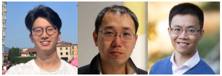
Bridge theory to practice: One-step full gradient can suffice for low-rank fine-tuning, provably and efficiently

Fanghui Liu
fanghui.liu@warwick.ac.uk

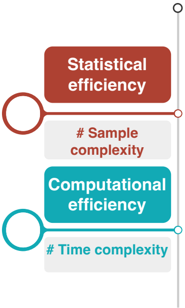
Department of Computer Science, University of Warwick, UK
Centre for Discrete Mathematics and its Applications (DIMAP), Warwick
[joint work with Yuanhe Zhang (Warwick) and Yudong Chen (UW-Madison)]



My research

❐ Research interests

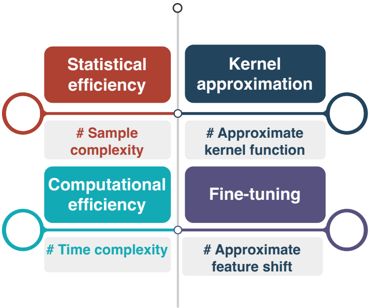
- Foundations of machine learning (ML)
- Theory-grounded efficient algorithm design
- Trustworthy ML



My research

❐ Research interests

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



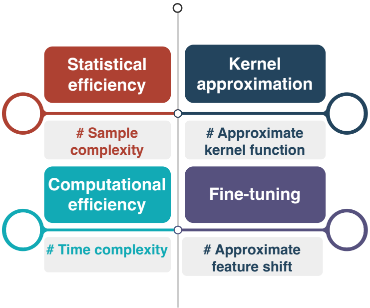
My research

- ❐ Research interests

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

- ❐ Research goal

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



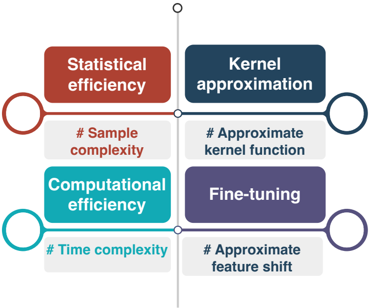
My research

❒ 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)

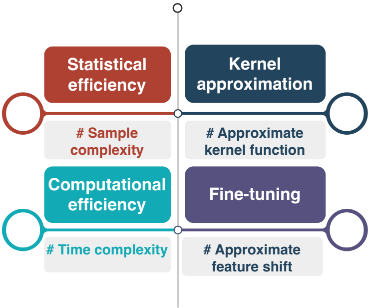
Machine learning works in **high dimensions** that can be a **curse**!

— 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

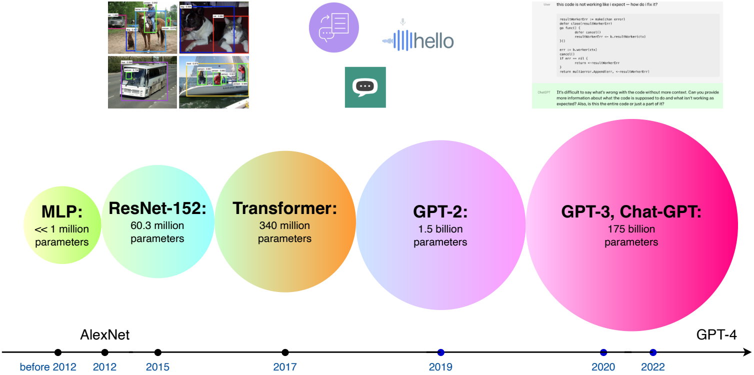
Learning efficiency (Curse of Dimensionality, CoD)

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

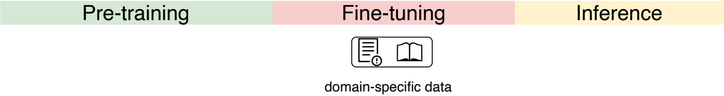


From pre-training to (parameter-efficient) fine-tuning

- GPT3: 175 billion parameters
- Llama3.1: $>$ 400 billion parameters
- Gemini 1.5 Pro 300–500 billion parameters (unconfirmed)
- Deepseek-v3: $>$ 600 billion parameters
- Llama 4 Behemoth: $>$ 2,000 billion parameters

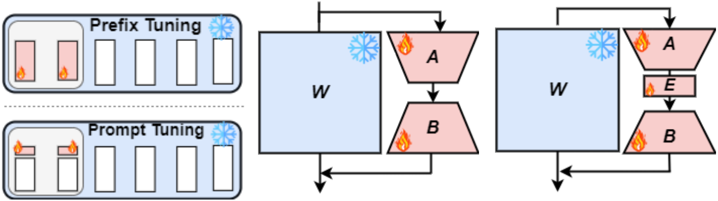
From pre-training to (parameter-efficient) fine-tuning

- GPT3: 175 billion parameters
- Llama3.1: $>$ 400 billion parameters
- Gemini 1.5 Pro 300–500 billion parameters (unconfirmed)
- Deepseek-v3: $>$ 600 billion parameters
- Llama 4 Behemoth: $>$ 2,000 billion parameters



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

- GPT3: 175 billion parameters
- Llama3.1: $>$ 400 billion parameters
- Gemini 1.5 Pro 300–500 billion parameters (unconfirmed)
- Deepseek-v3: $>$ 600 billion parameters
- Llama 4 Behemoth: $>$ 2,000 billion parameters



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

Low-rank adaption (LoRA) for fine-tuning [1]

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

LoRA

- Formulation:

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

- Initialization:

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

How theory guides practice (not limited to understanding)

