

ЛИНЕЙНАЯ РЕГРЕССИЯ

Сергей Николенко

СПбГУ — Санкт-Петербург

25 сентября 2026 г.

Random facts:



- 25 сентября 1303 г. в Хундуне случилось землетрясение магнитудой около 8 — самое раннее достоверно известное такой силы; погибло не меньше 270 тысяч человек, около 28% населения округа, а афтершоки шли ещё четыре года
- 25 сентября 1789 г. Конгресс США предложил штатам двенадцать поправок к Конституции; ратифицировали десять — Билль о правах
- 25 сентября 1799 г. Суворов во время перехода через Альпы взял штурмом высокогорный Чёртов мост
- 25 сентября 1956 г. заработал TAT-1, первый трансатлантический телефонный кабель: 35 одновременных разговоров на весь океан
- 25 сентября 1983 г. 38 заключённых сбежали из тюрьмы Мейз в Северной Ирландии — крупнейший побег в истории Британии
- 25 сентября 1992 г. NASA запустила Mars Observer; связь пропала за три дня до выхода на орбиту Марса

- В прошлый раз мы разобрали сопряжённые априорные распределения: бета для испытаний Бернулли, Дирихле для мультиномиального — и научились обновлять апостериорное, не беря ни одного интеграла.
- А потом взялись за первую настоящую модель машинного обучения, *линейную регрессию*, но дойти до конца не успели.
- Поэтому сегодня начнём её с начала и пройдем целиком: метод наименьших квадратов, байесовский вывод, оверфиттинг, регуляризация — и, если повезёт, предсказательное распределение.

ЛИНЕЙНАЯ РЕГРЕССИЯ

- Линейная регрессия: рассмотрим линейную функцию

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

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

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

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

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

- Как минимизировать?

- Можно на самом деле решить задачу точно – записать как

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

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: $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ 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?

- 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

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

Иными словами,

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

- Здесь пока y – любая функция.

- 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, $\phi_0(\mathbf{x}) = 1$).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

ПРИМЕР: ПОЛИНОМИАЛЬНАЯ
АППРОКСИМАЦИЯ

- Мы говорили о регрессии с базисными функциями:

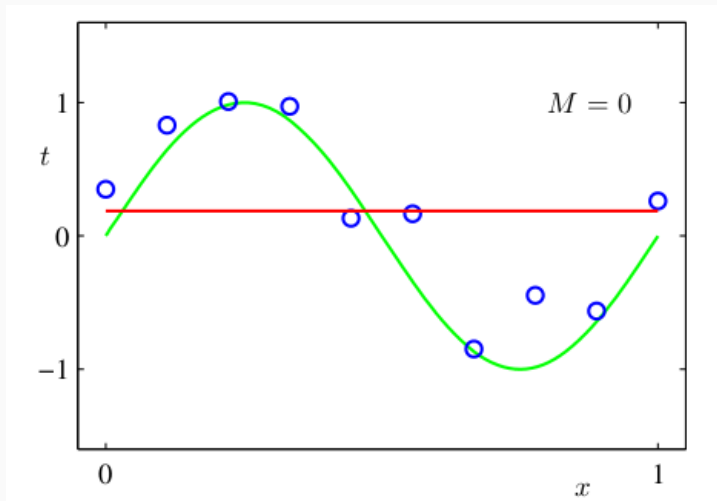
$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^M w_j \phi_j(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x}).$$

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

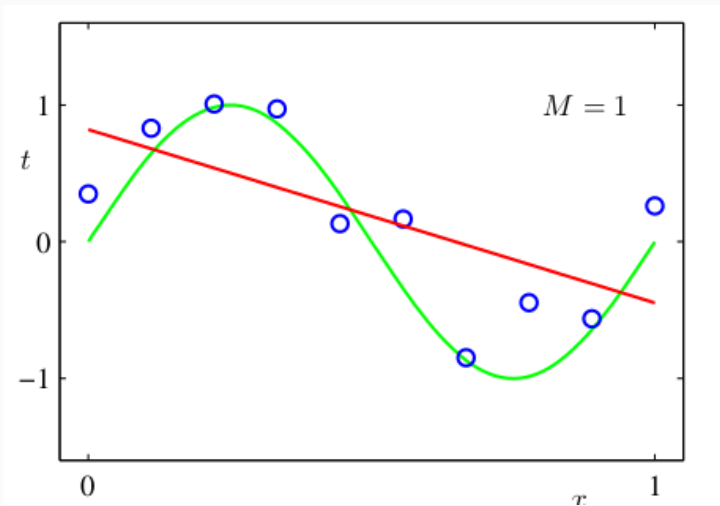
$$f(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \dots + w_M x^M.$$

