Towards Understanding the Dynamics of Low-Rank Adaptation

Shu Ding, Yang Peng, Hangan Zhou, Xinyu Lu, Shangwei Chen,
Junhua Huang, Mingxuan Yuan, Wei Wang
{dings, pengy, zhouha, luxy, chensw, wangw}@lamda.nju.edu.cn
huang.hjh@outlook.com; Yuan.Mingxuan@huawei.com

ICML 2026

Outline

Outline

LoRA: Low-Rank Adaptation

Standard LoRA Formulation (Hu et al., ICLR 2022)

$$\mathbf{W} = \mathbf{W}_0 + \Delta\mathbf{W} = \mathbf{W}_0 + \frac{\alpha}{r}\mathbf{B}\mathbf{A} \tag{1}$$

- $\mathbf{W}_0 \in \mathbb{R}^{m\times n}$: frozen pre-trained weights
- $\mathbf{B} \in \mathbb{R}^{m\times r}$, $\mathbf{A} \in \mathbb{R}^{r\times n}$: learnable low-rank matrices
- $r \ll \min(m, n)$, α: rank and the scaling factor

Why does this work? Pre-trained models have low intrinsic dimension (Li et al., 2018; Aghajanyan et al., 2021)

Previous Update Dynamics of LoRA

How does LoRA update the weights? For fixed $\mathbf{A}$, update $\mathbf{B}$ via SGD:

$$\mathbf{B}_{t+1} = \mathbf{B}_t - \eta \nabla_{\mathbf{B}} f(\mathbf{W}_t), \quad \nabla_{\mathbf{B}} f(\mathbf{W}_t) = \nabla f(\mathbf{W}_t) \mathbf{A}_t^\top \tag{2}$$

Substitute into $\mathbf{W} = \mathbf{W}_0 + \frac{\alpha}{r}\mathbf{B}\mathbf{A}$:

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta \frac{\alpha}{r} \nabla f(\mathbf{W}_t) \underbrace{\mathbf{A}_t^\top \mathbf{A}_t}_{\text{Key matrix!}} \tag{3}$$

The update dynamics are governed by the quadratic form $\mathbf{A}_t^\top \mathbf{A}_t$.

Why $\mathbf{A}^\top\mathbf{A}$ is Problematic

$\mathbf{A}^\top\mathbf{A}$ is NOT a projection matrix for a general $\mathbf{A}$.

- Unstable optimization: $\mathbf{A}^\top\mathbf{A}$ can have arbitrary eigenvalues (depending on $\mathbf{A}$), Large eigenvalues amplify gradient noise and small eigenvalues cause vanishing gradients
- Information loss depends on eigenvalues: Badly conditioned $\mathbf{A}$ can lose most gradient information, different directions in the gradient are **scaled by different factors**

Existing Heuristics: Gaussian and Stiefel Initialization

Common practice: Make $\mathbf{A}^\top\mathbf{A}$ behave like a projection *in expectation*.

- Gaussian initialization: $\mathbf{A}_{ij} \sim \mathcal{N}(0, 1)$
- Stiefel manifold: $\mathbf{A}\mathbf{A}^\top = \mathbf{I}_r$

But expectation is not enough:

$$\mathbb{E}[\mathbf{A}^\top\mathbf{A}] = \mathbf{I}_r \quad \text{but} \quad \mathbf{A}^\top\mathbf{A} \neq \mathbf{I}_r \text{ for any single sample} \tag{4}$$

Gaussian and Stiefel Initialization can produce misaligned subspaces with non-zero probability.

Key Challenge:

How to choose $\mathbf{A}$ to maximize the information preserved from the unknown gradient $\nabla f(\mathbf{W})$?

