

Algorithms for Solving Variational inequalities and Saddle Point Problems with some Generalizations of Lipschitz Property for Operators*

Alexander A. Titov^{1,3}[0000-0001-9672-0616], Fedor S.
Stonyakin^{1,2}[0000-0002-9250-4438], Mohammad S.
Alkousa^{1,3}[0000-0001-5470-0182], and Alexander V.
Gasnikov^{1,3,4,5}[0000-0002-5982-8983]

¹ Moscow Institute of Physics and Technology, Moscow, Russia

² V. I. Vernadsky Crimean Federal University, Simferopol, Russia

³ HSE University, Moscow, Russia

⁴ Institute for Information Transmission Problems RAS, Moscow, Russia

⁵ Caucasus Mathematical Center, Adyghe State University, Maikop, Russia
a.a.titov@phystech.edu, fedyor@mail.ru, mohammad.alkousa@phystech.edu,
gasnikov@yandex.ru

Abstract. The article is devoted to the development of numerical methods for solving saddle point problems and variational inequalities with simplified requirements for the smoothness conditions of functionals. Recently, some notable methods for optimization problems with strongly monotone operators were proposed. Our focus here is on newly proposed techniques for solving strongly convex-concave saddle point problems. One of the goals of the article is to improve the obtained estimates of the complexity of introduced algorithms by using accelerated methods for solving auxiliary problems. The second focus of the article is introducing an analogue of the boundedness condition for the operator in the case of arbitrary (not necessarily Euclidean) prox structure. We propose an analogue of the Mirror Descent method for solving variational inequalities with such operators, which is optimal in the considered class of problems.

Keywords: Strongly Convex Programming Problem. Relative Boundedness. Inexact Model. Variational Inequality. Saddle Point Problem.

Introduction

Modern numerical optimization methods are widely used in solving problems in various fields of science. The problem of finding the optimal value in the inves-

* The research in Introduction and Sections 2, 3 is supported by the Ministry of Science and Higher Education of the Russian Federation (Gos zadaniye in MIPT, project 075-00337-20-03). The research in Theorem 1 and partially in section 2.1 was supported by the Russian Science Foundation (project 18-71-10044).

tigated mathematical model naturally arises in machine learning, data analysis, economics, market equilibrium problems, electric power systems, control theory, optimal transport, molecular modeling, etc. This paper is devoted to the development and analysis of numerical methods for solving saddle point problems and variational inequalities. Both saddle point problems and variational inequalities play a critical role in structural analysis, resistive networks, image processing, zero-sum game, and Nash equilibrium problems [4,6,8,13].

Despite variational inequalities and saddle point problems are closely related (it will be discussed later), the article consists of two conditionally independent parts, each devoted to the special type of generalization of smoothness conditions and the corresponding problem.

Firstly, the paper focuses on recently proposed numerical methods for solving (μ_x, μ_y) -strongly convex-concave saddle point problems of the following form:

$$\min_x \max_y f. \quad (1)$$

In particular, in [1] authors considered the modification of some well-known scheme for speeding up methods for solving *smooth* saddle point problems. The smoothness, in such a setting, means the Lipschitz continuity of all partial gradients of the objective function f . We improve the obtained estimates of the complexity for the case of non-smooth objective.

In more details, we consider the saddle point problem under the assumption, that one of the partial gradients still satisfies the Lipschitz condition, while the other three satisfy simplified smoothness condition, namely, the Hölder continuity. Note, it is the Lipschitz continuity, which makes it possible to use accelerated methods for solving auxiliary problems. The Hölder continuity is an important generalization of the Lipschitz condition, it appears in a large number of applications [9,14]. In particular, if a function is uniformly convex, then its conjugated will necessarily have the Hölder-continuous gradient [9].

We prove, that the strongly convex-concave saddle point problem, modified in a functional form, admits an inexact (δ, L, μ_x) -model, where δ is comparable to ε , and apply the Fast Gradient Method to achieve the ε -solution of the considered problem. The total number of iterations is estimated as follows:

$$\mathcal{O} \left(\sqrt{\frac{L}{\mu_x}} \cdot \sqrt{\frac{L_{yy}}{\mu_y}} \cdot \log \frac{2L_{yy}R^2}{\varepsilon} \cdot \log \frac{2LD^2}{\varepsilon} \right),$$

$$\text{where } L = \tilde{L} \left(\frac{\tilde{L}(1-\nu)(2-\nu)}{2\varepsilon} \right)^{\frac{(1-\nu)(1+\nu)}{2-\nu}}, \tilde{L} = \left(L_{xy} \left(\frac{2L_{xy}}{\mu_y} \right)^{\frac{\nu}{2-\nu}} + L_{xx} D^{\frac{\nu-\nu^2}{2-\nu}} \right),$$

$L_{xx}, L_{xy}, L_{yy} > 0$, ν is the Hölder exponent of ∇f , D is a diameter of the domain of $f(x, \cdot)$, R denotes the distance between the initial point of the algorithm x^0 and the exact solution of the problem. Remind that (\tilde{x}, \tilde{y}) is called an ε -solution of the saddle point problem, if

$$\max_y f(\tilde{x}, y) - \min_x f(x, \tilde{y}) \leq \varepsilon. \quad (2)$$

Let us note, that due to the strong convexity-concavity of the considered saddle point problem, the convergence in argument takes place with the similar asymptotic behavior. Thus, we can consider the introduced definition of the ε -solution (2) for the strongly convex-concave saddle point problem (1).

The second part of the article continues with the study of accelerated methods and is devoted to the numerical experiments for some recently proposed [14] universal algorithms. In particular, we consider the restarted version of the algorithm and apply it to the analogue of the covering circle problem with non-smooth functional constraints. We show, that the method can work faster, than $O(\frac{1}{\varepsilon})$.

