OAK RIDGE National Laboratory

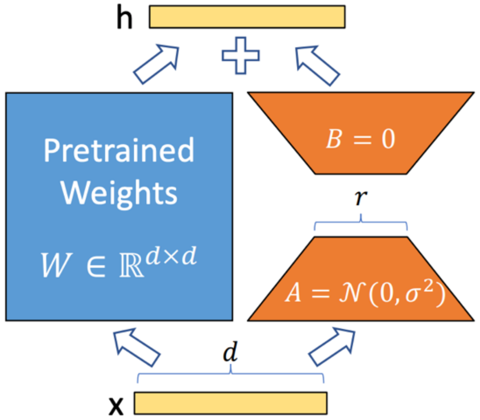
San Diego - Dec 5th

A geometric framework for momentum-based optimizers for low-rank training

Steffen Schotthöfer, Timon Klein, Jonas Kusch

ORNL, OVGU, NMBU

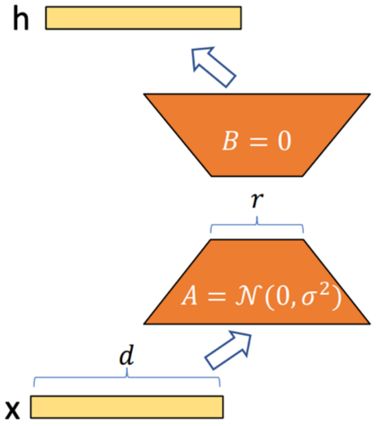
Low Rank Compression and Fine-Tuning of Neural Networks



Low Rank Finetuning

- Hu et al., *LoRA: Low-Rank Adaptation of Large Language Models. 2021*
- Zhang et al., *AdaLoRA: Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning.* 2023
- Liu et al., *DoRA: Weight-Decomposed Low-Rank Adaptation. 2024*

$$z(x) = \sigma(Wx + AB^\top x)$$



Low Rank Compression

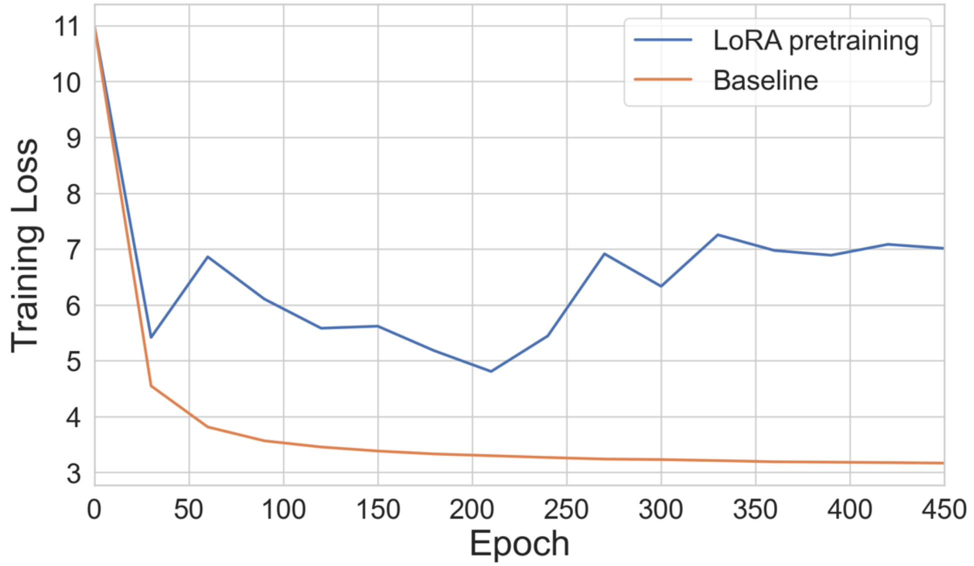
- Denton et al., *Exploiting Linear Structure Within Convolutional Networks.* 2014
- A. Novikov, et al., *Tensorizing neural networks*. 2015
- A. Tjandra, S. Sakti, and S. Nakamura. *Compressing recurrent neural network with tensor train*. 2017

$$z(x) = \sigma(AB^\top x)$$

Low Rank Attention

- DeepSeek-AI, *"Multihhead Latent Attention"*. 2024
- Ainslie, et al., *GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. '23*

$$z(x) = \sigma(xAB^\top x^\top)$$



GPT-2 on OWT, low-rank MLP: Low rank training stalls. Why?

