Линейная регрессия по-байесовски

Сергей Николенко

Samsung AI Center — Москва

30 марта 2020 г.

Random facts:

- 30 марта 1699 г. гуру Гобинд Сингх основал братство хальсы, ввел традицию пяти К — кеш (нетронутые волосы), канга (гребень), кара (стальной браслет), какча (трусы до колен), кирпан (кинжал под одеждой) — и тем самым основал современный сикхизм
- 30 марта 1847 г. Лев Толстой начал вести дневник; он делал это всю жизнь, а в его полном собрании сочинений дневники занимают 13 томов
- 30 марта 1867 г. в Вашингтоне был подписан договор, по которому США приобрели у России Аляску с Алеутскими островами за 7,2 млн долларов
- 30 марта 1885 г. произошёл бой при Кушке, в результате которого Великобритания чуть было не начала войну с Россией; но всё же договорились, и Кушка стала самым южным населённым пунктом Российской империи, а затем и СССР
- 30 марта 1934 г. в Мурманске была впервые проведена Полярная Олимпиада (или Праздник Севера); сначала соревновались только лыжники, потом добавили оленьи упряжки, горные лыжи, хоккей с мячом, парашютный спорт, биатлон, в наше время — скалолазание, зимний виндсёрфинг, футбол на снегу, натурбан и многое другое

Регуляризация 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:

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

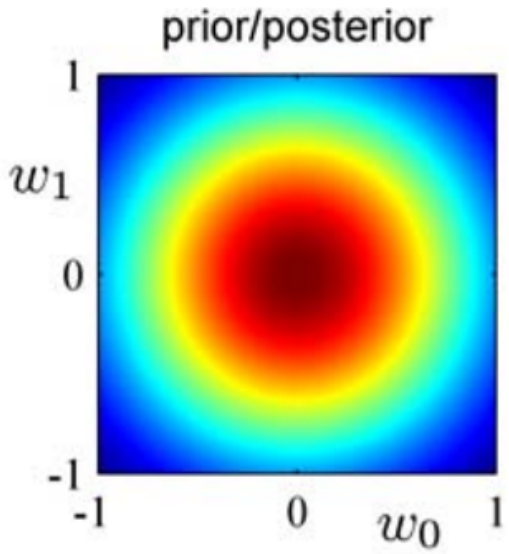
то логарифм правдоподобия получится

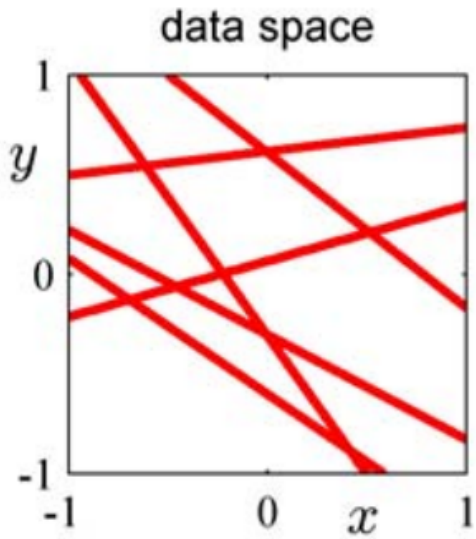
$$\ln p(\mathbf{w} \mid \mathbf{t}) = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} \left(t_n - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n)\right)^2 - \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w} + \text{const},$$

то есть в точности гребневая регрессия.

