ICLR

Paper
arXiv:2507.18553

Code
IST-DASLab/GPTQ-Babai

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

Institute of Science and Technology Austria
Red Hat
ETH zürich

Equivalence

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

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

Algorithm 1: GPTQ

Input: $\boldsymbol{W}, \boldsymbol{S}, \boldsymbol{X}, \boldsymbol{P}, \lambda, \mathbb{Z}_\dagger$
Output: $\boldsymbol{Z}, \boldsymbol{Q}$

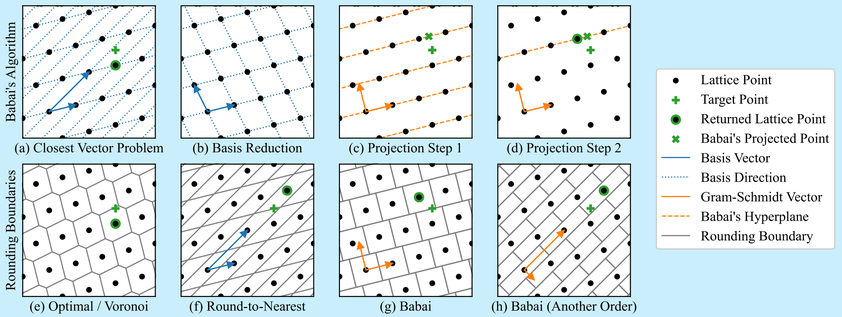
$\boldsymbol{H} \leftarrow \boldsymbol{P}^\top (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \mathbf{I})\, \boldsymbol{P}$
$\boldsymbol{L} \leftarrow \mathrm{LDL}\,(\boldsymbol{H}^{-1})$
$\boldsymbol{W}, \boldsymbol{S} \leftarrow \boldsymbol{P}^{-1}\boldsymbol{W}, \boldsymbol{P}^{-1}\boldsymbol{S}$
$\boldsymbol{Q}, \boldsymbol{Z} \leftarrow \boldsymbol{W}, \boldsymbol{0}$
for $j \leftarrow 1$ to c **do**
  $\boldsymbol{\zeta} \leftarrow \boldsymbol{W}[j,:]/\boldsymbol{S}[j,:]$
  $\boldsymbol{Z}[j,:] \leftarrow \textsc{Round}\,(\boldsymbol{\zeta}, \mathbb{Z}_\dagger)$
  $\boldsymbol{Q}[j,:] \leftarrow \boldsymbol{Z}[j,:] * \boldsymbol{S}[j,:]$
  $\boldsymbol{\varepsilon} \leftarrow \boldsymbol{Q}[j,:] - \boldsymbol{W}[j,:]$
  $\boldsymbol{W}[j:,:] \leftarrow \boldsymbol{W}[j:,:] + \boldsymbol{L}[j:,j]\boldsymbol{\varepsilon}$
end
$\boldsymbol{Z}, \boldsymbol{Q} \leftarrow \boldsymbol{PZ}, \boldsymbol{PQ}$

Algorithm 2: Babai's Nearest Plane

Input: $\boldsymbol{B}, \boldsymbol{y}$
Output: $\boldsymbol{z}$

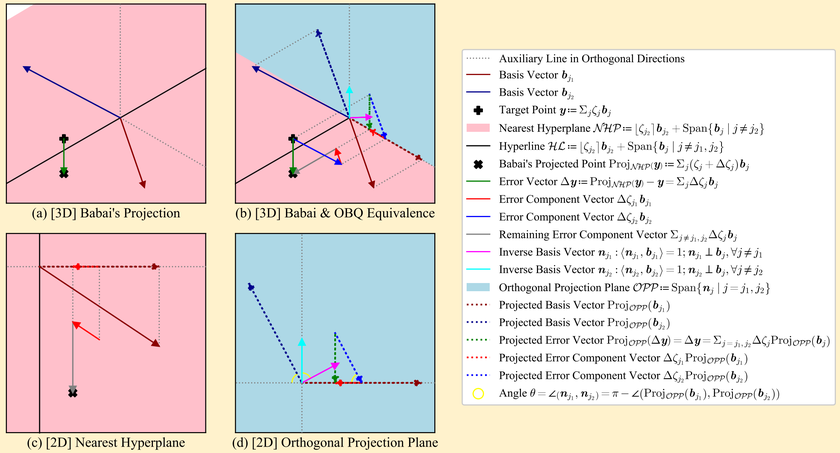
$\boldsymbol{T} \leftarrow \mathrm{LLL}\,(\boldsymbol{B})$ // transformation
$\boldsymbol{A} \leftarrow \boldsymbol{BT}$ // basis reduction
$\boldsymbol{\Phi} \leftarrow \mathrm{QR}\,(\boldsymbol{A})$ // orthogonalize
$\boldsymbol{y}', \boldsymbol{z} \leftarrow \boldsymbol{y}, \boldsymbol{0}$
for $j \leftarrow c$ to 1 **do**
  $\zeta \leftarrow \langle \boldsymbol{\Phi}[:,j], \boldsymbol{y}' \rangle \,/\, \langle \boldsymbol{\Phi}[:,j], \boldsymbol{A}[:,j] \rangle$
  $\boldsymbol{z}[j] \leftarrow \textsc{Round}\,(\zeta, \mathbb{Z})$
  $\boldsymbol{y}' \leftarrow \boldsymbol{y}' - \boldsymbol{A}[:,j]\boldsymbol{z}[j]$
end
$\boldsymbol{z} \leftarrow \boldsymbol{Tz}$

GPTQ Error Bound & Quantization Order

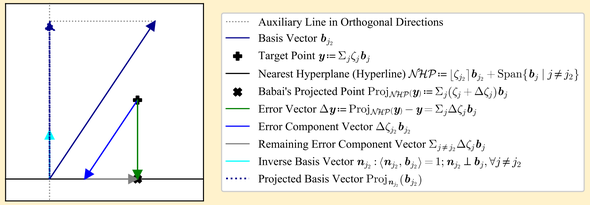


Proof of Equivalence

GPTQ/OBQ Error Propagation ⇔ Nearest Plane Projection



OBQ Quantization Order ⇔ "Nearest" Nearest Plane



Applications: Clip-Free GPTQ & Inference Kernel