

АДАПТИВНЫЕ АЛГОРИТМЫ ЗЕРКАЛЬНОГО СПУСКА
ДЛЯ ЗАДАЧ ВЫПУКЛОЙ И СИЛЬНО ВЫПУКЛОЙ
ОПТИМИЗАЦИИ С ФУНКЦИОНАЛЬНЫМИ
ОГРАНИЧЕНИЯМИ*)

Ф. С. Стонякин^{1,2,a}, М. Алкуса^{2,b}, А. Н. Степанов^{1,c}, А. А. Титов^{2,d}

¹Крымский федеральный университет им. В. И. Вернадского,
пр. Акад. Вернадского, 4, 295007 Симферополь, Россия

²Московский физико-технический институт (гос. университет),
Институтский пер., 9, 141701 Долгопрудный, Россия

E-mail: ^afedyor@mail.ru, ^bmohammad.alkousa@phystech.edu,
^cstepanov.student@gmail.com, ^da.a.titov@phystech.edu

Аннотация. Рассмотрены некоторые адаптивные алгоритмы зеркального спуска для задач минимизации выпуклого целевого функционала при наличии нескольких выпуклых липшицевых (вообще говоря, негладких) функциональных ограничений. Показано, что методы применимы к целевым функционалам различного уровня гладкости: выполнено условие Липшица либо для самого функционала, либо для его градиента или гессиана (при этом функционал может не удовлетворять свойству Липшица). Главная идея — адаптивная настройка метода на константу Липшица целевого функционала (градиента либо гессиана), а также ограничения. При этом рассмотрено два типа методов: адаптивный (не требует знания констант Липшица ни для целевого функционала, ни для ограничения) и частично адаптивный (требует знания константы Липшица для ограничения). С использованием техники рестартов (перезапусков) методов для задач выпуклой оптимизации предложены методы для задач сильно выпуклой минимизации. Получены оценки скорости сходимости всех рассматриваемых алгоритмов в зависимости от уровня гладкости целевого функционала. Приводятся численные эксперименты, иллюстрирующие для некоторых примеров преимущества предложенных методов. Табл. 3, библиогр. 22.

*) Исследование Ф. С. Стонякина (анализ алгоритмов 1 и 3) выполнено при поддержке гранта Российского фонда фундаментальных исследований (проект № 18–31–00219 мол_а).

Ключевые слова: адаптивный метод зеркального спуска, условие Липшица, липшицев градиент, липшицев гессиан, сильно выпуклая функция, техника рестартов.

Введение

Задачи минимизации выпуклых гладких и негладких функционалов с ограничениями возникают во многих проблемах современной large-scale оптимизации и её приложениях [1, 2]. Для таких задач известно множество методов, среди которых можно выделить метод уровней [3] и метод штрафных функций [4–6]. Метод зеркального спуска (МЗС) [7–9] 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 функционалами ограничений.

Отметим, что функционалы ограничений, вообще говоря, могут быть негладкими (недифференцируемыми) и поэтому естественно рассматривать субградиентные методы. Методы с использованием субградиентов негладких выпуклых функций активно разрабатываются уже несколько десятилетий. Эти исследования восходят к известным пионерским работам, одна из которых посвящена градиентному методу для безусловных задач при евклидовом расстоянии [10], а другая — его обобщению для задач с ограничениями [11]. В работе [11] предложена идея переключения шагов между направлением субградиента целевого функционала и направлением субградиента ограничения. Обобщение метода градиентного спуска для задач с неевклидовым расстоянием называют *методом зеркального спуска*. Этот метод был предложен в [8, 9] для задач без ограничений (см. также [7]). Зеркальный спуск для задач с функциональными ограничениями был предложен в [9] (см. также [12]). При этом, как правило, для нахождения величины шага и критерия остановки для зеркального спуска необходимо знать величину константы Липшица целевого функционала, а также функционала ограничения. Известны также методы с адаптивным выбором шага, рассмотренные в [13] для задач без ограничений, а в [12] — для задач с ограничениями. Недавно в [14] были предложены оптимальные с точки зрения нижних оракульных оценок алгоритмы зеркального спуска для задач выпуклого программирования с липшицевыми функционалами ограничений с адаптивным выбором шага, а также адаптивными критериями остановки. Модификации этих методов для задач в случае нескольких выпуклых функциональных ограничений были проанализированы в [15, 16].

В настоящей статье представлены некоторые алгоритмы зеркального спуска для задач минимизации выпуклого функционала f на некотором

выпуклом замкнутом множестве с ограничением, порождённым выпуклым, липшицевым и негладким функционалом $g(x) \leq 0$. Важно, что получены оценки скорости сходимости методов для целевых функционалов различного уровня 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 $x \in X$, где X — область определения задачи) и целевой функционал имеет вид

$$f(x) = \max_{i=\overline{1, m}} f_i(x), \quad (1)$$

где

$$f_i(x) = \frac{1}{2} \langle A_i x, x \rangle - \langle b_i, x \rangle + c_i, \quad i = \overline{1, m}, \quad (2)$$

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)–(2) возникают в задачах проектирования механических конструкций Truss Topology Design со взвешенными балками [17]. Для задач минимизации функционалов такого типа при наличии выпуклых липшицевых ограничений в [14, 15, 18] на базе методики работ Ю. Е. Нестерова [3, 17] были предложены некоторые новые адаптивные алгоритмы зеркального спуска, а также обоснована их оптимальность. Часть этих результатов опубликована в [18]. Настоящая статья посвящена изложению основных результатов доклада [18], а также развитию результатов [14, 15, 18] в следующих направлениях.

Во-первых, доказывалась оптимальность с точки зрения нижних оркульных оценок предложенных методов в [14, 15, 18] 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 [14, 18] (для выпуклых задач) предложены новые алгоритмы зеркального спуска аналогично для задач минимизации μ -сильно выпуклого функционала f с неположительным, μ -сильно выпуклым и липшицевым негладким функционалом ограничения g . Заметим, что техника рестартов (перезапусков) метода для выпуклых задач с целью ускорения сходимости для сильно выпуклых задач восходит к 1980-м гг. (см. [9, 19]). Техника такого типа была использована в [20] 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 метод, также оптимальный с точки зрения нижних оракульных оценок на классе задач с липшицевым целевым функционалом [14, п. 3.1]. Приведены расчёты, иллюстрирующие некоторые преимущества предлагаемых нами методов для сильно выпуклых задач.

