Convergence of Proximal Point and Extragradient-Based Methods Beyond Monotonicity: the Case of Negative Comonotonicity

Eduard Gorbunov [1] Adrien Taylor [2] Samuel Horváth [1] Gauthier Gidel [3]

[1]Mohamed bin Zayed University of Artificial Intelligence [2]INRIA & D.I. École Normale Supérieure, CNRS & PSL Research University [3]Mila, Université de Montréal, Canada CIFAR AI chair

1. Inclusion Problems

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

- $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is some (possibly set-valued) mapping
- $\mathrm{Gr}(F) := \{(u, F_u) \mid F_u \in F(u)\}$
- Generalization of minimization, saddle points, and variational inequalities problems
- Standard assumption is (maximal) monotonicity:

$$\langle F(x) - F(y), x - y\rangle \geq 0$$

- In many real-world problems, monotonicity does not hold
- *We focus on the structured non-monotone problems*

2. Negative Comonotonicity

Definition 1. ρ-Negative comonotonicity (cohypomonotonicity [1])

$$\langle F_x - F_y, x - y\rangle \geq -\rho\|F_x - F_y\|^2, \quad \forall x, y. \qquad (1)$$

Definition 2. Star-negative comonotonicity (weak Minty condition [2])

Operator $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is called ρ-star-negative comonotone for some $\rho \geq 0$ if $\forall$ $(x, F_x) \in \mathrm{Gr}(F)$ and x^* being a solution of (IP)

$$\langle F_x, x - x^*\rangle \geq -\rho\|F_x\|^2. \qquad (2)$$

- We assume that the mapping F is *maximal* in the sense that its graph is not strictly contained in the graph of any other ρ-negative comonotone operator (resp., ρ-star-negative comonotone)
- Some examples star-negative comonotone operators that are non-monotone can be found in [3]
- The next theorem provides a spectral viewpoint on NC

Theorem 1. Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be a continuously differentiable. Then, the following statements are equivalent:

- F is ρ-negative comonotone,
- $\Re(1/\lambda) \geq -\rho$ for all $\lambda \in \mathrm{Sp}(\nabla F(x)) := \{\lambda \in \mathbb{C} \mid \det(\nabla F(x) - \lambda I) = 0\}$, $\forall x \in \mathbb{R}^d$.

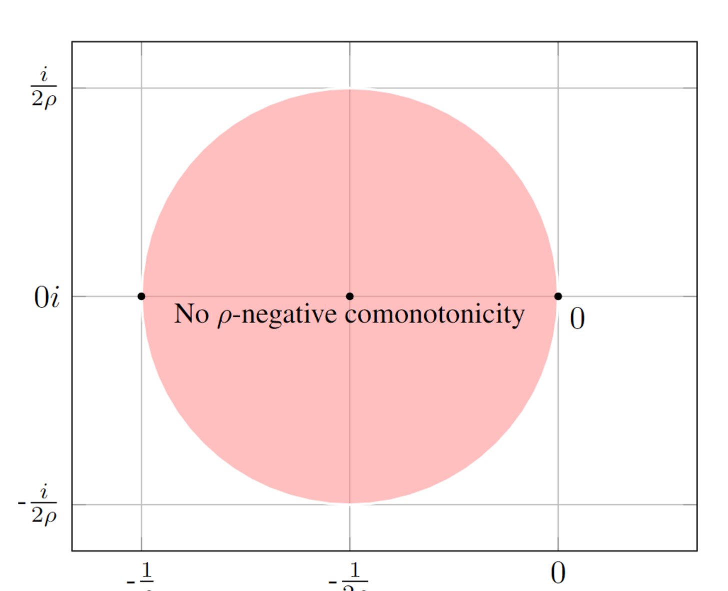

Figure: Visualization of Theorem 1. Red open disc corresponds to the constraint $\Re(1/\lambda) < -\rho$ that defines the set such that all eigenvalues the Jacobian of ρ-negative comonotone operator should lie outside this set.

Theorem 2 (Corollary 3.15 from [3]). If $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is maximally ρ-negative comonotone, then the solution set $X^* = F^{-1}(0)$ is convex.

