Федеральное государственное автономное образовательное
учреждение высшего образования
“Национальный исследовательский университет
“Высшая школа экономики”

Факультет математики

На правах рукописи

Самсонов Сергей Владимирович

Неравенства концентрации для функционалов от цепей Маркова и их приложения к снижению дисперсии MCMC алгоритмов

Резюме диссертации
на соискание ученой степени
кандидата математических наук

Научный руководитель
д.к.н.
Наумов Алексей Александрович

Москва – 2024

Оглавление

Введение

С ростом интереса к разработке новых методов машинного обучения возрастает также интерес к математическому аппарату, который предоставляет возможность анализа и оценки эффективности алгоритмов с учетом конечного объема доступных данных, например, конечного размера обучающей выборки. Результаты, основанные на явлении концентрации меры [1, 2], позволили производить неасимптотический анализ производительности различных алгоритмов в областях обучения с подкреплением (RL) [3], оптимизации [4], статистической теории обучения [5], методов Монте-Карло и Монте-Карло по схеме марковской цепи (MCMC, [6, 7]), и многих других. При этом неравенства концентрации для функций от независимых случайных величин и мартингалов сравнительно хорошо изучены, см., например, [2, 8, 9]. В то же время, иная ситуация складывается при изучении неравенств концентрации для функций от зависимых случайных величин. Хотя существует немало результатов для случайных процессов, удовлетворяющих различным типам условий перемешивания и слабой зависимости [10, 11], их применение даже к исследованию свойств аддитивных функционалов от цепей Маркова может оказаться затруднительным. В частности, они либо содержат неявные константы, либо недостаточно точны в отношении важных характеристик задачи, таких как дисперсия аддитивного функционала в неравенствах типа Бернштейна (см. соответствующие определения в Главе 1). Этот недостаток присущ многим существующим неравенствам концентрации, полученным для функционалов от цепей Маркова [12–15]. В то же время, именно такой тип зависимости возникает в большинстве алгоритмов машинного обучения. Особенно естественно марковские цепи возникают при неасимптотическом анализе алгоритмов стохастической аппроксимации [16, 17] или обучения с подкреплением [3, 18].

В Главе 1 данной диссертации получены новые аналоги классических неравенств Розенталя и Бернштейна для геометрически эргодических цепей Маркова с явной зависимостью от времени перемешивания соответствующих цепей. В данной главе рассматривается аддитивный функционал

$$S_n = \sum_{\ell=0}^{n-1} \{g(X_\ell) - \pi(g)\}, \tag{1}$$

где g — интегрируемая измеримая функция, а $(X_\ell)_{\ell=0}^{\infty}$ — цепь Маркова с марковским ядром P с единственным инвариантным распределением π. Для аддитивного функционала S_n получены неравенства концентрации, аналогичные приведенным в [12, 14, 19, 20]. При этом уточнена зависимость новых оценок от дисперсии S_n и времени перемешивания цепи. Предлагаемый метод доказательства основан на технике кумулянтного разложения, описанной в [21], и формуле Леонова-Ширяева [22], связывающей моменты и кумулянты.

В последующих частях диссертации рассматриваются применения неравенств концентрации для неасимптотического анализа производительности методов снижения дисперсии [23, 24], а также предлагаются новые методы снижения дисперсии для последовательностей зависимых случайных величин. Основная цель подобных методов заключается в уменьшении стохастической ошибки в оценках, получаемых с помощью метода Монте-Карло. Классические работы в этой области, к примеру [25] и [26], широко исследовали методы снижения дисперсии, основанные на использовании последовательностей независимых и одинаково

распределенных случайных величин [27]. Однако во многих практических задачах генерация независимых наблюдений невозможна, особенно в случае задач в большой размерности. В этом случае методы снижения дисперсии должны применяться к зависимым наблюдениям. Зачастую эти наблюдения формируют цепь Маркова, как, например, в случае алгоритмов MCMC [6]. Методы снижения дисперсии могут быть также применены в оптимизации и обучении с подкреплением, см. работы [28–31].

В Главе 2 предложен практический подход к снижению дисперсии для аддитивных функционалов от зависимых случайных величин. Этот подход обобщает предложенный в [32], и применим к более широкому классу цепей Маркова, удовлетворяющих условию эргодичности в метрике Канторовича-Васерштейна 1-го порядка, а также к последовательностям зависимых случайных величин, удовлетворяющих условию стационарности ковариаций. Предлагаемый метод основан на использовании контрольных переменных и минимизации эмпирической оценки асимптотической дисперсии. Для данного метода получены оценки скорости убывания избыточной асимптотической дисперсии с ростом размера обучающей выборки. Предлагаемый подход применен к алгоритмам MCMC на основе динамики Ланжевена с использованием стохастических градиентов (SGLD, [33]).

В Главе 3 рассмотрена задача снижения дисперсии для аддитивных функционалов от цепей Маркова в случае, когда аналитическое выражение для инвариантного распределения данной цепи неизвестно. Для решения данной задачи предложен подход, основанный на дискретном мартингальном разложении и обобщающий контрольные переменные, использующие разложения на основе ортогональных полиномов [34]. Этот подход не требует знания стационарного распределения цепи или наличия у него специальной структуры. Работа алгоритма проанализирована в модели нормального шума (см. Секцию 3.4), которая покрывает неадаптированный алгоритм Ланжевена (Unadjusted Langevin Algorithm, ULA), ранее изученный, в частности, в работах [35–37].

Цели и задачи исследования

Целью исследования является получение новых аналитических инструментов для изучения свойств концентрации функционалов от цепей Маркова и их применение для теоретического анализа методов пост-обработки оценок MCMC, основанных на контрольных переменных. Для решения этой задачи планируется выполнение следующих этапов:

1. Получить верхние оценки на кумулянты аддитивных функционалов от геометрически эргодических цепей Маркова с явной зависимостью от параметров соответствующего марковского ядра;

2. Получить новые аналоги неравенства Розенталя и неравенства Бернштейна с использованием полученных ранее оценок на кумулянты. Неравенства должны сохранять точную зависимость от дисперсии аддитивного функционала S_n из (1) и времени перемешивания цепи;

3. Обобщить указанные версии неравенства Розенталя на случай квадратичных форм от

функций от цепей Маркова, сходящихся к инвариантному распределению с геометрической скоростью в метрике Канторовича-Васерштейна первого порядка;

4. Разработать метод выбора контрольных переменных для уточнения оценок MCMC, основанный на минимизации оценок асимптотической дисперсии. Изучить статистические свойства предложенного метода;

5. Разработать метод снижения дисперсии для аддитивных функционалов от цепей Маркова, который не требует аналитического знания инвариантного распределения соответствующей цепи. Оценить дисперсию модифицированных при помощи контрольных переменных статистик в модели нормального шума, описанной в Секции 3.4.

Научная новизна

Все основные результаты диссертационной работы являются новыми. Получены новые неравенства концентрации типа Розенталя и Бернштейна для аддитивных функционалов от цепей Маркова. Данные неравенства обобщают известные в литературе оценки, при этом впервые получен результат, обобщающий неравенства Бернштейна на марковские ядра с условием эргодичности во взвешенной метрике Канторовича-Васерштейна. Также проведен неасимптотический анализ производительности нескольких методов снижения дисперсии для алгоритмов MCMC, получены оценки убывания избыточной дисперсии с ростом размера обучающей выборки. Также предложен и проанализирован новый подход к построению контрольных переменных с использованием дискретных мартингальных разложений. Данный метод является новым и применим, в частности, в ситуациях, в которых классические методы построения контрольных переменных, например, на основе оператора Стейна, не могут быть применены напрямую.

Теоретическая и практическая значимость результатов

Представленные результаты имеют теоретический и методологический характер. Полученные теоретические результаты предлагают новые неравенства концентрации для аддитивных функционалов от цепей Маркова, которые могут быть полезны для исследования методов Монте-Карло по схеме марковской цепи (MCMC). С методологической точки зрения предложены новые методы снижения дисперсии для алгоритмов MCMC, которые могут быть напрямую применены, в частности, в задачах Байесовской статистики.

Методология и методы исследования

В работе широко применяется аналитический аппарат теории вероятностей, в частности, метод каплинга и метод кумулянтов. Доказательства основных результатов опираются на теорию цепей Маркова и неравенства концентрации меры.

Публикации по результатам исследования

Основные результаты диссертации были опубликованы в трех статьях [38–40] в рецензируемых научных изданиях. Все три статьи входят в реферативные базы Scopus и Web of

Science.

1. A.Durmus, E. Moulines, A. Naumov, S. Samsonov. *Probability and Moment Inequalities for Additive Functionals of Geometrically Ergodic Markov Chains*, Journal of Theoretical Probability, 2024. https://doi.org/10.1007/s10959-024-01315-7;

2. D. Belomestny, L. Iosipoi, E. Moulines, A. Naumov, S. Samsonov. *Variance reduction for dependent sequences with applications to stochastic gradient MCMC*, SIAM/ASA Journal on Uncertainty Quantification, 9(2), 507-535, 2021. https://doi.org/10.1137/19M1301199;

3. D. Belomestny, E., Moulines, S. Samsonov. *Variance reduction for additive functionals of Markov chains via martingale representations*, Statistics and Computing, 32(1), 16, 2022. https://doi.org/10.1007/s11222-021-10073-z

Апробация работы

Результаты диссертации были представлены на следующих конференциях, школах и семинарах:

1. Зимняя школа-конференция "New frontiers in high-dimensional probability and statistics 2"("Новые рубежи в теории вероятностей и математической статистике 2"), Москва, 22–23 февраля 2019 г. Тема доклада: "Concentration inequalities for functionals of Markov Chains with applications to variance reduction"("Неравенства концентрации для функционалов от цепей Маркова и их приложения к снижению дисперсии");

2. Конференция "Structural Inference in High-Dimensional Models 2"("Структурный вывод в многомерных моделях 2"), Пушкин, Санкт- Петербург, 26–30 августа 2019 г. (постер). Тема постера: "Variance Reduction for Dependent Sequences via Empirical Variance Minimisation"("Снижение дисперсии для зависимых последовательностей с помощью минимизации эмпирической дисперсии");

3. Доклад на научно-исследовательском семинаре "Структурное обучение Москва, 3 декабря 2019 г. "Variance reduction for dependent sequences with applications to Stochastic Gradient MCMC ("Снижение дисперсии для зависимых последовательностей с приложениями к алгоритмам MCMC на основе стохастических градиентов");

4. Осенняя школа НИУ ВШЭ и Яндекса по генеративным моделям, Москва, 26–29 ноября 2019 г. Доклад "Variance reduction methods for MCMC algorithms"("Снижение дисперсии в алгоритмах MCMC");

5. Зимняя школа "Математика машинного обучения - 2020"("Math of Machine Learning 2020"), г. Сочи, Сириус, 19-22 февраля 2020. Тема постера: "Алгоритмы снижения дисперсии на основе дискретных мартингальных разложений";

6. Городской семинар по теории вероятностей и математической статистике, Санкт-Петербург, ПОМИ РАН, 20 октября 2020 года. Тема доклада: "Variance reduction methods for MCMC algorithms"("Снижение дисперсии в алгоритмах MCMC");

7. Конференция "New Trends in Mathematical Stochastics г. Санкт-Петербург, институт Эйлера, 30.08.2021-03.09.2021. Тема доклада: "Probability and moment inequalities for additive functionals of geometrically ergodic Markov chains"("Неравенства концентрации для аддитивных функционалов от геометрически эргодических цепей Маркова");

8. Доклад на исследовательском семинаре "Структурное обучение Москва, 28 февраля 2023 г. Тема доклада: "Rosenthal type inequalities for Markov chains and their applications to Linear Stochastic Approximation"("Неравенства Розенталя для цепей Маркова и их приложения в задаче линейной стохастической аппроксимации").

Положения, выносимые на защиту

1. В Главе 1 получены новые аналоги неравенств Розенталя и Бернштейна для аддитивных функционалов от эргодических марковских цепей, которые сходятся к стационарному распределению с экспоненциальной скоростью либо в V-норме полной вариации, либо в полуметрике Канторовича-Васерштейна. Использованный метод доказательства основан на кумулянтном разложении и связи между кумулянтами и центральными моментами, устанавливаемой с помощью формулы Леонова-Ширяева.

2. В Главе 2 предложено обобщение метода снижения дисперсии с использованием контрольных переменных на случай последовательностей зависимых случайных величин, удовлетворяющих условию стационарности ковариаций. Для данного алгоритма получены оценки скорости убывания избыточной асимптотической дисперсии с ростом обучающей выборки. Также получены неравенства концентрации для квадратичных форм от функций от цепей Маркова с условием сжимаемости в метрике Канторовича-Васерштейна 1-го порядка. Полученные результаты применены к алгоритмам MCMC на основе динамики Ланжевена с использованием стохастических градиентов (SGLD).

3. В Главе 3 предложен новый подход к снижению дисперсии для аддитивных функционалов от марковских цепей на основе дискретного мартингального разложения. Для специального случая модели нормального шума, покрывающей неадаптированный алгоритм Ланжевена (ULA), произведен неасимптотический анализ снижения дисперсии, достигаемого предложенным алгоритмом. Соответствующий теоретический анализ основан на неравенстве Пуанкаре для гауссовских случайных векторов.

Достоверность результатов

Все результаты диссертационной работы обоснованы с помощью математических доказательств. Результаты диссертационной работы докладывались на конференциях и научных семинарах.

Структура и объем работы

Диссертация состоит из введения, секции с обозначениями, трех глав, заключения и списка литературы. Объем диссертации составляет 113 страниц, включая 105 страниц текста,

2 таблицы и 12 рисунков. Список литературы занимает 8 страниц и включает в себя 119 наименований.

Личный вклад автора

Вклад диссертанта был определяющим в результатах, приведенных в Главах 1 и 3 диссертации. Результаты и основные положения, перечисленные в данных главах, отражают персональный вклад соискателя в опубликованные работы, исключая результат Теоремы 5. Данная теорема является результатом совместной работы диссертанта и прочих соавторов [38]. Для полноты изложения в текст Главы 2 данного исследования включены результаты, полученные совместно с соавторами, в частности, результаты Секции 2.4: Алгоритм 1 и Теорема 6. Они являются результатами совместной работы диссертанта и соавторов работы [39]. Диссертанту принадлежат результаты Главы 2, относящиеся к концентрации квадратичных форм для марковских цепей с условием сжимаемости в метрике Канторовича-Васерштейна 1-го порядка, а также их приложениям для алгоритма Ланжевена со стохастическими градиентами (SGLD). Данные результаты представлены в Секции 2.5 и Секции 2.6, при этом идея доказательства Утверждения 2 принадлежит А. Наумову.

