Last-Iterate Convergence of Extragradient-Based Methods

EUROPT 2024, Lund

Eduard Gorbunov\textsuperscript{1} Adrien Taylor\textsuperscript{2} Samuel Horváth\textsuperscript{1} Nicolas Loizou\textsuperscript{3} Gauthier Gidel\textsuperscript{4,5}

June 26, 2024

\textsuperscript{1} Mohamed bin Zayed University of Artificial Intelligence, UAE
\textsuperscript{2} INRIA & D.I. École Normale Supérieure, CNRS & PSL Research University, France
\textsuperscript{3} Johns Hopkins University
\textsuperscript{4} Université de Montréal and Mila, Canada
\textsuperscript{5} Canada CIFAR AI Chair
1. Variational Inequalities and Extragradient-Based Methods

2. Performance Estimation Problems and Last-Iterate Convergence of Extragradient

3. Last-Iterate Convergence of Optimistic Gradient

4. Last-Iterate Convergence Under Negative-Comonotonicity
The Talk is Based on Three Papers


Variational Inequalities and Extragradient-Based Methods
Variational Inequality Problem

find $x^* \in Q \subseteq \mathbb{R}^d$ such that $\langle F(x^*), x - x^* \rangle \geq 0$, $\forall x \in Q$ (VIP-C)

- $F : Q \to \mathbb{R}^d$ is $L$-Lipschitz operator: $\forall x, y \in Q$
  \[ \|F(x) - F(y)\| \leq L\|x - y\| \] (1)

- $F$ is monotone: $\forall x, y \in Q$
  \[ \langle F(x) - F(y), x - y \rangle \geq 0 \] (2)
• Min-max problems:

$$\min_{u \in U} \max_{v \in V} f(u, v)$$

(3)

If $f$ is convex-concave, then (3) is equivalent to finding $(u^*, v^*) \in U \times V$ such that

$$\forall (u, v) \in U \times V$$

$$\langle \nabla_u f(u^*, v^*), u - u^* \rangle \geq 0,$$

$$-\langle \nabla_v f(u^*, v^*), v - v^* \rangle \geq 0,$$

which is equivalent to (VIP-C) with $Q = U \times V$, $x = (u^\top, v^\top)^\top$, and $F(x) = \nabla_u f(u, v) - \nabla_v f(u, v)$.

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:

\[
\min_{u \in U} \max_{v \in V} f(u, v) \tag{3}
\]

If \( f \) is convex-concave, then (3) is equivalent to finding \((u^*, v^*) \in U \times V\) such that \(\forall (u, v) \in U \times V\)

\[
\langle \nabla_u f(u^*, v^*), u - u^* \rangle \geq 0, \quad -\langle \nabla_v f(u^*, v^*), v - v^* \rangle \geq 0,
\]

which is equivalent to (VIP-C) with \( Q = U \times V \), \( x = (u^T, v^T)^T \), and

\[
F(x) = \begin{pmatrix}
\nabla_u f(u, v) \\
-\nabla_v f(u, v)
\end{pmatrix}
\]

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:

\[
\min_{x \in Q} f(x) \quad (4)
\]

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

\[
F(x) = \nabla f(x)
\]
When $Q = \mathbb{R}^d$ (VIP-C) can be rewritten as

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

For simplicity, we first consider (VIP) rather than (VIP-C)
How to Solve VIP?

Naïve approach – Gradient Descent (GD):

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

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

✗ GD does not converge for simple convex-concave min-max problems
Figure 1: Behavior of \( \text{GD} \) on the problem \( \min_{u \in \mathbb{R}} \max_{v \in \mathbb{R}} uv \) (Gidel et al., 2019)

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.
• Extragradient method (EG) (Korpelevich, 1976)

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

• Optimistic Gradient method (OG) (Popov, 1980)

\[ x^{k+1} = x^k - 2\gamma F(x^k) + \gamma F(x^{k-1}) \]
Measures of Convergence

• Restricted gap function:
  \[ \text{Gap}_F(x^K) = \max_{y \in \mathbb{R}^d : \|y - x^*\| \leq R} \langle F(y), x^K - y \rangle, \text{ where } R \sim \|x^0 - x^*\| \]
  (Nesterov, 2007)
Measures of Convergence

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

  ✓ \( \text{Gap}_F(x^K) \) can be seen as a natural extension of optimization error for (VIP), when \( F \) is monotone

  ✗ It is unclear how to tightly estimate \( \text{Gap}_F(x^K) \) in practice and how to generalize it to non-monotone case
Measures of Convergence

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

  (Nesterov, 2007)

  ✓ 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 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 Gap$_F(x^K)$

  ✓ $\| F(x^K) \|^2$ is easier to compute than Gap$_F(x^K)$

In this part of the talk talk, we focus on the guarantees for $\| F(x^K) \|^2$
When $F$ is monotone and $L$-Lipschitz the following results are known for EG:

- Averaged- and best-iterate guarantees:
  \[
  \text{Gap } F(x^K) = O\left(\frac{1}{K}\right) \quad \text{for} \quad x^K = \frac{1}{K+1} \sum_{k=0}^{K} x_k \quad (\text{Nemirovski, 2004; Mokhtari et al., 2019; Hsieh et al., 2019; Monteiro and Svaiter, 2010; Auslender and Teboulle, 2005})
  \]

- \[
  \min_{k=0, 1, \ldots, K} \|F(x_k)\|_2 = O\left(\frac{1}{K}\right) \quad (\text{Solodov and Svaiter, 1999; Ryu et al., 2019})
  \]

- Lower bounds for the last-iterate (Golowich et al., 2020):
  \[
  \text{Gap } F(x^K) = \Omega\left(\frac{1}{\sqrt{K}}\right)
  \]
  \[
  \|F(x^K)\|_2 = \Omega\left(\frac{1}{K}\right)
  \]

- Upper bounds for the last-iterate (Golowich et al., 2020):
  If additionally the Jacobian $\nabla F(x)$ is $\Lambda$-Lipschitz, then
  \[
  \text{Gap } F(x^K) = O\left(\frac{1}{\sqrt{K}}\right)
  \]
  \[
  \|F(x^K)\|_2 = O\left(\frac{1}{K}\right)
  \]
When $F$ is monotone and $L$-Lipschitz the following results are known for EG:

- **Averaged- and best-iterate guarantees:**
  - $\text{Gap}_F(\bar{x}^K) = O(1/k)$ for $\bar{x}^K = \frac{1}{K+1} \sum_{k=0}^{K} x^k$
Convergence Guarantees for EG

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

- **Averaged- and best-iterate guarantees:**
  
  \[ \text{Gap}_F(\bar{x}^K) = \mathcal{O}(1/k) \text{ for } \bar{x}^K = \frac{1}{K+1} \sum_{k=0}^{K} x^k \]  
  (Nemirovski, 2004; Mokhtari et al., 2019; Hsieh et al., 2019; Monteiro and Svaiter, 2010; Auslender and Teboulle, 2005)
When $F$ is monotone and $L$-Lipschitz the following results are known for EG:

- **Averaged- and best-iterate guarantees:**
  
  - $\text{Gap}_F(\overline{x}^K) = \mathcal{O}(1/k)$ 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)
  
  - $\min_{k=0,1,...,K} \|F(x^k)\|^2 = \mathcal{O}(1/k)$
Convergence Guarantees for EG

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

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

- **Lower bounds for the last-iterate (Golowich et al., 2020):**
Convergence Guarantees for EG

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

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

- **Lower bounds for the last-iterate (Golowich et al., 2020):**
  - $\text{Gap}_F(x^K) = \Omega(1/\sqrt{K})$
Convergence Guarantees for EG

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

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

- **Lower bounds for the last-iterate (Golowich et al., 2020):**
  - $\text{Gap}_F(x^K) = \Omega(1/\sqrt{K})$
  - $\|F(x^K)\|^2 = \Omega(1/k)$
When $F$ is monotone and $L$-Lipschitz the following results are known for EG:

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

- **Lower bounds for the last-iterate (Golowich et al., 2020):**
  - $\text{Gap}_F(x^K) = \Omega(1/\sqrt{K})$
  - $\|F(x^K)\|^2 = \Omega(1/k)$

- **Upper bounds for the last-iterate (Golowich et al., 2020):** if additionally the Jacobian $\nabla F(x)$ is $\Lambda$-Lipschitz, then
When $F$ is monotone and $L$-Lipschitz the following results are known for EG:

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

- **Lower bounds for the last-iterate (Golowich et al., 2020):**
  - $\text{Gap}_F(x^K) = \Omega(1/\sqrt{k})$
  - $\|F(x^K)\|^2 = \Omega(1/k)$

- **Upper bounds for the last-iterate (Golowich et al., 2020): if additionally the Jacobian $\nabla F(x)$ is $\Lambda$-Lipschitz, then**
  - $\text{Gap}_F(x^K) = \mathcal{O}(1/\sqrt{k})$
  - $\|F(x^K)\|^2 = \mathcal{O}(1/k)$
Q1: Is it possible to prove last-iterate $\|F(x^K)\|^2 = O(1/K)$ convergence rate for EG when $F$ is monotone and $L$-Lipschitz without additional assumptions?

We address this question with the help of a computer.
Performance Estimation
Problems and Last-Iterate
Convergence of Extragradient
• A powerful technique for deriving tight convergence guarantees, obtaining proofs and even designing new optimal methods
• First work: (Drori and Teboulle, 2014)
• Some later works: (Kim and Fessler, 2016; Lessard et al., 2016; Taylor et al., 2017a,b; De Klerk et al., 2017; Ryu et al., 2020; Taylor and Bach, 2019)
PEP for method $\mathcal{M}$ applied to solve a problem $p$ from some class $\mathcal{P}$:

$$\max \; \text{Convergence Criterion}(x^K)$$

s.t. $p \in \mathcal{P}$, $x^0 \in \mathbb{R}^d$, $\text{Initial Conditions}(x^0)$,

$x^K$ is an output of method $\mathcal{M}$ after $K$ iterations
Example: PEP for the Last-Iterate of EG

\[ \begin{align*}
\max & \quad \| F(x^K) \|^2 \\
\text{s.t.} & \quad F \text{ is monotone and } L\text{-Lipschitz, } x^0 \in \mathbb{R}^d, \\
& \quad \| x^0 - x^* \|^2 \leq 1, \\
& \quad x^{k+1} = x^k - \gamma_2 F(x^k - \gamma_1 F(x^k)) , \quad k = 0, 1, \ldots, K - 1
\end{align*} \]
Another Example for \(\text{EG}\)

- Another example of what we could solve:
  - Check whether \(\|F(x^{k+1})\|^2 \leq \|F(x^k)\|^2\)

- Associated PEP problem:

\[
\Delta_{\text{EG}}(L, \gamma) = \max \|F(x^{k+1})\|^2 - \|F(x^k)\|^2 \\
\text{s.t.} \quad F \text{ is monotone and } L\text{-Lipschitz, } x^k \in \mathbb{R}^d, \\
x^{k+1/2} = x^k - \gamma F(x^k) \\
x^{k+1} = x^k - \gamma F(x^{k+1/2})
\] (7)

- Problems (7) and (6) are hard to solve since they are infinitely dimensional
Another Example for EG

• Another example of what we could solve:
  • Check whether $\|F(x^{k+1})\|^2 \leq \|F(x^k)\|^2$

• Associated PEP problem:

$$\Delta_{EG}(L, \gamma) = \max \|F(x^{k+1})\|^2 - \|F(x^k)\|^2$$

s.t. $F$ is monotone and $L$-Lipschitz, $x^k \in \mathbb{R}^d$,

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

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

• Problems (7) and (6) are hard to solve since they are infinitely dimensional
• Another example of what we could solve:
  • Check whether $\|F(x^{k+1})\|^2 \leq \|F(x^k)\|^2$

• Associated PEP problem:

$$\Delta_{EG}(L, \gamma) = \max \quad \|F(x^{k+1})\|^2 - \|F(x^k)\|^2 \quad (7)$$

$$\text{s.t.} \quad F \text{ is monotone and } L\text{-Lipschitz}, \ x^k \in \mathbb{R}^d,$$
$$x^{k+1/2} = x^k - \gamma F(x^k)$$
$$x^{k+1} = x^k - \gamma F(x^{k+1/2})$$

• Problems (7) and (6) are hard to solve since they are infinitely dimensional

• **Key idea:** replace the initial problem by an “easier” problem.

• The quantities ”mattering” are $x^k, x^{k+\frac{1}{2}}, x^{k+1}, F(x^k), F(x^{k+\frac{1}{2}})$ and $F(x^{k+1})$. 
• Another example of what we could solve:
  • Check whether $\|F(x^{k+1})\|^2 \leq \|F(x^k)\|^2$

• Associated PEP problem:

$$\Delta_{EG}(L, \gamma) = \max \|F(x^{k+1})\|^2 - \|F(x^k)\|^2$$

s.t.  $F$ is monotone and $L$-Lipschitz, $x^k \in \mathbb{R}^d,$

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

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

• Problems (7) and (6) are hard to solve since they are infinitely dimensional

• **Key idea:** replace the initial problem by an “easier” problem.