Статья состоит из введения и четырёх основных разделов. В разд. 1 приводим некоторые вспомогательные сведения, основные понятия для метода зеркального спуска, а также описываем адаптивный алгоритм 1 зеркального спуска из [14, п. 3.3] и частично адаптивный алгоритм 2 [18]. В разд. 2 рассматриваем новые утверждения об оценках скорости сходимости данных методов и обосновываем их оптимальность на рассматриваемом классе задач при различных допущениях на уровень гладкости целевого функционала. Разд. 3 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 разд. 4 приводим и обсуждаем результаты некоторых численных экспериментов, иллюстрирующих определённые преимущества предлагаемых нами методов.

1. Адаптивный и частично адаптивный алгоритм зеркального спуска для задач с выпуклыми функционалами

Начнём с постановки рассматриваемых задач оптимизации и приведём необходимые понятия и результаты. Пусть $(E, \|\cdot\|)$ — конечномерное нормированное векторное пространство и E^* — сопряжённое к E пространство 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 $x \in E$.

Всюду далее будем считать, что $X \subset E$ есть замкнутое выпуклое множество. Рассмотрим два выпуклых субдифференцируемых функционала $f, g: X \rightarrow \mathbb{R}$. Предположим, что функционал g удовлетворяет условию Липшица относительно нормы $\|\cdot\|$, т. е. существует $M_g > 0$ такое, что

$$|g(x) - g(y)| \leq M_g \|x - y\| \quad (3)$$

для всех $x, y \in X$. Это означает, что в любой точке $x \in X$ существует

субградиент $\nabla g(x)$, причём $\|\nabla g(x)\|_* \leq M_g$. Напомним, что для дифференцируемого функционала g субградиент $\nabla g(x)$ совпадает с обычным градиентом.

В настоящей работе будем рассматривать следующий тип задач оптимизации:

$$f(x) \rightarrow \min, \quad x \in X, \quad (4)$$

$$g(x) \leq 0, \quad (5)$$

если f и g удовлетворяют упомянутым условиям. Сделаем предположение о разрешимости задачи (4)–(5). Обозначим через x_* точное решение задачи (4)–(5). Наша цель — предложить метод, который позволяет за конечное число шагов найти ε -решение $\hat{x} \in X$ поставленной задачи:

$$f(\hat{x}) - f(x_*) \leq \varepsilon \quad \text{при} \quad g(\hat{x}) \leq \varepsilon.$$

Всюду далее предполагаем, что начальное приближение x^0 для всех методов выбирается так, что $d(x^0) = \min_{x \in X} d(x)$.

Отметим, что часть результатов работы (разд. 3) посвящена постановке задачи для μ -сильно выпуклых субдифференцируемых функционалов $f, g: X \rightarrow \mathbb{R}$, т. е. для произвольных $x, y \in X$ имеет место неравенство

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2 \quad (6)$$

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

Для дальнейших рассуждений потребуются вспомогательные понятия (см., например, [13]). Введём так называемую *прокс-функцию* $d: X \rightarrow \mathbb{R}$, обладающую свойством непрерывной дифференцируемости и 1-сильной выпуклости относительно нормы $\|\cdot\|$. Пусть существует такая константа $\Theta_0 > 0$, что $d(x_*) \leq \Theta_0^2$. Отметим, что если имеется множество решений X_* , то 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, [13]). Стандартно определим оператор проектирования:

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

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

Напомним одно известное утверждение (см., например, [13]).

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

$$h\langle \nabla f(y), y - x \rangle \leq \frac{h^2}{2} \|\nabla f(y)\|_*^2 + V(y, x) - V(z, x). \quad (7)$$

Перейдём к описанию рассматриваемых методов [14, 18] для задачи (4)–(5). Отметим, что необходимо найти подходящую точку \hat{x} , для которой $f(\hat{x}) - f(x_*) \leq \varepsilon$ (или $C \cdot \varepsilon$ для некоторой константы $C > 0$) при условии $g(x) \leq \varepsilon$.

Как видно из листингов алгоритмов, приведённых в данном разделе, эта точка выбирается среди точек x^k , для которых $g(x^k) \leq \varepsilon$. Поэтому будем называть *продуктивными* шаги, для которых $g(x^k) \leq \varepsilon$. Если выполнено обратное неравенство $g(x^k) > \varepsilon$, то такой шаг k будем называть *непродуктивным*. Напомним следующий алгоритм адаптивного зеркального спуска для задачи (4)–(5) из [14, п. 3.3].

Алгоритм 1 (адаптивный зеркальный спуск)

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

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

```

1:  $I := \emptyset$ 
2:  $N \leftarrow 0$ 
3: repeat
4:   if  $g(x^N) \leq \varepsilon$  then
5:      $h_N \leftarrow \frac{\varepsilon}{\|\nabla f(x^N)\|_*}$ 
6:      $x^{N+1} \leftarrow \text{Mirr}_{x^N}(h_N \nabla f(x^N))$  ▷ «продуктивные шаги»
7:      $N \rightarrow I$ 
8:   else
9:      $(g(x^N) > \varepsilon) \rightarrow$ 
10:     $h_N \leftarrow \frac{\varepsilon}{\|\nabla g(x^N)\|_*^2}$ 
11:     $x^{N+1} \leftarrow \text{Mirr}_{x^N}(h_N \nabla g(x^N))$  ▷ «непродуктивные шаги»
12:   end if
13:    $N \leftarrow N + 1$ 
14: until  $\Theta_0^2 \leq \frac{\varepsilon^2}{2} \left( |I| + \sum_{k \notin I} \frac{1}{\|\nabla g(x^k)\|_*^2} \right)$ 
```

Аналогично [3] введём для целевого функционала f в точке $y \in X$ следующую вспомогательную величину:

$$v_f(x, y) = \begin{cases} \left\langle \frac{\nabla f(x)}{\|\nabla f(x)\|_*}, x - y \right\rangle, & \nabla f(x) \neq 0, \\ 0, & \nabla f(x) = 0, \end{cases} \quad x \in X. \quad (8)$$

Допускаем, что в ходе работы метода можно использовать произвольный субградиент $\nabla f(x)$.