LoRA

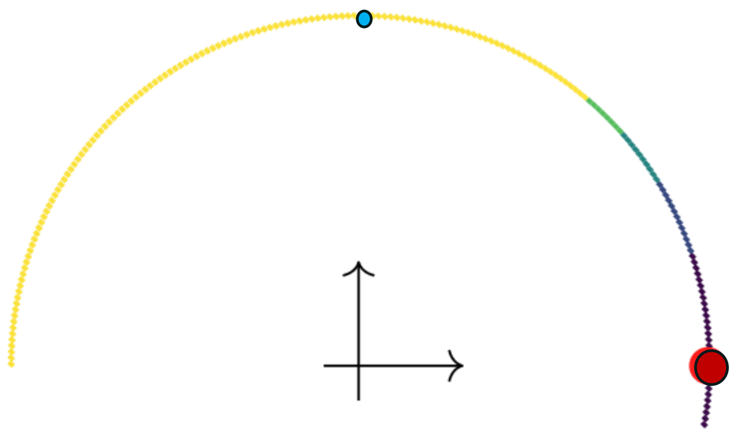
Memory savings

vs

Training stability

Oak Ridge National Laboratory

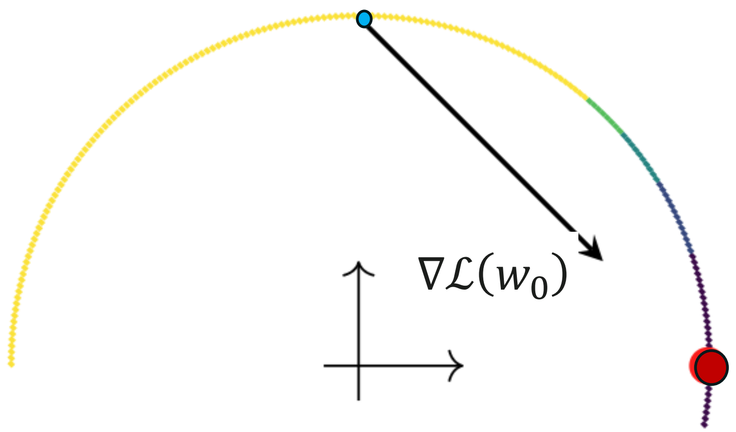
Thought Experiment: Manifold Constrained Optimization



$$\min_{w \in \mathcal{M}} \mathcal{L}(w) = \frac{1}{2} \|[1,0] - w\|_2^2$$

- Manifold: Unit circle $\mathcal{M} = \{w \in \mathbb{R}^2 : \|w\| = 1\}$
- Initialization at $w_0 = [0,1]$ solution at $w_* = [1,0]$

