КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2023 Т. 15 № 2 С. 413–432
DOI: 10.20537/2076-7633-2023-15-2-413-432

Ки&М

МАТЕМАТИЧЕСКИЕ ОСНОВЫ И ЧИСЛЕННЫЕ МЕТОДЫ МОДЕЛИРОВАНИЯ

УДК: 519.85

Аналоги условия относительной сильной выпуклости для относительно гладких задач и адаптивные методы градиентного типа

Ф. С. Стонякин[1,2,a], О. С. Савчук[1,2,b], И. В. Баран[2,c], М. С. Алкуса[1,3,d], А. А. Титов[1,e]

[1]Московский физико-технический институт (национальный исследовательский университет),
Россия, 141701, Московская обл., г. Долгопрудный, Институтский пер., 9
[2]Крымский федеральный университет им. В. И. Вернадского,
Россия, 295007, Республика Крым, г. Симферополь, проспект академика Вернадского, 4
[3]Национальный исследовательский университет «Высшая школа экономики»,
Россия, 101000, г. Москва, ул. Мясницкая, д. 20

E-mail: [a] fedyor@mail.ru, [b] oleg.savchuk19@mail.ru, [c] matemain@mail.ru,
[d] mohammad.alkousa@phystech.edu, [e] a.a.tytov@gmail.com

Получено 19.02.2023.
Принято к публикации 23.02.2023.

Данная статья посвящена повышению скоростных гарантий численных методов градиентного типа для относительно гладких и относительно липшицевых задач минимизации в случае дополнительных предположений о некоторых аналогах сильной выпуклости целевой функции. Рассматриваются два класса задач: выпуклые задачи с условием относительного функционального роста, а также задачи (вообще говоря, невыпуклые) с аналогом условия градиентного доминирования Поляка – Лоясиевича относительно дивергенции Брэгмана. Для первого типа задач мы предлагаем две схемы рестартов методов градиентного типа и обосновываем теоретические оценки сходимости двух алгоритмов с адаптивно подбираемыми параметрами, соответствующими относительной гладкости или липшицевости целевой функции. Первый из этих алгоритмов проще в части критерия выхода из итерации, но для него близкие к оптимальным вычислительные гарантии обоснованы только на классе относительно липшицевых задач. Процедура рестартов другого алгоритма, в свою очередь, позволила получить более универсальные теоретические результаты. Доказана близкая к оптимальной оценка сложности на классе выпуклых относительно липшицевых задач с условием функционального роста, а для класса относительно гладких задач с условием функционального роста получены гарантии линейной скорости сходимости. На классе задач с предложенным аналогом условия градиентного доминирования относительно дивергенции Брэгмана были получены оценки качества выдаваемого решения с использованием адаптивно подбираемых параметров. Также мы приводим результаты некоторых вычислительных экспериментов, иллюстрирующих работу методов для второго исследуемого в настоящей статье подхода. В качестве примеров мы рассмотрели линейную обратную задачу Пуассона (минимизация дивергенции Кульбака – Лейблера), ее регуляризованный вариант, позволяющий гарантировать относительную сильную выпуклость целевой функции, а также некоторый пример относительно гладкой и относительно сильно выпуклой задачи. В частности, с помощью расчетов показано, что относительно сильно выпуклая функция может не удовлетворять введенному относительному варианту условия градиентного доминирования.

Ключевые слова: относительная сильная выпуклость, относительная гладкость, относительный функциональный рост, относительное условие градиентного доминирования, адаптивный метод, рестарты

Работа выполнена при поддержке гранта Российского научного фонда и города Москвы № 22-21-20065 (https://rscf.ru/project/22-21-20065/).

COMPUTER RESEARCH AND MODELING
2023 VOL. 15 NO. 2 P. 413–432
DOI: 10.20537/2076-7633-2023-15-2-413-432

Ки&М

MATHEMATICAL MODELING AND NUMERICAL SIMULATION

UDC: 519.85

Analogues of the relative strong convexity condition for relatively smooth problems and adaptive gradient-type methods

F. S. Stonyakin[1,2,a], O. S. Savchuk[1,2,b], I. V. Baran[2,c], M. S. Alkousa[1,3,d], A. A. Titov[1,e]

[1]Moscow Institute of Physics and Technology,
9 Institutskiy per., Dolgoprudny, Moscow region, 141701, Russia
[2]V. I. Vernadsky Crimean Federal University,
4 Academician Vernadsky Avenue, Simferopol, Republic of Crimea, 295007, Russia
[3]HSE University,
20 Myasnitskaya st., Moscow, 101000, Russia

E-mail: [a] fedyor@mail.ru, [b] oleg.savchuk19@mail.ru, [c] matemain@mail.ru,
[d] mohammad.alkousa@phystech.edu, [e] a.a.tytov@gmail.com

Received 19.02.2023.
Accepted for publication 23.02.2023.

This paper is devoted to some variants of improving the convergence rate guarantees of the gradient-type algorithms for relatively smooth and relatively Lipschitz-continuous problems in the case of additional information about some analogues of the strong convexity of the objective function. We consider two classes of problems, namely, convex problems with a relative functional growth condition, and problems (generally, non-convex) with an analogue of the Polyak – Lojasiewicz gradient dominance condition with respect to Bregman divergence. For the first type of problems, we propose two restart schemes for the gradient type methods and justify theoretical estimates of the convergence of two algorithms with adaptively chosen parameters corresponding to the relative smoothness or Lipschitz property of the objective function. The first of these algorithms is simpler in terms of the stopping criterion from the iteration, but for this algorithm, the near-optimal computational guarantees are justified only on the class of relatively Lipschitz-continuous problems. The restart procedure of another algorithm, in its turn, allowed us to obtain more universal theoretical results. We proved a near-optimal estimate of the complexity on the class of convex relatively Lipschitz continuous problems with a functional growth condition. We also obtained linear convergence rate guarantees on the class of relatively smooth problems with a functional growth condition. For a class of problems with an analogue of the gradient dominance condition with respect to the Bregman divergence, estimates of the quality of the output solution were obtained using adaptively selected parameters. We also present the results of some computational experiments illustrating the performance of the methods for the second approach at the conclusion of the paper. As examples, we considered a linear inverse Poisson problem (minimizing the Kullback – Leibler divergence), its regularized version which allows guaranteeing a relative strong convexity of the objective function, as well as an example of a relatively smooth and relatively strongly convex problem. In particular, calculations show that a relatively strongly convex function may not satisfy the relative variant of the gradient dominance condition.

Keywords: relative strong convexity, relative smoothness, relative functional growth, adaptive method, restarts

Citation: *Computer Research and Modeling*, 2023, vol. 15, no. 2, pp. 413–432 (Russian).

This work was supported by Russian Science Foundation and Moscow city, project 22-21-20065 (https://rscf.ru/project/22-21-20065/).

1. Введение

Численные методы оптимизации играют ключевую роль в решении многих прикладных задач в пространствах больших размерностей. При этом во многих из них целевая функция не может считаться достаточно гладкой в стандартном смысле, чтобы гарантировать достаточно хорошую скорость сходимости вычислительных процедур. На сегодняшний день существует ряд подходов к обобщению стандартного свойства гладкости (условия Липшица градиента), которые всё же позволяют гарантировать неплохие оценки скорости сходимости численных методов для задач минимизации функций. Среди них можно выделить глобальные свойства непрерывности и гладкости относительно некоторой дивергенции Брэгмана: условие относительной гладкости [Bauschke, Bolte, Teboulle, 2017; Lu, Freund, Nesterov, 2018], а также условие относительной липшицевости (непрерывности) в [Nesterov, 2019; Lu, 2019]. Таким свойствам удовлетворяет ряд важных прикладных задач [Bauschke, Bolte, Teboulle, 2017; Lu, Freund, Nesterov, 2018; Lu, 2019]. Ключевая идея данных подходов — «дружественность» целевой функции по отношению к некоторой выпуклой (но не обязательно сильно выпуклой) функции, генерирующей расстояние на заданном пространстве (прокс-функция). Напомним необходимые вспомогательные понятия.

Пусть $(E, \|\cdot\|)$ — некоторое нормированное конечномерное векторное пространство, а E^* — его сопряженное с нормой:

$$\|y\|_* = \max_x\{\langle y,\ x\rangle,\ \|x\| \leqslant 1\},$$

где $\langle y,\ x\rangle$ — значение непрерывного линейного функционала y в точке $x \in E$.

Рассмотрим некоторое замкнутое выпуклое подмножество $Q \subset E$ и непрерывно дифференцируемую выпуклую функцию $d\colon Q \to \mathbb{R}$, порождающую расстояние на множестве Q. Напомним следующее важное понятие дивергенции (расстояния) Брэгмана, которое можно понимать как обобщенное расстояние между точками на множестве Q:

$$V(y,\ x) = d(y) - d(x) - \langle \nabla d(x),\ y - x\rangle. \tag{1}$$

Напомним также определения L-относительной гладкости (см. [Bauschke, Bolte, Teboulle, 2017; Stonyakin et al., 2021b]) и M-относительной липшицевости (см. [Lu, 2019; Nesterov, 2019]).

Определение 1. Выпуклая функция $f\colon Q \to \mathbb{R}$ называется *L-относительно гладкой* при некотором $L > 0$, если выполняется следующее неравенство:

$$f(y) \leqslant f(x) + \langle \nabla f(x),\ y - x\rangle + LV(y,\ x) \quad \forall x,\ y \in Q \tag{2}$$

для произвольного субградиента ∇f функции f.

