Shuffling Heuristic in Variational Inequalities: Establishing New Convergence Guarantees

Daniil Medyakov [1, 2] Gleb Molodtsov [1, 2] Grigoriy Evseev [1] Egor Petrov [1] Aleksandr Beznosikov [1, 2, 3]

[1] Moscow Institute of Physics and Technology [2] Institute for System Programming RAS [3] Innopolis University

The First School-Conference on Tensor Methods in Mathematics and Data Science, 2024, Shenzhen, China

Distributed problem

• **Variational Inequality (VI)**

$$\text{Find } z^* \in Z : \forall z \in Z \hookrightarrow \langle F(z^*), z - z^* \rangle + g(z) - g(z^*) \geqslant 0, \quad (1)$$

where F is a monotone operator and g is a proper convex lower semicontinuous function, which plays the role of regularizer.

• Training data describing F is **distributed** across n devices: $F(z) = \frac{1}{n}\sum_{i=1}^{n} F_i(z)$, where each F_i corresponds to an individual data point.

Example (Convex optimization)

$$\min_{z \in \mathbb{R}^d} [f(z) + g(z)]. \quad (2)$$

In this example, f is a smooth data representative term, and g is probably a non-smooth regularizer. In this setting, we define $F(z) = \nabla f(z)$. Then $z^* \in \operatorname{dom} g$ is the solution of (1) if and only if $z^* \in \operatorname{dom} g$ is the solution of (2). In this way, the problem (2) can be considered as a variational inequality.

Example (Convex-concave saddles)

We consider the following convex-concave saddle point problem:

$$\min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} [f(x, y) + g_1(x) - g_2(y)]. \quad (3)$$

There, f has the same interpretation as in Example 1, and g_1, g_2 can also be perceived as regularizers. In this setting, we define $F(z) = F(x, y) = [\nabla_x f(x, y), -\nabla_y f(x, y)]$. Then $z^* \in \operatorname{dom} g_1 \times \operatorname{dom} g_2$ is the solution of (1) if and only if $z^* \in \operatorname{dom} g_1 \times \operatorname{dom} g_2$ is the solution of (3). In this way, the problem (3) can be considered as a variational inequality.

Variance Reduction methods

We explore stochastic algorithms which are particularly suitable for practical extensive applications. The stochastic version of the Extragradient method select random independent indexes i_t, j_t at iteration t and performs the following updates:

$$z^{t+\frac{1}{2}} = z^t - \gamma F_{i_t}(z^t),$$
$$z^{t+1} = z^t - \gamma F_{j_t}(z^{t+\frac{1}{2}}).$$

The **variance reduction (VR)** technique was developed for a classical finite-sum minimization task. Considering convex optimization problem (see Example 1), we can formally write the stochastic reduced gradient at the point $z^{t+\frac{1}{2}}$ as

$$\nabla \hat{f}_{i_t}(z^{t+\frac{1}{2}}) = \nabla f_{i_t}(z^{t+\frac{1}{2}}) - \nabla f_{i_t}(\omega^t) + \nabla f(\omega^t).$$

Setup

Assumption 1: Each operator F_i is L-Lipschitz, i.e., it satisfies

$$\|F_i(z_1) - F_i(z_2)\| \leq L\|z_1 - z_2\|$$

for any $z_1, z_2 \in Z$.

Assumption 2: Each operator F_i is μ-strongly monotone, i.e., it satisfies

$$\langle F_i(z_1) - F_i(z_2), z_1 - z_2 \rangle \geqslant \mu\|z_1 - z_2\|^2$$

for any $z_1, z_2 \in Z$.

Assumption 3: Each stochastic operator F_i and full operator F is bounded at the point of the solution $z^* \in \operatorname{dom} g$, i.e.,

$$\mathbb{E}\|F_i(z^*)\|^2 \leqslant \sigma_*^2, \|F(z^*)\|^2 \leqslant \sigma_*^2.$$

Proximal Algorithm

We often encounter the need to minimize the function decomposed into two parts: a smooth differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ and a possibly non-smooth function $g : \mathbb{R}^n \to \mathbb{R}$ which is proximal friendly. To solve it, we can utilize the proximal gradient method:

$$\operatorname{prox}_g(z) = \arg\min_{y \in \mathbb{R}^n} \left\{ g(y) + \frac{1}{2}\|y - z\|^2 \right\}.$$

The update step for solving the problem can be written as

$$z^{t+1} = \operatorname{prox}_{\alpha_t g}\left(z^t - \alpha_t \nabla f(z^t)\right).$$

Table: Comparison of the convergence results for the methods for solving VI.

