ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2020, том 60, № 11, с. 1843–1866

ОПТИМАЛЬНОЕ УПРАВЛЕНИЕ

УДК 519.624

УСКОРЕННЫЕ МЕТОДЫ ДЛЯ СЕДЛОВЫХ ЗАДАЧ[1)]

© 2020 г. М. С. Алкуса[1,2,*], А. В. Гасников[1,2,3,], Д. М. Двинских[3,4,***], Д. А. Ковалев[5,****], Ф. С. Стонякин[6,1,*****]**

[1] 141700 Долгопрудный, М.о., Институтский пер., 9, НИУ МФТИ, Россия

[2] 101000 Москва, ул. Мясницкая, 18, НИУ ВШЭ, Россия

[3] 127051 Москва, Большой Каретный пер., 19, Ин-т проблем передачи информации РАН, Россия

[4] 10117 Berlin, Monhrenstr, 39, Weierstrass Institute for Applied Analysis and Stochastics, Германия

[5] 23955 Thuwal, King Abdullah University of Science and Technology, Саудовская Аравия

[6] 295007 Симферополь, пр-т Акад. Вернадского, 4, Крымский федеральный ун-т, Россия

*e-mail: mohammad.alkousa@phystech.edu

**e-mail: gasnikov@yandex.ru

***e-mail: darina.dvinskikh@wias-berlin.de

****e-mail: dmitry.kovalev@kaust.edu.sa

*****e-mail: fedyor@mail.ru

Поступила в редакцию 01.12.2019 г.
Переработанный вариант 20.12.2019 г.
Принята к публикации 07.07.2020 г.

В последнее время было показано, как на основе обычного ускоренного градиентного метода решения задач гладкой выпуклой оптимизации можно получить ускоренные методы для более сложных задач (со структурой) и задач, по ходу решения которых используется различная локальная информация о поведении функции (стохастический градиент, гессиан и т.п.). “Ускоренные” методы здесь означает, с одной стороны, наличие некоторого единого и достаточно общего способа ускорения. С другой стороны, это означает и “оптимальность” методов, что часто удается строго доказать. В настоящей работе предпринята попытка построить в том же духе теорию ускоренных методов решения гладких выпукло-вогнутых седловых задач со структурой. Основным результатом статьи является получение в некотором смысле необходимых и достаточных условий, при которых сложность решения нелинейных выпукло-вогнутых седловых задач со структурой по числу вычислений градиентов композитов по прямым переменным равна по порядку аналогичной сложности решения билинейных задач со структурой. Библ. 30.

Ключевые слова: седловая задача, ускоренный метод, слайдинг, проксимально-дружественная функция.

DOI: 10.31857/S0044466920110022

1. ВВЕДЕНИЕ

Одним из основных направлений в численных методах выпуклой оптимизации в последнее десятилетие стало повсеместное распространение конструкции ускорения обычного градиентного метода, предложенной в 1983 г. Ю.Е. Нестеровым [1], на различные другие численные методы оптимизации.

Напомним вкратце исходную конструкцию. Для решения задачи выпуклой оптимизации

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

[1)] Исследование (в § 1, 2) выполнено в рамках Программы фундаментальных исследований НИУ ВШЭ и финансировалось в рамках господдержки ведущих университетов Российскогй Федерации “5-100”. Работа выполнена при финансовой поддержке РФФИ 18-31-20005 мол-а-вед в § 3, РНФ 18-71-10108 в § 4. Работы над результатами приложения П1 и частично приложения П2 выполнены при поддержке гранта Президента Российской Федерации для государственной поддержки молодых российских ученых-кандидатов наук, код МК-15.2020.1.

было предложено использовать вместо стандартного метода градиентного спуска

$$x^{k+1} = x^k - \frac{1}{L}\nabla f(x^k), \quad k \geq 0,$$

где L – верхняя оценка константы Липшица градиента $f(x)$ в 2-норме (далее, для краткости, будем говорить, что $f(x)$ имеет L-липшицев градиент), следующий ускоренный метод

$$x^1 = x^0 - \frac{1}{L}\nabla f(x^0),$$

$$x^{k+1} = x^k - \frac{1}{L}\nabla f\left(x^k + \frac{k-1}{k+2}(x^k - x^{k-1})\right) + \frac{k-1}{k+2}(x^k - x^{k-1}), \quad k \geq 1.$$

Сложность итерации этого метода сопоставима со сложностью итерации градиентного спуска. Однако если градиентный спуск сходится в общем случае (на плохо обусловленных задачах [2]) не лучше, чем [3]

$$f(x^N) - f(x_*) \geq \frac{LR^2}{4N},$$

то ускоренный метод сходится как

$$f(x^N) - f(x_*) \leq \frac{4LR^2}{N^2}, \tag{2}$$

где x_* – решение задачи (1), а $R = \left\|x^0 - x_*\right\|_2$. Если решение не единственное, то под x_* можно понимать то решение, которое наиболее близко (относительно 2-нормы) к точке старта метода x^0 [4]. Оценка (2) с точностью до числового множителя уже не может быть улучшена в общем случае ни на каком другом возможном методе при фиксированном классе задач выпуклой оптимизации (1) с L-липшицевым градиентом, см., например, [5], [6]. Аналогичные рассуждения можно провести и в невырожденном случае, когда $f(x)$ есть μ-сильно выпуклая функция [7]. В следующих разделах мы рассмотрим именно такой случай.

За последние 15 лет описанная конструкция ускорения была успешно перенесена на гладкие задачи условной выпуклой оптимизации, на задачи со структурой (в частности, так называемые композитные задачи). Ускорение успешно перенеслось и на неполноградиентные методы (безградиентные методы, спуски по направлению, координатные спуски) и на методы, использующие старшие производные. Также удалось добиться ускорения рандомизированных методов (например, методов редукции дисперсии в задачах минимизации суммы функций) и методов решения задач гладкой стохастической оптимизации. Под успешностью переноса здесь, как и выше, понимается достижение с помощью соответствующих ускоренных методов (с точностью до числовых множителей) известных нижних оценок. Важно также отметить, что в основу схемы ускорения во всех описанных случаях положена идейно одна и та же конструкция. Детали и более подробный обзор литературы можно найти в работе [4].

Несмотря на отмеченные достижения, по-прежнему остается ряд достаточно важных для практики постановок задач, в которых пока не до конца ясно, как именно следует добиваться ускорения имеющихся методов. В частности, можно выделить специальный подкласс задач вида (1), включающий в себя седловые задачи

$$f(x) := r(x) + \underbrace{\max_{y \in Q_y}\{F(x,y) - h(y)\}}_{g(x) = F(x, y^*(x)) - h(y^*(x))} \to \min_{x \in Q_x}, \tag{3}$$

где $y^*(x) = \operatorname{argmax}_{y \in Q_y}\{F(x,y) - h(y)\}$. Заметим, что задачи такого типа встречаются в самых разных приложениях, в том числе в поиске равновесий в двухстадийных моделях транспортных потоков [8].

Если задача не сильно выпукла, то ее можно свести к сильно выпуклой применением техники регурялизации и (или) двойственного сглаживания. Без ограничения общности можно считать $\mu_x \geq \frac{\varepsilon}{2R_x^2}$ (регуляризация [4]) и (или) $\mu_y \geq \frac{\varepsilon}{2R_y^2}$ (двойственное сглаживание [4], [6], [9], [10]), где $R_x = \left\|x^0 - x_*\right\|_2$, $R_y = \left\|y^0 - y(x_*)\right\|_2$, (x^0, y^0) – стартовая точка для выбранного численного метода

решения задачи (3). Здесь и всюду далее в работе под $\langle x, y\rangle$мы понимаем обычное скалярное произведение векторов в конечномерном пространстве и под $\|x\|_2 = \langle x, x\rangle$ – евклидову норму.

С точки зрения ускоренных методов данный класс задач достаточно подробно исследован в основном в случае $F(x, y) = \langle Ax, y\rangle$ для некоторого линейного оператора A (см., например, [9]).

В настоящей работе получены аналоги результатов [9] для более общего класса операторов F, которые не имеют билинейную структуру. Попутно найдены и некоторые уточнения оценок и для случая $F(x, y) = \langle Ax, y\rangle$. Предлагается сводить рассматриваемые седловые задачи к комбинации вспомогательных задач гладкой сильно выпуклой минимизации отдельно по каждой из групп переменных. Каждая из таких вспомогательных задач может решаться упомянутым выше быстрым градиентным методом Ю.Е. Нестерова. При этом неточность решения вспомогательной (внутренней) подзадачи для одной из групп переменных приводит к необходимости использовать для внешней задачи (которая определяется решением внутренней) концепцию неточного оракула [11], для которой известны оценки скорости сходимости быстрого градиентого метода. В этом случае получены оценки сложности (необходимого количества обращений к подпрограмме для вычисления градиента) предложенного метода, аналогичные известному результату об ускоренном градиентном слайдинге Дж. Лана (см. [9]). При этом с одной стороны предложена новая относительно простая схема пояснения ускоренного градиентного слайдинга с помощью техники Каталист [12], а с другой стороны – результат об ускоренном градиентном слайдинге обобщен на случай, когда вместо обычных градиентов целевых функционалов используются некоторые их неточные аналоги.

Работа состоит из введения, заключения и трех основных разделов. В разд. 2 приводится постановка рассматриваемой задачи и кратко описывается подход к ней на базе известного проксимального зеркального метода А.С. Немировского для вариационных неравенств и седловых задач. Разд. 3 посвящен обзору известных результатов о возможности ускорений оценки скорости сходимости для рассматриваемого класса достаточно гладких сильно выпукло-вогнутых седловых задач, а также формулировкам основных результатов настоящей статьи (i)–(iii). В разд. 4 приводится схема обоснования основных утверждений работы, сформулированы необходимые вспомогательные утверждения (леммы 1, 2 и 3). Для леммы 3 об ускоренном градиентном слайдинге для минимизации суммы гладких выпуклых функционалов (один из которых сильно выпуклый) при использовании концепции неточного градиента в основной части статьи приведена схема рассуждений на базе недавно предложенной техники Каталист [12]. Полные доказательства лемм 1, 2 и 3 приводятся в приложении к работе.

2. ПОСТАНОВКА ЗАДАЧИ

Пусть $Q_x \subseteq \mathbb{R}^m$, $Q_y \subseteq \mathbb{R}^n$ – непустые, выпуклые и компактные множества, $r : Q_x \to \mathbb{R}$ и $h : Q_y \to \mathbb{R}$ есть μ_x-сильно выпуклая и μ_y-сильно выпуклая функции соответственно.

Всюду будем предполагать, что функционал $F : Q_x \times Q_y \to \mathbb{R}$ выпуклый по x и вогнутый по y и задан в некоторой окрестности множества $Q_x \times Q_y$. При этом F будем считать достаточно гладким на $Q_x \times Q_y$. Точнее говоря, для произвольных $x, x' \in Q_x$ и $y, y' \in Q_y$, верны следующие неравенства:

$$\|\nabla_x F(x, y) - \nabla_x F(x', y)\|_2 \le L_{xx}\|x - x'\|_2, \tag{4}$$

$$\|\nabla_x F(x, y) - \nabla_x F(x, y')\|_2 \le L_{xy}\|y - y'\|_2,$$

$$\|\nabla_y F(x, y) - \nabla_y F(x', y)\|_2 \le L_{xy}\|x - x'\|_2, \tag{5}$$

$$\|\nabla_y F(x, y) - \nabla_y F(x, y')\|_2 \le L_{yy}\|y - y'\|_2,$$

где $L_{xx}, L_{xy}, L_{yy} \ge 0$.

