ISSN 0005-1179, Automation and Remote Control, 2021, Vol. 82, No. 10, pp. 1679–1691. © Pleiades Publishing, Ltd., 2021.
Russian Text © The Author(s), 2021, published in Avtomatika i Telemekhanika, 2021, No. 10, pp. 60–75.

Solving Convex Min-Min Problems with Smoothness and Strong Convexity in One Group of Variables and Low Dimension in the Other

E. Gladin[1,2*], **M. Alkousa**[1**], **and A. Gasnikov**[1,2***]

[1] *Moscow Institute of Physics and Technology, Dolgoprudnyi, Moscow oblast, 141701 Russia*
[2] *Kharkevich Institute for Information Transmission Problems,*
Russian Academy of Sciences, Moscow, 127051 Russia
*e-mail: *gladin.el@phystech.edu, **mohammad.alkousa@phystech.edu, ***gasnikov.av@mipt.ru*

Received January 28, 2021; revised April 26, 2021; accepted June 30, 2021

Abstract—The article deals with some approaches to solving convex problems of the min-min type with smoothness and strong convexity in only one of the two groups of variables. It is shown that the proposed approaches based on Vaidya's method, the fast gradient method, and the accelerated gradient method with variance reduction have linear convergence. It is proposed to use Vaidya's method to solve the exterior problem and the fast gradient method to solve the interior (smooth and strongly convex) one. Due to its importance for applications in machine learning, the case where the objective function is the sum of a large number of functions is considered separately. In this case, the accelerated gradient method with variance reduction is used instead of the fast gradient method. The results of numerical experiments are presented that illustrate the advantages of the proposed procedures for a logistic regression problem in which the a priori distribution for one of the two groups of variables is available.

Keywords: convex optimization, cutting plane method, Vaidya's method, variance reduction, fast gradient method, logistic regression

DOI: 10.1134/S0005117921100064

1. INTRODUCTION

The widespread extension of the acceleration construction of the conventional gradient method proposed in 1983 by Nesterov [1] to various other numerical optimization methods has become one of the main trends in research on numerical convex optimization methods in the last decade. Over the past 15 years, the accelerated method has been successfully transferred to smooth conditional convex optimization problems, to problems with structure (in particular, the so-called composite problems), and to gradient-free and randomized methods (for example, the accelerated gradient method with variance reduction for problems of minimizing the sum of functions [2]). Acceleration has also been successfully transferred to methods using higher derivatives. Details and a more elaborate overview of publications can be found in [3].

Optimization problems of the min-max type and saddle point problems have been widely studied in the literature due to their broad range of applications in statistics, machine learning, computer graphics, game theory, and other fields. Recently, many researchers have been actively working on the topic of accelerated methods for solving these problems, taking into account their structure; see [4–8], and these are just some of the latest publications. In some applications, there is a problem, similar to the min-max problem, which remains largely unexplored; this is a problem of the min-min type,

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

where $Q_x \subset \mathbb{R}^d$ and $Q_y \subset \mathbb{R}^n$ are nonempty compact convex sets of relatively low dimension d ($d \ll n$) and the function $F(x, y)$ is jointly convex in all variables and L-smooth and μ-strongly convex with respect to y. By L-smoothness with respect to y we mean the property

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

This statement arises, for example, when searching for equilibria in transport networks [9]. In machine learning, problems of this type correspond to the case where regularization is applied to one of the two groups of model parameters (hence the strong convexity in only one group of variables out of the two). For example, if a large group of features in a dataset are sparse, then regularization can only be used for model weights corresponding to these features. Another example is a logistic regression in which an a priori distribution for some of the parameters is available. Several publications are devoted to the min-min problem, including [10–12]. For example, in [10], the authors proposed new algorithms with automatically adjusted steps for min-max problems, but the proposed methods also apply to min-min problems.

The present paper discusses two approaches to solving problem (1), which have a linear convergence rate. It is proposed to reduce the problem under consideration to a set of auxiliary (interior and exterior) problems. The exterior problem (minimization with respect to x) is solved by Vaidya's cutting plane method [13, 14].

In the case where the objective function F is simple, i.e., it is not the sum of a large number of functions, the interior problem (minimization over y) is solved by the fast gradient method for strongly convex optimization problems. As a result of this approach, an approximate solution of problem (1) can be produced in $\widetilde{\mathcal{O}}\left(d\right)$ calculations of $\partial_x F$ and $\widetilde{\mathcal{O}}\left(d\sqrt{\frac{L}{\mu}}\right)$ computations of $\nabla_y F$; see Theorem 5. Here and in what follows, $\widetilde{\mathcal{O}}(\cdot) = \mathcal{O}(\cdot)$ up to a small power of a logarithmic factor; usually this exponent is one or two.

