ECE 8873
Data Compression & Modeling

Lecture 6: Quantization Theory

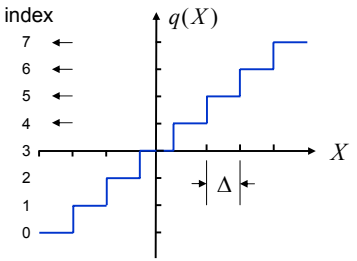
School of Electrical and Computer Engineering
Georgia Institute of Technology
Spring, 2004

Digital Representation of Analog Signals

- An analog signal has sample value that needs to be "rounded" to be represented in computers.
- Sheppard analyzed "round-off error" in 1898 (according to Bob Gray).
- Generally speaking, quantization means:
  - Divide the region (of value) into disjoint intervals;
  - For each value to be quantized (or re-quantized), find the interval it is in;
  - Represent the interval by its index for transmission and storage; same index also points to a typical value to be used at the receiver to recover the original value (with some discrepancy).

Uniform Quantizer

- A quantizer is called uniform if the intervals are equally and contagiously spaced.
- Conventionally, the typical value chosen for each interval for recovery at the receiver is the mid-point of the interval.



Quantization error:

$$\varepsilon = q(X) - X$$

$$X = q(X) - \varepsilon$$

Granular region: $|\varepsilon| < \Delta/2$

Overload or saturation region: quantization error is unbounded

Quantization Error - SQNR

- Suppose data is 0-mean uniformly distributed, no overload, i.e. $X \in [-G/2, G/2]$ and $\Delta = G/N$.

$$E[\varepsilon^2] = E[(q(X) - X)^2] = \frac{\Delta^2}{12}$$

- Entropy at the output (uniform q, uniform pdf)

$$H(q) = \ln N, \quad N \text{ is the number of intervals}$$

- In terms of SNR: "6dB gain per bit"

$$10\log_{10}\frac{E[X^2]}{E[\varepsilon^2]} = k + 10\log_2 N^2 \log_{10} 2 \approx k + 6R \ \text{(dB)}$$

- Bennett showed that for high rate, $N \to \infty$, the above average distortion result is approximately true for any pdf, not just uniform pdf.

Basics Again

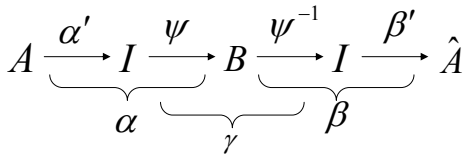
- A quantization/coding scheme consists of
  - An encoder $\alpha: \quad A^* \to B^* = \{0,1\}^*$

    If $x \in A$ and $\alpha(x)$ appear as a parsable units in their respective domain, we can define the average code length $E[l(\alpha(x))]$ per input "unit," which can be sample, pixel, word (variable number of letters), sequence, area, or time.
  - A decoder $\beta: \quad B^* \to \hat{A}^*, \ \beta(\alpha(x)) = \hat{x} \in \hat{A}$
  - A reproduction codebook $\hat{A} = \{\beta(i), \ i \in I\}$

    Each $\beta(i), \ i \in I$ is called a codeword and the codeset I has elements represented in integer index without loss of generality.

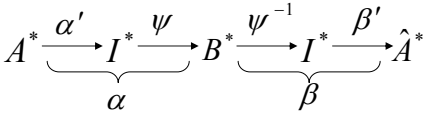
    Why?

Complete Coding Model



Source Encoder α' : maps letters in A to indices i in I.
Channel Encoder ψ : maps indices in I to channel code in B.
Channel Decoder ψ^{-1} : maps channel codes back to indices in I.
Source Decoder β' : maps indices in I to codewords $\beta(i)$ in C.
Channel/transmisson system γ is assumed to be lossless - $\psi\,\psi^{-1}$ is equal to unity.