- design new algorithm -> performance improvement (accuracy, efficiency)
- clarify some misconceptions in algorithm design

Low-rank adaption (LoRA) for fine-tuning [1]

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

LoRA

- Formulation:

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

- Initialization:

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

How theory guides practice (not limited to understanding)

- design new algorithm -> performance improvement (accuracy, efficiency)
- clarify some misconceptions in algorithm design

Low-rank adaption (LoRA) for fine-tuning [1]

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

LoRA

- Formulation:

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

- Initialization:

$$[A_0]_{ij} \sim \mathcal{N}(0, \alpha^2) \quad \text{and} \quad [B_0]_{ij} = 0\,, \quad \alpha > 0\,. \tag{LoRA-init.}$$

How theory guides practice (not limited to understanding)

- design new algorithm -> performance improvement (accuracy, efficiency)
- clarify some misconceptions in algorithm design

Motivation: non-linear dynamics and subspace alignment

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

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

- $\boldsymbol{G}^\natural$: 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...

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

- $\boldsymbol{G}^\natural$: 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...

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

- $\boldsymbol{G}^\natural$: 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...

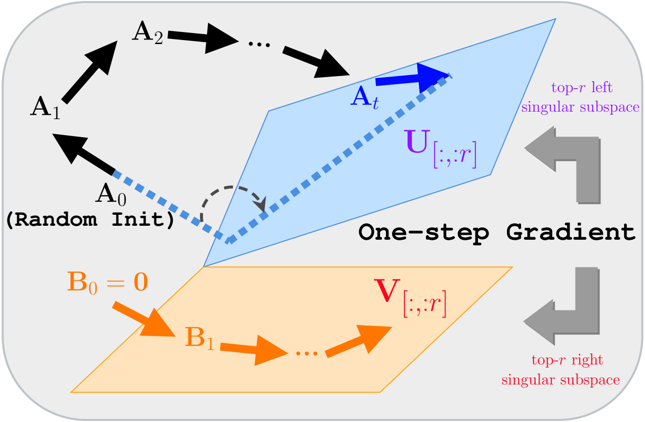
$$\begin{bmatrix} \boldsymbol{A}_{t+1} \\ \boldsymbol{B}_{t+1}^\top \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_d & \eta \boldsymbol{G}^\natural \\ \eta \boldsymbol{G}^{\natural\top} & \boldsymbol{I}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_t \\ \boldsymbol{B}_t^\top \end{bmatrix} + \text{nonlinear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

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

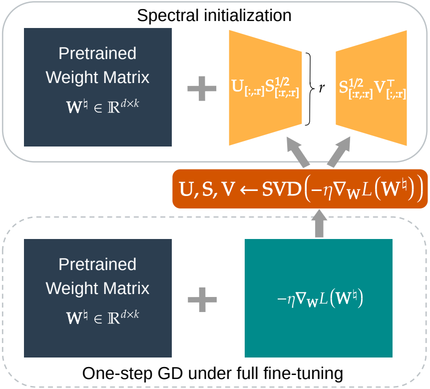
Target

- *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


$\mathbf{A}_2$
...
$\mathbf{A}_1$
$\mathbf{A}_t$
top-r left
singular subspace
$\mathbf{U}_{[:,:r]}$
$\mathbf{A}_0$
(Random Init)
One-step Gradient
$\mathrm{B}_0 = \mathbf{0}$
$\mathbf{V}_{[:,:r]}$
B_1
...
top-r right
singular subspace

Pipeline


Spectral initialization
Pretrained
Weight Matrix
$\mathbf{W}^\natural \in \mathbb{R}^{d\times k}$
$\mathbf{U}_{[:,:r]}\mathbf{S}^{1/2}_{[:r,:r]}$
r
$\mathbf{S}^{1/2}_{[:r,:r]}\mathbf{V}^\top_{[:,:r]}$
$\mathbf{U},\mathbf{S},\mathbf{V} \leftarrow \mathrm{SVD}\left(-\eta\nabla_{\mathbf{W}}L(\mathbf{W}^\natural)\right)$
Pretrained
Weight Matrix
$\mathbf{W}^\natural \in \mathbb{R}^{d\times k}$
$-\eta\nabla_{\mathbf{W}}L(\mathbf{W}^\natural)$
One-step GD under full fine-tuning

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:

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

- 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 step-size η

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

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

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

$$\mathbb{E}_{\widetilde{\boldsymbol{x}}} \left\| \widetilde{\boldsymbol{y}} - \sigma(\boldsymbol{W}^\natural + \boldsymbol{A}_t \boldsymbol{B}_t)^\top \widetilde{\boldsymbol{x}} \right\|_2^2 \lesssim \|\boldsymbol{A}_t \boldsymbol{B}_t - \Delta\|_{\mathrm{F}}^2$$

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 step-size η

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

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

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

$$\mathbb{E}_{\widetilde{\boldsymbol{x}}} \left\| \widetilde{\boldsymbol{y}} - \sigma(\boldsymbol{W}^\natural + \boldsymbol{A}_t \boldsymbol{B}_t)^\top \widetilde{\boldsymbol{x}} \right\|_2^2 \lesssim \|\boldsymbol{A}_t \boldsymbol{B}_t - \Delta\|_{\mathrm{F}}^2$$

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 step-size η

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

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

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

$$\mathbb{E}_{\widetilde{\boldsymbol{x}}} \left\| \widetilde{\boldsymbol{y}} - \sigma(\boldsymbol{W}^\natural + \boldsymbol{A}_t \boldsymbol{B}_t)^\top \widetilde{\boldsymbol{x}} \right\|_2^2 \lesssim \|\boldsymbol{A}_t \boldsymbol{B}_t - \Delta\|_{\mathrm{F}}^2$$