Рассмотрим класс выпукло-вогнутых седловых задач

$$\min_{x\in Q_x}\max_{y\in Q_y}\{S(x, y) := r(x) + F(x, y) - h(y)\}. \tag{6}$$

Пусть

$$\hat{S}(x, y) = F(x, y) - h(y). \tag{7}$$

Тогда возможно переписать задачу (6) следующим образом:

$$\min_{x\in Q_x}\{r(x) + \max_{y\in Q_y}\hat{S}(x, y)\} = \min_{x\in Q_x}\{r(x) + g(x)\},$$

т.е. задача (6) имеет вид

$$f(x) := r(x) + g(x) \to \min_{x\in Q_x},$$

где

$$g(x) = \max_{y\in Q_y}\hat{S}(x, y). \tag{8}$$

Поскольку функционал $\hat{S}(x,\cdot)$ есть μ_y-сильно вогнутый на Q_y, то задача максимизации (8) имеет единственное решение

$$y^*(x) := \arg\max_{y\in Q_y}\hat{S}(x, y) \quad \forall x \in Q_x, \tag{9}$$

откуда

$$g(x) = \hat{S}(x, y^*(x)) = F(x, y^*(x)) - h(y^*(x)).$$

Всюду далее для произвольного $x \in Q_x$ и некоторого $\delta \geq 0$ будем называть $\tilde{y}_\delta(x) \in Q_y$ δ-приближенным решением задачи (8), если

$$g(x) - \hat{S}(x, \tilde{y}_\delta(x)) = \hat{S}(x, y^*(x)) - \hat{S}(x, \tilde{y}_\delta(x)) \leq \delta. \tag{10}$$

Хорошо известно, что задача нахождения седловых точек выпукло-вогнутого функционала может сводиться к задаче решения вариационного неравенства с монотонным оператором

$$G(x) = \begin{pmatrix} \nabla_u f(u, v) \\ -\nabla_v f(u, v) \end{pmatrix}, \quad x = (u, v) \in Q := Q_1 \times Q_2. \tag{11}$$

Напомним общую постановку задачи решения вариационного неравенства (ВН). Для некоторого оператора $G : Q \to \mathbb{R}^n$, заданного на выпуклом компакте $Q \subset \mathbb{R}^n$ будем рассматривать сильные вариационные неравенства (ВН) вида

$$\langle G(x_*), x_* - x\rangle \leq 0 \quad \forall x \in Q, \tag{12}$$

где G удовлетворяет условию Липшица. Отметим, что в (12) требуется найти $x_* \in Q$ (где x_* – решение ВН), для которого

$$\max_{x\in Q}\langle G(x_*), x_* - x\rangle \leq 0.$$

Для решения задачи вариационного неравенства в последние годы весьма популярен проксимальный зеркальный метод А.С. Немировского [13]. Недавно был предложен также его вариант с адаптивным выбором шага [14].

Гладкость рассматриваемой седловой задачи приводит к липшицевости оператора G, с некоторой константой $L > 0$. В таком случае, как известно, для проксимального зеркального метода справедлива следующая оценка [14]:

$$\frac{1}{N}\sum_{k=1}^{N}\langle G(y^k), y^k - x\rangle \leq \frac{L\|x - x^0\|_2^2 - L\|x - x^N\|_2^2}{2N} \quad \forall x \in Q, \tag{13}$$

где

$$y^k := \arg\min_{x\in Q}\left\{\langle G(x^{k-1}), x - x^{k-1}\rangle + \frac{L}{2}\|x - x^{k-1}\|_2^2\right\}, \quad k = 1, 2, \ldots, N.$$

Заметим, что для всех $y^k \in Q$ имеем

$$\left\langle G(x_*), y^k - x_* \right\rangle \geq 0. \tag{14}$$

Из (13) следует, что

$$\frac{1}{N}\sum_{k=1}^{N}\left\langle G(y^k), y^k - x_* \right\rangle \leq \frac{L\left\|x_* - x^0\right\|_2^2}{2N}. \tag{15}$$

Объединяя неравенства (14) и (15), получаем

$$\frac{1}{N}\sum_{k=1}^{N}\left\langle G(y^k) - G(x_*), y^k - x_* \right\rangle \leq \frac{L\left\|x_* - x^0\right\|_2^2}{2N}.$$

Дополнительно предположим сильную монотонность оператора G, т.е. существование такого $\mu > 0$, что

$$\left\langle G(y) - G(x), y - x \right\rangle \geq \mu\left\|y - x\right\|_2^2 \quad \forall x, y \in Q.$$

С учетом выпуклости функции $\|x\|_2^2$ имеем

$$\mu\left\|\overline{y}^N - x_*\right\|_2^2 \leq \mu\sum_{k=1}^{N}\frac{1}{N}\left\|y^k - x_*\right\|_2^2 \leq \frac{1}{N}\sum_{k=1}^{N}\left\langle G(y^k) - G(x_*), y^k - x_* \right\rangle \leq \frac{L\left\|x_* - x^0\right\|_2^2}{2N},$$

где $\overline{y}^N = \frac{1}{N}\sum_{k=1}^{N} y^k$.

На основе последнего неравенства уже возможно организовать процедуру рестартов проксимального зеркального метода и тогда он будет сходиться с линейной скоростью. Общая оценка числа итераций, необходимого для достижения приемлемого качества решения, будет иметь вид

$$O\left(\frac{L}{\mu}\ln\left(\frac{\mu R^2}{\varepsilon}\right)\right),$$

где $R = \left\|x^0 - x_*\right\|_2$ и ε – точность решения x_*. Хорошо известно, что данная оценка не может быть улучшена для рассматриваемого класса вариационных неравенств никакими другими методами.

Если применять рассмотренный подход к поставленной задаче (6), то оператор G из (11) будет μ-сильно монотонным $\mu = \min\{\mu_x, \mu_y\}$, что приведет к следующей оценке сложности:

$$O\left(\frac{L}{\min\{\mu_x, \mu_y\}}\ln\left(\frac{\min\{\mu_x, \mu_y\}R^2}{\varepsilon}\right)\right) \tag{16}$$

достижения нужного качества решения. Если μ_x или μ_y близко к нулю, то величина в (16) может оказаться достаточно большой. В последующих разделах мы, в частности, рассмотрим альтернативные подходы, которые позволяют уточнить оценку (16) в случае, когда $\mu_x \ll \mu_y$ или $\mu_y \ll \mu_x$.

3. ОСНОВНЫЕ РЕЗУЛЬТАТЫ

Опишем наилучшие известные на данный момент результаты оценок скорости сходимости методов для задачи (3), а также сформулируем основные результаты нашей работы.

Будем говорить, что функция $r(x)$ – проксимально дружественная, если задача вида

$$\min_{x \in Q_x}\left\{\left\langle c_1, x \right\rangle + r(x) + c_2\left\|x\right\|_2^2\right\}, \tag{17}$$

где $c_1 \in Q_x$ и $c_2 > 0$, может быть решена явно. Аналогично $h(y)$ – проксимально дружественная функция, если задача вида

$$\min_{y \in Q_y}\left\{\left\langle c_1, y \right\rangle + h(y) + c_2\left\|y\right\|_2^2\right\} \tag{18}$$

может быть решена явно.

Под ε-решением задачи (3) будем понимать такую пару $(x^{N_x}, y^{N_y}) \in Q_x \times Q_y$, что

$$f(x^{N_x}) - f(x_*) \le \max_{y \in Q_y \cap B_m(2R_y)} \left\{ r(x^{N_x}) + F(x^{N_x}, y) - h(y) \right\} - \min_{x \in Q_x \cap B_n(2R_x)} \left\{ r(x) + F(x, y^{N_y}) - h(y^{N_y}) \right\} \le \varepsilon,$$

где $B_n(R)$ – евклидов шар радиуса R в $\mathbb{R}^n$.

Ниже приведены наилучшие известные нам результаты (см. [9], [10], [15], [16], [17] и цитированную в этих работах литературу) относительно сложности решения задачи (3), которые далее мы постараемся уточнить. Собственно, в пп. 2)–4) такое уточнение уже делается: ранее в приведенном виде данные результаты были известны только в предположении $F(x, y) = \langle Ax, y \rangle$ для некоторого линейного оператора A. В этом случае $L_{xx} = L_{yy} = 0$, $L_{xy} = L_{yx} = \sqrt{\lambda_{\max}(A^{\mathrm{T}}A)}$, где $\lambda_{\max}(A^{\mathrm{T}}A)$ – наибольшее собственное значение матрицы $A^{\mathrm{T}}A$.

1) Если $r(x)$ и $h(y)$ проксимально дружественны, то ε-решение задачи (3) может быть достигнуто за $\tilde{O}\left(\frac{L_{xy}}{\sqrt{\mu_x \mu_y}}\right)$ вычислений (17), $\nabla_x F(x, y)$ и (18), $\nabla_y F(x, y)$ при $F(x, y) = \langle Ax, y \rangle$ и за $\tilde{O}\left(\frac{\max\{L_{xx}, L_{xy}, L_{yy}\}}{\min\{\mu_x, \mu_y\}}\right)$ вычислений (17), $\nabla_x F(x, y)$ и (18), $\nabla_y F(x, y)$ в общем случае. Здесь и далее $\tilde{O}() = O()$ с точностью до небольшой степени логарифмического по ε, μ_x или μ_y, а также по R_x или R_y множителя. Как правило, показатель этой степени 1 или 2.

2) Если $r(x)$ имеет L_x-липшицев градиент, но не проксимально дружественна, то ε-решение задачи (3) может быть достигнуто за $\tilde{O}\left(\sqrt{\frac{L_x}{\mu_x}}\right)$ вычислений $\nabla r(x)$,

$$\tilde{O}\left(\sqrt{\frac{L_{xx} + \frac{L_{xy}^2}{\mu_y}}{\mu_x}}\right) \tag{19}$$

вычислений $\nabla_x F(x, y)$ и

$$\tilde{O}\left(\sqrt{\frac{L_{xx} + \frac{L_{xy}^2}{\mu_y}}{\mu_x}} \sqrt{\max\left\{\frac{L_{yy}}{\mu_y}, 1\right\}}\right) \tag{20}$$

вычислений (18), а также $\nabla_y F(x, y)$.

3) Если $h(y)$ имеет L_y-липшицев градиент, но не проксимально дружественна, то ε-решение задачи (3) может быть достигнуто за (19) вычислений (17), $\nabla_x F(x, y)$ и

$$\tilde{O}\left(\sqrt{\frac{L_{xx} + \frac{L_{xy}^2}{\mu_y}}{\mu_x}} \sqrt{\frac{L_y}{\mu_y}}\right) \tag{21}$$

вычислений $\nabla h(y)$, (20) вычислений $\nabla_y F(x, y)$.

4) Если $r(x)$ имеет L_x-липшицев градиент, $h(y)$ имеет L_y-липшицев градиент, но обе функции не проксимально дружественны, то ε-решение задачи (3) может быть достигнуто за $\tilde{O}\left(\sqrt{\frac{L_x}{\mu_x}}\right)$ вычислений $\nabla r(x)$, (19) вычислений $\nabla_x F(x, y)$ и (21) вычислений $\nabla h(y)$, (20) вычислений $\nabla_y F(x, y)$.

Отметим, что результаты всего п. 1) и пп. 2), 4) в части числа вычислений $\nabla r(x)$ (и, по-видимому, в части числа вычислений $\nabla_x F(x, y)$) в общем случае не могут быть улучшены [5], [18] (с точностью до логарифмического множителя), если

$$\dim(x) + \dim(y) \geqslant \frac{L_{xy}}{\sqrt{\mu_x \mu_y}}.$$

Заметим, что правую часть можно менять в зависимости от специфики постановки задачи. Если это условие не выполняется, то можно свести седловую задачу к негладкой задаче выпуклой оптимизации [19], [20] и решать ее методами типа центров тяжести, например, методом эллипсоидов или Вайды [20], [21]. Заметим, что методом эллипсоидов можно решать седловую задачу и напрямую [22].

Сформулируем результаты настоящей работы.

i) Основным результатом работы является обоснование возможности в пп. 2)–4) выше полностью убрать оговорку "при $F(x, y) = \langle Ax, y\rangle$".

При $\mu_x = \frac{\varepsilon}{R_x^2}$ частично это уже было сделано в работе [16]. Однако полученные в этой работе оценки в части числа вычислений $\nabla r(x)$ проигрывают оценкам, приведенным выше, в $\sim\frac{1}{\mu_y}$ раз.

ii) Другим результатом работы является уточнение приведенных выше утверждений в случае, когда $F(x, y) = \langle Ax, y\rangle$ для некоторого линейного оператора A, $Q_y = \mathbb{R}^m$, $h(y)$ имеет L_y-липшицев градиент и $\frac{\lambda_{\min}(A^{\mathrm{T}}A)}{L_y} \geqslant \mu_x$. В этом случае во всех приведенных выше формулах можно заменить μ_x на $\frac{\lambda_{\min}(A^{\mathrm{T}}A)}{L_y}$.

iii) Более того, если $F(x, y) = \langle Ax, y\rangle$ для некоторого линейного оператора A, $Q_x = \mathbb{R}^n$, $Q_y = \mathbb{R}^m$, $r(x)$ имеет L_x-липшицев градиент, $h(y)$ имеет L_y-липшицев градиент, то оценки на количество вычислений $\nabla_x F(x, y) = A^{\mathrm{T}}y$, $\nabla_y F(x, y) = Ax$ и (18) (здесь предполагается проксимальная дружественность $h(y)$) можно заменить (это имеет смысл, если новые оценки станут лучше имеющихся) на следующую оценку:

$$[\text{Число вычислений } \nabla r(x)] \cdot \tilde{O}\left(\sqrt{\frac{L_y \lambda_{\max}(A^{\mathrm{T}}A)}{\mu_y \lambda^{+}_{\min}(A^{\mathrm{T}}A)}}\right) = \tilde{O}\left(\sqrt{\frac{L_x L_y \lambda_{\max}(A^{\mathrm{T}}A)}{\mu_x \mu_y \lambda^{+}_{\min}(A^{\mathrm{T}}A)}}\right), \tag{22}$$

где $\lambda^{+}_{\min}(A^{\mathrm{T}}A)$ – минимальное положительное собственное значение матрицы $A^{\mathrm{T}}A$.

