Zeroth-Order Algorithms for Variational Inequalities: Theoretical Analysis and Practical Application in Machine Learning and Neural Networks

A. Sadiev, A. Beznosikov, P. Severilov, P. Dvurechensky, A. Gasnikov

Moscow Institute of Physics and Technology, Russia, Sirius University of Science and Technology, Russia

Sirius University of Science and Technology, Sochi, 2020
For a stochastic smooth monotone, propose a method that uses an oracle of zero order, that is, there is access only to the value of the function at a point.

Compare the proposed methods with each other.

Apply the proposed methods for training GANs.
Problem Setup

Variational inequalities

- find \( z^* \in \mathcal{Z} \) such that \( \langle F(z^*), z - z^* \rangle \geq 0 \ \forall z \in \mathcal{Z} \)
- \( \mathcal{Z} \) is a closed convex set.
- \( F : \mathbb{R}^n \rightarrow \mathbb{R}^n \) is an operator

Monotone Operators

Assumption 1(m). The operator \( F \) is monotone, i.e.

\[
\langle F(z_1) - F(z_2), z_1 - z_2 \rangle \geq 0, \quad \forall z_1, z_2 \in \mathcal{Z}.
\]

Assumption 1(sm). The operator \( F \) is \( \mu \)-strongly monotone w.r.t \( V.(\cdot) \), i.e.

\[
\langle F(z_1) - F(z_2), z_1 - z_2 \rangle \geq \frac{\mu}{2} (V_{z_1}(z_2) + V_{z_2}(z_1)), \quad \forall z_1, z_2 \in \mathcal{Z}
\]
Assumptions

Assumption 2(c). The operator $F$ is $L$-Lipschitz continuous w.r.t $\| \cdot \|_2$, i.e.

$$\| F(z_1, \xi) - F(z_2, \xi) \|_2 \leq L(\xi) \| z_1 - z_2 \|_2, \quad \mathbb{E}[L^2(\xi)] = L^2_2, \quad \forall z_1, z_2 \in \mathcal{Z}.$$  

Assumption 2(fc). The operator $F$ is $L$-firmly Lipschitz continuous w.r.t $\| \cdot \|_2$, i.e.

$$\| F(z_1, \xi) - F(z_2, \xi) \|_2^2 \leq L(\xi) \langle F(z_1) - F(z_2), z_1 - z_2 \rangle,$$

$$\mathbb{E}[L^2(\xi)] = L^2_2, \quad \forall z_1, z_2 \in \mathcal{Z}$$
Inexact Zero Order Stochastic Oracles

Projection Oracle

\[ G(z, e, \tau) = n\langle F(z), e \rangle e + \xi(z) + \delta(z), \]

\[ \mathbb{E}[\xi(z)] = 0, \quad \mathbb{E}[\|\xi(z)\|^2] \leq \sigma^2, \quad \|\delta(z)\|_2 \leq \Delta, \]

where random variable \( \xi(z) \) is responsible for unbiased stochastic noise and \( \delta(z) \) – for deterministic noise, the vector \( e \) is generated uniformly on the unit Euclidean sphere \( \mathcal{R}S^2(1) \).
Inexact Zero Order Stochastic Oracles

Random Direction Oracle

\[ g_d(z, e, \tau, \xi) = \frac{n}{\tau} (f(z + \tau e, \xi) + \delta(z + \tau e) - f(z, \xi) - \delta(z)) e, \]

\[ \mathbb{E}[F(z, \xi)] = F(z), \quad \mathbb{E}[\|F(z, \xi) - F(z)\|_2^2] \leq \sigma^2, \quad |\delta(z)| \leq \Delta, \]

Full Coordinates Oracle

\[ g_f(z, \tau, \xi) = \frac{1}{\tau} \sum_{i=1}^{n} (f(z + \tau h_i, \xi) + \delta(z + \tau h_i) - f(z, \xi) - \delta(z)) h_i, \]

where \( \{h_1, \ldots, h_n\} \) is a standard orthogonal normalized basis.
Function $d(z) : \mathcal{Z} \rightarrow \mathbb{R}$ is called prox-function if $d(z)$ is 1-strongly convex w.r.t. $\| \cdot \|$-norm and differentiable on $\mathcal{Z}$ function.

