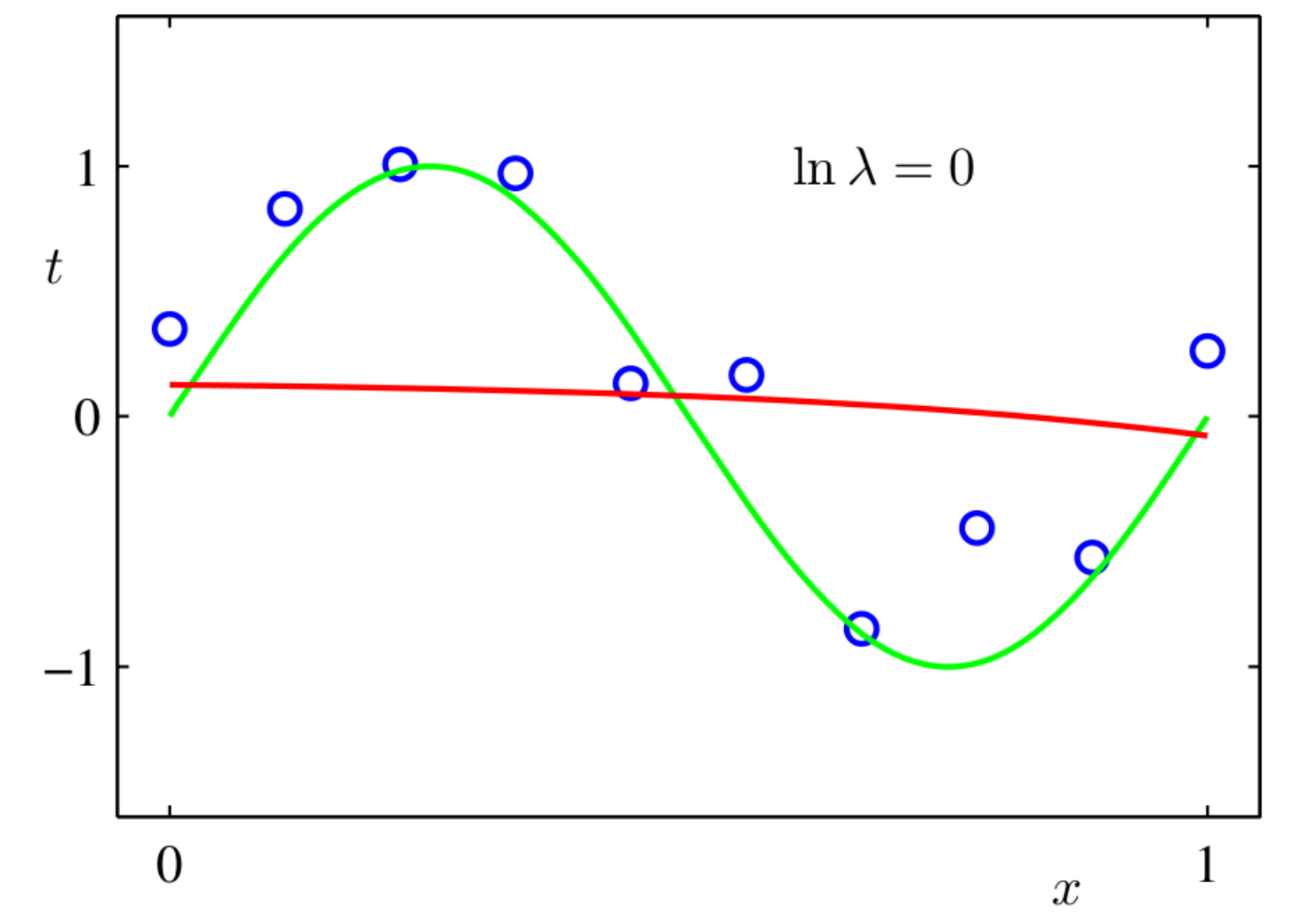
Гребневая регрессия: $\ln \alpha = -\infty$