✗ Negative comonotonicity is not satisfied for many practical tasks

✓ Studying the convergence of traditional methods under NC is a natural step towards understanding their behaviors in more complicated non-monotonic cases

Main Contributions

◇ **Closer look at Proximal Point method**

- $\mathcal{O}(1/N)$ last-iterate and best-iterate convergence rates under negative comonotonicity and star-negative comonotonicity assumptions, respectively
- Worst-case examples and counter-examples for the case when the stepsize is smaller than 2ρ

◇ **New results for Extragradient-based methods**

- $\mathcal{O}(1/N)$ last-iterate convergence of EG and OG under milder assumptions on the negative comonotonicity parameter ρ than in the prior work [5]
- Counter-examples showing that the range of ρ cannot be improved for EG and OG (for the best-iterate convergence)

3. Proximal Point Method

$$x^{k+1} = x^k - \gamma F(x^{k+1}). \qquad \text{(PP)}$$

- We analyze the worst-case behavior of (PP) using Performance Estimation Problems (PEPs) [6, 7, 8]

$$\max_{F, x^0} \ \|x^N - x^{N-1}\|^2 \qquad (3)$$
$$\text{s.t.} \quad F \text{ satisfies (2)},$$
$$\|x^0 - x^*\|^2 \leq R^2,\ 0 \in F(x^*),$$
$$x^{k+1} = x^k - \gamma F(x^{k+1}), \quad k = 0, 1, \ldots, N-1.$$

Theorem 3. Problem (3) can be reformulated as an SDP.

- Solving the resulting SDP numerically, we verified $\mathcal{O}(1/N)$ rate
- Using the trace heuristic, we found the worst-case example
- Finally, we constructed counter-example showing that (PP) is not necessary converging when $\gamma < 2\rho$

Theorem 4 (Upper bounds).

- Let $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be maximally ρ-star-negative comonotone. Then, for any $\gamma > 2\rho$ the iterates produced by PP are well-defined and satisfy $\forall N \geq 1$:

$$\frac{1}{N}\sum_{k=1}^{N}\|x^k - x^{k-1}\|^2 \leq \frac{\gamma\|x^0 - x^*\|^2}{(\gamma - 2\rho)N}. \qquad (4)$$

- If $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is maximally ρ-negative comonotone, then for any $\gamma > 2\rho$ and any $k \geq 1$ the iterates produced by PP satisfy $\|x^{k+1} - x^k\| \leq \|x^k - x^{k-1}\|$ and for any $N \geq 1$:

$$\|x^N - x^{N-1}\|^2 \leq \frac{\gamma\|x^0 - x^*\|^2}{(\gamma - 2\rho)N}. \qquad (5)$$

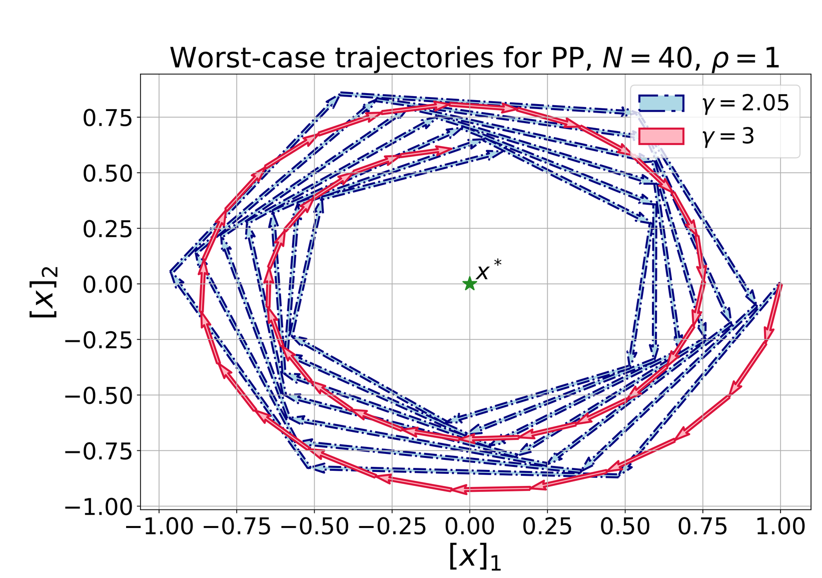

Figure: The worst-case trajectories of PP for $N = 40$. The form of trajectories hints that the worst-case operator is a rotation operator.

Comparison with Prior Work

Known and new $\mathcal{O}(1/N)$ convergence results for PP, EG and OG. Notation: NC = negative comonotonicity, SNC = star-negative comonotonicity, L-Lip. = L-Lipschitzness. Green color: the derived results are completely novel/extend the existing ones.