Определение 2. Выпуклая функция $f\colon Q \to \mathbb{R}$ называется *M-относительно липшицевой* при некотором $M > 0$, если выполняется следующее неравенство:

$$\langle \nabla f(x),\ y - x\rangle + M\sqrt{2V(y,\ x)} \geqslant 0 \quad \forall x,\ y \in Q \tag{3}$$

для произвольного субградиента ∇f функции f.

На классе относительно гладких и относительно сильно выпуклых задач известны результаты о сходимости методов градиентного типа со скоростью геометрической прогрессии [Lu, Freund, Nesterov, 2018; Stonyakin et al., 2021b]. В [Stonyakin et al., 2021a] предложены и методы с адаптивной настройкой параметров, соответствующих относительной гладкости и липшицевости целевой функции задачи.

В настоящей статье мы рассмотрим аналоги относительной сильной выпуклости (в том числе, возможно, и для некоторых типов невыпуклых задач), которые также позволят получить соответствующие скоростные гарантии, в том числе и для адаптивных методов.

Первая идея настоящей статьи — использование следующего условия κ-относительного функционального роста, предложенного в [Gutman, Pena, 2018].

Определение 3. Говорят, что выпуклая функция $f\colon Q \to \mathbb{R}$ удовлетворяет условию κ-относительного функционального роста, если для всякого $x \in Q$ верно неравенство

$$f(x) - f(x_*) \geqslant \kappa V(x_*, x) \tag{4}$$

при некотором фиксированном $\kappa > 0$.

В обозначении $V(x_*, x)$ здесь и всюду далее предполагается, что x_* — ближайшее к x точное решение задачи минимизации f на множестве Q. Предположение κ-относительного функционального роста целевой функции позволяет предложить схему рестартов некоторых алгоритмов из [Stonyakin et al., 2021a] с гарантированной оценкой сложности (достаточным для достижения ε-приближенного решения по аргументу (с точки зрения дивергенции Брэгмана) количеством обращений к подпрограмме для нахождения (суб)градиента:

$$O\left(\frac{M^2}{\kappa\varepsilon}\log_2\frac{2\kappa R^2}{\varepsilon}\right)$$

на классе выпуклых M-относительно липшицевых задач и

$$O\left(\frac{L}{\kappa}\log_2\frac{2\kappa R^2}{\varepsilon}\right)$$

на классе выпуклых L-относительно гладких задач. В данном случае условие κ-относительного функционального роста может пониматься как обобщение классической сильной выпуклости функции. Стоит отметить, что такое предположение позволяет рассмотреть некоторые важные прикладные задачи, а именно проблему бинарной классификации методом опорных векторов (SVM) и задачу об отыскании общей точки эллипсоидов (IEP) [Lu, 2019]. Действительно, достаточно рассмотреть основную лемму 5.1 из [Lu, 2019], гарантирующую выполнение условия относительной липшицевости для целевых функций в задачах SVM и IEP за счет соответствующего подбора прокс-функций и выполнения следующих неравенств:

$$d(x) = \frac{1}{3}\|x\|_2^3, \quad V(x_*, x) \leqslant \frac{1}{3}\|x_* - x\|_2^2\left(\|x_*\|_2 + 2\|x\|_2\right) \leqslant \frac{1}{\kappa}(f(x) - f(x_*)),$$
$$d(x) = \frac{1}{4}\|x\|_2^4, \quad V(x_*, x) \leqslant \frac{1}{4}\|x_* - x\|_2^2\left(\|x_* + x\|_2^2 + 2\|x\|_2^2\right) \leqslant \frac{1}{\kappa}(f(x) - f(x_*)),$$

в случае, например, если множество Q ограничено для евклидовой нормы $\|\cdot\|_2$.

Второй подход к повышению скорости сходимости методов градиентного типа на классе относительно гладких задач, рассмотренный в настоящей статье, основан на следующем аналоге известного условия Поляка – Лоясиевича [Поляк, 1963; Karimi, Nutini, Schmidt, 2016; Belkin, 2021] относительно дивергенции Брэгмана:

$$f(x) - f^* \leqslant \frac{p^2}{\mu}V(x, x_p) \quad \forall x \in Q \tag{5}$$

для некоторого фиксированного $\mu > 0$ и всякого $p \geqslant \mu$, которое, вообще говоря, гарантирует линейную скорость сходимости

$$f(x_{k+1}) - f^* \leqslant \left(1 - \frac{\mu}{L}\right)^{k+1}\left(f(x_0) - f^*\right)$$

даже для невыпуклых задач оптимизации, причем x_p определяет шаг метода зеркального спуска:

$$x_p := \operatorname*{argmin}_{y\in\mathbb{R}^n}\{\langle\nabla f(x), y - x\rangle + pV(y, x)\}.$$

Работа состоит из введения, заключения и трех основных параграфов.

Во втором параграфе рассматриваются выпуклые задачи с условием относительного функционального роста. Для этого типа задач предложены две схемы рестартов адаптивных градиентых алгоритмов 3 и 6 из [Stonyakin et al., 2021a] и доказаны теоретические оценки сходимости соответствующих алгоритмов с адаптивно подбираемыми параметрами, соответствующими относительной гладкости или липшицевости целевой функции. Для относительно гладких задач скорость сходимости гарантированно будет линейной. Доказана близкая к оптимальной оценка сложности на классе выпуклых относительно липшицевых задач с условием функционального роста.

В третьем параграфе исследованы задачи (вообще говоря, невыпуклые) с аналогом условия градиентного доминирования Поляка – Лоясиевича относительно дивергенции Брэгмана. На этом классе задач с аналогом условия градиентного доминирования относительно дивергенции Брэгмана получены оценки качества выдаваемого решения с использованием адаптивно подбираемых параметров.

В четвертом параграфе приведены результаты проведенных численных экспериментов для иллюстрации полученных результатов. В качестве примеров мы рассмотрели линейную обратную задачу Пуассона (минимизация дивергенции Кульбака – Лейблера), ее регуляризованный вариант, позволяющий гарантировать относительную сильную выпуклость целевой функции, а также пример относительно гладкой и относительно сильно выпуклой задачи, рассмотренной в подпараграфе 2.1 [Lu, Freund, Nesterov, 2018]. Полученные результаты показывают, что относительно сильно выпуклая функция может не удовлетворять относительному варианту условия градиентного доминирования.

2. Рестарты адаптивных методов градиентного типа для задач с относительным функциональным ростом

В данном параграфе предлагаются процедуры рестартов алгоритмов 3 и 6 из [Stonyakin et al., 2021a] для выпуклых задач в предположении, что целевой функционал удовлетворяет условию κ-относительного функционального роста при некотором $\kappa > 0$. Начнем со схемы рестартов алгоритма 3 из [Stonyakin et al., 2021a] (алгоритм 1 приведен ниже) с более простым критерием выхода из итерации, для которого удалось получить близкую к оптимальной оценку скорости сходимости на классе относительно сильно выпуклых и относительно липшицевых задач.

Algorithm 1. Адаптивный алгоритм для относительно липшицевых задач оптимизации

Require: $\varepsilon > 0$, x_0, $L_0 > 0$, R, удовл. $V(x_*, x_0) \leqslant R^2$.

1: $k = k + 1$, $L_{k+1} = \frac{L_k}{2}$.

2: Вычисляем

$$x_{k+1} = \operatorname*{argmin}_{x \in Q} \{\langle \nabla f(x_k), x\rangle + L_{k+1} V(x, x_k)\}. \tag{6}$$

3: **if**

$$0 \leqslant \langle \nabla f(x_k), x_{k+1} - x_k\rangle + L_{k+1} V(x_{k+1}, x_k) + \frac{\varepsilon}{2}, \tag{7}$$

then переходим к следующей итерации (п. 1).

4: **else**

$L_{k+1} = 2 \cdot L_{k+1}$ и переходим к п. 2.

5: **end if**

Ensure: $\widehat{x} = \frac{1}{S_N} \sum\limits_{k=0}^{N-1} \frac{x_{k+1}}{L_{k+1}}$.

Algorithm 2. Рестарты алгоритма 1 для относительно липшицевых задач оптимизации при условии κ-относительного функционального роста

Require: $\varepsilon > 0$, $\kappa > 0$, x_0, $L_0 > 0$, R, удовл. $V(x_*, x_0) \leqslant R^2$.
1: Set $p = 0$.
2: **repeat**
3: Set $\widehat{x}^{p-1}$ — результат работы алгоритма 1 с критерием остановки $S_{N_p} = \sum\limits_{k=0}^{N_p-1} \frac{1}{L_{k+1}} \geqslant \frac{2}{\kappa}$, где N_p — количество итераций на данном рестарте метода.
4: Set $x_0 = \widehat{x}^{p-1}_{\min}$.
5: Set $p = p + 1$.
6: **until** $p \geqslant \log_2 \frac{\kappa R^2}{\varepsilon}$.
Ensure: $\widehat{x}^{p-1}$.

Теорема 1. *Пусть $f\colon Q \to \mathbb{R}$ — выпуклая и M-относительно липшицева функция* (*см.* (3)) *и удовлетворяет условию κ-относительного функционального роста* (*см.* (4)), *а также $L_0 \leqslant \frac{2M^2}{\varepsilon}$. В таком случае после $p = \left\lceil \log_2 \frac{\kappa R^2}{\varepsilon} \right\rceil$ шагов алгоритма* 2 *будет выполняться следующее неравенство*:

$$V\left(x_*, \widehat{x}^{p-1}\right) \leqslant \frac{2\varepsilon}{\kappa}.$$

Для этого потребуется не более чем

$$N = O\left(\frac{M^2}{\kappa\varepsilon} \log_2 \frac{\kappa R^2}{\varepsilon}\right)$$

итераций алгоритма 1.

