# Online Learning in Games

Prof. Volkan Cevher
*volkan.cevher@epfl.ch*

*Lecture by Thomas Pethick*

*Lecture 5: A practitioner's guide to monotone operators (Part I)*

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

# Logistics

| | |
|---|---|
| Credits | 4 |
| Lectures | Monday 9:15-11:00 (ELG120) |
| Practical hours | Monday 9:15-12:00 and 14:15-17:00 starting 3rd of April (ELG116) |
| Prerequisites | Previous coursework in calculus, linear algebra, and probability is required. Familiarity with optimization is useful. |
| Grading | **Preparation & presentation of a lecture given in week 14 (cf., coursebook).** |
| Moodle | `https://moodle.epfl.ch/course/view.php?id=17204`. |
| Course book | `https://edu.epfl.ch/coursebook/en/online-learning-in-games-EE-735` |
| LIONS | Stratis Skoulakis, Kimon Antonakopoulos, Thomas Pethick, Igor Krawczuk |

## Introduction

- Offline minimax problems: Last week we showed $\mathcal{O}(1/\sqrt{T})$ rate using no-regret algorithms (FTRL/OGD).

### Goals of today

1. Show when we can obtain a $\mathcal{O}(1/T)$ rate with the gradient descent ascent method (GDA)
2. Extend the class for which we have a $\mathcal{O}(1/T)$ rate by using extragradient-like schemes

**Remarks:**

- For a basic exposure to extragradient-like schemes, see Math of Data.
- This material introduces monotone operators (the "right" abstraction).
- We will rediscover sufficient structures for $\mathcal{O}(1/T)$ rate through the convergence analysis.

# A motivating example

**Example (Unconstrained convex-concave minimax)**

Consider the following (unconstrained) minimax problem

$$\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^m} f(x, y), \tag{1}$$

where $f$ is differentiable, $f(\cdot, y)$ is convex $\forall y \in \mathbb{R}^m$ and $f(x, \cdot)$ is concave $\forall x \in \mathbb{R}^n$.

**Remarks:**

- There are many solution concepts for optimization problems. Here are two relevant ones:
  - first-order stationarity, i.e., for unconstrained a point $(x^\star, y^\star)$ such that

    $$\nabla_x f(x^\star, y^\star) = 0 \text{ and } \nabla_y f(x^\star, y^\star) = 0$$

  - saddle point or more generally the Nash equilibrium, i.e., a point $(x^\star, y^\star)$ such that

    $$f(x^\star, y) \leq f(x^\star, y^\star) \leq f(x, y^\star) \quad \forall x \in \mathbb{R}^n, y \in \mathbb{R}^m.$$

- For convex-concave problems, they coincide
- For this reason, we will start with the first-order stationarity and describe more later

## A motivating example

**Example (Unconstrained convex-concave minimax)**

Consider the following (unconstrained) minimax problem

$$\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^m} f(x, y), \tag{1}$$

where $f$ is differentiable, $f(\cdot, y)$ is convex $\forall y \in \mathbb{R}^m$ and $f(x, \cdot)$ is concave $\forall x \in \mathbb{R}^n$.

**An operator view:**

- The gradient $\nabla_x f(\cdot, y) : \mathbb{R}^n \to \mathbb{R}^n$ is an operator
- We can compactly write $z = (x, y)$ and $F(z) = (\nabla_x f(x, y), -\nabla_y f(x, y))$
- The operator is thus a mapping $F : \mathbb{R}^d \to \mathbb{R}^d$ where $d = n + m$.
- The first order stationary point can be written as

$$F(z) = 0. \tag{2}$$

- We will write $Fz := F(z)$ for short.
- **Note that $F$ is not necessarily a linear operator**: $F(z_1 + z_2) \neq Fz_1 + Fz_2$ in general.

# Gradient descent ascent: why we flip the sign for $y$ in $Fz = (\nabla_x f(x,y), -\nabla_y f(x,y))$

## Gradient descent ascent

Consider the (simultaneous) gradient descent ascent (GDA)

$$x^{t+1} = x^t - \gamma_t \nabla_x f(x^t, y^t),$$
$$y^{t+1} = y^t + \gamma_t \nabla_y f(x^t, y^t).$$

**Remarks:**

- Using $F$ we can compactly write the update as

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

- The average iterate of GDA converges for convex-concave minimax if $\gamma_t$ is diminishing

$$\gamma_t \propto 1/\sqrt{t}$$

# Gradient descent ascent: why we flip the sign for $y$ in $Fz = (\nabla_x f(x,y), -\nabla_y f(x,y))$

## Gradient descent ascent

Consider the (simultaneous) gradient descent ascent (GDA)

$$x^{t+1} = x^t - \gamma_t \nabla_x f(x^t, y^t),$$
$$y^{t+1} = y^t + \gamma_t \nabla_y f(x^t, y^t).$$