Для оценки скорости сходимости алгоритма 1 в [14] получена

Теорема 1. Пусть верно неравенство (3) и известна константа $\Theta_0 > 0$ такая, что $d(x_*) \leq \Theta_0^2$. Если $\varepsilon > 0$ — фиксированное число, то алгоритм 1 работает не более чем

$$N = \left\lceil \frac{2 \max\{1, M_g^2\} \Theta_0^2}{\varepsilon^2} \right\rceil \quad (9)$$

итераций, причём после его остановки справедливо неравенство

$$\min_{k \in I} v_f(x^k, x_*) < \varepsilon. \quad (10)$$

Можно предложить частично адаптивный алгоритм для задачи (4)–(5) [18]. Его отличие от алгоритма 1 в том, что адаптивно выбирается шаг лишь на продуктивных итерациях и критерий остановки неадаптивен.

Алгоритм 2 (частично адаптивная версия алгоритма 1)

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

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

```

1:  $x^0 = \arg \min_{x \in X} d(x)$ 
2:  $I := \emptyset$ 
3:  $N \leftarrow 0$ 
4: repeat
5:   if  $g(x^N) \leq \varepsilon$  then
6:      $h_N \leftarrow \frac{\varepsilon}{M_g \cdot \|\nabla f(x^N)\|_*}$ 
7:      $x^{N+1} \leftarrow \text{Mirr}_{x^N}(h_N \nabla f(x^N))$  ▷ «продуктивные шаги»
8:      $N \rightarrow I$ 
9:   else
10:     $(g(x^N) > \varepsilon) \rightarrow$ 
11:     $h_N \leftarrow \frac{\varepsilon}{M_g^2}$ 
12:     $x^{N+1} \leftarrow \text{Mirr}_{x^N}(h_N \nabla g(x^N))$  ▷ «непродуктивные шаги»
13:   end if
14:    $N \leftarrow N + 1$ 
15: until  $N \geq \left\lceil \frac{2M_g^2 \Theta_0^2}{\varepsilon^2} \right\rceil$ 
```

Справедлив следующий аналог теоремы 1 (см. также [18]).

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

$$N = \left\lceil \frac{2M_g^2\Theta_0^2}{\varepsilon^2} \right\rceil \quad (11)$$

итераций. Тогда

$$\min_{k \in I} v_f(x^k, x_*) < \frac{\varepsilon}{M_g}. \quad (12)$$

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

$$h_k(g(x^k) - g(x_*)) \leq \frac{h_k^2}{2} \|\nabla g(x^k)\|_*^2 + V(x^k, x_*) - V(x^{k+1}, x_*). \quad (13)$$

Если $\alpha_k = 0$ или $\alpha_k \rightarrow 0$, то не можем использовать неравенство (13) для

$$h_k = \frac{\varepsilon}{\|\nabla g(x^k)\|_*^2},$$

поскольку это может привести к большой погрешности. В таком случае неадаптивный выбор шага

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

в алгоритме 2 — более подходящий вариант для решения задачи (4)–(5).

2. Оценки скорости сходимости рассмотренных методов и их оптимальность

Рассмотрим конкретные оценки скорости сходимости рассмотренных методов, которые обосновывают их оптимальность с точки зрения теории оракульных оценок, восходящей к известной монографии А. С. Немировского и Д. Б. Юдина [13]. Функционалы ограничений предполагаются липшицевыми и, вообще говоря, негладкими. Поэтому для установления оптимальности метода с точки зрения нижних оракульных оценок нужно показать [13], что для достижения требуемой точности ε решения задачи (4)–(5) достаточно $O(\frac{1}{\varepsilon^2})$ итераций метода, в результате выполнения которых происходит вычисление (суб)градиента целевого функционала или ограничения. При этом в работе рассмотрены различные классы целевых функционалов. Будем использовать следующее вспомогательное

утверждение (см., например, [3, 17]). По-прежнему, через x_* обозначаем решение задачи (4)–(5).

Лемма 2. Введём функцию

$$\omega(\tau) = \max_{x \in X} \{f(x) - f(x_*) \mid \|x - x_*\| \leq \tau\}, \quad (14)$$

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

$$f(y) - f(x_*) \leq \omega(v_f(y, x_*)). \quad (15)$$

На базе леммы 2 в [18] показано, как с использованием теоремы 2 можно оценить скорость сходимости алгоритма 2, если целевой функционал f дифференцируем и его градиент удовлетворяет условию Липшица:

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\| \quad \forall x, y \in X. \quad (16)$$

Тогда для произвольного $x \in X$ верно неравенство [3]

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

откуда

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

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

Последнее неравенство позволяет сформулировать [18] для некоторого класса, вообще говоря, негладких целевых функционалов

Следствие 1. Предположим, что $f(x) = \max_{i=1, m} f_i(x)$, где f_i дифференцируемы в каждой точке $x \in X$ 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 будет выполнена следующая оценка:

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

где $L = \max_{i=1, m} L_i$.

Замечание 2. Мы рассматриваем условные задачи, поэтому совсем не обязательно, что $\|\nabla f(x_*)\|_* = 0$.

Опишем оценки скорости сходимости алгоритмов 1 и 2 для классов целевых функционалов, не рассмотренных в [18].

