Extragradient Method: $\mathcal{O}(1/K)$ 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

MTL MLOpt Internal Meeting

December 1, 2021

Outline

Variational Inequality Problem

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

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

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

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

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

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

How to Solve VIP?

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

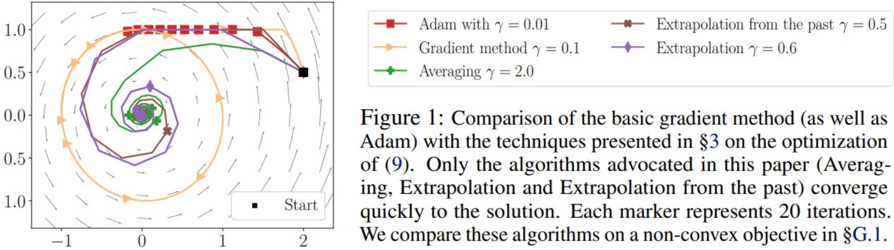


Figure 1: Comparison of the basic gradient method (as well as Adam) with the techniques presented in §3 on the optimization of (9). Only the algorithms advocated in this paper (Averaging, Extrapolation and Extrapolation from the past) converge quickly to the solution. Each marker represents 20 iterations.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

Figure: Behavior of GD on the problem $\min\limits_{u\in\mathbb{R}} \max\limits_{v\in\mathbb{R}} uv$ [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]

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

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

Measures of Convergence

- **Restricted gap function**: $\mathtt{Gap}_F(x^K) = \max_{y\in\mathbb{R}^d : \|y-x^*\|\le R} \langle F(y), x^K - y\rangle$, 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 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:

- **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]
  - $\min\limits_{k=0,1,\ldots,K} \|F(x^k)\|^2 = \mathcal{O}(1/K)$[Solodov and Svaiter, 1999, Ryu et al., 2019]
- **Lower bounds for the last-iterate [Golowich et al., 2020]:**
  - $\mathtt{Gap}_F(x^K) = \Omega(1/\sqrt{K})$
  - $\|F(x^K)\|^2 = \Omega(1/K)$
- **Upper bounds for the last-iterate [Golowich et al., 2020]:** *if additionally the Jacobian $\nabla F(x)$ is Λ-Lipschitz*, then
  - $\mathtt{Gap}_F(x^K) = \mathcal{O}(1/\sqrt{K})$
  - $\|F(x^K)\|^2 = \mathcal{O}(1/K)$

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 will give a positive answer to this question further in this talk!

Cocoercivity

Operator $F : \mathbb{R}^d \to \mathbb{R}^d$ is called ℓ-cocoercive if for all $x, y \in \mathbb{R}^d$

$$\|F(x) - F(y)\|^2 \leq \ell\langle F(x) - F(y), x - y\rangle \tag{5}$$

- F is ℓ-cocoercive $\Longrightarrow$ F is monotone and ℓ-Lipschitz
- F is monotone and ℓ-Lipschitz $\not\Longleftarrow$ F is ℓ-cocoercive
  - Counter-example: F corresponding to bilinear game $\min\limits_{u\in\mathbb{R}^{d_1}} \max\limits_{v\in R^{d_2}} x^\top Ay$
  - If $F = \nabla f$, then monotonicity and ℓ-Lipschitzness of F implies that F is ℓ-cocoercive

Operator $F : \mathbb{R}^d \to \mathbb{R}^d$ is called ℓ-star-cocoercive if for all $x \in \mathbb{R}^d$

$$\|F(x)\|^2 \leq \ell\langle F(x), x - x^*\rangle, \tag{6}$$

where x^* is such that $F(x^*) = 0$

GD Converges Under Star-Cocoercivity

Theorem 1 (Random-iterate convergence of GD)

Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be ℓ-**star-cocoercive**. Then for all $K \geq 0$ we have

$$\mathbb{E}\|F(\widehat{x}^K)\|^2 \leq \frac{\ell\|x^0 - x^*\|^2}{\gamma(K+1)}, \tag{7}$$

where $\widehat{x}^K$ is chosen uniformly at random from the set of iterates $\{x^0, x^1, \ldots, x^K\}$ produced by GD with $0 < \gamma \leq 1/\ell$.

... and the proof is trivial!

GD Converges Under Star-Cocoercivity

Proof of Theorem 1

Using the update rule of (GD) we derive

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &= \|x^k - \gamma F(x^k) - x^*\|^2 \\ &= \|x^k - x^*\|^2 - 2\gamma\langle x^k - x^*, F(x^k)\rangle + \gamma^2\|F(x^k)\|^2 \\ &\overset{(6)}{\leq} \|x^k - x^*\|^2 - \gamma\left(\frac{2}{\ell} - \gamma\right)\|F(x^k)\|^2. \end{aligned}$$

Rearranging the terms we get

$$\gamma\left(\frac{2}{\ell} - \gamma\right)\|F(x^k)\|^2 \leq \|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2. \tag{8}$$

It remains to average the above inequalitites for $k = 0, 1, \ldots, K$.

GD Converges Under Cocoercivity

Theorem 2 (Last-iterate convergence of GD)

Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be ℓ-**cocoercive**. Then for all $K \geq 0$ we have

$$\|F(x^K)\|^2 \leq \frac{\ell\|x^0 - x^*\|^2}{\gamma(K+1)}, \tag{9}$$

where x^K is produced by GD with $0 < \gamma \leq 1/\ell$.