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

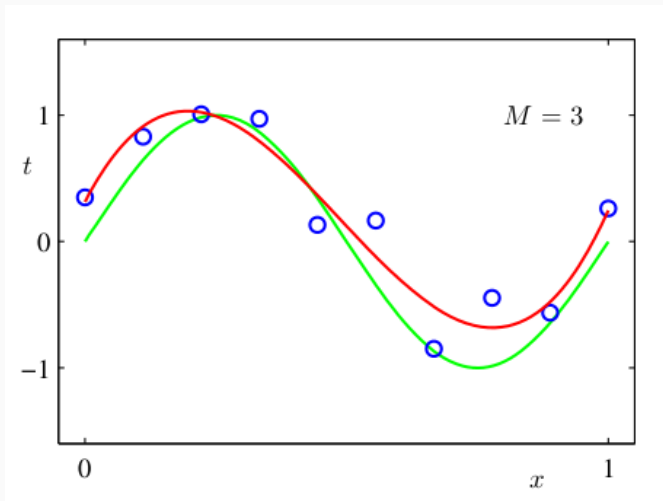
Полиномиальная аппроксимация



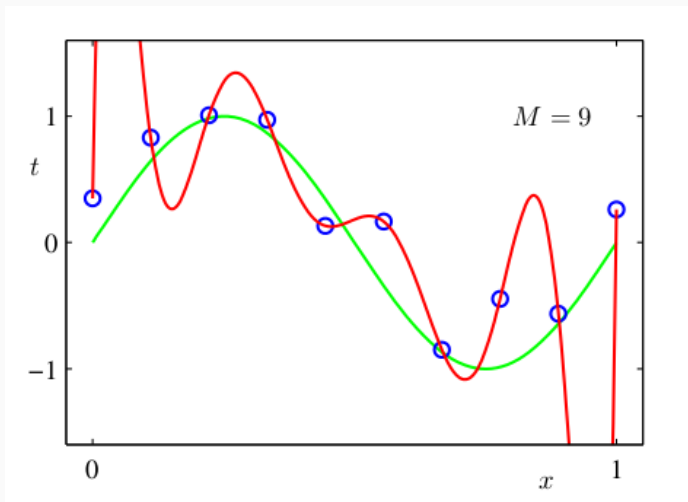
Полиномиальная аппроксимация



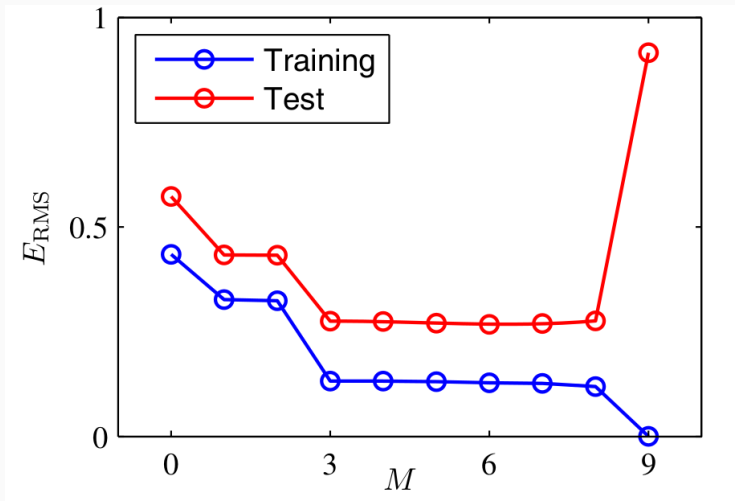
Полиномиальная аппроксимация



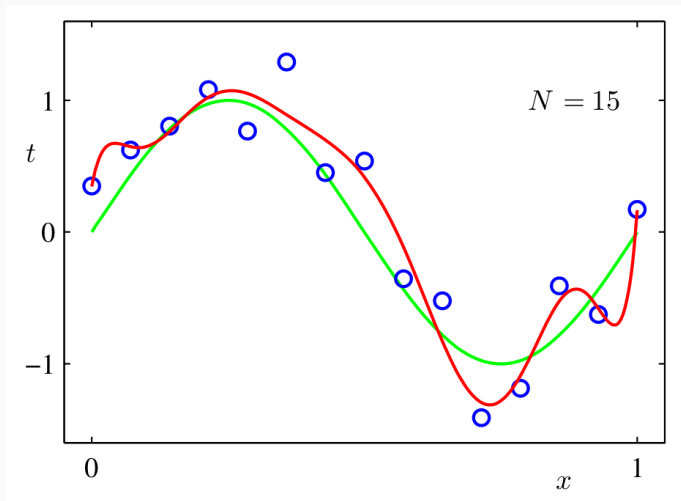
Полиномиальная аппроксимация



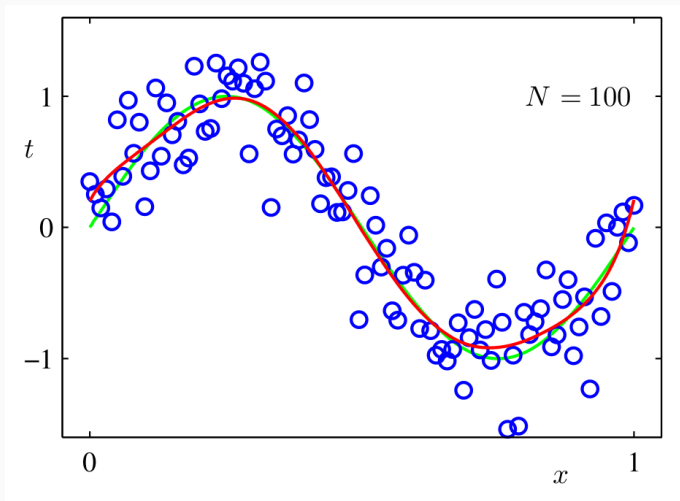
ЗНАЧЕНИЯ RMS



МОЖНО СОБРАТЬ БОЛЬШЕ ДАННЫХ...



МОЖНО СОБРАТЬ БОЛЬШЕ ДАННЫХ...



ЗНАЧЕНИЯ КОЭФФИЦИЕНТОВ

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

РЕГУЛЯРИЗАЦИЯ ПО-БАЙЕСОВСКИ

- 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:

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

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

$$p(\mathbf{w}) = N(\mathbf{w} \mid \mu_0, \Sigma_0).$$

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

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

- Тогда наша задача – посчитать

$$\begin{aligned} p(\mathbf{w} \mid \mathbf{t}) &\propto p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}, \sigma^2) p(\mathbf{w}) \\ &= N(\mathbf{w} \mid \mu_0, \Sigma_0) \prod_{n=1}^N N(t_n \mid \mathbf{w}^\top \phi(\mathbf{x}_n), \sigma^2). \end{aligned}$$

- Давайте подсчитаем.

- Получится

$$\begin{aligned}p(\mathbf{w} \mid \mathbf{t}) &= N(\mathbf{w} \mid \mu_N, \Sigma_N), \\ \mu_N &= \Sigma_N \left(\Sigma_0^{-1} \mu_0 + \frac{1}{\sigma^2} \Phi^\top \mathbf{t} \right), \\ \Sigma_N &= \left(\Sigma_0^{-1} + \frac{1}{\sigma^2} \Phi^\top \Phi \right)^{-1}.\end{aligned}$$

- Теперь давайте подсчитаем логарифм правдоподобия.

