AI709 Presentation:

Understanding Gradient Descent on Edge of Stability in Deep Learning

Sanjeev Arora, Zhiyuan Li, Abhishek Panigrahi

ICML 2022

Speaker: Hanseul Cho

Stableness

Definition 1.1

- Loss function $L : \mathbb{R}^D \to \mathbb{R}$, parameter $x \in \mathbb{R}^D$, learning rate (LR) $\eta > 0$.
- *Stableness*:

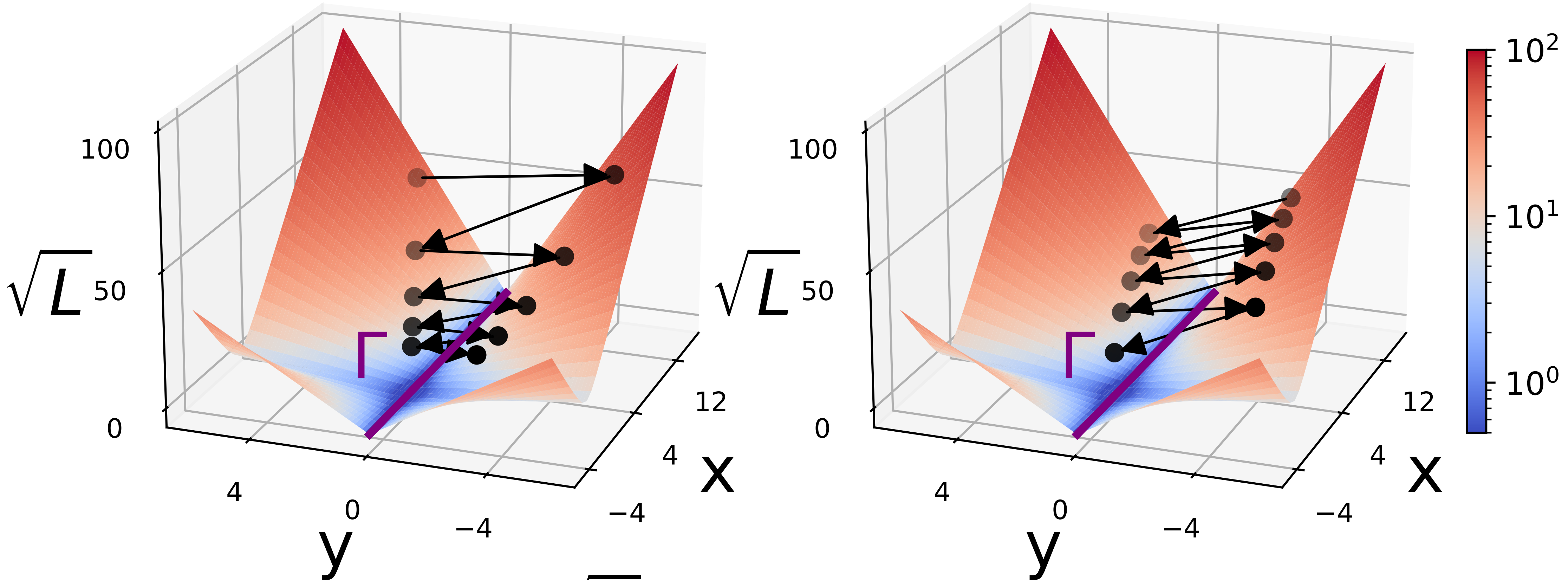
$$S_L(x, \eta) := \eta \cdot \sup_{s \in [0,\eta]} \lambda_1 \left(\nabla^2 L \left(x - s \nabla L(x) \right) \right)$$

  - LR×(supremum of sharpness at a point after a step of gradient descent (GD))
- L is *stable* at (x, η) iff $S_L(x, \eta) \leq 2$; otherwise, we say L is *unstable* at (x, η).
  - Note: L is $\left(\frac{S_L(x, \eta)}{\eta} \right)$-smooth on a line segment between x and $x - \eta \nabla L(x)$

