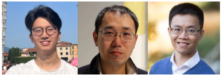
One-step full gradient suffices for low-rank fine-tuning, provably and efficiently

Fanghui Liu

fanghui.liu@warwick.ac.uk

Department of Computer Science, University of Warwick, UK
Centre for Discrete Mathematics and its Applications (DIMAP), Warwick

[joint work with Yuanhe Zhang (Warwick) and Yudong Chen (UW-Madison)]

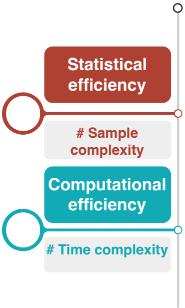


at HKUST@CS

My research

❒ Research interests

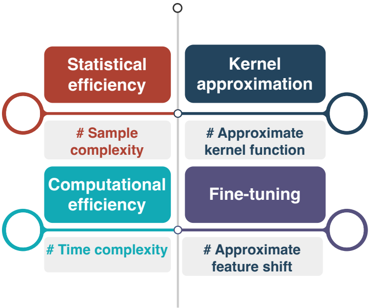
- Foundations of machine learning (ML)
- Theory-grounded efficient algorithm design
- Trustworthy ML



My research

❒ Research interests

- Foundations of machine learning (ML)
- Theory-grounded efficient algorithm design
- Trustworthy ML



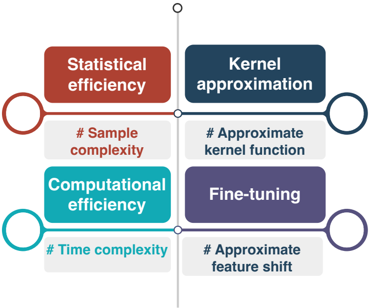
My research

- ❐ Research interests

- Foundations of machine learning (ML)
- Theory-grounded efficient algorithm design
- Trustworthy ML

- ❐ Research goal

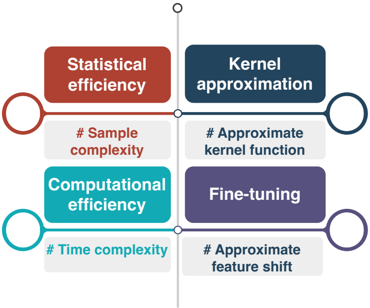
- characterize **learning efficiency** in theory
- contribute to practice



My research

❒ Research interests

- Foundations of machine learning (ML)
- Theory-grounded efficient algorithm design
- Trustworthy ML



❒ Research goal

- characterize **learning efficiency** in theory
- contribute to practice

Learning efficiency (Curse of Dimensionality, CoD)

Machine learning works in **high dimensions** that can be a **curse**!

— David Donoho, 2000. (Richard E. Bellman, 1957)

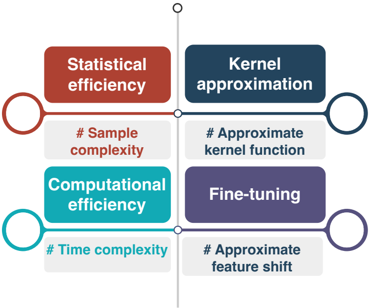
My research

❒ Research interests

- Foundations of machine learning (ML)
- Theory-grounded efficient algorithm design
- Trustworthy ML

❒ Research goal

- characterize **learning efficiency** in theory
- contribute to practice



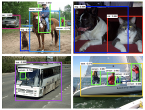
Learning efficiency (Curse of Dimensionality, CoD)

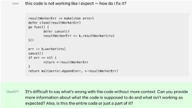
Machine learning works in **high dimensions** that can be a **curse**!

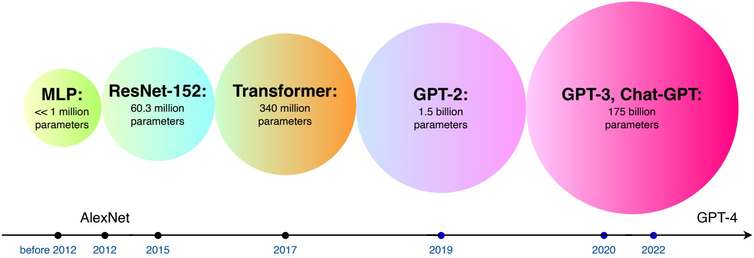
— David Donoho, 2000. (Richard E. Bellman, 1957)

In the era of machine learning (Pre-training)

relationship between data-centric, large model, huge compute resources





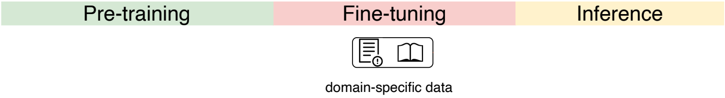


From pre-training to (parameter-efficient) fine-tuning

- GPT3: 175 billion parameters
- Llama3.1: $>$ 400 billion parameters
- Deepseek-v3: $>$ 600 billion parameters

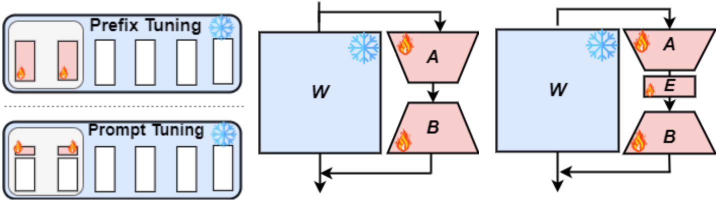
From pre-training to (parameter-efficient) fine-tuning

- GPT3: 175 billion parameters
- Llama3.1: $>$ 400 billion parameters
- Deepseek-v3: $>$ 600 billion parameters



From pre-training to (parameter-efficient) fine-tuning

- GPT3: 175 billion parameters
- Llama3.1: > 400 billion parameters
- Deepseek-v3: > 600 billion parameters



(a) Prefix & Prompt (b) LoRA (c) LoRA variants

Low-rank adaption (LoRA) for fine-tuning [2]

$$\boldsymbol{W}^{\mathrm{FT}} = \boldsymbol{W}^{\mathrm{pre}} + \Delta \in \mathbb{R}^{d \times k}$$

- $\Delta \approx \boldsymbol{A}\boldsymbol{B}$ with $\boldsymbol{A} \in \mathbb{R}^{d \times r}$ and $\boldsymbol{B} \in \mathbb{R}^{r \times k}$
- initialization

$$[\boldsymbol{A}_0]_{ij} \sim \mathcal{N}(0, \alpha^2) \quad \text{and} \quad [\boldsymbol{B}_0]_{ij} = 0\,, \quad \alpha > 0\,. \tag{LoRA-init}$$

- updated by gradient-based algorithms, e.g., SGD, AdamW
- obtain $(\boldsymbol{A}_t, \boldsymbol{B}_t)$

Low-rank adaption (LoRA) for fine-tuning [2]

$$\boldsymbol{W}^{\mathrm{FT}} = \boldsymbol{W}^{\mathrm{pre}} + \Delta \in \mathbb{R}^{d \times k}$$

- $\Delta \approx \boldsymbol{A}\boldsymbol{B}$ with $\boldsymbol{A} \in \mathbb{R}^{d \times r}$ and $\boldsymbol{B} \in \mathbb{R}^{r \times k}$
- initialization

$$[\boldsymbol{A}_0]_{ij} \sim \mathcal{N}(0, \alpha^2) \quad \text{and} \quad [\boldsymbol{B}_0]_{ij} = 0\,, \quad \alpha > 0\,. \tag{LoRA-init}$$

- updated by gradient-based algorithms, e.g., SGD, AdamW
- obtain $(\boldsymbol{A}_t, \boldsymbol{B}_t)$

Low-rank adaption (LoRA) for fine-tuning [2]

$$\boldsymbol{W}^{\mathrm{FT}} = \boldsymbol{W}^{\mathrm{pre}} + \Delta \in \mathbb{R}^{d \times k}$$

- $\Delta \approx \boldsymbol{AB}$ with $\boldsymbol{A} \in \mathbb{R}^{d \times r}$ and $\boldsymbol{B} \in \mathbb{R}^{r \times k}$
- initialization

$$[\boldsymbol{A}_0]_{ij} \sim \mathcal{N}(0, \alpha^2) \quad \text{and} \quad [\boldsymbol{B}_0]_{ij} = 0\,, \quad \alpha > 0\,. \tag{LoRA-init}$$

- updated by gradient-based algorithms, e.g., SGD, AdamW
- obtain $(\boldsymbol{A}_t, \boldsymbol{B}_t)$

Motivation: non-linear dynamics and subspace alignment

◦ Even for linear model (pre-training and fine-tuning), nonlinear dynamics...

$$\begin{bmatrix} \boldsymbol{A}_{t+1} \\ \boldsymbol{B}_{t+1}^\top \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_d & \eta_1 \boldsymbol{G}^\natural \\ \eta_2 \boldsymbol{G}^{\natural\top} & \boldsymbol{I}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_t \\ \boldsymbol{B}_t^\top \end{bmatrix} + \text{nonlinear term}\,.$$

- $\boldsymbol{G}^\natural$: one-step full gradient (from full fine-tuning)
- The dynamics $(\boldsymbol{A}_t, \boldsymbol{B}_t)$ heavily depends on $\boldsymbol{G}^\natural$!

Target

- *Q1: How to characterize low-rank dynamics of LoRA and the associated subspace alignment in theory?*
- *Q2: How can our theoretical results contribute to algorithm design for LoRA in practice?*

- Even for linear model (pre-training and fine-tuning), nonlinear dynamics...

$$\begin{bmatrix} \boldsymbol{A}_{t+1} \\ \boldsymbol{B}_{t+1}^{\top} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_d & \eta_1 \boldsymbol{G}^{\natural} \\ \eta_2 \boldsymbol{G}^{\natural\top} & \boldsymbol{I}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_t \\ \boldsymbol{B}_t^{\top} \end{bmatrix} + \text{nonlinear term}\,.$$

- $\boldsymbol{G}^{\natural}$: one-step full gradient (from full fine-tuning)
- The dynamics $(\boldsymbol{A}_t, \boldsymbol{B}_t)$ heavily depends on $\boldsymbol{G}^{\natural}$!

Target

- *Q1: How to characterize low-rank dynamics of LoRA and the associated subspace alignment in theory?*
- *Q2: How can our theoretical results contribute to algorithm design for LoRA in practice?*

Motivation: non-linear dynamics and subspace alignment

∘ Even for linear model (pre-training and fine-tuning), nonlinear dynamics...

$$\begin{bmatrix} \boldsymbol{A}_{t+1} \\ \boldsymbol{B}_{t+1}^{\top} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_d & \eta_1 \boldsymbol{G}^{\natural} \\ \eta_2 \boldsymbol{G}^{\natural\top} & \boldsymbol{I}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_t \\ \boldsymbol{B}_t^{\top} \end{bmatrix} + \text{nonlinear term}\,.$$

- $\boldsymbol{G}^{\natural}$: one-step full gradient (from full fine-tuning)
- The dynamics $(\boldsymbol{A}_t, \boldsymbol{B}_t)$ heavily depends on $\boldsymbol{G}^{\natural}$!

