



Гауссовская аппроксимация для линейной стохастической аппроксимации с приложением к обучению с подкреплением

Даниль Эльдарович Гайнанов¹, Илья Валерьевич Левин^{2✉}

^{1,2} Высшая школа экономики, Москва, Россия

[✉] ivlevin@hse.ru

Аннотация. Рассматривается задача статистического вывода для решения системы линейных уравнений $\mathbf{A}\theta^* = \mathbf{b}$, в которой матрица $\mathbf{A}$ и вектор $\mathbf{b}$ неизвестны, а вместо них доступны зашумлённые наблюдения. Исследуется метод линейной стохастической аппроксимации с постоянным шагом и марковским шумом. Усреднение Поляка–Рупперта (PR) снижает дисперсию, но сохраняет стационарное смещение порядка α , ухудшающее нормальную аппроксимацию. Для устранения ведущего смещения используется экстраполяция Ричардсона–Ромберга: средние двух рекурсий с шагами α и 2α , использующих одну и ту же траекторию марковской цепи, объединяются в одну оценку. Показано, что такая коррекция улучшает нормальную аппроксимацию без изменения предельной дисперсии.

Получена неасимптотическая оценка в расстоянии Колмогорова для скалярных проекций PR-усреднений с экстраполяцией Ричардсона–Ромберга. Для траектории длины n при выборе шага порядка $n^{-1/2}$ полученная верхняя граница убывает как $n^{-1/4}$ с точностью до логарифмических множителей. В качестве приложения исследуемой теории рассматривается построение доверительных интервалов для параметров, оцениваемых алгоритмом TD(0) в обучении с подкреплением. Численные эксперименты на синтетических данных подтверждают коррекцию среднего и улучшение покрытия доверительных интервалов без заметного увеличения их ширины.

Ключевые слова и фразы: линейная стохастическая аппроксимация, гауссовская аппроксимация, экстраполяция Ричардсона–Ромберга, доверительные интервалы, обучение с подкреплением

Для цитирования: Гайнанов Д. Э., Левин И. В. Гауссовская аппроксимация для линейной стохастической аппроксимации с приложением к обучению с подкреплением // Программные системы: теория и приложения. 2026. Т. 17. № 4(73). С. 33–56. https://psta.psir.ru/read/psta2026_4_33-56.pdf

Введение

В задаче линейной стохастической аппроксимации (LSA) требуется решить систему линейных уравнений $\bar{\mathbf{A}}\theta^* = \bar{\mathbf{b}}$, где $\theta^* \in \mathbb{R}^d$. Матрица $\bar{\mathbf{A}}$ и вектор $\bar{\mathbf{b}}$ неизвестны; вместо них доступны зашумлённые наблюдения $\{\mathbf{A}(Z_k), \mathbf{b}(Z_k)\}_{k \in \mathbb{N}}$, где $\{Z_k\}_{k \in \mathbb{N}}$ — последовательность случайных величин со значениями в измеримом пространстве $(Z, \mathcal{Z})$. В настоящей работе предполагается, что $\{Z_k\}_{k \in \mathbb{N}}$ является марковской цепью с инвариантным распределением π и

$$\mathbb{E}_\pi[\mathbf{A}(Z)] = \bar{\mathbf{A}}, \quad \mathbb{E}_\pi[\mathbf{b}(Z)] = \bar{\mathbf{b}}.$$

Такие задачи возникают, в частности, в обучении с подкреплением, например при анализе алгоритма TD-обучения (Temporal-Difference Learning) [1]. Этот алгоритм является одним из основных методов обучения с подкреплением.

Рассмотрим оценивание функции ценности фиксированной политики методом TD(0) с линейной аппроксимацией. Помимо точечной оценки, для количественной оценки её статистической точности нас интересуют доверительные интервалы для параметров линейного приближения. При последовательной обработке длинных траекторий такие интервалы позволяют оценивать статистическую точность оценок параметров по большому числу зависимых наблюдений.

Для приближённого решения этой системы рассмотрим алгоритм Роббинса—Монро [2]. Зададим последовательность шагов $\{\alpha_k\}_{k \in \mathbb{N}}$, индекс начала усреднения $n_0 \in \mathbb{N}$ и начальную точку $\theta_0 \in \mathbb{R}^d$. Итерации $\{\theta_k\}_{k \in \mathbb{N}}$ и их средние $\{\bar{\theta}_n\}_{n \geq n_0+1}$ определяются соотношениями

$$\begin{aligned} \theta_k &= \theta_{k-1} - \alpha_k \{\mathbf{A}(Z_k)\theta_{k-1} - \mathbf{b}(Z_k)\}, \quad k \geq 1, \\ (1) \quad \bar{\theta}_n &= (n - n_0)^{-1} \sum_{k=n_0}^{n-1} \theta_k, \quad n \geq n_0 + 1. \end{aligned}$$

Точное представление линейного TD(0) в форме LSA (1) описано в Приложении В.2.

Величина $\bar{\theta}_n$ является средним Поляка—Рупперта (PR) [3–5]. Такое усреднение широко используется для улучшения статистических свойств алгоритмов стохастической аппроксимации. Хотя классическая теория в значительной степени основана на убывающих шагах, на практике привлекательным является режим постоянного шага $\alpha_k \equiv \alpha$: такой шаг проще настраивать, а влияние начального условия затухает экспоненциально быстро [6].

Вместе с тем при марковском шуме постоянный шаг в общем случае приводит к стационарному смещению порядка α , которое не устраняется

усреднением Поляка–Рупшперта [7, 8]. Именно режим постоянного шага является основным объектом настоящей работы.

Для статистического вывода важна не только величина ошибки $\bar{\theta}_n - \theta^*$, но и её распределение. При стандартных условиях на шум и убывающий шаг PR-среднее удовлетворяет центральной предельной теореме [5, 9]: для любого фиксированного направления $u \in \mathbb{R}^d$

$$(2) \quad T_n(u) := \frac{\sqrt{n} u^\top (\bar{\theta}_n - \theta^*)}{\sigma(u)} \xrightarrow{d} \mathcal{N}(0, 1),$$

где $\sigma^2(u)$ — асимптотическая дисперсия соответствующей проекции. Доказательство такой сходимости опирается на выделение ведущего линейного члена и контроль остатка. При марковском шуме это приводит к изучению сумм функций от состояний марковской цепи. Уточнения предельных теорем и оценки скорости нормального приближения для таких сумм получены, в частности, в работах [10, 11]. Спектральный подход к доказательству центральной предельной теоремы для цепей с общим пространством состояний, в том числе без условия равномерной эргодичности, развит в работе [12]. В настоящей работе предполагается равномерная геометрическая эргодичность цепи.

Эта ЦПТ лежит в основе построения доверительных интервалов. При конечном n их точность определяется тем, насколько закон $T_n(u)$ близок к стандартному нормальному закону. В расстоянии Колмогорова такая близость измеряется величиной

$$d_K(T_n(u), \mathcal{N}(0, 1)) = \sup_{x \in \mathbb{R}} |\mathbb{P}\{T_n(u) \leq x\} - \Phi(x)|.$$

Поэтому для неасимптотического анализа алгоритма естественно получить явную верхнюю оценку на это расстояние, стремящуюся к нулю при росте n . В режиме постоянного шага основным препятствием для такой нормальной аппроксимации является стационарное смещение, сдвигающее среднее нормированной статистики.

Для усреднённого TD-обучения с убывающим шагом в [13] получена неасимптотическая оценка в расстоянии Вассерштейна; оптимизация приведённой там границы по скорости убывания шага даёт порядок $\sqrt{\log n} n^{-1/6}$. Для общей LSA с убывающим шагом и марковским шумом в [9] установлена оценка Берри–Эссеена для скалярных проекций PR-среднего: при $\alpha_k \asymp k^{-3/4}$ получена скорость порядка $n^{-1/4}$ с точностью до логарифмических множителей. Эта оценка имеет более быструю полиномиальную зависимость от n , однако прямое сравнение с [13] затруднено, поскольку результаты сформулированы в разных метриках.

