© 2021 г. Е.Л. ГЛАДИН (gladin.el@phystech.edu),
М. АЛКУСА, канд. физ.-мат. наук (mohammad.alkousa@phystech.edu),
А.В. ГАСНИКОВ, д-р физ.-мат. наук (gasnikov.av@mipt.ru)
(Московский физико-технический институт, Долгопрудный;
Институт проблем передачи информации им. А. А. Харкевича РАН, Москва)

О РЕШЕНИИ ВЫПУКЛЫХ MIN-MIN ЗАДАЧ С ГЛАДКОСТЬЮ И СИЛЬНОЙ ВЫПУКЛОСТЬЮ ПО ОДНОЙ ИЗ ГРУПП ПЕРЕМЕННЫХ И МАЛОЙ РАЗМЕРНОСТЬЮ ДРУГОЙ[1]

Статья посвящена некоторым подходам к решению выпуклых задач вида min-min с гладкостью и сильной выпуклостью только по одной из двух групп переменных. Показано, что предложенные подходы, основанные на методе Вайды, быстром градиентном методе и ускоренном градиентном методе с редукцией дисперсии, имеют линейную сходимость. Для решения внешней задачи предлагается использовать методы Вайды, для решения внутренней (гладкой и сильно выпуклой) — быстрый градиентный метод. Ввиду важности для приложений в машинном обучении отдельно рассмотрен случай, когда целевая функция является суммой большого числа функций. В этом случае вместо быстрого градиентного метода используется ускоренный градиентный метод с редукцией дисперсии. Приведены результаты численных экспериментов, иллюстрирующие преимущества предложенных процедур для задачи логистической регрессии, в которой есть априорное распределение на одну из двух групп переменных.

Ключевые слова: выпуклая оптимизация, метод секущей плоскости, метод Вайды, редукция дисперсии, быстрый градиентный метод, логистическая регрессия.

DOI: 10.31857/S0005231021100068

1. Введение

Одним из основных направлений исследований численных методов 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. За последние 15 лет ускоренный метод был успешно перенесен на гладкие задачи условной выпуклой оптимизации, на задачи со структурой

[1] Работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации (госзадание) № 075-00337-20-03, номер проекта 0714-2020-0005. Работа А.В. Гасникова была также частично поддержана Российским фондом фундаментальных исследований (проект № 18-29-03071 мк).

(в частности, так называемые композитные задачи), безградиентные и рандомизированные методы (например, ускоренный градиентный метод с редукцией дисперсии для задач минимизации суммы функций [2]). Также ускорение было успешно перенесено на методы, использующие старшие производные. Детали и более подробный обзор публикаций можно найти в [3].

Задачи оптимизации вида min-max и седловые задачи широко изучались в литературе из-за их широкого спектра приложений в статистике, машинном обучении, компьютерной графике, теории игр и других областях. В последнее время многие исследователи активно работают над темой ускоренных методов решения этих задач, учитывающих их структуру: [4–8] — и это лишь некоторые из последних публикаций. В некоторых приложениях существует задача, аналогичная задаче min-max, которая остается в значительной степени неизученной — это задача вида min-min:

$$\min_{x \in Q_x} \min_{y \in Q_y} F(x, y), \tag{1}$$

где $Q_x \subset \mathbb{R}^d$, $Q_y \subset \mathbb{R}^n$ — непустые компактные выпуклые множества, размерность d относительно небольшая ($d \ll n$), функция $F(x, y)$ — выпуклая по совокупности переменных, а также L-гладкая и μ-сильно выпуклая по y. Под L-гладкостью по y понимается свойство

$$\|\nabla_y F(x, y) - \nabla_y F(x, y')\|_2 \leqslant L\|y - y'\|_2 \quad \forall x \in Q_x,\ y, y' \in Q_y.$$

Такая постановка возникает, например, при поиске равновесий в транспортных сетях [9]. В машинном обучении задачи такого типа соответствуют случаю, когда регуляризация применяется к одной из двух групп параметров модели (отсюда сильная выпуклость только по одной группе переменных из двух). Например, когда в датасете большая группа признаков являются разреженными, то регуляризация может использоваться только для весов модели, соответствующих этим признакам. В качестве еще одного примера можно привести логистическую регрессию, в которой есть априорное распределение на часть параметров. Задаче min-min посвящено несколько публикаций, среди которых [10–12]. Например, в [10] авторы предложили новые алгоритмы для задач min-max, шаги которых настраиваются автоматически, но предложенные методы также применяются и к задачам min-min.

В данной статье рассматриваются два подхода к решению задачи (1), имеющие линейную скорость сходимости. Предлагается свести рассматриваемую задачу к совокупности вспомогательных задач (внутренней и внешней). Внешняя задача (минимизация по x) решается методом Вайды (метод секущей плоскости) [13, 14].

В случае когда целевая функция F простая, т.е. не является суммой большого количества функций, внутренняя задача (минимизация по y) решается быстрым градиентным методом для задач сильно выпуклой оптимизации. В результате такого подхода приближенное решение задачи (1) может быть достигнуто за $\widetilde{\mathcal{O}}(d)$ вычислений $\partial_x F$ и $\widetilde{\mathcal{O}}\left(d\sqrt{\frac{L}{\mu}}\right)$ вычислений $\nabla_y F$, см. тео-

рему 5. Здесь и далее $\widetilde{\mathcal{O}}(\cdot) = \mathcal{O}(\cdot)$ с точностью до небольшой степени логарифмического множителя, обычно эта степень равна единице или двум.

Оптимизация суммы большого количества функций в течение последних нескольких лет является предметом интенсивных исследований из-за широкого спектра приложений в машинном обучении, статистике, обработке изображений и других математических и инженерных приложениях. Поэтому отдельно рассматривается случай, когда целевая функция F представляет собой сумму (или среднее арифметическое) большого числа m функций, в котором использование быстрого градиентного метода для задач сильно выпуклой оптимизации потребовало бы вычисления градиентов m слагаемых на каждом шаге, что может занимать много времени. Вместо этого предлагаем использовать ускоренный градиентный метод с редукцией дисперсии [2, 15], который также имеет линейную сходимость. В результате такого подхода решение задачи может быть достигнуто за $\widetilde{\mathcal{O}}\left(md\right)$ вычислений $\partial_x F$ и за $\widetilde{\mathcal{O}}\left(md + d\sqrt{\frac{mL}{\mu}}\right)$ вычислений $\nabla_y F$, см. теорему 6.