Algorithm	Sampling	VR?	Strongly Monotone Complexity	Monotone Complexity
Extragradient (Korpelevich, 1976; Mokhtari, 2020)	D	✗	$\tilde{\mathcal{O}}\left(\frac{nL}{\mu}\right)$	$\mathcal{O}\left(\frac{nL}{\varepsilon}\right)$
Mirror-prox (Nemirovski, 2004)	D	✗	\	$\mathcal{O}\left(\frac{nL}{\varepsilon}\right)$
FBF (Tseng, 2000)	D	✗	\	$\mathcal{O}\left(\frac{nL}{\varepsilon}\right)$
FoRB (Malitsky, 2020)	D	✗	\	$\mathcal{O}\left(\frac{nL}{\varepsilon}\right)$
Mirror-prox (Juditsky, 2011)	I	✗	\	$\mathcal{O}\left(\frac{L}{\varepsilon} + \frac{1}{\varepsilon^2}\right)$
Extragradient (Beznosikov, 2020)	I	✗	$\tilde{\mathcal{O}}\left(\frac{L}{\mu} + \frac{1}{\mu^2\varepsilon}\right)$	$\mathcal{O}\left(\frac{L}{\varepsilon} + \frac{1}{\varepsilon^2}\right)$
REG (Mishchenko, 2020)	I	✗	$\tilde{\mathcal{O}}\left(\frac{L}{\mu} + \frac{1}{\mu^2\varepsilon}\right)$	$\mathcal{O}\left(\frac{L}{\varepsilon} + \frac{1}{\varepsilon^2}\right)$
Extragradient (Carmon, 2019)	I	✓	\	$\tilde{\mathcal{O}}\left(n + \frac{n\bar{L}}{\varepsilon}\right)$
Mirror-prox (Carmon, 2019)	I	✓	\	$\tilde{\mathcal{O}}\left(n + \frac{n\bar{L}}{\varepsilon}\right)$
FBF (Palaniappan, 2016)	I	✓	$\tilde{\mathcal{O}}\left(n + \frac{n\bar{L}}{\mu}\right)$	$\tilde{\mathcal{O}}\left(n + \frac{n\bar{L}}{\varepsilon}\right)^{(1)}$
Extragradient (Chavdarova, 2019)	I	✓	$\tilde{\mathcal{O}}\left(n + \frac{\bar{L}^2}{\mu^2}\right)$	$\tilde{\mathcal{O}}\left(n + \frac{\bar{L}^2}{\varepsilon^2}\right)^{(1)}$
FoRB (Alacaoglu, 2021)	I	✓	\	$\mathcal{O}\left(n + \frac{n\bar{L}}{\varepsilon}\right)$
Extragradient (Alacaoglu, 2022)	I	✓	\	$\mathcal{O}\left(n + \frac{n\bar{L}}{\varepsilon}\right)$
Mirror-prox (Alacaoglu, 2022)	I	✓	\	$\mathcal{O}\left(n + \frac{n\bar{L}}{\varepsilon}\right)$
Extragradient (this paper)	RR / SO	✗	$\tilde{\mathcal{O}}\left(n + \frac{L}{\mu} + \frac{n^2}{\mu^2\varepsilon}\right)$	$\tilde{\mathcal{O}}\left(n + \frac{L}{\varepsilon} + \frac{n^2}{\varepsilon^3}\right)^{(1)}$
Extragradient (this paper)	RR / SO	✓	$\tilde{\mathcal{O}}\left(n\frac{L^2}{\mu^2}\right)$	$\tilde{\mathcal{O}}\left(n\frac{L^2}{\varepsilon^2}\right)^{(1)}$

Columns: Sampling = D if considered deterministic method, I if method uses independent choice of operator's indexes, RR / SO if method uses shuffling heuristic, Assumption = assumption on operator F, VR? = whether the method uses variance reduction technique.

Notation: μ = constant of strong monotonicity, L = Lipschitz constant of F, $\overline{L}$ = Lipschitz in mean constant, i.e. $1/n\sum_{i=1}^{n}\|F_i(z_1) - F_i(z_2)\| \leqslant L\|z_1 - z_2\|\ \forall z_1, z_2 \in Z$, n = size of the dataset, ε = accuracy of the solution.
(1): This result is obtained with regularization trick: $\mu \sim \varepsilon/D^2$.

Main Contributions

• **Novel approach to proof.** We present a technique that allows us to "return" to the starting point of an epoch in which there is a property of unbiasedness.

• **Convergence estimates.** We provide the first theoretical convergence rates for shuffling methods applied to the finite-sum variational inequality problem considering both Extragradient (our linear term coincides with that for the method without shuffling) and Extragradient with VR (the first to obtain a linear convergence estimate for methods with shuffling in VIs).

• **Experiments.** Our experiments emphasize the superiority of shuffling over the random index selection heuristic. We also consider two classical practical applications: adversarial training and image denoising.

