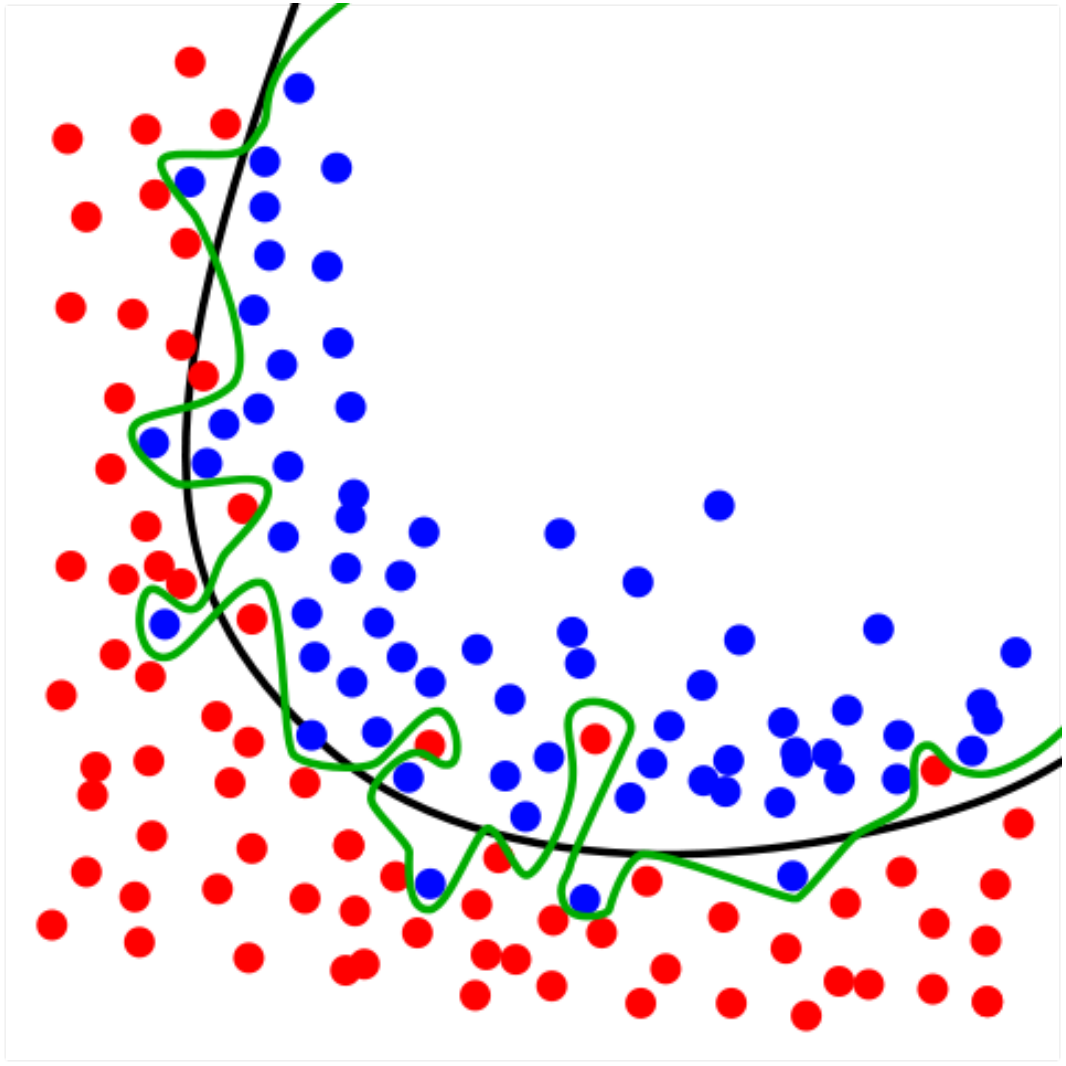
Регуляризация в МО

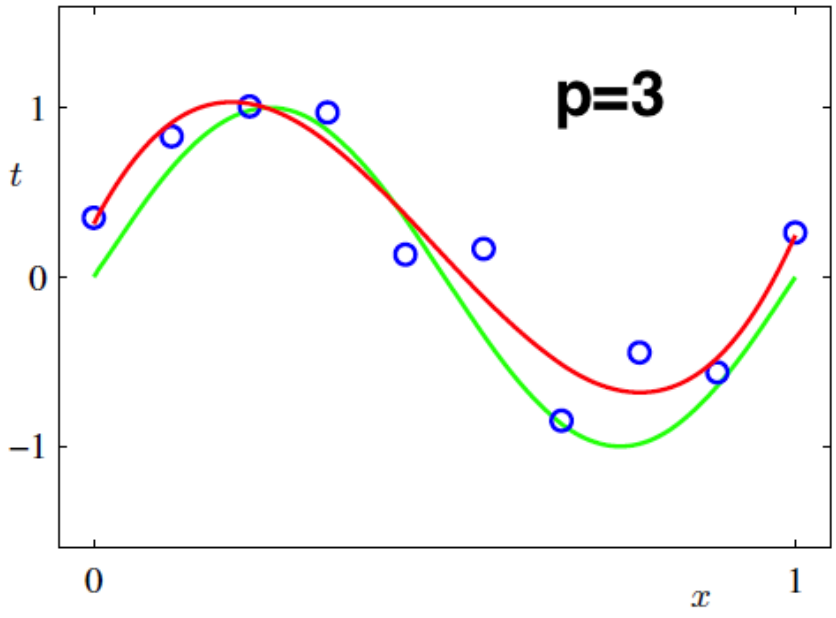
Евгений Бурнаев

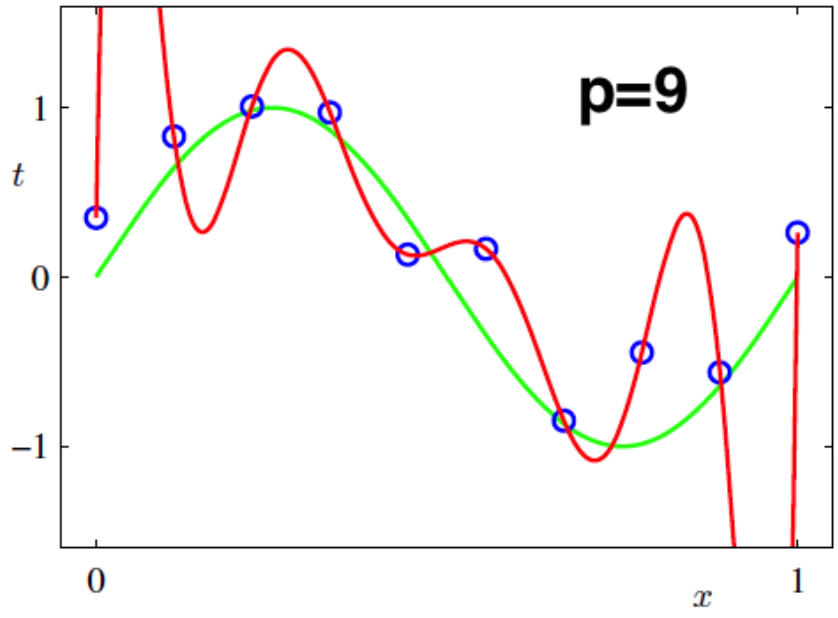
Сколтех, Москва, Россия

Переобучение: kNN



Переобучение: регрессия





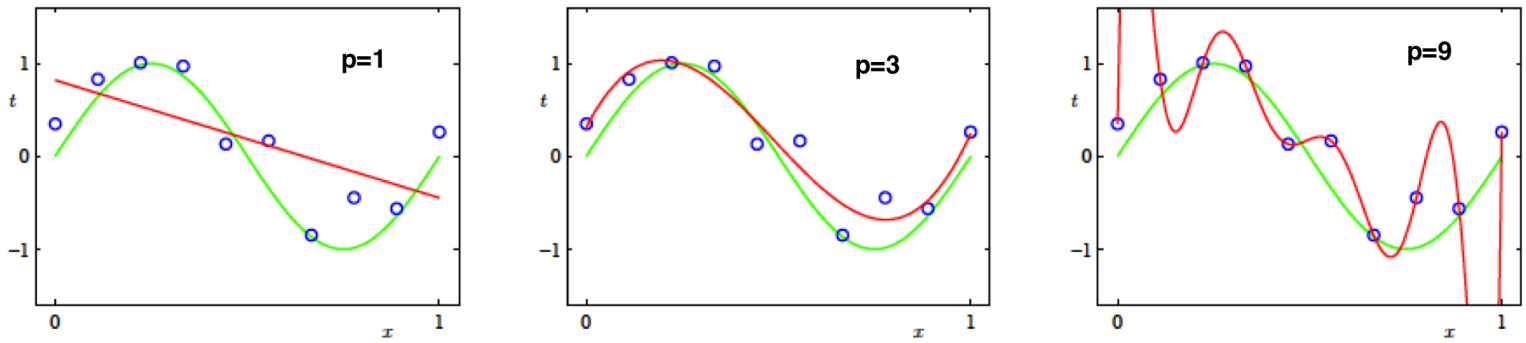
Разложение ошибки

- Ошибка аппроксимации/моделирования
  - аппроксимация реального поведения моделью
- Ошибка оценивания
  - обучение модели по конечной выборке
- Ошибка оптимизации
  - насколько хорошо решена оптимизационная задача
- Байесовская ошибка
  - реальность не идеальна (существует нижняя граница на ошибку для всех моделей, обычно ненулевая)

Компромисс между смещением и дисперсией

