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[3,4]

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

PEP Talks 2023, UCLovain, Belgium

Outline

Monotone Inclusion Problems

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

- $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is some (possibly set-valued) mapping

Monotone Inclusion Problems

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

- $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is some (possibly set-valued) mapping
- Classical assumption: F is maximally monotone, i.e., $\forall (x, X), (y, Y) \in \mathtt{Gr}(F) = \{(u, U) \in \mathbb{R}^{d\times d} \mid U \in F(u)\}$

$$\langle X - Y, x - y\rangle \geq 0 \tag{1}$$

and there is no other monotone operator H such that $\mathtt{Gr}(F) \subset \mathtt{Gr}(H)$

Monotone Inclusion Problems: Examples

- Min-max problems:

$$\min_{u \in \mathbb{R}^{d_u}} \max_{v \in \mathbb{R}^{d_v}} f(u, v). \tag{2}$$

If f is convex-concave, then (2) is equivalent to solving (IP) with

$$F(x) = \begin{pmatrix} \partial_u f(u, v) \\ \partial_v(-f(u, v)) \end{pmatrix}, \quad x = (u, v)$$

Monotone Inclusion Problems: Examples

- Min-max problems:

$$\min_{u\in\mathbb{R}^{d_u}} \max_{v\in\mathbb{R}^{d_v}} f(u,v). \tag{2}$$

If f is convex-concave, then (2) is equivalent to solving (IP) with

$$F(x) = \begin{pmatrix} \partial_u f(u,v) \\ \partial_v(-f(u,v)) \end{pmatrix}, \quad x = (u,v)$$

- Minimization problems:

$$\min_{x\in\mathbb{R}^d} f(x) \tag{3}$$

If f is convex, then (3) is equivalent to finding a solution of (IP) with

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

Methods for Monotone Inclusion Problems

- Proximal Point (PP) method [Martinet, 1970, Rockafellar, 1976]:

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

Methods for Monotone Inclusion Problems

- Proximal Point (PP) method [Martinet, 1970, Rockafellar, 1976]:

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

- Extragradient (EG) method [Korpelevich, 1976]:

$$\begin{aligned} \widetilde{x}^k &= x^k - \gamma_1 F(x^k), \\ x^{k+1} &= x^k - \gamma_2 F(\widetilde{x}^k), \end{aligned} \quad \forall k \geq 0, \tag{EG}$$

Methods for Monotone Inclusion Problems

- Proximal Point (PP) method [Martinet, 1970, Rockafellar, 1976]:

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

- Extragradient (EG) method [Korpelevich, 1976]:

$$\begin{aligned} \widetilde{x}^k &= x^k - \gamma_1 F(x^k), \\ x^{k+1} &= x^k - \gamma_2 F(\widetilde{x}^k), \end{aligned} \quad \forall k \geq 0, \tag{EG}$$

- Optimistic Gradient (OG) method [Popov, 1980]: $\widetilde{x}^0 = x^0$ and

$$\begin{aligned} \widetilde{x}^k &= x^k - \gamma_1 F(\widetilde{x}^{k-1}), \quad \forall k > 0, \\ x^{k+1} &= x^k - \gamma_2 F(\widetilde{x}^k), \quad \forall k \geq 0, \end{aligned} \tag{OG}$$

Methods for Monotone Inclusion Problems

- Proximal Point (PP) method [Martinet, 1970, Rockafellar, 1976]:

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

- Extragradient (EG) method [Korpelevich, 1976]:

$$\begin{aligned} \widetilde{x}^k &= x^k - \gamma_1 F(x^k), \\ x^{k+1} &= x^k - \gamma_2 F(\widetilde{x}^k), \end{aligned} \quad \forall k \geq 0, \tag{EG}$$

- Optimistic Gradient (OG) method [Popov, 1980]: $\widetilde{x}^0 = x^0$ and

$$\begin{aligned} \widetilde{x}^k &= x^k - \gamma_1 F(\widetilde{x}^{k-1}), \quad \forall k > 0, \\ x^{k+1} &= x^k - \gamma_2 F(\widetilde{x}^k), \quad \forall k \geq 0, \end{aligned} \tag{OG}$$

These methods are well-studied for **monotone** problems

Negative Comonotone Operators

Operator $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is maximally negative comonotone if $\forall (x, X), (y, Y) \in \mathtt{Gr}(F)$

$$\langle X - Y, x - y \rangle \geq -\rho \|X - Y\|^2 \tag{4}$$

and there is no other monotone operator H such that $\mathtt{Gr}(F) \subset \mathtt{Gr}(H)$

Negative Comonotone Operators

Operator $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is maximally negative comonotone if $\forall (x, X), (y, Y) \in \texttt{Gr}(F)$

$$\langle X - Y, x - y \rangle \geq -\rho \|X - Y\|^2 \tag{4}$$

and there is no other monotone operator H such that $\texttt{Gr}(F) \subset \texttt{Gr}(H)$

- Negative comonotonicity (also known as cohypomonotonicity) was introduced by Pennanen [2002]

Negative Comonotone Operators

Operator $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is maximally negative comonotone if $\forall (x, X), (y, Y) \in \mathtt{Gr}(F)$

$$\langle X - Y, x - y \rangle \geq -\rho \|X - Y\|^2 \tag{4}$$

and there is no other monotone operator H such that $\mathtt{Gr}(F) \subset \mathtt{Gr}(H)$

