Generalized Mirror Prox Algorithm for Monotone Variational Inequalities: Universality and Inexact Oracle [1]

Fedor Stonyakin [2] Alexander Gasnikov [3] Pavel Dvurechensky [4]
Alexander Titov [5] Mohammad S. Alkousa [6]

February 18, 2022

Abstract

We introduce an inexact oracle model for variational inequalities with monotone operator, propose a numerical method, which solves such variational inequalities, and analyze its convergence rate. As a particular case, we consider variational inequalities with Hölder-continuous operator and show that our algorithm is universal. This means that, without knowing the Hölder exponent and Hölder constant, the algorithm has the least possible in the worst-case sense complexity for this class of variational inequalities. We also consider the case of variational inequalities with strongly monotone operator and generalize the algorithm for variational inequalities with inexact oracle and our universal method for this class of problems. Finally, we show, how our method can be applied to convex-concave saddle point problems with Hölder-continuous partial subgradients.

Keywords: Variational inequality, monotone operator, Hölder continuity, inexact oracle, complexity estimate

AMS: 65K15, 90C33, 90C06, 68Q25, 65Y20, 68W40, 58E35

1 Introduction

This paper is devoted to Minty [31] (or weak [48]) variational inequalities with a monotone and continuous operator. Variational inequalities with monotone operators are closely con-

[1]Submitted to the editors DATE.

[2]V.I. Vernadsky Crimean Federal University, Simferopol; Moscow Institute of Physics and Technology, Moscow (fedyor@mail.ru).

[3]Moscow Institute of Physics and Technology, Moscow; Institute for Information Transmission Problems RAS, Moscow (gasnikov@yandex.ru).

[4]Weierstrass Institute for Applied Analysis and Stochastics, Berlin (pavel.dvurechensky@wias-berlin.de).

[5]Moscow Institute of Physics and Technology, Moscow (a.a.titov@phystech.edu).

[6]Moscow Institute of Physics and Technology, Moscow (mohammad.alkousa@phystech.edu).

nected with convex optimization problems and convex-concave saddle point problems. In the former case, the operator is just the subgradient of the objective function, and in the latter case the operator is composed from partial subgradients of the objective in the saddle point problem. Studying variational inequalities is important also for equilibrium and complementarity problems [37, 24] and saddle point problems has become an important part of research in machine learning [3, 41].

Our focus here is on numerical methods for such problems, their convergence rate and complexity estimates. Significant contribution to the development of numerical methods for solving variational inequalities was made in 1970's, when the extragradient method was proposed in [42]. More recently, [45] proposed a non-Euclidean variant of this method, called Mirror Prox algorithm, which can be applied for Lipschitz continuous operators.

Different methods with similar complexity were also proposed in [59, 4, 48, 44, 43, 28]. Besides that, in [48], Nesterov proposed a method for variational inequalities with bounded variation of the operator, i.e. with non-smooth operator. He raised also a question, whether it is possible to propose a method, which automatically "adjusts to the actual level of smoothness of the current problem instance". One of the goals of this paper is to propose such an algorithm.

To this aim, we consider a more general class of operators being so-called Hölder-continuous. This class covers both the case of operators with bounded variation and Lipschitz-continuous operators. Variational inequalities with
Hölder-continuous monotone operator were already considered in [45], where a special choice of the stepsize for the Mirror Prox algorithm led to the optimal complexity for this class of problems; see [46]. The authors of [14] consider variational inequalities with non-monotone Hölder-continuous operator. Unfortunately, both papers use Hölder constant and exponent to define the stepsize of their methods. This is in contrast to optimization, where so called universal algorithms were proposed, which do not use the information about the Hölder exponent and Hölder constant; see [49, 30, 29, 7, 19, 39, 35, 36, 51]. In this paper, we propose a universal method for variational inequalities with Hölder-continuous monotone operator. We also generalize this method for the case of variational inequalities with strongly monotone operator. Such problems were considered in [52], but only for the case of Lipschitz-continuous operator with known Lipschitz constant.

On the other hand, as it was shown for optimization problems in [16, 49], universal methods have a natural connection with methods for smooth problems with inexact oracle. Namely, it can be shown that a function with Hölder-continuous subgradient can be considered as a Lipschitz-smooth function with inexact oracle. Despite that there are many works on optimization methods with inexact oracle, see e.g. [15, 16, 20, 27, 11, 39, 13, 5, 61], we are not aware of any extensions of these non-stochastic definitions of inexact oracle to the variational inequality setting and methods for variational inequalities with inexactly given operator, except stochastic case. By this paper, we introduce a theory of methods for variational inequalities with deterministic inexact oracle, see also the follow-up work [60].

2 Preliminaries

We start with the general notations, problem statement, and description of proximal setup. Let E be a finite-dimensional real vector space and E^* be its dual. We denote the value of a linear function $u \in E^*$ at $x \in E$ by $\langle u, x\rangle$. Let $\|\cdot\|$ be a general norm on E, $\|\cdot\|_*$ be its dual, defined by
$\|u\|_* := \max\limits_x \{\langle u, x\rangle, \|x\| \leq 1\}$. We use $\nabla f(x)$ to denote any subgradient of a function f at a point $x \in \mathrm{dom} f$.

The main problem, we consider, is the following Minty variational inequality (VI)

$$\text{Find} \quad x_* \in Q: \quad \langle g(x), x_* - x\rangle \leq 0, \quad \forall x \in Q \tag{1}$$

where Q is convex (non-necessary compact) subset of finite-dimensional vector space E, $g : Q \to E^*$ is continuous, monotone operator

$$\langle g(x) - g(y), x - y\rangle \geq 0, \quad x, y \in Q,$$

satisfying Hölder condition on Q, i.e., for some $\nu \in [0, 1]$ and $L_\nu \geq 0$,

$$\|g(x) - g(y)\|_* \leq L_\nu \|x - y\|^\nu, \quad x, y \in Q. \tag{2}$$

We refer to ν as Hölder exponent and to L_ν as Hölder constant. We assume that the variational inequality (1) has a solution. Under the assumption of continuity and monotonicity of the operator g, the problem (1) is equivalent to a Stampacchia [31] (or strong [48]) variational inequality, in which the goal is to find $x_* \in Q$ such that

$$\langle g(x_*), x_* - x\rangle \leq 0, \quad \forall x \in Q. \tag{3}$$

Following [48, 2], to assess the quality of a candidate solution $\widehat{x}$, we use a convex non-empty compact subset C of the set Q and the following restricted gap (or merit) function

$$\mathrm{Gap}_C(\widehat{x}) = \max_{u \in C} \langle g(u), \widehat{x} - u\rangle. \tag{4}$$

Proposition 1 in [2] states that $\mathrm{Gap}_C(\widehat{x}) \geq 0$ whenever $\widehat{x} \in C$ and if $\mathrm{Gap}_C(\widehat{x}) = 0$ and C contains a neighborhood of $\widehat{x}$, then $\widehat{x}$ is a solution of (3). This motivates our goal that is to find an approximate solution of the problem, that is, a point $\widehat{x} \in Q$ such that, for some $\varepsilon > 0$

$$\mathrm{Gap}_C(\widehat{x}) = \max_{u \in C} \langle g(u), \widehat{x} - u\rangle \leq \varepsilon. \tag{5}$$

As already mentioned, [45] proposed Mirror Prox algorithm under the assumption of compactness of Q and L_1-Lipschitz continuity of the operator, i.e., g satisfying (2) with $\nu = 1$ and L_1. This method has complexity $O\left(\frac{L_1 R^2}{\varepsilon}\right)$, where R characterizes the diameter of the set Q and ε is the desired accuracy. By complexity we mean the number of iterations of an algorithm to find a point $\widehat{x} \in Q$ such that (5) holds. For the case of variational inequalities with bounded variation of the operator g, i.e., g satisfying (2) with $\nu = 0$ and L_0

[48] proposed a method with complexity $O\left(\frac{L_0^2 R^2}{\varepsilon^2}\right)$. The method for variational inequalities with Hölder-continuous monotone operator [45] has the complexity

$$O\left(\left(\frac{L_\nu}{\varepsilon}\right)^{\frac{2}{1+\nu}} R^2\right),$$

which is optimal for the case of $\nu = 1$ and for the case of $\nu = 0$ [54, 46].

Next we give several definitions, which are necessary for introducing the method. We choose a *prox-function* $d(x)$, which is continuous and convex on Q, and also is

1. continuously differentiable at the relative interior of Q;
2. 1-strongly convex on Q with respect to $\|\cdot\|$, i.e., for any $x \in Q^0, y \in Q$ $d(y) - d(x) - \langle \nabla d(x), y - x\rangle \geq \frac{1}{2}\|y - x\|^2$.

Without loss of generality, we assume that $\min\limits_{x\in Q} d(x) = 0$.

We define also the corresponding *Bregman divergence*

$$V[z](x) := d(x) - d(z) - \langle \nabla d(z), x - z\rangle,\ x \in Q, z \in Q^0.$$

Standard proximal setups, i.e., Euclidean, entropy, ℓ_1/ℓ_2, simplex, nuclear norm, spectahedron can be found in [9]. Below we use Bregman divergence in so-called *prox-mapping*

$$\min_{x\in Q} \{\langle g, x\rangle + MV[\bar{x}](x)\}, \tag{6}$$

where $M > 0$, $\bar{x} \in Q^0$, $g \in E^*$ are given. We allow this problem to be solved inexactly in the following sense inspired by [9]. Assume that we are given $\delta_{pu} > 0$, $M > 0$, $\bar{x} \in Q^0$, $g \in E^*$. Further, we assume that for an arbitrary $\delta_{pc} > 0$, we can calculate $\tilde{x} = \tilde{x}(\bar{x}, g, M, \delta_{pc}, \delta_{pu}) \in Q^0$ such that

$$\langle g + M\left[\nabla d(\tilde{x}) - \nabla d(\bar{x})\right], u - \tilde{x}\rangle \geq -\delta_{pc} - \delta_{pu}, \quad \forall u \in Q. \tag{7}$$

We call the point $\tilde{x}$ an *inexact prox-mapping* and write

$$\tilde{x} := \arg\min_{x\in Q}{}^{\delta_{pc}+\delta_{pu}} \{\langle g, x\rangle + MV[\bar{x}](x)\}. \tag{8}$$

Here δ_{pu} denotes the error of the prox-mapping, which is not controlled, and δ_{pc} denotes the error of the prox-mapping, which can be controlled and made as small as desired.

3 Inexact Oracle for Variational Inequalities

Our goal is to consider, in a unified manner, VIs with Hölder-continuous operator and VIs with inexact values of the operator. This can be done by considering Hölder-continuous operator as a particular case of Lipschitz-continuous operator with some inexactness. Thus, we introduce the following definition of inexact oracle for the operator g.

