Last-Iterate Convergence of Extragradient-Based Methods

EUROPT 2024, Lund

Eduard Gorbunov[1] Adrien Taylor[2] Samuel Horváth[1]
Nicolas Loizou[3] Gauthier Gidel[4,5]

June 26, 2024

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

Outline

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

- E. Gorbunov, N. Loizou, G. Gidel. *Extragradient Method: $O(1/K)$ Last-Iterate Convergence for Monotone Variational Inequalities and Connections With Cocoercivity*. AISTATS 2022
- E. Gorbunov, A. Taylor, G. Gidel. *Last-Iterate Convergence of Optimistic Gradient Method for Monotone Variational Inequalities*. NeurIPS 2022
- E. Gorbunov, A. Taylor, S. Horváth, G. Gidel. *Convergence of Proximal Point and Extragradient-Based Methods Beyond Monotonicity: the Case of Negative Comonotonicity*. ICML 2023

Variational Inequalities and Extragradient-Based Methods

Variational Inequality Problem

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

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

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

- F is monotone: $\forall x, y \in Q$

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

Variational Inequality Problem: Examples

- Min-max problems:

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

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 $Q = U\times V$, $x = (u^\top, v^\top)^\top$, 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).

Variational Inequality Problem: Examples

- Minimization problems:

$$\min_{x \in Q} f(x) \tag{4}$$

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

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

Variational Inequality Problem: Unconstrained Case

When $Q = \mathbb{R}^d$ (VIP-C) can be rewritten as

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

For simplicity, we first consider (VIP) rather than (VIP-C)

How to Solve VIP?

Naïve approach – Gradient Descent (GD):

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

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

✗ GD does not converge for simple convex-concave min-max problems

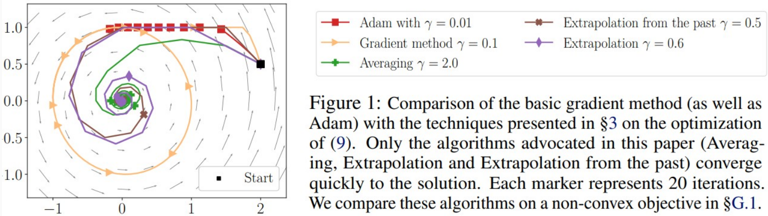

Figure 1: Comparison of the basic gradient method (as well as Adam) with the techniques presented in §3 on the optimization of (9). Only the algorithms advocated in this paper (Averaging, Extrapolation and Extrapolation from the past) converge quickly to the solution. Each marker represents 20 iterations.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

Figure 1: Behavior of GD on the problem $\min_{u \in \mathbb{R}} \max_{v \in \mathbb{R}} uv$ (Gidel et al., 2019)

Popular Alternatives to GD

- Extragradient method (EG) (Korpelevich, 1976)

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

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

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

Measures of Convergence

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

Measures of Convergence

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

Measures of Convergence

- **Restricted gap function**: $\mathtt{Gap}_F(x^K) = \max_{y \in \mathbb{R}^d : \|y - x^*\| \leq R} \langle F(y), x^K - y \rangle$, where $R \sim \|x^0 - x^*\|$ (Nesterov, 2007)
  - ✓ $\mathtt{Gap}_F(x^K)$ can be seen as a natural extension of optimization error for (VIP), when F is monotone
  - ✗ It is unclear how to tightly estimate $\mathtt{Gap}_F(x^K)$ 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 $\mathtt{Gap}_F(x^K)$
  - ✓ $\|F(x^K)\|^2$ is easier to compute than $\mathtt{Gap}_F(x^K)$

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

Convergence Guarantees for EG

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

Convergence Guarantees for EG

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

- **Averaged- and best-iterate guarantees:**
  - $\mathtt{Gap}_F(\overline{x}^K) = \mathcal{O}(1/K)$ for $\overline{x}^K = \frac{1}{K+1}\sum_{k=0}^{K} x^k$

Convergence Guarantees for $\texttt{EG}$

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

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

Convergence Guarantees for EG

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

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

Convergence Guarantees for EG

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

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

Convergence Guarantees for EG

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

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

Convergence Guarantees for EG

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

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

Convergence Guarantees for EG

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

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

Convergence Guarantees for EG

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

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

Convergence Guarantees for EG

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

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

Convergence Guarantees for EG: Resolved Question

Q1: *Is it possible to prove last-iterate* $\|F(x^K)\|^2 = \mathcal{O}(1/K)$ *convergence rate for* EG *when* F *is monotone and* L*-Lipschitz* *without additional assumptions?*

We address this question with the help of a computer

Performance Estimation Problems and Last-Iterate Convergence of Extragradient

Performance Estimation Problems

- 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)

Performance Estimation Problem: A General Form

PEP for method $\mathcal{M}$ applied to solve a problem p from some class $\mathcal{P}$:

$$\begin{aligned} \max \quad & \texttt{Convergence\_Criterion}(x^K) \\ \text{s.t.} \quad & p \in \mathcal{P},\ x^0 \in \mathbb{R}^d, \\ & \texttt{Initial\_Conditions}(x^0), \\ & x^K \text{ is an output of method } \mathcal{M} \text{ after } K \text{ iterations} \end{aligned} \tag{5}$$

