LoRA Done RITE: Robust Invariant Transformation Equilibration for Lora Optimization

Jui-Nan Yen *[1] Si Si [2] Zhao Meng [2] Felix Yu [2] Sai Surya Duvvuri *[3]
Inderjit S. Dhillon [2] Cho-Jui Hsieh [12] Sanjiv Kumar[2]

*Work done while at Google

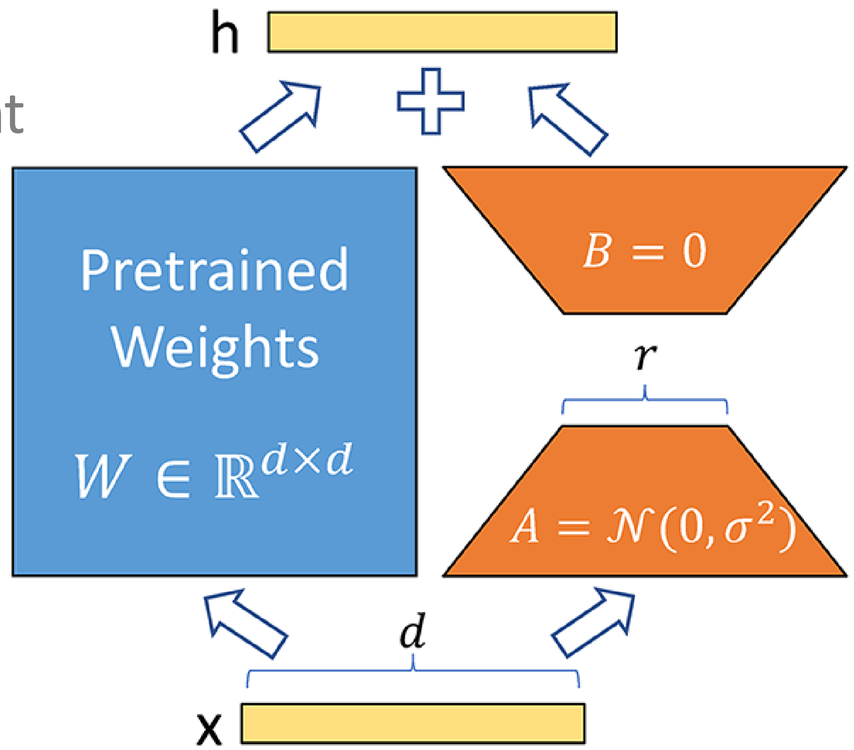
UCLA Samueli Computer Science [1] Google [2] TEXAS The University of Texas at Austin [3]

LoRA Architecture

- LoRA decomposes the model weight into two matrices W and H.
- H has a low-rank structure:

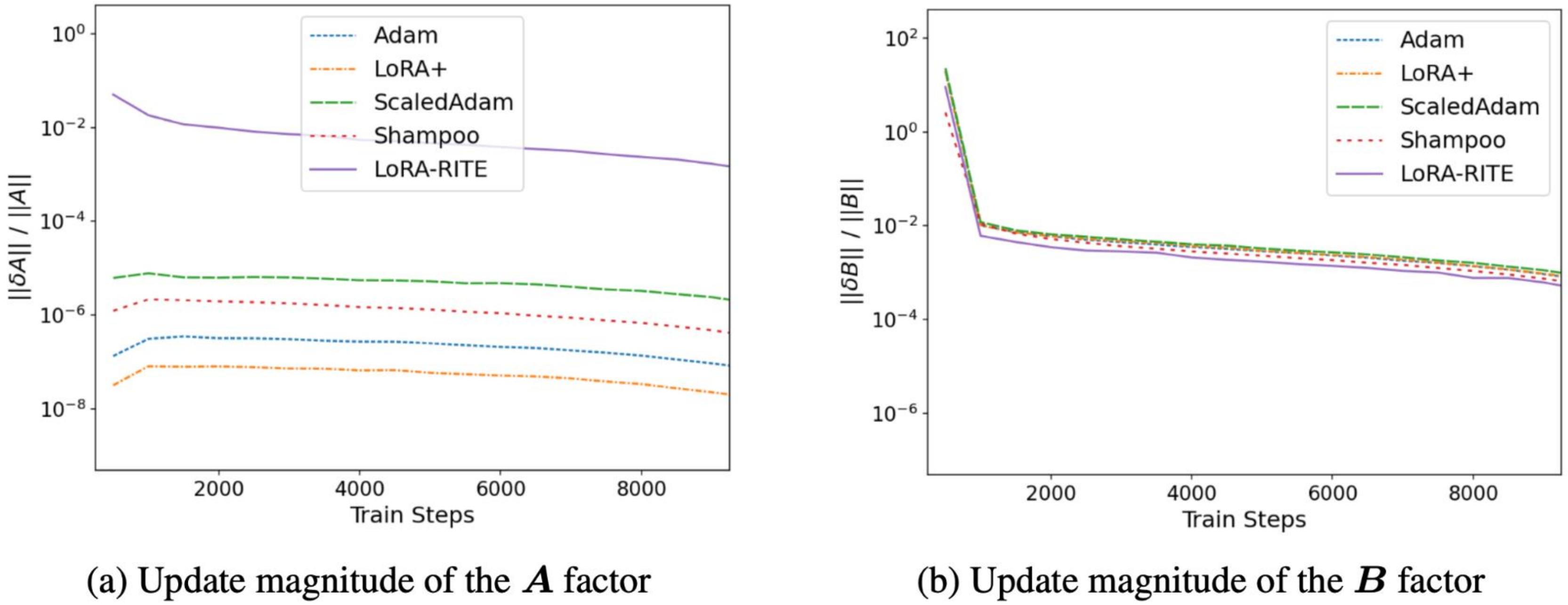
$$H = AB^T$$

- $h = (W + H)x = Wx + AB^Tx.$



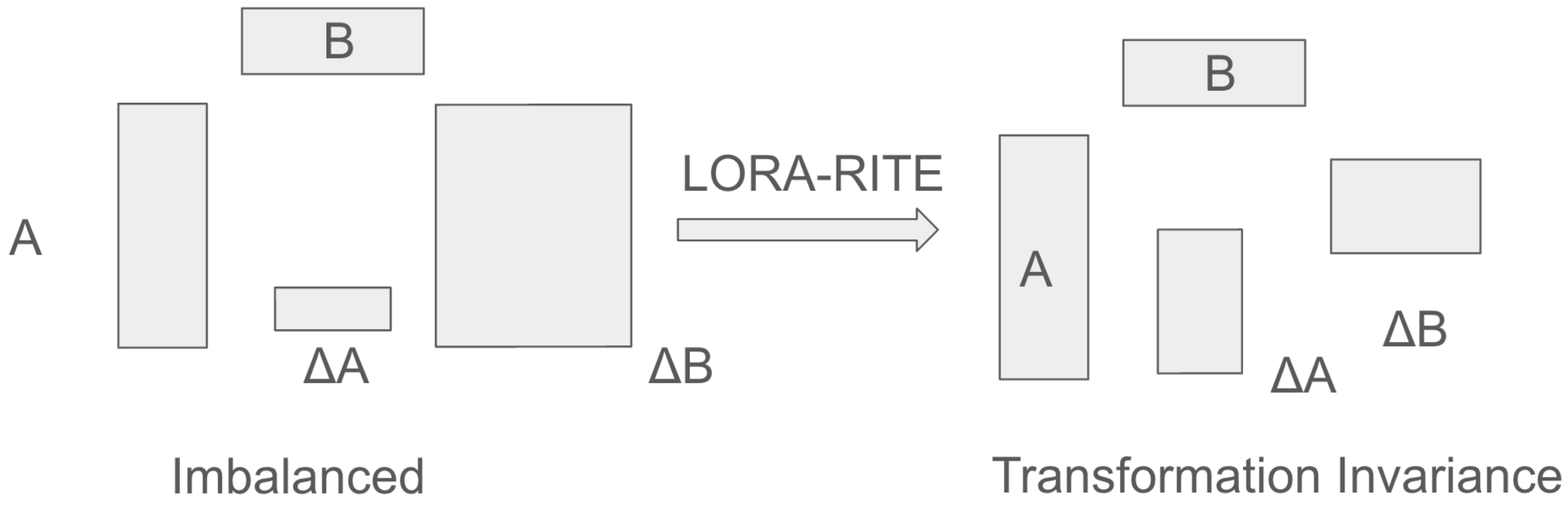
Issues of Lora Training

- Factor A gets **extremely small** updates compared to Factor B with traditional optimizers (blue line for Adam).
- Our new optimizer LoRA-RITE (purple line) addresses this issue.



