Smooth monotone stochastic variational inequalities and saddle point problems: A survey

Aleksandr Beznosikov, Boris Polyak[†], Eduard Gorbunov, Dmitry Kovalev and Alexander Gasnikov

This paper is a survey of methods for solving smooth, (strongly) monotone stochastic variational inequalities. To begin with, we present the deterministic foundation from which the stochastic methods eventually evolved. Then we review methods for the general stochastic formulation, and look at the finite-sum setup. The last parts of the paper are devoted to various recent (not necessarily stochastic) advances in algorithms for variational inequalities.

1 Introduction

In its long, more than half-century history of study (going back to the classical article [113]), variational inequalities have become one of the most popular and universal optimization formulations. Variational inequalities are used in various areas of applied mathematics. Here we can highlight both classical examples from game theory, economics, operator theory, convex analysis [6, 19, 106, 110, 113], as well as newer and even more recent applications in optimization and machine learning: non-smooth optimization [93], unsupervised learning [9, 22, 36], robust/adversarial optimization [11], GANs [47] and reinforcement learning [57, 100]. Modern times present new challenges to the community. The increase in scale of problems and the need to speed up solution processes have sparked a huge interest in *stochastic* formulations of applied tasks, including variational inequalities. This paper surveys stochastic methods for solving variational inequalities.

Structure of the paper. In Section 2, we give a formal statement of the variational inequality problem, basic examples, and main assumptions. Section 3 deals with deterministic methods, from which stochastic methods have evolved. Section 4 covers a variety of stochastic methods. Section 5 is devoted to the recent advances in (not necessarily stochastic) variational inequalities and saddle point problems.

2 Problem: Setting and assumptions

Notation. We use $\langle x, y \rangle \coloneqq \sum_{i=1}^d x_i y_i$ to denote the standard inner product of vectors $x, y \in \mathbb{R}^d$, where x_i is the i-th component of x in the standard basis of $\mathbb{R}^d$. It induces the ℓ_2-norm in $\mathbb{R}^d$ by $\|x\|_2 \coloneqq \sqrt{\langle x, x \rangle}$. We denote the ℓ_p-norm by $\|x\|_p \coloneqq \big(\sum_{i=1}^d |x_i|^p\big)^{1/p}$ for $p \in [1, \infty)$, and $\|x\|_\infty \coloneqq \max_{1 \le i \le d} |x_i|$ for $p = \infty$. The dual norm $\|\cdot\|_*$ corresponding to the norm $\|\cdot\|$ is defined by $\|y\|_* \coloneqq \max\{\langle x, y \rangle \mid \|x\| \le 1\}$. The symbol $\mathbb{E}[\cdot]$ stands for the total mathematical expectation. Finally, we need to introduce the symbols $\mathcal{O}$ and Ω to enclose numerical constants that do not depend on any parameters of the problem, and the symbols $\tilde{\mathcal{O}}$ and $\tilde{\Omega}$ to enclose numerical constants and logarithmic factors.

We study variational inequalities (VI) of the form

$$\text{find } z^* \in \mathcal{Z} \text{ such that } \langle F(z^*), z - z^* \rangle \ge 0 \quad \forall z \in \mathcal{Z}, \tag{1}$$

where $F\colon \mathcal{Z} \to \mathbb{R}^d$ is an operator and $\mathcal{Z} \subseteq \mathbb{R}^d$ is a convex set.

To emphasize the extensiveness of formulation (1), we give a few examples of variational inequalities arising in applied sciences.

Example 1 (Minimization)**.** Consider the minimization problem

$$\min_{z \in \mathcal{Z}} f(z). \tag{2}$$

Let $F(z) \coloneqq \nabla f(z)$. Then, if f is convex, one can prove that $z^* \in \mathcal{Z}$ is a solution of (1) if and only if $z^* \in \mathcal{Z}$ is a solution of problem (2).

Example 2 (Saddle point problem)**.** Consider the saddle point problem (SPP)

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} g(x, y). \tag{3}$$

Suppose that $F(z) \coloneqq F(x, y) = [\nabla_x g(x, y), -\nabla_y g(x, y)]$ and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} \subseteq \mathbb{R}^{d_x}$, $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$. Then, if g is convex-concave, one can prove that $z^* \in \mathcal{Z}$ is a solution of problem (1) if and only if $z^* \in \mathcal{Z}$ is a solution of problem (3).

The study of saddle point problems is often associated with variational inequalities.

Example 3 (Fixed point problem). Consider the fixed point problem

$$\text{find } z^* \in \mathbb{R}^d \text{ such that } T(z^*) = z^*, \tag{4}$$

where $T\colon \mathbb{R}^d \to \mathbb{R}^d$ is an operator. If we set $F(z) = z - T(z)$, then one can prove that $z^* \in \mathcal{Z} = \mathbb{R}^d$ is a solution of problem (1) if and only if $F(z^*) = 0$, i.e., $z^* \in \mathbb{R}^d$ is a solution of problem (4).

For the operator F from (1) we assume the following.

Assumption 1 (Lipschitzness). The operator F is L-Lipschitz continuous, i.e., for all $u, v \in \mathcal{Z}$, we have $\|F(u) - F(v)\|_* \leq L\|u - v\|$.

In the context of problems (2) and (3), L-Lipschitzness of the operator means that the functions $f(z)$ and $g(x,y)$ are L-smooth.

Assumption 2 (Strong monotonicity). The operator F is μ-strongly monotone, i.e., for all $u, v \in \mathcal{Z}$, we have $\langle F(u) - F(v), u - v\rangle \geq \mu\|u - v\|_2^2$. If $\mu = 0$, then the operator F is monotone.

In the context of problems (2) and (3), strong monotonicity of F means strong convexity of $f(z)$ and strong convexity-strong concavity of $g(x,y)$. In this paper we first focus on the strongly monotone and monotone cases. But there are also various assumptions relaxing monotonicity and strong monotonicity (e.g., see [55] and references therein).

We note that Assumptions 1 and 2 are sufficient for the existence of a solution to problem (1) (see, e.g., [37]).

Since we work on the set $\mathcal{Z}$, it is useful to introduce the Euclidean projection onto $\mathcal{Z}$,

$$P_{\mathcal{Z}}(z) = \arg\min_{v \in \mathcal{Z}} \|z - v\|_2.$$

To characterize the convergence of the methods for monotone variational inequalities we introduce the gap function,

$$\mathrm{Gap}_{\mathrm{VI}}(z) \coloneqq \sup_{u \in \mathcal{Z}} [\langle F(u), z - u\rangle]. \tag{5}$$

Such a gap function, regarded as a convergence criterion, is more suitable for the following variational inequality problem:

$$\text{find } z^* \in \mathcal{Z} \text{ such that } \langle F(z), z^* - z\rangle \leq 0 \quad \text{for } z \in \mathcal{Z}.$$

Such a solution is also called weak or Minty (whereas the solution of (1) is called strong or Stampacchia). However, in view of Assumption 1, F is single-valued and continuous on $\mathcal{Z}$, meaning that actually the two indicated formulations of the variational inequality problem are equivalent [37].

For the minimization problem (2), the functional distance to the solution, i.e., the difference $f(z) - f(z^*)$, can be used instead of (5). For saddle point problems (3), a slightly different gap function is used, namely,

$$\mathrm{Gap}_{\mathrm{SPP}}(z) \coloneqq \mathrm{gap}(x,y) = \max_{y' \in \mathcal{Y}} f(x,y') - \min_{x' \in \mathcal{X}} f(x',y). \tag{6}$$

For both functions (5) and (6) it is crucial that the feasible set is bounded (in fact it is not necessary to take the whole set $\mathcal{Z}$, which can be unbounded – it suffices to take a bounded convex subset $\mathcal{C}$ which contains some solution, see [95]). Therefore it is necessary to define a distance on the set $\mathcal{Z}$. Since this survey covers methods not only in the Euclidean setup, let us introduce a more general notion of distance.

Definition 1 (Bregman divergence). Let $v(z)$ be a function that is 1-strongly convex w.r.t. the norm $\|\cdot\|$ and differentiable on $\mathcal{Z}$. Then for any two points $z, z' \in \mathcal{Z}$ the Bregman divergence (or Bregman distance) $V(z,z')$ associated with $v(z)$ is defined as

$$V(z,z') \coloneqq v(z') - v(z) - \langle \nabla v(z), z' - z\rangle.$$

We denote the Bregman diameter of the set $\mathcal{Z}$ w.r.t. the divergence $V(z,z')$ as $D_{\mathcal{Z},V} \coloneqq \max\{\sqrt{2V(z,z')} \mid z, z' \in \mathcal{Z}\}$. In the Euclidean case, we simply write $D_{\mathcal{Z}}$ instead of $D_{\mathcal{Z},V}$. Using the definition of V, we introduce the so-called proximal operator as follows:

$$\mathrm{prox}_x(y) = \arg\min_{z \in \mathcal{Z}} \{\langle y, z\rangle + V(z,x)\}.$$

3 Deterministic foundation: Extragradient and other methods

The first and the simplest method for solving the variational inequality (1) is the iterative scheme (also known as the Gradient method)

$$z^{k+1} = P_{\mathcal{Z}}(z^k - \gamma F(z^k)), \tag{7}$$

where $\gamma > 0$ is a step size. Note that using the proximal operator associated with the Euclidean Bregman divergence this method can be rewritten in the form

$$z^{k+1} = \mathrm{prox}_{z^k}(\gamma F(z^k)).$$

The basic result asserts the convergence of the method to the unique solution of (1) for strongly monotones and L-Lipschitz operators F; it was obtained in the papers [19, 106, 110].

