EE269
Signal Processing and Quantization for Machine Learning

Quantization Noise

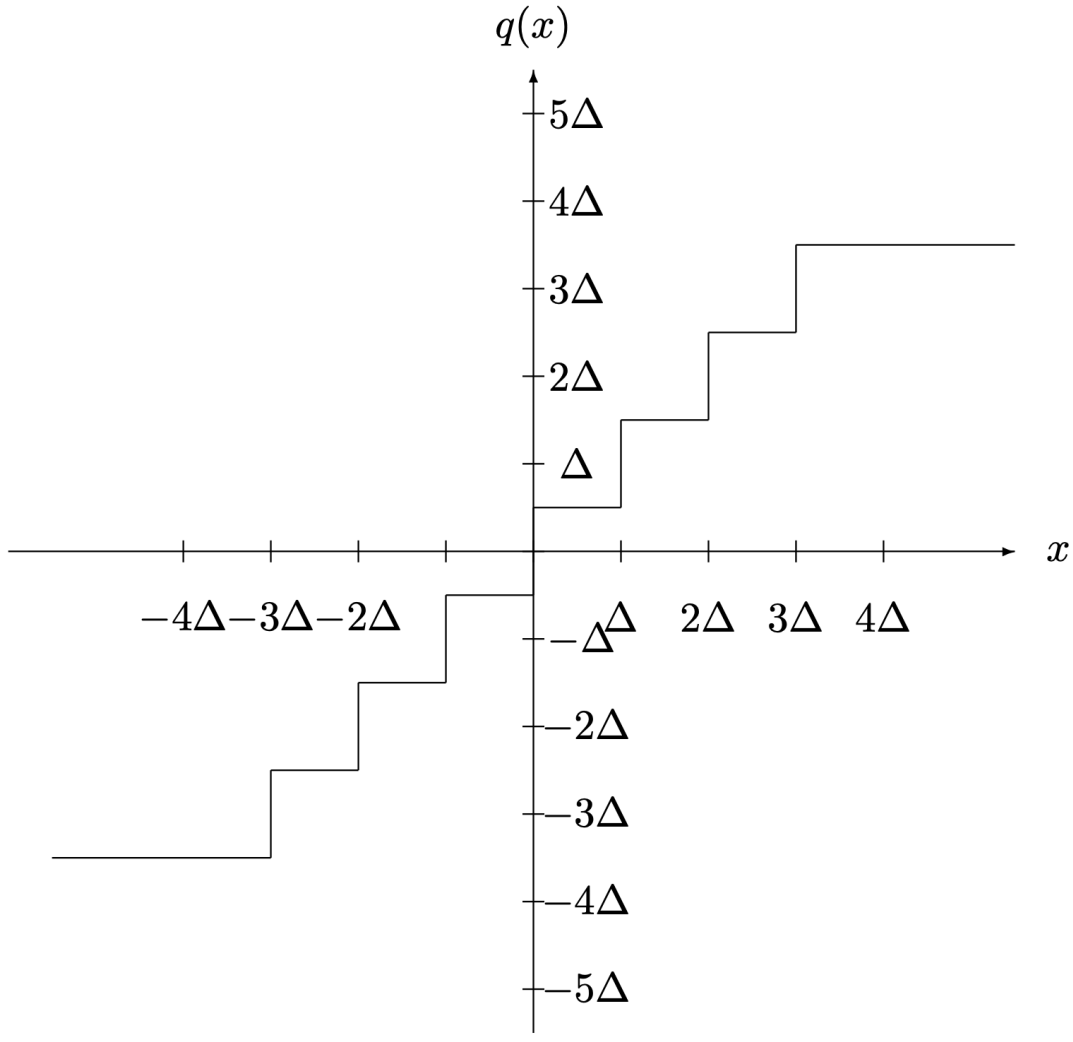
Instructor : Mert Pilanci

Stanford University

Outline

- uniform quantizer
- saturation
- asymptotic analysis (Bennett's Theorem)
- statistical properties
- 6dB rule
- practical takeaways