Target

- *Q1: How to characterize low-rank dynamics of LoRA and the associated subspace alignment in theory?*
- *Q2: How can our theoretical results contribute to algorithm design for LoRA in practice?*

Motivation: non-linear dynamics and subspace alignment

◦ Even for linear model (pre-training and fine-tuning), nonlinear dynamics...

$$\begin{bmatrix} \boldsymbol{A}_{t+1} \\ \boldsymbol{B}_{t+1}^\top \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_d & \eta_1 \boldsymbol{G}^\natural \\ \eta_2 \boldsymbol{G}^{\natural\top} & \boldsymbol{I}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_t \\ \boldsymbol{B}_t^\top \end{bmatrix} + \text{nonlinear term} \,.$$

- $\boldsymbol{G}^\natural$: one-step full gradient (from full fine-tuning)
- The dynamics $(\boldsymbol{A}_t, \boldsymbol{B}_t)$ heavily depends on $\boldsymbol{G}^\natural$!

Target

- *Q1: How to characterize low-rank dynamics of LoRA and the associated subspace alignment in theory?*
- *Q2: How can our theoretical results contribute to algorithm design for LoRA in practice?*

Alignment and theory-grounded algorithm

Problem setting and assumptions

- Pre-trained model: known $\boldsymbol{W}^\natural \in \mathbb{R}^{d\times k}$ and the ReLU activation σ

$$f_{\text{pre}}(\boldsymbol{x}) := \begin{cases} (\boldsymbol{x}^\top \boldsymbol{W}^\natural)^\top \in \mathbb{R}^k & \text{linear} \\ \sigma[(\boldsymbol{x}^\top \boldsymbol{W}^\natural)^\top] \in \mathbb{R}^k & \text{nonlinear} \end{cases}.$$

- Unknown low-rank feature shift Δ: $\widetilde{\boldsymbol{W}}^\natural := \boldsymbol{W}^\natural + \Delta$
- $\mathrm{Rank}(\Delta) = r^* < \min\{d, k\}$ with unknown r^*
- Downstream well-behaved data $\{(\widetilde{\boldsymbol{x}}_i, \widetilde{y}_i)\}_{i=1}^N$ for fine-tuning:

$$\widetilde{\boldsymbol{y}} := \begin{cases} (\widetilde{\boldsymbol{x}}^\top \widetilde{\boldsymbol{W}}^\natural)^\top \in \mathbb{R}^k, \quad \{\widetilde{\boldsymbol{x}}_i\}_{i=1}^N \overset{i.i.d.}{\sim} \text{sub-Gaussian}, & \text{linear} \\ \sigma[(\widetilde{\boldsymbol{x}}^\top \widetilde{\boldsymbol{W}}^\natural)^\top], \ \{\widetilde{\boldsymbol{x}}_i\}_{i=1}^N \overset{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{I}_d) & \text{nonlinear} \end{cases}.$$

- We assume $N > d$, e.g., MetaMathQA, Code-Feedback, $d = 1,024$ and $N \sim 10^5$

Problem setting and assumptions

- Pre-trained model: known $\boldsymbol{W}^\natural \in \mathbb{R}^{d\times k}$ and the ReLU activation σ

$$f_{\text{pre}}(\boldsymbol{x}) := \begin{cases} (\boldsymbol{x}^\top \boldsymbol{W}^\natural)^\top \in \mathbb{R}^k & \text{linear} \\ \sigma[(\boldsymbol{x}^\top \boldsymbol{W}^\natural)^\top] \in \mathbb{R}^k & \text{nonlinear} \end{cases}.$$

- Unknown low-rank feature shift Δ: $\widetilde{\boldsymbol{W}}^\natural := \boldsymbol{W}^\natural + \Delta$
- $\text{Rank}(\Delta) = r^* < \min\{d, k\}$ with unknown r^*
- Downstream well-behaved data $\{(\widetilde{\boldsymbol{x}}_i, \widetilde{y}_i)\}_{i=1}^N$ for fine-tuning:

$$\widetilde{\boldsymbol{y}} := \begin{cases} (\widetilde{\boldsymbol{x}}^\top \widetilde{\boldsymbol{W}}^\natural)^\top \in \mathbb{R}^k, \quad \{\widetilde{\boldsymbol{x}}_i\}_{i=1}^N \overset{i.i.d.}{\sim} \text{sub-Gaussian}, & \text{linear} \\ \sigma[(\widetilde{\boldsymbol{x}}^\top \widetilde{\boldsymbol{W}}^\natural)^\top], \ \{\widetilde{\boldsymbol{x}}_i\}_{i=1}^N \overset{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{I}_d) & \text{nonlinear} \end{cases}.$$

- We assume $N > d$, e.g., MetaMathQA, Code-Feedback, $d = 1,024$ and $N \sim 10^5$

Problem setting and assumptions

- Pre-trained model: known $\boldsymbol{W}^\natural \in \mathbb{R}^{d\times k}$ and the ReLU activation σ

$$f_{\text{pre}}(\boldsymbol{x}) := \begin{cases} (\boldsymbol{x}^\top \boldsymbol{W}^\natural)^\top \in \mathbb{R}^k & \text{linear} \\ \sigma[(\boldsymbol{x}^\top \boldsymbol{W}^\natural)^\top] \in \mathbb{R}^k & \text{nonlinear} \end{cases}.$$

- Unknown low-rank feature shift Δ: $\widetilde{\boldsymbol{W}}^\natural := \boldsymbol{W}^\natural + \Delta$
- $\text{Rank}(\Delta) = r^* < \min\{d, k\}$ with unknown r^*
- Downstream well-behaved data $\{(\widetilde{\boldsymbol{x}}_i, \widetilde{y}_i)\}_{i=1}^N$ for fine-tuning:

$$\widetilde{\boldsymbol{y}} := \begin{cases} (\widetilde{\boldsymbol{x}}^\top \widetilde{\boldsymbol{W}}^\natural)^\top \in \mathbb{R}^k, \quad \{\widetilde{\boldsymbol{x}}_i\}_{i=1}^N \overset{i.i.d.}{\sim} \text{sub-Gaussian}, & \text{linear} \\ \sigma[(\widetilde{\boldsymbol{x}}^\top \widetilde{\boldsymbol{W}}^\natural)^\top], \ \{\widetilde{\boldsymbol{x}}_i\}_{i=1}^N \overset{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{I}_d) & \text{nonlinear} \end{cases}.$$

- We assume $N > d$, e.g., MetaMathQA, Code-Feedback, $d = 1,024$ and $N \sim 10^5$

Problem setting and assumptions

- Pre-trained model: known $\boldsymbol{W}^\natural \in \mathbb{R}^{d\times k}$ and the ReLU activation σ

$$f_{\text{pre}}(\boldsymbol{x}) := \begin{cases} (\boldsymbol{x}^\top \boldsymbol{W}^\natural)^\top \in \mathbb{R}^k & \text{linear} \\ \sigma[(\boldsymbol{x}^\top \boldsymbol{W}^\natural)^\top] \in \mathbb{R}^k & \text{nonlinear} \end{cases}.$$

- Unknown low-rank feature shift Δ: $\widetilde{\boldsymbol{W}}^\natural := \boldsymbol{W}^\natural + \Delta$
- $\text{Rank}(\Delta) = r^* < \min\{d\,, k\}$ with unknown r^*
- Downstream well-behaved data $\{(\widetilde{\boldsymbol{x}}_i, \widetilde{y}_i)\}_{i=1}^N$ for fine-tuning:

$$\widetilde{\boldsymbol{y}} := \begin{cases} (\widetilde{\boldsymbol{x}}^\top \widetilde{\boldsymbol{W}}^\natural)^\top \in \mathbb{R}^k, \quad \{\widetilde{\boldsymbol{x}}_i\}_{i=1}^N \overset{i.i.d.}{\sim} \text{sub-Gaussian}, & \text{linear} \\ \sigma[(\widetilde{\boldsymbol{x}}^\top \widetilde{\boldsymbol{W}}^\natural)^\top],\ \{\widetilde{\boldsymbol{x}}_i\}_{i=1}^N \overset{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{I}_d) & \text{nonlinear} \end{cases}.$$

- We assume $N > d$, e.g., MetaMathQA, Code-Feedback, $d = 1,024$ and $N \sim 10^5$

Full fine-tuning and LoRA updates

- full fine-tuning (initialized at $\boldsymbol{W}_0 := \boldsymbol{W}^\natural$)

$$L(\boldsymbol{W}) := \frac{1}{2N} \begin{cases} \left\| \widetilde{\boldsymbol{X}} \boldsymbol{W} - \widetilde{\boldsymbol{Y}} \right\|_{\mathrm{F}}^2 & \text{linear} \\ \left\| \sigma(\widetilde{\boldsymbol{X}} \boldsymbol{W}) - \widetilde{\boldsymbol{Y}} \right\|_{\mathrm{F}}^2 & \text{nonlinear} \end{cases}$$

- LoRA update

$$\widetilde{L}(\boldsymbol{A}, \boldsymbol{B}) := \frac{1}{2N} \begin{cases} \left\| \widetilde{\boldsymbol{X}} (\boldsymbol{W}^\natural + \boldsymbol{A}\boldsymbol{B}) - \widetilde{\boldsymbol{Y}} \right\|_{\mathrm{F}}^2 & \text{linear} \\ \left\| \sigma\Big(\widetilde{\boldsymbol{X}} (\boldsymbol{W}^\natural + \boldsymbol{A}\boldsymbol{B})\Big) - \widetilde{\boldsymbol{Y}} \right\|_{\mathrm{F}}^2 & \text{nonlinear} \end{cases}$$

- Gradient descent with different step-size, e.g., LoRA+ [1]

$$\boldsymbol{A}_{t+1} = \boldsymbol{A}_t - \eta_1 \nabla_{\boldsymbol{A}} \widetilde{L}(\boldsymbol{A}_t, \boldsymbol{B}_t)$$

$$\boldsymbol{B}_{t+1} = \boldsymbol{B}_t - \eta_2 \nabla_{\boldsymbol{B}} \widetilde{L}(\boldsymbol{A}_t, \boldsymbol{B}_t)$$

- Evaluation by $\|\boldsymbol{A}_t \boldsymbol{B}_t - \Delta\|_{\mathrm{F}}$: optimization and generalization!

Full fine-tuning and LoRA updates

- full fine-tuning (initialized at $\boldsymbol{W}_0 := \boldsymbol{W}^\natural$)

$$L(\boldsymbol{W}) := \frac{1}{2N}\begin{cases} \left\|\widetilde{\boldsymbol{X}}\boldsymbol{W} - \widetilde{\boldsymbol{Y}}\right\|_{\mathrm{F}}^2 & \text{linear} \\ \left\|\sigma(\widetilde{\boldsymbol{X}}\boldsymbol{W}) - \widetilde{\boldsymbol{Y}}\right\|_{\mathrm{F}}^2 & \text{nonlinear} \end{cases}$$

- LoRA update

$$\widetilde{L}(\boldsymbol{A}, \boldsymbol{B}) := \frac{1}{2N}\begin{cases} \left\|\widetilde{\boldsymbol{X}}(\boldsymbol{W}^\natural + \boldsymbol{A}\boldsymbol{B}) - \widetilde{\boldsymbol{Y}}\right\|_{\mathrm{F}}^2 & \text{linear} \\ \left\|\sigma\Big(\widetilde{\boldsymbol{X}}(\boldsymbol{W}^\natural + \boldsymbol{A}\boldsymbol{B})\Big) - \widetilde{\boldsymbol{Y}}\right\|_{\mathrm{F}}^2 & \text{nonlinear} \end{cases}$$

- Gradient descent with different step-size, e.g., LoRA+ [1]

$$\boldsymbol{A}_{t+1} = \boldsymbol{A}_t - \eta_1 \nabla_{\boldsymbol{A}} \widetilde{L}(\boldsymbol{A}_t, \boldsymbol{B}_t)$$

$$\boldsymbol{B}_{t+1} = \boldsymbol{B}_t - \eta_2 \nabla_{\boldsymbol{B}} \widetilde{L}(\boldsymbol{A}_t, \boldsymbol{B}_t)$$

- Evaluation by $\|\boldsymbol{A}_t\boldsymbol{B}_t - \Delta\|_{\mathrm{F}}$: optimization and generalization!