Используя два предложенных подхода, получаем линейную скорость сходимости для задачи min-min (1). Отметим, что гладкость и сильная выпуклость требуются только по одной из двух групп переменных.

Статья состоит из 5 разделов и Приложения. В разделе 2 приводятся используемые алгоритмы и их сложность, а именно: быстрый градиентный метод, метод Вайды (метод секущей плоскости) и метод ускоренного градиентного спуска с редукцией дисперсии. В разделе 3 формулируется постановка задачи и приводятся подходы к рассматриваемой задаче для различных случаев целевой функции, в одном из которых целевая функция является суммой или средним арифметическим большого числа функций. В разделе 4 приводятся результаты вычислительных экспериментов и сравнение скорости работы предложенных подходов. Отметим, что полные доказательства теорем 4, 5, 6 и вспомогательного утверждения 1 приводятся в Приложении.

2. Используемые алгоритмы

Приведем алгоритмы, используемые в предлагаемых в статье подходах к решению задачи (1). Сначала приводится быстрый градиентный метод, затем метод Вайды (метод секущей плоскости) и, наконец, ускоренный градиентный метод с редукцией дисперсии.

2.1. Быстрый градиентный метод

В [16] предложен адаптивный алгоритм для решения задачи оптимизации

$$f(y) \to \min_{y \in Q_y}, \tag{2}$$

где $Q_y \subset \mathbb{R}^n$ — непустое компактное выпуклое множество, f — L-гладкая выпуклая функция. Этот алгоритм, получивший название быстрого градиентного метода, позволяет ускорить сходимость обычного градиентного спуска

с $\mathcal{O}\left(\frac{1}{N}\right)$ до $\mathcal{O}\left(\frac{1}{N^2}\right)$, где N — количество итерации алгоритма. Быстрый градиентный метод (не адаптивный вариант) приведен далее как алгоритм 1.

Алгоритм 1. Быстрый градиентный метод [16].

Вход: Количество шагов N, начальная точка $y^0 \in Q_y$, параметр $L > 0$.

1: 0-шаг: $z^0 := y^0, \quad u^0 := y^0, \quad \alpha_0 := 0, \quad A_0 := 0.$

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

3: Находим наибольший корень α_{k+1} такой, что $A_k + \alpha_{k+1} = L\alpha_{k+1}^2$,

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

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

6: $$u^{k+1} := \arg\min_{y \in Q_y} \left\{ \alpha_{k+1} \left\langle \nabla f\left(z^{k+1}\right), y - z^{k+1} \right\rangle + \frac{1}{2}\|y - u^k\|_2^2 \right\},$$

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

8: end for

Выход: y^N.

Следующая теорема дает оценку сложности (скорости сходимости) алгоритма 1.

Теорема 1 [16]. *Пусть функция $f : Q_y \to \mathbb{R}$ является L-гладкой и выпуклой, тогда алгоритм 1 возвращает такую точку y^N, что*

$$f\left(y^N\right) - f(y_*) \leqslant \frac{8LR^2}{(N+1)^2},$$

где y_* — решение задачи (2), $R^2 = \frac{1}{2}\|y^0 - y_*\|_2^2$.

Опишем далее технику рестартов (перезапусков) быстрого градиентного метода (алгоритм 1) для случая μ-сильно выпуклой функции.

Ввиду μ-сильной выпуклости f имеем

$$\frac{\mu}{2}\|z - y\|_2^2 \leqslant f(z) - (f(y) + \langle \nabla f(y), z - y\rangle) \leqslant \frac{L}{2}\|z - y\|_2^2 \quad \forall y, z \in Q_y.$$

Тогда после N_1 итераций алгоритма 1 с учетом теоремы 1 получаем

$$\frac{\mu}{2}\|y^{N_1} - y_*\|_2^2 \leqslant f\left(y^{N_1}\right) - f\left(y_*\right) \leqslant \frac{4L\|y^0 - y_*\|_2^2}{N_1^2}, \tag{3}$$

отсюда

$$\|y^{N_1} - y_*\|_2^2 \leqslant \frac{8L}{\mu N_1^2}\|y^0 - y_*\|_2^2.$$

Поэтому, выбирая $N_1 = \left\lceil 4\sqrt{\frac{L}{\mu}} \right\rceil$, где $\lceil \cdot \rceil$ — округление вверх, получим

$$\|y^{N_1} - y_*\|_2^2 \leqslant \frac{1}{2}\|y^0 - y_*\|_2^2.$$

После этого выберем для алгоритма 1 в качестве точки старта y^{N_1}, снова сделаем N_1 итераций и т.д. Для достижения приемлемого качества решения можно выбрать количество рестартов алгоритма 1 (параметр p алгоритма 2) следующим образом:

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

В таком случае общее число итераций алгоритма 2 будет

$$N = \left\lceil \frac{1}{2} \ln \left(\frac{\mu R^2}{\varepsilon} \right) \right\rceil \cdot \left\lceil 4\sqrt{\frac{L}{\mu}} \right\rceil,$$

т.е.

$$N = \mathcal{O}\left(\sqrt{\frac{L}{\mu}} \ln \left(\frac{\mu R^2}{\varepsilon} \right) \right) = \widetilde{\mathcal{O}}\left(\sqrt{\frac{L}{\mu}} \right). \tag{4}$$

Алгоритм 2. Быстрый градиентный метод для задач сильно выпуклой оптимизации, рестарты алгоритма 1.

Вход: начальная точка $y^0 \in Q_y$, $L > 0$, число рестартов $p = \left\lceil \frac{1}{2} \ln \left(\frac{\mu R^2}{\varepsilon} \right) \right\rceil$.

1: for $j = 1, \ldots, p$ do
2: Выполнить $N_j = \left\lceil 4\sqrt{\frac{L}{\mu}} \right\rceil$ итераций алгоритма 1,
3: $y^0 := y^{N_j}$.
4: end for

Выход: $\hat{y} := y^{N_p}$.

2.2. Метод Вайды

Метод Вайды (метод секущей плоскости) был предложен Вайдой в [13, 14] для решения условной задачи оптимизации

$$f(x) \to \min_{x \in Q_x}, \tag{5}$$

где $Q_x \subset \mathbb{R}^d$ — выпуклое компактное множество с непустой внутренностью, а целевая функция f, определенная на Q_x, непрерывна и выпукла.

