EE269
Signal Processing and Quantization for Machine Learning

Non-uniform Quantizers, Lloyd–Max Optimality and High-Rate Theory

Instructor : Mert Pilanci

Stanford University

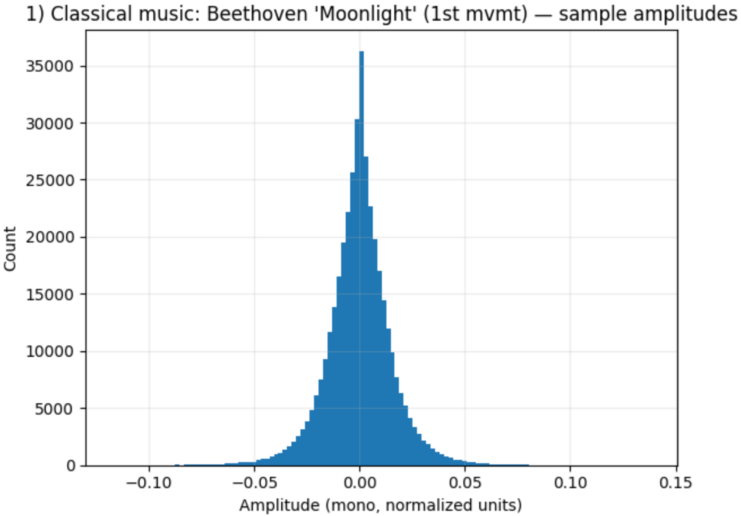
Outline

- from uniform to non-uniform quantization
- objective: distortion-rate tradeoffs (fixed-rate scalar quantization)
- optimality conditions (Lloyd–Max)
- "solutions" for Gaussian and other distributions
- Lloyd–Max iterative algorithm
- connection to 1D k-means (and vector quantization)
- practical non-uniform quantizers (companding, dead-zone, log)
- NormalFloat and quantizing Large Language Models

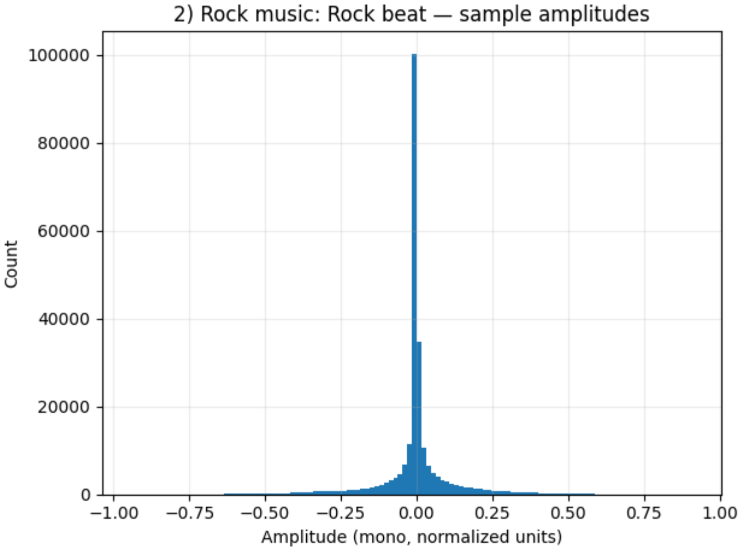
Why go beyond uniform quantizers?

- Uniform quantizers use a constant step size Δ across the dynamic range.
- Many real signals have *non-uniform* statistics:
    - speech and transform coefficients are often peaked near 0 (Laplacian-like)
    - sensor noise can be close to Gaussian
    - heavy tails / outliers are common in ML activations
- Uniform bins waste resolution where the pdf is small and under-resolve where the pdf is large.
- **Idea:** allocate more bins where $f_X(x)$ is high.

