Optimization for Data Science
ETH Zürich, FS 2022 261-5110-00L

Lecture 13: Min-Max Optimization, Part I

Bernd Gärtner
Niao He

https://www.ti.inf.ethz.ch/ew/courses/ODS22/index.html
May 23, 2022

Lecture Outline

Why Min-Max Optimization?

Saddle Points and Global Minimax Points

The Classical Setting: Convex-Concave Min-Max Optimization

First-order Methods

- Gradient Descent Ascent (GDA)
- Extragradient Method
- Optimistic GDA

Extension to Concave Games, Variational Inequalities

Lecture Outline

Why Min-Max Optimization?

Saddle Points and Global Minimax Points

The Classical Setting: Convex-Concave Min-Max Optimization

First-order Methods

- Gradient Descent Ascent (GDA)
- Extragradient Method
- Optimistic GDA

Extension to Concave Games, Variational Inequalities

Min-Max Optimization

Let $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{Y} \subset \mathbb{R}^p$ and $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Consider the min-max problem:

$$\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} \phi(\mathbf{x}, \mathbf{y})$$

Wide applications in machine learning:

- Zero-sum matrix games
- Nonsmooth optimization
- Generative adversarial networks
- Distributionally robust optimization
- ....

Zero-sum matrix games

- 2-players games where players have opposite evaluations of outcomes:
  - I (resp. J) non-empty finite set of strategies of player 1 (resp. player 2).
  - payoff of player 1 given by a real-valued $I \times J$ matrix $\mathbf{A}$ (resp. $-\mathbf{A}$ for player 2).
  - Set of mixed strategies $\Delta(I) = \{\mathbf{x} \in \mathbb{R}^{|I|} : \mathbf{x}_i \geq 0\,, i \in I\,, \sum_{i \in I} \mathbf{x}_i = 1\}$ of player 1 (resp. $\Delta(J)$ for player 2).

$$\min_{\mathbf{x} \in \Delta(I)} \max_{\mathbf{y} \in \Delta(J)} \mathbf{x}^T \mathbf{A}\, \mathbf{y}$$

- Example: "Matching Pennies", "Rock-Paper-Scissors".

Nonsmooth optimization

Let f, g be convex nonsmooth functions, $\mathbf{A} \in \mathbb{R}^{p\times d}$ a matrix and consider the problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) + g(\mathbf{A}\mathbf{x})\,.$$

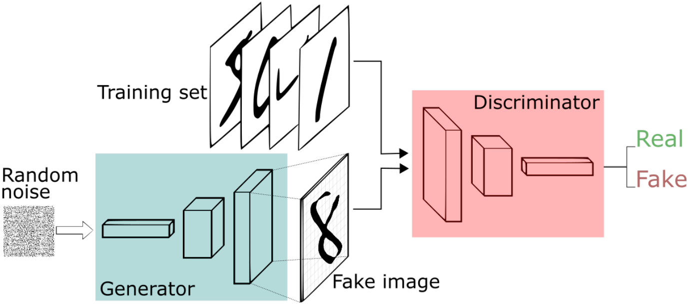
- Examples: $g(\mathbf{z}) = \|\mathbf{z}-\mathbf{b}\|_1$, $g(\mathbf{z}) = \|\mathbf{z}-\mathbf{b}\|_2^2$ or $g(\mathbf{z}) = \iota_{\{\mathbf{b}\}}(\mathbf{z})$ ($= 0$ if $\mathbf{z}=\mathbf{b}$, $+\infty$ otherwise) for which the Fenchel conjugate can be explicitly computed.
- Recall that $g(\mathbf{A}\mathbf{x}) = \max_{\mathbf{y}\in\mathbb{R}^p}\langle \mathbf{A}\mathbf{x},\mathbf{y}\rangle - g^*(\mathbf{y})$ where g^* is the Fenchel conjugate.

Min-Max reformulation:

$$\min_{\mathbf{x}\in\mathbb{R}^d}\max_{\mathbf{y}\in\mathbb{R}^p} f(\mathbf{x}) + \langle \mathbf{A}\mathbf{x},\mathbf{y}\rangle - g^*(\mathbf{y})$$

Generative Adversarial Networks (GANs)

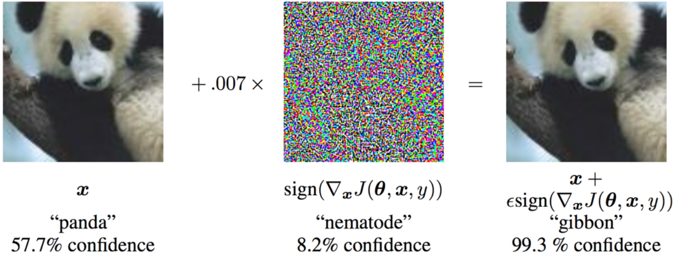
[Goodfellow et al., 2014]



$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \mathbb{E}_{\xi\sim p_{\mathsf{data}}}[\log D_{\mathbf{y}}(\xi)] + \mathbb{E}_{\zeta\sim p_\zeta}[\log(1 - D_{\mathbf{y}}(G_{\mathbf{x}}(\zeta)))]\,,$$

where $G_{\mathbf{x}}$ (resp. $D_{\mathbf{y}}$) is the generator (resp. discriminator) NN with parameters $\mathbf{x}$ (resp. $\mathbf{y}$).

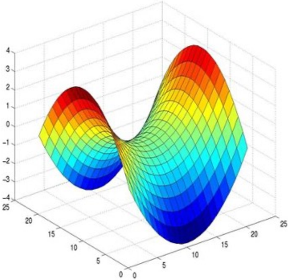
Distributionally Robust Optimization



$$\min_{\mathbf{x}} \max_{P \in \mathcal{P}} \mathbb{E}_{\boldsymbol{\xi} \sim P}[\ell(\mathbf{x}, \boldsymbol{\xi})]$$

where $\mathcal{P} := \{P : W_c(P, P_n) \leq \rho\}$ is some ambiguity set of the data distribution.

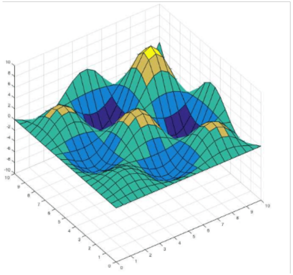
Three Critical Regimes



Convex-Concave

Nonconvex-Concave

Nonconvex-Nonconcave



Fundamental Questions

- What is the right notion of solution?
- When is the problem solvable?
- How to design gradient-based algorithms to solve it?
- How fast can the algorithm converge?
- What's the optimal complexity?
- ....

