The Lattice Geometry of Neural Network Quantization

A Short Equivalence Proof of GPTQ and Babai's Algorithm

Johann Birnick

recorded on 28 Mar 2026 for ICLR 2026

Quantization

The core of a neural network is a *neuron*, which is a linear map $\mathbb{R}^n \to \mathbb{R}$

$$x \mapsto x^\top w$$

where $w \in \mathbb{R}^n$ makes up the *parameters* and $x \in \mathbb{R}^n$ the *input data*.

Problem: Storing w in 32-bit or 16-bit precision takes a lot of memory, and computing with 32-bit or 16-bit floats is expensive.

Solution: We use low-bit data types to approximate w.

However: Naive rounding kills accuracy, we need *"smart rounding"*.

Problem Setup

We want to find $v \in \mathbb{Z}^n$ that approximates w well *on certain input data* $x_1, \ldots, x_k \in \mathbb{R}^n$.

> **Problem.** Given $X \in \mathbb{R}^{k \times n}$ and $w \in \mathbb{R}^n$, find $v \in \mathbb{Z}^n$ such that $\|Xw - Xv\|_2$ is as small as possible.

Here $X = \begin{pmatrix} — x_1 — \\ — x_2 — \\ \vdots \\ — x_k — \end{pmatrix}$ is tall, i.e., $k \gg n$, and is called *calibration data.*

Outline

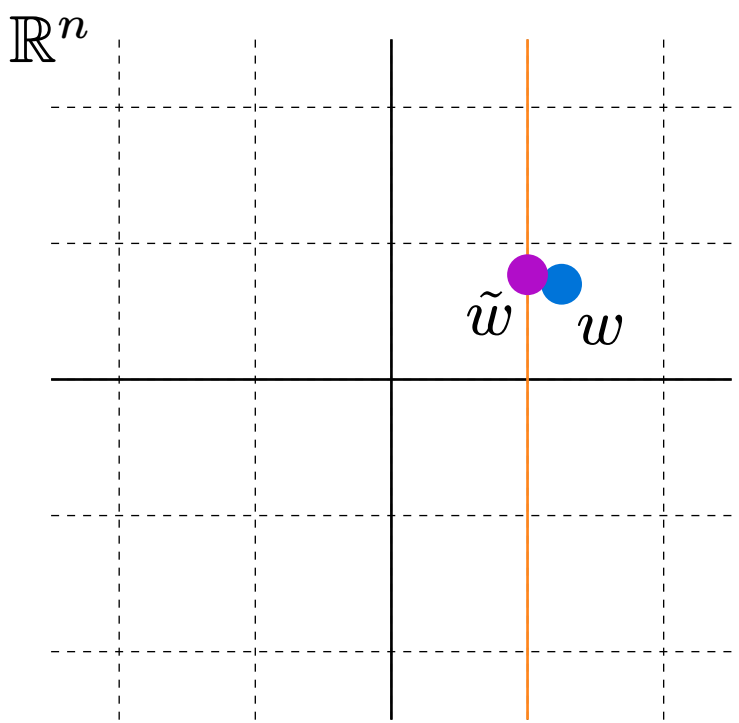
Problem. Want to find $v \in \mathbb{Z}^n$ such that $\|Xw - Xv\|_2$ is very small.

- The GPTQ Algorithm [1] from 2022 was created for quantization purposes, and remains widely used due to good quantization results.
- **Core contribution**: We prove that this algorithm is equivalent to *Babai's nearest plane algorithm* [2] from 1986, a core result from the theory of *lattices*.
- This yields a new geometric perspective, error guarantees for GPTQ, and ideas for improved quantization algorithms.

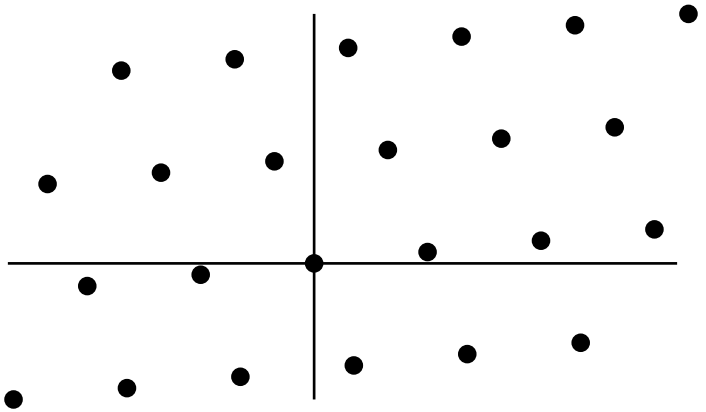
GPTQ Algorithm

Problem. Want to find $v \in \mathbb{Z}^n$ such that $\|Xw - Xv\|_2$ is very small.

