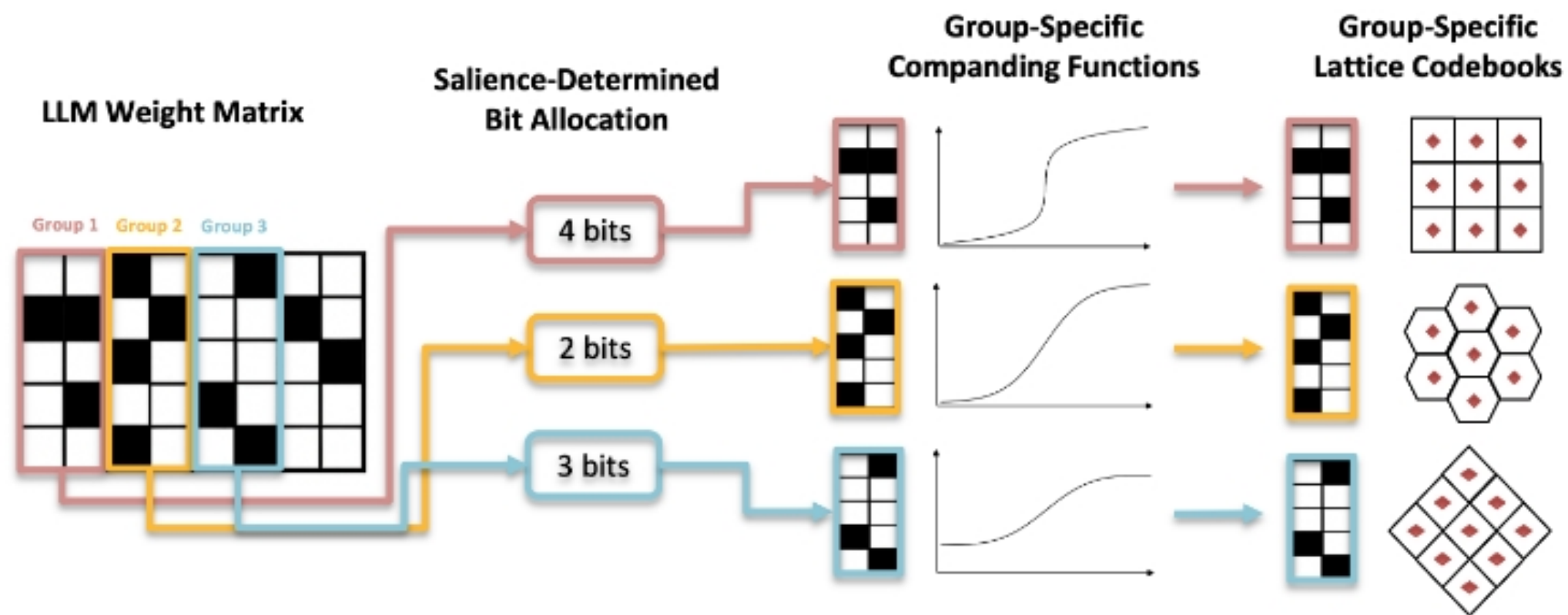


Learning Grouped Lattice Vector Quantizers for Low-Bit LLM Compression

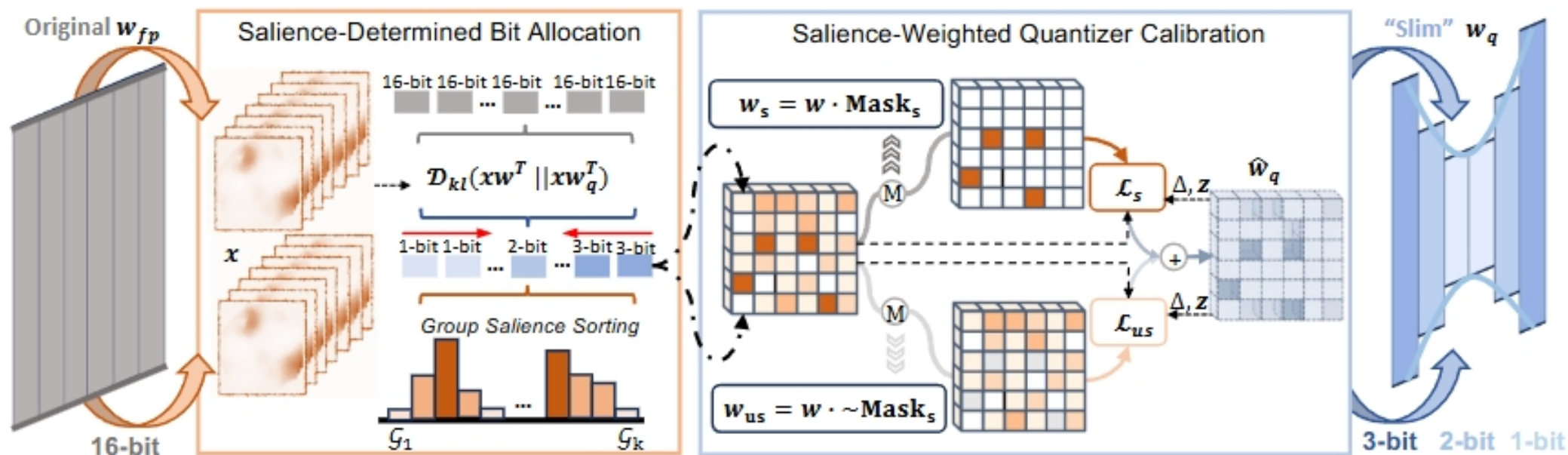
量化流程：

- 1.分组
- 2.SDBA分配比特
- 3.分组压扩
- 4.学习生成矩阵 Gg
- 5.得到整数索引 z
- 6.存储： $z+Gg+\mu g$



对权重 W 分组，每个组 g 分配bit