Example: PEP for the Last-Iterate of EG

$$
\begin{aligned}
\max \quad & \|F(x^K)\|^2 \\
\text{s.t.} \quad & F \text{ is monotone and } L\text{-Lipschitz},\ x^0 \in \mathbb{R}^d, \\
& \|x^0 - x^*\|^2 \leq 1, \\
& x^{k+1} = x^k - \gamma_2 F\left(x^k - \gamma_1 F(x^k)\right),\ k = 0, 1, \ldots, K-1
\end{aligned} \tag{6}
$$

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:

$$\begin{aligned}
\Delta_{\texttt{EG}}(L, \gamma) = \quad \max \quad & \|F(x^{k+1})\|^2 - \|F(x^k)\|^2 \\
\text{s.t.} \quad & F \text{ is monotone and } L\text{-Lipschitz}, \; x^k \in \mathbb{R}^d, \\
& x^{k+1/2} = x^k - \gamma F(x^k) \\
& x^{k+1} = x^k - \gamma F(x^{k+1/2})
\end{aligned} \tag{7}$$

- 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:

$$\begin{aligned} \Delta_{\texttt{EG}}(L,\gamma) = \quad \max \quad & \|F(x^{k+1})\|^2 - \|F(x^k)\|^2 \\ \text{s.t.} \quad & F \text{ is monotone and } L\text{-Lipschitz}, \; x^k \in \mathbb{R}^d, \\ & x^{k+1/2} = x^k - \gamma F(x^k) \\ & x^{k+1} = x^k - \gamma F(x^{k+1/2}) \end{aligned} \tag{7}$$

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

Another Example for $\texttt{EG}$

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

$$\begin{aligned} \Delta_{\texttt{EG}}(L, \gamma) = \quad \max \quad & \|F(x^{k+1})\|^2 - \|F(x^k)\|^2 \\ \text{s.t.} \quad & F \text{ is monotone and } L\text{-Lipschitz}, \; x^k \in \mathbb{R}^d, \\ & x^{k+1/2} = x^k - \gamma F(x^k) \\ & x^{k+1} = x^k - \gamma F(x^{k+1/2}) \end{aligned} \tag{7}$$

- Problems (7) and (6) are hard to solve since they are infinitely dimensional
- **Key idea:** replace the intial 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 for EG

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

$$\begin{aligned}
\Delta_{\mathtt{EG}}(L,\gamma) = \quad \max \quad & \|F(x^{k+1})\|^2 - \|F(x^k)\|^2 \\
\text{s.t.} \quad & F \text{ is monotone and } L\text{-Lipschitz}, \; x^k \in \mathbb{R}^d, \\
& x^{k+1/2} = x^k - \gamma F(x^k) \\
& x^{k+1} = x^k - \gamma F(x^{k+1/2})
\end{aligned} \tag{7}$$

- Problems (7) and (6) are hard to solve since they are infinitely dimensional
- **Key idea:** replace the intial 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 $F(x^{k+1})$.
- **Key point:** consider monotonicity and Lipchitzness at these points

Finite-Dimensional Relaxation

$$\begin{aligned}
\max\ & \|F_{k+1}\|^2 - \|F_k\|^2 && (8)\\
\text{s.t. } & d \text{ and } x^k, F_k, F_{k+1}, F_{k+\frac{1}{2}} \in \mathbb{R}^d, \\
& x^{k+\frac{1}{2}} = x^k - \gamma F_k, && \textit{(extrapolation step)}\\
& x^{k+1} = x^k - \gamma F_{k+\frac{1}{2}}, && \textit{(update step)}
\end{aligned}$$

Finite-Dimensional Relaxation

$$
\begin{aligned}
&\max \|F_{k+1}\|^2 - \|F_k\|^2 && (8)\\
&\text{s.t. } d \text{ and } x^k, F_k, F_{k+1}, F_{k+\frac{1}{2}} \in \mathbb{R}^d, \\
&\qquad x^{k+\frac{1}{2}} = x^k - \gamma F_k, && \textit{(extrapolation step)}\\
&\qquad x^{k+1} = x^k - \gamma F_{k+\frac{1}{2}}, && \textit{(update step)}\\
&\lambda_1 \ : 0 \leq \langle F_k - F_{k+\frac{1}{2}}, x^k - x^{k+\frac{1}{2}} \rangle, && \textit{(monotonicity in } (x^k, x^{k+\frac{1}{2}})\textit{)}\\
&\lambda_2 \ : 0 \leq \langle F_k - F_{k+1}, x^k - x^{k+1} \rangle, && \textit{(monotonicity in } (x^k, x^{k+1})\textit{)}\\
&\lambda_3 \ : 0 \leq \langle F_{k+1} - F_{k+\frac{1}{2}}, x^k - x^{k+1/2} \rangle, && \textit{(monotonicity in } (x^{k+1}, x^{k+\frac{1}{2}})\textit{)}
\end{aligned}
$$