A new optimizer for LoRA: LORA-RITE

- LoRA-RITE is **transformation invariant**, prevents imbalanced updates.



Definition of Invariance

- Equivalent Lora pairs:

$$H = A_1 B_1^T = A_2 B_2^T$$

- **Invariance**: for equivalent Lora pairs, an optimizer should produce the same update to *H*,

$$\begin{aligned} H + \Delta H &= (A_1 + \Delta A_1)(B_1 + \Delta B_1)^T \\ &= (A_2 + \Delta A_2)(B_2 + \Delta B_2)^T \end{aligned}$$

Transformation Invariance

- Transformation Invariance:

$$\boldsymbol{A}_2 = \boldsymbol{A}_1 \boldsymbol{R}, \boldsymbol{B}_2 = \boldsymbol{B}_1 \boldsymbol{R}^{-\top}$$

- Scale Invariance:

$$\boldsymbol{A}_2 = s\boldsymbol{A}_1, \boldsymbol{B}_2 = (1/s)\boldsymbol{B}_1$$

- None of the existing optimizers, e.g., Adam, are even scale invariant.

Lack of Invariance Can Lead to Imbalanced Updates

- For SGD, let

$$\boldsymbol{A}_2 = s\boldsymbol{A}_1, \boldsymbol{B}_2 = (1/s)\boldsymbol{B}_1,$$

we have

$$\Delta \boldsymbol{A}_2 = -\nabla \boldsymbol{H}\boldsymbol{B}_2 = -(1/s)\nabla \boldsymbol{H}\boldsymbol{B}_1 = (1/s)\Delta \boldsymbol{A}_1 \quad (\text{since } \boldsymbol{B}_2 = (1/s)\boldsymbol{B}_1)$$
$$\Delta \boldsymbol{B}_2 = -\nabla \boldsymbol{H}^\top \boldsymbol{A}_2 = -s\nabla \boldsymbol{H}^\top \boldsymbol{A}_1 = s\Delta \boldsymbol{B}_1 \quad (\text{since } \boldsymbol{A}_2 = s\boldsymbol{A}_1).$$

- This leads to imbalanced updates

$$\|\Delta \boldsymbol{A}_2\|/\|\boldsymbol{A}_2\| = (1/s^2)\|\Delta \boldsymbol{A}_1\|/\|\boldsymbol{A}_1\|$$
$$\|\Delta \boldsymbol{B}_2\|/\|\boldsymbol{B}_2\| = (s^2)\|\Delta \boldsymbol{B}_1\|/\|\boldsymbol{B}_1\|$$

Decomposition of Basis and Magnitude

- The lora factors can be decomposed into directions (basis) and magnitudes:

$$A = \underbrace{U_A}_{\text{basis}} \underbrace{R_A}_{\text{magnitude}}, B = U_B R_B$$

- Equivalent lora pairs

$$H = A_1 B_1^T = A_2 B_2^T$$

have equivalent directions (basis) but different magnitudes

$$\mathrm{span}(U_{A_1}) = \mathrm{span}(U_{A_2}) \quad \mathrm{span}(U_{B_1}) = \mathrm{span}(U_{B_2})$$

Unmagnified Gradient

- We define the unmagnified gradient,

$$\nabla A = \nabla H B \longrightarrow \bar{\nabla} A = \nabla H U_B$$

which removes the influence of the lora weight magnitudes.

- This is the same for all equivalent lora pairs up to the choice of the basis.
- Can be seen as the projection of the full gradient to the lora basis.

LoRA-Rite Update

- Our update

$$\Delta A = \underbrace{\boxed{\bar{\nabla} A(\bar{L}_A)^{-1/2}}}_{\text{without magnitude}} \underbrace{\boxed{R_B^{-T}}}_{\text{scaling}}, \quad \bar{L}_A = (\bar{\nabla} A)^T \bar{\nabla} A$$

- Scaling cancels out the multiplication of A/B and achieves transformation invariance

$$\begin{aligned} \Delta A_1 B_1^T &= \bar{\nabla} A_1 (\bar{L}_{A_1})^{-1/2} (R_{B_1}^{-T} B_1^T) = \bar{\nabla} A_1 (\bar{L}_{A_1})^{-1/2} U_{B_1}^T \\ &= \bar{\nabla} A_2 (\bar{L}_{A_2})^{-1/2} U_{B_2}^T = \Delta A_2 B_2^T \end{aligned}$$