- **Смещение**: разность между реальным поведением и тем поведением, которое мы ожидаем получить
  - Оценивает насколько ожидания расходятся с реальностью
  - Уменьшается с ростом сложности модели
- **Дисперсия**: разность между тем, что мы ожидаем обучить, и тем, что мы выучиваем на заданной выборке
  - Оценивает насколько чувствителен алгоритм к конкретной выборке
  - Увеличивается с ростом сложности модели

- Пример: полиномиальная регрессия $h(x) = \sum_{j=0}^{p} w_j x^j$



- Значения оптимальных регрессионных коэффициентов

	p=0	p=1	p=3	p=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

Что такое выбор модели?

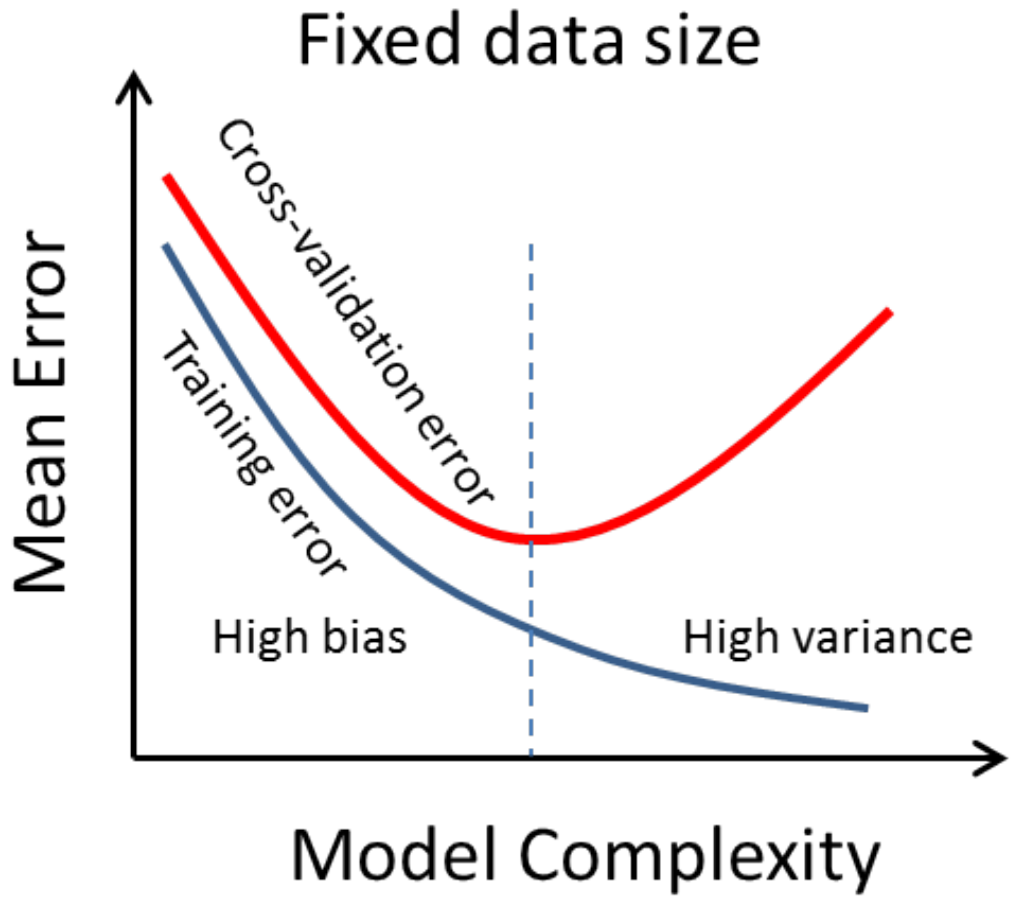
- Задано множество моделей $F = \{F_1, \ldots, F_K\}$. Выбрать модель с **наилучшим ожидаемым качеством на тестовых данных**
- F может состоять из
  1. моделей из одного класса, различающихся только по гиперпараметрам
     - Нелинейная регрессия: полиномы разной степени
     - k ближайших соседей: разные значения k
     - Решающие деревья: различная глубина/количество листьев
     - SVM: различные значения штрафа на неправильную классификацию C
     - Модели с регуляризацией: различные значения параметра регуляризации
     - Ядерные методы: различные ядра и т.д.
  2. Различные обучающие модели (SVM, kNN, решающие деревья и т.д.)
- Замечание: обычно выбор модели возникает в контексте обучения с учителем, но также встречается при обучении без учителя (например, “количество кластеров” при кластеризации)

