# Жадный метод оптимизации первого порядка с относительным шумом

Рубцов Денис
rubtsov.dn@phystech.edu

Корнилов Никита
kornilov.nm@phystech.edu

## Abstract

Работа посвящена жадному методу гладкой выпуклой оптимизации первого порядка с градиентами, известными лишь с некоторой относительной погрешностью. С помощью численной техники PEP получена эмпирическая гипотеза о влиянии этой погрешности на скорость сходимости метода.

Ключевые слова : методы оптимизации первого порядка, жадные методы, неточный градиент, относительный шум, Performance Estimation Problem, машинное обучение

## 1 Введение

В данной статье изучаются методы гладкой выпуклой оптимизации первого порядка. Их изучение актуально в связи с высоким успехом их применения во многих приложениях, в том числе в машинном обучении (см., например, в Bottou and Bousquet [2007]). Эти методы (например, Adam (Kingma and Ba [2014])) позволяют эффективно решать задачи многомерной оптимизации, в том числе и невыпуклой.

Во многих ситуациях алгоритмы не имеют доступа к точному градиенту. Так, например, бывает, когда для того, чтобы получить значение градиента, требуется решить другую сложную задачу (например, решить дифференциальное уравнение, см. в Matyukhin et al. [2021]). В данной статье мы сосредоточимся на случае, когда градиенты известны с некоторой относительной погрешностью $\varepsilon \in [0, 1]$:

$$\|\widetilde{\nabla} f(x) - \nabla f(x)\|_2 \leq \varepsilon \|\nabla f(x)\|_2.$$

Часто доказательства скорости сходимости методов оптимизации носят неинтуитивный, техничный характер. Они представляют собой цепочку длинных нетривиальных неравенств, оценивающих наихудший случай. Численное решение задачи поиска наихудшего случая может быть осуществлено с помощью техники Performance Estimation Problem (далее – PEP, см. в Goujaud et al. [2022], Taylor et al. [2017], Taylor [2017]), которая применяется с этой целью и в данной статье.

### 1.1 Определения и обозначения

В данной работе будут решаться задачи оптимизации в следующих предположениях на исследуемые функции $f$.

Предположение 1 Функция $f$ — $\mu$-сильно выпукла, если существует константа $\mu > 0$ такая, что:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2, \quad x, y \in \mathbb{R}^n. \tag{1}$$

Предположение 2 Функция $f$ — $L$-гладкая, если существует константа $L > 0$ такая, что:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2, \quad \forall x, y \in \mathbb{R}^n, \tag{2}$$

или (эквивалентно)

$$\|\nabla f(y) - \nabla f(x)\|_2 \leq L\|y - x\|_2, \quad \forall x, y \in \mathbb{R}^n. \tag{3}$$

Множество $\mu$-сильно выпуклых $L$-гладких функций будем обозначать $\mathcal{F}_{\mu,L}$.

Предположение 3 Будем говорить, что мы имеем доступ к градиенту функции $f$ с некоторой относительной погрешностью $\varepsilon \in [0,1]$, если

$$\|\widetilde{\nabla} f(x) - \nabla f(x)\|_2 \leq \varepsilon \|\nabla f(x)\|_2 \ \forall x \in \mathbb{R}^n. \tag{4}$$

Здесь $\nabla f(x)$ — точное значение градиента функции $f$ в точке $x$, а $\widetilde{\nabla} f(x)$ — доступное нам зашумленное значение градиента.

### 1.2 Жадный метод оптимизации первого порядка (GFOM)

Классические методы оптимизации первого порядка включают в себя градиентный спуск (GD, Cauchy et al. [1847]), метод тяжелого шарика Б.Т. Поляка (HB, Polyak [1963]) и ускоренный метод Ю.Е. Нестерова (NAG, Nesterov [1983]). Так, итерация метода градиентного спуска записывается в следующем виде:

$$x_{k+1} = x_k - \alpha \nabla f(x_k). \tag{GD}$$

Итерация метода тяжелого шарика:

$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}). \tag{HB}$$

Итерация ускоренного метода Ю.Е. Нестерова:

$$x_{k+1} = x_k - h\nabla f(x_k + \frac{k-1}{k+2}(x_k - x_{k-1})) + \frac{k-1}{k+2}(x_k - x_{k-1}). \tag{NAG}$$

В каждом из этих методов результат $k$-ой итерации $x_k$ линейно выражается через результаты предыдущих итераций $(x_s)_{s<k}$ и предыдущие ответы оракула $(\nabla f(x_s))_{s<k}$. В случае методов оптимизации первого порядка ответы оракула — это градиенты $\nabla f(x)$ или неточные градиенты $\widetilde{\nabla} f(x)$.