Доказательство. Мы отправляемся от полученной для алгоритма 1 в [Stonyakin et al., 2021a] следующей оценки для невязки по функции:

$$f(\widehat{x}) - f(x_*) \leqslant \frac{V(x_*, x_0)}{S_N} + \frac{\varepsilon}{2} \leqslant \frac{R^2}{S_N} + \frac{\varepsilon}{2}. \tag{8}$$

С учетом (4) и (8) получаем

$$\kappa V(x_*, \widehat{x}) \leqslant \frac{V(x_*, x_0)}{S_N} + \frac{\varepsilon}{2} \leqslant \frac{R^2}{S_N} + \frac{\varepsilon}{2},$$

или

$$V(x_*, \widehat{x}) \leqslant \frac{V(x_*, x_0)}{\kappa S_N} + \frac{\varepsilon}{2\kappa} \leqslant \frac{R^2}{\kappa S_N} + \frac{\varepsilon}{2\kappa}.$$

Потребуем теперь, чтобы при некотором достаточно большом N выполнялось условие

$$V(x_*, \widehat{x}) \leqslant \frac{1}{2} V(x_*, x_0) + \frac{\varepsilon}{2\kappa}, \tag{9}$$

откуда

$$\frac{1}{\kappa S_N} \leqslant \frac{1}{2}, \quad \kappa S_N \geqslant 2, \quad S_N \geqslant \frac{2}{\kappa}.$$

Поскольку функция f M-относительно липшицева, то

$$f(x) - f(y) \leqslant \langle \nabla f(x), x - y \rangle \leqslant M\sqrt{2V(y, x)} \leqslant \frac{M^2 V(y, x)}{\varepsilon} + \frac{\varepsilon}{2}. \tag{10}$$

Отсюда следует, что критерий выхода из итерации (7) будет гарантированно выполнен при $L_{k+1} \geqslant \frac{M^2}{\varepsilon}$. Далее, поскольку выход из итерации произойдет при $L_{k+1} \leqslant \frac{2M^2}{\varepsilon}$, имеем

$$S_N = \sum_{k=0}^{N-1} \frac{1}{L_{k+1}} \geqslant \frac{N\varepsilon}{2M^2},$$

откуда следует, что

$$\frac{V(x_*,\, x_0)}{\kappa S_N} \leqslant \frac{2M^2 V(x_*,\, x_0)}{N\kappa\varepsilon} \leqslant \frac{1}{2} V(x_*,\, x_0).$$

Таким образом, получаем, что для гарантированного достижения требования (9) количество итераций алгоритма 1 не превышает

$$N = \left\lceil \frac{4M^2}{\kappa\varepsilon} \right\rceil.$$

Теперь реализуем для алгоритма 1 предложенную схему рестартов (перезапусков). Итак, после N шагов алгоритма 1 имеем

$$V\left(x_*,\, \widehat{x}^0\right) \leqslant \frac{1}{2} V(x_*,\, x_0) + \frac{\varepsilon}{2\kappa}. \tag{11}$$

Положим $x^1 := \widehat{x}^0$ и запускаем алгоритм 1 из точки x^1 вместо x_0. После N шагов из (11) получим

$$V\left(x_*,\, \widehat{x}^1\right) \leqslant \frac{1}{2} V\left(x_*,\, x^1\right) + \frac{\varepsilon}{2\kappa} \leqslant \left(\frac{1}{2}\right)^2 V(x_*,\, x_0) + \frac{\varepsilon}{4\kappa} + \frac{\varepsilon}{2\kappa} = \left(\frac{1}{2}\right)^2 V(x_*,\, x_0) + \frac{3\varepsilon}{4\kappa}.$$

Далее, $x^2 := \widehat{x}^1$, и повторим процесс:

$$V\left(x_*,\, \widehat{x}^2\right) \leqslant \frac{1}{2} V\left(x_*,\, x^2\right) + \frac{\varepsilon}{2\kappa} \leqslant \left(\frac{1}{2}\right)^2 V\left(x_*,\, x^1\right) + \frac{3\varepsilon}{4\kappa} \leqslant$$
$$\leqslant \left(\frac{1}{2}\right)^3 V(x_*,\, x_0) + \frac{\varepsilon}{8\kappa} + \frac{3\varepsilon}{4\kappa} = \left(\frac{1}{2}\right)^3 V(x_*,\, x_0) + \frac{7\varepsilon}{8\kappa}.$$

Аналогично: $x^3 := \widehat{x}^2$, и снова запускаем алгоритм 1 из новой точки:

$$V\left(x_*,\, \widehat{x}^3\right) \leqslant \frac{1}{2} V\left(x_*,\, x^3\right) + \frac{\varepsilon}{2\kappa} \leqslant \left(\frac{1}{2}\right)^2 V\left(x_*,\, x^2\right) + \frac{3\varepsilon}{4\kappa} \leqslant \left(\frac{1}{2}\right)^3 V\left(x_*,\, x^1\right) + \frac{7\varepsilon}{8\kappa} \leqslant$$
$$\leqslant \left(\frac{1}{2}\right)^4 V(x_*,\, x_0) + \frac{\varepsilon}{16\kappa} + \frac{7\varepsilon}{8\kappa} = \left(\frac{1}{2}\right)^4 V(x_*,\, x_0) + \frac{15\varepsilon}{16\kappa}.$$

Как видим, в процессе перезапусков алгоритма 1 возникает убывающая геометрическая прогрессия со знаменателем $q = \frac{1}{2}$ и первым членом $b_1 = \frac{\varepsilon}{2\kappa}$. Найдем ее сумму:

$$S = \frac{b_1}{1-q} = \frac{\frac{\varepsilon}{2\kappa}}{1-\frac{1}{2}} = \frac{\varepsilon}{\kappa}. \tag{12}$$

Таким образом, мы получили систему последовательных перезапусков (рестартов) алгоритма 1. После p перезапусков с учетом (12) получим

$$V\left(x_*,\, \widehat{x}^{p-1}\right) \leqslant \left(\frac{1}{2}\right)^p V(x_*,\, x_0) + \frac{(2^p-1)\,\varepsilon}{2^p\kappa} \leqslant \left(\frac{1}{2}\right)^p R^2 + \frac{(2^p-1)\,\varepsilon}{2^p\kappa} \leqslant \left(\frac{1}{2}\right)^p R^2 + \frac{\varepsilon}{\kappa}. \tag{13}$$

Из (13) следует, что $V\left(x_*, \widehat{x}^{p-1}\right) \leqslant \frac{2\varepsilon}{\kappa}$ гарантированно при $\left(\frac{1}{2}\right)^p R^2 \leqslant \frac{\varepsilon}{\kappa}$, откуда $2^p \geqslant \frac{\kappa R^2}{\varepsilon}$ и $p \geqslant \log_2 \frac{\kappa R^2}{\varepsilon}$. А значит, $p = \left\lceil \log_2 \frac{\kappa R^2}{\varepsilon} \right\rceil$. Таким образом, итоговая сложность примененной схемы

$$N \cdot p = \left\lceil \frac{4M^2}{\kappa\varepsilon} \right\rceil \cdot \left\lceil \log_2 \frac{\kappa R^2}{\varepsilon} \right\rceil = O\left(\frac{M^2}{\kappa\varepsilon} \log_2 \frac{\kappa R^2}{\varepsilon} \right).$$

□

Замечание 1. Отметим, что с учетом относительной липшицевости функции f после $p \geqslant \geqslant \log_2 \frac{\kappa R^2}{\varepsilon}$ шагов алгоритма 2 будет иметь место следующая оценка для невязки по функции:

$$f\left(\widehat{x}^{p-1}\right) - f(x_*) \leqslant M\sqrt{\frac{4\varepsilon}{\kappa}}.$$

Действительно, с учетом неравенства (10) получаем

$$f\left(\widehat{x}^{p-1}\right) - f(x_*) \leqslant \left\langle \nabla f\left(\widehat{x}^{p-1}\right), \widehat{x}^{p-1} - x_* \right\rangle \leqslant M\sqrt{2V\left(x_*, \widehat{x}^{p-1}\right)} \leqslant M\sqrt{\frac{4\varepsilon}{\kappa}}.$$

Теперь перейдем к схеме рестартов алгоритма 6 из [Stonyakin et al., 2021a] (алгоритм 3 приведен ниже), для которого за счет предложенного критерия выхода из итерации возможно получить оценки скорости сходимости не только для относительно липшицевых, но и для относительно гладких задач.

Algorithm 3. Универсальный метод для относительно гладких и относительно липшицевых задач выпуклой оптимизации

Require: $\varepsilon > 0$, x_0, $L_0 > 0$, R, удовл. $V(x_*, x_0) \leqslant R^2$.

1: $k = k + 1$, $L_{k+1} = \frac{L_k}{2}$.

2: Вычисляем

$$x_{k+1} = \operatorname*{argmin}_{x \in Q}\{\langle \nabla f(x_k), x\rangle + L_{k+1} V(x, x_k)\}. \tag{14}$$

3: **If**

$$f(x_{k+1}) \leqslant f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + L_{k+1} V(x_{k+1}, x_k) + \frac{3\varepsilon}{4}, \tag{15}$$

then переходим к следующей итерации (п. 1).

4: **else**

set $L_{k+1} = 2 \cdot L_{k+1}$ и переходим к п. 2.

5: **end if**

Ensure: $\widehat{x} = \frac{1}{S_N} \sum\limits_{k=0}^{N-1} \frac{x_{k+1}}{L_{k+1}}$.

