Байесовские Методы. Лекция 4

БИВТ-23-9/10-ИСАД

29 сентября 2025 г.

Мотивация: интегралы в байесовских моделях

Предсказательное распределение

$$P(y^* \mid x^*, D) = \int P(y^* \mid x^*, \theta)\, P(\theta \mid D)\, d\theta.$$

- В простых случаях (сопряжённые априоры) интеграл берётся аналитически.

Мотивация: интегралы в байесовских моделях

Предсказательное распределение

$$P(y^* \mid x^*, D) = \int P(y^* \mid x^*, \theta)\, P(\theta \mid D)\, d\theta.$$

- В простых случаях (сопряжённые априоры) интеграл берётся аналитически.
- Но в реальных задачах:

Мотивация: интегралы в байесовских моделях

Предсказательное распределение

$$P(y^* \mid x^*, D) = \int P(y^* \mid x^*, \theta)\, P(\theta \mid D)\, d\theta.$$

- В простых случаях (сопряжённые априоры) интеграл берётся аналитически.
- Но в реальных задачах:
  - высокая размерность параметров θ,

Мотивация: интегралы в байесовских моделях

Предсказательное распределение

$$P(y^* \mid x^*, D) = \int P(y^* \mid x^*, \theta)\, P(\theta \mid D)\, d\theta.$$

- В простых случаях (сопряжённые априоры) интеграл берётся аналитически.
- Но в реальных задачах:
  - высокая размерность параметров θ,
  - сложные модели без закрытой формы.

Идея Монте-Карло

Заменим интеграл на среднее по случайным выборкам.

Мотивация: интегралы в байесовских моделях

Предсказательное распределение

$$P(y^* \mid x^*, D) = \int P(y^* \mid x^*, \theta)\, P(\theta \mid D)\, d\theta.$$

- В простых случаях (сопряжённые априоры) интеграл берётся аналитически.
- Но в реальных задачах:
  - высокая размерность параметров θ,
  - сложные модели без закрытой формы.
- Пример: линейная регрессия с тысячами признаков или нейросеть.

Идея Монте-Карло

Заменим интеграл на среднее по случайным выборкам.

Мотивация: интегралы в байесовских моделях

Предсказательное распределение

$$P(y^* \mid x^*, D) = \int P(y^* \mid x^*, \theta)\, P(\theta \mid D)\, d\theta.$$

- В простых случаях (сопряжённые априоры) интеграл берётся аналитически.
- Но в реальных задачах:
  - высокая размерность параметров θ,
  - сложные модели без закрытой формы.
- Пример: линейная регрессия с тысячами признаков или нейросеть.
- Вычислить интеграл «вручную» невозможно.

Идея Монте-Карло

Заменим интеграл на среднее по случайным выборкам.

Главные вызовы, которые решает Монте-Карло

- **Высокая размерность:** интегралы по $\theta \in \mathbb{R}^d$ при большом d. Классические квадратурные методы (Ньютон-Котс, метод Гаусса) взрываются по стоимости, так как количество операций растет экспоненциально размерности. Сложность МС ($\sim 1/\sqrt{N}$) не зависит от размерности.

Главные вызовы, которые решает Монте-Карло

- **Высокая размерность:** интегралы по $\theta \in \mathbb{R}^d$ при большом d. Классические квадратурные методы (Ньютон-Котс, метод Гаусса) взрываются по стоимости, так как количество операций растет экспоненциально размерности. Сложность MC ($\sim 1/\sqrt{N}$) не зависит от размерности.
- **Неизвестные нормировочные константы.** В Байесе часто важен $P(D)$ или $P(\theta \mid D) = \frac{P(D|\theta)P(\theta)}{Z}$, где Z — интеграл, который нельзя посчитать явно. MC позволяет оценить такие интегралы.

Главные вызовы, которые решает Монте-Карло

- **Высокая размерность:** интегралы по $\theta \in \mathbb{R}^d$ при большом d. Классические квадратурные методы (Ньютон-Котс, метод Гаусса) взрываются по стоимости, так как количество операций растет экспоненциально размерности. Сложность МС ($\sim 1/\sqrt{N}$) не зависит от размерности.
- **Неизвестные нормировочные константы.** В Байесе часто важен $P(D)$ или $P(\theta \mid D) = \frac{P(D|\theta)P(\theta)}{Z}$, где Z — интеграл, который нельзя посчитать явно. МС позволяет оценить такие интегралы.
- **Редкие события.** Вероятности вида $\mathbb{P}(\theta \in A)$, где A — область малой меры, трудно вычислить напрямую. Специальные варианты МС дают определенные оценки.

Главные вызовы, которые решает Монте-Карло

- **Высокая размерность:** интегралы по $\theta \in \mathbb{R}^d$ при большом d. Классические квадратурные методы (Ньютон-Котс, метод Гаусса) взрываются по стоимости, так как количество операций растет экспоненциально размерности. Сложность МС ($\sim 1/\sqrt{N}$) не зависит от размерности.
- **Неизвестные нормировочные константы.** В Байесе часто важен $P(D)$ или $P(\theta \mid D) = \frac{P(D|\theta)P(\theta)}{Z}$, где Z — интеграл, который нельзя посчитать явно. МС позволяет оценить такие интегралы.
- **Редкие события.** Вероятности вида $\mathbb{P}(\theta \in A)$, где A — область малой меры, трудно вычислить напрямую. Специальные варианты МС дают определенные оценки.
- **Сложные апостериоры.** В многомодальных или плохо обусловленных распределениях невозможно получить аналитическое решение. МС позволяет хотя бы *приблизить* эти распределения выборкой с вариативной точностью.