- Negative comonotonicity (also known as cohypomonotonicity) was introduced by Pennanen [2002]
- Star-negative comonotonicity (also known as weak Minty condition) was introduced by Diakonikolas et al. [2021]: instead of (4) this assumption requires

$$\langle X, x - x^* \rangle \geq -\rho \|X\|^2, \quad \text{where} \quad 0 \in F(x^*) \tag{5}$$

Negative Comonotone Operators

Operator $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is maximally negative comonotone if $\forall (x, X), (y, Y) \in \texttt{Gr}(F)$

$$\langle X - Y, x - y \rangle \geq -\rho \|X - Y\|^2 \tag{4}$$

and there is no other monotone operator H such that $\texttt{Gr}(F) \subset \texttt{Gr}(H)$

- Negative comonotonicity (also known as cohypomonotonicity) was introduced by Pennanen [2002]
- Star-negative comonotonicity (also known as weak Minty condition) was introduced by Diakonikolas et al. [2021]: instead of (4) this assumption requires
$$\langle X, x - x^* \rangle \geq -\rho \|X\|^2, \quad \text{where} \quad 0 \in F(x^*) \tag{5}$$
- These assumptions are the weakest known ones under which EG-type methods can be analyzed

Convergence under Negative Comonotonicity

- Proximal Point method
  - ✓ Asymptotic convergence [Combettes and Pennanen, 2004]

Convergence under Negative Comonotonicity

- Proximal Point method
  - ✓ Asymptotic convergence [Combettes and Pennanen, 2004]
  - ? Non-asymptotic analysis and convergence rates are missing

Convergence under Negative Comonotonicity

- Proximal Point method
  - ✓ Asymptotic convergence [Combettes and Pennanen, 2004]
  - ? Non-asymptotic analysis and convergence rates are missing
  - ? Worst-case examples are unknown

Convergence under Negative Comonotonicity

- Proximal Point method
  - ✓ Asymptotic convergence [Combettes and Pennanen, 2004]
  - ? Non-asymptotic analysis and convergence rates are missing
  - ? Worst-case examples are unknown
- Extragradient and Optimistic Gradient methods
  - ✓ Best-iterate $\mathcal{O}(1/N)$ convergence under star-negative comonotonicity for $\rho < 1/2L$ [Pethick et al., 2022, Böhm, 2022] (F is assumed to be L-Lipschitz)

Convergence under Negative Comonotonicity

- Proximal Point method
  - ✓ Asymptotic convergence [Combettes and Pennanen, 2004]
  - ? Non-asymptotic analysis and convergence rates are missing
  - ? Worst-case examples are unknown
- Extragradient and Optimistic Gradient methods
  - ✓ Best-iterate $\mathcal{O}(1/N)$ convergence under star-negative comonotonicity for $\rho < 1/2L$ [Pethick et al., 2022, Böhm, 2022] (F is assumed to be L-Lipschitz)
  - ✓ Last-iterate $\mathcal{O}(1/N)$ convergence under negative comonotonicity for $\rho < 1/16L$ (in the case of EG) and $\rho < 8/27\sqrt{6}L$ (in the case of OG) [Luo and Tran-Dinh, 2022]

Convergence under Negative Comonotonicity

- Proximal Point method
  - ✓ Asymptotic convergence [Combettes and Pennanen, 2004]
  - ? Non-asymptotic analysis and convergence rates are missing
  - ? Worst-case examples are unknown
- Extragradient and Optimistic Gradient methods
  - ✓ Best-iterate $\mathcal{O}(1/N)$ convergence under star-negative comonotonicity for $\rho < 1/2L$ [Pethick et al., 2022, Böhm, 2022] (F is assumed to be L-Lipschitz)
  - ✓ Last-iterate $\mathcal{O}(1/N)$ convergence under negative comonotonicity for $\rho < 1/16L$ (in the case of EG) and $\rho < 8/27\sqrt{6}L$ (in the case of OG) [Luo and Tran-Dinh, 2022]
  - ✓ Non-convergence of EG when $\gamma_1 = 1/L$ and $\rho \geq (1-L\gamma_2)/2L$ [Pethick et al., 2022]

Convergence under Negative Comonotonicity

- Proximal Point method
  - ✓ Asymptotic convergence [Combettes and Pennanen, 2004]
  - ? Non-asymptotic analysis and convergence rates are missing
  - ? Worst-case examples are unknown
- Extragradient and Optimistic Gradient methods
  - ✓ Best-iterate $\mathcal{O}(1/N)$ convergence under star-negative comonotonicity for $\rho < 1/2L$ [Pethick et al., 2022, Böhm, 2022] (F is assumed to be L-Lipschitz)
  - ✓ Last-iterate $\mathcal{O}(1/N)$ convergence under negative comonotonicity for $\rho < 1/16L$ (in the case of EG) and $\rho < 8/27\sqrt{6}L$ (in the case of OG) [Luo and Tran-Dinh, 2022]
  - ✓ Non-convergence of EG when $\gamma_1 = 1/L$ and $\rho \geq (1-L\gamma_2)/2L$ [Pethick et al., 2022]
  - ? Can the bound on ρ be improved when $\gamma_1 \neq 1/L$ in the case of EG?