Finally, the third part of the article explores Minty variational inequalities with the Relatively bounded operator. Recently Y. Nesterov proposed [10] a generalization of the Lipschitz condition, which consists in abandoning the classical boundedness of the norm of the objective's gradient ∇f in favor of a more complex structure. This structure allows one to take into account the peculiarities of the domain of the optimization problem, which can be effectively used in solving support vector machine (SVM) problem and intersection of n ellipsoids problem [7,15].

Remind the formulation of the Minty variational inequality. For a given operator $g(x) : X \rightarrow \mathbb{R}$, where X is a closed convex subset of some finite-dimensional vector space, we need to find a vector $x_* \in X$, such that

$$\langle g(x), x_* - x \rangle \leq 0, \quad \forall x \in X. \quad (3)$$

We consider the variational inequality problem under the assumption of Relative boundedness of the operator g , which is the modification of the aforementioned Relative Lipschitz condition for functionals. We introduce the modification of the Mirror Descent method to solve such variational inequality problems. The proposed method guarantees an $(\varepsilon + \sigma)$ -solution of the problem after no more than

$$\frac{2RM^2}{\varepsilon^2}$$

iterations, where M is the Relative boundedness constant, which depends on the characteristics of the operator g and R can be understood as the distance between the initial point of the algorithm and the exact solution of the variational inequality in some generalized sense. The constant σ reflects certain features of the monotonicity of the operator g . An $(\varepsilon + \sigma)$ -solution of the variational inequality is understood as the point \tilde{x} , such that

$$\max_{x \in X} \langle g(x), \tilde{x} - x \rangle \leq \varepsilon + \sigma. \quad (4)$$

The paper consists of the introduction and 3 main sections. In Sect. 1, we consider the strongly convex-concave saddle point problem in non-a smooth setting. Sect. 2, is devoted to some numerical experiments concerning the methods, recently proposed in [14] and the analysis of their asymptotics in comparison with the method proposed in Sect. 1. In Sect. 3, we consider the Minty variational inequality problem with Relatively Bounded operator.

To sum it up, the contributions of the paper can be formulated as follows:

- We consider the strongly convex-concave saddle point problem and its modified functional form. We prove, that the functional form admits an inexact (δ, L, μ_x) -model and apply the Fast Gradient Method to achieve the ε -solution of the considered problem. Moreover, we show, that $\delta = \mathcal{O}(\varepsilon)$.
- We show, that the total number of iterations of the proposed method does not exceed

$$\mathcal{O} \left(\sqrt{\frac{L}{\mu_x}} \cdot \sqrt{\frac{L_{yy}}{\mu_y}} \cdot \log \frac{2L_{yy}R^2}{\varepsilon} \cdot \log \frac{2LD^2}{\varepsilon} \right),$$

where $L = \tilde{L} \left(\frac{\tilde{L}}{2\varepsilon} \frac{(1-\nu)(2-\nu)}{2-\nu} \right)^{\frac{(1-\nu)(1+\nu)}{2-\nu}}$, $\tilde{L} = \left(L_{xy} \left(\frac{2L_{xy}}{\mu_y} \right)^{\frac{\nu}{2-\nu}} + L_{xx} D^{\frac{\nu-\nu^2}{2-\nu}} \right)$.

- We introduce the modification of the Mirror Descent method to solve Minty variational inequalities with Relatively Bounded and σ -monotone operators.
- We show, that the proposed method can be applied to obtain an $(\varepsilon + \sigma)$ -solution after no more than $\frac{2RM^2}{\varepsilon^2}$ iterations.

1 Accelerated Method for Saddle Point Problems with Generalized Smoothness Condition

Let $Q_x \subset \mathbb{R}^n$ and $Q_y \subset \mathbb{R}^m$ be nonempty, convex, compact sets and there exist $D > 0$ and $R > 0$, such that

$$\|x_1 - x_2\|_2 \leq D \quad \forall x_1, x_2 \in Q_x,$$

$$\|y_1 - y_2\|_2 \leq R \quad \forall y_1, y_2 \in Q_y,$$

Let $f : Q_x \times Q_y \rightarrow \mathbb{R}$ be a μ_x -strongly convex function for fixed $y \in Q_y$ and μ_y -strongly concave for fixed $x \in Q_x$. Remind, that differentiable function $h(x) : Q_x \rightarrow \mathbb{R}$ is called μ -strongly convex, if

$$\langle \nabla h(x_1) - \nabla h(x_2), x_1 - x_2 \rangle \geq \mu \|x_1 - x_2\|_2^2 \quad \forall x_1, x_2 \in Q_x.$$

Consider (μ_x, μ_y) -strongly convex-concave saddle point problem (1) under assumptions, that one of the partial gradients of f satisfies the Lipschitz condition, while other three gradients satisfy the Hölder condition. More formally, for any $x, x' \in Q_x, y, y' \in Q_y$ and for some $\nu \in [0, 1]$, the following inequalities hold:

$$\|\nabla_x f(x, y) - \nabla_x f(x', y)\|_2 \leq L_{xx} \|x - x'\|_2^\nu, \quad (5)$$

$$\|\nabla_x f(x, y) - \nabla_x f(x, y')\|_2 \leq L_{xy} \|y - y'\|_2^\nu, \quad (6)$$

$$\|\nabla_y f(x, y) - \nabla_y f(x', y)\|_2 \leq L_{xy} \|x - x'\|_2^\nu, \quad (7)$$

$$\|\nabla_y f(x, y) - \nabla_y f(x, y')\|_2 \leq L_{yy} \|y - y'\|_2. \quad (8)$$

