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

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

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

Accepted to AISTATS 2022

Rising Stars in AI Symposium 2022, KAUST

March 13, 2022

# Short Summary of Our Work

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

# Short Summary of Our Work

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

# Short Summary of Our Work

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

# Short Summary of Our Work

- We prove $\mathcal{O}(1/\kappa)$ *last-iterate* convergence rate for Extragardient method [Korpelevich, 1976] in terms of squared norm of the operator for monotone Lipschitz variational ineqiality problems (VIPs)
  - The proof is *obtained via computer*
- We establish new connections for several known methods with cocoercivity when the original operator is monotone and Lipschitz
  - In particular, our results empahsize the mathematical differences between Extragradient method and Optimistic Gradient method [Popov, 1980] that usualy considered as approximations of Proximal Point method
- Our code is available online: `https://github.com/eduardgorbunov/extragradient_last_iterate_AISTATS_2022`

# Outline

# Variational Inequality Problem

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

# Variational Inequality Problem

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

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

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

# Variational Inequality Problem

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

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

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

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:

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

  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:

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

# Variational Inequality Problem: Examples

- Min-max problems:

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

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

- Minimization problems:

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

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

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

In this talk, we focus on (VIP) rather than (VIP-C)

# How to Solve VIPs?

Naive approach – Gradient Descent (GD):

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

# How to Solve VIPs?

Naive approach – Gradient Descent (GD):

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

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

# How to Solve VIPs?

Naive approach – Gradient Descent (GD):

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

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

# Non-Convergence of GD

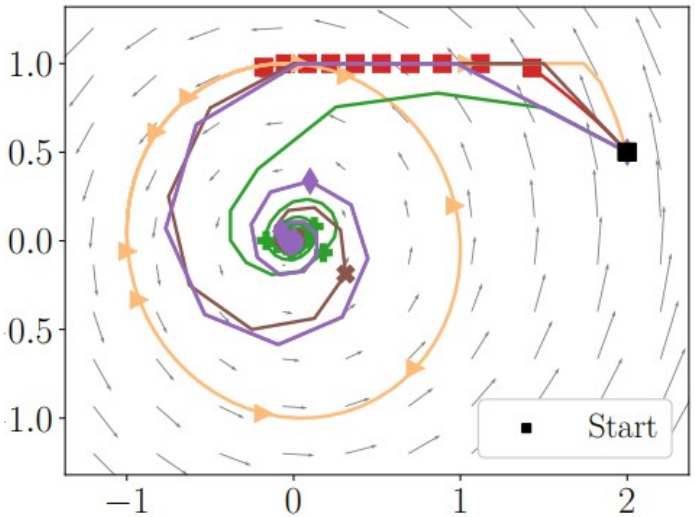



Figure 1: Comparison of the basic gradient method (as well as Adam) with the techniques presented in §3 on the optimization of (9). Only the algorithms advocated in this paper (Averaging, Extrapolation and Extrapolation from the past) converge quickly to the solution. Each marker represents 20 iterations.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

Figure: Behavior of GD 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]

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

# Popular Alternatives to GD

- Extragradient method (EG) [Korpelevich, 1976]

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

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

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

# Popular Alternatives to GD

- Extragradient method (EG) [Korpelevich, 1976]

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

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

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

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

# Measures of Convergence

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

# Measures of Convergence

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

# Measures of Convergence

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

# Measures of Convergence

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

# Measures of Convergence

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

# Measures of Convergence

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

# Measures of Convergence

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

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

# Convergence Guarantees for EG

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

# Convergence Guarantees for EG

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

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

# 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]

# 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]
  - $\min\limits_{k=0,1,\ldots,K} \|F(x^k)\|^2 = \mathcal{O}(1/K)$

# 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 $\Lambda$-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 $\Lambda$-Lipschitz*, then
  - $\texttt{Gap}_F(x^K) = \mathcal{O}(1/\sqrt{K})$
  - $\|F(x^K)\|^2 = \mathcal{O}(1/K)$

# Convergence Guarantees for EG: Resolved Question

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

# Convergence Guarantees for EG: Resolved Question

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

We give a positive answer to this question in our paper

# Performance Estimation Problems

We derive the result via solving a special *Performance Estimation Problem* **numerically**.

