Chapter 13 minimax

- Fomulation $\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{x}, \mathbf{y})$
- 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, $\Delta(I) = \left\{\mathbf{x} \in \mathbb{R}^{|I|} : \mathbf{x}_i \geq 0, i \in I, \sum_{i\in I} \mathbf{x}_i = 1\right\}$.
  - *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* $\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \mathbb{E}_{\xi\sim p_{\text{data}}}\left[\log D_{\mathbf{y}}(\xi)\right] + \mathbb{E}_{\zeta\sim p_\zeta}\left[\log\left(1 - D_{\mathbf{y}}\left(G_{\mathbf{x}}(\zeta)\right)\right)\right]$.
  - *Adversarial Robustness* $\min_{\mathbf{x}} \max_{P\in\mathcal{P}} \mathbb{E}_{\boldsymbol{\xi}\sim P}[\ell(\mathbf{x}, \boldsymbol{\xi})]$

Saddle Points and Global Minimax Points

- Definition of saddle point $(\mathbf{x}^*, \mathbf{y}^*)$ is a *saddle point* if $\forall \mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}, \phi\left(\mathbf{x}^*, \mathbf{y}\right) \leq \phi\left(\mathbf{x}^*, \mathbf{y}^*\right) \leq \phi\left(\mathbf{x}, \mathbf{y}^*\right)$.
  - corresponds to *Nash equilibrium*, simulaneous 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 $\forall \mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}, \phi\left(\mathbf{x}^*, \mathbf{y}\right) \leq \phi\left(\mathbf{x}^*, \mathbf{y}^*\right) \leq \max_{\mathbf{y}'\in\mathcal{Y}} \phi\left(\mathbf{x}, \mathbf{y}'\right)$.
  - Corresponds to *Stackelberg equilibrium*, sequential game, best responds to the best response.

Primal and Dual Problems

- Primal $\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{x}, \mathbf{y}) := \min_{\mathbf{x}\in\mathcal{X}} \overline{\phi}(\mathbf{x})$
- 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
    - (->)
      - Saddle point $\Leftrightarrow \min_{\mathbf{x}\in\mathcal{X}} \phi\left(\mathbf{x}, \mathbf{y}^*\right) \geq \phi\left(\mathbf{x}^*, \mathbf{y}^*\right) \geq \max_{\mathbf{y}\in\mathcal{Y}} \phi\left(\mathbf{x}^*, \mathbf{y}\right)$
      - Then by definition $\max_{\mathbf{y}\in\mathcal{Y}} \min_{\mathbf{x}\in\mathcal{X}} \phi\left(\mathbf{x}, \mathbf{y}\right) \geq \min_{\mathbf{x}\in\mathcal{X}} \phi\left(\mathbf{x}, \mathbf{y}^*\right) \geq \phi\left(\mathbf{x}^*, \mathbf{y}^*\right) \geq \max_{\mathbf{y}\in\mathcal{Y}} \phi\left(\mathbf{x}^*, \mathbf{y}\right) \geq \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \phi\left(\mathbf{x}, \mathbf{y}\right)$
      - By lemma A, we know left is no bigger than right, so every $\geq$ is $=$.
      - This also means $\mathbf{y}^* \in \operatorname{argmax}_{\mathbf{y}\in\mathcal{Y}} \underline{\phi}(\mathbf{y})$
    - (<-)
      - 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 $\overline{\phi}(\mathbf{x}^*) = \underline{\phi}(\mathbf{y}^*)$
      - by the fact of $\underline{\phi}(\mathbf{y}^*) \leq \phi(\mathbf{x}^*, \mathbf{y}^*) \leq \overline{\phi}(\mathbf{x}^*)$, we know every $\leq$ is $=$.
      - Since $\forall \mathbf{x}, \mathbf{y}, \phi(\mathbf{x}^*, \mathbf{y}) \leq \overline{\phi}(\mathbf{x}^*)$ 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 $\phi(x, y) = (x - y)^2, \mathcal{X} = [0, 1], \mathcal{Y} = [0, 1]$

