4F5 Advanced Wireless Communications, 2012

Examples Paper I

Question 1

Consider the four waveforms $x_m(t), m = 1, \ldots, 4$ shown in Figure 1.

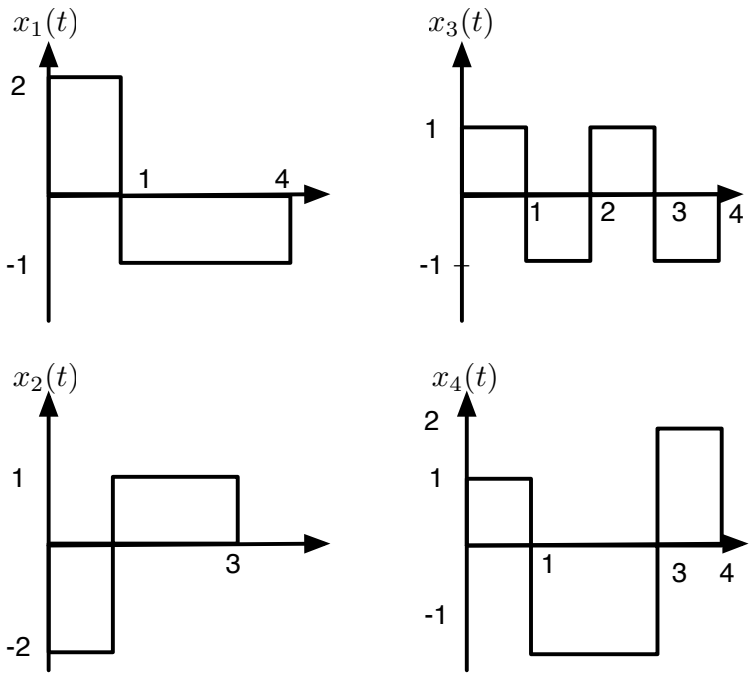


Figure 1: Signal set for Question 1.

(a) *Determine the dimensionality of the waveforms and a set of orthonormal basis functions.*

The four waveforms can be expressed as a linear combination of $f_i(t)$ $i = 1, \ldots, 4$ shown in Figure 2. Hence, the dimensionality of the signal space is 4. This question can also be solved using the Gram-Schmidt method.

(b) Use the basis functions to represent the four waveforms by vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4$.

Using the basis shown in Figure 2, we write the signal space vectors as

$$\boldsymbol{x}_1 = (2, -1, -1, -1), \boldsymbol{x}_2 = (-2, 1, 1, 0), \boldsymbol{x}_3 = (1, -1, 1, -1), \boldsymbol{x}_4 = (1, -2, -2, 2)$$

(c) Determine the minimum distance between any pair of vectors.

The distance between the first and the second vector is:

$$d_{1,2} = \sqrt{\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2} = \sqrt{\|(4, -2, -2, -1)\|^2} = \sqrt{25}$$

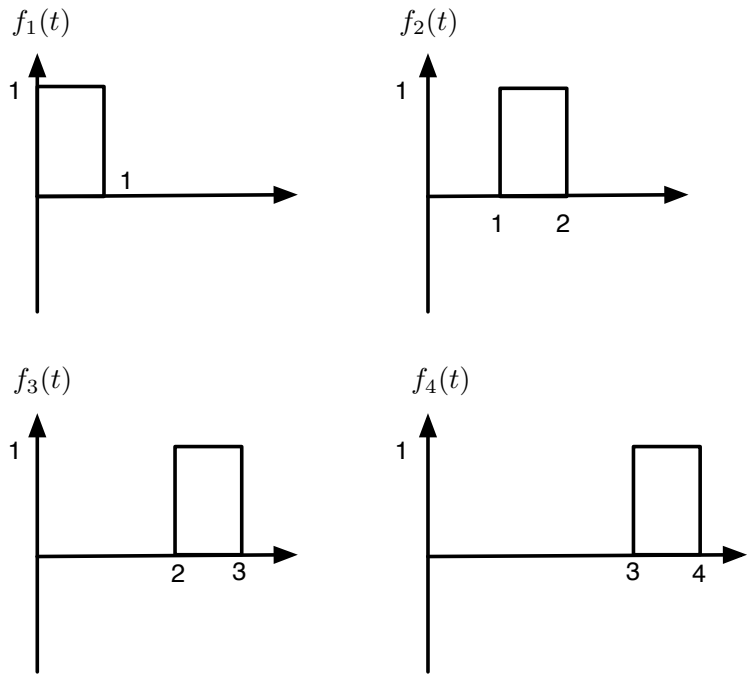


Figure 2: Basis set for Question 1.

Similarly, we find that

$$\begin{aligned}
d_{1,3} &= \sqrt{\|\boldsymbol{x}_1 - \boldsymbol{x}_3\|^2} = \sqrt{\|(1,0,-2,0)\|^2} = \sqrt{5} \\
d_{1,4} &= \sqrt{\|\boldsymbol{x}_1 - \boldsymbol{x}_4\|^2} = \sqrt{\|(1,1,1,-3)\|^2} = \sqrt{12} \\
d_{2,3} &= \sqrt{\|\boldsymbol{x}_2 - \boldsymbol{x}_3\|^2} = \sqrt{\|(-3,2,0,1)\|^2} = \sqrt{14} \\
d_{2,4} &= \sqrt{\|\boldsymbol{x}_2 - \boldsymbol{x}_4\|^2} = \sqrt{\|(-3,3,3,-2)\|^2} = \sqrt{31} \\
d_{3,4} &= \sqrt{\|\boldsymbol{x}_3 - \boldsymbol{x}_4\|^2} = \sqrt{\|(0,1,3,-3)\|^2} = \sqrt{19}
\end{aligned}$$

Therefore, the minimum distance is $\sqrt{5}$.

Question 2

Determine the signal space representation of the four signals $x_m(\cdot), m = 1, \ldots, 4$ shown in Figure 3 using as basis functions $f_1(\cdot)$ and $f_2(\cdot)$. Plot the signal space diagram and show that it is equivalent to that of QPSK modulation.

Using the basis given in the figure, we can represent the signals using the following vectors

$$\boldsymbol{x}_1 = (\sqrt{E_x}, 0), \boldsymbol{x}_2 = (-\sqrt{E_x}, 0), \boldsymbol{x}_3 = (0, \sqrt{E_x}), \boldsymbol{x}_4 = (0, -\sqrt{E_x})$$

This is precisely the signal set representation of QPSK modulation, using the standard basis (see Figure 4).

