Understanding the Learning Dynamics of LoRA

A Gradient Flow Perspective in Matrix Factorization

Enrique Mallada

Informs Optimization Society Conference

March 22nd, 2026

Acknowledgements

Ziqing Xu

Hancheng Min

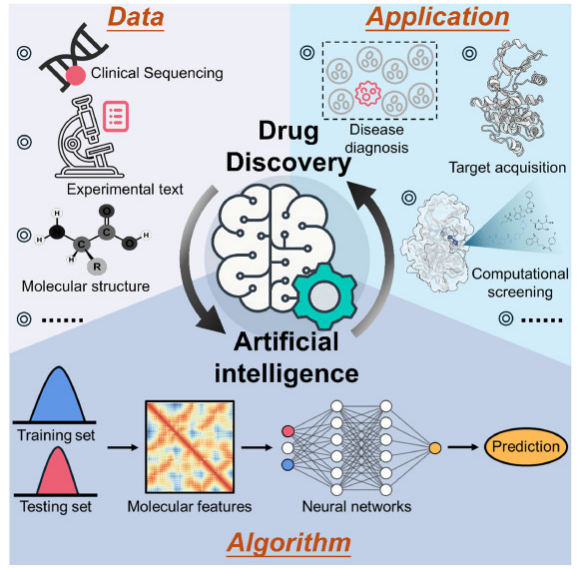
Salma Tarmoun

Jinqi Luo

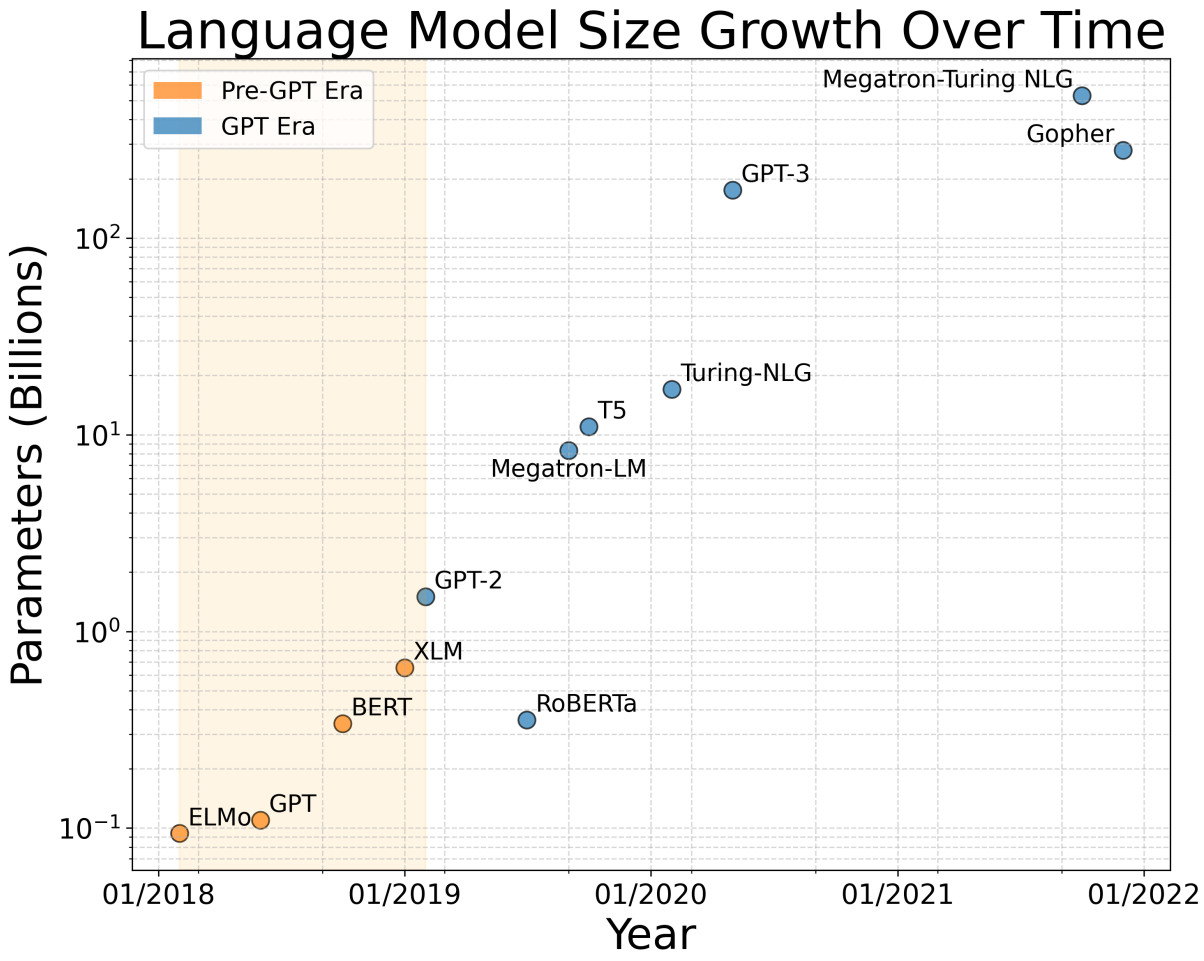
Lachlan Ewen MacDonald

Rene Vidal

Deep Learning's Transformative Impact

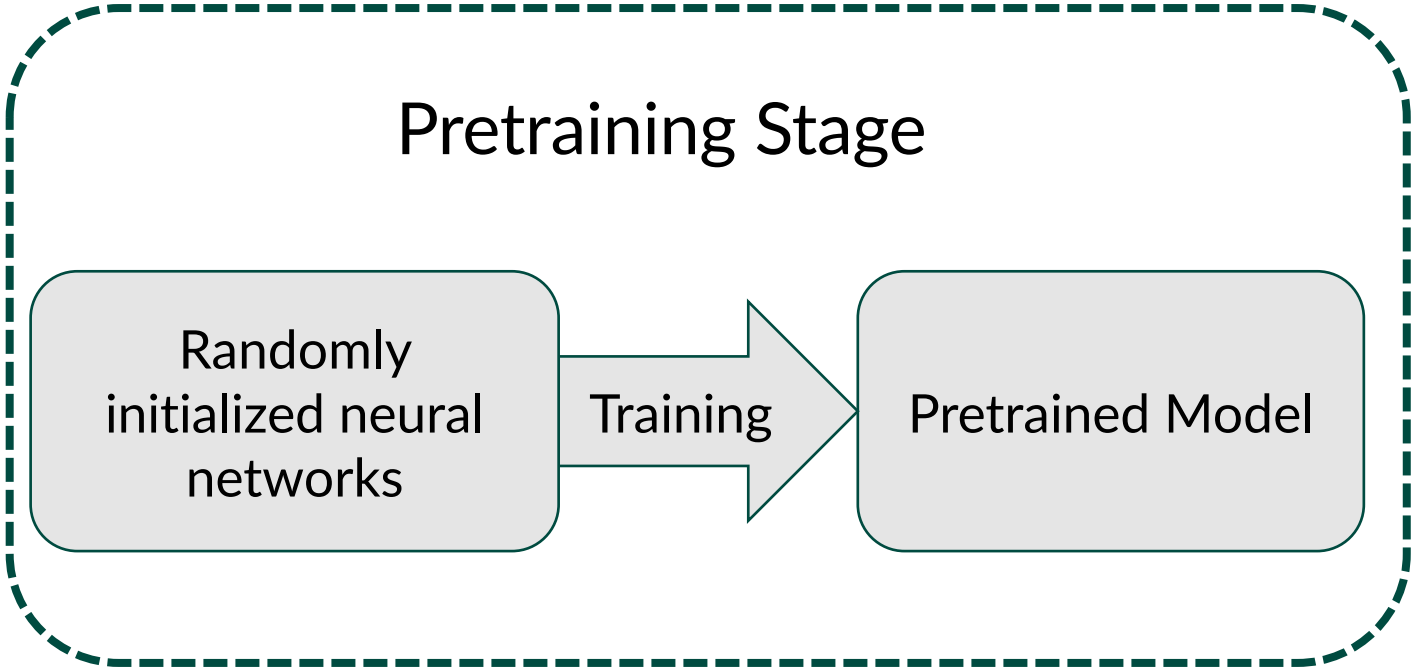


Through the power of scaling



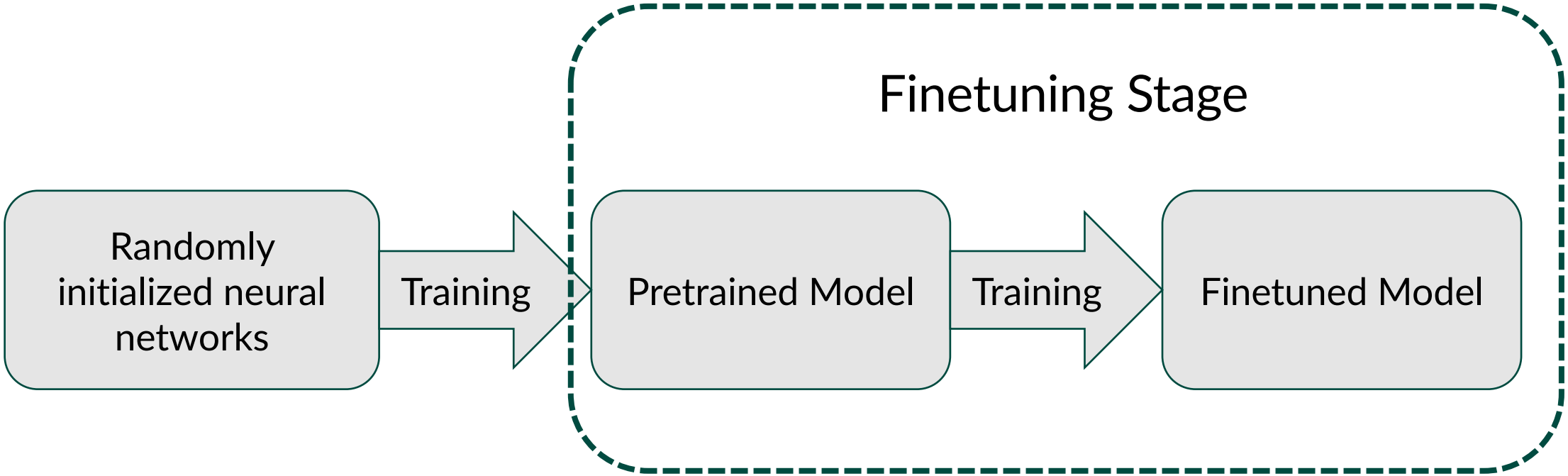
Pretraining & Fine-Tuning: The Modern Approach