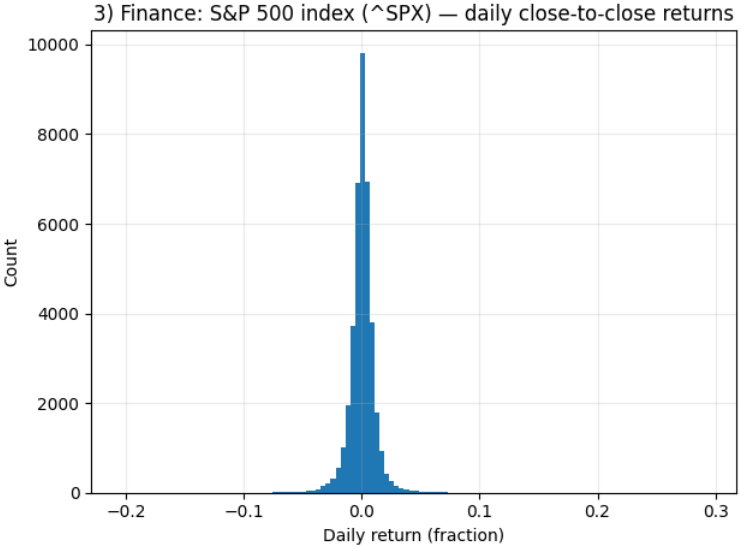
Histograms of amplitudes



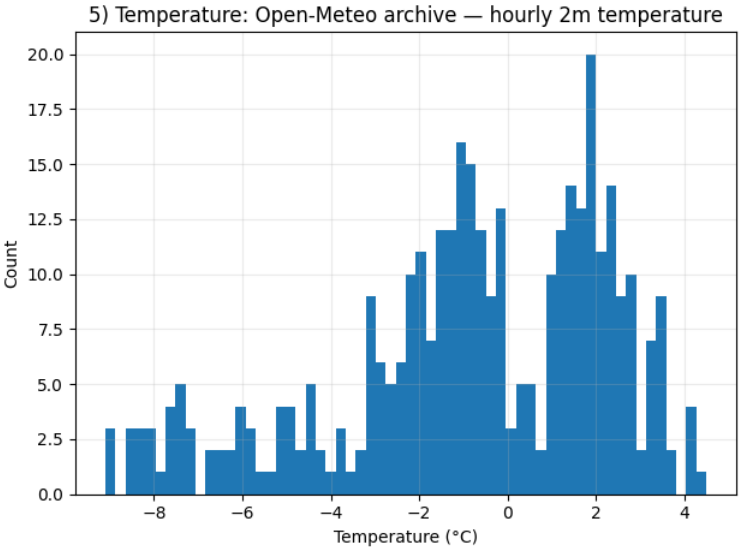
Histograms of amplitudes



Histograms of amplitudes

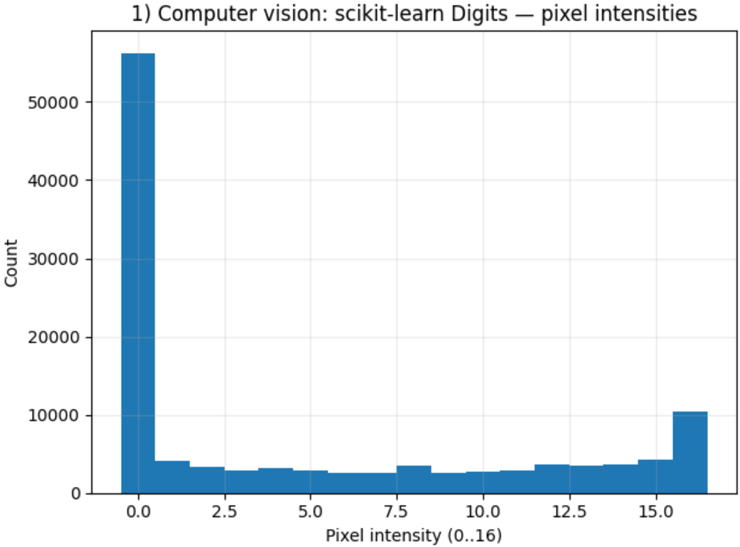


Histograms of amplitudes



Berlin 2025-12-22 to 2026-01-05

Histograms of amplitudes



Questions

- Intuition: we should use finer quantization where the source is more likely to take values.

Questions

- Intuition: we should use finer quantization where the source is more likely to take values.
- **Question 1 (Finite number of levels):** Given a fixed number of quantization levels, how should we choose the decision regions and reconstruction values to minimize distortion?

Questions

- Intuition: we should use finer quantization where the source is more likely to take values.
- **Question 1 (Finite number of levels):** Given a fixed number of quantization levels, how should we choose the decision regions and reconstruction values to minimize distortion?
- **Question 2 (Asymptotic behavior):** As the number of quantization levels becomes very large, how should these reconstruction values be distributed across the real line?

Theoretical Results

- We'll derive necessary conditions for optimal quantizers with a fixed number of levels, and a local search algorithm
- Asymptotically, the density of quantization centroids must be proportional to the density of data raised to power $1/3$

General scalar quantizer model

- A scalar quantizer is specified by
  - decision thresholds $t_0 < t_1 < \cdots < t_M$ with $t_0 = -\infty,\ t_M = +\infty$
  - reconstruction levels (codepoints) $y_0, \ldots, y_{M-1}$
- Mapping rule:

$$Q(x) = y_k \quad \text{if } x \in [t_k, t_{k+1})\,.$$

- Quantization error: $\varepsilon = X - Q(X)$.

Objective: fixed-rate MSE quantizer design

- Given $M = 2^b$ levels and pdf f_X, choose $\{t_k\}, \{y_k\}$ to minimize MSE.
- **Distortion (MSE):**

$$D \triangleq \mathbb{E}\big[(X - Q(X))^2\big] = \sum_{k=0}^{M-1} \int_{t_k}^{t_{k+1}} (x - y_k)^2 f_X(x)\, dx.$$

- **Fixed-rate** means: M is fixed (each symbol uses $\log_2 M$ bits).
- (Later: entropy coding gives variable-rate / expected bits.)

Geometry: Voronoi cells on the real line

- For squared error, the distortion contribution of cell k is

$$D_k = \int_{t_k}^{t_{k+1}} (x - y_k)^2 f_X(x)\, dx.$$

- Think of $\{y_k\}$ as "centers" and thresholds as 1D Voronoi boundaries.
- The design problem is non-convex (many local minima), but has simple necessary conditions.

Lloyd–Max: two necessary conditions (MSE)

- A locally optimal MSE quantizer must satisfy:
  - **(C1) Centroid condition:**

$$y_k = \mathbb{E}[X \mid X \in [t_k, t_{k+1})] = \frac{\int_{t_k}^{t_{k+1}} x f_X(x)\,dx}{\int_{t_k}^{t_{k+1}} f_X(x)\,dx}.$$

  - **(C2) Nearest-neighbor (boundary) condition:**

$$t_k = \frac{y_{k-1} + y_k}{2}, \quad k = 1, \ldots, M-1.$$

Lloyd–Max: two necessary conditions (MSE)

- A locally optimal MSE quantizer must satisfy:
  - **(C1) Centroid condition:**

$$y_k = \mathbb{E}[X \mid X \in [t_k, t_{k+1})] = \frac{\int_{t_k}^{t_{k+1}} x f_X(x)\,dx}{\int_{t_k}^{t_{k+1}} f_X(x)\,dx}.$$

  - **(C2) Nearest-neighbor (boundary) condition:**

$$t_k = \frac{y_{k-1} + y_k}{2}, \quad k = 1, \ldots, M-1.$$

Notes

- The optimal quantizer may not be unique!
- Even if there is only one optimal quantizer, there may be more than one quantizer that satisfies the properties, in which case the best quantizer is one of them. In other words, the optimality properties are necessary for optimality, but not sufficient. Example of two quantizers that satisfy the optimality criteria
- For log-concave densities (logarithm of the density is concave), such as Gaussian, there is a unique quantizer

Outline

- Non-uniform scalar quantizers
- Lloyd–Max optimality conditions
  - Proof of the centroid condition
  - Proof of the boundary condition
- Gaussian and other distributions
  - Exact solution: usually not closed-form
- Lloyd–Max iterative algorithm
- Connection to k-means in 1D
  - Numerical Lloyd–Max designs (examples)
- High-rate analysis for non-uniform quantizers
  - Examples: Gaussian and Laplacian
- Practical non-uniform quantizers
- Implementation details
- High-rate companding construction

Proof idea: optimize y_k with fixed thresholds

Fix the thresholds $\{t_k\}$ and minimize D over $\{y_k\}$.