Definition 1. *Assume that for some $\delta_u > 0$ (uncontrolled error) and for any number $\delta_c > 0$ (controlled error) there exists a constant $L(\delta_c) \in]0, +\infty[$ such that, for any points $x, y \in Q$, one can calculate $\tilde{g}(x, \delta_c, \delta_u)$ and*
$\tilde{g}(y, \delta_c, \delta_u) \in E^*$ satisfying

$$\langle \tilde{g}(y, \delta_c, \delta_u) - \tilde{g}(x, \delta_c, \delta_u), y - z \rangle \leq \frac{L(\delta_c)}{2} \left(\|y - x\|^2 + \|y - z\|^2 \right) + \delta_c + \delta_u, \tag{9}$$

$$\langle \tilde{g}(y, \delta_c, \delta_u) - g(y), y - z \rangle \geq -\delta_u, \quad \forall z \in Q. \tag{10}$$

Then, the operator $\tilde{g}(\cdot, \delta_c, \delta_u)$ is called inexact oracle for the operator g.

In this definition, δ_c represents the error of the oracle, which we can control and make as small as we would like to. On the opposite, δ_u represents the uncontrolled error, which can be understood as an error in the problem data, for example, when g is given as a solution to an auxiliary problem. We notice also that if the inequality (9) holds for some $L(\delta_c)$, then it holds also for any $\widetilde{L}(\delta_c) \geq L(\delta_c)$.

Example 1 below shows that this definition satisfies our goal of covering both the case of Hölder-continuous operator and the case of inexact values of the operator. The following technical lemma is the main clue for this example.

Lemma 1. *Let $a, b, c \geq 0$, $\nu \in [0, 1]$. Then, for any $\delta > 0$,*

$$ab^\nu c \leq \left(\frac{1}{\delta} \right)^{\frac{1-\nu}{1+\nu}} \frac{a^{\frac{2}{1+\nu}}}{2} \left(b^2 + c^2 \right) + \frac{\delta}{2}.$$

The proof of this lemma is given in the Appendix A.

Example 1. *(Hölder-continuous operator with inexact values on a bounded set). Let us assume that:*

1. *The operator $g(x)$ is Hölder-continuous on Q, i.e., satisfies (2).*
2. *The set Q is bounded with $\max_{x,y \in Q} \|x - y\| \leq D$.*
3. *There exist $\bar{\delta}_u > 0$ and at any point $x \in Q$, we can calculate approximation $\bar{g}(x)$ for $g(x)$ such that $\|\bar{g}(x) - g(x)\|_* \leq \bar{\delta}_u$.*

Then, for any $z \in Q$,

$$\begin{aligned}
\langle \bar{g}(y) - \bar{g}(x), y - z \rangle &= \langle \bar{g}(y) - g(y), y - z \rangle - \langle \bar{g}(x) - g(x), y - z \rangle + \langle g(y) - g(x), y - z \rangle \\
&\leq 2\bar{\delta}_u D + \|g(y) - g(x)\|_* \|y - z\| \leq 2\bar{\delta}_u D + L_\nu \|y - x\|^\nu \|y - z\| \\
&\leq 2\bar{\delta}_u D + \frac{1}{2} \left(\frac{1}{\delta} \right)^{\frac{1-\nu}{1+\nu}} L_\nu^{\frac{2}{1+\nu}} \left(\|x - y\|^2 + \|y - z\|^2 \right) + \frac{\delta}{2},
\end{aligned}$$

where Lemma 1 was used to get the last inequality.
Thus, we can set $\delta_u = 2\bar{\delta}_u D$, $\delta_c = \frac{\delta}{2}$, *and* $L(\delta_c) = \left(\frac{1}{2\delta_c}\right)^{\frac{1-\nu}{1+\nu}} L_\nu^{\frac{2}{1+\nu}}$ *to get* (9). *Further,*

$$|\langle \overline{g}(y) - g(y), y - z\rangle| \le \|\overline{g}(y) - g(y)\|_* \|y - z\| \le \overline{\delta}_u D = \frac{1}{2}\delta_u < \delta_u$$

and we have $\langle \overline{g}(y) - g(y), y - z\rangle > -\delta_u$, *which is* (10).

Example 2. *(Connection with* (δ, L)*-oracle in optimization). Let convex function* $f : Q \to \mathbb{R}$, *where* Q *is a convex compact, be endowed with* (δ, L)*-oracle [16]. This means that for some* $L > 0$ *at any* $y \in Q$ *there exists a pair* $(f_\delta(y), g_\delta(y)) \in \mathbb{R} \times \mathbb{R}^n$ *such that, for all* $y \in Q$,

$$f_\delta(y) + \langle g_\delta(y), x - y\rangle \le f(x) \le f_\delta(y) + \langle g_\delta(y), x - y\rangle + \frac{L\|x - y\|^2}{2} + \delta. \tag{11}$$

Let $g(y)$ *be any selector of the exact subgradients of* f. *Under an additional assumption that* $\|g_\delta(y) - g(y)\| \le \bar{\delta}_u$, *we show that* $g_\delta(y)$ *is an inexact oracle for the operator* $g(y)$. *This is similar to the exact case, when the subgradient of a convex function defines a monotone operator. It is easy to show* (10). *Indeed,*

$$\langle g_\delta(y) - g(y), y - x\rangle \ge -\bar{\delta}_u \|x - y\| \ge -\bar{\delta}_u D, \quad \forall x, y \in Q. \tag{12}$$

Let us show that (9) *is satisfied with* $L(\delta_c) = L$, $\delta_c = 0$,
$\delta_u = \max\{2\delta, \bar{\delta}_u D\}$, *and* $\tilde{g}(y, \delta_c, \delta_u) = g_\delta(y)$. *Indeed,*

$$\begin{aligned}
\langle g_\delta(z) - g_\delta(y), z - x\rangle &= \langle g_\delta(y) - g_\delta(z), x - z\rangle \\
&= \langle g_\delta(y), x - y\rangle - \langle g_\delta(z), x - z\rangle - \langle g_\delta(y), z - y\rangle \\
&= (f(x) - f_\delta(z) - \langle g_\delta(z), x - z\rangle) + (f(z) - f_\delta(y) - \langle g_\delta(y), z - y\rangle) - \\
&\qquad - (f(x) - f_\delta(y) - \langle g_\delta(y), x - y\rangle) + (f_\delta(z) - f(z)) \\
&\le \left(\frac{L}{2}\|x - z\|^2 + \delta\right) + \left(\frac{L}{2}\|z - y\|^2 + \delta\right),
\end{aligned}$$

where in the last inequality we used twice the right inequality in (11), *and twice the left inequality in* (11).

4 Generalized Mirror Prox

In this section, we introduce a new algorithm, which we call Generalized Mirror Prox (GMP), for problem (1) with inexact oracle for g in the sense of Definition 1. The algorithm is listed below as Algorithm 1.

Theorem 1. *Assume that* $g(\cdot)$ *and* $\tilde{g}(\cdot, \delta_c, \delta_u)$ *satisfy* (9) *and* (10). *Then, for any* $k \ge 1$ *and any* $u \in Q$,

$$\frac{1}{\sum_{i=0}^{k-1} M_i^{-1}} \sum_{i=0}^{k-1} M_i^{-1} \langle g(w_i), w_i - u\rangle \le \frac{1}{\sum_{i=0}^{k-1} M_i^{-1}} (V[z_0](u) - V[z_k](u)) + \frac{\varepsilon}{2} + \delta_u + 2\delta_{pu}.$$

Algorithm 1 Generalized Mirror Prox

Input: accuracy $\varepsilon > 0$, level $\delta_u > 0$ of the uncontrolled oracle error, level $\delta_{pu} > 0$ of the uncontrolled error of prox-mapping, initial guess M_{-1} for $L(\delta_c)$, prox-setup: $d(x)$, $V[z](x)$.

1: Set $k = 0$, $z_0 = \arg\min_{u \in Q} d(u)$.
2: **for** $k = 0, 1, ...$ **do**
3: Set $i_k = 0$, $\delta_{c,k} = \frac{\varepsilon}{4}$, $\delta_{pc,k} = \frac{\varepsilon}{8}$.
4: **repeat**
5: Set $M_k = 2^{i_k - 1} M_{k-1}$.
6: Calculate

$$w_k = \arg\min_{x \in Q}^{\delta_{pc,k} + \delta_{pu}} \left\{ \langle \tilde{g}(z_k, \delta_{c,k}, \delta_u), x \rangle + M_k V[z_k](x) \right\}. \tag{13}$$

$$z_{k+1} = \arg\min_{x \in Q}^{\delta_{pc,k} + \delta_{pu}} \left\{ \langle \tilde{g}(w_k, \delta_{c,k}, \delta_u), x \rangle + M_k V[z_k](x) \right\}. \tag{14}$$

7: $i_k = i_k + 1$.
8: **until**

$$\langle \tilde{g}(w_k, \delta_c, \delta_u) - \tilde{g}(z_k, \delta_c, \delta_u), w_k - z_{k+1} \rangle \leq \frac{M_k}{2} \left(\|w_k - z_k\|^2 + \|w_k - z_{k+1}\|^2 \right) + \delta_{c,k} + \delta_u. \tag{15}$$

9: Set $k = k + 1$.
10: **end for**

Output: $\widehat{w}_k = \frac{1}{\sum_{i=0}^{k-1} M_i^{-1}} \sum_{i=0}^{k-1} M_i^{-1} w_i$.

Proof As it follows from (9), if $M_k \geq L\left(\delta_{c,k}\right) = L\left(\frac{\varepsilon}{4}\right)$, (15) holds. Thus, Algorithm 1 is correctly defined.

Let us fix some iteration $k \geq 0$. For simplicity, we denote $\tilde{g}(z_k) = \tilde{g}(z_k, \delta_{c,k}, \delta_u)$ and $\tilde{g}(w_k) = \tilde{g}(w_k, \delta_{c,k}, \delta_u)$. By the definition of inexact prox-mapping (7)-(8) and (13), (14), we have, for any $u \in Q$,

$$\langle \tilde{g}(z_k) + M_k \nabla d(w_k) - M_k \nabla d(z_k), u - w_k \rangle \geq -\delta_{pu} - \frac{\varepsilon}{8}, \tag{16}$$

$$\langle \tilde{g}(w_k) + M_k \nabla d(z_{k+1}) - M_k \nabla d(z_k), u - z_{k+1} \rangle \geq -\delta_{pu} - \frac{\varepsilon}{8}. \tag{17}$$

Whence, for all $u \in Q$,

$$\begin{aligned}
\langle \tilde{g}(w_k), w_k - u\rangle &= \langle \tilde{g}(w_k), z_{k+1} - u\rangle + \langle \tilde{g}(w_k), w_k - z_{k+1}\rangle \\
&\overset{(17)}{\leq} M_k\langle \nabla d(z_k) - \nabla d(z_{k+1}), z_{k+1} - u\rangle + \langle \tilde{g}(w_k), w_k - z_{k+1}\rangle + \delta_{pu} + \frac{\varepsilon}{8} \\
&= M_k(d(u) - d(z_k) - \langle \nabla d(z_k), u - z_k\rangle) - M_k(d(u) - d(z_{k+1}) \\
&\quad - \langle \nabla d(z_{k+1}), u - z_{k+1}\rangle) - M_k(d\,(z_{k+1}) - d\,(z_k) \\
&\quad - \langle \nabla d\,(z_k)\,, z_{k+1} - z_k\rangle) + \langle \tilde{g}(w_k), w_k - z_{k+1}\rangle + \delta_{pu} + \frac{\varepsilon}{8} \\
&= M_k V[z_k](u) - M_k V[z_{k+1}](u) - M_k V[z_k](z_{k+1}) + \langle \tilde{g}(w_k), w_k - z_{k+1}\rangle \\
&\quad + \delta_{pu} + \frac{\varepsilon}{8}
\end{aligned}$$

