КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2022 Т. 14 № 2 С. 377–398
DOI: 10.20537/2076-7633-2022-14-2-377-398

Ки&М

МАТЕМАТИЧЕСКИЕ ОСНОВЫ И ЧИСЛЕННЫЕ МЕТОДЫ МОДЕЛИРОВАНИЯ

УДК: 519.853.62

Тензорные методы внутри смешанного оракула для решения задач типа min-min

П. А. Остроухов[1,2]

[1]Московский физико-технический институт,
Россия, 141701, Московская область, г. Долгопрудный, Институтский переулок, д. 9
[2]Институт проблем передачи информации им. А. А. Харкевича Российской академии наук,
Россия, 127051, г. Москва, Большой Каретный переулок, д. 19, стр. 1

E-mail: ostroukhov@phystech.edu

Получено 11.02.2022.
Принято к публикации 13.02.2022.

В данной статье рассматривается задача типа min-min: минимизация по двум группам переменных. Данная задача в чем-то похожа на седловую (min-max), однако лишена некоторых сложностей, присущих седловым задачам. Такого рода постановки могут возникать, если в задаче выпуклой оптимизации присутствуют переменные разных размерностей или если какие-то группы переменных определены на разных множествах. Подобная структурная особенность проблемы дает возможность разбивать ее на подзадачи, что позволяет решать всю задачу с помощью различных смешанных оракулов. Ранее в качестве возможных методов для решения внутренней или внешней задачи использовались только методы первого порядка или методы типа эллипсоидов. В нашей работе мы рассматриваем данный подход с точки зрения возможности применения алгоритмов высокого порядка (тензорных методов) для решения внутренней подзадачи. Для решения внешней подзадачи мы используем быстрый градиентный метод.

Мы предполагаем, что внешняя подзадача определена на выпуклом компакте, в то время как для внутренней задачи мы отдельно рассматриваем задачу без ограничений и определенную на выпуклом компакте. В связи с тем, что тензорные методы по определению используют производные высокого порядка, время на выполнение одной итерации сильно зависит от размерности решаемой проблемы. Поэтому мы накладываем еще одно условие на внутреннюю подзадачу: ее размерность не должна превышать 1000. Для возможности использования смешанного оракула нам необходимы некоторые дополнительные предположения. Во-первых, нужно, чтобы целевой функционал был выпуклым по совокупности переменных и чтобы его градиент удовлетворял условию Липшица также по совокупности переменных. Во-вторых, нам необходимо, чтобы целевой функционал был сильно выпуклый по внутренней переменной и его градиент по внутренней переменной удовлетворял условию Липшица. Также для применения тензорного метода нам необходимо выполнение условия Липшица p-го порядка ($p > 1$). Наконец, мы предполагаем сильную выпуклость целевого функционала по внешней переменной, чтобы иметь возможность использовать быстрый градиентный метод для сильно выпуклых функций.

Стоит отметить, что в качестве метода для решения внутренней подзадачи при отсутствии ограничений мы используем супербыстрый тензорный метод. При решении внутренней подзадачи на компакте используется ускоренный проксимальный тензорный метод для задачи с композитом.

В конце статьи мы также сравниваем теоретические оценки сложности полученных алгоритмов с быстрым градиентным методом, который не учитывает структуру задачи и решает ее как обычную задачу выпуклой оптимизации (замечания 1 и 2).

Ключевые слова: тензорные методы, гладкость высокого порядка, сильная выпуклость, смешанный оракул, неточный оракул

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

COMPUTER RESEARCH AND MODELING
2022 VOL. 14 NO. 2 P. 377–398
DOI: 10.20537/2076-7633-2022-14-2-377-398

Ки&М

MATHEMATICAL MODELING AND NUMERICAL SIMULATION

UDC: 519.853.62

Tensor methods inside mixed oracle for min-min problems

P. A. Ostroukhov[1,2]

[1]Moscow Institute of Physics and Technology,
9 Institutskiy per., Dolgoprudny, Moscow Region, 141701, Russia
[2]Institute for Information Transmission Problems of Russian Academy of Sciences,
19/1 Bol'shoy Karetnyy per., Moscow, 212705, Russia

E-mail: ostroukhov@phystech.edu

Received 11.02.2022.
Accepted for publication 13.02.2022.

In this article we consider min-min type of problems or minimization by two groups of variables. In some way it is similar to classic min-max saddle point problem. Although, saddle point problems are usually more difficult in some way. Min-min problems may occur in case if some groups of variables in convex optimization have different dimensions or if these groups have different domains. Such problem structure gives us an ability to split the main task to subproblems, and allows to tackle it with mixed oracles. However existing articles on this topic cover only zeroth and first order oracles, in our work we consider high-order tensor methods to solve inner problem and fast gradient method to solve outer problem.

We assume, that outer problem is constrained to some convex compact set, and for the inner problem we consider both unconstrained case and being constrained to some convex compact set. By definition, tensor methods use high-order derivatives, so the time per single iteration of the method depends a lot on the dimensionality of the problem it solves. Therefore, we suggest, that the dimension of the inner problem variable is not greater than 1000. Additionally, we need some specific assumptions to be able to use mixed oracles. Firstly, we assume, that the objective is convex in both groups of variables and its gradient by both variables is Lipschitz continuous. Secondly, we assume the inner problem is strongly convex and its gradient is Lipschitz continuous. Also, since we are going to use tensor methods for inner problem, we need it to be p-th order Lipschitz continuous ($p > 1$). Finally, we assume strong convexity of the outer problem to be able to use fast gradient method for strongly convex functions.

We need to emphasize, that we use superfast tensor method to tackle inner subproblem in unconstrained case. And when we solve inner problem on compact set, we use accelerated high-order composite proximal method.

Additionally, in the end of the article we compare the theoretical complexity of obtained methods with regular gradient method, which solves the mentioned problem as regular convex optimization problem and doesn't take into account its structure (Remarks 1 and 2).

Keywords: tensor methods, high-order smoothness, strong convexity, mixed oracle, inexact oracle

Citation: *Computer Research and Modeling*, 2022, vol. 14, no. 2, pp. 377–398 (Russian).

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

Введение

На данный момент в оптимизации существует множество методов для различных классических постановок (седловые задачи, выпуклая оптимизация), которые являются наиболее общими. Для этих задач известны нижние оценки [Nesterov, 2004; Nemirovsky, Yudin, 1983] и известны (суб)оптимальные алгоритмы, достигающие этих нижних оценок. Для дальнейшего ускорения имеющихся алгоритмов научное сообщество все больше начинает смотреть на структуру имеющейся задачи. Если говорить о выпуклой оптимизации, то сравнительно недавно возникла задача типа min-min, которая схожа с седловой задачей, хотя не обладает некоторыми сложностями, которые возникают при решении седловых задач. Однако задача с подобной структурой является относительно новой и еще недостаточно изучена [Jungers, Trélat, Abou-Kandil, 2011; Konur, Farhangi, 2017; Bolte et al., 2020; Gladin et al., 2021; Gladin, Alkousa, Gasnikov, 2021]. Основная мотивация для решения таких задач заключается в транспортных приложениях [Gasnikov, Gasnikova, 2020]. Формально стандартная постановка задачи типа min-min выглядит следующим образом:

$$\min_{x \in Q_x} \min_{y \in Q_y} F(x, y), \tag{1}$$

где $Q_x \subseteq R^m$ и $Q_y \subseteq \mathbb{R}^n$ — некоторые непустые выпуклые множества, $F(x, y)$ является выпуклой по совокупности переменных.

В работах [Gladin et al., 2021; Gladin, Alkousa, Gasnikov, 2021] рассматривается возможность применения смешанного оракула к данной задаче: задача разбивается на внутреннюю и внешнюю. Обе задачи решаются методами первого или нулевого порядка. Но, так как внутренняя задача решается с определенной точностью, в решении внешней задачи используется неточный оракул. В нашей работе мы тоже используем концепцию смешанного оракула, однако мы исследуем возможность решения внутренней задачи методом высокого порядка, а внешней задачи — методом первого порядка.

Как известно, методы нулевого порядка имеют более высокую скорость сходимости, однако при большой размерности стоимость одной итерации становится слишком высока по сравнению с градиентными методами. К примеру, у метода эллипсоидов скорость сходимости для выпуклых функций составляет $O(n^2 \ln(\varepsilon^{-1}))$ [Nemirovsky, Yudin, 1983]. Поэтому область применения методов типа эллипсоидов ограничивается размерностью задачи $n \lesssim 100$. С другой стороны, градиентные методы сходятся медленнее относительно точности, но их скорость сходимости не зависит от размерности. К примеру, скорость сходимости быстрого градиентного метода в выпуклом гладком случае составляет $O(\varepsilon^{-1/2})$ [Nesterov, 2004]. Таким образом, при $n \gg 100$ градиентные методы выигрывают у методов нулевого порядка.