Outline

Deriving the Update Dynamics of LoRA

From the L-smoothness condition $\|\nabla f(\mathbf{W}) - \nabla f(\mathbf{V})\| \leqslant L\|\mathbf{W} - \mathbf{V}\|$:

$$f(\mathbf{W}_{t+1}) \leqslant f(\mathbf{W}_t) + \langle \nabla f(\mathbf{W}_t), \Delta \mathbf{W}_t \rangle + \frac{L}{2}\|\Delta \mathbf{W}_t\|^2 \tag{5}$$

For fixed $\mathbf{A}_t$, minimizing the quadratic bound leads to:

$$\mathbf{B}_t^* = -\frac{r}{L\alpha}\nabla f(\mathbf{W}_t)\mathbf{A}_t^\top(\mathbf{A}_t\mathbf{A}_t^\top)^\dagger \tag{6}$$

Substituting back:

$$\Delta \mathbf{W}_t = -\frac{1}{L}\nabla f(\mathbf{W}_t)\underbrace{\mathbf{A}_t^\top(\mathbf{A}_t\mathbf{A}_t^\top)^\dagger\mathbf{A}_t}_{\mathbf{P}_{\mathbf{A}_t}} \tag{7}$$

The update dynamics are governed by the preconditioner $\mathbf{A}_t^\top(\mathbf{A}_t\mathbf{A}_t^\top)^\dagger\mathbf{A}_t$.

From $\mathbf{A}^\top\mathbf{A}$ to Projection Matrix $\mathbf{P}_\mathbf{A}$

Existing view (Hao et al., 2024):

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta\frac{\alpha}{r}\nabla f(\mathbf{W}_t)\mathbf{A}_t^\top\mathbf{A}_t \tag{8}$$

- $\mathbf{A}_t^\top\mathbf{A}_t$ is NOT a projection matrix
- Gradient scaling depends on eigenvalues of $\mathbf{A}_t$
- Ill-conditioned $\mathbf{A}_t$ causes unstable optimization

Our Prototypical LoRA:

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \frac{1}{L}\nabla f(\mathbf{W}_t)\mathbf{P}_{\mathbf{A}_t} \tag{9}$$

- $\mathbf{P}_{\mathbf{A}_t} = \mathbf{A}_t^\top(\mathbf{A}_t\mathbf{A}_t^\top)^\dagger\mathbf{A}_t$ is a projection matrix
- 0 or 1 eigenvalues, stable gradient update
- Matches LoRA with AdamW under proper regularization

The projection matrix $\mathbf{P}_\mathbf{A}$ defines the subspace where gradient information is preserved.

Which $\mathbf{A}$ Preserves the Most Gradient Information?

Now we have the formulation: $\mathbf{W}_{t+1} = \mathbf{W}_t - \frac{1}{L}\nabla f(\mathbf{W}_t)\mathbf{P}_{\mathbf{A}}$.

$\mathbf{P}_{\mathbf{A}}$ projects onto the **subspace** induced by $\mathbf{A}$. Different $\mathbf{A}$ give **different subspaces**.

For any gradient $\nabla f(\mathbf{W})$, the preserved energy is $\|\nabla f(\mathbf{W})\mathbf{P}_{\mathbf{A}}\|^2$.

- If we knew $\nabla f(\mathbf{W})$, we could set $\mathbf{P}_{\mathbf{A}}$ to project onto its top-r singular vectors via SVD
- $\nabla f(\mathbf{W})$ is unavailable during training!

Key Challenge:

How to choose $\mathbf{A}$ to maximize the information preserved from the unknown gradient $\nabla f(\mathbf{W})$?

What is an Equiangular Tight Frame (ETF)?

Definition (Equiangular Tight Frame (ETF))

$\mathbf{X} \in \mathbb{R}^{r \times n}$ is an ETF if:

1. **Unit-norm columns:** $\|\mathbf{X}_{:,i}\|_2 = 1, \forall i$
2. **Constant inner product:** $|\mathbf{X}_{:,i}^\top \mathbf{X}_{:,j}| = \delta = \sqrt{\frac{n-r}{r(n-1)}}, \forall i \neq j$
3. **Tight frame:** $\mathbf{X}\mathbf{X}^\top = \frac{n}{r}\mathbf{I}_r$

Why are ETFs special?

- They achieve the Welch bound — minimum possible coherence
- The columns are orthogonal and equally spaced (regular simplex in higher dimensions)

ETF Maximizes Worst-Case Gradient Preservation

Theorem (ETF Guarantees Maximum Gradient Preservation)