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 step-size η

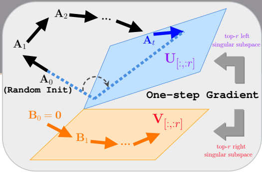
$$\boldsymbol{A}_{t+1} = \boldsymbol{A}_t - \eta \nabla_{\boldsymbol{A}} \widetilde{L}(\boldsymbol{A}_t, \boldsymbol{B}_t)$$

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

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

$$\mathbb{E}_{\widetilde{\boldsymbol{x}}}\left\|\widetilde{\boldsymbol{y}} - \sigma(\boldsymbol{W}^\natural + \boldsymbol{A}_t\boldsymbol{B}_t)^\top \widetilde{\boldsymbol{x}}\right\|_2^2 \lesssim \|\boldsymbol{A}_t\boldsymbol{B}_t - \Delta\|_{\mathrm{F}}^2$$

Our results: Alignment on B_t



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

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

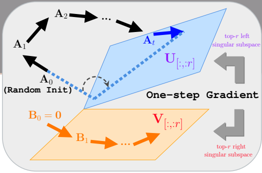
Theorem (Alignment between $G^\natural$ and B_t)

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

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

Remark: $B_1 = \eta_1 A_0^\top G^\natural$ with $\operatorname{Rank}(B_1) \le r^*$

Our results: Alignment on B_t



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

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

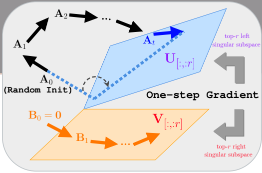
Theorem (Alignment between $G^\natural$ and B_t)

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 $\operatorname{Rank}(\boldsymbol{B}_1) \le r^*$

Our results: 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

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

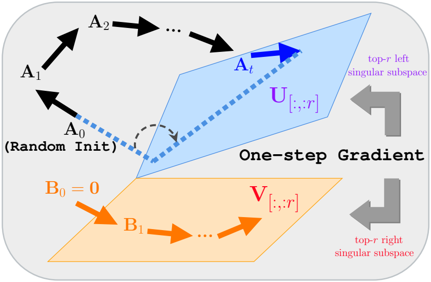
Remark: $\boldsymbol{B}_1 = \eta_1 \boldsymbol{A}_0^\top \boldsymbol{G}^\natural$ with $\operatorname{Rank}(\boldsymbol{B}_1) \le r^*$

Our results: Alignment on A_t

Theorem (Informal, LoRA initialization)

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

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



Our results: Alignment on A_t

Theorem (Informal, LoRA initialization)

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

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

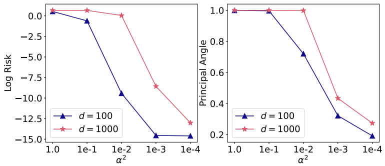


Figure 2: *Left*: the risk $\frac{1}{2}\|\boldsymbol{A}_t\boldsymbol{B}_t - \Delta\|_{\mathrm{F}}^2$. *Right*: the principal angle is $\min_t \|\boldsymbol{U}_{r^*,\perp}^{\top}(\boldsymbol{G}^\natural)\ \boldsymbol{U}_{r^*}(\boldsymbol{A}_t)\|_{op}$.

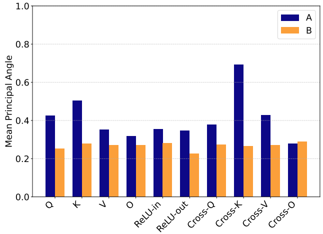


Figure 3: Principal angle of fine-tuning T5 on MRPC.

Key message: Algorithm design principle

Can we "escape" the alignment stage?

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

$$\boldsymbol{A}_0 = \left[\widetilde{\boldsymbol{U}}_{\boldsymbol{G}^\natural}\right]_{[:,1:r]} \left[\widetilde{\boldsymbol{S}}_{\boldsymbol{G}^\natural}^{1/2}\right]_{[1:r]} .$$

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

Message

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

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

The "best" initialization strategy!

Key message: Algorithm design principle

Can we "escape" the alignment stage?

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

$$\begin{aligned} \boldsymbol{A}_0 &= \left[\widetilde{\boldsymbol{U}}_{\boldsymbol{G}^\natural}\right]_{[:,1:r]} \left[\widetilde{\boldsymbol{S}}^{1/2}_{\boldsymbol{G}^\natural}\right]_{[1:r]} . \\ \boldsymbol{B}_0 &= \left[\widetilde{\boldsymbol{S}}^{1/2}_{\boldsymbol{G}^\natural}\right]_{[1:r]} \left[\widetilde{\boldsymbol{V}}_{\boldsymbol{G}^\natural}\right]^\top_{[:,1:r]} . \end{aligned} \quad \text{(Spec-init.)}$$

Message

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

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

The "best" initialization strategy!

Key message: Algorithm design principle

Can we "escape" the alignment stage?

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

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

Message

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

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

The "best" initialization strategy!

Key message: Algorithm design principle

Can we "escape" the alignment stage?

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

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

Message

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

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

The "best" initialization strategy!

Key message: Algorithm design principle

Can we "escape" the alignment stage?