The proof is also simple and consist of two steps:

1. Derivation of $\|F(x^{k+1})\| \leq \|F(x^k)\|$ using ℓ-cocoercivity at x^k and x^{k+1}
2. Application of the above inequality to the previous result

Idea: EG = GD with Special Operator

$$x^{k+1} = x^k - \gamma_2 \underbrace{F\left(x^k - \gamma_1 F(x^k)\right)}_{F_{\mathbf{EG},\gamma_1}(x^k)} = x^k - \gamma_2 F_{\mathrm{EG},\gamma_1}(x^k) \tag{EG}$$

Key idea: if we manage to show that $F_{\mathrm{EG},\gamma_1}(x^k)$ is ℓ-cocoercive with some $\ell > 0$ for any monotone and L-Lipschitz F and for a reasonable choice of γ_1,then we can simply apply the results for GD and we will get the desired last-iterate $\mathcal{O}(1/K)$ convergence rate.

Useful Facts on Cocoercivity

Lemma 1 (Proposition 4.2 from Bauschke et al. [2011])

For any operator $F : \mathbb{R}^d \to \mathbb{R}^d$ the following are equivalent

(i) $\mathrm{Id} - \frac{2}{\ell} F$ is non-expansive.

(ii) F is ℓ-cocoercive.

Lemma 2

For any operator $F : \mathbb{R}^d \to \mathbb{R}^d$ and x^* such that $F(x^*) = 0$ the following are equivalent:

(i) $\mathrm{Id} - \frac{2}{\ell} F$ is non-expansive around[a] x^*.

(ii) F is ℓ-star-cocoercive.

[a] Operator $U : \mathbb{R}^d \to \mathbb{R}^d$ is called non-expansive around x^* if for all $x \in \mathbb{R}^d$ it satisfies $\|U(x) - U(x^*)\| \leq \|x - x^*\|$.

Warm-up: Proximal-Point Method

Consider a simpler for the analysis but much less practical method called Proximal Point method (PP) [Martinet, 1970, Rockafellar, 1976]:

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

- x^{k+1} is defined implicitly for given x^k and $\gamma > 0$
- Define operator $F_{\mathrm{PP},\gamma} : \mathbb{R}^d \to \mathbb{R}^d$ such that $\forall x \in \mathbb{R}^d$

$$F_{\mathrm{PP},\gamma}(x) = F(y), \quad \text{where} \quad y = x - \gamma F(y) \tag{10}$$

- (PP) can be rewritten as GD for $F_{\mathrm{PP},\gamma}$:

$$x^{k+1} = x^k - \gamma F_{\mathrm{PP},\gamma}(x^k)$$

Warm-up: Proximal-Point Operator is Cocoercive

Theorem 3

Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be monotone and $\gamma > 0$. Then, $F_{\mathrm{PP},\gamma}(x)$ is $2/\gamma$-cocoercive.

Proof of Theorem 3

In view of Lemma 1, it is enough to prove that $\mathrm{Id} - \gamma F_{\mathrm{PP},\gamma}$ is non-expansive. Consider arbitrary $x, y \in \mathbb{R}^d$ and define $\widehat{x}$ and $\widehat{y}$ as follows:

$$\widehat{x} = x - \gamma F(\widehat{x}) = x - \gamma F_{\mathrm{PP},\gamma}(x), \quad \widehat{y} = y - \gamma F(\widehat{y}) = y - \gamma F_{\mathrm{PP},\gamma}(y).$$

Warm-up: Proximal-Point Operator is Cocoercive

Proof of Theorem 3

Using this notation, we derive

$$\begin{aligned}
\|\widehat{x}-\widehat{y}\|^2 &= \|x-y\|^2 - 2\gamma\langle x-y, F(\widehat{x})-F(\widehat{y})\rangle + \gamma^2\|F(\widehat{x})-F(\widehat{y})\|^2 \\
&= \|x-y\|^2 - 2\gamma\langle \widehat{x}+\gamma F(\widehat{x})-\widehat{y}-\gamma F(\widehat{y}), F(\widehat{x})-F(\widehat{y})\rangle \\
&\quad +\gamma^2\|F(\widehat{x})-F(\widehat{y})\|^2 \\
&= \|x-y\|^2 - 2\gamma\langle \widehat{x}-\widehat{y}, F(\widehat{x})-F(\widehat{y})\rangle - \gamma^2\|F(\widehat{x})-F(\widehat{y})\|^2 \\
&\overset{(2)}{\leq} \|x-y\|^2 - \gamma^2\|F(\widehat{x})-F(\widehat{y})\|^2 \\
&\leq \|x-y\|^2.
\end{aligned}$$

That is, $\mathrm{Id} - \gamma F_{\mathrm{PP},\gamma}$ is non-expansive, and, as a result, $F_{\mathrm{PP},\gamma}$ is $2/\gamma$-cocoercive.

Warm-up: Last-Iterate Convergence of PP

Applying Theorem 2, we derive last-iterate $\mathcal{O}(1/K)$ convergence rate for

$$x^{k+1} = x^k - \gamma F_{\mathrm{PP},2/\ell}(x^k) \tag{PP-γ-ℓ}$$

Theorem 4

Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be monotone, $\ell > 0$ and $0 < \gamma \leq 1/\ell$. Then for all $K \geq 0$ we have

$$\|F(\widehat{x}^K)\|^2 \leq \frac{\ell\|x^0 - x^*\|^2}{\gamma(K+1)}, \tag{11}$$

where $\widehat{x}^K = x^K - 2/\ell F(\widehat{x}^K) = x^K - 2/\ell F_{\mathrm{PP},2/\ell}(\widehat{x}^K)$ and x^K is produced by (PP-γ-ℓ).