В постановке с постоянным шагом в [14] получена явная неасимптотическая оценка расстояния Вассерштейна между законом нормированной

ошибки в стационарном режиме и гауссовским законом. Этот результат относится к распределению последней итерации после выхода в стационарный режим, а не к распределению среднего, построенного по траектории конечной длины. Наиболее близка к нашей постановке работа [7]: для PR-среднего LSA с марковским шумом и постоянным шагом в ней установлена асимптотическая ЦПТ с центрированием относительно стационарного среднего. Для уменьшения смещения при построении доверительных интервалов в той же работе предложена экстраполяция Ричардсона–Ромберга (RR). Экстраполяция по шагу применяется также для повышения порядка слабой сходимости численных методов решения стохастических дифференциальных уравнений; пример такого применения рассмотрен в [15].

Как показывают [6–8], усреднение Поляка–Рупшперта само по себе не устраняет стационарное смещение при постоянном шаге. Для сокращения его ведущего члена можно использовать экстраполяцию Ричардсона–Ромберга. Для этого две рекурсии (1), построенные на общем шуме и с шагами α и 2α , объединяются в комбинацию

$$2\bar{\theta}^{(\alpha)} - \bar{\theta}^{(2\alpha)}.$$

Такая комбинация сокращает ведущий член смещения порядка α . В [8] для соответствующей RR-оценки получены оценки моментов высокого порядка, позволяющие контролировать зависимость остаточного члена от α .

Цель настоящей работы — получить на основе этого контроля моментов неасимптотическую нормальную аппроксимацию для RR-оценки. В отличие от описанных выше результатов мы устанавливаем оценку в расстоянии Колмогорова для RR-комбинации PR-усреднённых траекторий с явной зависимостью от длины траектории n и шага α . Насколько нам известно, ранее неасимптотическая оценка такого типа в расстоянии Колмогорова для RR-усреднения LSA с постоянным шагом получена не была. Выбор шага и скорость, получаемая после его оптимизации, обсуждаются после формулировки основной теоремы. Численные эксперименты иллюстрируют лежащий в основе результата механизм: RR корректирует центр нормированной статистики практически без изменения ширины доверительного интервала.

1. Линейная стохастическая аппроксимация с марковским шумом и экстраполяция Ричардсона–Ромберга

Пусть $(Z_k)_{k \geq 0}$ — однородная марковская цепь с переходным ядром P на полном сепарабельном метрическом пространстве Z с борелевской сигма-алгеброй $\mathcal{Z}$ и инвариантным распределением π . Линейная стохастическая

аппроксимация с постоянным шагом $\alpha > 0$ имеет вид

$$(3) \quad \theta_k^{(\alpha)} = \theta_{k-1}^{(\alpha)} - \alpha(\mathbf{A}(Z_k)\theta_{k-1}^{(\alpha)} - \mathbf{b}(Z_k)).$$

Обозначим $\mathbb{E}_\pi[\mathbf{A}(Z)] = \bar{\mathbf{A}}$ и $\mathbb{E}_\pi[\mathbf{b}(Z)] = \bar{\mathbf{b}}$. Искомый параметр θ^* задаётся уравнением $\bar{\mathbf{A}}\theta^* = \bar{\mathbf{b}}$. Постоянный шаг обеспечивает геометрическое затухание начальной ошибки, но порождает стационарное смещение относительно θ^* .

Нам понадобятся три стандартных условия, совпадающие с базовыми условиями [8]: на перемешивание цепи, устойчивость системы и шум в оптимальной точке.

- (1) Цепь равномерно геометрически эргодична. В частности, её коэффициент Добрушина удовлетворяет

$$\Delta(P^k) := \sup_{z, z'} \frac{1}{2} \|P^k(z, \cdot) - P^k(z', \cdot)\|_{\text{TV}} \leq (1/4)^{\lfloor k/t_{\text{mix}} \rfloor}, \quad k \geq 1.$$

t_{mix} — число шагов, за которое коэффициент Добрушина уменьшается не менее чем до $1/4$; условие также влечёт единственность π .

- (2) Матрица $-\bar{\mathbf{A}}$ гурвицева, а функции $\mathbf{A}$ и $\mathbf{b}$ ограничены. Существует матрица $Q = Q^\top \succ 0$ и числа $a > 0$, $\alpha_\infty > 0$ такие, что, после возможного уменьшения α_∞ ,

$$(4) \quad \|I - \alpha\bar{\mathbf{A}}\|_Q^2 \leq 1 - \alpha a, \quad \alpha a \leq \frac{1}{2}, \quad 0 \leq \alpha \leq \alpha_\infty.$$

Гурвицевость обеспечивает единственность θ^* , а (4) даёт

$$\|(I - \alpha\bar{\mathbf{A}})^k\|_Q \leq (1 - \alpha a)^{k/2}$$

и тем самым контролирует вклад инициализации.

- (3) Центрированный шум в оптимальной точке $\varepsilon(z) = \tilde{\mathbf{A}}(z)\theta^* - \tilde{\mathbf{b}}(z)$ ограничен, где мы полагаем

$$\tilde{\mathbf{A}}(z) = \mathbf{A}(z) - \bar{\mathbf{A}}, \quad \tilde{\mathbf{b}}(z) = \mathbf{b}(z) - \bar{\mathbf{b}},$$

и по построению $\mathbb{E}_\pi[\varepsilon(Z)] = 0$.

Здесь $\|x\|_Q^2 = x^\top Q x$, матричная Q -норма индуцирована этой векторной нормой, а без индекса используется евклидова или спектральная норма. Все параметры задачи в дальнейшем считаются фиксированными. Условия 1 и 3, в частности, гарантируют абсолютную сходимость ряда, определяющего ковариационную матрицу

$$(5) \quad \Sigma_\varepsilon = \mathbb{E}_\pi[\varepsilon(Z_0)\varepsilon(Z_0)^\top] + 2 \sum_{\ell \geq 1} \mathbb{E}_\pi[\varepsilon(Z_0)\varepsilon(Z_\ell)^\top].$$

Будем считать матрицу Σ_ε положительно определённой и обозначим

$$(6) \quad C_\varepsilon := \lambda_{\min}(\Sigma_\varepsilon)^{-1/2}.$$

Асимптотическая ковариационная матрица усреднённой оценки имеет вид

$$(7) \quad \Sigma_\infty = \bar{\mathbf{A}}^{-1} \Sigma_\varepsilon \bar{\mathbf{A}}^{-\top}.$$

Это предельная ковариационная матрица РР-усреднения. Для направления $u \in \mathbb{S}^{d-1}$ положим

$$(8) \quad \sigma^2(u) := u^\top \Sigma_\infty u.$$

Пусть $0 \leq n_0 < n$. После разогрева усредняются последние $n - n_0$ итераций:

$$(9) \quad \bar{\theta}_{n,n_0}^{(\alpha)} = \frac{1}{n - n_0} \sum_{k=n_0}^{n-1} \theta_k^{(\alpha)}.$$

Параметр разогрева n_0 позволяет с экспоненциальной по n_0 скоростью забывать начальное состояние. Усреднение Поляка–Рупперта позволяет получать оптимальную асимптотическую дисперсию в задачах стохастической аппроксимации [5]. Две рекурсии (3) запускаются из одной детерминированной точки θ_0 , используют одну траекторию (Z_k) и объединяются в следующую оценку

$$(10) \quad \bar{\theta}_{n,n_0}^{(\alpha, \text{RR})} = 2\bar{\theta}_{n,n_0}^{(\alpha)} - \bar{\theta}_{n,n_0}^{(2\alpha)}.$$

Для достаточно малого α совместная цепь $(\theta_k^{(\alpha)}, Z_{k+1})$ имеет единственное стационарное распределение [8, Теорема 1]. Если $\theta_\infty^{(\alpha)}$ обозначает его θ -координату, то

$$(11) \quad \mathbb{E}[\theta_\infty^{(\alpha)}] = \theta^* + \alpha\Delta + O(\alpha^{3/2}),$$

где Δ — вектор, не зависящий от α . Следовательно,

$$2\mathbb{E}[\theta_\infty^{(\alpha)}] - \mathbb{E}[\theta_\infty^{(2\alpha)}] = \theta^* + O(\alpha^{3/2}).$$

Таким образом, RR сокращает смещение порядка α . При $\alpha \asymp n^{-1/2}$ нормированное смещение обычного среднего имеет порядок $\sqrt{n}\alpha \asymp 1$, а после RR — порядок $\sqrt{n}\alpha^{3/2} \asymp n^{-1/4}$.

