Правила остановки методов градиентного типа для невыпуклых задач при аддитивных ошибках градиента

Ф. С. Стонякин [1]

9 декабря 2022 г.

[1] На базе совместных работ с Курузовым И. А., Поляком Б. Т. и Алкусой М. С.

Введение. Постановка задачи. Невыпуклые задачи и условие градиентного доминирования

- **Наша цель** — исследование методов градиентного типа в случае приближённого значения градиента $\widetilde{\nabla} f(x)$ минимизируемой функции f в произвольной запрашиваемой точке $x \in \mathbb{R}^n$ при некотором фиксированном $\Delta > 0$:

$$\nabla f(x) = \widetilde{\nabla} f(x) + v(x), \quad \text{причём} \quad \|v(x)\| \leqslant \Delta \,^{1}. \qquad (1)$$

- Упор делается на **невыпуклые задачи**. В частности, на задачи минимизации функции удовлетворяющих (PL)-условию.

[1] Здесь и далее норма евклидва.

1. Рассмотрим простой пример не сильно выпуклой функции

$$f(x) = \langle Ax, x \rangle, \tag{2}$$

где $A = \mathrm{diag}(L, \mu, 0)$ — диагональная матрица 3-го порядка с ровно двумя положительными элементами $L > \mu > 0$. При $v(x) = (0, 0, \Delta)$, $x_0 = (0, 0, 0)$ и $h_k > 0$ последовательность $x_{k+1} = x_k - h_k \widetilde{\nabla} f(x_k)$ стремится к бесконечности при неограниченном увеличении k.

2. Далее, можно рассмотреть функцию Розенброка двух переменных $x = (x^{(1)}, x^{(2)})$

$$f(x) = 100 \left(x^{(2)} - \left(x^{(1)} \right)^2 \right)^2 + \left(1 - x^{(1)} \right)^2. \tag{3}$$

При $x_0 = (1, 1) = x_*$ и погрешности $v(x_k)$, такой, что $x_k^{(2)} = \left(x_k^{(1)} \right)^2$ траектория градиентного метода может уходить довольно далеко от точного решения x_*.

Цель работы:

1. Исследовать оценку расстояния $\|x_N - x_0\|$ для выдаваемых градиентным методом точек x_N.
2. Предложить правило ранней остановки градиентного метода, которое могло бы гарантировать некоторый компромисс между стремлением достичь приемлемого качества точки выхода по функции и целью обеспечить достаточно умеренное удаление такой точки от выбранного начального положения.

Правила ранней остановки: история вопроса

- По-видимому, впервые идеология раннего прерывания итераций предложена в статье [2], посвящённой методике для приближённого решения возникающих при регуляризации некорректных или плохо обусловленных задач.
- Известные подходы, связанные с ранней остановкой методов первого порядка в случае использования на итерациях неточной информации о градиенте (см. [3], гл. 6, параграф 1, а также к примеру недавний препринт [4])

---

[2] Емелин И. В., Красносельский М. А. Правило останова в итерационных процедурах решения некорректных задач. // Автомат. и телемех., 1978. Вып. 12. С. 59–63.

[3] Поляк Б. Т. Введение в оптимизацию. М.: Наука, 1983. 384 с.

[4] Vasin A., Gasnikov, A. and Spokoiny, V. Stopping rules for accelerated gradient methods with additive noise in gradient. // arxiv preprint (2021), Url: `https://arxiv.org/pdf/2102.02921.pdf`.

Правила ранней остановки: история вопроса

- Однако данные (известные нам) результаты для выпуклых (не сильно выпуклых) задач отличаются по сравнению с полученными нами, что обычно гарантируется достижение худшего уровня по функции (даже без погрешностей получается лишь сублинейная скорость сходимости) или оценок вида $\|x_N - x_*\| \leqslant \|x_0 - x_*\|$ без исследования $\|x_N - x_0\|$, где $\{x_k\}_{k \in \mathbb{N}}$ — образуемая методом последовательность, x_* — ближайшее к точке старта метода x_0 точное решение задачи минимизации f.

План

- Предположения о задаче
- Известные результаты: градиентный метод с аддитивными помехами
- Основные теоретические результаты работы
- Результаты вычислительных экспериментов

Stopping Rules for Gradient Methods for Non-Convex Problems with Additive Noise in Gradient. Fedor S. Stonyakin, Ilia A. Kuruzov, Boris T. Polyak. ArXiv preprint: arXiv:2205.07544 [math.OC], 2022, https://arxiv.org/abs/2205.07544.
Gradient-Type Methods for Optimization Problems with Polyak-Lojasiewicz Condition: Early Stopping and Adaptivity to Inexactness Parameter. Ilya A. Kuruzov, Fedor S. Stonyakin, and Mohammad S. Alkousa. // Advances in Optimization and Applications, 2022.

Предположения

Пусть норма $\|\cdot\|$ евклидова. Будем рассматривать задачи минимизации f, удовлетворяющие условию Липшица градиента

$$\|\nabla f(x) - \nabla f(y)\| \leqslant L\|x - y\| \quad \forall\, x,\, y \in \mathbb{R}^n \tag{4}$$

для евклидовой нормы, а также (PL)-условию [5, 6, 7]

$$f(x) - f^* \leqslant \frac{1}{2\mu}\|\nabla f(x)\|^2 \quad \forall\, x \in \mathbb{R}^n, \tag{5}$$

где x_* — одно из точных решений задачи минимизации f, $f^* = f(x_*)$, а $\mu > 0$ — некоторая константа.

---

[5] Поляк Б. Т. Градиентные методы минимизации функционалов, 1963.

[6] Karimi, H., Nutini, J., Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the Polyak–Lojasiewicz condition, 2016.

[7] Belkin, M. Fit Without Fear: Remarkable Mathematical Phenomena of Deep Learning Through the Prism of Interpolation, 2021.

Условие PL

1. Любая μ-сильно выпуклая функция удовлетворяет условию PL с константой μ;
2. Можно рассмотреть систему нелинейных уравнений [8] $g(x) = 0$ (записанную в векторном виде) и задачу нахождения какого-нибудь решения этой системы, где $g : \mathbb{R}^n \to \mathbb{R}^m$, $m \leqslant n$. Тогда при условии $\lambda_{min}\left(\frac{\partial g(x)}{\partial x} \cdot \left[\frac{\partial g(x)}{\partial x}\right]^T\right) \geqslant \mu, \forall x$ функция $f(x) = \|g(x)\|^2$ удовлетворяет условию (5) для произвольного x_* такого, что $f(x_*) = 0$. [9]

[8] Гасников А. В. Современные численные методы оптимизации. Метод универсального градиентного спуска. М.: МЦНМО, 2021. 272 с.

[9] Nesterov Y., Polyak B. Cubic regularization of Newton method and its global performance. // Math. Program. Ser. A. 2006. Vol. 108. P. 177–205.

