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

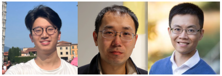
Fanghui Liu

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)]



Outline

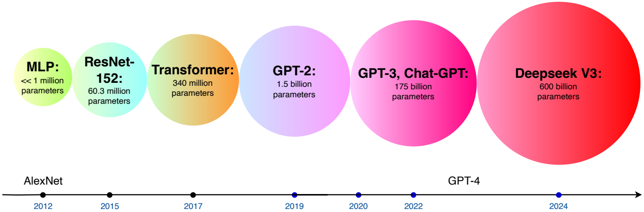
❒ Fine-tuning in LLMs

❒ How does theory contribute to practice?

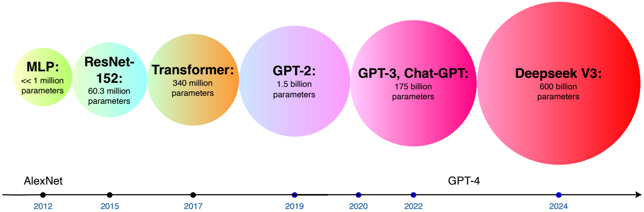
- understanding: subspace alignment
- theory-grounded algorithm for efficiency improvement
- clarify misconceptions

❒ Proofs

In the era of machine learning (Pre-training)



In the era of machine learning (Pre-training)



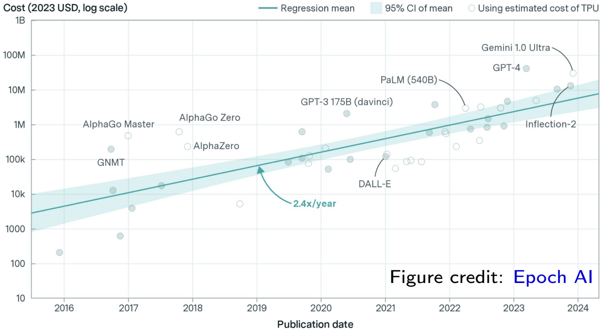
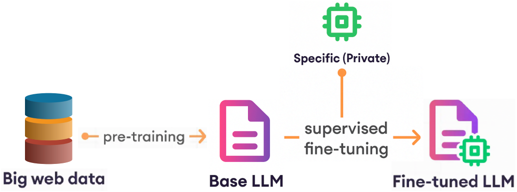
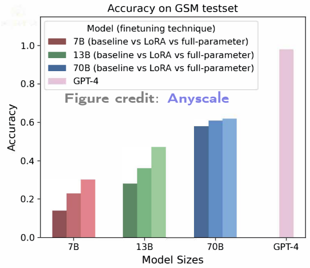


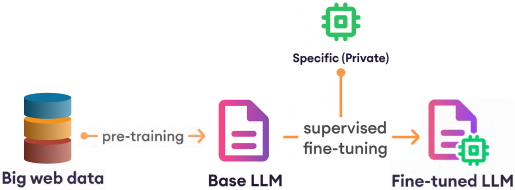
Figure credit: Epoch AI

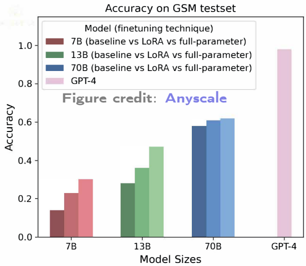
From pre-training to (parameter-efficient) fine-tuning





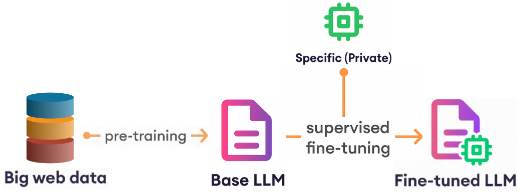
From pre-training to (parameter-efficient) fine-tuning

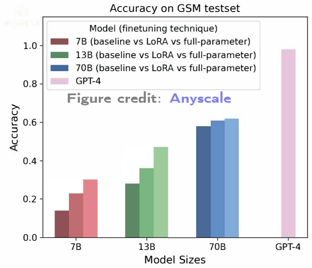




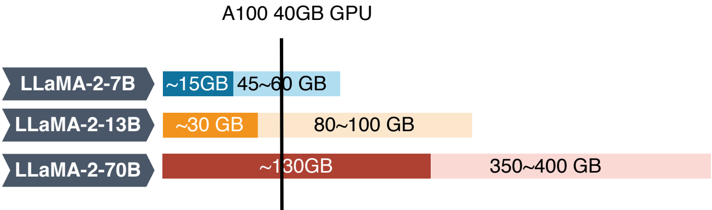
Full fine-tuning vs. parameter-efficient fine-tuning