Так и жадный метод оптимизации первого порядка (Greedy first-order method GFOM, см. в Drori and Taylor [2020]) результат $k$-ой итерации $x_k$ выражает через начальную точку $x_0$ и линейную комбинацию (span) градиентов на предыдущих шагах $(\nabla f(x_s))_{s<k}$

$$x_k = \arg\min_{x \in \mathbb{R}^n} \{f(x) : x \in x_0 + \mathrm{span}\{\nabla f(x_0), ..., \nabla f(x_{k-1})\}\}. \tag{GFOM}$$

Таким образом, на каждой итерации GFOM жадно находит оптимальную линейную комбинацию градиентов функции в точках, полученных на предыдущих итерациях. Оптимальность здесь имеется в виду в смысле минимальности значения функции на этой комбинации.

## 2 Постановка задачи

Мы решаем задачу безусловной оптимизации

$$x_\star = \arg\min_{x \in \mathbb{R}^n} f(x). \tag{5}$$

Конретнее, решаем задачу поиска точки минимума $x_\star$ функции $f(x)$ из класса $\mathcal{F}_{\mu,L}$ $\mu$-сильно выпуклых $L$-гладких функций на пространстве $\mathbb{R}^n$.

Эту задачу будем решать модифицированной версией (NoisyGFOM) метода GFOM, в которой мы имеем доступ к градиентам лишь с относительной погрешностью.

$$x_k = \arg\min_{x \in \mathbb{R}^n} \{f(x) : x \in x_0 + \mathrm{span}\{\ \widetilde{\nabla} f(x_0), ..., \widetilde{\nabla} f(x_{k-1})\}\} \tag{NoisyGFOM}$$

Интерес представляет скорость сходимости данного метода. Исследование сходимости мы будем проводить с помощью анализа наихудшего случая, то есть решая следующую задачу:

$$\max_{f \in \mathcal{F}_{\mu,L},\ (x_k)_{k \leq N} \in (\mathbb{R}^n)^{N+1}} |f(x_N) - f(x_\star)| \tag{6}$$

$$\text{s.t.} \begin{cases} x_0 \in \{x : \|x - x_\star\|_2^2 \leq R^2\} \\ (x_k)_{k \leq N} = x_0 + \mathrm{span}\{(\widetilde{\nabla} f(x_k))_{k \leq N}\} \\ \|\widetilde{\nabla} f(x) - \nabla f(x)\|_2 \leq \varepsilon \|\nabla f(x)\|_2 \end{cases}$$

Это задача максимизации невязки $|f(x_N) - f(x_\star)|$ на $N$-ой итерации, в случае когда начальная точка итерационного алгоритма $x_0$ лежит не очень далеко от искомой точки минимума $x_\star$, а следующие точки итерационного алгоритма сгенерированы с помощью NoisyGFOM. Обратим внимание, что это бесконечномерная задача оптимизации на классе функций $\mathcal{F}_{\mu,L}$. Оказывается, однако, что с помощью плодотворных идей интерполяции (Taylor et al. [2017]), такую задачу можно свести к конечномерной задаче полуопределенного программирования (semidefinite programming, далее — SDP). Эта техника называется Performance Estimation Problem (PEP). Используемый солвер — PEPit (Goujaud et al. [2022]).

# 3 Вычислительный эксперимент

В данной секции сформулирована гипотеза о зависимости сходимости метода NoisyGFOM в зависимости от шума градиента $\varepsilon$. Продемонстрированы результаты численных экспериментов и сформулированы выводы о непротиворечивости гипотезы.

## 3.1 Проверяемая гипотеза

Скорость сходимости обычного градиентного спуска (GD):

$$f(x_N) - f(x_\star) = O\left(\left(\frac{1-\frac{\mu}{L}}{1+\frac{\mu}{L}}\right)^N\right) \tag{7}$$

Скорость сходимости ускоренных методов (включая GFOM, см.Hildebrand et al. [2021]):

$$f(x_N) - f(x_\star) = O\left(\left(\frac{1-\sqrt{\frac{\mu}{L}}}{1+\sqrt{\frac{\mu}{L}}}\right)^N\right) \tag{8}$$

Сформулируем гипотезу о сходимости метода NoisyGFOM.