Lecture Outline

Why Min-Max Optimization?

Saddle Points and Global Minimax Points

The Classical Setting: Convex-Concave Min-Max Optimization

First-order Methods

- Gradient Descent Ascent (GDA)
- Extragradient Method
- Optimistic GDA

Extension to Concave Games, Variational Inequalities

Saddle Points and Global Minimax Points

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{x},\mathbf{y})$$

$(\mathbf{x}^*,\mathbf{y}^*)$ is a saddle point if

$$\phi(\mathbf{x}^*,\mathbf{y}) \leq \phi(\mathbf{x}^*,\mathbf{y}^*) \leq \phi(\mathbf{x},\mathbf{y}^*),$$

for any $\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}$.

- Game interpretation: Nash equilibrium
- No player has the incentive to make unilateral change at NE.
- Simultaneous game

$(\mathbf{x}^*,\mathbf{y}^*)$ is a global minimax point if

$$\phi(\mathbf{x}^*,\mathbf{y}) \leq \phi(\mathbf{x}^*,\mathbf{y}^*) \leq \max_{\mathbf{y}'\in\mathcal{Y}} \phi(\mathbf{x},\mathbf{y}'),$$

for any $\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}$.

- Game interpretation: Stackelberg equilibrium
- Best response to the best response.
- Sequential game

Primal and Dual Problems

Two induced problems:

$$(P): \quad \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{x},\mathbf{y}) := \min_{\mathbf{x}\in\mathcal{X}} \bar{\phi}(\mathbf{x})$$

$$(D): \quad \max_{\mathbf{y}\in\mathcal{Y}} \min_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x},\mathbf{y}) := \max_{\mathbf{y}\in\mathcal{Y}} \underline{\phi}(\mathbf{y})$$

Note that

$$\max_{\mathbf{y}\in\mathcal{Y}} \min_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x},\mathbf{y}) \leq \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{x},\mathbf{y})$$

Characterization of Saddle Points

Lemma 12.1

$(\mathbf{x}^*, \mathbf{y}^*)$ *is a saddle point if and only if*

$$\max_{\mathbf{y}\in\mathcal{Y}} \min_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x}, \mathbf{y}) = \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{x}, \mathbf{y})$$

and $\mathbf{x}^* \in \mathrm{argmin}_{\mathbf{x}\in\mathcal{X}} \bar{\phi}(\mathbf{x}), \quad \mathbf{y}^* \in \mathrm{argmax}_{\mathbf{y}\in\mathcal{Y}} \underline{\phi}(\mathbf{y}).$

Invoking the definition of saddle point, we have

$$\min_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x}, \mathbf{y}^*) \geq \phi(\mathbf{x}^*, \mathbf{y}^*) \geq \max_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{x}^*, \mathbf{y}).$$

Examples

Example 1: $\min_{\mathbf{x}\in\Delta(I)} \max_{\mathbf{y}\in\Delta(J)} \mathbf{x}^T \mathbf{A}\, \mathbf{y}$ (Rock-Paper-Scissors)

	rock	paper	scissor
rock	$0, 0$	$-1, 1$	$1, -1$
paper	$1, -1$	$0, 0$	$-1, 1$
scissor	$-1, 1$	$1, -1$	$0, 0$

- Only one saddle point: $\mathbf{x}^* = (1/3, 1/3, 1/3), \mathbf{y}^* = (1/3, 1/3, 1/3)$

Examples

Example 2: $\phi(x,y) = (x-y)^2, \mathcal{X} = [0,1], \mathcal{Y} = [0,1]$.

- Saddle point does not exist.
- $\bar{\phi}(x) = \max\{x^2, (x-1)^2\}$, $\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \phi(x,y) = \frac{1}{4}$.
- $\underline{\phi}(y) = 0$, $\max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} \phi(x,y) = 0$.

Lecture Outline

Convex-Concave Functions

Definition 12.2 (Convex-concave function)

A function $\phi(\mathbf{x}, \mathbf{y}) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is convex-concave if

- $\phi(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{x} \in \mathcal{X}$ for every fixed $\mathbf{y} \in \mathcal{Y}$;
- $\phi(\mathbf{x}, \mathbf{y})$ is concave in $\mathbf{y} \in \mathcal{Y}$ for every fixed $\mathbf{x} \in \mathcal{X}$.

Definition 12.3 (Strongly-convex-strongly-concave function)

A function $\phi(\mathbf{x}, \mathbf{y}) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is strongly-convex-strongly-concave if there exist constants $\mu_1, \mu_2 > 0$ such that

- $\phi(\mathbf{x}, \mathbf{y})$ is μ_1-strongly convex in $\mathbf{x} \in \mathcal{X}$ for every fixed $\mathbf{y} \in \mathcal{Y}$;
- $\phi(\mathbf{x}, \mathbf{y})$ is μ_2-strongly concave in $y \in \mathcal{Y}$ for every fixed $\mathbf{x} \in \mathcal{X}$.

Existence of Saddle Points

Theorem 12.4 (Minimax Theorem)

If $\mathcal{X}$ and $\mathcal{Y}$ are closed convex sets and one of them is bounded, and $\phi(\mathbf{x}, \mathbf{y})$ is a continuous convex-concave function, then there exists a saddle point on $\mathcal{X} \times \mathcal{Y}$ and

$$\max_{\mathbf{y} \in \mathcal{Y}} \min_{\mathbf{x} \in \mathcal{X}} \phi(\mathbf{x}, \mathbf{y}) = \min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} \phi(\mathbf{x}, \mathbf{y}).$$

- Minimax theorem was first proven and published in 1928 by John von Neumann.
- Can be extended to lower-semicontinous and quasi-convex functions.
- See general results in Chapter VI of Convex Analysis and Variational Problems.
- If $\phi(\mathbf{x}, \mathbf{y})$ is strongly-convex-strongly-concave, then we can remove the compactness assumption and the saddle point is unique.

Lecture Outline

Why Min-Max Optimization?

Saddle Points and Global Minimax Points

The Classical Setting: Convex-Concave Min-Max Optimization

First-order Methods

- Gradient Descent Ascent (GDA)
- Extragradient Method
- Optimistic GDA

Extension to Concave Games, Variational Inequalities

Accuracy Measure of Minimax Optimization: Duality Gap

For convex-concave minimax optimization, saddle points exist.

