КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2023 Т. 15 № 2 С. 259–280
DOI: 10.20537/2076-7633-2023-15-2-259-280

Ки&М

МАТЕМАТИЧЕСКИЕ ОСНОВЫ И ЧИСЛЕННЫЕ МЕТОДЫ МОДЕЛИРОВАНИЯ

УДК: 519.8

Влияние конечности мантиссы на точность безградиентных методов оптимизации

Д. Д. Востриков[a], Г. О. Конин[b], А. В. Лобанов[c], В. В. Матюхин[d]

Московский физико-технический институт (национальный исследовательский университет),
Россия, 141701, Московская обл., г. Долгопрудный, Институтский пер., 9

E-mail: [a] danonvostr@gmail.com, [b] koningeorgiy@gmail.com, [c] lobbsasha@mail.ru, [d] vladmatyukh@gmail.com

Получено 19.02.2023.
Принято к публикации 23.02.2023.

Безградиентные методы оптимизации, или методы нулевого порядка, широко применяются в обучении нейронных сетей, обучении с подкреплением, а также в промышленных задачах, где доступны лишь значения функции в точке (работа с неаналитическими функциями). В частности, метод обратного распространения ошибки в PyTorch работает именно по этому принципу. Существует общеизвестный факт, что при компьютерных вычислениях используется эвристика чисел с плавающей точкой, и из-за этого возникает проблема конечности мантиссы.

В этой работе мы, во-первых, сделали обзор наиболее популярных методов аппроксимации градиента: конечная прямая/центральная разность (FFD/FCD), покомпонентная прямая/центральная разность (FWC/CWC), прямая/центральная рандомизация на l_2 сфере (FSSG2/CFFG2); во-вторых, мы описали текущие теоретические представления шума, вносимого неточностью вычисления функции в точке: враждебный шум, случайный шум; в-третьих, мы провели серию экспериментов на часто встречающихся классах задач, таких как квадратичная задача, логистическая регрессия, SVM, чтобы попытаться определить, соответствует ли реальная природа машинного шума существующей теории. Оказалось, что в реальности (по крайней мере на тех классах задач, которые были рассмотрены в данной работе) машинный шум оказался чем-то средним между враждебным шумом и случайным, в связи с чем текущая теория о влиянии конечности мантиссы на поиск оптимума в задачах безградиентной оптимизации требует некоторой корректировки.

Ключевые слова: конечность мантиссы, безградиентные методы оптимизации, аппроксимация градиента, градиентный спуск, квадратичная задача, логистическая регрессия

Работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации (госзадание), № 075-00337-20-03, номер проекта 0714-2020-0005.

COMPUTER RESEARCH AND MODELING
2023 VOL. 15 NO. 2 P. 259–280
DOI: 10.20537/2076-7633-2023-15-2-259-280

Ки&М

MATHEMATICAL MODELING AND NUMERICAL SIMULATION

UDC: 519.8

Influence of the mantissa finiteness on the accuracy of gradient-free optimization methods

D. D. Vostrikov[a], G. O. Konin[b], A. V. Lobanov[c], V. V. Matyukhin[d]

Moscow Institute of Physics and Technology,
9 Institutskiy per., Dolgoprudny, Moscow region, 141701, Russia

E-mail: [a] danonvostr@gmail.com, [b] koningeorgiy@gmail.com, [c] lobbsasha@mail.ru, [d] vladmatyukh@gmail.com

Received 19.02.2023.
Accepted for publication 23.02.2023.

Gradient-free optimization methods or zeroth-order methods are widely used in training neural networks, reinforcement learning, as well as in industrial tasks where only the values of a function at a point are available (working with non-analytical functions). In particular, the method of error back propagation in PyTorch works exactly on this principle. There is a well-known fact that computer calculations use heuristics of floating-point numbers, and because of this, the problem of finiteness of the mantissa arises.

In this paper, firstly, we reviewed the most popular methods of gradient approximation: Finite forward/central difference (FFD/FCD), Forward/Central wise component (FWC/CWC), Forward/Central randomization on l_2 sphere (FSSG2/CFFG2); secondly, we described current theoretical representations of the noise introduced by the inaccuracy of calculating the function at a point: adversarial noise, random noise; thirdly, we conducted a series of experiments on frequently encountered classes of problems, such as quadratic problem, logistic regression, SVM, to try to determine whether the real nature of machine noise corresponds to the existing theory. It turned out that in reality (at least for those classes of problems that were considered in this paper), machine noise turned out to be something between adversarial noise and random, and therefore the current theory about the influence of the mantissa limb on the search for the optimum in gradient-free optimization problems requires some adjustment.

Keywords: mantissa finiteness, gradient-free optimization, gradient approximation, gradient descent, quadratic problem, logistic regression

Citation: *Computer Research and Modeling*, 2023, vol. 15, no. 2, pp. 259–280 (Russian).

This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Goszadaniye), 075-00337-20-03, project No. 0714-2020-0005.

Введение

Во многих прикладных задачах поиска оптимума функции, таких как обучение нейронных сетей [Chen et al., 2017], обучение с подкреплением [Li et al., 2019], физические эксперименты, бизнес-задачи от сторонних заказчиков, у которых повышенный уровень конфиденциальности, нам часто либо недоступно, либо просто невыгодно аналитически считать градиент исследуемой функции, и мы имеем доступ лишь к значениям функции в требуемой точке, то есть фактически нам доступен лишь оракул нулевого порядка. В таких случаях для совершения процедур типа градиентного спуска (GD) [Ruder, 2016] приходится использовать так называемые методы аппроксимации градиента, или методы безградиентной оптимизации [Gasnikov et al., 2022; Dvinskikh et al., 2022]. Также известно, что в компьютерных вычислениях используются числа с плавающей точкой, которые характеризуются конечной мантиссой [Leung, 2000], то есть при безградиентном подходе мы будем получать не точные значения функции в точке, а зашумленные, причем величина шума будет определяться меткой конечной мантиссы. И важным результатом в данной области будет выявление реальной зависимости точности определения оптимума задачи от точности выдаваемого значения функции при использовании безградиентных методов оптимизации. В данной работе предлагается как раз описать основную существующую теорию по этой теме, после чего проверить ее экспериментально на состоятельность на некоторых классах задач.

Основной вклад и структура статьи

В разделе «Постановка задачи» мы формально определили задачу и ограничения, в которых будем работать. В разделе «Методы аппроксимации градиента» мы описали существующие представления о машинном шуме: враждебном или случайном, перечислили основные методы аппроксимации градиента, такие как FFD (конечная прямая разность), CFD (конечная центральная разность), FWC (прямая покомпонентная разность), CWC (центральная покомпонентная разность), FSSG2 (прямая рандомизация на l_2-сфере), CSSG2 (центральная рандомизация на l_2-сфере), вывели оценки искомых зависимостей в предположении о враждебности машинного шума. В разделе «Эксперименты» мы провели экспериментальные исследования зависимости на таких классах задач, как квадратичная задача, логистическая регрессия, SVM, после чего попробовали интерпретировать результаты и получили, что по крайней мере на рассматриваемых классах задач машинный шум ведет себя как что-то среднее между враждебным и случайным.

Постановка задачи

В данной работе мы рассматриваем стандартную постановку задачи оптимизации

$$\min_{x \in \mathbb{R}^d} f(x) := \mathbb{E}_{\xi} f(x, \xi), \tag{1}$$

в настройке black-box оракула, где последнее означает, что мы предполагаем доступность только оракула нулевого порядка, который выдает значение в запрашиваемой точке с некоторым (враждебным) шумом $\delta(x)$:

$$x = (x_1, \ldots, x_d)$$
$$\downarrow$$
$$\boxed{\text{black-box}}$$
$$\downarrow$$
$$f_\delta(x) = f(x) + \delta(x).$$

Обозначения

Мы используем $\langle x, y\rangle := \sum\limits_{i=1}^{d} x_i y_i$, чтобы определить стандартное скалярное произведение $x, y \in \mathbb{R}^d$. Через $\|x\|_p := \left(\sum\limits_{i=1}^{d} |x_i|^p\right)^{1/p}$ мы обозначаем l_p-норму ($p \geqslant 1$). В частности, при $p = 2$ мы определяем l_2-норму в $\mathbb{R}^d$ следующим образом $\|x\|_2 := \sqrt{\langle x, x\rangle}$. l_p-сферу мы обозначаем как $S_p^d(r) := \left\{x \in \mathbb{R}^d : \|x\|_p = r\right\}$. Для обозначения точности решения задачи мы вводим ε такой, что $\mathbb{E}\left[f\left(\widehat{x}^N\right)\right] - \min\limits_{x\in\mathbb{R}^d} f(x) \leqslant \varepsilon$, где $\widehat{x}^N$ является решением задачи (1).

Предположения

Предположение 1 (выпуклость функции). Функция $f(x)$ является выпуклой, то есть $\forall x', x'' \in \operatorname{dom} f = \{x \in X : f(x) < +\infty\}$ и $\forall \alpha \in [0, 1]$ выполнено неравенство Йенсена:

$$f((1-\alpha)x' + \alpha x'') \leqslant (1-\alpha) f(x') + \alpha f(x''). \tag{2}$$

Предположение 2 (липшицивость функции). Функция $f(x)$ является M-липшицевой непрерывной в l_2-норме, то есть для всех $x, y \in \mathbb{R}^d$ выполнено

$$|f(x) - f(y)| \leqslant M\|x - y\|_2. \tag{3}$$

Последующие предположения актуальны только для гладких задач.

Предположение 3 (липшицивость градиента). Функция $f(x)$ непрерывно дифференцируема, а $\nabla f(x)$ является L-липшицевым непрерывным для всех $x \in \mathbb{R}^d$, то есть выполнено

$$\|\nabla f(x) - \nabla f(y)\|_2 \leqslant L\|x - y\|_2. \tag{4}$$

Предположение 4 (липшицевость гессиана). Функция $f(x)$ дважды непрерывно дифференцируема, а $\nabla^2 f(x)$ является M-липшицевым непрерывным для всех $x \in \mathbb{R}^d$, то есть выполнено

$$\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\|_2 \leqslant M\|x - y\|_2. \tag{5}$$

Предположение 5 (враждебный шум). Шум зависит от входа x, при этом наиболее сильно зашумляя аппроксимацию градиента; единственное ограничение, которое мы накладываем, — ограниченность шума по модулю:

$$\forall x \in X \to |\delta(x)| \leqslant \Delta. \tag{6}$$

Предположение 6 (случайный шум). Для любых выбранных x', x'' шум не зависит от траектории, и второй момент ограничен константой σ^2, т. е.

$$\forall x', x'' \in X \to \mathbb{E}\left[\delta(x')^2\right] \leqslant \sigma^2, \quad \mathbb{E}\left[\delta(x'')^2\right] \leqslant \sigma^2. \tag{7}$$

Методы аппроксимации градиента

Далее все оценки будем получать при предположении 5 (считаем шум враждебным, так как теория случайного шума предполагает, что ошибка накапливаться не будет вне зависимости от величины шума или размерности пространства).