Пусть $P = \{x \in \mathbb{R}^d : Ax \geqslant b\}$ — ограниченный d-мерный многогранник, где $A \in \mathbb{R}^{m\times d}$ и $b \in \mathbb{R}^m$. Логарифмический барьер множества P определяется как

$$Barr(x) = -\sum_{i=1}^{m} \log\left(a_i^\top x - b_i\right),$$

где $a_i^\top$ — i-я строка матрицы A. Гессиан $H(x)$ функции $Barr(x)$ равен

$$H(x) = \sum_{i=1}^{m} \frac{a_i a_i^\top}{\left(a_i^\top x - b_i\right)^2}.$$

Матрица $H(x)$ положительно определена для всех x из внутренности P. Волюметрический барьер (volumetric barrier) $\mathcal{V}$ определяется как

$$\mathcal{V}(x) = \frac{1}{2}\log\left(\det\left(H(x)\right)\right),$$

где $\det\left(H(x)\right)$ обозначает детерминант $H(x)$. Будем называть точку минимума функции $\mathcal{V}$ на P волюметрическим центром множества P.

Обозначим

$$\sigma_i(x) = \frac{a_i^\top \left(H(x)\right)^{-1} a_i}{\left(a_i^\top x - b_i\right)^2}, \quad 1 \leqslant i \leqslant m, \tag{6}$$

тогда градиент волюметрического барьера $\mathcal{V}$ может быть записан как

$$\nabla\mathcal{V}(x) = -\sum_{i=1}^{m} \sigma_i(x)\frac{a_i}{a_i{}^\top x - b_i}.$$

Пусть $\mathcal{Q}(x)$ определяется как

$$\mathcal{Q}(x) = \sum_{i=1}^{m} \sigma_i(x)\frac{a_i a_i^\top}{\left(a_i^\top x - b_i\right)^2}.$$

Заметим, что $\mathcal{Q}(x)$ положительно определена на внутренности P, а также $\mathcal{Q}(x)$ является хорошим приближением гессиана функции $\mathcal{V}(x)$, т.е. $\nabla^2\mathcal{V}(x)$.

Метод Вайды производит последовательность пар $(A_k, b_k) \in \mathbb{R}^{m\times d} \times \mathbb{R}^m$ таких, что соответствующие многогранники содержат решение. В качестве начального многогранника, задаваемого парой (A_0, b_0), обычно берется симплекс (алгоритм может начинать с любого выпуклого ограниченного n-мерного многогранника, для которого легко вычислить волюметрический центр — например, с n-прямоугольника).

Параметром алгоритма является небольшое число $\gamma \leqslant 0{,}006$, смысл которого более подробно раскрывается в книге [17]. Пусть x_k $(k \geqslant 0)$ обозначает волюметрический центр многогранника, заданного парой (A_k, b_k), и пусть для него вычислены величины $\{\sigma_i(x_k)\}_{1\leqslant i\leqslant m}$ (см. (6)). Следующий многогранник (A_{k+1}, b_{k+1}) получается из текущего в результате либо присоединения, либо удаления ограничения:

1) Если для некоторого $i \in \{1, \ldots, m\}$ выполняется $\sigma_i(x_k) = \min\limits_{1 \leqslant j \leqslant m} \sigma_j(x_k) < < \gamma$, тогда (A_{k+1}, b_{k+1}) получается исключением i-й строки из (A_k, b_k);
2) иначе $\left(\text{если } \min\limits_{1 \leqslant j \leqslant m} \sigma_j(x_k) \geqslant \gamma\right)$ оракул, вызванный в текущей точке x_k, возвращает вектор c_k такой, что $f(x) \leqslant f(x_k)\ \forall x \in \left\{z \in Q_x : : c_k^\top z \geqslant c_k^\top x_k\right\}$, т.е. $c_k \in -\partial f(x_k)$. Выберем $\beta_k \in \mathbb{R}$ таким, что

$$\frac{c_k^\top \left(H(x_k)\right)^{-1} c_k}{\left(x_k^\top c_k - \beta_k\right)^2} = \frac{1}{5}\sqrt{\gamma}.$$

Определим (A_{k+1}, b_{k+1}) добавлением строки (c_k, β_k) к (A_k, b_k).
Волюметрический барьер $\mathcal{V}_k$ является самосогласованной функцией, поэтому может быть эффективно минимизирован методом Ньютона. Достаточно одного шага метода Ньютона для $\mathcal{V}_k$, сделанного из x_{k-1}. Подробности и анализ метода Вайды можно найти в [13, 14, 17].

Следующая теорема дает оценку сложности алгоритма Вайды.

Теорема 2. Пусть $\mathcal{B}_\rho$ и $\mathcal{B}_\mathcal{R}$ — некоторые евклидовы шары радиусов ρ и $\mathcal{R}$ соответственно такие, что $\mathcal{B}_\rho \subseteq Q_x \subseteq \mathcal{B}_\mathcal{R}$, и пусть число $B > 0$ таково, что $|f(x) - f(x')| \leqslant B\ \forall x, x' \in Q_x$. Тогда метод Вайды находит ε-решение задачи (5) за $\mathcal{O}\left(d \log \frac{dB\mathcal{R}}{\rho\varepsilon}\right)$ шагов.

Замечание 1. Как показано в [18], метод Вайды можно использовать с неточным субградиентом без накопления ошибки.

Замечание 2. Помимо вычисления субградиента, в стоимость итерации метода Вайды входит стоимость обращения матрицы размера $d \times d$ и решения системы линейных уравнений.

2.3. Ускоренный градиентный метод с редукцией дисперсии

Рассмотрим задачу

$$f(y) \to \min_{y \in Q_y}, \tag{7}$$

где $Q_y \subseteq \mathbb{R}^n$ — замкнутое выпуклое множество, а целевая функция f представляет собой сумму (или среднее арифметическое) большого числа m гладких выпуклых функций f_i, т.е. $f(y) = \frac{1}{m}\sum_{i=1}^m f_i(y)$. При решении (7) с помощью быстрого градиентного метода для задач сильно выпуклой оптимизации (алгоритм 2) потребуется вычислять градиент m функций на каждой итерации, что очень дорого. Поэтому предпочтительнее вместо алгоритма 2 использовать рандомизированный градиентный метод, а именно ускоренный градиентный метод с редукцией дисперсии, также называемый Varag [2, 15]. Приведенный далее алгоритм 3 представляет собой ускоренный градиентный метод с редукцией дисперсии (Varag) для гладкой сильно выпуклой задачи оптимизации конечной суммы (7). Этот алгоритм был предложен Г. Ланом и др. в [15].