- We measure the optimality via the duality gap.

$$\text{duality gap} := \max_{\mathbf{y}\in\mathcal{Y}} \phi(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x}, \hat{\mathbf{y}}) \geq 0.$$

- When duality gap $= 0$, $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is a saddle point.
- When duality gap $\leq \epsilon$, $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is an ϵ-saddle point.

Gradient Descent Ascent (GDA)

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}(\mathbf{x}_t - \gamma \nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t))$$
$$\mathbf{y}_{t+1} = \Pi_{\mathcal{Y}}(\mathbf{y}_t + \gamma \nabla_{\mathbf{y}} \phi(\mathbf{x}_t, \mathbf{y}_t))$$

- Simplest gradient-based algorithm, widely used
- Simultaneous update of $\mathbf{x}$ and $\mathbf{y}$
- Q: Does it converge? If so, how fast?

Strongly-Convex-Strongly-Concave (SC-SC) Setting

- μ-strongly convex about $\mathbf{x}$ and strongly concave about $\mathbf{y}$:

$$\phi(\mathbf{x}_1, \mathbf{y}) \geq \phi(\mathbf{x}_2, \mathbf{y}) + \nabla_{\mathbf{x}} \phi(\mathbf{x}_2, \mathbf{y})^\top (\mathbf{x}_1 - \mathbf{x}_2)) + \frac{\mu}{2} \|\mathbf{x}_1 - \mathbf{x}_2\|^2,$$
$$-\phi(\mathbf{x}, \mathbf{y}_1) \geq -\phi(\mathbf{x}, \mathbf{y}_2) - \nabla_{\mathbf{y}} \phi(\mathbf{x}, \mathbf{y}_2)^\top (\mathbf{y}_1 - \mathbf{y}_2) + \frac{\mu}{2} \|\mathbf{y}_1 - \mathbf{y}_2\|^2$$

- L-Lipschitz smooth jointly in $\mathbf{x}$ and $\mathbf{y}$:

$$\|\nabla_{\mathbf{x}} \phi(\mathbf{x}_1, \mathbf{y}_1) - \nabla_{\mathbf{x}} \phi(\mathbf{x}_2, \mathbf{y}_2)\| \leq L\big(\|\mathbf{x}_1 - \mathbf{x}_2\| + \|\mathbf{y}_1 - \mathbf{y}_2\|\big)$$
$$\|\nabla_{\mathbf{y}} \phi(\mathbf{x}_1, \mathbf{y}_1) - \nabla_{\mathbf{y}} \phi(\mathbf{x}_2, \mathbf{y}_2)\| \leq L\big(\|\mathbf{x}_1 - \mathbf{x}_2\| + \|\mathbf{y}_1 - \mathbf{y}_2\|\big)$$

- There exists a unique saddle point $(\mathbf{x}^*, \mathbf{y}^*)$

GDA for SC-SC Setting

Theorem 12.5

In SC-SC setting, GDA with stepsize $\eta < \frac{\mu}{2L^2}$ converges linearly:

$$\|\mathbf{x}_{t+1}-\mathbf{x}^*\|^2+\|\mathbf{y}_{t+1}-\mathbf{y}^*\|^2 \leq (1+4\eta^2L^2-2\eta\mu)\left(\|\mathbf{x}_t-\mathbf{x}^*\|^2+\|\mathbf{y}_t-\mathbf{y}^*\|^2\right).$$

When $\eta = \frac{\mu}{4L^2}$,

$$\|\mathbf{x}_T-\mathbf{x}^*\|^2+\|\mathbf{y}_T-\mathbf{y}^*\|^2 \leq (1-4\mu^2/L^2)^T\left(\|\mathbf{x}_0-\mathbf{x}^*\|^2+\|\mathbf{y}_0-\mathbf{y}^*\|^2\right).$$

- It implies a complexity of $O\left(\kappa^2 \log \frac{1}{\epsilon}\right)$ with $\kappa = L/\mu$ being condition number.

Proof Sketch: GDA for SC-SC Setting

- By strong-convexity-strong-concavity,

$$(\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*))^\top (\mathbf{x}-\mathbf{x}^*) + (\nabla_{\mathbf{y}} f(\mathbf{x}^*,\mathbf{y}^*) - \nabla_{\mathbf{y}} f(\boldsymbol{x},\boldsymbol{y}))^\top (\mathbf{y}-\mathbf{y}^*) \\ \geq \mu\|\mathbf{x}-\mathbf{x}^*\|^2 + \mu\|\mathbf{y}-\mathbf{y}^*\|^2$$

  - This inequality instead of strong convexity (concavity) is enough for convergence.

- By Lipschitz smoothness,

$$\|\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*)\|^2 \leq 2L\|\mathbf{x}-\mathbf{x}^*\|^2 + 2L\|\mathbf{y}-\mathbf{y}^*\|^2,$$

$$\|\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{y}} f(\mathbf{x}^*,\mathbf{y}^*)\|^2 \leq 2L\|\mathbf{x}-\mathbf{x}^*\|^2 + 2L\|\mathbf{y}-\mathbf{y}^*\|^2$$