Гребневая регрессия: $\ln \alpha = -18$



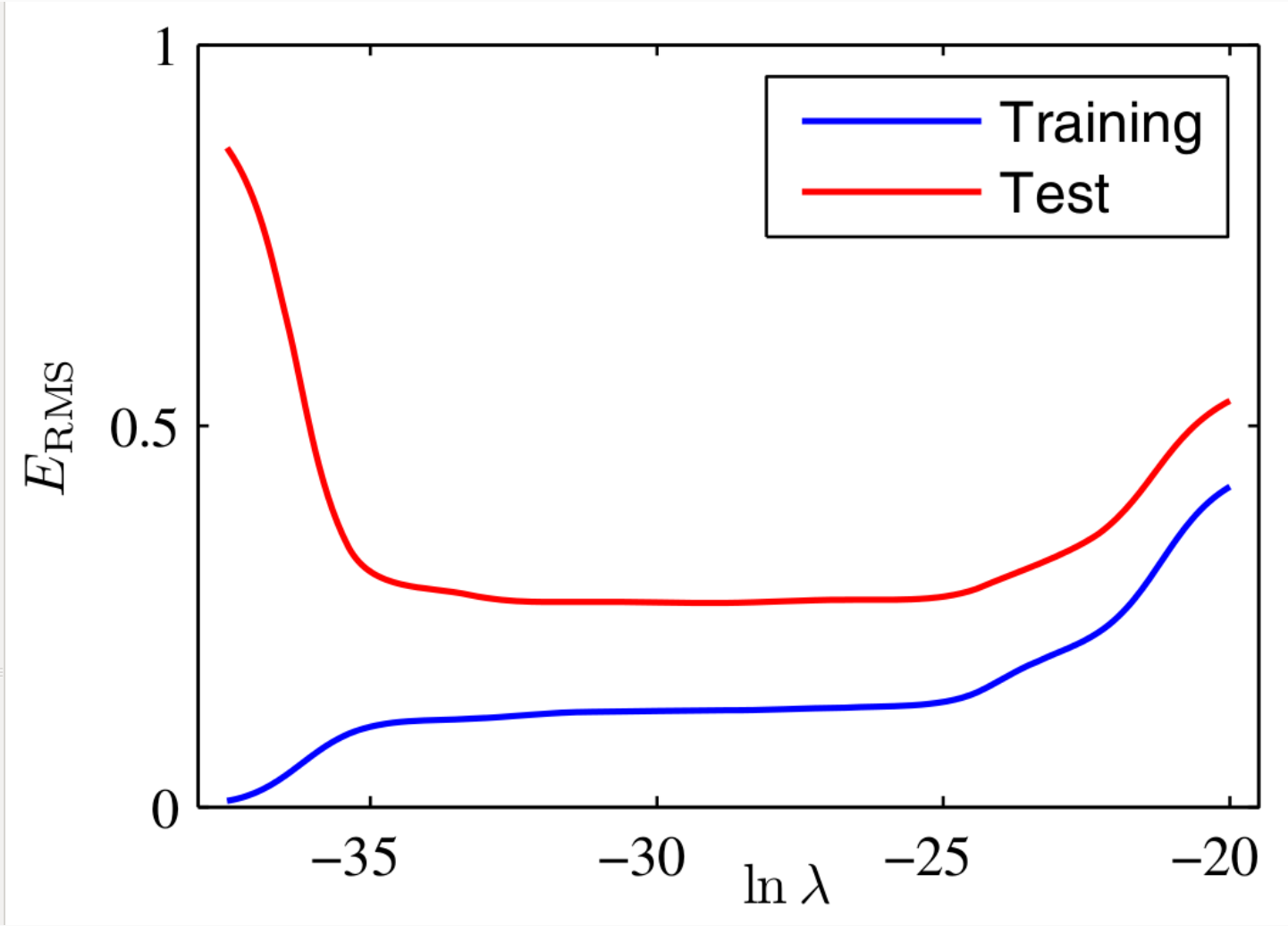
Гребневая регрессия: $\ln \alpha = 0$



Гребневая регрессия: коэффициенты

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^\star$	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

Гребневая регрессия: RMS



Другая регуляризация

- Почему именно так? Почему именно $\frac{\alpha}{2}\|\mathbf{w}\|^2$?
- Мы сейчас ответим на этот вопрос, но, вообще говоря, это не обязательно.
- Лассо-регрессия (lasso regression) регуляризует L_1-нормой, а не L_2:

$$\mathrm{RSS}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}(f(x_i, \mathbf{w}) - y_i)^2 + \alpha\sum_{j=0}^{M}|w_j|.$$

- Есть и другие типы; об этом будем говорить позже.