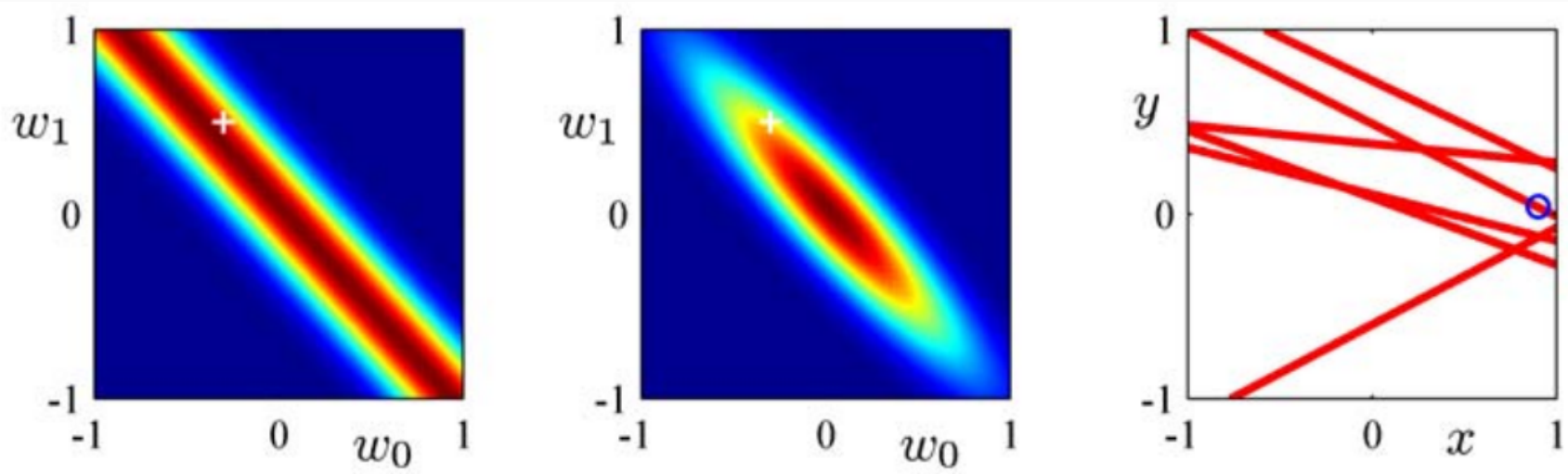
Пример

likelihood

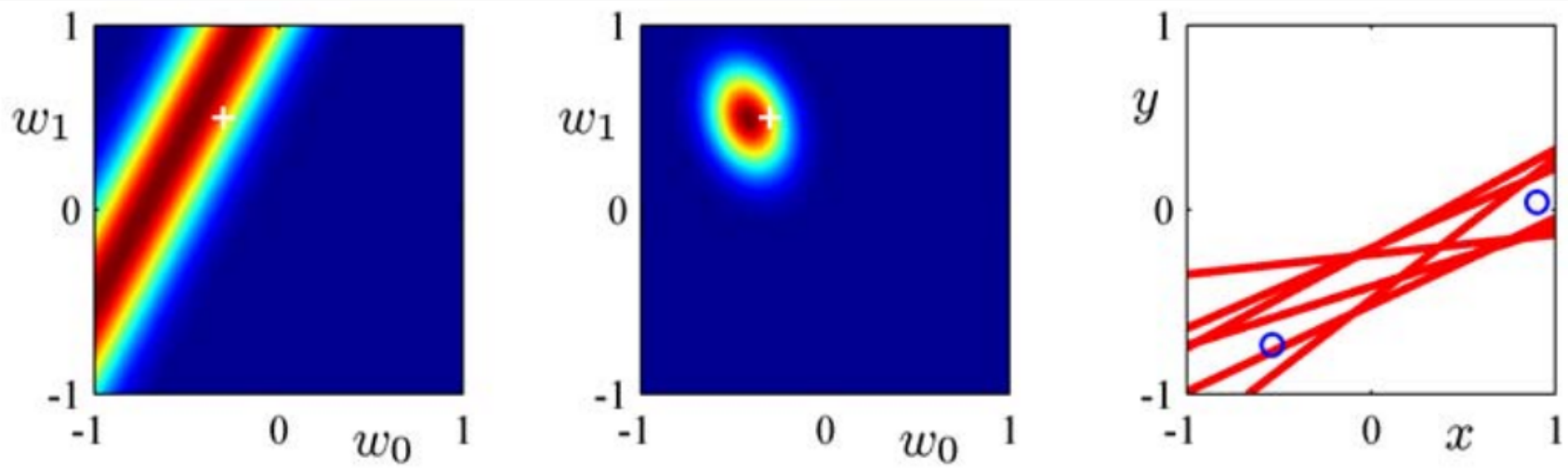




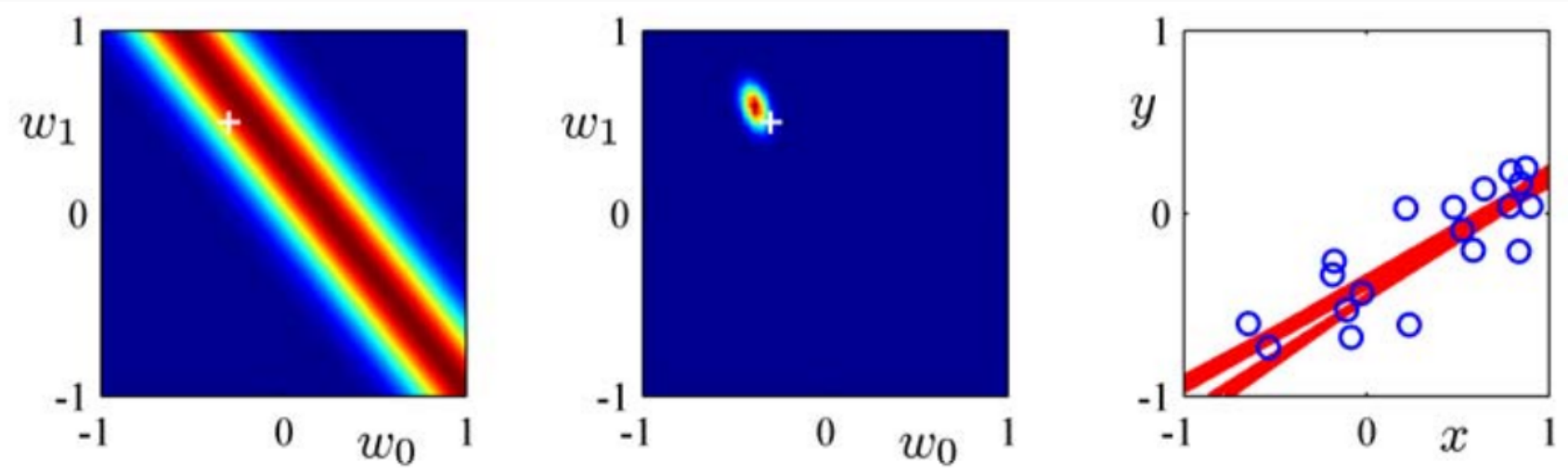
Пример



Пример



Пример



Обобщение

- Можно слегка обобщить – рассмотреть априорное распределение более общего вида

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

Упражнение. Подсчитайте логарифм правдоподобия.

И снова о регуляризации

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 побольше нулевых компонент в векторе $\mathbf{w}$.

