

# ЛИНЕЙНАЯ РЕГРЕССИЯ

---

Сергей Николенко

СПбГУ — Санкт-Петербург

08 сентября 2017 г.

---

*Random facts:*

- 8 сентября --- Международный день грамотности, установленный ООН, а также День финансиста России
- 8 сентября 1636 г. был основан Гарвардский университет

# РЕГУЛЯРИЗАЦИЯ В ЛИНЕЙНОЙ РЕГРЕССИИ

---

- 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.

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

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

- Стало:

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

где  $\alpha$  – коэффициент регуляризации (его надо будет как-нибудь выбрать).

- Как оптимизировать эту функцию ошибки?

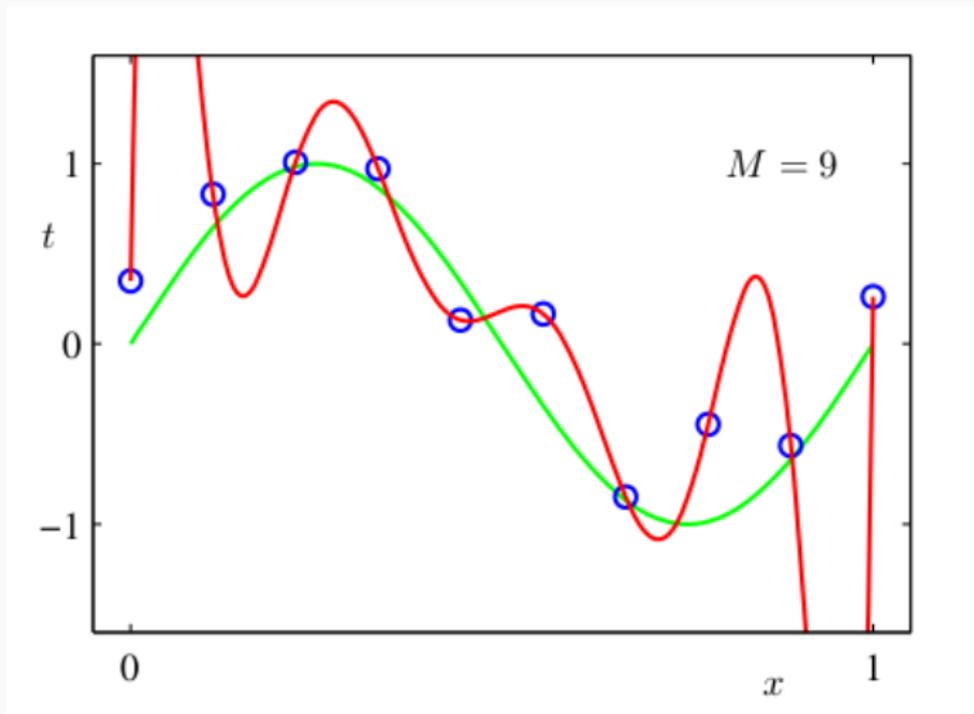
- Да точно так же – запишем как

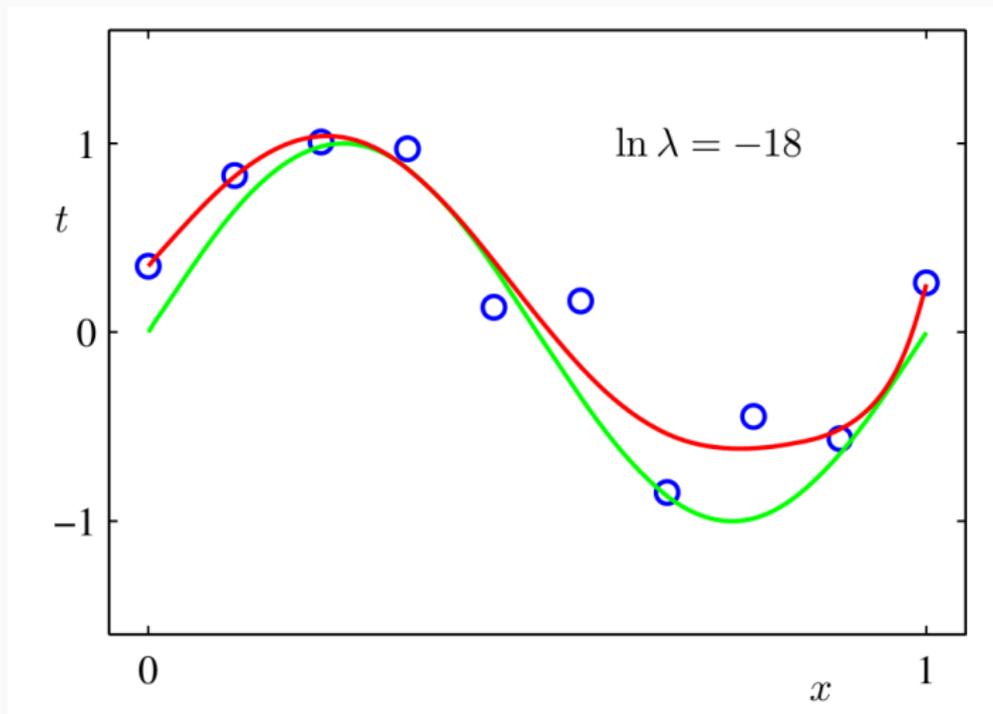
$$\text{RSS}(\mathbf{w}) = \frac{1}{2} (\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w}) + \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w}$$

