

# Байесовский выбор моделей: Гамильтоновы методы Монте-Карло по схеме марковских цепей (НМС)

Александр Адуенко

3е декабря 2024

## Содержание предыдущих лекций

- Формула Байеса и формула полной вероятности;
- Определение априорных вероятностей и selection bias;
- (Множественное) тестирование гипотез
- Экспоненциальное семейства. Достаточные статистики.
- Наивный байес. Связь целевой функции и вероятностной модели.
- Линейная регрессия: связь МНК и  $w_{ML}$ , регуляризации и  $w_{MAP}$ .
- Свойство сопряженности априорного распределения правдоподобию.
- Прогноз для одиночной модели:

$$p(\mathbf{y}_{test} | \mathbf{X}_{test}, \mathbf{X}_{train}, \mathbf{y}_{train}) = \int p(\mathbf{y}_{test} | \mathbf{w}, \mathbf{X}_{test}) p(\mathbf{w} | \mathbf{X}_{train}, \mathbf{y}_{train}) d\mathbf{w}.$$

- Связь апостериорной вероятности модели и обоснованности
- Обоснованность: понимание и связь со статистической значимостью.
- Логистическая регрессия: проблемы ML-оценки  $w$  и связь априорного распределения с отбором признаков.
- EM-алгоритм и отбор признаков в байесовской линейной регрессии.
- Вариационный EM-алгоритм. Смесь моделей лог. регрессии.
- Гауссовские процессы. Учёт эволюции моделей во времени.
- Построение адекватных мультимodelей.
- Сэмплирование. Схема Гиббса.

Пусть имеется однородная марковская цепь с функцией плотности вероятности перехода между состояниями  $q(\mathbf{Z}_{i+1}|\mathbf{Z}_i)$ .

- Возьмем некоторое  $p_0(\mathbf{Z})$  и сгенерируем  $\mathbf{Z}_0 \sim p_0(\mathbf{Z})$ ;
- Генерируем  $\mathbf{Z}_{i+1} \sim q(\mathbf{Z}_{i+1}|\mathbf{Z}_i)$ ,  $i = 0, 1, \dots$ ;
- Выбрасываем первые  $m_0$  наблюдений (и прореживаем, если нужна НОР (i.i.d) выборка).

**Вопрос:** при каких условиях такая схема приведет к получению выборки из  $p(\mathbf{Z})$  ?

**Условие 1:**  $p(\mathbf{Z})$  инвариантно относительно цепи, то есть

$$p(\mathbf{Z}_{i+1}) = \int p(\mathbf{Z}_i)q(\mathbf{Z}_{i+1}|\mathbf{Z}_i)d\mathbf{Z}_i \text{ (стационарное распределение).}$$

**Достаточное условие:**  $p(\mathbf{Z}_{i+1})q(\mathbf{Z}_i|\mathbf{Z}_{i+1}) = p(\mathbf{Z}_i)q(\mathbf{Z}_{i+1}|\mathbf{Z}_i)$ .

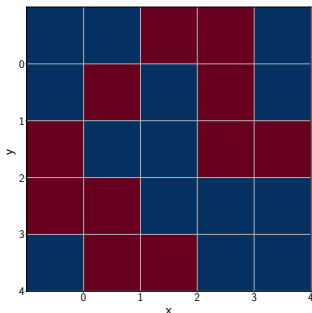
**Условие 2:** цепь эргодична, то есть стационарное распределение не зависит от начальных условий  $\forall p_0(\mathbf{Z}) p_i(\mathbf{Z}_i) \rightarrow p(\mathbf{Z})$  при  $i \rightarrow \infty$ .

**Достаточное условие:**  $\forall s \forall t : p(t) \neq 0 q(t|s) > 0$ .

# Модель Изинга (Ising model)

Пусть в каждой точке есть магнитный момент  $x_i \in \{\pm 1\}$  и  $p(\mathbf{x}) = \frac{1}{Z} \exp(-E(\mathbf{x})/T)$ , где  $E(\mathbf{x}) = - \sum_{(i,j) \in \varepsilon} x_i x_j - \sum_i h_i x_i$ .

