КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2022 Т. 14 № 2 С. 497–515
DOI: 10.20537/2076-7633-2022-14-2-497-515

Ки&M

МАТЕМАТИЧЕСКИЕ ОСНОВЫ И ЧИСЛЕННЫЕ МЕТОДЫ МОДЕЛИРОВАНИЯ

УДК: 519.8

Об адаптивных ускоренных методах и их модификациях для альтернированной минимизации

Н. К. Тупица[1,2,3]

[1]Московский физико-технический институт,
Россия, 141701, Московская обл., г. Долгопрудный, Институтский пер., 9
[2]Институт проблем передачи и обработки информации,
Россия, 127051, г. Москва, Большой Каретный пер. 19, стр. 1
[3]Национальный исследовательский университет «Высшая школа экономики»,
Россия, 101000, г. Москва, ул. Мясницкая, д. 20

Получено 15.03.2020, после доработки — 12.12.2021.
Принято к публикации 13.02.2022.

В первой части работы получена оценка скорости сходимости ранее известного ускоренного метода первого порядка AGMsDR на классе задач минимизации, вообще говоря, невыпуклых функций с M-липшицевым градиентом и удовлетворяющих условию Поляка – Лоясиевича. При реализации метода не требуется знать параметр $\mu^{PL} > 0$ из условия Поляка – Лоясиевича, при этом метод демонстрирует линейную скорость сходимости (сходимость со скоростью геометрической прогрессии со знаменателем $\left(1 - \frac{\mu^{PL}}{M}\right)$). Ранее для метода была доказана сходимость со скоростью $O\left(\frac{1}{k^2}\right)$ на классе выпуклых задач с M-липшицевым градиентом. А также сходимость со скоростью геометрической прогрессии, знаменатель которой $\left(1 - \sqrt{\frac{\mu^{SC}}{M}}\right)$, но только если алгоритму известно значение параметра сильной выпуклости $\mu^{SC} > 0$. Новизна результата заключается в том, что удается отказаться от использования методом значения параметра $\mu^{SC} > 0$ и при этом сохранить линейную скорость сходимости, но уже без корня в знаменателе прогрессии.

Во второй части представлена новая модификация метода AGMsDR для решения задач, допускающих альтернированную минимизацию (Alternating AGMsDR). Доказываются аналогичные оценки скорости сходимости на тех же классах оптимизационных задач.

Таким образом, представлены адаптивные ускоренные методы с оценкой сходимости $O\left(\min\left\{\frac{M}{k^2}, \left(1 - \frac{\mu^{PL}}{M}\right)^{(k-1)}\right\}\right)$ на классе выпуклых функций с M-липшицевым градиентом, которые удовлетворяют условию Поляка – Лоясиевича. При этом для работы метода не требуются значения параметров M и μ^{PL}. Если же условие Поляка – Лоясиевича не выполняется, то можно утверждать, что скорость сходимости равна $O\left(\frac{1}{k^2}\right)$, но при этом методы не требуют никаких изменений.

Также рассматривается адаптивная каталист-оболочка неускоренного градиентного метода, которая позволяет доказать оценку скорости сходимости $O\left(\frac{1}{k^2}\right)$. Проведено экспериментальное сравнение неускоренного градиентного метода с адаптивным выбором шага, ускоренного с помощью адаптивной каталист-оболочки с методами AGMsDR, Alternating AGMsDR, APDAGD (Adaptive Primal-Dual Accelerated Gradient Descent) и алгоритмом Синхорна для задачи, двойственной к задаче оптимального транспорта.

Проведенные вычислительные эксперименты показали более быструю работу метода Alternating AGMsDR по сравнению как с неускоренным градиентным методом, ускоренным с помощью адаптивной каталист-оболочки, так и с методом AGMsDR, несмотря на асимптотически одинаковые гарантии скорости сходимости $O\left(\frac{1}{k^2}\right)$. Это может быть объяснено результатом о линейной скорости сходимости метода Alternating AGMsDR на классе задач, удовлетворяющих условию Поляка – Лоясиевича. Гипотеза была проверена на квадратичных задачах. Метод Alternating AGMsDR показал более быструю сходимость по сравнению с методом AGMsDR.

Ключевые слова: выпуклая оптимизация, альтернированная минимизация, ускоренные методы, адаптивные методы, условие Поляка – Лоясиевича

Исследование выполнено за счет гранта Российского научного фонда (проект № 18-71-10108).

COMPUTER RESEARCH AND MODELING
2022 VOL. 14 NO. 2 P. 497–515
DOI: 10.20537/2076-7633-2022-14-2-497-515

Ки&М

MATHEMATICAL MODELING AND NUMERICAL SIMULATION

UDC: 519.8

On accelerated adaptive methods and their modifications for alternating minimization

N. K. Tupitsa[1,2,3]

[1]Moscow Institute of Physics and Technology,
9 Institutskiy per., Dolgoprudny, Moscow region, 141701, Russia
[2]Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute),
19/1 Bol'shoy Karetnyy pereulok, Moscow, 212705, Russia
[3]HSE University,
20 Myasnitskaya st., Moscow, 101000, Russia

Received 15.03.2020, after completion — 12.12.2021.
Accepted for publication 13.02.2022.

In the first part of the paper we present convergence analysis of AGMsDR method on a new class of functions — in general non-convex with M-Lipschitz-continuous gradients that satisfy Polyak – Lojasiewicz condition. Method does not need the value of $\mu^{PL} > 0$ in the condition and converges linearly with a scale factor $\left(1 - \frac{\mu^{PL}}{M}\right)$. It was previously proved that method converges as $O\left(\frac{1}{k^2}\right)$ if a function is convex and has M-Lipschitz-continuous gradient and converges linearly with a scale factor $\left(1 - \sqrt{\frac{\mu^{SC}}{M}}\right)$ if the value of strong convexity parameter $\mu^{SC} > 0$ is known. The novelty is that one can save linear convergence if $\frac{\mu^{PL}}{\mu^{SC}}$ is not known, but without square root in the scale factor.

The second part presents modification of AGMsDR method for solving problems that allow alternating minimization (Alternating AGMsDR). The similar results are proved.

As the result, we present adaptive accelerated methods that converge as $O\left(\min\left\{\frac{M}{k^2}, \left(1 - \frac{\mu^{PL}}{M}\right)^{(k-1)}\right\}\right)$ on a class of convex functions with M-Lipschitz-continuous gradient that satisfy Polyak – Lojasiewicz condition. Algorithms do not need values of M and μ^{PL}. If Polyak – Lojasiewicz condition does not hold, the convergence is $O\left(\frac{1}{k^2}\right)$, but no tuning needed.

We also consider the adaptive catalyst envelope of non-accelerated gradient methods. The envelope allows acceleration up to $O\left(\frac{1}{k^2}\right)$. We present numerical comparison of non-accelerated adaptive gradient descent which is accelerated using adaptive catalyst envelope with AGMsDR, Alternating AGMsDR, APDAGD (Adaptive Primal-Dual Accelerated Gradient Descent) and Sinkhorn's algorithm on the problem dual to the optimal transport problem.

Conducted experiments show faster convergence of alternating AGMsDR in comparison with described catalyst approach and AGMsDR, despite the same asymptotic rate $O\left(\frac{1}{k^2}\right)$. Such behavior can be explained by linear convergence of AGMsDR method and was tested on quadratic functions. Alternating AGMsDR demonstrated better performance in comparison with AGMsDR.

Keywords: convex optimization, alternating minimization, accelerated methods, adaptive methods, Polyak – Lojasiewicz condition

Citation: *Computer Research and Modeling*, 2022, vol. 14, no. 2, pp. 497–515 (Russian).
This research was funded by Russian Science Foundation (project 18-71-10108).

Введение

В данной работе рассматривается задача безусловной оптимизации

$$\min_{x \in \mathbb{R}^N} f(x), \tag{1}$$

где $f(x)$ выпуклая, с M-липшицевым градиентом ($M > 0$). Основные предположения данной статьи — возможность разделить пространство $\mathbb{R}^N$ на n непересекающихся подпространств $L_i \subset \mathbb{R}^N$ таких, что $\Pi_i L_i = \mathbb{R}^N$, и возможность явно минимизировать функцию f на каждом из этих подпространств при прочих фиксированных аргументах. Формально такое предположение означает, что f имеет блочную структуру $f(x) = f(x_1, \ldots, x_n)$, и известно явное решение каждой из задач:

$$x_i^* = \operatorname*{argmin}_{z \in \mathbb{R}^{N_i}} f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n),$$

N_i — размер i-го блока, $\sum N_i = N$, а значения x_j, $j \neq i$ фиксированы.