# Performance Estimation Problems

We derive the result via solving a special *Performance Estimation Problem* **numerically**.

PEP

- A powerful technique for deriving tight convergence guarantees, obtaining proofs and even designing new optimal methods

# Performance Estimation Problems

We derive the result via solving a special *Performance Estimation Problem* **numerically**.

PEP

- A powerful technique for deriving tight convergence guarantees, obtaining proofs and even designing new optimal methods
- First works: [Drori and Teboulle, 2014, Kim and Fessler, 2016, Lessard et al., 2016]
- Some later works: Taylor et al. [2017a,b], De Klerk et al. [2017], Ryu et al. [2020], Taylor and Bach [2019]

# Performance Estimation Problems

We derive the result via solving a special *Performance Estimation Problem* **numerically**.

PEP

- A powerful technique for deriving tight convergence guarantees, obtaining proofs and even designing new optimal methods
- First works: [Drori and Teboulle, 2014, Kim and Fessler, 2016, Lessard et al., 2016]
- Some later works: Taylor et al. [2017a,b], De Klerk et al. [2017], Ryu et al. [2020], Taylor and Bach [2019]
- For those who are interested in this topic, I recommend to read papers and slides by Adrien Taylor `https://www.di.ens.fr/~ataylor`

# Performance Estimation Problem: A General Form

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

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

# PEP for EG: Infinitely Dimensional Formulation

We consider the problem

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

# PEP for EG: Infinitely Dimensional Formulation

We consider the problem

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

- Problem (6) is hard to solve since it is infinitely dimensional

# PEP for EG: Infinitely Dimensional Formulation

We consider the problem

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

- Problem (6) is hard to solve since it is infinitely dimensional
- **Key idea:** replace the intial problem by an "easy" problem

# PEP for EG: Finitely Dimensional Formulation

- Introduce new variables $\{(x^k, g^k)\}_{k=0}^K$ and $\{(\widetilde{x}^k, \widetilde{g}^k)\}_{k=0}^{K-1}$ such that $\widetilde{x}^k = x^k - \gamma_1 g^k$, $x^{k+1} = x^k - \gamma_2 \widetilde{g}^k$

# PEP for EG: Finitely Dimensional Formulation

- Introduce new variables $\{(x^k, g^k)\}_{k=0}^K$ and $\{(\widetilde{x}^k, \widetilde{g}^k)\}_{k=0}^{K-1}$ such that $\widetilde{x}^k = x^k - \gamma_1 g^k$, $x^{k+1} = x^k - \gamma_2 \widetilde{g}^k$
- Add a constraint that $F(x^k) = g^k$, $F(\widetilde{x}^k) = \widetilde{g}^k$ for some monotone and $L$-Lipschitz operator $F$.

# PEP for EG: Finitely Dimensional Formulation

- Introduce new variables $\{(x^k, g^k)\}_{k=0}^K$ and $\{(\widetilde{x}^k, \widetilde{g}^k)\}_{k=0}^{K-1}$ such that $\widetilde{x}^k = x^k - \gamma_1 g^k$, $x^{k+1} = x^k - \gamma_2 \widetilde{g}^k$
- Add a constraint that $F(x^k) = g^k$, $F(\widetilde{x}^k) = \widetilde{g}^k$ for some monotone and $L$-Lipschitz operator $F$. The resulting PEP:

$$
\begin{aligned}
\max \quad & \|g^K\|^2 \\
\text{s.t.} \quad & \{x^k\}_{k=0}^K, \{\widetilde{x}^k\}_{k=0}^{K-1}, \{g^k\}_{k=0}^K, \{\widetilde{g}^k\}_{k=0}^{K-1} \in \mathbb{R}^d, \\
& \|x^0 - x^*\|^2 \leq 1, \\
& x^{k+1} = x^k - \gamma_2 \widetilde{g}^k, \ \widetilde{x}^k = x^k - \gamma_1 g^k, \ k = 0, 1, \ldots, K-1, \\
& \exists\, F : \mathbb{R}^d \to \mathbb{R}^d : \ F(x^k) = g^k, \ k = 0, 1, \ldots, K, \\
& F(\widetilde{x}^k) = \widetilde{g}^k, \ k = 0, 1, \ldots, K-1, \ F \text{ is monotone and } L - \text{Lipschitz}
\end{aligned}
\tag{7}
$$