Выбор модели

- Бритва Оккама: среди всех гипотез необходимо выбирать ту, в которой делается меньше предположений
- Слишком много переменных/параметров $\Rightarrow$ большая дисперсия и маленькое смещение на обучающей выборке и наоборот
- Получаем **две взаимосвязанные проблемы**
  - **Задача 1**. Оценивание целевой функции, характеризующей обобщающую способность
  - **Задача 2**. Вычислительно эффективно выбрать оптимальную модель с точки зрения критерия точности

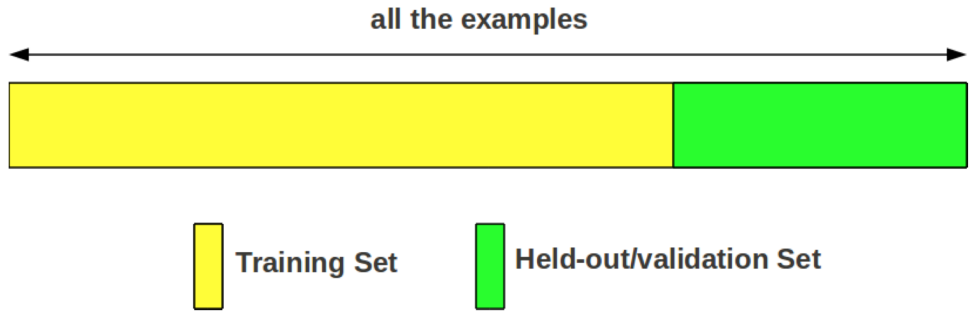
Обучающая и тестовая ошибки

- Обучающая ошибка уменьшается с ростом сложности модели
- Тестовая ошибка растет с ростом сложности модели



Валидационная выборка

- Отложить долю (к примеру $10\% - 20\%$) обучающей выборки
- Эта часть называется валидационной выборкой



- Запомните: валидационная выборка это НЕ тестовая выборка
- Обучить модель на оставшейся части обучающей выборки
- Оценить ошибку на валидационной выборке
- Выбрать модель с наименьшей ошибкой на валидационной выборке
- Проблемы:
  - Уменьшается обучающая выборка, поэтому обычно используется при больших размерах выборки
  - Валидационная выборка может быть не очень хорошей при неудачном разбиении (используйте случайное разбиение!)

