О НЕКОТОРЫХ АДАПТИВНЫХ АЛГОРИТМАХ ЗЕРКАЛЬНОГО СПУСКА ДЛЯ ЗАДАЧ ВЫПУКЛОЙ И СИЛЬНО ВЫПУКЛОЙ ОПТИМИЗАЦИИ С ФУНКЦИОНАЛЬНЫМИ ОГРАНИЧЕНИЯМИ

Ф. С. Стонякин, М. С. Алкуса, А. А. Титов

1 Введение

Задачи минимизации выпуклых гладких и негладких функционалов с ограничениями возникают во многих задачах современной large-scale оптимизации и её приложений [5, 19]. Для таких задач имеется множество методов, среди которых можно выделить метод уровней [6], метод штрафных функций [13, 20], метод множителей Лагранжа [11]. Метод зеркального спуска (МЗС) [10, 15] восходит к обычному градиентному спуску и вполне может считаться достаточно простым методом для задач негладкой выпуклой оптимизации. Предлагаемая работа посвящена некоторым адаптивным методам зеркального спуска для задач выпуклого программирования с липшицевыми функциональными ограничениями.

Отметим, что функциональные ограничения, вообще, могут быть негладким (недифференцируемыми) и поэтому мы рассматриваем субградиентные методы. Методы с использованием субградиентов негладких выпуклых функций разрабатываются уже несколько десятилетий и восходят к известным пионерским работам, одна из которых посвящена градиентному методу для безусловных задач при евклидовом расстоянии [18], а другая — его обобщению для задач с ограничениями [17]. В работе [17] предложена идея переключения шагов между направлением субградиента целевого функционала и направлением субградиента ограничения. Обобщение метода градиентного спуска на постановку задачи с неевклидовым расстоянием называют *методом зеркального спуска*. Этот метод был предложен в [14, 15] (см. также [10]). Зеркальный спуск для задач с функциональными ограничениями был предложен в [15] (см. также [9]). При этом, как правило, для нахождения величины шага и критерия остановки для зеркального спуска необходимо знать величину константы Липшица целевого функционала, а также ограничения. Известны также и методы с адаптивным выбором шага, рассмотрены в [4] для задач без ограничений, а в [9] — для задач с ограничениями. Недавно в [1] были предложены оптимальные алгоритмы зеркального спуска для задач выпуклого программирования с липшицевыми функциональными ограничениями с адаптивным выбором шага, а также адаптивными критериями остановки. Также модификации этих методов для задач в случае нескольких выпуклых функциональных ограничений были проанализированы в [3].

В настоящей статье мы рассматриваем некоторые алгоритмы зеркального спуска для задач минимизации выпуклого функционала f с неположительным, выпуклым и липшицевым негладким функциональным ограничением g. Важно, что целевой функционал может иметь разный уровень гладкости. В частности, целевой функционал f может не удовлетворять свойству Липшица, но иметь липшицев градиент. Например, квадратичные функционалы не удовлетворяют обычному свойству Липшица (или константа Липшица достаточно большая), но имеют липшицев градиент. Можно рассматривать и негладкие выпуклые функции, равные максимуму конечного набора дифференцируемых функционалов с липшицевым градиентом. Например, пусть $A_i (i \in \overline{1, m})$

— положительно полуопределённые матрицы ($x^T A_i x \geq 0$ для всякого $x \in X$) и целевой функционал имеет вид
$$f(x) = \max_{i=\overline{1,m}} f_i(x), \tag{1.1}$$
где
$$f_i(x) = \frac{1}{2}\langle A_i x, x\rangle - \langle b_i, x\rangle + c_i, \;\; i = \overline{1,m}. \tag{1.2}$$
для некоторых фиксированных $b_i \in \mathbb{R}^n$ и $c_i \in \mathbb{R}$, для всех $i = \overline{1,m}$. Отметим, что функционалы вида (1.1) – (1.2) возникают в задачах проектирования механических конструкций Truss Topology Design со взвешенными балками [7]. Для задач минимизации функционалов такого типа при наличии выпуклых липшицевых ограничений в [1, 2, 3] на базе методики работ Ю.Е. Нестерова [6, 7] были предложены некоторые новые адаптивные алгоритмы зеркального спуска, а также обоснована их оптимальность. Часть этих результатов (про частично адаптивный метод) была заявлена в качестве доклада на VII Международную конференцию «Проблемы оптимизации и их приложения» (OPTA-2018) [2]. Настоящая статья посвящена изложению основных результатов доклада [2], а также развитию результатов [1, 2, 3] в следующих направлениях.

Во-первых, доказывается оптимальность с точки зрения оракульных оценок предложенных методов в [1, 2, 3] для задач с выпуклым липшицевым целевым функционалом, а также для задач с липшицевым гессианом при наличии выпуклых липшицевых ограничений.

Во-вторых, на базе техники рестартов (перезапусков) методов из [1, 2] (для выпуклых задач) предложены новые алгоритмы зеркального спуска аналогично для задач минимизации μ-сильно выпуклых функционалов f с неположительным, μ-сильно выпуклым и липшицевым негладким функциональным ограничением g. Заметим, что техника рестартов метода для выпуклых задач с целью ускорения сходимости для сильно выпуклых задач восходит к 1980-м годам, см. [15, 16]. Техника такого типа была использована в [12] для обоснования более высокой скорости сходимости метода зеркального спуска для сильно выпуклого целевого функционала в задачах без ограничений.

В-третьих, мы приводим ряд численных экспериментов, иллюстрирующих преимущества предложенных нами методов перед их аналогами. В частности, показано, что для задачи Ферма-Торричелли-Штейнера (целевой функционал удовлетворяет условию Липшица с константой 1) при наличии квадратичных ограничений предлагаемый нами метод может работать существенно быстрее, чем аналогичный адаптивный и также оптимальный для класса задач с липшицевым целевым функционалом с точки зрения оракульных оценок метод ([1], п. 3.1). Также приведены расчёты, иллюстрирующие некоторые преимущества предлагаемых нами методов для сильно выпуклых задач.

Статья состоит из введения и 5 основных разделов. В разделе 2 мы приводим некоторые вспомогательных сведения, а также основные понятия для метода зеркального спуска. В разделе 3 мы описываем адаптивный алгоритм зеркального спуска (алгоритм 1) из ([1], п. 3.3) и частично адаптивный алгоритм 2 [2]. В разделе 4 мы доказываем оценки скорости сходимости данных методов и обосновываем их оптимальность на рассматриваемых классах задач при различных допущениях на уровень гладкости целевого функционала. Раздел 5 посвящён методам для задач минимизации сильно выпуклых функций с рестартами алгоритмов 1 (алгоритм 3) и 2 (алгоритм 4), а также соответствующим теоретическим оценкам скорости сходимости. В последнем разделе мы приводим некоторые численные эксперименты, иллюстрирующие некоторые преимущества предлагаемых нами методов.

2 Постановка задачи и основные понятия

Пусть $(E, ||\cdot||)$ — конечномерное нормированное векторное пространство и E^* — сопряженное пространство к E со стандартной нормой:

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

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

Пусть $X \subset E$ — замкнутое выпуклое множество. Рассмотрим два выпуклых субдифференцируемых функционала f и $g : X \to \mathbb{R}$. Также предположим, что функционал g удовлетворяет условию Липшица относительно нормы $\|\cdot\|$, т. е. существует $M_g > 0$, такое, что

$$|g(x) - g(y)| \leq M_g \|x - y\| \tag{2.1}$$

для всяких $x, y \in X$. Это означает, что в каждой точке $x \in X$ можно вычислить субградиент $\nabla g(x)$, причём $\|\nabla g(x)\|_* \leq M_g$. Напомним, что для дифференцируемого функционала g субградиент $\nabla g(x)$ есть обычный градиент.

В настоящей работе будем рассматривать следующий тип задач оптимизации:

$$f(x) \to \min_{x \in X}, \tag{2.2}$$

$$g(x) \leq 0. \tag{2.3}$$

если f и g удовлетворяют упомянутым предыдущим условиям. Сделаем предположение о разрешимости задачи (2.2) – (2.3).

Отметим, что часть результатов работы относятся к постановке задачи для μ-сильно выпуклых субдифференцируемых функционалов f и $g : X \to \mathbb{R}$, т.е. для произвольных $x, y \in X$ имеет место неравенство

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2, \tag{2.4}$$

и такое же неравенство верно для g (с тем же параметром сильной выпуклости μ).