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

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

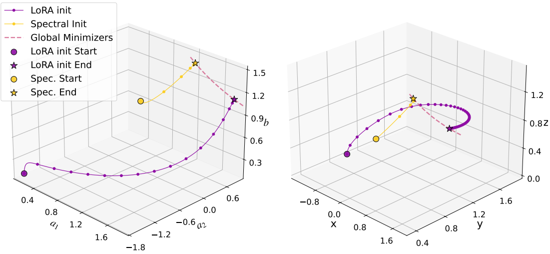
Message

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

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

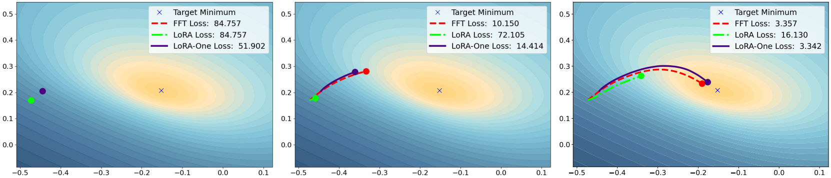
The "best" initialization strategy!

Toy example (I)



(a) Trajectory 1 **(b)** Trajectory 2

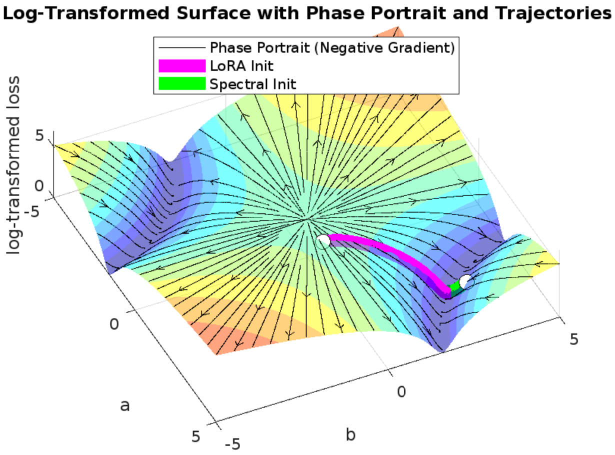
Figure 4: Comparison of the GD trajectories between LoRA and ours. (a) and (b): $\boldsymbol{A} \in \mathbb{R}^2$ and $\boldsymbol{B} \in \mathbb{R}$ with different initializations. The set of global minimizers is $\{a_1^* = 2/t\,, a_2^* = 1/t\,, b^* = t \mid t \in \mathbb{R}\}$.



(a) At initialization **(b)** Middle **(c)** End

Figure 5: Comparison of the GD trajectories between LoRA and ours. We use two-layer neural networks pre-trained on odd-labeled data and fine-tuned on even-labeled data on MNIST.

Toy example (III): Phase portrait



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

Table 1: Fine-tuning T5 model across NLP tasks from GLUE.

Dataset	**MNLI**	**SST-2**	**CoLA**	**QNLI**	**MRPC**
Size	393k	67k	8.5k	105k	3.7k
Pre-trained	-	89.79	59.03	49.28	63.48
One-step GD	-	90.48	73.00	76.64	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!

Table 1: Fine-tuning T5 model across NLP tasks from GLUE.

Dataset	**MNLI**	**SST-2**	**CoLA**	**QNLI**	**MRPC**
Size	393k	67k	8.5k	105k	3.7k
Pre-trained	-	89.79	59.03	49.28	63.48
One-step GD	-	90.48	73.00	76.64	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 [3]

make LoRA's gradients align to full fine-tuning!

- 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 $\boldsymbol{G}^\natural$.

Motivation [3]

make LoRA's gradients align to full fine-tuning!

◦ best-$2r$ approximation: $\operatorname{rank}(\nabla_{\boldsymbol{A}} \widetilde{L}(\boldsymbol{A}_t, \boldsymbol{B}_t)) + \operatorname{rank}(\nabla_{\boldsymbol{B}} \widetilde{L}(\boldsymbol{A}_t, \boldsymbol{B}_t)) \leq 2r$

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

◦ 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 [3]

make LoRA's gradients align to full fine-tuning!

- best-$2r$ approximation: $\operatorname{rank}(\nabla_{\boldsymbol{A}} \widetilde{L}(\boldsymbol{A}_t, \boldsymbol{B}_t)) + \operatorname{rank}(\nabla_{\boldsymbol{B}} \widetilde{L}(\boldsymbol{A}_t, \boldsymbol{B}_t)) \leq 2r$

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

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

Clarification on gradient alignment based work

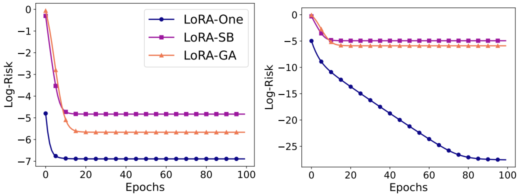
Motivation [3]

make LoRA's gradients align to full fine-tuning!

- 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 $\boldsymbol{G}^\natural$.



(a) $r < r^*$ **(b)** $r > r^*$

Experiments

Key features in our LoRA-One algorithm

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

$\boldsymbol{A}_0 \leftarrow \sqrt{\gamma} \cdot \boldsymbol{U}_{[:,1:r]}\boldsymbol{S}^{1/2}_{[:r,:r]}$

$\boldsymbol{B}_0 \leftarrow \sqrt{\gamma} \cdot \boldsymbol{S}^{1/2}_{[:r,:r]}\boldsymbol{V}^{\top}_{[:,1:r]}$