1. Конечная прямая разность (FFD)

Первый метод, который мы анализируем, — это стандартный метод конечных разностей. Приближение $\nabla f(x)$ прямой конечной разностью (FFD) при $x \in \mathbb{R}^d$ вычисляется с использованием множества $X = \{x + \gamma e_i\}_{i=1}^d \cup \{x\}$, где $\gamma > 0$ — шаг, а $e_i = (0, \ldots, 1, \ldots, 0)^T$, $i = \overline{1, d}$ (1 стоит на i-й позиции), следующим образом:

$$\frac{\partial f_\delta}{\partial x_i} \approx \frac{f_\delta(x + \gamma e_i) - f_\delta(x)}{\gamma} = [g(x)]_i. \tag{8}$$

Теорема 1. *Если верны предположения* 1, 3, 5 *и* $g(x)$ *— аппроксимация* $\nabla f(x)$ *конечной прямой разностью, то*

$$[g(x)]_i \approx \frac{\partial f}{\partial x_i} + O\left(\sqrt{\Delta}\right).$$

Доказательство. Оценим (8), используя определение оракула нулевого порядка $f_\delta(x)$:

$$f_\delta(x + \gamma e_i) - f_\delta(x) = f(x + \gamma e_i) - f(x) + \delta(x + \gamma e_i) - \delta(x) \leqslant f(x + \gamma e_i) - f(x) + 2\Delta.$$

Далее разложим $f(x)$ в ряд Тейлора:

$$f(x + \gamma e_i) - f(x) + 2\Delta \approx \langle f'(x),\, \gamma e_i\rangle + \frac{\langle \gamma e_i,\, f''(x)\gamma e_i\rangle}{2} + 2\Delta \approx \gamma \cdot \frac{\partial f}{\partial x_i} + \frac{\gamma^2 L}{2} + 2\Delta.$$

Тогда мы получаем, что

$$\frac{\partial f_\delta}{\partial x_i} \approx \frac{\gamma \cdot \frac{\partial f}{\partial x_i} + \frac{\gamma^2 L}{2} + 2\Delta}{\gamma} \approx \frac{\partial f}{\partial x_i} + \frac{\gamma L}{2} + \frac{2\Delta}{\gamma}.$$

Из приведенных выше оценок видно, что взаимосвязь между шагом γ и шумом Δ играет решающую роль в качестве аппроксимации. В частности, когда Δ равно нулю, то γ может быть выбрано сколь угодно малым и может быть получена близкая аппроксимация $\nabla f(x)$. С другой стороны, когда Δ большое, то малые значения γ приводят к очень неточным аппроксимациям градиента.

Найдем минимум по γ:

$$\frac{\partial\left(\frac{\partial f_\delta}{\partial x_i}\right)}{\partial \gamma} = \frac{L}{2} - \frac{2\Delta}{\gamma^2} = 0 \Rightarrow \gamma = 2\sqrt{\frac{\Delta}{L}}.$$

Значит,

$$[g(x)]_i = \frac{\partial f_\delta}{\partial x_i} = \frac{\partial f}{\partial x_i} + 2\sqrt{\Delta L} = \frac{\partial f}{\partial x_i} + O\left(\sqrt{\Delta}\right).$$

□

Далее перейдем к оценке нормы.

Следствие 1. *Если верны предположения* 1, 3, 5 *и* $g(x)$ *— аппроксимация* $\nabla f(x)$ *конечной прямой разностью, то* $\forall x \in \mathbb{R}^d$

$$\|g(x) - \nabla f(x)\| \leqslant \sqrt{d} \cdot \frac{L\gamma}{2} + \sqrt{d} \cdot \frac{2\Delta}{\gamma}.$$

При $\gamma = \sqrt{2\frac{\Delta}{L}}$ (*оптимальное для FFD*)*, получаем, что*

$$\|g(x) - \nabla f(x)\| \approx \sqrt{d} \cdot O\left(\sqrt{\Delta}\right).$$

Теперь надо понять, как зависит невязка по функции ε от величины шума Δ и размерности пространства d: для этого норму разности $\|g(x) - \nabla f(x)\|$ приравняем к $\frac{\varepsilon}{R}$, где R — расстояние до точки оптимума.

Лемма 1. *Если верны предположения* 1, 3, 5 *и* $g(x)$ *— аппроксимация* $\nabla f(x)$ *конечной прямой разностью, то*

$$\varepsilon = O\left(R\sqrt{\Delta d}\right).$$

Доказательство.

$$\|g(x) - \nabla f(x)\| \leqslant L \cdot \sqrt{\frac{d\Delta}{L}} + \frac{d\Delta}{\sqrt{\frac{d\Delta}{L}}} = \sqrt{L \cdot d \cdot \Delta} + \sqrt{L \cdot d \cdot \Delta} = 2\sqrt{L \cdot d \cdot \Delta} = \frac{\varepsilon}{R} \Rightarrow \varepsilon = O\left(R\sqrt{\Delta d}\right).$$

□

2. Конечная центральная разность (FCD)

Второй метод, который мы анализируем, — это стандартный метод конечных разностей. Приближение $\nabla f(x)$ центральной конечной разностью (FCD) при $x \in \mathbb{R}^d$ вычисляется с использованием множества $X = \{x + \gamma e_i\}_{i=1}^d \cup \{x - \gamma e_i\}_{i=1}^d$, где $\gamma > 0$ — шаг, а $e_i = (0, \ldots, 1 \ldots, 0)^T$, $i = \overline{1, d}$ (1 стоит на i-й позиции), следующим образом:

$$\frac{\partial f_\delta}{\partial x_i} \approx \frac{f_\delta(x + \gamma e_i) - f_\delta(x - \gamma e_i)}{2\gamma} = [g(x)]_i. \tag{9}$$

Теорема 2. *Если верны предположения* 1, 4, 5 *и* $g(x)$ *— аппроксимация* $\nabla f(x)$ *конечной прямой разностью, то*

$$[g(x)]_i \approx \frac{\partial f}{\partial x_i} + O\left(\Delta^{2/3}\right).$$

Доказательство. Оценим (9) по определению оракула нулевого порядка:

$$f_\delta(x + \gamma e_i) - f_\delta(x - \gamma e_i) = f(x + \gamma e_i) - f(x - \gamma e_i) + \delta(x + \gamma e_i) - \delta(x - \gamma e_i) \leqslant$$
$$\leqslant f(x + \gamma e_i) - f(x - \gamma e_i) + 2\Delta. \tag{10}$$

Далее, разложив в ряд Тейлора $f(x + \gamma e_i)$ и $f(x - \gamma e_i)$ и подставив в (10), мы получим

$$f(x + \gamma e_i) - f(x - \gamma e_i) + 2\Delta \approx 2\Delta + f(x) + \langle f'(x),\, \gamma e_i\rangle + \frac{1}{2}\langle \gamma e_i^T,\, f''(x)\gamma e_i\rangle + \frac{1}{6} -$$
$$- \left(f(x) + \langle f'(x),\, -\gamma e_i\rangle + \frac{1}{2}\langle -\gamma e_i,\, f''(x)(-\gamma e_i)\rangle + \frac{1}{6}\right) = 2\langle f'(x),\, \gamma e_i\rangle + \frac{1}{3} + 2\Delta =$$
$$= 2\gamma \cdot \frac{\partial f}{\partial x_i} + \frac{1}{3} + 2\Delta \leqslant 2\gamma \cdot \frac{\partial f}{\partial x_i} + \frac{1}{3}\gamma^3 \cdot M + 2\Delta.$$

Получаем, что

$$\frac{\partial f_\delta}{\partial x_i} \approx \frac{1}{2\gamma} \cdot \left(2\gamma \frac{\partial f}{\partial x_i} + \frac{1}{3}\gamma^3 M + 2\Delta\right) = \frac{\partial f}{\partial x_i} + \frac{1}{6}\gamma^2 M + \frac{\Delta}{\gamma}.$$

Аналогично предыдущему пункту взаимосвязь между шагом γ и шумом Δ играет решающую роль в качестве аппроксимации.

Найдем минимум по γ:

$$\frac{\partial\left(\frac{\partial f_\delta}{\partial x_i}\right)}{\partial \gamma} = \frac{1}{3}\gamma M - \frac{\Delta}{\gamma^2} = 0 \Rightarrow \gamma = \left(\frac{3\Delta}{M}\right)^{1/3}.$$

Значит,

$$[g(x)]_i = \frac{\partial f_\delta}{\partial x_i} = \frac{\partial f}{\partial x_i} + \frac{1}{6}\left(\frac{\Delta}{3M}\right)^{2/3} \cdot M + \frac{2\Delta^{2/3}}{(3M)^{1/3}} = \frac{\partial f}{\partial x_i} + O\left(\Delta^{2/3}\right).$$

□

Теперь перейдем к оценке нормы.

Следствие 2. *Если верны предположения* 1, 4, 5 *и* $g(x)$ *— аппроксимация* $\nabla f(x)$ *конечной прямой разностью, то* $\forall x \in \mathbb{R}^d$

$$\|g(x) - \nabla f(x)\| \leqslant \sqrt{d} \cdot \frac{M\gamma^2}{6} + \sqrt{d} \cdot \frac{\Delta}{\gamma}.$$

При $\gamma = \left(\frac{3\Delta}{M}\right)^{1/3}$ (*оптимальное для CFD*), *получаем, что*

$$\|g(x) - \nabla f(x)\| \approx \sqrt{d} \cdot O\left(\Delta^{2/3}\right).$$

Поймем, как зависит невязка по функции ε от величины шума Δ и размерности пространства d: для этого норму разности $\|g(x) - \nabla f(x)\|$ приравняем к $\frac{\varepsilon}{R}$, где R — это норма точки оптимума.

