EE5139/EE6139
Solution 1

Semester I, AY2021/2022
Marco Tomamichel

Exercise 1.1 Expectation value and variance [EE5139]

Let V and W be discrete random variables defined on some probability space with a joint pmf $P_{VW}(v,w)$. We do not assume independence.

a.) Prove that $\mathbb{E}[V+W] = \mathbb{E}[V] + \mathbb{E}[W]$.

Solution:

$$\begin{aligned}\mathbb{E}[V+W] &= \sum_{v,w}(v+w)p_{V,W}(v,w) = \sum_{v,w} v p_{V,W}(v,w) + \sum_{w} w p_{V,W}(v,w)\\ &= \sum_{v} v p_V(v) + \sum_{w} w p_W(w) = \mathbb{E}[V] + \mathbb{E}[W]\end{aligned}$$

b.) Prove that if V and W are independent, then $\mathbb{E}[VW] = \mathbb{E}[V]\mathbb{E}[W]$.

Solution:

$$\begin{aligned}\mathbb{E}[VW] &= \sum_{v,w} vw p_{V,W}(v,w) = \sum_{v,w} vw p_V(v) p_W(w)\\ &= \sum_{v} v p_V(v) \sum_{w} w p_W(w) = \mathbb{E}[V]\mathbb{E}[W]\end{aligned}$$

c.) Let V and W be independent and let σ_V^2 and σ_W^2 be their respective variances. Find the variance of $Z = V + W$.

Solution:

$$\begin{aligned}\sigma_Z^2 &= \mathbb{E}[Z^2] - \mathbb{E}[Z]^2 = \mathbb{E}[(V+W)^2] - \mathbb{E}[V+W]^2\\ &= \mathbb{E}[V^2 + W^2 + 2VW] - (\mathbb{E}[V]^2 + 2\mathbb{E}[V]\mathbb{E}[W] + \mathbb{E}[W]^2)\\ &= \mathbb{E}[V^2] - \mathbb{E}[V]^2 + \mathbb{E}[W^2] - \mathbb{E}[W]^2 = \sigma_V^2 + \sigma_W^2,\end{aligned}$$

where the second to last equality follows from the fact that $\mathbb{E}[VW] = \mathbb{E}[V]\mathbb{E}[W]$.

Exercise 1.2 Coin flips [EE5139]

Flip a fair coin four times. Let X be the number of Heads obtained, and let Y be the position of the first Heads i.e. if the sequence of coin flips is TTHT, then $Y = 3$, if it is THHH, then $Y = 2$. If there are no heads in the four tosses, then we define $Y = 0$.

a.) Model the experiment completely, i.e. define the sample space and the random variables X and Y as functions from that sample space.

Solution: The underlying sample space is the set

$$\Omega = \{TTTT, TTTH, \ldots, HHHH\}.$$

Each outcome $\omega \in \Omega$ can be mapped to $X(\omega)$ and $Y(\omega)$, e.g.,

$$\begin{aligned} X(TTTT) &= 0 \qquad & Y(TTTT) &= 0 \\ X(THHT) &= 2 \qquad & Y(THHT) &= 2 \\ X(TTTH) &= 1 \qquad & Y(TTTH) &= 4 \\ &\text{etc.} \end{aligned}$$

b.) Find the joint pmf of X and Y.

Solution: By listing all 16 elements of Ω, and computing X and Y for each we can see that, e.g.,

$$\{X = 2, Y = 2\} = \{THHT, THTH\}$$

and thus $P_{XY}(2,2) = \frac{1}{16} + \frac{1}{16} = \frac{1}{8}$. This can be repeated for all feasible values of X and Y as in the table below: We can now read off

			x		
y	0	1	2	3	4
0	$\frac{1}{16}$				
1		$\frac{1}{16}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{1}{16}$
2		$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	
3		$\frac{1}{16}$	$\frac{1}{16}$		
4		$\frac{1}{16}$			

c.) Using the joint pmf, find the marginal pmf of X. What is $\Pr[Y = 0|X = 1]$ and $\Pr[Y = 1|X = 3]$?

Solution: Summing over the columns of the table below, we get the pmf of X as

$$P_X(x) = \begin{cases} \frac{1}{16} & x = 0 \\ \frac{1}{4} & x = 1 \\ \frac{3}{8} & x = 2 \\ \frac{1}{4} & x = 3 \\ \frac{1}{16} & x = 4 \end{cases}$$

We can now use the Bayes' rule to compute

$$\Pr[Y = 0|X = 1] = \frac{P_{XY}(1,0)}{P_X(1)} = 0,$$

$$\Pr[Y = 1|X = 3] = \frac{P_{XY}(3,1)}{P_X(3)} = \frac{3}{16} \cdot \frac{4}{1} = \frac{3}{4}.$$

Exercise 1.3 Property of convex functions [EE5139]

Let f be convex on $[a, b]$. Using only the defining property of convex functions, show that for any $a \leq x_1 < x_2 \leq x_3 < x_4 \leq b$, we have

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq \frac{f(x_4) - f(x_3)}{x_4 - x_3}.$$

Solution: Recall that a function $f : (a,b) \to \mathbb{R}$ is convex on (a,b) if for all $x, y \in (a,b)$ with $x < y$ and $0 \le \lambda \le 1$,

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y). \tag{1}$$

For brevity in notation, we define $g(x,y) = (f(y) - f(x))/(y-x)$, so we must show that $g(x_1,x_2) \le g(x_3,x_4)$. We first show that for $x_1, x_2, x_3 \in (a,b)$ such that $x_1 < x_2 < x_3$, we have $g(x_1,x_2) \le g(x_2,x_3)$, i.e., that

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \le \frac{f(x_3) - f(x_2)}{x_3 - x_2}. \tag{2}$$

In (1), we let $x \leftarrow x_1$, $y \leftarrow x_3$ and $\lambda = (x_3 - x_2)/(x_3 - x_1) \in [0,1]$. Then we check that

$$\lambda x + (1-\lambda)y = \lambda x_1 + (1-\lambda)x_3 = \frac{x_3 - x_2}{x_3 - x_1} \cdot x_1 + \frac{x_2 - x_1}{x_3 - x_1} \cdot x_3 = x_2. \tag{3}$$

So we have

$$f(\lambda x + (1-\lambda)y) = f(x_2) \le \frac{x_3 - x_2}{x_3 - x_1} f(x_1) + \frac{x_2 - x_1}{x_3 - x_1} f(x_3) = \lambda f(x_1) + (1-\lambda) f(x_3). \tag{4}$$

This is equivalent to

$$\lambda f(x_2) + (1-\lambda) f(x_2) = \lambda f(x_1) + (1-\lambda) f(x_3), \tag{5}$$

or

$$\lambda(f(x_2) - f(x_1)) \le (1-\lambda)(f(x_3) - f(x_2)) \tag{6}$$

Recalling the definition of λ shows (2), i.e., that $g(x_1,x_2) \le g(x_2,x_3)$. By the same logic, for three points $x_2 < x_3 < x_4$, we also have $g(x_2,x_3) \le g(x_3,x_4)$. Putting these two inequalities yields $g(x_1,x_2) \le g(x_3,x_4)$ as desired.

Exercise 1.4 Finite fields [EE5139]

Derive the addition and multiplication tables for F_8 and F_9. You should use the construction described in the lecture notes and the irreducible polynomials $x^3 + x + 1$ for F_8 and $x^2 + 1$ for F_9.
Hint: You may want to use Matlab to solve this problem. However, you will need to compute some elements by hand to verify the computer-generated output.
Solution:

$+_{F_8}$	0	1	2	3	4	5	6	7
0	0	1	2	3	4	5	6	7
1	1	0	3	2	5	4	7	6
2	2	3	0	1	6	7	4	5
3	3	2	1	0	7	6	5	4
4	4	5	6	7	0	1	2	3
5	5	4	7	6	1	0	3	2
6	6	7	4	5	2	3	0	1
7	7	6	5	4	3	2	1	0

$+_{F_9}$	0	1	2	3	4	5	6	7	8
0	0	1	2	3	4	5	6	7	8
1	1	2	0	4	5	3	7	8	6
2	2	0	1	5	3	4	8	6	7
3	3	4	5	6	7	8	0	1	2
4	4	5	3	7	8	6	1	2	0
5	5	3	4	8	6	7	2	0	1
6	6	7	8	0	1	2	3	4	5
7	7	8	6	1	2	0	4	5	3
8	8	6	7	2	0	1	5	3	4

$\times_{F_8}$	0	1	2	3	4	5	6	7
0	0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6	7
2	0	2	4	6	3	1	7	5
3	0	3	6	5	7	4	1	2
4	0	4	3	7	6	2	5	1
5	0	5	1	4	2	7	3	6
6	0	6	7	1	5	3	2	4
7	0	7	5	2	1	6	4	3

$\times_{F_9}$	0	1	2	3	4	5	6	7	8
0	0	0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6	7	8
2	0	2	1	6	8	7	3	5	4
3	0	3	6	2	5	8	1	4	7
4	0	4	8	5	6	1	7	2	3
5	0	5	7	8	1	3	4	6	2
6	0	6	3	1	7	4	2	8	5
7	0	7	5	4	2	6	8	3	1
8	0	8	4	7	3	2	5	1	6

The MatLab Program generating above tables:

Listing 1: addition_table.m

```
p = 3;
degree = 1; % Thus the size of the field is p^(degree+1)
h = [1,0,1]; % The irridicible polynomial of degree 'degree+1'
% List all polynomials
F = zeros(p^(degree+1),degree+1);
for i = 1:p^(degree+1)
    m = i—1;
    for j = 1:degree+1
        F(i,j) = mod(m,p);
        m = (m—mod(m,p))/p;
    end
end
a_table = zeros(p^(degree+1));
for i = 1:p^(degree+1)
    for j = 1:p^(degree+1)
        f = mod(F(i,:)+F(j,:),p);
        [~,k] = ismember(f,F,'rows');
        a_table(i,j) = k—1;
    end
end
disp(a_table);
```

Listing 2: multiplication_table.m

```
p = 3;
degree = 1; % Thus the size of the field is p^(degree+1)
h = [1,0,1]; % The irridicible polynomial of degree 'degree+1'
% List all polynomials
F = zeros(p^(degree+1),degree+1);
for i = 1:p^(degree+1)
    m = i—1;
    for j = 1:degree+1
        F(i,j) = mod(m,p);
        m = (m—mod(m,p))/p;
    end
end
m_table = zeros(p^(degree+1));
```

```
for i = 1:p^(degree+1)
    for j = 1:p^(degree+1)
        f = polyMod(polyMultiply(F(i,:),F(j,:),p),h,p);
        [~,k] = ismember(f,F,'rows');
        m_table(i,j) = k-1;
    end
end
disp(m_table);
```

Listing 3: polyMultiply.m

```
function [ f ] = polyMultiply (g, h, p)
    % Multiply two polynomials on base field Fp where p is a prime number.
    g_degree = length(g)-1;
    h_degree = length(h)-1;
    f = zeros(1, g_degree + h_degree+1);
    for k = 0:(length(f)-1) % k-th order
        for a = max(0,k-h_degree):min(k,g_degree)
            b = k-a;
            f(k+1) = f(k+1) + g(a+1)*h(b+1);
        end
    end
    f = mod(f, p);
end
```

Listing 4: polyMod.m

```
function [ f ] = polyMod (g, h, p)
    % Find the modulo of polynomial g with respect to irriducible polynomial h on
         base field Fp where p is a prime number.
    g_length = length(g);
    h_length = length(h);
    while (max([g ~= 0].*[1:g_length]) >= h_length)
        d = max([g ~= 0].*[1:g_length]);
        C = g(d);
        h_shifted = zeros(size(g));
        h_shifted(d-h_length+1:d) = h;
        g = mod(g-C*h_shifted, p);
    end
    f = g(1: h_length-1);
end
```

Exercise 1.5 Continuous and discrete random variables [all]

Consider the following random experiment. A ball is thrown and lands after X meters, where X is distributed uniformly in the interval $[1, 2]$. It either stays there or bounces off and jumps again an additional distance of $\frac{1}{2}X$. The binary random variable $Y \in \{0, 1\}$ takes the value 0 (with probability 50%), indicating that the ball stays put, and 1 (with probability 50%), indicating that the ball jumps again. After this additional bounce the ball rests.

a.) Express the total distance Z that the ball travels in terms of X and Y. Compute and plot the pdf for Z.

Solution:

$$Z = X + \frac{1}{2}XY.$$

We may first compute the pmf for Z ending up in an interval around z, e.g. $[z \pm \epsilon] = [z-\epsilon, z+\epsilon]$.

$$\begin{aligned}
\Pr[Z \in [z \pm \epsilon]] &= \Pr[Z \in [z \pm \epsilon] | Y = 0] \Pr[Y = 0] + \Pr[Z \in [z \pm \epsilon] | Y = 1] \Pr[Y = 1] \\
&= \frac{1}{2} \Pr[X \in [z \pm \epsilon] | Y = 0] + \frac{1}{2} \Pr\left[\frac{3}{2}X \in [z \pm \epsilon] | Y = 1\right] \\
&= \frac{1}{2} \Pr[X \in [z \pm \epsilon]] + \frac{1}{2} \Pr\left[\frac{3}{2}X \in [z \pm \epsilon]\right].
\end{aligned}$$

For $[z \pm \epsilon] \subset [1, 3/2)$,

$$\Pr[Z \in [z \pm \epsilon]] = \frac{1}{2} \times 2\epsilon + 0 = \epsilon.$$

For $[z \pm \epsilon] \subset [3/2, 2]$,

$$\Pr[Z \in [z \pm \epsilon]] = \frac{1}{2} \times 2\epsilon + \frac{1}{2} \times \frac{4}{3}\epsilon = \frac{5}{3}\epsilon.$$

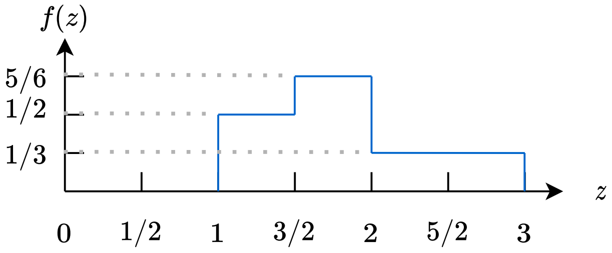
For $[z \pm \epsilon] \subset (2, 3]$,

$$\Pr[Z \in [z \pm \epsilon]] = \frac{1}{2} \times \frac{4}{3}\epsilon = \frac{2}{3}\epsilon.$$

The pdf can then be derived by taking the derivative (note that since we grow the interval on both sides a factor $\frac{1}{2}$ needs to be introduced).

$$f(z) = \lim_{\epsilon \to 0} \frac{\Pr\big[Z \in [z-\epsilon, z+\epsilon]\big]}{2\epsilon} = \begin{cases} \frac{1}{2}, & z \in (1, 3/2), \\ \frac{5}{6}, & z \in (3/2, 2), \\ \frac{1}{3}, & z \in (2, 3), \\ 0, & \text{otherwise.} \end{cases}$$

(Obviously, the pdf can also be found more directly.) On the boundary points the values of $f(z)$ are not uniquely defined — they depend on the convention used. If we define them as the limit above then it would simply be the average between the two regions adjoining the point. Plot:



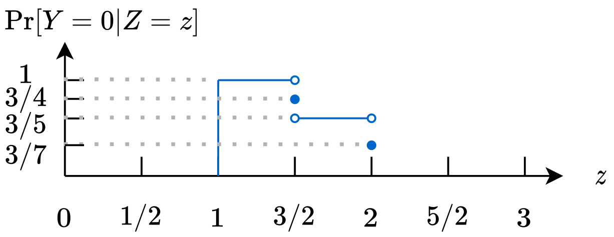
b.) Find the pmf for Y given $Z = z$. Plot $\Pr[Y = 0|Z = z]$ as a function of z.

Hint: We would be inclined to use Bayes' rule here, but the problem is that the $\Pr[Z = z] = 0$ for each z. To avoid this, consider an interval $z \pm \epsilon$ and compute the pmf for Y given $Z \in [z - \epsilon, z + \epsilon]$ and then let $\epsilon \to 0$.

Solution:

$$\begin{aligned}\Pr[Y = 0|Z = z] &= \lim_{\epsilon \to 0} \Pr[Y = 0|Z \in [z \pm \epsilon]] \\ &= \lim_{\epsilon \to 0} \frac{\Pr[Z \in [z \pm \epsilon]|Y = 0] \Pr[Y = 0]}{\Pr[Z \in [z \pm \epsilon]]} \\ &= \begin{cases} 1, & z \in [1, \frac{3}{2}) \\ \lim_{\epsilon \to 0} (2\epsilon \cdot \frac{1}{2})/\frac{4\epsilon}{3} = \frac{3}{4}, & z = \frac{3}{2} \\ \lim_{\epsilon \to 0} (2\epsilon \cdot \frac{1}{2})/\frac{5\epsilon}{3} = \frac{3}{5}, & z \in (\frac{3}{2}, 2) \\ \lim_{\epsilon \to 0} (\epsilon \cdot \frac{1}{2})/\frac{7\epsilon}{6} = \frac{3}{7}, & z = 2 \\ 0, & \text{otherwise.} \end{cases}\end{aligned}$$

Plot:



Exercise 1.6 Matrix representation of a communication channel [all]

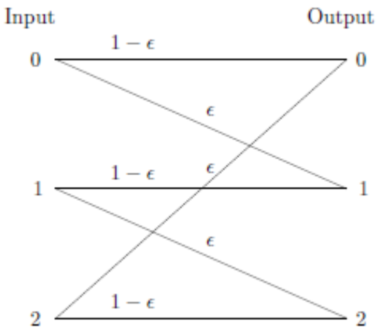


Figure 1: Ternary communication channel

A ternary communication channel is shown in Figure 1.

a.) Represent the channel as a matrix W such that the output distribution of the channel can be written as the matrix product pW, where p is a row vector containing the three input probabilities.

Solution:

$$W = \begin{bmatrix} 1-\epsilon & \epsilon & 0 \\ 0 & 1-\epsilon & \epsilon \\ \epsilon & 0 & 1-\epsilon \end{bmatrix}$$

b.) Suppose that the input probabilities are given by the vector $p = [\frac{1}{2},\ \frac{1}{4},\ \frac{1}{4}]$. Find the probabilities of the output symbols.

Solution:

$$pW = \left[\frac{1}{2} - \frac{\epsilon}{4},\quad \frac{1}{4} + \frac{\epsilon}{4},\quad \frac{1}{4}\right]$$

c.) Suppose that 1 was observed at the output. What's the probability that the input was 0? 1? 2?