Convergence under Negative Comonotonicity

- Proximal Point method
  - ✓ Asymptotic convergence [Combettes and Pennanen, 2004]
  - ? Non-asymptotic analysis and convergence rates are missing
  - ? Worst-case examples are unknown
- Extragradient and Optimistic Gradient methods
  - ✓ Best-iterate $\mathcal{O}(1/N)$ convergence under star-negative comonotonicity for $\rho < 1/2L$ [Pethick et al., 2022, Böhm, 2022] (F is assumed to be L-Lipschitz)
  - ✓ Last-iterate $\mathcal{O}(1/N)$ convergence under negative comonotonicity for $\rho < 1/16L$ (in the case of EG) and $\rho < 8/27\sqrt{6}L$ (in the case of OG) [Luo and Tran-Dinh, 2022]
  - ✓ Non-convergence of EG when $\gamma_1 = 1/L$ and $\rho \geq (1 - L\gamma_2)/2L$ [Pethick et al., 2022]
  - ? Can the bound on ρ be improved when $\gamma_1 \neq 1/L$ in the case of EG?
  - ? What is the upper bound on ρ in the case of OG?

Convergence under Negative Comonotonicity

- Proximal Point method
  - ✓ Asymptotic convergence [Combettes and Pennanen, 2004]
  - ? Non-asymptotic analysis and convergence rates are missing
  - ? Worst-case examples are unknown
- Extragradient and Optimistic Gradient methods
  - ✓ Best-iterate $\mathcal{O}(1/N)$ convergence under star-negative comonotonicity for $\rho < 1/2L$ [Pethick et al., 2022, Böhm, 2022] (F is assumed to be L-Lipschitz)
  - ✓ Last-iterate $\mathcal{O}(1/N)$ convergence under negative comonotonicity for $\rho < 1/16L$ (in the case of $\texttt{EG}$) and $\rho < 8/27\sqrt{6}L$ (in the case of $\texttt{OG}$) [Luo and Tran-Dinh, 2022]
  - ✓ Non-convergence of $\texttt{EG}$ when $\gamma_1 = 1/L$ and $\rho \geq (1-L\gamma_2)/2L$ [Pethick et al., 2022]
  - ? Can the bound on ρ be improved when $\gamma_1 \neq 1/L$ in the case of $\texttt{EG}$?
  - ? What is the upper bound on ρ in the case of $\texttt{OG}$?
  - ? Is it possible to achieve last-iterate convergence for $\rho \in [1/16L, 1/2L)$ (in the case of $\texttt{EG}$) and $\rho \geq [8/27\sqrt{6}L, 1/2L)$ (in the case of $\texttt{OG}$)?

Convergence under Negative Comonotonicity

- Proximal Point method
  - ✓ Asymptotic convergence [Combettes and Pennanen, 2004]
  - ? Non-asymptotic analysis and convergence rates are missing
  - ? Worst-case examples are unknown
- Extragradient and Optimistic Gradient methods
  - ✓ Best-iterate $\mathcal{O}(1/N)$ convergence under star-negative comonotonicity for $\rho < 1/2L$ [Pethick et al., 2022, Böhm, 2022] (F is assumed to be L-Lipschitz)
  - ✓ Last-iterate $\mathcal{O}(1/N)$ convergence under negative comonotonicity for $\rho < 1/16L$ (in the case of EG) and $\rho < 8/27\sqrt{6}L$ (in the case of OG) [Luo and Tran-Dinh, 2022]
  - ✓ Non-convergence of EG when $\gamma_1 = 1/L$ and $\rho \geq (1-L\gamma_2)/2L$ [Pethick et al., 2022]
  - ? Can the bound on ρ be improved when $\gamma_1 \neq 1/L$ in the case of EG?
  - ? What is the upper bound on ρ in the case of OG?
  - ? Is it possible to achieve last-iterate convergence for $\rho \in [1/16L, 1/2L)$ (in the case of EG) and $\rho \geq [8/27\sqrt{6}L, 1/2L)$ (in the case of OG)?

In our work, we address these questions (the last question is resolved partially)

PEP for the Analysis of PP

- PP: $x^{k+1} = x^k - \gamma F(x^{k+1})$
- Convergence metric: $\|x^N - x^{N-1}\|^2$

PEP for the Analysis of PP

- PP: $x^{k+1} = x^k - \gamma F(x^{k+1})$
- Convergence metric: $\|x^N - x^{N-1}\|^2$
- Let's PEP it!

$$\begin{aligned} \max_{F,d,x^0} \quad & \|x^N - x^{N-1}\|^2 \\ \text{s.t.} \quad & F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d \text{ is max. } \rho\text{-negative comonotone}, \\ & \|x^0 - x^*\|^2 \leq R^2,\ 0 \in F(x^*), \\ & x^{k+1} = x^k - \gamma F(x^{k+1}), \quad k = 0, 1, \ldots, N-1. \end{aligned} \tag{6}$$

PEP for the Analysis of PP

- PP: $x^{k+1} = x^k - \gamma F(x^{k+1})$
- Convergence metric: $\|x^N - x^{N-1}\|^2$
- Let's PEP it!

$$\begin{aligned} \max_{F,d,x^0} \quad & \|x^N - x^{N-1}\|^2 \\ \text{s.t.} \quad & F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d \text{ is max. } \rho\text{-negative comonotone}, \\ & \|x^0 - x^*\|^2 \leq R^2,\ 0 \in F(x^*), \\ & x^{k+1} = x^k - \gamma F(x^{k+1}), \quad k = 0, 1, \ldots, N-1. \end{aligned} \tag{6}$$

✗ Maximization over *infinitely-dimensional* space of ρ-negative comonotone operators