Лемма 2. *Если верны предположения* 1, 4, 5 *и* $g(x)$ *— аппроксимация* $\nabla f(x)$ *конечной прямой разностью, то*

$$\varepsilon = O\left(R\Delta^{2/3}\sqrt{d}\right).$$

Доказательство.

$$\|g(x) - \nabla f(x)\| \leqslant \sqrt{d} \cdot \frac{M\gamma^2}{6} + \sqrt{d} \cdot \frac{\Delta}{\gamma} = \sqrt{d}\left(\frac{M \cdot (\frac{3\Delta}{M})^{2/3}}{6} + \frac{\Delta}{(\frac{3\Delta}{M})^{1/3}}\right) =$$

$$= \sqrt{d}\left(M^{1/3} \cdot \frac{3^{2/3}}{6} \cdot \Delta^{2/3} + \Delta^{2/3} \cdot \frac{1}{3^{1/3}} \cdot M^{1/3}\right) = \frac{\varepsilon}{R} \Rightarrow \varepsilon = O\left(R\Delta^{2/3}\sqrt{d}\right).$$

□

3. Покомпонентные методы

FWC (forward wise component)

Рассмотрим

$$g(x) = \frac{d \cdot (f_\delta(x + \gamma e_i) - f_\delta(x)) \cdot e_i}{\gamma}, \tag{11}$$

где e_i — рандомный вектор с 1 на i-й из d позиций, $\gamma > 0$, d — коэффициент для несмещенности оценки.

Лемма 3. *Выражение* (11) *является несмещенной оценкой реального градиента.*

Доказательство.

$$\mathbb{E}[g(x, e_i)] = \sum_{i=1}^{d} \frac{1}{d} \cdot \frac{d \cdot (f_\delta(x + \gamma e_i) - f_\delta(x)) \cdot e_i}{\gamma} = \sum_{i=1}^{d} \frac{\partial f}{\partial x_i} \cdot e_i = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ 0 \end{pmatrix} + \ldots + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{pmatrix} = \nabla f(x).$$

□

При предположениях 1, 3, 5 оценки для покомпонентного метода FWC, очевидно, будут полностью аналогичны оценкам для конечного метода FFD.

CWC (central wise component)

Рассмотрим

$$g(x) = \frac{d \cdot (f_\delta(x + \gamma e_i) - f_\delta(x - \gamma e_i)) \cdot e_i}{2\gamma}, \tag{12}$$

где e_i — рандомный вектор с 1 на i-й из d позиций, $\gamma > 0$, d — коэффициент для несмещенности оценки.

Лемма 4. *Выражение* (12) *является несмещенной оценкой реального градиента.*

Доказательство.

$$\mathbb{E}[g(x, e_i)] = \sum_{i=1}^{d} \frac{1}{d} \cdot \frac{d \cdot (f_\delta(x + \gamma e_i) - f_\delta(x - \gamma e_i)) \cdot e_i}{2\gamma} = \sum_{i=1}^{d} \frac{\partial f}{\partial x_i} \cdot e_i =$$

$$= \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ 0 \end{pmatrix} + \ldots + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{pmatrix} = \nabla f(x).$$

□

При предположениях 1, 4, 5 оценки для покомпонентного метода CWC, очевидно, будут полностью аналогичны оценкам для конечного метода CFD.

4. Рандомизация на l_2-сфере

Теперь рассмотрим алгоритм, который использует рандомизированную аппроксимацию: рандомизация на l_2-сфере $S_2^d = \{a \in \mathbb{R}^d : \|a\|_2 = 1\}$.

FSSG2 (forward sphere smoothing gradients l_2)

Пусть $f(x)$ удовлетворяет условиям 1, 3, 5, тогда выберем следующую схему:

$$g(x) = d \cdot \frac{f_\delta(x + \gamma e) - f_\delta(x)}{\gamma} \cdot e, \tag{13}$$

где $\gamma > 0$, $e \sim \left(S_2^d\right)$.

Опять же, коэффициент d в формуле (13) нужен для несмещенности оценки относительно реального градиента функции $f(x)$.

Теорема 3 [Berahas et al., 2022].

$$\|\mathbb{E}[g(x)] - \nabla f(x)\| \leqslant L\gamma + \frac{d\Delta}{\gamma}.$$

Найдем оптимальное γ:

$$\frac{d}{d\gamma}\left(L\gamma + \frac{d\Delta}{\gamma}\right) = L - \frac{d\Delta}{\gamma^2} = 0 \Rightarrow \gamma = \sqrt{\frac{d\Delta}{L}}.$$

Найдем зависимость невязки по функции ε от величины шума Δ и размерности пространства d:

$$\|\mathbb{E}[g(x)] - \nabla f(x)\| \leqslant L \cdot \sqrt{\frac{d\Delta}{L}} + \frac{d\Delta}{\sqrt{\frac{d\Delta}{L}}} = \sqrt{L \cdot d \cdot \Delta} + \sqrt{L \cdot d \cdot \Delta} = 2\sqrt{L \cdot d \cdot \Delta} = \frac{\varepsilon}{R} \Rightarrow$$

$$\Rightarrow \varepsilon = 2R\sqrt{L \cdot d \cdot \Delta}.$$

CSSG2 (central sphere smoothing gradients l_2)

Пусть $f(x)$ удовлетворяет условиям 1, 4, 5, тогда выберем следующую схему:

$$g(x) = d \cdot \frac{f_\delta(x + \gamma e) - f_\delta(x - \gamma e)}{2\gamma} \cdot e, \tag{14}$$

где $\gamma > 0$, $e \sim \left(S_2^d\right)$.

Опять же, коэффициент d в формуле (14) нужен для несмещенности оценки относительно реального градиента функции $f(x)$.

Теорема 4 [Berahas et al., 2022].

$$\|\mathbb{E}[g(x)] - \nabla f(x)\| \leqslant M\gamma^2 + \frac{d\Delta}{\gamma}.$$

Найдем оптимальное γ:

$$\frac{d}{d\gamma}\left(M\gamma^2 + \frac{d\Delta}{\gamma}\right) = 2M\gamma - \frac{d\Delta}{\gamma^2} = 0 \Rightarrow \gamma = \left(\frac{d\Delta}{2M}\right)^{1/3}.$$

Найдем зависимость невязки по функции ε от величины шума Δ и размерности пространства d:

$$\|\mathbb{E}[g(x)] - \nabla f(x)\| = M \cdot \left(\frac{d\Delta}{2M}\right)^{2/3} + \frac{d\Delta}{\left(\frac{d\Delta}{2M}\right)^{1/3}} = \frac{M^{1/3}}{2^{2/3}} \cdot d^{2/3}\Delta^{2/3} + 2^{1/3} \cdot M^{1/3} \cdot d^{2/3} \cdot \Delta^{2/3} =$$

$$= \text{const} \cdot M^{1/3} d^{2/3} \Delta^{2/3} = \frac{\varepsilon}{R} \Rightarrow \varepsilon \sim R \cdot M^{1/3} \cdot d^{2/3} \cdot \Delta^{2/3}.$$

Эксперименты

В этой части работы будут экспериментально проверены теоретические зависимости, полученные в предыдущем параграфе, на некоторых классах задач, а именно: квадратичная задача, логистическая регрессия, метод опорных векторов (SVM).

1. Квадратичная задача

Определение 1. Будем рассматривать квадратичную задачу:

$$f(x) = \frac{1}{2}\langle Ax,\, x\rangle - \langle b,\, x\rangle \longrightarrow \min_{x \in \mathbb{R}^d}, \tag{15}$$

где A — симметричная матрица $d \times d$, $b \in \mathbb{R}^d$. Такие задачи являются выпуклыми и гладкими: константа Липшица L — это максимальное собственное значение матрицы A.

Как следует из определения, мы будем работать в рамках предположения 1 (о выпуклости) и предположения 3 (о липшицевости градиента), поэтому в экспериментах будут исследоваться прямые методы аппроксимации градиента (FFD, FWC, FSSG2).

Теперь следует обсудить метод семплирования квадратичных задач: при создании задачи нам важно задавать размерность задачи d, константу Липшица L (чтобы осуществлять градиентный спуск с оптимальным шагом h), а также точку оптимума x^*, чтобы считать точную невязку по функции ε и 2-норму точки оптимума R (так как эти величины присутствуют в экспериментальных оценках, которые мы будем проверять). В связи с этим получаем алгоритм семплирования квадратичных задач, их последующее решение методом GD с использованием соответствующих аппроксимаций.

1. Вначале создаем случайную ортогональную матрицу $O \in \mathbb{R}^{d\times d}$.

2. Далее создаем диагональную матрицу $D \in \mathbb{R}^{d\times d}$, где максимальный элемент на диагонали равен константе Липшица L, одна половина диагональных элементов имеет близкие к L значения, а другая половина диагональных элементов мала (на самом деле были опробованы совершенно разные варианты создания диагональной матрицы, но в итоге получались идентичные результаты, поэтому выбор остановили на самом простом и быстром варианте).

3. После этого получаем симметричную матрицу $A = O^T \cdot D \cdot O$ с максимальным собственным значением, равным L.

4. Затем задаем вектор $b = Ax^*$, где x^* — требуемая точка оптимума задачи. Эксперименты проводились при разных точках оптимума x^*, однако результаты всегда получались идентичными, поэтому для простоты (в том числе чтобы $R = \|x^*\|_2 = \sqrt{d}$) в качестве точки оптимума было выбрано следующее: $x^* = (1, 1, \ldots, 1)^T \in \mathbb{R}^d$. В итоге мы получили квадратичную задачу с требуемыми размерностью d, константой Липшица L и точкой оптимума x^*.

5. Начальную точку x_{start} будем выбирать равномерно из d-мерного куба $[-10, 10]^d$.

6. Градиентный спуск GD осуществляем с оптимальными $h = \frac{1}{L}$ для FFD, $h = \frac{1}{Ld}$ для FWC и FSSG2, используя разные методы аппроксимации градиента с оптимальным параметром $\gamma = \sqrt{\frac{\Delta}{L}}$, полученным в предыдущем параграфе.

7. Первое, что исследуем, — это зависимость невязки по функции ε от величины шума $\Delta = 10^{-m}$, где m — метка мантиссы (см. рис. 1, 2, 3, 4, 5 и 6).

8. Второе, что исследуем, — это зависимость невязки по функции ε от размерности задачи d (см. рис. 7 и 8). Данный эксперимент проведем лишь для метода FFD.

Теперь приступим к выводам, которые можно сделать на основе проведенных экспериментов.

1. Зависимость $\varepsilon(\Delta)$.

   Было проведено большое количество экспериментов с разными размерностями пространства d, разными точками оптимума x^*, разными константами Липшица L. При этом для всех рассматриваемых методов аппроксимации (FFD, FWC, FSSG2) во всех экспериментах были получены практически идентичные результаты: в то время как теория враждебного шума предсказывает зависимость типа $\varepsilon \sim O\left(\Delta^{0,5}\right)$, экспериментально для класса квадратичных задач была получена зависимость типа $\varepsilon \sim O(\Delta^t)$, где $t = 1 \pm 0{,}02$, то есть зависимость близка к линейной (см. рис. 2, 4, 6).

   Во-первых, это может означать, что существующая теория враждебного шума не слишком подходит для описания шума, создаваемого конечностью мантиссы при вычислениях, и требует доработки.

