ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
School of Computer and Communication Sciences

Handout 20 Information Theory and Coding
Solutions to Homework 8 Nov. 18, 2019

Problem 1.

Let $P'_{X,Y}(x,y) = P_{Y|X}(y|x)Q'(x)$, $P'_Y(y) = \sum_{x\in\mathcal{X}} P'_{X,Y}(x,y)$ and $P_Y(y) = \sum_{x\in\mathcal{X}} P_{Y|X}(y|x)Q(x)$. We then have for any Q'

$$\begin{aligned}
&\sum_{x\in\mathcal{X}} Q'(x) \sum_{y\in\mathcal{Y}} P_{Y|X}(y|x) \log\left(\frac{P_{Y|X}(y|x)}{\sum_{x'\in\mathcal{X}} P_{Y|X}(y|x')Q(x')}\right) - I(Q') \\
&= E_{P'_{X,Y}} \log \frac{P_{Y|X}}{P_Y} - I(Q') \\
&= E_{P'_{X,Y}} \left(\log \frac{P_{Y|X}}{P_Y} - \log \frac{P'_{X,Y}}{Q'_X P'_Y}\right) \\
&= E_{P'_{X,Y}} \log \frac{P'_Y}{P_Y} \\
&= E_{P'_Y} \log \frac{P'_Y}{P_Y} \\
&= D(P'_Y || P_Y) \geq 0
\end{aligned}$$

with equality if and only if $Q' = Q$. To prove (b), notice in the upper bound of part (a), that the inner summation is a function of x and that the outer summation is an average of this function with respect to the distribution $Q'(x)$. The average of a function is upper bounded by the maximum value that the function takes, and hence (b) follows.

Problem 2.

(a) By the chain rule

$$I(U,T;V) = I(U;V) + I(T;V|U) = I(U;V),$$

since $I(T;V|U) = 0$ from the Markov property. Also,

$$I(U,T;V) = I(T;V) + I(U;V|T) \geq I(U;V|T),$$

from the non-negativity of the mutual information. These together imply that $I(U;V) \geq I(U;V|T)$.

(b)

$$I(X;Y|W) = \Pr\{W=1\} I(X;Y|W=1) + \Pr\{W=2\} I(X;Y|W=2)$$

Conditional on $W = k$, the distribution of (X,Y) is $p_k(x)p(y|x)$, thus

$$I(X;Y|W) = \lambda I_1 + (1-\lambda) I_2.$$

(c) We obtain $p(x)$ by summing $p(w,x,y)$ over y and w. This gives

$$p(x) = \lambda p_1(x) + (1-\lambda) p_2(x).$$

(d) Note that

$$p(w, x, y) = p(w)p(x|w)p(y|x),$$

that is Y is independent of W when X is given. Thus from (a)

$$I(X; Y) \geq I(X; Y|W). \tag{1}$$

Letting $f(p_X)$ denote the value of $I(X; Y)$ as a function of the distribution of X we can rewrite (1) as

$$f(\lambda p_1 + (1-\lambda)p_2) \geq \lambda f(p_1) + (1-\lambda)f(p_2),$$

which says that f is concave.

Problem 3.

(a) Chain rule for mutual information.

(b) $I(W, Y^{i-1}; Y_i) = I(Y^{i-1}; Y_i) + I(W; Y_i|Y^{i-1}) \geq I(W; Y_i|Y^{i-1})$.

(c) $I(W, X_i, X^{i-1}, Y^{i-1}; Y_i) = I(W, Y^{i-1}; Y_i) + I(X_i, X^{i-1}; Y_i|W, Y^{i-1}) \geq I(W, Y^{i-1}; Y_i)$. Note that this inequality is in fact equality, unless the mapping f_i is randomized.

(d) $W \to (X_i, X^{i-1}, Y^{i-1}) \to Y_i$ is a Markov chain. This follows from the following facts:

- For all $1 \leq j \leq i$, X_j is a function of (W, Y^{j-1}).
- For all $1 \leq j \leq i$, Y_j depends on (W, X^j, Y^{j-1}) only through X_j since the channel is memoryless.

This means that the joint probability distribution of (W, X^i, Y^i) can be written as follows:

$$\begin{aligned} P_{W,X^i,Y^i}(w, x^i, y^i) &= P_W(w) \times P_{X_1|W}(x_1|w)P_{Y_1|X_1}(y_1|x_1) \\ &\times P_{X_2|W,Y_1}(x_2|w, y_1)P_{Y_2|X_2}(y_2|x_2) \times \ldots \times P_{X_i|W,Y^{i-1}}(x_i|w, x^{i-1})P_{Y_i|X_i}(y_i|x_i), \end{aligned}$$

which can be rewritten as

$$P_{W,X^i,Y^i}(w, x^i, y^i) = P_W(w)P_{X_i,X^{i-1},Y^{i-1}|W}(x_i, x^{i-1}, y^{i-1}|w)P_{Y_i|X_i}(y_i|x_i).$$

(e) Since the channel is stationary and memoryless, $(X^{i-1}, Y^{i-1}) \to X_i \to Y_i$ is a Markov chain.

(f) From the definition of the capacity.

This proof still works even when the mappings f_i are randomized. We conclude that feedback does not increase the capacity even if we are allowed to use a randomized encoder.

Problem 4.