Clear $\nabla_{\boldsymbol{W}} L(\boldsymbol{W}^{\natural})$

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

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

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

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

Method	MNLI	SST-2	CoLA	QNLI	MRPC
LoRA	$85.30_{\pm 0.04}$	$94.04_{\pm 0.09}$	$72.84_{\pm 1.25}$	$93.02_{\pm 0.07}$	$68.38_{\pm 0.01}$
LoRA+	$85.81_{\pm 0.09}$	$93.85_{\pm 0.24}$	$77.53_{\pm 0.20}$	$93.14_{\pm 0.03}$	$74.43_{\pm 1.39}$
P-LoRA	$85.28_{\pm 0.15}$	$93.88_{\pm 0.11}$	$79.58_{\pm 0.67}$	$93.00_{\pm 0.07}$	$83.91_{\pm 1.16}$
PiSSA	$85.75_{\pm 0.07}$	$94.07_{\pm 0.06}$	$74.27_{\pm 0.39}$	$93.15_{\pm 0.14}$	$76.31_{\pm 0.51}$
LoRA-GA	$85.70_{\pm 0.09}$	$94.11_{\pm 0.18}$	$80.57_{\pm 0.20}$	$93.18_{\pm 0.06}$	$85.29_{\pm 0.24}$
LoRA-Pro	$\mathbf{86.03}_{\pm 0.19}$	$94.19_{\pm 0.13}$	$81.94_{\pm 0.24}$	$\mathbf{93.42}_{\pm 0.05}$	$86.60_{\pm 0.14}$
LoRA-One	$85.89_{\pm 0.08}$	$\mathbf{94.53}_{\pm 0.13}$	$\mathbf{82.04}_{\pm 0.22}$	$93.37_{\pm 0.02}$	$\mathbf{87.83}_{\pm 0.37}$

Results on LLaMA 2-7B (for one epoch)

	GSM8K		MMLU	HumanEval
($r = 8$)	Direct	8s-CoT	Avg.	PASS@1
LoRA	$59.26_{\pm 0.76}$	$53.36_{\pm 0.77}$	$45.73_{\pm 0.30}$	$25.85_{\pm 1.75}$
LoRA-GA	$56.44_{\pm 1.37}$	$46.07_{\pm 1.01}$	$45.70_{\pm 0.77}$	$26.95_{\pm 1.30}$
LoRA-One	$\mathbf{60.44}_{\pm 0.17}$	$\mathbf{55.88}_{\pm 0.44}$	$\mathbf{47.12}_{\pm 0.12}$	$\mathbf{28.66}_{\pm 0.39}$

- One epoch, rank 8, three runs
- Hyperparameter optimized over learning rate, batch size

- Train: 100k subset from MetaMathQA
- Test: GSM8K, Accuracy (%)

Results on LLaMA 2-7B (for one epoch)

	GSM8K		MMLU	HumanEval
($r = 8$)	Direct	8s-CoT	Avg.	PASS@1
LoRA	$59.26_{\pm 0.76}$	$53.36_{\pm 0.77}$	$45.73_{\pm 0.30}$	$25.85_{\pm 1.75}$
LoRA-GA	$56.44_{\pm 1.37}$	$46.07_{\pm 1.01}$	$45.70_{\pm 0.77}$	$26.95_{\pm 1.30}$
LoRA-One	$\mathbf{60.44}_{\pm 0.17}$	$\mathbf{55.88}_{\pm 0.44}$	$\mathbf{47.12}_{\pm 0.12}$	$\mathbf{28.66}_{\pm 0.39}$

- One epoch, rank 8, three runs
- Hyperparameter optimized over learning rate, batch size

- Train: 100k subset from MetaMathQA
- Test: GSM8K, Accuracy (%)

LoRA: 6h 20min + 3 min

Results on LLaMA 2-7B (for one epoch)

	GSM8K		MMLU	HumanEval
($r = 8$)	Direct	8s-CoT	Avg.	PASS@1
LoRA	$59.26_{\pm 0.76}$	$53.36_{\pm 0.77}$	$45.73_{\pm 0.30}$	$25.85_{\pm 1.75}$
LoRA-GA	$56.44_{\pm 1.37}$	$46.07_{\pm 1.01}$	$45.70_{\pm 0.77}$	$26.95_{\pm 1.30}$
LoRA-One	$\mathbf{60.44}_{\pm 0.17}$	$\mathbf{55.88}_{\pm 0.44}$	$\mathbf{47.12}_{\pm 0.12}$	$\mathbf{28.66}_{\pm 0.39}$

- One epoch, rank 8, three runs
- Hyperparameter optimized over learning rate, batch size

- Train: 100k subset from MetaMathQA
- Test: GSM8K, Accuracy (%)

LoRA: 21.6 GB + 0.1 GB

Results on LLaMA 2-7B (for one epoch)

($r = 8$)	GSM8K		MMLU	HumanEval
	Direct	8s-CoT	Avg.	PASS@1
LoRA	$59.26_{\pm 0.76}$	$53.36_{\pm 0.77}$	$45.73_{\pm 0.30}$	$25.85_{\pm 1.75}$
LoRA-GA	$56.44_{\pm 1.37}$	$46.07_{\pm 1.01}$	$45.70_{\pm 0.77}$	$26.95_{\pm 1.30}$
LoRA-One	$\mathbf{60.44}_{\pm 0.17}$	$\mathbf{55.88}_{\pm 0.44}$	$\mathbf{47.12}_{\pm 0.12}$	$\mathbf{28.66}_{\pm 0.39}$

- One epoch, rank 8, three runs
- Hyperparameter optimized over learning rate, batch size
- Train: 100k subset from Code-Feedback
- Test: Humaneval, Pass@1

Results on LLaMA 2-7B (for one epoch)