Further, for all $u \in Q$,

$$\begin{aligned}
\langle \tilde{g}(w_k), w_k - z_{k+1}\rangle - M_k V[z_k](z_{k+1}) &= \langle \tilde{g}(w_k) - \tilde{g}(z_k), w_k - z_{k+1}\rangle \\
&\quad - M_k V[z_k](z_{k+1}) + \langle \tilde{g}(z_k), w_k - z_{k+1}\rangle \\
&\overset{(16)}{\leq} \langle \tilde{g}(w_k) - \tilde{g}(z_k), w_k - z_{k+1}\rangle + M_k\langle \nabla d(z_k) - \nabla d(w_k), w_k - z_{k+1}\rangle \\
&\quad - M_k V[z_k](z_{k+1}) + \delta_{pu} + \frac{\varepsilon}{8} \\
&= \langle \tilde{g}(w_k) - \tilde{g}(z_k), w_k - z_{k+1}\rangle + M_k\langle \nabla d(z_k) - \nabla d(w_k), w_k - z_{k+1}\rangle \\
&\quad - M_k(d(z_{k+1}) - d(z_k) - \langle \nabla d(z_k), z_{k+1} - z_k\rangle) + \delta_{pu} + \frac{\varepsilon}{8} \\
&= \langle \tilde{g}(w_k) - \tilde{g}(z_k), w_k - z_{k+1}\rangle - M_k\big(d(w_k) - d(z_k) \\
&\quad - \langle \nabla d(z_k), w_k - z_k\rangle\big) - M_k(d(z_{k+1}) - d(w_k) - \langle \nabla d(w_k), z_{k+1} - w_k\rangle) \\
&\quad + \delta_{pu} + \frac{\varepsilon}{8} = \langle \tilde{g}(w_k) - \tilde{g}(z_k), w_k - z_{k+1}\rangle - M_k V[z_k](w_k) \\
&\quad - M_k V[w_k](z_{k+1}) + \delta_{pu} + \frac{\varepsilon}{8} \leq \langle \tilde{g}(w_k) - \tilde{g}(z_k), w_k - z_{k+1}\rangle \\
&\quad - \frac{M_k}{2}(\|z_k - w_k\|^2 + \|z_{k+1} - w_k\|^2) + \delta_{pu} + \frac{\varepsilon}{8} \overset{(15)}{\leq} \frac{3\varepsilon}{8} + \delta_u + \delta_{pu},
\end{aligned}$$

where we also used that $\delta_{c,k} = \varepsilon/4$.

Thus, we obtain, for all $u \in Q$ and $i \geq 0$,

$$M_i^{-1}\langle \tilde{g}(w_i), w_i - u\rangle \leq V[z_i](u) - V[z_{i+1}](u) + \frac{M_i^{-1}\varepsilon}{2} + M_i^{-1}(\delta_u + 2\delta_{pu}).$$

Summing up these inequalities for i from 0 to $k-1$, we have

$$\frac{1}{\sum_{i=0}^{k-1} M_i^{-1}} \sum_{i=0}^{k-1} M_i^{-1}\langle \tilde{g}(w_i), w_i - u\rangle \leq \frac{1}{\sum_{i=0}^{k-1} M_i^{-1}}(V[z_0](u) - V[z_k](u)) + \frac{\varepsilon}{2} + \delta_u + 2\delta_{pu}.$$

By (10), we obtain the statement of the Theorem 1. □

Note that the compactness of the set Q was not used in the proof.

Corollary 1. *Assume that $g(\cdot)$ and $\tilde{g}(\cdot, \delta_c, \delta_u)$ satisfy (9) and (10). Also let $C \subseteq Q$ be a convex compact. Then, for all $k \geq 0$, we have*

$$\mathrm{Gap}_C(\widehat{w}_k) = \max_{u \in C} \langle g(u), \widehat{w}_k - u \rangle \leq \frac{1}{\sum_{i=0}^{k-1} M_i^{-1}} \max_{u \in Q} V[z_0](u) + \frac{\varepsilon}{2} + \delta_u + 2\delta_{pu}, \tag{18}$$

where $\widehat{w}_k = \left(\sum_{i=0}^{k-1} M_i^{-1}\right)^{-1} \sum_{i=0}^{k-1} M_i^{-1} w_i$.

Proof By monotonicity of g, we have, for all $i \geq 0$ and $u \in Q$,

$$\langle g(u), w_i - u \rangle = \langle g(w_i), w_i - u \rangle + \langle g(u) - g(w_i), w_i - u \rangle \leq \langle g(w_i), w_i - u \rangle,$$

Therefore, $\left(\sum_{i=0}^{k-1} M_i^{-1}\right)^{-1} \sum_{i=0}^{k-1} M_i^{-1} \langle g(w_i), w_i - u \rangle \geq \langle g(u), \widehat{w}_k - u \rangle$ for any $u \in Q$. Combining this with Theorem 1 and taking the maximum over all $u \in C$, we obtain the statement of the Corollary 1. □

If a number D satisfying $\max_{u \in C} V[z_0](u) \leq D$ is known, which is the case for most of the standard proximal setups [45], we can construct an adaptive stopping criterion: the algorithm stops whenever $D \left(\sum_{i=0}^{k-1} M_i^{-1}\right)^{-1} \leq \varepsilon/2$. This guarantees that the r.h.s. of (18) is no greater than $\varepsilon + \delta_u + 2\delta_{pu}$ and $\widehat{w}_k$ is an $(\varepsilon + \delta_u + 2\delta_{pu})$-solution to (1).

Next, we consider the case of Hölder-continuous operator g and show that Algorithm 1 is universal. For simplicity we assume that the prox-mapping is calculated exactly, i.e., $\delta_{pc} = \delta_{pu} = 0$ and $\delta_u = 0$. In this case, it is sufficient to set $\delta_{c,k} = \frac{\varepsilon}{2}$ at each iteration of Algorithm 1. □

Corollary 2 (Universal Method for VI)**.** *Assume that the operator g is Hölder-continuous with constant L_ν for some $\nu \in [0, 1]$ and that in Algorithm 1 we have $\delta_{c,k} = \varepsilon/2$, $\delta_u = 0$, $\delta_{pc,k} = 0$, $\delta_{pu} = 0$. Also let $C \subseteq Q$ be a convex compact. Then, for all $k \geq 0$, we have*

$$\mathrm{Gap}_C(\widehat{w}_k) = \max_{u \in C} \langle g(u), \widehat{w}_k - u \rangle \leq \frac{2 L_\nu^{\frac{2}{1+\nu}}}{k \varepsilon^{\frac{1-\nu}{1+\nu}}} \max_{u \in C} V[z_0](u) + \frac{\varepsilon}{2}. \tag{19}$$

Proof As it follows from (9), if $M_k \geq L(\delta_{c,k}) L(\frac{\varepsilon}{2})$, (15) holds. Here $L(\cdot)$ is defined in Example 1. Thus, for all
$i = 0, ..., k-1$, we have $M_i \leq 2 \cdot L(\frac{\varepsilon}{2})$ and

$$\frac{1}{\sum_{i=0}^{k-1} M_i^{-1}} \leq \frac{2L(\frac{\varepsilon}{2})}{k} = \frac{2 L_\nu^{\frac{2}{1+\nu}}}{k \varepsilon^{\frac{1-\nu}{1+\nu}}}.$$

Thus, (19) follows from (18). □

Let us make several comments on the universal method. Since Algorithm 1 does not use the values of parameters ν and L_ν, we take the infinum w.r.t. $\nu \in [0, 1]$ and obtain the following iteration complexity bound to find $\widehat{w}_k$ satisfying $\max_{u \in C} \langle g(u), \widehat{w}_k - u \rangle \leq \varepsilon$:

$$4 \inf_{\nu \in [0,1]} \left(\frac{L_\nu}{\varepsilon}\right)^{\frac{2}{1+\nu}} \cdot \max_{u \in C} V[z_0](u).$$

Using the same reasoning as in [49], we estimate the number of oracle calls for Algorithm 1. The number of oracle calls at each iteration k is equal to $2i_k$, where by i_k we mean the last value of i_k at the end of the inner cycle. So, $M_k = 2^{i_k-2}M_{k-1}$ and, hence, $i_k = 2 + \log_2 \frac{M_k}{M_{k-1}}$. Thus, the total number of oracle calls is

$$\sum_{j=0}^{k-1} i_j = 4k + 2\sum_{i=0}^{k-1} \log_2 \frac{M_j}{M_{j-1}} < 4k + 2\log_2 \left(2L\left(\frac{\varepsilon}{2}\right)\right) - 2\log_2(M_{-1}), \tag{20}$$

where we used that $M_k \leq 2L(\frac{\varepsilon}{2})$. Hence, the total number of oracle calls of the Algorithm 1 does not exceed

$$\inf_{\nu\in[0,1]} \left(16\left(\frac{L_\nu}{\varepsilon}\right)^{\frac{2}{1+\nu}} \cdot \max_{u\in C} V[z_0](u) + 2\log_2 2\left(\left(\frac{1}{\varepsilon}\right)^{\frac{1-\nu}{1+\nu}} L_\nu^{\frac{2}{1+\nu}}\right)\right) - 2\log_2(M_{-1}).$$

Next, we compare our algorithm with the algorithm in [6] which appeared after the first version of this paper appeared as an arxiv preprint. Their algorithm uses AdaGrad-type of stepsizes and, denoting $D^2 \geq \max_{u\in C} V[z_0](u)$, for the non-smooth case $\nu = 0$, achieves, up to logarithmic factors, complexity of $O\left(L_0^2 D^2 \varepsilon^{-2}\right)$, which is similar to ours. For the smooth case $\nu = 1$ it achieves complexity $O\left((L_0 D + L_1 D^2)\varepsilon^{-1}\right)$, which is similar to ours. Unlike our paper, they consider only two extreme cases $\nu \in \{0, 1\}$, but also consider stochastic setting. At the same time, for our algorithm, $1/M_k$ plays the role of stepsize at iteration k, and from (18) it is clear that, the smaller M_k, the smaller is the right hand side and the better is the convergence guarantee. Step 5 of our algorithm ensures that the stepsize may decrease in the course of the algorithm execution, leading to a better performance in practice. This is in contrast to the stepsize in [6] which is a decreasing sequence. The experiments in [60, Appendix 7] show that decreasing M_k (which is equivalent to increasing the stepsize) allows to obtain much faster convergence.

5 Solving Variational Inequalities with Strongly Monotone Operator

In this section, we assume, that the operator g in (1) is strongly monotone, which means that, for some $\mu > 0$,

$$\langle g(x) - g(y), x - y\rangle \geq \mu\|x - y\|^2 \quad \forall x, y \in Q. \tag{21}$$

We slightly modify the assumptions on the prox-function d. Namely, we assume that $0 = \arg\min_{x\in Q} d(x)$ and that d is bounded on the unit ball in the chosen norm $\|\cdot\|$, that is

$$d(x) \leq \frac{\Omega}{2}, \quad \forall x \in Q : \|x\| \leq 1, \tag{22}$$

where Ω is some known constant. Note that for standard proximal setups, $\Omega = O(\ln \dim E)$ [38]. Finally, we assume that we are given a starting point $x_0 \in Q$ and a number $R_0 > 0$

such that $\|x_0 - x_*\|^2 \leq R_0^2$, where x_* is the solution to (1). We show that the well-known in optimization restart technique [47, 38, 20] also works in the context of VIs. To the best of our knowledge, this is the first time when this technique is applied to VIs. The resulting Restarted Generalized Mirror Prox algorithm is listed below as Algorithm 2.

