# Weaker MVI Condition

## Extragradient Methods with Multi-Step Exploration

Yifeng Fan  Yongqiang Li  Bo Chen

Zhejiang University of Technology

ICLR 2024

# Min-max optimization

## Problem setting

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})$$

# Min-max optimization

## Problem setting

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})$$

- Saddle gradient operator
  $F\boldsymbol{z} = \begin{bmatrix} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}) \\ -\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}) \end{bmatrix}$, where $\boldsymbol{z} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}$

# Min-max optimization

**Problem setting**

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})$$

- Saddle gradient operator
  $F\boldsymbol{z} = \begin{bmatrix} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}) \\ -\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}) \end{bmatrix}$, where $\boldsymbol{z} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}$
- Scenaries: GANs training, adversarial training, robust learning

# Min-max optimization

## Problem setting

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})$$

- Saddle gradient operator
  $F\boldsymbol{z} = \begin{bmatrix} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}) \\ -\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}) \end{bmatrix}$, where $\boldsymbol{z} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}$
- Scenaries: GANs training, adversarial training, robust learning
- $F$ may not be monotone

# Min-max optimization

## Problem setting

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})$$

- Saddle gradient operator
  $F\boldsymbol{z} = \begin{bmatrix} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}) \\ -\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}) \end{bmatrix}$, where $\boldsymbol{z} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}$
- Scenaries: GANs training, adversarial training, robust learning
- $F$ may not be monotone
- Assumptions on $F$

# Min-max optimization

## Problem setting

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})$$

- Saddle gradient operator
  $F\boldsymbol{z} = \begin{bmatrix} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}) \\ -\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}) \end{bmatrix}$, where $\boldsymbol{z} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}$
- Scenaries: GANs training, adversarial training, robust learning
- $F$ may not be monotone
- Assumptions on $F$
  - $L$-Lipschitz $\quad \|F\boldsymbol{u} - F\boldsymbol{v}\| \leq L\|\boldsymbol{u} - \boldsymbol{v}\|$

# Min-max optimization

## Problem setting

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})$$

- Saddle gradient operator
  $F\boldsymbol{z} = \begin{bmatrix} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}) \\ -\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}) \end{bmatrix}$, where $\boldsymbol{z} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}$
- Scenaries: GANs training, adversarial training, robust learning
- $F$ may not be monotone
- Assumptions on $F$
  - $L$-Lipschitz $\|F\boldsymbol{u} - F\boldsymbol{v}\| \leq L\|\boldsymbol{u} - \boldsymbol{v}\|$
  - weak MVI condition $\langle F\boldsymbol{z}, \boldsymbol{z} - \boldsymbol{z}^* \rangle \geq \rho \|F\boldsymbol{z}\|^2$

# Min-max optimization

**Problem setting**

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})$$

- Saddle gradient operator
  $F\boldsymbol{z} = \begin{bmatrix} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}) \\ -\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}) \end{bmatrix}$, where $\boldsymbol{z} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}$
- Scenaries: GANs training, adversarial training, robust learning
- $F$ may not be monotone
- Assumptions on $F$
  - $L$-Lipschitz $\quad \|F\boldsymbol{u} - F\boldsymbol{v}\| \leq L\|\boldsymbol{u} - \boldsymbol{v}\|$
  - weak MVI condition $\quad \langle F\boldsymbol{z}, \boldsymbol{z} - \boldsymbol{z}^* \rangle \geq \rho\|F\boldsymbol{z}\|^2$
    - If $\rho = 0$, MVI (star-monotonicity)

# Min-max optimization

## Problem setting

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})$$

- Saddle gradient operator
  $F\boldsymbol{z} = \begin{bmatrix} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}) \\ -\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}) \end{bmatrix}$, where $\boldsymbol{z} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}$
- Scenaries: GANs training, adversarial training, robust learning
- $F$ may not be monotone
- Assumptions on $F$
  - $L$-Lipschitz $\quad \|F\boldsymbol{u} - F\boldsymbol{v}\| \leq L\|\boldsymbol{u} - \boldsymbol{v}\|$
  - weak MVI condition $\quad \langle F\boldsymbol{z}, \boldsymbol{z} - \boldsymbol{z}^* \rangle \geq \rho \|F\boldsymbol{z}\|^2$
    - If $\rho = 0$, MVI (star-monotonicity)
    - Weak MVI allows $\rho < 0$