- $\bar{\phi}(x) = \max\left\{x^2, (x-1)^2\right\}, \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \phi(x,y) = \frac{1}{4}$, 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 $\phi(\mathbf{x},\mathbf{y}) : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ 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 13.3 (Strongly-convex-concave function) $\phi(\mathbf{x},\mathbf{y}) : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ 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** $:= \max_{\mathbf{y}\in\mathcal{Y}} \phi(\hat{\mathbf{x}},\mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x},\hat{\mathbf{y}}) = \overline{\phi}(\hat{\mathbf{x}}) - \underline{\phi}(\hat{\mathbf{y}})$
  - If saddle point exists, $\mathrm{DG}(\hat{\mathbf{x}},\hat{\mathbf{y}}) \geq 0$, since $\overline{\phi}(\hat{\mathbf{x}}) - \underline{\phi}(\hat{\mathbf{y}}) = (\overline{\phi}(\hat{\mathbf{x}}) - \min_x \overline{\phi}(\mathbf{x})) + (\max \underline{\phi}(\mathbf{y}) - \underline{\phi}(\hat{\mathbf{y}})) \geq 0$
- Iteratively update may not be successfull.
  - Example $\min_{x\in[-1,1]} \max_{y\in[-1,1]} xy$, $x_0 = 1, y_0 = 1, x_1 = -1, y_1 = -1, \ldots$.

Gradient Descent Ascent (GDA)

- Algorithm $\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}\left(\mathbf{x}_t - \gamma\nabla_{\mathbf{x}}\phi\left(\mathbf{x}_t,\mathbf{y}_t\right)\right)$, $\mathbf{y}_{t+1} = \Pi_{\mathcal{Y}}\left(\mathbf{y}_t + \gamma\nabla_{\mathbf{y}}\phi\left(\mathbf{x}_t,\mathbf{y}_t\right)\right)$.
- 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, $\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 + \|\mathbf{y}_{t+1} - \mathbf{y}^*\|^2 \leq \left(1 + 4\eta^2L^2 - 2\eta\mu\right)\left(\|\mathbf{x}_t - \mathbf{x}^*\|^2 + \|\mathbf{y}_t - \mathbf{y}^*\|^2\right)$.
  - Proof
    - By SC-SC, $\left(\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{x}} f\left(\mathbf{x}^*,\mathbf{y}^*\right)\right)^\top \left(\mathbf{x} - \mathbf{x}^*\right) + \left(\nabla_{\mathbf{y}} f\left(\mathbf{x}^*,\mathbf{y}^*\right) - \nabla_{\mathbf{y}} f(\boldsymbol{x},\boldsymbol{y})\right)^\top \left(\mathbf{y} - \mathbf{y}^*\right) \geq \mu\|\mathbf{x} - \mathbf{x}^*\|^2 + \mu\|\mathbf{y} - \mathbf{y}^*\|^2$
    - By Lipschitz $\|\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{x}} f\left(\mathbf{x}^*,\mathbf{y}^*\right)\|^2 \leq 2L\|\mathbf{x} - \mathbf{x}^*\|^2 + 2L\|\mathbf{y} - \mathbf{y}^*\|^2$ and $\|\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{y}} f\left(\mathbf{x}^*,\mathbf{y}^*\right)\|^2 \leq 2L\|\mathbf{x} - \mathbf{x}^*\|^2 + 2L\|\mathbf{y} - \mathbf{y}^*\|^2$
    - Then $\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 + \|\mathbf{y}_{t+1} - \mathbf{y}^*\|^2 = \|\Pi_X\left(\mathbf{x}_t - \eta\nabla_{\mathbf{x}}\phi\left(\mathbf{x}_t,\mathbf{y}_t\right)\right) - \Pi_X\left(\mathbf{x}^* - \eta\nabla_{\mathbf{x}}\phi\left(\mathbf{x}^*,\mathbf{y}^*\right)\right)\|^2 + \|\Pi_Y\left(\mathbf{y}_t + \eta\nabla_{\mathbf{y}}\phi\left(\mathbf{x}_t,\mathbf{y}_t\right)\right) - \Pi_Y\left(\mathbf{y}^* + \eta\nabla_{\mathbf{y}}\phi\left(\mathbf{x}^*,\mathbf{y}^*\right)\right)\|^2$

