ЛИНЕЙНАЯ РЕГРЕССИЯ

Сергей Николенко

СПбГУ — Санкт-Петербург
15 сентября 2017 г.

Random facts:

- 15 сентября 1928 г. Александр Флеминг впервые в мире выделил пенициллин
- 15 сентября 1933 г. Ягода сообщил Сталину о раскрытии в Ленинграде «общества педерастов»; были арестованы более 150 человек
- 15 сентября 1974 г. на пересечении улиц Островитянова и Профсоюзной в Беляево работники по благоустройству и развитию лесопарка крайне нервно, с применением тяжелой техники отреагировали на попытку сорвать им субботник
- 15 сентября 1997 г. был зарегистрирован домен google.com
- 15 сентября 2007 г. близ деревни Каранкас около Титикаки упал необычный метеорит; в то же время 600 жителей деревни начали страдать от разных болезней с неясными причинами

Линейная регрессия

Метод наименьших квадратов

- Линейная регрессия: рассмотрим линейную функцию

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

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

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

Метод наименьших квадратов

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

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

- Как минимизировать?

Метод наименьших квадратов

- Можно на самом деле решить задачу точно – записать как

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

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

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

  если матрица $X^\top X$ невырожденная.
- Замечание: $\left(X^\top X\right)^{-1} X^\top$ называется *псевдообратной матрицей Мура–Пенроуза* (Moore–Penrose pseudo-inverse) матрицы 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

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

  Иными словами,

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

- 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:

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

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

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 $\phi_j(x) = e^{-\frac{(x-\mu_j)^2}{2s^2}}$);
  - ...

Байесовская регрессия

- Рассмотрим набор данных $X = \{x_1, \dots, x_N\}$ со значениями $t = \{t_1, \dots, t_N\}$.
- Будем предполагать, что данные взяты независимо по одному и тому же распределению:

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

- Прологарифмируем (опустим X, т.к. по нему всегда условная вероятность будет):

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

Байесовская регрессия

- Прологарифмируем (опустим X, т.к. по нему всегда условная вероятность будет):

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

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

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

Байесовская регрессия

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

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

- Здесь $\Phi = (\phi_j(x_i))_{i,j}$.

Байесовская регрессия

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

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}\left(t_n - w_{ML}^\top \phi(x_n)\right)^2,$$

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

Пример: полиномиальная аппроксимация

Полиномиальная аппроксимация

- Мы говорили о регрессии с базисными функциями:

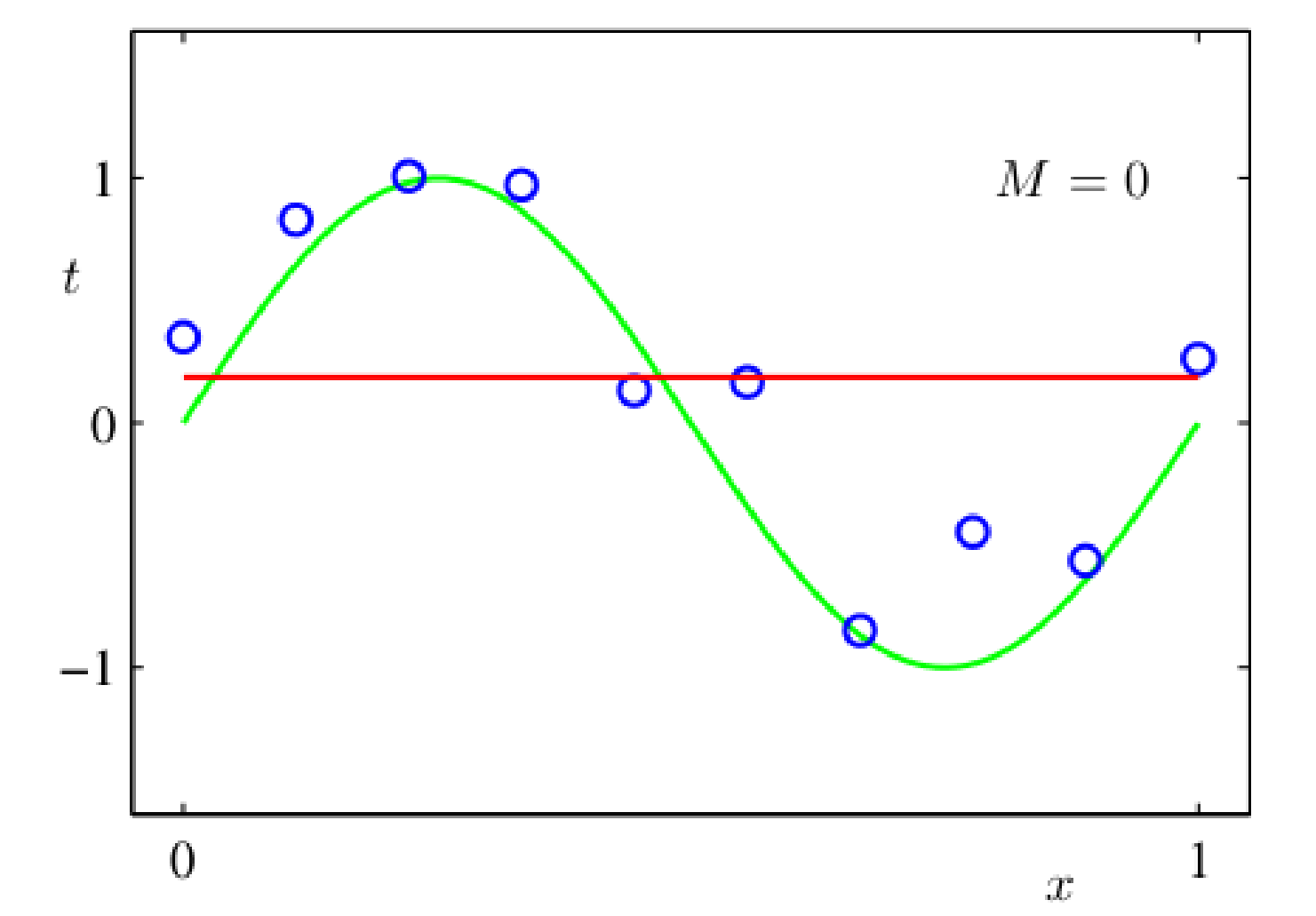
$$f(x, w) = w_0 + \sum_{j=1}^{M} w_j \phi_j(x) = w^\top \phi(x).$$

