Sampling methods

Urakov Mikhail

25 октября 2025 г.

Содержание

Постановка задачи сэмплирования

Основная задача

Генерация случайных выборок из заданного распределения вероятностей $p(x)$

- Дано: распределение $p(x)$ (возможно ненормированное)
- Найти: алгоритм генерации $\{x_i\}_{i=1}^{N} \sim p(x)$
- Применение: Монте-Карло методы, байесовский вывод, машинное обучение

$$\mathbb{E}_{p(x)}[f(x)] \approx \frac{1}{N} \sum_{i=1}^{N} f(x_i)$$

Мотивация

Почему сэмплирование важно?

- Численное интегрирование
- Байесовская статистика
- Обучение генеративных моделей
- Оптимизация
- Физическое моделирование
- **Normalizing Flows**

Типичные применения

- MCMC методы
- Вариационные автоэнкодеры
- **Bayesian Networks**
- Reinforcement Learning
- Computational physics

Преобразование Смирнова

Theorem (Преобразование Смирнова)

Если $U \sim Uniform(0,1)$ *и* F *- функция распределения, то* $X = F^{-1}(U)$ *имеет распределение с функцией распределения* F*.*

Алгоритм

1. Сгенерировать $u \sim U(0,1)$
2. Вычислить $x = F^{-1}(u)$
3. Вернуть x

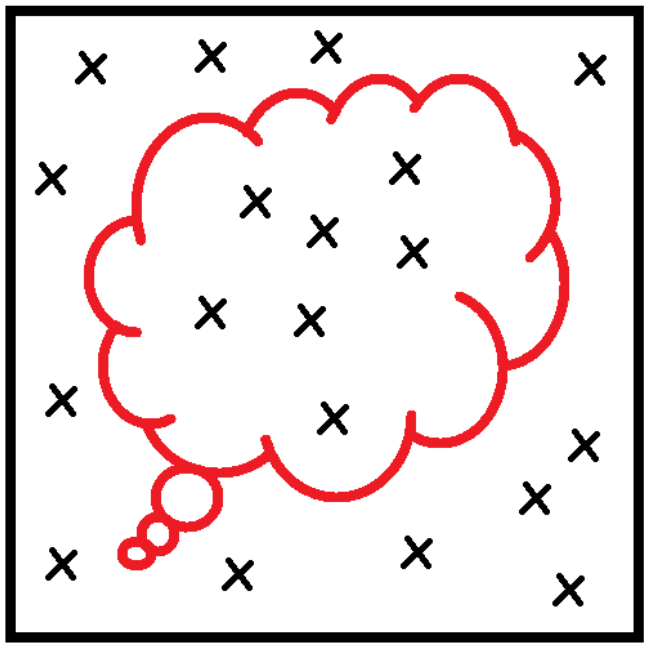
Example

Экспоненциальное распределение: $F(x) = 1 - e^{-\lambda x} \Rightarrow$x = -ln$(1-u)_{\overline{\lambda}}$

Сэмплирование из равномерного распределения на множестве

Задача

Сгенерировать точку равномерно из множества $A \subset \mathbb{R}^d$



Сэмплирование из подграфика плотности

Идея метода

Рассмотрим подграфик плотности:

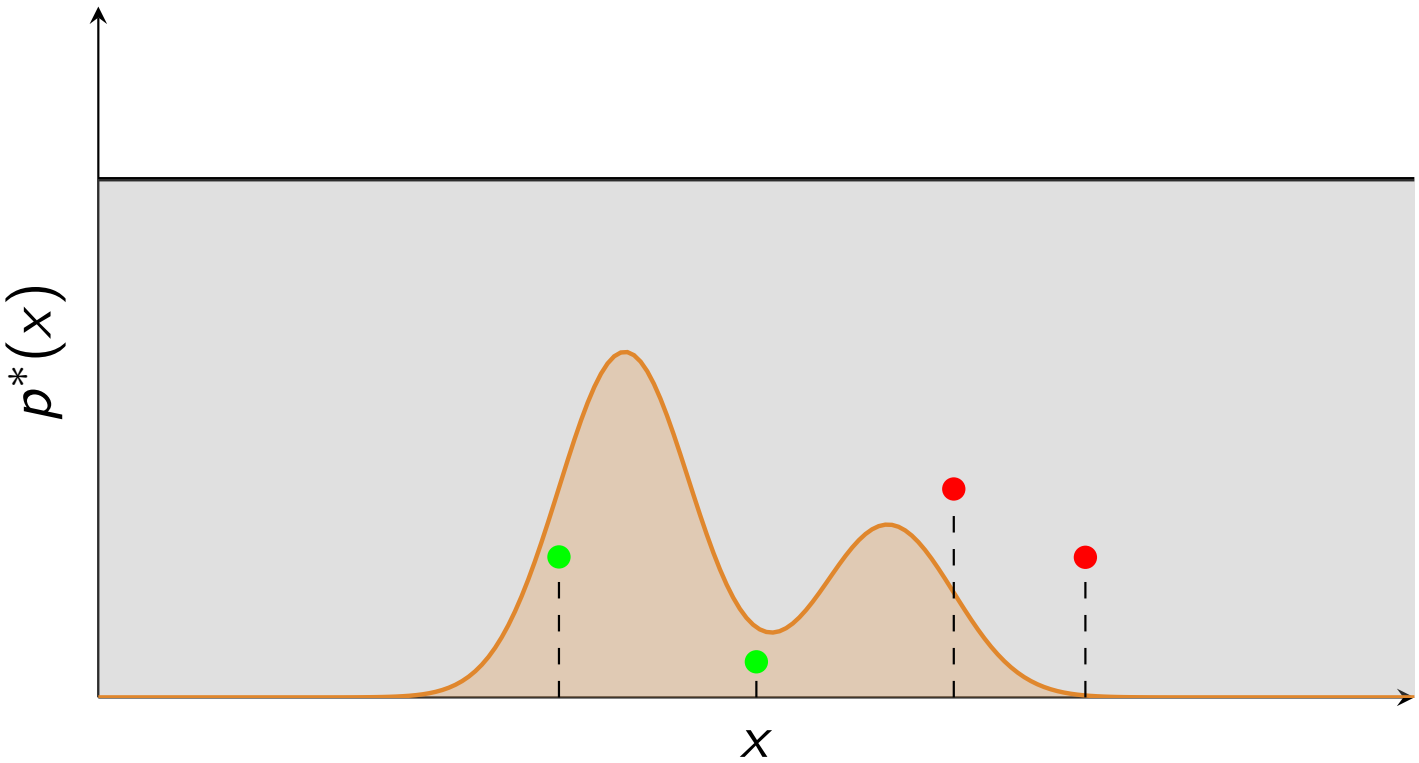
$$G = \{(x, y) : 0 \leq y \leq p^*(x)\}$$

Если (X, Y) равномерно распределена в G, то $X \sim p(x) \propto p^*(x)$

Алгоритм:

1. Выбрать область $[a, b] \times [0, M]$
2. Генерировать $(x, y) \sim U(G)$
3. Если $y \leq p^*(x)$, принять x
4. Иначе повторить

Сэмплирование из подграфика плотности



Rejection Sampling

Основная идея

Использование proposal distribution $q(x)$ для генерации кандидатов

Алгоритм:

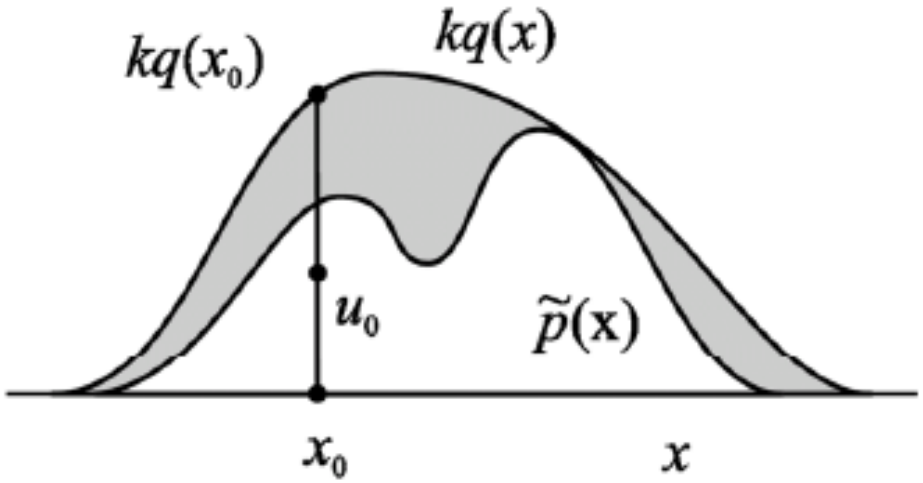
1. Выбрать $q(x)$ и константу M такие, что $p(x) \leq Mq(x)$
2. Сгенерировать $x \sim q(x)$
3. Сгенерировать $u \sim U(0, Mq(x))$
4. Если $u \leq p(x)$, принять x
5. Иначе вернуться к шагу 2