Uniform Quantizer

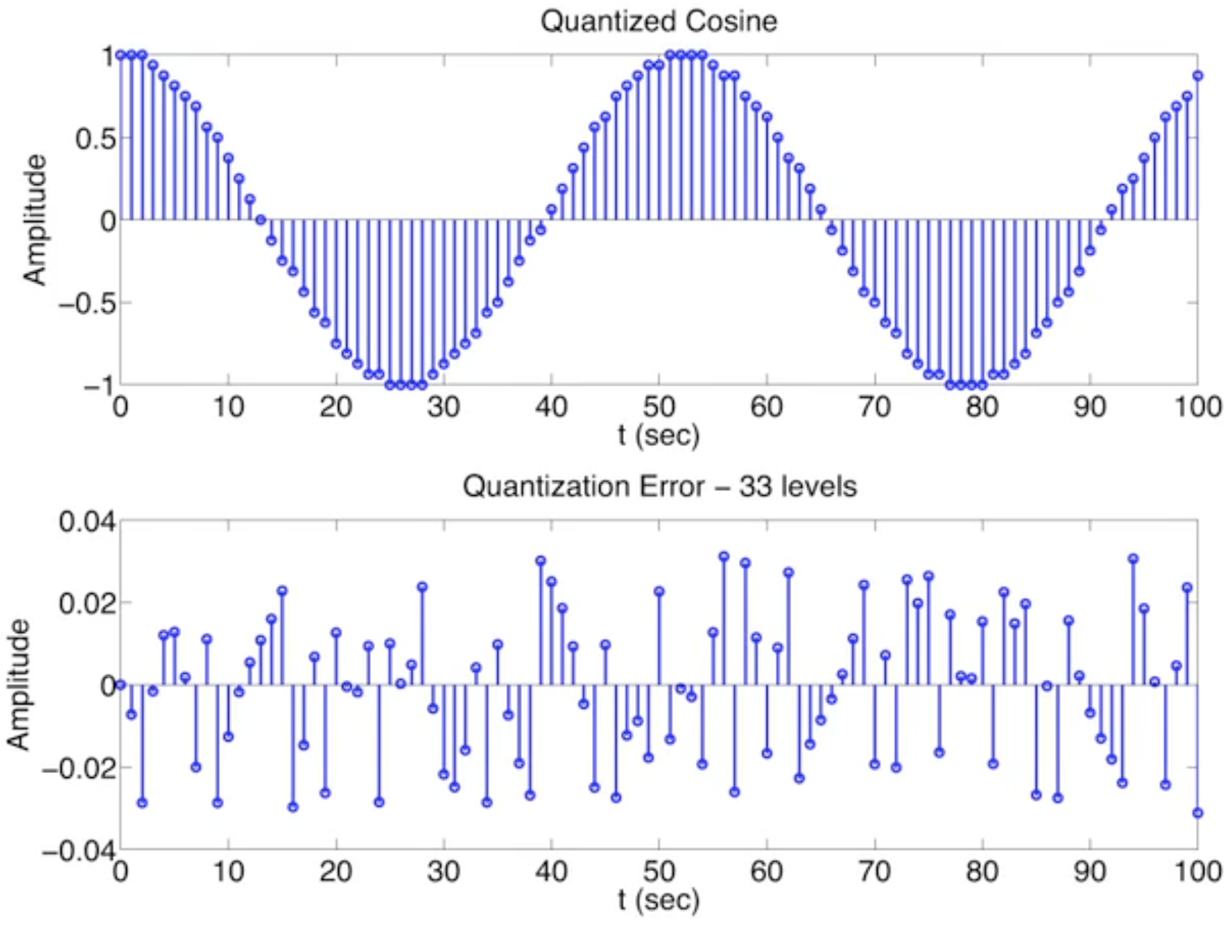


slide credit: B. Gray

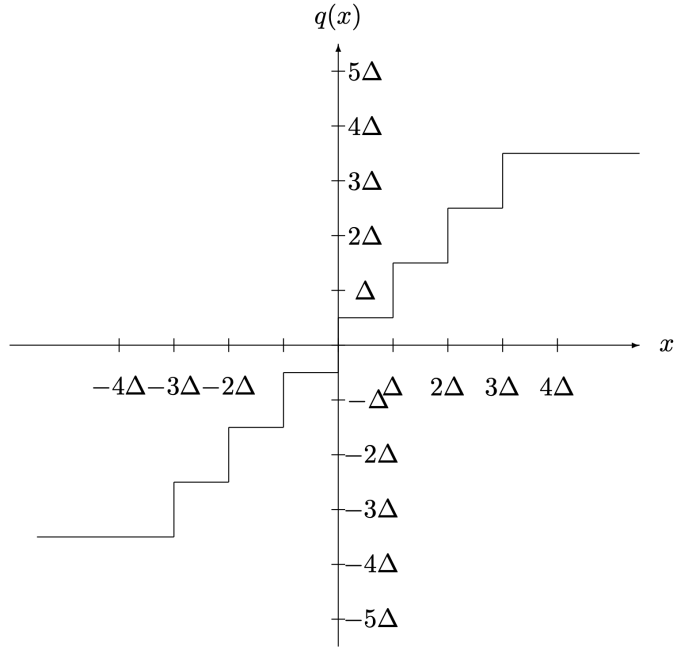
Quantizer mapping $q(x)$

$$q(x) = \begin{cases} \text{quantization levels} & \text{quantization cells} \\ y_{M-1} = (\frac{M}{2} - \frac{1}{2})\Delta; & u \in R_{M-1} = [(\frac{M}{2} - 1)\Delta, \infty) \\ y_k = (-\frac{M}{2} + k + \frac{1}{2})\Delta; & u \in R_k = \\ & [(-\frac{M}{2} + k)\Delta, (-\frac{M}{2} + k + 1)\Delta) \\ & k = 0, \cdots, \frac{M}{2} - 1 \\ y_0 = (-\frac{M}{2} + \frac{1}{2})\Delta; & u \in R_0 = (-\infty, (-\frac{M}{2} + 1)\Delta) \end{cases}$$

Quantized Cosine



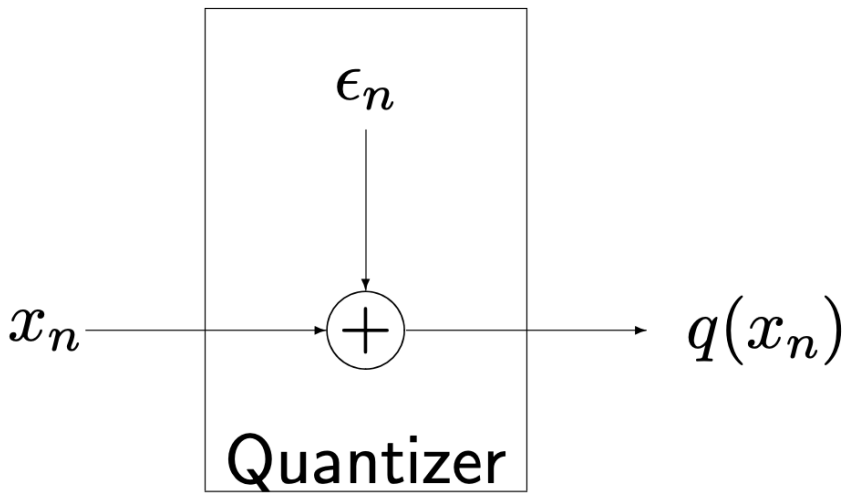
Saturation and Error



- **no-overload range:** input $\in [-\frac{M}{2}\Delta, \frac{M}{2}\Delta]$
  - an input within bin is mapped to the midpoint of the bin
  - error $\leq \frac{\Delta}{2}$
- **overload range:** input $\notin [-\frac{M}{2}\Delta, \frac{M}{2}\Delta]$
  - error is greater than $\frac{\Delta}{2}$

Quantization Noise

- $\epsilon \triangleq x - q(x)$ is the quantizer error
- given an input sequence x_n, we have a corresponding quantization noise sequence ϵ_n
- additive noise model



Analysis of Quantization Noise

two assumptions used in system analysis and design

- (A1) ϵ_n is uniformly distributed in $[-\Delta/2, \Delta/2]$