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

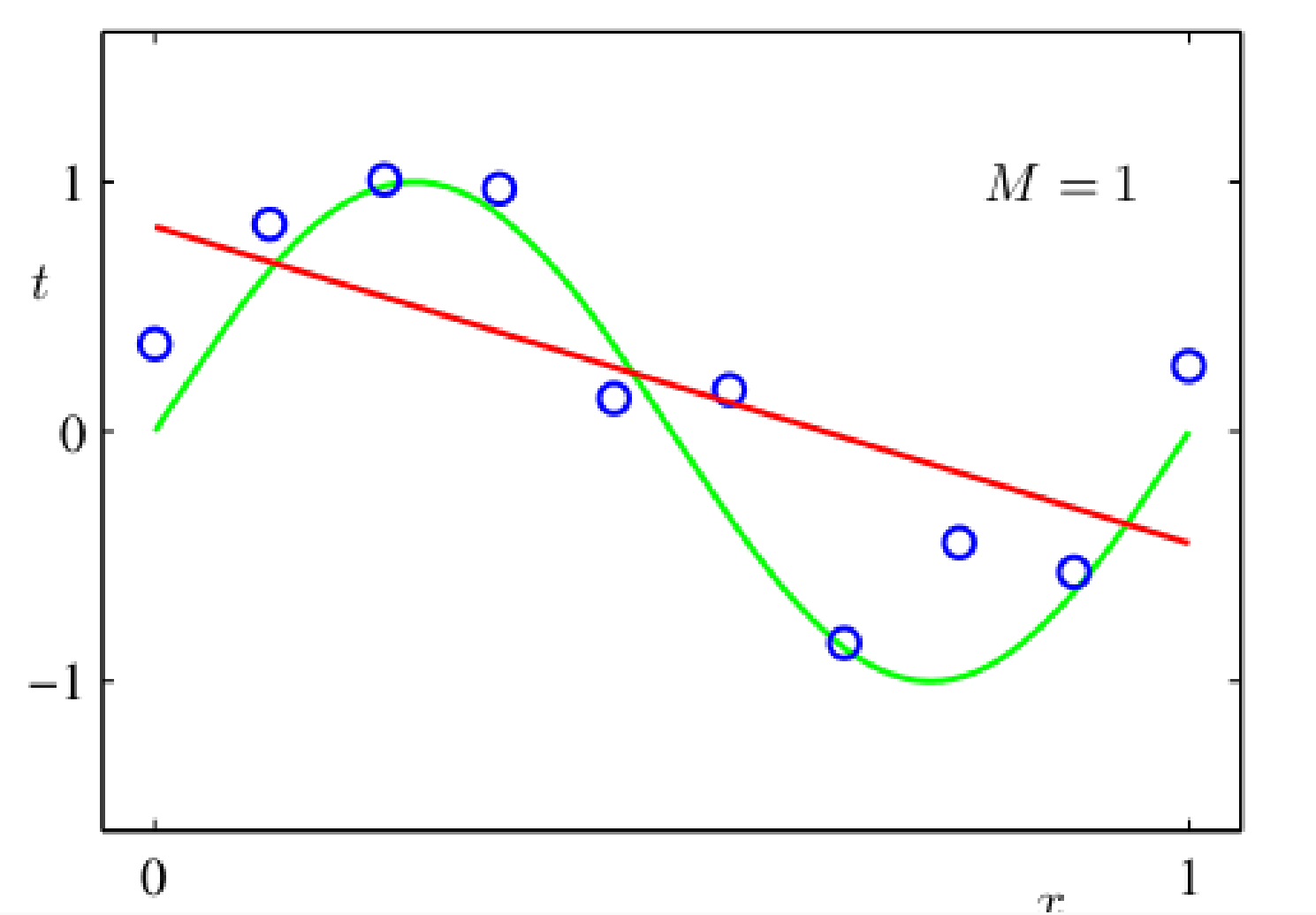
$$f(x, w) = w_0 + w_1 x + w_2 x^2 + ... + w_M x^M.$$

