БАЙЕСОВСКИЕ ПРЕДСКАЗАНИЯ

Сергей Николенко

СПбГУ — Санкт-Петербург
15 сентября 2017 г.

Random facts:

- 15 сентября --- Международный день демократии, который ООН отмечает «с целью повышения информирования общественности», а также Международный день распространения информации о лимфоме
- 15 сентября 1792 г. Наполеон прибыл в Аяччо и возглавил батальон волонтёров, 15 сентября 1805 г. Россия объявила Наполеону первую войну, а 15 сентября 1812 г. Наполеон вошёл в Московский Кремль

Ещё раз о регуляризации

- Мы знаем, что наименьшие квадраты не всегда хорошо работают. Две причины:
  1. плохая предсказательная сила – часто лучше регуляризовать, пожертвовав bias'ом в пользу variance;
  2. сложности в интерпретации – хотелось бы понимать, что происходит, если переменных с ненулевыми коэффициентами слишком много, не получится.
- Мораль: хотелось бы сделать так, чтобы было поменьше ненулевых компонент в векторе $\mathbf{w}$.

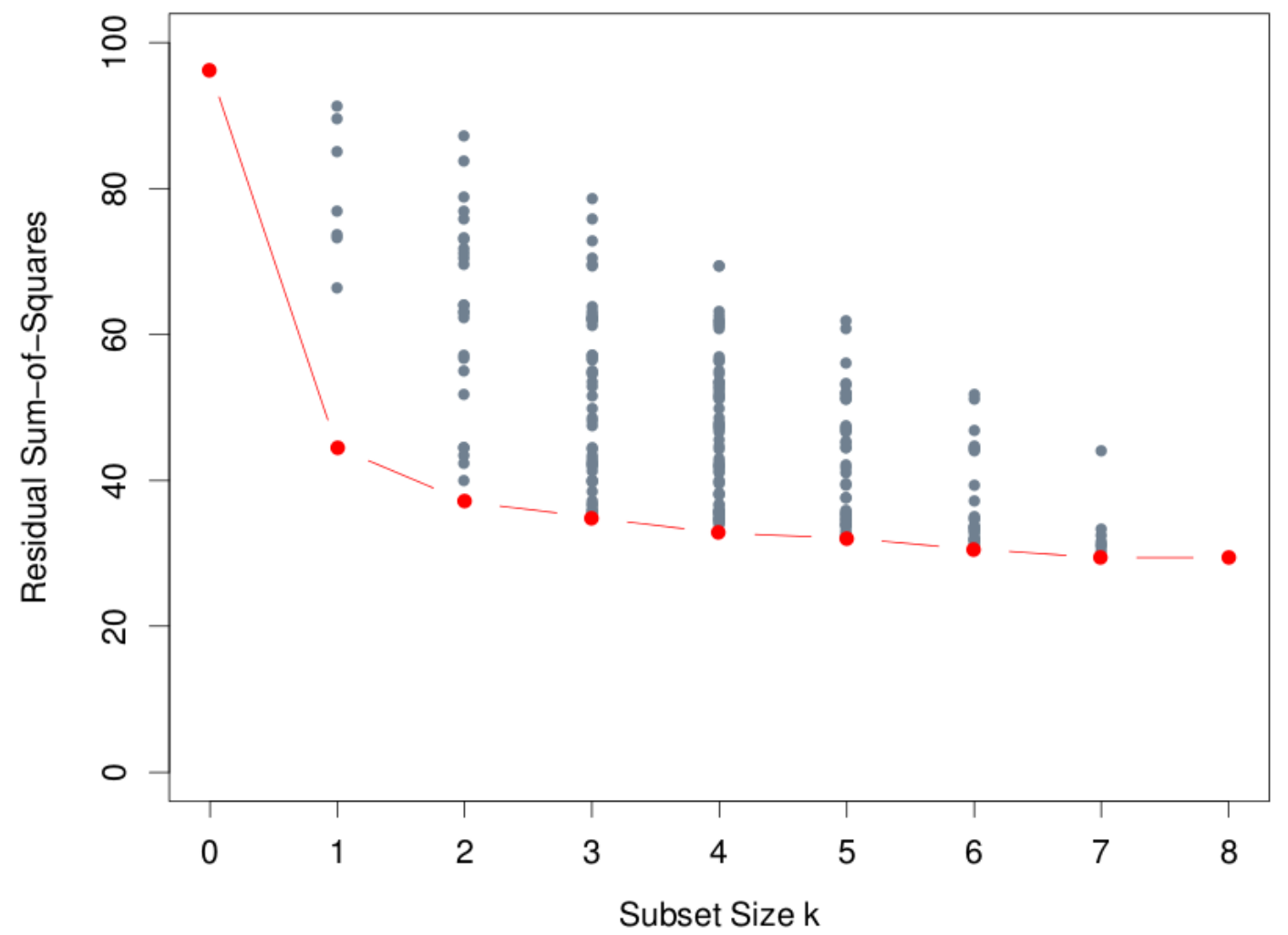
