

Chapter 13 minimax

- Formulation $\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} \phi(\mathbf{x}, \mathbf{y})$
- Motivation
 - Zero sum games $\min_{\mathbf{x} \in \Delta(I)} \max_{\mathbf{y} \in \Delta(J)} \mathbf{x}^T \mathbf{A} \mathbf{y}$, I, J finite set of strategies of player 1 and 2, \mathbf{A} is payoff for player 1, $-\mathbf{A}$ for player 2, $\Delta(I) = \{\mathbf{x} \in \mathbb{R}^{|I|} : \mathbf{x}_i \geq 0, i \in I, \sum_{i \in I} \mathbf{x}_i = 1\}$.
 - Nonsmooth optimization Original problem $\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) + g(\mathbf{A}\mathbf{x})$, plug in $g(\mathbf{A}\mathbf{x}) = \max_{\mathbf{y} \in \mathbb{R}^p} \langle \mathbf{A}\mathbf{x}, \mathbf{y} \rangle - g^*(\mathbf{y})$, we reformulate as $\min_{\mathbf{x} \in \mathbb{R}^d} \max_{\mathbf{y} \in \mathbb{R}^p} f(\mathbf{x}) + \langle \mathbf{A}\mathbf{x}, \mathbf{y} \rangle - g^*(\mathbf{y})$.
 - GANs $\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} \mathbb{E}_{\xi \sim p_{\text{data}}} [\log D_{\mathbf{y}}(\xi)] + \mathbb{E}_{\zeta \sim p_{\zeta}} [\log (1 - D_{\mathbf{y}}(G_{\mathbf{x}}(\zeta)))]$.
 - Adversarial Robustness $\min_{\mathbf{x}} \max_{P \in \mathcal{P}} \mathbb{E}_{\xi \sim P} [\ell(\mathbf{x}, \xi)]$

Saddle Points and Global Minimax Points

- Definition of saddle point $(\mathbf{x}^*, \mathbf{y}^*)$ is a *saddle point* if $\forall \mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}, \phi(\mathbf{x}^*, \mathbf{y}) \leq \phi(\mathbf{x}^*, \mathbf{y}^*) \leq \phi(\mathbf{x}, \mathbf{y}^*)$.
 - corresponds to *Nash equilibrium*, simultaneous game, no player has the incentive to make unilateral change at NE.
- Definition of global minimax point $(\mathbf{x}^*, \mathbf{y}^*)$ is a *global minimax point* if
 $\forall \mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}, \phi(\mathbf{x}^*, \mathbf{y}) \leq \phi(\mathbf{x}^*, \mathbf{y}^*) \leq \max_{\mathbf{y}' \in \mathcal{Y}} \phi(\mathbf{x}, \mathbf{y}')$.
 - Corresponds to *Stackelberg equilibrium*, sequential game, best responds to the best response.