Предположение 4 Скорость сходимости метода NoisyGFOM зависит от соотношения $\left(\frac{\mu}{L}\right)^{\alpha(\varepsilon)}$, где $\alpha(\varepsilon)$ - некоторая монотонно возрастающая функция, причем $\alpha|_{\varepsilon=0} = \frac{1}{2}$, а $\alpha|_{\varepsilon=1} = 1$.

$$f(x_N) - f(x_\star) = O\left(\left(\frac{1-\left(\frac{\mu}{L}\right)^{\alpha(\varepsilon)}}{1+\left(\frac{\mu}{L}\right)^{\alpha(\varepsilon)}}\right)^N\right) \tag{9}$$

То есть при возрастании относительного шума $\varepsilon$ от 0 до 1 сходимость метода ухудшается от сходимости ускоренных методов до сходимости обычного градиентного спуска.

## 3.2 Ход эксперимента

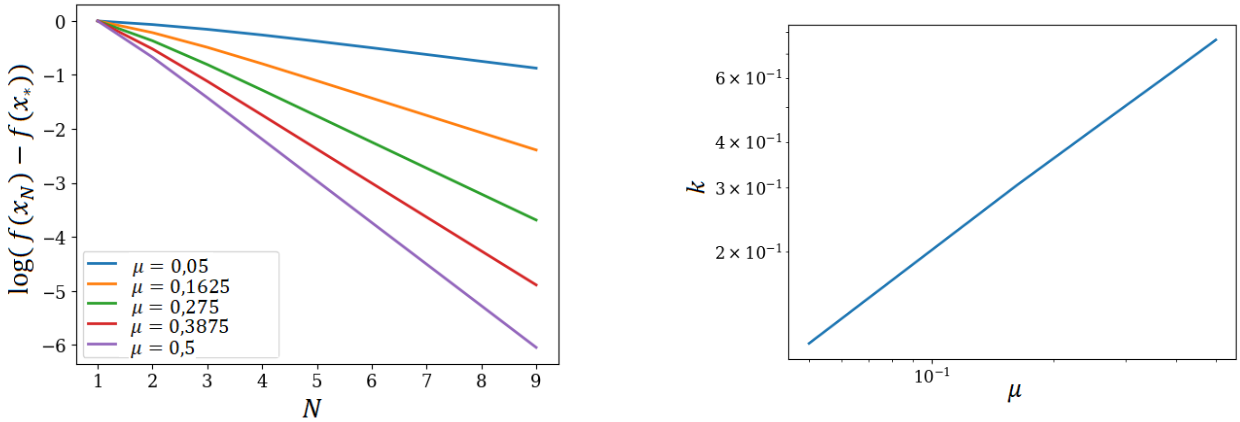
При небольших $\frac{\mu}{L}$ гипотеза принимает следующий вид:

$$f(x_N) - f(x_\star) = O\left(\left(1 - 2\left(\frac{\mu}{L}\right)^{\alpha(\varepsilon)}\right)^N\right) \tag{10}$$

Проверим линейную сходимость алгоритма. Для этого построим график невязки $f(x_N) - f(x_\star)$ в зависимости от числа итераций $N$ при фиксированном $\varepsilon$ и различных $\mu$. Видно, что при всех $\mu$ при больших номерах итераций графики являются прямыми, т.е. сходимость линейная.

При фиксированном $\varepsilon$ и $\frac{\mu}{L}$ $\log\|f(x_N) - f_\star\| \sim N \cdot k(\varepsilon,\mu)$, где $k(\varepsilon,\mu) = \log\left(1 - 2\left(\frac{\mu}{L}\right)^{\alpha(\varepsilon)}\right) \sim \left(\frac{\mu}{L}\right)^{\alpha(\varepsilon)}$. Тогда $\log k(\varepsilon,\mu) = \alpha(\varepsilon)\log\left(\frac{\mu}{L}\right)$. Из предыдущего графика найдем угловые коэффициенты прямых $k(\varepsilon,\mu)$ для различных $\mu$ (с помощью линейной регрессии). Построим график $k(\mu)$ в логарифмическом масштабе. С помощью линейной регрессии определим коэффициент $\alpha(\varepsilon)$.

Повторим эксперимент для различных $\varepsilon$ и построим график $\alpha(\varepsilon)$ (2).




(a) График зависимости невязки $(f(x_N) - f_\star)$ от номера итерации $N$ при $\varepsilon = 0.5$. Демонстрирует линейную сходимость.