Proof Sketch: GDA for SC-SC Setting

- By non-expansiveness of the projection,

$$
\begin{aligned}
& \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 + \|\mathbf{y}_{t+1} - \mathbf{y}^*\|^2 \\
= \quad & \|\Pi_X(\mathbf{x}_t - \eta \nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t)) - \Pi_X(\mathbf{x}^* - \eta \nabla_{\mathbf{x}} \phi(\mathbf{x}^*, \mathbf{y}^*))\|^2 + \\
& \|\Pi_Y(\mathbf{y}_t + \eta \nabla_{\mathbf{y}} \phi(\mathbf{x}_t, \mathbf{y}_t)) - \Pi_Y(\mathbf{y}^* + \eta \nabla_{\mathbf{y}} \phi(\mathbf{x}^*, \mathbf{y}^*))\|^2 \\
\leq \quad & \|\mathbf{x}_t - \eta \nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t) - \mathbf{x}^* + \eta \nabla_{\mathbf{x}} \phi(\mathbf{x}^*, \mathbf{y}^*)\|^2 + \\
& \|\mathbf{y}_t + \eta \nabla_{\mathbf{y}} \phi(\mathbf{x}_t, \mathbf{y}_t) - \mathbf{y}^* - \eta \nabla_{\mathbf{y}} \phi(\mathbf{x}^*, \mathbf{y}^*) \\
\leq \quad & \|\mathbf{x}_t - \mathbf{x}^*\|^2 + \eta^2 \|\nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} \phi(\mathbf{x}^*, \mathbf{y}^*)\|^2 - \\
& 2\eta(\nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} \phi(\mathbf{x}^*, \mathbf{y}^*))^\top (\mathbf{x}_t - \mathbf{x}^*) + \\
& \|\mathbf{y}_t - \mathbf{y}^*\|^2 + \eta^2 \|\nabla_{\mathbf{y}} \phi(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} \phi(\mathbf{x}^*, \mathbf{y}^*)\|^2 - \\
& 2\eta(\nabla_{\mathbf{y}} \phi(\mathbf{x}^*, \mathbf{y}^*) - \nabla_{\mathbf{y}} \phi(\mathbf{x}_t, \mathbf{y}_t))^\top (\mathbf{y}_t - \mathbf{y}^*).
\end{aligned}
$$

Plugging in the previous two inequalities leads to the desired result.

GDA for Convex-Concave (C-C) Setting

- Consider $\phi(x, y) = xy$. GDA update:

$$x_{t+1}^2 + y_{t+1}^2 = (x_t - \eta y_t)^2 + (y_t + \eta x_t)^2 = (1 + \eta^2)(x_t^2 + y_t^2)$$

  It does not converge to the saddle point $(0, 0)$.

- Even with different stepsizes for $\mathbf{x}$ and $\mathbf{y}$, GDA may still not converge for the bilinear games [Gidel et al., 2019].

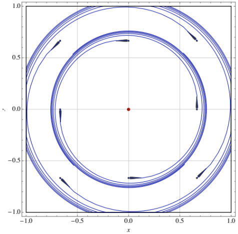


Figure: GDA for $\phi(x, y) = xy$ [Hsieh et al., 2021].

Extragradient

Extragradient Method

$$
\begin{aligned}
\mathbf{x}_{t+\frac{1}{2}} &= \Pi_{\mathcal{X}}\left(\mathbf{x}_t - \eta \nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t)\right), \\
\mathbf{y}_{t+\frac{1}{2}} &= \Pi_{\mathcal{Y}}\left(\mathbf{y}_t + \eta \nabla_{\mathbf{y}} \phi(\mathbf{x}_t, \mathbf{y}_t)\right) \\
\mathbf{x}_{t+1} &= \Pi_{\mathcal{X}}\left(\mathbf{x}_t - \eta \nabla_{\mathbf{x}} \phi\left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}\right)\right), \\
\mathbf{y}_{t+1} &= \Pi_{\mathcal{Y}}\left(\mathbf{y}_t + \eta \nabla_{\mathbf{y}} \phi\left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}\right)\right)
\end{aligned}
$$

- It is different from two steps of GDA!

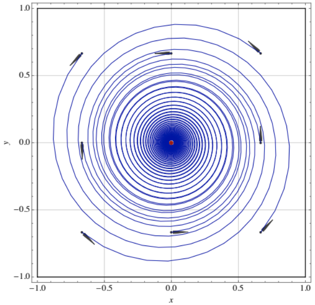


Figure: EG for $\phi(x, y) = xy$

EG for C-C Setting

Theorem 12.6

Assume ϕ is convex-concave, L-Lipschitz smooth, $\mathcal{X}$ has diameter $D_{\mathcal{X}}$, and $\mathcal{Y}$ has diameter $D_{\mathcal{Y}}$, then EG with stepsize $\eta \leq \frac{1}{2L}$ satisfies

$$\max_{\mathbf{y}\in\mathcal{Y}} \phi\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}\right) - \min_{\mathbf{x}\in\mathcal{X}} \phi\left(\mathbf{x}, \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_{t+\frac{1}{2}}\right) \leq \frac{D_{\mathcal{X}}^2 + D_{\mathcal{Y}}^2}{2\eta T}.$$

- $O(1/T)$ convergence rate for averaged iterates at "mid-point".
- $O(1/T)$ rate is optimal [Ouyang and Xu, 2021]
- Can use more general Bregman divergence – Mirror Prox [Nemirovski, 2004]

More about EG

- In C-C setting, EG has best-iterate and last-iterate convergence rate of $O(\frac{1}{\sqrt{T}})$ for primal-dual gap [Yang et al., 2022], which is slower than the averaged-iterate convergence. This is the best that can be achieved by EG [Golowich et al., 2020].

Theorem 12.7 (Mokhtari et al., 2020)

In SC-SC setting, EG with stepsize $\eta = \frac{1}{8L}$ converges linearly:

$$\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 + \|\mathbf{y}_{t+1} - \mathbf{y}^*\|^2 \leq \left(1 - \frac{\mu}{4L}\right) \{\|\mathbf{x}_t - \mathbf{x}^*\|^2 + \|\mathbf{y}_t - \mathbf{y}^*\|^2\}.$$

- This $O(\kappa \log \frac{1}{\epsilon})$ complexity is optimal for SC-SC setting [Zhang et al., 2021].

Optimistic GDA

- Optimistic GDA (OGDA, or Past EG [Popov, 1980]):

$$\mathbf{x}_{t+\frac{1}{2}} = \Pi_{\mathcal{X}}\left(\mathbf{x}_t - \eta\nabla_{\mathbf{x}}\phi(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}})\right), \qquad \mathbf{y}_{t+\frac{1}{2}} = \Pi_{\mathcal{Y}}\left(\mathbf{y}_t + \eta\nabla_{\mathbf{y}}\phi(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}})\right)$$

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}\left(\mathbf{x}_t - \eta\nabla_{\mathbf{x}}\phi\left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}\right)\right), \quad \mathbf{y}_{t+1} = \Pi_{\mathcal{Y}}\left(\mathbf{y}_t + \eta\nabla_{\mathbf{y}}\phi\left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}\right)\right)$$

  It is different from two steps of GDA!