Primal and Dual Problems

- Primal $\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} \phi(\mathbf{x}, \mathbf{y}) := \min_{\mathbf{x} \in \mathcal{X}} \bar{\phi}(\mathbf{x})$
- Dual $\max_{\mathbf{y} \in \mathcal{Y}} \min_{\mathbf{x} \in \mathcal{X}} \phi(\mathbf{x}, \mathbf{y}) := \max_{\mathbf{y} \in \mathcal{Y}} \underline{\phi}(\mathbf{y})$
- Lemma A $\max_{\mathbf{y} \in \mathcal{Y}} \min_{\mathbf{x} \in \mathcal{X}} \phi(\mathbf{x}, \mathbf{y}) \leq \min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} \phi(\mathbf{x}, \mathbf{y})$
 - Proof
 - $\forall x, y, \min_x \phi(x, y) \leq \phi(x, y)$ taking maximum $\forall x, \max_y \min_x \phi(x, y) \leq \max_y \phi(x, y)$, so is its minimum $\min_x \max_y \phi(x, y)$
- Lemma B $(\mathbf{x}^*, \mathbf{y}^*)$ 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 $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}} \bar{\phi}(\mathbf{x})$, $\mathbf{y}^* \in \operatorname{argmax}_{\mathbf{y} \in \mathcal{Y}} \underline{\phi}(\mathbf{y})$.
 - Proof
 - (\rightarrow)
 - Saddle point $\Leftrightarrow \min_{\mathbf{x} \in \mathcal{X}} \phi(\mathbf{x}, \mathbf{y}^*) \geq \phi(\mathbf{x}^*, \mathbf{y}^*) \geq \max_{\mathbf{y} \in \mathcal{Y}} \phi(\mathbf{x}^*, \mathbf{y})$
 - Then by definition $\max_{\mathbf{y} \in \mathcal{Y}} \min_{\mathbf{x} \in \mathcal{X}} \phi(\mathbf{x}, \mathbf{y}) \geq \min_{\mathbf{x} \in \mathcal{X}} \phi(\mathbf{x}, \mathbf{y}^*) \geq \phi(\mathbf{x}^*, \mathbf{y}^*) \geq \max_{\mathbf{y} \in \mathcal{Y}} \phi(\mathbf{x}^*, \mathbf{y}) \geq \min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} \phi(\mathbf{x}, \mathbf{y})$
 - By lemma A, we know left is no bigger than right, so every \geq is $=$.
 - This also means $\mathbf{y}^* \in \operatorname{argmax}_{\mathbf{y} \in \mathcal{Y}} \underline{\phi}(\mathbf{y})$
 - (\leftarrow)
 - If we have $\max_{\mathbf{y} \in \mathcal{Y}} \min_{\mathbf{x} \in \mathcal{X}} \phi(\mathbf{x}, \mathbf{y}) = \min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} \phi(\mathbf{x}, \mathbf{y})$, and define $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}} \bar{\phi}(\mathbf{x})$ and $\mathbf{y}^* \in \operatorname{argmax}_{\mathbf{y} \in \mathcal{Y}} \underline{\phi}(\mathbf{y})$, this means $\bar{\phi}(\mathbf{x}^*) = \underline{\phi}(\mathbf{y}^*)$
 - by the fact of $\underline{\phi}(\mathbf{y}^*) \leq \phi(\mathbf{x}^*, \mathbf{y}^*) \leq \bar{\phi}(\mathbf{x}^*)$, we know every \leq is $=$.
 - Since $\forall \mathbf{x}, \mathbf{y}, \phi(\mathbf{x}^*, \mathbf{y}) \leq \bar{\phi}(\mathbf{x}^*)$ and $\phi(\mathbf{x}, \mathbf{y}^*) \geq \underline{\phi}(\mathbf{y}^*)$, we prove that $(\mathbf{x}^*, \mathbf{y}^*)$ is saddle point.
- Example of Saddle Point not exists $\phi(x, y) = (x - y)^2$, $\mathcal{X} = [0, 1]$, $\mathcal{Y} = [0, 1]$
 - $\bar{\phi}(x) = \max \{x^2, (x - 1)^2\}$, $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \phi(x, y) = \frac{1}{4}$, while $\underline{\phi}(y) = 0$, $\max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} \phi(x, y) = 0$.

Minimal point for Convex-Concave min-max function

- Definition 13.2 (Convex-concave function) A function $\phi(\mathbf{x}, \mathbf{y}) : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ 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 13.3 (Strongly-convex-concave function) $\phi(\mathbf{x}, \mathbf{y}) : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ is *strongly-convex-strongly-concave* if $\exists \mu_1, \mu_2$ s.t.
 - $\phi(\mathbf{x}, \mathbf{y})$ 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 $\mathbf{y} \in \mathcal{Y}$ for every fixed $\mathbf{x} \in \mathcal{X}$.
- Theorem 13.4 (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 $\max_{\mathbf{y} \in \mathcal{Y}} \min_{\mathbf{x} \in \mathcal{X}} \phi(\mathbf{x}, \mathbf{y}) = \min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} \phi(\mathbf{x}, \mathbf{y})$.

- Remark

- First proven by John von Neumann in 1928, can also be proved by on-line learning algorithm.
- can be extended to lower-semicontinuous and quasi-convex function.
- If $\phi(\mathbf{x}, \mathbf{y})$ is strongly-convex-strongly-concave, then we can remove the compactness, and saddle point is unique.