Главные вызовы, которые решает Монте-Карло

- **Высокая размерность:** интегралы по $\theta \in \mathbb{R}^d$ при большом d. Классические квадратурные методы (Ньютон-Котс, метод Гаусса) взрываются по стоимости, так как количество операций растет экспоненциально размерности. Сложность МС ($\sim 1/\sqrt{N}$) не зависит от размерности.
- **Неизвестные нормировочные константы.** В Байесе часто важен $P(D)$ или $P(\theta \mid D) = \frac{P(D|\theta)P(\theta)}{Z}$, где Z — интеграл, который нельзя посчитать явно. МС позволяет оценить такие интегралы.
- **Редкие события.** Вероятности вида $\mathbb{P}(\theta \in A)$, где A — область малой меры, трудно вычислить напрямую. Специальные варианты МС дают определенные оценки.
- **Сложные апостериоры.** В многомодальных или плохо обусловленных распределениях невозможно получить аналитическое решение. МС позволяет хотя бы *приблизить* эти распределения выборкой с вариативной точностью.

Главные вызовы, которые решает Монте-Карло

- **Высокая размерность:** интегралы по $\theta \in \mathbb{R}^d$ при большом d. Классические квадратурные методы (Ньютон-Котс, метод Гаусса) взрываются по стоимости, так как количество операций растет экспоненциально размерности. Сложность МС ($\sim 1/\sqrt{N}$) не зависит от размерности.
- **Неизвестные нормировочные константы.** В Байесе часто важен $P(D)$ или $P(\theta \mid D) = \frac{P(D|\theta)P(\theta)}{Z}$, где Z — интеграл, который нельзя посчитать явно. МС позволяет оценить такие интегралы.
- **Редкие события.** Вероятности вида $\mathbb{P}(\theta \in A)$, где A — область малой меры, трудно вычислить напрямую. Специальные варианты МС дают определенные оценки.
- **Сложные апостериоры.** В многомодальных или плохо обусловленных распределениях невозможно получить аналитическое решение. МС позволяет хотя бы *приблизить* эти распределения выборкой с вариативной точностью.

Итого

МС-методы превращают невозможные аналитические задачи в численные, с контролируемой точностью.

Основные идеи методов Монте-Карло

- **Детерминированные преобразования** (например, метод обратной функции распределения).

Основные идеи методов Монте-Карло

- **Детерминированные преобразования** (например, метод обратной функции распределения).
- **Вероятностные механизмы «принять/отклонить»** (rejection sampling).

Основные идеи методов Монте-Карло

- **Детерминированные преобразования** (например, метод обратной функции распределения).
- **Вероятностные механизмы «принять/отклонить»** (rejection sampling).
- **Взвешивание выборки** (importance sampling).

Основные идеи методов Монте-Карло

- **Детерминированные преобразования** (например, метод обратной функции распределения).
- **Вероятностные механизмы «принять/отклонить»** (rejection sampling).
- **Взвешивание выборки** (importance sampling).
- **Итеративные/цепные методы** (локальное исследование распределений, будут рассмотрены позже).

Основные идеи методов Монте-Карло

- **Детерминированные преобразования** (например, метод обратной функции распределения).
- **Вероятностные механизмы «принять/отклонить»** (rejection sampling).
- **Взвешивание выборки** (importance sampling).
- **Итеративные/цепные методы** (локальное исследование распределений, будут рассмотрены позже).

Идея

Все методы Монте-Карло строятся на комбинации этих базовых принципов.

Метод обратной функции распределения

Идея

Если $U \sim \text{Uniform}(0, 1)$ и F — функция распределения (CDF) случайной величины X, то случайная величина

$$X = F^{-1}(U)$$

имеет распределение F.

- Универсальный способ получить выборку из произвольного распределения, если можно посчитать F^{-1}.

Метод обратной функции распределения

Идея

Если $U \sim \text{Uniform}(0, 1)$ и F — функция распределения (CDF) случайной величины X, то случайная величина

$$X = F^{-1}(U)$$

имеет распределение F.

- Универсальный способ получить выборку из произвольного распределения, если можно посчитать F^{-1}.
- Очень удобен для простых распределений.

Метод обратной функции распределения

Идея

Если $U \sim \text{Uniform}(0, 1)$ и F — функция распределения (CDF) случайной величины X, то случайная величина

$$X = F^{-1}(U)$$

имеет распределение F.

- Универсальный способ получить выборку из произвольного распределения, если можно посчитать F^{-1}.
- Очень удобен для простых распределений.

Пример: экспоненциальное распределение

CDF: $F(x) = 1 - e^{-\lambda x}$, $x \geq 0$.

$$X = F^{-1}(U) = -\frac{1}{\lambda} \log(1 - U).$$

Rejection sampling: интуиция

Задача

Хотим сэмплировать из сложной плотности $f(x)$, но напрямую это сделать трудно.

- Берём простое распределение $g(x)$, из которого легко сэмплировать (называется *proposal*).

Rejection sampling: интуиция

Задача

Хотим сэмплировать из сложной плотности $f(x)$, но напрямую это сделать трудно.

- Берём простое распределение $g(x)$, из которого легко сэмплировать (называется *proposal*).
- Масштабируем его вверх: подбираем M так, чтобы для всех x

$$f(x) \leq Mg(x).$$

То есть $Mg(x)$ — «крышка», которая накрывает целевую плотность $f(x)$.

Rejection sampling: интуиция

Задача

Хотим сэмплировать из сложной плотности $f(x)$, но напрямую это сделать трудно.

- Берём простое распределение $g(x)$, из которого легко сэмплировать (называется *proposal*).
- Масштабируем его вверх: подбираем M так, чтобы для всех x

$$f(x) \leq Mg(x).$$

  То есть $Mg(x)$ — «крышка», которая накрывает целевую плотность $f(x)$.
- Генерируем $x \sim g(x)$ и дополнительное $u \sim \mathrm{Uniform}(0, 1)$.

