

Lecture 21

Instructor: Sruthi Sekar

Scribe: Aakarsh Chaudhary

1 Information Theory

Lecture Overview

This lecture establishes the foundations of information theory, introducing entropy as a measure of uncertainty and mutual information as a measure of dependence between variables. In future lectures we'll connect the concepts learned in this lecture with communication complexity.

1.1 Shannon's Entropy

Example 1 (Wordle as Information Partitioning). *In Wordle, we guess a hidden 5-letter word by sequentially testing letters. Each guess partitions the set of possible words:*

- **First guess 's'**: Partitions words into:

- Set A: Contains 's' (probability p)
- Set B: No 's' (probability $1 - p$)

To specify which set contains the target word, we need $\log_2 2 = 1$ bit.

- **Second guess 't'**: Further partitions into 4 subsets:

Subset	Description
A_1	Contains both 's' and 't'
A_2	Contains 's' but not 't'
B_1	Contains 't' but not 's'
B_2	Contains neither

Now requiring $\log_2 4 = 2$ bits to identify the subset.

- **General Case**: For N equally likely partitions:

- Probability $p = 1/N$ for each partition
- Information needed: $\log_2 N = \log_2(1/p)$ bits

This motivates our formal definition of information content:

Definition 1 (Information Content). *For an event occurring with probability p , its **information content** is:*

$$I(p) = \log_2 \left(\frac{1}{p} \right) \text{ bits}$$

Observation 2. *The partitioning argument shows why logarithmic measure is natural:*

$$\boxed{A_1 \quad A_2 \quad B_1 \quad B_2} \longrightarrow \log_2 4 = 2 \text{ bits}$$

Definition 3 (Shannon's Entropy). For a discrete random variable X with PMF $p(x)$, entropy quantifies expected information:

$$H(X) = \mathbb{E}[I(X)] = \sum_x p(x) \log_2 \left(\frac{1}{p(x)} \right)$$

Note: Using Shannon's Coding Theorem, we can actually show that the minimum number of average bits needed to encode outcomes of X is $H(X)$

Example 2. Encode colors with given probabilities:

Color	Probability	Encoding
red	1/2	0
green	1/4	10
blue	1/8	110
yellow	1/8	111

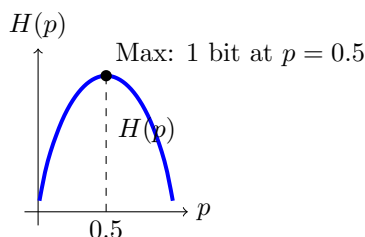
Calculating entropy:

$$H(X) = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3 = 1.75 \text{ bits}$$

Example 3 (Binary Entropy Function). For a Bernoulli random variable with $P(X = 1) = p$:

$$H(p) = p \log_2 \left(\frac{1}{p} \right) + (1 - p) \log_2 \left(\frac{1}{1 - p} \right)$$

Note: The entropy is 0 only when p is either 0 or 1



1.2 Properties of Entropy with Proofs

Theorem 4 (Entropy Properties). 1. **Non-negativity:** $H(X) \geq 0$ with equality iff X is deterministic.

Proof. Since $0 \leq p(x) \leq 1$, $\log_2(1/p(x)) \geq 0$. Equality occurs when one $p(x) = 1$ and others are 0. □

2. **Upper bound:** For N outcomes, $H(X) \leq \log_2 N$ with equality iff uniform.

Proof. By Jensen's inequality, since log is concave:

$$H(X) = \mathbb{E}[\log(1/p(x))] \leq \log(\mathbb{E}[1/p(x)]) = \log N$$

□

3. **Additivity:** For independent X, Y : $H(X, Y) = H(X) + H(Y)$

Proof. $p(x, y) = p(x)p(y) \Rightarrow \log(1/p(x, y)) = \log(1/p(x)) + \log(1/p(y))$

since, $\sum_{x,y} p(x)p(y) \log_2(1/p(x)) = \sum_x p(x) \log_2(1/p(x))$

$\therefore H(X, Y) = H(X) + H(Y)$

□

1.3 Mutual Information and Properties

Definition 5 (Mutual Information). For random variables X and Y :

$$I(X; Y) = \sum_{x,y} p(x,y) \log_2 \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

Theorem 6 (Properties of Mutual Information). 1. **Non-negativity:** $I(X; Y) \geq 0$ with equality iff $X \perp Y$

Proof. Follows from KL divergence non-negativity. □

2. **Symmetry:** $I(X; Y) = I(Y; X)$

3. **Relation to Entropy:**

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

4. **Chain Rule:**

$$I(X; Y, Z) = I(X; Y) + I(X; Z|Y)$$

5. **Conditional MI:**

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

6. **Example:** For independent X, Y and $Z = X \oplus Y$:

$$I(X; Y) = 0 \text{ but } I(X; Y|Z) = 1$$

1.4 Information Complexity

Say, Alice and Bob are communicating with their messages $f(x,y)$, where x and y are taken from a joint distribution $\mu_{(x,y)}$.

$D_\epsilon^\mu(f)$: minimum communication cost over all deterministic protocols π such that-

$$\text{err}_\mu = P_{(x,y) \sim \mu_{(x,y)}}[\pi(x,y) \neq f(x,y)] \leq \epsilon$$

Definition 7 (Information Complexity). For a protocol π computing $f(x,y)$ over input distribution μ :

$$IC_\mu(\pi) = I(T_\pi; Y|X) + I(T_\pi; X|Y)$$

where T_π is the transcript. This measures the information revealed about inputs.

In the next class we'll see-

- $IC_q^\mu(f) = \min_\pi \{IC_\mu(\pi) : \text{error}(\pi) \leq \epsilon\}$
- $D_\epsilon^\mu(f) \geq IC_\epsilon^\mu(f)$