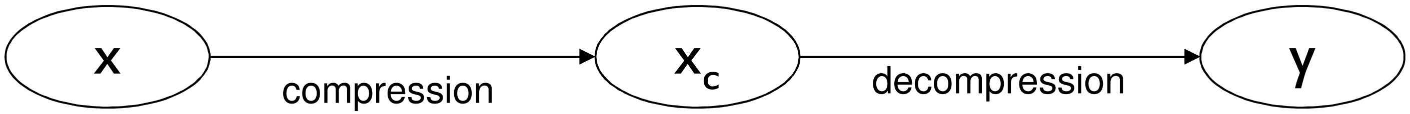
Mathematical Background on Lossy Data Compression

National Chiao Tung University
Chun-Jen Tsai
10/30/2014

Concept of Lossy Coding

- If x is the original data, x_c is the compressed representation, and y is the reconstructed data,



x must equals y for lossless coding

- For lossy coding, we want to find a way to modify x such that the entropy is reduced.

As a result, $x \approx x' = y'$, and $Rate(y') < Rate(y)$.

Distortion Criteria

- The difference between x and x' is the distortion
- Whether the distortion is acceptable or not depends on the applications:
  - A work of art?
  - Commercial photos?
  - Machine vision applications?
  - Audiophile entertainment?
  - Political speech broadcasting?
- If the target user of the distorted data is a human:
  - Difficult to incorporate the human response into mathematical design procedures
  - If a human is used to evaluate distortion, there is difficulty in objectively reporting the results

Objective Distortion Measures (1/2)

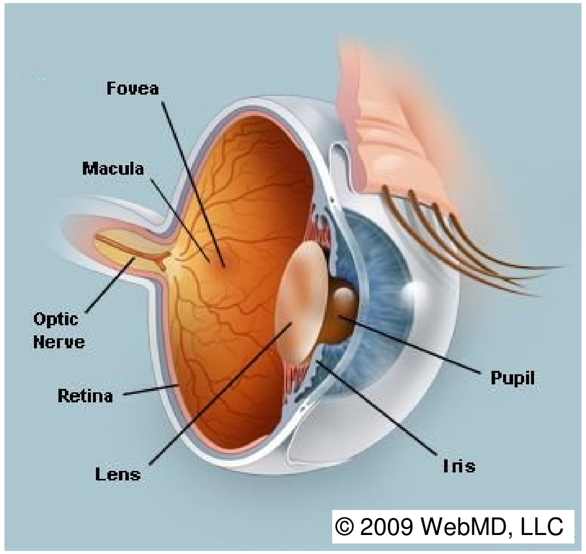
- If $\{x_n\}$ is the source and $\{y_n\}$ is the reconstructed data:
  - Squared error measure: $d(x, y) = (x - y)^2$
  - Absolute difference measure: $d(x, y) = |x - y|$.
- A scalar-value measure is "easier" to use:
  - Mean square error (MSE): $\sigma_d^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - y_n)^2.$
  - Mean absolute difference (MAD): $d_1 = \frac{1}{N}\sum_{n=1}^{N}|x_n - y_n|.$
  - Max error measure: $d_\infty = \max_n |x_n - y_n|.$

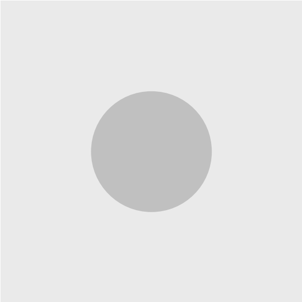
Objective Distortion Measures (2/2)

- Often, relative error measures (w.r.t. $\{x_n\}$) are more descriptive:
  - Signal-to-noise ratio: $SNR = 10\log_{10}\frac{\sigma_x^2}{\sigma_d^2}$ (dB).
  - Peak-signal-to-noise-ratio: $PSNR = 10\log_{10}\frac{x_{peak}^2}{\sigma_d^2}$ (dB).
- Until today, there is no perceptual measures (neither visual nor audio) that can represent human perceptions objectively