Theorem 1. *If Assumptions 1 and 2 hold and $0 < \gamma < 2\mu/L^2$, then after k iterations method (7) converges to z^* with a linear rate:*

$$\|z^k - z^*\|_2^2 = \mathcal{O}(R_0^2 q^k), \quad \textit{with } q = (1 - 2\gamma\mu + \gamma^2 L^2)$$

and R_0 denotes (here and everywhere in the sequel) the norm $\|z^0 - z^*\|_2$. For $\gamma = \mu/L^2$, we have $q = (1 - 1/\kappa^2)$, $\kappa = L/\mu$, thus *the upper bound on the number of iterations needed to achieve the ε-solution (i.e., $\|z^k - z^*\|_2^2 \le \varepsilon$) is $\mathcal{O}(\kappa^2 \log(R_0^2/\varepsilon))$.*

Various extensions of this statement (for the case when F is not Lipschitz, but with linear growth bounds, or when the values of F are corrupted by noise) can be found in [10, Theorem 1].

When F is a potential operator (see Example 1) method (7) coincides with the gradient projection algorithm. It converges for strongly monotone F. Moreover, the bounds for the admissible step size are less restrictive ($0 < \gamma < 2/L$) and the relevant complexity estimates are better ($O(\kappa \log(R_0^2/\varepsilon))$) than in Theorem 1; see [104, Theorem 2 in Section 1.4.2].

However, in the general monotone, but not strongly monotone case (for instance, for the convex-concave SPP, Example 2) convergence fails. The original statements on the convergence of Uzawa's method (a version of (7)) for saddle point problems [6] were wrong; there are numerous well-known examples where method (7) for F corresponding to a bilinear SPP diverges, see, e.g., [104, Figure 39].

There have been many other attempts to recover the convergence of gradient-like methods, not for VIs, but for saddle point problems. One of them is based on the transition to modified Lagrangians when $g(x, y)$ is a Lagrange function, see [45, 104]. However, we focus on the general VI case. A possible approach is based on the idea of *regularization*. Instead of the monotone variational inequality (1) one can deal with a regularized inequality, in which the monotone operator F is replaced by strongly monotone one $F + \varepsilon_k T$, where $T(z)$ is a strongly monotone operator and $\varepsilon_k > 0$ is a regularization parameter. If we denote by z^k the solution of the regularized VI, then one can prove that z^k converges to z^* as $\varepsilon_k \to 0$ (see [10]). However, usually the solution z^k is not easily available. To address this problem, an *iterative regularization* technique is proposed in [10], where one step of the basic method (7) is applied for the regularized problem. Step sizes and regularization parameters can be adjusted to guarantee convergence.

Another technique is based on the Proximal Point Method proposed independently by B. Martinet in [84] and by T. Rockafellar in [107]. At each iteration this methods requires the solution of the VI with the operator $F + cI$, where $c > 0$ and I is the identity operator. This is an implicit method (similar to the regularization method), however there exist numerous implementable versions of Proximal Point. For instance, some methods discussed below can be considered from this point of view.

The breakthrough in methods for solving (non-strongly) monotone variational inequalities was made by Galina Korpelevich [64]. She exploited the idea of extrapolation for the gradient method. How this works can be explained for the simplest example of a two-dimensional min-max problem with $g(x, y) = xy$ and $\mathcal{Z} = \mathbb{R}^2$. It has the unique saddle point $z = 0$, and in any point z^k the direction $F(z^k)$ is orthogonal to z^k; thus, the iteration (7) increases the distance to the saddle point. However, if we perform the step (7) and get the extrapolated point $z^{k+1/2}$, the direction $-F(z^{k+1/2})$ is attracted to the saddle point. Thus, the Extragradient method for solving (1) reads

$$\begin{aligned} z^{k+1/2} &= P_{\mathcal{Z}}(z^k - \gamma F(z^k)), \\ z^{k+1} &= P_{\mathcal{Z}}(z^k - \gamma F(z^{k+1/2})). \end{aligned}$$

Theorem 2. *Let F satisfy Assumptions* 1 *and* 2 *(with $\mu = 0$) and let $0 < \gamma < 1/L$. Then the sequence of iterates z^k generated by the Extragradient method converges to $z^\star$.*

For the particular cases of the zero-sum matrix game or the general bilinear problem with $g(x, y) = y^\top \mathbf{A} x - b^\top x + c^\top y$, the method converges linearly, provided that the optimal solution is unique (see [64, Theorem 3]). In this case, the rate of convergence is equal to $\mathcal{O}(\kappa \log(R_0^2/\varepsilon))$ with $\kappa = \lambda_{\max}(\mathbf{AA}^\top)/\lambda_{\min}(\mathbf{AA}^\top)$. More general upper bounds for the Extragradient method can be found in [119] and in the recent paper [87]. In particular, for the strongly monotone case the estimate $O(\kappa \log(R_0^2/\varepsilon))$ with $\kappa = L/\mu$ holds true (compare with the much worse bound $O(\kappa^2 \log(R_0^2/\varepsilon))$ for the Gradient method). An adaptive version of the Extragradient method (no knowledge of L is required) is proposed in [61].

Another version of the Extragradient method for finding saddle points is provided in [65]. Considering the setup of Example 2, we can exploit just one extrapolating step for the variables y:

$$\begin{aligned} y^{k+1/2} &= P_{\mathcal{Y}}(y^k + \gamma \nabla_y g(x^k, y^k)), \\ x^{k+1} &= P_{\mathcal{X}}(x^k - \gamma \nabla_x g(x^k, y^{k+1/2}), \\ y^{k+1} &= y^k + q(y^{k+1/2} - y^k), \end{aligned} \tag{8}$$

with $0 < \gamma < 1/(2L)$ and $0 < q < 1$. This method converges to the solution and if $g(x, y)$ is linear in y, then the rate of convergence is linear. If we set $q = 1$ in method (8), then $y^{k+1} = y^{k+1/2}$ and we get the so-called Alternating Gradient Method (alternating descent-ascent). In [123], it was proved that this method has *local* linear convergence with complexity $O(\kappa \log(R_0^2/\varepsilon))$, where $\kappa = L/\mu$.

L. Popov [105] proposed a version of extrapolation scheme (sometimes this type of scheme is referred to as *optimistic* or *single-call*):

$$\begin{aligned} z^{k+1/2} &= P_{\mathcal{Z}}(z^k - \gamma F(z^{k-1/2})), \\ z^{k+1} &= P_{\mathcal{Z}}(z^k - \gamma F(z^{k+1/2})). \end{aligned} \tag{9}$$

It requires the single calculation of F at each iteration vs two calculations in the Extragradient method. As shown in [105], method (9) converges for $0 < \gamma < 1/(3L)$. Rates of convergence for this method were derived recently in [41, 87], i.e., $O(\kappa \log(R_0^2/\varepsilon))$ with $\kappa = L/\mu$ for the strongly monotone case and $\kappa = \lambda_{\max}(\mathbf{AA}^\top)/\lambda_{\min}(\mathbf{AA}^\top)$ for the bilinear case. Note that in the general strongly monotone case this estimate is optimal [124], but for the bilinear problem the upper bounds available in the literature for both the Extragradient and optimistic methods are not tight [56]. Meanwhile, optimal estimates $O(\sqrt{\kappa} \log(R_0^2/\varepsilon))$ with $\kappa = \lambda_{\max}(\mathbf{AA}^\top)/\lambda_{\min}(\mathbf{AA}^\top)$ can be achieved using approaches from [4, 7].

An extension of the above schemes to an arbitrary proximal setup was obtained in the work of A. Nemirovsky [92]. He proposed the Mirror-Prox method for VIs, exploiting the Bregman divergence:

$$\begin{aligned} z^{k+1/2} &= \mathrm{prox}_{z^k}(\gamma F(z^k)), \\ z^{k+1} &= \mathrm{prox}_{z^k}(\gamma F(z^{k+1/2})). \end{aligned} \tag{10}$$

This yields the following rate-of-convergence result.

Theorem 3. *Let F satisfy Assumptions* 1 *and* 2 *(with* $\mu = 0$*), and let*

$$\hat{z}^k = \frac{1}{k}\sum_{i=1}^{k} z^{i+1/2}, \tag{11}$$

where $z^{i+1/2}$ *are generated by algorithm* (10) *with* $\gamma = 1/(\sqrt{2}L)$. *Then, after k iterations,*

$$\mathrm{Gap}_{\mathrm{VI}}(\hat{z}^k) = \mathcal{O}\Big(\frac{LD^2_{\mathcal{Z},V}}{k}\Big). \tag{12}$$

Numerous extensions of these original versions of iterative methods for solving variational inequalities were published later. One can highlight Tseng's Forward-Backward Splitting [120], Nesterov's Dual Extrapolation [95], Malitsky and Tam's Forward-Reflected-Backward [83]. All methods have convergence guarantees (12). It turns out that this rate is optimal [101].

4 Stochastic methods: Different setups and assumptions

In this section, we move from deterministic to stochastic methods, i.e., we consider problem (1) with an operator of the form

$$F(z) = \mathbb{E}_{\xi\sim\mathcal{D}}[F_\xi(z)], \tag{13}$$

where ξ is a random variable, $\mathcal{D}$ is some (typically unknown) probability distribution and $F_\xi\colon \mathcal{Z} \to \mathbb{R}^d$ is a stochastic operator. In this setup, calculating the value of the full operator F is computationally expensive or even intractable. Therefore, one has to work mainly with stochastic realizations F_ξ.

4.1 General case

The stochastic formulation (13) for problem (1) was first considered by the authors of [60]. They proposed a natural stochastic generalization of the Extragradient method (more precisely, of the Mirror-Prox methods):

$$\begin{aligned} z^{k+1/2} &= \mathrm{prox}_{z^k}(\gamma F_{\xi^k}(z^k)), \\ z^{k+1} &= \mathrm{prox}_{z^k}(\gamma F_{\xi^{k+1/2}}(z^{k+1/2})). \end{aligned} \tag{14}$$

Here it is important to note that the variables ξ^k and $\xi^{k+1/2}$ are independent and $F_\xi(z)$ is an unbiased estimator of $F(z)$. Moreover, $F_\xi(z)$ is assumed to satisfy the following condition.