Finitely-Dimensional PEP for the Analysis of PP

$$\max_{\substack{d \\ x^*, x^0, x^1, \ldots, x^N \in \mathbb{R}^d \\ g^*, g^0, g^1, \ldots, g^N \in \mathbb{R}^d}} \quad \|x^N - x^{N-1}\|^2 \tag{7}$$

$$\text{s.t.} \quad F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d \text{ is max. } \rho\text{-negative comonotone,} \tag{8}$$

$$g^k \in F(x^k), \quad k = *, 0, 1, \ldots, N, \quad g^* = 0, \tag{9}$$

$$\|x^0 - x^*\|^2 \leq R^2,$$

$$x^{k+1} = x^k - \gamma g^{k+1}, \quad k = 0, 1, \ldots, N-1.$$

Finitely-Dimensional PEP for the Analysis of PP

$$\max_{\substack{d \\ x^*, x^0, x^1, \ldots, x^N \in \mathbb{R}^d \\ g^*, g^0, g^1, \ldots, g^N \in \mathbb{R}^d}} \quad \|x^N - x^{N-1}\|^2 \tag{7}$$

$$\text{s.t.} \quad F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d \text{ is max. } \rho\text{-negative comonotone}, \tag{8}$$

$$g^k \in F(x^k), \quad k = *, 0, 1, \ldots, N, \quad g^* = 0, \tag{9}$$

$$\|x^0 - x^*\|^2 \leq R^2,$$

$$x^{k+1} = x^k - \gamma g^{k+1}, \quad k = 0, 1, \ldots, N-1.$$

- ✓ *Finitely-dimensional* problem
- ✓ Equivalent to the original PEP

Finitely-Dimensional PEP for the Analysis of PP

$$\max_{\substack{d\\ x^*,x^0,x^1,\ldots,x^N\in\mathbb{R}^d\\ g^*,g^0,g^1,\ldots,g^N\in\mathbb{R}^d}} \quad \|x^N - x^{N-1}\|^2 \tag{7}$$

$$\text{s.t.} \quad F:\mathbb{R}^d \rightrightarrows \mathbb{R}^d \text{ is max. } \rho\text{-negative comonotone}, \tag{8}$$

$$g^k \in F(x^k), \quad k = *, 0, 1, \ldots, N, \quad g^* = 0, \tag{9}$$

$$\|x^0 - x^*\|^2 \leq R^2,$$

$$x^{k+1} = x^k - \gamma g^{k+1}, \quad k = 0, 1, \ldots, N-1.$$

- ✓ *Finitely-dimensional* problem
- ✓ Equivalent to the original PEP
- ✗ Non-trivial constraints (8)-(9)

Interpolation Conditions for Negative Comonotonicity

Theorem 1

Let $\{(x^k, g^k)\}_{k=0}^N \subseteq \mathbb{R}^d \times \mathbb{R}^d$ be some finite set of pairs of points in $\mathbb{R}^d$.

Interpolation Conditions for Negative Comonotonicity

Theorem 1

Let $\{(x^k, g^k)\}_{k=0}^N \subseteq \mathbb{R}^d \times \mathbb{R}^d$ be some finite set of pairs of points in $\mathbb{R}^d$. There exists a maximal ρ-negative comonotone operator $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ such that $g^k \in F(x^k)$, $k = 0, \ldots, N$ if and only if

$$\langle g^i - g^j, x^i - x^j \rangle \geq -\rho \|g^i - g^j\|^2 \quad \forall i, j = 0, \ldots, N. \tag{10}$$

Finitely-Dimensional PEP for the Analysis of PP: Better Version

$$
\begin{aligned}
\max_{\substack{d \\ x^*, x^0, x^1, \ldots, x^N \in \mathbb{R}^d \\ g^*, g^0, g^1, \ldots, g^N \in \mathbb{R}^d}} \quad & \|x^N - x^{N-1}\|^2 \qquad (11) \\
\text{s.t.} \quad & \langle g^i - g^j, x^i - x^j \rangle \geq -\rho \|g^i - g^j\|^2, \; i, j = *, 0, 1, \ldots, N, \\
& g^* = 0, \\
& \|x^0 - x^*\|^2 \leq R^2, \\
& x^{k+1} = x^k - \gamma g^{k+1}, \quad k = 0, 1, \ldots, N-1
\end{aligned}
$$

Finitely-Dimensional PEP for the Analysis of PP: Better Version

$$
\begin{aligned}
\max_{\substack{x^*, x^0, x^1, \ldots, x^N \in \mathbb{R}^d \\ g^*, g^0, g^1, \ldots, g^N \in \mathbb{R}^d}} \quad & \|x^N - x^{N-1}\|^2 \\
\text{s.t.} \quad & \langle g^i - g^j, x^i - x^j \rangle \geq -\rho \|g^i - g^j\|^2, \; i, j = *, 0, 1, \ldots, N, \\
& g^* = 0, \\
& \|x^0 - x^*\|^2 \leq R^2, \\
& x^{k+1} = x^k - \gamma g^{k+1}, \quad k = 0, 1, \ldots, N-1
\end{aligned} \tag{11}
$$

- ✓ *Finitely-dimensional* problem
- ✓ Equivalent to the original PEP
- ✓ Can be reformulated as SDP using the standard steps for PEPs [Taylor et al., 2017, Ryu et al., 2020]

SDP Reformulation of PEP for PP

$$\begin{aligned} \max_{G \in \mathbb{S}_+^{N+3}} \quad & \mathrm{Tr}(M_0 G) \\ \text{s.t.} \quad & \mathrm{Tr}(M_i G) \leq 0, \quad i = 1, 2 \ldots, (N+2)(N+3), \\ & \mathrm{Tr}(M_{-1} G) \leq R^2 \end{aligned} \tag{12}$$

- $G = V^\top V$, where $V = (x^*, x^0, g^0, g^1, \ldots, g^N)$
- Matrices M_i encode the objective and constraints