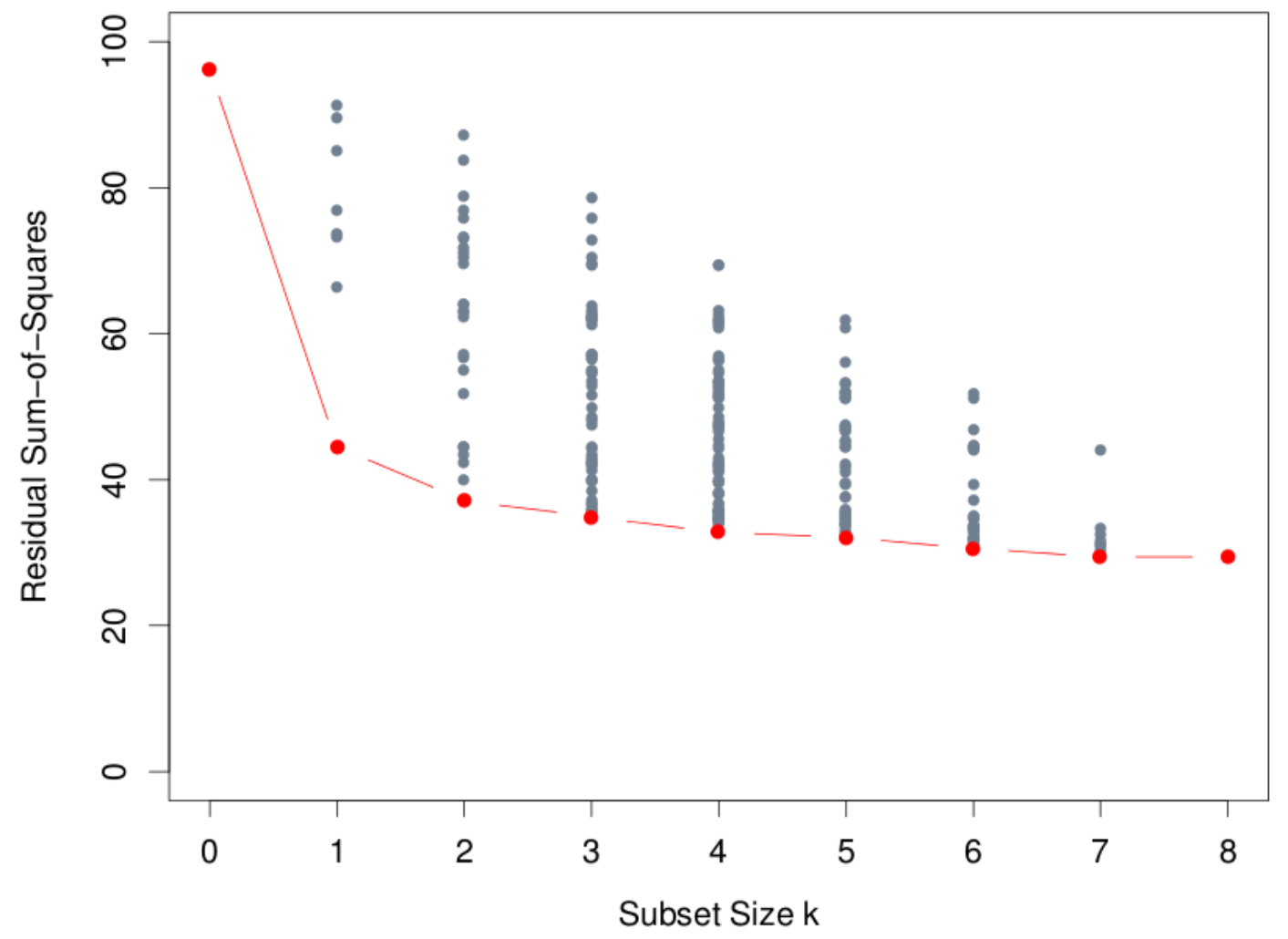
Subset selection

- 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:

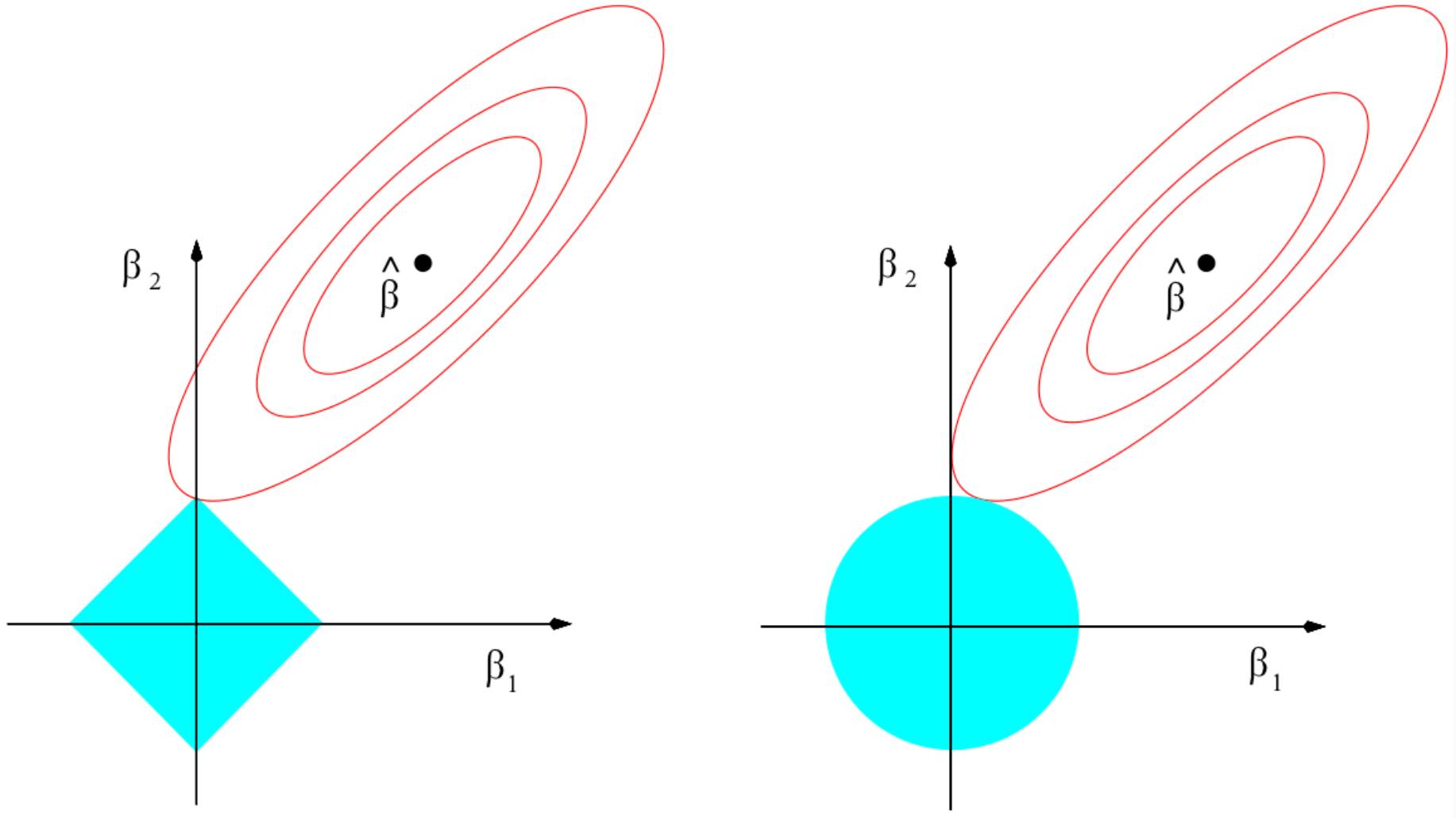
$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left\{ \frac{1}{2} \sum_{i=1}^{N} (f(x_i, \mathbf{w}) - y_i)^2 + \lambda \sum_{j=0}^{p} |w_j| \right\},$$