SUBSET SELECTION

- Может быть, давайте так и сделаем? Будем искать самые лучшие компоненты и делать их ненулевыми.
- Это называется subset selection.
- Можно просто делать best subset selection: выбирать подмножество из k входных переменных, которые дают самые лучшие результаты.

SUBSET SELECTION

- Это долго, даже если делать с умом, потому что subsets много.
- Forward-stepwise selection: начинаем со свободного члена, потом добавляет на каждом шаге предиктор, который максимально уменьшает ошибку.
- Т.е. подмножества тут получаются вложенные.
- Backward-stepwise selection: начинаем с полной регрессии и на каждом шаге убираем предиктор, который оказывает меньше всего влияния на ошибку.

SUBSET SELECTION



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:

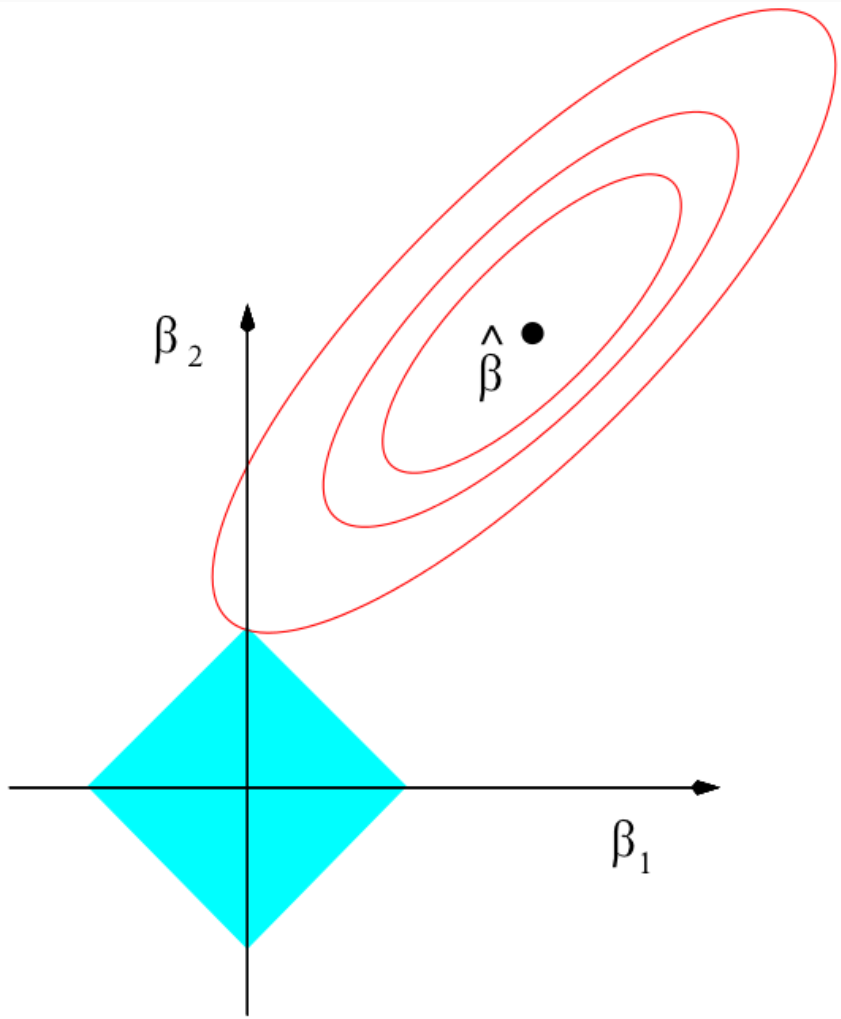
$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left\{ \frac{1}{2} \sum_{i=1}^{N} (f(x_i, \mathbf{w}) - y_i)^2 + \lambda \sum_{j=0}^{p} |w_j| \right\},$$