Define the following function:

$$g(x) = \max_{y \in Q_y} f, \quad x \in Q_x. \quad (9)$$

It is obvious, that the considered saddle point problem (1) can be rewritten in the following more simple way:

$$g(x) = \max_{y \in Q_y} f \rightarrow \min_{x \in Q_x}. \quad (10)$$

Since $f(x, \cdot)$ is μ_y -strongly concave on Q_y , the maximization problem (9) has the unique solution

$$y^*(x) = \arg \max_{y \in Q_y} f \quad \forall x \in Q_x,$$

so $g(x) = f(x, y^*(x))$. Moreover,

$$\|y^*(x_1) - y^*(x_2)\|_2^2 \leq \frac{2}{\mu_y} \left(f(x_1, y^*(x_1)) - f(x_1, y^*(x_2)) \right) \quad \forall x_1, x_2 \in Q_x. \quad (11)$$

Lemma 1. *Consider the problem 1 under assumptions (5)–(8). Define the function $g(x) : Q_x \rightarrow \mathbb{R}$ according to (9). Then $g(x)$ has the Hölder continuous gradient with Hölder constant $\left(L_{xy} \left(\frac{2L_{xy}}{\mu_y} \right)^{\frac{\nu}{2-\nu}} + L_{xx} D^{\frac{\nu-\nu^2}{2-\nu}} \right)$ and Hölder exponent $\frac{\nu}{2-\nu}$.*

Proof. Similarly to [1], for any $x_1, x_2 \in Q_x$, let us estimate the following difference:

$$\begin{aligned} & \left(f(x_1, y^*(x_1)) - f(x_1, y^*(x_2)) \right) - \left(f(x_2, y^*(x_1)) - f(x_2, y^*(x_2)) \right) \leq \\ & \leq \left\| \nabla_x f(x_1 + t(x_2 - x_1), y^*(x_1)) - \nabla_x f(x_1 + t(x_2 - x_1), y^*(x_2)) \right\|_2 \cdot \|x_2 - x_1\|_2 \leq \\ & \leq L_{xy} \|y^*(x_1) - y^*(x_2)\|_2^\nu \cdot \|x_2 - x_1\|_2, \quad t \in [0, 1]. \end{aligned}$$

Using (11), one can get

$$\|y^*(x_1) - y^*(x_2)\|_2 \leq \left(\frac{2L_{xy}}{\mu_y} \right)^{\frac{1}{2-\nu}} \|x_2 - x_1\|_2^{\frac{1}{2-\nu}},$$

which means, that $y^*(x)$ satisfies the Hölder condition on Q_x with Hölder constant $\left(\frac{2L_{xy}}{\mu_y} \right)^{\frac{1}{2-\nu}}$ and Hölder exponent $\frac{1}{2-\nu} \in [\frac{1}{2}, 1]$.

Further,

$$\begin{aligned} & \|\nabla g(x_1) - \nabla g(x_2)\|_2 = \|\nabla_x f(x_1, y^*(x_1)) - \nabla_x f(x_2, y^*(x_2))\|_2 \leq \\ & \leq \|\nabla_x f(x_1, y^*(x_1)) - \nabla_x f(x_1, y^*(x_2))\|_2 + \|\nabla_x f(x_1, y^*(x_2)) - \nabla_x f(x_2, y^*(x_2))\|_2 \leq \end{aligned}$$

$$\begin{aligned}
&\leq L_{xy}\|y^*(x_1) - y^*(x_2)\|_2^\nu + L_{xx}\|x_2 - x_1\|_2^\nu = \\
&\leq L_{xy} \left(\frac{2L_{xy}}{\mu_y} \right)^{\frac{\nu}{2-\nu}} \|x_2 - x_1\|_2^{\frac{\nu}{2-\nu}} + L_{xx}\|x_2 - x_1\|_2^\nu = \\
&= L_{xy} \left(\frac{2L_{xy}}{\mu_y} \right)^{\frac{\nu}{2-\nu}} \|x_2 - x_1\|_2^{\frac{\nu}{2-\nu}} + L_{xx}\|x_2 - x_1\|_2^{\frac{\nu}{2-\nu}} \cdot \|x_2 - x_1\|_2^{\frac{\nu-\nu^2}{2-\nu}}.
\end{aligned}$$

Since Q_x is bounded, we have

$$\|\nabla g(x_1) - \nabla g(x_2)\|_2 \leq \left(L_{xy} \left(\frac{2L_{xy}}{\mu_y} \right)^{\frac{\nu}{2-\nu}} + L_{xx} D^{\frac{\nu-\nu^2}{2-\nu}} \right) \|x_2 - x_1\|_2^{\frac{\nu}{2-\nu}}, \quad (12)$$

which means, that $g(x)$ has the Hölder continuous gradient.

Definition 1. A function $h(x) : Q_x \rightarrow \mathbb{R}$ admits (δ, L, μ) -model, if, for any $x_1, x_2 \in Q_x$, the following inequalities hold:

$$\begin{aligned}
&\frac{\mu}{2}\|x_2 - x_1\|_2^2 + \langle \nabla h(x_1), x_2 - x_1 \rangle + h(x_1) - \delta \leq h(x_2) \leq h(x_1) \\
&\quad + \langle \nabla h(x_1), x_2 - x_1 \rangle + \frac{L}{2}\|x_2 - x_1\|_2^2 + \delta
\end{aligned} \quad (13)$$

Remark 1. [5] Note, that if a function $h(x)$ has the Hölder-continuous gradient with Hölder constant $\tilde{L}_{\tilde{\nu}}$ and Hölder exponent $\tilde{\nu}$, then $h(x)$ admits (δ, L, μ) -model. More precisely, inequalities (13) hold with $L = \tilde{L}_{\tilde{\nu}} \left(\frac{\tilde{L}_{\tilde{\nu}}}{2\delta} \frac{1-\tilde{\nu}}{1+\tilde{\nu}} \right)^{\frac{1-\tilde{\nu}}{1+\tilde{\nu}}}$.

