L01

Plan:

(0) • Lectures outline
• preregs: no IT today, best rec — 5 min

(1) Motivation: tokenization, weight & act. quant. — 15 min
+ classical name: lossy compression.

(2) Scalar quant: Pauter-Dite
→ Smooth Quant error
→ dithering.

(3) Non-MSE error:
→ motivation from matmul
→ LDLQ
→ YAQA?

⓪ Lecture outline:

"weights-only"
- L01: Scalar quantization
- L02: Vector Q. (VQ-VAE +IT)
- L03: Info-theory of VQ

"wt + activations"
- L04: LLM & Matmul Quantization
  Open problems

① Motivation

→ Classic: signals → bits

→ Modern 1: LLMs spend most time doing this $Y = WX$.

- GPU does 600–1000 FLOPS during the time it takes to load 1 byte to the multipliers!

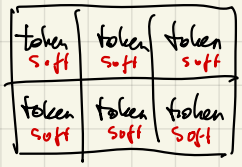
High Rate
R = 1...4

↳ want to store W and X compactly
(FP32 → BF16 → int8 → int4, FP4, ...)

Energy: load FP32 from DRAM = 640 pJ
one FP add-mult = .4 pJ.
(int add-mult = 3 pJ)

→ Modern?: tokenization of non-language modalities

text → <token 1> <token 2> ...

image → VAE Encoder → 

Some archs use soft tokens in MLM.

(DALL-E 2+, Gemma 3, Llama 3.2-Vision..., Llava, ... Audio: OpenAI whisper)

Others first convert them to discrete "hard" tokens:

(VQ-GAN, DALL-E 1, Byte Dance VAR

Audio: every one except whisper

Meta Encodec

Google Soundstream

Boson.AI Higgs2 ...)

Low Rate

R < 1

R = 0.05 VQ-GAN

R = 0.4 VQ-ViT

SoundStream?

→ Personal relevant theorem for LLMs.

(2) Scalar quantization:

(a) Simplest form.

$$x \to q_\varepsilon(x) = \varepsilon\left[x/\varepsilon\right]$$

x_.inf() RTN

Problem: ∞ number of integers

Classical: user's job is to make sure

$$x \in [-1/2; 1/2] \Rightarrow \#\text{bits} = R = \log_2 2/\varepsilon$$

Every input to your phone passes through A2D

Question: What is error?

If $X \sim P_X$ - cts density on $[-1;1]$ $\Rightarrow$ unif. cts

$\Rightarrow$ as $\varepsilon \to 0$ $P_X(q_\varepsilon(x) = j\varepsilon \to \text{Unif}[-\varepsilon/2; \varepsilon/2]$

$$\Rightarrow \mathbb{E}\left(x - q_\varepsilon(x)\right)^2 = \varepsilon^2 \frac{1+o(1)}{12} \approx \frac{1}{12} 2^{-2R}$$

Remark: In signal quality is usually measured in dB:

$$10 \log_{10} \frac{1}{12} 2^{-2R} \approx -11 + (6\,\text{dB}) \times R$$

I.e. 1 bit = 6dB ($4\times$ reduction of MSE)

So 16 bit digitized audio is often said to have SNR = 96dB

Uniform Q. is still going strong today (Tim Dettmers L[M.inf%() → bits and bytes ...)

(6) Dithering:

Problem: • When analysing error we had to take $\varepsilon \to 0$. Is there a way to make analysis rigorous w/o $\varepsilon \to 0$?

Yes! Dithering! Also helps with actual performance.

History: Invented by Royal Navy ≈WW1 when navigation device was noticed to have poor perf. at stationary.

• Given X let U be indep. dither $\mathrm{Unif}[-\varepsilon/2, \varepsilon/2]$

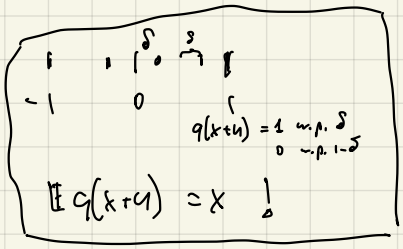
(In practice, pseudo random → shared b/w compressor and decompressor)

$$\hat{X} = q(X+U) - U.$$

(shared; gets stored/sent.)