# PEP for EG: Finitely Dimensional Formulation

- Introduce new variables $\{(x^k, g^k)\}_{k=0}^K$ and $\{(\widetilde{x}^k, \widetilde{g}^k)\}_{k=0}^{K-1}$ such that $\widetilde{x}^k = x^k - \gamma_1 g^k$, $x^{k+1} = x^k - \gamma_2 \widetilde{g}^k$
- Add a constraint that $F(x^k) = g^k$, $F(\widetilde{x}^k) = \widetilde{g}^k$ for some monotone and $L$-Lipschitz operator $F$. The resulting PEP:

$$
\begin{aligned}
\max \quad & \|g^K\|^2 \\
\text{s.t.} \quad & \{x^k\}_{k=0}^K, \{\widetilde{x}^k\}_{k=0}^{K-1}, \{g^k\}_{k=0}^K, \{\widetilde{g}^k\}_{k=0}^{K-1} \in \mathbb{R}^d, \\
& \|x^0 - x^*\|^2 \leq 1, \\
& x^{k+1} = x^k - \gamma_2 \widetilde{g}^k, \ \widetilde{x}^k = x^k - \gamma_1 g^k, \ k = 0, 1, \ldots, K-1, \\
& \exists\, F : \mathbb{R}^d \to \mathbb{R}^d : \ F(x^k) = g^k, \ k = 0, 1, \ldots, K, \\
& F(\widetilde{x}^k) = \widetilde{g}^k, \ k = 0, 1, \ldots, K-1, \ F \text{ is monotone and } L - \text{Lipschitz}
\end{aligned}
\tag{7}
$$

- Problem (7) is equivalent to (6), but it is still hard to solve.

# PEP for EG: Finitely Dimensional Formulation 2

Necessary conditions of the existence of monotone $L$-Lipschitz operator $F$ interpolating the pairs $\{(x^k, g^k)\}_{k=0}^K$, $\{(\widetilde{x}^k, \widetilde{g}^k)\}_{k=0}^{K-1}$:

# PEP for EG: Finitely Dimensional Formulation 2

Necessary conditions of the existence of monotone $L$-Lipschitz operator $F$ interpolating the pairs $\{(x^k, g^k)\}_{k=0}^K$, $\{(\widetilde{x}^k, \widetilde{g}^k)\}_{k=0}^{K-1}$:

$$\langle g - h, x - y \rangle \geq 0, \quad \|g - h\|^2 \leq L^2 \|x - y\|^2$$

for all pairs $(x, g)$, $(y, h)$ from $\{(x^k, g^k)\}_{k=0}^K$, $\{(\widetilde{x}^k, \widetilde{g}^k)\}_{k=0}^{K-1}$.

# PEP for EG: Finitely Dimensional Formulation 2

This leads us to the following formulation:

$$
\begin{aligned}
\max \quad & \|g^K\|^2 && (8)\\
\text{s.t.} \quad & \{x^k\}_{k=0}^K, \{\widetilde{x}^k\}_{k=0}^{K-1}, \{g^k\}_{k=0}^K, \{\widetilde{g}^k\}_{k=0}^{K-1} \in \mathbb{R}^d,\\
& \|x^0 - x^*\|^2 \leq 1,\\
& x^{k+1} = x^k - \gamma_2 \widetilde{g}^k,\ \widetilde{x}^k = x^k - \gamma_1 g^k,\ k = 0, 1, \ldots, K-1,\\
& \langle g - h, x - y\rangle \geq 0,\ \|g - h\|^2 \leq L^2 \|x - y\|^2, && (9)\\
& \text{for all pairs } (x, g), (y, h) \text{ from } \{(x^k, g^k)\}_{k=0}^K,\ \{(\widetilde{x}^k, \widetilde{g}^k)\}_{k=0}^{K-1} && (10)
\end{aligned}
$$

