Федеральное государственное автономное образовательное
учреждение высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»

Факультет математики

на правах рукописи

Иосипой Леонид Сергеевич

Снижение дисперсии оценок Монте-Карло

Резюме диссертации
на соискание ученой степени
кандидата математических наук

Научный руководитель:
кандидат физико-математических наук,
профессор Д. В. Беломестный

Москва – 2021

Введение

В данной работе изучается задача вычисления математического ожидания $\pi(f) := \mathsf{E}\,[f(X)]$ некоторой функции $f:\ \mathsf{X} \to \mathbb{R}$ от случайного вектора X с плотностью распределения $\pi(x)$ и значениями в множестве $\mathsf{X} \subset \mathbb{R}^d$ в предположении, что $f \in L^2(\pi)$. Мы будем предполагать, пока не сказано иное, что функция $f(x)$ и плотность $\pi(x)$ известны.

Сформулированную задачу можно рассмотреть как задачу вычисления интеграла от функции $f(x) \cdot \pi(x)$ по множеству X. Мы будем предполагать, что функция $f(x)$ и плотность $\pi(x)$ достаточно сложные, и данный интеграл явно не вычисляется. Поэтому мы будем оценивать его с помощью алгоритмов численного интегрирования. Известно, что любые детерминированные (даже адаптивные) алгоритмы численного интегрирования «страдают» от так называемого «проклятия размерности», которое означает, что сложность данных алгоритмов экспоненциально растет с ростом размерности d, см., например, классическую работу Н. С. Бахвалова [B71] и современный обзор Е. Новака [N16]. В свою очередь, рандомизированные алгоритмы численного интегрирования, к которым относятся методы Монте-Карло, оказываются эффективнее других подходов к решению данной задачи, если размерность d является большой и/или функция f является сложно вычислимой, см., например, обзор Г. Робертса и Д. Розенталя [RR04] или книгу К. Роберта и Д. Каселлы [RC99].

В случае, когда можно получить независимую выборку $X_1 \ldots, X_n$ из распределения с плотностью $\pi(x)$, математическое ожидание $\pi(f)$ можно оценить с помощью оценки Монте-Карло

$$\pi_n(f) := \frac{1}{n}\sum_{k=1}^{n} f(X_k).$$

Данная оценка является несмещенной, то есть $\mathrm{E}[\pi_n(f)] = \pi(f)$, и имеет дисперсию, равную $\mathrm{Var}(\pi_n(f)) = V(f)/n$, где, здесь и далее, $V(f)$ обозначает дисперсию функции f по отношению к π, то есть

$$V(f) := \mathrm{Var}(f(X)), \quad X \sim \pi.$$

С помощью центральной предельной теоремы можно построить асимптотический доверительный интервал для $\pi(f)$, который будет иметь вид

$$\pi_n(f) \pm c\sqrt{\frac{V(f)}{n}}, \tag{1}$$

где c — квантиль нормального распределения. С точки зрения приложений, интересна задача улучшения точности оценки Монте-Карло $\pi_n(f)$, что в свою очередь означает уменьшение длины вышеупомянутого доверительного интервала. Этого можно добиться, увеличив размер выборки n, но в некоторых случаях данное решение может быть нецелесообразно (например, в связи со сложностью генерирования случайных величин из плотности $\pi(x)$ или сложностью вычисления значений функции $f(x)$). Другим походом к уменьшению длины доверительного интервала является уменьшение величины $V(f)$ за счет перехода к другой оценке Монте-Карло, у которой бы сохранялось среднее, а дисперсия была бы меньше. Такие методы называются методами снижения дисперсии.

По методам снижения дисперсии доступна обширная литература, см., например, классические книги И. Димова [D08], П. Глассермана [G13], Р. Рубинштейна и Д. Кроуза [RK16]. Стоит отметить, что большая часть методов снижения дисперсии явно используют структуру функции $f(x)$ либо плотности $\pi(x)$, поэтому не применимы в самых общих постановках. Одним из немногих общих методов снижения дисперсии является метод контрольных функций. Его идея заключается в следующем. Допустим, нам дан класс функций $\mathcal{G} \subset L^2(\pi)$, который обладают свойством, что для любой $g \in \mathcal{G}$ выполняется $\pi(g) = 0$. Такие функции называются контрольными. Тогда можно рассмотреть новую оценку Монте-Карло

$$\pi_n\big(f - \bar{g}\big) := \frac{1}{n}\sum_{k=1}^{n}\big(f(X_k) - \widehat{g}(X_k)\big),$$

где $\bar{g}$ — функция из $\mathcal{G}$, на которой достигается (или почти достигается) минимум дисперсии $V\big(f - g\big) = \operatorname{Var}\big(f(X) - g(X)\big)$, $X \sim \pi$, то есть

$$\bar{g} \in \operatorname*{argmin}_{g \in \mathcal{G}} V\big(f - g\big). \tag{2}$$

В случае, когда класс контрольных функций $\mathcal{G}$ выбран хорошо, такой поход может значительно снизить дисперсию. Здесь возникают два ключевых вопроса. Как конструктивно построить класс контрольных функций $g \in \mathcal{G}$ и как численно решить задачу (2), так как дисперсия $V(f-g)$ не предполагается известной, а ее вычисление — более сложная задача, чем исходная.

Вопрос построения класса контрольных функций $\mathcal{G}$ был изучен ранее в работе А. Миры, Р. Солги, Д. Импарато [MSI13] и работах К. Оутса, М. Джиролами, Н. Шопена [OGC17] и [OGC19], поэтому на нем подробно мы останавливаться не будем. Мы лишь упомянем, что популярным методом построения функций $g \in \mathcal{G}$ является подстановка в

$$g_\Phi = \langle \Phi, \nabla \log \pi \rangle + \operatorname{div}(\Phi) \tag{3}$$

различных функций $\Phi : \mathsf{X} \to \mathbb{R}^d$; здесь $\langle \cdot, \cdot \rangle$ — стандартное скалярное произведение в $\mathbb{R}^d$, а $\operatorname{div}(\Phi)$ — дивергенция Φ. При достаточно слабых условиях на π и Φ, с помощью интегрирования по частям можно показать, что $\pi(g_\Phi) = 0$, см. Предложения 1 и 2 в [MSI13]. Контрольные функции вида (3) называются контрольными функциями Стейна.

Что касается вопроса о численном решении задачи (2), кажется естественным заменить истинную дисперсию $V(f - g)$ на выборочную дисперсию $V_n(f - g)$, построенную по выборке

$X_1, \ldots, X_n$, то есть на

$$V_n(f-g) := \frac{1}{n-1} \sum_{k=1}^{n} \bigl(f(X_k) - g(X_k) - \pi_n(f-g)\bigr)^2. \tag{4}$$

Кроме того, мы также заменим класс $\mathcal{G}$ на его дискретную конечную аппроксимацию $\mathcal{G}'$, так как численная оптимизация по бесконечному множеству функций является сложной задачей, которую в общем случае решить невозможно. Возникает естественный вопрос: сколько мы «проиграем» за счет замены истинной дисперсии на эмпирическую и множества $\mathcal{G}$ на его аппроксимацию $\mathcal{G}'$? Ответ на этот вопрос является основной целью данной работы.