Thought experiment: Manifold Constrained Optimization

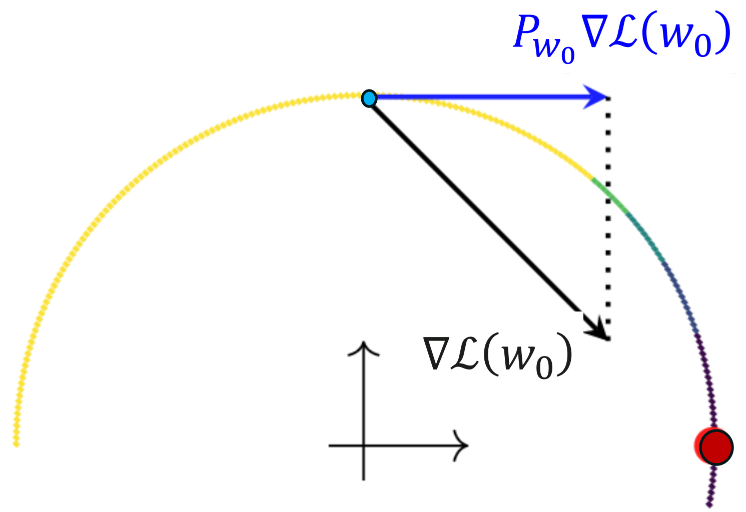


$$\min_{w \in \mathcal{M}} \mathcal{L}(w) = \frac{1}{2} \|[1,0] - w\|_2^2$$

- Manifold: Unit circle $\mathcal{M} = \{w \in \mathbb{R}^2 : \|w\| = 1\}$
- Initialization at $w_0 = [0,1]$ solution at $w_* = [1,0]$
- Gradient $\nabla\mathcal{L}(w_0) = [-1,1]$

OAK RIDGE National Laboratory

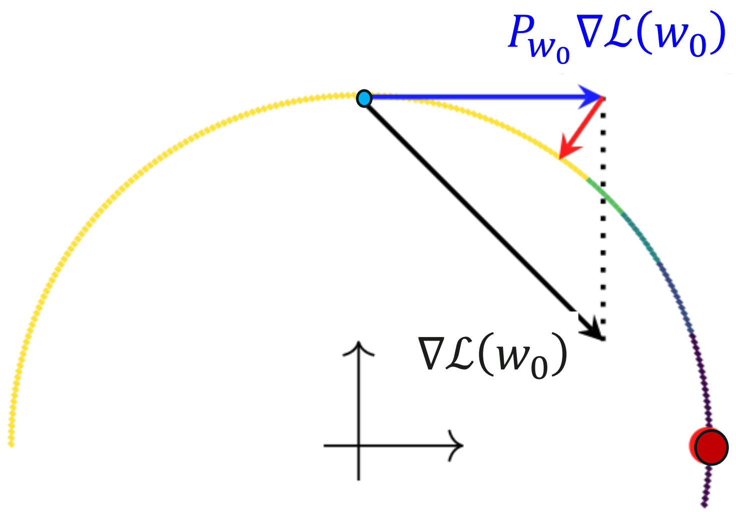
Thought experiment: Manifold Constrained Optimization



$$\min_{w \in \mathcal{M}} \mathcal{L}(w) = \frac{1}{2} \|[1,0] - w\|_2^2$$

- Manifold: Unit circle $\mathcal{M} = \{w \in \mathbb{R}^2 : \|w\| = 1\}$
- Initialization at $w_0 = [0,1]$ solution at $w_* = [1,0]$
- Gradient $\nabla\mathcal{L}(w_0) = [-1,1]$
- Riemannian gradient $P_{w_0}\nabla\mathcal{L}(w_0) = [0,1]$
  ➔ orthogonal projection P_{w_0}

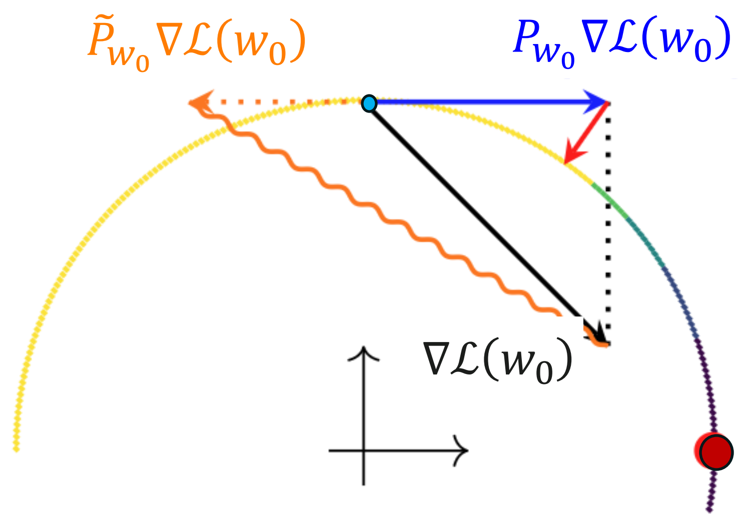
Thought experiment: Manifold Constrained Optimization