Solution: We use the formula

$$P_{X|Y}(x|y) = \frac{P_{Y|X}(y|x)P_X(x)}{P_Y(y)}$$

with $y = 1$. Now $P_Y(1) = \frac{1}{4} + \frac{1}{4}\epsilon$ and so

$$\begin{aligned} P_{X|Y}(0|1) &= \frac{P_{Y|X}(1|0)P_X(0)}{\frac{1}{4}+\frac{1}{4}\epsilon} = \frac{\frac{1}{2}\epsilon}{\frac{1}{4}+\frac{1}{4}\epsilon} \\ P_{X|Y}(1|1) &= \frac{P_{Y|X}(1|1)P_X(1)}{\frac{1}{4}+\frac{1}{4}\epsilon} = \frac{\frac{1}{4}(1-\epsilon)}{\frac{1}{4}+\frac{1}{4}\epsilon} \\ P_{X|Y}(2|1) &= \frac{P_{Y|X}(1|2)P_X(2)}{\frac{1}{4}+\frac{1}{4}\epsilon} = 0 \end{aligned}$$

We may check that $\sum_{x=0}^{2} P_{X|Y}(x|1) = 1$.

Exercise 1.7 Further tail bounds [EE6139]

a.) For a nonnegative integer-valued random variable N, show that $\mathbb{E}[N] = \sum_{n>0} \Pr(N \geq n)$.

Solution: Solution:

$$\begin{aligned} \mathbb{E}[N] &= \sum_{n=0}^{\infty} np_N(n) = 0 \cdot p_N(0) + 1 \cdot p_N(1) + 2 \cdot p_N(2) + 3 \cdot p_N(3) + \dots && (7) \\ &= [p_N(1) + p_N(2) + p_N(3) + \dots] \\ &\quad + [p_N(2) + p_N(3) + p_N(4) + \dots] \\ &\quad + [p_N(3) + p_N(4) + p_N(5) + \dots] + \dots && (8) \\ &= \Pr(N \geq 1) + \Pr(N \geq 2) + \Pr(N \geq 3) + \dots = \sum_{n>0} \Pr(N \geq n) && (9) \end{aligned}$$

b.) Derive the Cauchy-Schwarz inequality, which says that $\mathbb{E}[AB] \leq \sqrt{\mathbb{E}[A^2]\mathbb{E}[B^2]}$.

Hint: Consider the non-negative random variable $(X - \alpha Y)^2$ and compute its expectation, then choose α appropriately.

Solution:

We have

$$0 \leq \mathbb{E}\left[(X - \alpha Y)^2\right] = \mathbb{E}[X^2] - 2\alpha\mathbb{E}[XY] + \alpha^2\mathbb{E}[Y^2]\,. \tag{10}$$

Choosing $\alpha = \mathbb{E}[XY]/\mathbb{E}[Y^2]$ yields

$$0 \leq \mathbb{E}[X^2] - 2\frac{\mathbb{E}[XY]^2}{\mathbb{E}[Y^2]} + \frac{\mathbb{E}[XY]^2}{\mathbb{E}[Y^2]}, \tag{11}$$

which equals the desired statement after multiplication with $\mathbb{E}[Y^2]$ on both sides.

c.) Derive the one-sided Cheybyshev inequality, which says that $\Pr(Y \geq a) \leq \sigma_Y^2/(\sigma_Y^2 + a^2)$ if $\mathbb{E}[Y] = 0$ and $a > 0$.

Solution: Since Y has zero mean, we have

$$a = \mathbb{E}[a - Y]$$

Consider the expectation above: We have

$$\mathbb{E}[a - Y] = \sum_y P_Y(y)(a - y) = \sum_{y:y<a} P_Y(y)(a - y) + \sum_{y:y\geq a} P_Y(y)(a - y) \leq \sum_{y:y<a} P_Y(y)(a - y) \tag{12}$$

because the second sum is non-positive. This can be written as

$$a \leq \sum_y P_Y(y)(a - y)\mathbf{1}\{y < a\} \tag{13}$$

where $\mathbf{1}\{\text{statement}\}$ returns 1 if the statement is true and 0 otherwise. Thus, we have

$$a \leq \mathbb{E}[(a - Y)\mathbf{1}\{Y < a\}]\mathbf{1} \tag{14}$$

Now square both sides,

$$a^2 \leq (\mathbb{E}[(a - Y)\mathbf{1}\{Y < a\}])^2 \tag{15}$$

Apply Cauchy-Schwarz inequality to the expectation,

$$a^2 \leq \mathbb{E}[(a - Y)^2]\mathbb{E}[\mathbf{1}\{Y < a\}^2] = \mathbb{E}[(a - Y)^2]\mathbb{E}[\mathbf{1}\{Y < a\}] = (a^2 + \mathbb{E}[Y^2])\Pr(Y < a) \tag{16}$$

Since $\mathbb{E}[Y^2] = \sigma_Y^2$, rearrangement of the above inequality yields the one-sided Chebyshev inequality as desired.

d.) Derive the reverse Markov inequality: Let X be a random variable such that $\Pr(X \leq a) = 1$ for some constant a. Then for $d < \mathbb{E}[X]$, we have

$$\Pr(X > d) \geq \frac{\mathbb{E}[X] - d}{a - d}.$$

Solution: Apply Markov's inequality to the non-negative random variable $\tilde{X} := a - X$. Then one has

$$\Pr(X \le d) = \Pr(a - \tilde{X} \le d) = \Pr(\tilde{X} \ge a - d) \le \frac{\mathbb{E}[\tilde{X}]}{a-d} = \frac{\mathbb{E}[a-X]}{a-d}$$

Hence,

$$\Pr(X < d) \ge 1 - = \frac{\mathbb{E}[a-X]}{a-d} = \frac{\mathbb{E}[X]-d}{a-d}.$$

e.) **Chernoff Bound**: Let $X_1, \ldots, X_n$ be a sequence of i.i.d. rvs with zero-mean and moment generating function $M_X(s) := \mathbb{E}[e^{sX}]$. Show that for any $\epsilon > 0$,

$$\Pr\left(\frac{1}{n}(X_1 + \ldots + X_n) > \epsilon\right) \le \exp\left[-n \max_{s \ge 0}(\epsilon s - \log M_X(s))\right].$$

Hint: Note that the event $\{\frac{1}{n}(X_1 + \ldots + X_n) > \epsilon\}$ occurs if and only if $\{\exp(s(X_1 + \ldots + X_n)) > \exp(ns\epsilon)\}$ occurs for any fixed $s \ge 0$. Now apply Markov's inequality.

Solution: Using the hint, we know that

$$\Pr\left(\frac{1}{n}(X_1 + \ldots + X_n) > \epsilon\right) = \Pr\left(\exp(s(X_1 + \ldots + X_n)) > \exp(ns\epsilon)\right) \tag{17}$$

$$\le \frac{\mathbb{E}[\exp(s(X_1 + \ldots + X_n))]}{\exp(n\epsilon s)} \tag{18}$$

where the inequality is from Markov's inequality. Since the X_i's are independent,

$$\Pr\left(\frac{1}{n}(X_1 + \ldots + X_n) > \epsilon\right) \le \frac{\prod_{i=1}^n \mathbb{E}[\exp(sX_i)]}{\exp(n\epsilon s)} = \exp[-n(\epsilon s - \log M_X(s))]$$

where the final equality is due to the fact that the X_i are identically distributed. Since $s \ge 0$ is arbitrary, we can minimize over this parameter.

EE5139/EE6139
Solution 2

Semester I, AY2021/2022
Marco Tomamichel

Exercise 2.1 Upper bound on entropy [EE5139]

In the lecture notes we show that $H(X) \leq \log|\mathcal{X}|$ for binary random variables. Show this statement for general discrete random variables on any (finite) alphabet $\mathcal{X}$.
Solution: Let us define $f(t) = -t \log t$, which is strictly concave in t.

$$\begin{aligned} H(X) &= -\sum_{x \in \mathcal{X}} p(x) \log p(x) \\ &= \sum_{x \in \mathcal{X}} f(p(x)) \\ &= |\mathcal{X}| \sum_{x \in \mathcal{X}} \frac{1}{|\mathcal{X}|} f(p(x)) \\ &\leq |\mathcal{X}| f\left(\sum_{x \in \mathcal{X}} \frac{1}{|\mathcal{X}|} p(x) \right) \\ &= |\mathcal{X}| f\left(\frac{1}{|\mathcal{X}|} \right) \\ &= \log|\mathcal{X}|, \end{aligned}$$

where the inequality follows from the concavity of f.

Exercise 2.2 Relative entropy as a parent quantity [all]

Let X and Y be random variables on alphabets $\mathcal{X}$ and $\mathcal{Y}$ with joint pmf P_{XY}. Moreover, let U be a uniform random variable on $\mathcal{X}$. Show the following relations:

a.) $H(X) = \log|\mathcal{X}| - D(P_X \| U_X)$.

Solution:

$$H(X) = -\sum_{x \in \mathcal{X}} P_X(x) \log P_X(x) = -\sum_{x \in \mathcal{X}} P_X(x) \left(\log \frac{P_X(x)}{1/|\mathcal{X}|} + \log \frac{1}{|\mathcal{X}|} \right) = \log|\mathcal{X}| - D(P_X \| U_X).$$

b.) $H(X|Y) = \log|\mathcal{X}| - D(P_{XY} \| U_X \times P_Y)$.

Solution:

$$\begin{aligned} H(X|Y) &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{XY}(x,y) \log P_{X|Y}(x|y) \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{XY}(x,y) \left(\log \frac{P_{XY}(x,y)}{P_Y(y)/|\mathcal{X}|} + \log \frac{1}{|\mathcal{X}|} \right) \\ &= \log|\mathcal{X}| - D(P_{XY} \| U_X \times P_Y). \end{aligned}$$

c.) $I(X:Y) = D(P_{XY} \| P_X \times P_Y)$.