For $\mathbf{A}_t \in \mathbb{R}^{r\times n}$, *define* $\bar{\mathbf{A}}_t = (\mathbf{A}_t\mathbf{A}_t^\top)^{1/2}\mathbf{A}_t$ *and* $\mathbf{P}_{\mathbf{A}_t} = \bar{\mathbf{A}}_t^\top\bar{\mathbf{A}}_t$. *Let* $\nabla f(\mathbf{W})$ *have* s*-effective freedom with* $s < 1 + 1/\delta$, *where* $\delta = \sqrt{\frac{n-r}{r(n-1)}}$. *Define* $\Omega(\bar{\mathbf{A}}_t) = \inf_{\mathbf{W}} \frac{\|\nabla f(\mathbf{W})\mathbf{P}_{\mathbf{A}_t}\|^2}{\|\nabla f(\mathbf{W})\|^2}$, *then:*

$$\Omega(\bar{\mathbf{A}}_t) \leqslant \frac{r}{n}(1-\epsilon), \quad \epsilon = (s-1)\sqrt{\frac{n-r}{r(n-1)}}. \tag{10}$$

Equality holds iff $\bar{\mathbf{A}}_t$ *is a normalized ETF. When* $\mathbf{A}_t$ *is an ETF,* $\bar{\mathbf{A}}_t$ *is a normalized ETF.*

Worst-Case Information Preservation

- The projection matrix $\mathbf{P}_{\mathbf{A}_t}$ preserves at least a fixed fraction of the gradient energy
- Unlike Gaussian/Stiefel initialization, there is no risk of catastrophic energy loss

Computational Advantage

When $\mathbf{A}_t$ is an ETF, it gives:

$$(\mathbf{A}_t\mathbf{A}_t^\top)^\dagger = \left(\frac{n}{r}\mathbf{I}_r\right)^\dagger = \frac{r}{n}\mathbf{I}_r \tag{11}$$

- The update rule simplifies to $\mathbf{B}_t \leftarrow \mathbf{B}_t - \eta_{\mathbf{B}}\nabla_{\mathbf{B}} f(\mathbf{W}_t)\cdot\frac{r}{n}$
- No matrix inverse or pseudo-inverse needs to be computed during training
- Same computational cost as standard LoRA with Gaussian initialization

Convergence Analysis

Theorem (Convergence of LoRA for smooth loss functions)

Under L-smoothness with ETF initialization, the Prototypical LoRA satisfies:

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{W}_t)\|^2 \leqslant \frac{2nL}{r(1-\epsilon)T}(f(\mathbf{W}_0)-f^*)$$

Theorem (Convergence of LoRA for smooth loss functions under the PL-condition)

Under L-smoothness and PL condition with ETF initialization, the Prototypical LoRA satisfies:

$$f(\mathbf{W}_t)-f^* \leqslant \left(1-\frac{\mu r}{nL}(1-\epsilon)\right)^t (f(\mathbf{W}_0)-f^*)$$

Outline

Experimental Setup

Tasks:

- NLU: GLUE benchmark (CoLA, SST-2, MRPC, QQP, MNLI, QNLI, RTE)
- NLG: GSM8K (math), HumanEval/MBPP (code), MT-Bench (instruction)

Models:

- RoBERTa — NLU tasks
- Llama-3.1-8B — NLU + NLG tasks
- Mistral-7B-v0.1 — NLG tasks

Baselines: LoRA, LoRA-GA, PiSSA, FLORA, AsymLoRA, ReLoRA, RAC-LoRA, GaLore

NLU Results: RoBERTa on GLUE

Table 1. Test accuracy(%) of RoBERTa with LoRA variants (with/without ETF) on GLUE benchmark datasets (rank 8).

