Generalized Mirror Prox Algorithm for Variational Inequalities

Pavel Dvurechensky
Weierstrass Institute for Applied Analysis and Stochastics
pavel.dvurechensky@wias-berlin.de

Alexander Gasnikov
Moscow Institute of Physics and Technology,
Institute for Information Transmission Problems RAS
gasnikov@yandex.ru

Fedor Stonyakin
V.I. Vernadsky Crimean Federal University
fedyor@mail.ru

Alexander Titov
Moscow Institute of Physics and Technology
a.a.titov@phystech.edu

Abstract

Recently saddle-point problems and variational inequalities (VI's) gained a new interest in the deep learning community, e.g. in application to Generative Adversary Networks. We consider VI's with Hölder continuous monotone operator and propose a new algorithm which is universal for VI's with smooth, non-smooth and intermediate operator. This means that without knowing the Hölder parameter ν and Hölder constant L_ν our method has the best possible complexity for this class of VI's, namely our algorithm has complexity $O\left(\inf_{\nu\in[0,1]}\left(\frac{L_\nu}{\varepsilon}\right)^{\frac{2}{1+\nu}}R^2\right)$, where R is the size of the feasible set and ε is the desired accuracy of the solution. We also consider the case of VI's with strongly monotone operator and generalize our method for such VI's. [1]

1 Introduction

The main problem, we consider, is the following weak 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 $Q \subseteq E$ is a closed convex set and continuous operator $g(x): Q \to E^*$ is monotone

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

Under the assumption of continuity and monotonicity of the operator, this problem is equivalent to strong variational inequality, in which the goal is to find $x_* \in Q$ s.t. $\langle g(x_*), x_* - x\rangle \leq 0, \quad \forall x \in Q$. Variational inequalities with monotone operators are strongly connected with convex optimization problems and convex-concave saddle-point problems. In the former case, operator g is the subgradient of the objective, and in the latter case operator g is composed from partial subgradients of the objective in the saddle-point problem. Studying VI's is also important for equilibrium and complementarity

[1] The full text of the paper, including the case of VI's with inexactly given operator and application to saddle point problems, can be found as [7] at https://arxiv.org/abs/1806.05140

Abstract for: Smooth Games Optimization and Machine Learning Workshop (NIPS 2018), Montréal, Canada.

problems [12]. Recently saddle-point problems and variational inequalities became of interest in the deep learning community [10, 14].

Our focus here is on numerical methods for VI problems, their convergence rate and complexity estimates. Numerical methods for VI's are known since 1970's when the extragradient method was proposed in [13]. More recently, [16] proposed a non-Euclidean variant of this method, called Mirror Prox algorithm. Under the assumption of L_1-Lipschitz continuity of the operator g, i.e. g satisfies $\|g(x) - g(y)\|_* \le L_1\|x - y\|$, $x, y \in Q$, this method has complexity $O\left(\frac{L_1 R^2}{\varepsilon}\right)$, where R characterizes the diameter of the set Q, ε is the desired accuracy. By complexity we mean the number of iterations of an algorithm to find a point $\widehat{x} \in Q$ s.t.

$$\max_{u \in Q} \langle g(u), \widehat{x} - u\rangle \le \varepsilon.$$

Different methods with similar complexity were also proposed in [21, 1, 18, 15]. In [18], the author proposed a method for VI's with bounded variation of the operator g, i.e. g satisfying $\|g(x) - g(y)\|_* \le L_0$, $x, y \in Q$, and complexity $O\left(\frac{L_0^2 R^2}{\varepsilon^2}\right)$. He also raised 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 work is to propose such an algorithm.

To do so, we consider a more general class of operators g being Hölder continuous on Q, i.e., for some $\nu \in [0, 1]$ and $L_\nu \ge 0$,

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

This class covers both the case of g with bounded variation ($\nu = 0$) and Lipschitz continuous g ($\nu = 1$). Variational inequalities with Hölder continuous monotone operator were already considered in [16], where a special choice of the stepsize for the Mirror Prox algorithm leads to complexity

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

which is optimal for this class of problems, see [17]. [4] consider VI's with non-monotone Hölder continuous operator. Unfortunately, both papers use ν and L_ν to define the stepsize of their methods. This is in sharp contrast to optimization, where so called universal algorithms were proposed, which do not use the information about the Hölder class and Hölder constant, see [19, 9, 8, 2, 5, 20, 6, 11]. In this work, we close this gap and propose a universal method for VI's with Hölder continuous monotone operator.

