L7: GDA with negative step size

Shugart, Henry, and Jason M. Altschuler. "Negative Stepsizes Make Gradient-Descent-Ascent Converge." *arXiv preprint arXiv:2505.01423* (2025).

- Toy problem

$$\min_{x \in \mathbb{R}} \max_{y \in \mathbb{R}} f(x,y) = xy \qquad (1)$$

GDA $\{\alpha_t, \beta_t\}_t$ $\quad t = 0, 1, 2, \dots$

$$\begin{cases} x_{t+1} = x_t - \alpha_t \nabla_x f(x_t, y_t) \\ y_{t+1} = y_t + \beta_t \nabla_y f(x_t, y_t) \end{cases}$$

For (1), $\begin{bmatrix} x_{t+1} \\ y_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & -\alpha_t \\ \beta_t & 1 \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix}$,

GDA fail to converge when any of the following

- $\alpha_t, \beta_t \geq 0$
- $\alpha_t = \beta_t$ (lemma 2.1)
- $\alpha_t = \alpha, \beta_t = \beta$

- Bi-linear problem

$$\min_{x \in \mathbb{R}^m} \max_{y \in \mathbb{R}^n} f(x,y) = x^T B y$$

$$m I \preceq B B^T, B^T B \preceq M I$$

(B has non-zero singular values)

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} I & -\alpha_1 B \\ \beta_1 B^T & I \end{bmatrix} \begin{bmatrix} I & -\alpha_0 B \\ \beta_0 B^T & I \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}$$

$$\boxed{\alpha_0 = \beta_1 = h, \quad \alpha_1 = \beta_0 = -h} \qquad (1.5)$$

$$\overbrace{\begin{bmatrix} I & hB \\ hB^T & I \end{bmatrix}}^{u_1} \overbrace{\begin{bmatrix} I & -hB \\ -hB^T & I \end{bmatrix}}^{u_2}$$

$$= \begin{bmatrix} I - h^2 BB^T & 0 \\ 0 & I - h^2 B^T B \end{bmatrix}$$

u_1, u_2 expansive, but $\| u_1 u_2 \| < 1$ when h small enough

$$\begin{bmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{bmatrix} \begin{bmatrix} 1 & -\varepsilon \\ -\varepsilon & 1 \end{bmatrix} = \begin{bmatrix} 1-\varepsilon^2 & \\ & 1-\varepsilon^2 \end{bmatrix}$$

Observation: for bilinear problem, two GDA steps with $\alpha_1 = \beta_0 = -h$ is equivalent to one GD step with step size h^2 on

$$\min_{x,y} \Phi(x,y) = \frac{1}{2} \| \nabla f(x,y) \|^2$$

pf. $\Phi(x,y) = \frac{1}{2}\left(x^T B B^T x + y^T B^T B y \right)$.

Then, GDA converges by using standard GD stepsize.

- Nonlinear case (convex - concave f)

$$\begin{cases} \text{w.p. } \frac{1}{2}: \ \alpha_0 = \beta_1 = h, \ \alpha_1 = 0, \ \beta_0 = -h & (1.6) \\ \text{w.p. } \frac{1}{2}: \ \alpha_1 = \beta_0 = h, \ \alpha_0 = -h, \ \beta_1 = 0 & \end{cases}$$

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} + \frac{h}{2} \underbrace{\begin{bmatrix} -\nabla_x f(x_0, y_0) \\ \nabla_y f(x_0, y_0) \end{bmatrix}}_{\text{GDA}} - \frac{h^2}{2} \overbrace{\nabla^2 f(x_0, y_0) \nabla f(x_0, y_0)} + O(h^3)$$

② is an h^2-GD step to minimise Hamiltonian

For bilinear case

$$\begin{bmatrix} I - h^2 BB^T & -\frac{h}{2} B \\ \frac{h}{2} B^T & I - h^2 B^T B \end{bmatrix}$$

- Rates

1. bilinear $O(\sqrt{\kappa} \log \frac{1}{\varepsilon})$ tight (info-theoretical optimal)

2. convex-concave $O(1/\varepsilon^2)$ lower-bound: $\Omega(1/\varepsilon)$
   $\|\nabla f\| \le \varepsilon$

3. SC-SC $O(\kappa \log \frac{1}{\varepsilon})$ tight

Optimality (bilinear, quadratic)

algo: assymmetric Krylov-subspace

$$x_t \in x_0 + \text{span}\{\nabla_x f(x_s, y_s), s < t\}$$

$$y_t \in y_0 + \text{span}\{\nabla_y f(x_s, y_s), s < t\}$$

called symmetric if

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} \in \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} + \text{span}\left\{ \begin{bmatrix} \nabla_x f(x_s, y_s) \\ \nabla_y f(x_s, y_s) \end{bmatrix}, s < t \right\}$$