Human Visual System

- Human eyes
  - Retina: has two types of sensors
    - Rod – sensitive to magnitude
    - Cone – sensitive to wavelengths
  - Fovea
    - A small area of the retina where cones concentrate
    - High resolution area of retina
  - Just noticeable difference (JND)
    - If the background intensity is I, the center intensity is $I + \Delta I$, JND is the minimal ΔI which makes the center square visible

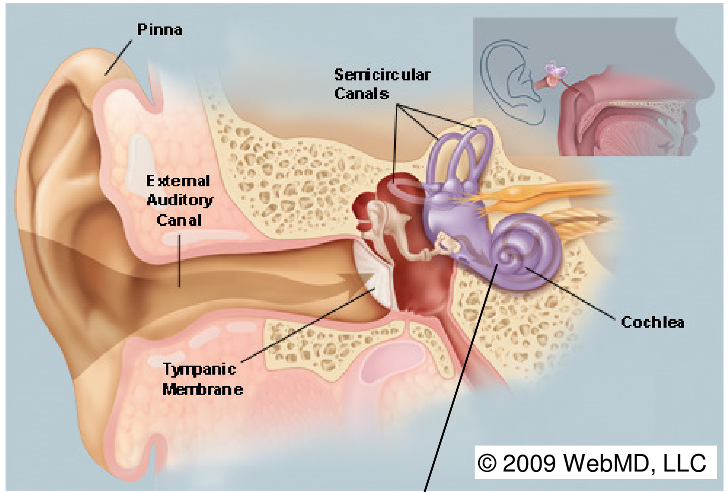


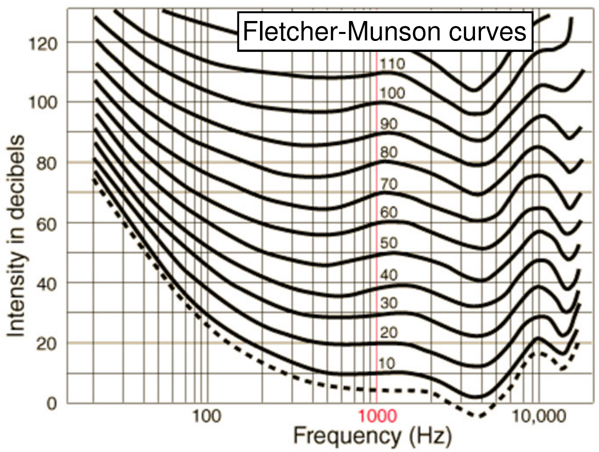


contrast sensitivity test

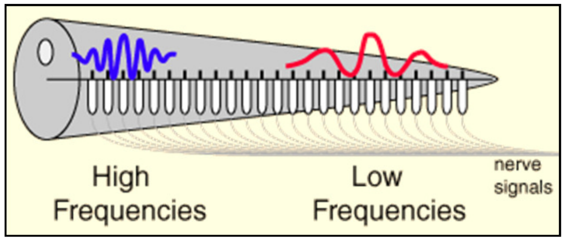
Human Auditory Perception

- Human auditory system model (Basilar Membrane):
  - A bandpass filterbank
  - 25 overlapping critical bands covering 20~20k Hz
  - Masking: a loud sound will mask the audibility of another sound of nearby frequency (in the same critical band)





Critical band effect



Formulation of Lossy Compression

- Assume that the source alphabet $X = \{x_0, x_1, \ldots, x_{N-1}\}$ and the reconstructed alphabet $Y = \{y_0, y_1, \ldots, y_{M-1}\}$ are different:
  - What is the information relationship between two different (but correlated) random variables?
- Note that the entropies of the source and the reconstruction are:

$$H(X) = -\sum_{i=0}^{N-1} P(x_i)\log_2 P(x_i)$$

$$H(Y) = -\sum_{j=0}^{M-1} P(y_j)\log_2 P(y_j).$$

Conditional Self-Information

- A measure of the relationship between two random variables is the *conditional entropy* (the average value of the conditional self-information)
- The conditional self-information of an event A, given that another event B has occurred, can be defined as

$$i(A \mid B) = \log \frac{1}{P(A \mid B)} = -\log P(A \mid B).$$

  - B : the event "something is barking"
    A : the event "there is a dog"
    $\rightarrow P(A \mid B)$ should be close to one, which means that the conditional self-information $i(A \mid B)$ would be close to zero