# Weak MVI Condition

**Weak MVI Condition (star-cohypomonotonicity)**

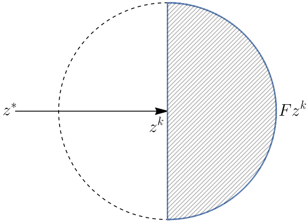
$$\langle F\boldsymbol{z}, \boldsymbol{z} - \boldsymbol{z}^* \rangle \geq \rho \|F\boldsymbol{z}\|^2$$

# Weak MVI Condition

**Weak MVI Condition (star-cohypomonotonicity)**

$$\langle F\boldsymbol{z}, \boldsymbol{z} - \boldsymbol{z}^* \rangle \geq \rho \|F\boldsymbol{z}\|^2$$

- Minty variational inequality $\langle F\boldsymbol{z}, \boldsymbol{z} - \boldsymbol{z}^* \rangle \geq 0$



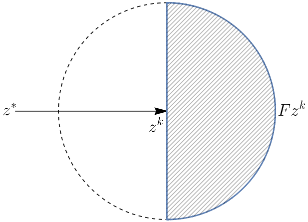

MVI: negative gradient points towards zero

# Weak MVI Condition

**Weak MVI Condition (star-cohypomonotonicity)**

$$\langle F\boldsymbol{z}, \boldsymbol{z} - \boldsymbol{z}^* \rangle \geq \rho \|F\boldsymbol{z}\|^2$$

- Minty variational inequality
  $\langle F\boldsymbol{z}, \boldsymbol{z} - \boldsymbol{z}^* \rangle \geq 0$
- monotone
  $\langle F\boldsymbol{z} - F\boldsymbol{v}, \boldsymbol{z} - \boldsymbol{v} \rangle \geq 0$



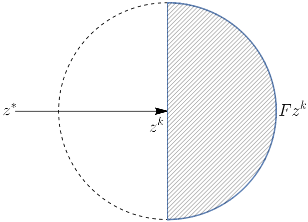

MVI: negative gradient points towards zero

# Weak MVI Condition

## Weak MVI Condition (star-cohypomonotonicity)

$$\langle F\boldsymbol{z}, \boldsymbol{z} - \boldsymbol{z}^* \rangle \geq \rho \|F\boldsymbol{z}\|^2$$

- Minty variational inequality
  $\langle F\boldsymbol{z}, \boldsymbol{z} - \boldsymbol{z}^* \rangle \geq 0$
- monotone
  $\langle F\boldsymbol{z} - F\boldsymbol{v}, \boldsymbol{z} - \boldsymbol{v} \rangle \geq 0$
- $\rho$-comonotone
  $\langle F\boldsymbol{z} - F\boldsymbol{v}, \boldsymbol{z} - \boldsymbol{v} \rangle \geq \rho \|F\boldsymbol{z} - F\boldsymbol{v}\|^2$



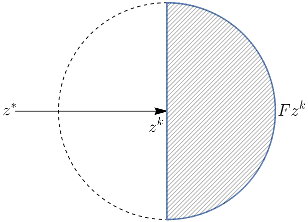

MVI: negative gradient points towards zero

# Weak MVI Condition

## Weak MVI Condition (star-cohypomonotonicity)

$$\langle F\boldsymbol{z}, \boldsymbol{z} - \boldsymbol{z}^* \rangle \geq \rho \|F\boldsymbol{z}\|^2$$

