Extragradient Sliding for Composite Non-Monotone Variational Inequalities

Roman Emelyanov¹, Andrey Tikhomirov¹, Aleksandr Beznosikov^{1,2,3}, and Alexander Gasnikov^{3,1,2}

¹ Moscow Institute of Physics and Technology, Moscow, Russian Federation

Institute for Information Transmission Problems RAS, Moscow, Russian Federation

³ Innopolis University, Innopolis, Russian Federation

Abstract. Variational inequalities offer a versatile and straightforward approach to analyzing a broad range of equilibrium problems in both theoretical and practical fields. In this paper, we consider a composite generally non-monotone variational inequality represented as a sum of L_q -Lipschitz monotone and L_p -Lipschitz generally non-monotone operators. We applied a special sliding version of the classical Extragradient method to this problem and obtain better convergence results. In particular, to achieve ε -accuracy of the solution, the oracle complexity of the nonmonotone operator Q for our algorithm is $\mathcal{O}\left(L_p^2/\varepsilon^2\right)$ in contrast to the basic Extragradient algorithm with $\mathcal{O}\left((L_p + L_q)^2/\varepsilon^2\right)$. The results of numerical experiments confirm the theoretical findings and show the superiority of the proposed method.

Keywords: Variational inequality \cdot Extragradient \cdot Composite problem \cdot Sliding \cdot Minty assumption

1 Introduction

Variational inequalities (VIs) are a popular class of optimization problems, which despite its relative youth has an extensive history of research, both in terms of different formulations, and of effective methods of their solution. The variational inequalities paradigm has gained particular popularity due to its generality and its ability to describe and represent various optimization problems in a unified way [1,2]. Nowadays VI problems can be found in a large number of fields from economics, game theory, physics and modelling of transport flows to machine learning and rapidly developing deep learning [3,4,5]. For instance, the development of optimization methods aimed at solving variational inequalities has recently attracted the attention of many researchers in the context of optimizing loss functions of generative adversarial networks [6] and reinforcement learning [7].

The most straightforward and widely known method for dealing with problems posed as variational inequalities is Gradient Descent-Ascent [8,9,10] developed by analogy with methods of optimization of a single objective [11]. Its serious drawbacks [12], which include weaker convergence compared to ordinary gradient 2

descent and the problem of rotation around the optimum [11,12] led researchers to create more advanced methods. One of the most well-known of these algorithms is Extragradient, proposed by G. Korpelevich [13]. Over time, the research around this method evolved and there are now various modifications (e.g. Optimistic Gradient with one oracle call per iteration [14,15]) and generalizations to an arbitrary Bregman setup (e.g. Mirror-Prox [16]). Moreover, for the monotone variational inequalities, the optimality of the Extragradient method was also shown. Meanwhile, there is a large body of theoretical studies related to various kinds of relaxations of monotonicity and transition to non-monotone operators [17,18,19,20,21]. The aim of this paper is also to delve into the non-monotone setting, but to look deeper into it and consider composite operators, i.e., operators that are represented as the sum of two operators. It seems that using the additional structure of the target problem can give improvements in terms of convergence theory. For monotone and strongly monotone operators such approaches already exist and indeed give theoretical and practical improvements [22,23,24].

Our contribution. We consider the variational inequality with a composite operator R := P + Q, both components of which are Lipschitz-continuous and only one of the component Q is monotone (the operator P can be generally nonmonotone). We additionally assume that the whole target operator R satisfies Minty assumption [25] – by far the most common and well-known relaxation of non-monotonicity. For this kind of problem, the classical Extragradient method requires [17]

$$\mathcal{O}\left(\frac{(L_p + L_q)^2 \|x^0 - x^*\|^2}{\varepsilon^2}\right)$$
computations of the operator P ,

where L_p and L_q are Lipschitz constants of P and Q, x^0 is a starting point, x^* is a solution of the VI problem, ε is required accuracy of the obtained numerical solution. On the other hand, motivated by the fact that the non-monotone operator P is likely to be more computationally expensive than Q, we consider a modification of Extragradient using the sliding technique [24], which allows to take into account the composite structure of the problem. This makes it possible to compute one of the operators less frequently. In particular, we obtain improvements on the estimate of operator P calls. For our method it is

$$\mathcal{O}\left(\frac{L_p^2 \|x^0 - x^*\|^2}{\varepsilon^2}\right)$$

which can be much better for some relations on L_p and L_q compared to the Extragradient method. To support the theoretical results, a series of experiments are set up and the results confirm the improvements.

