ЛИНЕЙНАЯ РЕГРЕССИЯ ПО-БАЙЕСОВСКИ

Сергей Николенко

TRA Robotics — Санкт-Петербург
19 апреля 2018 г.

Random facts:

- 19 апреля 1825 г. группа повстанцев под руководством Хуана Антонио Лавальехи высадились на пляже Аграсиада и водрузили там флаг Тридцати трёх Ориенталес; позже из этого возник независимый Уругвай
- 19 апреля 1948 г. Альберт Хофман поехал из лаборатории домой на велосипеде
- 19 апреля 1965 г. Гордон Мур сформулировал свой знаменитый закон
- 19 апреля 1987 г. в Tracey Ullman Show появилась первая короткометражка о семье Симпсонов

Регрессия 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):

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

- 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:

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

Байесовская регуляризация

- Тогда наша задача – посчитать

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

- Давайте подсчитаем.

Байесовская регуляризация

- 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.

Байесовская регуляризация

- Если мы возьмём априорное распределение около нуля:

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

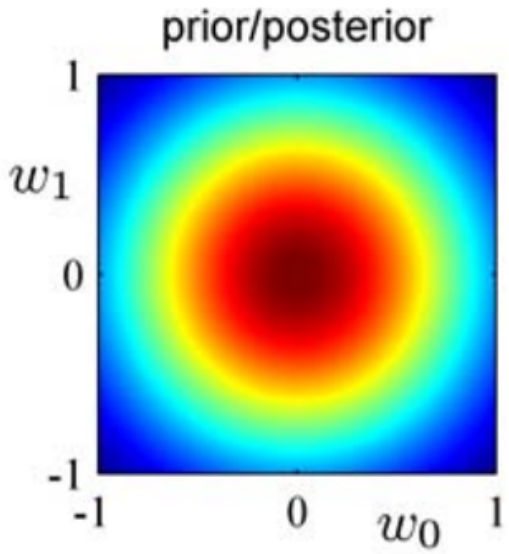
то логарифм правдоподобия получится

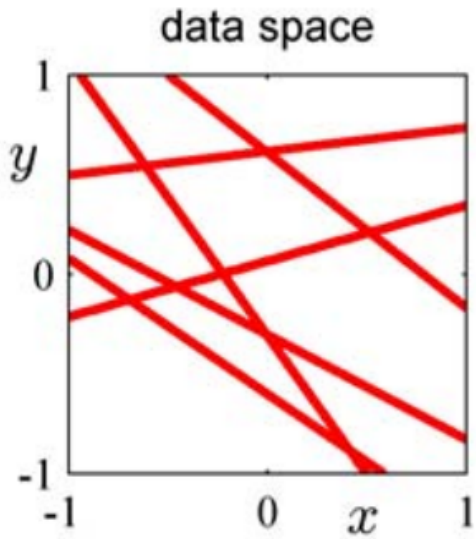
$$\ln p(\mathbf{w} \mid \mathbf{t}) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^\top \phi(\mathbf{x}_n)\right)^2 - \frac{\alpha}{2}\mathbf{w}^\top\mathbf{w} + \text{const},$$

то есть в точности гребневая регрессия.

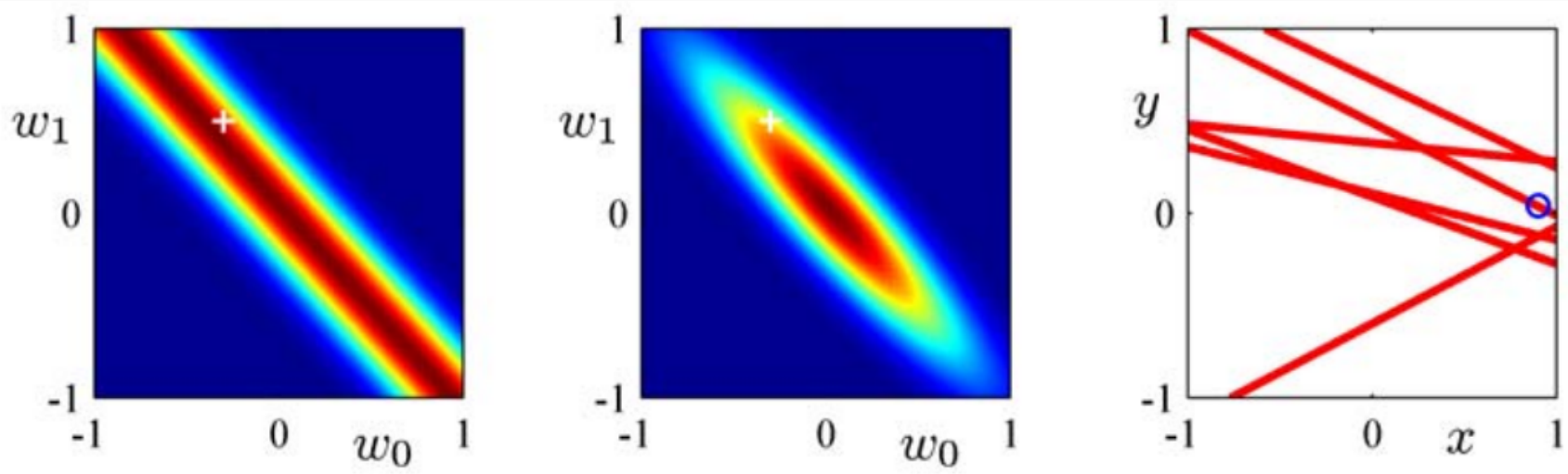
Пример

likelihood

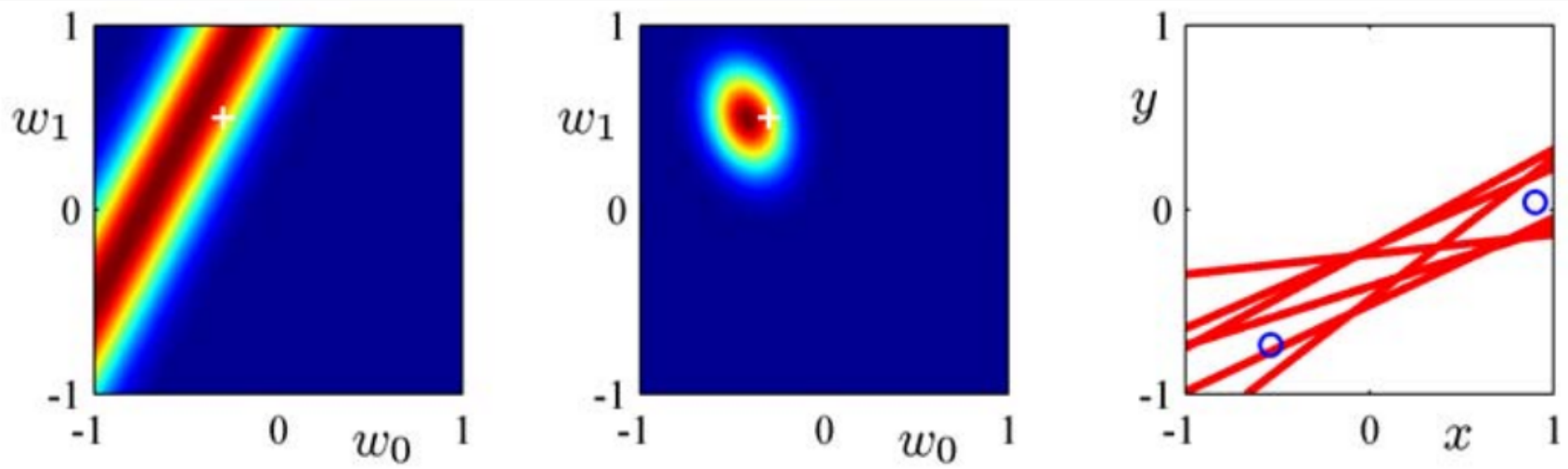




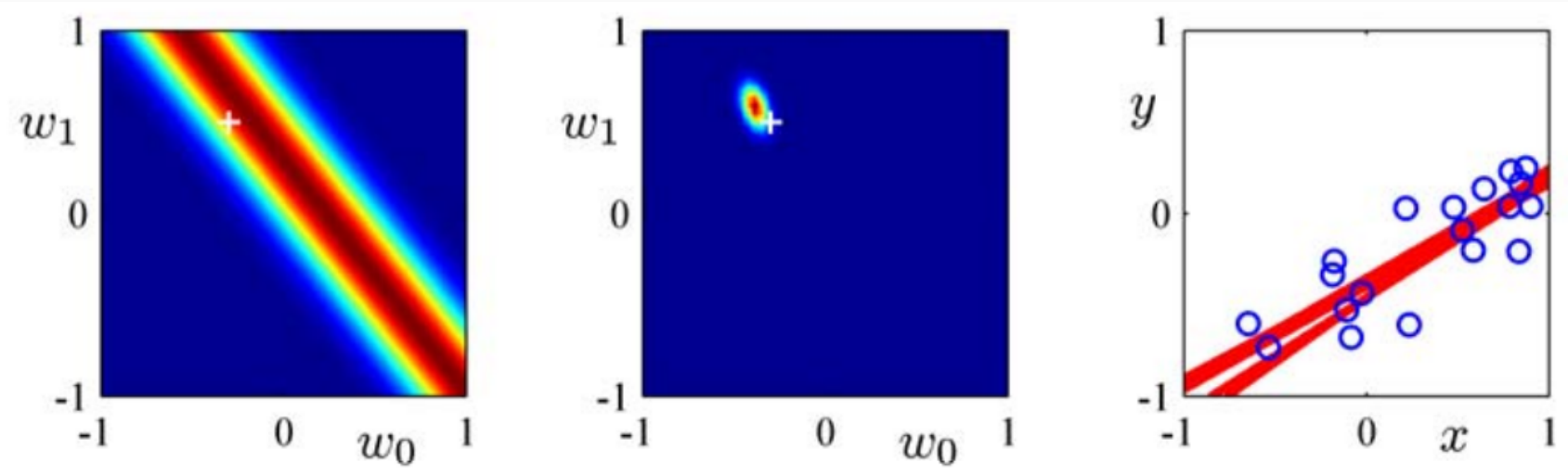
Пример



Пример



Пример



Обобщение

- Можно слегка обобщить – рассмотреть априорное распределение более общего вида

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

Упражнение. Подсчитайте логарифм правдоподобия.