Question 3

A binary digital communications system employs the signals

$$\begin{aligned}
x_0(t) &= 0, \quad t \in \mathbb{R} \\
x_1(t) &= \begin{cases} A, & 0 \le t \le T \\ 0, & \text{otherwise} \end{cases}
\end{aligned}$$

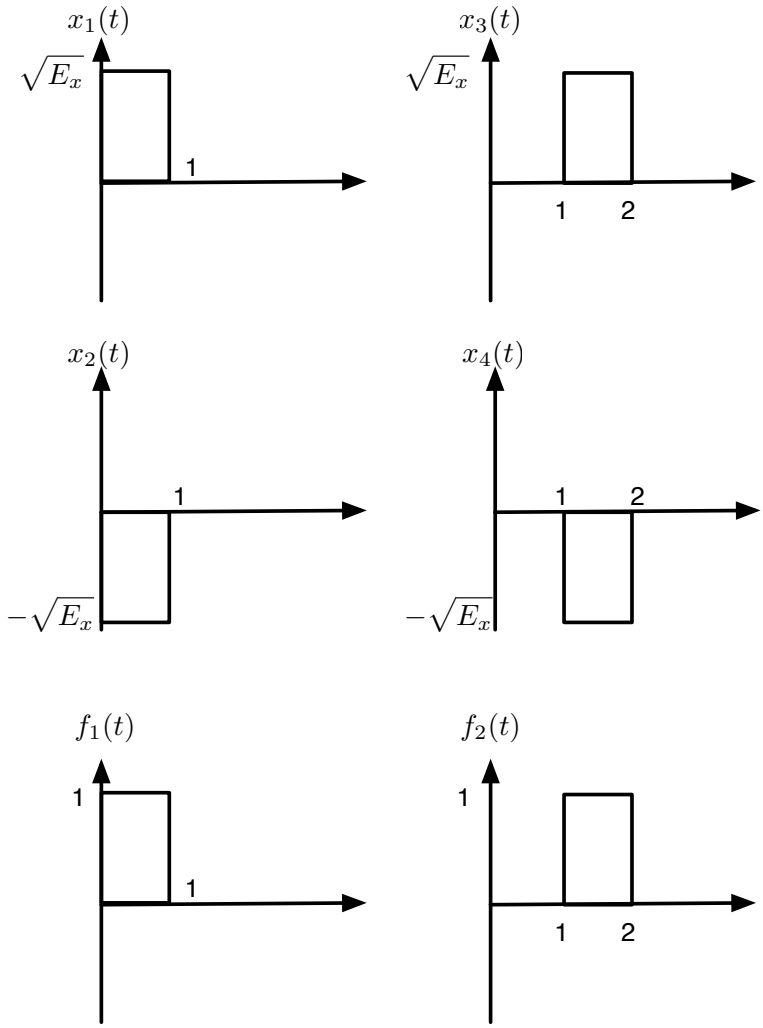


Figure 3: Signal set for Question 2.

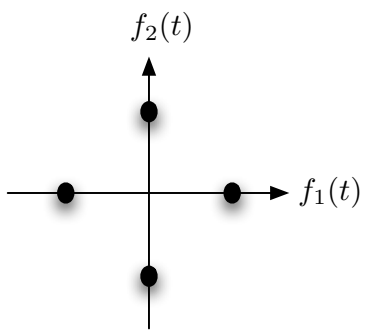


Figure 4: Signal space for Question 2.

for transmission of information. This is called on-off *signalling.*

(a) *What are optimum receiver structures?*

The possible receiver structures consist of either

(a) A bank of correlators that calculate $\int_{-\infty}^{\infty} y(t) f_k(t) dt$ for $k = 1, \ldots, K$

(b) A bank of K matched filters with responses $f_k(-t)$, $t \in \mathbb{R}$ and samplers at $t = 0$

together with an optimum detector (see Figure 5).

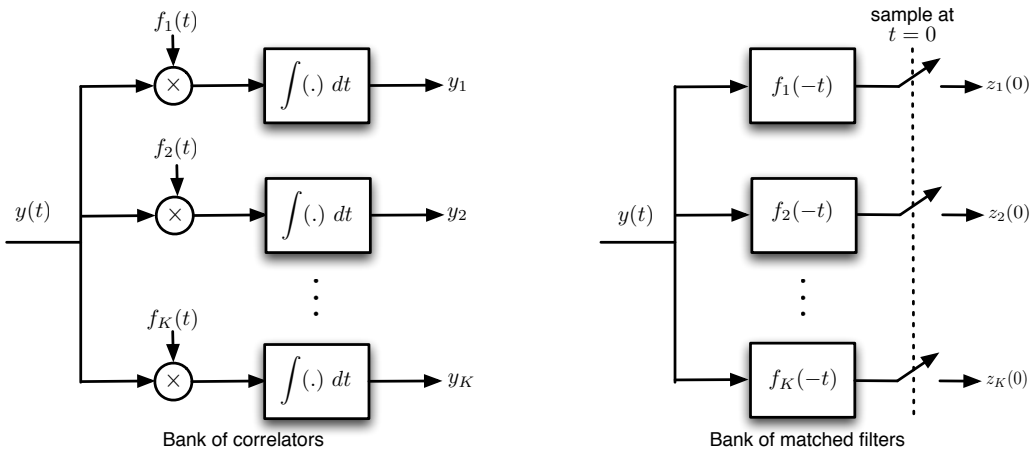


Figure 5: Receiver structures.

In this case, the basis function is

$$f_1(t) = \begin{cases} \frac{1}{\sqrt{T}} & 0 \le t \le T \\ 0 & \text{otherwise} \end{cases}$$

(b) *Determine the optimum detector for the AWGN channel, assuming that the signals are equiprobable.*

At the output of the correlator we have

$$y = x_m + n$$

where $x_m \in \{0, A\sqrt{T}\}$ and the noise has zero mean and variance $\frac{N_0}{2}$. Then, the channel transition probability densities are

$$p(y|x=0) = \frac{1}{\sqrt{\pi N_0}} e^{-\frac{1}{N_0} y^2}, \quad p(y|x = A\sqrt{T}) = \frac{1}{\sqrt{\pi N_0}} e^{-\frac{1}{N_0}(y - A\sqrt{T})^2}$$

Since the signals are equally likely, the optimum detector decides in favour of $x = 0$ whenever

$$\frac{p(y|x=0)}{p(y|x = A\sqrt{T})} > 1$$
$$y > \frac{A\sqrt{T}}{2}.$$

This implies that the optimum decision threshold is $\frac{A\sqrt{T}}{2}$, as intuition suggests.

(c) *Determine the probability of error as a function of SNR.*

The error probability is determined following the standard steps

$$
\begin{aligned}
P_e &= \frac{1}{2}\Pr\{\text{error}|x=0\} + \frac{1}{2}\Pr\{\text{error}|x=A\sqrt{T}\} \\
&= \frac{1}{2}\int_{\frac{A\sqrt{T}}{2}}^{\infty} p(y|x=0)dy + \frac{1}{2}\int_{-\infty}^{\frac{A\sqrt{T}}{2}} p(y|x=A\sqrt{T})dy \\
&= \frac{1}{2}\int_{\frac{A\sqrt{T}}{2}}^{\infty} \frac{1}{\sqrt{\pi N_0}} e^{-\frac{1}{N_0}y^2} dy + \frac{1}{2}\int_{-\infty}^{\frac{A\sqrt{T}}{2}} \frac{1}{\sqrt{\pi N_0}} e^{-\frac{1}{N_0}(y-A\sqrt{T})^2} dy \\
&\frac{1}{2}\int_{\frac{A\sqrt{T}}{2}\sqrt{\frac{2}{N_0}}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx + \frac{1}{2}\int_{-\infty}^{-\frac{A\sqrt{T}}{2}\sqrt{\frac{2}{N_0}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx \\
&= Q\left(\frac{A\sqrt{T}}{2}\sqrt{\frac{2}{N_0}}\right) = Q\left(\sqrt{\frac{A^2T}{2N_0}}\right) = Q\left(\sqrt{\frac{\text{SNR}}{4}}\right)
\end{aligned}
$$

where $\text{SNR} = \frac{A^2T}{N_0/2}$.