$$\min_{w \in \mathcal{M}} \mathcal{L}(w) = \frac{1}{2} \|[1,0] - w\|_2^2$$

- Manifold: Unit circle $\mathcal{M} = \{w \in \mathbb{R}^2 : \|w\| = 1\}$
- Initialization at $w_0 = [0,1]$ solution at $w_* = [1,0]$
- Gradient $\nabla\mathcal{L}(w_0) = [-1,1]$
- Riemannian gradient $P_{w_0}\nabla\mathcal{L}(w_0) = [0,1]$
  ➔ orthogonal projection P_{w_0}
- Retraction onto unit circle $\mathcal{M}$

Thought experiment: Manifold Constrained Optimization



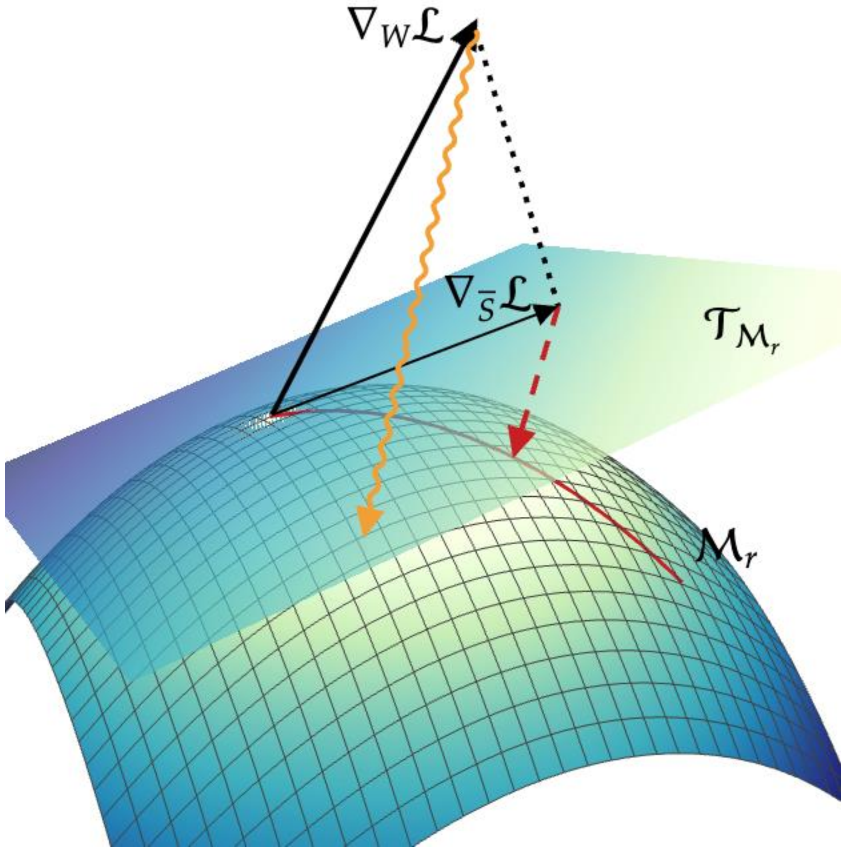
$$\min_{w\in\mathcal{M}} \mathcal{L}(w) = \frac{1}{2}\|[1,0] - w\|_2^2$$

- Manifold: Unit circle $\mathcal{M} = \{w \in \mathbb{R}^2 : \|w\| = 1\}$
- Initialization at $w_0 = [0,1]$ solution at $w_* = [1,0]$
- Gradient $\nabla\mathcal{L}(w_0) = [-1,1]$
- Riemannian gradient $P_{w_0}\nabla\mathcal{L}(w_0) = [0,1]$
  ➔ orthogonal projection P_{w_0}
- Retraction onto unit circle $\mathcal{M}$
- Non orthogonal $\tilde{P}_{w_0}$ breaks the structure