Для дальнейших рассуждений нам также потребуются вспомогательные понятия (см., например, [4]). Введём так называемую *прокс-функцию* $d : X \to \mathbb{R}$, обладающую свойством непрерывной дифференцируемости и 1-сильной выпуклости относительно нормы $\|\cdot\|$, и предположим, что $\min\limits_{x \in X} d(x) = d(0)$. Будем полагать, что существует такая константа $\Theta_0 > 0$, что $d(x_*) \leq \Theta_0^2$, где x_* — точное решение задачи (2.2)–(2.3). Отметим, что если имеется множество решений X_*, то мы предполагаем, что для константы Θ_0

$$\min_{x_* \in X_*} d(x_*) \leq \Theta_0^2.$$

Для произвольных $x, y \in X$ рассмотрим соответствующую дивергенцию Брэгмана

$$V(x, y) = d(y) - d(x) - \langle \nabla d(x), y - x\rangle.$$

В зависимости от постановки конкретной задачи возможны различные подходы к определению прокс-структуры задачи и соответствующей дивергенции Брэгмана: евклидова, энтропийная и многие другие (см., например, [4]). Стандартно определим оператор проектирования

$$\mathrm{Mirr}_x(p) = \arg\min_{u\in X}\left\{\langle p,u\rangle + V(x,u)\right\} \text{ для каждого } x\in X \text{ и } p\in E^*.$$

Сделаем предположение о том, что оператор $\mathrm{Mirr}_x(p)$ легко вычислим.

Напомним одно известное утверждение (см., например [4]).

Лемма 1. *Пусть $f : X \to \mathbb{R}$ — выпуклый субдифференцируемый функционал на выпуклом множестве X и $z = Mirr_y(h\nabla f(y))$ для некоторого $y \in X$. Тогда для произвольных $x\in X$ и $h > 0$ справедливо неравенство*

$$h\langle\nabla f(y), y-x\rangle \le \frac{h^2}{2}||\nabla f(y)||_*^2 + V(y,x) - V(z,x). \tag{2.5}$$

3 Адаптивный и частично адаптивный алгоритм зеркального спуска задач с выпуклыми функционалами

Перейдём к описанию рассматриваемых методов [1, 2] для задач (2.2) – (2.3).

Напомним следующий алгоритм адаптивного зеркального спуска для задач (2.2) – (2.3) из ([1], п. 3.3).

Algorithm 1 Адаптивный зеркальный спуск (нестандартные условия роста)

Require: точность $\varepsilon > 0$; начальная точка x^0; Θ_0; X; $d(\cdot)$.

$I =: \emptyset$

$N \leftarrow 0$

repeat

  if $g(x^N) \leqslant \varepsilon$ **then**

    $h_N \leftarrow \frac{\varepsilon}{\|\nabla f(x^N)\|_*}$

    $x^{N+1} \leftarrow Mirr_{x^N}(h_N\nabla f(x^N))$ ("продуктивные шаги")

    $N \to I$

  else

    $(g(x^N) > \varepsilon) \to$

    $h_N \leftarrow \frac{\varepsilon}{\|\nabla g(x^N)\|_*^2}$

    $x^{N+1} \leftarrow Mirr_{x^N}(h_N\nabla g(x^N))$ ("непродуктивные шаги")

  end if

  $N \leftarrow N+1$

until $\Theta_0^2 \leqslant \frac{\varepsilon^2}{2}\left(|I| + \sum\limits_{k\notin I}\frac{1}{\|\nabla g(x^N)\|_*^2}\right)$

Ensure: $\bar{x}^N := \arg\min_{x^k, k\in I} f(x^k)$

Нам потребуется ввести для целевого функционала f по аналогии с [6], определим для некоторого субградиента $\nabla f(x)$ (мы допускаем, что в ходе работы метода можно использовать произвольный субградиент) в точке $y \in X$ следующую вспомогательную величину:

$$v_f(x,y) = \begin{cases} \left\langle \dfrac{\nabla f(x)}{\|\nabla f(x)\|_*}, x - y \right\rangle, & \nabla f(x) \neq 0 \\ 0 & \nabla f(x) = 0 \end{cases}, \quad x \in X. \tag{3.1}$$

Для оценки скорости сходимости алгоритма 1 в [1] получен следующий результат.

Теорема 1. *Пусть верно неравенство* (2.1) *и известна константа* $\Theta_0 > 0$ *такова, что* $d(x_*) \leq \Theta_0^2$. *Если* $\varepsilon > 0$ *— фиксированное число, то алгоритм 1 работает не более*

$$N = \left\lceil \frac{2\max\{1, M_g^2\}\Theta_0^2}{\varepsilon^2} \right\rceil \tag{3.2}$$

итераций, причём после его остановки справедливо неравенство

$$\min_{k \in I} v_f(x^k, x_*) < \varepsilon. \tag{3.3}$$

Возможно [2] предложить также и частично адаптивный метод для задачи (2.2) – (2.3). Его отличие от алгоритма 1 в том, что адаптивно выбирается шаг лишь на продуктивных итерациях и критерий остановки неадаптивен.

Algorithm 2 Частично адаптивная версия Алгоритма 1

Require: точность $\varepsilon > 0$; начальная точка x^0; Θ_0; X; $d(\cdot)$.

$x^0 = \mathrm{argmin}_{x \in X} d(x)$
$I =: \emptyset$
$N \leftarrow 0$
repeat
  if $g(x^N) \leqslant \varepsilon \rightarrow$ **then**
    $h_N \leftarrow \frac{\varepsilon}{M_g \cdot \|\nabla f(x^N)\|_*}$
    $x^{N+1} \leftarrow Mirr_{x^N}(h_N \nabla f(x^N))$ ("продуктивные шаги")
    $N \rightarrow I$
  else
    $(g(x^N) > \varepsilon) \rightarrow$
    $h_N \leftarrow \frac{\varepsilon}{M_g^2}$
    $x^{N+1} \leftarrow Mirr_{x^N}(h_N \nabla g(x^N))$ ("непродуктивные шаги")
  end if
  $N \leftarrow N + 1$
until $N \geq \left\lceil \dfrac{2M_g^2\Theta_0^2}{\varepsilon^2} \right\rceil$

Ensure: $\bar{x}^N := \arg\min_{x^k, k \in I} f(x^k)$

Пусть $[N] = \{k \in \overline{0, N-1}\}$, $J = [N]/I$, где I набор индексов продуктивных шагов

$$h_k = \frac{\varepsilon}{M_g||\nabla f(x^k)||_*}, \tag{3.4}$$

и $|I|$ — количество "продуктивных шагов". Аналогично, для "непродуктивных шагов"из множества J аналогичная переменная определяется следующим образом:

$$h_k = \frac{\varepsilon}{M_g^2}, \tag{3.5}$$

и $|J|$ — количество "непродуктивных шагов". Очевидно,

$$|I| + |J| = N. \tag{3.6}$$

Справедлив следующий аналог теоремы 1 (см. также [2]).

Теорема 2. *Пусть $\varepsilon > 0$ — фиксированное число и алгоритм 2 работает*

$$N = \left\lceil \frac{2M_g^2\Theta_0^2}{\varepsilon^2} \right\rceil \tag{3.7}$$

итераций. Тогда

$$\min_{k\in I} v_f(x^k, x_*) < \frac{\varepsilon}{M_g}. \tag{3.8}$$

Доказательство. 1) Для продуктивных шагов из (2.5), (3.4) можно получить, что

$$h_k\langle \nabla f(x^k), x^k - x\rangle \leq \frac{h_k^2}{2}||\nabla f(x^k)||_*^2 + V(x^k, x) - V(x^{k+1}, x).$$

Принимая во внимание $\frac{h_k^2}{2}||\nabla f(x^k)||_*^2 = \frac{\varepsilon^2}{2M_g^2}$, мы имеем

$$h_k\langle \nabla f(x^k), x^k - x\rangle = \frac{\varepsilon}{M_g}\left\langle \frac{\nabla f(x^k)}{||\nabla f(x^k)||_*}, x^k - x\right\rangle = \frac{\varepsilon}{M_g}v_f(x^k, x). \tag{3.9}$$

2) Аналогично, для непродуктивных шагов $k \in J$:

$$h_k(g(x^k) - g(x)) \leq \frac{h_k^2}{2}||\nabla g(x^k)||_*^2 + V(x^k, x) - V(x^{k+1}, x).$$

Используя (2.1) и $||\nabla g(x)|| \leq M_g$, получаем

$$h_k(g(x^k) - g(x)) \leq \frac{\varepsilon^2}{2M_g^2} + V(x^k, x) - (x^{k+1}, x). \tag{3.10}$$

3) Из (3.9) и (3.10) для $x = x_*$, мы имеем

$$\frac{\varepsilon}{M_g}\sum_{k\in I} v_f(x^k, x_*) + \sum_{k\in J}\frac{\varepsilon}{M_g^2}(g(x^k) - g(x_*)) \leq$$

$$\leq N\frac{\varepsilon^2}{2M_g^2} + \sum_{k=0}^{N-1}(V(x^k, x_*) - V(x^{k+1}, x_*)). \tag{3.11}$$