**Remarks:**

- Using $F$ we can compactly write the update as

$$z^{t+1} = z^t - \gamma_t F z^t \qquad \text{(GDA)}$$

- The average iterate of GDA converges for convex-concave minimax if $\gamma_t$ is diminishing

$$\gamma_t \propto 1/\sqrt{t}$$

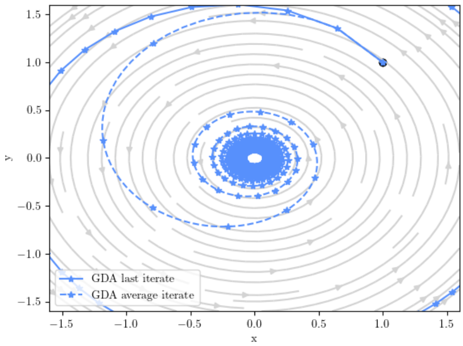
**Exercise:**

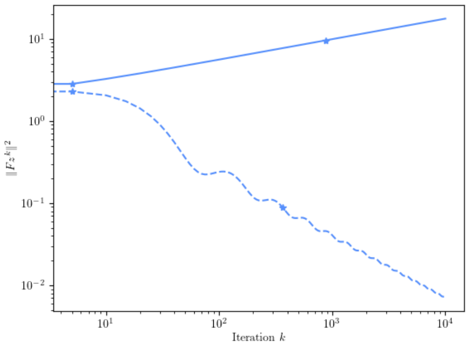
- *What online algorithms reduce to GDA in the unconstrained case?*
  - Deduce GDA from simultaneously played no-regret algorithms.

## An informative example: unconstrained bilinear game

$\min_{x\in\mathbb{R}^n} \max_{y\in\mathbb{R}^m} \langle x, My\rangle$

- Bilinear game are linear in both players: $Fz = (My, -M^\top x)$
- Captures the core problem: *rotation*



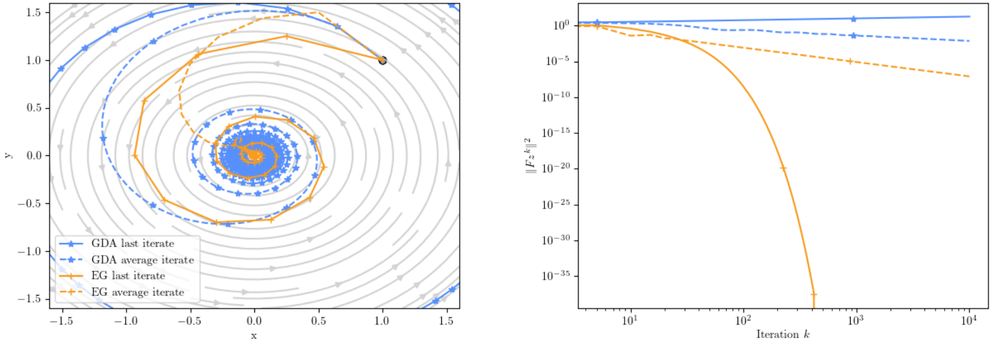

**Remarks:**

- The last iterate diverges!
- The *average* iterate converges as $\mathcal{O}(1/\sqrt{T})$ if we take $\gamma_t \propto 1/\sqrt{t}$

# Can we improve on the $\mathcal{O}(1/\sqrt{T})$ rate for the unconstrained bilinear game?

- Extragradient (EG) [2] takes an extrapolated step:

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




**Remarks:**

- The average iterate converges at a faster $\mathcal{O}(1/T)$.
- *Warning!* Bilinear can be misleading—rate for last iterate is linear (see next week).

## Warm-up to operator view: Analyzing GDA $z^{t+1} = z^t - \gamma F z^t$

- Under what conditions can we take the GDA stepsize $\gamma$ constant (and improve the rate to $\mathcal{O}(1/T)$)?
- **Goal**: find $z^\star \in \operatorname{zer} F$ where

$$\operatorname{zer} F := \{z \in \mathbb{R}^d \mid Fz = 0\}. \tag{3}$$

- To answer, we will begin by analyzing one step of the algorithm.

**Proof.**

$$\begin{aligned} \|z^{t+1} - z^\star\|^2 &= \|z^t - \gamma F z^t - z^\star\|^2 \\ &= \|z^t - z^\star\|^2 + \gamma^2 \|F z^t\|^2 - 2\gamma \langle F z^t, z^t - z^\star \rangle \end{aligned} \tag{4}$$

... (to be continued)

**Remark:**

- We need a way to convert $\langle F z^t, z^t - z^\star \rangle$ into $\|F z^t\|^2$. Then we would decrease:

$$\|z^{t+1} - z^\star\|^2 \overset{?}{\leq} \|z^t - z^\star\|^2 - \epsilon \|F z^t\|^2. \tag{5}$$

## Cocoerciveness

- Cocoercivity assumption can "convert" $\langle Fz^t, z^t - z^\star \rangle$ into $\|Fz^t\|^2$ (i.e., what we will need)

**Definition (Cocoercivity)**

*An operator $F : \mathbb{R}^d \to \mathbb{R}^d$ is said to be $\beta$-cocoercive for $\beta > 0$ if*

$$\langle Fz - Fz', z - z' \rangle \geq \beta \|Fz - Fz'\|^2 \quad \forall z, z' \in \mathbb{R}^d. \tag{6}$$

**Remarks:**

- Relationship to other structural assumptions:
  - $\beta$-cocoercivity implies monotonicity and $\frac{1}{\beta}$-Lipschitz continuity (defined later).
  - For a convex function $f$, $\nabla f$ is $L$-Lipschitz continuous iff $\nabla f$ is $\frac{1}{L}$-cocoercivity.
  - A $\mu$-strongly-monotone and $L$-Lipschitz continuous operator is also $\frac{\mu}{L^2}$-cocoercive.
- Due to the second point the result we are proving will apply to smooth convex minimization.

