Titov A.

Universal Proximal Method for Variational Inequalities

co-authors:

Stonyakin F.
Gasnikov A.
Dvurechensky P.
Alkousa M.

June 2018

1. Variational Inequalities: Problem Statement

For some operator $g : Q \to \mathbb{R}^n$, defined on a convex compact set $Q \subset \mathbb{R}^n$ we consider *variational inequalities*:

$$\langle g(x_*),\ x_* - x\rangle \leqslant 0, \tag{1}$$

We need to find such point $x_* \in Q$ (solution of the VI), that

$$\max_{x \in Q} \langle g(x_*),\ x_* - x\rangle \leqslant 0. \tag{2}$$

Aim: to extend the ideology of universal gradient methods to the formulation of the problem of solving the VI

Yu. Nesterov, Universal gradient methods for convex optimization problems, // Math. Program., 2015, Ser. A, 152, pp. 381–404. Available at https://doi.org/10.1007/s10107-014-0790-0**.**

Examples of arising of Variational Inequalities

1. The problem of convex (differentiable and non-differentiable) conditional and unconditional optimization

$$g(x) = \nabla f(x), \quad x \in Q. \tag{3}$$

2. Finding saddle points of a convex-concave functional:

$$f(x_*, y) \leqslant f(x_*, y_*) \leqslant f(x, y_*) \quad \forall x \in Q_x, y \in Q_y. \tag{4}$$

3. The problem of finding a fixed point of the projection map:

$$P(x) = \pi_Q(x - \alpha g(x)), \quad \alpha > 0. \tag{5}$$

Applications: game theory, mathematical programming problems.

Works concerning methods for solving VI

1. F. Facchinei, J. S. Pang, Finite-Dimensional Variational Inequality and Complementarity Problems, vols. 1 and 2, // Springer-Verlag, New York, 2003.
2. F. Yousefian, A. Nedic and U. V. Shanbhag Optimal robust smoothing extragradient algorithms for stochastic variational inequality problems // https://arxiv.org/pdf/1403.5591.pdf
3. Antipin A. S., Acimovic V., Acimovic M. Dynamics and variational inequality // Zh. Vychisl. mod. and mod. NAT., 57: 5 (2017), 783-800; Comput. Math. Math. Phys., 57: 5 (2017), Pp. 784-801.
4. 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, 2004, 15, pp. 229–251.
5. Yu. E. 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.

Let's call the operator $g : Q \to \mathbb{R}^n$ Hölder-continuous on Q, if:

$$\|f(x) - f(y)\|_* \leqslant L_\nu \|x - y\|^\nu \quad \forall\, x, y \in Q, \tag{6}$$

Some results

1. We suggested universal method for VI's with Hölder continuous monotone operator with complexity

$$O\left(\inf_{\nu \in [0;1]} \left(\frac{L_\nu}{\varepsilon}\right)^{\frac{2}{1+\nu}}\right) R^2 \tag{7}$$

2. We show, how our method can be applied to convex-concave saddle-point problems.

2. The basic idea goes back to the universal gradient descent

Yu. Nesterov, Universal gradient methods for convex optimization problems, // Math. Program., 2015, Ser. A, 152, pp. 381–404. Available at urlhttps://doi.org/10.1007/s10107-014-0790-0.

A.V. Gasnikov. Modern numerical optimization methods. The universal gradient descent method. // M.: MIPT: 2018.

Problem Statement: $f^* = \min\limits_{x \in Q} f(x)$

$$\|\nabla f(x) - \nabla f(y)\|_* \leqslant L_\nu \|x - y\|^\nu \quad \forall\, x, y \in Q, \tag{8}$$

where $L_\nu < +\infty$ for some $\nu \in [0; 1]$.

Algorithm 1: UGD (Yu. E. Nesterov, 2015)

(a) $N := N + 1$, $L^{N+1} := L^N/2$.

(b) Calculate

$$x^{N+1} := \arg\min_{x \in Q} \left\{ f(x^N) + \left\langle \nabla f(x^N), x - x^N \right\rangle + L^{N+1} V(x, x^N) \right\}.$$

(c) If

$$f(x^{N+1}) \leqslant f(x^N) + \left\langle \nabla f(x^N), x^{N+1} - x^N \right\rangle + L^{N+1} V(x^{N+1}, x^N) + \frac{\varepsilon}{2}, \tag{9}$$

then proceed to the next iteration (step 1).

