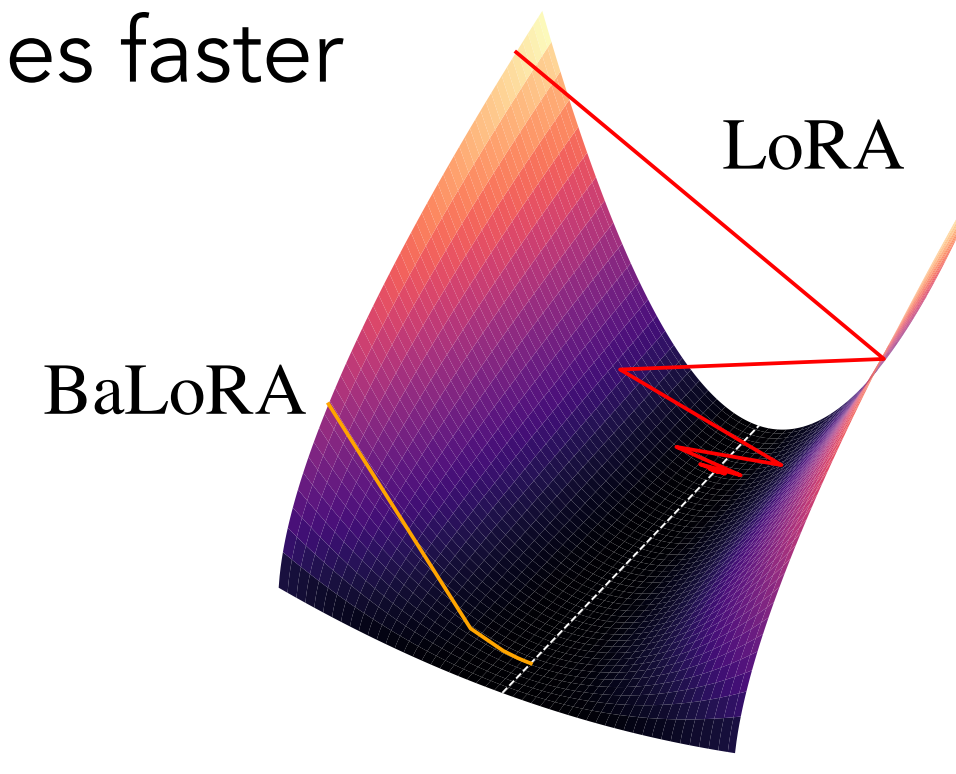


## BALORA IN A NUTSHELL

We propose a theoretically-grounded variant of LoRA that converges faster

- The LoRA loss  $f(A, B)$  has a continuous manifold of optimizers
- **Balanced** minimizers  $A^T A = B B^T$  are **optimally conditioned**
- When training LoRA, **projecting** the iterates to a **balanced manifold accelerates convergence**



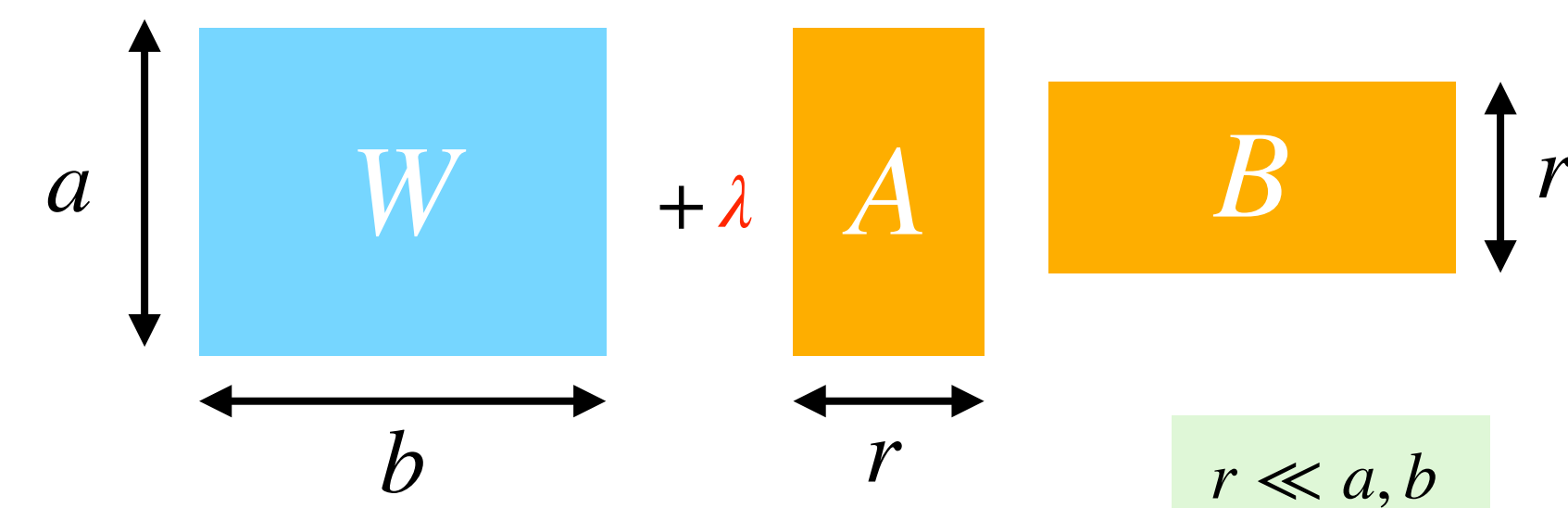
## LOW-RANK ADAPTATION (LoRA) [Hu et al., 2022]