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

$$|f(x) - f(y)| \leq M_f \|x - y\| \quad \forall x, y \in X. \quad (17)$$

Тогда для произвольного $x \in X$ верно неравенство

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

откуда

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

Поэтому имеет место

Следствие 2. Пусть f удовлетворяет условию Липшица (17) на X . Тогда

1) после

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

шагов работы алгоритма 1 будет верно неравенство

$$\min_{k \in I} f(x^k) - f(x_*) \leq M_f \varepsilon;$$

2) после

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

шагов работы алгоритма 2 будет справедлива оценка

$$\min_{k \in I} f(x^k) - f(x_*) \leq \frac{M_f}{M_g} \varepsilon.$$

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

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_* \leq L \|x - y\| \quad \forall x, y \in X. \quad (18)$$

В таком случае для всякого $x \in X$ верно следующее неравенство (см. [3, лемма 1.2.4]):

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

откуда

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

Поэтому для произвольного $x \in X$

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

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

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

Комбинируя утверждения теоремы 1 и леммы 2, получаем

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

и аналогично — из теоремы 2

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

Поэтому справедливо

Следствие 3. Пусть f дважды дифференцируем на X и имеет липшицев гессиан, т. е. верно (18). Тогда

1) после

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

шагов работы алгоритма 1 будет выполнено неравенство

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

2) после

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

шагов работы алгоритма 2 будет верно неравенство

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

Аналогичные оценки можно выписать и для некоторого класса задач с негладкими целевыми функционалами.

Следствие 4. Предположим, что $f(x) = \max_{i=1, m} f_i(x)$, 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

1) после

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

шагов работы алгоритма 1 будет справедливо неравенство

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

где $L = \max_{i=1, m} L_i$;

2) после

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

шагов работы алгоритма 2 будет выполнена следующая оценка:

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

где $L = \max_{i=1, m} L_i$.

3. Об ускорении рассматриваемых методов зеркального спуска для сильно выпуклых задач

В этом разделе рассмотрим задачу

$$f(x) \rightarrow \min, \quad g(x) \leq 0, \quad x \in X, \quad (19)$$

в предположениях (3), а также сильной выпуклости f и g с одинаковым параметром $\mu > 0$. Слегка модифицируем предположения на прокс-функцию и допустим, что $d(x)$ ограничена на единичном шаре относительно выбранной нормы $\|\cdot\|$:

$$d(x) \leq \Theta_0^2 \quad \forall x \in X: \|x\| \leq 1. \quad (20)$$

Также предположим, что для начальной точки $x^0 \in X$ существует такое $R_0 > 0$, что $\|x^0 - x_*\|^2 \leq R_0^2$.

Будем рассматривать методы для нахождения ε -решения \hat{x} поставленной задачи (19):

$$f(\hat{x}) - f(x_*) \leq \varepsilon, \quad g(\hat{x}) \leq \varepsilon.$$

Для построения методов решения задачи (19) при заданных предположениях используем идею рестартов (перезапусков) алгоритмов 1 и 2. Рассмотрим вспомогательное утверждение (см., например, [21]).

Лемма 3. Пусть f и g — μ -сильно выпуклые функционалы относительно нормы $\|\cdot\|$ на X , $x_* = \arg \min_{x \in X} f(x)$, $g(x) \leq 0$ при $x \in X$ и для некоторых $\varepsilon_f > 0$ и $\varepsilon_g > 0$ верно

$$f(x) - f(x_*) \leq \varepsilon_f, \quad g(x) \leq \varepsilon_g. \quad (21)$$

Тогда

$$\frac{\mu}{2} \|x - x_*\|^2 \leq \max\{\varepsilon_f, \varepsilon_g\}. \quad (22)$$

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

$$\|\nabla f_i(x) - \nabla f_i(y)\|_* \leq L_i \|x - y\| \quad \forall x, y \in X. \quad (23)$$

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

$$\tau(\delta) = \max \left\{ \delta \|\nabla f(x_*)\|_* + \frac{\delta^2 L}{2}, \delta \right\}, \quad (24)$$

где

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

Ясно, что τ возрастает, $\tau(0) = 0$, поэтому для всякого $\varepsilon > 0$ существует $\widehat{\varphi}(\varepsilon) > 0$ такая, что $\tau(\widehat{\varphi}(\varepsilon)) = \varepsilon$.

Предложим адаптивный алгоритм 3 для задачи (19).

Алгоритм 3 (адаптивный алгоритм зеркального спуска для сильно выпуклых функционалов)

Вход: точность $\varepsilon > 0$; начальная точка x^0 ; Θ_0 такое, что $d(x) \leq \Theta_0^2$ $\forall x \in X$: $\|x\| \leq 1$; X ; $d(\cdot)$; параметр сильной выпуклости μ ; R_0 удовлетворяет оценке $\|x^0 - x_*\|^2 \leq R_0^2$.

- 1: Set $d_0(x) = d(\frac{x-x^0}{R_0})$.
 - 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 $\widehat{\varphi}(\varepsilon_p)$, прокс-
функцией $d_{p-1}(\cdot)$ и Θ_0^2
 - 7: $d_p(x) \leftarrow d(\frac{x-x^p}{R_p})$.
 - 8: Set $p = p + 1$.
 - 9: **until** $p > \log_2 \frac{\mu R_0^2}{2\varepsilon}$
-

Теорема 3. Пусть f и g — μ -сильно выпуклые на $X \subset \mathbb{R}^n$ функционалы, f имеет липшицев градиент, удовлетворяющий (23), 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 $\hat{p} = \lceil \log_2 \frac{\mu R_0^2}{2\varepsilon} \rceil$ выход $x^{\hat{p}}$ есть ε -решение задачи (19), а также верно неравенство

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

При этом общее число итераций алгоритма 1 при работе алгоритма 3 не превышает

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

ДОКАЗАТЕЛЬСТВО. Функция $d_p(x) = d(\frac{x-x^p}{R_p})$, определённая в алгоритме 3, является 1-сильно выпуклой функцией относительно нормы $\frac{\|\cdot\|}{R_p}$ при всех $p \geq 0$. Методом математической индукции можно доказать, что

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

Для $p = 0$ это утверждение очевидно ввиду выбора x^0 и R_0 . Предположим, что для некоторого p верно $\|x^p - x_*\|^2 \leq R_p^2$. Докажем, что $\|x^{p+1} - x_*\|^2 \leq R_{p+1}^2$. Поскольку $d_p(x_*) \leq \Theta_0^2$, согласно 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 для $x^{p+1} = \bar{x}^{N_{p+1}}$ верны следующие неравенства:

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

Тогда по лемме 3

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

Итак, для всякого $p \geq 0$ доказано, что

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

Поэтому при $p = \hat{p} = \lceil \log_2 \frac{\mu R_0^2}{2\varepsilon} \rceil$ выход $x^{\hat{p}}$ есть ε -решение задачи (19) и справедливы следующие неравенства:

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

Пусть K — общее число итераций алгоритма 1 при работе алгоритма 3 согласно п. 6 листинга, а N_p — общее количество итераций алгоритма 1 на p -м рестарте. Вспомним, что функция $\tau: \mathbb{R}^+ \rightarrow \mathbb{R}^+$ возрастает и для каждого $\varepsilon > 0$ существует $\widehat{\varphi}(\varepsilon) > 0$ такая, что $\tau(\widehat{\varphi}(\varepsilon)) = \varepsilon$. Стало быть,

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

Теорема 3 доказана.

