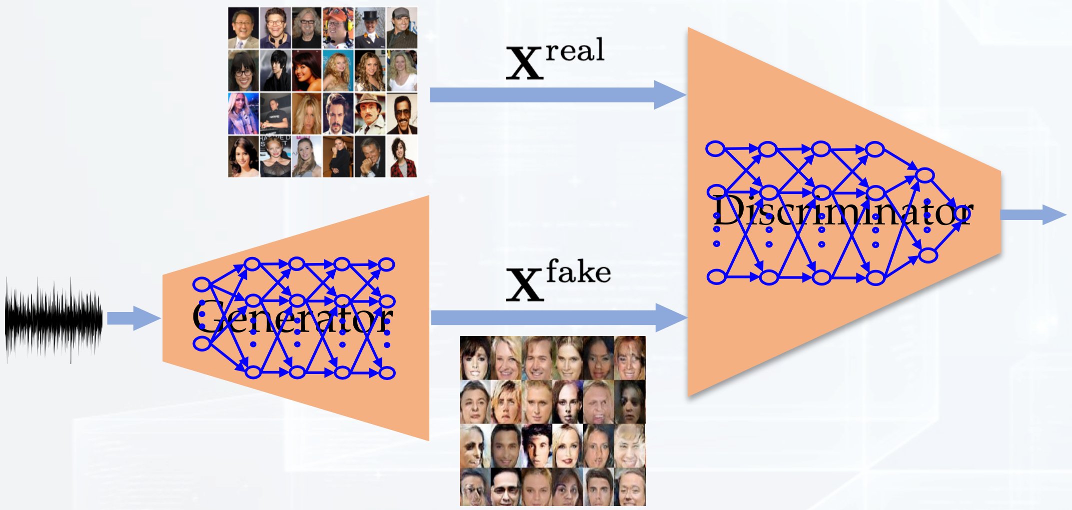
Train simultaneously, generalize better: Stability of gradient-based minimax learners