Algorithm 2 Generalized Mirror Prox with restarts

Input: accuracy $\varepsilon > 0$, $\delta_u > 0$, $\delta_{pu} > 0$, $\mu > 0$, Ω such that $d(x) \leq \frac{\Omega}{2}\ \forall x \in Q : \|x\| \leq 1$; x_0, R_0 such that $\|x_0 - x_*\|^2 \leq R_0^2$.

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 Algorithm 1 for monotone case with accuracy $\mu\varepsilon/2$, δ_u, δ_{pu}, prox-function $d_p(\cdot)$ and stopping criterion $\sum_{i=0}^{k-1} M_i^{-1} \geq \frac{\Omega}{\mu}$.
4: Set $R_{p+1}^2 = R_0^2 \cdot 2^{-(p+1)} + 2(1 - 2^{-(p+1)})(\frac{\varepsilon}{4} + \delta_u + 2\delta_{pu})$.
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** $p > \log_2 \frac{2R_0^2}{\varepsilon}$.

Output: x_p.

Theorem 2. *Assume that g is strongly monotone with parameter μ. Also assume that the prox-function d satisfies (22) and the starting point $x_0 \in Q$ and a number $R_0 > 0$ are such that $\|x_0 - x_*\|^2 \leq R_0^2$, where x_* is the solution to (1). Then, for $p \geq 0$, the sequence x_p generated by Algorithm 2 satisfies*

$$\|x_p - x_*\|^2 \leq R_0^2 \cdot 2^{-p} + \frac{\varepsilon}{2} + \frac{2\delta_u + 4\delta_{pu}}{\mu},$$

and the point x_p returned by Algorithm 2 satisfies $\|x_p - x_*\|^2 \leq \varepsilon + \frac{2\delta_u + 4\delta_{pu}}{\mu}$.

Proof Let us denote $\Delta = \frac{\varepsilon}{4} + \frac{\delta_u + 2\delta_{pu}}{\mu}$. We show by induction that, for $p \geq 0$,

$$\|x_p - x_*\|^2 \leq R_0^2 \cdot 2^{-p} + 2(1 - 2^{-p})\Delta,$$

which leads to the statement of the Theorem 2. For $p = 0$ this inequality holds by the theorem assumption. Assuming that it holds for some $p \geq 0$, our goal is to prove it for $p+1$ considering the outer iteration $p+1$. Observe that the function d_p defined in Algorithm 2 is 1-strongly convex w.r.t. the norm $\|\cdot\|$. Using the definition of $d_p(\cdot)$ and (22), we have, since $x_p = \arg\min_{x \in Q} d_p(x)$ and $\|x_p - x_*\| \leq R_p$

$$V_p[x_p](x_*) = d_p(x_*) - d_p(x_p) - \langle \nabla d_p(x_p), x_* - x_p \rangle \leq d_p(x_*) = R_p^2 d\left(\frac{x_* - x_p}{R_p}\right) \leq \frac{\Omega R_p^2}{2}.$$

Thus, by Theorem 1, taking $u = x_*$, we obtain

$$\frac{1}{\sum_{i=0}^{k-1} M_i^{-1}} \sum_{i=0}^{k-1} M_i^{-1} \langle g(w_i), w_i - x_* \rangle \leq \frac{V_p[x_p](x_*)}{\sum_{i=0}^{k-1} M_i^{-1}} + \frac{\mu\varepsilon}{4} + \delta_u + 2\delta_{pu} \leq \frac{\Omega R_p^2}{2\sum_{i=0}^{k-1} M_i^{-1}} + \mu\Delta.$$

Since the operator g is continuous and monotone, the solution of the Minty VI (1) is also the solution of the Stampacchia variational inequality [48, 24], i.e., $\langle g(x_*), x_* - w_i \rangle \leq 0, \quad i = 0, ..., k-1$. This and the strong monotonicity of g, see (21), give, for all $i = 0, ..., k-1$,

$$\langle g(w_i), w_i - x_* \rangle \geq \langle g(w_i) - g(x_*), w_i - x_* \rangle \geq \mu \|w_i - x_*\|^2.$$

Thus, by convexity of the squared norm, we obtain

$$\begin{aligned}
\mu\|x_{p+1} - x_*\|^2 &= \mu \left\| \frac{1}{\sum_{i=0}^{k-1} M_i^{-1}} \sum_{i=0}^{k-1} M_i^{-1} w_i - x_* \right\|^2 \\
&\leq \frac{\mu}{\sum_{i=0}^{k-1} M_i^{-1}} \sum_{i=0}^{k-1} M_i^{-1} \|w_i - x_*\|^2 \\
&\leq \frac{1}{\sum_{i=0}^{k-1} M_i^{-1}} \sum_{i=0}^{k-1} M_i^{-1} \langle g(w_i), w_i - x_* \rangle \leq \frac{\Omega R_p^2}{2\sum_{i=0}^{k-1} M_i^{-1}} + \mu\Delta.
\end{aligned}$$

Using the stopping criterion $\sum_{i=0}^{k-1} M_i^{-1} \geq \frac{\Omega}{\mu}$, we get

$$\|x_{p+1} - x_*\|^2 \leq \frac{R_p^2}{2} + \Delta = \frac{1}{2}(R_0^2 \cdot 2^{-p} + 2(1 - 2^{-p})\Delta) + \Delta = R_0^2 \cdot 2^{-(p+1)} + 2(1 - 2^{-(p+1)})\Delta,$$

which finishes the induction proof. □

Corollary 3. *Assume that the operator g is Hölder-continuous with constant L_ν for some $\nu \in [0, 1]$ and strongly monotone with parameter μ. Then, Algorithm 2 returns a point x_p such that $\|x_p - x_*\|^2 \leq \varepsilon + \frac{2\delta_u + 4\delta_{pu}}{\mu}$ and the total number of iterations of the inner Algorithm 1 does not exceed*

$$\inf_{\nu \in [0,1]} \left\lceil \left(\frac{L_\nu}{\mu} \right)^{\frac{2}{1+\nu}} \frac{2^{\frac{2}{1+\nu}} \Omega}{\varepsilon^{\frac{1-\nu}{1+\nu}}} \cdot \log_2 \frac{2R_0^2}{\varepsilon} \right\rceil. \tag{23}$$

Proof Let us denote $\hat{p} = \left\lceil \log_2 \frac{2R_0^2}{\varepsilon} \right\rceil$. As it was shown in Corollary 2, at each inner iteration, $M_i \leq 2L(\frac{\mu\varepsilon}{4}) = 2\left(\frac{2}{\mu\varepsilon}\right)^{\frac{1-\nu}{1+\nu}} L_\nu^{\frac{2}{1+\nu}}$. Thus, by the stopping criterion $\sum_{i=0}^{k-1} M_i^{-1} \geq \frac{\Omega}{\mu}$, the inner cycle stops at the latest when

$$k_p = \left\lceil \left(\frac{L_\nu}{\mu} \right)^{\frac{2}{1+\nu}} \frac{2^{\frac{2}{1+\nu}} \Omega}{\varepsilon^{\frac{1-\nu}{1+\nu}}} \right\rceil$$

and we have

$$N = \sum_{p=1}^{\hat{p}} k_p \leq \left\lceil \left(\frac{L_\nu}{\mu}\right)^{\frac{2}{1+\nu}} \frac{2^{\frac{2}{1+\nu}}\Omega}{\varepsilon^{\frac{1-\nu}{1+\nu}}} \cdot \log_2 \frac{2R_0^2}{\varepsilon} \right\rceil.$$

Since Algorithm 1 does not need to know ν and L_ν, we can take the infimum w.r.t. $\nu \in [0,1]$. □The obtained complexity estimate is optimal for the case $\nu = 1$ [65] and is optimal up to a logarithmic factor for the case $\nu = 0$, [46]. For the intermediate case $\nu \in (0,1)$ we are not aware of any lower bounds. As a remark, we note that the complexity estimate for the case $\nu = 0$ is $O\left(\frac{L_0^2}{\mu^2\varepsilon}\right)$, whereas one would expect $O\left(\frac{L_0^2}{\mu\varepsilon}\right)$. The reason is that we measure the error in terms of the distance to the solution $\|x_p - x_*\|$, but not in terms of the residual in VI, i.e. $\max_{u\in Q}\langle g(u), x_p - u\rangle$, as in Corollary 2.

6 Applications to Saddle Point Problems

In this section, we consider saddle point problems and show, how Generalized Mirror Prox can be applied to such problems. The problem, we consider is

$$f^* = \min_{u\in Q_1}\max_{v\in Q_2} f(u,v), \tag{24}$$

where $Q_1 \subset E_1$ and $Q_2 \subset E_2$ are convex and closed subsets of normed spaces E_1 and E_2 with norms $\|\cdot\|_1$ and $\|\cdot\|_2$, respectively. Based on the norms in E_1 and E_2, we define the norm on their product $E_1 \times E_2$ as
$\|x\| = \max\{\|u\|_1, \|v\|_2\}$, $x = (u,v) \in E_1 \times E_2$ with the corresponding dual norm $\|s\|_* = \|z\|_{1,*} + \|w\|_{2,*}$, $s = (z,w) \in E^*$, where $\|\cdot\|_{1,*}$ and $\|\cdot\|_{2,*}$ are the norms on the conjugate spaces E_1^* and E_2^*, dual to $\|\cdot\|_1$ and $\|\cdot\|_2$ respectively.

The function f in (24) is assumed to be convex in u and concave in v. As it is usually done, we consider the operator

$$g(x) = \begin{pmatrix} \nabla_u f(u,v) \\ -\nabla_v f(u,v) \end{pmatrix}, \; x = (u,v) \in Q := Q_1 \times Q_2. \tag{25}$$

By the convexity of f in u and the concavity in v, the operator g is monotone

$$\langle g(x) - g(y), x - y\rangle \geq 0 \quad \forall x, y \in Q \subset E, \tag{26}$$

where $x = (u_1, v_1), y = (u_2, v_2)$.

The following lemma gives sufficient conditions for g to be Hölder-continuous, i.e., to satisfy (2).