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

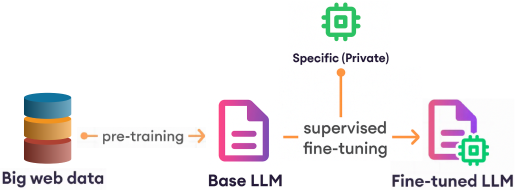


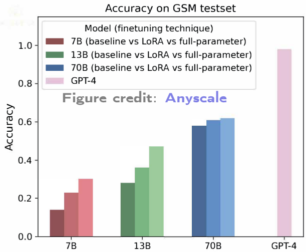


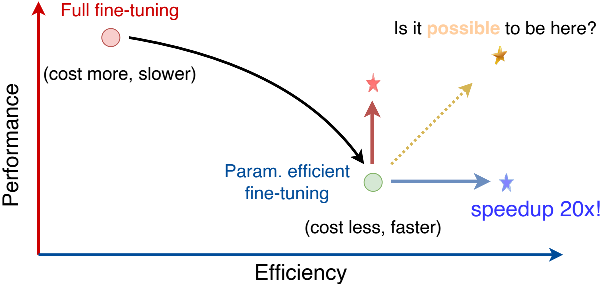
Full fine-tuning vs. parameter-efficient fine-tuning



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



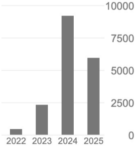




Published as a conference paper at ICLR 2022

LoRA: Low-Rank Adaptation of Large Language Models

Edward Hu* **Yelong Shen*** **Phillip Wallis** **Zeyuan Allen-Zhu**
Yuanzhi Li **Shean Wang** **Lu Wang** **Weizhu Chen**
Microsoft Corporation
edward.hu@mila.quebec
{yeshe, phwallis, zeyuana, swang, luw, wzchen}@microsoft.com
yuanzhil@andrew.cmu.edu

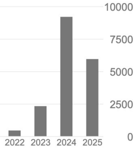


LoRA: Low-rank adaption

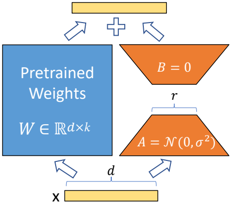
Published as a conference paper at ICLR 2022

LoRA: Low-Rank Adaptation of Large Language Models

Edward Hu* **Yelong Shen*** **Phillip Wallis** **Zeyuan Allen-Zhu**
Yuanzhi Li **Shean Wang** **Lu Wang** **Weizhu Chen**
Microsoft Corporation
`edward.hu@mila.quebec`
`{yeshe, phwallis, zeyuana, swang, luw, wzchen}@microsoft.com`
`yuanzhil@andrew.cmu.edu`



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



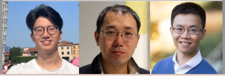
- Formulation:

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

- Initialization:

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

Today's talk: Improve "sub-optimal" LoRA



How can theory guide practice

- understanding: training dynamics of $(\boldsymbol{A}_t, \boldsymbol{B}_t)$
- design theory-grounded algorithm (LoRA-One)
- clarify some misconceptions in previous algorithm designs

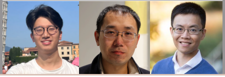
❒ Even for linear model (pre-training and fine-tuning), nonlinear dynamics...

$$\begin{bmatrix} \boldsymbol{A}_{t+1} \\ \boldsymbol{B}_{t+1}^{\top} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_d & \eta \boldsymbol{G} \\ \eta \boldsymbol{G}^{\top} & \boldsymbol{I}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_t \\ \boldsymbol{B}_t^{\top} \end{bmatrix} + \text{nonlinear term} \qquad \begin{cases} [\boldsymbol{A}_0]_{ij} \sim \mathcal{N}(0, \alpha^2) \\ [\boldsymbol{B}_0]_{ij} = 0\,. \end{cases}$$

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

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

Today's talk: Improve "sub-optimal" LoRA



How can theory guide practice

- understanding: training dynamics of $(\boldsymbol{A}_t, \boldsymbol{B}_t)$
- design theory-grounded algorithm (LoRA-One)
- clarify some misconceptions in previous algorithm designs

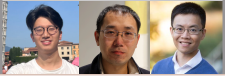
❐ Even for linear model (pre-training and fine-tuning), nonlinear dynamics...

$$\begin{bmatrix} \boldsymbol{A}_{t+1} \\ \boldsymbol{B}_{t+1}^{\top} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_d & \eta \boldsymbol{G} \\ \eta \boldsymbol{G}^{\top} & \boldsymbol{I}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_t \\ \boldsymbol{B}_t^{\top} \end{bmatrix} + \text{nonlinear term} \qquad \begin{cases} [\boldsymbol{A}_0]_{ij} \sim \mathcal{N}(0, \alpha^2) \\ [\boldsymbol{B}_0]_{ij} = 0\,. \end{cases}$$