Accuracy Measure of Minimax: Duality Gap

- Definition of Duality Gap **duality gap** := $\max_{\mathbf{y} \in \mathcal{Y}} \phi(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x} \in \mathcal{X}} \phi(\mathbf{x}, \hat{\mathbf{y}}) = \bar{\phi}(\hat{\mathbf{x}}) - \underline{\phi}(\hat{\mathbf{y}})$
 - If saddle point exists, $DG(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \geq 0$, since $\bar{\phi}(\hat{\mathbf{x}}) - \underline{\phi}(\hat{\mathbf{y}}) = (\bar{\phi}(\hat{\mathbf{x}}) - \min_{\mathbf{x}} \bar{\phi}(\mathbf{x})) + (\max_{\mathbf{y}} \underline{\phi}(\mathbf{y}) - \underline{\phi}(\hat{\mathbf{y}})) \geq 0$
- Iteratively update may not be successfull.
 - Example $\min_{x \in [-1,1]} \max_{y \in [-1,1]} xy, x_0 = 1, y_0 = 1, x_1 = -1, y_1 = -1, \dots$

Gradient Descent Ascent (GDA)

- Algorithm $\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}(\mathbf{x}_t - \gamma \nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t)), \mathbf{y}_{t+1} = \Pi_{\mathcal{Y}}(\mathbf{y}_t + \gamma \nabla_{\mathbf{y}} \phi(\mathbf{x}_t, \mathbf{y}_t))$.
- Not converging example $\min_x \max_y xy, x_{t+1} = x_t - \gamma y_t, y_{t+1} = x_t + \gamma x_t, (x_{t+1}^2 + y_{t+1}^2) = (1 + \gamma^2)(x_t^2 + y_t^2)$, saddle point $(0, 0)$.

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

- Assumption μ -strongly convex in \mathbf{x} and μ -strongly concave in \mathbf{y} , gradient $\nabla_{\mathbf{x}} f$ and $\nabla_{\mathbf{y}} f$, L -Lipschitz smooth in (\mathbf{x}, \mathbf{y}) separately.
 - Then saddle point is unique.
- Theorem 13.5 In SC-SC setting, GDA with stepsize $\eta < \frac{\mu}{2L^2}$ converges linearly,

$$\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 + \|\mathbf{y}_{t+1} - \mathbf{y}^*\|^2 \leq (1 + 4\eta^2 L^2 - 2\eta\mu) (\|\mathbf{x}_t - \mathbf{x}^*\|^2 + \|\mathbf{y}_t - \mathbf{y}^*\|^2).$$
 - Proof
 - By SC-SC,
 - $$(\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*))^\top (\mathbf{x} - \mathbf{x}^*) + (\nabla_{\mathbf{y}} f(\mathbf{x}^*, \mathbf{y}^*) - \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}))^\top (\mathbf{y} - \mathbf{y}^*) \geq \mu \|\mathbf{x} - \mathbf{x}^*\|^2 + \mu \|\mathbf{y} - \mathbf{y}^*\|^2$$
 - By Lipschitz $\|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*)\|^2 \leq 2L \|\mathbf{x} - \mathbf{x}^*\|^2 + 2L \|\mathbf{y} - \mathbf{y}^*\|^2$ and $\|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{y}} f(\mathbf{x}^*, \mathbf{y}^*)\|^2 \leq 2L \|\mathbf{x} - \mathbf{x}^*\|^2 + 2L \|\mathbf{y} - \mathbf{y}^*\|^2$
 - Then $\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 + \|\mathbf{y}_{t+1} - \mathbf{y}^*\|^2 =$

$$\begin{aligned} & \|\Pi_X(\mathbf{x}_t - \eta \nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t)) - \Pi_X(\mathbf{x}^* - \eta \nabla_{\mathbf{x}} \phi(\mathbf{x}^*, \mathbf{y}^*))\|^2 + \|\Pi_Y(\mathbf{y}_t + \eta \nabla_{\mathbf{y}} \phi(\mathbf{x}_t, \mathbf{y}_t)) - \Pi_Y(\mathbf{y}^* + \eta \nabla_{\mathbf{y}} \phi(\mathbf{x}^*, \mathbf{y}^*))\|^2 \\ & \leq \|\mathbf{x}_t - \eta \nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t) - \mathbf{x}^* + \eta \nabla_{\mathbf{x}} \phi(\mathbf{x}^*, \mathbf{y}^*)\|^2 + \|\mathbf{y}_t + \eta \nabla_{\mathbf{y}} \phi(\mathbf{x}_t, \mathbf{y}_t) - \mathbf{y}^* - \eta \nabla_{\mathbf{y}} \phi(\mathbf{x}^*, \mathbf{y}^*)\|^2 \\ & = \|\mathbf{x}_t - \mathbf{x}^*\|^2 + \eta^2 \|\nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} \phi(\mathbf{x}^*, \mathbf{y}^*)\|^2 - 2\eta (\nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} \phi(\mathbf{x}^*, \mathbf{y}^*))^\top (\mathbf{x}_t - \mathbf{x}^*) + \\ & \quad \|\mathbf{y}_t - \mathbf{y}^*\|^2 + \eta^2 \|\nabla_{\mathbf{y}} \phi(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} \phi(\mathbf{x}^*, \mathbf{y}^*)\|^2 - 2\eta (\nabla_{\mathbf{y}} \phi(\mathbf{x}^*, \mathbf{y}^*) - \nabla_{\mathbf{y}} \phi(\mathbf{x}_t, \mathbf{y}_t))^\top (\mathbf{y}_t - \mathbf{y}^*) \end{aligned}$$
 - Plug in the two inequality, we get the proof.
 - Setting $\eta = \frac{\mu}{4L^2}$, then $\|\mathbf{x}_T - \mathbf{x}^*\|^2 + \|\mathbf{y}_T - \mathbf{y}^*\|^2 \leq (1 - 4\mu^2/L^2)^T (\|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \|\mathbf{y}_0 - \mathbf{y}^*\|^2)$, complexity of $\mathcal{O}(\kappa^2 \log \frac{1}{\epsilon})$.