Remark 2. According to Lemma 1, $g(x)$ has the Hölder continuous gradient, so $g(x)$ admits (δ_0, L, μ_x) -model (δ was replaced by δ_0 to simplify notation) with

$$L = \tilde{L} \left(\frac{\tilde{L}}{2\delta_0} \frac{(1-\nu)(2-\nu)}{2-\nu} \right)^{\frac{(1-\nu)(1+\nu)}{2-\nu}},$$

where $\tilde{L} = \left(L_{xy} \left(\frac{2L_{xy}}{\mu_y} \right)^{\frac{\nu}{2-\nu}} + L_{xx} D^{\frac{\nu-\nu^2}{2-\nu}} \right)$.

Let us now assume, that instead of the ordinary gradient $\nabla g(x)$ we are given an inexact one $\tilde{\nabla} g(x) := \nabla_x f(x, \tilde{y})$, such that

$$\|y_1 - \tilde{y}\|_2 \leq \tilde{\Delta}, \quad (14)$$

where $f(x, y_1) = \max_{y \in Q_y} f, \tilde{\Delta} = \text{const} > 0$.

Theorem 1. Consider the strongly convex-concave saddle point problem (1) under assumptions (5)-(8). Define the function $g(x)$ according to (10). Then $g(x)$ admits an inexact (δ, L, μ_x) -model with $\delta = (D\tilde{\Delta} + \delta_0)$ and L defined in Remark 2. Applying k steps of the Fast Gradient Method to the "outer" problem (10) and

solving the "inner" problem (14) in linear time, we obtain an ε -solution (2) to the problem (1), where $\delta = \mathcal{O}(\varepsilon)$. The total number of iterations does not exceed

$$\mathcal{O} \left(\sqrt{\frac{L}{\mu_x}} \cdot \sqrt{\frac{L_{yy}}{\mu_y}} \cdot \log \frac{2L_{yy}R^2}{\varepsilon} \cdot \log \frac{2LD^2}{\varepsilon} \right),$$

$$\text{where } L = \tilde{L} \left(\frac{\tilde{L}}{2\varepsilon} \frac{(1-\nu)(2-\nu)}{2-\nu} \right)^{\frac{(1-\nu)(1+\nu)}{2-\nu}}, \tilde{L} = \left(L_{xy} \left(\frac{2L_{xy}}{\mu_y} \right)^{\frac{\nu}{2-\nu}} + L_{xx} D^{\frac{\nu-1}{2-\nu}} \right).$$

Proof. Since the problem $\max_{y \in Q_y} f(x_1, y)$ is smooth ((8) holds) and μ_y -strongly concave, one can achieve any arbitrary accuracy in (14), furthermore, in linear time [5]. More formally, for any $\tilde{\Delta} > 0$, in particular, $\tilde{\Delta} = \left(\frac{\Delta}{L_{xy}} \right)^{\frac{1}{\nu}}$, there exists \tilde{y} , such that

$$\|\tilde{y} - y_1\|_2 \leq \left(\frac{\Delta}{L_{xy}} \right)^{\frac{1}{\nu}}.$$

Then

$$\|\tilde{\nabla}g(x_1) - \nabla g(x_1)\|_2 = \|\nabla_x f(x_1, \tilde{y}) - \nabla_x f(x_1, y_1)\|_2 \leq L_{xy} \|\tilde{y} - y_1\|_2 \leq \Delta. \quad (15)$$

Thus, taking into account that

$$\langle \nabla g(x_1), x_2 - x_1 \rangle = \langle \nabla g(x_1) - \tilde{\nabla}g(x_1), x_2 - x_1 \rangle + \langle \tilde{\nabla}g(x_1), x_2 - x_1 \rangle,$$

we get:

$$\begin{aligned} & \frac{\mu_x}{2} \|x_2 - x_1\|_2^2 + \langle \tilde{\nabla}g(x_1), x_2 - x_1 \rangle + g(x_1) - D\Delta - \delta_0 \leq \\ & \leq g(x_2) \leq g(x_1) + \langle \tilde{\nabla}g(x_1), x_2 - x_1 \rangle + \frac{L}{2} \|x_2 - x_1\|_2^2 + D\Delta + \delta_0. \end{aligned} \quad (16)$$

The inequality (16) means, that $g(x)$ admits the inexact (δ, L, μ_x) -model with $\delta = (D\Delta + \delta_0)$ and L defined in Remark 2.

It is well known [5], that using Fast Gradient Method for the described construction, after k iterations, one can obtain the following accuracy of the solution:

$$g(x^k) - g(x^*) \leq LR^2 \exp \left(-\frac{k}{2} \sqrt{\frac{\mu_x}{L}} \right) + \delta \left(1 + \sqrt{\frac{L}{\mu_x}} \right).$$

As noted above, it is possible to achieve an arbitrarily small value of $\tilde{\Delta}$, more precisely, such $\tilde{\Delta}$, that

$$DL_{xy}\tilde{\Delta}^\nu + \delta_0 = D\Delta + \delta_0 = \delta \leq \frac{\varepsilon}{2 \left(1 + \sqrt{\frac{L}{\mu_x}} \right)}.$$

In other words, we can obtain the appropriate $\tilde{\Delta}$, such that δ will be comparable to ε . For example, we can put

$$\delta_0 \leq \frac{\varepsilon}{4 \left(1 + \sqrt{\frac{L}{\mu_x}}\right)}, \quad \Delta \leq \frac{\varepsilon}{4D \left(1 + \sqrt{\frac{L}{\mu_x}}\right)}.$$

Thus, after no more than $k = 2\sqrt{\frac{L}{\mu_x}} \log \frac{2LR^2}{\varepsilon}$ iterations we obtain an ε -solution (2) of the considered saddle point problem (1).