Problem setting: Algorithms

1. Normalized GD on L: $x_{t+1} = x_t - \frac{\eta}{\|\nabla L(x_t)\|} \nabla L(x_t)$ +noise?

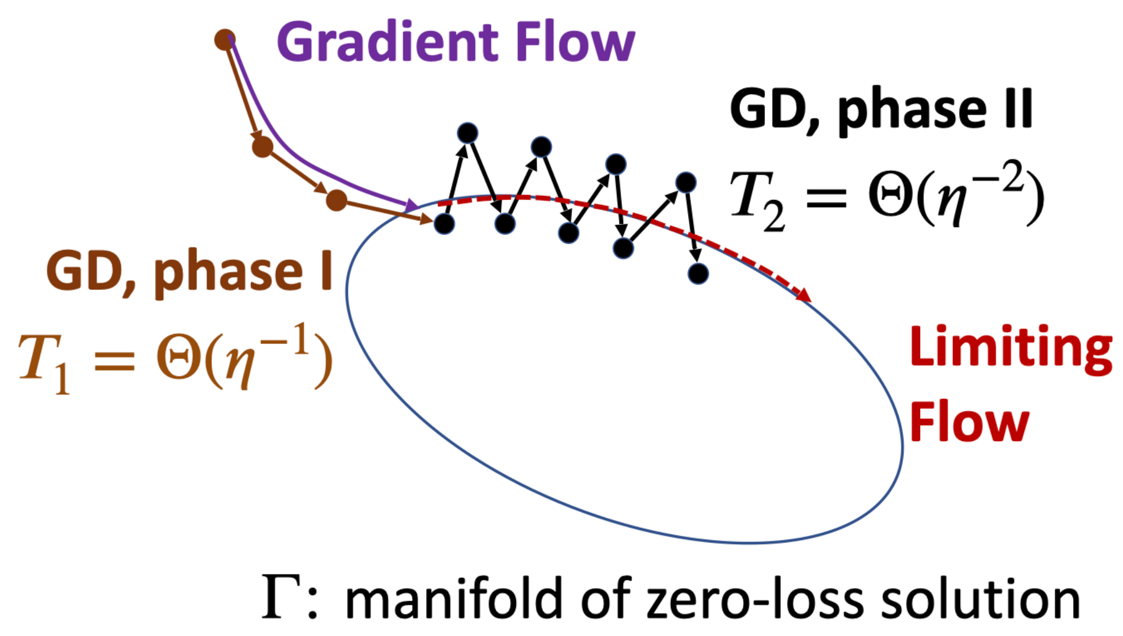
2. GD on $\sqrt{L} - \sqrt{L_{\min}}$: $x_{t+1} = x_t - \eta \nabla \sqrt{L}(x_t)$ +noise?



(a) GD on $\sqrt{L}$ (b) Normalized GD on L

Contribution

Two-phase dynamics of GD variants with small LR η



- Phase I
  - Starting from a neighborhood of the manifold Γ of the minimizers of the loss,
  - GD tracks a gradient flow (GF) governed by L (monotone decrease in L).
  - GD gets $\mathcal{O}(\eta)$-close to the manifold Γ.
- Phase 2
  - (slightly perturbed) GD tracks another flow on Γ which decreases the loss sharpness
  - Unstable: *stableness* at least in one step of every two consecutive steps is > 2
  - The loss non-monotonically decreases (proportionally to the loss sharpness)

Warm-up: Quadratic Loss

$L(x) = \frac{1}{2}x^\top Ax$ **where A is PSD**

- Normalized GD on L: $x_{t+1} = x_t - \frac{\eta}{\|Ax_t\|}Ax_t$
- GD on $\sqrt{L}$: $x_{t+1} = x_t - \frac{\eta}{\sqrt{2x_t^\top Ax_t}}Ax_t$
- If we set $\tilde{x}_t = \frac{1}{\eta}Ax_t$ for Normalized GD and $\tilde{x}_t = \frac{1}{\eta}(2A)^{1/2}x_t$ for GD on $\sqrt{L}$, both $\tilde{x}_t$'s satisfy the same update rule

$$\tilde{x}_{t+1} = \tilde{x}_t - A\frac{\tilde{x}_t}{\|\tilde{x}_t\|}.$$

Warm-up: Quadratic Loss

$\tilde{x}_t$ oscillates & aligns to $\pm v_1$

- Consider $A \in \mathbb{R}^{D\times D}$ with eigenvalues $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_D > 0$ and $v_1, \cdots, v_D$ are the corresponding eigenvectors.
- **Theorem 3.1.** If $\left|\langle v_1, \tilde{x}_t\rangle\right| \neq 0$ for $t \geq 0$, then $\exists C \in (0,1)$ and $\exists s \in \{\pm 1\}$ such that $\lim_{t\to\infty} \tilde{x}_{2t} = Cs\lambda_1 v_1$ and $\lim_{t\to\infty} \tilde{x}_{2t+1} = -(1-C)s\lambda_1 v_1$.
- The angle θ_t between $\tilde{x}_t$ and v_1 converges to 0 (“alignment”), while the direction of $\tilde{x}_t$ flips back and forth near the minima.