Extragradient Method in Convex-Concave Setting

- Algorithm
 - $\mathbf{x}_{t+\frac{1}{2}} = \Pi_{\mathcal{X}}(\mathbf{x}_t - \eta \nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t)), \mathbf{y}_{t+\frac{1}{2}} = \Pi_{\mathcal{Y}}(\mathbf{y}_t + \eta \nabla_{\mathbf{y}} \phi(\mathbf{x}_t, \mathbf{y}_t))$
 - $\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}(\mathbf{x}_t - \eta \nabla_{\mathbf{x}} \phi(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}})), \mathbf{y}_{t+1} = \Pi_{\mathcal{Y}}(\mathbf{y}_t + \eta \nabla_{\mathbf{y}} \phi(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}))$
- Example $\min_x \max_y xy$ gives $x_{t+1} = x_t - \eta(y_t + \eta x_t), y_{t+1} = y_t + \eta(x_t - \eta y_t)$,
 - then $\|x_{t+1}\|^2 + \|y_{t+1}\|^2 = (1 - \eta^2 + \eta^4)(\|x_t\|^2 + \|y_t\|^2)$
- Theorem 13.6 Assume ϕ is convex-concave, L -Lipschitz smooth, \mathcal{X} has diameter $D_{\mathcal{X}}$ and \mathcal{Y} has diameter $D_{\mathcal{Y}}$, then ExtraGradient with stepsize $\eta < 1/2L$ satisfies $\max_{\mathbf{y} \in \mathcal{Y}} \phi\left(\frac{1}{T} \sum_{t=1}^T \mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}\right) - \min_{\mathbf{x} \in \mathcal{X}} \phi\left(\mathbf{x}, \frac{1}{T} \sum_{t=1}^T \mathbf{y}_{t+\frac{1}{2}}\right) \leq \frac{D_{\mathcal{X}}^2 + D_{\mathcal{Y}}^2}{2\eta T}$.
 - Proof
 - For complexity, we denote the unprojected point with $\tilde{\mathbf{x}}_a$, then $\mathbf{x}_a = \Pi_{\mathcal{X}}(\tilde{\mathbf{x}}_a)$.
 - We deal with $\nabla_{\mathbf{x}} \phi(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}})^\top (\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x})$ for arbitrary \mathbf{x} , by fact of $\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x} = (\mathbf{x}_{t+1} - \mathbf{x}) + (\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1})$, we also need $\nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t)$ as reference point, then
 - $$\nabla_{\mathbf{x}} \phi(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}})^\top (\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}) = \nabla_{\mathbf{x}} \phi(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}})^\top (\mathbf{x}_{t+1} - \mathbf{x}) + \nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t)^\top (\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1}) +$$
 - $$(\nabla_{\mathbf{x}} \phi(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}) - \nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t))^\top (\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1})$$