- Starting from a massive pretrained model, then fine-tuning for downstream tasks, is now standard in deep learning



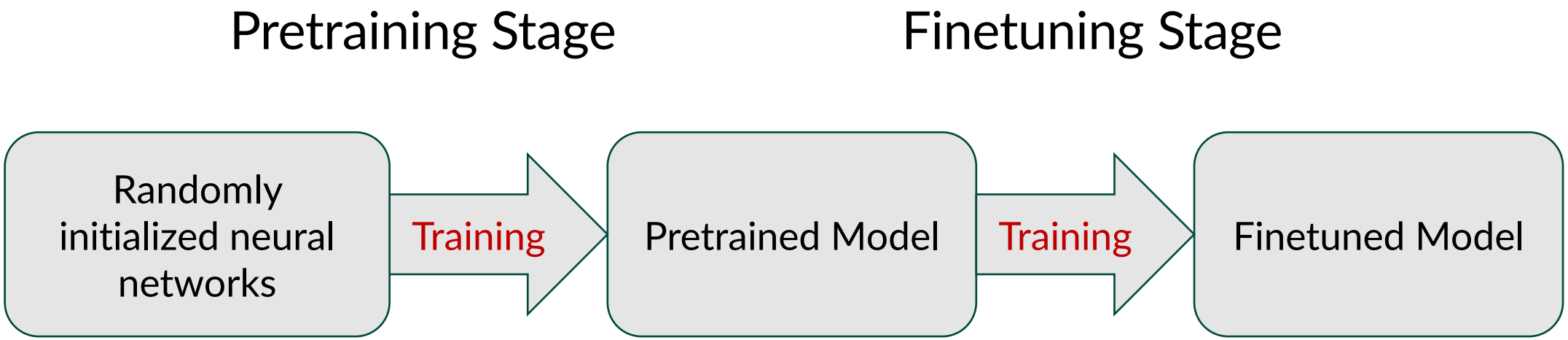
Pretraining & Fine-Tuning: The Modern Approach

- Starting from a massive pretrained model, then fine-tuning for downstream tasks, is now standard in deep learning

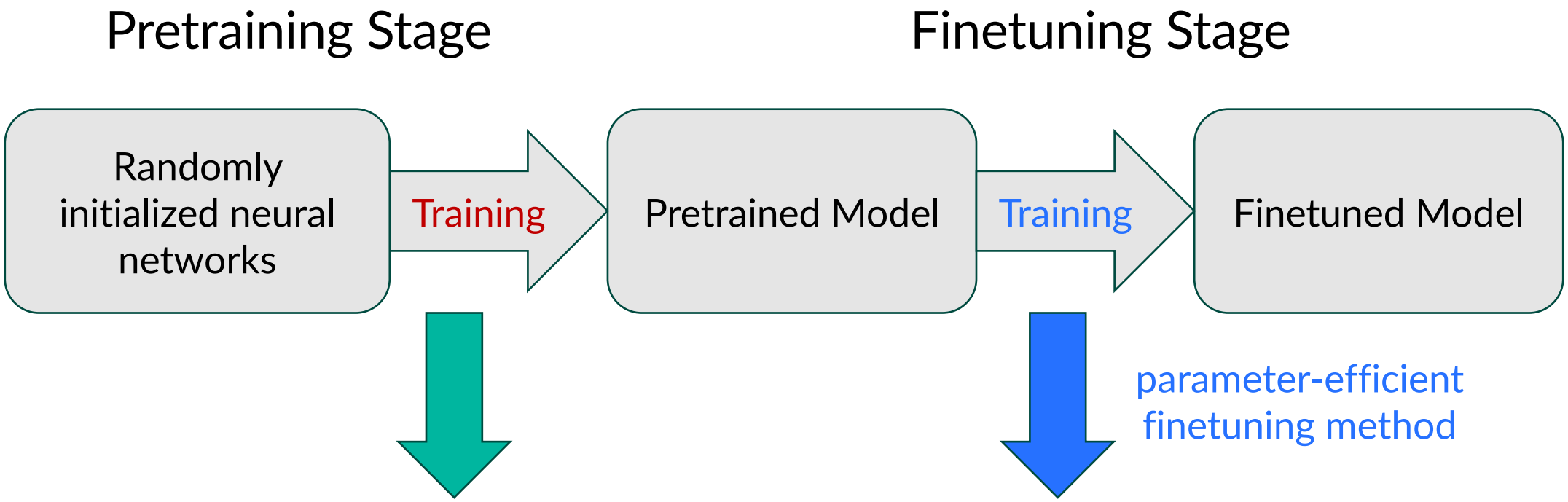


Pretraining & Fine-Tuning: The Modern Approach

- Starting from a massive pretrained model, then fine-tuning for downstream tasks, is now standard in deep learning

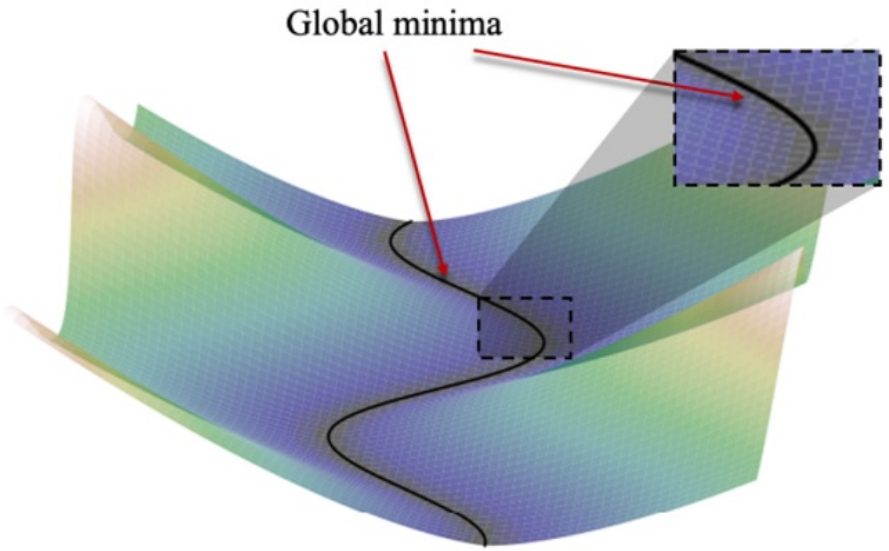


Comparing Training in Pretraining and Finetuning Stage

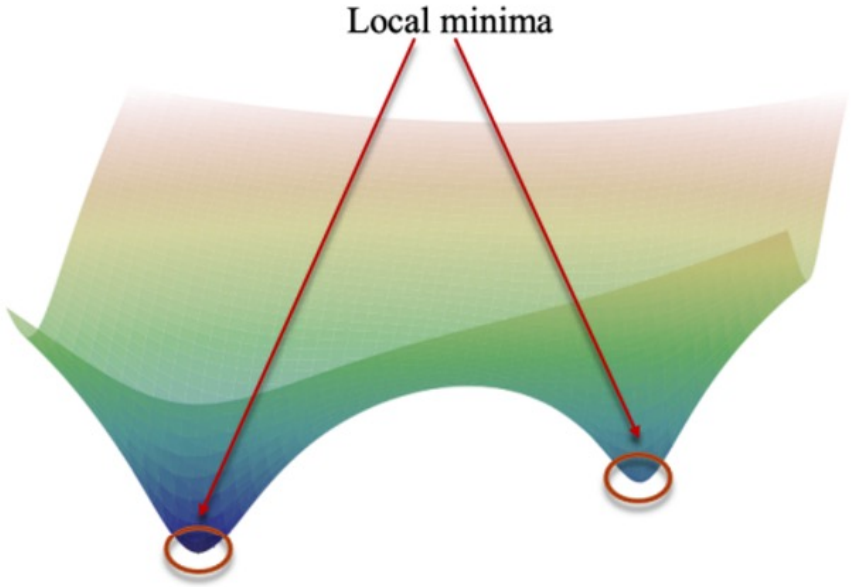


overparametrized models versus non-overparametrized models

Loss Landscape of Pretraining and Finetuning



Loss landscape of overparametrized model

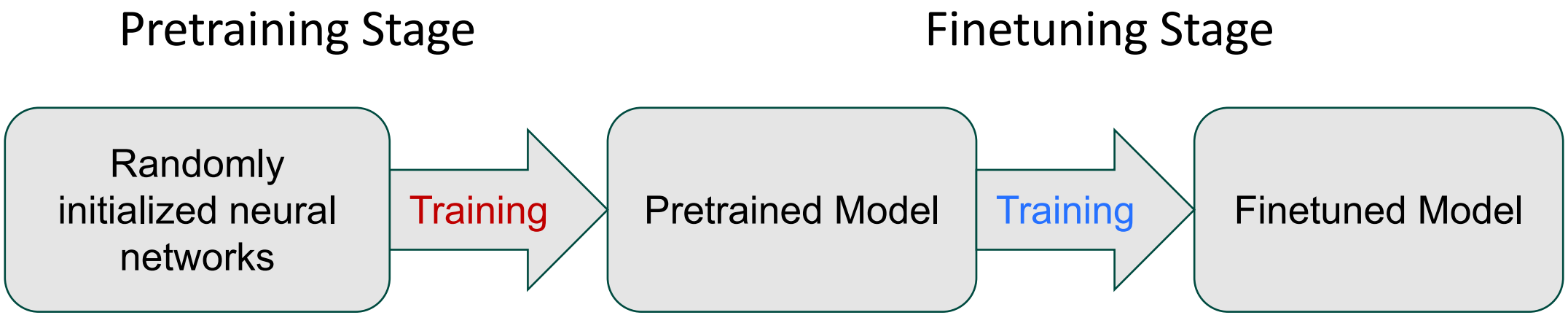


Loss landscape of non-overparametrized model

Figure from Liu et al., 2022

Loss Landscape of overparametrized and non-overparametrized model are different

Research Goal and Questions



Goal: understand optimization process of gradient-based method for both pretraining and finetuning stage.