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

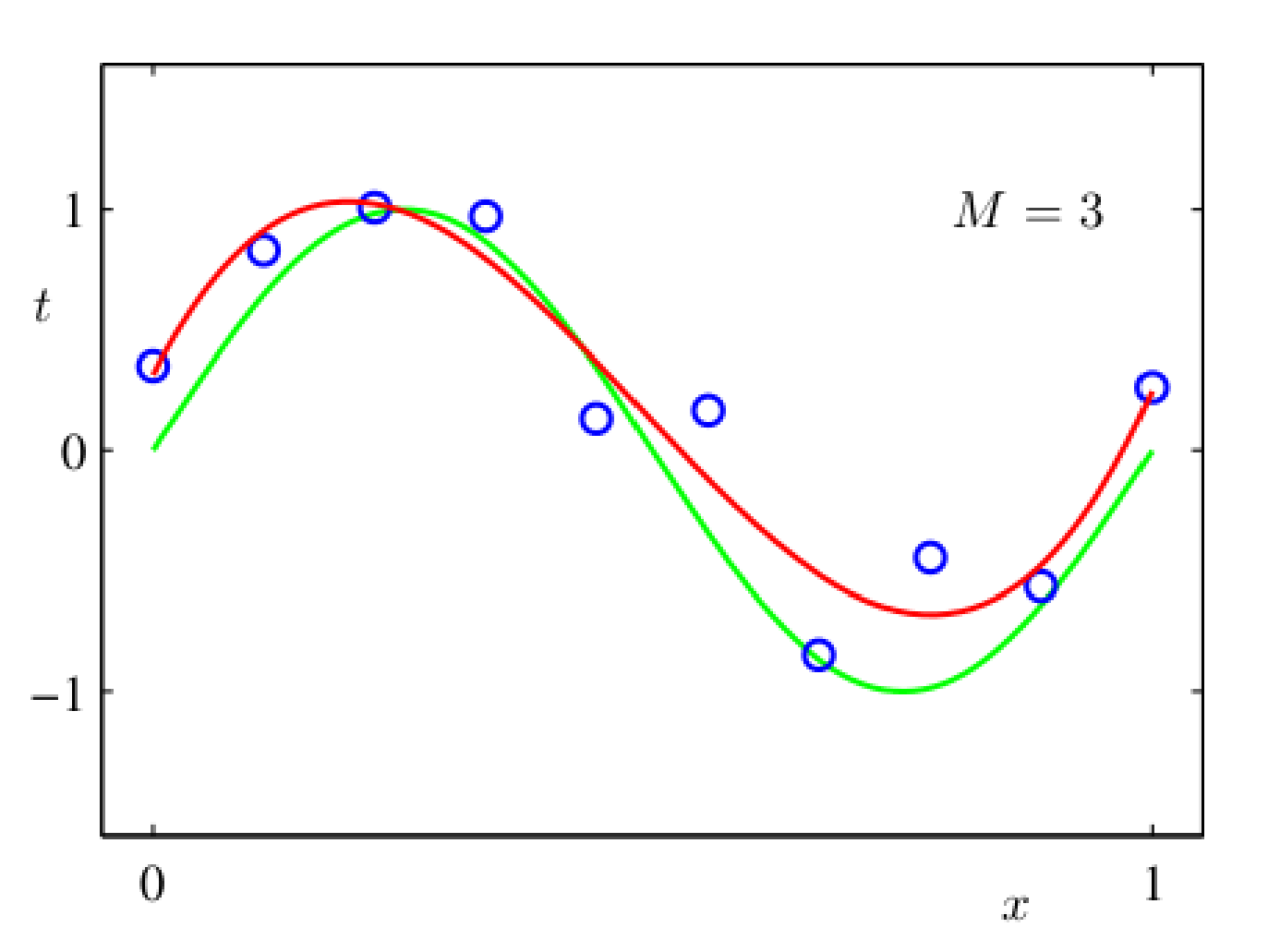
Полиномиальная аппроксимация



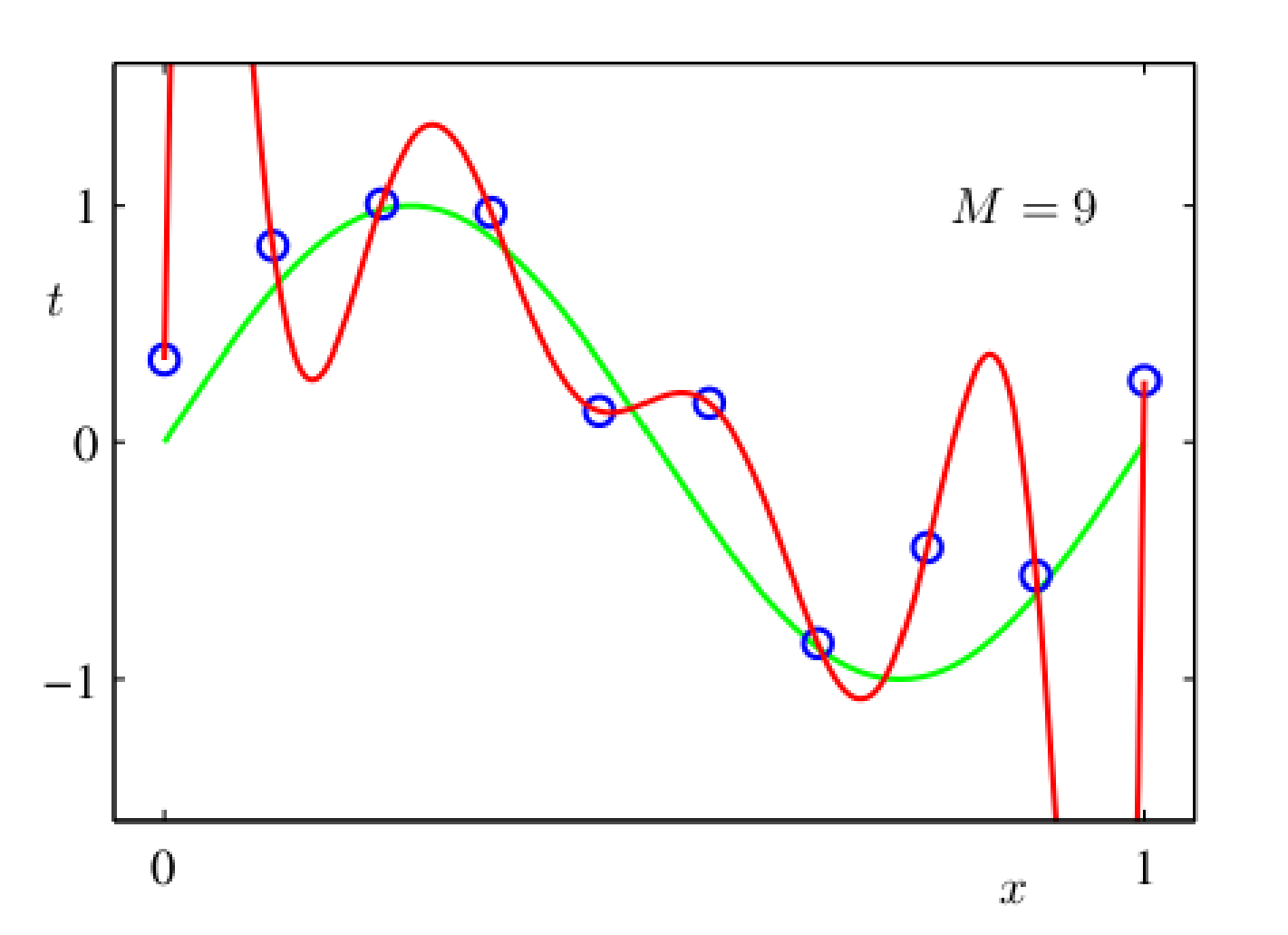
Полиномиальная аппроксимация



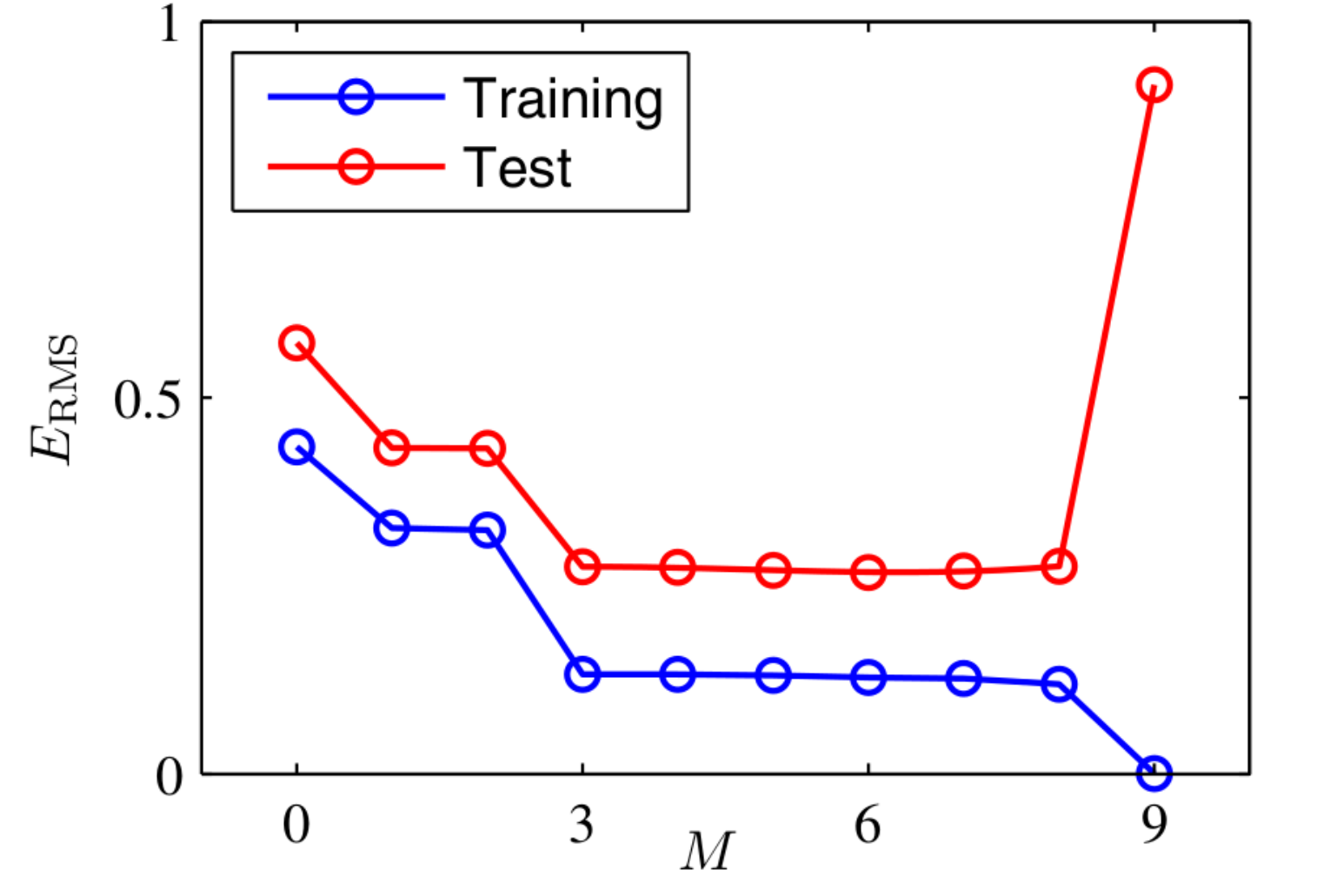
