COALA: Numerically Stable and Efficient Framework for Context-Aware Low-Rank Approximation

Uliana Parkina, Maxim Rakhuba

HSE University

NeurIPS 2025

Introduction

To compress a weight matrix W of an LLM, we consider the **W**eighted **L**ow-rank **A**pproximation (WLA) problem:

$$\min_{\operatorname{rank}(W') \leq r} \|(W - W')X\|_F, \tag{1}$$

where $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ with a low-rank W', $X \in \mathbb{R}^{d_{\text{in}} \times M}$ – calibration data matrix.

Our contribution:

- We propose inversion-free formulas for improved numerical stability.
- We add regularization to boost approximation quality in all settings.
- We develop a QR-based solution that is efficient and GPU-parallelizable.

WLA does not require matrix inversion!

Example of conventional route to solve problem[1, 2, 3]

$$W' = \mathrm{SVD}_r(WS)\, S^{-1}, \qquad \text{where } S \text{ is any matrix that } SS^\top = XX^\top.$$

- Only works for X of full column rank.
- Evaluating S^{-1} may lead to a precision loss for nearly singular S.

[1] Xin Wang et al. "SVD-LLM: Truncation-aware Singular Value Decomposition for Large Language Model Compression". In: *The Thirteenth International Conference on Learning Representations*. 2025.

[2] Zhiteng Li et al. "AdaSVD: Adaptive Singular Value Decomposition for Large Language Models". In: *CoRR* abs/2502.01403 (Feb. 2025). URL: `https://doi.org/10.48550/arXiv.2502.01403`.

[3] Patrick Chen et al. "Drone: Data-aware low-rank compression for large nlp models". In: *Advances in neural information processing systems* 34 (2021), pp. 29321–29334.

WLA does not require matrix inversion!

Our route to solve problem

$$W' = U_r U_r^\top W, \quad \text{where} \quad U_r \Sigma_r V_r^\top = \mathrm{SVD}_r(WX).$$

In exact arithmetic, these are identical – even if you can't see it!

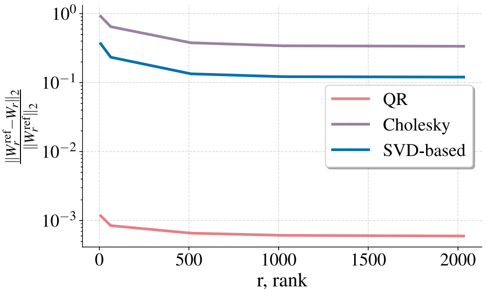

Figure: Relative error vs. rank for various methods.

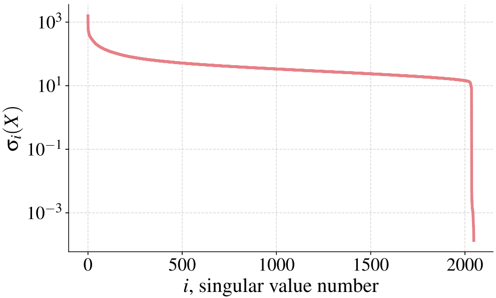

Figure: Behavior of singular values $\sigma_i(X)$.

What if the calibration matrix X is huge?

Forming WX and evaluating SVD explicitly becomes expensive.
Solution:

$$\|(W - W')X\|_F = \|(W - W')R^\top\|_F, \quad \text{where} \quad QR = X^T,$$

To handle massive X, we employ a block-wise QR scheme, known as *TSQR*[4].

$$\begin{array}{ccccccc}
X_0 \rightarrow R_0 & \searrow & & & & & \\
 & & R_{01} & \searrow & & & \\
X_1 \rightarrow R_1 & \nearrow & & & & & \\
 & & & & R_{0123} & & \\
X_2 \rightarrow R_2 & \searrow & & & & & \\
 & & R_{23} & \nearrow & & & \\
X_3 \rightarrow R_3 & \nearrow & & & & &
\end{array}$$

[4] James Demmel et al. "Communication-optimal parallel and sequential QR and LU factorizations". In: *SIAM Journal on Scientific Computing* 34.1 (2012), A206–A239.

Efficiency. Experiments

QR is still the quickest initializer for large-model compression, even with highly unbalanced matrices.

