ECE 5654 Lecture 13
Information Theory & Channel Capacity

Ravi Tandon

Virginia Tech

March 24, 2014

Learning Objectives

At the end of this lecture, the student should be able to:

- Describe Entropy, Mutual Information and provide its definition
- Describe the concept of capacity and how it relates to digital communications and define it for the binary symmetric channel (BSC) and the AWGN channel

Mutual Information

- Let us consider two random variables X and Y (with possible outcomes in alphabets $\mathcal{X}$ and $\mathcal{Y}$)
- We wish to define a measure of the information which Y tells us about the value of X
  - Clearly, if X and Y were statistically independent, then we would expect the event $Y = y$ to provide no information about the occurrence of the event $X = x$
  - On the other hand, if X and Y were fully dependent, then the event $Y = y$ exactly determines the event $X = x$
- Based on this intuition, we define a measure as the logarithm of the ratio of the conditional probability

$$p(X = x|Y = y) = p(x|y)$$

to the probability

$$p(X = x) = p(x)$$

Mutual Information

- That is, the information content provided by occurrence of the event $Y = y$ about the event $X = x$ is

$$I(x; y) = \log \frac{p(x|y)}{p(x)}$$

- $I(x; y)$ is called the mutual information between the events x and y
- Using this, we define the mutual information between random variables X and Y as

$$I(X; Y) = \sum_{(x,y)} p(x, y) I(x; y) = \sum_{(x,y)} p(x, y) \log \left(\frac{p(x|y)}{p(x)} \right)$$

$$= \sum_{(x,y)} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right)$$

Properties of MI

- Units of MI are determined by base of logarithm.
  - if $\log_2$, $I(X;Y)$ measured in bits
  - if $\log_e$, $I(X;Y$ measured in nats (natural units)
- Properties of Mutual Information
  - Symmetry: $I(X;Y) = I(Y;X)$ (directly verifiable by its definition)
  - Non-negativity: $I(X;Y) \geq 0$, with equality if and only if X, Y are independent
  - $I(X;Y) \leq \min\{|\mathcal{X}|, |\mathcal{Y}|\}$

  where $\mathcal{X}$ denotes the size of the alphabet of X

Entropy

- If X is identical to Y, then the mutual information is:

$$\begin{aligned} I(X;Y) &= \sum_{(x,y)} p(x,y) \log\left(\frac{1}{p(x)}\right) \\ &= -\sum_{x} p(x) \log(p(x)) \\ &= H(X) \end{aligned}$$

- $H(X)$ is termed as the entropy of the random variable X
- $H(X)$ is a measure of the uncertainty (or intrinsic randomness) in X
- NOTE: We use the convention $0 \times \log(0) = 0$ in this definition
- $0 \leq H(X) \leq \log(|\mathcal{X}|)$
- Suggested reading: Section 6.2, 6.3 (Proakis Salehi, 4th Edition)

Joint Entropy, Conditional Entropy

- The joint entropy of X and Y is defined as

$$H(X,Y) = -\sum_{(x,y)} p(x,y)\log(p(x,y))$$

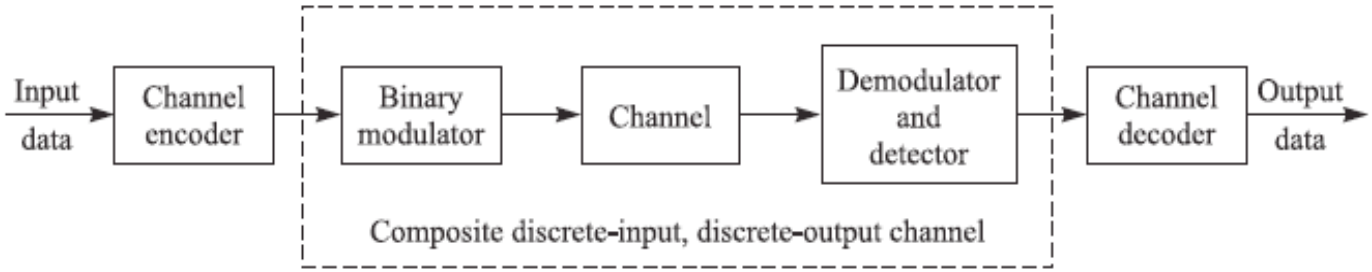
- The conditional entropy of X given Y is defined as