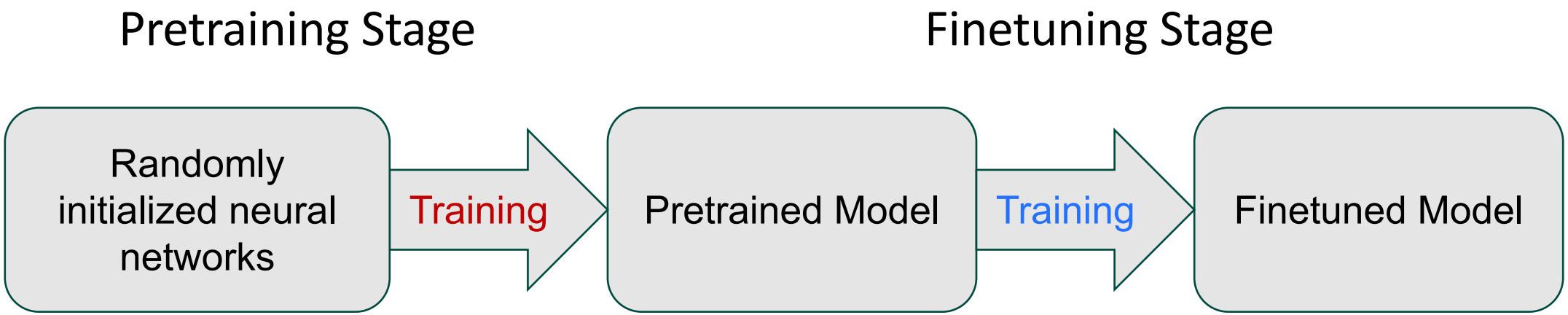
- **Q1:** How overparametrized models converge during pretraining? [1], [2]
- **Q2:** How pretrained models adapt quickly and effectively during fine-tuning? [3]

[1] Xu, Min, Tarmoun, M, Vidal, Linear Convergence of Gradient Descent for Finite-Width Over-Parametrized Linear Networks with General Initialization, AISTATS 2023

[2] Xu, Min, Tarmoun, M, Vidal, A Local Polyak–Łojasiewicz and Descent Lemma of Gradient Descent for Overparameterized Linear Models, TMLR 2025

[3] Xu, Min, MacDonald, Luo, Tarmoun, M, Vidal, Understanding the Learning Dynamics of LoRA: A Gradient Flow Perspective on Low-Rank Adaptation in Matrix Factorization, AISTATS 2025

Research Goal and Questions



Goal: understand optimization process of gradient-based method for both pretraining and finetuning stage.

- **Q1:** How overparametrized models converge during pretraining? [1], [2]
- **Q2:** How pretrained models adapt quickly and effectively during fine-tuning? [3]

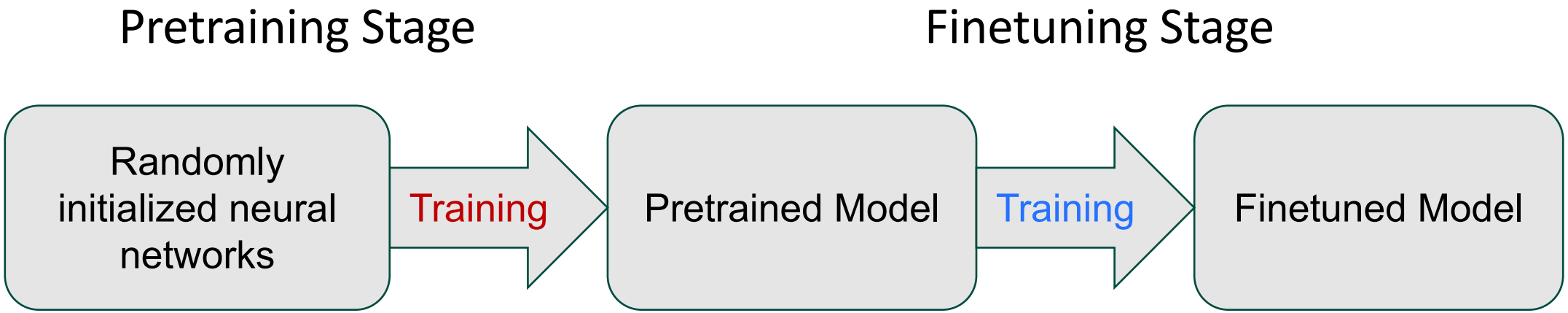
Case study: gradient flow w/ two-layer **linear networks** (matrix factorization)

[1] Xu, Min, Tarmoun, M, Vidal, Linear Convergence of Gradient Descent for Finite-Width Over-Parametrized Linear Networks with General Initialization, AISTATS 2023

[2] Xu, Min, Tarmoun, M, Vidal, A Local Polyak–Łojasiewicz and Descent Lemma of Gradient Descent for Overparameterized Linear Models, TMLR 2025

[3] Xu, Min, MacDonald, Luo, Tarmoun, M, Vidal, Understanding the Learning Dynamics of LoRA: A Gradient Flow Perspective on Low-Rank Adaptation in Matrix Factorization, AISTATS 2025

Research Goal and Questions



Goal: understand optimization process of gradient-based method for both pretraining and finetuning stage.

- **Q1:** How overparametrized models converge during pretraining? [1], [2]
- **Q2: How pretrained models adapt quickly and effectively during fine-tuning? [3]**

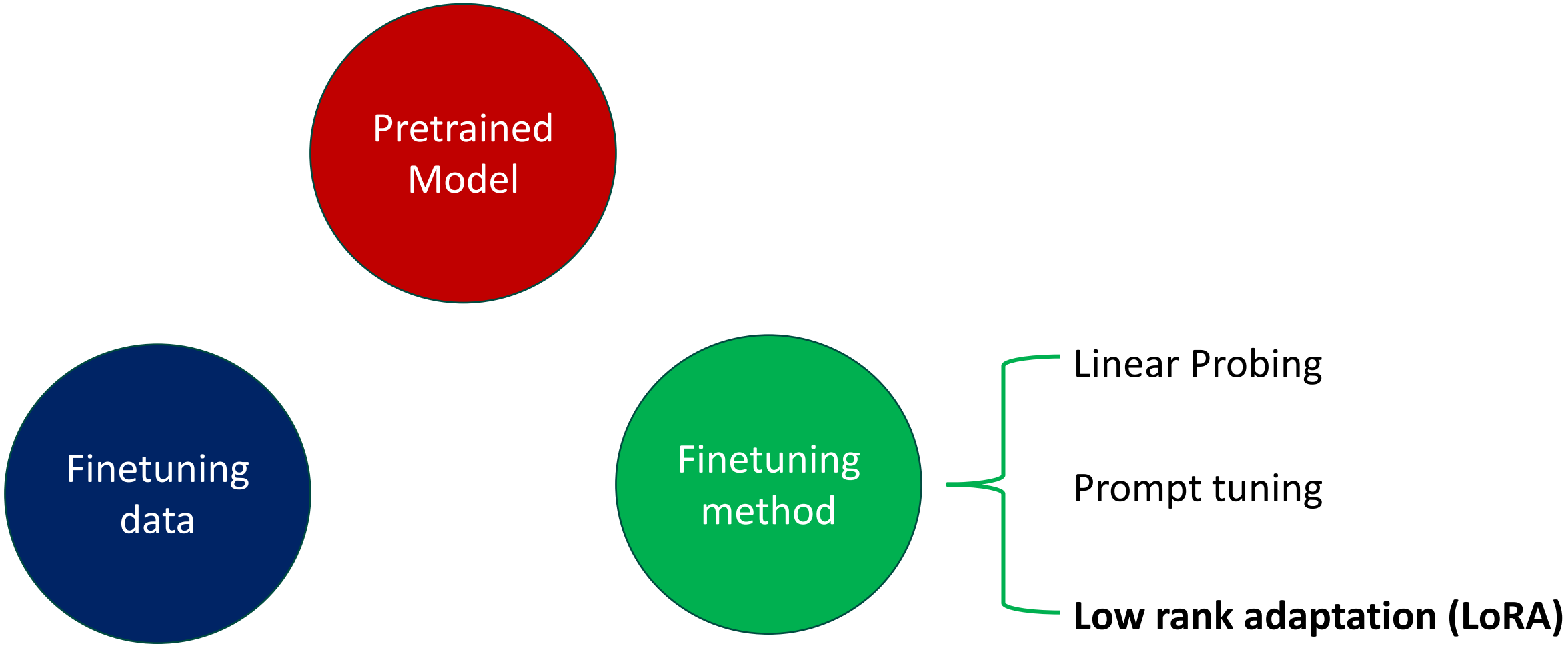
Case study: gradient flow w/ two-layer **linear networks** (matrix factorization)