• The quantities ”mattering” are $x^k, x^{k+1/2}, x^{k+1}, F(x^k), F(x^{k+1/2})$ and $F(x^{k+1}).$

• **Key point:** consider monotonicity and Lipchitzness at these points
Finite-Dimensional Relaxation

\[
\max \|F_{k+1}\|^2 - \|F_k\|^2
\]

s.t. \(d\) and \(x^k, F_k, F_{k+1}, F_{k+\frac{1}{2}} \in \mathbb{R}^d\),

\[
x^{k+\frac{1}{2}} = x^k - \gamma F_k,
\]

\[
x^{k+1} = x^k - \gamma F_{k+\frac{1}{2}},
\]

(8) (extrapolation step) (update step)
Finite-Dimensional Relaxation

\[ \begin{align*}
\max & \quad \| F_{k+1} \|^2 - \| F_k \|^2 \\
\text{s.t.} & \quad d \text{ and } x^k, F_k, F_{k+1}, F_{k+\frac{1}{2}} \in \mathbb{R}^d, \\
& \quad x^{k+\frac{1}{2}} = x^k - \gamma F_k, \\
& \quad x^{k+1} = x^k - \gamma F_{k+\frac{1}{2}}, \\
\lambda_1 : & \quad 0 \leq \langle F_k - F_{k+\frac{1}{2}}, x^k - x^{k+\frac{1}{2}} \rangle, \\
\lambda_2 : & \quad 0 \leq \langle F_k - F_{k+1}, x^k - x^{k+1} \rangle, \\
\lambda_3 : & \quad 0 \leq \langle F_{k+1} - F_{k+\frac{1}{2}}, x^k - x^{k+1/2} \rangle,
\end{align*} \]

(8)

(extrapolation step)

(update step)

(monotonicity in \((x^k, x^{k+\frac{1}{2}}))\)

(monotonicity in \((x^k, x^{k+1})\))

(monotonicity in \((x^{k+1}, x^{k+\frac{1}{2}}))\)
Finite-Dimensional Relaxation

\[
\begin{align*}
\text{max} & \quad \|F_{k+1}\|^2 - \|F_k\|^2 \\
\text{s.t.} & \quad d \text{ and } x^k, F_k, F_{k+1}, F_{k+\frac{1}{2}} \in \mathbb{R}^d, \\
\quad & x^{k+\frac{1}{2}} = x^k - \gamma F_k, \\
\quad & x^{k+1} = x^k - \gamma F_{k+\frac{1}{2}}, \\
\lambda_1 & : 0 \leq \langle F_k - F_{k+\frac{1}{2}}, x^k - x^{k+\frac{1}{2}} \rangle, \\
\lambda_2 & : 0 \leq \langle F_k - F_{k+1}, x^k - x^{k+1} \rangle, \\
\lambda_3 & : 0 \leq \langle F_{k+1} - F_{k+\frac{1}{2}}, x^k - x^{k+1/2} \rangle, \\
\lambda_4 & : \|F_k - F_{k+\frac{1}{2}}\|^2 \leq L^2 \gamma^2 \|x^k - x^{k+\frac{1}{2}}\|^2, \\
\lambda_5 & : \|F_k - F_{k+1}\|^2 \leq L^2 \gamma^2 \|x^k - x^{k+1}\|^2, \\
\lambda_6 & : \|F_{k+1} - F_{k+\frac{1}{2}}\|^2 \leq L^2 \gamma^2 \|x^k - x^{k+\frac{1}{2}}\|^2.
\end{align*}
\]
Problem (8) is not equivalent to (7)

- There might exist a solution of (8) such that no monotone Lipschitz operator $F$ can interpolate it (Ryu et al., 2020)
- In general, for the class of monotone Lipschitz operators interpolation conditions are unknown
✗ Problem (8) is not equivalent to (7)
  • There might exist a solution of (8) such that no monotone Lipschitz operator $F$ can interpolate it (Ryu et al., 2020)
  • In general, for the class of monotone Lipschitz operators interpolation conditions are unknown
✓ But we can still solve (8) numerically
Towards SDP Formulation

• The unknown parameters are \((x^k, x^{k+\frac{1}{2}}, x^{k+1}, F_k, F_{k+\frac{1}{2}}, F_{k+1})\).

• Consider the Gram matrix of these vectors:

\[
G = \begin{pmatrix}
(x^k)^T \\
(x^{k+\frac{1}{2}})^T \\
(x^{k+1})^T \\
(F_k)^T \\
(F_{k+\frac{1}{2}})^T \\
(F_{k+1})^T
\end{pmatrix} \cdot \begin{pmatrix}
x^k & x^{k+\frac{1}{2}} & x^{k+1} & F_k & F_{k+\frac{1}{2}} & F_{k+1}
\end{pmatrix}
\]
Towards SDP Formulation

- The unknown parameters are \((x^k, x^{k+\frac{1}{2}}, x^{k+1}, F_k, F_{k+\frac{1}{2}}, F_{k+1})\).
- Consider the Gram matrix of these vectors:

\[
G = \begin{pmatrix}
(x^k)^\top \\
(x^{k+\frac{1}{2}})^\top \\
(x^{k+1})^\top \\
(F_k)^\top \\
(F_{k+\frac{1}{2}})^\top \\
(F_{k+1})^\top 
\end{pmatrix} \cdot \begin{pmatrix}
x^k & x^{k+\frac{1}{2}} & x^{k+1} & F_k & F_{k+\frac{1}{2}} & F_{k+1}
\end{pmatrix}
\]

- One can easily show that for all \(d \geq 4\)

\[
G \in S_+^6 \iff \exists x^k, x^{k+\frac{1}{2}}, x^{k+1}, F_k, F_{k+\frac{1}{2}}, F_{k+1} \in \mathbb{R}^d : G \text{ is Gram matrix.}
\]
• Therefore, problem (8) is equivalent to the following SDP problem:

\[
\begin{align*}
\text{max} & \quad \operatorname{Tr}(M_0G) \\
\text{s.t.} & \quad G \in S^4_+, \\
& \quad \operatorname{Tr}(M_iG) \leq 0, \ i = 1, 2, \ldots, 6,
\end{align*}
\]

where $M_0, \ldots, M_6$ are some symmetric matrices.
Therefore, problem (8) is equivalent to the following SDP problem:

\[
\begin{align*}
\text{max} & \quad \text{Tr}(M_0G) \\
\text{s.t.} & \quad G \in \mathbb{S}^4_+,
\end{align*}
\]

\[\text{Tr}(M_iG) \leq 0, \quad i = 1, 2, \ldots, 6,\]

where \(M_0, \ldots, M_6\) are some symmetric matrices.

In that case, the dual problem is very simple:

Find \(\lambda_1, \ldots, \lambda_6 \geq 0\) such that

\[\sum_{i=1}^{6} \lambda_i M_i \succeq M_0\]
Therefore, problem (8) is equivalent to the following SDP problem:

\[
\begin{align*}
\text{max} & \quad \text{Tr}(M_0G) \\
\text{s.t.} & \quad G \in \mathbb{S}^4_+, \\
& \quad \text{Tr}(M_iG) \leq 0, \ i = 1, 2, \ldots, 6,
\end{align*}
\]

where \(M_0, \ldots, M_6\) are some symmetric matrices.