Rejection sampling: интуиция

Задача

Хотим сэмплировать из сложной плотности $f(x)$, но напрямую это сделать трудно.

- Берём простое распределение $g(x)$, из которого легко сэмплировать (называется *proposal*).
- Масштабируем его вверх: подбираем M так, чтобы для всех x

$$f(x) \leq Mg(x).$$

  То есть $Mg(x)$ — «крышка», которая накрывает целевую плотность $f(x)$.
- Генерируем $x \sim g(x)$ и дополнительное $u \sim \mathrm{Uniform}(0, 1)$.
- Принимаем x, если

$$u \leq \frac{f(x)}{Mg(x)}.$$

Rejection sampling: интуиция

Задача

Хотим сэмплировать из сложной плотности $f(x)$, но напрямую это сделать трудно.

- Берём простое распределение $g(x)$, из которого легко сэмплировать (называется *proposal*).
- Масштабируем его вверх: подбираем M так, чтобы для всех x

$$f(x) \leq Mg(x).$$

  То есть $Mg(x)$ — «крышка», которая накрывает целевую плотность $f(x)$.
- Генерируем $x \sim g(x)$ и дополнительное $u \sim \mathrm{Uniform}(0, 1)$.
- Принимаем x, если

$$u \leq \frac{f(x)}{Mg(x)}.$$

- В среднем каждая точка принимается с вероятностью $1/M$.

Rejection sampling: интуиция

Задача

Хотим сэмплировать из сложной плотности $f(x)$, но напрямую это сделать трудно.

- Берём простое распределение $g(x)$, из которого легко сэмплировать (называется *proposal*).
- Масштабируем его вверх: подбираем M так, чтобы для всех x

$$f(x) \leq Mg(x).$$

  То есть $Mg(x)$ — «крышка», которая накрывает целевую плотность $f(x)$.
- Генерируем $x \sim g(x)$ и дополнительное $u \sim \mathrm{Uniform}(0, 1)$.
- Принимаем x, если

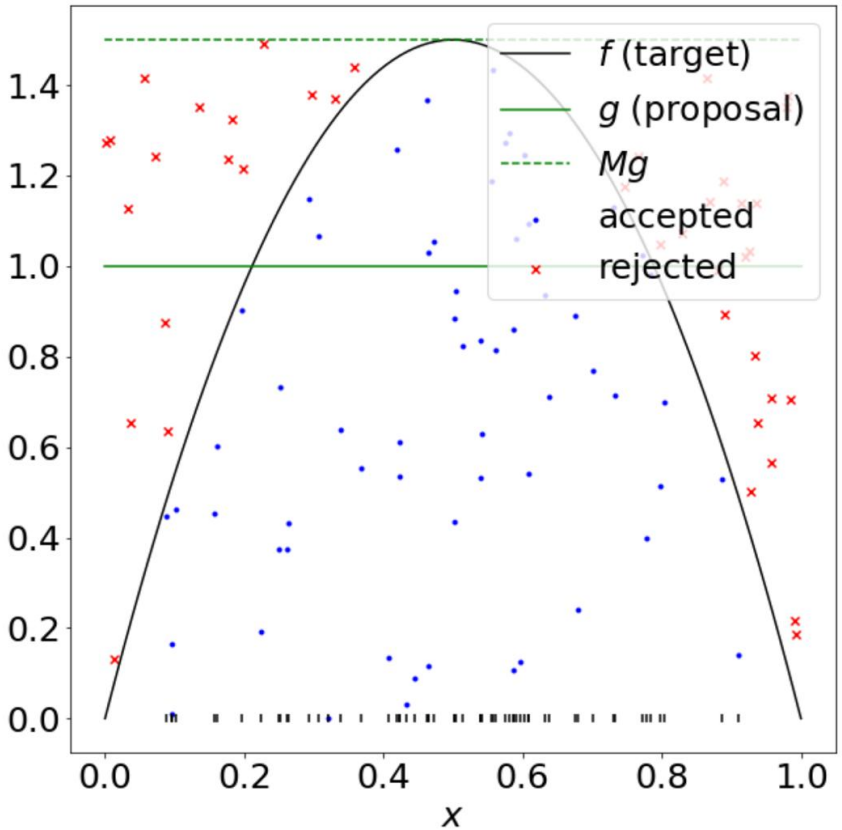
$$u \leq \frac{f(x)}{Mg(x)}.$$

- В среднем каждая точка принимается с вероятностью $1/M$.

Смысл

Мы равномерно «разбрасываем точки» под кривой $Mg(x)$. Оставляем только те, что попали под $f(x)$ → их распределение в точности f.

Пример rejection sampling: Beta(2,2)



- Целевая плотность $f(x) = \mathrm{Beta}(2, 2)$ (чёрная кривая).
- Proposal: $g(x) = \mathrm{Uniform}(0, 1)$ (зелёная линия).
- Масштабированное $Mg(x)$ (зелёный пунктир).
- Синие точки — принятые выборки $\sim \mathrm{Beta}(2, 2)$.
- Красные крестики — отклонённые точки.

Envelope (Delayed-acceptance) Rejection Sampling

Идея

Если вычислять $f(x)$ дорого, используем дешёвую нижнюю оценку $\ell(x) \le f(x)$, чтобы часть выборок принимать «заранее», без обращения к f.

- Сэмплируем $z \sim g,\ u \sim \mathrm{Unif}(0,1)$.
- Если $u \le \dfrac{\ell(z)}{Mg(z)}$, принять z (без вычисления f).
- Иначе вычислить $f(z)$ и принять, если $u \le \dfrac{f(z)}{Mg(z)}$.