Теорема 2. *Пусть $f\colon Q \to \mathbb{R}$ — выпуклая функция и удовлетворяет условию κ-относительного функционального роста* (*см.* (4)). *Если f M-относительно липшицева функция и $L_0 \leqslant \frac{2M^2}{\varepsilon}$, то после $p = \left\lceil \log_2 \frac{2\kappa R^2}{\varepsilon} \right\rceil$ шагов алгоритма* 4 *неравенство*

$$V\left(x_*, \widehat{x}^{p-1}\right) \leqslant \frac{2\varepsilon}{\kappa} \tag{16}$$

будет выполняться после не более чем

$$N = O\left(\frac{M^2}{\kappa\varepsilon} \log_2 \frac{2\kappa R^2}{\varepsilon} \right)$$

итераций алгоритма 3.

Algorithm 4. Рестарты универсального алгоритма 3 для относительно гладких и относительно липшицевых задач выпуклой оптимизации при условии κ-относительного функционального роста

Require: $\varepsilon > 0$, $\kappa > 0$, x_0, $L_0 > 0$, R, удовл. $V(x_*, x_0) \leqslant R^2$.
1: Set $p = 0$.
2: **repeat**
3: Set $\widehat{x}^{p-1}$ — результат работы алгоритма 3 с критерием остановки $S_{N_p} = \sum\limits_{k=0}^{N_p-1} \frac{1}{L_{k+1}} \geqslant \frac{2}{\kappa}$, где N_p — количество итераций на данном рестарте метода.
4: Set $x_0 = \widehat{x}^{p-1}_{\min}$.
5: Set $p = p + 1$.
6: **until** $p \geqslant \log_2 \frac{2\kappa R^2}{\varepsilon}$.
Ensure: $\widehat{x}^{p-1}$.

Если f L-относительно гладкая функция и $L_0 \leqslant 2L$, то после $p = \left\lceil \log_2 \frac{2\kappa R^2}{\varepsilon} \right\rceil$ шагов алгоритма 4 *для выполнения неравенства* (16) *нам потребуется не более*

$$N = O\left(\frac{L}{\kappa} \log_2 \frac{2\kappa R^2}{\varepsilon}\right)$$

итераций алгоритма 3.

Доказательство. С учетом оценки для невязки по функции, полученной в [Stonyakin et al., 2021a] для алгоритма 3:

$$f(\widehat{x}) - f(x_*) \leqslant \frac{V(x_*, x_0)}{S_N} + \frac{3\varepsilon}{4} \leqslant \frac{R^2}{S_N} + \frac{3\varepsilon}{4}, \tag{17}$$

а также условия относительного κ-функционального роста (4), имеем

$$\kappa V(x_*, \widehat{x}) \leqslant \frac{V(x_*, x_0)}{S_N} + \frac{3\varepsilon}{4} \leqslant \frac{R^2}{S_N} + \frac{3\varepsilon}{4},$$

откуда

$$V(x_*, \widehat{x}) \leqslant \frac{V(x_*, x_0)}{\kappa S_N} + \frac{3\varepsilon}{4\kappa} \leqslant \frac{R^2}{\kappa S_N} + \frac{3\varepsilon}{4\kappa}.$$

Потребуем теперь, чтобы при некотором N было верно неравенство

$$V(x_*, \widehat{x}) \leqslant \frac{1}{2} V(x_*, x_0) + \frac{3\varepsilon}{4\kappa}. \tag{18}$$

Из (18) получаем, что

$$\frac{1}{\kappa S_N} \leqslant \frac{1}{2}, \quad \kappa S_N \geqslant 2, \quad S_N \geqslant \frac{2}{\kappa}.$$

Также в предположении об M-относительной липшицевости целевой функции в [Stonyakin et al., 2021a] было получено следующее неравенство:

$$f(x_{k+1}) \leqslant f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{2M^2}{\varepsilon} V(x_{k+1}, x_k) + \frac{3\varepsilon}{4}, \tag{19}$$

из которого следует, что критерий выхода из итерации (15) гарантированно выполнится при $L_{k+1} \geqslant \frac{2M^2}{\varepsilon}$. Далее, поскольку выход из итерации обязательно произойдет при $L_{k+1} \leqslant \frac{4M^2}{\varepsilon}$, то

$$S_N = \sum_{k=0}^{N-1} \frac{1}{L_{k+1}} \geqslant \frac{N\varepsilon}{4M^2}$$

и

$$\frac{V(x_*, x_0)}{\kappa S_N} \leqslant \frac{4M^2 V(x_*, x_0)}{N\kappa\varepsilon} \leqslant \frac{1}{2}V(x_*, x_0),$$

т. е. гарантированное для достижения (18) количество итераций алгоритма 3 не превосходит

$$N = \left\lceil \frac{8M^2}{\kappa\varepsilon} \right\rceil.$$

Если f — L-относительно гладкая функция, то критерий выхода из итерации (15) заведомо выполнится при $L_{k+1} \geqslant L$. Выход из итерации гарантированно произойдет для $L_{k+1} \leqslant 2L$, откуда

$$S_N \geqslant \frac{N}{2L} \quad \text{и} \quad \frac{V(x_*, x_0)}{\kappa S_N} \leqslant \frac{2LV(x_*, x_0)}{\kappa N} \leqslant \frac{1}{2}V(x_*, x_0).$$

Из последних неравенств следует, что

$$N \geqslant \frac{4L}{\kappa}, \quad \text{т. е. } N = \left\lceil \frac{4L}{\kappa} \right\rceil.$$

Применим схему рестартов для алгоритма 3. После N шагов алгоритма 3 имеем

$$V\left(x_*, \widehat{x}^0\right) \leqslant \frac{1}{2}V(x_*, x_0) + \frac{3\varepsilon}{4\kappa}. \tag{20}$$

Положим $x^1 := \widehat{x}^0$ и запустим алгоритм 3 из точки x^1. После N шагов из (20) получим

$$V\left(x_*, \widehat{x}^1\right) \leqslant \frac{1}{2}V\left(x_*, x^1\right) + \frac{3\varepsilon}{4\kappa} \leqslant \left(\frac{1}{2}\right)^2 V(x_*, x_0) + \frac{3\varepsilon}{8\kappa} + \frac{3\varepsilon}{4\kappa} = \left(\frac{1}{2}\right)^2 V(x_*, x_0) + \frac{9\varepsilon}{8\kappa}.$$

Далее, $x^2 := \widehat{x}^1$, и перезапускаем алгоритм 3 из указанной точки. Снова после N шагов алгоритма 3 будем иметь

$$V\left(x_*, \widehat{x}^2\right) \leqslant \frac{1}{2}V\left(x_*, x^2\right) + \frac{3\varepsilon}{4\kappa} \leqslant \left(\frac{1}{2}\right)^2 V\left(x_*, x^1\right) + \frac{9\varepsilon}{8\kappa} \leqslant$$
$$\leqslant \left(\frac{1}{2}\right)^3 V(x_*, x_0) + \frac{3\varepsilon}{16\kappa} + \frac{9\varepsilon}{8\kappa} = \left(\frac{1}{2}\right)^3 V(x_*, x_0) + \frac{21\varepsilon}{16\kappa}.$$

Аналогично: $x^3 := \widehat{x}^2$, и продолжаем процесс:

$$V\left(x_*, \widehat{x}^3\right) \leqslant \frac{1}{2}V\left(x_*, x^3\right) + \frac{3\varepsilon}{4\kappa} \leqslant \left(\frac{1}{2}\right)^2 V\left(x_*, x^2\right) + \frac{9\varepsilon}{8\kappa} \leqslant \left(\frac{1}{2}\right)^3 V\left(x_*, x^1\right) + \frac{21\varepsilon}{16\kappa} \leqslant$$
$$\leqslant \left(\frac{1}{2}\right)^4 V(x_*, x_0) + \frac{3\varepsilon}{32\kappa} + \frac{21\varepsilon}{16\kappa} = \left(\frac{1}{2}\right)^4 V(x_*, x_0) + \frac{45\varepsilon}{32\kappa}.$$

Найдем сумму бесконечно убывающей геометрической прогрессии со знаменателем $q = \frac{1}{2}$ и первым членом $b_1 = \frac{3\varepsilon}{4\kappa}$:

$$S = \frac{b_1}{1-q} = \frac{\frac{3\varepsilon}{4\kappa}}{1-\frac{1}{2}} = \frac{3\varepsilon}{2\kappa}.$$

Таким образом, получена система последовательных перезапусков алгоритма 3. После p перезапусков будем иметь

$$V\left(x_*, \widehat{x}^{p-1}\right) \leqslant \left(\frac{1}{2}\right)^p V(x_*, x_0) + \frac{3\varepsilon}{2^{p+1}\kappa} + \frac{3\varepsilon}{2^p\kappa} + \ldots + \frac{3\varepsilon}{4\kappa} \leqslant \left(\frac{1}{2}\right)^p V(x_*, x_0) + \frac{3\varepsilon}{2\kappa} \leqslant \left(\frac{1}{2}\right)^p R^2 + \frac{3\varepsilon}{2\kappa}.$$

