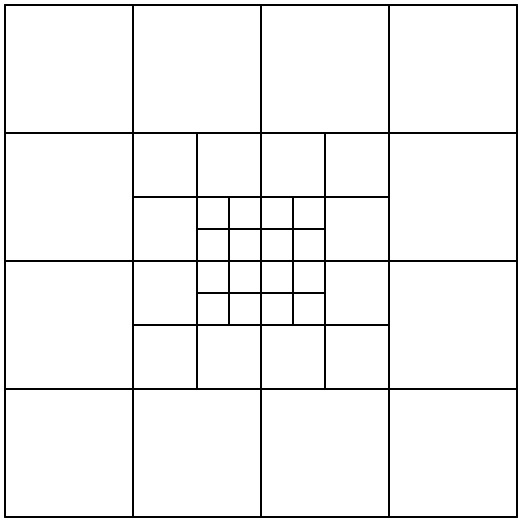
Quantization

- **Prof. Brian L. Evans**
- **Dept. of Electrical and Computer Engineering**
- **The University of Texas at Austin**

Lecture 8

Resolution

- **Human eyes**
  - Sample received light on 2-D grid
  - Photoreceptor density in retina falls off exponentially away from fovea (point of focus)
  - Respond logarithmically to intensity (amplitude) of light



Foveation grid: point of focus in the middle

- **Human ears**
  - Respond to frequencies in 20 Hz to 20 kHz range
  - Respond logarithmically in both intensity (amplitude) of sound (pressure waves) and frequency (octaves)
  - Log-log plot for hearing response vs. frequency

Types of Quantizers

- **Quantization is *an interpretation of a continuous quantity by a finite set of discrete values***
- **Amplitude quantization approximates its input by a discrete amplitude taken from finite set of values**

System Property	*Amplitude Quantizer*	*Sampler*	*Sampler + Quantizer*
Linearity			
Time-invariance			
Causality			
Memoryless			

For the sampler, stay in continuous time domain at input and output to decide on time invariance

Public Switched Telephone Network

- **Sample voice signals at 8000 samples/s**
- **Quantize voice to 8 bits/sample**

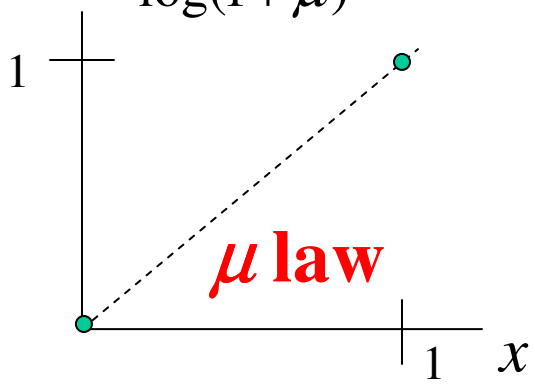
Maximum data rate? → ____ kbps

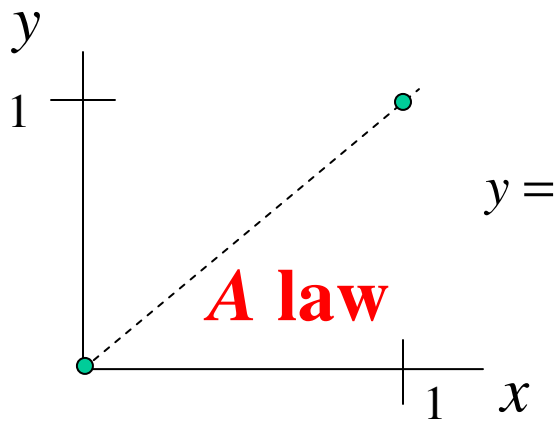
Uniformly quantize to 8 bits/sample, or

Companding by uniformly quantizing to 12 bits and map 12 bits logarithmically to 8 bits (by lookup table) to allocate more bits in quiet segments (where ear is more sensitive)

