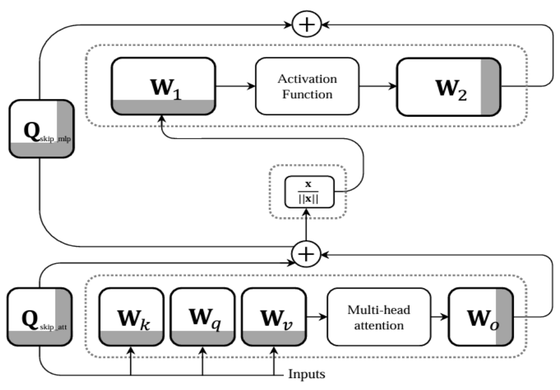
Getting free Bits Back from Rotational Symmetries in LLMs

Jiajun He, Gergely Flamich, José Miguel Hernández-Lobato
University of Cambridge

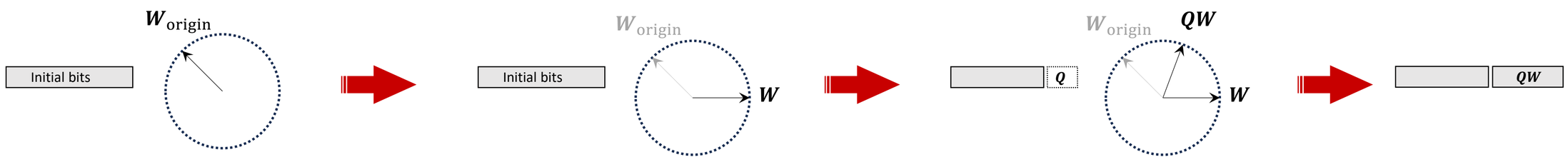
SliceGPT and rotational symmetry



👉 Transformer with SliceGPT has **rotational symmetries; directly saving weights cause redundancy**

Bits-back coding for rotational symmetry

Encoding



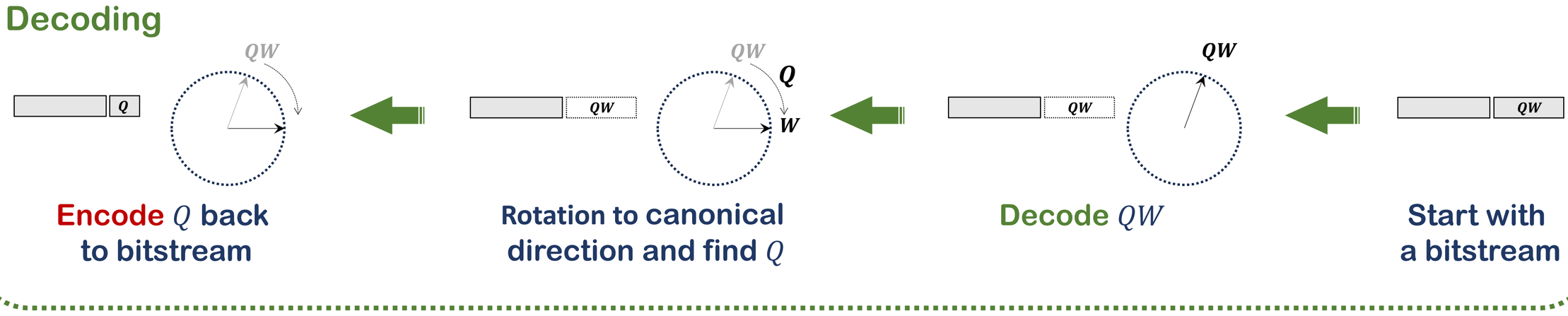
start with W_{origin} a nd initial bitstream