For sequences:



Objectives & Concerns

- Given $\|I\|$, find the coder and the decoder such that the discrepancy between x and $\hat{x}$ is minimum.
- How to measure discrepancy? Use "distortion measure" $d(x,\hat{x})$.
- With $\|I\|$ fixed, instead of looking at $E[l(\alpha(x))]$, we are interested in $E[d(X,\hat{X})] = E[d(X,\beta(\alpha(X)))] = D(\alpha,\beta)$

$d(x,\hat{x})$ carries the implication that x and $\hat{x}$ are parsable units and the function is defined over x and $\hat{x}$.

Often $d(x,\hat{x}) \to d(g(\mathbf{x}), g(\hat{\mathbf{x}}))$ involves a window/frame of data or requires some alignment in the sequential order due to possibility in insertion and deletion of source symbols.

Further Notations & Terminology

Quantizer: $Q = (\alpha, \beta, l)$ l is the "length function" which also determines the rate.

Distortion: $d(x,\hat{x}) = d(x,\beta(\alpha(x)))$

Average or expected distortion:

$$E[d(X,\hat{X})] = E[d(X,\beta(\alpha(X)))] = D(\alpha,\beta)$$

Average or expected rate: $E[l(\alpha(x))]$

Encoder partition: $S = \{S_i, i \in I\}$

where $S_i = \{x: \ \alpha'(x) = i \in I\}$

S_i are disjoint $\Rightarrow \quad S_i \cap S_j = \phi$ if $i \neq j$

$\cup_{i,i\in I} S_i = A$

Centroid

Let $d(x,\hat{x}) = d_2^2 = (x-\hat{x})^t(x-\hat{x})$

$\arg\min_{\hat{x}} E[(X-\hat{x})^t(X-\hat{x})] = E[X]$

$E[X]$ = centroid of $X = \overline{X}$

(Since X is a random variable, this notation also implies the pdf and the space/domain of X, which could be the entire space or a cell of a partition.)

If training data are used instead of the real pdf, the expectation is replaced by the sample mean or sample average and

$\overline{X} = \frac{1}{L}\sum_{i=1}^{L} x_i$ is the "centroid" or Euclidean center of gravity.

High Rate Quantizers

- Relevant literature: Bennett (1948), Lloyd (1957), Zador (1963), Gersho (1979)

Random vector X with sooth pdf f.

Quanitzer: $Q = (\alpha, \beta, l)$

Encoder partition: $S = \{S_i, i \in I\}, \quad I = \{1, 2, \cdots, N\}$

Reproduction codebook: $\hat{A} = \{\beta(i),\ i \in I\}$

Average distortion: $D(Q) = E[d(X,\hat{X})] = E[\| X - \beta(i(X)) \|^2]$

$$= \sum_{i=1}^{N} \int_{S_i} \| x - \beta(i(x)) \|^2 \ f(x)dx$$

Average rate: $\log_2 N$ (assuming fixed rate)

Finite & Infinite Cells & Bennett's Assumptions

- Define volume of a cell $V(S_i) = \int_{S_i} dx$

  In granular region, $V(S_i) < \infty$.

  In overload region, $V(S_i) = \infty$.

$$D(Q) = \sum_{i:V(S_i)<\infty} \int_{S_i} \| x - \beta(i) \|^2 \ f(x)dx + \sum_{i:V(S_i)=\infty} \int_{S_i} \| x - \beta(i) \|^2 \ f(x)dx$$

Bennett's Assumptions:

- N is VERY large;
- f is smooth (Riemann sums approach Riemann integrals and mean value theorem applies);
- Total overload distortion is negligible;
- Bounded cells have only tiny volumes;
- Reproduction codewords are centroids corresponding to uniform pdf.