- Equivalent formulation (self check):

$$\begin{cases} \mathbf{x}_{t+1} = \mathbf{x}_t - 2\eta\nabla_{\mathbf{x}}\phi(\mathbf{x}_t, \mathbf{y}_t) + \eta\nabla_{\mathbf{x}}\phi(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \\ \mathbf{y}_{t+1} = \mathbf{y}_t - 2\eta\nabla_{\mathbf{y}}\phi(\mathbf{x}_t, \mathbf{y}_t) + \eta\nabla_{\mathbf{y}}\phi(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \end{cases}$$

- Query the gradient only once each iteration (vs. EG)
- Similar convergence guarantees as EG for SC-SC and C-C settings.

Proximal Point Algorithm (PPA)

- Proximal Point Algorithm (PPA):

$$(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) \leftarrow \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}} \operatorname*{argmax}_{\mathbf{y}\in\mathcal{Y}} \left\{ \phi(\mathbf{x}, \mathbf{y}) + \frac{1}{2\eta}\|\mathbf{x} - \mathbf{x}_t\|^2 - \frac{1}{2\eta}\|\mathbf{y} - \mathbf{y}_t\|^2 \right\}$$

- In the "hardest" problem $\phi(x, y) = x^\top y$, GDA may not converge, while PPA is provably convergent (self check)
- PPA has been shown to converge with $\mathcal{O}(1/T)$ rate in convex-concave case (Nemirovski, 2004, Mokhtari et al., 2019)

Connections between PPA, EG and OGDA

- **Implicit update of PPA:**

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}\left(\mathbf{x}_t - \eta \nabla_{\mathbf{x}} \phi\left(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}\right)\right)$$
$$\mathbf{y}_{t+1} = \Pi_{\mathcal{Y}}\left(\mathbf{y}_t + \eta \nabla_{\mathbf{y}} \phi\left(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}\right)\right)$$

- How to compute $\nabla_{\mathbf{x}} \phi\left(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}\right)$ and $\nabla_{\mathbf{y}} \phi\left(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}\right)$?
- EG and OGDA can be viewed as approximate PPA with error

$$\mathbf{z}_{t+1} = \Pi_{\mathcal{Z}}\left(\mathbf{z}_t - \eta F\left(\mathbf{z}_{t+1}\right) + \epsilon_t\right)$$

where $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, $F(z) = \left(\nabla_{\mathbf{x}} \phi\left(\mathbf{x}, \mathbf{y}\right), -\nabla_{\mathbf{y}} \phi\left(\mathbf{x}, \mathbf{y}\right)\right)$

- EG: $\epsilon_t = \eta \left[F(\mathbf{z}_{t+1}) - F(\mathbf{z}_{t+\frac{1}{2}})\right]$
- OGDA: $\epsilon_t = \eta \left[(F(\mathbf{z}_{t+1}) - F(\mathbf{z}_t)) - (F(\mathbf{z}_t) - F(\mathbf{z}_{t-1}))\right]$

Lecture Outline

Extensions

Beyond the Min-Max optimization problem:

- Concave games and convex Nash equilibrium problems.
- Variational Inequalities with monotone operators.
- ....

Concave Games

- Finite number of players $i \in \mathcal{N} = \{1, \cdots, N\}$.
- Action profile $\mathbf{x} = (\mathbf{x}_i, \mathbf{x}_{-i}) = (\mathbf{x}_1, \cdots, \mathbf{x}_N) \in \mathcal{X} = \prod_i \mathcal{X}_i$ where:
  - $\mathcal{X}_i$ is a compact convex subset $\mathbb{R}^{d_i}$.
  - $\mathbf{x}_i \in \mathcal{X}_i$ is the action of player i.
- Payoff (or utility) function $u_i : \mathcal{X} \to \mathbb{R}$.
- **Assumption:** $u_i(\mathbf{x}_i, \mathbf{x}_{-i})$ is concave in $\mathbf{x}_i$ for all $\mathbf{x}_{-i} \in \mathcal{X}_{-i} = \prod_{j \neq i} \mathcal{X}_j, \, , i \in \mathcal{N}$.

Nash Equilibrium for Concave Games

- Nash equilibrium: Any action profile $x^* \in \mathcal{X}$ resilient to unilateral deviations, i.e.,

$$u_i(\mathbf{x}_i^*, \mathbf{x}_{-i}^*) \geq u_i(\mathbf{x}_i, \mathbf{x}_{-i}^*) \quad \forall\, \mathbf{x}_i \in \mathcal{X}_i\,,\, i \in \mathcal{N}\,.$$

- Existence theorem [Debreu, 1952]: every concave game admits a Nash equilibrium.

- First-order characterization: under the concavity assumption on the game's payoff functions, Nash equilibria can also be characterized via first-order optimality:

$$\langle \nabla_i\, u_i(\mathbf{x}_i^*, \mathbf{x}_{-i}^*), \mathbf{x}_i - \mathbf{x}_i^* \rangle \leq 0 \quad \forall \mathbf{x}_i \in \mathcal{X}_i\,,$$

  where ∇_i refers to differentiation with respect to $\mathbf{x}_i$.

Variational Inequalities

Let $\mathcal{Z} \subset \mathbb{R}^d$ be a nonempty subset and consider a mapping $F : \mathcal{Z} \to \mathbb{R}^d$.

Variational Inequality Problem (VI)

Find $\mathbf{z}^* \in \mathcal{Z}$ such that $\langle F(\mathbf{z}^*), \mathbf{z} - \mathbf{z}^* \rangle \geq 0$ for all $\mathbf{z} \in \mathcal{Z}$.

- Existence [Stampacchia, 1966]: If $\mathcal{Z}$ is a nonempty convex compact subset of $\mathbb{R}^d$ and $F : \mathcal{Z} \to \mathbb{R}^d$ is continuous, then there exists a solution z^* to (VI).
- **NB:** Compactness can be replaced by a "coercivity condition" and the result can be generalized to a set valued mapping F.