Notation. We use $\langle x, y \rangle := \sum_{i=1}^{d} x_i y_i$ to introduce inner product of $x, y \in \mathbb{R}^d$, where x_i corresponds to the *i*-th component of x in the standard basis in \mathbb{R}^d . It induces ℓ_2 -norm in \mathbb{R}^d according to $||x|| := \sqrt{\langle x, x \rangle}$.

2 Problem setting

We consider the composite variational inequality in the following form:

Find
$$x^* \in \mathbb{R}^d : R(x^*) = 0$$
 with $R(x) := Q(x) + P(x),$ (1)

where $Q(x), P(x) : \mathbb{R}^d \to \mathbb{R}^d$ are operators.

Many known problems can be reformulated using the language of composite variational inequalities. Let us consider two common use cases for VIs:

1. Minimization problem. One can notice, that solving $\min_{x \in \mathbb{R}^d} r(x)$, where r(x) := q(x) + p(x) is a convex function, is equivalent to solving the VI problem (1) with $Q(x) := \nabla q(x)$, $P(x) := \nabla p(x)$, $R(x) := \nabla r(x) = \nabla q(x) + \nabla p(x)$.

2. Saddle point problem. Let us consider the saddle-point problem:

$$\min_{y \in \mathbb{R}^{d_y}} \max_{z \in \mathbb{R}^{d_z}} [r(y, z) := q(y, z) + p(y, z)].$$
(2)

If we take x = [y, z], $Q(x) := Q(y, z) = [\nabla_y q(y, z), -\nabla_z q(y, z)]$, $P(x) := P(y, z) = [\nabla_y p(y, z), -\nabla_z p(y, z)]$, then it can be proved for the convexconcave function r(y, z) that $x^* = (y^*, z^*)$ is a solution for (1) if and only if the following inequality holds

$$r(y^*, z) \le r(y^*, z^*) \le r(y, z^*) \ \forall y \in \mathbb{R}^{d_y}, z \in \mathbb{R}^{d_z}.$$

Equivalently, it means that $x^* = [y^*, z^*]$ is a solution for (2).

Despite the fact that a minimization problem is a special case of variational inequalities, they are usually studied separately. This is due to the fact that a more optimistic convergence theory can be constructed for minimization problems [11,26] compared to the general results for variational inequalities. Meanwhile, the study of saddle point problems is often conducted through the prism of variational inequalities [16,27], particularly in recent years, theoretical studies of variational inequalities have been associated with solving the practical minimax learning problem of GANs training [28,20,29].

We study problem (1) under the following commonly used assumptions:

Assumption 1 R(x) satisfies Minty assumption:

$$\exists x^* \in \mathbb{R}^d : \forall x \in \mathbb{R}^d \hookrightarrow \langle R(x), x - x^* \rangle > 0.$$

This assumption, also called the variational stability condition is considered as an option to structurally constrain a non-monotone problem. Minty assumption is widely used in the literature [17,30,31,32,19,33].

Next, we also introduce two standard assumptions for the analysis of variational inequalities.

Assumption 2 Q(x) is L_q -Lipschitz and monotone:

$$\begin{aligned} \forall x_1, x_2 \in \mathbb{R}^d &\hookrightarrow \|Q(x_1) - Q(x_2)\| \le L_q \|x_1 - x_2\|, \\ \forall x_1, x_2 \in \mathbb{R}^d &\hookrightarrow \langle Q(x_1) - Q(x_2), x_1 - x_2 \rangle \ge 0. \end{aligned}$$

R. Emelyanov, A. Tikhomirov, A. Beznosikov, A. Gasnikov

Assumption 3 P(x) is L_p -Lipschitz:

 $\forall x_1, x_2 \in \mathbb{R}^d \hookrightarrow \|P(x_1) - P(x_2)\| \le L_p \|x_1 - x_2\|.$

Once again we emphasize the key detail that only the operator Q, but not P, is monotone, and hence the full operator R can be non-monotone in general.