$$H(X|Y) = -\sum_{x}\sum_{y} p(x,y)\log(p(y|x))$$

- Mutual information can be related to entropy and conditional entropy as

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}$$

Channel Models



- Channel encoder introduces redundancy (in a controlled manner)
- Take k bits and map to n bits (Rate $= R = k/n$)
- Output of channel encoder is fed to modulator
- Passes through the channel and then demodulator
- Channel decoder then exploits the redundancy to correct channel disturbances

Channel Models (Continued)

- Channel Models provide a mathematical model to relate the transmitted (random) symbols and the observed (random) symbols/values
- Using Information theory, we can examine the information provided to the receiver by the observation about the channel input
- Let us the define the input symbols by the set of values $\mathcal{X}$ and the output values by the set $\mathcal{Y}$
- Further define a series of inputs and received (observed) outputs using the vectors $\mathbf{x}$ and $\mathbf{y}$

Discrete Memoryless Channels (DMCs)

- A memoryless channel is one for which we have

$$p(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n} p(y_i|x_i)$$

- This describes the probability of receiving $\mathbf{y}$ given that $\mathbf{x}$ is transmitted
- Binary Symmetric Channel (BSC)
  - Simplest memoryless channel
  - Two (possible) inputs, two (possible) outputs
- General discrete memoryless channel
  - $|\mathcal{X}|$ possible inputs
  - $|\mathcal{Y}|$ possible inputs
  - Often different values

Discrete-Time AWGN Channel

- Let $\mathcal{X} = \mathcal{Y} = \mathcal{R}$, where $\mathcal{R}$ is the set of reals
- The discrete time AWGN channel is

$$y_i = x_i + n_i$$

  where n_i are zero mean Gaussian random variables with variance σ^2
- It is usually assumed that

$$E[X^2] \leq P$$

- Also note that x_i is a continuous valued random variable

The AWGN Waveform Channel

- Inputs and outputs are waveforms
- Channel has a given bandwidth W, $C(f) = 1$ in $[-W, W]$

$$y(t) = x(t) + n(t)$$

where $n(t)$ is sample function of the AWGN process with PSD $N_0/2$

- Power constraint: $E[X(t)^2] \leq P$
- We can convert the waveform channel to a discrete time AWGN channel using an appropriate set of basis functions

$$y(t) = \sum_j y_j \phi_j(t)$$

$$x(t) = \sum_j x_j \phi_j(t)$$

$$n(t) = \sum_j n_j \phi_j(t)$$

Channel Capacity

- Let X represent the transmitted symbol at the input to a channel and let Y represent the symbol received at the output of a channel.
- The channel is described by the transition probabilities: $\{p(y|x)\}$, defined for each (x, y) pair
- Then, we define a quantity called Channel Capacity:

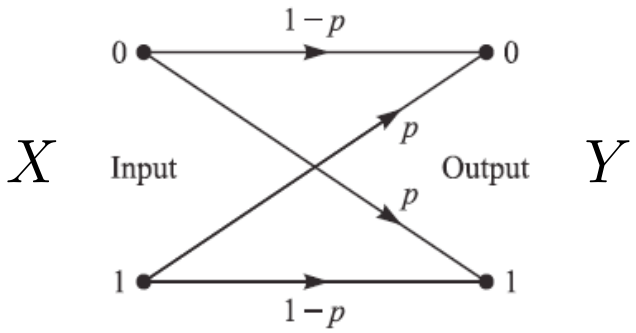
$$C = \max_{p(x)} I(X;Y)$$

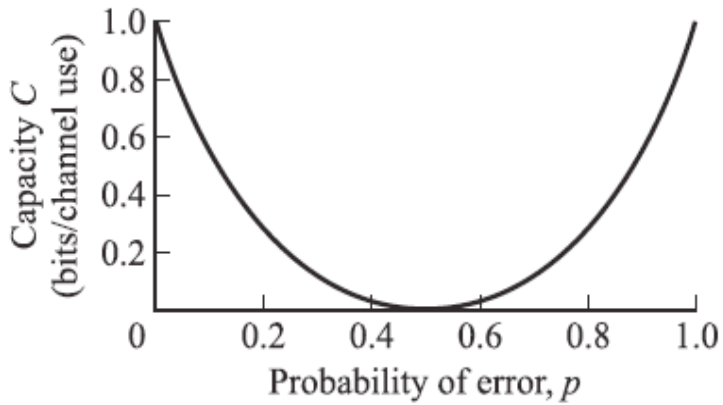
  where $I(X;Y)$ is the mutual information between X and Y
- Shannon's Noisy Channel Coding Theorem states that reliable communication is possible over a discrete memoryless channel if and only if the communication rate R satisfies $R < C$, where C is the channel capacity.

Channel Capacity

- Channel capacity C has units of information rate (bits/sec or bits/sec/Hz)
- We are interested in designing a system to operate as close to this fundamental limit as possible

Capacity of Binary Symmetric Channel (BSC)





- p = probability of error for a binary modulation scheme
- Shannon's theorem tells us that

$$C = 1 + p\log_2(p) + (1-p)\log_2(1-p) \text{ bits/symbol}$$

- Note that error free transmission is possible even for non zero values of p (however at a smaller rate)
- Channel capacity is symmetric around $p = 0.5$, where no information can be transmitted

Capacity of Bandlimited AWGN Channel with an Input Power Constraint

- Consider a band limited AWGN channel where the waveform $x(t)$ is the input and the waveform $y(t)$ is the output

$$y(t) = x(t) + n(t)$$

- We know that we can express these waveforms in terms of $N = 2WT$ orthonormal basis functions $\phi_i(t)$ where W is the bandwidth and T is the time interval
- The equivalent channel model can be expressed as $y_i = x_i + n_i$, $i = 1, 2, \ldots, N$
- The channel conditional probability can then be written as:

$$p(y_1, \ldots, y_N | x_1, \ldots, x_N) = \prod_{i=1}^{N} p(y_i|x_i) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i}} e^{-\frac{(y_i - x_i)^2}{2\sigma_i^2}}$$

Capacity (continued)

- The capacity is defined as

$$C = \lim_{T\to\infty} \max_{p(x)} \frac{1}{T} I(X;Y)$$

- Further, the mutual information is defined as

$$\begin{aligned} I(X_N;Y_N) &= \int_{\mathbf{x}_N} \int_{\mathbf{y}_N} p(\mathbf{y}_N, \mathbf{x}_N) \log\left(\frac{p(\mathbf{y}_N|\mathbf{x}_N)}{p(\mathbf{x}_N)}\right) dx_N dy_N \\ &= \sum_{i=1}^{N} \int_{x_i} \int_{y_i} p(y_i, x_i) \log\left(\frac{p(y_i|x_i)}{p(x_i)}\right) dx_i dy_i \end{aligned}$$

- This is maximized when X_i is Gaussian, i.e., $p(x_i) = \frac{1}{\sqrt{2\pi\sigma_x}} e^{-x_i^2/2\sigma_x^2}$

Capacity (continued)