Предсказания в линейной регрессии

Предсказание в линейной регрессии

- Вспомним задачи байесовского вывода:
  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

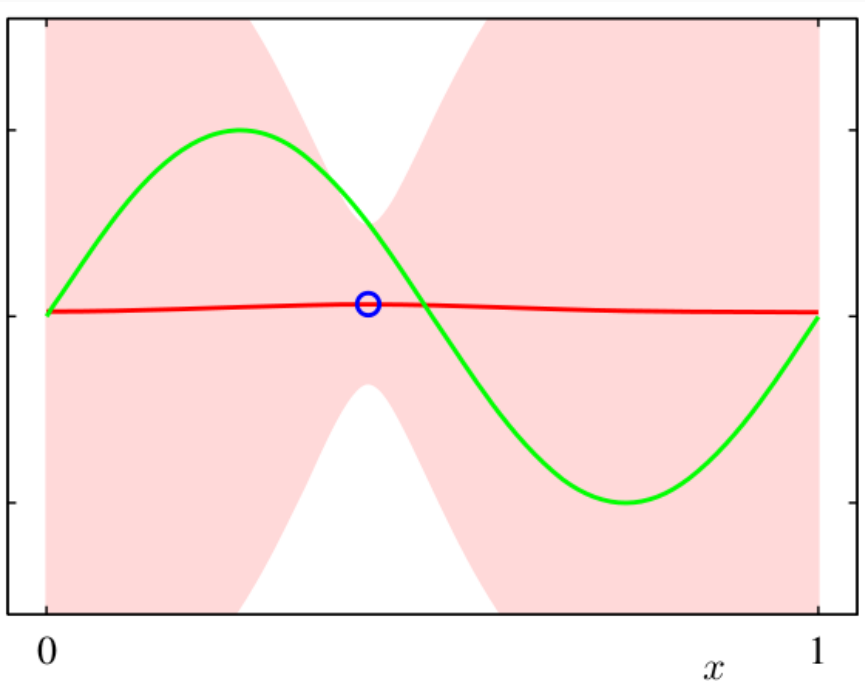
- ...тоже гауссиан:

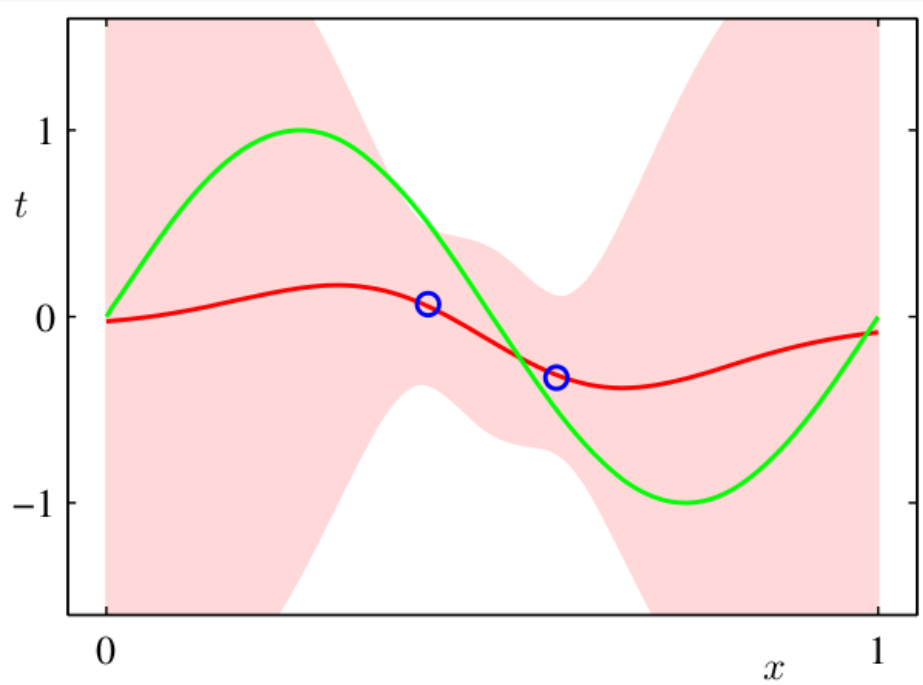
$$p(t \mid \mathbf{t}, \alpha, \beta) = \mathcal{N}(t \mid \mu_N^\top \phi(\mathbf{x}), \sigma_N^2),$$
$$\text{где } \sigma_N^2 = \frac{1}{\beta} + \phi(\mathbf{x})^\top \Sigma_N \phi(\mathbf{x}).$$

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

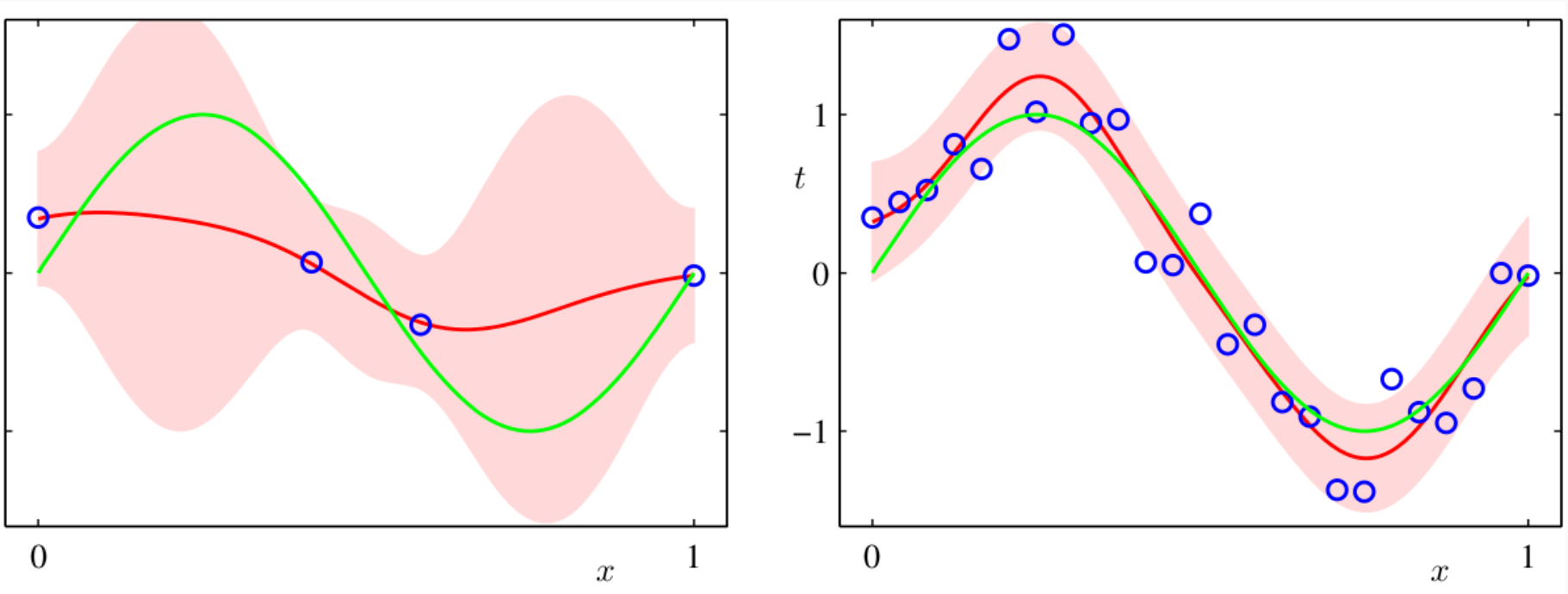
Упражнение. Оценка всё время уточняется: $\sigma_{N+1}^2 \leq \sigma_N^2$.