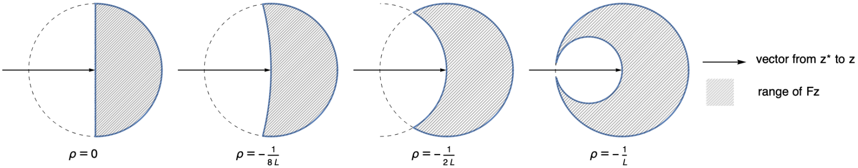
- Minty variational inequality
  $\langle F\boldsymbol{z}, \boldsymbol{z} - \boldsymbol{z}^* \rangle \geq 0$
- monotone
  $\langle F\boldsymbol{z} - F\boldsymbol{v}, \boldsymbol{z} - \boldsymbol{v} \rangle \geq 0$
- $\rho$-comonotone
  $\langle F\boldsymbol{z} - F\boldsymbol{v}, \boldsymbol{z} - \boldsymbol{v} \rangle \geq \rho \|F\boldsymbol{z} - F\boldsymbol{v}\|^2$
- star: $\boldsymbol{v} = \boldsymbol{z}^*$, hypo: $\rho < 0$




MVI: negative gradient points towards zero

# Weak MVI Condition: Intuitions

- Weak MVI: negative gradient might point away from zero




weak MVI with different parameters

# Existing Works

- Extragradient (MVI)
  - same stepsizes

$$\bar{\boldsymbol{z}}^k = \boldsymbol{z}^k - \alpha_k F \boldsymbol{z}^k$$
$$\boldsymbol{z}^{k+1} = \boldsymbol{z}^k - \alpha_k F \bar{\boldsymbol{z}}^k$$

# Existing Works

- Extragradient (MVI)
  - same stepsizes

$$\bar{\boldsymbol{z}}^k = \boldsymbol{z}^k - \alpha_k F\boldsymbol{z}^k$$
$$\boldsymbol{z}^{k+1} = \boldsymbol{z}^k - \alpha_k F\bar{\boldsymbol{z}}^k$$

- EG+/AdaptiveEG+ (weak MVI, $\rho > -1/2L$)
  - larger extrapolation stepsize $\gamma_k$

$$\bar{\boldsymbol{z}}^k = \boldsymbol{z}^k - \gamma_k F\boldsymbol{z}^k$$
$$\boldsymbol{z}^{k+1} = \boldsymbol{z}^k - \alpha_k F\bar{\boldsymbol{z}}^k$$

# Our Contributions

- First known first-order algorithm that converges for weak MVI problems with $\rho < -1/2L$.
  We provide convergence guarantee for $\rho > -0.632/L$.

# Our Contributions

- First known first-order algorithm that converges for weak MVI problems with $\rho < -1/2L$.
  We provide convergence guarantee for $\rho > -0.632/L$.

- Adaptive method effectively resolves problems with limit cycles.

# Multi-step Extragradient

## 2-step Extragradient

$$
\begin{aligned}
\boldsymbol{z}_1^k &= \boldsymbol{z}^k - \gamma_{k,1} F \boldsymbol{z}^k \\
\bar{\boldsymbol{z}}^k &= \boldsymbol{z}_1^k - \gamma_{k,2} F \boldsymbol{z}_1^k \\
\boldsymbol{z}^{k+1} &= \boldsymbol{z}^k - \alpha_k F \bar{\boldsymbol{z}}^k
\end{aligned}
$$

# Multi-step Extragradient

## 2-step Extragradient

$$\begin{aligned} \boldsymbol{z}_1^k &= \boldsymbol{z}^k - \gamma_{k,1} F \boldsymbol{z}^k \\ \bar{\boldsymbol{z}}^k &= \boldsymbol{z}_1^k - \gamma_{k,2} F \boldsymbol{z}_1^k \\ \boldsymbol{z}^{k+1} &= \boldsymbol{z}^k - \alpha_k F \bar{\boldsymbol{z}}^k \end{aligned}$$

Take stepsizes $\gamma_{k,1} = {}^{\delta_1}/_L$, $\gamma_{k,2} = {}^{\delta_2}/_L$

$$\rho > \begin{cases} -\frac{1}{L}\left[1 - \frac{1}{(1+\delta_1)(1+\delta_2)}\right] & \text{if } \delta_1 + \delta_2 \le 1 \\ -\frac{1}{L}\left[\frac{\delta_1(1-\delta_1^2-\delta_2^2)}{2(1-\delta_1^2)(1-\delta_2^2)} + \frac{\delta_2}{1+\delta_2}\right] & \text{if } \delta_1 + \delta_2 > 1 \end{cases}$$