EG is "an Approximation" of PP

$$
\begin{array}{rl}
\text{PP}: \; x^{k+1} = & x^k - \gamma F(x^{k+1}) = x^k - \gamma F_{\text{PP},\gamma}(x^k) \\
\text{EG}: \; x^{k+1} = & x^k - \gamma F\left(x^k - \gamma F(x^k)\right) = x^k - \gamma F_{\text{EG},\gamma}(x^k)
\end{array}
$$

- **Informal explanation:** gradient step $x^k - \gamma F(x^k)$ "approximates" the next point x^{k+1}
- **Formal explanation:** if F is L-Lipschitz and x^{k+1} is obtained via EG, then

$$
\begin{aligned}
\left\|F(x^{k+1}) - F\left(x^k - \gamma F(x^k)\right)\right\| &\leq L\|x^{k+1} - x^k - \gamma F(x^k)\| \\
&= L\gamma \left\|F\left(x^k - \gamma F(x^k)\right) - F(x^k)\right\| \\
&\leq L^2\gamma^2\|F(x^k)\|,
\end{aligned}
$$

so, for the difference between update directions decreases quadratically in γ

EG and Cocoercivity: Resolved Question

Q2: *Is operator $F_{\mathrm{EG},\gamma}$ cocoercive when F is monotone and L-Lipschitz?*

We give the following answer: in some cases it is true,
but in general it is not the case!

EG and Cocoercivity: What We Obtained

Assume that F is monotone and L-Lipschitz and consider

$$x^{k+1} = x^k - \gamma_2 \underbrace{F\left(x^k - \gamma_1 F(x^k)\right)}_{F_{\mathbf{EG},\gamma_1}(x^k)} = x^k - \gamma_2 F_{\mathrm{EG},\gamma_1}(x^k)$$

✓ If F is linear, i.e., for any $\alpha, \beta \in \mathbb{R}$ and $x, y \in \mathbb{R}^d$ the operator satisfies $F(\alpha x + \beta y) = \alpha F(x) + \beta F(y)$, then operator $F_{\mathrm{EG},\gamma_1}(x)$ with $\gamma_1 \leq 1/L$ is $2/\gamma_1$-cocoercive $\Longrightarrow \|F_{\mathrm{EG},\gamma_1}(x^K)\|^2 = \mathcal{O}(1/K)$

✓ If $F(x) = \mathrm{A}x + b$ for some $\mathrm{A} \in \mathbb{R}^{d\times d}, b \in \mathbb{R}^d$, then operator $F_{\mathrm{EG},\gamma_1}(x)$ with $\gamma_1 \leq 1/L$ is $2/\gamma_1$-cocoercive $\Longrightarrow \|F_{\mathrm{EG},\gamma_1}(x^K)\|^2 = \mathcal{O}(1/K)$

✓✗ If $F(x)$ is not necessarily affine but is star-monotone, i.e., $\langle F(x), x - x^* \rangle \geq 0$ for all $x \in \mathbb{R}^d$, then operator $F_{\mathrm{EG},\gamma_1}(x)$ with $\gamma_1 \leq 1/L$ is $2/\gamma_1$-star-cocoercive $\Longrightarrow \min\limits_{k=0,1,\ldots,K} \|F_{\mathrm{EG},\gamma_1}(x^k)\|^2 = \mathcal{O}(1/K)$

Proofs are relatively simple and based mainly on Lemmas 1 and 2

EG and Cocoercivity: What Else We Obtained

Assume that F is monotone and L-Lipschitz and consider

$$x^{k+1} = x^k - \gamma_2 \underbrace{F\left(x^k - \gamma_1 F(x^k)\right)}_{F_{\mathbf{EG},\gamma_1}(x^k)} = x^k - \gamma_2 F_{\mathrm{EG},\gamma_1}(x^k)$$

✗ For all $L > 0$ and $\gamma_1 \in (0, 1/L]$ there exists a monotone and L-Lipschitz operator F such that operator $F_{\mathrm{EG},\gamma_1}(x)$ is non-cocoercive

The proof of this fact was obtained via numerical solutions of so-called Performance Estimation Problems (PEP)

Performance Estimation Problems

- 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, my biased personal recomendation: read papers and slides by Adrien Taylor `https://www.di.ens.fr/~ataylor`

Performance Estimation Problem and Expansiveness

- In view of Lemma 1, it is sufficient to show that for any $L > 0$ and any $\gamma_1, \gamma_2 > 0$ there exists a monotone and L-Lipschitz operator F such that $\mathrm{Id} - \gamma_2 F_{\mathrm{EG},\gamma_1}$ is not non-expansive
- In other words, our goal is to show that for all $L, \gamma_1, \gamma_2 > 0$ the quantity

$$\begin{aligned} \rho_{\mathrm{EG}}(L, \gamma_1, \gamma_2) = \quad \max \quad & \frac{\|\widehat{x} - \widehat{y}\|^2}{\|x - y\|^2} \\ \text{s.t.} \quad & F \text{ is mon. \& } L\text{-Lip.}, \\ & x, y \in \mathbb{R}^d,\ x \neq y, \\ & \widehat{x} = x - \gamma_2 F(x - \gamma_1 F(x)), \\ & \widehat{y} = y - \gamma_2 F(y - \gamma_1 F(y)) \end{aligned} \tag{12}$$

is bigger than 1, i.e., $\rho_{\mathrm{EG}}(L, \gamma_1, \gamma_2) > 1$.

PEP for Expansiveness

- Problem (15) is hard to solve since it is infinitely dimensional
- Let us try to come up with an equivalent finite-deminsional formulation. **Naive idea №1:** consider the following problem

$$\begin{aligned} \max \quad & \frac{\|x - \gamma_2 x_{F_2} - y + \gamma_2 y_{F_2}\|^2}{\|x-y\|^2} \\ \text{s.t.} \quad & F \text{ is mon. \& } L\text{-Lip.}, \; x, y \in \mathbb{R}^d, \; x \neq y, \\ & x_{F_2} = F(x - \gamma_1 x_{F_1}), \; x_{F_1} = F(x), \\ & y_{F_2} = F(y - \gamma_1 y_{F_1}), \; y_{F_1} = F(y) \end{aligned} \tag{13}$$

- It is equivalent to (12) but the new problem is finite-dimensional. However, it is stil unclear how to check that there exists a monotone and L-Lipschitz operator F such that $F(x) = x_{F_1}, F(y) = y_{F_1}, F(x - \gamma_1 x_{F_1}) = x_{F_2}, F(y - \gamma_1 y_{F_1}) = y_{F_2}$