- (A2) ϵ_n is uncorrelated, i.e., $\mathbb{E}[\epsilon_i \epsilon_j] = 0$ for $i \neq j$

Analysis of Quantization Noise

two assumptions used in system analysis and design

- (A1) ϵ_n is uniformly distributed in $[-\Delta/2, \Delta/2]$
- (A2) ϵ_n is uncorrelated, i.e., $\mathbb{E}[\epsilon_i \epsilon_j] = 0$ for $i \neq j$

these assumptions are usually false!

Asymptotic Analysis

- it can be shown that as the number of bins grow to infinity and Δ goes to zero, these assumptions are correct.

  Bennett's Theorem: Suppose that

  1. the input is in no-overload region
  2. M (number of bins) is asymptotically large
  3. Δ is asymptotically small
  4. probability density of the signal is smooth

  then, assumptions A1 and A2 hold.

Consequences

- since ϵ_n is uniformly distributed in $[-\Delta/2, \Delta/2]$
- quantization error is zero mean

$$\mathbb{E}[\epsilon_n] = \mathbb{E}[q(x_n) - x_{(n)}] = 0$$

- variance

$$\mathbf{Var}[\epsilon_n] = \mathbb{E}[\epsilon_n^2] = \frac{1}{\Delta} \int_{-\Delta/2}^{\Delta/2} e^2 \, de = \frac{\Delta^2}{12}$$

Signal-to-quantization-noise ratio (SQNR)

- define SQNR$\triangleq \frac{\mathbb{E}[x^2]}{\mathbb{E}[\epsilon^2]}$
- suppose signal variance is σ^2, then SQNR$\approx \frac{\sigma^2}{\Delta^2/12}$
- in dB:

$$\mathrm{SQNR}_{dB} \approx 10 \log_{10} \left(\frac{12\sigma^2}{\Delta^2} \right)$$