Finite-Dimensional Relaxation

$$
\begin{aligned}
\max\ & \|F_{k+1}\|^2 - \|F_k\|^2 && (8)\\
\text{s.t. } & d \text{ and } x^k, F_k, F_{k+1}, F_{k+\frac{1}{2}} \in \mathbb{R}^d, \\
& x^{k+\frac{1}{2}} = x^k - \gamma F_k, && \textit{(extrapolation step)}\\
& x^{k+1} = x^k - \gamma F_{k+\frac{1}{2}}, && \textit{(update step)}\\
\lambda_1\ &: 0 \le \langle F_k - F_{k+\frac{1}{2}}, x^k - x^{k+\frac{1}{2}} \rangle, && \textit{(monotonicity in } (x^k, x^{k+\frac{1}{2}})\textit{)}\\
\lambda_2\ &: 0 \le \langle F_k - F_{k+1}, x^k - x^{k+1} \rangle, && \textit{(monotonicity in } (x^k, x^{k+1})\textit{)}\\
\lambda_3\ &: 0 \le \langle F_{k+1} - F_{k+\frac{1}{2}}, x^k - x^{k+1/2} \rangle, && \textit{(monotonicity in } (x^{k+1}, x^{k+\frac{1}{2}})\textit{)}\\
\lambda_4\ &: \|F_k - F_{k+\frac{1}{2}}\|^2 \le L^2\gamma^2 \|x^k - x^{k+\frac{1}{2}}\|^2, && \textit{(Lipschitzness in } (x^k, x^{k+\frac{1}{2}})\textit{)}\\
\lambda_5\ &: \|F_k - F_{k+1}\|^2 \le L^2\gamma^2 \|x^k - x^{k+1}\|^2, && \textit{(Lipschitzness in } (x^k, x^{k+1})\textit{)}\\
\lambda_6\ &: \|F_{k+1} - F_{k+\frac{1}{2}}\|^2 \le L^2\gamma^2 \|x^k - x^{k+\frac{1}{2}}\|^2. && \textit{(Lipschitzness in } (x^k, x^{k+\frac{1}{2}})\textit{)}
\end{aligned}
$$

Finite-Dimensional Relaxation: Bad News

✗ 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

Finite-Dimensional Relaxation: Bad News

✗ 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:

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

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:

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

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

$$\mathbf{G} \in \mathbb{S}^6_+ \iff \exists\, x^k, x^{k+\frac{1}{2}}, x^{k+1}, F_k, F_{k+\frac{1}{2}}, F_{k+1} \in \mathbb{R}^d : \ \mathbf{G} \text{ is Gram matrix.}$$

Primal SDP

- Therefore, problem (8) is equivalent to the following SDP problem:

$$
\begin{aligned}
\max \quad & \mathrm{Tr}(\mathbf{M}_0 \mathbf{G}) \\
\text{s.t.} \quad & \mathbf{G} \in \mathbb{S}^4_+, \\
& \mathrm{Tr}(\mathbf{M}_i \mathbf{G}) \leq 0, \ i = 1, 2, \ldots, 6,
\end{aligned}
\tag{9}
$$

where $\mathbf{M}_0, \ldots, \mathbf{M}_6$ are some symmetric matrices.

Primal SDP

- Therefore, problem (8) is equivalent to the following SDP problem:

$$\begin{aligned} \max \quad & \mathrm{Tr}(\mathbf{M}_0\mathbf{G}) \\ \text{s.t.} \quad & \mathbf{G} \in \mathbb{S}_+^4, \\ & \mathrm{Tr}(\mathbf{M}_i\mathbf{G}) \leq 0,\ i = 1, 2, \ldots, 6, \end{aligned} \tag{9}$$

  where $\mathbf{M}_0, \ldots, \mathbf{M}_6$ are some symmetric matrices.
- In that case, the dual problem is very simple:

$$\text{Find} \quad \lambda_1, \ldots, \lambda_6 \geq 0 \quad \text{such that} \quad \sum_{i=1}^{6} \lambda_i M_i \succeq M_0 \tag{10}$$

Primal SDP

- Therefore, problem (8) is equivalent to the following SDP problem:

$$\begin{aligned} \max \quad & \mathrm{Tr}(\mathbf{M}_0\mathbf{G}) \\ \text{s.t.} \quad & \mathbf{G} \in \mathbb{S}^4_+, \\ & \mathrm{Tr}(\mathbf{M}_i\mathbf{G}) \leq 0,\ i = 1, 2, \ldots, 6, \end{aligned} \tag{9}$$

  where $\mathbf{M}_0, \ldots, \mathbf{M}_6$ are some symmetric matrices.

- In that case, the dual problem is very simple:

$$\text{Find} \quad \lambda_1, \ldots, \lambda_6 \geq 0 \quad \text{such that} \quad \sum_{i=1}^{6} \lambda_i M_i \succeq M_0 \tag{10}$$

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

$$0 \geq \lambda_i \mathrm{Tr}(M_i G) \geq \mathrm{Tr}(M_0 G) \tag{11}$$

Analysis of the Solution

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

Analysis of the Solution

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

$$\lambda_1 \approx \tfrac{1}{2}\,, \quad \lambda_2 \approx 2\,, \quad \lambda_3 = 0\,, \quad \lambda_4 = 0\,, \quad \lambda_5 = 0\,, \quad \lambda_6 \approx \tfrac{3}{2}\,,$$