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

$$k_{p+1} = \left\lceil \frac{2\Theta_0^2 \max\{1, M_g^2\} R_p^2}{\widehat{C}^2 \varepsilon_{p+1}^2} \right\rceil \quad (25)$$

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, для 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 для любых $p \geq 1$ верны 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 $(\max\{1, \widehat{C}\}\varepsilon)$ -решением рассматриваемой задачи, при этом

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

Оценим общее число N итераций алгоритма 1. Пусть $\hat{p} = \lceil \log_2 \frac{\mu R_0^2}{2\varepsilon} \rceil$. Тогда согласно (25) с точностью до умножения на константу имеем

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

Рассмотрим также частично адаптивную версию алгоритма 3 для задачи (19), предложенную в [18].

Алгоритм 4 (частично адаптивный алгоритм зеркального спуска для сильно выпуклых функционалов)

Вход: точность $\varepsilon > 0$; начальная точка x^0 ; Θ_0 такое, что $d(x) \leq \Theta_0^2$ $\forall x \in X$: $\|x\| \leq 1$; X ; $d(\cdot)$; параметр сильно выпуклости μ ; R_0 удовлетворяет оценке $\|x^0 - x_*\|^2 \leq R_0^2$.

- 1: Set $d_0(x) = d(\frac{x-x^0}{R_0})$
 - 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 $\varphi(\varepsilon_p)$, прокс-функцией $d_{p-1}(\cdot)$ и Θ_0^2 .
 - 7: $d_p(x) \leftarrow d(\frac{x-x^p}{R_p})$
 - 8: Set $p = p + 1$.
 - 9: **until** $p > \log_2 \frac{\mu R_0^2}{2\varepsilon}$
-

В условиях следствия 1 после остановки алгоритма 4 будут выполняться неравенства (21) при

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

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

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

Ясно, что τ возрастает, $\tau(0) = 0$, поэтому для всякого $\varepsilon > 0$ существует $\varphi(\varepsilon) > 0$ такая, что $\tau(\varphi(\varepsilon)) = \varepsilon$.

Теорема 4 [18]. Пусть f и g — μ -сильно выпуклые на $X \subset \mathbb{R}^n$ функционалы, удовлетворяющие условиям следствия 1, 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 $\hat{p} = \left\lceil \log_2 \frac{\mu R_0^2}{2\varepsilon} \right\rceil$ выход $x^{\hat{p}}$ есть ε -решение задачи (19) и верно неравенство

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

При этом общее число итераций алгоритма 2 при работе алгоритма 4 не превышает

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

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

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

Для другого случая ($\|\nabla f(x_*)\|_* > M_g$) для всех $\varepsilon > 0$

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

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

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

Замечание 8. В силу следствий 2 и 4 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 .

4. Численные эксперименты

4.1. Сравнение скорости работы методов для задачи Ферма — Торричелли — Штейнера с ограничениями. Отметим, что в [14, п. 3.1] предложен также следующий адаптивный метод, оптимальный с точки зрения нижних оракульных оценок в случае задач с липшицевым целевым функционалом.

Алгоритм 5 (адаптивный зеркальный спуск (липшицев целевой функционал))

Вход: $\varepsilon > 0$; Θ_0 такое, что $d(x_*) \leq \Theta_0^2$.

Выход: $\bar{x}^N := \frac{\sum_{k \in I} x^k h_k}{\sum_{k \in I} h_k}$

```

1:  $x^0 = \arg \min_{x \in X} d(x)$ 
2:  $I := \emptyset$ 
3:  $N \leftarrow 0$ 
4: repeat
5:   if  $g(x^N) \leq \varepsilon$  then
6:      $M_N = \|\nabla f(x^N)\|_*$ ,  $h_N = \frac{\varepsilon}{M_N^2}$ 
7:      $x^{N+1} = \text{Mirr}_{x^N}(h_N \nabla f(x^N))$  ▷ «продуктивные шаги»
8:      $N \rightarrow I$ 
9:   else
10:     $M_N = \|\nabla g(x^N)\|_*$ ,  $h_N = \frac{\varepsilon}{M_N^2}$ 
11:     $x^{N+1} = \text{Mirr}_{x^N}(h_N \nabla g(x^N))$  ▷ «непродуктивные шаги»
12:   end if
13:    $N \leftarrow N + 1$ 
14: until  $\sum_{j=0}^{N-1} \frac{1}{M_j^2} \geq 2 \frac{\Theta_0^2}{\varepsilon^2}$ 

```

В настоящей работе рассматривается альтернативный метод (алгоритм 1), оптимальность которого 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 $A_k = (a_{1k}, a_{2k}, \dots, a_{10k})$ в 10-мерном евклидовом пространстве \mathbb{R}^{10} необходимо найти точку $x = (x_1, \dots, x_{10})$ такую, что целевая функция

$$f(x) := \sum_{k=1}^{10} \sqrt{(x_1 - a_{1k})^2 + (x_2 - a_{2k})^2 + \dots + (x_{10} - a_{10k})^2}$$

принимает наименьшее значение на множестве X , которое задаётся следующими ограничениями:

$$g_1(x_1, \dots, x_{10}) = 2x_1^2 + x_2^2 + \dots + x_{10}^2 - 1 \leq 0,$$

$$g_2(x_1, \dots, x_{10}) = x_1^2 + 2x_2^2 + \dots + x_{10}^2 - 1 \leq 0,$$

...