Full fine-tuning and LoRA updates

- full fine-tuning (initialized at $\boldsymbol{W}_0 := \boldsymbol{W}^\natural$)

$$L(\boldsymbol{W}) := \frac{1}{2N}\begin{cases} \left\| \widetilde{\boldsymbol{X}}\boldsymbol{W} - \widetilde{\boldsymbol{Y}} \right\|_{\mathrm{F}}^2 & \text{linear} \\ \left\| \sigma(\widetilde{\boldsymbol{X}}\boldsymbol{W}) - \widetilde{\boldsymbol{Y}} \right\|_{\mathrm{F}}^2 & \text{nonlinear} \end{cases}$$

- LoRA update

$$\widetilde{L}(\boldsymbol{A}, \boldsymbol{B}) := \frac{1}{2N}\begin{cases} \left\| \widetilde{\boldsymbol{X}}(\boldsymbol{W}^\natural + \boldsymbol{A}\boldsymbol{B}) - \widetilde{\boldsymbol{Y}} \right\|_{\mathrm{F}}^2 & \text{linear} \\ \left\| \sigma\Big(\widetilde{\boldsymbol{X}}(\boldsymbol{W}^\natural + \boldsymbol{A}\boldsymbol{B})\Big) - \widetilde{\boldsymbol{Y}} \right\|_{\mathrm{F}}^2 & \text{nonlinear} \end{cases}$$

- Gradient descent with different step-size, e.g., LoRA+ [1]

$$\boldsymbol{A}_{t+1} = \boldsymbol{A}_t - \eta_1 \nabla_{\boldsymbol{A}} \widetilde{L}(\boldsymbol{A}_t, \boldsymbol{B}_t)$$

$$\boldsymbol{B}_{t+1} = \boldsymbol{B}_t - \eta_2 \nabla_{\boldsymbol{B}} \widetilde{L}(\boldsymbol{A}_t, \boldsymbol{B}_t)$$

- Evaluation by $\|\boldsymbol{A}_t\boldsymbol{B}_t - \Delta\|_{\mathrm{F}}$: optimization and generalization!

Full fine-tuning and LoRA updates

- full fine-tuning (initialized at $\boldsymbol{W}_0 := \boldsymbol{W}^\natural$)

$$L(\boldsymbol{W}) := \frac{1}{2N}\begin{cases} \left\|\widetilde{\boldsymbol{X}}\boldsymbol{W} - \widetilde{\boldsymbol{Y}}\right\|_{\mathrm{F}}^2 & \text{linear} \\ \left\|\sigma(\widetilde{\boldsymbol{X}}\boldsymbol{W}) - \widetilde{\boldsymbol{Y}}\right\|_{\mathrm{F}}^2 & \text{nonlinear} \end{cases}$$

- LoRA update

$$\widetilde{L}(\boldsymbol{A}, \boldsymbol{B}) := \frac{1}{2N}\begin{cases} \left\|\widetilde{\boldsymbol{X}}(\boldsymbol{W}^\natural + \boldsymbol{A}\boldsymbol{B}) - \widetilde{\boldsymbol{Y}}\right\|_{\mathrm{F}}^2 & \text{linear} \\ \left\|\sigma\Big(\widetilde{\boldsymbol{X}}(\boldsymbol{W}^\natural + \boldsymbol{A}\boldsymbol{B})\Big) - \widetilde{\boldsymbol{Y}}\right\|_{\mathrm{F}}^2 & \text{nonlinear} \end{cases}$$

- Gradient descent with different step-size, e.g., LoRA+ [1]

$$\boldsymbol{A}_{t+1} = \boldsymbol{A}_t - \eta_1 \nabla_{\boldsymbol{A}} \widetilde{L}(\boldsymbol{A}_t, \boldsymbol{B}_t)$$

$$\boldsymbol{B}_{t+1} = \boldsymbol{B}_t - \eta_2 \nabla_{\boldsymbol{B}} \widetilde{L}(\boldsymbol{A}_t, \boldsymbol{B}_t)$$

- Evaluation by $\|\boldsymbol{A}_t\boldsymbol{B}_t - \Delta\|_{\mathrm{F}}$: optimization and generalization!

Alignment on B_t

○ one-step full gradient: $\boldsymbol{G}^\natural \in \mathbb{R}^{d\times k}$ and $\mathrm{rank}(\boldsymbol{G}^\natural) = r^*$

$$\boldsymbol{G}^\natural := -\nabla_{\boldsymbol{W}} L(\boldsymbol{W}^\natural) = \frac{1}{N}\widetilde{\boldsymbol{X}}^\top(\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{W}^\natural) = \frac{1}{N}\widetilde{\boldsymbol{X}}^\top\widetilde{\boldsymbol{X}}\Delta\,.$$

Theorem (Alignment between $G^\natural$ and B_t)

For the linear setting, consider the LoRA updates with (LoRA-init). *We have*

$$\left\|\boldsymbol{V}_{r^*,\perp}^\top\left(\boldsymbol{G}^\natural\right)\boldsymbol{V}_{r^*}(\boldsymbol{B}_t)\right\|_{op} = 0\,,\quad \forall t \in \mathbb{N}_+\,.$$

Remark: $\boldsymbol{B}_1 = \eta_1\boldsymbol{A}_0^\top\boldsymbol{G}^\natural$ with $\mathrm{Rank}(\boldsymbol{B}_1) \le r^*$

Alignment on B_t

- one-step full gradient: $\boldsymbol{G}^{\natural} \in \mathbb{R}^{d \times k}$ and $\operatorname{rank}(\boldsymbol{G}^{\natural}) = r^*$

$$\boldsymbol{G}^{\natural} := -\nabla_{\boldsymbol{W}} L(\boldsymbol{W}^{\natural}) = \frac{1}{N} \widetilde{\boldsymbol{X}}^{\top} (\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}} \boldsymbol{W}^{\natural}) = \frac{1}{N} \widetilde{\boldsymbol{X}}^{\top} \widetilde{\boldsymbol{X}} \Delta \,.$$

Theorem (Alignment between $\boldsymbol{G}^{\natural}$ and $\boldsymbol{B}_t$)

For the linear setting, consider the LoRA updates with (LoRA-init). We have

$$\left\| \boldsymbol{V}_{r^*,\perp}^{\top} \left(\boldsymbol{G}^{\natural} \right) \boldsymbol{V}_{r^*}(\boldsymbol{B}_t) \right\|_{op} = 0 \,, \quad \forall t \in \mathbb{N}_+ \,.$$

Remark: $\boldsymbol{B}_1 = \eta_1 \boldsymbol{A}_0^{\top} \boldsymbol{G}^{\natural}$ with $\operatorname{Rank}(\boldsymbol{B}_1) \leq r^*$

Alignment on B_t

∘ one-step full gradient: $\boldsymbol{G}^\natural \in \mathbb{R}^{d\times k}$ and $\mathrm{rank}(\boldsymbol{G}^\natural) = r^*$

$$\boldsymbol{G}^\natural := -\nabla_{\boldsymbol{W}} L(\boldsymbol{W}^\natural) = \frac{1}{N}\widetilde{\boldsymbol{X}}^\top(\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{W}^\natural) = \frac{1}{N}\widetilde{\boldsymbol{X}}^\top\widetilde{\boldsymbol{X}}\Delta\,.$$

Theorem (Alignment between $\boldsymbol{G}^\natural$ and $\boldsymbol{B}_t$)

For the linear setting, consider the LoRA updates with (LoRA-init). *We have*

$$\left\| \boldsymbol{V}_{r^*,\perp}^\top\left(\boldsymbol{G}^\natural\right)\boldsymbol{V}_{r^*}(\boldsymbol{B}_t)\right\|_{op} = 0\,, \quad \forall t \in \mathbb{N}_+\,.$$

Remark: $\boldsymbol{B}_1 = \eta_1 \boldsymbol{A}_0^\top \boldsymbol{G}^\natural$ with $\mathrm{Rank}(\boldsymbol{B}_1) \le r^*$

Theorem (Informal)

For $r \geq r^*$, recall $[\boldsymbol{A}_0]_{ij} \sim \mathcal{N}(0, \alpha^2)$ in (LoRA-init), for any $\epsilon \in (0,1)$, choosing $\alpha = \mathcal{O}\left(\epsilon d^{-\frac{3}{4}\kappa^\natural - \frac{1}{2}}\right)$, running GD with $t^* \asymp \frac{\ln d}{\sqrt{\eta_1 \eta_2}}$, then we have

$$\left\| \boldsymbol{U}^\top_{r^*,\perp}(\boldsymbol{G}^\natural)\ \boldsymbol{U}_{r^*}(\boldsymbol{A}_{t^*}) \right\|_{op} \lesssim \epsilon\,, w.h.p.$$

- small initialization: better alignment and better generalization performance
- imbalanced step-size finishes alignment earlier

Theorem (Informal)

For $r \geq r^*$, recall $[\mathbf{A}_0]_{ij} \sim \mathcal{N}(0, \alpha^2)$ in (LoRA-init), for any $\epsilon \in (0,1)$, choosing $\alpha = \mathcal{O}\left(\epsilon d^{-\frac{3}{4}\kappa^\natural - \frac{1}{2}}\right)$, running GD with $t^* \asymp \frac{\ln d}{\sqrt{\eta_1 \eta_2}}$, then we have

$$\left\| \mathbf{U}_{r^*,\perp}^\top(\mathbf{G}^\natural)\ \mathbf{U}_{r^*}(\mathbf{A}_{t^*}) \right\|_{op} \lesssim \epsilon\,, w.h.p.$$

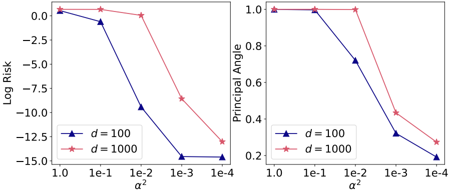


Figure 1: *Left*: the risk $\frac{1}{2}\|\mathbf{A}_t\mathbf{B}_t - \Delta\|_{\mathrm{F}}^2$. *Right*: the principal angle is defined as $\min_t \|\mathbf{U}_{r^*,\perp}^\top(\mathbf{G}^\natural)\ \mathbf{U}_{r^*}(\mathbf{A}_t)\|_{op}$.