**Interpretation:**

- Geometrically, $\langle Fz^t, z^t - z^\star \rangle \geq \beta \|Fz^t\|^2$ ensure $-Fz^t$ points towards the solution set.

# Analysis GDA (continued)

## Proof (Cont.)

We can convert the inner product in (4) into $\|Fz^t\|^2$, by using cocoercivity on $z^t, z^\star$ and recalling that $Fz^\star = 0$ by assumption,

$$\langle Fz^t, z^t - z^\star\rangle = \langle Fz^t - Fz^\star, z^t - z^\star\rangle \geq \beta\|Fz^t - Fz^\star\|^2 = \|Fz^t\|^2,$$

such that (4) reduces to

$$\|z^{t+1} - z^\star\|^2 \leq \|z^t - z^\star\|^2 - (2\gamma\beta - \gamma^2)\|Fz^t\|^2.$$

Then, it is just a matter of summing and telescoping as follows:

$$\begin{aligned}\textstyle\sum_{t=0}^{T-1}(2\gamma\beta - \gamma^2)\|Fz^t\|^2 &\leq \textstyle\sum_{t=0}^{T-1}\|z^t - z^\star\|^2 - \|z^{t+1} - z^\star\|^2\\ &= \|z^0 - z^\star\|^2 - \|z^{T-1} - z^\star\|^2 \leq \|z^0 - z^\star\|^2,\end{aligned}$$

from which it immediately follows that

$$\frac{1}{T}\textstyle\sum_{t=0}^{T-1}\|Fz^t\|^2 \leq \frac{\|z^0 - z^\star\|^2}{(2\gamma\beta - \gamma^2)T}.$$

The proof is complete by noting that the minimum is always smaller than the average.

## GDA convergence under cocoercivity

**Theorem (Best iterate of (GDA))**

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

$$\min_{t\in\{0,\ldots,T-1\}} \|Fz^t\|^2 \le \frac{\|z^0 - z^\star\|^2}{\gamma(2\beta-\gamma)T}. \tag{7}$$

**Remarks:**

- The full range $\gamma \in (0, 2\beta)$ is allowed but the "optimal" choice is $\gamma = \beta$.
- The convergence rate is $\mathcal{O}(1/T)$.
- Implies convergence of (fixed stepsize) gradient descent for convex and $\frac{1}{\beta}$-Lipschitz.
- The *best* iterate can be hard to select in practice
  - next week we will derive *last* iterate convergence results

## Beyond cocoercivity: Lipschitz and monotone

- We have seen that cocoercivity implies monotone and Lipschitz, but the converse does not hold.
- Can we still get a $\mathcal{O}(1/T)$-rate in this more general setting?

### Definition (Monotone)

An operator $F : \mathbb{R}^d \to \mathbb{R}^d$ is said to be monotone if $\langle Fz - Fz', z - z' \rangle \geq 0 \quad \forall z, z' \in \mathbb{R}^d$.

**Examples:**

- For $F = \nabla f$ monotonicity reduces to convexity: $\langle \nabla f(z) - \nabla f(z'), z - z' \rangle \geq 0$
- For $F = (\nabla_x f, -\nabla_y f)$ monotonicity reduces to convex-concavity of $f(x, y)$

### Definition (Lipschitz)

An operator $F : \mathbb{R}^d \to \mathbb{R}^d$ is said to be $L$-Lipschitz for $L > 0$ if $\|Fz - Fz'\| \leq L\|z - z'\| \quad \forall z, z' \in \mathbb{R}^d$.

### Example (Bilinear game)

A simple example of an operator which is monotone and Lipschitz but is not cocoercive, is a skew-symmetric linear operator $F = M \in \mathbb{R}^{d \times d}$ (aka bilinear game).

**Exercise:**

- Convince yourself.

## Cocoercivity of $H = \mathrm{id} - \gamma F$

- Cocoercivity of $F$ is the key to convergence for GDA. What operator is cocoercive when $F$ is only Lipschitz?
- First attempt: the GDA update rule (also known as the *forward* operator)

$$Hz = z - \gamma F z. \tag{8}$$

- A quick computation shows that $H$ is indeed cocoercive!