Предсказания

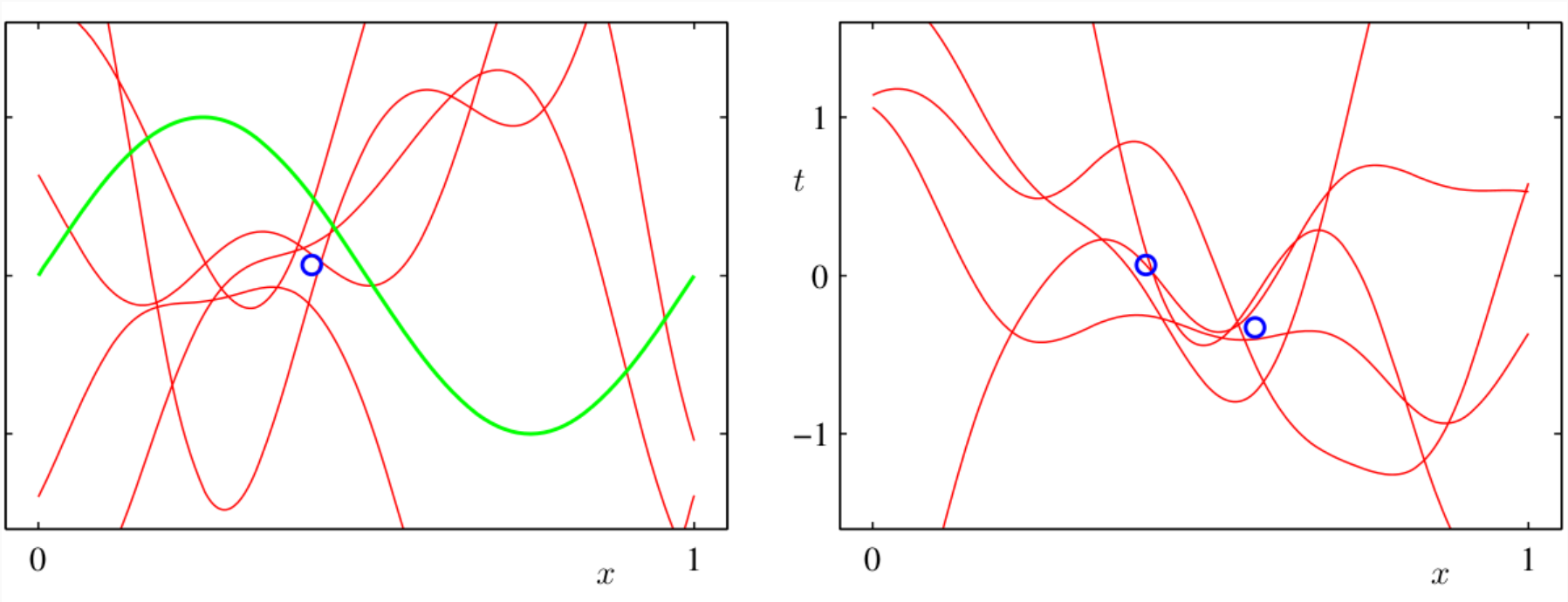




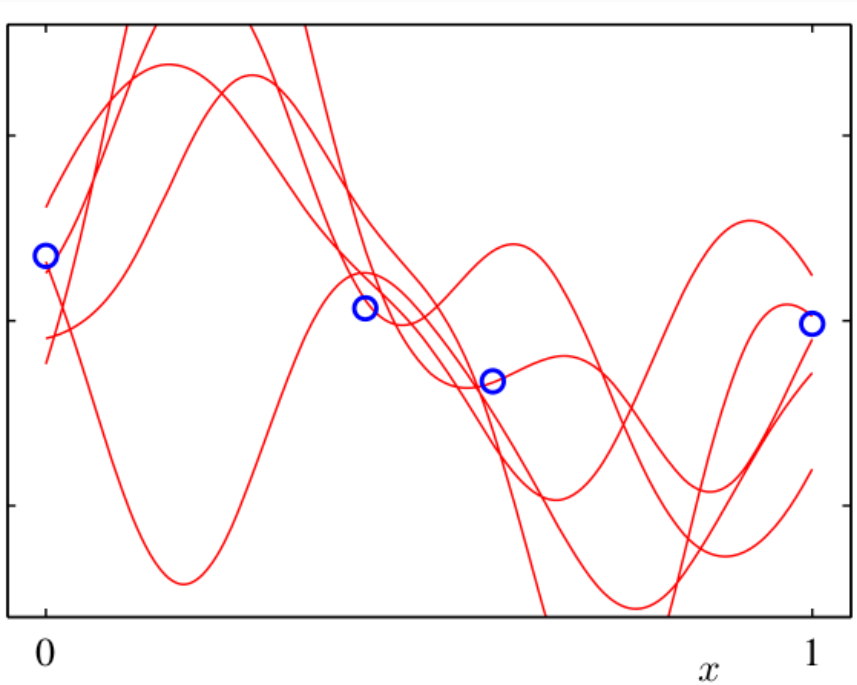
Предсказания

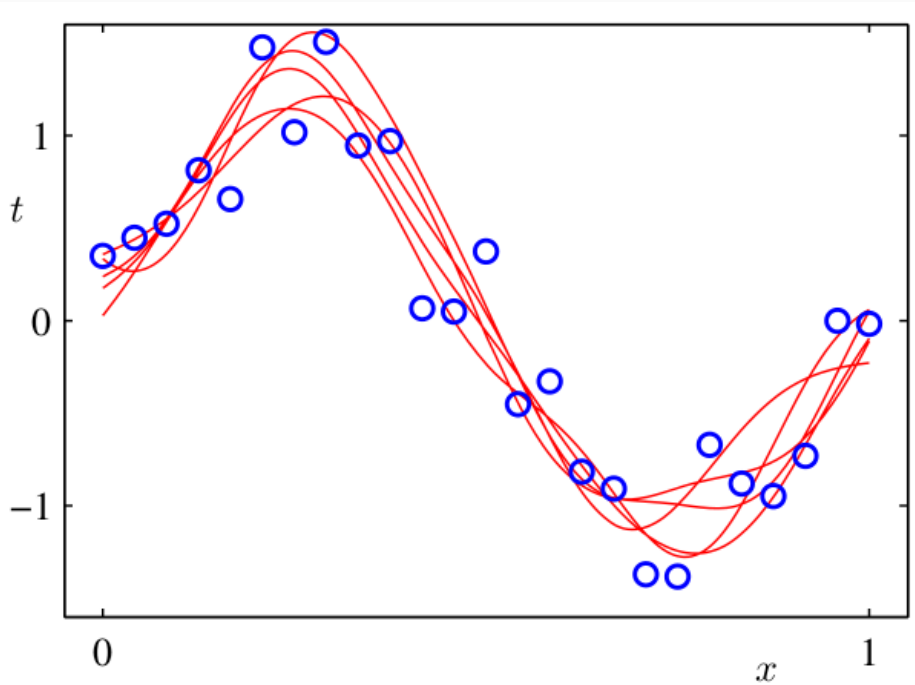


Предсказания



Предсказания





Байесовский вывод для гауссиана

Нормальное распределение: фиксируем σ

- На самом деле всё это — байесовский вывод для нормального распределения:

$$p(x_1, \dots, x_n \mid \mu, \sigma^2) \propto \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2\right).$$

- Хотим: найти сопряжённое априорное распределение, подсчитать правдоподобие, решить задачу предсказания.
- Для начала зафиксируем σ^2 и будем в качестве параметра рассматривать только μ.

Нормальное распределение: фиксируем σ

- Сопряжённое априорное распределение для μ при фиксированном σ^2 тоже нормальное и выглядит как

$$p(\mu \mid \mu_0, \sigma_0^2) \propto \frac{1}{\sigma_0^n} \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right).$$

- Обычно выбирают $\mu_0 = 0$, $\sigma_0^2 \to \infty$ (порой буквально).
- Давайте рассмотрим сначала случай ровно одного наблюдения x и найдём $p(\mu \mid x)$.

Нормальное распределение: фиксируем σ

- При нашем априорном распределении у μ и x совместное нормальное распределение:

$$x = \mu + \sigma\epsilon, \quad \mu = \mu_0 + \sigma_0\delta, \quad \epsilon, \delta \sim \mathcal{N}(0, 1).$$

Упражнение. Пусть (z_1, z_2) – случайные величины с совместным нормальным распределением. Докажите, что случайная величина $z_1 \mid z_2$ распределена нормально с параметрами

$$E(z_1 \mid z_2) = E(z_1) + \frac{Cov(z_1, z_2)}{Var(z_2)}\left(z_2 - E(z_2)\right),$$

$$Var(z_1 \mid z_2) = Var(z_1) - \frac{Cov^2(z_1, z_2)}{Var(z_2)}$$

$(Var(x) = E\left[(x - Ex)^2\right], Cov(x, y) = E\left[(x - Ex)(y - Ey)\right])$.

Нормальное распределение: фиксируем σ

- В нашем случае:

$$
\begin{aligned}
x = \mu + \sigma\epsilon, \quad \mu = \mu_0 + \sigma_0\delta, \quad \epsilon, \delta \sim \mathcal{N}(0,1),\\
E(x) = \mu_0,\\
Var(x) = E(Var(x \mid \mu)) + Var(E(x \mid \mu)) = \sigma^2 + \sigma_0^2,\\
Cov(x, \mu) = E\left[(x - \mu_0)(\mu - \mu_0)\right] = \sigma_0^2.
\end{aligned}
$$

- Применив упражнение, получаем:

$$
\begin{aligned}
E(\mu \mid x) = \mu_0 + \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2}(x - \mu_0) = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2}x + \frac{\sigma^2}{\sigma_0^2 + \sigma^2}\mu_0,\\
Var(\mu \mid x) = \frac{\sigma^2\sigma_0^2}{\sigma_0^2 + \sigma^2} = \frac{1}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}}.
\end{aligned}
$$

Нормальное распределение: фиксируем σ

- Итого:

$$p(\mu \mid x) \sim \mathcal{N}\left(\frac{\sigma_0^2}{\sigma_0^2 + \sigma^2} x + \frac{\sigma^2}{\sigma_0^2 + \sigma^2}\mu_0, \left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}\right)^{-1}\right).$$

- Опять же, сложные вычисления можно забыть и пользоваться этими формулами.
- Замечание: часто используют $\tau = \frac{1}{\sigma^2}$ как параметр нормального распределения (precision). Тогда

$$\tau_{\mu|x} = \tau_\mu + \tau.$$

Нормальное распределение: фиксируем σ

- А что, если данных больше, $x_1, \dots, x_n$?
- Тогда можно повторить всё то же самое, а можно заметить, что набор данных описывается своим средним.