The total number of iterations (including achieving the appropriate precision of $\tilde{\Delta}$) is expressed as follows:

$$\mathcal{O} \left(\sqrt{\frac{L}{\mu_x}} \cdot \sqrt{\frac{L_{yy}}{\mu_y}} \cdot \log \frac{2L_{yy}R^2}{\varepsilon} \cdot \log \frac{2LD^2}{\varepsilon} \right),$$

$$\text{where } L = \tilde{L} \left(\frac{\tilde{L} (1-\nu)(2-\nu)}{2\varepsilon} \right)^{\frac{(1-\nu)(1+\nu)}{2-\nu}}, \quad \tilde{L} = \left(L_{xy} \left(\frac{2L_{xy}}{\mu_y} \right)^{\frac{\nu}{2-\nu}} + L_{xx} D^{\frac{\nu-\nu^2}{2-\nu}} \right).$$

2 Comparison of Theoretical Results for Accelerated Method and the Universal Proximal Method for Saddle-Point Problems with Generalized Smoothness

Let us investigate the effectiveness of the proposed method for solving strongly convex–concave saddle point problems in comparison with the Universal Algorithm, which was recently proposed in [14] to solve variational inequalities. We have to start with the problem statement and define all basics concerning variational inequalities and Proximal Setup. Note, that in Sect. 1 we used exclusively the Euclidean norm, while results of Sect. 2 and Sect. 3 hold for an arbitrary norm.

Let E be some finite-dimensional vector space, E^* be its dual. Let us choose some norm $\|\cdot\|$ on E . Define the dual norm $\|\cdot\|_*$ as follows:

$$\|\phi\|_* = \max_{\|x\| \leq 1} \{\langle \phi, x \rangle\},$$

where $\langle \phi, x \rangle$ denotes the value of the linear function $\phi \in E^*$ at the point $x \in E$.

Let $X \subset E$ be a closed convex set and $g(x) : X \rightarrow E^*$ be a monotone operator, i.e.

$$\langle g(x) - g(y), x - y \rangle \geq 0 \quad \forall x, y \in X. \quad (17)$$

We also need to choose a so-called prox-function $d(x)$, which is continuously differentiable and convex on X , and the corresponding Bregman divergence, which can be understood as an analogue of the distance, defined as follows:

$$V(y, x) = V_d(y, x) = d(y) - d(x) - \langle \nabla d(x), y - x \rangle, \quad \forall x, y \in X. \quad (18)$$

In [14], the authors proposed the Universal Proximal method (UMP, this method is listed as Algorithm 1 below) for solving the problem (3) with in- exactly given operator. In details, suppose that there exist some $\delta > 0$, $L(\delta) > 0$, such that, for any points $x, y, z \in X$, we are able to calculate $\tilde{g}(x, \delta), \tilde{g}(y, \delta) \in E^*$, satisfying

$$\langle \tilde{g}(y, \delta) - \tilde{g}(x, \delta), y - z \rangle \leq \frac{L(\delta)}{2} (\|y - x\|^2 + \|y - z\|^2) + \delta, \quad \forall z \in Q. \quad (19)$$

In addition, there was considered the possibility of using the restarted version of the method (**Restarted** UMP, see Algorithm 2 below) in the case of μ -strongly monotone operator g :

$$\langle g(x) - g(y), x - y \rangle \geq \mu \|x - y\|^2, \quad \forall x, y \in X.$$

We additionally assume that $\arg \min_{x \in X} d(x) = 0$ and $d(\cdot)$ is bounded on the unit ball in the chosen norm $\|\cdot\|$, more precisely

$$d(x) \leq \frac{\Omega}{2}, \quad \forall x \in X : \|x\| \leq 1,$$

where Ω is a known constant.

Algorithm 1 Universal Mirror Prox (UMP)

Require: $\varepsilon > 0$, $\delta > 0$, $x_0 \in X$, initial guess $L_0 > 0$, prox-setup: $d(x)$, $V(x, z)$.

1: Set $k = 0$, $z_0 = \arg \min_{u \in Q} d(u)$.

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

3: Set $M_k = L_k/2$.

4: Set $\delta = \frac{\varepsilon}{2}$.

5: **repeat**

6: Set $M_k = 2M_k$.

7: Calculate $\tilde{g}(z_k, \delta)$ and

$$w_k = \arg \min_{x \in Q} \{ \langle \tilde{g}(z_k, \delta), x \rangle + M_k V(x, z_k) \}. \quad (20)$$

8: Calculate $\tilde{g}(w_k, \delta)$ and

$$z_{k+1} = \arg \min_{x \in Q} \{ \langle \tilde{g}(w_k, \delta), x \rangle + M_k V(x, z_k) \}. \quad (21)$$

9: **until**

$$\langle \tilde{g}(w_k, \delta) - \tilde{g}(z_k, \delta), w_k - z_{k+1} \rangle \leq \frac{M_k}{2} (\|w_k - z_k\|^2 + \|w_k - z_{k+1}\|^2) + \frac{\varepsilon}{2} + \delta. \quad (22)$$

10: Set $L_{k+1} = M_k/2$, $k = k + 1$.

11: **end for**

Ensure: $z^k = \frac{1}{\sum_{i=0}^{k-1} M_i^{-1}} \sum_{i=0}^{k-1} M_i^{-1} w_i$.

Algorithm 2 Restarted Universal Mirror Prox (Restarted UMP).

Require: $\varepsilon > 0$, $\mu > 0$, $\Omega : d(x) \leq \frac{\Omega}{2} \forall x \in Q : \|x\| \leq 1$; $x_0, R_0 : \|x_0 - x_*\|^2 \leq R_0^2$.

1: Set $p = 0$, $d_0(x) = R_0^2 d\left(\frac{x-x_0}{R_0}\right)$.

2: **repeat**

3: Set x_{p+1} as the output of UMP for monotone case with prox-function $d_p(\cdot)$ and stopping criterion $\sum_{i=0}^{k-1} M_i^{-1} \geq \frac{\Omega}{\mu}$.

4: Set $R_{p+1}^2 = R_0^2 \cdot 2^{-(p+1)} + 2(1 - 2^{-(p+1)}) \frac{\varepsilon}{4\mu}$.