Because the cells are disjoint, we can minimize each D_k separately:

$$D_k(y_k) = \int_{t_k}^{t_{k+1}} (x - y_k)^2 f_X(x)\, dx.$$

- $D_k(y_k)$ is a convex quadratic function of y_k.

Centroid condition: calculus

Differentiate $D_k(y_k)$ w.r.t. y_k:

$$\frac{\partial D_k}{\partial y_k} = \int_{t_k}^{t_{k+1}} 2(y_k - x) f_X(x)\, dx =$$

$$2\left(y_k \int_{t_k}^{t_{k+1}} f_X(x)\, dx - \int_{t_k}^{t_{k+1}} x f_X(x)\, dx \right).$$

Set derivative to zero:

$$y_k \int_{t_k}^{t_{k+1}} f_X(x)\, dx = \int_{t_k}^{t_{k+1}} x f_X(x)\, dx.$$

Hence

$$\boxed{y_k = \frac{\int_{t_k}^{t_{k+1}} x f_X(x)\, dx}{\int_{t_k}^{t_{k+1}} f_X(x)\, dx} = \mathbb{E}[X \mid X \in [t_k, t_{k+1})].}$$

Outline

- Non-uniform scalar quantizers
- Lloyd–Max optimality conditions
    - Proof of the centroid condition
    - Proof of the boundary condition
- Gaussian and other distributions
    - Exact solution: usually not closed-form
- Lloyd–Max iterative algorithm
- Connection to k-means in 1D
    - Numerical Lloyd–Max designs (examples)
- High-rate analysis for non-uniform quantizers
    - Examples: Gaussian and Laplacian
- Practical non-uniform quantizers
- Implementation details
- High-rate companding construction

Proof idea: optimize t_k with fixed codepoints

Fix the codepoints $\{y_k\}$ and optimize the thresholds.

Only two adjacent cells depend on an interior boundary t_k:

$$D_{k-1} + D_k = \int_{t_{k-1}}^{t_k} (x - y_{k-1})^2 f_X(x)\, dx + \int_{t_k}^{t_{k+1}} (x - y_k)^2 f_X(x)\, dx.$$

Boundary condition: "equal distortion" at the boundary

Differentiate w.r.t. t_k using Leibniz' rule:

$$\frac{\partial}{\partial t_k}(D_{k-1} + D_k) = (t_k - y_{k-1})^2 f_X(t_k) - (t_k - y_k)^2 f_X(t_k).$$

Set derivative to zero (assuming $f_X(t_k) > 0$):

$$(t_k - y_{k-1})^2 = (t_k - y_k)^2.$$

Since $y_{k-1} < y_k$ for an ordered quantizer,

$$\boxed{t_k = \frac{y_{k-1} + y_k}{2}.}$$

- Interpretation: at the boundary, both neighboring reconstructions are equally good.

Summary: Lloyd–Max conditions

- For MSE-optimal scalar quantization, a local optimum satisfies
  - **nearest-neighbor boundaries:** $t_k = (y_{k-1} + y_k)/2$
  - **centroids:** $y_k = \mathbb{E}[X \mid X \in [t_k, t_{k+1})]$
- This suggests an *alternating minimization* (update t then y).
- Global optimality is not guaranteed (non-convex); initialization matters.

Outline

- Non-uniform scalar quantizers
- Lloyd–Max optimality conditions
    - Proof of the centroid condition
    - Proof of the boundary condition
- Gaussian and other distributions
    - Exact solution: usually not closed-form
- Lloyd–Max iterative algorithm
- Connection to k-means in 1D
    - Numerical Lloyd–Max designs (examples)
- High-rate analysis for non-uniform quantizers
    - Examples: Gaussian and Laplacian
- Practical non-uniform quantizers
- Implementation details
- High-rate companding construction

Closed-form designs are rare

- The Lloyd–Max conditions are implicit and coupled.
- For most pdfs and most M, there is no closed-form $\{t_k, y_k\}$.
- Two useful notions of “solution”:
  - **Numerical Lloyd–Max** design for a given pdf and M
  - **High-rate approximation** that gives analytic bin spacing (companding)

Lloyd–Max algorithm (fixed-rate, MSE)

Given M levels and pdf f_X:

1. Initialize reconstruction levels $y_0 < \cdots < y_{M-1}$ (or thresholds).
2. **Boundary update:**
$$t_k \leftarrow \frac{y_{k-1} + y_k}{2}, \quad k = 1, \ldots, M-1.$$
3. **Centroid update:**
$$y_k \leftarrow \frac{\int_{t_k}^{t_{k+1}} x f_X(x)\, dx}{\int_{t_k}^{t_{k+1}} f_X(x)\, dx}, \quad k = 0, \ldots, M-1.$$
4. Repeat until convergence (small change in D, t, or y).

- Each step does not increase distortion D (coordinate descent).
- Converges to a local optimum (stationary point).

Implementation notes

- For analytic pdfs: integrals may be computed in closed form (rare) or numerically.
- For empirical data samples $\{x_i\}_{i=1}^n$:
  - replace integrals by sums (sample means)
  - yields exactly the 1D k-means algorithm
- Initialization matters (multiple restarts can help).
- Constraints are easy to add in practice: clipping / overload region, symmetry, etc.

Empirical quantization objective = 1D k-means

Given samples $x_1, \ldots, x_n$ and M clusters/levels, define assignments $a(i) \in \{0, \ldots, M-1\}$.

The empirical distortion is

$$\hat{D} = \frac{1}{n} \sum_{i=1}^{n} (x_i - y_{a(i)})^2.$$

- This is exactly the k-means algorithm (with M centers) on the line.
- k-means is a popular iterative algorithm for clustering data. Start with random centers, then repeat: Assign each observation to the cluster with the nearest mean center. Recalculate centers for observations assigned to each cluster.