Регрессия по-байесовски

Байесовская регуляризация

- А теперь давайте посмотрим на регрессию с совсем байесовской стороны.
- Напомним основу байесовского подхода:
  1. найти апостериорное распределение на гипотезах/параметрах:

$$p(\theta \mid D) \propto p(D|\theta)p(\theta)$$

     (и/или найти максимальную апостериорную гипотезу $\arg\max_\theta p(\theta \mid D)$);
  2. найти апостериорное распределение исходов дальнейших экспериментов:

$$p(x \mid D) \propto \int_{\theta \in \Theta} p(x \mid \theta)p(D|\theta)p(\theta)\mathrm{d}\theta.$$

Байесовская регуляризация

- В нашем рассмотрении пока не было никаких априорных распределений.
- Давайте какое-нибудь введём; например, нормальное (почему так – позже):

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0).$$

- Рассмотрим набор данных $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ со значениями $\mathbf{t} = \{t_1, \ldots, t_N\}$. В этой модели мы предполагаем, что данные независимы и одинаково распределены:

$$p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(t_n \mid \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n), \sigma^2\right).$$

Байесовская регуляризация

- Тогда наша задача – посчитать

$$p(\mathbf{w} \mid \mathbf{t}) \propto p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}, \sigma^2) p(\mathbf{w})$$
$$= \mathcal{N}(\mathbf{w} \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \prod_{n=1}^{N} \mathcal{N}\left(t_n \mid \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n), \sigma^2\right).$$

- Давайте подсчитаем.

Байесовская регуляризация

- Получится

$$p(\mathbf{w} \mid \mathbf{t}) = \mathcal{N}(\mathbf{w} \mid \boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N),$$
$$\mu_N = \boldsymbol{\Sigma}_N \left(\boldsymbol{\Sigma}_0^{-1} \mu_0 + \frac{1}{\sigma^2} \boldsymbol{\Phi}^\top \mathbf{t} \right),$$
$$\boldsymbol{\Sigma}_N = \left(\boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma^2} \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \right)^{-1}.$$

- Теперь давайте подсчитаем логарифм правдоподобия.

Байесовская регуляризация

- Если мы возьмём априорное распределение около нуля:

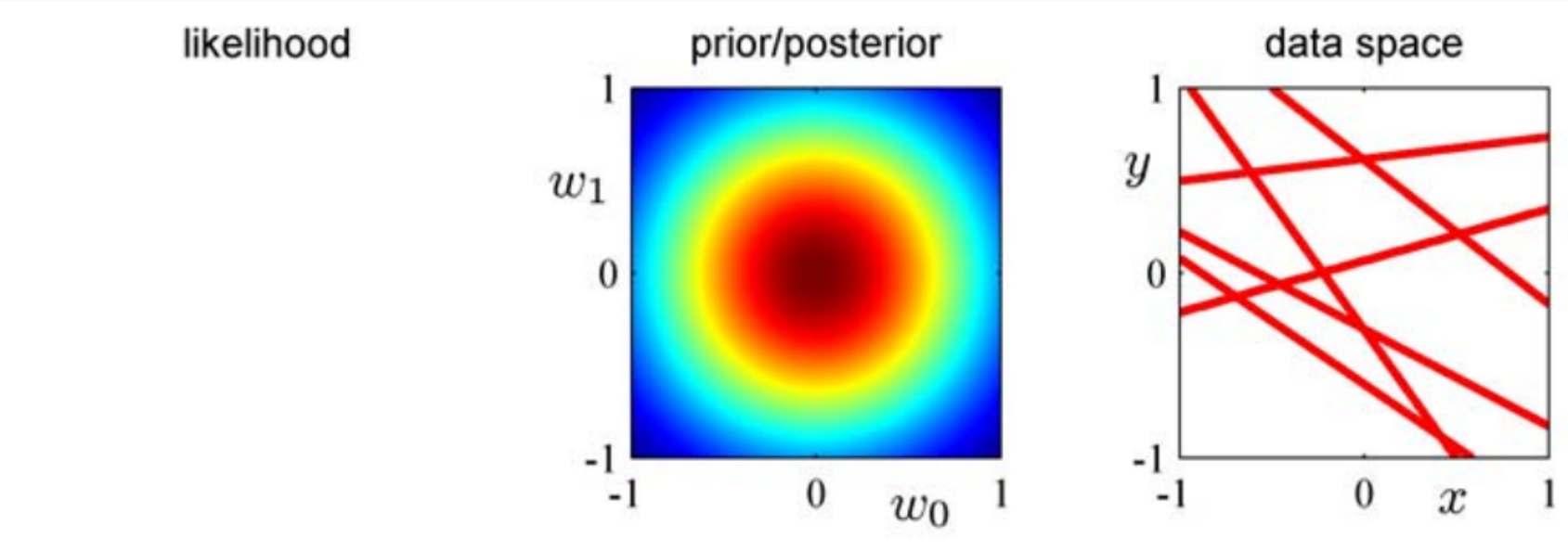
$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \frac{1}{\alpha}\mathbf{I}),$$