Используемые обозначения

Мы по умолчанию предполагаем, что все случайные величины, о которых идет речь в Главах 1 и 2 данной диссертации, принимают значения в полном сепарабельном метрическом пространстве (X, d) с борелевской сигма-алгеброй $\mathcal{X}$. Ситуации, когда $\mathsf{X} = \mathbb{R}^d$, отдельно оговариваются в тексте. Для меры со знаком (заряда) ξ на $(\mathsf{X}, \mathcal{X})$ и измеримой функции $g : \mathsf{X} \to \mathbb{R}$ обозначим

$$\xi(g) = \int_{\mathsf{X}} g(x)\xi(\mathrm{d}x)\,.$$

Для измеримой функции $V : \mathsf{X} \to [1, \infty)$ определим L_V как множество всех измеримых функций $g : \mathsf{X} \to \mathbb{R}$, для которых $\|g\|_V = \sup_{x\in\mathsf{X}}\left\{\frac{|g(x)|}{V(x)}\right\} < \infty$. V-норма (также называемая V-нормой полной вариации) меры со знаком ξ определяется как $\|\xi\|_V = \int_{\mathsf{X}} V(x)\,|\xi|\,(\mathrm{d}x)$, где $|\xi|$ — полная вариация меры ξ. В случае $V \equiv 1$, V-норма является нормой полной вариации и обозначается $\|\cdot\|_{\mathrm{TV}}$. V-норму для меры ξ можно также определить соотношением $\|\xi\|_V = \sup\{\xi(g) : \|g\|_V \le 1\}$ (см. [41, теорема D.3.2]). Обозначим $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$.

При рассмотрении марковской цепи $\{X_n\}_{n\in\mathbb{N}_0}$ с начальным распределением ξ и марковским ядром P на пространстве $(\mathsf{X}, \mathcal{X})$, без потери общности предполагается, что $\{X_n\}_{n\in\mathbb{N}_0}$ является соответствующим каноническим процессом, определенным на каноническом пространстве $(\mathsf{X}^{\mathbb{N}_0}, \mathcal{X}^{\otimes\mathbb{N}_0})$. Для вероятностной меры ξ на $(\mathsf{X}, \mathcal{X})$ обозначим соответствующие вероятность и математическое ожидание при начальном распределении ξ как P_ξ и E_ξ, соответственно. Положим $\mathsf{E}_x = \mathsf{E}_{\delta_x}$ и $\mathsf{P}_x = \mathsf{P}_{\delta_x}$ для всех $x \in \mathsf{X}$. Следующие определения можно найти, например, в книге [41]:

Определение 1. *Множество C называется малым множеством для марковского ядра* P*, если существуют $\epsilon \in (0, 1]$, $m \in \mathbb{N}$ и вероятностная мера ν, такие что для всех $x \in C$ и $\mathsf{A} \in \mathcal{X}$ выполняется*

$$\mathrm{P}^m(x, \mathsf{A}) \ge \epsilon\nu(\mathsf{A})\,. \tag{2}$$

В этом случае множество C называется $(m, \epsilon\nu)$-малым множеством.

Когда нас не интересует, какая именно вероятностная мера ν выбрана в (2), будем просто говорить, что C является (m, ϵ)-малым множеством. Для множества $\mathsf{A} \in \mathcal{X}$ определим время возвращения σ_A в множество A как $\sigma_\mathsf{A} = \inf\{n \ge 1 : X_n \in \mathsf{A}\}$. Множество A называется *достижимым множеством* для марковского ядра P, если $\mathsf{P}_x(\sigma_\mathsf{A} < \infty) > 0$ для любого $x \in \mathsf{X}$. Введем определение *сильно апериодического* ядра:

Определение 2. *Марковское ядро* P *называется сильно апериодическим, если для него существует достижимое $(1, \epsilon\nu)$-малое множество C, причем $\nu(C) > 0$.*

Важным классом марковских ядер P, изучаемым в Главе 1, является класс V-геометрически эргодических ядер, где $V : \mathsf{X} \to [1, \infty)$ — измеримая функция.

Определение 3. *Марковское ядро* P *называется V-геометрически эргодическим, если* P *допускает единственное инвариантное распределение π, такое, что $\pi(V) < \infty$, и существуют $c > 0$ и $\rho \in (0, 1)$, такие что для всех $x \in \mathsf{X}$*

$$\|\delta_x\mathrm{P}^n - \pi\|_V \le c\rho^n V(x)\,. \tag{3}$$

Если марковское ядро P является V-геометрически эргодическим, будем говорить, что и соответствующая марковская цепь $\{X_n\}_{n\in\mathbb{N}_0}$ является V-геометрически эргодической.

Для вероятностных мер ξ и ξ' на $(\mathsf{X},\mathcal{X})$ скажем, что вероятностная мера ν на $(\mathsf{X}^2,\mathcal{X}^{\otimes 2})$ является каплингом ξ и ξ', если для каждого $\mathsf{A}\in\mathcal{X}$, $\nu(\mathsf{A}\times\mathsf{X})=\xi(\mathsf{A})$ и $\nu(\mathsf{X}\times\mathsf{A})=\xi'(\mathsf{A})$. Обозначим через $\Pi(\xi,\xi')$ множество каплингов мер ξ и ξ' на $(\mathsf{X},\mathcal{X})$. Пусть $\mathsf{c}:\mathsf{X}\times\mathsf{X}\to\mathbb{R}_+$ есть полунепрерывная снизу, симметричная функция, такая что $\mathsf{c}(x,x')=0$ при $x=x'$ и существует такое $p_{\mathsf{c}}\in\mathbb{N}$, что для всех $x,x'\in\mathsf{X}$, $(\mathsf{d}(x,x')\wedge 1)^{p_{\mathsf{c}}}\le\mathsf{c}(x,x')$. Тогда полуметрика Канторовича-Васерштейна $\mathbf{W}_{\mathsf{c}}(\xi,\xi')$, ассоциированная с ценовой функцией c, определяется как

$$\mathbf{W}_{\mathsf{c}}(\xi,\xi')=\inf_{\nu\in\Pi(\xi,\xi')}\int_{\mathsf{X}\times\mathsf{X}}\mathsf{c}(x,x')\nu(\mathrm{d}x\mathrm{d}x')\,. \tag{4}$$

В англоязычной литературе данный объект обычно называется полуметрикой Вассерштейна. $\mathbf{W}_{\mathsf{c}}(\xi,\xi')$ является полуметрикой в том смысле, что она удовлетворяет аксиомам симметричности, неотрицательности и тождества, но, вообще говоря, не удовлетворяет неравенству треугольника. Далее, пусть $\mathbb{M}_1(\mathsf{X})$ - множество вероятностных мер на X. Обозначим для $p\ge 1$ класс вероятностных мер с конечным p-м моментом

$$\mathbb{S}_p(\mathsf{X},\mathsf{d}):=\{\xi\in\mathbb{M}_1(\mathsf{X}):\int_{\mathsf{X}}\mathsf{d}^p(x,x')\xi(\mathrm{d}x')<\infty\text{ для всех }x\in\mathsf{X}\}\,.$$

Для $p\ge 1$ и $\xi,\xi'\in\mathbb{S}_p(\mathsf{X},\mathsf{d})$, определим расстояние Канторовича-Васерштейна порядка p между ξ и ξ' как

$$\mathbf{W}_{\mathsf{d},p}(\xi,\xi'):=\inf_{\nu\in\Pi(\xi,\xi')}\Big\{\int_{\mathsf{X}\times\mathsf{X}}\mathsf{d}^p(x,x')\,\nu(\mathrm{d}x\mathrm{d}x')\Big\}^{1/p}.$$

Заметим, что $\mathbf{W}_{\mathsf{d},p}(\xi,\xi')$ является метрикой на $\mathbb{S}_p(\mathsf{X},\mathsf{d})$. Пусть $\mathcal{W}:\mathsf{X}\to[1,\infty)$ - измеримая функция, и $\bar{\mathcal{W}}(x,y)=(\mathcal{W}(x)+\mathcal{W}(y))/2$. Для $\beta\ge 0$ определим взвешенную липшицеву норму функции f относительно $\mathcal{W}$ в соответствии с формулой

$$[f]_{\beta,\mathcal{W}}=\max\left\{\sup_{x,x'\in\mathsf{X},\,x\neq x'}\frac{|f(x)-f(x')|}{\mathsf{c}^{1/2}(x,x')\bar{\mathcal{W}}^{\beta}(x,x')},\sup_{x\in\mathsf{X}}\frac{|f(x)|}{\mathcal{W}^{\beta}(x)}\right\},\tag{5}$$

и обозначим соответствующий класс функций $\mathcal{L}_{\beta,\mathcal{W}}=\{f:\mathsf{X}\to\mathbb{R}:[f]_{\beta,\mathcal{W}}<\infty\}$. Липшицева норма для функции $h:\mathsf{X}\to\mathbb{R}$ определяется как $\|h\|_{\mathsf{Lip}}=\sup_{x\neq y\in\mathsf{X}}\frac{|h(y)-h(x)|}{\mathsf{d}(x,y)}$. Обозначим через $\mathrm{Lip}_{\mathsf{d}}(L)$ и $\mathrm{Lip}_{b,\mathsf{d}}(L,\mathsf{B})$ класс липшицевых функций (соответственно, ограниченных липшицевых функций) на X с $\|h\|_{\mathsf{Lip}}\le L$ (соответственно, $\|h\|_{\mathsf{Lip}}\le L$ и $|h|_\infty\le\mathsf{B}$).

Обозначим для $q\in[1,\infty)$ $2q$-й момент стандартного нормального распределения

$$\mathrm{m}_{\mathrm{G},q}=(2q)!/(q!2^q)=2^q\mathbf{\Gamma}((2q+1)/2)/\pi^{1/2}\,,\tag{6}$$

где $\mathbf{\Gamma}$ — гамма-функция. Для мульти-индекса $\mathbf{k}=(k_1,\ldots,k_d)$ используем обозначения $\|\mathbf{k}\|=\max\limits_{i\in\{1,\ldots,d\}}k_i\,,\quad|\mathbf{k}|=\sum_{i=1}^d k_i\,,\quad\mathbf{k}!:=k_1!\ldots k_d!$. Для функции $f:\mathbb{R}^d\to\mathbb{R}$ обозначим через $\nabla f(x)$ и $\nabla^2 f(x)$ её градиент и гессиан в точке x.

Глава 1

Неравенства Розенталя и Бернштейна для линейных статистик от цепей Маркова

1.1. Введение

В данной главе диссертации рассматриваются свойства концентрации аддитивных функционалов от цепей Маркова, то есть свойства сумм вида

$$S_n = \sum\nolimits_{\ell=0}^{n-1} \{g(X_\ell) - \pi(g)\}\,, \tag{1.1}$$

где $\{X_\ell\}_{\ell=0}^{\infty}$ — цепь Маркова с ядром P, начальным распределением ξ и единственным инвариантным распределением π, а $g : \mathsf{X} \to \mathbb{R}$ — функция, удовлетворяющая условию $\pi(|g|) < \infty$. Целью данной главы является получение неравенств концентрации типа Бернштейна и Розенталя для S_n. Результаты данной главы опубликованы в работе [38].

Начнем с краткого обзора упомянутых неравенств для сумм независимых случайных величин. Пусть $(Y_\ell)_{\ell=0}^{n-1}$ — независимые случайные величины с $\mathsf{E}[Y_\ell] = 0$ и положим

$$\bar{S}_n = \sum\nolimits_{\ell=0}^{n-1} Y_\ell\,. \tag{1.2}$$

Предположим, что для некоторого $c > 0$, для всех $\ell \in \{0, \dots, n-1\}$ и целых $k \geq 3$ выполнена оценка $|\mathsf{E}[Y_\ell^k]| \leq (k!/2)\operatorname{Var}(Y_\ell)c^{k-2}$. Это условие известно как *условие Бернштейна* и влечет одноименное неравенство: для любого $t > 0$ и $n \in \mathbb{N}$,

$$\mathsf{P}(|\bar{S}_n| \geq t) \leq 2\exp\left\{-\frac{t^2/2}{\operatorname{Var}(\bar{S}_n) + ct}\right\}. \tag{1.3}$$

Условие и неравенство Бернштейна могут быть обобщены с помощью кумулянтного разложения, как предложено в [21]. Напомним, что k-й кумулянт случайной величины Y есть

$$\Gamma_k(Y) = \frac{1}{\mathrm{i}^k}\,\frac{\mathrm{d}^k}{\mathrm{d}t^k}\big(\log \mathsf{E}[\mathrm{e}^{\mathrm{i}tY}]\big)\Big|_{t=0}.$$

Бенткус в [21] предложил следующее обобщение неравенства Бернштейна: если существуют $\gamma \geq 0$ и $B \geq 0$ такие, что для всех $k \in \mathbb{N}\,, k \geq 2$, выполняется

$$|\Gamma_k(\bar{S}_n)| \leq (k!/2)^{1+\gamma}\operatorname{Var}(\bar{S}_n)B^{k-2}\,, \tag{1.4}$$

то для всех $t \geq 0$,

$$\mathsf{P}(|\bar{S}_n| \geq t) \leq 2\exp\left\{-\frac{t^2/2}{\operatorname{Var}(\bar{S}_n) + B^{1/(1+\gamma)}t^{2-1/(1+\gamma)}}\right\}. \tag{1.5}$$

Можно показать, что условие (1.4) также следует из некоторых обобщений условия Бернштейна, см., к примеру, Теорему 3.1 в [42]. Заметим, что неравенство Бернштейна влечет

экспоненциальную скорость убывания хвостов распределения $\bar{S}_n$. В то же время, важно исследовать и свойства сумм случайных величин, имеющих лишь конечное число моментов. В таких случаях предметом интереса обычно являются оценки (полиномиальных) моментов для $\bar{S}_n$. В частности, в работе [43] получена следующая версия неравенства Розенталя (см. также оригинальную статью [44]): для $q \geq 2$,

$$\mathsf{E}[|\bar{S}_n|^q] \leq C^q \big\{ q^{q/2} \operatorname{Var}(\bar{S}_n)^{q/2} + q^q \mathsf{E}[|\max_{\ell \in \{0,\dots,n-1\}} Y_\ell|^q] \big\}, \tag{1.6}$$

где C - абсолютная константа. Если условие (1.6) выполнено для всех $q \geq 2$, можно показать, что $\bar{S}_n$ удовлетворяет неравенству типа Бернштейна, при условии, что $|\max_{\ell \in \{1,\dots,n\}} Y_\ell|$ имеет конечную ψ_1-норму Орлича, см. например [12].