3 Algorithm

4

The algorithm studied in this paper is a version of the Extragradient algorithm, but with additional sliding technique [24]:

Algorithm 1 Extragradient Sliding
1: Input: starting point $x^0 \in \mathbb{R}^d$
2: Parameters: stepsizes $\eta, \theta > 0$, number iterations $K \in \mathbb{N}$
3: for $k = 0, 1, 2, \dots, K - 1$ do
4: Find $u^k \approx \tilde{u}^k$ where \tilde{u}^k is solution for
Find $\tilde{u}^k \in \mathbb{R}^d$: $B^k_\theta(\tilde{u}^k) = 0$ with $B^k_\theta(x) := P(x^k) + Q(x) + \frac{1}{\theta}(x - x^k)$
5: $x^{k+1} = x^k - \eta R(u^k)$
6: end for

Let us give a high-level intuition of how the above algorithm works. The main idea of this algorithm is to move away from equal number of calls of P and Q, as it happens in the classical Extragradient method. The more computationally expensive P is called twice per iteration of the algorithm, when selecting the optimal \tilde{u}^k in line 4 and when computing R in line 5. In turn, the computationally simpler Q is called some number of times in the inner problem (line 4) and also when computing R. This is the idea behind the sliding technique: fix one of the operators and vary the other due to its cheapness. Thus in line 4 the full operator R(x) is approximated by Q(x) and a slightly outdated version $P(x^k)$.

4 Convergence analysis

Theorem 1. Consider Algorithm 1 for the problem (1) under Assumptions 1–3, with the following tuning:

$$\theta = \frac{1}{2L_p}, \ \eta = \frac{\theta}{2}.$$

Assume that u^k (line 4) satisfies:

$$\|B_{\theta}^{k}(u^{k})\|^{2} \leq \frac{L_{p}^{2}}{3} \|x^{k} - \tilde{u}^{k}\|^{2}.$$
(3)

Then, we have the following convergence estimate:

$$\min_{0 \le j \le K-1} \|R(u^j)\|^2 \le \frac{16L_p^2}{K} \|x_0 - x^*\|^2}{K}$$

This results means sublinear convergence. To prove the theorem we first deal with the auxiliary lemma.

Lemma 1. Consider Algorithm 1. Let θ be defined as $\theta = \frac{1}{2L_p}$. Then, under Assumptions 1, 2, 3, the following inequality holds:

$$2\langle x^* - x^k, R(u^k) \rangle \le -\theta \|R(u^k)\|^2 + 3\theta \left(\|B^k_\theta(u^k)\|^2 - \frac{L_p^2}{3} \|x^k - \tilde{u}^k\|^2 \right).$$
(4)

Proof of Lemma 1. Using Assumption 1, we get

$$\begin{split} 2\langle x^k - x^*, R(u^k) \rangle &= 2\langle x^* - u^k, R(u^k) \rangle + 2\langle u^k - x^k, R(u^k) \rangle \\ &\leq 2\langle u^k - x^k, R(u^k) \rangle = 2\theta \left\langle \frac{1}{\theta} (u^k - x^k), R(u^k) \right\rangle. \end{split}$$

The definition of $B^k_{\theta}(x)$ (line 4 of Algorithm 1) gives

$$\begin{aligned} 2\theta \left\langle \frac{1}{\theta} (u^k - x^k), R(u^k) \right\rangle &= \theta \left\| \frac{1}{\theta} (u^k - x^k) + R(u^k) \right\|^2 - \frac{1}{\theta} \|u^k - x^k\|^2 - \theta \|R(u^k)\|^2 \\ &= -\frac{1}{\theta} \|u^k - x^k\|^2 - \theta \|R(u^k)\|^2 \\ &+ \theta \|B_{\theta}^k(u^k) - P(x^k) + P(u^k)\|^2. \end{aligned}$$