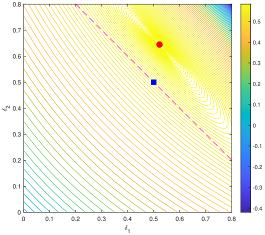
# Multi-step Extragradient

## 2-step Extragradient

$$\begin{aligned} z_1^k &= z^k - \gamma_{k,1} F z^k \\ \bar{z}^k &= z_1^k - \gamma_{k,2} F z_1^k \\ z^{k+1} &= z^k - \alpha_k F \bar{z}^k \end{aligned}$$

Take stepsizes $\gamma_{k,1} = {}^{\delta_1}/_{L}$, $\gamma_{k,2} = {}^{\delta_2}/_{L}$

$$\rho > \begin{cases} -\frac{1}{L}\left[1 - \frac{1}{(1+\delta_1)(1+\delta_2)}\right] & \text{if } \delta_1 + \delta_2 \leq 1 \\ -\frac{1}{L}\left[\frac{\delta_1(1-\delta_1^2-\delta_2^2)}{2(1-\delta_1^2)(1-\delta_2^2)} + \frac{\delta_2}{1+\delta_2}\right] & \text{if } \delta_1 + \delta_2 > 1 \end{cases}$$




Relation between the upper bound of $-\rho L$ and $\delta_1$, $\delta_2$

# Multi-step Extragradient

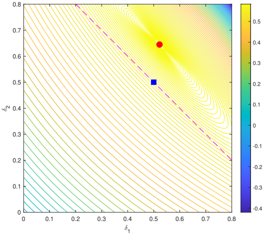
## 2-step Extragradient

$$z_1^k = z^k - \gamma_{k,1} F z^k$$
$$\bar{z}^k = z_1^k - \gamma_{k,2} F z_1^k$$
$$z^{k+1} = z^k - \alpha_k F \bar{z}^k$$

Take stepsizes $\gamma_{k,1} = {}^{\delta_1}/_L$, $\gamma_{k,2} = {}^{\delta_2}/_L$

$$\rho > \begin{cases} -\frac{1}{L}\left[1 - \frac{1}{(1+\delta_1)(1+\delta_2)}\right] & \text{if } \delta_1 + \delta_2 \leq 1 \\ -\frac{1}{L}\left[\frac{\delta_1(1-\delta_1^2-\delta_2^2)}{2(1-\delta_1^2)(1-\delta_2^2)} + \frac{\delta_2}{1+\delta_2}\right] & \text{if } \delta_1 + \delta_2 > 1 \end{cases}$$

- $\delta_1 = \delta_2 = {}^1/_2$, $\rho > -{}^5/_{9L}$




Relation between the upper bound of $-\rho L$ and $\delta_1$, $\delta_2$

# Multi-step Extragradient

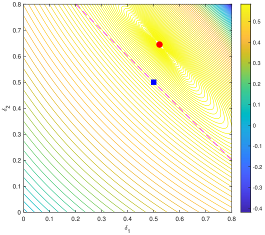
## 2-step Extragradient

$$\begin{aligned} \boldsymbol{z}_1^k &= \boldsymbol{z}^k - \gamma_{k,1} F \boldsymbol{z}^k \\ \bar{\boldsymbol{z}}^k &= \boldsymbol{z}_1^k - \gamma_{k,2} F \boldsymbol{z}_1^k \\ \boldsymbol{z}^{k+1} &= \boldsymbol{z}^k - \alpha_k F \bar{\boldsymbol{z}}^k \end{aligned}$$

Take stepsizes $\gamma_{k,1} = {}^{\delta_1}/_L$, $\gamma_{k,2} = {}^{\delta_2}/_L$

$$\rho > \begin{cases} -\frac{1}{L}\left[1 - \frac{1}{(1+\delta_1)(1+\delta_2)}\right] & \text{if } \delta_1 + \delta_2 \le 1 \\ -\frac{1}{L}\left[\frac{\delta_1(1-\delta_1^2-\delta_2^2)}{2(1-\delta_1^2)(1-\delta_2^2)} + \frac{\delta_2}{1+\delta_2}\right] & \text{if } \delta_1 + \delta_2 > 1 \end{cases}$$