$\dot{w} = -\nabla\mathcal{L}$ vs $\dot{w} = -P_w\nabla\mathcal{L}$ vs $\dot{w} = -\tilde{P}_w\nabla\mathcal{L}$

gradient flow — Riemannian gradient flow

Low Rank training is manifold constrained optimization



$$\min_{W \in \mathcal{M}} \mathcal{L}(X, Y; W)$$

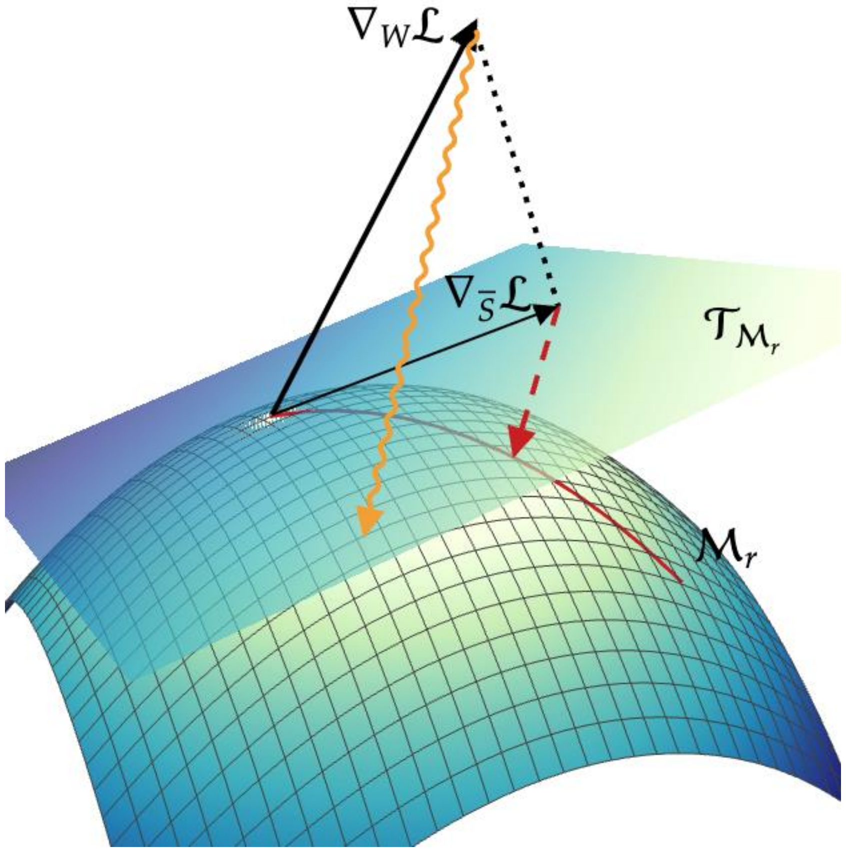
- Manifold $\mathcal{M} = \{W \in \mathbb{R}^{n \times n} : \text{rank}(W) = r\}$
- LoRA ansatz: $W = AB^\top$ with $A, B \in \mathbb{R}^{n \times r}$
- Gradient flow: $\dot{W} = \dot{A}B^\top + A\dot{B}^\top = \tilde{P}_W \nabla_W \mathcal{L}$ (chain rule)
- $\tilde{P}_W$ is defined by $[A, \dot{A}]$, and $[B, \dot{B}]$

➔ not orthogonal

➔ no steepest descent on $\mathcal{M}$

Oak Ridge National Laboratory

Low Rank training is manifold constrained optimization



$$\min_{W \in \mathcal{M}} \mathcal{L}(X, Y; W)$$

- Manifold $\mathcal{M} = \{W \in \mathbb{R}^{n\times n} : \text{rank}(W) = r\}$
- AdaLoRA ansatz: $W = USV^\top$ with $U, V \in \mathbb{R}^{n\times r}, S \in \mathbb{R}^{r\times r}$
- Gradient flow: $\dot{W} = \dot{U}SV^\top + U\dot{S}V^\top + US\dot{V}^\top = \widetilde{P}_W \nabla_W \mathcal{L}$ (chain rule)
- $\widetilde{P}_W$ is defined by $[U, \dot{U}]$, and $[V, \dot{V}]$