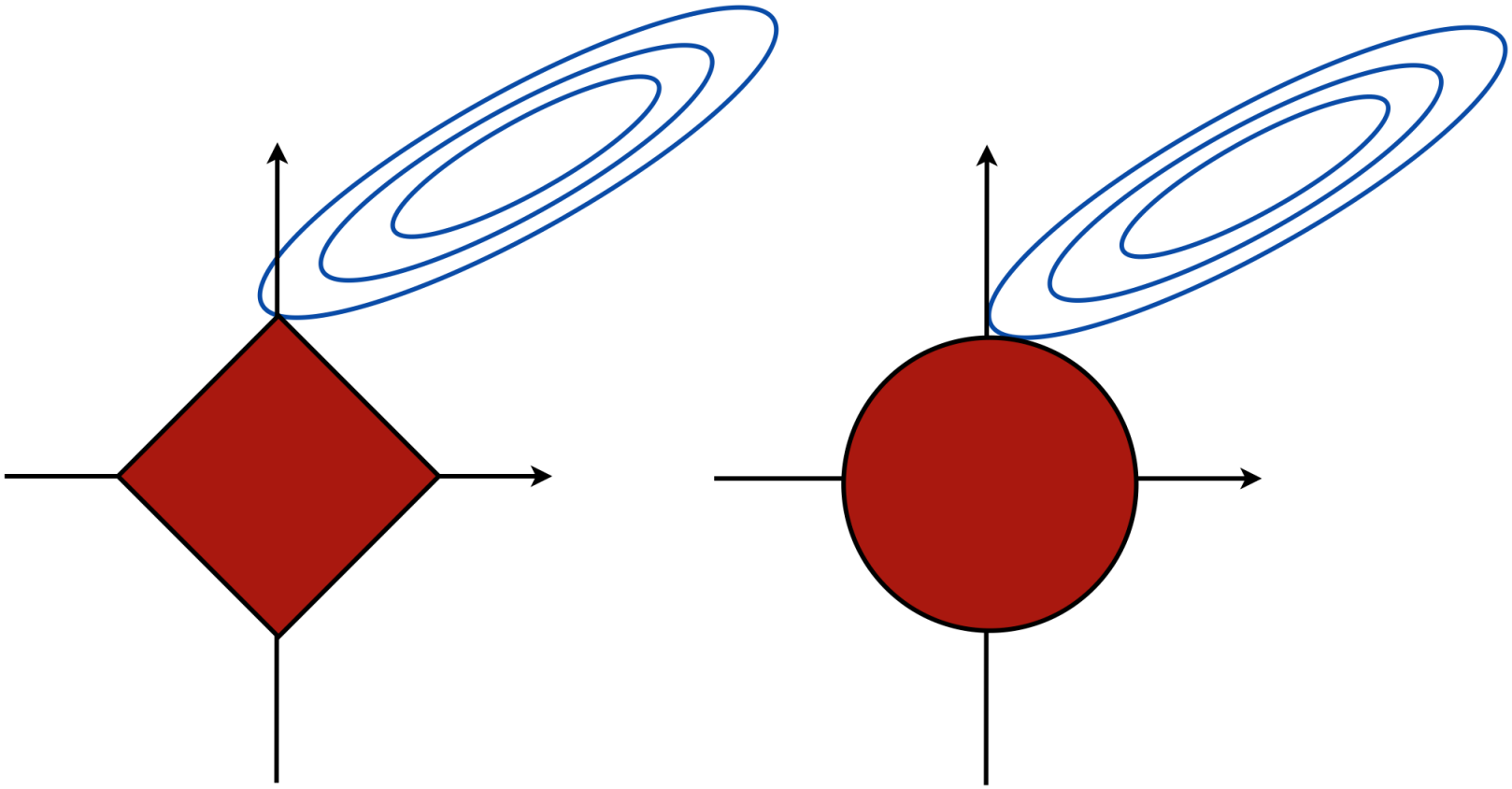
- Обучающая выборка $S_m = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, $\mathbf{x} \in X$, $y \in Y$
- $\mathbf{X} = \{\mathbf{x}_i, i = 1, \ldots, m\}$ это матрица входных значений
- Рассмотрим линейную модель $f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \cdot \mathbf{x} + b$, $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{x} \in \mathbb{R}^d$, $y = f(\mathbf{x}) + \varepsilon$ (ε это белый шум)

- **Оптимизационная задача**: ("Least Absolute Shrinkage and Selection Operator")

$$\min_{\mathbf{w},b} F(\mathbf{w},b) = \lambda\|\mathbf{w}\|_1 + \sum_{i=1}^{m}(\mathbf{w}\cdot\mathbf{x}_i + b - y_i)^2,$$

  где $\lambda \geq 0$ это параметр регуляризации
- **Решение**: эквивалентно выпуклому квадратичному программированию (QP)
  - общее: стандартные QP солверы
  - специальные алгоритмы: LARS (регрессия на основе наименьших углов), полный путь решения

Разреженность L_1 регуляризации



L1 regularization

L2 regularization

Комментарии

- Преимущества
  - строгие теоретические гарантии
  - разреженное решение
  - отбор признаков
- Недостатки
  - нет возможности использовать ядра
  - нет решения в явной форме (необязательно, но могло бы быть удобно для теоретического анализа)
- Другие семейства алгоритмов включают в себя
  - нейронные сети, гауссовские процессы
  - решающие правила
  - бустинг решающих деревьев
- **Эмпирическое правило** для улучшения качества предсказания:
  - Сначала необходимо сделать отбор признаков с помощью LASSO
  - Затем оценить параметры модели заново с помощью гребневой регрессии используя отобранные признаки

