# Байесовские методы. Лекция 5.
# Семплирование. MCMC.

Целищев М.А.

МГУ им. М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математической статистики

весна 2021

## Список литературы


- J. Kruschke. Doing Bayesian Data Analysis, Second Edition. A Tutorial with R, JAGS, and Stan. Academic Press, 2014.
- R. McElreath. Statistical Rethinking. A Bayesian Course with Examples in R and Stan. Chapman and Hall CRC, 2015.
- O. Martin. Bayesian Analysis with Python. Introduction to Statistical Modeling and Probabilistic Programming using PyMC3 and ArviZ, Second Edition. Packt, 2018.
- K. P. Murphy. Machine Learning: A Probabilistic Perspective. The MIT Press, 2012.
- A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, D. Rubin. Bayesian Data Analysis, Third Edition. CRC Press, 2013.
- S. Brooks, A. Gelman, G. L. Jones and X.-L. Meng. Handbook of Markov Chain Monte Carlo. Chapman & Hall/CRC, 2011.

# Напоминание

Итак, если $D$ — наблюдаемые данные, а $\theta$ — неизвестные параметры, то хотят найти апостериорное распределение $p(\theta|D)$:

$$p(\theta|D) = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}} = \frac{p(D|\theta)\,p(\theta)}{p(D)} = \frac{p(D|\theta)\,p(\theta)}{\int p(D|\theta)\,p(\theta)d\theta}.$$

Проблема в подсчёте интеграла в знаменателе.

Если априорное *сопряжено* к правдоподобию (как это было в прошлых лекциях), то $p(\theta|D)$ удаётся посчитать аналитически (в явном виде), но так бывает нечасто.

Можно попробовать численно посчитать интеграл в знаменателе, но такой подход крайне неэффективен в случае, когда параметр $\theta$ большой размерности.

Рассмотрим чуть более общую задачу: подсчёт мат. ожидания функции от случайной величины (или вектора) $\xi$ с плотностью $p(x)$:

$$\mathbf{E}f(\xi) = \int f(x)\,p(x)\,dx.$$

# Метод Монте-Карло

$$\mathbf{E}f(\xi) = \int f(x)p(x)\,dx.$$

Если мы умеем генерировать выборку $X_1, X_2, \dots, X_n$ из распределения $\xi$, то несмещённой и состоятельной оценкой для $\mathbf{E}f(\xi)$ будет

$$\eta = \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

Это следует из закона больших чисел. Более точно, по центральной предельной теореме распределение случайной величины $\eta$ близко к $\mathcal{N}\left(\mathbf{E}f(\xi), \frac{1}{n}\mathbf{D}f(\xi)\right)$.

Мораль: хорошо бы уметь генерировать выборку (семплировать) из распределения с известной плотностью $p(x)$.

## Семплирование

Сначала рассмотрим одномерный случай, т.е. $\xi$ — случайная величина с плотностью $p(x)$, $x \in \mathbb{R}$, и функцией распределения $F(x) = \mathbf{P}(\xi < x) = \int\limits_{-\infty}^{x} p(y)\,dy$, $x \in \mathbb{R}$.

### Теорема

Если $F$ — строго возрастающая и непрерывная функция распределения случайной величины $\xi$ и $U \sim \mathcal{U}[0,1]$, то

$$F^{-1}(U) \overset{\mathrm{d}}{=} \xi,$$

то есть $F^{-1}(U)$ имеет то же распределение, что и $\xi$.

*Доказательство.* Для каждого $x \in \mathbb{R}$ имеем:

$$\mathbf{P}(F^{-1}(U) < x) = \mathbf{P}(U < F(x)) = F(x). \quad \square$$

**ДЗ:** показать, что результат остаётся в силе, даже если не требовать непрерывности и строгого возрастания $F$ (под $F^{-1}$ при этом понимают квантильную функцию).

Иными словами, если умеем семплировать из равномерного распределения, то теоретически умеем генерировать выборку из любого распределения на прямой.