Method	ETF	CoLA	SST-2	MRPC	QQP	MNLI	QNLI	RTE
LoRA	w/o	$57.95_{\pm 1.23}$	$94.19_{\pm 0.05}$	$87.83_{\pm 0.99}$	$90.38_{\pm 0.04}$	$86.27_{\pm 0.06}$	$91.53_{\pm 0.05}$	$73.05_{\pm 1.47}$
	w/	$62.18_{\pm 0.87}$ (+4.23)	$95.22_{\pm 0.11}$ (+1.03)	$88.89_{\pm 1.29}$ (+1.06)	$91.47_{\pm 0.06}$ (+1.09)	$87.48_{\pm 0.08}$ (+1.21)	$92.65_{\pm 0.11}$ (+1.12)	$74.97_{\pm 1.73}$ (+1.92)
LoRA-GA	w/o	$58.70_{\pm 0.82}$	$94.43_{\pm 0.23}$	$87.52_{\pm 0.62}$	$90.45_{\pm 0.02}$	$86.37_{\pm 0.16}$	$91.88_{\pm 0.12}$	$73.48_{\pm 1.81}$
	w/	$62.18_{\pm 0.87}$ (+3.48)	$95.22_{\pm 0.11}$ (+0.79)	$88.89_{\pm 1.29}$ (+1.37)	$91.47_{\pm 0.06}$ (+1.02)	$87.48_{\pm 0.08}$ (+1.11)	$92.65_{\pm 0.11}$ (+0.77)	$74.97_{\pm 1.73}$ (+1.49)
PiSSA	w/o	$58.14_{\pm 0.71}$	$94.31_{\pm 0.19}$	$87.49_{\pm 1.26}$	$90.32_{\pm 0.02}$	$85.94_{\pm 0.05}$	$91.56_{\pm 0.14}$	$74.37_{\pm 1.64}$
	w/	$62.18_{\pm 0.87}$ (+4.04)	$95.22_{\pm 0.11}$ (+0.91)	$88.89_{\pm 1.29}$ (+1.40)	$91.47_{\pm 0.06}$ (+1.15)	$87.48_{\pm 0.08}$ (+1.54)	$92.65_{\pm 0.11}$ (+1.09)	$74.97_{\pm 1.73}$ (+0.60)
FLORA	w/o	$59.87_{\pm 1.27}$	$93.96_{\pm 0.30}$	$87.01_{\pm 0.53}$	$90.29_{\pm 0.06}$	$86.51_{\pm 0.04}$	$91.92_{\pm 0.15}$	$72.08_{\pm 0.55}$
	w/	$62.42_{\pm 1.18}$ (+2.55)	$95.22_{\pm 0.11}$ (+1.26)	$88.89_{\pm 0.64}$ (+1.88)	$91.37_{\pm 0.04}$ (+1.08)	$87.52_{\pm 0.06}$ (+1.01)	$92.94_{\pm 0.08}$ (+1.02)	$74.85_{\pm 0.90}$ (+2.77)
AsymLoRA	w/o	$58.47_{\pm 0.67}$	$94.15_{\pm 0.12}$	$87.52_{\pm 0.75}$	$90.03_{\pm 0.18}$	$85.86_{\pm 0.13}$	$91.43_{\pm 0.14}$	$72.20_{\pm 1.28}$
	w/	$60.57_{\pm 0.54}$ (+2.10)	$95.11_{\pm 0.14}$ (+0.96)	$88.64_{\pm 0.42}$ (+1.12)	$91.05_{\pm 0.05}$ (+1.02)	$86.96_{\pm 0.10}$ (+1.10)	$92.50_{\pm 0.11}$ (+1.07)	$73.65_{\pm 1.56}$ (+1.45)
ReLoRA	w/o	$57.14_{\pm 1.94}$	$91.97_{\pm 0.34}$	$87.01_{\pm 0.87}$	$89.66_{\pm 0.02}$	$82.79_{\pm 0.62}$	$90.56_{\pm 0.26}$	$71.24_{\pm 1.11}$
	w/	$62.11_{\pm 0.90}$ (+4.97)	$93.35_{\pm 0.23}$ (+1.38)	$88.07_{\pm 0.90}$ (+1.06)	$90.70_{\pm 0.02}$ (+1.04)	$85.16_{\pm 0.02}$ (+2.37)	$91.97_{\pm 0.10}$ (+1.41)	$72.44_{\pm 0.90}$ (+1.20)
RAC-LoRA	w/o	$55.35_{\pm 1.23}$	$93.69_{\pm 0.30}$	$86.93_{\pm 0.37}$	$90.02_{\pm 0.05}$	$85.89_{\pm 0.01}$	$91.48_{\pm 0.10}$	$71.84_{\pm 1.64}$
	w/	$58.84_{\pm 0.87}$ (+3.49)	$94.72_{\pm 0.19}$ (+1.03)	$88.56_{\pm 0.42}$ (+1.63)	$91.01_{\pm 0.02}$ (+0.99)	$86.92_{\pm 0.11}$ (+1.03)	$92.49_{\pm 0.13}$ (+1.01)	$73.65_{\pm 1.35}$ (+1.81)

Figure 1: LoRA variants with ETF **consistently outperform** their default initialization counterparts across all GLUE datasets.