Analysis of the Solution

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

$$\lambda_1 \approx \tfrac{1}{2}\,, \quad \lambda_2 \approx 2\,, \quad \lambda_3 = 0\,, \quad \lambda_4 = 0\,, \quad \lambda_5 = 0\,, \quad \lambda_6 \approx \tfrac{3}{2}\,,$$

- Duality of SDPs says that

$$\begin{aligned}
\|F(x^k)\|^2 - \|F(x^{k+1})\|^2 &\leq \lambda_6(\|F(x^{k+1}) - F(x^{k+\frac{1}{2}})\|^2 - L^2\gamma^2\|x^k - x^{k+\frac{1}{2}}\|^2) \\
&\quad - \lambda_1\langle F(x^k) - F(x^{k+\frac{1}{2}}), x^k - x^{k+\frac{1}{2}}\rangle \\
&\quad - \lambda_2\langle F(x^k) - F(x^{k+1}), x^k - x^{k+1}\rangle \\
&\leq 0
\end{aligned}$$

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

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

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

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

Using this result, it is quite trivial to derive last-iterate $\mathcal{O}(1/K)$ 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 $0 < \gamma \leq 1/\sqrt{2}L$. Moreover,

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

Last-Iterate Convergence of Optimistic Gradient

Going Back to the Constrained Setting

- Problem:

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

Going Back to the Constrained Setting

- Problem:

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

- Projected Extragradient:

$$\widetilde{x}^k = \mathrm{proj}[x^k - \gamma F(x^k)], \quad x^{k+1} = \mathrm{proj}[x^k - \gamma F(\widetilde{x}^k)] \tag{Proj-EG}$$

- Projected Past Extragradient:

$$\widetilde{x}^k = \mathrm{proj}[x^k - \gamma F(\widetilde{x}^{k-1})], \quad x^{k+1} = \mathrm{proj}[x^k - \gamma F(\widetilde{x}^k)] \tag{Proj-PEG}$$

Going Back to the Constrained Setting

- Problem:

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

- Projected Extragradient:

$$\widetilde{x}^k = \mathrm{proj}[x^k - \gamma F(x^k)], \quad x^{k+1} = \mathrm{proj}[x^k - \gamma F(\widetilde{x}^k)] \quad (\texttt{Proj-EG})$$

- Projected Past Extragradient:

$$\widetilde{x}^k = \mathrm{proj}[x^k - \gamma F(\widetilde{x}^{k-1})], \quad x^{k+1} = \mathrm{proj}[x^k - \gamma F(\widetilde{x}^k)] \quad (\texttt{Proj-PEG})$$

- Convergence metric: $\|x^{k+1} - x^k\|^2$.

Past Extragradient and Optimistic Gradient

In the unconstrained case, `PEG` and `OG` are equivalent

- Past Extragradient (`PEG`)

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

- Optimistic Gradient method (`OG`)

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

PEP for $\mathtt{PEG}$

$$G_{\mathtt{PEG}}(\gamma, L, N) = \max_{\substack{F,d,x^* \\ \tilde{x}^0,\ldots,\tilde{x}^N \\ x^0,\ldots,x^N}} \quad \frac{\|F(x^N)\|^2}{\|x^0 - x^*\|^2} \tag{13}$$

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

PEP for PEG: Relaxation

$$\widetilde{G}_{\mathrm{PEG}}(\gamma, L, N) = \max_{\substack{d\in\mathbb{N}, x^*\in\mathbb{R}^d \\ \{(x^k,g^k)\}_{k=0}^N \subset \mathbb{R}^d\times\mathbb{R}^d \\ \{(\widetilde{x}^k,\widetilde{g}^k)\}_{k=0}^N \subseteq \mathbb{R}^d\times\mathbb{R}^d}} \|g^N\|^2 \tag{14}$$

$$\text{s.t.} \langle g-h, x-y\rangle \geq 0 \quad \forall (x,g),(y,h)\in S \tag{15}$$

$$\|g-h\|^2 \leq L^2\|x-y\|^2 \quad \forall (x,g),(y,h)\in S \tag{16}$$

$$\widetilde{x}^0 = x^0 \in \mathbb{R}^d, x^1 = x^0 - \gamma g^0$$

$$\widetilde{x}^k = x^k - \gamma \widetilde{g}^{k-1}, \text{ for } k = 1,\dots,N,$$

$$x^{k+1} = x^k - \gamma \widetilde{g}^k, \text{ for } k = 1,\dots,N-1,$$

$$\|x^0 - x^*\|^2 \leq 1 \tag{17}$$

- Following the same steps as in the previous examples, one can reformulate this problem as SDP

PEP for OG

$$
\begin{aligned}
G_{\texttt{OG}}(\gamma, L, N) = \quad & \max_{\substack{F, d, x^* \\ \tilde{x}^0, \ldots, \tilde{x}^N}} \quad \frac{\|F(\widetilde{x}^N)\|^2}{\|\widetilde{x}^0 - x^*\|^2} \\
& \text{s.t.} \quad F \text{ is monotone and } L\text{-Lipschitz}, \\
& \qquad \widetilde{x}^0 \in \mathbb{R}^d,\ \widetilde{x}^1 = \widetilde{x}^0 - \gamma F(\widetilde{x}^0), \\
& \qquad \widetilde{x}^{k+1} = \widetilde{x}^k - 2\gamma F(\widetilde{x}^k) + \gamma F(\widetilde{x}^{k-1}), \\
& \qquad\qquad \text{for } k = 1, \ldots, N-1,
\end{aligned}
\tag{18}
$$