Envelope (Delayed-acceptance) Rejection Sampling

Идея

Если вычислять $f(x)$ дорого, используем дешёвую нижнюю оценку $\ell(x) \le f(x)$, чтобы часть выборок принимать «заранее», без обращения к f.

- Сэмплируем $z \sim g,\ u \sim \mathrm{Unif}(0,1)$.
- Если $u \le \dfrac{\ell(z)}{Mg(z)}$, принять z (без вычисления f).
- Иначе вычислить $f(z)$ и принять, если $u \le \dfrac{f(z)}{Mg(z)}$.

Эффект

Доля быстрых приёмов $= \frac{1}{M}\int \ell(x)\,dx$, общая вероятность приёма также $1/M$. Экономим вызовы $f(x)$.

Выбор proposal распределения $g(x)$

- **Основное правило:** $g(x)$ должно «похоже» аппроксимировать $f(x)$.

Выбор proposal распределения $g(x)$

- **Основное правило:** $g(x)$ должно «похоже» аппроксимировать $f(x)$.
- Чем ближе $g(x)$ к форме $f(x)$, тем меньше M и выше эффективность.

Выбор proposal распределения $g(x)$

- **Основное правило:** $g(x)$ должно «похоже» аппроксимировать $f(x)$.
- Чем ближе $g(x)$ к форме $f(x)$, тем меньше M и выше эффективность.
- Хорошие кандидаты:

Выбор proposal распределения $g(x)$

- **Основное правило:** $g(x)$ должно «похоже» аппроксимировать $f(x)$.
- Чем ближе $g(x)$ к форме $f(x)$, тем меньше M и выше эффективность.
- Хорошие кандидаты:
  - Простые распределения, у которых есть аналитический inverse CDF (Uniform, Exponential, Normal).

Выбор proposal распределения $g(x)$

- **Основное правило:** $g(x)$ должно «похоже» аппроксимировать $f(x)$.
- Чем ближе $g(x)$ к форме $f(x)$, тем меньше M и выше эффективность.
- Хорошие кандидаты:
  - Простые распределения, у которых есть аналитический inverse CDF (Uniform, Exponential, Normal).
  - Адаптивные распределения, подобранные под конкретную задачу.

Выбор proposal распределения $g(x)$

- **Основное правило:** $g(x)$ должно «похоже» аппроксимировать $f(x)$.
- Чем ближе $g(x)$ к форме $f(x)$, тем меньше M и выше эффективность.
- Хорошие кандидаты:
  - Простые распределения, у которых есть аналитический inverse CDF (Uniform, Exponential, Normal).
  - Адаптивные распределения, подобранные под конкретную задачу.
- Плохой выбор: слишком «плоское» $g(x) \to M$ большое, почти все выборки отклоняются.

Rejection sampling и проклятие размерности

Одномерный случай

Если $f(x) \leq Mg(x)$, то вероятность принять сэмпл = $1/M$.

Многомерный случай

Пусть $f_d(x) = \prod_{i=1}^{d} f(x_i)$, $g_d(x) = \prod_{i=1}^{d} g(x_i)$, где $x = (x_1, \ldots, x_d)$.
Тогда

$$\frac{f_d(x)}{g_d(x)} = \prod_{i=1}^{d} \frac{f(x_i)}{g(x_i)} \leq M^d.$$

- Значит, в d-мерном случае нужно брать знаменатель $M^d g_d(x)$.

Rejection sampling и проклятие размерности

Одномерный случай

Если $f(x) \leq Mg(x)$, то вероятность принять сэмпл = $1/M$.

Многомерный случай

Пусть $f_d(x) = \prod_{i=1}^{d} f(x_i)$, $g_d(x) = \prod_{i=1}^{d} g(x_i)$, где $x = (x_1, \ldots, x_d)$.
Тогда

$$\frac{f_d(x)}{g_d(x)} = \prod_{i=1}^{d} \frac{f(x_i)}{g(x_i)} \leq M^d.$$

- Значит, в d-мерном случае нужно брать знаменатель $M^d g_d(x)$.
- Вероятность принять сэмпл = $1/M^d$.

Rejection sampling и проклятие размерности

Одномерный случай

Если $f(x) \leq Mg(x)$, то вероятность принять сэмпл = $1/M$.

Многомерный случай

Пусть $f_d(x) = \prod_{i=1}^{d} f(x_i)$, $g_d(x) = \prod_{i=1}^{d} g(x_i)$, где $x = (x_1, \ldots, x_d)$.
Тогда

$$\frac{f_d(x)}{g_d(x)} = \prod_{i=1}^{d} \frac{f(x_i)}{g(x_i)} \leq M^d.$$

- Значит, в d-мерном случае нужно брать знаменатель $M^d g_d(x)$.
- Вероятность принять сэмпл = $1/M^d$.
- $\Rightarrow$ эффективность убывает *экспоненциально* с размерностью d.

Rejection sampling и проклятие размерности

Одномерный случай

Если $f(x) \leq Mg(x)$, то вероятность принять сэмпл = $1/M$.

Многомерный случай

Пусть $f_d(x) = \prod_{i=1}^{d} f(x_i)$, $g_d(x) = \prod_{i=1}^{d} g(x_i)$, где $x = (x_1, \ldots, x_d)$.
Тогда

$$\frac{f_d(x)}{g_d(x)} = \prod_{i=1}^{d} \frac{f(x_i)}{g(x_i)} \leq M^d.$$

- Значит, в d-мерном случае нужно брать знаменатель $M^d g_d(x)$.
- Вероятность принять сэмпл = $1/M^d$.
- $\Rightarrow$ эффективность убывает *экспоненциально* с размерностью d.