$$g_{10}(x_1, \dots, x_{10}) = x_1^2 + x_2^2 + \dots + 2x_{10}^2 - 1 \leq 0.$$

Для $n = 10$ приведём пример начального приближения $x^0 = (1, 1, \dots, 1)$ с параметром $\Theta_0 = 3$ при выборе стандартной евклидовой прокс-структуры. Координаты точек $A_k = (a_{1k}, a_{2k}, \dots, a_{10k})$ при $k = 1, 2, \dots, 10$ выбираем как строки следующей матрицы A :

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

Отметим также, что возможно некоторое ускорение работы алгоритма 1 в случае нескольких ограничений за счёт возможности выбора подходящего ограничения на непродуктивных итерациях (см. алгоритм 6, предложенный в [15]). Соответствующие результаты приведены в табл. 1.

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

В табл. 2 приведём сравнение скорости работы методов при тех же параметрах, но уже с негладкими функциональными ограничениями:

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

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

...

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

Алгоритм 6 (модифицированный адаптивный зеркальный спуск)**Вход:** $\varepsilon > 0$; Θ_0 такое, что $d(x_*) \leq \Theta_0^2$.**Выход:** $\bar{x}^N := \arg \min_{x^k, k \in I} f(x^k)$.

```

1:  $x^0 = \arg \min_{x \in X} d(x)$ 
2:  $I := \emptyset$ 
3:  $N \leftarrow 0$ 
4: repeat
5:   if  $g(x^N) \leq \varepsilon$  then
6:      $h_N \leftarrow \frac{\varepsilon}{\|\nabla f(x^N)\|_*}$ 
7:      $x^{N+1} \leftarrow \text{Mirr}_{x^N}(h_N \nabla f(x^N))$  ▷ «продуктивные шаги»
8:      $N \rightarrow I$ 
9:   else ▷  $(g_{m(N)}(x^N) > \varepsilon)$  для некоторого  $m(N) \in \{1, \dots, M\}$ 
10:     $h_N \leftarrow \frac{\varepsilon}{\|\nabla g_{m(N)}(x^N)\|_*^2}$ 
11:     $x^{N+1} \leftarrow \text{Mirr}_{x^N}(h_N \nabla g_{m(N)}(x^N))$  ▷ «непродуктивные шаги»
12:   end if
13:    $N \leftarrow N + 1$ 
14: until  $\Theta_0^2 \leq \frac{\varepsilon^2}{2} \left( |I| + \sum_{k \notin I} \frac{1}{\|\nabla g_{m(k)}(x^k)\|_*^2} \right)$ 

```

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

4.2. О преимуществах использования метода с рестартами в сильно выпуклом случае. Для демонстрации преимуществ алгоритма 3 по сравнению с алгоритмом 1 был проведён 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. $f(x) = \max\{f_1(x), f_2(x), f_3(x)\}$, где

$$\begin{aligned} f_1(x) &= \frac{1}{2}(x_1^2 + x_2^2 + 2x_3^2 + 4x_4^2 + x_5^2 + 5x_6^2 + 3x_7^2 + 2x_8^2 + 4x_9^2 + 8x_{10}^2) \\ &\quad - \sum_{i=1}^{10} ix_i + 5, \\ f_2(x) &= \frac{1}{2}(2x_1^2 + x_2^2 + 3x_3^2 + 4x_4^2 + 2x_5^2 + 5x_6^2 + x_7^2 + 6x_8^2 + 7x_9^2 + 2x_{10}^2) \\ &\quad - \sum_{i=1}^{10} (10+i)x_i + 6, \\ f_3(x) &= \frac{1}{2}(x_1^2 + x_2^2 + 2x_3^2 + 3x_4^2 + 5x_5^2 + x_6^2 + 4x_7^2 + 2x_8^2 + 3x_9^2 + 6x_{10}^2) \\ &\quad - \sum_{i=1}^{10} (20+i)x_i + 7. \end{aligned}$$

Пример 3 (задача регрессии [22]).

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

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

при $b = (1, 2, 3)^T$, $\mu = 1$.

Пример 4 [22].

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

Пример 5 [22]. Следующий тест выполнен для сглаженной сильно выпуклой версии задачи подавления шумов

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

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

$$b = (1, 2)^T, \quad \mu = 1, \quad \lambda = 0,05, \quad \tau = 0,0001$$

и $\|\cdot\|_{l_1, \tau}$ задаётся следующим образом:

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

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

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

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

а константы β_i суть нули.

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

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

начальное приближение $x^0 = \frac{(1,1,\dots,1)}{\|(1,1,\dots,1)\|}$, $\Theta_0 = 3$, $R_0 = 2$ и точность ε равна 0,05.

Таблица 3

Сравнение результатов работы алгоритмов 1 и 3

	Итерации	Время	Итерации	Время
	Алгоритм 1		Алгоритм 3	
Пример 1	115 973	9:16	95 447	7:37
Пример 2	57 798	7:01	45 455	5:14
Пример 3	56 874	5:02	50 747	4:18
Пример 4	13 720	1:15	6 764	0:38
Пример 5	64 324	6:04	55 073	4:52

Результаты выполнения алгоритмов 1 и 3 представлены в табл. 3, где приводится количество итераций и время работы (в минутах и секундах) каждого из алгоритмов 1 и 3.

Все вычисления были произведены с помощью программного обеспечения Python 3.4 на компьютере, оснащённом Intel(R) Core(TM) i7-8550U CPU @ 1,80 GHz, 1992 MHz, 4 Core(s), 8 Logical Processor(s). ОЗУ компьютера составляло 8 ГБ.

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

Авторы выражают огромную признательность Ю. Е. Нестерову, А. В. Гасникову и П. Е. Двуреченскому за плодотворные обсуждения и комментарии.