	GSM8K		MMLU	HumanEval
$(r = 8)$	Direct	8s-CoT	Avg.	PASS@1
LoRA	$59.26_{\pm 0.76}$	$53.36_{\pm 0.77}$	$45.73_{\pm 0.30}$	$25.85_{\pm 1.75}$
LoRA-GA	$56.44_{\pm 1.37}$	$46.07_{\pm 1.01}$	$45.70_{\pm 0.77}$	$26.95_{\pm 1.30}$
LoRA-One	$\mathbf{60.44}_{\pm 0.17}$	$\mathbf{55.88}_{\pm 0.44}$	$\mathbf{47.12}_{\pm 0.12}$	$\mathbf{28.66}_{\pm 0.39}$

- One epoch, rank 8, three runs
- Hyperparameter optimized over learning rate, batch size

- Train: 100k subset from Code-Feedback
- Test: Humaneval, Pass@1

LoRA: 6h 24 min + 2 min

Results on LLaMA 2-7B (for one epoch)

	GSM8K		MMLU	HumanEval
($r = 8$)	Direct	8s-CoT	Avg.	PASS@1
LoRA	$59.26_{\pm 0.76}$	$53.36_{\pm 0.77}$	$45.73_{\pm 0.30}$	$25.85_{\pm 1.75}$
LoRA-GA	$56.44_{\pm 1.37}$	$46.07_{\pm 1.01}$	$45.70_{\pm 0.77}$	$26.95_{\pm 1.30}$
LoRA-One	$\mathbf{60.44}_{\pm 0.17}$	$\mathbf{55.88}_{\pm 0.44}$	$\mathbf{47.12}_{\pm 0.12}$	$\mathbf{28.66}_{\pm 0.39}$

- One epoch, rank 8, three runs
- Hyperparameter optimized over learning rate, batch size

- Train: 100k subset from Code-Feedback
- Test: Humaneval, Pass@1

LoRA: 22.6 GB - 1.1 GB

Results on LLaMA 2-7B (for more epochs)

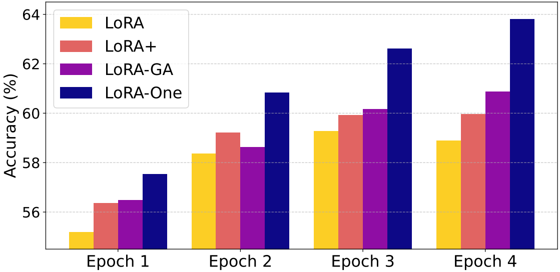


Figure 7: Accuracy comparison across different methods over epochs on GSM8K.

Theory and proof...

Model	Algorithm	Initialization	Results
Linear	GD	(LoRA-init.)	Subspace alignment of $\boldsymbol{B}_t$
	GD	(LoRA-init.)	Subspace alignment of $\boldsymbol{A}_t$
	GD	(Spec-init.)	$\|\boldsymbol{A}_0\boldsymbol{B}_0 - \Delta\|_{\mathrm{F}}$ is small
	GD	(Spec-init.)	Linear convergence of $\|\boldsymbol{A}_t\boldsymbol{B}_t - \Delta\|_{\mathrm{F}}$
	Precondition GD	(Spec-init.)	Linear convergence rate independent of $\kappa(\Delta)$
Nonlinear	Precondition GD	(Spec-init.)	Linear convergence rate independent of $\kappa(\Delta)$

- 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 $\boldsymbol{A}_t$ [2] (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 $\boldsymbol{Z}_t^{\mathtt{lin}} := \boldsymbol{H}^t \boldsymbol{Z}_0$

- 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 $\boldsymbol{A}_t$ [2] (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 $\boldsymbol{Z}_t^{\mathtt{lin}} := \boldsymbol{H}^t \boldsymbol{Z}_0$

- 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 $\boldsymbol{A}_t$ [2] (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, $w.h.p.$

Global convergence on nonlinear models

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

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

- 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

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

- additional assumptions on $\widetilde{\boldsymbol{W}}^\natural$, e.g., adapted weight has smaller signal than pre-trained model

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

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

- additional assumptions on $\widetilde{\boldsymbol{W}}^\natural$, e.g., adapted weight has smaller signal than pre-trained model

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

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

- additional assumptions on $\widetilde{\boldsymbol{W}}^\natural$, e.g., adapted weight has smaller signal than pre-trained model

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

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

- additional assumptions on $\widetilde{\boldsymbol{W}}^\natural$, e.g., adapted weight has smaller signal than pre-trained model

Global convergence

Theorem (Linear convergence rate)

Under (Spec-init.) and $\boldsymbol{J}_{\boldsymbol{W}_t}$ for gradient update (adding preconditioners), choose constant step-size $\eta < 1$, we have

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

$$\left\| \boldsymbol{A}_0 \boldsymbol{B}_0 - \Delta \right\|_{op} \leq \left\| \boldsymbol{A}_0 \boldsymbol{B}_0 - 2\boldsymbol{G}^\natural \right\|_{op} + 2 \left\| \boldsymbol{G}^\natural - \mathbb{E}_{\widetilde{\boldsymbol{x}}}\left[\boldsymbol{G}^\natural\right] \right\|_{op} + \left\| 2\mathbb{E}_{\widetilde{\boldsymbol{x}}}\left[\boldsymbol{G}^\natural\right] - \Delta \right\|_{op}$$

- low-rank approximation error $\leq 2\lambda_{r^*+1}(\boldsymbol{G}^\natural)$
- population error: using $\mathbb{E}_{\widetilde{\boldsymbol{x}}}[-\boldsymbol{J}_{\boldsymbol{W}_t}] = \frac{1}{2}(\boldsymbol{A}_t \boldsymbol{B}_t - \Delta) + \mathcal{O}(\frac{1}{\kappa r^*})$
- concentration error