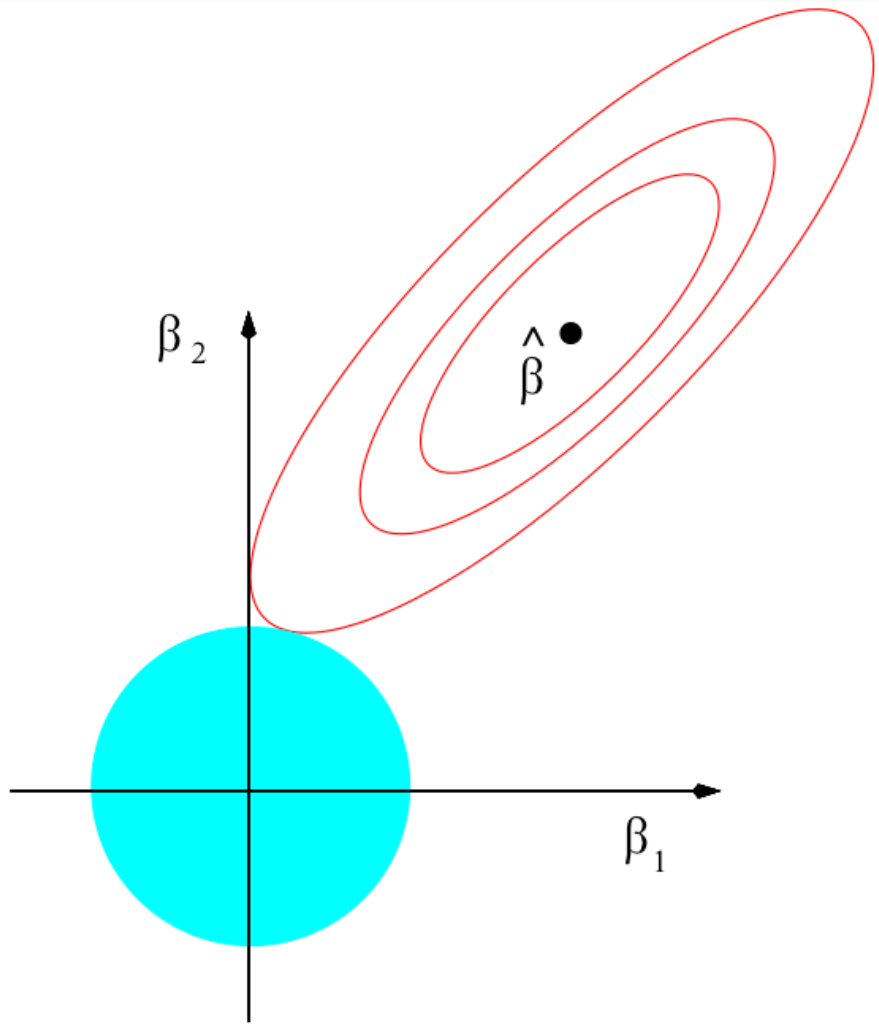
эквивалентно

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