Известные результаты: градиентный метод с аддитивными помехами

Теорема 1 (1/3)

Пусть для численного решения задачи минимизации

$$f(x) \to \min_{x \in \mathbb{R}^n}$$

в случае, если f удовлетворяет условию L-гладкости (4), используется градиентный спуск вида

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k). \tag{6}$$

Тогда верно неравенство

$$\min_{k=0,\ldots,N-1} \|\nabla f(x_k)\|_2 \leqslant \sqrt{\frac{2L(f(x_0) - f(x_*))}{N}}. \tag{7}$$

Теорема 1 (2/2)

Пусть f L-гладкая и верно (PL)-условие (5). Тогда имеет место сходимость градиентного метода со скоростью геометрической прогрессии

$$\begin{aligned} f(x_N) - f(x_*) &\leqslant \left(1 - \frac{\mu}{L}\right)^N (f(x_0) - f(x_*)) \leqslant \\ &\leqslant \exp\left(-\frac{\mu}{L} N\right) (f(x_0) - f(x_*)), \end{aligned} \tag{8}$$

причём

$$\|x_* - x_0\| \leqslant \frac{\sqrt{2L(f(x_0) - f^*)}}{\mu}, \tag{9}$$

где x_* — ближайшее к x_0 решение задачи минимизации f.

Нижняя оценка: ON THE LOWER BOUND OF MINIMIZING POLYAK-ŁOJASIEWICZ FUNCTIONS. // Under review as a conference paper at ICLR 2023.

Замечание 1

Существенно, что оценки при использовании градиентного метода (6) с неточным градиентом вида (1)

$$\min_{k=0,N-1} \|\nabla f(x_k)\| \leqslant \sqrt{\Delta^2 + \frac{2L(f(x_0) - f(x_*))}{N}} \leqslant \leqslant \Delta + \sqrt{\frac{2L(f(x_0) - f(x_*))}{N}} \tag{10}$$

и

$$f(x_{k+1}) - f^* \leqslant \left(1 - \frac{\mu}{L}\right)^{k+1} (f(x_0) - f^*) + \frac{\Delta^2}{2\mu}, \tag{11}$$

вообще говоря, не улучшаемы. К примеру, известны нижние оценки уровня точности по функции $O\left(\frac{\Delta^2}{2\mu}\right)$ для градиентного метода с аддитивно неточным градиентом (см., например, раздел 2.11.1 пособия [10], а также имеющиеся там ссылки) даже на классе сильно выпуклых функций.

[10] Воронцова Е. А., Хильдебранд Р. Ф., Гасников А В., Стонякин Ф. С. Выпуклая оптимизация 2021.

Пример

Действительно, рассмотрим такой пример:

$$\min_{x \in \mathbb{R}^n} f(x) := \frac{1}{2} \sum_{i=1}^{n} \lambda_i (x^i)^2, \tag{12}$$

где $0 \leqslant \mu = \lambda_1 \leqslant \lambda_2 \leqslant ... \leqslant \lambda_n = L$, $L \geqslant 2\mu$. Решением задачи (12) будет $x^* = 0$. Предположим, что неточность имеется только в вычислении первой компоненты градиента. То есть вместо $\partial f(x)/\partial x^1 = \mu x^1$ нам доступно только $\widetilde{\partial} f(x)/\partial x^1 = \mu x^1 - \Delta$.

Для простейшей динамики градиентного спуска (в этом разделе мы обозначаем номер компоненты вектора верхним индексом, потому что далее будет использоваться одновременно номер итерации и номер компоненты (13)):

$$x_{k+1} = x_k - \frac{1}{L}\widetilde{\nabla} f(x_k),$$

можно получить, что при $x_0^1 \geqslant 0$ и достаточно больших $k \in \mathbb{N}$ $(k \gg L/\mu)$

$$x_k^1 \geqslant \frac{\Delta}{L} \cdot \frac{1-(1-\mu/L)^k}{1-(1-\mu/L)} \simeq \frac{\Delta}{\mu}. \tag{13}$$

Следовательно, при $x_0^1 \geqslant 0$ и достаточно больших $k \in \mathbb{N}$

$$f(x_k) - f(x_*) \gtrsim \frac{\Delta^2}{2\mu}.$$

Основные результаты

$$x_{k+1} = x_k - \frac{1}{L}\widetilde{\nabla} f(x_k) \tag{14}$$

$$f(x_{k+1}) - f(x_k) \leqslant \frac{\Delta^2}{L} - \frac{1}{4L}\|\widetilde{\nabla} f(x_k)\|^2. \tag{15}$$

Неравенство (15) показывает, что в случае достаточно большого значения $\|\widetilde{\nabla} f(x_k)\|$ можно гарантировать, что $f(x_{k+1}) < f(x_k)$. Тем самым для всякого $C > 2$ возникает альтернатива: или верно неравенство $\|\widetilde{\nabla} f(x_k)\| \leqslant C\Delta$, и это само по себе гарантирует достижение приемлемого качества точки выхода x_k по функции в силу (PL)-условия, или же

$$f(x_{k+1}) - f(x_k) < -\frac{\Delta^2}{L}\left(\frac{C^2}{4} - 1\right). \tag{16}$$

Тем самым за конечное число шагов градиентного метода (20) возможно получить x_k такое, что значение $f(x_k)$ достаточно близко к минимальному f^*. Выберем для определённости $C = \sqrt{6}$ (чтобы получить «удобный» коэффициент) и будем рассматривать 2 сценария:

1. $\|\widetilde{\nabla} f(x_k)\| > \Delta\sqrt{6}$, что с учётом (15) влечёт неравенство

$$f(x_{k+1}) - f(x_k) < -\frac{\Delta^2}{2L}, \tag{17}$$

указыв. на убывание значения f при переходе от x_k к x_{k+1}.

2.

$$\|\widetilde{\nabla} f(x_k)\| \leqslant \Delta\sqrt{6}, \tag{18}$$

откуда

$$f(x_k) - f^* \leqslant \frac{7\Delta^2}{\mu}. \tag{19}$$

Теорема 2

Пусть на N-й итерации градиентного метода

$$x_{k+1} = x_k - \frac{1}{L}\widetilde{\nabla} f(x_k) \tag{20}$$

впервые выполнен критерий остановки

$$\|\widetilde{\nabla} f(x_k)\| \leqslant \Delta\sqrt{6}. \tag{21}$$

Тогда для точки выхода $\widehat{x} = x_N$ гарантированно будет верно неравенство

$$f(\widehat{x}) - f^* \leqslant \frac{7\Delta^2}{\mu}. \tag{22}$$

При этом справедлива следующая оценка количества итераций до остановки

$$N < \frac{2L}{\Delta^2}(f(x_0) - f^*). \tag{23}$$

Теорема 3

Пусть градиентный метод (20) работает либо

$$N_* = \left\lceil \frac{L}{\mu} \ln \frac{\mu(f(x_0) - f^*)}{6\Delta^2} \right\rceil. \tag{24}$$

шагов, либо при некотором $N \leqslant N_*$ на N-й итерации метода (20) впервые выполнен критерий остановки (21). Тогда для точки выхода $\widehat{x}$ ($\widehat{x} = x_N$ или $\widehat{x} = x_{N_*}$) метода (20) гарантированно будет верно неравенство

$$f(\widehat{x}) - f^* \leqslant \frac{7\Delta^2}{\mu},$$

причём

$$\|\widehat{x} - x_0\| \leqslant \frac{2\Delta}{\mu} \sqrt{1 + \frac{L}{\mu} \left\lceil \ln \frac{\mu(f(x_0) - f^*)}{6\Delta^2} \right\rceil} + \frac{4\sqrt{L(f(x_0) - f^*)}}{\mu}. \tag{25}$$

