Extragradient Method: $\mathcal{O}(1/\kappa)$ Last-Iterate Convergence for Monotone Variational Inequalities and Connections With Cocoercivity

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

[1] Moscow Institute of Physics and Technology, Russian Federation
[2] Mila, Université de Montréal, Canada
[3] Canada CIFAR AI Chair

AISTATS 2022

March 29, 2022

Short Summary of Our Work

- We prove $\mathcal{O}(1/\kappa)$ *last-iterate* convergence rate for Extragardient method [Korpelevich, 1976] in terms of squared operator norm for monotone Lipschitz variational ineqiality problems (VIPs)

Short Summary of Our Work

- We prove $\mathcal{O}(1/\kappa)$ *last-iterate* convergence rate for Extragardient method [Korpelevich, 1976] in terms of squared operator norm for monotone Lipschitz variational ineqiality problems (VIPs)
  - The proof is *obtained via computer*

Short Summary of Our Work

- We prove $\mathcal{O}(1/\kappa)$ *last-iterate* convergence rate for Extragardient method [Korpelevich, 1976] in terms of squared operator norm for monotone Lipschitz variational ineqiality problems (VIPs)
  - The proof is *obtained via computer*
- We establish new connections of several known methods with cocoercivity when the original operator is monotone and Lipschitz

Short Summary of Our Work

- We prove $\mathcal{O}(1/\kappa)$ *last-iterate* convergence rate for Extragardient method [Korpelevich, 1976] in terms of squared operator norm for monotone Lipschitz variational ineqiality problems (VIPs)
  - The proof is *obtained via computer*
- We establish new connections of several known methods with cocoercivity when the original operator is monotone and Lipschitz
- Our code is available online: `https://github.com/eduardgorbunov/extragradient_last_iterate_AISTATS_2022`

Outline

Variational Inequality Problem

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

- $F : Q \to \mathbb{R}^d$ is L-Lipschitz operator: $\forall x, y \in \mathbb{R}^d$

$$\|F(x) - F(y)\| \leq L\|x - y\| \tag{1}$$

- F is monotone: $\forall x, y \in \mathbb{R}^d$

$$\langle F(x) - F(y), x - y\rangle \geq 0 \tag{2}$$

Variational Inequality Problem: Examples

- Min-max problems:

$$\min_{u \in \mathbb{R}^{d_1}} \max_{v \in \mathbb{R}^{d_2}} f(u, v) \tag{3}$$

These problems appear in various applications such as robust optimization [Ben-Tal et al., 2009] and control [Hast et al., 2013], adversarial training [Goodfellow et al., 2015, Madry et al., 2018] and generative adversarial networks (GANs) [Goodfellow et al., 2014].

Variational Inequality Problem: Examples

- Min-max problems:

$$\min_{u \in \mathbb{R}^{d_1}} \max_{v \in \mathbb{R}^{d_2}} f(u, v) \tag{3}$$

  These problems appear in various applications such as robust optimization [Ben-Tal et al., 2009] and control [Hast et al., 2013], adversarial training [Goodfellow et al., 2015, Madry et al., 2018] and generative adversarial networks (GANs) [Goodfellow et al., 2014].

- Minimization problems:

$$\min_{x \in \mathbb{R}^d} f(x) \tag{4}$$

  If f is convex, then (4) is equivalent to finding a solution of (VIP) with

$$F(x) = \nabla f(x)$$

How to Solve VIPs?

Naive approach – Gradient Descent (GD):

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

✓ GD seems very natural and it is well-studied for minimization

How to Solve VIPs?