эквивалентно

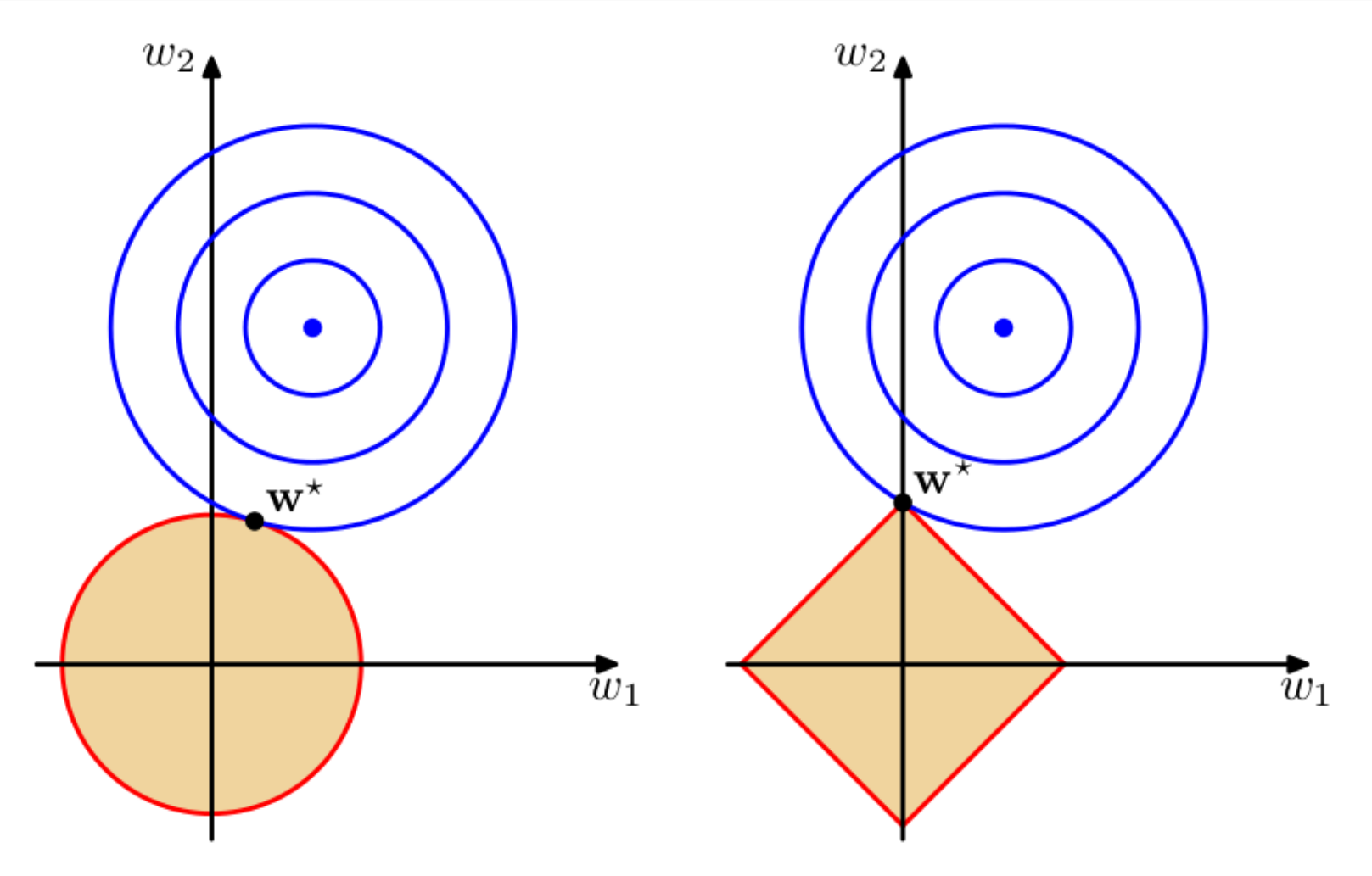
$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left\{ \frac{1}{2} \sum_{i=1}^{N} (f(x_i, \mathbf{w}) - y_i)^2 \right\} \text{ при } \sum_{j=0}^{p} |w_j| \leq t.$$

Упражнение. Докажите это.