- small initialization: better alignment and better generalization performance
- imbalanced step-size finishes alignment earlier

- Take the SVD of $\boldsymbol{G}^\natural$: $\boldsymbol{G}^\natural = \widetilde{\boldsymbol{U}}_{\boldsymbol{G}^\natural} \widetilde{\boldsymbol{S}}_{\boldsymbol{G}^\natural} \widetilde{\boldsymbol{V}}_{\boldsymbol{G}^\natural}^\top$

$$\boldsymbol{A}_0 = \sqrt{\gamma}\left[\widetilde{\boldsymbol{U}}_{\boldsymbol{G}^\natural}\right]_{[:,1:r]}\left[\widetilde{\boldsymbol{S}}_{\boldsymbol{G}^\natural}^{1/2}\right]_{[1:r]}. \qquad \text{(Spectral-initialization)}$$
$$\boldsymbol{B}_0 = \sqrt{\gamma}\left[\widetilde{\boldsymbol{S}}_{\boldsymbol{G}^\natural}^{1/2}\right]_{[1:r]}\left[\widetilde{\boldsymbol{V}}_{\boldsymbol{G}^\natural}\right]_{[:,1:r]}^\top.$$

Message

If we choose (Spectral-initialization), for both linear/nonlinear models, we can directly achieve the alignment at initialization.

$$\|\boldsymbol{A}_0\boldsymbol{B}_0 - \Delta\|_{\mathrm{F}} \leq \epsilon\|\Delta\|_{op}, \quad w.p.\ 1 - C\exp(-\epsilon^2 N)$$

Key message: Algorithm design principle

- Take the SVD of $\boldsymbol{G}^\natural$: $\boldsymbol{G}^\natural = \widetilde{\boldsymbol{U}}_{\boldsymbol{G}^\natural} \widetilde{\boldsymbol{S}}_{\boldsymbol{G}^\natural} \widetilde{\boldsymbol{V}}_{\boldsymbol{G}^\natural}^\top$

$$\begin{aligned} \boldsymbol{A}_0 &= \sqrt{\gamma}\Big[\widetilde{\boldsymbol{U}}_{\boldsymbol{G}^\natural}\Big]_{[:,1:r]}\Big[\widetilde{\boldsymbol{S}}_{\boldsymbol{G}^\natural}^{1/2}\Big]_{[1:r]}. \\ \boldsymbol{B}_0 &= \sqrt{\gamma}\Big[\widetilde{\boldsymbol{S}}_{\boldsymbol{G}^\natural}^{1/2}\Big]_{[1:r]}\Big[\widetilde{\boldsymbol{V}}_{\boldsymbol{G}^\natural}\Big]_{[:,1:r]}^\top. \end{aligned} \qquad \text{(Spectral-initialization)}$$

Message

If we choose (Spectral-initialization), for both linear/nonlinear models, we can directly achieve the alignment at initialization.

$$\|\boldsymbol{A}_0\boldsymbol{B}_0 - \Delta\|_{\mathrm{F}} \le \epsilon\|\Delta\|_{op}\,, \quad w.p.\ 1 - C\exp(-\epsilon^2 N)$$

Key message: Algorithm design principle

- Take the SVD of $\boldsymbol{G}^\natural$: $\boldsymbol{G}^\natural = \widetilde{\boldsymbol{U}}_{\boldsymbol{G}^\natural} \widetilde{\boldsymbol{S}}_{\boldsymbol{G}^\natural} \widetilde{\boldsymbol{V}}_{\boldsymbol{G}^\natural}^\top$

$$\boldsymbol{A}_0 = \sqrt{\gamma}\left[\widetilde{\boldsymbol{U}}_{\boldsymbol{G}^\natural}\right]_{[:,1:r]}\left[\widetilde{\boldsymbol{S}}_{\boldsymbol{G}^\natural}^{1/2}\right]_{[1:r]}\,.$$
$$\boldsymbol{B}_0 = \sqrt{\gamma}\left[\widetilde{\boldsymbol{S}}_{\boldsymbol{G}^\natural}^{1/2}\right]_{[1:r]}\left[\widetilde{\boldsymbol{V}}_{\boldsymbol{G}^\natural}\right]_{[:,1:r]}^\top\,. \qquad \text{(Spectral-initialization)}$$

Message

If we choose (Spectral-initialization), for both linear/nonlinear models, we can directly achieve the alignment at initialization.

$$\|\boldsymbol{A}_0\boldsymbol{B}_0 - \Delta\|_{\mathrm{F}} \leq \epsilon\|\Delta\|_{op}\,, \quad w.p.\ 1 - C\exp(-\epsilon^2 N)$$

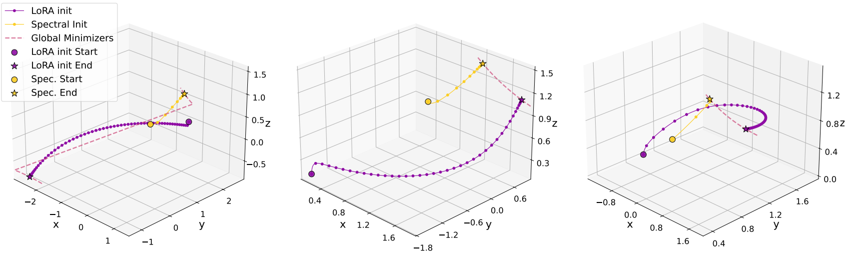
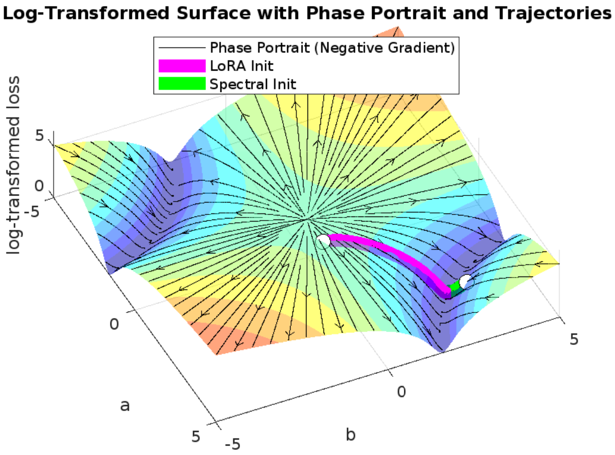


Figure 2: Comparison of the GD trajectories between LoRA and ours. $\boldsymbol{A} \in \mathbb{R}^2$ and $\boldsymbol{B} \in \mathbb{R}$. The set of global minimizers is $\{a_1^* = 2/t\,, a_2^* = 1/t\,, b^* = t \mid t \in \mathbb{R}\}$

Toy example (II): Phase portrait



One-step full gradient may suffice for low-rank fine-tuning!

Dataset	**MNLI**	**SST-2**	**CoLA**	**QNLI**	**MRPC**
Size	393k	67k	8.5k	105k	3.7k
Full	$86.33_{\pm 0.00}$	$94.75_{\pm 0.21}$	$80.70_{\pm 0.24}$	$93.19_{\pm 0.22}$	$84.56_{\pm 0.73}$
Pre-trained	-	89.79	59.03	49.28	63.48
One-step GD	-	90.48	73.00	69.13	68.38
LoRA_8	$85.30_{\pm 0.04}$	$94.04_{\pm 0.09}$	$72.84_{\pm 1.25}$	$93.02_{\pm 0.07}$	$68.38_{\pm 0.01}$

Time cost

- **CoLA** LoRA: 47s, one-step: $<$1s
- **MRPC** LoRA: 25s, one-step: $<$1s

One-step full gradient may suffice for low-rank fine-tuning!

Dataset	**MNLI**	**SST-2**	**CoLA**	**QNLI**	**MRPC**
Size	393k	67k	8.5k	105k	3.7k
Full	$86.33_{\pm 0.00}$	$94.75_{\pm 0.21}$	$80.70_{\pm 0.24}$	$93.19_{\pm 0.22}$	$84.56_{\pm 0.73}$
Pre-trained	-	89.79	59.03	49.28	63.48
One-step GD	-	90.48	73.00	69.13	68.38
LoRA_8	$85.30_{\pm 0.04}$	$94.04_{\pm 0.09}$	$72.84_{\pm 1.25}$	$93.02_{\pm 0.07}$	$68.38_{\pm 0.01}$

Time cost

- **CoLA** LoRA: 47s, one-step: $<$1s
- **MRPC** LoRA: 25s, one-step: $<$1s

Clarification on gradient alignment based work

- Motivation: make LoRA's gradients align to full fine-tuning [5]
- best-$2r$ approximation: $\operatorname{rank}(\nabla_{\boldsymbol{A}} \widetilde{L}(\boldsymbol{A}_t, \boldsymbol{B}_t)) + \operatorname{rank}(\nabla_{\boldsymbol{B}} \widetilde{L}(\boldsymbol{A}_t, \boldsymbol{B}_t)) \leq 2r$

$$\boldsymbol{A}_0 \leftarrow \left[\widetilde{\boldsymbol{U}}_{\boldsymbol{G}^\natural}\right]_{[:,1:r]}, \boldsymbol{B}_0 \leftarrow \left[\widetilde{\boldsymbol{V}}_{\boldsymbol{G}^\natural}\right]^\top_{[:,r+1:2r]}. \qquad \text{(LoRA-GA)}$$

- But! $\boldsymbol{B}_t$ will align to the right-side rank-r^* singular subspace of $\boldsymbol{G}^\natural$.

Clarification on gradient alignment based work

- Motivation: make LoRA's gradients align to full fine-tuning [5]
- best-$2r$ approximation: $\mathrm{rank}(\nabla_{\boldsymbol{A}} \widetilde{L}(\boldsymbol{A}_t, \boldsymbol{B}_t)) + \mathrm{rank}(\nabla_{\boldsymbol{B}} \widetilde{L}(\boldsymbol{A}_t, \boldsymbol{B}_t)) \leq 2r$

$$\boldsymbol{A}_0 \leftarrow \left[\widetilde{\boldsymbol{U}}_{\boldsymbol{G}^\natural}\right]_{[:,1:r]}, \boldsymbol{B}_0 \leftarrow \left[\widetilde{\boldsymbol{V}}_{\boldsymbol{G}^\natural}\right]_{[:,r+1:2r]}^\top. \qquad \text{(LoRA-GA)}$$

- But! $\boldsymbol{B}_t$ will align to the right-side rank-r^* singular subspace of $\boldsymbol{G}^\natural$.

Clarification on gradient alignment based work

- Motivation: make LoRA's gradients align to full fine-tuning [5]
- best-$2r$ approximation: $\operatorname{rank}(\nabla_{\boldsymbol{A}} \widetilde{L}(\boldsymbol{A}_t, \boldsymbol{B}_t)) + \operatorname{rank}(\nabla_{\boldsymbol{B}} \widetilde{L}(\boldsymbol{A}_t, \boldsymbol{B}_t)) \leq 2r$

$$\boldsymbol{A}_0 \leftarrow \left[\widetilde{\boldsymbol{U}}_{\boldsymbol{G}^\natural}\right]_{[:,1:r]}, \boldsymbol{B}_0 \leftarrow \left[\widetilde{\boldsymbol{V}}_{\boldsymbol{G}^\natural}\right]^\top_{[:,r+1:2r]}. \qquad \text{(LoRA-GA)}$$

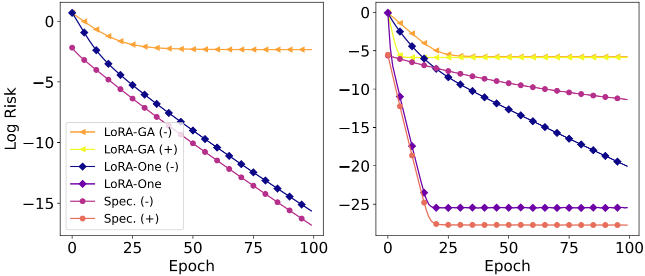
- But! $\boldsymbol{B}_t$ will align to the right-side rank-r^* singular subspace of $\boldsymbol{G}^\natural$.