Lemma 2. *Assume that for f in (24) there exist a number $\nu \in [0,1]$ and constants $L_{11,\nu}, L_{12,\nu}, L_{21,\nu}, L_{22,\nu} < +\infty$ such that*

$$\|\nabla_u f(u+\Delta u, v+\Delta v) - \nabla_u f(u,v)\|_{1,*} \leq L_{11,\nu}\|\Delta u\|_1^\nu + L_{12,\nu}\|\Delta v\|_2^\nu, \tag{27}$$

$$\|\nabla_v f(u+\Delta u, v+\Delta v) - \nabla_v f(u,v)\|_{2,*} \le L_{21,\nu}\|\Delta u\|_1^\nu + L_{22,\nu}\|\Delta v\|_2^\nu \tag{28}$$

for all $u, u+\Delta u \in Q_1, v, v+\Delta v \in Q_2$. *Then* g *defined in* (25) *is Hölder-continuous, i.e., satisfies* (2) *with the same* ν *and*

$$L_\nu = L_{11,\nu} + L_{12,\nu} + L_{21,\nu} + L_{22,\nu}.$$

Proof Indeed, for each $x = (u_1, v_1), y = (u_2, v_2) \in Q$ we have:

$$\begin{aligned}
\|g(x) - g(y)\|_* &= \|\nabla_u f(u_1,v_1) - \nabla_u f(u_2,v_2)\|_{1,*} + \|\nabla_v f(u_1,v_1) - \nabla_v f(u_2,v_2)\|_{2,*} \\
&\le L_{11,\nu}\|u_1-u_2\|_1^\nu + L_{12,\nu}\|v_1-v_2\|_2^\nu + L_{21,\nu}\|u_1-u_2\|_1^\nu + L_{22,\nu}\|v_1-v_2\|_2^\nu \\
&= (L_{11,\nu}+L_{21,\nu})\|u_1-u_2\|_1^\nu + (L_{12,\nu}+L_{22,\nu})\|v_1-v_2\|_2^\nu \\
&\le (L_{11,\nu}+L_{12,\nu}+L_{21,\nu}+L_{22,\nu})\max\{\|u_1-u_2\|_1^\nu, \|v_1-v_2\|_2^\nu\} \\
&= (L_{11,\nu}+L_{12,\nu}+L_{21,\nu}+L_{22,\nu})\|x-y\|^\nu. \qquad \square
\end{aligned}$$

Remark 1. As an alternative, one can consider the following primal and dual pair of norms for $E = E_1 \times E_2$: $\|x\| = \sqrt{\|u\|_1^2 + \|v\|_2^2}$, $x = (u,v) \in E_1 \times E_2$, and $\|s\|_* = \sqrt{\|z\|_{1,*}^2 + \|w\|_{2,*}^2}$, $s = (z,w) \in E^*$, where $\|\cdot\|_{1,*}$ and $\|\cdot\|_{2,*}$ are the norms on the conjugate spaces E_1^* and E_2^*, dual to $\|\cdot\|_1$ and $\|\cdot\|_2$ respectively. We have, for each $x = (u_1, v_1), y = (u_2, v_2) \in Q$,

$$\begin{aligned}
\|g(x) - g(y)\|_*^2 &= \|\nabla_u f(u_1,v_1) - \nabla_u f(u_2,v_2)\|_{1,*}^2 + \|\nabla_v f(u_1,v_1) - \nabla_v f(u_2,v_2)\|_{2,*}^2 \\
&\le 2\big(L_{11,\nu}^2\|u_1-u_2\|_1^{2\nu} + L_{12,\nu}^2\|v_1-v_2\|_2^{2\nu} + L_{21,\nu}^2\|u_1-u_2\|_1^{2\nu} + L_{22,\nu}^2\|v_1-v_2\|_2^{2\nu}\big) \\
&= 2(L_{11,\nu}^2+L_{21,\nu}^2)\|u_1-u_2\|_1^{2\nu} + 2(L_{12,\nu}^2+L_{22,\nu}^2)\|v_1-v_2\|_2^{2\nu} \\
&\le 2(L_{11,\nu}^2+L_{12,\nu}^2+L_{21,\nu}^2+L_{22,\nu}^2)\max\{\|u_1-u_2\|_1^{2\nu}, \|v_1-v_2\|_2^{2\nu}\}) \\
&\le 2(L_{11,\nu}^2+L_{12,\nu}^2+L_{21,\nu}^2+L_{22,\nu}^2)\|x-y\|^{2\nu}
\end{aligned}$$

and

$$\|g(x) - g(y)\|_* \le \sqrt{2(L_{11,\nu}^2+L_{12,\nu}^2+L_{21,\nu}^2+L_{22,\nu}^2)}\|x-y\|^\nu.$$

Remark 2. Generally speaking, if the set Q is bounded, one can consider different level of smoothness in (27) and (28). More precisely, assume that for some numbers $\nu_{11}, \nu_{12}, \nu_{21}, \nu_{22} \in [0;1]$, we have

$$\|\nabla_u f(u+\Delta u, v+\Delta v) - \nabla_u f(u,v)\|_{1,*} \le \widehat{L}_{11}\|\Delta u\|_1^{\nu_{11}} + \widehat{L}_{12}\|\Delta v\|_2^{\nu_{12}}, \tag{29}$$

$$\|\nabla_v f(u+\Delta u, v+\Delta v) - \nabla_v f(u,v)\|_{2,*} \le \widehat{L}_{21}\|\Delta u\|_1^{\nu_{21}} + \widehat{L}_{22}\|\Delta v\|_2^{\nu_{22}} \tag{30}$$

for all $u, u+\Delta u \in Q_1, v, v+\Delta v \in Q_2$. Then the statement of Lemma 2 holds for $\nu = \min\{\nu_{11}, \nu_{12}, \nu_{21}\, \nu_{22}\} \in [0;1]$. Indeed, from (29), (30), we have

$$\|\nabla_u f(u+\Delta u, v+\Delta v) - \nabla_u f(u,v)\|_{1,*} \le \widehat{L}_{11} \cdot D_Q^{\nu_{11}-\nu} \cdot \|\Delta u\|_1^\nu + \widehat{L}_{12} \cdot D_Q^{\nu_{12}-\nu} \cdot \|\Delta v\|_2^\nu,$$

$$\|\nabla_v f(u+\Delta u, v+\Delta v) - \nabla_v f(u,v)\|_{2,*} \le \widehat{L}_{21} \cdot D_Q^{\nu_{21}-\nu} \cdot \|\Delta u\|_1^\nu + \widehat{L}_{22} \cdot D_Q^{\nu_{22}-\nu} \cdot \|\Delta v\|_2^\nu$$

for all $u, u+\Delta u \in Q_1, v, v+\Delta v \in Q_2$, $D_Q = \sup\{\|x-y\| \,|\, x, y \in Q\}$.

The next theorem shows, how Algorithm 1 can be applied to solve the saddle point problem (24).

Theorem 3. *Let the assumptions of Lemma 2 hold, the set Q be bounded, and L_ν be given in Lemma 2. Assume also that Algorithm 1 with accuracy ε is applied to the operator g defined in (25). Let $w_i = (u^i, v^i)$ be the sequence generated by this algorithm. Then,*

$$\max_{v \in Q_2} f(\widehat{u}_k, v) - \min_{u \in Q_1} f(u, \widehat{v}_k) \leq \frac{2L_\nu^{\frac{2}{1+\nu}}}{k\varepsilon^{\frac{1-\nu}{1+\nu}}} \max_{x \in Q} V[w_0](x) + \frac{\varepsilon}{2},$$

$$\text{where} \quad (\widehat{u}_k, \widehat{v}_k) = \frac{1}{S_k} \sum_{i=0}^{k-1} M_i^{-1}(u^i, v^i), \quad S_k = \sum_{i=0}^{k-1} M_i^{-1}.$$

Moreover, in the number of iterations

$$O\left(\inf_{\nu \in [0,1]} \left(\frac{L_\nu}{\varepsilon} \right)^{\frac{2}{1+\nu}} \cdot \max_{x \in Q} V[w_0](x) \right),$$

the algorithm finds a pair $(\widehat{u}, \widehat{v})$ such that $\max_{v \in Q_2} f(\widehat{u}, v) - \min_{u \in Q_1} f(u, \widehat{v}) \leq \varepsilon$.

Proof. By convexity of f in u and concavity of f in v, we have, for all $u \in Q_1$,

$$\begin{aligned} \frac{1}{S_k} \sum_{i=0}^{k-1} \left\langle \nabla_u f(u^i, v^i), u^i - u \right\rangle_1 &\geq \frac{1}{S_k} \sum_{i=0}^{k-1} M_i^{-1}(f(u^i, v^i) - f(u, v^i)) \\ &\geq \frac{1}{S_k} \sum_{i=0}^{k-1} M_i^{-1} f(u^i, v^i) - f(u, \widehat{v}_k). \end{aligned}$$

In the same way, we obtain, for all $v \in Q_2$,

$$\frac{1}{S_k} \sum_{i=0}^{k-1} M_i^{-1} \left\langle -\nabla_v f(u^i, v^i), v^i - v \right\rangle_2 \geq -\frac{1}{S_k} \sum_{i=0}^{k-1} M_i^{-1} f(u^i, v^i) + f(\widehat{u}_k, v).$$

Summing these inequalities, using (25) and Theorem 1, we obtain that, for all $u \in Q_1$, $v \in Q_2$ and $x = (u, v)$,

$$f(\widehat{u}_k, v) - f(u, \widehat{v}_k) \leq \frac{1}{S_k} \sum_{i=0}^{k-1} M_i^{-1} \langle g(w_i), w_i - x \rangle \leq \frac{1}{\sum_{i=0}^{k-1} M_i^{-1}} V[w_0](x) + \frac{\varepsilon}{2}.$$

Since $M_i \leq 2L\left(\frac{\varepsilon}{2}\right)$, where $L(\cdot)$ is given in Example 1, and the set Q is bounded, we obtain

$$\max_{v \in Q_2} f(\widehat{u}_k, v) - \min_{u \in Q_1} f(u, \widehat{v}_k) \leq \frac{2L_\nu^{\frac{2}{1+\nu}}}{k\varepsilon^{\frac{1-\nu}{1+\nu}}} \max_{x \in Q} V[w_0](x) + \frac{\varepsilon}{2}.$$

The iteration complexity follows from the requirement for k to make the first term in r.h.s. smaller than ε. □

Remark 3. Since, for a saddle point $(u_*, v_*) \in Q$, $\max\limits_{v \in Q_2} f(u_*, v) = \min\limits_{u \in Q_1} f(u, v_*)$, the inequality $\max_{v \in Q_2} f(\widehat{u}, v) - \min_{u \in Q_1} f(u, \widehat{v}) \leq \varepsilon$ means that $(\widehat{u}, \widehat{v})$ is an ε-optimal solution.

Remark 4. For μ-strongly convex-concave saddle point problems, Algorithm 2 returns a point x_p such that $\|x_p - x_*\|^2 \leq \varepsilon + \frac{2\delta_u + 4\delta pu}{\mu}$ (for the exact solution x_*) with the complexity estimate given in (23).