Гребень и лассо



Гребень и лассо



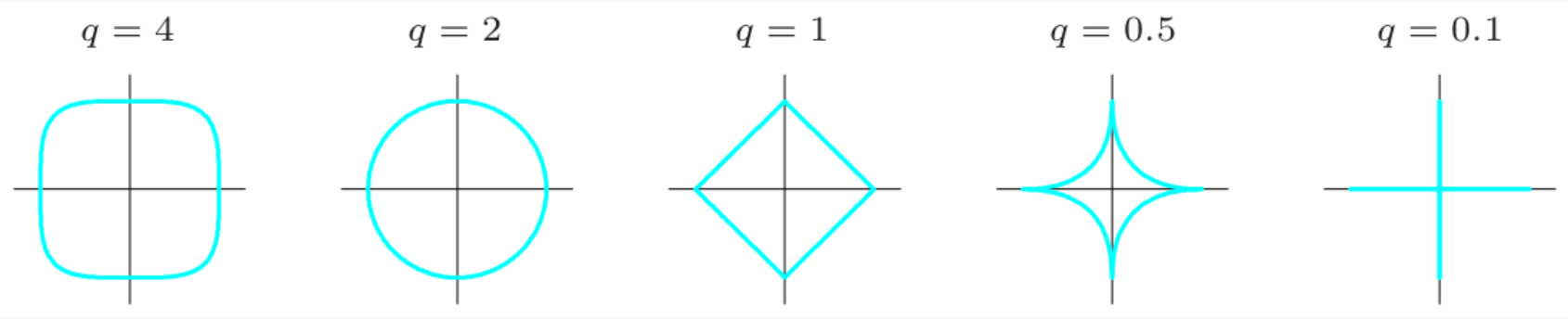
Обобщение

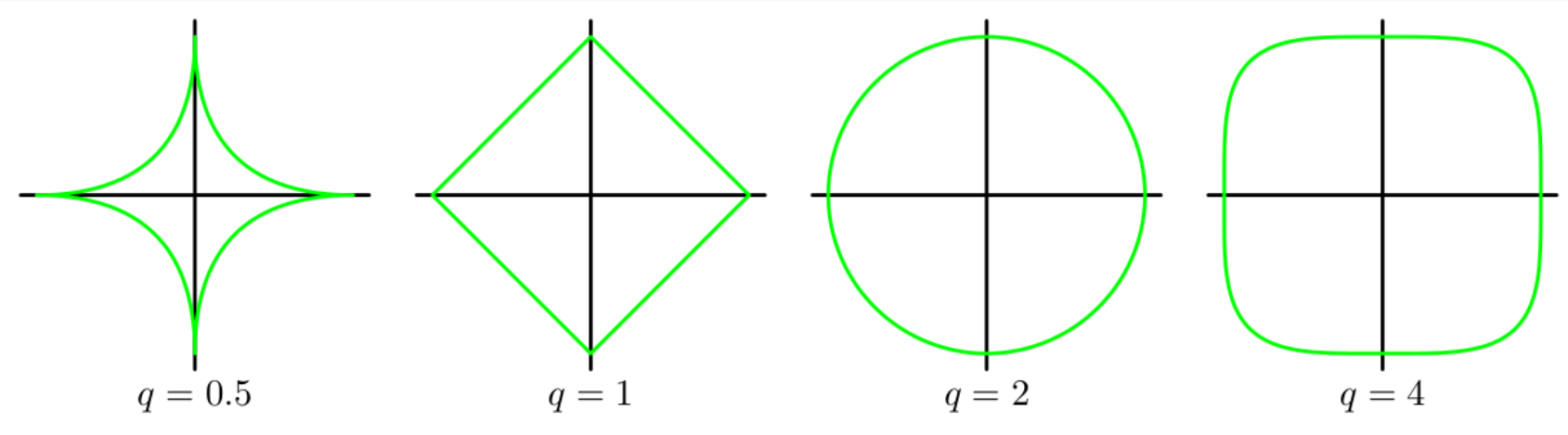
- Можно рассмотреть обобщение гребневой и лассо-регрессии:

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

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

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

мы нашли

$$\begin{aligned} p(\mathbf{w} \mid \mathbf{t}, \alpha, \beta) &= \mathcal{N}(\mathbf{w} \mid \boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N), \\ \boldsymbol{\mu}_N &= \boldsymbol{\Sigma}_N \left(\boldsymbol{\Sigma}_0^{-1} \mu_0 + \beta \boldsymbol{\Phi}^\top \mathbf{t} \right), \\ \boldsymbol{\Sigma}_N &= \left(\boldsymbol{\Sigma}_0^{-1} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \right)^{-1}, \end{aligned}$$

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

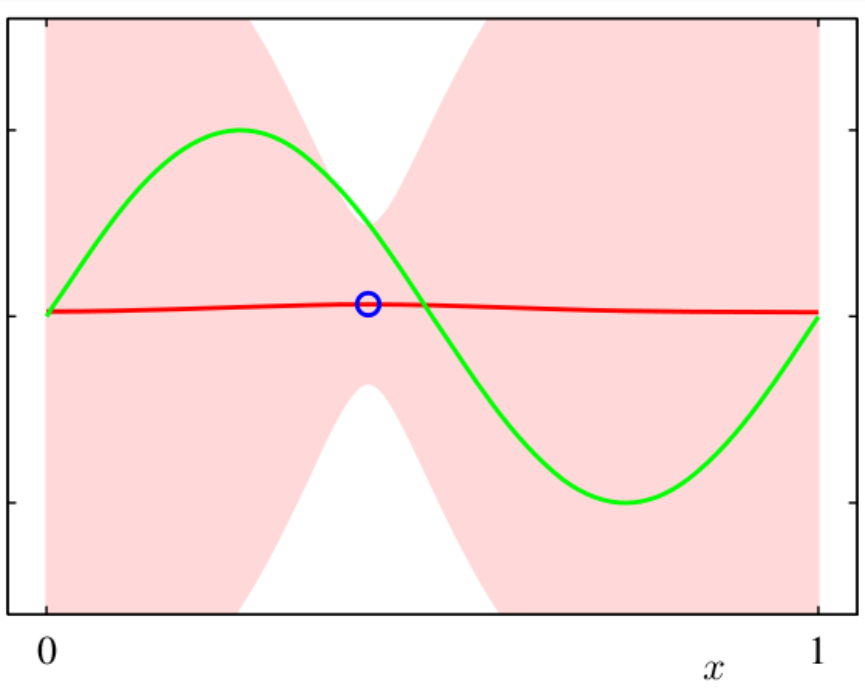
- ...тоже гауссиан:

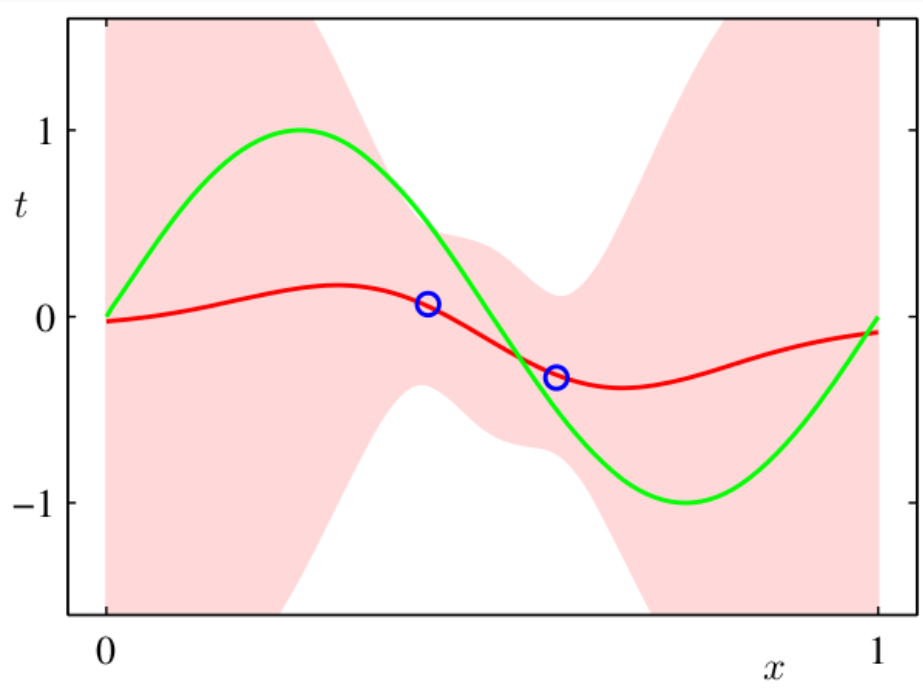
$$p(t \mid \mathbf{t}, \alpha, \beta) = \mathcal{N}(t \mid \boldsymbol{\mu}_N^\top \boldsymbol{\phi}(\mathbf{x}), \sigma_N^2),$$
$$\text{где } \sigma_N^2 = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\Sigma}_N \boldsymbol{\phi}(\mathbf{x}).$$