В этих предположениях классической и естественной является идея альтернированной минимизации [Ortega, Rheinboldt, 2000; Bertsekas, Tsitsiklis, 1989], где функция минимизируется на каждом подпространстве по очереди. Для гладких сильно выпуклых задач при некоторых дополнительных предположениях был получен результат о линейной скорости сходимости в работе [Luo, Tseng, 1993]. В работе [Beck, 2015] изучен алгоритм альтернированной минимизации для двухблочной функции в достаточно общих предположениях. Эти предположения заключались в гладкости целевой функции на хотя бы одном подпространстве L_i. Также показано, что негладкость допустима в виде композитного слагаемого, что не влияет на оценки скорости сходимости. Поскольку среди этих предположений нет предположения о сильной выпуклости целевой функции, удалось обосновать лишь сублинейную скорость сходимости $O\left(\frac{1}{k}\right)$, где k — номер итерации. Похожий результат получен для произвольного числа блоков [Hong et al., 2017; Sun, Hong, 2015]. В предположениях гладкости и сильной выпуклости в работе [Nutini et al., 2015] получен результат о линейной скорости сходимости в случае произвольного числа блоков, скорость сходимости пропорциональна эффективному значению $\left(1 - \frac{\mu^{SC}}{M}\right)$ (понятие эффективного значения определено в той же работе). В работе [Chambolle, Tan, Vaiter, 2017] представлен ускоренный алгоритм альтернированной минимизации для задачи специального вида и для двух блоков, а именно с задачей вида суммы квадратичных функций с композитными членами проксимального вида. Получен результат о скорости сходимости $O\left(\frac{1}{k^2}\right)$ для выпуклых постановок и сходимости со скоростью геометрической прогрессии, знаменатель которой $\left(1 - \sqrt{\frac{\mu^{SC}}{M}}\right)$ для сильно выпуклых. В работе [Diakonikolas, Orecchia, 2018] рассмотрен неускоренный алгоритм альтернированной минимизации и получена скорость сходимости $O\left(\frac{1}{k}\right)$ в выпуклом случае и сходимость со скоростью геометрической прогрессии, знаменатель которой $\left(1 - \frac{\mu^{SC}}{M}\right)$ в сильно выпуклом случае. Также предложен ускоренный метод для задачи в общей выпуклой постановке со сходимостью $O\left(\frac{1}{k^2}\right)$ и высказано предположение о возможности обобщения подхода на сильно выпуклый случай. В работе [Guminov et al., 2019a] рассматривается ускоренный метод в общей выпуклой постановке со скоростью сходимости $O\left(\frac{1}{k^2}\right)$ в гладком случае и $\left(1 - \sqrt{\frac{\mu^{SC}}{M}}\right)$ — в сильно выпуклом случае для произвольного числа блоков. Также стоит упомянуть обзорную работу [Hong et al., 2016], в которой собраны основные результаты об оценках скорости сходимости для рассматриваемой задачи.

В первой части данной работы приведено доказательство сходимости со скоростью геометрической прогрессии, знаменатель которой $\left(1 - \frac{\mu^{PL}}{M}\right)$ на классе функций, удовлетворяющих

условию Поляка – Лоясиевича [Polyak, 1963] с константой $\mu^{PL} > 0$ (или на классе сильно выпуклых функций с константой $\mu^{SC} > 0$) алгоритма 1 из [Nesterov et al., 2020]. При этом существенно, что значение параметра $\mu^{PL/SC} > 0$ не используется при реализации метода. Этот результат, по всей видимости, ранее не был опубликован, несмотря на обширные исследования различных модификаций метода [Guminov et al., 2019b]. Таким образом, представлен метод со скоростью сходимости $O\left(\frac{1}{k^2}\right)$ в гладком выпуклом случае и сходимостью со скоростью геометрической прогрессии, знаменатель которой $\left(1 - \sqrt{\frac{\mu^{SC}}{M}}\right)$ в сильно выпуклом случае, если значение параметра сильной выпуклости $\mu^{SC} > 0$ используется алгоритмом. Также приводится доказательство аналогичных утверждений о скорости сходимости для обобщения алгоритма на задачи, допускающие альтернированную минимизацию.

В следующей секции работы приводится обобщение ускоренного алгоритма альтернированной минимизации из работы [Guminov et al., 2019a] (алгоритм 1) на сильно выпуклый случай.

Также рассматривается другой подход к ускорению альтернированной минимизации, основанный на работе [Ivanova et al., 2019]. С помощью экспериментов проводится сравнение его работы с алгоритмом 1 из [Nesterov et al., 2020] и неускоренным методом альтернированной минимизации на примере задачи, двойственной к задаче оптимального транспорта.

Сходимость ускоренного градиентного метода (AGMsDR) на классе функций, удовлетворяющих условию Поляка – Лоясиевича

Рассмотрим алгоритм 3 из работы [Nesterov et al., 2020]. Для его работы требуется указать значение параметра сильной выпуклости оптимизируемой функции, тогда его сложность описывается выражением

$$f(x_k) - f(x^*) \leqslant \min\left\{\frac{2MR^2}{k^2}, \left(1 - \sqrt{\frac{\mu^{SC}}{M}}\right)^{k-1} MR^2\right\},$$

где $R = \|x_0 - x^*\|_2$, а x_0 — точка старта, x^* — решение задачи.

Если же значение параметра сильной выпуклости неизвестно, то алгоритм 3 из работы [Nesterov et al., 2020] необходимо запустить с $\mu^{in} = 0$. В таком случае алгоритм 3 из [Nesterov et al., 2020] будет в точности совпадать с алгоритмом 1 из той же работы. В этом случае авторы работы [Nesterov et al., 2020] гарантируют лишь сходимость со скоростью

$$f(x_k) - f(x_*) \leqslant \frac{2MR^2}{k^2}.$$

Ниже приведен алгоритм 3 из [Nesterov et al., 2020] — алгоритм 1 данной работы.

Следующая лемма объясняет поведение алгоритма 1, если оптимизируемая функция является сильно выпуклой, но значение параметра не используется алгоритмом. Более того, оказывается, что возможно обосновать оценку скорости сходимости не только для сильно выпуклых функций, но и для более широкого класса задач с так называемым условием Поляка – Лоясиевича [Polyak, 1963]:

$$\|\nabla f(y)\|_2^2 \geqslant 2\mu^{PL}(f(y) - f(x^*)).$$

Теорема 1. *Алгоритм* 1*, запущенный с параметром* $\mu^{in} = 0$*, демонстрирует линейную сходимость на классе функций, удовлетворяющих условию Поляка – Лоясиевича с константой* $\mu^{PL} > 0$*.*

$$f(x^{k+1}) - f(x^*) \leqslant \Pi_{i=0}^{k-1}\left(1 - \frac{\mu^{PL}}{\widehat{M}_i}\right) \cdot (f(x^0) - f(x^*)),$$

где $\widehat{M}_i = \frac{A_i + a_{i+1}}{a_{i+1}^2} \leqslant M$ *— оценка константы Липшица градиента функции M на i-й итерации.*

Алгоритм 1. Accelerated gradient method with small-dimensional relaxation (AGMsDR)

Вход: $x^0 = v^0$, $\mu^{in} \geqslant 0$

Выход: x^k

1: Полагаем $k = 0$, $A_0 = 0$, м^0 = ч^0, $\psi_0(x) = \frac{1}{2}\|x - x^0\|_2$

2: **for** $k \geqslant 0$ **do**

3:
$$\beta_k = \underset{\beta\in[0,\,1]}{\operatorname{argmin}} f\left(v^k + \beta(x^k - v^k)\right), \quad y^k = v^k + \beta_k(x^k - v^k). \tag{2}$$

4: Вариант а): значение M известно,
$$x^{k+1} = \arg\min_{x\in E}\left\{f(y^k) + \langle\nabla f(y^k),\, x - y^k\rangle + \frac{M}{2}\|x - y^k\|_2^2\right\}. \tag{3}$$
Найти a_{k+1} из уравнения $\frac{a_{k+1}^2}{(A_k+a_{k+1})(\tau_k+\mu^{in}a_{k+1})} = \frac{1}{M}$.
Вариант б):
$$h_{k+1} = \arg\min_{h\geqslant 0} f\left(y^k - h\nabla f(y^k)\right), \quad x^{k+1} = y^k - h_{k+1}\nabla f(y^k). \tag{4}$$
Найти a_{k+1} из уравнения
$$f(y^k) - \frac{a_{k+1}^2}{2(A_k + a_{k+1})(\tau_k + \mu^{in}a_{k+1})}\|\nabla f(y^k)\|_2^2 + \frac{\mu^{in}\tau_k a_{k+1}}{2(A_k + a_{k+1})(\tau_k + \mu^{in}a_{k+1})}\|v^k - y^k\|_2^2 = f(x^{k+1}). \tag{5}$$

5: Полагаем $A_{k+1} = A_k + a_{k+1}$.

6: Полагаем $\psi_{k+1}(x) = \psi_k(x) + a_{k+1}\{f(y^k) + \langle\nabla f(y^k),\, x - y^k\rangle\}$.

7: $v^{k+1} = \arg\min_x \psi_{k+1}(x)$

8: $k = k + 1$

9: **end for**

Заметим, что запустить алгоритм можно с любыми значениями μ^{in}, такими, что $0 \leqslant \mu^{in} \leqslant \mu^{SC}$,[1] и при этом гарантировать работу алгоритма согласно оценке $f(x^{k+1}) - f(x^*) \leqslant \Pi_{i=0}^{k-1}\left(1 - \frac{\mu^{PL}}{\widetilde{M}_i}\right)\cdot(f(x^0) - f(x^*))$. Но при запуске алгоритма со значением $\mu^{in} = 0$ также гарантируется сходимость со скоростью $f(x_k) - f(x^*) \leqslant \frac{2MR^2}{k^2}$ для выпуклой функции с липшицевым градиентом.

