

Markov chain Monte Carlo

Селиханович Даниил

МФТИ

ИППИ им. Харкевича РАН

selihanovich.do@phystech.edu

6 мая 2019 г.

План доклада

- 1 Мотивация
- 2 Геометрия многомерного пространства
- 3 Марковские цепи, напоминание
- 4 Метод Монте-Карло
- 5 Метод MCMC
- 6 Алгоритм Метрополиса-Гастингса
- 7 Частные случаи алгоритма МН
- 8 Свойства МН

Постановка задачи

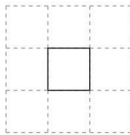
- В статистике больших размерностей возникает необходимость численного подсчёта многомерных интегралов, например, выражающих матожидание/дисперсию.
- Методы вычислительной математики могут быть неустойчивы.
- Вопрос: как бороться с этой проблемой?

Явление концентрации меры

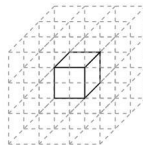
Факт: в многомерных пространствах мера шара концентрируется на границе.



(a)



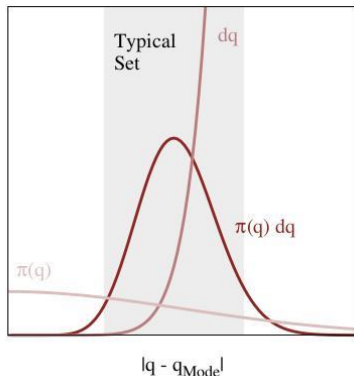
(b)



(c)

Typical set

Наибольший вклад в матожидание $\mathbb{E}_{\pi}[f] = \int_{\mathcal{Q}} f(q)\pi(q)dq$ вносят так называемые *typical sets*, на которых произведение плотности вероятности и меры наибольшее.



Definition (Эргодическая марковская цепь)

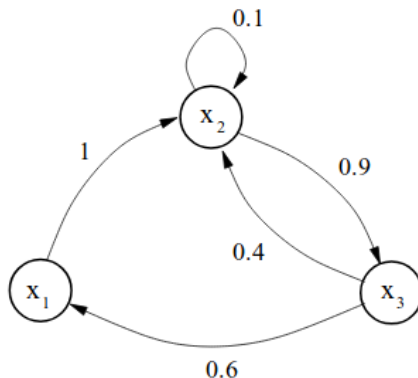
Дискретная марковская цепь называется эргодичной, если существует независимое от начального состояния ненулевое предельное распределение вероятностей её состояний, причём все предельные вероятности положительны.

Theorem (Эргодическая теорема)

Конечная однородная дискретная марковская цепь является эргодичной тогда и только тогда, когда она неразложима и непериодическая.

Наблюдение

- Если есть какое-то вероятностное распределение, то можно попробовать подобрать для него эргодичную марковскую цепь, для которой оно будет стационарным распределением.
- Вопрос: как?



Метод Монте-Карло

- Пусть сгенерированы i.i.d. данные $\{x^{(i)}\}_{i=1}^N$ из распределения $p(x)$, определённого на пространстве χ высокой размерности. Нам дана функция f , матожидание которой мы хотим оценить.

- УЗБЧ даёт сходимость почти наверное среднего к своему матожиданию

$$I_N(f) = \frac{1}{N} \sum_{i=1}^N f(x^{(i)}) \xrightarrow[N \rightarrow \infty]{a.s.} I(f) = \int_{\chi} f(x) p(x) dx.$$

- Для случая одной переменной и конечности дисперсии f , равной $\sigma_f^2 \triangleq \mathbb{E}_{p(x)}(f^2(x)) - I^2(f) < \infty$, ЦПТ даёт сходимость к нормальному распределению:

$$\sqrt{N}(I_N(f) - I(f)) \xrightarrow[N \rightarrow \infty]{d} \mathcal{N}(0, \sigma_f^2)$$

- Введём дискретную марковскую цепь, состояниями которой будут сгенерированные данные: $x^{(i)} \in \mathcal{X} = \{x_1, x_2, \dots, x_N\}$.
- Введём матрицу переходов T для такой марковской цепи. В предыдущем примере $T = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0.1 & 0.9 \\ 0.6 & 0.4 & 0 \end{bmatrix}$
- Условие стационарного распределения:
$$p(x^{(i)}) = \sum_{x^{(i-1)}} p(x^{(i-1)}) T(x^{(i)} | x^{(i-1)}).$$
- Скорость сходимости к стационарному распределению определяется $\lambda_2(T)$, которое должно быть как можно меньше (теорема Фробениуса-Перрона).