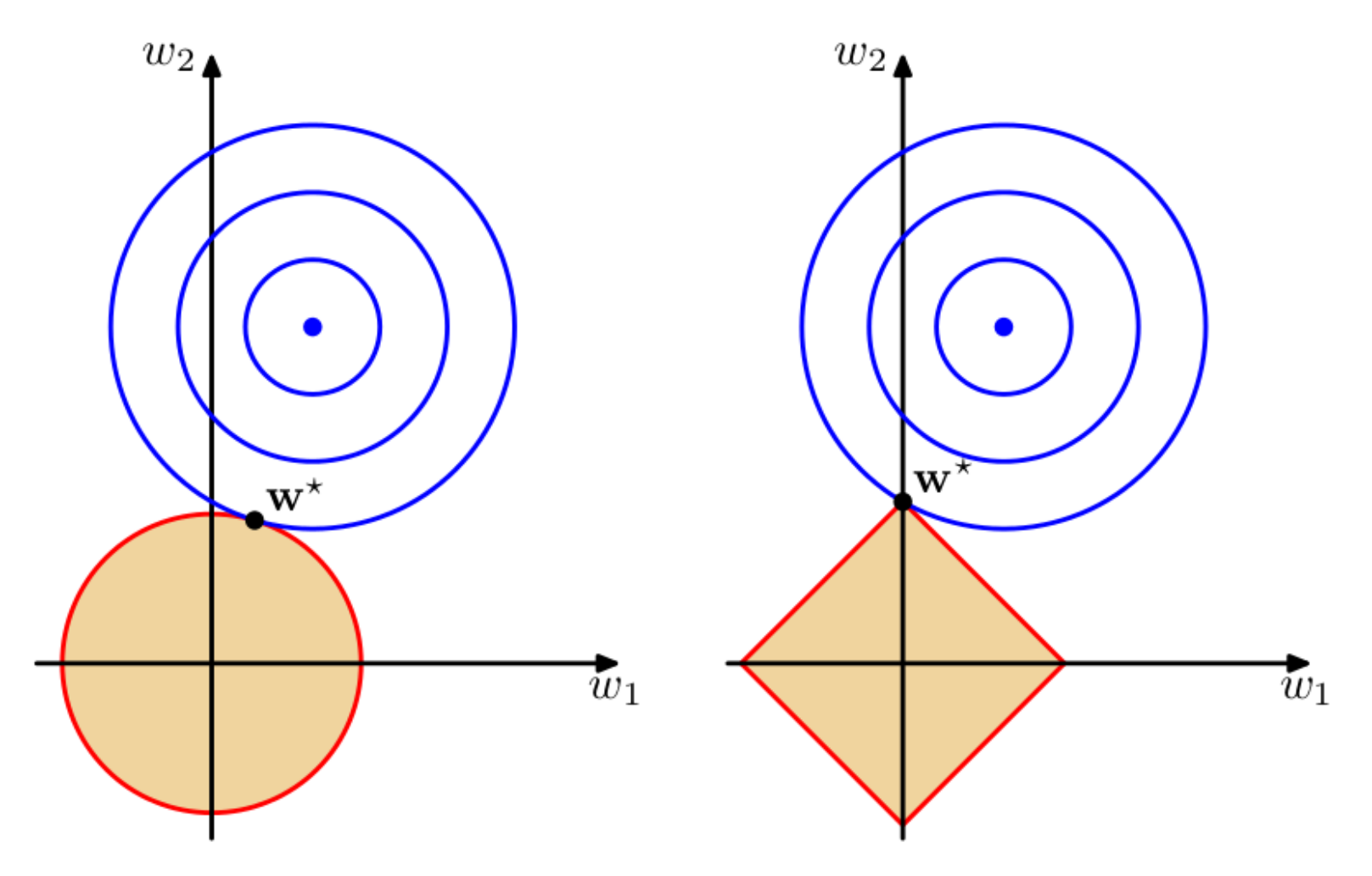
Упражнение. Докажите это.

