Online Learning in Games

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture by Thomas Pethick

Lecture 3: A practitioner's guide to monotone operators (Part II)

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

EE-735 (Spring 2024)

License Information for Online Learning in Games Slides

- This work is released under a Creative Commons License with the following terms:
- **Attribution**
  - The licensor permits others to copy, distribute, display, and perform the work. In return, licensees must give the original authors credit.
- **Non-Commercial**
  - The licensor permits others to copy, distribute, display, and perform the work. In return, licensees may not use the work for commercial purposes – unless they get the licensor's permission.
- **Share Alike**
  - The licensor permits others to distribute derivative works only under a license identical to the one that governs the licensor's work.
- Full Text of the License

Outline of this lecture

Overview of today

- Last week,
  - we have derived best iterate and average iterate results;
  - we have arrived at the extragradient (EG) type updates via our analysis.
- This week,
  - we will obtain $\|Fz^T\|^2 = \mathcal{O}(1/T)$ last iterate convergence rates for monotone and Lipschitz operators;
  - we will show that *linear* convergence is possible in particular cases.
- For the EG type methods, we will also answer the following questions:
  - How would we more intuitively motivate the scheme?
  - Can we avoid querying an extra gradient?

Summary of FBF: The descent inequality

- Let $H := \mathrm{id} - \gamma F$ and recall the update:

$$\begin{aligned} \bar{z}^t &= (\mathrm{id} + \gamma A)^{-1} H z^t, \\ z^{t+1} &= z^t - \alpha(H z^t - H\bar{z}^t). \end{aligned} \tag{FBF}$$

- The descent inequality is all you need!
- The three convergence results for FBF are a consequence of the following inequality:

Descent inequality of FBF

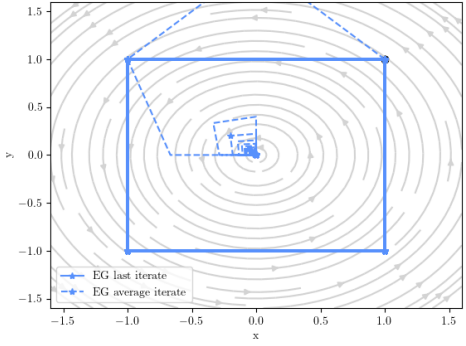
Let F be monotone and L-Lipschitz and A maximally monotone. Then the sequence generated by FBF satisfies

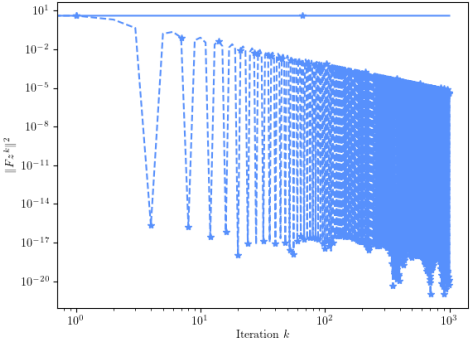
$$\|z^{t+1} - z^\star\|^2 \leq \|z^t - z^\star\|^2 \underbrace{-\alpha(1-\alpha)\|Hz^t - H\bar{z}^t\|^2}_{\text{best } \|F\bar{z}^t\|^2} \underbrace{-2\alpha\langle Hz^t - H\bar{z}^t, \bar{z}^t - z^\star\rangle}_{\text{gap for average: } \langle Fz^\star, \hat{z}^T - z^\star\rangle} \underbrace{-\alpha(1-\gamma^2 L^2)\|\bar{z}^t - z^t\|^2}_{\text{best } \|Fz^t\|^2}.$$

Exercise:
- Convince yourself that we can arrive at the above inequality.

Remarks:
- When we telescope, we were not using the remaining "good" terms so far.
- We will make use of these "good" terms today.

A (subtle) difference between last iterate and average iterate





Observations:

- If $\gamma = 1/L$ *exactly*, the last iterate will actually cycle!
  - Notice that we need $\gamma < 1/L$ strictly even for the best iterate convergence of $\|Fz^t\|$.
- The average iterate still converges (through the gap since $\gamma = 1/L$ is allowed).

Last iterate of GDA under cocoercivity

- Let us first recall GDA:

$$z^{t+1} = z^t - \gamma F z^t \tag{GDA}$$

- It suffices to show that $\|Fz^t\|^2$ is monotonically decreasing, i.e.,

$$\|Fz^{t+1}\|^2 \leq \|Fz^t\|^2. \tag{1}$$

- It turns out that this is directly implied by cocoercivity!

Theorem (Last iterate of (GDA))

Suppose $F : \mathbb{R}^d \to \mathbb{R}^d$ is β-cocoercive. Consider the sequence $(z^t)_{t\in\mathbb{N}}$ generated by (GDA) *with $\gamma < 2\beta$. Then, for all $z^\star \in \operatorname{zer} F$, the following holds*

$$\|Fz^T\|^2 \leq \frac{\|z^0 - z^\star\|^2}{\gamma(2\beta - \gamma)T}. \tag{2}$$

Last iterate of GDA under cocoercivity

Proof.