Полиномиальная аппроксимация



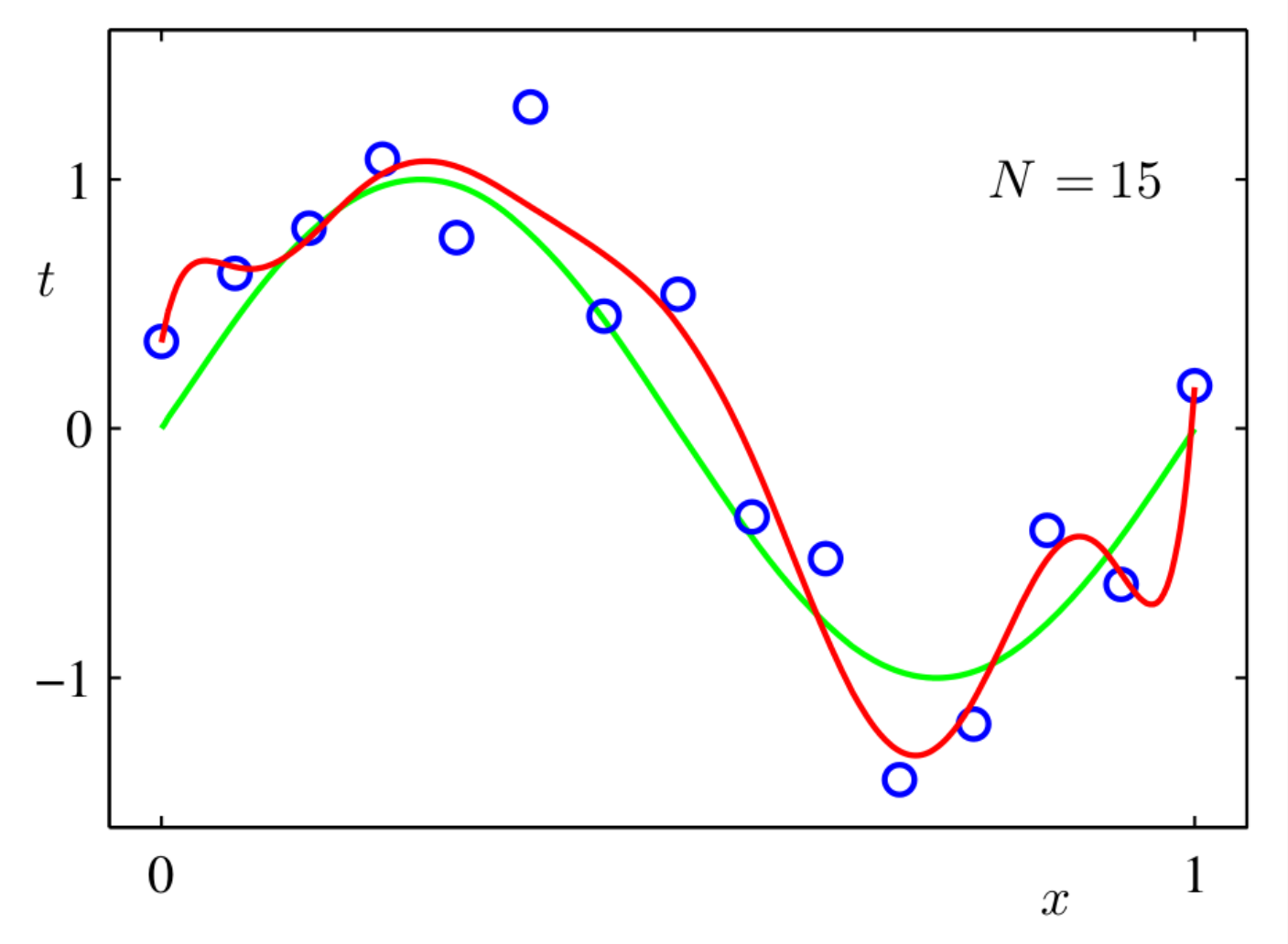
Полиномиальная аппроксимация



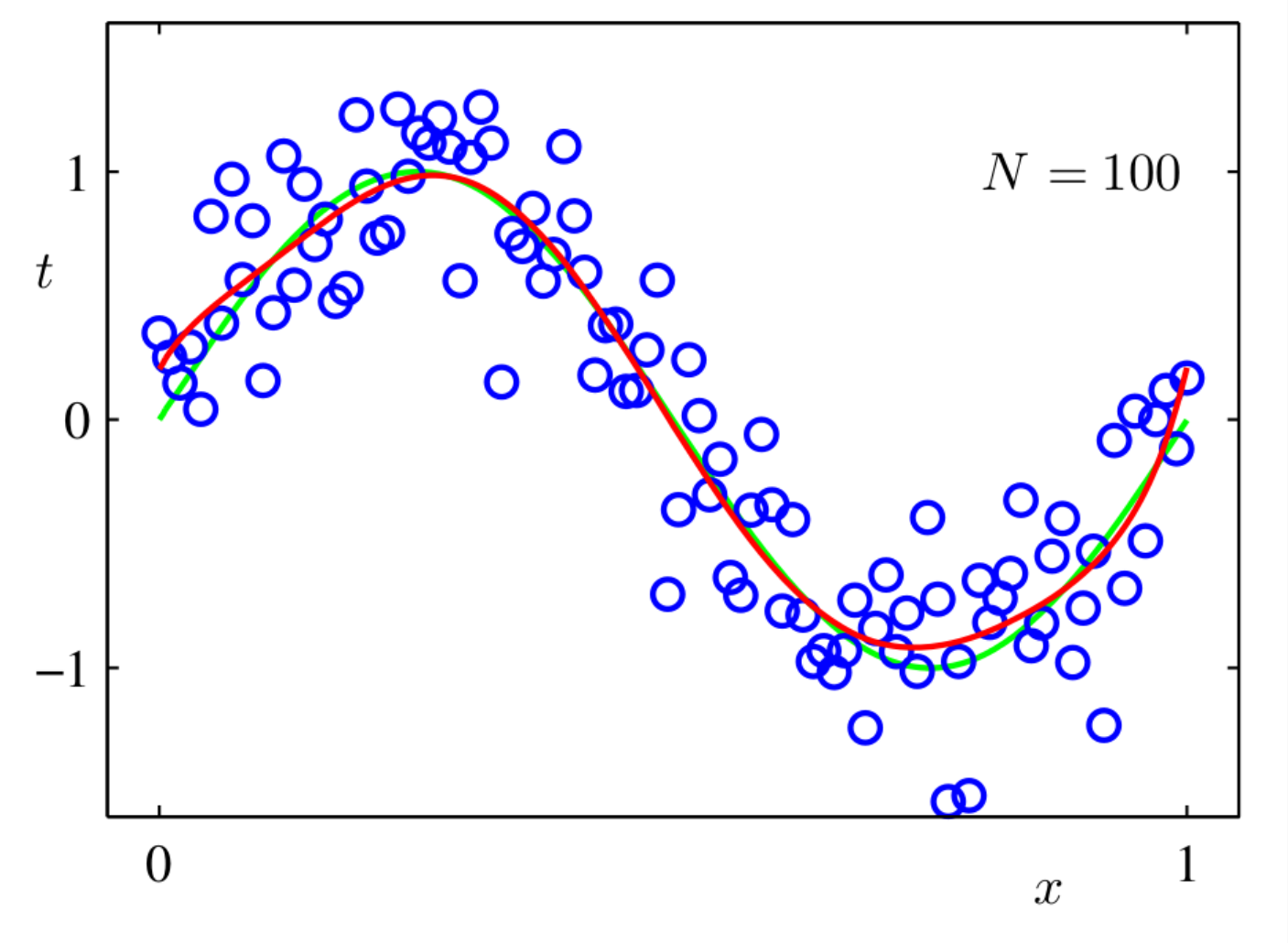
Значения RMS



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