Variational Inequalities with Monotone Operators

The operator $F : \mathcal{Z} \to \mathbb{R}^d$ is:

- monotone if

$$\langle F(\mathbf{u}) - F(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \geq 0 \quad \forall\, \mathbf{u}, \mathbf{v} \in \mathcal{Z}\,.$$

- μ-strongly-monotone $(\mu > 0)$ if

$$\langle F(\mathbf{u}) - F(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \geq \mu \left\| \mathbf{u} - \mathbf{v} \right\|^2 \quad \forall\, \mathbf{u}, \mathbf{v} \in \mathcal{Z}\,.$$

Weak solution of VI and error metric

- (Strong) solution (of Stampacchia VI): find $\mathbf{z}^* \in \mathcal{Z}$ such that:

$$\langle F(\mathbf{z}^*), \mathbf{z} - \mathbf{z}^* \rangle \geq 0 \ \forall \mathbf{z} \in \mathcal{Z} \,.$$

- Weak solution (of Minty VI): find $\mathbf{z}^* \in \mathcal{Z}$ such that:

$$\langle F(\mathbf{z}), \mathbf{z} - \mathbf{z}^* \rangle \geq 0 \ \forall \mathbf{z} \in \mathcal{Z} \,.$$

- If F is monotone, then a strong solution is also a weak solution (Exercise).
- If F is continuous, then a weak solution is also a strong solution (Exercise).

We use $\epsilon_{\mathrm{VI}}(\hat{\mathbf{z}}) := \max_{\mathbf{z} \in \mathcal{Z}} \langle F(\mathbf{u}), \hat{\mathbf{z}} - \mathbf{z} \rangle$ to measure the inaccuracy of a solution $\hat{\mathbf{z}}$.

Examples of Variational Inequality problems

- Convex minimization: $F = \nabla f$ for some convex function f .

  The (VI) solutions are the minimizers of the function f.

- Min-Max problems: $F = (\nabla_{\mathbf{x}}\phi, -\nabla_{\mathbf{y}}\phi)$ for some convex-concave $\phi(\mathbf{x}, \mathbf{y})$.

  The (VI) solutions are the global saddle points of ϕ, i.e., $\mathbf{z}^* = (\mathbf{x}^*, \mathbf{y}^*)$ is a solution of (VI) if and only if:

$$\phi(\mathbf{x}^*, \mathbf{y}) \leq \phi(\mathbf{x}^*, \mathbf{y}^*) \leq \phi(\mathbf{x}, \mathbf{y}^*)\,, \quad \forall\, \mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}\,.$$

- Concave (or monotone) games: $F(\mathbf{z}) = (-\nabla_i u_i(\mathbf{z}_i, \mathbf{z}_{-i}))_{i \in \mathcal{N}}$.

  The (VI) solutions coincide with Nash equilibria.

Assumptions

We make the following assumptions under which we will present general (convergent) algorithms for solving monotone VIs:

- The set $\mathcal{Z}$ is a closed convex subset of $\mathbb{R}^d$.
- The solution set of (VI) is nonempty.
- The mapping F is monotone.
- The mapping F is Lipschitz continuous with constant $L > 0$, i.e.,

$$\|F(\mathbf{u}) - F(\mathbf{v})\| \leq L \|\mathbf{u} - \mathbf{v}\| \ , \quad \forall \mathbf{u}, \mathbf{v} \in \mathcal{Z} \, .$$

Extragradient Algorithm for VIs

[Korpelevich, 1976]

Extragradient:

$$\tilde{\mathbf{z}}_{t+1} = \Pi_{\mathcal{Z}}(\mathbf{z}_t - \eta_t F(\mathbf{z}_t))$$
$$\mathbf{z}_{t+1} = \Pi_{\mathcal{Z}}(\mathbf{z}_t - \eta_t F(\tilde{\mathbf{z}}_{t+1}))$$

- Unconstrained setting ($\mathcal{Z} = \mathbb{R}^d$): $\mathbf{z}_{t+1} = \mathbf{z}_t - \eta_t F(\mathbf{z}_t - \eta_t F(\mathbf{z}_t))$.

Convergence Rates for EG for VIs

Theorem 12.8 (Nemirovski, 2004)

Let the previously stated assumptions (mainly F is a monotone and L-Lipschitz operator) hold. Set $\eta_t = \eta = \frac{1}{\sqrt{2}L}$. Then the sequence $(\tilde{z}_t)$ generated by EG with step size η satisfies:

$$\max_{\mathbf{z}\in\mathcal{Z}} \left\langle F(\mathbf{z}), \frac{1}{T}\sum_{t=1}^{T} \tilde{\mathbf{z}}_t - \mathbf{z} \right\rangle \leq \frac{\sqrt{2}LD_{\mathcal{Z}}^2}{T},$$

where $D_{\mathcal{Z}} = \max_{\mathbf{z},\mathbf{z}'} \|\mathbf{z}-\mathbf{z}'\|_2$ is the $\|\cdot\|_2$-diameter of $\mathcal{Z}$.

- We recover the previous EG theorem for the convex-concave min-max problem.
- Remark: EG is a particular case (in the Euclidean setting) of the Prox-Method for VIs with monotone Lipschitz operators in [Nemirovski, 2004] (see for proofs).

Other Algorithms for VIs

Consider unconstrained setting ($\mathcal{Z} = \mathbb{R}^d$) for simplicity.

- GDA: $\mathbf{z}_{t+1} = \mathbf{z}_t - \eta_t F(\mathbf{z}_t)$.
- PPA: $\mathbf{z}_{t+1} = \mathbf{z}_t - \eta_t F(\mathbf{z}_{t+1})$.
- OGDA: $\mathbf{z}_{t+1} = \mathbf{z}_t - \eta_t(2F(\mathbf{z}_t) - F(\mathbf{z}_{t-1}))$.
- Reflected Gradient [Malitsky, 2015]: $\mathbf{z}_{t+1} = \mathbf{z}_t - \eta_t F(2\mathbf{z}_t - \mathbf{z}_{t-1})$.
- Halpern iteration [Halpern, 1967]: $\mathbf{z}_{t+1} = \lambda_k \mathbf{z}_0 + (1 - \lambda_k)(\mathbf{z}_t - \eta_t F(\mathbf{z}_t))$.
- ...

Bibliography

G. Debreu.
A social equilibrium existence theorem.
Proceedings of the National Academy of Sciences, 38(10):886-893, 1952.

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

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

Y. Malitsky.
Projected reflected gradient methods for monotone variational inequalities.
SIAM Journal on Optimization, 25(1):502-520, 2015.

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.

Bibliography

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.

Gidel, G., Hemmat, R.A., Pezeshki, M., Le Priol, R., Huang, G., Lacoste-Julien, S. and Mitliagkas, I.
Negative momentum for improved game dynamics.
AISTATS 2019.

Gidel, G., Berard, H., Vignoud, G., Vincent, P., Lacoste-Julien, S.
A variational inequality perspective on generative adversarial networks.
ICLR 2019.