$\mu = 256$ in US/Japan and $A = 87.6$ in Europe

$$y = \frac{\log(1+\mu|x|)}{\log(1+\mu)}$$



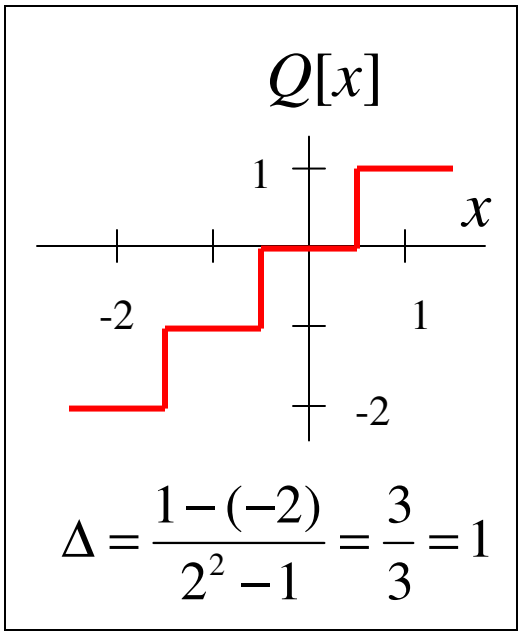


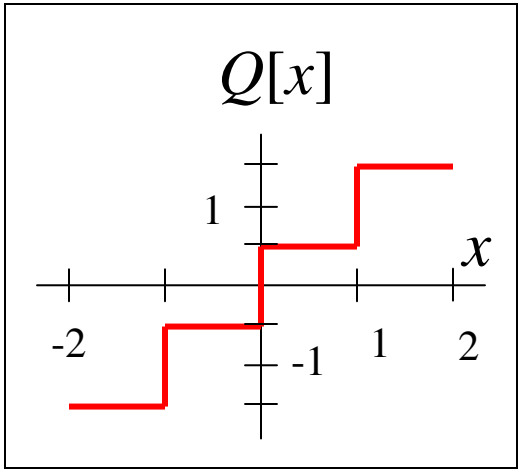
$$y = \begin{cases} \dfrac{A|x|}{1+\log A} & \text{if } 0 \le |x| \le \dfrac{1}{A} \\ \dfrac{1+\log A|x|}{1+\log A} & \text{if } \dfrac{1}{A} \le |x| \le 1 \end{cases}$$

Uniform Quantization

- **Round to nearest integer (midtread)**
  - Quantize amplitude to levels {-2, -1, 0, 1}
  - Step size Δ for linear region of operation
  - Represent levels by {00, 01, 10, 11} or {10, 11, 00, 01} …
  - Latter is two's complement representation
- **Rounding with offset (midrise)**
  - Quantize to levels {-3/2, -1/2, 1/2, 3/2}
  - Represent levels by {11, 10, 00, 01} …
  - Step size $\Delta = \dfrac{\frac{3}{2} - \left(-\frac{3}{2}\right)}{2^2 - 1} = \dfrac{3}{3} = 1$

Used in slide 8-10





Handling Overflow

- **Example: Consider set of integers {-2, -1, 0, 1}**

  Represented in two's complement system {10, 11, 00, 01}.

  Add (–1) + (–1) + (–1) + 1 + 1

  Intermediate computations are – 2, 1, –2, –1 for wraparound arithmetic and –2, –2, –1, 0 for saturation arithmetic

- **Saturation: *When to use it?***

  If input value greater than maximum, set it to maximum; if less than minimum, set it to minimum

  Used in quantizers, filtering, other signal processing operators

Native support in MMX and DSPs

- **Wraparound: *When to use it?***

  Addition performed modulo set of integers

  Used in address calculations, array indexing

Standard two's complement behavior

Audio Compact Discs (CDs)

- **Sampled at 44.1 kHz**

  Analog signal bandwidth of 20 kHz

  Analog bandwidth from 20 kHz to 22.05 kHz is for anti-aliasing filter to rolloff from passband to stopband (10% of maximum passband frequency)

- **Amplitude is uniformly quantized to B = 16 bits to yield dynamic range (signal-to-noise ratio) of**

  $$1.76 \text{ dB} + 6.02 \text{ dB/bit} * B = 98.08 \text{ dB}$$

  This loose upper bound is derived later in slides 8-11 to 8-15

  In practice, audio CDs have dynamic range of about 95 dB

- **Dynamic range helps set filter design specifications**

Dynamic Range in Audio

- **Sound Pressure Level (SPL)**

  Reference in dB SPL is 20 μPa (threshold of hearing)

  Typical living room has 40 dB SPL of noise

  Sound intensity of 120 dB SPL is threshold of pain

  Dynamic range is 80 dB SPL, which audio CDs far exceed

- **In linear systems, SNR = dynamic range**

  (a) Find maximum RMS output of the system with some specified amount of distortion, typically 1%