Incorporating Momentum

- To incorporate momentum while maintaining transformation invariance, one must account for the varying basis at each step.
- We use

  $$\boldsymbol{P}_{\boldsymbol{A}_t} := (\boldsymbol{U}_{\boldsymbol{B}_t})^\top \boldsymbol{U}_{\boldsymbol{B}_{t-1}}$$

  to transform the momentum to the new basis

  $$\bar{\boldsymbol{M}}_{\boldsymbol{A}_t} = \beta_1 \bar{\boldsymbol{M}}_{\boldsymbol{A}_{t-1}} \boldsymbol{P}_{\boldsymbol{A}_t}^\top + (1-\beta_1) \bar{\boldsymbol{S}}_{\boldsymbol{A}_t}$$

  before the accumulation of the preconditioned updates $\bar{\boldsymbol{S}}_{\boldsymbol{A}_t}$
- Update with momentum

  $$\Delta A = \bar{\boldsymbol{M}}_{\boldsymbol{A}_t} R_B^{-T}$$

Algorithmic Comparison with Adam

- We use unmagnified version of the gradients to remove the influence of weight magnitudes.
- We transform the momentum at each step for correct accumulation.

Complexity

- Let *r* be the rank of LoRA, and *n* is the dimension of the weight matrix, *n>>r.*
- LoRA-RITE is faster than Shampoo, with time/space complexity similar to Adam.

	Time complexity	Space complexity
Forward/Backward	$\Omega(n^2)$	$\Omega(n^2)$
Adam (first-order)	$O(nr)$	$O(nr)$
Shampoo (second-order)	$O(n^3+r^3)$	$O(n^2+r^2)$
LoRA-RITE (second-order)	$O(nr^2)$	$O(nr+r^2)$

Experiment - Public LLM Benchmarks

Table 2: Experimental results on LLM benchmarking datasets.

Model	Optimizer	HellaSwag	ArcChallenge	GSM8K	OpenBookQA	Avg.
Gemma-2B	Adam	83.76	45.31	24.26	64.0	54.33
	LoRA+	83.75	45.31	23.65	64.4	54.28
	ScaledAdam	83.52	45.22	23.96	64.8	54.38
	Shampoo	83.26	44.88	23.35	63.6	53.77
	Lamb	86.60	47.35	26.76	68.0	57.18
	LoRA-RITE	**87.28**	**49.06**	**30.10**	**68.8**	**58.81**
Gemma-7B	Adam	94.07	54.78	48.37	77.60	68.71
	LoRA+	93.99	54.01	48.75	77.60	68.59
	ScaledAdam	93.31	52.90	48.07	75.80	67.52
	Shampoo	94.15	52.47	49.05	76.80	68.12
	Lamb	95.11	69.80	50.64	83.20	74.69
	LoRA-RITE	**95.59**	**71.76**	**55.50**	**84.80**	**76.91**

Experiment - Public LLM Benchmarks

Table 2: Experimental results on LLM benchmarking datasets.

Model	Optimizer	HellaSwag	ArcChallenge	GSM8K	OpenBookQA	Avg.
Gemma-2B	Adam	83.76	45.31	24.26	64.0	54.33
	LoRA+	83.75	45.31	23.65	64.4	54.28
	ScaledAdam	83.52	45.22	23.96	64.8	54.38
	Shampoo	83.26	44.88	23.35	63.6	53.77
	Lamb	86.60	47.35	26.76	68.0	57.18
	LoRA-RITE	**87.28**	**49.06**	**30.10**	**68.8**	**58.81**
Gemma-7B	Adam	94.07	54.78	48.37	77.60	68.71
	LoRA+	93.99	54.01	48.75	77.60	68.59
	ScaledAdam	93.31	52.90	48.07	75.80	67.52
	Shampoo	94.15	52.47	49.05	76.80	68.12
	Lamb	95.11	69.80	50.64	83.20	74.69
	LoRA-RITE	**95.59**	**71.76**	**55.50**	**84.80**	**76.91**

3.5 and 8.2 percentage point of accuracy gain over Adam!

Thank you!