Clarification on gradient alignment based work

- Motivation: make LoRA's gradients align to full fine-tuning [5]
- best-$2r$ approximation: $\operatorname{rank}(\nabla_{\boldsymbol{A}} \widetilde{L}(\boldsymbol{A}_t, \boldsymbol{B}_t)) + \operatorname{rank}(\nabla_{\boldsymbol{B}} \widetilde{L}(\boldsymbol{A}_t, \boldsymbol{B}_t)) \leq 2r$

$$\boldsymbol{A}_0 \leftarrow \left[\widetilde{\boldsymbol{U}}_{\boldsymbol{G}^\natural}\right]_{[:,1:r]}, \boldsymbol{B}_0 \leftarrow \left[\widetilde{\boldsymbol{V}}_{\boldsymbol{G}^\natural}\right]^\top_{[:,r+1:2r]}. \qquad \text{(LoRA-GA)}$$

- But! $\boldsymbol{B}_t$ will align to the right-side rank-r^* singular subspace of $\boldsymbol{G}^\natural$.



Experiments

Algorithm 1 LoRA-One training for a specific layer

Input: Pre-trained weight $\boldsymbol{W}^{\natural}$, batched data $\{\mathcal{D}_t\}_{t=1}^{T}$, LoRA rank r, LoRA alpha α, loss function L

Output: $\boldsymbol{W}^{\natural} + \frac{\alpha}{\sqrt{r}} \boldsymbol{A}_T \boldsymbol{B}_T$

Compute $\nabla_{\boldsymbol{W}} L(\boldsymbol{W}^{\natural})$ and $\boldsymbol{U}, \boldsymbol{S}, \boldsymbol{V} \leftarrow \mathrm{SVD}\Big(\nabla_{\boldsymbol{W}} L(\boldsymbol{W}^{\natural})\Big)$

$\boldsymbol{A}_0 \leftarrow \sqrt{\gamma} \cdot \boldsymbol{U}_{[:,1:r]}$

$\boldsymbol{B}_0 \leftarrow \sqrt{\gamma} \cdot \boldsymbol{V}^{\top}_{[:,1:r]}$

$\boldsymbol{W}^{\natural} \leftarrow \boldsymbol{W}^{\natural} - \frac{\alpha}{\sqrt{r}} \boldsymbol{A}_0 \boldsymbol{B}_0$

for $t = 1, \ldots, T$ **do**

$\boldsymbol{G}_t^{\boldsymbol{A}} \leftarrow \nabla_{\boldsymbol{A}} \widetilde{L}(\boldsymbol{A}_{t-1}, \boldsymbol{B}_{t-1}) \Big(\boldsymbol{B}_{t-1} \boldsymbol{B}_{t-1}^{\top} + \lambda \boldsymbol{I}_r\Big)^{-1}$

$\boldsymbol{G}_t^{\boldsymbol{B}} \leftarrow \Big(\boldsymbol{A}_{t-1}^{\top} \boldsymbol{A}_{t-1} + \lambda \boldsymbol{I}_r\Big)^{-1} \nabla_{\boldsymbol{B}} \widetilde{L}(\boldsymbol{A}_{t-1}, \boldsymbol{B}_{t-1})$

Update $\boldsymbol{A}_t, \boldsymbol{B}_t \leftarrow \mathrm{AdamW}\Big(\boldsymbol{G}_t^{\boldsymbol{A}}, \boldsymbol{G}_t^{\boldsymbol{B}}\Big)$

end

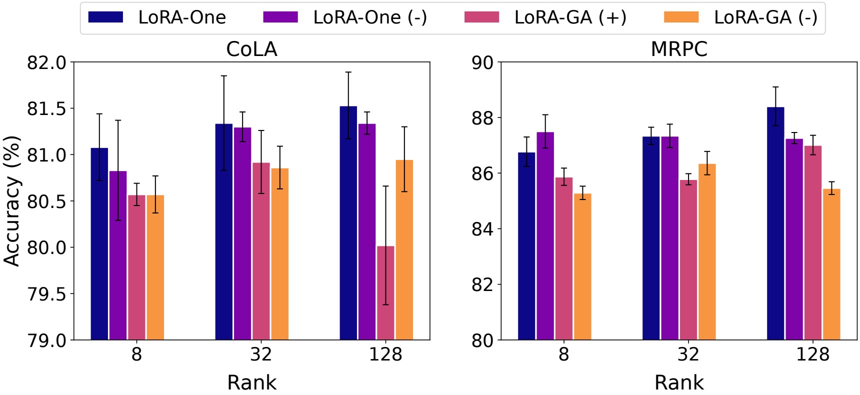
Experiments on NLP tasks from GLUE

Dataset	**MNLI**	**SST-2**	**CoLA**	**QNLI**	**MRPC**
Size	393k	67k	8.5k	105k	3.7k
Full	$86.33_{\pm 0.00}$	$94.75_{\pm 0.21}$	$80.70_{\pm 0.24}$	$93.19_{\pm 0.22}$	$84.56_{\pm 0.73}$
Pre-trained	-	89.79	59.03	49.28	63.48
One-step GD	-	90.48	73.00	69.13	68.38
LoRA_8 (Hu et al., 2022)	$85.30_{\pm 0.04}$	$94.04_{\pm 0.09}$	$72.84_{\pm 1.25}$	$93.02_{\pm 0.07}$	$68.38_{\pm 0.01}$
LoRA_{32}	$85.23_{\pm 0.11}$	$94.08_{\pm 0.05}$	$70.66_{\pm 0.41}$	$92.87_{\pm 0.05}$	$67.24_{\pm 0.58}$
LoRA_{128}	$85.53_{\pm 0.13}$	$93.96_{\pm 0.05}$	$69.45_{\pm 0.25}$	$92.91_{\pm 0.13}$	$65.36_{\pm 0.31}$
LoRA+_8 (Hayou et al., 2024)	$85.81_{\pm 0.09}$	$93.85_{\pm 0.24}$	$77.53_{\pm 0.20}$	$93.14_{\pm 0.03}$	$74.43_{\pm 1.39}$
LoRA+_{32}	$85.88_{\pm 0.16}$	$94.15_{\pm 0.25}$	$79.29_{\pm 0.96}$	$\mathbf{93.25}_{\pm 0.08}$	$79.49_{\pm 0.64}$
LoRA+_{128}	$86.07_{\pm 0.15}$	$94.08_{\pm 0.30}$	$78.59_{\pm 0.73}$	$93.06_{\pm 0.23}$	$78.76_{\pm 0.12}$
P-LoRA_8 (Zhang & Pilanci, 2024)	$85.28_{\pm 0.15}$	$93.88_{\pm 0.11}$	$79.58_{\pm 0.67}$	$93.00_{\pm 0.07}$	$83.91_{\pm 1.16}$
P-LoRA_{32}	$85.07_{\pm 0.11}$	$94.08_{\pm 0.14}$	$76.54_{\pm 1.29}$	$93.00_{\pm 0.08}$	$79.49_{\pm 0.50}$
P-LoRA_{128}	$85.38_{\pm 0.11}$	$93.96_{\pm 0.24}$	$72.04_{\pm 1.89}$	$92.98_{\pm 0.06}$	$79.66_{\pm 1.44}$
LoRA-GA_8 (Wang et al., 2024a)	$85.70_{\pm 0.09}$	$94.11_{\pm 0.18}$	$80.57_{\pm 0.20}$	$93.18_{\pm 0.06}$	$85.29_{\pm 0.24}$
LoRA-GA_{32}	$83.32_{\pm 0.10}$	$94.49_{\pm 0.32}$	$80.86_{\pm 0.23}$	$93.06_{\pm 0.14}$	$86.36_{\pm 0.42}$
LoRA-GA_{128}	$84.75_{\pm 0.06}$	$94.19_{\pm 0.14}$	$80.95_{\pm 0.35}$	$93.12_{\pm 0.11}$	$85.46_{\pm 0.23}$
LoRA-One_8 (Ours)	$\mathbf{85.81}_{\pm 0.03}$	$\mathbf{94.69}_{\pm 0.05}$	$\mathbf{81.08}_{\pm 0.36}$	$\mathbf{93.22}_{\pm 0.12}$	$\mathbf{86.77}_{\pm 0.53}$
LoRA-One_{32}	$\mathbf{86.08}_{\pm 0.01}$	$\mathbf{94.73}_{\pm 0.37}$	$\mathbf{81.34}_{\pm 0.51}$	$93.19_{\pm 0.02}$	$\mathbf{87.34}_{\pm 0.31}$
LoRA-One_{128}	$\mathbf{86.22}_{\pm 0.08}$	$\mathbf{94.65}_{\pm 0.19}$	$\mathbf{81.53}_{\pm 0.36}$	$\mathbf{93.34}_{\pm 0.11}$	$\mathbf{88.40}_{\pm 0.70}$

Experimental results on fine-tuning Llama 2-7B

	GSM8K	Human-eval
Full	$59.36_{\pm 0.85}$	$35.31_{\pm 2.13}$
LoRA_8	$46.89_{\pm 0.05}$ (**6.33h**)	$15.67_{\pm 0.60}$ (**6.75h**)
LoRA_{32}	$47.44_{\pm 0.74}$	$16.02_{\pm 0.85}$
LoRA_{128}	$47.33_{\pm 0.32}$	$15.57_{\pm 0.75}$
LoRA-GA_8	$53.60_{\pm 0.13}$	$20.45_{\pm 0.92}$
LoRA-GA_{32}	$55.12_{\pm 0.30}$	$20.18_{\pm 0.19}$
LoRA-GA_{128}	$55.07_{\pm 0.18}$	$23.05_{\pm 0.37}$
LoRA-One_8	$\mathbf{53.80}_{\pm 0.44}$ (**+0.5h**)	$\mathbf{21.02}_{\pm 0.01}$ (**+0.25h**)
LoRA-One_{32}	$\mathbf{56.61}_{\pm 0.29}$	$\mathbf{23.86}_{\pm 0.01}$
LoRA-One_{128}	$\mathbf{58.10}_{\pm 0.10}$	$\mathbf{26.79}_{\pm 0.21}$

Ablation study



- (+): with preconditioners
- (-): no preconditioners

Theory and proof...