Alternating minimization = Lloyd–Max / k-means

- **Assignment step:** for fixed centers $\{y_k\}$,

$$a(i) \leftarrow \arg\min_k (x_i - y_k)^2,$$

  which in 1D produces contiguous intervals and boundaries $t_k = (y_{k-1} + y_k)/2$.
- **Update step:** for fixed assignments,

$$y_k \leftarrow \frac{1}{|\{i : a(i) = k\}|} \sum_{i:a(i)=k} x_i.$$

- In the population limit ($n \to \infty$), sample means become conditional expectations.

Outline

Non-uniform scalar quantizers

Lloyd–Max optimality conditions

- Proof of the centroid condition
- Proof of the boundary condition

Gaussian and other distributions

- Exact solution: usually not closed-form

Lloyd–Max iterative algorithm

Connection to k-means in 1D

- Numerical Lloyd–Max designs (examples)

High-rate analysis for non-uniform quantizers

- Examples: Gaussian and Laplacian

Practical non-uniform quantizers

Implementation details

High-rate companding construction

Numerical Lloyd–Max: standard normal example

For $X \sim \mathcal{N}(0,1)$ and $b = 3$ bits ($M = 8$), Lloyd–Max yields symmetric thresholds/levels.

k	threshold t_k	level y_k
0	$t_0 = -\infty$	$y_0 \approx -2.1519$
1	$t_1 \approx -1.7479$	$y_1 \approx -1.3439$
2	$t_2 \approx -1.0500$	$y_2 \approx -0.7560$
3	$t_3 \approx -0.5005$	$y_3 \approx -0.2451$
4	$t_4 = 0$	$y_4 \approx 0.2451$
5	$t_5 \approx 0.5005$	$y_5 \approx 0.7560$
6	$t_6 \approx 1.0500$	$y_6 \approx 1.3439$
7	$t_7 \approx 1.7479$	$y_7 \approx 2.1519$
8	$t_8 = +\infty$	

Table: Lloyd–Max thresholds/levels for $\mathcal{N}(0,1)$ with $M = 8$ (rounded).

Numerical Lloyd–Max: unit-variance Laplacian example

For a unit-variance Laplacian and $b = 3$ bits ($M = 8$):

k	threshold t_k	level y_k
0	$t_0 = -\infty$	$y_0 \approx -3.0867$
1	$t_1 \approx -2.3796$	$y_1 \approx -1.6725$
2	$t_2 \approx -1.2527$	$y_2 \approx -0.8330$
3	$t_3 \approx -0.5332$	$y_3 \approx -0.2334$
4	$t_4 = 0$	$y_4 \approx 0.2334$
5	$t_5 \approx 0.5332$	$y_5 \approx 0.8330$
6	$t_6 \approx 1.2527$	$y_6 \approx 1.6725$
7	$t_7 \approx 2.3796$	$y_7 \approx 3.0867$
8	$t_8 = +\infty$	

Table: Lloyd–Max thresholds/levels for a unit-variance Laplacian with $M = 8$ (rounded).

Takeaways

- Both designs are symmetric (because the pdf is symmetric).
- Compared to Gaussian, Laplacian has:
  - tighter inner thresholds (more bins near 0)
  - more spread-out outer reconstruction levels (longer tails)
- Uniform sources yield uniform quantizers.
- In general: **bin widths adapt to the shape of** f_X.

Beyond 1D: vector quantization (preview)

- Quantizing $x \in \mathbb{R}^d$ with a codebook $\{y_k\}_{k=1}^M$ leads to

$$\min \mathbb{E}[\|X - Q(X)\|_2^2], \quad Q(X) \in \{y_1, \dots, y_M\}.$$

- Boundaries become Voronoi cells in $\mathbb{R}^d$.
- Lloyd's algorithm (a.k.a. k-means) generalizes directly.

High-rate distortion for non-uniform scalar quantizers

Assume "locally uniform" cells with a *cell size function* $\Delta(x)$.

Heuristic Bennett-style approximation:

$$D \approx \frac{1}{12} \int \Delta(x)^2 \, f_X(x) \, dx.$$

- Compare to uniform case: $\Delta(x) \equiv \Delta$ gives $D \approx \Delta^2/12$.
- Now we can choose $\Delta(x)$ to trade distortion across x.

Point density and the fixed-rate constraint

Instead of $\Delta(x)$, use *point density* $\lambda(x)$:

$\lambda(x) \triangleq$ (fraction of centroids per unit length near x), $\displaystyle\int_{\mathbb{R}} \lambda(x)\,dx = 1.$

$$\lambda(x)dx = \frac{\text{number of centroids in } [x, x+dx]}{M}$$

For large number of centroids M (high-rate), local cell width satisfies

$$\Delta(x) \approx \frac{1}{M\lambda(x)}.$$

Plugging into the high-rate distortion:

$$D \approx \frac{1}{12M^2} \int \frac{f_X(x)}{\lambda(x)^2}\,dx.$$

Optimal high-rate non-uniform quantizer (sketch proof)

We minimize

$$\min_{\lambda \geq 0} \int \frac{f_X(x)}{\lambda(x)^2}\, dx \quad \text{s.t.} \quad \int \lambda(x)\, dx = 1.$$

Form the Lagrangian

$$\mathcal{L}(\lambda) = \int \left(\frac{f_X(x)}{\lambda(x)^2} + \nu \lambda(x) \right) dx.$$

Stationarity (pointwise):

$$\frac{\partial}{\partial \lambda} \left(\frac{f}{\lambda^2} + \nu \lambda \right) = -\frac{2f}{\lambda^3} + \nu = 0 \quad \Rightarrow \quad \lambda(x)^3 \propto f_X(x).$$

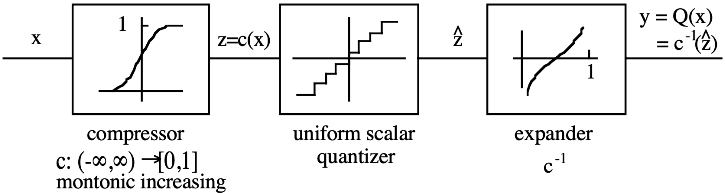
Therefore

$$\boxed{\lambda^\star(x) \propto f_X(x)^{1/3}, \qquad \Delta^\star(x) \propto f_X(x)^{-1/3}.}$$

Interpretation

- The density of quantization centroids must be proportional to the density of data raised to power $1/3$.
- **Companding:** We can find a nonlinear scalar transformation to warp the signal axis so that optimal non-uniform bins in become uniform bins after transformation.
  - practical implication: we can apply the nonlinear transformation and then quantize uniformly

Companding



- One can implement any nonuniform quantizer with a compander.
- Conversely, any compander is the implementation of some quantizer.

