
Optimal Gradient Methods with Relative Inexactness

Рубцов Денис
rubtsov.dn@phystech.edu

Корнилов Никита
kornilov.nm@phystech.edu

Abstract

Работа посвящена ускоренным методам гладкой выпуклой оптимизации первого порядка с градиентами, известными лишь с некоторой относительной погрешностью. Проведен обзор полученных ранее теоретических результатов об оценках максимально допустимой погрешности, сохраняющей линейную сходимость методов. С помощью анализа численного решения эквивалентной задачи полуопределенного программирования (техника PER) показана достижимость этих оценок.

Ключевые слова методы первого порядка, ускоренные методы, неточный градиент, относительный шум, Performance Estimation Problem

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. При этом во многих ситуациях алгоритмы не имеют доступа к точной информации о градиенте. Так, например, бывает, когда для того, чтобы получить значение градиента, требуется решить другую сложную задачу (например, решить дифференциальное уравнение Matyukhin et al. [2021]). В данной статье мы сосредоточимся на ситуации, когда градиенты известны с некоторой относительной погрешностью $\hat{\varepsilon} \in [0, 1]$

$$\|\tilde{\nabla} f(x) - \nabla f(x)\|_2 \leq \hat{\varepsilon} \|\nabla f(x)\|_2$$

Особое место в данной статье отведено технике Performance Estimation Problem (далее – PER, см. в Goujaud et al. [2022], Taylor et al. [2017], Taylor [2017]). С помощью этого инструмента исследования в данной статье демонстрируется точность полученной ранее в статье Kornilov et al. [2023] теоретической оценки максимальной погрешности, сохраняющей сходимость метода ISTM. Более того, с помощью этой техники проводится поиск оптимальных методов первого порядка, основанных на принципе жадных алгоритмов (Goujaud et al. [2023], Kim and Fessler [2018]).

1.1 Определения и обозначения

Определение 1.1 Скалярное произведение элементов $x, y \in \mathbb{R}^n$: $\langle x, y \rangle := \sum_{k=1}^n x_k y_k$.

Определение 1.2 ℓ_2 -норма элемента $x \in \mathbb{R}^n$: $\|x\|_2 := \left(\sum_{k=1}^n x_k^2 \right)^{1/2}$.

Определение 1.3 Функция f выпукла, если

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall x, y \in \mathbb{R}^n. \quad (1)$$

Определение 1.4 Функция f - μ -сильно выпукла, если существует константа $\mu > 0$ такая, что:

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

Определение 1.5 Функция f — L -гладкая, если существует константа $L > 0$ такая, что:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|_2^2, \quad \forall x, y \in \mathbb{R}^n, \quad (3)$$

или (эквивалентно)

$$\|\nabla f(y) - \nabla f(x)\|_2 \leq L \|y - x\|_2, \quad \forall x, y \in \mathbb{R}^n. \quad (4)$$

Определение 1.6 Будем говорить, что мы имеем доступ к градиенту функции f с некоторой относительной погрешностью $\hat{\varepsilon}$, если

$$\|\tilde{\nabla} f(x) - \nabla f(x)\|_2 \leq \hat{\varepsilon} \|\nabla f(x)\|_2, \quad \forall \hat{\varepsilon} \in [0, 1]. \quad (5)$$

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

Определение 1.7 Класс всех L -гладких, μ -сильно выпуклых функций будем обозначать $\mathcal{F}_{\mu,L}$. Класс всех L -гладких, выпуклых функций будем обозначать $\mathcal{F}_{0,L}$.