**Пример**. Пусть $\xi \sim \mathrm{Exp}(\lambda)$, $\lambda > 0$. Тогда $p(x) = \lambda e^{-\lambda x}$, $x > 0$, и $F(x) = 1 - e^{-\lambda x}$, $x > 0$. Как найти $F^{-1}$?

$$u = 1 - e^{-\lambda x} \qquad \Longrightarrow \qquad x = -\frac{1}{\lambda}\ln(1-u),$$

то есть $F^{-1}(u) = -\frac{1}{\lambda}\ln(1-u)$. Таким образом, если $U \sim \mathcal{U}[0,1]$, то

$$-\frac{1}{\lambda}\ln U \sim \mathrm{Exp}(\lambda).$$

**ДЗ:** Этим же способом построить алгоритм семплирования из распределения $\mathrm{Cauchy}(a,b)$, $a \in \mathbb{R}$, $b > 0$.

# Семплирование из $\mathcal{N}(0,1)$

Как семплировать из $\mathcal{N}(0,1)$? Функция стандартного нормального распределения есть

$$\Phi(x) = \int\limits_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, dy,$$

так что формально $\Phi^{-1}(U) \sim \mathcal{N}(0,1)$, где $U \sim \mathcal{U}[0,1]$.

Проблема в том, что ни $\Phi$, ни $\Phi^{-1}$ не выражаются через элементарные функции, и их подсчёт не всегда тривиален.

Можно воспользоваться ЦПТ, и если $U_1, U_2, \ldots \overset{\text{i.i.d.}}{\sim} \mathcal{U}[0,1]$, то тогда распределение с.в.

$$\eta = \sum_{i=1}^{12} U_i - 6$$

будет близко к $\mathcal{N}(0,1)$. Почему 12?

# Семплирование из $\mathcal{N}(0, 1)$

Но на практике обычно делают по-другому.

А именно, берут независимые $\mathcal{E} \sim \mathrm{Exp}(0.5)$ и $U \sim \mathcal{U}[0, 2\pi]$. Тогда пара случайных величин

$$\begin{cases} \xi_1 = \sqrt{\mathcal{E}} \cos U, \\ \xi_2 = \sqrt{\mathcal{E}} \sin U, \end{cases}$$

будут независимы и обе $\xi_1, \xi_2 \sim \mathcal{N}(0, 1)$.

**ДЗ(*):** доказать этот факт.

Итак, чтобы сгерерировать наблюдение из $\mathcal{N}(0, 1)$,
нужно взять независимые $U_1, U_2 \sim \mathcal{U}[0, 1]$ и вычислить

$$\xi_1 = \sqrt{-2 \ln U_1} \cos(2\pi U_2).$$

## Семплирование из $\mathcal{N}(\mu, \Sigma)$, $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d\times d}$

Вычислим спектральное разложение матрицы $\Sigma$:

$$\Sigma = Q\Lambda Q^T,$$

где $\Lambda = \operatorname{diag}\{\lambda_1, \ldots, \lambda_d\}$, $\lambda_j \geqslant 0$ — собственные значения $\Sigma$,
а $j$-ый столбец ортогональной матрицы $Q \in \mathbb{R}^{d\times d}$ есть соответствующий собственный вектор. Обозначим $\Lambda^{1/2} = \operatorname{diag}\{\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_d}\}$.

### Теорема

Если $\xi \sim \mathcal{N}(0, I)$, то $\eta = \mu + Q\Lambda^{1/2}\xi \sim \mathcal{N}(\mu, \Sigma)$.

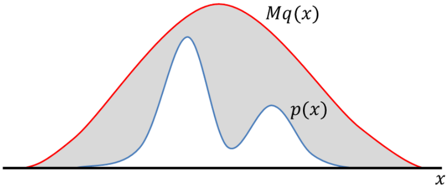
*Доказательство.* Напомним, что х.ф. $\mathcal{N}(\mu, \Sigma)$ есть $\varphi(t) = \exp\left(it^T\mu - \frac{1}{2}t^T\Sigma t\right)$.
Тогда $\varphi_\xi(t) = \mathbf{E}\exp\left(it^T\xi\right) = \exp\left(-\frac{1}{2}t^Tt\right)$, $t \in \mathbb{R}^d$, и потому

$$\varphi_\eta(t) = \mathbf{E}\exp\left(it^T\left(\mu + Q\Lambda^{1/2}\xi\right)\right) = \exp\left(it^T\mu\right)\cdot\varphi_\xi(\Lambda^{1/2}Q^Tt) =$$

$$= \exp\left(it^T\mu\right)\cdot\exp\left(-\frac{1}{2}t^TQ\Lambda^{1/2}\Lambda^{1/2}Q^Tt\right) = \exp\left(it^T\mu - \frac{1}{2}t^T\Sigma t\right). \quad \square$$

## Rejection sampling

Пусть не получается в явном виде семплировать из распределения с плотностью $p(x)$. Но допустим, что умеем сэмплировать из распределения с плотностью $q(x)$ (proposal distribution), причём $p(x) \leqslant Mq(x)\ \forall x \in \mathbb{R}$ для некоторго числа $M \geqslant 1$.




