ENEE 627
SPRING 2011
INFORMATION THEORY

DATA PROCESSING

Markov chains

Consider a collection of n rvs, say $X_1, \ldots, X_n$, defined on the same probability triple. For each $i = 1, \ldots, n$, the rv X_i is $\mathcal{X}_i$-valued with $\mathcal{X}_i$ a finite set. We shall write

$$\mathcal{X}^n = \times_{i=1}^n \mathcal{X}_i.$$

The rvs $X_1, \ldots, X_n$ are said to form a *Markov chain* if the conditions

$$\begin{aligned} & \mathbb{P}[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] \\ &= \mathbb{P}[X_1 = x_1] \prod_{k=1}^{n-1} p_{k+1}(x_{k+1}|x_k), \quad (x_1, \ldots, x_n) \in \mathcal{X}^n \end{aligned} \tag{1}$$

all hold where for each $k = 1, \ldots, n-1$, we require

$$\left. \begin{array}{c} 0 \leq p_{k+1}(x_{k+1}|x_k) \leq 1 \\ \sum_{x_{k+1} \in \mathcal{X}_{k+1}} p_{k+1}(x_{k+1}|x_k) = 1 \end{array} \right., \quad x_k \in \mathcal{X}_k, \ x_{k+1} \in \mathcal{X}_{k+1}.$$

The Markov chain property of the rvs $X_1, \ldots, X_n$ is concisely represented through

$$X_1 \to X_2 \to \ldots \to X_{n-1} \to X_n.$$

Simple facts

Here are some simple facts concerning Markov chains as needed in the context of Information Theory.

Fact 0.1 *If* $X_1 \to X_2 \to \ldots \to X_{n-1} \to X_n$*, then for each* $k = 2, \ldots, n-1$*, the rvs* $\{X_i, \ i = 1, 2, \ldots, k-1\}$ *and* $\{X_j, \ j = k+1, k+2, \ldots, n\}$ *are mutually independent given* X_k*.*

The Markov property is inherited by taking subsets.

Fact 0.2 *If $X_1 \to X_2 \to \ldots \to X_{n-1} \to X_n$, then for any subset I of $\{1, \ldots, n\}$ with $|I| \geq 2$, the collection of rvs $\{X_i,\ i \in I\}$ is also a Markov chain, namely if $I = \{i_1, \ldots, i_k\}$ with $i_1 < i_2 < \ldots < i_k$ for some $k = 2, \ldots, n$, then*

$$X_{i_1} \to X_{i_2} \to \ldots \to X_{i_k}.$$

Proof. Note that if $X_1 \to X_2 \to \ldots \to X_{n-1} \to X_n$, then

$$\begin{aligned}
&\mathbb{P}\left[X_1 = x_1, X_2 = x_2, \ldots, X_{n-1} = x_{n-1}\right] \\
&\quad = \sum_{x_n \in \mathcal{X}_n} \mathbb{P}\left[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right] \\
&\quad = \sum_{x_n \in \mathcal{X}_n} \mathbb{P}\left[X_1 = x_1\right] \prod_{k=1}^{n-1} p_{k+1}(x_{k+1}|x_k) \\
&\quad = \sum_{x_n \in \mathcal{X}_n} \mathbb{P}\left[X_1 = x_1\right] \prod_{k=1}^{n-2} p_{k+1}(x_{k+1}|x_k) p_n(x_n|x_{n-1}) \\
&\quad = \mathbb{P}\left[X_1 = x_1\right] \prod_{k=1}^{n-2} p_{k+1}(x_{k+1}|x_k) \left(\sum_{x_n \in \mathcal{X}_n} p_n(x_n|x_{n-1}) \right) \\
&\quad = \mathbb{P}\left[X_1 = x_1\right] \prod_{k=1}^{n-2} p_{k+1}(x_{k+1}|x_k), \quad (x_1, \ldots, x_{n-1}) \in \mathcal{X}^{n-1} \qquad (2)
\end{aligned}$$

since

$$\sum_{x_n \in \mathcal{X}_n} p_n(x_n|x_{n-1}), \quad x_{n-1} \in \mathcal{X}_{n-1},$$

and it is now plain that $X_1 \to X_2 \to \ldots \to X_{n-1}$.

Similarly, if $X_1 \to X_2 \to \ldots \to X_{n-1} \to X_n$, then

$$\begin{aligned}
&\mathbb{P}\left[X_2 = x_2, \ldots, X_{n-1} = x_{n-1}, X_n = x_n\right] \\
&\quad = \sum_{x_1 \in \mathcal{X}_1} \mathbb{P}\left[X_1 = x_1, X_2 = x_2, \ldots, X_{n-1} = x_{n-1}, X_n = x_n\right] \\
&\quad = \sum_{x_1 \in \mathcal{X}_1} \mathbb{P}\left[X_1 = x_1\right] \prod_{k=1}^{n-1} p_{k+1}(x_{k+1}|x_k)
\end{aligned}$$

$$= \left(\sum_{x_1 \in \mathcal{X}_1} \mathbb{P}\left[X_1 = x_1\right] p_2(x_2|x_1) \right) \prod_{k=1}^{n-1} p_{k+1}(x_{k+1}|x_k)$$

$$= \mathbb{P}\left[X_2 = x_2\right] \prod_{k=2}^{n-1} p_{k+1}(x_{k+1}|x_k), \quad (x_2, \ldots, x_n) \in \mathcal{X}_2 \times \ldots \times \mathcal{X}_n \tag{3}$$

as we note that

$$\mathbb{P}\left[X_2 = x_2\right] = \sum_{x_1 \in \mathcal{X}_1} \mathbb{P}\left[X_1 = x_1\right] p_2(x_2|x_1).$$