В следующем разделе будет использоваться та же схема разогрева, что и в [8, Теорема 2]. Считаем n чётным и полагаем

$$(12) \quad n_0 = n/2.$$

Далее для краткости будем писать

$$(13) \quad \bar{\theta}_n^{(\alpha)} := \bar{\theta}_{n,n/2}^{(\alpha)}, \quad \bar{\theta}_n^{(\alpha, \text{RR})} := \bar{\theta}_{n,n/2}^{(\alpha, \text{RR})}.$$

Таким образом, первые $n/2$ итераций отбрасываются, а последние $n/2$ усредняются. Для ошибки

$$\bar{\mathbf{A}}(\bar{\theta}_n^{(\alpha, \text{RR})} - \theta^*)$$

непосредственно применима оценка [8, Теорема 2]. В следующем разделе эта оценка вместе с мартингальной аппроксимацией марковского шума используется для получения неасимптотической ЦПТ для итераций Ричардсона–Ромберга.

2. Основной результат: гауссовская аппроксимация

В этом разделе запись $x \lesssim_{\log n} y$ означает, что $x \leq C(\log n)^q y$ для некоторых не зависящих от n и α констант $C, q \geq 0$. Иными словами, такое неравенство выполнено с точностью до постоянного и логарифмического множителей; C может зависеть от фиксированных параметров задачи.

Для $p \geq 2$ через α_p^L обозначим верхнюю границу шага из [8, Теорема 2]. Положим $L_\alpha = \log(1/(\alpha a))$; через $c_{\text{tr}} > 0$ обозначим постоянную, определяющую скорость затухания начальной ошибки.

Для $u \in \mathbb{S}^{d-1}$ обозначим нормированную ошибку через

$$T_{n,\alpha}^{\text{RR}}(u) := \frac{\sqrt{n/2} u^\top (\bar{\theta}_n^{(\alpha, \text{RR})} - \theta^*)}{\sigma(u)}.$$

Наша цель — получить неасимптотическую оценку расстояния

$$d_K(T_{n,\alpha}^{\text{RR}}(u), \mathcal{N}(0, 1)).$$

Точная правая часть оценки, включая логарифмические множители, далее записывается через величину

$$(14) \quad \begin{aligned} r_{n,\alpha} := & n^{-1/4} + \frac{L_\alpha^{1/2}}{\sqrt{\alpha n}} + \sqrt{\alpha} + \sqrt{n} \alpha^{3/2} L_\alpha^{3/2} \\ & + \frac{\sqrt{\alpha} L_\alpha^{3/2}}{\sqrt{n}} + \frac{1}{\alpha \sqrt{n}} \exp(-c_{\text{tr}} \alpha n). \end{aligned}$$

Кроме того, обозначим

$$\alpha_n^L := \alpha_{\max\{2, \lceil \log n \rceil\}}^L.$$

ТЕОРЕМА 1 (Гауссовская аппроксимация). *Пусть выполнены условия 1–3 и n чётно. Предположим также, что $\Sigma_\varepsilon \succ 0$, и зафиксируем $u \in \mathbb{S}^{d-1}$. Если n достаточно велико и $0 < \alpha \leq \alpha_n^L$, то*

$$(15) \quad d_K\left(\frac{\sqrt{n/2} u^\top (\bar{\theta}_n^{(\alpha, \text{RR})} - \theta^*)}{\sigma(u)}, \mathcal{N}(0, 1)\right) \lesssim_{\log n} C_G r_{n,\alpha}.$$

Величина $r_{n,\alpha}$ определена в (14). Здесь $C_G \geq 1$ не зависит от n , α , u и начального распределения. Эта константа может зависеть от размерности d , времени смешивания t_{mix} , параметров устойчивости Q , a и α_∞ , C_ε , равномерных верхних границ для $\mathbf{A}$, $\mathbf{b}$ и ε , а также от $\|\theta_0 - \theta^*\|$. Оценка равномерна по $u \in \mathbb{S}^{d-1}$ и начальному распределению ξ .

Доказательство теоремы приведено в [Приложении](#).

При условиях Теоремы 1 и $L_\alpha \lesssim \log n$ полученная оценка имеет

структуру

$$(16) \quad d_K(T_{n,\alpha}^{\text{RR}}(u), \mathcal{N}(0, 1)) \lesssim_{\log n} \underbrace{n^{-1/4}}_{\text{мартингальная ЦПТ}} + \underbrace{(\alpha n)^{-1/2}}_{\text{слишком малый шаг}} + \underbrace{\frac{1}{\alpha\sqrt{n}} \exp(-c_{\text{tr}}\alpha n)}_{\text{начальная ошибка}},$$

остаточный член RR

где $c_{\text{tr}} > 0$. Первое слагаемое задаётся мартингальной нормальной аппроксимацией, второе отражает ухудшение оценки при слишком малом шаге, а два следующих возникают при контроле остаточного члена ошибки RR. Последнее слагаемое описывает скорость забывания начального условия. При $\alpha \asymp n^{-1/2}$ все четыре полиномиальных слагаемых в (16) имеют порядок не больше $n^{-1/4}$, а начальная ошибка экспоненциально мала. Следовательно, основной результат сразу даёт

$$d_K(T_{n,\alpha}^{\text{RR}}(u), \mathcal{N}(0, 1)) \lesssim_{\log n} n^{-1/4},$$

что соответствует порядку сходимости с убывающим шагом в работе [9]. При этом такой выбор шага является оптимальным для полученной оценки.

3. Доверительное оценивание: алгоритм и численные результаты

В этом разделе описано построение доверительных интервалов для RR-оценки. Численные эксперименты показывают влияние коррекции смещения, выбора шага и способа оценивания дисперсии на точность нормальной аппроксимации и покрытие интервалов.

3.1. Построение доверительных интервалов

Для заранее выбранного направления u рассмотрим доверительный интервал номинального уровня 95% для $u^\top \theta^*$:

$$(17) \quad u^\top \hat{\theta}_{n,n_0}^{(\alpha, \text{RR})} \pm z_{0.975} \frac{\hat{\sigma}(u)}{\sqrt{n - n_0}},$$

где $z_{0.975}$ — квантиль стандартного нормального распределения, а $\hat{\sigma}^2(u)$ оценивает асимптотическую дисперсию проекции $u^\top (2\theta_k^{(\alpha)} - \theta_k^{(2\alpha)})$.

При марковском шуме обычная выборочная дисперсия не учитывает зависимость между последовательными наблюдениями. Поэтому в экспериментах используются методы неперекрывающихся (NBM) и перекрывающихся (ОВМ) блочных средних [16, 17]. Отдельно рассматривается интервал с известной асимптотической дисперсией, вычисленной аналитически.

Теоретические гарантии в Теореме 1 получены для интервалов с известной асимптотической дисперсией, тогда как интервалы с оценённой дисперсией в настоящей работе исследуются численно в Приложении В.1.

3.2. Постановка экспериментов

Во всех экспериментах размерность параметра равна $d = 5$, а шум задаётся эргодической марковской цепью с десятью состояниями.

Переходная матрица строится посредством построчной нормировки независимых величин $M_{ij} \sim U(0, 1)$. Средняя матрица и её марковские возмущения имеют вид

$$\bar{\mathbf{A}} = U \operatorname{diag}(\lambda_1, \dots, \lambda_d) U^\top, \quad \mathbf{A}(z) = \bar{\mathbf{A}} - s \left(R_z - \sum_x \pi_x R_x \right),$$

где U — случайная ортогональная матрица, $\lambda_j \sim U(0.25, 0.60)$, а элементы R_z распределены как $U(-1, 1)$. Коэффициент s выбирается так, чтобы $\|\mathbf{A}(z)\|_2 \leq 1$. Координаты $\mathbf{b}(z)$ независимы и равномерно распределены на $(-1, 1)$; $\theta^* = \bar{\mathbf{A}}^{-1} \bar{\mathbf{b}}$. Рекурсии с шагами α и 2α строятся по одной траектории цепи. Все результаты о покрытии относятся к фиксированной скалярной проекции.