Optimizing the sum of a large number of functions has been the subject of intense research over the past few years due to its wide range of applications in machine learning, statistics, image processing, and other mathematical and engineering fields. Therefore, a separate case is considered where the objective function F is the sum (or the arithmetic mean) of a large number m of functions. In this case, the use of the fast gradient method for strongly convex optimization problems would require calculating the gradients of m terms at each step; this may take a long time. Instead, we propose to use the accelerated gradient method with variance reduction [2, 15], which also has linear convergence. As a result of this approach, the solution of the problem can be achieved in $\widetilde{\mathcal{O}}\left(md\right)$ calculations of $\partial_x F$ and in $\widetilde{\mathcal{O}}\left(md + d\sqrt{\frac{mL}{\mu}}\right)$ computations of $\nabla_y F$; see Theorem 6.

Using the two proposed approaches, we obtain the linear convergence rate for the min-min problem (1). Note that smoothness and strong convexity are required only in one of the two groups of variables.

The paper consists of five sections and the Appendix. Section 2 lists the algorithms used and their complexity, namely the fast gradient method, Vaidya's cutting plane method, and the accelerated gradient descent method with variance reduction. Section 3 formulates the problem statement and provides approaches to the problem under consideration for various cases of the objective function, in one of which the objective function is the sum or the arithmetic mean of a large number of functions. Section 4 contains the results of computational experiments and a comparison of the rates of the approaches proposed. Note that the complete proofs of Theorems 4, 5, and 6 as well as the auxiliary Assertion 1 are given in the Appendix.

2. THE METHODS USED

Let us present the algorithms used in the approaches to solving problem (1) proposed in the present paper. First, the fast gradient method is presented, then Vaidya's cutting plane method, and finally, the fast gradient method with variance reduction.

2.1. Fast Gradient Method

The paper [16] proposes an adaptive algorithm for solving the optimization problem

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

where $Q_y \subset \mathbb{R}^n$ is a nonempty compact convex set and f is an L-smooth convex function. This algorithm, dubbed the fast gradient method, permits one to accelerate the convergence of the conventional gradient descent from $\mathcal{O}\left(\frac{1}{N}\right)$ to $\mathcal{O}\left(\frac{1}{N^2}\right)$, where N is the number of algorithm iterations. The fast gradient method (its nonadaptive variant) is given below as Algorithm 1.

Algorithm 1. The fast gradient method [16].

```
Input: number of steps $N$, initial point $y^0 \in Q_y$, and parameter $L > 0$.
0-step: $z^0 := y^0, \quad u^0 := y^0, \quad \alpha_0 := 0, \quad A_0 := 0$.
for $k = 0, 1, \dots, N-1$ do
    find the greatest root $\alpha_{k+1}$ such that $A_k + \alpha_{k+1} = L\alpha_{k+1}^2$,
    $A_{k+1} := A_k + \alpha_{k+1}$,
    $z^{k+1} := \frac{\alpha_{k+1}u^k + A_k y^k}{A_{k+1}}$,
    $u^{k+1} := \arg\min_{y\in Q_y}\left\{\alpha_{k+1}\left\langle \nabla f(z^{k+1}), y - z^{k+1}\right\rangle + \frac{1}{2}\|y - u^k\|_2^2\right\}$,
    $y^{k+1} := \frac{\alpha_{k+1}u^{k+1} + A_k y^k}{A_{k+1}}$,
end for
Output: $y^N$.
```

The following theorem gives an estimate for the complexity (rate of convergence) of Algorithm 1.

Theorem 1 [16]. *Let a function $f : Q_y \to \mathbb{R}$ be L-smooth and convex; then algorithm 1 returns a point y^N such that*

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

where y_* is a solution of problem (2) and $R^2 = \frac{1}{2}\|y^0 - y_*\|_2^2$.

Next, we describe the technique of restarts of the fast gradient method (Algorithm 1) for the case of a μ-strongly convex function.

In view of the μ-strong convexity of f, we have

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

Then, after N_1 iterations of Algorithm 1, in view of Theorem 1 we obtain

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

and hence

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

Therefore, choosing $N_1 = \left\lceil 4\sqrt{\frac{L}{\mu}} \right\rceil$, where $\lceil \cdot \rceil$ is the ceiling, we obtain

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

After this, for Algorithm 1 we choose y^{N_1} for the start point and again perform N_1 iterations, and so on. To achieve an acceptable quality of solution, we can choose the number of restarts of Algorithm 1 (the parameter p in Algorithm 2) as follows:

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

In this case, the total number of iterations in Algorithm 2 will be

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

i.e.,

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

Algorithm 2. The fast gradient method for strongly convex optimization problems, restarts of Algorithm 1.

Input: initial point $y^0 \in Q_y$, $L > 0$, number of restarts $p = \left\lceil \frac{1}{2} \ln\left(\frac{\mu R^2}{\varepsilon}\right) \right\rceil$.

```
for $j = 1, \ldots, p$ do
    run $N_j = \left\lceil 4\sqrt{\frac{L}{\mu}} \right\rceil$ iterations of Algorithm 1,
    $y^0 := y^{N_j}$.
end for
```

Output: $\hat{y} := y^{N_p}$.