Just apply (1) and use the conditions

$$\sum_{x_{k+1} \in \mathcal{X}_{k+1}} p_{k+1}(x_{k+1}|x_k) = 1, \qquad \begin{array}{c} k = 1, \ldots, n-1 \\ x_k \in \mathcal{X}_k \end{array}$$

and we get $X_2 \rightarrow X_2 \rightarrow \ldots \rightarrow X_{n-1} \rightarrow X_n$.

Finally, assuming $n \geq 3$, pick $2 \leq \ell \leq n-1$. Similar arguments show that removing X_ℓ does not change the Markov property of the remaining rvs. Thus removing any one of the rvs does not change the Markov property. Iterating this operation k times with the rvs with index in I yields the result. ■

Reversing time does not change the Markov property.

Fact 0.3 *If* $X_1 \rightarrow X_2 \rightarrow \ldots \rightarrow X_{n-1} \rightarrow X_n$*, then it is also the case that* $X_n \rightarrow X_{n-1} \rightarrow \ldots \rightarrow X_2 \rightarrow X_1$.

Proof. From (1) we note that

$$\begin{aligned} &\mathbb{P}\left[X_n = y_1, X_{n-1} = y_2, \ldots, X_1 = y_n\right] \\ &\quad = \mathbb{P}\left[X_1 = y_n, X_2 = y_{n-1}, \ldots, X_{n-1} = y_2, X_n = y_1\right] \\ &\quad = \mathbb{P}\left[X_1 = y_n\right] \prod_{k=1}^{n-1} p_{k+1}(y_k|y_{k-1}), \quad (y_n, \ldots, y_1) \in \mathcal{X}^n \end{aligned} \tag{4}$$

■

Data Processing Inequalities

We begin with the Data Processing Inequality in its standard form.

Lemma 0.1 *For any Markov chain $X \to Y \to Z$, it is the case that*

$$I(X;Z) \le I(X;Y) \tag{5}$$

and

$$I(X;Z) \le I(Z;Y). \tag{6}$$

Proof. By the chain rule for mutual informations applied to $I(X;(Y,Z))$ twice, we find

$$I(X;(Y,Z)) = I(X;Y) + I(X;Z|Y) \tag{7}$$

and

$$I(X;(Y,Z)) = I(X;Z) + I(X;Y|X) \tag{8}$$

The Markov property $X \to Y \to Z$ implies that X and Z are conditionally independent given Y, whence $I(X;Z|Y) = 0$. Thus,

$$I(X;(Y,Z)) = I(X;Y) \tag{9}$$

and we conclude that

$$I(X;Z) + I(X;Y|X) = I(X;Y). \tag{10}$$

The desired conclusion (5) now follows since $I(X;Y|X) \ge 0$.

By Fact 0.3 we note that $Z \to Y \to X$ since $X \to Y \to Z$, and applying (5) (this time with $X \leftarrow Z$, $Y \leftarrow Y$ and $Z \leftarrow X$) yields (6). ■

In the context of the Channel Coding Theorem, the following version of the Data Processing Inequality is needed.

Lemma 0.2 *For any Markov chain $X \to U \to V \to Y$, we have*

$$I(X;Y) \le I(U;V). \tag{11}$$

Proof. By Fact 0.2 the Markov property $X \to U \to V \to Y$ implies both

(12) $$X \to V \to Y$$

and

(13) $$X \to U \to V.$$

Applying Lemma 0.1 to (12) and (13) we get

(14) $$I(X;Y) \leq I(X;V)$$

and

(15) $$I(X;V) \leq I(U;V).$$

The conclusion (11) follows by combining (14) and (15). ■

The Markov property and the DMC

Consider the DMC with channel matrix $\boldsymbol{P} = (p(y|x),\ x \in \mathcal{X},\ y \in \mathcal{Y})$. The message W to be sent is selected from a set of M distinct messages $\mathcal{M} \equiv \{1, 2, \ldots, M\}$ with M some positive integer. For each $n = 1.2. \ldots$, consider the (M, n)-code $C_n = (f_n, g_n)$ with encoding function $f_n : \mathcal{M} \to \mathcal{X}^n$ and decoding function $g_n : \mathcal{Y}^n \to \mathcal{M}$.