Cocoercivity applied at z^t and z^{t+1} yields the following

$$\begin{aligned}\beta\|Fz^{t+1} - Fz^t\|^2 &\leq \langle Fz^{t+1} - Fz^t, z^{t+1} - z^t\rangle \\ &= -\gamma\langle Fz^{t+1} - Fz^t, Fz^t\rangle \\ &= -\frac{\gamma}{2}\|Fz^t - Fz^{t+1} + Fz^t\|^2 - \frac{\gamma}{2}\|Fz^t\|^2 + \frac{\gamma}{2}\|Fz^{t+1} - Fz^t\|^2 \\ &= -\frac{\gamma}{2}\|Fz^{t+1}\|^2 + \frac{\gamma}{2}\|Fz^t\|^2 + \frac{\gamma}{2}\|Fz^{t+1} - Fz^t\|^2\end{aligned} \tag{3}$$

where the first equality follows from the update rule of (GDA). If we assume $\gamma \leq 2\beta$ then the above inequality reduces to

$$\|Fz^{t+1}\|^2 \leq \|Fz^t\|^2, \tag{4}$$

and we have shown monotonicity. We can use this to show

$$\|Fz^T\|^2 \leq \frac{1}{T}\sum_{t=1}^{T-1}\|Fz^t\|^2, \tag{5}$$

which we can subsequently upper bound using the argument for the best iterate. This completes the proof. □

Last iterate convergence rate for extragradient

- Recall the extragradient (EG) algorithm:

$$\begin{aligned} \bar{z}^t &= z^t - \gamma F z^t, \\ z^{t+1} &= z^t - \gamma F \bar{z}^t. \end{aligned} \tag{EG}$$

- Similarly to GDA, the EG iterates satisfy the following:

$$\|F z^{t+1}\|^2 \leq \|F z^t\|^2. \tag{6}$$

Theorem (Last z-iterate of (EG))

Suppose $F : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz and monotone. Consider the sequence $(z^t)_{t\in\mathbb{N}}$ generated by EG with $\gamma < 1/L$. Then, for all $z^\star \in \operatorname{zer} F$, the following holds

$$\|F z^T\|^2 \leq \frac{\|z^0 - z^\star\|^2}{\gamma^2(1-\gamma^2 L^2)(T-1)}. \tag{7}$$

Remarks:

- In contrast, the gap of the last iterate has a $\mathcal{O}(1/\sqrt{T})$ rate with a matching lower bound [7].
- This is easily attained by converting (7) using Cauchy-Schwarz inequality [2, Lm. 2].

Last iterate convergence rate for extragradient

Proof.

Monotonicity, Lipschitz and the update rule give us,

$$0 \leq \langle Fz^t - Fz^{t+1}, z^t - z^{t+1} \rangle = \langle Fz^t - Fz^{t+1}, \gamma F\bar{z}^t \rangle \tag{8}$$

$$\|Fz^{t+1} - F\bar{z}^t\|^2 \leq L^2 \|z^{t+1} - \bar{z}^t\|^2 = L^2\gamma^2 \|Fz^t - F\bar{z}^t\|^2 \tag{9}$$

Adding (8) and (9) with $\frac{2}{\gamma}$ and 1 respectively and rewriting the inner product,

$$\begin{aligned} \|F\bar{z}^t - Fz^{t+1}\|^2 \leq{}& \|Fz^t\|^2 + \|F\bar{z}^t\|^2 - \|F\bar{z}^t - Fz^t\|^2 \\ & - \|Fz^{t+1}\|^2 - \|F\bar{z}^t\|^2 + \|F\bar{z}^t - Fz^{t+1}\|^2 \\ & + L^2\gamma^2 \|F\bar{z}^t - Fz^t\|^2 \end{aligned} \tag{10}$$

Observing that terms cancels,

$$0 \leq \|Fz^t\|^2 - \|Fz^{t+1}\|^2 + (L^2\gamma^2 - 1)\|F\bar{z}^t - Fz^t\|^2 \tag{11}$$

We conclude that $\|Fz^t\|^2$ is monotonically decreasing for (EG) as well, which completes the proof. □

Contraction: fixed point iterations

- So far, the fastest rate we have seen is the $\mathcal{O}(1/T)$-rate. When can we improve significantly on this?
- Let us redefine (GDA) through an operator $S : \mathbb{R}^d \to \mathbb{R}^d$ as follows

$$z^{t+1} = Sz^t \quad \text{where} \quad S := \mathrm{id} - \gamma F.$$

- We can ask when the algorithm no longer "moves" the iterates:

Fixed point

Find $z \in \operatorname{fix} S$ where

$$\operatorname{fix} S := \{z \mid z = Sz\}.$$

Remark:

- By construction it is equivalent to finding a zero of F, i.e., $\operatorname{zer} F = \operatorname{fix} S$.
- We will instead ask what we need of S to converge geometrically.

Contraction: getting linear convergence

Definition (Contraction)

The operator $S : \mathbb{R}^d \to \mathbb{R}^d$ is an contraction if S is ℓ-Lipschitz with $\ell < 1$.

Remark: ◦ When γF is strongly monotone and 1-Lipschitz then $S := \mathrm{id} - \gamma F$ is a contraction.

Theorem

Suppose $S : \mathbb{R}^d \to \mathbb{R}^d$ is an contraction and $z^\star \in \operatorname{fix} S$. Then, the iterates generated by $z^{t+1} = S(z^t)$ satisfies

$$\|z^T - z^\star\|^2 \leq \ell^{2T} \|z^0 - z^\star\|^2. \tag{12}$$

Proof.