Замечание 3

В силу теоремы 1 и (25) величину $\|\widehat{x} - x_0\|$ можно считать сопоставимой с $\|x_* - x_0\|$ при достаточно малом $\Delta > 0$, поскольку

$$\|x_* - x_0\| \leqslant \frac{\sqrt{2L(f(x_0) - f^*)}}{\mu}. \tag{26}$$

Замечание 4

Ввиду (25) достаточно потребовать выполнения (4) и (5) (L-гладкость и условие PL)только в R–окрестности x_0, где

$$R = \frac{2\Delta}{\mu}\sqrt{1 + \frac{L}{\mu}}\left\lceil \ln \frac{\mu(f(x_0) - f^*)}{6\Delta^2} \right\rceil + \frac{4\sqrt{L(f(x_0) - f^*)}}{\mu}.$$

Characterization of Gradient Dominance and Regularity Conditions for Neural Networks. Yi Zhou, Yingbin Liang. ArXiv preprint: arXiv:2205.07544 [math.OC], 2017, https://arxiv.org/abs/1710.06910.

Вариант градиентного метода с адаптивно подбираемым шагом

Идея метода с адаптивным шагом: подбирать величину L_k так, чтобы было выполнено условие

$$\widetilde{f}(x_{k+1}) \leqslant \widetilde{f}(x_k) + \langle \widetilde{\nabla} f(x_k), x_{k+1} - x_k \rangle + L_k \|x_{k+1} - x_k\|^2 + \frac{\Delta^2}{2L_k} + 2\delta,$$

где $\widetilde{f}(x)$ есть неточное значение функции f в точке x.

Алгоритм 1. Адаптивный вариант градиентного метода.

Require: $L_{\min} \geqslant 0, L_0 \geqslant L_{\min}, \delta \geqslant 0, \Delta \geqslant 0$

1: $k := 0$

2: $L_k = \max\left(\frac{L_{k-1}}{2}, L_{\min}\right)$

3: Вычисляем новое значение x:

$$x_{k+1} = x_k - \frac{1}{2L_k}\widetilde{\nabla} f(x_k) \tag{27}$$

4: Если выполнено

$$\widetilde{f}(x_{k+1}) \leqslant \widetilde{f}(x_k) + \langle \widetilde{\nabla} f(x_k), x_{k+1} - x_k \rangle + L_k \|x_{k+1} - x_k\|^2 + \frac{\Delta^2}{2L_k} + 2\delta \tag{28}$$

то увеличиваем $k := k + 1$ и переходим к шагу 2. Иначе $L_k := 2L_k$ и переходим к шагу 3.

5: **return** $\mathbf{x}_k$

Теорема 4 (1/2)

Пусть в алгоритме 1 $L_{\min} \geqslant \frac{\mu}{4}$, а также верны (PL)-условие (5) и

$$|f(x) - \widetilde{f}(x)| \leqslant \delta, \tag{29}$$

причем $\Delta^2 \geqslant 16L\delta$. Тогда алгоритм 1 работает либо

$$N_* = \left\lceil \frac{8L}{\mu} \log \frac{\mu(f(x_0) - f^*)}{\Delta^2} \right\rceil \tag{30}$$

шагов, либо при некотором $N \leqslant N_*$ на N-й итерации алгоритма 1 выполнен критерий остановки

$$\|\widetilde{\nabla} f(x_k)\| \leqslant 2\Delta. \tag{31}$$

Теорема 4 (2/2)

Тогда для точки выхода $\widehat{x}$ ($\widehat{x} = x_N$ или $\widehat{x} = x_{N_*}$) алгоритма 1 гарантированно будет верно неравенство

$$f(\widehat{x}) - f^* \leqslant \frac{5\Delta^2}{\mu},$$

причём

$$\|\widehat{x} - x_0\| \leqslant 8\frac{\Delta}{\mu}\sqrt{\frac{1}{2}\gamma^2 + 4\gamma\frac{L}{\mu}\log\frac{\mu(f(x_0) - f^*)}{\Delta^2}} + 16\frac{\sqrt{\gamma L(f(x_0) - f^*)}}{\mu}, \tag{32}$$

где $\gamma = \frac{L}{L_{\min}}$. Также суммарное количество обращений к подпрограмме для вычисления неточных значений функции и шага (27) не более чем $2N + \log\frac{2L}{L_0}$.

Замечание 5

Условие

$$\|\widetilde{\nabla} f(x_k)\| \leqslant 2\Delta \tag{33}$$

выполнено для любых $L_k \geqslant L$ и по построению получаем, что $L_k \leqslant 2L$. В оценках выше величина $2L$ оценивает максимальное значение параметра L_k. Оценки выше останутся верными, если заменить L на $\frac{1}{2}\max_{j\leqslant k} L_j$ и γ на $\frac{\max_{j\leqslant k} L_j}{2\min_{j\leqslant k} L_j}$. Аналогично можно заменить параметр алгоритма $L_{\min}$ на $\min_{j\leqslant k} L_j$.

Замечание 6

Оценка на количество итераций

$$N_* = \left\lceil \frac{8L}{\mu} \log \frac{\mu(f(x_0) - f^*)}{\Delta^2} \right\rceil \tag{34}$$

в теореме 4 говорит о конечности процесса, однако является сильно завышенной. На практике более интересной оценкой является

$$N_* = \left\lceil \frac{4\widehat{L}}{\mu} \log \frac{\mu(f(x_0) - f^*)}{\Delta^2} \right\rceil,$$

где $\widehat{L} = \frac{\mu}{4} \frac{1}{1 - \left(\prod_{j=0}^{N_*-1} \left(1 - \frac{\mu}{4L_j} \right) \right)^{\frac{1}{N_*}}}$ — величина, зависящая от подобранных параметров L_j в процессе работы алгоритма 1.

Замечание 7

Возможно ослабить требование $L_{\min} \geqslant \frac{\mu}{4}$ до $L_{\min} > 0$. В таком случае будет верна оценка

$$\|x_N - x_0\| \leqslant N\Delta\sqrt{\frac{1}{2L_{\min}^2} + \frac{4}{\mu L_{\min}}} + 16\sqrt{\frac{L}{L_{\min}}}\frac{\sqrt{L(f(x_0) - f^*)}}{\mu}$$

на N-ой итерации градиентного метода, однако уже не верна оценка

$$\|\widehat{x} - x_0\| \leqslant 8\frac{\Delta}{\mu}\sqrt{\frac{1}{2}\gamma^2 + 4\gamma\frac{L}{\mu}}\log\frac{\mu(f(x_0) - f^*)}{\Delta^2} + 16\frac{\sqrt{\gamma L(f(x_0) - f^*)}}{\mu} \tag{36}$$

из теоремы 4. В таком случае возможно оценить достаточное для гарантированного выполнения критерия $\|\widetilde{\nabla} f(x_k)\| \leqslant 2\Delta$ количество итераций алгоритма 1

$$N < \frac{2L}{\Delta^2 - 16L\delta}(f(x_0) - f^*), \tag{37}$$

где $\Delta^2 > 16L\delta$.