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

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

Today's talk: Improve "sub-optimal" LoRA



How can theory guide practice

- understanding: training dynamics of $(\boldsymbol{A}_t, \boldsymbol{B}_t)$
- design theory-grounded algorithm (LoRA-One)
- clarify some misconceptions in previous algorithm designs

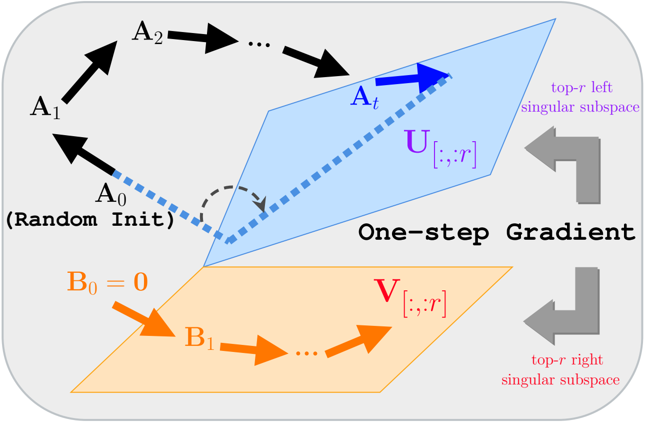
❐ Even for linear model (pre-training and fine-tuning), nonlinear dynamics...

$$\begin{bmatrix} \boldsymbol{A}_{t+1} \\ \boldsymbol{B}_{t+1}^\top \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_d & \eta \boldsymbol{G} \\ \eta \boldsymbol{G}^\top & \boldsymbol{I}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_t \\ \boldsymbol{B}_t^\top \end{bmatrix} + \text{nonlinear term} \qquad \begin{cases} [\boldsymbol{A}_0]_{ij} \sim \mathcal{N}(0, \alpha^2) \\ [\boldsymbol{B}_0]_{ij} = 0\,. \end{cases}$$

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

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

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
$\mathbf{B}_0 = \mathbf{0}$
$\mathbf{V}_{[:,:r]}$
$\mathbf{B}_1$
...
top-r right singular subspace

Problem setting and assumptions

- (Downstream) data : with unknown low-rank feature shift $\mathrm{rank}(\Delta) = r^*$

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

Problem setting and assumptions

- (Downstream) data : with unknown low-rank feature shift $\text{rank}(\Delta) = r^*$

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

- Model : LoRA starting from **known** pre-trained $\boldsymbol{W}^\natural \in \mathbb{R}^{d \times k}$

$$f(\boldsymbol{x}; \boldsymbol{A}, \boldsymbol{B}) := \begin{cases} (\boldsymbol{W}^\natural + \boldsymbol{A}\boldsymbol{B})^\top \widetilde{\boldsymbol{x}} & \text{linear} \\ \sigma[(\boldsymbol{W}^\natural + \boldsymbol{A}\boldsymbol{B})^\top \widetilde{\boldsymbol{x}}] & \text{nonlinear} \end{cases}.$$

Problem setting and assumptions

- (Downstream) data : with unknown low-rank feature shift $\text{rank}(\Delta) = r^*$

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

- Model : LoRA starting from **known** pre-trained $\boldsymbol{W}^\natural \in \mathbb{R}^{d \times k}$

$$f(\boldsymbol{x}; \boldsymbol{A}, \boldsymbol{B}) := \begin{cases} (\boldsymbol{W}^\natural + \boldsymbol{A}\boldsymbol{B})^\top \widetilde{\boldsymbol{x}} & \text{linear} \\ \sigma[(\boldsymbol{W}^\natural + \boldsymbol{A}\boldsymbol{B})^\top \widetilde{\boldsymbol{x}}] & \text{nonlinear} \end{cases}.$$

- Algorithm : Squared loss, gradient descent

$$\begin{cases} L(\boldsymbol{W}) = \frac{1}{2N} \sum_{i=1}^N \|f(\boldsymbol{x}_i; \boldsymbol{W}) - \widetilde{\boldsymbol{y}}_i\|_2^2, & \text{Full fine-tuning} \\ L(\boldsymbol{A}, \boldsymbol{B}) = \frac{1}{2N} \sum_{i=1}^N \|f(\boldsymbol{x}_i; \boldsymbol{A}, \boldsymbol{B}) - \widetilde{\boldsymbol{y}}_i\|_2^2, & \text{LoRA} \end{cases}.$$

Problem setting and assumptions

❒ (Downstream) data : with unknown low-rank feature shift $\operatorname{rank}(\Delta) = r^*$