Предположим, что для каждого $i \in \{1, \ldots, m\}$, существует $L_i > 0$ такое, что

$$\|\nabla f_i(y) - \nabla f_i(z)\|_2 \leqslant L_i \|y - z\|_2 \quad \forall y, z \in Q_y.$$

Ясно, что f имеет липшицев градиент с константой не более $L := \frac{1}{m}\sum_{i=1}^m L_i$. Предположим также, что целевая функция f сильно выпуклая с константой $\mu > 0$, т.е.

$$f(z) \geqslant f(y) + \langle \nabla f(y), z - y\rangle + \frac{\mu}{2}\|y - z\|_2 \quad \forall y, z \in Q_y.$$

Определение 1. Случайный вектор $\bar{y}$, принимающий значения из Q_y, называется стохастическим ε-решением задачи (7)*, если $\mathbb{E}[f(\bar{y}) - f(y_*)] \leqslant \leqslant \varepsilon$, где y_* — точное решение задачи* (7)*.*

Алгоритм Varag содержит вложенные циклы — внешний и внутренний (индексируемые переменными s и t соответственно). На каждой итерации внешнего цикла вычисляется полный градиент $\nabla f(\tilde{y})$ в точке $\tilde{y}$, который затем используется во внутреннем цикле для определения оценок градиента G_t. Каждая итерация внутреннего цикла требует информацию о градиенте только одного случайно выбранного слагаемого f_{i_t} и содержит три основные последовательности: $\{\underline{y}_t\}, \{y_t\}$ и $\{\bar{y}_t\}$.

Обозначим $s_0 := \lfloor \log_2 m \rfloor + 1$, где $\lfloor \cdot \rfloor$ — округление вниз. Параметры алгоритма 3 $\{q_1, \ldots, q_m\}$, $\{\theta_t\}$, $\{\alpha_s\}$, $\{\gamma_s\}$, $\{p_s\}$ и $\{T_s\}$ описываются следующим образом:

- Вероятности $q_i = \frac{1}{\sum_{i=1}^m L_i} L_i \ \forall i \in \{1, \ldots, m\}$;
- Веса $\{\theta_t\}$ при $1 \leqslant s \leqslant s_0$ или $s_0 < s \leqslant s_0 + \sqrt{\frac{12L}{m\mu}} - 4$, $m < \frac{3L}{4\mu}$ равны

$$\theta_t = \begin{cases} \dfrac{\gamma_s}{\alpha_s}\left(\alpha_s + p_s\right), & 1 \leqslant t \leqslant T_s - 1, \\ \dfrac{\gamma_s}{\alpha_s}, & t = T_s. \end{cases} \tag{8}$$

  В остальных случаях они равны

$$\theta_t = \begin{cases} \Gamma_{t-1} - (1 - \alpha_s - p_s)\,\Gamma_t, & 1 \leqslant t \leqslant T_s - 1, \\ \Gamma_{t-1}, & t = T_s, \end{cases} \tag{9}$$

  где $\Gamma_t = (1 + \mu\gamma_s)^t$;
- Параметры $\{T_s\}, \{\gamma_s\}$ и $\{p_s\}$ определяются как

$$T_s = \begin{cases} 2^{s-1}, & s \leqslant s_0, \\ T_{s_0}, & s > s_0, \end{cases} \qquad \gamma_s = \frac{1}{3L\alpha_s}, \quad p_s = \frac{1}{2}; \tag{10}$$

- Наконец,

$$\alpha_s = \begin{cases} \dfrac{1}{2}, & s \leqslant s_0, \\ \max\left\{\dfrac{2}{s - s_0 + 4}, \min\left\{\sqrt{\dfrac{m\mu}{3L}}, \dfrac{1}{2}\right\}\right\}, & s > s_0. \end{cases} \tag{11}$$

Алгоритм 3. Ускоренный градиентный метод с редукцией дисперсии (Varag) [15].

Вход: $y^0 \in Q_y, \{T_s\}$, $\{\gamma_s\}$, $\{\alpha_s\}$, $\{p_s\}$, $\{\theta_t\}$ и распределение вероятностей $\{q_1, \ldots, q_m\}$ на $\{1, \ldots, m\}$.

$\tilde{y}^0 := y^0$.
for $s = 1, 2, \ldots,$ do
  $\tilde{y} := \tilde{y}^{s-1}$, $\tilde{g} := \nabla f(\tilde{y})$.
  $y_0 := y^{s-1}$, $\bar{y}_0 = \tilde{y}$, $T := T_s$.
  for $t = 1, 2, \ldots, T$ do
    Выбрать $i_t \in \{1, \ldots, m\}$ случайным образом согласно $\{q_1, \ldots, q_m\}$.
    $\underline{y}_t := \dfrac{1}{(1+\mu\gamma_s(1-\alpha_s))}[(1+\mu\gamma_s)(1-\alpha_s-p_s)\bar{y}_{t-1} + \alpha_s y_{t-1} + (1+\mu\gamma_s)p_s\tilde{y}]$.
    $G_t := \dfrac{1}{(q_{i_t} m)} \left(\nabla f_{i_t}\left(\underline{y}_t\right) - \nabla f_{i_t}(\tilde{y})\right) + \tilde{g}$.
    $y_t := \arg\min\limits_{y \in Q_y} \left\{\gamma_s \left(\langle G_t, y\rangle + \dfrac{\mu}{2}\|\underline{y}_t - y\|_2^2\right) + \dfrac{1}{2}\|y_{t-1} - y\|_2^2\right\}$.
    $\bar{y}_t := (1 - \alpha_s - p_s)\,\bar{y}_{t-1} + \alpha_s y_t + p_s\tilde{y}$.
  end for
  $y^s := y_T, \tilde{y}^s := \dfrac{1}{\sum_{t=1}^T \theta_t} \sum\limits_{t=1}^T (\theta_t \bar{y}_t)$.
end for

Следующий результат дает оценку сложности алгоритма 3.