Conditional Entropy

- The conditional entropies of the source and reconstruction are given as

$$H(X|Y) = -\sum_{i=0}^{N-1}\sum_{j=0}^{M-1} P(x_i|y_j)P(y_j)\log_2 P(x_i|y_j)$$

- The conditional entropy $H(X|Y)$ is the amount of uncertainty about X, given that we know what value the reconstruction Y took. Note that $H(X|Y) \le H(X)$.

Note: $H(X|Y) = \sum_{j=0}^{N-1} P(Y=j)H(X|Y=j) = -\sum_{j=0}^{N-1} P(Y=j)\sum_{i=0}^{M-1} P(X=i|Y=j)\log_2 P(X=i|Y=j).$

Example: Uniform Quantization (1/2)

- Let X = $\{0, 1, \ldots, 15\}$, Y = $\{0, 2, \ldots, 14\}$, $y_i = \lfloor x_i/2 \rfloor \times 2$.
  - Assume $P(X = i) = 1/16$, for $i \in$ X, then $H(X) = 4$ bits.
  - $P(Y = j) = P(X = j) + P(X = j+1) = 1/8 \rightarrow H(Y) = 3$ bits.

$$P(X = i \mid Y = j) = \begin{cases} \frac{1}{2} & \text{if } i = j \text{ or } i = j+1, \text{ for } j = 0, 2, 4, \ldots, 14 \\ 0 & \text{otherwise.} \end{cases}$$

- Thus, the conditional entropy $H(X \mid Y)$ is

$$\begin{aligned} H(X \mid Y) &= -\sum_i \sum_j P(X = i \mid Y = j) P(Y = j) \log_2 P(X = i \mid Y = j) \\ &= -\sum_j [P(X = j \mid Y = j) P(Y = j) \log_2 P(X = j \mid Y = j) + \\ &\quad + P(X = j+1 \mid Y = j) P(Y = j) \log_2 P(X = j+1 \mid Y = j)] = 1. \end{aligned}$$

Example: Uniform Quantization (2/2)

- Note that, the conditional entropy $H(X \mid Y) = 1$ means that the uncertainty of X given Y is 1 bit

- On the other hand, since

$$P(Y = j \mid X = i) = \begin{cases} 1 & \text{if } i = j \text{ or } i = j+1, \text{ for } j = 0, 2, 4, ..., 14 \\ 0 & \text{otherwise.} \end{cases}$$

we have that $H(Y \mid X) = 0$ bits.

Average Mutual Information (1/2)

- Mutual information: the amount of joint information contained by both X and Y. For the joint event x_i and y_j, the mutual information is defined as

$$i(x_i; y_j) = \log \frac{1}{P(x_i, y_j)} - \log \frac{1}{P(x_i)P(y_j)} = \log \frac{P(x_i \mid y_j)}{P(x_i)}.$$

- Average mutual information:

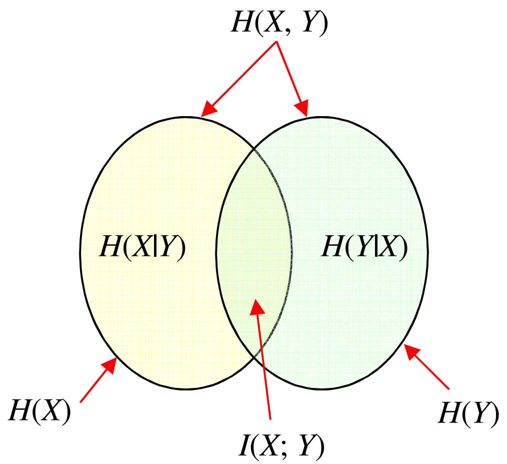
$$I(X;Y) = \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} P(x_i, y_j) i(x_i; y_j)$$

$$= \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} P(x_i \mid y_j) P(y_j) \log \frac{P(x_i \mid y_j)}{P(x_i)}.$$