Model	Results	Algorithm	Initialization	Conclusion
Linear	Theorem 3.1	GD	(LoRA-init)	Subspace alignment of $\boldsymbol{B}_t$
	Theorem 3.2	GD	(LoRA-init)	Subspace alignment of $\boldsymbol{A}_t$
	Proposition 3.3	GD	(Spectral-init)	$\lVert \boldsymbol{A}_0 \boldsymbol{B}_0 - \Delta \rVert_{\mathrm{F}}$ is small
	Theorem 3.5	GD	(Spectral-init)	Linear convergence of $\lVert \boldsymbol{A}_t \boldsymbol{B}_t - \Delta \rVert_{\mathrm{F}}$
	Theorem 3.6	Precondition GD	(Spectral-init)	Linear convergence rate independent of $\kappa(\Delta)$
Nonlinear	Theorem 4.3	Precondition GD	(Spectral-init)	Linear convergence rate independent of $\kappa(\Delta)$
	Theorem C.15	Smoothed Precondition GD	(Spectral-init)	Better convergence performance with less assumptions

- subspace alignment
- global convergence

Proof of sketch: Control the dynamics for alignment

$$\underbrace{\begin{bmatrix} \boldsymbol{A}_{t+1} \\ \boldsymbol{B}_{t+1}^\top \end{bmatrix}}_{:=\boldsymbol{Z}_{t+1}} = \underbrace{\begin{bmatrix} \boldsymbol{I}_d & \eta_1 \boldsymbol{G}^\natural \\ \eta_2 \boldsymbol{G}^{\natural\top} & \boldsymbol{I}_k \end{bmatrix}}_{:=\boldsymbol{H}} \underbrace{\begin{bmatrix} \boldsymbol{A}_t \\ \boldsymbol{B}_t^\top \end{bmatrix}}_{:=\boldsymbol{Z}_t} - \frac{1}{N} \begin{bmatrix} 0 & \eta_1 \widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}} \boldsymbol{A}_t \boldsymbol{B}_t \\ \eta_2 \boldsymbol{B}_t^\top \boldsymbol{A}_t^\top \widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_t \\ \boldsymbol{B}_t^\top \end{bmatrix}.$$

∘ Approximated linear dynamical system $\boldsymbol{Z}_t^{\mathtt{lin}} := \boldsymbol{H}^t \boldsymbol{Z}_0$

- Schur decomposition of $\boldsymbol{H}$
- obtain the dynamics of $\boldsymbol{Z}_t^{\mathtt{lin}}$ (decouple $\boldsymbol{A}_t^{\mathtt{lin}}$ and $\boldsymbol{B}_t^{\mathtt{lin}}$ and obtain the alignment to $\boldsymbol{G}^\natural$)
- Define the residual term $\boldsymbol{E}_t := \boldsymbol{Z}_t - \boldsymbol{Z}_t^{\mathtt{lin}}$, control $\|\boldsymbol{E}_t\|_{op}$ in early stage $t < T_1 \sim \ln\left(\frac{\|\boldsymbol{G}^\natural\|_{op}}{\|\boldsymbol{A}_0\|_{op}^2}\right)$

∘Transfer the alignment from $\boldsymbol{A}_t^{\mathtt{lin}}$ to $\boldsymbol{A}_t$ [4] (Stöger & Soltanolkotabi)

$\|\boldsymbol{U}_{r^*,\perp}^\top(\boldsymbol{G}^\natural)\boldsymbol{U}_{r^*}(\boldsymbol{A}_t)\|_{op} \lesssim \|\boldsymbol{U}_{r^*,\perp}^\top(\boldsymbol{P}_t^{\boldsymbol{A}})\boldsymbol{U}_{r^*}(\boldsymbol{P}_t^{\boldsymbol{A}}\boldsymbol{A}_0 + \boldsymbol{E}_t)\|_{op}$ is small , $w.h.p.$

Proof of sketch: Control the dynamics for alignment

$$\underbrace{\begin{bmatrix} \boldsymbol{A}_{t+1} \\ \boldsymbol{B}_{t+1}^\top \end{bmatrix}}_{:=\boldsymbol{Z}_{t+1}} = \underbrace{\begin{bmatrix} \boldsymbol{I}_d & \eta_1 \boldsymbol{G}^\natural \\ \eta_2 \boldsymbol{G}^{\natural\top} & \boldsymbol{I}_k \end{bmatrix}}_{:=\boldsymbol{H}} \underbrace{\begin{bmatrix} \boldsymbol{A}_t \\ \boldsymbol{B}_t^\top \end{bmatrix}}_{:=\boldsymbol{Z}_t} - \frac{1}{N} \begin{bmatrix} 0 & \eta_1 \widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}} \boldsymbol{A}_t \boldsymbol{B}_t \\ \eta_2 \boldsymbol{B}_t^\top \boldsymbol{A}_t^\top \widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_t \\ \boldsymbol{B}_t^\top \end{bmatrix}.$$

∘ Approximated linear dynamical system $\boldsymbol{Z}_t^{\texttt{lin}} := \boldsymbol{H}^t \boldsymbol{Z}_0$

- Schur decomposition of $\boldsymbol{H}$
- obtain the dynamics of $\boldsymbol{Z}_t^{\texttt{lin}}$ (decouple $\boldsymbol{A}_t^{\texttt{lin}}$ and $\boldsymbol{B}_t^{\texttt{lin}}$ and obtain the alignment to $\boldsymbol{G}^\natural$)
- Define the residual term $\boldsymbol{E}_t := \boldsymbol{Z}_t - \boldsymbol{Z}_t^{\texttt{lin}}$, control $\|\boldsymbol{E}_t\|_{op}$ in early stage $t < T_1 \sim \ln\left(\frac{\|\boldsymbol{G}^\natural\|_{op}}{\|\boldsymbol{A_0}\|_{op}^{\boldsymbol{2}}}\right)$

∘Transfer the alignment from $\boldsymbol{A}_t^{\texttt{lin}}$ to $\boldsymbol{A}_t$ [4] (Stöger & Soltanolkotabi)

$$\|\boldsymbol{U}_{r^*,\perp}^\top(\boldsymbol{G}^\natural)\boldsymbol{U}_{r^*}(\boldsymbol{A}_t)\|_{op} \lesssim \|\boldsymbol{U}_{r^*,\perp}^\top(\boldsymbol{P}_t^{\boldsymbol{A}})\boldsymbol{U}_{r^*}(\boldsymbol{P}_t^{\boldsymbol{A}}\boldsymbol{A}_0 + \boldsymbol{E}_t)\|_{op} \text{ is small}, \ w.h.p.$$

Proof of sketch: Control the dynamics for alignment

$$\underbrace{\begin{bmatrix} \boldsymbol{A}_{t+1} \\ \boldsymbol{B}_{t+1}^\top \end{bmatrix}}_{:=\boldsymbol{Z}_{t+1}} = \underbrace{\begin{bmatrix} \boldsymbol{I}_d & \eta_1 \boldsymbol{G}^\natural \\ \eta_2 \boldsymbol{G}^{\natural\top} & \boldsymbol{I}_k \end{bmatrix}}_{:=\boldsymbol{H}} \underbrace{\begin{bmatrix} \boldsymbol{A}_t \\ \boldsymbol{B}_t^\top \end{bmatrix}}_{:=\boldsymbol{Z}_t} - \frac{1}{N} \begin{bmatrix} 0 & \eta_1 \widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}} \boldsymbol{A}_t \boldsymbol{B}_t \\ \eta_2 \boldsymbol{B}_t^\top \boldsymbol{A}_t^\top \widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_t \\ \boldsymbol{B}_t^\top \end{bmatrix}.$$

◦ Approximated linear dynamical system $\boldsymbol{Z}_t^{\texttt{lin}} := \boldsymbol{H}^t \boldsymbol{Z}_0$

- Schur decomposition of $\boldsymbol{H}$
- obtain the dynamics of $\boldsymbol{Z}_t^{\texttt{lin}}$ (decouple $\boldsymbol{A}_t^{\texttt{lin}}$ and $\boldsymbol{B}_t^{\texttt{lin}}$ and obtain the alignment to $\boldsymbol{G}^\natural$)
- Define the residual term $\boldsymbol{E}_t := \boldsymbol{Z}_t - \boldsymbol{Z}_t^{\texttt{lin}}$, control $\|\boldsymbol{E}_t\|_{op}$ in early stage $t < T_1 \sim \ln\left(\frac{\|\boldsymbol{G}^\natural\|_{op}}{\|\boldsymbol{A}_0\|_{op}^2}\right)$

◦Transfer the alignment from $\boldsymbol{A}_t^{\texttt{lin}}$ to $\boldsymbol{A}_t$ [4] (Stöger & Soltanolkotabi)

$$\|\boldsymbol{U}_{r^*,\perp}^\top(\boldsymbol{G}^\natural)\boldsymbol{U}_{r^*}(\boldsymbol{A}_t)\|_{op} \lesssim \|\boldsymbol{U}_{r^*,\perp}^\top(\boldsymbol{P}_t^{\boldsymbol{A}})\boldsymbol{U}_{r^*}(\boldsymbol{P}_t^{\boldsymbol{A}}\boldsymbol{A}_0 + \boldsymbol{E}_t)\|_{op} \text{ is small}, \ w.h.p.$$

Global convergence on nonlinear models

Recall problem setting and assumptions for nonlinear model

- Pre-trained model $f_{\mathrm{pre}}(\boldsymbol{x}) = \sigma[(\boldsymbol{x}^\top \boldsymbol{W}^\natural)^\top] \in \mathbb{R}^k$
- Unknown low-rank feature shift Δ: $\widetilde{\boldsymbol{W}}^\natural := \boldsymbol{W}^\natural + \Delta$ with $\mathrm{Rank}(\Delta) = r^*$
- We assume $r = r^*$.
- Downstream well-behaved data $\widetilde{\boldsymbol{y}} = \sigma[(\widetilde{\boldsymbol{x}}^\top \widetilde{\boldsymbol{W}}^\natural)^\top]$, $\{\widetilde{\boldsymbol{x}}_i\}_{i=1}^N \overset{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{I}_d)$
- training loss

$$\widetilde{L}(\boldsymbol{A}, \boldsymbol{B}) := \frac{1}{2N}\left\| \sigma\left(\widetilde{\boldsymbol{X}}(\boldsymbol{W}^\natural + \boldsymbol{A}\boldsymbol{B}) \right) - \widetilde{\boldsymbol{Y}} \right\|_{\mathrm{F}}^2 .$$

- gradient updates

$$\nabla_{\boldsymbol{A}} \widetilde{L}(\boldsymbol{A}_t, \boldsymbol{B}_t) = -\boldsymbol{J}_{\boldsymbol{W}_t} \boldsymbol{B}_t^\top, \quad \nabla_{\boldsymbol{B}} \widetilde{L}(\boldsymbol{A}_t, \boldsymbol{B}_t) = -\boldsymbol{A}_t^\top \boldsymbol{J}_{\boldsymbol{W}_t},$$

where we define

$$\boldsymbol{J}_{\boldsymbol{W}_t} := \underbrace{\frac{1}{N}\widetilde{\boldsymbol{X}}^\top \left[\sigma(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{W}}^\natural) - \frac{1}{N}\widetilde{\boldsymbol{X}}^\top \sigma(\widetilde{\boldsymbol{X}}\boldsymbol{W}_t) \right]}_{\boldsymbol{J}^{\mathrm{GLM}}_{\boldsymbol{W}_t}} \odot \sigma'(\widetilde{\boldsymbol{X}}\boldsymbol{W}_t) .$$

- GLM-tron style: [3, 6]