Настоящее резюме диссертации организовано следующим образом. В Разделе 1 будут сформулированы базовые определения, которые будут использоваться в следующих разделах. В Разделе 2, мы оценим с большой вероятностью разность

$$V\bigl(f - \widehat{g}\,\bigr) - \inf_{g \in \mathcal{G}} V\bigl(f - g\bigr), \tag{5}$$

где

$$\widehat{g} \in \operatorname*{argmin}_{g \in \mathcal{G}'} V_n(f-g), \quad \mathcal{G}' \text{ является } \varepsilon\text{-сетью для } \mathcal{G},$$

а дисперсия $V(f - \widehat{g}\,)$ является условной по выборке $X_1, \ldots, X_n$, на которой вычислялась оценка $\widehat{g}$. Подобные оценки часто встречаются в статистической литературе и называются оценками на «решетках» (в английской литературе skeleton estimates или sieve estimates), см., например, работы В. Вонг и К. Шен [WS95], Л. Девроя, Л. Гёрфи, Г. Лугоши [DGL96] или С. ван де Гир [G00]. Далее, в Разделе 3 мы рассмотрим случай, когда независимую выборку $X_1 \ldots, X_n$ из распределения с плотностью $\pi(x)$ явно получить нельзя, но сама $\pi(x)$ является известной (возможно, с точностью до нормировочной константы). В таких случаях часто можно воспользоваться алгоритмами Монте-Карло на основе марковских цепей (MCMC в английской литературе) и получить марковскую цепь $\{X_k\}_{k=1}^{\infty}$, которая сходится по распределению к распределению с плотностью $\pi(x)$. Для марковских цепей, при некоторых условиях, также верна центральная предельная теорема и, следовательно, с ее помощью можно построить асимптотический доверительный интервал

$$\pi_n(f) \pm c\sqrt{\frac{V^{\infty}(f)}{n}}, \tag{6}$$

где $V^{\infty}(f)$ — асимптотическая дисперсия, которая определяется как

$$V^{\infty}(f) := \lim_{n \to \infty} n \cdot \mathrm{E}\left[\left(\frac{1}{n}\sum_{k=1}^{n} f(X_k) - \pi(f)\right)^2\right]. \tag{7}$$

Ввиду возможных попарных корреляций у цепи $\{X_k\}_{k=1}^{\infty}$, асимптотическая дисперсия $V^{\infty}(f)$ не равна в общем случае дисперсии $V(f)$. Следовательно, в данном случае целесообразно подбирать контрольные функции $g \in \mathcal{G}$, оптимизируя уже асимптотическую дисперсию $V^{\infty}(f-g)$. Так как асимптотическая дисперсия так же, как и обычная дисперсия, не предполагается известной, мы ее заменим на ее оценку по выборке. Итак, в Разделе 3 мы оценим с большой вероятностью разность

$$V^{\infty}\big(f - \widehat{g}\,\big) - \inf_{g \in \mathcal{G}} V^{\infty}\big(f - g\big), \tag{8}$$

где уже для некоторой оценки V_n^{∞} асимптотической дисперсии V^{∞}

$$\widehat{g} \in \operatorname*{argmin}_{g \in \mathcal{G}'} V_n^{\infty}\left(f - g\right), \quad \mathcal{G}' \text{ является } \varepsilon\text{-сетью для } \mathcal{G},$$

а дисперсия $V^{\infty}(f - \widehat{g}\,)$ является условной по выборке $X_1, \ldots, X_n$, на которой вычислялась оценка $\widehat{g}$. Данное обобщение оказывается нетривиальным по двум причинам. Во-первых, в связи с попарными корреляциями, оценивание асимптотической дисперсии возможно несколькими способами и требует специальных техник, таких как спектральные методы или разделение марковской цепи на слабо коррелирующие блоки, см. обзор методов оценки асимптотической дисперсии в работе Д. Флегала и Д. Джонса [FJ10]. Во-вторых, неасиптотический анализ оценки $f - \widehat{g}$ требует обобщения используемых в доказательстве неравенств концентрации на цепи Маркова, которое также не является тривиальным. Мы проведем данный анализ для достаточно широкого класса марковских цепей, который включает в себя многие MCMC алгоритмы. Наконец, в Разделе 4 мы рассмотрим приложения полученных результатов.

Цель работы

Цель настоящей работы — изучение теоретических свойств описанного выше метода снижения дисперсии оценок Монте-Карло (метода контрольных функций) и, в частности, получение неасимптотических оценок на избыточный риск (5) и (8) в независимом и зависимом случае соотвественно.

1. Обозначения и определения

Перед тем как сформулировать основные результаты, введем еще некоторые обозначения. Обозначим класс функций $h(x) = f(x) - g(x)$ для $g \in \mathcal{G}$ через $\mathcal{H}$:

$$\mathcal{H} := \{f - g : \ g \in \mathcal{G}\},$$

где $\mathcal{G}$ — класс контрольных функции (то есть таких, что $\pi(g) = 0$ для всех $g \in \mathcal{G}$). Заметим, что у всех функций $h \in \mathcal{H}$ будет одинаковое математическое ожидание, равное $\pi(f)$. Как уже упоминалось, вместо оптимизации по всему классу $\mathcal{H}$ мы будем искать минимум в его конечной аппроксимации. Зафиксируем $\varepsilon > 0$ и положим $r = 1$ или $r = 2$. Предполагая, что класс $\mathcal{H}$ вполне ограничен, обозначим через $\mathcal{H}_{\varepsilon,r}$ произвольную минимальную по мощности ε-сеть множества $\mathcal{H}$ в $L^r(\pi)$-норме, то есть такой минимальный набор функций $\mathcal{H}_{\varepsilon,r} = \{h_1, \ldots, h_m\} \subset \mathcal{H}$, что для произвольной $h \in \mathcal{H}$ существует $h_* \in \mathcal{H}_{\varepsilon,r}$, что расстояние между h и h_* в $L^r(\pi)$-норме не превосходит ε. Обозначим метрическую энтропию $\mathcal{H}$ через $H_{L^r(\pi)}(\mathcal{H}, \varepsilon) := \log |\mathcal{H}_{\varepsilon,r}|$, где $|\mathcal{H}_{\varepsilon,r}|$ — это мощность $\mathcal{H}_{\varepsilon,r}$. Далее, определим величину, которая будет использована в основных результатах данной работы:

$$\gamma_{L^r(\pi)}(\mathcal{H}, n) := \inf\bigl\{\eta > 0 : H_{L^r(\pi)}(\mathcal{H}, \eta) \le n\eta^r\bigr\}.$$

Обратим внимание, что $\eta > 0$, удовлетворяющее неравенству $H_{L^r(\pi)}(\mathcal{H}, \eta) \le n\eta^r$, существует и конечно, так как метрическая энтропия $H_{L^r(\pi)}(\mathcal{H}, \eta)$ является убывающей функцией по η, а отображение $\eta \mapsto n\eta^r$ является строго возрастающим. Значение $\gamma_{L^r(\pi)}(\mathcal{H}, n)$ будет использовано для того, чтобы оценить мощность класса $\mathcal{H}_{\varepsilon,r}$. Действительно, выбирая $\varepsilon \ge \gamma_{L^r(\pi)}(\mathcal{H}, n)$ мы получим $|\mathcal{H}_{\varepsilon,r}| \le \mathrm{e}^{n\varepsilon^r}$. Легко также видеть из определения, что $\gamma_{L^r(\pi)}(\mathcal{H}, n)$ является убывающей функцией по n.