使用SDBA方法自动分配bit

SDBA:高显著组给 $N+1$ bit,低显著组给 $N-1$ bi,数量相等保证平均 N bit



2. 格型码本学习

每个权重组 g 学习一个生成矩阵 G_g ，以及一个整数向量 $\mathbf{z}_g$ 。

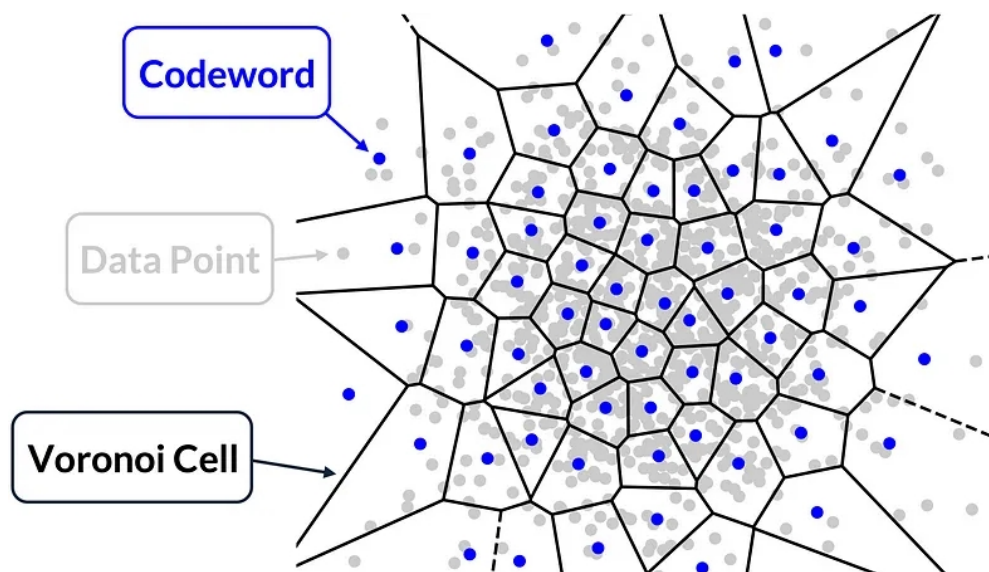
通过矩阵 G_g 定义的格结构对每个子块进行近似 $\widehat{\mathbf{W}}_g = G_g \mathbf{Z}_g$ ：