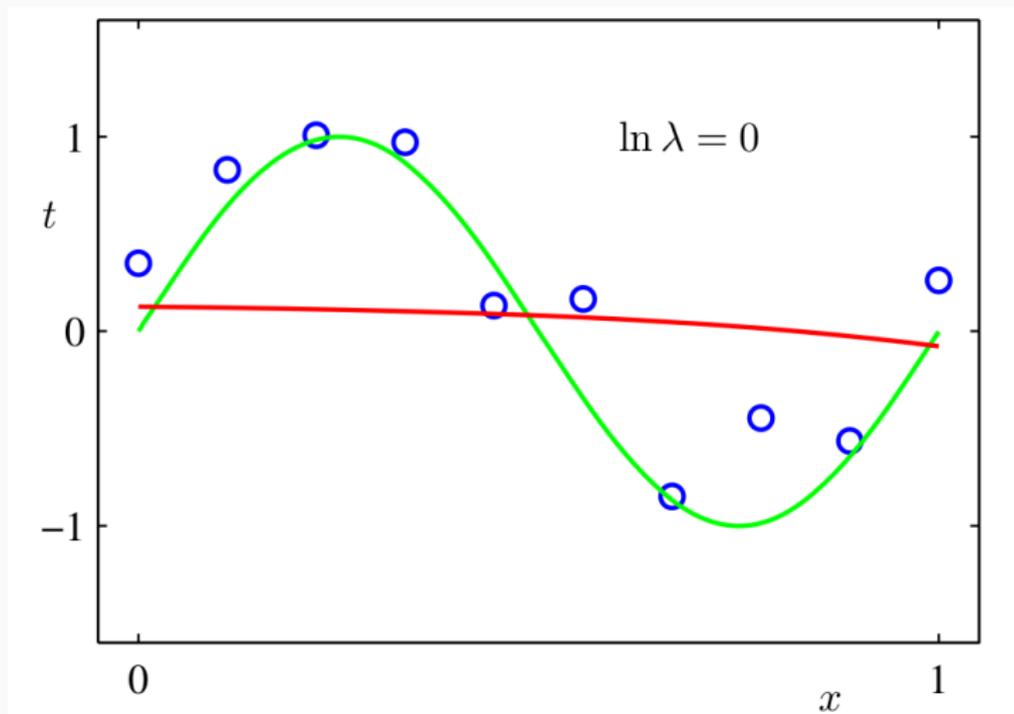
и возьмём производную; получится

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

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

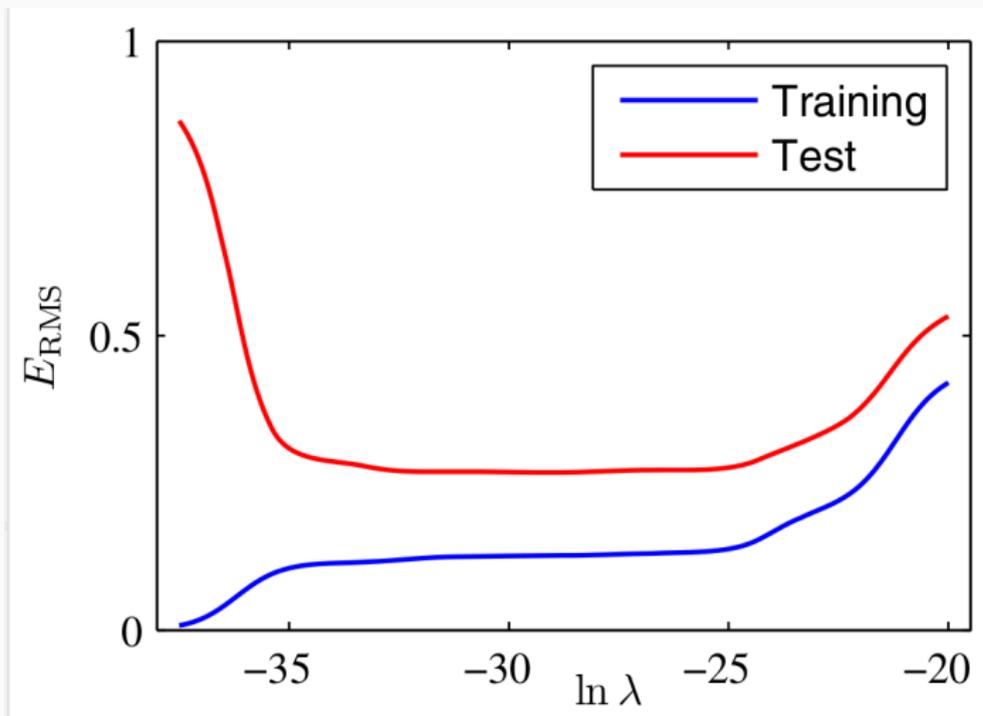






## ГРЕБНЕВАЯ РЕГРЕССИЯ: КОЭФФИЦИЕНТЫ

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



- 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  $L_2$ :

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

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

# РЕГРЕССИЯ ПО-БАЙЕСОВСКИ

---

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

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

(и/или найти максимальную апостериорную гипотезу  $\arg \max_{\theta} p(\theta | D)$ );

2. найти апостериорное распределение исходов дальнейших экспериментов:

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