Naive approach – Gradient Descent (GD):

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

- ✓ GD seems very natural and it is well-studied for minimization
- ✗ GD does not converge for simple convex-concave min-max problems

Non-Convergence of GD

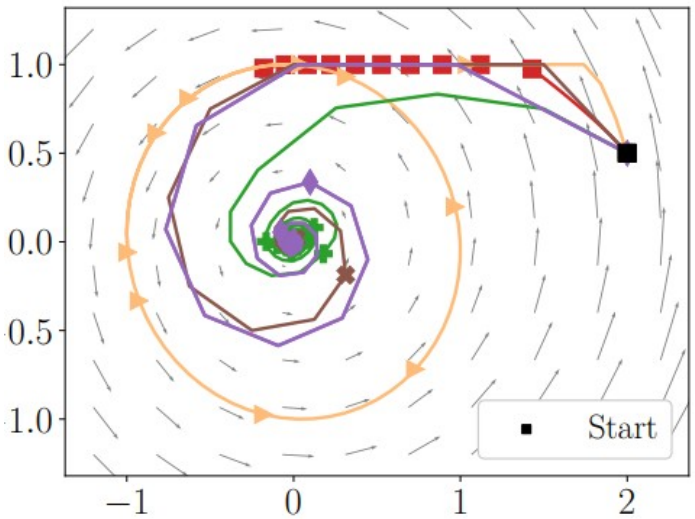

Figure 1: Comparison of the basic gradient method (as well as Adam) with the techniques presented in §3 on the optimization of (9). Only the algorithms advocated in this paper (Averaging, Extrapolation and Extrapolation from the past) converge quickly to the solution. Each marker represents 20 iterations. We compare these algorithms on a non-convex objective in §G.1.

Figure: Behavior of GD (and other methods) on the problem $\min_{u \in \mathbb{R}} \max_{v \in \mathbb{R}} uv$ [Gidel et al., 2019]

Popular Alternatives to GD

- Extragradient method (EG) [Korpelevich, 1976]

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

- Optimistic Gradient method (OG) [Popov, 1980]

$$x^{k+1} = x^k - 2\gamma F(x^k) + \gamma F(x^{k-1})$$

Popular Alternatives to GD

- Extragradient method (EG) [Korpelevich, 1976]

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

- Optimistic Gradient method (OG) [Popov, 1980]

$$x^{k+1} = x^k - 2\gamma F(x^k) + \gamma F(x^{k-1})$$

In this talk, we focus on EG and, in particular, on its convergence properties

Measures of Convergence

- **Restricted gap function**: $\texttt{Gap}_F(x^K) = \max\limits_{y \in \mathbb{R}^d : \|y - x^*\| \leq R} \langle F(y), x^K - y \rangle$, where $R \sim \|x^0 - x^*\|$ [Nesterov, 2007]

Measures of Convergence

- **Restricted gap function**: $\texttt{Gap}_F(x^K) = \max\limits_{y \in \mathbb{R}^d : \|y - x^*\| \leq R} \langle F(y), x^K - y \rangle$, where $R \sim \|x^0 - x^*\|$ [Nesterov, 2007]
  - ✓ $\texttt{Gap}_F(x^K)$ can be seen as a natural extension of optimization error for (VIP), when F is monotone
  - ✗ It is unclear how to tightly estimate $\texttt{Gap}_F(x^K)$ in practice and how to generalize it to non-monotone case

Measures of Convergence

- **Restricted gap function**: $\texttt{Gap}_F(x^K) = \max\limits_{y \in \mathbb{R}^d : \|y - x^*\| \leq R} \langle F(y), x^K - y \rangle$, where $R \sim \|x^0 - x^*\|$ [Nesterov, 2007]
  - ✓ $\texttt{Gap}_F(x^K)$ can be seen as a natural extension of optimization error for (VIP), when F is monotone
  - ✗ It is unclear how to tightly estimate $\texttt{Gap}_F(x^K)$ in practice and how to generalize it to non-monotone case
- **Squared norm of the operator**: $\|F(x^K)\|^2$

Measures of Convergence

- **Restricted gap function**: $\texttt{Gap}_F(x^K) = \max\limits_{y \in \mathbb{R}^d : \|y - x^*\| \leq R} \langle F(y), x^K - y \rangle$, where $R \sim \|x^0 - x^*\|$ [Nesterov, 2007]
  - ✓ $\texttt{Gap}_F(x^K)$ can be seen as a natural extension of optimization error for (VIP), when F is monotone
  - ✗ It is unclear how to tightly estimate $\texttt{Gap}_F(x^K)$ in practice and how to generalize it to non-monotone case
- **Squared norm of the operator**: $\|F(x^K)\|^2$
  - ✗ In general, it provides weaker guarantees than $\texttt{Gap}_F(x^K)$
  - ✓ $\|F(x^K)\|^2$ is easier to compute than $\texttt{Gap}_F(x^K)$