- Если мы возьмём априорное распределение около нуля:

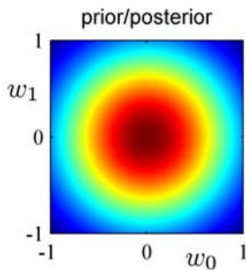
$$p(\mathbf{w}) = N(\mathbf{w} \mid 0, \frac{1}{\alpha} \mathbf{I}),$$

то логарифм правдоподобия получится

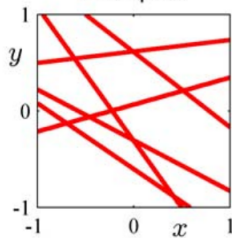
$$\ln p(\mathbf{w} \mid \mathbf{t}) = -\frac{1}{2\sigma^2} \sum_{n=1}^N (t_n - \mathbf{w}^\top \phi(\mathbf{x}_n))^2 - \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w} + \text{const},$$

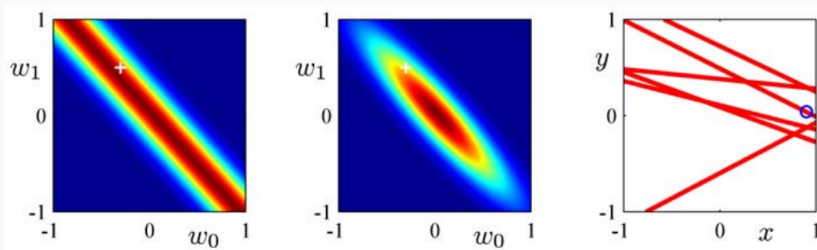
то есть в точности гребневая регрессия.

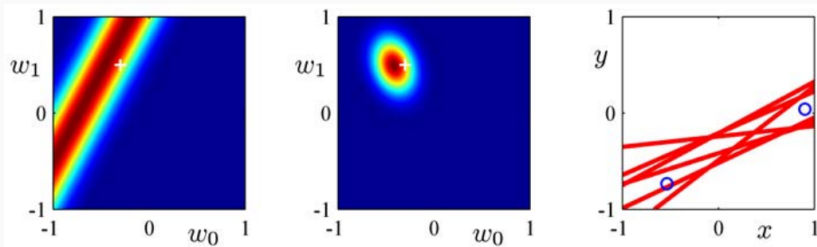
likelihood



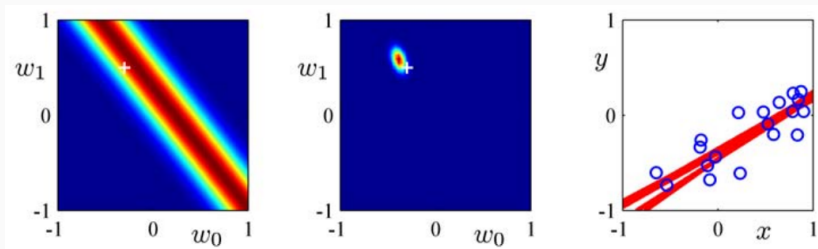
data space







ПРИМЕР



- Можно слегка обобщить – рассмотреть априорное распределение более общего вида

$$p(\mathbf{w} \mid \alpha) = \left[\frac{q}{2} \left(\frac{\alpha}{2} \right)^{1/q} \frac{1}{\Gamma(1/q)} \right]^M e^{-\frac{\alpha}{2} \sum_{j=1}^M |w_j|^q}.$$

Упражнение. Подсчитайте логарифм правдоподобия.

- 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»?

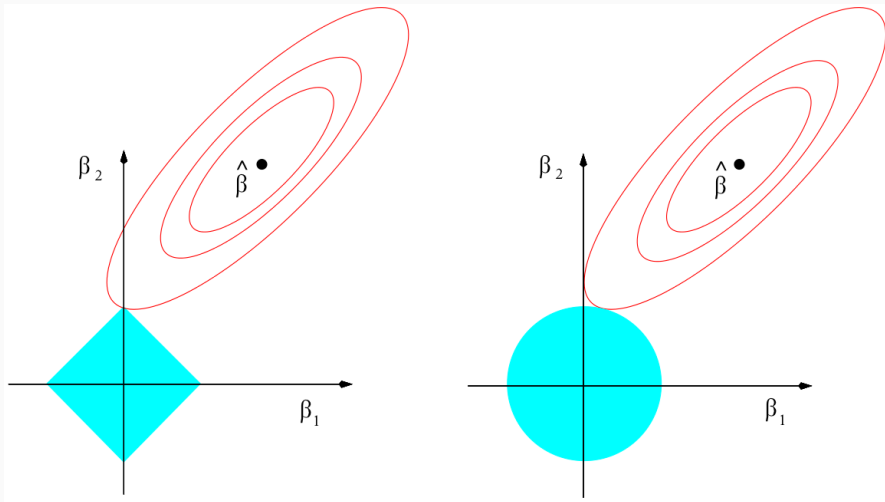
- Мы можем переписать регрессию с регуляризатором по-другому:

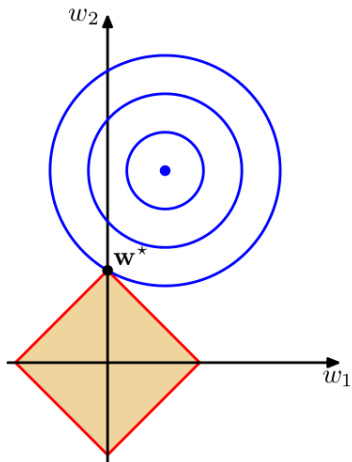
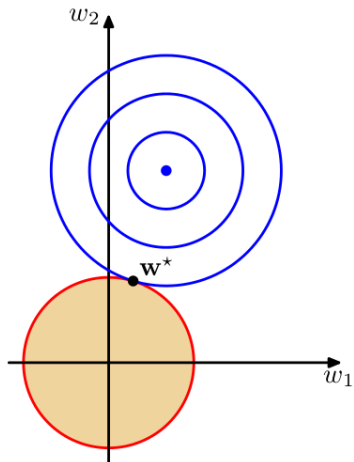
$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \left\{ \frac{1}{2} \sum_{i=1}^N (f(x_i, \mathbf{w}) - y_i)^2 + \lambda \sum_{j=0}^p |w_j| \right\},$$

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.



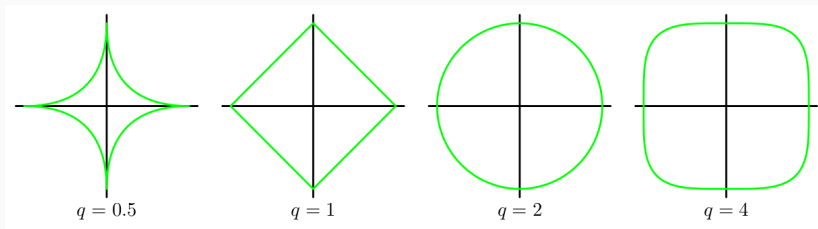
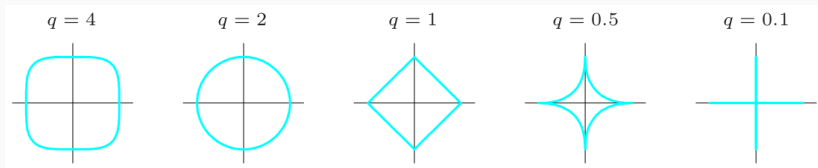


- Можно рассмотреть обобщение гребневой и лассо-регрессии:

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

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

РАЗНЫЕ q



ПРЕДСКАЗАНИЯ В ЛИНЕЙНОЙ РЕГРЕССИИ

- Теперь давайте вернёмся к байесовской постановке:

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:

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

- В прошлый раз мы нашли апостериорное распределение: для гауссовского априорного

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

мы нашли

$$\begin{aligned} p(\mathbf{w} \mid \mathbf{t}, \alpha, \beta) &= N(\mathbf{w} \mid \mu_N, \Sigma_N), \\ \mu_N &= \Sigma_N \left(\Sigma_0^{-1} \mu_0 + \beta \Phi^\top \mathbf{t} \right), \\ \Sigma_N &= \left(\Sigma_0^{-1} + \beta \Phi^\top \Phi \right)^{-1}, \end{aligned}$$

где $\beta = \frac{1}{\sigma^2}$ (precision нормального распределения).

- Теперь сделаем следующий шаг – найдём апостериорное распределение наших предсказаний

$$p(t \mid \mathbf{t}, \alpha, \beta) = \int p(t \mid \mathbf{w}, \beta) p(\mathbf{w} \mid \mathbf{t}, \alpha, \beta) d\mathbf{w}.$$

- Это свёртка двух гауссианов, и получается...