Для обеспечения линейной скорости сходимости, соответствующей нижним оценкам, необходимо запустить алгоритм со значением $\mu^{in} = \mu^{SC}$.

Доказательство. В оригинальном доказательстве теоремы 1 из [Nesterov et al., 2020] значение a_{k+1} выбиралось таким, чтобы гарантировать выполнение неравенства
$$A_{k+1}f(y^k) - \frac{a_{k+1}^2}{2}\|\nabla f(y^k)\|_2^2 \geqslant A_{k+1}f(x^{k+1}). \tag{6}$$
Так, для варианта а) для a_{k+1} выполняется $\frac{a_{k+1}^2}{A_{k+1}} = \frac{1}{M}$, и это означает
$$f(y^k) - \frac{1}{2M}\|\nabla f(y^k)\|_2^2 \geqslant f(x^{k+1}) \geqslant f(y^{k+1}). \tag{7}$$

[1] У читателя может возникнуть вопрос, будет ли в таком случае уравнение (5) иметь неотрицательное решение. Доказательство этого факта для алгоритма 1 не отличается от аналогичного для уравнения (10) в алгоритме 2.

Для варианта б) a_{k+1} является наибольшим решением уравнения

$$A_{k+1}f(y^k) - \frac{a_{k+1}^2}{2}\|\nabla f(y^k)\|_2^2 = A_{k+1}f(x^{k+1}). \tag{8}$$

Покажем, что $\widehat{M}^{k+1} := \frac{A_{k+1}}{a_{k+1}^2} \leqslant M$. M-липшицевость градиента позволяет записать соотношения

$$f(x^{k+1}) \leqslant \min_x \left(f(y^k) + \langle \nabla f(y^k),\, x - y^k\rangle + \frac{M}{2}\|x - y^k\|_2^2 \right) = f(y^k) - \frac{1}{2M}\|\nabla f(y^k)\|_2^2.$$

Тогда из уравнения (8) получим

$$f(y^k) - \frac{1}{2\widehat{M}^{k+1}}\|\nabla f(y^k)\|_2^2 = f(x^{k+1}) \leqslant f(y^k) - \frac{1}{2M}\|\nabla f(y^k)\|_2^2,$$

откуда и следует требуемое неравенство $\widehat{M}^{k+1} \leqslant M$.

Итак, в обоих вариантах работы алгоритма имеем

$$f(y^k) - \frac{1}{2\widehat{M}^{k+1}}\|\nabla f(y^k)\|_2^2 \geqslant f(x^{k+1}) \geqslant f(y^{k+1}),$$

причем $\widehat{M}^{k+1} \leqslant M$.

Далее рассмотрим условие Поляка – Лоясиевича[1] в точке y^k:

$$\|\nabla f(y^k)\|_2^2 \geqslant 2\mu^{PL}\left(f(y^k) - f(x^*)\right).$$

Скомпоновав последнее неравенство с (1), получим

$$f(y^{k+1}) - f(x^*) \leqslant \left(1 - \frac{\mu^{PL}a_{k+1}^2}{A_k + a_{k+1}}\right)\left(f(y^k) - f(x^*)\right) \leqslant \Pi_{i=0}^k \left(1 - \frac{\mu^{PL}a_{i+1}^2}{A_i + a_{i+1}}\right)\left(f(x^0) - f(x^*)\right).$$

И наконец, воспользуемся условием (4), которое обеспечивает $f(x^{k+1}) \leqslant f(y^k)$, и получим линейную скорость сходимости:

$$f(x^{k+1}) - f(x^*) \leqslant \Pi_{i=0}^{k-1} \left(1 - \frac{\mu^{PL}a_{i+1}^2}{A_i + a_{i+1}}\right)(f(x^0) - f(x^*)).$$

□

Таким образом, наблюдается сходимость со скоростью геометрической прогрессии, если функция (вообще говоря, невыпуклая) имеет M-липшицев градиент и удовлетворяет условию Поляка – Лоясиевича, но алгоритму информация об этом недоступна (алгоритм 1 запускается с параметром $\mu^{in} = 0$).

Заметим также, что для глобальной константы Липшица градиента функции может выполняться $M > \frac{A_i + a_{i+1}}{a_{i+1}^2}$, а также для μ_k может выполняться $\mu_k^{PL} \geqslant \mu^{PL}$, что означает, что, согласно обоснованной в предыдущей теореме оценке, скорость сходимости может оказаться лучше, чем со знаменателем $\left(1 - \frac{\mu^{PL}}{M}\right)$, где M — глобальная константа Липшица градиента, а μ^{PL} — константа в условии Поляка – Лоясиевича.

[1] Данное условие выполняется в произвольной точке, но может оказаться, что в некоторых точках условие будет верно с константой $\mu_k \geqslant \mu^{PL}$, то есть $\|\nabla f(y^k)\|_2^2 \geqslant 2\mu_k^{PL}\left(f(y^k) - f(x^*)\right)$. Поэтому на практике сходимость может оказаться более быстрой, чем предписывает данная теорема.

Таким образом, данный алгоритм демонстрирует нижние оценки сложности на классе выпуклых функций с липшицевым градиентом — $O\left(\frac{1}{k^2}\right)$, а также на классе сильно выпуклых функций, если константа сильной выпуклости известна, — сходимость со скоростью геометрической прогрессии, знаменатель которой $\left(1 - \sqrt{\frac{\mu^{SC}}{M}}\right)$. Кроме этого, данный алгоритм демонстрирует сходимость со скоростью геометрической прогрессии на классе функций, удовлетворяющих условию Поляка – Лоясиевича (в том числе и для сильно выпуклых задач), если константа $\mu^{PL} > 0$ неизвестна, но в этом случае доказана сходимость со скоростью геометрической прогрессии лишь со знаменателем $\left(1 - \frac{\mu^{PL}}{M}\right)$ (алгоритм в этом случае запускается с $\mu^{in} = 0$).

Альтернированная минимизация

В этом разделе рассмотрим вариацию алгоритма 1 для задач, допускающих альтернированную минимизацию. В работе [Guminov et al., 2019a] рассмотрена такая вариация для функций с липшицевым градиентом. Далее приводится модификация алгоритма для решения сильно выпуклых задач, допускающих альтернированную минимизацию. Введем необходимые обозначения. Множество $\{1, \ldots, N\}$ векторов $\{e_i\}_{i=1}^N$ ортонормированного базиса разделено на n непересекающихся блоков I_k, $k \in \{1, \ldots, n\}$ (см. введение). Пусть $S_k(x) = x + \text{span}\{e_i : i \in I_k\}$ — подпространство, содержащее x, построенное на базисных векторах k-го блока.

Алгоритм 2. Accelerated alternating minimization

Вход: Начальная точка x_0, $\mu^{in} \geqslant 0$
Выход: x^k

1: Полагаем $A_0 = 0$, $x^0 = v^0$, $\tau_0 = 1$
2: **for** $k \geqslant 0$ **do**
3: Полагаем

$$\beta_k = \underset{\beta \in [0,\,1]}{\operatorname{argmin}} f\left(x^k + \beta(v^k - x^k)\right) \tag{9}$$

4: Полагаем $y^k = x^k + \beta_k(v^k - x^k)$
5: Выбираем $i_k = \underset{i \in \{1, \ldots, n\}}{\operatorname{argmax}} \|\nabla_i f(y^k)\|_2^2$
6: Полагаем $x^{k+1} = \underset{x \in S_{i_k}(y^k)}{\operatorname{argmin}} f(x)$
7: Находим a_{k+1} из уравнения

$$f(y^k) - \frac{a_{k+1}^2}{2(A_k + a_{k+1})(\tau_k + \mu^{in} a_{k+1})}\|\nabla f(y^k)\|_2^2 + \frac{\mu^{in}\tau_k a_{k+1}}{2(A_k + a_{k+1})(\tau_k + \mu^{in} a_{k+1})}\|v^k - y^k\|_2^2 = f(x^{k+1}) \tag{10}$$

8: Полагаем $A_{k+1} = A_k + a_{k+1}$, $\tau_{k+1} = \tau_k + \mu^{in} a_{k+1}$
9: Полагаем $v^{k+1} = \underset{x \in \mathbb{R}^N}{\operatorname{argmin}} \psi_{k+1}(x)$
10: **end for**

Заметим, что запустить алгоритм можно с любыми значениям μ^{in}, такими, что $0 \leqslant \mu^{in} \leqslant \mu^{SC}$, но гарантировать работу алгоритма, согласно оценкам, которые будут получены в теореме 2, можно, если $\mu^{in} = \mu^{SC}$, что существенным образом используется при доказательстве далее.

Введем вспомогательную последовательность функций

$$\psi_0(x) = \frac{1}{2}\|x - x^0\|_2^2,$$

$$\psi_{k+1}(x) = \psi_k(x) + a_{k+1}\left\{f(y^k) + \langle \nabla f(y^k),\ x - y^k\rangle\ + \frac{\mu^{SC}}{2}\|x - y^k\|_2^2\right\}.$$

Используя следующее обозначение:

$$l_k(x) = \sum_{i=0}^{k} a_{i+1}\left\{f(y^i) + \langle \nabla f(y^i),\ x - y^i\rangle + \frac{\mu^{SC}}{2}\|x - y^i\|_2^2\right\},$$

получим нерекуррентное представление

$$\psi_{k+1}(x) = \psi_0(x) + l_k(x).$$

Заметим, что $\psi_k(x)$ является сильно выпуклой функцией со значением параметра сильной выпуклости $\tau_k = 1 + \mu^{SC}\sum\limits_{i=0}^{k} a_i = 1 + \mu^{SC}A_k$.