Average Mutual Information (2/2)

- $$I(X;Y) = \sum_{i=0}^{N-1}\sum_{j=0}^{M-1} P(x_i, y_j) \log \frac{P(x_i \mid y_j)}{P(x_i)}$$
  $$= \sum_{i=0}^{N-1}\sum_{j=0}^{M-1} P(x_i, y_j) \log P(x_i \mid y_j) - \sum_{i=0}^{N-1}\sum_{j=0}^{M-1} P(x_i, y_j) \log P(x_i)$$
  $$= H(X) - H(X \mid Y).$$

- $I(X; Y) = I(Y; X)$



Differential Entropy

- The concept of entropy can be extended to sources with continuous distributions. The differential entropy of a random variable X with pdf $f_X(x)$ is defined to be:

$$h(X) = -\int_{-\infty}^{\infty} f_X(x) \log f_X(x) dx.$$

- With this definition, the relation between mutual information and entropies of X and Y still holds:

$$I(X; Y) = h(X) - h(X \mid Y).$$

Rate-Distortion Theory

- Rate distortion theory is concerned with the trade-offs between distortion and rate in lossy compression schemes
- Rate distortion function $R(D)$:
  - A function that specifies the lowest rate at which the output of a source can be encoded while keeping the distortion less than or equal to D.
  - Given a source X, a reconstruction Y, and a distortion constraint D^*, if the distortion measure is $d(x, y)$, then

$$D = \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} P(y_j \mid x_i) P(x_i) d(x_i, y_j).$$

  But, what is the lowest R for $D \le D^*$? Is it minimal $H(Y)$?

Minimal Rate R Given D and Codec

- Note that, if the distortion constraint D^* is large, random guesses on the decoder side (which has $R = 0$) may still satisfy the rate constraint $D \le D^*$.
- In 1959, Shannon showed that the minimal rate for a given distortion is given by

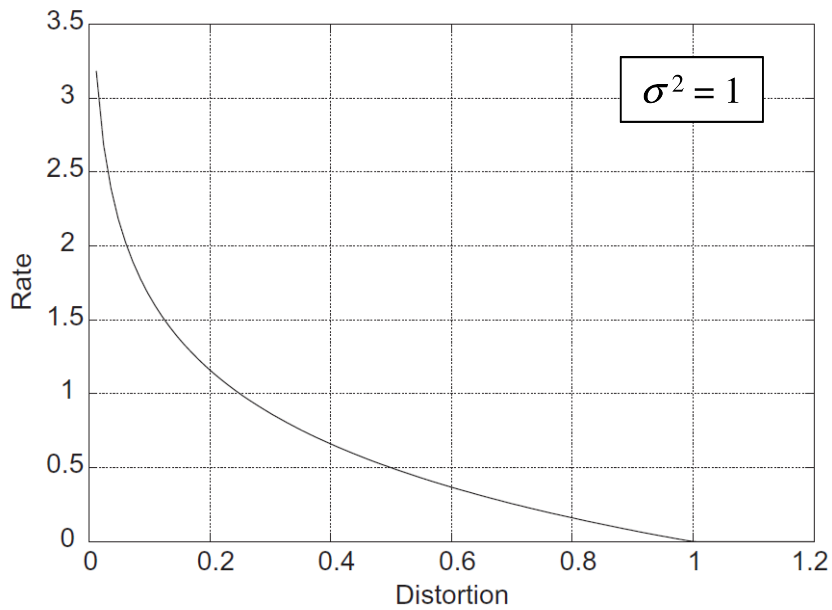
$$R(D) = \min_{\{P(y_j|x_i)\}\in\Gamma} I(X;Y),$$

  where $\Gamma = \left\{ \{P(y_j \mid x_i)\} \text{ such that } D(\{P(y_j \mid x_i)\}) \le D^* \right\}$ is determined by the compression scheme
  - $H(Y \mid X) = 0 \rightarrow I(X\,;\,Y) = H(Y)$
  - $H(Y \mid X) = H(Y) \rightarrow I(X\,;\,Y) = 0$