ЛИТЕРАТУРА

1. **Ben-Tal A., Nemirovski A.** Robust truss topology design via semidefinite programming // *SIAM J. Optim.* 1997. Vol. 7, No. 4. P. 991–1016.
2. **Nesterov Yu., Shpirko S.** Primal-dual subgradient methods for huge-scale linear conic problem // *SIAM J. Optim.* 2014. Vol. 24, No. 3. P. 1444–1457.
3. **Nesterov Yu.** Introductory lectures on convex optimization: A basic course. Dordrecht: Kluwer Acad. Publ., 2004 (Appl. Optim.; Vol. 87).
4. **Васильев Ф. П.** Методы оптимизации. Кн. 1. М.: МЦНМО, 2011. 624 с.
5. **Васильев Ф. П.** Методы оптимизации. Кн. 2. М.: МЦНМО, 2011. 434 с.
6. **Lan G.** Gradient sliding for composite optimization // *Math. Program.* 2016. Vol. 159, No. 1–2. P. 201–235.
7. **Beck A., Teboulle M.** Mirror descent and nonlinear projected subgradient methods for convex optimization // *Oper. Res. Lett.* 2003. Vol. 31, No. 3. P. 167–175.
8. **Немировский А. С., Юдин Д. Б.** Эффективные методы решения задач выпуклого программирования большой размерности // *Экономика и мат. методы.* 1979. № 2. С. 135–152.
9. **Немировский А. С., Юдин Д. Б.** Сложность задач и эффективность методов оптимизации. М.: Наука, 1979. 384 с.
10. **Шор Н. З.** Применение обобщённого градиентного спуска в блочном программировании // *Кибернетика.* 1967. № 3. С. 53–55.
11. **Polyak B. T.** A general method of solving extremum problems // *Sov. Math. Dokl.* 1967. Vol. 8, No. 3. P. 593–597.
12. **Beck A., Ben-Tal A., Guttman-Beck N., Tetruashvili L.** The CoMirror algorithm for solving nonsmooth constrained convex problems // *Oper. Res. Lett.* 2010. Vol. 38, No. 6. P. 493–498.
13. **Ben-Tal A., Nemirovski A.** Lectures on modern convex optimization. Philadelphia, PA: SIAM, 2001.
14. **Bayandina A., Dvurechensky P., Gasnikov A., Stonyakin F., Titov A.** Mirror descent and convex optimization problems with non-smooth inequality constraints // *Large-Scale and Distributed Optimization.* Cham: Springer, 2018. P. 181–213 (Lect. Notes Math.; Vol. 2227).
15. **Стонякин Ф. С., Алкуза М., Степанов А. Н., Баринов М. А.** Адаптивные алгоритмы зеркального спуска в задачах выпуклого программирования с липшицевыми ограничениями // *Тр. Ин-та математики и механики УрО РАН.* 2018. Т. 24, № 2. С. 266–279.

16. **Titov A. A., Stonyakin F. S., Gasnikov A. V., Alkousa M. S.** Mirror descent and constrained online optimization problems // Optimization and Applications : Rev. Sel. Pap. 9th Int. Conf. OPTIMA-2018 (Petrovac, Montenegro, Oct. 1–5, 2018). Cham: Springer, 2019. P. 64–78 (Commun. Comput. Inform. Sci.; Vol. 974).
17. **Nesterov Yu.** Subgradient methods for convex functions with nonstandard growth properties. 2016. Available at http://mathnet.ru/PresentFiles/16179/growthbm_nesterov.pdf.
18. **Stonyakin F. S., Titov A. A.** One mirror descent algorithm for convex constrained optimization problems with non-standard growth properties // Proc. School-Seminar on Optimization Problems and Their Applications (Omsk, Russia, July 8–14, 2018). Aachen: RWTH Aachen Univ., 2018. P. 372–384 (CEUR Workshop Proc., Vol. 2098). Available at <http://ceur-ws.org/Vol1-2098>.
19. **Nesterov Yu.** A method of solving a convex programming problem with convergence rate $O(1/k^2)$ // Sov. Math. Dokl. 1983. Vol. 27, No. 2. P. 372–376.
20. **Juditsky A., Nemirovski A.** First order methods for non-smooth convex large-scale optimization, I: General purpose methods // Optimization for machine learning. Cambridge, MA: MIT Press, 2012. P. 121–148.
21. **Баяндина А. С., Гасников А. В., Гасникова Е. В., Мациевский С. В.** Прямо-двойственный метод зеркального спуска для условных задач стохастической оптимизации // Журн. вычисл. математики и мат. физики 2018. Т. 58, № 11. С. 1728–1736.
22. **Meng X., Chen H.** Accelerating Nesterov’s method for strongly convex functions with Lipschitz gradient. 2011 (Cornell Univ. Libr. e-Print Archive; arXiv: 1109.6058).

Стонякин Фёдор Сергеевич
Алкуса Мохаммад
Степанов Алексей Николаевич
Титов Александр Александрович

Статья поступила
17 октября 2018 г.
После доработки —
24 февраля 2019 г.
Принята к публикации
27 февраля 2019 г.

ADAPTIVE MIRROR DESCENT ALGORITHMS
FOR CONVEX AND STRONGLY CONVEX OPTIMIZATION
PROBLEMS WITH FUNCTIONAL CONSTRAINTS^(*)*F. S. Stonyakin*^{1,2,a}, *M. Alkousa*^{2,b},
A. N. Stepanov^{1,c}, and *A. A. Titov*^{2,d}¹Vernadsky Crimean Federal University,
4 Vernadsky Avenue, 295007 Simferopol, Russia²Moscow Institute of Physics and Technologies,
9 Institutskii Bystreet, 141701 Dolgoprudnyi, RussiaE-mail: ^a*fedyor@mail.ru*, ^b*mohammad.alkousa@phystech.edu*,
^c*stepanov.student@gmail.com*, ^d*a.a.titov@phystech.edu*

Abstract. Under consideration are some adaptive mirror descent algorithms for the problems of minimization of a convex objective functional with several convex Lipschitz (generally, nonsmooth) functional constraints. It is demonstrated that the methods are applicable to the objective functionals of various levels of smoothness: The Lipschitz condition holds either for the objective functional itself or for its gradient or Hessian (while the functional itself can fail to satisfy the Lipschitz condition). The main idea is the adaptive adjustment of the method with respect to the Lipschitz constant of the objective functional (its gradient or Hessian), as well as the Lipschitz constant of the constraint. The two types of methods are considered: adaptive (not requiring the knowledge of the Lipschitz constants neither for the objective functional nor for constraints) and partially adaptive (requiring the knowledge of the Lipschitz constant for constraints). Using the restart technique, some methods are proposed for strongly convex minimization problems. Some estimates of the rate of convergence are obtained for all algorithms under consideration in dependence on the level of smoothness of the objective functional. Numerical experiments are presented to illustrate the advantages of the proposed methods for some examples. Tab. 3, bibliogr. 22.

