8.1. Preliminaries

8.1.1. Markov Chains

When $X$, $Y$, and $Z$ are random variables that model the signals in a cascade of systems as in Figure 8.1, the joint distribution of $X$, $Y$, and $Z$ usually implies that $X$ and $Z$ are conditionally independent given $Y$. This is sometimes summarized by writing $X \rightarrow Y \rightarrow Z$ and saying that the triple $X$, $Y$, and $Z$ forms a Markov chain.

8.1.2. Data Processing Inequalities

Data processing inequalities say that if $X \rightarrow Y \rightarrow Z$ is a Markov chain, then “the mutual information between close things is greater than or equal to the mutual information between far things.” More precisely, it is shown in the problems that

$$X \rightarrow Y \rightarrow Z \Rightarrow \begin{cases} I(X \wedge Y) \geq I(X \wedge Z), \\
\text{and} \\
I(X \wedge Z) \leq I(Y \wedge Z).
\end{cases}$$

Example 8.1. Let $X$ and $Y$ have any joint pmf, and suppose $Z = \varphi(Y)$ for some deterministic function $\varphi$. In this case we show that $X \rightarrow Y \rightarrow Z$ is a Markov chain. To begin, write

$$P(Z = z|Y = y, X = x) = P(\varphi(Y) = z|Y = y, X = x) = P(\varphi(y) = z|Y = y, X = x)$$

$$= P(\varphi(y) = z),$$
where we have used the law of substitution and the fact that a deterministic event is independent of everything. Since the probability of a deterministic condition is one if the condition is true and zero otherwise,

\[ P(Z = z | Y = y, X = x) = \begin{cases} 
1, & \text{if } \varphi(y) = z, \\
0, & \text{otherwise},
\end{cases} \]

which does not depend on \( x \).

**Example 8.2.** Suppose \( M \) and \( Y \) are random variables with \( P(Y \in B, M = i) = W_n(B \mid x_i)q(i) \), where \( q \) is a pmf on \( \{1, \ldots, N\} \), and \( x_1, \ldots, x_N \) are given. Put \( f(i) := x_i \), and define a new random variable \( X := f(M) = x_M \). Then

\[
P(Y \in B, X = x, M = i) = P(Y \in B, f(M) = x, M = i) = P(Y \in B, f(i) = x, M = i) = \begin{cases} 
P(Y \in B, M = i), & \text{if } x = x_i, \\
0, & \text{otherwise},
\end{cases} = W_n(B \mid x) I_{\{x_i\}}(x)q(i).
\]

Taking \( B = Y \) shows that

\[
P(X = x, M = i) = I_{\{x_i\}}(x)q(i).
\]

Thus

\[
P(Y \in B \mid X = x, M = i) = \frac{P(Y \in B, X = x, M = i)}{P(X = x, M = i)} = \frac{W_n(B \mid x) I_{\{x_i\}}(x)q(i)}{I_{\{x_i\}}(x)q(i)} = W_n(B \mid x),
\]

which does not depend on \( i \).

---

### 8.2. Fano’s Inequality

**Theorem 8.3** (Fano’s Inequality). Let \( M \) and \( V \) be two random variables taking values in \( \{1, \ldots, N\} \). Then

\[
H(M \mid V) \leq P(M \neq V) \log(N - 1) + h(P(M \neq V)),
\]

\[\text{a} \]

The value assigned to \( P(Y \in B \mid X = x, M = i) \) when the denominator is zero can be any probability measure on \( B \); what is important about conditional probability is that its product with the denominator always equal the joint probability; i.e., we need

\[
P(Y \in B \mid X = x, M = i)P(X = x, M = i) = P(Y \in B, X = x, M = i).
\]

This is clearly holds for all \( x \) — check the two cases \( x = x_i \) and \( x \neq x_i \).
where
\[ h(\theta) := \begin{cases} 
\theta \log \frac{1}{\theta} + (1 - \theta) \log \frac{1}{1 - \theta}, & 0 < \theta < 1, \\
0, & \text{otherwise},
\end{cases} \]
is the binary entropy function.

**Corollary 8.4.** \( H(M) \leq P(M \neq V) \log N + \log 2 + I(M \wedge V). \)

**Proof of Corollary.** Since \( h \) is the entropy of a pmf on two points, \( h(P(M \neq V)) \leq \log 2. \) Since \( N - 1 < N, \) Fano’s inequality implies
\[ 0 \leq P(M \neq v) \log N + \log 2 - H(M|V). \]
Now add \( H(M) \) to both sides to obtain the results, since \( H(M) - H(M|V) = I(M \wedge V). \) \( \square \)