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

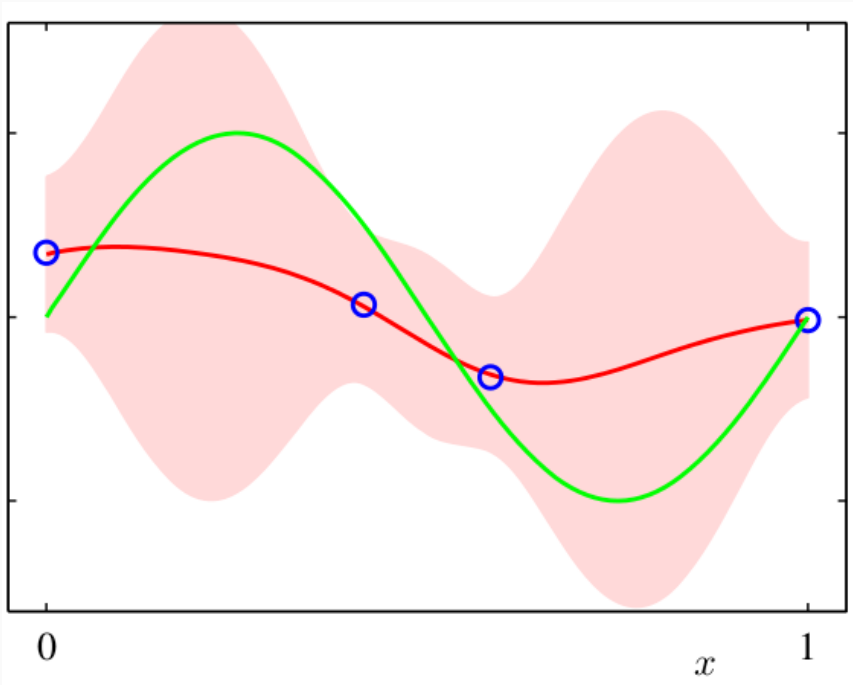
Упражнение. Оценка всё время уточняется: $\sigma_{N+1}^2 \leq \sigma_N^2$.

