**ECE 8833**
***Data Compression and Modeling***

# Lecture 1: Introduction to Data Compression

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



## Signal & Coding

- Signal
  - Continuous-time or discrete-time function
  - Scalar- or vector-valued
  - Any information-bearing representations
- Coding (due to Shannon)
  - Source coding: conversion of signal into efficient digital representation for conservation of resources needed for transmission or storage of the signal.
  - Channel coding or error control coding: transformation of signal (or data) so as to permit reliable communication in presence of noise or distortion.

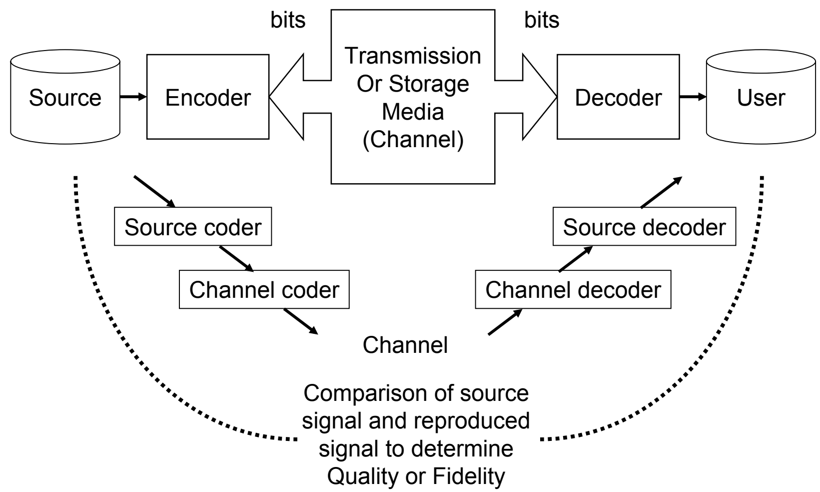


## Morse Code Alphabet

| | | | | |
|---|---|---|---|---|
| **A** .- | **I** .. | **Q** --.- | **Y** -.-- | **6** -.... |
| **B** -... | **J** .--- | **R** .-. | **Z** --.. | **7** --... |
| **C** -.-. | **K** -.- | **S** ... | **0** ----- | **8** ---.. |
| **D** -.. | **L** .-.. | **T** - | **1** .---- | **9** ----. |
| **E** . | **M** -- | **U** ..- | **2** ..--- | **Fullstop** .-.-.- |
| **F** ..-. | **N** -. | **V** ...- | **3** ...-- | **Comma** --..-- |
| **G** --. | **O** --- | **W** .-- | **4** ....- | **Query** ..--.. |
| **H** .... | **P** .--. | **X** -..- | **5** ..... | |



## A Framework for Data Compression



## Data Compression

- Various practical concepts related to time:
  - Time compression
  - Time scale modification with or without changing the signal characteristics
  - Garvey, W.D. "*The intelligibility of speeded speech*." Journal of Experimental Psychology, 45:102-108, 1953.
- Our interest:
  - Encoding or representation of information for storage or transmission at the lowest cost in resources (bandwidth, storage area, etc.) and without significant loss of information upon reconstruction.



## Coding as a Task

- Representation of analog signal for digital transmission or storage; often integrated with A/D conversion
- Compression of digital information to reduce transmission or storage requirement; compression can also be realized in analog domain
- efficiency is defined in terms of bandwidth or storage required for the delivery of a fixed amount of information such as a second of speech, a video frame
- Result of coding is a sequence of digital, often binary, symbols
- The sequence of digital symbols may or may not have explicit "delimiter."



## From Shannon Information Theory

- If the minimum achievable source coding rate of a given source is strictly below the capacity of the channel, then the source can be transmitted reliably by appropriate encoding-decoding; implicitly, reliable transmission can be accomplished by separate source and channel coding.
- If the source coding rate is strictly greater than the channel capacity, then reliable transmission is impossible; but, we can still strive to reduce the negative impact of the rate excess by joint source-channel coding.
- Memoryless block source codes can achieve minimum average distortion for a constrained rate, in the absence of complexity constraint – i.e. source coding subject to a fidelity criterion.



## Issues in Source Coding

- Coding algorithm design
- [Bit] rate and distortion relationship; lossy or lossless coding
- Implementation complexity
- Memory and delay requirement
- Robustness in performance against source variation
- Choice and significance of performance metric
- Impact of errors in code upon fidelity performance

Practical coding algorithms often involve detailed tradeoffs among these issues.