1. Pick $v_1 := \text{round}(w_1)$.
2. Replace w by $\tilde{w}$ which minimizes $\|Xw - X\tilde{w}\|_2$ under the constraint $\tilde{w}_1 = v_1$.
3. Continue in the same fashion with $w_{\geq 2}$.

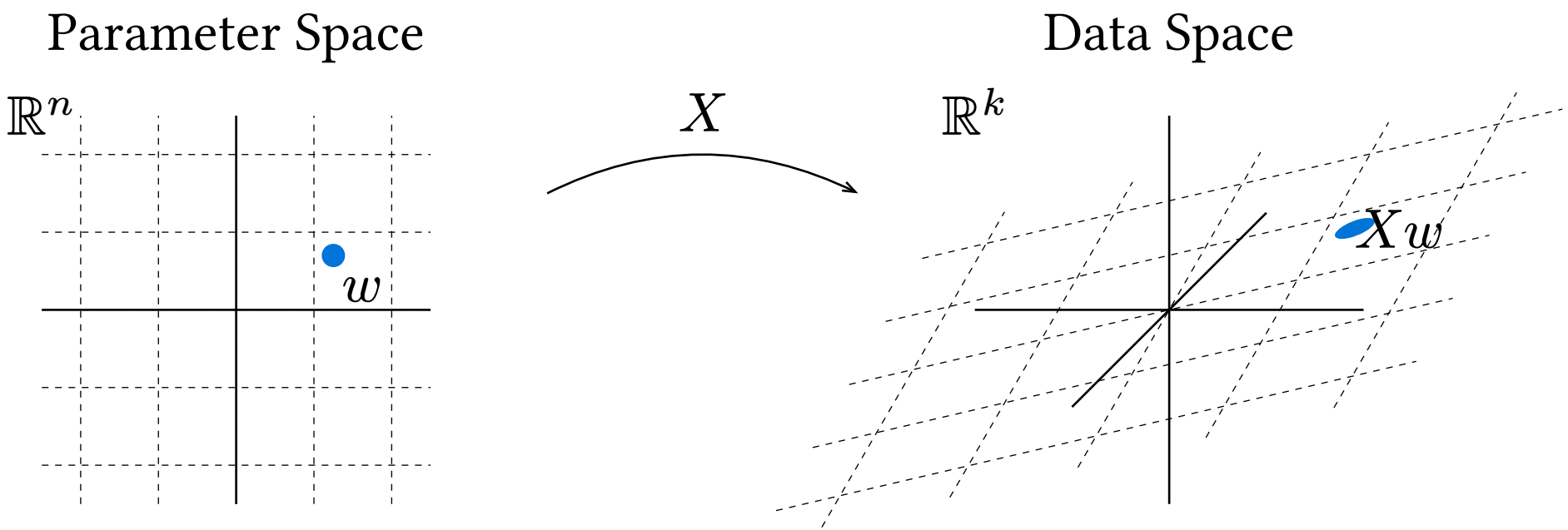


Lattices



- A *lattice* is the $\mathbb{Z}$-span $\mathbb{Z}b_1 + \ldots + \mathbb{Z}b_n$ of a set of $\mathbb{R}$-linearly independent vectors $b_1, \ldots, b_n$ in $\mathbb{R}^k$.
- The vectors $b_1, \ldots, b_n$ are called a *basis* of the lattice.
- Multiple different bases can produce the same lattice.