Assumption 3 (Bounded variance). The unbiased operator F_ξ has uniformly bounded variance, i.e., for all $\xi \sim \mathcal{D}$ and $u \in \mathcal{Z}$, we have $\mathbb{E}\|F_\xi(u) - F(u)\|_*^2 \le \sigma^2$.

Under this assumption, the following result was established in [60].

Theorem 4. *Let* F_ξ *satisfy Assumptions* 1, 2 *(with* $\mu = 0$*) and* 3, *and let* $\hat{z}^k$ *be defined as in* (11), *where* $z^{i+1/2}$ *are generated by algorithm* (14) *with* $\gamma = \min\{\frac{1}{\sqrt{3}L}, D_{\mathcal{Z},V}\sqrt{\frac{1}{7k\sigma^2}}\}$. *Then, after k iterations, one can guarantee that*

$$\mathbb{E}[\mathrm{Gap}_{\mathrm{VI}}(\hat{z}^k)] = \mathcal{O}\Big(\frac{LD^2_{\mathcal{Z},V}}{k} + D_{\mathcal{Z},V}\sqrt{\frac{\sigma^2}{k}}\Big). \tag{15}$$

In [17], the authors carried out an analysis of algorithm (14) for strongly monotone VIs in the Euclidean case. In particular, under Assumptions 1, 2 and 3 one can guarantee that after k iterations of method (14) one has that (here and below we omit numerical constants in the exponential multiplier)

$$\mathbb{E}\|z^k - z^*\|_2^2 = \tilde{\mathcal{O}}\Big(R_0^2\exp\Big(-\frac{\mu k}{L}\Big) + \frac{\sigma^2}{\mu^2 k}\Big). \tag{16}$$

Also in [17], the authors obtained lower complexity bounds for solving VIs satisfying Assumptions 1, 2 and 3 with stochastic methods. It turns out that the conclusions of Theorem 4 in the monotone case and estimate (16) are optimal and meet lower bounds up to numerical constants.

Optimistic-like (or single-call) methods were also investigated in the stochastic setting. The work [41] applies the following update scheme:

$$\begin{aligned} z^{k+1/2} &= P_{\mathcal{Z}}(z^k - \gamma F_{\xi^{k-1/2}}(z^{k-1/2})), \\ z^{k+1} &= P_{\mathcal{Z}}(z^k - \gamma F_{\xi^{k+1/2}}(z^{k+1/2})). \end{aligned} \tag{17}$$

For this method, in the monotone Euclidean case, the authors proved an estimate similar to (15). In the strongly monotone case, method (17) was investigated in the paper [54], but the estimates obtained there do not meet the lower bounds. The optimal estimates for this scheme were obtained later in [14].

The work [66] deals with a slightly different single-call approach in the non-Euclidean case:

$$z^{k+1} = \mathrm{prox}_{z^k}(\gamma_k F_{\xi^k}(z^k) + \gamma_k\alpha_k[F_{\xi^k}(z^k) - F_{\xi^{k-1}}(z^{k-1})]). \tag{18}$$

This update rule is a modification of the Forward-Reflected-Backward approach, namely, here α_k is a parameter, while in [83], $\alpha_k \equiv 1$. The analysis of method (18) gives optimal estimates in both the strongly monotone and monotone cases.

The theoretical results and guarantees discussed above rely in significant manner on the bounded variance assumption (Assumption 3). This assumption is quite restrictive (especially when the domain is unbounded) and does not hold for many popular machine learning problems. Moreover, one can even design a strongly monotone variational inequality on an unbounded domain such that method (14) *diverges* exponentially fast [26]. The authors of [48, 55] consider a relaxed form of the bounded variance condition and assume that $\mathbb{E}\|F_\xi(u) - F(u)\|_2^2 \le \sigma^2 + \delta\|u - z^*\|_2^2$ with $\delta \ge 0$ in the Euclidean case. Under this condition and Assumptions 1 and 2, it is proved in [48] that after k iterations of algorithm (14) (when $\mathcal{Z} = \mathbb{R}^d$) it holds that

$$\mathbb{E}\|z^k - z^*\|_2^2 = \mathcal{O}\Big(\kappa R_0^2 \exp\Big(-\frac{k}{\kappa}\Big) + \frac{\sigma^2}{\mu^2 k}\Big), \tag{19}$$

where $\kappa = \max\{\frac{\delta}{\mu^2}; \frac{L+\sqrt{\delta}}{\mu}\}$. The same assumption on stochastic realizations was considered in [67], where method (18) was used, yielding the estimate

$$\mathbb{E}\|z^k - z^*\|_2^2 = \mathcal{O}\Big(R_0^2 \exp\Big(-\frac{\mu k}{L}\Big) + \frac{\sigma^2 + \delta^2 D_{\mathcal{Z}}^2}{\mu^2 k}\Big). \tag{20}$$

Estimates (19) and (20) are competitive: the former is superior in terms of the stochastic term (second term), while the latter is superior in terms of the deterministic term (first term). However, none of these results deals completely with the issue of bounded noise, because the condition considered above is not general. The key to avoiding the bounded variance assumption on F_ξ lies in the way how stochasticity is generated in method (14). Method (14) is sometimes called Independent Samples Stochastic Extragradient (I-SEG). To address the bounded variance issue, K. Mishchenko et al. [86] proposed another stochastic modification of the Extragradient algorithm, called Same Sample Stochastic Extragradient (S-SEG):

$$z^{k+1/2} = z^k - \gamma F_{\xi^k}(z^k),$$
$$z^{k+1} = z^k - \gamma F_{\xi^k}(z^{k+1/2}).$$

For simplicity, we present the above method for the case when $\mathcal{Z} = \mathbb{R}^d$ ($F(x^*) = 0$), and refer the reader to [86] for a more general case of regularized VIs. In contrast to I-SEG, S-SEG uses the same sample ξ^k for both steps at iteration k. Although such a strategy cannot be implemented in some scenarios (streaming oracle), it can be applied to finite-sum problems, which have been gaining an increasing attention in the recent years. Moreover, S-SEG relies in significant manner on the following assumption.

Assumption 4. The operator $F_\xi(z)$ is L-Lipschitz and μ-strongly monotone almost surely in ξ, i.e., $\|F_\xi(z) - F_\xi(z')\|_2 \le L\|z - z'\|_2$ and $\langle F_\xi(z) - F_\xi(z'), z - z'\rangle \ge \mu\|z - z'\|_2^2$ for all $z, z' \in \mathbb{R}^d$, almost surely in ξ.

The evident difference between the I-SEG and S-SEG setups can be explained through the connection between the Extragradient and the Proximal Point (PP) methods [84, 107]. In the rest of this subs-section we assume that $\mathcal{Z} = \mathbb{R}^d$ ($F(z^*) = 0$). In this setup, PP has the update rule

$$z^{k+1} = z^k - \gamma F(z^{k+1}).$$

The method converges for any monotone operator F and any $\gamma > 0$. However, the update rule of PP is implicit and in many situations it cannot be computed efficiently. The Extragradient method can be seen as a natural approximation of PP that substitutes z^{k+1} in the right-hand side by one gradient step from z^k:

$$z^{k+1} = z^k - \gamma F(z^k - \gamma F(z^k)).$$

In addition, when F is L-Lipschitz, one can estimate how good the approximation is. Consider $z^{k+1} = z^k - \gamma F(z^k - \gamma F(z^k))$ (the Extragradient step) and $\tilde{z}^{k+1} = z^k - \gamma F(\tilde{z}^{k+1})$ (the PP step). Then $\|z^{k+1} - \tilde{z}^{k+1}\|_2$ can be estimated as follows [86]:

$$\begin{aligned}\|z^{k+1} - \tilde{z}^{k+1}\|_2 &= \gamma\|F(z^k - \gamma F(z^k)) - F(\tilde{z}^{k+1})\|_2 \\ &\le \gamma L\|z^k - \gamma F(z^k) - \tilde{z}^{k+1}\|_2 = \gamma^2 L\|F(z^k) - F(\tilde{z}^{k+1})\|_2 \\ &\le \gamma^2 L^2\|z^k - \tilde{z}^{k+1}\|_2 = \gamma^3 L^2\|F(\tilde{z}^{k+1})\|_2 \\ &\le \gamma^3 L^3\|\tilde{z}^{k+1} - z^*\|_2.\end{aligned}$$

That is, the difference between the Extragradient and PP steps is of the order $\mathcal{O}(\gamma^3)$ rather than $\mathcal{O}(\gamma^2)$. Since the latter corresponds to the difference between PP and the simple gradient step (7), the Extragradient method approximates PP better than gradient steps, which are known to be non-convergent for general monotone Lipschitz variational inequalities. This approximation feature of the Extragradient method is crucial for its convergence and, as the above derivation implies, the approximation argument significantly relies on the Lipschitzness of the operator F.

Let us go back to the differences between I-SEG and S-SEG. In S-SEG, the k-th iteration can be regarded as a single Extragradient step for the operator $F_{\xi^k}(z)$. Therefore, Lipschitzness and monotonicity of $F_{\xi^k}(z)$ (Assumption 4) are important for the analysis of S-SEG. In contrast, I-SEG uses different operators for the extrapolation and update steps. In this case, there is no effect from the Lipschitzness/monotonicity of individual $F_\xi(z)$s. Therefore, the analysis of I-SEG naturally relies on the Lipschitzness and monotonicity of $F(z)$ as well as on the closeness (on average) of $F_\xi(z)$ and $F(z)$ (Assumption 3).

The convergence of I-SEG was discussed earlier in this section. Regarding S-SEG, one has the following result [86].

