Метод Монте-Карло на марковских цепях в задачах оценивания малых вероятностей

Логинов Андрей Сергеевич, группа 18.Б04-мм

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Вычислительная стохастика и статистические модели
Научный руководитель: к. ф.-м. н., доцент Коробейников А. И.
Рецензент: программист-исследователь Зенкова Н В.

8 июня 2022г.

Введение

X — вещественная случайная величина.
Рассмотрим задачу оценки следующей вероятности:

$$p = \mathbb{P}(X \geq a).$$

При достаточно больших a вероятность $p \to 0$.
Событие $\{X \geq a\}$ будем называть **редким**, а вероятность p — **малой**.

Примеры:

- Задачи молекулярной динамики
- Цифровые водяные знаки
- Задачи, связанные с выравниванием последовательностей в биоинформатике

Введение

$\mathbb{S}^{(n)}$ — множество строк длины n над алфавитом $\mathfrak{A}$ мощности l.

- Рассмотрим расстояние $d : \mathbb{S}^{(n)} \times \mathbb{S}^{(n)} \to \mathbb{Z}^+$
  $s_0 \in \mathbb{S}^{(n)}$ — фиксированная строка
  $s \in \mathbb{S}^{(n)}$ — случайная строка
- Представляет интерес задача оценки вероятности p события

$$A = \{d(s_0, s) \geq a\}.$$

- Частный случай — расстояние Хэмминга:

$$d_{\mathcal{H}}(s_0, s) = \sum_{k=1}^{n} [s_0^k = s^k].$$

Формула для нахождения p известна, что может быть полезно при отладке.

Метод Монте-Карло

$(X_i)_{i \in \mathbb{N}}$ — последовательность независимых одинаково распределенных случайных величин.

- Для любого $N \in \mathbb{N}$ определена несмещенная оценка p:

$$\hat{p}_{MC} = \frac{1}{N} \sum_{k=1}^{N} [X_i \geq a].$$

- Относительная ошибка:

$$\mathrm{RE}(\hat{p}_{MC}) = \frac{\sqrt{\mathbb{D}\hat{p}_N}}{p} = \sqrt{\frac{1-p}{Np}}.$$

- При $p \approx 10^{-6}$ для достижения относительной ошибки $\varepsilon = 1\%$ требуется объем выборки $N \sim 10^9$.

Постановка задачи

Проблемы оценивания малых вероятностей

- Необходимо добиться уменьшения дисперсии при фиксированном объеме выборки.
- Множество $\mathbb{S}^{(n)}$ дискретно, поэтому не все существующие методы уменьшения дисперсии применимы, например, алгоритм адаптивного многоуровневого расщепления (Bréhier, Lelievre и Rousset 2014).

Задача:

- Исследовать некоторые подходы метода Монте-Карло на марковских цепях для оценивания малых вероятностей событий вида $d(s_0, s) \geq a$.

Метод существенной выборки

Пусть $\mathcal{G}$, $\mathcal{Q}$ — распределения на $\mathbb{S}^{(n)}$:

- g, q — их плотности относительно считающей меры,
- $\mathcal{G} \prec \mathcal{Q}$.

Рассмотрим выборку $s_1, \dots, s_N \sim \mathcal{Q}$ и фиксированную строку $s_0 \in \mathbb{S}^{(n)}$.

Оценка вероятности $p = \mathbb{P}(d(s_0, s) \geq a)$ по методу существенной выборки:

$$\hat{p}_{IS} = \frac{1}{N} \sum_{i=1}^{N} \frac{g(s_i)}{q(s_i)} [d(s_0, s_i) \geq a].$$

Моделирующее распределение

В качестве моделирующего рассмотрим распределение с плотностью вида

$$q(s) = \frac{1}{Z}\tilde{q}(s) = \frac{1}{Z}g(s)w(s),$$

где Z — нормирующая константа, $w(s) = \exp\{\gamma \cdot d(s_0, s)\}$ — весовая функция.