Parameter-efficient method to adapt a pretrained model to a new dataset

❄️ freeze pretrained weights  $W$

🔥 train low-rank update to  $W$  as

$$\underbrace{W}_{\text{frozen}} + \lambda \underbrace{AB}_{\text{trained low rank}}$$



Training: optimize  $\min_{A_1, B_1, \dots, A_L, B_L} \ell(W_1 + A_1 B_1, \dots, W_L + A_L B_L)$  with Adam

Loss with one adapter:  $f(A, B) = \ell(W + AB)$

- ✓ reduced memory cost when training
- ✓ allows to ship easily several fine-tuned models

## CONDITIONING OF MINIMIZERS and ASYMPTOTIC CONVERGENCE RATE

LoRA is overparameterized:  $f(AR, R^{-1}B) = f(A, B)$  for  $R \in GL(r)$

$\rightarrow r^2$ -dimensional manifold of minimizers

Define the conditioning of  $(A, B)$  as  $\kappa = L/\mu$  with  $\begin{cases} L := \lambda_{\max}(\nabla^2 f(A, B)) \\ \mu := \lambda_{\min \neq 0}(\nabla^2 f(A, B)) \end{cases}$



✓ well-conditioned



✗ ill-conditioned

$\kappa$  controls LoRA's **asymptotic convergence rate** for GD and sign-GD (Adam without momentum)

$$\theta_t := (A_t, B_t) \quad \limsup_{t \rightarrow \infty} \frac{f(\theta_{t+1}) - f(\theta^*)}{f(\theta_t) - f(\theta^*)} \leq \begin{cases} \left(\frac{\kappa-1}{\kappa+1}\right)^2 & \text{if GD } \theta_{t+1} = \theta_t - \gamma \nabla f(\theta_t) \text{ with } \gamma = 2/(L + \mu) \\ 1 - \frac{1}{r(a+b)\kappa} & \text{if sign-GD } \theta_{t+1} = \theta_t - \gamma \|\nabla f(\theta_t)\|_1 s_t \text{ with } \gamma \text{ well-chosen} \end{cases}$$

**Q1:** Are some minimizers better conditioned than others?

**Q2:** How to steer the training dynamics towards an optimally-conditioned minimizer?

## THEORETICAL ANALYSIS of the CONDITIONING

**Model 1:**  $f(A, B) = \frac{1}{2} \|Z - AB\|_F^2$  (1-layer linear NN)

**Proposition:** for any minimizer  $(A, B)$ ,

$$\frac{\sigma_1(A)^2 + \sigma_1(B)^2}{\min(\sigma_r(A)^2, \sigma_r(B)^2)} \leq \kappa(\nabla^2 f(A, B)) \leq \frac{\sigma_1(A)^2 + \sigma_1(B)^2}{\min(\sigma_r(A)^2, \sigma_r(B)^2) - \sigma_{r+1}(Z)}$$

**Model 2:**  $f(A, B) = \frac{1}{2} \|h(AB) - Z\|_F^2$  ( $h$  deep non-linear model)

**Proposition:** if  $h(AB) = Z$  (interpolating regime), then for any minimizer  $(A, B)$ ,

$$\kappa(\nabla^2 f(A, B)) \leq \underbrace{\kappa(J_h(AB)^T J_h(AB))^{1/2}}_{\text{additional factor}} \underbrace{\frac{\sigma_1(A)^2 + \sigma_1(B)^2}{\min(\sigma_r(A)^2, \sigma_r(B)^2)}}_{\text{bound for linear case}}$$