- halve $\Delta \implies$ SQNR improves by $20 \log_{10} 2 \approx 6.02 dB$
- uniform quantizer with total dynamic range $2A$ has error $\Delta = \frac{2A}{2^b}$ where b is the number of bits
- therefore, each bit increases SQNR by ≈ 6.02 dB

Bits	SQNR (dB)	Typical application / perceptual quality
4	≈ 25	Severe audible distortion
8	≈ 50	Telephone PCM, AM radio quality
12	≈ 74	High-quality speech, FM radio, early digital audio
16	≈ 98	CD-quality audio, professional distribution
24	≈ 146	Studio recording

Table: Approximate signal-to-quantization-noise ratio (SQNR) versus bit-depth for uniform PCM quantization, assuming full-scale utilization and no overload. Each additional bit improves SQNR by approximately 6 dB.

Some compression standards have variable bits

- MP3, AAC, MP4 do not have a fixed bit-depth. They use adaptive, psychoacoustically driven quantization, where the effective number of bits varies in time, frequency, and channel.
- FLAC, WAV, AIFF preserve a true bit-depth.

Quantization in Images

- grayscale image quantization
  - 6 dB/bit $\implies$ 8 bits $\approx$ 48 dB SQNR
  - human vision saturates around this level for uniform noise
- PNG/JPEG use 8 bits per channel
- 10-bit and 12-bit exist only for HDR or professional post-processing

Proof of Bennett's Theorem

Setup. Let X_n be a continuous random variable with pdf $f_{X_n}(x)$. Consider a *uniform midrise quantizer* with step size Δ and quantization regions

$$R_k = \left[-\frac{M\Delta}{2} + k\Delta,\ -\frac{M\Delta}{2} + (k+1)\Delta \right), \quad k = 0, 1, \ldots, M-1.$$

Define the quantization error

$$\varepsilon_n = X_n - Q(X_n), \quad \varepsilon_n \in \left(-\frac{\Delta}{2}, \frac{\Delta}{2} \right).$$

We study the marginal distribution of ε_n.

CDF Decomposition

Define the cumulative distribution function

$$F_{\varepsilon_n}(\alpha) = \Pr(\varepsilon_n \leq \alpha), \quad \alpha \in \left(-\frac{\Delta}{2}, \frac{\Delta}{2}\right).$$

By conditioning on the quantization regions,

$$\Pr(\varepsilon_n \leq \alpha) = \sum_{k=0}^{M-1} \Pr(\varepsilon_n \leq \alpha \ \cap \ X_n \in R_k).$$

Integral Approximation

For a fixed region R_k, the event $\{\varepsilon_n \leq \alpha\}$ corresponds to

$$X_n \in \left[-\frac{M\Delta}{2} + k\Delta,\ -\frac{M\Delta}{2} + k\Delta + \alpha \right).$$

Hence,

$$\Pr(\varepsilon_n \leq \alpha \cap X_n \in R_k) = \int_{-\frac{M\Delta}{2}+k\Delta}^{-\frac{M\Delta}{2}+k\Delta+\alpha} f_{X_n}(\beta)\, d\beta.$$

Assuming f_{X_n} is smooth relative to Δ, the mean value theorem yields

$$\Pr(\varepsilon_n \leq \alpha \cap X_n \in R_k) \approx f_{X_n}(y_k)\, \alpha,$$

for some $y_k \in R_k$.

Riemann Sum Argument

Summing over all quantization cells,

$$\Pr(\varepsilon_n \leq \alpha) \approx \alpha \sum_{k=0}^{M-1} f_{X_n}(y_k).$$

Multiply and divide by Δ:

$$\Pr(\varepsilon_n \leq \alpha) \approx \frac{\alpha}{\Delta} \sum_{k=0}^{M-1} f_{X_n}(y_k)\,\Delta.$$

The sum is a Riemann approximation of

$$\int f_{X_n}(x)\,dx = 1.$$

Therefore,

$$\Pr(\varepsilon_n \leq \alpha) \approx \frac{\alpha}{\Delta}.$$

Resulting PDF

Differentiating the CDF,

$$f_{\varepsilon_n}(\alpha) = \frac{d}{d\alpha} F_{\varepsilon_n}(\alpha) \approx \frac{1}{\Delta}, \quad \alpha \in \left(-\frac{\Delta}{2}, \frac{\Delta}{2}\right).$$

Conclusion (Bennett's Theorem, heuristic): The quantization error is approximately

$$\varepsilon_n \sim \mathcal{U}\left(-\frac{\Delta}{2}, \frac{\Delta}{2}\right),$$

independent of the input signal.