Вывод

В высоких размерностях rejection sampling становится практически бесполезным. Требуется экспоненциальное число попыток, чтобы получить хоть один сэмпл.

Rejection sampling для постериора

Напоминание

Постериор:

$$p(\theta \mid D) \propto p(D \mid \theta)\, p(\theta).$$

Часто нормировочная константа $p(D)$ недоступна.

- Для rejection sampling нормировка не нужна — достаточно знать $f(\theta) \propto p(D \mid \theta)p(\theta)$.

Rejection sampling для постериора

Напоминание

Постериор:

$$p(\theta \mid D) \propto p(D \mid \theta)\, p(\theta).$$

Часто нормировочная константа $p(D)$ недоступна.

- Для rejection sampling нормировка не нужна — достаточно знать $f(\theta) \propto p(D \mid \theta)p(\theta)$.
- Выбираем простое proposal $g(\theta)$ (например, гауссиан), и M так, что

$$f(\theta) \leq Mg(\theta) \quad \forall \theta.$$

Rejection sampling для постериора

Напоминание

Постериор:

$$p(\theta \mid D) \propto p(D \mid \theta)\, p(\theta).$$

Часто нормировочная константа $p(D)$ недоступна.

- Для rejection sampling нормировка не нужна — достаточно знать $f(\theta) \propto p(D \mid \theta)p(\theta)$.
- Выбираем простое proposal $g(\theta)$ (например, гауссиан), и M так, что

$$f(\theta) \leq Mg(\theta) \quad \forall \theta.$$

- Алгоритм:

Rejection sampling для постериора

Напоминание

Постериор:

$$p(\theta \mid D) \propto p(D \mid \theta)\, p(\theta).$$

Часто нормировочная константа $p(D)$ недоступна.

- Для rejection sampling нормировка не нужна — достаточно знать $f(\theta) \propto p(D \mid \theta)p(\theta)$.
- Выбираем простое proposal $g(\theta)$ (например, гауссиан), и M так, что

$$f(\theta) \leq Mg(\theta) \quad \forall \theta.$$

- Алгоритм:
  - Сэмплируем $\theta \sim g,\ u \sim \mathrm{Unif}(0, 1)$.

Rejection sampling для постериора

Напоминание

Постериор:

$$p(\theta \mid D) \propto p(D \mid \theta)\, p(\theta).$$

Часто нормировочная константа $p(D)$ недоступна.

- Для rejection sampling нормировка не нужна — достаточно знать $f(\theta) \propto p(D \mid \theta)p(\theta)$.
- Выбираем простое proposal $g(\theta)$ (например, гауссиан), и M так, что

$$f(\theta) \leq Mg(\theta) \quad \forall \theta.$$

- Алгоритм:
  - Сэмплируем $\theta \sim g,\ u \sim \mathrm{Unif}(0,1)$.
  - Принимаем, если $u \leq f(\theta)/(Mg(\theta))$.

Ограничение

Работает для малых размерностей; в больших d из-за проклятия размерности эффективность стремится к нулю.

Rejection sampling для постериора

Напоминание

Постериор:

$$p(\theta \mid D) \propto p(D \mid \theta)\, p(\theta).$$

Часто нормировочная константа $p(D)$ недоступна.

- Для rejection sampling нормировка не нужна — достаточно знать $f(\theta) \propto p(D \mid \theta)p(\theta)$.
- Выбираем простое proposal $g(\theta)$ (например, гауссиан), и M так, что

$$f(\theta) \leq Mg(\theta) \quad \forall \theta.$$

- Алгоритм:
  - Сэмплируем $\theta \sim g,\ u \sim \mathrm{Unif}(0,1)$.
  - Принимаем, если $u \leq f(\theta)/(Mg(\theta))$.
- Принятые сэмплы распределены как $p(\theta \mid D)$.

Ограничение

Работает для малых размерностей; в больших d из-за проклятия размерности эффективность стремится к нулю.

Классический метод Монте-Карло

Задача

Хотим вычислить математическое ожидание

$$\mu = \mathbb{E}_p[f(x)] = \int f(x)\, p(x)\, dx,$$

когда интеграл аналитически не берётся.

- Генерируем $x_1, \ldots, x_N \sim p(x)$.
- Оценка:

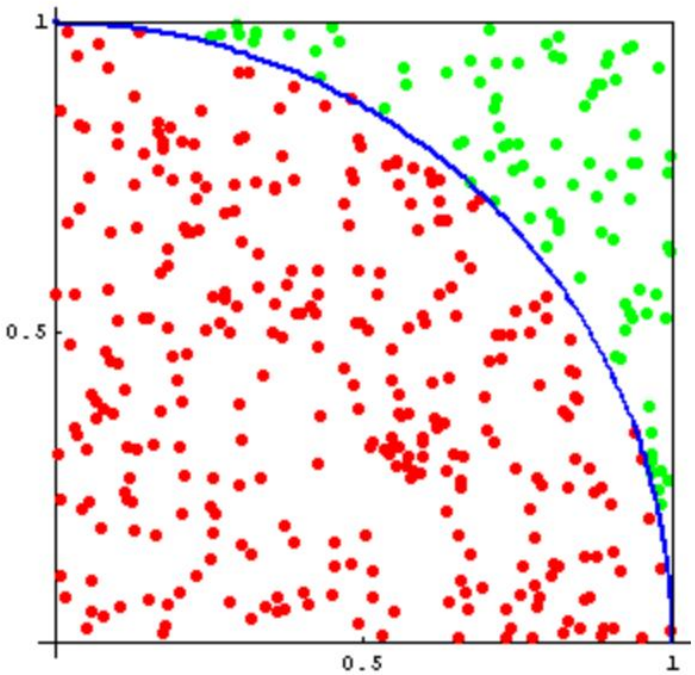
$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} f(x_i).$$

- По закону больших чисел $\hat{\mu} \to \mu$ при $N \to \infty$.
- Стандартная ошибка:

$$\mathrm{SE}(\hat{\mu}) = \frac{\sigma}{\sqrt{N}}, \quad \sigma^2 = \mathrm{Var}_p[f(x)].$$

Пример: оценка числа π

- Рассмотрим единичный квадрат $[0, 1] \times [0, 1]$.
- Площадь четверти круга радиуса 1 = $\pi/4$.
- Алгоритм:
  - Сэмплируем N случайных точек (x, y) из $[0, 1]^2$.
  - Считаем долю точек внутри круга $x^2 + y^2 \leq 1$.
  - Умножаем её на 4 → получаем приближение π.
- При $N \to \infty$ оценка сходится к π.