Companding interpretation (analytic "solution")

Define a monotone function g (a compressor) with derivative proportional to the centroid density:

$$g'(x) = \lambda^\star(x) = \frac{f_X(x)^{1/3}}{\int f_X(u)^{1/3} du}.$$

Then

$$g(x) = \int_{-\infty}^{x} g'(u)\, du \in [0,1].$$

- Quantize $u = g(x)$ **uniformly** into M bins: $u \in [\frac{k}{M}, \frac{k+1}{M})$.
- Thresholds in x-domain are approximately

$$t_k \approx g^{-1}\left(\frac{k}{M}\right).$$

- This is a principled non-uniform quantizer derived from high-rate analysis.

Why does this work?

- Thresholds $t_k = g^{-1}\left(\frac{k}{M}\right)$, imply $g(t_{k+1}) - g(t_k) = \frac{1}{M}$
- By the mean value theorem, for some $\xi_k \in [t_k, t_{k+1}]$,

$$g(t_{k+1}) - g(t_k) = g'(\xi_k)\,(t_{k+1} - t_k).$$

- Substituting $g'(x) = \lambda^\star(x)$ gives

$$t_{k+1} - t_k \approx \frac{1}{M\lambda^\star(\xi_k)}.$$

- hence the induced cell width satisfies

$$\boxed{\Delta(x) \approx \frac{1}{M\lambda^\star(x)}}$$

which is exactly the spacing required by high-rate optimality.

Outline

- Non-uniform scalar quantizers
- Lloyd–Max optimality conditions
  - Proof of the centroid condition
  - Proof of the boundary condition
- Gaussian and other distributions
  - Exact solution: usually not closed-form
- Lloyd–Max iterative algorithm
- Connection to k-means in 1D
  - Numerical Lloyd–Max designs (examples)
- High-rate analysis for non-uniform quantizers
  - Examples: Gaussian and Laplacian
- Practical non-uniform quantizers
- Implementation details
- High-rate companding construction

Example: Gaussian pdf $X \sim \mathcal{N}(0,1)$

For $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$,

$$f(x)^{1/3} \propto e^{-x^2/6}.$$

Hence the optimal high-rate compressor is

$$g(x) \propto \int_{-\infty}^{x} e^{-u^2/6}\,du \;=\; \tfrac{1}{2}\Big(1 + \operatorname{erf}(x/\sqrt{6})\Big) \;=\; \Phi\left(\frac{x}{\sqrt{3}}\right).$$

- High-rate thresholds: $t_k \approx g^{-1}(k/M) = \sqrt{3}\,\Phi^{-1}(k/M)$.
- Interpretation: compared to equal-probability bins (Φ^{-1}), this is "less aggressive" near the tails.

Example: Laplacian pdf (unit variance)

For a unit-variance Laplacian,

$$f(x) = \frac{\sqrt{2}}{2} e^{-\sqrt{2}|x|}.$$

Then

$$f(x)^{1/3} \propto e^{-\beta|x|}, \qquad \beta = \frac{\sqrt{2}}{3}.$$

The optimal high-rate compressor (normalized to $[0,1]$) becomes

$$g(x) = \begin{cases} \frac{1}{2} e^{\beta x}, & x \leq 0, \\ 1 - \frac{1}{2} e^{-\beta x}, & x \geq 0. \end{cases}$$

- High-rate thresholds follow $t_k \approx g^{-1}(k/M)$.
- Compared to Gaussian, Laplacian puts even *more* resolution near 0.

Sanity check: uniform pdf (unit variance)

If $X \sim \mathrm{Unif}[-\sqrt{3}, \sqrt{3}]$ then $f(x)$ is constant on its support.

High-rate result gives

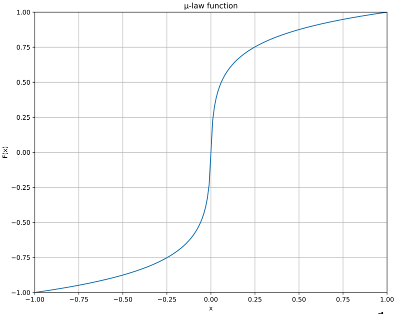
$$\lambda^{\star}(x) \propto f(x)^{1/3} = \text{constant} \quad \Rightarrow \quad \Delta^{\star}(x) = \text{constant}.$$

- For uniform sources, the optimal quantizer is (approximately) uniform.

Companding in practice: μ-law (speech telephony)

- Goal: allocate more resolution near small amplitudes.
- Compressor for $|x| \leq X_{\max}$:

$$g(x) = \operatorname{sign}(x) \, \frac{\ln\left(1 + \mu|x|/X_{\max}\right)}{\ln(1+\mu)}.$$



- Quantize $g(x)$ uniformly, then expand via g^{-1}.
- Large μ $\Rightarrow$ more non-uniformity (stronger compression).
- $\mu = 255$ in the North American and Japanese standards

A-law and piecewise companders

- A-law (common in Europe) is a piecewise log/linear compander.
- Motivation: approximate an “optimal” non-uniform quantizer with simple hardware.
- General design pattern:
  - choose a monotone compressor g (log-like, power-law, learned)
  - uniform quantization in the compressed domain
  - optional entropy coding on the indices
- μ-law (used in North America) offers better dynamic range but worse distortion for small signals, while A-law (used in Europe) provides more resolution for low-level signals, making it better for international calls

Dead-zone quantizers (transform coding)

- Transform coefficients (DCT/MDCT/wavelets) are often strongly peaked at 0.
- A **dead-zone** quantizer uses a larger bin around 0:
  - encourages many zeros $\Rightarrow$ compressible symbol stream
  - matches perceptual metrics (small coefficients are less important)
- Canonical in JPEG (DCT $+$ quantization matrices) and audio codecs.