ГРЕБЕНЬ И ЛАССО





ГРЕБЕНЬ И ЛАССО



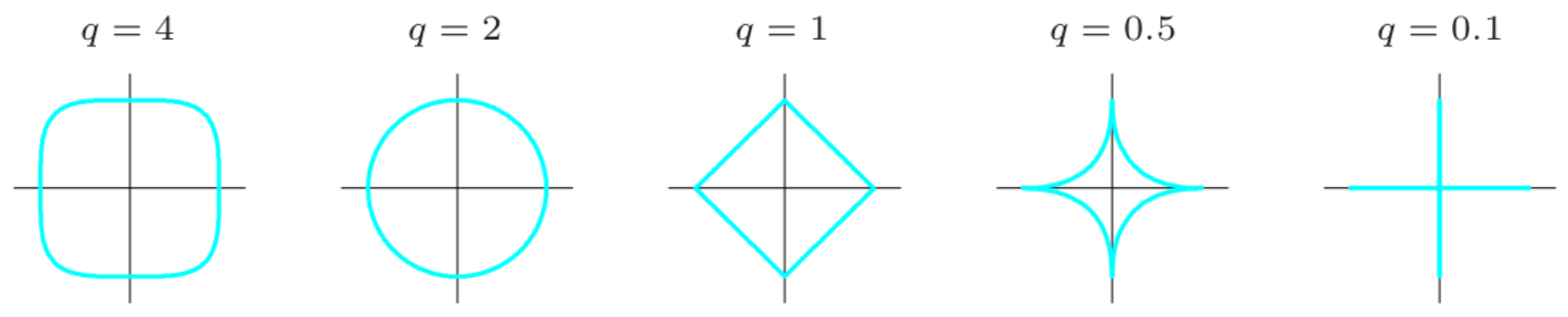
ОБОБЩЕНИЕ

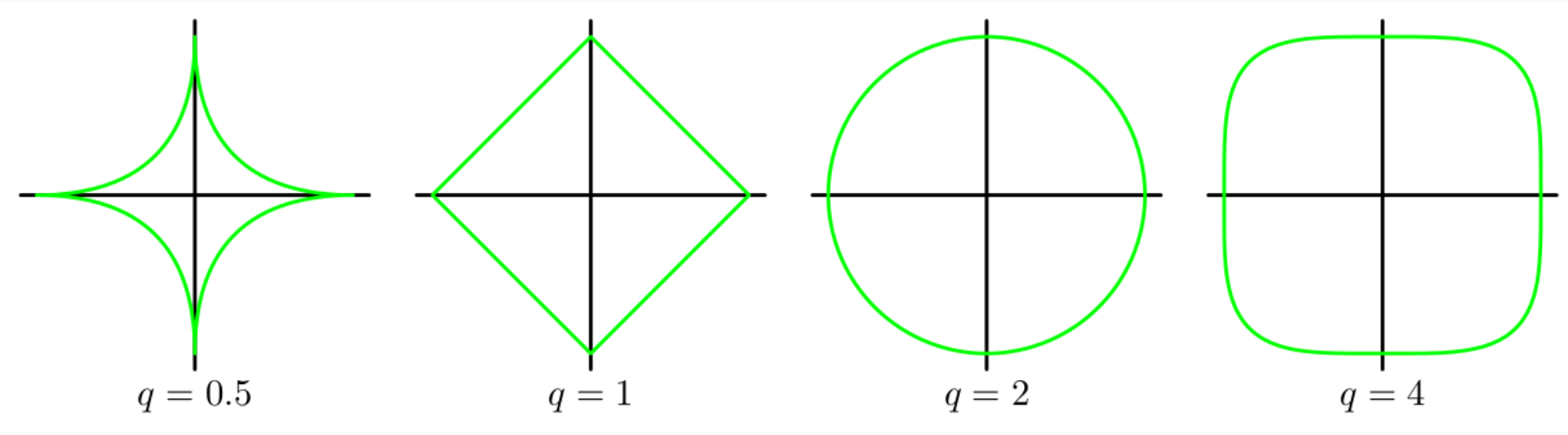
- Можно рассмотреть обобщение гребневой и лассо-регрессии:

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

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

РАЗНЫЕ q





ПРЕДСКАЗАНИЯ В ЛИНЕЙНОЙ РЕГРЕССИИ

Предсказание 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:

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

ПРЕДСКАЗАНИЕ В ЛИНЕЙНОЙ РЕГРЕССИИ

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

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

мы нашли

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

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

ПРЕДСКАЗАНИЕ В ЛИНЕЙНОЙ РЕГРЕССИИ

- 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...

Предсказание в линейной регрессии

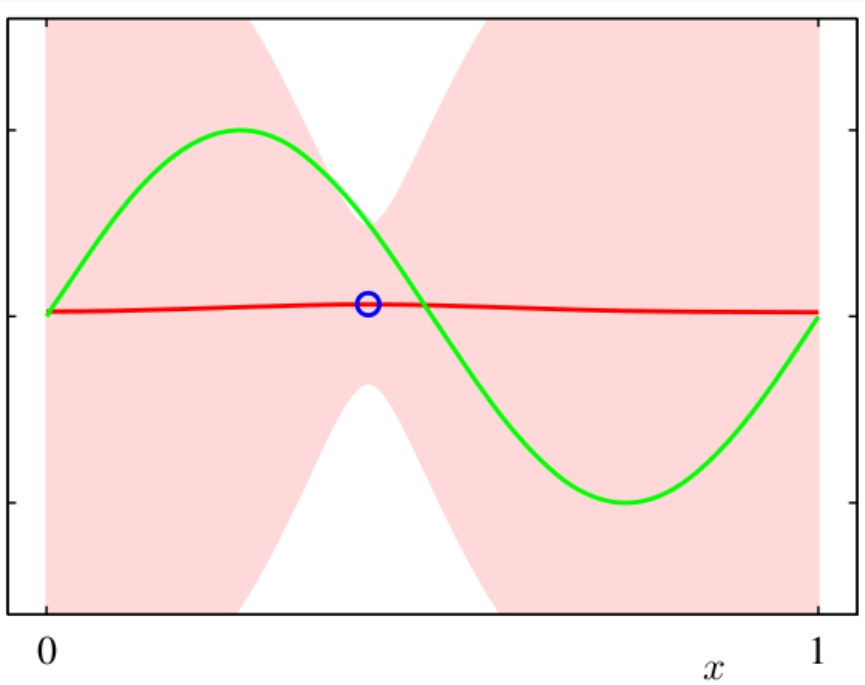
- ...тоже гауссиан:

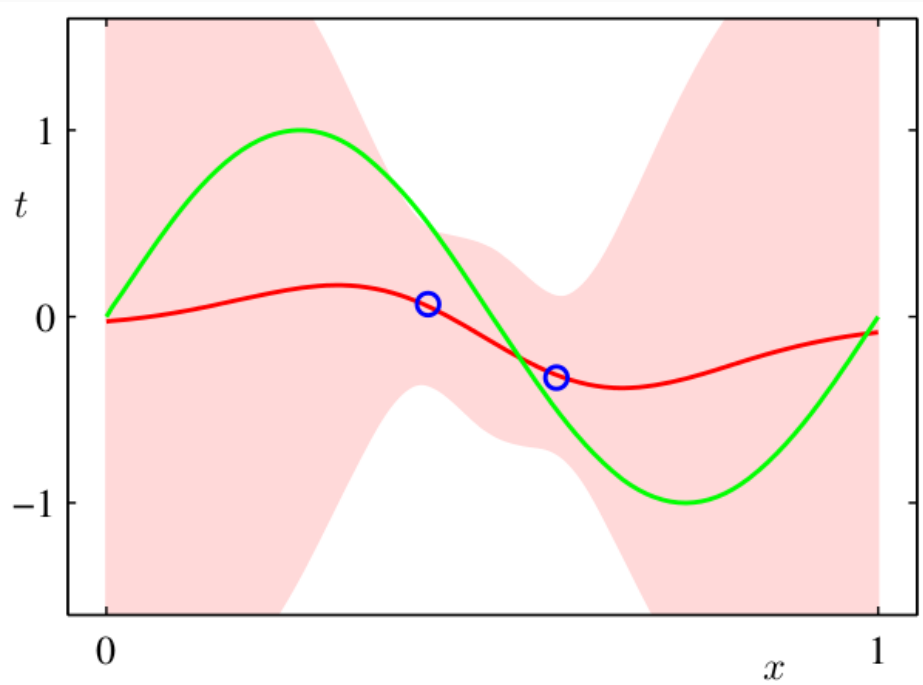
$$p(t \mid \mathbf{t}, \alpha, \beta) = \mathcal{N}(t \mid \mu_N^\top \phi(\mathbf{x}), \sigma_N^2),$$
$$\text{где } \sigma_N^2 = \frac{1}{\beta} + \phi(\mathbf{x})^\top \Sigma_N \phi(\mathbf{x}).$$

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

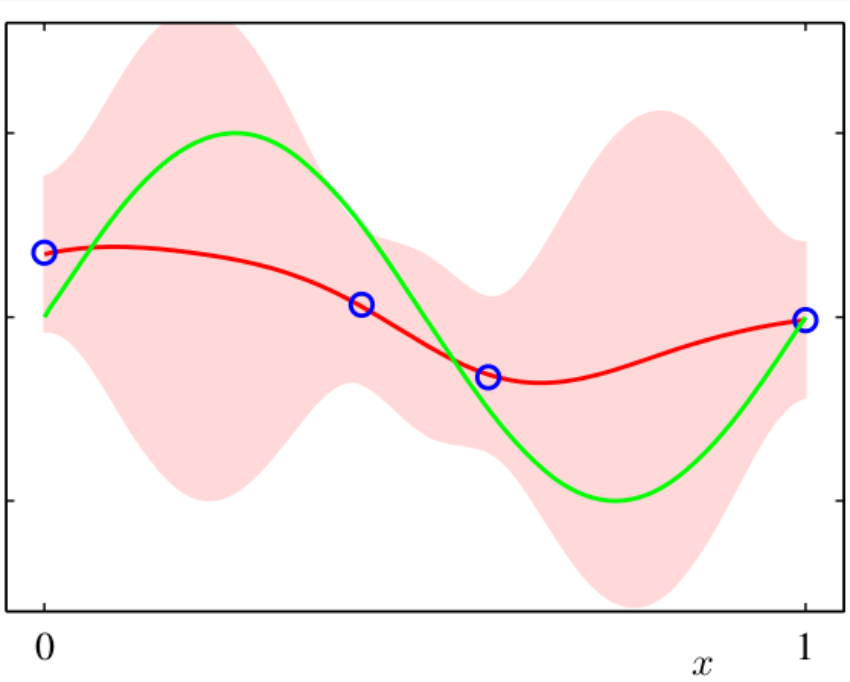
Упражнение. Оценка всё время уточняется: $\sigma_{N+1}^2 \leq \sigma_N^2$.