Using the Cauchy-Schwarz inequality and L_p -Lipschitzness of P(x) (Assumption 3), we get

$$\begin{aligned} 2\theta \left\langle \frac{1}{\theta}(u^{k} - x^{k}), R(u^{k}) \right\rangle &\leq -\frac{1}{\theta} \|u^{k} - x^{k}\| - \theta \|R(u^{k})\|^{2} + 2\theta \|B_{\theta}^{k}(u^{k})\|^{2} \\ &\quad +2\|P(u^{k}) - P(x^{k})\|^{2} \\ &\leq -\frac{1}{\theta} \|u^{k} - x^{k}\| - \theta \|R(u^{k})\|^{2} + 2\theta \|B_{\theta}^{k}(u^{k})\|^{2} \\ &\quad +2\theta L_{p}^{2}\|u^{k} - x^{k}\|^{2} \\ &= -\frac{1}{\theta} \left(1 - 2\theta^{2}L_{p}^{2}\right) \|u^{k} - x^{k}\|^{2} - \theta \|R(u^{k})\|^{2} + 2\theta \|B_{\theta}^{k}(u^{k})\|^{2} \end{aligned}$$

With $\theta = \frac{1}{2L_p}$ and the Cauchy-Schwarz inequality in the form $-\|a\|^2 \le \|b\|^2 - \frac{1}{2}\|a+b\|^2$, one can obtain

$$\begin{aligned} 2\langle x^* - x^k, R(u^k) \rangle &\leq -\theta \|R(u^k)\|^2 + 2\theta \|B_{\theta}^k(u^k)\|^2 - \frac{1}{2\theta} \|u^k - x^k\|^2 \\ &\leq -\theta \|R(u^k)\|^2 + 2\theta \|B_{\theta}^k(u^k)\|^2 \\ &\quad + \frac{1}{2\theta} \|u^k - \tilde{u}^k\|^2 - \frac{1}{4\theta} \|x^k - \tilde{u}^k\|^2. \end{aligned}$$

Additionally, we can observe that $B^k_{\theta}(x)$ is $\frac{1}{\theta}$ -strongly monotone. It follows from the definition of the operator: the operator $B^k_{\theta}(x)$ is a sum of the monotone

operator Q (Assumption 2) and the strong monotone linear operator $\frac{1}{\theta}(x-x^k)$. Together with the Cauchy-Schwarz inequality, it gives that

$$\|u^k - \tilde{u}^k\|^2 \le \theta \langle B^k_\theta(u^k) - B^k_\theta(\tilde{u}^k), u^k - \tilde{u}^k \rangle \le \theta \|B^k_\theta(u^k) - B^k_\theta(\tilde{u}^k)\| \cdot \|u^k - \tilde{u}^k\|.$$

With $B^k_{\theta}(\tilde{u}^k) = 0$ (\tilde{u}^k is the solution of the subproblem from line 4), we get

$$||u^k - \tilde{u}^k||^2 \le \theta^2 ||B^k_\theta(u^k)||^2$$

Applying this to the upper inequality, we finalize the proof:

$$\begin{aligned} 2\langle x^* - x^k, R(u^k) \rangle &\leq -\theta \|R(u^k)\|^2 + \frac{5}{2}\theta \|B^k_{\theta}(u^k)\|^2 - \frac{1}{4\theta} \|x^k - \tilde{u}^k\|^2 \\ &\leq -\theta \|R(u^k)\|^2 + 3\theta \|B^k_{\theta}(u^k)\|^2 - \frac{3\theta}{12\theta^2} \|x^k - \tilde{u}^k\|^2 \\ &= -\theta \|R(u^k)\|^2 + 3\theta \left(\|B^k_{\theta}(u^k)\|^2 - \frac{L^2_p}{3} \|x^k - \tilde{u}^k\|^2 \right). \end{aligned}$$

6

Here we substitute $\theta = \frac{1}{2L_p}$. \Box Now we are ready to prove the main theorem. Proof of Theorem 1. Line 5 of Algorithm 1 gives

$$\begin{split} \|x^{k+1} - x^*\|^2 &= \|x^{k+1} - x^k\|^2 + 2\langle x^{k+1} - x^k, x^k - x^*\rangle + \|x^k - x^*\|^2 \\ &= \|x^{k+1} - x^k\|^2 + \|x^k - x^*\|^2 - 2\eta\langle R(u^k), x^k - x^*\rangle. \end{split}$$

Using the results of Lemma 1 and the condition (3) on $||B_{\theta}^{k}(u^{k})||^{2}$, we get

$$\begin{split} \|x^{k+1} - x^*\|^2 &\leq \|x^{k+1} - x^k\|^2 + \|x^k - x^*\|^2 - \eta\theta \|R(u^k)\|^2 \\ &+ 3\eta\theta \left(\|B^k_\theta(u^k)\|^2 - \frac{L_p^2}{3} \|x^k - \tilde{u}^k\|^2 \right) \\ &\leq \|x^{k+1} - x^k\|^2 + \|x^k - x^*\|^2 - \eta\theta \|R(u^k)\|^2. \end{split}$$