Recall problem setting and assumptions for nonlinear model

- Pre-trained model $f_{\text{pre}}(\boldsymbol{x}) = \sigma[(\boldsymbol{x}^\top \boldsymbol{W}^\natural)^\top] \in \mathbb{R}^k$
- Unknown low-rank feature shift Δ: $\widetilde{\boldsymbol{W}}^\natural := \boldsymbol{W}^\natural + \Delta$ with $\text{Rank}(\Delta) = r^*$
- We assume $r = r^*$.
- Downstream well-behaved data $\widetilde{\boldsymbol{y}} = \sigma[(\widetilde{\boldsymbol{x}}^\top \widetilde{\boldsymbol{W}}^\natural)^\top]$, $\{\widetilde{\boldsymbol{x}}_i\}_{i=1}^N \overset{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{I}_d)$
- training loss

$$\widetilde{L}(\boldsymbol{A}, \boldsymbol{B}) := \frac{1}{2N}\left\|\sigma\Big(\widetilde{\boldsymbol{X}}(\boldsymbol{W}^\natural + \boldsymbol{A}\boldsymbol{B})\Big) - \widetilde{\boldsymbol{Y}}\right\|_{\mathrm{F}}^2 .$$

- gradient updates

$$\nabla_{\boldsymbol{A}} \widetilde{L}(\boldsymbol{A}_t, \boldsymbol{B}_t) = -\boldsymbol{J}_{\boldsymbol{W}_t} \boldsymbol{B}_t^\top, \quad \nabla_{\boldsymbol{B}} \widetilde{L}(\boldsymbol{A}_t, \boldsymbol{B}_t) = -\boldsymbol{A}_t^\top \boldsymbol{J}_{\boldsymbol{W}_t},$$

where we define

$$\boldsymbol{J}_{\boldsymbol{W}_t} := \underbrace{\frac{1}{N}\widetilde{\boldsymbol{X}}^\top \left[\sigma(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{W}}^\natural) - \frac{1}{N}\widetilde{\boldsymbol{X}}^\top \sigma(\widetilde{\boldsymbol{X}}\boldsymbol{W}_t)\right]}_{\boldsymbol{J}^{\mathrm{GLM}}_{\boldsymbol{W}_t}} \odot \sigma'(\widetilde{\boldsymbol{X}}\boldsymbol{W}_t) .$$

- GLM-tron style: [3, 6]

Global convergence

Theorem (Linear convergence rate)

Under (Spectral-initialization) and $\boldsymbol{J}^{\mathrm{GLM}}_{\boldsymbol{W}_t}$ for gradient update (adding preconditioners), choose constant step-size $\eta < 1$, we have

$$\|\boldsymbol{A}_t\boldsymbol{B}_t - \Delta\|_{\mathrm{F}} \lesssim \left(1 - \frac{\eta}{4}\right)^t \lambda_{r^*}(\Delta)\,, w.h.p$$

- holds for standard gradient update, but requires more assumptions.
- $\|\boldsymbol{A}_0\boldsymbol{B}_0 - \Delta\|_{\mathrm{F}} \leq \epsilon\|\Delta\|_{op}, w.h.p.$

A bit proof sketch at

- recover at initialization:

  $\|\boldsymbol{A}_0\boldsymbol{B}_0 - \Delta\|_{\mathrm{F}} \leq \|\boldsymbol{A}_0\boldsymbol{B}_0 - \gamma\boldsymbol{G}^\natural\|_{\mathrm{F}} +$ concentration on $\boldsymbol{G}^\natural + \rho\lambda^*_{r^*}, w.h.p$
- $\mathbb{E}_{\widetilde{\boldsymbol{x}}}\left[-\boldsymbol{J}^{\mathrm{GLM}}_{\boldsymbol{W}_t}\right] = \frac{1}{2}(\boldsymbol{A}_t\boldsymbol{B}_t - \Delta)$ by Stein's lemma $\Rightarrow \mathbb{E}_{\widetilde{\boldsymbol{x}}}[\boldsymbol{G}^\natural] = \mathbb{E}_{\widetilde{\boldsymbol{x}}}\left[\boldsymbol{J}^{\mathrm{GLM}}_{\boldsymbol{W}^\natural}\right] = \frac{1}{2}\Delta$
- concentration:

$$\left\|\boldsymbol{J}^{\mathrm{GLM}}_{\boldsymbol{W}_t} - \mathbb{E}_{\widetilde{\boldsymbol{x}}}\left[\boldsymbol{J}^{\mathrm{GLM}}_{\boldsymbol{W}_t}\right]\right\|_{\mathrm{F}} \lesssim \sqrt{d}\epsilon\|\boldsymbol{A}_t\boldsymbol{B}_t - \Delta\|_{\mathrm{F}}\,, w.h.p. \Rightarrow \text{control}\,\boldsymbol{G}^\natural$$

Global convergence

Theorem (Linear convergence rate)

Under (Spectral-initialization) and $\boldsymbol{J}^{\mathrm{GLM}}_{\boldsymbol{W}_t}$ for gradient update (adding preconditioners), choose constant step-size $\eta < 1$, we have

$$\|\boldsymbol{A}_t \boldsymbol{B}_t - \Delta\|_{\mathrm{F}} \lesssim \left(1 - \frac{\eta}{4}\right)^t \lambda_{r^*}(\Delta)\,, w.h.p$$

- holds for standard gradient update, but requires more assumptions.
- $\|\boldsymbol{A}_0 \boldsymbol{B}_0 - \Delta\|_{\mathrm{F}} \leq \epsilon \|\Delta\|_{op}, w.h.p.$

A bit proof sketch at

- recover at initialization:
  $\|\boldsymbol{A}_0 \boldsymbol{B}_0 - \Delta\|_{\mathrm{F}} \leq \|\boldsymbol{A}_0 \boldsymbol{B}_0 - \gamma \boldsymbol{G}^\natural\|_{\mathrm{F}} +$ concentration on $\boldsymbol{G}^\natural + \rho \lambda^*_{r^*}, w.h.p$
- $\mathbb{E}_{\widetilde{\boldsymbol{x}}}\left[-\boldsymbol{J}^{\mathrm{GLM}}_{\boldsymbol{W}_t}\right] = \frac{1}{2}(\boldsymbol{A}_t \boldsymbol{B}_t - \Delta)$ by Stein's lemma $\Rightarrow \mathbb{E}_{\widetilde{\boldsymbol{x}}}[\boldsymbol{G}^\natural] = \mathbb{E}_{\widetilde{\boldsymbol{x}}}\left[\boldsymbol{J}^{\mathrm{GLM}}_{\boldsymbol{W}^\natural}\right] = \frac{1}{2}\Delta$
- concentration:

$$\left\|\boldsymbol{J}^{\mathrm{GLM}}_{\boldsymbol{W}_t} - \mathbb{E}_{\widetilde{\boldsymbol{x}}}\left[\boldsymbol{J}^{\mathrm{GLM}}_{\boldsymbol{W}_t}\right]\right\|_{\mathrm{F}} \lesssim \sqrt{d}\epsilon \|\boldsymbol{A}_t \boldsymbol{B}_t - \Delta\|_{\mathrm{F}}\,, w.h.p. \Rightarrow \text{control}\, \boldsymbol{G}^\natural$$

Global convergence

Theorem (Linear convergence rate)

Under (Spectral-initialization) and $J^{\mathrm{GLM}}_{W_t}$ for gradient update (adding preconditioners), choose constant step-size $\eta < 1$, we have

$$\|A_t B_t - \Delta\|_{\mathrm{F}} \lesssim \left(1 - \frac{\eta}{4}\right)^t \lambda_{r^*}(\Delta)\,, w.h.p$$

- holds for standard gradient update, but requires more assumptions.
- $\|A_0 B_0 - \Delta\|_{\mathrm{F}} \leq \epsilon \|\Delta\|_{op}, w.h.p.$

A bit proof sketch at

- recover at initialization:
  $\|A_0 B_0 - \Delta\|_{\mathrm{F}} \leq \|A_0 B_0 - \gamma G^\natural\|_{\mathrm{F}} +$ concentration on $G^\natural + \rho \lambda^*_{r^*}, w.h.p$
- $\mathbb{E}_{\widetilde{x}}\left[-J^{\mathrm{GLM}}_{W_t}\right] = \frac{1}{2}(A_t B_t - \Delta)$ by Stein's lemma $\Rightarrow \mathbb{E}_{\widetilde{x}}[G^\natural] = \mathbb{E}_{\widetilde{x}}\left[J^{\mathrm{GLM}}_{W^\natural}\right] = \frac{1}{2}\Delta$
- concentration:
  $$\left\|J^{\mathrm{GLM}}_{W_t} - \mathbb{E}_{\widetilde{x}}\left[J^{\mathrm{GLM}}_{W_t}\right]\right\|_{\mathrm{F}} \lesssim \sqrt{d}\epsilon \|A_t B_t - \Delta\|_{\mathrm{F}}\,, w.h.p. \Rightarrow \text{control} G^\natural$$

Global convergence

Theorem (Linear convergence rate)

Under (Spectral-initialization) and $\boldsymbol{J}^{\mathtt{GLM}}_{\boldsymbol{W}_t}$ for gradient update (adding preconditioners), choose constant step-size $\eta < 1$, we have

$$\|\boldsymbol{A}_t \boldsymbol{B}_t - \Delta\|_{\mathrm{F}} \lesssim \left(1 - \frac{\eta}{4}\right)^t \lambda_{r^*}(\Delta)\,, w.h.p$$

- holds for standard gradient update, but requires more assumptions.
- $\|\boldsymbol{A}_0 \boldsymbol{B}_0 - \Delta\|_{\mathrm{F}} \leq \epsilon \|\Delta\|_{op}, w.h.p.$

A bit proof sketch at

- recover at initialization:
  $\|\boldsymbol{A}_0 \boldsymbol{B}_0 - \Delta\|_{\mathrm{F}} \leq \|\boldsymbol{A}_0 \boldsymbol{B}_0 - \gamma \boldsymbol{G}^\natural\|_{\mathrm{F}} +$ concentration on $\boldsymbol{G}^\natural + \rho\lambda^*_{r^*}, w.h.p$
- $\mathbb{E}_{\widetilde{\boldsymbol{x}}}\left[-\boldsymbol{J}^{\mathtt{GLM}}_{\boldsymbol{W}_t}\right] = \frac{1}{2}(\boldsymbol{A}_t \boldsymbol{B}_t - \Delta)$ by Stein's lemma $\Rightarrow \mathbb{E}_{\widetilde{\boldsymbol{x}}}[\boldsymbol{G}^\natural] = \mathbb{E}_{\widetilde{\boldsymbol{x}}}\left[\boldsymbol{J}^{\mathtt{GLM}}_{\boldsymbol{W}^\natural}\right] = \frac{1}{2}\Delta$
- concentration:
  $$\left\|\boldsymbol{J}^{\mathtt{GLM}}_{\boldsymbol{W}_t} - \mathbb{E}_{\widetilde{\boldsymbol{x}}}\left[\boldsymbol{J}^{\mathtt{GLM}}_{\boldsymbol{W}_t}\right]\right\|_{\mathrm{F}} \lesssim \sqrt{d}\epsilon \|\boldsymbol{A}_t \boldsymbol{B}_t - \Delta\|_{\mathrm{F}}\,, w.h.p. \Rightarrow \text{control} \boldsymbol{G}^\natural$$