то логарифм правдоподобия получится

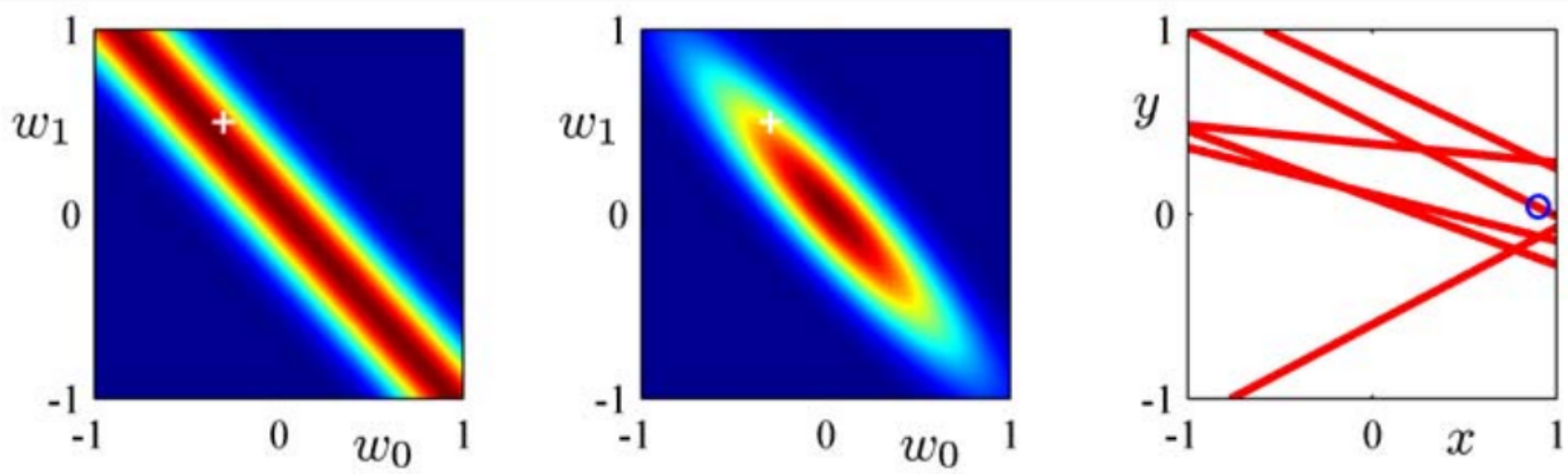
$$\ln p(\mathbf{w} \mid \mathbf{t}) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n)\right)^2 - \frac{\alpha}{2}\mathbf{w}^\top\mathbf{w} + \text{const},$$

то есть в точности гребневая регрессия.