Приведем теперь оценку на $\gamma_{L^r(\pi)}(\mathcal{H}, n)$ в достаточно общем случае, когда $\mathcal{H}$ является подмножеством взвешенного соболевского пространства. Напомним, что соболевское пространство $W^{s,p}(\mathsf{X})$ определяется как $W^{s,p}(\mathsf{X}) = \bigl\{h \in L^p(\lambda) :\ D^\alpha h \in L^p(\lambda),\ \forall |\alpha| \leqslant s\bigr\}$, где λ является мерой Лебега, $\alpha = (\alpha_1, \ldots, \alpha_d)$ является мульти-индексом с $|\alpha| = \alpha_1 + \ldots + \alpha_d$, а D^α является дифференциальным оператором вида $D^\alpha = \partial^{|\alpha|}/\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}$. Здесь все производные понимаются в слабом смысле. Взвешенное соболевское пространство $W^{s,p}(\mathsf{X}, \langle x\rangle^\beta)$ для полиномиальной весовой функции $\langle x\rangle^\beta = \bigl(1 + \|x\|^2\bigr)^{\beta/2}$, $\beta \in \mathbb{R}$, определяется следующим образом:

$$W^{s,p}(\mathsf{X}, \langle x\rangle^\beta) = \bigl\{h \in L^p(\lambda) :\ h \cdot \langle x\rangle^\beta \in W^{s,p}(\mathsf{X})\bigr\}.$$

Норма в данном взвешенном соболевском пространстве определяется как

$$\|h\|_{W^{s,p}_\beta} = \sum_{|\alpha| \le s} \bigl\|D^\alpha\bigl(h \cdot \langle x\rangle^\beta\bigr)\bigr\|_{L^p(\lambda)}.$$

Мы будем говорить, что $\mathcal{H} \subset W^{s,p}(\mathsf{X}, \langle x\rangle^\beta)$ является ограниченным подмножеством пространства $W^{s,p}(\mathsf{X}, \langle x\rangle^\beta)$, если $\mathcal{H}$ вполне ограничено и существует константа $c > 0$, такая что $\|h\|_{W^{s,p}_\beta} \le c$ для всех $h \in \mathcal{H}$. Верно следующее предложение.

Предложение 1. *Пусть $\mathcal{H}$ будет (непустым) ограниченным подмножеством $W^{s,p}(\mathbb{R}^d, \langle x\rangle^\beta)$, где $1 < p < \infty$, $\beta \in \mathbb{R}$ и $s - d/p > 0$. Пусть также для некоторого $\kappa > 0$, $\|\langle x\rangle^{\kappa-\beta}\|_{L^r(\pi)} < \infty$. Тогда для $r = 1, 2$,*

$$\gamma_{L^r(\pi)}(\mathcal{H}, n) \lesssim \begin{cases} n^{-\frac{1}{r+d/s}} & \text{при } \kappa > s - d/p, \\ n^{-\frac{1}{r+(\kappa/d+1/p)^{-1}}} & \text{при } \kappa < s - d/p, \end{cases}$$

где символ $\lesssim$ означает неравенство с точностью до некоторой константы, не зависящей от n.

Данный результат получен подстановкой в определение $\gamma_{L^r(\pi)}(\mathcal{H}, n)$ оценки на энтропию ограниченных подмножеств $W^{s,p}(\mathsf{X}, \langle x\rangle^\beta)$ из Следствия 4 работы Р. Никля и Б. Петшера [NP07]. Предложение 1 не является основным результатом диссертационной работы и приведено здесь в качестве примера возможных значений величины $\gamma_{L^r(\pi)}(\mathcal{H}, n)$, которая далее будет использоваться в основных результатах.

Везде далее распределение с плотностью $\pi(x)$ и соответствующую ему вероятностную меру мы также будем обозначать буквой π, но это не приведет к двусмысленным трактовкам. Если не оговорено иное, символ $\lesssim$ будет означать неравенство с точностью до некоторой абсолютной константы, а символ $\propto$ будет означать равенство с точностью до абсолютной константы.

2. Снижение дисперсии в независимом случае

Перед тем как сформулировать основные результаты для независимого случая, напомним, что для класса $\mathcal{H} = \{f - g :\ g \in \mathcal{G}\}$ оценку, свойства которой мы будем изучать, можно записать как

$$\widehat{h}_{\varepsilon,r} \in \operatorname*{argmin}_{h\in\mathcal{H}_{\varepsilon,r}} V_n(h),$$

где $V_n(h)$ — эмпирическая дисперсия, определенная в (4), а $\mathcal{H}_{\varepsilon,r}$ — ε-сеть множества $\mathcal{H}$ в норме $L^r(\pi)$, $r = 1, 2$. Подробные определения даны в Разделе 1. Здесь и далее мы будем предполагать, что функция $\widehat{h}_{\varepsilon,r}$ существует (все утверждения и их доказательства легко переносятся на случай, когда минимум в классе $\mathcal{H}_{\varepsilon,r}$ не достигается).

Теперь мы можем перейти к основным результатам раздела. Мы начнем со случая, когда функции $h \in \mathcal{H}$ являются ограниченными и класс $\mathcal{H}$ является замкнутым и выпуклым множеством.

Теорема 2. *Пусть класс $\mathcal{H}$ является замкнутым и выпуклым множеством. Пусть также $\sup_{h\in\mathcal{H}} \sup_{x\in\mathsf{X}} |h(x)| \le b$ для некоторого $b > 0$ и пусть $\pi(h) = c$ для всех $h \in \mathcal{H}$ и некоторой*

$c \in \mathbb{R}$. Тогда для некоторого $\varepsilon > 0$ (значение которого определено в доказательстве) и любого $\delta \in (0,1)$, с вероятностью не менее $1-\delta$,

$$V\Big(\widehat{h}_{\varepsilon,1}\Big) - \inf_{h \in \mathcal{H}} V\big(h\big) \lesssim b^2 \gamma_{L^1(\pi)}(\mathcal{H}, n) + \frac{b^2 \log(1/\delta)}{n},$$

где первая дисперсия является условной по выборке $X_1, \ldots, X_n$, на которой считалась оценка $\widehat{h}_{\varepsilon,1}$, а символ $\lesssim$ означает неравенство с точностью до некоторой абсолютной константы.

Теперь сформулируем подобный результат, когда вместо выпуклости класса $\mathcal{H}$, мы потребуем, чтобы $\mathcal{H}$ содержал функцию, тождественно равную константе. Так как по построению $\pi(h) = \pi(f)$ для всех $h \in \mathcal{H}$, эта константа должна быть равна $\pi(f)$ и, следовательно, в этом случае $\inf_{h \in \mathcal{H}} V(h) = 0$.

Теорема 3. *Пусть класс $\mathcal{H}$ содержит константную функцию, то есть $h^*(x) \equiv c$ для некоторой $h^* \in \mathcal{H}$ и некоторого $c \in \mathbb{R}$. Пусть также $\sup_{h \in \mathcal{H}} \sup_{x \in \mathsf{X}} |h(x)| \leq b$ для некоторого $b > 0$ и пусть $\pi(h) = c$ для всех $h \in \mathcal{H}$ и некоторой $c \in \mathbb{R}$. Тогда для некоторого $\varepsilon > 0$ (значение которого определено в доказательстве) и любого $\delta \in (0,1)$, с вероятностью не менее $1-\delta$,*

$$V\Big(\widehat{h}_{\varepsilon,1}\Big) \lesssim b^2 \gamma_{L^1(\pi)}(\mathcal{H}, n) + \frac{b^2 \log(1/\delta)}{n}$$

и

$$V\Big(\widehat{h}_{\varepsilon,2}\Big) \lesssim b^2 \big(\gamma_{L^2(\pi)}(\mathcal{H}, n)\big)^2 + \frac{b^2 \log(1/\delta)}{n},$$

где дисперсии являются условными по выборке $X_1, \ldots, X_n$, на которой считались оценки $\widehat{h}_{\varepsilon,1}$ и $\widehat{h}_{\varepsilon,2}$, а символ $\lesssim$ означает неравенство с точностью до некоторой абсолютной константы.