Method	Setup	$\rho \in$	Convergence	Reference	Counter-/Worst-case examples?
PP[(1)]	NC	$[0, +\infty)$	Last-iterate	Theorem 4	Theorem 5 (worst-case example & divergence for $\gamma \leq 2\rho$)
	SNC	$[0, +\infty)$	Best-iterate	Theorem 4	Theorem 5 (worst-case example & divergence for $\gamma \leq 2\rho$)
EG	NC + L-Lip.	$[0, 1/16L)$	Last-iterate	[5]	✗
	NC + L-Lip.	$[0, 1/8L)$	Last-iterate	Theorem 6	Theorem 6 (diverge for $\rho \geq 1/2L$ and any $\gamma_1, \gamma_2 > 0$)
	SNC + L-Lip.	$[0, 1/8L)$	Best-iterate	[2]	✗
	SNC + L-Lip.	$[0, 1/2L)$	Best-iterate	[3]	Theorem 3.4 (diverge for $\gamma_1 = 1/L$ and $\rho \geq (1-L\gamma_2)/2L$)
	SNC + L-Lip.	$[0, 1/2L)$	Best-iterate	Theorem 6 [(2)]	Theorem 6 (diverge for $\rho \geq 1/2L$ and any $\gamma_1, \gamma_2 > 0$)
OG	NC + L-Lip.	$[0, 8/(27\sqrt{6}L))$	Last-iterate	[5]	✗
	NC + L-Lip.	$[0, 5/62L)$	Last-iterate	Theorem 7	Theorem 7 (diverge for $\rho \geq 1/2L$ and any $\gamma_1, \gamma_2 > 0$)
	SNC + L-Lip.	$[0, 1/2L)$	Best-iterate	[9]	✗
	SNC + L-Lip.	$[0, 1/2L)$	Best-iterate	Theorem 7 [(2)]	Theorem 7 (diverge for $\rho \geq 1/2L$ and any $\gamma_1, \gamma_2 > 0$)

[(1)] The best-iterate convergence result can be obtained from Lemma 2 [10], and the last-iterate convergence result can also be derived from the non-expansiveness of PP update, see Proposition 3.13 (iii) [4]. At the moment of writing our paper, we were not aware of these results.

[(2)] Although these results are not new for the best-iterate convergence of EG and OG, the proof techniques differ from prior works.

Theorem 5 (Wosrt-case example and counter-examples).

- For any $\rho > 0, \gamma > 2\rho$, and $N \geq \max\{\rho^2/\gamma(\gamma-2\rho), 1\}$ consider two-dimensional $F : \mathbb{R}^2 \to \mathbb{R}$: $F(x) = \alpha A x$ with

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \quad \alpha = \frac{|\cos\theta|}{\rho}$$

for $\theta \in (\pi/2, \pi)$ such that $\cos\theta = -\frac{\rho}{\sqrt{N\gamma(\gamma-2\rho)}}$. Then, F is ρ-negative comonotona and after N iterations PP with stepsize γ produces x^{N+1} satisfying

$$\|F(x^{N+1})\|^2 \geq \frac{\|x^0 - x^*\|^2}{\gamma(\gamma - 2\rho)N\left(1 + \frac{1}{N}\right)^{N+1}}. \qquad (6)$$

- For any $\rho > 0$ there exists ρ-negatively comonotone single-valued operator $F : \mathbb{R}^d \to \mathbb{R}^d$ (e.g., $F(x) = -x/\rho$) such that PP does not converge to the solution of IP for any $0 < \gamma \leq 2\rho$.

4. Extragradient

$$\widetilde{x}^k = x^k - \gamma_1 F(x^k), \quad \forall k \geq 0. \qquad \text{(EG)}$$
$$x^{k+1} = x^k - \gamma_2 F(\widetilde{x}^k),$$

Theorem 6.