Алгоритм 2. Адаптивный градиентный метод для неизвестных L и Δ.

Require: x_0, $L_{\min} \geqslant \frac{\mu}{4} > 0, L_0 \geqslant L_{\min}, \Delta_0 > 0, \Delta_{\min} > 0$.

1: $k := 0$, $L_k := \max\left\{\frac{L_{k-1}}{2}, L_{\min}\right\}$.

2: Если условие останова выполнено для x_k, то переходим к шагу 6. Иначе вычисляем $x_{k+1} = x_k - \frac{1}{2L_k}\widetilde{\nabla} f(x_k)$

3: Если

$$f(x_{k+1}) \leqslant f(x_k) + \langle \widetilde{\nabla} f(x_k), x_{k+1} - x_k \rangle + \Delta_k \|x_{k+1} - x_k\| + \frac{L_k}{2}\|x_{k+1} - x_k\|^2, \tag{35}$$

то переходим к шагу 4. Иначе $L_k := 2L_k$ and $\Delta_k = 2\Delta_k$ и переходим к шагу 2.

4: Переопределим Δ_k так, чтобы было верно (35), а также $\Delta_k \geqslant \Delta_{\min}$ и $\Delta_k \geqslant \max_{1 \leqslant j < k} \Delta_j$.

5: Обновляем $k := k + 1$ и переходим к шагу 1.

6: **return** x_k.

Выполняется одна из следующих альтернатив:

1. Алгоритм 2 работает N_* шагов
2. Для некоторого $k_* \leqslant N_*$ на k_*ой итерации выполняется критерий останова

$$\|\widetilde{\nabla} f(x_{k_*})\| \leqslant 2 \max_{j \leqslant k} \Delta_j, \tag{38}$$

$$N_* = \left\lceil \frac{8L_{\max}}{\mu} \log \left(\frac{\mu(f(x_0) - f^*)}{4\Delta_{\max}^2} \right) \right\rceil.$$

Теоретические результаты для алгоритма 2

Теорема 5

Пусть либо алгоритм 2 работает

$$N_* = \left\lceil \frac{8L_{\max}}{\mu} \log \left(\frac{\mu(f(x_0) - f^*)}{4\Delta_{\max}^2} \right) \right\rceil \tag{39}$$

шагов, либо для некоторого $k_* \leqslant N_*$ на k_*-ой итерации, выполнен критерий останова

$$\|\widetilde{\nabla} f(x_{k_*})\| \leqslant 2 \max_{j \leqslant k} \Delta_j.$$

Тогда для точки выхода $\widehat{x}$ ($\widehat{x} = x_{k_*}$ or $\widehat{x} = x_{N_*}$) алгоритма 2, верно следующее соотношение

$$f(\widehat{x}) - f^* \leqslant \frac{5 \max_{j \leqslant k} \Delta_j^2}{\mu} \leqslant \frac{5\Delta_{\max}^2}{\mu}. \tag{40}$$

Теоретические результаты

Теорема 6

Более того,

$$\|\widehat{x} - x_0\| \leqslant \frac{8\Delta_{\max}}{\mu}\sqrt{\frac{\gamma^2}{2} + \frac{2\gamma L_{\max}}{\mu}}\log\left(\frac{\mu(f(x_0) - f^*)}{4\Delta_{\max}^2}\right) + \frac{16\sqrt{\gamma L_{\max}(f(x_0) - f^*)}}{\mu},$$

где $\gamma = \frac{4L_{\max}}{\mu}$, $\Delta_{\max} = 2\Delta \max\left\{\frac{L}{L_{\min}}, 1\right\}$, $L_{\max} = L\max\left\{\frac{\Delta}{\Delta_{\min}}, 1\right\}$.
Также, общее количество вызовов подпрограммы для вычисления функции f будет не больше, чем

$$N_* \log_2\left(\frac{4L}{L_{\min}}\max\left\{\frac{L}{L_{\min}}, \frac{\Delta}{\Delta_{\min}}\right\}\right).$$

Теоретические результаты

1. Величина $L_{\max}$ оценивает максимальное значение параметра L_k. Оценки сверху остаются корректными, если заменить $L_{\max}$ на $\max_{j \leqslant k} L_j$.
2. Пусть в каждой точке x мы имеем доступ к модели $(\widetilde{f}, \widetilde{\nabla} f)$ функции f, такой что следующие условия выполняются:

$$\widetilde{f}(x) \leqslant \widetilde{f}(y) + \langle \widetilde{\nabla} f(y), x - y \rangle + \frac{L}{2}\|x - y\|^2 + \Delta\|x - y\| + \delta,$$

$$\widetilde{f}(x) - f^* \leqslant \frac{1}{\mu}\left(\|\widetilde{\nabla} f(x)\| + \Delta^2\right) + \delta.$$

В таком случае, оценки на число итераций и точность решения, и расстояние от стартовой точки до точки выхода остаются верными, если изменить параметры $\Delta_{\max}$ и $L_{\max}$ соответствующим образом.

Численные эксперименты

1. Задача минимизации квадратичной формы

Сравним количество итераций, требуемых для достижения условия (21) и оценку N_* из теоремы 3.

- Задача минимизации следующей квадратичной формы

$$\min_{x \in \mathbb{R}^n} \sum_{j=p+1}^{n} d_j x_j^2, \tag{41}$$

где p — количество нулевых собственных значений матрицы квадратичной формы, d_j — некоторые положительные константы. Т.е. имеем квадратичную форму с неотрицательно определенной вырожденной диагональной матрицей.

- Случайно генерируемая неточность: $v(x) \sim \mathcal{U}\left(S_1^n(0)\right)$
- $\mu = \min\limits_{j=\overline{p+1,n}}(d_j), L = \max\limits_{j=\overline{p+1,n}}(d_j)$
- $X^* = \{x | x_j = 0, j = \overline{p+1, n}\}$

Численные эксперименты

Параметры эксперимента:

- $L = 1$
- μ варьируется от 0 до 1
- Параметры d_j будут браться равномерно из отрезка $[\mu, L]$
- $n = 100$
- $p = 10$

Численные эксперименты

μ	Δ	N	N_*
0.01	10^{-7}	1528	3817
	10^{-4}	841	2436
	10^{-1}	155	1054
0.1	10^{-7}	169	406
	10^{-4}	104	267
	10^{-1}	40	129
0.9	10^{-7}	10	48
	10^{-4}	8	33
	10^{-1}	5	17
0.99	10^{-7}	6	44
	10^{-4}	5	30
	10^{-1}	3	16

Таблица 1: Сравнение номера итерации N для достижения условия (21) и оценки N_* из теоремы 3.