SDP Reformulation of PEP for PP

$$
\begin{aligned}
\max_{G \in \mathbb{S}_+^{N+3}} \quad & \mathrm{Tr}(M_0 G) \\
\text{s.t.} \quad & \mathrm{Tr}(M_i G) \leq 0, \quad i = 1, 2 \ldots, (N+2)(N+3), \\
& \mathrm{Tr}(M_{-1} G) \leq R^2
\end{aligned}
\tag{12}
$$

- $G = V^\top V$, where $V = (x^*, x^0, g^0, g^1, \ldots, g^N)$
- Matrices M_i encode the objective and constraints
- Using the *trace heuristic* [Taylor et al., 2017] one can generate low-dimensional worst-case examples

Worst-case Examples for PP

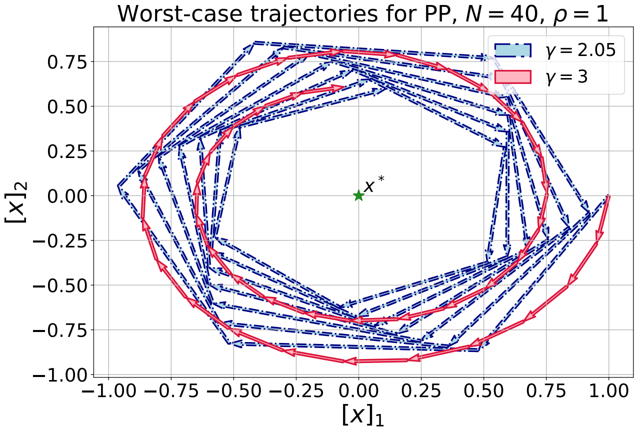
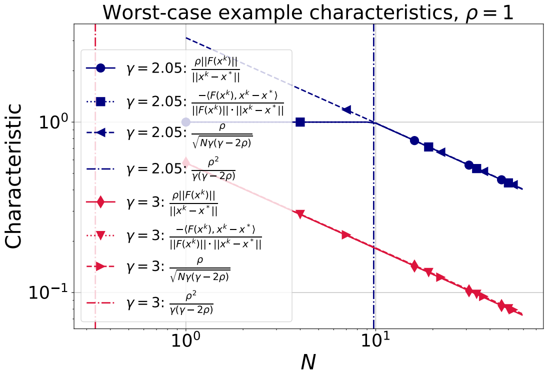


Figure: The worst-case trajectories of PP for $N = 40$. The form of trajectories hints that the worst-case operator is a rotation operator.

Worst-case Example Characteristics



- $\rho\|F(x^k)\|/\|x^k-x^*\|$ and $-\langle F(x^k),x^k-x^*\rangle/(\|F(x^k)\|\cdot\|x^k-x^*\|)$ remain the same during the run of the method
- These characteristics coincide with $\rho/\sqrt{N\gamma(\gamma-2\rho)}$ as long as the total number of steps N is sufficiently large ($N \geq \max\{\rho^2/\gamma(\gamma-2\rho), 1\}$)

Worst-case Example: Explicit Expression

Theorem 2

For any $\rho > 0, \gamma > 2\rho$, and $N \geq \max\{\rho^2/\gamma(\gamma-2\rho), 1\}$ there exists ρ-negatively comonotone single-valued operator $F : \mathbb{R}^d \to \mathbb{R}^d$ such that after N iterations PP with stepsize γ produces x^{N+1} satisfying

$$\|F(x^{N+1})\|^2 \geq \frac{\|x^0 - x^*\|^2}{\gamma(\gamma - 2\rho)N\left(1 + \frac{1}{N}\right)^{N+1}}. \tag{13}$$

Worst-case Example: Explicit Expression

Theorem 2

For any $\rho > 0, \gamma > 2\rho$, and $N \geq \max\{\rho^2/\gamma(\gamma-2\rho), 1\}$ there exists ρ-negatively comonotone single-valued operator $F : \mathbb{R}^d \to \mathbb{R}^d$ such that after N iterations PP with stepsize γ produces x^{N+1} satisfying

$$\|F(x^{N+1})\|^2 \geq \frac{\|x^0 - x^*\|^2}{\gamma(\gamma - 2\rho)N\left(1 + \frac{1}{N}\right)^{N+1}}. \tag{13}$$

Indeed, one can pick the 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 $\cos\theta = -\frac{\rho}{\sqrt{N\gamma(\gamma-2\rho)}}$.

Convergence Results for PP

Theorem 3

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

$$\frac{1}{N}\sum_{k=1}^{N}\|x^k - x^{k-1}\|^2 \leq \frac{\gamma\|x^0 - x^*\|^2}{(\gamma - 2\rho)N}. \tag{14}$$

Convergence Results for PP

Theorem 3

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

$$\frac{1}{N}\sum_{k=1}^{N}\|x^k - x^{k-1}\|^2 \leq \frac{\gamma\|x^0 - x^*\|^2}{(\gamma - 2\rho)N}. \quad (14)$$

If $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is 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 $N \geq 1$:

$$\|x^N - x^{N-1}\|^2 \leq \frac{\gamma\|x^0 - x^*\|^2}{(\gamma - 2\rho)N}. \quad (15)$$

Large Stepsize is Mandatory for PP

Theorem 4

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

Large Stepsize is Mandatory for PP

Theorem 4

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

This a relatively rare phenomenon when decreasing the stepsize leads to non-convergence