Реализация магнитных моментов



Намагниченность:  $\mu = \left| \frac{1}{N} \sum_i x_i \right|$ ,

где  $N$  – число атомов в решетке.

Вопрос 1: Как оценить среднюю

намагниченность:  $E_p \mu$ ?

Вариационное приближение

$$p(\mathbf{x}) \approx q(\mathbf{x}) = \prod_{i=1}^N q_i(x_i).$$

$$\log q_i(x_i) \propto E_{q_{\setminus i}} \log p(\mathbf{x}) = -\frac{1}{T} E_{q_{\setminus i}} E(\mathbf{x}).$$

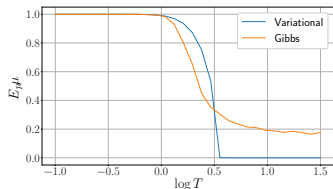
$$q_i(x_i = 1) = \frac{1}{1 + \exp\left(-\frac{2}{T}(h_i + \sum_{j: (i,j) \in \varepsilon} E_{q_j} x_j)\right)}.$$

Вопрос 2: Насколько хороша вариационная аппроксимация?

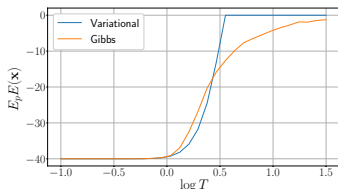
Вопрос 3: Какую альтернативу можно предложить?

# Сравнение вариационной аппроксимации и схемы Гиббса

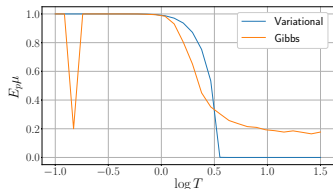
## Намагниченность с температурой



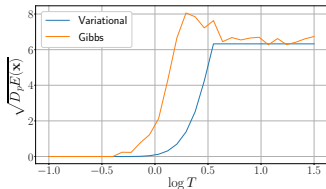
## Средняя энергия с температурой



## Намагниченность с температурой (проблема)



## Стандартная ошибка энергии с температурой

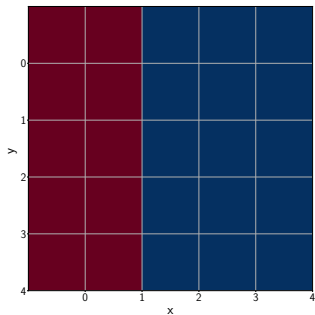


Вопрос: Что вызывает провал графика намагниченности?

## Свойства:

- (+) Подходит и для дискретных, и для непрерывных распределений;
- (+) Нет настраиваемых параметров;
- (-) Неэффективна в пространствах большой размерности;
- (-) Возможна очень долгая сходимость цепи к стационарному распределению.

## Реализация магнитных моментов



**Вопрос:** Что произойдет, если начальный элемент MCMC такой и  $T \ll 1$ ?

# Схема Метрополиса-Хастингса (Metropolis-Hastings)

$p(\mathbf{Z}) \propto \tilde{p}(\mathbf{Z})$ ,  $r(\mathbf{Z}|\mathbf{Z}_i)$  – предположеное распределение.

- Имеем  $\mathbf{Z}_i$ , сэмплируем  $\mathbf{Z}^* \sim r(\mathbf{Z}|\mathbf{Z}_i)$ ;
- Вычисляем  $P(\mathbf{Z}^*, \mathbf{Z}_i) = \min\left(1, \frac{\tilde{p}(\mathbf{Z}^*)r(\mathbf{Z}_i|\mathbf{Z}^*)}{\tilde{p}(\mathbf{Z}_i)r(\mathbf{Z}^*|\mathbf{Z}_i)}\right)$
- $\mathbf{Z}_{i+1} = \mathbf{Z}^*$  с вероятностью  $P(\mathbf{Z}^*, \mathbf{Z}_i)$ ,  
 $\mathbf{Z}_{i+1} = \mathbf{Z}_i$  с вероятностью  $1 - P(\mathbf{Z}^*, \mathbf{Z}_i)$ .