Bregman divergence $V_z(w)$ associated with $d(z)$:

$$V_z(w) = d(z) - d(w) - \langle \nabla d(w), z - w \rangle.$$ 

Prox-operator: $\text{prox}_x(\xi) = \arg \min_{y \in \mathcal{Z}} (V_x(y) + \langle \xi, y \rangle)$

the Bregman-diameter $D_p$ of set $\mathcal{Z}$ w.r.t. $V_{z_1}(z_2)$:

$$D_p = \max \{ \sqrt{2V_{z_1}(z_2)} : z_1, z_2 \in \mathcal{Z} \}$$
Algorithm zoVIA

Algorithm 1 zoVIA

**Input:** $z_0$, $N$, $\gamma$, $\tau$.
Choose oracle grad from $G$, $g_d$, $g_f$.

for $k = 0, 1, 2, \ldots, N$ do
  Sample indep. $e_k$, $\xi_k$.
  $d_k = \text{grad}(z_k, e_k, \tau, \xi_k)$.
  $z_{k+1} = \text{prox}_{z_k}(\gamma \cdot d_k)$.
end for

**Output:** $z_{N+1}$ or $\bar{z}_{N+1}$.

where

$$
\bar{z}_{N+1} = \frac{1}{N+1} \left( \sum_{k=0}^{N} z_k \right).
$$
Theorem 1

For Algorithm 1 with Random direction oracle under Assumptions 1(sm), 2(c) and with \( \gamma = \frac{\mu}{96n^2/q \rho_n L^2} \),

\[
\tau = \mathcal{O} \left( \min \left\{ \frac{\varepsilon \mu^2 N}{n^{2/q+1} \rho_n L^3 D_p}, \sqrt{\frac{\varepsilon \mu N}{n^{2/q+1} \rho_n L^2}} \right\} \right),
\]

\[
\Delta = \mathcal{O} \left( \min \left\{ L \left( \frac{\varepsilon \mu^2 N}{n^{2/q+1} \rho_n L^3 D_p} \right)^2, \frac{\varepsilon \mu N}{n^{2/q+1} \rho_n L^2} \right\} \right),
\]

then the number of iterations (coincides with oracle complexity) to find \( \varepsilon \)-solution

\[
N = \tilde{\mathcal{O}} \left( \max \left\{ \frac{n^{2/q} \rho_n L^2}{\mu^2} \log \left( \frac{1}{\varepsilon} \right), \frac{n^{2/q} \rho_n \sigma^2}{\mu \varepsilon} \right\} \right).
\]
Algorithm zoESVIA

Algorithm 2 zoESVIA

**Input:** $z_0$, $N$, $\gamma$, $\tau$.
Choose oracle grad from $G, g_d, g_f$.

for $k = 0, 1, 2, \ldots, N$ do

Sample indep. $e_k$, $e_{k+1/2}$, $\xi_k$, $\xi_{k+1/2}$.

$d_k = \text{grad}(z_k, e_k, \tau, \xi_k)$.

$z_{k+1/2} = \text{prox}_{z_k}(\gamma \cdot d_k)$.

$d_{k+1/2} = \text{grad}(z_{k+1/2}, e_{k+1/2}, \tau, \xi_{k+1/2})$.

$z_{k+1} = \text{prox}_{z_k}(\gamma \cdot d_{k+1/2})$.

end for

**Output:** $z_{N+1}$ or $\bar{z}_{N+1}$.

where

$$\bar{z}_{N+1} = \frac{1}{N + 1} \left( \sum_{k=0}^{N} z_{k+1/2} \right).$$
Convergence analysis

**Theorem 2**

Let $\varepsilon$ – accuracy of the solution. For Algorithm 2 with Full coordinates oracle under Assumptions 1, 2, 3(m), 4 and with $\gamma = \min \left\{ \frac{1}{2L}, \frac{D_p}{\sigma \sqrt{N}} \right\}$ and additionally if

$$
\tau = \mathcal{O} \left( \min \left\{ \frac{\varepsilon}{\sqrt{nL_2D_p}}, \sqrt{\frac{\varepsilon L \sqrt{N}}{nL_2^2}} \right\} \right),
$$

$$
\Delta = \mathcal{O} \left( \min \left\{ \frac{\varepsilon^2}{nL_2D_p^2}, \frac{\varepsilon L \sqrt{N}}{nL_2} \right\} \right),
$$

then the oracle complexity to find $\varepsilon$-solution

$$
N = \mathcal{O} \left( \max \left\{ \frac{LD_p^2}{\varepsilon}, \frac{\sigma^2 D_p^2}{\varepsilon^2} \right\} \right).
$$
Algorithm 3 zoscESVIA