(d) *How does on-off signalling compare with antipodal signalling, $x_0(t) = -A$, $x_1(t) = A$ for $0 \le t \le T$ and $x_0(t) = x_1(t) = 0$ otherwise?*

Following the same steps for antipodal signalling we obtain that

$$P_e = Q\left(\sqrt{\text{SNR}}\right).$$

Therefore, on-off signalling requires a factor of four more energy than antipodal signalling. This is not surprising, since on-off signalling with amplitude A is equivalent to antipodal signalling with amplitude $A/2$ (just shift the whole signal constellation by $A/2$ to the left).

Note, however, that this comparison is only on the basis of peak SNR and not average SNR. If we normalize the amplitudes such that both signalling schemes use the same average SNR, then on-off signalling requires a factor of two more energy than antipodal signalling. Indeed, the average SNR is given by

$$
\begin{aligned}
\text{SNR} &= \frac{A^2T}{N_0}, \quad \text{on-off signalling} \\
\text{SNR} &= \frac{A^2T}{N_0/2}, \quad \text{antipodal signalling.}
\end{aligned}
$$

Thus, in order to have the same average SNR, the squared amplitude of on-off signalling needs to be twice as large as the squared magnitude of antipodal signalling. The error probability is then given by

$$
\begin{aligned}
P_e &= Q\left(\sqrt{\frac{2A^2T}{2N_0}}\right) = Q\left(\sqrt{\frac{\text{SNR}}{2}}\right), \quad \text{on-off signalling} \\
P_e &= Q\left(\sqrt{\frac{A^2T}{N_0/2}}\right) = Q\left(\sqrt{\text{SNR}}\right), \quad \text{antipodal signalling.}
\end{aligned}
$$

Question 4

Suppose the energy limited signal $x(\cdot)$ is corrupted by the AWGN $n(\cdot)$. Hence, the observed signal is

$$y(t) = x(t) + n(t), \quad t \in \mathbb{R}.$$

The received signal is passed through a filter whose impulse response is $t \mapsto h(t)$. Find the filter $h(\cdot)$ that maximises the SNR at its output at $t = 0$.

Let the filter output be denoted b $\tilde{y}(\cdot)$. Then, the SNR is given by

$$\text{SNR} = \frac{|\tilde{y}(0)|^2}{\mathbb{E}[|\tilde{n}(0)|^2]}$$

where $\tilde{n}(\cdot)$ is the filtered noise. The numerator is expressed as

$$|\tilde{y}(0)|^2 = \left| \int_{-\infty}^{\infty} x(t)h(-t)dt \right|^2$$

while the denominator is

$$\mathbb{E}[|\tilde{n}(0)|^2] = \int_{-\infty}^{-\infty} \int_{-\infty}^{\infty} \mathbb{E}[n(a)n^*(b)]h(-a)h^*(-b)dadb = 2N_0 \int_{-\infty}^{\infty} |h(-t)|^2 dt.$$

Therefore we want to maximise

$$\text{SNR} = \frac{\left| \int_{-\infty}^{\infty} x(t)h(-t)dt \right|^2}{2N_0 \int_{-\infty}^{\infty} |h(-t)|^2 dt}.$$

Using the Cauchy-Schwarz inequality we obtain that

$$\left| \int_{-\infty}^{\infty} x(t)h(-t)dt \right|^2 \leq \int_{-\infty}^{\infty} |h(-t)|^2 dt \int_{-\infty}^{\infty} |x(t)|^2 dt.$$

Therefore

$$\text{SNR} \leq \frac{1}{2N_0} \int_{-\infty}^{\infty} |x(t)|^2 dt = \frac{E_x}{N_0} = \text{SNR}_{\max}$$

so the maximum occurs when

$$x(t) = h^*(-t), \quad t \in \mathbb{R} \qquad \text{or equivalently} \qquad h(t) = x^*(-t), \quad t \in \mathbb{R}.$$

Question 5

Suppose a signal set $x_m(\cdot), m = 1, \ldots, M$ is transmitted over an AWGN channel with noise variance $\sigma^2 = \frac{N_0}{2}$. Hence, the received signal is

$$y(t) = x_m(t) + n(t), \quad t \in \mathbb{R}.$$

A correlator receiver using the K signal-space basis functions $f_k(\cdot)$, $k = 1, \ldots, K$ is employed, producing

$$y_k = \langle y(t), f_k(t) \rangle = x_{m,k} + n_k, \quad k = 1, \ldots, K.$$

Show that

(a) $\mathbb{E}[n_k] = 0$

This is bookwork from the notes:

$$\mathbb{E}[n_k] = \int_{-\infty}^{\infty} \mathbb{E}[n(t)] f_k(t) dt = 0.$$

(b) $\mathbb{E}[n_k n_l] = \frac{N_0}{2}\delta_{k,l}$

This is bookwork from the notes:

$$\mathbb{E}[n_k n_l] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbb{E}[n(t)n(\tau)] f_k(t) f_l(\tau) dt\, d\tau = \frac{N_0}{2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_k(t) f_l(\tau) dt\, d\tau = \frac{N_0}{2}\delta_{k,l}.$$

(c) $\mathbb{E}[n'(t)y_k] = 0$ *where* $n'(t) = n(t) - \sum_{k=1}^{K} n_k f_k(t)$ *is the AWGN component that does not lie within the signal space.*

This is bookwork from the notes:

$$\begin{aligned}\mathbb{E}[n'(t)y_k] = \mathbb{E}[n'(t)n_k] &= \mathbb{E}\left[\left(n(t) - \sum_{j=1}^{K} n_j f_j(t)\right) n_k\right] \\ &= \int_{-\infty}^{\infty} \mathbb{E}[n(t)n(\tau)] f_k(\tau) d\tau - \sum_{j=1}^{K} \mathbb{E}[n_j n_k] f_j(t) = \frac{N_0}{2} f_k(t) - \frac{N_0}{2} f_k(t) = 0.\end{aligned}$$

(d) *Show that the optimum demodulator/detector can be obtained as shown in Figure 6, where* $E_m = \|x_m(t)\|^2$, $m = 1, \ldots, M$.