PEP for Expansiveness

- **Naive idea №2:** consider the following problem

$$
\begin{aligned}
\max \quad & \frac{\|x - \gamma_2 x_{F_2} - y + \gamma_2 y_{F_2}\|^2}{\|x - y\|^2} \\
\text{s.t.} \quad & \|z_1 - z_1'\|^2 \le L^2 \|z - z'\|^2, \\
& \langle z_1 - z_1', z - z' \rangle \ge 0, \\
& \text{for each two pairs } (z, z_1), (z', z_1') \\
& \text{from } \{(x, x_{F_1}), (y, y_{F_1}), (x - \gamma_1 x_{F_1}, x_{F_2}), (y - \gamma_1 y_{F_1}, y_{F_2})\}
\end{aligned}
\tag{14}
$$

- **Bad news:** problem (14) is not equivalent to (13) [Ryu et al., 2020]: feasible set in (14) contains some points that are not feasible for (13), i.e., some feasible points for (14) cannot be interpolated by any monotone and L-Lipschitz operator.

PEP for Expansiveness

- **Good news:** one can circumvent this issue if we focus on a different problem. Let us try to show that for any $\ell > 0$ and any $\gamma_1, \gamma_2 > 0$ there exists a ℓ-cocoercive operator F such that $\mathrm{Id} - \gamma F_{\mathsf{EG},\gamma_1}$ is not non-expansive.
- In other words, our goal is to show that for all $\ell, \gamma_1, \gamma_2 > 0$ the quantity

$$\begin{aligned} \rho_{\mathsf{EG}}(\ell, \gamma_1, \gamma_2) = \quad \max \quad & \frac{\|\widehat{x} - \widehat{y}\|^2}{\|x - y\|^2} \\ \text{s.t.} \quad & F \text{ is } \ell\text{-cocoercive}, \\ & x, y \in \mathbb{R}^d,\ x \neq y, \\ & \widehat{x} = x - \gamma_2 F(x - \gamma_1 F(x)), \\ & \widehat{y} = y - \gamma_2 F(y - \gamma_1 F(y)) \end{aligned} \tag{15}$$

is bigger than 1, i.e., $\rho_{\mathsf{EG}}(\ell, \gamma_1, \gamma_2) > 1$.

PEP for Expansiveness

- Consider an equivalent finite-dimensional problem:

$$\begin{aligned} \max \quad & \frac{\|x - \gamma_2 x_{F_2} - y + \gamma_2 y_{F_2}\|^2}{\|x - y\|^2} \\ \text{s.t.} \quad & F \text{ is } \ell\text{-cocoercive},\ x, y \in \mathbb{R}^d,\ x \neq y, \\ & x_{F_2} = F(x - \gamma_1 x_{F_1}),\ x_{F_1} = F(x), \\ & y_{F_2} = F(y - \gamma_1 y_{F_1}),\ y_{F_1} = F(y). \end{aligned} \tag{16}$$

- Next, for all $\alpha > 0$ the following equivalence holds:

$$F \text{ is } \ell\text{-cocoercive} \quad \Longleftrightarrow \quad \left(\alpha^{-1}\mathrm{Id}\right) \circ F \circ \left(\alpha \mathrm{Id}\right) \text{ is } \ell\text{-cocoercive.}$$

PEP for Expansiveness

- Therefore, in problem (16) one can apply the change of variables

$$x := \alpha^{-1}x, \quad y := \alpha^{-1}y, \quad x_{F_1} := \alpha^{-1}x_{F_1}, \quad y_{F_1} := \alpha^{-1}y_{F_1},$$
$$x_{F_2} := \alpha^{-1}x_{F_2}, \quad y_{F_2} := \alpha^{-1}y_{F_2}, \quad F := \left(\alpha^{-1}\mathrm{Id}\right) \circ F \circ (\alpha\mathrm{Id}),$$

where $\alpha = \|x - y\|$, and get another equivalent problem

$$\begin{aligned} \max \quad & \|x - \gamma_2 x_{F_2} - y + \gamma_2 y_{F_2}\|^2 \\ \text{s.t.} \quad & F \text{ is } \ell\text{-cocoercive}, \ x, y \in \mathbb{R}^d, \ \|x - y\| = 1, \\ & x_{F_2} = F(x - \gamma_1 x_{F_1}), \ x_{F_1} = F(x), \\ & y_{F_2} = F(y - \gamma_1 y_{F_1}), \ y_{F_1} = F(y). \end{aligned} \tag{17}$$

PEP for Expansiveness

- Proposition 2 from Ryu et al. [2020] implies that (17) is equivalent to

$$
\begin{aligned}
\max \quad & \|x - \gamma_2 x_{F_2} - y + \gamma_2 y_{F_2}\|^2 \\
\text{s.t.} \quad & x, y, x_{F_1}, y_{F_1}, x_{F_2}, y_{F_2} \in \mathbb{R}^d,\ \|x - y\|^2 = 1, \\
& \ell\langle x_{F_1} - x_{F_2}, \gamma_1 x_{F_1}\rangle \geq \|x_{F_1} - x_{F_2}\|^2, \\
& \ell\langle x_{F_1} - y_{F_1}, x - y\rangle \geq \|x_{F_1} - y_{F_1}\|^2, \\
& \ell\langle x_{F_1} - y_{F_2}, x - y + \gamma_1 y_{F_1}\rangle \geq \|x_{F_1} - y_{F_2}\|^2, \\
& \ell\langle x_{F_2} - y_{F_1}, x - \gamma_1 x_{F_1} - y\rangle \geq \|x_{F_2} - y_{F_1}\|^2, \\
& \ell\langle x_{F_2} - y_{F_2}, x - \gamma_1 x_{F_1} - y + \gamma_1 y_{F_1}\rangle \geq \|x_{F_2} - y_{F_2}\|^2, \\
& \ell\langle y_{F_1} - y_{F_2}, \gamma_1 y_{F_1}\rangle \geq \|y_{F_1} - y_{F_2}\|^2.
\end{aligned}
\tag{18}
$$