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

❒ Model : LoRA starting from **known** pre-trained $\boldsymbol{W}^\natural \in \mathbb{R}^{d \times k}$

$$f(\boldsymbol{x}; \boldsymbol{A}, \boldsymbol{B}) := \begin{cases} (\boldsymbol{W}^\natural + \boldsymbol{A}\boldsymbol{B})^\top \widetilde{\boldsymbol{x}} & \text{linear} \\ \sigma[(\boldsymbol{W}^\natural + \boldsymbol{A}\boldsymbol{B})^\top \widetilde{\boldsymbol{x}}] & \text{nonlinear} \end{cases}.$$

❒ Algorithm : Squared loss, gradient descent

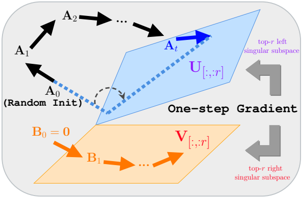
$$\begin{cases} L(\boldsymbol{W}) = \frac{1}{2N} \sum_{i=1}^N \|f(\boldsymbol{x}_i; \boldsymbol{W}) - \widetilde{\boldsymbol{y}}_i\|_2^2, & \text{Full fine-tuning} \\ L(\boldsymbol{A}, \boldsymbol{B}) = \frac{1}{2N} \sum_{i=1}^N \|f(\boldsymbol{x}_i; \boldsymbol{A}, \boldsymbol{B}) - \widetilde{\boldsymbol{y}}_i\|_2^2, & \text{LoRA} \end{cases}.$$

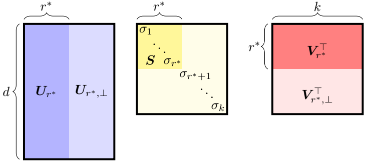
❒ Evaluate $\|\boldsymbol{A}_t \boldsymbol{B}_t - \Delta\|_{\mathrm{F}}$: optimization and generalization!

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

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

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





Theorem (Alignment between G and B_t)

For the linear setting, LoRA trained by gradient descent admits

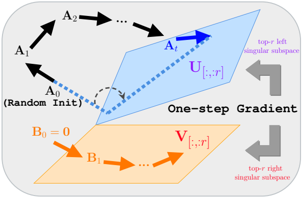
$$\angle(\boldsymbol{V}_{r^*}(\boldsymbol{B}_t), \boldsymbol{V}_{r^*}(\boldsymbol{G})) = 0\,, \quad \forall t \in \mathbb{N}_+\,.$$

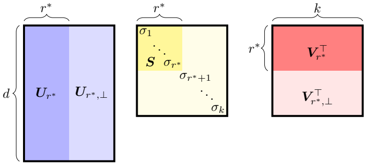
Remark: $B_1 = \eta_1 A_0^\top G$ with Rank$(B_1) \le r^*$

Our results: Alignment on B_t

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

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





Theorem (Alignment between G and B_t)

For the linear setting, LoRA trained by gradient descent admits

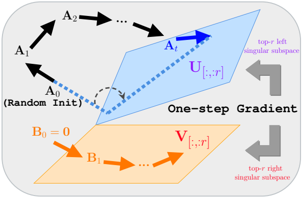
$$\angle(\boldsymbol{V}_{r^*}(\boldsymbol{B}_t), \boldsymbol{V}_{r^*}(\boldsymbol{G})) = 0\,, \quad \forall t \in \mathbb{N}_+\,.$$

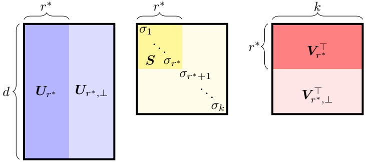
Remark: $B_1 = \eta_1 A_0^\top G$ with $\mathrm{Rank}(B_1) \le r^*$

Our results: Alignment on B_t

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

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





Theorem (Alignment between G and B_t)

For the linear setting, LoRA trained by gradient descent admits

$$\angle(\boldsymbol{V}_{r^*}(\boldsymbol{B}_t), \boldsymbol{V}_{r^*}(\boldsymbol{G})) = 0\,, \quad \forall t \in \mathbb{N}_+\,.$$

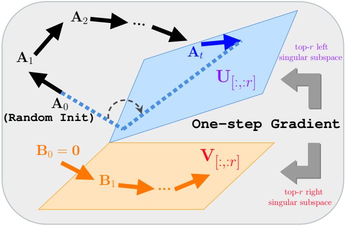
Remark: $\boldsymbol{B}_1 = \eta_1 \boldsymbol{A}_0^\top \boldsymbol{G}$ with Rank$(\boldsymbol{B}_1) \leq r^*$

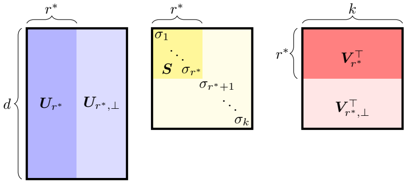
Our results: Alignment on A_t

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

$$\angle(\boldsymbol{U}_{r^*}(\boldsymbol{A}_{t^*}), \boldsymbol{U}_{r^*}(\boldsymbol{G})) \lesssim \epsilon\,, w.h.p.$$