Лемма 1. *После k итераций алгоритма 2 выполняется*

$$A_k f(x^k) \leqslant \min_{x\in\mathbb{R}^N} \psi_k(x) = \psi_k(v^k). \tag{11}$$

Более того, если функция имеет M-липшицев градиент и является сильно выпуклой, с константой $\mu^{SC} \geqslant 0$, то

$$A_k \geqslant \max\left\{\frac{k^2}{4nM},\ \frac{1}{nM}\left(1 - \sqrt{\frac{\mu^{SC}}{nM}}\right)^{-k+1}\right\},$$

где n — количество блоков, по которым допускается явная минимизация.

Доказательство. Докажем (11) индукцией по k. При $k = 0$ это неравенство верно. Предположим

$$A_k f(x^k) \leqslant \min_{x\in\mathbb{R}^N} \psi_k(x) = \psi_k(v^k).$$

Тогда

$$\psi_{k+1}(v^{k+1}) = \min_{x\in\mathbb{R}^N}\left\{\psi_k(x) + a_{k+1}\left\{f(y^k) + \langle \nabla f(y^k),\ x - y^k\rangle + \frac{\mu^{SC}}{2}\|x - y^k\|_2^2\right\}\right\} \geqslant$$

$$\geqslant \min_{x\in\mathbb{R}^N}\left\{\psi_k(v^k) + \frac{\tau_k}{2}\|x - v^k\|_2^2 + a_{k+1}\left\{f(y^k) + \langle \nabla f(y^k),\ x - y^k\rangle + \frac{\mu^{SC}}{2}\|x - y^k\|_2^2\right\}\right\} \geqslant$$

$$\geqslant \min_{x\in\mathbb{R}^N}\left\{A_k f(x^k) + \frac{\tau_k}{2}\|x - v^k\|_2^2 + a_{k+1}\left\{f(y^k) + \langle \nabla f(y^k),\ x - y^k\rangle + \frac{\mu^{SC}}{2}\|x - y^k\|_2^2\right\}\right\},$$

где было использовано, что ψ_k сильно выпуклая и имеет минимум в v^k, и то, что $f(y^k) \leqslant f(x^k)$.

Условия оптимальности для $\min\limits_{\beta\in[0,\,1]} f\left(x^k + \beta(v^k - x^k)\right)$ гарантируют выполнение одного из следующих условий:

1) $\beta_k = 1$, $\langle \nabla f(y^k),\ x^k - v^k\rangle \geqslant 0$, $y^k = v^k$;

2) $\beta_k \in (0,\ 1)$ и $\langle \nabla f(y^k),\ x^k - v^k\rangle = 0$, $y^k = v^k + \beta_k(x^k - v^k)$;

3) $\beta_k = 0$ и $\langle \nabla f(y^k),\ x^k - v^k\rangle \leqslant 0$, $y^k = x^k$.

Во всех случаях выполняется неравенство $\langle \nabla f(y^k),\ v^k - y^k\rangle \geqslant 0$.

Таким образом,

$$\psi_{k+1}(v^{k+1}) \geqslant \min_{x\in\mathbb{R}^n}\left\{A_k f(y^k) + \frac{\tau_k}{2}\|x-v^k\|_2^2 + a_{k+1}\left\{f(y^k) + \langle\nabla f(y^k),\, x-y^k\rangle + \frac{\mu^{SC}}{2}\|x-y^k\|_2^2\right\}\right\}. \quad (12)$$

Решение задачи поиска минимума в (12) записывается явно:

$$x = \frac{1}{\tau_{k+1}}(\tau_k v^k + \mu^{SC} a_{k+1} y^k - a_{k+1}\nabla f(y^k)).$$

Значение минимизируемой функции в (12) в точке минимума, с учетом неравенства $\langle\nabla f(y^k),\, v^k - y^k\rangle \geqslant 0$, позволяет получить неравенство

$$\psi_{k+1}(v^{k+1}) \geqslant A_{k+1} f(y^k) - \frac{a_{k+1}^2}{2\tau_{k+1}}\|\nabla f(y^k)\|_2^2 + \frac{\mu^{SC}\tau_k a_{k+1}}{2\tau_{k+1}}\|v^k - y^k\|_2^2.$$

Далее, покажем, что

$$A_{k+1} f(y^k) - \frac{a_{k+1}^2}{2\tau_{k+1}}\|\nabla f(y^k)\|_2^2 + \frac{\mu^{SC}\tau_k a_{k+1}}{2\tau_{k+1}}\|v^k - y^k\|_2^2 \geqslant A_{k+1} f(x^{k+1}),$$

что завершит индукционный переход.

Для этого, принимая во внимание, что f имеет M-липшицев градиент, получим $\forall i$

$$f(y^k) - \frac{1}{2M}\|\nabla_i f(y^k)\|_2^2 \geqslant f(x_i^{k+1}),$$

где $x_i^{k+1} = \underset{x\in S_i}{\operatorname{argmin}} f(x)$. Так как $i_k = \underset{i}{\operatorname{argmax}}\,\|\nabla_i f(y^k)\|_2^2$, то

$$\|\nabla_{i_k} f(y^k)\|_2^2 \geqslant \frac{1}{n}\|\nabla f(y^k)\|_2^2$$

и

$$f(y^k) - \frac{1}{2nM}\|\nabla f(y^k)\|_2^2 \geqslant f(y^k) - \frac{1}{2M}\|\nabla_{i_k} f(y^k)\|_2^2 \geqslant f(x^{k+1}).$$

Выбор a_{k+1}, согласно неравенству (10), означает, что

$$-\frac{a_{k+1}^2}{2(A_k + a_{k+1})(\tau_k + \mu^{SC} a_{k+1})}\|\nabla f(y^k)\|_2^2 + \frac{\mu^{SC}\tau_k a_{k+1}}{2(A_k + a_{k+1})(\tau_k + \mu^{SC} a_{k+1})}\|v^k - y^k\|_2^2 = f(x^{k+1}) - f(y^k). \quad (13)$$

Покажем, что (10) и (13) имеют положительный корень, для этого перепишем (10) в виде

$$\left(2\mu^{in}\delta_k + \|\nabla f(y^k)\|_2^2\right)\alpha_{k+1}^2 + \left(2\delta_k\left(\tau_k + \mu^{in} A_k\right)\right) - \mu^{in}\tau_k\|v^k - y^k\|_2^2\alpha_{k+1} + 2\tau_k A_k\delta_k = 0,$$

где $\delta_k = f(x^{k+1}) - f(y_k) < 0$ (в противном случае y_k — точка минимума). Сильная выпуклость f означает, что $f(y_k) - f(x^*) \leqslant \frac{1}{2\mu^{SC}}\|\nabla f(y^k)\|_2^2$. Случай $\mu^{in} = 0$ рассмотрим отдельно далее. Последнее неравенство остается верным и для μ^{in} такого, что $0 < \mu^{in} \leqslant \mu^{SC}$. Случай $\mu^{in} = 0$ рассмотрим отдельно далее. То есть также верно и $f(y_k) - f(x^*) \leqslant \frac{1}{2\mu^{in}}\|\nabla f(y^k)\|_2^2$, так как $f(y_k) - f(x^*) \leqslant \leqslant \frac{1}{2\mu^{SC}}\|\nabla f(y^k)\|_2^2 \leqslant \frac{1}{2\mu^{in}}\|\nabla f(y^k)\|_2^2$. Итак, имеем два неравенства:

$$f(y_k) - f(x^*) \leqslant \frac{1}{2\mu^{SC}}\|\nabla f(y^k)\|_2^2$$

и

$$f(y_k) - f(x^*) \leqslant \frac{1}{2\mu^{in}}\|\nabla f(y^k)\|_2^2.$$

К обеим частям этих неравенств добавим $f(x^{k+1})$, а затем умножим обе части на μ^{SC} и μ^{in} соответственно и получим следующие два неравенства:

$$2\mu^{SC}\delta_k + \|\nabla f(y^k)\|_2^2 \geqslant 2\mu^{SC}(f(x_{k+1}) - f(x^*)) \geqslant 0$$

и

$$2\mu^{in}\delta_k + \|\nabla f(y^k)\|_2^2 \geqslant 2\mu^{in}(f(x_{k+1}) - f(x^*)) \geqslant 0.$$

Если $\mu^{in} = 0$ или $\mu^{SC} = 0$, то последние два неравенства также, очевидно, останутся верными. Так как

$$2\mu^{SC}\delta_k + \|\nabla f(y^k)\|_2^2 \geqslant 0$$

и

$$2\mu^{in}\delta_k + \|\nabla f(y^k)\|_2^2 \geqslant 0,$$

а также что $\delta_k = f(x^{k+1}) - f(y_k) < 0$, то, в силу неотрицательности A_k и τ_k,

$$-A_k\delta_k\tau_k(2\delta_k\mu^{SC} + \|\nabla f(y^k)\|_2^2) \geqslant 0$$

и, соответственно,

$$-A_k\delta_k\tau_k(2\delta_k\mu^{in} + \|\nabla f(y^k)\|_2^2) \geqslant 0.$$

Таким образом, неотрицательное решение (13) действительно существует и записывается в виде

$$\alpha_{k+1} = \frac{-S_k + \sqrt{S_k^2 - 8A_k\delta_k\tau_k(2\delta_k\mu^{SC} + \|\nabla f(y^k)\|_2^2)}}{2(2\mu^{SC}\delta_k + \|\nabla f(y^k)\|_2^2)};$$