Global convergence

Theorem (Linear convergence rate)

Under (Spectral-initialization) and $\boldsymbol{J}_{\boldsymbol{W}_t}^{\mathrm{GLM}}$ for gradient update (adding preconditioners), choose constant step-size $\eta < 1$, we have

$$\|\boldsymbol{A}_t\boldsymbol{B}_t - \Delta\|_{\mathrm{F}} \lesssim \left(1 - \frac{\eta}{4}\right)^t \lambda_{r^*}(\Delta)\,, w.h.p$$

- holds for standard gradient update, but requires more assumptions.
- $\|\boldsymbol{A}_0\boldsymbol{B}_0 - \Delta\|_{\mathrm{F}} \leq \epsilon\|\Delta\|_{op}, w.h.p.$

A bit proof sketch at

- recover at initialization:
  $\|\boldsymbol{A}_0\boldsymbol{B}_0 - \Delta\|_{\mathrm{F}} \leq \|\boldsymbol{A}_0\boldsymbol{B}_0 - \gamma\boldsymbol{G}^\natural\|_{\mathrm{F}} +$ concentration on $\boldsymbol{G}^\natural + \rho\lambda_{r^*}^*, w.h.p$
- $\mathbb{E}_{\widetilde{\boldsymbol{x}}}\left[-\boldsymbol{J}_{\boldsymbol{W}_t}^{\mathrm{GLM}}\right] = \frac{1}{2}(\boldsymbol{A}_t\boldsymbol{B}_t - \Delta)$ by Stein's lemma $\Rightarrow \mathbb{E}_{\widetilde{\boldsymbol{x}}}[\boldsymbol{G}^\natural] = \mathbb{E}_{\widetilde{\boldsymbol{x}}}\left[\boldsymbol{J}_{\boldsymbol{W}^\natural}^{\mathrm{GLM}}\right] = \frac{1}{2}\Delta$
- concentration:

$$\left\|\boldsymbol{J}_{\boldsymbol{W}_t}^{\mathrm{GLM}} - \mathbb{E}_{\widetilde{\boldsymbol{x}}}\left[\boldsymbol{J}_{\boldsymbol{W}_t}^{\mathrm{GLM}}\right]\right\|_{\mathrm{F}} \lesssim \sqrt{d}\epsilon\|\boldsymbol{A}_t\boldsymbol{B}_t - \Delta\|_{\mathrm{F}}\,, w.h.p. \Rightarrow \text{control}\boldsymbol{G}^\natural$$

Proof of sketch on $\boldsymbol{A}_t\boldsymbol{B}_t - \Delta$

$$
\begin{aligned}
\left\|\boldsymbol{A}_{t+1}\boldsymbol{B}_{t+1} - \Delta\right\|_{\mathtt{F}} \lesssim & \left\|\boldsymbol{J}^{\mathtt{GLM}}_{\boldsymbol{W}_t} - c_{\mathtt{H}}(\boldsymbol{A}_t\boldsymbol{B}_t - \Delta)\right\|_{\mathtt{F}} \ \textcolor{red}{\text{[concentration+Hermite]}} \\
& + (1-\eta)\left\|\boldsymbol{U}_{\boldsymbol{A}_t}\boldsymbol{U}_{\boldsymbol{A}_t}^{\top}(\boldsymbol{A}_t\boldsymbol{B}_t - \Delta)\boldsymbol{V}_{\boldsymbol{B}_t}\boldsymbol{V}_{\boldsymbol{B}_t}^{\top}\right\|_{\mathtt{F}} \\
& + \left\|\left(\boldsymbol{I}_d - \boldsymbol{U}_{\boldsymbol{A}_t}\boldsymbol{U}_{\boldsymbol{A}_t}^{\top}\right)(\boldsymbol{A}_t\boldsymbol{B}_t - \Delta)\left(\boldsymbol{I}_k - \boldsymbol{V}_{\boldsymbol{B}_t}\boldsymbol{V}_{\boldsymbol{B}_t}^{\top}\right)\right\|_{\mathtt{F}} \\
& + \text{other projections}
\end{aligned}
$$

$$
\boldsymbol{L} = \begin{bmatrix} \boldsymbol{U}_{\boldsymbol{A}_t} & 0_{d\times r} \\ 0_{k\times r} & \boldsymbol{V}_{\boldsymbol{B}_t} \end{bmatrix} \in \mathbb{R}^{(d+k)\times 2r},
$$

then $\boldsymbol{L}\boldsymbol{L}^{\top}$ is a projection matrix, $\boldsymbol{I}_{d+k} - \boldsymbol{L}\boldsymbol{L}^{\top} = \boldsymbol{L}_{\perp}\boldsymbol{L}_{\perp}^{\top}$

- transformed to lower bound $\left\|\boldsymbol{L}_{\perp}^{\top}\boldsymbol{\Delta}\boldsymbol{L}\right\|_{\mathtt{F}}^2$
- upper bound $\left\|\boldsymbol{L}_{\perp}^{\top}\boldsymbol{U}\right\|_{op} < 1$ by Wedin's sin-θ theorem

Proof of sketch on $A_t B_t - \Delta$

$$
\begin{aligned}
\|\boldsymbol{A}_{t+1}\boldsymbol{B}_{t+1} - \Delta\|_{\mathrm{F}} \lesssim{} & \|\boldsymbol{J}^{\mathtt{GLM}}_{\boldsymbol{W}_t} - c_{\mathrm{H}}(\boldsymbol{A}_t\boldsymbol{B}_t - \Delta)\|_{\mathrm{F}} \ \text{[concentration+Hermite]} \\
& + (1-\eta)\left\|\boldsymbol{U}_{\boldsymbol{A}_t}\boldsymbol{U}_{\boldsymbol{A}_t}^\top(\boldsymbol{A}_t\boldsymbol{B}_t - \Delta)\boldsymbol{V}_{\boldsymbol{B}_t}\boldsymbol{V}_{\boldsymbol{B}_t}^\top\right\|_{\mathrm{F}} \\
& + \left\|\left(\boldsymbol{I}_d - \boldsymbol{U}_{\boldsymbol{A}_t}\boldsymbol{U}_{\boldsymbol{A}_t}^\top\right)(\boldsymbol{A}_t\boldsymbol{B}_t - \Delta)\left(\boldsymbol{I}_k - \boldsymbol{V}_{\boldsymbol{B}_t}\boldsymbol{V}_{\boldsymbol{B}_t}^\top\right)\right\|_{\mathrm{F}} \\
& + \text{other projections}
\end{aligned}
$$

$$
\boldsymbol{L} = \begin{bmatrix} \boldsymbol{U}_{\boldsymbol{A}_t} & 0_{d\times r} \\ 0_{k\times r} & \boldsymbol{V}_{\boldsymbol{B}_t} \end{bmatrix} \in \mathbb{R}^{(d+k)\times 2r},
$$

then $\boldsymbol{L}\boldsymbol{L}^\top$ is a projection matrix, $\boldsymbol{I}_{d+k} - \boldsymbol{L}\boldsymbol{L}^\top = \boldsymbol{L}_\perp\boldsymbol{L}_\perp^\top$

- transformed to lower bound $\left\|\boldsymbol{L}_\perp^\top \boldsymbol{\Delta} \boldsymbol{L}\right\|_{\mathrm{F}}^2$
- upper bound $\left\|\boldsymbol{L}_\perp^\top \boldsymbol{U}\right\|_{op} < 1$ by Wedin's sin-θ theorem

Takeaway messages

- arXiv: 2502.01235 and code

Model	Results	Algorithm	Initialization	Conclusion
Linear	Theorem 3.1	GD	(LoRA-init)	Subspace alignment of $\boldsymbol{B}_t$
	Theorem 3.2	GD	(LoRA-init)	Subspace alignment of $\boldsymbol{A}_t$
	Proposition 3.3	GD	(Spectral-init)	$\\|\boldsymbol{A}_0\boldsymbol{B}_0 - \Delta\\|_{\mathrm{F}}$ is small
	Theorem 3.5	GD	(Spectral-init)	Linear convergence of $\\|\boldsymbol{A}_t\boldsymbol{B}_t - \Delta\\|_{\mathrm{F}}$
	Theorem 3.6	Precondition GD	(Spectral-init)	Linear convergence rate independent of $\kappa(\Delta)$
Nonlinear	Theorem 4.3	Precondition GD	(Spectral-init)	Linear convergence rate independent of $\kappa(\Delta)$
	Theorem C.15	Smoothed Precondition GD	(Spectral-init)	Better convergence performance with less assumptions

- subspace alignment: $\boldsymbol{G}^{\natural}$ and $(\boldsymbol{A}_t, \boldsymbol{B}_t) \Rightarrow$ theory-grounded algorithm design
- clarification on gradient alignment based algorithms

Target

- How to handle **nonlinearity** at a theoretical level (e.g., training dynamics)
- How to precisely and efficiently approximate **nonlinearity** at a practical level under theoretical guidelines

Takeaway messages

- arXiv: 2502.01235 and code

Model	Results	Algorithm	Initialization	Conclusion
Linear	Theorem 3.1	GD	(LoRA-init)	Subspace alignment of $\boldsymbol{B}_t$
	Theorem 3.2	GD	(LoRA-init)	Subspace alignment of $\boldsymbol{A}_t$
	Proposition 3.3	GD	(Spectral-init)	$\\|\boldsymbol{A}_0\boldsymbol{B}_0 - \Delta\\|_{\mathrm{F}}$ is small
	Theorem 3.5	GD	(Spectral-init)	Linear convergence of $\\|\boldsymbol{A}_t\boldsymbol{B}_t - \Delta\\|_{\mathrm{F}}$
	Theorem 3.6	Precondition GD	(Spectral-init)	Linear convergence rate independent of $\kappa(\Delta)$
Nonlinear	Theorem 4.3	Precondition GD	(Spectral-init)	Linear convergence rate independent of $\kappa(\Delta)$
	Theorem C.15	Smoothed Precondition GD	(Spectral-init)	Better convergence performance with less assumptions

- subspace alignment: $\boldsymbol{G}^\natural$ and $(\boldsymbol{A}_t, \boldsymbol{B}_t) \Rightarrow$ theory-grounded algorithm design
- clarification on gradient alignment based algorithms

Target

- How to handle **nonlinearity** at a theoretical level (e.g., training dynamics)
- How to precisely and efficiently approximate **nonlinearity** at a practical level under theoretical guidelines

Soufiane Hayou, Nikhil Ghosh, and Bin Yu.
LoRA+: Efficient low rank adaptation of large models.
arXiv preprint arXiv:2402.12354, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen.
LoRA: Low-rank adaptation of large language models.
In *International Conference on Learning Representations*, 2022.

Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai.
Efficient learning of generalized linear and single index models with isotonic regression.
In *Advances in Neural Information Processing Systems*, 2011.

Dominik Stöger and Mahdi Soltanolkotabi.
Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction.
In *Advances in Neural Information Processing Systems*, pages 23831–23843, 2021.

Shaowen Wang, Linxi Yu, and Jian Li.
LoRA-GA: Low-rank adaptation with gradient approximation.
In *Advances in Neural Information Processing Systems*, 2024.

Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Sham M Kakade.
Finite-sample analysis of learning high-dimensional single relu neuron.
In *International Conference on Machine Learning*, pages 37919–37951, 2023.