# PEP for EG: Finitely Dimensional Formulation 2

This leads us to the following formulation:

$$
\begin{aligned}
\max \quad & \|g^K\|^2 && (8)\\
\text{s.t.} \quad & \{x^k\}_{k=0}^K, \{\widetilde{x}^k\}_{k=0}^{K-1}, \{g^k\}_{k=0}^K, \{\widetilde{g}^k\}_{k=0}^{K-1} \in \mathbb{R}^d,\\
& \|x^0 - x^*\|^2 \leq 1,\\
& x^{k+1} = x^k - \gamma_2 \widetilde{g}^k,\ \widetilde{x}^k = x^k - \gamma_1 g^k,\ k = 0, 1, \ldots, K-1,\\
& \langle g - h, x - y\rangle \geq 0,\ \|g - h\|^2 \leq L^2 \|x - y\|^2, && (9)\\
& \text{for all pairs } (x, g), (y, h) \text{ from } \{(x^k, g^k)\}_{k=0}^K,\ \{(\widetilde{x}^k, \widetilde{g}^k)\}_{k=0}^{K-1} && (10)
\end{aligned}
$$

- Unfortunately, this problem is not equivalent to (6) since it is possible to construct the set of points $\{(x^k, g^k)\}_{k=0}^K$, $\{(\widetilde{x}^k, \widetilde{g}^k)\}_{k=0}^{K-1}$ satisfying (9)-(10) such that there are no monotone $L$-Lipschitz operators interpolating these points (see Proposition 3 from Ryu et al. [2020])

# PEP for EG: Finitely Dimensional Formulation 2

This leads us to the following formulation:

$$
\begin{aligned}
\max \quad & \|g^K\|^2 && (8)\\
\text{s.t.} \quad & \{x^k\}_{k=0}^K, \{\widetilde{x}^k\}_{k=0}^{K-1}, \{g^k\}_{k=0}^K, \{\widetilde{g}^k\}_{k=0}^{K-1} \in \mathbb{R}^d,\\
& \|x^0 - x^*\|^2 \leq 1,\\
& x^{k+1} = x^k - \gamma_2 \widetilde{g}^k,\ \widetilde{x}^k = x^k - \gamma_1 g^k,\ k = 0, 1, \ldots, K-1,\\
& \langle g - h, x - y\rangle \geq 0,\ \|g - h\|^2 \leq L^2\|x - y\|^2, && (9)\\
& \text{for all pairs } (x, g), (y, h) \text{ from } \{(x^k, g^k)\}_{k=0}^K,\ \{(\widetilde{x}^k, \widetilde{g}^k)\}_{k=0}^{K-1} && (10)
\end{aligned}
$$

- Unfortunately, this problem is not equivalent to (6) since it is possible to construct the set of points $\{(x^k, g^k)\}_{k=0}^K$, $\{(\widetilde{x}^k, \widetilde{g}^k)\}_{k=0}^{K-1}$ satisfying (9)-(10) such that there are no monotone $L$-Lipschitz operators interpolating these points (see Proposition 3 from Ryu et al. [2020])
- Nevertheless, it gives a valid upper bound for $\|F(x^K)\|^2$

# PEP for EG: SDP Formulation

Problem (8) can be reformulated as an SDP problem.
This means that one can easily find its solutions using standard methods.