Theorem 5. *Let Assumption* 4 *hold. Then there exists a choice of step size γ (see* [48]*) such that the output of S-SEG after k iterations satisfies*

$$\mathbb{E}\|z^k - z^*\|_2^2 = \mathcal{O}\Big(\frac{LR_0^2}{\mu} \exp\Big(-\frac{\mu k}{L}\Big) + \frac{\sigma_*^2}{\mu^2 k}\Big),$$

where $\sigma_*^2 = \mathbb{E}\|F_\xi(z^*)\|_2^2$.

This rate is similar to the one known for I-SEG, with the following differences. First, instead of the uniform bound on the variance σ^2, the rate depends on σ_*^2, which is the variance of F_ξ measured at the solution. In many cases, $\sigma^2 = \infty$, while σ_*^2 is finite. From this perspective, S-SEG enjoys a better rate of convergence than I-SEG. However, this comes at a price: while the rate of I-SEG depends on the Lipschitz and strong-monotonicity constants of F, the rate of S-SEG depends on *the worst* constants of F_ξ, which can be much worse than those for F. In particular, consider the finite-sum setup with uniform sampling of ξ: $F(x) = \frac{1}{n}\sum_{i=1}^n F_i(x)$, where F_i is L_i-Lipschitz and μ_i-strongly monotone, and $\mathbb{P}\{\xi = i\} = \frac{1}{n}$. Then Assumption 4 holds with $L = \max_{1\le i\le n} L_i$ and $\mu = \min_{1\le i\le n}\mu_i$ and these constants appear in the rate from Theorem 3. The authors of [48] tighten this rate. In particular, they prove that for S-SEG with different step sizes for the extrapolation and update steps one has that

$$\mathbb{E}\|z^k - z^*\|_2^2 = \mathcal{O}\Big(\frac{LR_0^2}{\overline{\mu}}\exp\Big(-\frac{\overline{\mu}k}{L}\Big) + \frac{\sigma_*^2}{\overline{\mu}^2 k}\Big),$$

where $\sigma_*^2 = \frac{1}{n}\sum_{i=1}^n \|F_i(z^*)\|_2^2$ and $\overline{\mu} = \frac{1}{n}\sum_{i=1}^n \mu_i$. Since $\overline{\mu}$ is (sometimes considerably) larger than μ, the improvement is noticeable. Moreover, when the constants $\{L_i\}_{i=1}^n$ are known, one can consider the so-called *importance sampling* [52]: $\mathbb{P}\{\xi = i\} = L_i/(n\overline{L})$, where $\overline{L} = \frac{1}{n}\sum_{i=1}^n L_i$. As the authors of [48] show, importance sampling can be combined with S-SEG by allowing the extrapolation and update step sizes at the k-th iteration to depend on the sample ξ^k. In particular, for the proposed modification of S-SEG they derive the estimate

$$\mathbb{E}\|z^k - z^*\|_2^2 = \mathcal{O}\Big(\frac{\overline{L}R_0^2}{\overline{\mu}}\exp\Big(-\frac{\overline{\mu}k}{\overline{L}}\Big) + \frac{\hat{\sigma}_*^2}{\overline{\mu}^2 k}\Big),$$

where $\hat{\sigma}_*^2 = \frac{1}{n}\sum_{i=1}^n \frac{\overline{L}}{L_i}\|F_i(z^*)\|_2^2$. The exponentially decaying term is always better than the corresponding one for S-SEG with uniform sampling. This usually implies faster convergence during the initial stage. Next, typically, a larger norm of $F_i(z^*)$ implies larger L_i, e.g., $\|F_i(z^*)\|_2^2 \sim L_i^2$. In this case, $\hat{\sigma}_*^2 \le \sigma_*^2$, because

$$\hat{\sigma}_*^2 \sim (\overline{L})^2 \quad \text{and} \quad \sigma_*^2 \sim \overline{L^2} = \frac{1}{n}\sum_{i=1}^n L_i^2 \ge (\overline{L})^2.$$

Moreover, one can allow other sampling strategies and cover the case when some μ_i are negative, see [48] for the details.

4.2 Finite-sum case

As noted earlier, when we deal with problem (13), it is often the case (especially in practical problems) that the distribution $\mathcal{D}$ is unknown, but nevertheless some samples from $\mathcal{D}$ are available. Then one can replace problem (13) by a finite-sum approximation:

$$F(z) = \frac{1}{n}\sum_{i=1}^n F_i(z).$$

This approximation is sometimes also called Monte Carlo approximation. For machine learning problems the term empirical risk is often encountered. Although calls of the full operator are now tractable, they remain expensive in practice. Therefore, it is worth avoiding frequent computation of F and mainly use calls to single F_i operators or small batches of them.

Before presenting the results, let us introduce the appropriate analogue of the Lipschitzness assumption.

Assumption 5 (Lipschitzness in the mean). The operator F is L_{avg}-Lipschitz continuous in mean, i.e., for all $u, v \in \mathcal{Z}$, we have

$$\mathbb{E}[\|F_\xi(u) - F_\xi(v)\|_*^2] \le L_{\mathrm{avg}}^2\|u - v\|^2.$$

For example, if F_i is L_i-Lipschitz for all i and we draw the index $\xi = i$ with probability $p_i = L_i/\sum_j L_j$, then

$$L_{\mathrm{avg}} = \frac{1}{n}\sum_j L_j.$$

The study of finite-sum problems in stochastic optimization is connected, first of all, with classical methods for minimization problems such as SVRG [59] and SAGA [29]. For the saddle point problems, these methods were adopted in [102] (in fact, these results are also valid for variational inequalities). The authors considered strongly convex-strongly concave saddles in the Euclidean case and proved the following estimates for SVRG and SAGA:

$$\mathbb{E}\|z^k - z^*\|_2^2 = \mathcal{O}\Big(R_0^2\exp\Big(-\min\Big\{\frac{1}{n}, \frac{\mu^2}{L_{\mathrm{avg}}^2}\Big\}k\Big)\Big).$$

Since this last bound is not tight in terms of L_{avg}/μ, the authors proposed accelerating SVRG and SAGA via the Catalyst envelope [76]. In this case, they obtain the bound

$$\mathbb{E}\|z^k - z^*\|_2^2 = \mathcal{O}\Big(R_0^2\exp\Big(-\min\Big\{\frac{1}{n}; \frac{\mu}{\sqrt{n}L_{\mathrm{avg}}}\Big\}\frac{k}{\log[L_{\mathrm{avg}}/\mu]}\Big)\Big). \tag{21}$$

The same estimates for methods for saddle point problems based on accelerating envelopes were also presented in [118].

An important step in the study of the finite-sum stochastic setup was taken in the work [25], which is primarily focused on bilinear games. For this class of problems, the authors improved estimate (21) and removed the additional logarithmic factor. For general problems (saddle point and variational inequalities) the results of [25] are very similar to those in (21) and also include an additional logarithmic factor. The authors also considered the convex-concave/monotone case in the non-Euclidean setting and found that for their method after k iterations it holds that

$$\mathbb{E}[\mathrm{Gap}_{\mathrm{VI}}(\hat{z}^k)] = \tilde{\mathcal{O}}\Big(\frac{\sqrt{n}L_{\mathrm{avg}}D_{\mathcal{Z},V}^2}{k}\Big). \tag{22}$$

The issue of the additional logarithmic factor was resolved in [2], where the following modification of the Extragradient method was proposed:

$$
\begin{aligned}
z^{k+1/2} &= P_{\mathcal{Z}}(z^k + \tau(w^k - z^k) - \gamma F(w^k)),\\
\Delta^k &= F_{\xi^k}(z^{k+1/2}) - F_{\xi^k}(w^k) + F(w^k),\\
z^{k+1} &= P_{\mathcal{Z}}(z^k + \tau(w^k - z^k) - \gamma \Delta^k) \qquad (23)\\
w^{k+1} &= \begin{cases} z^{k+1}, & \text{with probability } p,\\ w^k, & \text{with probability } 1-p. \end{cases}
\end{aligned}
$$

This algorithm is a combination of the extra step technique from the theory of VIs and the loopless approach [73] for finite-sum problems. An interesting ingredient of the method is the randomized negative momentum: $\tau(w^k - z^k)$. While for minimization problems it is usual to apply a positive/heavy-ball momentum, the opposite approach proves useful for saddle point problems and variational inequalities. This effect was noticed earlier [3, 42, 122] and and is encountered now in the theory of stochastic methods for VIs. Also, in [2], the authors presented modifications for the Forward-Backward, Forward-Reflected-Backward as well as for the Extragradient methods in the non-Euclidean case.

As we noted earlier, the results of [2] give estimates (21) and (22), but without additional logarithmic factors. That is, to achieve

$$
\begin{aligned}
\mathbb{E}\|z^k - z^*\|_2^2 \le \varepsilon \quad &\text{in the strongly monotone case,}\\
\mathbb{E}[\mathrm{Gap}_{VI}(\hat{z}^k)] \le \varepsilon \quad &\text{in the monotone case,}
\end{aligned}
$$

the methods from [2] require

$$
\mathcal{O}\Big(\max\Big\{n; \frac{\sqrt{n}L_{\mathrm{avg}}}{\mu}\Big\}\log\frac{R_0^2}{\varepsilon}\Big) \qquad (24)
$$

and

$$
\mathcal{O}\Big(\frac{\sqrt{n}L_{\mathrm{avg}}D^2_{\mathcal{Z},V}}{\varepsilon}\Big) \qquad (25)
$$

stochastic oracle calls, respectively. It remains to discuss the effect of batching on the method from (23), i.e., see how the oracle complexity bounds change if instead a single sample F_{ξ^k} at each iteration we use but a batch size of b: $\frac{1}{b}\sum_{i\in S^k} F_i$, where $S_k \subseteq \{1,\dots,n\}$ is the set of cardinality b of indices in the mini-batch. In this case, the methods from [2] give estimates (24) and (25), but multiplied by an additional factor $\sqrt{b}$. This extra multiplier issue was resolved in [69] using the following method:

$$
\begin{aligned}
\Delta^k &= \frac{1}{b}\sum_{i\in S^k}\big[F_i(z^k) - F_i(w^{k-1})\\
&\qquad + \alpha(F_i(z^k) - F_i(z^{k-1}))\big] + F(w^{k-1}),\\
z^{k+1} &= P_{\mathcal{Z}}(z^k + \tau(w^k - z^k) - \gamma\Delta^k),\\
w^{k+1} &= \begin{cases} z^{k+1}, & \text{with probability } p,\\ w^k, & \text{with probability } 1-p. \end{cases}
\end{aligned}
$$

The authors proved that in the strongly monotone case this method gives estimate (24), i.e., without additional logarithmic factors and without factors depending on b.