5: Set $d_{p+1}(x) \leftarrow R_{p+1}^2 d\left(\frac{x-x_{p+1}}{R_{p+1}}\right)$.

6: Set $p = p + 1$.

7: **until** $p > \log_2\left(\frac{2R_0^2}{\varepsilon}\right)$.

Ensure: x_p .

Remark 3. (Connection between saddle point problem and VI). Any saddle point problem

$$\min_{x \in Q_x} \max_{y \in Q_y} f$$

can be reduced to a variational inequality problem by considering the following operator:

$$g(z) = \begin{pmatrix} \nabla_x f \\ -\nabla_y f \end{pmatrix}, \quad z = (x, y) \in Q := Q_x \times Q_y. \quad (23)$$

Remark 4. Restarted UMP method returns a point x_p such that

$$\|x_p - x_*\|^2 \leq \varepsilon + \frac{2\delta}{\mu}.$$

The total number of iterations of the inner Algorithm UMP does not exceed

$$\inf_{\nu \in [0,1]} \left[\left(\frac{L_\nu}{\mu} \right)^{\frac{2}{1+\nu}} \cdot \frac{2^{\frac{2}{1+\nu}} \Omega}{\varepsilon^{\frac{1-\nu}{1+\nu}}} \cdot \log_2 \frac{2R_0^2}{\varepsilon} \right].$$

Remark 5. Note, that for $\nu = 0$ the convergence rate of the Restarted UMP and the accelerated method, introduced in Sect. 1, coincides, while for $\nu > 0$ the asymptotic of the proposed accelerated method is better.

2.1 Numerical Experiments for Restarted Universal Mirror Prox Method

The problem of constrained minimization of convex functionals arises and attracts widespread interest in many areas of modern large-scale optimization and its applications [3,12].

In this subsection, in order to demonstrate the performance of the Restarted UMP, we consider an example of the Lagrange saddle point problem induced by a problem with geometrical nature, namely, an analogue of the well-known *smallest covering ball problem* with non-smooth functional constraints. This example

is equivalent to the following non-smooth convex optimization problem with functional constraints

$$\min_{x \in Q} \left\{ f(x) := \max_{1 \leq k \leq N} \|x - A_k\|_2^2; \varphi_p(x) \leq 0, p = 1, \dots, m \right\}, \quad (24)$$

where $A_k \in \mathbb{R}^n, k = 1, \dots, N$ are given points and Q is a convex compact set. Functional constraints φ_p , for $p = 1, \dots, m$, have the following form:

$$\varphi_p(x) := \sum_{i=1}^n \alpha_{pi} x_i^2 - 5, p = 1, \dots, m. \quad (25)$$

For solving such problems there are various first-order methods, which guarantee achieving an acceptable precision ε by function with complexity $\mathcal{O}(\varepsilon^{-1})$. We present the results of some numerical experiments, which demonstrate the effectiveness of the **Restarted UMP**. Remind, that we analyze how the restart technique can be used to improve the convergence rate of the **UMP**, and show that in practice it works with a convergence rate smaller than $\mathcal{O}(\varepsilon^{-1})$, for some different randomly generated data associated with functional constraints (25).

The corresponding Lagrange saddle point problem of the problem (24) is defined as follows

$$\min_{x \in Q} \max_{\vec{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_m)^T \in \mathbb{R}_+^m} L(x, \lambda) := f(x) + \sum_{p=1}^m \lambda_p \varphi_p(x) - \frac{1}{2} \sum_{p=1}^m \lambda_p^2.$$

This problem is satisfied to (5) – (8) for $\nu = 0$ and equivalent to the variational inequality with the monotone bounded operator

$$G(x, \lambda) = \left(\begin{array}{c} \nabla f(x) + \sum_{p=1}^m \lambda_p \nabla \varphi_p(x), \\ (-\varphi_1(x) + \lambda_1, -\varphi_2(x) + \lambda_2, \dots, -\varphi_m(x) + \lambda_m)^T \end{array} \right),$$

where ∇f and $\nabla \varphi_p$ are subgradients of f and φ_p . For simplicity, let us assume that there exists (potentially very large) bound for the optimal Lagrange multiplier $\vec{\lambda}^*$. Thus, we are able to compactify the feasible set for the pair $(x, \vec{\lambda})$ to be an Euclidean ball of some radius.

To demonstrate the independence on the choice of experimental data, the coefficients α_{pi} in (25) are drawn randomly from four different distributions.

- **Case 1:** the standard exponential distribution.
- **Case 2:** the Gumbel distribution with mode and scale equal to zero and 1, respectively.
- **Case 3:** the inverse Gaussian distribution with mean and scale equal to 1 and 2, respectively.
- **Case 4:** from the discrete uniform distribution in the half open interval $[1, 6)$.

We run **Restarted UMP** for different values of n and m with standard Euclidean prox-structure and the starting point $(x^0, \vec{\lambda}^0) = \frac{1}{\sqrt{m+n}} \mathbf{1} \in \mathbb{R}^{n+m}$, where $\mathbf{1}$ is the vector of all ones. Points A_k , $k = 1, \dots, N$, are chosen randomly from the uniform distribution over $[0, 1)$. For each value of the parameters the random data was drawn 5 times and the results were averaged. The results of the work of **Restarted UMP** are presented in Table 1. These results demonstrate the number of iterations produced by **Restarted UMP** to reach the ε -solution of the problem (24) with (25), the running time of the algorithm in seconds, qualities of the solution with respect to the objective function f ($f^{\text{best}} := f(x_{\text{out}})$) and the functional constraints g ($g^{\text{out}} := g(x_{\text{out}})$), where x_{out} denotes the output of the compared algorithms, with different values of $\varepsilon \in \{1/2^i, i = 1, 2, 3, 4, 5, 6\}$.