1.2. Научный вклад

В данной главе получены аналоги неравенств Розенталя (1.6) и Бернштейна (1.5) для аддитивных функционалов от цепей Маркова вида (1.1). Рассмотрены марковские ядра P, итерации которых $\xi \mathrm{P}^n$ сходятся к стационарному распределению π либо в V-норме полной вариации, либо в полуметрике Канторовича-Васерштейна. Доказательство основано на технике кумулянтного разложения, изложенной в [21], [42], и далее развитой в [45]. В стационарном случае (при совпадении стартового распределения ξ с π) ключевым шагом доказательства является оценка центральных моментов, связанных с $\{g(X_\ell)\}_{\ell=0}^{n-1}$. Данная оценка получена при помощи формулы Леонова-Ширяева [22]. Предложенная техника позволяет получить оценки с явными и вычислимыми константами. Также рассмотрен случай нестационарной цепи с произвольным стартовым распределением ξ. Результаты для нестационарного случая получены с использованием методов каплинга, см. [46] и главу 19 в [41].

1.3. Обзор литературы

Ниже мы обсудим основные известные в литературе результаты, связанные с концентрацией аддитивных функционалов от цепей Маркова, и соответствующие подходы к доказательствам. Так, в работах [47–49] получены неравенства типа Азумы-Хёфдинга для (1.1). Оценки для случая марковских ядер, которые являются сжимающими в метрике Канторовича-Васерштейна, получены в [50]. Однако результаты [50] требуют проверки дополнительных условий, связанных с такими величинами, как *гранулярность* и *локальная размерность*, которые сложно оценить на практике. В работах [51–53] установлены аналоги неравенств Хёфдинга и Бернштейна для ограниченных функций g с использованием спектральных методов. При этом предполагается, что соответствующее марковское ядро P имеет положительный спектральный зазор. Отметим однако, что V-геометрическая эргодичность P не обязательно влечет наличие спектрального зазора (см. [54]). Более того, неравенства в этих статьях формулируются с использованием верхних оценок на дисперсию S_n.

Подход, исследованный в [12, 14, 19, 20, 55, 56], заключается в использовании техники регенерации для получения моментных оценок и неравенств концентрации в предположении V-геометрической эргодичности. Эти техники основаны на конструкции Нуммелина [57] и [58], позволяющей разделить сумму S_n на блоки случайного размера, являющиеся

1-зависимыми случайными величинами. Особенно стоит отметить статьи [14] и [20]. В теореме 1 из [14] получено неравенство типа Бернштейна для V-геометрически эргодических сильно апериодических марковских ядер (см. Определение 2) и неограниченных функций. Теорема 1 из [20] обобщает этот результат на апериодические марковские цепи, но это обобщение применимо только к ограниченными функциям. Кроме того, упомянутые результаты неприменимы к марковским цепям, которые геометрически эргодичны в смысле полуметрики Канторовича-Васерштейна и не охватывают результаты, представленные в Секции 1.5.

Моментные оценки и неравенства типа Бернштейна были также получены для слабо зависимых случайных величин и для случайных величин, удовлетворяющих условиям перемешивания; см. [10, 45, 59]. Эти результаты, вообще говоря, несравнимы с представленными в диссертации, так как зависят от коэффициентов слабой зависимости или перемешивания вместо условий сноса и локальных условий миноризации (см. **A** 1-**A** 2).

1.4. Результаты для V-геометрически эргодических цепей Маркова

Начнем со случая, когда марковское ядро P является V-геометрически эргодическим (см. Определение 3). А именно, потребуем, чтобы P удовлетворяло следующим условиям:

A 1. *Существуют измеримая функция* $V : \mathsf{X} \to [\mathrm{e}, \infty)$, $\lambda \in (0,1)$ *и* $b \geq 0$, *такие что для любого* $x \in \mathsf{X}$, $\mathrm{P}V(x) \leq \lambda V(x) + b$.

A 2. *Существуют целое число* $m \geq 1$, $\epsilon \in (0,1)$ *и* $d \geq 0$, *такие что множество* $\{x \in \mathsf{X} : V(x) \leq d\}$ *является* (m, ϵ)*-малым и* $\lambda + 2b/(1+d) < 1$. *Здесь* λ *и* b *определены в* ***A*** *1.*

В отличие от стандартного определения функции Ляпунова V, в **A** 1 предполагается, что V принимает значения в $[\mathrm{e}, \infty)$, а не в $[1, \infty)$. Такой выбор позволяет избежать технических проблем при рассмотрении норм, связанных с $W(x) = \log V(x)$. При выполнении **A** 1 и **A** 2 марковское ядро P является V-геометрически эргодическим. Обозначим через π его единственное инвариантное распределение. Тогда по теореме 19.4.1 из [41], для любой вероятностной меры ξ, удовлетворяющей $\xi(V) < \infty$, и для всех $n \in \mathbb{N}$,

$$\|\xi \mathrm{P}^n - \pi\|_{\mathrm{TV}} \leq \|\xi \mathrm{P}^n - \pi\|_V \leq c\{\xi(V) + \pi(V)\}\rho^n \,. \tag{1.7}$$

Явные выражения для ρ и c в зависимости от параметров из **A** 1 и **A** 2 можно найти, например, в [41, теорема 19.4.1]. Перед тем как перейти к основным результатам, введем несколько обозначений. Для каждого $q \in \mathbb{N}, u \in \{1, \ldots, q-1\}$ и $\gamma \geq 0$, определим

$$\mathrm{B}_\gamma(u, q) = \frac{(2q)!}{u!} \sum_{(k_1, \ldots, k_u) \in \mathscr{E}_{u,q}} \prod_{i=1}^{u} (k_i!)^{\gamma+2} \,, \tag{1.8}$$

где $\mathscr{E}_{u,q} = \{(k_1, \ldots, k_u) \in \mathbb{N}^u : \sum_{i=1}^u k_i = 2q \,,\, k_i \geq 2\}$. Теперь сформулируем неравенство типа Розенталя для V-геометрически эргодической марковской цепи с ведущим (относительно n) членом оценки пропорциональным $\mathrm{Var}_\pi(S_n)$, умноженной на соответствующий момент гауссовского распределения $\mathrm{m}_{\mathrm{G},q}$ (см. (6)). Всюду в этой главе положим $\bar{g}(x) = g(x) - \pi(g)$.

Теорема 1. *Пусть выполнены* ***A*** *1,* ***A*** *2, и* $q \in \mathbb{N}$. *Тогда для любой функции* $g \in \mathrm{L}_{V^{1/(2q)}}$,

$$\mathsf{E}_\pi[|S_n|^{2q}] \le \mathrm{m}_{\mathrm{G},q}\{\mathrm{Var}_\pi(S_n)\}^q + \mathsf{C}_0^{2q}\|\bar{g}\|^{2q}_{V^{1/(2q)}} \sum\nolimits_{u=1}^{q-1} \frac{\mathrm{B}_0(u,q)n^u}{\rho^{u/2}\log^{2q-u}(1/\rho)}\,, \tag{1.9}$$

где $\mathsf{C}_0 = 2c\pi(V)$, *и* c *определено в* (1.7). *Более того, для любой вероятностной меры* ξ *на* $(\mathsf{X},\mathcal{X})$, *такой что* $\xi(V) < \infty$, *выполнено*

$$\mathsf{E}_\xi\big[\big|S_n\big|^{2q}\big] \le 2^{2q-1}\mathsf{E}_\pi\big[\big|S_n\big|^{2q}\big] + 2^{6q-1}\|\bar{g}\|^{2q}_{V^{1/(2q)}} c\{\xi(V)+\pi(V)\}\frac{q^{2q}}{\rho(\log(1/\rho))^{2q}}\,. \tag{1.10}$$

Доказательство данного результата основано на методе кумулянтного разложения, подробно описанном в [45] для случая слабо зависимых случайных величин. Для сумм независимых случайных величин данный метод был ранее описан в [21] и [42]. Обобщение на произвольное начальное распределение (1.10) выполнено с использованием методов каплинга [41, глава 19]. Стоит отметить, что результат Теоремы 1 не требует сильной апериодичности ядра P, в отличие от результатов [14]. Напомним, что сильная апериодичность P в частности влечет, что **A** 2 выполняется с $m = 1$. Ранее требование сильной апериодичности было ослаблено в [20], но только при требовании ограниченности функций g в S_n.

В результатах выше мы задали $q \in \mathbb{N}$ и рассмотрели $g \in \mathrm{L}_{V^{1/(2q)}}$. При этих предположениях мы не можем контролировать экспоненциальные моменты S_n. Далее рассмотрим случай $g \in \mathrm{L}_{W^\gamma}$, где $W = \log V$ и $\gamma \ge 0$. В этом случае, помимо неравенства типа Розенталя (1.9), мы можем сформулировать аналог неравенства Бернштейна (1.5).

Теорема 2. *Пусть выполнены* ***A*** *1,* ***A*** *2, и* $\gamma \ge 0$. *Тогда для всех* $g \in \mathrm{L}_{W^\gamma}$ *и* $t \ge 0$ *выполнено*

$$\mathsf{P}_\pi(|S_n| \ge t) \le 2\exp\left\{-\frac{t^2/2}{\mathrm{Var}_\pi(S_n) + \mathrm{J}_{n,W^\gamma}^{1/(\gamma+3)}\, t^{2-1/(\gamma+3)}}\right\},\ \textit{где}\ \mathrm{J}_{n,W^\gamma}\ \textit{имеет вид} \tag{1.11}$$

$$\mathrm{J}_{n,W^\gamma} = \left(\frac{n\rho^{-1/2}\{\log(1/\rho)\}^{-1}\mathsf{C}_0^2\|\bar{g}\|^2_{W^\gamma}}{\mathrm{Var}_\pi(S_n)} \vee 1\right)\frac{2^{1+3\gamma}\gamma^{3\gamma}\mathsf{C}_0\|\bar{g}\|_{W^\gamma}}{\log(1/\rho)}\,. \tag{1.12}$$

Более того, для любого стартового распределения ξ, *такого что* $\xi(V) < \infty$, *и функции* g, *такой что* $\|\bar{g}\|_{W^\gamma} = 1$, *выполнено*

$$\mathsf{P}_\xi(|S_n| \ge t) \le \mathsf{P}_\pi(|S_n| \ge \tfrac{t}{4}) + \big(\tfrac{\exp\{-h_1(\gamma,\rho)t^{\varpi\gamma}\}}{\rho^{1/2}} + \tfrac{\exp\{-h_2(\gamma)t^{\varpi\gamma}\}}{1-\rho}\big)c\{\xi(V)+\pi(V)\}\,, \tag{1.13}$$

где $\varpi_\gamma = 1/(1+\gamma)$, $h_1(\gamma,\rho) = \log(1/\rho)/(4^{1+\varpi_\gamma}\varpi_\gamma)$, *и* $h_2(\gamma) = (1+\gamma)/(2^{1+2\varpi_\gamma}\gamma)$.

Сравнивая (1.11) с (1.5), можно отметить, что в субэкспоненциальном режиме показатель $t^{1/(\gamma+1)}$ заменяется на $t^{1/(\gamma+3)}$. Аналогичное поведение хвостов оценки ранее было отмечено в [45]. Заметим также, что J_{n,W^γ} зависит от $\mathrm{Var}_\pi(S_n)$, более того, возможно, что J_{n,W^γ} растет с увеличением n. Этот недостаток имеют и другие результаты, ранее полученные с использованием кумулянтного разложения [45]. Доказательство неравенства типа Бернштейна (1.13) использует метод каплинга [41, глава 19].

Можно сравнить результаты Теоремы 2 с результатами, полученными в теоремах 1.1-1.3 [14] в дополнительном предположении сильной апериодичности P. Тогда (1.11) есть версия

неравенства Бернштейна с точной константой 2 перед дисперсией и с явной зависимостью от параметров из **A** 1 и **A** 2, что улучшает результаты теоремы 1.1 из [14]. Однако оценка [14] точнее для больших t и убывает как $\exp\{-t^{1/(1+\gamma)}\}$ по сравнению с $\exp\{-t^{1/(3+\gamma)}\}$ в (1.11). Стоит отметить, что показатель степени в членах, отражающих зависимость от начального условия (1.13), равен $1/(1+\gamma)$, как и в теореме 1.1 из [14]. Но в отличие от результата [14], зависимость от стартового распределения ξ появляется в качестве множителя, а не показателя степени.

1.5. Результаты для цепей Маркова, геометрически эргодических относительно полуметрики Канторовича-Васерштейна

В этой главе мы обобщаем результаты, полученные в Секции 1.4, на случай марковских ядер, которые являются сжимающими в смысле полуметрики Канторовича-Васерштейна. Это позволяет рассмотреть задачи, в которых марковское ядро P не является неприводимым. Таким образом, мы более не предполагаем **A** 2, и техника регенерации, лежащая в основе [14, 20], становится неприменимой. Марковские цепи с подобными ядрами P типичны в случае бесконечномерных пространств, см., например, [60, 61] и [41, глава 20]. В Секции 1.4 комбинация предположений **A** 1 и **A** 2 позволяла доказать существование и единственность инвариантного распределения π для P вместе с геометрической скоростью сходимости $\xi\mathrm{P}^n$ к π. Чтобы обобщить эти результаты без опоры на **A** 2, необходимо определить несколько объектов, связанных с полуметрикой Канторовича-Васерштейна. Рассмотрим ценовую функцию $\mathsf{c} : \mathsf{X} \times \mathsf{X} \to \mathbb{R}_+$, удовлетворяющую следующему условию:

C 1. c *полунепрерывна снизу и симметрична, при этом* $\mathsf{c}(x, x') = 0$ *для* $x = x'$. *Также существует* $p_{\mathsf{c}} \in \mathbb{N}$, *такое что для любых* $x, x' \in \mathsf{X}$, $(\mathsf{d}(x, x') \wedge 1)^{p_{\mathsf{c}}} \leq \mathsf{c}(x, x') \leq 1$.

Будем говорить, что K является каплинг-ядром для P, если для всех $(x, x') \in \mathsf{X}^2$ и $\mathsf{A} \in \mathcal{X}$, $\mathrm{K}((x, x'), \mathsf{A} \times \mathsf{X}) = \mathrm{P}(x, \mathsf{A})$ и $\mathrm{K}((x, x'), \mathsf{X} \times \mathsf{A}) = \mathrm{P}(x', \mathsf{A})$. Рассмотрим следующее предположение, ослабляющее условие малых множеств **A** 2:

A 3. *Существуют каплинг-ядро* K *для* P, $m \in \mathbb{N}, \varepsilon \in (0, 1)$, $\kappa_{\mathrm{K}} \geq 1$, *такие что*

$$\mathrm{K}\mathsf{c}(x, x') \leq \kappa_{\mathrm{K}}\mathsf{c}(x, x')\,, \qquad \mathrm{K}^m\mathsf{c}(x, x') \leq (1 - \varepsilon \mathbb{1}_{\bar{\mathsf{C}}}(x, x'))\mathsf{c}(x, x')\,, \tag{1.14}$$

где $\bar{\mathsf{C}} = \{V(x) \leq d\} \times \{V(x') \leq d\}$, *и параметр* d *удовлетворяет* $\lambda + 2b/(1 + d) < 1$. *Здесь* λ *и* b *даны в* ***A*** *1, а функция* c *определена в* ***C*** *1.*

Можно показать, что предположения **A** 1 и **A** 3 влекут существование и единственность инвариантного распределения π и геометрическую скорость сходимости $\xi\mathrm{P}^n$ к π для любого начального распределения ξ в смысле $\mathbf{W}_{\mathsf{c}^{1/2}\bar{V}^{1/2}}$.