Наконец, рассмотрим методы высокого порядка. Так как для любой функции построение алгоритма оптимизации, основанного на производных как первого так и более высоких порядков, подразумевает аппроксимацию целевой функции многочленом Тейлора, возникает вопрос о выпуклости полученной аппроксимации. До недавнего времени научное сообщество весьма пессимистично смотрело на данную проблему [Baes, 2009]. Однако Юрий Нестеров в своей работе [Nesterov, 2021a] показал, как можно правильным образом регуляризовать аппроксимацию Тейлора, чтобы сделать ее выпуклой. В этой же статье он предложил нижние оценки скорости сходимости тензорных методов для выпуклой оптимизации. Это породило огромный поток работ на эту тему, к примеру [Gasnikov et al., 2019; Bubeck et al., 2019; Jiang, Wang, Zhang, 2019; Dvurechensky et al., 2019; Ostroukhov et al., 2020]. В частности, в работах [Gasnikov et al., 2019; Bubeck et al., 2019; Jiang, Wang, Zhang, 2019] авторы практически одновременно предложили ускоренные тензорные методы со скоростью сходимости $\widetilde{O}\left(\varepsilon^{-2/(3p+1)}\right)$, где под тильдой в $\widetilde{O}$ обозначен мультипликативный логарифмический фактор. Данная оценка является почти оптимальной и совпадает с нижней оценкой из [Nesterov, 2021a], не считая логарифма. В следующей

своей работе [Nesterov, 2021b] Нестеров показал, как можно в предположении о липшицевости производной третьего порядка решать задачу методом второго порядка и предложил «супербыстрый тензорный метод». Полученный таким образом метод имеет сложность, которая оказывается лучше, чем существующие оценки для методов второго порядка. Связано это с тем, что раньше нижние оценки для алгоритмов p-го порядка предлагались исходя из предположения о липшицевости также p-й производной целевого функционала. Таким образом, предположение о липшицевости производной $(p + 1)$-го порядка нас выводит из рассматриваемого класса. Одним из продолжений этой работы послужила статья [Ahookhosh, Nesterov, 2021a]. В ней авторы показали, как можно с помощью ускоренных тензорных методов решать задачи с композитом типа

$$\min_{x \in \operatorname{dom} \psi} \{F(x) \equiv f(x) + \psi(x)\},$$

где $f : \mathbb{E} \to \mathbb{R}$ — выпуклая замкнутая и, возможно, недифференцируемая функция, $\psi : \mathbb{E} \to \mathbb{R}$ — простая выпуклая замкнутая функция, $\operatorname{dom} \psi \subseteq \operatorname{int}(\operatorname{dom} \psi)$, $\mathbb{E}$ — конечномерное вещественное векторное пространство. Это позволяет, к примеру, решать задачи с ограничениями на простых замкнутых множествах, если в качестве композита $\psi(x)$ использовать индикатор этого множества.

Очевидно, что скорость сходимости тензорных методов превышает скорость сходимости градиентного метода с точки зрения количества итераций. Но стоимость одной итерации возрастает ввиду необходимости использования производных высокого порядка. Обычно предполагается, что тензорные методы показывают свою эффективность при размерности задачи $n < 1000$. Таким образом, можно сделать вывод, что тензорные методы занимают некую нишу между методами типа эллипсоидов ($n \lesssim 100$) и градиентными методами ($n \gg 100$). В связи с этим кажется целесообразным исследовать применение методов высокого порядка к задаче (1).

Итак, в исследуемой нами постановке задачи (1) мы предполагаем $m \gg n$, $100 < n < 1000$. Дополнительно мы предполагаем, что $Q_x \subset R^m$ — компактное множество, для Q_y мы рассматриваем два случая: $Q_y = \mathbb{R}^n$ и $Q_y \subset \mathbb{R}^n$ — компакт. $F(x, y)$ является $L_{p,y}$-гладкой по y (см. определение 1), L_{xy}-гладкой по совокупности переменных (см. определение 2), μ_x-сильно выпуклой по x и μ_y-сильно выпуклой по y. Предлагается, подобно работе [Gladin, Alkousa, Gasnikov, 2021], разбить данную задачу на две подзадачи: внешнюю ($\min\limits_{x \in \mathbb{R}^m}$) и внутреннюю ($\min\limits_{y \in \mathbb{R}^n}$). Внутренняя задача неточно решается тензорным методом. Внешняя задача, с использованием неточного оракула, полученного из решения внутренней задачи, решается быстрым градиентным методом.

Структура работы выглядит следующим образом. В следующем параграфе приводятся некоторые используемые по ходу статьи обозначения и общие определения. Далее описываются используемые в нашей работе алгоритмы: супербыстрый тензорный метод, неточный проксимальный тензорный метод и быстрый градиентный метод на простых множествах. Наши основные результаты об объединении упомянутых алгоритмов в смешанный оракул приводятся в последующем параграфе. Тут же сравниваются теоретические сложности полученных методов и обычного быстрого градиентного метода. В конце вкратце еще раз описываются полученные нами результаты и обсуждаются возможные дальнейшие направления развития этой темы.

Обозначения и определения

В данном разделе мы опишем некоторые общеизвестные определения. Также мы введем используемые обозначения, которые понадобятся в дальнейшем.

Для некоторой функции f обозначим производную порядка p в точке $x \in \operatorname{dom} f$ по направлениям $h_i \in \mathbb{R}^n$, $i = 1, \ldots, p$, через $\nabla^p f(x)[h_1, \ldots, h_p]$, $p \geqslant 1$. Тогда норма p-й производной определяется как

$$\|\nabla^p f(x)\|_2 := \max_{h_1, \ldots, h_p \in \mathbb{R}^n} \{|\nabla^p f(x)[h_1, \ldots, h_p]| : \|h_i\| = 1,\ i = 1, \ldots, p\}$$

или

$$\|\nabla^p f(x)\|_2 := \max_{h\in\mathbb{R}^n}\{|\nabla^p f(x)[h]^p : \|h\|_2 \leqslant 1\}.$$

Обозначим аппроксимацию Тейлора некоторой функции f в точке $\widehat{x} \in \operatorname{dom} f$ вплоть до p-го порядка ($p \geqslant 1$) через

$$f(x) = \Omega^f_{\widehat{x},p}(x) + o(\|x - \widehat{x}\|_2^p) \quad \forall x \in \operatorname{dom} f,$$

$$\Omega^f_{\widehat{x},p}(x) := \sum_{i=0}^{p} \frac{1}{i!}\nabla^i f(\widehat{x})[x - \widehat{x}]^p.$$

Также нам понадобится обозначение регуляризованной аппроксимации Тейлора порядка $p \geqslant 1$:

$$\widehat{\Omega}^f_{\widehat{x},p,H}(x) := \Omega^f_{\widehat{x},p}(x) + \frac{H}{(p+1)!}\|x - \widehat{x}\|_2^{p+1} \quad \forall x \in \operatorname{dom} f.$$

Для простоты, если из контекста будет понятно, верхний индекс над Ω мы будем опускать: $\Omega^f_{\widehat{x},p}(x) \equiv \Omega_{\widehat{x},p}(x)$, $\widehat{\Omega}^f_{\widehat{x},p}(x) \equiv \widehat{\Omega}_{\widehat{x},p}(x)$. Заметим, что $\widehat{\Omega}^f_{\widehat{x},p,H}(x)$ является выпуклой функцией, в случае если f является выпуклой, $L_{p,x}$-гладкой (см. определение 1) и $H \geqslant pL_{p,x}$ [Nesterov, 2021a].

Определение 1. Пусть $f(x)$ — некоторая $p \geqslant 1$ раз дифференцируемая функция. Тогда f удовлетворяет условию Липшица $p \geqslant 1$ порядка (является $L_{p,x}$-гладкой), если

$$\forall x,\ x' \in Q_x \Rightarrow \|\nabla^p f(x) - \nabla^p f(x')\|_2 \leqslant L_{p,x}\|x - x'\|_2. \tag{2}$$

В дальнейшем, если будет понятно из контекста, вместо $L_{p,x}$ будем писать L_p.

Определение 2. Пусть $F(x, y)$ — некоторая дифференцируемая по обеим переменным функция. Тогда $F(x, y)$ удовлетворяет условию Липшица первого порядка по совокупности переменных (является L_{xy}-гладкой по совокупности переменных), если

$$\forall x,\ x' \in Q_x,\ y,\ y' \in Q_y \Rightarrow \|\nabla F(x, y) - \nabla F(x', y')\|_2 \leqslant L_{xy}\|(x, y) - (x', y')\|_2. \tag{3}$$

Также при описании супербыстрого тензорного метода нам понадобится определение дивергенции Брегмана и прокс-функции.

Определение 3. Дивергенция Брегмана для функции $f(x)$ определяется следующим образом:

$$\beta_f(x, y) := f(y) - f(x) - \langle \nabla f(x);\ y - x\rangle \quad \forall x,\ y \in \operatorname{dom} f.$$

По сути, дивергенция Брегмана показывает разницу между значением функции в точке y и значением ее линейной аппроксимации в точке y относительно точки x.

Определение 4. Прокс-функцией $d_p(x)$ порядка p для некоторого $x \in Q_x$ называется некоторая p раз непрерывно дифференцируемая, 1-сильно выпуклая функция. В нашей работе мы выбираем

$$d_p(x) := \frac{1}{p}\|x\|_2^p.$$

Также нам понадобятся определения гладкости и сильной выпуклости относительно некоторой функции.

Определение 5. Пусть $f(x)$ — некоторая выпуклая функция. Тогда $f(\cdot)$ является L_h-гладкой относительно некоторой функции $\rho(\cdot)$, если существует такая константа $L_h > 0$, что $(L_h\rho - h)(\cdot)$ выпукла.

Определение 6. Пусть $f(x)$ — некоторая выпуклая функция. Тогда $f(\cdot)$ является μ_h-сильно выпуклой относительно некоторой функции $\rho(\cdot)$, если существует такая константа $\mu_h > 0$, что $(h - \mu_h\rho)(\cdot)$ выпукла.