The rate is an immediate consequence of the definition:

$$\|z^T - z^\star\|^2 = \|Sz^{T-1} - Sz^\star\|^2 \leq \ell^2 \|z^{T-1} - z^\star\|^2 = \ell^4 \|Sz^{T-1} - Sz^\star\|^2 \leq \ldots \leq \ell^{2T} \|z^0 - z^\star\|^2. \tag{13}$$

□

Error bound condition: generalizing the linear convergence result

- Using EG we can expand the class for which we have this geometric convergence result.
- Let us return to the descent inequality in the proof of EG:

$$\begin{aligned} \|z^{t+1} - z^\star\|^2 &\leq \|z^t - z^\star\|^2 - \alpha(1-\gamma^2 L^2)\|z^t - \bar{z}^t\|^2 \\ &= \|z^t - z^\star\|^2 - \alpha\gamma^2(1-\gamma^2 L^2)\|Fz^t\|^2 \end{aligned} \tag{14}$$

- To apply a similar recursive argument as for GDA under contraction:
  - We only need to convert $\|Fz^t\|^2$ to $\|z^t - z^\star\|^2$.

Definition (Error bound)

The operator $F : \mathbb{R}^d \to \mathbb{R}^d$ satisfies, for all $z \in \mathbb{R}^d$ and some $z^\star \in \operatorname{zer} F$, with $\tau > 0$

$$\|Fz\| \geq \tau \|z - z^\star\|. \tag{15}$$

Remarks:

- This class includes both strongly monotone *and* affine operators (i.e. the bilinear game).
- (local) error bound, Polyak-Lojasiewicz, and growth conditions can be all equivalent [17].

Error bound condition: convergence of EG

Theorem

Suppose $F : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz and monotone and satisfies the error bound condition. Consider the sequence $(z^t)_{t\in\mathbb{N}}$ generated by EG with $\gamma < 1/L$. Then, for all $z^\star \in \operatorname{zer} F$, it holds that

$$\|z^T - z^\star\|^2 \leq (1 - \tau^2\gamma^2(1 - \gamma^2 L^2))^{2T}\|z^0 - z^\star\|^2. \tag{16}$$

Remarks:

- See Tseng [20] for a generalization to constraint settings when the set is a polyhedral.
- The error bound condition was first proposed in Luo et al. [11].
- For other sufficient conditions for linear convergence (in minimization) see Karimi et al. [8].

Error bound condition: proof

Proof.

We continue from the descent inequality of EG for a monotone and Lipschitz operator F as follows:

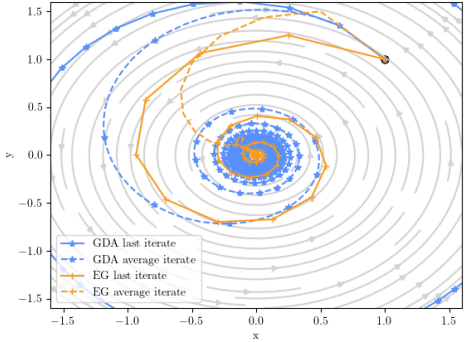
$$\begin{aligned} \|z^{t+1} - z^\star\|^2 &\leq \|z^t - z^\star\|^2 - (1 - \gamma^2 L^2)\|z^t - \bar{z}^t\|^2 \\ &= \|z^t - z^\star\|^2 - \gamma^2(1 - \gamma^2 L^2)\|Fz^t\|^2 \\ &\leq (1 - \tau^2\gamma^2(1 - \gamma^2 L^2))\|z^t - z^\star\|^2, \end{aligned} \tag{17}$$

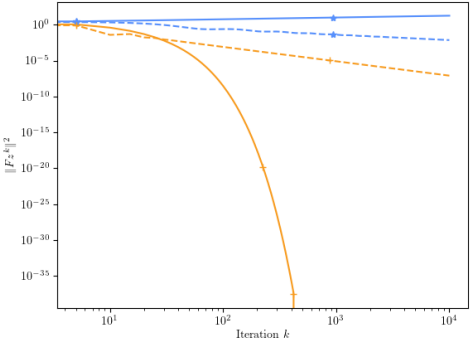
given that $\gamma^2(1 - \gamma^2 L^2) \geq 0$ and where the error bound condition is used in the last inequality. □

Error bound condition: seeing it in action

- Consider the bilinear game:

$$\min_{x \in \mathbb{R}} \max_{y \in \mathbb{R}} xy$$





Observation:

- We have linear convergence for EG!

Different perspectives on extragradient

- We arrived at EG through the analysis last week.
- Is there a more intuitive motivation?
- Two powerful ideas:
  - Each step of EG is a projection onto a particular hyperplane.
  - EG can be seen as approximating to the proximal point method.

GDA as a fixed point iteration

- The update of FBF can be cast as an iterative projection onto a set containing the solution set.
- We will now build up to this view.