Утверждение 1. *Пусть выполнены* ***A*** *1,* ***A*** *3, и* ***C*** *1. Тогда* P *допускает единственное инвариантное распределение* π, *такое что* $\pi(V) < \infty$. *Более того, для всех начальных распределений* ξ *и* $n \in \mathbb{N}$,

$$\mathbf{W}_{\mathsf{c}}\left(\xi\mathrm{P}^n, \pi\right) \leq \mathbf{W}_{\mathsf{c}^{1/2}\bar{V}^{1/2}}\left(\xi\mathrm{P}^n, \pi\right) \leq c_1\varrho^n\left[\xi(V^{1/2}) + \pi(V^{1/2})\right]\,, \tag{1.15}$$

где явные выражения для $c_1 > 0$ *и* $\varrho \in (0;1)$ *можно найти в Утверждении 1 работы [38].*

Первый основной результат этого раздела — аналог неравенства Розенталя (1.9). Рассмотрим функции g из взвешенного липшицевого класса $\mathcal{L}_{\beta,V}$ с подходящим $\beta > 0$ (см. определение $\mathcal{L}_{\beta,V}$ в (5)). Требование лишь конечной V^β-нормы, как в Секции 1.4, было бы недостаточным. Аналогично Теореме 1, ведущим членом оценки является $\mathrm{Var}_\pi(S_n)$, умноженная на момент гауссовской случайной величины.

Теорема 3. *Пусть выполнены* ***A*** *1,* ***A*** *3,* ***C*** *1, и* $q \in \mathbb{N}$. *Тогда для любой функции* $g \in \mathcal{L}_{1/(4q),V}$,

$$\mathsf{E}_\pi[|S_n|^{2q}] \le \mathrm{m}_{\mathrm{G},q}\{\mathrm{Var}_\pi(S_n)\}^q + \mathsf{C}_1^{2q}[\bar{g}]^{2q}_{1/(4q),V} \sum\nolimits_{u=1}^{q-1} \frac{\mathrm{B}_0(u,q) n^u}{\varrho^{u/2}\{\log(1/\varrho)\}^{2q-u}}\,, \tag{1.16}$$

где $\mathrm{B}_0(u,q)$ *определено в* (1.8), *и* $\mathsf{C}_1 = 4c_1\{\pi(V)\}^{1/2}$.

Доказательство следует той же общей идее, что и доказательство Теоремы 1, и также опирается на метод кумулянтов. Данный результат допускает обобщение на нестационарный случай аналогично (1.10). При этом вместо максимального каплинга используется ядерный каплинг. Сформулируем следствие из данного результата для функций $g \in \mathcal{L}_{1,W^\gamma}$ в виде неравенства Бернштейна:

Теорема 4. *Пусть выполнены* ***A*** *1,* ***A*** *3,* ***C*** *1. Тогда для любого* $\gamma \ge 0$, $g \in \mathcal{L}_{1,W^\gamma}$, *и* $t \ge 0$,

$$\mathsf{P}_\pi(|S_n| \ge t) \le 2\exp\left\{-\frac{t^2/2}{\mathrm{Var}_\pi(S_n) + \mathfrak{J}_{n,W^\gamma}^{1/(\gamma+3)} t^{2-1/(\gamma+3)}}\right\},\ \textit{где}\ \mathfrak{J}_{n,W^\gamma}\ \textit{имеет вид}$$

$$\mathfrak{J}_{n,W^\gamma} = \left(\frac{n\varrho^{-1/2}\{\log(1/\varrho)\}^{-1}\mathsf{C}_1^2(2\gamma)^{4\gamma}[\bar{g}]^2_{1,W^\gamma}}{\mathrm{Var}_\pi(S_n)} \vee 1\right)\frac{2(2\gamma)^{2\gamma}\mathsf{C}_1[\bar{g}]_{1,W^\gamma}}{\log(1/\varrho)}\,. \tag{1.17}$$

Теорема 5. *В предположениях Теоремы 4 для любой вероятностной меры* ξ *на* $(\mathsf{X},\mathcal{X})$, *удовлетворяющей* $\xi(V^{1/2}) < \infty$, *верно, что для всех* $t \ge 0$

$$\begin{aligned}\mathsf{P}_\xi(|S_n| \ge t) \le \mathsf{P}_\pi(|S_n| \ge t/2) &+ \exp\left(-\frac{\log(1/\varrho)t^{\varpi_\gamma}}{2^{3+\varpi_\gamma}[\bar{g}]^{\varpi_\gamma}_{1,W^\gamma}\varpi_\gamma}\right)c_1^{1/2}\{\pi(V^{1/2}) + \xi(V^{1/2})\}^{1/2}h_1(\varrho)\\ &+ \exp\left(-\frac{(1+\gamma)\upsilon_\gamma t^{\varpi_\gamma}}{2^{5+\varpi_\gamma}[\bar{g}]^{\varpi_\gamma}_{1,W^\gamma}\gamma}\right)c_1^{\upsilon_\gamma}\{\pi(V^{1/2}) + \xi(V^{1/2})\}^{\upsilon_\gamma}h_2(\varrho)\,,\end{aligned}$$

где $\varpi_\gamma = 1/(1+\gamma)$, $\upsilon_\gamma = 1 \wedge (2\gamma)^{-1}$, *и выражения для* $h_1(\varrho)$, $h_2(\varrho)$ *приведены в [38, Теорема 12].*

Насколько известно автору, в Теореме 4 впервые получено неравенство типа Бернштейна для функций g из взвешенного липшицева класса $\mathcal{L}_{1,W^\gamma}$ без условия **A** 2 или его аналогов. Предыдущие результаты такого рода для неограниченных функций и слабо зависимых последовательностей [10] покрывали лишь случай функций g линейного роста. Также Теорема 5 позволяет сохранить ту же скорость убывания члена, отражающего зависимость от начальных условий (то есть, $\exp\{-t^{1/(1+\gamma)}\}$), которая была ранее получена для V-геометрически эргодического случая, рассмотренного в (1.13). Это первый результат такого рода для ядер, которые геометрически эргодичны в полуметрике Канторовича-Васерштейна.

Глава 2

Снижение дисперсии для зависимых последовательностей и его приложения к алгоритмам MCMC со стохастическими градиентами

2.1. Введение

В данной главе диссертации предложен и проанализирован новый метод снижения дисперсии для аддитивных функционалов от зависимых случайных величин. Ключевым техническим элементом анализа являются неравенства типа Розенталя, подобные тем, что были получены ранее в Секции 1. Основные результаты главы опубликованы в [39].

Начнем с описания постановки задачи снижения дисперсии в методах Монте-Карло. Основной целью данных методов является вычисление интеграла $\pi(f) = \int_{\mathsf{X}} f(x)\pi(\mathrm{d}x)$ относительно вероятностной меры π для некоторой интегрируемой функции $f : \mathsf{X} \mapsto \mathbb{R}$, заданной на измеримом пространстве $(\mathsf{X}, \mathcal{X})$. Чаще всего на практике рассматривается случай пространства $\mathsf{X} \subseteq \mathbb{R}^d$ и меры π, допускающей плотность относительно меры Лебега на $\mathbb{R}^d$. Для упрощения обозначений, будем также обозначать данную плотность через π. В этом случае рассмотренная нами задача сводится к вычислению интеграла

$$\int_{\mathsf{X}} f(x)\pi(x)\,\mathrm{d}x \tag{2.1}$$

Известно, что в случае $\mathsf{X} = [0,1]^d$ приближение интеграла (2.1) с заданной степенью точности при помощи детерминированных алгоритмов требует вычислить экспоненциально много (в зависимости от размерности d) значений функции f, см. к примеру [62]. Поэтому с ростом d подобные методы быстро становятся неприменимы. Альтернативой квадратурным формулам являются стохастические оценки (2.1), получаемые с помощью метода Монте-Карло и его модификаций [26]. Заметим, что $\pi(f) = \mathsf{E}_\pi[f(X)]$, где случайная величина X имеет распределение π. Следовательно, состоятельной оценкой $\pi(f)$ в силу закона больших чисел является выборочное среднее

$$\pi_N(f) := N^{-1}\sum\nolimits_{k=0}^{N-1} f(X_k)\,, \quad N \in \mathbb{N},$$

где $(X_k)_{k=0}^{N-1}$ - независимые, одинаково распределенные случайные величины с распределением π. Более того, если $\pi(f^2) < \infty$, можно построить для $\pi(f)$ асимптотический доверительный интервал с уровнем доверия $1-\alpha$ на основе центральной предельной теоремы:

$$\left[\pi_N(f) - \mathsf{q}_{1-\alpha/2}\sqrt{\tfrac{\mathrm{Var}_\pi(f)}{N}}, \pi_N(f) + \mathsf{q}_{1-\alpha/2}\sqrt{\tfrac{\mathrm{Var}_\pi(f)}{N}}\right], \tag{2.2}$$

где $\mathsf{q}_{1-\alpha/2}$ - квантиль стандартного нормального распределения уровня $1-\alpha/2$. Таким образом, более точную оценку для $\pi(f)$ можно получить либо путем увеличения размера выборки N, либо путем снижения дисперсии $\mathrm{Var}_\pi(f)$. Одним из наиболее популярных методов, направленных на снижение $\mathrm{Var}_\pi(f)$, является метод *контрольных переменных*, см. [23], [24].

Целью метода контрольных переменных является построение легко вычислимой случайной величины Y (контрольной переменной), такой, что $\mathsf{E}[Y] = 0$, $\mathsf{E}[Y^2] < \infty$, и дисперсия случайной величины $f(X) - Y$ мала (напомним, что случайная величина X имеет распределение π). Чаще всего на практике [27] используются контрольные переменные вида $Y = g(X)$, где функция g удовлетворяет условию $\pi(g) = 0$. Обычно практическое применение метода контрольных переменных в описанной постановке состоит из двух этапов. На первом этапе выбирается класс функций $\mathcal{G} = \{g : \pi(g) = 0\}$, называемый классом контрольных переменных. Затем на основе независимых одинаково распределенных величин $(X_k)_{k=0}^{n-1}$ с распределением π и некоторой оптимизационной задачи выбирается $\widehat{g}_n \in \mathcal{G}$. После этого для новой выборки $(X'_k)_{k=0}^{N-1}$ из π, не зависящей от $(X_k)_{k=0}^{n-1}$, строится оценка

$$\pi_N(f - \widehat{g}_n) = N^{-1} \sum\nolimits_{k=0}^{N-1} \{f(X'_k) - \widehat{g}_n(X'_k)\}\,. \tag{2.3}$$

Заметим, что $\pi(f - \widehat{g}_n) = \pi(f)$, и удачный выбор функции $\widehat{g}_n$ может обеспечить более точные асимптотические доверительные интервалы для $\pi(f)$, имеющие вид

$$\left[\pi_N(f - \widehat{g}_n) - \mathsf{q}_{1-\alpha/2} \sqrt{\tfrac{\mathrm{Var}_\pi(f-\widehat{g}_n)}{N}}, \pi_N(f - \widehat{g}_n) + \mathsf{q}_{1-\alpha/2} \sqrt{\tfrac{\mathrm{Var}_\pi(f-\widehat{g}_n)}{N}}\right]. \tag{2.4}$$

Возможный выигрыш от применения метода контрольных переменных зависит от того, как изменяется длина асимптотического доверительного интервала в модифицированном эксперименте (2.4) по сравнению с исходным экспериментом (2.2).

2.2. Метод контрольных переменных для зависимых наблюдений

Зачастую оказывается, что генерация независимых случайных величин с распределением π невозможна, либо вычислительно неэффективна [7]. Такая ситуация типична для $\mathsf{X} = \mathbb{R}^d$ и высокой размерности d. В таких случаях оказывается проще построить последовательность зависимых случайных величин $(X_k)_{k=0}^{\infty}$, распределение которых сходится к π с ростом k. При условии выполнения центральной предельной теоремы, асимптотический доверительный интервал для $\pi(f)$ в подобном эксперименте принимает вид

$$\left[\pi_N(f) - \mathsf{q}_{1-\alpha/2} \sqrt{\tfrac{V_\infty(f)}{N}}, \pi_N(f) + \mathsf{q}_{1-\alpha/2} \sqrt{\tfrac{V_\infty(f)}{N}}\right], \tag{2.5}$$

где $V_\infty(f)$ - асимптотическая дисперсия, определяемая соотношением

$$V_\infty(f) := \lim_{N\to\infty} N \cdot \mathsf{E}[\big(\pi_N(f) - \pi(f)\big)^2]\,. \tag{2.6}$$

Таким образом, естественной целью методов снижения дисперсии для зависимых случайных величин является разработка эксперимента с меньшей асимптотической дисперсией $V_\infty(\cdot)$.

Чаще всего зависимая последовательность $(X_k)_{k=0}^{\infty}$ строится при помощи алгоритмов Монте-Карло по схеме марковской цепи (Markov Chain Monte Carlo, MCMC). Подобные алгоритмы строят $(X_k)_{k=0}^{\infty}$, являющуюся цепью Маркова с единственным инвариантным распределением π, после чего оценивают $\pi(f)$ при помощи эргодического среднего $\pi_N(f)$. При

этом $\pi_N(f)$ может иметь большую дисперсию из-за неизбежных корреляций между соседними элементами цепи. Особенно популярны в приложениях алгоритмы MCMC, основанные на динамике Ланжевена, см. [36, 37, 63]. Особенно стоит отметить алгоритм SGLD [33], см. Секцию 2.6. Существуют также модификации алгоритма SGLD, которые генерируют зависимые последовательности $(X_k)_{k=0}^{\infty}$, не являющиеся цепями Маркова, например [29].

Важным вопросом, связанным с методом контрольных переменных, является выбор критерия, используемого для выбора контрольной переменной. В случае независимых наблюдений обычно используется критерий на основе метода наименьших квадратов [27], [64] или эмпирической дисперсии [65]. Последний приводит к методу минимизации эмпирической дисперсии (EVM), подробно изученному в [65]. В случае зависимых наблюдений $(X_k)_{k=0}^{\infty}$ ситуация более сложная, что связано со сложностью оценивания $V_\infty(\cdot)$ в (2.6). В частности, в случае $(X_k)_{k=0}^{\infty}$, являющейся цепью Маркова, в недавних работах [66–68] используется метод, основанный на минимизации оценки наименьших квадратов, цель которого — минимизация маргинальной дисперсии $\mathrm{Var}_\pi(\cdot)$ вместо V_∞. Альтернативный подход, предложенный в [32], использует минимизацию оценки V_∞, аналогичную рассматриваемой в данной работе. Однако в [32] рассматриваются цепи Маркова с ядром, удовлетворяющим либо условию V-геометрической эргодичности, либо L^p-транспортному неравенству (см. Определение 3.1 в [39]), что ограничивает класс возможных приложений.