• Equiva to randomly shifting ε-grid

• helps when $X_1 = X_2 = \dots = X_{100} \approx \varepsilon/3$

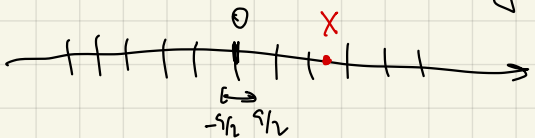


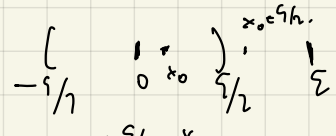
Prop. (Crypto Lemma)

$$(\hat{X}, X) \overset{(d)}{=} (X+\tilde{U}, X) \qquad \tilde{U} \perp\!\!\!\perp X$$

Pf: $\hat{X} - X \mid X$ is $\mathrm{Unif}[-\varepsilon/2, \varepsilon/2]$

Indeed, fix X then $\hat{X}$ = n.n. on a ε-grid randomly displaced by $(-\varepsilon/2, \varepsilon/2)$





$$q(x_0+u) = \begin{cases} 0, & u < \varepsilon/2 - x_0 \\ \varepsilon, & u \geq \varepsilon/2 - x_0. \end{cases}$$

$$\hat{X} - x = q(x+u) - u - x = -(\text{rounding error of } x+u)$$

But $x+u$ given x is equally likely to appear anywhere in the bin.

- This is helpful, because now for Dithered Unif. Quant. we can rigorously state

$$\mathbb{E}(x-\hat{x})^2 = \varepsilon^2/12.$$

and do things like: what is error of x^2?

$$\mathbb{E}\left(x^2 - \hat{x}^2\right)^2 = \mathbb{E}\left(x^2 - (x+\tilde{u})^2\right)^2 =$$

$$= \mathbb{E}\left(-2x\tilde{u} - \tilde{u}^2\right)^2 = (\mathbb{E}x^2)\varepsilon^2/3 + \varepsilon^4/80.$$

We will use later: $\mathbb{E}\left(xy - \hat{x}\hat{y}\right)^2 = \mathbb{E}\left(u_2 x + u_1 y + u_1 u_2\right)^2$

$$= \frac{\varepsilon^2}{12}\left(\mathbb{E}x^2 + \mathbb{E}y^2 + \frac{\varepsilon^2}{12}\right)$$

- Empirical fact: DUQ analysis usually perfectly matches empirical perf. of UQ on real world data, but is fully rigorous.

This is important trick to do rigorous proofs in this field

(c) Can we do better than uniform q.?

NB! Do not talk about MX part

Example: 2024 MXFP4 standard (GPT-OSS q.)

$(X_1, \dots, X_{32}) = 2^{0 \dots 255} \times (\underbrace{\tilde{x}_1}_{\text{FP4}}, \dots, \underbrace{\tilde{x}_{32}}_{\text{FP4}})$

$R = 4.25$ (vector q.)

FP4 = 1S, 2E, 1M : $\pm 2^{\{0,1,2,3\}} (1 + 2^{-1}\{0,1\})$

$\pm (2^{-1}\{0,1\})$

so we get: $\pm 0, \pm\frac{1}{2}, \pm 1, \pm 3/2, \pm 2, \pm 3, \pm 4, \pm 6$.

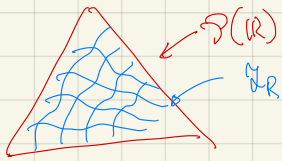
The above format was chosen for legacy & computational reasons.

Let us understand how can we do better?

$$D_{\text{scalar}}(R, P_X) = \inf_{|q(x)| \le 2^R} \mathbb{E}|X - q(X)|^2$$

- Ugly non-convex problem.
- Solution unknown even for $P_X = \mathrm{Unif}(0,1)$ and $2^R = 3$
- Wasserstein projection.
- But some optimality claims can still be made.