- We can bound each term,
 - $\nabla_{\mathbf{x}} \phi \left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}} \right)^{\top} (\mathbf{x}_{t+1} - \mathbf{x}) = \frac{1}{\eta} (\mathbf{x}_t - \tilde{\mathbf{x}}_{t+1})^{\top} (\mathbf{x}_{t+1} - \mathbf{x}) \leq \frac{1}{\eta} (\mathbf{x}_t - \mathbf{x}_{t+1})^{\top} (\mathbf{x}_{t+1} - \mathbf{x})$
 $= \frac{1}{2\eta} [\|\mathbf{x} - \mathbf{x}_t\|^2 - \|\mathbf{x} - \mathbf{x}_{t+1}\|^2 - \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2]$ by projection Fact 4.1
 - $\nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t)^{\top} (\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1}) = \frac{1}{\eta} (\mathbf{x}_t - \tilde{\mathbf{x}}_{t+\frac{1}{2}})^{\top} (\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1}) \leq \frac{1}{\eta} (\mathbf{x}_t - \mathbf{x}_{t+\frac{1}{2}})^{\top} (\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1})$
 $= \frac{1}{2\eta} [\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 - \|\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1}\|^2 - \|\mathbf{x}_t - \mathbf{x}_{t+\frac{1}{2}}\|^2]$ by same reason,
 - $(\nabla_{\mathbf{x}} \phi \left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}} \right) - \nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t))^{\top} (\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1}) \leq \|\nabla_{\mathbf{x}} \phi \left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}} \right) - \nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t)\| \|\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1}\|$
 $\leq L [\|\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_t\| + \|\mathbf{y}_{t+\frac{1}{2}} - \mathbf{y}_t\|] \|\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1}\| \leq \frac{L}{2} \|\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_t\|^2 + \frac{L}{2} \|\mathbf{y}_{t+\frac{1}{2}} - \mathbf{y}_t\|^2 + L \|\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1}\|^2$ by smoothness and Young's equality
- Add them together and we get $\nabla_{\mathbf{x}} \phi \left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}} \right)^{\top} (\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x})$
 $\leq \frac{2}{\eta} [\|\mathbf{x}_t - \mathbf{x}\|^2 - \|\mathbf{x} - \mathbf{x}_{t+1}\|^2] + \left(L - \frac{1}{2\eta} \right) \|\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1}\|^2 + \left(\frac{L}{2} - \frac{1}{2\eta} \right) \|\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_t\|^2 + \frac{L}{2} \|\mathbf{y}_{t+\frac{1}{2}} - \mathbf{y}_t\|^2$
- Similarly $-\nabla_{\mathbf{y}} \phi \left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}} \right)^{\top} (\mathbf{y}_{t+\frac{1}{2}} - \mathbf{y}) \leq \frac{2}{\eta} [\|\mathbf{y}_t - \mathbf{y}\|^2 - \|\mathbf{y} - \mathbf{y}_{t+1}\|^2] + \left(L - \frac{1}{2\eta} \right) \|\mathbf{y}_{t+\frac{1}{2}} - \mathbf{y}_{t+1}\|^2 + \left(\frac{L}{2} - \frac{1}{2\eta} \right) \|\mathbf{y}_{t+\frac{1}{2}} - \mathbf{y}_t\|^2 + \frac{L}{2} \|\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_t\|^2$
- Then $\nabla_{\mathbf{x}} \phi \left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}} \right)^{\top} (\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}) - \nabla_{\mathbf{y}} \phi \left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}} \right)^{\top} (\mathbf{y}_{t+\frac{1}{2}} - \mathbf{y}) \leq \frac{2}{\eta} [\|\mathbf{x}_t - \mathbf{x}\|^2 - \|\mathbf{x} - \mathbf{x}_{t+1}\|^2] + \left(L - \frac{1}{2\eta} \right) [\|\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1}\|^2 + \|\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_t\|^2 + \|\mathbf{y}_{t+\frac{1}{2}} - \mathbf{y}_{t+1}\|^2 + \|\mathbf{y}_{t+\frac{1}{2}} - \mathbf{y}_t\|^2]$
 - choosing $L \leq 1/2\eta$ we can omit all $(L - 1/2\eta)$ term, and only leave the $a_t - a_{t+1}$ like term.
- We then use the convex concave condition, so that
 $\phi \left(\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t, \mathbf{y} \right) - \phi \left(\mathbf{x}, \frac{1}{T} \sum_{t=1}^T \mathbf{y}_t \right) \leq \frac{1}{T} \sum_{t=1}^T \phi \left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y} \right) - \phi \left(\mathbf{x}, \mathbf{y}_{t+\frac{1}{2}} \right)$
 $\leq \frac{1}{T} \sum_{t=1}^T -\nabla_{\mathbf{y}} \phi \left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}} \right)^{\top} (\mathbf{y}_{t+\frac{1}{2}} - \mathbf{y}) + \nabla_{\mathbf{x}} \phi \left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}} \right)^{\top} (\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x})$

- Remark

- $\mathcal{O}(\frac{1}{T})$ convergence rate for average iterates at *mid-point*, this rate is optimal with no further assumption [Ouyang and Xu, 2021].
- Can be extend to Bregman divergence, Mirror Prox.
- Best-iterate and last-iterate convergence rate of $\mathcal{O}(\frac{1}{\sqrt{T}})$ for primal-dual gap [Yang et al., 2022]. This is the lower bound by EG [Golowich et al., 2020].