(d) If (9) is not satisfied, then increase L^{N+1} 2 times: $L^{N+1} := 2L^{N+1}$ and proceed to step 2.

(e) Stopping Criterion of the method:

$$\sum_{k=0}^{N-1} \frac{1}{L^{k+1}} \geqslant \frac{2R^2}{\varepsilon}, \tag{10}$$

where $R^2 = \max_{x \in Q} V(x, x^0)$.

Theorem 1. *For the universal gradient descent after*

$$N = \inf_{\nu \in [0,1]} \left(\frac{2L_\nu R^{1+\nu}}{\varepsilon} \right)^{\frac{2}{1+\nu}} \tag{11}$$

iterations the following inequation holds:

$$f(\overline{x}^N) - f^* \leqslant \varepsilon, \tag{12}$$

where

$$\overline{x}^N = \frac{1}{\sum\limits_{k=0}^{N} 1/L^k} \sum_{k=0}^{N} \frac{x^k}{L^k}. \tag{13}$$

3. Universal method for VI

A. V. Gasnikov. Modern numerical optimization methods. The universal gradient descent method. // M.: MIPT: 2018.

A.V. Gasnikov, P. E. Dvurechensky, F. S. Stonyakin, A. A. Titov. Generalized Mirror Prox: Solving Variational Inequalities with Monotone Operator, Inexact Oracle, and Unknown Hölder Parameters // arXiv:1806.05140v1 [math.OC] 13 Jun 2018

Basic ideas:

(a) 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, 2004, 15, pp. 229–251.

(b) Yu. Nesterov, Universal gradient methods for convex optimization problems, // Math. Program., 2015, Ser. A, 152, pp. 381–404. Available at https://doi.org/10.1007/s10107-014-0790-0.

We assume that the set Q is convex and compact in the space $\mathbb{R}^n$ with the norm $\|\cdot\|$ (generally speaking, not Euclidean), and $\|\cdot\|_*$ - conjugate to $\|\cdot\|$ norm.

Consider a 1-strongly convex function d relatively to $\|\cdot\|$, which is differentiable at all points $x \in Q$ and the corresponding d *divergence of the Bregman*

$$V(x,y) = d(x) - d(y) - \langle \nabla d(y), x - y \rangle \forall x, y \in Q, \tag{14}$$

where $\langle \cdot, \cdot \rangle$ - inner product in $\mathbb{R}^n$.

Examples

$$a = \frac{2 \cdot \ln n}{2 \cdot \ln n - 1}$$

$Q = B_p^n(1)$	$1 \le p \le a$	$a \le p \le 2$	$2 \le p \le \infty$
$\lVert \ \rVert$	$\lVert \ \rVert_1$	$\lVert \ \rVert_p$	$\lVert \ \rVert_2$
$d(x)$	$d(x) = \frac{1}{2(a-1)} \lVert x \rVert_a^2$	$d(x) = \frac{1}{2(p-1)} \lVert x \rVert_p^2$	$\frac{1}{2} \lVert x \rVert_2^2$
R^2	$O(\ln n)$	$O\left((p-1)^{-1}\right)$	$O(1)$

Table 1. "Prox-functions and divergences of Bregman"

4. Proximal method of A. S. Nemirovsky

Consider

$$\|g(x) - g(y)\|_* \leqslant L_\nu \|x - y\|^\nu \quad \forall x, y \in Q, \tag{15}$$

where $L_\nu < \infty$ and $\nu \in [0, 1]$, fixed number $\varepsilon > 0$ (accuracy of solution) and the initial approximation $x^0 = \arg\min_{x \in Q} d(x) \in Q$. Let's describe $(N + 1)$st iteration $(N = 0, 1, 2, \ldots)$, putting initially $N := 0$.