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 $\{(x_i, y_i)\}_{i=1}^N$):

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

- Стало:

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

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

- Как оптимизировать эту функцию ошибки?

Регуляризация

- Да точно так же – запишем как

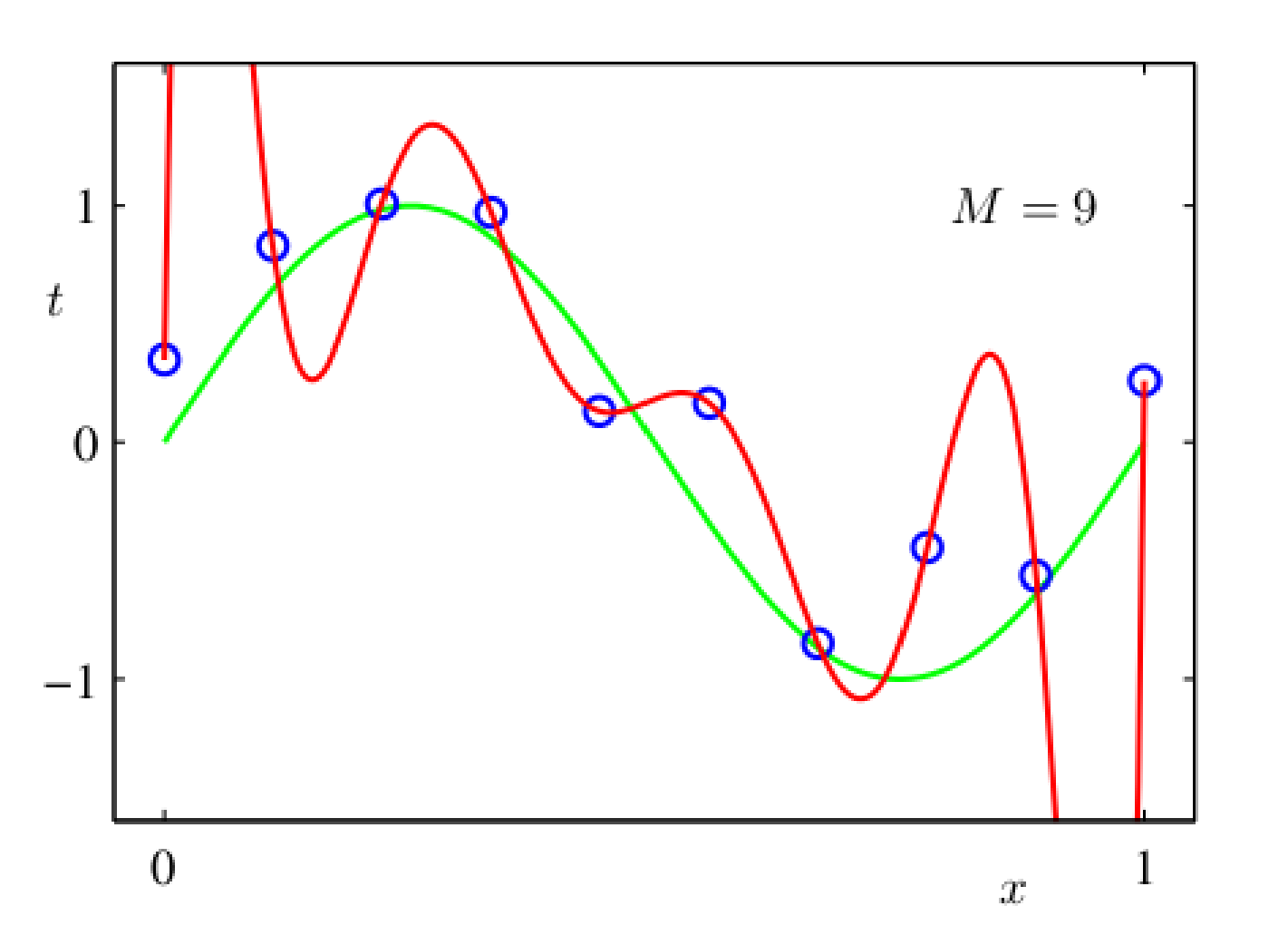
$$\mathrm{RSS}(w) = \frac{1}{2}\left(y - Xw\right)^\top \left(y - Xw\right) + \frac{\alpha}{2} w^\top w$$