- Theorem 12.7 (Mokhtari et al., 2020) In SC-SC setting, EG with stepsize $\eta = 1/8L$ converges linearly,
 $\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 + \|\mathbf{y}_{t+1} - \mathbf{y}^*\|^2 \leq \left(1 - \frac{\mu}{4L}\right) \{\|\mathbf{x}_t - \mathbf{x}^*\|^2 + \|\mathbf{y}_t - \mathbf{y}^*\|^2\}$
◦ Remark This $O(\kappa \log \frac{1}{\epsilon})$ complexity is optimal for SC-SC setting [Zhang et al., 2021]

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

- Algorithm
 - $\mathbf{x}_{t+\frac{1}{2}} = \Pi_{\mathcal{X}} \left(\mathbf{x}_t - \eta \nabla_{\mathbf{x}} \phi \left(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}} \right) \right)$, $\mathbf{y}_{t+\frac{1}{2}} = \Pi_{\mathcal{Y}} \left(\mathbf{y}_t + \eta \nabla_{\mathbf{y}} \phi \left(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}} \right) \right)$
 - $\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} \left(\mathbf{x}_t - \eta \nabla_{\mathbf{x}} \phi \left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}} \right) \right)$, $\mathbf{y}_{t+1} = \Pi_{\mathcal{Y}} \left(\mathbf{y}_t + \eta \nabla_{\mathbf{y}} \phi \left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}} \right) \right)$
- If we ignore projection, equivalently we have $\omega_{t+\frac{1}{2}} = \omega_t - \eta F(\omega_{t-\frac{1}{2}}) = \omega_{t-\frac{1}{2}} + (\omega_t - \omega_{t-1}) + (\omega_{t-1} - \omega_{t-\frac{1}{2}}) - \eta F(\omega_{t-\frac{1}{2}})$
 $= \omega_{t-\frac{1}{2}} - 2\eta F(\omega_{t-\frac{1}{2}}) + \eta F(\omega_{t-\frac{3}{2}})$
◦ reformulate as $\omega_{t+1} = \omega_t - 2\eta F(\omega_t) + \eta F(\omega_{t-1}) = \omega_t - \eta (F(\omega_t) - (F(\omega_{t-1}) - F(\omega_t)))$, negative momentum
- Query gradient only once each iteration, comparing with extra-gradient
- similar convergence guarantees as EG for SC-SC and C-C setting.

Proximal Point Algorithm (PPA)

- $(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) \leftarrow \underset{\mathbf{x} \in \mathcal{X}}{\text{argmin}} \underset{\mathbf{y} \in \mathcal{Y}}{\text{argmax}} \left\{ \phi(\mathbf{x}, \mathbf{y}) + \frac{1}{2\eta} \|\mathbf{x} - \mathbf{x}_t\|^2 - \frac{1}{2\eta} \|\mathbf{y} - \mathbf{y}_t\|^2 \right\}$
- The problem of $\phi = \mathbf{x}^{\top} \mathbf{y}$ gives $\mathbf{x}_{t+1} = \frac{\mathbf{x}_t - \eta \mathbf{y}_t}{1+\eta^2}$, $\mathbf{y}_{t+1} = \frac{\mathbf{y}_t + \eta \mathbf{x}_t}{1+\eta^2}$, $\|\mathbf{x}_{t+1}\|^2 + \|\mathbf{y}_{t+1}\|^2 = \frac{1}{1+\eta^2} (\|\mathbf{x}_t\|^2 + \|\mathbf{y}_t\|^2)$, converge.
- PPA has been shown to converge with $\mathcal{O}(1/T)$ rate in convex-concave case.
- Implicit update of PPA $\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} (\mathbf{x}_t - \eta \nabla_{\mathbf{x}} \phi(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}))$, $\mathbf{y}_{t+1} = \Pi_{\mathcal{Y}} (\mathbf{y}_t + \eta \nabla_{\mathbf{y}} \phi(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}))$
 - EG and OGDA can be viewed as approximate PPA with error $\mathbf{z}_{t+1} = \Pi_{\mathcal{Z}} (\mathbf{z}_t - \eta F(\mathbf{z}_{t+1}) + \epsilon_t)$
 - EG: $\epsilon_t = \eta [F(\mathbf{z}_{t+1}) - F(\mathbf{z}_{t+\frac{1}{2}})]$, OGDA: $\epsilon_t = \eta [(F(\mathbf{z}_{t+1}) - F(\mathbf{z}_t)) - (F(\mathbf{z}_t) - F(\mathbf{z}_{t-1}))]$.