Algorithm 2: A. S. Nemirovsky method for VI, 2004

1. $N := N + 1$,
2. Calculate

$$y^{N+1} := \arg\min_{x\in Q}\left\{\gamma_{N+1}\cdot\left\langle g(x^N), x - x^N\right\rangle + V(x,x^N)\right\},$$

$$x^{N+1} := \arg\min_{x\in Q}\left\{\gamma_{N+1}\cdot\left\langle g(y^{N+1}), x - x^N\right\rangle + V(x,x^N)\right\},$$

where

$$\gamma_{N+1} = 0.7L_\nu^{-1}\left(\frac{R^2}{N+1}\right)^{\frac{1-\nu}{2}}\varepsilon^{\frac{1+\nu}{2}}, \quad R^2 = \max_{x\in Q} V(x,x^0).$$

3.Output:

$$\widetilde{y} = \left(\sum_\tau \gamma_\tau\right)^{-1}\sum_\tau \gamma_\tau y_\tau.$$

Theorem 2. *After N iterations of working of the Algorithm 2 the next inequality takes place:*

$$\max_{x\in Q}\langle g(x), \widetilde{y} - x\rangle \leqslant O(1)L_\nu\left(\frac{R^2}{\varepsilon\cdot(N+1)}\right)^{\frac{1+\nu}{2}} \tag{16}$$

with the absolute constant $O(1)$.

5. Universal method for VI

Let's describe the $(N+1)$-t iteration of the proposed algorithm ($N = 0, 1, 2, \ldots$), putting initially $N := 0$.

Algorithm 3: APMD

---

1. $N := N+1$, $L^{N+1} := L^N/2$.
2. Calculate

$$y^{N+1} := \arg\min_{x\in Q} \left\{ \left\langle g(x^N), x - x^N \right\rangle + L^{N+1} V(x, x^N) \right\},$$

$$x^{N+1} := \arg\min_{x\in Q} \left\{ \left\langle g(y^{N+1}), x - x^N \right\rangle + L^{N+1} V(x, x^N) \right\}.$$

3. If $\left\langle g(y^{N+1}) - g(x^N), y^{N+1} - x^{N+1} \right\rangle \leqslant$

$$\leqslant L^{N+1} V(y^{N+1}, x^N) + L^{N+1} V(x^{N+1}, y^{N+1}) + \frac{\varepsilon}{2}, \qquad (17)$$

then proceed to the next iteration (step 1).

4. If (17) is not satisfied, then increase L^{N+1} 2 times: $L^{N+1} := 2L^{N+1}$ and proceed to step 2.
5. The stopping criterion of the method is $\sum\limits_{k=0}^{N-1} \frac{1}{L^{k+1}} \geqslant \frac{2R^2}{\varepsilon}$, where $R^2 = \max\limits_{x\in Q} V(x, x^0)$.

Let's define $S_N = \sum\limits_{k=0}^{N-1} \frac{1}{L^{k+1}}$.

Theorem 3. *After stopping of the algorithm 3 the inequality takes place:*

$$\frac{1}{S_N}\sum_{k=0}^{N-1}\frac{1}{L^{k+1}}\left\langle g(y^{k+1}), y^{k+1}-x\right\rangle \leqslant \frac{V(x,x^0)-V(x^N)}{S_N}+\frac{\varepsilon}{2}. \quad (18)$$

Theorem 4. *If the Algorithm 3 works*

$$N=\left\lceil \inf_{\nu\in[0;1]}\left(\frac{2L_\nu R^{1+\nu}}{\varepsilon}\right)^{\frac{2}{1+\nu}}\right\rceil \quad (19)$$

iterations, conditions

$$\|g(x)-g(y)\|_* \leqslant L_\nu\|x-y\|^\nu \quad (20)$$

with $\nu\in[0;1]$, $x,y\in Q$, *and* $L_0<+\infty$ *(other constants* L_ν *may be infinite) and*

$$L^0\leqslant 2L=2\inf_{\nu\in[0;1]}L_\nu\cdot\left(\frac{2L_\nu}{\varepsilon}\right)^{\frac{1-\nu}{1+\nu}} \quad (21)$$

take place, then

$$\max_{x\in Q}\min_{k=\overline{0,N-1}}\left\{\left\langle g(y^{k+1}), y^{k+1}-x\right\rangle\right\}\leqslant\frac{1}{S_n}\max_{x\in Q}\sum_{k=0}^{N-1}\frac{1}{L^{k+1}}\left\langle g(y^{k+1}), y^{k+1}-x\right\rangle\leqslant\varepsilon. \quad (22)$$