Solution:

$$
\begin{aligned}
I(X:Y) = H(Y) - H(Y|X) &= \sum_{y\in\mathcal{Y}} P_Y(y)\log\frac{1}{P_Y(y)} - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(x,y)\log\frac{P_X(x)}{P_{XY}(x,y)} \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(x,y)\log\frac{1}{P_Y(y)} - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(x,y)\log\frac{P_X(x)}{P_{XY}(x,y)} \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(x,y)\log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \\
&= D(P_{XY}\|P_X\times P_Y).
\end{aligned}
$$

Exercise 2.3 Example correlations [EE5139]

For each item, find an example of random variables X, Y and Z (you can restrict the alphabet size to at most 2 bits) such that the desired relations holds:

a.) $H(X|YZ) = 0$ but $H(X|Y) = H(X|Z) = 1$.

Solution: We may choose binary random variables satisfying $X = Y \oplus Z$ with Y and Z uniform and independent.

b.) $I(X:Y|Z) = 1$ but $I(X:Y) = 0$.

Solution: We may choose X and Y uniform and independent with $Z = X \oplus Y$.

c.) $I(X:Y) = 1$ but $I(X:Y|Z) = 0$.

Solution: We may choose binary $X = Y = Z$ uniform.

d.) $I(X:Y) = I(X:Z) = 1$ but $I(Y:Z) = 0$.

Solution: We may choose $X = (Y, Z)$ with Y and Z independent and uniform.

Exercise 2.4 Information spectrum [EE6139]

Given a random variable X governed by the pmf P or an alternative pmf Q, the log-likelihood ratio is defined as the random variable $Z(X) = \log\frac{P(X)}{Q(X)}$.

a.) We have seen that the expectation value of Z (under P) is the relative entropy

$$
\mathbb{E}[Z] = \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)} = D(P\|Q)\,. \tag{1}
$$

Give an expression for $\mathrm{Var}[Z]$ (under P). This quantity is called the relative entropy variance and denoted by $V(P\|Q)$.

Solution:

$$
V(P\|Q) = \mathrm{Var}[Z] = \sum_{x\in\mathcal{X}} P(x)\left(\log\frac{P(x)}{Q(x)} - D(P\|Q)\right)^2.
$$

Consider now a sequence of i.i.d. random variables $X^n = (X_1, X_2, \ldots, X_n)$ on $\mathcal{X}^n$ where each X_i is governed by the pmf P or an alternative pmf Q. We are interested in pmf of the log-likelihood ratio $Z(X^n)$.

b.) Show that $Z(X^n) = \sum_{i=1}^n Z(X_i)$. What is $\mathbb{E}[Z^n]$ and $\mathrm{Var}[Z^n]$?

Solution:

$$Z(X^n) = \log \frac{P(X^n)}{Q(X^n)} = \log \frac{\prod_{i=1}^n P(X_i)}{\prod_{i=1}^n Q(X_i)} = \sum_{i=1}^n \log \frac{P(X)}{Q(X)} = \sum_{i=1}^n Z(X_i).$$

$$\mathbb{E}[Z^n] = \sum_{i=1}^n \mathbb{E}[Z(X_i)] = nD(P\|Q).$$

We use independence to write

$$\mathrm{Var}[Z^n] = \sum_{i=1}^n \mathrm{Var}[Z(X_i)] = nV(P\|Q).$$

c.) Let us now consider the quantity $\Pr[Z(X^n) \leq nR]$ in the limit of large n for different values of R. Show that

$$\lim_{n\to\infty} \Pr[Z(X^n) \leq nR] = \begin{cases} 0 & \text{if} \quad R < D(P\|Q) \\ 1 & \text{if} \quad R > D(P\|Q) \end{cases} . \tag{2}$$

Hint: Argue using the weak law of large numbers.

Solution: Using the weak law of large numbers, we have for any positive number $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr\left[\left|\frac{\sum_{i=1}^n Z(X_i)}{n} - D(P\|Q)\right| > \epsilon\right] = 0,$$

and

$$\lim_{n\to\infty} \Pr\left[\left|\frac{\sum_{i=1}^n Z(X_i)}{n} - D(P\|Q)\right| \leq \epsilon\right] = 1.$$

Then, in particular,

$$\lim_{n\to\infty} \Pr[Z(X^n) \leq n(D(P\|Q) - \epsilon)] = 0,$$
$$\lim_{n\to\infty} \Pr[Z(X^n) \leq n(D(P\|Q) + \epsilon)] = 1.$$

Now, if $R < D(P\|Q)$ then there also exists an $\epsilon > 0$ such that $R \leq D(P\|Q) - \epsilon$. And similalry, if $R > D(P\|Q)$ then there exists an $\epsilon > 0$ such that $R \geq D(P\|Q) + \epsilon$. Hence, the above inequalities imply the desired result.

d.) Later on in the lecture we will encounter the quantity

$$D_s^\epsilon(P^n\|Q^n) := \sup\{k \in \mathbb{R} : \Pr[Z(X^n) \leq k] \leq \epsilon\}, \tag{3}$$

which, in words, is asking the largest k such that the tail of the distribution of Z that lies below k has cumulative probability at most ϵ. Show that $D_s^\epsilon(P^n\|Q^n) = nD(P\|Q) + o(n)$, or equivalently,

$$\lim_{n\to\infty} \frac{1}{n} D_s^\epsilon(P^n\|Q^n) = D(P\|Q) . \tag{4}$$

Hint: Verify that $\frac{1}{n} D_s^\epsilon(P^n\|Q^n) = \sup\{k \in \mathbb{R} : \Pr[\frac{1}{n} Z(X^n) \leq k] \leq \epsilon\}$.

Solution: From the previous item and the definition of the limit we know that for any $R < D(P\|Q)$ and $\epsilon > 0$, there exists an $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ we have

$$\Pr[Z(X^n) \leq nR] \leq \epsilon\,. \tag{5}$$

This implies that in the definition of (3) we are allowed to choose any $k \leq nR$ and thus by taking the supremum we get

$$D_s^\epsilon(P^n\|Q^n) \geq \sup\{k \in \mathbb{R} : k < nR\} = nR\,. \tag{6}$$

Taking the limit $n \to \infty$ yields the desired lower bound, for all $R < D(P\|Q)$,

$$\lim_{n\to\infty} \frac{1}{n} D_s^\epsilon(P^n\|Q^n) \geq R, \tag{7}$$

And hence $\lim_{n\to\infty} \frac{1}{n} D_s^\epsilon(P^n\|Q^n) \geq D(P\|Q)$ since this holds for all $R < D(P\|Q)$.

We can argue similarly in the opposite direction. For any $R > D(P\|Q)$ and $\mu > 0$, there exists an $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ we have

$$\Pr[Z(X^n) \leq nR] \geq 1 - \mu\,. \tag{8}$$

If we choose μ small enough so that $1 - \mu > \epsilon$ then this implies that any $k \geq nR$ violates the constraint on the probability in the definition of $D_s^\epsilon(P^n\|Q^n)$, and thus we must have

$$D_s^\epsilon(P^n\|Q^n) \leq nR \tag{9}$$

Taking again the limit we find that

$$\lim_{n\to\infty} \frac{1}{n} D_s^\epsilon(P^n\|Q^n) \leq R, \tag{10}$$

Since this holds for any $R > D(P\|Q)$ we deduce that $\lim_{n\to\infty} \frac{1}{n} D_s^\epsilon(P^n\|Q^n) \leq D(P\|Q)$.

e.) Optional: Show that in the next order in n, we have

$$D_s^\epsilon(P^n\|Q^n) = nD(P\|Q) + \sqrt{nV(P\|Q)}\,\Phi^{-1}(\epsilon) + o\left(\sqrt{n}\right) \tag{11}$$

Can we even say something more about the $o\left(\sqrt{n}\right)$ term?

Hint: The statement can be shown using the central limit theorem. A quantitative version of the central limit theorem is the Berry-Esseen theorem. Look it up to make even stronger statements about the remainder term.