Junyu Zhang and Mingyi Hong and Shuzhong Zhang.
On lower iteration complexity bounds for the convex concave saddle point problems .
Mathematical Programming, 2021.

Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis.
Tight last-iterate convergence rates for no-regret learning in multi-player games .
NeurIPS 2020.

Bibliography

Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman E. Ozdaglar.
Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems.
COLT 2020.

Cai, Yang, Argyris Oikonomou, and Weiqiang Zheng.
Tight Last-Iterate Convergence of the Extragradient Method for Constrained Monotone Variational Inequalities.
arXiv:2204.09228 (2022).

Ouyang, Yuyuan, and Yangyang Xu.
Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems.
Mathematical Programming, 2021.

F. Facchinei and J-S. Pang.
Finite-dimensional variational inequalities and complementarity problems.
Springer, 2003.

Appendix: More Details on GDA for SC-SC Setting

To show the first inequality from strong-convexity-strongly-concavity,

$$
\begin{aligned}
\phi(\boldsymbol{x}_2, \boldsymbol{y}_1) &\geq \phi(\boldsymbol{x}_1, \boldsymbol{y}_1) + \nabla_{\mathbf{x}} \phi(\boldsymbol{x}_1, \boldsymbol{y}_1)^\top (\boldsymbol{x}_2 - \boldsymbol{x}_1) + \frac{\mu}{2} \left\| \boldsymbol{x}_2 - \boldsymbol{x}_1 \right\|^2, \\
\phi(\boldsymbol{x}_1, \boldsymbol{y}_2) &\geq \phi(\boldsymbol{x}_2, \boldsymbol{y}_2) + \nabla_{\mathbf{x}} \phi(\boldsymbol{x}_2, \boldsymbol{y}_2)^\top (\boldsymbol{x}_1 - \boldsymbol{x}_2) + \frac{\mu}{2} \left\| \boldsymbol{x}_1 - \boldsymbol{x}_2 \right\|^2, \\
-\phi(\boldsymbol{x}_1, \boldsymbol{y}_2) &\geq -\phi(\boldsymbol{x}_1, \boldsymbol{y}_1) - \nabla_{\mathbf{y}} \phi(\boldsymbol{x}_1, \boldsymbol{y}_1)^\top (\boldsymbol{y}_2 - \boldsymbol{y}_1) + \frac{\mu}{2} \left\| \boldsymbol{y}_2 - \boldsymbol{y}_1 \right\|^2, \\
-\phi(\boldsymbol{x}_2, \boldsymbol{y}_1) &\geq -\phi(\boldsymbol{x}_2, \boldsymbol{y}_2) - \nabla_{\mathbf{y}} \phi(\boldsymbol{x}_2, \boldsymbol{y}_2)^\top (\boldsymbol{y}_1 - \boldsymbol{y}_2) + \frac{\mu}{2} \left\| \boldsymbol{y}_1 - \boldsymbol{y}_2 \right\|^2.
\end{aligned}
$$

Summing four equations together, we get:

$$
\begin{aligned}
(\nabla_x f(\boldsymbol{x}_1, \boldsymbol{y}_1) - \nabla_x f(\boldsymbol{x}_2, \boldsymbol{y}_2))^\top (\boldsymbol{x}_1 - \boldsymbol{x}_2) + (\nabla_y f(\boldsymbol{x}_2, \boldsymbol{y}_2) - \nabla_y f(\boldsymbol{x}_1, \boldsymbol{y}_1))^\top (\boldsymbol{y}_1 - \boldsymbol{y}_2) \\
\geq \mu \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2 + \mu \|\boldsymbol{y}_1 - \boldsymbol{y}_2\|^2.
\end{aligned}
$$

Proof of EG for C-C Setting

For convenience, write the update as the following:

$$
\begin{aligned}
\widetilde{\mathbf{x}}_{t+\frac{1}{2}} &= \mathbf{x}_t - \eta \nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t), \qquad \widetilde{\mathbf{y}}_{t+\frac{1}{2}} = \mathbf{y}_t + \eta \nabla_{\mathbf{y}} \phi(\mathbf{x}_t, \mathbf{y}_t) \\
\mathbf{x}_{t+\frac{1}{2}} &= \Pi_{\mathcal{X}}\left(\widetilde{\mathbf{x}}_{t+\frac{1}{2}}\right), \qquad \mathbf{y}_{t+\frac{1}{2}} = \Pi_{\mathcal{Y}}\left(\widetilde{\mathbf{y}}_{t+\frac{1}{2}}\right) \\
\widetilde{\mathbf{x}}_{t+1} &= \mathbf{x}_t - \eta \nabla_{\mathbf{x}} \phi\left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}\right), \qquad \widetilde{\mathbf{y}}_{t+1} = \mathbf{y}_t + \eta \nabla_{\mathbf{y}} \phi\left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}\right) \\
\mathbf{x}_{t+1} &= \Pi_{\mathcal{X}}(\widetilde{\mathbf{x}}_{t+1}), \qquad \mathbf{y}_{t+1} = \Pi_{\mathcal{Y}}(\widetilde{\mathbf{y}}_{t+1})
\end{aligned}
$$

First, we note that

$$
\begin{aligned}
\nabla_{\mathbf{x}} \phi\left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}\right)^\top (\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}) = \nabla_{\mathbf{x}} \phi\left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}\right)^\top (\mathbf{x}_{t+1} - \mathbf{x}) + \\
\nabla_{\mathbf{x}} \phi\left(\mathbf{x}_t, \mathbf{y}_t\right)^\top (\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1}) + \\
\left(\nabla_{\mathbf{x}} \phi\left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}\right) - \nabla_{\mathbf{x}} \phi\left(\mathbf{x}_t, \mathbf{y}_t\right)\right)^\top (\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1}).
\end{aligned}
$$

Proof of EG for C-C Setting II

We will bound each term of the right hand side.

$$
\begin{aligned}
\nabla_{\mathbf{x}} \phi\left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}\right)^{\top}\left(\mathbf{x}_{t+1}-\mathbf{x}\right) &= \frac{1}{\eta}(\mathbf{x}_t - \widetilde{\mathbf{x}}_{t+1})^{\top}(\mathbf{x}_{t+1}-\mathbf{x}) \\
&\leq \frac{1}{\eta}(\mathbf{x}_t - \mathbf{x}_{t+1})^{\top}(\mathbf{x}_{t+1}-\mathbf{x}) \\
&= \frac{1}{2\eta}\left[\|\mathbf{x}-\mathbf{x}_t\|^2 - \|\mathbf{x}-\mathbf{x}_{t+1}\|^2 - \|\mathbf{x}_t-\mathbf{x}_{t+1}\|^2\right]
\end{aligned}
$$

