Lecture 14: Information Theory

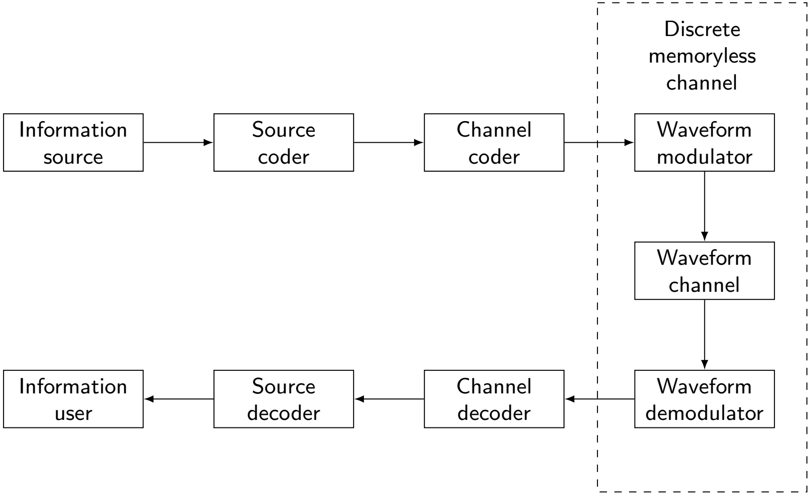
Prof. Deniz Gunduz

Department of Electrical & Electronic Engineering
Imperial College London

Outline

- Introduction to information theory
- Discrete memoryless source (DMS) and source entropy
- Discrete memoryless channel (DMC) and conditional entropy
- Mutual information and channel coding theorem
- Binary symmetric channel (BSC) and additive white Gaussian noise (AWGN) channel capacities
- Reference
  - ✓ [Haykin] Chapter 10

Model of a Digital Communication System



What is Information?

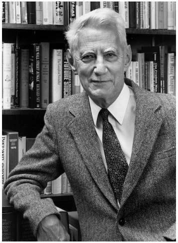
- Information: any new knowledge about something
  - ✓ How to store information efficiently?
  - ✓ How to transmit information over noisy channels?
- Information is everywhere
  - ✓ Collected by sensory system, transmitted via nervous system, processed in brain, . . .
  - ✓ Stored in DNA, in hard-drives, in books, . . .
  - ✓ Transmitted over the phone line, over the air, over generations, . . .
- How to quantify information?

What is Information Theory?

- C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, 1948

  "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point."

- Two fundamental questions in information theory:
  - ✓ ultimate limit on data compression? (source coding)
  - ✓ ultimate transmission rate of reliable communication over noisy channels? (channel coding)



What about non-uniform distribution?

Example: Almost all students at Imperial are smart

- Event that an imperial student is smart: *not so informative*
- Event that an imperial student is not smart: *very informative*

- Messages containing knowledge of a *high* probability of occurrence $\Rightarrow$ Not very informative
- Messages containing knowledge of *low* probability of occurrence $\Rightarrow$ More informative
- A small change in the probability of a certain output should not change the information delivered by that output by a large amount (it seems like a continuous function of the probability distribution)

Definition

- Amount of information in a symbol s with probability p:

$$I(s) = \log \frac{1}{p}$$

- Properties
  - ✓ $p = 1 \Rightarrow I(s) = 0$: a deterministic symbol contains no information
  - ✓ $0 \leq p \leq 1 \Rightarrow 0 \leq I(s) \leq \infty$: information measure is monotonic and non-negative
  - ✓ $p = p_1 \times p_2 \Rightarrow I(s) = I(s_1) + I(s_2)$: information from statistically independent events is additive

- Logarithm base 2 is commonly used, resulting in bits

Example

Discrete Memoryless Source (DMS)

- Suppose we have an information source emitting a sequence of symbols from a finite alphabet:

$$\mathcal{S} = \{s_1, s_2, \ldots, s_N\}$$

- Discrete memoryless source: The successive symbols are statistically independent and identically distributed (i.i.d.)

- Example: $\mathcal{S} = \{0, 1\}$, symbol sequence $= 001011000110\ldots$

- Assume that each symbol has probability p_n for $n = 1, \ldots, N$, such that $\sum_{n=1}^{N} p_n = 1$

Source Entropy

- We know that if symbol s_n has occurred, this corresponds to amount of information,

$$I(s_n) = \log_2 \frac{1}{p_n} = -\log_2 p_n \text{ bits of information}$$

- For random variable $S \in \mathcal{S}$, expected value of $I(S)$ over the source alphabet

$$\mathbb{E}\{I(S)\} = \sum_{n=1}^{N} p_n I(s_n) = -\sum_{n=1}^{N} p_n \log_2 p_n$$

- Source entropy: average amount of information per source symbol

$$H(S) = -\sum_{n=1}^{N} p_n \log_2 p_n$$

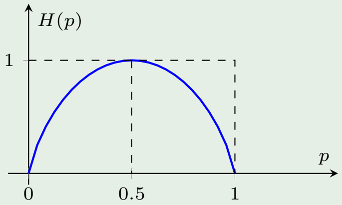
- Units: bits/symbol

Meaning of Entropy

- What does *entropy* tell us about the source?
- It is the amount of uncertainty before we receive it
- It tells us how many bits of information per symbol we get on the average by learning the source realization
- Relation with thermodynamic entropy
  - ✓ In thermodynamics: entropy measures disorder and randomness
  - ✓ In information theory: entropy measures uncertainty