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

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

- 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:

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

- Тогда наша задача – посчитать

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

- Давайте подсчитаем.

- Получится

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

- Теперь давайте подсчитаем логарифм правдоподобия.

- Если мы возьмём априорное распределение около нуля:

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

то логарифм правдоподобия получится

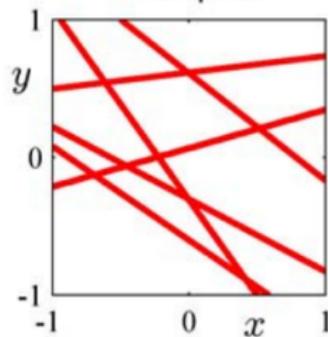
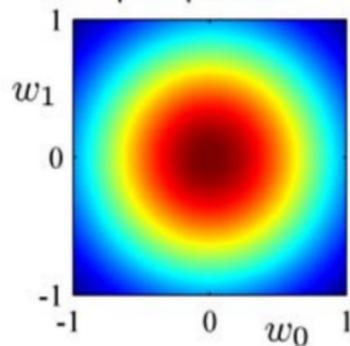
$$\ln p(\mathbf{w} \mid \mathbf{t}) = -\frac{1}{2\sigma^2} \sum_{n=1}^N (t_n - \mathbf{w}^\top \phi(\mathbf{x}_n))^2 - \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w} + \text{const},$$