$\leq \|\mathbf{x}_t - \eta \nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t) - \mathbf{x}^* + \eta \nabla_{\mathbf{x}} \phi(\mathbf{x}^*, \mathbf{y}^*)\|^2 + \|\mathbf{y}_t + \eta \nabla_{\mathbf{y}} \phi(\mathbf{x}_t, \mathbf{y}_t) - \mathbf{y}^* - \eta \nabla_{\mathbf{y}} \phi(\mathbf{x}^*, \mathbf{y}^*)\|$
$= \|\mathbf{x}_t - \mathbf{x}^*\|^2 + \eta^2 \|\nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} \phi(\mathbf{x}^*, \mathbf{y}^*)\|^2 - 2\eta (\nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} \phi(\mathbf{x}^*, \mathbf{y}^*))^\top (\mathbf{x}_t - \mathbf{x}^*) +$
$\|\mathbf{y}_t - \mathbf{y}^*\|^2 + \eta^2 \|\nabla_{\mathbf{y}} \phi(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} \phi(\mathbf{x}^*, \mathbf{y}^*)\|^2 - 2\eta (\nabla_{\mathbf{y}} \phi(\mathbf{x}^*, \mathbf{y}^*) - \nabla_{\mathbf{y}} \phi(\mathbf{x}_t, \mathbf{y}_t))^\top (\mathbf{y}_t - \mathbf{y}^*)$

- Plug in the two inequality, we get the proof.

- Setting $\eta = \frac{\mu}{4L^2}$, then $\|\mathbf{x}_T - \mathbf{x}^*\|^2 + \|\mathbf{y}_T - \mathbf{y}^*\|^2 \leq \left(1 - 4\mu^2/L^2\right)^T \left(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \|\mathbf{y}_0 - \mathbf{y}^*\|^2\right)$, complexity of $\mathcal{O}\left(\kappa^2 \log \frac{1}{\epsilon}\right)$.