An important particular case of saddle point problem is the Lagrange saddle point problem for a constrained minimization problem. Let us consider the following convex optimization problem

$$\min\{f(x) \quad : \quad x \in Q, \quad \phi_j(x) \leq 0, \quad j = 1, ..., m\}, \tag{31}$$

where Q is a convex and compact set, f and ϕ_j are convex and have Hölder-continuous subgradients

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L_{\nu_0} \|x - y\|^{\nu_0},$$

$$\|\nabla \phi_j(x) - \nabla \phi_j(y)\|_* \leq L_{\nu_j} \|x - y\|^{\nu_j} \quad \forall x, y \in Q, j = 1, ..., m$$

for some $\nu_0, ..., \nu_m \in [0, 1]$ and $L_{\nu_0}, ..., L_{\nu_m} > 0$. The corresponding Lagrange function for this problem is $L(x, \lambda) = f(x) + \sum\limits_{j=1}^{m} \lambda_j \phi_j(x)$, where $\lambda_j \geq 0$, $j = 1, ..., m$ are Lagrange multipliers. If a point (x_*, λ_*) is a saddle point of the convex-concave Lagrange function $L(x, \lambda)$, then x_* is a solution to (31). Assume also that the Slater's constraint qualification condition holds, i.e., there exists a point $\bar{x}$ such that $\phi_j(\bar{x}) < 0$, $j = 1, ..., m$. Then it can be shown that the optimal Lagrange multiplier λ_* is bounded. Thus, instead of minimization problem (31), one can consider the saddle point problem $\min_{x \in Q} \max_{\lambda \in \Lambda} L(x, \lambda)$, which is a convex-concave problem on a bounded set. Using Lemma 2 and the Hölder continuity assumption for the subgradients of f and ϕ_j, we see that Algorithm 1 and Theorem 3 can be applied. We underline that by its nature, the smoothness level of the primal problem and the dual problem is different. Thus, it is important for the algorithm to adapt to the actual level of smoothness.

Next we introduce the concept inexact oracle for saddle point problems.

Definition 2. *Assume that for some $\delta_u > 0$ (uncontrolled error) and for any number $\delta_c > 0$ (controlled error) there exists a constant $L(\delta_c) \in]0, +\infty[$ such that, for any points $x, y \in Q$, one can calculate $\tilde{g}(x, \delta_c, \delta_u)$ and*
$\tilde{g}(y, \delta_c, \delta_u) \in E^*$ satisfying

$$\langle \tilde{g}(y, \delta_c, \delta_u) - \tilde{g}(x, \delta_c, \delta_u), y - z \rangle \leq \frac{L(\delta_c)}{2} \left(\|y - x\|^2 + \|y - z\|^2 \right) + \delta_c + \delta_u, \tag{32}$$

and, for each $x = (u_x, v_x), y = (u_y, v_y) \in Q$,

$$f(u_y, v_x) - f(u_x, v_y) \leq \langle \tilde{g}(y, \delta_c, \delta_u), y - x \rangle + \delta_u. \tag{33}$$

Then the operator $\tilde{g}(\cdot, \delta_c, \delta_u)$ is called inexact oracle for the problem (24).

Remark 5. Recall (see Definition 1), that δ_c represents the error of the oracle, which one can control and make as small as we would like to. On the opposite, δ_u represents the error, which one can not control.

Example 3 (*saddle point problems and (δ, L)-oracle in optimization*)**.** *Assume that we have access to the operator*

$$g_\delta(x) = \begin{pmatrix} g_{\delta,u}(u,v) \\ -g_{\delta,v}(u,v) \end{pmatrix}, x = (u,v) \in Q,$$

where $(f_{\delta,u}, g_{\delta,u})$ is a (δ, L)-oracle for f as a function of u, and $-(f_{\delta,v}, g_{\delta,v})$ is a (δ, L)-oracle for $(-f)$ as a function of v, see (11).
Define $\|x\| = \|(u,v)\| := \sqrt{\|u\|^2 + \|v\|^2}$. By Example 2, for each $x = (u_x, v_x)$, $y = (u_y, v_y)$, $z = (u_z, v_z) \in Q$,

$$\langle g_{\delta,u}(u_y, v_y) - g_{\delta,u}(u_x, v_x), u_y - u_z\rangle \le \frac{L}{2}(\|u_y - u_z\|^2 + \|u_y - u_x\|^2) + 2\delta,$$
$$\langle -g_{\delta,v}(u_y, v_y) + g_{\delta,v}(u_x, v_x), v_y - v_z\rangle \le \frac{L}{2}(\|v_y - v_z\|^2 + \|v_y - v_x\|^2) + 2\delta,$$

and we have

$$\langle g_\delta(y) - g_\delta(x), y - z\rangle \le \frac{L}{2}(\|y - z\|^2 + \|y - x\|^2) + 4\delta. \tag{34}$$

Further, from inequalities

$$f(u_y, v_y) - f(u_x, v_y) \le \langle g_{\delta,u}(y), u_y - u_x\rangle + \delta,$$
$$f(u_y, v_x) - f(u_y, v_y) \le \langle -g_{\delta,v}(y), v_y - v_x\rangle + \delta$$

we have $f(u_y, v_x) - f(u_x, v_y) \le \langle g_\delta(y), y - x\rangle + 2\delta$. So, $\tilde{g}(y, \delta_c, \delta_u) = g_\delta(y)$ satisfies Definition 2 with $\delta_u = 4\delta$, $\delta_c = 0$ and $L(\delta_c) = L$.

Similarly to Theorem 3 it is sufficient to make

$$O\left(\inf_{\nu\in[0,1]}\left(\frac{L}{\varepsilon}\right) \cdot \max_{x\in Q} V[w_0](x)\right)$$

iterations of Algorithm 1, to find a pair $(\widehat{u}, \widehat{v})$ satisfying

$$\max_{v\in Q_2} f(\widehat{u}, v) - \min_{u\in Q_1} f(u, \widehat{v}) \le \varepsilon + O(\delta_u + \delta_c).$$

7 Conclusions

In this paper we introduced a definition of inexact oracle for VIs with monotone operator and provided several examples, where such inexactness naturally arises. In particular, we showed, that Hölder-continuous operator is covered by our general framework of inexact oracle. In order to solve VIs with inexact oracle, we generalized Mirror Prox algorithm [45] and provided theoretical guarantees for its convergence rate. As a corollary, we proved

that this method is universal for VIs with Hölder-continuous monotone operator, i.e., has complexity

$$O\left(\inf_{\nu\in[0,1]}\left(\frac{L_\nu}{\varepsilon}\right)^{\frac{2}{1+\nu}}R^2\right)$$

and, unlike existing methods, does not require any knowledge of L_ν or ν. We generalized our algorithm for the case of μ-strongly monotone operators and obtain complexity

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

to find a point $\widehat{x}\in Q$ such that $\|\widehat{x}-x_*\|\leq\varepsilon$. Finally, we showed, how our method can be applied to convex-concave saddle point problems. In the follow-up work [60] we extended the proposed here methods for the case of inexact relative smoothness and strong convexity.

As a future work it would be interesting to generalize this method for the case of stochastic VIs using the techniques in [18, 33] and apply it for the Wasserstein barycenter problem [62], apply this method for solving differential games in the spirit of [22, 23], extend this algorithm to the case of VIs with operator having higher-order Hölder-continuous derivatives [50, 25, 12, 53], propose a generalization for zeroth order methods for saddle point problems [32, 57] using the techniques in [34, 58, 21], for accelerated methods for saddle-point problems [26, 64, 1, 63], for decentralized distributed setup by combining with the ideas of [17, 56, 10].

Acknowledgements The authors are grateful to Yurii Nesterov for fruitful discussions. The research by P. Dvurechensky and A. Gasnikov in Section 3 was supported by the Ministry of Science and Higher Education of the Russian Federation (Goszadaniye) No. 075-00337-20-03, project No. 0714-2020-0005. The research by F. Stonyakin in Section 6 and Appendix B was supported by Russian Science Foundation (project 18-71-00048).

References

[1] M. S. Alkousa, A. V. Gasnikov, D. M. Dvinskikh, D. A. Kovalev, and F. S. Stonyakin, *Accelerated methods for saddle-point problem*, Computational Mathematics and Mathematical Physics, 60 (2020), pp. 1787–1809.

[2] K. Antonakopoulos, V. Belmega, and P. Mertikopoulos, *Adaptive extra-gradient methods for min-max optimization and games*, in International Conference on Learning Representations, 2021, `https://openreview.net/forum?id=R0a0kFI3dJx`.

[3] M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein GAN*, arXiv:1701.07875, (2017).

[4] A. Auslender and M. Teboulle, *Interior projection-like methods for monotone variational inequalities*, Mathematical programming, 104 (2005), pp. 39–68.

[5] N. S. Aybat, A. Fallah, M. Gurbuzbalaban, and A. Ozdaglar, *Robust accelerated gradient methods for smooth strongly convex functions*, SIAM J. Optim., 30 (2020), pp. 717–751.

[6] F. Bach and K. Y. Levy, *A universal algorithm for variational inequalities adaptive to smoothness and noise*, in Proceedings of the Thirty-Second Conference on Learning Theory, A. Beygelzimer and D. Hsu, eds., vol. 99 of Proceedings of Machine Learning Research, Phoenix, USA, 25–28 Jun 2019, PMLR, pp. 164–194, `http://proceedings.mlr.press/v99/bach19a.html`. arXiv:1902.01637.

[7] D. R. Baimurzina, A. V. Gasnikov, E. V. Gasnikova, P. E. Dvurechensky, E. I. Ershov, M. B. Kubentaeva, and A. A. Lagunovskaya, *Universal method of searching for equilibria and stochastic equilibria in transportation networks*, Computational Mathematics and Mathematical Physics, 59 (2019), pp. 19–33.

[8] A. Bayandina, P. Dvurechensky, A. Gasnikov, F. Stonyakin, and A. Titov, *Mirror descent and convex optimization problems with non-smooth inequality constraints*, in Large-Scale and Distributed Optimization, P. Giselsson and A. Rantzer, eds., Springer International Publishing, 2018, ch. 8, pp. 181–215.

[9] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization (Lecture Notes)*, Personal web-page of A. Nemirovski, 2015, `https://www2.isye.gatech.edu/~nemirovs/LMCO_LN.pdf`.

[10] A. Beznosikov, P. Dvurechensky, A. Koloskova, V. Samokhin, S. U. Stich, and A. Gasnikov, *Decentralized local stochastic extra-gradient for variational inequalities*, arXiv:2106.08315, (2021).

[11] L. Bogolubsky, P. Dvurechensky, A. Gasnikov, G. Gusev, Y. Nesterov, A. M. Raigorodskii, A. Tikhonov, and M. Zhukovskii, *Learning supervised pagerank with gradient-based and gradient-free optimization methods*, in Advances in Neural Information Processing Systems 29, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, eds., Curran Associates, Inc., 2016, pp. 4914–4922. arXiv:1603.00717.

[12] B. Bullins and K. A. Lai, *Higher-order methods for convex-concave min-max optimization and monotone variational inequalities*, arXiv:2007.04528, (2020).

[13] M. Cohen, J. Diakonikolas, and L. Orecchia, *On acceleration with noise-corrupted gradients*, in Proceedings of the 35th International Conference on Machine Learning, J. Dy and A. Krause, eds., vol. 80 of Proceedings of Machine Learning Research, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018, PMLR, pp. 1019–1028. arXiv:1805.12591.

[14] C. D. Dang and G. Lan, *On the convergence properties of non-Euclidean extragradient methods for variational inequalities with generalized monotone operators*, Computational Optimization and Applications, 60 (2015), pp. 277–310.

[15] A. d'Aspremont, *Smooth optimization with approximate gradient*, SIAM J. on Optimization, 19 (2008), pp. 1171–1183.