Теорема 3 [15]. *Если параметры алгоритма 3* $\{\theta_t\}$, $\{\alpha_s\}$, $\{\gamma_s\}$, $\{p_s\}$ *и* $\{T_s\}$ *заданы согласно формулам* (8), (9), (10) *и* (11), *то общее количество вычислений градиентов функций* f_i, *выполняемых алгоритмом 3 для нахождения стохастического* ε*-решения задачи* (7), *ограничено*

$$N := \begin{cases} \mathcal{O}\left\{m \log \dfrac{D_0}{\varepsilon}\right\}, & m \geqslant \dfrac{D_0}{\varepsilon} \ \text{ или } \ m \geqslant \dfrac{3L}{4\mu}, \\ \mathcal{O}\left\{m \log m + \sqrt{\dfrac{mD_0}{\varepsilon}}\right\}, & m < \dfrac{D_0}{\varepsilon} \leqslant \dfrac{3L}{4\mu}, \\ \mathcal{O}\left\{m \log m + \sqrt{\dfrac{mL}{\mu}} \log \dfrac{D_0/\varepsilon}{3L/4\mu}\right\}, & m < \dfrac{3L}{4\mu} \leqslant \dfrac{D_0}{\varepsilon}, \end{cases} \tag{12}$$

где $D_0 = 2\left(f(y^0) - f(y_*)\right) + \frac{3L}{2}\|y^0 - y_*\|_2^2$, *где* y_* *— решение задачи* (7).

Заметим, что оценку (12) можно записать как $N = \widetilde{\mathcal{O}}\left(m + \sqrt{\frac{mL}{\mu}}\right)$, где $\widetilde{\mathcal{O}}(\cdot) = \mathcal{O}(\cdot)$ с точностью до логарифмического множителя по m, L, μ, ε и D_0.

3. Постановка задачи и полученные результаты

Рассмотрим задачу

$$\min_{x\in Q_x}\min_{y\in Q_y} F(x,y), \tag{13}$$

где $Q_x \subset \mathbb{R}^d$, $Q_y \subset \mathbb{R}^n$ — непустые компактные выпуклые множества, размерность d относительно небольшая ($d \ll n$), функция $F(x,y)$ — выпуклая по совокупности переменных, а также L-гладкая и μ-сильно выпуклая по y. Под L-гладкостью по y понимается свойство

$$\|\nabla_y F(x,y) - \nabla_y F(x,y')\|_2 \leqslant L\|y-y'\|_2 \quad \forall x \in Q_x,\ y,y' \in Q_y.$$

Введем функцию

$$f(x) = \min_{y\in Q_y} F(x,y). \tag{14}$$

Задачу (13) можно переписать в виде

$$f(x) \to \min_{x\in Q_x}. \tag{15}$$

При решении (15) некоторым итерационным методом необходимо на каждом его шаге решать вспомогательную задачу (14), чтобы приближенно находить субградиент $\partial f(x)$. Обратимся к следующему определению.

Определение 2 ([19], с. 123). *Пусть $\delta \geqslant 0$, $Q_x \subseteq \mathbb{R}^d$ — выпуклое множество, $f: Q_x \to \mathbb{R}$ — выпуклая функция. Вектор $g \in \mathbb{R}^d$ называется δ-субградиентом f в точке $x' \in Q_x$, если*

$$f(x) \geqslant f(x') + \langle g, x-x'\rangle - \delta \quad \forall x \in Q_x.$$

Множество δ-субградиентов f в точке x' обозначается $\partial_\delta f(x')$.

Обозначим $D := \max_{y,z\in Q_y}\|y-z\|_2$, $y(x) := \arg\min_{y\in Q_y} F(x,y)$. Следующая теорема говорит о том, как вычислить δ-субградиент функции $f(x)$, приближенно решая вспомогательную задачу (15).

Теорема 4. Пусть найден такой $\tilde{y} \in Q_x$, что $F(x,\tilde{y}) - f(x) \leqslant \varepsilon$, тогда

$$\partial_x F(x,\tilde{y}) \in \partial_\delta f(x), \quad \delta = \left(LD + \|\nabla_y F\left(x, y(x)\right)\|_2\right)\sqrt{\frac{2\varepsilon}{\mu}}.$$

Эта теорема непосредственно следует из двух утверждений.

Утверждение 1. Пусть $g: Q_y \to \mathbb{R}$ — L-гладкая μ-сильно выпуклая функция, точка $\tilde{y} \in Q_y$ такова, что $g(\tilde{y}) - g(y_*) \leqslant \varepsilon$, тогда

$$\max_{y\in Q_y}\langle \nabla g(\tilde{y}), \tilde{y}-y\rangle \leqslant \delta, \quad \delta = \left(LD + \|\nabla g\left(y_*\right)\|_2\right)\sqrt{\frac{2\varepsilon}{\mu}},$$

где $y_* = \arg\min_{y\in Q_y} g(y)$.

Утверждение 2 ([20], с. 12). *Пусть найден такой* $\tilde{y} \in Q_y$, *что*

$$\max_{y \in Q_y} \langle \nabla_y F(x, \tilde{y}), \tilde{y} - y \rangle \leqslant \delta,$$

тогда $\partial_x F(x, \tilde{y}) \in \partial_\delta f(x)$.

Интуитивно теорема 4 говорит о том, что, решив вспомогательную задачу (14) достаточно точно, получим хорошее приближение субградиента $\partial f(x)$, которое может быть использовано для решения внешней задачи (15). На этой идее основан предлагаемый подход к решению (13).

Подход 1 (основной случай). Внешняя задача (15) решается методом Вайды. Вспомогательная задача (14) решается быстрым градиентным методом для задач сильно выпуклой оптимизации (алгоритм 2).

Теорема 5. Подход 1 позволяет получить ε*-решение задачи* (13) *после* $\widetilde{\mathcal{O}}(d)$ *вычислений* $\partial_x F$ *и обращений матриц размера* $d \times d$, *а также* $\widetilde{\mathcal{O}}\left(d\sqrt{\frac{L}{\mu}}\right)$ *вычислений* $\nabla_y F$.

Замечание 3. Обращение матриц появляется в сложности предлагаемого подхода из-за того, что оно производится на каждом шаге метода Вайды.