При больших γ выше вероятность попасть в область с высоким $d(s_0, s)$, но выше ошибка от оценки нормирующий константы Z.

Оценка по методу существенной выборки примет вид

$$\hat{p}_{IS} = \frac{\sum\limits_{i=1}^{N} [s_i \in A](w(s_i))^{-1}}{\sum\limits_{i=1}^{N} (w(s_i))^{-1}}.$$

Алгоритм Метрополиса-Гастингса (Hastings 1970)

Для построения оценки $\hat{p}_{IS}$ необходимо моделировать распределение Q с плотностью специального вида.

- Алгоритм Метрополиса-Гастингса позволяет строить цепь Маркова, стационарное распределение которой есть нужное $\mathcal{Q}$.
- Чтобы построить такую марковскую цепь, достаточно знать плотность распределения $\mathcal{Q}$ с точностью до нормирующей константы.

Обозначим $\hat{p}_{MH}$ оценку вероятности p, построенную с помощью алгоритма Метрополиса-Гастингса.

Метод зацепления (Geyer 1991)

- Параллельно моделируются r траекторий $S_0, \dots, S_{r-1}$ марковской цепи, весовые функции стационарного распределения которых отличаются только значением γ_i.
- Каждые T итераций случайным образом производится попарный обмен текущих состояний цепей, как шаг алгоритма Метрополиса-Гастингса.
- В конце по траектории S_0 строится оценка.
- Нормирующая константа Z оценивается по независимой одномерной цепи.
- Обозначим $\hat{p}_{MC3}$ оценку вероятности p, построенную с использованием метода зацепления.

Программная реализация

- Описанные ранее методы были реализованы на языке R.
- Для определения оптимальных параметров метода зацепления было необходимо провести большой объем моделирования, требующий много времени.
- Ввиду специфики рассматриваемых методов и языка R добиться на нем существенного сокращения времени не представлялось возможным.
- Использование языка C++ и библиотеки Rcpp (Eddelbuettel и François 2011; Eddelbuettel 2013; Eddelbuettel и Balamuta 2018) для реализации алгоритма Метрополиса-Гастингса и метода зацепления позволило значительно увеличить производительность.

Определение параметров метода зацепления

Для определения оптимальных параметров было проведено моделирование для оценки вероятностей $\mathbb{P}(d_{\mathcal{H}}(s_0, s) \geq 15)$ и $\mathbb{P}(d_{\mathcal{H}}(s_0, s) \geq 18)$, где s_0, $s \in \mathbb{S}^{(20)}$, $l = 4$.

Выводы:

- Среднеквадратичная ошибка и относительная ошибка убывают с ростом r.
- Частота обмена состояний не оказывает существенного влияния на точность.
- Свойства оценок сильно зависят от набора параметров $(\gamma_j)_{j=1}^{r}$.
- Оптимальным является выбор всех γ_j равными, причем для оцениваемых вероятностей оптимальное значение γ различно.

Оценка дисперсии и доверительные интервалы

- Оценка дисперсии $\hat{p}_{MC2}$ с помощью метода batch means (Jones и др. 2006):

$$\mathrm{Var}(\hat{p}_{MC2}) = \frac{\hat{\sigma}^2_{BM}}{N} = \frac{v}{(u-1)N} \sum_{j=1}^{u} (p_j - \hat{p}_{MC2})^2,$$

где $u \cdot v = N$, а p_j находится по формуле:

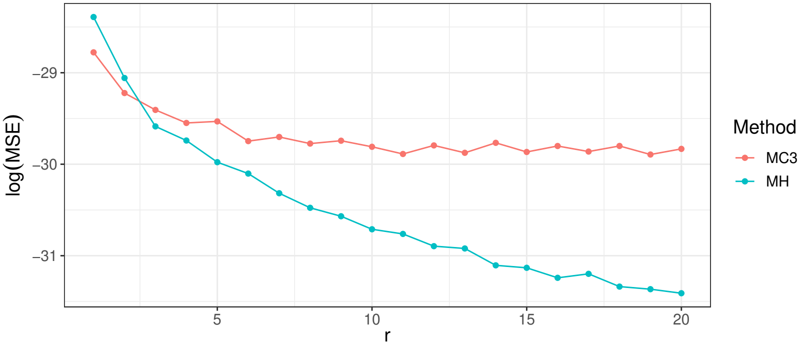
$$p_j = \frac{1}{v} \sum_{i=(j-1)v}^{jv-1} \frac{g(s_i)}{q(s_i)} [s_i \in A],$$

- Доверительные интервалы получаются из нормальности оценки (Owen 2013).