Отсюда  $q(\mathbf{Z}_{n+1}|\mathbf{Z}_n) = \begin{cases} r(\mathbf{Z}_{n+1}|\mathbf{Z}_n)P(\mathbf{Z}_{n+1}, \mathbf{Z}_n), & \mathbf{Z}_{n+1} \neq \mathbf{Z}_n, \\ 1 - \int r(\mathbf{Z}^*|\mathbf{Z}_n)P(\mathbf{Z}^*, \mathbf{Z}_n)d\mathbf{Z}^*, & \mathbf{Z}_{n+1} = \mathbf{Z}_n. \end{cases}$

**Достаточное условие эргодичности:**  $\forall s \forall t : \tilde{p}(t) > 0, q(t|s) > 0$ .

**Замечание 1:** для выполнения этого требования достаточно

$r(t|s) > 0 \forall s \forall t$ .

**Достаточное условие инвариантности:**  $\forall s \forall t \tilde{p}(s)q(t|s) = \tilde{p}(t)q(s|t)$ .

**Замечание 2:** Убеждаемся в выполнении условия подстановкой.

Для  $s = t$  очевидно. Пусть  $s \neq t$ , тогда  $\tilde{p}(s)q(t|s) = \tilde{p}(s)r(t|s) \min\left(1, \frac{\tilde{p}(t)r(s|t)}{\tilde{p}(s)r(t|s)}\right) = \min(\tilde{p}(s)r(t|s), \tilde{p}(t)r(s|t)) = \tilde{p}(t)q(s|t)$ .

# Схема Метрополиса-Хастингса (продолжение)

- Имеем  $\mathbf{Z}_i$ , сэмплируем  $\mathbf{Z}^* \sim r(\mathbf{Z}|\mathbf{Z}_i)$ ;
- Вычисляем  $P(\mathbf{Z}^*, \mathbf{Z}_i) = \min\left(1, \frac{\tilde{p}(\mathbf{Z}^*)r(\mathbf{Z}_i|\mathbf{Z}^*)}{\tilde{p}(\mathbf{Z}_i)r(\mathbf{Z}^*|\mathbf{Z}_i)}\right)$
- $\mathbf{Z}_{i+1} = \mathbf{Z}^*$ ,  $P(\mathbf{Z}^*, \mathbf{Z}_i)$ ,  
 $\mathbf{Z}_{i+1} = \mathbf{Z}_i$ ,  $1 - P(\mathbf{Z}^*, \mathbf{Z}_i)$ .

Если  $r(\mathbf{Z}^*|\mathbf{Z}) = r(\mathbf{Z}|\mathbf{Z}^*)$ , то  $P(\mathbf{Z}^*, \mathbf{Z}_i) = \min\left(1, \frac{\tilde{p}(\mathbf{Z}^*)}{\tilde{p}(\mathbf{Z}_i)}\right)$ .

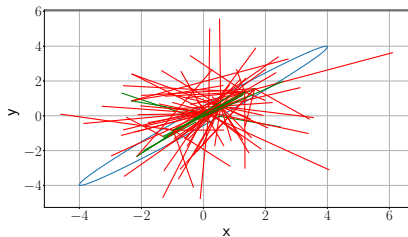
## Пример

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{0}, \sigma^2 \begin{pmatrix} 1 & 0.95 \\ 0.95 & 1 \end{pmatrix}\right),$$

$$\sigma = 2,$$

$$r(\mathbf{Z}^*|\mathbf{Z}) = \mathcal{N}(\mathbf{Z}^*|\mathbf{Z}, \sigma^2\mathbf{I}).$$

## Результат сэмплирования

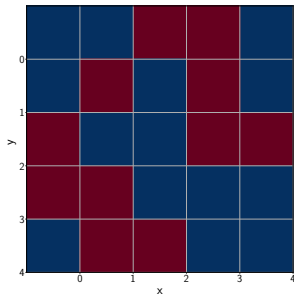




# Релаксация модели Изинга

Пусть в каждой точке есть магнитный момент  $x_i \in [-1, 1]$  и  $p(\mathbf{x}) = \frac{1}{Z} \exp(-E(\mathbf{x})/T)$ , где  $E(\mathbf{x}) = - \sum_{(i,j) \in \epsilon} x_i x_j - \sum_i h_i x_i$ .