3.1. Минимизация суммы большого числа функций

Пусть в задаче (13)

$$F(x, y) = \frac{1}{m} \sum_{i=1}^{m} F_i(x, y), \tag{16}$$

где функции F_i являются выпуклыми по совокупности переменных и L_i-гладкими по y, а F является μ-сильно выпуклой по y. Из этого следует, что F является выпуклой по совокупности переменных и гладкой по y с константой гладкости не более $L := \frac{1}{m} \sum_{i=1}^{m} L_i$.

Подход 2 (сумма функций). Внешняя задача (15) решается методом Вайды. Вспомогательная задача (14) решается ускоренным градиентным методом с редукцией дисперсии (алгоритм 3).

Теорема 6. Подход 2 позволяет получить ε*-решение задачи* (13) *за* $\widetilde{\mathcal{O}}(md)$ *вычислений* $\partial_x F_i$, $\widetilde{\mathcal{O}}(d)$ *обращений матриц размера* $d \times d$ *и* $\widetilde{\mathcal{O}}\left(dm + d\sqrt{\frac{mL}{\mu}}\right)$ *вычислений* $\nabla_y F_i$.

4. Эксперименты

Рассмотрим модель логистической регрессии для задачи бинарной классификации. Ошибка модели с параметрами w на обучающем объекте с вектором признаков z, принадлежащем классу $t \in \{-1, 1\}$, записывается как

$$\ell_z(w) = \log\left(1 + e^{-t\langle w, z \rangle}\right).$$

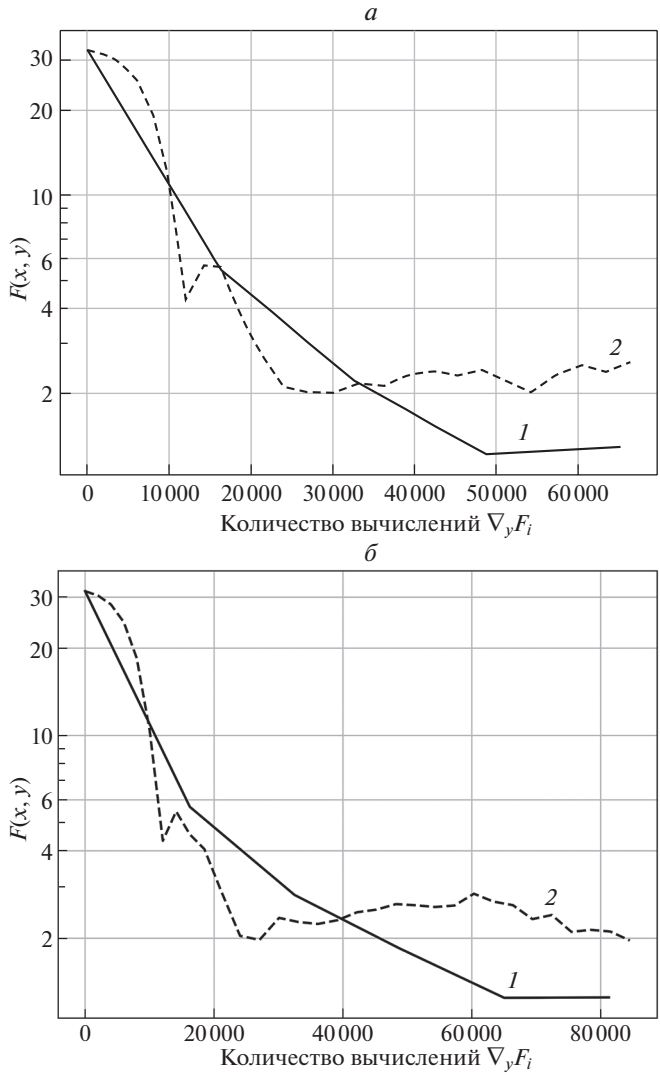

Рис. 1. *а* и *б* соответствуют размерностям $d = 20$ и $d = 30$ соответственно. Графики *1* и *2* показывают сходимость предлагаемого подхода и метода Varag соответственно.

Пусть параметры модели состоят из двух групп: $w = (x, y),\ x \in \mathbb{R}^d,\ y \in \mathbb{R}^n$, причем для группы y задано гауссовское априорное распределение:

$$y \sim \mathcal{N}\left(0, \sigma^2 I_n\right),$$

где I_n — единичная матрица размера n. Максимизация апостериорной вероятности приведет (см. [21], § 4.5.1) к задаче

$$\min_{x \in Q_x} \min_{y \in Q_y} \left\{ F(x,y) := \frac{1}{m} \sum_{i=1}^{m} \ell_{z_i}(x,y) + \frac{1}{\sigma^2} \|y\|_2^2 \right\}, \tag{17}$$

где в качестве Q_x и Q_y можно взять евклидовы шары достаточно большого радиуса.

Будем решать задачу (17) при помощи подхода 2 и сравним его работу с работой метода Varag (алгоритм 3). Заметим, что эта задача не является сильно выпуклой по совокупности переменных. Для такой постановки можно использовать Varag, задавая параметры θ_t по формуле (8), а все остальные параметры по формулам для сильно выпуклого случая, положив $\mu = 0$, см. [15]. При этом стохастическое ε-решение будет найдено за $\mathcal{O}\left(\sqrt{\frac{mD_0}{\varepsilon}} + m\log m\right)$ вычислений градиентов функций F_i, где $D_0 = 2\left(F(x^0, y^0) - F(x_*, y_*)\right) + \frac{3L}{2}\|(x^0, y^0) - (x_*, y_*)\|_2^2$, (x_*, y_*) — решение задачи (17). Эта сублинейная оценка уступает предлагаемому в статье подходу, см. теорему 2.

Для экспериментов использовался датасет madelon, представленный 2000 объектов, имеющих 500 признаков. Был выбран небольшой коэффициент регуляризации $\frac{1}{\sigma^2} = 0{,}005$ и проведены эксперименты для двух размерностей d, равных 20 и 30.

На рис. 1 отражены результаты эксперимента. По оси x откладывается количество вычислений градиентов $\nabla_y F_i$, которое для Varag совпадает с количеством вычислений $\nabla_x F_i$. Отметим, что предложенный подход требует меньше вычислений $\nabla_x F_i$, поскольку они выполняются только во внешнем цикле. Так, график *1* на рис. 1,*а* соответствует четырем итерациям внешнего цикла (т.е. 8000 вычислений $\nabla_x F_i$), а график *1* на рис. 1,*б* — пяти итерациям (т.е. 10 000 вычислений $\nabla_x F_i$). В данном эксперименте подход 2 позволил достичь меньших значений целевой функции.