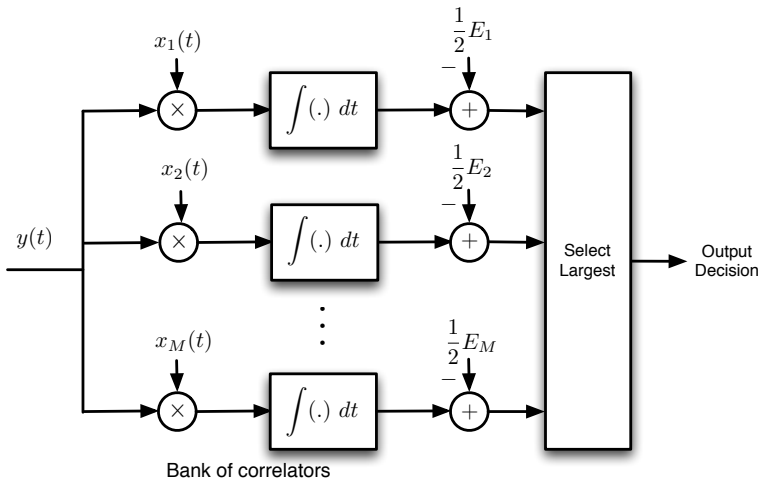


Figure 6: Detector for Question 5.

The optimum detector decides in favour of

$$\hat{\boldsymbol{x}} = \arg\min_{m=1,\ldots,M} \|\boldsymbol{y} - \boldsymbol{x}_m\|^2 = \arg\min_{m=1,\ldots,M} \|\boldsymbol{y}\|^2 + \|\boldsymbol{x}_m\|^2 - 2\boldsymbol{y}^T\boldsymbol{x}_m.$$

Since $\|\boldsymbol{y}\|^2$ is common to all terms, we have that

$$\begin{aligned}\hat{\boldsymbol{x}} &= \arg\min_{m=1,\ldots,M} \|\boldsymbol{x}_m\|^2 - 2\boldsymbol{y}^T\boldsymbol{x}_m = \arg\max_{m=1,\ldots,M} 2\boldsymbol{y}^T\boldsymbol{x}_m - \|\boldsymbol{x}_m\|^2 \\ &= \arg\max_{m=1,\ldots,M} \int_{-\infty}^{\infty} y(t)x_m(t)dt - \frac{1}{2}\int_{-\infty}^{\infty} x_m^2(t)dt\end{aligned}$$

where the last equality follows from the orthonormality of the basis functions when expanding the transmitted and received signals as a linear combination of the basis functions, i.e., $y(t) = \sum_{k=1}^{K} y_k f_k(t)$ and $x_m(t) = \sum_{k=1}^{K} x_{m,k} f_k(t)$. The block-diagram therefore follows from the last equation.

Question 6

Show the following results:

(a) *Chain rule for entropy:*

$$H(X,Y) = H(X) + H(Y|X)$$

$$\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{X,Y}(x,y)\log_2 P_{X,Y}(x,y) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{X,Y}(x,y)\log_2\big(P_{Y|X}(y|x)P_X(x)\big) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{X,Y}(x,y)\left(\log_2 P_{Y|X}(y|x) + \log_2 P_X(x)\right) \\
&= H(Y|X) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{X,Y}(x,y)\log_2 P_X(x) \\
&= H(Y|X) - \sum_{x\in\mathcal{X}} P_X(x)\log_2 P_X(x) \\
&= H(X) + H(Y|X)
\end{aligned}$$

(b) *Positiveness*

$$I(X;Y) \geq 0$$

Note that

$$I(X;Y) = D(P_{X,Y}(X,Y)\|P_X(X)P_Y(Y)).$$

It thus suffices to show that the Kullback-Leibler divergence is positive:

$$\begin{aligned}
-D(P\|Q) &= -\sum_{x\in\mathcal{X}} P_X(x)\log_2\frac{P_X(x)}{Q_X(x)} \\
&= \sum_{x\in\mathcal{X}} P_X(x)\log_2\frac{Q_X(x)}{P_X(x)} \\
&\leq \log_2\left(\sum_{x\in\mathcal{X}} P_X(x)\frac{Q_X(x)}{P_X(x)}\right) \\
&= \log_2\left(\sum_{x\in\mathcal{X}} Q_X(x)\right) = \log_2 1 = 0
\end{aligned}$$

(c) *Conditioning reduces entropy:*

$$H(X) \geq H(X|Y)$$

Based on the positiveness of mutual information, this property can be easily seen from the following

$$0 \leq I(X;Y) = H(X) - H(X|Y)$$

(d) *Chain rule for mutual information:*

$$I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y|X_1)$$

Using the chain rule for entropy, we write

$$\begin{aligned} I(X_1, X_2; Y) &= H(X_1, X_2) - H(X_1, X_2|Y) \\ &= H(X_1) + H(X_2|X_1) - H(X_1|Y) - H(X_2|Y, X_1) \\ &= I(X_1; Y) + I(X_2; Y|X_1) \end{aligned}$$

(e) *Data Processing Inequality: Let $X \to Y \to Z$ form a Markov chain. Show that*

$$I(X; Z) \leq I(X; Y).$$

The Markov chain $X \to Y \to Z$ is equivalent to $P_{X,Y,Z}(x, y, z) = P_X(x)P_{Y|X}(y|x)P_{Z|Y}(z|y)$. Using the chain rule for mutual information, we expand the mutual information in two different ways, i.e.,

$$\begin{aligned} I(X; Y, Z) &= I(X; Z) + I(X; Y|Z) \geq I(X; Z) \\ &= I(X; Y) + I(X; Z|Y) = I(X; Y) \end{aligned}$$

where the second term of the first equation is positive (from the positiveness of mutual information) and the second term of the second equation is zero from the Markov chain property. This proves the Data Processing Inequality.

(f) *Let $\boldsymbol{X}, \boldsymbol{Y}$ be n-dimensional random vectors over $\mathcal{X}^n, \mathcal{Y}^n$. Assume that*

$$P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} P_{Y|X}(y_i|x_i).$$

Show that, for any distribution on $\boldsymbol{X}$,

$$I(\boldsymbol{X}; \boldsymbol{Y}) \leq \sum_{i=1}^{n} I(X_i; Y_i).$$

We write the mutual information as

$$\begin{aligned} I(\boldsymbol{X}; \boldsymbol{Y}) &= H(\boldsymbol{Y}) - H(\boldsymbol{Y}|\boldsymbol{X}) \\ &= \sum_{i=1}^{n} H(Y_i|Y_{i-1}, \ldots, X_1) - \sum_{i=1}^{n} H(Y_i|Y_{i-1}, \ldots, Y_1, \boldsymbol{X}) \\ &\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|Y_{i-1}, \ldots, Y_1, \boldsymbol{X}) \\ &= \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \\ &= \sum_{i=1}^{n} I(X_i; Y_i) \end{aligned}$$

where the third line follows from (c), and the fourth line follows because, by assumption, Y_i only depends on X_i and not on $(Y_{i-1}, \ldots, Y_1, \boldsymbol{X})$.

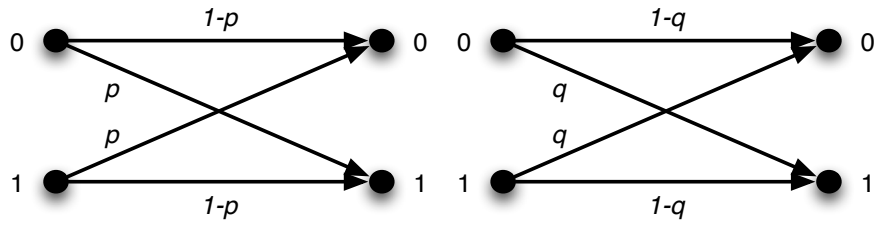


Figure 7: BSC channel cascade for Question 7.