Упражнение. Докажите, что если $p(x_i \mid \mu) \sim \mathcal{N}(\mu, \sigma^2)$ и x_i независимы, то $p(\bar{x} \mid \mu) \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

Нормальное распределение: фиксируем σ

- Для апостериорной вероятности будет

$$p(\mu \mid x_1, \dots, x_n) \propto p(x_1, \dots, x_n \mid \mu) p(\mu) \propto p(\bar{x} \mid \mu) p(\mu) \propto p(\mu \mid \bar{x}).$$

- Подставляя в наш предыдущий результат, получим:

$$p(\mu \mid x_1, \dots, x_n) \sim \mathcal{N}\left(\frac{\sigma_0^2}{\sigma_0^2 + \frac{\sigma^2}{n}} x + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0, \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1} \right).$$

Нормальное распределение: фиксируем μ

- Если зафиксировать μ и менять σ^2, то сопряжённым априорным распределением будет обратное гамма-распределение:

$$p(\sigma^2 \mid \alpha, \beta) \propto IG(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{-\alpha-1} \exp\left(\frac{-\beta}{z}\right).$$

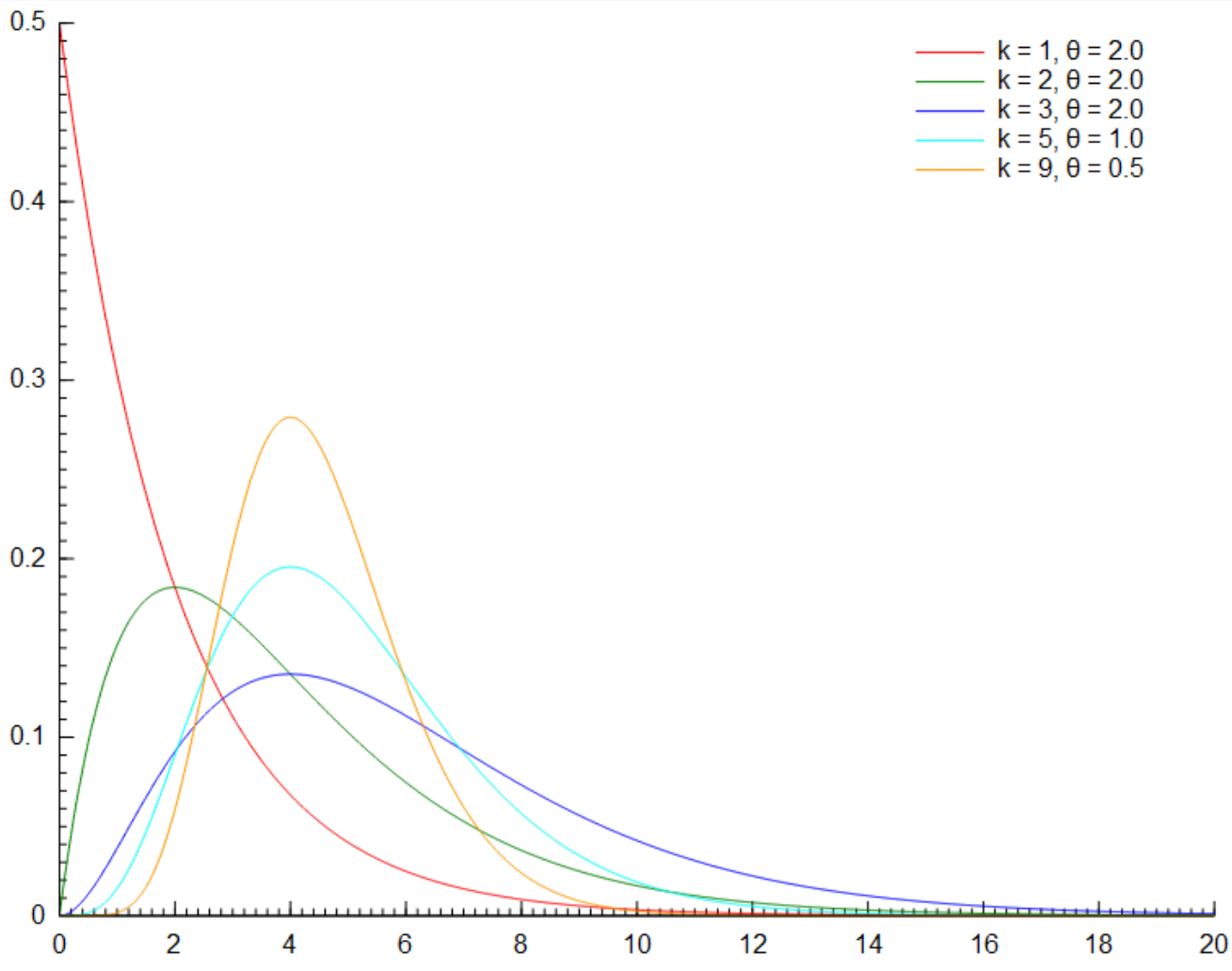
- Тогда в апостериорном распределении будет

$$p(\sigma^2 \mid x_1, \dots, x_n, \alpha, \beta) \propto IG\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum(x_i - \mu)\right).$$

- А в терминах $\tau = \frac{1}{\sigma^2}$ будет обычное гамма-распределение:

$$p(\tau \mid x_1, \dots, x_n, \alpha, \beta) \propto \text{Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum(x_i - \mu)\right).$$

Гамма--распределение



Когда и μ, и σ^2 меняются

- Что делать, когда и μ, и σ^2 меняются?
- Можно было бы предположить, что μ и σ^2 независимы; тогда просто априорное распределение будет

$$p(\mu, \sigma \mid \mu_0, \sigma_0, \alpha, \beta) \propto \mathcal{N}(\mu_0, \sigma_0^2) \cdot IG(\alpha, \beta).$$

- К сожалению, это распределение не будет сопряжённым к нормальному. Почему?

Когда и μ, и σ^2 меняются

- Что делать, когда и μ, и σ^2 меняются?
- Можно было бы предположить, что μ и σ^2 независимы; тогда просто априорное распределение будет

$$p(\mu, \sigma \mid \mu_0, \sigma_0, \alpha, \beta) \propto \mathcal{N}(\mu_0, \sigma_0^2) \cdot IG(\alpha, \beta).$$

- К сожалению, это распределение не будет сопряжённым к нормальному. Почему?
- Потому что μ и σ^2 зависимы. :) Новая точка x вводит зависимость между ними.

Когда и μ, и σ^2 меняются

- В настоящем сопряжённом априорном распределении будут:

$$\begin{aligned} x \mid \mu, \tau &\sim \mathcal{N}(\mu, \tau), \\ \mu \mid \tau &\sim \mathcal{N}(\mu_0, n_0\tau), \\ \tau &\sim G(\alpha, \beta). \end{aligned}$$

- Давайте выясним, как изменятся параметры, и заодно докажем.

Когда и μ, и σ^2 меняются

- Самое простое – это, по уже известным результатам,

$$\mu \mid x, \tau \sim \mathcal{N}\left(\frac{n\tau}{n\tau + n_0\tau}\bar{x} + \frac{n_0\tau}{n\tau + n_0\tau}\mu_0, n\tau + n_0\tau\right).$$

- Затем давайте разберёмся с $\tau \mid x$:

$$p(\tau, \mu \mid x) \propto p(\tau) \cdot p(\mu \mid \tau) \cdot p(x \mid \tau, \mu),$$

  и мы хотим это распределение маргинализовать по μ...

Когда и μ, и σ^2 меняются

- Подсчитаем:

$$
\begin{aligned}
p(\tau, \mu \mid x) &\propto p(\tau) \cdot p(\mu \mid \tau) \cdot p(x \mid \tau, \mu) \\
&\propto \tau^{\alpha-1} e^{-\tau\beta} \cdot \tau^{\frac{1}{2}} e^{-\frac{n_0 \tau}{2}(\mu-\mu_0)^2} \cdot \tau^{\frac{n}{2}} e^{-\frac{\tau}{2}\sum(x_i-\mu)^2} \\
&\propto \tau^{\alpha+\frac{n}{2}-\frac{1}{2}} e^{-\tau\left(\beta+\frac{1}{2}\sum(x_i-\bar{x})^2\right)} e^{-\frac{\tau}{2}\left(n_0(\mu-\mu_0)^2+n(\bar{x}-\mu)^2\right)}
\end{aligned}
$$

(простой трюк: $x_i - \mu = x_i - \bar{x} + \bar{x} - \mu$).