Remark 1. For VI with monotone operator:

$$\langle g(x), y^{k+1} - x\rangle \leqslant \left\langle g(y^{k+1}), y^{k+1} - x\right\rangle,$$

so the criterion (22) can be replaced by

$$\max_{x\in Q} \langle g(x), \widetilde{y} - x\rangle \leqslant \varepsilon, \text{ where } \widetilde{y} = \frac{\sum\limits_{k=0}^{N-1} \frac{1}{L^{k+1}} \cdot y^{k+1}}{S_N}. \tag{23}$$

Often (23) is used as a criterion for the quality of the solution of the variational inequality.

Remark 2. Let i_k be the number of oracle calls on the iteration N. The total number of oracle calls during the work of the Algorithm 3 satisfies:

$$\sum_{N=0}^{N-1} i_N < 4N + 2\log_2\left(2L\left(\frac{\varepsilon}{2}\right)\right) - 2\log_2(L_0) \tag{24}$$

Remark 3. If $g \not\equiv Const$, the inequality $L^0 \leqslant 2L$ can be reached by choosing

$$L^0 := \frac{\|g(x) - g(y)\|_*}{\|x - y\|} \ \text{при} \ \ g(x) \neq g(y).$$

Remark 4. The estimate (19) of the number of iterations is optimal for variational inequalities.

In the case of the monotone operator g

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

one can also consider *weak variational inequalities*

$$\langle g(x), \ x_* - x \rangle \leqslant 0. \tag{25}$$

Usually in (25) it is necessary to find $x_* \in Q$ for which (25) is true for all $x \in Q$.

6. Some numerical experiments

$$g(x_1, x_2, \ldots, x_N) = \left(e^{x_1 + \frac{x_2}{e^3}}, e^{x_2 + \frac{x_3}{e^3}}, \ldots, e^{x_N + \frac{x_1}{e^3}}\right) \tag{26}$$

Apply the method to the variational inequality for the operator g of (26). For the set Q we choose a unit ball with center at zero

$$Q = B_1(0) = \{x = (x_1, x_2, ..., x_N) : \; x_1^2 + x_2^2 + ... + x_N^2 \leq 1\}$$

for the standard Euclidean norm in $\mathbb{R}^N$.

For both experiments $N = 1000$.

ε	N	Time
10^{-1}	3	16
$5 \cdot 10^{-2}$	4	23
10^{-2}	6	33
$5 \cdot 10^{-3}$	7	40
10^{-3}	9	50
$5 \cdot 10^{-4}$	10	56
10^{-4}	13	73
$5 \cdot 10^{-5}$	14	80
10^{-5}	16	95
$5 \cdot 10^{-6}$	17	98

Table 2. "Results for the proposed method"

ε	N	Time
10^{-1}	2	13
$5 \cdot 10^{-2}$	2	18
10^{-2}	3	22
$5 \cdot 10^{-3}$	3	25
10^{-3}	4	27
$5 \cdot 10^{-4}$	4	27
10^{-4}	5	30
$5 \cdot 10^{-5}$	5	35
10^{-5}	5	35
$5 \cdot 10^{-6}$	6	39

Table 3. "Results for the proposed method in case, if step 1 of the Algorithm 3 is defined as follows: $L^{N+1} := L^N/16$"

The method works faster then theoretical estimate $O(1/\varepsilon)$.