Question 7

Consider the cascade of channels shown in Figure 7. Show that the channel capacity of the cascade of channels cannot be larger than the channel capacity of the individual binary symmetric channels. Determine the channel capacity of the cascade of channels.

Let X denote the input of the first channel, let Y denote the output of the first channel (and hence also the input of the second channel), and let Z denote the output of the second. Furthermore, let $X_\star$ denote the input corresponding to the capacity-achieving input distribution of the cascade of channels, and let $Y_\star$ denote the Y that corresponds to $X_\star$. Then

$$C = I(X_\star; Z) \leq I(X_\star; Y) \leq \max_{P_X(\cdot)} I(X; Y) = C(\text{first channel})$$

where the first inequality follows from the Data Processing Inequality, and the second inequality follows by maximizing over $P_X(\cdot)$. Likewise, we obtain

$$C = I(X_\star; Z) \leq I(Y_\star; Z) \leq \max_{P_Y(\cdot)} I(Y; Z) = C(\text{second channel}).$$

This proves the first part.

To prove the second part, we note that the cascade of channels is equivalent to a BSC with crossover probability

$$p' = (1-p)q + p(1-q).$$

Hence the capacity is $C = 1 - H_b(p')$.

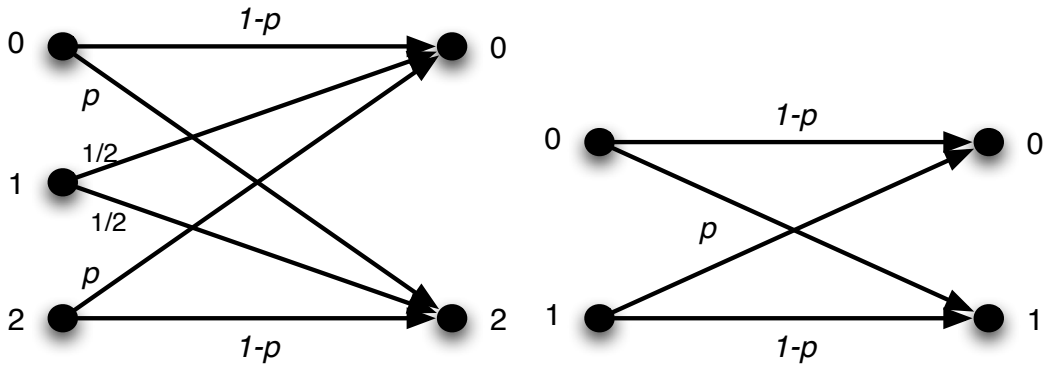
Question 8

What is the channel capacity and the capacity achieving input distribution of the ternary-input binary-output channel given in Figure 8(a)? Compare this capacity to the channel capacity of the binary symmetric channel depicted in Figure 8(b). (Recall that the channel capacity of the binary symmetric channel with cross-over probability p is given by $C = 1 - H_b(p)$, where $H_b(\cdot)$ is the binary entropy function.) What is the capacity ad capacity-achieving distribution of the ternary-input binary-output channel given in Figure 8(a).

We note that symbol 1 in Figure 8(a) is always mistaken by either 0 or 2. Hence the optimum input distribution is the one that never uses that symbol to transmit information ($P_X(1) = 0$) converting the channel into a BSC with crossover probability p. Thus, the capacity of both channels is $C = 1 - H_b(p)$.

Question 9

Consider the three channels that are depicted in Figure 9.



(a) Ternary-input binary-output channel (b) BSC

Figure 8: BSC and ternary-input binary-output channel for Question 8.

(a) *Find the capacity of channel 1. What input distribution achieves the capacity?*

The capacity is $C = \max_{P_X(\cdot)} I(X;Y) = \max_{P_X(\cdot)}\big(H(Y) - H(Y|X)\big)$. Since $H(Y|X) = 0$ (given X there is no uncertainty about Y), the capacity is $C_1 = \max_{P_X(\cdot)} H(Y) = 1$, which is achieved for equiprobable inputs.

(b) *Find the capacity of channel 2. What input distribution achieves the capacity?*

Let $P_X(0) = q$ and $P_X(1) = 1 - q$. We have

$$\begin{aligned} H(Y|X) &= \sum_x P_X(x) H(Y|X = x) \\ &= q\underbrace{H(Y|X = 0)}_{=0} + (1-q)H(Y|X = 1) \\ &= (1-q)H(Y|X = 1) = (1-q)\underbrace{H_b\left(\frac{1}{2}\right)}_{=1} = 1 - q \end{aligned}$$

We will now calculate $H(Y)$. The probability mass function of the output is

$$\begin{aligned} P_Y(c) &= qP(Y = c|X = 0) + (1-q)P(Y = c|X = 1) = \frac{1}{2} + \frac{q}{2} \\ P_Y(d) &= \frac{1}{2}(1-q) = \frac{1}{2} - \frac{q}{2} \end{aligned}$$

Then, $H(Y) = H_b\left(\frac{1}{2} + \frac{q}{2}\right)$ and therefore

$$C_2 = \max_q \left\{ h\left(\frac{1}{2} + \frac{q}{2}\right) - (1-q) \right\}$$

Computing the derivative of C_2 and setting it to zero yields $q_\star = \frac{3}{5}$. Hence the capacity is $C_2 = 0.3219$ bits/channel use.

(c) *Let C_3 denote the capacity of the third channel, and let C_1 and C_2 represent the capacities of the first and second channel. Which of the following relations holds true and why?*

(i) $C < \frac{1}{2}(C_1 + C_2)$

(ii) $C = \frac{1}{2}(C_1 + C_2)$

(iii) $C > \frac{1}{2}(C_1 + C_2)$

The transition probability matrix of the third channel can be written as $\boldsymbol{P}_3 = \frac{1}{2}\boldsymbol{P}_1 + \frac{1}{2}\boldsymbol{P}_2$, where $\boldsymbol{P}_1, \boldsymbol{P}_2$ are the transition probability matrices of channel 1 and channel 2 respectively. Here we have assumed that the output space of both channels is $\{a, b, c, d\}$. (For $\boldsymbol{P}_1$, the transition probabilities to c and d are zero, whereas for $\boldsymbol{P}_2$ the transition probabilities to a and b are zero.) Using the convexity of the mutual information with respect to the transition matrix $\boldsymbol{P}$ we have that

$$\begin{aligned} C_3 &= \max_{P_X(\cdot)} I(X;Y) = \max_{P_X(\cdot)} I(P_X(X), \boldsymbol{P}) \\ &= \max_{P_X(\cdot)} I\left(P_X, \frac{1}{2}\boldsymbol{P}_1 + \frac{1}{2}\boldsymbol{P}_2\right) \\ &\leq \max_{P_X(\cdot)} \frac{1}{2} I\left(P_X, \boldsymbol{P}_1\right) + \max_{P_X(\cdot)} \frac{1}{2} I\left(P_X, \boldsymbol{P}_2\right) \\ &\leq \frac{1}{2}C_1 + \frac{1}{2}C_2 \end{aligned}$$

Since $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are different, the inequality is strict, i.e., $C_3 < \frac{1}{2}C_1 + \frac{1}{2}C_2$.