3.3. Точность гауссовской аппроксимации

В первом эксперименте фиксируются одна задача и одно случайное единичное направление u . Для каждого из 11 значений длины траектории n в диапазоне $[2 \cdot 10^4, 10^6]$ моделируются $N_{\text{traj}} = 10\,000$ независимых траекторий. Используется сбалансированный режим

$$\alpha_n = 20/\sqrt{n}, \quad n_0 = \lfloor \{\alpha_n \lambda_{\min}(\bar{\mathbf{A}})\}^{-1} \log^2 n \rfloor.$$

Здесь $\theta_0 = 0$, $Z_0 = 0$, а для нормировки используется аналитическое асимптотическое стандартное отклонение, причём $\sigma^2(u) = 1.6539$. Для PR-средних с шагами α_n , $2\alpha_n$ и их RR-комбинации рассматривается статистика

$$(18) \quad Z_n = \frac{\sqrt{n - n_0} u^\top (\bar{\theta}_{n, n_0} - \theta^*)}{\sigma(u)}.$$

Это практическое правило разогрева отличается от $n_0 = n/2$ в Теореме 1 и непосредственно ею не покрывается. Покрытие в таблице 1 и ниже означает эмпирическую вероятность события $\{|Z_n| \leq z_{0.975}\}$, усреднённую по нескольким запускам эксперимента.

На рисунке 1 показаны распределения при $n = 10^6$ и эмпирическое расстояние Колмогорова

$$\hat{D}_n = \sup_x \left| \hat{F}_n(x) - \Phi(x) \right|$$

Таблица 1. Эмпирические характеристики Z_n при $n = 10^6$ на основе 10 000 траекторий.

Метод	Среднее	Дисперсия	$\widehat{D}_n$	Покрытие
PR, 2α	0.957	1.059	0.3626	82.96%
PR, α	0.482	1.056	0.1886	91.70%
RR	0.006	1.054	0.0129	94.75%

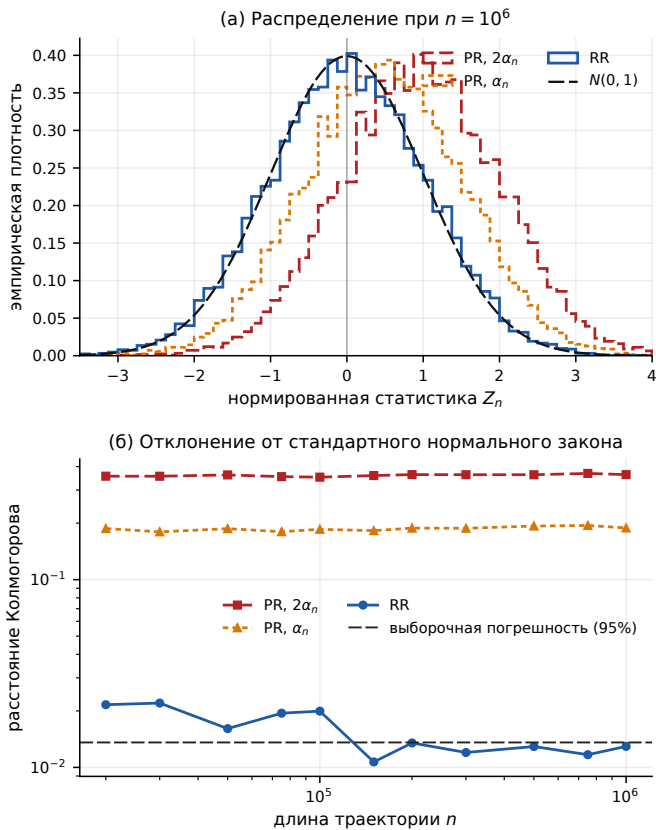


Рисунок 1. Главный эксперимент на сбалансированном шаге $\alpha_n = 20/\sqrt{n}$. Сверху: распределения трёх нормированных статистик при $n = 10^6$. Снизу: расстояние Колмогорова до $N(0, 1)$ на 11 горизонтах; пунктирная линия показывает 95%-й ориентир выборочной погрешности при 10 000 траекториях.

на всей сетке значений n , где $\Phi(x)$ — функция распределения стандартного нормального закона.

При $n = 10^6$ экстраполяция Ричардсона–Ромберга уменьшает среднее нормированной статистики с 0.482 и 0.957 до 0.006, тогда как её дисперсия практически не меняется. Поэтому улучшение нормальной аппроксимации связано именно с коррекцией среднего, а не с увеличением разброса. Покрытие возрастает до 94.75%, а расстояние Колмогорова уменьшается до 0.0129.

Та же картина сохраняется на всей сетке: среднее RR-статистики лежит в интервале $[-0.0207, 0.0134]$, тогда как средние отдельных PR-оценок остаются отделёнными от нуля. Их расстояния Колмогорова также не убывают, в то время как для RR они находятся в диапазоне 0.0107–0.0220. При $n \geq 150\,000$ наблюдаемое отклонение RR сопоставимо с погрешностью для 10 000 траекторий. Это не доказывает точную нормальность, но подтверждает предсказываемый теорией механизм: RR устраняет ведущий сдвиг, не увеличивая дисперсию случайных флуктуаций.

Дополнительные эксперименты на 100 случайных задачах показывают устойчивость RR-оценки к выбору шага, тогда как для отдельных PR-средних покрытие интервалов снижается. Использование перекрывающихся блочных средних также даёт покрытие, близкое к номинальному. Постановка этих экспериментов, формулы оценок дисперсии и подробные результаты приведены в Приложении В.1.

Заключение

Проведённый анализ показывает, что стационарное смещение PR-усреднения с постоянным шагом может существенно влиять на точность статистического вывода. Применение экстраполяции Ричардсона–Ромберга к средним двух последовательностей итераций с шагами α и 2α позволяет устранить ведущую составляющую этого смещения и тем самым повысить точность нормальной аппроксимации. Как следствие, RR-оценка позволяет улучшить скорость гауссовской аппроксимации усреднённых оценок линейного $TD(0)$ и установить статистические гарантии для параметров приближения функции ценности по конечной траектории марковской модели принятия решений. Это позволяет характеризовать статистическую неопределённость оценок по длинным траекториям зависимых наблюдений.

Численные эксперименты на синтетических данных показывают, что RR-коррекция уменьшает смещение и улучшает покрытие доверительных интервалов без заметного увеличения их ширины. Дальнейшее развитие результатов может быть связано с обоснованием нормальной аппроксимации для RR-оценки при замене $\sigma^2(u)$ на выборочную оценку асимптотической дисперсии и применением метода мультипликативного бутстрапа [9].

Список использованных источников

- [1] Sutton R. S., Barto A. G. *Reinforcement Learning: An Introduction*, 2nd edition.– The MIT Press.– 2018.– ISBN 978-0-262-03924-6.
- [2] Robbins H., Monro S. *A stochastic approximation method* // The Annals of Mathematical Statistics.– 1951.– Vol. **22**.– No. 3.– Pp. 400–407 .
- [3] Ruppert D. *Efficient estimations from a slowly convergent Robbins–Monro process*.– Cornell University Operations Research and Industrial Engineering.– 1988 (Cornell eCommons, Technical Report 781).
- [4] Поляк Б. Т. *Новый метод типа стохастической аппроксимации* // Автоматика и телемеханика.– 1990.– № 7.– С. 98–107.
- [5] Polyak B. T., Juditsky A. B. *Acceleration of stochastic approximation by averaging* // SIAM Journal on Control and Optimization.– 1992.– Vol. **30**.– No. 4.– Pp. 838–855 .
- [6] Dieuleveut A., Durmus A., Bach F. *Bridging the gap between constant step size stochastic gradient descent and Markov chains* // The Annals of Statistics.– 2020.– Vol. **48**.– No. 3.– Pp. 1348–1382 .
- [7] Huo D., Chen Y., Xie Q. *Effectiveness of constant stepsize in Markovian LSA and statistical inference* // Proceedings of the AAAI Conference on Artificial Intelligence.– 2024.– Vol. **38**.– No. 18.– Pp. 20447–20455 .
- [8] Levin I., Naumov A., Samsonov S. *High-order error bounds for Markovian LSA with Richardson–Romberg extrapolation* // Proceedings of the AAAI Conference on Artificial Intelligence.– 2026.– Vol. **40**.– No. 43.– Pp. 36696–36704 (Extended version: arXiv: 2508.05570) .
- [9] Samsonov S., Sheshukova M., Moulines E., Naumov A. *Statistical inference for linear stochastic approximation with Markovian noise* // Advances in Neural Information Processing Systems.– 2025.– Vol. **38** (Proceedings version) .
- [10] Нагаев С. В. *Уточнение предельных теорем для однородных цепей Маркова* // Теория вероятн. и ее примен.– 1961.– Т. **6**.– № 1.– С. 67–86.
- [11] Сираждинов С. Х., Форманов Ш. К. *Об оценках скорости сходимости в центральной предельной теореме для однородных цепей Маркова* // Теория вероятн. и ее примен.– 1983.– Т. **28**.– № 2.– С. 219–228.
- [12] Нагаев С. В. *Спектральный метод и центральная предельная теорема для общих цепей Маркова* // Изв. РАН. Сер. матем.– 2017.– Т. **81**.– № 6.– С. 114–157.
- [13] Srikant R. *Rates of convergence in the central limit theorem for Markov chains, with an application to TD learning*.– 2024 (arXiv: 2401.15719, version 4 revised in 2026).
- [14] Wang Z., Wang Y., Narang I., Wang F., Wang Y., Maguluri S. T. *Steady-state behavior of constant-stepsize stochastic approximation: Gaussian approximation and tail bounds*.– 2026 (arXiv: 2602.13960).