To sum up, our contributions in this paper are as follows.

- We generalize Mirror Prox algorithm for VI's with Hölder continuous monotone operator and provide theoretical analysis of its convergence rate showing that it 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 about L_ν or ν.
- We generalize our algorithm for the case of μ-strongly monotone operator g 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 $\widehat{x} \in Q$ s.t. $\|\widehat{x} - x_*\| \le \varepsilon$.

2 Main results

2.1 Monotone operator

We start with the general notation 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 $g \in E^*$ at $x \in E$ by $\langle g, x\rangle$. Let $\|\cdot\|$ be some norm on E, $\|\cdot\|_*$ be its dual, defined by $\|g\|_* = \max\limits_x \big\{\langle g, x\rangle, \|x\| \le 1\big\}$.

We choose a *prox-function* $d(x)$, which is 1-strongly convex on Q with respect to $\|\cdot\|$, i.e., for any $x, y \in Q$ $d(y) - d(x) - \langle \nabla d(x), y - x\rangle \geq \frac{1}{2}\|y - x\|^2$. 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 [3].

Algorithm 1 Generalized Mirror Prox

Input: accuracy $\varepsilon > 0$, initial guess $M_{-1} > 0$, prox-setup: $d(x)$, $V[z](x)$.

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

for $k = 0, 1, ...$ **do**

Set $i_k = 0$.

repeat

Set $M_k = 2^{i_k - 1} M_{k-1}$.

Calculate

$$w_k = \arg\min_{x\in Q}\left\{\langle g(z_k), x\rangle + M_k V[z_k](x)\right\}. \tag{3}$$

Calculate

$$z_{k+1} = \arg\min_{x\in Q}\left\{\langle g(w_k), x\rangle + M_k V[z_k](x)\right\}. \tag{4}$$

$i_k = i_k + 1$.

until

$$\langle g(w_k) - g(z_k), w_k - z_{k+1}\rangle \leq \frac{M_k}{2}\left(\|w_k - z_k\|^2 + \|w_k - z_{k+1}\|^2\right) + \frac{\varepsilon}{2}. \tag{5}$$

Set $k = k + 1$.

end for

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

Theorem 1 (Universal Method for VI)**.** *Assume that the operator g is Hölder continuous with constant L_ν for some $\nu \in [0, 1]$ and $M_{-1} \leq \left(\frac{2}{\varepsilon}\right)^{\frac{1-\nu}{1+\nu}} L_\nu^{\frac{2}{1+\nu}}$. Also assume that the set Q is bounded. Then, for all $k \geq 0$, we have*

$$\max_{u\in Q}\langle g(u), \widehat{w}_k - u\rangle \leq \frac{(2L_\nu)^{\frac{2}{1+\nu}}}{k\varepsilon^{\frac{1-\nu}{1+\nu}}} \max_{u\in Q} V[z_0](u) + \frac{\varepsilon}{2}. \tag{6}$$

Since the algorithm does not use the values of parameters ν and L_ν, we obtain the following iteration complexity bound to achieve $\max_{u\in Q}\langle g(u), \widehat{w}_k - u\rangle \leq \varepsilon$

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

Using the same reasoning as in [19], we estimate the number of oracle calls for Algorithm 1. The number of oracle calls on each iteration k is equal to $2i_k$. At the same time, $M_k = 2^{i_k - 2} M_{k-1}$ and, hence, $i_k = 2 + \log_2 \frac{M_k}{M_{k-1}}$. Thus, the total number of oracle calls is

$$\sum_{j=0}^{k-1} i_j = 4k + 2\sum_{i=0}^{k-1} \log_2 \frac{M_j}{M_{j-1}} < 4k + 2\log_2\left(2L\left(\frac{\varepsilon}{2}\right)\right) - 2\log_2(M_{-1}), \tag{7}$$

where we used that $M_k \leq 2L(\frac{\varepsilon}{2})$.