Классический Monte Carlo: особенности

- **Плюсы:**

Классический Monte Carlo: особенности

- **Плюсы:**
  - Простая и универсальная идея.

Классический Monte Carlo: особенности

- **Плюсы:**
  - Простая и универсальная идея.
  - Точность не зависит от размерности d напрямую.

Классический Monte Carlo: особенности

- **Плюсы:**
  - Простая и универсальная идея.
  - Точность не зависит от размерности d напрямую.
- **Минусы:**

Классический Monte Carlo: особенности

- **Плюсы:**
  - Простая и универсальная идея.
  - Точность не зависит от размерности d напрямую.
- **Минусы:**
  - Нужно уметь напрямую сэмплировать из $p(x)$.

Классический Monte Carlo: особенности

- **Плюсы:**
  - Простая и универсальная идея.
  - Точность не зависит от размерности d напрямую.
- **Минусы:**
  - Нужно уметь напрямую сэмплировать из $p(x)$.
  - Сходимость медленная: ошибка $\sim 1/\sqrt{N}$.

Классический Monte Carlo: особенности

- **Плюсы:**
  - Простая и универсальная идея.
  - Точность не зависит от размерности d напрямую.
- **Минусы:**
  - Нужно уметь напрямую сэмплировать из $p(x)$.
  - Сходимость медленная: ошибка $\sim 1/\sqrt{N}$.
  - Высокая дисперсия возможна для сложных $f(x)$.

Классический Monte Carlo: особенности

- **Плюсы:**
  - Простая и универсальная идея.
  - Точность не зависит от размерности d напрямую.
- **Минусы:**
  - Нужно уметь напрямую сэмплировать из $p(x)$.
  - Сходимость медленная: ошибка $\sim 1/\sqrt{N}$.
  - Высокая дисперсия возможна для сложных $f(x)$.
- Поэтому используются более умные методы: *importance sampling, MCMC*.

Зачем нужен importance sampling?

Проблемы классического MC

1. **Высокая дисперсия оценок.** Если $f(x)$ сильно меняется, одни сэмплы дают огромный вклад, а другие почти ноль → медленная сходимость.

Зачем нужен importance sampling?

Проблемы классического MC

1. **Высокая дисперсия оценок.** Если $f(x)$ сильно меняется, одни сэмплы дают огромный вклад, а другие почти ноль → медленная сходимость.
2. **Невозможность сэмплировать из $p(x)$.** Целевая плотность может быть известна лишь с точностью до константы (как постериор в Байесе: $p(\theta \mid D) \propto p(D \mid \theta)p(\theta)$). Напрямую сэмплировать из неё нельзя.

Зачем нужен importance sampling?

Проблемы классического MC

1. **Высокая дисперсия оценок.** Если $f(x)$ сильно меняется, одни сэмплы дают огромный вклад, а другие почти ноль → медленная сходимость.
2. **Невозможность сэмплировать из $p(x)$.** Целевая плотность может быть известна лишь с точностью до константы (как постериор в Байесе: $p(\theta \mid D) \propto p(D \mid \theta)p(\theta)$). Напрямую сэмплировать из неё нельзя.

Идея importance sampling

Сэмплировать из простого распределения $q(x)$ и *компенсировать разницу* весами.

Importance sampling: основная идея

Задача

Оценить ожидание

$$\mu = \mathbb{E}_p[f(x)] = \int f(x)\, p(x)\, dx,$$

но из $p(x)$ сэмплировать трудно.

- Выбираем proposal $q(x)$, из которого легко сэмплировать.

Importance sampling: основная идея

Задача

Оценить ожидание

$$\mu = \mathbb{E}_p[f(x)] = \int f(x)\, p(x)\, dx,$$

но из $p(x)$ сэмплировать трудно.

- Выбираем proposal $q(x)$, из которого легко сэмплировать.
- Переписываем интеграл:

$$\mu = \int f(x)\, \frac{p(x)}{q(x)}\, q(x)\, dx.$$

Importance sampling: основная идея

Задача

Оценить ожидание

$$\mu = \mathbb{E}_p[f(x)] = \int f(x)\, p(x)\, dx,$$

но из $p(x)$ сэмплировать трудно.

- Выбираем proposal $q(x)$, из которого легко сэмплировать.
- Переписываем интеграл:

$$\mu = \int f(x)\, \frac{p(x)}{q(x)}\, q(x)\, dx.$$

- Алгоритм:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} f(x_i)\, w(x_i), \quad x_i \sim q,\ w(x) = \frac{p(x)}{q(x)}.$$

Автонормализованный importance sampling

Проблема

В Байесе $p(x)$ часто известна лишь с точностью до константы:

$$p(x) = \frac{\tilde{p}(x)}{Z}, \quad Z \text{ неизвестно.}$$

- Обычные веса: $w(x) = p(x)/q(x)$ — посчитать нельзя (нужен Z).