1.2 Intermediate Similar Triangle Method with Relative Noise in Gradient

В статье Kornilov et al. [2023] рассматривается алгоритм Intermediate Similar Triangle Method. В этой работе была получена наилучшая из известных на данный момент оценок максимальной относительной погрешности, сохраняющей сходимость этого метода в случае сильной выпуклости $\hat{\varepsilon} \lesssim (\mu/L)^{1/2}$. Более того, в выпуклом случае было показано, что при малом числе итераций N алгоритм сходится со скоростью $\sim \frac{1}{N^p}$. Однако при числе итераций $N : N^p \hat{\varepsilon}^2 \gtrsim 1$, близость к оптимальному значению функции выходит на плато $\hat{\varepsilon}^2 L R_0^2$, где $R_0 = \|x^0 - x^*\|_2$, а L — константа гладкости. Эти оценки были подтверждены с помощью численных экспериментов, использующих технику PER, в той же статье. В нашей статье проводятся дополнительные численные эксперименты.

Algorithm 1 Intermediate Similar Triangle Method (ISTM).

Require: Initial point x^0 , number of iterations N , smoothness constant $L > 0$, and step size parameter $a \geq 1$, intermediate parameter $p \in [1, 2]$.

- 1: Set $A_0 = \alpha_0 = 0, y^0 = z^0 = x^0$.
 - 2: for $k = 0, 1, \dots, N - 1$ do
 - 3: Set $\alpha_{k+1} = \frac{(k+2)^{p-1}}{2aL}, A_{k+1} = \alpha_{k+1} + A_k$.
 - 4: $x^{k+1} = \frac{1}{A_{k+1}} (A_k y^k + \alpha_{k+1} z^k)$.
 - 5: $z^{k+1} = z^k - \alpha_{k+1} \tilde{\nabla} f(x^{k+1})$.
 - 6: $y^{k+1} = \frac{1}{A_{k+1}} (A_k y^k + \alpha_{k+1} z^{k+1})$.
 - 7: end for
- Ensure: y^N .
-

1.3 PER

Часто доказательства скорости сходимости методов оптимизации носят неинтуитивный, техничный характер. Они представляют собой цепочку длинных нетривиальных неравенств, оценивающих наихудший случай. Если сразу задаться поиском этого наихудшего случая, то мы придем к бесконечномерной задаче оптимизации на некотором классе функций (например, на $\mathcal{F}_{\mu,L}$). 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 и называется PER.

1.4 О поиске оптимальных методов

Во всех статьях (в том числе, в Kornilov et al. [2023]) техника PER применяется для конкретных методов оптимизации первого порядка. Более интересной и важной задачей, которой посвящена данная статья, является поиск оптимального метода первого порядка с зашумленными градиентами. Эта задача является задачей оптимизации на пространстве методов. Мы будем рассматривать лишь

определенный класс «линейных» методов так, что эта задача превратится в задачу поиска коэффициентов $\alpha, \alpha_0, \alpha_1, \dots, \alpha_{k-1}$ линейной комбинации, выражающей k -ое приближение точки минимума $x^k = \alpha x^0 + \alpha_0 * \tilde{\nabla} f(x^0) + \dots + \alpha_{k-1} * \tilde{\nabla} f(x^{k-1})$ через начальную точку итерационного процесса x^0 и ответы оракула $\tilde{\nabla} f(x^i)$. Вообще говоря, α_i зависят от номера итерации k и могут быть найдены жадным алгоритмом на каждой итерации.

2 Постановка задачи

Мы решаем задачу безусловной оптимизации

$$x^* = \arg \min_{x \in \mathbb{R}^d} f(x)$$

Функция $f(x) \in \mathcal{F}_{\mu, L}$. Мы имеем доступ к оракулу первого порядка, предоставляющему информацию о приближенном значении градиента $\mathcal{O}^{(f)}(x) = \tilde{\nabla} f(x)$.

3 Вычислительный эксперимент

Для алгоритма ISTM в статье Kornilov et al. [2023] была доказана следующая теорема.

