Extragradient Method: *O(1/K)* Last-Iterate Convergence for Monotone Variational Inequalities and Connections With Cocoercivity

Eduard Gorbunov[1,2], Nicolas Loizou[3], Gauthier Gidel[2,4]

[1]Moscow Institute of Physics and Technology (Russia), [2]Mila & Université de Montréal (Canada), [3]Johns Hopkins University (USA), [4]Canada CIFAR AI Chair

1. Preliminaries

Problem: unconstrained variational inequality problem (VIP)

$$\text{find } x^* \in \mathbb{R}^d \quad \text{such that} \quad F(x^*) = 0$$

Examples:

- Min-max problems $\qquad \min_{u\in\mathbb{R}^{d_1}} \max_{v\in\mathbb{R}^{d_2}} f(u,v)$
- Minimization problems $\qquad \min_{x\in\mathbb{R}^d} f(x)$

Assumptions:

- Lipschitzness $\quad \|F(x) - F(x')\| \leq L\|x - x'\| \quad \forall x, x' \in \mathbb{R}^d$
- Monotonicity $\quad \langle F(x) - F(x'), x - x'\rangle \geq 0 \quad \forall x, x' \in \mathbb{R}^d$

Measures of convergence:

- $\texttt{Gap}_F(x^K) = \max_{y\in\mathbb{R}^d: \|y - x^*\| \leq \|x^0 - x^*\|} \langle F(y), x^K - y\rangle$
  - ✓ Natural extension of optimization error for VIPs
  - ✗ Hard to estimate in practice and to generalize to non-monotone case
- $\|F(x^K)\|^2$
  - ✗ Provides weaker guarantees than the Gap-function
  - ✓ Easier to compute than the Gap-function

2. Extragradient Method

Extragradient method (EG) is one of the most popular methods for VIPs

$$x^{k+1} = x^k - \gamma F\left(x^k - \gamma F\left(x^k\right)\right)$$

Known convergence results

- Averaged- and best-iterate rates

$$\texttt{Gap}_F\left(\overline{x}^K\right) = \mathcal{O}(1/K)$$

$$\min_{k=0,1,\ldots,K} \left\|F\left(x^k\right)\right\|^2 = \mathcal{O}(1/K)$$

- Last-iterate rates [1]

$$\texttt{Gap}_F\left(x^K\right) = \mathcal{O}\left(1/\sqrt{K}\right)$$

$$\left\|F\left(x^K\right)\right\|^2 = \mathcal{O}(1/K)$$

Lipschitz continuity of the Jacobian is assumed

3. Our Contributions

- ***O(1/K)* last-iterate convergence rate** for EG in terms of the squared norm of the operator for monotone and Lipschitz VIPs
  - **No additional assumptions** are used
  - The proof is *obtained via computer*
- New connections for several known methods with cocoercivity when the original operator is monotone and Lipschitz
  - **Fundamental mathematical differences between EG and Optimistic Gradient method** that are usualy considered as approximations of Proximal Point method

4. Last-Iterate Convergence of EG

Key Lemma: if $0 < \gamma \leq \frac{1}{\sqrt{2}L}$ then for all $k \geq 0$

$$\left\|F\left(x^{k+1}\right)\right\| \leq \left\|F\left(x^k\right)\right\|$$

- This result was unknown even under Lipschitzness of the Jacobian

Using this lemma in the standard analysis of EG, we derive the new result

Main Theorem: if $0 < \gamma \leq \frac{1}{\sqrt{2}L}$ then for all $k \geq 0$

$$\left\|F\left(x^K\right)\right\|^2 \leq \frac{\left\|x^0 - x^*\right\|^2}{\gamma^2\left(1 - L^2\gamma^2\right)(K+1)}$$

- The above theorem implies $\texttt{Gap}_F\left(x^K\right) = \mathcal{O}\left(1/\sqrt{K}\right)$
- Rates match the lower bounds for EG from [1]

The proof is guided by numerical results

$$\begin{aligned} &\max \left\|F\left(x^1\right)\right\|^2 - \left\|F\left(x^0\right)\right\|^2 \\ &\text{s.t. } F \text{ is monotone and } L\text{-Lipschitz, } x^0 \in \mathbb{R}^d, \\ &\qquad \left\|x^0 - x^*\right\|^2 \leq 1 \\ &\qquad x^1 = x^0 - \gamma F\left(x^0 - \gamma F\left(x^0\right)\right) \end{aligned}$$

→ SDP problem (easy to solve)

→ Dual solution

→ **The proof**: sum up the inequalities corresponding to the constraints with weights being equal to the dual solution and rearrange the terms

5. Connections with Cocoercivity

Cocoercivity: $\quad \|F(x) - F(x')\|^2 \leq \ell\langle F(x) - F(x'), x - x'\rangle$

Gradient Descent (GD) $\; x^{k+1} = x^k - \gamma F(x^k)$ has a simple proof of *O(1/K)* last-iterate convergence rate when operator *F* is cocoercive.

Idea: consider EG as GD with a special operator

$$x^{k+1} = x^k - \gamma_2 F_{\mathrm{EG},\gamma_1}(x^k), \qquad F_{\mathrm{EG},\gamma_1}(x) = F(x - \gamma_1 F(x))$$

If we manage to prove that EG-operator is cocoercive, then *O(1/K)* last-iterate convergence rate will easily follow from the corresponding result for GD!

Good news:

- When *F* is affine, EG-operator is cocoercive
- EG-operator is star-cocoercive $\quad \|F_{\mathrm{EG},\gamma_1}(x)\|^2 \leq \frac{2}{\gamma_1}\langle F_{\mathrm{EG},\gamma_1}(x), x - x^*\rangle$
- Operator corresponding to the update of Proximal Point method

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

is cocoercive

Bad news: EG-operator can be non-cocoercive even if *F* is cocoercive! The proof is based on the following fact from [2]:

$$F \text{ is } \ell\text{-cocoercive} \iff \mathrm{Id} - \tfrac{2}{\ell}F \text{ is non-expansive}$$

That is, it is sufficient to find cocoercive operator *F* and points *x*, *y* such that

$$\|x - \gamma_2 F_{\mathrm{EG},\gamma_1}(x) - (y - \gamma_2 F_{\mathrm{EG},\gamma_1}(y))\| > \|x - y\|$$

for any meaningful choices of the stepsizes.

Non-cocoercivity of EG-operator: for all $\ell > 0$ and $\gamma_1 \in (0, 1/\ell]$ there exists ℓ-cocoercive operator *F* such that EG-operator is non-cocoercive.

- The proof is obtained numerically via solving a special SDP [3,4]

References

[1] N. Golowich, S. Pattathil, C. Daskalakis, and A. Ozdaglar. **Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems.** In *Conference on Learning Theory*, pages 1758–1784. PMLR, 2020.

[2] H. H. Bauschke, P. L. Combettes, et al. **Convex analysis and monotone operator theory in Hilbert spaces**, volume 408. Springer, 2011.

[3] A. B. Taylor, J. M. Hendrickx, and F. Glineur. **Performance estimation toolbox (PESTO): automated worst-case analysis of first-order optimization methods**. *In 2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 1278–1283. IEEE, 2017.

[4] E. K. Ryu, A. B. Taylor, C. Bergeling, and P. Giselsson. **Operator splitting performance estimation: Tight contraction factors and optimal parameter selection.** *SIAM Journal on Optimization*, 30(3):2251–2271, 2020