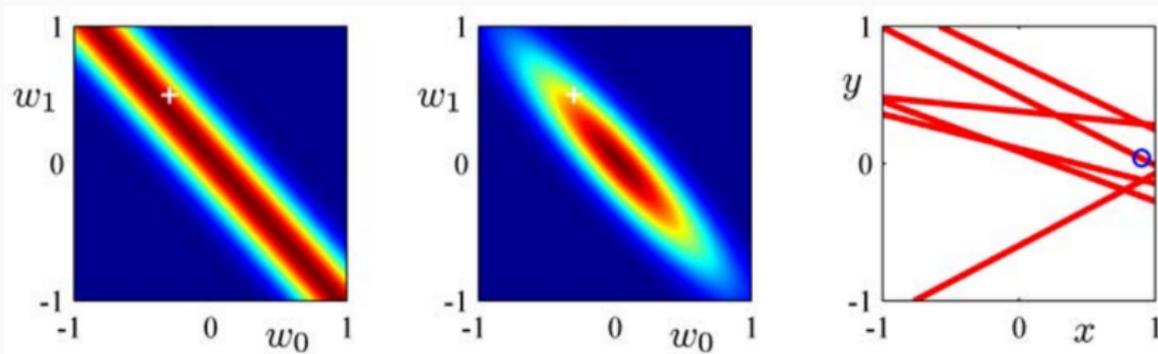
то есть в точности гребневая регрессия.

likelihood

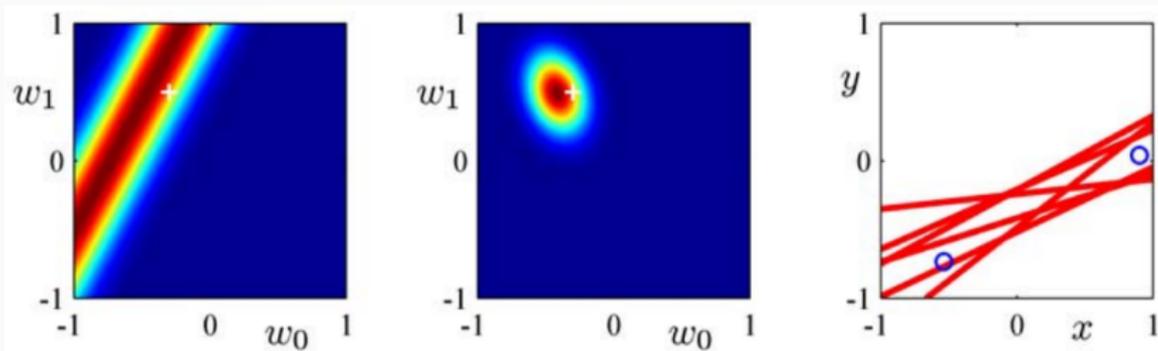
prior/posterior

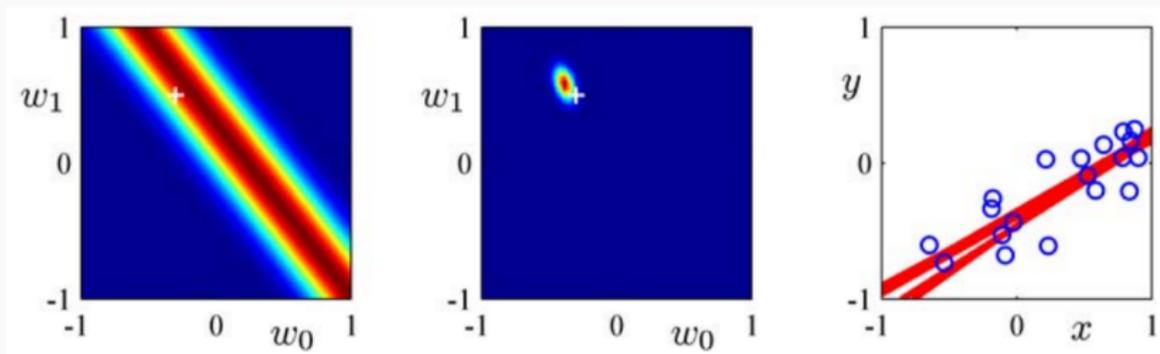
data space





# 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.

Спасибо за внимание!