Since X and Z are both in the interval $[-1, 1]$, their sum $X + Z$ lies in the interval $[-2, +2]$. If we could *choose* the distribution of $X + Z$ as we wished (without the constraint that it has to be the sum of two independent random variables, one of which is uniform) we would have chosen it to be uniform on the interval $[-2, +2]$ to have the largest entropy. Observe now that if we choose X as the random variable that equals $+1$ with probability $1/2$ and -1 with probability $1/2$, then $X + Z$ is uniform in $[-2, +2]$ and thus this distribution maximizes the entropy. An alternate derivation is as follows: note that since X and Z are independent, the moment generating functions of the random variables involved satisfy $E[e^{s(X+Z)}] = E[e^{sX}]E[e^{sZ}]$. Now, we know that $E[e^{sZ}] = \int e^{sz} f_Z(z)\,dz = \int_{-1}^{+1} \frac{1}{2}e^{sz}\,dz = [e^s - e^{-s}]/(2s)$. Similarly, if we want $X + Z$ to be uniform on $[-2, 2]$, we can compute $E[e^{s(X+Z)}] = [e^{2s} - e^{-2s}]/(4s)$. This then requires $E[e^{sX}] = \frac{1}{2}[e^{2s} - e^{-2s}]/[e^s - e^{-s}] = \frac{1}{2}[e^s + e^{-s}]$ which is the moment generating function of a random variable which takes on the values $+1$ and -1, each with probability $1/2$.

Similarly, under the constraint XZ lies in the interval $[-1, +1]$, and the best we could hope is that XZ is uniform on this interval. But this can be achieved by making sure that X only takes on the values $+1$ or -1.

Problem 5.

$$
\begin{aligned}
h(X) &= \frac{1}{2}\log\left(2\pi e\sigma_x^2\right) \\
h(Y) &= \frac{1}{2}\log\left(2\pi e\sigma_y^2\right) \\
h(X,Y) &= \frac{1}{2}\log\left((2\pi e)^2\det(K)\right) = \frac{1}{2}\log\left((2\pi e)^2\left(\sigma_x^2\sigma_y^2 - \rho^2\sigma_x^2\sigma_y^2\right)\right) \\
I(X,Y) &= h(X) + h(Y) - h(X,Y) = \frac{1}{2}\log\frac{1}{1-\rho^2}
\end{aligned}
$$

Note that the result does not depend on σ_x, σ_y, which says that normalization does not change the mutual information.

Problem 6.

(a) Since X and Y are independent, $H(X,Y) = H(X) + H(Y)$. $H(X)$ can be found by the following steps.

$$
\begin{aligned}
H(X) &= -\sum_{i=1}^{\infty}(1-p)^{i-1}p\log\left((1-p)^{i-1}p\right) \\
&= -\sum_{i=1}^{\infty}(1-p)^{i-1}p\left((i-1)\log(1-p) + \log p\right) \\
&= -p\log(1-p)\sum_{i=1}^{\infty}(1-p)^{i-1}(i-1) - p\log p\sum_{i=1}^{\infty}(1-p)^{i-1} \\
&= -(1-p)\log(1-p)/p - p\log p/p \\
&= h_2(p)/p.
\end{aligned}
$$

Similarly, $H(Y) = h_2(q)/q$, and $H(X,Y) = h_2(p)/p + h_2(q)/q$.

(b) Since $(X,Y) \to (2X+Y, X-2Y)$ is a 1-to-1 transformation, $H(X,Y) = H(2X + Y, X - 2Y)$. To see this, we can write

$$
\begin{aligned}
H(X,Y,2X+Y,X-2Y) &= H(X,Y|2X+Y,X-2Y) + H(2X+Y,X-2Y) \\
&= H(2X+Y,X-2Y|X,Y) + H(X,Y)
\end{aligned}
$$

As $H(X,Y|2X+Y,X-2Y) = H(2X+Y,X-2Y|X,Y) = 0$, we obtain $H(2X + Y, X - 2Y) = H(X,Y)$.

(c) Similar to part (a), $h(X,Y) = h(X) + h(Y)$. To find $h(X)$, observe the following steps.

$$\begin{aligned} h(X) &= -\int_0^\infty \lambda_X e^{-\lambda_X t} \log\left(\lambda_X e^{-\lambda_X t}\right) dt \\ &= -\int_0^\infty \lambda_X e^{-\lambda_X t} \log \lambda_X dt + \int_0^\infty \lambda_X^2 t e^{-\lambda_X t} dt \\ &= -\log \lambda_X - \lambda_X E[X] \\ &= 1 - \log \lambda_X \end{aligned}$$

as $E[X] = 1/\lambda_X$. Similarly, $h(Y) = 1 - \log \lambda_Y$ and $h(X,Y) = 2 - \log \lambda_X \lambda_Y$

(d) Here, we cannot use the fact that $(X,Y) \to (2X+Y, X-2Y)$ is a 1-to-1 transformation as $h(X,Y) \neq h(f(X,Y), g(X,Y))$ in general even if it is a 1-to-1 transformation. Here, we have to use the fact that $h(\mathbf{Ax}) = h(\mathbf{x}) + \log|\mathbf{A}|$. Since we know that

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & -2 \end{bmatrix}$$

$$h(2X+Y, X-2Y) = h(X,Y) + \log|\mathbf{A}| = 2 - \log \frac{\lambda_X \lambda_Y}{5}$$