Отметим, что для любого $k \in J$

$$g(x^k) - g(x_*) \geq g(x^k) > \varepsilon$$

и с учетом

$$\sum_{k=1}^{N} (V(x^k, x_*) - V(x^{k+1}, x_*)) \leq \Theta_0^2$$

неравенство (3.11) можно преобразовать следующим образом:

$$\frac{\varepsilon}{M_g} \sum_{k \in I} v_f(x^k, x_*) \leq N \frac{\varepsilon^2}{2M_g^2} + \Theta_0^2 - \frac{\varepsilon^2}{M_g^2}|J|.$$

С другой стороны,

$$\sum_{k \in I} v_f(x^k, x_*) \geq |I| \min_{k \in I} v_f(x^k, x_*).$$

Предположим, что

$$\frac{\varepsilon^2}{2M_g^2} N \geq \Theta_0^2, \text{ или } N \geq \frac{2M_g \Theta_0^2}{\varepsilon^2}. \tag{3.12}$$

Таким образом

$$|I| \frac{\varepsilon}{M_g} \min v_f(x^k, x_*) < N \frac{\varepsilon^2}{2M_g^2} - \frac{\varepsilon^2}{M_g^2}|J| + \Theta_0^2 \leq$$

$$\leq \frac{N\varepsilon^2}{M_g^2} - \frac{\varepsilon^2}{M_g^2}|J| = \frac{\varepsilon^2}{M_g^2}|I|,$$

откуда

$$|I| \frac{\varepsilon}{M_g} \min v_f(x^k, x_*) < \frac{\varepsilon^2}{M_g^2}|I| \Rightarrow \min v_f(x^k, x_*) < \frac{\varepsilon}{M_g}. \tag{3.13}$$

Чтобы закончить доказательство, мы должны показывать что $|I| \neq 0$. Предположим наоборот, что $|I| = 0 \Rightarrow |J| = N$, т. е. все шаги непродуктивны, поэтому после использования

$$g(x^k) - g(x_*) \geq g(x^k) > \varepsilon$$

мы можем видеть, что

$$\sum_{k=0}^{N-1} h_k (g(x^k) - g(x_*)) \leq \sum_{k=0}^{N-1} \frac{\varepsilon^2}{2M_g^2} + \Theta_0^2 \leq N \frac{\varepsilon^2}{2M_g^2} + N \frac{\varepsilon^2}{2M_g^2} = N \frac{\varepsilon^2}{M_g^2}.$$

Итак,

$$\frac{\varepsilon}{M_g^2} \sum_{k=0}^{N-1} (g(x^k) - g(x_*)) \leq \frac{N\varepsilon^2}{M_g^2}$$

и

$$N\varepsilon < \sum_{k=0}^{N-1} (g(x^k) - g(x_*)) \leq N\varepsilon.$$

Итак, мы получили противоречие и поэтому множество I непусто. $\square$

Замечание 1. Поясним ситуацию, когда частично адаптивная версия алгоритма может оказаться более выгодной, чем адаптивная. Например, пусть имеется ситуация, когда нет возможности точного нахождения нормы (суб)градиента ограничения $\|\nabla g(x^k)\|_*$ для одного или нескольких непродуктивных шагов ($k \in J$), а известно лишь его некоторое приближение по норме: т.е. $\|\nabla g(x^k)\|_* = \alpha_k \pm \delta_k$, где δ_k — точность приближения. По лемме 1 на всяком непродуктивном шаге x^k верно неравенство

$$h_k \left(g(x^k) - g(x_*)\right) \leq \frac{h_k^2}{2}||\nabla g(x^k)||_*^2 + V(x^k, x_*) - V(x^{k+1}, x_*). \tag{3.14}$$

Если $\alpha_k = 0$ или $\alpha_k \to 0$, то мы не можем использовать неравенство (3.14), поскольку это может привести к большой погрешности его правой части. В таком случае неадаптивный выбор шага

$$h_k = \frac{\varepsilon}{M_g^2}$$

в алгоритме 2 — более подходящий вариант для решения задачи (2.2) – (2.3).

4 Оценки скорости сходимости рассмотренных методов и их оптимальность

В данном разделе работы мы рассмотрим конкретные оценки скорости сходимости рассмотренных методов, которые обоснуют их оптимальность с точки зрения оракульных оценок (с точки зрения теории А.С. Немировского и Д.Б. Юдина). Точнее говоря ввиду липшицевости и, вообще говоря, негладкости функциональных ограничений для оптимальности метода с точки зрения нижних оракульных оценок этого достаточно показать [4], что для достижения требуемой точности ε решения задачи (2.2)–(2.3) для каждого из рассмортренных в данном разделе статьи класса целевых функционалов достаточно

$$O\left(\frac{1}{\varepsilon^2}\right)$$

итераций метода, предполагающих вычисление (суб)градиента целевого функционала или ограничения. Будем использовать следующее вспомогательное утверждение (см. например [6, 7]). Пусть x_* — решение задачи (2.2) — (2.3).

Лемма 2. *Введём следующую функцию:*

$$\omega(\tau) = \max_{x \in X}\{f(x) - f(x_*) : ||x - x_*|| \leq \tau\}, \tag{4.1}$$

где τ положительное число. Тогда для всякого $y \in X$

$$f(y) - f(x_*) \leq \omega(v_f(y, x_*)). \tag{4.2}$$

Теперь мы можем показать (см. также доклад [2]), как с использованием предыдущего утверждения и теоремы 2, можно оценить скорость сходимости алгоритма 2, если целевой функционал f дифференцируем и его градиент удовлетворяет условию Липшица:

$$||\nabla f(x) - \nabla f(y)||_* \leq L||x - y|| \quad \forall x, y \in X. \tag{4.3}$$

Используя следующий известный факт

$$f(x) \leq f(x_*) + ||\nabla f(x_*)||_* ||x - x_*|| + \frac{1}{2} L ||x - x_*||^2,$$

мы можем получить

$$\min_{k \in I} f(x^k) - f(x_*) \leq \min_{k \in I} \left\{ ||\nabla f(x_*)||_* ||x^k - x_*|| + \frac{1}{2} L ||x^k - x_*||^2 \right\}.$$

Итак

$$f(x) - f(x_*) \leq ||\nabla f(x_*)||_* \frac{\varepsilon}{M_g} + \frac{L\varepsilon^2}{2M_g}.$$

Поэтому имеет место следующий результат [2].

Следствие 1. *Пусть f дифференцируем на X и верно* (4.3). *Тогда после*

$$N = \left\lceil \frac{2M_g^2 \Theta_0^2}{\varepsilon^2} \right\rceil$$

шагов работы алгоритма 2 выполнена следующая оценка:

$$\min_{0 \leq k \leq N} f(x^k) - f(x_*) \leq ||\nabla f(x_*)||_* \frac{\varepsilon}{M_g} + \frac{L}{2} \frac{\varepsilon^2}{M_g^2}.$$

Мы можем применить наш метод к некоторому классу задач с негладкими целевыми функционалами специального типа [2].

Следствие 2. *Предположим, что $f(x) = \max\limits_{i=\overline{1,m}} f_i(x)$, где f_i дифференцируемы на каждой $x \in X$ и*

$$||\nabla f_i(x) - \nabla f_i(y)||_* \leq L_i ||x - y|| \quad \forall x, y \in X.$$

Тогда после

$$N = \left\lceil \frac{2M_g^2 \Theta_0^2}{\varepsilon^2} \right\rceil$$

шагов работы Алгоритма 2 выполнена следующая оценка:

$$\min_{0 \leq k \leq N} f(x^k) - f(x_*) \leq ||\nabla f(x_*)||_* \frac{\varepsilon}{M_g} + \frac{L}{2} \frac{\varepsilon^2}{M_g^2},$$

где $L = \max\limits_{i=\overline{1,m}} L_i$.

Замечание 2. Вообще $||\nabla f(x_*)||_* \neq 0$, поскольку мы рассматриваем некоторый класс условных задач.

Замечание 3. Пусть целевой функционал $f : X \to \mathbb{R}$ удовлетворяет условию Липшица:

$$|f(x) - f(y)| \leqslant M_f ||x - y|| \quad \forall x, y \in X. \tag{4.4}$$

Итак

$$f(x) \leq f(x_*) + M_f ||x - x_*||,$$

мы можем получить

$$\min_{k\in I} f(x^k) - f(x_*) \leq \min_{k\in I} \left\{ M_f ||x^k - x_*|| \right\}.$$

Итак, комбинируя утверждения теоремы 1 и леммы 2, мы можем гарантировать после остановки алгоритма 1 выполнение неравенства

$$f(x) - f(x_*) \leq M_f \varepsilon,$$

и аналогично из теоремы 2 для алгоритма 2:

$$f(x) - f(x_*) \leq \frac{M_f}{M_g}\varepsilon.$$

Поэтому имеет место следующий результат.