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

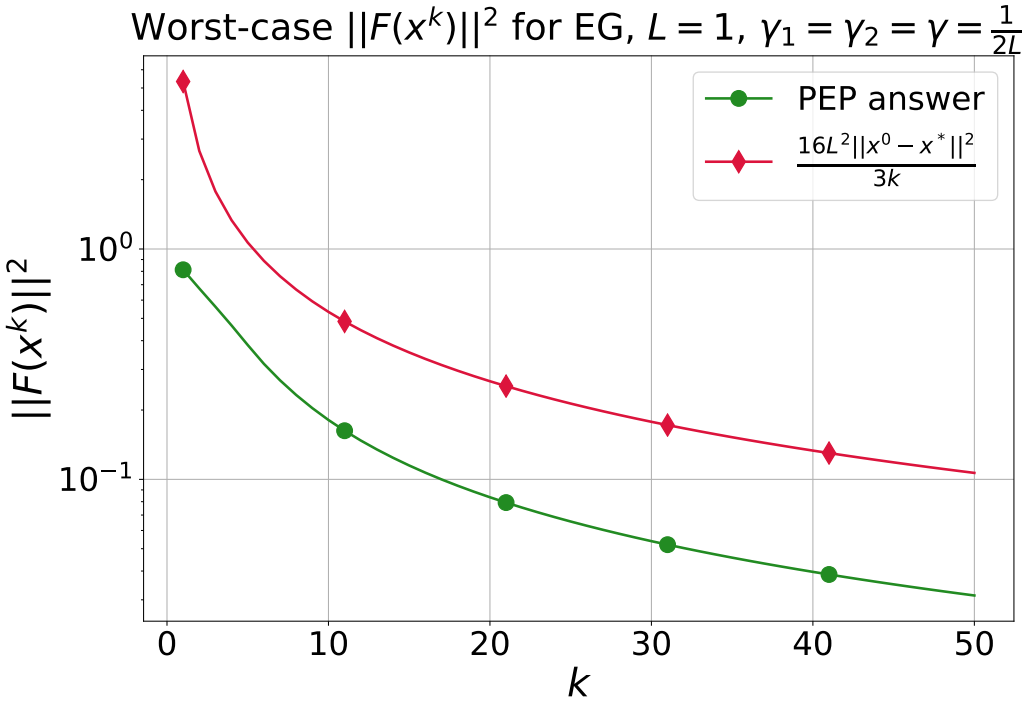


Figure: Comparison of the worst-case rate of EG obtained via solving PEP and the guessed upper-bound $16L^2\|x^0-x^*\|^2/k$. Vertical axis is shown in logarithmic scale and after iteration $k = 20$ the curves are almost parallel, i.e., PEP answer and $16L^2\|x^0-x^*\|^2/k$ differ almost by a constant factor. In view of Proposition 3 from Ryu et al. [2020], PEP may give the answer that is not tight for the class of monotone and Lipschitz operators. However, in this particular case, it turns out to be quite tight.

# The Recipe for Deriving the Proof

Using standard duality theory for SDP [De Klerk, 2006] one can show that the solution of the dual problem to the SDP obtained from (6) gives the proof of convergence.

# The Recipe for Deriving the Proof

Using standard duality theory for SDP [De Klerk, 2006] one can show that the solution of the dual problem to the SDP obtained from (6) gives the proof of convergence.

The recipe [De Klerk et al., 2017]:

- Solve the dual problem numerically for different parameters of the problem

# The Recipe for Deriving the Proof

Using standard duality theory for SDP [De Klerk, 2006] one can show that the solution of the dual problem to the SDP obtained from (6) gives the proof of convergence.

The recipe [De Klerk et al., 2017]:

- Solve the dual problem numerically for different parameters of the problem
- Guess the analytical form of the dual solution

# The Recipe for Deriving the Proof

Using standard duality theory for SDP [De Klerk, 2006] one can show that the solution of the dual problem to the SDP obtained from (6) gives the proof of convergence.

The recipe [De Klerk et al., 2017]:

- Solve the dual problem numerically for different parameters of the problem
- Guess the analytical form of the dual solution
- Sum up the constraints of the primal problem with weights corresponding to the solution of the dual problem