Concave Games

- Definition
 - Finite number of players, $i \in \mathcal{N} = \{1, \dots, N\}$,
 - action profile $\mathbf{x} = (\mathbf{x}_i, \mathbf{x}_{-i}) = (\mathbf{x}_1, \dots, \mathbf{x}_N) \in \mathcal{X} = \prod_i \mathcal{X}_i$, where \mathcal{X}_i is a compact convex subset \mathbb{R}^{d_i} and $\mathbf{x}_i \in \mathcal{X}_i$ is the action of player i ,
 - Payoff (or utility) function $u_i : \mathcal{X} \rightarrow \mathbb{R}$, (which player i want to maximize)
- Assumption $u_i(\mathbf{x}_i, \mathbf{x}_{-i})$ is concave in \mathbf{x}_i
- Definition of Nash equilibrium The action profile $\mathbf{x}^* \in \mathcal{X}$ that is resilient to unilateral derivations, which means $u_i(\mathbf{x}_i^*, \mathbf{x}_{-i}^*) \geq u_i(\mathbf{x}_i, \mathbf{x}_{-i}^*), \forall \mathbf{x}_i \in \mathcal{X}_i, i \in \mathcal{N}$.
- Existence theorem [Debreu, 1952] every concave game admits a Nash equilibrium.
- First-order characterization $\langle \nabla_{\mathbf{x}_i} u_i(\mathbf{x}_i^*, \mathbf{x}_{-i}^*), \mathbf{x}_i - \mathbf{x}_i^* \rangle \leq 0, \forall \mathbf{x}_i \in \mathcal{X}_i$

Variational Inequalities

- Definition Let $\mathcal{Z} \subset \mathbb{R}^d$ be a nonempty subset and consider a mapping $\mathbf{F} : \mathcal{Z} \rightarrow \mathbb{R}^d$, then *variational inequality problem (VI)* is to find $\mathbf{z}^* \in \mathcal{Z}$ such that $\langle \mathbf{F}(\mathbf{z}^*), \mathbf{z} - \mathbf{z}^* \rangle \geq 0, \forall \mathbf{z} \in \mathcal{Z}$.
 - This is not just optimization problem, \mathbf{F} can be any vector field.
- Existence [Stampacchia, 1966] If \mathcal{Z} is a nonempty convex compact subset of \mathbb{R}^d and $\mathbf{F} : \mathcal{Z} \rightarrow \mathbb{R}^d$ is continuous, then there exists a solution \mathbf{z}^* to (VI).
 - Note compactness can be replaced by a *coercivity condition* and the result can be generalized to a set valued mapping \mathbf{F} .

