On Convergence of Gradient Descent Ascent: A Tight Local Analysis

Haochuan Li (MIT), Farzan Farnia (CUHK),

Subhro Das (IBM Research), Ali Jadbabaie (MIT)

Background

- Minimax Optimization:

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x};\boldsymbol{y})$$

  - Applications: GAN, Adversarial Training, RL, etc.
- Gradient Descent Ascent (GDA) Algorithm

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \eta_{\boldsymbol{x}} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^k;\boldsymbol{y}^k),$$
$$\boldsymbol{y}^{k+1} = \boldsymbol{y}^k + \eta_{\boldsymbol{y}} \nabla_{\boldsymbol{y}} f(\boldsymbol{x}^k;\boldsymbol{y}^k).$$

A gap between theory and practice on the ratio $r = \frac{\eta_{\boldsymbol{y}}}{\eta_{\boldsymbol{x}}}$

Empirical Observations

- Practical GAN training often chooses similar stepsizes for both players, i.e., $\eta_x \approx \eta_y$ and $r = \Theta(1)$

(a) (b)

Figure 1. Generated images of the learned generator on MNIST (a) and CIFAR10 (b). For both MNIST and CIFAR10, we train WGAN-GP models (Gulrajani et al., 2017) using simultaneous GDA with $\eta_x = \eta_y = 0.001$.

Existing Theory

Theorem ((Lin et al., 2020a), informal). Suppose f is L smooth and $f(\boldsymbol{x}, \cdot)$ is μ strongly concave for any fixed $\boldsymbol{x}$. Choosing $\eta_{\boldsymbol{x}} = \Theta(\frac{1}{\kappa^2 L})$ and $\eta_{\boldsymbol{y}} = \Theta(\frac{1}{L})$. The gradient complexity of GDA is bounded by $\mathcal{O}\left(\kappa^2/\epsilon^2\right)$ where $\kappa = L/\mu$ is the condition number and ϵ is the level of stationarity.

- Suggested stepsize ratio: $r = \Theta(\kappa^2)$

Motivation

- Gap between theory and practice:
  - Practice: $r = \Theta(1)$
  - Theory: $r = \Theta(\kappa^2)$
- Question:
  - *What is the best stepsize ratio that ensures convergence of GDA and what is the corresponding convergence rate?*
- This work:
  - A tight local analysis on convergence of GDA near a Stackelberg Equilibrium

Preliminaries

Definition (Stackelberg Equilibrium, informal) A point $\boldsymbol{z}^* = (\boldsymbol{x}^*, \boldsymbol{y}^*)$ is a differential Stackelberg Equilibrium if

- $\boldsymbol{z}^*$ is a stationary point of f;
- $f(\boldsymbol{x}^*, \cdot)$ is locally μ strongly concave;
- The primal function $\Phi(\cdot) = \max_{\boldsymbol{y} \in \mathcal{Y}} f(\cdot; \boldsymbol{y})$ is locally $\mu_{\boldsymbol{x}}$ strongly convex.

Assume f is L smooth and denote condition numbers

$$\kappa = L/\mu, \quad \kappa_{\boldsymbol{x}} = L/\mu_{\boldsymbol{x}}.$$

Theoretical Results

- Phase transition point:
  - If $r \leq \kappa$, there exist a hard function s.t. GDA diverges
  - If $r > \kappa$, GDA provably converges

- Optimal rate: (with matching upper/lower bounds)

$$\tilde{\mathcal{O}}(\kappa\kappa_{\boldsymbol{x}})$$

  - under stepsize choice $\eta_{\boldsymbol{x}} = \Theta(1/\kappa L), \ \ \eta_{\boldsymbol{y}} = \Theta(1/L), \ \ r = \Theta(\kappa)$
  - Same rate as running GD directly on the primal function

$$\Phi(\cdot) = \max_{\boldsymbol{y}\in\mathcal{Y}} f(\cdot;\boldsymbol{y})$$

- Extension to SGDA and EG

Takeaway

- A wide gap between theory and practice on the stepsize ratio of GDA
- A tight local analysis on GDA near a Stackelberg Equilibrium