Предполагая дополнительно к условиям Теоремы 2 или Теоремы 3, что $\mathcal{H}$ является ограниченным подмножеством некоторого взвешенного соболевского пространства $W^{s,p}(\mathsf{X}, \langle x \rangle^{\beta})$, и используя результат Предложения 1, мы получим окончательную оценку с большой вероятностью вида:

$$V\Big(\widehat{h}_{\varepsilon,1}\Big) - \inf_{h \in \mathcal{H}} V\big(h\big) \lesssim n^{-\alpha}, \quad \alpha \in (0,1), \tag{9}$$

где значение параметра α будет зависеть от размерности d и параметров соболевского пространства s, p и β. Подобные порядки сходимости называются «быстрыми» в статистической литературе, так как позволяют получать скорость сходимости быстрее стандартных порядков $1/\sqrt{n}$ при параметрическом оценивании. В качестве следствия стоит отметить, что порядок длины асимптотического доверительного интервала (1) с помощью данной процедуры снижения дисперсии можно уменьшить с $n^{-1/2}$ до $n^{-1/2-\alpha/2}$, если класс $\mathcal{H}$ выбран так, что $\inf_{h \in \mathcal{H}} V\big(h\big) \lesssim n^{-\alpha}$.

Насколько автору известно, Теорема 2 и Теорема 3 являются первыми доступными в литературе результатами об оценках, полученных минимизацией эмпирической дисперсии $V_n(h)$. Отличительной особенностью данных результатов является тот факт, что эмпирическая дисперсия $V_n(h)$ не является суммой независимых случайных величин $h(X_1), \ldots, h(X_n)$, как часто бывает в теории эмпирических процессов. Вместо этого, $V_n(h)$ является U-статистикой. «Быстрые» порядки в задачах минимизации U-статистик были ранее получены, насколько известно автору, только в работе С. Клеменсона, Г. Лугоши, Н. Ваятиса [CLV08] (и некоторых других последующих работах авторов), где рассматривалась задача ранжирования случайных величин. В доказательствах Теоремы 2 и Теоремы 3 автором используется другая техника, основанная на дискретизации множества $\mathcal{H}$ и концентрации U-статистик, что позволяет избежать многих технических проблем, которые присутствовали в [CLV08].

Приведем теперь здесь два следствия Теоремы 3, в которых мы явно вычислим значение параметра α из оценки (9). Для простоты мы рассмотрим одномерный случай $\mathsf{X} = \mathbb{R}$ и в качестве контрольных функций мы возьмем контрольные функции Стейна (3), которые в одномерном случае можно записать как

$$g_\phi(x) := \phi(x) \cdot \big(\log \pi(x)\big)' + \phi'(x) = \frac{1}{\pi(x)}\big(\phi(x)\pi(x)\big)',$$

где $\phi : \mathbb{R} \to \mathbb{R}$ — функции из некоторого класса Φ, который мы определим позднее. Обозначим класс s раз непрерывно дифференцируемых функций на $\mathbb{R}$ через $C^s(\mathbb{R})$ и класс функций из $C^s(\mathbb{R})$, производные которых растут не быстрее некоторого полинома через $C^s_{poly}(\mathbb{R})$, через

$$C^s_{poly}(\mathbb{R}) := \big\{\phi \in C^s(\mathbb{R}) : \ \exists m \in \mathbb{N}, \text{ такой что } |\phi^{(k)}(x)| \lesssim |x|^m \text{ при } |x| \to \infty, \ \forall k = 0, \ldots, s\big\}.$$

В следующем следствии мы рассмотрим случай, когда плотность π убывает экспоненциально.

Следствие 4. *Пусть $\pi(x) \propto e^{-c|x|^q}$ для некоторых $c \in \mathbb{R}$ и $q \in \mathbb{N}$. Пусть также $f \in C^s_{poly}(\mathbb{R})$ для некоторого $s \in \mathbb{N}$. Зафиксируем произвольное $\Phi \subset C^{s+1}_{poly}(\mathbb{R})$, такое что соответствующий класс $\mathcal{H} = \{f - g_\phi : \ \phi \in \Phi\}$ является ограниченным, содержит константную функцию и $\sup_{h\in\mathcal{H}} \sup_{x\in\mathsf{X}} |h(x)| \le b$ для некоторого $b > 0$. Тогда для некоторого $\varepsilon > 0$ и любого $\delta \in (0, 1)$, с вероятностью не менее $1 - \delta$,*

$$V\Big(\widehat{h}_{\varepsilon,1}\Big) \lesssim \left(\frac{1}{n}\right)^{\frac{1}{1+1/s}} + \frac{\log(1/\delta)}{n},$$

где символ $\lesssim$ означает неравенство с точностью до некоторой константы, не зависящей от n, но, возможно, зависящей от других параметров задачи.

Рассмотрим теперь случай, когда плотность π убывает полиномиально. Для этого сначала обозначим класс функций из $C^s(\mathbb{R})$, производные которых растут не быстрее полинома порядка $m \in \mathbb{N}$, через

$$C^s_{poly<m}(\mathbb{R}) := \big\{\phi \in C^s(\mathbb{R}) : \ |\phi^{(k)}(x)| \lesssim |x|^m \text{ при } |x| \to \infty, \forall k = 0, \ldots, s\big\}.$$

Следствие 5. *Пусть $\pi(x) \propto (1+x^2)^{-q}$ для $q \in \mathbb{N}$. Пусть также $f \in C^s_{poly<m}(\mathbb{R})$ для некоторых $s, m \in \mathbb{N}$. Зафиксируем произвольное $\Phi \subset C^{s+1}_{poly}(\mathbb{R})$, такое что соответствующий класс $\mathcal{H} = \{f - g_\phi : \ \phi \in \Phi\}$ является ограниченным, содержит константную функцию и $\sup_{h\in\mathcal{H}} \sup_{x\in\mathsf{X}} |h(x)| \leq b$ для некоторого $b > 0$. Тогда для некоторого $\varepsilon > 0$ и любого $\delta \in (0,1)$, с вероятностью не менее $1-\delta$,*

$$V\Big(\widehat{h}_{\varepsilon,1}\Big) \lesssim \begin{cases} \left(\dfrac{1}{n}\right)^{\frac{1}{1+1/s}} + \dfrac{\log(1/\delta)}{n} & \text{при } s < 2q - m + 1/p - 3, \\ \left(\dfrac{1}{n}\right)^{\frac{1}{1+(2q-m+1/p-3)^{-1}}} + \dfrac{\log(1/\delta)}{n} & \text{при } s > 2q - m + 1/p - 3, \end{cases}$$

где символ $\lesssim$ означает неравенство с точностью до некоторой константы, не зависящей от n, но, возможно, зависящей от других параметров задачи.

Следствие 4 и Следствие 5 демонстрируют интересный эффект. Если хвосты π убывают экспоненциально и производные f растут не быстрее некоторого полинома, то чем более гладкой является f, тем более быстрые порядки мы получаем. Если же убывание π и рост f являются полиномиальными, то существует критическая точка для гладкости f, после которой порядки сходимости не улучшаются.

3. Снижение дисперсии в зависимом случае

В этом разделе мы сформулируем аналоги Теоремы 2 и Теоремы 3 в зависимом случае. Пусть $\{X_k\}_{k=0}^\infty$ будет марковской цепью (в этом разделе мы нумеруем последовательность для удобства с 0) с переходным ядром P и начальным значением $X_0 = x_0$ для некоторого $x_0 \in \mathsf{X}$. Мы будем предполагать, что марковское ядро P имеет единственную инвариантную меру π.