In that case, the dual problem is very simple:

Find \(\lambda_1, \ldots, \lambda_6 \geq 0\) such that

\[
\sum_{i=1}^{6} \lambda_i M_i \succeq M_0
\]

If we solve the dual, we get "a proof":

\[
0 \geq \lambda_i \text{Tr}(M_iG) \geq \text{Tr}(M_0G)
\]
In the numerical tests, we observed that $\Delta_{EG}(L, \gamma) \approx 0$ for all tested pairs of $L$ and $\gamma$.
Analysis of the Solution

- In the numerical tests, we observed that $\Delta_{EG}(L, \gamma) \approx 0$ for all tested pairs of $L$ and $\gamma$
- Moreover, we have

$$\lambda_1 \approx \frac{1}{2}, \quad \lambda_2 \approx 2, \quad \lambda_3 = 0, \quad \lambda_4 = 0, \quad \lambda_5 = 0, \quad \lambda_6 \approx \frac{3}{2},$$
Analysis of the Solution

- In the numerical tests, we observed that $\Delta_{EG}(L, \gamma) \approx 0$ for all tested pairs of $L$ and $\gamma$
- Moreover, we have
  \[
  \lambda_1 \approx \frac{1}{2}, \quad \lambda_2 \approx 2, \quad \lambda_3 = 0, \quad \lambda_4 = 0, \quad \lambda_5 = 0, \quad \lambda_6 \approx \frac{3}{2},
  \]
- Duality of SDPs says that
  \[
  \|F(x^k)\|^2 - \|F(x^{k+1})\|^2 \leq \lambda_6 (\|F(x^{k+1}) - F(x^{k+\frac{1}{2}})\|^2 - L^2 \gamma^2 \|x^k - x^{k+\frac{1}{2}}\|^2)
  
  - \lambda_1 \langle F(x^k) - F(x^{k+\frac{1}{2}}), x^k - x^{k+\frac{1}{2}} \rangle
  
  - \lambda_2 \langle F(x^k) - F(x^{k+1}), x^k - x^{k+1} \rangle
  
  \leq 0
  \]
Theorem 1
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 $\|F(x^{k+1})\| \leq \|F(x^k)\|$. 
**Theorem 1**

Let $F: \mathbb{R}^d \to \mathbb{R}^d$ be monotone and $L$-Lipschitz, $0 < \gamma \leq \frac{1}{\sqrt{2}L}$. 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 $O(1/k)$ rate.

**Theorem 2**

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

$$\|F(x^K)\|^2 \leq \frac{\|x^0 - x^*\|^2}{\gamma^2(1 - L^2\gamma^2)(K + 1)}, \tag{12}$$

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

$$\text{Gap}_F(x^K) = \max_{y \in \mathbb{R}^d: \|y - x^*\| \leq \|x^0 - x^*\|} \langle F(y), x^K - y \rangle \leq \frac{2\|x^0 - x^*\|^2}{\gamma \sqrt{1 - L^2\gamma^2} \sqrt{K + 1}}.$$
Last-Iterate Convergence of Optimistic Gradient
Going Back to the Constrained Setting

- **Problem:**

  \[
  \text{find } x^* \in Q \subseteq \mathbb{R}^d \text{ such that } \langle F(x^*), x - x^* \rangle \geq 0, \quad \forall x \in Q
  \]  
  \[
  \text{(VIP-C)}
  \]
Going Back to the Constrained Setting

- Problem:

\[
\text{find } x^* \in Q \subseteq \mathbb{R}^d \text{ such that } \langle F(x^*), x - x^* \rangle \geq 0, \forall x \in Q
\]

(VIP-C)

- Projected Extragradient:

\[
\tilde{x}^k = \text{proj}[x^k - \gamma F(x^k)], \quad x^{k+1} = \text{proj}[x^k - \gamma F(\tilde{x}^k)] \quad \text{(Proj-EG)}
\]

- Projected Past Extragradient:

\[
\tilde{x}^k = \text{proj}[x^k - \gamma F(\tilde{x}^{k-1})], \quad x^{k+1} = \text{proj}[x^k - \gamma F(\tilde{x}^k)] \quad \text{(Proj-PEG)}
\]
Going Back to the Constrained Setting

- Problem:
  \[ \text{find } x^* \in Q \subseteq \mathbb{R}^d \text{ such that } \langle F(x^*), x - x^* \rangle \geq 0, \forall x \in Q \]  
  (VIP-C)

- Projected Extragradient:
  \[ \tilde{x}^k = \text{proj}[x^k - \gamma F(x^k)], \quad x^{k+1} = \text{proj}[x^k - \gamma F(\tilde{x}^k)] \]  
  (Proj-EG)

- Projected Past Extragradient:
  \[ \tilde{x}^k = \text{proj}[x^k - \gamma F(\tilde{x}^{k-1})], \quad x^{k+1} = \text{proj}[x^k - \gamma F(\tilde{x}^k)] \]  
  (Proj-PEG)

- Convergence metric: \[ \|x^{k+1} - x^k\|^2. \]
In the unconstrained case, PEG and OG are equivalent

- Past Extragradient (PEG)
  \[
  \tilde{x}^k = x^k - \gamma F(\tilde{x}^{k-1}), \quad x^{k+1} = x^k - \gamma F(\tilde{x}^k)
  \]

- Optimistic Gradient method (OG)
  \[
  \tilde{x}^{k+1} = \tilde{x}^k - 2\gamma F(\tilde{x}^k) + \gamma F(\tilde{x}^{k-1})
  \]
\[ G_{\text{PEG}}(\gamma, L, N) = \max_{F,d,x^*} \frac{\| F(x^N) \|^2}{\| x^0 - x^* \|^2} \] (13)

s.t. \( F \) is monotone and \( L \)-Lipschitz,
\[ \tilde{x}^0 = x^0 \in \mathbb{R}^d, \quad x^1 = x^0 - \gamma F(x^0) \]
\[ \tilde{x}^k = x^k - \gamma F(\tilde{x}^{k-1}), \text{ for } k = 1, \ldots, N, \]
\[ x^{k+1} = x^k - \gamma F(\tilde{x}^k), \text{ for } k = 1, \ldots, N - 1. \]
PEP for PEG: Relaxation

\[ \tilde{G}_{\text{PEG}}(\gamma, L, N) = \max_{d \in \mathbb{N}, x^* \in \mathbb{R}^d} \|g^N\|^2 \]

\begin{align*}
\text{s.t. } & \langle g - h, x - y \rangle \geq 0 \quad \forall (x, g), (y, h) \in S \\
& \|g - h\|^2 \leq L^2 \|x - y\|^2 \quad \forall (x, g), (y, h) \in S \\
& \tilde{x}^0 = x^0 \in \mathbb{R}^d, x^1 = x^0 - \gamma g^0 \\
& \tilde{x}^k = x^k - \gamma \tilde{g}^{k-1}, \text{ for } k = 1, \ldots, N, \\
& x^{k+1} = x^k - \gamma \tilde{g}^k, \text{ for } k = 1, \ldots, N - 1, \\
& \|x^0 - x^*\|^2 \leq 1
\end{align*}