Farzan Farnia, Asu Ozdaglar

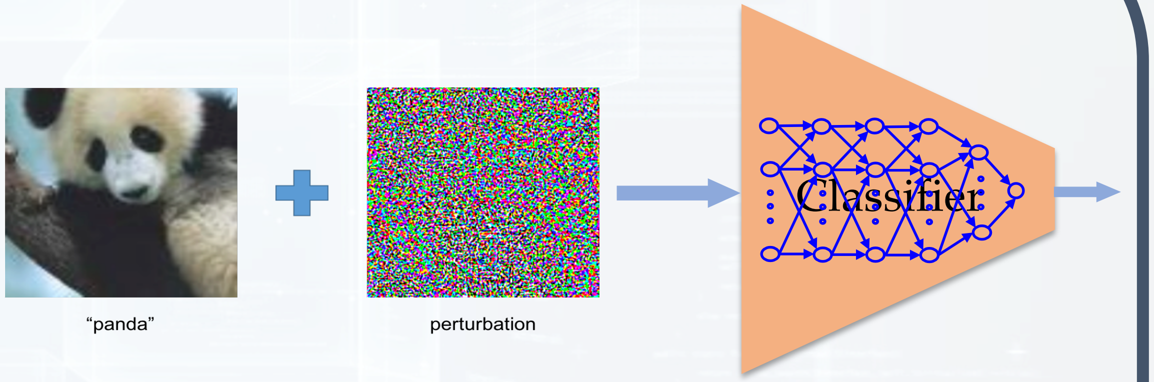
Massachusetts Institute of Technology

International Conference on Machine Learning, July 2021

Minimax Deep Learning Frameworks



Generative Adversarial Nets



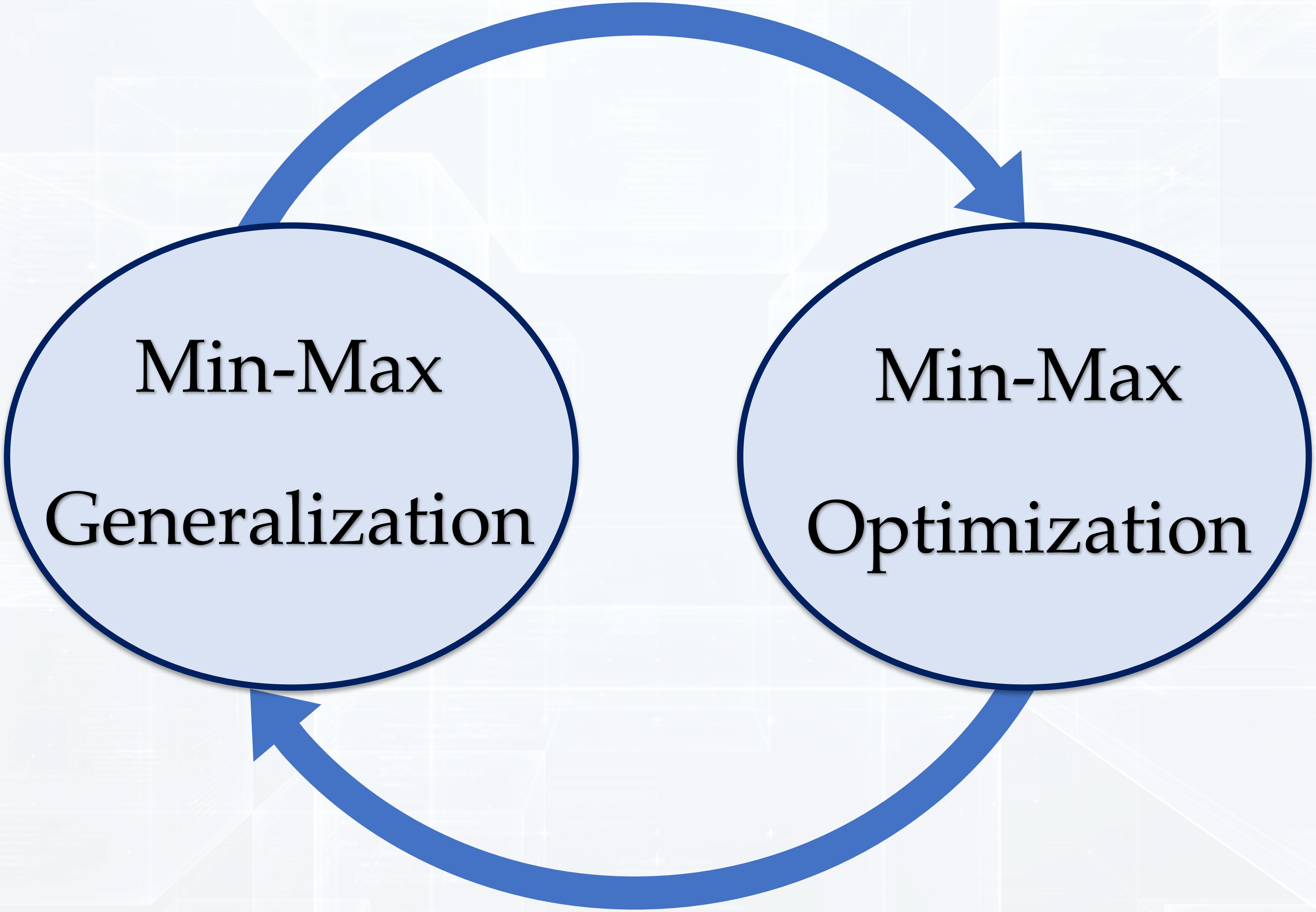
Adversarial Training

Minimax Deep Learning Frameworks

$$\min_{\mathbf{w}} \max_{\boldsymbol{\theta}} \boxed{\mathbb{E}_{Z \sim P_Z}\left[f(\mathbf{w}, \boldsymbol{\theta}; Z)\right]}$$

$$\approx \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{w}, \boldsymbol{\theta}; z_i)$$

Generalization Error


Min-Max
Generalization
Min-Max
Optimization

Gradient-based Min-Max Learners

GDA
$$\begin{cases} \mathbf{w}_{k+1} = \mathbf{w}_k - \eta_w \nabla_w \mathcal{L}(\mathbf{w}_k, \boldsymbol{\theta}_k) \\ \boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \eta_\theta \nabla_\theta \mathcal{L}(\mathbf{w}_k, \boldsymbol{\theta}_k) \end{cases}$$

Simultaneous Optimization Methods

GDmax
$$\begin{cases} \mathbf{w}_{k+1} = \mathbf{w}_k - \eta_w \nabla_w \mathcal{L}(\mathbf{w}_k, \boldsymbol{\theta}_k) \\ \boldsymbol{\theta}_{k+1} = \operatorname{argmax}_{\tilde{\boldsymbol{\theta}}} \mathcal{L}(\mathbf{w}_k, \tilde{\boldsymbol{\theta}}) \end{cases}$$

Non-Simultaneous Optimization Methods

Generalization Analysis in Convex-Concave Settings

- For convex-concave minimax objectives, our generalization bounds suggest a similar performance for simultaneous and non-simultaneous update algorithms.