NLU Results: Llama on GLUE

Table 7. Test accuracy(%) of Llama-3.1-8B with LoRA variants (with/without ETF) on GLUE benchmark datasets (rank 8).

Method	ETF	CoLA	SST-2	MRPC	QQP	MNLI	QNLI	RTE
LoRA	w/o	$67.56_{\pm 0.23}$	$95.46_{\pm 0.05}$	$87.75_{\pm 0.35}$	$90.54_{\pm 0.11}$	$90.34_{\pm 0.03}$	$94.19_{\pm 0.05}$	$84.44_{\pm 0.85}$
	w/	$68.75_{\pm 0.20}$ (+1.19)	$96.44_{\pm 0.18}$ (+0.98)	$89.95_{\pm 0.44}$ (+2.20)	$91.56_{\pm 0.20}$ (+1.02)	$91.57_{\pm 0.08}$ (+1.23)	$95.21_{\pm 0.05}$ (+1.02)	$86.48_{\pm 0.10}$ (+2.04)
LoRA-GA	w/o	$67.27_{\pm 0.88}$	$95.22_{\pm 0.12}$	$88.89_{\pm 0.69}$	$90.49_{\pm 0.86}$	$90.64_{\pm 0.08}$	$94.42_{\pm 0.04}$	$86.20_{\pm 1.30}$
	w/	$68.75_{\pm 0.20}$ (+1.48)	$96.44_{\pm 0.18}$ (+1.22)	$89.95_{\pm 0.44}$ (+1.06)	$91.56_{\pm 0.20}$ (+1.07)	$91.57_{\pm 0.08}$ (+0.93)	$95.21_{\pm 0.05}$ (+0.79)	$86.48_{\pm 0.10}$ (+0.28)
PiSSA	w/o	$66.86_{\pm 1.42}$	$94.90_{\pm 0.25}$	$87.50_{\pm 0.31}$	$90.45_{\pm 0.06}$	$90.47_{\pm 0.09}$	$94.07_{\pm 0.16}$	$85.92_{\pm 1.33}$
	w/	$68.75_{\pm 0.20}$ (+1.89)	$96.44_{\pm 0.18}$ (+1.54)	$89.95_{\pm 0.44}$ (+2.45)	$91.56_{\pm 0.20}$ (+1.11)	$91.57_{\pm 0.08}$ (+1.10)	$95.21_{\pm 0.05}$ (+1.14)	$86.48_{\pm 0.10}$ (+0.56)
FLORA	w/o	$65.87_{\pm 1.38}$	$95.56_{\pm 0.06}$	$87.99_{\pm 0.98}$	$89.68_{\pm 0.05}$	$89.22_{\pm 0.11}$	$93.39_{\pm 0.05}$	$85.30_{\pm 0.54}$
	w/	$67.30_{\pm 1.77}$ (+1.43)	$96.90_{\pm 0.03}$ (+1.34)	$89.95_{\pm 0.12}$ (+1.96)	$90.99_{\pm 0.05}$ (+1.31)	$91.41_{\pm 0.23}$ (+2.19)	$94.98_{\pm 0.29}$ (+1.59)	$86.32_{\pm 1.81}$ (+1.02)
AsymLoRA	w/o	$55.65_{\pm 1.10}$	$94.42_{\pm 0.05}$	$75.00_{\pm 0.81}$	$88.07_{\pm 0.01}$	$89.99_{\pm 0.03}$	$92.38_{\pm 0.08}$	$73.18_{\pm 1.87}$
	w/	$60.69_{\pm 1.87}$ (+5.04)	$95.54_{\pm 0.06}$ (+1.12)	$77.22_{\pm 0.12}$ (+2.22)	$89.44_{\pm 0.06}$ (+1.37)	$90.94_{\pm 0.01}$ (+0.95)	$93.46_{\pm 0.12}$ (+1.08)	$75.08_{\pm 1.43}$ (+1.90)
ReLoRA	w/o	$66.31_{\pm 0.56}$	$95.33_{\pm 0.16}$	$85.50_{\pm 0.12}$	$89.21_{\pm 0.06}$	$90.29_{\pm 0.01}$	$92.36_{\pm 0.18}$	$84.43_{\pm 1.43}$
	w/	$67.33_{\pm 0.82}$ (+1.02)	$96.44_{\pm 0.22}$ (+1.11)	$89.22_{\pm 0.62}$ (+3.72)	$90.43_{\pm 0.44}$ (+1.22)	$91.39_{\pm 0.23}$ (+1.10)	$93.54_{\pm 0.07}$ (+1.18)	$85.56_{\pm 0.36}$ (+1.13)
RAC-LoRA	w/o	$58.11_{\pm 0.83}$	$94.30_{\pm 0.05}$	$75.73_{\pm 0.12}$	$89.92_{\pm 0.06}$	$89.52_{\pm 0.02}$	$92.46_{\pm 0.13}$	$73.54_{\pm 1.72}$
	w/	$59.19_{\pm 1.38}$ (+1.08)	$95.50_{\pm 0.19}$ (+1.20)	$80.02_{\pm 0.56}$ (+4.29)	$90.83_{\pm 0.65}$ (+0.91)	$90.63_{\pm 0.16}$ (+1.11)	$94.51_{\pm 0.32}$ (+2.05)	$75.08_{\pm 0.61}$ (+1.54)