Когда и μ, и σ^2 меняются

- Теперь надо проинтегрировать

$$\int_{\mu} e^{-\frac{\tau}{2}\left(n_0(\mu-\mu_0)^2+n(\bar{x}-\mu)^2\right)} d\mu.$$

Упражнение. Проинтегрируйте. :) Должна получиться нормировочная константа

$$\tau^{-\frac{1}{2}} e^{\frac{-nn_0\tau}{2(n+n_0)}(\bar{x}-\mu_0)^2}.$$

Когда и μ, и σ^2 меняются

- Таким образом, получается апостериорное распределение

$$p(\tau \mid x) \propto \tau^{\alpha+\frac{n}{2}-1} e^{-\tau\left(\beta+\frac{1}{2}\sum(x_i-\bar{x})^2+\frac{nn_0}{2(n+n_0)}(\bar{x}-\mu_0)^2\right)}.$$

- Итого результаты такие:

$$\begin{aligned} \mu \mid \tau, x &\sim \mathcal{N}\left(\tfrac{n\tau}{n\tau+n_0\tau}\bar{x}+\tfrac{n_0\tau}{n\tau+n_0\tau}\mu_0, n\tau+n_0\tau\right), \\ \tau \mid x &\sim G\left(\alpha+\tfrac{n}{2}, \beta+\tfrac{1}{2}\sum(x_i-\bar{x})^2+\tfrac{nn_0}{2(n+n_0)}(\bar{x}-\mu_0)^2\right). \end{aligned}$$

Предсказание

- Теперь предсказание нового x_{new}:

$$p(x_{\text{new}} \mid x) = \int\int \underbrace{\text{Gamma}}_{\tau|x} \cdot \underbrace{\text{Gaussian}}_{\mu|\tau,x} \cdot \underbrace{\text{Gaussian}}_{x_{\text{new}}|\tau,\mu} d\tau d\mu =$$

$$= \int \underbrace{\text{Gamma}}_{\tau|x} \int \underbrace{\text{Gaussian}}_{\mu|\tau,x} \cdot \underbrace{\text{Gaussian}}_{x_{\text{new}}|\tau,\mu} d\tau d\mu =$$

$$= \int \underbrace{\text{Gamma}}_{\tau|x} \cdot \underbrace{\text{Gaussian}}_{x_{\text{new}}|\tau,x} d\tau = \ldots$$

- В результате получится распределение Стьюдента.

Упражнение. Проведите эти вычисления.

Экспоненциальное семейство

- Вообще говоря, всё, о чём мы говорили – частные случаи *экспоненциального семейства* распределений:

$$p(\mathbf{x} \mid \eta) = h(\mathbf{x})g(\eta)e^{\eta^\top \mathbf{u}(\mathbf{x})}.$$

- η называются *естественными параметрами* (natural parameters).

Экспоненциальное семейство

- Например, распределение Бернулли:

$$p(x \mid \mu) = \mu^x (1-\mu)^{1-x} = e^{x \ln \mu + (1-x)\ln(1-\mu)} = \\ = (1-\mu) e^{\ln\left(\frac{\mu}{1-\mu}\right) x},$$

и естественный параметр получился $\eta = \ln\left(\frac{\mu}{1-\mu}\right)$:

$$p(x \mid \eta) = \sigma(-\eta) e^{-\eta x},$$

где $\sigma(y) = \frac{1}{1+e^{-y}}$ – *сигмоид-функция*.

Экспоненциальное семейство

- Для мультиномиального распределения с параметрами $\mu_1, \dots, \mu_{M-1}$ получаются

$$\eta_k = \ln\left(\frac{\mu_k}{1 - \sum_j \mu_j}\right) \quad \text{и}$$

$$p(\mathbf{x} \mid \eta) = \left(1 + \sum_{k=1}^{M-1} e^{\eta_k}\right)^{-1} e^{\eta^\top \mathbf{x}}.$$

Упражнение. Проверьте!

Экспоненциальное семейство

- Так вот, для распределений из экспоненциального семейства

$$p(\mathbf{x} \mid \eta) = h(\mathbf{x})g(\eta)e^{\eta^\top \mathbf{u}(\mathbf{x})}$$

можно сразу оптом найти сопряжённые априорные распределения:

$$p(\eta \mid \chi, \nu) = f(\chi, \nu)g(\eta)^\nu e^{\nu\eta^\top \chi},$$

где χ – гиперпараметры, а g то же самое, что в исходном распределении.

Упражнение. Проверьте это и получите вышеописанные примеры как частные случаи.

Важные распределения

Важные распределения

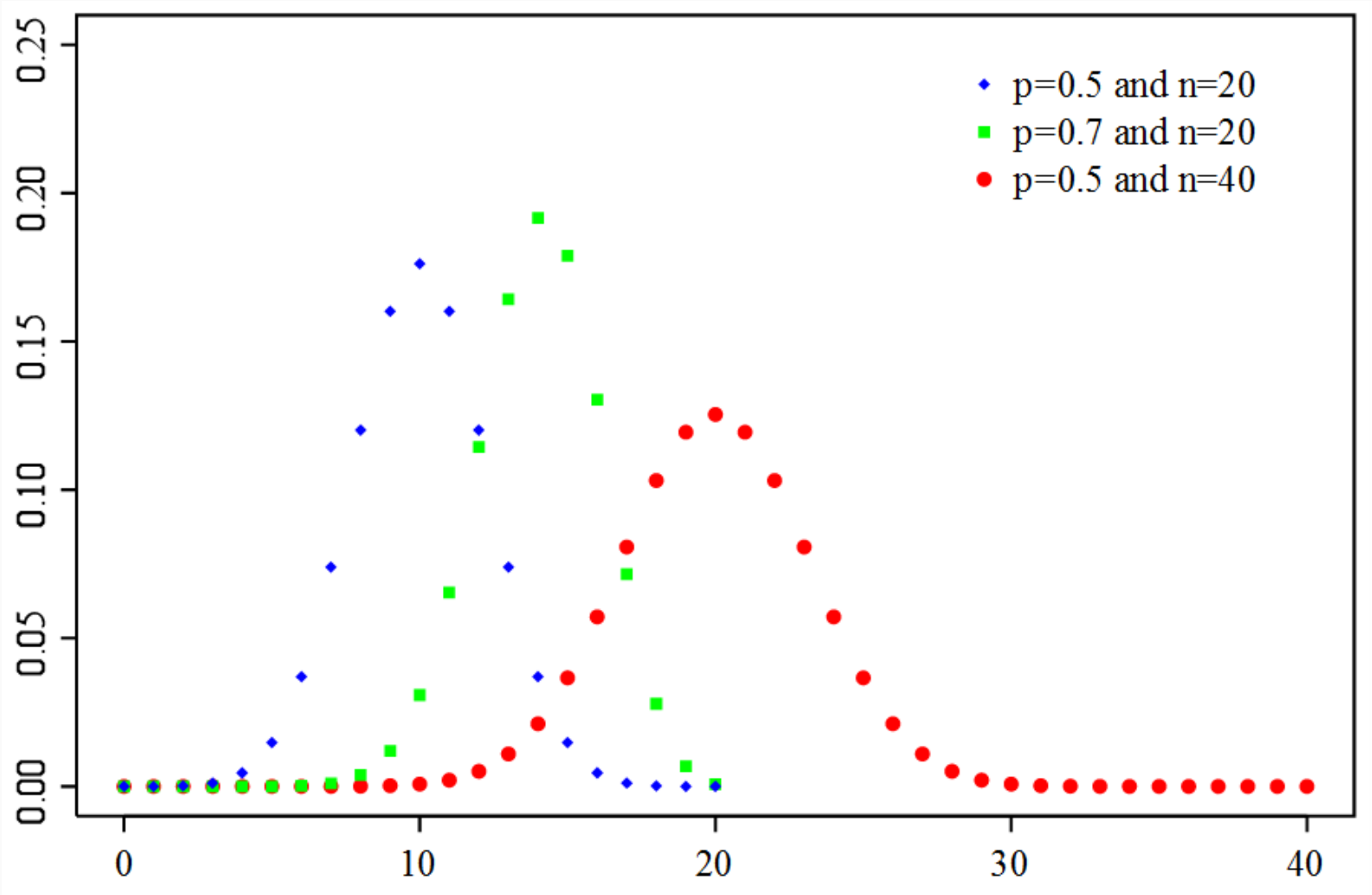
- Есть некоторое количество распределений, которые часто появляются в разных задачах и являются наиболее полезными на практике.
- Их полезно... ну, если не знать, то хотя бы быть знакомыми.
- Сейчас по ним и пробежимся.