Solution: We give here actually an even stronger bound, using the Berry–Eseen theorem, which tells us how quickly the renormalised distribution approaches the Gaussian distribution in the central limit theorem. To make this precise we define

$$T(P\|Q) := \sum_{x\in\mathcal{X}} P(x) \left| \log \frac{P(x)}{Q(x)} - D(P\|Q) \right|^3 .$$

By the Berry-Esseen theorem, we have

$$\left| \Pr[Z(X^n) \leq k] - \Phi\!\left(\frac{\sqrt{n}(k/n - D(P\|Q))}{\sqrt{V(P\|Q)}} \right) \right| \leq \frac{6T(P\|Q)}{\sqrt{nV^3(P\|Q)}}.$$

By letting

$$\Phi\left(\frac{\sqrt{n}(k/n - D(P\|Q))}{\sqrt{V(P\|Q)}}\right) + \frac{6T(P\|Q)}{\sqrt{nV^3(P\|Q)}} \geq \epsilon$$

and

$$\Phi\left(\frac{\sqrt{n}(k/n - D(P\|Q))}{\sqrt{V(P\|Q)}}\right) - \frac{6T(P\|Q)}{\sqrt{nV^3(P\|Q)}} \leq \epsilon,$$

we can constrain k as follows:

$$\begin{aligned} &nD(P\|Q) + \sqrt{nV(P\|Q)}\Phi^{-1}\left(\epsilon - \frac{6T(P\|Q)}{\sqrt{nV^3(P\|Q)}}\right) \\ &\leq k \\ &\leq nD(P\|Q) + \sqrt{nV(P\|Q)}\Phi^{-1}\left(\epsilon + \frac{6T(P\|Q)}{\sqrt{nV^3(P\|Q)}}\right). \end{aligned}$$

This implies the two bounds

$$D_s^\epsilon(P^n\|Q^n) \geq nD(P\|Q) + \sqrt{nV(P\|Q)}\Phi^{-1}\left(\epsilon - \frac{6T(P\|Q)}{\sqrt{nV^3(P\|Q)}}\right) \tag{12}$$

$$D_s^\epsilon(P^n\|Q^n) \leq nD(P\|Q) + \sqrt{nV(P\|Q)}\Phi^{-1}\left(\epsilon + \frac{6T(P\|Q)}{\sqrt{nV^3(P\|Q)}}\right) \tag{13}$$

Equivalently,

$$\lim_{n\to\infty} \frac{1}{n} D_s^\epsilon(P^n\|Q^n) = D(P\|Q)\,. \tag{14}$$

If $V(P\|Q) > 0$ and $T(P\|Q) < \infty$, the term $\frac{6T(P\|Q)}{\sqrt{nV^3(P\|Q)}}$ is equal to $c/\sqrt{n}$ for some finite $c > 0$. By Taylor expansions,

$$\Phi^{-1}\left(\epsilon \pm \frac{c}{\sqrt{n}}\right) = \Phi^{-1}(\epsilon) + O\left(\frac{1}{\sqrt{n}}\right).$$

By plugging in the Taylor expansion, we can get the result.

Exercise 2.5 Independence and mutual information [all]

Consider two sequences of random variables $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$. Show that if $X_1, \ldots, X_n$ are mutually independent, then

$$I(X_1, \ldots, X_n : Y_1, \ldots, Y_n) \geq \sum_{i=1}^n I(X_i : Y_i)$$

while if given Y_i the random variable X_i is conditionally independent of all the remaining random variables for all $i = 1, \ldots, n$, then

$$I(X_1, \ldots, X_n : Y_1, \ldots, Y_n) \leq \sum_{i=1}^n I(X_i : Y_i)$$

Solution: For the first claim, consider

$$\begin{aligned}
I(X_1,\dots,X_n : Y_1,\dots,Y_n) &= \sum_{i=1}^{n} I(X_i : Y_1,\dots,Y_n|X_1,\dots,X_{i-1}) \\
&= \sum_{i=1}^{n} I(X_i : Y_1,\dots,Y_n, X_1,\dots,X_{i-1}) - I(X_i : X_1,\dots,X_{i-1}) \\
&= \sum_{i=1}^{n} I(X_i : Y_1,\dots,Y_n, X_1,\dots,X_{i-1}) \\
&\geq \sum_{i=1}^{n} I(X_i : Y_i)
\end{aligned}$$

where the third equality is by independence.
For the second claim, consider

$$\begin{aligned}
I(X_1,\dots,X_n : Y_1,\dots,Y_n) &= H(X_1,\dots,X_n) - H(X_1,\dots,X_n|Y_1,\dots,Y_n) \\
&= H(X_1,\dots,X_n) - \sum_{i=1}^{n} H(X_i|Y_1,\dots,Y_n, X_1,\dots,X_{i-1}) \\
&= H(X_1,\dots,X_n) - \sum_{i=1}^{n} H(X_i|Y_i) \\
&\leq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|Y_i) \\
&= \sum_{i=1}^{n} I(X_i : Y_i)
\end{aligned}$$

where the third equality is by the fact that given Y_i, X_i is conditionally independent of all other random variables for $i = 1,\dots,n$ so

$$H(X_i|Y_1,\dots,Y_n, X_1,\dots,X_{i-1}) = H(X_i|Y_i).$$

National University of Singapore

EE5139/EE6139
Solution 3

Semester I, AY2021/2022
Marco Tomamichel

Exercise 3.1 Huffman code [EE5139]

Consider a source that outputs independent letters according to the frequency that they appear in the English language (use the frequencies listed in Table 1).

a.) Calculate the entropy of this source.

Solution: Let $\mathcal{X}$ denote the alphabet and $X \in \mathcal{X}$. Then the entropy of this source is given by

$$H(X) = \sum_{x\in\mathcal{X}} p(x) \log \frac{1}{p(x)} \approx 4.19.$$

b.) Construct a Huffman code for this source. You may either construct the Huffman code manually according to the algorithm discussed in the lecture, or write a compute program that does this for you.

Solution: One possible Huffman code: (See the Matlab code “huffman.m”)

a	0111	b	001000	c	10100	d	01101	e	110	f	10101
g	001010	h	1111	i	0011	j	111000010	k	111001	l	01100
m	10110	n	0001	o	0100	p	001001	q	111000011	r	0101
s	0000	t	100	u	11101	v	1110001	w	10111	x	111000000
y	001011	z	111000001								

Listing 1: huffman.m

```
function [code,compression,avelen]=huffman5(p)
%HUFFMAN5
%HUFFMAN CODER FOR V5
% Format [CODE,COMPRESSION,AVELEN]=HUFFMAN5(P)
%
% P is the probability (or number of occurences) of each alphabet symbol
% CODE gives the huffman code in a string format of ones and zeros
% COMPRESSION gives the compression rate
% AVELEN gives the expected length of the code
%
% Huffman5 works by first building up a binary tree (eg p =[ .5 .2 .15 .15])
%
%      a_1     a_4
%    1/      1/
%    /       /
%  b3      b1
```

a	8.4%	b	1.5%	c	2.2%	d	4.2%	e	11.0%	f	2.2%	g	2.0%
h	6.0%	i	7.4%	j	0.1%	k	1.3%	l	4.0%	m	2.4%	n	6.7%
o	7.4%	p	1.9%	q	0.1%	r	7.5%	s	6.2%	t	9.2%	u	2.7%
v	0.9%	w	2.5%	x	0.1%	y	2.0%	z	0.1%				

Table 1: Statistical distribution of letters in the English language. Source: `https://en.wikipedia.org/wiki/Letter_frequency`, but normalized so that they add up to 100%.

```
%    \    /  \
%    0\ 1/   0\
%      b2      a_3
%        \
%        0\
%          a_2
%
% Such that the tree always terminates at an alphabet symbol and the
% symbols furthest away from the root have the lowest probability.
% The branches at each level are  labeled 0 and 1.
% For this example CODE would be
%     1
%     00
%     010
%     011
% and the compression rate 1.1111
% Sean Danaher University of Northumbria at Newcastle UK 98/6/4

p=p(:)/sum(p);    %normalises probabilities
c=huff5(p);
code=char(getcodes(c,length(p)));
compression=ceil(log(length(p))/log(2))/ (sum(code' ~= ' ')*p);
avelen=sum(sum(code' ~= ' ').*p);
%----------------------------------------------------------------
function c=huff5(p)
% HUFF5 Creates Huffman Tree
% Simulates a tree structure using a nested cell structure
% P is a vector with the probability (number of occurences)
%   of each alphabet symbol
% C is the Huffman tree. Note Matlab 5 version
% Sean Danaher 98/6/4        University of Northumbria, Newcastle UK

c=cell(length(p),1);                      % Generate cell structure
for i=1:length(p)                                 % fill cell structure with 1,2,3...n
   c{i}=i;                                                % (n=number of symbols in alphabet)
end
while size(c)-2                                   % Repeat till only two branches
        [p,i]=sort(p);                                  % Sort to ascending probabilities
        c=c(i);                                                 % Reorder tree.
        c{2}={c{1},c{2}};c(1)=[];       % join branch 1 to 2 and prune 1
        p(2)=p(1)+p(2);p(1)=[];         % Merge Probabilities
end
%----------------------------------------------------------------
function y= getcodes(a,n)
% Y=GETCODES(A,N)
% Pulls out Huffman Codes for V5
% a is the nested cell structure created by huffcode5
% n is the number of symbols
% Sean Danaher 98/6/4   University of Northumbria, Newcastle UK
global y
y=cell(n,1);
getcodes2(a,[])
%---------------------------------------------------------------------
function getcodes2(a,dum)
% GETCODES(A,DUM)
%getcodes2
% called by getcodes
% uses Recursion to pull out codes
% Sean Danaher 98/6/4   University of Northumbria, Newcastle UK

global y
if isa(a,'cell')
         getcodes2(a{1},[dum 0]);
         getcodes2(a{2},[dum 1]);
else
   y{a}=setstr(48+dum);
end
```

c.) Compute the expected length of the codeword for this code. How does this compare to the entropy computed in a)?

Solution: Expected length: $4.22 \geq H(X) = 4.19$.

Exercise 3.2 Huffman code [all]

Which of the following sets of codewords can never be valid Huffman codes (for any assignment of probabilities)? Argue why.

a.) $\{0, 10, 11\}$,

Solution: Can.

b.) $\{00, 01, 10, 11\}$,

Solution: Can.

c.) $\{00, 01, 10, 110\}$,

Solution: Cannot. 110 can be shortened to 11.

d.) $\{01, 10\}$,

Solution: Cannot. The codewords can be shortened to $\{0, 1\}$.

e.) $\{1, 01, 10\}$.

Solution: Cannot. It is not a prefix code.