PEP for OG: Relaxation

$$
\begin{aligned}
\widetilde{G}_{\mathtt{OG}}(\gamma, L, N) = \max_{\substack{d\in\mathbb{N}, x^*\in\mathbb{R}^d \\ \{(\widetilde{x}^k, \widetilde{g}^k)\}_{k=0}^N \subseteq \mathbb{R}^d\times\mathbb{R}^d}} \quad & \|g^N\|^2 && (19)\\
\text{s.t.} & \langle g-h, x-y\rangle \geq 0 \quad \forall (x,g),(y,h)\in S && (20)\\
& \|g-h\|^2 \leq L^2\|x-y\|^2 \quad \forall (x,g),(y,h)\in S && (21)\\
& \widetilde{x}^0 \in \mathbb{R}^d,\ \widetilde{x}^1 = \widetilde{x}^0 - \gamma\widetilde{g}^0, \\
& \widetilde{x}^{k+1} = \widetilde{x}^k - 2\gamma\widetilde{g}^k + \gamma\widetilde{g}^{k-1}, \\
& \qquad \text{for } k = 1, \ldots, N-1, \\
& \|x^0 - x^*\|^2 \leq 1 && (22)
\end{aligned}
$$

- Following the same steps as in the previous examples, one can reformulate this problem as SDP

Both Relaxations Show $\mathcal{O}(1/N)$ Convergence

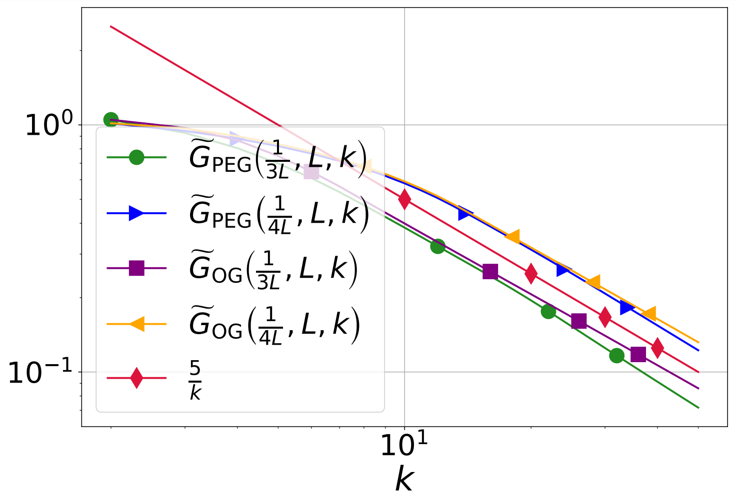

Figure 2: $\widetilde{G}_{\mathrm{PEG}}(\gamma, L, N)$ and $\widetilde{G}_{\mathrm{OG}}(\gamma, L, N)$ for different values of γ and N

PEP for `PEG` and `OG`: Existence of Simple Proofs?

- Typical proof – some clever combination of inequalities

PEP for `PEG` and `OG`: Existence of Simple Proofs?

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

PEP for PEG and OG: Existence of Simple Proofs?

- 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 $\widetilde{G}_{\texttt{PEG}}(\gamma, L, N, t)$ and $\widetilde{G}_{\texttt{OG}}(\gamma, L, N, t)$

PEP for PEG and OG: Existence of Simple Proofs?

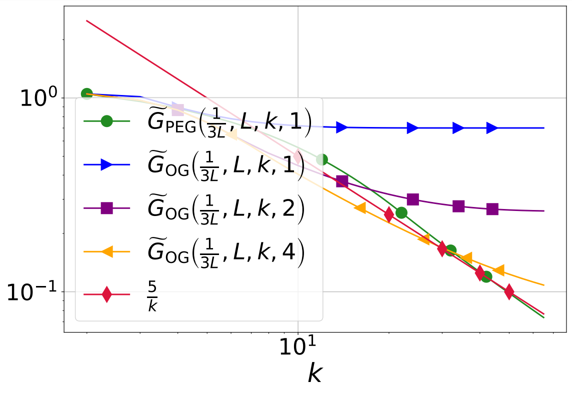

Figure 3: We report $\widetilde{G}_{\texttt{PEG}}(\gamma, L, N, 1)$ and $\widetilde{G}_{\texttt{OG}}(\gamma, L, N, t)$ for $t = 1, 2, 4$. It suggest that $\widetilde{G}_{\texttt{PEG}}(\gamma, L, N, 1) \sim 1/N$ but not $\widetilde{G}_{\texttt{OG}}(\gamma, L, N, t)$ (even for $t = 4$).

PEP for PEG and OG: Existence of Simple Proofs?