Численные эксперименты

Квадратичная функция

μ	Δ	Alg. constant	Adaptive L	Full Adaptive
0.01	10^{-7}	2.29	2.03	1.84
	10^{-4}	2.26	2.31	2.43
	10^{-1}	2.27	1.95	1.07
0.1	10^{-7}	2.17	1.97	0.87
	10^{-4}	2.18	1.68	0.80
	10^{-1}	2.14	2.26	0.83
0.9	10^{-7}	0.92	1.58	0.69
	10^{-4}	0.91	0.96	0.79
	10^{-1}	0.96	0.95	0.72
0.99	10^{-7}	0.96	0.95	0.71
	10^{-4}	0.95	0.92	0.68
	10^{-1}	1.05	0.93	0.69

Таблица 2: Сравнение достигнутой величины $\frac{\|\widetilde{\nabla} f(x_N)\|}{\Delta}$.

Численные эксперименты

2. Задача минимизации функции логистической регрессии

Рассмотрим задачу минимизации логистической регрессии:

$$f(x) = \frac{1}{m} \sum_{i=1}^{m} \log\left(1 + \exp\left(-y_i \langle w_i, x \rangle\right)\right), \tag{42}$$

где $y = (y_1 \dots y_m)^\top \in [-1, 1]^m$ — вектор целевой переменной, $W = [w_1 \dots w_m] \in \mathbb{R}^{n \times m}$ — матрица признаков, где вектор $w_i \in \mathbb{R}^n$ из того же пространства, что и оптимизируемый вектор весов x.

Заметим, что в общем случае эта задача может не иметь конечного решения, поэтому мы создадим такой искусственный датасет, что существует конечный вектор x^*, минимизирующий заданную функцию.

Численные эксперименты

Параметры эксперимента:

- Размерность задачи: $n = 200$
- Количество элементов в сумме: $m = 700$
- Функция удовлетворяет (PL)-условию локально
- Случайно генерируемая неточность: $v(x) \sim \mathcal{U}\left(S_1^n(0)\right)$

Численные эксперименты

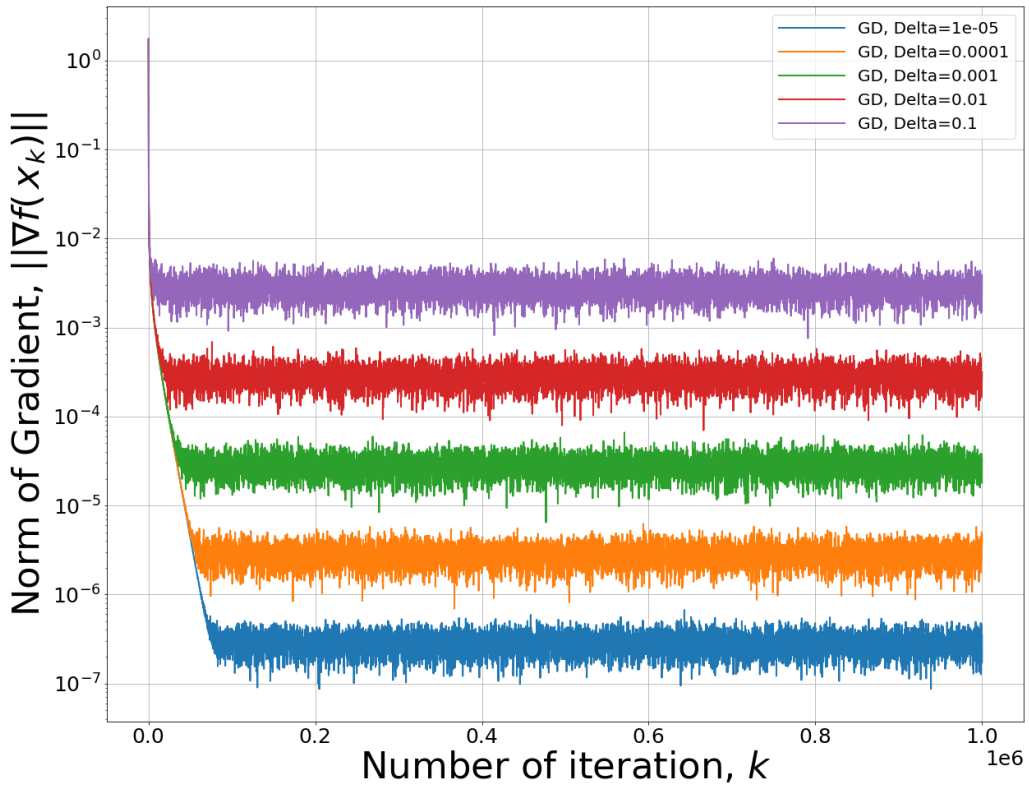

Рис. 1: Скорость сходимости градиентного метода по норме градиента для различных значений неточностей градиента Δ на задаче минимизации логистической регрессии на первых $N = 10^5$ итераций без использования правила остановки.

Численные эксперименты

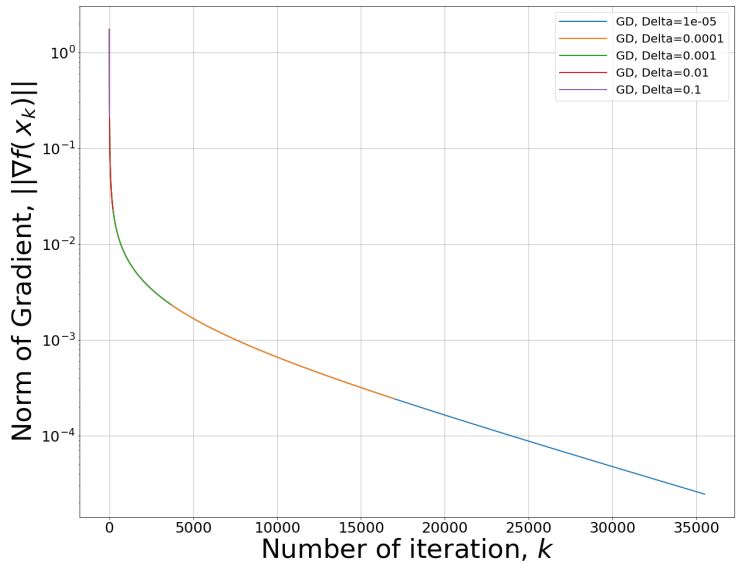

Рис. 2: Скорость сходимости градиентного метода по норме градиента для различных значений неточностей градиента Δ на задаче минимизации логистической регрессии при использовании правила остановки (21).

Как можно видеть, метод останавливается, когда метод достигает точности $\|\widetilde{\nabla} f(x_k)\| \sim \Delta$. Также можно заметить, что на данном примере траектории методов практически совпадают до достижения соответствующей точности.