Отметим, что аналогичное ii) утверждение можно сформулировать и в случае, когда $r(x)$ имеет L_x-липшицев градиент.

Стоит отметить, что результат ii) позволяет, среди прочего, элементарным образом объяснить, как решать матричную игру за $\tilde{O}\left(\sqrt{\frac{\lambda_{\max}(A^{\mathrm{T}}A)}{\lambda_{\min}(A^{\mathrm{T}}A)}}\right)$ матрично-векторных умножений. Сначала необходимо двойственно сгладить (должным образом) исходную постановку задачи [4], [6], [9], [10], т.е. ввести проксимально-дружественную функцию $h(y) = \frac{\varepsilon\|y\|_2^2}{4R_y^2}$. Отметим, что при этом $Q_x = Q_y = \mathbb{R}^n$. Далее, полученную в результате функцию $g(x)$ можно минимизировать обычным ускоренным (быстрым градиентным) методом [1], [4]–[7], [9], [10] (в сильно выпуклом случае). Полученная при этом сложность решения задачи существенно лучше популярных и активно исследуемых процедур типа экстраградиентного метода, для которого удалось лишь получить такую оценку трудоемкости $\tilde{O}\left(\frac{\lambda_{\max}(A^{\mathrm{T}}A)}{\lambda_{\min}(A^{\mathrm{T}}A)}\right)$ [23].

Отметим также, что результаты, приведенные в п. iii), нашли важное приложение в построении оптимальных ускоренных методов для гладких выпуклых задач децентрализованной распределенной оптимизации [24]. Собственно, наш изначальный интерес к этой проблематике и был связан с развитием идей работы [24].

4. НЕОБХОДИМЫЕ ВСПОМОГАТЕЛЬНЫЕ УТВЕРЖДЕНИЯ И СХЕМА ОБОСНОВАНИЯ ОСНОВНЫХ РЕЗУЛЬТАТОВ РАБОТЫ

Для того чтобы получить отмеченные в предыдущем пункте результаты, приведем необходимые и, в основной части, уже известные результаты.

Лемма 1 (см. [4], [25], [26]). *В обозначениях п. 2, если* $F(x,y) = \langle Ax, y\rangle$, *то* $g(x)$ *будет иметь* L*-липшицев градиент, где* $L = \dfrac{\lambda_{\max}(A^{\mathrm{T}}A)}{\mu_y}$. *Если, дополнительно,* $h(y)$ *имеет* L_y*-липшицев градиент и* $Q_y = \mathbb{R}^m$, *то* $g(x)$ *есть* μ*-сильно выпуклая функция на* $(\operatorname{Ker} A)^{\perp}$, *где* $\mu = \dfrac{\lambda^{+}_{\min}(A^{\mathrm{T}}A)}{L_y}$. *При этом* $\nabla g(x) \in (\operatorname{Ker} A)^{\perp}$.

Следующее утверждение демонстрирует необходимость использования концепции неточного градиента для рассматриваемого подхода к седловым задачам.

Лемма 2 (см. [11], [16], [17]). *В обозначениях п. 2,* $g(x)$ *будет иметь* L*-липшицев градиент* $\nabla g(x) = \nabla_x F(x, y^*(x))$ (*теорема Демьянова–Данскина*), *где* $L = L_{xx} + \dfrac{2L_{xy}^2}{\mu_y}$. *Более того, для всех* $x, z \in Q_x$ *выполняется следующее условие* (*словами*: $\nabla_x F(x, \tilde{y}_\delta(x)) - (2\delta, 2L)$ *– градиент для* $g(x)$):

$$0 \le g(z) - \left[\{F(x, \tilde{y}_\delta(x)) - h(\tilde{y}_\delta(x))\} + \langle \nabla_x F(x, \tilde{y}_\delta(x)), z - x\rangle\right] \le \frac{2L}{2}\|z - x\|_2^2 + 2\delta, \tag{23}$$

где $\tilde{y}_\delta(x)$ *такой, что*

$$\underbrace{\max_{y \in Q_y}\{F(x,y) - h(y)\}}_{g(x)} - \{F(x, \tilde{y}_\delta(x)) - h(\tilde{y}_\delta(x))\} \le \delta,$$

который определен в (10).

Рассмотрим наиболее тонкое из используемых нами вспомогательных утверждений [9]. Пусть нужно решить задачу сильно выпуклой композитной оптимизации

$$r(x) + g(x) \to \min_{x \in \mathbb{R}^n}, \tag{24}$$

с точностью ε по функции. Считаем, что функции $r(x)$ и $g(x)$ имеют константы Липшица градиента L_r и L_g, и хотя бы одна из этих функций μ-сильно выпуклая. К такой задаче (24) можно применить результат о так называемом ускоренном градиентном слайдинге из п. 8.2 [9].

Лемма 3. *Необходимого качества решения задачи решения* (24) *можно достичь за* $N_r = \tilde{O}\left(\sqrt{\dfrac{L_r}{\mu}}\right)$ *вычислений градиента первого слагаемого* $\nabla r(x)$ *и* $N_g = \tilde{O}\left(\sqrt{\dfrac{L_g}{\mu}}\right)$ *вычислений градиента второго слагаемого* $\nabla g(x)$. *Результат останется верным, если вместо* $\nabla g(x)$ *использовать* $\left(O\left(\dfrac{\varepsilon}{N_g}\right), O(L_g)\right)$*-градиент*.

Подробное доказательство леммы 3 приведено в приложении. Здесь мы лишь приведем некоторую схему рассуждений. Для наглядности полагаем μ-сильно выпуклым именно функционал g. Если это не так и r есть μ_r-сильно выпуклый, а g просто выпуклый, то можно заменить $r(x)$ на $r(x) - \dfrac{\mu_g}{2}\|x\|_2^2$ и $g(x)$ на $g(x) + \dfrac{\mu_g}{2}\|x\|_2^2$ для некоторого $\mu_g < \mu_r$. Тогда оба функционала будут как гладкими, так и сильно выпуклыми.

К рассмотренной задаче (24) можно применить технику Каталист (алгоритм 1) из [12] в предположении $L_r \ll L_g$ и μ-сильной выпуклости g в 2-норме, причем $\mu \ll L_r$.

Алгоритм 1. Каталист.

1: **Вход:** $x^0 \in \mathbb{R}^n$, параметр L.

2: $y^0 := x^0$.

3: **while** желаемая точность не достигнута **do**

4: Найти x^k с некоторой точностью применением алгоритма 2

$$x^k \approx \arg\min_{x\in\mathbb{R}^n}\left\{r(x) + g(x) + \frac{L}{2}\left\|x - y^{k-1}\right\|_2^2\right\}.$$

5: Вычисляем y^k, используя шаг экстраполяции, с $\beta_k \in (0,1)$

$$y^k = x^k + \beta_k(x^k - x^{k-1}).$$

6: **end while**

7: **Выход:** x^k.

Алгоритм 2. Неускоренный градиентный метод для задач композитной оптимизации.

1: **Вход:** $x^0 \in \mathbb{R}^n$, параметр L_r.

2: **for** $k = 0, 1, 2, \ldots$ **do**

3:

$$\phi_{k+1}(x) := \left\langle \nabla r(x^k), x - x^k\right\rangle + g(x) - g(x^k) + \frac{L_r}{2}\left\|x - x^k\right\|_2^2,$$

$$x^{k+1} := \arg\min_{x\in\mathbb{R}^n}\phi_{k+1}(x).$$

4: **end for**

5: **Выход:** $\bar{x}_N = \frac{1}{N}\sum_{k=0}^{N-1} x^{k+1}$.

Тогда (см. [12]) вместо исходной задачи потребуется $\tilde{O}\left(\sqrt{\frac{L}{\mu}}\right)$ раз решать задачу вида

$$r(x) + g(x) + \frac{L}{2}\left\|x - y^{k-1}\right\|_2^2 \to \min_{x\in\mathbb{R}^n}, \tag{25}$$

где $L > 0$– некоторый параметр регуляризации, а последовательность $y^0, y^1, \ldots$ образуется согласно схеме алгоритма 1. При этом задачу (25) можно решать неускоренным композитным градиентным методом (алгоритм 2), считая $g(x) + \frac{L}{2}\left\|x - y^{k-1}\right\|_2^2$ композитом. Как известно, число итераций такого метода будет совпадать с количеством вычислений $\nabla r(x)$ и равно $\tilde{O}\left(\frac{L_r}{L+\mu}\right)$. Но при этом не предполагается проксимальная дружественность функции $g(x)$ и поэтому необходимо учитывать сложность решения возникающей на каждой итерации неускоренного композитного градиентного метода задачи вида

$$\left\langle \nabla r(\tilde{x}^l), x - \tilde{x}^l\right\rangle + \frac{L_r}{2}\left\|x - \tilde{x}^l\right\|_2^2 + g(x) + \frac{L}{2}\left\|x - y^{k-1}\right\|_2^2 \to \min_{x\in\mathbb{R}^n}.$$

Для решения уже этой вспомогательной задачи можно использовать ускоренный композитный градиентный метод для задач сильно выпуклой оптимизации [27] (см. также алгоритмы 3, 4 да-

лее). При этом слагаемое $\frac{L_r}{2}\left\|x-\tilde{x}^l\right\|_2^2+\frac{L}{2}\left\|x-y^{k-1}\right\|_2^2$ считается композитом. Число итераций такого метода будет $\tilde{O}\left(\sqrt{\frac{L_g}{L_r+L+\mu}}\right)$. Таким образом, общее количество вычислений $\nabla g(x)$ будет следующим

$$\tilde{O}\left(\sqrt{\frac{L}{\mu}}\right)\cdot\tilde{O}\left(\frac{L_r}{L+\mu}\right)\cdot\tilde{O}\left(\sqrt{\frac{L_g}{L_r+L+\mu}}\right).$$

Выберем параметр регуляризации L так, чтобы последнее выражение было минимальным. Тогда с учетом сделанных предположений $L_r \ll L_g$ и $\mu \ll L_r$ получим, что $L \simeq L_r$. Следовательно, общее число вычислений $\nabla g(x)$ будет действительно равно $\tilde{O}\left(\sqrt{\frac{L_g}{\mu}}\right)$.

Заметим, что полностью новым является утверждение из последнего предложения формулировки леммы 3. Доказательство этого ключевого наблюдения проводится аналогично оригинальной статье [11], см. также [9], [16]. Действительно, если функционал $g:\mathbb{R}^n\to\mathbb{R}$ допускает (δ,L)-градиент $\nabla_\delta g(x)$ в любой запрошенной точке x, т.е. верно неравенство

$$g(y)\le g(x)+\langle\nabla_\delta g(x),y-x\rangle+\frac{L}{2}\|y-x\|_2^2+\delta.$$

Последнее неравенство отличается от стандартного условия L-липшицевости градиента выпуклого функционала g

$$g(y)\le g(x)+\langle\nabla g(x),y-x\rangle+\frac{L}{2}\|y-x\|_2^2 \tag{26}$$

лишь постоянной величиной погрешности δ. Поэтому применение указанного неравенства для модификации результата об ускоренном градиентном слайдинге [9] (который в стандартном случае основан на комбинации оценок, связанных с неравенством (26)) для решения N_g вспомогательных подзадач приведет к накоплению в итоговой оценке погрешности, сопоставимой с величиной $O(N_g\delta)$. Это обстоятельство и приводит к необходимости вместо $\nabla g(x)$ использовать именно $\left(O\left(\frac{\varepsilon}{N_g}\right),O(L_g)\right)$-градиент.

Утверждения i) и ii) получаются сочетанием лемм 1–3. Приведем схему доказательства первого из результатов. Напомним, что мы рассматриваем задачу вида

$$\min_{x\in Q_x}\max_{y\in Q_y}\{S(x,y)=r(x)+F(x,y)-h(y)\}. \tag{27}$$

Мы проводим доказательство в предположении, что функционалы r и h проксимально дружественны (простой структуры, т.е. композиты).

Замечание 1. Если некоторый функционал $S(x,y)$ является μ_x-сильно выпуклым по x и μ_y-сильно вогнутым по y, то можно представить S в виде

$$S(x,y)=F(x,y)+\frac{\mu_x}{2}\|x\|_2^2-\frac{\mu_y}{2}\|y\|_2^2,$$

причем функционал F сохранит свойство выпуклости по x и вогнутости по y. Отметим, что ввиду достаточной гладкости функционалов $\|x\|_2^2$ и $\|y\|_2^2$ (липшицевость градиентов), если S гладкий, то F будет гладким. Таким образом, приводимая ниже схема рассуждений применима к седловым задачам для произвольного μ_x-сильно выпуклого по x и μ_y-сильно вонутого по y, а также достаточно гладкого функционала $S(x,y)$.