Log and floating-like quantization

- Some applications care about *relative* error: $|\varepsilon|/|x|$.
- Log quantization (for $x > 0$):

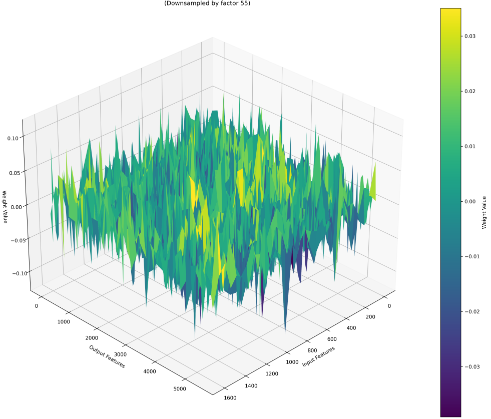
$$Q(x) = \exp\left(Q_{\mathrm{uni}}(\ln x)\right)$$

  - roughly constant relative error
  - connects to floating point and block floating point formats
- Used in: audio amplitude coding, dynamic range compression, some ML activation quantizers.

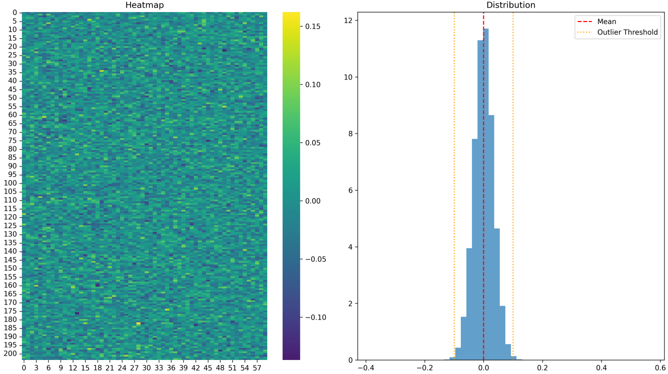
Non-uniform quantization in modern ML systems (preview)

- Non-uniformity appears in several ways:
  - per-channel scales (different Δ per channel)
  - learned codebooks (vector quantization / product quantization)
  - clipping + non-linear companders to handle outliers
- Many methods can be interpreted as "match the quantizer to the data distribution".

Visualization of LLM Weights: Hymba (Nvidia) Nov 2024


Weight Matrix 3D Surface - model.layers.26.moe.experts.0.up_proj
(Downsampled by factor 55)

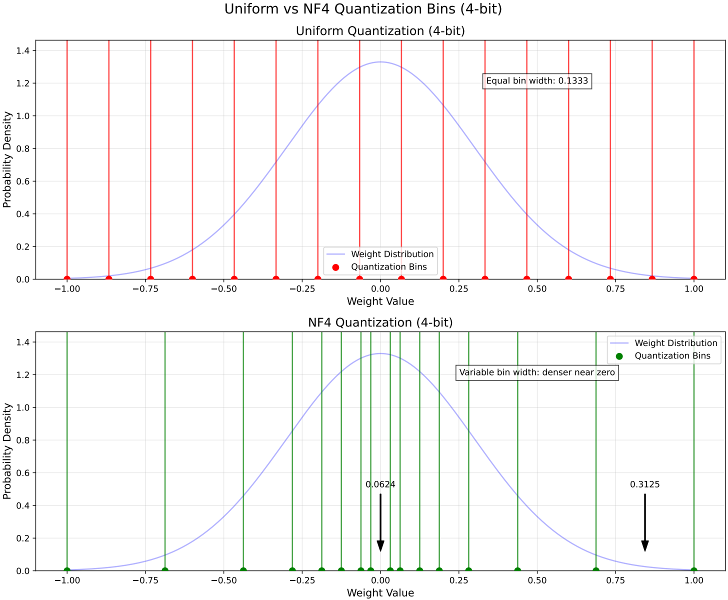
Layer Analysis: model.layers.26.moe.experts.0.up_proj

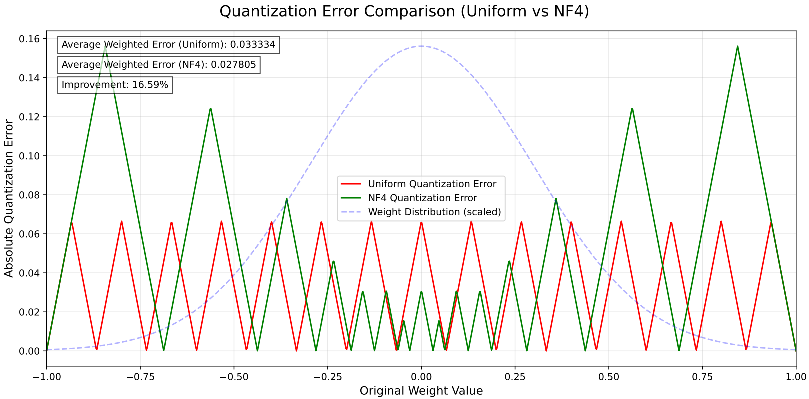


```
Shape: (5504, 1600)
Mean: -0.000
Std: 0.033
Min: -0.381
Max: 0.566
Median: -0.000
Q1: -0.022
Q3: 0.022
1st percentile: -0.079
99th percentile: 0.078
Outliers: 29776
Outlier %: 0.34%
```

NormalFloat NF4 (Dettmers et al. 2023)

- NF4 is a 4-bit quantization format employing non-uniform quantization across 16 discrete levels.
- It better approximates the weight distribution—typically near a normal distribution—by allocating more quantization levels near zero.
- This non-uniform allocation minimizes quantization error, especially for small but critical weight values.
- NF4 is widely used in QLoRA for efficient fine-tuning of large language models.