- Following the same steps as in the previous examples, one can reformulate this problem as SDP
\[ G_{OG}(\gamma, L, N) = \max_{F, d, x^*} \frac{\|F(\tilde{x}^N)\|^2}{\|\tilde{x}^0 - x^*\|^2} \]  

\[ \text{s.t. } F \text{ is monotone and } L\text{-Lipschitz,} \]
\[ \tilde{x}^0 \in \mathbb{R}^d, \quad \tilde{x}^1 = \tilde{x}^0 - \gamma F(\tilde{x}^0), \]
\[ \tilde{x}^{k+1} = \tilde{x}^k - 2\gamma F(\tilde{x}^k) + \gamma F(\tilde{x}^{k-1}), \]
\[ \text{for } k = 1, \ldots, N - 1, \]
PEP for 0G: Relaxation

\[ \tilde{G}_{0G}(\gamma, L, N) = \max_{d \in \mathbb{N}, x^* \in \mathbb{R}^d} \| g^N \|^2 \]

\[ \{(\tilde{x}^k, \tilde{g}^k)\}_{k=0}^N \subseteq \mathbb{R}^d \times \mathbb{R}^d \]

s.t. \( \langle g - h, x - y \rangle \geq 0 \quad \forall (x, g), (y, h) \in S \) (20)

\[ \| g - h \|^2 \leq L^2 \| x - y \|^2 \quad \forall (x, g), (y, h) \in S \] (21)

\( \tilde{x}^0 \in \mathbb{R}^d, \quad \tilde{x}^1 = \tilde{x}^0 - \gamma \tilde{g}^0, \)

\( \tilde{x}^{k+1} = \tilde{x}^k - 2\gamma \tilde{g}^k + \gamma \tilde{g}^{k-1}, \)

for \( k = 1, \ldots, N - 1, \)

\[ \| x^0 - x^* \|^2 \leq 1 \] (22)

- Following the same steps as in the previous examples, one can reformulate this problem as SDP
Both Relaxations Show $\mathcal{O}(1/N)$ Convergence

Figure 2: $\tilde{G}_{\text{PEG}}(\gamma, L, N)$ and $\tilde{G}_{\text{OG}}(\gamma, L, N)$ for different values of $\gamma$ and $N$
PEP for PEG and OG: Existence of Simple Proofs?

- Typical proof – some clever combination of inequalities

We can explicitly drop constraints for the iterates with indices $i, j$ such that $|i - j| \leq t$. We denote the corresponding problems as $e_{PEG}(\gamma, L, N, t)$ and $e_{OG}(\gamma, L, N, t)$. 

30
• Typical proof – some clever combination of inequalities
• Those inequalities typically involve consecutive iterates and/or $x^*$
Typical proof – some clever combination of inequalities

Those inequalities typically involve consecutive iterates and/or $x^*$

We can explicitly drop constraints for the iterates with indices $i, j$ such that $|i - j| \leq t$

- We denote the corresponding problems as $\tilde{G}_{\text{PEG}}(\gamma, L, N, t)$ and $\tilde{G}_{\text{OG}}(\gamma, L, N, t)$
Figure 3: We report $\tilde{G}_{\text{PEG}}(\gamma, L, N, 1)$ and $\tilde{G}_{\text{OG}}(\gamma, L, N, t)$ for $t = 1, 2, 4$. It suggest that $\tilde{G}_{\text{PEG}}(\gamma, L, N, 1) \sim \frac{1}{N}$ but not $\tilde{G}_{\text{OG}}(\gamma, L, N, t)$ (even for $t = 4$).
Figure 3: We report $\tilde{G}_{\text{PEG}}(\gamma, L, N, 1)$ and $\tilde{G}_{\text{OG}}(\gamma, L, N, t)$ for $t = 1, 2, 4$. It suggest that $\tilde{G}_{\text{PEG}}(\gamma, L, N, 1) \sim \frac{1}{N}$ but not $\tilde{G}_{\text{OG}}(\gamma, L, N, t)$ (even for $t = 4$).

Since interpolation is not guaranteed, extra points are crucial!
Can We Prove that the Norm Monotonically Decrease?

\[ \Delta(\gamma, L, N) = \max_{F, d, x^*} \frac{\| F(x^{N+1}) \|^2 - \| F(x^N) \|^2}{\| x^0 - x^* \|^2} \]  

(23)

s.t. \( F \) is monotone and \( L \)-Lipschitz,

\[ \tilde{x}^0 = x^0 \in \mathbb{R}^d, \ x^1 = x^0 - \gamma F(x^0), \]

\[ \tilde{x}^k = x^k - \gamma F(\tilde{x}^{k-1}), \text{ for } k = 1, \ldots, N, \]

\[ x^{k+1} = x^k - \gamma F(\tilde{x}^k), \text{ for } k = 1, \ldots, N - 1, \]
Can We Prove that the Norm Monotonically Decrease?

\[ \Delta(\gamma, L, N) = \max_{F, d, x^*} \frac{\| F(x^{N+1}) \|^2 - \| F(x^N) \|^2}{\| x^0 - x^* \|^2} \]  \hspace{1cm} (23)

\[
\text{s.t. } F \text{ is monotone and } L\text{-Lipschitz,}
\]

\[ \tilde{x}^0 = x^0 \in \mathbb{R}^d, \ x^1 = x^0 - \gamma F(x^0), \]

\[ \tilde{x}^k = x^k - \gamma F(\tilde{x}^{k-1}), \text{ for } k = 1, \ldots, N, \]

\[ x^{k+1} = x^k - \gamma F(\tilde{x}^k), \text{ for } k = 1, \ldots, N - 1, \]

- \( \tilde{\Delta}(\gamma, L, N) \) – value of SDP relaxation
Can We Prove that the Norm Monotonically Decrease?

\[ \Delta(\gamma, L, N) = \max_{F, d, x^*} \frac{\| F(x^{N+1}) \|^2 - \| F(x^N) \|^2}{\| x^0 - x^* \|^2} \]

\[ \| x_0 - \cdots - x_N \| \]

s.t. \quad \begin{align*}
F \text{ is monotone and } L\text{-Lipschitz}, \\
\tilde{x}^0 &= x^0 \in \mathbb{R}^d, \quad x^1 = x^0 - \gamma F(x^0), \\
\tilde{x}^k &= x^k - \gamma F(\tilde{x}^{k-1}), \text{ for } k = 1, \ldots, N, \\
x^{k+1} &= x^k - \gamma F(\tilde{x}^k), \text{ for } k = 1, \ldots, N - 1,
\end{align*} \]

\[ \Delta(\gamma, L, N) - \text{value of SDP relaxation} \]

- We also consider another version of (23) for \( L \)-cocoercive operator \( F \) (i.e., \( \langle g - h, x - y \rangle \geq 0 \) and \( \| g - h \|^2 \leq L^2 \| x - y \|^2 \) are replaced by \( \| g - h \|^2 \leq L \langle g - h, x - y \rangle \)), the corresponding values are denoted as \( \delta(\gamma, L, N) \) and \( \tilde{\delta}(\gamma, L, N) \)
- Guaranteed interpolation (Ryu et al., 2020): \( \delta(\gamma, L, N) = \tilde{\delta}(\gamma, L, N) \)
Can We Prove that the Norm Monotonically Decrease? No

Figure 4: Evolution of $\Delta(\gamma, L, N)$, $\tilde{\Delta}(\gamma, L, N)$, $\delta(\gamma, L, N)$, $\tilde{\delta}(\gamma, L, N)$

Need to find other potentials to prove the last-iterate convergence.
Can We Prove that the Norm Monotonically Decrease? No