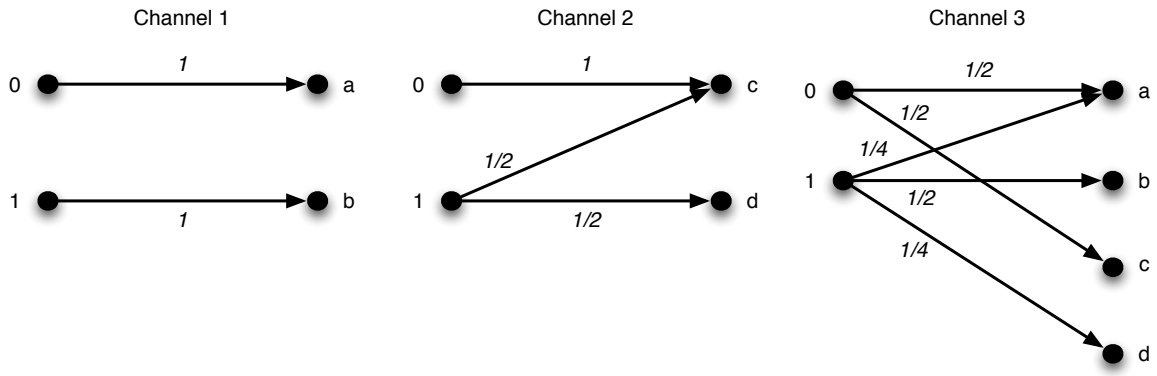


Figure 9: Channels for Question 9.

Question 10

Let C denote the capacity of a discrete memoryless channel with input alphabet $\mathcal{X} = \{x_1, \ldots, x_M\}$ and output alphabet $\mathcal{Y} = \{y_1, \ldots, y_N\}$. Show that $C \leq \min\{\log_2 M, \log_2 N\}$.

The capacity of a channel is

$$C = \max_{P_X(\cdot)} I(X;Y) = \max_{P_X(\cdot)} \big(H(Y) - H(Y|X)\big) = \max_{P_X(\cdot)} \big(H(X) - H(X|Y)\big).$$

Since $H(Y|X) \geq 0$ and $H(X|Y) \geq 0$, we obtain that

$$C \leq \min\{\max_{P_X(\cdot)} H(X), \max_{P_X(\cdot)} H(Y)\}$$

Note that the the maxima of the entropies $H(X), H(Y)$ are attained when the corresponding random variables are uniformly distributed with probabilities $P_X(x) = \frac{1}{|\mathcal{X}|}$ and $P_Y(y) = \frac{1}{|\mathcal{Y}|}$. The value of these maxima are $\max_{P_X(\cdot)} H(X) = \log_2 |\mathcal{X}|, \max_{P_X(\cdot)} H(Y) \leq \log_2 |\mathcal{Y}|$. Hence the result follows.

Question 11

(Converse to Fano's inequality). Suppose we want to estimate $X \in \mathcal{X}$ from the observation $Y \in \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are some finite sets. Fano's inequality states that, for any estimator $\hat{X} = g(Y)$ such that $X \to Y \to \hat{X}$ forms a Markov chain, we have that

$$H(X|Y) \leq H_b(P_e) + P_e \log_2(|\mathcal{X}| - 1)$$

where $P_e = \Pr(\hat{X} \neq X)$, and where $H_b(\cdot)$ is the binary entropy function. Using the graphical approach presented in the lecture notes, can you state a lower bound on $H(X|Y)$ in function of P_e to complement Fano's inequality?

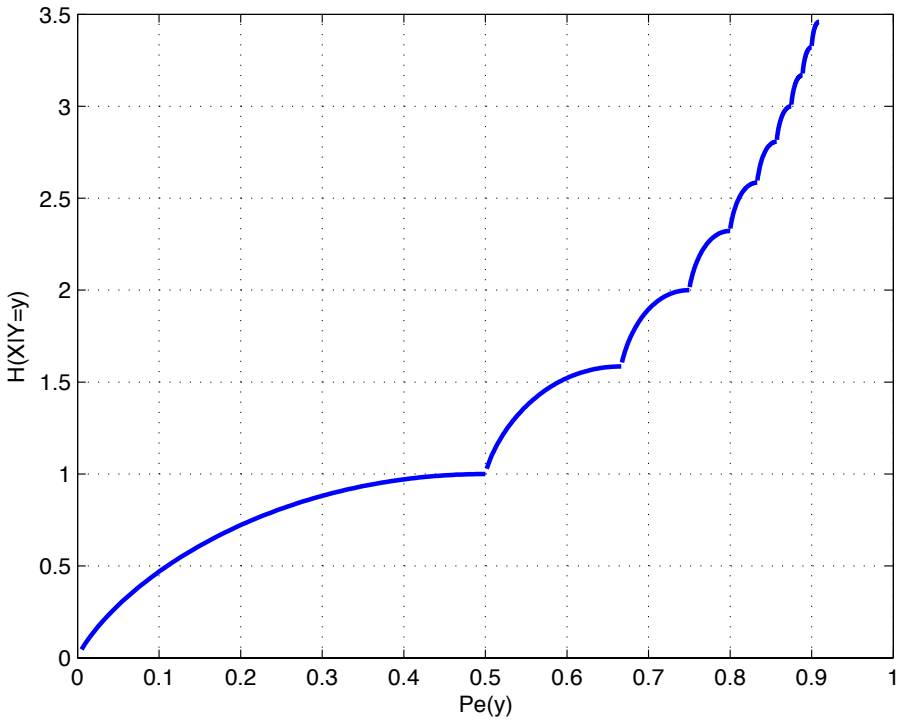


Figure 10: Lower bound on $H(X|Y = y)$ in function of $P_e(y)$

We pick up the derivation in the lecture notes on page 19, where it is stated that "Maximising $H(X|Y = y)$ for a given $P_e(y)$ is equivalent to maximising over all distributions for which $\max_{x \in \mathcal{X}} P_{X|Y}(x|y) = 1 - P_e$". We are now interested in minimising $H(X|Y = y)$ instead. If $1 - P_e(y) \geq 0.5$, then the distribution minimising $H(X|Y = y)$ is the distribution that concentrates all remaining probability on the next largest symbol, i.e., re-ordered to be monotone decreasing, the probability distribution $(1 - P_e(y), P_e(y), 0, 0, \ldots)$. Its conditional entropy is $H_2(P_e(y))$.