Алгоритм выборки с отклонением:

1. Семплируем $X$ из $q(x)$
2. Семплируем $U \sim \mathcal{U}[0, Mq(X)]$
3. Если $U < p(X)$, то возвращаем $X$, иначе goto (1)

Тогда $\mathbf{P}(\text{Accept}) = \mathbf{E}\frac{p(X)}{Mq(X)} = \int \frac{p(x)}{Mq(x)} q(x)\,dx = 1/M$.
Условная плотность $X$ в случае, если приняли наблюдение есть

$$\lim_{\delta\to+0} \frac{\mathbf{P}\left(X \in [x, x+\delta) | \text{Accept}\right)}{\delta} = \lim_{\delta\to+0} \frac{\mathbf{P}\left(X \in [x, x+\delta)\right) \mathbf{P}\left(\text{Accept} | X \in [x, x+\delta)\right)}{\delta\, \mathbf{P}(\text{Accept})} =$$

$$= \lim_{\delta\to+0} \frac{q(x)\delta\ p(x) / \left(Mq(x)\right)}{\delta/M} = p(x)$$ — что и требовалось.

**Недостаток:** если неудачно подобрали $q$, то $M$ будет большим и отклонения будут очень часто...

**Замечание.** Эта техника остаётся справедливой, если плотность $p(x)$ мы знаем только с точностью до нормировочной константы, т.е. $\tilde{p}(x)$ — ненормированная плотность и $\tilde{p}(x) \leqslant Mq(x)\ \forall x \in \mathbb{R}$. На третьем шаге алгоритма надо возвращать $X$, если $U < \tilde{p}(X)$. **ДЗ:** убедиться в корректности этого замечания.

Wait, what??? То есть существует теоретическая возможность генерировать выборку из распределения, плотность которого нам известна с точностью до нормировочной константы??? Ведь это как раз случай байесовского вывода, когда $\quad p(\theta|D) \ \propto\ p(D|\theta)\, p(\theta)$.

Если мы теперь сможем генерировать выборку $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)}$ из $p(\theta|D)$, то можем оценивать $p(D_{\text{new}}|D) = \int p(D_{\text{new}}|\theta)p(\theta|D)\, d\theta$ с помощью $\frac{1}{N}\sum\limits_{i=1}^{N} p(D_{\text{new}}|\theta^{(i)})$, то есть усредняя правдоподобие по сэмплам из апостериорного распределения $\theta$.

## Importance sampling

Рассмотрим ещё один метод оценки $\mathbf{E}f(\xi) = \int f(x)p(x)\,dx$, если не умеем семплировать из $p(x)$, но умеем семплировать из $q(x)$.

$$\mathbf{E}f(\xi) = \int f(x)p(x)\,dx = \int f(x)\frac{p(x)}{q(x)}q(x)\,dx = \mathbf{E}\left[f(\eta)\frac{p(\eta)}{q(\eta)}\right],$$

где $\eta$ имеет плотность $q(x)$. Последний интеграл оценивается по выборке $X_1, \ldots, X_n$ из $q(x)$ методом Монте-Карло: $\frac{1}{n}\sum\limits_{i=1}^{n}\frac{p(X_i)}{q(X_i)}f(X_i)$. **Проблема:** если $p$ и $q$ сильно отличаются, то многие веса $w_i = \frac{p(X_i)}{q(X_i)}$ будут пренебрежимо малы...