Численные эксперименты

(a) (b)

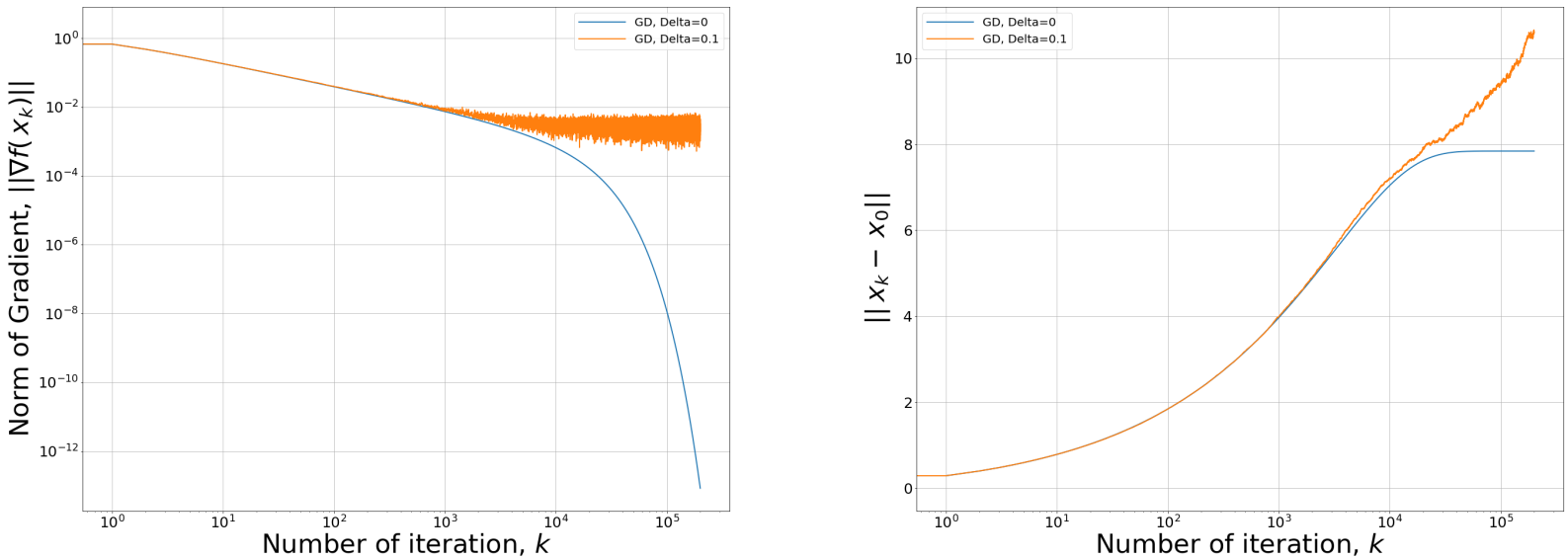

Рис. 3: Результаты для градиентного метода по норме градиента без использования критерия останова для $\Delta = 0.1$ на задаче минимизации логистической регрессии для $v = -\frac{\Delta}{\|\nabla f(x)\|}\nabla f(x)$: (a) скорость сходимости по норме градиента; (b) расстояние от начальной точки до x_k.

Численные эксперименты

Логистическая регрессия

	Alg. constant			Adaptive L		
Δ	Iters	Time, ms	$\frac{\|\widetilde{\nabla} f(\hat{x})\|}{\Delta}$	Iters	Time, ms	$\frac{\|\widetilde{\nabla} f(\hat{x})\|}{\Delta}$
10^{-5}	–	–	–	902	2856.98	2.22
10^{-4}	9700	2605.02	2.42	472	1678.02	2.16
10^{-2}	83	49.86	2.29	17	68.36	2.10

	Full Adaptive		
Δ	Iters	Time, ms	$\frac{\|\widetilde{\nabla} f(\hat{x})\|}{\Delta}$
10^{-5}	23	449.68	3.37
10^{-4}	25	370.09	3.62
10^{-2}	17	161.27	0.84

Таблица 3: Результаты алгоритмов для (42) с различными величинами Δ при использовании правила $\|\widetilde{\nabla} f(x)\| \leqslant \sqrt{6}\Delta$.

Функция Розенброка

Исследуем скорость сходимости алгоритма 1 на двумерной функции Розенброка:

$$f(x_1, x_2) = 100(x_2 - (x_1)^2)^2 + (x_1 - 1)^2.$$

В наших экспериментах мы будем варьировать Δ и брать $\delta = \Delta^2$. Стартовая точка для всех параметров $x_1 = 1, x_2 = 2$. Расстояние от начальной точки до оптимальной $\mathbf{1}$ равно 1. В таблице 4 представлены результаты сходимости для различных типов шумов. В качестве постоянного шума брался вектор $v = (1, 0)^\top$. В данном и следующем эксперименте мы будем пользоваться критерием останова

$$\|\widetilde{\nabla} f(x_k)\| \leqslant 2\Delta. \tag{43}$$

Inexactness	Δ	Iters	Time, ms	$\|x_N - x_0\|$	$\frac{\|\widetilde{\nabla} f(x_N)\|}{\Delta}$	$f(x_N) - f_*$
Random	10^{-4}	7266	2273.26	0.999	2.70	$0.89 \cdot 10^{-7}$
	10^{-3}	5412	2594.61	0.994	2.90	$1.00 \cdot 10^{-5}$
	10^{-2}	3690	3163.32	0.940	2.61	$0.89 \cdot 10^{-3}$
Antigradient	10^{-4}	7188	2615.61	0.999	2.99	$0.11 \cdot 10^{-6}$
	10^{-3}	5493	2490.56	0.993	3.00	$0.11 \cdot 10^{-4}$
	10^{-2}	3536	3031.05	0.931	2.99	$0.12 \cdot 10^{-2}$
Constant	10^{-4}	7491	2301.32	1.000	1.54	$0.27 \cdot 10^{-7}$
	10^{-3}	5697	2490.32	0.997	1.87	$0.24 \cdot 10^{-5}$
	10^{-2}	3965	3485.89	0.965	1.93	$0.30 \cdot 10^{-3}$

Таблица 4: Результаты работы адаптивного градиентного спуска для двумерной функции Розенброка при использовании условия останова (43).