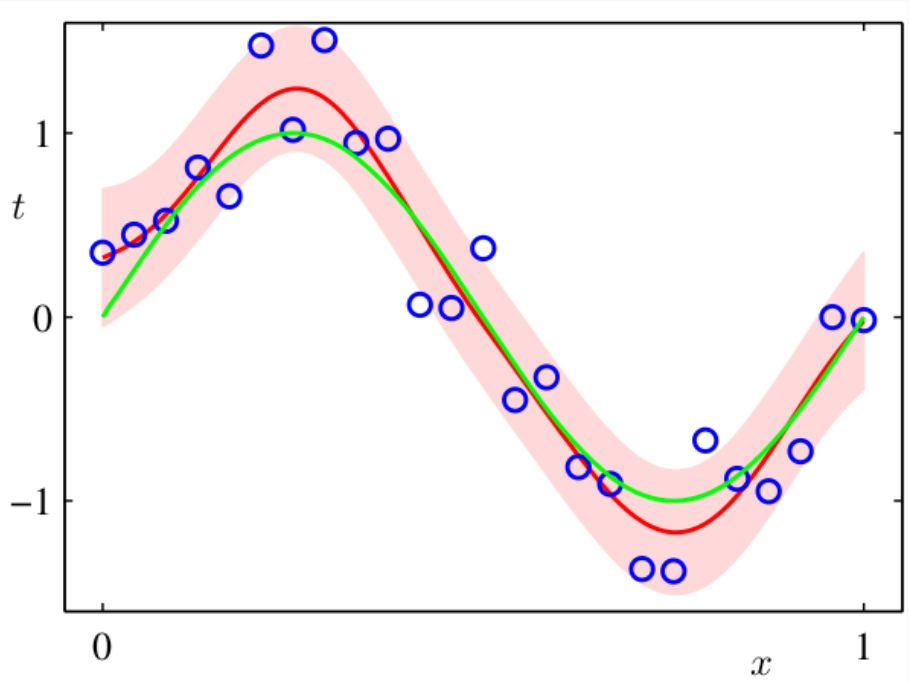
ПРЕДСКАЗАНИЯ



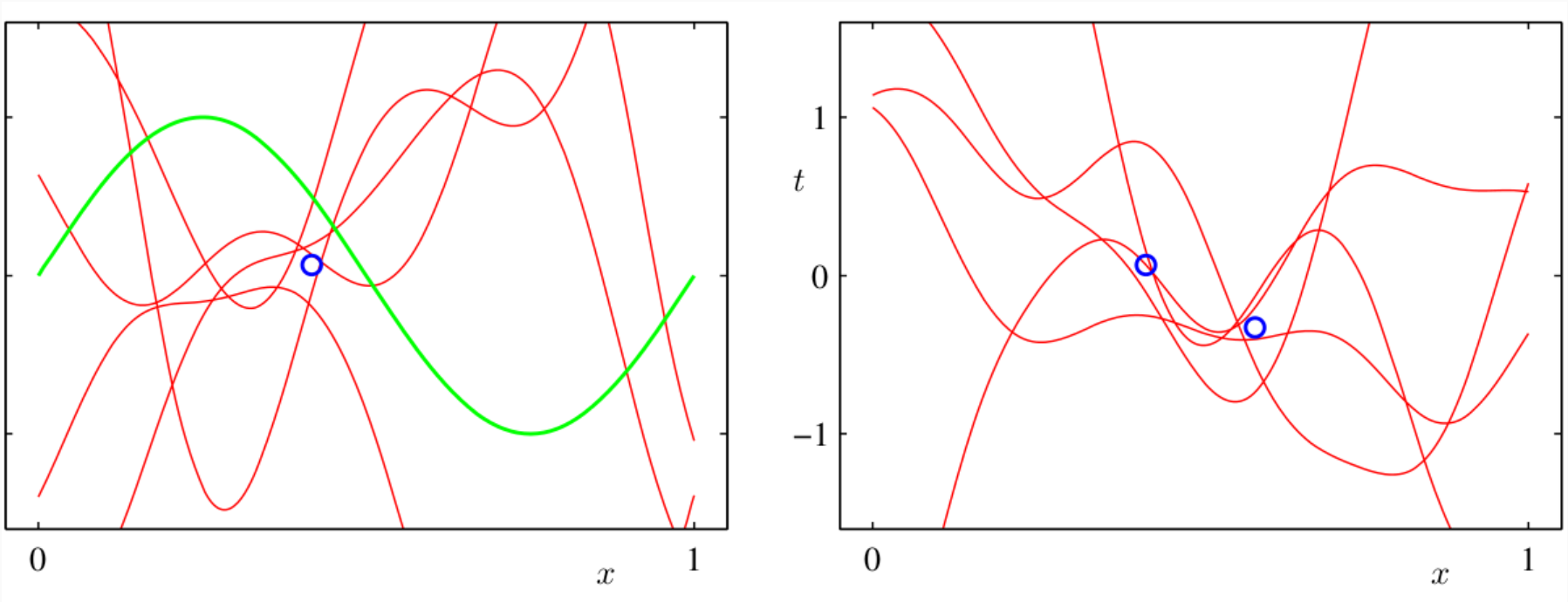


ПРЕДСКАЗАНИЯ