Используемые алгоритмы

В этом разделе мы вкратце описываем используемые методы, опуская многие детали, важные для понимания, и отсылая читателя к указанным первоисточникам.

Супербыстрый тензорный метод

Рассмотрим задачу

$$\min_{y\in\mathbb{R}^n} f(y), \tag{4}$$

где f — L_3-гладкая, μ-сильно выпуклая функция.

Для начала стоит еще раз акцентировать внимание на том, почему супербыстрый тензорный метод носит такое название. Этот метод использует предположение об L_3-гладкости, хотя по факту решает задачу, используя оракул второго порядка. Это стало возможным благодаря замене третьей производной по направлению на ее разностную аппроксимацию производными первого порядка (см. лемму 4.2 в [Nesterov, 2021b]). Рассматривая таким образом поставленную задачу, алгоритм получает оценку сходимости лучше, чем нижние оценки для методов второго порядка в традиционной постановке, когда предполагается только L_2-гладкость.

В данной работе мы будем использовать ускоренный вариант супербыстрого метода, названного ATMI_3 (Inexact Accelerated 3rd-Order Tensor Method). Перед тем как описывать его псевдокод, опишем подзадачу, которую этот метод решает.

Внутри рассматриваемого метода на каждом шаге необходимо найти неточный минимум функции $\widehat{\Omega}_{y,p,H}(\cdot)$ в следующем смысле. Осуществляется поиск точек во вложенных окрестностях:

$$\mathcal{N}^{\gamma}_{p,H}(y) = \{T \in \mathbb{R}^n \colon \|\nabla\widehat{\Omega}_{y,p,H}(T)\|_2 \leqslant \gamma\|\nabla f(T)\|_2\}, \tag{5}$$

где $\gamma \in [0, 1)$ — некоторый параметр точности. Зафиксируем $\gamma = \frac{1}{2p}$, $H = 2pL_p$ и для простоты обозначим

$$\mathcal{N}_p(y) \equiv \mathcal{N}^{1/(2p)}_{p,2pL_p}(y).$$

При этом $p = 3$.

Для решения описанной подзадачи будет использоваться дополнительный алгоритм, который носит название Bregman Distance Gradient Method (BDGM). В [Grapiglia, Nesterov, 2021] было доказано, что BDGM решает данную проблему за линейное время, а в [Nesterov, 2021b] было показано, как можно улучшить данный алгоритм, чтобы он работал с неточными градиентами. Данная модификация позволяет избежать вычисления $\nabla^3 f(\widehat{y}_k)[y-\widehat{y}_k]^2$, $\forall\widehat{y}_k$, $y \in \mathbb{R}^n$, заменив ее разностными аналогами с использованием градиентов:

$$g^{\tau}_{\widehat{y}_k}(y) := \frac{1}{\tau^2}(\nabla f(\widehat{y}_k + \tau(y-\widehat{y}_k)) + \nabla f(\widehat{y}_k - \tau(y-\widehat{y}_k)) - 2\nabla f(\widehat{y}_k)). \tag{6}$$

Введем дополнительное обозначение: $\phi_k(y) \equiv \nabla\widehat{\Omega}_{\widehat{y}_k,3,6L_p}(y)$. Тогда аппроксимацию $\phi(y)$ можно переписать в виде

$$g_{\phi_k,\tau}(y) := \nabla f(\widehat{y}_k) + \nabla^2 f(\widehat{y}_k)[y-\widehat{y}_k] + \frac{1}{2}g^{\tau}_{\widehat{y}_k}(y) + L_3\|y-\widehat{y}_k\|_2^2(y-\widehat{y}_k). \tag{7}$$

Псевдокод BDGM можно увидеть в алгоритме 1. Авторы в своей работе показывают, что при выборе погрешности при решении задачи (5) в виде

$$\delta = O\left(\frac{\varepsilon^{3/2}}{\frac{\|\nabla f(\widehat{y}_k)\|_2^{1/2} + \|\nabla^2 f(\widehat{y}_k)\|^{3/2}}{L_3^{1/2}}}\right)$$

Algorithm 1. Bregman distance gradient method (BDGM) [Nesterov, 2021b]

Вход: $\delta > 0, \widehat{y}_k \in \mathbb{R}^n$.

$z_0 = \widehat{y}_k$.

$\tau = \frac{3\delta}{8\left(2+\sqrt{2}\right)\|\nabla f(\widehat{y}_k)\|_2}$.

Определим множество

$$S_k = \left\{ z \colon \|z - \widehat{y}_k\| \leqslant 2\left(\frac{2+\sqrt{2}}{L_3}\|\nabla f(\widehat{y}_k)\|_2\right)^{1/3} \right\}.$$

Определим функцию

$$\rho_k(z) = \frac{1}{2}\left\langle \nabla^2 f(\widehat{y}_k)^T (z - \widehat{y}_k);\, z - \widehat{y}_k \right\rangle + L_3 d_4(z - \widehat{y}_k).$$

$k = 0$.

while $\|g_{\phi_k,\tau}(z_k)\| > \frac{1}{6}\|\nabla f(z_k)\|_2 - \delta$ **do**

  Вычислить $g_{\phi_k,\tau}(z_k)$ через (7).

$$z_{k+1} = \arg\min_{z \in S_k} \left\{ \langle g_{\phi_k,\tau}(z_k);\, z - z_k \rangle + 2\left(1 + \frac{1}{\sqrt{2}}\right)\beta_{\rho_k}(z_k,\, z) \right\}.$$

  $k = k + 1$.

return z_k.

алгоритму 1 нужно будет сделать

$$T_k(\delta) = O\left(\ln \frac{G+H}{\varepsilon}\right)$$

итераций, где G и H — равномерные верхние границы норм градиентов и гессианов, вычисленных в точках, сгенерированных основным алгоритмом. Следовательно, алгоритму 1 нужно будет один раз вычислить $\nabla^2 f(\cdot)$ и $T_k(\delta)$ раз вычислить $\nabla f(\cdot)$.

Algorithm 2. Inexact accelerated 3rd-order tensor method

Вход: $y_0 \in \mathbb{R}^n$, $N \in \mathbb{N}$

$c_3 = \left(\frac{5}{7L_3}\right)^{1/3}$

$\psi_0(y) = d_4(y - y_0)$

for $i = 0, \ldots, N-1$ **do**

  $v_i = \arg\min_{y \in \mathbb{R}^n} \psi_i(y)$.

$$A_i = 2\left(\frac{2}{3}c_3\right)^3 \left(\frac{i}{4}\right)^4, \quad a_{i+1} = A_{i+1} - A_i.$$

$$z_i = \frac{A_i}{A_{i+1}} y_i + \frac{a_i}{A_{i+1}} v_i.$$

  Вычислить $y_{i+1} = \mathcal{N}_3(z_i)$ с помощью алгоритма 1.

  $\psi_{i+1}(y) = \psi_i(y) + a_{i+1}\left(f(y_{i+1}) + \left\langle \nabla f(y_{i+1});\, y - y_{i+1} \right\rangle\right)$.

return y_N.

Обычный (неускоренный) супербыстрый метод, по сути, представляет из себя алгоритм 1, запущенный некоторое предопределенное количество раз. При помощи стандартной техники оценивающих последовательностей этот метод можно ускорить и получить алгоритм 2. Скорость сходимости данного метода приводится в следующей теореме.

Теорема 1 (с. 26 в [Nesterov, 2021b]). *Пусть некоторая функция $f\colon \mathbb{R}^n \to \mathbb{R}$ является выпуклой и L_3-гладкой. Тогда для алгоритма 2 выполняется*

$$f(y_N) - f(y^*) \leqslant \frac{7}{60}\left(\frac{6}{N}\right)^4 \cdot L_3 R^4, \tag{8}$$

где $R = \|y_0 - y^*\|_2$. Соответственно, для нахождения $y_\varepsilon \in \operatorname{dom} f\colon f(y_\varepsilon) - f(y^*) \leqslant \varepsilon$ алгоритму нужно

$$K = O\left(R\left(\frac{L_3}{\varepsilon}\right)^{1/4}\right) \tag{9}$$

вычислений гессианов и

$$O\left(R\left(\frac{L_3}{\varepsilon}\right)^{1/4} \log\frac{G+H}{\varepsilon_g}\right) \tag{10}$$

вычислений градиентов, где ε_g — нижняя граница для норм всех градиентов, посчитанных во время решения подзадачи.

Так как в этой работе мы предполагаем μ_y-сильную выпуклость внутренней задачи, воспользуемся стандартной процедурой рестартов, чтобы получить еще более ускоренный вариант алгоритма.

Algorithm 3. Restarted inexact accelerated 3rd-order tensor method

Вход: $\varepsilon > 0$, $y_0 \in \mathbb{R}^n$, $R \geqslant \|y_0 - y^*\|_2$

1: **for** $i = 0, \ldots, \left\lceil \log\frac{\mu R^2}{\varepsilon} \right\rceil - 1$ **do**

2: $R_i = \frac{R}{2^i}$.

3:
$$N_i = 6\left\lceil \sqrt[4]{\frac{7L_3R_i^2}{15\mu}} \right\rceil.$$

4: Запустить алгоритм 2 в течение N_i итераций, на выходе получить y_{N_i}.

