KurTail : Kurtosis-based LLM Quantization

Sadegh Akhondzadeh, Aleksandar Bojchevski, Evangelos Eleftheriou, Martino Dazzi
Email: akhondzadeh@cs.uni-koeln.de

TL;DR.

The Problem

Large Language Models exhibit extreme activation outliers making uniform quantization ineffective.

Our Solution

KurTail is a PTQ method for LLMs that learns an orthogonal rotation by minimizing Kurtosis to flatten the activation tails before quantizing.

Rotations are trained layer-wise from a few hundred samples and then fused, without any retraining or fine-tuning.

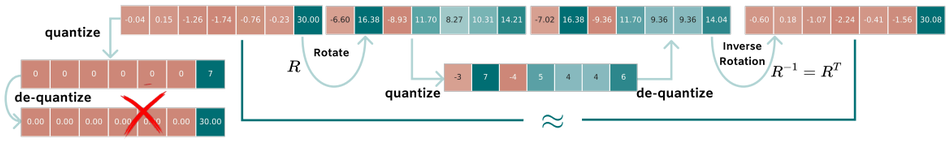
We quantize weights, activations & KV-cache all down to 4 bits.

Advantages

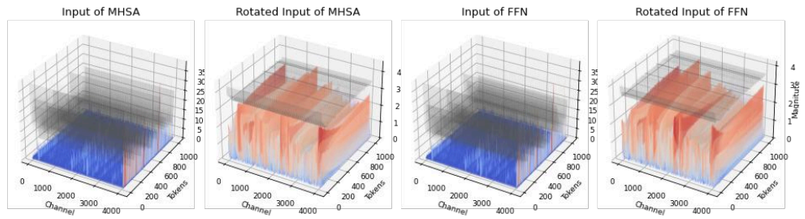
+13 % MMLU / −15 % PPL over QuaRot, and still tops SpinQuant.

Single H100 is enough to quantize Llama-3-70B, SpinQuant needs 4×.

Rotation (Channel-mixing) Improves Quantization Error



Rotation Flattens Distributions and Removes Outliers



Results

Method	Llama-2-7b			Llama-2-13b			Llama-3-8b		
	Wiki (↓)	0-shot (↑)	MMLU (↑)	Wiki (↓)	0-shot (↑)	MMLU (↑)	Wiki (↓)	0-shot (↑)	MMLU (↑)
16-bit	5.5	64.1	42.1	4.9	66.5	52.7	6.1	67.2	63.2
GPTQ	9600.0s	38.9	23.8	3120.0	33.8	24.8	166.3	39.8	23.3
QuaRot	6.2	60.6	32.3	5.4	64.7	46.83	8.50	60.1	47.4
SpinQuant	6.0	61.0	34.8	5.2	64.8	47.8	7.4	63.8	56.2
Kurtail	**5.9**	**61.3**	**32.9**	**5.2**	**65.2**	**49.1**	**7.2**	**64.6**	**57.3**
Method	**Llama-3-70b**			**Llama-3.2-1b**			**Llama-3.2-3b**		
	Wiki (↓)	**0-shot (↑)**	**MMLU (↑)**	**Wiki (↓)**	**0-shot (↑)**	**MMLU (↑)**	**Wiki (↓)**	**0-shot (↑)**	**MMLU (↑)**
16-bit	2.8	73.1	76.3	9.75	54.9	37.9	7.8	62.7	54.8
GPTQ	452.7	45.5	23.2	108.9	38.0	24.9	178.3	40.3	24.8
QuaRot	6.19	65.1	62.9	17.4	49.0	23.8	10.1	56.1	42.0
SpinQuant	6.2	65.7	59.4	13.6	48.8	25.6	9.2	57.9	44.2
Kurtail	**4.2**	**70.7**	**73.1**	**12.9**	**50.1**	**27.2**	**9.0**	**59.0**	**47.8**

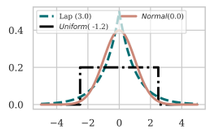
Learning the Rotation Improves Quantization

End-to-end optimization (SpinQuant) is too expensive.

We minimize Kurtosis as a useful proxy. This makes activations more uniform, ideal for uniform quantization.

Kurtosis measures the tailedness of a distribution.

$$\kappa = \frac{\mathbb{E}[(x-\mu)^4]}{(\mathbb{E}[(x-\mu)^2])^2} = \frac{\mu_4}{\sigma^4}$$



Higher Kurtosis → heavy tails → many outliers.

Implementing Layer-wise Optimization

We conduct optimization per-layer, significantly reducing memory and computational requirements.

Informal Theorem

Layer-wise optimization = End-to-end optimization.

Our loss pushes activations toward the uniform distribution by minimizing the gap to uniform Kurtosis.

$$\mathcal{L}_\kappa = \frac{1}{L}\sum_{i=1}^{L} \left|\kappa\left(\bigoplus_{j=1}^{N} a_{ij}\right) - \kappa_u\right|$$

layers; activation of token j in layer i; Kurtosis of activations; Kurtosis of the uniform distribution

Why a Rotation (Orthogonal Transformation)?

1- Weight Fusion → Efficiency

To enable weight fusion we need equivariance for RMSNorm. Orthogonal transformations ensure this.

$$\mathrm{RMSNorm}(XR) = \mathrm{RMSNorm}(X)R \text{ if and only if } R^T R = I$$

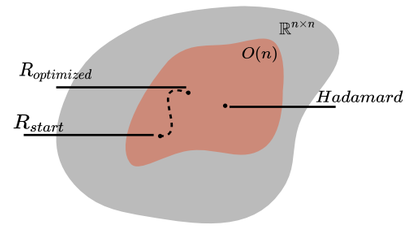
2- Arbitrary linear transformations can amplify errors!

Informal Theorem

Quantization error $\propto$ Condition number of transformation matrix.

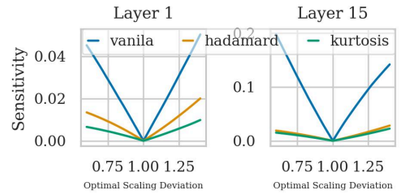
Orthogonal transform has lowest condition number and thus lowest error.

We use the Caley-Adam optimiser to efficiently navigates the space of orthogonal matrices.



Rotations Reduces Uniform Quantization Sensitivity

Earlier layers benefit more from learnable rotations.



Fusing the Rotations