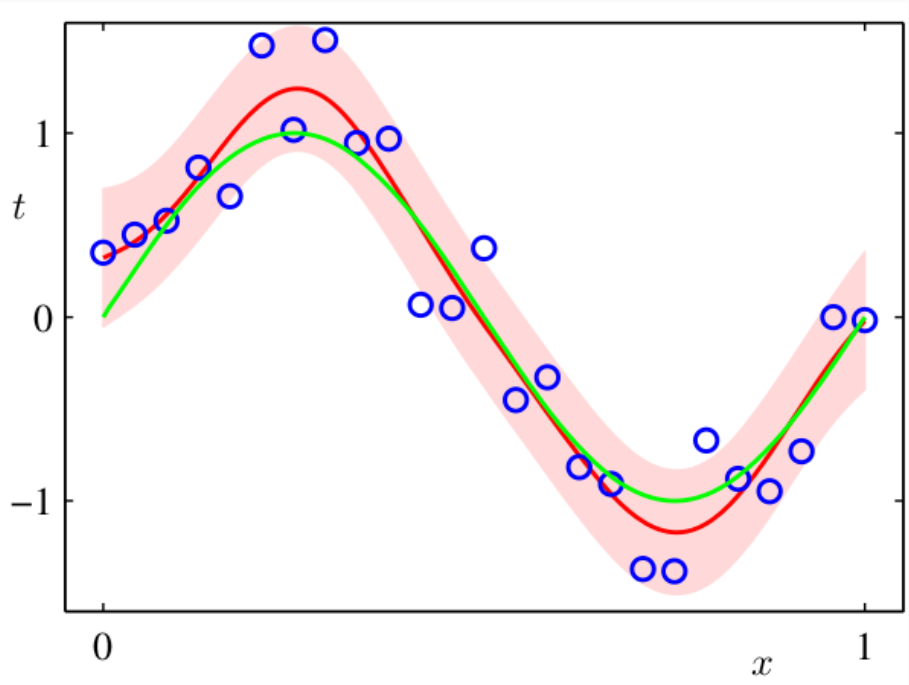
Предсказания



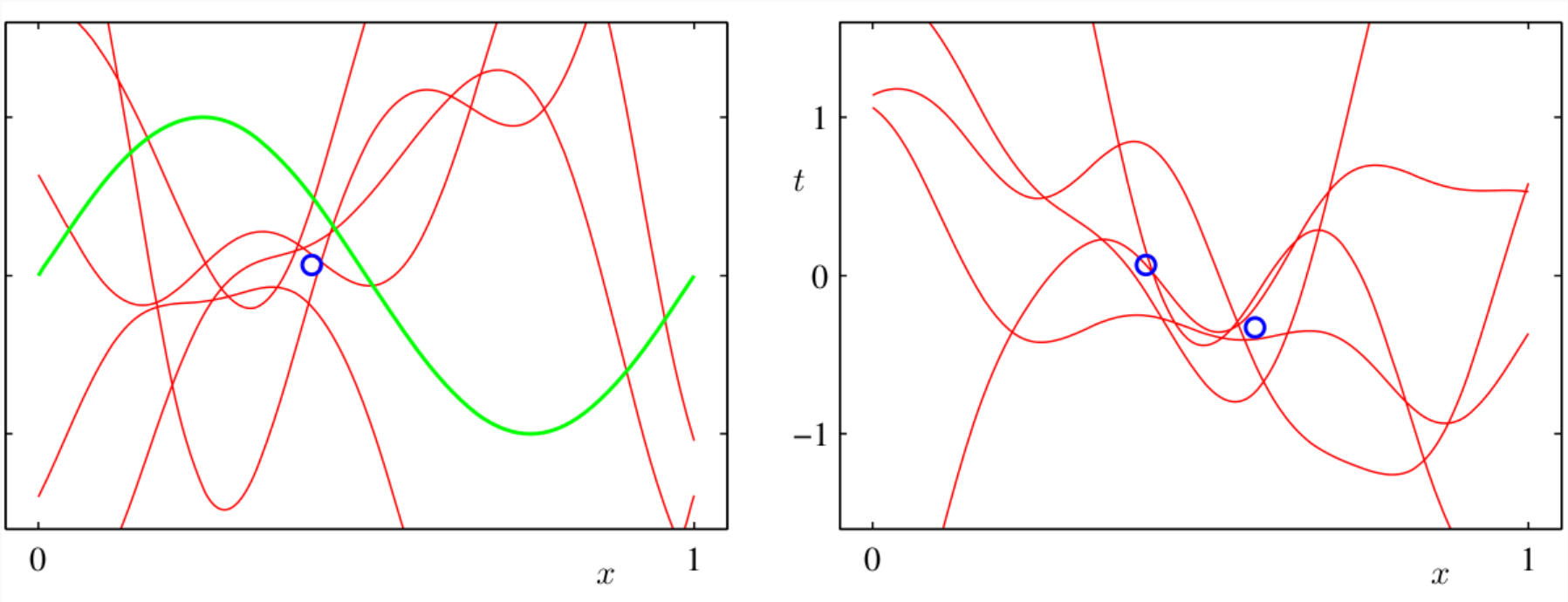


Предсказания

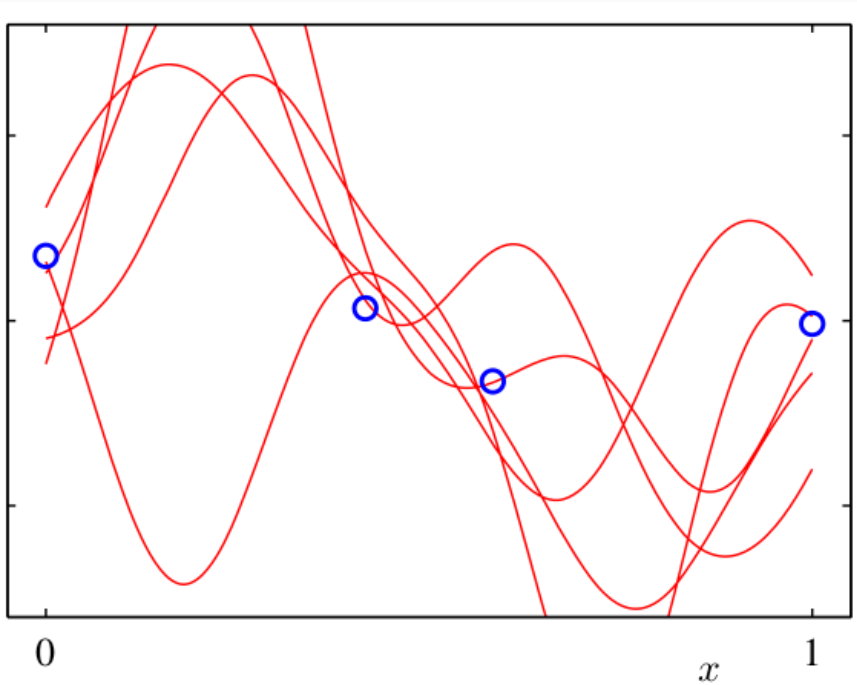


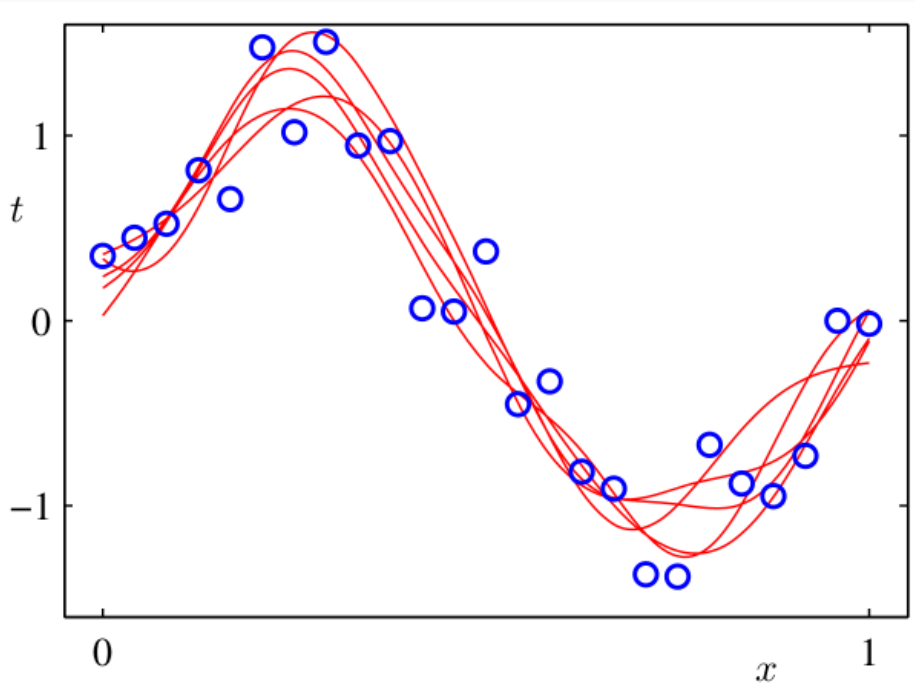


Предсказания



Предсказания





Параметрические и непараметрические модели

- Последнее замечание: модели бывают параметрические и непараметрические.
- Мы в основном будем заниматься моделями с фиксированным числом параметров, которые делают сильные предположения.
- Но есть класс непараметрических моделей, которые не делают предположений почти никаких (это не совсем правда), а основаны непосредственно на данных; они в некоторых ситуациях очень хороши, но плохо обобщаются на высокие размерности и большие датасеты.

Метод ближайших соседей

- Пример непараметрической модели: метод ближайших соседей.
- Давайте на примере задачи классификации.
- Не будем строить вообще никакой модели, а будем классифицировать новые примеры как

$$\hat{y}(\mathbf{x}) = \frac{1}{k} \sum_{\mathbf{x}_i \in N_k(\mathbf{x})} y_i,$$

где $N_k(\mathbf{x})$ – множество k ближайших соседей точки $\mathbf{x}$ среди имеющихся данных $(\mathbf{x}_i, y_i)_{i=1}^N$.