**Lemma**

*Suppose $F$ is $L$-Lipschitz and $\gamma \leq 1/L$. Then, the mapping $H = \mathrm{id} - \gamma F$ is $1/2$-cocoercive for all $u \in \mathbb{R}^d$, where $\mathrm{id}$ is the identity operator. Specifically, it holds that*

$$\langle Hz - H\bar{z}, z - \bar{z} \rangle \geq \frac{1}{2}\|Hz - H\bar{z}\|^2 + \frac{1}{2}(1 - \gamma^2 L^2)\|z - \bar{z}\|^2 \quad \forall \bar{z}, z \in \mathbb{R}^d. \tag{9}$$

## Cocoercivity of $H = \mathrm{id} - \gamma F$

- Cocoercivity of $F$ is the key to convergence for GDA. What operator is cocoercive when $F$ is only Lipschitz?
- First attempt: the GDA update rule (also known as the *forward* operator)

$$Hz = z - \gamma F z. \tag{8}$$

- A quick computation shows that $H$ is indeed cocoercive!

### Lemma

*Suppose $F$ is $L$-Lipschitz and $\gamma \leq 1/L$. Then, the mapping $H = \mathrm{id} - \gamma F$ is $1/2$-cocoercive for all $u \in \mathbb{R}^d$, where $\mathrm{id}$ is the identity operator. Specifically, it holds that*

$$\langle Hz - H\bar{z}, z - \bar{z} \rangle \geq \tfrac{1}{2}\|Hz - H\bar{z}\|^2 + \tfrac{1}{2}(1 - \gamma^2 L^2)\|z - \bar{z}\|^2 \quad \forall \bar{z}, z \in \mathbb{R}^d. \tag{9}$$

### Proof.

$$\begin{aligned} \langle Hz - H\bar{z}, z - \bar{z} \rangle &= \langle Hz - H\bar{z}, Hz - H\bar{z} + \gamma F\bar{z} - \gamma F z \rangle \\ &= \frac{1}{2}\|Hz - H\bar{z}\|^2 - \frac{\gamma^2}{2}\|F\bar{z} - Fz\|^2 + \frac{1}{2}\|\bar{z} - z\|^2 \\ &\geq \frac{1}{2}\|Hz - H\bar{z}\|^2, \end{aligned} \tag{10}$$

where the last line follows from Lipschitzness of $F$ and from assuming $\gamma \leq 1/L$. □

## Using $H = \mathrm{id} - \gamma F$: Convergence for monotone and Lipschitz

∘ Building on the cocoercivity of the forward operator $H$, we can motivate the forward-forward method:

$$\begin{aligned} \bar{z}^t &= Hz^t \\ z^{t+1} &= z^t - \alpha(Hz^t - H\bar{z}^t), \end{aligned} \tag{FF}$$

where $\alpha > 0$ is a step-size

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

*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 FF with $\gamma \le 1/L$ and $\alpha \in (0, 1)$. Then, for all $z^\star \in \operatorname{zer} F$, it holds that*

$$\min_{t\in\{0,\ldots,T-1\}} \|Hz^t - H\bar{z}^t\|^2 \le \frac{\|z^0 - z^\star\|^2}{\alpha(1-\alpha)T}. \tag{11}$$

**Remark:**

- ∘ By the update rule, $Hz^t - H\bar{z}^t = \gamma F\bar{z}^t$, so convergence is given in terms of $\gamma^2\|F\bar{z}^t\|^2$.
- ∘ For this proof, we need to have $\alpha < 1$.
- ∘ When $\alpha = 1$, we stumble upon EG.

## Proof using $H = \mathrm{id} - \gamma F$

- For the GDA analysis before, we use cocoercivity of $F$ to cancel $\|z^{t+1} - z^t\|^2 = \gamma^2 \|Fz^t\|^2$
- Cocoercivity of $H$ gives us $-\|Hz - H\bar{z}\|^2$, motivating the following update rule

$$z^{t+1} = z^t - \alpha(Hz^t - H\bar{z}^t),$$

  where $\alpha > 0$ is a step-size and $\bar{z}^t$ is to be defined.

**Proof.**

Let us attempt to prove convergence by expanding the iterate as in the cocoercive case. Hence, we have

$$\|z^{t+1} - z^\star\|^2 = \|z^t - z^\star\|^2 + \alpha^2 \|Hz^t - H\bar{z}^t\|^2 - 2\alpha \langle Hz^t - H\bar{z}^t, z^t - z^\star \rangle. \tag{12}$$

We cannot immediately apply cocoercivity to the last term so we expand as follows

$$\begin{aligned} \langle Hz^t - H\bar{z}^t, z^t - z^\star \rangle &= \langle Hz^t - H\bar{z}^t, z^t - \bar{z}^t \rangle + \langle Hz^t - H\bar{z}^t, \bar{z}^t - z^\star \rangle \\ \text{(cocoercive } H) &\geq \tfrac{1}{2}\|Hz^t - H\bar{z}^t\|^2 + \langle Hz^t - H\bar{z}^t, \bar{z}^t - z^\star \rangle \\ \text{(monotone } F\text{—see remark)} &\geq \tfrac{1}{2}\|Hz^t - H\bar{z}^t\|^2 \end{aligned} \tag{13}$$

... (to be continued, also see Slide 24)

**Remark:**
- Pick $Hz^t - H\bar{z}^t = \gamma F\bar{z}^t$ so monotonicity applies to the last term.
- Equivalently, we can choose $\bar{z}^t = Hz^t$.