аналогично: неотрицательное решение (13) существует и записывается в виде

$$\alpha_{k+1} = \frac{-S_k + \sqrt{S_k^2 - 8A_k\delta_k\tau_k(2\delta_k\mu^{in} + \|\nabla f(y^k)\|_2^2)}}{2(2\mu^{in}\delta_k + \|\nabla f(y^k)\|_2^2)},$$

где $S_k = 2\delta_k(\tau_k + \mu^{SC}A_k) - \mu^{SC}\tau_k\|v^k - y^k\|_2^2$ и $S_k = 2\delta_k(\tau_k + \mu^{in}A_k) - \mu^{in}\tau_k\|v^k - y^k\|_2^2$ соответственно. Необходимо выбрать неотрицательное решение, так как далее будет показано, что чем быстрее растет последовательность α_k, тем быстрее будет расти последовательность A_k и тем быстрее будет сходиться алгоритм.

Равенство (13) в совокупности с неравенством

$$f(y^k) - \frac{1}{2nM}\|\nabla f(y^k)\|_2^2 \geqslant f(x^{k+1})$$

означает, что

$$-\frac{a_{k+1}^2}{2(A_k + a_{k+1})(\tau_k + \mu^{SC}a_{k+1})} \leqslant -\frac{1}{2nM},$$

или

$$\frac{a_{k+1}^2}{2A_{k+1}\tau_{k+1}} \geqslant \frac{1}{2nM}. \tag{14}$$

Преобразовав выражение a_{k+1},[1] получим $\frac{a_{k+1}^2}{(A_k+a_{k+1})(\tau_k+\mu^{SC}a_{k+1})} \geqslant \frac{1}{nM}$.

Теперь оценим скорость роста последовательности A_k. Выпишем выражение для

$$\tau_k = 1 + \mu^{SC}\sum_{i=0}^{k} a_i = 1 + \mu^{SC}A_k.$$

Из (14) следует, что

$$a_k^2 \geqslant \frac{A_k\tau_k}{nM} = \frac{A_k + \mu^{SC}A_k^2}{nM}.$$

Извлекая квадратный корень из a_k^2 и используя $A_k \geqslant 0$, получим

$$a_k \geqslant \frac{1}{\sqrt{nM}}\sqrt{A_k + \mu^{SC}A_k^2} \geqslant \sqrt{\frac{\mu^{SC}}{2nM}}A_k. \tag{15}$$

Выполним следующие преобразования:

$$\sqrt{A_i} - \sqrt{A_{i-1}} \geqslant \frac{A_i - A_{i-1}}{\sqrt{A_i} + \sqrt{A_{i-1}}} \geqslant \frac{a_i}{2\sqrt{A_i}} \geqslant \frac{\sqrt{1+\mu^{SC}A_i}}{2\sqrt{nM}},$$

где последнее неравенство следует из (15), а также использовано, что $a_i = A_i - A_{i-1}$ и $A_i \geqslant A_{i-1}$. Просуммировав последние неравенства по $i = 1, \ldots, k$, получим

$$A_k \geqslant \frac{k^2}{4nM}. \tag{16}$$

Далее,

$$A_{k+1} = A_k + a_{k+1} \geqslant A_k + \sqrt{\frac{\mu^{SC}}{nM}}A_{k+1},$$

что означает

$$A_{k+1} \geqslant \left(1 - \sqrt{\frac{\mu^{SC}}{nM}}\right)^{-1} A_k. \tag{17}$$

Остается оценить A_1:

$$A_1 = \frac{a_1^2}{A_1} \geqslant \frac{a_1^2}{(1+\mu^{SC}A_1)A_1} \geqslant \frac{a_1^2}{A_1\tau_1} \geqslant \frac{1}{nM}.$$

Рекурсивно применяя оценки (17) и (16), приходим к утверждению леммы:

$$A_k \geqslant \max\left\{\frac{k^2}{4nM}, \frac{1}{nM}\left(1 - \sqrt{\frac{\mu^{SC}}{nM}}\right)^{-k+1}\right\}.$$

□

[1] Если константа Липшица градиента известна, то также можно найти a_{k+1} из уравнения $\frac{a_{k+1}^2}{(A_k+a_{k+1})(\tau_k+\mu^{SC}a_{k+1})} = \frac{1}{nM}$ аналогично AGMsDR. При этом неравенство (14) будет удовлетворено, и дальнейшее доказательство останется верным. Также заметим, что неравенство (14) позволяет локально оценить константу Липшица градиента, и метод, таким образом, настраивается на локальную гладкость задачи. Вариант метода с известной константой Липшица градиента использует глобальную информацию о задаче — константу Липшица градиента, что на практике приводит к более медленной сходимости.

Теорема 2. *После k итераций алгоритма* 2 *выполняется*

$$f(x^k) - f(x_*) \leqslant nMR^2 \min\left\{\frac{4}{k^2}, \left(1 - \sqrt{\frac{\mu^{SC}}{nM}}\right)^{k-1}\right\}, \tag{18}$$

где $R = \|x_0 - x^*\|_2$.

Снова заметим, что запустить алгоритм можно с любыми значениями μ^{in}, но гарантировать работу алгоритма согласно оценке $f(x^k) - f(x_*) \leqslant nMR^2 \min\left\{\frac{4}{k^2}, \left(1 - \sqrt{\frac{\mu^{in}}{nM}}\right)^{k-1}\right\}$ возможно, только если $0 \leqslant \mu^{in} \leqslant \mu^{SC}$, так как используемые неравенства — следствия сильной выпуклости — остаются верными при замене μ^{SC} на μ^{in} при $0 \leqslant \mu^{in} \leqslant \mu^{SC}$.[1] В доказательстве используется именно μ^{SC}, как значение, обеспечивающее наиболее быструю сходимость.

Доказательство. Выпуклость функции $f(x)$ позволяет получить неравенство

$$l_k(x_*) = \sum_{i=0}^{k} a_{i+1}\left(f(y^i) + \langle \nabla f(y^i),\, x_* - y^i\rangle + \frac{\mu^{SC}}{2}\|x_* - y^i\|_2^2\right) \leqslant A_{k+1} f(x_*).$$

По лемме 1,

$$A_k f(x^k) \leqslant \psi_k(v^k) \leqslant \psi_k(x_*) = \frac{1}{2}\|x_* - x^0\|_2^2 + \\ + \sum_{i=0}^{k-1} a_{i+1}\left(f(y^i) + \langle \nabla f(y^i),\, x_* - y^i\rangle + \frac{\mu^{SC}}{2}\|x_* - y^i\|_2^2\right) \leqslant A_k f(x_*) + \frac{1}{2}\|x_* - x^0\|_2^2.$$

И наконец,

$$f(x^k) - f(x_*) \leqslant \frac{R^2}{2A_k} \leqslant nMR^2 \min\left\{\frac{4}{k^2}, \left(1 - \sqrt{\frac{\mu^{SC}}{nM}}\right)^{k-1}\right\}.$$

□

В альтернированной версии алгоритма минимизируется только один блок, в отличие от не альтернированной, где шаг выполняется во всем пространстве. Поэтому можно гарантировать убывание функции на каждой итерации на меньшую величину

$$\frac{1}{2nM}\|\nabla f(y^k)\|_2^2 \leqslant \frac{1}{2M}\|\nabla_{i_k} f(y^k)\|_2^2 \leqslant f(y^k) - f(x^{k+1})$$

для альтернированной версии и

$$\frac{1}{2M}\|\nabla f(y^k)\|_2^2 \leqslant f(y^k) - f(x^{k+1})$$

для не альтернированной. Можно видеть, что оценки сходимости одинаковы с точностью до умножения M на n. Поэтому в альтернированной версии сходимость по итерациям медленнее, но и асимптотическая стоимость одной итерации остается такой же — $O(N)$ — за счет обновления вектора v, хотя и является менее дорогой за счет менее вычислительно сложного шага явной минимизации — $O(N_i)$.

[1] В теоремах используется неравенство $f(x) \geqslant f(y^k) + \langle \nabla f(y^k),\, x - y^k\rangle + \frac{\mu^{SC}}{2}\|x - y^k\|_2^2$, которое вместе с $0 \leqslant \mu^{in} \leqslant \mu^{SC}$ и означает, что верно будет также и $f(x) \geqslant f(y^k) + \langle \nabla f(y^k),\, x - y^k\rangle + \frac{\mu^{in}}{2}\|x - y^k\|_2^2$.