Figure 2: LoRA variants with ETF **consistently outperform** their default initialization counterparts across all GLUE datasets.

ETF benefits both encoder-only (RoBERTa) and decoder-only (Llama) models.

Comparison with GaLore

Table 2. Test accuracy(%) of RoBERTa with Galore (with/without ETF) on GLUE benchmark datasets.

Rank	ETF	CoLA	SST-2	MRPC	QQP	MNLI	QNLI	RTE
4	w/o	$58.46_{\pm 0.52}$	$95.03_{\pm 0.05}$	$86.85_{\pm 0.61}$	$91.08_{\pm 0.04}$	$87.32_{\pm 0.11}$	$92.64_{\pm 0.09}$	$68.83_{\pm 1.33}$
	w/	$59.55_{\pm 0.48}$ (+1.09)	$94.88_{\pm 0.26}$ (-0.15)	$88.32_{\pm 0.95}$ (+1.47)	$90.15_{\pm 0.04}$ (-0.93)	$87.18_{\pm 0.08}$ (-0.14)	$92.35_{\pm 0.19}$ (-0.29)	$71.72_{\pm 1.12}$ (+2.89)
8	w/o	$59.05_{\pm 0.94}$	$95.15_{\pm 0.22}$	$87.09_{\pm 0.12}$	$91.25_{\pm 0.06}$	$87.47_{\pm 0.02}$	$92.92_{\pm 0.07}$	$71.84_{\pm 1.84}$
	w/	$61.61_{\pm 0.16}$ (+2.56)	$95.12_{\pm 0.06}$ (-0.03)	$88.48_{\pm 0.20}$ (+1.39)	$91.29_{\pm 0.04}$ (+0.04)	$87.66_{\pm 0.03}$ (+0.19)	$92.58_{\pm 0.17}$ (-0.34)	$74.01_{\pm 0.78}$ (+2.17)
16	w/o	$59.91_{\pm 1.00}$	$95.22_{\pm 0.18}$	$87.13_{\pm 0.86}$	$91.42_{\pm 0.04}$	$87.64_{\pm 0.10}$	$92.71_{\pm 0.12}$	$73.65_{\pm 2.21}$
	w/	$61.76_{\pm 0.67}$ (+1.85)	$95.26_{\pm 0.13}$ (+0.04)	$88.40_{\pm 1.10}$ (+1.27)	$91.45_{\pm 0.13}$ (+0.03)	$87.67_{\pm 0.03}$ (+0.03)	$92.95_{\pm 0.06}$ (+0.24)	$74.13_{\pm 1.33}$ (+0.48)

Figure 3: GaLore performs SVD on $\nabla f(\mathbf{W})$ periodically to define the projection subspace. Our ETF initialization achieves comparable performance without computing the full gradient.

ETF is a universally optimal initialization strategy; SVD is instance-specific.

NLG Results

Table 3. Test score(%) of Llama-3.1-8B and Mistral-7B-v0.1 with LoRA variants (with/without ETF) on NLG datasets (rank 8).