Следствие 3. *Если f удовлетворяет условию Липшица (4.4) на X. Тогда*

- *после*
$$N = \left\lceil \frac{2\max\{1, M_g^2\}\Theta_0^2}{\varepsilon^2} \right\rceil$$
шагов работы алгоритма 1, выполнена следующая оценка:
$$\min_{1\leq k\leq N} f(x^k) - f(x_*) \leq M_f \varepsilon;$$

- *после*
$$N = \left\lceil \frac{2M_g^2\Theta_0^2}{\varepsilon^2} \right\rceil$$
шагов работы алгоритма 2, выполнена следующая оценка:
$$\min_{1\leq k\leq N} f(x^k) - f(x_*) \leq \frac{M_f}{M_g}\varepsilon.$$

Замечание 4. Пусть целевой функционал $f : X \to \mathbb{R}$ дважды дифференцируем на X и имеет липшицев гессиан, т.е. справедливо следующее неравенство

$$||\nabla^2 f(x) - \nabla^2 f(y)||_* \leq L||x - y|| \quad \forall x, y \in X. \tag{4.5}$$

Используя следующее неравенство (см. [6], лемма 1.2.4)

$$|f(x) - f(x_*) - \langle \nabla f(x_*), x - x_* \rangle - \frac{1}{2}\left\langle \nabla^2 f(x_*)(x - x_*), x - x_* \right\rangle| \leq \frac{L}{6}||x - x_*||^3,$$

мы можем видеть, что

$$f(x) \leq f(x_*) + ||\nabla f(x_*)|| \cdot ||x - x_*|| + \frac{1}{2}||\nabla^2 f(x_*)(x - x_*)|| \cdot ||x - x_*|| + \frac{L}{6}||x - x_*||^3$$

Итак

$$f(x) \leq f(x_*) + ||\nabla f(x_*)|| \cdot ||x - x_*|| + \frac{1}{2}||\nabla^2 f(x_*)||_{Fro} \cdot ||x - x_*||^2 + \frac{L}{6}||x - x_*||^3$$

где $||A||_{Fro} = tr(A^TA)$ норма Фробениуса матрицы $A \in \mathbb{R}^{m\times n}$. Тогда

$$\min_{k\in I} f(x^k)-f(x_*) \leq \min_{k\in I}\left\{||\nabla f(x_*)||\cdot||x^k-x_*|| + \frac{1}{2}||\nabla^2 f(x_*)||_{Fro}\cdot||x^k-x_*||^2 + \frac{L}{6}||x^k-x_*||^3\right\}.$$

Итак, комбинируя утверждение теоремы 1 и леммы 2, возможно получить

$$f(x)-f(x_*) \leq ||\nabla f(x_*)||_*\cdot\varepsilon + \frac{1}{2}||\nabla^2 f(x_*)||_{Fro}\cdot\varepsilon^2 + \frac{L}{6}\varepsilon^3,$$

а также аналогично из теоремы 2

$$f(x)-f(x_*) \leq ||\nabla f(x_*)||_*\cdot\frac{\varepsilon}{M_g} + \frac{1}{2}||\nabla^2 f(x_*)||_{Fro}\cdot\frac{\varepsilon^2}{M_g^2} + \frac{L}{6}\frac{\varepsilon^3}{M_g^3}.$$

Поэтому имеет место следующий результат.

Следствие 4. *Пусть f дважды дифференцируем на X и имеет липшицев гессиан, т.е. верно* (4.5). *Тогда*

- *после*
$$N = \left\lceil\frac{2\max\{1, M_g^2\}\Theta_0^2}{\varepsilon^2}\right\rceil$$
шагов работы алгоритма 1 выполнена следующая оценка:
$$\min_{1\leq k\leq N} f(x^k)-f(x_*) \leq ||\nabla f(x_*)||_*\cdot\varepsilon + \frac{1}{2}||\nabla^2 f(x_*)||_{Fro}\cdot\varepsilon^2 + \frac{L}{6}\varepsilon^3;$$
- *после*
$$N = \left\lceil\frac{2M_g^2\Theta_0^2}{\varepsilon^2}\right\rceil$$
шагов работы алгоритма 2 выполнена следующая оценка:
$$\min_{1\leq k\leq N} f(x^k)-f(x_*) \leq ||\nabla f(x_*)||_*\cdot\frac{\varepsilon}{M_g} + \frac{1}{2}||\nabla^2 f(x_*)||_{Fro}\cdot\frac{\varepsilon^2}{M_g^2} + \frac{L}{6}\frac{\varepsilon^3}{M_g^3}.$$

Мы можем применить наши методы к некоторому классу задач с негладкими целевыми функционалами.

Следствие 5. *Предположим, что $f(x) = \max\limits_{i=\overline{1,m}} f_i(x)$, где f_i дважды дифференцируемы в каждой точке $x\in X$ и*

$$||\nabla^2 f(x) - \nabla^2 f(y)||_* \leq L_i||x-y|| \quad \forall x,y\in X.$$

Тогда

- *после*
$$N = \left\lceil\frac{2\max\{1, M_g^2\}\Theta_0^2}{\varepsilon^2}\right\rceil$$
шагов работы алгоритма 1 выполнена следующая оценка:
$$\min_{1\leq k\leq N} f(x^k)-f(x_*) \leq ||\nabla f(x_*)||_*\cdot\varepsilon + \frac{1}{2}||\nabla^2 f(x_*)||_{Fro}\cdot\varepsilon^2 + \frac{L}{6}\varepsilon^3,$$
где $L = \max\limits_{i=\overline{1,m}} L_i$;

- *после*
$$N = \left\lceil \frac{2M_g^2\Theta_0^2}{\varepsilon^2} \right\rceil$$
шагов работы алгоритма 2 выполнена следующая оценка:
$$\min_{1\le k\le N} f(x^k) - f(x_*) \le ||\nabla f(x_*)||_* \cdot \frac{\varepsilon}{M_g} + \frac{1}{2}||\nabla^2 f(x_*)||_{Fro} \cdot \frac{\varepsilon^2}{M_g^2} + \frac{L}{6}\frac{\varepsilon^3}{M_g^3},$$
где $L = \max\limits_{i=\overline{1,m}} L_i$.

5 Об ускорении рассматриваемых методов зеркального спуска для сильно выпуклых задач

В этом разделе работы мы рассмотрим задачу

$$f(x) \to \min, \ \ g(x) \le 0, \ \ x \in X \tag{5.1}$$

с предположениями (2.1), а также сильной выпуклости f и g с одинаковым параметром $\mu > 0$. Мы также слегка модифицируем предположения на прокс-функцию $d(x)$. А именно, предположим, что $0 = \arg\min\limits_{x\in X} d(x)$ и что d ограничено на единичном шаре в выбранной норме $\|\cdot\|$, т. е.

$$d(x) \le \Theta_0^2, \quad \forall x \in X : \|x\| \le 1, \tag{5.2}$$

Наконец, мы допускаем, что нам дана начальная точка $x^0 \in X$ и число $R_0 > 0$ такое, что $\|x_0 - x_*\|^2 \le R_0^2$. Для построения метода решения задачи (5.1) при заданных предположениях мы используем идею рестартов (перезапусков) алгоритма 1 и алгоритма 2. Рассмотрим вспомогательное утверждение (см., например [8]).

Лемма 3. *Если f и g — μ-сильно выпуклые функционалы относительно нормы $\|\cdot\|$ на X, $x_* = arg\min\limits_{x\in X} f(x)$, $g(x) \le 0$ ($\forall x \in X$) и для некоторых $\varepsilon_f > 0$, а также $\varepsilon_g > 0$ верно:*

$$f(x) - f(x_*) \le \varepsilon_f, \ \ g(x) \le \varepsilon_g. \tag{5.3}$$

Тогда

$$\frac{\mu}{2}\|x - x_*\|^2 \le \max\{\varepsilon_f, \varepsilon_g\}. \tag{5.4}$$

Предположим, что $f(x) = \max\limits_{i=\overline{1,m}} f_i(x)$, где f_i дифференцируемы во всякой точке $x \in X$ и имеют с липшицев градиент, т. е. существуют $L_i > 0$ такие, что

$$\|\nabla f_i(x) - \nabla f_i(y)\|_* \le L_i\|x - y\| \quad \forall x, y \in X. \tag{5.5}$$

Рассмотрим функцию $\tau : \mathbb{R}^+ \to \mathbb{R}^+$:

$$\tau(\delta) = \max\left\{\delta\|\nabla f(x_*)\|_* + \frac{\delta^2 L}{2}, \ \delta\right\}, \tag{5.6}$$

где

$$L := \max_{i=\overline{1,m}}\{L_i\}.$$

Ясно, что функция τ возрастает и поэтому для всякого $\varepsilon > 0$ существует

$$\hat{\varphi}(\varepsilon) > 0 : \ \ \tau(\hat{\varphi}(\varepsilon)) = \varepsilon.$$

Рассмотрим следующий адаптивный алгоритм 3 для задачи (5.1).