Numerical Verification of the Upper Bound

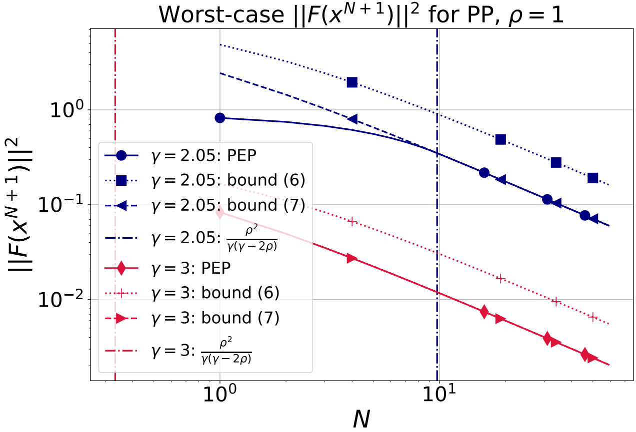


Figure: The solution of PEP compared with the lower bound ("bound (6)" in the plot) and the upper bound ("bound (7)" in the plot) for different values of γ and N.

Summary of the Obtained Results

- Proximal Point method
  - Last-iterate $\mathcal{O}(1/N)$ convergence under negative comonotonicity for $\gamma > 2\rho$
  - Best-iterate $\mathcal{O}(1/N)$ convergence under star-negative comonotonicity $\gamma > 2\rho$
  - Worst-case examples matching (up to numerical factor) the upper bound
  - Counter-examples for $\gamma \leq 2\rho$
- Extragradient and Optimistic Gradient methods
  - Last-iterate $\mathcal{O}(1/N)$ convergence under negative comonotonicity for $\rho \leq 1/8L$ (in the case of EG) and $\rho \leq 5/62L$ (in the case of OG)
  - Counter-examples for $\rho \geq 1/2L$ and any stepsizes

References I

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

Y. Cai, A. Oikonomou, and W. Zheng. Tight last-iterate convergence of the extragradient method for constrained monotone variational inequalities. *arXiv preprint arXiv:2204.09228 version 1*, 2022.

P. L. Combettes and T. Pennanen. Proximal methods for cohypomonotone operators. *SIAM journal on control and optimization*, 43(2):731–742, 2004.

J. Diakonikolas, C. Daskalakis, and M. I. Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021.

N. Golowich, S. Pattathil, and C. Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. *arXiv preprint arXiv:2010.13724*, 2020a.

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, 2020b.

E. Gorbunov, N. Loizou, and G. Gidel. Extragradient method: $O(1/k)$ last-iterate convergence for monotone variational inequalities and connections with cocoercivity. pages 366–402, 2022a.

E. Gorbunov, A. Taylor, and G. Gidel. Last-iterate convergence of optimistic gradient method for monotone variational inequalities. *arXiv preprint arXiv:2205.08446*, 2022b.

G. Gu and J. Yang. Optimal nonergodic sublinear convergence rate of proximal point algorithm for maximal monotone inclusion problems. *arXiv preprint arXiv:1904.05495*, 2019.

B. He and X. Yuan. On the convergence rate of douglas–rachford operator splitting method. *Mathematical Programming*, 153(2):715–722, 2015.

References III

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

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

B. Martinet. Regularisation d'inequations variationelles par approximations successives. *Revue Francaise d'Informatique et de Recherche Operationelle*, 4: 154–159, 1970.

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

T. Pethick, P. Latafat, P. Patrinos, O. Fercoq, and V. Cevherå. Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems. In *International Conference on Learning Representations*, 2022.

References IV

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.

R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

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.

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.

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.

Methods for **Monotone** Inclusions are Well-Studied

- Proximal Point (PP) method
  - ✓ Last-iterate guarantees: $\|x^N - x^{N-1}\|^2 = \mathcal{O}(1/N)$ [He and Yuan, 2015]

Methods for **Monotone** Inclusions are Well-Studied

- Proximal Point (PP) method
  - ✓ Last-iterate guarantees: $\|x^N - x^{N-1}\|^2 = \mathcal{O}(1/N)$ [He and Yuan, 2015]
  - ✓ Worst-case examples and matching lower-bounds for PP [Gu and Yang, 2019]

Methods for **Monotone** Inclusions are Well-Studied

- Proximal Point (PP) method
  - ✓ Last-iterate guarantees: $\|x^N - x^{N-1}\|^2 = \mathcal{O}(1/N)$ [He and Yuan, 2015]
  - ✓ Worst-case examples and matching lower-bounds for PP [Gu and Yang, 2019]
- Extragradient and Optimistic Gradient (EG) methods:
  - ✓ Last-iterate guarantees when F and ∇F are Lipschitz: $\|F(x^N)\|^2 = \mathcal{O}(1/N)$ [Golowich et al., 2020b,a]

Methods for **Monotone** Inclusions are Well-Studied

- Proximal Point (PP) method
  - ✓ Last-iterate guarantees: $\|x^N - x^{N-1}\|^2 = \mathcal{O}(1/N)$ [He and Yuan, 2015]
  - ✓ Worst-case examples and matching lower-bounds for PP [Gu and Yang, 2019]
- Extragradient and Optimistic Gradient (EG) methods:
  - ✓ Last-iterate guarantees when F and ∇F are Lipschitz: $\|F(x^N)\|^2 = \mathcal{O}(1/N)$ [Golowich et al., 2020b,a]
  - ✓ Matching lower bounds [Golowich et al., 2020b,a]