**Замечание:** Если $p(x) = \frac{1}{c}\,\tilde{p}(x)$, где $c$ — неизвестная нормировочная константа, то $c = \int \tilde{p}(x)\,dx = \int \frac{\tilde{p}(x)}{q(x)}q(x)\,dx = \mathbf{E}\left[\frac{\tilde{p}(\eta)}{q(\eta)}\right]$ и

$$\mathbf{E}f(\xi) = \mathbf{E}\left[f(\eta)\frac{\tilde{p}(\eta)}{q(\eta)}\right] \Big/ \mathbf{E}\left[\frac{\tilde{p}(\eta)}{q(\eta)}\right],$$

что тоже оценивается методом Монте-Карло по выборке из $q(x)$.

# Цепи Маркова

Последовательность случайных величин (или векторов) $X_1, X_2, \dots$ называют *цепью Маркова* с областью значений $\mathscr{X}$, если

$$p_n(x_n|x_1, x_2, \dots, x_{n-1}) = p_n(x_n|x_{n-1}) \quad \forall n = 2, 3, \dots,$$

где $p_n(x_n|x_j) = \mathbf{P}(X_n = x_n|X_j = x_j)$ в случае дискретного $\mathscr{X}$, или $p_n(x_n|x_j)$ — условная плотность $X_n$ при условии $X_j = x_j$ в абсолютно непрерывном случае.

Иначе говоря, динамика Марковской цепи такова, что при фиксированном настоящем $(x_{n-1})$ будущее $(x_n)$ не зависит от прошлого $(x_1, \dots, x_{n-2})$.

Этот эффект также называют *«отсутствием памяти»*.

Цепь Маркова называется *однородной*, если

$$p_2(y|x) = p_3(y|x) = \dots =: p(y|x).$$

$p(y|x)$ называют вероятностями (или плотностями) перехода из $x$ в $y$.

## Стационарное распределение

Распределение $\pi(x)$ на $\mathscr{X}$ называется *стационарным* для однородной цепи Маркова с вероятностями перехода $p(y|x)$, если

$$\int_{\mathscr{X}} p(y|x)\,\pi(x)\,dx = \pi(y).$$

Пусть $p(x_1)$ — произвольное начальное распределение марковской цепи. Будет ли распределение $X_n$

$$p(x_n) = \int p(x_n|x_{n-1})\ldots p(x_2|x_1)\,p(x_1)\,dx_1\,dx_2\ldots dx_{n-1}$$

сходиться к стационарному при $n \to \infty$? Не всегда...

**Теорема (достаточное условие сходимости цепи Маркова)**

Если $p(y|x) > 0$ для всех $x, y \in \mathscr{X}$, то существует стационарное распределение $\pi$, причём распределение $p(x_n)$ величины $X_n$ будет сходиться к $\pi$ при $n \to \infty$, каково бы ни было начальное распределение $p(x_1)$.

# Markov Chain Monte Carlo

**Задача:** сэмплировать из некоторого распределения $\pi(x)$.

**Идея:** если у нас есть цепь Маркова со стационарным распределением $\pi(x)$, то рано или поздно цепь *выйдет на стационарный режим*, то есть с некоторого номера $X_N, X_{N+1}, \ldots$ можно считать сэмплами из $\pi(x)$.

**Проблема 1**: $X_N, X_{N+1}, \ldots$ будут зависимы. Но так ли это плохо?

### Эргодическая теорема (УЗБЧ для марковских цепей)

Если $X_1, X_2, X_3 \ldots$ — цепь Маркова с $p(y|x) > 0 \;\; \forall x, y \in \mathscr{X}$, со стационарным распределением $\pi$, и $f \colon \mathscr{X} \to \mathbb{R}$ такова, что $\mathbf{E}_\pi f := \int\limits_{\mathscr{X}} f(x)\,\pi(x)\,dx$ конечно. Тогда

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \xrightarrow{\text{п.в.}} \mathbf{E}_\pi f \quad \text{при } n \to \infty.$$

Итак, зависимость сэмплов — не очень сильная проблема для подсчёта $\mathbf{E}_\pi f$.

**Проблема 2**: Как построить цепь Маркова, чтобы её стационарное распределение было $\pi(x)$? Иными словами, как выбрать вероятности перехода $p(y|x)$?

### Теорема (условие детального баланса)

Пусть дана однородная цепь Маркова с вероятностями перехода $p(y|x)$. Тогда если для некоторого распределения $\pi(x)$ выполнено *условие детального баланса*:

$$p(y|x)\,\pi(x) = p(x|y)\,\pi(y) \quad \forall x, y \in \mathscr{X},$$

то $\pi(x)$ — стационарное распределение этой цепи Маркова.