Algorithm 3 Адаптивный алгоритм зеркального спуска для сильно выпуклых функционалов.

Require: точность $\varepsilon > 0$; начальная точка x_0; Θ_0 s.t. $d(x) \leq \Theta_0^2 \quad \forall x \in X : \|x\| \leq 1$; $X; d(\cdot)$; параметр сильно выпуклости μ; R_0 s.t. $\|x^0 - x_*\|^2 \leq R_0^2$.

1: Set $d_0(x) = d\left(\frac{x-x^0}{R_0}\right)$.
2: Set $p = 1$.
3: **repeat**
4: Set $R_p^2 = R_0^2 \cdot 2^{-p}$.
5: Set $\varepsilon_p = \frac{\mu R_p^2}{2}$.
6: Set x^p — выход алгоритма 1 с точностью ε_p, прокс-функцией $d_{p-1}(\cdot)$ и Θ_0^2.
7: $d_p(x) \leftarrow d\left(\frac{x-x^p}{R_p}\right)$.
8: Set $p = p + 1$.
9: **until** $p > \log_2 \frac{\mu R_0^2}{2\varepsilon}$.

Ensure: x^p.

Теорема 3. *Пусть f имеет липшицев градиент, удовлетворяющий* (5.5). *Если f и g — μ-сильно выпуклые функционалы на $X \subset \mathbb{R}^n$ и $d(x) \leq \Theta_0^2$ для всех $x \in X$, таких, что $\|x\| \leq 1$. Пусть начальное приближение $x^0 \in X$ и число $R_0 > 0$ заданы так, что*

$$\|x^0 - x_*\|^2 \leq R_0^2.$$

Тогда для $\widehat{p} = \left\lceil \log_2 \dfrac{\mu R_0^2}{2\varepsilon} \right\rceil$ выход $x_{\widehat{p}}$ есть ε-решение задачи (5.1) *(т. е. $f(x^{\widehat{p}}) - f(x_*) < \varepsilon$ и $g(x^{\widehat{p}}) < \varepsilon$), где*

$$\|x^{\widehat{p}} - x_*\|^2 \leq \frac{2\varepsilon}{\mu}.$$

При этом, количество итераций алгоритма 1 не более

$$\widehat{p} + \sum_{p=1}^{\widehat{p}} \frac{2\Theta_0^2 \max\{1, M_g^2\}}{\hat{\varphi}^2(\varepsilon_p)}, \ \ \text{где} \ \ \varepsilon_p = \frac{\mu R_0^2}{2^{p+1}}$$

итераций.

Доказательство. Функция $d_p(x) = d\left(\dfrac{x - x^p}{R_p}\right)$, которая определена в алгоритме 3, является 1-сильно выпуклой функцией относительно нормы $\dfrac{\|.\|}{R_p}$, для всех $p \geq 0$. Математической индукцией мы покажем, что

$$\|x^p - x_*\|^2 \leq R_p^2 \quad \forall p \geq 0.$$

Для $p = 0$ это утверждение очевидно из-за выбора x^0 и R_0. Предположим, что для некоторого p, у нас $\|x^p - x_*\|^2 \leq R_p^2$, и давайте докажем, что $\|x^{p+1} - x_*\|^2 \leq R_{p+1}^2$. Имеем $\|x^p - x_*\|^2 \leq R_p^2$. Докажем, что $\|x^{p+1} - x_*\|^2 \leq R_{p+1}^2$. У нас $d_p(x_*) \leq \Theta_0^2$, таким образом, по теореме 1, на $(p+1)$-м рестарте после не более чем

$$N_{p+1} = \left\lceil \frac{2\Theta_0^2 \max\{1, M_g^2\}}{\hat{\varphi}^2(\varepsilon_{p+1})} \right\rceil$$

итераций алгоритма 1, следующие неравенства верны для $x^{p+1} = \bar{x}^{N_{p+1}}$:

$$f(x^{p+1}) - f(x_*) \leq \varepsilon_{p+1}, \quad g(x^{p+1}) \leq \varepsilon_{p+1} \quad \text{for} \quad \varepsilon_{p+1} = \frac{\mu R_{p+1}^2}{2}.$$

Тогда, согласно лемме 3

$$\|x^{p+1} - x_*\|^2 \leq \frac{2\varepsilon_{p+1}}{\mu} = R_{p+1}^2.$$

Итак, для всех $p \geq 0$ мы доказали, что

$$\|x^p - x_*\|^2 \leq R_p^2 = \frac{R_0^2}{2^p}, \quad f(x^p) - f(x_*) \leq \frac{\mu R_0^2}{2^{p+1}}, \quad g(x^p) \leq \frac{\mu R_0^2}{2^{p+1}}.$$

и так, для $p = \widehat{p} = \left\lceil \log_2 \frac{\mu R_0^2}{2\varepsilon} \right\rceil$, x_p это ε-решение задачи (5.1) и справедливо следующее соотношение

$$\|x^p - x_*\|^2 \leq R_p^2 = \frac{R_0^2}{2^p} \leq \frac{2\varepsilon}{\mu}.$$

Итак, пусть K обозначим общее число итераций алгоритма 1, и N_p к общему числу итераций алгоритма 1 на p-м рестарте. Поскольку функция $\tau : \mathbb{R}^+ \to \mathbb{R}^+$, возрастает и для каждого $\varepsilon > 0$ существует $\hat{\varphi}(\varepsilon) > 0 : \ \tau(\hat{\varphi}(\varepsilon)) = \varepsilon$. Поэтому мы имеем

$$K = \sum_{p=1}^{\widehat{p}} N_p = \sum_{p=1}^{\widehat{p}} \left\lceil \frac{2\Theta_0^2 \max\{1, M_g^2\}}{\hat{\varphi}^2(\varepsilon_p)} \right\rceil \leq \widehat{p} + \sum_{p=1}^{\widehat{p}} \frac{2\Theta_0^2 \max\{1, M_g^2\}}{\hat{\varphi}^2(\varepsilon_p)}.$$

□

Замечание 5. Предыдущую оценку количества итераций работы алгоритма 1 можно несколько конкретизировать в случае $\varepsilon < 1$. В этом случае при всяком $\delta < 1$ имеем $\tau(\delta) \leq C\delta$ для некоторой константы C. Поэтому можно считать, что $\hat{\varphi}(\varepsilon) = \widehat{C} \cdot \varepsilon$ для соответствующей константы $\widehat{C} > 0$. Это означает, что на $p+1$-м рестарте алгоритма 1 после не более, чем

$$k_{p+1} = \left\lceil \frac{\Omega \max\{1, M_g^2\} R_p^2}{\varepsilon_{p+1}^2} \right\rceil \tag{5.7}$$

итераций работы алгоритма 1, выход x_{p+1} гарантированно удовлетворяет неравенству

$$f(x^{p+1}) - f(x_*) \leq \widehat{C} \cdot \varepsilon_{p+1}, \quad g(x^{p+1}) \leq \varepsilon_{p+1},$$

где $\varepsilon_{p+1} = \frac{\mu R_{p+1}^2}{2}$. Тогда по лемме 3,

$$\|x^{p+1} - x_*\|^2 \leq \frac{2 \max\{1, \widehat{C}\} \varepsilon_{p+1}}{\mu} = \max\{1, \widehat{C}\} \cdot R_{p+1}^2.$$

Таким образом, всех $p \geq 0$,

$$\|x^p - x_*\|^2 \leq \max\{1, \widehat{C}\} \cdot R_p^2 = \max\{1, \widehat{C}\} \cdot R_0^2 \cdot 2^{-p}.$$

В то же время мы имеем для всяких $p \geq 1$ имеют место неравенства:

$$f(x^p) - f(x_*) \leq \max\{1, \widehat{C}\} \cdot \frac{\mu R_0^2}{2} \cdot 2^{-p}, \quad g(x_p) \leq \max\{1, \widehat{C}\} \cdot \frac{\mu R_0^2}{2} \cdot 2^{-p}.$$

Таким образом, если $p > \log_2 \frac{\mu R_0^2}{2\varepsilon}$, то x_p будет $\max\{1, \widehat{C}\} \cdot \varepsilon$-решением для поставленной задачи, причём:

$$\|x^p - x_*\|^2 \leq \max\{1, \widehat{C}\} \cdot R_0^2 \cdot 2^{-p} \leq \frac{2\varepsilon}{\mu}.$$