All experiments were implemented in Python 3.4, on a computer fitted with Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz, 1992 Mhz, 4 Core(s), 8 Logical Processor(s). RAM of the computer is 8 GB.

Table 1: The results of **Restarted UMP**, for the problem (24) with constraints (25) for **Cases 1 – 4**.

$\frac{1}{\varepsilon}$	Case 1: $n = 1000, m = 50, N = 10$				Case 2: $n = 1000, m = 50, N = 10$			
	Iter.	Time (sec.)	f^{best}	g^{out}	Iter.	Time (sec.)	f^{best}	g^{out}
2	9	0.392	324.066325	-4.349744	12	0.517	324.889649	-3.965223
4	12	0.519	324.066312	-4.349732	16	0.630	324.889634	-3.965213
8	15	0.599	324.066305	-4.349692	20	0.807	324.889621	-3.965197
16	18	0.972	324.066295	-4.349680	24	0.977	324.889598	-3.965165
32	21	0.984	324.066291	-4.349653	28	1.224	324.889540	-3.965223
64	24	1.357	324.066286	-4.349598	32	1.317	324.889527	-3.965111
	Case 3: $n = 500, m = 25, N = 10$				Case 4: $n = 500, m = 25, N = 10$			
2	624	20.321	153.846214	-4.398782	832	26.629	158.210875	-2.610276
4	1319	39.045	153.842306	-4.398106	1702	51.258	158.201645	-2.601072
8	2158	58.919	153.830012	-4.397387	4144	96.136	158.190455	-2.599223
16	4298	117.383	153.827731	-4.397271	6145	174.713	158.188211	-2.598865
32	8523	264.777	153.826829	-4.397226	12081	351.52	158.187255	-2.598713
64	17584	554.480	153.826382	-4.397204	30186	768.861	158.186744	-2.598628

From the results in Table 1 we can see that the work of **Restarted UMP**, does not only depend on the dimension n of the problem, the number of the constraints φ_p , $p = 1, \dots, m$ and the number of the points A_k , $k = 1, \dots, N$, but also depends on the shape of the data, that generated to the coefficients α_{pi} in (25). Also, these results demonstrate how the **Restarted UMP**, due to its adaptivity to the level of smoothness of the problem, in practice works with a convergence rate smaller than $\mathcal{O}(\varepsilon^{-1})$, for all different cases 1–4.

3 Mirror Descent for Variational Inequalities with Relatively Bounded Operator

Let us remind the definition of the boundedness of an operator.

Definition 2. (*The classical boundedness*). Let $g(x) : X \rightarrow E^*$, be an operator, given on some convex set $X \subset E$. $g(x)$ is bounded on X , if there exists $M > 0$, such that

$$\|g(x)\|_* \leq M, \quad \forall x \in X.$$

We can replace the classical concept of the boundedness of an operator by the so-called Relative boundedness condition as following.

Definition 3. (*The Relative boundedness*). An operator $g(x) : X \rightarrow E^*$ is Relatively bounded on X , if there exists $M > 0$, such that

$$\langle g(x), y - x \rangle \leq M \sqrt{2V(y, x)}, \quad \forall x, y \in X, \quad (26)$$

where $V(y, x)$ is the Bregmann divergence, defined in (18).

Let us note the following special case of the definition.

Remark 6. The Relative boundedness condition can be rewritten in the following way:

$$\|g(x)\|_* \leq \frac{M \sqrt{2V(y, x)}}{\|y - x\|}, \quad y \neq x.$$

In addition to the Relative boundedness condition, suppose that the operator $g(x)$ is σ -monotone.

Definition 4. (*σ -monotonicity*). Let $\sigma > 0$. The operator $g(x) : X \rightarrow E^*$ is σ -monotone, if the following inequality holds:

$$\langle g(y) - g(x), y - x \rangle \geq -\sigma, \quad \forall x, y \in X. \quad (27)$$

Let us propose an analogue of the Mirror Descent algorithm for variational inequalities with Relatively bounded and σ -monotone operator.

For any $x \in X$ and $p \in E^*$, we define the Mirror Descent step $Mirr_x(p)$ as follows:

$$Mirr_x(p) = \arg \min_{y \in X} \left\{ \langle p, y \rangle + V(y, x) \right\}.$$

The following theorem describes the effectiveness of the proposed Algorithm 3.

Theorem 2. Let $g : X \rightarrow E^*$ be Relatively bounded and σ -monotone operator, i.e. (26) and (27) hold. Then after no more than

$$N = \frac{2RM^2}{\varepsilon^2}$$

iterations of Algorithm 3, one can obtain an $(\varepsilon + \sigma)$ -solution (4) of the problem (3).

Algorithm 3 Mirror Descent method for variational inequalities.

Require: $\varepsilon > 0, \sigma > 0, M > 0; x_0$ and R such that $V(x_*, x_0) \leq R^2$.

1: Set $h = \frac{\varepsilon}{M^2}$.

2: Initialization $k = 0$.

3: **repeat**

4: $x^{k+1} = \text{Mirr}_{x^k}(hg(x^k))$.

5: Set $k = k + 1$.

6: **until** $k \geq N = \frac{2RM^2}{\varepsilon}$.

Ensure: $\tilde{x} = \frac{1}{N} \sum_{k=0}^{N-1} x^k$.