Krasnosel'skii-Mann (KM) iteration

Let $S : \mathbb{R}^d \to \mathbb{R}^d$ be an operator and $\lambda > 0$. The KM iteration is given by

$$z^{t+1} = (1 - \lambda)z^t + \lambda S z^t \tag{KM}$$

Remarks:

- GDA can be seen as an instance of Krasnosel'skii-Mann (KM) iteration with $S = \mathrm{id} - \gamma F$.
- We were able to show convergence when γF was $\frac{1}{2}$-cocoercive.
- What is the equivalent condition expressed in terms of S instead?

GDA convergence for cocoercive F through KM

Definition (Nonexpansiveness)

An operator $S : \mathbb{R}^n \to \mathbb{R}^d$ is said to be nonexpansive if the following holds

$$\|Sz - Sz'\| \leq \|z - z'\| \quad \forall z, z' \in \mathbb{R}^d. \tag{18}$$

Remark:

- The operator $S = \mathrm{id} - \gamma F$ is nonexpansive iff γF is $\frac{1}{2}$-cocoercive.
- With this definition we can reprove the convergence of GDA.

Theorem (Best iterate of KM)

Assume $S : \mathbb{R}^d \to \mathbb{R}^d$ is nonexpansive. Consider the sequence $(z^t)_{t\in\mathbb{N}}$ generated by (KM) *with $\lambda \in (0, 1)$. Then, for all $z^\star \in \operatorname{fix} S$, the following holds*

$$\min_{t\in\{0,\dots,T-1\}} \|Sz^t - z^t\|^2 \leq \frac{\|z^0 - z^\star\|^2}{\lambda(1-\lambda)T}. \tag{19}$$

GDA convergence for cocoercive F through KM

Proof.

We proceed as usual with a one-step analysis:

$$\begin{aligned}\|z^{t+1}-z^\star\|^2 &= (1-\lambda)\|z^t-z^\star\|^2+\lambda\|Sz^t-z^\star\|^2-\lambda(1-\lambda)\|Sz^t-z^t\|^2\\ &= (1-\lambda)\|z^t-z^\star\|^2+\lambda\|Sz^t-Sz^\star\|^2-\lambda(1-\lambda)\|Sz^t-z^t\|^2\\ &\le (1-\lambda)\|z^t-z^\star\|^2+\lambda\|z^t-z^\star\|^2-\lambda(1-\lambda)\|Sz^t-z^t\|^2\\ &= \|z^t-z^\star\|^2-\lambda(1-\lambda)\|Sz^t-z^t\|^2,\end{aligned} \tag{20}$$

where we have used nonexpansiveness of S and that $Sz^\star = z^\star$. Telescoping completes the proof. □

KM of firmly nonexpansive operator

- We can improve when S is *firmly* nonexpansive.

Definition (Firmly nonexpansive)

An operator $S : \mathbb{R}^d \to \mathbb{R}^d$ is said to be firmly nonexpansive if

$$\|Sz - Sz'\|^2 + \|(\mathrm{id} - S)z - (\mathrm{id} - S)z'\|^2 \leq \|z - z'\|^2 \quad \forall z, z' \in \mathbb{R}^d. \tag{21}$$

Remark:

- If S is firmly nonexpansive so is $\mathrm{id} - S$ and both are 1-cocoercive.
- A projection is firmly nonexpansive when the set is convex **(This will be crucial!)**.

Theorem (Best iterate of KM)

Suppose $S : \mathbb{R}^d \to \mathbb{R}^d$ is firmly nonexpansive. Consider the sequence $(z^t)_{t \in \mathbb{N}}$ generated by KM with $\lambda \in (0, 2)$. Then, for all $z^\star \in \operatorname{fix} S$, it holds that

$$\min_{t \in \{0, \ldots, T-1\}} \|Sz^t - z^t\|^2 \leq \frac{\|z^0 - z^\star\|^2}{\lambda(2 - \lambda)T}. \tag{22}$$

Remarks:

- This implies convergence of GDA by taking $S = \mathrm{id} - \eta F$ when ηF is 1-cocoercive.
- We can now crucially take $\lambda \in (0, 2)$ instead of only $\lambda \in (0, 1)$.

KM of firmly nonexpansive operator

Proof.

The steps are the same as for the nonexpansive case, except we have an additional (good) term,

$$\begin{aligned}\|z^{t+1}-z^\star\|^2 &= (1-\lambda)\|z^t-z^\star\|^2+\lambda\|Sz^t-z^\star\|^2-\lambda(1-\lambda)\|Sz^t-z^t\|^2\\ &= (1-\lambda)\|z^t-z^\star\|^2+\lambda\|Sz^t-Sz^\star\|^2-\lambda(1-\lambda)\|Sz^t-z^t\|^2\\ &\leq (1-\lambda)\|z^t-z^\star\|^2+\lambda\|z^t-z^\star\|^2-\lambda\|Sz^t-z^t\|^2-\lambda(1-\lambda)\|Sz^t-z^t\|^2\\ &= \|z^t-z^\star\|^2-\lambda(2-\lambda)\|Sz^t-z^t\|^2.\end{aligned} \tag{23}$$

We have used firmly nonexpansiveness of S and that $Sz^\star = z^\star$. Notice that we can now crucially take $\lambda \in (0,2)$ instead of only $\lambda \in (0,1)$. Telescoping completes the proof. □

Hyperplane projection: capturing extragradient and forward-backward-forward (FBF)

- Recall that we wish to find $z^\star \in \mathbb{R}^d$ such that

$$z^\star \in \operatorname{zer}(F + A). \tag{24}$$

- We will now use convergence of KM to reprove convergence of EG/FBF.
- It turns out that FBF can be seen as running KM on an iterative projection:

Projected interpretation of EG/FBF

Let $H := \mathrm{id} - \gamma F$. Consider the sequence generated by

$$z^{t+1} = (1-\lambda)\, z^t + \lambda \mathbf{\Pi}_{\mathcal{D}(z^t)}(z^t) \tag{25}$$

which projects onto the half-space $\mathcal{D}(z) := \{w \mid \langle Hz - H\bar{z}, \bar{z} - w\rangle \geq 0\}$ with $\bar{z} := (\mathrm{id} + \gamma A)^{-1} Hz$.