2.2. Vaidya's Method

Vaidya's cutting plane method was proposed in [13, 14] to solve the constrained optimization problem

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

where $Q_x \subset \mathbb{R}^d$ is a convex compact set with a nonempty interior and the objective function f defined on Q_x is continuous and convex.

Let $P = \{x \in \mathbb{R}^d : Ax \geqslant b\}$ be a bounded d-dimensional polyhedron, where $A \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$. The logarithmic barrier of the set P is defined as

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

where $a_i^\top$ is the ith row of the matrix A. The Hessian $H(x)$ of the function $Barr(x)$ is

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

The matrix $H(x)$ is positive definite for all x in the interior of P. The volumetric barrier $\mathcal{V}$ is defined as

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

where $\det\left(H(x)\right)$ designates the determinant of $H(x)$. The point of minimum of the function $\mathcal{V}$ on P will be referred to as the volumetric center of the set P.

Denote

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

then the gradient of the volumetric barrier $\mathcal{V}$ can be written as

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

Let $\mathcal{Q}(x)$ be defined as

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

Note that $\mathcal{Q}(x)$ is positive definite on the interior of P and also that $\mathcal{Q}(x)$ is a good approximation to the Hessian of the function $\mathcal{V}(x)$; i.e., $\nabla^2\mathcal{V}(x)$.

Vaidya's method generates a sequence of pairs $(A_k, b_k) \in \mathbb{R}^{m\times d} \times \mathbb{R}^m$ such that the corresponding polyhedra contain the solution. For the initial polyhedron, defined by the pair (A_0, b_0), one usually takes a simplex (the algorithm can start from any convex bounded n-dimensional polyhedron that easily yields to the calculation of its volumetric center, for example, from the n-rectangle).

One of the algorithm parameters is a small number $\gamma \leqslant 0.006$, the meaning of which is revealed in more detail in the book [17]. Let x_k $(k \geqslant 0)$ denote the volumetric center of the polyhedron defined by the pair (A_k, b_k), and suppose that the quantities $\{\sigma_i(x_k)\}_{1\leqslant i\leqslant m}$ have been calculated for this polyhedron (see (6)). The next polyhedron (A_{k+1}, b_{k+1}) is obtained from the current one as a result of either joining or removing a constraint:

1. If for some $i \in \{1, \ldots, m\}$ one has $\sigma_i(x_k) = \min\limits_{1\leqslant j\leqslant m} \sigma_j(x_k) < \gamma$, then (A_{k+1}, b_{k+1}) is obtained by eliminating the ith row from (A_k, b_k).
2. Otherwise, $\big($if $\min\limits_{1\leqslant j\leqslant m} \sigma_j(x_k) \geqslant \gamma\big)$, the oracle called up at the current point x_k returns a vector c_k such that $f(x) \leqslant f(x_k)$ $\forall x \in \left\{z \in Q_x : c_k^\top z \geqslant c_k^\top x_k\right\}$; i.e., $c_k \in -\partial f(x_k)$. Select $\beta_k \in \mathbb{R}$ such that

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

   Determine (A_{k+1}, b_{k+1}) by adding the row (c_k, β_k) to (A_k, b_k).

The volumetric barrier $\mathcal{V}_k$ is a self-concordant function; therefore, it can be efficiently minimized by the Newton method—one step of the Newton method for $\mathcal{V}_k$ made from x_{k-1} is sufficient. The details and analysis of Vaidya's method can be found in [13, 14, 17].

The following theorem gives an estimate for the complexity of Vaidya's algorithm.

Theorem 2. *Let $\mathcal{B}_\rho$ and $\mathcal{B}_\mathcal{R}$ be some Euclidean balls of radii ρ and $\mathcal{R}$, respectively, such that $\mathcal{B}_\rho \subseteq Q_x \subseteq \mathcal{B}_\mathcal{R}$, and let $B > 0$ be a number such that $|f(x) - f(x')| \leqslant B\ \forall x, x' \in Q_x$. Then Vaidya's method finds an ε-solution of problem (5) in $\mathcal{O}\left(d \log \frac{dB\mathcal{R}}{\rho\varepsilon}\right)$ steps.*

Remark 1. As shown in [18], Vaidya's method can be used with an imprecise subgradient without accumulating errors.

Remark 2. In addition to calculating the subgradient, the cost of iterating Vaidya's method includes the cost of inverting a $d \times d$ matrix and solving a system of linear equations.