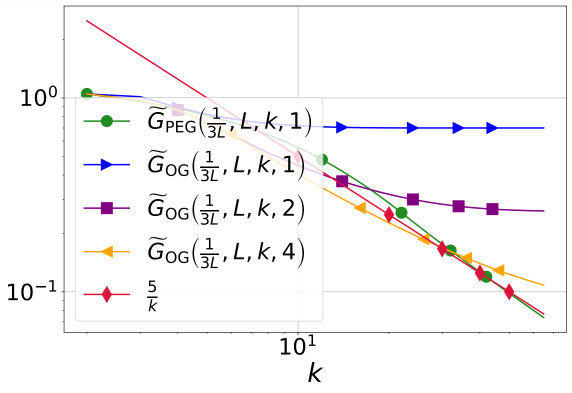

Figure 3: We report $\widetilde{G}_{\texttt{PEG}}(\gamma, L, N, 1)$ and $\widetilde{G}_{\texttt{OG}}(\gamma, L, N, t)$ for $t = 1, 2, 4$. It suggest that $\widetilde{G}_{\texttt{PEG}}(\gamma, L, N, 1) \sim {}^1\!/\!{}_N$ but not $\widetilde{G}_{\texttt{OG}}(\gamma, L, N, t)$ (even for $t = 4$).

Since interpolation is not guaranteed, extra points are crucial!

Can We Prove that the Norm Monotonically Decrease?

$$\Delta(\gamma, L, N) = \max_{\substack{F, d, x^* \\ \tilde{x}^0, \ldots, \tilde{x}^N \\ x^0, \ldots, x^N}} \frac{\|F(x^{N+1})\|^2 - \|F(x^N)\|^2}{\|x^0 - x^*\|^2} \tag{23}$$

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

Can We Prove that the Norm Monotonically Decrease?

$$
\begin{aligned}
\Delta(\gamma, L, N) = \quad & \max_{\substack{F, d, x^* \\ \tilde{x}^0, \ldots, \tilde{x}^N \\ x^0, \ldots, x^N}} \quad \frac{\|F(x^{N+1})\|^2 - \|F(x^N)\|^2}{\|x^0 - x^*\|^2} \\
& \text{s.t.} \quad F \text{ is monotone and } L\text{-Lipschitz}, \\
& \qquad \widetilde{x}^0 = x^0 \in \mathbb{R}^d,\ x^1 = x^0 - \gamma F(x^0), \\
& \qquad \widetilde{x}^k = x^k - \gamma F(\widetilde{x}^{k-1}),\ \text{for } k = 1, \ldots, N, \\
& \qquad x^{k+1} = x^k - \gamma F(\widetilde{x}^k),\ \text{for } k = 1, \ldots, N-1,
\end{aligned}
\tag{23}
$$

- $\widetilde{\Delta}(\gamma, L, N)$ – value of SDP relaxation

Can We Prove that the Norm Monotonically Decrease?

$$\Delta(\gamma, L, N) = \max_{\substack{F, d, x^* \\ \tilde{x}^0, \dots, \tilde{x}^N \\ x^0, \dots, x^N}} \quad \frac{\|F(x^{N+1})\|^2 - \|F(x^N)\|^2}{\|x^0 - x^*\|^2} \tag{23}$$

$$\begin{aligned} \text{s.t.} \quad & F \text{ is monotone and } L\text{-Lipschitz}, \\ & \widetilde{x}^0 = x^0 \in \mathbb{R}^d,\ x^1 = x^0 - \gamma F(x^0), \\ & \widetilde{x}^k = x^k - \gamma F(\widetilde{x}^{k-1}),\ \text{for } k = 1, \dots, N, \\ & x^{k+1} = x^k - \gamma F(\widetilde{x}^k),\ \text{for } k = 1, \dots, N-1, \end{aligned}$$

- $\widetilde{\Delta}(\gamma, L, N)$ – 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 $\widetilde{\delta}(\gamma, L, N)$
  - Guaranteed interpolation (Ryu et al., 2020): $\delta(\gamma, L, N) = \widetilde{\delta}(\gamma, L, N)$

Can We Prove that the Norm Monotonically Decrease? No

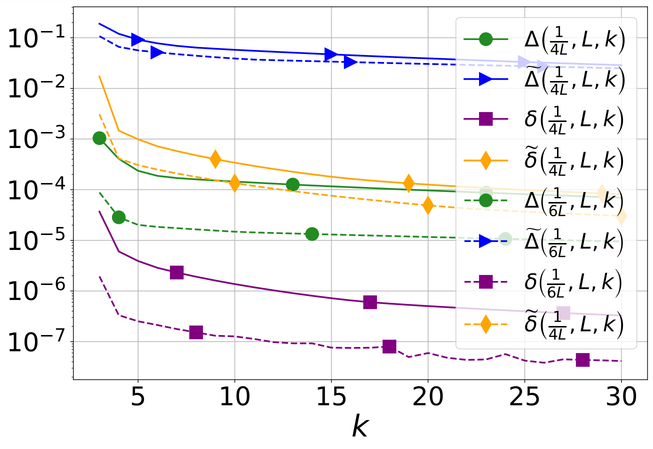

Figure 4: Evolution of $\Delta(\gamma, L, N), \widetilde{\Delta}(\gamma, L, N), \delta(\gamma, L, N), \widetilde{\delta}(\gamma, L, N)$