Remark:

- The proof (and construction) has two key components:
  - first show that when a solution is found we will stay at the solution
  - and otherwise we will make progress towards the solution set.
- Then we only need to show equivalence with EG/FBF.

Hyperplane projection: proof step I

Proof.

Step 1. The set $\mathcal{D}(z)$ is constructed to contain the solution set defined as

$$\mathcal{S}^\star = \{z^\star \mid \langle Fz + Az, z - z^\star \rangle \geq 0 \ \forall z, z^\star\}.$$

Let us verify this claim. From the definition of $\bar{z}$ we have

$$\tfrac{1}{\gamma}\left(Hz - H\bar{z}\right) \in A\bar{z} + F\bar{z}. \tag{26}$$

So by monotonicity of $A + F$,

$$\tfrac{1}{\gamma}\langle Hz - H\bar{z}, \bar{z} - z^\star \rangle \geq 0. \tag{27}$$

This confirms that $\mathcal{S}^\star \subseteq \mathcal{D}\left(z\right)$. Thus, any solution $z^\star \in \mathcal{S}^\star$ is a fixed point of the projection $\mathbf{\Pi}_{\mathcal{D}(z)}$, i.e. $z^\star \in \operatorname{fix} \mathbf{\Pi}_{\mathcal{D}(z)}$ for any $z \in \mathbb{R}^n$.

Hyperplane projection: proof step II

Proof (Cont.)

Step 2. To find the closed form solution for the projection we invoke the lemma below, concerning general hyperplane projections, with $a = H\bar{z} - Hz$ and $b = -\langle Hz - H\bar{z}, \bar{z}\rangle$. By simple substitution, we have

$$\mathbf{\Pi}_{\mathcal{D}(z^t)}(z^t) = z^t + \alpha_t(H\bar{z}^t - Hz^t), \quad \text{with} \quad \alpha_t = \frac{\langle H\bar{z}^t - Hz^t, \bar{z}^t - z^t\rangle}{\|H\bar{z}^t - Hz^t\|^2}. \tag{28}$$

This recovers FBF (modulo the adaptive parameter choice).

○ We have used the following lemma.

Lemma

The projection $\mathbf{\Pi}_{\mathcal{D}}(x) := \arg\min_{z\in\mathcal{D}} \frac{1}{2}\|z - x\|^2$ *onto the set* $\mathcal{D} = \{z \mid \langle a, z\rangle \geq b\}$ *is given for* $x \notin \mathcal{D}$ *as,*

$$\mathbf{\Pi}_{\mathcal{D}}(x) = x - \frac{\langle a, x\rangle - b}{\|a\|^2}a. \tag{29}$$

Hyperplane projection: proof step III

Proof (Cont.)

Step 3. We finally need to argue that we always improve (if we are not at a solution).
Notice that (25) is an instance of KM with $S = \mathbf{\Pi}_{\mathcal{D}(z^t)}$. By using that S is firmly nonexpansive we get

$$\min_{t\in\{0,\dots,T-1\}} \|Sz^t - z^t\|^2 \leq \frac{\|z^0 - z^\star\|^2}{\lambda(2-\lambda)T}, \tag{30}$$

where $z^\star \in \mathcal{S}^\star$ due to step 1. Convergence follows by noting that

$$\|Sz^t - z^t\|^2 = \alpha_t^2 \|H\bar{z}^t - Hz^t\|^2 \tag{31}$$

and that α_t is bounded away from zero by $\alpha_t \geq \frac{1}{2}$ due to cocoercivity of H.

Hyperplane projection

- The scheme:

$$\begin{aligned} \bar{z}^t &= (\mathrm{id} + \gamma A)^{-1} H z^t \\ z^{t+1} &= z^t + \lambda \alpha_t (H\bar{z}^t - Hz^t), \quad \text{with} \quad \alpha_t = \frac{\langle H\bar{z}^t - Hz^t, \bar{z}^t - z^t \rangle}{\|H\bar{z}^t - Hz^t\|^2}. \end{aligned} \tag{32}$$

with $\lambda \in (0, 2)$.

Theorem (Best $\bar{z}$-iterate of (25))

Assume $F : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz and monotone and $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is maximally monotone. Consider the sequence $(z^t)_{t\in\mathbb{N}}$ generated by (32) with $\gamma \leq 1/L$ and $\lambda \in (0, 2)$. Then for $z^\star \in \mathrm{zer}(F + A)$,

$$\min_{t\in\{0,\dots,T-1\}} \|Hz^t - H\bar{z}^t\|^2 \leq \frac{4\|z^0 - z^\star\|^2}{\lambda(2-\lambda)T}. \tag{33}$$

Remark:

- The stepsize is adaptive (but we can infer the constant stepsize result)
- Hyperplane projections are powerful due to their generality (see [10, 6, 15] for extension).
- The idea dates back to at least Solodov & Tseng [18] and Solodov & Svaiter [19].