Methods for **Monotone** Inclusions are Well-Studied

- Proximal Point (PP) method
  - ✓ Last-iterate guarantees: $\|x^N - x^{N-1}\|^2 = \mathcal{O}(1/N)$ [He and Yuan, 2015]
  - ✓ Worst-case examples and matching lower-bounds for PP [Gu and Yang, 2019]
- Extragradient and Optimistic Gradient (EG) methods:
  - ✓ Last-iterate guarantees when F and ∇F are Lipschitz: $\|F(x^N)\|^2 = \mathcal{O}(1/N)$ [Golowich et al., 2020b,a]
  - ✓ Matching lower bounds [Golowich et al., 2020b,a]
  - ✓ Last-iterate guarantees when F is Lipschitz: $\|F(x^N)\|^2 = \mathcal{O}(1/N)$ [Gorbunov et al., 2022a,b, Cai et al., 2022]

Methods for **Monotone** Inclusions are Well-Studied

- Proximal Point (PP) method
  - ✓ Last-iterate guarantees: $\|x^N - x^{N-1}\|^2 = \mathcal{O}(1/N)$ [He and Yuan, 2015]
  - ✓ Worst-case examples and matching lower-bounds for PP [Gu and Yang, 2019]
- Extragradient and Optimistic Gradient (EG) methods:
  - ✓ Last-iterate guarantees when F and ∇F are Lipschitz: $\|F(x^N)\|^2 = \mathcal{O}(1/N)$ [Golowich et al., 2020b,a]
  - ✓ Matching lower bounds [Golowich et al., 2020b,a]
  - ✓ Last-iterate guarantees when F is Lipschitz: $\|F(x^N)\|^2 = \mathcal{O}(1/N)$ [Gorbunov et al., 2022a,b, Cai et al., 2022]

Can we relax the monotonicity assumption to achieve similar results?

Spectral Viewpoint on Negative Comonotonicity

Theorem 8

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

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

Spectral Viewpoint on Negative Comonotonicity

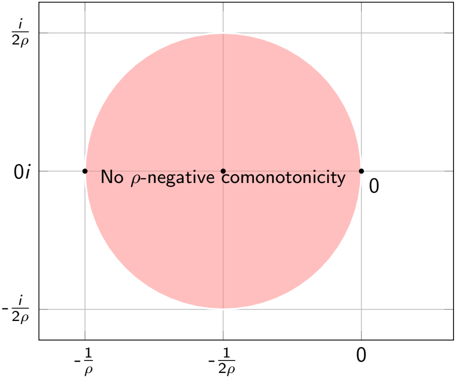


Figure: Visualization of Theorem 8. Red open disc corresponds to the constraint $\mathrm{Re}(1/\lambda) < -\rho$ that defines the set such that all eigenvalues the Jacobian of ρ-negative comonotone operator should lie outside this set.

Negative Comonotonicity Does Not Allow Separated Optima

Theorem 9

If $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is ρ-negative comonotone, then the solution set $X^* = F^{-1}(0)$ is convex.

Negative Comonotonicity Does Not Allow Separated Optima

Theorem 9

If $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is ρ-negative comonotone, then the solution set $X^* = F^{-1}(0)$ is convex.

Proof sketch

- F and $(F^{-1} + \rho \cdot \mathrm{Id})^{-1}$ have the same set of solutions

Negative Comonotonicity Does Not Allow Separated Optima

Theorem 9

If $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is ρ-negative comonotone, then the solution set $X^* = F^{-1}(0)$ is convex.

Proof sketch

- F and $(F^{-1} + \rho \cdot \mathrm{Id})^{-1}$ have the same set of solutions
- F is maximally negative comonotone $\iff$ $F^{-1} + \rho \cdot \mathrm{Id}$ is maximally monotone $\iff$ $(F^{-1} + \rho \cdot \mathrm{Id})^{-1}$ is maximally monotone

Negative Comonotonicity Does Not Allow Separated Optima

Theorem 9

If $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is ρ-negative comonotone, then the solution set $X^* = F^{-1}(0)$ is convex.

Proof sketch

- F and $(F^{-1} + \rho \cdot \mathrm{Id})^{-1}$ have the same set of solutions
- F is maximally negative comonotone $\Longleftrightarrow$ $F^{-1} + \rho \cdot \mathrm{Id}$ is maximally monotone $\Longleftrightarrow$ $(F^{-1} + \rho \cdot \mathrm{Id})^{-1}$ is maximally monotone
- $((F^{-1} + \rho \cdot \mathrm{Id})^{-1})^{-1}(0)$ is convex as the set of zeros of maximally comonotone operator

Negative Comonotonicity Does Not Allow Separated Optima

Theorem 9

If $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is ρ-negative comonotone, then the solution set $X^* = F^{-1}(0)$ is convex.

Proof sketch

- F and $(F^{-1} + \rho \cdot \mathrm{Id})^{-1}$ have the same set of solutions
- F is maximally negative comonotone $\iff F^{-1} + \rho \cdot \mathrm{Id}$ is maximally monotone $\iff (F^{-1} + \rho \cdot \mathrm{Id})^{-1}$ is maximally monotone
- $((F^{-1} + \rho \cdot \mathrm{Id})^{-1})^{-1}(0)$ is convex as the set of zeros of maximally comonotone operator

Nevertheless, studying the convergence of traditional methods under negative comonotonicity can be seen as a natural step towards understanding their behaviors in more complicated non-monotonic cases