2.3. Accelerated Gradient Method with Variance Reduction

Consider the problem

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

where $Q_y \subseteq \mathbb{R}^n$ is a closed convex set and the objective function f is the sum (or the arithmetic mean) of a large number m of smooth convex functions f_i; i.e., $f(y) = \frac{1}{m}\sum_{i=1}^m f_i(y)$. When solving (7) using the fast gradient method for strongly convex optimization problems (Algorithm 2), it will be required to calculate the gradient of m functions on each iteration; this is very expensive. Therefore, it is preferable to use a randomized gradient method instead of Algorithm 2, namely, the accelerated gradient method with variance reduction, also called Varag [2, 15]. The following Algorithm 3 is an accelerated gradient method with variance reduction (Varag) for a smooth strongly convex finite sum optimization problem (7). This algorithm was proposed by Lan et al. in [15].

Assume that for each $i \in \{1, \ldots, m\}$ there exists an $L_i > 0$ such that

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

It is clear that f has a Lipschitz gradient with constant at most $L := \frac{1}{m}\sum_{i=1}^m L_i$. Let us also assume that the objective function f is strongly convex with constant $\mu > 0$; i.e.,

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

Definition 1. A random vector $\bar{y}$ ranging in Q_y is called a stochastic ε-solution of problem (7) if $\mathbb{E}[f(\bar{y}) - f(y_*)] \leqslant \varepsilon$, where y_* is an exact solution of problem (7).

The Varag algorithm contains nested—exterior and interior—cycles (indexed by the variables s and t, respectively). On each iteration in the exterior cycle, the full gradient $\nabla f(\tilde{y})$ is calculated at the point $\tilde{y}$; it is then used in the inner loop for determining estimates for the gradient G_t. Each iteration in the inner loop requires information about the gradient of only one randomly selected term f_{i_t} and contains three main sequences $\{\underline{y}_t\}$, $\{y_t\}$, and $\{\bar{y}_t\}$.

Denote $s_0 := \lfloor \log_2 m \rfloor + 1$, where $\lfloor \cdot \rfloor$ is the floor. The parameters $\{q_1, \ldots, q_m\}$, $\{\theta_t\}$, $\{\alpha_s\}$, $\{\gamma_s\}$, $\{p_s\}$, and $\{T_s\}$ of Algorithm 3 are described as follows:

– The probabilities $q_i = \frac{1}{\sum_{i=1}^m L_i} L_i\ \forall i \in \{1, \ldots, m\}$.

– The weights $\{\theta_t\}$ for $1 \leqslant s \leqslant s_0$ or $s_0 < s \leqslant s_0 + \sqrt{\frac{12L}{m\mu}} - 4,\ m < \frac{3L}{4\mu}$ are equal to

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

In the remaining cases, they are equal to

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

where $\Gamma_t = (1+\mu\gamma_s)^t$.

– The parameters $\{T_s\}$, $\{\gamma_s\}$, and $\{p_s\}$ are defined as

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

– Finally,

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

Algorithm 3. The accelerated gradient method with variance reduction (Varag) [15].

Input: $y^0 \in Q_y$, $\{T_s\}$, $\{\gamma_s\}$, $\{\alpha_s\}$, $\{p_s\}$, $\{\theta_t\}$, and a probability distribution $\{q_1, \dots, q_m\}$ on $\{1, \dots, m\}$.

```
$\tilde{y}^0 := y^0$.
for $s = 1, 2, \dots$ 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 $t = 1, 2, \dots, T$ do
        choose $i_t \in \{1, \dots, m\}$ in a random way according to $\{q_1, \dots, q_m\}$.
        $\underline{y}_t := \frac{1}{(1+\mu\gamma_s(1-\alpha_s))}[(1+\mu\gamma_s)(1-\alpha_s-p_s)\bar{y}_{t-1} + \alpha_s y_{t-1} + (1+\mu\gamma_s)p_s\tilde{y}]$.
        $G_t := \frac{1}{(q_{i_t} m)}\left(\nabla f_{i_t}\left(\underline{y}_t\right) - \nabla f_{i_t}(\tilde{y})\right) + \tilde{g}$.
        $y_t := \arg\min_{y \in Q_y} \{\gamma_s\left(\langle G_t, y\rangle + \frac{\mu}{2}\|\underline{y}_t - y\|_2^2\right) + \frac{1}{2}\|y_{t-1} - y\|_2^2\}$.
        $\bar{y}_t := (1-\alpha_s-p_s)\,\bar{y}_{t-1} + \alpha_s y_t + p_s\tilde{y}$.
    end for
    $y^s := y_T, \tilde{y}^s := \frac{1}{\sum_{t=1}^{T}\theta_t}\sum_{t=1}^{T}(\theta_t\bar{y}_t)$.
end for
```

The following result gives an estimate for the complexity of Algorithm 3.