- Let F be L-Lipschitz and ρ-star-negative comonotone with $\rho < 1/2L$. Then, for any $2\rho < \gamma_1 < 1/L$ and $0 < \gamma_2 \leq \gamma_1 - 2\rho$ the iterates produced by EG after $N \geq 0$ iteration satisfy

$$\frac{1}{N+1}\sum_{k=0}^{N}\|F(x^k)\|^2 \leq \frac{\|x^0 - x^*\|^2}{\gamma_1\gamma_2(1 - L^2\gamma_1^2)(N+1)}. \qquad (7)$$

- If, in addition, F is ρ-negative comonotone with $\rho \leq 1/8L$ and $\gamma_1 = \gamma_2 = \gamma$ such that $4\rho \leq \gamma \leq 1/2L$, then for any $k \geq 0$ the iterates produced by EG satisfy $\|F(x^{k+1})\| \leq \|F(x^k)\|$ and for any $N \geq 1$

$$\|F(x^N)\|^2 \leq \frac{28\|x^0 - x^*\|^2}{N\gamma^2 + 320\gamma\rho}. \qquad (8)$$

- For $\rho \geq 1/2L$ and any choice of stepsizes $\gamma_1, \gamma_2 > 0$ EG does not necessary converges on solving IP with this operator F. In particular, for $\gamma_1 > 1/L$ it is sufficient to take $F(x) = Lx$, and for $0 < \gamma_1 \leq 1/L$ one can take $F(x) = LAx$, where $x \in \mathbb{R}^2$,

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \quad \theta = \frac{2\pi}{3}.$$

5. Optimistic Gradient

$$\widetilde{x}^k = x^k - \gamma_1 F(\widetilde{x}^{k-1}), \quad \forall k > 0, \qquad \text{(OG)}$$
$$x^{k+1} = x^k - \gamma_2 F(\widetilde{x}^k), \quad \forall k \geq 0,$$

Theorem 7.

- Let F be L-Lipschitz and ρ-star-negative comonotone with $\rho < 1/2L$. Then, for any $2\rho < \gamma_1 < 1/L$ and $0 < \gamma_2 \leq \min\{1/L - \gamma_1, \gamma_1 - 2\rho\}$ the iterates produced by OG after $N \geq 0$ iteration satisfy

$$\frac{1}{N+1}\sum_{k=0}^{N}\|F(x^k)\|^2 \leq \frac{\|x^0 - x^*\|^2}{\gamma_1\gamma_2(1 - L^2(\gamma_1 + \gamma_2)^2)(N+1)}. \qquad (9)$$

- If, in addition, F is ρ-negative comonotone with $\rho \leq 5/62L$ and $\gamma_1 = \gamma_2 = \gamma$ such that $4\rho \leq \gamma \leq 10/31L$, then for any $N \geq 1$ the iterates produced by OG satisfy

$$\|F(x^N)\|^2 \leq \frac{717\|x^0 - x^*\|^2}{N\gamma(\gamma - 3\rho) + 800\gamma^2}. \qquad (10)$$

- For $\rho \geq 1/2L$ and any choice of stepsizes $\gamma_1, \gamma_2 > 0$ OG does not necessary converges on solving IP with this operator F. In particular, for $\gamma_1 > 1/L$ it is sufficient to take $F(x) = Lx$, and for $0 < \gamma_1 \leq 1/L$ one can take $F(x) = LAx$, where $x \in \mathbb{R}^2$,

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \quad \theta = \frac{2\pi}{3}.$$

- The proofs for (EG) and (OG) are potential-based proof and were discovered via PEP

References

[1] Pennanen, T. (2002). Local convergence of the proximal point algorithm and multiplier methods without monotonicity. Mathematics of Operations Research, 27(1):170–191

[2] Diakonikolas, J., Daskalakis, C., and Jordan, M. I. (2021). Efficient methods for structured nonconvex-nonconcave min-max optimization. ICML 2021.

[3] Pethick, T., Latafat, P., Patrinos, P., Fercoq, O., and Cevher, V. Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems. ICLR 2022.

[4] Bauschke, H. H., Moursi, W. M., and Wang, X. Generalized monotone operators and their averaged resolvents. Mathematical Programming, 189:55–74, 2021.

[5] Luo, Y. and Tran-Dinh, Q. Last-iterate convergence rates and randomized block-coordinate variant of extragradient-type methods for co-monotone equations. preprint, 2022.

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

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

[8] Taylor, A. B., Hendrickx, J. M., and Glineur, F. Exact worst-case performance of first-order methods for composite convex optimization. SIAM Journal on Optimization, 27 (3):1283–1313, Jan 2017.

[9] Böhm, A. Solving nonconvex-nonconcave min-max problems exhibiting weak minty solutions. arXiv preprint arXiv:2201.12247, 2022.

[10] Iusem, A. N., Pennanen, T., and Svaiter, B. F. Inexact variants of the proximal point algorithm without monotonicity. SIAM Journal on Optimization, 13(4):1080–1097, 2003.