This is Wasserstein projection of P_X on $\Sigma_R = \{Q_X : |\mathrm{supp}\, Q_X| = 2^{nR}\}$

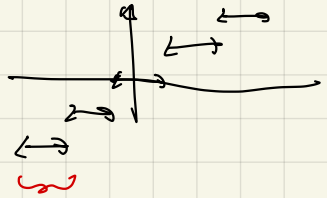


Prop

$D_{\text{scalar}} = \inf_{P_{X,Y}} \{\mathbb{E}|X-Y|^2 : |\mathrm{supp}| \le 2^{nR}$

- Given $q(\mathbb{R}) = \{c_1, \dots, c_M\}$ – codebook.

the map q is clear: $q(x) =$ nearest nb in C

$\Rightarrow$ $q =$ 

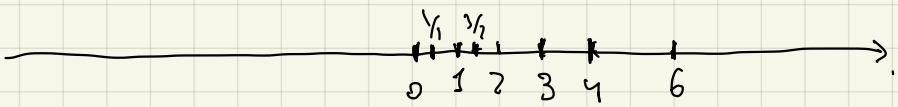
monotone, piecewise constant

quantization bin

- Conversely if quant. bins are chosen

then $c_i = \underset{u}{\operatorname{argmin}} \mathbb{E}\left[(X-u)^2 \mid X \in B_i\right] = \mathbb{E}[X \mid B_i]$

- Immediate implication. Suppose you quantize X via FPq.



If $q(x) = 6$ Is it a good idea to say $X \approx 6$ vs $X \approx \mathbb{E}[X \mid X > 5]$?

(Note that block scaling tries to move $P_X \subset [-6, 6]$)

I am waiting for a 6000 c.f. paper from DeepMind on this.

(d) High-rate quantization (Panter-Dite).

Th. Suppose $X \sim P_X$ with cts bdd pdf f_X then

$$D_{\text{scalar}}(R) = \frac{1}{12}\left(\int f_X^{1/3}\,dx\right)^3 2^{-2R}(1+o(1)),\ R\to\infty$$

Pf:

Introduce. $\cdot\ \lambda(x)$ → point density, $\int\lambda(x)=1$

$\cdot$ put N points c_j at s.t. $\int_{-\infty}^{c_j}\lambda(x)=j/N$.

$\cdot$ ⇒ in interval Δ_j $I\approx N\int_I\lambda(x)$ pts.

$$\mathbb{E}|X-q(X)|^2 = \sum_j \mathbb{E}\left[|X-c_j|^2 \mid X\in\Delta_j\right]\mathbb{P}[X\in\Delta_j]$$

$$\approx \sum_j \frac{\Delta_j^2}{12}\,\mathbb{P}[X\in\Delta_j].$$

$$\approx \sum_j \frac{1}{12N^2}\lambda(c_j)^{-2}\,\mathbb{P}[X\in\Delta_j]$$

$$\approx \frac{1}{12N^2}\int \frac{P_X(x)}{\lambda^2(x)}\,dx$$

Δ_j … $\frac{1}{N}$; $\int_{\Delta_j}\lambda(x)=1/N$; $\Delta_j\approx\lambda(c_j)^{-1}\frac{1}{N}$

→ $\min_{\lambda:\int\lambda=1}$ $\lambda^*=\frac{P_X^{1/3}}{\int P_X^{1/3}}$

⇒

$$D_{\text{scalar}}(P_X,R)\approx\frac{1}{12}2^{-2R}\left(\int P_X^{1/3}\right)^3,\quad R\to\infty.$$

For Gsn: $P_X=\mathcal{N}(0,\sigma^2)$: $D\approx\sigma^2 2^{-2R}\cdot\frac{\pi\sqrt{3}}{2}$.