Extragradient Method in Convex-Concave Setting

- Algorithm
  - $\mathbf{x}_{t+\frac{1}{2}} = \Pi_{\mathcal{X}}(\mathbf{x}_t - \eta \nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t))$, $\mathbf{y}_{t+\frac{1}{2}} = \Pi_{\mathcal{Y}}(\mathbf{y}_t + \eta \nabla_{\mathbf{y}} \phi(\mathbf{x}_t, \mathbf{y}_t))$,
  - $\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}\left(\mathbf{x}_t - \eta \nabla_{\mathbf{x}} \phi\left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}\right)\right)$, $\mathbf{y}_{t+1} = \Pi_{\mathcal{Y}}\left(\mathbf{y}_t + \eta \nabla_{\mathbf{y}} \phi\left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}\right)\right)$
- 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 $\widetilde{\mathbf{x}}_a$, then $\mathbf{x}_a = \Pi_{\mathcal{X}}(\widetilde{\mathbf{x}}_a)$.
    - We deal with $\nabla_{\mathbf{x}} \phi\left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}\right)^\top \left(\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}\right)$ for arbitary $\mathbf{x}$, 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
      - $\nabla_{\mathbf{x}} \phi\left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}\right)^\top \left(\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}\right) = \nabla_{\mathbf{x}} \phi\left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}\right)^\top (\mathbf{x}_{t+1} - \mathbf{x}) + \nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t)^\top \left(\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1}\right) + \left(\nabla_{\mathbf{x}} \phi\left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}\right) - \nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t)\right)^\top \left(\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1}\right)$
    - We can bound each term,
      - $\nabla_{\mathbf{x}} \phi\left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}\right)^\top (\mathbf{x}_{t+1} - \mathbf{x}) = \frac{1}{\eta}(\mathbf{x}_t - \widetilde{\mathbf{x}}_{t+1})^\top (\mathbf{x}_{t+1} - \mathbf{x}) \leq \frac{1}{\eta}(\mathbf{x}_t - \mathbf{x}_{t+1})^\top (\mathbf{x}_{t+1} - \mathbf{x}) = \frac{1}{2\eta}\left[\|\mathbf{x} - \mathbf{x}_t\|^2 - \|\mathbf{x} - \mathbf{x}_{t+1}\|^2 - \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2\right]$ by projection Fact 4.1
      - $\nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t)^\top \left(\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1}\right) = \frac{1}{\eta}\left(\mathbf{x}_t - \widetilde{\mathbf{x}}_{t+\frac{1}{2}}\right)^\top \left(\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1}\right) \leq \frac{1}{\eta}\left(\mathbf{x}_t - \mathbf{x}_{t+\frac{1}{2}}\right)^\top \left(\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1}\right) = \frac{1}{2\eta}\left[\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 - \|\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1}\|^2 - \|\mathbf{x}_t - \mathbf{x}_{t+\frac{1}{2}}\|^2\right]$ by same reason,
      - $\left(\nabla_{\mathbf{x}} \phi\left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}\right) - \nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t)\right)^\top \left(\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1}\right) \leq \|\nabla_{\mathbf{x}} \phi\left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}\right) - \nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t)\| \|\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1}\| \leq L\left[\|\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_t\| + \|\mathbf{y}_{t+\frac{1}{2}} - \mathbf{y}_t\|\right] \|\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1}\|$ $\leq \frac{L}{2}\|\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_t\|^2 + \frac{L}{2}\|\mathbf{y}_{t+\frac{1}{2}} - \mathbf{y}_t\|^2 + L\|\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1}\|^2$ by smoothness and Young's equality
    - Add them togethoer and we get $\nabla_{\mathbf{x}} \phi\left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}\right)^\top \left(\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}\right) \leq \frac{2}{\eta}\left[\|\mathbf{x}_t - \mathbf{x}\|^2 - \|\mathbf{x} - \mathbf{x}_{t+1}\|^2\right] + \left(L - \frac{1}{2\eta}\right)\|\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1}\|^2 +$ $\left(\frac{L}{2} - \frac{1}{2\eta}\right)\|\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_t\|^2 + \frac{L}{2}\|\mathbf{y}_{t+\frac{1}{2}} - \mathbf{y}_t\|^2$
    - Similarly $-\nabla_{\mathbf{y}} \phi\left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}\right)^\top \left(\mathbf{y}_{t+\frac{1}{2}} - \mathbf{y}\right) \leq \frac{2}{\eta}\left[\|\mathbf{y}_t - \mathbf{y}\|^2 - \|\mathbf{y} - \mathbf{y}_{t+1}\|^2\right] + \left(L - \frac{1}{2\eta}\right)\|\mathbf{y}_{t+\frac{1}{2}} - \mathbf{y}_{t+1}\|^2 + \left(\frac{L}{2} - \frac{1}{2\eta}\right)\|\mathbf{y}_{t+\frac{1}{2}} - \mathbf{y}_t\|^2 + \frac{L}{2}\|\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_t\|^2$
    - Then $\nabla_{\mathbf{x}} \phi\left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}\right)^\top \left(\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}\right) - \nabla_{\mathbf{y}} \phi\left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}\right)^\top \left(\mathbf{y}_{t+\frac{1}{2}} - \mathbf{y}\right) \leq \frac{2}{\eta}\left[\|\mathbf{x}_t - \mathbf{x}\|^2 - \|\mathbf{x} - \mathbf{x}_{t+1}\|^2\right] + \frac{2}{\eta}\left[\|\mathbf{y}_t - \mathbf{y}\|^2 - \|\mathbf{y} - \mathbf{y}_{t+1}\|^2\right] +$ $\left(L - \frac{1}{2\eta}\right)\left[\|\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_{t+1}\|^2 + \|\mathbf{x}_{t+\frac{1}{2}} - \mathbf{x}_t\|^2 + \|\mathbf{y}_{t+\frac{L}{2}} - \mathbf{y}_{t+1}\|^2 + \|\mathbf{y}_{t+\frac{1}{2}} - \mathbf{y}_t\|^2\right]$
      - choosing $L \leq 1/2\eta$ we can ommit 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 $\phi\left(\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t, \mathbf{y}\right) - \phi\left(\mathbf{x}, \frac{1}{T} \sum_{t=1}^T \mathbf{y}_t\right) \leq \frac{1}{T} \sum_{t=1}^T \phi\left(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}\right) - \phi\left(\mathbf{x}, \mathbf{y}_{t+\frac{1}{2}}\right)$