The only issue that remains to be understood is whether the current state-of-the-art methods with best complexities from [2, 69] are optimal. The lower bounds from [53] claim that under Assumptions 5 and 2, the methods above are optimal. However, under $L_{\max}$-Lipschitzness of all F_i, $i \in \{1,\dots,n\}$ and Assumption 2, the lower bound from [53] is

$$
\mathbb{E}\|z^k - z^*\|_2^2 = \Omega\Big(R_0^2\exp\Big(-\min\Big\{\frac{1}{n}, \frac{\mu}{L_{\max}}\Big\}k\Big)\Big).
$$

The question whether this lower bound is tight remains open.

4.3 *Cocoercivity assumption*

In some papers, the following assumption is used instead of Assumption 1.

Assumption 6 (Cocoercivity). The operator F is ℓ-cocoercive, i.e., for all $u, v \in \mathcal{Z}$, we have $\|F(u) - F(v)\|_2^2 \le \ell\langle F(u) - F(v), u - v\rangle$.

Cocoercivity is stronger than monotonicity + Lipschitzness, i.e., not all monotone Lipschitz operators are cocoercive. Note, for instance, that the operator for the bilinear SPP ($\min_x \max_y x^\top Ay$) is not cocoercive. However, if F is L-Lipschitz and μ-strongly monotone, then it is (L^2/μ)-cocoercive. Moreover, the operator corresponding to a convex L-smooth minimization problem is L-cocoercive.

There is no need to use an Extragradient for cocoercive operators. One can apply the iterative scheme (7) and its modifications for the stochastic case. In spite of this, the first work on cocoercive operators in the stochastic cases used the Extragradient as the basic method [26]. In this paper, the authors investigated methods for finite-sum problems. The subsequent results from [15, 81] give an almost complete picture of stochastic algorithms based on method (7) for operators under Assumption 6. In particular, the work [15] provides a unified analysis for a large number of popular stochastic methods currently known for minimization problems [51].

4.4 *High-probability convergence*

Up to this point, we focused on convergence-in-expectation guarantees for stochastic methods, i.e., bounds on $\mathbb{E}[\mathrm{Gap}_{\mathrm{VI}}(\hat{z}^k)]$ and/or $\mathbb{E}\|z^k - z^*\|_2^2$. However, *high-probability convergence guarantees*, i.e., bounds on $\mathrm{Gap}_{\mathrm{VI}}(\hat{z}^k)$ and/or $\|z^k - z^*\|_2^2$ that hold with probability at least $1-\beta$ for a given confidence level $\beta \in (0,1)$, reflect the real behavior of the methods more accurately [50]. Despite this fact, high-probability convergence of stochastic methods for solving VIs is studied only in a couple of works.

It is worth mentioning that one can always deduce the high-probability bound from the in-expectation one via Markov's inequality. However, in this case, the derived rate of convergence will have a negative-power dependence on β^{-1}. Such guarantees are not desirable and the goal is to derive the rates that have a (poly-) logarithmic dependence on the confidence level, i.e., β should appear only in the $\mathcal{O}(\mathrm{poly}(\log(\frac{1}{\beta})))$ factor.

The first, and for many years the only high-probability guarantees of this type for solving stochastic VIs were derived in [60]. The authors assume that F is monotone and L-Lipschitz, the underlying domain is bounded, and F_ξ is an unbiased estimator with sub-Gaussian (light) tails of the distribution:

$$\mathbb{E}\Big[\exp\Big(\frac{\|F_\xi(x) - F(x)\|_2^2}{\sigma^2}\Big)\Big] \le \exp(1).$$

The above condition is much stronger than Assumption 3. Under the listed assumptions, the authors of [60] prove that after k iterations of Mirror-Prox with probability at least $1-\beta$ (for any $\beta \in (0,1)$) the following inequality is in force:

$$\mathrm{Gap}_{\mathrm{VI}}(\hat{z}^k) = \mathcal{O}\Big(\frac{LD_{\mathcal{Z}}^2}{k} + \frac{\sigma D_{\mathcal{Z}} \log(1/\beta)}{\sqrt{k}}\Big).$$

Up to the logarithmic factor this result coincides with in-expectation one and, thus, it is optimal (up to the logarithms). However, the result is derived under the restrictive light-tails assumption.

This last limitation was recently addressed in [49], where the authors derive the high-probability rates for the considered problem under just the bounded variance assumption. In particular, they consider the clipped-SEG for problems with $\mathcal{Z} = \mathbb{R}^d$:

$$z^{k+1/2} = z^k - \gamma \cdot \mathrm{clip}(F_{\xi^k}(z^k), \lambda_k),$$
$$z^{k+1} = z^k - \gamma \cdot \mathrm{clip}(F_{\xi^{k+1/2}}(z^{k+1/2}), \lambda_{k+1/2}),$$

where $\mathrm{clip}(x,\lambda) = \min\{1, \lambda/\|x\|_2\}x$ is the clipping operator, a popular tool in deep learning [46,103]. In the setup when F is monotone and L-Lipschitz and Assumption 3 holds, in [49] it is proved that after k iterations of clipped-SEG with probability at least $1-\beta$ (for any $\beta \in (0,1)$) the following inequality holds:

$$\mathrm{Gap}_{\mathrm{VI}}(\hat{z}^k) = \mathcal{O}\Big(\frac{LR_0^2 \log(k/\beta)}{k} + \frac{\sigma R_0 \sqrt{\log(k/\beta)}}{\sqrt{k}}\Big).$$

Up to the differences in logarithmic factors, the definition of σ, and the difference between $D_{\mathcal{Z}}$ and R_0, the rate coincides with the one from [60], but it was derived without the light-tails assumption. The key algorithmic tool that allows removing the light-tails assumption is clipping: with a proper choice of the clipping level λ the authors cut heavy tails without making the bias too large. It is worth mentioning that the result for clipped-SEG is derived for the unconstrained case and the rate depends on R_0, while in [60], the analysis relies on the boundedness of the domain, the diameter of which appears explicitly in the rate obtained. To remove the dependence on the diameter of the domain, the authors of [48] show that with high probability the iterates produced by clipped-SEG stay in the ball around x^* with a radius proportional to R_0. Using this trick, they also show that it is sufficient that all the assumptions (monotonicity and Lipschitzness of F and bounded variance) hold just on this ball. Such a degree of generality allows them to cover problems that are non-Lipschitz on $\mathbb{R}^d$ (e.g., for certain monotone polynomially growing operators) and also the situation when the variance is bounded only on a compact set, which is common for many finite-sum problems. Finally, [48] contains high-probability convergence results for strongly monotone VIs and VIs with structured non-monotonicity.

5 Recent advances

In this section, we report briefly on a few recent theoretical advances with practical impacts.

5.1 Saddle point problems with different constants of strong convexity and strong concavity

Saddle point problems with different constants of strong convexity and strong concavity started gaining interest a few years ago, see e.g., [4, 77]. However, even for the particular case

$$\min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} g(x,y) = f(x) + y^\top \mathbf{A} x - h(y),$$

where the function f is μ_x-strongly convex ($\mu_x > 0$) and L_x-smooth, and the function h is μ_y-strongly convex ($\mu_y > 0$) and L_y-smooth, optimal algorithms have been proposed only recently [58, 72, 116]. These algorithms have the convergence rates

$$\mathcal{O}\Big(\Big(\sqrt{\frac{L_x}{\mu_x}} + \sqrt{\frac{\lambda_{\max}(\mathbf{A}^\top \mathbf{A})}{\mu_x \mu_y}} + \sqrt{\frac{L_y}{\mu_y}}\Big) \log \frac{1}{\varepsilon}\Big)$$

and attain the lower bound, which was obtained in [56, 124] (here one needs to assume that $\lambda_{\min}(\mathbf{A}^\top \mathbf{A}) \le \sqrt{\mu_x \mu_y}$; without this assumption no optimal methods are known).

Note that the algorithm from [58] is built upon a technique related to the analysis of primal-dual Extragradient methods via relative Lipschitzness [28, 115]. As a by-product, this technique makes it possible to obtain Nesterov's accelerated method as a particular case of primal-dual Extragradient method with relative Lipschitzness [28].

For the non-bilinear SPP, optimal methods, based on the accelerated Monteiro–Svaiter proximal envelope, were developed only in the non-composite case [24, 71]. For the non-bilinear SPP with composite terms, there is a poly-logarithmic gap between the lower bound and the best known upper bounds [118]. A gap also appears for the SPP with stochastic finite-sum structure [58, 82, 118]. The stochastic setting with bounded variance was considered in [32, 85, 125].

Further deterministic "cutting-plane" improvements are connected with the additional assumptions about small dimension of the involved vectors x or/and y (see [43, 44, 91]) or with different structural (e.g., SPP on balls in 1- or ∞-norms) and sparsity assumptions, see e.g., [21, 111, 112] and references therein. Here lower bounds are mostly unknown.

In this subsection we mentioned many works dealing with (sub-)optimal algorithms for different variants of SPP. We note that, in contrast to convex optimization, where the oracle call is uniquely associated with the gradient call $\nabla f(x)$, for SPP we have two criteria: the number of $\nabla_x g(x,y)$-calls and that of $\nabla_y g(x,y)$-calls (and more variants for SPP with composites). "Optimality" in the most of the aforementioned papers means that the method is optimal according to the worst of the criteria. In [4, 118], the authors consider these criteria separately. However, the development of the lower bounds and optimal methods for a multi-criterion setup is still an open problem.