➔ not orthogonal

➔ no steepest descent on $\mathcal{M}$

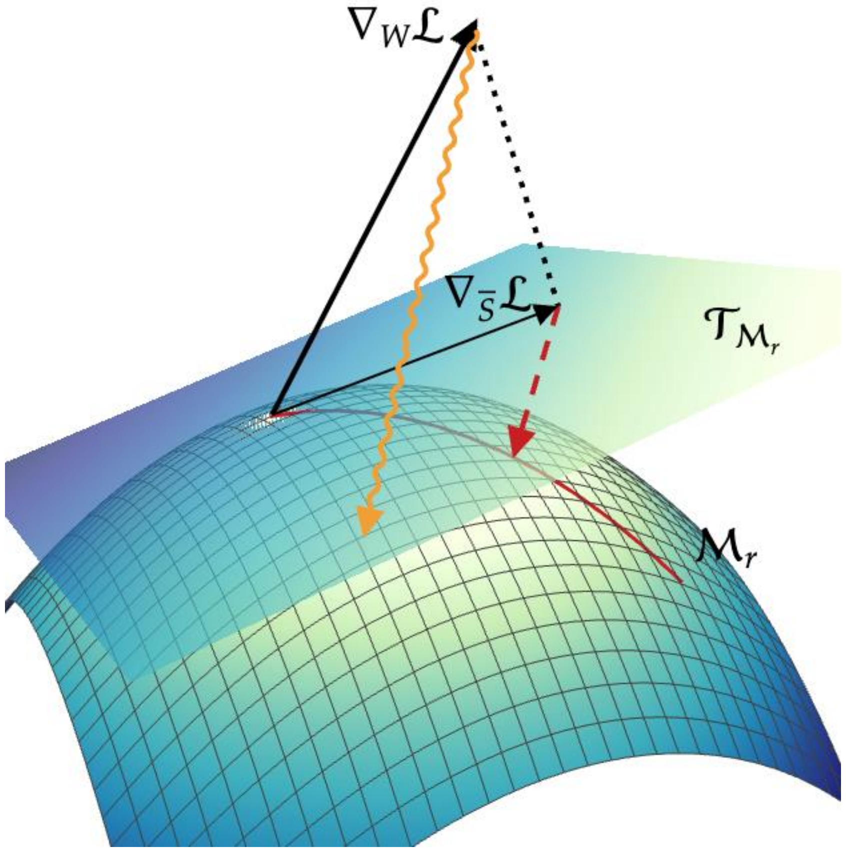
DLRT: Dynamical Low Rank Training

Efficient evolution of projected gradient flow

S.S., Zangrando, Kusch, Ceruti, Tudisco; *Low-Rank Lottery Tickets ...* ; NeurIPS 2022

$$\dot{W} = P_W \nabla \mathcal{L}$$

$$\min_{W \in \mathcal{M}} \mathcal{L}(X, Y; W)$$



- Manifold $\mathcal{M} = \{W \in \mathbb{R}^{n \times n} : \mathrm{rank}(W) = r\}$
- DLRT ansatz: $W = USV^\top$ with $U, V \in \mathbb{R}^{n \times r}, S \in \mathbb{R}^{r \times r}$
- Gradient flow: $\dot{W} = \dot{U}SV^\top + U\dot{S}V^\top + US\dot{V}^\top = P_W \nabla_W \mathcal{L}$
- Construct P_W with orthogonal bases
  $\bar{U} = \mathrm{ortho}\{[U, \dot{U}]\}$, and $\bar{V} = \mathrm{ortho}\{[V, \dot{V}]\}$

➔ Basis for tangent space $\mathcal{T}_{\mathcal{M}}$
➔ enables steepest descent on $\mathcal{M}$

Construct $\bar{U}$ / Construct $\bar{V}$ → Optimize $\bar{S}$ → Retract onto $\mathcal{M}$

Memory cost – slightly better than LoRA

- Weights: $\mathcal{O}(2nr + r^2)$
- Gradients: $\mathcal{O}(2nr)$ for basis update $\mathcal{O}(r^2)$ for optimization
- Optimizer states: $\mathcal{O}(r^2)$