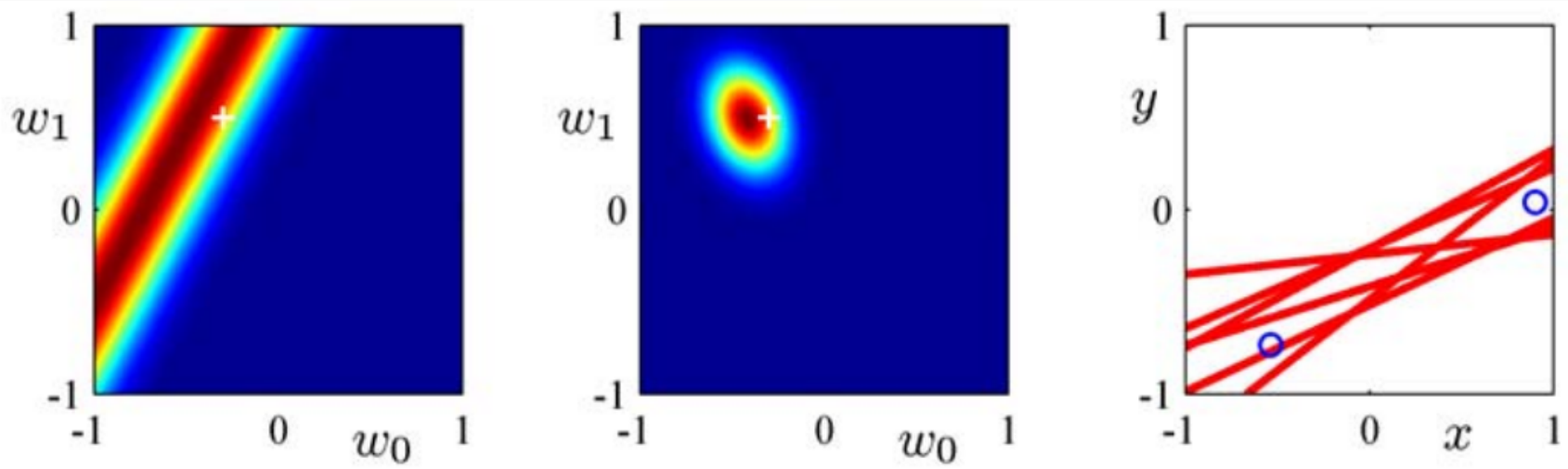
Пример


likelihood
prior/posterior
data space
1
w_1
0
-1
-1
0
w_0
1
1
y
0
-1
-1
0
x
1

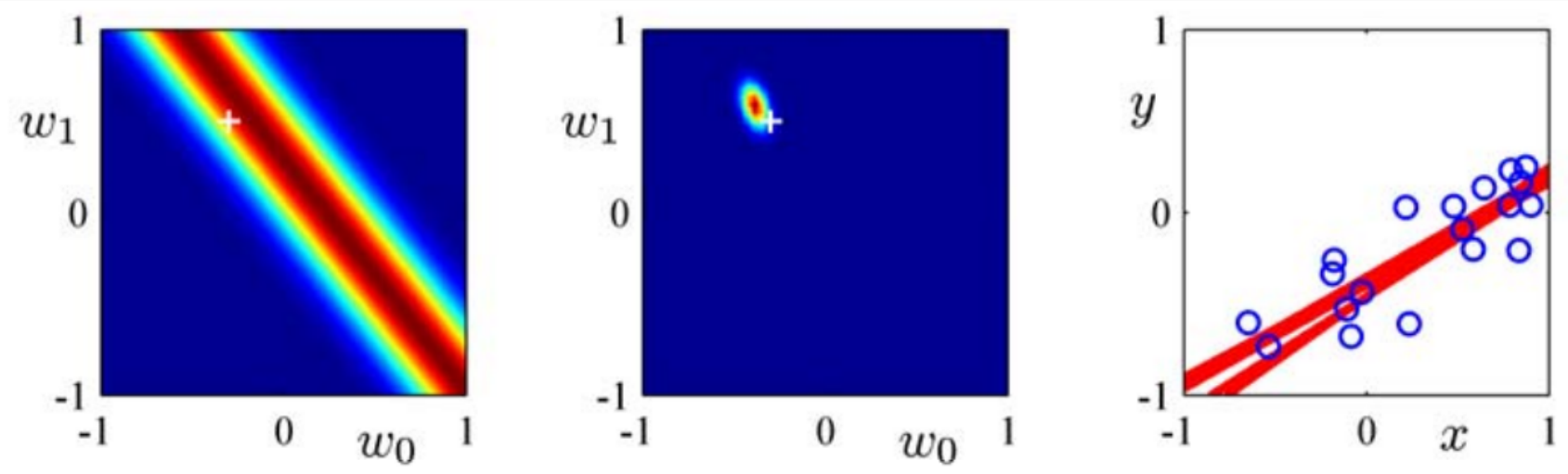
Пример



Пример



Пример



Лассо

- Теперь давайте рассмотрим лассо-регрессию:

$$L(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}(f(x_i, \mathbf{w}) - y_i)^2 + \lambda \sum_{j=0}^{p} |w_j|.$$

- Главное отличие – теперь форма ограничений (т.е. форма априорного распределения) такова, что весьма вероятно получить строго нулевые w_j.
- Кстати, что значит «форма ограничений»?

