The first half of this lecture covers the core concepts of information theory, the study of the quantification, storage, and communication of information.

1 Information Theory Definitions

First, we provide the definitions of several core concepts in information theory, along with some notable facts claims regarding these concepts.

1.1 Discrete Distributions

$p$ is a discrete distribution over a finite support of size $n$ if:

- $p = (p_1, p_2, ..., p_n)$
- $p_i \in [0, 1]$ for all $i \in [n]$
- $\sum_i p_i = 1$

$X$ is a random variable with distribution $p$ if $\Pr[X = i] = p_i$.

1.2 Entropy

**Definition** (Entropy). $H(X) = \sum_i p_i \log_2 \left( \frac{1}{p_i} \right)$

Intuitively, entropy $H(X)$ is a measurement of the uncertainty of $X$. It has the following characteristics:

- If $p_i = 0$, then $p_i \log_2 \left( \frac{1}{p_i} \right) = 0$.
- $H(X) \leq \log_2 n$. Equality holds when $p_i = \frac{1}{n}$ for all $i$.
- If $B$ is a bit with bias $p$, then

$$H(B) = p \log_2 \frac{1}{p} + (1 - p) \log_2 \frac{1}{1 - p}$$
1.3 Conditional and Joint Entropy

**Definition** (Conditional Entropy). \( H(X \mid Y) = \sum_y H(X \mid Y = y) \Pr[Y = y] \)

**Definition** (Joint Entropy). \( H(X, Y) = \sum_{x,y} \Pr[(X, Y) = (x, y)] \log(\frac{1}{\Pr[(X,Y)=(x,y)]}) \)

**Claim 1** (Chain Rule). \( H(X,Y) = H(X) + H(Y \mid X) \)

**Proof.**

\[
H(X,Y) = \sum_{x,y} \Pr[(X,Y) = (x,y)] \log\left(\frac{1}{\Pr[(X,Y) = (x,y)]}\right) \quad \text{(By definition)}
\]

\[
= \sum_{x,y} \Pr[X = x] \Pr[Y = y \mid X = x] \log\left(\frac{1}{\Pr[X = x] \Pr[Y = y \mid X = x]}\right) \quad \text{(By chain rule for probabilities)}
\]

\[
= \sum_{x,y} \Pr[X = x] \Pr[Y = y \mid X = x] \left( \log\left(\frac{1}{\Pr[X = x]}\right) + \log\left(\frac{1}{\Pr[Y = y \mid X = x]}\right) \right)
\]

\[
= H(X) + H(Y \mid X) \quad \text{(By definition)}
\]

■

2
Claim 2 (Conditioning Cannot Increase Entropy). Let $X$ and $Y$ be random variables. Then $H(X \mid Y) \leq H(X)$.

Proof. For this proof, we need Jensen’s inequality:

Let $f$ be a continuous, concave function, and let $p_1, \ldots, p_n$ be non-negative reals that sum to 1. For any $x_1, \ldots, x_n$,

$$\sum_{i=1,\ldots,n} p_if(x_i) \leq f(\sum_{i=1,\ldots,n} p_ix_i)$$

Then $H(X \mid Y) - H(X) = \sum_{x,y} \Pr[Y = y] \Pr[X = x \mid Y = y] \log \left( \frac{1}{\Pr[X = x \mid Y = y]} \right)$

$$- \sum_x \Pr[X = x] \log \left( \frac{1}{\Pr[X = x]} \right)$$

$$= \sum_{x,y} \Pr[Y = y] \Pr[X = x \mid Y = y] \log \left( \frac{1}{\Pr[X = x \mid Y = y]} \right)$$

$$- \sum_x \Pr[X = x] \log \left( \frac{1}{\Pr[X = x]} \right) \sum_y \Pr[Y = y \mid X = x]$$

$$= \sum_{x,y} \Pr[X = x, Y = y] \log \left( \frac{\Pr[X = x]}{\Pr[X = x \mid Y = y]} \right)$$