Consequences

- Pdf is smooth and the cell volumes are "tiny,"

$$f(x) \approx f(\beta(i)) \quad \text{for } x \in S_i$$

$$P(S_i) = \int_{S_i} f(x)dx \approx f(\beta(i)) \int_{S_i} dx = f(\beta(i))V(S_i)$$

which leads to $f(\beta(i)) = P(S_i)V^{-1}(S_i)$ and

$$D(Q) \approx \sum_{i=1}^{N} P(S_i) \int_{S_i} V^{-1}(S_i) \| x - \beta(i) \|^2 \ dx$$

In other words, under the conditions, centroid w.r.t. true pdf is approximately the same as the centroid for uniform pdf. That is, we can focus on $\min \int_{S_i} \| x - \beta(i) \|^2 \ dx$ for each cell, rather than $\min \int_{S_i} \| x - \beta(i) \|^2 \ f(x)dx$

Cell Shapes

- Now, if local pdf variation does not matter (i.e., roughly uniform), the factor that affect the centroid (and the distortion) the most is the shape.
- But remember $\cup_{i,i\in I} S_i = A$; so, no unattended space allowed ⟹ How to pack the space tightly and minimize D?

Introduce dimensionless 2nd moment of a cell S, $M(S)$

$$M(S) = \frac{1}{k}\frac{1}{V(S)^{2/k}}\int_S \frac{\| x-\bar{x}(S)\|^2}{V(S)}dx$$

k is dimension of x.

$x(S)$ is the Euclidean centroid of S.

$M(S)$ depends on shape and not scale.

$$M(S) = M(cS) \qquad cS = \{cx : c > 0, x \in S\}$$

Average Distortion

With $$M(S_i) = \frac{1}{k}\frac{1}{V(S_i)^{2/k}}\int_{S_i}\frac{\| x-\bar{x}(S_i)\|^2}{V(S_i)}dx$$

$$D(Q) \approx \sum_{i=1}^{N} P(S_i)\int_{S_i} V^{-1}(S_i)\,\| x-\beta(i)\|^2\, dx$$

$$= \sum_{i=1}^{N} P(S_i)kM(S_i)V(S_i)^{2/k}$$

$$\approx \sum_{i=1}^{N} f(\beta(i))kM(S_i)V(S_i)^{1+2/k}$$

recall $P(S_i) = f(\beta(i))V(S_i)$

In this expression, the effect of the partition shape is somewhat isolated.

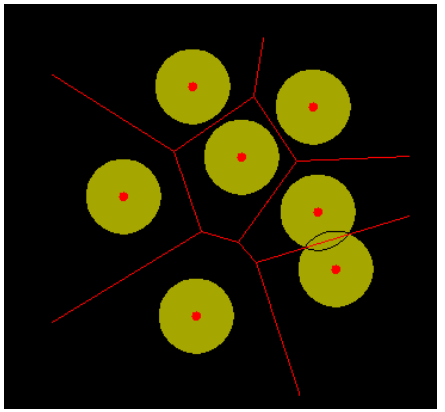
What kind of partition would minimize $D(Q)$?

Consider the bin packing problem.

Voronoi Region, Cell, Partition

- (AMS definition) Consider any collection of non-overlapping discs of equal size (possibly just isolated points). We associate to this collection a partition of the plane into regions called **Voronoi cells**. The Voronoi cell of a disc in the distribution is the set of points in the plane which are as close or closer to the center of that disc than to the center of any other disc in the distribution. The partition we associate to any distribution of non-overlapping discs in the plane is its Voronoi partition.
  - http://www.ams.org/new-in-math/cover/cass4.html

Voronoi Partition



Lattice

- A lattice $\mathcal{L}$ in R^k is a set of all vectors represented by a linear combination of integer multiples of the generator vectors.

Let $\{\mathbf{u}_i; i = 1, 2, \cdots\}$ be the set of generating vectors.