The problem is linear in terms of the pairwise inner products of $x, y, x_{F_1}, y_{F_1}, x_{F_2}, y_{F_2}$

PEP for Expansiveness

- Consider a Grammian representation of $(x^\top, y^\top, x_{F_1}^\top, y_{F_1}^\top, x_{F_2}^\top, y_{F_2}^\top)^\top$:

$$G = \begin{pmatrix} x^\top \\ y^\top \\ x_{F_1}^\top \\ y_{F_1}^\top \\ x_{F_2}^\top \\ y_{F_2}^\top \end{pmatrix} \cdot \begin{pmatrix} x & y & x_{F_1} & y_{F_1} & x_{F_2} & y_{F_2} \end{pmatrix}$$

- One can easily show that for all $d \geq 6$

$$G \in \mathbb{S}_+^6 \quad \Longleftrightarrow \quad \exists\, x, y, x_{F_1}, y_{F_1}, x_{F_2}, y_{F_2} \in \mathbb{R}^d : \text{ G is Gram matrix for them}$$

PEP for Expansiveness

- Therefore, problem (18) is equivalent to the following SDP problem:

$$\begin{aligned} \max \quad & \mathrm{Tr}(\mathsf{M}_0\mathsf{G}) \\ \text{s.t.} \quad & \mathsf{G} \in \mathbb{S}^6_+, \\ & \mathrm{Tr}(\mathsf{M}_i\mathsf{G}) \geq 0,\ i = 1, 2, \ldots, 6, \\ & \mathrm{Tr}(\mathsf{M}_7\mathsf{G}) = 1, \end{aligned} \tag{19}$$

where $\mathsf{M}_0, \ldots, \mathsf{M}_7$ are some symmetric matrices.

PEP for Expansiveness: M_0

$$M_0 = \begin{pmatrix} 1 & -1 & 0 & 0 & -\gamma_2 & \gamma_2 \\ -1 & 1 & 0 & 0 & \gamma_2 & -\gamma_2 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -\gamma_2 & \gamma_2 & 0 & 0 & \gamma_2^2 & -\gamma_2^2 \\ \gamma_2 & -\gamma_2 & 0 & 0 & -\gamma_2^2 & \gamma_2^2 \end{pmatrix}$$

PEP for Expansiveness: M_1

$$
M_1 = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \ell\gamma_1 - 1 & 0 & 1 - \frac{\ell\gamma_1}{2} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 - \frac{\ell\gamma_1}{2} & 0 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
$$

PEP for Expansiveness: M_2

$$M_2 = \begin{pmatrix} 0 & 0 & \frac{\ell}{2} & -\frac{\ell}{2} & 0 & 0 \\ 0 & 0 & -\frac{\ell}{2} & \frac{\ell}{2} & 0 & 0 \\ \frac{\ell}{2} & -\frac{\ell}{2} & -1 & 1 & 0 & 0 \\ -\frac{\ell}{2} & \frac{\ell}{2} & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

PEP for Expansiveness: M_3

$$M_3 = \begin{pmatrix} 0 & 0 & \frac{\ell}{2} & 0 & 0 & -\frac{\ell}{2} \\ 0 & 0 & -\frac{\ell}{2} & 0 & 0 & \frac{\ell}{2} \\ \frac{\ell}{2} & -\frac{\ell}{2} & -1 & \frac{\ell\gamma_1}{2} & 0 & 1 \\ 0 & 0 & \frac{\ell\gamma_1}{2} & 0 & 0 & -\frac{\ell\gamma_1}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -\frac{\ell}{2} & \frac{\ell}{2} & 1 & -\frac{\ell\gamma_1}{2} & 0 & -1 \end{pmatrix}$$

PEP for Expansiveness: M_4

$$
M_4 = \begin{pmatrix}
0 & 0 & 0 & -\frac{\ell}{2} & \frac{\ell}{2} & 0 \\
0 & 0 & 0 & \frac{\ell}{2} & -\frac{\ell}{2} & 0 \\
0 & 0 & 0 & \frac{\ell\gamma_1}{2} & -\frac{\ell\gamma_1}{2} & 0 \\
-\frac{\ell}{2} & \frac{\ell}{2} & \frac{\ell\gamma_1}{2} & -1 & 1 & 0 \\
\frac{\ell}{2} & -\frac{\ell}{2} & -\frac{\ell\gamma_1}{2} & 1 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
$$

PEP for Expansiveness: M_5

$$
M_5 = \begin{pmatrix}
0 & 0 & 0 & 0 & \frac{\ell}{2} & -\frac{\ell}{2} \\
0 & 0 & 0 & 0 & -\frac{\ell}{2} & \frac{\ell}{2} \\
0 & 0 & 0 & 0 & -\frac{\ell\gamma_1}{2} & \frac{\ell\gamma_1}{2} \\
0 & 0 & 0 & 0 & \frac{\ell\gamma_1}{2} & -\frac{\ell\gamma_1}{2} \\
\frac{\ell}{2} & -\frac{\ell}{2} & -\frac{\ell\gamma_1}{2} & \frac{\ell\gamma_1}{2} & -1 & 1 \\
-\frac{\ell}{2} & \frac{\ell}{2} & \frac{\ell\gamma_1}{2} & -\frac{\ell\gamma_1}{2} & 1 & -1
\end{pmatrix}
$$

PEP for Expansiveness: M_6

$$M_6 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \ell\gamma_1 - 1 & 0 & 1 - \frac{\ell\gamma_1}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 - \frac{\ell\gamma_1}{2} & 0 & -1 \end{pmatrix}$$