# Example of the Proof [De Klerk et al., 2017]

Set $f_i = f(\mathbf{x}_i)$ and $\mathbf{g}_i = \nabla f(\mathbf{x}_i)$ for $i \in \{*, 0, 1\}$. Note that $\mathbf{g}_* = \mathbf{0}$. The following five inequalities are now satisfied:

$$
\begin{aligned}
1: &\quad f_0 \geq f_1 + \mathbf{g}_1^\top(\mathbf{x}_0 - \mathbf{x}_1) + \frac{1}{2(1-\mu/L)}\left(\frac{1}{L}\|\mathbf{g}_0 - \mathbf{g}_1\|^2 + \mu\|\mathbf{x}_0 - \mathbf{x}_1\|^2 - 2\frac{\mu}{L}(\mathbf{g}_1 - \mathbf{g}_0)^\top(\mathbf{x}_1 - \mathbf{x}_0)\right)\\
2: &\quad f_* \geq f_0 + \mathbf{g}_0^\top(\mathbf{x}_* - \mathbf{x}_0) + \frac{1}{2(1-\mu/L)}\left(\frac{1}{L}\|\mathbf{g}_* - \mathbf{g}_0\|^2 + \mu\|\mathbf{x}_* - \mathbf{x}_0\|^2 - 2\frac{\mu}{L}(\mathbf{g}_0 - \mathbf{g}_*)^\top(\mathbf{x}_0 - \mathbf{x}_*)\right)\\
3: &\quad f_* \geq f_1 + \mathbf{g}_1^\top(\mathbf{x}_* - \mathbf{x}_1) + \frac{1}{2(1-\mu/L)}\left(\frac{1}{L}\|\mathbf{g}_* - \mathbf{g}_1\|^2 + \mu\|\mathbf{x}_* - \mathbf{x}_1\|^2 - 2\frac{\mu}{L}(\mathbf{g}_1 - \mathbf{g}_*)^\top(\mathbf{x}_1 - \mathbf{x}_*)\right)\\
4: &\quad -\mathbf{g}_0^\top\mathbf{g}_1 \geq 0\\
5: &\quad \mathbf{g}_1^\top(\mathbf{x}_0 - \mathbf{x}_1) \geq 0.
\end{aligned}
$$

Indeed, the first three inequalities are the $\mathcal{F}_{\mu,L}$-interpolability conditions, the fourth inequality is a relaxation of (4), and the fifth inequality is a relaxation of (3).

We aggregate these five inequalities by defining the following positive multipliers,

$$
y_1 = \frac{L-\mu}{L+\mu}, \quad y_2 = 2\mu\frac{(L-\mu)}{(L+\mu)^2}, \quad y_3 = \frac{2\mu}{L+\mu}, \quad y_4 = \frac{2}{L+\mu}, \quad y_5 = 1, \tag{9}
$$

and adding the five inequalities together after multiplying each one by the corresponding multiplier.

The result is the following inequality (as may be verified directly):

$$
\begin{aligned}
f_1 - f_* \quad \leq \quad & \left(\tfrac{L-\mu}{L+\mu}\right)^2 (f_0 - f_*) - \tfrac{\mu L(L+3\mu)}{2(L+\mu)^2}\left\|\mathbf{x}_0 - \tfrac{L+\mu}{L+3\mu}\mathbf{x}_1 - \tfrac{2\mu}{L+3\mu}\mathbf{x}_* - \tfrac{3L+\mu}{L^2+3\mu L}\mathbf{g}_0 - \tfrac{L+\mu}{L^2+3\mu L}\mathbf{g}_1\right\|^2\\
& - \tfrac{2L\mu^2}{L^2+2L\mu-3\mu^2}\left\|\mathbf{x}_1 - \mathbf{x}_* - \tfrac{(L-\mu)^2}{2\mu L(L+\mu)}\mathbf{g}_0 - \tfrac{L+\mu}{2\mu L}\mathbf{g}_1\right\|^2.
\end{aligned} \tag{10}
$$

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

- However, guessing the dependencies is not always an easy task: the dependencies on the parameters of the problem like $L, \gamma_1, \gamma_2$ might be quite tricky