$$= \sum_{x,y} \Pr[X = x, Y = y] \log \left( \frac{\Pr[X = x] \Pr[Y = y]}{\Pr[(X, Y) = (x, y)]} \right)$$

$$\leq \log(\sum_{x,y} \Pr[X = x, Y = y] \cdot \frac{\Pr[X = x] \Pr[Y = y]}{\Pr[(X, Y) = (x, y)]}) \quad \text{(By Jensen’s inequality)}$$

$$= \log(\sum_{x,y} \Pr[X = x] \Pr[Y = y])$$

$$= 0$$

Equality holds when $X$ and $Y$ are independent.

1.4 Mutual Information

Definition (Mutual Information). $I(X ; Y) = H(X) - H(X \mid Y)$

Note that $I(X ; X) = H(X) - H(X \mid X) = H(X)$

Definition (Conditional Mutual Information). $I(X ; Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z)$

This raises the question. Does conditioning on $Z$ increase or decrease the mutual information of $X$ and $Y$? It turns out that both can be true.
Claim 3. For certain $X, Y, Z$, we can have $I(X ; Y \mid Z) \leq I(X ; Y)$

Proof. Consider $X = Y = Z$. Then,

- $I(X ; Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z) = 0 - 0 = 0$
- $I(X ; Y) = H(X) - H(X \mid Y) = H(X) - 0 = H(X)$

Intuitively, $Y$ only reveals information that $Z$ already revealed, and we are conditioning on $Z$ being revealed.

Claim 4. For certain $X, Y, Z$, we can have $I(X ; Y \mid Z) \geq I(X ; Y)$

Proof. Consider $X = Y + Z \mod 2$, where $X$ and $Y$ are uniform in $\{0, 1\}$ Then,

- $I(X ; Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z) = 1 - 0 = 1$
- $I(X ; Y) = H(X) - H(X \mid Y) = 1 - 1 = 0$

Intuitively, $Y$ only reveals useful information about $X$ after also conditioning on $Z$.

Claim 5 (Chain Rule for Mutual Information). $I(X, Y ; Z) = I(X ; Z) + I(Y ; Z \mid X)$

Proof.

\[
I(X, Y ; Z) = H(X, Y) - H(X, Y \mid Z) \\
= H(X) + H(Y \mid X) - H(X \mid Z) - H(X, Y \mid Z) \\
= I(X ; Z) + I(Y ; Z \mid X)
\]

By induction, it follows that

\[
I(X_1, X_2, ..., X_n ; Z) = \sum_{i} I(X_i ; Z \mid X_1, ..., X_{i-1})
\]
2 Proving Fano’s Inequality

Fano’s Inequality is as follows:

For any estimator \( X' : X \to Y \to X' \) with \( P_e = \Pr[X' \neq X] \), where \( X \to Y \to X' \) is a Markov Chain, that is, \( X' \) and \( X \) are independent given \( Y \), we have the following:

\[
H(X \mid Y) \leq H(P_e) + P_e \cdot \log(|X| - 1)
\]

To prove Fano’s Inequality, we need to use the data processing inequality.

Claim 6 (Data Processing Inequality). Suppose \( X \to Y \to Z \) is a Markov Chain. Then,

\[
I(X ; Y) \geq I(X ; Z)
\]

That is, no clever combination of the data can improve our estimation of \( X \).

Proof. Note that \( I(X ; Y \mid Z) = I(X ; Z) + I(X ; Y \mid Z) = I(X ; Y) + I(X ; Z \mid Y) \). Thus, it suffices to show that \( I(X ; Z \mid Y) = 0 \), since we know that \( I(X ; Y \mid Z) \geq 0 \).

\[
I(X ; Z \mid Y) = H(X \mid Y) - H(X \mid Y, Z).
\]

By the Markov Chain requirement, given \( Y \), \( X \) and \( Z \) are independent. Thus, \( H(X \mid Y, Z) = H(X \mid Y) \).