Автонормализованный importance sampling

Проблема

В Байесе $p(x)$ часто известна лишь с точностью до константы:

$$p(x) = \frac{\tilde{p}(x)}{Z}, \quad Z \text{ неизвестно.}$$

- Обычные веса: $w(x) = p(x)/q(x)$ — посчитать нельзя (нужен Z).
- Решение: использовать *ненормированные* веса

$$\tilde{w}(x) = \frac{\tilde{p}(x)}{q(x)}.$$

Автонормализованный importance sampling

Проблема

В Байесе $p(x)$ часто известна лишь с точностью до константы:

$$p(x) = \frac{\tilde{p}(x)}{Z}, \quad Z \text{ неизвестно.}$$

- Обычные веса: $w(x) = p(x)/q(x)$ — посчитать нельзя (нужен Z).
- Решение: использовать *ненормированные* веса

$$\tilde{w}(x) = \frac{\tilde{p}(x)}{q(x)}.$$

- Нормализуем их в выборке:

$$\hat{\mu} = \sum_{i=1}^{N} f(x_i) \, \frac{\tilde{w}(x_i)}{\sum_j \tilde{w}(x_j)}.$$

Автонормализованный importance sampling

Проблема

В Байесе $p(x)$ часто известна лишь с точностью до константы:

$$p(x) = \frac{\tilde{p}(x)}{Z}, \quad Z \text{ неизвестно.}$$

- Обычные веса: $w(x) = p(x)/q(x)$ — посчитать нельзя (нужен Z).
- Решение: использовать *ненормированные* веса

$$\tilde{w}(x) = \frac{\tilde{p}(x)}{q(x)}.$$

- Нормализуем их в выборке:

$$\hat{\mu} = \sum_{i=1}^{N} f(x_i) \frac{\tilde{w}(x_i)}{\sum_j \tilde{w}(x_j)}.$$

- Это называется **auto-normalized importance sampling**.

Автонормализованный importance sampling

Проблема

В Байесе $p(x)$ часто известна лишь с точностью до константы:

$$p(x) = \frac{\tilde{p}(x)}{Z}, \quad Z \text{ неизвестно.}$$

- Обычные веса: $w(x) = p(x)/q(x)$ — посчитать нельзя (нужен Z).
- Решение: использовать *ненормированные* веса

$$\tilde{w}(x) = \frac{\tilde{p}(x)}{q(x)}.$$

- Нормализуем их в выборке:

$$\hat{\mu} = \sum_{i=1}^{N} f(x_i) \frac{\tilde{w}(x_i)}{\sum_j \tilde{w}(x_j)}.$$

- Это называется **auto-normalized importance sampling**.

Замечание

Метод даёт смещённую, но состоятельную оценку: при $N \to \infty$ получаем μ.

Почему importance sampling снижает дисперсию?

Классический MC

$$\hat{\mu} = \frac{1}{N} \sum f(x_i), \quad x_i \sim p.$$

Если $f(x)$ очень вариативна, дисперсия оценки велика.

Importance sampling

$$\hat{\mu} = \frac{1}{N} \sum f(x_i) \frac{p(x_i)}{q(x_i)}, \quad x_i \sim q.$$

- Выбираем $q(x)$ так, чтобы $f(x)p(x)$ стало более «плоским». Тогда веса $w(x)$ компенсируют, и итоговая дисперсия меньше.

Почему importance sampling снижает дисперсию?

Классический MC

$$\hat{\mu} = \frac{1}{N} \sum f(x_i), \quad x_i \sim p.$$

Если $f(x)$ очень вариативна, дисперсия оценки велика.

Importance sampling

$$\hat{\mu} = \frac{1}{N} \sum f(x_i) \frac{p(x_i)}{q(x_i)}, \quad x_i \sim q.$$

- Выбираем $q(x)$ так, чтобы $f(x)p(x)$ стало более «плоским». Тогда веса $w(x)$ компенсируют, и итоговая дисперсия меньше.
- Оптимально: $q(x) \propto |f(x)|p(x)$ (тогда дисперсия = 0, все сэмплы дают одинаковый вклад).

Почему importance sampling снижает дисперсию?

Классический MC

$$\hat{\mu} = \frac{1}{N} \sum f(x_i), \quad x_i \sim p.$$

Если $f(x)$ очень вариативна, дисперсия оценки велика.

Importance sampling

$$\hat{\mu} = \frac{1}{N} \sum f(x_i) \frac{p(x_i)}{q(x_i)}, \quad x_i \sim q.$$

- Выбираем $q(x)$ так, чтобы $f(x)p(x)$ стало более «плоским». Тогда веса $w(x)$ компенсируют, и итоговая дисперсия меньше.
- Оптимально: $q(x) \propto |f(x)|p(x)$ (тогда дисперсия = 0, все сэмплы дают одинаковый вклад).
- На практике выбирают q «похожим» на p, но с более толстыми хвостами.

Effective Sample Size (ESS)

Проблема importance sampling

При сильной разбалансированности весов w_i по сути «работает» только несколько точек.

Определение

$$N_{\text{eff}} = \frac{\left(\sum_{i=1}^{N} w_i\right)^2}{\sum_{i=1}^{N} w_i^2}.$$

- Если все веса равны ($w_i = 1$) $\rightarrow N_{\text{eff}} = N$.

Effective Sample Size (ESS)

Проблема importance sampling

При сильной разбалансированности весов w_i по сути «работает» только несколько точек.

Определение

$$N_{\text{eff}} = \frac{\left(\sum_{i=1}^{N} w_i\right)^2}{\sum_{i=1}^{N} w_i^2}.$$

- Если все веса равны ($w_i = 1$) $\rightarrow N_{\text{eff}} = N$.
- Если один вес огромный, а остальные малы $\rightarrow N_{\text{eff}} \approx 1$.