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

- However, guessing the dependencies is not always an easy task: the dependencies on the parameters of the problem like $L, \gamma_1, \gamma_2$ might be quite tricky
- To simplify the process of guessing the proof, we consider a simpler problem:

$$\begin{aligned} \Delta_{\text{EG}}(L, \gamma_1, \gamma_2) = \quad & \max \quad \|F(x^1)\|^2 - \|F(x^0)\|^2 \\ & \text{s.t.} \quad F \text{ is monotone and } L\text{-Lipschitz}, \ x^0 \in \mathbb{R}^d, \\ & \|x^0 - x^*\|^2 \leq 1, \\ & x^1 = x^0 - \gamma_2 F\left(x^0 - \gamma_1 F(x^0)\right) \end{aligned} \tag{11}$$

with $\gamma_1 = \gamma_2 = \gamma$

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

- In the numerical tests, we observed that $\Delta_{\text{EG}}(L, \gamma_1, \gamma_2) \approx 0$ for all tested pairs of $L$ and $\gamma$

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

- In the numerical tests, we observed that $\Delta_{\text{EG}}(L, \gamma_1, \gamma_2) \approx 0$ for all tested pairs of $L$ and $\gamma$
- Moreover, the dual variables $\lambda_1, \lambda_2, \lambda_3$ that correspond to the constraints

$$0 \leq \frac{1}{\gamma}\langle F(x^k) - F(x^{k+1}), x^k - x^{k+1}\rangle,$$

$$0 \leq \frac{1}{\gamma}\langle F(x^k - \gamma F(x^k)) - F(x^{k+1}), x^k - \gamma F(x^k) - x^{k+1}\rangle,$$

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

are always close to the constants $2, 1/2$, and $3/2$

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

- In the numerical tests, we observed that $\Delta_{\text{EG}}(L, \gamma_1, \gamma_2) \approx 0$ for all tested pairs of $L$ and $\gamma$
- Moreover, the dual variables $\lambda_1, \lambda_2, \lambda_3$ that correspond to the constraints

$$0 \leq \frac{1}{\gamma}\langle F(x^k) - F(x^{k+1}), x^k - x^{k+1}\rangle,$$

$$0 \leq \frac{1}{\gamma}\langle F(x^k - \gamma F(x^k)) - F(x^{k+1}), x^k - \gamma F(x^k) - x^{k+1}\rangle,$$

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

are always close to the constants $2, 1/2$, and $3/2$

  - Although $\lambda_2$ and $\lambda_3$ were sometimes slightly smaller, e.g., sometimes we had $\lambda_2 \approx 3/5$ and $\lambda_3 \approx 13/20$, we simplified these dependencies and simply summed up the corresponding inequalities with weights $\lambda_1 = 2$, $\lambda_2 = 1/2$ and $\lambda_3 = 3/2$ respectively

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

- In the numerical tests, we observed that $\Delta_{\text{EG}}(L, \gamma_1, \gamma_2) \approx 0$ for all tested pairs of $L$ and $\gamma$
- Moreover, the dual variables $\lambda_1, \lambda_2, \lambda_3$ that correspond to the constraints

$$0 \leq \frac{1}{\gamma}\langle F(x^k) - F(x^{k+1}), x^k - x^{k+1}\rangle,$$

$$0 \leq \frac{1}{\gamma}\langle F(x^k - \gamma F(x^k)) - F(x^{k+1}), x^k - \gamma F(x^k) - x^{k+1}\rangle,$$

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

  are always close to the constants $2, 1/2$, and $3/2$
  - Although $\lambda_2$ and $\lambda_3$ were sometimes slightly smaller, e.g., sometimes we had $\lambda_2 \approx 3/5$ and $\lambda_3 \approx 13/20$, we simplified these dependencies and simply summed up the corresponding inequalities with weights $\lambda_1 = 2$, $\lambda_2 = 1/2$ and $\lambda_3 = 3/2$ respectively
  - After that it was just needed to rearrange the terms and apply Young's inequality to some inner products.