5: **return** y_{N_i}

Теорема 2. *Пусть решается задача* (4)*. Тогда для нахождения $y_\varepsilon \in \operatorname{dom} f\colon f(y_\varepsilon) - f(y^*) \leqslant \varepsilon$ алгоритму* 3 *нужно*

$$O\left(\left(\frac{L_3R^2}{\mu}\right)^{1/4} \log\frac{\mu R^2}{\varepsilon}\right) \tag{11}$$

вычислений гессианов и

$$O\left(\left(\frac{L_3R^2}{\mu}\right)^{1/4} \log\frac{\mu R^2}{\varepsilon} \log\frac{G+H}{\varepsilon_g}\right) \tag{12}$$

вычислений градиентов, где $R = \|y_0 - y^*\|_2$, а ε_g — нижняя граница для норм всех градиентов, посчитанных во время решения подзадачи.

Доказательство. Из (8) и сильной выпуклости получаем

$$\frac{\mu}{2}\|y_N - y^*\|_2^2 \leqslant f(y_N) - f^* \leqslant \frac{7}{60}\left(\frac{6}{N}\right)^4 \cdot L_3 R^4.$$

Выберем N_1: $\|y_{N_1} - y^*\|_2^2 \leqslant \frac{1}{2}\|y_0 - y^*\|_2^2 = \frac{1}{2}R^2$. Тогда

$$N_1 = 6\left\lceil \sqrt[4]{\frac{7L_3R^2}{15\mu}} \right\rceil.$$

Далее, аналогично будем выбирать N_i, каждый раз уменьшая квадрат расстояния вдвое.

Обозначим $R_k = \frac{R}{2^{k-1}}$. Так как мы хотим получить $f(y_{N_k}) - f^* \leqslant \varepsilon$, то оценим количество необходимых рестартов:

$$f(y_{N_k}) - f^* \leqslant \frac{7}{60}\left(\frac{6}{N_k}\right)^4 \cdot L_3 R_k^4 \leqslant \frac{\mu R^2}{2^k} = \varepsilon.$$

В итоге получаем

$$k = \left\lceil \log \frac{\mu R^2}{\varepsilon} \right\rceil.$$

Тогда общее число итераций

$$N = \sum_{i=1}^{k} N_i \leqslant kN_1 = O\left(\left(\frac{L_3R^2}{\mu}\right)^{1/4} \log \frac{\mu R^2}{\varepsilon}\right).$$

□

Неточный проксимальный тензорный метод

Предположим, что мы решаем задачу с ограничениями на некотором простом замкнутом множестве:

$$\min_{y \in Q_y} f(y), \tag{13}$$

где f — L_p-гладкая, выпуклая функция. Эту задачу также можно переписать в виде

$$\min_{y \in \operatorname{dom}\psi} \{\Phi(y) := f(y) + \psi(y)\}, \tag{14}$$

где $\psi(y)$ — индикаторная функция Q_y:

$$\psi(y) = \begin{cases} 0, & y \in Q_y, \\ +\infty, & y \notin Q_y. \end{cases}$$

Введем следующее определение.

Определение 7. Композитным проксимальным оператором p-го порядка для некоторой функции Ф из (14) называется следующая функция:

$$\operatorname{prox}^p_{\Phi/H}(\widetilde{y}) = \arg\min_{y \in E}\left\{\Phi(y) + \frac{H}{p+1}\|y - \widetilde{y}\|_2^{p+1}\right\}. \tag{15}$$

В [Ahookhosh, Nesterov, 2021a] авторы исследуют проксимальные методы высокого порядка через аппроксимацию оператора (15) и его неточное решение на каждом шаге в виде подзадачи. Множество возможных решений (15) описывается через

$$\mathcal{A}_H^p(\widetilde{y},\,\gamma) = \left\{(y,\,g) \in \operatorname{dom}\psi \times \mathbb{R}^n \colon g \in \partial\psi(x),\ \|\nabla f_{\widetilde{y},H}^p(y) + g\|_2 \leqslant \gamma\|\nabla f(y) + g\|_2\right\}, \tag{16}$$

где

$$f_{\widetilde{y},H}^p(y) = f(y) + Hd_{p+1}(y - \widetilde{y}), \tag{17}$$

где $\gamma \in [0,\,1)$ — параметр точности. Приведем вспомогательную лемму, описывающую свойства решений, входящих в (16).

Лемма 1 (лемма 2.2 в [Ahookhosh, Nesterov, 2021a]). *Пусть* $(T,\,g) \in \mathcal{A}_H^p(\widehat{x},\,\gamma)$ *для некоторого* $g \in \partial\psi(T)$. *Тогда выполняется*

$$(1-\gamma)\|\nabla f(T) + g\|_2 \leqslant H\|T - \widehat{x}\|_2 \leqslant (1+\gamma)\|\nabla f(T) + g\|_2, \tag{18}$$

$$\langle \nabla f(T) + g;\, \widehat{x} - T\rangle \geqslant \frac{H}{1-\gamma}\|T - \widehat{x}\|_2^{p+1}. \tag{19}$$

Если дополнительно известно, что $\gamma \leqslant \frac{1}{p}$, *то*

$$\langle \nabla f(T) + g;\, \widehat{x} - T\rangle \geqslant \left(\frac{1-\gamma}{H}\right)^{1/p} \|\nabla f(T) + g\|_2^{(p+1)/p}. \tag{20}$$

Определим функцию

$$\rho_{\widehat{y}_k,H}(z) = \sum_{k=1}^{\lfloor p/2\rfloor} \frac{1}{(2k)!} D^{2k} f(\widehat{y}_k)[z - \widehat{y}_k]^{2k} + Hd_{p+1}(z - \widehat{y}_k). \tag{21}$$

Тогда для нахождения решения из (16) авторы предлагают алгоритм 4, где для простоты используется обозначение $\rho \equiv \rho_{\widehat{y}_k,H}$.

Algorithm 4. Non-euclidean composite gradient algorithm [Ahookhosh, Nesterov, 2021a]

Вход: $\widehat{y}_k \in \operatorname{dom}\psi$, $\gamma \in \left[0,\,\frac{1}{p}\right]$, $L > 0$.

1: $z_0 = \widehat{y}_k$.
2: $i = 0$.
3: **while** $z_i \notin \mathcal{A}_H^p(\widehat{y}_k,\,\gamma)$ **do**
4: Вычислить

$$z_{i+1} = \arg\min_{z \in \operatorname{dom}\psi} \left\{\left\langle \nabla f_{y_k,H}^p(z_i);\, z - z_i\right\rangle + \psi(z) + 2L\beta_\rho(z_i,\,z)\right\}.$$

5: $g = L(\nabla\rho(z_i) - \nabla\rho(z_{i+1})) - \nabla f_{y_k,H}^p(z_i)$, $g \in \partial\psi(z_{i+1})$.
6: $i = i + 1$.
7: **return** $(z_i,\,g)$.

Согласно теореме 3.4 из [Ahookhosh, Nesterov, 2021a], данный алгоритм сходится линейно относительно точности решения основной задачи.

Далее, для решения исходной задачи (14) авторы оборачивают алгоритм 4 в итерационный процесс и ускоряют с помощью техники оценивающих последовательностей. Псевдокод полученного метода приведен в алгоритме 5. Его сложность описана в следующей теореме.

Algorithm 5. Bi-level high-order algorithm

Вход: $y_0 \in \operatorname{dom}\psi$, $\gamma \in \left[0, \frac{1}{p}\right]$

$H = \frac{6}{(p-1)!}L_p$, $A_0 = 0$, $c_p = \left(\frac{1-\gamma}{H}\right)^{1/p}$.
$\Psi_0(y) = d_{p+1}(y - y_0)$
$k = 0$
while $\Phi(y_k) - \Phi(y^*) > \varepsilon$ **do**
  $v_k = \arg\min\limits_{x \in \operatorname{dom}\psi} \Psi_k(x)$

$$A_k = \left(\frac{c_p}{2}\right)^p \left(\frac{k}{p+1}\right)^{p+1}, \quad a_{k+1} = A_{k+1} - A_k.$$

$$z_k = \frac{A_k}{A_{k+1}} y_k + \frac{a_{k+1}}{A_{k+1}} v_k.$$

  Вычислить $(T_k, g) \in \mathcal{A}_H^p(z_k, \gamma)$ с помощью алгоритма 4.
  Найти $y_{k+1} : \Phi(y_{k+1}) \leqslant \Phi(T_k)$.
  $\Psi_{k+1}(y) = \Psi_k(x) + a_{k+1}(\Phi(y_{k+1}) + \langle \nabla\Phi(y_{k+1}); y - y_{k+1}\rangle)$.
  $k = k + 1$.
return y_k.

Теорема 3 (теорема 3.11 в [Ahookhosh, Nesterov, 2021a]). *Пусть решается задача* (14)*,* $H \geqslant pL_p$*,* $p \geqslant 2$*,* $q = \left\lfloor \frac{p}{2} \right\rfloor$*,* $\gamma \in \left[0, \frac{1}{p}\right]$*,* ρ *определяется из* (21)*. Тогда сложность алгоритма* 5 *определяется как*

$$\widetilde{O}\left(\frac{L_p R^{p+1}}{\varepsilon}\right)^{1/(p+1)}, \tag{22}$$

где $R = \|y_0 - y^*\|_2$.

Дополнительно предположив сильную выпуклость $f(x)$, мы можем улучшить полученный результат с помощью техники рестартов аналогично теореме 2. Как итог, мы получим следующий результат.