*Доказательство.*

$$\int\limits_{\mathscr{X}} p(y|x)\,\pi(x)\,dx = \int\limits_{\mathscr{X}} p(x|y)\,\pi(y)\,dx = \pi(y). \quad \square$$

## Metropolis algorithm

(1949) Metropolis N. and Ulam S.

(1953) Metropolis N., Rosenbluth A. & M., Teller A. & E.

Будем строить марковскую цепь по следующему алгоритму:

1. Пусть уже есть наблюдение $X_n = x_n$.
2. Генерируем $X_*$ из некоторого $q(x_*|x_n) > 0$ (proposal), из которого умеем семплировать, например $\mathcal{N}(x_n, \sigma^2 I)$.
3. $X_{n+1} = \begin{cases} X_* & \text{с вероятностью } \alpha = \min\left(\dfrac{\pi(X_*)}{\pi(X_n)}, 1\right) \\ X_n & \text{иначе (т.е. отбрасывание } X_*\text{, reject)} \end{cases}$

### Теорема

Если $q(x|y)$ — симметрично, то есть $q(x|y) = q(y|x)$, то для построенной цепи $X_1, X_2, \ldots$ распределение $\pi$ удовлетворяет условию детального баланса и потому стационарно.

*Доказательство.* $\forall x, y \in \mathscr{X}$, $x \neq y$:

$$p(y|x)\pi(x) = q(y|x)\min\left(\frac{\pi(y)}{\pi(x)}, 1\right)\pi(x) = q(x|y)\min(\pi(y), \pi(x)) = p(x|y)\pi(y).$$

Hastings W. (1970)

## Metropolis-Hastings algorithm

Можно изменить алгоритм, избавившись от симметричности $q(x|y)$:

1. Пусть уже есть наблюдение $X_n = x_n$.
2. Генерируем $X_*$ из некоторого $q(x_*|x_n) > 0$ (proposal), из которого умеем семплировать, например $\mathcal{N}(x_n, \sigma^2 I)$.
3. $X_{n+1} = \begin{cases} X_* & \text{с вер-тью } \alpha = \min\left(\dfrac{\pi(X_*)}{\pi(X_n)} \cdot \dfrac{q(X_n|X_*)}{q(X_*|X_n)}, 1\right) \\ X_n & \text{иначе (т.е. отбрасывание } X_*\text{, reject)} \end{cases}$

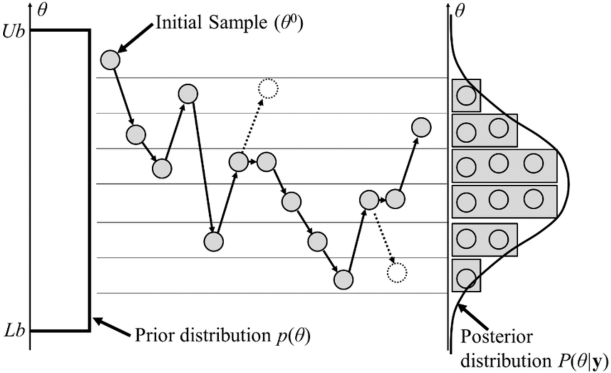
ДЗ: показать, что для такой цепи распределение $\pi$ удовлетворяет условию детального баланса и потому стационарно.

**Замечание**: для применения этого метода целевое распределение $\pi = \frac{\tilde{\pi}}{c}$ можно знать с точностью до нормировочной константы $c$, поскольку

$$\frac{\tilde{\pi}(y)}{\tilde{\pi}(x)} = \frac{\pi(y)}{\pi(x)}.$$

Это позволяет использовать метод Метрополиса-Гастингса для семплирования из апостериорного распределения в байесовском выводе!

## MCMC




Метод Метрополиса-Гастингса будет эффективен, если reject происходит редко. Если целевое распределение $\pi$ имеет «гребневую структуру», то число reject-ов будет велико.

# Gibbs Sampler

Geman and Geman (1984)

Пусть $X_n = (X_n^{[1]}, X_n^{[2]}, \dots, X_n^{[d]})$ — $d$-мерные случайные векторы и $\pi(x) = \pi(x^{[1]}, \dots, x^{[d]})$ — целевое распределение на пространстве $\mathscr{X} \subset \mathbb{R}^d$.