Пусть функция f – выпуклая и L -гладкая с относительным шумом $\hat{\varepsilon} \in [0, 1]$. Тогда после $N \geq 1$ итераций алгоритма ISTM с промежуточным параметром $p \in [1, 2]$ и

$$a = C \cdot \max \{1, N^p \hat{\varepsilon}^2\}, \quad C = 2304 \quad (6)$$

имеем невязку на N -ой итерации, равную

$$f(y^N) - f(x^*) \leq \frac{8aLR_0^2}{(N+1)^p}, \quad (7)$$

где $R_0 = \|x^0 - x^*\|_2$ или, если подставить параметр a из (6), имеем

$$f(y^N) - f(x^*) \leq 8C \cdot \max \left\{ \frac{LR_0^2}{N^p}, \hat{\varepsilon}^2 LR_0^2 \right\}. \quad (8)$$

Теорема говорит о том, что при номерах итераций N таких, что $N^p \hat{\varepsilon}^2 \leq 1$, алгоритм сходится со скоростью $\sim \frac{1}{N^p}$. При больших номерах итераций, из-за наличия относительной погрешности в определении градиента, алгоритм выходит на вынужденное плато по невязке, равное $f(y^N) - f(x^*) = 8C \cdot \hat{\varepsilon}^2 LR_0^2$. Далее мы будем рассматривать случай $p = 2$ в силу его наиболее быстрой сходимости.

В этой теореме смущает довольно большая константа $C = 2304$. Однако она была получена авторами теоремы с применением не самых жестких неравенств. Численные эксперименты с применением техники РЕР демонстрируют, что фактически эта константа значительно меньше. Цель нашего вычислительного эксперимента — уточнить эту константу.

Для этого мы решим следующую задачу оптимизации с фиксированными $N, \hat{\varepsilon}$ и a .

$$\tau^N := \max_{n, f, x^0} f(x^N) - f(x^*), \quad (9)$$

с условиями выпуклости и L -гладкости функции $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $0 \in \nabla f(x^*)$ и

$$\|x^0 - x^*\|_2^2 \leq R^2, \quad (10)$$

$$\|\tilde{\nabla} f(x^k) - \nabla f(x^k)\|_2^2 \leq \hat{\varepsilon}^2 \|\nabla f(x^k)\|_2^2, \quad k = \overline{0, N-1}, \quad (11)$$

$$x^{k+1}, y^{k+1}, z^{k+1} = \text{ISTMstep}(x^k, y^k, z^k, \tilde{\nabla} f(x^k)), \quad k = \overline{0, N-1}. \quad (12)$$

Таким образом, мы ищем функцию f из класса $\mathcal{F}_{0, L}$, дающую на N -ой итерации алгоритма ISTM с относительным шумом $\hat{\varepsilon}$ наибольшее значение невязки τ^N . Как уже ранее в этой статье обсуждалось, такая задача может быть переформулирована в конечномерную задачу полуопределенного программирования (техника РЕР) и решена численно с помощью специальных решателей. Мы будем использовать MOSEK solver ApS [2019] и PEPit framework Goujaud et al. [2022].

3.1 Качественная демонстрация сходимости

С помощью решателей исследовалась задача (9) для $p = 2$. Было показано, что теоретическое плато по невязке действительно достигается. Плато тем больше, чем больше шум.

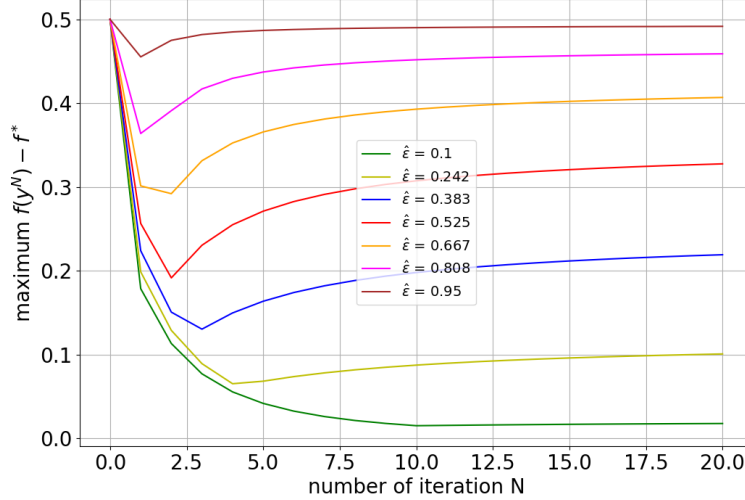


Рис. 1: Графики зависимостей максимальной невязки τ^N в зависимости от номера итерации N алгоритма ISTM для различных значений шумов $\hat{\varepsilon}$ и при фиксированных $p = 2, L = 1, R = 1$.