[1] Xu, Min, Tarmoun, M, Vidal, Linear Convergence of Gradient Descent for Finite-Width Over-Parametrized Linear Networks with General Initialization, AISTATS 2023

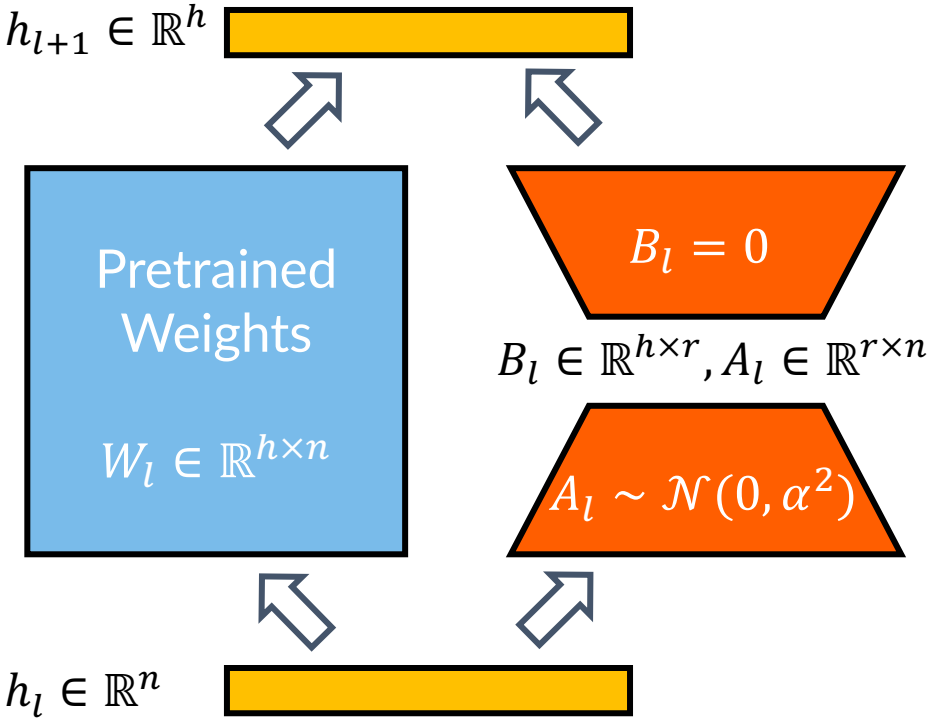
[2] Xu, Min, Tarmoun, M, Vidal, A Local Polyak–Łojasiewicz and Descent Lemma of Gradient Descent for Overparameterized Linear Models, TMLR 2025

[3] Xu, Min, MacDonald, Luo, Tarmoun, M, Vidal, Understanding the Learning Dynamics of LoRA: A Gradient Flow Perspective on Low-Rank Adaptation in Matrix Factorization, AISTATS 2025

Learning Dynamics of Finetuning Model



Overview of LoRA



- LoRA modifies the output of each layer as

$$h_{l+1} = \phi(W_l h_l) \rightarrow h_{l+1} = \phi\big((W_l + B_l A_l) h_l\big)$$

Figure from Hu et. al., 2022

Overview of LoRA

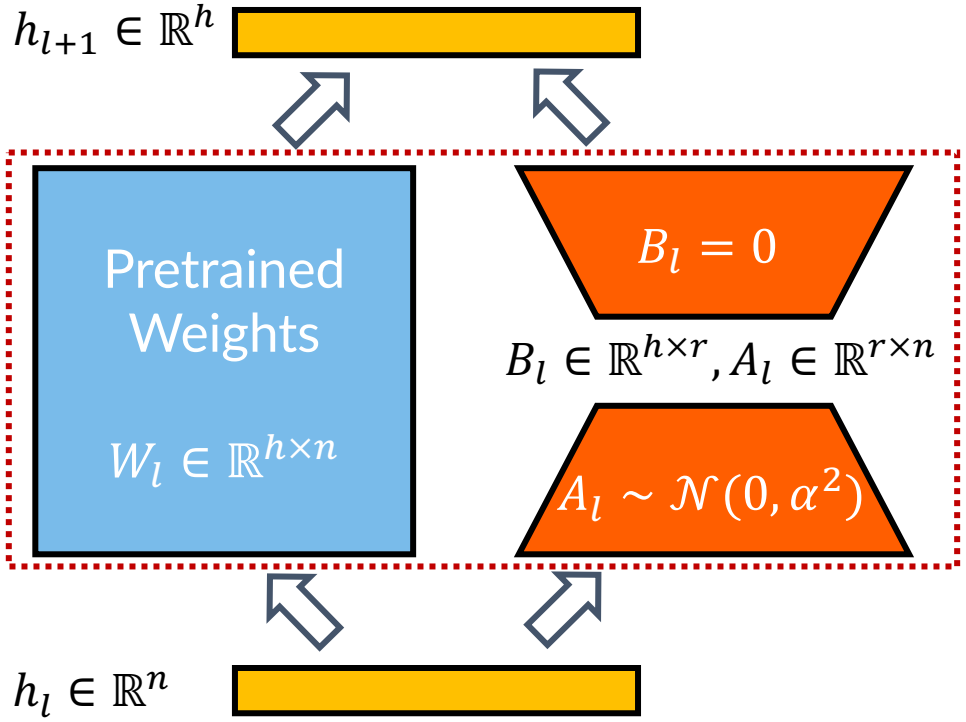


Figure from Hu et. al., 2022

- LoRA modifies the output of each layer as

$$h_{l+1} = \phi\big((W_l + B_l A_l)h_l\big)$$

During the training:

- Pretrained weights W_l are fixed
- LoRA weights B_l, A_l are trainable

Overview of LoRA

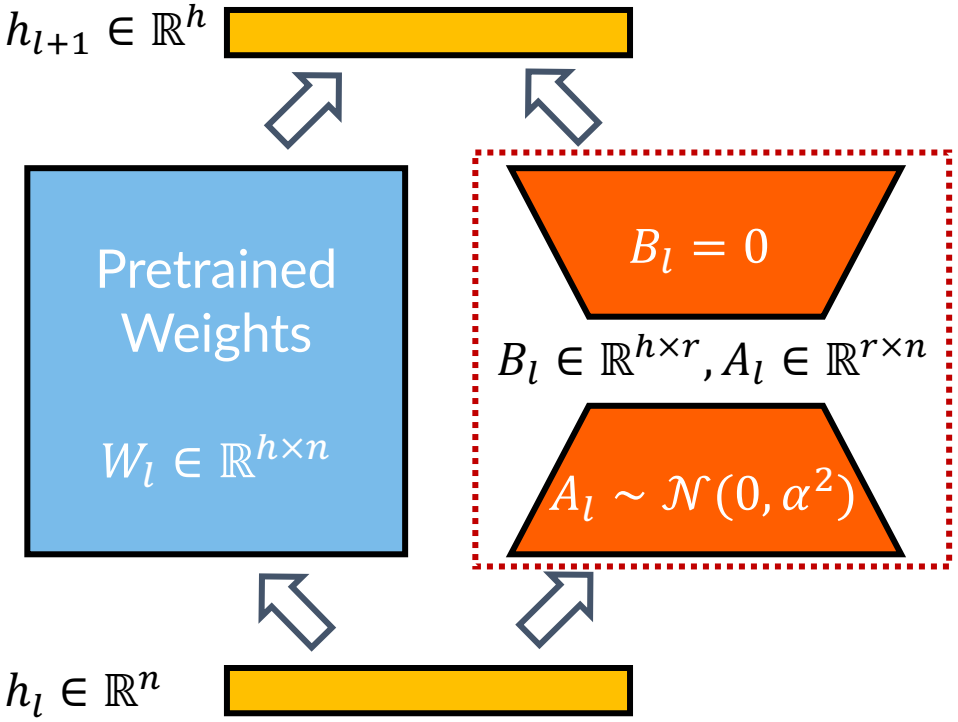


Figure from Hu et. al., 2022

- LoRA modifies the output of each layer as

$$h_{l+1} = \phi\big((W_l + B_l A_l)h_l\big)$$

During the training:

- Pretrained weights W_l are fixed
- LoRA weights B_l, A_l are trainable

- Special initialization: $A_l \sim \mathcal{N}(0, \alpha^2)$

Overview of LoRA

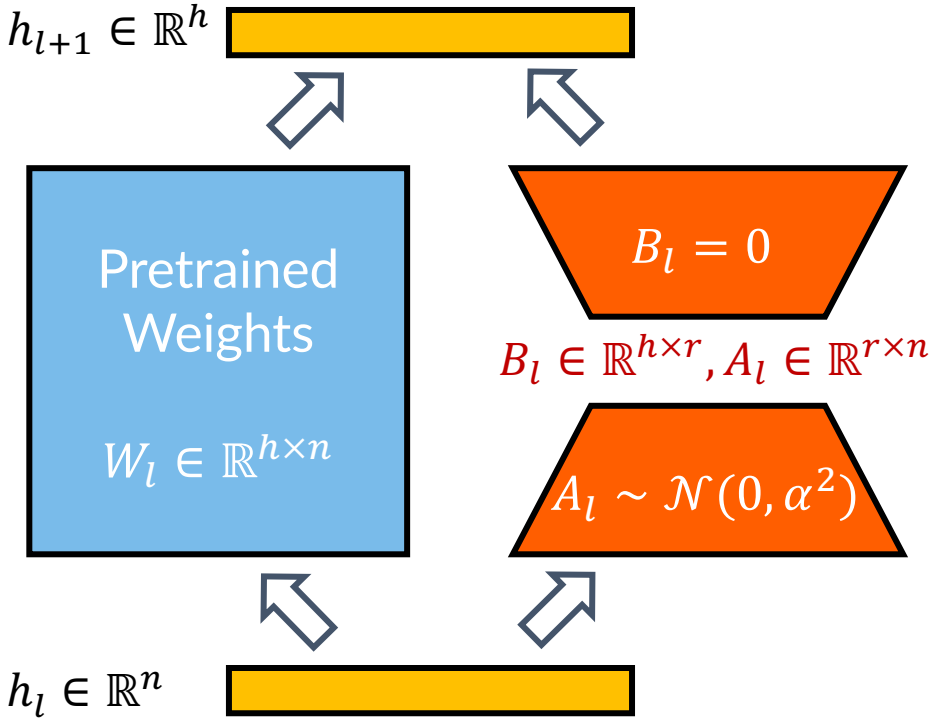


Figure from Hu et. al., 2022

- LoRA modifies the output of each layer as

$$h_{l+1} = \phi\big((W_l + B_l A_l)h_l\big)$$

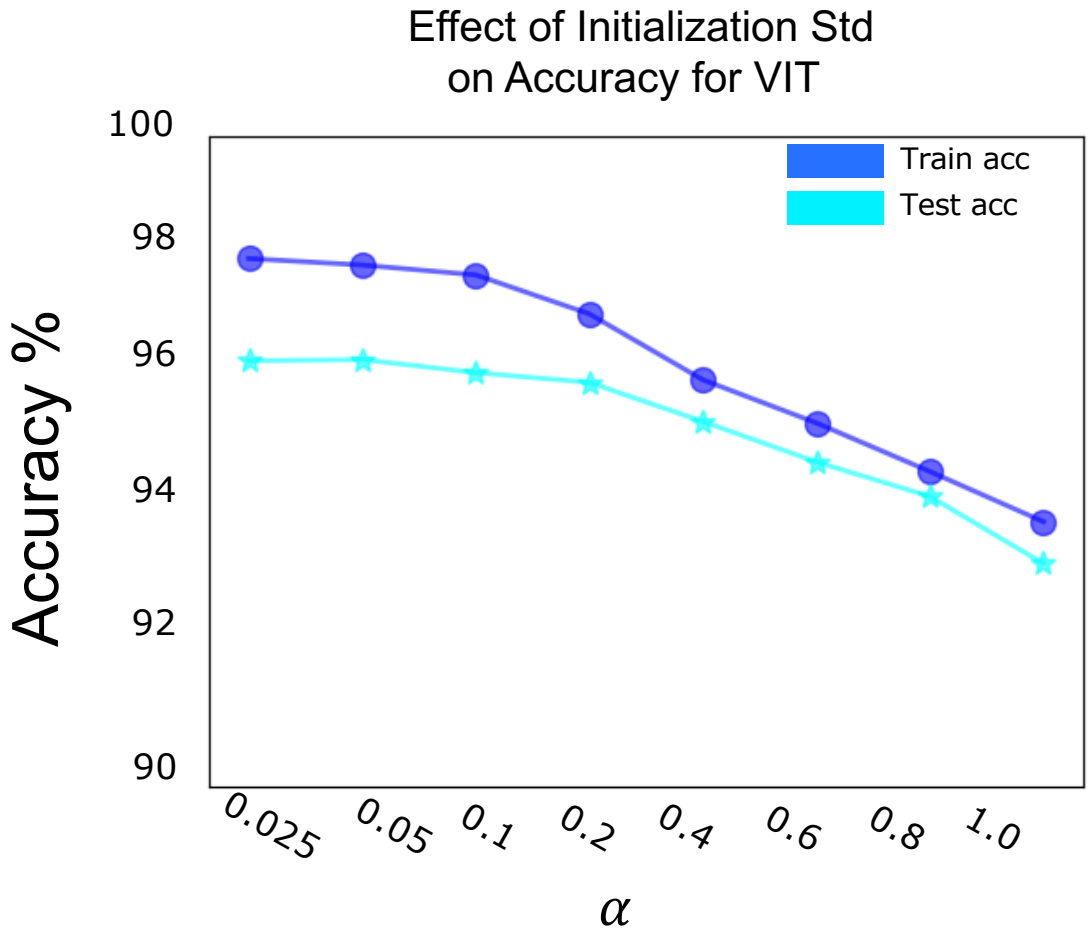
During the training:

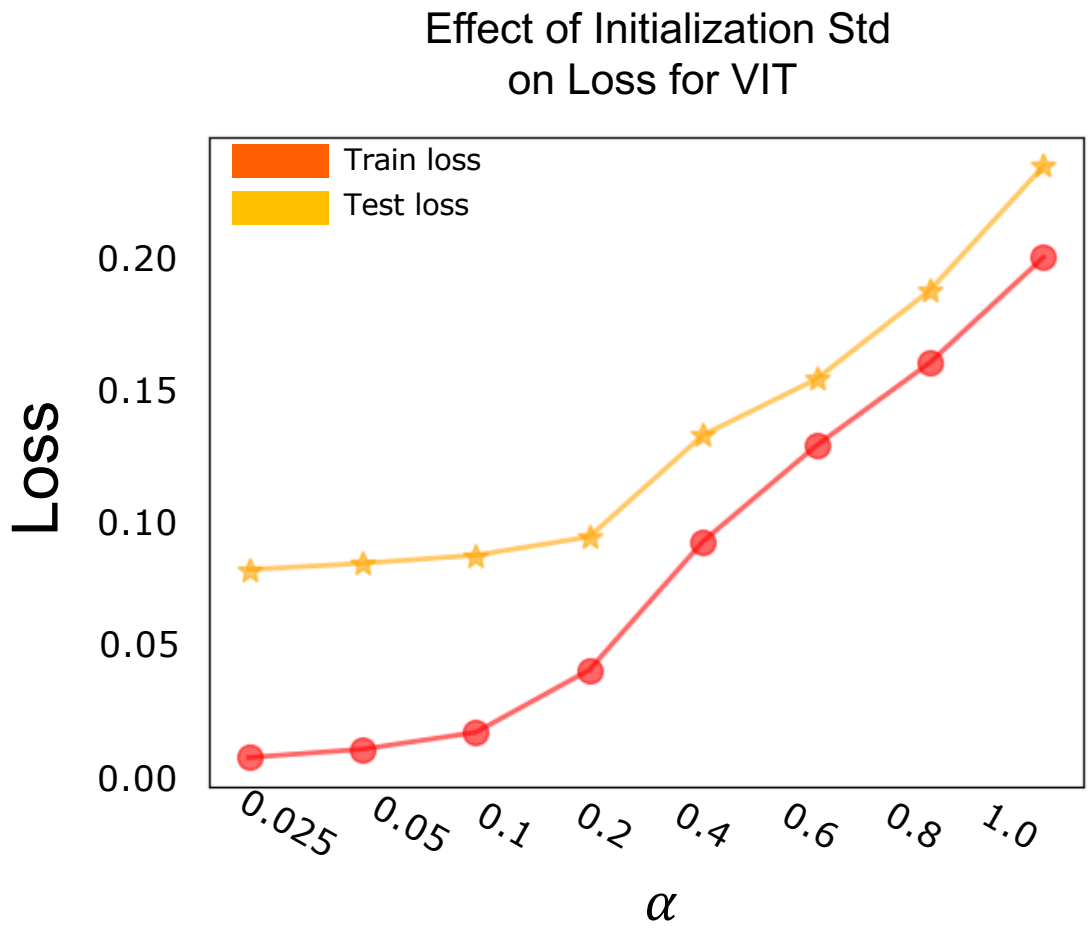
- Pretrained weights W_l are fixed
- LoRA weights B_l, A_l are trainable

- Special initialization: $A_l \sim \mathcal{N}(0, \alpha^2)$
- Low rank condition: $r \ll \min(n, h)$.

Initialization scale of LoRA matters

- Setting: we finetune a VIT, pretrained on ImageNet, on CIFAR-10





Problem Setting

- **Pretrained model**: $Y_{\text{pre}} \in \mathbb{R}^{m\times n}, U \in \mathbb{R}^{m\times h}, V \in \mathbb{R}^{h\times n}$

$$(W_2, W_1) \in \operatorname{argmin}_{U,V} \frac{1}{2}||Y_{\text{pre}} - UV||_{\text{F}}^2,$$

- **Finetuning objective**: $Y_{\text{ft}} \in \mathbb{R}^{m\times n}, A_1 \in \mathbb{R}^{r\times n}, B_1 \in \mathbb{R}^{h\times r}, A_2 \in \mathbb{R}^{r\times h}, B_2 \in \mathbb{R}^{m\times r}$

$$\min_{A_1,B_1,A_2,B_2} \frac{1}{2}||Y_{\text{ft}} - (W_2 + B_2A_2)(W_1 + B_1A_1)||_{\text{F}}^2 := \mathcal{L}$$

- **Training:** Gradient flow (GF): $\frac{d}{dt}B_i = -\nabla_{B_i}\mathcal{L}, \ \frac{d}{dt}A_i = -\nabla_{A_i}\mathcal{L}, \quad j = 1, 2$

  Small initialization: $B_2, B_1 = 0, A_1, A_2 \sim N(0, \alpha^2), \ \alpha \ll 1$

- **Goal:** Understand the learning dynamics of Gradient Flow

Finetuning model

- **Finetuning objective:** $Y_{\text{ft}} \in \mathbb{R}^{m\times n}, A_1 \in \mathbb{R}^{r\times n}, B_1 \in \mathbb{R}^{h\times r}, A_2 \in \mathbb{R}^{r\times h}, B_2 \in \mathbb{R}^{m\times r}$

$$\min_{A_1,B_1,A_2,B_2} \frac{1}{2}||Y_{\text{ft}} - (W_2 + B_2A_2)(W_1 + B_1A_1)||_{\text{F}}^2 := \mathcal{L}$$

$$= \frac{1}{2}||\underbrace{(Y_{\text{ft}} - W_2W_1)}_{\Delta Y := U_{\Delta Y}\Sigma_{\Delta Y}V_{\Delta Y}^\top} - \underbrace{(W_2B_1A_1 + B_2A_2W_1 + B_2A_2B_1A_1)}_{F := U_F\Sigma_F V_F^\top}||_{\text{F}}^2$$

Remarks/Questions:

- Loss converges to zero $\iff^*$ $U_F \to U_{\Delta Y}, V_F \to V_{\Delta Y}, \Sigma_F \to \Sigma_{\Delta Y}$
  - Do they all converge simultaneously?