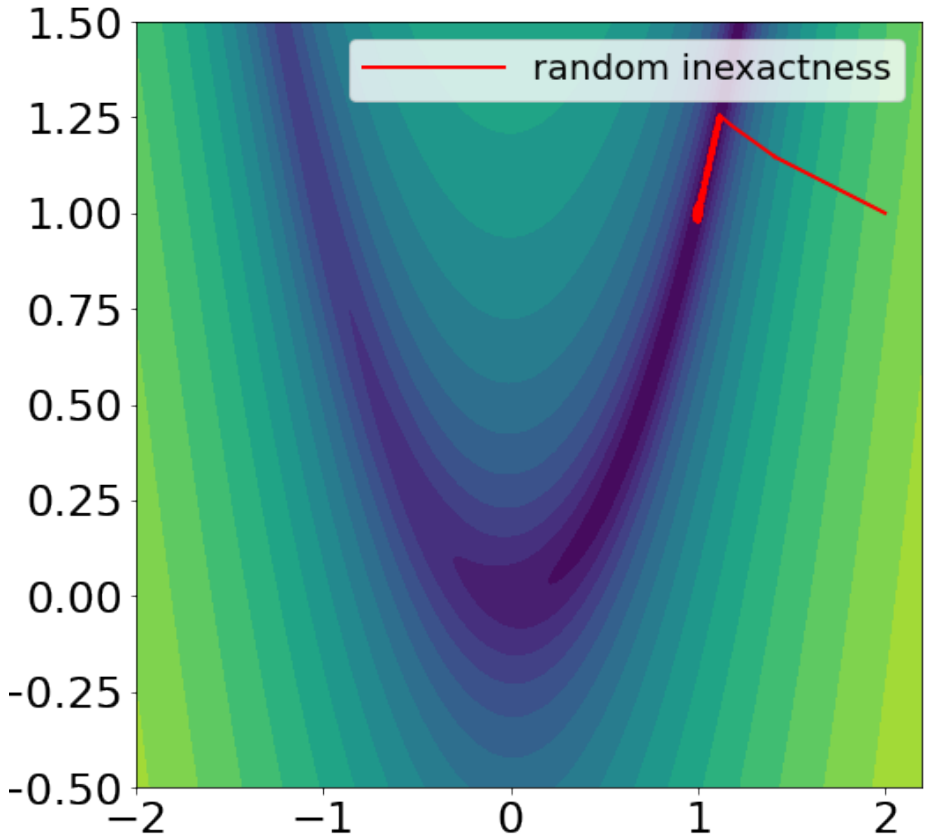

Рис. 4: Траектория градиентного метода со случайной неточностью в погрешности градиента для функции Розенброка без условия останова. Градиентный метод, начиная с некоторой итерации, начинает осцилировать вокруг некоторой точки, и дальнейшие итерации бессмысленны.

Функция Нестерова-Скокова

Рассмотрим систему нелинейных уравнений $g(x) = 0$, где $g_1 = \frac{1}{2}(x_1 - 1), g_i = x_i - 2x_{i-1}^2 + 1, i = \overline{2, n}$. Задача решения этой системы эквивалентна минимизации функции Нестерова-Скокова[11]

$$f(x) = \frac{1}{4}(1 - x_1)^2 + \sum_{i=1}^{n-1} \left(x_{i+1} - 2x_i^2 + 1\right)^2. \qquad (44)$$

Данная функция является обобщением функции Розенброка. Она так же является невыпуклой и удовлетворяет условию Липшица для градиента только локально. Также функция (44) обладает глобальным минимумом в точке $(1, 1 \ldots 1, 1)^\top$ и оптимальным значением $f^* = 0$.

[11] Нестеров Ю. Е., Скоков В. А. К вопросу о тестировании алгоритмов безусловной оптимизации. — в книге "Численные методы математического программирования" — М.: ЦЭМИ, 1980. — С. 77–91.

Как было показано в предыдущих экспериментах, предложенный нами критерий останова одинаково хорошо работает для всех опробованных типов шумов. В данном эксперименте мы будем использовать только шум, равномерно распределенный на сфере. Все наши эксперименты мы будем начинать со стартовой точки $(-1, 1, \dots 1, 1)^\top$. Таким образом, $\|x_0 - x_*\| = 2$. Мы будем варьировать точность градиента Δ и размерность задачи n.

n	Δ	Iters	Time, ms	$\|x_N - x_0\|$	$\frac{\|\widetilde{\nabla} f(x_N)\|}{\Delta}$	$f(x_N) - f^*$
3	10^{-4}	14097	230.58	1.996	2.86	$0.20 \cdot 10^{-4}$
	10^{-3}	2477	247.64	2.155	2.93	$0.11 \cdot 10^{-2}$
	10^{-2}	606	383.63	2.650	2.19	$0.87 \cdot 10^{-2}$
5	10^{-4}	73028	275.03	2.930	2.93	$0.30 \cdot 10^{-3}$
	10^{-3}	15765	292.39	3.312	2.65	$0.49 \cdot 10^{-2}$
	10^{-2}	6	200.87	0.036	1.45	0.98
7	10^{-4}	2898	316.51	0.049	2.69	0.98
	10^{-3}	103	164.23	0.036	2.07	0.98
	10^{-2}	17	104.77	0.036	1.42	0.98

Таблица 5: Результаты работы адаптивного градиентного спуска для функции Нестерова-Скокова при использовании условия останова (43).

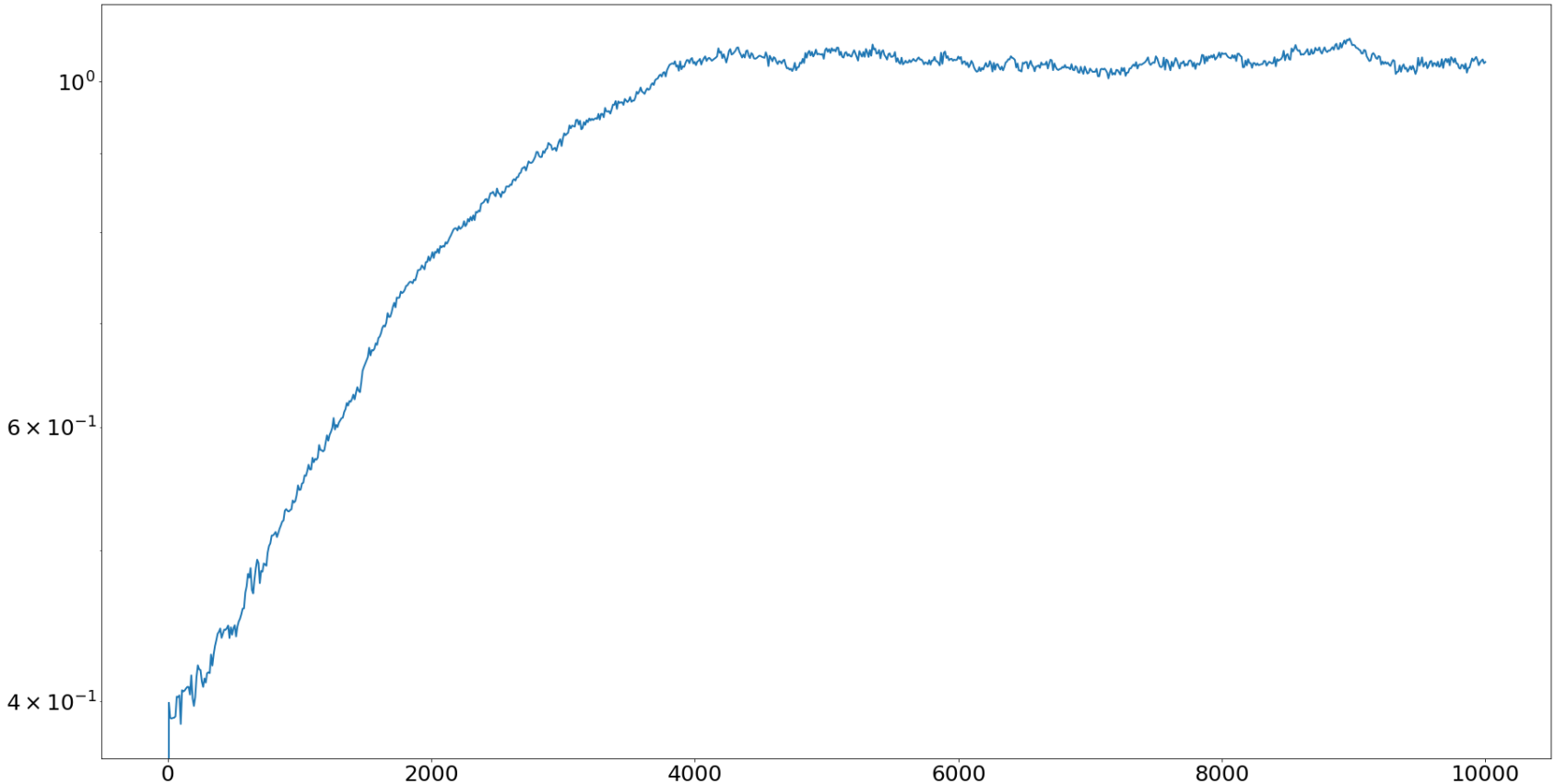

Рис. 5: Расстояние $\|x_N - x_0\|$ без критерия останова выходит в какой-то момент просто на плато.