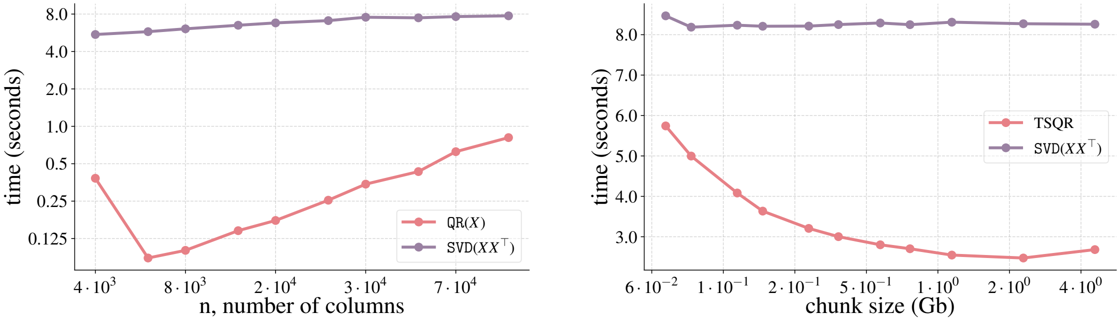

Figure: Runtimes for computing S: $SS^\top = XX^\top$ using two approaches. *Left:* Matrix $X \in \mathbb{R}^{4096\times n}$ for different n. *Right:* Matrix $X \in \mathbb{R}^{4096\times 3\cdot 10^5}$ split into chunks of different size. In this case, QR is computed using the TSQR method and the Gram matrix using $XX^\top = \sum_{i=1}^{p} X_i X_i^\top$.

Let's add regularization

$$\min_{\text{rank}(W') \leq r} \|WX - W'X\|_F^2 + \mu \|W - W'\|_F^2 \tag{2}$$

In practice, we also want to adapt the model to fit the available examples, but not excessively, as we aim to avoid overfitting and preserve the model's knowledge in other domains:

What is the limit of W_μ as $\mu \to 0$?

Let's add regularization

Theorem

Suppose that X has $\operatorname{rank}(X) = k \geq r$ and that $\sigma_r(WX) \neq \sigma_{r+1}(WX)$. Then, if the solution W_0 to the problem (1), and if W_μ denotes the solution to the regularized problem (2):

$$\|W_0 - W_\mu\|_F \leq \frac{2\|W\|_2^2\|W\|_F\left(\frac{\sigma_1(X)}{\sigma_k(X)} + \max\left(1,\ \frac{\mu}{4\,\sigma_k^2(X)}\right)\right)}{\sigma_r^2\,(WX) - \sigma_{r+1}^2\,(WX)} \cdot \mu.$$

In practice, we observe a this linear dependence on μ, and the proportionality constant correlates with the singular-value gap.

Experiments

Table: Metric values of various compression methods. Experiments were conducted using the *Mistral-7B* model on the WikiText2 dataset and commonsense reasoning used for validation.

Ratio	Method	MMLU	BoolQ	PIQA	WiNoG	HSweg	ARC-E	ARC-C	OBQA
0%	Mistral-7B	62.50	83.98	82.05	73.95	81.02	79.55	53.92	44.00
80%	FLAP	25.90	<u>62.26</u>	72.31	<u>64.09</u>	<u>55.94</u>	51.05	31.91	36.80
	SliceGPT	28.60	37.86	60.66	59.43	45.10	48.15	30.03	32.00
	SVD-LLM	<u>41.80</u>	68.29	<u>73.39</u>	68.43	61.75	<u>71.34</u>	40.53	36.60
	SoLA	**44.20**	66.09	<u>73.67</u>	<u>68.75</u>	<u>63.32</u>	69.99	39.76	<u>39.20</u>
	COALA	41.20	**78.07**	**77.04**	**68.82**	**65.06**	**72.13**	**43.43**	**40.20**
70%	FLAP	26.40	<u>65.26</u>	<u>69.59</u>	64.80	<u>55.61</u>	<u>48.91</u>	30.55	<u>35.80</u>
	SliceGPT	25.00	37.83	54.41	51.62	32.54	35.02	22.95	26.80
	SVD-LLM	<u>28.20</u>	64.62	64.91	<u>64.17</u>	47.36	58.25	30.72	34.20
	SoLA	**33.80**	62.57	<u>68.39</u>	**64.48**	<u>53.00</u>	<u>60.90</u>	<u>32.76</u>	<u>37.60</u>
	COALA	27.35	**63.82**	**70.40**	62.43	**51.02**	**63.63**	**35.49**	**36.00**

More stuff!

- Sensitivity analysis with respect to μ
- More experiments
- We also show that an analogous problem arises in PEFT when initializing adapter layers

Learn this and more in our paper!