Where the second inequality uses the property of projection.

$$
\begin{aligned}
\nabla_{\mathbf{x}} \phi\left(\mathbf{x}_t, \mathbf{y}_t\right)^{\top}\left(\mathbf{x}_{t+\frac{1}{2}}-\mathbf{x}_{t+1}\right) &= \frac{1}{\eta}(\mathbf{x}_t - \widetilde{\mathbf{x}}_{t+\frac{1}{2}})^{\top}(\mathbf{x}_{t+\frac{1}{2}}-\mathbf{x}_{t+1}) \\
&\leq \frac{1}{\eta}(\mathbf{x}_t - \mathbf{x}_{t+\frac{1}{2}})^{\top}(\mathbf{x}_{t+\frac{1}{2}}-\mathbf{x}_{t+1}) \\
&= \frac{1}{2\eta}\left[\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 - \|\mathbf{x}_{t+\frac{1}{2}}-\mathbf{x}_{t+1}\|^2 - \|\mathbf{x}_t-\mathbf{x}_{t+\frac{1}{2}}\|^2\right]
\end{aligned}
$$

Proof of EG for C-C Setting III

$$
\begin{aligned}
&\left(\nabla_{\mathbf{x}}\phi\left(\mathbf{x}_{t+\frac{1}{2}},\mathbf{y}_{t+\frac{1}{2}}\right)-\nabla_{\mathbf{x}}\phi\left(\mathbf{x}_t,\mathbf{y}_t\right)\right)^{\top}\left(\mathbf{x}_{t+\frac{1}{2}}-\mathbf{x}_{t+1}\right)\\
&\leq\left\|\nabla_{\mathbf{x}}\phi\left(\mathbf{x}_{t+\frac{1}{2}},\mathbf{y}_{t+\frac{1}{2}}\right)-\nabla_{\mathbf{x}}\phi\left(\mathbf{x}_t,\mathbf{y}_t\right)\right\|\left\|\mathbf{x}_{t+\frac{1}{2}}-\mathbf{x}_{t+1}\right\|\\
&\leq L\left[\left\|\mathbf{x}_{t+\frac{1}{2}}-\mathbf{x}_t\right\|+\left\|\mathbf{y}_{t+\frac{1}{2}}-\mathbf{y}_t\right\|\right]\left\|\mathbf{x}_{t+\frac{1}{2}}-\mathbf{x}_{t+1}\right\|\\
&\leq\frac{l}{2}\left\|\mathbf{x}_{t+\frac{1}{2}}-\mathbf{x}_t\right\|^2+\frac{L}{2}\left\|\mathbf{y}_{t+\frac{1}{2}}-\mathbf{y}_t\right\|^2+l\left\|\mathbf{x}_{t+\frac{1}{2}}-\mathbf{x}_{t+1}\right\|^2.
\end{aligned}
$$

Combine three inequalites above,

$$
\begin{aligned}
&\nabla_{\mathbf{x}}\phi\left(\mathbf{x}_{t+\frac{1}{2}},\mathbf{y}_{t+\frac{1}{2}}\right)^{\top}\left(\mathbf{x}_{t+\frac{1}{2}}-\mathbf{x}\right)\\
\leq&\frac{2}{\eta}\left[\|\mathbf{x}_t-\mathbf{x}\|^2-\|\mathbf{x}-\mathbf{x}_{t+1}\|^2\right]+\left(L-\frac{1}{2\eta}\right)\left\|\mathbf{x}_{t+\frac{L}{2}}-\mathbf{x}_{t+1}\right\|^2+\\
&\left(\frac{L}{2}-\frac{1}{2\eta}\right)\left\|\mathbf{x}_{t+\frac{1}{2}}-\mathbf{x}_t\right\|^2+\frac{L}{2}\left\|\mathbf{y}_{t+\frac{1}{2}}-\mathbf{y}_t\right\|^2
\end{aligned}
$$

Proof of EG for C-C Setting IV

Similarly, we can show

$$
\begin{aligned}
&-\nabla_{\mathbf{y}}\phi\left(\mathbf{x}_{t+\frac{1}{2}},\mathbf{y}_{t+\frac{1}{2}}\right)^{\top}\left(\mathbf{y}_{t+\frac{1}{2}}-\mathbf{y}\right)\\
\leq\ &\frac{2}{\eta}\left[\|\mathbf{y}_t-\mathbf{y}\|^2-\|\mathbf{y}-\mathbf{y}_{t+1}\|^2\right]+\left(L-\frac{1}{2\eta}\right)\left\|\mathbf{y}_{t+\frac{L}{2}}-\mathbf{y}_{t+1}\right\|^2+\\
&\left(\frac{L}{2}-\frac{1}{2\eta}\right)\left\|\mathbf{y}_{t+\frac{1}{2}}-\mathbf{y}_t\right\|^2+\frac{L}{2}\left\|\mathbf{x}_{t+\frac{1}{2}}-\mathbf{x}_t\right\|^2
\end{aligned}
$$

Lastly, note that

$$
\begin{aligned}
&\phi\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t,\mathbf{y}\right)-\phi\left(\mathbf{x},\frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t\right)\leq\frac{1}{T}\sum_{t=1}^{T}\phi\left(\mathbf{x}_{t+\frac{1}{2}},\mathbf{y}\right)-\phi\left(\mathbf{x},\mathbf{y}_{t+\frac{1}{2}}\right)\\
\leq\ &\frac{1}{T}\sum_{t=1}^{T}-\nabla_{\mathbf{y}}\phi\left(\mathbf{x}_{t+\frac{1}{2}},\mathbf{y}_{t+\frac{1}{2}}\right)^{\top}\left(\mathbf{y}_{t+\frac{1}{2}}-\mathbf{y}\right)+\nabla_{\mathbf{x}}\phi\left(\mathbf{x}_{t+\frac{1}{2}},\mathbf{y}_{t+\frac{1}{2}}\right)^{\top}\left(\mathbf{x}_{t+\frac{1}{2}}-\mathbf{x}\right).
\end{aligned}
$$