If follows that \( I(X ; Z \mid Y) = 0 \).

Now, we can proceed with the proof for Fano’s Inequality.

Let \( E = 1 \) if \( X' \neq X \), and \( E = 0 \) otherwise. It is an indicator variable of whether we have an error on estimating \( X \).

\[
H(E, X \mid X) = H(X \mid X') + H(E \mid X, X') \tag{By chain rule}
\]

\[
= H(X \mid X') + 0 \tag{As X and X’ together determine E}
\]

\[
H(E, X \mid X) = H(E \mid X') + H(X \mid E, X') \tag{By chain rule}
\]

\[
\leq H(P_e) + H(X \mid E, X') \tag{As conditioning cannot increase entropy}
\]

\[
= H(P_e) + \Pr[E = 0]H(X \mid X', E = 0) + \Pr[E = 1]H(X \mid X', E = 1)
\]

\[
= H(P_e) + (1 - P_e) \cdot 0 + (P_e) \cdot H(X \mid X', E = 1)
\]

\[
\leq H(P_e) + P_e \cdot H(X \mid X', E = 1)
\]

Given \( X' \) and \( E \), there are \(|X| - 1\) possible values for \( X \), as the only condition is that it must be different from \( X' \). The conditional entropy \( H(X \mid X', E = 1) \) is upper bounded by the case of uniform distribution, where \( H(X \mid X', E = 1) = \log_2(|X| - 1) \). Thus, we can conclude that:

\[
H(E, X \mid X) \leq H(P_e) + P_e \cdot H(X \mid X', E = 1) \leq H(P_e) + P_e \cdot \log_2(|X| - 1)
\]
Combining the above, we get

\[ H(X \mid X') \leq H(P_e) + P_e \cdot \log_2(|X| - 1) \]  

(A)

By the data processing inequality, we have:

\[ I(X \mid Y) \geq I(X \mid X') \]

\[ \implies H(X) - H(X \mid Y) \geq H(X) - H(X \mid X') \]  

(By definition)

\[ \implies H(X \mid Y) \leq H(X \mid X') \]

Combining with (A), we can conclude that

\[ H(X \mid Y) \leq H(X \mid X') \leq H(P_e) + P_e \cdot \log_2(|X| - 1) \]

\[ \blacksquare \]

### 2.1 Showing Tightness

Suppose the distribution \( p \) of \( X \) satisfies \( p_1 \leq p_2 \leq .. \leq p_n \).

Suppose \( Y \) is a constant, so \( I(X \mid Y) = H(X) - H(X \mid Y) = 0 \).

As \( p_1 \) is the largest discrete probability, the best predictor \( X' \) of \( X \) is \( X' = 1 \).

Then, \( P_e = \Pr[X' \neq X] = 1 - p_1 \).

Fano’s Inequality gives the following:

\[ H(X \mid Y) \leq H(P_1) + (1 - p_1) \cdot \log_2(n - 1) \]

Here, we can let \( p_2 = p_3 = ... = p_n = \frac{1-p_1}{n-1} \).

Then, the RHS can be simplified as follows:

\[ H(P_1) + (1 - p_1) \cdot \log_2(n - 1) = p_1 \log_2 \frac{1}{p_1} + (1 - p_1) \log_2 \frac{1}{1 - p_1} + (1 - p_1) \cdot \log_2(n - 1) \]

\[ = p_1 \log_2 \frac{1}{p_1} + (1 - p_1)(\log_2 \frac{n - 1}{1 - p_1}) \]

\[ = p_1 \log_2 \frac{1}{p_1} + \sum_{i=2,...,n} \frac{1-p_1}{n-1} (\log_2 \frac{n - 1}{1 - p_1}) \]

\[ = \sum_{i=1,...,n} p_i \log_2 \frac{1}{p_i} \]

\[ = H(X) \]

\[ = H(X \mid Y) \]  

(As \( X \) and \( Y \) independent)

Thus, the inequality is tight in this case.