Theorem 3 [15]. *If the parameters $\{\theta_t\}$, $\{\alpha_s\}$, $\{\gamma_s\}$, $\{p_s\}$, and $\{T_s\}$ of Algorithm 3 are given according to formulas (8), (9), (10), and (11), then the total number of calculations of the gradients of the functions f_i performed by Algorithm 3 for finding the stochastic ε-solution of problem (7) is* *bounded,*

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

where $D_0 = 2\left(f(y^0) - f(y_*)\right) + \frac{3L}{2}\|y^0 - y_*\|_2^2$, *with* y_* *being a solution of problem* (7).

Note that the estimate (12) can be written as $N = \widetilde{\mathcal{O}}\left(m + \sqrt{\frac{mL}{\mu}}\right)$, where $\widetilde{\mathcal{O}}(\cdot) = \mathcal{O}(\cdot)$ up to a logarithmic factor in m, L, μ, ε, and D_0.

3. STATEMENT OF THE PROBLEM AND THE RESULTS OBTAINED

Consider the problem

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

where $Q_x \subset \mathbb{R}^d$ and $Q_y \subset \mathbb{R}^n$ are nonempty compact convex sets; the dimension d is relatively small ($d \ll n$) and the function $F(x, y)$ is jointly convex in all variables and L-smooth and μ-strongly convex with respect to y. By L-smoothness with respect to y we mean the property

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

We introduce the function

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

Problem (13) can be rewritten in the form

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

Solving (15) by some iterative method involves solving the auxiliary problem (14) on each step so as to approximately find the subgradient $\partial f(x)$. Let us proceed to the following definition.

Definition 2 [19, p. 123]. Let $\delta \geqslant 0$, let $Q_x \subseteq \mathbb{R}^d$ be a convex set, and let $f : Q_x \to \mathbb{R}$ be a convex function. A vector $g \in \mathbb{R}^d$ is called a δ-subgradient of f at a point $x' \in Q_x$ if

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

The set of δ-subgradients of f at x' is denoted by $\partial_\delta f(x')$.

Denote $D := \max_{y,z \in Q_y} \|y - z\|_2$ and $y(x) := \arg\min_{y \in Q_y} F(x, y)$. The next theorem tells us how to calculate the δ-subgradient of the function $f(x)$ by approximately solving the auxiliary problem (15).

Theorem 4. *Suppose that a* $\tilde{y} \in Q_x$ *such that* $F(x, \tilde{y}) - f(x) \leqslant \varepsilon$ *has been found. Then*

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

This theorem is implied directly by the following two assertions.

Assertion 1. *Let $g : Q_y \to \mathbb{R}$ be an L-smooth μ-strongly convex function, and let a point $\tilde{y} \in Q_y$ be such that $g(\tilde{y}) - g(y_*) \leqslant \varepsilon$. Then*

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

where $y_* = \arg\min_{y \in Q_y} g(y)$.

Assertion 2 [20, p. 12]. *Assume that a $\tilde{y} \in Q_y$ has been found such that*

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

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

Intuitively, Theorem 4 says that, having solved the auxiliary problem (14) sufficiently accurately, we obtain a good approximation to the subgradient $\partial f(x)$, which can be used to solve the exterior problem (15). The proposed approach to solving (13) is based on this idea.

Approach 1 (the main case). The exterior problem (15) is solved by Vaidya's method. The auxiliary problem (14) is solved by the fast gradient method for strongly convex optimization problems (Algorithm 2).

Theorem 5. *Approach 1 allows one to obtain an ε-solution of problem (13) after $\widetilde{\mathcal{O}}\,(d)$ calculations of $\partial_x F$ and inversions of matrices of size $d \times d$ and $\widetilde{\mathcal{O}}\left(d\sqrt{\frac{L}{\mu}}\right)$ calculations of $\nabla_y F$.*

Remark 3. The inversion of matrices occurs in the complexity of the proposed approach owing to the fact that it is performed at each step of Vaidya's method.

3.1. Minimizing the Sum of a Large Number of Functions

Suppose that in problem (13) we have

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

where the functions F_i are jointly convex in all variables and L_i-smooth with respect to y, while F is μ-strongly convex in y. It follows that F is jointly convex in all variables and smooth with respect to y with smoothness constant at most $L := \frac{1}{m} \sum_{i=1}^{m} L_i$.

Approach 2 (the sum of functions). The exterior problem (15) is solved by Vaidya's method. The auxiliary problem (14) is solved by the accelerated gradient method with variance reduction (Algorithm 3).