## Proof using $H = \mathrm{id} - \gamma F$

### Proof (Cont.)

Continuing from (12) and using (13), we have

$$\begin{aligned} \|z^{t+1} - z^\star\|^2 &= \|z^t - z^\star\|^2 + \alpha^2 \|Hz^t - H\bar{z}^t\|^2 - 2\alpha \langle Hz^t - H\bar{z}^t, z^t - z^\star \rangle \\ &\overset{(13)}{\leq} \|z^t - z^\star\|^2 - \alpha(1-\alpha)\|Hz^t - H\bar{z}^t\|^2. \end{aligned} \tag{14}$$

Summing and telescoping 14 completes the proof.

## Convergence of best $z^t$ and equivalence to extragradient EG

- What if we want to characterize another commonly used criterion $\|Fz^t\|^2 \leq \varepsilon$ instead?
- We can use the additional "good" term ($\|z^t - \bar{z}^t\|^2$) from cocoercivity of $H$:

**Theorem (Best $z$-iterate of FF)**

*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 FF with $\gamma < 1/L$ and $\alpha \in (0, 1]$. Then, for all $z^\star \in \operatorname{zer} F$, it holds that*

$$\min_{t\in\{0,\dots,T-1\}} \|z^t - \bar{z}^t\|^2 \leq \frac{\|z^0 - z^\star\|^2}{\alpha(1-\gamma^2L^2)T}. \tag{15}$$

**Remarks:**

- From the update rule, $z^t - \bar{z}^t = \gamma F z^t$, so we get convergence of $\gamma^2\|Fz^t\|^2$.
- The scheme reduces to extragradient (EG) for $\alpha = 1$

$$\bar{z}^t = Hz^t = z^t - \gamma F z^t,$$
$$z^{t+1} = z^t - \alpha(Hz^t - H\bar{z}^t) = z^t - \alpha\gamma F\bar{z}^t.$$

- Using $H$ will help us generalize to the constrained cases.

## Constrained problems as monotone inclusions

- So far the performance measure has been: $\|Fz\| \leq \varepsilon$.
- How can we treat constraints (and more generally a nonsmooth objective term $g$)?

### The proximal operator

$$\operatorname{prox}_{\lambda g}(z) := \arg\min_{z' \in \mathbb{R}^d} \left\{ \lambda g(z') + \frac{1}{2}\|z' - z\|^2 \right\}.$$

**Remark:**

- The proximal operator reduces to a projection on $\mathcal{Z} \subseteq \mathbb{R}^d$ when $g$ is the indictor function

$$g(z) = \delta_{\mathcal{Z}}(z) := \begin{cases} 0 & z \in \mathcal{Z} \\ \infty & \text{otherwise} \end{cases}$$

- Under convexity we can equivalent express the prox using the first order stationarity condition

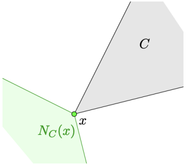
$$0 \in \lambda\partial g(z') + z' - z \quad \Leftrightarrow \quad z' = (\mathrm{id} + \lambda\partial g)^{-1}(z) =: \underbrace{J_{\lambda\partial g}}_{\text{resolvent}}(z)$$

- **Note that $\partial g$ is a *set-valued* operator ($A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$):** for any $z$ it assigns a subset $\partial g(z) \subseteq \mathbb{R}^d$.

# Constrained problems as monotone inclusions: an overview

| | | |
|---|---|---|
| $\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y)$ | $\min_{x\in\mathbb{R}^n} \max_{y\in\mathbb{R}^m} f(x,y) + g_1(x) - g_2(y)$ | $0 \in Sz := Fz + Az$ with $z = (x,y)$ |
| $x \in \mathcal{X}$ | $g_1(x) = \delta_{\mathcal{X}}(x) = \begin{cases} 0 & x \in \mathcal{X} \\ \infty & \text{otherwise} \end{cases}$ | |
| | $\partial g_1(x) = N_{\mathcal{X}}(x) = \{v \mid \langle v, x' - x\rangle \leq 0 \ \forall x' \in \mathcal{X}\}$ | $Az = (\partial g_1(x), \partial g_2(y))$ |
| $\Pi_{\mathcal{X}}(x)$ | $\mathrm{prox}_{\lambda g_1}(x) = (\mathrm{id} + \lambda\partial g)^{-1}(x)$ | $(\mathrm{id} + \lambda A)^{-1}z = (\mathrm{prox}_{\lambda g_1}(x), \mathrm{prox}_{\lambda g_2}(y))$ |
| $\mathcal{X}$ convex | $g_1$ is proper lsc convex | $A$ is *maximally* monotone |

**Remarks:**
- So far implicitly assumed *single-valued* operators (e.g., $F : \mathbb{R}^d \to \mathbb{R}^d$)
- The operator $A$ is *set-valued* (consider for instance the normal cone $N_{\mathcal{X}}$)
- To indicate $A$ assigns a subset of $\mathbb{R}^d$, we write $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$