[16] O. Devolder, F. Glineur, and Y. Nesterov, *First-order methods of smooth convex optimization with inexact oracle*, Mathematical Programming, 146 (2014), pp. 37–75.

[17] D. Dvinskikh and A. Gasnikov, *Decentralized and parallel primal and dual accelerated methods for stochastic convex programming problems*, Journal of Inverse and Ill-posed Problems, 29 (2021), pp. 385–405, https://doi.org/doi:10.1515/jiip-2020-0068, https://doi.org/10.1515/jiip-2020-0068.

[18] D. Dvinskikh, A. Ogaltsov, A. Gasnikov, P. Dvurechensky, and V. Spokoiny, *On the line-search gradient methods for stochastic optimization*, IFAC-PapersOnLine, 53 (2020), pp. 1715–1720, https://doi.org/https://doi.org/10.1016/j.ifacol.2020.12.2284. 21th IFAC World Congress, arXiv:1911.08380.

[19] P. Dvurechensky, *Gradient method with inexact oracle for composite non-convex optimization*, arXiv:1703.09180, (2017).

[20] P. Dvurechensky and A. Gasnikov, *Stochastic intermediate gradient method for convex problems with stochastic inexact oracle*, Journal of Optimization Theory and Applications, 171 (2016), pp. 121–145.

[21] P. Dvurechensky, E. Gorbunov, and A. Gasnikov, *An accelerated directional derivative method for smooth stochastic convex optimization*, European Journal of Operational Research, 290 (2021), pp. 601 – 621, https://doi.org/https://doi.org/10.1016/j.ejor.2020.08.027, http://www.sciencedirect.com/science/article/pii/S0377221720307402.

[22] P. Dvurechensky, Y. Nesterov, and V. Spokoiny, *Primal-dual methods for solving infinite-dimensional games*, Journal of Optimization Theory and Applications, 166 (2015), pp. 23–51.

[23] P. E. Dvurechensky and G. E. Ivanov, *Algorithms for computing Minkowski operators and their application in differential games*, Computational Mathematics and Mathematical Physics, 54 (2014), pp. 235–264.

[24] F. Facchinei and J.-S. Pang, *Finite-dimensional variational inequalities and complementarity problems*, Springer Science & Business Media, 2007.

[25] A. Gasnikov, P. Dvurechensky, E. Gorbunov, E. Vorontsova, D. Selikhanovych, C. A. Uribe, B. Jiang, H. Wang, S. Zhang, S. Bubeck, Q. Jiang, Y. T. Lee, Y. Li, and A. Sidford, *Near optimal methods for minimizing convex functions with lipschitz p-th derivatives*, in Proceedings of the Thirty-Second Conference on Learning Theory, A. Beygelzimer and D. Hsu, eds., vol. 99 of Proceedings of Machine Learning Research, Phoenix, USA, 25–28 Jun 2019, PMLR, pp. 1392–1393.

[26] A. V. Gasnikov, D. M. Dvinskikh, P. E. Dvurechensky, D. I. Kamzolov, V. V. Matyukhin, D. A. Pasechnyuk, N. K. Tupitsa, and A. V. Chernov, *Accelerated meta-algorithm for convex optimization problems*, Computational Mathematics and Mathematical Physics, 61 (2021), pp. 17–28, https://doi.org/10.1134/S096554252101005X, https://doi.org/10.1134/S096554252101005X.

[27] A. V. Gasnikov and P. E. Dvurechensky, *Stochastic intermediate gradient method for convex optimization problems*, Doklady Mathematics, 93 (2016), pp. 148–151.

[28] A. V. Gasnikov, P. E. Dvurechensky, F. S. Stonyakin, and A. A. Titov, *An adaptive proximal method for variational inequalities*, Computational Mathematics and Mathematical Physics, 59 (2019), pp. 836–841.

[29] A. V. Gasnikov and Y. E. Nesterov, *Universal method for stochastic composite optimization problems*, Computational Mathematics and Mathematical Physics, 58 (2018), pp. 48–64.

[30] S. Ghadimi, G. Lan, and H. Zhang, *Generalized uniformly optimal methods for nonlinear programming*, Journal of Scientific Computing, 79 (2019), pp. 1854–1881.

[31] F. Giannessi, *On Minty variational principle*, New Trends in Mathematical Programming. Applied Optimization, 13 (1997), pp. 93–99.

[32] E. Gladin, A. Sadiev, A. Gasnikov, P. Dvurechensky, A. Beznosikov, and M. Alkousa, *Solving smooth min-min and min-max problems by mixed oracle algorithms*, in Mathematical Optimization Theory and Operations Research: Recent Trends, A. Strekalovsky, Y. Kochetov, T. Gruzdeva, and A. Orlov, eds., Cham, 2021, Springer International Publishing, pp. 19–40. arXiv:2103.00434.

[33] E. Gorbunov, M. Danilova, I. Shibaev, P. Dvurechensky, and A. Gasnikov, *Near-optimal high probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise*, arXiv:2106.05958, (2021).

[34] E. Gorbunov, P. Dvurechensky, and A. Gasnikov, *An accelerated method for derivative-free smooth stochastic convex optimization*, arXiv:1802.09022, (2018).

[35] S. Guminov, A. Gasnikov, A. Anikin, and A. Gornov, *A universal modification of the linear coupling method*, Optimization Methods and Software, 34 (2019), pp. 560–577.

[36] S. V. Guminov, Y. E. Nesterov, P. E. Dvurechensky, and A. V. Gasnikov, *Accelerated primal-dual gradient descent with linesearch for convex, nonconvex, and nonsmooth optimization problems*, Doklady Mathematics, 99 (2019), pp. 125–128.

[37] P. T. Harker and J.-S. Pang, *Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications*, Mathematical programming, 48 (1990), pp. 161–220.

[38] A. Juditsky and A. Nemirovski, *First order methods for nonsmooth convex large scale optimization, i: general purpose methods*, Optimization for Machine Learning, (2011), pp. 121–148.

[39] D. Kamzolov, P. Dvurechensky, and A. Gasnikov, *Universal intermediate gradient method for convex problems with inexact oracle*, Optimization Methods and Software, (2020), pp. 1–28.

[40] P. D. Khanh and P. T. Vuong, *Modified projection method for strongly pseudomonotone variational inequalities*, Journal of Global Optimization, 58 (2014), pp. 341–350.

[41] V. V. Kniaz, V. A. Knyaz, V. Mizginov, A. Papazyan, N. Fomin, and L. Grodzitsky, *Adversarial dataset augmentation using reinforcement learning and 3d modeling*, in Advances in Neural Computation, Machine Learning, and Cognitive Research IV, B. Kryzhanovsky, W. Dunin-Barkowski, V. Redko, and Y. Tiumentsev, eds., Cham, 2021, Springer International Publishing, pp. 316–329.

[42] G. Korpelevich, *The extragradient method for finding saddle points and other problems*, Eknomika i Matematicheskie Metody, 12 (1976), pp. 747–756.

[43] J. Koshal, A. Nedić, and U. Shanbhag, *Multiuser optimization: Distributed algorithms and error analysis*, SIAM Journal on Optimization, 21 (2011), pp. 1046–1081.

[44] R. D. Monteiro and B. F. Svaiter, *On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean*, SIAM Journal on Optimization, 20 (2010), pp. 2755–2787.

[45] A. Nemirovski, *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 (2004), pp. 229–251.

[46] A. Nemirovsky and D. Yudin, *Problem Complexity and Method Efficiency in Optimization*, J. Wiley & Sons, New York, 1983.

[47] Y. Nesterov, *A method of solving a convex programming problem with convergence rate $o(1/k^2)$*, Soviet Mathematics Doklady, 27 (1983), pp. 372–376.

[48] Y. Nesterov, *Dual extrapolation and its applications to solving variational inequalities and related problems*, Mathematical Programming, 109 (2007), pp. 319–344. First appeared in 2003 as CORE discussion paper 2003/68.

[49] Y. Nesterov, *Universal gradient methods for convex optimization problems*, Mathematical Programming, 152 (2015), pp. 381–404.

[50] Y. Nesterov, *Implementable tensor methods in unconstrained convex optimization*, Mathematical Programming, (2019).

[51] Y. Nesterov, A. Gasnikov, S. Guminov, and P. Dvurechensky, *Primal-dual accelerated gradient methods with small-dimensional relaxation oracle*, Optimization Methods and Software, (2020), pp. 1–28.

[52] Y. Nesterov and L. Scrimali, *Solving strongly monotone variational and quasi-variational inequalities*, Discrete & Continuous Dynamical Systems - A, 31 (2011), pp. 1383–1396.

[53] P. Ostroukhov, R. Kamalov, P. Dvurechensky, and A. Gasnikov, *Tensor methods for strongly convex strongly concave saddle point problems and strongly monotone variational inequalities*, arXiv:2012.15595, (2020).

[54] Y. Ouyang and Y. Xu, *Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems*, Mathematical Programming, 185 (2021), pp. 1–35, https://doi.org/10.1007/s10107-019-01420-0, https://doi.org/10.1007/s10107-019-01420-0.

[55] B. Polyak, *A general method of solving extremum problems*, Soviet Mathematics Doklady, 8 (1967), pp. 593–597.

[56] A. Rogozin, A. Beznosikov, D. Dvinskikh, D. Kovalev, P. Dvurechensky, and A. Gasnikov, *Decentralized distributed optimization for saddle point problems*, arXiv:2102.07758, (2021).

[57] A. Sadiev, A. Beznosikov, P. Dvurechensky, and A. Gasnikov, *Zeroth-order algorithms for smooth saddle-point problems*, in Mathematical Optimization Theory and Operations Research: Recent Trends, A. Strekalovsky, Y. Kochetov, T. Gruzdeva, and A. Orlov, eds., Cham, 2021, Springer International Publishing, pp. 71–85. arXiv:2009.09908.

[58] I. Shibaev, P. Dvurechensky, and A. Gasnikov, *Zeroth-order methods for noisy Hölder-gradient functions*, Optimization Letters, (2021), https://doi.org/10.1007/s11590-021-01742-z. (accepted), arXiv:2006.11857.

[59] M. Solodov and B. Svaiter, *A hybrid approximate extragradient–proximal point algorithm using the enlargement of a maximal monotone operator*, Set-Valued Analysis, 7 (1999), pp. 323–345.

[60] F. Stonyakin, A. Tyurin, A. Gasnikov, P. Dvurechensky, A. Agafonov, D. Dvinskikh, M. Alkousa, D. Pasechnyuk, S. Artamonov, and V. Piskunova, *Inexact model: A framework for optimization and variational inequalities*, Optimization Methods and Software, (2021), https://doi.org/10.1080/10556788.2021.1924714. (accepted), WIAS Preprint No. 2709, arXiv:2001.09013, arXiv:1902.00990.

[61] F. S. Stonyakin, D. Dvinskikh, P. Dvurechensky, A. Kroshnin, O. Kuznetsova, A. Agafonov, A. Gasnikov, A. Tyurin, C. A. Uribe, D. Pasechnyuk, and S. Artamonov, *Gradient methods for problems with inexact model of the objective*, in Mathematical Optimization Theory and Operations Research, M. Khachay, Y. Kochetov, and P. Pardalos, eds., Cham, 2019, Springer International Publishing, pp. 97–114. arXiv:1902.09001.