Theorem 6. *Approach 2 allows one to obtain an ε-solution of problem (13) in $\widetilde{\mathcal{O}}(md)$ calculations of $\partial_x F_i$, $\widetilde{\mathcal{O}}\,(d)$ inversions of matrices of size $d \times d$, and $\widetilde{\mathcal{O}}\left(dm + d\sqrt{\frac{mL}{\mu}}\right)$ calculations of $\nabla_y F_i$.*

4. EXPERIMENTS

Consider the model of logistic regression for the binary classification problem. The error of the model with parameters w on a training object with a feature vector z belonging to the class $t \in \{-1, 1\}$ is written as

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

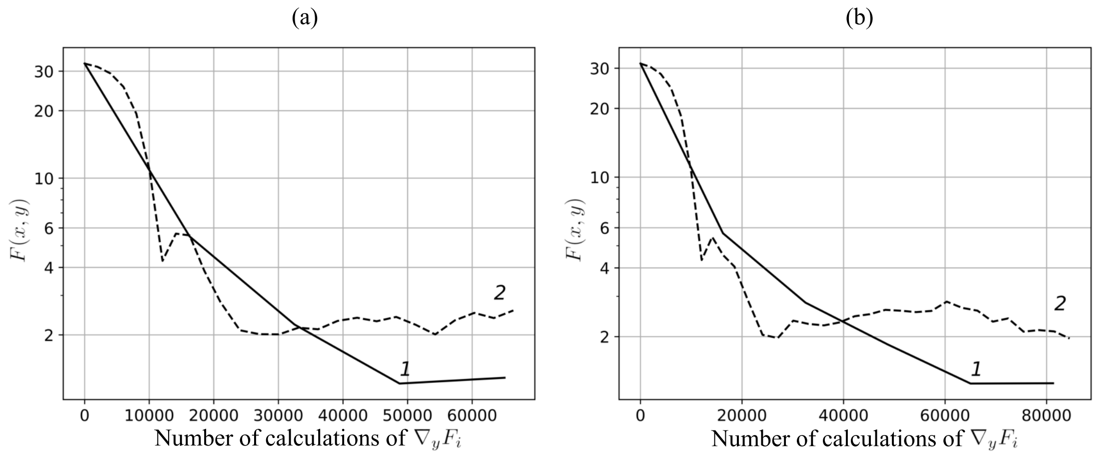

Fig. 1. Panels (a) and (b) correspond to dimensions $d = 20$ and $d = 30$, respectively. Graphs *1* and *2* demonstrate the convergence of the proposed approach and the Varag method, respectively.

Let the model parameters consist of two groups, $w = (x, y)$, $x \in \mathbb{R}^d$, $y \in \mathbb{R}^n$, with the Gaussian a priori distribution

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

being given for the group y, where I_n is the identity matrix of size n. The maximization of the a posteriori probability will lead (see [21, Sec. 4.5.1]) to the problem

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

where for Q_x and Q_y we can take Euclidean balls of sufficiently large radii.

We will solve problem (17) using Approach 2 and compare its operation with that of the Varag method (Algorithm 3). Note that this problem is not jointly strongly convex in all variables. For such a statement, one can use the Varag, setting the parameters θ_t by the formula (8) and all the remaining parameters by the formulas for the strongly convex case with $\mu = 0$; see [15]. In this case, the stochastic ε-solution will be found in $\mathcal{O}\left(\sqrt{\frac{mD_0}{\varepsilon}} + m \log m\right)$ calculations of the gradients of the functions F_i, where $D_0 = 2\left(F(x^0, y^0) - F(x_*, y_*)\right) + \frac{3L}{2}\|(x^0, y^0) - (x_*, y_*)\|_2^2$, (x_*, y_*) being a solution of problem (17). This sublinear estimate is inferior to the approach proposed in the paper; see Theorem 2.

In experiments we used the dataset madelon, containing 2000 objects with 500 features. The small regularization coefficient $\frac{1}{\sigma^2} = 0.005$ was chosen, and the experiments were run for two dimensions d equal to 20 and 30.

Figure 1 shows the results of the experiment. The x-axis represents the number of calculations of the gradient $\nabla_y F_i$, which for the Varag is the same as the number of calculations of $\nabla_x F_i$. Note that the proposed approach requires less computations of $\nabla_x F_i$, since they are performed only in the outer loop. For example, graph *1* in Fig. 1a corresponds to four iterations of the outer loop (i.e., 8 000 calculations of $\nabla_x F_i$), and graph *1* in Fig. 1b, to five iterations (i.e., 10 000 calculations of $\nabla_x F_i$). In this experiment, Approach 2 made it possible to achieve lower values of the objective function.

The source code and experimental results can be found in the repository https://github.com/egorgladin/min_min.

5. CONCLUSIONS

In the present paper, we have considered the min-min type problem

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

where $Q_x \subset \mathbb{R}^d$ and $Q_y \subset \mathbb{R}^n$ are nonempty compact convex sets, the dimension d is relatively small ($d \ll n$), and the function $F(x,y)$ is jointly convex in all variables and also L-smooth and μ-strongly convex with respect to y.