Как упоминалось во Введении, известно несколько оценок для асимптотической дисперсии $V^\infty(h)$, см. работу Д. Флегала и Д. Джонса [FJ10]. Для простоты изложения мы рассмотрим только спектральную оценку, так как она является наиболее общей. Тем не менее, результаты, аналогичные приведенным ниже, можно получить и для других оценок. Начнем с определения. Пусть $\mathcal{H} = \{f - g : \ g \in \mathcal{G}\}$, где $\mathcal{G}$ — класс контрольных функций, см. Раздел 1. Обозначим автоковариационную функцию процесса $\{h(X_k)\}_{k=0}^\infty$ через

$$\rho(s) := \mathrm{E}\big[(h(X_0) - \pi(h))(h(X_k) - \pi(h))\big],$$

где предполагается, что $X_0 \sim \pi$. Обозначим также выборочную автоковариационную функцию через

$$\hat{\rho}_n(s) := \frac{1}{n} \sum_{k=0}^{n-s-1} \big(h(X_k) - \pi_n(h)\big)\big(h(X_{k+s}) - \pi_n(h)\big).$$

Легко видеть, что асимптотическую дисперсию $V^{\infty}(h)$, определенную в (7), можно переписать как

$$V^{\infty}(h) = \sum_{s=-\infty}^{+\infty} \rho(|s|).$$

Спектральная оценка асимптотической дисперсии основана на взвешенных выборочных автоковариационных функциях:

$$V_n^{\infty}(h) := \sum_{s=-(b_n-1)}^{b_n-1} w_n(s)\hat{\rho}_n(|s|), \tag{10}$$

где w_n — скользящее окно, а b_n — ширина окна. Ширина окна b_n при любом значении n является натуральным числом, а скользящее окно w_n является ядром вида $w_n(s) = w(s/b_n)$, где w — симметричная неотрицательная функция с носителем $[-1, 1]$, равная

$$w(s) = \begin{cases} 2s+2, & -1 \le s < -1/2, \\ 1, & -1/2 \le s \le 1/2, \\ -2s+2, & 1/2 < s \le 1. \end{cases}$$

Возможен и другой выбор $w(s)$, см. подробнее в [FJ10]. В данном разделе мы будем изучать свойства оценки

$$\widehat{h}_{\varepsilon,2} := \operatorname*{argmin}_{h \in \mathcal{H}_{\varepsilon,2}} V_n^{\infty}(h),$$

где $\mathcal{H}_{\varepsilon,2}$ — ε-сеть множества $\mathcal{H}$ в норме $L^2(\pi)$, см. подробные определения в Разделе 1.

Напомним, что для двух марковских ядер P и Q на $\mathsf{X} \times \mathcal{X}$, где $\mathcal{X}$ — борелевская σ-алгебра на X, их произведение тоже является марковским ядром и определяется для $x \in \mathsf{X}$ и $A \in \mathcal{X}$ как $PQ(x, A) = \int P(x, \mathrm{d}y)Q(y, A)$. Кроме того, ядро P^n определяется рекуррентным соотношением $P^n = PP^{n-1}$. Пусть теперь $W : \mathsf{X} \to [1, \infty)$ — измеримая функция. W-норма произвольной функции $h : \mathsf{X} \to \mathbb{R}$ определяется как $\|h\|_W = \sup_{x \in \mathsf{X}} \{|h(x)|/W(x)\}$. Для двух вероятностных мер μ и ν на $(\mathsf{X}, \mathcal{X})$, удовлетворяющих $\mu(W) < \infty$ и $\nu(W) < \infty$, W-нормой разности $\mu - \nu$ является по определению $\|\mu - \nu\|_W = \sup_{\|h\|_W \le 1} |\mu(h) - \nu(h)|$.

Перейдем теперь к предположениям, необходимым для формулировки основных результатов раздела. Нашим первым предположением будет геометрическая эргодичность $\{X_k\}_{k=0}^{\infty}$.

(GE) Марковское ядро P имеет единственную инвариантную меру π. Кроме того, для некоторой измеримой функции $W : \mathsf{X} \to [1, \infty)$, такой что $\pi(W) < \infty$, и некоторых чисел $\varsigma > 0$, $\rho \in (0, 1)$ выполняется для всех $x \in \mathsf{X}$ и всех $n \in \mathbb{N}$

$$\|P^n(x, \cdot) - \pi\|_W \leq \varsigma W(x) \rho^n.$$

Известно, что условие (GE) влечет центральную предельную теорему для последовательности $\{h(X_k)\}_{k=0}^{\infty}$, $h \in \mathcal{H}$, если дополнительно $\pi(h^{2+\kappa}) < \infty$ для некоторого $\kappa > 0$, см. работу И. А. Ибрагимова и Ю. В. Линника [IL71] или Д. Джонса [J04]. Чтобы асимптотический доверительный интервал (6) был корректно определен, далее мы будем предполагать, что $\mathcal{H} \subset L^{2+\kappa}$ для некоторого $\kappa > 0$.

Напомним, что L^2-метрикой Канторовича (в зарубежной литературе — L^2-метрикой Васерштейна) между вероятностными мерами μ и ν называется величина

$$W_2(\mu, \nu) = \inf_{\zeta} \left(\int_{\mathsf{X} \times \mathsf{X}} \|x - y\|^2 \, \mathrm{d}\zeta(x, y) \right)^{1/2},$$

где $\|\cdot\|$ — евклидова норма в $\mathbb{R}^d$, а инфимум взят по всем вероятностным мерам ζ на декартовом произведении $\mathsf{X} \times \mathsf{X}$, у которых маргинальными мерами являются μ и ν соответственно. Расстояние Кульбака-Лейблера между мерами μ и ν определяется как

$$\mathrm{KL}(\mu \| \nu) = \begin{cases} \int \log(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}) \mathrm{d}\mu, & \text{если } \mu \ll \nu, \\ \infty, & \text{иначе.} \end{cases}$$

Мы будем говорить, что вероятностная мера μ удовлетворяет (информационному) транспортному неравенству $\mathrm{T}_2(C)$, если существует $C > 0$, такая что для любой вероятностной меры ν выполняется

$$W_2(\mu, \nu) \leq \sqrt{2C \, \mathrm{KL}(\mu \| \nu)}.$$

Для доказательства основных результатов нам необходима концентрация величины $V_n^{\infty}(h)$. Важно отметить, что $V_n^{\infty}(h)$ является квадратичной формой по $\{h(X_j)\}_{j=0}^{n-1}$. Нами была доказана гауссовская концентрация $V_n^{\infty}(h)$ в случае, когда функций $h \in \mathcal{H}$ являются липшицевыми, а марковское ядро P является сжимающим отображением в L^2-метрике Канторовича.

(L) Функции $h \in \mathcal{H}$ являются L-липшицевыми, то есть $|h(x) - h(y)| \leq L\|x - y\|$ для некоторого $L > 0$ и всех $x, y \in \mathsf{X}$.

(CW) Марковское ядро $P(x, \cdot)$ удовлетворяет $\mathrm{T}_2(C)$ для всех $x \in \mathsf{X}$ и некоторого $C > 0$. И, кроме того, существует $r \in (0, 1)$, такое что $W_2(P(x, \cdot), P(y, \cdot)) \leq r\|x - y\|$ для всех $x, y \in \mathsf{X}$.

Теперь мы можем сформулировать следующую теорему.