Исходный код и результаты экспериментов могут быть найдены в репозитории https://github.com/egorgladin/min_min.

5. Заключение

В статье рассмотрена задача вида min-min:

$$\min_{x\in Q_x}\min_{y\in Q_y} F(x, y), \tag{18}$$

где $Q_x \subset \mathbb{R}^d$, $Q_y \subset \mathbb{R}^n$ — непустые компактные выпуклые множества, размерность d относительно небольшая ($d \ll n$), функция $F(x, y)$ — выпуклая по совокупности переменных, а также L-гладкая и μ-сильно выпуклая по y.

Предложено два подхода к решению задачи (18), в которых она сводится к совокупности вспомогательных задач (внутренней и внешней). Внешняя задача (минимизация по x) решается методом Вайды, а внутренняя (минимизация по y) — быстрым градиентным методом для задач сильно выпуклой оптимизации или, если минимизируется сумма большого количества функций, ускоренным градиентным методом с редукцией дисперсии. Это позволяет достигать приближенного решения задачи (18) за $\widetilde{\mathcal{O}}(d)$ вычислений $\partial_x F$ и $\widetilde{\mathcal{O}}\left(d\sqrt{\frac{L}{\mu}}\right)$ вычислений $\nabla_y F$, см. теорему 5. Для сравнения, если бы задача (18) была гладкой по совокупности переменных, то ее решение при

использовании только быстрого градиентного метода имело бы сложность $\mathcal{O}\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)$, где R — расстояние от начального приближения до решения. В случае суммы с m слагаемыми решение задачи может быть достигнуто за $\widetilde{\mathcal{O}}\left(md\right)$ вычислений $\partial_x F$ и за $\widetilde{\mathcal{O}}\left(md + d\sqrt{\frac{mL}{\mu}}\right)$ вычислений $\nabla_y F$, см. теорему 6.

Проведен численный эксперимент, в котором один из предлагаемых подходов применен к задаче логистической регрессии с регуляризацией, применяемой к одной из двух групп параметров модели. По сравнению с алгоритмом Varag, предложенный подход достиг меньших значений функции при меньшем числе вызовов оракулов.

Отметим также, что если функция $F(x, y)$ — μ-сильно выпуклая по совокупности переменных, то функция $g(y) = \min_{x \in Q_x} F(x, y)$ также будет μ-сильно выпуклая. Более того, все это можно сформулировать в терминах (δ, μ, L)-оракула (см. [3] и цитированную там литературу). При $\mu = 0$ это сделано в [20], при $\mu > 0$ доказательство практически дословно повторяет утверждения 1 и 3 из [20] (см. также [9]). Приведенное наблюдение позволяет обоснованно (с теоретической проработкой) использовать для решения внутренней задачи метод Вайды, а для решения внешней задачи использовать, например, быстрый градиентный метод. Однако такой подход будет предпочтительнее рассмотренного в данной статье только при весьма специальных (как правило, трудно выполнимых) условиях [5].

ПРИЛОЖЕНИЕ

Доказательство утверждения 1. Рассмотрим произвольный $y \in Q_y$

$$\langle \nabla g(\tilde{y}), \tilde{y} - y \rangle = \langle \nabla g(\tilde{y}) - \nabla g(y_*), \tilde{y} - y \rangle + \langle \nabla g(y_*), \tilde{y} - y \rangle . \tag{П.1}$$

Оценим сверху первое слагаемое, используя неравенство Коши–Буняковского и определение липшицевости градиента:

$$\begin{aligned} \langle \nabla g(\tilde{y}) - \nabla g(y_*), \tilde{y} - y \rangle &\leqslant \|\nabla g(\tilde{y}) - \nabla g(y_*)\|_2 \, \|\tilde{y} - y\|_2 \leqslant \\ &\leqslant L \, \|\tilde{y} - y_*\|_2 \, \|\tilde{y} - y\|_2 . \end{aligned} \tag{П.2}$$

Из сильной выпуклости следует, что

$$g(\tilde{y}) \geqslant g(y_*) + \langle \nabla g(y_*), \tilde{y} - y_* \rangle + \frac{\mu}{2} \|\tilde{y} - y_*\|_2^2.$$

Воспользовавшись неравенствами $g(\tilde{y}) - g(y_*) \leqslant \varepsilon$ и $\langle \nabla g(y_*), y - y_* \rangle \geqslant 0 \; \forall y \in Q_y$, получим

$$\|\tilde{y} - y_*\|_2 \leqslant \sqrt{\frac{2\varepsilon}{\mu}} \overset{\text{(П.2)}}{\Longrightarrow} \langle \nabla g(\tilde{y}) - \nabla g(y_*), \tilde{y} - y \rangle \leqslant L \, \|\tilde{y} - y\|_2 \sqrt{\frac{2\varepsilon}{\mu}}. \tag{П.3}$$

Теперь оценим сверху второе слагаемое в (П.1)

$$\langle \nabla g(y_*), \tilde{y} - y \rangle = \langle \nabla g(y_*), \tilde{y} - y_* \rangle + \langle \nabla g(y_*), y_* - y \rangle .$$

Снова воспользовавшись критерием оптимальности точки y_* и неравенством Коши–Буняковского, получим

$$\langle \nabla g(y_*), \tilde{y} - y \rangle \leqslant \|\nabla g(y_*)\|_2 \|\tilde{y} - y_*\|_2 \overset{(\text{П.3})}{\leqslant} \|\nabla g(y_*)\|_2 \sqrt{\frac{2\varepsilon}{\mu}}.$$

Объединив верхние оценки для обоих слагаемых, получим

$$\langle \nabla g(\tilde{y}), \tilde{y} - y \rangle \leqslant \left(L \|\tilde{y} - y\|_2 + \|\nabla g(y_*)\|_2\right) \sqrt{\frac{2\varepsilon}{\mu}},$$

откуда следует доказываемое утверждение 1.

Доказательство теоремы 4. Зафиксировав $x \in Q_x$, применим утверждение 1 к функции $g(y) := F(x, y)$ и утверждение 2. Теорема 4 доказана.