By Reverse Hölder:

$$\int P_X\frac{1}{\lambda^2}\ge\|P_X\|_p\left\|\frac{1}{\lambda^2}\right\|_q$$

set $q=-1/2$, $p=\frac{1}{3}$

$$\Rightarrow \mathbb{E}|X-q(X)|^2\ge\frac{1}{12N^2}\left(\int P_X^{1/3}\right)^3$$

Mention naive arguments: each bin $\approx 1/N$ prob-ty to maximize $H(X)$, (used in SmoothQuant etc) completely wrong!

Preview: Shannon's breakthrough was to show that $D_{\text{vector}}(R;\mathcal{N}(0,\sigma^2)) = \sigma^2\cdot 2^{-2R} < D_{\text{scalar}}$!

Note: if domain of X is $[-1/2, 1/2]$ then.

$$\left(\int p_x^{1/3}\right)^3 \leq \left(\int p_x\right) = 1 \text{ !}$$ by Jensen's

$\Rightarrow$ Non-Uniform quant. provably wins.

(3) What about quantizing a vector?

(a) Uniform case

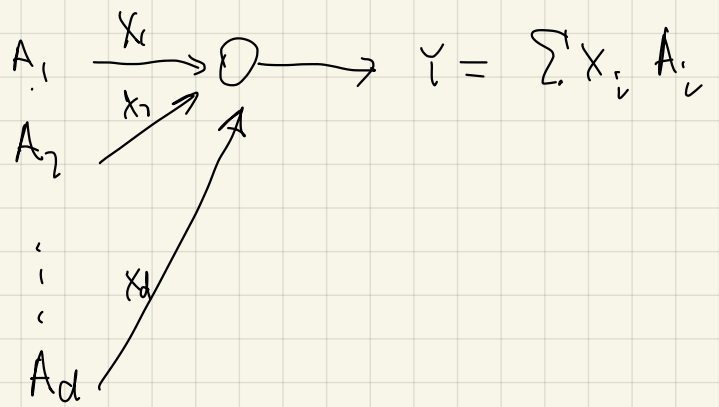
$$\begin{bmatrix} X_1 \\ \vdots \\ X_d \end{bmatrix} \rightarrow \begin{bmatrix} q(x_1) \\ \vdots \\ q(x_d) \end{bmatrix}$$

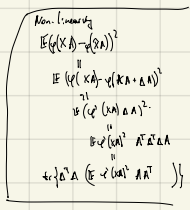
Is that it?

(b) Modern case.

X – weights of linear layer.

A – activations from prev. layer.





We want to find $\hat{X} \in \mathbb{Z}^d$ s.t.

$$\min \mathbb{E}\left(X^T A - \hat{X}^T A\right)^2 \cdot$$

$$\Updownarrow$$

$$\min_{\hat{X} \in \mathbb{Z}^d} (X-\hat{X})^T H (X-\hat{X})$$

$$H = \mathbb{E}\, AA^T$$

or $\min \mathbb{E}\left(\sigma(XA)-\sigma(\hat{X}A)\right)^2$

or $H = \mathbb{E}\,\sigma'(X^T A)^2\, AA^T$

(c) $\min_{\hat{x} \in \mathbb{Z}^d} (x - \hat{x})^T M (x - \hat{x})$.

Claim: This is post-quantum hard.

Pf: $\min_{\hat{x}} (x - \hat{x})^T Q^T Q (x - \hat{x}) \iff \min_{\hat{y} \in \Lambda} \|y - \hat{y}\|^2$

Λ = $\mathbb{Z}$-span of col's of Q

Nearest vec. search in Lattice....

(d) Popular heuristic (CPTQ):

LDLQ notes:

Target: $\min_{\hat{x} \in \mathbb{Z}^d} (x-\hat{x})^T H (x-\hat{x})$, $H \succ 0$.

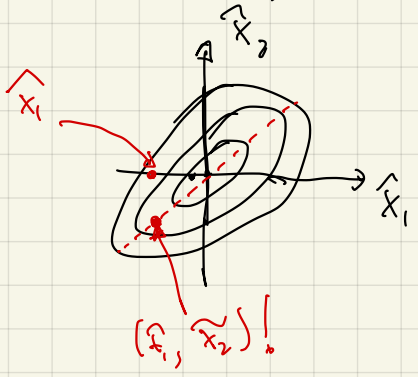
- Claim 1: NP-hard

  Pf: Reduce to NN Lattice quant.