- Substituting, we obtain

$$\max_{p(\mathbf{x})} I(X_N; Y_N) = \sum_{i=1}^{N} \frac{1}{2} \log\left(1 + \frac{2\sigma_x^2}{N_0}\right) = \frac{N}{2} \log\left(1 + \frac{2\sigma_x^2}{N_0}\right)$$
$$= WT \log\left(1 + \frac{2\sigma_x^2}{N_0}\right)$$

- Average Power constraint:

$$P_{avg} = \frac{1}{T} \int_0^T E[x^2(t)]dt = \frac{1}{T} \sum_{i=1}^{N} E(x_i^2) = \frac{N\sigma_x^2}{T} = 2W\sigma_x^2$$

- Substituting, we get

$$C = W \log\left(1 + \frac{P_{avg}}{WN_0}\right) \text{ bits/sec}$$

Behavior of Channel Capacity

$$C = W \log \left(1 + \frac{P}{WN_0}\right) \text{ bits/sec}$$

- Impact of power, P
  - For fixed W, C increases as P increases, and $C \to \infty$ as $P \to \infty$
  - The rate of increase is logarithmic in P
  - I.e, capacity increases very slowly by increasing power
  - But it can grow indefinitely

Behavior of Channel Capacity

$$C = W \log\left(1 + \frac{P}{WN_0}\right) \text{ bits/sec}$$

- Impact of bandwidth, W
  - Increasing W, plays a dual role
  - Clearly, it causes capacity to increase as higher W implies more transmissions over the channel per unit time
  - However, increasing W also decreases the SNR (defined by $\frac{P}{WN_0}$)
  - If we take the limit $W \to \infty$, we get

$$C_\infty = \lim_{W\to\infty} W \log\left(1 + \frac{P}{WN_0}\right) = (\log_2 e)\frac{P}{N_0} \approx 1.44\frac{P}{N_0} bits/sec$$

  - For a fixed P, capacity cannot increase indefinitely by increasing W

Can Modulation Alone Achieve Channel Capacity ?

- Standard modulation falls well short of channel capacity
- If we use a very large dimension for our signal constellation, it may be possible to achieve channel capacity
- E.g., M-ary FSK achieves channel capacity but it is very bandwidth inefficient

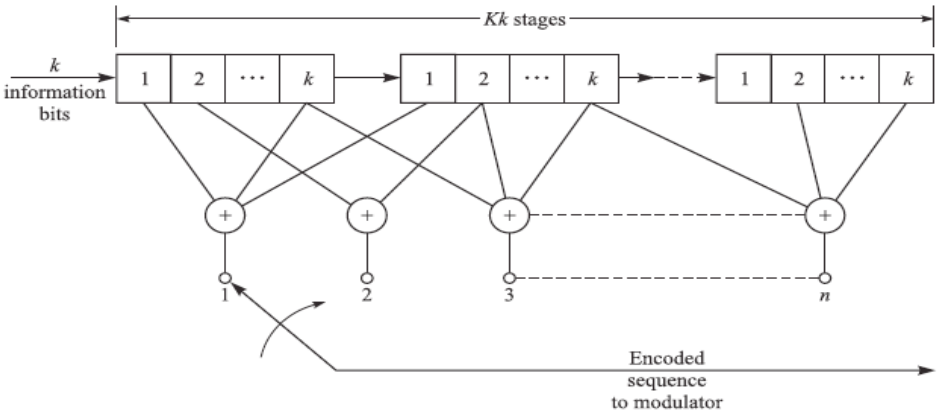
Channel Codes

- Major classes of channel codes:
- Block codes
- Convolutional codes
- Trellis codes
- Turbo codes
- Concatenated codes
- ...