   Во-вторых, так как на классе квадратичных задач экспериментально была выявлена линейная зависимость, то для этого класса задач может быть предложен алгоритм для ускорения нахождения точного оптимума: считаем оптимум с 2 наименьшими мантиссами, а далее по 2 точкам восстанавливаем зависимость $f^*(\Delta)$ и получаем точное значение оптимума $f^*(0)$. Таким образом, получили довольно красивый и изящный алгоритм.

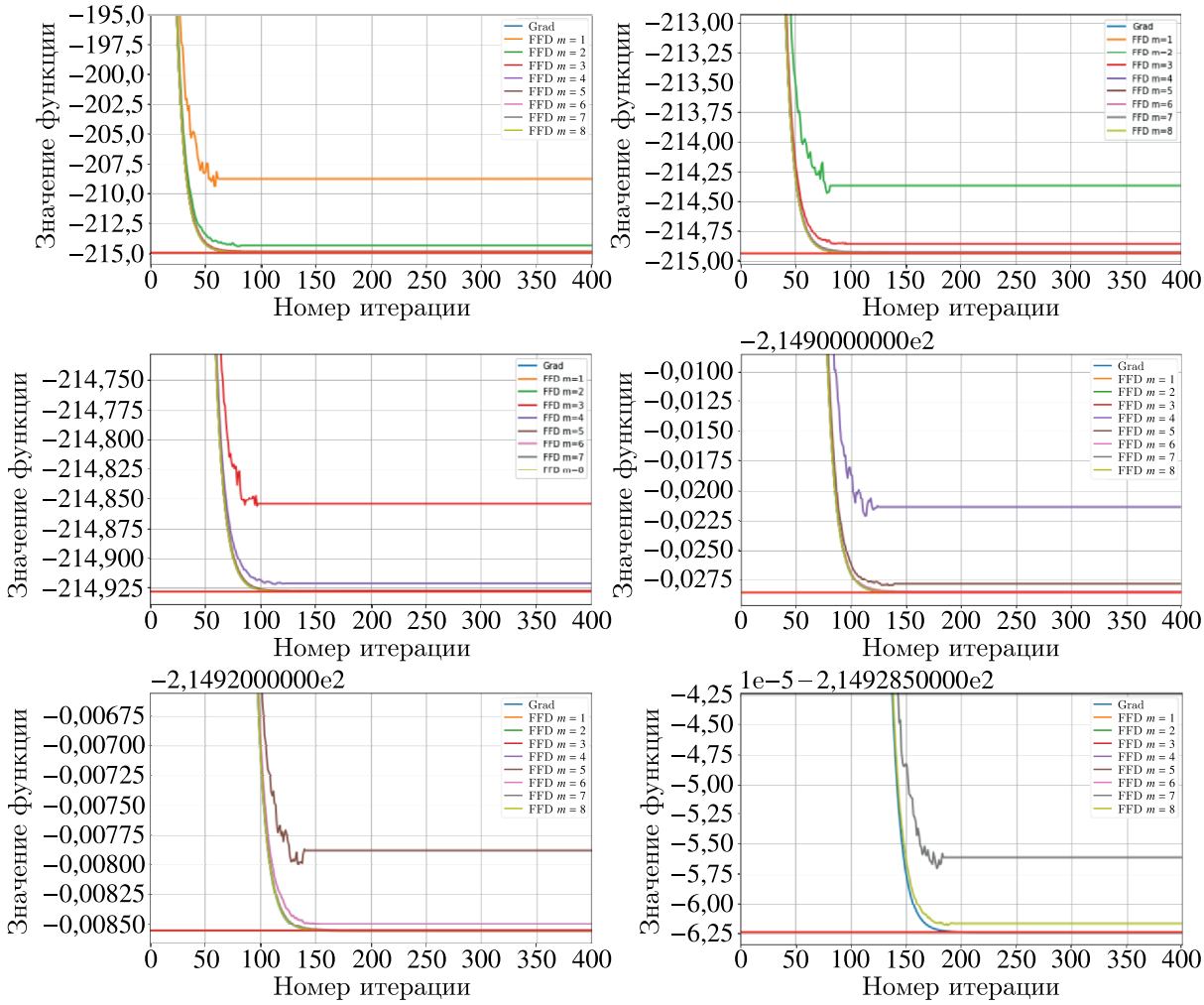

Рис. 1. Графики динамики процедуры GD, использующей аппроксимацию градиента FFD, при разных метках мантиссы $m = \overline{1, 8}$. Также на графиках изображена динамика процедуры GD с аналитической производной. 6 представленных графиков отличаются лишь масштабом для наглядности

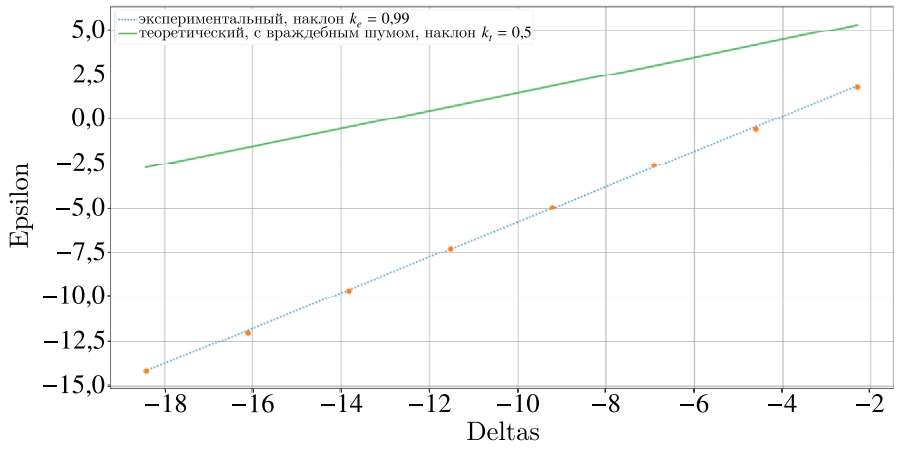

Рис. 2. Зависимость невязки по функции ε от величины шума $\Delta = 10^{-m}$ для FFD в логарифмическом масштабе (зеленая — теоретическая, голубая — экспериментальная). Также на графике указаны коэффициенты наклона кривых

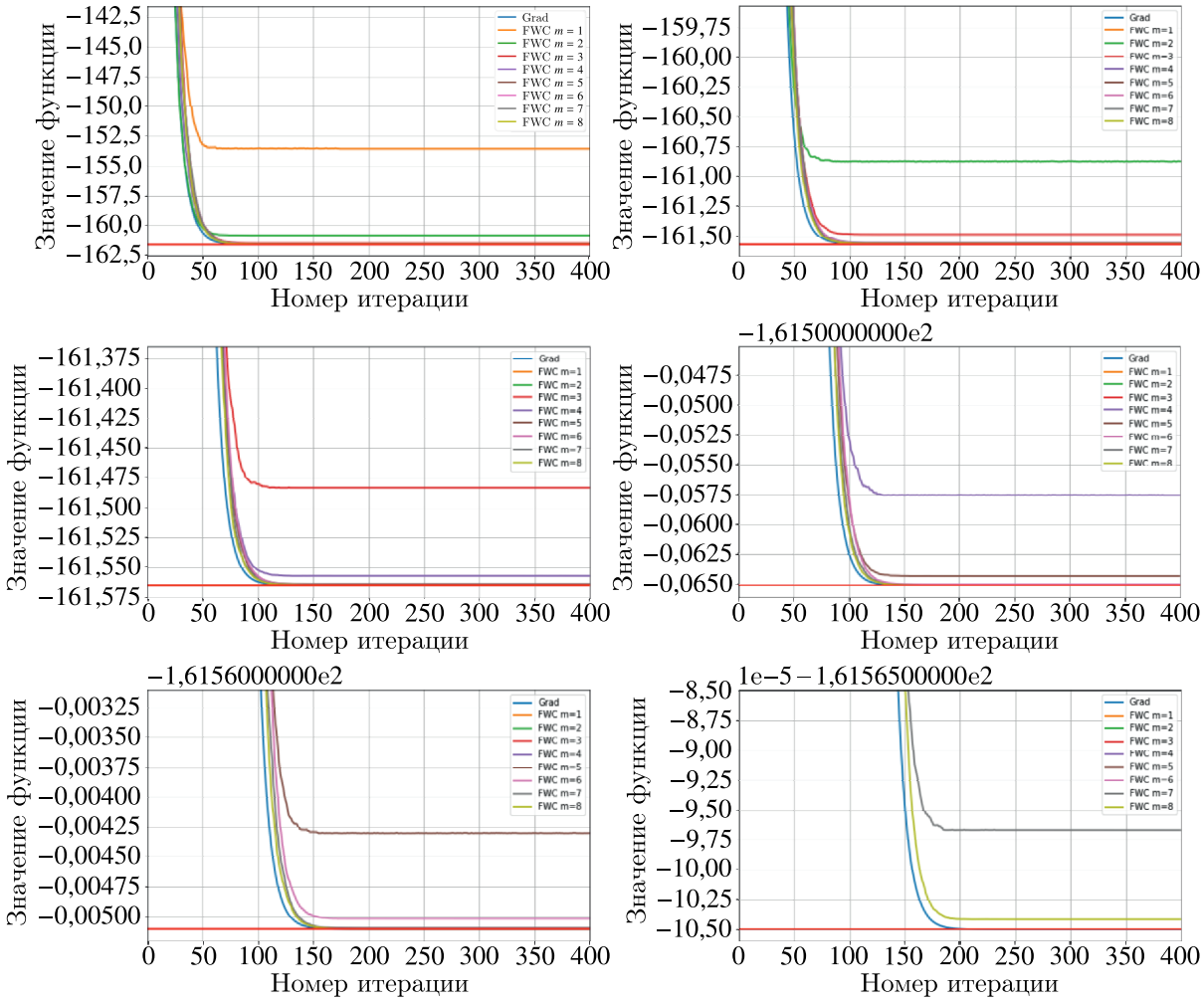

Рис. 3. Графики динамики процедуры GD, использующей аппроксимацию градиента FWC, при разных метках мантиссы $m = \overline{1, 8}$. Также на графиках изображена динамика процедуры GD с аналитической производной. 6 представленных графиков отличаются лишь масштабом для наглядности, а масштаб по итерациям уменьшен в d раз $\left(\text{так как шаг } h = \frac{1}{Ld}\right)$