Extendable to tensors

Zangrando , Schotthöfer, Ceruti, Kusch, Tudisco; *Geometry-aware training [...] in tensor Tucker format* ; NeurIPS '24

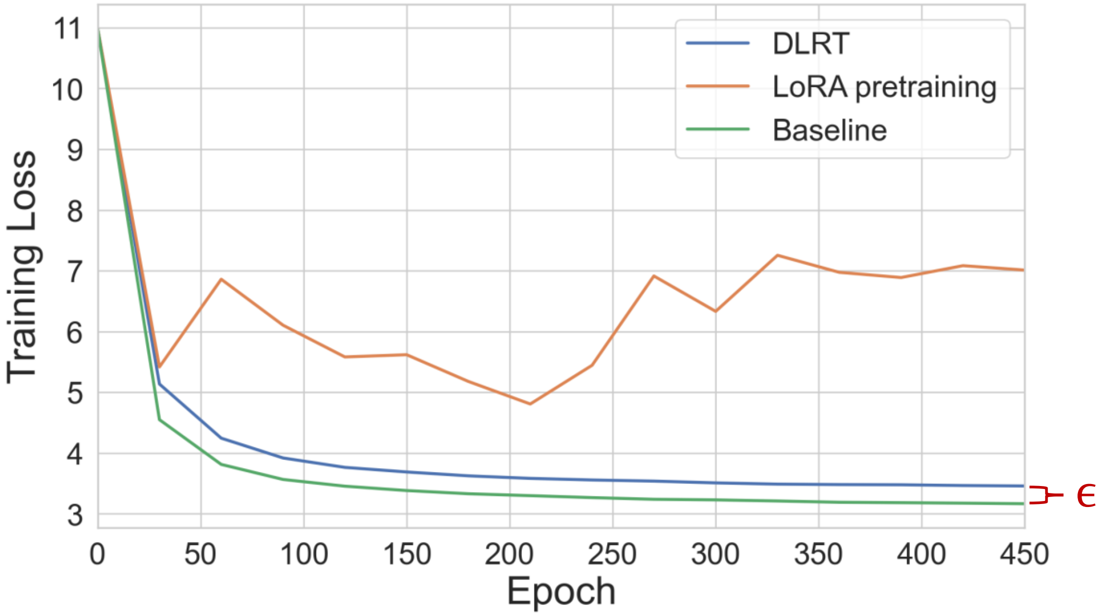
Upgrade to single step scheme

Schotthöfer, Zangrando Ceruti, Tudisco, Kusch; *GeoLoRA [...]*; ICLR '25

OAK RIDGE National Laboratory

DLRT: Dynamical Low Rank Training

S.S., Zangrando, Kusch, Ceruti, Tudisco; *Low-Rank Lottery Tickets ...* ; NeurIPS 2022
Zangrando , S.S., Ceruti, Kusch, Tudisco; *Geometry-aware training [...] in tensor Tucker format* ; NeurIPS '24
S.S., Zangrando Ceruti, Kusch, Tudisco; *GeoLoRA [...]*; ICLR '25