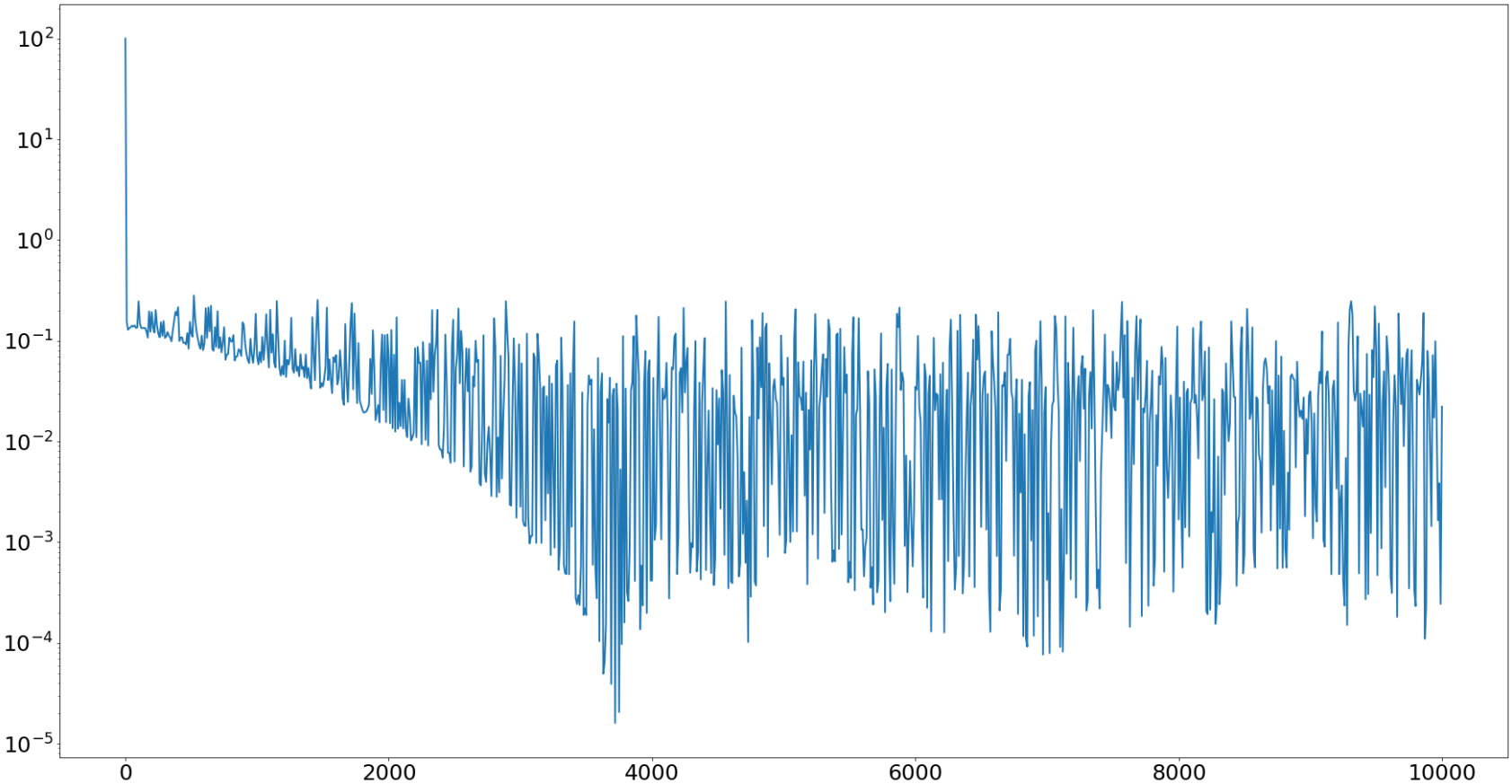

Рис. 6: Это плато приблизительно одинаково для всех уровней шума, и различается только насколько сильно на этом плато значение колеблется. Похожим образом ведет себя и величина $f(x_N) - f^*$.

Численные эксперименты

Нелинейная система

Рассмотрим задачу решения следующей системы нелинейных уравнений

$$g_i(x) = \sum_{j=1}^{n} A_{ij} \sin(x_j) + B_{ij} \cos(x_j) := E_i, \quad i = 1, \ldots, m, \quad (45)$$

где $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ и
$A_{ij}, B_{ij} \in \mathbb{R}$, $\forall i = 1, \ldots, m$; $j = 1, \ldots, n$. Задача выше может быть переписана в виде следующей задачи оптимизации:

$$\min_{x \in \mathbb{R}^n} f(x) := \sum_{i=1}^{m} \left(g_i(x) - E_i\right)^2. \quad (46)$$

Численные эксперименты

Нелинейная система

			Alg. constant		Adaptive L		Full Adaptive	
m	$\frac{L}{\mu}$	Δ	Dist	Grad	Dist	Grad	Dist	Grad
8	$2.1 \cdot 10^5$	10^{-4}	5.1	2.41	5.1	2.25	5.1	3.07
		10^{-1}	4.8	2.34	4.8	2.09	4.9	2.28
32	$5.0 \cdot 10^6$	10^{-4}	7.4	2.41	7.5	2.27	7.5	3.31
		10^{-1}	7.4	2.37	7.4	2.24	7.4	2.72
128	$7.7 \cdot 10^8$	10^{-4}	14.4	2.43	14.4	2.33	14.4	15.67
		10^{-1}	14.3	2.43	14.3	2.32	14.4	1.84

Таблица 6: Результаты для задачи (46) ля различных значений Δ.

Мы сравнили расстояние $\|x_n - x_0\|$ и величину $\frac{\|\widetilde{\nabla} f(x_N)\|}{\Delta}$.

Пока нерешённые задачи

- Стохастические варианты градиентного метода и правила остановки
- Бесконечномерные задачи
- Тестирование на реальных задачах обучения

Заключение

1. Предложены критерии ранней остановки для градиентного метода, гарантирующее с нашей точки зрения компромисс между стремлением достичь приемлемого качества точки выхода по функции и целью обеспечить не столь существенное удаление этой точки от выбранного начального положения.
2. Рассмотрен как градиентный метод с постоянным шагом, так и его вариант с адаптивной регулировкой шага, что полезно в ситуации неизвестного (или даже бесконечного) значения параметра L.
3. Получены теоретические оценки количества итераций для достижения приемлемого качества приближённого решения.
4. Проведены численные эксперименты, демонстрирующие работу критерия останова на практике для задач логистической регрессии, минимизации функции Розенброка и её многомерного аналога (функции Нестерова-Скокова).

Спасибо за внимание!