Зафиксируем $x\in Q_x$ и введем вспомогательный функционал

$$f(x):=\max_{y\in Q_y}S(x,y), \tag{28}$$

что позволяет рассматривать седловую задачу (27) как задачу выпуклой минимизации

$$f(x)\to\min_{x\in Q_x}. \tag{29}$$

Ясно, что

$$f(x) = \max_{y\in Q_y}\{r(x) + F(x,y) - h(y)\} = r(x) + \underbrace{\max_{y\in Q_y}\{F(x,y) - h(y)\}}_{g(x)=F(x,y^*(x))-h(y^*(x)}.$$

По лемме 2 функционал f (см. (28)) является μ_x-сильно выпуклым и g имеет L-липщицев градиент

$$\|\nabla f(x_2) - \nabla f(x_1)\|_2 \le L\|x_2 - x_1\|_2 \quad \forall x_1, x_2 \in Q_x,$$

где

$$L = L_{xx} + \frac{2L_{xy}^2}{\mu_y}.$$

Также согласно лемме 2 (см. также [16]) для функционала f выполнено неравенство

$$f(x_2) \le S(x_1, \tilde{y}_\gamma(x_1)) + \left\langle \nabla_x S(x_1, \tilde{y}_\gamma(x_1)), x_2 - x_1\right\rangle + L\|x_2 - x_1\|_2^2 + 2\gamma, \tag{30}$$

а γ – точность решения вспомогательной задачи максимизации $S(x_1, y)$ по y:

$$f(x_1) - S(x_1, \tilde{y}_\gamma(x_1)) \le \gamma. \tag{31}$$

Задача μ_y-сильно вогнутой композитной (в силу предположения для h выше) максимизации

$$F(x,y) - h(y) \to \max_{y\in Q_y}$$

(при фиксированном $x \in Q_x$) может быть решена с точностью γ в (31) за

$$O\left(\sqrt{\frac{L_{yy}}{\mu_y}}\ln\frac{1}{\gamma}\right) \tag{32}$$

итераций быстрого градиентного метода.

Алгоритм 3. Быстрый градиентный метод с (δ, L)-оракулом для задач композитной оптимизации (33).

1: **Вход:** $x^0 \in Q$ – начальная точка, N – количество шагов, $L > 0$ и $\delta > 0$.

2: $y^0 := x^0$, $u^0 := x^0$, $\alpha_0 := 0$, $A_0 := 0$.

3: **for** $k = 1, 2, \ldots, N$ **do**

4: Находим наибольший корень, α_{k+1} так, что

$$A_k + \alpha_{k+1} = L\alpha_{k+1}^2,$$

5:

$$A_{k+1} := A_k + \alpha_{k+1},$$

6:

$$y^{k+1} := \frac{\alpha_{k+1}u^k + A_k x^k}{A_{k+1}},$$

7:

$$\phi_{k+1}(x) = \frac{1}{2}\|x - u^k\|_2^2 + \alpha_{k+1}\left(\left\langle \nabla r(y^{k+1}), x - y^{k+1}\right\rangle + g(x) - g(y^{k+1})\right),$$

$$u^{k+1} := \arg\min_{x\in Q}\phi_{k+1}(x),$$

8:

$$x^{k+1} := \frac{\alpha_{k+1}u^{k+1} + A_k x^k}{A_{k+1}},$$

9: **end for**

10: **Выход:** x^N.

Покажем, как с использованием рестартов быстрого градиентного метода в концепции (δ, L)-оракула можно достичь заданного качества решения по функции для задач композитной оптимизации с требуемой оценкой сложности. Для задачи композитной оптимизации

$$f(x) := r(x) + g(x) \to \min_{x \in Q}. \tag{33}$$

можно применить предложенный в [28] алгоритм 3.

Для алгоритма 3 в [28] доказана следующая оценка скорости сходимости (мы предполагаем, что вспомогательные задачи решаются точно):

$$f(x^N) - f(x_*) \le \frac{8LR^2}{(N+1)^2} + 2N\delta,$$

где $R^2 = \frac{1}{2}\left\|x_* - x^0\right\|_2^2$, а x_* – ближайшая точка минимума к точке x^0. Введем вспомогательное обозначение

$$\psi(x, y) := \left\langle \nabla r(y), x - y \right\rangle + g(x) - g(y).$$

Тогда ввиду μ_x-сильной выпуклости g имеем

$$\frac{\mu_x}{2}\|x - y\|_2^2 \le f(x) - \left(f(y) + \psi(x, y)\right) \le \frac{L}{2}\|y - x\|_2^2 + \delta.$$

Алгоритм 4. Быстрый градиентный метод для задач сильно выпуклой композитной оптимизации с (δ, L)-оракулом, рестарты алгоритма 3.

1: **Вход:** $x^0 \in Q$ – начальная точка, $L > 0$, $\mu > 0$, $\delta > 0$, ε – точность решения, R и $p = \left\lceil \log_2\left(\frac{\mu R^2}{\varepsilon}\right)\right\rceil$ – количество рестартов.

2: **for** $j = 1, \ldots, p$ **do**

3: Выполнить $N_j = \left\lceil 3e\sqrt{\frac{L}{\mu}}\right\rceil$ итераций алгоритма 3.

4: $x^0 := x^{N_j}$.

5: **end for**

6: **Выход:** $\hat{x} := x^{N_p}$.

Если сделать естественное предположение $\psi(x, x_*) \ge \left\langle \nabla(r+g)(x_*), x - x_* \right\rangle \ge 0$, то после N_1 итераций алгоритма 3 имеем

$$\frac{\mu_x}{2}\left\|x^{N_1} - x_*\right\|_2^2 \le f(x^{N_1}) - f(x_*) \le \frac{4L\left\|x^0 - x_*\right\|_2^2}{N_1^2} + 2N_1\delta.$$

Если для числа итераций N_1 алгоритма 3 выбрать δ так, чтобы выполнялось неравенство

$$2N_1\delta \le \frac{L\left\|x^0 - x_*\right\|_2^2}{2N_1^2},$$

то получим

$$\left\|x^{N_1} - x_*\right\|_2^2 \le \frac{9L}{\mu_x N_1^2}\left\|x^0 - x_*\right\|_2^2.$$

Поэтому, выбирая $N_1 = \left\lceil 3e\sqrt{\frac{L}{\mu_x}} \right\rceil$, получаем

$$\left\|x^{N_1} - x_*\right\|_2^2 \le \frac{1}{2}\left\|x^0 - x_*\right\|_2^2.$$

(Данный способ выбора количества итераций авторам подсказал Роланд Хильдебранд.)

После этого выберем для алгоритма 3 в качестве точки старта x^{N_1} и снова сделаем N_1 итераций и т.д. Ясно, что для достижения приемлемого качества решения можно выбрать количество рестартов p алгоритма 3 для алгоритма 4 следующим образом:

$$p = \left\lceil \frac{1}{2}\ln\left(\frac{\mu_x R^2}{\varepsilon}\right) \right\rceil,$$

где $R^2 = \frac{1}{2}\left\|x^0 - x_*\right\|_2^2$. В таком случае общее число итераций алгоритма 4 в предложенной схеме будет

$$N = \left\lceil \ln\left(\frac{\mu_x R^2}{\varepsilon}\right) \right\rceil \cdot \left\lceil \frac{3e}{2}\sqrt{\frac{L}{\mu_x}} \right\rceil,$$

т.е.

$$N = O\left(\sqrt{\frac{L}{\mu_x}}\left\lceil \ln\left(\frac{\mu_x R^2}{\varepsilon}\right) \right\rceil\right).$$

Итак, ясно, что при $\delta = O\left(\mu_x \varepsilon \sqrt{\frac{\mu_x}{L}}\right)$ предложенной схемой возможно получить точность $f(\hat{x}) - f(x_*) \le \varepsilon$ для выхода $\hat{x}$ алгоритма 4.

Далее, неравенство (30) (см. также неравенство (2.20) из [16]) означает, что функционал f в произвольной точке допускает $(2\gamma, L, \mu_x)$-оракул в смысле [29].

Таким образом, точность γ для решения внутренней задачи (аналог точности δ в концепции (δ, L)-градиента) нужно выбирать как $O(\varepsilon)$ и тогда для внешней задачи будет гарантирована точность ε по функции. Это означает, что общее количество вызовов оракула для $\nabla_x S(\cdot,\cdot)$ при решении внешней задачи (29) будет равно

$$O\left(\sqrt{\frac{L_{xx}}{\mu_x} + \frac{2L_{xy}^2}{\mu_x\mu_y}}\ln\frac{1}{\varepsilon}\right).$$

Количество же вызовов оракула для $\nabla_y S(\cdot,\cdot)$ при решении внутренней задачи будет равно в силу (32)

$$O\left(\sqrt{\frac{L_{yy}}{\mu_y}}\sqrt{\frac{L_{xx}}{\mu_x} + \frac{2L_{xy}^2}{\mu_x\mu_y}}\ln^2\frac{1}{\varepsilon}\right).$$

Замечание 2. Согласно известной теореме о минимаксе сильная выпукло-вогнутость функционала $S(\cdot,\cdot)$ означает, что

$$\min_{x\in Q_x}\{f(x) = \max_{y\in Q_y} S(x,y)\} = \max_{y\in Q_y}\{l(x) = \min_{x\in Q_x} S(x,y)\}$$

и можно свести задачу (27) к задаче вогнутой максимизации

$$\tilde{f}(y) := \min_{x\in Q_x} S(x,y) \to \max_{y\in Q_y}.$$

В таком случае количество вызовов $\nabla_y S(\cdot,\cdot)$ будет

$$O\left(\sqrt{\frac{L_{yy}}{\mu_y} + \frac{2L_{xy}^2}{\mu_x\mu_y}}\ln\frac{1}{\varepsilon}\right),$$

а количество вызовов $\nabla_x S(\cdot,\cdot)$

$$O\left(\sqrt{\frac{L_{xx}}{\mu_x}}\sqrt{\frac{L_{yy}}{\mu_y}+\frac{2L_{xy}^2}{\mu_x^2\mu_y}}\ln^2\frac{1}{\varepsilon}\right).$$

Утверждение i) в случае, когда какая-то из функций $r(x)$ или $h(y)$ не проксимально дружественна, получается похожей схемой рассуждений. Но при этом мы получим вспомогательные подзадачи вида (24) с сильно выпуклой целевой функцией, которые можно решать с использованием результата об ускоренном градиентном слайдинге. Сформулированные оценки утверждения i), таким образом, вытекают из леммы 3.

Отметим, что для утверждения леммы 3 о сложности решения вспомогательной подзадачи (24) в случае сильной выпуклости r и g не имеет значения, параметр сильной выпуклости какого из этих функционалов использовать в оценках из леммы 3. Поэтому утверждение ii) теперь будет следовать из леммы 1 (где приводится оценка на параметр сильной выпуклости g).

Скажем несколько слов про обоснование утверждения iii). Мы отправляемся от [24].

Ясно, что решаемую задачу минимизаци функции $f(x) = r(x) + g(x)$ на множестве $Q_x = \mathbb{R}^n$ можно рассматривать как задачу композитной оптимизации с композитом g. Точнее говоря, ввиду μ_x-сильной выпуклости и L_x-липшицевости градиента r можно свести задачу минимизации f к рассмотрению семейства $k = \tilde{O}\left(\sqrt{\frac{L_x}{\mu_x}}\right)$ вспомогательных задач вида

$$\alpha_i(\langle\nabla r(x_i), x - x_i\rangle + g(x) - g(x_i)) + \frac{L_x\|x - x_i\|_2^2}{2} \to \min_{x\in Q_x}, \quad i = 1,2,\ldots,k. \tag{34}$$

Поскольку g уже, вообще говоря, не имеет простой структуры, то необходимо учитывать сложность каждой из k задач (34). Поскольку вдоль ядра $\operatorname{Ker} A$ функция g принимает постоянное значение, то без уменьшения общности рассуждений можно считать, что $x \in (\operatorname{Ker} A)^\perp$. Дело в том, что вдоль всякого направления, ортогонального $\operatorname{Ker} A$, вспомогательная задача будет иметь обусловленность 1 и не представляет трудоемкости в предположении, что известно $\operatorname{Ker} A$ (а стало быть, и ортогональные направления). Тогда согласно лемме 1 трудоемкость нахождения приемлемого качества решения каждой из задач (34) (линейные слагаемые не влияют на нее) будет определяться числом обусловленности

$$\frac{\alpha_i\frac{\lambda_{\max}(A^{\mathrm{T}}A)}{\mu_y} + L_x}{\alpha_i\frac{\lambda^+_{\min}(A^{\mathrm{T}}A)}{L_y} + L_x} \le \frac{L_y\lambda_{\max}(A^{\mathrm{T}}A)}{\mu_y\lambda^+_{\min}(A^{\mathrm{T}}A)},$$

поскольку для произвольных $a \ge b > 0$ и $c > 0$ верно неравенство $\frac{a+c}{b+c} \le \frac{a}{b}$. Таким образом, верна оценка (22).