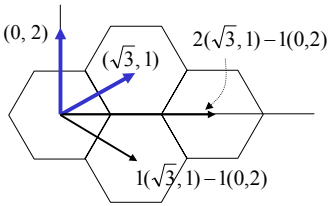
$\mathbf{U} = [\mathbf{u}_1\ \mathbf{u}_2 \cdots \mathbf{u}_k]$

$\mathbf{m} = [m_1\ m_2 \cdots m_k]^t, \quad m_i \in Z, \quad \mathbf{m} \in Z^k$

$\mathcal{L} = \{\mathbf{v};\ \mathbf{v} = \mathbf{U}\mathbf{m}\}$

Hexagonal lattice:

$$\mathbf{U} = \begin{bmatrix} \sqrt{3} & 0 \\ 1 & 2 \end{bmatrix}$$



Lattice from Voronoi Partition

- Let Ξ_0 be the Voronoi cell of a lattice with its Euclidean centroid located at the origin.
- It has volume $V(\Xi_0)$ and normalized moment of inertia $M(\Xi_0)$.
- Let $M(S_i) = M(\Xi_0)$ for all i .

$$D(Q) \approx kM(\Xi_0)\sum_{i=1}^{N} f(\beta(i))V(\Xi_0)^{1+2/k}$$

$$= kM(\Xi_0)V(\Xi_0)^{2/k}\sum_{i=1}^{N} f(\beta(i))V(\Xi_0)$$

$$\approx kM(\Xi_0)V(\Xi_0)^{2/k}\sum_{i=1}^{N} P(S_i) = kM(\Xi_0)V(\Xi_0)^{2/k}$$

Scaled Partition & Bounded Partition

- Scaled partition: $\mathcal{L} \to a\mathcal{L}$

The normalized moment of inertia does not change $M(\mathcal{L}_0) = M(a\mathcal{L}_0)$

The average distortion $D(Q)$ decreases as the volume shrinks in $V(\Xi_0)^{2/k}$

- Bounded Partition
  - Input pdf is concentrated in a bounded set, say A
  - Since N is large and the cell is small, $V(A) \approx NV(\Xi_0)$

$$D(Q) \approx kM(\Xi_0)V(A)^{2/k}N^{-2/k}$$

  - The best lattice quantizer is the one that minimizes $M(\Xi_0)$ over all lattices – cell shape is the key.

Gersho's Conjecture (Gray)

- For optimal quantization, all cells are well approximated asymptotically by polytopes which are scaled, rotated or translated versions of a single tesselating (space filling) convex polytope Ξ^* with minimum normalized moment of inertia $M(\Xi^*) = Q_k$, and $\lambda(x)$, the quantizer point density function

$$D(Q) \approx k\sum_{i=1}^{N} f(\beta(i))M(S_i)V(S_i)^{1+2/k}$$

$$\approx kM(\Xi^*)\sum_{i=1}^{N} f(\beta(i))\frac{V(S_i)}{[N\lambda(\beta(i))]^{2/k}}$$

$$\approx kM(\Xi^*)\int f(x)\frac{dx}{[N\lambda(x)]^{2/k}} = kM(\Xi^*)E\left\{\frac{1}{[N\lambda(x)]^{2/k}}\right\}$$

Q_k only depends on dimension k.

$E\left\{\frac{1}{[N\lambda(x)]^{2/k}}\right\}$ involves distributions of data and quantization points; in lattice quantizer, the point density is uniform and only the shape matters.

Common Minimum NMI

- k=1 Ξ^* = line segments $M(\Xi^*) = Q_1 = \frac{1}{12} = 0.08333...$
- k=2 Ξ^* = regular hexagonal [Fejes Toth (1959)]

$$M(\Xi^*) = Q_2 = \frac{5}{36\sqrt{3}} = 0.08019...$$

- k=3 Ξ^* = regular truncated octohedron

$$M(\Xi^*) = Q_3 = \frac{19}{192 \times 2^{1/3}} = 0.07855...$$

In general, lower bound = sphere bound

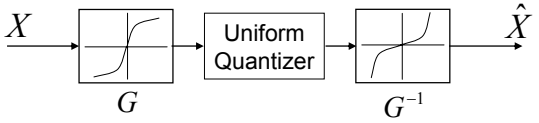
$$Q_k \geq M(sphere) = \frac{V_k^{-2/k}}{k+2}$$

$$= \frac{\text{volume of unit - radius sphere in k dimensions}}{k+2}$$

Non-uniform Quantizer – Impact of Distribution

k=1 Back to scalar case

- If data statistics is known, it is possible to tailor the quantizer to the input statistics for superior SQNR.