Algorithms and convergence analysis

The **unbiasedness** of stochastic operators complicates the analysis. However, the equality holds at two points: z_s^0, the first point of the epoch, and z^*. Thus, we can leverage the unbiased operators by **"going back"** to the start of the epoch. This approach is also useful for methods involving Markov chains, where the unbiased property only holds at the chain's correlation point.

Extragradient

Algorithm 1. RR Extragradient

Input: Starting point $z_0^0 \in \mathbb{R}^d$
Parameter: Stepsize γ
for $s = 0, 1, 2, \ldots, S-1$ **do**
  Generate a permutation $\pi_0, \pi_1, \ldots, \pi_{n-1}$ of sequence $\{1, 2, \ldots, n\}$
  for $t = 0, 1, 2, \ldots, n-1$ **do**
    $z_s^{t+\frac{1}{2}} = \operatorname{prox}_{\gamma g}\left(z_s^t - \gamma F_{\pi_t^s}(z_s^t)\right)$
    $z_s^{t+1} = \operatorname{prox}_{\gamma g}\left(z_s^t - \gamma F_{\pi_s^t}(z_s^{t+\frac{1}{2}})\right)$
  end for
  $z_s^n = z_{s+1}^0$
end for
Output: z_S^n

Algorithm 2. SO Extragradient

Input: Starting point $z_0^0 \in \mathbb{R}^d$
Parameter: Stepsize γ
Generate a permutation $\pi_0, \pi_1, \ldots, \pi_{n-1}$ of sequence $\{1, 2, \ldots, n\}$
for $s = 0, 1, 2, \ldots, S-1$ **do**
  for $t = 0, 1, 2, \ldots, n-1$ **do**
    $z_s^{t+\frac{1}{2}} = \operatorname{prox}_{\gamma g}\left(z_s^t - \gamma F_{\pi_s^t}(z_s^t)\right)$
    $z_s^{t+1} = \operatorname{prox}_{\gamma g}\left(z_s^t - \gamma F_{\pi_s^t}(z_s^{t+\frac{1}{2}})\right)$
  end for
  $z_s^n = z_{s+1}^0$
end for
Output: z_S^n

Theorem 1

Suppose Assumptions 1, 2, 3 hold. Then for Algorithms 1, 2 with $\gamma \leqslant \min\left\{\frac{1}{2\mu n}, \frac{1}{6L}\right\}$ after S epochs,

$$\|z_S^n - z^*\|^2 \leqslant (1 - \frac{\gamma\mu}{2})^{Sn}\|z_0^0 - z^*\|^2 + \frac{256\gamma n^2\sigma_*^2}{\mu}.$$

Remark 1

To transform the obtained estimation for the case of monotone stochastic operators, we use a regularization trick with $\mu \sim \frac{\varepsilon}{D}$. Thus, we obtain $\tilde{\mathcal{O}}\left(n + \frac{L}{\varepsilon} + \frac{n^2}{\varepsilon^3}\right)$ iteration and oracle complexity. This is convergence in argument, it differs from the classical form.

Our result is a great achievement in the shuffling theory, since despite the deterioration on n in the sublinear term, the estimation on the **linear term coincides** with that in the classical setting with independent choice of stochastic operators.

Extragradient with Variance Reduction

Previously, authors used a more classical version and compute $F(\omega_s^t)$ at the beginning of each epoch. We consider another option and compute this full operator randomly with probability p. We put $p = \frac{1}{n}$ not to increase the oracle complexity and obtain that on average we also update the full operator once per epoch.

Theorem 2

Suppose that Assumptions 1, 2 hold. Then for Algorithm 3 with $\gamma \leqslant \frac{(1-\alpha)\mu}{6L^2}, p = \frac{1}{n}$ and $V_s^t = \mathbb{E}\|z_s^t - z^*\|^2 + \mathbb{E}\|\omega_s^t - z^*\|^2$ after T iterations,

$$V_S^n \leqslant \left(1 - \frac{\gamma\mu}{4}\right)^T V_0^0.$$

Remark 2

Similarly to Remark 1, we can use our result in the monotone case by the regularization trick and obtain $\tilde{\mathcal{O}}\left(n\frac{L^2}{\varepsilon^2}\right)$.

We remove the variance that arose in Theorem 1 and **obtain linear convergence**. However, according to current theory, methods with the shuffling heuristic are inferior to those with independent sampling for variance reduction methods.

Experiments (Adversarial Training)

We address an adversarial training problem. Let us formulate it in the following way:

$$\min_{w \in \mathbb{R}^d} \max_{\|r_i\| \leqslant D} \left[\frac{1}{2N}\sum_{i=1}^{N}\left(w^T(x_i + r_i) - y_i\right)^2 + \frac{\lambda}{2}\|w\|^2 - \frac{\beta}{2}\|r\|^2\right], \quad (4)$$

where the samples corresponds to features x_i and targets y_i. The results are presented in Figure 1.