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

$$\angle(\boldsymbol{U}_{r^*}(\boldsymbol{A}_{t^*}), \boldsymbol{U}_{r^*}(\boldsymbol{G})) \lesssim \epsilon\,, w.h.p.$$

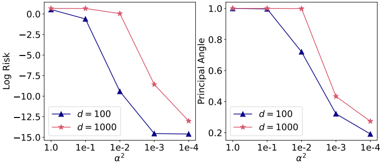


Figure 2: *Left*: the risk $\frac{1}{2}\|\boldsymbol{A}_t\boldsymbol{B}_t - \Delta\|_{\mathrm{F}}^2$. *Right*: the principal angle is $\min_t \|\boldsymbol{U}_{r^*,\perp}^{\top}(\boldsymbol{G})\ \boldsymbol{U}_{r^*}(\boldsymbol{A}_t)\|_{op}$.

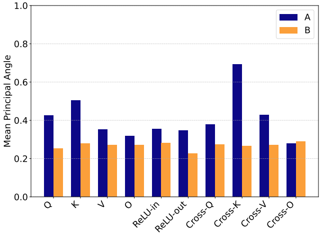
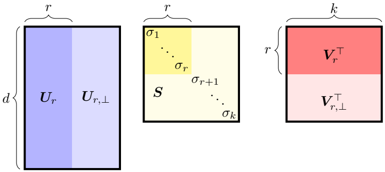


Figure 3: Principal angle of fine-tuning T5 on MRPC.

Algorithm design principle

- SVD: $\boldsymbol{G} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$

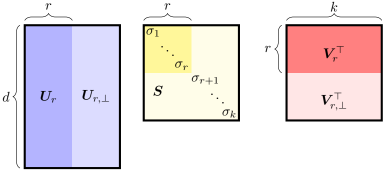
$$
\begin{aligned}
\boldsymbol{A}_0 &= \boldsymbol{U}_{[:,1:r]}\boldsymbol{S}^{\frac{1}{2}}_{[1:r]}\,.\\
\boldsymbol{B}_0 &= \boldsymbol{S}^{\frac{1}{2}}_{[1:r]}\boldsymbol{V}^\top_{[:,1:r]}\,.
\end{aligned}
\qquad \text{(Spec-init.)}
$$



Algorithm design principle

❒ SVD: $\boldsymbol{G} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$

$$\boldsymbol{A}_0 = \boldsymbol{U}_{[:,1:r]}\boldsymbol{S}_{[1:r]}^{\frac{1}{2}}\,,\qquad \boldsymbol{B}_0 = \boldsymbol{S}_{[1:r]}^{\frac{1}{2}}\boldsymbol{V}_{[:,1:r]}^\top\,. \qquad \text{(Spec-init.)}$$



Key Message: we can "escape" the alignment stage

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

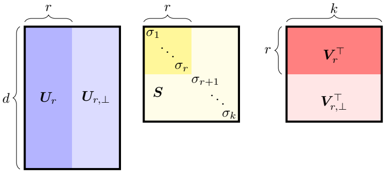
$$\|\boldsymbol{A}_0\boldsymbol{B}_0 - \Delta\|_{\mathrm{F}} \text{ is small, } w.h.p.$$

Algorithm design principle

- SVD: $\boldsymbol{G} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$

$$\boldsymbol{A}_0 = \boldsymbol{U}_{[:,1:r]}\boldsymbol{S}^{\frac{1}{2}}_{[1:r]}\,.$$

$$\boldsymbol{B}_0 = \boldsymbol{S}^{\frac{1}{2}}_{[1:r]}\boldsymbol{V}^\top_{[:,1:r]}\,. \qquad \text{(Spec-init.)}$$



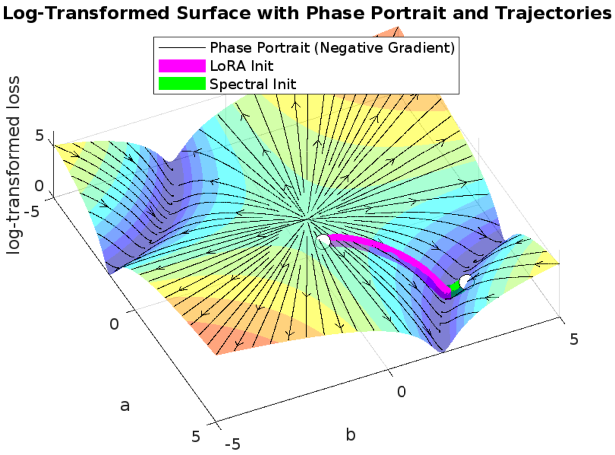
Key Message: we can "escape" the alignment stage