- Multiple ways to fit: $\Delta Y = F$
  - $W_2B_1A_1 = \Delta Y$ and $B_2A_2W_1 + B_2A_2B_1A_1 = 0$?
  - $B_2A_2W_1 = \Delta Y$ and $W_2B_1A_1 + B_2A_2B_1A_1 = 0$?
  - $W_2B_1A_1, B_2A_2W_1, B_2A_2B_1A_1$ all contribute

*modulus a sign flip…

Main Contribution

- **Finetuning objective:** $Y_{\text{ft}} \in \mathbb{R}^{m\times n}, A_1 \in \mathbb{R}^{r\times n}, B_1 \in \mathbb{R}^{h\times r}, A_2 \in \mathbb{R}^{r\times h}, B_2 \in \mathbb{R}^{m\times r}$

$$\min_{A_1,B_1,A_2,B_2} \frac{1}{2}\left|\left|\underbrace{(Y_{\text{ft}} - W_2W_1)}_{\Delta Y \coloneqq U_{\Delta Y}\Sigma_{\Delta Y}V_{\Delta Y}^\top} - \underbrace{(W_2B_1A_1 + B_2A_2W_1 + B_2A_2B_1A_1)}_{F \coloneqq U_F\Sigma_F V_F^\top}\right|\right|_{\text{F}}^2$$

Characterization of LoRA Learning Dynamics:

- Order of convergence: $U_F \to U_{\Delta Y}, V_F \to V_{\Delta Y}$ first, followed by $\Sigma_F \to \Sigma_{\Delta Y}$
- Effect of initialization: smaller α leads to better alignment of $U_F, V_F \to U_{\Delta Y}, V_{\Delta Y}$
- Multiple ways to fit: $\Delta Y = F$
  - $W_2B_1A_1 = \Delta Y$ and $B_2A_2W_1 + B_2A_2B_1A_1 = 0$
  - $B_2A_2W_1 = \Delta Y$ and $W_2B_1A_1 + B_2A_2B_1A_1 = 0$
  - $W_2B_1A_1, B_2A_2W_1, B_2A_2B_1A_1$ all contribute

*modulus a sign flip…

Illustration of Evolution Σ_F, U_F, V_F

$$F := W_2 B_1 A_1 + B_2 A_2 W_1 + B_2 A_2 B_1 A_1$$
$$\Delta Y := Y_{\text{ft}} - Y_{\text{pre}}$$

- Data: $\Delta Y = \sigma u v^\top \in \mathbb{R}^{2\times 2}$
- SVD of model: $F = \sigma_1 u_1 v_1^\top + \sigma_2 u_2 v_2^\top$

σ, u, v $\quad$ 0

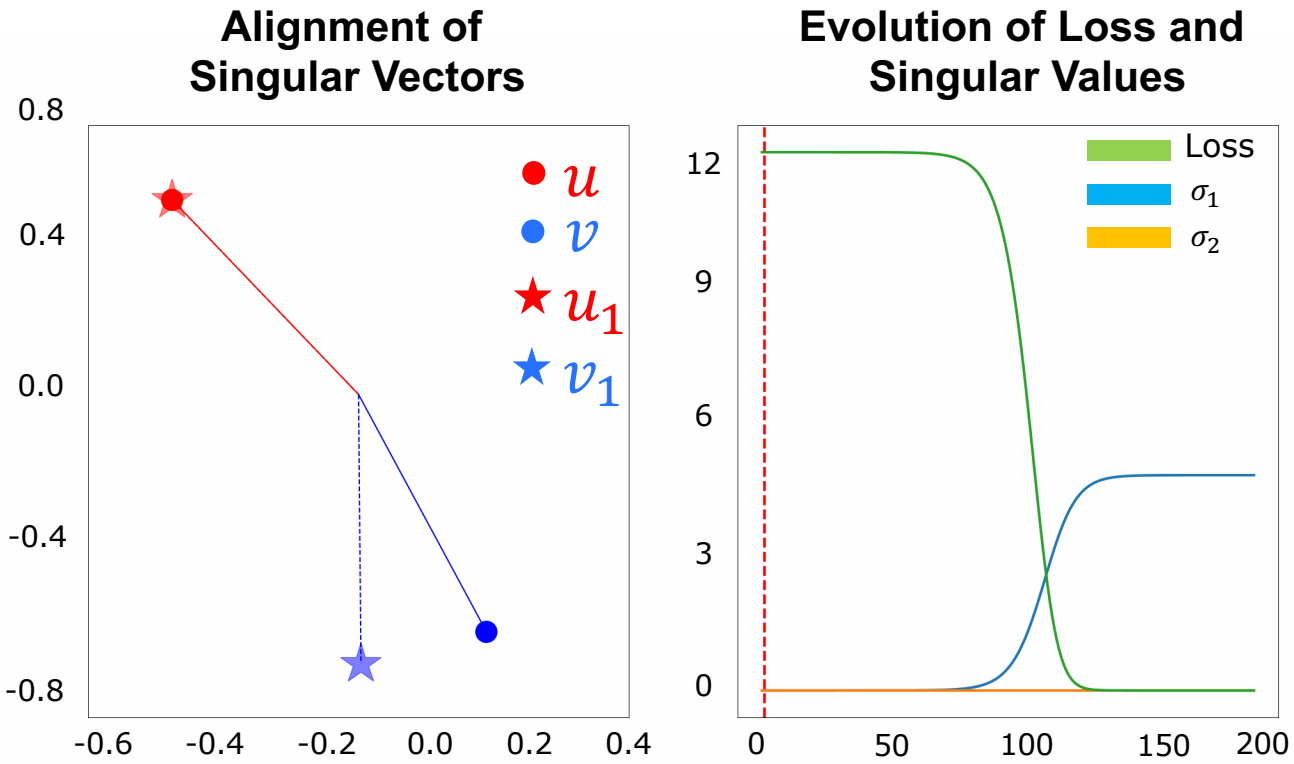
Early Stage: Alignment

$$F := W_2B_1A_1 + B_2A_2W_1 + B_2A_2B_1A_1$$
$$\Delta Y := Y_{\text{ft}} - Y_{\text{pre}}$$

- Data: $\Delta Y = \sigma u v^\top \in \mathbb{R}^{2\times 2}$
- SVD of model: $F = \sigma_1 u_1 v_1^\top + \sigma_2 u_2 v_2^\top$
- Initialization: $\alpha = 10^{-3}$

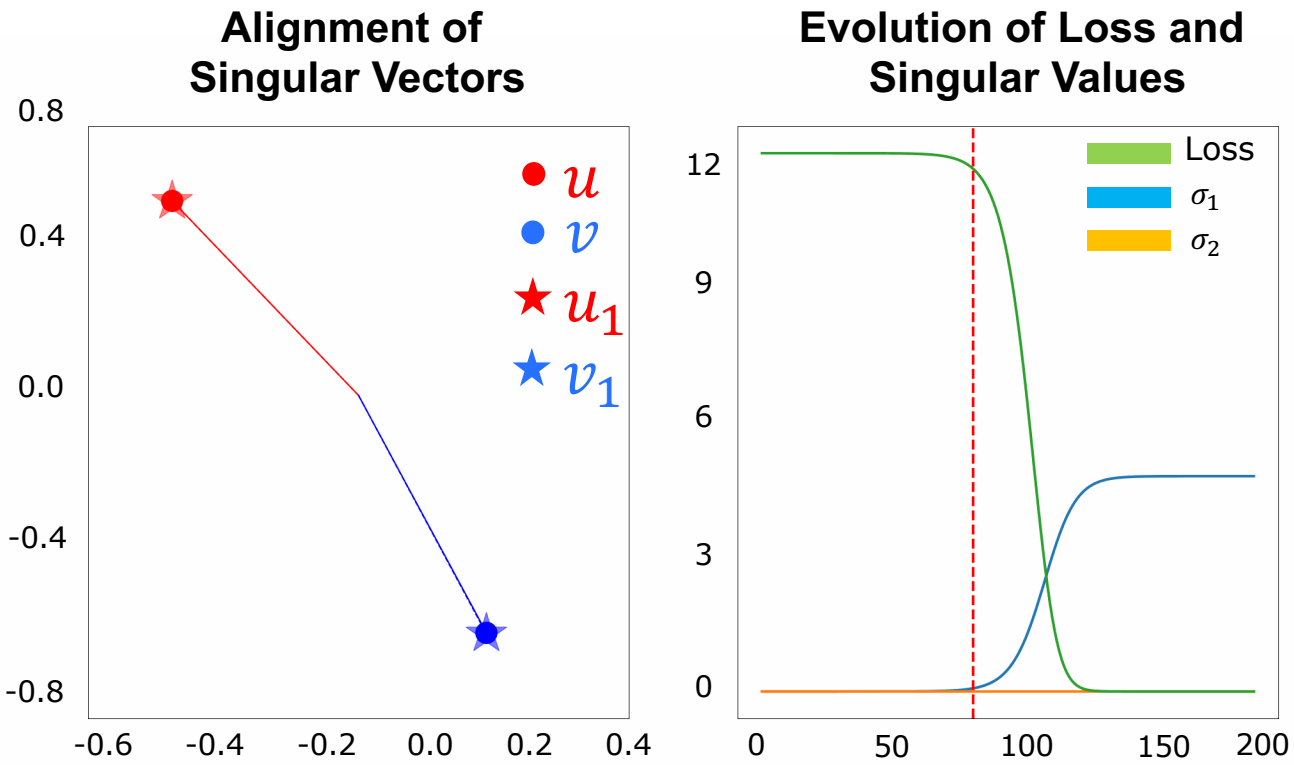
First Stage:

- u_1, v_1 align with u, v
- σ_1, σ_2 stays small
- No significant decrease in loss



Second Stage: Local Convergence

- Data: $\Delta Y = \sigma u v^\top \in \mathbb{R}^{2\times 2}$
- SVD of model: $F = \sigma_1 u_1 v_1^\top + \sigma_2 u_2 v_2^\top$
- Initialization: $\alpha = 10^{-3}$



First Stage:

- u_1, v_1 align with u, v
- σ_1, σ_2 stays small
- No significant decrease in loss

Second stage:

- σ_1 grows, σ_2 stays small
- Loss decreases fast

Two-stage Training

	First stage: Alignment Phase	**Second stage: Local Convergence Phase**
Changes in norm	Small	Large until loss is small
Changes in direction	Large until good alignment	Small and preserve good alignment from first stage

Two-stage Training

	First stage: Alignment Phase	Second stage: Local Convergence Phase
Changes in norm	Small	Large until loss is small
Changes in direction	Large until good alignment	Small and preserve good alignment from first stage

Decoupling Dynamics in Alignment Phase

$$F := W_2 B_1 A_1 + B_2 A_2 W_1 + B_2 A_2 B_1 A_1$$
$$\Delta Y := Y_{\text{ft}} - Y_{\text{pre}}$$

Gradient flow dynamics

$$\dot{A}_i = -\frac{\partial L}{\partial A_i}, \dot{B}_i = -\frac{\partial L}{\partial B_i}$$

???

Norm dynamics

$$\frac{d}{dt}||F||_2^2$$

Singular vector dynamics

$$\frac{d}{dt}u_1, \frac{d}{dt}v_1$$

Decoupling Characterizes:

1. **Model selection:** induced by the relative values of pretrained weights W_1 and W_2
2. **Alignment:** of singular vectors of F with the data ΔY

Alignment Phase: LoRA Adapters Selection

Gradient flow dynamics

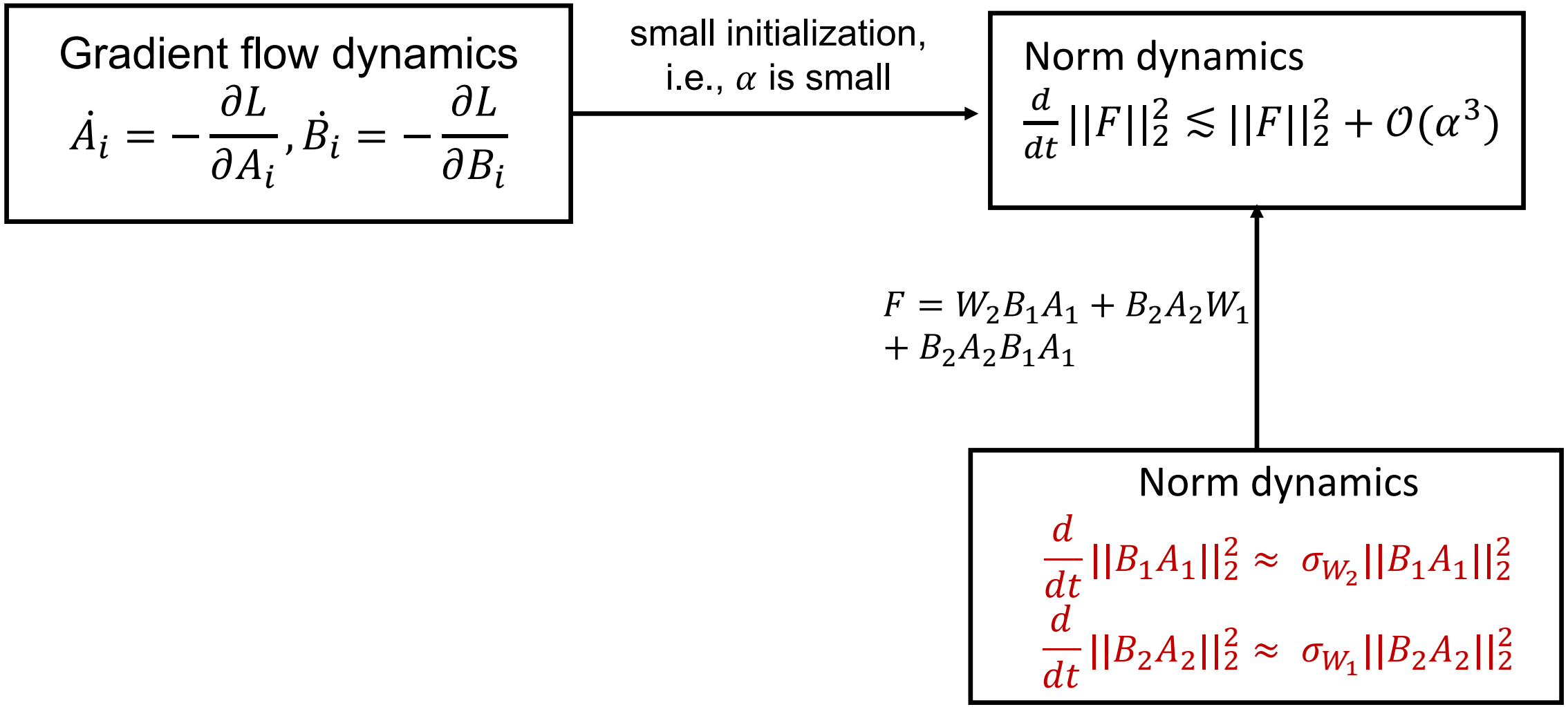
$$\dot{A}_i = -\frac{\partial L}{\partial A_i}, \dot{B}_i = -\frac{\partial L}{\partial B_i}$$

small initialization, i.e., α is small

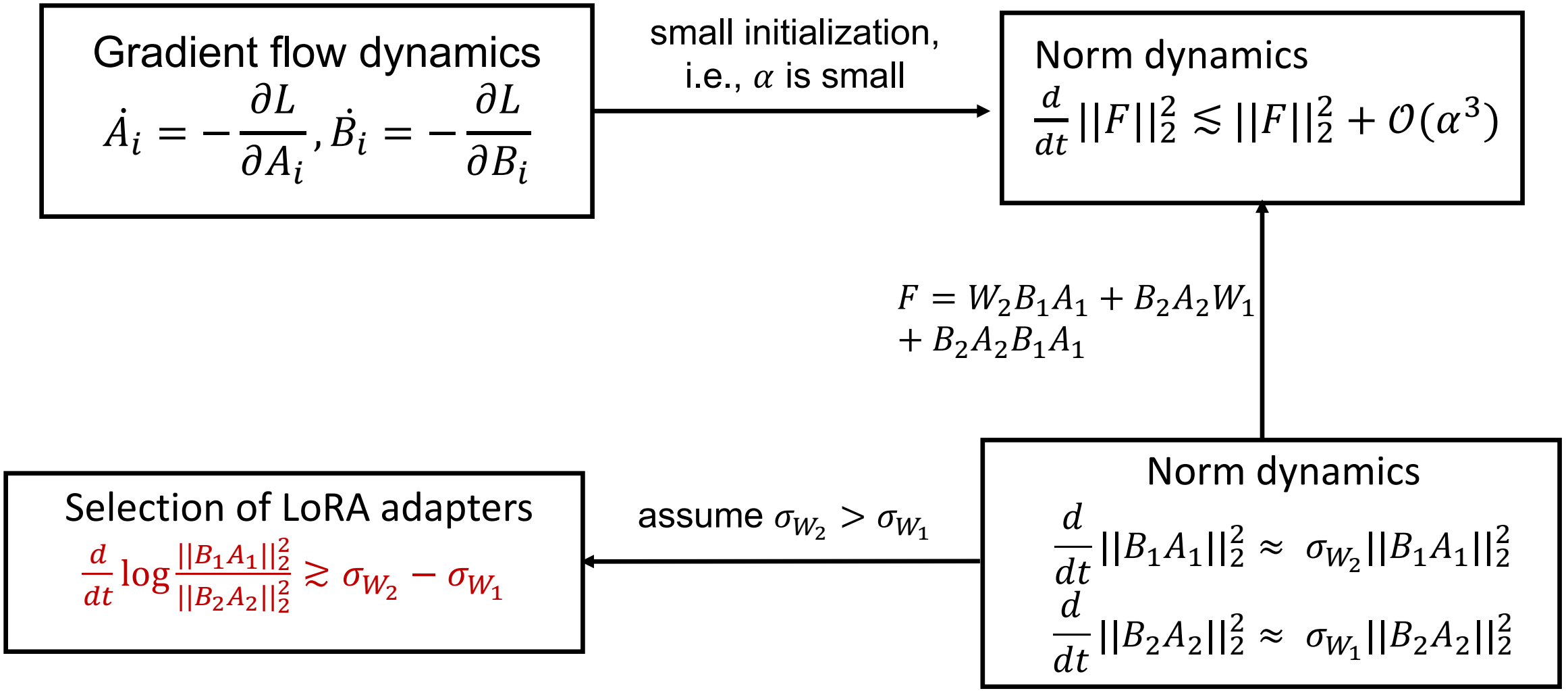
Norm dynamics: change is small

$$\frac{d}{dt}||F||_2^2 \lesssim ||F||_2^2 + \mathcal{O}(\alpha^3)$$

Alignment Phase: LoRA Adapters Selection



Alignment Phase: LoRA Adapters Selection



Alignment Phase: Singular Vector Alignment

Gradient flow dynamics

$$\dot{A}_i = -\frac{\partial L}{\partial A_i}, \dot{B}_i = -\frac{\partial L}{\partial B_i}$$

small initialization, i.e., α is small

assume $\sigma_{W_2} > \sigma_{W_1}$

Singular vector dynamics

$$\frac{d}{dt}\begin{pmatrix} u_1 \\ v_1 \end{pmatrix} \approx \sigma\sigma_{W_2}\mathcal{P}_{\gamma_1}^{\perp}\begin{pmatrix} u \\ v \end{pmatrix}$$

$$\Delta Y = \sigma u v^{\top}, \gamma_1 = \begin{pmatrix} u_1 \\ v_1 \end{pmatrix}$$

Decouple Dynamics in Alignment Phase

- LoRA weights are small at initialization, so does model F and its time derivative.
- GF dynamically selects LoRA adapters based on singular values of pretrained weights.
- Singular vectors of model F move quickly towards the target, i.e., $\begin{pmatrix} u_1 \\ v_1 \end{pmatrix} \to \begin{pmatrix} u \\ v \end{pmatrix}$.