Лассо

- Мы можем переписать регрессию с регуляризатором по-другому:

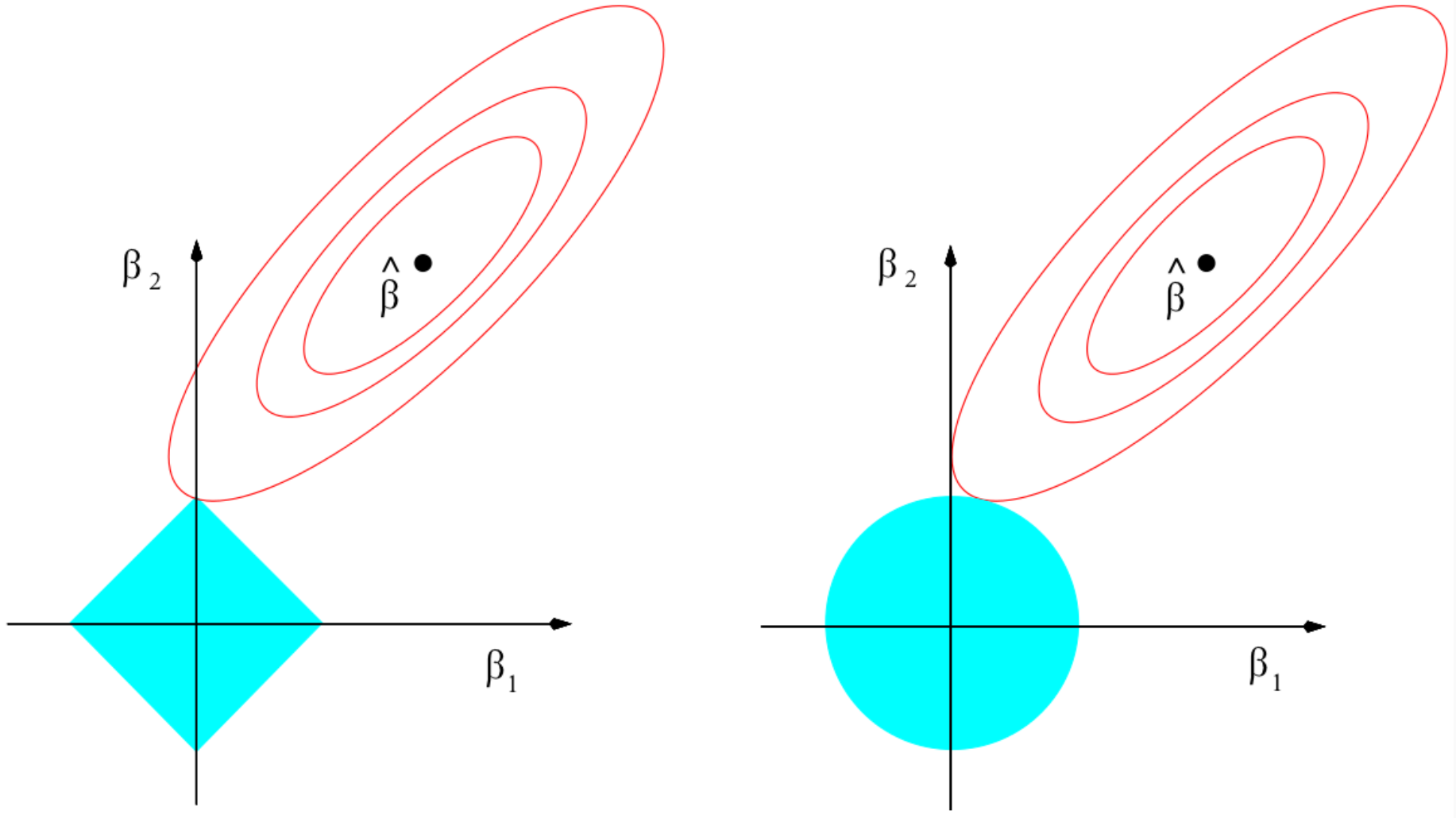
$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left\{ \frac{1}{2} \sum_{i=1}^{N} (f(x_i, \mathbf{w}) - y_i)^2 + \lambda \sum_{j=0}^{p} |w_j| \right\},$$