## Properties of the resolvent $J_{\lambda A} := (\mathrm{id} + \lambda A)^{-1}$

- To treat constraints, we can consider inclusions: find $z \in \mathbb{R}^d$ such that

$$0 \in Sz := Fz + Az$$

- Unlike $F : \mathbb{R}^d \to \mathbb{R}^d$, we cannot use $Az^t$ in the algorithm since $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is set-valued!
- Instead, we use the resolvent to evaluate $A$ (it helps to think of it as a projection),

$$z' = (\mathrm{id} + \lambda A)^{-1} z \quad \Leftrightarrow \quad z' \in z - \lambda A z'.$$

**Lemma**

*When $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is maximally monotone, then the resolvent $J_{\lambda A}$ is*

(i) *single-valued, i.e., $J_{\lambda A} : \mathbb{R}^d \to \mathbb{R}^d$,*

(ii) *defined everywhere on $\mathbb{R}^d$.*

**Remarks:**

- From an algorithmic perspective both are crucial.
- Claim (i) is easy to prove, while (ii) is difficult to prove (in general)
- *Maximality* is a technical (but important) requirement (reviewed in the appendix)

## Modifying FF to handle a more general inclusion

- Let $\operatorname{zer} S := \{z \in \mathbb{R}^d \mid 0 \in Sz\}$.

**Inclusion problem**

Find $z \in \operatorname{zer} S$ where $S := A + F$.

**Remarks:**

- We could still run FF which only involves the operator $F$.
- *Issue*: Theorem 6 shows convergence of $\|Hz^t - H\bar{z}^t\| = \|\gamma F\bar{z}^t\|$ and not $\|F\bar{z}^t + A\bar{z}^t\|$.
- We can modify the update rule for $\bar{z}^t$ to satisfy this requirement

$$
\begin{aligned}
Hz^t - H\bar{z}^t \in \gamma F\bar{z}^t + \gamma A\bar{z}^t &\Leftrightarrow Hz^t \in \bar{z}^t + \gamma A\bar{z}^t \\
&\Leftrightarrow Hz^t \in (\mathrm{id} + \gamma A)\bar{z}^t \\
\text{(resolvent lemma)} &\Leftrightarrow \bar{z}^t = (\mathrm{id} + \gamma A)^{-1} Hz^t.
\end{aligned}
$$

## Convergence under constraints

- Based on the previous derivation, it is clear that we should modify FF as follows:

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

- We almost immediately obtain the following theorem.

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

*Suppose $F : \mathbb{R}^d \to \mathbb{R}^d$ 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 FBF with $\gamma \leq 1/L$ and $\alpha \in (0,1)$. Then, for $z^\star \in \operatorname{zer} S$, where $\operatorname{zer}$ denotes the optimality set, we have the following hold*

$$\min_{t \in \{0,\ldots,T-1\}} \|H z^t - H \bar{z}^t\|^2 \leq \frac{\|z^0 - z^\star\|^2}{\alpha(1-\alpha)T}. \tag{best iterate guarantee}$$

**Remarks:**

- We could rewrite as $\operatorname{dist}^2(0, \gamma S \bar{z}^t) \leq \|H z^t - H \bar{z}^t\|^2$ where $\operatorname{dist}(v, \mathcal{V}) := \min_{v' \in \mathcal{V}} \|v - v'\|$.
- Compare such a guarantee to a "average iterate guarantee"

## Proof for FBF

- The proof is almost immediate following the unconstrained result.

**Proof.**

- Cocoercivity of $H$ holds regardless of the redefinition of $\bar{z}^t$.
- The only step to re-verify is the use of monotonicity in (13).
- We will use that $S = F + A$ is monotone when $F$ and $A$ are monotone.
- Together with the definition, $Hz^t - H\bar{z}^t \in \gamma S\bar{z}^t$, and the fact that $0 \in Sz^\star$, it follows that

$$\langle Hz^t - H\bar{z}^t, \bar{z}^t - z^\star \rangle \geq 0.$$

□

# Solution concepts: Monotone inclusions and variational inequalities

## Monotone inclusion (MI)

So far we have considered the monotone inclusions using

$$\text{find } z^\star \in \mathbb{R}^d \text{ such that } 0 \in Fz^\star + Az^\star, \tag{MI}$$

where $F : \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz and $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is maximally monotone.

**Remark:**

- For constrained problems take $A = N_{\mathcal{Z}} := \{v \mid \langle v, z' - z\rangle \leq 0,\ \forall z' \in \mathcal{Z}\}$ (i.e., normal cone).

## Variational inequality

- The Stampacchia variational inequality

$$\text{find } z^\star \in \mathbb{R}^d \text{ such that } \langle Fz^\star, z - z^\star\rangle \geq 0 \quad \forall z \in \mathcal{Z}. \tag{SVI}$$

- The Minty variational inequality

$$\text{find } z^\star \in \mathbb{R}^d \text{ such that } \langle Fz, z - z^\star\rangle \geq 0 \quad \forall z \in \mathcal{Z}. \tag{MVI}$$

**Remarks:**

- (SVI) is the first order condition of a (possibly nonconvex) constrained problem.
- A star-convex function satisfies (MVI): Notice the close resemblance to linearized regret.

