On the Crucial Role of Initialization for Matrix Factorization and LoRA

Bingcong Li[1]

Liang Zhang[1], Aryan Mokhtari[2], Niao He[1]

[1]ETH Zurich, [2]University of Texas at Austin

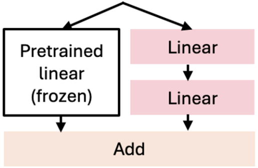
ICLR 2025

Singapore

A recap of low-rank adapters (LoRA)

❑ LoRA[1] recap

- The default choice for PEFT of LLMs
- Trainable parameters per layer $\mathcal{O}(mn) \rightarrow \mathcal{O}((m+n)r)$
- Great savings in memory and training time
- Scalable to serve on various downstream tasks



$$\boxed{\text{Optimization:} \quad \min_{\{\mathbf{X}_l, \mathbf{Y}_l\}} f(\{\mathbf{W}_l + \mathbf{X}_l \mathbf{Y}_l^\top\}_l)}$$

❑ A precursor to LoRA in the pre-LLM era: Burer-Monteiro (BM) factorization[2]

- Large-scale semidefinite programmings (SDPs)
- Matrix factorization, sensing, and completion problems

❑ **Goal:** To reveal the underlying optimization dynamics of LoRA, identify its limitations, and propose solutions

---

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang W. Chen. LoRA: Low-rank adaptation of large language models.ICLR 2022

S. Burer, R.D. Monteiro. A nonlinear programming alg. for solving SDPs via low-rank factorization. MP 2003

Matrix factorization as a proxy for LoRA optimization

❑ Matrix factorization for understanding LoRA dynamics

$$\textbf{Symmetric:} \quad \min_{\mathbf{X}\in\mathbb{R}^{m\times r}} \frac{1}{4}\|\mathbf{X}\mathbf{X}^\top - \mathbf{A}\|_F^2$$

$$\textbf{Asymmetric:} \quad \min_{\mathbf{X}\in\mathbb{R}^{m\times r},\mathbf{Y}\in\mathbb{R}^{n\times r}} \frac{1}{2}\|\mathbf{X}\mathbf{Y}^\top - \mathbf{A}\|_F^2$$

- Whitened data: $\mathbf{A}$ is the cross-covariance between original features and labels[1]
- A "reversed" view: given one-hot features, generate features
- Exact-, over-, under- parametrization (EP, OP, UP) when $\text{rank}(\mathbf{A}) =, <, > r$
- While solvable, solving them is not the central focus

❑ (Minor) challenges for quartic optimization

- Commonly adopted assumptions are not satisfied (globally)

$$\text{Example: } f(x, y) = (xy - 1)^2$$

- Non-convexity
- Non-smoothness: Hessian at $x = 0$ is $[y^2, -1; -1, 0]$
- Polyak-Lojasiewicz (PL) condition cannot be met due to existence of saddles

S. Arora, N. Cohen, N. Golowich, and W. Hu. A convergence analysis of gradient descent for deep linear neural networks. ICLR 2019

Matrix factorization/LoRA optimized via ScaledGD

- We will only talk about **symmetric problem** for simplicity
  - Our results extend to asymmetric problems

$$\min_{\mathbf{X}\in\mathbb{R}^{m\times r}} \frac{1}{4}\|\mathbf{X}\mathbf{X}^\top - \mathbf{A}\|_F^2$$

- Recap of ScaledGD[1]

$$\mathbf{X}_{t+1} = \mathbf{X}_t - \eta \underbrace{(\mathbf{X}_t\mathbf{X}_t^\top - \mathbf{A})\mathbf{X}_t}_{\text{gradient}} \cdot \underbrace{(\mathbf{X}_t^\top\mathbf{X}_t)^{-1}}_{\textit{preconditioner}}$$

  - EP: linear convergence under small initialization $[\mathbf{X}_0]_{ij} \sim \mathcal{N}(0, \zeta^2)$ (for small ζ^2)
  - OP: (modified version) linear convergence when using small initialization
  - UP: convergence is unclear yet

- **Our contributions**
  - Nyström initialization enables ScaledGD to achieve **quadratic** rates for EP and OP, and guarantees **linear** convergence for UP in (a)symmetric problems
  - Initialization can exponentially impact convergence!
  - Nyström initialization also helps for finetuning LLMs with LoRA

[1] T. Tong, C. Ma, and Y. Chi, Accelerating ill-conditioned low-rank matrix estimation via ScaledGD. *JMLR 2021*

Nyström initialization enables quadratic convergence in EP

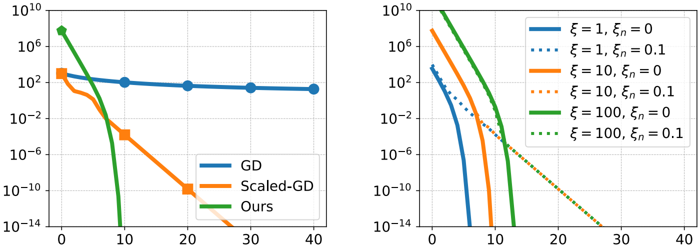
❑ Nyström initialization

- Nyström initialization enables $\mathbf{X}_0$ and $\mathbf{A}$ share the same column space, i.e., it eliminates the "noise" space

$$\mathbf{X}_0 = \mathbf{A}\boldsymbol{\Omega}, \quad \text{where } [\boldsymbol{\Omega}]_{ij} \sim \mathcal{N}(0, \xi^2), \forall i, \forall j$$

❑ Noise space converges exponentially slower

- Slightly perturbing Nyström initialization, the quadratic rates vanish!



