ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

School of Computer and Communication Sciences

Handout 7 — Information Theory and Coding

Solutions to homework 3 — Oct. 08, 2013

PROBLEM 1.

(a) $H(X) = \frac{2}{3}\log\frac{3}{2} + \frac{1}{3}\log 3 = 0.918$ bits $= H(Y)$.

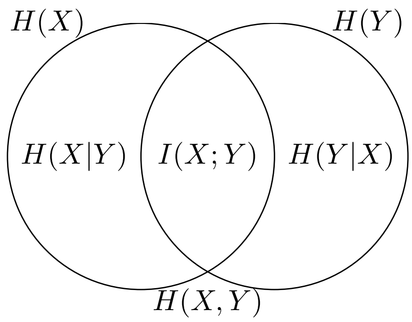
(b) $H(X|Y) = \frac{1}{3}H(X|Y=0) + \frac{2}{3}H(X|Y=1) = 0.667$ bits $= H(Y|X)$.

(c) $H(X,Y) = 3 \times \frac{1}{3}\log 3 = 1.585$ bits.

(d) $H(Y) - H(Y|X) = 0.251$ bits.

(d) $I(X;Y) = H(Y) - H(Y|X) = 0.251$ bits.

(f)



PROBLEM 2.

$$\begin{aligned} H(X) &= -\sum_{k=1}^{M} P_X(a_k)\log P_X(a_k) \\ &= -\sum_{k=1}^{M-1}(1-\alpha)P_Y(a_k)\log[(1-\alpha)P_Y(a_k)] - \alpha\log\alpha \\ &= (1-\alpha)H(Y) - (1-\alpha)\log(1-\alpha) - \alpha\log\alpha \end{aligned}$$

Since Y is a random variable that takes $M-1$ values $H(Y) \le \log(M-1)$ with equality if and only if Y takes each of its possible values with equal probability.

PROBLEM 3.

(a) Using the chain rule for mutual information,

$$I(X,Y;Z) = I(X;Z) + I(Y;Z \mid X) \ge I(X;Z),$$

with equality iff $I(Y;Z \mid X) = 0$, that is, when Y and Z are conditionally independent given X.

(b) Using the chain rule for conditional entropy,

$$H(X,Y \mid Z) = H(X \mid Z) + H(Y \mid X,Z) \ge H(X \mid Z),$$

with equality iff $H(Y \mid X,Z) = 0$, that is, when Y is a function of X and Z.

(c) Using first the chain rule for entropy and then the definition of conditional mutual information,

$$
\begin{aligned}
H(X,Y,Z) - H(X,Y) = H(Z \mid X,Y) &= H(Z \mid X) - I(Y;Z \mid X) \\
&\le H(Z \mid X) = H(X,Z) - H(X)\,,
\end{aligned}
$$

with equality iff $I(Y;Z \mid X) = 0$, that is, when Y and Z are conditionally independent given X.

(d) Using the chain rule for mutual information,

$$I(X;Z \mid Y) + I(Z;Y) = I(X,Y;Z) = I(Z;Y \mid X) + I(X;Z)\,,$$

and therefore

$$I(X;Z \mid Y) = I(Z;Y \mid X) - I(Z;Y) + I(X;Z)\,.$$

We see that this inequality is actually an equality in all cases.

Problem 4. Let X^i denote $X_1, \ldots, X_i$.

(a) By the chain rule for entropy,

$$
\begin{aligned}
\frac{H(X_1, X_2, \ldots, X_n)}{n} &= \frac{\sum_{i=1}^{n} H(X_i|X^{i-1})}{n} && (1)\\
&= \frac{H(X_n|X^{n-1}) + \sum_{i=1}^{n-1} H(X_i|X^{i-1})}{n} && (2)\\
&= \frac{H(X_n|X^{n-1}) + H(X_1, X_2, \ldots, X_{n-1})}{n}. && (3)
\end{aligned}
$$

From stationarity it follows that for all $1 \le i \le n$,

$$H(X_n|X^{n-1}) \le H(X_i|X^{i-1}),$$

which further implies, by summing both sides over $i = 1, \ldots, n-1$ and dividing by $n-1$, that,

$$
\begin{aligned}
H(X_n|X^{n-1}) &\le \frac{\sum_{i=1}^{n-1} H(X_i|X^{i-1})}{n-1} && (4)\\
&= \frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1}. && (5)
\end{aligned}
$$

Combining (3) and (5) yields,

$$
\begin{aligned}
\frac{H(X_1, X_2, \ldots, X_n)}{n} &\le \frac{1}{n}\left[\frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1} + H(X_1, X_2, \ldots, X_{n-1})\right] && (6)\\
&= \frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1}. && (7)
\end{aligned}
$$