## Solution concepts: relationships

- We have the following relations (see [5, 1] and [3] for additional discussion)

**Lemma**

*For $F : \mathbb{R}^d \to \mathbb{R}^d$ and $A = N_{\mathcal{Z}}$, the following holds*

1. (MI) $\Leftrightarrow$ (SVI)
2. (SVI) $\Leftarrow$ (MVI) *if $F$ is Lipschitz and $\mathcal{Z}$ is convex*
3. (SVI) $\Rightarrow$ (MVI) *if $F$ is monotone*

**Proof.**

The equivalence between (MI) and (SVI) follows immediately from the following argument

$$0 \in Fz^\star + N_{\mathcal{Z}}(z^\star) \;\Leftrightarrow\; -Fz^\star \in N_{\mathcal{Z}}(z^\star) \;\Leftrightarrow\; \langle Fz^\star, z - z^\star \rangle \geq 0 \quad \forall z \in \mathcal{Z}.$$

The third claim follows directly from the definition of monotonicity. □

**Remark:**

- Consequently we can translate our (best iterate) convergence results for (MI) into (SVI).
- We will now show convergence for the average iterate.
- As we will see, it is easier to show convergence to (MVI).
- Through monotonicity we directly obtain the (SVI).

# Restricted gap function

- For VIs, we will use the *restricted* gap function [4] as a measure of progress

$$\mathrm{Gap}_{\mathcal{C}}(z) = \sup_{z^\star \in \mathcal{C}} \langle F z^\star, z - z^\star \rangle + g(z) - g(z^\star), \tag{16}$$

  where $F : \mathbb{R}^d \to \mathbb{R}^d$, $g$ is a proper convex lower semicontinuous function, and $\mathcal{C}$ is a compact subset of $\mathbb{R}^d$.
- Without restricting $z^\star$ to $\mathcal{C}$: the gap could be infinite everywhere (except at the solution)!

**Example**

Consider the unconstrained problem $\min_{x\in\mathbb{R}} \max_{y\in\mathbb{R}} xy$ where $Fz = (\nabla_x f(x,y), -\nabla_y f(x,y)) = (y, -x)$.

**Remark:**
- If not restricted:
  - the gap would be infinite except at the unique solution $z^\star = (0,0)$.
  - Thus useless as a measure of progress for approximate methods.

## Restricted gap function: conversion

- We will now show a "last" iterate guarantee on the *average* iterate in VIs.
- A guarantee on the iterates *on average* is sufficient... (due to monotonicity!)

**Lemma**

*Let* $S := F + A$, $F : \mathbb{R}^d \to \mathbb{R}^d$ *be monotone and* $A = \partial g$, *where* $g$ *is proper lsc convex. Generate* $(z^t)_{t\in\mathbb{N}}$ *and take* $\hat{z}^T = \frac{1}{T}\sum_{t=0}^{T-1} z^t$. *Then for all* $z^\star \in \operatorname{zer} S$ *which also belongs to* $\mathcal{C}$,

$$\operatorname{Gap}_{\mathcal{C}}(\hat{z}^T) \le \frac{1}{T}\sum_{t=0}^{T-1}\langle v^t, z^t - z^\star\rangle \quad \forall v^t \in S z^t. \qquad \text{(average iterate guarantee)}$$

**Remarks:**

- Compare this guarantee with "best iterate guarantee"

## Proof of restricted gap function conversion lemma

◦ The lemma below follows directly from the convexity of $\partial g$ and the monotonicity of $F$.

**Proof.**

Let $z^\star \in \operatorname{zer} S$ and $u^t \in \partial g(z^t)$. From convexity of $\partial g$, we have that for all $u \in \partial g(z)$

$$g(z) \leq g(z^\star) + \langle u, z - z^\star \rangle, \tag{17}$$

from which it immediately follows that

$$\begin{aligned}\sum_{t=0}^{T-1} \langle Fz^t + u^t, z^t - z^\star \rangle &= \sum_{t=0}^{T-1} \langle Fz^t, z^t - z^\star \rangle + \langle u^t, z^t - z^\star \rangle \\ &\overset{(17)}{\geq} \sum_{t=0}^{T-1} \langle Fz^t, z^t - z^\star \rangle + g(\hat{z}^T) - g(z^\star) \geq \langle Fz^\star, \hat{z}^T - z^\star \rangle + g(\hat{z}^T) - g(z^\star),\end{aligned}$$

where the last line follows from monotonicity of $F$. The proof is complete by restricting $z^\star$ to $\mathcal{C}$. □

## Restricted gap function: warmup by revisiting GDA

- Applying cocoercivity to the inner product in (4) is a choice. We can also apply cocoercivity to the norm:

$$\|z^{t+1} - z^\star\|^2 \le \|z^t - z^\star\|^2 - (2\gamma - \frac{\gamma^2}{\beta})\langle Fz^t, z^t - z^\star\rangle.$$

- Summing and telescoping, we get convergence on a different performance measure (assuming $\gamma < 2\beta$):

$$\frac{1}{T}\sum_{t=0}^{T-1}\langle Fz^t, z^t - z^\star\rangle \le \frac{\|z^0 - z^\star\|^2}{\gamma(2 - \frac{\gamma}{\beta})T},$$

- The above can be converted into a guarantee on the restricted gap using the previous lemma:

**Theorem (Gap for average iterate of (GDA))**

*Suppose* $F : \mathbb{R}^d \to \mathbb{R}^d$ *is* $\beta$*-cocoercive. Consider the sequence* $(z^t)_{t\in\mathbb{N}}$ *generated by* (GDA) *with* $\gamma < 2\beta$*. Then, for all* $z^\star \in \mathrm{zer}\, F$ *and any compact neighborhood* $\mathcal{C} \subseteq \mathbb{R}^d$ *of* $z^\star$*, it holds that*

$$\mathrm{Gap}_{\mathcal{C}}(\hat{z}^T) \le \frac{\|z^0 - z^\star\|^2}{\gamma(2 - \frac{\gamma}{\beta})T},$$

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

## Restricted gap: forward-backward-forward (FBF)

◦ A similar argument as in the analysis for (GDA) also applies to (FBF).

### Theorem (Gap for average iterate of FBF)

*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 FBF with* $\gamma \leq 1/L$ *and* $\alpha \in (0, 1]$*. Then, for all* $z^\star \in \operatorname{zer} S$ *and any compact neighborhood* $\mathcal{C} \subseteq \mathbb{R}^d$ *of* $z^\star$*,*

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

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

**Remark:** ◦ Since (SVI) ⇔ (MI) under monotonicity, the average iterate also converges in norm.

# Proof of restricted gap for FBF

**Proof.**

- The descent inequality before applying monotonicity and keeping "good" term from cocoercivity of $H$:

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

- The norms can be made negative, so by summing and telescoping we get:

$$\sum_{t=0}^{T-1}\langle v^t, z^t - z^\star\rangle \leq \frac{\|z^0 - z^\star\|^2}{2\alpha\gamma T} \quad \forall v^t \in Sz^t.$$

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

□

# Summary

We have seen:

- Best iterate and average iterate (next week last iterate)
- Descent inequality $\Rightarrow$ convergence of residual, operator norm, gap (next week iterate distance)
- How we arrived at extragradient-like schemes is still mysterious (next week we will see two elegant derivations)

# Appendix

# Monotone operators

- A set-valued mappings, $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^d$, maps each input $x \in \mathbb{R}^n$ to a subset $Sx \subseteq \mathbb{R}^d$.
- The domain of $S$ is defined as

$$\operatorname{dom} S = \{x \in \mathbb{R}^n \mid Sx \neq \emptyset\}. \tag{18}$$

- All input-value pairs are called the graph of $S$, denoted as

$$\operatorname{gph} S = \{(x, y) \mid y \in Sx\}, \tag{19}$$

and the inverse of $S$ is defined through its graph via

$$\operatorname{gph} S^{-1} = \{(y, x) \mid y \in Sx\}. \tag{20}$$

Notice that, by the definition, the inverse always exists.

# Maximal monotonicity

## Definition (Monotone)

An operator $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^d$ is said to be monotone if $\langle u - v, x - y \rangle \geq 0 \quad \forall (x, u), (y, v) \in \operatorname{gph} S$.

**Remark:**

- A more stringent condition, which might seem technical at first, is the notion of maximality.

## Definition (Maximally monotone)

An operator $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^d$ is said to be maximally monotone if it is not strictly contained within the graph of another monotone operator.

**Remarks:**

- We should not be able to add $(x, u) \notin \operatorname{gph} A$ to $\operatorname{gph} S$ without violating monotonicity.
- Geometrically, for 1-dim ($A : \mathbb{R} \rightrightarrows \mathbb{R}$), it corresponds to having no "holes" in the line characterizing the graph.

## Maximal monotonicity is important algorithmically

- *Monotonicity* ensures resolvent $J_A = (\operatorname{id} + A)^{-1}$ is single valued. If $J_A$ was instead set-valued, then one update of the iteration $z^{t+1} = J_A(z^t)$ could potentially leave us with a *set* of iterates!
- *Maximality* ensures $\operatorname{dom} J_A = \mathbb{R}^d$. If the domain was further restricted then the update rule $z^{t+1} = Jz^t$ would be undefined for some input.

# References I

[1] Sándor Komlósi.
On the stampacchia and minty variational inequalities.
*Generalized Convexity and Optimization for Economic and Financial Decisions*, pages 231–260, 1999.
(Cited on page 29.)

[2] Galina M Korpelevich.
The extragradient method for finding saddle points and other problems.
*Matecon*, 12:747–756, 1976.
(Cited on page 10.)

[3] Panayotis Mertikopoulos, Ya-Ping Hsieh, and Volkan Cevher.
Learning in games from a stochastic approximation viewpoint.
*arXiv preprint arXiv:2206.03922*, 2022.
(Cited on page 29.)

[4] Yurii Nesterov.
Dual extrapolation and its applications to solving variational inequalities and related problems.
*Mathematical Programming*, 109(2):319–344, 2007.
(Cited on page 30.)

# References II

[5] Michael Patriksson and R Tyrrell Rockafellar.
Variational geometry and equilibrium.
In *Equilibrium Problems and Variational Models*, pages 347–368. Springer, 2003.
(Cited on page 29.)