Figure 4: Evolution of $\Delta(\gamma, L, N), \tilde{\Delta}(\gamma, L, N), \delta(\gamma, L, N), \tilde{\delta}(\gamma, L, N)$

Need to find other potentials to prove the last-iterate convergence
Some Intuition from the Numerical Results

- Inequality $\|F(x^N)\|^2 \leq \|F(x^{N-1})\|^2$ does not hold for PEG...
Some Intuition from the Numerical Results

- Inequality $\| F(x^N) \|^2 \leq \| F(x^{N-1}) \|^2$ does not hold for PEG...
- ... but we see that $\| F(x^N) \|^2 - \| F(x^{N-1}) \|^2$ decreases
Some Intuition from the Numerical Results

- Inequality $\|F(x^N)\|^2 \leq \|F(x^{N-1})\|^2$ does not hold for PEG...
- ... but we see that $\|F(x^N)\|^2 - \|F(x^{N-1})\|^2$ decreases
- Idea: try to find such sequence $\{A_N\}_{N \geq 0}$ that
  \[ \|F(x^N)\|^2 + A_N \leq \|F(x^{N-1})\|^2 + A_{N-1} \]
Some Intuition from the Numerical Results

- Inequality $\|F(x^N)\|^2 \leq \|F(x^{N-1})\|^2$ does not hold for PEG...
- ... but we see that $\|F(x^N)\|^2 - \|F(x^{N-1})\|^2$ decreases
- Idea: try to find such sequence $\{A_N\}_{N \geq 0}$ that
  $\|F(x^N)\|^2 + A_N \leq \|F(x^{N-1})\|^2 + A_{N-1}$
- After several (educated) guess and trials we found numerically:

  $\|F(x^{N+1})\|^2 + 2\|F(x^{N+1}) - F(\tilde{x}^N)\|^2 \leq \|F(x^N)\|^2 + 2\|F(x^N) - F(\tilde{x}^{N-1})\|^2$
Potential for PEG: Unconstrained Problems

Theorem 3
Let $F : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be monotone and $L$-Lipschitz, $0 < \gamma$. Then for all $k \geq 0$ the iterates produced by PEG satisfy

$$\|F(x^{k+1})\|^2 + 2\|F(x^{k+1}) - F(\bar{x}^k)\|^2 \leq \|F(x^k)\|^2 + 2\|F(x^k) - F(\bar{x}^{k-1})\|^2$$

$$+ 3 \left( L^2 \gamma^2 - \frac{2}{9} \right) \|F(\bar{x}^k) - F(\bar{x}^{k-1})\|^2.$$

The last term is non-positive for $0 < \gamma \leq \sqrt{2}/3L$. 
Theorem 4

Let \( F : \mathbb{R}^d \to \mathbb{R}^d \) be monotone and \( L \)-Lipschitz, \( 0 < \gamma \). Then for all \( k \geq 0 \) the iterates produced by PEG with \( \gamma \leq 1/3L \) satisfy \( \Phi_{k+1} \leq \Phi_k \) with \( \Phi_k \) defined as

\[
\Phi_k = \| x^k - x^* \|^2 + \frac{k + 32}{3} \gamma^2 \left( \| F(x^k) \|^2 + 2 \| F(x^k) - F(\tilde{x}^{k-1}) \|^2 \right). \tag{24}
\]
Theorem 4

Let \( F : \mathbb{R}^d \to \mathbb{R}^d \) be monotone and \( L \)-Lipschitz, \( 0 < \gamma \). Then for all \( k \geq 0 \) the iterates produced by PEG with \( \gamma \leq \frac{1}{3L} \) satisfy \( \Phi_{k+1} \leq \Phi_k \) with \( \Phi_k \) defined as

\[
\Phi_k = \|x^k - x^*\|^2 + \frac{k + 32}{3} \gamma^2 \left( \|F(x^k)\|^2 + 2\|F(x^k) - F(\tilde{x}^{k-1})\|^2 \right).
\]

(24)

In particular, for all \( N \geq 0 \) and \( \gamma \leq \frac{1}{3L} \) the above formula implies

\[
\|F(x^N)\|^2 \leq \frac{3(1 + 32L^2\gamma^2)\|x^0 - x^*\|^2}{\gamma^2(N + 32)},
\]

\[
\text{Gap}_F(x^k) \leq \frac{2\sqrt{41}(1 + 32L^2\gamma^2)\|x^0 - x^*\|^2}{\gamma\sqrt{3N + 96}}.
\]
• Different metric: instead of $\|F(x^N)\|^2$ we consider $\|x^N - x^{N-1}\|^2$
From Unconstrained to Constrained Problems

- Different metric: instead of $\| F(x^N) \|^2$ we consider $\| x^N - x^{N-1} \|^2$
- It is non-trivial to directly extend previous potential to the constrained case

We need more help from the computer
Guided by the approach from Taylor and Bach (2019) of computer-aided search of the potentials for the methods applied to stochastic minimization problems, we consider the following problem

\[
\text{find } F, d, x^*, \tilde{x}^0, \ldots, \tilde{x}^N, x^0, \ldots, x^N, \Phi_N, \Phi_{N-1} \\
\text{s.t. } F \text{ is monotone and } L\text{-Lipschitz,} \\
\Phi_N \text{ and } \Phi_{N-1} \text{ are quadratic w.r.t. iterates and operator values,} \\
\Phi_N \text{ and } \Phi_{N-1} \text{ have same structure,} \\
\|x^N - x^{N-1}\|^2 \leq \Phi_N, \Phi_N - \Phi_{N-1} \leq 0, \\
\tilde{x}^0 = x^0 \in \mathbb{R}^d, x^1 = x^0 - \gamma F(x^0), \\
\tilde{x}^k = x^k - \gamma F(\tilde{x}^{k-1}), \text{ for } k = 1, \ldots, N, \\
x^{k+1} = x^k - \gamma F(\tilde{x}^k), \text{ for } k = 1, \ldots, N - 1
\]
Guided by the approach from Taylor and Bach (2019) of computer-aided search of the potentials for the methods applied to stochastic minimization problems, we consider the following problem

\[
\text{find } F, d, x^*, \tilde{x}^0, \ldots, \tilde{x}^N, x^0, \ldots, x^N, \Phi_N, \Phi_{N-1} \\
\text{s.t. } F \text{ is monotone and } L\text{-Lipschitz,}
\]
\[
\Phi_N \text{ and } \Phi_{N-1} \text{ are quadratic w.r.t. iterates and operator values,}
\]
\[
\Phi_N \text{ and } \Phi_{N-1} \text{ have same structure,}
\]
\[
\|x^N - x^{N-1}\|^2 \leq \Phi_N, \; \Phi_N - \Phi_{N-1} \leq 0,
\]
\[
\tilde{x}^0 = x^0 \in \mathbb{R}^d, \; x^1 = x^0 - \gamma F(x^0),
\]
\[
\tilde{x}^k = x^k - \gamma F(\tilde{x}^{k-1}), \; \text{for } k = 1, \ldots, N,
\]
\[
x^{k+1} = x^k - \gamma F(\tilde{x}^k), \; \text{for } k = 1, \ldots, N - 1
\]

The dual SDP relaxation can be efficiently solved!
Solving the corresponding dual SDP relaxation and imposing additional constraints (to make potential and proof easier), we found the following potential (for $\gamma \leq \frac{1}{\sqrt{5L}}$).