VI with monotone operator

- Definition $\mathbf{F} : \mathcal{Z} \rightarrow \mathbb{R}^d$ is
 - *monotone* if $\langle \mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \geq 0, \forall \mathbf{u}, \mathbf{v} \in \mathcal{Z}$,
 - μ -*strongly-monotone* if $\langle \mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \geq \mu \|\mathbf{u} - \mathbf{v}\|^2, \forall \mathbf{u}, \mathbf{v} \in \mathcal{Z}$
- Definition
 - The (*strong*) solution (of Stampacchia VI) is to find $\mathbf{z}^* \in \mathcal{Z}$ s.t. $\langle \mathbf{F}(\mathbf{z}^*), \mathbf{z} - \mathbf{z}^* \rangle \geq 0 \forall \mathbf{z} \in \mathcal{Z}$
 - The *Weak solution* (of Minty VI) is to find $\mathbf{z}^* \in \mathcal{Z}$ s.t. $\langle \mathbf{F}(\mathbf{z}), \mathbf{z} - \mathbf{z}^* \rangle \geq 0 \forall \mathbf{z} \in \mathcal{Z}$.
- If \mathbf{F} monotone, then strong solution is also a weak solution since
 - $\langle \mathbf{F}(\mathbf{z}) - \mathbf{F}(\mathbf{z}^*), \mathbf{z} - \mathbf{z}^* \rangle \geq 0$, then $\langle \mathbf{F}(\mathbf{z}), \mathbf{z} - \mathbf{z}^* \rangle = \langle \mathbf{F}(\mathbf{z}^*), \mathbf{z} - \mathbf{z}^* \rangle + \langle \mathbf{F}(\mathbf{z}) - \mathbf{F}(\mathbf{z}^*), \mathbf{z} - \mathbf{z}^* \rangle \geq 0$
- If \mathbf{F} continuous, then a weak solution is also a strong solution, let $\mathbf{z} = \mathbf{z}^* + \lambda \mathbf{d}$, if there is a open ball inside \mathcal{Z} , then this is provable.
- We use $\epsilon_{VI}(\hat{\mathbf{z}}) := \max_{\mathbf{z} \in \mathcal{Z}} \langle \mathbf{F}(\mathbf{z}), \hat{\mathbf{z}} - \mathbf{z} \rangle$ to measure the inaccuracy of a solution $\hat{\mathbf{z}}$.
- VI is a generality over a wide range of problems
 - *Convex minimization* $\mathbf{F} = \nabla f$, with f convex
 - *Min-Max* $\mathbf{F} = (\nabla_{\mathbf{x}} \phi, -\nabla_{\mathbf{y}} \phi)$, solution exists only when saddle point exists.
 - *Concave or monotone games* $\mathbf{F}(\mathbf{z}) = (-\nabla_i u_i(\mathbf{z}_i, \mathbf{z}_{-i}))_{i \in \mathcal{N}}$
- Some possible assumptions
 - \mathcal{Z} is a closed convex subset of \mathbb{R}^d
 - Solution of VI exists
 - Mapping \mathbf{F} is monotone
 - \mathbf{F} also Lipschitz continuous, $\|\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v})\| \leq L \|\mathbf{u} - \mathbf{v}\|, \forall \mathbf{u}, \mathbf{v} \in \mathcal{Z}$.
- Extragradient Algorithm first $\tilde{\mathbf{z}}_{t+1} = \Pi_{\mathcal{Z}}(\mathbf{z}_t - \eta_t \mathbf{F}(\mathbf{z}_t))$, then $\mathbf{z}_{t+1} = \Pi_{\mathcal{Z}}(\mathbf{z}_t - \eta_t \mathbf{F}(\tilde{\mathbf{z}}_{t+1}))$.
- Theorem 13.8 (Nemirovski, 2004) If \mathbf{F} nbitibe and L -Lipschitz, and other previous assumption hold, then setting $\eta_t = \eta = \frac{1}{\sqrt{2}L}$ and EG gives $\max_{\mathbf{z} \in \mathcal{Z}} \left\langle \mathbf{F}(\mathbf{z}), \frac{1}{T} \sum_{t=1}^T \tilde{\mathbf{z}}_t - \mathbf{z} \right\rangle \leq \frac{\sqrt{2}LD_{\mathcal{Z}}^2}{T}$, where $D_{\mathcal{Z}} = \max_{\mathbf{z}, \mathbf{z}'} \|\mathbf{z} - \mathbf{z}'\|_2$ is the 2-norm diameter of \mathcal{Z} .
- Algorithm for VI
 - *GDA* $\mathbf{z}_{t+1} = \mathbf{z}_t - \eta_t \mathbf{F}(\mathbf{z}_t)$
 - *PPA* $\mathbf{Z}_{t+1} = \mathbf{Z}_t - \eta_t \mathbf{F}(\mathbf{Z}_{t+1})$
 - *OGDA* $\mathbf{z}_{t+1} = \mathbf{z}_t - \eta_t (2\mathbf{F}(\mathbf{z}_t) - \mathbf{F}(\mathbf{z}_{t-1}))$
 - *Reflected Gradient* $\mathbf{z}_{t+1} = \mathbf{z}_t - \eta_t \mathbf{F}(2\mathbf{z}_t - \mathbf{z}_{t-1})$
 - *Halpern iteration* $\mathbf{z}_{t+1} = \lambda_k \mathbf{z}_0 + (1 - \lambda_k)(\mathbf{z}_t - \eta_t \mathbf{F}(\mathbf{z}_t))$, use \mathbf{z}_0 as arching, this gives last iterate convergence.