Effective Sample Size (ESS)

Проблема importance sampling

При сильной разбалансированности весов w_i по сути «работает» только несколько точек.

Определение

$$N_{\text{eff}} = \frac{\left(\sum_{i=1}^{N} w_i\right)^2}{\sum_{i=1}^{N} w_i^2}.$$

- Если все веса равны ($w_i = 1$) → $N_{\text{eff}} = N$.
- Если один вес огромный, а остальные малы → $N_{\text{eff}} \approx 1$.
- Интерпретация: *сколько независимых «честных» сэмплов* у нас реально осталось.

Effective Sample Size (ESS)

Проблема importance sampling

При сильной разбалансированности весов w_i по сути «работает» только несколько точек.

Определение

$$N_{\text{eff}} = \frac{\left(\sum_{i=1}^{N} w_i\right)^2}{\sum_{i=1}^{N} w_i^2}.$$

- Если все веса равны ($w_i = 1$) → $N_{\text{eff}} = N$.
- Если один вес огромный, а остальные малы → $N_{\text{eff}} \approx 1$.
- Интерпретация: *сколько независимых «честных» сэмплов* у нас реально осталось.

Применение

ESS используют для оценки качества importance sampling и для остановки/диагностики сходимости алгоритмов.

Antithetic sampling

Идея

Сэмплы берём парами, симметрично, чтобы ошибки компенсировались.

- Пример: $U \sim \mathrm{Uniform}(0, 1)$. Вместо независимых u_1, u_2 берём пару u, $1 - u$.

Antithetic sampling

Идея

Сэмплы берём парами, симметрично, чтобы ошибки компенсировались.

- Пример: $U \sim \mathrm{Uniform}(0, 1)$. Вместо независимых u_1, u_2 берём пару u, $1 - u$.
- Для монотонных функций $f(x)$ $f(u)$ и $f(1 - u)$ имеют ошибки «в разные стороны».

Antithetic sampling

Идея

Сэмплы берём парами, симметрично, чтобы ошибки компенсировались.

- Пример: $U \sim \mathrm{Uniform}(0,1)$. Вместо независимых u_1, u_2 берём пару u, $1-u$.
- Для монотонных функций $f(x)$ $f(u)$ и $f(1-u)$ имеют ошибки «в разные стороны».
- Среднее по паре:

$$\hat{\mu} = \frac{f(u) + f(1-u)}{2}$$

имеет дисперсию меньше, чем по независимым сэмплам.

Antithetic sampling

Идея

Сэмплы берём парами, симметрично, чтобы ошибки компенсировались.

- Пример: $U \sim \mathrm{Uniform}(0, 1)$. Вместо независимых u_1, u_2 берём пару u, $1 - u$.
- Для монотонных функций $f(x)$ $f(u)$ и $f(1 - u)$ имеют ошибки «в разные стороны».
- Среднее по паре:

$$\hat{\mu} = \frac{f(u) + f(1 - u)}{2}$$

имеет дисперсию меньше, чем по независимым сэмплам.

Вывод

Antithetic sampling снижает дисперсию за счёт введения отрицательной корреляции между сэмплами. Скажем, одна часть идет от толстого хвоста, а другая из тонкого.

Control variates

Идея

Использовать вспомогательную функцию $h(x)$, для которой $\mathbb{E}[h(x)]$ известно точно.

- Новая оценка:

$$\hat{\mu}_{\text{cv}} = \frac{1}{N} \sum_{i=1}^{N} \Big(f(x_i) - c(h(x_i) - \mathbb{E}[h]) \Big).$$

Control variates

Идея

Использовать вспомогательную функцию $h(x)$, для которой $\mathbb{E}[h(x)]$ известно точно.

- Новая оценка:

$$\hat{\mu}_{\text{cv}} = \frac{1}{N} \sum_{i=1}^{N} \Big(f(x_i) - c(h(x_i) - \mathbb{E}[h]) \Big).$$

- Оптимальный коэффициент:

$$c^* = \frac{\text{Cov}(f, h)}{\text{Var}(h)}.$$

Control variates

Идея

Использовать вспомогательную функцию $h(x)$, для которой $\mathbb{E}[h(x)]$ известно точно.

- Новая оценка:

$$\hat{\mu}_{\text{cv}} = \frac{1}{N} \sum_{i=1}^{N} \Big(f(x_i) - c(h(x_i) - \mathbb{E}[h]) \Big).$$

- Оптимальный коэффициент:

$$c^* = \frac{\text{Cov}(f, h)}{\text{Var}(h)}.$$

- Если $h(x)$ коррелирует с $f(x)$, вариации компенсируются и дисперсия $\hat{\mu}_{\text{cv}}$ уменьшается, так как $f(x_i) - ch(x_i)$ меньше изменяется

Control variates

Идея

Использовать вспомогательную функцию $h(x)$, для которой $\mathbb{E}[h(x)]$ известно точно.

- Новая оценка:

$$\hat{\mu}_{\text{cv}} = \frac{1}{N} \sum_{i=1}^{N} \Big(f(x_i) - c(h(x_i) - \mathbb{E}[h]) \Big).$$

- Оптимальный коэффициент:

$$c^* = \frac{\text{Cov}(f, h)}{\text{Var}(h)}.$$

- Если $h(x)$ коррелирует с $f(x)$, вариации компенсируются и дисперсия $\hat{\mu}_{\text{cv}}$ уменьшается, так как $f(x_i) - ch(x_i)$ меньше изменяется

Интуиция

Можно рассмотреть сам variate $c(h(x_i) - \mathbb{E}[h])$ как добавку с нулевым матожиданием, которая может только изменить дисперсию, исходя из корреляции $f(x)$ и $h(x)$.