7. Saddle point problems

The method we consider for VI can be applied to finding saddle points:

$$f^* = \min_{u \in Q_1} \max_{v \in Q_2} f(u, v), \tag{27}$$

where $Q_1 \subseteq E_1$ and $Q_2 \subseteq E_2$ - convex and closed subsets of normed spaces $E_{1,2}$ with norms $\|\cdot\|_{1,2}$ respectively, the function f is convex by u and concave by v. Consider the operator from the gradients (derivatives) of f on u and v:

$$g(x) := \begin{pmatrix} \nabla_u f(u, v) \\ -\nabla_v f(u, v) \end{pmatrix}, \quad x = (u, v) \in Q = Q_1 \times Q_2, \tag{28}$$

where $Q = \{(u, v) \mid u \in Q_1, v \in Q_2\} \subset E_1 \times E_2$ — the direct product of the spaces E_1 and E_2 with the norm

$$\|x\| = \max\{\|u\|_1, \|v\|_2\}, \quad x = (u, v) \in E_1 \times E_2. \tag{29}$$

In the conjugated to $E = E_1 \times E_2$ space consider the norm:

$$\|s\| = \|s^{(1)}\|_{1,*} + \|s^{(2)}\|_{2,*}, \quad s = (s^{(1)}, s^{(2)}) \in E^*, \tag{30}$$

where $\|\cdot\|_{1,*}$ and $\|\cdot\|_{2,*}$ — norms of conjugated spaces E_1^* and E_1^* respectively.

Due to the convexity of f at u and concavity in v, the operator g is monotone:

$$\langle g(x) - g(y), x - y\rangle \geqslant 0 \quad \forall x, y \in Q \subset E, \tag{31}$$

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

The problem (27) under these assumptions brings us to the problem of finding a solution to VI.

Theorem 5. *Let for f and fixed $\nu \in [0, 1]$ there exist $L_{11,\nu}, L_{12,\nu}, L_{21,\nu}, L_{22,\nu} \leqslant +\infty$:*

$$\|\nabla_u f(u + \Delta u, v + \Delta v) - \nabla_u f(u, v)\|_{1,*} \leq L_{11,\nu}\|\Delta u\|_1^\nu + L_{12,\nu}\|\Delta v\|_2^\nu, \tag{32}$$

$$\|\nabla_v f(u + \Delta u, v + \Delta v) - \nabla_v f(u, v)\|_{2,*} \leq L_{21,\nu}\|\Delta u\|_1^\nu + L_{22,\nu}\|\Delta v\|_2^\nu \tag{33}$$

for all $u, u + \Delta u \in Q_1, v, v + \Delta v \in Q_2, \nu \in [0, 1]$. Then the operator g of (28) satisfies the Hölder condition with ν.

Remark 5. If the set Q is bounded, one can consider different level of smoothness in Theorem 5. Assume that for some numbers $\nu_{11}, \nu_{12}, \nu_{21}, \nu_{22} \in [0; 1]$:

$$\|\nabla_u f(u + \Delta u, v + \Delta v) - \nabla_u f(u, v)\|_{1,*} \leq \widehat{L_{11}} \|\Delta u\|_1^{\nu_{11}} + \widehat{L_{12}} \|\Delta v\|_2^{\nu_{12}}, \tag{34}$$

$$\|\nabla_v f(u + \Delta u, v + \Delta v) - \nabla_v f(u, v)\|_{2,*} \leq \widehat{L_{21}} \|\Delta u\|_1^{\nu_{21}} + \widehat{L_{22}} \|\Delta v\|_2^{\nu_{22}} \tag{35}$$

for all $u, u + \Delta u \in Q_1, v, v + \Delta v \in Q_2$. Then the statement of Theorem 5 holds for $\nu = \min\{\nu_{11}, \nu_{12}, \nu_{21}, \nu_{22}\}$.

After the algorithm 3 stops for the corresponding VI:

$$\widetilde{y} := \frac{1}{S_N} \sum_{k=0}^{N+1} \frac{y^{k+1}}{L^{k+1}}. \tag{36}$$

Theorem 6. *For $\widetilde{y} = (\widetilde{u}, \widetilde{v})$ the inequality holds:*

$$\max_{v \in Q_2} f(\widetilde{u}, v) - \min_{u \in Q_1} f(u, \widetilde{v}) \leqslant \varepsilon \tag{37}$$

for the desired saddle point $(u_*, v_*) \in Q$:

$$\max_{v \in Q_2} f(u_*, v) = \min_{u \in Q_1} f(u, v_*). \tag{38}$$

Thank you for attention!