Теорема 6. *Предположим **(GE)**, **(L)** и **(CW)**. Положим $b_n = 2\log(n)/\log(1/\rho)$. Тогда для некоторого $\varepsilon > 0$ (значение которого определено в доказательстве), любого начального значения $x_0 \in X$ и любого $\delta \in (0,1)$, с вероятностью не менее $1-\delta$,*

$$V^{\infty}\Big(\widehat{h}_{\varepsilon,2}\Big) - \inf_{h\in\mathcal{H}} V^{\infty}\big(h\big) \lesssim A_1 \log(n)\gamma_{\mathrm{L}^2(\pi)}(\mathcal{H},n) + A_2 \frac{\log(n)\log(1/\delta)}{\sqrt{n}},$$

где первая асимптотическая дисперсия является условной по выборке $X_1,\ldots,X_n$, на которой вычислялась оценка $\widehat{h}_{\varepsilon,1}$, символ $\lesssim$ означает неравенство с точностью до некоторой абсолютной константы и

$$A_1 = \frac{\sqrt{C}L^2}{(1-r)\log(1/\rho)}, \quad A_2 = \frac{\sqrt{\varsigma}(\pi(W)+W(x_0))}{\sqrt{1-\rho}\log(1/\rho)}\left(\frac{\sqrt{C}L^2}{1-r} + \sup_{h\in\mathcal{H}}\|h\|^2_{W^{1/2}}\right).$$

Предполагая дополнительно к условиям Теоремы 6, что $\mathcal{H}$ является ограниченным подмножеством некоторого взвешенного соболевского пространства $W^{s,p}(\mathsf{X},\langle x\rangle^{\beta})$, и подставляя результат Предложения 1, мы получим окончательную оценку с большой вероятностью вида:

$$V^{\infty}\Big(\widehat{h}_{\varepsilon,2}\Big) - \inf_{h\in\mathcal{H}} V^{\infty}\big(h\big) \lesssim n^{-\alpha}, \quad \alpha \in (0,1/2),$$

где значение параметра α будет зависеть от размерности d и параметров соболевского пространства s, p и β. В статистической литературе такие порядки сходимости называются «медленными».

Порядок сходимости может быть улучшен до $n^{-\alpha}$ для $\alpha \in (0,1)$, если дополнительно предположить, что $\mathcal{H}$ содержит константную функцию. Так как $\pi(h) = \pi(f)$ для всех $h \in \mathcal{H}$, эта константа должна быть равна $\pi(f)$ и, следовательно, в этом случае $\inf_{h\in\mathcal{H}} V^{\infty}(h) = 0$.

Теорема 7. *Предположим **(GE)**, **(L)** и **(CW)**. Предположим также, что $\mathcal{H}$ содержит константную функцию, то есть $h^*(x) \equiv c$ для некоторой $h^* \in \mathcal{H}$ и некоторого $c \in \mathbb{R}$. Положим $b_n = 2\log(n)/\log(1/\rho)$. Тогда для некоторого $\varepsilon > 0$ (значение которого определено в доказательстве), любого начального значения $x_0 \in X$ и любого $\delta \in (0,1)$, с вероятностью не менее $1-\delta$,*

$$V^{\infty}\Big(\widehat{h}_{\varepsilon,2}\Big) \lesssim A_1 \log(n)\big(\gamma_{\mathrm{L}^2(\pi)}(\mathcal{H},n)\big)^2 + A_2 \frac{\log(n)\log(1/\delta)}{n},$$

где асимптотическая дисперсия является условной по выборке $X_1,\ldots,X_n$, на которой вычислялась оценка $\widehat{h}_{\varepsilon,1}$, символ $\lesssim$ означает неравенство с точностью до некоторой абсолютной константы и

$$A_1 = \frac{CL^2}{(1-r)^2\log(1/\rho)}, \quad A_2 = \frac{CL^2}{(1-r)^2\log(1/\rho)} + \frac{\varsigma(\pi(W)+W(x_0))}{(1-\rho)^{1/2}\log(1/\rho)} \sup_{h\in\mathcal{H}}\|h\|^2_{W^{1/2}}.$$

Предполагая теперь, что $\mathcal{H}$ является ограниченным подмножеством $W^{s,p}(\mathsf{X}, \langle x \rangle^{\beta})$, мы получим оценку с большой вероятностью вида:

$$V^{\infty}\Big(\widehat{h}_{\varepsilon,2}\Big) \lesssim n^{-\alpha}, \quad \alpha \in (0,1).$$

В свою очередь, порядок длины асимптотического доверительного интервала (6) с помощью данной процедуры снижения дисперсии можно уменьшить с $n^{-1/2}$ до $n^{-1/2-\alpha/2}$.

В доказательстве Теоремы 6 и Теоремы 7 использовалось следующее неравенство концентрации, которое, как кажется автору, представляет самостоятельный интерес.

Предложение 8. *Пусть $\{X_k\}_{k=0}^{\infty}$ является марковской цепью с переходным ядром P и начальным значением $X_0 = x_0$ для некоторого $x_0 \in \mathsf{X}$. Предположим, что существует $C > 0$, такое что $P(x,\cdot)$ удовлетворяет $\mathrm{T}_2(C)$ для всех $x \in \mathsf{X}$. Предположим также, что существует число $r \in (0,1)$, такое что для всех $x, y \in \mathsf{X}$ выполняется*

$$W_2\big(P(x,\cdot), P(y,\cdot)\big) \le r\|x-y\|.$$

Для произвольной L-липшицевой функции $h : \mathsf{X} \to \mathbb{R}$ обозначим $Z_n(h) = (h(X_0), \ldots, h(X_{n-1}))^{\top}$. Тогда для любой матрицы A размера $n \times n$ и любого $t > 0$,

$$\mathrm{P}\Big(\big|Z_n(h)^{\top} A Z_n(h) - \mathrm{E}\big[Z_n(h)^{\top} A Z_n(h)\big]\big| > t\Big) \le 2\exp\left(-\frac{t^2}{cK^2\left(\mathrm{E}\|AZ_n(h)\|^2 + t\|A\|\right)}\right),$$

где $K^2 = CL^2/(1-r)^2$, а $c > 0$ — некоторая абсолютная константа.

Доказательство данного утверждения состоит из двух этапов. Сначала, используя результат из статьи Х. Джеллоута, А. Гийена и Л. Ву [DGW04], мы показываем, что совместное распределение $\{X_k\}_{k=0}^{n-1}$ принадлежит классу $\mathrm{T}_2(C/(1-r)^2)$, что, в свою очередь, влечет гауссовскую концентрацию для всех липшицевых функций. Далее, используя технику из работы Р. Адамчака [A15], мы доказываем концентрацию для квадратичных форм.

4. Алгоритм Ланжевена

Простым примером MCMC алгоритма, для которого выполняются условия (GE) и (CW) является алгоритм Ланжевена (в английской литературе Unadjusted Langevin Algorithm или ULA). Пусть плотность $\pi(x)$ известна с точностью до нормировочной константы и выражается в виде $\pi(x) \propto \mathrm{e}^{-U(x)}$ для некоторой неотрицательной функции $U(x)$, которую принято называть потенциалом. В алгоритме Ланжевена марковская цепь $\{X_k\}_{k\ge 0}$ строится с помощью следующего рекуррентного соотношения

$$X_{k+1} = X_k - \gamma \nabla U(X_k) + \sqrt{2\gamma}\, Z_{k+1}, \tag{11}$$

где $\{Z_k\}_{k\geq 1}$ — последовательность независимых стандартных нормальных векторов в $\mathbb{R}^d$, а $\gamma > 0$ — шаг алгоритма. Данный алгоритм является дискретизацией первого порядка (по схеме Эйлера-Маруямы) стохастического дифференциального уравнения Ланжевена:

$$\mathrm{d}Y_t = -\nabla U(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t,$$

где $\{B_t\}_{t\geq 0}$ — d-мерное броуновское движение. При некоторых условиях, единственным стационарным распределением динамики Ланжевена $\{Y_t\}_{t\geq 0}$ является π, а марковская цепь $\{X_k\}_{k\geq 0}$ сходится к единственному стационарному распределению π_γ, которое близко к распределению π в некоторых метриках, таких как расстояние по вариации или L^2-метрике Канторовича. Подробное описание алгоритма и теоремы сходимости доступны в работах Г. Робертса и Р. Твиди [RT96], А. Далаляна [D17], А. Дурмуса и Э. Мулине [DM17]. В случае, когда плотность $\pi(x)$ не является известной, для использования алгоритма Ланжевена можно воспользоваться методами из статей Д. В. Беломестного и автора [BI19] и [BI20].