Rotation to canonical direction: WW^{T} is diagonal

Decode a rotation matrix Q and rotate W

Encode QW

Decoding



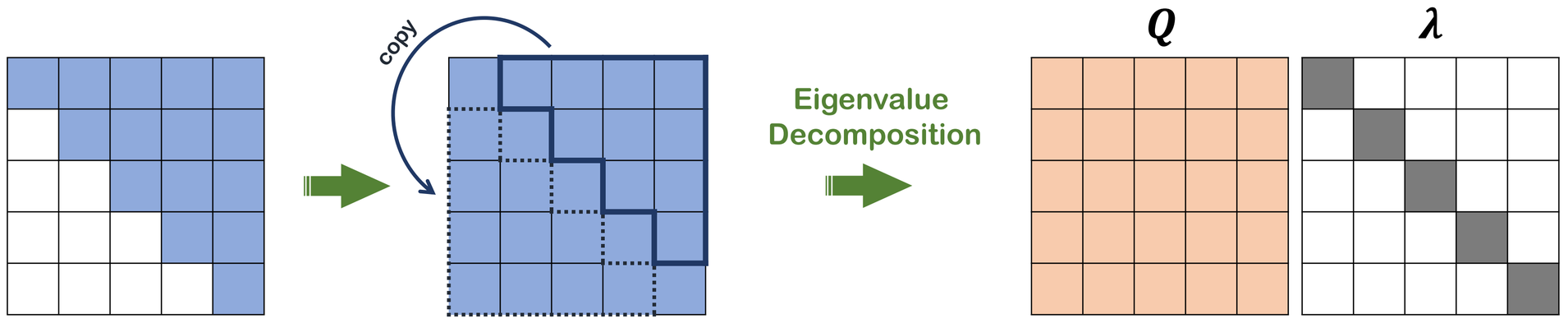
Encode Q back to bitstream

Rotation to canonical direction and find Q

Decode QW

Start with a bitstream

Decoding and encoding of a rotation matrix Q



Decode $D(D+1)/2$ float

Form a symmetric matrix

Return eigenvectors as the rotation matrix
Encode D eigenvalues to bitstream

$$Q \operatorname{diag}(\lambda) Q^{\mathrm{T}}$$

Encode the upper triangular and diagonal elements

Decode D eigenvalues (λ) from bitstream

Performance

Model	SliceGPT Slicing	Compress Rate after SliceGPT	Compress Rate after bits-back	Performance (before/after bits-back)			
				PPL (↓)	PIQA (%, ↑)	WinoGrande (%, ↑)	HellaSwag (%, ↑)
OPT-1.3B	20%	-9.53%	-13.77%	**16.59**/16.60	**64.91**/64.80	**54.78**/54.38	45.26/**45.32**
	25%	-14.84%	-18.61%	**17.78**/17.86	**63.55**/63.33	52.80/**53.28**	**43.20**/43.11
	30%	-20.53%	-23.81%	**19.60**/19.66	**60.88**/60.50	52.88/**53.28**	**40.25**/40.06
OPT-2.7B	20%	-9.19%	-13.84%	**13.89**/13.95	**68.44**/68.12	**58.88**/58.72	**51.35**/51.17
	25%	-15.07%	-19.09%	**14.85**/14.87	**66.70**/66.76	57.30/**57.70**	**48.41**/48.38
	30%	-20.88%	-24.43%	**16.31**/16.33	64.64/**64.69**	55.80/**56.04**	44.52/**44.57**
OPT-6.7B	20%	-9.29%	-14.07%	**11.63**/11.71	72.91/**73.01**	**61.33**/61.17	60.53/**60.55**
	25%	-15.16%	-19.29%	**12.12**/12.15	71.00/**71.22**	60.30/**60.77**	**57.76**/57.55
	30%	-21.18%	-24.84%	**12.81**/12.91	69.31/**69.42**	**59.75**/59.59	**53.64**/52.94
OPT-13B	20%	-9.18%	-14.01%	**10.75**/10.77	74.27/74.27	**64.96**/64.88	65.74/**65.79**
	25%	-15.27%	-19.51%	11.08/**11.07**	**74.27**/73.72	63.46/**63.93**	**63.48**/63.09
	30%	-21.29%	-24.97%	**11.55**/11.59	72.69/**73.01**	61.96/**62.43**	**60.12**/60.05
Llama-2-7B	20%	-9.38%	-14.13%	**6.86**/6.98	**69.53**/69.42	64.17/**64.72**	**58.96**/58.89
	25%	-15.34%	-19.53%	**7.56**/7.59	67.03/**67.57**	62.98/**63.38**	**54.29**/53.93
	30%	-21.45%	-25.09%	**8.63**/8.69	**64.69**/64.09	**62.75**/62.12	**49.13**/49.07

Conclusion:

👉 3-5% bits saving

👉 almost no influence on performance