(b) График зависимости график $k(\mu)$ при $\varepsilon = 0.5$. Демонстрирует зависимость скорости сходимости только лишь от $\left(\frac{\mu}{L}\right)^{\alpha(\varepsilon)}$.

Рис. 1: Вычислительные эксперименты

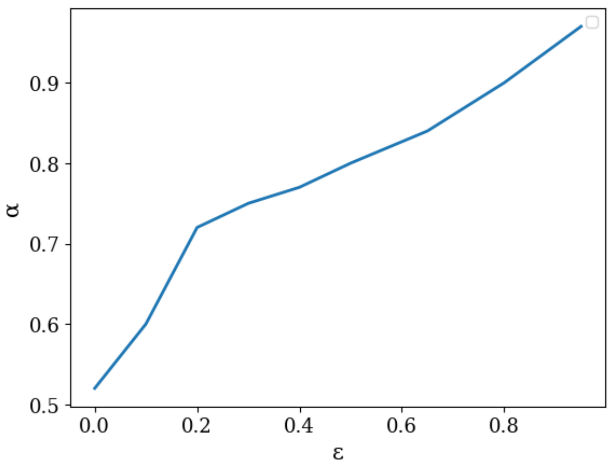



Рис. 2: График зависимости $\alpha(\varepsilon)$.

### 3.3 Результаты и их обсуждение

Результаты вычислительных экспериментов не противоречат гипотезе. Скорость сходимости метода NoisyGFOM зависит от шума $\varepsilon$. При увеличении шума скорость сходимости падает от быстрой (8) до свойственной обычному градиентному спуску (7). Итак,

$$f(x_N) - f_\star = O\left(\left(\frac{1 - \left(\frac{\mu}{L}\right)^{\alpha(\varepsilon)}}{1 + \left(\frac{\mu}{L}\right)^{\alpha(\varepsilon)}}\right)^N\right), \ \alpha|_{\varepsilon=0} = \frac{1}{2}, \ \alpha|_{\varepsilon=1} = 1 \tag{11}$$

## 4 Заключение

В данной работе было исследовано влияние зашумленного градиента (3) на жадный метод оптимизации первого порядка NoisyGFOM. Так, при появлении шума в градиентах метод сохраняет свою линейную сходимость при $\varepsilon \in [0, 1]$, уменьшая скорость сходимости с увеличением шума. При шуме $\varepsilon \approx 1$ скорость сходимости NoisyGFOM совпадает со скоростью сходимости градиентного спуска (7), а при $\varepsilon \approx 0$ - со скоростью сходимости ускоренных методов (8). В наши дальнейшие планы входит теоретическое обоснование полученных экспериментально результатов.

## Список литературы

Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. Advances in neural information processing systems, 20, 2007.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

Vladislav Matyukhin, Sergey Kabanikhin, Maxim Shishlenin, Nikita Novikov, Artem Vasin, and Alexander Gasnikov. Convex optimization with inexact gradients in hilbert space and applications to elliptic inverse problems. In International Conference on Mathematical Optimization Theory and Operations Research, pages 159–175. Springer, 2021.

Baptiste Goujaud, Céline Moucer, François Glineur, Julien Hendrickx, Adrien Taylor, and Aymeric Dieuleveut. Pepit: computer-assisted worst-case analyses of first-order optimization methods in python. arXiv preprint arXiv:2201.04040, 2022.

Adrien B Taylor, Julien M Hendrickx, and François Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. Mathematical Programming, 161:307–345, 2017.

Adrien B Taylor. Convex interpolation and performance estimation of first-order methods for convex optimization. PhD thesis, Catholic University of Louvain, Louvain-la-Neuve, Belgium, 2017.

Augustin Cauchy et al. Méthode générale pour la résolution des systemes d'équations simultanées. Comp. Rend. Sci. Paris, 25(1847):536–538, 1847.

Boris T Polyak. Gradient methods for the minimisation of functionals. USSR Computational Mathematics and Mathematical Physics, 3(4):864–878, 1963.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate o (1/k** 2). Doklady Akademii Nauk SSSR, 269(3):543, 1983.

Yoel Drori and Adrien B Taylor. Efficient first-order methods for convex minimization: a constructive approach. Mathematical Programming, 184(1):183–220, 2020.

Roland Hildebrand, Eugenia Vorontsova, Alexander Gasnikov, and Fyodor Stonyakin. Выпуклая оптимизация. 2021.