Two approaches to solving problem (18) are proposed. In these approaches, the problem is reduced to a set of auxiliary (interior and exterior) problems. The exterior problem (minimization over x) is solved by Vaidya's method, and the interior one (minimization over y) is solved by the fast gradient method for strongly convex optimization problems or, if the sum of a large number of functions is being minimized, by the accelerated gradient method with variance reduction. This allows achieving an approximate solution of problem (18) in $\widetilde{\mathcal{O}}\left(d\right)$ calculations of $\partial_x F$ and $\widetilde{\mathcal{O}}\left(d\sqrt{\frac{L}{\mu}}\right)$ calculations of $\nabla_y F$; see Theorem 5. For comparison, were problem (18) jointly smooth in all variables, its solution using only the fast gradient method would have had the complexity $\mathcal{O}\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)$, where R is the distance from the initial approximation to the solution. In the case of a sum with m terms, the solution of the problem can be achieved in $\widetilde{\mathcal{O}}\left(md\right)$ calculations of $\partial_x F$ and in $\widetilde{\mathcal{O}}\left(md + d\sqrt{\frac{mL}{\mu}}\right)$ calculations of $\nabla_y F$; see Theorem 6.

A numerical experiment has been carried out in which one of the proposed approaches is used for the problem of logistic regression with regularization applied to one of the two groups of model parameters. Compared to the Varag algorithm, the proposed approach achieves lower function values with fewer oracle calls.

Note also that if the function $F(x,y)$ is jointly μ-strongly convex in all variables, then the function $g(y) = \min_{x\in Q_x} F(x,y)$ will be μ-strongly convex as well. Moreover, all this can be stated in terms of the (δ,μ,L)-oracle (see [3] and the literature cited therein). This is done in [20] for $\mu = 0$, and for $\mu > 0$ the proof almost word for word reproduces Assertions 1 and 3 in [20] (see also [9]). The above observation permits reasonably (with theoretical elaboration) using Vaidya's method to solve the interior problem, and employ, for example, the fast gradient method to solve the exterior problem. However, this approach will be preferable to the one discussed in this paper only under very special (usually difficult to implement) conditions [5].

APPENDIX

Proof of Assertion 1. Consider an arbitrary $y \in Q_y$,

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

Let us produce an upper bound for the first term using the Cauchy–Schwarz inequality and the definition of the Lipschitz property of gradient,

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

It follows from the strong convexity that

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

Using the inequalities $g(\tilde{y}) - g(y_*) \leqslant \varepsilon$ and $\langle \nabla g(y_*), y - y_* \rangle \geqslant 0 \ \forall y \in Q_y$, we obtain

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

Now let us estimate the second term in (A.1) from above,

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

Let us use the criterion of optimality of the point y_* and the Cauchy–Schwarz inequality one more time to obtain

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

Combining the upper bounds for both terms, we obtain

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

which implies the desired Assertion 1. ■

Proof of Theorem 4. Fixing $x \in Q_x$, we apply Assertion 1 to the function $g(y) := F(x,y)$ and Assertion 2. The proof of Theorem 4 is complete. ■

Proof of Theorem 5. According to (4), Algorithm 2 converges linearly; therefore, we can assume that the auxiliary problem $\min\limits_{y \in Q_y} F(x,y)$ can be solved arbitrarily precisely in time $\widetilde{\mathcal{O}}\left(\sqrt{\frac{L}{\mu}}\right)$. According to Theorem 4, this allows using the δ-subgradient, where δ decreases exponentially. For the exterior problem we use Vaidya's method, which also converges linearly and has complexity $\widetilde{\mathcal{O}}\left(d\right)$. Thus, to solve problem (13) it suffices to perform $\widetilde{\mathcal{O}}\left(d\right)$ calculations of $\partial_x F$ and inversions of matrices of size $d \times d$ as well as $\widetilde{\mathcal{O}}\left(d\sqrt{\frac{L}{\mu}}\right)$ calculations of $\nabla_y F$. The proof of Theorem 5 is complete. ■

Proof of Theorem 6. According to Theorem 3, Varag converges linearly; therefore, we can assume that the auxiliary problem $\min\limits_{y \in Q_y} F(x,y)$ is solved arbitrarily precisely in time $\widetilde{\mathcal{O}}\left(m + \sqrt{\frac{mL}{\mu}}\right)$. According to Theorem 4, this allows using the δ-subgradient, where δ decreases exponentially. For the exterior problem we use Vaidya's method, which also converges linearly and has the complexity of $\widetilde{\mathcal{O}}\left(d\right)$ iterations. On each of its iterations, we need to calculate the subgradients of all m terms $\partial_x F_i$. Thus, to solve the problem, it suffices to perform $\widetilde{\mathcal{O}}\left(md\right)$ calculations of $\partial_x F_i$, $\widetilde{\mathcal{O}}\left(d\right)$ inversions of matrices of size $d \times d$, and $\widetilde{\mathcal{O}}\left(dm + d\sqrt{\frac{mL}{\mu}}\right)$ calculations of $\nabla_y F_i$. The proof of Theorem 6 is complete. ■

FUNDING

This work was supported by the Ministry of Science and Higher Education of the Russian Federation, state assignment no. 075-00337-20-03, project no.0714-2020-0005. The work of E.L. Gladin was also supported by an A.M. Raigorodskii scholarship in the field of numerical optimization methods. The work of A.V. Gasnikov was also partly supported by the Russian Foundation for Basic Research, project no. 18-29-03071 mk.