Теорема 4. *Пусть решается задача* (14)*,* $H \geqslant pL_p$*,* $p \geqslant 2$*,* $q = \left\lfloor \frac{p}{2} \right\rfloor$*,* $\gamma \in \left[0, \frac{1}{p}\right]$*,* ρ *определяется из* (21)*. Предположим также, что* $f(x)$ *является* μ*-сильно выпуклой. Тогда, применив технику рестартов к алгоритму* 5*, получим метод со сложностью*

$$O\left(\left(\frac{L_p R^{p-1}}{\mu}\right)^{1/(p+1)} \log \frac{\mu R^2}{\varepsilon}\right). \tag{23}$$

Доказательство. Доказывается аналогично теореме 2. □

Быстрый градиентный метод на простом множестве

Рассмотрим следующую постановку задачи:

$$\min_{x \in Q_x} f(x), \tag{24}$$

где $Q_x \subset \mathbb{R}^m$ — некоторое простое множество, f — L-гладкая, μ-сильно выпуклая функция. Под простым множеством подразумевается такое множество, проекцию градиента на которое можно явно вычислить. Определим градиентное отображение на Q_x.

Определение 8 (определение 2.2.3 в [Nesterov, 2004]). Зафиксируем некоторое $\gamma > 0$. Обозначим

$$x_{Q_x}(\widehat{x}, \gamma) := \arg\min_{Q_x} f(\widehat{x}) + \langle \nabla f(\widehat{x}); x - \widehat{x} \rangle + \frac{\gamma}{2}\|x - \widehat{x}\|_2^2,$$
$$g_{Q_x}(\widehat{x}, \gamma) := \gamma(\widehat{x} - x_{Q_x}(\widehat{x}, \gamma).$$

Тогда $g_{Q_x}(\widehat{x}, \gamma)$ называется градиентным отображением f в точке $\widehat{x}$ на множество Q_x.

Псевдокод быстрого градиентного метода для задачи (24) приведен в алгоритме 6. Сложность данного метода приведена в следующей теореме.

Algorithm 6. Быстрый градиентный метод на простом множестве

Вход: $x_0 \in Q_x$, размер шага $\alpha \in (0, 1)$, $N \in \mathbb{N}$.

1: $y_0 = x_0$, $q = \frac{\mu}{L}$
2: **for** $i = 1, \ldots, N - 1$ **do**
3: $x_{i+1} = x_{Q_x}(y_k, L)$
4: Вычислить $\alpha_{i+1} \in (0, 1)$: $\alpha_{i+1}^2 = (1 - \alpha_{i+1})\alpha_i^2 + q\alpha_{i+1}$.
5: $\beta_i = \frac{\alpha_i(1-\alpha_i)}{\alpha_i^2+\alpha_{i+1}}$
6: $y_{i+1} = x_{i+1} + \beta_k(x_{i+1} - x_i)$.
7: **return** x_N

Теорема 5 (с. 119 в [Nesterov, 2004]). *Пусть решается задача* (24). *Тогда для нахождения* $x_\varepsilon \in Q_x$: $f(x_\varepsilon) - f(x^*)$ *алгоритму* 6 *необходимо*

$$O\left(\frac{LR^2}{\varepsilon}\right)^{1/2} \tag{25}$$

итераций в выпуклом случае и

$$O\left(\left(\frac{L}{\mu}\right)^{1/2} \log \frac{LR^2}{\varepsilon}\right) \tag{26}$$

в сильно выпуклом случае, где $R = \|x_0 - x^*\|_2$.

Полученные результаты

Опишем еще раз рассматриваемую нами постановку задачи.

$$\min_{x \in Q_x} \min_{y \in Q_y} F(x, y), \tag{27}$$

где выполняются следующие предположения:

1) $Q_x \subset \mathbb{R}^m$ — выпуклое компактное множество, $m > 1000$;
2) $Q_y = \mathbb{R}^n$ или $Q_y \subset \mathbb{R}^n$ — компакт, $100 < n < 1000$;
3) $F(x, y)$ — μ_y-сильно выпуклая по y;
4) $F(x, y)$ — μ_x-сильно выпуклая по x;
5) $F(x, y)$ — $L_{p,y}$-гладкая по y;
6) $F(x, y)$ — L_{xy}-гладкая по совокупности переменных.

Как уже упоминалось ранее, для решения задачи (27) мы используем смешанный оракул. На первом этапе мы неточно решаем внутреннюю задачу: для некоторого фиксированного $x_k \in Q_x$ находим

$$\widetilde{y}_{k+1} = \widetilde{y}(x_k) \in Q_y \colon F(x_k, \widetilde{y}_{k+1}) - F(x_k, y^*_{k+1}) \leqslant \varepsilon, \tag{28}$$

где $y^*_{k+1} = y^*(x_{k+1}) = \min\limits_{y \in Q_y} F(x_k, y)$. Упомянутые предположения о функции позволяют нам использовать для решения супербыстрые методы для сильно выпуклых функций (см. теорему 2), если мы решаем внутреннюю задачу на всем пространстве. Если же мы решаем внутреннюю задачу на компакте, то мы можем использовать проксимальный тензорный метод для задачи с композитом (см. теорему 4). На следующем этапе, исходя из полученного решения внутренней задачи, мы осуществляем итерацию внешнего метода для решения внешней задачи и на выходе получаем x_{k+1}. Ввиду предположений о (27) здесь мы используем быстрый градиентный метод для сильно выпуклых функций на простых множествах. Далее мы вновь решаем внутреннюю задачу уже при фиксированном x_{k+1} и т. д.

Так как мы решаем внутреннюю задачу неточно, необходимо эту неточность учитывать при обращении к оракулу во время решения внешней задачи. Для этого мы используем концепцию (δ, L)-оракула.

Определение 9 (определение 1 в [Gasnikov et al., 2015]). (δ, L)-оракул выдает на запрос, в котором указана одна точка $x \in Q_x$, такие $(f_\delta(x), g_\delta(x))$, что

$$\forall x' \in Q_x \Rightarrow 0 \leqslant f(x') - f_\delta(x) - \langle g_\delta(x), x - x' \rangle \leqslant \frac{L}{2}\|x' - x\|_2^2 + \delta.$$

Прежде чем описать, как можно получить неточный оракул из неточного решения внутренней задачи, введем вспомогательное предложение.

Предложение 1 (утверждение 3 в [Gasnikov et al., 2015]). *Пусть* $F \colon Q_x \times Q_y \to \mathbb{R}$, $f(x) = \min\limits_{y \in Q_y} F(x, y)$, *где* $F(x, y)$ *— выпуклая,* L_{xy}*-гладкая по совокупности переменных функция, а* $Q_x \subseteq \mathbb{R}^m$, $Q_y \subseteq \mathbb{R}^n$ *— некоторые выпуклые множества. Пусть*

$$\forall x \in Q_x \ \exists y_\delta = y_\delta(x) \in Q_y \colon \max_{y' \in Q_y} \langle \nabla_y F(x, y_\delta); y_\delta - y' \rangle \leqslant \delta. \tag{29}$$

Тогда

$$\forall x' \in Q_x \Rightarrow \|\nabla f(x') - \nabla f(x)\|_2 \leqslant L_{xy}\|x' - x\|_2,$$

и $(F(x, y_\delta) - 2\delta, \nabla_x F(x, y_\delta))$ *будет* $(6\delta, 2L_{xy})$*-оракулом для* $f(x)$ *на* Q_x.

Как видно, для выполнения этого предложения нам необходимо выполнение условия (29). Так как мы рассматриваем внутреннюю задачу на двух разных доменах, оба случая нам нужно рассмотреть по отдельности.

Внутренняя задача на всем пространстве

Предложение 2. *Пусть* $f \colon \mathbb{R}^n \to \mathbb{R}$ *— некоторая* μ*-сильно выпуклая,* L*-гладкая функция. Пусть* $\exists x_\varepsilon \in \mathbb{R}^n \colon f(x_\varepsilon) - f(x^*) \leqslant \varepsilon$, *где* $x^* = \arg\min\limits_{x \in \mathbb{R}^n} f(x)$. *Тогда*

$$\forall x \in \mathbb{R}^n \Rightarrow \langle \nabla f(x_\varepsilon); x_\varepsilon - x \rangle \leqslant L\|x_\varepsilon - x\|_2 \sqrt{\frac{2\varepsilon}{\mu}}.$$

Доказательство. По неравенству Коши – Буняковского – Шварца

$$\langle \nabla f(x_\varepsilon); x_\varepsilon - x \rangle \leqslant \|\nabla f(x_\varepsilon)\|_2 \|x_\varepsilon - x\|_2. \tag{30}$$

Из условия гладкости (2)

$$\|\nabla f(x_\varepsilon)\|_2 = \|\nabla f(x_\varepsilon) - \nabla f(x^*)\|_2 \leqslant L\|x_\varepsilon - x^*\|_2.$$

Далее, из сильной выпуклости и определения x_ε имеем

$$\frac{\mu}{2}\|x_\varepsilon - x^*\|_2^2 \leqslant f(x_\varepsilon) - f(x^*) \leqslant \varepsilon \Leftrightarrow \|x_\varepsilon - x^*\|_2 \leqslant \sqrt{\frac{2\varepsilon}{\mu}}.$$

Таким образом,

$$\|\nabla f(x_\varepsilon)\|_2 \leqslant L\sqrt{\frac{2\varepsilon}{\mu}}.$$

Подставляя данное неравенство в (30), получаем искомый результат. □

Очевидно, что для выполнения предложения 1 в условиях нашей задачи необходимо оценить $\|x_\varepsilon - x\|_2$ в предложении 2. Мы сделаем это в следующей теореме, аналогичной теореме 4 в [Gladin et al., 2021].