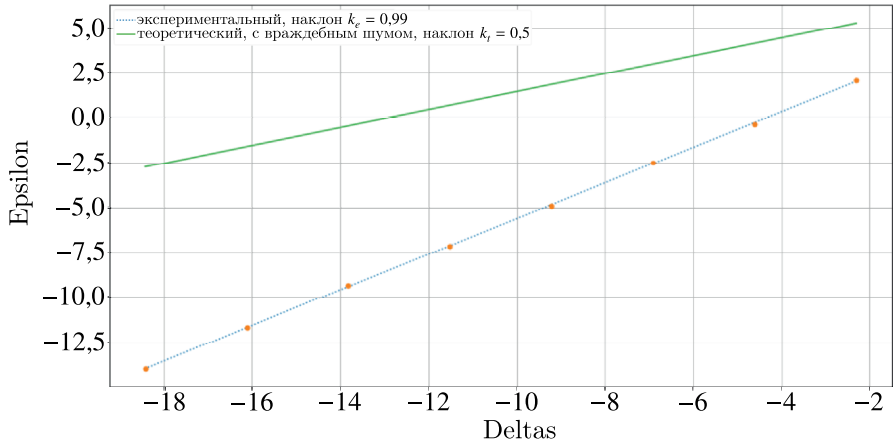

Рис. 4. Зависимость невязки по функции ε от величины шума $\Delta = 10^{-m}$ для FWC в логарифмическом масштабе (зеленая — теоретическая, голубая — экспериментальная). Также на графике указаны коэффициенты наклона кривых

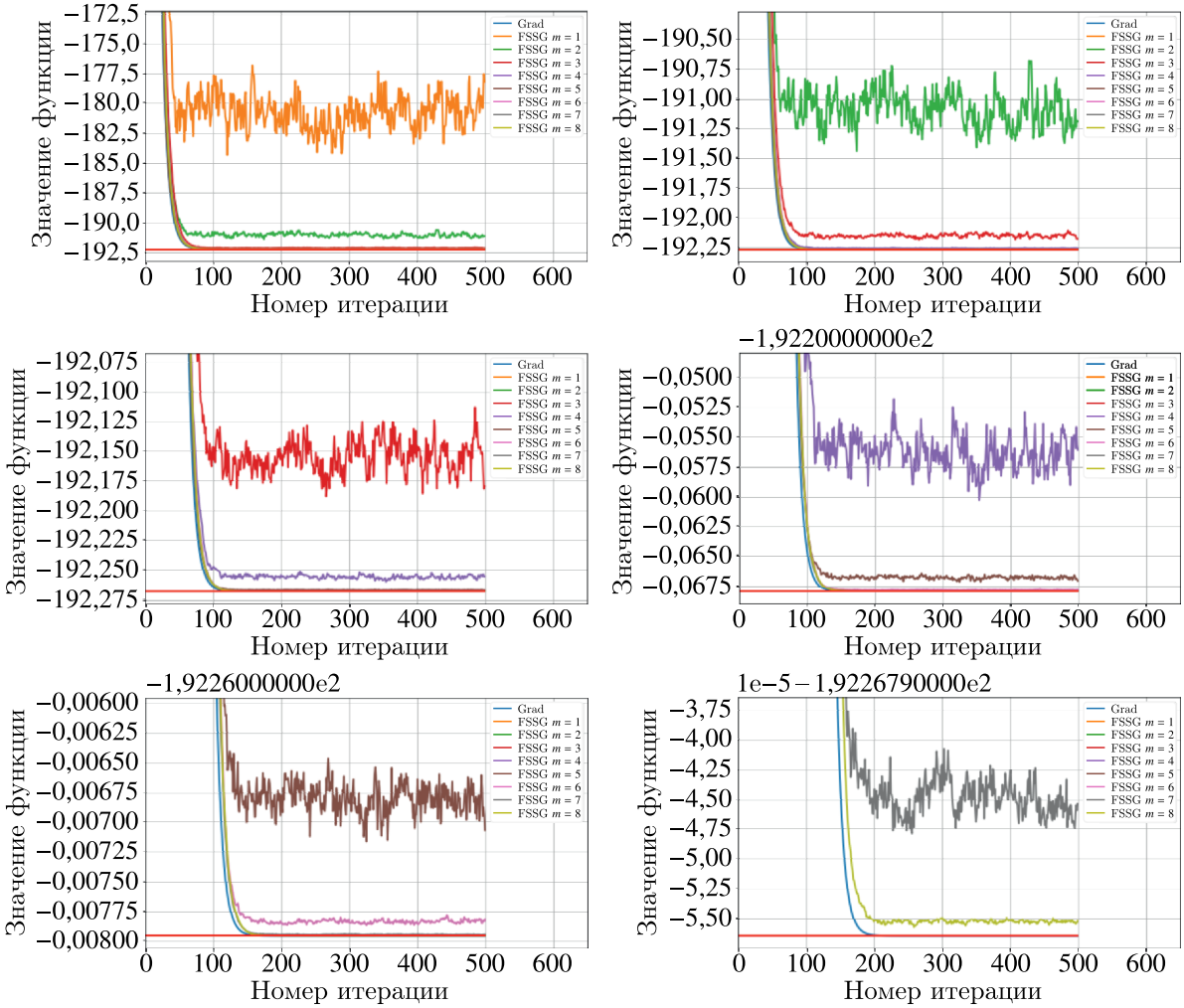

Рис. 5. Графики динамики процедуры GD, использующей аппроксимацию градиента FSSG2, при разных метках мантиссы $m = \overline{1, 8}$. Также на графиках изображена динамика процедуры GD с аналитической производной. 6 представленных графиков отличаются лишь масштабом для наглядности, а масштаб по итерациям уменьшен в d раз $\left(\text{так как шаг } h = \frac{1}{Ld}\right)$

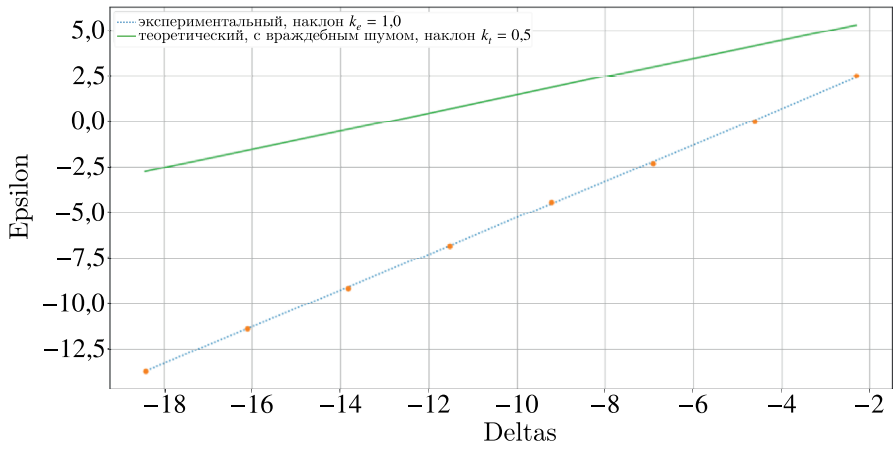

Рис. 6. Зависимость невязки по функции ε от величины шума $\Delta = 10^{-m}$ для FSSG2 в логарифмическом масштабе (зеленая — теоретическая, голубая — экспериментальная). Также на графике указаны коэффициенты наклона кривых

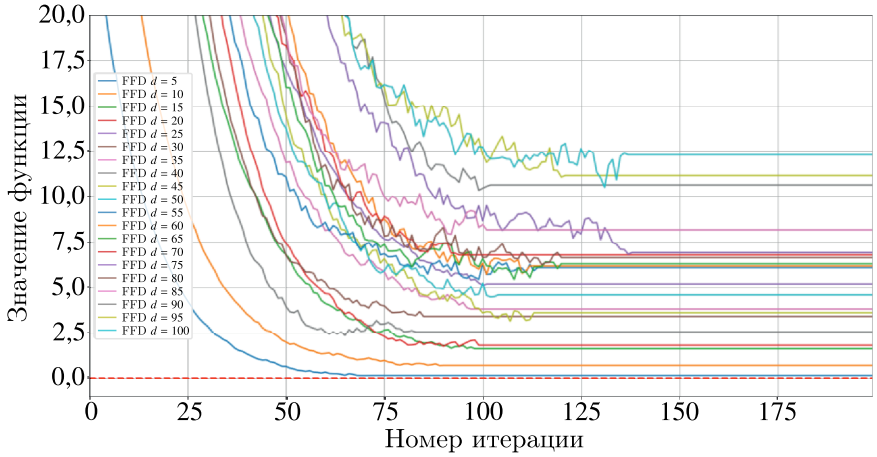

Рис. 7. Графики динамики процедуры GD, использующей аппроксимацию градиента FFD, при одной и той же метке мантиссы $m = 1$ и разных размерностях пространства $d = \overline{5, 100}$

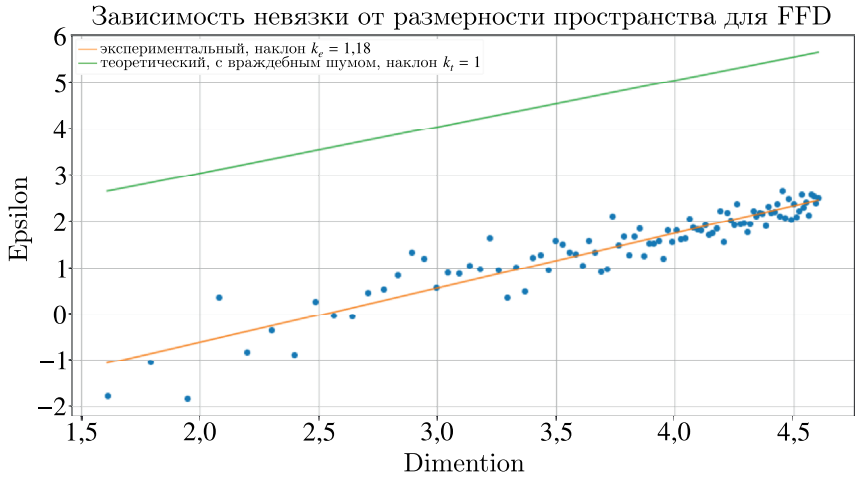

Рис. 8. Зависимость невязки по функции ε от размерности пространства d для FFD в логарифмическом масштабе (зеленая — теоретическая, оранжевая — экспериментальная). Также на графике указаны коэффициенты наклона кривых

2. Зависимость $\varepsilon(d)$.

Опять же, было проведено большое количество экспериментов с разными входными параметрами. И для метода аппроксимации FFD всегда получались похожие результаты: экспериментально для класса квадратичных задач была выявлена зависимость типа $\varepsilon \sim O(d^t)$, где $t = 1 \pm 0{,}2$, то есть она близка к линейной (см. рис. 8). Но и описанная теория предполагает линейную зависимость, ведь, по теореме 2, $\varepsilon = O\left(R\sqrt{d}\right) = \left[x^* = (1, 1, \ldots, 1)^T \in \mathbb{R}^d\right] = O\left(\sqrt{d} \cdot \sqrt{d}\right) = O(d)$. Таким образом, в данном пункте мы не получили никаких расхождений с теорией.