Comparison with existing work

setting		alg.	ref.	init.	rate
Asymmetric	EP	GD	(Ye & Du 2021)	small	$\mathcal{O}\left(\kappa^4 + \kappa^4 \log(1/\epsilon)\right)$
		AltGD	(Ward & Kolda 2023)	special	$\mathcal{O}\left(\kappa^2 \log(1/\epsilon)\right)$
		ScaledGD	(Cong et al., 2023)	local	$\mathcal{O}(\log(1/\epsilon))$
		ScaledGD	**Theorem 3**	Nyström	$\mathcal{O}(1)$
	OP	AltGD	(Ward & Kolda 2023)	special	$\mathcal{O}\left(\kappa^2 \log(1/\epsilon)\right)$
		ScaledGD	**Theorem 6**	Nyström	$\mathcal{O}(1)$
	UP	GD	(Du et al., 2018)	small	asymptotic
		ScaledGD	**Theorem 4**	Nyström	$\mathcal{O}(1)$
Symmetric	EP	GD*	(Stöger et al. 2021)	small	$\mathcal{O}\left(\kappa^8 + \kappa^2 \log(1/\epsilon)\right)$
		ScaledGD	**Theorem 1**	Nyström	$\mathcal{O}\left(\kappa^3 + \log\log(1/\epsilon)\right)$
	OP	GD*	(Stöger et al. 2021)	small	$\mathcal{O}\left(\kappa^8 + \kappa^6 \log(\kappa/\epsilon)\right)$
		ScaledGD-λ*	(Xu et al. 2023)	small	$\mathcal{O}\left(\log^2 \kappa + \log(1/\epsilon)\right)$
		ScaledGD	**Theorem 5**	Nyström	$\mathcal{O}\left(\kappa^3 + \log\log(1/\epsilon)\right)$
	UP	ScaledGD	**Theorem 2**	Nyström	$\mathcal{O}(1/\epsilon \cdot \log(1/\epsilon))$

- 1-step convergence (asymmetric) uses very particular structure of ScaledGD
- A quadratic rate also holds for asymmetric problems

Implications to finetuning LLMs

❑ Intuition for fintuning LLMs with LoRA

- Inject column/row space at initialization and get rid of noise space as much as possible

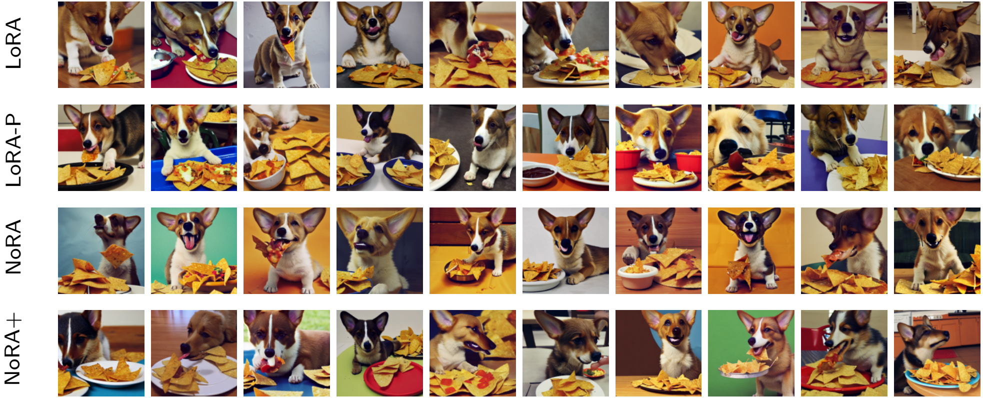
❑ What are the good directions for initializing LoRA?

- A set of well-performed adapters share the same col. space with pretrained weights[1]
- We can apply Nyström initialization directly!
- NoRA: Nyström initialized LoRA

V. Lingam, et al. SVFT: Parameter-Efficient Fine-Tuning with Singular Vectors.*NeurIPS 2024*.

NoRA on StableDiffusion

- Subject-driven image generation (finetuning diffusion models with my own pics)
  - Goal: "A dog eating nachos"
  - StableDiffusion-v1.4 (0.98B params), LoRA ($r = 4$) applied on the U-Net (0.8M trainable params)
  - Initialization improves the quality (NoRA vs. LoRA, and NoRA+ vs. LoRA-P)



NoRA on LLaMA

- Common-sense reasoning with LLaMA-7B and LLaMA2-7B
  - Mix all the datasets for training; evaluate separately
  - LoRA with $r = 32$, leading to 56M trainable params
  - Nyström initialization is beneficial for tasks beyond pattern recognition, where commonsense and knowledge is needed for proper inferences

	Alg.	BoolQ	PIQA	SIQA	HS	WG	ARC-e	ARC-c	OBQA	avg (↑)
LLaMA	LoRA	66.42	80.03	77.84	82.88	81.85	79.92	63.40	77.20	76.19
	LoRA-P	68.96	80.95	77.43	81.54	80.27	78.83	64.16	79.20	76.41
	NoRA	68.20	80.79	78.40	85.09	80.27	79.17	62.80	78.80	**76.69**
	NoRA+	69.85	81.83	77.38	82.09	80.03	79.67	64.25	78.60	**76.71**
LLaMA2	LoRA‡	69.8	79.9	79.5	83.6	82.6	79.8	64.7	81.0	77.6
	LoRA-P	71.47	81.50	78.81	85.97	80.43	81.14	66.55	81.00	78.35
	NoRA	71.16	83.08	79.53	85.90	81.85	80.64	66.13	81.80	78.76
	NoRA+	70.52	81.94	79.07	87.66	82.24	82.70	67.06	80.20	**78.92**