Exercise 3.3 A code that allows a prefix [EE5139]

As we have seen in the lecture, we like codes to be prefix-free as otherwise we do not know when a codeword ends and decoding might no longer be unique. One simple (but not necessarily very efficient) way to overcome this problem is to simply add a special symbol that indicates the end of a codeword. In this exercise we will thus consider codewords that are comprised of “0” and “1” and always end with a special space character “_”.

a.) Construct a code for this source in the following way: the two most frequent laters are assigned codewords of length $1+1$, the next four most frequent letters codewords of lenght $2+1$, etc.

Solution: (See the Matlab code “ex3_2.m”)

a	b	c	d	e	f	g	h	i	j	k
00_	0101_	0000_	011_	0_	0001_	0010_	010_	10_	1000_	0110_
l	m	n	o	p	q	r	s	t	u	v
100_	111_	000_	11_	0100_	1001_	01_	001_	1_	101_	0111_
w	x	y	z							
110_	1010_	0011_	1011_							

Listing 2: ex3_2.m

```
%%%Exercise 3.2 prefix code%%%

pp=[8.4,1.5,2.2,4.2,11,2.2,2.0,6,7.4,0.1,1.3,4,2.4,6.7,7.4,1.9,0.1,7.5,...
    6.2,9.2,2.7,0.9,2.5,0.1,2,0.1];
A='a':'z';
AA=mat2cell(A,1,ones(26,1));
AAA=[];

%sort probabilities and record the index
[pp0,pindex]=sort(pp,'descend');

%generate codewords
code={};
for i=1:4
    code(2^i-1:(2^i-1)*2)=cellstr(dec2bin(0:2^i-1));
end
code=code(1:26);

%combine symbols and codewords
for i=1:26
    AA(2,pindex(i))=cellstr([cell2mat(code(i)),'_']);
end

%expected length
Sum=0;
for i=1:26
    Sum=Sum+pp(i)*length(cell2mat(AA(2,i)));
end
Sum=Sum/100;

%expected time
time=0;
for i=1:26
   t0=2*length(find(cell2mat(AA(2,i))=='0'));
   t1=4*length(find(cell2mat(AA(2,i))=='1'));
   time=time+pp(i)/100*(t0+t1);
end
time=time+3
```

b.) Compute the expected length of the codewords.

Solution: Expected length: 3.433

c.) The code you arrived at is very similar to a very famous code that has been in use since the 1830s. Can you figure out which code that is? Compare the expected length of codewords of your code to that of this historical code.

Solution: Morse code. The expected length of Morse code is 3.558, which is longer than that of our code.

Hint: Replace “0” by “·” and “1” by “−” in your codewords.

Exercise 3.4 Kraft–McMillan inequality for uniquely decodable codes [all]

Assume a uniquely decodable code has codeword lengths $l_1, \ldots, l_M$. Our goal is to derive Kraft's inequality for uniquely decodable codes:

$$\sum_{j=1}^{M} 2^{-l_j} \leq 1 .$$

a.) Prove the following identity (this is easy):

$$\left(\sum_{j=1}^{M} 2^{-l_j} \right)^n = \sum_{j_1=1}^{M} \sum_{j_2=1}^{M} \cdots \sum_{j_n=1}^{M} 2^{-(l_{j_1}+l_{j_2}+\ldots+l_{j_n})} .$$

Solution: Perform the multiplication...

b.) Let A_l be the number of concatenations of n codewords that have overall length $l = l_{j_1}+l_{j_2}+\ldots+l_{j_n}$ and let $l_{\max} = \max\{l_1, l_2, \ldots, l_M\}$ be the maximum length of a codeword. Show that

$$\left(\sum_{j=1}^{M} 2^{-l_j} \right)^n = \sum_{l=n}^{nl_{\max}} A_l \, 2^{-l} .$$

Solution: The smallest value this exponent can take is n, which would happen if all code words had the length 1. The largest value the exponent can take is $nl_{\max}$ where $l_{\max}$ is the maximal codeword length. The summation can then be written as above.

c.) Using unique decodability, show that $A_i \leq 2^i$ and hence

$$\left(\sum_{j=1}^{M} 2^{-l_j} \right)^n \leq nl_{\max} .$$

Use this to derive the desired inequality.

Solution: The number of possible binary sequences of length i is 2^i. Since the code is uniquely decodable, we must have $A_i \leq 2^i$ in order to be able to decode. Plugging this into the above bound yields

$$\left(\sum_{j=1}^{M} 2^{-l_j} \right)^n \leq \sum_{i=n}^{nl_{\max}} 2^i 2^{-i} = nl_{\max} - n + 1 \leq nl_{\max} .$$

We then have

$$\sum_{j=1}^{M} 2^{-l_j} \leq \left[n(l_{\max} - 1) \right]^{1/n} = \exp\left[\frac{1}{n} \log(n(l_{\max} - 1)) \right]$$

The exponent goes to zero as $n \to \infty$ and hence, $\sum_{j=1}^{M} 2^{-l_j} \leq 1$, Kraft's inequality for uniquely decodable codes.

Exercise 3.5 Shannon code [all]

Let X be a random variable distributed on $\{0, 1, 2, \ldots, d-1\}$ $(d>1)$, and let P_X be its probability mass function. Without loss of generality, we assume $P_X(0) \geq P_X(1) \geq \cdots \geq P_X(d-1) > 0$. A Shannon code for this source is constructed as follows. For each $x \in \mathcal{X}$, we consider the binary representation of the real number $\sum_{x'<x} P_X(x')$, and use the first $\lceil \log_2 \frac{1}{P_X(x)} \rceil$ bits in its fractional part to represent x. For example, the binary representation of the real number $\frac{1}{3}$ is '0.01010101...', and the first 2 bits in its fractional part are '01'.

a.) Show that the above code is unique decodable.

Solution: Denote the codeword representing x as $C(x)$ (namely, for this specific random variable X, Shannon code maps each x to $C(x)$). It suffices to show the above code to be a prefix code. Suppose the opposite, i.e., one of $C(x)$ and $C(\tilde{x})$ is a prefix to the other for some $x < \tilde{x}$. Note that $P_X(x) \geq P_X(\tilde{x})$. Thus, $\lceil \log_2 \frac{1}{P_X(x)} \rceil \leq \lceil \log_2 \frac{1}{P_X(\tilde{x})} \rceil$; namely, $C(x)$ is shorter and must be a prefix to $C(\tilde{x})$. On the other hand, denoting $\ell = \lceil \log_2 \frac{1}{P_X(x)} \rceil$ the length of $C(x)$, we have

$$\left| \sum_{x'<x} P_X(x') - \sum_{x'<\tilde{x}} P_X(x') \right| \geq P_X(x) \geq 2^{-\ell}.$$

However, this implies a contradiction since the first ℓ bits of these two numbers are same. Hence, the above code must be a prefix code.

b.) Produce a Shannon code table for the following source

x	0	1	2	3
$P_X(x)$	1/2	1/6	1/6	1/6

and compute the expected code length.

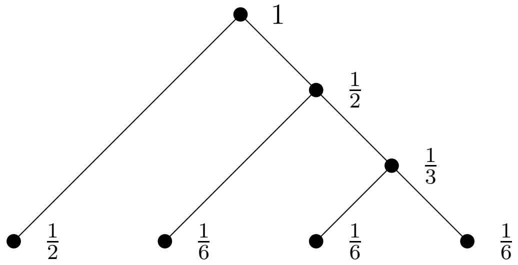
Solution:

x	$\sum_{x'<x} P_X(x)$	LHS in binary	$\lceil \log_2 \frac{1}{P_X(x)} \rceil$	codeword
0	0	0.00000000	1	0
1	1/2	0.10000000	3	100
2	2/3	0.10101010...	3	101
3	5/6	0.11010101...	3	110

The expected code length is $\frac{1}{2} \cdot 1 + 3 \cdot \frac{1}{6} \cdot 3 = 2$.

c.) Show that the expected length of the Huffman code of the above source is shorter.

Solution: The Huffman tree constructed for this case looks like the following:



The expected code length is $\frac{1}{2} \cdot 1 + \frac{1}{6} \cdot 2 + 2 \cdot \frac{1}{6} \cdot 3 = 1\frac{5}{6} < 2$.

Exercise 3.6 Huffman code algorithm [EE6139]

(from 2013/2014 final exam)

Consider a discrete memoryless source X with alphabet $\{1, 2, \ldots, M\}$. Suppose that the symbol probabilities are ordered and satisfy $p_1 > p_2 > \ldots > p_M$ and also satisfy $p_1 < p_{M-1} + p_M$. Let $l_1, l_2, \ldots, l_M$ be the lengths of a prefix-free code of minimum expected length for such a source.

a.) **(1 point)** Is the following statement true or false? $l_1 \leq l_2 \leq \ldots \leq l_M$. Argue why.

Solution: For an optimal code, if $p_i > p_j$, then $l_i \leq l_j$. Suppose otherwise, i.e., that $l_i > l_j$, then I can create a new code with lengths $l'_i = l_j$ and $l'_j = l_i$ for symbols i and j respectively. It is then easy to see that the expected cost will decrease. Indeed, the contribution to the expected cost from symbols i and j in the old code is $c_{ij}^{\text{old}} = p_i l_i + p_j l_j$. The contribution to the expected cost from symbols i and j in the new code is $c_{ij}^{\text{new}} = p_i l_j + p_j l_i$. Because $p_i > p_j$ and $l_i > l_j$, we have that $c_{ij}^{\text{old}} > c_{ij}^{\text{new}}$ which is a contradiction and so the initial assumption that $l_i > l_j$ must be false. Hence, $l_i \leq l_j$. This argument can be repeated for all symbols.

b.) **(2 points)** Show that if the Huffman algorithm is used to generate the above code, then $l_M \leq l_1 + 1$.