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

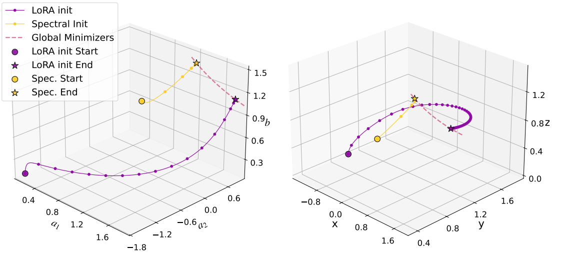
$$\|\boldsymbol{A}_0\boldsymbol{B}_0 - \Delta\|_{\mathrm{F}} \text{ is small, } w.h.p.$$

The "best" initialization strategy!

"Best" initialization: phase portrait



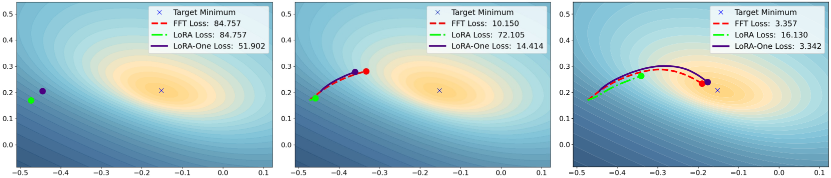
Toy example (I)



(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 $\{a_1^* = 2/t\,, a_2^* = 1/t\,, b^* = t \mid t \in \mathbb{R}\}$.

Toy example (II)



(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, see GIF illustration.

One-step gradient can suffice on small-scale datasets!

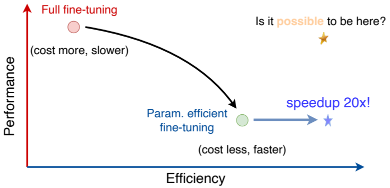
Dataset	**MNLI**	**SST-2**	**CoLA**	**QNLI**	**MRPC**
Size	393k	67k	8.5k	105k	3.7k
Pre-trained	-	89.79	59.03	49.28	63.48
Spectral init.	-	90.48	73.00	76.64	68.38
LoRA_8	$85.30_{\pm 0.04}$	$94.04_{\pm 0.09}$	$72.84_{\pm 1.25}$	$93.02_{\pm 0.07}$	$68.38_{\pm 0.01}$

One-step gradient can suffice on small-scale datasets!

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
Spectral init.	-	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 (sec.)	LoRA	Spectral init.
CoLA	47s	$<$1s
MRPC	25s	$<$1s

❒ memory-efficient [1] + randomized SVD + parallel initialization



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**

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

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

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

end

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 (%)

Time cost	LoRA: 6h 20min	+ 3min
Memory	LoRA: 21.6 GB	+ 0.1GB

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

($r = 8$)	GSM8K Direct	GSM8K 8s-CoT	MMLU Avg.	HumanEval 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

Time cost LoRA: 6h 24min + 2min

Memory LoRA: 22.6 GB - 1.1GB

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

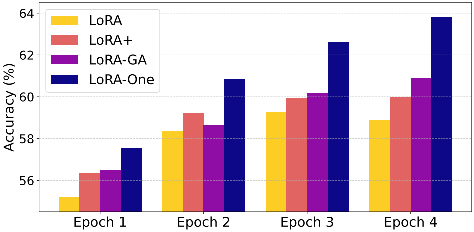


Figure 6: Accuracy comparison across different methods over epochs on GSM8K.

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

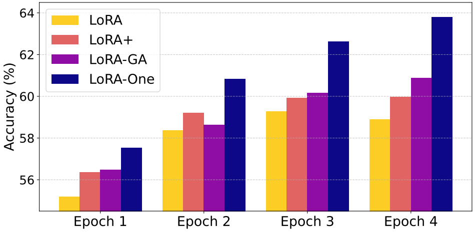
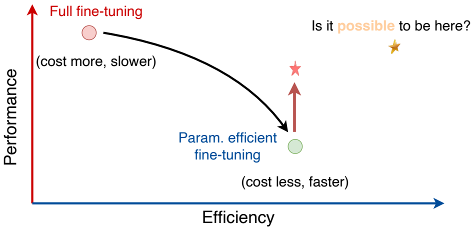


Figure 6: Accuracy comparison across different methods over epochs on GSM8K.



Clarification on gradient alignment based work

LoRA-GA (Wang et al, 2024): make LoRA's gradients align to full fine-tuning!

Clarification on gradient alignment based work

LoRA-GA (Wang et al, 2024): make LoRA's gradients align to full fine-tuning!