Extra Slide on Trace Heuristic

- First one needs to solve SDP (12) (denote the optimal value as v_*)
- Then one can solve another SDP

$$\min_{G \in \mathbb{S}_+^{N+3}} \quad \mathrm{Tr}(G) \tag{16}$$

$$\text{s.t.} \quad \mathrm{Tr}(M_i G) \leq 0, \quad i = 1, 2 \ldots, (N+2)(N+3),$$

$$\mathrm{Tr}(M_{-1} G) \leq R^2,$$

$$\mathrm{Tr}(M_0 G) = v_* \tag{17}$$

in the hope of finding a low-rank solution

Last-Iterate Convergence of EG

Theorem 5

If $F : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz and 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

$$\|F(x^{k+1})\| \leq \|F(x^k)\| \quad (18)$$

Last-Iterate Convergence of EG

Theorem 5

If $F : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz and 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

$$\|F(x^{k+1})\| \leq \|F(x^k)\| \tag{18}$$

and for any $N \geq 1$

$$\|F(x^N)\|^2 \leq \frac{28\|x^0 - x^*\|^2}{N\gamma^2 + 320\gamma\rho}. \tag{19}$$

Last-Iterate Convergence of EG

Theorem 5

If $F : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz and 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

$$\|F(x^{k+1})\| \leq \|F(x^k)\| \tag{18}$$

and for any $N \geq 1$

$$\|F(x^N)\|^2 \leq \frac{28\|x^0 - x^*\|^2}{N\gamma^2 + 320\gamma\rho}. \tag{19}$$

- ✓ Potential from (18) and the proof of (18) were found via computer (the approach was motivated by [Taylor and Bach, 2019])

Last-Iterate Convergence of EG

Theorem 5

If $F : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz and 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

$$\|F(x^{k+1})\| \leq \|F(x^k)\| \tag{18}$$

and for any $N \geq 1$

$$\|F(x^N)\|^2 \leq \frac{28\|x^0 - x^*\|^2}{N\gamma^2 + 320\gamma\rho}. \tag{19}$$

- ✓ Potential from (18) and the proof of (18) were found via computer (the approach was motivated by [Taylor and Bach, 2019])
- ? **Open question:** is it possible to show $\mathcal{O}(1/N)$ last-iterate convergence when $\rho \in (1/8L, 1/2L)$?

Last-Iterate Convergence of OG

Theorem 6

If $F : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz and 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 $k \geq 0$ the iterates produced by OG satisfy

$$\begin{aligned} \|F(x^{k+1})\|^2 + \|F(x^{k+1}) - F(\widetilde{x}^k)\|^2 \quad &\leq \quad \|F(x^k)\|^2 + \|F(x^k) - F(\widetilde{x}^{k-1})\|^2 \\ &\quad - \frac{1}{100}\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2. \end{aligned} \tag{20}$$

Last-Iterate Convergence of OG

Theorem 6

If $F : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz and 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 $k \geq 0$ the iterates produced by OG satisfy

$$\|F(x^{k+1})\|^2 + \|F(x^{k+1}) - F(\widetilde{x}^k)\|^2 \leq \|F(x^k)\|^2 + \|F(x^k) - F(\widetilde{x}^{k-1})\|^2 - \frac{1}{100}\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2. \quad (20)$$

and for any $N \geq 1$

$$\|F(x^N)\|^2 \leq \frac{717\|x^0 - x^*\|^2}{N\gamma(\gamma - 3\rho) + 800\gamma^2}. \quad (21)$$

Last-Iterate Convergence of OG

Theorem 6

If $F : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz and 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 $k \geq 0$ the iterates produced by OG satisfy

$$\|F(x^{k+1})\|^2 + \|F(x^{k+1}) - F(\widetilde{x}^k)\|^2 \leq \|F(x^k)\|^2 + \|F(x^k) - F(\widetilde{x}^{k-1})\|^2 - \frac{1}{100}\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2. \quad (20)$$

and for any $N \geq 1$

$$\|F(x^N)\|^2 \leq \frac{717\|x^0 - x^*\|^2}{N\gamma(\gamma - 3\rho) + 800\gamma^2}. \quad (21)$$

✓ Potential from (18) and the proof of (18) were found via computer

Last-Iterate Convergence of OG

Theorem 6

If $F : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz and 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 $k \geq 0$ the iterates produced by OG satisfy

$$\|F(x^{k+1})\|^2 + \|F(x^{k+1}) - F(\widetilde{x}^k)\|^2 \leq \|F(x^k)\|^2 + \|F(x^k) - F(\widetilde{x}^{k-1})\|^2 - \frac{1}{100}\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2. \quad (20)$$

and for any $N \geq 1$

$$\|F(x^N)\|^2 \leq \frac{717\|x^0 - x^*\|^2}{N\gamma(\gamma - 3\rho) + 800\gamma^2}. \quad (21)$$

- ✓ Potential from (18) and the proof of (18) were found via computer
- ? **Open question:** is it possible to show $\mathcal{O}(1/N)$ last-iterate convergence when $\rho \in (5/62L, 1/2L)$?

No Convergence when $\rho \geq 1/2L$

Theorem 7

For any $L > 0$, $\rho \geq 1/2L$, and any choice of stepsizes $\gamma_1, \gamma_2 > 0$ there exists ρ-negative comonotone L-Lipschitz operator F such that EG/OG does not necessary converges on solving IP with this operator F.

No Convergence when $\rho \geq 1/2L$

Theorem 7

For any $L > 0$, $\rho \geq 1/2L$, and any choice of stepsizes $\gamma_1, \gamma_2 > 0$ there exists ρ-negative comonotone L-Lipschitz operator F such that EG/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 $x \in \mathbb{R}^2$,

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