Оценим теперь общее число N итераций алгоритма 1. Пусть $\hat{p} = \left\lceil \log_2 \frac{\mu R_0^2}{2\varepsilon} \right\rceil$. Тогда согласно (5.7), мы имеем с точностью до умножения на константу:

$$N = \sum_{p=1}^{\hat{p}} k_p \leq \sum_{p=1}^{\hat{p}} \left(1 + \frac{2\Theta_0^2 \max\{1, M_g^2\} R_p^2}{\varepsilon_{p+1}^2}\right) = \sum_{p=1}^{\hat{p}} \left(1 + \frac{32\Theta_0^2 \max\{1, M_g^2\} 2^p}{\mu^2 R_0^2}\right)$$
$$\leq \hat{p} + \frac{64\Theta_0^2 \max\{1, M_g^2\} 2^{\hat{p}}}{\mu^2 R_0^2} \leq \hat{p} + \frac{64\Theta_0^2 \max\{1, M_g^2\}}{\mu\varepsilon}.$$

Замечание 6. Вообще говоря, $\varphi(\varepsilon)$ зависит от $\|\nabla f(x_*)\|_*$ и константа Липшица L для ∇f. Если $\|\nabla f(x_*)\|_* < M_g$, тогда $\varphi(\varepsilon) = \varepsilon$ для небольших достаточно ε:

$$\varepsilon < \frac{2(M_g - \|\nabla f(x_*)\|_*)}{L}.$$

Для другого случая ($\|\nabla f(x_*)\|_* > M_g$) у нас $\forall \varepsilon > 0$:

$$\varphi(\varepsilon) = \frac{\sqrt{\|\nabla f(x_*)\|_*^2 + 2\varepsilon L} - \|\nabla f(x_*)\|_*}{L}.$$

Рассмотрим также следующую частично адаптивную версию алгоритма 4 для задачи (5.1) [2].

В условиях следствия 2 после остановки алгоритма 4 будут верными неравенства (5.3) для

$$\varepsilon_f = \frac{\varepsilon}{M_g}\|\nabla f(x_*)\|_* + \frac{\varepsilon^2 L}{2M_g^2}$$

и $\varepsilon_g = \varepsilon$. Рассмотрим функцию $\tau : \mathbb{R}^+ \to \mathbb{R}^+$:

$$\tau(\delta) = \max\left\{\delta\|\nabla f(x_*)\|_* + \frac{\delta^2 L}{2};\ \delta M_g\right\}.$$

Ясно, что функция τ возрастает и поэтому для каждого $\varepsilon > 0$ существует

$$\varphi(\varepsilon) > 0 : \ \ \tau(\varphi(\varepsilon)) = \varepsilon.$$

Справедлива следующая

Algorithm 4 Частично адаптивный алгоритм зеркального спуска для сильно выпуклых функционалов

Require: точность $\varepsilon > 0$; начальная точка x^0; Θ_0 s.t. $d(x) \leq \Theta_0^2 \quad \forall x \in X : \|x\| \leq 1$; $X; d(\cdot)$; параметр сильно выпуклости $\mu; R_0$ s.t. $\|x^0 - x_*\|^2 \leq R_0^2$.

1: Set $d_0(x) = d\left(\frac{x-x^0}{R_0}\right)$.
2: Set $p = 1$.
3: **repeat**
4: Set $R_p^2 = R_0^2 \cdot 2^{-p}$.
5: Set $\varepsilon_p = \frac{\mu R_p^2}{2}$.
6: Set x^p — выход алгоритма 2 с точностью ε_p, прокс-функцией $d_{p-1}(\cdot)$ и Θ_0^2.
7: $d_p(x) \leftarrow d\left(\frac{x-x^p}{R_p}\right)$.
8: Set $p = p + 1$.
9: **until** $p > \log_2 \frac{\mu R_0^2}{2\varepsilon}$.

Ensure: x^p.

Теорема 4. *Пусть f и g удовлетворяют условиям следствия 2. Если f и g — μ-сильно выпуклые функционалы на $X \subset \mathbb{R}^n$ и $d(x) \leq \Theta_0^2 \;\; \forall\, x \in X,\; \|x\| \leq 1$. Пусть начальное приближение $x^0 \in X$ и число $R_0 > 0$ заданы так, что $\|x^0 - x_*\|^2 \leq R_0^2$. Тогда для $\widehat{p} = \left\lceil \log_2 \dfrac{\mu R_0^2}{2\varepsilon} \right\rceil$ выход $x^{\widehat{p}}$ есть ε-решение задачи (5.1), где*

$$\|x^{\widehat{p}} - x_*\|^2 \leq \frac{2\varepsilon}{\mu}.$$

При этом общее количество итераций алгоритма 2 не превышает

$$\widehat{p} + \sum_{p=1}^{\widehat{p}} \frac{2\Theta_0^2 M g^2}{\varphi^2(\varepsilon_p)}, \quad \text{где} \;\; \varepsilon_p = \frac{\mu R_0^2}{2^{p+1}}.$$

Доказательство. Функция $d_p(x)$ $(p = 0, 1, 2, \ldots)$ 1-сильно выпукла относительно нормы $\dfrac{\|\cdot\|}{R_p}$, для всех $p \geq 0$. Методом математической индукции покажем, что

$$\|x^{\widehat{p}} - x_*\|^2 \leq R_p \;\; \forall p \geq 0.$$

Для $p = 0$ это утверждение очевидно в силу выбора x_0 и R_0. Предположим, что для некоторого p: $\|x^p - x_*\|^2 \leq R_p^2$. Докажем, что $\|x^{p+1} - x_*\|^2 \leq R_{p+1}^2$. У нас $d_p(x_*) \leq \Theta_0^2$, и на $(p+1)$-м рестарте после не более чем

$$\left\lceil \frac{2\Theta_0^2 M_g^2}{\varphi^2(\varepsilon_{p+1})} \right\rceil$$

итераций алгоритма 2 будут выполняться следующие неравенства:

$$f(x^{p+1}) - f(x_*) \leq \varepsilon_{p+1}, \quad g(x^{p+1}) \leq \varepsilon_{p+1} \;\; \text{для} \;\; \varepsilon_{p+1} = \frac{\mu R_{p+1}^2}{2}.$$

Тогда, согласно лемме 3

$$\|x^{p+1} - x_*\|^2 \leq \frac{2\varepsilon_{p+1}}{\mu} = R_{p+1}^2.$$

Итак, для произвольного $p \geq 0$

$$\|x^p - x_*\|^2 \leq R_p^2 = \frac{R_0^2}{2^p}, \quad f(x^p) - f(x_*) \leq \frac{\mu R_0^2}{2} 2^{-p}, \quad g(x_p) \leq \frac{\mu R_0^2}{2} 2^{-p}.$$

Для $p = \widehat{p} = \left\lceil \log_2 \frac{\mu R_0^2}{2\varepsilon} \right\rceil$ верное следующее соотношение:

$$\|x^p - x_*\|^2 \leq R_p^2 = R_0^2 \cdot 2^{-p} \leq \frac{2\varepsilon}{\mu}.$$

Остается лишь заметить, что количество итераций работы алгоритма 2 не превосходит

$$\sum_{p=1}^{\widehat{p}} \left\lceil \frac{2\Theta_0^2 M_g^2}{\varphi^2(\varepsilon_{p+1})} \right\rceil \leq \widehat{p} + \sum_{p=1}^{\widehat{p}} \frac{2\Theta_0^2 M_g^2}{\varphi^2(\varepsilon_{p+1})}.$$

□

Замечание 7. По аналогии с рассуждениями замечания 5, при $\varepsilon < 1$ с точностью до умножения на константу можно уточнить верхнюю оценку количества итераций 2:

$$N = \hat{p} + \frac{64\Theta_0^2 M_g^2 \cdot 2^{\hat{p}}}{\mu^2 R_0^2} \leq \hat{p} + \frac{64\Theta_0^2 \cdot M_g^2}{\mu\varepsilon}.$$

Замечание 8. Обратив внимание на следствия 3 и 5, нетрудно понять, что при условии $\varepsilon < 1$ утверждения замечаний 5 и 7 нетрудно перенести и на случаи, когда целевой функционал f удовлетворяет условию Липшица или условию Липшица для гессиана f.

6 Численные эксперименты

6.1 Сравнение скорости работы методов для задачи Ферма-Торричелли-Штейнера с ограничениями.

Отметим, что в ([1], п. 3.1) предложен также следующий адаптивный метод, оптимальный с точки зрения нижних оракульных оценок в случае задач с липшицевым целевым функционалом.