Key definitions (1)

Gradient flow (GF), its limiting map, & attraction set of Γ

- GF on L can be described through a mapping $\phi : \mathbb{R}^D \times [0,\infty) \to \mathbb{R}^D$ s.t.

$$\phi(x,t) = x - \int_0^t \nabla L(\phi(x,s))\mathrm{d}s$$

  - Satisfies $\phi(x,0) = x,\ \partial_t \phi(x,t) = -\nabla L(\phi(x,t))$

- The **limiting map** $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ of GF: $\Phi(x) = \lim_{t\to\infty} \phi(x,t)$

- Attraction set U of Γ: an open neighborhood of Γ s.t. for all $x \in U$, $\Phi(x) \in \Gamma$

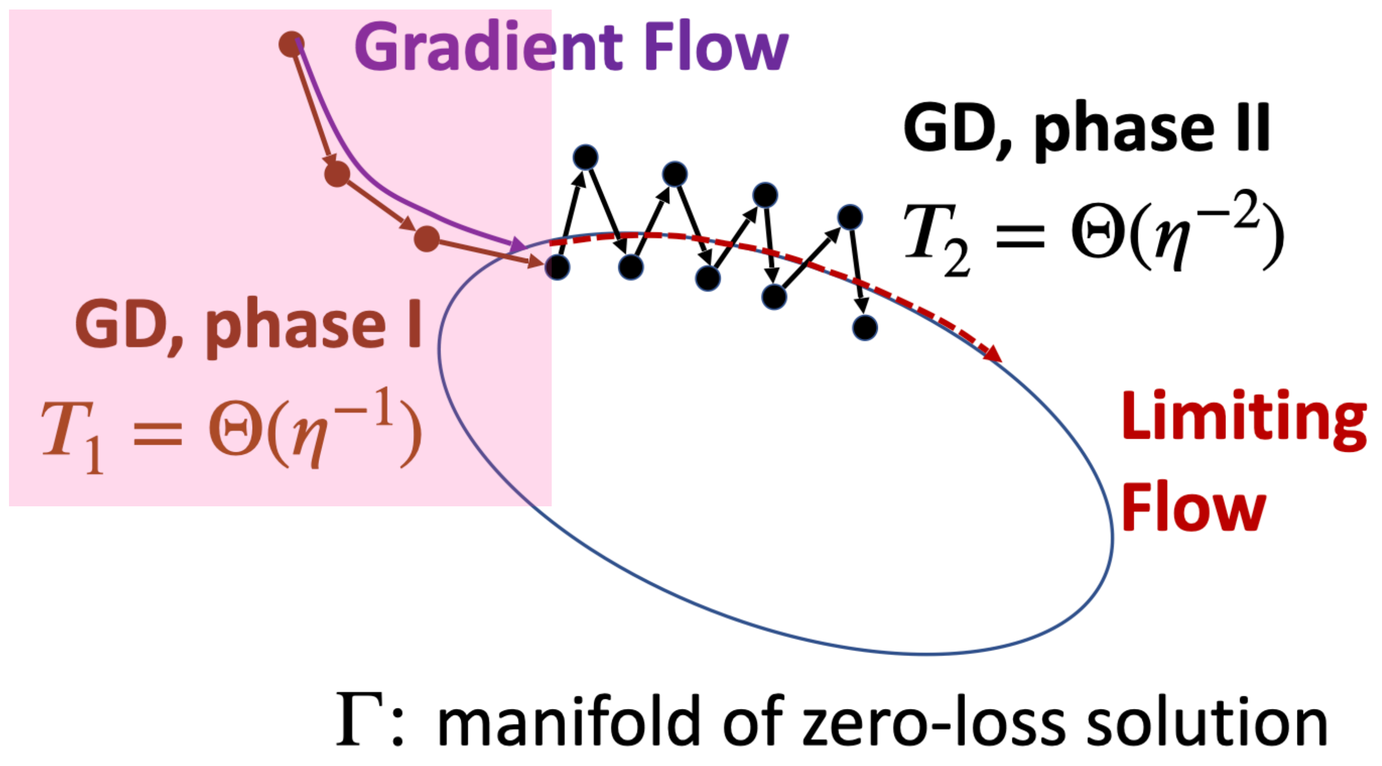
Key Definitions (2)

transformed iterate $\tilde{x}_t$ (motivated by quadratic case)