  (b) Find RMS output of system with small input signal (e.g. -60 dB of full scale) with input signal removed from output

  (c) Divide (b) into (a) to find the dynamic range

Contribution by Dr. Thomas D. Kite, Audio Precision

Digital vs. Analog Audio

- **An audio engineer claims to notice differences between analog vinyl master recording and the remixed CD version. *Is this possible?***

  When digitizing an analog recording, the maximum voltage level for the quantizer is the maximum volume in the track

  Samples are uniformly quantized (to 2^{16} levels in this case although early CDs circa 1982 were recorded at 14 bits)

  Problem on a track with both loud and quiet portions, which occurs often in classical pieces

  When track is quiet, relative error in quantizing samples grows

  Contrast this with analog media such as vinyl which responds linearly to quiet portions

Digital vs. Analog Audio

- **Analog and digital media response to voltage *v***

$$A(v) = \begin{cases} V_0 + (v - V_0)^{1/3} & \text{for } v > V_0 \\ v & \text{for } -V_0 \le v \le V_0 \\ -V_0 - (V_0 - v)^{1/3} & \text{for } v < -V_0 \end{cases} \qquad D(v) = \begin{cases} V_0 & \text{for } v > V_0 \\ v & \text{for } -V_0 \le v \le V_0 \\ -V_0 & \text{for } v < -V_0 \end{cases}$$

- **For a large dynamic range**

  Analog media: records voltages above V_0 with distortion

  Digital media: clips voltages above V_0 to V_0

- **Audio CDs use delta-sigma modulation**

  Effective dynamic range of 19 bits over lower frequencies but lower than 16 bits for higher frequencies

  Human hearing is more sensitive at lower frequencies

Quantization Error (Noise) Analysis

- **Quantization output**
  - Input signal plus noise
  - Noise is difference of output and input signals
- **Signal-to-noise ratio (SNR) derivation**
  - Quantize to B bits

  - Quantization error

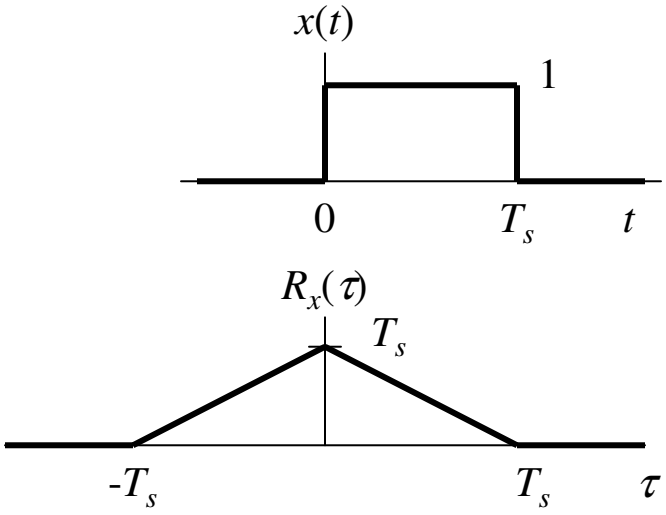
$$q = Q_B[m] - m = v - m$$

- **Assumptions**
  - $m \in (-m_{\max}, m_{\max})$
  - Uniform midrise quantizer
  - Input does not overload quantizer
  - Quantization error (noise) is uniformly distributed
  - Number of quantization levels $L = 2^B$ is large enough so that $\frac{1}{L-1} \approx \frac{1}{L}$

Quantization Error (Noise) Analysis

- **Deterministic signal $x(t)$ w/ Fourier transform $X(f)$**
  - Power spectrum is square of absolute value of magnitude response (phase is ignored)
    $P_x(f) = |X(f)|^2 = X(f)\,X^*(f)$
  - Multiplication in Fourier domain is convolution in time domain
  - Conjugation in Fourier domain is reversal and conjugation in time
    $X(f)\,X^*(f) = F\{x(\tau) * x^*(-\tau)\}$

- **Autocorrelation of $x(t)$**
  $R_x(\tau) = x(\tau) * x^*(-\tau)$
  - Maximum value at $R_x(0)$
  - $R_x(\tau)$ is even symmetric, i.e. $R_x(\tau) = R_x(-\tau)$



Quantization Error (Noise) Analysis

- **Power spectrum for signal $x(t)$ is $P_x(f) = F\{ R_x(\tau) \}$**
  - Autocorrelation of random signal $n(t)$

$$R_n(\tau) = E\{ n(t)\, n^*(t+\tau) \} = \int_{-\infty}^{\infty} n(t)\, n^*(t+\tau)\, dt$$