(b) By stationarity we have for all $1 \le i \le n$,

$$H(X_n|X^{n-1}) \le H(X_i|X^{i-1}),$$

which implies that,

$$
\begin{aligned}
H(X_n|X^{n-1}) &= \frac{\sum_{i=1}^{n} H(X_n|X^{n-1})}{n} && (8)\\
&\le \frac{\sum_{i=1}^{n} H(X_i|X^{i-1})}{n} && (9)\\
&= \frac{H(X_1, X_2, \ldots, X_n)}{n}. && (10)
\end{aligned}
$$

Problem 5. By the chain rule for entropy,

$$H(X_0|X_{-1},\ldots,X_{-n}) = H(X_0,X_{-1},\ldots,X_{-n}) - H(X_{-1},\ldots,X_{-n}) \quad (11)$$
$$= H(X_0,X_1,\ldots,X_n) - H(X_1,\ldots,X_n) \quad (12)$$
$$= H(X_0|X_1,\ldots,X_n), \quad (13)$$

where (12) follows from stationarity.

Problem 6. For a Markov chain, given X_0 and X_n are independent given X_{n-1}. Thus

$$H(X_0|X_nX_{n-1}) = H(X_0|X_{n-1})$$

But, since conditioning reduces entropy,

$$H(X_0|X_nX_{n-1}) \le H(X_0|X_n).$$

Putting the above together we see that $H(X_0|X_{n-1}) \le H(X_0|X_n)$.

Problem 7.

$X_1, X_2, \ldots$ are i.i.d. with distribution $p(x)$. Hence $\log p(X_i)$ are also i.i.d. and

$$\begin{aligned}
\lim(p(X_1,\ldots,X_n))^{\frac{1}{n}} &= \lim 2^{\log(p(X_1,X_2,\ldots,X_n))^{\frac{1}{n}}}\\
&= 2^{\lim \frac{1}{n}\sum \log p(X_i)}\\
&= 2^{E(\log(p(X)))} \text{ a.e.}\\
&= 2^{-H(X)}
\end{aligned}$$

by the strong law of large numbers (assuming of course that $H(X)$ exists). Note: The abbreviation a.e. stands for 'almost everywhere', which is synonymous with 'with probability 1'.

For the second part of the problem we had intended to ask a question for which taking limit and taking the expectation do not commute (i.e., the order you take them matters). This, however, is not the case here: Let G_n be the set of $(x_1,\ldots,x_n)$ for which $|p(x_1,\ldots,x_n)^{1/n} - 2^{-H(X)}| < \epsilon$. We know from the first part that $\Pr(G_n) \to 1$ as n gets large. Since $0 \le p(x_1,\ldots,x_n)^{1/n} \le 1$ for any $x_1,\ldots,x_n$, we see that

$$\begin{aligned}
|E[p(X_1,\ldots,X_n)^{1/n}] - 2^{-H(X)}| &= \left|\sum_{x_1,\ldots,x_n} p(x_1,\ldots,x_n)\left[p(x_1,\ldots,x_n)^{1/n} - 2^{-H(X)}\right]\right|\\
&\le \sum_{x_1,\ldots,x_n} p(x_1,\ldots,x_n)\left|p(x_1,\ldots,x_n)^{1/n} - 2^{-H(X)}\right|\\
&= \sum_{(x_1,\ldots,x_n)\in G_n} p(x_1,\ldots,x_n)\left|p(x_1,\ldots,x_n)^{1/n} - 2^{-H(X)}\right|\\
&\quad + \sum_{(x_1,\ldots,x_n)\notin G_n} p(x_1,\ldots,x_n)\left|p(x_1,\ldots,x_n)^{1/n} - 2^{-H(X)}\right|\\
&\overset{\text{(a)}}{\le} \sum_{(x_1,\ldots,x_n)\in G_n} p(x_1,\ldots,x_n)\epsilon + \sum_{(x_1,\ldots,x_n)\notin G_n} p(x_1,\ldots,x_n)\\
&= P(G_n)\epsilon + (1 - P(G_n))\\
&\le \epsilon + (1 - P(G_n)).
\end{aligned}$$

Where (a) follows from the definition of G_n and the fact that $|p(x_1,\ldots,x_n)^{1/n} - 2^{-H}| \le 1$. Now, as n gets large $1 - P(G_n)$ approaches zero, and we see that the difference between $E[p(X_1,\ldots,X_n)^{1/n}]$ and $2^{-H(X)}$ gets smaller than any arbitrary $\epsilon > 0$, and thus

$$\lim_{n\to\infty} E[p(X_1,\ldots,X_n)^{1/n}] = 2^{-H(X)}.$$