Theorem: Consider an ℓ-smooth and L-Lipschitz minimax objective that is μ-strongly convex-concave in the min and max variables. Then, the expected minimax generalization risk of GDA and GDmax are bounded:

$$\epsilon_{\text{gen}}(\text{GDA}) \le \frac{2L^2(\ell/\mu + 1)}{(\mu - \frac{\ell^2 \eta_w}{2})n}, \quad \epsilon_{\text{gen}}(\text{GDmax}) \le \frac{2L^2(\ell/\mu + 1)}{\mu n}$$

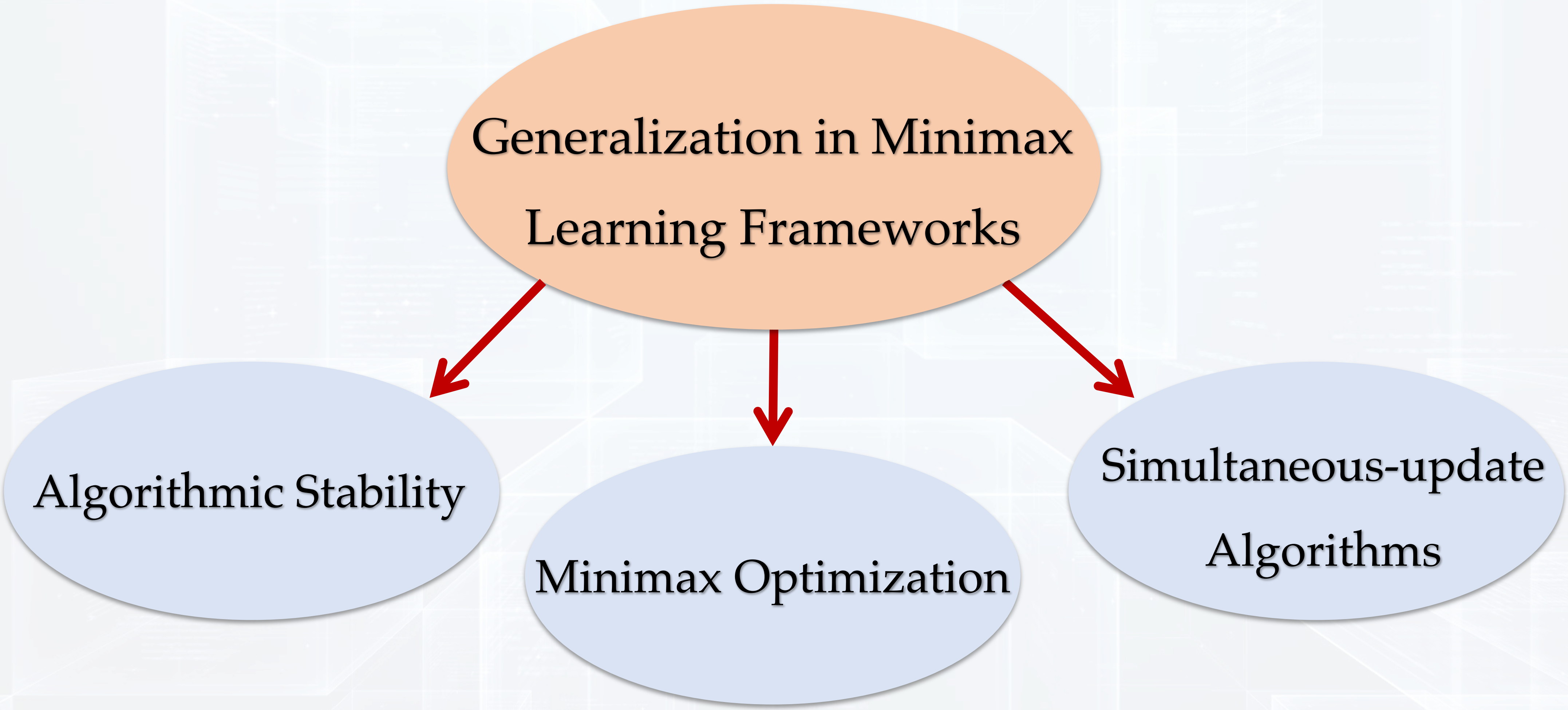
Generalization Analysis in Non-Convex-Concave Settings

- For general non-convex-concave minimax objectives, our generalization bounds indicate a different performance for simultaneous and non-simultaneous update algorithms.

Theorem: Consider an ℓ-smooth and L-Lipschitz objective that is μ-strongly concave in maximization variable. Under $\eta_{w,t} \leq c/t$, the minimax generalization risk for of GDA with stepsize ratio r and GDmax satisfy:

$$\epsilon_{\text{gen}}(\text{GDA}) \leq \mathcal{O}\left(T^{\frac{1}{1+1/(\ell r c)}}\right), \quad \epsilon_{\text{gen}}(\text{GDmax}) \leq \mathcal{O}\left(T^{\frac{1}{1+1/(\ell^2 c/\mu)}}\right)$$

Summary


Generalization in Minimax Learning Frameworks
Algorithmic Stability
Minimax Optimization
Simultaneous-update Algorithms