The Geometry

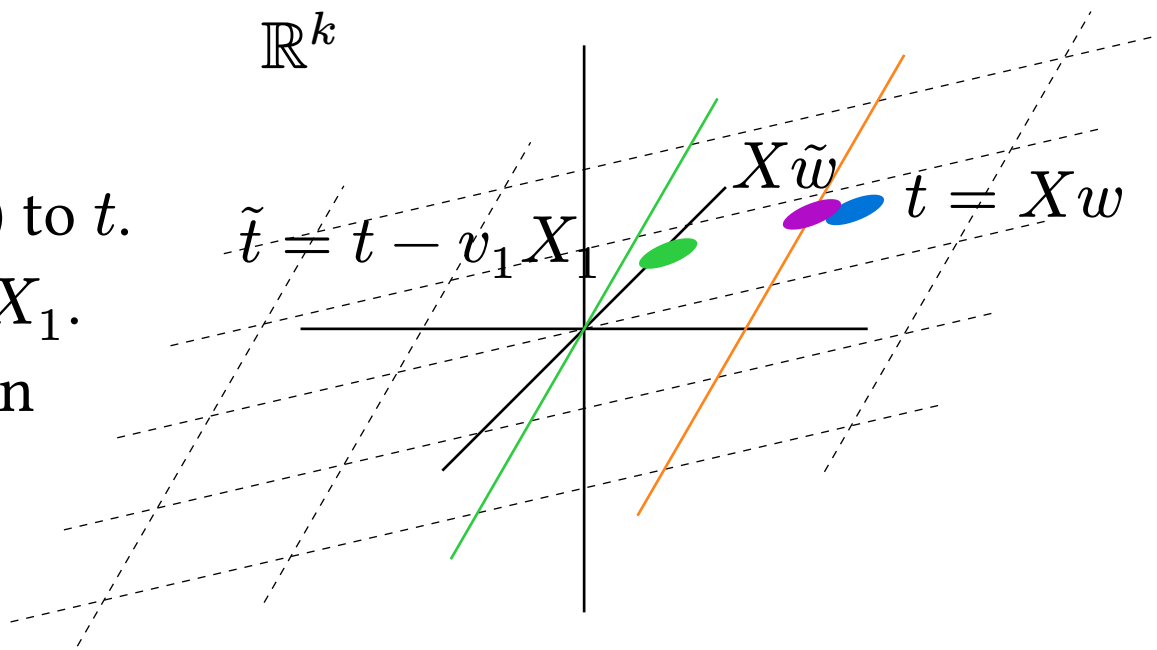


X maps the quantization grid $\mathbb{Z}^n$ to a lattice in $\mathbb{R}^k$. GPTQ works in $\mathbb{R}^n$, on the left. Babai's algorithm works in $\mathbb{R}^k$, on the right.

Babai's Algorithm

Problem. Want to find $v \in \mathbb{Z}^n$ such that $\|Xw - Xv\|_2$ is very small.

0. Compute target $t := Xw$.
1. Pick v_1 as the index of the nearest plane (certain sense) to t.
2. Update target to $\tilde{t} := t - v_1 X_1$.
3. Continue in the same fashion with $v_{\geq 2}$.



The Equivalence

Theorem. The procedures GPTQ and Babai are equivalent.

procedure GPTQ$(X,\, w)$
  ▷ Compute $QL = X$.
  $\tilde{L} \leftarrow L^{-1}$
  $w^{(0)} \leftarrow w$
  for $i = 1, \ldots, n$ **do**
    $v_i \leftarrow \text{round}\left[w_i^{(i-1)}\right]$
    $\Delta_i \leftarrow v_i - w_i^{(i-1)}$
    $w^{(i)} \leftarrow w^{(i-1)} + \frac{\Delta_i}{\tilde{L}_{i,i}} \cdot \tilde{L}_i$
  end
  return v
end

procedure Babai$(X,\, w)$
  ▷ Compute $QL = X$.
  $t^{(0)} \leftarrow Xw$
  for $i = 1, \ldots, n$ **do**
    $v_i \leftarrow \text{round}\left[\frac{\langle t^{(i-1)}, Q_i \rangle}{L_{i,i}}\right]$
    $t^{(i)} \leftarrow t^{(i-1)} - v_i X_i$
  end
  return v
end

Consequences (see paper for details)

Error Guarantees. There are well-known error bounds for Babai's algorithm, which directly carry over to GPTQ.

Handling of Quantization over Multiple Layers. The problem then becomes to minimize $\left\|Xw - \hat{X}v\right\|_2$ where $X \neq \hat{X}$. Babai's perspective shows how one should deal with this correctly, and it has been experimentally verified by Qronos [3] to outperform GPTQ.

Use of Lattice Reduction. A classic technique to solve the closest vector problem is to pre-process the lattice basis before running Babai's algorithm, and one could potentially leverage this in quantization too.

Check out the paper!

SCAN ME

The Lattice Geometry of Neural Network Quantization –
A Short Equivalence Proof of GPTQ and Babai's algorithm

arxiv.org/abs/2508.01077

References

[1] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, "Gptq: Accurate post-training quantization for generative pre-trained transformers," *arXiv preprint arXiv:2210.17323*, 2022.

[2] L. Babai, "On Lovász'lattice reduction and the nearest lattice point problem," *Combinatorica*, vol. 6, no. 1, pp. 1–13, 1986.

[3] S. Zhang, H. Zhang, I. Colbert, and R. Saab, "Qronos: Correcting the Past by Shaping the Future... in Post-Training Quantization," *arXiv preprint arXiv:2505.11695*, 2025.