Эффективность

Вероятность принятия: $\frac{1}{M}$

$$\text{Стараться минимизировать } M = \sup_x \frac{p(x)}{q(x)}$$

Rejection Sampling



Rejection Sampling

Совместное распределение (X, U)

Совместная плотность предложенной пары (x, u):

$$f_{X,U}(x, u) = q(x) \cdot \mathbb{I}_{[0,1]}(u)$$

Условное распределение принятых образцов

Обозначим событие принятия:

$$A = \left\{ u \leq \frac{p^*(x)}{M \cdot q(x)} \right\}$$

Совместная плотность принятых пар (x, u):

$$f_{X,U|A}(x, u) = \frac{f_{X,U}(x, u) \cdot \mathbb{I}_A}{P(A)}$$

где вероятность принятия:

$$P(A) = \iint f_{X,U}(x, u) \cdot \mathbb{I}_A dxdu = \int q(x) \int_0^{\frac{p^*(x)}{Mq(x)}} dudx = \frac{1}{M} \int p^*(x)dx$$

Маргинальное распределение принятых x

Маргинализуем по u для принятых образцов:

$$f_{X|A}(x) = \int_0^1 f_{X,U|A}(x,u)du = \frac{\int_0^{\frac{p^*(x)}{Mq(x)}} q(x)du}{P(A)}$$

$$= \frac{q(x) \cdot \frac{p^*(x)}{Mq(x)}}{P(A)} = \frac{\frac{p^*(x)}{M}}{P(A)}$$

Подставляя вероятность принятия

$$P(A) = \frac{1}{M}\int p^*(x)dx = \frac{c}{M}\int p(x)dx = \frac{c}{M}$$

Следовательно:

$$f_{X|A}(x) = \frac{\frac{p^*(x)}{M}}{\frac{c}{M}} = \frac{p^*(x)}{c} = p(x)$$

Importance Sampling

Основная идея

Оценка матожидания через взвешенные samples из proposal distribution

$$\mathbb{E}_{p(x)}[f(x)] = \int f(x)p(x)dx = \int \left(f(x)\frac{p(x)}{q(x)}\right)q(x)dx = \mathbb{E}_{q(x)}\left[f(x)\frac{p(x)}{q(x)}\right]$$

Алгоритм:

1. Выбрать proposal distribution $q(x)$
2. Сгенерировать $\{x_i\}_{i=1}^{N} \sim q(x)$
3. Вычислить веса $w_i = \frac{p(x_i)}{q(x_i)}$
4. Оценка: $\mathbb{E}_{p(x)}[f(x)] \approx \frac{1}{N}\sum_{i=1}^{N} w_i f(x_i)$

Importance Sampling

А если у нас нет распределений $q(x), p(x)$, но есть $q^*(x) = Mq(x), p^*(x) = Cp(x)$? Тогда

$$\mathbb{E}_{p(x)}[f(x)] = \mathbb{E}_{q(x)}\left[f(x)\frac{p(x)}{q(x)}\right] = \mathbb{E}_{q(x)}\left[f(x)\frac{p^*(x)}{q^*(x)}\frac{C}{M}\right]$$

В свою очередь

$$\frac{C}{M} = \int \frac{p^*(x)}{M}dx = \int \frac{p^*(x)}{q^*(x)}q(x)dx = \mathbb{E}_{q(x)}\left[\frac{p^*(x)}{q^*(x)}\right] \approx \frac{1}{N}\sum_{i=1}^{N}\frac{p^*(x_i)}{q^*(x_i)}$$

Сравнение методов

Метод	Простота	Эффективность	Применимость
Смирнова	Высокая	Высокая	Ограниченная
Rejection	Средняя	Переменная	Широкая
Importance	Средняя	Переменная	Очень широкая

Таблица: Сравнение методов сэмплирования

Ключевые моменты

- Выбор метода зависит от задачи и распределения
- Rejection sampling требует знания верхней границы
- Importance sampling может иметь большую дисперсию
- Для сложных распределений используются MCMC методы

Заключение

Основные выводы

- Сэмплирование - фундаментальная задача в статистике и ML
- Простые методы работают для простых распределений
- Для сложных случаев нужны продвинутые методы (MCMC, VI)
- Правильный выбор proposal distribution критически важен

Дальнейшее изучение

- Markov Chain Monte Carlo (MCMC)
- Normalizing Flows