Биномиальное распределение

- *Биномиальное распределение* возникает, когда мы подбрасываем нечестную монетку (вероятность решки p) n раз и хотим найти вероятность появления r решек.

$$p(r|p,n) = \binom{n}{r} p^r (1-p)^{n-r}.$$

Биномиальное распределение



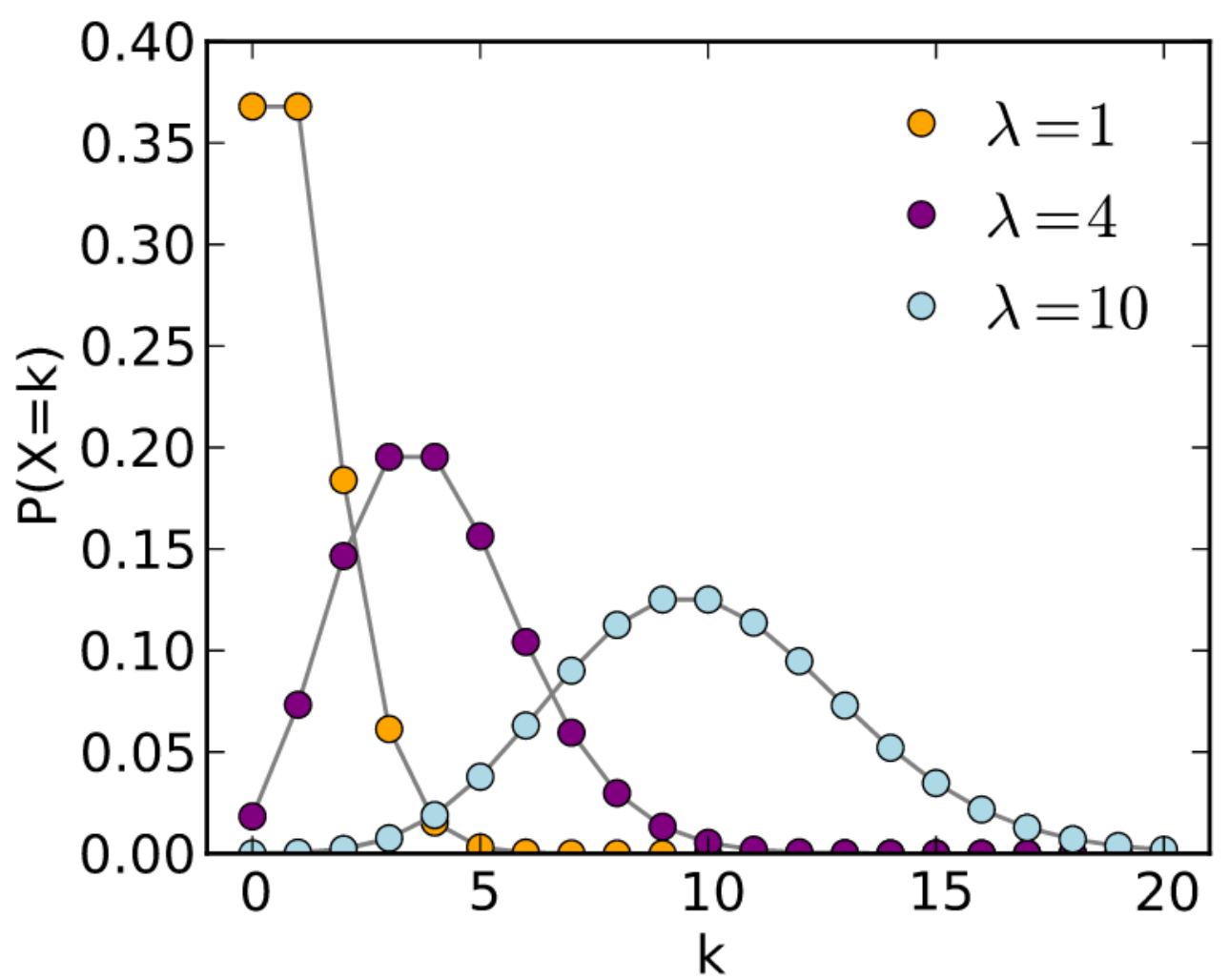
Пуассоновское распределение

- *Распределение Пуассона* возникает, когда мы хотим подсчитать количество событий за фиксированный интервал, если нам дана средняя интенсивность этих событий.
- Если ожидается в среднем λ событий за этот интервал, то вероятность того, что произойдут ровно r событий, равна

$$p(r|\lambda) = e^{-\lambda}\frac{\lambda^r}{r!}.$$

- Распределение Пуассона – это предельный случай биномиального распределения. Если n очень велико, а p очень мало, $\mathrm{Binomial}(n, p)$ будет очень похоже на $\mathrm{Poisson}(np)$.

Пуассоновское распределение



Нормальное распределение

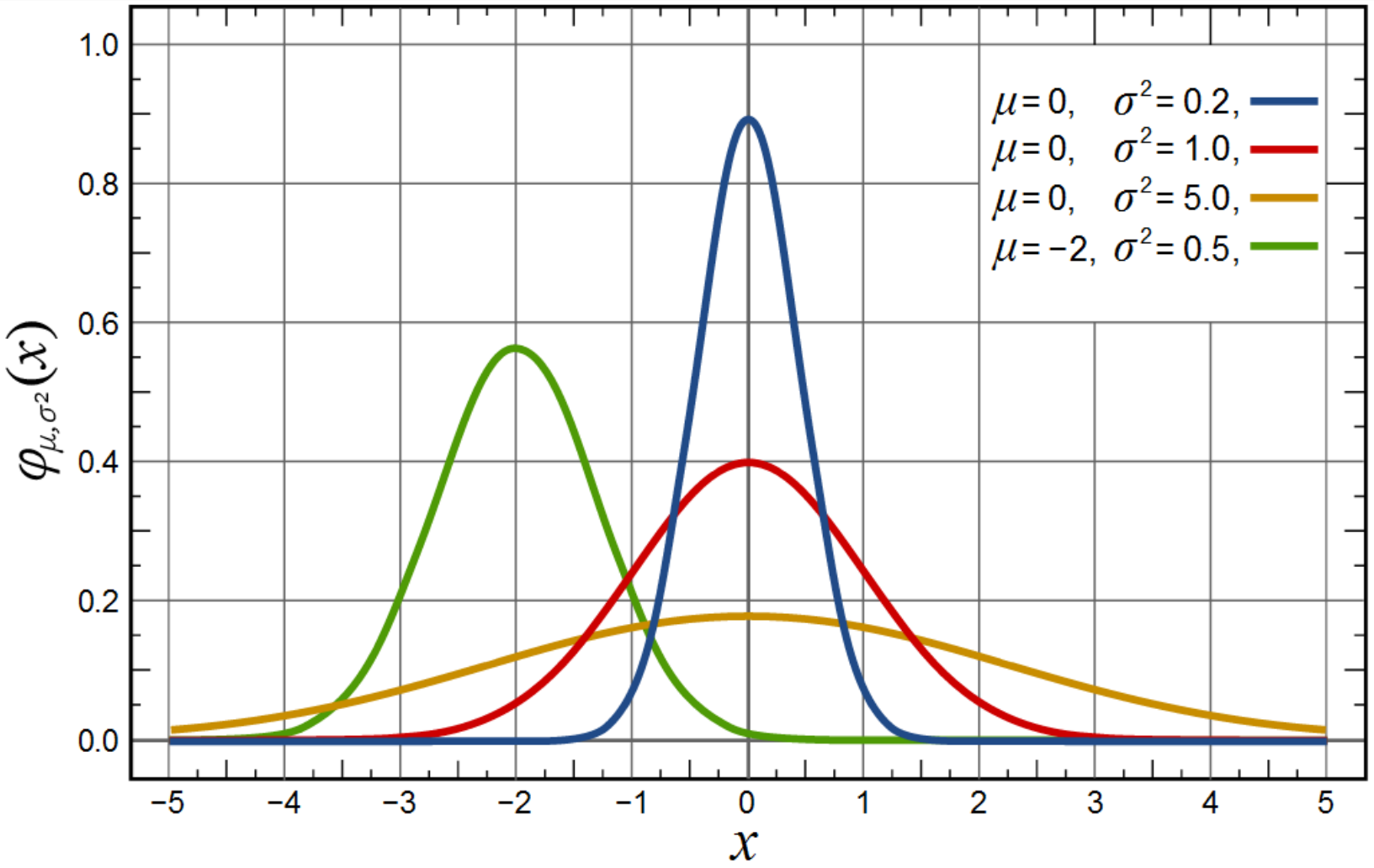
- Мы уже давно знаем нормальное распределение:

$$p(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

- Очень многие процессы могут моделироваться нормальным (гауссовским) распределением; обычно возникает, когда есть некое среднее значение μ и шум вокруг него.
- Функция правдоподобия данных $x_1, \dots, x_n$:

$$p(x_1, \dots, x_n|\mu,\sigma) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}}.$$

Нормальное распределение



Гауссиан: достаточные статистики

- Заметим, что функция эта зависит от двух параметров, а не от n:

$$p(x_1, \dots, x_n|\mu, \sigma) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{S+n(\bar{x}-\mu)^2}{2\sigma^2}},$$

  где

$$\bar{x} = \frac{x_1 + \dots + x_n}{n}, \quad S = \sum_{i=1}^{n} (\bar{x} - x_n)^2.$$

- Параметры $\bar{x}$ и S называются *достаточными статистиками* (sufficient statistics).

Гауссиан: ГМП

- Какие параметры лучше всего описывают данные?
- Перейдём, как водится, к логарифму:

$$\ln p(x_1, \dots, x_n | \mu, \sigma) = -n \ln(\sigma\sqrt{2\pi}) - \frac{S + n(\bar{x} - \mu)^2}{2\sigma^2}.$$

- Как выяснить, при каких параметрах функция правдоподобия максимизируется?

Гауссиан: ГМП

- Какие параметры лучше всего описывают данные?
- Перейдём, как водится, к логарифму:

$$\ln p(x_1, \dots, x_n | \mu, \sigma) = -n \ln(\sigma\sqrt{2\pi}) - \frac{S + n(\bar{x} - \mu)^2}{2\sigma^2}.$$

- Как выяснить, при каких параметрах функция правдоподобия максимизируется?
- Взять частные производные и приравнять нулю.