Отсюда следует, что $V\left(x_*, \widehat{x}^{p-1}\right) \leqslant \frac{2\varepsilon}{\kappa}$ гарантированно при $\left(\frac{1}{2}\right)^p R^2 \leqslant \frac{\varepsilon}{2\kappa}$, т. е. $2^p \geqslant \frac{2\kappa R^2}{\varepsilon}$ и $p \geqslant$ $\geqslant \log_2 \frac{2\kappa R^2}{\varepsilon}$. Таким образом,

$$p = \left\lceil \log_2 \frac{2\kappa R^2}{\varepsilon} \right\rceil.$$

Итоговая сложность данной схемы рестартов для M-относительно липшицевого случая имеет вид

$$N \cdot p = \left\lceil \frac{8M^2}{\kappa\varepsilon} \right\rceil \cdot \left\lceil \log_2 \frac{2\kappa R^2}{\varepsilon} \right\rceil = O\left(\frac{M^2}{\kappa\varepsilon} \log_2 \frac{2\kappa R^2}{\varepsilon}\right).$$

Соответственно, для L-относительно гладкого случая

$$N \cdot p = \left\lceil \frac{4L}{\kappa} \right\rceil \cdot \left\lceil \log_2 \frac{2\kappa R^2}{\varepsilon} \right\rceil = O\left(\frac{L}{\kappa} \log_2 \frac{2\kappa R^2}{\varepsilon}\right).$$

□

Замечание 2. Если функция f M-относительно липшицева и $L_0 \leqslant \frac{2M^2}{\varepsilon}$, то после $p \geqslant \log_2 \frac{2\kappa R^2}{\varepsilon}$ шагов алгоритма 4 верна следующая оценка для невязки по функции:

$$f\left(\widehat{x}^{p-1}\right) - f(x_*) \leqslant M\sqrt{\frac{4\varepsilon}{\kappa}}. \tag{21}$$

Действительно,

$$f\left(\widehat{x}^{p-1}\right) - f(x_*) \leqslant \left\langle \nabla f\left(\widehat{x}^{p-1}\right), \widehat{x}^{p-1} - x_* \right\rangle \leqslant M\sqrt{2V\left(x_*, \widehat{x}^{p-1}\right)} \leqslant M\sqrt{\frac{4\varepsilon}{\kappa}}.$$

3. Об аналоге условия градиентного доминирования для относительно гладких задач

Теперь рассмотрим иной подход к аналогу относительной сильной выпуклости, позволяющий сохранить гарантии линейной скорости сходимости. Мы отправляемся от классического условия градиентного доминирования Поляка – Лоясиевича (далее — (PL)-условие). Оно было введено в [Поляк, 1963] и, как оказалось, покрывает существенно более общий класс задач по сравнению с условием сильной выпуклости. При этом такое условие гарантирует для гладких задач сходимость градиентного метода со скоростью геометрической прогрессии (см. также недавние работы [Belkin, 2021; Karimi, Nutini, Schmidt, 2016] и имеющиеся в них ссылки).

Определение 4. Пусть функция $f\colon Q \to \mathbb{R}$ дифференцируема и $\mu > 0$. Говорят, что f удовлетворяет условию Поляка – Лоясиевича ((PL)-условию) с константой μ, если верно неравенство

$$f(x) - f^* \leqslant \frac{1}{2\mu}\|\nabla f(x)\|_*^2 \quad \forall x \in \mathbb{R}^n, \tag{22}$$

где $f^* = f(x_*)$ есть значение функции f в одном из точных решений x_* задачи гладкой оптимизации $\min\limits_{x\in\mathbb{R}^n} f(x)$.

В последние годы возник интерес к задачам с (PL)-условием, в частности ввиду приложений к перепараметризованным нелинейным задачам глубокого обучения [Belkin, 2021].

Предложим аналог (PL)-условия относительно дивергенции Брэгмана и покажем, что для относительно гладких (вообще говоря, невыпуклых) задач оптимизации оно по-прежнему гарантирует сходимость метода градиентного типа (зеркального спуска) со скоростью геометрической прогрессии. Для всякого $p > 0$ напомним обозначение шага зеркального спуска:

$$x_p := \operatorname*{argmin}_{y\in\mathbb{R}^n}\{\langle\nabla f(x),\, y - x\rangle + pV(y,\, x)\}.$$

В таком случае

$$\nabla d(x_p) = \nabla d(x) - \frac{1}{p}\nabla f(x) \quad \forall x \in \mathbb{R}^n. \tag{23}$$

Теперь введем следующий аналог (PL)-условия:

$$f(x) - f^* \leqslant \frac{p^2}{\mu}V(x,\, x_p) \quad \forall x \in \mathbb{R}^n \tag{24}$$

для некоторого фиксированного $\mu > 0$ и всякого $p \geqslant \mu$. Обсудим некоторые примеры.

Пример 1. Пусть дифференцируемая прокс-функция d имеет l-липшицев градиент

$$\|\nabla d(x) - \nabla d(y)\|_* \leqslant l\|x - y\| \quad \forall x,\, y \in \mathbb{R}^n.$$

Тогда, как известно,

$$V(y,\, x) \geqslant \frac{1}{2l}\|\nabla d(x) - \nabla d(y)\|_*^2 \quad \forall x,\, y \in \mathbb{R}^n.$$

А это означает, что при произвольном x

$$V(x,\, x_p) \geqslant \frac{1}{2l}\|\nabla d(x) - \nabla d(y)\|_*^2 \overset{(23)}{=} \frac{1}{2lp^2}\|\nabla f(x)\|_*^2,$$

откуда при всяком x верно

$$\frac{1}{2l\mu}\|\nabla f(x)\|_*^2 \leqslant \frac{p^2}{\mu}V(x,\, x_p),$$

где $p \geqslant \mu > 0$.

Поэтому для (24) достаточно выполнения обычного (PL)-условия для неевклидовой нормы (относительно которой d гладкая):

$$f(x) - f^* \leqslant \frac{1}{2l\mu}\|\nabla f(x)\|_*^2 \tag{25}$$

для всякого $x \in \mathbb{R}^n$ при фиксированном $\mu > 0$. Тем не менее, как показали результаты экспериментов для регуляризованной линейной обратной задачи Пуассона далее, условие относительной

сильной выпуклости, вообще говоря, не влечет (24) при $p > \mu$. Тем самым в отличие от обычной сильной выпуклости и градиентного доминирования в «относительном» случае необходима разработка методов для каждого из этих классов задач отдельно.

Пример 2. Пусть f относительно μ-сильно выпукла, т. е.

$$f(y) \geqslant f(x) + \langle \nabla f(x),\, y - x\rangle + \mu V(y,\, x) \quad \forall x,\, y \in \mathbb{R}^n. \tag{26}$$

Тогда при всяком $x \in \mathbb{R}^n$

$$\begin{aligned}\min_{y\in\mathbb{R}^n} f(y) &\geqslant f(x) + \min_{y\in\mathbb{R}^n}\{\langle \nabla f(x),\, y - x\rangle + \mu V(y,\, x)\} = f(x) + \langle \nabla f(x),\, x_\mu - x\rangle + \mu V(x_\mu,\, x) = \\ &= f(x) + \langle \nabla f(x),\, x_\mu - x\rangle + \mu d(x_\mu) - \mu d(x) - \mu\langle \nabla d(x),\, x_\mu - x\rangle = \\ &= f(x) + \langle \nabla f(x) - \mu\nabla d(x),\, x_\mu - x\rangle + \mu d(x_\mu) - \mu d(x) \overset{(23)}{=} \\ &\overset{(23)}{=} f(x) - \mu\langle \nabla d(x_\mu),\, x_\mu - x\rangle + \mu d(x_\mu) - \mu d(x) = f(x) - \mu V(x,\, x_\mu),\end{aligned}$$

откуда

$$f(x) - f^* \leqslant \mu V(x,\, x_\mu) \quad \forall x \in \mathbb{R}^n.$$

Это означает, что (26) влечет выполнение условия (24) при $p = \mu$. Поэтому в таком случае для гарантии (24) при всяком $p \geqslant \mu$ достаточно дополнительно к (26) потребовать для всякого $x \in \mathbb{R}^n$, чтобы

$$\mu^2 V(x,\, x_\mu) \leqslant p^2 V(x,\, x_p) \quad \forall p \geqslant \mu. \tag{27}$$

Допустим теперь, что f относительно L-гладкая, то есть

$$f(y) \leqslant f(x) + \langle \nabla f(x),\, y - x\rangle + LV(y,\, x) \quad \forall x,\, y \in \mathbb{R}^n. \tag{28}$$

Тогда для градиентного метода

$$x_{k+1} := (x_k)_L \tag{29}$$

получим

$$f(x_k) - f(x_{k+1}) \geqslant LV(x_k,\, x_{k+1}).$$

Если верно (24), то

$$f(x_k) - f(x_{k+1}) \geqslant LV(x_k,\, x_{k+1}) = LV\left(x_k,\, (x_k)_L\right) \geqslant \frac{\mu}{L}\left(f(x_k) - f^*\right),$$

и тогда

$$f(x_{k+1}) - f^* \leqslant \left(1 - \frac{\mu}{L}\right)\left(f(x_k) - f^*\right),$$

то есть

$$f(x_{k+1}) - f^* \leqslant \left(1 - \frac{\mu}{L}\right)^{k+1}\left(f(x_0) - f^*\right). \tag{30}$$

Отметим, что недавно в [Yueet, Fang, Lin, 2022] получены нижние оценки для гладких оптимизационных задач с целевыми функционалами, удовлетворяющими (PL)-условию. Это указывает на оптимальность оценки (30) на более широком классе относительно гладких задач с рассматриваемым аналогом (PL)-условия относительно соответствующей дивергенции Брэгмана.