Теорема 3. *Алгоритм* 2*, запущенный с параметром* $\mu^{in} = 0$*, демонстрирует линейную сходимость на классе функций, удовлетворяющих условию Поляка–Лоясиевича с константой* $\mu^{PL} > 0$*.*

$$f(x^{k+1}) - f(x^*) \leqslant \Pi_{i=0}^{k-1}\left(1 - \frac{\mu^{PL}}{\widehat{M}_i}\right)\cdot(f(x^0) - f(x^*)),$$

где $\widehat{M}_i = \frac{A_i + a_{i+1}}{a_{i+1}^2} \leqslant M$ *является оценкой константы Липшица градиента функции* M *на* i*-й итерации.*

Доказательство полностью повторяет доказательство леммы 1, если заменить M в доказательстве леммы 1 на nM. □

Снова заметим, что запустить алгоритм можно с любым значениями μ^{in} и при этом гарантировать работу алгоритма согласно оценке $f(x^{k+1}) - f(x^*) \leqslant \Pi_{i=0}^{k-1}\left(1 - \frac{\mu^{PL}}{\widehat{M}_i}\right)\cdot(f(x^0) - f(x^*))$. Но при запуске алгоритма со значением $\mu^{in} = 0$ также гарантируется сходимость со скоростью $f(x_k) - f(x^*) \leqslant \frac{4nMR^2}{k^2}$ для выпуклой функции с липшицевым градиентом.

Таким образом, наблюдается сходимость со скоростью геометрической прогрессии, если функция (вообще говоря, невыпуклая) имеет M-липшицев градиент и удовлетворяет условию Поляка – Лоясиевича, но алгоритму информация об этом недоступна (алгоритм 2 запускается с параметром $\mu^{in} = 0$).

Следует сказать, что, несмотря на отсутствие различий в доказательствах, последовательности α_k и A_k различаются в представленных модификациях алгоритмов. Способ генерации последовательностей одинаковый, но последовательности для одной и той же задачи будут разными, так как точки x^{k+1} различаются. В первом случае точка x^{k+1} получается выполнением градиентного шага, а во втором — шага блочной минимизации, а значение $f(x^{k+1})$ используется при генерации $f(y^k) - \frac{a_{k+1}^2}{2(A_k + a_{k+1})}\|\nabla f(y^k)\|_2^2 = f(x^{k+1})$.

Численные эксперименты

Большое количество исследований посвящено решению задач оптимального транспорта [Cuturi, 2013; Dvurechensky, Gasnikov, Kroshnin, 2018; Guminov et al., 2019a], поиску барицентров Вассерштейна [Kroshnin et al., 2019; Dvinskikh et al., 2019; Uribe et al., 2018; Dvurechensky et al., 2018; Lin et al., 2020], а также многомаргинального оптимального транспорта [Lin et al., 2019; Tupitsa et al., 2020] двойственными методами. Все эти задачи допускают альтернированную минимизацию. Далее проводится экспериментальное сравнение представленных алгоритмов на примере задачи, двойственной к энтропийно регуляризованной задаче дискретного оптимального транспорта (ЭОТ) [Guminov et al., 2019a].

Как известно [Guminov et al., 2019a], задача ЭОТ выглядит следующим образом:

$$\min_{u,\, v \in \mathbb{R}^N} f(u,\, v) = \gamma(\ln\left(\mathbf{1}^T B(u,\, v)\mathbf{1}\right) - \langle u,\, r\rangle - \langle v,\, c\rangle), \tag{19}$$

где $[B(u,\, v)]^{ij} = \exp\left(u^i + v^j - \frac{C^{ij}}{\gamma}\right)$, $B,\, C \in \mathbb{R}_+^{N\times N}$, $\gamma \in \mathbb{R}_+$. Переменные в этой задаче естественным образом разделяются на два блока, и при фиксированных переменных одного блока удается явно выписать решение условий оптимальности по другому блоку. Таким образом, в результате обновления переменных и получается алгоритм Синхорна.

В работе [Dvurechensky, Gasnikov, Kroshnin, 2018] представлен алгоритм APDAGD, который показал более быструю сходимость по сравнению с алгоритмом Синхорна.

В работе [Guminov et al., 2019b] алгоритм 2, примененный к этой задаче, показал себя наиболее стабильным и быстрым в сравнении с другими алгоритмами образом.

На рис. 1 приводится экспериментальное сравнение алгоритмов 1 и 2, а также алгоритма Синхорна и алгоритма APDAGD на примере задачи ЭОТ.

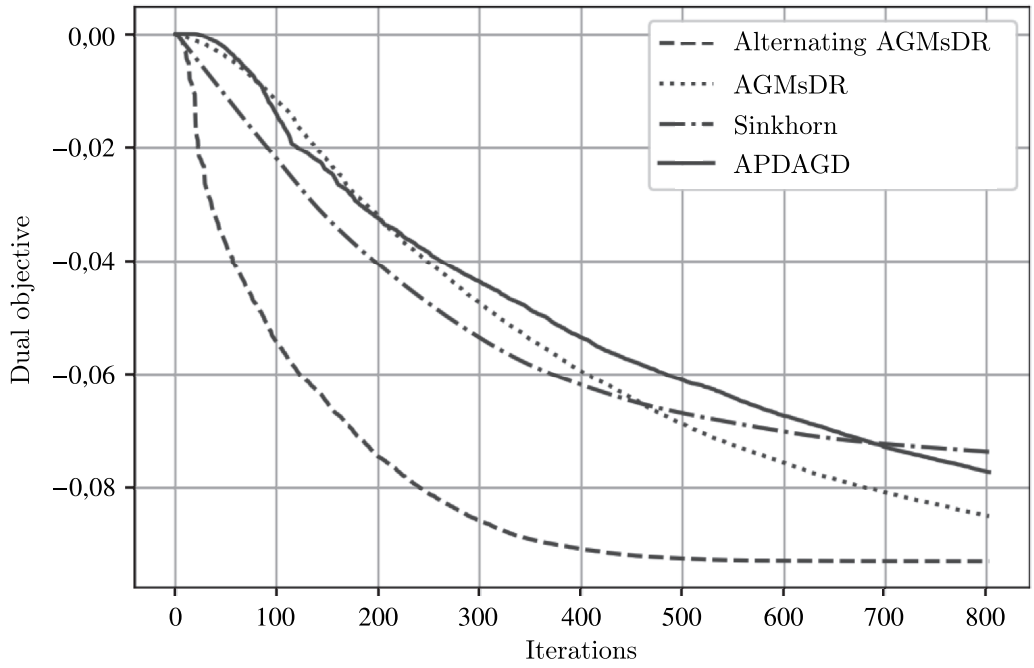

Рис. 1. Сходимость алгоритма Синхорна и методов AGMsDR, Alternating AGMsDR и APDAGD, примененных к ЭОТ

Следует заметить, что теоретические оценки сложности для Alternating AGMsDR показывают, что для достижения заданной точности ему требуется в $\sqrt{n}$ раз большее число итераций, чем для AGMsDR, при одинаковой асимптотической стоимости одной итерации, где n — число блоков, по которым возможна альтернированная минимизация, причем для ЭОТ $n = 2$. Экспериментальное сравнение дает противоположные результаты. Одной из целей данной работы является попытка объяснения этого явления. Предполагается, что такое поведение связано с адаптивностью методов Alternating AGMsDR и AGMsDR к сильной выпуклости. Как известно, задача ЭОТ не является сильно выпуклой, так как значение целевой функции инвариантно на прямых, параллельных вектору $(\mathbf{1}, -\mathbf{1})$. Поэтому схожая сходимость методов AGMsDR и APDAGD является ожидаемой, так как в не сильно выпуклом случае скорость сходимости имеет порядок $O\left(\frac{1}{k^2}\right)$. Alternating AGMsDR также имеет сходимость порядка $O\left(\frac{1}{k^2}\right)$ для не сильно выпуклых функций, но, вероятно, распознает сильную выпуклость на подпространствах переменных u и v, ортогональных вектору $(\mathbf{1}, -\mathbf{1})$, что может приводить к наблюдаемой более быстрой сходимости.

Данная гипотеза была также проверена на квадратичных задачах

$$\min_z f(z) = \|Wz - b\|_2^2. \tag{20}$$

Матрица W в (20) является симметричной и положительно определенной, и тогда f является сильно выпуклой с константой $\mu^{SC} = \lambda_{\min}(W^T W)$.

Последняя задача может быть решена с помощью алгоритма 1.

Построим эквивалентную задачу, допускающую альтернированную минимизацию путем разделения вектора z на два блока одинакового размера:

$$z = \begin{pmatrix} x \\ y \end{pmatrix}.$$

Также разделим матрицу W на 4 блока одинакового размера:

$$W = \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right), \quad b = \begin{pmatrix} d \\ c \end{pmatrix},$$

вектор b — на два вектора:

$$b = \begin{pmatrix} c \\ d \end{pmatrix}.$$

Тогда задача (20) будет эквивалентна задаче

$$\min_{x,\,y} \|Ax + By - c\|_2^2 + \|Cx + Dy - d\|_2^2. \tag{21}$$

Для этой задачи можно выписать формулы явной минимизации по переменных x и y:

$$x = \operatorname*{argmin}_x \|Ax + By - c\|_2^2 + \|Cx + Dy - d\|_2^2 = (A^TA + C^TC)^{-1}\left[A^T(c - By) + C^T(d - Dy)\right],$$
$$y = \operatorname*{argmin}_y \|Ax + By - c\|_2^2 + \|Cx + Dy - d\|_2^2 = (B^TB + D^TD)^{-1}\left[B^T(c - Ax) + D^T(d - Cx)\right].$$

Поэтому задача (21) допускает альтернированную минимизацию

$$x^{k+1} = (A^TA + C^TC)^{-1}\left[A^T(c - By^k) + C^T(d - Dy^k)\right],$$
$$y^{k+1} = (B^TB + D^TD)^{-1}\left[B^T(c - Ax^k) + D^T(d - Cx^k)\right]$$

и может быть решена с помощью алгоритма 2.