5. ЗАКЛЮЧЕНИЕ

В данной работе рассмотрен класс гладких седловых задач со структурой. Такие задачи возникают, например, в обработке изображений и при решении различных обратных задач [9]. С помощью оценок ускоренного слайдинга Дж. Лана [9] и результатов о том, как преобразуются свойства гладкости и сильной выпуклости (кривизны графика функции, определяемые экстремальными значениями ее гессиана) при преобразовании Фенхеля–Лежандра [26], были получены новые (существенно лучшие) оценки на число вычислений градиента по прямым переменным в нелинейных седловых задачах со структурой. По сути, были получены достаточно общие условия, при которых удается перенести основные результаты для билинейных седловых задач на общий класс нелинейных гладких выпукло-вогнутых седловых задач.

Следует отметить, что в тексте статьи мы намерено избегали максимальной общности изложения для компактности и удобства восприятия. Тем не менее отметим, что приведенные в статье результаты можно обобщить на случай более общих оракулов [4]: стохастических, рандомизированных (в том числе инкрементальных, возникающих при работе с целевыми функционала-

ми вида суммы функций), неполноградиентных, модельных (в том числе с дополнительными композитными членами); вместо евклидовой нормы, можно было рассматривать более общие нормы и прокс-структуры (впрочем, все же не в такой общности, как в не сильно выпуклом случае); наконец, можно попробовать рассмотреть более чем два слагаемых в структуре целевого функционала. Также весьма интересен вопрос о том, в какой степени можно перенести полученные результаты на классы негладких седловых задач. Многое упирается в возможность должным образом обобщить лемму 3. Нам представляется, что на этом пути может быть получено достаточно много новых интересных результатов.

Отметим также, что если в приведенных в статье постановках задач $\dim(y)$ и(или) $\dim(x)$ небольшие, то у использованных в статье ускоренных методов появляются конкуренты в виде методов типа центров тяжести [20], [21]. В качестве примера можно посмотреть случай $\dim(y) = 2$, разобранный в [27].

Несмотря на полученное в работе улучшение известных верхних оценок, мы все же по-прежнему не знаем, оптимальны ли приведенные в этой статье оценки по совокупности критериев: число вычислений градиента по прямым переменным и по двойственным (не только по одному из них)? Недавно появился препринт [30], в котором получены нижние оценки для гладких сильно выпукло-вогнутых седловых задач. Однако при этом была показана достижимость этих оценок только на специальном подклассе задач. В общем случае предложенная в отмеченном препринте [30] методика (Section 4.2) приводит к оценкам, аналогичным полученным нами. Однако мы обосновываем указанные верхние оценки для более широкого класса задач, не предполагающих проксимальную дружественность h и r.

Авторы выражают благодарность А.С. Немировскому, Ю.Е. Нестерову и Р. Хильдебранду за ценное обсуждение части материала статьи.

ПРИЛОЖЕНИЕ

П1. Доказательство леммы 1

Ясно, что $g(x) = h^*(Ax)$, где h^* – сопряженная функция к h. По теореме Демьянова–Данскина $\nabla g(x) = A^{\mathrm{T}} y(x)$, где $y(x) = \langle Ax, y(x)\rangle - h(y(x))$ (т.е. $y(x) = \arg\max_y \{\langle Ax, y\rangle - h(y)\}$).

Пусть $h(y)$ есть μ_y – сильно выпуклая. Тогда в силу выбора $y(x)$ для всяких $x_1, x_2 \in Q_x$ получим

$$\langle Ax_1, y(x_2)\rangle - h(y(x_2)) \le \langle Ax_1, y(x_1)\rangle - hy(x_1) - \frac{\mu_y}{2}\|y(x_2) - y(x_1)\|_2^2,$$

$$\langle Ax_2, y(x_1)\rangle - h(y(x_1)) \le \langle Ax_2, y(x_2)\rangle - h(y(x_2)) - \frac{\mu_y}{2}\|y(x_1) - y(x_2)\|_2^2.$$

После сложения этих двух неравенств имеем

$$\langle Ax_1 - Ax_2, y(x_2) - y(x_1)\rangle \le -\mu_y \|y(x_2) - y(x_1)\|_2^2,$$

откуда

$$\mu_y \|y(x_2) - y(x_1)\|_2^2 \le \langle Ax_2 - Ax_1, y(x_2) - y(x_1)\rangle \le \|Ax_2 - Ax_1\|_2 \cdot \|y(x_2) - y(x_1)\|_2,$$

т.е. для нормы матрицы $\|A\|_2$ получим

$$\|y(x_2) - y(x_1)\|_2 \le \frac{\|A\|_2 \|x_2 - x_1\|_2}{\mu_y}.$$

Поэтому

$$\|\nabla g(x_1) - \nabla g(x_2)\|_2 \le \|A^{\mathrm{T}}\|_2 \|y(x_1) - y(x_2)\|_2 \le \frac{\|A^{\mathrm{T}}A\|_2}{\mu_y}\|x_1 - x_2\|_2 = \frac{\lambda_{\max}(A^{\mathrm{T}}A)}{\mu_y}\|x_1 - x_2\|_2.$$

Проверим теперь вторую часть утверждения. Пусть теперь $x_1, x_2 \in (\operatorname{Ker} A)^{\perp}$.

Хорошо известно, что для сопряженной функции

$$h^*(x) = \max_y \{\langle x, y\rangle - h(y)\} = \langle x, \hat{y}_x\rangle - h(\hat{y}_x),$$

верно следующее:

$$\hat{y}_x \in \partial h^*(x) \Leftrightarrow x \in \partial h(\hat{y}_x),$$

откуда ($x \to Ax, \hat{y}_x \to y(x)$) верно

$$y(x) \in \partial h^*(Ax) \Leftrightarrow Ax \in \partial h(y(x)).$$

Тогда имеем

$$\langle \nabla g(x_1) - \nabla g(x_2), x_1 - x_2\rangle = \langle A^{\mathrm{T}} y(x_1) - A^{\mathrm{T}} y(x_2), x_1 - x_2\rangle = \langle y(x_1) - y(x_2), Ax_1 - Ax_2\rangle =$$
$$= \langle Ax_1 - Ax_2, y(x_1) - y(x_2)\rangle \geq \{\text{из } L_y - \text{гладкости } h\} \geq \frac{1}{L_y}\|Ax_1 - Ax_2\|_2^2 =$$
$$= \frac{1}{L_y}\langle A^{\mathrm{T}} A(x_1 - x_2), x_1 - x_2\rangle \geq \{\text{из } x_1 - x_2 \notin \operatorname{Ker} A \text{ (т.к. } x_1, x_2 \in \operatorname{Ker} A^{\perp})\} \geq \frac{\lambda^+_{\min}(A^{\mathrm{T}}A)}{L_y}\|x_1 - x_2\|_2^2,$$

что обосновывает -сильную выпуклость $g(x)$ при $x \in (\operatorname{Ker} A)^{\perp}$.

П2. Доказательство леммы 2

Функция $\hat{S}(x, \cdot)$ есть μ_y-сильно вогнута на Q_y, и $\hat{S}(\cdot, y)$ дифференцируема на Q_x. Поэтому по теореме Демьянова–Данскина, для любого $x \in Q_x$ имеем

$$\nabla g(x) = \nabla_x \tilde{S}(x, y^*(x)) = \nabla_x F(x, y^*(x)). \tag{П1}$$

Чтобы доказать, что $g(\cdot)$ имеет L–липшицев градиент при $L = L_{xx} + \frac{2L_{xy}^2}{\mu_y}$, покажем условие Липшица для $y^*(\cdot)$ (функция y^* определена в (9)) с константой $\frac{2L_{xy}}{\mu_y}$.

Ввиду того, что $\hat{S}(x_1, \cdot)$ есть μ_y-сильно вогнута на Q_y, для произвольных $x_1, x_2 \in Q_x$:

$$\|y^*(x_1) - y^*(x_2)\|_2^2 \leq \frac{2}{\mu_y}\left(\hat{S}(x_1, y^*(x_1)) - \hat{S}(x_1, y^*(x_2))\right). \tag{П2}$$

С другой стороны, $\hat{S}(x_2, y^*(x_1)) - \hat{S}(x_2, y^*(x_2)) \leq 0$, т.к. $y^*(x_2)$ доставляет максимальное значение $\hat{S}(x_2, .)$ на Q_y. Имеем

$$\hat{S}(x_1, y^*(x_1)) - \hat{S}(x_1, y^*(x_2)) \leq \left(\hat{S}(x_1, y^*(x_1)) - \hat{S}(x_1, y^*(x_2))\right) - \left(\hat{S}(x_2, y^*(x_1)) + \hat{S}(x_2, y^*(x_2))\right) \overset{\text{из (7)}}{=}$$
$$= \left(F(x_1, y^*(x_1)) - F(x_1, y^*(x_2))\right) - \left(F(x_2, y^*(x_1)) - F(x_2, y^*(x_2))\right) =$$
$$= \int_0^1 \langle \nabla_x F(x_1 + t(x_2 - x_1), y^*(x_1)) - \nabla_x F(x_1 + t(x_2 - x_1), y^*(x_2)), x_2 - x_1\rangle dt \leq \tag{П3}$$
$$\leq \|\nabla_x F(x_1 + t(x_2 - x_1), y^*(x_1)) - \nabla_x F(x_1 + t(x_2 - x_1), y^*(x_1))\|_2 \cdot \|x_2 - x_1\|_2 \overset{\text{из (5)}}{\leq}$$
$$\leq L_{xy}\|y^*(x_1) - y^*(x_2)\|_2 \cdot \|x_2 - x_1\|_2.$$

Таким образом, из (П2) и (П3) вытекает неравенство

$$\|y^*(x_2) - y^*(x_1)\|_2 \leq \frac{2L_{xy}}{\mu_y}\|x_2 - x_1\|_2, \tag{П4}$$

т.е. функция $y^*(\cdot)$ удовлетворяет условию Липшица с константой $\frac{2L_{xy}}{\mu_y}$. Далее, из (П1) получаем

$$\begin{aligned}\left\|\nabla g(x_1) - \nabla g(x_2)\right\|_2 &= \left\|\nabla_x F(x_1, y^*(x_1)) - \nabla_x F(x_2, y^*(x_2))\right\|_2 = \\ &= \left\|\nabla_x F(x_1, y^*(x_1)) - \nabla_x F(x_1, y^*(x_2)) + \nabla_x F(x_1, y^*(x_2)) - \nabla_x F(x_2, y^*(x_2))\right\|_2 \le \\ &\le \left\|\nabla_x F(x_1, y^*(x_1)) - \nabla_x F(x_1, y^*(x_2))\right\|_2 + \left\|\nabla_x F(x_1, y^*(x_2)) - \nabla_x F(x_2, y^*(x_2))\right\|_2 \overset{\text{из (4) и (5)}}{\le} \\ &\le L_{xy}\left\|y^*(x_1) - y^*(x_2)\right\|_2 + L_{xx}\left\|x_2 - x_1\right\|_2 \overset{\text{из (П4)}}{=} \left(L_{xx} + \frac{2L_{xy}^2}{\mu_y}\right)\left\|x_2 - x_1\right\|_2.\end{aligned}$$

Это означает, что $g(\cdot)$ имеет L-липшицев градиент при $L = L_{xx} + \frac{2L_{xy}^2}{\mu_y}$.

Проверим теперь неравенства из (23). Сначала докажем, что для любых $\delta \ge 0$ и $x \in Q_x$ верно

$$\left\|\nabla_x \hat{S}(x, \tilde{y}_\delta(x)) - \nabla g(x)\right\|_2 \le L_{xy}\sqrt{\frac{2\delta}{\mu_y}}. \tag{П5}$$

Для всякого $x \in Q_x$ верно $\nabla_x \hat{S}(x, \tilde{y}_\delta(x)) = \nabla_x F(x, \tilde{y}_\delta(x))$. Тогда

$$\begin{aligned}\left\|\nabla_x \hat{S}(x, \tilde{y}_\delta(x)) - \nabla g(x)\right\|_2^2 &= \left\|\nabla_x F(x, \tilde{y}_\delta(x)) - \nabla_x F(x, y^*(x))\right\|_2^2 \overset{\text{из (5)}}{\le} L_{xy}^2\left\|y^*(x) - \tilde{y}_\delta(x)\right\|_2^2 \overset{\text{из (П2)}}{\le} \\ &\le \frac{2L_{xy}^2}{\mu_y}\left(\hat{S}(x, y^*(x)) - \hat{S}(x, \tilde{y}_\delta(x))\right) \overset{\text{из (10)}}{\le} \frac{2\delta L_{xy}^2}{\mu_y},\end{aligned}$$

что обосновывает неравенство (П5).