Can We Prove that the Norm Monotonically Decrease? No

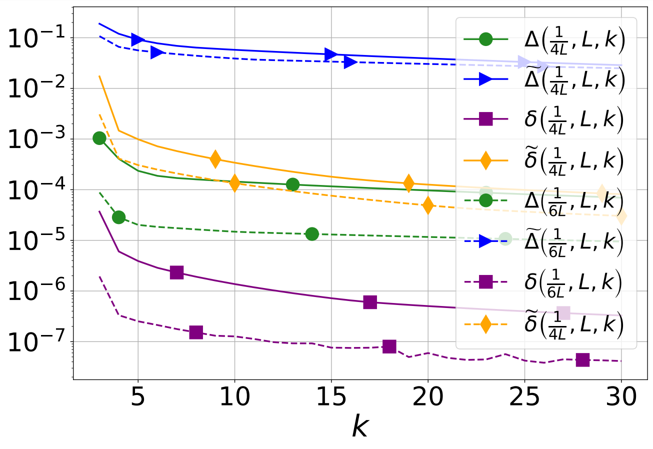

Figure 4: Evolution of $\Delta(\gamma, L, N), \widetilde{\Delta}(\gamma, L, N), \delta(\gamma, L, N), \widetilde{\delta}(\gamma, L, N)$

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:

$$\|F(x^{N+1})\|^2+2\|F(x^{N+1})-F(\widetilde{x}^N)\|^2 \leq \|F(x^N)\|^2+2\|F(x^N)-F(\widetilde{x}^{N-1})\|^2$$

Potential for PEG: Unconstrained Problems

Theorem 3

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

$$\begin{aligned}\|F(x^{k+1})\|^2 + 2\|F(x^{k+1}) - F(\widetilde{x}^k)\|^2 \leq \|F(x^k)\|^2 + 2\|F(x^k) - F(\widetilde{x}^{k-1})\|^2 \\ + 3\left(L^2\gamma^2 - \frac{2}{9}\right)\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2.\end{aligned}$$

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

Last-Iterate Convergence of PEG for Unconstrained Problems

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 Φ_k defined as

$$\Phi_k = \|x^k - x^*\|^2 + \frac{k+32}{3}\gamma^2 \left(\|F(x^k)\|^2 + 2\|F(x^k) - F(\widetilde{x}^{k-1})\|^2\right). \quad (24)$$

Last-Iterate Convergence of PEG for Unconstrained Problems

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 Φ_k defined as

$$\Phi_k = \|x^k - x^*\|^2 + \frac{k+32}{3}\gamma^2 \left(\|F(x^k)\|^2 + 2\|F(x^k) - F(\widetilde{x}^{k-1})\|^2\right). \quad (24)$$

In particular, for all $N \geq 0$ and $\gamma \leq 1/3L$ the above formula implies

$$\begin{aligned} \|F(x^N)\|^2 &\leq \frac{3(1+32L^2\gamma^2)\|x^0 - x^*\|^2}{\gamma^2(N+32)}, \\ \mathrm{Gap}_F(x^k) &\leq \frac{2\sqrt{41}(1+32L^2\gamma^2)\|x^0 - x^*\|^2}{\gamma\sqrt{3N+96}}. \end{aligned}$$

From Unconstrained to Constrained Problems

- 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

PEP for Searching Potentials

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

$$\begin{aligned}
\text{find} \quad & F, d, x^*, \tilde{x}^0, \ldots, \tilde{x}^N, x^0, \ldots, x^N, \Phi_N, \Phi_{N-1} \\
\text{s.t.} \quad & F \text{ is monotone and } L\text{-Lipschitz}, \\
& \Phi_N \text{ and } \Phi_{N-1} \text{ are quadratic w.r.t. iterates and operator values}, \\
& \Phi_N \text{ and } \Phi_{N-1} \text{ have same structure}, \\
& \|x^N - x^{N-1}\|^2 \leq \Phi_N,\ \Phi_N - \Phi_{N-1} \leq 0, \\
& \widetilde{x}^0 = x^0 \in \mathbb{R}^d,\ x^1 = x^0 - \gamma F(x^0), \\
& \widetilde{x}^k = x^k - \gamma F(\widetilde{x}^{k-1}),\ \text{for } k = 1, \ldots, N, \\
& x^{k+1} = x^k - \gamma F(\widetilde{x}^k),\ \text{for } k = 1, \ldots, N-1
\end{aligned} \tag{25}$$

PEP for Searching Potentials

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

$$
\begin{aligned}
\text{find} \quad & F, d, x^*, \tilde{x}^0, \ldots, \tilde{x}^N, x^0, \ldots, x^N, \Phi_N, \Phi_{N-1} \\
\text{s.t.} \quad & F \text{ is monotone and } L\text{-Lipschitz}, \\
& \Phi_N \text{ and } \Phi_{N-1} \text{ are quadratic w.r.t. iterates and operator values}, \\
& \Phi_N \text{ and } \Phi_{N-1} \text{ have same structure}, \\
& \|x^N - x^{N-1}\|^2 \leq \Phi_N,\ \Phi_N - \Phi_{N-1} \leq 0, \\
& \widetilde{x}^0 = x^0 \in \mathbb{R}^d,\ x^1 = x^0 - \gamma F(x^0), \\
& \widetilde{x}^k = x^k - \gamma F(\widetilde{x}^{k-1}),\ \text{for } k = 1, \ldots, N, \\
& x^{k+1} = x^k - \gamma F(\widetilde{x}^k),\ \text{for } k = 1, \ldots, N-1
\end{aligned}
\tag{25}
$$

The dual SDP relaxation can be efficiently solved!