Thus, the number of oracle calls of Algorithm 1 does not exceed:

$$4 \inf_{\nu\in[0,1]} \left(\frac{2\cdot L_\nu}{\varepsilon}\right)^{\frac{2}{1+\nu}} \cdot \max_{u\in Q} V[z_0](u) + 2 \inf_{\nu\in[0,1]} \log_2 2\left(\left(\frac{2}{\varepsilon}\right)^{\frac{1-\nu}{1+\nu}} L_\nu^{\frac{2}{1+\nu}}\right) - 2\log_2(M_{-1}).$$

2.2 Strongly monotone operator

Now we assume, that g in (1) is a strongly monotone operator, which means that, for some $\mu > 0$,

$$\langle g(x) - g(y), x - y \rangle \geq \mu \|x - y\|^2 \quad \forall x, y \in Q. \tag{8}$$

We slightly modify the assumptions on prox-function $d(x)$. 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

$$d(x) \leq \frac{\Omega}{2}, \quad \forall x \in Q : \|x\| \leq 1, \tag{9}$$

where Ω is some known constant. Note that for standard proximal setups, $\Omega = O(\ln \dim E)$. 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).

Algorithm 2 Restarted Generalized Mirror Prox

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

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

2: **repeat**

3: Set x_{p+1} as the output of Algorithm 1 for monotone case with accuracy $\mu\varepsilon/2$, prox-function $d_p(\cdot)$ and stopping criterion $\sum_{i=0}^{k-1} M_i^{-1} \geq \frac{\Omega}{\mu}$.

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

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

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 $d(x)$ satisfies* (9) *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* $p \geq 0$

$$\|x_p - x_*\|^2 \leq R_0^2 \cdot 2^{-p} + \frac{\varepsilon}{2}$$

and the point x_p returned by Algorithm 2 satisfies $\|x_p - x_*\|^2 \leq \varepsilon$.

Theorem 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 s.t. $\|x_p - x_*\|^2 \leq \varepsilon$ and the total number of iterations of the inner Algorithm 1 does not exceed*

$$\inf_{\nu \in [0,1]} \left\lceil \left(\frac{L_\nu}{\mu}\right)^{\frac{2}{1+\nu}} \frac{2^{\frac{3-\nu}{1+\nu}} \Omega}{\varepsilon^{\frac{1-\nu}{1+\nu}}} \cdot \log_2 \frac{2R_0^2}{\varepsilon} \right\rceil. \tag{10}$$

References

[1] A. Auslender and M. Teboulle. Interior projection-like methods for monotone variational inequalities. *Mathematical programming*, 104(1):39–68, 2005.

[2] D. Baimurzina, A. Gasnikov, E. Gasnikova, P. Dvurechensky, E. Ershov, M. Kubentaeva, and A. Lagunovskaya. Universal similar triangulars method for searching equilibriums in traffic flow distribution models. *arXiv:1701.02473*, 2017.

[3] A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization (Lecture Notes)*. Personal web-page of A. Nemirovski, 2015.

[4] 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(2):277–310, 2015.

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

[6] P. Dvurechensky, A. Gasnikov, and D. Kamzolov. Universal intermediate gradient method for convex problems with inexact oracle. *arXiv:1712.06036*, 2017.

[7] P. Dvurechensky, A. Gasnikov, F. Stonyakin, and A. Titov. Generalized Mirror Prox: Solving variational inequalities with monotone operator, inexact oracle, and unknown Hölder parameters. *arXiv:1806.05140*, 2018.

[8] A. V. Gasnikov and Y. E. Nesterov. Universal method for stochastic composite optimization problems. *Computational Mathematics and Mathematical Physics*, 58(1):48–64, 2018. First appeared in arXiv:1604.05275.

[9] S. Ghadimi, G. Lan, and H. Zhang. Generalized uniformly optimal methods for nonlinear programming. *arXiv:1508.07384*, 2015.

[10] G. Gidel, H. Berard, P. Vincent, and S. Lacoste-Julien. A variational inequality perspective on generative adversarial networks. *arXiv:1802.10551*, 2018.

[11] S. Guminov, A. Gasnikov, A. Anikin, and A. Gornov. A universal modification of the linear coupling method. *arXiv:1711.01850*, 2017.

[12] 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(1-3):161–220, 1990.

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

[14] L. Mescheder, S. Nowozin, and A. Geiger. The numerics of gans. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1825–1835. Curran Associates, Inc., 2017.

[15] 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(6):2755–2787, 2010.

[16] 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(1):229–251, 2004.

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

[18] Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007. First appeared in 2003 as CORE discussion paper 2003/68.

[19] Y. Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1):381–404, 2015.

[20] V. Roulet and A. d' Aspremont. Sharpness, restart and acceleration. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1119–1129. Curran Associates, Inc., 2017. arXiv:1702.03828.

[21] M. Solodov and B. Svaiter. A hybrid approximate extragradient–proximal point algorithm using the enlargement of a maximal monotone operator. *Set-Valued Analysis*, 7(4):323–345, 1999.