❒ best $2r$ approximation

$$\begin{cases} \mathrm{rank}(\nabla_{\boldsymbol{A}} L(\boldsymbol{A}_t, \boldsymbol{B}_t)) \leq r \\ \mathrm{rank}(\nabla_{\boldsymbol{B}} L(\boldsymbol{A}_t, \boldsymbol{B}_t)) \leq r \end{cases}$$

Clarification on gradient alignment based work

LoRA-GA (Wang et al, 2024): make LoRA's gradients align to full fine-tuning!

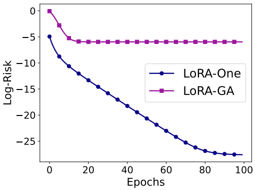
❒ best $2r$ approximation

$$\begin{cases} \operatorname{rank}(\nabla_{\boldsymbol{A}} L(\boldsymbol{A}_t, \boldsymbol{B}_t)) \leq r \\ \operatorname{rank}(\nabla_{\boldsymbol{B}} L(\boldsymbol{A}_t, \boldsymbol{B}_t)) \leq r \end{cases}$$

Method	Init. on $\boldsymbol{A}$	Init. on $\boldsymbol{B}$	Calibration
LoRA	$\mathcal{N}(0, \alpha^2)$	0	-
LoRA-GA	$\boldsymbol{U}_{[:,1:r]}$	$\boldsymbol{V}^\top_{[:,r+1:2r]}$	$\boldsymbol{W}^{\text{pre}} - \boldsymbol{A}_0 \boldsymbol{B}_0$
LoRA-One	$\boldsymbol{U}_{[:,1:r]} \boldsymbol{S}^{1/2}_{[1:r]}$	$\boldsymbol{S}^{1/2}_{[1:r]} \boldsymbol{V}^\top_{[:,1:r]}$	-

Clarification on gradient alignment based work

LoRA-GA (Wang et al, 2024): make LoRA's gradients align to full fine-tuning!



Method	Init. on A	Init. on B	Calibration
LoRA	$\mathcal{N}(0, \alpha^2)$	0	-
LoRA-GA	$\boldsymbol{U}_{[:,1:r]}$	$\boldsymbol{V}^\top_{[:,r+1:2r]}$	$\boldsymbol{W}^{\text{pre}} - \boldsymbol{A}_0\boldsymbol{B}_0$
LoRA-One	$\boldsymbol{U}_{[:,1:r]}\boldsymbol{S}^{1/2}_{[1:r]}$	$\boldsymbol{S}^{1/2}_{[1:r]}\boldsymbol{V}^\top_{[:,1:r]}$	-

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	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 $\kappa(\Delta)$

- subspace alignment
- global convergence

Proof of sketch: Control the dynamics for alignment

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

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

- Schur decomposition of $\boldsymbol{H}$
- obtain the dynamics of $\boldsymbol{Z}_t^{\mathtt{lin}}$
- Define $\boldsymbol{E}_t := \boldsymbol{Z}_t - \boldsymbol{Z}_t^{\mathtt{lin}}$, control $\|\boldsymbol{E}_t\|_{op}$ in $\mathcal{O}(\log d)$ steps

❒Transfer the alignment from $\boldsymbol{A}_t^{\mathtt{lin}}$ to $\boldsymbol{A}_t$ [2] (Stöger & Soltanolkotabi, 2021)

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

Proof of sketch: Control the dynamics for alignment

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

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

- Schur decomposition of $\boldsymbol{H}$
- obtain the dynamics of $\boldsymbol{Z}_t^{\texttt{lin}}$
- Define $\boldsymbol{E}_t := \boldsymbol{Z}_t - \boldsymbol{Z}_t^{\texttt{lin}}$, control $\|\boldsymbol{E}_t\|_{op}$ in $\mathcal{O}(\log d)$ steps

❒Transfer the alignment from $\boldsymbol{A}_t^{\texttt{lin}}$ to $\boldsymbol{A}_t$ [2] (Stöger & Soltanolkotabi, 2021)

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

Proof of sketch: Control the dynamics for alignment

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

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

- Schur decomposition of $\boldsymbol{H}$
- obtain the dynamics of $\boldsymbol{Z}_t^{\mathtt{lin}}$
- Define $\boldsymbol{E}_t := \boldsymbol{Z}_t - \boldsymbol{Z}_t^{\mathtt{lin}}$, control $\|\boldsymbol{E}_t\|_{op}$ in $\mathcal{O}(\log d)$ steps

❒Transfer the alignment from $\boldsymbol{A}_t^{\mathtt{lin}}$ to $\boldsymbol{A}_t$ [2] (Stöger & Soltanolkotabi, 2021)

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

Global convergence of nonlinear models

Theorem (Informal, linear convergence rate)