Теперь в силу μ_x-сильной выпуклости $\hat{S}(\cdot, \tilde{y}_\delta(x))$ на Q_x для произвольных $x, z \in Q_x$ верно

$$g(z) \overset{\text{из (8)}}{\le} \hat{S}(z, \tilde{y}_\delta(x)) \ge \hat{S}(x, \tilde{y}_\delta(x)) + \left\langle \nabla_x \hat{S}(x, \tilde{y}_\delta(x)), z - x\right\rangle.$$

Таким образом,

$$0 \ge \hat{S}(x, \tilde{y}_\delta(x)) - g(z) + \left\langle \nabla_x \hat{S}(x, \tilde{y}_\delta(x)), z - x\right\rangle,$$

что доказывает левую часть (23). Чтобы доказать правую часть (23), отметим, что g выпуклая и имеет L-липшицев градиент на Q_x. Поэтому для произвольных $x, z \in Q_x$ имеем

$$\begin{aligned}g(z) &\le g(x) + < \nabla g(x), z - x > + \frac{L}{2}\left\|z - x\right\|_2^2 \overset{\text{из (10)}}{\le} \hat{S}(x, \tilde{y}_\delta(x)) + \delta + \frac{L}{2}\left\|z - x\right\|_2^2 + \\ &+ \left\langle \nabla g(x), z - x\right\rangle + \left\langle \nabla_x \hat{S}(x, \tilde{y}_\delta(x)), x - z\right\rangle - \left\langle \nabla_x \hat{S}(x, \tilde{y}_\delta(x)), x - z\right\rangle = \hat{S}(x, \tilde{y}_\delta(x)) + \delta + \\ &+ \left\langle \nabla_x \hat{S}(x, \tilde{y}_\delta(x)), z - x\right\rangle + \left\langle \nabla_x \hat{S}(x, \tilde{y}_\delta(x)) - \nabla g(x), x - z\right\rangle + \frac{L}{2}\left\|z - x\right\|_2^2 \overset{\text{из (П5)}}{\le} \\ &\le \hat{S}(x, \tilde{y}_\delta(x)) + \delta + \left\langle \nabla_x \hat{S}(x, \tilde{y}_\delta(x)), z - x\right\rangle + L_{xy}\sqrt{\frac{2\delta}{\mu_y}} \cdot \left\|z - x\right\|_2 + \frac{L}{2}\left\|z - x\right\|_2^2.\end{aligned}$$

Однако

$$L_{xy}\sqrt{\frac{2\delta}{\mu_y}} \cdot \left\|z - x\right\|_2 \le \frac{2\sqrt{\delta}L_{xy}}{\sqrt{\mu_y}}\left\|z - x\right\|_2 = 2\sqrt{\frac{L_{xy}^2}{\mu_y}\left\|z - x\right\|_2^2 \cdot \delta} \le \frac{L_{xy}^2}{\mu_y}\left\|z - x\right\|_2^2 + \delta$$

ввиду классического неравенства между средним арифметическим и средним геометрическим. Поэтому

$$g(z) \le \hat{S}(x, \tilde{y}_\delta(x)) + 2\delta + \left\langle \nabla_x \hat{S}(x, \tilde{y}_\delta(x)), z - x\right\rangle + \frac{L_{xy}^2}{\mu_y}\left\|z - x\right\|_2^2 + \frac{L}{2}\left\|z - x\right\|_2^2,$$

и поскольку $L = L_{xx} + \dfrac{2L_{xy}^2}{\mu_y}$, то $\dfrac{L_{xy}^2}{\mu_y} \le \dfrac{L}{2}$ и поэтому

$$g(z) \le \hat{S}(x, \tilde{y}_\delta(x)) + \left\langle \nabla_x \hat{S}(x, \tilde{y}_\delta(x)), z - x \right\rangle + 2\delta + L\|z - x\|_2^2.$$

Итак, имеем

$$g(z) - \hat{S}(x, \tilde{y}_\delta x)) - \left\langle \nabla_x \hat{S}(x, \tilde{y}_\delta(x)), z - x \right\rangle \le L\|z - x\|_2^2 + 2\delta,$$

откуда вытекает справедливость левой части неравенства (23).

П3. Доказательство леммы 3

Напомним, что в утверждении леммы 3 рассматривается задача минимизации

$$\min_{x \in \mathbb{R}^n} P(x) := r(x) + g(x), \tag{П6}$$

где функция $r(x)$ есть μ_r-сильно выпуклая и L_r-гладкая для $L_r \ge \mu_r \ge 0$, функция $g(x)$ есть μ_g-сильно выпуклая и L_g-гладкая для $L_g \ge \mu_g \ge 0$, функция $P(x)$ есть μ-сильно выпуклая и L-гладкая при $L = L_r + L_g \ge \mu = \mu_r + \mu_g > 0$. Обозначим через x^* искомую точку минимума функционала P.

Докажем лемму 3 в предположении, что функция $r(x)$ допускает в произвольной запрошенной точке (δ_r, L_r, μ_r)-градиент $\nabla r_{\delta_r}(x)$, а функция $g(x)$ допускает – (δ_g, L_g, μ_g)-градиент $\nabla g_{\delta_g}(x)$. Это означает, что для произвольных $x, y \in \mathbb{R}^n$ выполнены следующие неравенства:

$$\begin{aligned} \frac{\mu_r}{2}\|x - y\|_2^2 - \delta_r &\le r(x) - r(y) - \left\langle \nabla r_{\delta_r}(y), x - y \right\rangle \le \frac{L_r}{2}\|x - y\|_2^2 + \delta_r, \\ \frac{\mu_g}{2}\|x - y\|_2^2 - \delta_g &\le g(x) - g(y) - \left\langle \nabla g_{\delta_g}(y), x - y \right\rangle \le \frac{L_g}{2}\|x - y\|_2^2 + \delta_g, \end{aligned} \tag{П7}$$

где $\delta_r \ge 0$ и $\delta_g \ge 0$.

На самом деле для обоснования основных результатов работы достаточно утверждение леммы 3 для менее ограничительной концепции (δ, L)-градиента в предположении сильной выпуклости g и r. Поскольку предполагается, что как r, так и g допускает неточные значения градиентов в запрашиваемых точках, то для определенности можно положить $L_r \le L_g$.

Алгоритм 5. Ускоренный проксимальный градиентный метод с неточными значениями градиентов.

1: **Параметры:** $x^0 \in \mathbb{R}^n$, шаги $\alpha, \beta \in (0,1)$, $\eta > 0$.

2: $y^0 = z^0 = x^0$.

3: **for** $k = 0, 1, 2, \ldots$ **do**

4: $\quad x^k = \alpha z^k + (1 - \alpha) y^k$

5: $\quad y^{k+1} \approx \hat{y}^{k+1} := \mathrm{prox}_{\frac{1}{L_r} g(\cdot)}\left(x^k - \dfrac{1}{L_r}\nabla r_{\delta_r}(x^k)\right)$, ($y^{k+1}$ – приближенное значение данного оператора, найденное посредством решения вспомогательной задачи оптимизации быстрым градиентным методом).

6: $\quad z^{k+1} = \beta z^k + (1 - \beta) x^k + \eta(y^{k+1} - x^k)$.

7: **end for**

Будем применять к рассмотренной задаче (П6) следующий метод, который подразумевает решение вспомогательной подзадачи быстрым градиентным методом при условии неточно заданного (δ_g, L_g, μ_g)-градиента g.

Докажем необходимую вспомогательную оценку для параметров x^k и y^{k+1} при произвольном $x \in \mathbb{R}^n$.

Утверждение 1. *Для всякого $x \in \mathbb{R}^n$ выполнено следующее неравенство*:

$$\left\langle x^k - y^{k+1}, x - x^k \right\rangle \le \frac{1}{L_r + \mu_g}\left[P(x) - P(y^{k+1}) - \frac{\mu}{4}\left\| x - x^k \right\|_2^2 - \frac{L_r + \mu_g}{4}\left\| y^{k+1} - x^k \right\|_2^2 + 2\delta_r \right] + c_1 \left\| y^{k+1} - y^{k+1} \right\|_2^2, \quad (\text{П8})$$

где константа c_1 определяется следующим выражением:

$$c_1 = 2\left[\frac{L_r}{\mu} + 1 \right]\left[\frac{L_g^2}{L_r^2} + 1 \right].$$

Доказательство. Из определения $\hat{y}^{k+1}$ следует, что

$$\hat{y}^{k+1} = x^k - \frac{1}{L_r}\nabla r_{\delta_r}(x^k) - \frac{1}{L_r}\nabla g(\hat{y}^{k+1}).$$

В силу предположения (П7) и μ_g-сильной выпуклости функции $g(x)$ имеем

$$\left\langle x^k - y^{k+1}, x - x^k \right\rangle = \left\langle x^k - \hat{y}^{k+1} + \hat{y}^{k+1} - y^{k+1}, x - x^k \right\rangle = \frac{1}{L_r}\left\langle \nabla r_{\delta_r}(x^k) + \nabla g(\hat{y}^{k+1}), x - x^k \right\rangle +$$

$$+ \left\langle \hat{y}^{k+1} - y^{k+1}, x - x^k \right\rangle = \frac{1}{L_r}\left\langle \nabla r_{\delta_r}(x^k) + \nabla g(y^{k+1}), x - x^k \right\rangle +$$

$$+ \left\langle \frac{1}{L_r}[\nabla g(\hat{y}^{k+1}) - \nabla g(y^{k+1})] + \hat{y}^{k+1} - y^{k+1}, x - x^k \right\rangle = \frac{1}{L_r}\left\langle \nabla r_{\delta_r}(x^k), x - x^k \right\rangle + \frac{1}{L_r}\left\langle \nabla g(y^{k+1}), x - y^{k+1} \right\rangle +$$

$$+ \frac{1}{L_r}\left\langle \nabla g(y^{k+1}), y^{k+1} - x^k \right\rangle + \left\langle \frac{1}{L_r}[\nabla g(\hat{y}^{k+1}) - \nabla g(y^{k+1})] + \hat{y}^{k+1} - y^{k+1}, x - x^k \right\rangle \le$$

$$\le \frac{1}{L_r}\left[r(x) - r(x^k) - \frac{\mu_r}{2}\left\| x - x^k \right\|_2^2 + \delta_r \right] + \frac{1}{L_r}\left[g(x) - g(y^{k+1}) - \frac{\mu_g}{2}\left\| x - y^{k+1} \right\|_2^2 \right] +$$

$$+ \frac{1}{L_r}\left\langle \nabla g(y^{k+1}), y^{k+1} - x^k \right\rangle + \left\langle \frac{1}{L_r}[\nabla g(\hat{y}^{k+1}) - \nabla g(y^{k+1})] + \hat{y}^{k+1} - y^{k+1}, x - x^k \right\rangle.$$

Далее, применим правую часть неравенства (П7) для $r(x)$:

$$r(y^{k+1}) \le r(x^k) + \left\langle \nabla r_{\delta_r}(x^k), y^{k+1} - x^k \right\rangle + \frac{L_r}{2}\left\| y^{k+1} - x^k \right\|_2^2 + \delta_r, v$$

откуда следует

$$\left\langle x^k - y^{k+1}, x - x^k \right\rangle \le \frac{1}{L_r}\left[r(x) - r(y^{k+1}) - \frac{\mu_r}{2}\left\| x - x^k \right\|_2^2 + 2\delta_r \right] + \frac{1}{L_r}\left[g(x) - g(y^{k+1}) - \frac{\mu_g}{2}\left\| x - y^{k+1} \right\|_2^2 \right] +$$

$$+ \frac{1}{L_r}\left\langle \nabla r_{\delta_r}(x^k) + \nabla g(y^{k+1}), y^{k+1} - x^k \right\rangle + \frac{1}{2}^{k+1} - x^k 2^2 + \left\langle \frac{1}{L_r}[\nabla g(\hat{y}^{k+1}) - \nabla g(y^{k+1})] + \hat{y}^{k+1} - y^{k+1}, x - x^k \right\rangle =$$

$$= \frac{1}{L_r}\left[P(x) - P(y^{k+1}) - \frac{\mu_r}{2}\left\| x - x^k \right\|_2^2 - \frac{\mu_g}{2}\left\| x - y^{k+1} \right\|_2^2 - \frac{L_r}{2}\left\| y^{k+1} - x^k \right\|_2^2 + 2\delta_r \right] +$$

$$+ \frac{1}{L_r}\left\langle \nabla r_{\delta_r}(x^k) + \nabla g(y^{k+1}), y^{k+1} - x^k \right\rangle + \left\| y^{k+1} - x^k \right\|_2^2 + \left\langle \frac{1}{L_r}[\nabla g(\hat{y}^{k+1}) - \nabla g(y^{k+1})] + \hat{y}^{k+1} - y^{k+1}, x - x^k \right\rangle =$$

$$= \frac{1}{L_r}\left[P(x) - P(y^{k+1}) - \frac{\mu_r}{2}\left\| x - x^k \right\|_2^2 - \frac{\mu_g}{2}\left\| x - y^{k+1} \right\|_2^2 - \frac{L_r}{2}\left\| y^{k+1} - x^k \right\|_2^2 + 2\delta_r \right] +$$

$$+ \left\langle \frac{1}{L_r}[\nabla g(\hat{y}^{k+1}) - \nabla g(y^{k+1})] + \hat{y}^{k+1} - y^{k+1}, x^k - y^{k+1} \right\rangle + \left\langle \frac{1}{L_r}[\nabla g(\hat{y}^{k+1}) - \nabla g(y^{k+1})] + \hat{y}^{k+1} - y^{k+1}, x - x^k \right\rangle.$$