Again from line 5 it follows that

$$||x^{k+1} - x^*||^2 \le \eta^2 ||R(u^k)||^2 + ||x^k - x^*||^2 - \eta\theta ||R(u^k)||^2.$$

Let us substitute the choice of parameters: $\theta = \frac{1}{2L_p}, \ \eta = \frac{\theta}{2}$:

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &\leq \|x^k - x^*\|^2 - \eta(\theta - \eta) \|R(u^k)\|^2 \\ &\leq \|x^k - x^*\|^2 - \frac{\theta^2}{4} \|R(u^k)\|^2. \end{aligned}$$

Summing from 1 to K - 1, one can obtain

$$\sum_{j=0}^{K-1} \frac{\theta^2}{4} \|R(u^j)\|^2 \le \sum_{j=0}^{K-1} \left(\|x^j - x^*\|^2 - \|x^{j+1} - x^*\|^2 \right)$$
$$= \|x^0 - x^*\|^2 - \|x^K - x^*\|^2$$
$$\le \|x^0 - x^*\|^2.$$

Thus, we have

$$\sum_{j=0}^{K-1} \|R(u^j)\|^2 \le 16L_p^2 \|x^0 - x^*\|^2,$$

and finally,

$$\min_{0 \leq j \leq K-1} \|R(u^j)\|^2 \leq \frac{16L_p^2 \|x^0 - x^*\|^2}{K}.$$

This ends the proof. \Box

Corollary 1. Under the assumptions of Theorem 1, to achieve a ε -solution in terms $\varepsilon \sim ||R(u)||$, Algorithm 1 needs

$$\mathcal{O}\left(\frac{L_p^2 \|x^0 - x^*\|^2}{\varepsilon^2}\right)$$
 computations of the operator P.

As noted in Section 1, this result is better than for the Extragradient method, which provides an estimate: $\mathcal{O}\left(\frac{(L_p+L_q)^2 ||x^0-x^*||^2}{\varepsilon^2}\right)$.

Remark 1. Meanwhile, line 4 of the algorithm requires an additional algorithm to efficiently solve the resulting subproblem. The Extra Anchored Gradient algorithm proposed in [34] can be used for these purposes. In fact, any method of solving variational inequalities with a Lipschitz monotone operator can be used here, but the method from [34] has convergence guarantees necessary for our theoretical analysis (see (3)). Moreover, it is also shown in [34] that these guarantees are optimal and unimprovable.

5 Numerical Experiments

In this part, we conduct three experiments: with a generated bilinear problem, with a logarithmic logistic regression, and with non-convex least squares (NLLSQ) on the **mushrooms** dataset from LibSVM [35,36]. It turns out we evaluate the performance of the algorithm on both synthetic and real data.

5.1 Bilinear problem

A bilinear problem is a classical and keystone example of the saddle:

$$\min_{x \in [-1;1]^d} \max_{y \in [-1;1]^d} \left[f(x,y) := (x - b_x)^T A(y - b_y) + \frac{1}{2} \|x - b_x\|^2 - \frac{1}{2} \|y - b_y\|^2 \right].$$
(5)

(5) In this setting, P refers to the main part $(x - b_x)^T A(y - b_y)$ of the gradient while Q represents the gradient from the regularization terms $\frac{1}{2} ||x - b_x||^2 - \frac{1}{2} ||y - b_y||^2$ (see the second example from Section 2).

For the purposes of the experiment, a random bilinear saddle point problem is generated. The dimension d of the problem is set equal to 1000. The matrix A

7

are sampled as a positive definite matrix with uniformly distributed eigenvalues from μ to L, where μ and L are chosen as 0.1 and 100 correspondingly. Biases b_x and b_y are both sampled from $\mathcal{U}(-1,1)$ as well as the starting point. In Figure 1, one can see the plots comparing the convergence of the algorithm presented in this paper and Extragradient in terms of the number of iterations and oracle calls.