For nonlinear model with $r = r^*$ and gradient descent (with preconditioners), choose constant step-size $\eta < 1$, we have

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

$$\|\boldsymbol{A}_0 \boldsymbol{B}_0 - \Delta\|_{op} \le \|\boldsymbol{A}_0 \boldsymbol{B}_0 - 2\boldsymbol{G}\|_{op} + 2\|\boldsymbol{G} - \mathbb{E}_{\widetilde{\boldsymbol{x}}}[\boldsymbol{G}]\|_{op} + \|2\mathbb{E}_{\widetilde{\boldsymbol{x}}}[\boldsymbol{G}] - \Delta\|_{op}$$

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

- low-rank approximation error $\le 2\lambda_{r^*+1}(\boldsymbol{G})$
- 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

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

Global convergence of nonlinear models

Theorem (Informal, linear convergence rate)

For nonlinear model with $r = r^*$ and gradient descent (with preconditioners), choose constant step-size $\eta < 1$, we have

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

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

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

- low-rank approximation error $\leq 2\lambda_{r^*+1}(\boldsymbol{G})$
- 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

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

Global convergence of nonlinear models

Theorem (Informal, linear convergence rate)

For nonlinear model with $r = r^*$ and gradient descent (with preconditioners), choose constant step-size $\eta < 1$, we have

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

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

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

- low-rank approximation error $\leq 2\lambda_{r^*+1}(\boldsymbol{G})$
- 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

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

Theorem (Informal, linear convergence rate)

For nonlinear model with $r = r^*$ and gradient descent (with preconditioners), choose constant step-size $\eta < 1$, we have

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

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

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

- low-rank approximation error $\leq 2\lambda_{r^*+1}(\boldsymbol{G})$
- 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

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

Global convergence of nonlinear models

Theorem (Informal, linear convergence rate)

For nonlinear model with $r = r^*$ and gradient descent (with preconditioners), choose constant step-size $\eta < 1$, we have

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

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

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

- low-rank approximation error $\leq 2\lambda_{r^*+1}(\boldsymbol{G})$
- 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

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

Proof of sketch on $A_t B_t - \Delta$

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

❒ projection

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

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

❒ lower bound $\left\|\boldsymbol{L}_\perp^\top \boldsymbol{\Delta} \boldsymbol{L}\right\|_{\mathrm{F}}^2$, upper bound $\left\|\boldsymbol{L}_\perp^\top \boldsymbol{U}\right\|_{op} < 1$

Proof of sketch on $A_t B_t - \Delta$

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

❒ projection

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

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

❒ lower bound $\left\|\boldsymbol{L}_\perp^\top \Delta \boldsymbol{L}\right\|_{\mathrm{F}}^2$, upper bound $\left\|\boldsymbol{L}_\perp^\top \boldsymbol{U}\right\|_{op} < 1$

Proof of sketch on $A_t B_t - \Delta$

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

❒ projection

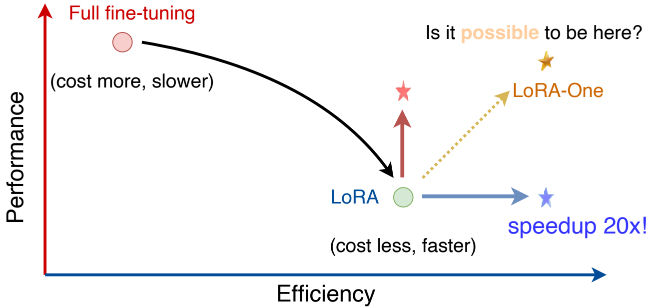
$$
\boldsymbol{L} = \begin{bmatrix} \boldsymbol{U}_{\boldsymbol{A}_t} & 0_{d\times r} \\ 0_{k\times r} & \boldsymbol{V}_{\boldsymbol{B}_t} \end{bmatrix} \in \mathbb{R}^{(d+k)\times 2r},
$$

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

❒ lower bound $\left\|\boldsymbol{L}_\perp^\top \boldsymbol{\Delta}\boldsymbol{L}\right\|_{\mathrm{F}}^2$, upper bound $\left\|\boldsymbol{L}_\perp^\top\boldsymbol{U}\right\|_{op} < 1$

Takeaway messages: speedup via spectral initialization

LoRA-One



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

- subspace alignment: $\boldsymbol{G}$ and $(\boldsymbol{A}_t, \boldsymbol{B}_t) \Rightarrow$ theory-grounded algorithm design
- "optimal" non-zero initialization strategy
- enables feature learning, precondition helps convergence...

Go beyond...

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

Go beyond...

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

Kai Lv, Yuqing Yang, Tengxiao Liu, Qipeng Guo, and Xipeng Qiu.
Full Parameter Fine-tuning for Large Language Models with Limited Resources.
In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8187–8198, 2024.

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.