2.3. Научный вклад

Ниже перечислены основные результаты данной главы:

- Предложено обобщение метода снижения дисперсии ESVM, предложенного в [32], для зависимых случайных последовательностей, удовлетворяющих условию стационарности ковариаций (см. **(CS)** в Секции 2.4). Для метода получены оценки избыточной асимптотической дисперсии $V_\infty(f-\widehat{g}_n) - \inf_{g\in\mathcal{G}} V_\infty(f-g)$, где контрольная переменная $\widehat{g}_n$ построена с использованием алгоритма ESVM;
- Получены неравенства концентрации для квадратичных форм от функций от цепей Маркова, удовлетворяющих условию равномерной геометрической эргодичности в метрике Канторовича-Васерштейна $\mathbf{W}_{\mathsf{d},1}$. Данные результаты применены к итерациям алгоритма SGLD (см. Секцию 2.6). В частности, показано, что с выбором параметрического класса контрольных переменных $\mathcal{G}$, верно

$$V_\infty(f-\widehat{g}_n) - \inf_{g\in\mathcal{G}} V_\infty(f-g) \lesssim n^{-1/2}\log^{5/2}(n),$$

где n - число наблюдений $X_0,\dots,X_{n-1}$, использованных для оценки $\widehat{g}_n$.

2.4. Минимизация спектральной оценки асимптотической дисперсии

Пусть $(\Omega, \mathfrak{F}, (\mathfrak{F}_k)_{k\geq 0}, \mathsf{P})$ - вероятностное пространство с фильтрацией $(\mathfrak{F}_k)_{k\geq 0}$, $(X_k)_{k=0}^{\infty}$ - случайный процесс, согласованный с $(\mathfrak{F}_k)_{k\geq 0}$ и принимающий значения в X. На данном этапе предполагается, что X - полное сепарабельное метрическое пространство. Обозначим через $\mathcal{G}$ класс контрольных переменных, то есть функций $g\in\mathcal{G}$, удовлетворяющих условиям $\pi(g^2) <$

∞, $\pi(g) = 0$ и $\mathsf{E}[g^2(X_k)] < \infty$ для всех $k \in \mathbb{N}$. Обозначим класс функций $\mathcal{H} := \{f - g : \, g \in \mathcal{G}\}$. Напомним, что $\bar{h} = h - \pi(h)$. Предположим, что класс $\mathcal{H}$ удовлетворяет условию **(CS)**.

(CS). *Для всех $h \in \mathcal{H}$ существует симметричная, суммируемая и положительно полуопределённая последовательность $(\rho^{(h)}(\ell))_{\ell\in\mathbb{Z}}$, удовлетворяющая условиям:*

- $\rho^{(h)}(0) = \mathrm{Var}_\pi(h)$;
- *Существует константа $R > 0$, не зависящая от h и ℓ, такая что для всех $\ell \in \mathbb{N}_0$*

$$\sum\nolimits_{k\in\mathbb{N}_0} \left| \mathsf{E}\big[\bar{h}(X_k)\bar{h}(X_{k+\ell})\big] - \rho^{(h)}(\ell) \right| \le R\,,$$

- $\lim_{\ell\to\infty} \sum_{k\in\mathbb{N}_0} \left| \mathsf{E}\big[\bar{h}(X_k)\bar{h}(X_{k+\ell})\big] - \rho^{(h)}(\ell) \right| = 0.$

Условие **(CS)** гарантирует существование $V_\infty(h)$, определенной в (2.6) для любой функции $h \in \mathcal{H}$, см. Утверждение 2.1 в [39]. Более того, $V_\infty(h)$ может быть записана в виде

$$V_\infty(h) = \sum_{\ell\in\mathbb{Z}} \rho^{(h)}(\ell). \tag{2.7}$$

Поскольку прямое вычисление $V_\infty(h)$ обычно невозможно, предложенный алгоритм снижения дисперсии должен опираться на ее эмпирический аналог, $V_n(h)$. Среди существующих оценок асимптотической дисперсии (см. [69]) мы используем спектральную оценку

$$V_n(h) := \sum_{|\ell|<b_n} w_n(\ell)\rho_n^{(h)}(|\ell|)\,, \quad \rho_n^{(h)}(|\ell|) = n^{-1}\sum_{k=0}^{n-\ell-1} \big(h(X_k) - \pi_n(h)\big)\big(h(X_{k+\ell}) - \pi_n(h)\big)\,. \tag{2.8}$$

В (2.8) параметр b_n обычно увеличивается с ростом n, а веса $w_n(\ell)$ имеют вид $w_n(\ell) = w(\ell/b_n)$ для симметричной неотрицательной функции w, такой что $\sup_{y\in[0,1]} |w(y)| \le 1$, и $w(y) = 1$ для $y \in [-1/2, 1/2]$. Теперь сформулируем алгоритм ESVM для зависимых последовательностей, удовлетворяющих **(CS)**. Основываясь на спектральной оценке $V_n(h)$ из (2.8), будем выбирать $\widehat{g}_n$ (эквивалентно, $\widehat{h}_n = f - \widehat{g}_n$) как минимизатор спектральной оценки асимптотической дисперсии

$$\widehat{h}_n \in \arg\min\nolimits_{h\in\mathcal{H}} V_n(h)\,. \tag{2.9}$$

Псевдокод процедуры ESVM приведен в Алгоритме 1.

Алгоритм 1 Минимизация спектральной оценки асимптотической дисперсии (ESVM)

Входные данные: Независимые последовательности: $\mathbf{X}_n = (X_k)_{k=0}^{n-1}$ и $\mathbf{X}'_N = (X'_k)_{k=0}^{N-1}$.

1. Фиксируем класс $\mathcal{G}$ функций с $\pi(g) = 0$ для всех $g \in \mathcal{G}$;

2. Вычисляем $\widehat{g}_n \in \arg\min_{g\in\mathcal{G}} V_n(f - g)$, где V_n вычисляется на основе траектории $\mathbf{X}_n$.

Результат: Оценка $\pi_N(f - \widehat{g}_n)$, вычисляемая на последовательности $\mathbf{X}'_N$.

Построение контрольной переменной. Если π известно хотя бы с точностью до нормализующей константы, можно построить контрольные переменные, зависящие только от градиента $\nabla \log \pi$, используя оператор Стейна, как предложено в [70, 71]. Данный подход

был подробно исследован в литературе, см. [27, 64, 72]. Соответствующие контрольные переменные (часто называемые стейновскими контрольными переменными) имеют вид

$$g_\phi(\theta) = \langle \phi(\theta), \nabla \log \pi(\theta) \rangle + \operatorname{div}\big(\phi(\theta)\big), \tag{2.10}$$

где $\phi : \mathbb{R}^d \to \mathbb{R}^d$ - гладкая функция, а $\operatorname{div}(\phi)$ — дивергенция ϕ. Тогда при выполнении соответствующих условий на $\log \pi(\theta)$ можно гарантировать, что $\pi(g_\phi) = 0$, см. [72].

Теоретический анализ. Для упрощения задачи оптимизации в (2.9), предположим, что $\mathcal{H}$ является вполне ограниченным, и обозначим через $\mathcal{H}_\varepsilon$ минимальную ε-сеть относительно $L^2(\pi)$-нормы, то есть наименьшее возможное множество функций $\mathcal{H}_\varepsilon \subset \mathcal{H}$, такое что для всех $h \in \mathcal{H}$ существует $h_\varepsilon \in \mathcal{H}_\varepsilon$, такое что $L^2(\pi)$-расстояние между h и h_ε не превышает ε. Рассмотрим задачу минимизации по классу $\mathcal{H}_\varepsilon$:

$$\widehat{h}_{n,\varepsilon} \in \arg\min_{h \in \mathcal{H}_\varepsilon} V_n(h)\,. \tag{2.11}$$

Можно показать, что асимптотическая дисперсия $V_\infty(\widehat{h}_{n,\varepsilon})$ близка к наименьшей асимптотической дисперсии элементов класса $\mathcal{H}$, т.е. к $\inf_{h\in\mathcal{H}} V_\infty(h)$. Для этого необходимо предположение о скорости убывания $(\rho^{(h)}(\ell))_{\ell\in\mathbb{Z}}$ из **(CS)**:

(CD). *Существуют* $\varsigma > 0$ *и* $\lambda \in [0,1)$, *такие что для всех* $h \in \mathcal{H}$ *и* $\ell \in \mathbb{N}_0$, $\big|\rho^{(h)}(\ell)\big| \le \varsigma\lambda^\ell$.

Теорема 6. *Пусть выполнены условия* ***(CS)*** *и* ***(CD)****. Дополнительно предположим, что для любого* $n \in \mathbb{N}$ *существует непрерывная убывающая функция* α_n, *удовлетворяющая*

$$\sup_{h\in\mathcal{H}} \mathsf{P}\Big(\big|V_n(h) - \mathsf{E}[V_n(h)]\big| > t\Big) \le \alpha_n(t), \quad t > 0. \tag{2.12}$$

Тогда, для любого $\delta \in (0,1)$ *и* $\varepsilon > 0$, *с вероятностью не менее* $1-\delta$, *выполняется*

$$V_\infty(\widehat{h}_{n,\varepsilon}) - \inf_{h\in\mathcal{H}} V_\infty(h) \lesssim \alpha_n^{-1}\left(\frac{\delta}{2|\mathcal{H}_\varepsilon|}\right) + \Big(\sqrt{R}n^{-1/2} + \sqrt{D}\Big)\, b_n\varepsilon + \sqrt{RD}\, b_n n^{-1/2} \\ + \big(R + \varsigma(1-\lambda)^{-1}\big)\, b_n n^{-1} + \varsigma(1-\lambda)^{-2}n^{-1} + \varsigma(1-\lambda)^{-1}\lambda^{b_n/2},$$

где α_n^{-1} - *обратная функция к* α_n, *и* $D = \sup_{h\in\mathcal{H}} \operatorname{Var}_\pi(h)$, *и первая асимптотическая дисперсия является условной по случайным величинам* $X_0, \dots, X_{n-1}$, *по которым была вычислена* $\widehat{h}_{n,\varepsilon}$ *в* (2.11).

Теорема 6 обобщает аналогичные результаты, ранее полученные в [65] для независимых последовательностей $(X_k)_{k=0}^\infty$ и в [32] для марковских цепей с условием геометрической эргодичности в V-норме полной вариации и метрике Канторовича-Васерштейна $\mathbf{W}_{\mathsf{d},2}$. Данная теорема применима к любой зависимой последовательности, для которой можно проверить выполнение условий **(CS)** и **(CD)**, а также неравенство концентрации (2.12) для квадратичной формы $V_n(h)$. Далее мы исследуем эти свойства для цепей Маркова, удовлетворяющих условию равномерной эргодичности относительно метрики Канторовича-Васерштейна $\mathbf{W}_{\mathsf{d},1}$.

2.5. Приложения к марковским ядрам, эргодическим в смысле метрики Канторовича-Васерштейна

В дальнейшем предположим, что $(X_k)_{k=0}^{\infty}$ - цепь Маркова на полном сепарабельном метрическом пространстве (X,d) с борелевской σ-алгеброй $\mathcal{X}$ и марковским ядром P. Рассмотрим ядро P, удовлетворяющее условию $\mathbf{W}_{\mathsf{d},p}$-равномерной эргодичности для некоторого $p \geq 1$:

(WE)$\cdot p$ *Существует* $x_0 \in \mathsf{X}$, *такое что* $\int_{\mathsf{X}} \mathsf{d}(x_0,x)\mathrm{P}(x_0,\mathrm{d}x) < \infty$ *и такая константа* $\Delta_p \in [0,1)$ *(возможно, зависящая от* p*), что*

$$\sup_{(x,x')\in\mathsf{X}^2,\, x\neq x'} \frac{\mathbf{W}_{\mathsf{d},p}(\delta_x\mathrm{P},\delta_{x'}\mathrm{P})}{\mathsf{d}(x,x')} = \Delta_p\,.$$

Известно (см. теорему 20.3.4 в [41]), что выполнение условия **(WE)**-p с некоторым $p \geq 1$, означает, что P имеет единственное инвариантное распределение π. Более того, для любой вероятностной меры ξ с конечным p-м моментом,

$$\mathbf{W}_{\mathsf{d},p}(\xi\mathrm{P}^n,\pi) \leq \Delta_p^n \mathbf{W}_{\mathsf{d},p}(\xi,\pi)\,, \quad n\in\mathbb{N}. \tag{2.13}$$

Случай марковских ядер, удовлетворяющих **(WE)**-2, уже рассматривался ранее в [32]. Однако исследование свойств концентрации квадратичных форм $V_n(h)$ в таком случае требует дополнительных ограничительных свойств от марковского ядра $\mathrm{P}(x,\cdot)$, связанных с выполнением L^2-транспортного неравенства, см. Утверждение 3.2 в [39]. Далее мы сосредоточимся на более общем условии **(WE)**-1. В этом случае условия **(CS)** и **(CD)** могут быть проверены для класса $\mathcal{H}$, являющегося подмножеством класса ограниченных липшицевых функций.

Утверждение 2. *Пусть* $\mathcal{H} \subset \mathrm{Lip}_{b,\mathsf{d}}(L,\mathsf{B})$ *и выполнено* ***(WE)***-1. *Тогда для любого начального распределения* $\xi \in \mathbb{S}_1(\mathsf{X},\mathsf{d})$ *предположения* ***(CS)*** *и* ***(CD)*** *выполнены с*

$$\rho^{(h)}(\ell) = \mathsf{E}_\pi\big[\bar{h}(X_0)\bar{h}(X_{|\ell|})\big], \quad \lambda = \Delta_1, \tag{2.14}$$

и константами R, ς *указанными в [39, Утверждение 3.3]. Более того, для любого* $p\in\mathbb{N}$,

$$\mathsf{P}_\xi\big(\big|V_n(h) - \mathsf{E}_\xi[V_n(h)]\big| \geq t\big) \leq \frac{\bar{C}^p_{\mathsf{R},1} B^{2p} b_n^{3p/2} p^p}{n^{p/2}t^p} + \frac{\bar{C}^p_{\mathsf{R},2} B^{2p} b_n^{2p} p^{2p}}{n^{p-1}t^p}, \tag{2.15}$$

где $\bar{C}_{\mathsf{R},1}$ *и* $\bar{C}_{\mathsf{R},2}$ *- константы, приведенные в [39, уравнение (A.28)].*

Доказательство Утверждения 2 основано на версии неравенства Розенталя, адаптированной из [45]. Отметим, что $V_n(h)$ является квадратичной формой от $h(X_0),\ldots,h(X_{n-1})$. Изучение свойств концентрации таких объектов является сложной задачей даже для независимых случайных величин [73]. Недавние результаты [74] и [75] для квадратичных форм от цепей Маркова покрывают только случай равномерной геометрической эргодичности (относительно нормы полной вариации). Данное условие является значительно более ограничительным по сравнению с условиями Утверждения 2. Более того, случай равномерно геометрически эргодических цепей Маркова не покрывает алгоритмы, рассмотренные в Секции 2.6.