Keywords: adaptive Mirror Descent, Lipschitz condition, Lipschitz gradient, Lipschitz Hessian, strongly convex function, the technique of restarts.

^(*)The research by F. S. Stonyakin (analysis of Algorithms 1 and 3) was supported by the Russian Foundation for Basic Research (Project 18–31–00219 mol_a).

REFERENCES

1. **A. Ben-Tal** and **A. Nemirovski**, Robust truss topology design via semidefinite programming, *SIAM J. Optim.* **7** (4), 991–1016 (1997).
2. **Yu. Nesterov** and **S. Shpirko**, Primal-dual subgradient methods for huge-scale linear conic problem, *SIAM J. Optim.* **24** (3), 1444–1457 (2014).
3. **Y. Nesterov**, *Introductory Lectures on Convex Optimization: A Basic Course* (Kluwer Acad. Publ., Massachusetts, 2004).
4. **F. P. Vasil'ev**, *Optimization Methods*, Vol. 1 (MTsNMO, Moscow, 2011) [Russian].
5. **F. P. Vasil'ev**, *Optimization Methods*, Vol. 2 (MTsNMO, Moscow, 2011) [Russian].
6. **G. Lan**, Gradient sliding for composite optimization, *Math. Program.* **159** (1–2), 201–235 (2016).
7. **A. Beck** and **M. Teboulle**, Mirror descent and nonlinear projected subgradient methods for convex optimization, *Oper. Res. Lett.* **31** (3), 167–175 (2003).
8. **A. Nemirovskii** and **D. Yudin**, Efficient methods for large-scale convex optimization problems, *Ekonomika Mat. Metody*, No. 2, 135–152 (1979) [Russian].
9. **A. Nemirovskii** and **D. Yudin**, *Problem Complexity and Method Efficiency in Optimization* (Nauka, Moscow, 1979 [Russian]; J. Wiley & Sons, New York, 1983).
10. **N. Z. Shor**, Application of generalized gradient descent in block programming, *Kibernetika* **3** (3), 53–55 (1967) [Russian].
11. **B. T. Polyak**, A general method of solving extremum problems, *Soviet Math. Dokl.* **8** (3), 593–597 (1967).
12. **A. Beck**, **A. Ben-Tal**, **N. Guttman-Beck**, and **L. Tetruashvili**, The CoMirror Algorithm for solving nonsmooth constrained convex problems, *Oper. Res. Lett.* **38** (6), 493–498 (2010).
13. **A. Ben-Tal** and **A. Nemirovski**, *Lectures on Modern Convex Optimization* (SIAM, Philadelphia, 2001).
14. **A. Bayandina**, **P. Dvurechensky**, **A. Gasnikov**, **F. Stonyakin**, and **A. Titov**, Mirror descent and convex optimization problems with non-smooth inequality constraints, in *Large-Scale and Distributed Optimization: Lecture Notes in Mathematics*, Vol. 2227 (Springer, Cham, 2018), pp. 181–213.
15. **F. S. Stonyakin**, **M. S. Alkousa**, **A. N. Stepanov**, and **M. A. Barinov**, Adaptive mirror descent algorithms in convex programming problems with Lipschitz constraints, *Trudy Inst. Mat. Mekh.* **24** (2), 266–279 (2018) [Russian].
16. **A. A. Titov**, **F. S. Stonyakin**, **A. V. Gasnikov**, and **M. S. Alkousa**, Mirror descent and constrained online optimization problems optimization and applications, in *Optimization and Applications: Revised Selected Papers, 9th International Conference OPTIMA-2018 (Petrovac, Montenegro, October 1-5, 2018)*. Ser. *Communications in Computer and Information Science*, Vol. 974 (Springer, Cham, 2019), pp. 64–78.
17. **Yu. Nesterov**, Subgradient methods for convex functions with nonstandard growth properties, (2016). Available at www.mathnet.ru:8080/PresentFiles/16179/growthbm_nesterov.pdf.

18. **F. S. Stonyakin** and **A. 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), Omsk, Russia, July 8–14, 2018, CEUR Workshop Proceedings*, Vol. 2098 (RWTH Aachen Univ., Aachen, 2018), pp. 372–384. Available at <http://ceur-ws.org/Vol-2098>.
19. **Yu. Nesterov**, A method of solving a convex programming problem with convergence rate $O(1/k^2)$, *Soviet Math. Doklady* **27** (2), 372–376 (1983).
20. **A. Juditsky** and **A. Nemirovski**, First order methods for non-smooth convex large-scale optimization. I: General purpose methods, in *Optimization for Machine Learning*, (MIT Press, Cambridge, MA, 2012), pp. 121–148.
21. **A. S. Bayandina**, **A. V. Gasnikov**, **E. V. Gasnikova**, and **S. V. Matsievsky**, Primal-dual mirror descent for the stochastic programming problems with functional constraints, *Comput. Math. Math. Phys.* **58** (11), 1728–1736 (2018).
22. **X. Meng** and **H. Chen**, Accelerating Nesterov’s method for strongly convex functions with Lipschitz gradient, (Cornell Univ. Libr. e-Print Archive, 2011; arXiv: 1109.6058). Available at <https://arxiv.org/pdf/1109.6058.pdf>.

Fedor S. Stonyakin
Mohammad Alkousa
Aleksey N. Stepanov
Aleksandr A. Titov

Received October 17, 2018
 Revised February 24, 2019
 Accepted February 27, 2019