## Preliminaries

- Probability Theory
- Random Variables and Processes
- Linear systems
- Information Theory
- Entropy and measurement of information



## Shannon's Self-Information

- Let $X$ be an event of a random experiment and $P(A)$ denotes the probability that event $X$ will occur.
- Self-information associated with event $X$ is given by

$$i(X) = -\log_b P(X)$$

- If $X$ and $Y$ are independent events,

$P(XY) = P(X)P(Y)$

and thus

$i(XY) = -\log_b P(X)P(Y) = -\log_b P(X) - \log_b P(Y) = i(X) + i(Y)$

- When $b$=2, the unit of information is called bit; if the base is $e$, the unit is nat; if $b$=10, the unit is hartley.



## Information Source

- A source is an origin of information. A random source is equivalent to a random experiment, which generates outcomes for observation or reception.
- The mechanism that a random source uses to generate information is usually unknown to the observer, who sees only the outcomes of the experiment or the signals the source puts out.
- As in random experiments, an information source is associated with a probability measure, from which one can calculate the entropy of the source.
- When symbols or signals are generated in sequence, the sequential experiments may or may not be independent.



## Fundamental Dimensions of Source Coding

- Structure of information (modeling)
  - How is information generated by the source?
  - How to approximate the information-generation process?
  - How to represent this process?
- Random nature of information
  - Efficiency of codes depends on how precise the knowledge the encoder has about the source.
  - How to estimate the source distribution?
  - How to design codes to achieve maximum efficiency given prescribed constraints?

## Two Components of Information

- Structure – deterministic component; may or may not be known; may or may not be easily represented
- Entropy – random component; never known completely in real world

$$X(t) = A\cos(\omega t + \Theta) + V(t)$$

Many (incomplete) ways to view it:
- Treat every time sample as the outcome of an independent random experiment
- Treat the amplitude of the sinusoid as random variable
- Treat the phase as random variable
- Treat the signal not as a sinusoid but a general random process



## Defining a Source – Parallel to Pr Space

- Sample space, observation space, or signal space built upon a symbol set $A = \{\alpha_i\}_{i=1}^{M}$
  which is also called an alphabet without loss of generality, the symbols $\alpha_i$ are referred to as letters, and $m$ the size of the alphabet.
- Let $\mathbf{X} = (X_1, X_2, X_3, \cdots, X_n)$ be a signal sequence generated by the source. A sequence of length $n$ so generated can be considered as an outcome of a combined experiment with the observation space formed by the cartesian product of the original alphabet: $A^n = A \times A \times \cdots \times A$ and $X_i = \alpha_j \in A$

  *Again, the experiments may not be independent.*



## Source Entropy

- The average self-information of such a length-$n$ sequence is

$$G_n = -\sum_{i_1=1}^{m}\sum_{i_2=1}^{m}\cdots\sum_{i_n=1}^{m} \Pr(X_1 = \alpha_{i_1}, X_2 = \alpha_{i_2}, \cdots, X_n = \alpha_{i_n}) \bullet \log \Pr(X_1 = \alpha_{i_1}, X_2 = \alpha_{i_2}, \cdots, X_n = \alpha_{i_n})$$

- The entropy of the source (per symbol) is defined as

$$H(S) = \lim_{n\to\infty} \frac{G_n}{n}$$

- In the lack of complete knowledge of the experiment, assumptions are often made to facilitate entropy calculation; e.g., iid, Markov, …



## Source Entropy

- If $X_i$ are iid (independent & identically distributed), with $X$ denoting a generic random variable as $X_i$

$$G_n = -\sum_{i_1=1}^{m}\sum_{i_2=1}^{m}\cdots\sum_{i_n=1}^{m} \Pr(X_1 = \alpha_{i_1}, X_2 = \alpha_{i_2}, \cdots, X_n = \alpha_{i_n}) \bullet \log \Pr(X_1 = \alpha_{i_1}, X_2 = \alpha_{i_2}, \cdots, X_n = \alpha_{i_n})$$

$$= -\sum_{i_1=1}^{m}\sum_{i_2=1}^{m}\cdots\sum_{i_n=1}^{m} \Pr(X_1 = \alpha_{i_1})\Pr(X_2 = \alpha_{i_2})\cdots\Pr(X_n = \alpha_{i_n}) \bullet \sum_{k=1}^{n}\log \Pr(X_k = \alpha_{i_k})$$

$$= -n\sum_{i=1}^{m} \Pr(X = \alpha_i)\log \Pr(X = \alpha_i)$$

$$H(S) = \lim_{n\to\infty} \frac{G_n}{n} = -\sum_{i=1}^{m} \Pr(X = i)\log \Pr(X = i)$$

If the condition of iid is assumed, rather than a given fact, then the above $H(S)$ is called 1st order entropy.

## Source Entropy

- True source entropy
  - defined over the true probability space (and the true probability measure of the source, structure included); a characteristic quantity of the source.
- Estimated source entropy
  - Source distribution is usually not completely or precisely known (particularly in sequences resulted from non-independent combined experiments);
  - Source entropy is normally calculated using an estimated source distribution with certain assumed conditions;
  - A “better” distribution estimate often leads to lower estimated entropy.