Реализация магнитных моментов



Намагниченность:  $\mu = \left| \frac{1}{N} \sum_i x_i \right|$ ,

где  $N$  – число атомов в решетке.

**Вопрос 1:** как оценить среднюю намагниченность:  $E_p \mu$ ?

Уравнения Гамильтона

Пусть есть частица, которая движется в поле с потенциалом  $U(\mathbf{x})$ .

$\mathbf{x}$ ,  $\mathbf{p}$  – координаты и импульс частицы.

$K(\mathbf{p}) = \frac{\mathbf{p}^T \mathbf{p}}{2m}$  – кинетическая энергия.

$H(\mathbf{x}, \mathbf{p}) = U(\mathbf{x}) + K(\mathbf{p})$  – гамильтониан.

$$\frac{dx_i}{dt} = \frac{\partial H}{\partial p_i}, \quad \frac{dp_i}{dt} = -\frac{\partial H}{\partial x_i}.$$

**Вопрос 2:** зачем нам уравнения Гамильтона?

# Свойства гамильтоновой динамики (HD)

$\frac{dx_i}{dt} = \frac{\partial H}{\partial p_i}; \frac{dp_i}{dt} = -\frac{\partial H}{\partial x_i}$  – уравнения Гамильтона

$$\frac{dz}{dt} = \mathbf{J} \nabla_{\mathbf{z}} H(\mathbf{z}), \text{ где } \mathbf{z} = (\mathbf{x}, \mathbf{p}), \mathbf{J} = \begin{pmatrix} \mathbf{0}_n & \mathbf{I}_n \\ -\mathbf{I}_n & \mathbf{0}_n \end{pmatrix}.$$

Закон сохранения полной энергии:  $\frac{dH(\mathbf{x}, \mathbf{p})}{dt} = 0$ .

Обратимость:  $[\mathbf{x}(t+s), \mathbf{p}(t+s)] = \text{HD}(\mathbf{x}(t), \mathbf{p}(t), s)$ ;

$[\mathbf{x}(t), \mathbf{p}(t)] = \text{HD}(\mathbf{x}(t+s), \mathbf{p}(t+s), -s)$ ;

$[\mathbf{x}(t), -\mathbf{p}(t)] = \text{HD}(\mathbf{x}(t+s), -\mathbf{p}(t+s), s)$ .

Сохранение фазового объема:  $d\mathbf{p}d\mathbf{x} = \text{const}$ .

Симплектичность:  $\mathbf{B}^T \mathbf{J}^{-1} \mathbf{B} = \mathbf{J}^{-1}$ , где  $\mathbf{B}$  – якобиан HD по паре  $(\mathbf{x}, \mathbf{p})$ .

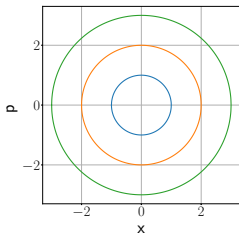
## Пример

$$U(x) = \frac{x^2}{2}, K(p) = \frac{p^2}{2}.$$

$$\frac{dx}{dt} = p, \frac{dp}{dt} = -x.$$

$$x(t) = r \sin(t + \varphi),$$

$$p(t) = r \cos(t + \varphi).$$



# Гамильтоновы методы Монте-Карло (НМС)

$p(\mathbf{x}, \mathbf{p}) = \frac{1}{Z} \exp(-U(\mathbf{x}) - \frac{1}{2}\mathbf{p}^T \mathbf{p})$ ,  $U(\mathbf{x}) = -\log \tilde{p}(\mathbf{x})$  (считаем  $m = 1$ ).

Идея: сэмплируем  $(\mathbf{x}_1, \mathbf{p}_1), \dots, (\mathbf{x}_m, \mathbf{p}_m)$  из  $p(\mathbf{x}, \mathbf{p})$  и выбрасываем  $\mathbf{p}_i$ .