In this talk, we focus on the guarantees for $\|F(x^K)\|^2$

Convergence Guarantees for EG

When F is monotone and L-Lipschitz the following results are known for EG:

Convergence Guarantees for EG

When F is monotone and L-Lipschitz the following results are known for EG:

- **Averaged- and best-iterate guarantees:**
  - $\texttt{Gap}_F(\overline{x}^K) = \mathcal{O}(1/K)$ for $\overline{x}^K = \frac{1}{K+1}\sum_{k=0}^K x^k$ [Nemirovski, 2004, Mokhtari et al., 2019, Hsieh et al., 2019, Monteiro and Svaiter, 2010, Auslender and Teboulle, 2005]
  - $\min\limits_{k=0,1,\ldots,K} \|F(x^k)\|^2 = \mathcal{O}(1/K)$ [Solodov and Svaiter, 1999, Ryu et al., 2019]

Convergence Guarantees for EG

When F is monotone and L-Lipschitz the following results are known for EG:

- **Averaged- and best-iterate guarantees:**
  - $\texttt{Gap}_F(\overline{x}^K) = \mathcal{O}(1/K)$ for $\overline{x}^K = \frac{1}{K+1}\sum_{k=0}^K x^k$ [Nemirovski, 2004, Mokhtari et al., 2019, Hsieh et al., 2019, Monteiro and Svaiter, 2010, Auslender and Teboulle, 2005]
  - $\min\limits_{k=0,1,\ldots,K} \|F(x^k)\|^2 = \mathcal{O}(1/K)$ [Solodov and Svaiter, 1999, Ryu et al., 2019]
- **Lower bounds for the last-iterate [Golowich et al., 2020]:**
  - $\texttt{Gap}_F(x^K) = \Omega(1/\sqrt{K})$
  - $\|F(x^K)\|^2 = \Omega(1/K)$

Convergence Guarantees for EG

When F is monotone and L-Lipschitz the following results are known for EG:

- **Averaged- and best-iterate guarantees:**
  - $\texttt{Gap}_F(\overline{x}^K) = \mathcal{O}(1/K)$ for $\overline{x}^K = \frac{1}{K+1}\sum_{k=0}^K x^k$ [Nemirovski, 2004, Mokhtari et al., 2019, Hsieh et al., 2019, Monteiro and Svaiter, 2010, Auslender and Teboulle, 2005]
  - $\min\limits_{k=0,1,\ldots,K} \|F(x^k)\|^2 = \mathcal{O}(1/K)$ [Solodov and Svaiter, 1999, Ryu et al., 2019]
- **Lower bounds for the last-iterate [Golowich et al., 2020]:**
  - $\texttt{Gap}_F(x^K) = \Omega(1/\sqrt{K})$
  - $\|F(x^K)\|^2 = \Omega(1/K)$
- **Upper bounds for the last-iterate [Golowich et al., 2020]:** *if additionally the Jacobian $\nabla F(x)$ is Λ-Lipschitz*, then
  - $\texttt{Gap}_F(x^K) = \mathcal{O}(1/\sqrt{K})$
  - $\|F(x^K)\|^2 = \mathcal{O}(1/K)$

Convergence Guarantees for EG: Resolved Question

Is it possible to prove last-iterate $\|F(x^K)\|^2 = \mathcal{O}(1/K)$ *convergence rate for* EG *when* F *is monotone and* L*-Lipschitz without additional assumptions?*

Convergence Guarantees for EG: Resolved Question

Is it possible to prove last-iterate $\|F(x^K)\|^2 = \mathcal{O}(1/K)$ *convergence rate for* EG *when* F *is monotone and* L*-Lipschitz without additional assumptions?*

We give a positive answer to this question in our paper

Last-Iterate $\mathcal{O}(1/\kappa)$ Rate for EG

Theorem 6

Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be monotone and L-Lipschitz, $0 < \gamma \leq 1/\sqrt{2}L$. Then for all $k \geq 0$ the iterates produced by EG satisfy $\|F(x^{k+1})\| \leq \|F(x^k)\|$.