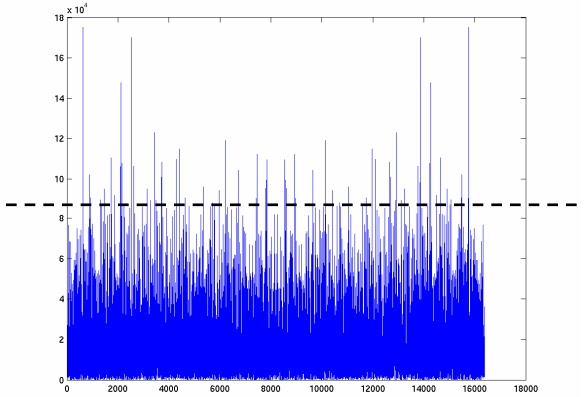
$$R_n(-\tau) = E\{ n(t)\, n^*(t-\tau) \} = \int_{-\infty}^{\infty} n(t)\, n^*(t-\tau)\, dt = n(\tau) * n^*(-\tau)$$

  - For zero-mean Gaussian $n(t)$ with variance σ^2

$$R_n(\tau) = E\{ n(t)\, n^*(t+\tau) \} = \sigma^2\, \delta(\tau) \;\Leftrightarrow\; P_n(f) = \sigma^2$$

- **Estimate noise power spectrum in Matlab**

```
N = 16384;    % number of samples
gaussianNoise = randn(N,1);
plot( abs(fft(gaussianNoise)) .^ 2 );
```



noise floor

Quantization Error (Noise) Analysis

- **Quantizer step size**

$$\Delta = \frac{2\,m_{\max}}{L-1} \approx \frac{2\,m_{\max}}{L}$$

- **Quantization error**

$$-\frac{\Delta}{2} \le q \le \frac{\Delta}{2}$$

q is sample of zero-mean random process Q

q is uniformly distributed

$$\sigma_Q^2 = E\{Q^2\} - \underbrace{\mu_Q^2}_{zero}$$

$$\sigma_Q^2 = \frac{\Delta^2}{12} = \frac{1}{3} m_{\max}^2\, 2^{-2B}$$

- **Input power: $P_{\text{average,m}}$**

$$\text{SNR} = \frac{\text{Signal Power}}{\text{Noise Power}}$$

$$\text{SNR} = \frac{P_{\text{average,m}}}{\sigma_Q^2} = \left(\frac{3P_{\text{average,m}}}{m_{\max}^2}\right) 2^{2B}$$

- **SNR exponential in *B***
- **Adding 1 bit increases SNR by factor of 4**
- **Derivation of SNR in deciBels on next slide**

Quantization Error (Noise) Analysis

- **SNR in dB = constant + 6.02 dB/bit * *B***

Loose upper bound

$$10\log_{10}\text{SNR} = 10\log_{10}\left(\left(\frac{3P_{\text{average,m}}}{m_{\max}^2}\right)2^{2B}\right)$$

$$= 10\log_{10}3 + 10\log_{10}\left(P_{\text{average,m}}\right) - 20\log_{10}\left(m_{\max}\right) + 20\,\text{B}\log_{10}(2)$$

$$= \underbrace{0.477 + 10\log_{10}\left(P_{\text{average,m}}\right) - 20\log_{10}\left(m_{\max}\right)}_{\text{1.76 and 1.17 are common constants used in audio}} + 6.02\,B$$

- **What is maximum number of bits of resolution for**

  Landline telephone speech signal of SNR of 35 dB

  Audio CD signal with SNR of 95 dB

Noise Immunity at Receiver Output

- **Depends on modulation, average transmit power, transmission bandwidth, channel noise, demod**
- **Analog communications (*receiver output SNR*)**

  "When the carrier to noise ratio is high, an increase in the transmission bandwidth B_T provides a corresponding quadratic increase in the output signal-to-noise ratio or figure of merit of the [wideband] FM system."
  – Simon Haykin, *Communication Systems*, 4th ed., p. 147.

- **Digital communications (*receiver symbol error*)**

  For code division multiple access (CDMA) spread spectrum communications, probability of symbol error decreases exponentially with transmission bandwidth B_T
  – Andrew Viterbi, *CDMA: Principles of Spread Spectrum Communications*, 1995, pp. 34-36.