## Схема сэмплирования НМС:

- 1 Выбираем  $\mathbf{x}^0$ ;
- 2 На шаге  $j \geq 1$  сэмплируем  $\mathbf{p}^j = \mathcal{N}(\mathbf{0}, \mathbf{I})$ ;
- 3 Решаем уравнения Гамильтона за интервал  $\Delta t = T\varepsilon$ , стартуя из  $(\mathbf{x}^{j-1}, \mathbf{p}^j)$  и получаем  $(\mathbf{x}^j, \mathbf{p}_{\text{new}}^j)$ , переворачиваем импульс:  $(\mathbf{x}^j, -\mathbf{p}_{\text{new}}^j)$ ;
- 4 Принимаем новую точку с вероятностью  $A = \min(1, \exp(-H(\mathbf{x}^j, -\mathbf{p}_{\text{new}}^j) + H(\mathbf{x}^{j-1}, \mathbf{p}^j)))$ ;
- 5 Переходим на шаг 2.

**Замечание:** НМС есть частный случай схемы Метрополиса-Хастингса.

**Вопрос 1:** Чем задается предположное распределение в НМС?

**Вопрос 2:** Зачем переворачивать знак импульса?

**Вопрос 3:** Что было использовано при получении формулы шага 4?

**Вопрос 4:** Какие значения может принимать  $-H(\mathbf{x}^1, \mathbf{p}_{\text{new}}^1) + H(\mathbf{x}^0, \mathbf{p}^1)$ ?

**Вопрос 5:** Гарантируется ли эргодичность для такой цепи?

# Решение уравнений Гамильтона

$$\frac{dx_i}{dt} = \frac{\partial H}{\partial p_i}, \quad \frac{dp_i}{dt} = -\frac{\partial H}{\partial x_i}$$

– уравнения Гамильтона.

## Настройка НМС

- $\mathbf{x}_0 : U(\mathbf{x}_0) \approx \max_{\mathbf{x}} U(\mathbf{x})$ ;
- Запуск нескольких цепей;
- Выбор шага  $\varepsilon : A(1 - A) \gg 0$ ;
- Оценка эффективной размерности выборки.

**Вопрос 1:** Зачем

$$U(\mathbf{x}_0) \approx \max_{\mathbf{x}} U(\mathbf{x})?$$

**Вопрос 2:** Как определить, что цепь сошлась?

**Вопрос 3:** Что дает условие  $A(1 - A) \gg 0$  на вероятность принятия объекта?

**Вопрос 4:** Как оценить эффективную размерность?

## Метод Эйлера

$$\begin{cases} p_i(t + \varepsilon) = p_i(t) - \varepsilon \frac{\partial U(\mathbf{x}(t))}{\partial x_i}, \\ x_i(t + \varepsilon) = x_i(t) + \varepsilon p_i(t). \end{cases}$$

**Проблема:** Невязка через время  $T$  есть  $O(T\varepsilon)$  (метод первого порядка).

## Метод leapfrog

$$\begin{cases} p_i(t + \frac{\varepsilon}{2}) = p_i(t) - \frac{\varepsilon}{2} \frac{\partial U(\mathbf{x}(t))}{\partial x_i}, \\ x_i(t + \varepsilon) = x_i(t) + \varepsilon p_i(t + \frac{\varepsilon}{2}), \\ p_i(t + \varepsilon) = p_i(t + \frac{\varepsilon}{2}) - \frac{\varepsilon}{2} \frac{\partial U(\mathbf{x}(t + \varepsilon))}{\partial x_i}. \end{cases}$$

**Замечание 1:** Невязка через время  $T$  есть  $O(T\varepsilon^2)$  (метод II порядка).

**Замечание 2:** Метод обратим: если сначала идти вперед по времени  $T$ , а затем назад, то через  $T$  вернемся в исходную точку.