REFERENCES

1. Nesterov, Yu.E., Method of minimizing convex functions with convergence rate $O(1/k^2)$, *Dokl. Akad. Nauk SSSR*, 1983, vol. 269, no. 3, pp. 543–547.

2. Lan, G., *First-order and Stochastic Optimization Methods for Machine Learning*, Atlanta: Springer, 2020.

3. Gasnikov, A.V., *Sovremennye chislennye metody optimizatsii. Metod universal'nogo gradientnogo spuska* (Modern Numerical Optimization Methods. Universal Gradient Descent Method), Moscow: MTsNMO, 2020.

4. Alkousa, M.S., Dvinskikh, D.M., Stonyakin, F.S., Gasnikov, A.V., and Kovalev, D., Accelerated methods for saddle point problems, *Comput. Math. Math. Phys.*, 2020, vol. 60, no. 11, pp. 1787–1809.

5. Gladin, E., Kuruzov, I., Stonyakin, F., Pasechnyuk, D., Alkousa, M., and 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., and Michael, I.J., Near-Optimal Algorithms for Minimax Optimization. https://arxiv.org/pdf/2002.02417v5.pdf.

7. Yuanhao, W. and Jian, L., Improved Algorithms for Convex-Concave Minimax Optimization. https://arxiv.org/pdf/2006.06359.pdf.

8. Zhongruo Wang, Krishnakumar Balasubramanian, Shiqian Ma, and Meisam Razaviyayn, Zeroth-Order Algorithms for Nonconvex Minimax Problems with Improved Complexities. https://arxiv.org/pdf/2001.07819.pdf.

9. Gasnikov, A.V. and Gasnikova, E.V., *Modeli ravnovesnogo raspredeleniya transportnykh potokov v bol'shikh setyakh. Uch. pos.* (Equilibrium Distribution Models of Traffic Flows in Large Networks. A Handbook), Moscow: Mosk. Fiz.-Tekh. Inst., 2020.

10. Bolte, J., Glaudin, L., Pauwels, E., and 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., and Abou-Kandil, H., Min-max and min-min Stackelberg strategies with closed-loop information structure, *J. Dyn. Control Syst. Springer,* 2011, no. 17(3), pp. 387–425.

12. Farhangi, H. and Konur, D., Set-based min-max and min-min robustness for multi-objective robust optimization, *Proc. 67th Annu. Conf. Expo Inst. Ind. Eng.* (Pittsburgh, PA, 2017), Inst. Ind. Eng. (IIE), 2017, pp. 1217–1222.

13. Vaidya, P.M., A new algorithm for minimizing convex functions over convex sets, *Foundations of Computer Science. 30th Annu. Symp.* (1989), pp. 338–343.

14. Vaidya, P.M., A new algorithm for minimizing convex functions over convex sets, *Math. Program.*, 1996, vol. 73, pp. 291–341.

15. Lan, G., Zhize Li, and Yi, Zhou., A unified variance-reduced accelerated gradient method for convex optimization, *33rd Conf. Neural Inf. Process. Syst. (NeurIPS 2019)* (Vancouver, Canada, 2019). https://arxiv.org/pdf/1905.12412.pdf.

16. Tyurin, A.I. and 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, vol. 59, no. 7, pp. 1137–1150.

17. Bubeck, S., Convex optimization: algorithms and complexity, *Found. Trends Mach. Learn.,* 2015, vol. 8, nos. 3–4, pp. 231–357

18. Gladin, E., Sadiev, A., Gasnikov, A., Stonyakin, F., Dvurechensky, P., Beznosikov, A., and Alkousa, M., Solving Smooth Min-Min and Min-Max Problems by Mixed Oracle Algorithms. https://arxiv.org/pdf/2103.00434.pdf.

19. Polyak, B.T., *Vvedenie v optimizatsiyu* (Introduction to Optimization), Moscow: Nauka, 1983.

20. Gasnikov, A.V., Dvurechenskii, P.E., Kamzolov, D.I., Nesterov, Yu.E., Spokoinyi, V.G., Stetsyuk, P.I., Suvorikova, A.L., and Chernov, A.V., Finding equilibria in multistage transport models, *Tr. Mosk. Fiz.-Tekh. Inst.*, 2015, vol. 7, no. 4(28), pp. 143–155.

21. Bishop, C., *Pattern Recognition and Machine Learning*, New York: Springer, 2006.

This paper was recommended for publication by A.A. Lazarev, a member of the Editorial Board