Uniform vs NF4 Quantization Bins (4-bit)
Uniform Quantization (4-bit)
Probability Density
Weight Value
Equal bin width: 0.1333
Weight Distribution
Quantization Bins
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
NF4 Quantization (4-bit)
Variable bin width: denser near zero
0.0624
0.3125


Quantization Error Comparison (Uniform vs NF4)
Average Weighted Error (Uniform): 0.033334
Average Weighted Error (NF4): 0.027805
Improvement: 16.59%
Uniform Quantization Error
NF4 Quantization Error
Weight Distribution (scaled)
Absolute Quantization Error
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Original Weight Value

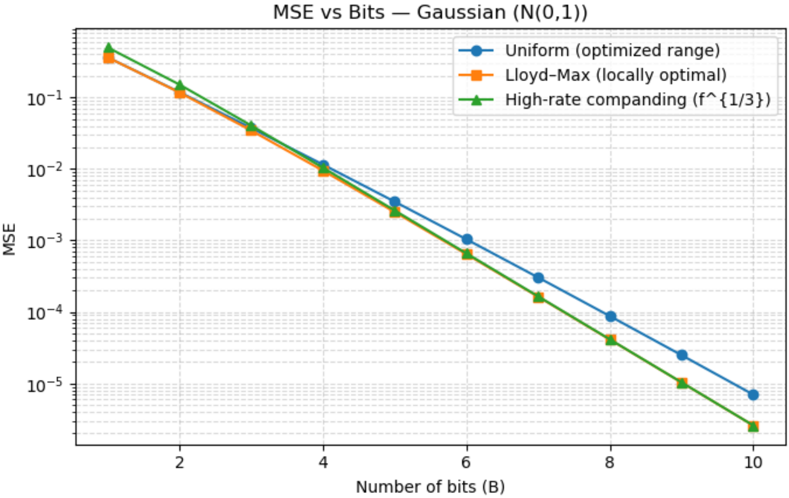
Key takeaways

- Uniform quantization is rarely optimal when f_X is non-uniform.
- MSE-optimal scalar quantizers satisfy Lloyd–Max conditions:
    - boundaries are midpoints between adjacent codepoints
    - codepoints are conditional means (centroids)
- High-rate theory gives analytic guidance:

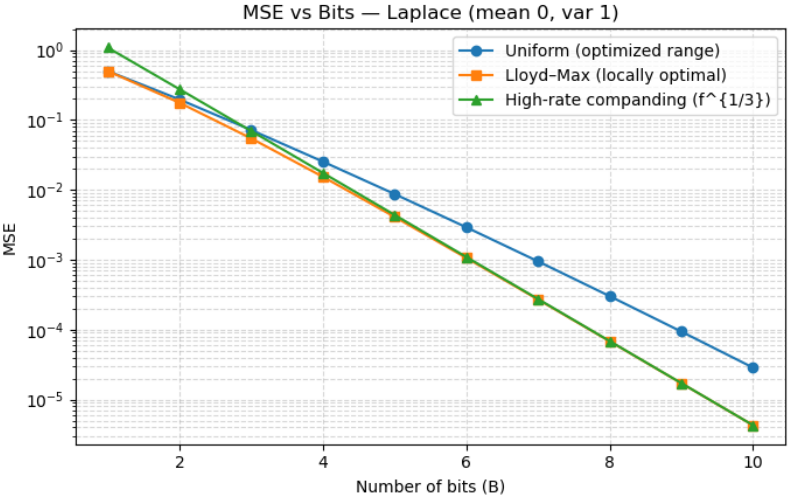
$$\lambda^\star(x) \propto f_X(x)^{1/3}, \quad \Delta^\star(x) \propto f_X(x)^{-1/3}.$$

- Lloyd–Max is alternating minimization $=$ 1D k-means.
- Practical designs: companding (μ-law), dead-zone quantizers, log quantization, NormalFloat

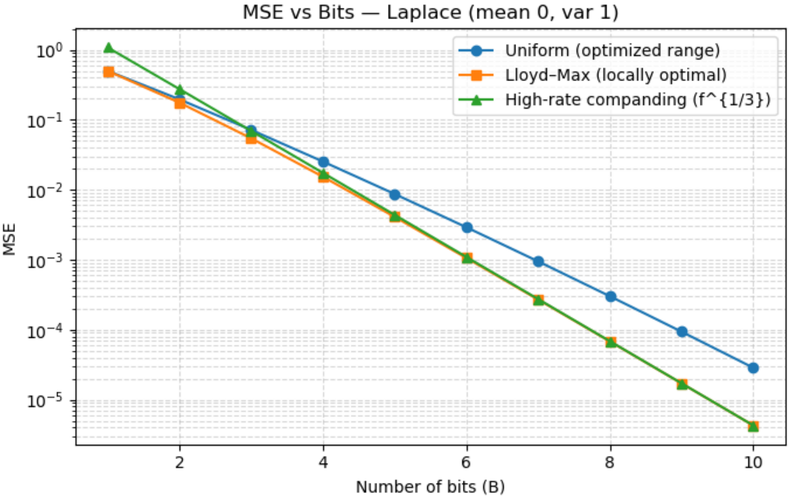
Numerical Comparisons



Numerical Comparisons



Numerical Comparisons



Implementation details

- Fixed-rate scalar quantization: b bits $\Rightarrow M = 2^b$ levels.
- In the plots, we compared three *constructive* designs:
  - **(A) Optimized uniform (with clipping)**: choose a range parameter $X_{\max}$ to minimize MSE.
  - **(B) Lloyd–Max (nonuniform)**: alternate boundary $+$ centroid updates until convergence.
  - **(C) High-rate compander (analytic approx)**: derive $\lambda^\star(x) \propto f_X(x)^{1/3}$, build g, then set thresholds via g^{-1}.
- We can use (C) as a **good initialization** for (B).