- great heuristic: (OBC, GPTQ, LDLQ) 90's arxiv:2210.17323 arxiv:2307.13304

  1) $\hat{x}_1 = [x_1]$

  2)



Now the minimizer $\tilde{x}_2^d$ $\quad f(\hat{x}_1, \tilde{x}_2, \dots, \tilde{x}_d) \neq x_2^d$!

3) $\operatorname{argmin}_{\tilde{x}_2^d} \begin{bmatrix} (x_1-\hat{x}_1) \\ e \end{bmatrix}^T H \begin{bmatrix} x-\hat{x}_1 \\ e \end{bmatrix} = \begin{bmatrix} H_{11} & H_{21}^T \\ H_{21} & H_{22} \end{bmatrix}$

$= \Delta_1^2 H_{11} + 2\Delta_1 H_{21}^T e + e^T H_{22} e.$

$= \text{const} + (e-e^*)^T H_{22} (e-e^*)$

$e^* = -2H_{22}^{-1} H_{21} \Delta_1$. $\quad 2\Delta_1 H_{21} + H_{22} e = 0.$

$x_2^d - \tilde{x}_2^d = -2H_{22}^{-1} H_{21} \Delta_1$

$\hat{x}_2^d = x_2^d + 2H_{22}^{-1} H_{21} \Delta_1.$

4) From P.O.V. of $\hat{x}_2^q$ now we have the

$$f(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_d) = \text{const} + \underbrace{(\tilde{x}_2^d - \hat{x}_2^d)^T H_{22} (\tilde{x}_2^d - \hat{x}_2^d)}$$

5) Thus

$d \leftarrow d-1$.

$X \leftarrow \tilde{x}_2^d$

Repeat !

6) Does it matter? GPTQ paper on OPT-175B, 4-bit

	FP16	8.34
perp.W2:	RTN	10.54
	GPTQ	8.37

→ 98.5% reduction in gap!!

OPT-66B is even worse: RTN fails! GPTQ gives 9.55 vs 9.34

7) Note: this time $\Delta = X - \hat{X}$ has non-diagonal cov. matrix. (errors are correlated). How to find this error dist.?

$$\Delta_k = X_k - q\left(X_k + \sum_{j=1}^{k-1} a_{kj} \Delta_j\right)$$

$$\overset{(*)}{\approx} X_k - \left(X_k + \sum_{j=1}^{k-1} a_{kj} \Delta_j + Z_k\right)$$

(Z_k: uniform $[-\frac{s}{2}, \frac{s}{2}]$)

$$= -\left(A\Delta\right)_k + Z_k$$

(A: strictly lower-triang.)

$$\Delta = (I + A)^{-1} Z$$

$$\mathbb{E}\, \Delta^T H \Delta = \frac{1}{12} \operatorname{tr} H \Sigma_\Delta = \frac{1}{12} \operatorname{tr} H (I+A)^{-1} (I+A)^{-T}$$

($(I+A)^{-1}$: lower triang.)

$$\begin{bmatrix} 0 & x \\ 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & x & x \\ 0 & 1 & x \\ 0 & 0 & 1 \end{bmatrix}$$

$U_{tr}^{-1} \in U_{tr}$

Note: If $H = (I+B)^T D (I+B)$

(B: strictly lower tri.)

$$\operatorname{tr} D \underbrace{(I+B)(I+A)^{-1}}_{E} E^T = \sum_{i,j} d_i e_{ij}^2 \geq \sum_i d_i$$

↳ so we see that if $H = (I+B)^T D (I+B)$ then $A = B$! this allows to skip submatrix inversions. B is computed via LDL decomposition.

History: OBC was via inversion
GPTQ / LDLQ → via LDL / Cholesky