Algorithm 3. RR/SO Extragradient with variance reduction

Input: Parameters: z_0^0, ω_0^0
Parameter: Stepsize γ, $\alpha \in (0, 1)$
Generate a permutation $\pi_0, \pi_1, \ldots, \pi_{n-1}$ of sequence $\{1, 2, \ldots, n\}$ // SO heuristic
for $s = 0, 1, \ldots$ **do**
  Generate a permutation $\pi_0, \pi_1, \ldots, \pi_{n-1}$ of sequence $\{1, 2, \ldots, n\}$ // RR heuristic
  for $t = 0, 1, \ldots, n-1$ **do**
    $\overline{z}_s^t = \alpha z_s^t + (1 - \alpha)\omega_s^t$
    $z_s^{t+1/2} = \operatorname{prox}_{\gamma g}\left(\overline{z}_s^t - \gamma F(\omega_s^t)\right)$
    $\hat{F}(z_s^{t+1/2}) = F_{\pi_s^t}(z_s^{t+1/2}) - F_{\pi_s^t}(\omega_s^t) + F(\omega_s^t)$
    $z_s^{t+1} = \operatorname{prox}_{\gamma g}\left(\overline{z}_s^t - \gamma\hat{F}\left(z_s^{t+1/2}\right)\right)$
    $\omega_s^{t+1} = \begin{cases} z_s^t, & \text{with probability } p \\ \omega_s^t, & \text{with probability } 1 - p \end{cases}$
  end for
  $z_{s+1}^0 = z_s^n$
  $\omega_{s+1}^0 = \omega_s^n$
end for
Output: z_S^n

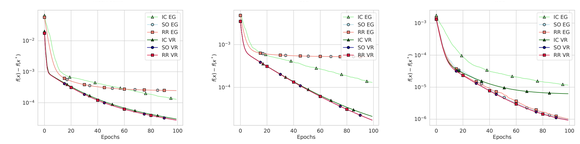

Figure: Extragradient with and without VR compared using various shuffling heuristics on mushrooms, a9a and w8a datasets on the problem (4).

Experiments (Image Denoising)

We consider the classic saddle point problem as in Example 2:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} [\langle Kx, y\rangle + G_1(x) - G_2(y)],$$

where G_1 and G_2 are proper convex lower semicontinuous regularizers, and K is a continuous linear operator. Let g be a given noisy image and u – a solution we seek. Thus, for the image denoising,

$$\min_{u \in \mathcal{X}} \max_{p \in \mathcal{Y}} \left[\langle \nabla u, p\rangle_{\mathcal{Y}} + \lambda/2\|u - g\|_2^2 - \delta_P(p)\right]$$

is the considered problem with p being a dual variable and $\delta_P(p)$ – the indicator function of the set $P = \{p \in \mathcal{Y} : \|p(x)\| \leq 1\}$. Using duality, we can write the final formulation of considering problem as

$$\min_{u \in \mathcal{X}} \max_{p \in \mathcal{Y}} \left[-\langle u, \operatorname{div} p\rangle_{\mathcal{X}} + \lambda/2\|u - g\|_2^2 - \delta_P(p)\right]. \quad (5)$$

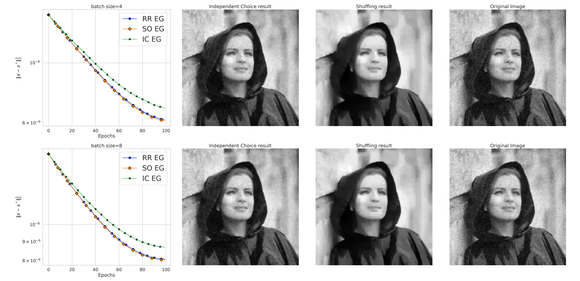

Figure: Extragradient on image with $\sigma = 0.05$ on the problem (5).

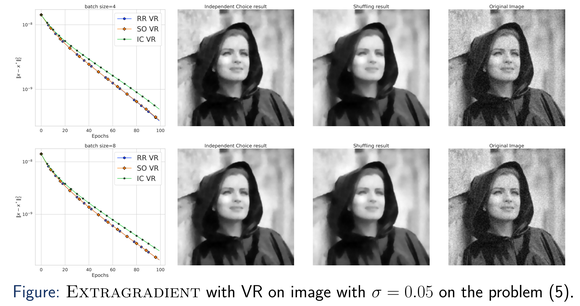

Figure: Extragradient with VR on image with $\sigma = 0.05$ on the problem (5).