- $\tilde{x} = \begin{cases} \nabla^2 L(\Phi(x))(x - \Phi(x)) & \text{for Normalized GD on } L \\ \left(2\,\nabla^2 L(\Phi(x))\right)^{1/2}(x - \Phi(x)) & \text{for GD on } \sqrt{L} \end{cases}$
- $\theta_t \in \left[0, \frac{\pi}{2}\right]$: **angle** between $\tilde{x}_t$ & top eigenspace of $\nabla^2 L(\Phi(x_t))$
- $R_j(x) := \sqrt{\sum_{i=j}^{M} \langle v_i(x), \tilde{x}\rangle^2} - \lambda_j(x)\eta$, for $j \in [D]$
  - $M = \mathrm{rank}(\nabla^2 L(x))$ for all $x \in \Gamma$ (so that Γ is a $(D - M)$-dimensional manifold)
  - $\{(\lambda_i(x), v_i(x))\}_{i=1}^{D}$: eigenvalue-eigenvector pairs of $\nabla^2 L(\Phi(x))$ $(\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_D)$
  - the first square root term: length of the projection of $\tilde{x}$ onto the bottom-$(D - j)$ eigenspace of $\nabla^2 L(\Phi(x))$

Results for Normalized GD (1)

Phase I



- **Theorem 4.3.** Let $x_0 = x_{\text{init}} \in U$. Then, there is a constant $T_1 > 0$ such that for any $T_1' > T_1$ and a sufficiently small LR $\eta > 0$, the following holds:

(1) $$\max_{t \in [T_1/\eta, T_1'/\eta]} \left\| x_t - \Phi(x_{\text{init}}) \right\| \leq \mathcal{O}(\eta)$$

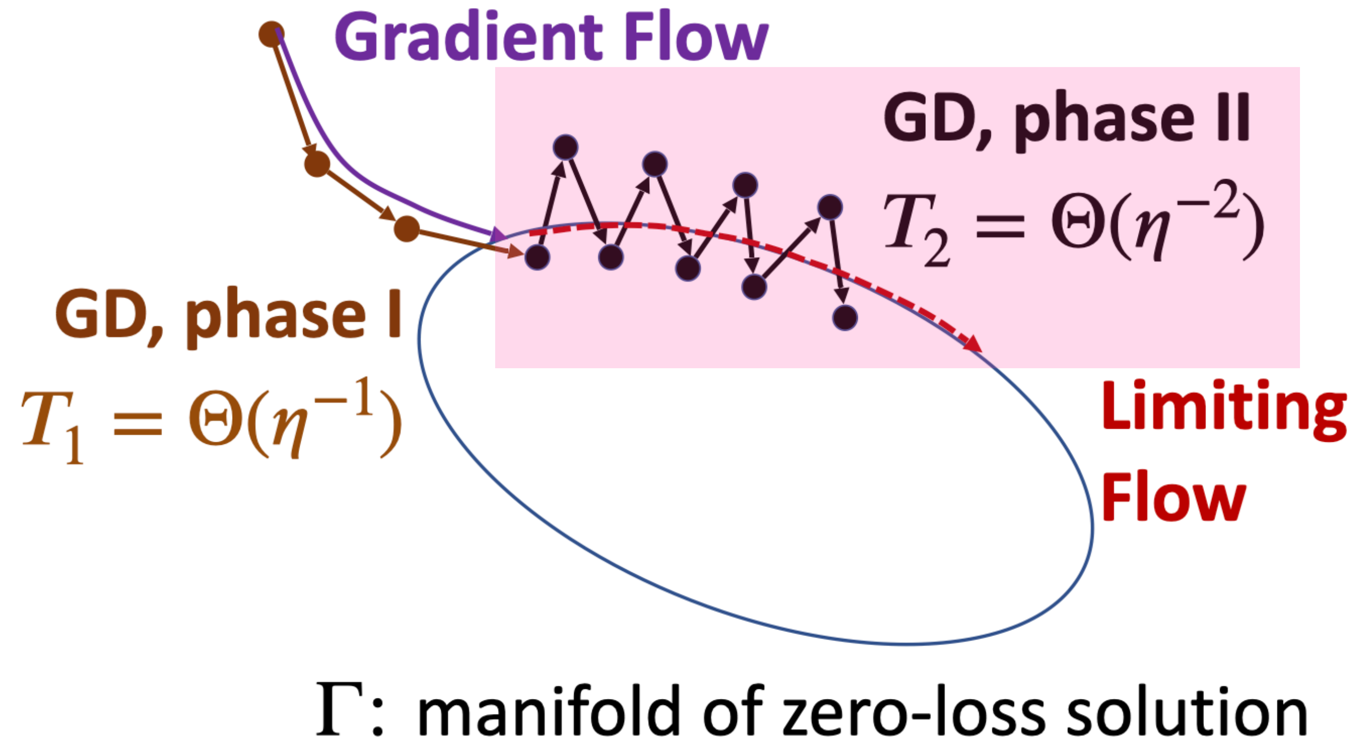
- (iterates track the GF & get $\mathcal{O}(\eta)$-close to the minimizer manifold Γ)