2.6. Приложения к алгоритмам MCMC на основе динамики Ланжевена

Рассмотрим ситуацию, когда интересующее нас распределение π является вероятностной мерой на $\mathbb{R}^d$ с плотностью относительно меры Лебега, которую мы также обозначим через π. Предположим, что существует функция $U(\theta) : \mathbb{R}^d \to \mathbb{R}$, такая, что $\tilde{C} = \int_{\mathbb{R}^d} e^{-U(\theta)} \mathrm{d}\theta < \infty$, и $\pi(\theta) = e^{-U(\theta)}/\tilde{C}$ для $\theta \in \mathbb{R}^d$. Предполагается, что функция U известна, но не нормирующая константа $\tilde{C}$. Популярный класс алгоритмов MCMC для генерации данных из подобной плотности π основан на диффузии Ланжевена

$$\mathrm{d}Y_t = -\nabla U(Y_t)\,\mathrm{d}t + \sqrt{2}\mathrm{d}W_t\,, \tag{2.16}$$

где $(W_t)_{t\geq 0}$ - d-мерный Винеровский процесс. При соответствующих технических условиях на U (см. [36]), уравнение (2.16) имеет единственное сильное решение. Более того, распределение Y_t сходится к π с экспоненциальной скоростью, см., например, [63]. Используя схему Эйлера для дискретизации (2.16), мы получим неадаптированный алгоритм Ланжевена (Unadjusted Langevin Algorithm, ULA). Обозначим через $\gamma > 0$ соответствующий шаг дискретизации, и через $(\xi_k)_{k\geq 1}$ последовательность независимых стандартных нормальных d-мерных векторов. Тогда итерации алгоритма ULA определяются соотношением

$$\theta_{k+1} = \theta_k - \gamma\nabla U(\theta_k) + \sqrt{2\gamma}\,\xi_{k+1}\,. \tag{2.17}$$

Анализу теоретических свойств алгоритма ULA посвящены недавние работы [36, 37, 76]. В числе недостатков данного алгоритма отметим, что вычисление градиента ∇U может требовать значительного объема вычислений, например, если $U(\theta) = U_0(\theta) + \sum_{i=1}^K U_i(\theta)$, и число слагаемых K велико. Следуя [33], в этом случае можно использовать алгоритм стохастической градиентной динамики Ланжевена (Stochastic Gradient Langevin Dynamics, SGLD):

$$\theta_{k+1} = \theta_k - \gamma G(\theta_k, S_{k+1}) + \sqrt{2\gamma}\,\xi_{k+1}, \quad G(\theta, S) = \nabla U_0(\theta) + KM^{-1}\sum_{i\in S}\nabla U_i(\theta). \tag{2.18}$$

Случайная величина $S_{k+1} \in \mathsf{S}_M$ называется мини-батчем. При этом S_M есть множество всех подмножеств $\{1,\dots,K\}$ мощности M, и S_{k+1} выбирается независимо от $\mathcal{F}_k = \sigma(\{(\theta_\ell, S_\ell)\}_{0\leq\ell\leq k})$. Для теоретического анализа процедуры ESVM, примененной к наблюдениям, получаемым при помощи (2.18), мы наложим следующие предположения на потенциал U:

(SGLD). *Функция $U(\theta) = U_0(\theta) + \sum_{i=1}^K U_i(\theta)$ удовлетворяет следующим условиям:*

1) *Липшицев градиент: для любого $i \in \{0,\dots,K\}$, U_i непрерывно дифференцируема на $\mathbb{R}^d$ с градиентом, удовлетворяющим условию Липшица с константой $\widetilde{L}_U$;*
2) *Выпуклость: для любого $i \in \{0,\dots,K\}$, U_i является выпуклой;*
3) *Сильная выпуклость: существует константа $m_U > 0$, такая что для любых $\theta, \theta' \in \mathbb{R}^d$ выполняется $U(\theta') \geq U(\theta) + \langle\nabla U(\theta), \theta' - \theta\rangle + (m_U/2)\|\theta' - \theta\|^2$.*

Данные предположения являются классическими для анализа SGLD, см. [77, 78]. Обозначим через $\mathrm{P}_{\mathsf{SGLD}}$ переходное ядро SGLD, а через Υ_M — равномерное распределение над

S_M. Пусть $\overline{\mathrm{P}} := \mathrm{P}_{\mathsf{SGLD}} \otimes \Upsilon_M$, и заметим, что теперь рассматриваемая цепь Маркова имеет вид $X_k = (\theta_k, S_{k+1})$. В Утверждении 3.7 работы [39] показано, что $\overline{\mathrm{P}}$ удовлетворяет условию **(WE)**-1 с $\Delta_1 = \sqrt{1-\gamma m_U}$ и имеет единственное инвариантное распределение. Обозначим соответствующую асимптотическую дисперсию итераций SGLD как $V_\infty^{(\mathsf{SGLD})}(\cdot)$. Тогда мы получаем следующий результат:

Теорема 7. *Пусть $\mathcal{H} \subseteq \mathrm{Lip}_{b,\mathsf{d}}(L, \mathsf{B})$ и выполнено условие **(SGLD)**. Зафиксируем любое $\gamma \in \left(0, \widetilde{L}_U^{-1}(K+1)^{-1}\right)$ и положим $b_n = 2\lceil \log(n)/\log(1/\Delta_1)\rceil$ с $\Delta_1 = \sqrt{1-\gamma m_U}$. Тогда, для любого $\varepsilon > 0$ и $\delta \in (0,1)$, с вероятностью не менее $1-\delta$,*

$$V_\infty^{(\mathsf{SGLD})}(\widehat{h}_{n,\varepsilon}) - \inf_{h\in\mathcal{H}} V_\infty^{(\mathsf{SGLD})}(h) \lesssim \bar{C}_4\, \varepsilon \log(n) + \bar{C}_5 \sqrt{\frac{\log^5(n)}{n}} \left(\frac{|\mathcal{H}_\varepsilon|}{\delta}\right)^{1/\log(n)} + \bar{C}_6\, \frac{\log n}{n}, \tag{2.19}$$

где $\bar{C}_4$, $\bar{C}_5$, и $\bar{C}_6$ — константы, приведенные в Теореме 3.8 работы [39], и первая асимптотическая дисперсия является условной по случайным величинам $X_0, \ldots, X_{n-1}$, по которым была вычислена $\widehat{h}_{n,\varepsilon}$ в (2.11).

Следствие 1. *Предположим, что в условиях Теоремы 7 класс $\mathcal{H}$ выбран параметрическим, то есть существуют такие $C_\rho, \rho > 0$, что $|\mathcal{H}_\varepsilon| \le C_\rho \varepsilon^{-\rho}$ для всех $\varepsilon \in (0,1)$. Тогда с вероятностью не менее $1 - 1/n$ верно*

$$V_\infty^{(\mathsf{SGLD})}(\widehat{h}_{n,\varepsilon}) - \inf_{h\in\mathcal{H}} V_\infty^{(\mathsf{SGLD})}(h) \lesssim n^{-1/2} \log^{5/2}(n),$$

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

Если класс $\mathcal{H}$ построен на основе стейновских контрольных переменных, мы можем обеспечить $\mathcal{H} \subseteq \mathrm{Lip}_{b,\mathsf{d}}(L, \mathsf{B})$ за счет выбора гладких функций ϕ с компактным носителем. Как результат Следствия 1, асимптотические доверительные интервалы, построенные нашим методом для алгоритма SGLD, имеют вид

$$\pi_N(\widehat{h}_{n,\varepsilon}) \pm \mathsf{q}_{1-\alpha/2} \sqrt{\frac{\inf_{h\in\mathcal{H}} V_\infty(h) + Cn^{-1/2}}{N}} \tag{2.20}$$

для некоторой константы $C > 0$. Предполагая, что число тестовых наблюдений $N = n$ (см. Алгоритм 1), ESVM позволяет получить длину асимптотического доверительного интервала порядка $n^{-3/4}$, при условии, что класс $\mathcal{H}$ выбран таким образом, что $\inf_{h\in\mathcal{H}} V_\infty(h)$ достаточно мал. Этот результат интересно сравнить с ранее полученным в [65]. Результат (2.20) сопоставим с "медленными" порядками, полученными в этой работе для случая независимых наблюдений.

Примеры работы предлагаемого алгоритма для улучшения работы SGLD и его модификаций можно найти в работе [39].

Глава 3

Снижение дисперсии на основе мартингальных разложений

3.1. Введение

Основное внимание в Главе 2 уделено задаче снижения дисперсии в задаче (2.1) в случае аналитически известного распределения π. Эта постановка идеально подходит для использования метода контрольных переменных, см. [32, 70, 71, 79, 80]. Особенно популярны при этом стейновские контрольные переменные [81], [71], описанные в Секции 2.4. Основная проблема этого подхода заключается в том, что он требует возможности вычисления π и $\nabla \log \pi$, что не всегда возможно. Однако, оказывается, что если π не известно аналитически и прямое применение стейновских контрольных переменных невозможно, всё ещё можно предложить альтернативные конструкции контрольных переменных, чему и посвящена данная глава. Результаты этой главы опубликованы в [40].

Аналогично Главе 2, наша цель заключается в оценке $\pi(f) := \int_{\mathbb{R}^d} f(x)\pi(\mathrm{d}x)$, где $f : \mathbb{R}^d \to \mathbb{R}$, $f \in \mathrm{L}^2(\pi)$, и π имеет гладкую и всюду положительную плотность относительно меры Лебега. Будем обозначать одной и той же буквой π как саму вероятностную меру, так и её плотность. В дальнейшем будем отталкиваться от оценки $\pi(f)$, имеющей вид $\pi_n^x(f) = \frac{1}{n}\sum_{p=1}^n f(X_p^x)$, где $(X_p^x)_{p\in\mathbb{N}_0}$ - цепь Маркова, удовлетворяющая рекуррентному соотношению

$$X_p^x = \Phi(X_{p-1}^x, \xi_p), \quad p = 1, 2, \ldots, \quad X_0 = x\,. \tag{3.1}$$

Здесь $(\xi_p)_{p=1}^\infty$ - последовательность независимых одинаково распределенных случайных векторов $\xi_p \in \mathbb{R}^m$ с распределением P_ξ, а $\Phi\colon \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ - борелевская функция. Мы используем верхний индекс x у случайных величин X_p^x, чтобы подчеркнуть зависимость от начального условия. Класс цепей Маркова, удовлетворяющих (3.1), является довольно общим (см. [41, теорема 1.3.6]), и покрывает такие алгоритмы MCMC, как неадаптированный алгоритм Ланжевена, и его версию с поправкой Метрополиса (ULA, MALA, см. [35], [36] или [41, глава 2]). В дальнейшем предполагается известной полная ортонормированная система в $\mathrm{L}^2(\mathbb{R}^m, P_\xi)$, обозначаемая как $(\phi_k)_{k\geq 0}$. Поскольку зависимость от начального условия явно отражена в π_n^x и X_p^x, мы используем обозначения E и Var вместо E_x и Var_x, когда это применимо.

3.2. Научный вклад

Ниже перечислены основные результаты данной главы:

- Предложен метод снижения дисперсии для аддитивных функционалов от цепей Маркова, удовлетворяющих рекуррентному представлению (3.1). В отличие от методов, основанных на использовании стейновских контрольных переменных, предлагаемый подход не требует знания инвариантного распределения рассматриваемой марковской цепи.
- Проведен неасимптотический анализ снижения дисперсии, достигаемого нашим алгоритмом в модели с нормальным шумом (см. Раздел 3.4). Данная модель покрывает в частности алгоритмы MCMC на основе динамики Ланжевена [36, 37].

3.3. Мартингальные разложения

Начнем с вывода мартингального разложения для функций от марковских цепей вида (3.1). Затем это разложение будет использовано для построения эффективного алгоритма снижения дисперсии. Пусть $(\phi_k)_{k\in\mathbb{Z}_+}$ является полной ортонормированной системой в $\mathrm{L}^2(\mathbb{R}^m, P_\xi)$ при $\phi_0 \equiv 1$. В частности, $\mathsf{E}[\phi_i(\xi)\phi_j(\xi)] = \delta_{ij}, \quad i,j \in \mathbb{N}$, где $\xi \sim P_\xi$, а случайные величины $\phi_k(\xi)$, $k \geq 1$, имеют нулевое среднее $\mathsf{E}[\phi_k(\xi)] = 0$. Пусть $(\xi_p)_{p\in\mathbb{N}}$ - независимые и одинаково распределённые m-мерные случайные векторы с распределением P_ξ. Обозначим через $(\mathcal{G}_p)_{p\in\mathbb{N}_0}$ фильтрацию, порождённую $(\xi_p)_{p\in\mathbb{N}}$, где $\mathcal{G}_0 = \mathrm{triv}$. Тогда получим следующий результат:

Теорема 8. *Для всех $q \in \mathbb{N}$, $j < q$, ограниченной измеримой функции f и $x \in \mathbb{R}^d$, верно*

$$f(X_q^x) = \mathsf{E}\left[f(X_q^x)\middle|\,\mathcal{G}_j\right] + \sum\nolimits_{k=1}^{\infty}\sum\nolimits_{l=j+1}^{q} \bar{a}_{q-l+1,k}(X_{l-1}^x)\phi_k\left(\xi_l\right)\,, \tag{3.2}$$

где равенство понимается в смысле $\mathrm{L}^2\big(\mathbb{R}^{mq}, P_\xi^{\otimes q}\big)$, и коэффициенты $\bar{a}_{r,k}(y)$ имеют вид

$$\bar{a}_{r,k}(y) = \mathsf{E}\left[f(X_r^y)\phi_k\left(\xi_1\right)\right] \quad r,k \in \mathbb{N}. \tag{3.3}$$

Коэффициенты $\bar{a}_{r,k}$ в (3.3) могут быть также записаны в виде

$$\bar{a}_{r,k}(x) = \mathsf{E}\left[\phi_k\left(\xi\right) Q_{r-1}\left(\Phi(x,\xi)\right)\right]\,, \text{ где } Q_r(y) = \mathsf{E}\left[f(X_r^y)\right], \quad r \in \mathbb{N}. \tag{3.4}$$