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

Theorem (Linear convergence rate)

Under (Spec-init.) and $\boldsymbol{J}_{\boldsymbol{W}_t}$ for gradient update (adding preconditioners), choose constant step-size $\eta < 1$, we have

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

$$\|\boldsymbol{A}_0\boldsymbol{B}_0 - \Delta\|_{op} \leq \left\|\boldsymbol{A}_0\boldsymbol{B}_0 - 2\boldsymbol{G}^\natural\right\|_{op} + 2\left\|\boldsymbol{G}^\natural - \mathbb{E}_{\widetilde{\boldsymbol{x}}}\left[\boldsymbol{G}^\natural\right]\right\|_{op} + \left\|2\mathbb{E}_{\widetilde{\boldsymbol{x}}}\left[\boldsymbol{G}^\natural\right] - \Delta\right\|_{op}$$

- low-rank approximation error $\leq 2\lambda_{r^*+1}(\boldsymbol{G}^\natural)$
- population error: using $\mathbb{E}_{\widetilde{\boldsymbol{x}}}[-\boldsymbol{J}_{\boldsymbol{W}_t}] = \frac{1}{2}(\boldsymbol{A}_t\boldsymbol{B}_t - \Delta) + \mathcal{O}(\frac{1}{\kappa r^*})$
- concentration error

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

Theorem (Linear convergence rate)

Under (Spec-init.) and $\boldsymbol{J}_{\boldsymbol{W}_t}$ for gradient update (adding preconditioners), choose constant step-size $\eta < 1$, we have

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

$$\|\boldsymbol{A}_0\boldsymbol{B}_0 - \Delta\|_{op} \leq \left\|\boldsymbol{A}_0\boldsymbol{B}_0 - 2\boldsymbol{G}^\natural\right\|_{op} + 2\left\|\boldsymbol{G}^\natural - \mathbb{E}_{\widetilde{\boldsymbol{x}}}\left[\boldsymbol{G}^\natural\right]\right\|_{op} + \left\|2\mathbb{E}_{\widetilde{\boldsymbol{x}}}\left[\boldsymbol{G}^\natural\right] - \Delta\right\|_{op}$$

- low-rank approximation error $\leq 2\lambda_{r^*+1}(\boldsymbol{G}^\natural)$
- population error: using $\mathbb{E}_{\widetilde{\boldsymbol{x}}}[-\boldsymbol{J}_{\boldsymbol{W}_t}] = \frac{1}{2}(\boldsymbol{A}_t\boldsymbol{B}_t - \Delta) + \mathcal{O}(\frac{1}{\kappa r^*})$
- concentration error

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

Theorem (Linear convergence rate)

Under (Spec-init.) and $\boldsymbol{J}_{\boldsymbol{W}_t}$ for gradient update (adding preconditioners), choose constant step-size $\eta < 1$, we have

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

$$\|\boldsymbol{A}_0\boldsymbol{B}_0 - \Delta\|_{op} \leq \left\|\boldsymbol{A}_0\boldsymbol{B}_0 - 2\boldsymbol{G}^\natural\right\|_{op} + 2\left\|\boldsymbol{G}^\natural - \mathbb{E}_{\widetilde{\boldsymbol{x}}}\left[\boldsymbol{G}^\natural\right]\right\|_{op} + \left\|2\mathbb{E}_{\widetilde{\boldsymbol{x}}}\left[\boldsymbol{G}^\natural\right] - \Delta\right\|_{op}$$

- low-rank approximation error $\leq 2\lambda_{r^*+1}(\boldsymbol{G}^\natural)$
- population error: using $\mathbb{E}_{\widetilde{\boldsymbol{x}}}[-\boldsymbol{J}_{\boldsymbol{W}_t}] = \frac{1}{2}(\boldsymbol{A}_t\boldsymbol{B}_t - \Delta) + \mathcal{O}(\frac{1}{\kappa r^*})$
- concentration error

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

Theorem (Linear convergence rate)

Under (Spec-init.) and $\boldsymbol{J}_{\boldsymbol{W}_t}$ for gradient update (adding preconditioners), choose constant step-size $\eta < 1$, we have

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

$$\|\boldsymbol{A}_0 \boldsymbol{B}_0 - \Delta\|_{op} \leq \left\|\boldsymbol{A}_0 \boldsymbol{B}_0 - 2\boldsymbol{G}^\natural\right\|_{op} + 2\left\|\boldsymbol{G}^\natural - \mathbb{E}_{\widetilde{\boldsymbol{x}}}\left[\boldsymbol{G}^\natural\right]\right\|_{op} + \left\|2\mathbb{E}_{\widetilde{\boldsymbol{x}}}\left[\boldsymbol{G}^\natural\right] - \Delta\right\|_{op}$$

- low-rank approximation error $\leq 2\lambda_{r^*+1}(\boldsymbol{G}^\natural)$
- population error: using $\mathbb{E}_{\widetilde{\boldsymbol{x}}}[-\boldsymbol{J}_{\boldsymbol{W}_t}] = \frac{1}{2}(\boldsymbol{A}_t \boldsymbol{B}_t - \Delta) + \mathcal{O}(\frac{1}{\kappa r^*})$
- concentration error

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

Theorem (Linear convergence rate)

Under (Spec-init.) and $\boldsymbol{J}_{\boldsymbol{W}_t}$ for gradient update (adding preconditioners), choose constant step-size $\eta < 1$, we have