эквивалентно

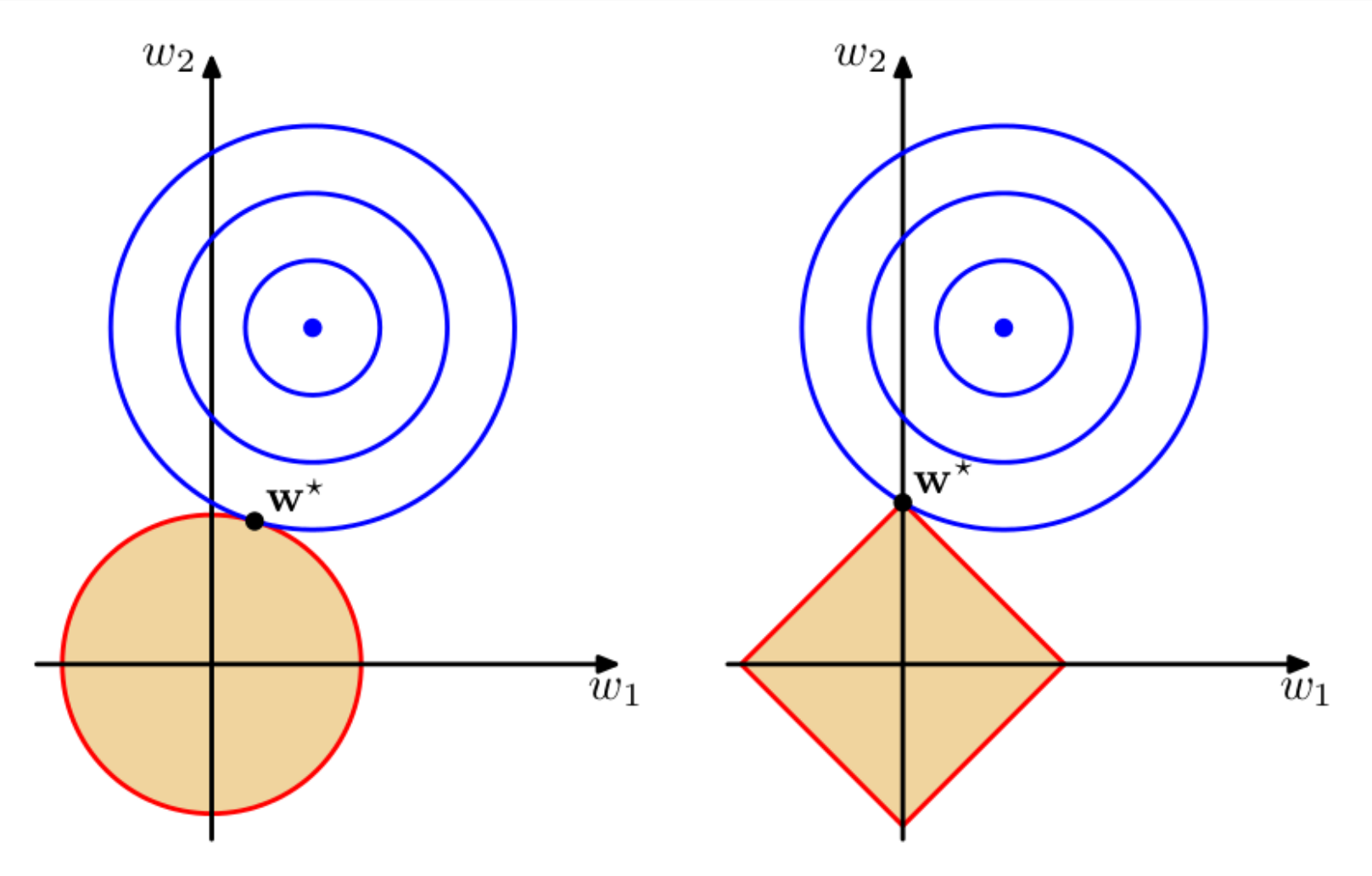
$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left\{ \frac{1}{2} \sum_{i=1}^{N} (f(x_i, \mathbf{w}) - y_i)^2 \right\} \text{ при } \sum_{j=0}^{p} |w_j| \leq t.$$

Упражнение. Докажите это.