Используя формулу (3.2), получим следующее разложение, верное для ограниченной измеримой функции f:

$$\pi_n^x(f) = \tfrac{1}{n}\sum\nolimits_{q=1}^{n}\mathsf{E}[f(X_q^x)] + \tfrac{1}{n}\sum\nolimits_{k=1}^{\infty} M_{n,k}^x\,, \text{ где } M_{n,k}^x = \sum\nolimits_{l=1}^{n}\sum\nolimits_{r=1}^{n-l+1}\bar{a}_{r,k}(X_{l-1}^x)\phi_k(\xi_l)\,. \tag{3.5}$$

Конструкция контрольной переменной. Покажем, как (3.5) можно использовать для построения оценок $\pi(f)$ с дисперсией, меньшей, чем у базовой оценки $\pi_n^x(f)$. Ввиду (3.5) естественно было бы рассмотреть оценку

$$\pi_n^{(x,K)}(f) = \pi_n^x(f) - n^{-1}\sum\nolimits_{k=1}^{K} M_{n,k}^x\,, \tag{3.6}$$

где $K \geq 0$ - параметр срезки. Однако сложность вычисления $\pi_n^{(x,K)}(f)$ зависит *квадратично* от количества наблюдений n. Чтобы преодолеть эту проблему, мы устанавливаем второй уровень срезки n_0 - максимальное количество оцениваемых коэффициентов $\bar{a}_{r,k}, r \in \{1,\ldots,n_0\}$ в представлении для $M_{n,k}^x$. Соответствующая оценка имеет вид

$$\pi_{n,n_0}^{(x,K)}(f) = \pi_n^x(f) - n^{-1}\sum_{k=1}^{K} M_{n,k,n_0}^x\,, \quad M_{n,k,n_0}^x = \sum_{l=1}^{n}\sum_{r=1}^{\{n-\ell+1\}\wedge n_0}\bar{a}_{r,k}(X_{l-1}^x)\phi_k(\xi_l)\,. \tag{3.7}$$

Остаётся оценить $\bar{a}_{r,k}$. Для этого сначала приблизим $Q_r(\cdot)$ из (3.4) функциями вида $Q_{r,\boldsymbol{\beta}}(y) = \sum_{b=1}^{b_0}\boldsymbol{\beta}_b\psi_b(y)$ с некоторыми базисными функциями $\{\psi_b\}_{b=1}^{b_0}$ и $\boldsymbol{\beta} \in \mathcal{B} \subset \mathbb{R}^{b_0}$. Вектор $\boldsymbol{\beta}$ оценива-

ется с помощью метода наименьших квадратов, то есть для $r \in \{0, \dots, n_0 - 1\}$ мы находим

$$\hat{\boldsymbol{\beta}}_r \in \arg\min_{\beta \in \mathbb{R}_0^b} \sum_{s=1}^{n-r} \left| f(X_{r+s}^x) - Q_{r,\boldsymbol{\beta}}(X_s^x) \right|^2 , \tag{3.8}$$

и затем вычисляем оценки $\widehat{a}_{r,k}$ функций $\bar{a}_{r,k}$ согласно формулам

$$\widehat{a}_{r+1,k}(y) = \int \phi_k(z) Q_{\hat{\boldsymbol{\beta}}_r, r}(\Phi(y,z)) P_\xi(\mathrm{d}z) , \tag{3.9}$$

где Φ определено в (3.1). Оценка, полученная подстановкой (3.9) в (3.7), называется оценкой MAD-CV (MArtingale Decomposition Control Variate). Полученная оценка

$$\widehat{\pi}_{n,n_0}^{(x,K)}(f) = \pi_n^x(f) - n^{-1} \sum_{k=1}^{K} \widehat{M}_{n,k,n_0}^x , \quad \widehat{M}_{n,k,n_0}^x = \sum_{l=1}^{n} \sum_{r=1}^{(n-l+1)\wedge n_0} \widehat{a}_{r,k}(X_{l-1}^x) \phi_k(\xi_l) \tag{3.10}$$

остаётся несмещённой для $\pi(f)$ при вычислении на новой траектории, независимой от данных, на которых оценены $\widehat{a}_{r,k}$. Псевдокод процедуры приведен в Алгоритме 2.

Алгоритм 2 Контрольные переменные на основе мартингального разложения (MAD-CV)

Входные данные: Независимые последовательности $\mathbf{X}_N = (X_k^x)_{k=0}^{N-1}$ и $\tilde{\mathbf{X}}_n = (\tilde{X}_k^x)_{k=0}^{n-1}$, удовлетворяющие рекуррентному соотношению (3.1); параметры срезки n_0, K.

1. Решить задачу регрессии на r шагов вперёд для $\hat{\boldsymbol{\beta}}_r$ по формуле (3.8), используя $\mathbf{X}_N$;

2. Вычислить оценки $\widehat{a}_{r,k}$ согласно $\widehat{a}_{r+1,k}(y) = \int \phi_k(z) Q_{\hat{\boldsymbol{\beta}}_r, r}(\Phi(y,z)) P_\xi(\mathrm{d}z)$

Результат: Оценка MAD-CV $\widehat{\pi}_{n,n_0}^{(x,K)}(f)$ для $\pi(f)$, вычисленная по новой траектории $\tilde{\mathbf{X}}_n$.

3.4. Модель на основе нормального шума

Проанализируем алгоритм MAD-CV для цепей Маркова $(X_p^x)_{p \geq 0}$ вида

$$X_p^x = \Phi(X_{p-1}^x, Z_p), \quad Z_p \sim \mathcal{N}(0, \mathrm{I}_d), \quad p = 1, 2, \dots, \quad X_0^x = x \tag{3.11}$$

Для мульти-индекса $\mathbf{k} = (k_i) \in \mathbb{N}_0^d$ обозначим через $\mathbf{H}_\mathbf{k}(x)$ нормированный полином Эрмита на $\mathbb{R}^d$, то есть, $\mathbf{H}_\mathbf{k}(x) := \prod_{i=1}^d H_{k_i}(x_i)$, $x = (x_i) \in \mathbb{R}^d$, где $H_{k_i}(\cdot)$ - одномерные полиномы Эрмита степени k_i. Тогда оценка (3.6) принимает вид

$$\pi_n^{(x,K)}(f) = \pi_n^x(f) - n^{-1} \sum_{0 < \|\mathbf{k}\| \leq K} \sum_{l=1}^{n} \sum_{r=1}^{n-l+1} \bar{a}_{r,k}(X_{l-1}^x) \mathbf{H}_\mathbf{k}(Z_l). \tag{3.12}$$

Применим MAD-CV для оценки математического ожидания относительно стационарного распределения эргодического диффузионного процесса. Пусть $b(x) : \mathbb{R}^d \to \mathbb{R}^d$ - функция сноса, $(\mathsf{W}_t)_{t \geq 0}$ — $d-$мерный Винеровский процесс, и предположим, что стохастическое дифференциальное уравнение

$$\mathrm{d}\mathsf{X}_t^x = -b(\mathsf{X}_t^x)\,\mathrm{d}t + \mathrm{d}\mathsf{W}_t, \; \mathsf{X}_0 = x \tag{3.13}$$

имеет единственное сильное решение $(\mathsf{X}_t^x)_{t \geq 0}$ для любого $x \in \mathbb{R}^d$. Рассмотрим дискретизацию (3.13) по схеме Эйлера-Маруямы, то есть однородную цепь Маркова $(X_k^x)_{k \geq 0}$ с $X_0^x = x \in \mathbb{R}^d$

и задаваемую следующей рекуррентной схемой: для любого $k \in \mathbb{N}$,

$$X_{k+1}^x = X_k^x - \gamma b(X_k^x) + \sqrt{\gamma} Z_{k+1}, \tag{3.14}$$

где $\gamma > 0$ — шаг дискретизации, а $(Z_k)_{k\in\mathbb{N}}$ — последовательность независимых $d-$мерных стандартных нормальных векторов. Заметим, что рекурсия (3.14) является частным случаем общей схемы (3.11) с $\Phi(x,z) = x - \gamma b(x) + \sqrt{\gamma} z$. Наложим некоторые технические условия на функцию сноса b, следуя [82], а именно, предположим, что:

A 4. *Существует $L > 0$, такое что $\|b(x) - b(y)\| \le L\|x - y\|$ для всех $x, y \in \mathbb{R}^d$.*

A 5. *Существует $m > 0$, такое что $\langle b(x) - b(y), x - y\rangle \ge m\|x - y\|^2$ для всех $x, y \in \mathbb{R}^d$.*

При условиях **A** 4 и **A** 5 можно получить следующую оценку дисперсии аддитивных функционалов от марковских цепей вида (3.14):

Теорема 9. *Пусть $(X_k^x)_{k\ge 0}$ — марковская цепь, заданная рекуррентным соотношением (3.14), и предположим, что выполнены условия **A** 4 и **A** 5. Пусть $f : \mathbb{R}^d \to \mathbb{R}$ — $K \times d$ раз непрерывно дифференцируемая функция для некоторого $K \in \mathbb{N}$. Дополнительно предположим, что существуют константы C_f и C_b, такие что для каждого $x \in \mathbb{R}^d$, мультииндекса $\mathbf{k} \in \mathbb{N}_0^d$ с $0 < \|\mathbf{k}\| \le K$, и любого $u \in \{1, \dots, d\}$,*

$$|f^{(\mathbf{k})}(x)| \le C_f, \quad |b_u^{(\mathbf{k})}(x)| \le C_b\,.$$

Тогда, при $0 < \gamma < \min(1/C_b, m/\mathrm{L}^2)$ и любом $n \in \mathbb{N}$,

$$\mathrm{Var}\left[\pi_n^{(x,K)}(f)\right] \lesssim \tfrac{\gamma^{K-2}}{n}.$$

Более того, выбирая $n_0(\gamma) = \lceil K \log \gamma^{-1}/(2m\gamma)\rceil$, для усеченной оценки $\pi_{n,n_0(\gamma)}^{(x,K)}(f)$ выполнено

$$\mathrm{Var}\left[\pi_{n,n_0(\gamma)}^{(x,K)}(f)\right] \lesssim \tfrac{\gamma^{K-2}}{n},$$

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

Для доказательства Теоремы 9 требуется установить скорость убывания коэффициентов $\bar{a}_{r,k}$ с ростом r, а затем связать $\mathrm{Var}\left[\pi_n^{(x,K)}(f)\right]$ с $\bar{a}_{r,k}(\cdot)$, используя соответствующую версию неравенства Пуанкаре для гауссовских векторов [2]. Заметим, что в условиях Теоремы 9 дисперсия оценки $\pi_n^{(x,K)}(f)$ в случае дискретизации диффузии (3.14) удовлетворяет

$$\mathrm{Var}\left[\pi_n^{(x,K)}(f)\right] \lesssim \tfrac{\gamma^{K-2}}{n}\,.$$

В то же время, дисперсия стандартной оценки метода Монте-Карло $\pi_n^x(f)$ имеет порядок $1/(n\gamma)$, и этот порядок в общем случае не может быть улучшен. Таким образом, для $K \ge 2$ и достаточно малых γ мы явно наблюдаем эффект снижения дисперсии.

Замечание 1. *В частном случае неадаптированного алгоритма Ланжевена ULA (см. (2.17)), условия Теоремы 9 выполнены для гладкого и сильно выпуклого потенциала U, то есть для*

$U \in C^2(\mathbb{R}^d)$, такого что $m_U\|x\|^2 \le \langle \nabla^2 U(y)x, x\rangle \le M_U\|x\|^2$ для некоторых $m_U > 0$, $M_U > 0$ и всех $x, y \in \mathbb{R}^d$.

3.5. Численные эксперименты

Сравним снижение дисперсии, достигаемое MAD-CV по сравнению с обычными оценками MCMC на основе ULA (2.17). Рассмотрим целевую плотность π, являющуюся смесью двух d-мерных стандартных нормальных распределений

$$\pi(x) = \frac{1}{2\sqrt{(2\pi)^d}} \left(\mathrm{e}^{-(1/2)\|x-\mu\|^2} + \mathrm{e}^{-(1/2)\|x+\mu\|^2} \right). \tag{3.15}$$

Фиксируем $d = 2$, $\mu = (0.5, 0.5)$ и оценим $\pi(f)$ для $f(x) = x_1 + x_2$ и $f(x) = x_1^2 + x_2^2$. Используя ULA с шагом $\gamma = 0.2$, сгенерируем траекторию длиной 5×10^4 с начальной точкой $X_0 = (1, 1)$. Затем решим задачу метода наименьших квадратов (3.8) с регрессорами $\{x_1, x_2, x_1^2, x_1x_2, x_2^2\}$ для различного выбора срезки $n_0 \in [2, 20]$. Наконец, оценим степень снижения дисперсии относительно вычислительных затрат:

$$\mathcal{R}(f, K, n, n_0) = \frac{\mathrm{cost}\{\pi_n^x(f)\}\,\mathrm{Var}[\pi_n^x(f)]}{\mathrm{cost}\{\pi_{n,n_0}^{(x,K)}\}\,\mathrm{Var}[\pi_{n,n_0}^{(x,K)}(f)]}. \tag{3.16}$$

Заметим, что $\mathcal{R}(f, K, n, n_0) > 1$ указывает на то, что снижение дисперсии является более эффективным с вычислительной точки зрения по сравнению с простым увеличением длины траектории n. Вычислим приближенное значение $\mathcal{R}(f, K, n, n_0)$ на основе 100 независимых траекторий, каждая из которых имеет длину $n = 5 \times 10^4$. При этом

$$\mathrm{cost}\{\pi_{n,n_0}^{(x,K)}(f)\} = \mathrm{cost}\{\pi_n^x(f)\} \times n_0 \times t(K),$$

где $t(K)$ — количество оцененных коэффициентов $\hat{a}_{r,\mathbf{k}}(x)$. Средние значения $\mathcal{R}(f, K, n, n_0)$ для различных n_0 приведены на Рисунке 3.1. Дополнительные численные примеры можно найти в работе [40].