- $\delta_1 = \delta_2 = {}^1/_2$, $\rho > -{}^5/_{9L}$
- $\delta_1 \approx 0.52212$, $\delta_2 \approx 0.644793$, $\rho > -{}^{0.5834}/_L$




Relation between the upper bound of $-\rho L$ and $\delta_1$, $\delta_2$

# Multi-step Extragradient

## $n$-step Extragradient

$$
\begin{aligned}
\mathbf{z}_1^k &= \mathbf{z}^k - \gamma_{k,1} F \mathbf{z}^k \\
\mathbf{z}_2^k &= \mathbf{z}_1^k - \gamma_{k,2} F \mathbf{z}_1^k \\
&\vdots \\
\bar{\mathbf{z}}^k &= \mathbf{z}_{n-1}^k - \gamma_{k,n-1} F \mathbf{z}_{n-1}^k \\
\mathbf{z}^{k+1} &= \mathbf{z}^k - \alpha_k F \bar{\mathbf{z}}^k
\end{aligned}
$$

# Multi-step Extragradient

## $n$-step Extragradient

$$
\begin{aligned}
\mathbf{z}_1^k &= \mathbf{z}^k - \gamma_{k,1} F\mathbf{z}^k \\
\mathbf{z}_2^k &= \mathbf{z}_1^k - \gamma_{k,2} F\mathbf{z}_1^k \\
&\vdots \\
\bar{\mathbf{z}}^k &= \mathbf{z}_{n-1}^k - \gamma_{k,n-1} F\mathbf{z}_{n-1}^k \\
\mathbf{z}^{k+1} &= \mathbf{z}^k - \alpha_k F\bar{\mathbf{z}}^k
\end{aligned}
$$

- Take $\gamma_{k,i} = {}^{\delta_i}/_{L}$, $i \in [n]$. If $\sum_{i=1}^n \delta_i \leq 1$,

$$
\rho > -\frac{1}{L}\left(1 - \prod_{i=1}^{n} \frac{1}{1+\delta_i}\right)
$$

# Multi-step Extragradient

**$n$-step Extragradient**

$$
\begin{aligned}
\boldsymbol{z}_1^k &= \boldsymbol{z}^k - \gamma_{k,1} F \boldsymbol{z}^k \\
\boldsymbol{z}_2^k &= \boldsymbol{z}_1^k - \gamma_{k,2} F \boldsymbol{z}_1^k \\
&\vdots \\
\bar{\boldsymbol{z}}^k &= \boldsymbol{z}_{n-1}^k - \gamma_{k,n-1} F \boldsymbol{z}_{n-1}^k \\
\boldsymbol{z}^{k+1} &= \boldsymbol{z}^k - \alpha_k F \bar{\boldsymbol{z}}^k
\end{aligned}
$$

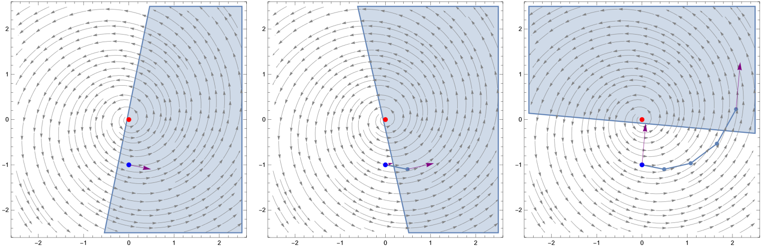
- Take $\gamma_{k,i} = {}^{\delta_i}/_{L}, i \in [n]$. If $\sum_{i=1}^n \delta_i \leq 1$,

$$
\rho > -\frac{1}{L}\left(1 - \prod_{i=1}^{n} \frac{1}{1+\delta_i}\right)
$$

- Convergence guarantee for $\rho > -{}^{(1-1/e)}/_{L} \approx -{}^{0.632}/_{L}$ with invariant stepsizes $\gamma_{k,i} = {}^{1}/_{nL}, i \in [n]$ and large enough $n$