**Theorem 5**

Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be monotone and $L$-Lipschitz, $0 < \gamma$. Then for all $k \geq 0$ the iterates produced by PEG satisfy

$$\Psi_{k+1} \leq \Psi_k - (1 - 5L^2\gamma^2) \|x^{k+1} - \tilde{x}^k\|^2 - \gamma^2 \|F(x^{k+1}) - F(\tilde{x}^k)\|^2, \quad (26)$$

where

$$\Psi_k = \|x^k - x^{k-1}\|^2 + \|x^k - x^{k-1} - 2\gamma(F(x^k) - F(\tilde{x}^{k-1}))\|^2.$$
Theorem 6

Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be monotone and $L$-Lipschitz, $0 < \gamma$. Then for all $k \geq 0$ the iterates produced by PEG with $\gamma \leq 1/4L$ satisfy $\Phi_{k+1} \leq \Phi_k$ with $\Phi_k$ defined as

$$\Phi_k = \|x^k - x^*\|^2 + \frac{1}{16}\|\tilde{x}^{k-1} - \tilde{x}^{k-2}\|^2 + \frac{3k + 32}{24}\psi_k, \quad (27)$$

where $\psi_k = \|x^k - x^{k-1}\|^2 + \|x^k - x^{k-1} - 2\gamma(F(x^k) - F(\tilde{x}^{k-1}))\|^2$. 

40
Theorem 6
Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be monotone and $L$-Lipschitz, $0 < \gamma$. Then for all $k \geq 0$ the iterates produced by PEG with $\gamma \leq \frac{1}{4L}$ satisfy $\Phi_{k+1} \leq \Phi_k$ with $\Phi_k$ defined as

$$
\Phi_k = \|x^k - x^*\|^2 + \frac{1}{16}\|\tilde{x}^{k-1} - \tilde{x}^{k-2}\|^2 + \frac{3k + 32}{24}\Psi_k,
$$

where $\Psi_k = \|x^k - x^{k-1}\|^2 + \|x^k - x^{k-1} - 2\gamma(F(x^k) - F(\tilde{x}^{k-1}))\|^2$. In particular, for all $N \geq 0$ and $\gamma \leq \frac{1}{4L}$ the above formula implies

$$
\|x^N - x^{N-1}\|^2 \leq \frac{24H_{0, \gamma}^2}{3N+32}, \quad \text{Gap}_F(x^N) \leq \frac{8\sqrt{3}H_{0, \gamma} \cdot H_0}{\gamma \sqrt{3N+32}}, \quad \forall N \geq 2,
$$

where $H_0, H_{0, \gamma} > 0$ are such that

$$
H_{0, \gamma}^2 = 2(1 + 3\gamma^2 L^2 + 4\gamma^4 L^4)\|x^0 - x^*\|^2 + (\frac{41}{12} + \frac{19}{3}\gamma^2 L^2)\gamma^2\|F(x^0)\|^2,
$$

$$
H_0^2 = 3\|x^0 - x^*\|^2 + \frac{1}{30L^2}\|F(x^0)\|^2.
$$
Last-Iterate Convergence Under Negative-Comonotonicity
find \( x^* \in \mathbb{R}^d \) such that \( F(x^*) = 0 \) \hspace{1cm} (VIP)

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

\[
\|F(x) - F(y)\| \leq L\|x - y\| \tag{28}
\]
Negative Comonotonicity for the Unconstrained Case

find $x^* \in \mathbb{R}^d$ such that $F(x^*) = 0$ \hspace{1cm} (VIP)

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

\[ \|F(x) - F(y)\| \leq L\|x - y\| \hspace{1cm} (28) \]

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

\[ \langle F(x) - F(y), x - y \rangle \geq -\rho\|F(x) - F(y)\| \hspace{1cm} (29) \]
Negative Comonotonicity for the Unconstrained Case

find $x^* \in \mathbb{R}^d$ such that $F(x^*) = 0$ \hspace{1cm} (VIP)

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

  $$\|F(x) - F(y)\| \leq L\|x - y\|$$  \hspace{1cm} (28)

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

  $$\langle F(x) - F(y), x - y \rangle \geq -\rho\|F(x) - F(y)\|$$  \hspace{1cm} (29)

  - $\rho < 0$ – cocoercivity
  - $\rho = 0$ – monotonicity
  - $\rho > 0$ – cohypomonotonicity (Pennanen, 2002)
Theorem 7

If $F : \mathbb{R}^d \to \mathbb{R}^d$ is $L$-Lipschitz and $\rho$-negative comonotone with $\rho \leq \frac{1}{8L}$ and $\gamma_1 = \gamma_2 = \gamma$ such that $4\rho \leq \gamma \leq \frac{1}{2L}$, then for any $k \geq 0$ the iterates produced by EG satisfy

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

(30)
Theorem 7

If $F : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is $L$-Lipschitz and $\rho$-negative comonotone with $\rho \leq \frac{1}{8L}$ and $\gamma_1 = \gamma_2 = \gamma$ such that $4\rho \leq \gamma \leq \frac{1}{2L}$, then for any $k \geq 0$ the iterates produced by EG satisfy

$$\|F(x^{k+1})\| \leq \|F(x^k)\|$$  \hspace{1cm} (30)

and for any $N \geq 1$

$$\|F(x^N)\|^2 \leq \frac{28\|x^0 - x^*\|^2}{N\gamma^2 + 320\gamma\rho}.$$  \hspace{1cm} (31)
Last-Iterate Convergence of \( \text{EG} \)

**Theorem 7**

If \( F : \mathbb{R}^d \to \mathbb{R}^d \) is \( L \)-Lipschitz and \( \rho \)-negative comonotone with \( \rho \leq \frac{1}{8L} \) and \( \gamma_1 = \gamma_2 = \gamma \) such that \( 4\rho \leq \gamma \leq \frac{1}{2L} \), then for any \( k \geq 0 \) the iterates produced by \( \text{EG} \) satisfy

\[
\| F(x^{k+1}) \| \leq \| F(x^k) \| \tag{30}
\]

and for any \( N \geq 1 \)

\[
\| F(x^N) \|^2 \leq \frac{28\| x^0 - x^* \|^2}{N\gamma^2 + 320\gamma\rho}. \tag{31}
\]

✓ Again, we found the potential via computer
Theorem 7

If $F : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is $L$-Lipschitz and $\rho$-negative comonotone with $\rho \leq \frac{1}{8L}$ and $\gamma_1 = \gamma_2 = \gamma$ such that $4\rho \leq \gamma \leq \frac{1}{2L}$, then for any $k \geq 0$ the iterates produced by EG satisfy

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

and for any $N \geq 1$

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

✓ Again, we found the potential via computer

✓ Previous result is derived for $\rho < \frac{1}{16L}$ (Luo and Tran-Dinh, 2022)
**Theorem 7**