The string of symbols $\boldsymbol{X}^n$ to be sent over the channel is specified by

$$\boldsymbol{X}^n = f_n(W),$$

and upon receiving the string of symbols $\boldsymbol{Y}^n$, the estimate $\widehat{W}$ is generated according to

$$\widehat{W} = g_n(\boldsymbol{Y}^n).$$

As usual the DMC assumption is encapsulated through

$$\mathbb{P}\left[\boldsymbol{Y}^n = \boldsymbol{y}^n | \boldsymbol{X}^n = \boldsymbol{x}^n\right] = \prod_{k=1}^{n} p(y_k|x_k), \quad \boldsymbol{x}^n \in \mathcal{X}^n, \boldsymbol{y}^n \in \mathcal{Y}^n.$$

Lemma 0.3 *With the usual notation, for each $n = 1, 2, \ldots$, we have*

(16) $$W \to \boldsymbol{X}^n \to \boldsymbol{Y}^n \to \widehat{W}$$

provided W is selected independently of the operation of the DMC.

Proof. Select w in $\mathcal{M}$, $\boldsymbol{x}^n$ in $\mathcal{X}^n$, $\boldsymbol{y}^n$ in $\mathcal{Y}^n$ and v in $\mathcal{M}$. Note that

$$\begin{aligned}
& \mathbb{P}\left[W = w, \boldsymbol{X}^n = \boldsymbol{x}^n, \boldsymbol{Y}^n = \boldsymbol{y}^n, \widehat{W} = v\right] \\
&= \mathbb{P}\left[W = w, \boldsymbol{X}^n = \boldsymbol{x}^n, \boldsymbol{Y}^n = \boldsymbol{y}^n, g_n(\boldsymbol{Y}^n) = v\right] \\
&= \mathbb{P}\left[W = w, \boldsymbol{X}^n = \boldsymbol{x}^n, \boldsymbol{Y}^n = \boldsymbol{y}^n, g_n(\boldsymbol{y}^n) = v\right] \\
&= \mathbb{P}\left[W = w, \boldsymbol{X}^n = \boldsymbol{x}^n, \boldsymbol{Y}^n = \boldsymbol{y}^n\right] \cdot \delta(g_n(\boldsymbol{y}^n), v) \\
&= \mathbb{P}\left[W = w, \boldsymbol{X}^n = \boldsymbol{x}^n\right] \cdot \prod_{k=1}^n p(y_k|x_k) \cdot \delta(g_n(\boldsymbol{y}^n), v) \\
&= \mathbb{P}\left[W = w, f_n(W) = \boldsymbol{x}^n\right] \cdot \prod_{k=1}^n p(y_k|x_k) \cdot \delta(g_n(\boldsymbol{y}^n), v) \\
&= \mathbb{P}\left[W = w\right] \cdot \delta(f_n(w), \boldsymbol{x}^n) \cdot \prod_{k=1}^n p(y_k|x_k) \cdot \delta(g_n(\boldsymbol{y}^n), v). \qquad (17)
\end{aligned}$$

■

Basic arguments in the converse of the CCT

The converse to the Channel Coding Theorem results from the following chain of arguments:

Assume W to be uniformly distributed over the message set $\mathcal{M}$, and independent of the operation of the DMC. Thus,

$$\begin{aligned}
& \log_2 M \\
&= H(W) \\
&= H(W|\widehat{W}) + I(W;\widehat{W}) \\
&\leq 1 + \log_2 M \cdot \mathbb{P}\left[\widehat{W} \neq W\right] + I(W;\widehat{W}) \quad \text{(Fano's Inequality)} \\
&\leq 1 + \log_2 M \cdot \mathbb{P}\left[\widehat{W} \neq W\right] + I(\boldsymbol{X}^n;\boldsymbol{Y}^n) \quad \text{(Data Processing)} \\
&= 1 + \log_2 M \cdot \mathbb{P}\left[\widehat{W} \neq W\right] + H(\boldsymbol{Y}^n) - H(\boldsymbol{Y}^n|\boldsymbol{X}^n) \\
&= 1 + \log_2 M \cdot \mathbb{P}\left[\widehat{W} \neq W\right] + H(\boldsymbol{Y}^n) - \sum_{i=1}^n H(Y_i|\boldsymbol{X}^n, \boldsymbol{Y}^{i-1})
\end{aligned}$$

(Chain rule for conditional entropy)

$$
\begin{aligned}
&= 1 + \log_2 M \cdot \mathbb{P}\left[\widehat{W} \neq W\right] + H(\boldsymbol{Y}^n) - \sum_{i=1}^{n} H(Y_i|X_i) \quad \text{(DMC)} \\
&\leq 1 + \log_2 M \cdot \mathbb{P}\left[\widehat{W} \neq W\right] + \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \\
&= 1 + \log_2 M \cdot \mathbb{P}\left[\widehat{W} \neq W\right] + \sum_{i=1}^{n} I(X_i;Y_i) \\
&\leq 1 + \log_2 M \cdot \mathbb{P}\left[\widehat{W} \neq W\right] + nC \quad \text{(Definition of channel capacity for the DMC)}
\end{aligned}
$$

In short,

$$
\frac{\log_2 M}{n} \leq \frac{1}{n} + \frac{\log_2 M}{n} \cdot \mathbb{P}\left[\widehat{W} \neq W\right] + C \tag{18}
$$