Solution: For the Huffman algorithm, we will first merge symbols $M-1$ and M to form a symbol with probability $p_{M-1} + p_M$. The codeword for this new symbol is $l_M - 1$. Since $p_{M-1} + p_M > p_1$, by part (a), we must have that $l_1 \geq l_M - 1$ which is the desired result.

c.) **(2 points)** Show that $l_M \leq l_1 + 1$ for *any* (not necessarily Huffman generated) prefix-free code of minimum expected length. **Hint:** A minimum-expected-length code must be full.

Solution: A minimum-expected-length code must be full, and thus the codeword for letter M must have a sibling, say letter j. Since $p_j \geq p_{M-1}$, we have $p_j + p_M > p_1$ because of the condition $p_1 < p_{M-1} + p_M$. Let l' be the length of the intermediate node that is parent to j and M. Now $l' \leq l_1$ since otherwise the codeword for 1 could be interchanged with this intermediate node for a reduction in $\bar{L}$. Again $l_M - 1 \leq l_1$.

d.) **(2 points)** Suppose $M = 2^k$ for some integer k. Show that all codewords must have the same length. **Hint:** What does the Kraft inequality look like for an optimal code? Consider the three cases $l_1 = k$, $l_1 < k$ and $l_1 > k$.

Solution: An optimal prefix-free code must be full. For full codes, Kraft must be satisfied with equality, i.e., $\sum_{j=1}^{2^k} 2^{-l_j} = 1$. First assume that $l_1 = k$. Then all codewords have length k or $k+1$, but as mentioned above, the Kraft's inequality can be satisfied with equality (i.e., the code can be full) only if all codewords have length k. If $l_1 > k$, then $l_j > k$ for all j and the Kraft inequality can not be satisfied with equality. Finally, if $l_1 < k$ with $l_j \leq k$, then the Kraft inequality cannot be met. Thus all codewords have length k.

Exercise 4.1 Infinite alphabet optimal code [all]

Let X be an i.i.d. random variable with an infinite alphabet, $\mathcal{X} = \{1, 2, 3, \ldots\}$. In addition, let $P(X = i) = 2^{-i}$.

a.) What is the entropy of X?

Solution: By direct calculation,

$$H(X) = \sum_{i=1}^{\infty} -2^{-i} \log(2^{-i}) = \sum_{i=1}^{n} i2^{-i} = 2.$$

This is because

$$\sum_{i=1}^{\infty} ix^{i-1} = \frac{1}{(1-x)^2}$$

for all $|x| < 1$. This can be shown by differentiation of the identity $\sum_{i=1}^{\infty} x^i = \frac{1}{1-x}$.

b.) Find an optimal variable length code, and show that it is indeed optimal.

Solution: Take the code lengths to be $-\log(2^{-1}), -\log(2^{-2}), -\log(2^{-3}), \ldots$. Codewords can be

$$\begin{aligned} C(1) &= 0 \\ C(2) &= 10 \\ C(3) &= 110 \\ &\vdots \end{aligned}$$

Exercise 4.2 Codeword lengths for Huffmann codes [all]

(from 2019/2020 quiz)

Consider a random variable X which takes on four possible values with probabilities $(\frac{1}{3}, \frac{1}{3}, \frac{1}{4}, \frac{1}{12})$.

a.) (5 points) Construct a Huffman code for this source.

Solution: Applying the Huffman algorithm gives us the table above, which gives codeword lengths (1, 2, 3, 3) or (2,2,2,2) for the different codewords.

Code1	Code2	Symbol	Probability
0	00	1	1/3
11	01	2	1/3
101	10	3	1/4
100	11	4	1/12

b.) (5 points) Show that there exist two different sets of optimal lengths for the codewords, namely, show that codeword length assignments $(1, 2, 3, 3)$ and $(2, 2, 2, 2)$ are both optimal.

Solution: Both set of lengths $(1, 2, 3, 3)$ and $(2, 2, 2, 2)$ satisfy the Kraft inequality, and they both achieve the same expected length (2 bits) for the above distribution. Therefore they are both optimal.

c.) (5 points) Are there optimal codes with codeword lengths for some symbols that exceed the Shannon code length $\lceil \log \frac{1}{p(x)} \rceil$?

Solution: The symbol with probability 1/4 has an Huffman code of length 3, which is greater than $\left\lceil \log \frac{1}{p(x)} \right\rceil$. Thus the Huffman code for a particular symbol may be longer than the Shannon code for that symbol. But on the average, the Huffman code cannot be longer than the Shannon code.

Exercise 4.3 A better Morse code [EE5139]

In the previous exercise we designed a code for English letters that had slightly lower expected length than the Morse code. Now let us look at this problem a bit more closely. For the Morse code, the codeword symbol "-" requires 4 time units to send, whereas "." only requires 2 time units. The end-of-letter symbol "_" requires 3 time units. So what we actually want to optimise is not the codeword length but the the *time* require to send it, e.g. for the codeword ".-._" this would be 11.

a.) Devise an algorithm that produces an alternative Morse code that optimises the expected time for a source producing English letters.

Solution: An idea is to list all codeword within certain transmission time limit, and assign faster signals to letters with higher frequencies. An algorithm is listed below. We denote the frequency of a letter in Table 1 by $f(\cdot)$, e.g., $f(\mathrm{a}) = 0.084$.

```
Let $\{A_1, A_2, \cdots, A_{26}\} = \{\mathrm{a}, \mathrm{b}, \ldots, \mathrm{z}\}$, and $f(A_i) \geq f(A_{i+1})$ for each $i \in \{1, 2, \ldots, 25\}$;
Initialize $t \leftarrow 2$, $j \leftarrow 0$;
do
    $\mathcal{C}_t \leftarrow \{(s_1, \ldots, s_n, \_) | s_1, \ldots, s_n \in \{., \text{-}\}, \sum_{i=1}^{n} 2 \cdot \mathbf{1}\{s_i = .\} + 4 \cdot \mathbf{1}\{s_i = \text{-}\} = t, n \in \mathbb{N}\}$;
    for $i = 1, \ldots, |\mathcal{C}_t|$ do
        if $j < 26$ then
            Assign the $i$-th element of $\mathcal{C}_t$ to $A_{i+j}$;
            $j \leftarrow i$;
        end
    end
    $t \leftarrow t + 1$;
while $j < 26$;
```

A code table is listed below

a	-_	b	......_	c	.-.._	d	-.._	e	._	f	-..._	g	--._
h	.-._	i	.-_	j	..-.._	k	....-_	l	--	m	..-._	n	...._
o	-._	p	.--_	q	.-..._	r	..._	s	..-_	t	.._	u	....._
v	...-._	w	...-_	x	-...._	y	-.-_	z	--.._				

b.) Compute the expected time of transmission of the above code. Compute the expected time of the code produced in Exercise 3.3a, where we treat 0 as '.' and 1 as '-'.

a	8.4%	b	1.5%	c	2.2%	d	4.2%	e	11.0%	f	2.2%	g	2.0%
h	6.0%	i	7.4%	j	0.1%	k	1.3%	l	4.0%	m	2.4%	n	6.7%
o	7.4%	p	1.9%	q	0.1%	r	7.5%	s	6.2%	t	9.2%	u	2.7%
v	0.9%	w	2.5%	x	0.1%	y	2.0%	z	0.1%				

Table 1: Statistical distribution of letters in the English language. Source: `https://en.wikipedia.org/wiki/Letter_frequency`, but normalized so that they add up to 100%.

Solution: Expected time of transmission of the above code: 9.712.
Expected time of transmission of the code in Exercise 3.3a: 9.872.
(Note that we included ‘_’ into the transmission time.)

c.) Can you show that it is optimal?

Solution: Let vector $\mathbf{f} = (f(A_1), f(A_2), \dots, f(A_{26}), 0, 0, \dots)$. Let set $\mathfrak{A}$ denote the set of all finite sequences made up of ‘.’ and ‘-’, ending with ‘_’, namely

$$\mathfrak{A} = \{(s_1, s_2, \dots, s_n, \_) : s_i \in \{.,\text{-}\} \text{ for each } i, n \in \mathbb{N}.\}$$

For each $\mathbf{s} \in \mathfrak{A}$, let $t(\mathbf{s})$ denote the transmission time of $\mathbf{s}$. We order the set $\mathfrak{A} = \{\mathbf{s}_i : i \in \mathbb{N}\}$ s.t. $t(\mathbf{s}_i) \leq t(\mathbf{s}_{i+1})$ for all $i \in \mathbb{N}$. Define the vector $\mathbf{t} = (t(\mathbf{s}_1), t(\mathbf{s}_2), \dots)$.

For any code described in this exercise, its expected transmission time can be expressed as

$$\sum_i \mathbf{f}_i \cdot t(\mathbf{s}_{\pi_i}) = \sum_i \mathbf{f}_i \cdot \mathbf{t}_{\pi_i}$$

for some permutation π of natural numbers. Since $\mathbf{f} \geq 0$, $\mathbf{t} \geq 0$, and one is non-increasing, and the other is non-decreasing,

$$\sum_i \mathbf{f}_i \cdot \mathbf{t}_{\pi_i} \geq \sum_i \mathbf{f}_i \cdot \mathbf{t}_i.$$

Since the right-hand side is the expected transmission time of the code we constructed in a), we have shown the optimality.

d.) Can you come up with a prefix code that attempts to minimise the expected transmission time? This code does not need the end-of-letter symbol. Your algorithm can be heuristic (you do not need to prove optimality). Compute the expected transmission time and compare it to the code from a.