Пример 1 выше показывает, что оценка (30) может быть верна даже для невыпуклых f. Для (30) достаточно иметь уверенность в выполнении (24) только при $p = L \geqslant \mu$. Если рассмотреть адаптивный вариант метода (29) вида

$$x_{k+1} := (x_k)_{L_{k+1}}, \quad L_{k+1} \geqslant \mu > 0, \tag{31}$$

при

$$f(x_{k+1}) \leqslant f(x_k) + \langle \nabla f(x_k),\, x_{k+1} - x_k \rangle + L_{k+1} V(x_{k+1},\, x_k), \tag{32}$$

то имеем

$$f(x_k) - f(x_{k+1}) \geqslant L_{k+1} V(x_k,\, x_{k+1}).$$

Тогда (24) при $p = L_{k+1}$ гарантирует

$$f(x_k) - f(x_{k+1}) \geqslant \frac{\mu}{L_{k+1}} \Big(f(x_k) - f^*\Big),$$
$$f(x_{k+1}) - f^* \leqslant \left(1 - \frac{\mu}{L_{k+1}}\right)\Big(f(x_k) - f^*\Big).$$

Поэтому (30) принимает в этом случае вид

$$f(x_{k+1}) - f^* \leqslant \prod_{i=0}^{k} \left(1 - \frac{\mu}{L_{i+1}}\right)\Big(f(x_0) - f^*\Big). \tag{33}$$

Таким образом, справедлива следующая теорема.

Теорема 3. *Пусть f — L-относительно гладкая функция, т. е. верно неравенство* (28)*. Если, кроме того, f удовлетворяет относительному аналогу условия градиентного доминирования* (24)*, то для градиентного метода* (29) *верно неравенство* (30)*, означающее сходимость со скоростью геометрической прогрессии. При этом для адаптивного метода градиентного типа* (31) *в предположении* (24) *справедлива оценка* (33)*.*

Замечание 3. Для гарантии оценки (33) вполне достаточно иметь уверенность в выполнении (24) только при $p = L_{i+1}$ $\forall i = \overline{0,\, k}$.

Отметим следующее потенциально полезное для практической реализации методов наблюдение.

Замечание 4. Если известно, что $f^* \geqslant 0$, то условие (24) заведомо выполняется при

$$f(x) \leqslant \frac{p^2}{\mu} V(x,\, x_p).$$

Более того, для метода (31) при всяком $k = 0,\, 1,\, 2,\, \ldots$ можно ввести величину

$$\mu_{k+1} := \frac{L_{k+1}^2 V(x_k,\, x_{k+1})}{f(x_k)},$$

и тогда получим следующий аналог (33):

$$f(x_{k+1}) - f^* \leqslant \left(1 - \frac{\mu_{k+1}}{L_{k+1}}\right)\Big(f(x_k) - f^*\Big) \leqslant \prod_{i=0}^{k} \left(1 - \frac{\mu_{i+1}}{L_{i+1}}\right)\Big(f(x_0) - f^*\Big). \tag{34}$$

Аналогичное уточнение оценки (30) можно сделать и для метода (29) с постоянным шагом

$$f(x_{k+1}) - f^* \leqslant \prod_{i=0}^{k} \left(1 - \frac{\mu_{i+1}}{L}\right)\Big(f(x_0) - f^*\Big), \tag{35}$$

где $\mu_{k+1} := \frac{L^2 V(x_k,\, x_{k+1})}{f(x_k)}$. Существенно, что оценки (34) и (35) применимы даже при отсутствии гарантий для (24) при всех $x \in \mathbb{R}^n$.

4. Результаты вычислительных экспериментов

В данном параграфе с целью иллюстрации работы неадаптивного алгоритма (29) и его адаптивной версии (31), а также динамики изменения оценки качества выдаваемого этими алгоритмами приближенного решения рассматриваются некоторые численные эксперименты для линейной обратной задачи Пуассона и примера задачи с относительно сильно выпуклой целевой функцией. Все эксперименты проводились на Python 3.4 на компьютере с Intel(R) Core(TM) i7-8550U CPU (1,80 GHz, 4 ядра, 8 потоков). Оперативная память компьютера составляла 8 ГБ. Всюду далее под нормой $\|x\|_p$ вектора $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ будем понимать число $\|x\|_p = \left(|x_1|^p + |x_2|^p + \ldots + |x_n|^p\right)^{1/p}$, а под евклидовой нормой матрицы A — $\|A\|_2 = \max\{\|Ax\|_2 \mid \|x\|_2 \leqslant 1\}$.

Рассмотрим результаты расчетов для линейной обратной задачи Пуассона. Пусть $x \mapsto \sum\limits_{i=1}^{n} x_i \log(x_i)$ $\forall x \in \mathbb{R}^n_{++}$ — энтропийная функция Больцмана – Шеннона. Дивергенция Брэгмана, связанная с этой функцией, представляет собой дивергенцию Кульбака – Лейблера (KL), которая имеет следующий вид:

$$D_{\mathrm{KL}}(u, v) = \sum_{i=1}^{n} \left(u_i \log\left(\frac{u_i}{v_i}\right) - u_i + v_i \right) \quad \forall u, v \in \mathbb{R}^n_{++}. \tag{36}$$

Как известно, в обратной задаче Пуассона [Bauschke, Bolte, Teboulle, 2017; Bertero et al., 2009; Csiszar, 1991] заданы неотрицательная матрица наблюдений $A \in \mathbb{R}^{m \times n}_{+}$ и зашумленный вектор измерений $b \in \mathbb{R}^m_{++}$. Цель состоит в том, чтобы восстановить сигнал $x \in \mathbb{R}^n_{+}$ так, чтобы $Ax \approx b$. Естественной мерой близости двух неотрицательных векторов является KL-дивергенция (36). В [Bauschke, Bolte, Teboulle, 2017] было показано, что функция

$$f_1(x) = \frac{1}{n} D_{\mathrm{KL}}(b, Ax) \quad \forall x \in \mathbb{R}^n_{+} \tag{37}$$

является $\frac{1}{n}\|b\|_1$-гладкой относительно $d(x) = -\sum\limits_{i=1}^{n} \log(x_i)$. Пусть $\mu \geqslant 0$, в нашем эксперименте мы рассматриваем следующую оптимизационную задачу:

$$\min_{x \in Q} \left\{ f(x) = \frac{1}{n} D_{\mathrm{KL}}(b, Ax) + \mu d(x) \right\}, \tag{38}$$

которая представляет собой линейную задачу Пуассона с регуляризацией. При $\mu = 0$ берем $Q = \mathbb{R}^n_{+}$, а при $\mu > 0$ берем $Q = \{x \in \mathbb{R}^n_{+};\ \|x\|_2 \leqslant 1\}$, чтобы гарантировать, что $f^* \geqslant 0$. Функция f является $\left(\frac{1}{n}\|b\|_1 + \mu\right)$-гладкой и μ-сильно выпуклой относительно прокс-функции d.

Для задачи (38) мы запускаем алгоритмы (29) и (31) с начальной точкой $x_0 = \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right) \in \mathbb{R}^n$. Элементы матрицы A и вектор b в (38) генерируются случайным образом с равномерным распределением на [0, 1).

Результаты проведенных экспериментов представлены на рис. 1 и 2. Эти результаты описывают динамику значений $f(x_k)$ и оценок (34), (35) с ростом числа итераций. Кроме того, для случая, когда $\mu \neq 0$, они демонстрируют значения μ_{k+1} —

$$\mu_{k+1} := \frac{L^2_{k+1} V(x_k, x_{k+1})}{f(x_k)}, \tag{39}$$

$$\mu_{k+1} := \frac{L^2 V(x_k, x_{k+1})}{f(x_k)} \tag{40}$$

— как функцию от k.

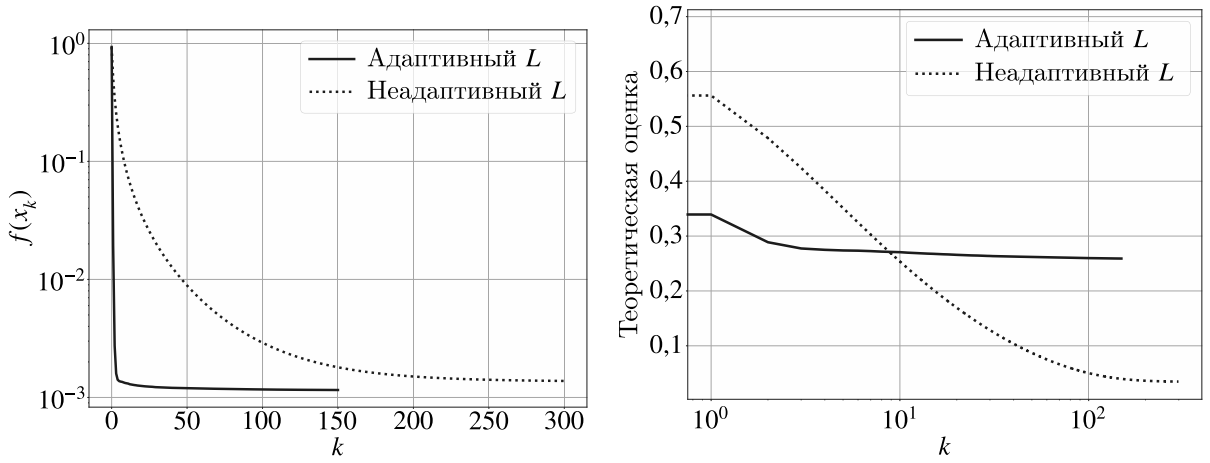

Рис. 1. Результаты алгоритмов: адаптивный (29) и неадаптивный (31) для задачи (38) с $m = 200$, $n = 100$ и $\mu = 0$. На рисунках представлены значения целевой функции f в точках x_k, а также динамика изменения теоретических оценок (34) и (35) с ростом количества итераций

