ICLR

The Geometry of LLM Quantization: GPTQ as Babai's Nearest Plane Algorithm

Jiale Chen (jiale.chen@ist.ac.at)

Yalda Shabanzadeh

Elvir Crnčević

Torsten Hoefler

Dan Alistarh

Red Hat

ETH zürich

Background & Motivation

- LLMs are expensive to store and run
- Post-training quantization reduces weights from 16-bit to low bitwidth
- GPTQ is widely used, but its success was not well understood theoretically
- **Main contribution:** GPTQ can be interpreted geometrically as Babai's nearest plane algorithm for the closest vector problem (CVP) on a lattice

From LLM Quantization to Lattice Problem

LLM Linear Layer Quantization

Find $\boldsymbol{Q}$ or $\boldsymbol{z}_i$ to minimize

$$\|\boldsymbol{XQ} - \boldsymbol{XW}\|_{\mathrm{F}}^2 = \sum_i \|\boldsymbol{X}\,\mathrm{diag}(\boldsymbol{s}_i)\,\boldsymbol{z}_i - \boldsymbol{X}\boldsymbol{w}_i\|^2$$

Closest Vector Problem (CVP)

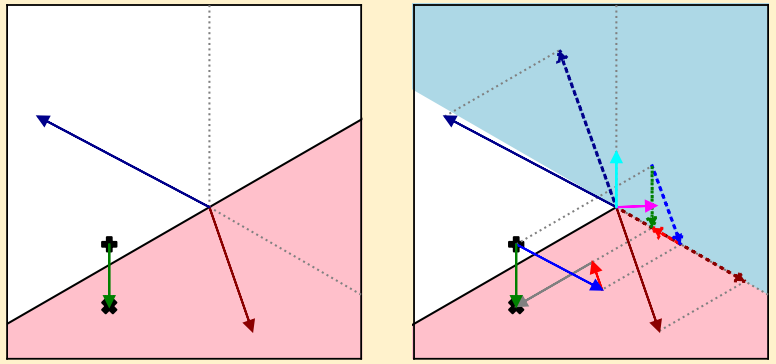
Find $\boldsymbol{z}$ to minimize

$$\|\boldsymbol{B}\boldsymbol{z} - \boldsymbol{y}\|^2$$

- Activations $\boldsymbol{X}$ → Basis directions
- Scales $\boldsymbol{s}_i$ → Basis stretches
- Quantized integers $\boldsymbol{z}_i$ → Lattice coordinates $\boldsymbol{z}$
- Target $\boldsymbol{X}\boldsymbol{w}_i$ → Point to approximate $\boldsymbol{y}$

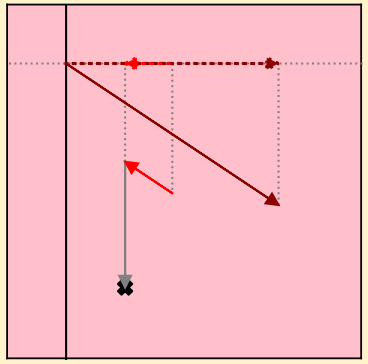
Main Theorem: GPTQ is Babai's Algorithm

- OBQ/GPTQ error propagation ⇔ Projection onto the nearest hyperplane
- Geometrically, GPTQ performs an orthogonal walk through affine subspaces