- [15] Аверина Т. А., Алифиренко А. А. *Анализ устойчивости линейного осциллятора с мультипликативным шумом* // Сиб. журн. вычисл. матем.– 2007.– Т. 10.– № 2.– С. 127–145.
- [16] Flegal J. M., Jones G. L. *Batch means and spectral variance estimators in Markov chain Monte Carlo* // The Annals of Statistics.– 2010.– Vol. 38.– No. 2.– Pp. 1034–1070 .
- [17] Liu Y., Vats D., Flegal J. M. *Batch size selection for variance estimators in MCMC* // Methodology and Computing in Applied Probability.– 2022.– Vol. 24.– No. 1.– Pp. 65–93 .

Приложение А: доказательства

А.1. Оценка остаточного члена

Для начального распределения ξ цепи $(Z_k)_{k \geq 0}$ обозначим через $\mathbb{E}_\xi$ соответствующее математическое ожидание, то есть $Z_0 \sim \xi$.

Разложения (26)–(27), оценка (25) и её точная версия (87) в приложении работы [8] дают представление

$$(19) \quad \bar{\mathbf{A}}(\bar{\theta}_n^{(\alpha, \text{RR})} - \theta^*) = -\frac{2}{n} \sum_{k=n/2}^{n-1} \varepsilon(Z_{k+1}) + \mathcal{E}_{n,p,\alpha}^{\text{RR}},$$

где при $\alpha \leq \alpha_p^L$ для любого начального распределения ξ ,

$$(20) \quad \{\mathbb{E}_\xi [\|\mathcal{E}_{n,p,\alpha}^{\text{RR}}\|^p]\}^{1/p} \leq C_L(t_{\text{mix}}, d) r_{n,p,\alpha},$$

и, с $L_\alpha = \log(1/(\alpha a))$,

$$(21) \quad \begin{aligned} r_{n,p,\alpha} = & p \log(2p) n^{-3/4} + p^3 (\alpha n)^{-1/2} L_\alpha^{1/2} n^{-1/2} \\ & + p^{3/2} \alpha^{1/2} n^{-1/2} + p^{7/2} \alpha^{3/2} L_\alpha^{3/2} \\ & + p^{7/2} n^{-1} \alpha^{1/2} L_\alpha^{3/2} \\ & + \frac{\|\theta_0 - \theta^*\|}{\alpha n} \exp\left(-\frac{\alpha a n}{24}\right). \end{aligned}$$

В формуле (87) работы [8] соответствующий множитель записан как $\log_2(2p)$, то есть как логарифм по основанию 2, а не как $\log^2(2p)$. Здесь использовано тождество $\log_2(2p) = \log(2p)/\log 2$, и постоянный множитель $1/\log 2$ поглощён константой $C_L(t_{\text{mix}}, d)$. Здесь и далее константы могут зависеть от фиксированных параметров задачи, но не от n , p и α . Формула (20) — единственная необходимая далее оценка момента высокого порядка.

А.2. Вспомогательные леммы

Поясним схему доказательства. Положим $m = n/2$ и

$$v := \bar{\mathbf{A}}^{-\top} u, \quad g(z) := v^\top \varepsilon(z).$$

Поскольку $v^\top \bar{\mathbf{A}} = u^\top$, разложение (19) представляет нормированную ошибку в виде

$$(22) \quad T_{n,\alpha}^{\text{RR}}(u) = -\frac{1}{\sqrt{m}\sigma(u)} \sum_{k=m}^{n-1} g(Z_{k+1}) + \frac{\sqrt{m}}{\sigma(u)} v^\top \mathcal{E}_{n,p,\alpha}^{\text{RR}}.$$

Первое слагаемое является ведущей линейной частью, а второе — остаточным членом, для которого доступна оценка моментов высокого порядка (20).

Ведущую сумму, в свою очередь, можно разложить с помощью уравнения Пуассона. Пусть $\hat{g}$ — ограниченное решение $\hat{g} - P\hat{g} = g$. Тогда

$$(23) \quad \sum_{k=m}^{n-1} g(Z_{k+1}) = \sum_{k=m}^{n-1} \{\hat{g}(Z_{k+1}) - P\hat{g}(Z_k)\} + P\hat{g}(Z_m) - P\hat{g}(Z_n).$$

Сумма в правой части состоит из мартингалльных разностей; её нормальная аппроксимация устанавливается в Лемме 1. Крайние члены пуассоновского разложения объединяются с $\mathcal{E}_{n,p,\alpha}^{\text{RR}}$ в остаточную случайную величину. Её момент контролируется с помощью (20) и ограниченности $\hat{g}$. Для их объединения используем неравенство [9, Предложение 1]: для любых вещественных случайных величин X, Y и любого $p \geq 1$

$$(24) \quad d_K(X + Y, \mathcal{N}(0, 1)) \leq d_K(X, \mathcal{N}(0, 1)) + 2 \{\mathbb{E}[|Y|^p]\}^{1/(p+1)}.$$

Таким образом, достаточно отдельно оценить расстояние Колмогорова для мартингалльной части и моменты остаточной случайной величины.

ЛЕММА 1 (Мартингалльная ЦПТ). *Пусть выполнены условия 1–3, $\Sigma_\varepsilon \succ 0$, n чётно, $n \geq \max\{4, 2t_{\text{mix}}\}$ и $u \in \mathbb{S}^{d-1}$. Определим*

$$(25) \quad X_n(u) := -\frac{1}{\sqrt{n/2}\sigma(u)} \sum_{k=n/2}^{n-1} \{\hat{g}(Z_{k+1}) - P\hat{g}(Z_k)\}.$$

Тогда существуют не зависящие от n, p, u и ξ константы C_1, C_2 такие, что для любого начального распределения ξ и любого $p \geq 2$, если положить

$$(26) \quad \mathcal{B}_M(n, p) := \frac{C_1 p^{(3p+1)/(2(2p+1))}}{(n/2)^{p/(2(2p+1))}} + C_2 \left\{ \frac{\log n}{\sqrt{n/2}} + \frac{p}{(n/2)^{p/(2p+1)}} \right\}$$

то выполнено неравенство

$$(27) \quad d_K(X_n(u), \mathcal{N}(0, 1)) \leq \mathcal{B}_M(n, p).$$

ДОКАЗАТЕЛЬСТВО. Из (7) и (8) следует

$$(28) \quad \sigma^2(u) = v^\top \Sigma_\varepsilon v.$$

Таким образом, $\sigma(u) > 0$, а $\sigma^2(u)$ является асимптотической дисперсией процесса $g(Z_k)$. Кроме того,

$$(29) \quad \frac{\|v\|}{\sigma(u)} \leq C_\varepsilon, \quad \frac{\|g\|_\infty}{\sigma(u)} \leq C_\varepsilon \|\varepsilon\|_\infty.$$

Условие равномерной геометрической эргодичности обеспечивает оценку

$$\|\hat{g}\|_\infty \lesssim t_{\text{mix}} \|g\|_\infty.$$

Положим $m = n/2$ и для $j = 0, \dots, m-1$ обозначим

$$D_{j+1} := \hat{g}(Z_{m+j+1}) - P\hat{g}(Z_{m+j}), \quad M_m := \sum_{j=0}^{m-1} D_{j+1}.$$

Относительно фильтрации $\mathcal{F}_{m+j} = \sigma(Z_0, \dots, Z_{m+j})$ величины D_{j+1} являются мартингальными разностями. Если

$$h(z) := P(\hat{g}^2)(z) - \{P\hat{g}(z)\}^2,$$

то

$$\mathbb{E}_\xi[D_{j+1}^2 \mid \mathcal{F}_{m+j}] = h(Z_{m+j}), \quad \langle M \rangle_m = \sum_{j=0}^{m-1} h(Z_{m+j}).$$

При стационарном начальном распределении из уравнения Пуассона следует тождество

$$\mathbb{E}_\pi[h(Z)] = \mathbb{E}_\pi[\{\hat{g}(Z_1) - P\hat{g}(Z_0)\}^2] = \sigma^2(u).$$

Следовательно, $m\sigma^2(u)$ является стационарным средним квадратичной характеристики M_m . При произвольном начальном распределении он, вообще говоря, не совпадает с её математическим ожиданием.