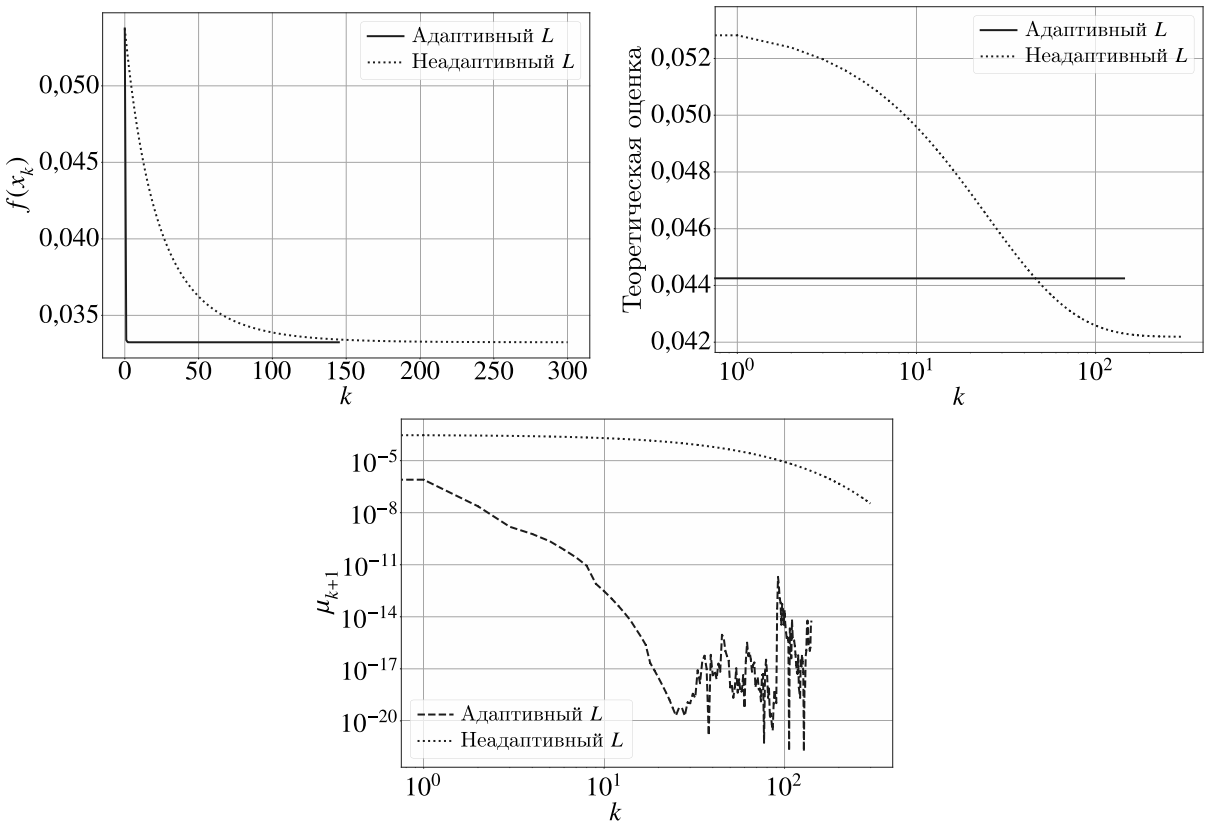

Рис. 2. Результаты алгоритмов: адаптивный (29) и неадаптивный (31) для задачи (38) с $m = 200$, $n = 100$ и $\mu = 0{,}001$. На рисунках представлены значения целевой функции f в точках x_k, а также динамика изменения теоретических оценок (34), (35) и значений параметров μ_{k+1} (39), (40) с ростом количества итераций

Из рис. 1 и 2 видно, что адаптивный алгоритм (31) хорошо работает только для первых итераций (например, для первых 10 итераций), после этого он работает медленно и не дает хорошего качества решения, в то время как неадаптивный алгоритм (29) работает лучше. Также мы видим, что, когда $\mu \neq 0$, при увеличении k значения μ_{k+1} (см. (39) и (40)) становятся значительно

меньше значения $\mu = 0{,}001$. Таким образом, мы видим, что класс задач с целевыми функциями, удовлетворяющими предлагаемому варианту (PL)-условия, и класс задач с относительными сильно выпуклыми целевыми функциями — это разные функциональные классы задач (в отличие от обычных условий сильной выпуклости и градиентного доминирования), и для каждого из них необходимо разрабатывать подходящие методы.

Теперь рассмотрим задачу с относительно сильно выпуклой целевой функцией [Lu, Freund, Nesterov, 2018]:

$$f(x) = \frac{1}{4}\|\widehat{E}x\|_2^4 + \frac{1}{4}\|\widehat{A}x - \widehat{b}\|_4^4 + \frac{1}{2}\|\widehat{C}x - \widehat{d}\|_2^2 \quad \forall x \in \mathbb{R}^n, \tag{41}$$

где $\widehat{E}$, $\widehat{A}$, $\widehat{C}$ — вещественные матрицы $n \times n$ и $\widehat{b}$, $\widehat{d} \in \mathbb{R}^n$, и пусть $\sigma_{\widehat{E}}$, $\sigma_{\widehat{C}}$ обозначают наименьшие сингулярные значения $\widehat{E}$, $\widehat{C}$ соответственно, и предположим, что $\sigma_{\widehat{E}} > 0$ и $\sigma_{\widehat{C}} > 0$. В [Lu, Freund, Nesterov, 2018] показано, что функция f является L-гладкой и μ-сильно выпуклой относительно

$$d(x) = \frac{1}{4}\|x\|_2^4 + \frac{1}{2}\|x\|_2^2 \quad \forall x \in \mathbb{R}^n \tag{42}$$

при

$$L = 3\|\widehat{E}\|_2^4 + 3\|\widehat{A}\|_2^4 + 6\|\widehat{A}\|_2^3 \cdot \|\widehat{b}\|_2 + 3\|\widehat{A}\|_2^2 \cdot \|\widehat{b}\|_2^2 + \|\widehat{C}\|_2^2 \quad \text{и} \quad \mu = \min\left\{\frac{\sigma_{\widehat{E}}^4}{3}, \sigma_{\widehat{C}}^2\right\}.$$

Для рассматриваемой оптимизационной задачи с целевой функцией (41) мы запускаем алгоритмы (31) и (29) с прокс-функцией (42), начальная точка $x_0 = \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right) \in \mathbb{R}^n$ и $L_0 = 0{,}01$.

Матрицы $\widehat{A}$, $\widehat{C}$, $\widehat{E}$ и векторы $\widehat{b}$, $\widehat{d}$ генерируются случайным образом с равномерным распределением на [0, 1). При такой постановке задачи константа μ очень мала ($\mu \approx 10^{-21}$) и задача плохо обусловлена. Результаты проведенных экспериментов представлены на рис. 3 ниже. На данных рисунках указаны динамика значений $f(x_k)$, оценки (34), (35) и значения μ_{k+1} (см. (39) и (40)) в зависимости от k.

Из этих рисунков видно, что, как и в предыдущих результатах для задачи Пуассона, адаптивный алгоритм (31) работает очень хорошо только для первых итераций, а затем лучше взять неадаптивный алгоритм (29) для такого класса задач, где это дает лучшее качество решения. Также мы можем видеть некоторую пользу предложенной процедуры подбора μ_{k+1} в (39) и (40), поскольку значения μ_{k+1} значительно больше, чем глобальное значение параметра μ.

5. Заключение

В данной статье был предложен аналог понятия относительного функционального роста (в случае общей прокс-структуры, введенной в [Gutman, Pena, 2018]), который обобщает известное понятие квадратичного роста целевой функции. Отметим, что недавно в [Stonyakin et al., 2021a] был предложен адаптивный алгоритм, гарантирующий решение задачи оптимизации с относительно липшицевым целевым функционалом (алгоритм 3) за оптимальное (с точностью до умножения на постоянный множитель) количество итераций. На базе алгоритмов 3 и 6 из [Stonyakin et al., 2021a] в настоящей статье была предложена процедура улучшения оценок скорости сходимости с помощью техники рестартов (перезапусков) в случае дополнительного предположения относительного функционального роста (которое может пониматься как относительное обобщение понятия сильной выпуклости функции или аналог условия градиентного доминирования), позволяющая предложить методы рестартов адаптивного и универсального методов. Данный подход для универсального метода позволяет доказать линейную скорость сходимости для относительно гладких выпуклых задач.