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

- 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, $\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 + \|\mathbf{y}_{t+1} - \mathbf{y}^*\|^2 \leq \left(1 - \frac{\mu}{4L}\right)\left\{\|\mathbf{x}_t - \mathbf{x}^*\|^2 + \|\mathbf{y}_t - \mathbf{y}^*\|^2\right\}$
  - Remark This $O\left(\kappa \log \frac{1}{\epsilon}\right)$ 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 $\omega_{t+1} = \omega_t - 2\eta F(\omega_t) + \eta F(\omega_{t-1}) = \omega_t - \eta\left(F(\omega_t) - (F(\omega_{t-1}) - F(\omega_t))\right)$, 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)

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

Concave Games

- Definition
  - 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 $\mathbb{R}^{d_i}$ and $\mathbf{x}_i \in \mathcal{X}_i$ is the action of player i,
  - Payoff (or utility) function $u_i : \mathcal{X} \to \mathbb{R}$, (which player i want to maximize)
- Assumption $u_i\left(\mathbf{x}_i, \mathbf{x} - i\right)$ is concave in $\mathbf{x}_i$
- Definition of Nash equilibrium The action profile $x^* \in \mathcal{X}$that is resilient to unilateral derivations, which means $u_i\left(\mathbf{x}_i^*, \mathbf{x}_{-i}^*\right) \geq u_i\left(\mathbf{x}_i, \mathbf{x}_{-i}^*\right), \forall \mathbf{x}_i \in \mathcal{X}_i, i \in \mathcal{N}$.

- Existence theorem [Debreu, 1952] every concave game admits a Nash equilibrium.
- First-order characterization $\left\langle\nabla_i u_i\left(\mathbf{x}_i^*, \mathbf{x}_{-i}^*\right), \mathbf{x}_i-\mathbf{x}_i^*\right\rangle \leq 0,\ \forall \mathbf{x}_i \in \mathcal{X}_i$