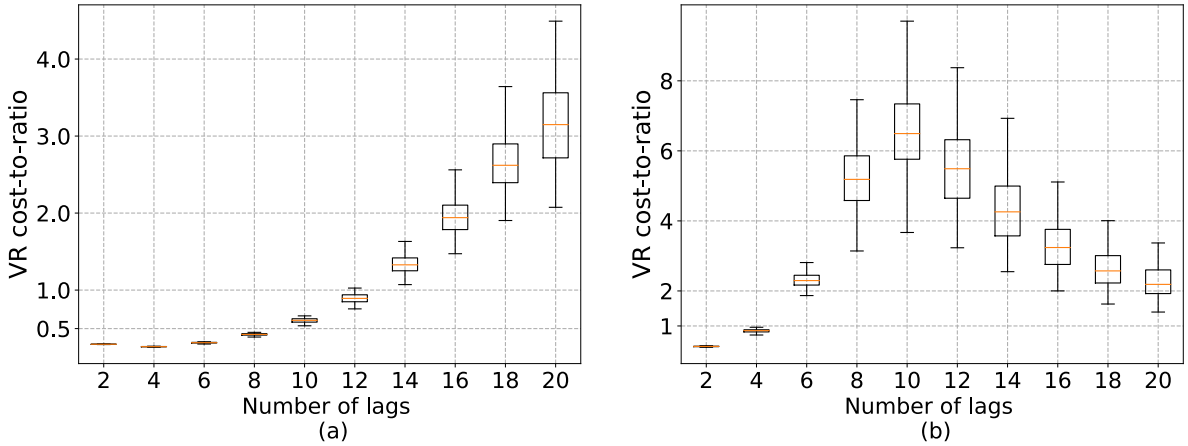

Рис. 3.1. Зависимость $\mathcal{R}(f, K, n, n_0)$ из (3.16) от уровня срезки n_0 для смеси (3.15) двумерных нормальных распределений. Рисунок (a) : $f(x) = x_1 + x_2$, рисунок (b) : $f(x) = x_1^2 + x_2^2$.

Заключение

1. В Главе 1 получены новые аналоги неравенств Розенталя и Бернштейна для аддитивных функционалов от эргодических марковских цепей, которые сходятся к стационарному распределению с экспоненциальной скоростью либо в V-норме полной вариации, либо в полуметрике Канторовича-Васерштейна. Использованный метод доказательства основан на кумулянтном разложении и связи между кумулянтами и центральными моментами, устанавливаемой с помощью формулы Леонова-Ширяева.

2. В Главе 2 предложено обобщение метода снижения дисперсии с использованием контрольных переменных на случай последовательностей зависимых случайных величин, удовлетворяющих условию стационарности ковариаций, произведен анализ избыточной асимптотической дисперсии алгоритма. Получены неравенства концентрации для квадратичных форм от функций от цепей Маркова, удовлетворяющих условию равномерной геометрической эргодичности в метрике Канторовича-Васерштейна $\mathbf{W}_{\mathsf{d},1}$. Полученные результаты применены к алгоритмам MCMC на основе динамики Ланжевена с использованием стохастических градиентов (SGLD).

3. В Главе 3 предложен новый подход к снижению дисперсии для аддитивных функционалов от марковских цепей на основе дискретного мартингального разложения. Для специального случая модели нормального шума, покрывающей неадаптированный алгоритм Ланжевена (ULA), произведен неасиптотический анализ снижения дисперсии, достигаемого предложенным алгоритмом. Теоретический анализ основан на неравенстве Пуанкаре для гауссовских случайных векторов.

Список литературы

1. M. Ledoux. *The Concentration of Measure Phenomenon*, volume 89. AMS Surveys and Monographs, 2001.
2. S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.
3. Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
4. Ron Dorfman and Kfir Yehuda Levy. Adapting to mixing time in stochastic optimization with Markovian data. In *International Conference on Machine Learning*, pages 5429–5446. PMLR, 2022.
5. Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.
6. G. O. Roberts and J. S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probab. Surv.*, 1:20–71, 2004.
7. Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to MCMC for machine learning. *Machine learning*, 50:5–43, 2003.
8. Bernard Bercu, Bernard Delyon, and Emmanuel Rio. *Concentration inequalities for sums and martingales*. Springer, 2015.
9. Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Number 47 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
10. Florence Merlevède, Magda Peligrad, and Emmanuel Rio. A bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151(3-4):435–474, 2011.
11. Emmanuel Rio et al. *Asymptotic theory of weakly dependent random processes*, volume 80. Springer, 2017.
12. Radoslaw Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability*, 13:1000–1034, 2008.
13. Błażej Miasojedow. Hoeffding's inequalities for geometrically ergodic markov chains on general state space. *Statistics & Probability Letters*, 87:115–120, 2014.
14. Radosław Adamczak and Witold Bednorz. Exponential concentration inequalities for additive functionals of Markov chains. *ESAIM: Probability and Statistics*, 19:440–481, 2015.
15. Jianqing Fan, Bai Jiang, and Qiang Sun. Hoeffding's inequality for general markov chains and its applications to statistical learning. *The Journal of Machine Learning Research*, 22(1):6185–6219, 2021.
16. Alexandros G. Dimakis, Soummya Kar, José M. F. Moura, Michael G. Rabbat, and Anna Scaglione. Gossip algorithms for distributed signal processing. *Proceedings of the IEEE*, 98(11):1847–1864, 2010.
17. Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with

convergence rate o(1/n). In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

18. Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pages 1691–1692. PMLR, 2018.
19. Stéphan JM Clémençon. Moment and probability inequalities for sums of bounded additive functionals of regular Markov chains via the Nummelin splitting technique. *Statistics & probability letters*, 55(3):227–238, 2001.
20. Michał Lemańczyk. General Bernstein-like inequality for additive functionals of Markov chains. *Journal of Theoretical Probability*, 34(3):1426–1454, 2021.
21. R. Bentkus and R. Rudzkis. Exponential estimates for the distribution of random variables. *Litovsk. Mat. Sb.*, 20(1):15–30, 216, 1980.
22. V. P. Leonov and A. N. Sirjaev. On a method of semi-invariants. *Theor. Probability Appl.*, 4:319–329, 1959.
23. Reuven Y. Rubinstein and Dirk P. Kroese. *Simulation and the Monte Carlo Method*, volume 10. John Wiley & Sons, 2016.
24. Emmanuel Gobet. *Monte-Carlo Methods and Stochastic Processes*. CRC Press, Boca Raton, FL, 2016.
25. Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, New York, 1999.
26. Paul Glasserman. *Monte Carlo Methods in Financial Engineering*, volume 53. Springer Science & Business Media, 2013.
27. Chris J Oates, Mark Girolami, and Nicolas Chopin. Control functionals for monte carlo integration. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):695–718, 2017.
28. Rie Johnson and Tong Zhang. Accelerating Stochastic Gradient Descent Using Predictive Variance Reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.
29. Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A Fast Incremental Gradient Method with Support for Non-Strongly Convex Composite Objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.
30. Niladri S Chatterji, Nicolas Flammarion, Yi-An Ma, Peter L Bartlett, and Michael I Jordan. On the Theory of Variance Reduction for Stochastic Gradient Monte Carlo. *Proceedings of Machine Learning Research*, 80, 2018.
31. Jack Baker, Paul Fearnhead, Emily B Fox, and Christopher Nemeth. Control variates for stochastic gradient MCMC. *Statistics and Computing*, 29(3):599–615, 2019.
32. D. Belomestny, L. Iosipoi, E. Moulines, A. Naumov, and S. Samsonov. Variance reduction for Markov chains with application to MCMC. *Statistics and Computing*, 30(4):973–997, 2020.
33. M. Welling and Y. W. Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011.

34. Denis Belomestny, Stefan Häfner, and Mikhail Urusov. Variance reduction for discretised diffusions via regression. *Journal of Mathematical Analysis and Applications*, 458:393–418, 2018.
35. K. L. Mengersen and R. L. Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics*, 24(1):101–121, 02 1996.
36. Arnak 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.
37. A. Durmus and É. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 2017.
38. Alain Durmus, Eric Moulines, Alexey Naumov, and Sergey Samsonov. Probability and moment inequalities for additive functionals of geometrically ergodic Markov chains. *Journal of Theoretical Probability*, pages 1–50, 2024.
39. Denis Belomestny, Leonid Iosipoi, Eric Moulines, Alexey Naumov, and Sergey Samsonov. Variance reduction for dependent sequences with applications to stochastic gradient MCMC. *SIAM/ASA Journal on Uncertainty Quantification*, 9(2):507–535, 2021.
40. Denis Belomestny, Eric Moulines, and Sergey Samsonov. Variance reduction for additive functionals of Markov chains via martingale representations. *Statistics and Computing*, 32(1):16, 2022.
41. R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov chains*. Springer Series in Operations Research and Financial Engineering. Springer, 2018.
42. L. Saulis and V. A. Statulevičius. *Limit theorems for large deviations*, volume 73 of *Mathematics and its Applications (Soviet Series)*. Kluwer Academic Publishers Group, Dordrecht, 1991. Translated and revised from the 1989 Russian original.
43. Iosif Pinelis. Optimum Bounds for the Distributions of Martingales in Banach Spaces. *The Annals of Probability*, 22(4):1679 – 1706, 1994.
44. Haskell P. Rosenthal. On the subspaces of L^p $(p > 2)$ spanned by sequences of independent random variables. *Israel J. Math.*, 8:273–303, 1970.
45. Paul Doukhan and Michael H Neumann. Probability and moment inequalities for sums of weakly dependent random variables, with applications. *Stochastic Processes and their Applications*, 117(7):878–903, 2007.
46. Hermann Thorisson. On maximal and distributional coupling. *The Annals of Probability*, pages 873–876, 1986.
47. Peter W Glynn and Dirk Ormoneit. Hoeffding's inequality for uniformly ergodic markov chains. *Statistics & probability letters*, 56(2):143–146, 2002.
48. Katalin Marton. A measure concentration inequality for contracting Markov chains. *Geometric & Functional Analysis GAFA*, 6(3):556–571, 1996.
49. Jérôme Dedecker, Sébastien Gouëzel, et al. Subgaussian concentration inequalities for geometrically ergodic Markov chains. *Electronic Communications in Probability*, 20, 2015.
50. A. Joulin and Y. Ollivier. Curvature, concentration and error estimates for Markov chain Monte Carlo. *The Annals of Probability*, 38(6):2418 – 2442, 2010.
51. Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral

methods. *Electronic Journal of Probability*, 20(none):1 – 32, 2015.

52. J. Fan, B. Jiang, and Q. Sun. Hoeffding's lemma for Markov chains and its applications to statistical learning. *arXiv preprint arXiv:1802.00211*, 2018.
53. J. Fan, B. Jiang, and Q. Sun. Bernstein's inequality for general Markov chains. *arXiv preprint arXiv:1805.10721*, 2018.
54. Ioannis Kontoyiannis and Sean P Meyn. Geometric ergodicity and the spectral gap of non-reversible Markov chains. *Probability Theory and Related Fields*, 154(1-2):327–339, 2012.
55. Patrice Bertail and Stéphan Clémençon. Sharp bounds for the tails of functionals of Markov chains. *Theory of Probability & Its Applications*, 54(3):505–515, 2010.
56. Gabriela Ciołek and Patrice Bertail. New Bernstein and Hoeffding type inequalities for regenerative Markov chains. *Latin American journal of probability and mathematical statistics*, 16:1–19, 02 2019.
57. Krishna B Athreya and Peter Ney. A new approach to the limit theory of recurrent Markov chains. *Transactions of the American Mathematical Society*, 245:493–501, 1978.
58. E. Nummelin. A splitting technique for harris recurrent Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 43:309–318, 1978.
59. Paul Doukhan and Sana Louhichi. A new weak dependence condition and applications to moment inequalities. *Stochastic Process. Appl.*, 84(2):313–342, 1999.
60. M. Hairer, J.C. Mattingly, and M. Scheutzow. Asymptotic coupling and a general form of Harris' theorem with applications to stochastic delay equations. *Probability theory and related fields*, 149(1-2):223–259, 2011.
61. M. Hairer, A.M. Stuart, and S.J. Vollmer. Spectral gaps for Metropolis-Hastings algorithms in infinite dimensions. *Ann. Appl. Probab.*, 24:2455–290, 2014.
62. Nikolai Sergeevich Bakhvalov. On the optimality of linear methods for operator approximation in convex classes of functions. *USSR Computational Mathematics and Mathematical Physics*, 11(4):244–249, 1971.
63. G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.
64. Leah F South, Chris J Oates, Antonietta Mira, and Christopher Drovandi. Regularized zero-variance control variates. *Bayesian Analysis*, 1(1):1–24, 2022.
65. D Belomestny, L Iosipoi, and N Zhivotovskiy. Variance reduction via empirical variance minimization: convergence and complexity. *arXiv preprint, arXiv:1712.04667*, 2017.
66. Zhuo Sun, Chris J Oates, and François-Xavier Briol. Meta-learning control variates: Variance reduction with limited data. In *Uncertainty in Artificial Intelligence*, pages 2047–2057. PMLR, 2023.
67. L F South, T Karvonen, C Nemeth, M Girolami, and C J Oates. Semi-exact control functionals from Sard's method. *Biometrika*, 109(2):351–367, 09 2021.
68. Leah F South, Marina Riabiz, Onur Teymur, and Chris J Oates. Postprocessing of mcmc. *Annual Review of Statistics and Its Application*, 9:529–555, 2022.
69. James M. Flegal and Galin L. Jones. Batch means and spectral variance estimators in Markov chain monte carlo. *Ann. Statist.*, 38(2):1034–1070, 04 2010.
70. Roland Assaraf and Michel Caffarel. Zero-variance principle for Monte Carlo algorithms.

Physical review letters, 83(23):4682, 1999.

71. Antonietta Mira, Reza Solgi, and Daniele Imparato. Zero variance Markov chain Monte Carlo for Bayesian estimators. *Statistics and Computing*, 23(5):653–662, 2013.
72. Chris J. Oates, Jon Cockayne, François-Xavier Briol, and Mark Girolami. Convergence rates for a class of estimators based on Stein's method. *Bernoulli*, 25(2):1141 – 1159, 2019.
73. Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of u-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.
74. Quentin Duchemin, Yohann De Castro, and Claire Lacour. Concentration inequality for u-statistics of order two for uniformly ergodic markov chains. *Bernoulli*, 29(2):929–956, 2023.
75. Quentin Duchemin, Yohann De Castro, and Claire Lacour. Three rates of convergence or separation via u-statistics in a dependent framework. *Journal of Machine Learning Research*, 23(201):1–59, 2022.
76. Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm. *Bernoulli*, 25(4A):2854–2882, 11 2019.
77. Yi-An Ma, Tianqi Chen, and Emily Fox. A Complete Recipe for Stochastic Gradient MCMC. In *Advances in Neural Information Processing Systems*, pages 2917–2925, 2015.
78. Arnak S. Dalalyan and Avetik G. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stoch. Proc. Appl.*, 129(12):5278–5311, 2019.
79. Shane G Henderson. *Variance reduction via an approximating Markov process*. PhD thesis, Stanford University, 1997.
80. P. Dellaportas and I. Kontoyiannis. Control variates for estimation based on reversible Markov chain Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1), 2012.
81. Roland Assaraf and Michel Caffarel. Zero-variance principle for Monte Carlo algorithms. *Phys. Rev. Lett.*, 83(23):4682–4685, 1999.
82. Valentin De Bortoli and Alain Durmus. Convergence of diffusions and their discretizations: from continuous to discrete processes and back. *arXiv preprint arXiv:1904.09808*, 2019.