However, if $1 - P_e(y) < 0.5$, then the distribution $(1 - P_e(y), P_e(y), 0, 0 \ldots)$ cannot be used because we assumed by definition that $1 - P_e(y)$ is the maximum likelihood decision probability, but in this case $P_e(y) > 1 - P_e(y)$ so the second symbol in the distribution $(1 - P_e(y), P_e(y), 0, 0, \ldots)$ is more probable than the supposed maximum likelihood symbol. For $1/3 \leq 1 - P_e(y) < 0.5$, we can instead concentrate all remaining probability on the next two symbols, yielding the re-ordered distribution $(1 - P_e(y), 1 - P_e(y), 2P_e(y) - 1, 0, 0, \ldots)$. In general, for $\frac{1}{k+1} \leq 1 - P_e(y) < \frac{1}{k}$, the distribution minimising $H(X|Y = Y)$ is the distribution

$$(\underbrace{1 - P_e(y), 1 - P_e(y), \ldots, 1 - P_e(y)}_{k \text{ times}}, kP_e(y) - k + 1, 0, 0, \ldots)$$

whose entropy is a lower bound for the conditional entropy, i.e.,

$$H(X|Y=y) \geq -k(1-P_e(y))\log(1-P_e(y)) - (kP_e - k + 1)\log(kP_e - k + 1).$$

Figure 10 plots this lower bound in function of P_e. Going back to the derivation of Fano's

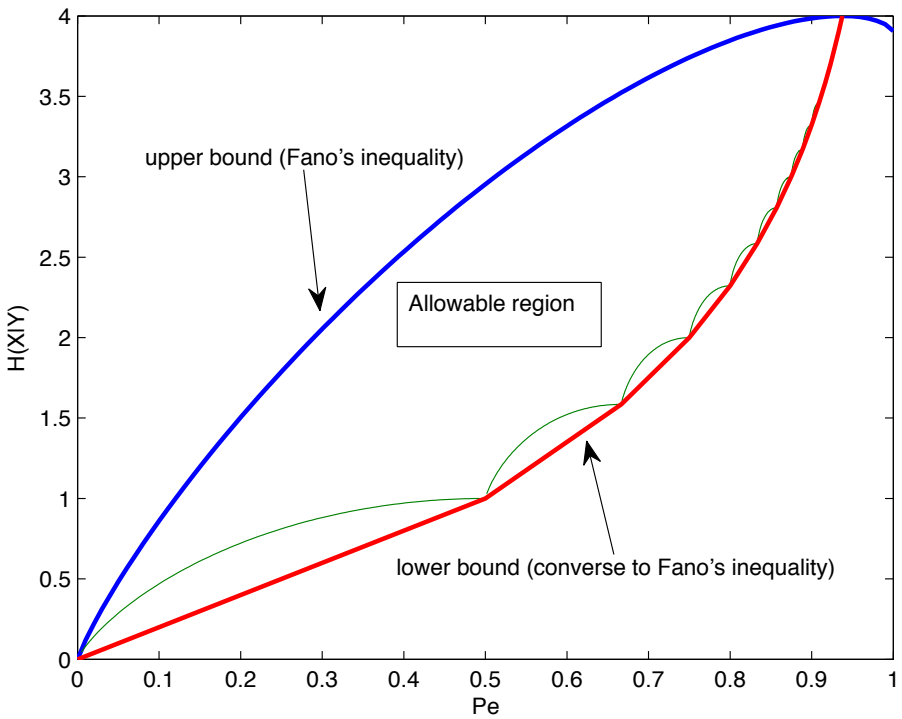


Figure 11: Lower and upper bound on $H(X|Y)$ in function of P_e

inequality in the lecture notes, we see that the step from an upper bound on the conditional entropy $H(X|Y=y)$ to an upper bound on the equivocation $H(X|Y)$ relied on the concavity of the upper bound. To mirror this derivation in view of providing a lower bound, we would now need the lower bound on $H(X|Y=y)$ to be *convex*. The plot in Figure 10 clearly shows that the lower bound is not convex, so the lower bound we computed for $H(X|Y=y)$ does not apply as is to $H(X|Y)$. In order to provide a lower bound, we need to take a convex envelope of our lower bound. This can be done by drawing the piecewise linear function joining up the points where the derivative is discontinuous, i.e., the points for which $P_e = 1 - 1/k$ for any k and for which the resulting entropy is simply $\log_2 k$, corresponding to the uniform distribution where the k most likely symbols all have probability $1 - P_e$ and the remaining symbols have probability 0. Thus, we have, for $\frac{1}{k+1} \leq 1 - P_e < \frac{1}{k}$,

$$H(X|Y) \geq \log k + P_e k(k+1)\log\left(1 + \frac{1}{k}\right).$$

This bound can be weakened to the logarithmic curve that passes through the points where the derivative is discontinuous, i.e.

$$H(X|Y) \geq -\log(1-P_e).$$

Figure 11 shows the piecewise linear lower bound on top of the original non-convex bound, as well as the upper bound derived in the lecture notes, both for an alphabet size of 16. Note that the lower bound correctly restricts the allowable region to $P_e \leq 15/16$, since it is not

possible that the maximum likelihood decision have probability less than 1/16 as there would then necessarily be another symbol with a higher probability. Thus, the conventional plots of Fano's inequality in most information theory textbooks (including your 4F5 lecture notes) showing it to rise monotonely to $\log|\mathcal{X}|$ then dipping back to $\log|\mathcal{X}-1|$ are in fact misguided and the correct upper bound should stop at $P_e = (|\mathcal{X}|-1)/|\mathcal{X}|$ and is monotone non-increasing in the relevant interval (i.e., the larger the error probability, the larger the equivocation).

Question 12

(a) *Suppose that X is a discrete random variable taking only nonnegative values. Show that for every $\delta > 0$*

$$\Pr(X \geq \delta) \leq \frac{\mathbb{E}[X]}{\delta}.$$

This is called Markov's inequality.

We have for every $\delta > 0$

$$\mathbb{E}[X] = \sum_{x\in\mathcal{X}} xP_X(x) = \sum_{x<\delta} xP_X(x) + \sum_{x\geq\delta} xP_X(x) \geq \sum_{x\geq\delta} xP_X(x) \geq \delta \sum_{x\geq\delta} P_X(x) = \delta \Pr(X \geq \delta)$$

where the first inequality follows because X is nonnegative, so the first sum is nonnegative, too; and the second inequality follows because we only sum over those x's that are greater or equal to δ. This proves Markov's inequality.

(b) *Use Markov's inequality to show that for any discrete random variable Y of variance σ_Y^2 we have for every $\delta > 0$*

$$\Pr\Big(\big(Y - \mathbb{E}[Y]\big)^2 \geq \delta\Big) \leq \frac{\sigma_Y^2}{\delta}.$$

This is called Chebyshev's inequality.

Let $X = (Y - \mathbb{E}[Y])^2$. Since X is a discrete random variable of mean $\mathbb{E}\big[(Y-\mathbb{E}[Y])^2\big] = \sigma_Y^2$ taking only nonnegative values, we can apply Markov's inequality to obtain

$$\Pr\Big(\big(Y - \mathbb{E}[Y]\big)^2 \geq \delta\Big) = \Pr(X \geq \delta) \leq \frac{\mathbb{E}[X]}{\delta} = \frac{\sigma_Y^2}{\delta}.$$

This proves Chebyshev's inequality.