Применим теперь неравенство Юнга, а также L_g-липшицевость градиента $\nabla g(x)$, $\left\|\nabla g(y^{k+1}) - \nabla g(\hat{y}^{k+1})\right\|_2^2 \le L_g^2 \left\|\hat{y}^{k+1} - y^{k+1}\right\|_2^2$:

$$\left\langle x^k - y^{k+1}, x - x^k\right\rangle \le \frac{1}{L_r}\left[P(x) - P(y^{k+1}) - \frac{\mu_r}{2}\left\|x - x^k\right\|_2^2 - \frac{\mu_g}{2}\left\|x - y^{k+1}\right\|_2^2 - \frac{L_r}{2}\left\|y^{k+1} - x^k\right\|_2^2\right] +$$

$$+ \frac{2\delta_r}{L_r} + \frac{\mu}{4L_r}\left\|x - x^k\right\|_2^2 + \frac{L_r + \mu_g}{4L_r}\left\|y^{k+1} - x^k\right\|_2^2 + \left[\frac{L_r}{\mu} + \frac{L_r}{L_r + \mu_g}\right]\left\|\frac{1}{L_r}[\nabla g(\hat{y}^{k+1}) - \nabla g(y^{k+1})] + \hat{y}^{k+1} - y^{k+1}\right\|_2^2 \le$$

$$\le \frac{1}{L_r}\left[P(x) - P(y^{k+1}) - \frac{\mu_r - \mu_g}{4}\left\|x - x^k\right\|_2^2 - \frac{\mu_g}{2}\left\|x - y^{k+1}\right\|_2^2 - \frac{L_r - \mu_g}{4}\left\|y^{k+1} - x^k\right\|_2^2\right] +$$

$$+ \frac{2\delta_r}{L_r} + 2\left[\frac{L_r}{\mu} + \frac{L_r}{L_r + \mu_g}\right]\left[\frac{L_g^2}{L_r^2} + 1\right]\left\|\hat{y}^{k+1} - y^{k+1}\right\|_2^2.$$

Наконец, проведем финальные преобразования:

$$\left\langle x^k - y^{k+1}, x - x^k\right\rangle = \frac{L_r}{L_r + \mu_g}\left\langle x^k - y^{k+1}, x - x^k\right\rangle + \frac{\mu_g}{L_r + \mu_g}\left\langle x^k - y^{k+1}, x - x^k\right\rangle \le$$

$$\le \frac{1}{L_r + \mu_g}\left[P(x) - P(y^{k+1}) - \frac{\mu_r - \mu_g}{4}\left\|x - x^k\right\|_2^2 - \frac{L_r - \mu_g}{4}\left\|y^{k+1} - x^k\right\|_2^2 + 2\delta_r\right] - \frac{\mu_g}{2\left(L_r + \mu_g\right)}\left\|x - y^{k+1}\right\|_2^2 +$$

$$+ \frac{2L_r}{L_r + \mu_g}\left[\frac{L_r}{\mu} + \frac{L_r}{L_r + \mu_g}\right]\left[\frac{L_g^2}{L_r^2} + 1\right]\left\|\hat{y}^{k+1} - y^{k+1}\right\|_2^2 + \frac{\mu_g}{2\left(L_r + \mu_g\right)}\left[\left\|x - y^{k+1}\right\|_2^2 - \left\|x^k - y^{k+1}\right\|_2^2 - \left\|x - x^{k2}_2\right\|\right] \le$$

$$\le \frac{1}{L_r + \mu_g}\left[P(x) - P(y^{k+1}) - \frac{\mu_r + \mu_g}{4}\left\|x - x^k\right\|_2^2 - \frac{L_r + \mu_g}{4}\left\|y^{k+1} - x^k\right\|_2^2 + 2\delta_r\right] +$$

$$+ 2\left[\frac{L_r}{\mu} + 1\right]\left[\frac{L_g^2}{L_r^2} + 1\right]\left\|\hat{y}^{k+1} - y^{k+1}\right\|_2^2.$$

Утверждение 2. *Пусть выбраны следующие значения параметров для алгоритма* 5:

$$\eta = \frac{2(L_r + \mu_g)}{8\alpha(L_r + \mu_g) + (1 - \alpha)\mu},$$

$$\beta = 1 - \frac{\eta\mu}{2(L_r + \mu_g)} = 1 - \frac{\mu}{8\alpha(L_r + \mu_g) + (1 - \alpha)\mu},$$

$$\alpha = \frac{1}{4}\sqrt{\frac{\mu}{L_r + \mu_g}} \le \frac{1}{4}.$$

Тогда выполнено следующее неравенство:

$$\left\|z^{k+1} - x^*\right\|_2^2 + c_2[P(y^{k+1}) - P(x^*)] \le (1 - \alpha)\left(\left\|z^k - x^*\right\|_2^2 + c_2[P(y^{k+1}) - P(x^*)]\right) +$$
$$+ c_3\left[8c_1\left\|y^{k+1} - \hat{y}^{k+1}\right\|_2^2 - \left\|y^{k+1} - x^k\right\|_2^2\right] + 4c_2\delta_r. \qquad (\text{П9})$$

где c_2 и c_3 – некоторые положительные константы.

Доказательство. Оценим величину $\left\|z^{k+1} - x^*\right\|_2^2$:

$$\left\|z^{k+1} - x^*\right\|_2^2 = \left\|\beta z^k + (1 - \beta)x^k - x^* + \eta(y^{k+1} - x^k)\right\|_2^2 = \left\|\beta(z^k - x^*) + (1 - \beta)(x^k - x^*)\right\|_2^2 +$$
$$+ \eta^2\left\|y^{k+1} - x^k\right\|_2^2 + 2\eta\left\langle\beta z^k + (1 - \beta)x^k - x^*, y^{k+1} - x^k\right\rangle \le \beta\left\|z^k - x^*\right\|_2^2 + (1 - \beta)\left\|x^k - x^*\right\|_2^2 +$$

$$+\eta^2\left\|y^{k+1}-x^k\right\|_2^2+2\eta\beta\left\langle z^k-x^k,y^{k+1}-x^k\right\rangle+2\eta\left\langle x^k-x^*,y^{k+1}-x^k\right\rangle\le\beta\left\|z^k-x^*\right\|_2^2+$$
$$+(1-\beta)\left\|x^k-x^*\right\|_2^2+\eta^2\left\|y^{k+1}-x^k\right\|_2^2+2\eta\beta\frac{1-\alpha}{\alpha}\left\langle x^k-y^k,y^{k+1}-x^k\right\rangle+2\eta\left\langle x^k-x^*,y^{k+1}-x^k\right\rangle.$$

Далее, два раза применим неравенство (П8):

$$\left\|z^{k+1}-x^*\right\|_2^2\le\beta\left\|z^k-x^*\right\|_2^2+(1-\beta)\left\|x^k-x^*\right\|_2^2+\eta^2\left\|y^{k+1}-x^k\right\|_2^2+$$
$$+2\beta\frac{\eta}{L_r+\mu_g}\frac{1-\alpha}{\alpha}\left[P(y^k)-P(y^{k+1})-\frac{L_r+\mu_g}{4}\left\|y^{k+1}-x^k\right\|_2^2+2\delta_r\right]+$$
$$+2\frac{\eta}{L_r+\mu_g}\left[P(x^*)-P(y^{k+1})-\frac{\mu}{4}\left\|x^*-x^k\right\|_2^2-\frac{L_r+\mu_g}{4}\left\|y^{k+1}-x^k\right\|_2^2+2\delta_r\right]+$$
$$+2\eta c_1\left[\beta\frac{1-\alpha}{\alpha}+1\right]\left\|y^{k+1}-\hat y^{k+1}\right\|_2^2=\beta\left\|z^k-x^*\right\|_2^2+\left[1-\beta-\frac{\eta\mu}{2(L_r+\mu_g)}\right]\left\|x^k-x^*\right\|_2^2+$$
$$+\left[\eta^2-\frac{\eta\beta}{4}\frac{1-\alpha}{\alpha}-\frac{\eta}{4}\right]\left\|y^{k+1}-x^k\right\|_2^2+2\beta\frac{\eta}{L_r+\mu_g}\frac{1-\alpha}{\alpha}[P(y^k)-P(y^{k+1})]+$$
$$+2\frac{\eta}{L_r+\mu_g}[P(x^*)-P(y^{k+1})]+\frac{\eta}{4}\left[\beta\frac{1-\alpha}{\alpha}+1\right]\left[8c_1\left\|y^{k+1}-\hat y^{k+1}\right\|_2^2-\left\|y^{k+1}-x^k\right\|_2^2\right]+$$
$$+2\delta_r\left[2\beta\frac{\eta}{L_r+\mu_g}\frac{1-\alpha}{\alpha}+2\frac{\eta}{L_r+\mu_g}\right].$$

С учетом выбранных значений параметров β и η, а также

$$c_2=\frac{2\eta\beta}{\alpha(L_r+\mu_g)},\quad c_3=\frac{\eta}{4}\left[\beta\frac{1-\alpha}{\alpha}+1\right],$$

получаем

$$\left\|z^{k+1}-x^*\right\|_2^2\le\beta\left\|z^k-x^*\right\|_2^2+2\beta\frac{\eta}{L_r+\mu_g}\frac{1-\alpha}{\alpha}[P(y^k)-P(y^{k+1})]+2\frac{\eta}{L_r+\mu_g}[P(x^*)-P(y^{k+1})]+$$
$$+c_3\left[8c\left\|y_1^{k+1}-\hat y^{k+1}\right\|_2^2-\left\|y^{k+1}-x^k\right\|_2^2\right]+\frac{4\delta_r\eta}{\alpha(L_r+\mu_g)}\le\beta\left\|z^k-x^*\right\|_2^2+\frac{2\beta\eta}{L_r+\mu_g}\frac{1-\alpha}{\alpha}[P(y^k)-P(y^{k+1})]+$$
$$+\frac{2\beta\eta}{L_r+\mu_g}[P(x^*)-P(y^{k+1})]+c_3\left[8c_1\left\|y^{k+1}-\hat y^{k+1}\right\|_2^2-\left\|y^{k+1}-x^k\right\|_2^2\right]+\frac{4\delta_r\eta}{\alpha(L_r+\mu_g)}=$$
$$=\beta\left\|z^k-x^*\right\|_2^2+c_2(1-\alpha)[P(y^k)-P(y^{k+1})]+c_2\alpha[P(x^*)-P(y^{k+1})]+$$
$$+c_3\left[8c_1\left\|y^{k+1}-\hat y^{k+1}\right\|_2^2-\left\|y^{k+1}-x^k\right\|_2^2\right]+\frac{2c_2\delta_r}{\beta}=\beta\left\|z^k-x^*\right\|_2^2+c_2(1-\alpha)[P(y^k)-P(x^*)]+$$
$$+c_2[P(x^*)-P(y^{k+1})]+c_3\left[8c_1\left\|y^{k+1}-\hat y^{k+1}\right\|_2^2-\left\|y^{k+1}-x^k\right\|_2^2\right]+\frac{2c_2\delta_r}{\beta}.$$

Используя значение параметра α, получаем

$$\frac12\le\beta=1-\frac{\mu}{2\sqrt{(L_r+\mu_g)\mu}+(1-\alpha)\mu}\le1-\frac13\sqrt{\frac{\mu}{L_r+\mu_g}}\le1-\alpha,$$

откуда следует

$$\left\|z^{k+1}-x^*\right\|_2^2+c_2[P(y^{k+1})-P(x^*)]\le(1-\alpha)\left(\left\|z^k-x^*\right\|_2^2+c_2[P(y^k)-P(x^*)]\right)+$$
$$+c_3\left[8c_1\left\|y^{k+1}-\hat y^{k+1}\right\|_2^2-\left\|y^{k+1}-x^k\right\|_2^2\right]+4c_2\delta_r.$$

Теперь учтем, что вспомогательная задача строки 5 листинга алгоритма 5 решается быстрым градиентным методом с неточным заданием градиента g. Оценим необходимую точность δ_g градиента g для получения требуемого качества решения задачи по функции.