Another perspective: arriving at EG through proximal point

- For simplicity we will consider the unconstrained case where $F : \mathbb{R}^d \to \mathbb{R}^d$ (single-valued) and $A \equiv 0$.
- Consider the (implicit) proximal point method (PP)

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

- We immediately have descent

$$\begin{aligned} \|z^{t+1} - z^\star\|^2 &= \|z^t - z^\star\|^2 - \|z^{t+1} - z^t\|^2 - 2\gamma\langle F(z^{t+1}), z^{t+1} - z^\star\rangle \\ &= \|z^t - z^\star\|^2 - \gamma^2\|Fz^{t+1}\|^2 - 2\gamma\langle F(z^{t+1}), z^{t+1} - z^\star\rangle \end{aligned} \tag{34}$$

- However, PP is an implicit scheme. Can we approximate it?

The proximal point method as a resolvent

- The proximal point update is an instance of the resolvent.

Definition (The resolvent)

Given an operator $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$, we define the associated resolvent:

$$J_A = (\mathrm{id} + A)^{-1}. \tag{35}$$

Remarks:

- If A is maximally monotone then J_A is firmly nonexpansive.
- By taking $S = J_A$ in KM:
  - PP convergence through the KM theorem when A is maximally monotone ($\lambda = 1$ is allowed).
- The resolvent might not be known in closed form: Let us approximate it!

Extragradient as approximating the resolvent

- One step of the resolvent requires us to find

$$z' = (\mathrm{id} + \gamma F)^{-1} z \quad \Leftrightarrow \quad z' = z - \gamma F z'.$$

- In other word, we seek to find

$$w^\star \in \operatorname{fix} C_z \quad \text{where} \quad C_z : w \mapsto z - \gamma F w. \tag{36}$$

- Apply C_z repeatedly: A fast geometric rate is immediate by establishing that C_z is a contraction.

Lemma

If γF is a contraction then the sequence $(w^k)_{k\in\mathbb{N}}$ generated by repeatedly applying C_z converges geometrically.

Proof.

The operator $C_z : \mathbb{R}^d \to \mathbb{R}^d$ as defined in (36) is contractive when γF is a contractive. That is,

$$\|C_z(x) - C_z(y)\| = \|\gamma F x - \gamma F y\| < \|x - y\|. \tag{37}$$

Thus, the contraction argument applies, which completes the proof. □

Extragradient as approximating the resolvent

- Let $C_z(w) := z - \gamma F w$ and define k inner iterations as:

$$C^k(z) := \underbrace{C_z \circ \cdots \circ C_z}_{k \text{ times}}(z).$$

- The outer iterations are then given as:

$$z^{t+1} = (1-\lambda)z^t + \lambda C^k(z^t).$$

Example

Consider the resolvent $J_{\gamma F} := (\mathrm{id} + \gamma F)^{-1}$ where F is monotone and L-Lipschitz F:

- With stepsize $\gamma < 1/L$ the operator γF is a contraction.
- The inner loop (approximate $J_{\gamma F}$) has linear rate, so we only need $\log T$ number of inner steps.

Remarks:

- To shave off the logarithmic factor in the complexity we want constant inner iterations.
- This motivates EG, since $(S \circ S)(z^t)$ and $\lambda = 1$ exactly corresponds to one step of (EG).
- Original motivation behind MirrorProx Nemirovski [14] (a generalization of extragradient).
- The above argument is made precise in Cevher et al. [3].

How to use the left-over "good" term: Single-call variant

- The method FBF we have derived so far requires two operator evaluations of F.
- Can we construct a method that only uses a *single call* per iteration?
- There is hope since there is an unused "good" term ($\|z^t - \bar{z}^t\|^2$) as long as the stepsize $\gamma < 1/L$.
- Let us alter the forward operator:

$$H_z(\bar{z}) = z - \gamma F\bar{z}.$$

- We can now modify FBF to reuse the *past* operator evaluation $F\bar{z}^{t-1}$:

$$\begin{aligned} \bar{z}^t &= (\mathrm{id} + \gamma A)^{-1} H_{z^t}(\bar{z}^{t-1}) \\ z^{t+1} &= z^t - \alpha(H_{z^t}(\bar{z}^{t-1}) - H\bar{z}^t). \end{aligned} \tag{PFBF}$$

Single-call variant

- The proof for the single-call variant only requires one additional triangle inequality and Lipschitzness.

Theorem (Gap for average iterate of PFBF)

Suppose $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is maximally monotone and $F : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz and monotone. Consider the sequence $(z^t)_{t\in\mathbb{N}}$ generated by PFBF with $\gamma \leq \frac{1}{2L}$ and $\alpha = 1$. Then, for all $z^\star \in \operatorname{zer}(A+F)$ and any compact neighborhood $\mathcal{C} \subseteq \mathbb{R}^d$ of $z^\star$, it holds that

$$\operatorname{Gap}_{\mathcal{C}}(\hat{z}^T) \leq \frac{\|z^0 - z^\star\|^2}{2\gamma T}.$$

where $\hat{z}^T = \frac{1}{T}\sum_{t=0}^{T-1} \bar{z}^t$.