Начальным распределением сдвинутой цепи $(Z_{m+j})_{j \geq 0}$ служит закон Z_m при $Z_0 \sim \xi$. Применяя к функции $h - \pi(h)$ аргумент из доказательства [9, Лемма 2], равномерный по начальному распределению, получаем

$$(30) \quad \{\mathbb{E}_\xi[|\langle M \rangle_m - m\sigma^2(u)|^p]\}^{1/p} \lesssim \|g\|_\infty^2 p^{1/2} t_{\text{mix}}^{5/2} \sqrt{m}.$$

Кроме того,

$$|D_{j+1}| \leq 2\|\hat{g}\|_\infty \lesssim t_{\text{mix}} \|g\|_\infty.$$

Для корректного применения [9, Лемма 1] введём фактическую безусловную дисперсию мартингальной суммы

$$(31) \quad s_{m,\xi}^2 := \sum_{j=0}^{m-1} \mathbb{E}_\xi[D_{j+1}^2] = \mathbb{E}_\xi[\langle M \rangle_m].$$

Из (30), неравенства Йенсена и случая $p = 2$ следует

$$(32) \quad |s_{m,\xi}^2 - m\sigma^2(u)| \leq \{\mathbb{E}_\xi[|\langle M \rangle_m - m\sigma^2(u)|^2]\}^{1/2} \lesssim \|g\|_\infty^2 t_{\text{mix}}^{5/2} \sqrt{m}.$$

Разделив (32) на $t\sigma^2(u)$ и применив (29), получаем

$$(33) \quad \left| \frac{s_{m,\xi}^2}{t\sigma^2(u)} - 1 \right| \leq \delta_m, \quad \delta_m := C_\delta t_{\text{mix}}^{5/2} m^{-1/2},$$

где C_δ зависит только от фиксированных параметров задачи.

Если $\delta_m < 1/2$, то $s_{m,\xi}^2 \geq t\sigma^2(u)/2$. Применим [9, Лемма 1] к мартингалльным разностям D_{j+1} , положив в обозначениях этой леммы $s_m^2 = s_{m,\xi}^2$. Вместе с (29) это даёт

$$d_K\left(-\frac{M_m}{s_{m,\xi}}, \mathcal{N}(0, 1)\right) \leq \mathcal{B}_M(n, p).$$

Положим $r_{m,\xi} = s_{m,\xi}/(\sqrt{m}\sigma(u))$. Тогда $|r_{m,\xi} - 1| \leq |r_{m,\xi}^2 - 1| \leq \delta_m$, и из неравенства треугольника следует

$$\begin{aligned} d_K\left(-\frac{M_m}{\sqrt{m}\sigma(u)}, \mathcal{N}(0, 1)\right) &\leq d_K\left(-\frac{M_m}{s_{m,\xi}}, \mathcal{N}(0, 1)\right) + \sup_{x \in \mathbb{R}} |\Phi(x/r_{m,\xi}) - \Phi(x)| \\ &\leq \mathcal{B}_M(n, p) + C\delta_m. \end{aligned}$$

Если же $\delta_m \geq 1/2$, то тривиальная оценка даёт

$$d_K\left(-\frac{M_m}{\sqrt{m}\sigma(u)}, \mathcal{N}(0, 1)\right) \leq 1 \leq 2\delta_m.$$

Следовательно, в обоих случаях

$$d_K\left(-\frac{M_m}{\sqrt{m}\sigma(u)}, \mathcal{N}(0, 1)\right) \leq \mathcal{B}_M(n, p) + Ct_{\text{mix}}^{5/2} m^{-1/2}.$$

Поскольку $m = n/2$ и $\log n \geq \log 2$, второе слагаемое является слагаемым меньшего порядка. Наконец,

$$X_n(u) = -\frac{M_m}{\sqrt{m}\sigma(u)},$$

поэтому получаем (27) без дополнительного ограничения на m . $\square$

ЛЕММА 2 (Остаточный член). Пусть выполнены условия леммы 1. Пусть также $p \geq 2$ и $0 < \alpha \leq \alpha_p^L$. Тогда существует не зависящая от n , p , α и u константа $C_P > 0$ такая, что для

$$(34) \quad \mathcal{R}_{n,p,\alpha} := C_P \frac{C_\varepsilon t_{\text{mix}} \|\varepsilon\|_\infty}{\sqrt{n/2}} + C_\varepsilon C_L(t_{\text{mix}}, d) \sqrt{n/2} r_{n,p,\alpha}$$

выполнено неравенство

$$(35) \quad d_K(T_{n,\alpha}^{\text{RR}}(u), \mathcal{N}(0, 1)) \leq \mathcal{B}_M(n, p) + 2\mathcal{R}_{n,p,\alpha}^{p/(p+1)},$$

где $\mathcal{B}_M(n, p)$ определён в (26).

ДОКАЗАТЕЛЬСТВО. Из (22) и (23) следует разложение

$$(36) \quad T_{n,\alpha}^{\text{RR}}(u) = X_n(u) + Y_{n,p,\alpha}(u),$$

где

$$Y_{n,p,\alpha}(u) := -\frac{P\hat{g}(Z_{n/2}) - P\hat{g}(Z_n)}{\sqrt{n/2}\sigma(u)} + \frac{\sqrt{n/2}}{\sigma(u)} v^\top \mathcal{E}_{n,p,\alpha}^{\text{RR}}.$$

Оценки (20) и (29), а также ограниченность решения уравнения Пуассона дают

$$\{\mathbb{E}_\xi[|Y_{n,p,\alpha}(u)|^p]\}^{1/p} \leq \mathcal{R}_{n,p,\alpha}$$

для любого начального распределения ξ . Применяя (24) к $X = X_n(u)$ и $Y = Y_{n,p,\alpha}(u)$ и используя лемму 1, получаем

$$\begin{aligned} d_K(T_{n,\alpha}^{\text{RR}}(u), \mathcal{N}(0, 1)) &\leq d_K(X_n(u), \mathcal{N}(0, 1)) + 2\{\mathbb{E}_\xi[|Y_{n,p,\alpha}(u)|^p]\}^{1/(p+1)} \\ &\leq \mathcal{B}_M(n, p) + 2\{\mathcal{R}_{n,p,\alpha}\}^{p/(p+1)}. \end{aligned}$$

Это доказывает (35). $\square$

А.3. Доказательство теоремы 1

ДОКАЗАТЕЛЬСТВО. Положим

$$p = p_n := \max\{2, \lceil \log n \rceil\}.$$

По определению $\alpha_n^L = \alpha_p^L$. Введём используемую только в доказательстве вспомогательную величину

$$(37) \quad \rho_{n,p,\alpha} := \sqrt{n/2} r_{n,p,\alpha}.$$

Из (21) следует

$$\begin{aligned} \rho_{n,p,\alpha} &\lesssim p \log(2p) n^{-1/4} + \frac{p^3 L_\alpha^{1/2}}{\sqrt{\alpha n}} + p^{3/2} \sqrt{\alpha} + p^{7/2} \sqrt{n} \alpha^{3/2} L_\alpha^{3/2} \\ (38) \quad &+ p^{7/2} n^{-1/2} \alpha^{1/2} L_\alpha^{3/2} + \frac{\|\theta_0 - \theta^*\|}{\alpha \sqrt{n}} \exp\left(-\frac{\alpha a n}{24}\right). \end{aligned}$$

Следовательно,

$$(39) \quad \rho_{n,p,\alpha} \lesssim_{\log n} r_{n,\alpha},$$

где $r_{n,\alpha}$ определяется в (14).