Утверждение 3. *Пусть приближение* y^{k+1} *прокс-оператора* $\hat{y}^{k+1} = \operatorname{prox}_{\frac{1}{L_r}g(\cdot)}\left(x^k - \frac{1}{L_r}\nabla r_{\delta_r}(x^k)\right)$ (*строка* 5 *листинга алгоритма* 5) *вычисляется быстрым градиентным методом в предположении, что доступен* (δ_g, μ_g, L_g)*-градиент* g *в произвольной запрошенной точке* [29]. *Решаемая задача минимизации при этом имеет вид*

$$\min_{x\in\mathbb{R}^n} g(x) + \frac{L_r}{2}\left\|x^k - \frac{1}{L_r}\nabla r_{\delta_r}(x^k) - x\right\|_2^2, \tag{П10}$$

где x^k *— начальное приближение. Тогда известно* [29], *что для произвольного* $\delta \in (0;1)$ *после*

$$T = O\left(\sqrt{\frac{L_r + L_g}{L_r + \mu_g}}\log\frac{L_r + L_g}{\delta(L_r + \mu_g)}\right),$$

итераций указанного метода гарантированно будет выполнено неравенство

$$\left\|y^{k+1} - \hat{y}^{k+1}\right\|_2^2 \le \delta\left\|x^k - \hat{y}^{k+1}\right\|_2^2 + c_4\delta_g, \tag{П11}$$

где константа c_4 *может быть задана выражением*

$$c_4 = \frac{4\sqrt{L_r + L_g}}{(L_r + \mu_g)\sqrt{L_r + \mu_g}}.$$

Доказательство. Отметим, что целевая функция задачи (П10) $(L_r + \mu_g)$-сильно выпуклая и $(L_r + L_g)$-гладкая, а $\hat{y}^{k+1}$ — точное решение задачи (П10). Неравенство (П11) следует из соответствующего результата для быстрого градиентного метода в концепции (δ_g, μ_g, L_g)-оракула для g [29].

Доказательство леммы 3. Выбрав в неравенстве (П11) $\delta = \frac{1}{32c_1} \le \frac{1}{4}$, получим

$$\left\|y^{k+1} - \hat{y}^{k+1}\right\|_2^2 \le 2\delta\left(\left\|x^k - y^{k+1}\right\|_2^2 + \left\|y^{k+1} - \hat{y}^{k+1}\right\|_2^2\right) + c_4\delta_g \le 2\delta\left\|x^k - y^{k+1}\right\|_2^2 + \frac{1}{2}\left\|y^{k+1} - \hat{y}^{k+1}\right\|_2^2 + c_4\delta_g,$$

откуда следует

$$\left\|y^{k+1} - \hat{y}^{k+1}\right\|_2^2 \le 4\delta\left\|x^k - y^{k+1}\right\|_2^2 + 2c_4\delta_g \le \frac{1}{8c_1}\left\|x^k - y^{k+1}\right\|_2^2 + 2c_4\delta_g.$$

С учетом доказанных неравенств (П9) означает, что

$$\left\|z^{k+1} - x^*\right\|_2^2 + c_2[P(y^{k+1}) - P(x^*)] \le (1-\alpha)\left(\left\|z^k - x^*\right\|_2^2 + c_2[P(y^k) - P(x^*)]\right) + 4c_2\delta_r + 2c_3c_4\delta_g,$$

откуда после телескопирования имеем

$$\left\|z^k - x^*\right\|_2^2 + c_2[P(y^k) - P(x^*)] \le (1-\alpha)^k\left(\left\|x^0 - x^*\right\|_2^2 + c_2[P(x^0) - P(x^*)]\right) + \frac{4c_2\delta_r + 2c_3c_4\delta_g}{\alpha}.$$

С учетом μ-сильной выпуклости функции $P(x)$ имеем

$$P(y^k) - P(x^*) \le (1-\alpha)^k\left(1 + \frac{2}{\mu c_2}\right)[P(x^0) - P(x^*)] + \frac{4\delta_r}{\alpha} + \frac{2c_3c_4\delta_g}{c_2\alpha} \le$$
$$\le 2(1-\alpha)^k[P(x^0) - P(x^*)] + \frac{4\delta_r}{\alpha} + \frac{2c_3c_4\delta_g}{c_2\alpha}.$$

Выбирая число итераций внешнего метода

$$k = \frac{1}{\alpha}\log\frac{4(P(x^0) - P(x^*))}{\varepsilon} = O\left(\sqrt{\frac{L_r + \mu_g}{\mu}}\log\frac{1}{\varepsilon}\right)$$

и точность (δ_r, L_r, μ_r)-градиента $\nabla r_{\delta_r}(x)$

$$\delta_r = \frac{\alpha\varepsilon}{16} = O\left(\sqrt{\frac{\mu}{L_r + \mu_g}}\varepsilon\right),$$

а также (δ_g, L_g, μ_g)-градиента $\nabla g_{\delta_g}(x)$

$$\delta_g = \frac{\alpha c_2\varepsilon}{8c_3c_4} = \frac{\alpha\varepsilon}{8c_4}\frac{2\eta\beta}{\alpha(L_r + \mu_g)}\frac{4\alpha}{\eta[(1-\alpha)\beta + \alpha]} \le \frac{\alpha\varepsilon}{c_4(1-\alpha)(L_r + \mu_g)} = \frac{\sqrt{L_r + \mu_g}\alpha\varepsilon}{4(1-\alpha)\sqrt{L_r + L_g}} =$$
$$= \frac{\sqrt{L_r + \mu_g}\sqrt{\mu}\varepsilon}{16(1-\alpha)\sqrt{L_r + L_g}\sqrt{L_r + \mu_g}} \le \frac{\varepsilon}{12}\sqrt{\frac{\mu}{L_r + L_g}} = O\left(\sqrt{\frac{\mu}{L_g + \mu_r}}\varepsilon\right),$$

где в последнем равенстве использовалось предположение $L_r \le L_g$ и $\alpha \le \frac{1}{4}$, получаем требуемое качество решения

$$P(y^k) - P(x^*) \le \varepsilon.$$

При этом количество вызовов (δ_r, L_r, μ_r)-градиента $\nabla r_{\delta_r}(x)$

$$k = O\left(\sqrt{\frac{L_r + \mu_g}{\mu}}\log\frac{1}{\varepsilon}\right),$$

а количество вызовов (δ_g, L_g, μ_g)-градиента $\nabla_{\delta_g} g(x)$

$$k \times T = O\left(\sqrt{\frac{L_r + \mu_g}{\mu}}\log\frac{1}{\varepsilon}\right) \times O\left(\sqrt{\frac{L_r + L_g}{L_r + \mu_g}}\log\frac{L_r + L_g}{\delta(L_r + \mu_g)}\right) = \tilde{O}\left(\sqrt{\frac{L_r + L_g}{\mu}}\log\frac{1}{\varepsilon}\right) = \tilde{O}\left(\sqrt{\frac{L_g + \mu_r}{\mu}}\log\frac{1}{\varepsilon}\right)$$

в силу допущения $L_r \le L_g$ (данное допущение не существенно ввиду симметричности найденных оценок на δ_r и δ_g).

СПИСОК ЛИТЕРАТУРЫ

1. *Нестеров Ю.Е.* Метод минимизации выпуклых функций со скоростью сходимости $O(1/k^2)$ // Докл. АН СССР. 1983. Т. 269. № 3. С. 543–547.
2. *Поляк Б.Е.* Введение в оптимизацию. М.: Наука, 1983. 384 с.
3. *Drori Y., Teboulle M.* Performance of first-order methods for smooth convex minimization: a novel approach // Math. Program. 2014. V. 145. № 1–2. P. 451–482.
4. *Гасников А.В.* Современные численные методы оптимизации. Метод универсального и градиентного спуска. М.: Изд-во МФТИ: 2018. 160 с. https://arxiv.org/abs/1711.00394
5. *Nemirovski A.* Lectures on Modern Convex Optimization analysis, algorithms, and engineering applications. Philadelphia: SIAM, 2015. http://www2.isye.gatech.edu/~nemirovs/Lect_ModConvOpt.pdf.
6. *Nesterov Yu.* Lectures on convex optimization. Switzerland: Springer Optimization and Its Applications, 2018.
7. *Taylor A.B., Hendrickx J.M., Glineur F.* Smooth strongly convex interpolation and exact worst-case performance of first-order methods // Math. Program. 2017. V. 161. № 1–2. P. 307–345.
8. *Гасников А.В.* Эффективные численные методы поиска равновесий в больших транспортных сетях: Дис. ... докт. физ.-матем. наук. М.: МФТИ, 2016, 487 с.
9. *Lan G.* First-order and Stochastic Optimization Methods for Machine Learning. Switzerland: Springer Series in the Data Sciences, 2020.
10. *Нестеров Ю.Е.* Алгоритмическая выпуклая оптимизация: Дис. ... докт. физ.-матем. наук. М.: МФТИ, 2013. 367 с.
11. *Devolder O., Glineur F., Nesterov Yu.* First-order methods of smooth convex optimization with inexact oracle // Math. Program. Ser. A. 2014. V. 146. № 1–2. P. 37–75.

12. *Hongzhou Lin*, *Julien Mairal, Zaid Harchaoui.* Catalyst Acceleration for First-order Convex Optimization: from Theory to Practice // J. of Machine Learning Research 2018. V. 18. P. 1–54.
13. *Nemirovski A.* Prox-method with rate of convergence $O(1/T)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems // SIAM Journal on Optim. 2004. V. 15. № 1. P. 229–251.
14. *Gasnikov A.V., Dvurechensky P.E., Stonyakin F.S., Titov A.A.* Adaptive proximal method for variational inequalities // Comput. Math. Math. Phys. 2019. V. 59. № 5. P. 836–841.
15. *Azizian W., Mitliagkas I., Lacoste-Julien S., Gidel G.* A Tight and Unified Analysis of Extragradient for a Whole Spectrum of Differentiable Games // Proceedings of Machine Learning Research. 2020. V. 108. P. 2863–2873.
16. *Hien L.T.K., Zhao R., Haskell W.B.* An inexact primal-dual framework for large-scale non bilinear saddle point problem // arxiv e-print 2019. https://arxiv.org/pdf/1711.03669.pdf.
17. *Гасников А.В., Двуреченский П.Е., Нестеров Ю.Е.* Стохастические градиентные методы с неточным оракулом // Тр. МФТИ. М., 2016. Т. 8. № 1. С. 41–91.
18. *Ouyang Y., Xu Y.* Lower complexity bounds of first-order methods for convexconcave bilinear saddle-point problems // Math. Program. 2019.
https://doi.org/10.1007/s10107-019-01420-0
19. *Жадан В.Г.* Методы оптимизации. Ч. 3. М.: МФТИ, 2017. 244 с.
20. *Немировский А.С., Юдин Д.Б.* Сложность задач и эффективность методов оптимизации. М.: Наука, 1979. 384 с.
21. *Bubeck S.* Convex optimization: algorithms and complexity // Foundations and Trends in Machine Learning. 2015. V. 8. № 3–4. P. 231–357. https://arxiv.org/pdf/1405.4980.pdf.
22. *Nemirovski A., Onn S., Rothblum U.G.* Accuracy certificates for computational problems with convex structure // Math. Oper. Res. 2010. V. 35. № 1. P. 52–78.
23. *Mokhtari A., Ozdaglar A., Pattathil S.* A Unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: proximal point approach // Proceedings of Machine Learning Research, 2020. V. 108. P. 1497–1507.
24. *Dvinskikh D., Gasnikov A.* Decentralized and Parallelized Primal and Dual Accelerated Methods for Stochastic Convex Programming Problems // arxiv e-print 2019. https://arxiv.org/pdf/1904.09015.pdf.
25. *Kakde S.M., Shalev-Shwartz S., Tewari A.* On the duality of strong convexity and strong smoothness: learning applications and matrix regularization // J. of Machine Learning Research 2012. V. 13. P. 1865–1890.
26. *Rockafellar R.T.* Convex analysis. Princeton: Princeton University Press, 1996.
27. *Гасников А.В., Камзалов Д.И., Мендель М.А.* Основные конструкции над алгоритмами выпуклой оптимизации и их приложения к получению новых оценок для сильно выпуклых задач // Тр. МФТИ. М., 2016. Т. 8. № 3. С. 25–42.
28. *Gasnikov A.V., Tyurin A.I.* Fast Gradient Descent for Convex Minimization Problems with an Oracle Producing a (δ, L)-Model of Function at the Requested Point // Comput. Math. and Math. Phys. 2019. V. 59. № 7. P. 1085–1097.
29. *Devolder O.* Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization. Louvain: CORE UCL, Ph.D. thesis, 2013.
30. *Zhang J., Hong M., Zhang S.* On Lower Iteration Complexity Bounds for the Saddle Point Problems // arxiv e-print 2019. https://arxiv.org/pdf/1912.07481.pdf.