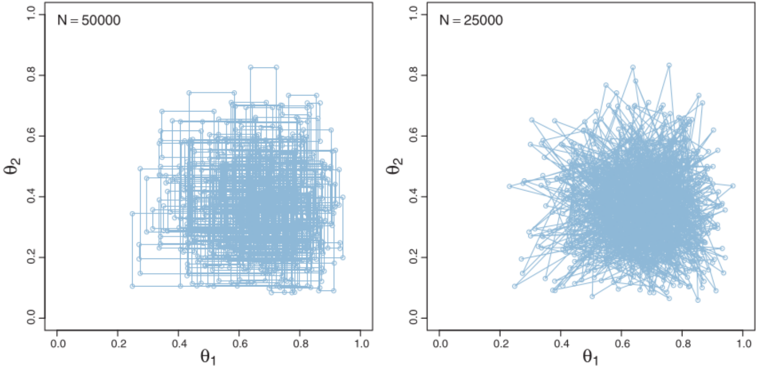
Очередной шаг однородной цепи Маркова (когда уже сгенерировано $X_n$), построенной методом Гиббса выглядит так:

$$
\begin{aligned}
X_{n+1}^{[1]} &\sim \pi(x^{[1]} \mid X_n^{[2]}, X_n^{[3]}, \dots, X_n^{[d]}) \\
X_{n+1}^{[2]} &\sim \pi(x^{[2]} \mid X_{n+1}^{[1]}, X_n^{[3]}, \dots, X_n^{[d]}) \\
&\dots \\
X_{n+1}^{[d]} &\sim \pi(x^{[d]} \mid X_{n+1}^{[1]}, X_{n+1}^{[2]}, \dots, X_{n+1}^{[d-1]})
\end{aligned}
$$

Очевидно, что распределение $\pi(x)$ будет стационарным для построенной цепи, поскольку для каждого подшага $j = 1, \dots, d$:

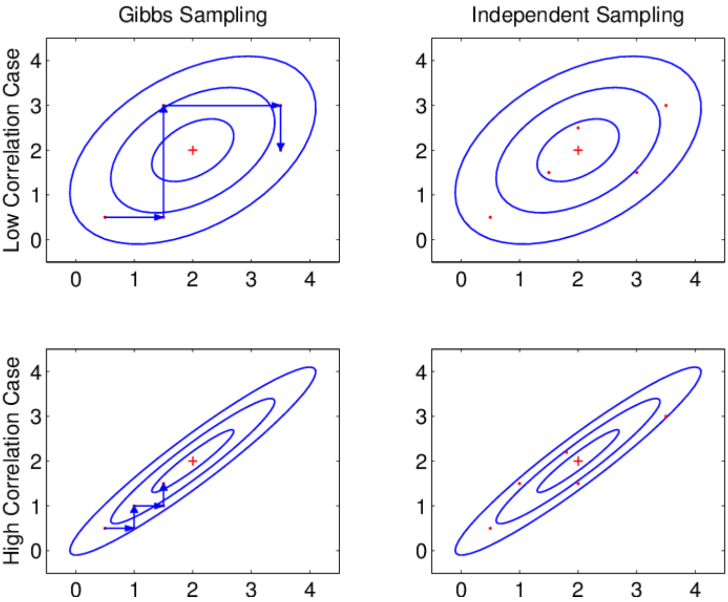
$$\int \pi(y^{[j]} \mid x^{[1]}, \dots, x^{[j-1]}, x^{[j+1]}, \dots, x^{[d]}) \pi(x)\, dx = \pi(y^{[j]}).$$

# Gibbs Sampler




Метод Гиббса удобен, если умеем качественно сэмплировать одномерные распределения $\pi(x^{[j]} \mid x^{[-j]}) \propto \pi(x^{[1]}, \ldots, x^{[j]}, \ldots, x^{[d]})$.

**ДЗ(*):** Показать, что метод Гиббса есть частный случай алгоритма Метрополиса-Гастингса для специального proposal distribution $q$, причём для него reject-ы не происходят никогда (т.е. $\alpha = 1$).



Gibbs Sampling
Independent Sampling
Low Correlation Case
High Correlation Case
0
1
2
3
4