给定校准输入 $\mathbf{X}$ ，得到量化后输出的重构误差

$$\mathcal{L}_g = \|\mathbf{W}_g \mathbf{X} - \widehat{\mathbf{W}}_g \mathbf{X}\|_2^2 = \|\mathbf{W}_g \mathbf{X} - G_g \mathbf{Z}_g \mathbf{X}\|_2^2.$$

两步高效求解：1. $\mathbf{z}_i = \lfloor G_g^{-1} \mathbf{w}_i \rfloor$

$$2. \nabla_{G_g} \mathcal{L}_g = -2(\mathbf{W}_g \mathbf{X} - G_g \mathbf{Z}_g \mathbf{X})(\mathbf{Z}_g \mathbf{X})^\top$$



3.组专属扩压

为了解决权重分布的重尾特性，GLVQ在格型量化前插入了逐元素压缩扩展

对权重组 g 采用参数为 μ 的变换：

$$F_g(x) = \text{sgn}(x) \frac{\ln(1 + \mu_g |x|)}{\ln(1 + \mu_g)}, \quad F_g^{-1}(y) = \text{sgn}(y) \frac{(1 + \mu_g)^{|y|} - 1}{\mu_g}.$$

于是完整的编码—解码过程变为

$$\widetilde{\mathbf{W}}_g = F_g(\mathbf{W}_g) \rightarrow \mathbf{Z}_g = \lfloor \mathbf{G}_g^{-1} \widetilde{\mathbf{W}}_g \rfloor \rightarrow \widehat{\mathbf{W}}_g = F_g^{-1}(\mathbf{G}_g \mathbf{Z}_g).$$

Algorithm 1 Pseudo code of GLVQ algorithm.

Require: Group weights $\mathbf{W}_g \in \mathbb{R}^{m_g \times n_g}$, calibration inputs $\mathbf{X} \in \mathbb{R}^{n_g \times T}$, initial generation matrix $\mathbf{G}_g^{(0)} \in \mathbb{R}^{d \times d}$, initial companding parameter $\mu_g^{(0)}$

- 1: $\mathbf{G}_g \leftarrow \mathbf{G}_g^{(0)}, \quad \mu_g \leftarrow \mu_g^{(0)}$
- 2: **repeat**
- 3: $\widetilde{\mathbf{W}}_g \leftarrow F_{\mu_g}(\text{reshape}(\mathbf{W}_g, [d, \ell]))$ { apply companding; $\ell = m_g n_g / d$ }
- 4: $\mathbf{Z}_g \leftarrow \text{round}(\mathbf{G}_g^{-1} \widetilde{\mathbf{W}}_g)$ { Babai lattice rounding }
- 5: $\widehat{\mathbf{W}}_g \leftarrow \text{reshape}(F_{\mu_g}^{-1}(\mathbf{G}_g \mathbf{Z}_g), [m_g, n_g])$ { inverse compand & reshape }
- 6: $\mathcal{L}_g \leftarrow \|\mathbf{W}_g \mathbf{X} - \widehat{\mathbf{W}}_g \mathbf{X}\|_F^2 + \lambda \|\mathbf{G}_g - \mathbf{G}_g^{(0)}\|_F^2$ { reconstruction + regularizer }
- 7: $\mathbf{G}_g \leftarrow \mathbf{G}_g - \eta_G \nabla_{\mathbf{G}_g} \mathcal{L}_g$ { update generation matrix }
- 8: $\mu_g \leftarrow \mu_g - \eta_\mu \nabla_{\mu_g} \mathcal{L}_g$ { update companding parameter }
- 9: **until** $|\mathcal{L}_g^{(t)} - \mathcal{L}_g^{(t-1)}| / \mathcal{L}_g^{(t-1)} < \varepsilon$ { convergence check }
- 10: **return** $\widehat{\mathbf{W}}_g$

实验

Table 1: Perplexity ($\downarrow$) of competing methods on Wikitext2 and C4 for Llama 1 and 2. All results use a context length of 2048 for Llama 1 and 4096 for Llama 2.