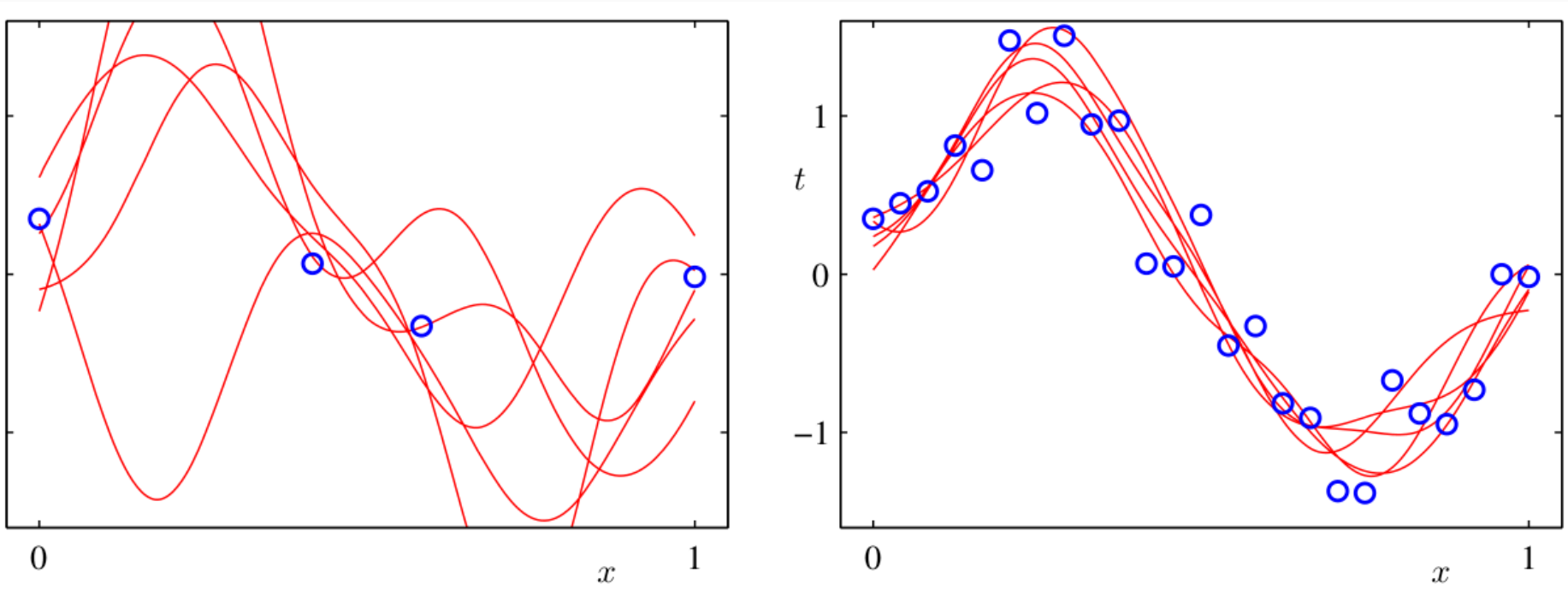




ПРЕДСКАЗАНИЯ



ПРЕДСКАЗАНИЯ



СПАСИБО!

Спасибо за внимание!