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

## Theorem 6

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

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

## Theorem 6

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

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

## Theorem 7

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

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

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

$$\texttt{Gap}_F(x^K) = \max_{y \in \mathbb{R}^d : \|y - x^*\| \leq \|x^0 - x^*\|} \langle F(y), x^K - y \rangle \leq \frac{2\|x^0 - x^*\|^2}{\gamma\sqrt{1 - L^2\gamma^2}\sqrt{K+1}}. \tag{13}$$

# In the Paper We Also Have

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

# References I

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

D. Balduzzi, S. Racaniere, J. Martens, J. Foerster, K. Tuyls, and T. Graepel. The mechanics of n-player differentiable games. In *International Conference on Machine Learning*, pages 354–363. PMLR, 2018.

H. H. Bauschke, P. L. Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

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

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

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

# References II

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

M. Fazel, H. Hindi, and S. P. Boyd. Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices. In *Proceedings of the 2003 American Control Conference, 2003.*, volume 3, pages 2156–2162. IEEE, 2003.

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

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

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

# References III

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

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

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

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

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

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

# References IV

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

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

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

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

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

# References V

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

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

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

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

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

# References VI

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

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

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

A. B. Taylor, J. M. Hendrickx, and F. Glineur. Performance estimation toolbox (pesto): automated worst-case analysis of first-order optimization methods. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 1278–1283. IEEE, 2017a.

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

# Details on SDP and Its Dual

- **Primal problem.** For given symmetric matrices $\mathbf{C}, \mathbf{A}_1, \ldots, \mathbf{A}_m \in \mathbb{S}^n$, vectors $a_1, \ldots, a_m \in \mathbb{R}^l$, and numbers $b_1, \ldots, b_m \in \mathbb{R}$, we consider a primal SDP:

$$\begin{aligned} \max_{\mathbf{X} \in \mathbb{S}^n, u \in \mathbb{R}^l} \quad & \operatorname{Tr}(\mathbf{C}\mathbf{X}) + c^\top u \\ \text{s.t.} \quad & \operatorname{Tr}(\mathbf{A}_k \mathbf{X}) + a_k^\top u \leq b_k \quad \text{for } k = 1, \ldots, m \\ & \mathbf{X} \succeq 0 \end{aligned}$$

- **Dual problem** can be written as (for $b = (b_1, \ldots, b_m)^\top \in \mathbb{R}^m$)

$$\begin{aligned} \min_{y \in \mathbb{R}^m} \quad & b^\top y \\ \text{s.t.} \quad & \sum_{k=1}^m y_k \mathbf{A}_k - \mathbf{C} \succeq 0 \quad \text{and} \quad \sum_{k=1}^m y_k a_k = c \\ & y \geq 0 \text{ (component-wise)} \end{aligned}$$

# Details on SDP and Its Dual

- **Strong duality.** For PEPs one can prove

$$\mathrm{Tr}(\mathbf{C}\mathbf{X}^*) + c^\top u^* = b^\top y^*$$

- Summing up the constraints from the primal problem with weights $y_1^*, \ldots, y_m^*$ we get

$$\sum_{k=1}^m y_k^* \left(\mathrm{Tr}(\mathbf{A}_k\mathbf{X}) + a_k^\top u\right) - b^\top y^* \leq 0,$$

which is equivalent to

$$\mathrm{Tr}\left(\left(\sum_{k=1}^m y_k^* \mathbf{A}_k\right)\mathbf{X}\right) + \left(\sum_{k=1}^m y_k^* a_k\right)^\top u - b^\top y^* \leq 0,$$

# Details on SDP and Its Dual

- For any $\mathbf{A} \succeq 0$ and $\mathbf{B} \succeq 0$ we have $\mathrm{Tr}(\mathbf{AB}) \geq 0$
- Since $\sum\limits_{k=1}^{m} y_k^* \mathbf{A}_k - \mathbf{C} \succeq 0$, we have $\mathrm{Tr}\left(\left(\sum\limits_{k=1}^{m} y_k^* \mathbf{A}_k\right)\mathbf{X}\right) \geq \mathrm{Tr}(\mathbf{CX})$
- Putting all together, **we derive**

$$\mathrm{Tr}(\mathbf{CX}) + c^\top u \leq b^\top y^* = \mathrm{Tr}(\mathbf{CX}^*) + c^\top u^*$$

The result is trivial but the **derivation** gives a recipe of getting the proof!