Algorithm 5 Адаптивный зеркальный спуск (липшицев целевой функционал)

Require: $\varepsilon > 0, \Theta_0 : d(x_*) \leqslant \Theta_0^2$

$x^0 = argmin_{x \in X}\, d(x)$

$I =: \emptyset$

$N \leftarrow 0$

repeat

  if $g(x^N) \leqslant \varepsilon$ **then**

    $M_N = ||\nabla f(x^N)||_*,\ h_N = \frac{\varepsilon}{M_N^2}$

    $x^{N+1} = Mirr_{x^N}(h_N \nabla f(x^N))$ // *"продуктивные шаги"*

    $N \to I$

  else

    $M_N = ||\nabla g(x^N)||_*,\ h_N = \frac{\varepsilon}{M_N^2}$

    $x^{N+1} = Mirr_{x^N}(h_N \nabla g(x^N))$ // *"непродуктивные шаги"*

  end if

  $N \leftarrow N + 1$

until $\sum\limits_{j=0}^{N-1} \frac{1}{M_j^2} \geqslant 2\frac{\Theta_0^2}{\varepsilon^2}$

Ensure: $\bar{x}^N := \frac{\sum\limits_{k \in I} x^k h_k}{\sum\limits_{k \in I} h_k}$

В настоящей работе мы рассматриваем альтернативный метод (алгоритм 1), оптимальность которого уже удаётся установить для условных задач с более широким классом целевых функционалов (имеющих липшицев градиент или липшицев гессиан). Но оказывается, что и в случае липшицевого целевого функционала, когда применим алгоритм 5, алгоритм 1 может работать быстрее. В качестве примера приведём расчёты для известной *задачи Ферма-Торричелли-Штейнера с ограничениями.*

Задача. *Для заданных точек $A_k = (a_{1k}, a_{2k}, \ldots, a_{nk},)$ в n-мерном евклидовом пространстве $\mathbb{R}^n$ необходимо найти такую точку $X = (x_1, x_2, \ldots, x_n)$, чтобы целевая функция*

$$f(x) := \sum_{k=1}^{n} \sqrt{(x_1 - a_{1k})^2 + (x_2 - a_{2k})^2 + \ldots + (x_n - a_{nk})^2}$$

принимала наименьшее значение на множестве X, которое задаётся несколькими ограничениями:

$$g_1((x_1, \ldots, x_{10})) = 2x_1^2 + x_2^2 + \ldots + x_{10}^2 - 1 \leqslant 0,$$
$$g_2((x_1, \ldots, x_{10})) = x_1^2 + 2x_2^2 + \ldots + x_{10}^2 - 1 \leqslant 0,$$
$$\ldots$$
$$g_{10}((x_1, \ldots, x_{10})) = x_1^2 + x_2^2 + \ldots + 2x_{10}^2 - 1 \leqslant 0.$$

Мы приведём пример для $n = 10$, начального приближения $x^0 = (1, 1, ..., 1)$ с параметром $\Theta = 3$ при выборе стандартной евклидовой прокс-структуры. Координаты точек $A_k = (a_{1k}, a_{2k}, \ldots, a_{10k})$ при $k = 1, 2, \ldots, 10$ мы выбираем как строки следующей матрицы A:

$$\begin{pmatrix} 1 & 2 & 1 & 4 & 1 & 0 & 4 & 4 & 4 & 3 \\ 2 & 4 & 3 & 1 & 0 & 2 & 4 & 0 & 4 & 0 \\ 3 & 2 & 3 & 4 & 3 & 0 & 3 & 4 & 2 & 3 \\ 0 & 0 & 2 & 0 & 2 & 4 & 4 & 1 & 0 & 0 \\ 3 & 3 & 4 & 4 & 3 & 0 & 1 & 0 & 4 & 4 \\ 2 & 2 & 4 & 0 & 4 & 0 & 2 & 2 & 1 & 1 \\ 0 & 4 & 3 & 4 & 2 & 3 & 3 & 4 & 0 & 2 \\ 2 & 2 & 1 & 4 & 2 & 1 & 4 & 3 & 0 & 3 \\ 4 & 1 & 2 & 2 & 3 & 3 & 2 & 1 & 3 & 1 \\ 3 & 3 & 2 & 2 & 0 & 0 & 4 & 0 & 3 & 4 \end{pmatrix}$$

Отметим также, что возможно некоторое ускорение метода в случае нескольких ограничений за счёт возможности выбора подходящего ограничения на непродуктивных итерациях (см. алгоритм 6 ниже [3]), что видно из таблицы 1 ниже.

Algorithm 6 Модифицированный адаптивный зеркальный спуск

Require: $\varepsilon > 0, \Theta_0 : d(x_*) \leqslant \Theta_0^2$

$x^0 = argmin_{x \in X} \, d(x)$

$I =: \emptyset$

$N \leftarrow 0$

repeat

  if $g(x^N) \leqslant \varepsilon$ **then**

    $h_N \leftarrow \frac{\varepsilon}{||\nabla f(x^N)||_*}$

    $x^{N+1} \leftarrow Mirr_{x^N}(h_N \nabla f(x^N))$ // *"продуктивные шаги"*

    $N \rightarrow I$

  else

    // $(g_{m(N)}(x^N) > \varepsilon)$ *для некоторого* $m(N) \in \{1, \ldots, M\}$

    $h_N \leftarrow \frac{\varepsilon}{||\nabla g_{m(N)}(x^N)||_*^2}$

    $x^{N+1} \leftarrow Mirr_{x^N}(h_N \nabla g_{m(N)}(x^N))$ // *"непродуктивные шаги"*

  end if

  $N \leftarrow N + 1$

until $\Theta_0^2 \leqslant \frac{\varepsilon^2}{2} \left(|I| + \sum\limits_{k \notin I} \frac{1}{||\nabla g_{m(k)}(x^k)||_*^2} \right)$

Ensure: $\bar{x}^N := argmin_{x^k,\, k \in I} \, f(x^k)$

Таблица 1. Сравнение алгоритмов 1, 5 и 6

ε	Итерации	Время, с	Итерации	Время, с	Итерации	Время, с
	Алгоритм 5		Алгоритм 1		Алгоритм 6	
1/2	1659	97	283	15	231	6
1/4	5951	336	899	49	774	22
1/8	22356	1491	3159	180	2850	100

Приведём также сравнение скорости работы методов при тех же параметрах, но уже с негладкими функциональными ограничениями:

$$g_1((x_1, \ldots, x_{10})) = 2|x_1| + |x_2| + |\ldots + x_{10}| - 1 \leqslant 0,$$

$$g_2((x_1,\dots,x_{10})) = |x_1| + 3|x_2| + \ldots + |x_{10}| - 1 \leqslant 0,$$

$$\dots$$

$$g_{10}((x_1,\dots,x_{10})) = |x_1| + |x_2| + \ldots + 11|x_{10}| - 1 \leqslant 0.$$

Таблица 2. Сравнение алгоритмов 1, 5 и 6

ε	Итерации	Время, с	Итерации	Время, с	Итерации	Время, с
	Алгоритм 5		Алгоритм 1		Алгоритм 6	
1/2	3709	279	671	29	437	21
1/4	14212	833	2418	103	1970	95
1/8	54655	2980	8979	455	8329	344

6.2 О преимуществах использования метода с рестартами в сильно выпуклом случае.

Для демонстрации преимуществ алгоритма 3 по сравнению с алгоритмом 1, был проведен ряд численных экспериментов. Рассмотрим различные 1-сильно выпуклые целевые функционалы f, которые имеют липшицев градиент.

- **Пример 1.**

$$f(x) = \frac{L-\mu}{4}\left\{\frac{1}{2}\left[x_1^2 + \sum_{i=1}^{n-1}(x_i - x_{i+1})^2\right] - x_1\right\} + \frac{\mu}{2}\|x\|^2,$$

  где $\mu = 1, L = 10\,000$ и $n = 10$.

- **Пример 2.**

$$f(x) = \max\{f_1(x), f_2(x), f_3(x)\}, \text{где}$$

$$f_1(x) = \frac{1}{2}\left(x_1^2 + x_2^2 + 2x_3^2 + 4x_4^2 + x_5^2 + 5x_6^2 + 3x_7^2 + 2x_8^2 + 4x_9^2 + 8x_{10}^2\right) - \sum_{i=1}^{10} ix_i + 5,$$

$$f_2(x) = \frac{1}{2}\left(2x_1^2 + x_2^2 + 3x_3^2 + 4x_4^2 + 2x_5^2 + 5x_6^2 + x_7^2 + 6x_8^2 + 7x_9^2 + 2x_{10}^2\right) - \sum_{i=11}^{20} ix_i + 6,$$

$$f_3(x) = \frac{1}{2}\left(x_1^2 + x_2^2 + 2x_3^2 + 3x_4^2 + 5x_5^2 + x_6^2 + 4x_7^2 + 2x_8^2 + 3x_9^2 + 6x_{10}^2\right) - \sum_{i=21}^{30} ix_i + 7.$$

- **Пример 3, задача регрессии** [21].

$$f(x) = \frac{1}{2}\|Ax - b\|^2 + \frac{\mu}{2}\|x\|^2, \text{ где}$$

$$A = \begin{pmatrix} 5 & 3 & 3 & 5 & 4 & 4 & 3 & 3 & 5 & 1 \\ 2 & 4 & 3 & 5 & 3 & 4 & 2 & 2 & 5 & 4 \\ 5 & 2 & 1 & 4 & 1 & 1 & 2 & 3 & 5 & 5, \end{pmatrix}$$

  при $b = (1,2,3)^T$, $\mu = 1$.