Непрерывный случай

- В непрерывных пространствах состояний матрица переходов T становится ядром интегрального оператора $K(x^{(i)}|x^{(i-1)})$.
- Условие стационарного распределения переписывается так:
$$\int p(x^{(i)}) K(x^{(i+1)}|x^{(i)}) dx^{(i)} = p(x^{(i+1)}).$$
- Стационарное распределение становится собственной функцией ядра интегрального оператора.

Алгоритм Метрополиса-Гастингса

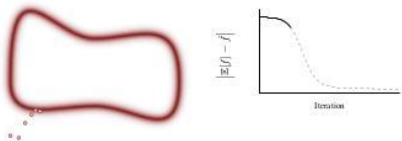
Algorithm 1: Псевдокод

Input: Вспомогательная функция распределения $q(x^*|x)$, для которой генерировать выборку просто (например, нормальное распределение), функция распределения $p(x)$

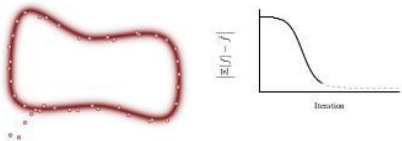
Output: последовательность, сгенерированная из распределения $p(x)$

- 1 Инициализировать начальное состояние цепи $x^{(0)}$
 - 2 **for** $k \leftarrow 0$ **to** $N - 1$ **do**
 - 3 Сгенерировать из равномерного распределения $u \sim \mathcal{U}_{[0,1]}$
 - 4 Сгенерировать $x^* \sim q(x^*|x^{(i)})$
 - 5 **if** $u < \mathcal{A}(x^{(i)}, x^*) = \min \left\{ 1, \frac{p(x^*)q(x^{(i)}|x^*)}{p(x^{(i)})q(x^*|x^{(i)})} \right\}$ **then**
 - 6 $x^{(i+1)} = x^*$
 - 7 **else**
 - 8 $x^{(i+1)} = x^{(i)}$
 - 9 **return** $x^{(0)}, x^{(1)}, \dots, x^{(N-1)}$
-

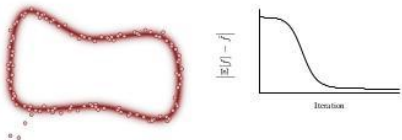
Поведение марковской цепи



(a)



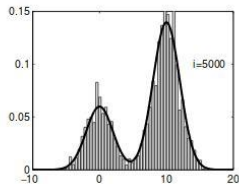
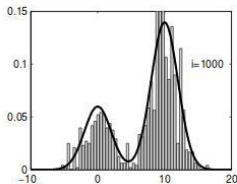
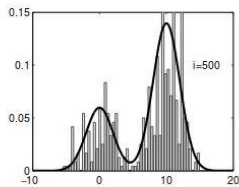
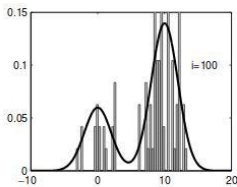
(b)



(c)

Пример

Гистограммы, полученные алгоритмом Метрополиса-Гастингса, на разных итерациях, $N = 5000$. Здесь $q(x^*|x^{(i)}) = \mathcal{N}(x^{(i)}, 100)$, а целевое распределение $p(x) \propto 0.3 \exp(-0.2x^2) + 0.7 \exp(-0.2(x-10)^2)$.



Почему построенная цепь нужная?

- По построению, интегральное ядро для алгоритма Метрополиса-Гастингса имеет вид: $K_{\text{МН}}(x^{(i+1)}|x^{(i)}) = q(x^{(i+1)}|x^{(i)}) \mathcal{A}(x^{(i)}, x^{(i+1)}) + \delta_{x^{(i)}}(x^{(i+1)}) r(x^{(i)})$, где
- $r(x^{(i)}) = \int_{\mathcal{X}} q(x^*|x^{(i)}) (1 - \mathcal{A}(x^{(i)}, x^*)) dx^*$ - член, связанный с возможностью отклонения.
- Отсюда можно получить соотношение:
 $p(x^{(i)}) K_{\text{МН}}(x^{(i+1)}|x^{(i)}) = p(x^{(i+1)}) K_{\text{МН}}(x^{(i)}|x^{(i+1)})$ - условие детального баланса.
- Если проинтегрировать обе части равенства по $dx^{(i)}$:
 $\int p(x^{(i)}) K(x^{(i+1)}|x^{(i)}) dx^{(i)} = p(x^{(i+1)})$ - значит, $p(x)$ является стационарным распределением построенной марковской цепи.