Fig. 1: Comparison of Extragradient and Extragradient Sliding for the generated bilinear saddle point problem (5).



5.2 Logarithmic loss problem

8

The second experiment uses the *mushrooms* dataset and a more complex to compute logarithmic regression problem. To create a saddle point problem, the adversarial noise [37] technique is used. In this setting, the model weights are adjusted in parallel with the trained noise, which makes the model more robust. The resulting problem is formulated as follows:

$$\min_{x \in \mathbb{R}^d} \max_{\|y_i\| \le \delta} \left[f(x, y_1, \dots, y_i, \dots, y_N) := \frac{1}{N} \sum_{i=1}^N \ln\left(1 + \exp(-b_i x^T (A_i + y_i)) \right) + \frac{\beta_x}{2} \|x\|^2 - \frac{\beta_y}{2} \|y\|^2 \right].$$
(6)

Here A, b are data, x represents the model's weights, y_i stands for adversarial noise, β_x and β_y determine the degree of regularization, δ defines the constraint imposed on adversarial noise. As in the previous case, the starting point is sampled from a uniform distribution $\mathcal{U}(-1, 1)$. β_x and β_y are both set to 0.1, δ is set to 0.1.

Plots showing comparison of the convergence for Algorithm 1 and Extragradient are presented in Figure 2. As can be seen, the algorithm presented in the paper outperforms baseline simultaneously in terms of P and Q oracle calls, despite the fact that solving the inner subproblem involves more Q oracle calls than in the standard Extragradient.



Fig. 2: Comparison of Extragradient and Extragradient Sliding for the log loss saddle point problem (6)

5.3 NLLSQ loss problem

The last experiment is performed on a non-convex loss function in order to demonstrate the success of the method for the non-monotone operator P. As in the case of the logistic loss function, a saddle problem with adversarial noise adding robustness is posed:

$$\min_{x \in \mathbb{R}^d} \max_{\|y_i\| \le \delta} \left[f(x, y_1, \dots, y_i, \dots, y_N) := \frac{1}{N} \sum_{i=1}^N \left(b_i - \frac{1}{1 + \exp(-x^T (A_i + y_i))} \right)^2 + \frac{\beta_x}{2} \|x\|^2 - \frac{\beta_y}{2} \|y\|^2 \right].$$
(7)

The mushrooms dataset is used again for the experiment. The notation used in the formula is similar to the previous case. The starting point is sampled from a uniform distribution $\mathcal{U}(-1, 1)$. β_x and β_y are both set to 0.1, δ is set to 0.1.

The comparison of algorithms in this setting, presented in Figure 3, emphasizes the practical value of the considered algorithm. In this experiment it again wins the baseline on the calls of both oracles.

Fig. 3: Comparison of Extragradient and Extragradient Sliding for the NLLSQ loss saddle point problem (7)



References

- P. T. Harker and J.-S. Pang. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. *Mathematical programming*, 1990.
- Francisco Facchinei and Jong-Shi Pang. Finite-Dimensional Variational Inequalities and Complementarity Problems. Springer Series in Operations Research. Springer, 2003.
- Alejandro Jofré, R Terry Rockafellar, and Roger JB Wets. Variational inequalities and economic equilibrium. *Mathematics of Operations Research*, 32(1):32–50, 2007.
- Francisco Facchinei and Jong-Shi Pang. Finite-dimensional variational inequalities and complementarity problems. Springer, 2003.
- Soroosh Shafieezadeh Abadeh, Peyman M Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. Advances in Neural Information Processing Systems, 28, 2015.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Advances in neural information processing systems, 27, 2014.
- Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pages 2817–2826. PMLR, 2017.
- F. E. Browder. Existence and approximation of solutions of nonlinear variational inequalities. *Proceedings of the National Academy of Sciences*, 56(4):1080–1086, 1966.
- R.T. Rockafellar. Convex functions, monotone operators and variational inequalities. theory and applications of monotone operators. *Theory and applications of monotone* operators, pages 13–65, 1969.
- M. Sibony. Méthodes itératives pour les équations et inéquations aux dérivées partielles non linéaires de type monotone. *Calcolo*, 7(1):65–183, 1970.
- 11. B. Polyak. Introduction to optimization. Optimization Software, 1987.
- Aleksandr Beznosikov, Boris Polyak, Eduard Gorbunov, Dmitry Kovalev, and Alexander Gasnikov. Smooth monotone stochastic variational inequalities and saddle point problems: A survey. *European Mathematical Society Magazine*, (127):15–28, 2023.
- G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.
- Leonid Denisovich Popov. A modification of the arrow-hurwicz method for search of saddle points. Mathematical notes of the Academy of Sciences of the USSR, 28:845–848, 1980.
- Aryan Mokhtari, Asuman E Ozdaglar, and Sarath Pattathil. Convergence rate of o(1/k) for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. SIAM Journal on Optimization, 30(4):3230–3251, 2020.
- Arkadi Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convexconcave saddle point problems. SIAM Journal on Optimization, 15(1):229–251, 2004.
- Cong D Dang and Guanghui 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.