(2) $$\max_{t \in [T_1/\eta, T_1'/\eta], j \in [D]} R_j(x_t) \leq \mathcal{O}(\eta^2)$$

- (projected length of $\tilde{x}_t$ onto eigenspace of $\nabla^2 L(\Phi(x_t))$ is not too large)

Results for Normalized GD (2)

Phase II



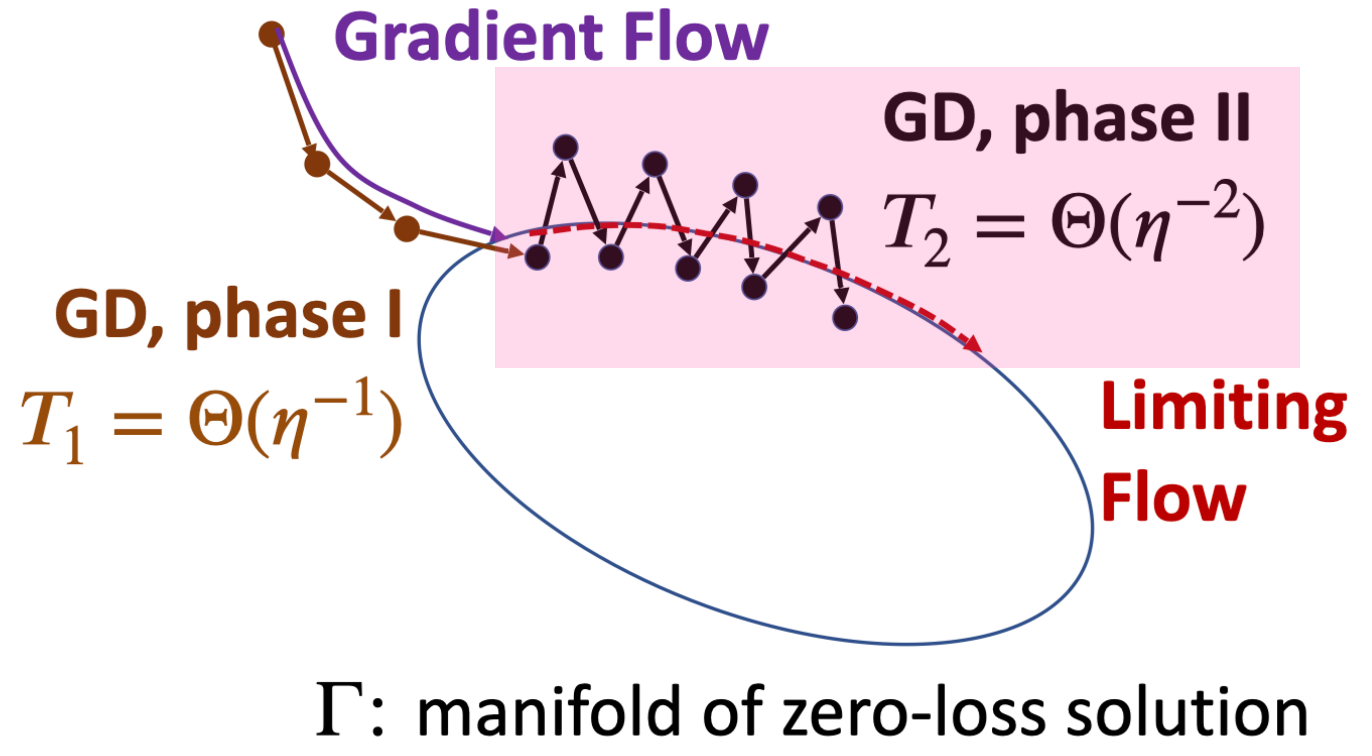
- Restart the algorithm from the end of Phase I: $x_0 = x_t^{\text{Phase I}}$ $(t \geq T_1/\eta)$
- Assume that $\left\| x_0 - \Phi(x_{\text{init}}) \right\| \leq \mathcal{O}(\eta)$ and $\max_{j \in [D]} R_j(x_0) \leq \mathcal{O}(\eta^2)$ hold for x_0.
- + Assume that the initial alignment of $\tilde{x}_0$ and $v_1(x_0)$ is not too small. (formal description is omitted)
- x_t will eventually track the following Riemannian gradient flow on Γ:

Limiting Flow: $$X(\tau) = \Phi(x_{\text{init}}) - \frac{1}{4}\int_0^\tau P^{\perp}_{X(s),\Gamma} \nabla \log \lambda_1(X(s))\mathrm{d}s, \quad X(\tau) \in \Gamma$$

  - $P^{\perp}_{x,\Gamma} : \Gamma \to \mathbb{R}^D$: projection operator onto the tangent space of Γ at x
  - The sharpness $\lambda_1(X(\tau))$ decreases!

Results for Normalized GD (3)

Phase II



- To make the theoretical analysis feasible, the alignment between $\tilde{x}_t$ and $v_1(x_t)$ should not vanish.

- To this end, we add a (uniform) noise of magnitude $\mathcal{O}(\eta^{100})$ occasionally.

- **Theorem 4.4.** For any constant time $T_2 > 0$ till which the solution of the "limiting flow" X exists, for sufficiently small $\eta > 0$, with probability at least $1 - \mathcal{O}(\eta^{10})$, the iterates of <u>*perturbed*</u> Normalized GD satisfies that

(1) $$\left\| \Phi\left(x_{\lfloor T_2/\eta^2 \rfloor}\right) - X(T_2) \right\| = \mathcal{O}(\eta),$$ (tracking the limiting flow)

(2) $$\frac{1}{\lfloor T_2/\eta^2 \rfloor} \sum_{t=0}^{\lfloor T_2/\eta^2 \rfloor} \theta_t \leq \mathcal{O}(\eta)$$ (alignment in average)

Results for Normalized GD (4)

Phase II → Edge of Stability: High stableness, non-monotonic decrease of loss

- **Theorem 4.7.** Under the setting of Phase II, by viewing Normalized GD as GD with time-varying LR $\eta_t = \frac{\eta}{\|\nabla L(x_t)\|}$, we have

(1) $$\frac{1}{S_L(x_t, \eta_t)} + \frac{1}{S_L(x_{t+1}, \eta_{t+1})} = 1 + \mathcal{O}(\theta_t + \eta)$$

  - Stableness $\gtrsim 2$ at least in one of every two consecutive steps.

(2) $$\sqrt{L(x_t)} + \sqrt{L(x_{t+1})} = \eta \sqrt{\frac{\lambda_1(\nabla^2 L(x_t))}{2}} + \mathcal{O}(\eta\theta_t)$$

  - Loss (non-monotonically) decreases as the loss sharpness decreases via limiting flow.

Results for GD on $\sqrt{L}$

Phase II → Edge of Stability: High stableness, non-monotonic decrease of loss

- **Theorem 4.8.** Under the setting of Phase II, Running GD on $\sqrt{L}$, we eventually have

(1) $S_L(x_t, \eta_t) \geq \Omega\left(\frac{1}{\theta_t}\right)$

  - Stableness is large.

(2) $\sqrt{L(x_t)} + \sqrt{L(x_{t+1})} = \eta\lambda_1(\nabla^2 L(x_t)) + \mathcal{O}(\eta\theta_t)$

  - Loss (non-monotonically) decreases as the loss sharpness decreases via limiting flow.

Discussion

- Different setting from Cohen et al. [2021]
  - Discrepancy in algorithms.
  - The sharpness should decrease to near zero to ensure the convergence in loss (↔ the sharpness hovers around $2/\eta$ [Cohen et al., 2021])
  - Although the analysis allows some non-smoothness in loss ($\sqrt{L}$ case), the manifold Γ of minimizers must be smooth enough ("C^2-submanifold of $\mathbb{R}^d$")
- Locality of the analysis
  - The analysis only applies when the initialization is close enough to Γ
- Non-vanishing but small learning rate η