Potential for PEG: Constrained Problems

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

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

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

where

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

Last-Iterate Convergence of PEG for Constrained Problems

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 Φ_k defined as

$$\Phi_k = \|x^k - x^*\|^2 + \frac{1}{16}\|\widetilde{x}^{k-1} - \widetilde{x}^{k-2}\|^2 + \frac{3k+32}{24}\Psi_k, \tag{27}$$

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

Last-Iterate Convergence of PEG for Constrained Problems

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 Φ_k defined as

$$\Phi_k = \|x^k - x^*\|^2 + \frac{1}{16}\|\widetilde{x}^{k-1} - \widetilde{x}^{k-2}\|^2 + \frac{3k+32}{24}\Psi_k, \tag{27}$$

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

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

where $H_0, H_{0,\gamma} > 0$ are such that
$H_{0,\gamma}^2 = 2(1 + 3\gamma^2L^2 + 4\gamma^4L^4)\|x^0 - x^*\|^2 + \left(\frac{41}{12} + \frac{19}{3}\gamma^2L^2\right)\gamma^2\|F(x^0)\|^2$,
$H_0^2 = 3\|x^0 - x^*\|^2 + \frac{1}{30L^2}\|F(x^0)\|^2$.

Last-Iterate Convergence Under Negative-Comonotonicity

Negative Comonotonicity for the Unconstrained Case

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

- $F : \mathbb{R}^d \to \mathbb{R}^d$ 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

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

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

$$\|F(x) - F(y)\| \leq L\|x - y\| \tag{28}$$

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

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

Negative Comonotonicity for the Unconstrained Case

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

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

$$\|F(x) - F(y)\| \leq L\|x - y\| \tag{28}$$

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

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

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

Theorem 7

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

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

Last-Iterate Convergence of EG

Theorem 7

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

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

and for any $N \geq 1$

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

Last-Iterate Convergence of EG

Theorem 7

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

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

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

Last-Iterate Convergence of $\texttt{EG}$

Theorem 7

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

Last-Iterate Convergence of EG

Theorem 7

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

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

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
- ✓ Previous result is derived for $\rho < 1/16L$ (Luo and Tran-Dinh, 2022)
- ? **Open question:** is it possible to show $\mathcal{O}(1/N)$ last-iterate convergence for EG when $\rho \in (1/8L, 1/2L)$?

Last-Iterate Convergence of OG

Theorem 8

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

$$\begin{aligned} \|F(x^{k+1})\|^2 + \|F(x^{k+1}) - F(\widetilde{x}^k)\|^2 \quad &\leq \quad \|F(x^k)\|^2 + \|F(x^k) - F(\widetilde{x}^{k-1})\|^2 \\ &\quad - \frac{1}{100}\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2. \quad (32) \end{aligned}$$

Last-Iterate Convergence of `OG`

Theorem 8

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

$$\begin{aligned} \|F(x^{k+1})\|^2 + \|F(x^{k+1}) - F(\widetilde{x}^k)\|^2 \quad &\leq \quad \|F(x^k)\|^2 + \|F(x^k) - F(\widetilde{x}^{k-1})\|^2 \\ &\quad - \frac{1}{100}\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2. \end{aligned} \tag{32}$$

and for any $N \geq 1$

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

Last-Iterate Convergence of OG

Theorem 8

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

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

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 OG

Theorem 8

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

$$\begin{aligned} \|F(x^{k+1})\|^2 + \|F(x^{k+1}) - F(\widetilde{x}^k)\|^2 \quad &\leq \quad \|F(x^k)\|^2 + \|F(x^k) - F(\widetilde{x}^{k-1})\|^2 \\ &\qquad - \frac{1}{100}\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2. \end{aligned} \tag{32}$$

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 $\rho < {}^8/_{27\sqrt{6}L}$ (Luo and Tran-Dinh, 2022)

Theorem 8

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

$$\begin{aligned} \|F(x^{k+1})\|^2 + \|F(x^{k+1}) - F(\widetilde{x}^k)\|^2 \quad \le \quad & \|F(x^k)\|^2 + \|F(x^k) - F(\widetilde{x}^{k-1})\|^2 \\ & - \frac{1}{100}\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2. \end{aligned} \quad (32)$$

and for any $N \ge 1$

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

- ✓ Again, we found the potential via computer
- ✓ Previous result is derived for $\rho < {}^8/_{27\sqrt{6}L}$ (Luo and Tran-Dinh, 2022)
- ? **Open question:** is it possible to show $\mathcal{O}({}^1/_N)$ last-iterate convergence for OG when $\rho \in ({}^5/_{62L}, {}^1/_{2L})$?

No Convergence for 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 ρ-negative comonotone L-Lipschitz operator F such that EG/OG does not necessary converges on solving (VIP) with this operator F.

No Convergence for 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 ρ-negative comonotone L-Lipschitz operator F such that EG/OG does not necessary 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 γ 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Nemirovski, A. (2004). Prox-method with rate of convergence $O(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251.

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

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

Popov, L. D. (1980). 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.

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

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

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

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

Taylor, A. B., Hendrickx, J. M., and Glineur, F. (2017a). 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.

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