5.2 *Adaptive methods for VI and SPP*

Interest in adaptive algorithms for stochastic convex optimization mainly arose in 2011 after the development of the AdaGrad (adaptive gradient) [33] and Adam (adaptive moment estimation) [63] algorithms. For variational inequalities and saddle point problems, people became interested in adaptive methods only in the last few years, see, e.g., [8,40] (see also [61]). Currently, this area of research is well developed. One can mention here works devoted to both adaptive step sizes [5, 34, 35, 114, 117] and adaptive scaling/preconditioning [12, 31, 80]. Approaches from the second group are based on the idea of a proper combination of AdaGrad/Adam with Extragradient or its modifications. All of the mentioned adaptive methods have no better (typically the same) theoretical rates of convergence than their non-adaptive analogues, but require less input information or demonstrate better performance in practice.

5.3 *Quasi-Newton and tensor methods for VI and SPP*

Quasi-Newton methods for solving nonlinear equations (unconstrained VI) and SPP are proposed in [75, 121] and [79], respectively. In these papers, local superlinear rates of convergence are derived for the modifications of the Broyden-type methods for solving nonlinear equations with Lipschitz Jacobian and SPP with Lipschitz Hessian. Stochastic versions of these methods for VI and SPP still await to be developed.

Tensor methods for convex optimization problems are currently quite well developed. In particular, starting with [99] it has been shown that optimal second- and third-order methods can be implemented with almost the same complexity of each iteration as the Newton method [39, 89, 97]. Moreover, optimal p-order methods (which use p-order derivatives) significantly reduce the rate of convergence from k^{-2} to $k^{-(3p+1)/2}$ (see [23, 70]). For VI and SPP, the study was initiated in [88, 94] and optimal p-order methods reduce the rate of convergence from k^{-1} to $k^{-(p+1)/2}$ (see [1, 78]) (for k^{-1}, see Theorem 3). However, in contrast to convex optimization, the use of tensor methods for sufficiently smooth monotone VIs and convex-concave saddle point problems is not expected to be as effective. Note that in [1, 78] one can also find optimal rates for strongly monotone VIs and strongly convex-concave SPP. Stochastic tensor methods for variational inequalities and saddle point problems still await to be developed.

5.4 *Convergence in terms of the gradient norm for SPP*

Several recent advances in the development of optimal algorithms are based on accelerated proximal envelopes with proper stopping rules for inner loop algorithms [68, 70, 71, 109]. Such rules are built upon the norm of the gradient calculated for the target function of the inner problem.

For smooth convex optimization problems, Yu. Nesterov in 2012 posed the problem of making the gradient norm small with the same rate of convergence as a gap in the function values, i.e., proportional to k^{-2} (see [96]). To address this problem, in [96] he proposed an optimal (up to a logarithmic factor) algorithm. This question was further investigated, leading to optimal results without additional logarithmic factors [62, 98] (see also [30] for explanations and a survey). In the stochastic case, algorithms were presented in [38].

For smooth convex-concave saddle point problems an optimal algorithm with $\|\nabla_{x,y} f(x^k, y^k)\|_2$ proportional to k^{-1} was proposed in [122] (see also [30] and [71] for monotone inclusion). For the stochastic case, see [20, 27, 74].

5.5 *Decentralized VI and SPP*

In practice, in order to solve a variational inequality problem more efficiently and quickly, one usually resorts to distributed methods. In particular, methods that work on arbitrary (possibly time-varying) decentralized communication networks between computing devices are popular.

While the field of decentralized algorithms for minimization problems has been extensively investigated, results for broader classes of problems have only begun to appear in recent years. Such works are primarily focused on saddle point problems [16–18, 90, 108], but we note that most of these results can easily be extended to variational inequalities. Let us emphasize two works that were from the outset devoted to VIs. In [13], the authors proposed a decentralized method with local steps, and [69] presented optimal decentralized methods for stochastic (finite-sum) variational inequalities on fixed and varying networks.

Acknowledgements. The work was supported by Russian Science Foundation (project No. 21-71-30005).

References

[1] D. Adil, B. Bullins, A. Jambulapati and S. Sachdeva, Line search-free methods for higher-order smooth monotone variational inequalities, preprint, arXiv:2205.06167 (2022)

[2] A. Alacaoglu and Y. Malitsky, Stochastic variance reduction for variational inequality methods, preprint, arXiv:2102.08352 (2021)

[3] A. Alacaoglu, Y. Malitsky and V. Cevher, Forward-reflected-backward method with variance reduction. *Comput. Optim. Appl.* **80**, 321–346 (2021)

[4] M. S. Alkousa, A. V. Gasnikov, D. M. Dvinskikh, D. A. Kovalev and F. S. Stonyakin, Accelerated methods for saddle-point problem. *Comput. Math. Math. Phys.* **60**, 1787–1809 (2020)

[5] K. Antonakopoulos, E. V. Belmega and P. Mertikopoulos, Adaptive extra-gradient methods for min-max optimization and games, preprint, arXiv:2010.12100 (2020)

[6] K. J. Arrow, L. Hurwicz and H. Uzawa, *Studies in linear and non-linear programming*. Stanford Mathematical Studies in the Social Sciences, II, Stanford University Press, Stanford (1958)

[7] W. Azizian, D. Scieur, I. Mitliagkas, S. Lacoste-Julien and G. Gidel, Accelerating smooth games by manipulating spectral shapes. In *International Conference on Artificial Intelligence and Statistics*, Proc. Mach. Learn. Res., 1705–1715 (2020)

[8] F. Bach and K. Y. Levy, A universal algorithm for variational inequalities adaptive to smoothness and noise. In *Conference on Learning Theory*, Proc. Mach. Learn. Res., 164–194 (2019)

[9] F. Bach, J. Mairal and J. Ponce, Convex sparse matrix factorizations, preprint, arXiv:0812.1869 (2008)

[10] A. Bakushinskii and B. Polyak, On the solution of variational inequalities. *Sov. Math. Dokl.* **15**, 1705–1710 (1974)

[11] A. Ben-Tal, L. El Ghaoui and A. Nemirovski, *Robust optimization*. Princeton Ser. Appl. Math., Princeton University Press, Princeton (2009)

[12] A. Beznosikov, A. Alanov, D. Kovalev, M. Takáč and A. Gasnikov, On scaled methods for saddle point problems, preprint, arXiv:2206.08303 (2022)

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

[14] A. Beznosikov, A. Gasnikov, K. Zainulina, A. Maslovskiy and D. Pasechnyuk, A unified analysis of variational inequality methods: Variance reduction, sampling, quantization and coordinate descent, preprint, arXiv:2201.12206 (2022)

[15] A. Beznosikov, E. Gorbunov, H. Berard and N. Loizou, Stochastic gradient descent-ascent: Unified theory and new efficient methods, preprint, arXiv:2202.07262 (2022)

[16] A. Beznosikov, A. Rogozin, D. Kovalev and A. Gasnikov, Near-optimal decentralized algorithms for saddle point problems over time-varying networks. In *Optimization and applications*, Lecture Notes in Comput. Sci. 13078, Springer, Cham, 246–257 (2021)

[17] A. Beznosikov, V. Samokhin and A. Gasnikov, Distributed saddle-point problems: Lower bounds, optimal and robust algorithms, preprint arXiv:2010.13112 (2020)

[18] A. Beznosikov, G. Scutari, A. Rogozin and A. Gasnikov, Distributed saddle-point problems under data similarity. *Adv. Neural Inf. Process. Syst.* **34**, 8172–8184 (2021)

[19] F. E. Browder, Existence and approximation of solutions of nonlinear variational inequalities. *Proc. Nat. Acad. Sci. U.S.A.* **56**, 1080–1086 (1966)

[20] X. Cai, C. Song, C. Guzmán and J. Diakonikolas, A stochastic halpern iteration with variance reduction for stochastic monotone inclusion problems, preprint, arXiv:2203.09436 (2022)

[21] Y. Carmon, Y. Jin, A. Sidford and K. Tian, Coordinate methods for matrix games. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science*, IEEE Computer Soc., Los Alamitos, 283–293 (2020)

[22] A. Chambolle and T. Pock, A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision* **40**, 120–145 (2011)

[23] Y. Carmon, D. Hausler, A. Jambulapati, Y. Jin and A. Sidford, Optimal and adaptive Monteiro–Svaiter acceleration, preprint, arXiv:2205.15371 (2022)

[24] Y. Carmon, A. Jambulapati, Y. Jin and A. Sidford, Recapp: Crafting a more efficient catalyst for convex optimization. In *International Conference on Machine Learning*, Proc. Mach. Learn. Res., 2658–2685 (2022)

[25] Y. Carmon, Y. Jin, A. Sidford and K. Tian, Variance reduction for matrix games, preprint, arXiv:1907.02056 (2019)

[26] T. Chavdarova, G. Gidel, F. Fleuret and S. Lacoste-Julien, Reducing noise in GAN training with variance reduced extragradient. *Adv. Neural Inf. Process. Syst.* **32**, 393–403 (2019)

[27] L. Chen and L. Luo, Near-optimal algorithms for making the gradient small in stochastic minimax optimization, preprint, arXiv:2208.05925 (2022)

[28] M. B. Cohen, A. Sidford and K. Tian, Relative Lipschitzness in extragradient methods and a direct recipe for acceleration. In *12th Innovations in Theoretical Computer Science Conference*, LIPIcs. Leibniz Int. Proc. Inform. 185, Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, Article No. 62, (2021)

[29] A. Defazio, F. Bach and S. Lacoste-Julien, SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Adv. Neural Inf. Process. Syst.* **27**, 1646–1654 (2014)