Рассмотрим следующие условия на потенциал $U(x)$, которые будут гарантировать выполнение условий (GE) и (CW):

(U1) Потенциал U является дважды непрерывно дифференцируемым $U \in C^2(\mathbb{R}^d)$ и имеет липшицевый градиент, то есть существует $L > 0$, такое что для всех $x, y \in \mathbb{R}^d$

$$\|\nabla U(x) - \nabla U(y)\| \leq L\|x - y\|.$$

(U2) Потенциал U является строго выпуклым, то есть существует константа $m > 0$, такая что для всех $x, y \in \mathbb{R}^d$ выполняется

$$U(y) \geq U(x) + \langle \nabla U(x), y - x\rangle + \frac{m}{2}\|x - y\|^2.$$

Предложение 9. *Пусть выполнены условия (U1) и (U2). Тогда при любом* $\gamma \in (0, 2/(m+L))$, *марковское ядро* P_γ, *соответствующее цепи* $\{X_k\}_{k\geq 0}$ *из* (11), *удовлетворяет условиям (GE) и (CW) при некотором* $\rho \in (0,1)$,

$$W(x) = \|x\|^2, \quad C = 2\gamma \quad \text{и} \quad r = \sqrt{1 - 2\gamma mL/(m+L)}.$$

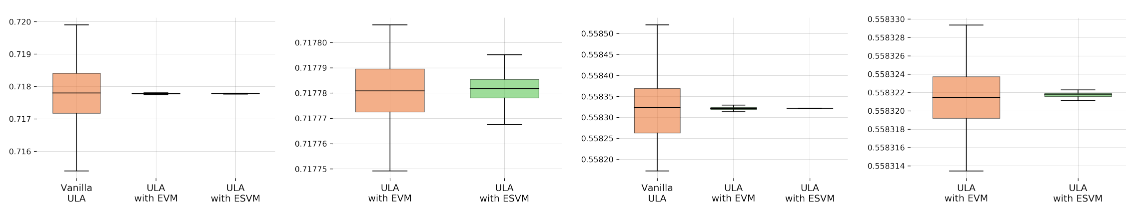
Данный факт является непосредственным следствием результатов из двух работ А. Дурмуса и Э. Мулине [DM17] и [DM19].

Байесовская логистическая регрессия. Рассмотрим также численный пример для алгоритма Ланжевена. Мы сравним два описанных метода выбора контрольных функций: минимизацию эмпирической дисперсии (EVM метод), см. (4), и минимизацию спектральной дисперсии (ESVM метод), см. (10). Сравнение будем проводить на двух наборах данных для задачи классификации из базы данных UCI. Первый набор данных называется Pima и содержит $N = 768$ наблюдений в размерности $d = 9$. Второй набор данных называется EEG и содержит $N = 14\,980$ наблюдений в размерности $d = 15$.

В логистической регрессии вероятность i-ого ответа $y_i \in \{-1, 1\}$ для $i = 1, \dots, N$, вычисляется как $\mathrm{p}(y_i|\mathbf{x}_i, \theta) = (1 + \mathrm{e}^{-y_i\langle\theta,\mathbf{x}_i\rangle})^{-1}$, где $\mathbf{x}_i \in \mathbb{R}^d$ — это вектор признаков (факторов) и $\theta \in \mathbb{R}^d$ — вектор неизвестных регрессионных коэффициентов. Сначала мы разобьем данные на обучающую выборку $\mathcal{T}^{\text{train}} = \{(y_i, \mathbf{x}_i)\}_{i=1}^{N-K}$ и тестовую выборку $\mathcal{T}^{\text{test}} = \{(y'_i, \mathbf{x}'_i)\}_{i=1}^{K}$, выбрав случайно $K = 100$ наблюдений из данных (и удалив их из обучающей выборки). Далее, мы наложим априорное стандартное нормальное распределение на параметр θ, обозначим его $\mathrm{p}_0(\theta)$, и используем алгоритм ULA для получения выборки из апостериорного распределения $\mathrm{p}(\theta|\mathcal{T}^{\text{train}}) \propto \mathrm{p}_0(\theta)\prod_{(y_i,\mathbf{x}_i)\in\mathcal{T}^{\text{train}}} \mathrm{p}(y_i|\mathbf{x}_i, \theta)$. С помощью полученной выборки $\{\theta_k\}_{k=0}^{n-1}$ мы оценим предсказание для фиксированной точки из тестовой выборки $(y', \mathbf{x}')$, то есть $\mathrm{p}(y'|\mathbf{x}') = \int_{\mathbb{R}^d} \mathrm{p}(y'|\mathbf{x}', \theta)\,\mathrm{p}(\theta|\mathcal{T}^{\text{train}})\,\mathrm{d}\theta$, вычислив оценку $n^{-1}\sum_{k=0}^{n-1} f(\theta_k)$ для $f(\theta) = \mathrm{p}(y'|\mathbf{x}', \theta)$. Чтобы избавиться от случайности, мы также оценим среднее предсказаний по всему тестовому набору $\mathcal{T}^{\text{test}}$, вычислив оценку для $f(\theta) = K^{-1}\sum_{(y_i,\mathbf{x}_i)\in\mathcal{T}^{\text{test}}} \mathrm{p}(y'_i|\mathbf{x}'_i, \theta)$. Кроме того, аналогичные оценки мы посчитаем для EVM и ESVM методов, в которых мы будем использовать контрольные функции Стейна, см. (3), где $\Phi(x) \equiv b$ для некоторого $b \in \mathbb{R}^d$.

Графики с разбросом полученных оценок даны на Рис. 1. Обратим внимание, что оба метода EVM и ESVM дают значительный выигрыш в дисперсии, но метод ESVM работает лучше, так как учитывает марковскую структуру цепи.

Рис. 1: Оценивание среднего предсказания в байесовской логистической регрессии. Первые два графика относятся к набору данных Pima, а последние два — к набору данных EEG. В первых графиках для каждого набора данных отражены 3 «ящика с усами» для 100 оценок полученных соответсвенно обычным алгоритмом ULA, алгоритмом ULA с EVM и алгоритмом ULA с ESVM, а следующий график — увеличение масштаба для последних двух алгоритмов.



Основные результаты диссертационной работы

В рамках диссертационной работы получены следующие основные результаты:

1) Неасимптотические оценки на избыточный риск (5) в независимом случае для произвольного класса контрольных функций $\mathcal{G}$. Данные результаты сформулированы в Теореме 2 и Теореме 3. Кроме того, продемонстрировано, какие получаются оценки в случае контрольных функций Стейна, см. Следствие 4 и Следствие 5.

2) Неасимптотические оценки на избыточный риск (8) в зависимом случае для произвольного класса контрольных функций $\mathcal{G}$. Данные результаты сформулированы в Теореме 6 и Теореме 7. Кроме того, в данной работе впервые было предложено минимизировать вместо эмпирической дисперсии (4) спектральной оценку асимптотической дисперсии (10). Данное решение лучше работает на практике в случае зависимых случайных величин.

3) Концентрационное неравенство для квадратичной формы от марковской цепи, когда ее марковское ядро является сжимающим отображением в L^2-метрике Канторовича, см. Предложение 8.

Данные результаты имеют теоретический и практический характер. Методы, развитые в представленной работе, могут быть использованы для дальнейшего исследования алгоритмов снижения дисперсии оценок Монте-Карло. Кроме этого, полученные результаты могут быть полезны в рамках исследования ряда прикладных задач байесовской статистики.