Optimality of convex minimization methods

Nemirovsky A. S., Yudin D. B. Complexity of problems and efficiency of optimization methods. M.: 1979.

Bakhvalov-Nemirovsky Optimality: the number of calls to the gradient calculation can be reduced on the whole class of problems only by a numerical factor independent from the problem parameters and the dimension of the space.

Let's say that $f \in C^{1,\nu}(\mathbb{R}^n)$, $\nu \in [0;1]$ if

$$\|\nabla f(x) - \nabla f(y)\|_2 \leqslant L_\nu \|x - y\|_2^\nu \ \forall x, y \in \mathbb{R}^n. \tag{39}$$

Theorem 7. *Achieving ε - exact solution is possible for at least*

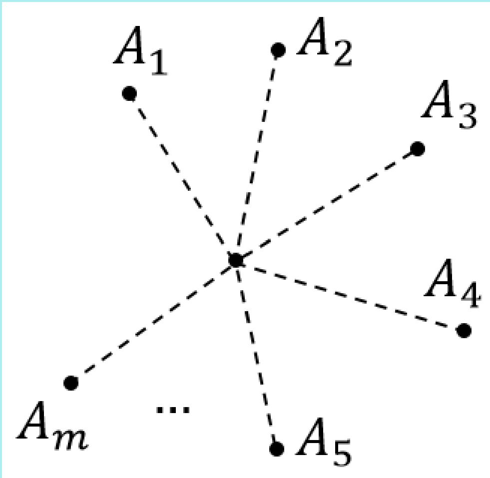
$$O\left(\left(\frac{L\nu R^{1+\nu}}{\varepsilon}\right)^{\frac{2}{1+3\nu}}\right) \tag{40}$$

iterations, where $R = \|x^0 - x_*\|$ (x^0 - initial approximation).

Special cases:

1. $\nu = 0$, i.e. $f \in \mathrm{Lip}: \ O\left(\frac{L_0^2 R^2}{\varepsilon^2}\right)$;
2. $\nu = 1$, i.e. $\nabla f \in \mathrm{Lip}: \ O\left(\frac{L_1^{1/2} \nu R}{\varepsilon^{1/2}}\right)$.

The Fermat-Torricelli-Steiner problem



Pic. 1.

$$\min_{x \in Q} f(x) \stackrel{\text{def}}{=} \sum_{i=1}^{m} \|x - a_i\| \in C^{1,0}(\mathbb{R}^n). \tag{41}$$

Eps	FGM$_{Euclid}$			PGM$_{Euclid}$		
	Iter	Gap	Lip	Iter	Gap	Lip
2^{-5}	205	$3.1 \cdot 10^{-2}$	$2.6 \cdot 10^{2}$	9,925	$3.1 \cdot 10^{-2}$	$2.6 \cdot 10^{2}$
2^{-6}	307	$1.5 \cdot 10^{-2}$	$5.1 \cdot 10^{2}$	19,895	$1.5 \cdot 10^{-2}$	$5.1 \cdot 10^{2}$
2^{-7}	277	$6.8 \cdot 10^{-3}$	$2.6 \cdot 10^{2}$	39,803	$7.8 \cdot 10^{-3}$	$2.6 \cdot 10^{2}$
2^{-8}	611	$3.9 \cdot 10^{-3}$	$5.1 \cdot 10^{2}$	77,138	$3.9 \cdot 10^{-3}$	$5.1 \cdot 10^{2}$
2^{-9}	827	$1.9 \cdot 10^{-3}$	$5.1 \cdot 10^{2}$	155,038	$2.0 \cdot 10^{-3}$	$2.6 \cdot 10^{2}$
2^{-10}	1,226	$9.8 \cdot 10^{-4}$	$2.6 \cdot 10^{2}$		out of time	
2^{-11}	1,655	$4.8 \cdot 10^{-4}$	$2.6 \cdot 10^{2}$			
2^{-12}	2,385	$2.4 \cdot 10^{-4}$	$5.1 \cdot 10^{2}$			
2^{-13}	3,388	$1.2 \cdot 10^{-4}$	$5.1 \cdot 10^{2}$			

Universal method works faster than theoretical estimates