- Aswin Kannan and Uday V Shanbhag. Optimal stochastic extragradient schemes for pseudomonotone stochastic variational inequality problems and their variants. *Computational Optimization and Applications*, 74(3):779–820, 2019.
- Jelena Diakonikolas, Constantinos Daskalakis, and Michael I Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2746–2754. PMLR, 2021.
- Aleksandr Beznosikov, Pavel Dvurechenskii, Anastasiia Koloskova, Valentin Samokhin, Sebastian U Stich, and Alexander Gasnikov. Decentralized local stochastic extra-gradient for variational inequalities. Advances in Neural Information Processing Systems, 35:38116–38133, 2022.
- Aleksandr Beznosikov, Peter Richtárik, Michael Diskin, Max Ryabinin, and Alexander Gasnikov. Distributed methods with compressed communication for solving variational inequalities, with theoretical guarantees. Advances in Neural Information Processing Systems, 35:14013–14029, 2022.
- 22. Guanghui Lan and Yuyuan Ouyang. Mirror-prox sliding methods for solving a class of monotone variational inequalities. *arXiv preprint arXiv:2111.00996*, 2021.
- Aleksandr Beznosikov, Gesualdo Scutari, Alexander Rogozin, and Alexander Gasnikov. Distributed saddle-point problems under data similarity. Advances in Neural Information Processing Systems, 34:8172–8184, 2021.
- Dmitry Kovalev, Aleksandr Beznosikov, Ekaterina Borodich, Alexander Gasnikov, and Gesualdo Scutari. Optimal gradient sliding and its application to optimal distributed optimization under similarity. Advances in Neural Information Processing Systems, 35:33494–33507, 2022.
- George J. Minty. Monotone (nonlinear) operators in Hilbert space. Duke Mathematical Journal, 29(3):341 – 346, 1962.
- 26. Yurii Nesterov et al. Lectures on convex optimization, volume 137. Springer.
- Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.
- Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. arXiv preprint arXiv:1802.10551, 2018.
- 29. Aleksandr Beznosikov, Aibek Alanov, Dmitry Kovalev, Martin Takáč, and Alexander Gasnikov. On scaled methods for saddle point problems. *arXiv preprint arXiv:2206.08303*, 2022.
- Alfredo N Iusem, Alejandro Jofré, Roberto Imbuzeiro Oliveira, and Philip Thompson. Extragradient method with variance reduction for stochastic variational inequalities. SIAM Journal on Optimization, 27(2):686–724, 2017.
- Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. arXiv preprint arXiv:1807.02629, 2018.
- 32. Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. Advances in Neural Information Processing Systems, 33:16223-16234, 2020.
- Aleksandr Beznosikov, Valentin Samokhin, and Alexander Gasnikov. Distributed saddle-point problems: Lower bounds, near-optimal and robust algorithms. arXiv preprint arXiv:2010.13112, 2022.
- 34. TaeHo Yoon and Ernest K. Ryu. Accelerated algorithms for smooth convex-concave minimax problems with $\mathcal{O}(1/k^2)$ rate on squared gradient norm, 2021.

- 12 R. Emelyanov, A. Tikhomirov, A. Beznosikov, A. Gasnikov
- Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1-27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
- Mushroom. UCI Machine Learning Repository, 1987. DOI: https://doi.org/10.24432/C5959T.
- Dawei Zhou, Nannan Wang, Bo Han, and Tongliang Liu. Modeling adversarial noise for adversarial training. In *International Conference on Machine Learning*, pages 27353–27366. PMLR, 2022.