Теорема 6. *Пусть* $F\colon Q_x \times \mathbb{R}^n \to \mathbb{R}$, $f(x) = \min\limits_{y\in\mathbb{R}^n} F(x, y)$. *При этом*

- $Q_x \subset \mathbb{R}^m$ *— некоторое компактное множество,*
- $F(x, y)$ *выпуклая по* x*,*
- $F(x, y)$ μ_y*-сильно выпуклая по* y*,*
- $F(x, y)$ L_y*-гладкая по* y*,*
- $F(x, y)$ *—* L_{xy}*-гладкая по совокупности переменных функция.*

Тогда $\forall x \in Q_x\ \exists y_\varepsilon = y_\varepsilon(x) \in \mathbb{R}^n$:

1)
$$F(x, y_\varepsilon) - f(x) \leqslant \varepsilon,$$

2) $\forall x' \in Q_x \Rightarrow$
$$\|\nabla f(x') - \nabla f(x)\|_2 \leqslant L_{xy}\|x' - x\|_2,$$

3) $(F(x, y_\varepsilon) - 2\delta,\ \nabla_x F(x, y_\varepsilon))$ *будет* $(6\delta,\ 2L_{xy})$*-оракулом для* $f(x)$ *на* Q_x*, где*

$$\delta = \frac{2L_y}{\mu_y}\left(\varepsilon + \sqrt{D\varepsilon}\right), \quad D = \max_{x\in Q_x}\left(F(x, y(x)) - F(x, y^*(x))\right).$$

Доказательство. Из предложения 2 мы знаем, что

$$\forall x \in Q_x,\ \forall y = y(x) \in \mathbb{R}^n\ \exists y_\varepsilon = y_\varepsilon(x) \in \mathbb{R}^n:$$
$$\langle \nabla_y F(x, y_\varepsilon);\ y_\varepsilon - y\rangle \leqslant L_y\|y_\varepsilon - y\|_2\sqrt{\frac{2\varepsilon}{\mu_y}}. \tag{31}$$

Из неравенства треугольника

$$\|y_\varepsilon - y\|_2 \leqslant \|y_\varepsilon - y^*\|_2 + \|y - y^*\|_2.$$

Из сильной выпуклости по y и определения y_ε имеем

$$\|y_\varepsilon - y^*\|_2 \leqslant \sqrt{\frac{2}{\mu_y}(F(x, y_\varepsilon) - F(x, y^*))} \leqslant \sqrt{\frac{2}{\mu}\varepsilon}.$$

Теперь оценим $\|y - y^*\|_2 = \|y(x) - y^*(x)\|_2$:

$$\|y(x) - y^*(x)\|_2 \leqslant \sqrt{\frac{2}{\mu_y}(F(x, y(x)) - F(x, y^*(x)))}.$$

Обозначим $\Delta(x) = F(x, y(x)) - F(x, y^*(x))$. Так как функция $\Delta(x)$ определена на компакте и непрерывна на нем, то она ограничена на нем:

$$\exists D \in \mathbb{R}\colon \Delta(x) \leqslant D.$$

Таким образом, получаем

$$\|y - y^*\|_2 = \|y(x) - y^*(x)\|_2 \leqslant \sqrt{\frac{2}{\mu}D}.$$

Итого,

$$\langle \nabla_y F(x, y_\varepsilon);\, y_\varepsilon - y\rangle \leqslant \left(\sqrt{\frac{2}{\mu_y}D} + \sqrt{\frac{2}{\mu_y}\varepsilon}\right) L_y \sqrt{\frac{2}{\mu_y}\varepsilon} = \frac{2L_y}{\mu_y}\left(\varepsilon + \sqrt{D\varepsilon}\right).$$

Осталось применить предложение 1, и мы получаем искомый результат. □

Внутренняя задача на компакте

Так как в алгоритме 5 мы можем выбирать $y_\varepsilon \equiv y_k\colon \Phi(y_\varepsilon) = \Phi(T_k)$ (строчка 9), просто будем выбирать $y_k = T_{k-1}$. Тогда $(y_\varepsilon, g) = (T_{k-1}, g) \in \mathcal{A}_H^p(z_{k-1}, \gamma)$.

Предложение 3. *Пусть решается задача* (14) *с помощью метода из теоремы* 4*, где в строчке* 9 *выбираем* $y_{k+1} = T_k$*. При этом* $f(y)$ *μ-сильно выпуклая,* $Q_y \subset \mathbb{R}^n$ *— некоторый компакт. Пусть* $\exists y_\varepsilon \in Q_y\colon f(y_\varepsilon) - f(y^*) \leqslant \varepsilon$*, где* $y^* = \arg\min\limits_{y\in Q_y} f(y)$*. Тогда*

$$\forall y \in Q_y \Rightarrow \langle \nabla\Phi(y_\varepsilon);\, y_\varepsilon - y\rangle \leqslant \frac{H}{1-\gamma}\left(\sqrt{\frac{2}{\mu}\varepsilon} + \sqrt{\frac{2}{\mu}D}\right)^{p+1}, \tag{32}$$

где $D := \max\limits_{y\in Q_y} f(y) - f(y^*)$.

Доказательство. Из неравенства Коши – Буняковского – Шварца

$$\langle \nabla\Phi(y_\varepsilon);\, y_\varepsilon - y\rangle \leqslant \|\nabla\Phi(y_\varepsilon)\|_2 \|y_\varepsilon - y\|_2. \tag{33}$$

В первую очередь оценим $\|\nabla\Phi(y_\varepsilon)\|_2$. По определению,

$$\|\nabla\Phi(y_\varepsilon)\|_2 = \|\nabla f(y_\varepsilon) + g\|_2, \quad g \in \partial\psi(y_\varepsilon).$$

Так как мы выбираем $y_k = T_{k-1}$, то $(y_\varepsilon, g) = (T_{k-1}, g)$, и согласно (18) имеем

$$\|\nabla\Phi(y_\varepsilon)\|_2 = \|\nabla f(y_\varepsilon) + g\|_2 \leqslant \frac{H}{1-\gamma}\|y_\varepsilon - z_{k-1}\|_2^p. \tag{34}$$

Оценим $\|y_\varepsilon - z_{k-1}\|_2$. Из неравенства треугольника

$$\|y_\varepsilon - z_{k-1}\|_2 \leqslant \|y_\varepsilon - y^*\|_2 + \|z_{k-1} - y^*\|_2.$$

Так как $f(y)$ μ-сильно выпукла, $\psi(y)$ выпукла, то $\Phi(y)$ μ-сильно выпукла, и для нее выполняется

$$\|y_\varepsilon - y^*\|_2 \leqslant \sqrt{\frac{2}{\mu}(\Phi(y_\varepsilon) - \Phi(y^*) - \langle \nabla\Phi(y^*);\, y_\varepsilon - y^*\rangle)}.$$

Из определения y_ε и

$$\forall y \in Q_y \Rightarrow \langle \nabla\Phi(y^*);\, y - y^*\rangle \geqslant 0 \tag{35}$$

получаем оценку

$$\|y_\varepsilon - y^*\|_2 \leqslant \sqrt{\frac{2}{\mu}\varepsilon}. \tag{36}$$

Далее, снова из сильной выпуклости получаем

$$\|z_{k-1} - y^*\|_2 \leqslant \sqrt{\frac{2}{\mu}\left(\Phi(z_{k-1}) - \Phi(y^*) - \langle \nabla\Phi(y^*);\, z_{k-1} - y^*\rangle\right)}.$$

Обозначим $\Delta(y) = \Phi(y) - \Phi(y^*) = f(y) - f(y^*)$ $\forall y \in Q_y$. Так как $\Delta(y)$ непрерывна на компакте Q_y, то она ограничена на нем, то есть $\exists D \in \mathbb{R}\colon D = \max\limits_{y \in Q_y} \Delta(y)$. Тогда

$$\forall y \in Q_y \Rightarrow \Delta(y) \leqslant D.$$

Из этого факта и (35) получаем

$$\|z_{k-1} - y^*\|_2 \leqslant \sqrt{\frac{2}{\mu}D}. \tag{37}$$

Таким образом, из (34), (36), (37) получаем

$$\|\nabla\Phi(y_\varepsilon)\|_2 \leqslant \frac{H}{1-\gamma}\left(\sqrt{\frac{2}{\mu}D} + \sqrt{\frac{2}{\mu}\varepsilon}\right)^p. \tag{38}$$

Чтобы оценить $\|y_\varepsilon - y\|_2$, также воспользуемся неравенством треугольника, (35), (36), (37) и получим

$$\|y_\varepsilon - y\|_2 \leqslant \sqrt{\frac{2}{\mu}D} + \sqrt{\frac{2}{\mu}\varepsilon}. \tag{39}$$

В итоге из (33), (38), (39) получаем

$$\langle \nabla\Phi(y_\varepsilon);\, y_\varepsilon - y\rangle \leqslant \frac{H}{1-\gamma}\left(\sqrt{\frac{2}{\mu}D} + \sqrt{\frac{2}{\mu}\varepsilon}\right)^{p+1}.$$

□

Как видно, в данном случае у нас сразу выполняется условие, необходимое нам в предложении 1. Так что мы можем привести следующую теорему без доказательства.