  и возьмём производную; получится

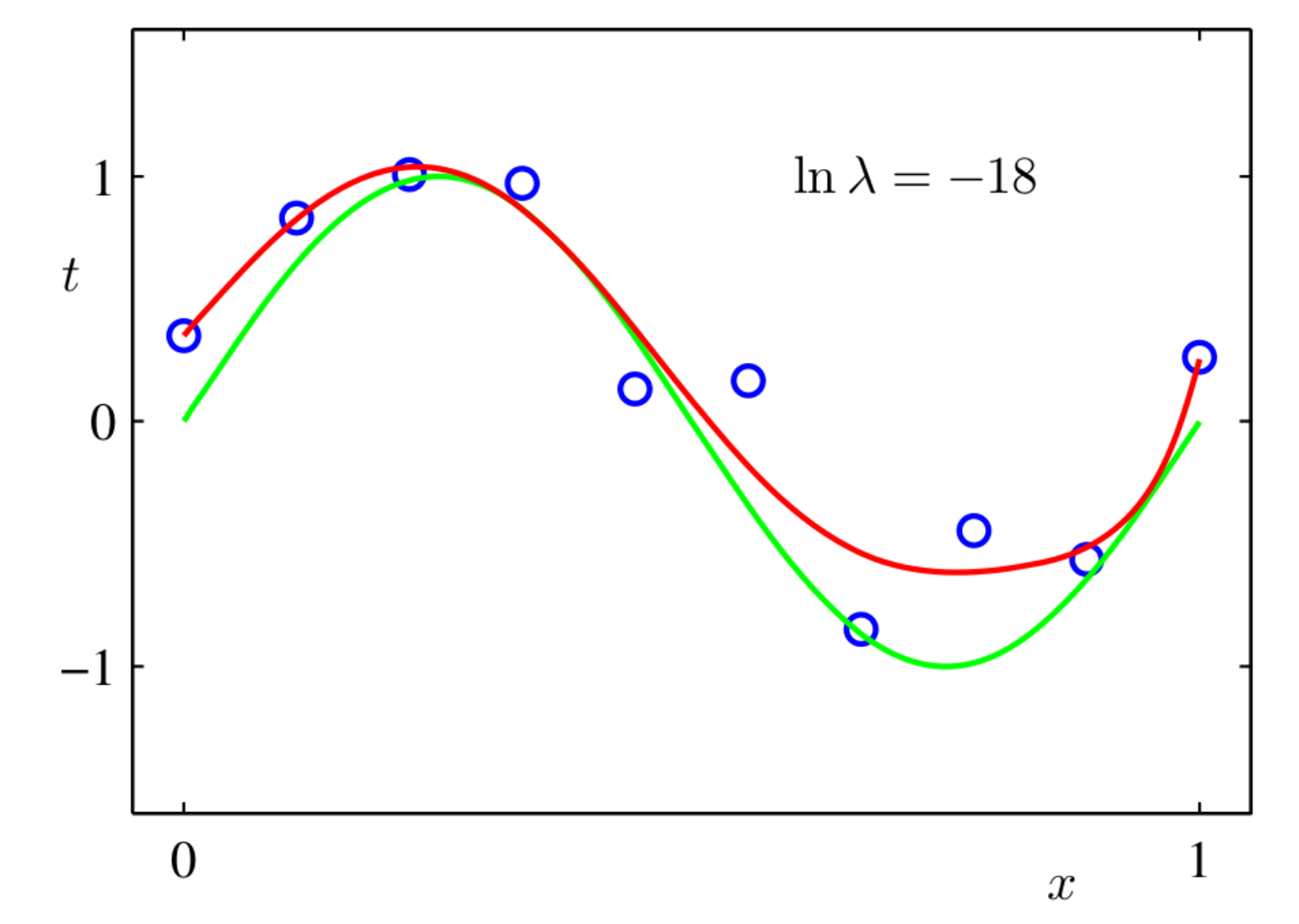
$$w^* = \left(X^\top X + \alpha I\right)^{-1} X^\top y.$$

- Это *гребневая регрессия* (ridge regression); кстати, добавление αI к матрице неполного ранга делает её обратимой; это и есть *регуляризация*, и это и было исходной мотивацией для гребневой регрессии.

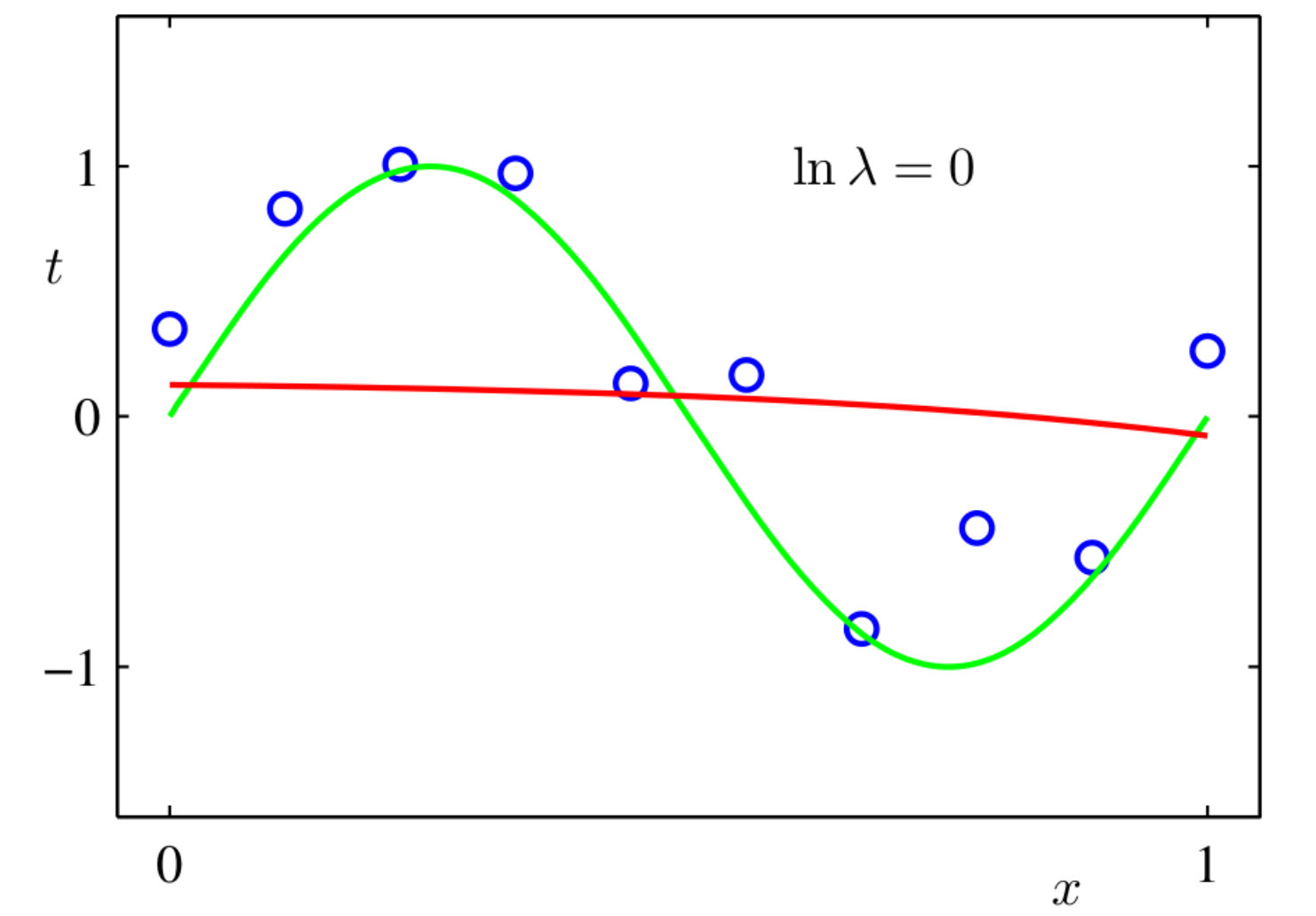
Гребневая регрессия: $\ln \alpha = -\infty$