Theoretical Rate-Distortion Function

- Assume that the data source is zero mean Gaussian with variance σ^2. If the distortion function is $d(x, y) = (x - y)^2$, the R-D function is:

$$R(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D}, & D < \sigma^2 \\ 0, & D > \sigma^2 \end{cases}$$

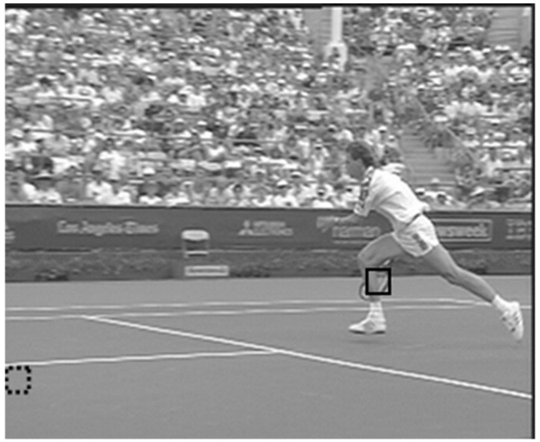


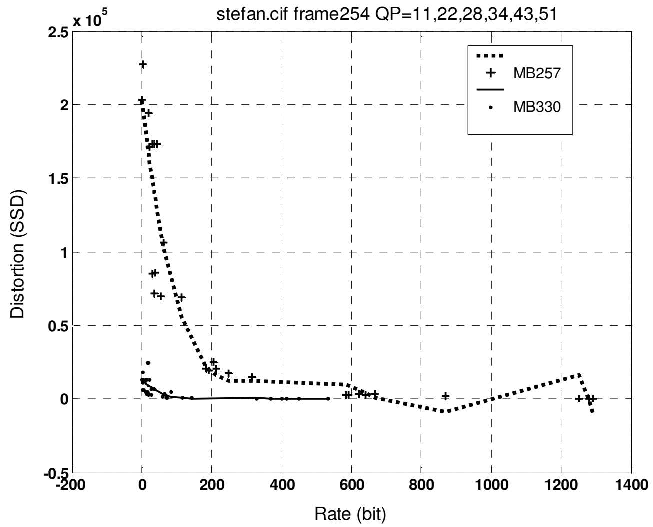
Rate-Distortion Functions in Practice

- The simplest (yet effective) first-order R-D model for video data:

$$R = \alpha \cdot \frac{C}{D}$$

where R is the rate, C the video complexity, D the distortion, and α the R-D model parameter.