If $F : \mathbb{R}^d \to \mathbb{R}^d$ is $L$-Lipschitz and $\rho$-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 $N \geq 1$

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

✓ Again, we found the potential via computer

✓ Previous result is derived for $\rho < 1/16L$ (Luo and Tran-Dinh, 2022)

? **Open question:** is it possible to show $O(1/N)$ last-iterate convergence for EG when $\rho \in (1/8L, 1/2L)$?
Theorem 8

If $F: \mathbb{R}^d \to \mathbb{R}^d$ is $L$-Lipschitz and $\rho$-negative comonotone with $\rho \leq \frac{5}{62}L$ and $\gamma_1 = \gamma_2 = \gamma$ such that $4\rho \leq \gamma \leq \frac{10}{31}L$, then for any $k \geq 0$ the iterates produced by OG satisfy

$$\|F(x^{k+1})\|^2 + \|F(x^{k+1}) - F(\tilde{x}^k)\|^2 \leq \|F(x^k)\|^2 + \|F(x^k) - F(\tilde{x}^{k-1})\|^2$$

$$- \frac{1}{100} \|F(\tilde{x}^k) - F(\tilde{x}^{k-1})\|^2. \quad (32)$$
Theorem 8

If $F : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is $L$-Lipschitz and $\rho$-negative comonotone with $\rho \leq \frac{5}{62}L$ and $\gamma_1 = \gamma_2 = \gamma$ such that $4\rho \leq \gamma \leq \frac{10}{31}L$, then for any $k \geq 0$ the iterates produced by OG satisfy

$$
\|F(x^{k+1})\|^2 + \|F(x^{k+1}) - F(\tilde{x}^k)\|^2 \leq \|F(x^k)\|^2 + \|F(x^k) - F(\tilde{x}^{k-1})\|^2
- \frac{1}{100} \|F(\tilde{x}^k) - F(\tilde{x}^{k-1})\|^2. \quad (32)
$$

and for any $N \geq 1$

$$
\|F(x^N)\|^2 \leq \frac{717\|x^0 - x^*\|^2}{N\gamma(\gamma - 3\rho) + 800\gamma^2}. \quad (33)
$$
Theorem 8

If $F : \mathbb{R}^d \to \mathbb{R}^d$ is $L$-Lipschitz and $\rho$-negative comonotone with $\rho \leq \frac{5}{62}L$ and $\gamma_1 = \gamma_2 = \gamma$ such that $4\rho \leq \gamma \leq \frac{10}{31}L$, then for any $k \geq 0$ the iterates produced by OG satisfy

$$\|F(x^{k+1})\|^2 + \|F(x^{k+1}) - F(\tilde{x}^k)\|^2 \leq \|F(x^k)\|^2 + \|F(x^k) - F(\tilde{x}^{k-1})\|^2$$

$$- \frac{1}{100} \|F(\tilde{x}^k) - F(\tilde{x}^{k-1})\|^2. \quad (32)$$

and for any $N \geq 1$

$$\|F(x^N)\|^2 \leq \frac{717\|x^0 - x^*\|^2}{N\gamma(\gamma - 3\rho) + 800\gamma^2}. \quad (33)$$

✓ Again, we found the potential via computer
Last-Iterate Convergence of $\Omega$G

**Theorem 8**

If $F : \mathbb{R}^d \to \mathbb{R}^d$ is $L$-Lipschitz and $\rho$-negative comonotone with $\rho \leq \frac{5}{62}L$ and $\gamma_1 = \gamma_2 = \gamma$ such that $4\rho \leq \gamma \leq \frac{10}{31}L$, then for any $k \geq 0$ the iterates produced by $\Omega$G satisfy

$$\|F(x^{k+1})\|^2 + \|F(x^{k+1}) - F(\tilde{x}^k)\|^2 \leq \|F(x^k)\|^2 + \|F(x^k) - F(\tilde{x}^{k-1})\|^2 - \frac{1}{100} \|F(\tilde{x}^k) - F(\tilde{x}^{k-1})\|^2.$$ (32)

and for any $N \geq 1$

$$\|F(x^N)\|^2 \leq \frac{717\|x^0 - x^*\|^2}{N\gamma(\gamma - 3\rho) + 800\gamma^2}.$$ (33)

✓ Again, we found the potential via computer
✓ Previous result is derived for $\rho < \frac{8}{27\sqrt{6}}L$ (Luo and Tran-Dinh, 2022)
Theorem 8

If \( F : \mathbb{R}^d \to \mathbb{R}^d \) is \( L \)-Lipschitz and \( \rho \)-negative comonotone with \( \rho \leq \frac{5}{62}L \) and \( \gamma_1 = \gamma_2 = \gamma \) such that \( 4\rho \leq \gamma \leq \frac{10}{31}L \), then for any \( k \geq 0 \) the iterates produced by \( \text{OG} \) satisfy

\[
\| F(x^{k+1}) \|^2 + \| F(x^{k+1}) - F(\tilde{x}^k) \|^2 \leq \| F(x^k) \|^2 + \| F(x^k) - F(\tilde{x}^{k-1}) \|^2 - \frac{1}{100} \| F(\tilde{x}^k) - F(\tilde{x}^{k-1}) \|^2. \tag{32}
\]

and for any \( N \geq 1 \)

\[
\| F(x^N) \|^2 \leq \frac{717\| x^0 - x^* \|^2}{N\gamma(\gamma - 3\rho) + 800\gamma^2}. \tag{33}
\]

✓ Again, we found the potential via computer

✓ Previous result is derived for \( \rho < \frac{8}{27}\sqrt{6}L \) (Luo and Tran-Dinh, 2022)

? Open question: is it possible to show \( O(1/N) \) last-iterate convergence for \( \text{OG} \) when \( \rho \in (\frac{5}{62}L, \frac{1}{2}L) \)?
No Convergence for $\text{EG and OG}$ when $\rho \geq \frac{1}{2L}$

**Theorem 9**
For any $L > 0$, $\rho \geq \frac{1}{2L}$, and any choice of stepsizes $\gamma_1, \gamma_2 > 0$ there exists $\rho$-negative comonotone $L$-Lipschitz operator $F$ such that $\text{EG/OG}$ does not necessary converges on solving (VIP) with this operator $F$. 
No Convergence for $\text{EG and OG when } \rho \geq 1/2L$

**Theorem 9**

For any $L > 0$, $\rho \geq 1/2L$, and any choice of stepsizes $\gamma_1, \gamma_2 > 0$ there exists $\rho$-negative comonotone $L$-Lipschitz operator $F$ such that EG/OG does not necessarily converges on solving (VIP) 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 $x \in \mathbb{R}^2$,

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

- PEP requires to specify numeric values for $\gamma$ and $L$. **Not a formal proof**, requires post-processing.
- PEP does not try to find a "simple proof".
- Can try to remove some of the constraints
  - The problem has more freedom (looser upper bound).
  - Simpler proof (uses lesser inequality).
- For the analysis of $EG$ in the constrained case we refer to Cai et al. (2022).
- More generally, you can "force" the value of any dual constant and see if PEP still works.


References