Результаты сравнения алгоритмов AGMsDR и Alternating AGMsDR, запущенных с $\mu^{in} = 0$, представлены на рис. 2 и 3 для различных чисел обусловленности. κ, κ_1, κ_2 — числа обусловленности матриц W, A и D соответственно. По всей видимости, более быстрая сходимость связана с тем, что один из блоков (или оба блока) обусловлены лучше, чем вся задача.

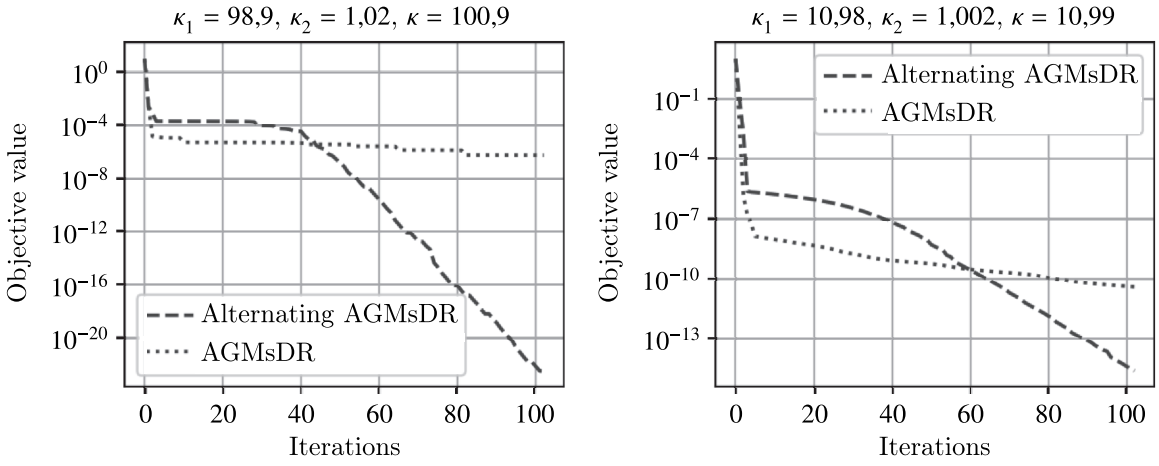

Рис. 2. Сходимость алгоритмов AGMsDR и Alternating AGMsDR, примененных к задачам (20) и (21) соответственно

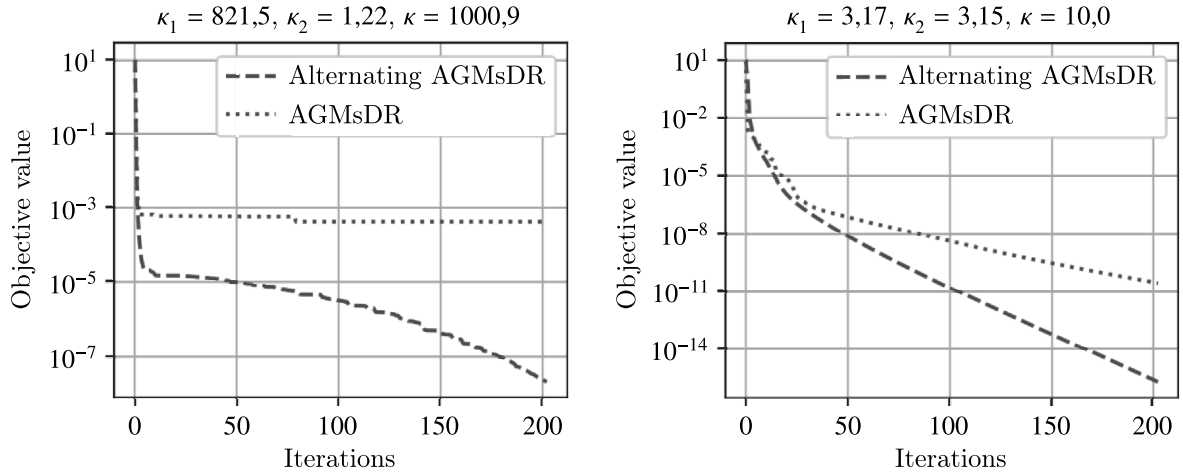

Рис. 3. Сходимость алгоритмов AGMsDR и Alternating AGMsDR, примененных к задачам (20) и (21) соответственно

Адаптивная каталист-оболочка

В работе [Ivanova et al., 2019] рассматривается так называемая адаптивная каталист-оболочка неускоренных методов, позволяющая ускорять эти методы. Обозначим:

$$F_{L,x}(y) = f(y) + \frac{L}{2}\|y - x\|_2^2,$$

где f — минимизируемая функция. Алгоритм ускорения некоторого метода $\mathcal{M}$ с помощью адаптивной каталист-оболочки представлен ниже (алгоритм 3).

Algorithm 3. Adaptive catalyst

Вход: Начальная точка x^0, оценка $M_0 > 0$, параметры $\alpha > \beta > \gamma > 0$ и метод $\mathcal{M}$.

1: Полагаем $y^0 = z^0 = x^0$
2: **for** $k = 0, 1, \ldots, N-1$ **do**
3: $M_{k+1} = \beta \cdot \min\left\{\alpha M_k, M_u\right\}$
4: $t = 0$
5: **repeat**
6: $t := t + 1$
7: $M_{k+1} := \max\left\{\frac{M_{k+1}}{\beta}, M_d\right\}$
8: Вычисляем:

$$a_{k+1} = \frac{\frac{1}{M_{k+1}} + \sqrt{\frac{1}{M_{k+1}^2} + \frac{4A_k}{M_{k+1}}}}{2},$$
$$A_{k+1} = A_k + a_{k+1},$$
$$x^{k+1} = \frac{A_k}{A_{k+1}} y^k + \frac{a_{k+1}}{A_{k+1}} z^k.$$

9: Вычисляем приближенное решение следующей задачи с помощью вспомогательного неускоренного метода $\mathcal{M}$:

$$y^{k+1} \approx \operatorname*{argmin}_y F_{M,x^{k+1}}(y).$$

Для этого, запуская из точки x^{k+1} и ожидая на выходе точку y^{k+1}, делаем N_t итераций метода $\mathcal{M}$ и проверяем адаптивный критерий остановки:

$$\|\nabla F_{M,x^{k+1}}(y^{k+1})\|_2 \leqslant \frac{M_{k+1}}{2}\|y^{k+1} - x^{k+1}\|_2. \tag{22}$$

10: **until** $t > 1$ and $N_t \geqslant \gamma \cdot N_{t-1}$ or $M_{k+1} = M_d$
11: $z^{k+1} = z^k - a_{k+1}\nabla f\left(y^{k+1}\right)$
12: **end for**
13: **Output:** y^N

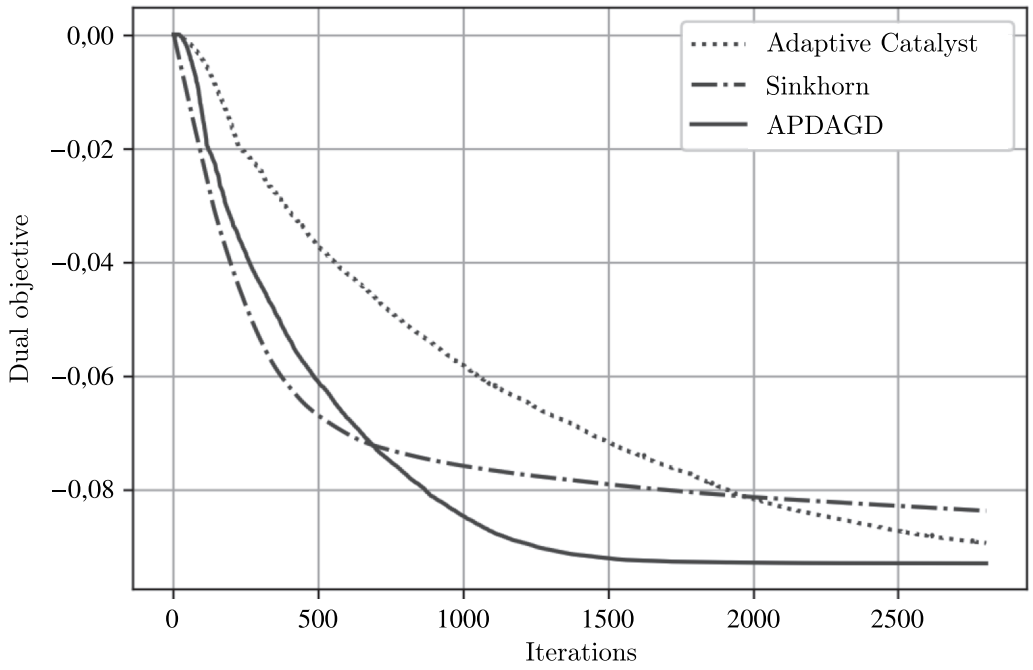

Рис. 4. Сходимость алгоритма Синхорна и градиентного метода, ускоренного с помощью адаптивной каталист-оболочки, примененной к ЭОТ

Данный алгоритм был представлен и исследован в работе [Hu, Koren, Volinsky, 2008]. Для алгоритма были проведены и оказались успешными численные эксперименты для задачи ALS.