Method	ETF	Llama-3.1-8B				Mistral-7B-v0.1			
		GSM8K	HumanEval	MBPP	MT-Bench	GSM8K	HumanEval	MBPP	MT-Bench
LoRA	w/o	$68.33_{\pm 0.27}$	$42.48_{\pm 0.58}$	$50.23_{\pm 0.34}$	$6.06_{\pm 0.04}$	$69.61_{\pm 0.19}$	$29.26_{\pm 1.36}$	$40.21_{\pm 0.16}$	$6.13_{\pm 0.03}$
	w/	$71.07_{\pm 0.19}$ (+2.74)	$46.34_{\pm 1.22}$ (+3.86)	$51.45_{\pm 0.57}$ (+1.22)	$6.45_{\pm 0.11}$ (+0.39)	$71.40_{\pm 0.36}$ (+1.79)	$32.31_{\pm 0.53}$ (+3.05)	$42.07_{\pm 0.58}$ (+1.86)	$6.53_{\pm 0.06}$ (+0.40)
LoRA-GA	w/o	$70.94_{\pm 0.12}$	$43.39_{\pm 0.66}$	$49.01_{\pm 0.12}$	$6.32_{\pm 0.05}$	$70.07_{\pm 0.27}$	$31.09_{\pm 0.75}$	$41.35_{\pm 0.16}$	$6.06_{\pm 0.03}$
	w/	$71.07_{\pm 0.19}$ (+0.13)	$46.34_{\pm 1.22}$ (+2.95)	$51.45_{\pm 0.57}$ (+2.44)	$6.45_{\pm 0.11}$ (+0.13)	$71.40_{\pm 0.36}$ (+1.33)	$32.31_{\pm 0.53}$ (+1.22)	$42.07_{\pm 0.58}$ (+0.72)	$6.53_{\pm 0.06}$ (+0.47)
PiSSA	w/o	$67.08_{\pm 0.24}$	$42.31_{\pm 0.73}$	$49.84_{\pm 0.48}$	$6.37_{\pm 0.09}$	$70.96_{\pm 0.91}$	$30.87_{\pm 0.25}$	$40.24_{\pm 0.66}$	$6.15_{\pm 0.04}$
	w/	$71.07_{\pm 0.19}$ (+3.99)	$46.34_{\pm 1.22}$ (+4.03)	$51.45_{\pm 0.57}$ (+1.61)	$6.45_{\pm 0.11}$ (+0.08)	$71.40_{\pm 0.36}$ (+0.44)	$32.31_{\pm 0.53}$ (+1.44)	$42.07_{\pm 0.58}$ (+1.83)	$6.53_{\pm 0.06}$ (+0.38)
FLORA	w/o	$71.22_{\pm 0.12}$	$43.90_{\pm 0.75}$	$46.15_{\pm 0.31}$	$6.13_{\pm 0.06}$	$70.34_{\pm 1.37}$	$38.31_{\pm 1.33}$	$40.83_{\pm 0.29}$	$5.96_{\pm 0.07}$
	w/	$73.40_{\pm 0.04}$ (+2.18)	$46.95_{\pm 1.22}$ (+3.05)	$48.67_{\pm 0.21}$ (+2.52)	$6.26_{\pm 0.03}$ (+0.13)	$71.94_{\pm 0.57}$ (+1.60)	$39.63_{\pm 0.85}$ (+1.32)	$42.52_{\pm 0.14}$ (+1.69)	$6.07_{\pm 0.05}$ (+0.11)
AsymLoRA	w/o	$63.03_{\pm 0.35}$	$42.07_{\pm 0.86}$	$50.07_{\pm 0.41}$	$5.91_{\pm 0.05}$	$64.93_{\pm 0.32}$	$26.37_{\pm 1.06}$	$39.63_{\pm 0.45}$	$6.21_{\pm 0.03}$
	w/	$66.06_{\pm 0.41}$ (+3.03)	$43.59_{\pm 0.30}$ (+1.52)	$51.22_{\pm 0.19}$ (+1.15)	$6.25_{\pm 0.10}$ (+0.34)	$66.84_{\pm 0.50}$ (+1.91)	$29.87_{\pm 1.22}$ (+3.50)	$40.62_{\pm 0.09}$ (+0.99)	$6.26_{\pm 0.02}$ (+0.05)
ReLoRA	w/o	$67.25_{\pm 0.30}$	$43.29_{\pm 2.17}$	$50.53_{\pm 0.33}$	$6.38_{\pm 0.06}$	$67.60_{\pm 0.55}$	$29.41_{\pm 0.67}$	$40.05_{\pm 0.51}$	$6.03_{\pm 0.01}$
	w/	$70.35_{\pm 0.30}$ (+3.10)	$46.95_{\pm 0.61}$ (+3.66)	$51.04_{\pm 0.52}$ (+0.51)	$6.42_{\pm 0.07}$ (+0.04)	$70.88_{\pm 0.38}$ (+3.28)	$30.48_{\pm 0.56}$ (+1.07)	$41.03_{\pm 0.05}$ (+0.98)	$6.38_{\pm 0.02}$ (+0.35)
RAC-LoRA	w/o	$60.30_{\pm 0.27}$	$41.16_{\pm 0.91}$	$50.45_{\pm 0.46}$	$6.02_{\pm 0.01}$	$64.52_{\pm 0.90}$	$27.89_{\pm 0.52}$	$40.28_{\pm 0.36}$	$5.93_{\pm 0.09}$
	w/	$63.33_{\pm 0.14}$ (+3.03)	$43.44_{\pm 1.80}$ (+2.28)	$52.26_{\pm 0.30}$ (+1.81)	$6.07_{\pm 0.08}$ (+0.05)	$66.61_{\pm 0.71}$ (+2.09)	$29.26_{\pm 0.44}$ (+1.37)	$41.47_{\pm 0.16}$ (+1.19)	$6.16_{\pm 0.07}$ (+0.23)
GaLore	w/o	$76.21_{\pm 0.25}$	$49.26_{\pm 0.30}$	$50.27_{\pm 0.13}$	$6.40_{\pm 0.02}$	$71.53_{\pm 1.28}$	$43.46_{\pm 1.74}$	$46.35_{\pm 0.25}$	$6.20_{\pm 0.01}$
	w/	$78.71_{\pm 0.75}$ (+2.50)	$52.43_{\pm 0.61}$ (+3.17)	$50.42_{\pm 0.42}$ (+0.15)	$6.54_{\pm 0.07}$ (+0.14)	$72.93_{\pm 0.37}$ (+1.40)	$46.95_{\pm 1.22}$ (+3.49)	$48.18_{\pm 0.41}$ (+1.83)	$6.43_{\pm 0.04}$ (+0.23)