# Introducing Adaptive Exploration

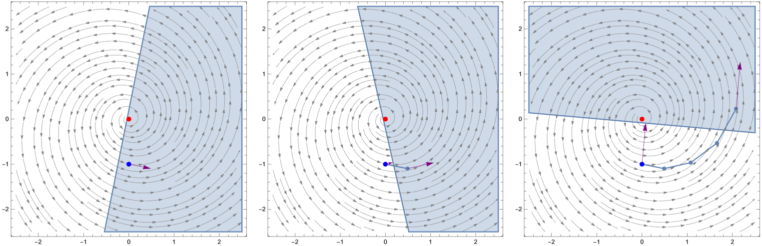
- Perspective of projection



(a) GDA (b) AdaptiveEG+ (c) MDEG

# Introducing Adaptive Exploration

- Perspective of projection

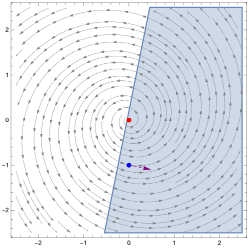



(a) GDA (b) AdaptiveEG+ (c) MDEG

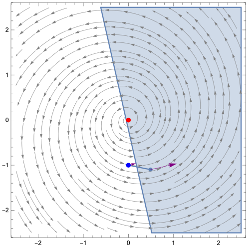
- MDEG: Max Distance Extragradient

# Introducing Adaptive Exploration

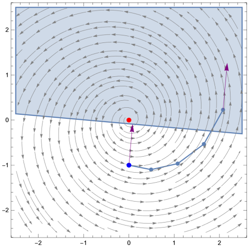
- Perspective of projection




(a) GDA




(b) AdaptiveEG+

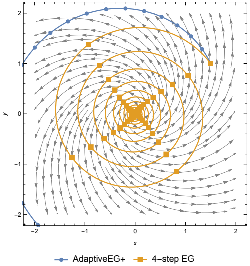



(c) MDEG

- MDEG: Max Distance Extragradient
- Sub-iteration stops when projection distance decreases

# Examples

- *n*-step EG converges for $\rho < -1/2L$.

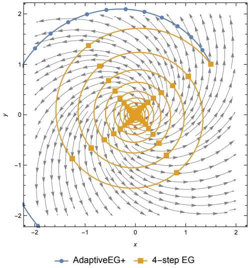



(a) $\rho L \approx -0.577$

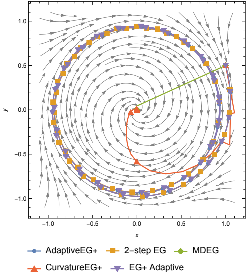
(b)

(c)

- $n$-step EG converges for $\rho < -1/2L$.
- MDEG converges break out of the limit cycle and converges to the stationary point.




(a) $\rho L \approx -0.577$

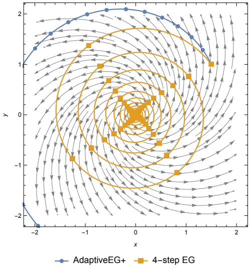



(b) $\rho L \approx -0.885521$

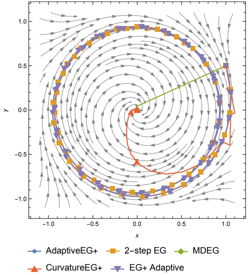
(c)

# Examples

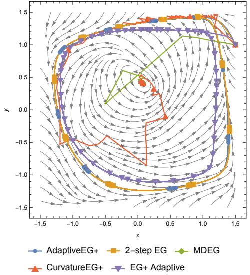
- *n*-step EG converges for $\rho < -1/2L$.
- MDEG converges break out of the limit cycle and converges to the stationary point.
- MDEG bypasses the highly nonmonotonic regions with local $\rho L \approx -3.04076$ and converges to the stationary point.




(a) $\rho L \approx -0.577$




(b) $\rho L \approx -0.885521$




(c) $\rho L \approx -3.04076$

*Thank you!*