Применим лемму 2 при $p = p_n$. Поскольку $p \geq \log n$, из (26), (39) и (34) следуют оценки

$$(40) \quad \mathcal{B}_M(n, p) \lesssim_{\log n} C_G n^{-1/4}, \quad \mathcal{R}_{n,p,\alpha} \lesssim_{\log n} C_G r_{n,\alpha}.$$

Степень $p/(p+1)$ в (35) также можно убрать из окончательной формулировки. Если $\mathcal{R}_{n,p,\alpha} \geq n^{-1/4}$, то

$$\mathcal{R}_{n,p,\alpha}^{p/(p+1)} \leq e^{1/4} \mathcal{R}_{n,p,\alpha};$$

если же $\mathcal{R}_{n,p,\alpha} < n^{-1/4}$, то

$$\mathcal{R}_{n,p,\alpha}^{p/(p+1)} \leq n^{-p/(4(p+1))} \leq e^{1/4} n^{-1/4}.$$

Подстановка этих двух оценок в (35) даёт (15). $\square$

Приложение В. Дополнительные эксперименты

В.1. Устойчивость по шагу и по задачам

Во втором эксперименте генерируются 100 независимых задач, по 100 траекторий длины $T = 10^6$ для каждой. Цепь стартует из стационарного распределения, $\theta_0 = 0$, разогрев равен $n_0 = 1000$. Для каждой задачи фиксируется случайное единичное направление, а пары шагов сравниваются на общих траекториях. Пусть $m = T - n_0$ — число итераций после разогрева и

$$Y_t = u^\top (2\theta_t^{(\alpha)} - \theta_t^{(2\alpha)}), \quad t = 1, \dots, m,$$

где для краткости сохранённая часть траектории перенумерована начиная с единицы. Для оценки $\sigma(u)$ методом неперекрывающихся блочных средних (NBM) выберем число блоков K , положим $\ell = \lfloor m/K \rfloor$ и определим

$$\bar{Y}_r = \frac{1}{\ell} \sum_{t=(r-1)\ell+1}^{r\ell} Y_t, \quad \bar{Y}_{\text{NBM}} = \frac{1}{K} \sum_{r=1}^K \bar{Y}_r.$$

Тогда оценка асимптотической дисперсии и стандартная ошибка имеют вид

$$(41) \quad \hat{\sigma}_{\text{NBM}}^2 = \frac{\ell}{K} \sum_{r=1}^K (\bar{Y}_r - \bar{Y}_{\text{NBM}})^2, \quad \hat{\text{se}}_{\text{NBM}} = \sqrt{\frac{\hat{\sigma}_{\text{NBM}}^2}{K\ell}}.$$

В реализации статистика (18) строится с оценкой асимптотического стандартного отклонения вместо $\sigma(u)$. Для NBM эффективная длина усреднения равна $K\ell$, поэтому статистика, по которой вычисляется покрытие в Таблице 2, имеет вид

$$(42) \quad Z_{\text{NBM}}(u) = \frac{\sqrt{K\ell} u^\top (\bar{\theta}_{T,n_0}^{(\alpha, \text{RR})} - \theta^*)}{\hat{\sigma}_{\text{NBM}}} = \frac{u^\top (\bar{\theta}_{T,n_0}^{(\alpha, \text{RR})} - \theta^*)}{\hat{\text{se}}_{\text{NBM}}}.$$

Таким образом, NBM-интервал для $u^\top \theta^*$ имеет вид

$$(43) \quad \left[u^\top \bar{\theta}_{T,n_0}^{(\alpha, \text{RR})} - z_{0.975} \hat{\text{se}}_{\text{NBM}}, u^\top \bar{\theta}_{T,n_0}^{(\alpha, \text{RR})} + z_{0.975} \hat{\text{se}}_{\text{NBM}} \right].$$

В основном сравнении берётся $K = 63$; если $K\ell < m$, последние $m - K\ell$ наблюдений не используются при вычислении NBM-оценки. В таблице приведены медианы по задачам; значение ошибки для каждой задачи усредняется по траекториям.

ТАБЛИЦА 2. Медианы для 100 случайных задач при $T = 10^6$

Пара $(2\alpha, \alpha)$	Покрытие без RR		RR		
	2α	α	$10^3 \times \text{MSE}$	$10^3 \times \text{ширина}$	Покрытие
(0.04, 0.02)	92.0%	94.0%	2.97	5.36	94.0%
(0.10, 0.05)	86.0%	91.5%	2.97	5.37	94.0%
(0.20, 0.10)	67.5%	86.0%	2.97	5.38	94.0%

С ростом шага покрытие отдельных PR-средних снижается с 92.0% до 67.5% для ветви 2α и с 94.0% до 86.0% для ветви α . Для RR оно остаётся равным 94.0% при практически неизменных MSE и ширине интервала. На наибольшей паре шагов (0.20, 0.10) экстраполяция уменьшает медианное MSE с 7.53 и 4.46 до 2.97 без расширения интервала.

Для оценки методом перекрывающихся блочных средних (ОВМ) выберем длину блока b и положим $N_b = m - b + 1$. Среднее в блоке, начинающемся в момент j , и среднее по всей сохранённой части траектории равны соответственно

$$\bar{Y}_{j,b} = \frac{1}{b} \sum_{t=j}^{j+b-1} Y_t, \quad \bar{Y} = \frac{1}{m} \sum_{t=1}^m Y_t.$$

Соответствующие ОВМ-оценки имеют вид

$$(44) \quad \hat{\sigma}_{\text{ОВМ}}^2 = \frac{b}{N_b} \sum_{j=1}^{N_b} (\bar{Y}_{j,b} - \bar{Y})^2, \quad \hat{\text{se}}_{\text{ОВМ}} = \sqrt{\frac{\hat{\sigma}_{\text{ОВМ}}^2}{m}}.$$

В этом случае статистика (18) принимает вид

$$(45) \quad Z_{\text{ОВМ}}(u) = \frac{\sqrt{m} u^\top (\bar{\theta}_{T,n_0}^{(\alpha, \text{RR})} - \theta^*)}{\hat{\sigma}_{\text{ОВМ}}} = \frac{u^\top (\bar{\theta}_{T,n_0}^{(\alpha, \text{RR})} - \theta^*)}{\hat{\text{se}}_{\text{ОВМ}}}.$$

Соответствующий ОВМ-интервал получается из (43) заменой $\hat{\text{se}}_{\text{NBM}}$ на $\hat{\text{se}}_{\text{ОВМ}}$. Замена NBM на ОВМ с $b = \lfloor T^{0.6} \rfloor$ практически не меняет результат: для пары (0.20, 0.10) медианное и среднее покрытия равны 95% и 94.6%, а ширина составляет $5.42 \cdot 10^{-3}$. Для использованного PR-алгоритма с убывающим шагом соответствующие покрытия равны 92% и 88.9%. Таким образом, результаты RR устойчивы как к выбору шага, так и к способу оценки асимптотической дисперсии.

В.2. Алгоритм TD(0) как линейная стохастическая аппроксимация

Рассмотрим марковскую модель принятия решений (MDP) $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, где $\mathcal{S}$ и $\mathcal{A}$ — пространства состояний и действий, $P(\cdot | s, a)$ — ядро перехода, $r(s, a)$ — ограниченная награда, а $\gamma \in [0, 1)$ — коэффициент дисконтирования. При фиксированной стационарной политике $\pi(\cdot | s)$ наблюдается траектория $a_k \sim \pi(\cdot | s_k)$, $s_{k+1} \sim P(\cdot | s_k, a_k)$. Цепь состояний

имеет переходное ядро

$$P_\pi(B \mid s) = \int_{\mathcal{A}} P(B \mid s, a) \pi(da \mid s).$$

Задача оценивания политики состоит в приближении её функции ценности

$$V^\pi(s) = \mathbb{E} \left[\sum_{k=0}^{\infty} \gamma^k r(s_k, a_k) \mid s_0 = s \right].$$

При большом пространстве состояний используют линейное приближение $V_{\theta}^\pi(s) = \phi(s)^\top \theta$, где $\phi: \mathcal{S} \rightarrow \mathbb{R}^d$ — заданное отображение признаков, а $\theta \in \mathbb{R}^d$ — оцениваемый параметр.