PEP for Expansiveness: M_7

$$M_7 = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

F_{EG,γ_1} is Not Cocoercive: Numerical Proof

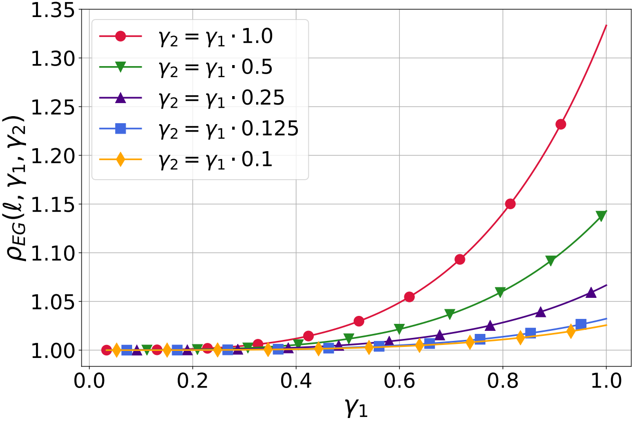


Figure: Numerical estimation of $\rho_{\mathrm{EG}}(\ell, \gamma_1, \gamma_2)$ defined in (15) for $\ell = 1$ and different γ_1, γ_2.

F_{EG,γ_1} is Not Cocoercive?

✓ We obtained the answer numerically for different choices of γ_1 and γ_2

✗ It is not a rigorous proof: probably, for smaller stepsize F_{EG,γ_1} is cocoercive, but we cannot check it because of the numerical inaccuracies

Analytical example is required

How to Construct Analytical Example?

- Try to solve the problem symbolicaly after some simplifications of the problem
  - ✓ Ryu et al. [2020] use this trick and obtained quite impressive results that are almost impossible to obtain by hands
  - ✗ Unfortunately, this approach does not always work and it did not help us to get the example
- Try to solve the problem numerically for different parameters γ_1, γ_2 and ℓ to guess the dependencies using visualization
  - ✓ Gu and Yang [2019] successfuly applied this technique to derive worst-case examples for PP
  - ✗ It is hard to visualize d-dimensional examples with $d \geq 3$

Low-Dimensional Examples: Trace Heuristic

Instead of solving

$$\begin{aligned} \max \quad & \mathrm{Tr}(M_0 G) \\ \text{s.t.} \quad & G \in \mathbb{S}^6_+, \\ & \mathrm{Tr}(M_i G) \geq 0, \; i = 1, 2, \ldots, 6, \\ & \mathrm{Tr}(M_7 G) = 1, \end{aligned}$$

which gives us 5-dimensional examples of G, we consider another problem [Taylor et al., 2017a]:

$$\begin{aligned} \min \quad & \mathrm{Tr}(G) \\ \text{s.t.} \quad & G \in \mathbb{S}^6_+, \\ & \mathrm{Tr}(M_0 G) \geq 1.0005, \\ & \mathrm{Tr}(M_i G) \geq 0, \; i = 1, 2, \ldots, 6, \\ & \mathrm{Tr}(M_7 G) = 1. \end{aligned}$$

Low-Dimensional Examples: Trace Heuristic

$$
\begin{aligned}
\min \quad & \mathrm{Tr}(G) \\
\text{s.t.} \quad & G \in \mathbb{S}^6_+, \\
& \mathrm{Tr}(M_0 G) \geq 1.0005, \\
& \mathrm{Tr}(M_i G) \geq 0,\ i = 1, 2, \ldots, 6,\ \mathrm{Tr}(M_7 G) = 1
\end{aligned}
$$

- It gives low-rank solutions (a heuristic)
- This shows non-$^2/_{\gamma_2}$-cocoercivity of F_{EG,γ_1}: we ensure this via the constraint $\mathrm{Tr}(M_0 G) \geq a = 1.0005 > 1$
- In theory, any $a > 1$ can be used but due to the inevitability of the numerical errors in practice we used $a = 1.0005$
- We obtained solutions of rank 3, i.e., 3-dimensional examples
  - ✗ Unfortunately, visualizations did not help to construct an analytical example

Low-Dimensional Examples: Log-Det Heuristic

To overcome this issue, we consider another problem with so called Log-det heurisctic [Fazel et al., 2003]:

$$\begin{aligned} \min \quad & \log\det\left(\mathrm{G} + \delta \mathrm{I}\right) \\ \text{s.t.} \quad & \mathrm{G} \in \mathbb{S}^6_+, \\ & \mathrm{Tr}(\mathrm{M}_0\mathrm{G}) \geq 1.0005, \\ & \mathrm{Tr}(\mathrm{M}_i\mathrm{G}) \geq 0,\ i = 1, 2, \ldots, 6, \\ & \mathrm{Tr}(\mathrm{M}_7\mathrm{G}) = 1, \end{aligned} \tag{20}$$

where $\delta > 0$ is some small positive regularization parameter. For simplicity we used $\gamma_2 = \gamma_1$ in some interval and $\ell = 1$.