Input: $z_0$, $N$, $\gamma$, $\tau$.
Choose oracle grad from $G, g_d, g_f$.
for $k = 0, 1, 2, \ldots, N$ do
    Sample independent $e_k$, $\xi_k$.
    Take $d_{k-1}$ from previous step.
    $z_{k+1/2} = \text{prox}_{z_k}(\gamma \cdot d_{k-1})$.
    $d_k = \text{grad}(z_{k+1/2}, e_{k+1/2}, \tau, \xi_k)$.
    $z_{k+1} = \text{prox}_{z_k}(\gamma \cdot d_k)$.
end for
Output: $z_{N+1}$ or $\bar{z}_{N+1}$.

where

$$\bar{z}_{N+1} = \frac{1}{N+1} \left( \sum_{k=0}^{N} z_{k+1/2} \right).$$
Theorem 3

For Algorithm 3 for Full coordinates oracle under Assumptions 1, 2, 3(sm), 4 and with $p = 2$ and $V_x(y) = \frac{1}{2} \|x - y\|_2^2$, $\gamma = \frac{1}{6L}$ and additionally if

\[ \tau = \mathcal{O} \left( \min \left\{ \frac{\epsilon \mu N}{\sqrt{nL^2D_2}}, \sqrt{\frac{\epsilon \mu N}{nL_2^2}} \right\} \right), \]

\[ \Delta = \mathcal{O} \left( \min \left\{ \frac{\epsilon^2 \mu^2 N^2}{nL^2D_2^2}, \frac{\epsilon \mu N}{nL_2^2} \right\} \right), \]

then the oracle complexity to find $\epsilon$-solution

\[ N = \tilde{\mathcal{O}} \left( \max \left\{ \frac{nL}{\mu} \log \left( \frac{1}{\epsilon} \right), \frac{n\sigma^2}{\mu \epsilon} \right\} \right). \]
Algorithm zoESVIA (same direction)

Algorithm 4 zoESVIA (same direction)

Input: $z_0, N, \gamma, \tau$.
Choose oracle grad from $G, g_d, g_f$.

for $k = 0, 1, 2, \ldots, N$ do
   Sample indep. $e_k, \xi_k$.
   $d_k = \text{grad}(z_k, e_k, \tau, \xi_k)$.
   $z_{k+1/2} = \text{prox}_{z_k}(\gamma \cdot d_k)$.
   $d_{k+1/2} = \text{grad}(z_{k+1/2}, e_k, \tau, \xi_k)$.
   $z_{k+1} = \text{prox}_{z_k}(\gamma \cdot d_{k+1/2})$.
end for

Output: $z_{N+1}$ or $\bar{z}_{N+1}$.  

Convergence analysis

Theorem 4

Let $\varepsilon$ – accuracy of the solution. For Algorithm 4 under Assumptions 1(m), 2(c) with $\gamma = \min \{1/2nL_2, D_2/\sigma\sqrt{nN}\}$ and additionally, if $F(z^*) = 0$ and

$$
\tau = \mathcal{O} \left( \min \left\{ \frac{\varepsilon}{nLD_2}, \sqrt{\frac{\varepsilon L_2 \sqrt{N}}{nL}} \right\} \right),
$$

$$
\Delta = \mathcal{O} \left( \min \left\{ \frac{\varepsilon^2}{n^2 LD_2^2}, \frac{\varepsilon L_2 \sqrt{N}}{nL} \right\} \right),
$$

then the number of iterations (coincides with oracle complexity) to find $\varepsilon$-solution

$$
N = \widetilde{\mathcal{O}} \left( \max \left\{ \frac{nL_2 D_2^2}{\varepsilon}, \frac{n\sigma^2 D_2^2}{\varepsilon^2} \right\} \right).
$$
## Convergence analysis