Method	Bits	Wikitext 2							C4						
		1-7	1-13	1-30	1-65	2-7	2-13	2-70	1-7	1-13	1-30	1-65	2-7	2-13	2-70
Fp16	16	5.68	5.09	4.10	3.53	5.12	4.57	3.12	7.08	6.61	5.98	5.62	6.63	6.05	4.97
OmniQ	2	15.5	13.2	8.71	7.58	–	–	–	24.9	18.3	13.9	10.8	–	–	–
QulP#	2	6.86	5.97	5.02	4.36	6.19	5.35	3.91	8.36	7.48	6.71	6.19	8.16	7.20	5.71
QTIP	2	6.52	5.80	4.83	4.21	5.91	5.26	3.78	7.99	7.31	6.56	6.08	7.76	6.99	5.56
GLVQ-8D	2	6.28	5.64	4.57	4.01	5.69	5.02	3.62	7.83	7.17	6.48	5.94	7.33	6.56	5.25
GLVQ-32D	2	6.00	5.38	4.32	3.81	5.41	4.80	3.36	7.61	6.95	6.22	5.80	7.14	6.33	5.04

Table 2: Zero-shot accuracy (*acc* in LM Eval, not *acc_norm*) for Llama 2 models.

	2-70					2-13					2-7				
Method	Bits	ArcC	ArcE	PIQA	WINO	Bits	ArcC	ArcE	PIQA	WINO	Bits	ArcC	ArcE	PIQA	WINO
FP16	16	51.1	77.7	81.1	77.0	16	45.6	73.3	73.5	69.6	16	40.0	69.3	78.5	67.3
OmniQ	4	49.8	77.9	80.7	75.8	4	43.1	70.2	78.4	67.8	4	37.9	67.8	77.1	67.0
QulP	4	47.0	74.3	80.3	76.0	4	44.9	73.3	79.0	69.7	4	–	–	–	–
AQLM	4	51.0	78.1	81.4	76.9	4	43.9	72.2	78.6	70.4	4	40.3	68.9	77.7	67.3
QulP#	4	50.6	78.1	81.4	77.1	4	45.5	73.9	78.9	69.9	4	40.5	69.1	78.4	67.6
QTIP	4	50.0	77.6	81.5	77.0	4	45.9	73.2	78.6	69.9	4	40.4	69.2	78.5	67.4
GLVQ-8D	4	51.2	78.0	81.6	77.3	4	46.0	74.0	79.2	70.5	4	40.7	69.5	78.6	67.8
OmniQ	3	47.6	75.7	79.7	73.5	3	42.0	69.0	77.7	65.9	3	35.3	62.6	73.6	63.6
QulP	3	46.3	73.2	80.0	74.6	3	41.5	70.4	76.9	69.9	3	–	–	–	–
AQLM	3	50.0	77.6	81.3	77.2	3	43.6	73.5	77.8	67.6	3	38.7	67.8	76.6	68.4
QulP#	3	50.9	77.7	81.4	76.4	3	44.0	72.5	78.4	69.1	3	39.2	68.4	77.3	66.5
QTIP	3	49.8	77.5	81.2	76.3	3	43.6	72.5	78.2	69.5	3	39.8	69.7	78.0	66.8
GLVQ-8D	3	50.8	78.0	81.6	77.1	3	46.5	74.2	79.3	71.0	3	40.9	69.8	78.8	67.9
OmniQ	2	28.7	55.4	68.8	53.2	2	23.0	44.4	62.6	52.6	2	21.6	35.2	57.5	51.5
QulP	2	34.0	62.2	74.8	67.5	2	23.5	45.2	62.0	52.8	2	19.4	26.0	54.6	51.8
AQLM	2	47.9	77.7	80.4	75.9	2	38.5	67.0	75.1	69.5	2	33.6	62.8	73.5	64.6
QulP#	2	48.7	77.3	80.3	75.9	2	39.5	69.3	77.3	67.7	2	34.6	64.6	75.1	64.9
QTIP	2	48.1	76.9	80.1	76.5	2	39.2	70.6	77.8	71.0	2	35.3	63.9	75.3	66.7
GLVQ-8D	2	49.0	77.5	80.5	76.1	2	40.0	70.8	78.0	71.3	2	35.5	64.5	75.5	67.2