- ✓ We obtained solutions of rank 2, i.e., we obtained $x, y, x_{F_1}, y_{F_1}, x_{F_2}, y_{F_2}$ in $\mathbb{R}^2$
- ✓ We observed that $x = -y$ for all tested values of γ_1

Low-Dimensional Examples: Log-Det Heuristic

✗ However, numerical solutions were not consistent enough to guess the right dependencies

✓ To overcome this issue, we

- rotated $x, y, x_{F_1}, y_{F_1}, x_{F_2}, y_{F_2}$ in such a way that $x = (-1/2, 0)^\top$, $y = (1/2, 0)^\top$,
- plotted the components of $x_{F_1}, y_{F_1}, x_{F_2}, y_{F_2}$ for different γ_1

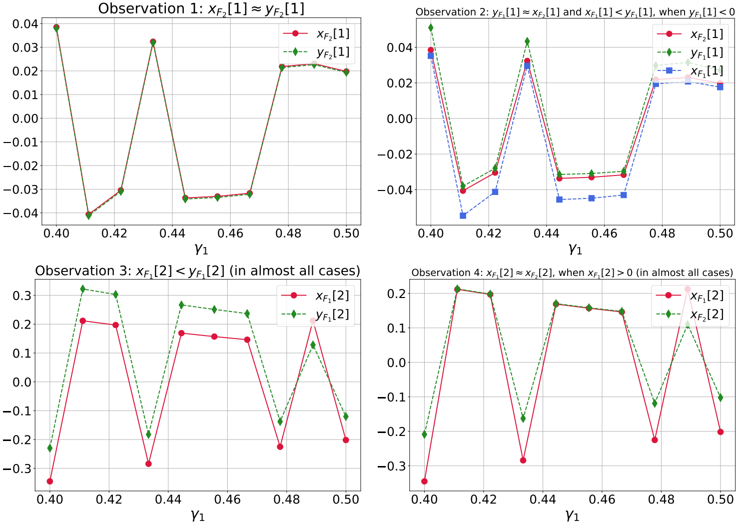
✓ Although the resulting dependencies were not perfect, the obtained plots helped us to sequentially construct the needed example:

$$x = \begin{pmatrix} -\frac{1}{2} \\ 0 \end{pmatrix}, \quad y = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix}, \quad x_{F_1} = \begin{pmatrix} -\frac{1}{2\gamma_1} \\ \frac{1}{2\gamma_1} \end{pmatrix}, \quad y_{F_1} = \begin{pmatrix} -\frac{1-\gamma_1\ell}{2\gamma_1} \\ \frac{1+\gamma_1\ell}{2\gamma_1} \end{pmatrix},$$

$$x_{F_2} = \begin{pmatrix} -\frac{1-\gamma_1\ell}{2\gamma_1} \\ \frac{1}{2\gamma_1} \end{pmatrix}, \quad y_{F_2} = \begin{pmatrix} -\frac{1-\gamma_1\ell}{2\gamma_1} \\ \frac{1-\gamma_1^2\ell^2}{2\gamma_1} \end{pmatrix} \tag{21}$$

- Required several days of playing with plots to get the needed insights

Non-Cocoercivity of F_{EG,γ_1}: Four Observations



Non-Cocoercivity of F_{EG,γ_1}: Four Observations

Mimicking the observed dependencies, we assumed that

$$\begin{gathered} x_{F_2}[1] = y_{F_2}[1], \\ y_{F_1}[1] = x_{F_2}[1] \quad \text{and} \quad x_{F_1}[1] < y_{F_1}[1] < 0, \\ 0 < x_{F_1}[2] < y_{F_1}[2], \\ x_{F_1}[2] = x_{F_2}[2] \end{gathered}$$

Non-Cocoercivity of F_{EG,γ_1}: Handling Four Observations

After that, we plugged them in the interpolation conditions from (18), and obtained the following inequalities:

$$
\begin{aligned}
y_{F_1}[1] &\leq (1-\gamma_1)x_{F_1}[1], \\
y_{F_1}[2] &\leq \frac{y_{F_2}[2]}{1-\gamma_1}, \\
y_{F_1}[2] &\leq (1+\gamma_1)x_{F_2}[2], \\
x_{F_2}[2] &\leq \frac{y_{F_2}[2]}{1-\gamma_1^2}
\end{aligned}
$$

Non-Cocoercivity of F_{EG,γ_1}: Making Some Assumptions

To fulfill these constraints, we simply assumed that they hold as equalities and got:

$$x_{F_2}[2] = x_{F_1}[2] = \frac{y_{F_2}[2]}{1-\gamma_1^2}, \quad y_{F_1}[2] = \frac{y_{F_2}[2]}{1-\gamma_1},$$
$$y_{F_1}[1] = x_{F_2}[1] = y_{F_2}[1] = (1-\gamma_1)x_{F_1}[1].$$

Non-Cocoercivity of F_{EG,γ_1}: Making More Assumptions

- Using these dependencies in the remaining interpolation conditions, we derived

$$x_{F_1}[1] + \gamma_1 (x_{F_1}[1])^2 + \frac{\gamma_1 (y_{F_2}[2])^2}{(1-\gamma_1^2)^2} \leq 0.$$

- After that, we assumed that

$$y_{F_2}[2] = -x_{F_1}[1](1-\gamma_1^2).$$

- Together with previous inequality it gives

$$x_{F_1}[1] + 2\gamma_1 (x_{F_1}[1])^2 \leq 0.$$

- Next, we chose $x_{F_1}[1] = -1/2\gamma_1$ and put it in all previously derived dependencies.
- Finally, we generalized the example to the case of non-unit ℓ using "physical-dimension" arguments and got (21).

F_{EG,γ_1} is Not Cocoercive!

Putting all together we derive the following result:

Theorem 5

For all $\ell > 0$ and $\gamma_1 \in (0, 1/\ell]$ there exists ℓ-cocoercive operator F such that $F(x) = x_{F_1}, F(y) = y_{F_1}, F(x - \gamma_1 x_{F_1}) = x_{F_2}, F(y - \gamma_1 y_{F_1}) = y_{F_2}$ for $x, y, x_{F_1}, y_{F_1}, x_{F_2}, y_{F_2}$ defined in (21) and

$$\|x - \gamma_2 F(x - \gamma_1 F(x)) - y + \gamma_2 F(y - \gamma_1 F(y))\| > 1 = \|x - y\| \tag{22}$$

for all $\gamma_2 > 0$, i.e., $F_{\mathrm{EG},\gamma_1} = F(\mathrm{Id} - \gamma_1 F)$ is non-cocoercive.