Remarks:

- The maximal stepsize is half of FBF (a worse rate by a constant factor 2)
- However, PFBF only uses half the operator evaluations per iteration

Single-call variant

Proof.

We can arrive at almost the same descent inequality as for FBF:

$$
\begin{aligned}
&\|z^{t+1} - z^\star\|^2 \\
&= \|z^t - z^\star\|^2 + \|H_{z^t}(\bar{z}^{t-1}) - H\bar{z}^t\|^2 - 2\langle H_{z^t}(\bar{z}^{t-1}) - H\bar{z}^t, z^t - z^\star\rangle \\
&= \|z^t - z^\star\|^2 + \|H_{z^t}(\bar{z}^{t-1}) - H\bar{z}^t\|^2 - 2\langle H_{z^t}(\bar{z}^{t-1}) - H\bar{z}^t, z^t - \bar{z}^t\rangle - 2\langle H_{z^t}(\bar{z}^{t-1}) - H\bar{z}^t, \bar{z}^t - z^\star\rangle \\
&= \|z^t - z^\star\|^2 - \|z^t - \bar{z}^t\|^2 + \|H_{z^t}(\bar{z}^{t-1}) - H\bar{z}^t - z^t + \bar{z}^t\|^2 - 2\langle H_{z^t}(\bar{z}^{t-1}) - H\bar{z}^t, \bar{z}^t - z^\star\rangle \\
&= \|z^t - z^\star\|^2 - \|z^t - \bar{z}^t\|^2 + \gamma^2\|F\bar{z}^{t-1} - F\bar{z}^t\|^2 - 2\langle H_{z^t}(\bar{z}^{t-1}) - H\bar{z}^t, \bar{z}^t - z^\star\rangle \\
&\le \|z^t - z^\star\|^2 - \|z^t - \bar{z}^t\|^2 + \gamma^2 L^2\|\bar{z}^{t-1} - \bar{z}^t\|^2 - 2\langle H_{z^t}(\bar{z}^{t-1}) - H\bar{z}^t, \bar{z}^t - z^\star\rangle
\end{aligned}
\tag{38}
$$

However, the third ("bad") term does not match the second ("good") term. Using triangle inequality,

$$
\begin{aligned}
\tfrac{1}{2}\|\bar{z}^{t-1} - \bar{z}^t\|^2 \le \|\bar{z}^{t-1} - z^t\|^2 + \|\bar{z}^t - z^t\|^2 &= \gamma^2\|F\bar{z}^{t-2} - F\bar{z}^{t-1}\|^2 + \|\bar{z}^t - z^t\|^2 \\
&\le \gamma^2 L^2\|\bar{z}^{t-2} - \bar{z}^{t-1}\|^2 + \|\bar{z}^t - z^t\|^2
\end{aligned}
\tag{39}
$$

where the equality follows from the update rule and the last inequality follows from Lipschitz.

Single-call variant

Proof (Cont.)

Adding (39) to (38) and rearranging,

$$\|z^{t+1} - z^\star\|^2 + (\tfrac{1}{2} - \gamma^2 L^2)\|\bar{z}^{t-1} - \bar{z}^t\|^2 \leq \|z^t - z^\star\|^2 + \gamma^2 L^2 \|\bar{z}^{t-2} - \bar{z}^{t-1}\|^2 - 2\langle H_{z^t}(\bar{z}^{t-1}) - H\bar{z}^t, \bar{z}^t - z^\star\rangle.$$

It will clearly no longer suffice to simply telescope $\|z^{t+1} - z^\star\|^2$.

Instead we now rely on the following potential function

$$\mathcal{U}_{t+1} = \|z^{t+1} - z^\star\|^2 + C_{t+1}\|\bar{z}^{t-1} - \bar{z}^t\|^2.$$

For $\mathcal{U}_{t+1}$ to telescope we need $(\frac{1}{2} - \gamma^2 L^2) \geq \gamma^2 L^2$, which is identical to requiring $\gamma \leq \frac{1}{2L}$. Telescoping, picking $\bar{z}^{-1} = \bar{z}^0$, applying monotonicity of $S = F + A$,

$$\langle v, \hat{z}^T - z^\star\rangle \leq \frac{\|z^0 - z^\star\|^2}{2\gamma T}. \quad \forall v \in S\hat{z}^T.$$

Converting into the restricted gap function through the gap lemma finishes the proof.

Overview of methods

Table: Overview of splitting methods for the maximally monotone inclusion $0 \in Az + Fz$ where only F is Lipschitz.