<table>
<thead>
<tr>
<th>Method</th>
<th>Assumptions</th>
<th>Complexity in deterministic setup</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZO-GDMSA [5]</td>
<td>NC-SC, UCst-Cst, S</td>
<td>$\tilde{\mathcal{O}}\left(\frac{\kappa^2 n}{\varepsilon^2}\right)$</td>
</tr>
<tr>
<td>ZO-Min-Max [4]</td>
<td>NC-SC, Cst-Cst, S</td>
<td>$\tilde{\mathcal{O}}\left(\frac{n}{\varepsilon^6}\right)$</td>
</tr>
<tr>
<td>zoSPA [1]</td>
<td>C-C, Cst-Cst, BG</td>
<td>$\mathcal{O}\left(\frac{M^2D^2}{\varepsilon^2}n^{2/q}\right)$</td>
</tr>
<tr>
<td>[this paper]</td>
<td>SC-SC, Cst-Cst, S</td>
<td>$\tilde{\mathcal{O}}\left(\min\left[\kappa^2 n^{2/q}, \kappa n\right] \cdot \log\left(\frac{1}{\varepsilon}\right)\right)$</td>
</tr>
<tr>
<td>[this paper]</td>
<td>C-C, Cst-Cst, S</td>
<td>$\tilde{\mathcal{O}}\left(\frac{nLD^2}{\varepsilon}\right)$</td>
</tr>
<tr>
<td>[this paper]</td>
<td>C-C, Cst-Cst, FS</td>
<td>$\tilde{\mathcal{O}}\left(\frac{n^{2/q}L^2D^2}{\varepsilon}\right)$</td>
</tr>
</tbody>
</table>
### Convergence analysis

<table>
<thead>
<tr>
<th>Method</th>
<th>Order</th>
<th>Assumptions</th>
<th>Complexity for stochastic part</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGMP [3]</td>
<td>1st</td>
<td>C-C, Cst-Cst, S</td>
<td>$\mathcal{O} \left( \frac{\sigma^2 D^2}{\varepsilon^2} \right)$</td>
</tr>
<tr>
<td>PEG [2]</td>
<td>1st</td>
<td>SC-SC, Cst-Cst, S</td>
<td>$\mathcal{O} \left( \frac{\sigma^2}{\mu^2 \varepsilon} \right)$</td>
</tr>
<tr>
<td>ZO-SGDMSA[5]</td>
<td>0th</td>
<td>NC-SC, UCst-Cst, S</td>
<td>$\tilde{\mathcal{O}} \left( \frac{\kappa^2 n \sigma^2}{\epsilon^4} \right)$</td>
</tr>
<tr>
<td>[this paper]</td>
<td>0th</td>
<td>SC-SC, Cst-Cst, S</td>
<td>$\mathcal{O} \left( \frac{n^2/q \sigma^2}{\mu^2 \varepsilon} \right)$</td>
</tr>
<tr>
<td>[this paper]</td>
<td>0th</td>
<td>C-C, Cst-Cst, S</td>
<td>$\mathcal{O} \left( \frac{n^2/q \sigma^2 D^2}{\varepsilon^2} \right)$</td>
</tr>
<tr>
<td>[this paper]</td>
<td>0th</td>
<td>C-C, Cst-Cst, FS</td>
<td>$\mathcal{O} \left( \frac{n^2/q \sigma^2 D^2}{\varepsilon^2} \right)$</td>
</tr>
</tbody>
</table>
We consider the classical saddle-point problem on a probability simplex:

$$\min_{x \in \Delta_n} \max_{y \in \Delta_k} [y^T Cx],$$

where $\Delta_n = \{w \in \mathbb{R}^n : \forall i \rightarrow w_i \geq 0, \sum_{i=1}^n w_i = 1\}$ - probability simplex.

**Proximal setup**

- Prox-function is $d(x) = \sum_{i=1}^n x_i \log x_i$ (entropy)
- Bregman divergence is $V_x(y) = \sum_{i=1}^n x_i \log \frac{x_i}{y_i}$ (KL divergence)
Algorithms zoVIA, zoESVIA, zoscESVIA, zoESVIA(same $e$) with different oracles are applied to solve saddle point problem (Matrix Game).


Wang, Z., Balasubramanian, K., Ma, S., Razaviyayn, M.: Zeroth-order algorithms for nonconvex minimax problems with improved complexities (2020)
Thank you for your attention!