**Theorem:** the loss is optimally conditioned if  $(A, B)$  is **balanced**, i.e.  $A^T A = B B^T$ .

## BALANCED LOW-RANK ADAPTATION (BaLoRA)

After each optimizer step, enforce balancing of  $(A, B)$  while preserving  $f(A, B)$

$$\begin{cases} \tilde{A}_{t+1} = A_t - \gamma \nabla_A f(A_t, B_t) \\ \tilde{B}_{t+1} = B_t - \gamma \nabla_B f(A_t, B_t) \\ (A_{t+1}, B_{t+1}) = P(\tilde{A}_{t+1}, \tilde{B}_{t+1}) \end{cases}$$

balanced unbalanced

« Projection » step:

- compute polar decomposition  $A = R_A S_A$ ,  $B = S_B R_B$
- compute SVD  $S_A S_B = U S V^T$
- apply  $P(A, B) = (R_A U S^{1/2}, S^{1/2} V^T R_B)$

$\downarrow$  sends  $(A, B)$  to the hyperbalanced manifold

$$\mathcal{H} = \{(US^{1/2}, S^{1/2}V^T) \mid U^T U = V^T V = I, \text{ and } S > 0 \text{ diagonal}\}$$

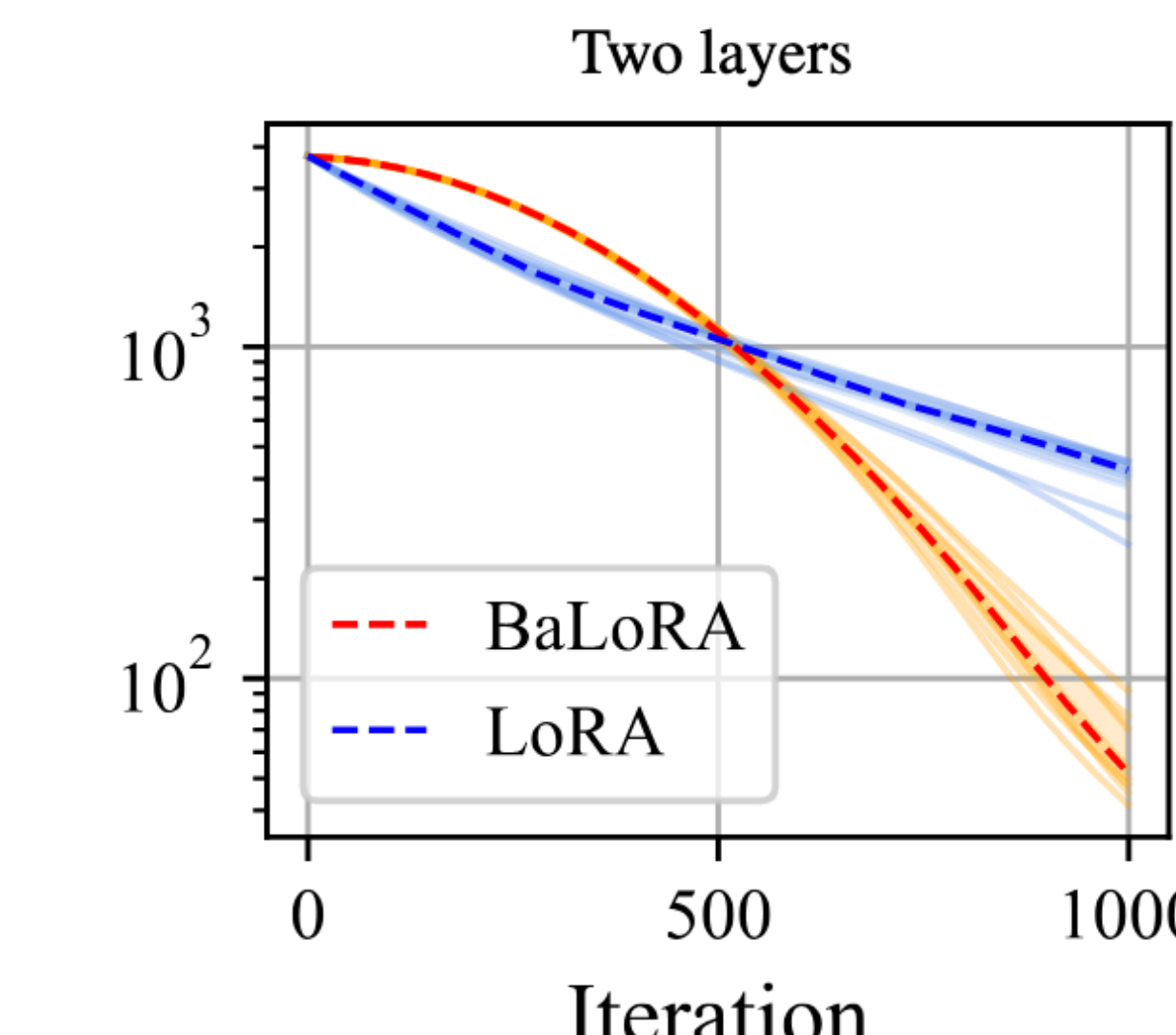
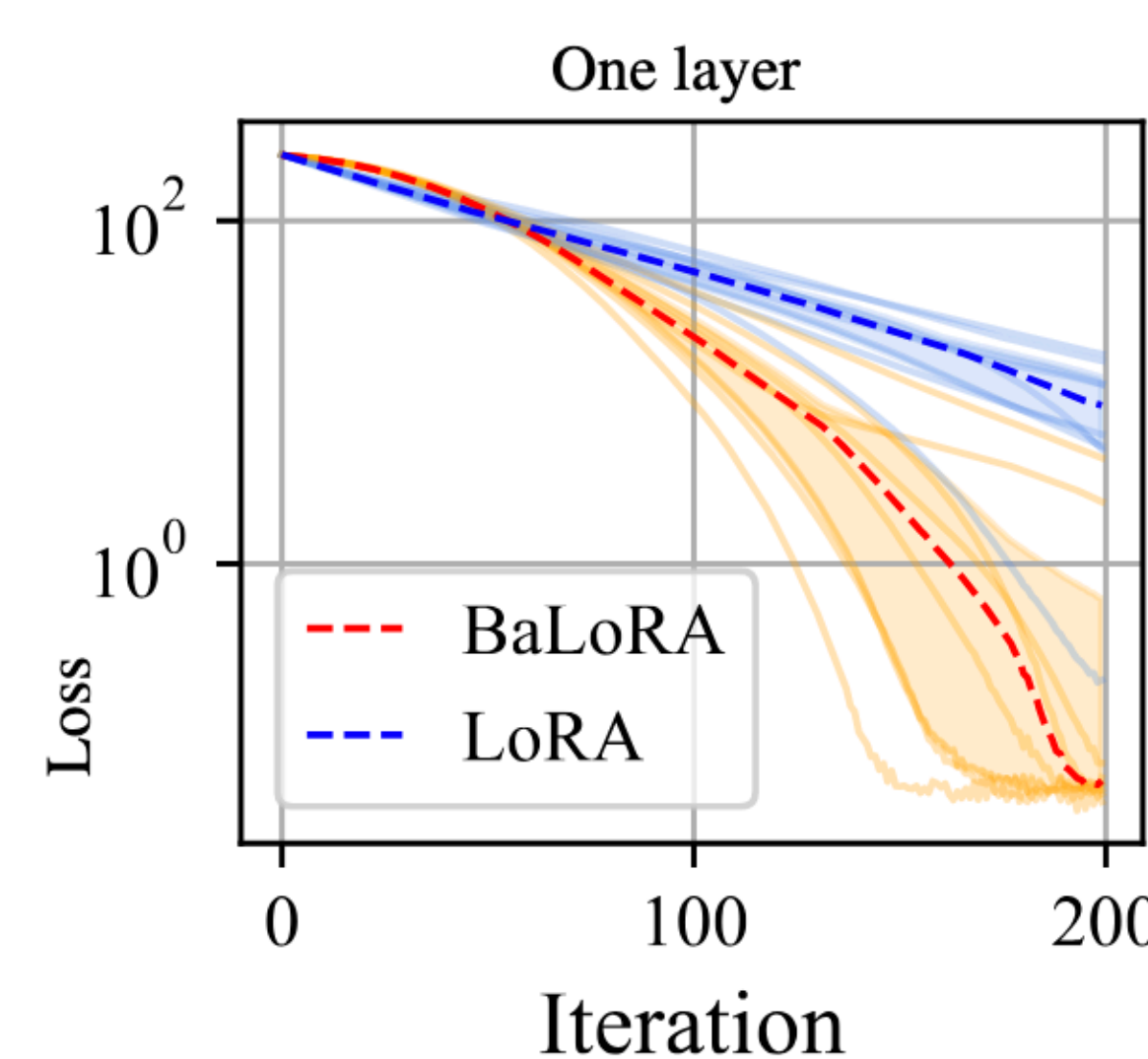
$\rightarrow$  optimizing in  $\mathcal{H}$  essentially removes LoRA invariances