В связи с этим было произведено сравнение ускоренного с помощью адаптивной каталист-оболочки градиентного метода с адаптивным выбором шага [Nesterov et al., 2014; Kamzolov, Dvurechensky, Gasnikov, 2020] для ЭОТ с алгоритмом Синхорна и методом APDAGD. Результаты представлены на рис. 4.

Заключение

Представлен специальный тип оценок сходимости для ранее известного метода AGMsDR из [Nesterov et al., 2020] и для обобщения этого метода на задачи, допускающие альтернированную минимизацию из [Guminov et al., 2019a]. А именно, сходимости со скоростью геометрической прогрессии в случае, вообще говоря, невыпуклых задач с M-липшицевым градиентом, для которых выполняется условие Поляка – Лоясиевича. При этом существенно, что значение параметра $\mu^{PL} > 0$ не используется при реализации метода.

Проведено экспериментальное сравнение AGMsDR и Alternating AGMsDR и выявлено расхождение с теоретическими оценками этих методов на примере задачи ЭОТ и задачи минимизации квадратичной функции. Гипотеза, объясняющая такое поведение, заключается в том, что в рассмотренных задачах реализуется специальный тип сходимости, а именно линейная сходимость на блоках переменных для метода Alternating AGMsDR, это требует более детального изучения.

Также проведены численные эксперименты для градиентного метода с адаптивным выбором шага, ускоренного с помощью адаптивной каталист-оболочки, для задачи ЭОТ и показана неоправданность применения такого подхода на практике для данной задачи.

Список литературы (References)

Beck A. On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes // SIAM Journal on Optimization. — 2015. — Vol. 25, No. 1. — P. 185–209. — https://doi.org/10.1137/13094829X

Bertsekas D. P., Tsitsiklis J. N. Parallel and distributed computation: numerical methods. — Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1989.

Chambolle A., Tan P., Vaiter S. Accelerated alternating descent methods for Dykstra-like problems // Journal of Mathematical Imaging and Vision. — 2017. — Vol. 59, No. 3. — P. 481–497. — https://doi.org/10.1007/s10851-017-0724-6

Cuturi M. Sinkhorn distances: lightspeed computation of optimal transport // Advances in Neural Information Processing Systems 26 / eds. C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Q. Weinberger. — Curran Associates, Inc., 2013. — P. 2292–2300. — http://papers.nips.cc/paper/4927-sinkhorn-distances-lightspeed-computation-of-optimal-transport.pdf

Diakonikolas J., Orecchia L. Alternating randomized block coordinate descent // Proceedings of the 35th International Conference on Machine Learning / eds. J. Dy, A. Krause. — PMLR, 2018. — Vol. 80. — P. 1224–1232. — http://proceedings.mlr.press/v80/diakonikolas18a/diakonikolas18a.pdf

Dvinskikh D., Gorbunov E., Gasnikov A., Dvurechensky P. Uribe C. A. On primal and dual approaches for distributed stochastic convex optimization over networks // IEEE 58th Conference on Decision and Control (CDC). — 2019. — P. 7435–7440. — https://doi.org/10.1109/CDC40024.2019.9029798

Dvurechensky P., Dvinskikh D., Gasnikov A., Uribe C. A., Nedić A. Decentralize and randomize: faster algorithm for Wasserstein barycenters // Proceedings of the 32th Conference on Neural Information Processing Systems / eds. S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett. — Curran Associates, Inc., 2018. — P. 10783–10793. — http://papers.nips.cc/paper/8274-decentralize-and-randomize-faster-algorithm-for-wasserstein-barycenters.pdf

Dvurechensky P., Gasnikov A., Kroshnin A. Computational optimal transport: complexity by accelerated gradient descent is better than by Sinkhorn's algorithm // Proceedings of the 35th International Conference on Machine Learning / eds. J. Dy, A. Krause. — PMLR, 2018. — Vol. 80. — P. 1367–1376. — http://proceedings.mlr.press/v80/dvurechensky18a/dvurechensky18a.pdf

Guminov S., Dvurechensky P., Tupitsa N., Gasnikov A. Accelerated alternating minimization, accelerated Sinkhorn's algorithm and accelerated iterative Bregman projections // arXiv preprint. — 2019a. — https://arxiv.org/pdf/1906.03622

Guminov S. V., Nesterov Yu. E., Dvurechensky P. E., Gasnikov A. V. Accelerated primal-dual gradient descent with linesearch for convex, nonconvex, and nonsmooth optimization problems // Doklady Mathematics. — 2019b. — Vol. 99, No. 2. — P. 125–128. — https://doi.org/10.1134/S1064562419020042

Hong M., Razaviyayn M., Luo Z., Pang J. A unified algorithmic framework for block-structured optimization involving Big Data: with applications in machine learning and signal processing // IEEE Signal Processing Magazine. — 2016. — Vol. 33, No. 1. — P. 57–77. — https://doi.org/10.1109/MSP.2015.2481563

Hong M., Wang X., Razaviyayn M., Luo Z.-Q. Iteration complexity analysis of block coordinate descent methods // Mathematical Programming. — 2017. — Vol. 163, No. 1. — P. 85–114. — https://doi.org/10.1007/s10107-016-1057-8

Hu Y., Koren Y., Volinsky C. Collaborative filtering for implicit feedback datasets // Eighth IEEE International Conference on Data Mining. — 2008. — P. 263–272. — https://doi.org/10.1109/ICDM.2008.22

Ivanova A., Pasechnyuk D., Grishchenko D., Shulgin E., Gasnikov A. Adaptive catalyst for smooth convex optimization // arXiv preprint. — 2019. — https://arxiv.org/pdf/1911.11271

Kamzolov D., Dvurechensky P., Gasnikov A. V. Universal intermediate gradient method for convex problems with inexact oracle // Optimization Methods and Software. — 2020. — P. 1–28. — https://doi.org/10.1080/10556788.2019.1711079

Kroshnin A., Tupitsa N., Dvinskikh D., Dvurechensky P., Gasnikov A., Uribe C. On the complexity of approximating Wasserstein barycenters // Proceedings of the 36th International Conference on Machine Learning / eds. K. Chaudhuri, R. Salakhutdinov. — PMLR, 2019. — Vol. 97. — P. 3530–3540. — http://proceedings.mlr.press/v97/kroshnin19a/kroshnin19a.pdf

Lin T., Ho N., Chen X., Cuturi M., Jordan M. I. Computational hardness and fast algorithm for fixed-support Wasserstein barycenter // arXiv preprint. — 2020. — https://arxiv.org/pdf/2002.04783

Lin T., Ho N., Cuturi M., Jordan M. I. On the complexity of approximating multimarginal optimal transport // arXiv preprint. — 2019. — https://arxiv.org/pdf/1910.00152

Luo Z.-Q., Tseng P. Error bounds and convergence analysis of feasible descent methods: a general approach // Annals of Operations Research. — 1993. — Vol. 46, No. 1. — P. 157–178. — https://doi.org/10.1007/BF02096261

Nesterov Yu. Universal gradient methods for convex optimization problems // Mathematical Programming. — 2014. — Vol. 152. — https://doi.org/10.1007/s10107-014-0790-0

Nesterov Yu., Gasnikov A., Guminov S., Dvurechensky P. Primal-dual accelerated gradient methods with small-dimensional relaxation oracle // Optimization Methods and Software. — 2020. — P. 1–28. — https://doi.org/10.1080/10556788.2020.1731747

Nutini J., Schmidt M., Laradji I., Friedlander M., Koepke H. Coordinate descent converges faster with the Gauss – Southwell rule than random selection // Proceedings of the 32nd International Conference on Machine Learning / eds. F. Bach, D. Blei. — PMLR, 2015. — Vol. 37. — P. 1632–1641. — http://proceedings.mlr.press/v37/nutini15.pdf

Ortega J. M., Rheinboldt W. C. Iterative solution of nonlinear equations in several variables. — Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2000.

Polyak B. T. Gradient methods for minimizing functionals // Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki. — 1963. — Vol. 3, No. 4. — P. 643–653.

Sun R., Hong M. Improved iteration complexity bounds of cyclic block coordinate descent for convex problems // Proceedings of the 28th International Conference on Neural Information Processing Systems. — Vol. 1. — Cambridge, MA, USA: MIT Press, 2015. — P. 1306–1314. — http://dl.acm.org/citation.cfm?id=2969239.2969385

Tupitsa N., Dvurechensky P., Gasnikov A., Uribe C A. Multimarginal optimal transport by accelerated alternating minimization // arXiv preprint. — 2020. — https://arxiv.org/pdf/2004.02294

Uribe C. A., Dvinskikh D., Dvurechensky P., Gasnikov A., Nedić A. Distributed computation of Wasserstein barycenters over networks // IEEE Conference on Decision and Control (CDC). — 2018. — P. 6544–6549. — https://doi.org/10.1109/CDC.2018.8619160