Доказательство теоремы 5. Согласно (4) алгоритм 2 сходится линейно, поэтому можно считать, что вспомогательная задача $\min_{y \in Q_y} F(x, y)$ решается сколь угодно точно за время $\widetilde{\mathcal{O}}\left(\sqrt{\frac{L}{\mu}}\right)$. Согласно теореме 4 это позволяет использовать δ-субградиент, где δ убывает со скоростью геометрической прогрессии. Для внешней задачи используется метод Вайды, который также сходится линейно и имеет сложность $\widetilde{\mathcal{O}}(d)$. Таким образом, для решения задачи (13) достаточно $\widetilde{\mathcal{O}}(d)$ вычислений $\partial_x F$ и обращений матриц размера $d \times d$, а также $\widetilde{\mathcal{O}}\left(d\sqrt{\frac{L}{\mu}}\right)$ вычислений $\nabla_y F$. Теорема 5 доказана.

Доказательство теоремы 6. Согласно теореме 3 Varag сходится линейно, поэтому можно считать, что вспомогательная задача $\min_{y \in Q_y} F(x, y)$ решается сколь угодно точно за время $\widetilde{\mathcal{O}}\left(m + \sqrt{\frac{mL}{\mu}}\right)$. Согласно теореме 4 это позволяет использовать δ-субградиент, где δ убывает со скоростью геометрической прогрессии. Для внешней задачи используется метод Вайды, который также сходится линейно и имеет сложность $\widetilde{\mathcal{O}}(d)$ итераций. На каждой его итерации необходимо вычислять субградиенты всех m слагаемых $\partial_x F_i$. Таким образом, для решения задачи достаточно $\widetilde{\mathcal{O}}(md)$ вычислений $\partial_x F_i$, $\widetilde{\mathcal{O}}(d)$ обращений матриц размера $d \times d$ и $\widetilde{\mathcal{O}}\left(dm + d\sqrt{\frac{mL}{\mu}}\right)$ вычислений $\nabla_y F_i$. Теорема 6 доказана.

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. *Lan G.* First-order and Stochastic Optimization Methods for Machine Learning. Atlanta: Springer, 2020.
3. *Гасников А.В.* Современные численные методы оптимизации. Метод универсального градиентного спуска. М.: МЦНМО, 2020.
4. *Alkousa M.S., Dvinskikh D.M., Stonyakin F.S., Gasnikov A.V., Kovalev D.* Accelerated Methods for Saddle Point Problems // Comput. Math. Math. Phys. 2020. V. 60. No. 11. P. 1787–1809.

5. *Gladin E., Kuruzov I., Stonyakin F., Pasechnyuk D., Alkousa M., Gasnikov A.* Solving strongly convex-concave composite saddle point problems with a small dimension of one of the variables. https://arxiv.org/pdf/2010.02280.pdf
6. *Tianyi L., Chi J., Michael I.J.* Near-Optimal Algorithms for Minimax Optimization. https://arxiv.org/pdf/2002.02417v5.pdf
7. *Yuanhao W., Jian L.* Improved Algorithms for Convex-Concave Minimax Optimization. https://arxiv.org/pdf/2006.06359.pdf
8. *Zhongruo Wang, Krishnakumar Balasubramanian, Shiqian Ma, Meisam Razaviyayn.* Zeroth-Order Algorithms for Nonconvex Minimax Problems with Improved Complexities. https://arxiv.org/pdf/2001.07819.pdf
9. *Гасников А.В., Гасникова Е.В.* Модели равновесного распределения транспортных потоков в больших сетях. Уч. пос. М.: МФТИ, 2020.
10. *Bolte J., Glaudin L., Pauwels E., Serrurier M.* A Hölderian backtracking method for min-max and min-min problems. https://arxiv.org/pdf/2007.08810.pdf
11. *Jungers M., Trélat E., Abou-Kandil H.* Min-Max and Min-Min Stackelberg Strategies with Closed-Loop Information Structure // J. Dynamical and Control Syst. Springer Verlag, 2011. No. 17 (3). P. 387–425.
12. *Konur D., Farhangi H.* Set-based Min-max and Min-min Robustness for Multi-objective Robust Optimization // Proc. 2017 Industrial and Systems Engineering Research Conf. K. Coperich, E. Cudney, H. Nembhard, eds.
13. *Vaidya P.M.* A New Algorithm for Minimizing Convex Functions over Convex Sets // Foundations of Computer Science, 1989. 30th Annual Sympos. 1989. P. 338–343.
14. *Vaidya P.M.* A new algorithm for minimizing convex functions over convex sets // Mathematical Programming 73. Springer, 1996. P. 291–341.
15. *Lan G., Zhize Li, Yi Zhou.* A unified variance-reduced accelerated gradient method for convex optimization // 33rd Conf. on Neural Information Processing Systems (NeurIPS 2019). Vancouver, Canada. https://arxiv.org/pdf/1905.12412.pdf
16. *Tyurin A.I., Gasnikov A.V.* Fast Gradient Descent Method for Convex Optimization Problems with an Oracle That Generates a (δ, L)-model of a Function in a Requested Point // Comput. Math. Math. Phys. 2019. V. 59. No. 7. P. 1137–1150.
17. *Bubeck S.* Convex Optimization: Algorithms and Complexity // Foundations and Trends in Machine Learning. 2015. V. 8. No. 3-4. P. 231–357.
18. *Gladin E., Sadiev A., Gasnikov A., Stonyakin F., Dvurechensky P., Beznosikov A., Alkousa M.* Solving smooth min-min and min-max problems by mixed oracle algorithms. https://arxiv.org/pdf/2103.00434.pdf
19. *Поляк Б.Т.* Введение в оптимизацию. М.: Наука, 1983.
20. *Гасников А.В., Двуреченский П.Е., Камзолов Д.И., Нестеров Ю.Е., Спокойный В.Г., Стецюк П.И., Суворикова А.Л., Чернов А.В.* Поиск равновесий в многостадийных транспортных моделях // Тр. Московского физико-технического института. 2015. № 7.4 (28).
21. *Bishop C.* Pattern recognition and machine learning. Springer, 2006.

Статья представлена к публикации членом редколлегии А.А. Лазаревым.

Поступила в редакцию 28.01.2021

После доработки 26.04.2021

Принята к публикации 30.06.2021