Example

Entropy of a Binary Source



Discrete Memoryless Channel (DMC)

- Input alphabet: $\mathcal{X} = \{x_0, x_1, \ldots, x_{J-1}\}$
- Output alphabet: $\mathcal{Y} = \{y_0, y_1, \ldots, y_{K-1}\}$
- Transition probabilities (characterizing channel):

$$p(y_k|x_j) = P(Y = y_k|X = x_j), \quad \forall j, k$$

- Input probability distribution:

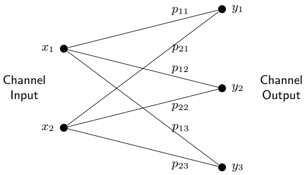
$$p(x_j) = P(X = x_j), \quad \forall j$$

- Joint probability distribution:

$$p(x_j, y_k) = p(y_k|x_j)p(x_j), \quad \forall j, k$$

- Marginal distribution of the channel output:

$$p(y_k) = P(Y = y_k) = \sum_{j=0}^{J-1} p(y_k|x_j)p(x_j), \quad \forall k$$



$$p_{jk} = p(y_k|x_j)$$

Conditional Entropy

- Conditional entropy:

$$H(X|Y=y_k) = \sum_{j=0}^{J-1} p(x_j|y_k)\log_2 \frac{1}{p(x_j|y_k)}$$

- Probability of $H(X|Y=y_k)$:

$$\begin{array}{cccc} H(X|Y=y_0) & H(X|Y=y_1) & \dots & H(X|Y=y_{K-1}) \\ \downarrow & \downarrow & & \downarrow \\ p(y_0) & p(y_1) & \dots & p(y_{K-1}) \end{array}$$

- Average entropy:

$$\begin{aligned} H(X|Y) &= \sum_{k=0}^{K-1} H(X|Y=y_k)p(y_k) = \sum_{k=0}^{K-1}\sum_{j=0}^{J-1} p(x_j|y_k)p(y_k)\log_2 \frac{1}{p(x_j|y_k)} \\ &= \sum_{k=0}^{K-1}\sum_{j=0}^{J-1} p(x_j, y_k)\log_2 \frac{1}{p(x_j|y_k)} \end{aligned}$$

- Interpretation: amount of uncertainty after observing the channel output

Mutual Information

- Mutual information $I(X;Y)$: the uncertainty resolved by observing channel output

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

- Also,

$$\begin{aligned} I(X;Y) &= \sum_{j=0}^{J-1} \sum_{k=0}^{K-1} p(x_j, y_k) \log_2 \frac{p(y_k|x_j)}{p(y_k)} \\ &= \sum_{k=0}^{K-1} \sum_{j=0}^{J-1} p(x_j, y_k) \log_2 \frac{p(x_j, y_k)}{p(x_j)p(y_k)} \end{aligned}$$

- Mutual information is
  - ✓ Non-negative: $I(X;Y) \geq 0$
  - ✓ Symmetric: $I(X;Y) = I(Y;X)$

- For a given channel $p(y_k|x_j)$ for $x_1, \ldots, x_J$ and $y_1, \ldots, y_K$, $I(X;Y)$ depends on $p(x_j)$ for $x_1, \ldots, x_J$

Channel Capacity and Coding Theorem

- Capacity of a discrete memoryless channel is the maximum mutual information between the input and output, where the maximization is over all possible input probability distributions

$$C = \max_{p(x_0),\ldots,p(x_{J-1})} I(X;Y)$$

- How to calculate?
  - ✓ usually very complicated if analytically, except some symmetrical cases
  - ✓ easily to calculate numerically

Channel Coding Theorem

- If the transmission rate $R \leq C$, then there exists a coding scheme such that R bits per channel use can be transmitted over the channel with an arbitrarily small probability of error.
- Conversely, if $R > C$, error probability is always bounded above zero when the transmission rate is above the capacity.

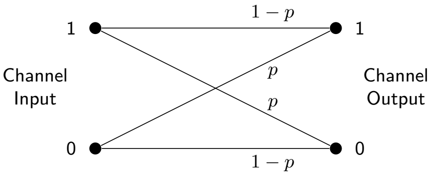
- How to code? We only know its existence but do not know how.
  - ✓ Polar code is the first code with an explicit construction to provably achieve the channel capacity for *symmetric binary-input*, discrete, memoryless channels (B-DMC) with polynomial dependence on the gap to capacity.

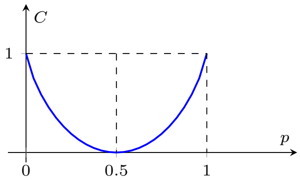
- Capacity of BSC

$$C = \max_{p(x_0),p(x_1)} I(X;Y) = 1 - h(p)$$

where

$$h(p) = -p\log_2 p - (1-p)\log_2(1-p)$$





Additive White Gaussian Noise (AWGN) channel

- Capacity of an additive white Gaussian noise (AWGN) channel:

$$C = B\log_2(1+\mathsf{SNR}) = B\log_2\left(1+\frac{P}{N_0 B}\right) \text{ bps}$$

  - ✓ B: bandwidth of the channel
  - ✓ P: average signal power at the receiver
  - ✓ N_0: single-sided PSD of noise

- How can we achieve this rate?
  - ✓ Design powerful error correcting codes to correct as many errors as possible
  - ✓ Use good modulation schemes that do not lose information in the detection process
  - ✓ No simple way!

Example

Capacity of bandwidth-limited AWGN channel

Note