- **Пример 4.** Рассмотрим функцию следующего вида [21]:

$$f(x) = \sum_{i=1}^{10} ix_i^4 + \frac{1}{2}\|x\|^2$$

- **Пример 5.** Следующий тест выполнен для сглаженной сильно выпуклой версии задачи подавления шумов [21]

$$f(x) = \frac{1}{2}\|Ax-b\|^2 + \lambda\|x\|_{l_1,\tau} + \frac{\mu}{2}\|x\|^2, \text{ где}$$

$$A = \begin{pmatrix} 9 & 2 & 4 & 2 & 2 & 3 & 6 & 3 & 5 & 5 \\ 6 & 7 & 2 & 4 & 8 & 6 & 8 & 8 & 5 & 1 \end{pmatrix}, b = (1,2)^T, \mu = 1, \lambda = 0.05, \tau = 0.0001$$

и $\|.\|_{l_1,\tau}$ задается следующим образом:

$$\|x\|_{l_1,\tau} = \begin{cases} |x| - \dfrac{\tau}{2} & \text{if } \ |x| \geq \tau \\ \dfrac{1}{2\tau}x^2 & \text{if } \ |x| < \tau \end{cases}$$

если x — скаляр и $\|x\|_{l_1,\tau} = \sum_{i=1}^{n} \|x_i\|_{l_1,\tau}$ если $x = (x_1, x_2, ..., x_n)$ — вектор в $\mathbb{R}^n$. Отметим, что квадратичное слагаемое $\frac{\mu}{2}\|x\|^2$ гарантирует сильную выпуклость целевой функции.

Рассмотрим функциональные ограничения вида $g(x) = G(x) + S(x)$, где $S(x) = \frac{1}{2}\|x\|^2$ и $G(x) = \max\limits_{i \in \overline{1,m}} g_i(x)$, так, что $g_i(x) = \langle \alpha_i, x\rangle + \beta_i$, где α_i^T — строки матрицы

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 7 & 8 & 6 & 2 & 9 & 2 & 3 & 3 & 2 & 6 \\ 6 & 3 & 4 & 3 & 5 & 1 & 6 & 3 & 2 & 8 \\ 3 & 5 & 2 & 7 & 8 & 3 & 2 & 1 & 5 & 2 \\ 2 & 3 & 1 & 8 & 1 & 2 & 1 & 1 & 5 & 8 \\ 1 & 8 & 9 & 1 & 3 & 5 & 1 & 3 & 5 & 2 \\ 1 & 7 & 8 & 5 & 5 & 9 & 3 & 1 & 6 & 4 \\ 7 & 3 & 5 & 8 & 9 & 1 & 8 & 7 & 8 & 8 \\ 6 & 4 & 6 & 2 & 9 & 2 & 3 & 1 & 6 & 3 \\ 2 & 3 & 4 & 4 & 2 & 1 & 9 & 1 & 1 & 8 \end{pmatrix}$$

и константы β_i есть нули.

Считаем, что имеется стандартное евклидово расстояние и соответствующая прокс-структура, и

$$X = B_1(0) = \left\{x = (x_1, x_2, \ldots, x_{10}) \in \mathbb{R}^{10} \mid x_1^2 + x_2^2 + ... + x_{10}^2 \leq 1\right\},$$

начальное приближение $x^0 = \frac{(1,1,...,1)}{\|(1,1,...,1)\|}$, $\Theta_0 = 3$, $R_0 = 2$, и точность $\varepsilon = 0.05$.

Результаты выполнения алгоритмов 1 и 3 представлены в таблице 3. Приводится количество итераций и время (указано в минутах и в секундах) работы каждого алгоритма 1 и 3.

Все вычисления были произведены с помощью программного обеспечения Python 3.4, на компьютере оснащенном Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz, 1992 Mhz, 4 Core(s), 8 Logical Processor(s). ОЗУ компьютера составляла 8 ГБ.

Таблица 3. Сравнение результатов работы алгоритмов 1 и 3.

	Итерации	Время	Итерации	Время
	Алгоритм 1		Алгоритм 3	
Пример 1	115 973	09:16	95 447	07:37
Пример 2	57 798	07:01	45 455	05:14
Пример 3	56 874	05:02	50 747	04:18
Пример 4	13 720	01:15	6 764	00:38
Пример 5	64 324	06:04	55 073	04:52

Из таблицы 3 видно, что алгоритм 3 работает быстрее алгоритма 1.

Благодарности. Авторы выражают огромную признательность Юрию Евгеньевичу Нестерову, Александру Владимировичу Гасникову и Павлу Евгеньевичу Двуреченскому за плодотворные обсуждения и комментарии.

Список литературы

[1] A. Bayandina, P. Dvurechensky, A. Gasnikov, F. Stonyakin, A. Titov (2017). Mirror Descent and Convex Optimization Problems With Non-Smooth Inequality Constraints. In LCCC Focus Period on Large-Scale and Distributed Optimization, June 14-16, 2017. Lund, Sweden: Lund Center for Control of Complex Engineering Systems, Lund University.

[2] Fedor S. Stonyakin and Alexander A. Titov. One Mirror Descent Algorithm for Convex Constrained Optimization Problems with Non-Standard Growth Properties. In Proceedings of the School-Seminar on Optimization Problems and their Applications (OPTA-SCL 2018) Omsk, Russia, July 8-14, 2018. CEUR Workshop Proceedings, vol. 2098, pp. 372-384 (2018).

[3] F.S. Stonyakin, M. S. Alkousa, A. N. Stepanov, M. A. Barinov.: Adaptive mirror descent algorithms in convex programming problems with Lipschitz constraints. Trudy Instituta Matematiki i Mekhaniki URO RAN, vol. 24, no. 2, pp. 266 – 279 (2018).

[4] A. Ben-Tal and A. Nemirovski, Lectures on Modern Convex Optimization. Philadelphia: SIAM, 2001.

[5] A. Ben-Tal and A. Nemirovski, Robust Truss Topology Design via Semidefinite Programming, SIAM J. Optim., vol. 7, no. 4, pp. 991–1016, Nov., 1997.

[6] Y. Nesterov. Introductory Lectures on Convex Optimization: a basic course. Kluwer Academic Publishers, Massachusetts, 2004.

[7] Y. Nesterov. Subgradient methods for convex functions with nonstandard growth properties, 2016.

[8] Bayandina, A., Gasnikov, A., Gasnikova, E., Matsievsky, S.: Primal-dual mirror descent for the stochastic programming problems with functional constraints. Computational Mathematics and Mathematical Physics. (Accepted) (2018) https://arxiv.org/pdf/1604.08194.pdf (in Russian)

[9] A. Beck, A. Ben-Tal, N. Guttmann-Beck, and L. Tetruashvili.The comirror algorithm for solving nonsmooth constrained convex problems. Operations Research Letters, 38(6): 493–498, 2010. ISSN: 0167-6377.

[10] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. Oper. Res. Lett., 31(3): 167 – 175, May 2003. ISSN: 0167–6377.

[11] S. Boyd and L. Vandenberghe,Convex Optimization. New York, NY: Cambridge University Press, 2004.

[12] A. Juditsky and A. Nemirovski, First Order Methods for Non-smooth Convex Large-scale Optimization, I: General purpose methods, in Optimization for Machine Learning, S. Sra et al, Eds., Cambridge, MA: MIT Press, 2012, pp. 121–184.

[13] G. Lan, Gradient Sliding for Composite Optimization, Math. Program., vol. 159, no. 1-2, pp. 201–235, 2016.

[14] A.Nemirovskii. Efficient methods for large-scale convex optimization problems. Ekonomika i Matematicheskie Metody, 15, 1979. In Russian.

[15] A. Nemirovsky and D. Yudin. Problem Complexity and Method Efficiency in Optimization. J. Wiley & Sons, New York, 1983.

[16] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Soviet Mathematics Doklady, 27(2): 372–376, 1983.

[17] B. Polyak. A general method of solving extremum problems. Soviet Mathematics Doklady, 8(3): 593–597, 1967.

[18] N. Z. Shor. Generalized gradient descent with application to block programming. Kibernetika, 3(3): 53–55, 1967.

[19] S. Shpirko and Yu. Nesterov,Primal-dual Subgradient Methods for Huge-scale Linear Conic Problem, SIAM Journal on Optimization, no. 24, pp. 1444–1457, 2014.

[20] F. Vasilyev, Optimization Methods. Moscow, Russia: FP, 2002.

[21] Xiangrui Meng and Hao Chen. Accelerating Nesterov's Method for Strongly Convex Functions with Lipschitz Gradient. https://arxiv.org/pdf/1109.6058.pdf