Last-Iterate $\mathcal{O}(1/K)$ Rate for EG

Theorem 6

Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be monotone and L-Lipschitz, $0 < \gamma \leq 1/\sqrt{2}L$. Then for all $k \geq 0$ the iterates produced by EG satisfy $\|F(x^{k+1})\| \leq \|F(x^k)\|$.

Using this result, it is quite trivial to derive last-iterate $\mathcal{O}(1/K)$ rate.

Theorem 7

Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be monotone and L-Lipschitz. Then for all $K \geq 0$

$$\|F(x^K)\|^2 \leq \frac{\|x^0 - x^*\|^2}{\gamma^2(1 - L^2\gamma^2)(K+1)}, \tag{5}$$

where x^K is produced by EG with stepsize $0 < \gamma \leq 1/\sqrt{2}L$. Moreover,

$$\texttt{Gap}_F(x^K) = \max_{y \in \mathbb{R}^d : \|y - x^*\| \leq \|x^0 - x^*\|} \langle F(y), x^K - y \rangle \leq \frac{2\|x^0 - x^*\|^2}{\gamma\sqrt{1 - L^2\gamma^2}\sqrt{K+1}}. \tag{6}$$

The Proof

We derive this result by solving a special SDP **numerically**

In the Paper We Also Have

- Several connections with cocoercivity of operators corresponding to Extragradient method, Optimistic Gradient method and Hamiltonian method
- Non-trivial negative results established via PEP
- Link to the code: `https://github.com/eduardgorbunov/extragradient_last_iterate_AISTATS_2022`

References I

A. Auslender and M. Teboulle. Interior projection-like methods for monotone variational inequalities. *Mathematical programming*, 104(1):39–68, 2005.

A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton university press, 2009.

E. De Klerk. *Aspects of semidefinite programming: interior point algorithms and selected applications*, volume 65. Springer Science & Business Media, 2006.

E. De Klerk, F. Glineur, and A. B. Taylor. On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions. *Optimization Letters*, 11(7):1185–1199, 2017.

Y. Drori and M. Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1):451–482, 2014.

G. Gidel, H. Berard, P. Vincent, and S. Lacoste-Julien. A variational inequality perspective on generative adversarial nets. In *ICLR*, 2019.

References II

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.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *ICLR 2015*, 2015.

M. Hast, K. J. Åström, B. Bernhardsson, and S. Boyd. Pid design by convex-concave optimization. In *2013 European Control Conference (ECC)*, pages 4460–4465. IEEE, 2013.

Y.-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. *Advances in Neural Information Processing Systems*, 32:6938–6948, 2019.

References III

D. Kim and J. A. Fessler. Optimized first-order methods for smooth convex minimization. *Mathematical programming*, 159(1):81–107, 2016.

G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26 (1):57–95, 2016.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *ICLR 2018*, 2018.

A. Mokhtari, A. Ozdaglar, and S. Pattathil. Proximal point approximations achieving a convergence rate of $O(1/k)$ for smooth convex-concave saddle point problems: Optimistic gradient and extra-gradient methods. *arXiv preprint arXiv:1906.01115*, 2019.

References IV

R. D. Monteiro and B. F. Svaiter. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM Journal on Optimization*, 20(6):2755–2787, 2010.

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 Journal on Optimization*, 15(1): 229–251, 2004.

Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344, 2007.

L. D. Popov. A modification of the arrow-hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28(5): 845–848, 1980.

References V

E. K. Ryu, K. Yuan, and W. Yin. Ode analysis of stochastic gradient methods with optimism and anchoring for minimax problems. *arXiv preprint arXiv:1905.10899*, 2019.

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.

M. V. Solodov and B. F. Svaiter. A hybrid approximate extragradient–proximal point algorithm using the enlargement of a maximal monotone operator. *Set-Valued Analysis*, 7(4):323–345, 1999.

A. Taylor and F. Bach. Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions. In *Conference on Learning Theory*, pages 2934–2992. PMLR, 2019.

References VI

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, 2017a.

A. B. Taylor, J. M. Hendrickx, and F. Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1-2):307–345, 2017b.