2. Логистическая регрессия

Определение 2. Будем рассматривать функцию потерь задачи логистической регрессии:

$$\text{loss}(\omega) = -\sum_{i=1}^{n}(y_i \log(p_i) + (1 - y_i)\log(1 - p_i)), \tag{16}$$

где $p_i = \omega^T x_i$ — вероятность класса 1, y_i — метка класса i-го элемента, x_i — i-й элемент. Эта функция является выпуклой, гладкой, а также обладает липшицевым гессианом.

Из определения следует, что мы работаем рамках предположения 1 (о выпуклости), предположения 3 (о липшицевости градиента) и предположения 4 (о липшицевости гессиана), поэтому в экспериментах будут исследоваться центральные методы апроксимации градиента (FCD, CWC, CSSG2).

Обсудим способ сэмплирования задачи логистической регресси, а также метод нахождения константы Липшица градиента и гессиана.

1. Генерируем задачу бинарной классификации путем создания объектов 2 классов, распределенных по нормальному закону, вокруг своих центров, выбранных случайным образом. d — размерность признакового пространства.

2. Для того чтобы осуществлять градиентный спуск с оптимальным шагом, нам нужно знать константу Липшица градиента. Чтобы ее определить, генерируем 100 векторов ω из d-мерного куба $[-10, 10]^d$ и далее для каждого вектора найдем максимальное собственное значение матрицы гессиана. Максимальное из этих собственных значений и будет оценкой константы Липшица градиента — L.

3. Для того чтобы определить оптимальный параметр γ, нам потребуется константа Липшица гессиана. Чтобы ее определить, генерируем 100 векторов ω из d-мерного куба $[-10, 10]^d$ и далее для каждой пары ω_i, ω_j считаем $\frac{\|\nabla^2 \text{loss}(\omega_i) - \nabla^2 \text{loss}(\omega_j)\|}{\|\omega_i - \omega_j\|}$. Максимальное из таких частных и будет оценкой константы Липшица гессиана — M.

4. Каждая компонента начальной точки ω_{start} берется из нормального распределения $N(0, 1)$.

5. Градиентный спуск GD осуществляем с оптимальными $h = \frac{1}{L}$ для FCD, $h = \frac{1}{Ld}$ для CWC и CSSG2, используя разные методы аппроксимации градиента с оптимальным параметром $\gamma = \sqrt[1/3]{\frac{3\Delta}{M}}$.

6. В данной постановке задача не имеет заранее известного значения оптимума. Для нахождения оптимального значения функции используем пакетный метод Ньютона.

7. В этом эксперименте мы исследуем зависимость невязки по функции ε от величины шума $\Delta = 10^{-m}$, где m — метка мантиссы.

Приступим к выводам, которые можно сделать на основе экспериментов.

Зависимость $\varepsilon(\Delta)$.

Было проведено большое количество экспериментов с разными размерностями пространства d, разными точками оптимума x^*, разными константами Липшица градиента (L) и гессиана (M). При этом для всех рассматриваемых методов аппроксимации (CFD, CWC, CSSG2) во всех экспериментах были получены практически идентичные результаты: в то время как теория враждебного шума предсказывает зависимость типа $\varepsilon \sim O\left(\Delta^{2/3}\right)$, экспериментально для функций потерь задач бинарной классификации была получена зависимость типа $\varepsilon \sim O(\Delta^t)$, где $t = 1 \pm 0{,}04$, то есть зависимость близка к линейной (см. рис. 10, 12, 14).

Таким образом, выводы, сделанные в эксперименте с квадратичными задачами, справедливы и для задач оптимизации функции логистических потерь. Это значит, что для задач такого типа также может быть использован предложенный в прошлом эксперименте алгоритм нахождения точного оптимума.

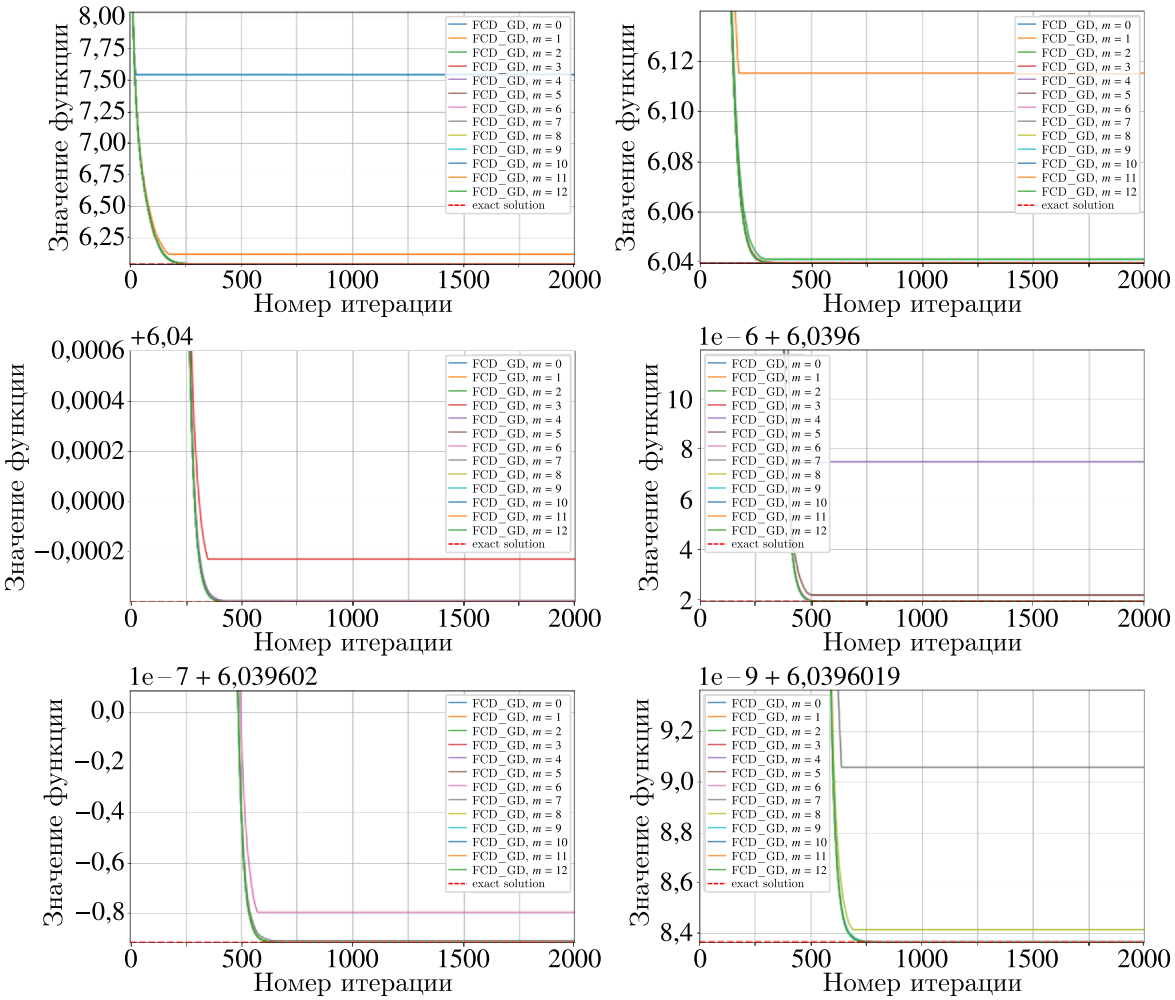

Рис. 9. Графики динамики процедуры GD, использующей аппроксимацию градиента FCD, при разных метках мантиссы $m = \overline{0, 12}$. 6 представленных графиков отличаются лишь масштабом для наглядности. Также на графиках пунктирной линией отмечено точное решение задачи, подсчитанное пакетным методом

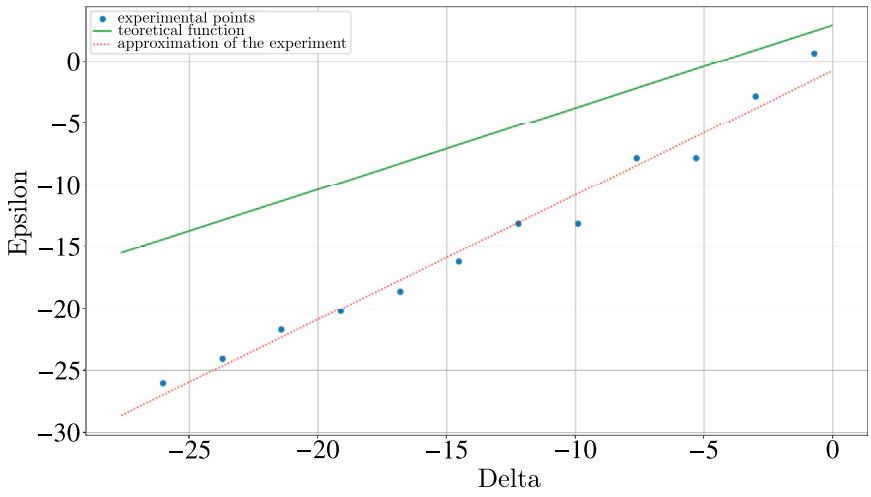

Рис. 10. Зависимость невязки по функции ε от величины шума $\Delta = 10^{-m}$ для FCD в логарифмическом масштабе