Также был предложен аналог условия Поляка – Лоясиевича относительно дивергенции Брэгмана. Был обоснован теоретический результат о сходимости со скоростью геометрической

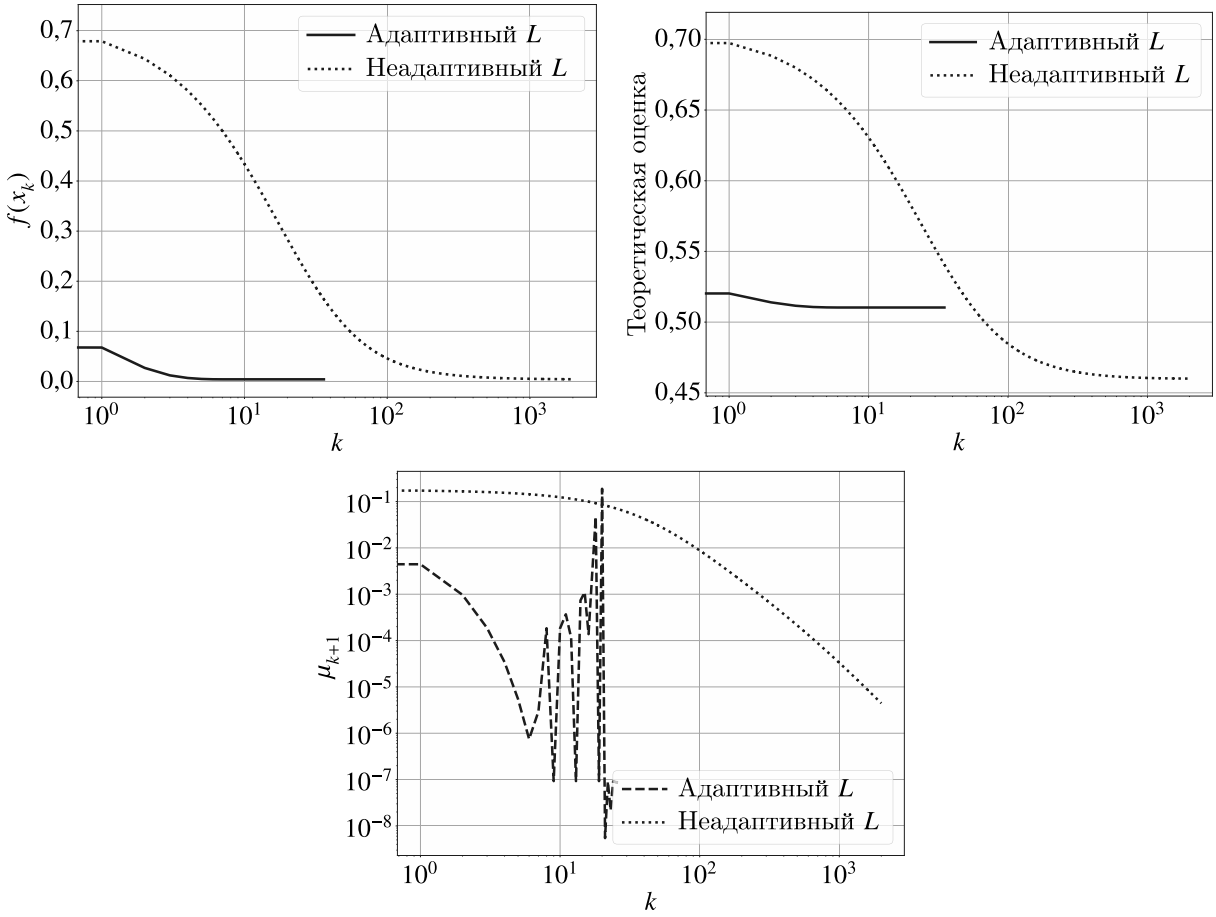

Рис. 3. Результаты алгоритмов: адаптивный (29) и неадаптивный (31) для задачи минимизации с (41) и $n = 1000$, $\mu \approx 10^{-21}$. На рисунках представлены значения целевой функции f в точках x_k, динамика изменения теоретических оценок (34) и (35), а также значений параметров μ_{k+1} из (39), (40) с ростом количества итераций

прогрессии метода градиентного типа при такого типа допущениях о варианте условия градиентного доминирования и относительной гладкости. Таким образом, были расширены границы применимости известного результата о сходимости градиентного метода со скоростью геометрической прогрессии для гладких задач с условием Поляка – Лоясиевича. Более того, были получены оценки скорости сходимости градиентного метода и в случае адаптивно подбираемых параметров. Отдельно отметим, что в случае неотрицательности искомого минимального значения функции возможно использовать и подход с адаптивно подбираемыми на итерациях метода параметрами, соответствующими параметру μ рассматриваемого варианта условия градиентного доминирования (Поляка – Лоясиевича). Последнее позволило применять найденные оценки скорости сходимости к выпуклым относительно гладким задачам без проверки условия градиентного доминирования. Были выполнены иллюстрирующие этот момент расчеты для следующих задач.

1. Линейная обратная задача Пуассона (минимизация расхождения/дивергенции Кульбака – Лейблера между Ax и b, где $A \in \mathbb{R}^{m\times n}_{+}$ — неотрицательная матрица и $b \in \mathbb{R}^{m}_{++}$ — зашумленный вектор, с целью нахождения $\widehat{x}$ такого, что $A\widehat{x}$ и b приближенно равны). Известно, что такая задача выпуклая и гладкая относительно прокс-функции, заданной логарифмическим барьером.

2. Регуляризованная линейная обратная задача Пуассона (добавлен логарифмический барьер). В таком случае целевая функция является относительно сильно выпуклой (с параметром регуляризации) в смысле соответствующего определения из [Lu, Freund, Nesterov, 2018].

3. Пример относительно гладкой и относительно сильно выпуклой задачи из подпараграфа 2.1 статьи [Lu, Freund, Nesterov, 2018].

Для первого примера показано, что адаптивность по μ может приводить к хорошей динамике оценки качества решения на начальных итерациях, а далее лучше себя проявляет адаптивная оценка для выпуклого случая. Для второго примера было показано, что относительная сильная выпуклость не влечет выполнения описанного варианта условия Поляка – Лоясиевича для соответствующего расхождения Брэгмана: адаптивно подбираемые параметры μ_{k+1} даже в случае шага с постоянным L могут оказаться существенно меньше параметра относительной сильной выпуклости. Расчеты по третьему приведенному выше примеру показали, что адаптивно подбираемые параметры μ_{k+1} в ходе работы методов могут быть существенно лучше по сравнению с очень малым глобальным значением параметра относительной сильной выпуклости [Lu, Freund, Nesterov, 2018], определяющего знаменатель геометрической прогрессии, с которой сходится градиентный метод.

Стоит сказать, что, несмотря на доказанные в статье гарантии линейной скорости сходимости представленных методов на классе относительно гладких задач, применять эти теоретические оценки на практике представляется довольно проблематичным. Дело в том, что оценки используют параметры относительного функционального роста κ и относительного варианта условия градиентного доминирования μ (хотя для этого случая предложены подходы «адаптивного» подбора, в некоторых ситуациях применимые). Тем не менее особенности реализации методов (выбор шага, проверки адаптивных правил окончания итерации) не требуют знания этих параметров, и методы можно применять на практике, не отвлекаясь на сложности оценивания параметров κ и μ. Эти величины важны лишь для теоретических оценок скорости сходимости.

Список литературы (References)

Поляк Б. Т. Градиентные методы минимизации функционалов // Журн. вычисл. математики и мат. физики. — 1963. — Т. 3, № 4. — С. 643–653.
Polyak B. T. Gradientnye metody minimizatsii funktsionalov // Zhurn. vychisl. matematiki i mat. fiziki. — 1963. — Vol. 3, No. 4. — P. 643–653 (in Russian).

Bauschke H. H., Bolte J., Teboulle M. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications // Mathematics of Operations Research. — 2017. — Vol. 42, No. 2. — P. 330–348.

Belkin M. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation // Acta Numerica. — 2021. — Vol. 30. — P. 203–248.

Bertero M., Boccacci P., Desidera G., Vicidomini G. Image deblurring with Poisson data: from cells to galaxies // Inverse Problems. — 2009. — Vol. 25, No. 12. — DOI: 10.1088/0266-5611/25/12/123006

Csiszar I. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems // The Annals of Statistics. — 1991. — Vol. 19, No. 4. — P. 2032–2066.

Gutman D. H., Pena J. F. A unified framework for Bregman proximal methods: subgradient, gradient, and accelerated gradient schemes // arXiv preprint. — 2018. — https://arxiv.org/pdf/1812.10198v1.pdf

Karimi H., Nutini J., Schmidt M. Linear convergence of gradient and proximal-gradient methods under the Polyak – Lojasiewicz condition // Joint European Conference on Machine Learning and Knowledge Discovery in Databases. — Springer, 2016. — P. 795–811.

Lu H., Freund R., Nesterov Yu. Relatively smooth convex optimization by first-order methods, and applications // SIOPT. — 2018. — Vol. 28, No. 1. — P. 333–354.

Lu H. Relative Continuity for Non-Lipschitz Nonsmooth Convex Optimization Using Stochastic (or Deterministic) Mirror Descent // Informs Journal on Optimization. — 2019. — Vol. 1, No. 4. — P. 288–303.

Nesterov Yu. Relative Smoothness: New Paradigm in Convex Optimization // EUSIPCO-2019, A Coruna, Spain, September 4. — Conference report. — 2019.

Roulet V., d'Aspremont A. Sharpness, restart and acceleration // 31st Conference on Neural Information Processing Systems (NIPS 2017). — https://papers.nips.cc/paper/2017/file/2ca65f58e35d9ad45bf7f3ae5cfd08f1-Paper.pdf

Stonyakin F., Titov A., Alkousa M., Savchuk O., Gasnikov A. Gradient-type adaptive methods for relatively Lipschitz convex optimization problems // arXiv preprint. — 2021a. — https://arxiv.org/pdf/2107.05765v10.pdf

Stonyakin F., Tyurin A., Gasnikov A., Dvurechensky P., Agafonov A., Dvinskikh D., Alkousa M., Pasechnyuk D., Artamonov S., Piskunova V. Inexact model: a framework for optimization and variational inequalities // Optimization Methods and Software. — 2021b. — Vol. 36, No. 6. — P. 1155–1201. — DOI: 10.1080/10556788.2021.1924714

Yue P., Fang C., Lin Z. On the lower bound of minimizing Polyak – Lojasiewicz functions // arXiv preprint. — 2022. — https://arxiv.org/pdf/2212.13551v1.pdf