[30] J. Diakonikolas and P. Wang, Potential function-based framework for minimizing gradients in convex and min-max optimization. *SIAM J. Optim.* **32**, 1668–1697 (2022)

[31] Z. Dou and Y. Li, On the one-sided convergence of Adam-type algorithms in non-convex non-concave min-max optimization, preprint, arXiv:2109.14213 (2021)

[32] S. S. Du, G. Gidel, M. I. Jordan and C. J. Li, Optimal extragradient-based bilinearly-coupled saddle-point optimization, preprint, arXiv:2206.08573 (2022)

[33] J. Duchi, E. Hazan and Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **12**, 2121–2159 (2011)

[34] A. Ene and H. L. Nguyen, Adaptive and universal algorithms for variational inequalities with optimal convergence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 36, 6559–6567 (2022)

[35] A. Ene, H. L. Nguyen and A. Vladu, Adaptive gradient methods for constrained convex optimization and variational inequalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 35, 7314–7321 (2021)

[36] E. Esser, X. Zhang and T. F. Chan, A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM J. Imaging Sci.* **3**, 1015–1046 (2010)

[37] F. Facchinei and J.-S. Pang, *Finite-dimensional variational inequalities and complementarity problems*. Springer Series in Operations Research, Springer, New York (2003)

[38] D. J. Foster, A. Sekhari, O. Shamir, N. Srebro, K. Sridharan and B. Woodworth, The complexity of making the gradient small in stochastic convex optimization. In *Conference on Learning Theory*, Proc. Mach. Learn. Res., 1319–1345 (2019)

[39] A. Gasnikov, P. Dvurechensky, E. Gorbunov, E. Vorontsova, D. Selikhanovych, C. A. Uribe, B. Jiang, H. Wang, S. Zhang, S. Bubeck et al., Near optimal methods for minimizing convex functions with Lipschitz p-th derivatives. In *Conference on Learning Theory*, Proc. Mach. Learn. Res., 1392–1393 (2019)

[40] A. V. Gasnikov, P. E. Dvurechensky, F. S. Stonyakin and A. A. Titov, An adaptive proximal method for variational inequalities. *Comput. Math. Math. Phys.* **59**, 836–841 (2019)

[41] G. Gidel, H. Berard, G. Vignoud, P. Vincent and S. Lacoste-Julien, A variational inequality perspective on generative adversarial networks, preprint, arXiv:1802.10551 (2018)

[42] G. Gidel, R. A. Hemmat, M. Pezeshki, R. Le Priol, G. Huang, S. Lacoste-Julien and I. Mitliagkas, Negative momentum for improved game dynamics. In *The 22nd International Conference on Artificial Intelligence and Statistics*, Proc. Mach. Learn. Res., 1802–1811 (2019)

[43] E. Gladin, I. Kuruzov, F. Stonyakin, D. Pasechnyuk, M. Alkousa and A. Gasnikov, Solving strongly convex-concave composite saddle point problems with a small dimension of one of the variables, preprint, arXiv:2010.02280 (2022)

[44] 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*, Commun. Comput. Inf. Sci. 1476, Springer, Cham, 19–40 (2021)

[45] E. G. Gol′šteĭn, Convergence of the gradient method for finding the saddle points of modified Lagrangian functions. *Èkonom. i Mat. Metody* **13**, 322–329 (1977)

[46] I. Goodfellow, Y. Bengio and A. Courville, *Deep learning*. Adaptive Computation and Machine Learning, MIT Press, Cambridge (2016)

[47] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, Generative adversarial networks. *Commun. ACM* **63**, 139–144 (2020)

[48] E. Gorbunov, H. Berard, G. Gidel and N. Loizou, Stochastic extragradient: General analysis and improved rates. In *International Conference on Artificial Intelligence and Statistics*, Proc. Mach. Learn. Res., 7865–7901 (2022)

[49] E. Gorbunov, M. Danilova, D. Dobre, P. Dvurechensky, A. Gasnikov and G. Gidel, Clipped stochastic methods for variational inequalities with heavy-tailed noise, preprint, arXiv:2206.01095 (2022)

[50] E. Gorbunov, M. Danilova and A. Gasnikov, Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Adv. Neural Inf. Process. Syst.* **33**, 15042–15053 (2020)

[51] E. Gorbunov, F. Hanzely and P. Richtárik, A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent. In *International Conference on Artificial Intelligence and Statistics*, Proc. Mach. Learn. Res., 680–690 (2020)

[52] R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin and P. Richtárik, SGD: General analysis and improved rates. In *Proceedings of the 36th International Conference on Machine Learning*, Proc. Mach. Learn. Res. 97, 5200–5209 (2019)

[53] Y. Han, G. Xie and Z. Zhang, Lower complexity bounds of finite-sum optimization problems: The results and construction, preprint, arXiv:2103.08280 (2021)

[54] Y.-G. Hsieh, F. Iutzeler, J. Malick and P. Mertikopoulos, On the convergence of single-call stochastic extra-gradient methods. *Adv. Neural Inf. Process. Syst.* **32**, 6938–6948 (2019)

[55] Y.-G. Hsieh, F. Iutzeler, J. Malick and P. Mertikopoulos, Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. *Adv. Neural Inf. Process. Syst.* **33**, 16223–16234 (2020)

[56] A. Ibrahim, W. Azizian, G. Gidel and I. Mitliagkas, Linear lower bounds and conditioning of differentiable games. In *International Conference on Machine Learning*, Proc. Mach. Learn. Res., 4583–4593 (2020)

[57] Y. Jin and A. Sidford, Efficiently solving MDPs with stochastic mirror descent. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, Proc. Mach. Learn. Res. 119, 4890–4900 (2020)

[58] Y. Jin, A. Sidford and K. Tian, Sharper rates for separable minimax and finite sum optimization via primal-dual extragradient methods, preprint, arXiv:2202.04640 (2022)

[59] R. Johnson and T. Zhang, Accelerating stochastic gradient descent using predictive variance reduction. *Adv. Neural Inf. Process. Syst.* **26**, 315–323 (2013)

[60] A. Juditsky, A. Nemirovski and C. Tauvel, Solving variational inequalities with stochastic mirror-prox algorithm. *Stoch. Syst.* **1**, 17–58 (2011)

[61] E. N. Khobotov, Modification of the extra-gradient method for solving variational inequalities and certain optimization problems. *U.S.S.R. Comput. Math. Math. Phys.* **27**, 120–127 (1987)

[62] D. Kim and J. A. Fessler, Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. *J. Optim. Theory Appl.* **188**, 192–219 (2021)

[63] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2305–2313 (2015)

[64] G. M. Korpelevič, An extragradient method for finding saddle points and for other problems. *Èkonom. i Mat. Metody* **12**, 747–756 (1976)

[65] G. M. Korpelevich, Extrapolation gradient methods and their relation to modified Lagrange functions. *Èkonom. i Mat. Metody* **19**, 694–703 (1983)

[66] G. Kotsalis, G. Lan and T. Li, Simple and optimal methods for stochastic variational inequalities. I: Operator extrapolation. *SIAM J. Optim.* **32**, 2041–2073 (2022)

[67] G. Kotsalis, G. Lan and T. Li, Simple and optimal methods for stochastic variational inequalities. II: Markovian noise and policy evaluation in reinforcement learning. *SIAM J. Optim.* **32**, 1120–1155 (2022)

[68] D. Kovalev, A. Beznosikov, E. Borodich, A. Gasnikov and G. Scutari, Optimal gradient sliding and its application to distributed optimization under similarity, preprint, arXiv:2205.15136 (2022)

[69] D. Kovalev, A. Beznosikov, A. Sadiev, M. Persiianov, P. Richtárik and A. Gasnikov, Optimal algorithms for decentralized stochastic variational inequalities, preprint, arXiv:2202.02771 (2022)

[70] D. Kovalev and A. Gasnikov, The first optimal acceleration of high-order methods in smooth convex optimization, preprint, arXiv:2205.09647 (2022)

[71] D. Kovalev and A. Gasnikov, The first optimal algorithm for smooth and strongly-convex-strongly-concave minimax optimization, preprint, arXiv:2205.05653 (2022)

[72] D. Kovalev, A. Gasnikov and P. Richtárik, Accelerated primal-dual gradient method for smooth and convex-concave saddle-point problems with bilinear coupling, preprint, arXiv:2112.15199 (2021)

[73] D. Kovalev, S. Horváth and P. Richtárik, Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, edited by A. Kontorovich and G. Neu, Proc. Mach. Learn. Res. 117, 451–467 (2020)

[74] S. Lee and D. Kim, Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. *Adv. Neural Inf. Process. Syst.* **34**, 22588–22600 (2021)

[75] D. Lin, H. Ye and Z. Zhang, Explicit superlinear convergence rates of Broyden's methods in nonlinear equations, preprint, arXiv: 2109.01974 (2021)

[76] H. Lin, J. Mairal and Z. Harchaoui, A universal catalyst for first-order optimization. *Adv. Neural Inf. Process. Syst.* **28**, 3384–3392 (2015)

[77] T. Lin, C. Jin and M. I. Jordan, Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, Proc. Mach. Learn. Res., 2738–2779 (2020)

[78] T. Lin, M. Jordan et al., Perseus: A simple high-order regularization method for variational inequalities, preprint, arXiv:2205.03202 (2022)

[79] C. Liu and L. Luo, Quasi-Newton methods for saddle point problems, preprint, arXiv:2111.02708 (2021)

[80] M. Liu, Y. Mroueh, J. Ross, W. Zhang, X. Cui, P. Das and T. Yang, Towards better understanding of adaptive gradient algorithms in generative adversarial nets, preprint, arXiv:1912.11940 (2019)

[81] N. Loizou, H. Berard, G. Gidel, I. Mitliagkas and S. Lacoste-Julien, Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity. *Adv. Neural Inf. Process. Syst.* **34**, 19095–19108 (2021)