Гребень и лассо



Гребень и лассо



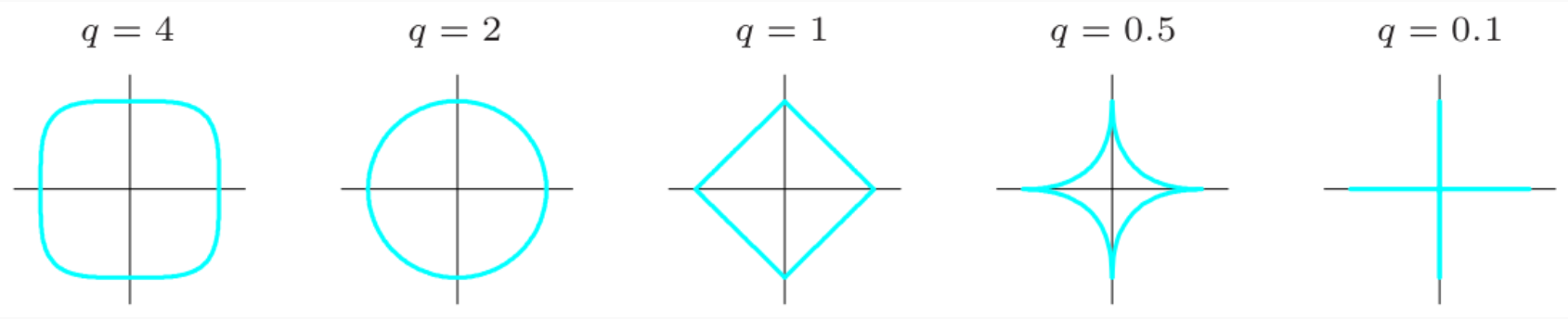
Обобщение

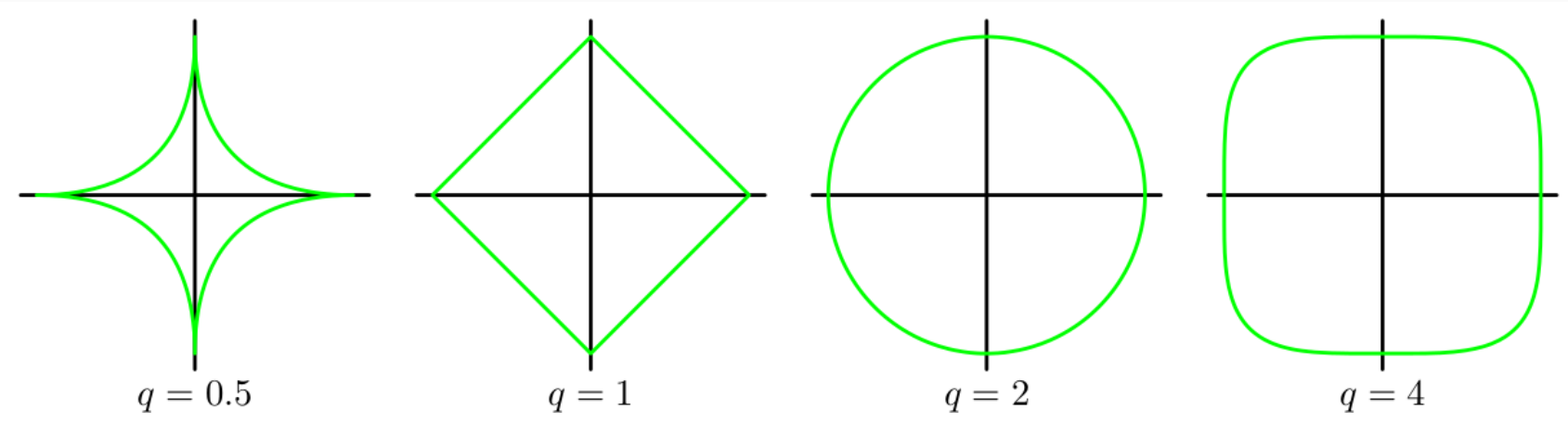
- Можно рассмотреть обобщение гребневой и лассо-регрессии:

$$L(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}(f(x_i, \mathbf{w}) - y_i)^2 + \lambda \sum_{j=0}^{p}(|w_j|)^q.$$

Упражнение. Какому априорному распределению на параметры $\mathbf{w}$ соответствует эта задача?

Разные q





Спасибо!

Спасибо за внимание!