(A) Optimized uniform quantizer: 1D search over $X_{\max}$

- Uniform quantizer with design range $[-X_{\max}, X_{\max}]$:

$$\Delta = \frac{2X_{\max}}{M}, \qquad y_k = -X_{\max}+\left(k+\tfrac{1}{2}\right)\Delta, \quad k = 0, \ldots, M-1,$$

$$t_k = -X_{\max}+k\Delta, \quad k = 1, \ldots, M-1, \qquad t_0 = -\infty,\ t_M = +\infty.$$

- Define the MSE as a scalar function of $X_{\max}$:

$$D(X_{\max}) \triangleq \mathbb{E}\left[(X-Q_{X_{\max}}(X))^2\right] = \sum_{k=0}^{M-1} \int_{t_k}^{t_{k+1}} (x-y_k)^2 f_X(x)\, dx.$$

- Optimize the clipping range:

$$X_{\max}^{\star} = \arg\min_{X_{\max}>0} D(X_{\max}).$$

- Implementation:
  - Evaluate $D(X_{\max})$ by analytic integrals or sample based approximation.
  - Use a 1D optimization (grid search / golden-section, e.g. `minimize_scalar` of scipy).

(B) Lloyd–Max (nonuniform) quantizer: alternating minimization

- Input: f_X (or samples), number of levels M.
- Initialize ordered codepoints $y_0 < \cdots < y_{M-1}$ (uniform spacing, or compander init).
- Repeat until convergence:
    - **Boundary update (nearest-neighbor):**

      $$t_k \leftarrow \frac{y_{k-1} + y_k}{2}, \quad k = 1, \dots, M-1, \quad t_0 = -\infty,\ t_M = +\infty.$$

    - **Centroid update:**

      $$y_k \leftarrow \mathbb{E}[X \mid X \in [t_k, t_{k+1})] = \frac{\int_{t_k}^{t_{k+1}} x f_X(x)\, dx}{\int_{t_k}^{t_{k+1}} f_X(x)\, dx}.$$

- Facts: each step does not increase D (coordinate descent) ⇒ convergence to a *local* optimum.
- Sample-based version is exactly 1D k-means (assignment + sample means).

(C) High-rate compander: how to get g and the thresholds

- Optimize over point densities $\lambda(x)$:

$$\lambda^\star(x) = \frac{f_X(x)^{1/3}}{\int f_X(u)^{1/3}\,du}, \qquad \Delta^\star(x) \propto f_X(x)^{-1/3}.$$

- Build the compressor as the integral of $\lambda^\star$:

$$g(x) = \int_{-\infty}^{x} \lambda^\star(u)\,du \in [0,1] \quad \Rightarrow \quad g'(x) = \lambda^\star(x).$$

- Uniformly quantize $u = g(x)$ into M bins, then map back:

$$t_k \approx g^{-1}\Big(\frac{k}{M}\Big),\ k = 0,\dots,M, \qquad y_k \approx g^{-1}\Big(\frac{k+\frac{1}{2}}{M}\Big),\ k = 0,\dots,M$$

- Practical refinement: do a few Lloyd–Max iterations starting from these $\{t_k, y_k\}$.

Closed-form g^{-1} for Gaussian and Laplace (unit variance)

- **Gaussian $X \sim \mathcal{N}(0,1)$:**

  $$f_X(x)^{1/3} \propto e^{-x^2/6} \;\Rightarrow\; \lambda^\star(x) = \mathcal{N}(0,3).$$

  Using Gaussian CDF

  $$g(x) = \Phi\Big(\frac{x}{\sqrt{3}}\Big), \qquad g^{-1}(u) = \sqrt{3}\,\Phi^{-1}(u).$$

- **Laplace (unit variance):** $f_X(x) = \frac{1}{2s}e^{-|x|/s}$ with $s = 1/\sqrt{2}$. Then $f_X(x)^{1/3} \propto e^{-|x|/(3s)}$, so $\lambda^\star$ is Laplace with scale $3s$, and

  $$g^{-1}(u) = \begin{cases} (3s)\ln(2u), & 0 < u < \frac{1}{2}, \\ -(3s)\ln(2(1-u)), & \frac{1}{2} \le u < 1. \end{cases}$$

- **Uniform source:** f_X constant on its support ⇒ $\lambda^\star$ constant ⇒ uniform quantizer is already optimal.

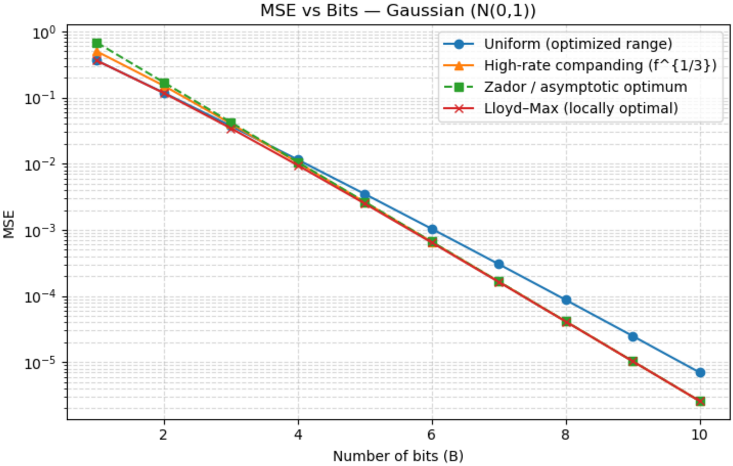
Zador's Formula for Asymptotic MSE

- In the high-rate regime, we have seen that $D(\lambda) \approx \frac{1}{12L^2} \int \frac{f(x)}{\lambda(x)^2} dx$ where $L = 2^b$ and $\lambda^\star(x) \propto f_X(x)^{1/3}$ therefore

$$\boxed{D^\star(b) \approx \frac{2^{-2b}}{12} \Big(\int f^{1/3}(x) dx \Big)^3 .}$$

- Zador's formula relating b (bits) and quantization MSE

Asymptotic Formula and Empirical MSE



References

- Least squares quantization in PCM, S. Lloyd, 1982
- Spectra of quantized signals, W.R. Bennett, 1948
- Quantization distortion in pulse-count modulation with nonuniform spacing of levels, P.F. Panter, W. Dite, 2006
- Vector quantization and signal compression, A. Gersho, R.M. Gray, 2012