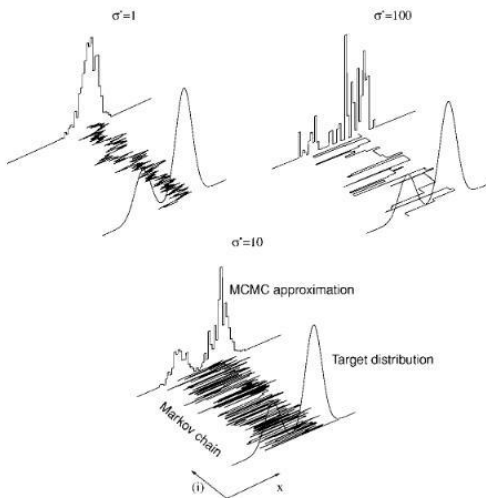
Частные случаи алгоритма МН

- Алгоритм *independent sampler* предполагает независимость вспомогательного распределения $q(x^*|x)$ от текущего состояния цепи: $q(x^*|x^{(i)}) = q(x^*)$
- Но генерируемые данные становятся зависимыми = (
- Алгоритм Метрополиса предполагает симметричность вспомогательного распределения: $q(x^*|x^{(i)}) = q(x^{(i)}|x^*)$
- Это позволяет упростить выражение для $\mathcal{A}(x^{(i)}, x^*)$:

$$\mathcal{A}(x^{(i)}, x^*) = \min \left\{ 1, \frac{p(x^*)}{p(x^{(i)})} \right\}$$

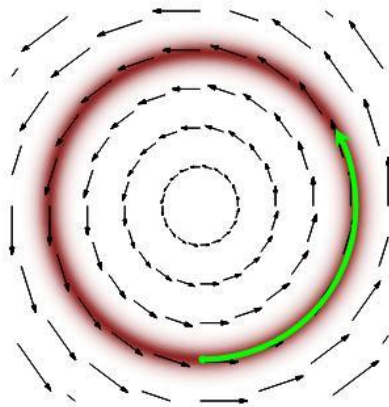
- Можно параллелить
- Достаточно значит целевое распределению с точностью до константы. Это видно из выражения для $\mathcal{A}(x^{(i)}, x^*)$.
- Разный выбор дисперсии вспомогательного распределения σ^* сильно влияет на результаты алгоритма.
- Узкая дисперсия - заметна только одна мода распределения $p(x)$
- Большая дисперсия - цепь чаще остаётся в том же состоянии, данные на выходе скоррелированы.

Иллюстрация зависимости результатов от дисперсии



Обобщения: Hamilton Markov Chain

Идея: ввести векторное поле на *typical sets*, которое будет указывать динамику для переходов марковской цепи.



Методы дифференциальной геометрии

- Можно рассматривать марковскую цепь как физическую систему в фазовом вероятностном пространстве переменных q_n .
- Эту систему можно дополнить до гамильтоновой, введя переменные «импульса» $q_n \rightarrow (q_n, p_n)$
- Затем ввести вероятностное распределение на расширенном пространстве: $\pi(q, p) = \pi(p|q)\pi(q)$ - каноническое распределение
- Ввести гамильтониан с кинетической и потенциальной энергией, который будет характеризовать динамику системы

Динамика марковской системы




- Гамильтониан системы: $H(q, p) \equiv -\log \pi(q, p) = -\log \pi(p|q) - \log \pi(q) \equiv K(p, q) + V(q)$
- Член $K(p, q) = -\log \pi(p|q)$ - кинетическая энергия, $V(q) = -\log \pi(q)$ - потенциальная энергия.
- Уравнения Гамильтона:

$$\begin{aligned}\frac{dq}{dt} &= +\frac{\partial H}{\partial p} = \frac{\partial K}{\partial p} \\ \frac{dp}{dt} &= -\frac{\partial H}{\partial q} = -\frac{\partial K}{\partial q} - \frac{\partial V}{\partial q}\end{aligned}$$

- Проблема: как выбрать кинетическую энергию?

Примеры выбора кинетической энергии

- Для евклидовой метрики - Euclidean-Gaussian Kinetic Energies: $K(q, p) = \frac{1}{2} p^T \cdot M^{-1} \cdot p + \log |M| + \text{const}$. Здесь $M = R \cdot S \cdot g \cdot S^T \cdot R^T$ - матрица, обратная матрице Грама в пространстве импульсов p , g - матрица Грама в пространстве координат q , S - диагональная масштабирующая матрица, R - ортогональная матрица.
- Для римановской метрики - Riemannian-Gaussian Kinetic Energies: $K(q, p) = \frac{1}{2} p^T \cdot \Sigma^{-1}(q) \cdot p + \frac{1}{2} \log |\Sigma(q)| + \text{const}$. Здесь матрица $\Sigma(q)$ имеет смысл: $\pi(p|q) = \mathcal{N}(p|0, \Sigma(q))$.
- Какую метрику выбирать в конкретной задаче - хороший вопрос.

-  *Andrieu, C., De Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. Machine learning, 50(1-2), 5-43.*
-  *Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. arXiv preprint arXiv:1701.02434.*
-  *David S. Markov chain Monte Carlo, Lecture 9.*

Спасибо за внимание! Вопросы?