GPT-2 on OWT, low-rank MLP: DLRT beats LoRA

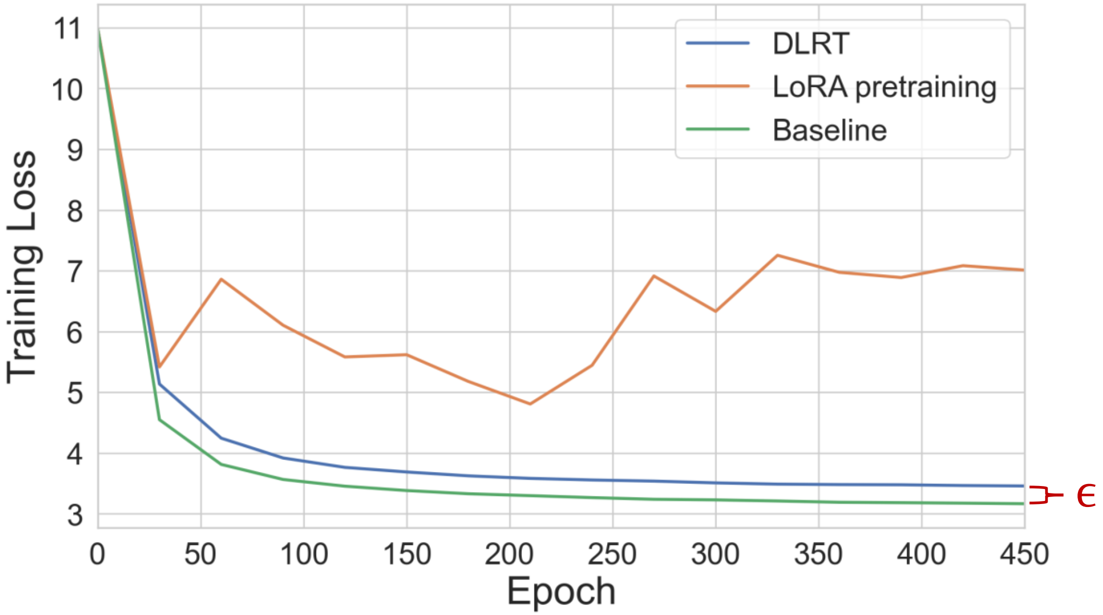
- DLRT inherits training robustness from full training
  - Provable: Optimality, loss descent, convergence
  - Hyperparameter can be transferred from full training
- Provable error bound to full rank training

$$\|W_{\text{full rank}}(t) - W_{\text{DLRT}}(t)\| < \epsilon(\lambda, \vartheta)$$

- Automatic rank selection (like AdaLoRA)

DLRT: Dynamical Low Rank Training

Schotthöfer, Zangrando, Kusch, Ceruti, Tudisco; *Low-Rank Lottery Tickets ...* ; NeurIPS 2022
Zangrando , Schotthöfer, Ceruti, Kusch, Tudisco; *Geometry-aware training [...] in tensor Tucker format* ; NeurIPS '24
Schotthöfer, Zangrando Ceruti,Tudisco, Kusch; *GeoLoRA [...]*; ICLR '25



GPT-2 on OWT, low-rank MLP: DLRT beats LoRA

- DLRT inherits training robustness from full training
  - Provable: Optimality, loss descent, convergence
  - Hyperparameter can be transferred from full training
- Provable error bound to full rank training

$$\|W_{\text{full rank}}(t) - W_{\text{DLRT}}(t)\| < \epsilon(\lambda, \vartheta)$$

- Automatic rank selection (like AdaLoRA)

What about momentum methods/Adam?

Schotthöfer, Klein, Kusch; *A geometric framework for momentum-based optimizers for low-rank training*; NeurIPS '25

Momentum gradient flow

$$\dot{W} = \mathcal{V}$$
$$\dot{\mathcal{V}} + \gamma\mathcal{V} = -\nabla_W \mathcal{L}$$

DLRT Momentum gradient flow

$$\dot{W} = P_W \mathcal{V}$$
$$\dot{\mathcal{V}} + \gamma\mathcal{V} = -P_W \nabla_W \mathcal{L}$$

➔ Bases U, V for W can be re-used for momentum terms
- Extendable for Adam, AdamW

Visit us at
- Our poster: Fri 5 Dec 4:30 p.m. PST — 7:30 p.m. PST @ Hall C,D,E
- Workshop Negel **Oral**: Sun 7 Dec 4:00 p.m. PST — 5 p.m. PST @ Upper Level Room 8

Using this method for adversarially robust compression
- Poster: Thu 4 Dec 11 a.m. PST — 2 p.m. PST @ Hall C,D,E
- **Oral**: Thu 4 Dec 10:20 a.m. — 10:40 a.m. PST @ Oral Session C
- Workshop COML: Sun 7 Dec 8 a.m. PST — 5 p.m. PST@ Upper Level Ballroom 6DE

OAK RIDGE National Laboratory