	First stage: Alignment Phase
Changes in norm	Small
Changes in direction	Large until good alignment

Norm dynamics

$$\frac{d}{dt}||F||_2^2 \lesssim ||F||_2^2 + \mathcal{O}(\alpha^3)$$

Selection of LoRA adapters

$$\frac{d}{dt}\log\frac{||B_1A_1||_F^2}{||B_2A_2||_F^2} \gtrsim \sigma_{W_2} - \sigma_{W_1}$$

Singular vector dynamics

$$\frac{d}{dt}\begin{pmatrix} u_1 \\ v_1 \end{pmatrix} \approx \sigma\sigma_{W_2}\mathcal{P}_{\gamma_1}^{\perp}\begin{pmatrix} u \\ v \end{pmatrix}$$

Alignment Phase: Theorem

Informal theorem: Let $\boldsymbol{\delta_w = \frac{\sigma_{W_1}}{\sigma_{W_2}}}$. Suppose $\mathbf{rank}(\boldsymbol{\Delta Y}) = \mathbf{1}$, and assume $\delta_w < 1$. For any LoRA rank r, when α is sufficiently small, at the end of $\boldsymbol{T_1 = \frac{1}{\sigma\sigma_{W_2}} log\left(\frac{1}{\alpha}\right)}$, we have

- Good alignment: $\cos\left(\begin{pmatrix}u_1\\v_1\end{pmatrix}, \begin{pmatrix}u\\v\end{pmatrix}\right) \geq 1 - \mathcal{O}\left(\alpha^{\frac{1+\delta_w}{3+\delta_w}}\right)$.
- Norm stays small: $||F||_2 \leq \mathcal{O}(\alpha)$.
- Sufficient imbalance: $\frac{||B_1A_1||_2^2}{||B_2A_2||_2^2} \geq \mathcal{O}\left(\alpha^{-\frac{4(1-\delta_w)}{5-\delta_w}}\right)$.

Selection of LoRA adapters

$$\frac{d}{dt}\log\frac{||B_1A_1||_2^2}{||B_2A_2||_2^2} \gtrsim \sigma_{W_2} - \sigma_{W_1}$$

	First stage: Alignment Phase
Changes in norm	Small
Changes in direction	Large until good alignment

Two-stage Training

	First stage: Alignment Phase	Second stage: Local Convergence Phase
Changes in norm	Small	Large until loss is small
Changes in direction	Large until good alignment	Small and preserve good alignment from first stage

Two-stage Training

	First stage: Alignment Phase	Second stage: Local Convergence Phase
Changes in norm	Small	Large until loss is small
Changes in direction	Large until good alignment	Small and preserve good alignment from first stage

Local Convergence Phase

- How the result in alignment phase help?
- Simplification of the problem!

Local Convergence Phase

- How the result in alignment phase help?
- Objective: $\frac{1}{2}||\Delta Y - (W_2 B_1 A_1 + B_2 A_2 W_1 + B_2 A_2 B_1 A_1)||_{\mathrm{F}}^2$

$$\approx \frac{1}{2}||\Delta Y - W_2 B_1 A_1||_{\mathrm{F}}^2$$

$$\approx \frac{1}{2}\left(\sigma - \sigma_{W_2}\sigma_{B_1}\sigma_{A_1}\right)^2$$

- The dynamics are simplified to scalar dynamics.

Sufficient imbalance

$$\frac{||B_1 A_1||_2^2}{||B_2 A_2||_2^2} \geq \mathcal{O}\left(\alpha^{-\frac{4(1-\delta_w)}{5-\delta_w}}\right).$$

Good alignment:

$$\cos\left(\begin{pmatrix} u_1 \\ v_1 \end{pmatrix}, \begin{pmatrix} u \\ v \end{pmatrix}\right) \geq 1 - \mathcal{O}\left(\alpha^{\frac{1+\delta_w}{3+\delta_w}}\right).$$

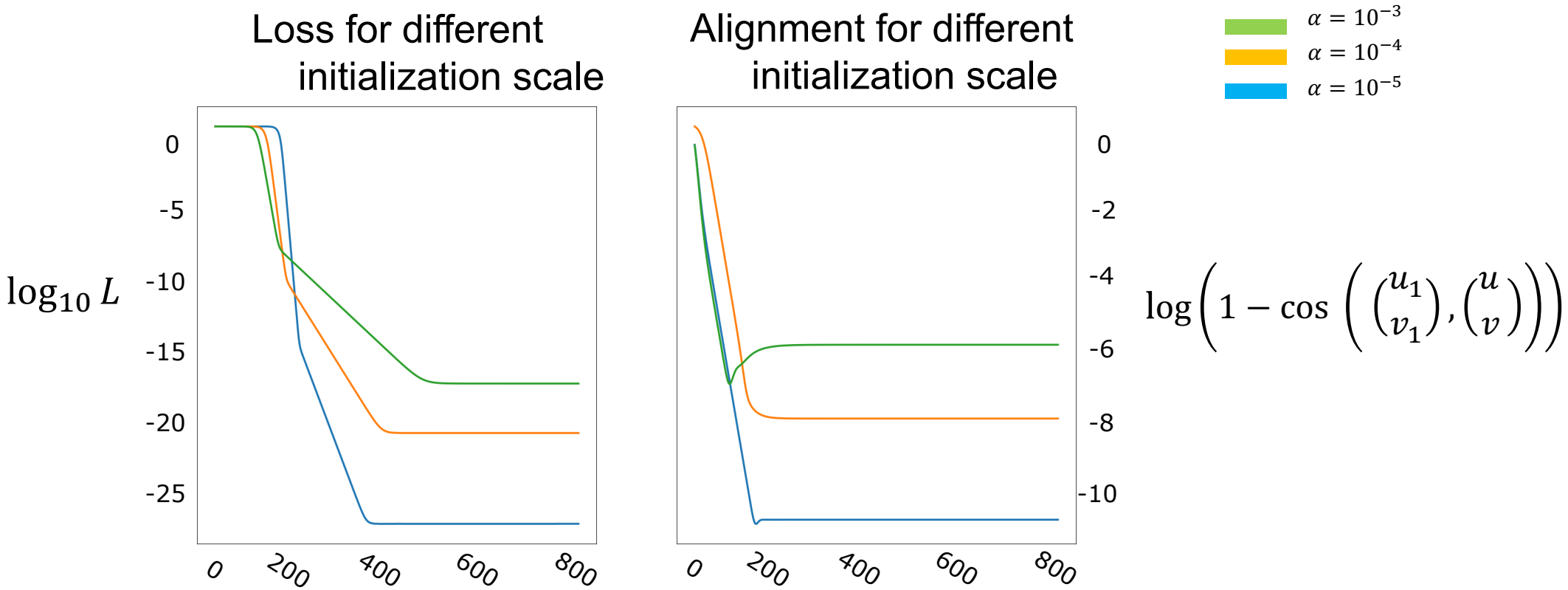
Local Convergence Phase

Informal theorem (Local Convergence): loss decreases linearly

$$L(t) \leq 2\exp\left(-\frac{(1-\delta_w)\sigma\sigma_{W_2}}{32}(t-T_1)\right)L(0) + 2\alpha^c$$

Loss for different initialization scale

Alignment for different initialization scale



Conclusion

- Studied learning dynamics of LoRA under GF for two-layer linear networks

- Our analysis shows that the learning dynamics of LoRA has two phases:
  - **Alignment phase:** GF select adapters, and orients the singular vectors of LoRA weights to correct the misalignment between the model and the finetuning task
  - **Local convergence phase:** loss decreases linearly until reaching a threshold determined by the initialization

- For small initialization, we show that GF converges to a neighborhood of the optimal solution, with smaller initialization leading to lower final loss
  - Matching empirical observation in DL models (ViT)

Thanks!

Related Publications:

1. Xu, Min, Tarmoun, M, Vidal, Linear Convergence of Gradient Descent for Finite-Width Over-Parametrized Linear Networks with General Initialization, **AISTATS 2023**
2. Xu, Min, Tarmoun, M, Vidal, A Local Polyak–Łojasiewicz and Descent Lemma of Gradient Descent for Overparameterized Linear Models, **TMLR 2025**
3. Xu, Min, MacDonald, Luo, Tarmoun, M, Vidal, Understanding the Learning Dynamics of LoRA: A Gradient Flow Perspective on Low-Rank Adaptation in Matrix Factorization, **AISTATS 2025**

Enrique Mallada
mallada@jhu.edu
http://mallada.ece.jhu.edu