Обозначим через μ инвариантное распределение ядра P_π . Целевой параметр линейного TD(0) [1, Раздел 9.4] задаётся системой $\bar{\mathbf{A}}\theta^* = \bar{\mathbf{b}}$, где

$$\begin{aligned} \bar{\mathbf{A}} &= \mathbb{E}_{s \sim \mu, s' \sim P_\pi(\cdot \mid s)} \left[\phi(s) (\phi(s) - \gamma \phi(s'))^\top \right], \\ \bar{\mathbf{b}} &= \mathbb{E}_{s \sim \mu, a \sim \pi(\cdot \mid s)} [\phi(s) r(s, a)]. \end{aligned}$$

Функция $V_{\theta^*}^\pi$ служит приближением V^π ; их разность составляет ошибку аппроксимации.

Для наблюдения $Z_k = (s_k, a_k, s'_k)$, где $s'_k = s_{k+1}$, положим

$$\begin{aligned} A_k &:= \mathbf{A}(Z_k) = \phi(s_k) (\phi(s_k) - \gamma \phi(s'_k))^\top, \\ b_k &:= \mathbf{b}(Z_k) = \phi(s_k) r(s_k, a_k). \end{aligned}$$

Тогда TD(0) с постоянным шагом α имеет вид

$$\theta_k^{(\alpha)} = \theta_{k-1}^{(\alpha)} - \alpha (A_k \theta_{k-1}^{(\alpha)} - b_k),$$

то есть совпадает с LSA (3). Для оценки функции ценности в заданном состоянии используется проекция $\hat{V}^\pi(s) := \phi(s)^\top \hat{\theta}_n^{(\alpha)}$, где $\hat{\theta}_n^{(\alpha)}$ это некоторая оценка искомого вектора параметров по выборке $(Z_1, \dots, Z_n)$ и шагом α .

Поступила в редакцию	18.08.2026;
одобрена после рецензирования	15.09.2026;
принята к публикации	21.09.2026;
опубликована онлайн	13.10.2026.

Рекомендовал к публикации

к.т.н. С. А. Амелёвкин

Информация об авторах:



Даниль Эльдарович Гайнанов

Исследователь в лаборатории Kandinsky. Область научных интересов: машинное обучение, обучение с подкреплением, методы синтеза изображений и видео, стохастический анализ.



0009-0006-9596-9587

e-mail: eviltsundera@gmail.com



Илья Валерьевич Левин

Младший научный сотрудник Международной лаборатории стохастических алгоритмов и анализа многомерных данных НИУ ВШЭ. Область научных интересов: стохастическая аппроксимация, обучение с подкреплением.



0000-0002-9708-8189

e-mail: ivlevin@hse.ru

Авторы внесли равный вклад в подготовку публикации.

Декларация об отсутствии личной заинтересованности: благополучие авторов не зависит от результатов исследования.



Gaussian approximation for linear stochastic approximation with an application to reinforcement learning

Danil Eldarovich **Gainanov**¹, Ilya Valerievich **Levin**²

^{1,2} Higher School of Economics, Moscow, Russia

²ivlevin@hse.ru

Abstract. We consider The problem of statistical inference for solving a system of linear equations $\mathbf{A}\theta^* = \mathbf{b}$, where the matrix $\mathbf{A}$ and the vector $\mathbf{b}$ are unknown, and noisy observations are available instead. we use linear stochastic approximation with a constant stepsize and Markovian noise. Polyak–Ruppert averaging reduces variance but retains a stationary bias of order α , which degrades a Gaussian approximation centered at θ^* . To remove the leading bias, Richardson–Romberg extrapolation combines averages of two recursions with step-sizes α and 2α , driven by the same Markov-chain trajectory. This correction is shown to improve the Gaussian approximation without changing the limiting variance.

We establish a non-asymptotic bound in Kolmogorov distance for scalar projections of the averaged iterates with Richardson–Romberg extrapolation. For a trajectory of length n , choosing a stepsize of order $n^{-1/2}$ yields the rate $n^{-1/4}$ up to logarithmic factors. As an application of our theory, we consider the construction of confidence intervals for the parameters estimated by the TD(0) algorithm in reinforcement learning. Numerical experiments on synthetic data confirm the correction of the center and improved confidence-interval coverage without a material increase in width. (*In Russian*).

Key words and phrases: Linear stochastic approximation, Gaussian approximation, Richardson–Romberg extrapolation, confidence intervals, reinforcement learning

2020 *Mathematics Subject Classification:* 62L20; 60J20, 65C40

For citation: Danil E. Gainanov, Ilya V. Levin. *Gaussian approximation for linear stochastic approximation with an application to reinforcement learning*. Program Systems: Theory and Applications, 2026, **17**:4(73), pp. 33–56. (*In Russ.*). https://psta.psir.ru/read/psta2026_4_33-56.pdf

References

- [1] R. S. Sutton, A. G. Barto. *Reinforcement Learning: An Introduction*, 2nd edition, The MIT Press, 2018, ISBN 978-0-262-03924-6.
- [2] H. Robbins, S. Monro. “A stochastic approximation method”, The Annals of Mathematical Statistics, **22**:3 (1951), pp. 400–407 [doi](#).
- [3] D. Ruppert. *Efficient estimations from a slowly convergent Robbins–Monro process*, Cornell University Operations Research and Industrial Engineering, 1988 (Cornell eCommons, Technical Report 781).
- [4] B. T. Polyak. “Novyj metod tipa stoxasticheskoy approssimacii”, Avtomatika i telexanika, 1990, no. 7, pp. 98–107.
- [5] B. T. Polyak, A. B. Juditsky. “Acceleration of stochastic approximation by averaging”, SIAM Journal on Control and Optimization, **30**:4 (1992), pp. 838–855 [doi](#).
- [6] A. Dieuleveut, A. Durmus, F. Bach. “Bridging the gap between constant step size stochastic gradient descent and Markov chains”, The Annals of Statistics, **48**:3 (2020), pp. 1348–1382 [doi](#).
- [7] D. Huo, Y. Chen, Q. Xie. “Effectiveness of constant stepsize in Markovian LSA and statistical inference”, Proceedings of the AAAI Conference on Artificial Intelligence, **38**:18 (2024), pp. 20447–20455 [doi](#).
- [8] I. Levin, A. Naumov, S. Samsonov. “High-order error bounds for Markovian LSA with Richardson–Romberg extrapolation”, Proceedings of the AAAI Conference on Artificial Intelligence, **40**:43 (2026), pp. 36696–36704 (Extended version: [arXiv: 2508.05570](#)) [doi](#).
- [9] S. Samsonov, M. Sheshukova, E. Moulines, A. Naumov. “Statistical inference for linear stochastic approximation with Markovian noise”, Advances in Neural Information Processing Systems, **38** (2025) (Proceedings version) [doi](#).
- [10] S. V. Nagaev. “Utochnenie predel’nyx teorem dlya odnorodnyx cepej Markova”, Teoriya veroyatn. i ee primen., **6**:1 (1961), pp. 67–86.
- [11] S. X. Sirazhdinov, Sh. K. Formanov. “Ob ocenkax skorosti sxodimosti v central’noj predel’noj teoreme dlya odnorodnyx cepej Markova”, Teoriya veroyatn. i ee primen., **28**:2 (1983), pp. 219–228.
- [12] S. V. Nagaev. “Spektral’nyj metod i central’naya predel’naya teorema dlya obshhix cepej Markova”, Izv. RAN. Ser. matem., **81**:6 (2017), pp. 114–157.
- [13] R. Srikant. “Rates of convergence in the central limit theorem for Markov chains, with an application to TD learning”, 2024 ([arXiv: 2401.15719](#), version 4 revised in 2026).

- [14] Z. Wang, Y. Wang, I. Narang, F. Wang, Y. Wang, S. T. Maguluri. “Steady-state behavior of constant-stepsizes stochastic approximation: Gaussian approximation and tail bounds”, 2026 ([arXiv: 2602.13960](#)).
- [15] T. A. Averina, A. A. Alifirenko. “Analiz ustojchivosti linejnogo oscillyatora s mul’tiplikativnym shumom”, Sib. zhurn. vychisl. matem., **10**:2 (2007), pp. 127–145.
- [16] J. M. Flegal, G. L. Jones. “Batch means and spectral variance estimators in Markov chain Monte Carlo”, The Annals of Statistics, **38**:2 (2010), pp. 1034–1070 .
- [17] Y. Liu, D. Vats, J. M. Flegal. “Batch size selection for variance estimators in MCMC”, Methodology and Computing in Applied Probability, **24**:1 (2022), pp. 65–93 .