G: Monotonic non-linearity, a compressor

G^{-1}: Monotonic non-linearity, an expander

G relates to quantizer point density:

$\lambda(x) = kG'(x), \quad G'(x) = \text{derivative of } G(x)$

k = normalizing constant related to the range of the quantizer.

How to design the non-linearity?

The Distortion Integral – High Rate Again

$$D(\alpha, \beta) = D = \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} (x - y_i)^2 f_X(x) dx \qquad y_i = \beta(i)$$

For Large N, $f_X(x) = f_i = const.; \; x \in S_i$

And $P_i = P(S_i) = \Pr(X \in S_i) = \int_{x_{i-1}}^{x_i} f_X(x) dx \approx (x_i - x_{i-1}) f_i$

$$f_i = \frac{P_i}{x_i - x_{i-1}} = \frac{P_i}{\Delta_i} \qquad D = \sum_{i=1}^{N} P_i \int_{x_{i-1}}^{x_i} \frac{(x - y_i)^2}{\Delta_i} dx \approx \sum_{i=1}^{N} \frac{P_i \Delta_i^2}{12}$$

Now, recall quantizer point density function $\lambda(x) = \lim_{N \to \infty} \frac{N(x)}{N}$

$$\Delta_i \equiv \frac{\text{interval length}}{\text{number of levels in the interval}} = \frac{\Delta x}{N\lambda(x)\Delta x} \approx \frac{1}{N\lambda(y_i)}$$

$$D \approx \sum_{i=1}^{N} \frac{P_i \Delta_i^2}{12} \approx \sum_{i=1}^{N} \frac{P_i}{12[N\lambda(y_i)]^2} \approx \sum_{i=1}^{N} \frac{f_X(y_i)\Delta_i}{12[N\lambda(y_i)]^2} \approx \frac{1}{12N^2} \int_{x_1}^{x_{N-1}} f_X(y) \lambda^{-2}(y) dy$$

Quantization Point Density

$$D \approx \frac{1}{12N^2} \int_{x_1}^{x_{N-1}} f_X(y) \lambda^{-2}(y) dy$$

Hölder's inequality

For any positive a and b with $a^{-1} + b^{-1} = 1$

$$\left(\int u(x)v(x)dx\right) \leq \left(\int u^a(x)dx\right)^{1/a} \left(\int v^b(x)dx\right)^{1/b}$$

Equality holds if $u(x) \propto v(x)$

$$u(x) = \left(f_X(x)\lambda^{-2}(x)\right)^{1/3}, \quad v(x) = \lambda^{2/3}(x)$$

$$\int \left(f_X(x)\lambda^{-2}(x)\right)^{1/3} \lambda^{2/3}(x)dx \leq \left(\int f_X(x)\lambda^{-2}(x)dx\right)^{1/3} \left(\int \lambda(x)dx\right)^{2/3}$$

$$\int f_X^{1/3}(x)dx \leq \left(\int f_X(x)\lambda^{-2}(x)dx\right)^{1/3}$$

$$D \approx \frac{1}{12N^2} \int_{x_1}^{x_{N-1}} f_X(y)\lambda^{-2}(y)dy \geq \frac{1}{12N^2} \left(\int_{x_1}^{x_{N-1}} f_X^{1/3}(y)dy\right)^3$$

with equality if $\lambda(x) = f_X^{1/3}(y) \left(\int_{x_1}^{x_{N-1}} f_X^{1/3}(y)dy\right)^{-1}$ $\quad G'(x) \propto f_X^{1/3}(x)$