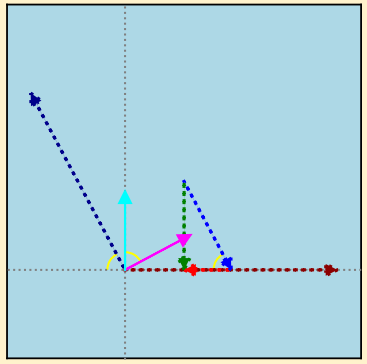


(a) [3D] Babai's Projection

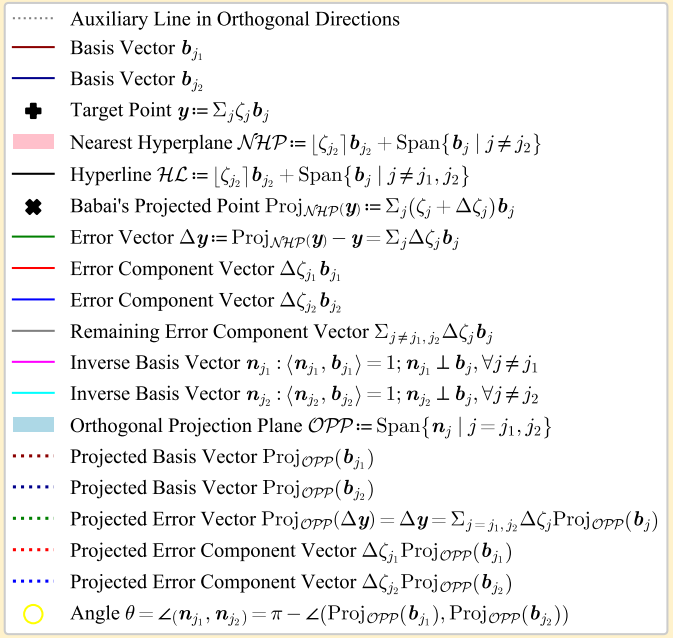
(b) [3D] Babai & OBQ Equivalence



(c) [2D] Nearest Hyperplane

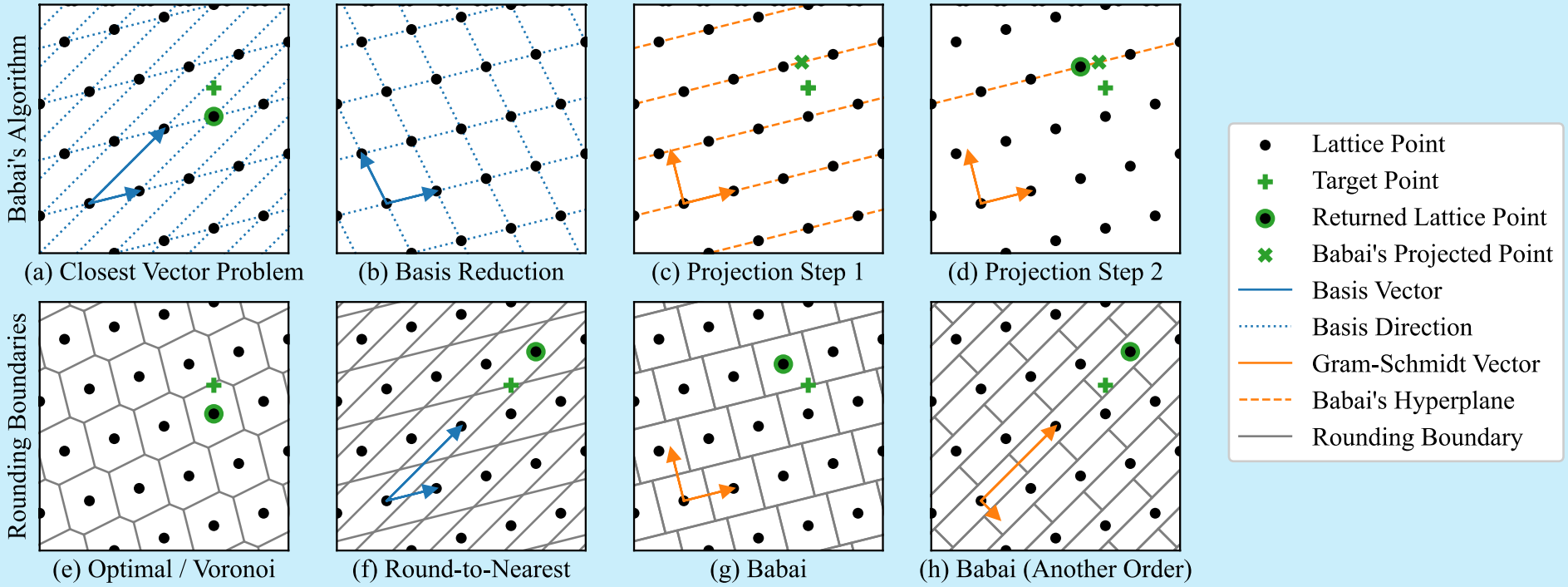


(d) [2D] Orthogonal Projection Plane

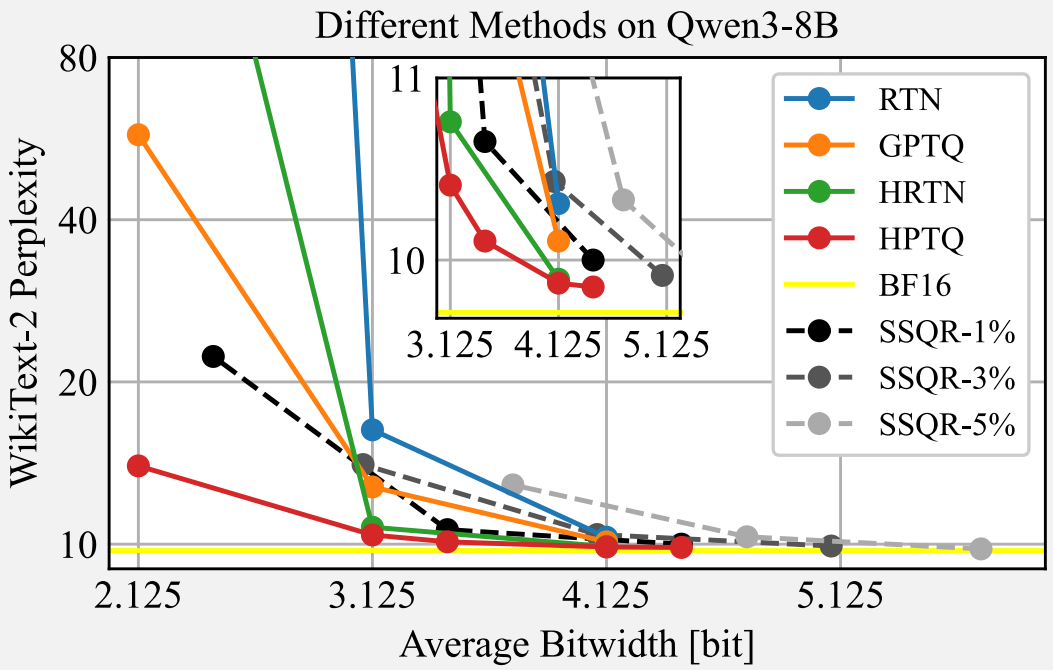


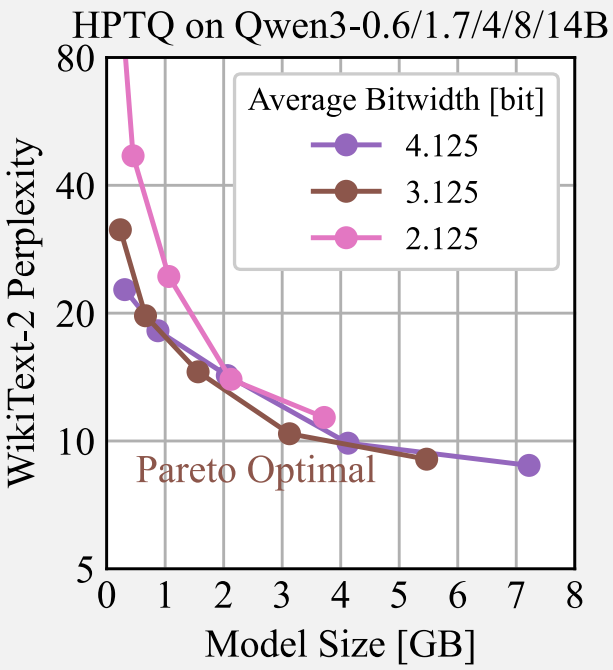