Личный вклад автора

Содержание диссертации и основные результаты, перечисленные в списке выше, отражают персональный вклад автора в опубликованные работы. В статьях с соавторами вклад диссертанта был определяющим во всех перечисленных выше результатах за исключением результата Предложения 8. Утверждения, не перечисленные в списке выше, но присутствующие в настоящем тексте, являются либо простыми следствиями результатов, не принадлежащих автору и его соавторам, либо были получены соавторами и приведены здесь лишь для полноты изложения.

Апробация результатов

Результаты диссертационной работы докладывались на следующих конференциях и научно-исследовательских семинарах:

- Конференция «Информационные технологии и системы» в ИППИ РАН, Санкт-Петербург, сентябрь 2016 года. Тема доклада: «Концентрация нормы изотропного логарифмически вогнутого случайного вектора».

- Аспирантский коллоквиум кафедры теории вероятностей в МГУ, Москва, декабрь 2017 года. Тема доклада: «Снижение дисперсии оценки Монте-Карло».

- Конференция «New frontiers in high-dimensional probability and statistics» в НИУ ВШЭ, Москва, февраль 2018 года. Тема доклада: «Снижение дисперсии оценок Монте-Карло с помощью минимизации выборочной дисперсии».

- Конференция «Ломоносов» в МГУ, Москва, апрель 2018 года. Тема доклада: «Снижение дисперсии оценок Монте-Карло с помощью минимизации выборочной дисперсии».

- Городской семинар по теории вероятностей и математической статистике в ПОМИ РАН, Санкт-Петербург, сентябрь 2018 года. Тема доклада: «Снижение дисперсии оценок Монте-Карло с помощью минимизации выборочной дисперсии».

- Зимняя школа «New frontiers in high-dimensional probability and statistics 2» в НИУ ВШЭ, Москва, февраль 2019 года. Тема доклада: «MCMC алгоритмы для распределений с известной характеристической функцией».

- Семинар центра прикладной математики в Политехнической школе (École polytechnique), Париж, июнь 2019 года. Тема доклада: «Снижение дисперсии оценок Монте-Карло с помощью минимизации выборочной дисперсии».

- Конференция «European Meeting of Statisticians» (EMS), проводимая обществом Бернулли по математической статистике и вероятности, Палермо, июль 2019 года. Тема доклада: «Снижение дисперсии для марковских цепей с помощью минимизации выборочной дисперсии».

- Конференция «Structural Inference in High-Dimensional Models 2», Санкт-Петербург, сентябрь 2019 года. Тема доклада: «MCMC алгоритмы для распределений с тяжелыми хвостами».

- Третья Санкт-Петербургская зимняя молодежная конференция по теории вероятностей и математической физике в ПОМИ РАН, Санкт-Петербург, декабрь 2019 года. Тема доклада: «Снижение дисперсии для зависимых последовательностей».

Публикации

Результаты диссертационной работы составляют содержание трех опубликованных статьей.

[1] Беломестный Д., Иосипой Л., Животовский Н. *Снижение дисперсии оценки Монте-Карло методом минимизации эмпирической дисперсии.* Doklady Mathematics, 98:2, 494–497, 2018.

[2] Беломестный Д., Иосипой Л. *Об оценке плотности распределения с помощью ряда Фурье.* Large-Scale Systems Control, 82, 28–43, 2019.

[3] Беломестный Д., Иосипой Л., Мулине Э., Наумов А., Самсонов С. *Снижение дисперсии для марковских цепей с помощью минимизации эмпирической дисперсии с приложением к MCMC алгоритмам.* Statistics and Computing, 30:4, 973–997, 2020.

Список литературы

[A15] R. Adamczak. *A note on the Hanson-Wright inequality for random vectors with dependencies.* Electron. Commun. Probab., 20:71, 1–13, 2015.

[B71] N. S. Bakhvalov. *On the approximate calculation of multiple integrals.* Vestnik MGU, Ser. Math. Mech. Astron. Phys. Chem., 4, 3–18, 1959.

[BI19] D. Belomestny, L. Iosipoi. *On density estimation via Fourier series.* Large-Scale Systems Control, 82, 28–43, 2019.

[BI20] D. Belomestny, L. Iosipoi. *Fourier transform MCMC, heavy tailed distributions, and geometric ergodicity.* ArXiv preprint, arXiv:1909.00698, 2020.

[CLV08] S. Clémençon, G. Lugosi, N. Vayatis. *Ranking and Empirical Minimization of U-statistics.* Ann. Statist., 36:2, 844–874, 2008.

[D17] A. Dalalyan. *Theoretical guarantees for approximate sampling from smooth and log-concave densities.* Journal of the Royal Statistical Society Series B (Statistical Methodology), 79:3, 651–676, 2017.

[DGL96] L. Devroye, L. Györfi, G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer, New York, 1996.

[DGW04] H. Djellout, A. Guillin, L. Wu. *Transportation cost-information inequalities and applications to random dynamical systems and diffusions. Ann. Probab.*, 32:3B, 2702–2732, 2004.

[D08] I. Dimov. *Monte Carlo methods for applied scientists.* World Scientific, 2008.

[DM17] A. Durmus, É. Moulines. *Non-asymptotic convergence analysis for the unadjusted Langevin algorithm.* Ann. Appl. Probab., 27:3, 1551–1587, 2017.

[DM19] A. Durmus, É. Moulines. *High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm.* Bernoulli, 4A, 2854–2882, 2019.

[FJ10] J. Flegal, G. Jones. *Batch means and spectral variance estimators in Markov chain Monte Carlo.* Ann. Statist., 38:2, 1034–1070, 2010.

[G00] S. van de Geer. *Empirical Processes in M-Estimation.* Cambridge, 2000.

[G13] P. Glasserman. *Monte Carlo methods in financial engineering.* Springer Science & Business Media, 2013.

[IL71] I. A. Ibragimov, Y. V. Linnik. *Independent and Stationary Sequences of Random Variables.* Wolters-Noordhoff, Groningen, 1971.

[J04] G. Jones. *On the Markov chain central limit theorem.* Probab. Surveys, 1, 299–320, 2004.

[MSI13] A. Mira, R. Solgi, D. Imparato. *Zero variance Markov chain Monte Carlo for Bayesian Estimators.* Statistics and Computing, 23:5, 653–662, 2013.

[NP07] R. Nickl, B. Pötscher. *Bracketing Metric Entropy Rates and Empirical Central Limit Theorems for Function Classes of Besov- and Sobolev-Type.* Journal of Theoretical Probability, 20:2, 177–199, 2007.

[N16] E. Novak. *Some Results on the Complexity of Numerical Integration.* Monte Carlo and Quasi-Monte Carlo Methods. Springer Proceedings in Mathematics & Statistics, 163, 161–183, 2016.

[OGC17] C. Oates, M. Girolami, N. Chopin. *Control functionals for Monte Carlo integration.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79:3, 695–718, 2017.

[OGC19] C. Oates, M. Girolami, N. Chopin. *Convergence rates for a class of estimators based on Stein's identity.* Bernoulli, 25:2, 1141–1159, 2019.

[RC99] C. Robert, G. Casella. *Monte Carlo Statistical Methods.* Springer, New York, 1999.

[RT96] G. Roberts, R. Tweedie. *Exponential convergence of Langevin distributions and theirdiscrete approximations.* Bernoulli, 2:4, 341–363, 1996.

[RR04] G. Roberts, J. Rosenthal. *General state space Markov chains and MCMC algorithms.* Probab. Surveys, 1, 20–71, 2004.

[RK16] R. Rubinstein, D. Kroese. *Simulation and the Monte Carlo method.* John Wiley & Sons, 2016.

[WS95] W. H. Wong, X. Shen. *Probability Inequalities for Likelihood Ratios and Convergence Rates of Sieve MLES.* Ann. Statist., 23:2, 339–362, 1995.