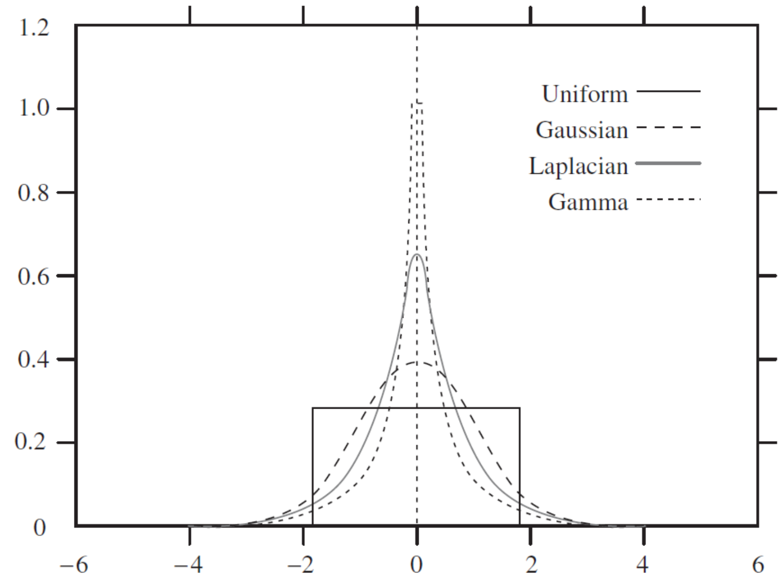
Source Models

- If the sources can be modeled accurately, we would be able to derive more accurate R-D relationships for coding decisions
  - In practice, tractable model that performs generally ok is better than precise model that works well for specific output samples
- Popular models
  - Probability models
  - Linear system models

Data Source Probability Models

- Uniform distribution
  - Used when we know nothing about the source

- Gaussian distribution
  - Mathematically simple
  - Sample mean approaches Gaussian
- Laplacian Distribution
  - Has higher concentration at zero than Gaussian model
  - Most de-correlated multimedia data has this characteristic
- Gamma Distribution
  - Even more peaked at zero than Laplacian model



Linear System Models

- Autoregressive Moving Average Model: ARMA(N, M)

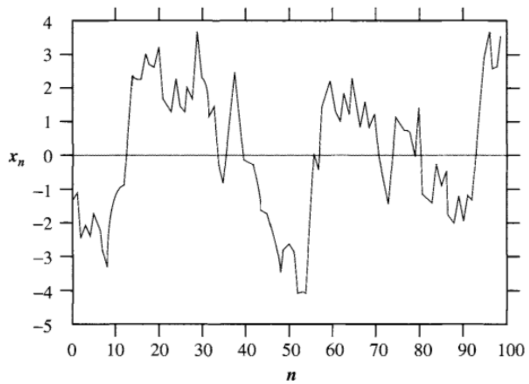
$$x_n = \sum_{i=1}^{N} a_i x_{n-i} + \sum_{j=1}^{M} b_j \varepsilon_{n-j} + \varepsilon_n.$$

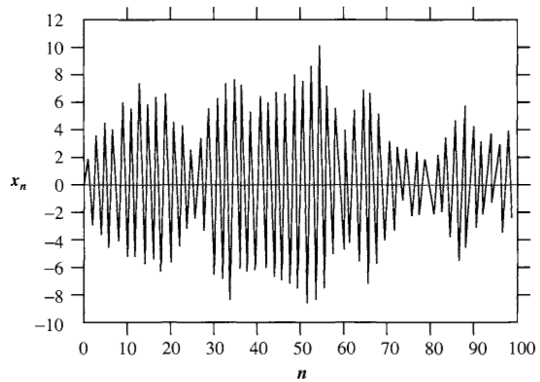
- Autoregressive Model: AR(N)

$$x_n = \sum_{i=1}^{N} a_i x_{n-i} + \varepsilon_n.$$

  - AR(N) is a Markov Model of order N.

- Examples of AR(1) sources:





Auto Correlation Function

- The autocorrelation function for the $AR(N)$ process can be obtained as follows:

$$R_{xx}(k) = E[x_n x_{n-k}] = E\left[\left(\sum_{i=1}^{N} a_i x_{n-i} + \varepsilon_n\right) x_{n-k}\right]$$

$$= E\left[\sum_{i=1}^{N} a_i x_{n-i} x_{n-k}\right] + E[\varepsilon_n x_{n-k}] = \begin{cases} \sum_{i=1}^{N} a_i R_{xx}(k-i), & k > 0 \\ \sum_{i=1}^{N} a_i R_{xx}(i) + \sigma_\varepsilon^2, & k = 0 \end{cases}$$

- Autocorrelation function of a process tells us the sample-to-sample behavior of a sequence
  - Slowly decay w.r.t. $k \rightarrow$ high sample-to-sample correlation
  - Fast decay w.r.t. $k \rightarrow$ low sample-to-sample correlation
  - No sample-to-sample correlation $\rightarrow$ zero (except when $k = 0$).