	1 forward call	2 forward calls
1 backward call	PFBF [5, 13, 1], Reflected-forward-backward (RFB) [12, 4]	FBF [21]
2 backward calls	Popov's method [16]	EG [9]

Historical notes: PFBF was studied under many names:

- *Optimistic gradient descent ascent* [5]
- *Forward-reflected-backward* [13]
- *Forward-backward-forward-past* [1]

Connection to optimistic methods in online learning

- Simplifying the update:

$$\begin{aligned} \bar{z}^t &= (\mathrm{id} + \gamma A)^{-1}(z^t - \gamma F \bar{z}^{t-1}) \\ z^{t+1} &= \bar{z}^t - \gamma(F\bar{z}^t - F\bar{z}^{t-1}) \end{aligned} \tag{PFBF}$$

- Condensing into one update:

$$\bar{z}^t = (\mathrm{id} + \gamma A)^{-1}(\bar{z}^t - 2\gamma F \bar{z}^{t-1} + \gamma F \bar{z}^{t-2}) \tag{PFBF}$$

- We have derived what is essentially known as *optimism* in online learning (albeit with fixed stepsize)
- By combining with a particular adaptive stepsize we can get:
  - (i) the optimal $\mathcal{O}(1/\sqrt{T})$ in the online setting
  - (ii) while maintaining the optimal $\mathcal{O}(1/T)$ in the deterministic offline monotone setting.

References I

[1] Axel Böhm, Michael Sedlmayer, Ernö Robert Csetnek, and Radu Ioan Bot.
Two steps at a time—taking GAN training in stride with Tseng's method.
SIAM Journal on Mathematics of Data Science, 4(2):750–771, 2022.
(Cited on page 36.)

[2] Yang Cai, Argyris Oikonomou, and Weiqiang Zheng.
Tight last-iterate convergence of the extragradient method for constrained monotone variational inequalities.

arXiv preprint arXiv:2204.09228, 2022.
(Cited on page 9.)

[3] Volkan Cevher, Georgios Piliouras, Ryann Sim, and Stratis Skoulakis.
Min-max optimization made simple: Approximating the proximal point method via contraction maps.
In *Symposium on Simplicity in Algorithms (SOSA)*, pages 192–206. SIAM, 2023.
(Cited on page 31.)

[4] Volkan Cevher and Bang Cong Vu.
A reflected forward-backward splitting method for monotone inclusions involving lipschitzian operators.
Set-Valued and Variational Analysis, pages 1–12, 2020.
(Cited on page 36.)

References II

[5] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng.
Training GANs with optimism.
arXiv preprint arXiv:1711.00141, 2017.
(Cited on page 36.)

[6] Pontus Giselsson.
Nonlinear forward-backward splitting with projection correction.
SIAM Journal on Optimization, 31(3):2199–2226, 2021.
(Cited on page 27.)

[7] Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis.
Tight last-iterate convergence rates for no-regret learning in multi-player games.
Advances in neural information processing systems, 33:20766–20778, 2020.
(Cited on page 9.)

[8] Hamed Karimi, Julie Nutini, and Mark Schmidt.
Linear convergence of gradient and proximal-gradient methods under the polyak-undefinedojasiewicz condition.
In *European Conference on Machine Learning and Knowledge Discovery in Databases - Volume 9851*, ECML PKDD 2016, pages 795–811, Berlin, Heidelberg, 2016. Springer-Verlag.
(Cited on page 14.)

References III

[9] Galina M Korpelevich.
The extragradient method for finding saddle points and other problems.
Matecon, 12:747–756, 1976.
(Cited on page 36.)

[10] Puya Latafat and Panagiotis Patrinos.
Asymmetric forward–backward–adjoint splitting for solving monotone inclusions involving three operators.
Computational Optimization and Applications, 68(1):57–93, Sep 2017.
(Cited on page 27.)

[11] Zhi-Quan Luo and Paul Tseng.
Error bounds and convergence analysis of feasible descent methods: a general approach.
Annals of Operations Research, 1993.
(Cited on page 14.)

[12] Yu Malitsky.
Projected reflected gradient methods for monotone variational inequalities.
SIAM Journal on Optimization, 25(1):502–520, 2015.
(Cited on page 36.)

References IV

[13] Yura Malitsky and Matthew K Tam.
A forward-backward splitting method for monotone inclusions without cocoercivity.
SIAM Journal on Optimization, 30(2):1451–1472, 2020.
(Cited on page 36.)

[14] Arkadi 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.
(Cited on page 31.)

[15] Thomas Pethick, Puya Latafat, Panagiotis Patrinos, Olivier Fercoq, and Volkan Cevher.
Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems.
arXiv preprint arXiv:2302.09831, 2023.
(Cited on page 27.)

[16] Leonid Denisovich 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.
(Cited on page 36.)

References V

[17] Quentin Rebjock and Nicolas Boumal.
Fast convergence to non-isolated minima: four equivalent conditions for C^2 functions.
arXiv preprint arXiv:2303.00096, 2023.
(Cited on page 13.)

[18] M. V. Solodov and P. Tseng.
Modified projection-type methods for monotone variational inequalities.
SIAM Journal on Control and Optimization, 34(5):1814–1830, 1996.
(Cited on page 27.)

[19] Mikhail V Solodov and Benar F Svaiter.
A hybrid projection-proximal point algorithm.
Journal of convex analysis, 6(1):59–70, 1999.
(Cited on page 27.)

[20] Paul Tseng.
On linear convergence of iterative methods for the variational inequality problem.
Journal of Computational and Applied Mathematics, 60(1-2):237–252, 1995.
(Cited on page 14.)

References VI

[21] Paul Tseng.
A modified forward-backward splitting method for maximal monotone mappings.
SIAM Journal on Control and Optimization, 38(2):431–446, 2000.
(Cited on page 36.)