- ✓ negligible computational overhead
- ✓ retraction-like properties
- ✓ can be written as an intrinsic GD on the manifold of rank- $r$  matrices

## EXPERIMENTS WITH SYNTHETIC DATA

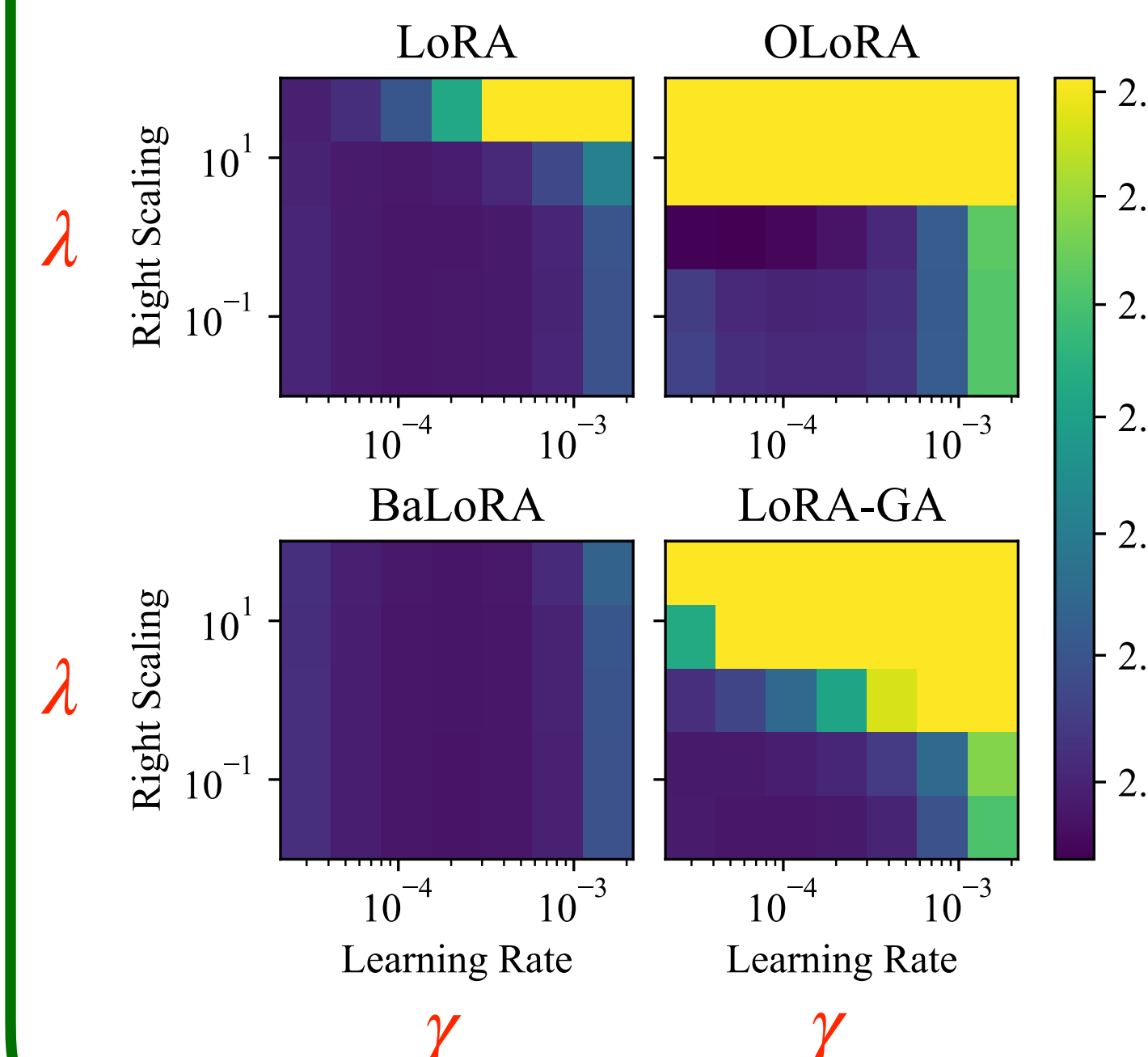
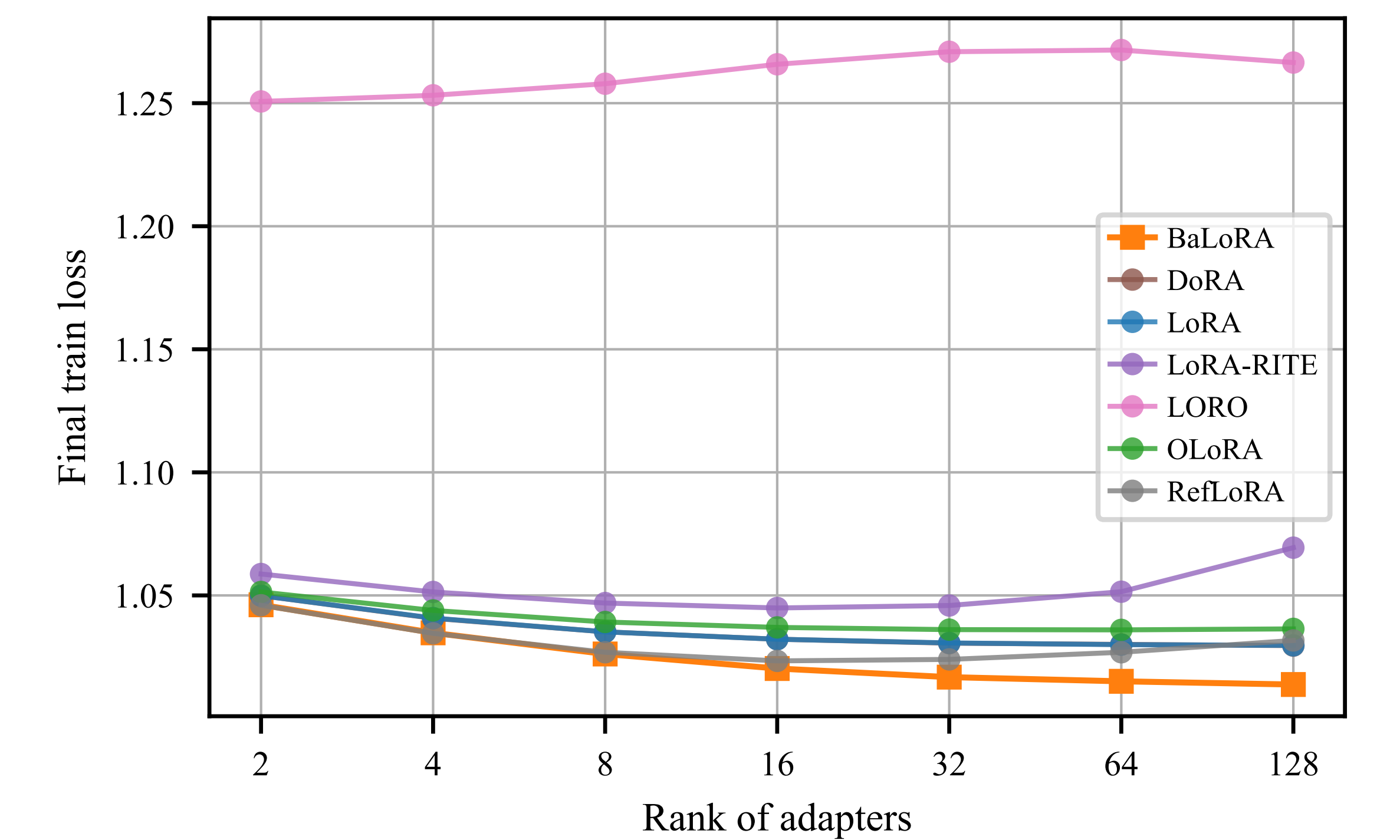
$$f(A, B) = \frac{1}{2} \|Z - AB\|_F^2$$

$$f((A_1, B_1), (A_2, B_2)) = \frac{1}{2} \|W^* - (W_{\text{frozen},1} + A_1 B_1)(W_{\text{frozen},2} + A_2 B_2)\|_F^2$$



## EXPERIMENTS WITH LLMs

Fine-tuning Qwen-3.2-3B on DeepMind-Mathematics:



- Balanced methods outperform the other baselines
- BaLoRA has an edge for larger ranks
- BaLoRA works with a wide range of hyperparameters