[82] L. Luo, G. Xie, T. Zhang and Z. Zhang, Near optimal stochastic algorithms for finite-sum unbalanced convex-concave minimax optimization, preprint, arXiv:2106.01761 (2021)

[83] Y. Malitsky and M. K. Tam, A forward-backward splitting method for monotone inclusions without cocoercivity. *SIAM J. Optim.* **30**, 1451–1472 (2020)

[84] B. Martinet, Régularisation d'inéquations variationnelles par approximations successives. *Rev. Française Informat. Recherche Opérationnelle* **4**, 154–158 (1970)

[85] D. Metelev, A. Rogozin, A. Gasnikov and D. Kovalev, Decentralized saddle-point problems with different constants of strong convexity and strong concavity, preprint, arXiv:2206.00090 (2022)

[86] K. Mishchenko, D. Kovalev, E. Shulgin, P. Richtárik and Y. Malitsky, Revisiting stochastic extragradient. In *International Conference on Artificial Intelligence and Statistics*, Proc. Mach. Learn. Res., 4573–4582 (2020)

[87] A. Mokhtari, A. Ozdaglar and S. Pattathil, A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, Proc. Mach. Learn. Res., 1497–1507 (2020)

[88] R. D. C. Monteiro and B. F. Svaiter, Iteration-complexity of a Newton proximal extragradient method for monotone variational inequalities and inclusion problems. *SIAM J. Optim.* **22**, 914–935 (2012)

[89] R. D. C. Monteiro and B. F. Svaiter, An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM J. Optim.* **23**, 1092–1125 (2013)

[90] S. Mukherjee and M. Chakraborty, A decentralized algorithm for large scale min-max problems. In *2020 59th IEEE Conference on Decision and Control (CDC)*, 2967–2972 (2020)

[91] A. Nemirovski, Efficient methods in convex programming. Lecture notes, https://www2.isye.gatech.edu/~nemirovs/Lect_EMCO.pdf (1994)

[92] 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 J. Optim.* **15**, 229–251 (2004)

[93] Y. Nesterov, Smooth minimization of non-smooth functions. *Math. Program.* **103**, 127–152 (2005)

[94] Y. Nesterov, Cubic regularization of Newton's method for convex problems with constraints. CORE Discussion Paper No. 2006/39, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=921825 (2006)

[95] Y. Nesterov, Dual extrapolation and its applications to solving variational inequalities and related problems. *Math. Program.* **109**, 319–344 (2007)

[96] Y. Nesterov, How to make the gradients small. *Optima. Mathematical Optimization Society Newsletter* **88**, 10–11 (2012)

[97] Y. Nesterov, Implementable tensor methods in unconstrained convex optimization. *Math. Program.* **186**, 157–183 (2021)

[98] Y. Nesterov, A. Gasnikov, S. Guminov and P. Dvurechensky, Primal-dual accelerated gradient methods with small-dimensional relaxation oracle. *Optim. Methods Softw.* **36**, 773–810 (2021)

[99] Y. Nesterov and B. T. Polyak, Cubic regularization of Newton method and its global performance. *Math. Program.* **108**, 177–205 (2006)

[100] S. Omidshafiei, J. Pazis, C. Amato, J. P. How and J. Vian, Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, Proc. Mach. Learn. Res. 70, 2681–2690 (2017)

[101] Y. Ouyang and Y. Xu, Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Math. Program.* **185**, 1–35 (2021)

[102] B. Palaniappan and F. Bach, Stochastic variance reduction methods for saddle-point problems. *Adv. Neural Inf. Process. Syst.*, 1416–1424 (2016)

[103] R. Pascanu, T. Mikolov and Y. Bengio, On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, 1310–1318 (2013)

[104] B. T. Polyak, *Introduction to optimization*. Translations Series in Mathematics and Engineering, Optimization Software, Inc., Publications Division, New York (1987)

[105] L. D. Popov, A modification of the Arrow–Hurwicz method for search of saddle points. *Math. Notes* **28**, 845–848 (1981)

[106] R. T. Rockafellar, Convex functions, monotone operators and variational inequalities. In *Theory and Applications of Monotone Operators (Proc. NATO Advanced Study Inst., Venice, 1968)*, Edizioni "Oderisi", Gubbio, 35–65 (1969)

[107] R. T. Rockafellar, Monotone operators and the proximal point algorithm. *SIAM J. Control Optim.* **14**, 877–898 (1976)

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

[109] A. Sadiev, D. Kovalev and P. Richtárik, Communication acceleration of local gradient methods via an accelerated primal-dual algorithm with inexact prox. arXiv:2207.03957 (2022)

[110] M. Sibony, Méthodes itératives pour les équations et inéquations aux dérivées partielles non linéaires de type monotone. *Calcolo* **7**, 65–183 (1970)

[111] C. Song, C. Y. Lin, S. J. Wright and J. Diakonikolas, Coordinate linear variance reduction for generalized linear programming, preprint, arXiv:2111.01842 (2021)

[112] C. Song, S. J. Wright and J. Diakonikolas, Variance reduction via primal-dual accelerated dual averaging for nonsmooth convex finite-sums. In *International Conference on Machine Learning*, Proc. Mach. Learn. Res., 9824–9834 (2021)

[113] G. Stampacchia, Formes bilinéaires coercitives sur les ensembles convexes. *C. R. Acad. Sci. Paris* **258**, 4413–4416 (1964)

[114] F. Stonyakin, A. Gasnikov, P. Dvurechensky, A. Titov and M. Alkousa, Generalized mirror prox algorithm for monotone variational inequalities: Universality and inexact oracle. *J. Optim. Theory Appl.* **194**, 988–1013 (2022)

[115] 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. *Optim. Methods Softw.* **36**, 1155–1201 (2021)

[116] K. K. Thekumparampil, N. He and S. Oh, Lifted primal-dual method for bilinearly coupled smooth minimax optimization, preprint, arXiv:2201.07427 (2022)

[117] A. A. Titov, S. S. Ablaev, M. S. Alkousa, F. S. Stonyakin and A. V. Gasnikov, Some adaptive first-order methods for variational inequalities with relatively strongly monotone operators and generalized smoothness, preprint, arXiv:2207.09544 (2022)

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

[119] P. Tseng, On linear convergence of iterative methods for the variational inequality problem. *J. Comput. Appl. Math.* **60**, 237–252 (1995)

[120] P. Tseng, A modified forward-backward splitting method for maximal monotone mappings. *SIAM J. Control Optim.* **38**, 431–446 (2000)

[121] H. Ye, D. Lin and Z. Zhang, Greedy and random Broyden's methods with explicit superlinear convergence rates in nonlinear equations, preprint, arXiv:2110.08572 (2021)

[122] T. Yoon and E. K. Ryu, Accelerated algorithms for smooth convex-concave minimax problems with $o(1/k^2)$ rate on squared gradient norm. In *International Conference on Machine Learning*, Proc. Mach. Learn. Res., 12098–12109 (2021)

[123] G. Zhang, Y. Wang, L. Lessard and R. B. Grosse, Near-optimal local convergence of alternating gradient descent-ascent for minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, Proc. Mach. Learn. Res., 7659–7679 (2022)

[124] J. Zhang, M. Hong and S. Zhang, On lower iteration complexity bounds for the convex concave saddle point problems. *Math. Program.* **194**, 901–935 (2022)

[125] X. Zhang, N. S. Aybat and M. Gurbuzbalaban, Robust accelerated primal-dual methods for computing saddle points, preprint, arXiv: 2111.12743 (2021)

Aleksandr Beznosikov is a PhD student at the Moscow Institute of Physics and Technology (Moscow, Russia). He is also a researcher at the Laboratory of Mathematical Methods of Optimization and at Laboratory of Advanced Combinatorics and Network Applications in the Moscow Institute of Physics and Technology, a junior researcher at the International Laboratory of SA and HDI in the Higher School of Economics (Moscow), a research intern at Yandex Research (Moscow). His current research interests are concentrated around variational inequalities, saddle point problems, distributed optimization, stochastic optimization, machine learning and federated learning.

anbeznosikov@gmail.com

Boris Polyak (1935–2023) was head of the Ya. Z. Tsypkin Laboratory of the Institute for Control Science of the Russian Academy of Sciences in Moscow and a professor at the Moscow University of Physics and Engineering. He received a PhD degree in mathematics from Moscow State University in 1963 and a Doctor of Science degree (habilitation) in engineering from the Institute for Control Science of the Russian Academy of Sciences in Moscow in 1977. He authored or coauthored more than 250 papers in peer-reviewed journals as well as four monographs, including "Introduction to Optimization". He was an IFAC Fellow, a recipient of the EURO-2012 Gold Medal and the INFORMS Optimization Society Khyachyan Prize. His main area of research was optimization algorithms and optimal control.

Eduard Gorbunov is a researcher at the Laboratory of Mathematical Methods of Optimization in the Moscow Institute of Physics and Technology. His current research interests are concentrated around stochastic optimization and its applications to machine learning, distributed optimization, derivative-free optimization, and variational inequalities.

ed-gorbunov@yandex.ru

Dmitry Kovalev is a PhD student at King Abdullah University of Science and Technology (Thuwal, Saudi Arabia). In 2020 and 2021 he received the Yandex Award (Ilya Segalovich Award). His current research interests include continuous optimization and machine learning.

dakovalev1@gmail.com

Alexander Gasnikov is a professor at the Moscow Institute of Physics and Technology, head of the Laboratory Mathematical Methods of Optimization and head of the department Mathematical Foundations of Control. He received a Doctor of Science degree (habilitation) in mathematics in 2016 from the Faculty of Control and Applied Mathematics of the Moscow Institute of Physics and Technology. In 2019 he received an award from the Yahoo Faculty Research and Engagement Program. In 2020 he received the Yandex Award (Ilya Segalovich Award). In 2021 he received the Award for Young Scientists from the Moscow government. His main area of research is optimization algorithms.

gasnikov@yandex.ru