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

$$\|\boldsymbol{A}_0\boldsymbol{B}_0 - \Delta\|_{op} \leq \left\|\boldsymbol{A}_0\boldsymbol{B}_0 - 2\boldsymbol{G}^\natural\right\|_{op} + 2\left\|\boldsymbol{G}^\natural - \mathbb{E}_{\widetilde{\boldsymbol{x}}}\left[\boldsymbol{G}^\natural\right]\right\|_{op} + \left\|2\mathbb{E}_{\widetilde{\boldsymbol{x}}}\left[\boldsymbol{G}^\natural\right] - \Delta\right\|_{op}$$

- low-rank approximation error $\leq 2\lambda_{r^*+1}(\boldsymbol{G}^\natural)$
- population error: using $\mathbb{E}_{\widetilde{\boldsymbol{x}}}[-\boldsymbol{J}_{\boldsymbol{W}_t}] = \frac{1}{2}(\boldsymbol{A}_t\boldsymbol{B}_t - \Delta) + \mathcal{O}(\frac{1}{\kappa r^*})$
- concentration error

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

Proof of sketch on $A_t B_t - \Delta$

$$
\begin{aligned}
\|\boldsymbol{A}_{t+1}\boldsymbol{B}_{t+1} - \Delta\|_{\mathrm{F}} \lesssim\ & \|\boldsymbol{J}^{\mathtt{GLM}}_{\boldsymbol{W}_t} - \frac{1}{2}(\boldsymbol{A}_t\boldsymbol{B}_t - \Delta)\|_{\mathrm{F}} \ \textcolor{red}{\text{[concentration+population]}} \\
& + (1-\eta)\left\| \boldsymbol{U}_{\boldsymbol{A}_t}\boldsymbol{U}_{\boldsymbol{A}_t}^\top (\boldsymbol{A}_t\boldsymbol{B}_t - \Delta)\boldsymbol{V}_{\boldsymbol{B}_t}\boldsymbol{V}_{\boldsymbol{B}_t}^\top \right\|_{\mathrm{F}} \\
& + \left\| \left(\boldsymbol{I}_d - \boldsymbol{U}_{\boldsymbol{A}_t}\boldsymbol{U}_{\boldsymbol{A}_t}^\top\right)(\boldsymbol{A}_t\boldsymbol{B}_t - \Delta)\left(\boldsymbol{I}_k - \boldsymbol{V}_{\boldsymbol{B}_t}\boldsymbol{V}_{\boldsymbol{B}_t}^\top\right) \right\|_{\mathrm{F}} \\
& + \text{cross terms}
\end{aligned}
$$

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

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

Proof of sketch on $A_t B_t - \Delta$

$$
\begin{aligned}
\|\boldsymbol{A}_{t+1}\boldsymbol{B}_{t+1} - \Delta\|_{\mathrm{F}} \lesssim\ & \|\boldsymbol{J}^{\mathtt{GLM}}_{\boldsymbol{W}_t} - \frac{1}{2}(\boldsymbol{A}_t\boldsymbol{B}_t - \Delta)\|_{\mathrm{F}} \ \textcolor{red}{\text{[concentration+population]}} \\
& + (1-\eta)\left\|\boldsymbol{U}_{\boldsymbol{A}_t}\boldsymbol{U}_{\boldsymbol{A}_t}^\top(\boldsymbol{A}_t\boldsymbol{B}_t - \Delta)\boldsymbol{V}_{\boldsymbol{B}_t}\boldsymbol{V}_{\boldsymbol{B}_t}^\top\right\|_{\mathrm{F}} \\
& + \left\|\left(\boldsymbol{I}_d - \boldsymbol{U}_{\boldsymbol{A}_t}\boldsymbol{U}_{\boldsymbol{A}_t}^\top\right)(\boldsymbol{A}_t\boldsymbol{B}_t - \Delta)\left(\boldsymbol{I}_k - \boldsymbol{V}_{\boldsymbol{B}_t}\boldsymbol{V}_{\boldsymbol{B}_t}^\top\right)\right\|_{\mathrm{F}} \\
& + \text{cross terms}
\end{aligned}
$$

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

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

◦ *LoRA-One: One-step full gradient could suffice for fine-tuning large language models, provably and efficiently.* ICML'25 spotlight. code

- subspace alignment: $\boldsymbol{G}^{\natural}$ and $(\boldsymbol{A}_t, \boldsymbol{B}_t) \Rightarrow$ theory-grounded algorithm design
- "optimal" non-zero initialization strategy
- 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

Thank you!

fanghui.liu@warwick.ac.uk

www.lfhsgre.org

Takeaway messages

◦ *LoRA-One: One-step full gradient could suffice for fine-tuning large language models, provably and efficiently.* ICML'25 spotlight. code

- subspace alignment: $\boldsymbol{G}^{\natural}$ and $(\boldsymbol{A}_t, \boldsymbol{B}_t) \Rightarrow$ theory-grounded algorithm design
- "optimal" non-zero initialization strategy
- 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

Thank you!

fanghui.liu@warwick.ac.uk

www.lfhsgre.org

Takeaway messages

◦ *LoRA-One: One-step full gradient could suffice for fine-tuning large language models, provably and efficiently.* ICML'25 spotlight. code

- subspace alignment: $\boldsymbol{G}^{\natural}$ and $(\boldsymbol{A}_t, \boldsymbol{B}_t) \Rightarrow$ theory-grounded algorithm design
- "optimal" non-zero initialization strategy
- 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

Thank you!

fanghui.liu@warwick.ac.uk

www.lfhsgre.org

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.

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.