**Замечание 3:** Сохраняет модифицированную полную энергию.

# Сравнение HD со схемой Метрополиса-Хастингса с диагональным предложным распределением

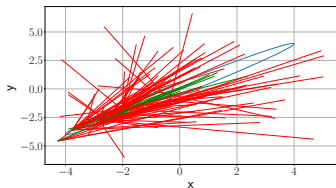
$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{0}, \sigma^2 \begin{pmatrix} 1 & 0.99 \\ 0.99 & 1 \end{pmatrix}\right), \sigma = 2.$$

Метрополис-Хастингс

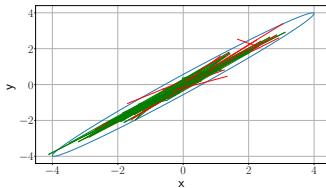
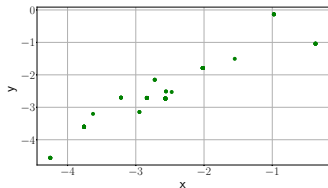
$$r(\mathbf{Z}^*|\mathbf{Z}) = \mathcal{N}(\mathbf{Z}^*|\mathbf{Z}, \sigma^2\mathbf{I})$$

Гамильтонов МС

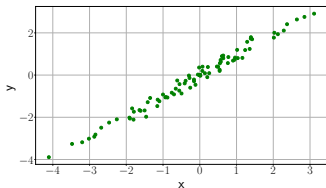
Leapfrog,  $\varepsilon = 0.25$ ,  $T = 6.25$ .



Сгенерированные сэмплы



Сгенерированные сэмплы



# Методы Монте-Карло по Ланжевену (Langevin Monte-Carlo)

Эквивалентная запись метода leapfrog:

$$\begin{cases} x_i(t + \varepsilon) = x_i(t) - \frac{\varepsilon^2}{2} \frac{\partial U(\mathbf{x}(t))}{\partial x_i} + \varepsilon p_i(t), \\ p_i(t + \varepsilon) = p_i(t) - \frac{\varepsilon}{2} \left( \frac{\partial U(\mathbf{x}(t))}{\partial x_i} + \frac{\partial U(\mathbf{x}(t+\varepsilon))}{\partial x_i} \right). \end{cases}$$

Схема сэмплирования LMC:

- 1 Выбираем  $\mathbf{x}_0$ ;
- 2 На шаге  $j \geq 1$  сэмплируем  $\mathbf{p}^j = \mathcal{N}(\mathbf{0}, \mathbf{I})$ ;
- 3 Решаем уравнения Гамильтона за интервал малый интервал  $\varepsilon$ :  
$$\begin{cases} \mathbf{x}^j = \mathbf{x}^{j-1} - \frac{\varepsilon^2}{2} \nabla_{\mathbf{x}} U(\mathbf{x}^{j-1}) + \varepsilon \mathbf{p}^j, \\ \mathbf{p}_{\text{new}}^j = \mathbf{p}^j - \frac{\varepsilon}{2} (\nabla_{\mathbf{x}} U(\mathbf{x}^{j-1}) + \nabla_{\mathbf{x}} U(\mathbf{x}^j)) \end{cases}$$
и переворачиваем импульс  $(\mathbf{x}^j, -\mathbf{p}_{\text{new}}^j)$ ;
- 4 Принимаем новую точку с вероятностью 1 в силу малости шага  $\varepsilon$ ;
- 5 Переходим на шаг 2.

**Замечание 1:** Обычно LMC работает хуже, чем общий НМС.

**Вопрос 1:** В чём преимущество НМС по сравнению со схемой Гиббса?

**Hint:** Подумайте, может ли НМС генерировать из разных мод

мультимодального распределения?

- 1 Bishop, Christopher M. "Pattern recognition and machine learning". Springer, New York (2006). Pp. 523-556.
- 2 Steve Brooks et al. «Handbook of Markov Chain Monte Carlo», Chapter 5. URL: <http://www.mcmchandbook.net/HandbookChapter5.pdf>
- 3 MacKay, David JC. Bayesian methods for adaptive models. Diss. California Institute of Technology, 1992.
- 4 MacKay, David JC. "The evidence framework applied to classification networks." Neural computation 4.5 (1992): 720-736.
- 5 Gelman, Andrew, et al. Bayesian data analysis, 3rd edition. Chapman and Hall/CRC, 2013.
- 6 Дрейпер, Норман Р. Прикладной регрессионный анализ. Рипол Классик, 2007.