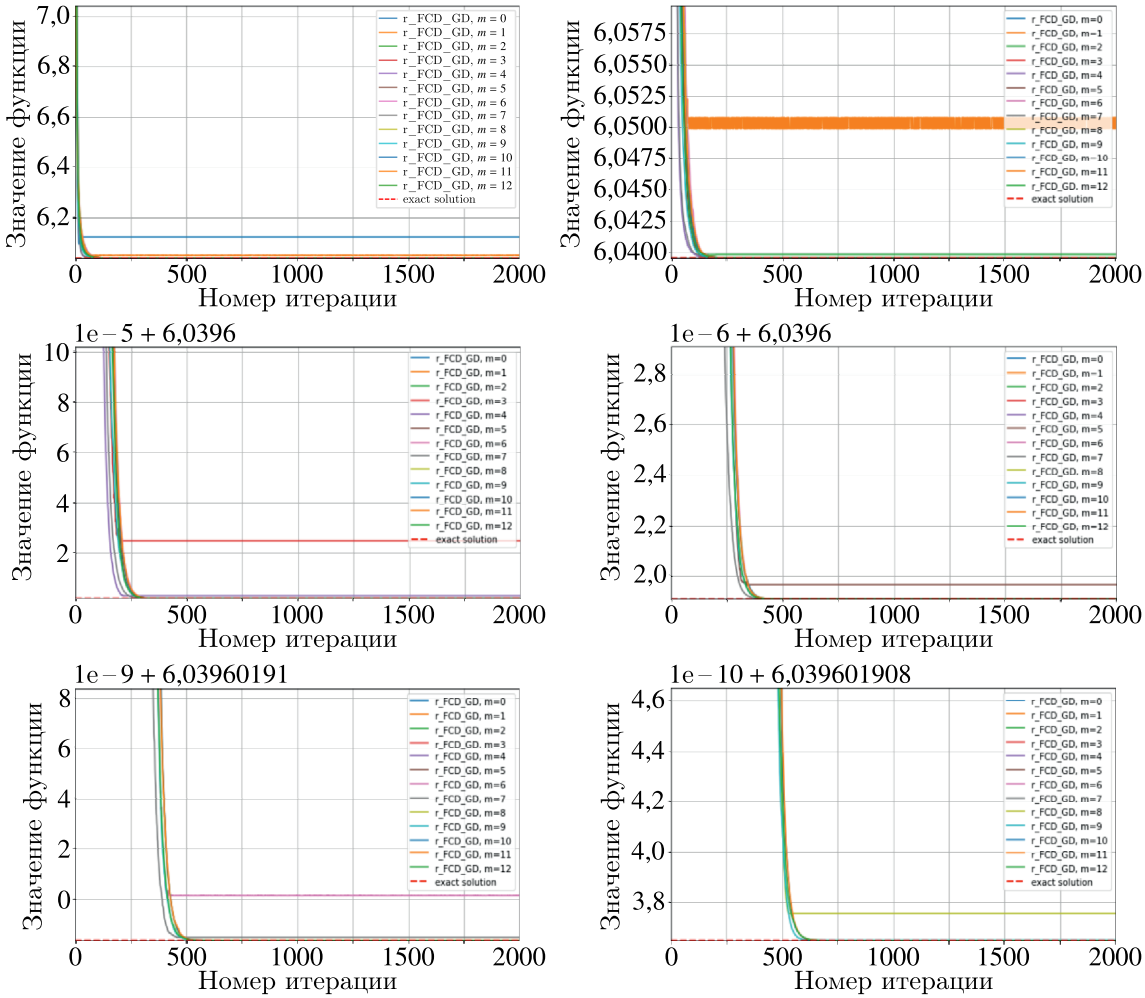

Рис. 11. Графики динамики процедуры GD, использующей аппроксимацию градиента CWC, при разных метках мантиссы $m = \overline{0, 12}$. 6 представленных графиков отличаются лишь масштабом для наглядности, а масштаб по итерациям уменьшен в d раз $\left(\text{так как шаг } h = \frac{1}{Ld}\right)$. Также на графиках пунктирной линией отмечено точное решение задачи, подсчитанное пакетным методом

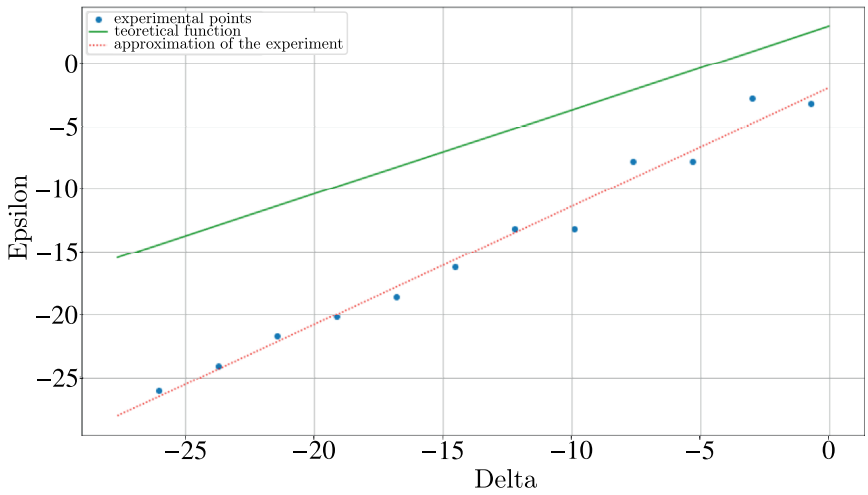

Рис. 12. Зависимость невязки по функции ε от величины шума $\Delta = 10^{-m}$ для CWC в логарифмическом масштабе

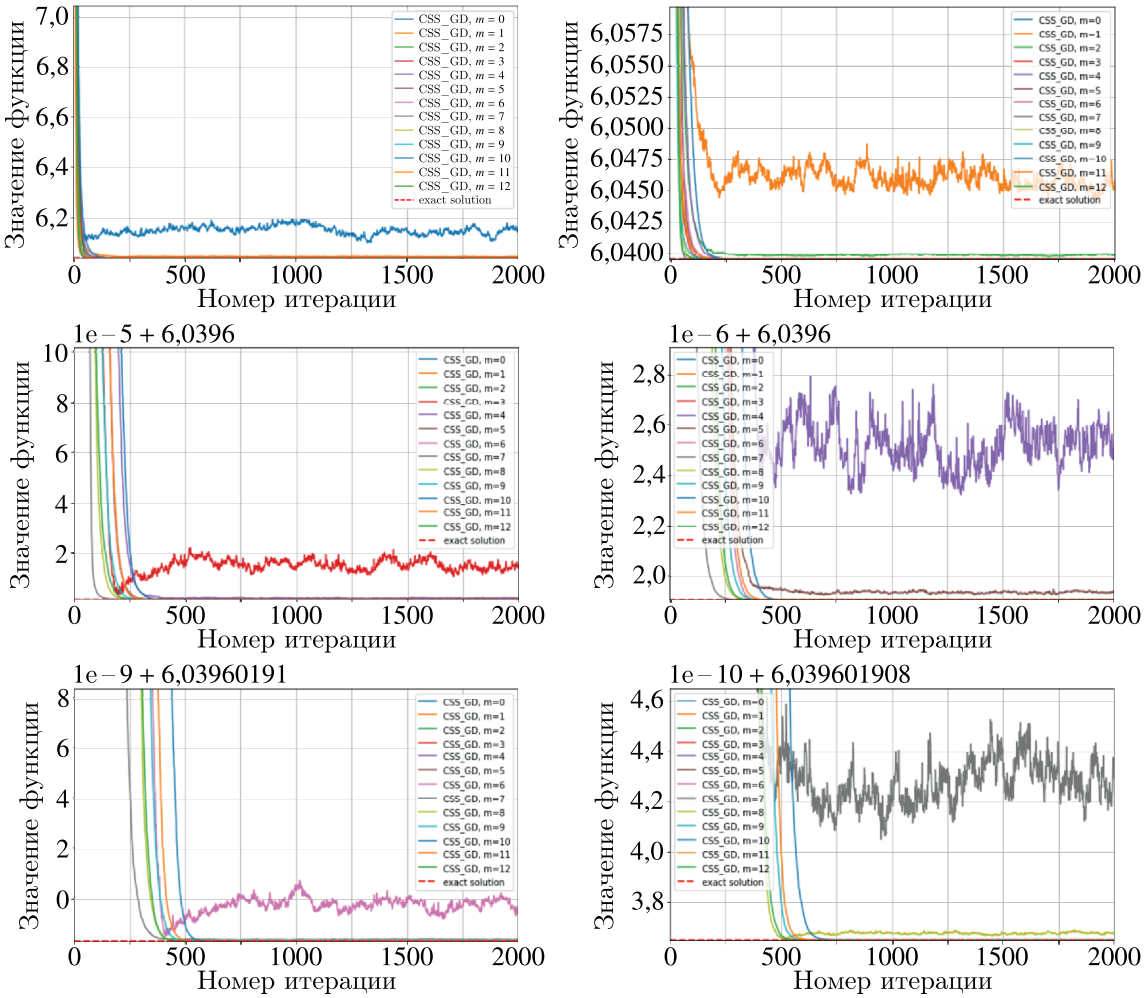

Рис. 13. Графики динамики процедуры GD, использующей аппроксимацию градиента CSSG2, при разных метках мантиссы $m = \overline{0, 12}$. 6 представленных графиков отличаются лишь масштабом для наглядности, а масштаб по итерациям уменьшен в d раз $\left(\text{так как шаг } h = \frac{1}{Ld}\right)$. Также на графиках пунктирной линией отмечено точное решение задачи, подсчитанное пакетным методом

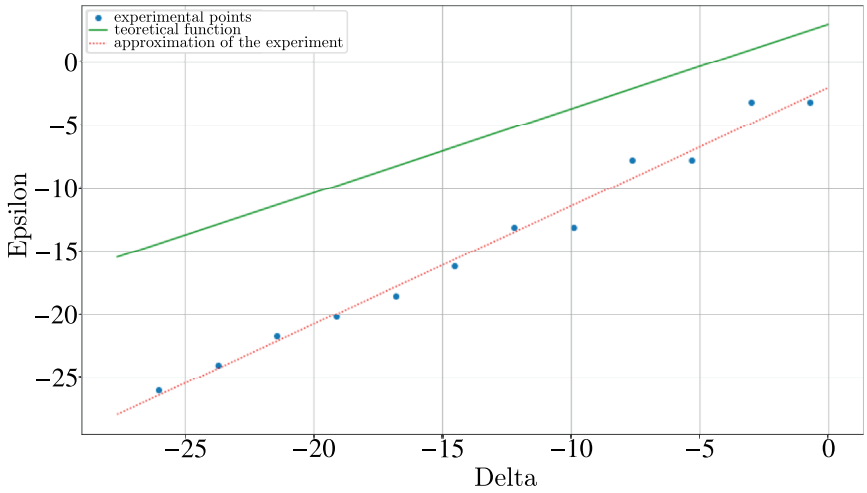

Рис. 14. Зависимость невязки по функции ε от величины шума $\Delta = 10^{-m}$ для CSSG2 в логарифмическом масштабе

3. Метод опорных векторов (SVM)

Определение 3. Будем рассматривать функцию потерь задачи SVM (hinge loss):

$$\frac{1}{2}\sum_{i=1}^{n}\max(0,\ 1 - y_i(\omega X_i - b)) + \lambda\|\omega\|_2, \tag{17}$$

где λ — коэффициент регуляризации. Эта функция является выпуклой, негладкой, но является липшицевой.