(c) *Use Markov's inequality to show that for every $\beta > 0$*

$$\Pr\big(i(\bar{\boldsymbol{X}};\boldsymbol{Y}) > \log_2\beta\big) \leq \frac{1}{\beta}$$

where $i(x;y) \triangleq \log_2 \frac{P_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})}{P_{\boldsymbol{X}}(\boldsymbol{x})P_{\boldsymbol{Y}}(\boldsymbol{y})}$, and where $P_{\bar{\boldsymbol{X}},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}) = P_{\boldsymbol{X}}(\boldsymbol{x})P_{\boldsymbol{Y}}(\boldsymbol{y})$.

Let $Z = \frac{P_{\boldsymbol{X},\boldsymbol{Y}}(\bar{\boldsymbol{X}},\boldsymbol{Y})}{P_{\boldsymbol{X}}(\bar{\boldsymbol{X}})P_{\boldsymbol{Y}}(\boldsymbol{Y})}$. We first note that Z is a discrete random variable taking only nonnegative values and whose mean is equal to

$$\mathbb{E}[Z] = \sum_{\boldsymbol{x},\boldsymbol{y}} \underbrace{P_{\bar{\boldsymbol{X}},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})}_{=P_{\boldsymbol{X}}(\boldsymbol{x})P_{\boldsymbol{Y}}(\boldsymbol{y})} \frac{P_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})}{P_{\boldsymbol{X}}(\boldsymbol{x})P_{\boldsymbol{Y}}(\boldsymbol{y})} = \sum_{\boldsymbol{x},\boldsymbol{y}} P_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}) = 1.$$

We thus have

$$\Pr\big(i(\bar{\boldsymbol{X}};\boldsymbol{Y}) > \log_2\beta\big) = \Pr(Z > \beta) \leq \frac{\mathbb{E}[Z]}{\beta} = \frac{1}{\beta}$$

where the inequality follows from Markov's inequality.

(d) *Use Chebyshev's inequality to prove the weak law of large numbers. Thus, show that for every* $\epsilon > 0$

$$\lim_{n\to\infty} \Pr\left(\left|\frac{Z_1 + \ldots + Z_n}{n} - \mu\right| \geq \epsilon\right) = 0$$

where $Z_1, \ldots, Z_n$ *is a sequence of i.i.d. random variables of mean* μ *and variance* σ^2*:*

(i) *Compute the mean and variance of the random variable*

$$X_n \triangleq \frac{1}{n}\sum_{k=1}^{n} Z_k.$$

The mean of X_n is equal to

$$\mathbb{E}[X_n] = \frac{1}{n}\sum_{k=1}^{n} \underbrace{\mathbb{E}[Z_k]}_{=\mu} = \mu.$$

Since $Z_1, \ldots, Z_n$ are i.i.d., we have that the variance of $Z_1 + \ldots + Z_n$ is given by $n\sigma^2$. It follows that the variance of $X_n = \frac{Z_1+\ldots+Z_n}{n}$ is given by the variance of $Z_1 + \ldots + Z_n$ divided by n^2. Consequently, the variance of X_n is given by $\frac{\sigma^2}{n}$.

(ii) *For every* $n = 1, 2, \ldots,$ *upper-bound the probability*

$$\Pr\left(\left|\frac{Z_1 + \ldots + Z_n}{n} - \mu\right| \geq \epsilon\right), \quad \epsilon > 0$$

using Chebyshev's inequality. Show that the weak law of large numbers follows from this upper bound by letting n *tend to infinity.*

For every $n = 1, 2, \ldots$, the random variable X_n is discrete. Chebyshev's inequality yields therefore

$$\Pr\left(\left|\frac{Z_1 + \ldots + Z_n}{n} - \mu\right| \geq \epsilon\right) = \Pr\big((X_n - \mathbb{E}[X_n])^2 \geq \epsilon^2\big) \leq \frac{\sigma^2}{n\epsilon^2}$$

where we have used that $X_n = \frac{Z_1+\ldots+Z_n}{n}$ has mean μ and variance $\frac{\sigma^2}{n}$. Thus, for every $\epsilon > 0$, we obtain

$$0 \leq \lim_{n\to\infty} \Pr\left(\left|\frac{Z_1 + \ldots + Z_n}{n} - \mu\right| \geq \epsilon\right) \leq \lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} = 0$$

which proves the weak law of large numbers.

Question 13

Consider the parallel discrete memoryless channels shown in Figure 12. Show that the capacity of the parallel channels is equal to the sum of the capacities of the individual channels. Thus, show that

$$C_{1,\ldots,K} = \max_{P_{X_1,\ldots,X_K}} I(X_1, \ldots, X_K; Y_1, \ldots, Y_K) = \sum_{k=1}^{K} C_k$$

where $C_k = \max_{P_{X_k}} I(X_k; Y_k)$.

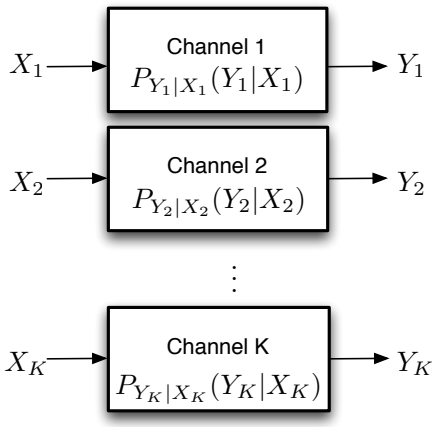


Figure 12: Channels for Question 13.

We first note that

$$P_{Y_1,\ldots,Y_K|X_1,\ldots,X_K}(\boldsymbol{y}|\boldsymbol{x}) = \prod_{k=1}^{K} P_{Y_k|X_k}(y_k|x_k).$$

We thus have by the chain rule (see also Part (f) of Question 6)

$$\begin{aligned}
I(X_1,\ldots,X_K;Y_1,\ldots,Y_K) &= \sum_{k=1}^{K} H(Y_k|Y_1,\ldots,Y_{k-1}) - \sum_{k=1}^{K} H(Y_k|Y_1,\ldots,Y_{k-1},X_1,\ldots,X_K) \\
&\leq \sum_{k=1}^{K} H(Y_k) - \sum_{k=1}^{K} H(Y_k|Y_1,\ldots,Y_{k-1},X_1,\ldots,X_K) \\
&= \sum_{k=1}^{K} H(Y_k) - \sum_{k=1}^{K} H(Y_k|X_k) \\
&= \sum_{k=1}^{K} I(X_k;Y_k).
\end{aligned}$$

Note that the inequality holds with equality if $Y_1,\ldots,Y_K$ are independent. Further note that this is the case, if $X_1,\ldots,X_K$ are independent. We thus obtain

$$C_{1,\ldots,K} = \max_{P_{X_1,\ldots,X_K}} I(X_1,\ldots,X_K;Y_1,\ldots,Y_K) \leq \max_{P_{X_1,\ldots,X_K}} \sum_{k=1}^{K} I(X_k;Y_k) \leq \sum_{k=1}^{K} \max_{P_{X_k}} I(X_k;Y_k) = \sum_{k=1}^{K} C_k$$

where we have equality if we choose $P_{X_1,\ldots,X_K}(x_1,\ldots,x_K) = \prod_{k=1}^{K} P_{X_k}(x_k)$ and if $P_{X_k}(\cdot)$ is such that it achieves the capacity of Channel k.