Data Processing and Mutual Information

Random variables $X$, $Y$, $Z$ are said to form a **Markov chain** if their joint pmf can be written as

$$P_{XYZ} = P_X P_{Y|X} P_{Z|Y}.$$ 

In other words, the conditional distribution of $Z$ given $(X, Y)$ depends only on $Y$, i.e., $P_{Z|XY} = P_{Z|Y}$.

Markov chains often occur in engineering problems, e.g.,

1. $Y$ is a noisy version of $X$, and $Z = f(Y)$ is an estimator of $X$ based only on $Y$.
2. The output of the $X \to Y$ channel is fed into the $Y \to Z$ channel.

**Data-Processing Inequality**

If $X$, $Y$, $Z$ form a Markov chain, then $I(X; Y) \geq I(X; Z)$.

*Proof:* Q.7, Examples Paper I.

“Processing the data $Y$ cannot increase the information about $X$.”
Fano’s inequality

\[ X \xrightarrow{P_{Y|X}} Y \xrightarrow{\text{Estimator}} \hat{X} = g(Y) \]

- We want to estimate \( X \) by observing a correlated random variable \( Y \)
- The probability of error of an estimator \( \hat{X} = g(Y) \) is \( P_e = \Pr(\hat{X} \neq X) \)
- We wish to bound \( P_e \)

### Fano’s Inequality

For any estimator \( \hat{X} \) such that \( X - Y - \hat{X} \), the probability of error \( P_e = \Pr(\hat{X} \neq X) \) satisfies

\[
1 + P_e \log |\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y) \quad \text{or} \quad P_e \geq \frac{H(X|Y) - 1}{\log |\mathcal{X}|}
\]

### Proof of Fano

- Define an error random variable
  \[
  E = \begin{cases} 
    1 & \text{if } \hat{X} \neq X \\
    0 & \text{if } \hat{X} = X 
  \end{cases}
  \]
- Use chain rule to expand \( H(E, X|\hat{X}) \) in two different ways:
  \[
  H(E, X|\hat{X}) = H(X|\hat{X}) + H(E|X, \hat{X}) = H(E|\hat{X}) + H(X|\hat{X}, E)
  \]

### Claims:

1. \( H(E|X, \hat{X}) = 0 \). (because \( E \) is a function of \( (X, \hat{X}) \))
2. \( H(E|\hat{X}) \leq H(E) = H_2(P_e) \). (conditioning can only reduce \( H \))
3. \( H(X|\hat{X}, E) \leq P_e \log |\mathcal{X}| \) because

\[
H(X|\hat{X}, E) = \Pr(E = 0)H(X|\hat{X}, E = 0) + \Pr(E = 1)H(X|\hat{X}, E = 1) \\
\leq (1 - P_e) 0 + P_e \log |\mathcal{X}|
\]

Using the three claims in (1), we get . . .
\[ \hat{X} = g(Y) \]

\[ H(X|\hat{X}) \leq H_2(P_e) + P_e \log |X| \]

Note that \( H_2(P_e) \leq 1 \). Therefore

\[ H(X|\hat{X}) \leq 1 + P_e \log |X| \cdot \]

We have proved one side of Fano.

For the other side, the data-processing inequality tells us that

\[ I(X; Y) = H(X) - H(X|Y) \geq I(X; \hat{X}) = H(X) - H(X|\hat{X}) \]

Thus \( H(X|\hat{X}) \geq H(X|Y) \).

Back to the Channel Coding problem . . .

\[ W \xrightarrow{Encoder} X^n \xrightarrow{DMC P_{Y|X}} Y^n \xrightarrow{Decoder} \hat{W} \]

Fano’s Inequality applied to a channel code:

- Consider a \( (2^{nR}, n) \) channel code
- \( \hat{W} \) is a guess of \( W \) based on \( Y^n \)
- \( W \) uniformly distributed in \( \{1, \ldots, 2^{nR}\} \)
- \( P_e = \Pr(\hat{W} \neq W) = \frac{1}{2^{nR}} \sum_{k=1}^{2^{nR}} \Pr(\hat{W} \neq k|W = k) \)

Fano’s inequality applied to this problem gives:

\[ H(W|\hat{W}) \leq 1 + P_e \log 2^{nR} = 1 + P_e nR \]

We will use this to show that any sequence of \( (2^{nR}, n) \) codes with \( P_e \to 0 \) must have \( R \leq C \).
A Little Lemma

Let $Y^n$ be the result of passing a sequence $X^n$ through a DMC of channel capacity $C$. Then

$$I(X^n; Y^n) \leq nC$$

regardless of the distribution of $X^n$.

Proof: $I(X^n; Y^n) = H(Y^n) - H(Y^n|X^n)$

$$= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_{i-1}, \ldots, Y_1, X^n)$$

$$\overset{(a)}{=} H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i)$$

$$\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i)$$

$$= \sum_{i=1}^{n} I(X_i; Y_i) \overset{(c)}{=} nC.$$

Justification for steps (a) – (c):

(a) The channel is assumed to be memoryless. This means that given $X_i$, $Y_i$ is conditionally independent of everything else.

(b) We have

$$H(Y^n) = H(Y_1) + H(Y_2|Y_1) + \ldots + H(Y_n|Y_{n-1}, \ldots, Y_1)$$

$$\leq H(Y_1) + H(Y_2) + \ldots + H(Y_n)$$

as conditioning can only reduce entropy.

(c) From the definition of capacity, $C$ is the maximum of $I(X; Y)$ over all joint pmfs over $(X, Y)$ where $P_{Y|X}$ is fixed by the channel.
Consider any \((2^{nR}, n)\) channel code with average probability of error \(P_e\). We have:

\[
nR \overset{(a)}{=} H(W) \\
\overset{(b)}{=} H(W|\hat{W}) + I(W; \hat{W}) \\
\overset{(c)}{\leq} 1 + P_e nR + I(W; \hat{W}) \\
\overset{(d)}{\leq} 1 + P_e nR + I(X^n; Y^n) \\
\overset{(e)}{\leq} 1 + P_e nR + nC.
\]

This implies:

\[
P_e \geq 1 - \frac{C}{R} - \frac{1}{nR}
\]

Thus, unless \(R \leq C\), \(P_e\) is bounded away from 0 as \(n \to \infty\). \(\square\)

Justification for steps \((a) - (e)\):

- \((a)\) \(W\) is uniform over \(\{1, \ldots, 2^{nR}\}\)
- \((b)\) \(I(W; \hat{W}) = H(W) - H(W|\hat{W})\)
- \((c)\) Fano applied to \(H(W|\hat{W})\) (see Slide 6)
- \((d)\) Data processing inequality applied to \(W - X^n - Y^n - \hat{W}\).
- \((e)\) From the lemma on Slide 7
Summary

$\mathcal{C}$ is a sharp threshold!

- For all rates $R < \mathcal{C}$, there exists a sequence of $(2^{nR}, n)$ codes whose $P_e \to 0$.
- For $R > \mathcal{C}$, you cannot find a sequence of $(2^{nR}, n)$ codes whose $P_e \to 0$.

**Given a channel, do we have a practical way to communicate reliably at any rate $R < \mathcal{C}$?**

No, because

1. Joint typical decoding is too complex to be feasible
2. An $2^{nR} \times n$ codebook too large to store

In the next six lectures (by Jossy), you will learn how to design good channel codes with

- Compact codebook representation
- Fast encoding and decoding algorithms

You can now do all the questions in Examples Paper 1