[62] D. Tiapkin, A. Gasnikov, and P. Dvurechensky, *Stochastic saddle-point optimization for wasserstein barycenters*, arXiv:2006.06763, (2020).

[63] A. Titov, F. Stonyakin, M. Alkousa, and A. Gasnikov, *Algorithms for solving variational inequalities and saddle point problems with some generalizations of lipschitz property for operators*, in Mathematical Optimization Theory and Operations Research, A. Strekalovsky, Y. Kochetov, T. Gruzdeva, and A. Orlov, eds., Cham, 2021, Springer International Publishing, pp. 86–101.

[64] V. Tominin, Y. Tominin, E. Borodich, D. Kovalev, A. Gasnikov, and P. Dvurechensky, *On accelerated methods for saddle-point problems with composite structure*, arXiv:2103.09344, (2021).

[65] J. Zhang, M. Hong, and S. Zhang, *On lower iteration complexity bounds for the saddle point problems*, arXiv:1912.07481, (2019).

Appendix A

Proof of Lemma 1

Proof Let us fix some $\nu \in [0,1]$. Then, for any $x \in [0,1]$, $x^{2\nu} \leq 1$. On the other hand, for any $x \geq 1$, $x^{2\nu} \leq x^2$. Thus, for any $x \geq 0$, $x^{2\nu} \leq x^2 + 1$. Hence, for any $\alpha, \beta \geq 0$,

$$\alpha^\nu \beta \leq \frac{\alpha^{2\nu}}{2} + \frac{\beta^2}{2} \leq \frac{\alpha^2}{2} + \frac{\beta^2}{2} + \frac{1}{2}.$$

Substituting $\alpha = \frac{b a^{\frac{1}{1+\nu}}}{\delta^{\frac{1}{1+\nu}}}$ and $\beta = \frac{c a^{\frac{1}{1+\nu}}}{\delta^{\frac{1}{1+\nu}}}$, we obtain

$$\frac{b^\nu a^{\frac{\nu}{1+\nu}}}{\delta^{\frac{\nu}{1+\nu}}} \frac{c a^{\frac{1}{1+\nu}}}{\delta^{\frac{1}{1+\nu}}} \leq \frac{b^2 a^{\frac{2}{1+\nu}}}{2\delta^{\frac{2}{1+\nu}}} + \frac{c^2 a^{\frac{2}{1+\nu}}}{2\delta^{\frac{2}{1+\nu}}} + \frac{1}{2}$$

and

$$ab^\nu c \leq \left(\frac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}} \frac{a^{\frac{2}{1+\nu}}}{2}\left(b^2+c^2\right)+\frac{\delta}{2}.$$

□

Appendix B

To show the practical performance of the proposed Algorithm 1, we performed a series of numerical experiments for the Lagrange saddle point problem induced by the Fermat-Torricelli-Steiner problem.

All experiments were made using Python 3.4, on a computer with Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz, 1992 Mhz, 4 Core(s), 8 Logical Processor(s), and 8 GB RAM.

We consider an example of a variational inequality with a non-smooth, i.e., with $\nu = 0$, and non-strongly monotone operator. For this VI, the proposed universal method, due to its adaptivity to the smoothness level of the problem, works in practice with iteration complexity much smaller than the one predicted by the theory. This example is inspired by the well-known *Fermat-Torricelli-Steiner problem*, in which we add some non-smooth functional constraints. This problem can be solved by a switching subgradient scheme [55, 8] with complexity $O(1/\varepsilon^2)$, but as we will see, our method allows to obtain much faster convergence in practice than the one given by this bound.

More precisely, for a given set of N points $A_k \in \mathbb{R}^n, k = 1, ..., N$ consider the optimization problem

$$\min_{x\in Q}\left\{ f(x) := \sum_{k=1}^{N} \|x - A_k\|_2 \;\middle|\; \varphi_p(x) := \sum_{i=1}^{n} \alpha_{pi}|x_i| - 1 \leq 0\ ,\ p = 1, ..., m \right\},$$

where Q is a convex compact, α_{pi} are drawn from the standard normal distribution and then truncated to be positive. The corresponding Lagrange saddle point problem is defined as

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

As it was described in (6), this problem is equivalent to the variational inequality with monotone non-smooth operator

$$G(x,\lambda) = \begin{pmatrix} \nabla f(x) + \sum\limits_{p=1}^{m}\lambda_p\nabla\varphi_p(x), \\ (-\varphi_1(x), -\varphi_2(x), \ldots, -\varphi_m(x))^T \end{pmatrix}.$$

For simplicity, we assume that there exists (potentially very large) bound for the optimal Lagrange multiplier λ^*, which allows us to compactify the feasible set for the pair (x, λ) to be a Euclidean ball of some radius. We believe that the approach from [44, 14] to deal with unbounded feasible sets can be extended to our setting and we leave this for future work.

We run Algorithm 1 for different values of n, m, and N with standard Euclidean prox-setup and the starting point $(x^0, \lambda^0) = \frac{1}{\sqrt{m+n}}\mathbf{1} \in \mathbb{R}^{n+m}$, where $\mathbf{1}$ is the vector of all ones. The points A_k, $k = 1, ..., N$ are drawn randomly from the standard normal distribution. For each value of the parameters, the random data was drawn 10 times and the results were averaged. The results of the work of Algorithm 1 are represented in Fig. 1. For different values of the accuracy $\varepsilon \in \{1/2^i, i = 1, 2, 3, 4, 5, 6\}$, we report the number of iterations and the running time in seconds required by Algorithm 1 to reach an ε-solution of the considered problem.

As it is known [46], for a VI with a non-smooth operator, the theoretical iteration complexity estimate $O\left(\frac{1}{\varepsilon^2}\right)$ is optimal. However, experimentally we see from slope of the lines in Fig. 1 that, due to the adaptivity, the proposed Algorithm 1 has iteration complexity $O\left(\frac{1}{\sqrt[4]{\varepsilon}}\right)$.

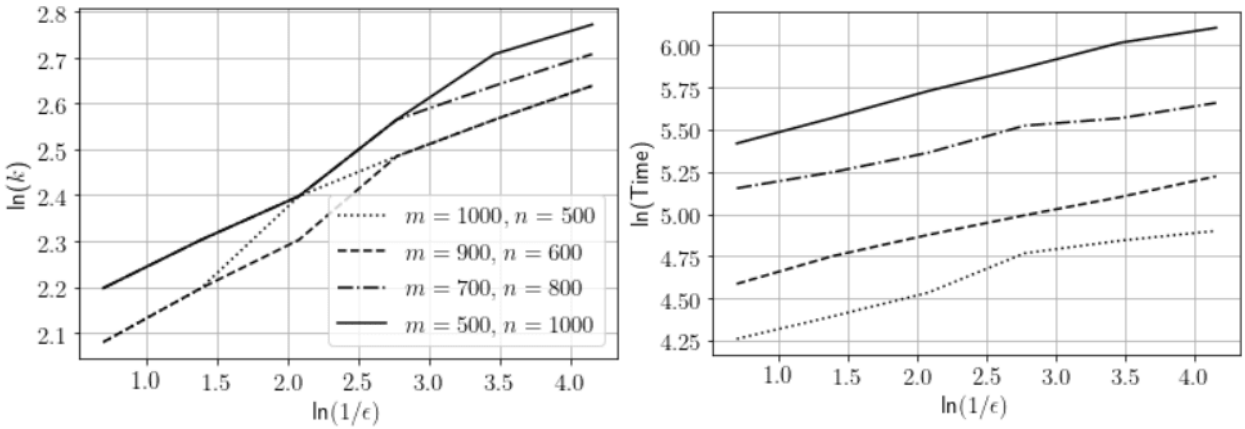

Figure 1: Results of Algorithm 1 for Fermat–Torricelli–Steiner problem with different values of m and n.

Appendix C

In this appendix, in order to demonstrate the performance of the Generalized Mirror Prox with restarts (Algorithm 2), we consider the variational inequality with Lipschitz-continuous strongly monotone operator (see Example 5.2 in [40])

$$g : Q \subset \mathbb{R}^n \to \mathbb{R}^n, \quad g(x) = x. \tag{35}$$

We compare the work of the proposed Algorithm 2 with Modified Projection Method, which was proposed in [40]. We run Algorithm 2 with different values of the accuracy $\varepsilon \in \{10^{-i}, i = 3, 4, \ldots, 10\}$ and for the dimension $n = 10^7$. We take $Q = \{x \in \mathbb{R}^n, \|x\|_2 \leq 2\}$. The results of the comparison are presented in Fig. 2, which illustrates the norm $\|x_{\text{out}} - x_*\|_2$, as a function of iterations, where x_{out} is the output of each algorithm, and x_* is the solution of the problem (1), for the operator (35). Note that $x_* = \mathbf{0} \in \mathbb{R}^n$. In the conducted experiments, at the

first, we run Algorithm 2, and calculate $\|x_{\text{out}} - x_*\|_2$ for the different previously mentioned values of ε and the corresponding number of iterations, resulted by the working of algorithm. For the calculated number of iteration of Algorithm 2, we rum Modified Projection Method and calculate the corresponding values $\|x_{\text{out}} - x_*\|_2$. From Fig. 2, we can see the higher efficiency of the proposed Algorithm 2, and the big difference between the results of the compared algorithms.

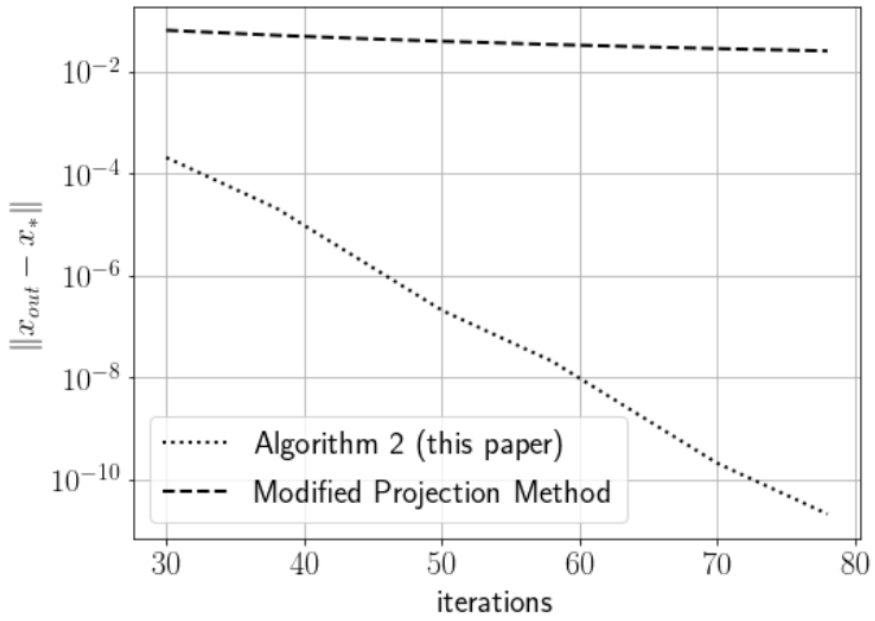

Figure 2: Results of Algorithm 2 and Modified Projection Method with $n = 10^7$.