Figure 4: ETF improves the performance on math reasoning (GSM8K), code generation (HumanEval, MBPP), and instruction following (MT-Bench).

Visualization

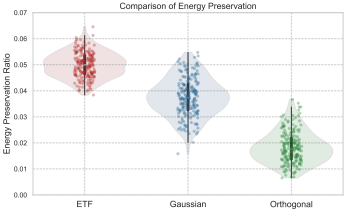

Figure 5: ETF consistently achieves higher average preservation and **higher minimum (worst-case) preservation** than Gaussian or Orthogonal initialization.

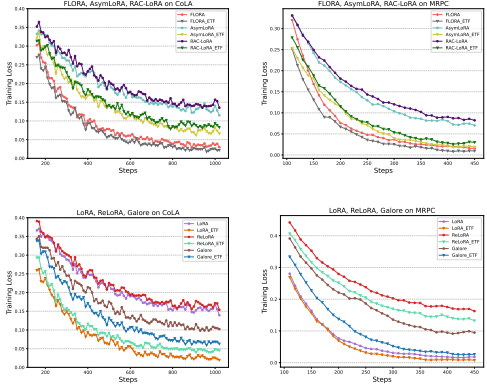

Figure 6: ETF preserves more gradient information per iteration, resulting in more efficient descent.

Outline

Summary of Contributions

1. **Theoretical dynamics:** Derived the appropriate update dynamics of LoRA as $\mathbf{W}_{t+1} = \mathbf{W}_t - \frac{1}{L}\nabla f(\mathbf{W}_t)\mathbf{P}_{\mathbf{A}_t}$, where $\mathbf{P}_{\mathbf{A}_t}$ is a projection matrix
2. **Optimal initialization:** Proved that ETF achieves the maximum lower bound for gradient information preservation when ∇f is unknown (Theorem 3.3)
3. **Convergence guarantees:** Established $\mathcal{O}(1/T)$ convergence for smooth losses and exponential convergence under PL condition
4. **Empirical validation:** ETF initialization is a drop-in replacement for existing initialization: no architectural changes, no additional hyperparameters, significant performance gains.

Thank you!