Теорема 7. *Пусть* $F\colon Q_x \times Q_y \to \mathbb{R}$, $f(x) = \min\limits_{y \in \mathbb{R}^n} F(x, y)$. *При этом*

- $Q_x \subset \mathbb{R}^m$ *и* $Q_y \subset \mathbb{R}^n$ *— некоторые компактные множества,*
- $F(x, y)$ *выпуклая по* x,
- $F(x, y)$ μ_y*-сильно выпуклая по* y,
- $F(x, y)$ L_y*-гладкая по* y,
- $F(x, y)$ $L_{p,y}$*-гладкая по* y,
- $F(x, y)$ L_{xy}*-гладкая по совокупности переменных.*

Если решать внутреннюю задачу с помощью метода из теоремы 4, *то* $\forall x \in Q_x$ $\exists y_\varepsilon = y_\varepsilon(x) \in Q_y$:

1)
$$F(x, y_\varepsilon) - f(x) \leqslant \varepsilon,$$

2) $\forall x' \in Q_x \Rightarrow$
$$\|\nabla f(x') - \nabla f(x)\|_2 \leqslant L_{xy}\|x' - x\|_2,$$

3) $(F(x, y_\varepsilon) - 2\delta, \nabla_x F(x, y_\varepsilon))$ *будет* $(6\delta, 2L_{xy})$*-оракулом для* $f(x)$ *на* Q_x, *где*

$$\delta = \frac{H}{1-\gamma}\left(\sqrt{\frac{2}{\mu_y}D} + \sqrt{\frac{2}{\mu_y}\varepsilon}\right)^{p+1}, \quad D = \max_{y \in Q_y}(F(x, y(x)) - F(x, y^*(x))),$$

где $H \geqslant pL_{p,y}$, $\gamma \in \left[0, \frac{1}{p}\right]$.

Внешняя задача

Итак, мы получили, что решение внутренней задачи с точностью ε дает нам возможность использовать $(6\delta, 2L_{xy})$-оракул для внешней задачи. В связи с этим можно для внешней задачи применять быстрый градиентный метод. Скорость сходимости быстрого градиентного метода для выпуклой функции в условиях (δ, L)-оракула описывается следующей теоремой.

Теорема 8 (теорема 4.9 в [Devolder, 2013]). *Пусть функция* f *выпукла и наделена* (δ, L)*-оракулом. Тогда последовательность* x_k, *которая генерируется быстрым градиентным методом с использованием этого оракула, удовлетворяет*

$$f(x_k) - f^* \leqslant \frac{2LR^2}{(k+1)^2} + \frac{1}{3}(k+3)\delta, \tag{40}$$

где $R = \|x_0 - x^*\|_2$.

Псевдокод быстрого градиентного метода с неточным оракулом совпадает с алгоритмом 6, где надо заменить выходы оракула первого порядка на их неточные аналоги. Чтобы учесть μ_x-сильную выпуклость $F(x, y)$, можно воспользоваться методом рестартов.

Теорема 9. *Пусть функция* $f\colon Q_x \to \mathbb{R}$, *где* $Q_x \subset \mathbb{R}^m$ — *выпуклое компактное множество,* μ-*сильно выпукла и наделена* (δ, L)-*оракулом. Предположим, что мы можем варьировать* δ. *Тогда из алгоритма* 6 *с неточным оракулом при* $q = 0$ *с помощью рестартов можно получить алгоритм со сложностью*

$$O\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{\mu R^2}{\varepsilon - \left(\sqrt{\frac{10L}{\mu}} + 3\right)\delta}\right)\right), \tag{41}$$

где $R = \|x_0 - x^*\|_2$, *а* δ *выбирается из условия*

$$\delta \leqslant \frac{\mu R^2}{10\left(\sqrt{\frac{10L}{\mu}} + 3\right)}.$$

Доказательство. Использовать технику рестартов в ее традиционном виде нам мешает аддитивный член в (25), линейно зависящий от δ. В первую очередь избавимся от него. Так как мы имеем возможность выбирать δ, выберем его таким, что

$$\frac{1}{\mu}(N+3)\delta \leqslant \frac{LR^2}{\mu N^2} \Leftrightarrow \delta \leqslant \frac{LR^2}{N^2(N+3)}. \tag{42}$$

Тогда мы можем получить следующую оценку из (40) и сильной выпуклости:

$$\frac{4LR^2}{\mu N^2} + \frac{1}{\mu}(N+3)\delta \leqslant \frac{5LR^2}{\mu N^2}.$$

Далее применяем традиционную технику рестартов аналогично доказательству теоремы 2. В итоге, получим $N_i = \sqrt{\frac{10L}{\mu}}$, $\forall i = 1, \ldots, k$

$$N = kN_1 = O\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{\mu R^2}{\varepsilon - \left(\sqrt{\frac{10L}{\mu}} + 3\right)\delta}\right)\right),$$

где δ выбирается исходя из (42). □

Можно заметить, что в (40) присутствует линейное накопление ошибки, возникающей из-за неточности δ. Однако в (41), помимо лучшей скорости сходимости, эта проблема отсутствует. Неточность содержится под логарифмом, в связи с чем пренебрежимо мала.

Теперь оценим сложность смешанного оракула, который мы описывали в начале параграфа, для решения задачи (27). Для начала рассмотрим случай, когда внутренняя задача определена на всем пространстве.

Теорема 10. *Пусть* $F\colon Q_x \times \mathbb{R}^n \to \mathbb{R}$, $f(x) = \min\limits_{y \in \mathbb{R}^n} F(x, y)$. *Обозначим через* $\widetilde{\varepsilon}$ *точность решения внутренней задачи* (28): $F(x, y_{\widetilde{\varepsilon}}) - f(x) \leqslant \widetilde{\varepsilon}$. *При этом*

- $Q_x \subset \mathbb{R}^m$ — *некоторое компактное множество,*
- $F(x, y)$ μ_x-*сильно выпуклая по* x,
- $F(x, y)$ μ_y-*сильно выпуклая по* y,
- $F(x, y)$ L_y-*гладкая по* y,

- $F(x, y)$ $L_{3,y}$-*гладкая по* y,
- $F(x, y)$ — L_{xy}-*гладкая по совокупности переменных функция,*
- *обозначим* $\delta(\widetilde{\varepsilon}) = \frac{2L_y}{\mu_y}\left(\widetilde{\varepsilon} + \sqrt{D\widetilde{\varepsilon}}\right)$, $D = \max\limits_{x\in Q_x}(F(x, y(x)) - F(x, y^*(x)))$ *и выберем* $\widetilde{\varepsilon}$ *таким образом, что*

$$\delta(\widetilde{\varepsilon}) \leqslant \frac{\mu_x\|x_0 - x^*\|_2^2}{10\left(\sqrt{\frac{10L_{xy}}{\mu_x}} + 3\right)}.$$

Будем решать внутреннюю задачу (28) *супербыстрым методом для сильно выпуклых функций* (*теорема* 2). *Внешнюю задачу будем решать быстрым градиентным методом с неточным оракулом для сильно выпуклых функций* (*теорема* 9). *Тогда сложность конечного алгоритма составит*

$$O\left(\left(\frac{L_{xy}}{\mu_x}\right)^{1/2}\left(\frac{L_{3,y}R_y^2}{\mu_y}\right)^{1/4}\log\frac{\mu_y R_y^2}{\widetilde{\varepsilon}}\log\frac{G+H}{\varepsilon_g}\log\left(\frac{\mu R_x^2}{\varepsilon - \left(\sqrt{\frac{10L}{\mu}} + 3\right)\delta}\right)\right) \tag{43}$$

или

$$\widetilde{O}\left(\left(\frac{L_{xy}}{\mu_x}\right)^{1/2}\left(\frac{L_{3,y}R_y^2}{\mu_y}\right)^{1/4}\right), \tag{44}$$

где $R_x = \|x_0 - x^*\|_2$, $R_y = \|y_0 - y^*\|_2$, *а под тильду в* $\widetilde{O}$ *мы занесли все логарифмические факторы.*

Доказательство. Следует непосредственно из теорем 2, 6, 9. □

И для внутренней задачи, определенной на компакте, мы получим следующее.