Partial Distortion Theorem

$$D = \sum_{i=1}^{N} D_i = \sum_{i=1}^{N} E\{(X - \alpha(X))^2 \mid X \in S_i\} P\{X \in S_i\}$$

Partial Distortion Theorem: For quantization with the asymptotically optimal (quantizer) point density function, the partial distortion $D_i = E\{(X - \alpha(X))^2 \mid X \in S_i\}$ are asymptotically constant with value D/N as N approaches infinity.

$$D = \sum_{i=1}^{N} P(S_i) \int_{x_{i-1}}^{x_i} \frac{(x - y_i)^2}{\Delta_i} dx \approx \sum_{i=1}^{N} \frac{P(S_i)\Delta_i^2}{12}$$

$$P(S_i) = \Delta_i f_X(y_i) \quad \text{and} \quad G'(y_i) \approx \Delta / \Delta_i$$

$$D_i \approx \frac{\Delta^3}{12} \frac{f_X(y_i)}{G'(y_i)^3} = \frac{c\Delta^3}{12}$$ independent of i.

Average distortion in each quantization region is the same if point density is optimally chosen.

Quantization Designs

- High rate analysis provides guidance on choice of shapes of partition.
- Lattice quantizer based on "space-filling polytope with minimum NMI" has the advantage of:
  - Efficient coding (fast search of closest region)
  - Performance $\propto N^{-2/k}$
- If smaller distortion is desired, then need to look at non-uniform quantizers with point density function to minimize $E\left\{[N\lambda(x)]^{-2/k}\right\}$
- The expectation can be evaluated by integration if the data pdf is known (e.g., the cubic root result for the scalar case); otherwise, use empirical methods.

Non-uniform Quantizer

- Coding principle
  - The best representation value in an interval is one that minimizes average distortion in that interval, i.e, the centroid
  - The best interval an input value is assigned to is the one whose centroid is closest to the input value.
- Lloyd's methods
  - Lloyd's method I (Lloyd algorithm)
  - Lloyd's method II
    - Reported in 1957.
    - Rediscovered by Max in 1960; led to the name Lloyd-Max quantizer.

Lloyd's Method II

- For scalar quantizer design
  - Initialization: set the lowest threshold $\tau_0 = -\infty$; choose a smallest reproduction value $\hat{x}_1$
  - Since $\hat{x}_1$ must be the centroid of (τ_0, τ_1), τ_1 can be computed.
  - But τ_1 must be mid-point between $\hat{x}_1$ and $\hat{x}_2$ due to the nearest neighbor principle for minimizing coding distortion given any codebook, $\hat{x}_2$ is thus determined as $\hat{x}_2 = \tau_1 + (\tau_1 - \hat{x}_1) = 2\tau_1 - \hat{x}_1$.
  - Iterate until final codeword is chosen. If $\hat{x}_N$ is not close to the centroid of (τ_{N-1}, ∞) , adjust reproduction initial value $\hat{x}_1$. If close enough, stop.

Lloyd Principle and Consequences

- Suppose all reconstruction values satisfy (Lloyd's) centroid conditions – they minimize squared error in respective intervals.
- Then, $\varepsilon = \alpha(X) - X$

$$E[\varepsilon] = 0$$

$$E[\alpha(X)\varepsilon] = 0$$

$$E[\varepsilon^2] = \sigma_\varepsilon^2 = \sigma_X^2 - \sigma_{q(X)}^2$$

$$E[X\varepsilon] = E[(\alpha(X) - \varepsilon)\varepsilon] = -E[\varepsilon^2] = \sigma_\varepsilon^2$$

$$E[\varepsilon^2] = \sigma_\varepsilon^2 = \sigma_X^2 - \sigma_{q(X)}^2 \geq 0 \quad \therefore \sigma_X^2 \geq \sigma_{q(X)}^2$$