Гауссиан: ГМП

- По μ:
$$\frac{\partial \ln p}{\partial \mu} = -\frac{n}{\sigma^2}(\mu - \bar{x}).$$
- То есть в гипотезе максимального правдоподобия $\mu_{ML} = \bar{x}$, независимо от S.
- Теперь нужно найти σ из гипотезы максимального правдоподобия.
- Для этого мы продифференцируем по $\ln \sigma$ — полезный приём на будущее. Кстати, $\frac{\mathrm{d}x^n}{\mathrm{d}(\ln x)} = nx^n$.

Гауссиан: ГМП

- $$\frac{\partial \ln p}{\partial \ln \sigma} = -n + \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{\sigma^2}.$$
- Следовательно, в гипотезе максимального правдоподобия

$$\sigma_{ML} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}}.$$

Упражнение. Докажите, что это смещённая оценка, т.е. ожидание этой оценки по настоящему нормальному распределению не равно σ^2.

Распределение Стьюдента

- Распределение Стьюдента применяется, когда нужно искать доверительные интервалы на параметры нормального распределения.
- Величина $T = \frac{\bar{x}_n - \mu}{S_n/\sqrt{n}}$ распределена по закону

$$f(t) = \frac{\Gamma(n/2)}{\sqrt{\pi(n-1)}\Gamma(\frac{n-1}{2})}\left(1 + \frac{t^2}{n-1}\right)^{-n/2}.$$

- Если мы выберем число A так, чтобы $p(-A < T < A) = 1 - \alpha$, то

$$\left[\bar{x}_n - A\frac{S_n}{\sqrt{n}}, \bar{x}_n + A\frac{S_n}{\sqrt{n}}\right]$$

будет доверительным интервалом для μ с вероятностью ошибки α.

Экспоненциальное распределение

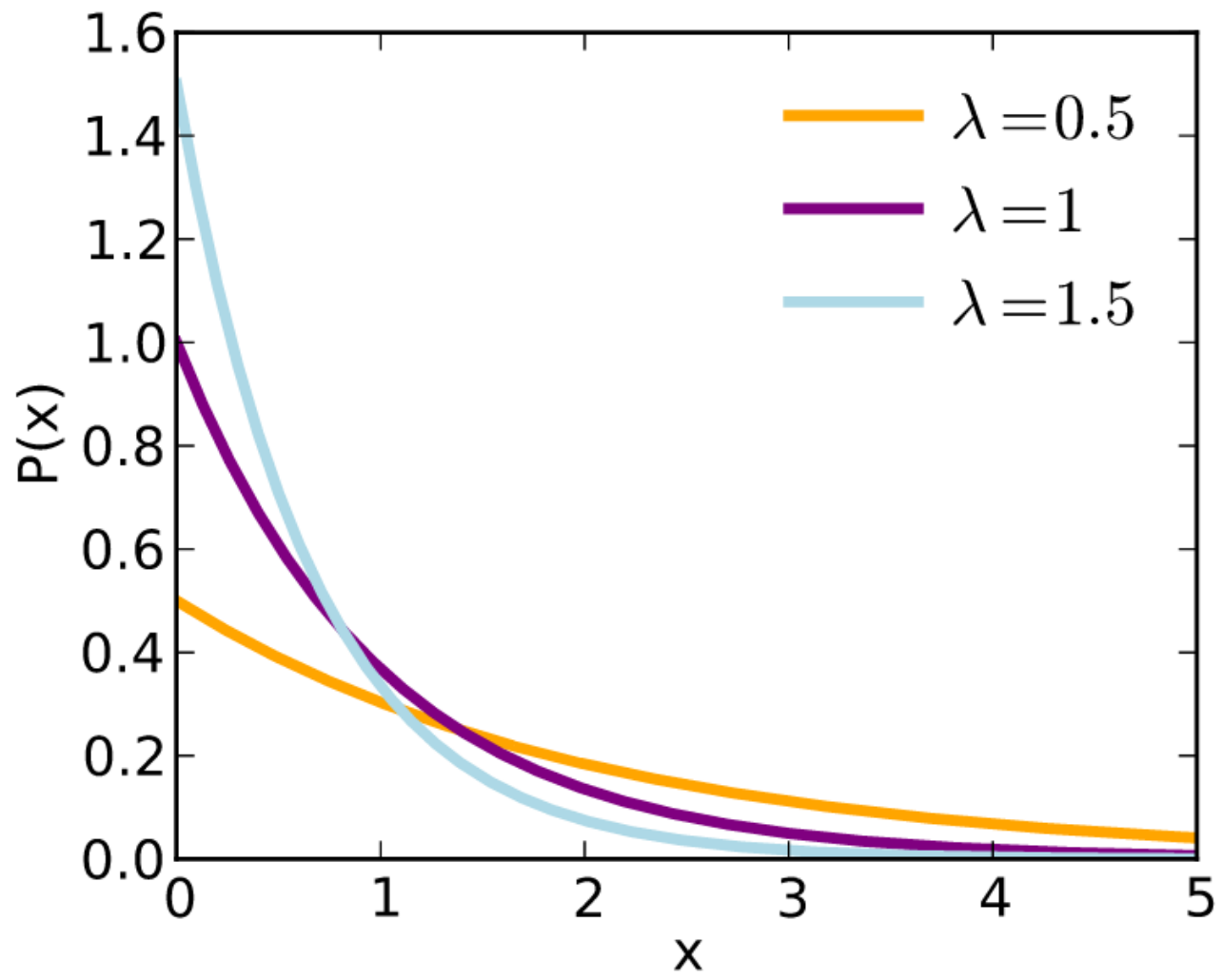
- Обычно возникает в ответах на вопрос «сколько надо ждать события».
- Дискретный вариант: вероятность того, что выпадения орла на нечестной монетке придётся ждать ровно r шагов, равна

$$p(r|p) = p^r(1-p) = (1-p)e^{-\lambda r}, \text{ где } \lambda = \ln\frac{1}{p}.$$

- Непрерывный вариант: если мы ждём события, которое происходит в среднем каждые $1/\lambda$ единиц времени (в пуассоновском процессе с интенсивностью λ), то распределение времени ожидания

$$p(x|s) = \lambda e^{-\lambda x}.$$

Экспоненциальное распределение

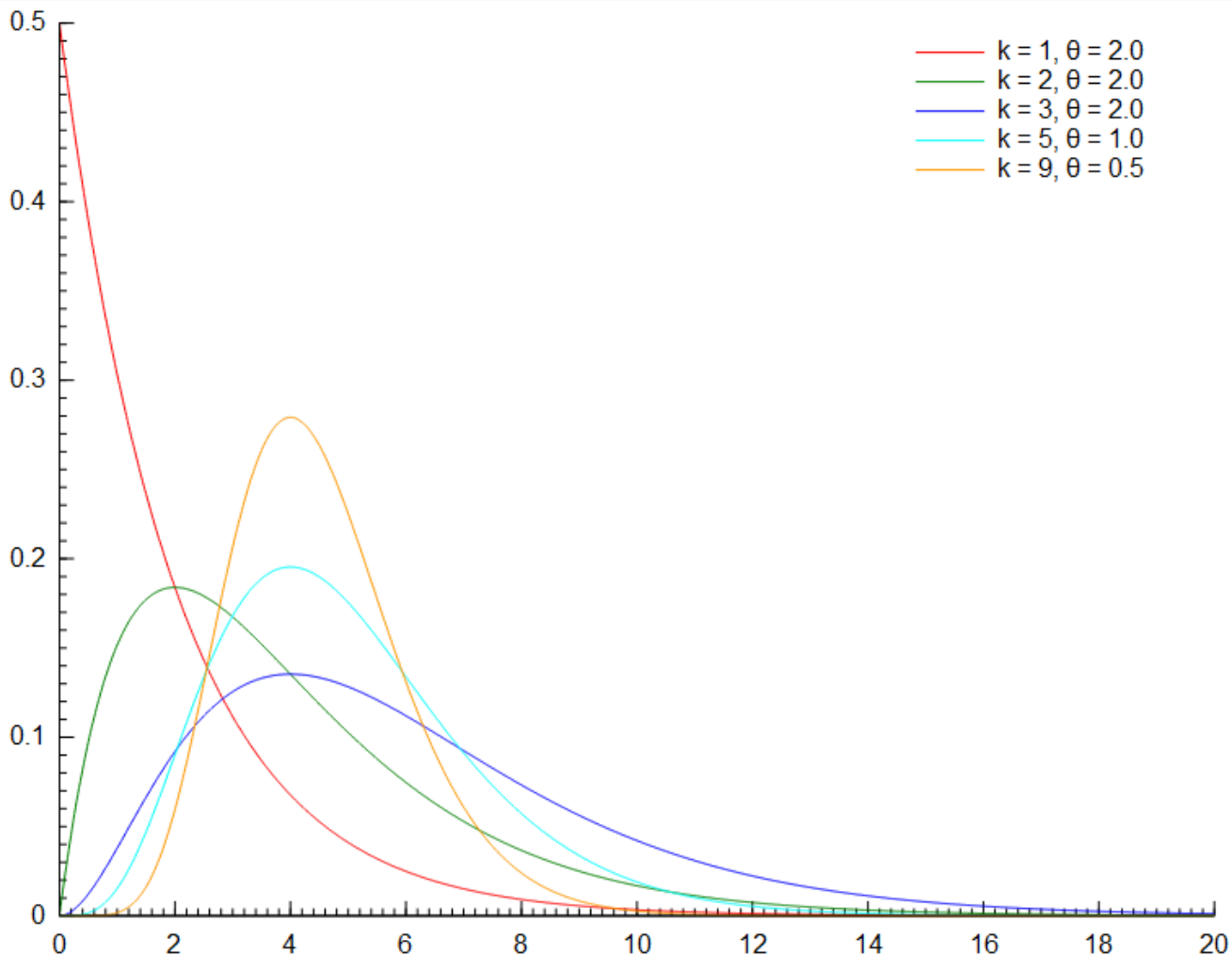


Гамма--распределение

- Может возникать как сумма нескольких экспоненциальных распределений.
- Плотность распределения