Proof. According to the listing of the Algorithm 3, the following inequality holds:

$$h\langle g(x^k), x^k - x \rangle \leq \frac{h^2 M^2}{2} + V(x, x^k) - V(x, x^{k+1}).$$

Taking summation over $k = 0, 1, \dots, N - 1$, we get

$$\begin{aligned} h \sum_{k=0}^{N-1} \langle g(x^k), x^k - x \rangle &\leq \sum_{k=0}^{N-1} \frac{h^2 M^2}{2} + V(x, x^0) - V(x, x^N) \leq \\ &\leq \frac{N}{2} h^2 M^2 + V(x, x^0). \end{aligned}$$

Due to the σ -monotonicity of $g(x)$, we have

$$\langle g(x^k), x^k - x \rangle \geq \langle g(x), x^k - x \rangle - \sigma.$$

Whence,

$$h \sum_{k=0}^{N-1} (\langle g(x^k), x^k - x \rangle + \sigma) \geq h \left\langle g(x), \sum_{k=0}^{N-1} (x^k - x) \right\rangle = h \langle g(x), N(\tilde{x} - x) \rangle,$$

where $\tilde{x} = \frac{1}{N} \sum_{k=0}^{N-1} x^k$. Since

$$Nh \langle g(x), \tilde{x} - x \rangle \leq \frac{N}{2} h^2 M^2 + V(x, x^0) + Nh\sigma,$$

we get

$$\langle g(x), \tilde{x} - x \rangle \leq \frac{M^2 h}{2} + \frac{V(x, x^0)}{Nh} + \sigma.$$

As $h = \frac{\varepsilon}{M^2}$, after the stopping criterion $N \geq \frac{2RM^2}{\varepsilon}$ is satisfied, we obtain an $(\varepsilon + \sigma)$ -solution of the problem (3), i.e.

$$\max_{x \in X} \langle g(x), \tilde{x} - x \rangle \leq \varepsilon + \sigma.$$

4 Conclusions

In the first part of the article we considered strongly convex-concave saddle point problems; using the concept of (δ, L, μ_x) -model we applied the Fast Gradient Method to obtain an ε -solution of the problem. We proved, that the proposed method can generate an ε -solution of the saddle point problem after no more than

$$\mathcal{O} \left(\sqrt{\frac{L}{\mu_x}} \cdot \sqrt{\frac{L_{yy}}{\mu_y}} \cdot \log \frac{2L_{yy}R^2}{\varepsilon} \cdot \log \frac{2LD^2}{\varepsilon} \right) \text{ iterations,}$$

where $L = \tilde{L} \left(\frac{\tilde{L}}{2\varepsilon} \frac{(1-\nu)(2-\nu)}{2-\nu} \right)^{\frac{(1-\nu)(1+\nu)}{2-\nu}}$, $\tilde{L} = \left(L_{xy} \left(\frac{2L_{xy}}{\mu_y} \right)^{\frac{\nu}{2-\nu}} + L_{xx}D^{\frac{\nu-\nu^2}{2-\nu}} \right)$.

We conducted some numerical experiments for the Universal Proximal algorithm and its restarted version and analyzed their asymptotics in comparison with the proposed method.

Also, we considered Minty variational inequalities with Relatively bounded and σ -monotone operator. We introduced the modification of the Mirror Descent method and proved, that obtaining an $(\varepsilon + \sigma)$ -solution will take no more than $\frac{2RM^2}{\varepsilon^2}$ iterations.

References

1. Alkousa, M.S., Gasnikov, A.V., Dvinskikh, D.M., Kovalev, D.A., Stonyakin, F.S.: Accelerated methods for saddle-point problems. *Comput. Math. and Math. Phys.*, **60**(11), 1843–1866 (2020)
2. Antonakopoulos, K., Belmega, V., Mertikopoulos, P.: An adaptive Mirror-Prox method for variational inequalities with singular operators. *Advances in Neural Information Processing Systems*. 8455–8465 (2019)
3. Ben-Tal, A., Nemirovski, A.: Robust Truss Topology Design via Semidefinite Programming. *SIAM J. Optim.* **7**(4), 991–1016 (1997)
4. Benzi, M., Golub, G. H., Liesen, J.: Numerical solution of saddle point problems. *Acta numerica*, **14**(1), 1–137 (2005)
5. Gasnikov, A.V.: Modern numerical optimization methods. Method of universal gradient descent. *MCCME*, (2020)
6. Juditsky, A., Nemirovski, A., Tauvel, C.: Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, **1**(1), 17–58 (2011)
7. Lu, H.: Relative Continuity for Non-Lipschitz Nonsmooth Convex Optimization Using Stochastic (or Deterministic) Mirror Descent. *Inform. Journal on Optimization*, **1**(4), 288–303 (2019)
8. 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 Optimization*, **15**(1), 229–251 (2004)
9. Nesterov, Yu.: Universal gradient methods for convex optimization problems. *Math. Program.*, **152**(1), 381–404 (2015)
10. Nesterov, Yu.: Relative Smoothness: New Paradigm in Convex Optimization. Conference report, EUSIPCO-2019, A Coruna, Spain, September 4, 2019. <http://eusipco2019.org/wp-content/uploads/2019/10/Relative-Smoothness-New-Paradigm-in-Convex.pdf>

11. Polyak, B.T.: Introduction to optimization. Optimization Software, Inc, New York (1987)
12. Shpirko, S., Nesterov, Yu.: Primal-dual subgradient methods for huge-scale linear conic problem. *SIAM J. Optim.* **24**(3), 1444–1457 (2014).
13. Scutari, G., Palomar, D. P., Facchinei, F., Pang, J. S. Convex optimization, game theory, and variational inequality theory. *IEEE Signal Processing Magazine*, 27(3), 35-49 (2010)
14. Stonyakin, F., Gasnikov, A., Dvurechensky, P., Alkousa, M., Titov, A.: Generalized mirror prox for monotone variational inequalities: Universality and inexact oracle (2018). <https://arxiv.org/pdf/1806.05140.pdf>
15. Titov, A.A., Stonyakin, F.S., Alkousa, M.S., Ablayev, S.S., Gasnikov, A.V.: Analogues of Switching Subgradient Schemes for Relatively Lipschitz-Continuous Convex Programming Problems. In: Kochetov Y., Bykadorov I., Gruzdeva T. (eds) *Mathematical Optimization Theory and Operations Research. MOTOR 2020. Communications in Computer and Information Science, Springer, Cham.* **1275**, 133–149 (2020)