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



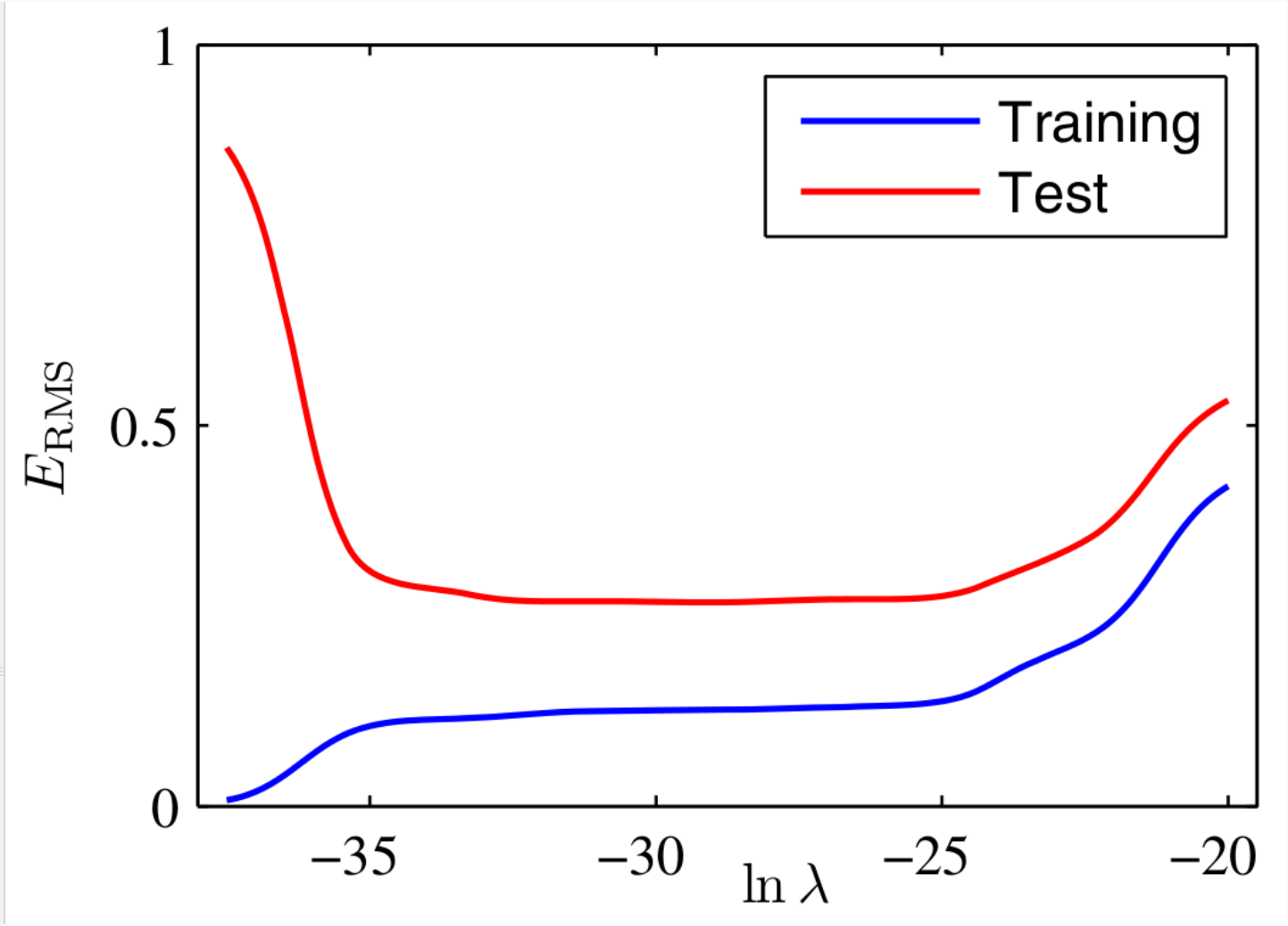
Гребневая регрессия: $\ln \alpha = 0$



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

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

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

Thank you!

Thank you for your attention!