$$p(x|k,\theta) = x^{k-1}\frac{e^{-x/\theta}}{\theta^k\Gamma(k)}, \quad x > 0.$$

Гамма-распределение



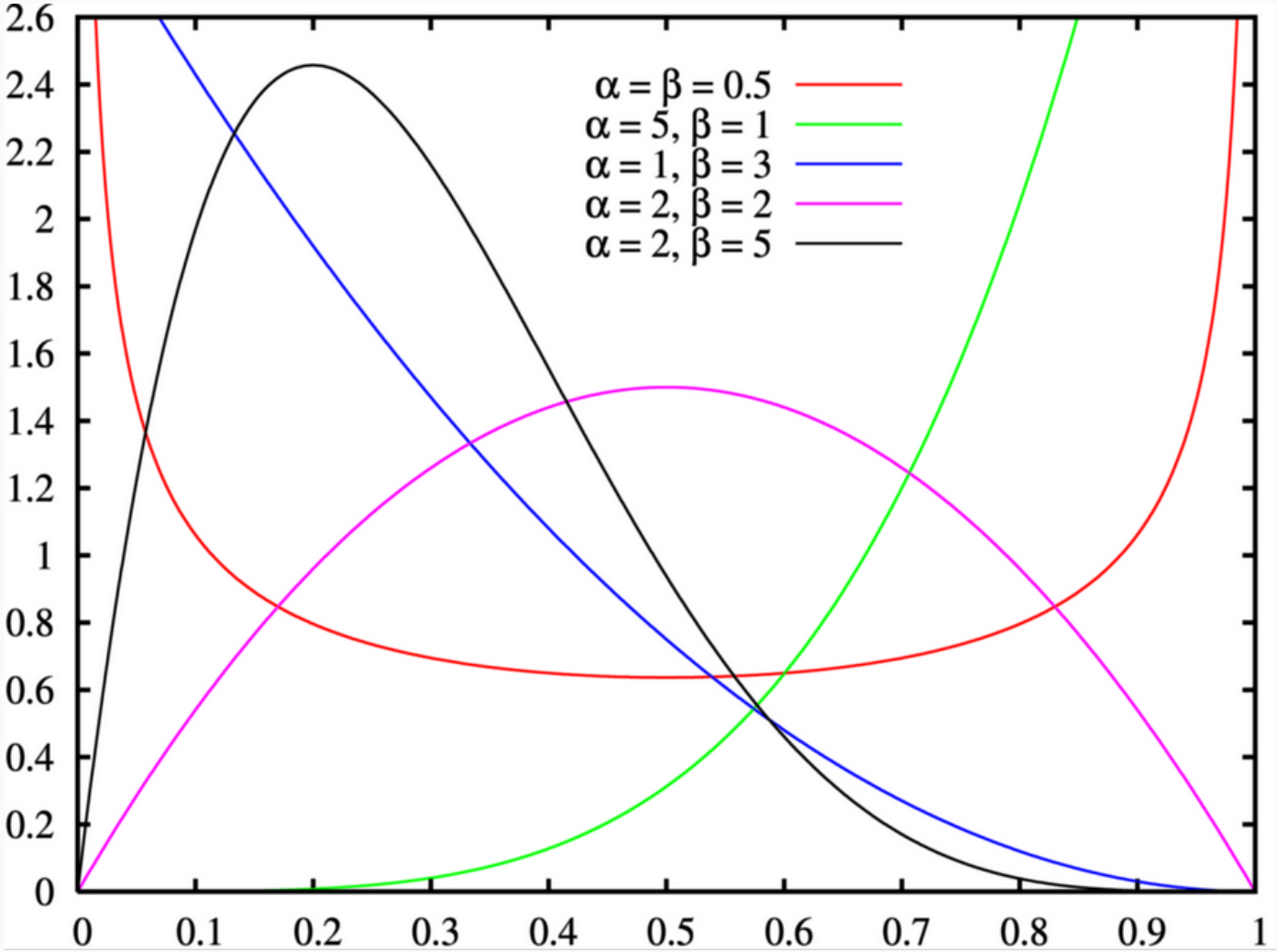
Бета--распределение

- Бета-распределение определяется над вероятностями, т.е. на интервале $[0, 1]$.
- Часто служит априорным распределением для каких-либо вероятностей; является сопряжённым априорным распределением для испытаний Бернулли:

$$\mathrm{Beta}(x|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad \beta(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

- $\mathrm{Beta}(i, j)$ — это распределение i-й по величине случайной величины из $i + j - 1$ случайных величин, распределённых равномерно на $[0, 1]$.

Бета--распределение



Порядковые статистики

- Бета–распределение — частный случай *порядковой статистики* (order statistics).
- Если n случайных величин $X_1, \dots, X_n$ распределены одинаково с функцией распределения F, то k-я порядковая статистика $Y_k^{(n)}(y)$ — это функция распределения k-й сверху величины из этих n величин.
- Например, очевидно, что

$$Y_1^{(n)}(y) = F(y)^{n-1}.$$

Упражнение. Вывести формулу для второй порядковой статистики $Y_2^{(n)}$.

Распределение Дирихле

- Обобщение бета–распределения на многомерный случай.
- На k-мерном симплексе $\{x \mid \sum_{i=1}^{k} x_i = 1\}$ плотность распределения Дирихле с параметром $\alpha = (\alpha_1, \dots, \alpha_k)$ равна

$$p(x|\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}, \quad B(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}.$$

Проклятие размерности

Параметрические и непараметрические модели

- Прежде чем узнать, что делать, общее 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

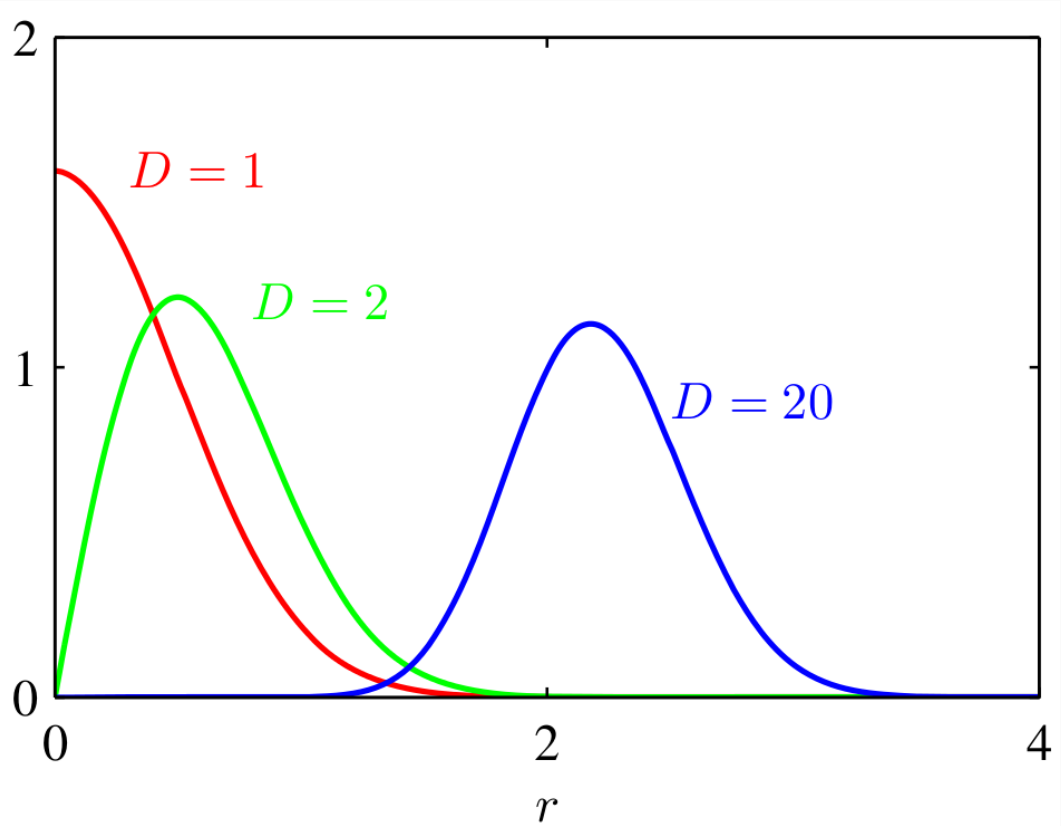
- Третье проявление: проблемы в оптимизации, которые и имел в виду Беллман.
- Если нужно примерно оптимизировать функцию от d переменных, на решётке с шагом ϵ понадобится примерно $\left(\frac{1}{\epsilon}\right)^d$ вычислений функции.
- В численном интегрировании – чтобы интегрировать функцию с точностью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