Variational Inequalites

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

VI with monotone operator

- Definition $F: \mathcal{Z} \rightarrow \mathbb{R}^d$ is
  - *monotone* if $\langle F(\mathbf{u})-F(\mathbf{v}), \mathbf{u}-\mathbf{v}\rangle \geq 0,\ \forall \mathbf{u}, \mathbf{v} \in \mathcal{Z}$,
  - *μ-strongly-monotone* if $\langle F(\mathbf{u})-F(\mathbf{v}), \mathbf{u}-\mathbf{v}\rangle \geq \mu\|\mathbf{u}-\mathbf{v}\|^2,\ \forall \mathbf{u}, \mathbf{v} \in \mathcal{Z}$
- Definition
  - The *(strong) solution (of Stampacchia VI)* is to find $\mathbf{z}^* \in \mathcal{Z}$ s.t. $\left\langle F\left(\mathbf{z}^*\right), \mathbf{z}-\mathbf{z}^*\right\rangle \geq 0 \forall \mathbf{z} \in \mathcal{Z}$
  - The *Weak solution (of Minty VI)* is to find $\mathbf{z}^* \in \mathcal{Z}$ s.t. $\left\langle F(\mathbf{z}), \mathbf{z}-\mathbf{z}^*\right\rangle \geq 0 \forall \mathbf{z} \in \mathcal{Z}$.
- If F monotone, then strong solution is also a weak solution since
  - $\left\langle F(\mathbf{z})-F(\mathbf{z}^*), \mathbf{z}-\mathbf{z}^*\right\rangle \geq 0$, then $\left\langle F(\mathbf{z}), \mathbf{z}-\mathbf{z}^*\right\rangle=\left\langle F\left(\mathbf{z}^*\right), \mathbf{z}-\mathbf{z}^*\right\rangle+\left\langle F(\mathbf{z})-F(\mathbf{z}^*), \mathbf{z}-\mathbf{z}^*\right\rangle \geq 0$
- If 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 , then this is provable.
- We use $\epsilon_{\mathrm{VI}}(\hat{\mathbf{z}}):=\max_{\mathbf{z} \in \mathcal{Z}}\langle F(\mathbf{z}), \hat{\mathbf{z}}-\mathbf{z}\rangle$ to measure the inaccuracy of a solution $\hat{\mathbf{z}}$.
- VI is a generality over a wide range of problems
  - *Convex minimization* $F=\nabla f$, with f convex
  - *Min-Max* $F=\left(\nabla_{\mathbf{x}} \phi,-\nabla_{\mathbf{y}} \phi\right)$, solution exists only when saddle point exists.
  - *Concave or monotone games* $F(\mathbf{z})=\left(-\nabla_i u_i\left(\mathbf{z}_i, \mathbf{z}_{-i}\right)\right)_{i \in \mathcal{N}}$
- Some possible assumptions
  - $\mathcal{Z}$ is a closed convex subset of $\mathbb{R}^d$
  - Solution of VI exists
  - Mapping F is monotone
  - F also Lipschitz continuous, $\|F(\mathbf{u})-F(\mathbf{v})\| \leq L\|\mathbf{u}-\mathbf{v}\|, \quad \forall \mathbf{u}, \mathbf{v} \in \mathcal{Z}$.
- Extragradient Algorithm first $\tilde{\mathbf{z}}_{t+1}=\Pi_{\mathcal{Z}}\left(\mathbf{z}_t-\eta_t F\left(\mathbf{z}_t\right)\right)$, then $\mathbf{z}_{t+1}=\Pi_{\mathcal{Z}}\left(\mathbf{z}_t-\eta_t F\left(\tilde{\mathbf{z}}_{t+1}\right)\right)$.
- Theorem 13.8 (Nemirovski, 2004) If F nibitibe and L-Lipschitz, and other previous assumption hold, then setting $\eta_t=\eta=\frac{1}{\sqrt{2} L}$ and EG gives $\max_{\mathbf{z} \in \mathcal{Z}}\left\langle F(\mathbf{z}), \frac{1}{T} \sum_{t=1}^T \tilde{\mathbf{z}}_t-\mathbf{z}\right\rangle \leq \frac{\sqrt{2} L D_{\mathcal{Z}}^2}{T}$, where $D_{\mathcal{Z}}=\max_{\mathbf{z}, \mathbf{z}^{\prime}}\|\mathbf{z}-\mathbf{z}^{\prime}\|_2$ is the 2-norm diameter of $\mathcal{Z}$.
- Algorithm for VI
  - *GDA* $\mathbf{z}_{t+1}=\mathbf{z}_t-\eta_t F\left(\mathbf{z}_t\right)$
  - *PPA* $\mathbf{Z}_{t+1}=\mathbf{Z}_t-\eta_t F\left(\mathbf{Z}_{t+1}\right)$

- *OGDA* $\mathbf{z}_{t+1} = \mathbf{z}_t - \eta_t \left(2F\left(\mathbf{z}_t\right) - F\left(\mathbf{z}_{t-1}\right)\right)$
- *Reflected Gradient* $\mathbf{z}_{t+1} = \mathbf{z}_t - \eta_t F\left(2\mathbf{z}_t - \mathbf{z}_{t-1}\right)$
- *Halpern iteration* $\mathbf{z}_{t+1} = \lambda_k \mathbf{z}_0 + (1 - \lambda_k)\left(\mathbf{z}_t - \eta_t F\left(\mathbf{z}_t\right)\right)$, use $\mathbf{z}_0$ as arching, this gives last iterate convergence.