3.2 Уточнение константы C из теоремы 3

Согласно теореме 3, алгоритм ISTM перестает монотонно убывать на итерации $N_{theory}(a, \hat{\varepsilon}, p) = (\frac{a}{C\hat{\varepsilon}^2})^{\frac{1}{p}}$. Проведем численный эксперимент, который при фиксированном $p = 2$ и для разных параметров a будет искать номер N_{per} , при котором последовательность невязок $\{\tau^N\}$, полученных из (9), перестает убывать. Построим график $N_{per}^2 \hat{\varepsilon}^2(a) = \frac{1}{C}a$. Его линейность будет подтверждать теоретический характер зависимости из 3, а угловой коэффициент позволит уточнить (уменьшить) константу C .

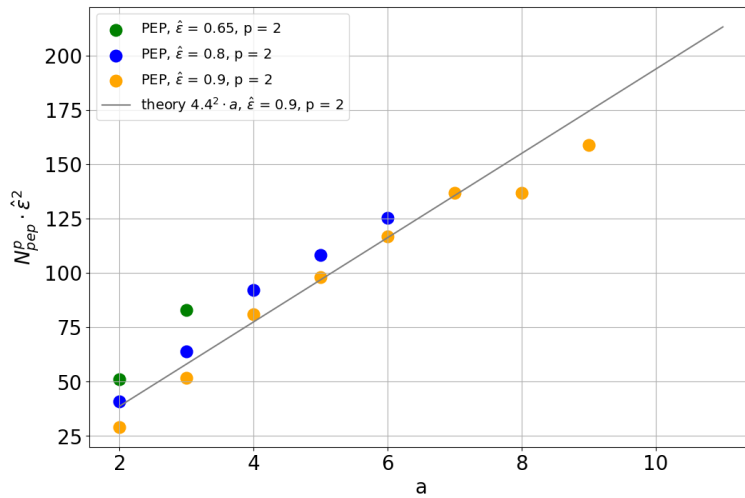


Рис. 2: Графики зависимости номера итерации N_{per} , при которой начинается расхождение метода ISTM, в зависимости от параметра a для разных значений шума $\hat{\varepsilon}$. Для наглядного подтверждения теории, график построен в координатах $N_{per}^2 \hat{\varepsilon}^2(a)$, в которых график становится линейным. Параметры $p = 2, L = 1, R = 1$.

Параметры прямой линии, проходящей через точки, были определены с помощью линейной регрессии (метода наименьших квадратов). Полученный результат $C \approx \frac{1}{4.4^2}$ существенно меньше завышенной теоретической константы $C = 2304$.

Список литературы

- 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.
- Nikita Kornilov, Eduard Gorbunov, Mohammad Alkousa, Fedor Stonyakin, Pavel Dvurechensky, and Alexander Gasnikov. Intermediate gradient methods with relative inexactness. arXiv preprint arXiv:2310.00506, 2023.
- Baptiste Goujaud, Aymeric Dieuleveut, and Adrien Taylor. On fundamental proof structures in first-order optimization. In 2023 62nd IEEE Conference on Decision and Control (CDC), pages 3023–3030. IEEE, 2023.
- Donghwan Kim and Jeffrey A Fessler. Generalizing the optimized gradient method for smooth convex minimization. SIAM Journal on Optimization, 28(2):1920–1950, 2018.
- Mosek ApS. Mosek optimization toolbox for matlab. User’s Guide and Reference Manual, Version, 4(1), 2019.