Метод ближайших соседей

- Единственный «параметр» – это k, но от него многое зависит.
- Для разумно большого k у нас в нашем примере стало меньше ошибок.
- Но это не предел – для $k = 1$ на тестовых данных вообще никаких ошибок нету!
- Что это значит? В чём недостаток метода ближайших соседей при $k = 1$?
- Как выбрать k? Можно ли просто подсчитать ошибку классификации и минимизировать её?

Проклятие размерности

- В прошлый раз k-NN давали гораздо более разумные результаты, чем линейная модель, особенно если хорошо выбрать k.
- Может быть, нам в этой жизни больше ничего и не нужно?
- Давайте посмотрим, как k-NN будет вести себя в более высокой размерности (что очень реалистично).

Проклятие размерности

- Давайте поищем ближайших соседей у точки в единичном гиперкубе. Предположим, что наше исходное распределение равномерное.
- Чтобы покрыть долю α тестовых примеров, нужно (ожидаемо) покрыть долю α объёма, и ожидаемая длина ребра гиперкуба-окрестности в размерности p будет $e_p(\alpha) = \alpha^{1/p}$.
- Например, в размерности 10 $e_{10}(0.1) = 0.8$, $e_{10}(0.01) = 0.63$, т.е. чтобы покрыть 1% объёма, нужно взять окрестность длиной больше половины носителя по каждой координате!
- Это скажется и на k-NN: трудно отвергнуть по малому числу координат, быстрые алгоритмы хуже работают.

Проклятие размерности

- Второе проявление the curse of dimensionality: пусть N точек равномерно распределены в единичном шаре размерности p. Тогда среднее расстояние от нуля до точки равно

$$d(p, N) = \left(1 - \frac{1}{2}^{1/N}\right)^{1/p},$$

  т.е., например, в размерности 10 для $N = 500$ $d \approx 0.52$, т.е. больше половины.
- Большинство точек в результаты ближе к границе носителя, чем к другим точкам, а это для ближайших соседей проблема – придётся не интерполировать внутри существующих точек, а экстраполировать наружу.

Проклятие размерности

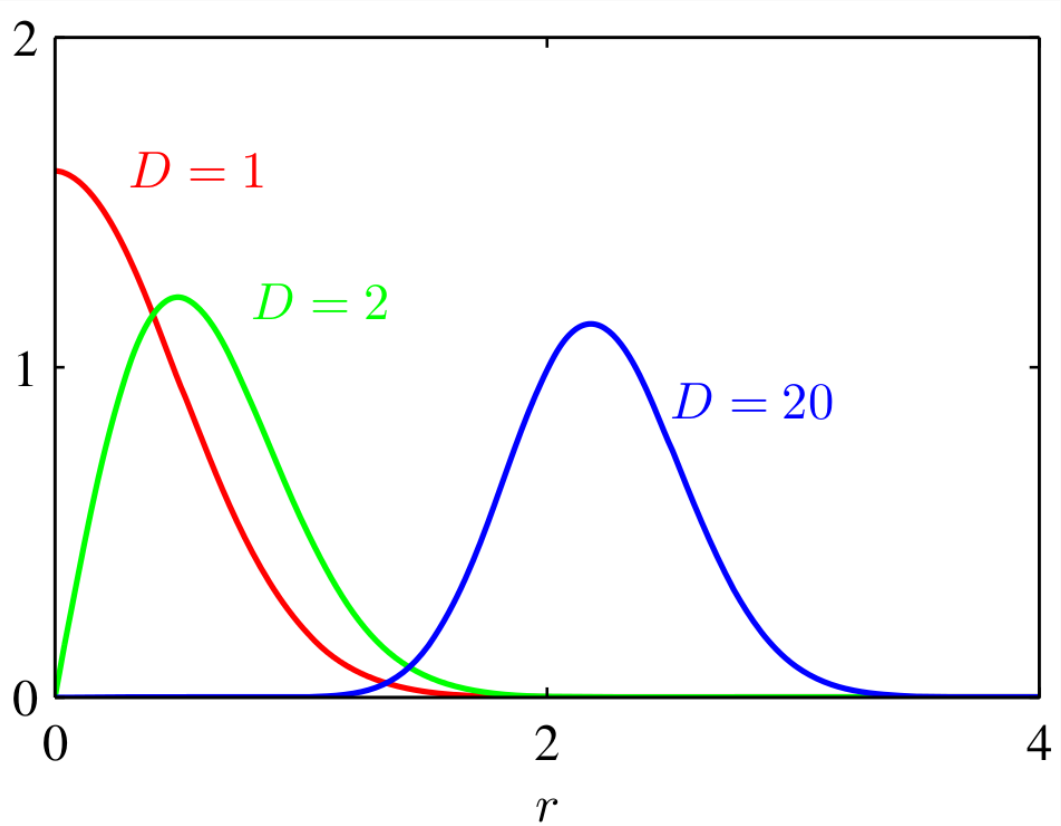
- Третье проявление: проблемы в оптимизации, которые и имел в виду Беллман.
- Если нужно примерно оптимизировать функцию от d переменных, на решётке с шагом ϵ понадобится примерно $\left(\frac{1}{\epsilon}\right)^d$ вычислений функции.
- В численном интегрировании – чтобы интегрировать функцию с точностью ϵ, нужно тоже примерно $\left(\frac{1}{\epsilon}\right)^d$ вычислений.

Проклятие размерности

- Плотные множества становятся очень разреженными. Например, чтобы получить плотность, создаваемую в размерности 1 при помощи $N = 100$ точек, в размерности 10 нужно будет 100^{10} точек.
- Поведение функций тоже усложняется с ростом размерности – чтобы строить регрессии в высокой размерности с той же точностью, может потребоваться экспоненциально больше точек, чем в низкой размерности.
- А у линейной модели ничего такого не наблюдается, она не подвержена проклятию размерности.

Проклятие размерности

- Ещё пример: нормально распределённая величина будет сосредоточена в тонкой оболочке.



Упражнение. Переведите плотность нормального распределения в полярные координаты и проверьте это утверждение.

Спасибо!

Спасибо за внимание!