EG and Cocoercivity: Preliminary Conclusions

- Now we know that one cannot apply analysis of GD to prove last-iterate $\mathcal{O}(1/K)$ convergence for EG
- We observed another significant difference between PP and EG: $F_{\mathrm{PP},\gamma}$ is cocoercive and $F_{\mathrm{EG},\gamma}$ is not
- But does it mean that it is impossible to prove $\mathcal{O}(1/K)$ convergence rate for EG in the considered setup (F is monotone and L-Lipschitz)? No, it does not!

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

We consider the problem

$$\begin{aligned} \max \quad & \|F(x^K)\|^2 \\ \text{s.t.} \quad & F \text{ is monotone and } L\text{-Lipschitz}, \; x^0 \in \mathbb{R}^d, \\ & \|x^0 - x^*\|^2 \leq 1, \\ & x^{k+1} = x^k - \gamma_2 F\left(x^k - \gamma_1 F(x^k)\right), \; k = 0, 1, \ldots, K-1 \end{aligned} \tag{23}$$

- Using similar steps as in the previous example, we construct a special SDP using the definitions of monotonicity (2) and (1) as interpolation conditions
- The resulting SDP gives just *an upper bound* for the value of (23) (see Proposition 3 from Ryu et al. [2020])
- Nevertheless, we solved the resulting SDP using PESTO [Taylor et al., 2017a] for $L = 1$, $\gamma_1 = \gamma_2 = 1/2L$, and various values of K.

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

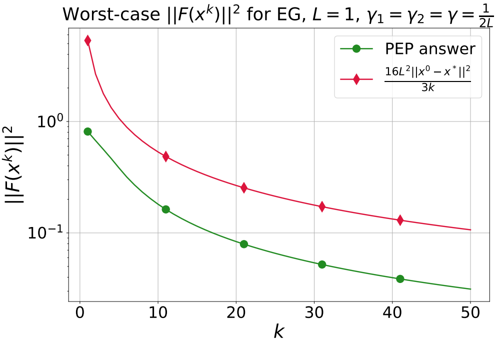


Figure: Comparison of the worst-case rate of EG obtained via solving PEP and the guessed upper-bound $^{16L^2\|x^0 - x^*\|^2}/_k$. Vertical axis is shown in logarithmic scale and after iteration $k = 20$ the curves are almost parallel, i.e., PEP answer and $^{16L^2\|x^0 - x^*\|^2}/_k$ differ almost by 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 (23) 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 $\mathcal{O}(1/K)$ Rate for EG

- However, guessing the dependencies is not always an easy task: the dependencies on the parameters of the problem like L, γ_1, γ_2 might be quite tricky
- To simplify the process of guessing the proof, we consider a simpler problem:

$$\begin{aligned} \Delta_{\mathrm{EG}}(L, \gamma_1, \gamma_2) = \quad \max \quad & \|F(x^1)\|^2 - \|F(x^0)\|^2 \\ \text{s.t.} \quad & F \text{ is monotone and } L\text{-Lipschitz}, \; x^0 \in \mathbb{R}^d, \\ & \|x^0 - x^*\|^2 \leq 1, \\ & x^1 = x^0 - \gamma_2 F\left(x^0 - \gamma_1 F(x^0)\right) \end{aligned} \tag{24}$$

with $\gamma_1 = \gamma_2 = \gamma$

Last-Iterate $\mathcal{O}(1/K)$ 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 γ
- 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 λ_2 and λ_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 $\mathcal{O}(1/K)$ 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 $K \geq 0$

$$\|F(x^K)\|^2 \leq \frac{\|x^0 - x^*\|^2}{\gamma^2(1 - L^2\gamma^2)(K+1)}, \tag{25}$$

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

$$\texttt{Gap}_F(x^K) = \max_{y \in \mathbb{R}^d : \|y - x^*\| \leq \|x^0 - x^*\|} \langle F(y), x^K - y \rangle \leq \frac{2\|x^0 - x^*\|^2}{\gamma\sqrt{1 - L^2\gamma^2}\sqrt{K+1}}. \tag{26}$$

On Our Failures Towards the Proof

The proof that I presented was not the first idea of what we tried to obtain. Here are some of the claims that we tried to prove first:

- We tried to show that $\|F_{\mathrm{EG},\gamma_1}(x^{k+1})\| \leq \|F_{\mathrm{EG},\gamma_1}(x^k)\|$ for a reasonable choice of γ_1 and γ_2
  - ✗ Perhaps, surprisingly, but this is not true even for L-cocoercive F: we observed this phenomenon via PEP
- We also tried to show that $\|F(x^{k+1})\| \leq \|F(x^k)\|$ when $\gamma_2 < \gamma_1$
  - ✗ Again, via PEP we oberved that this is not true even for L-cocoercive F

The usage of PEP helped us to save a lot of time from trying to prove the claims that do not hold in general!

Other Results in the Paper

- We obtain some “pessimistic” results on Optimistic Gradient method (OG):
  - ✗ for the two popular representations of OG (for the classical one and for the extrapolation from the past) we proved that corresponding operators are not even star-cocoercive
- For Hamiltonian Gradient method (HGM) [Balduzzi et al., 2018] we also
  - ✗ showed non-cocoercivity of corresponding operator when F is monotone and Lipschitz
  - ✓ derived best-iterate $\mathcal{O}\left(1/K\right)$ convergence rate in terms of the squared norm of the gradient of the Hamiltonian function $\mathcal{H}(x) = \frac{1}{2}\|F(x)\|^2$ when F and ∇F are Lipschitz-continuous but F is not necessary monotone

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 $\mathrm{C}, \mathrm{A}_1, \ldots, \mathrm{A}_m \in \mathbb{S}^n$, 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:

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

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

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

Details on SDP and Its Dual

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

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

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

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

which is equivalent to

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

Details on SDP and Its Dual

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

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

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