Для оценок $\hat{p}_{MC3}$ аналогичные результаты в данной работе были получены экспериментально.

Численные эксперименты. Сравнение метода зацепления и алгоритма Метрополиса-Гастингса

Сравнение $\hat{p}_{MC3}$ и $\hat{p}_{MH}$ при $N = 10^5$ и $r \in \{1, \dots 20\}$ для $p \approx 3.81 \cdot 10^{-6}$ по выборкам оценок объема $M = 1000$.



- При одинаковой трудоемкости метод зацепления существенно уступает алгоритму Метрополиса-Гастингса в точности.

Сравнение со стандартным методом Монте-Карло

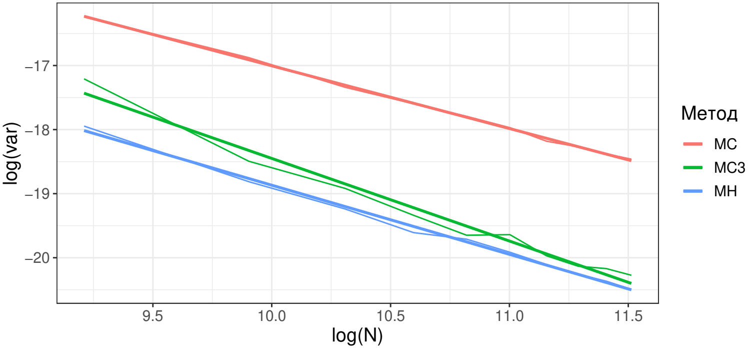
Для оценок построенных при одинаковом порядке трудоемкости:

p	$RE\hat{p}_{MC}$	$RE\hat{p}_{MH}$	$RE\hat{p}_{MC з}$
$1.84{\cdot}10^{-4}$	0.239	0.054	0.060
$2.95 \cdot 10^{-5}$	0.591	0.080	0.091
$3.81 \cdot 10^{-6}$	1.337	0.127	0.147
$3.87 \cdot 10^{-7}$	3.659	0.227	0.269
$2.96 \cdot 10^{-8}$	-	0.238	0.280
$1.61 \cdot 10^{-9}$	-	0.535	0.671
$9.10 \cdot 10^{-13}$	-	2.389	8.489

- Метод зацепления, хотя и уступает в точности алгоритму Метрополиса-Гастингса, но значительно превосходит стандартный метод Монте-Карло при оценивании малых вероятностей.

Сравнение скорости убывания дисперсии

Зависимость дисперсии оценок от объема выборки (трудоемкости в случае $\hat{p}_{MC3}$):



- Скорости убывания дисперсий оценок с ростом объема выборок имеют одинаковый порядок, однако метод Монте-Карло на марковских цепях существенно выигрывает в точности.

Заключение

В работе был изучен метод зацепления и его применение для оценивания малых вероятностей в дискретном случае.

- Алгоритм Метрополиса-Гастингса и метод зацепления были реализованы на языках R и C++.
- Экспериментально проверена асимптотическая нормальность построенных оценок.
- Экспериментально проверена состоятельность batch means, как метода оценивания дисперсии оценок, использующих метод зацепления.
- Алгоритм Метрополиса-Гастингса при фиксированной трудоемкости позволяет добиться большой точности в сравнении с методом зацепления.
- Использование обоих рассмотренных методов обеспечивает заметный выигрыш в точности в сравнении со стандартным методом Монте-Карло.