Block Codes

- One of $M = 2^k$ messages (i.e., a binary sequence of length k) is mapped to a sequence of length n (codeword)
- We have $n > k$, and this introduces redundancy
- The codeword sequence is transmitted over the channel (for e.g., using BPSK, QPSK, BFSK etc)
- Block coding is memoryless
- Current codeword only depends on the current k bits and independent of all past codewords

Convolutional Codes



- Described in terms of finite-state machines
- At each time instance i, k information bits enter the encoder
- n binary symbols are generated by the encoder and the state of the encoder changes from σ_{i-1} to σ_i
- Set of possible states is finite and denoted by Σ
- The n binary symbols generated and the next state σ_i depend on k information bits and the past state σ_{i-1}

Code Rate

- The Code rate is defined by

$$R_c = \frac{k}{n}$$

- Codeword of length n is sent using N-dim constellation of size M
- $L = \frac{n}{\log_2(M)}$ is the # of symbols per codeword
- If symbol duration is T_s, then transmission time for k bits is LT_s and transmission rate is

$$R = \frac{k}{LT_s} = \frac{k}{n} \times \frac{\log_2(M)}{T_s} = R_c \frac{\log_2(M)}{T_s}$$

- Minimum required transmission bandwidth is

$$W = \frac{N}{2T_s} = \frac{RN}{2R_c \log_2(M)}$$

Code Rate

- The Code rate is defined by

$$R_c = \frac{k}{n}$$

- Minimum required transmission bandwidth is

$$W = \frac{N}{2T_s} = \left(\frac{RN}{2\log_2(M)}\right) \times \frac{1}{R_c}$$

- Resulting spectral bit rate

$$r = \frac{R}{W} = \left(\frac{2\log_2(M)}{N}\right) R_c$$

- Comparing to an uncoded system ($n = k$), for a coded system:
  - bit rate, spectral bit rate are decreased
  - bandwidth is increased

Typical Values

- For practical block codes, $20 \leq n \leq 100$ is typical
- For practical block codes, $\frac{1}{4} \leq r \leq 1$ is typical
- We will refer to an (n, k) block code

(n, k) Block Codes

- n is also known as code length; k is also known as code size
- In addition to length and size, we are also interested in quantifying how much the codewords differ from one another
- Let F be an alphabet (elements of codewords are from this alphabet)
- The Hamming distance between two words $\mathbf{x}, \mathbf{y} \in F^n$ is the number of coordinates on which $\mathbf{x}$ and $\mathbf{y}$ differ.
- We refer to the Hamming distance by $d(\mathbf{x}, \mathbf{y})$
- Hamming distance satisfies the following properties:
  - $d(\mathbf{x}, \mathbf{y}) \geq 0$, with equality iff $\mathbf{x} = \mathbf{y}$
  - Symmetry: $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$
  - Triangle inequality: $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$

Hamming Weight & Minimum Distance

- For any $\mathbf{x} \in F^n$, the Hamming weight of $\mathbf{x}$ is the number of non-zero entries in $\mathbf{e}$
- We denote the Hamming weight by $w(\mathbf{x})$
- It then follows that the Hamming distance for any two $\mathbf{x}, \mathbf{y}$ is then

$$d(\mathbf{x}, \mathbf{y}) = w(\mathbf{x} - \mathbf{y})$$

- Minimum Distance of a Code (denoted in short by d) is the minimum Hamming distance between any two distinct codewords, i.e.,

$$d = \min_{c_1 \neq c_2, c_1, c_2 \in \mathcal{C}} w(c_1 - c_2)$$

- Sometimes, we also append the minimum distance and write the code as an (n, k, d) block code

Examples

- The binary $(n, k, d) = (3, 1, 3)$ repetition code:

$$0 \rightarrow 000$$
$$1 \rightarrow 111$$

  rate $= 1/3$, minimum distance $= 3$

- The binary $(n, k, d) = (3, 2, 2)$ parity code:

$$00 \rightarrow 000$$
$$01 \rightarrow 011$$
$$10 \rightarrow 101$$
$$11 \rightarrow 110$$

  rate $= 2/3$, minimum distance $= 2$

- Next Lecture, we will talk about Linear Block Codes
- Encoding/Decoding of Linear Block Codes
- Probability of error comparisons (coded vs uncoded)