Теорема 11. *Пусть* $F\colon Q_x \times Q_y \to \mathbb{R}$, $f(x) = \min\limits_{y\in Q_y} F(x, y)$. *Обозначим через* $\widetilde{\varepsilon}$ *точность решения внутренней задачи* (28): $F(x, y_{\widetilde{\varepsilon}}) - f(x) \leqslant \widetilde{\varepsilon}$. *При этом*

- $Q_x \subset \mathbb{R}^m$ *и* $Q_y \subset \mathbb{R}^n$ — *некоторые компактные множества,*
- $F(x, y)$ μ_x-*сильно выпуклая по* x,
- $F(x, y)$ μ_y-*сильно выпуклая по* y,
- $F(x, y)$ L_y-*гладкая по* y,
- $F(x, y)$ $L_{p,y}$-*гладкая по* y,
- $F(x, y)$ — L_{xy}-*гладкая по совокупности переменных функция,*
- *обозначим* $\delta(\widetilde{\varepsilon}) = \frac{H}{1-\gamma}\left(\sqrt{\frac{2}{\mu_y}D} + \sqrt{\frac{2}{\mu_y}\widetilde{\varepsilon}}\right)^{p+1}$, $D = \max\limits_{y\in Q_y}(F(x, y(x)) - F(x, y^*(x)))$ *и выберем* $\widetilde{\varepsilon}$ *таким образом, что*

$$\delta(\widetilde{\varepsilon}) \leqslant \frac{\mu_x\|x_0 - x^*\|_2^2}{10\left(\sqrt{\frac{20L_{xy}}{\mu_x}} + 3\right)}.$$

Будем решать внутреннюю задачу (28) *проксимальным тензорным методом для сильно выпуклых функций* (*теорема* 4). *Внешнюю задачу будем решать быстрым градиентным методом с неточным оракулом для сильно выпуклых функций* (*теорема* 9). *Тогда сложность конечного алгоритма составит*

$$\widetilde{O}\left(\left(\frac{L_{xy}}{\mu_x}\right)^{1/2}\left(\frac{L_{p,y}R_y^{p-1}}{\mu_y}\right)^{1/(p+1)}\right), \tag{45}$$

где $R_x = \|x_0 - x^*\|_2$, $R_y = \|y_0 - y^*\|_2$, *а под тильду в* $\widetilde{O}$ *мы занесли все логарифмические факторы.*

Доказательство. Следует непосредственно из теорем 4, 7, 9. □

Замечание 1. Сравним полученный результат с традиционным подходом в выпуклой оптимизации — быстрым градиентным методом по совокупности переменных. Предположим, что $\mu_x \geqslant \mu_y$. Если решать всю задачу (27) быстрым градиентным методом, то сложность составит

$$\widetilde{O}\left(\sqrt{\frac{L_{xy}}{\min\{\mu_x, \mu_y\}}}\right) = \widetilde{O}\left(\sqrt{\frac{L_{xy}}{\mu_y}}\right). \tag{46}$$

Допустим, внутренняя переменная определена на всем пространстве. Сравним данную оценку с (44) с точки зрения коэффициентов сильной выпуклости:

$$\frac{1}{\mu_y^{1/2}} \vee \frac{1}{\mu_x^{1/2}\mu_y^{1/4}} \Leftrightarrow \frac{1}{\mu_y^{1/4}} \vee \frac{1}{\mu_x^{1/2}}.$$

Таким образом, предлагаемый нами в теореме 10 метод выигрывает по сложности у быстрого градиентного метода, если $\mu_x \geqslant \sqrt{\mu_y}$.

Замечание 2. Теперь, допустим, внутренняя переменная определена на компакте. Сравним оценку быстрого градиентного метода с (45):

$$\frac{1}{\mu_y^{1/2}} \vee \frac{1}{\mu_x^{1/2}\mu_y^{1/(p+1)}} \Leftrightarrow \frac{1}{\mu_y^{(p-1)/(2(p+1))}} \vee \frac{1}{\mu_x^{1/2}}.$$

Таким образом, предлагаемый нами в теореме 11 метод выигрывает по сложности у быстрого градиентного метода, если $\mu_x \geqslant \mu_y^{(p-1)/(p+1)}$.

К примеру, для $p = 3$ получаем $\mu_x \geqslant \sqrt{\mu_y}$, так же как и в замечании 1. Однако здесь для решения внутренней подзадачи нам нужно будет вычислять производные третьего порядка в отличие от метода из замечания 1.

Заключение

В данной работе мы предложили методы для решения задач типа min-min для сильно выпуклых по обеим переменным функций. Мы разделили задачу на внутреннюю и внешнюю подзадачи. В случае если внутренняя задача определена на всем пространстве, мы воспользовались супербыстрым тензорным методом Нестерова [Nesterov, 2021b], если же она определена на компакте — то проксимальным тензорным методом [Ahookhosh, Nesterov, 2021a]. Внешняя задача решалась быстрым градиентным методом с неточным оракулом на компакте. Сложность полученного метода позволяет говорить о его преимуществах по сравнению с обычным быстрым градиентным методом в некотором специальном сеттинге (см. замечания 1 и 2).

В конце хотелось бы отметить некоторые моменты, которые не удалось проработать в рамках данной статьи. Во-первых, вместо (δ, L)-оракула можно попробовать рассмотреть (δ, L, μ)-оракул. Тогда не было бы необходимости в применении рестартов к быстрому градиентному методу. Ведь, как известно, рестарты на практике дают сильно завышенную оценку количества итераций. Также при применении (δ, L, μ)-оракула не возникает накопление ошибки (см.

теорему 5.14 в [Devolder, 2013]), которое мы наблюдали в (40). Во-вторых, также имеет смысл использовать методы нулевого порядка для решения внешней задачи вместо градиентных методов. В-третьих, в случае, если внутренняя задача определена на компакте, можно для ее решения использовать супербыстрый тензорный метод (см. [Ahookhosh, Nesterov, 2021b]). В конце концов можно рассмотреть различные обобщения постановки: равномерную выпуклость, условия Гёльдера вместо условий Липшица. Если рассматривать возможные направления исследований задачи min-min и абстрагироваться от конкретных методов, использованных внутри смешанного оракула, можно обратить внимание на отсутствие нижних оценок для данной задачи.

Список литературы (References)

Ahookhosh M., Nesterov Yu. High-order methods beyond the classical complexity bounds, I: inexact high-order proximal-point methods // arXiv preprint. — 2021a. — https://arxiv.org/pdf/2107.05958

Ahookhosh M., Nesterov Yu. High-order methods beyond the classical complexity bounds, II: inexact high-order proximal-point methods with segment search // arXiv preprint. — 2021b. — https://arxiv.org/pdf/2109.12303

Baes M. Estimate sequence methods:extensions and approximations. — 2009. — http://www.optimization-online.org/DB_FILE/2009/08/2372.pdf

Bolte J., Glaudin L., Pauwels E., Serrurier M. A Hölderian backtracking method for min-max and min-min problems // arXiv preprint. — 2020. — https://arxiv.org/pdf/2007.08810

Bubeck S., Jiang Q., Lee Y. T., Li Y., Sidford A. Near-optimal method for highly smooth convex optimization // Conference on Learning Theory. — PMLR, 2019. — P. 492–507.

Devolder O. Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization // PhD thesis. — 2013.

Dvurechensky P., Gasnikov A., Ostroukhov P., Uribe C. A., Ivanova A. Near-optimal tensor methods for minimizing the gradient norm of convex function // arXiv preprint. — 2019. — https://arxiv.org/pdf/1912.03381

Gasnikov A., Dvurechensky P., Gorbunov E., Vorontsova E., Selikhanovych D., Uribe C. A. Optimal tensor methods in smooth convex and uniformly convex optimization // Conference on Learning Theory. — PMLR, 2019. — P. 1374–1391.

Gasnikov A., Dvurechensky P., Kamzolov D., Nesterov Yu., Spokoiny V., Stetsyuk P., Suvorikova A., Chernov A. Universal method with inexact oracle and its applications for searching equillibriums in multistage transport problems // arXiv preprint. — 2015. — https://arxiv.org/pdf/1506.00292

Gasnikov A., Gasnikova E. Traffic assignment models. Numerical aspects // arXiv preprint. — 2020. — https://arxiv.org/pdf/2003.12160

Gladin E., Alkousa M., Gasnikov A. On solving convex min-min problems with smoothness and strong convexity in one variable group and small dimension of the other // arXiv preprint. — 2021. — https://arxiv.org/pdf/2102.00584

Gladin E., Sadiev A., Gasnikov A., Dvurechensky P., Beznosikov A., Alkousa M. Solving smooth min-min and min-max problems by mixed oracle algorithms // arXiv preprint. — 2021. — https://arxiv.org/pdf/2103.00434

Grapiglia G. N., Nesterov Yu. On inexact solution of auxiliary problems in tensor methods for convex optimization // Optimization Methods and Software. — 2021. — Vol. 36, No. 1. — P. 145–170.

Jiang B., Wang H., Zhang S. Near-optimal method for highly smooth convex optimization // Conference on Learning Theory. — PMLR, 2019. — P. 1799–1801.

Jungers M., Trélat E., Abou-Kandil H. Min-max and min-min Stackelberg strategies with closed-loop information structure // Journal of dynamical and control systems. — 2011. — Vol. 17, No. 3. — P. 387.

Konur D., Farhangi H. Set-based min-max and min-min robustness for multiobjective robust optimization // Proceedings of the 2017 Industrial and Systems Engineering Research Conference. — 2017. — P. 1–6.

Nemirovsky A. S., Yudin D. B. Problem Complexity and Method Efficiency in Optimization. — New York: J. Wiley & Sons, 1983.

Nesterov Yu. Implementable tensor methods in unconstrained convex optimization // Mathematical Programming. — 2021a. — Vol. 186, No. 1. — P. 157–183.

Nesterov Yu. Introductory Lectures on Convex Optimization: a basic course. — Massachusetts: Kluwer Academic Publishers, 2004.

Nesterov Yu. Superfast second-order methods for unconstrained convex optimization // Journal of Optimization Theory and Applications. — 2021b. — Vol. 191, No. 1. — P. 1–30.

Ostroukhov P., Kamalov R., Dvurechensky P., Gasnikov A. Tensor methods for strongly convex strongly concave saddle point problems and strongly monotone variational inequalities // arXiv preprint. — 2020. — https://arxiv.org/pdf/2012.15595