What the Equivalence Brings: Error Bounds

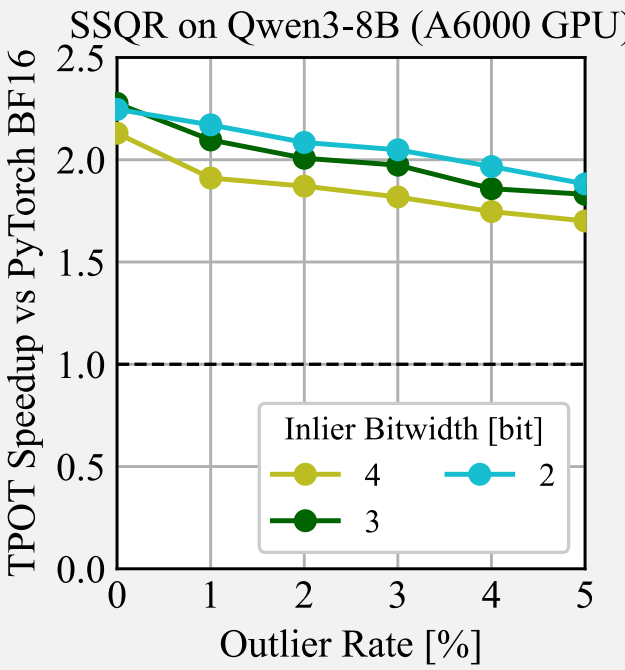
- In the no-clipping setting, GPTQ inherits Babai's tight error bound
- The bound depends on the LDL decomposition of the permuted Hessian
- Quantization order matters



Practical Impact: Better Quantizations







Paper
arXiv:2507.18553

Code
IST-DASLab/GPTQ-Babai