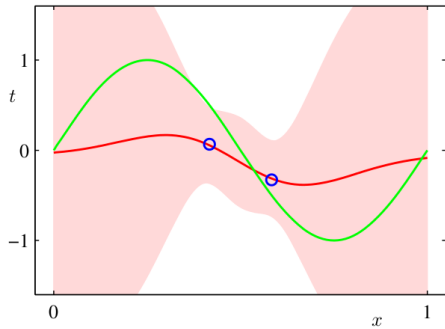
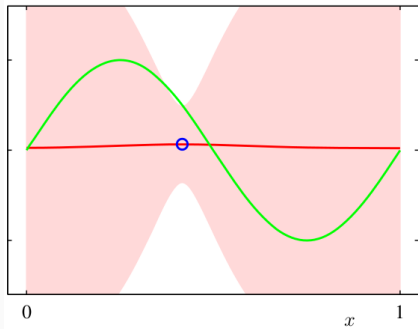
- ...тоже гауссиан:

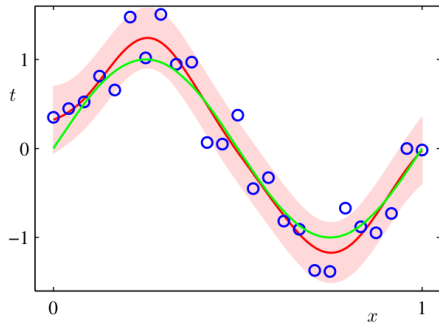
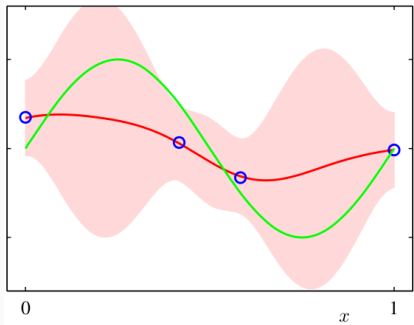
$$p(t \mid \mathbf{t}, \alpha, \beta) = N(t \mid \mu_N^\top \phi(\mathbf{x}), \sigma_N^2),$$

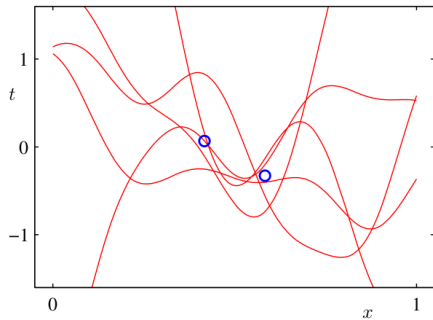
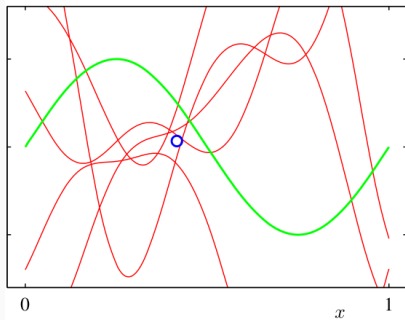
$$\text{где } \sigma_N^2 = \frac{1}{\beta} + \phi(\mathbf{x})^\top \Sigma_N \phi(\mathbf{x}).$$

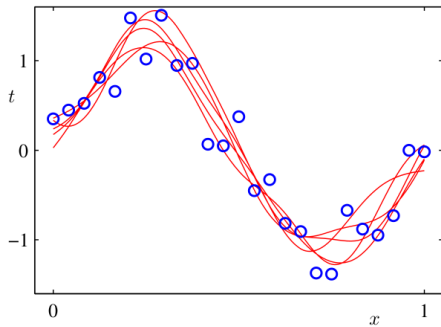
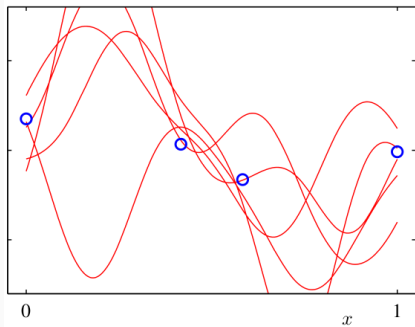
- Т.е. дисперсия складывается из шума в данных β и дисперсии параметров $\mathbf{w}$; гауссианы независимы, и их дисперсии просто складываются.

Упражнение. Оценка всё время уточняется: $\sigma_{N+1}^2 \leq \sigma_N^2$.









СПАСИБО!

Спасибо за внимание!