Эмпирическая ошибка

- $L(f(\mathbf{x}), y)$ это потери для пары $(\mathbf{x}, y)$ и модели f
- $\widehat{R}(f; S_m) = \frac{1}{m} \sum_{i=1}^{m} L(f(\mathbf{x}_i), y_i)$ это потери f на S_m
- Эмпирическая ошибка на обучающей выборке

$$\widehat{R}_{\mathcal{A}}(S_m) = \widehat{R}(f; S_m), \;\; f(\cdot) = \mathcal{A}(S_m)$$

  Эта ошибка это смещенная оценка обобщающего риска
- Эмпирическая ошибка на тестовой выборке оценивается на валидационной выборке S^t

$$\widehat{R}_{\mathcal{A}}(S_m; S^t) = \widehat{R}(f; S^t), \;\; f(\cdot) = \mathcal{A}(S_m)$$

  — нам необходима дополнительная тестовая выборка S^t или
  — необходимо разбить S_m на обучающую и валидационную выборки (результаты зависят от этого разбиения)

- Верхняя граница на вероятность переобучения для любой выборки S_m, довольно общего класса гипотез F и обучающего алгоритма $\mathcal{A}$:

$$\mathbb{P}\left(\widehat{R}(f;S^t) - \widehat{R}(f;S_m) \geq \varepsilon\right) \leq \delta(\varepsilon, F),\ \ f(\cdot) = \mathcal{A}(S_m)$$

- Тогда для любой S_m, F, $\mathcal{A}$ и $\delta \in (0,1)$ с вероятностью больше, чем $(1-\delta)$, получаем

$$\widehat{R}(f;S^t) \leq \widehat{R}(f;S_m) + \varepsilon(\delta, F)$$

- Эмпирический риск с поправкой

$$\widehat{R}(f;S_m) + \varepsilon(\delta, F) \to \min_{f,F}$$

- Регуляризация штрафует сложность модели F

$$\widehat{R}_{\text{pen}}(f; S_m) = \widehat{R}(f; S_m) + \text{pen}(F)$$

- Рассмотрим линейные модели $F = \{f(\mathbf{x}) = \text{sign}(\mathbf{w}^{\text{T}} \cdot \mathbf{x})\}$ (классификация) или $F = \{f(\mathbf{x}) = (\mathbf{w}^{\text{T}} \cdot \mathbf{x})\}$ (регрессия)
- Тогда
  a) L_2-регуляризация $\text{pen}(F) = \lambda \sum_{j=1}^{p} w_j^2$
  b) L_1-регуляризация $\text{pen}(F) = \lambda \sum_{j=1}^{p} |w_j|$
  c) L_0-регуляризация $\text{pen}(F) = \lambda \sum_{j=1}^{p} 1_{w_j \neq 0}$
- AIC и BIC это частные случаи L_0-регуляризации

- Рассмотрим линейную регрессию с гауссовским н.о.р. шумом
- Лог-правдоподобие на S_m имеет вид

$$\mathcal{L}(\mathbf{w}) = m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - \mathbf{w}^{\mathrm{T}} \mathbf{x}_i)^2$$

- Предположим, что

$$\mathbf{w} \sim \mathcal{N}(0, \tau^2 \mathbb{I})$$

- Апостериорное распределение $\mathbf{w}$ имеет вид

$$p(\mathbf{w}|S_m) \propto p(S_m|\mathbf{w}) p(\mathbf{w})$$
$$= \mathrm{C} \cdot \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - \mathbf{w}^{\mathrm{T}} \mathbf{x}_i)^2 \right\} \exp\left\{ -\frac{\mathbf{w}^{\mathrm{T}} \mathbf{w}}{2\tau^2} \right\}$$

- Апостериорное лог-правдоподобие

$$\begin{aligned}\mathcal{L}_{MAP}(\mathbf{w}|S_m) &= -\frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - \mathbf{w}^{\mathrm{T}}\mathbf{x}_i)^2 - \frac{1}{2\tau^2}\sum_{k=1}^{p} w_k^2 \\ &= -\frac{m}{2\sigma^2}\left(\frac{1}{m}\sum_{i=1}^{m}(y_i - \mathbf{w}^{\mathrm{T}}\mathbf{x}_i)^2 + \frac{\sigma^2}{m\tau^2}\sum_{k=1}^{p} w_k^2\right) \\ &= -\frac{m}{2\sigma^2}\left(\widehat{R}(f;S_m) + \lambda\|\mathbf{w}\|^2\right),\ \lambda = \frac{\sigma^2}{m\tau^2}\end{aligned}$$

- Таким образом, MAP оценка совпадает с L_2-регуляризацией