Из определения следует, что мы работаем в рамках предположения 1 (о выпуклости) и предположения 2 (о липшицевости функции). Существует несколько подходов для безградиентной оптимизации негладких функций, среди которых рандомизация на l_1-сфере, рандомизация на l_2-сфере, сглаживание и применение методов, характерных для оптимизации гладких задач. В этих экспериментах мы будем исследовать подход, основанный на рандомизации на l_2-сфере.

Обсудим способ сэмплирование задачи.

1. Генерируем две выборки 2 классов, распределенных по нормальному закону, вокруг своих центров. Центры выбираем так, чтобы выборки были линейно разделимыми. d — размерность пространства.

2. Для простоты в этом эксперименте берем γ порядка $1e-2$. Опытным путем было получено, что это значение наиболее близко к оптимальному.

3. Начальную точку ω_{start} берем как вектор размерности d, каждая компонента которого равна $1e-3$.

4. Градиентный спуск GD с рандомизацией на l_2-сфере осуществляем с шагом $h = 1e-2$.

5. В данной постановке задача не имеет заранее известного значения оптимума. Для нахождения оптимального значения функции используем градиентный спуск с рандомизацией на l_2-сфере, который использует незашумленное значение функции.

6. В этом эксперименте мы исследуем зависимость невязки по функции ε от величины шума $\Delta = 10^{-m}$, где m — метка мантиссы.

Приступим к выводам, которые можно сделать на основе экспериментов.

Зависимость $\varepsilon(\Delta)$.

Было проведено большое количество экспериментов с разными размерностями пространства d, разными точками оптимума x^*. По результатам этих экспериментов можно сделать вывод о том, что для задачи SVM теория враждебного шума является несостоятельной. Также зависимость не совпадает с той, что мы получили для гладких задач. Попробуем проинтерпретировать графики (рис. 16, 17).

1. Правая часть графиков характеризуется слишком грубым вычислением значения функции, поэтому как таковой оптимизации не происходит.

2. Левая часть графиков характеризуется уже слишком большим значением мантиссы, которую перекрывает ограничение, вносимое самой вычислительной машиной.

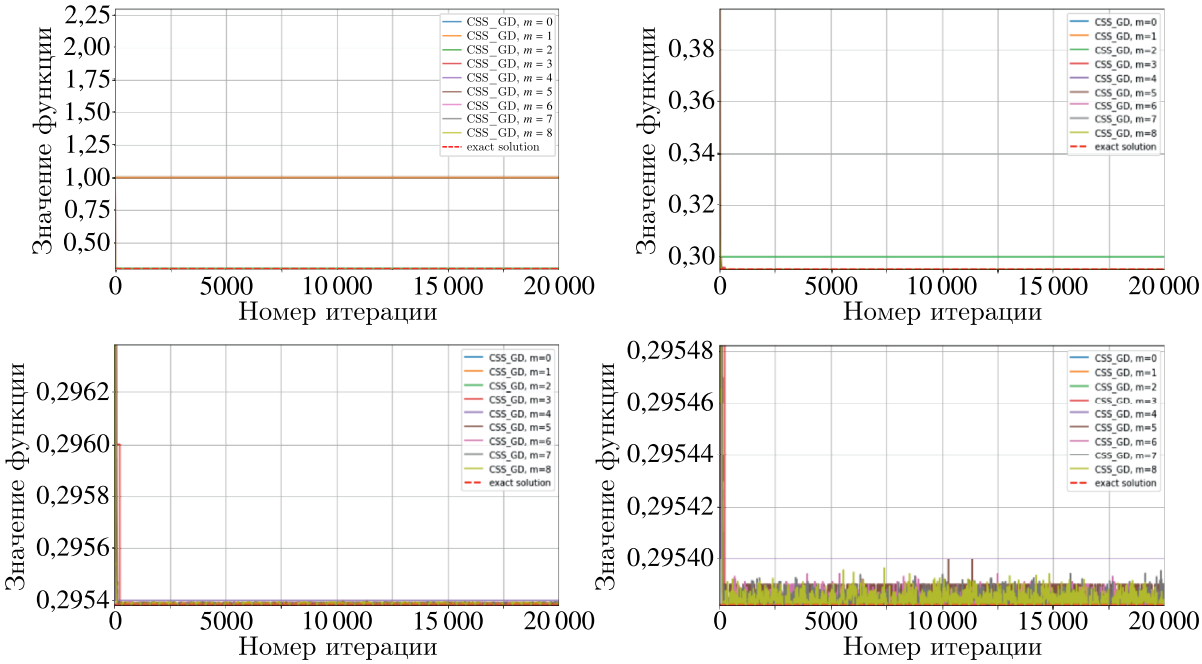

Рис. 15. Графики динамики процедуры GD, использующей аппроксимацию градиента CSSG2, при разных метках мантиссы $m = \overline{0, 8}$. 4 представленных графика отличаются лишь масштабом для наглядности. Также на графиках пунктирной линией отмечено точное решение задачи, подсчитанное на незашумленной задаче

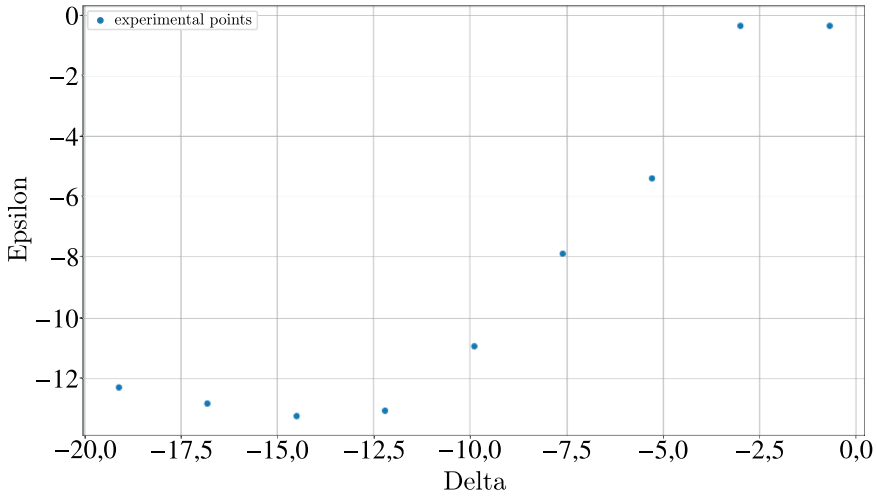

Рис. 16. Зависимость невязки по функции ε от величины шума $\Delta = 10^{-m}$ для CSSG2 в логарифмическом масштабе для задачи SVM

Также стоит отметить особенность графика, при построении которого значение функции бралось от усредненного по всем итерациям значения аргумента (рис. 17). В этом случае мы при не слишком большой мантиссе достигаем оптимального значения функции, подсчитанного без ограничений. В связи с этим можно предложить алгоритм для ускорения нахождения точного оптимума: считаем оптимум, используя неточную функцию, и, посчитав значение неточной функции от усредненного по всем итерациям значения аргумента, получаем оптимальное значение реальной задачи.

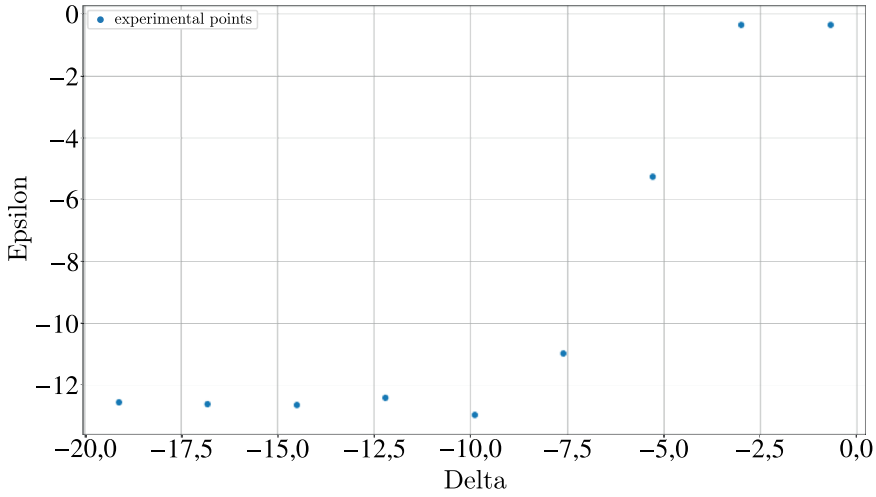

Рис. 17. Зависимость невязки по функции ε от величины шума $\Delta = 10^{-m}$ для CSSG2 в логарифмическом масштабе для задачи SVM. В данном случае значение функции берется от усредненного по всем итерациям значения аргумента

Заключение

В данной работе мы рассмотрели основные подходы безградиентной оптимизации для гладких и негладких задач. По результатам экспериментов было установлено, что природа машинного шума, возникающего из-за конечности мантиссы, довольно плохо изучена, и существующая теория пока что не может полностью его описать. Также было предложено два подхода, позволяющих в ускоренном режиме находить точный оптимум задачи, использующие неточное значение функции.

Список литературы (References)

Akhavan A., Pontil M., Chzhen E., Tsybakov A. A gradient estimator via L1-randomization for online zero-order optimization with two point feedback // arXiv preprint arXiv:2205.13910. — 2022.

Berahas A., Cao L., Choromanski K., Scheinberg K. A theoretical and empirical comparison of gradient approximations in derivative-free optimization // Foundations of Computational Mathematics. — 2022. — Vol. 22, No. 2. — P. 507–560.

Chen P., Zhang H., Sharma Y., Yi J., Hsieh C. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models // Proceedings of the 10th ACM workshop on artificial intelligence and security. — 2017. — P. 15–26.

Dvinskikh D., Tominin V., Tominin I., Gasnikov A. Noisy zeroth-order optimization for non-smooth saddle point problems // Mathematical Optimization Theory and Operations Research: 21st International Conference, MOTOR 2022, Petrozavodsk, Russia, July 2–6, 2022, Proceedings. — Cham: Springer International Publishing, 2022. — P. 18–33.

Gasnikov A., Novitskii A., Novitskii V., Abdukhakimov F., Kamzolov D., Beznosikov A., Takáč M., Dvurechensky P., Gu B. The power of first-order smooth optimization for black-box non-smooth problems // arXiv preprint arXiv:2201.12289. — 2022.

Gasnikov A., Dvinskikh D., Dvurechensky P., Gorbunov E., Beznosikov A., Lobanov A. Randomized gradient-free methods in convex optimization // arXiv preprint arXiv:2211.13566. — 2022.

Leung K. M. Floating-point numbers in digital computers. — 2000.

Li Y., Tang Y., Zhang R., Li N. Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach // arXiv preprint arXiv:1912.09135. — 2019.

Ruder S. An overview of gradient descent optimization algorithms // arXiv preprint arXiv:1609.04747. — 2016.