Solution: We present a heuristic algorithm here. (The idea of the algorithm is to view a prefix code as a process to divide the alphabet recursively until each subset has only one element. Intuitively speaking, for prefix codes with two balanced symbols, it makes sense to have each division to be as balanced as possible so that each symbol reduces maximum amount of uncertainty. In our case, one of the symbol costs twice as that of the other symbol. Thus, it makes sense to design the division so that choosing the shorter symbol twice should reduce the same amount of uncertainty as choosing the longer symbol.)

To solve this problem exactly, we refer to Golin and Rote’s 1998 paper “A dynamic programming algorithm for constructingoptimal prefix-free code with Unequal letter cost”, in which they cleverly converted the problem of finding the optimal coding tree into finding the shortest path in a “signature tree”. For your reference, we also include a MatLab implementation of their algorithm for our problem.

Listing 1: Unbalanced_prefix_code.m

```
A = 1:26;
p = [8.4, 1.5, 2.2, 4.2, 11.0, 2.2, 2.0, 6.0, 7.4, 0.1, 1.3, 4.0, 2.4, 6.7, 7.4, 1.9, 0.1, 7.5, 6.2, 9.2, 2.7, 0.9, 2.5, 0.1, 2.0, 0.1];
p = p./sum(p);

P = {A, p};
q = (sqrt(5)-1)/2;

[C, X] = unbalanced_prefix_code (P, q);

```

```matlab
C_sorted = cell(length(X),1);
for k = 1:length(X)
    C_sorted{X(k)} = C{k};
end
disp(C_sorted);

t = 0;
for x = 1:26
    t = t + p(x)*2*(length(C_sorted{x})+sum(C_sorted{x}));
end
disp(t);

function [C, X] = unbalanced_prefix_code (P, q)
    if length(P{1}) == 1
        C = {[]};
        X = P{1};
    else
        C = cell(length(P{2}),1);
        X = [];
        [Pl, Pr] = divide(P,q);
        [Cl, Xl] = unbalanced_prefix_code(Pl, q);
        [Cr, Xr] = unbalanced_prefix_code(Pr, q);
        X = [Xl, Xr];
        for k = 1:length(Xl)
            C{k} = [0,Cl{k}];
        end
        for k = 1:length(Xr)
            C{k+length(Xl)} = [1,Cr{k}];
        end
    end
end

function [Pl, Pr] = divide (P, q)
% We attempt to divide the normalized weighted set P={A,p} into Pl and Pr so that the total weight of Pl is as close to q as possible
% However, checking all subsets of A take exponential amount of time. Thus, here, we use a greedy algorithm.
    Al = [];
    pl = [];
    Ar = P{1};
    pr = P{2}/sum(P{2});
    % Al cannot be empty: Pick at least one element
    q_current = sum(pl);
    d_current = abs(q_current - q);
    j = 0;
    d = inf;
    for k = 1:length(Ar)
        if abs(q_current + pr(k) - q) < d
            d = abs(q_current + pr(k) - q);
            j = k;
        end
    end
    Al = [Al, Ar(k)];
    pl = [pl, pr(k)];
    Ar = [Ar(1:k-1), Ar(k+1:length(Ar))];
    pr = [pr(1:k-1), pr(k+1:length(Ar))];
   % Pick additional elements
    while (1)
        q_current = sum(pl);
        d_current = abs(q_current - q);
        if length(Ar) == 1
            break;
        end
        j = 0;
        d = inf;
        for k = 1:length(Ar)
            if abs(q_current + pr(k) - q) < d
                d = abs(q_current + pr(k) - q);
                j = k;
            end
        end
        if d < d_current
            Al = [Al, Ar(k)];
            pl = [pl, pr(k)];
            Ar = [Ar(1:k-1), Ar(k+1:length(Ar))];
            pr = [pr(1:k-1), pr(k+1:length(Ar))];
        else
            break;
        end
    end
    Pl = {Al, pl};
    Pr = {Ar, pr};
end
```

The code table generated by the above program looks like the below.

a	---	b	--..	c	--.-	d	-..-	e	-...	f	-.-.-	g	-.-..
h	-.--	i	..--	j	..-..-	k	..-...	l	..-.-	m	....-	n	......
o	.....-	p	...-.-	q	...-..	r	...--	s	-.-.	t	.-..	u	.--.-
v	.--...	w	.--..-	x	-...-.	y	-...--	z	.----				

The expected time of transmission of the above code is 12.802.

Listing 2: Golin_Rote.m

```
% Here, we implement the algorithm in the 1998 paper by Golin and Rote for
% the special case r=2, and n = 26
n = 26;
p = [8.4, 1.5, 2.2, 4.2, 11.0, 2.2, 2.0, ...
     6.0, 7.4, 0.1, 1.3, 4.0, 2.4, 6.7, ...
     7.4, 1.9, 0.1, 7.5, 6.2, 9.2, ...
     2.7, 0.9, 2.5, 0.1, 2.0, 0.1]/100;
 P = p;
 A = {'a', 'b', 'c', 'd', 'e', 'f', 'g', ...
     'h', 'i', 'j', 'k', 'l', 'm', 'n', ...
     'o', 'p', 'q', 'r', 's', 't', ...
     'u', 'v', 'w', 'x', 'y', 'z'};
[p, idx] =  sort(p,'descend');
A_num = idx;
% Now we implement the shortest-path-in-signature-tree algorithm
cost = zeros(n+1,n+1,n+1);
pre = cell(n+1,n+1,n+1);
update = zeros(n+1,n+1,n+1);
for k = 1:length(cost(:))
    cost(k) = Inf;
end
cost(0+1,1+1,1+1) = 0;
update(0+1,1+1,1+1) = 1;

while (sum(update(:)) > 0)
    [M, L1, L2] = ind2sub([n+1,n+1,n+1],find(update));
    M = M - 1; L1 = L1 - 1; L2 = L2 - 1;
    update = zeros(n+1,n+1,n+1);
    for k = length(M)
        m = M(k);
        l1 = L1(k);
        l2 = L2(k);
        new_cost = cost(m+1, l1+1, l2+1);
        for t = m+1:1:n
            new_cost = new_cost + p(t);
        end
        for q = 0:1:l1
            m_next = m+l1-q;
            l1_next = l2+q;
            l2_next = q;
            if m_next+l1_next+l2_next <=n
                if new_cost < cost(m_next+1, l1_next+1, l2_next+1)
                    cost(m_next+1, l1_next+1, l2_next+1) = new_cost;
                    pre{m_next+1, l1_next+1, l2_next+1} = [m, l1, l2];
                    update(m_next+1, l1_next+1, l2_next+1) = 1;
                end
            end
        end
    end
end

expected_time = cost(n+1, 1, 1)*2;

% Now list the path from (0,1,1) to (n,0,0)
path = [n,0,0];
m = n; l1 = 0; l2 = 0;
while (~isempty(pre{m+1, l1+1, l2+1}))
    PRE = pre{m+1, l1+1, l2+1};
    m = PRE(1);
    l1 = PRE(2);
    l2 = PRE(3);
    path = [m, l1, l2; path];
end
```

```
% The first row of path must be (0,1,1).

% Create the prefix code graph from the path
signature = [0,1,1];
graph = {[0], [1]}; % We only need to keep track of leaves.
D = 0;
for k = 2:size(path,1)
    D = D + 1;
    next_signature = path(k,:);
    q = next_signature(3);
    % change q-number of depth-D leaves in graph into parents
    populated = 0;
    new_graph = {};
    for j = 1:length(graph)
        leaf = graph{j};
        depth = sum(leaf+1);
        if depth == D && populated < q
            new_graph{length(new_graph)+1} = [leaf,0];
            new_graph{length(new_graph)+1} = [leaf,1];
            populated = populated + 1;
        else
            new_graph{length(new_graph)+1} = leaf;
        end
    end
    graph = new_graph;
end
% Sort the leaves in the graph (n == length(graph))
time = zeros(1,n);
for k = 1:n
    time(k) = 2*sum(graph{k}+1);
end
[time, idx] = sort(time);
code = cell(1,n);
for k = 1:n
    codeword = graph{idx(k)};
    code{A_num(k)} = codeword;
end
for k = 1:n
    disp([A{k},': ',num2str(code{k}), ' (', num2str(P(k)), ')']);
end
disp(expected_time);
```

The code table generated by the above program looks like the below.

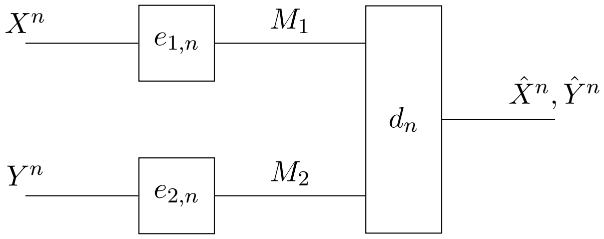
a	.--.	b	-.--	c	..-.-	d	.....-	e	.-...	f	..--.	g	.-..-
h	....	i	-..-	j	....--	k	--.-	l	....-.	m	..-...	n	--..
o	-.-.	p	-...-	q	...-.-	r	-....	s	---	t	.-.-	u	...-..
v	......-	w	...--	x	..-..-	y	.---	z	..---				

The expected time of transmission of the above code is 12.750.

Exercise 4.4 Converse for the Slepian–Wolf coding problem [all]

Let X and Y be a pair of jointly distributed random variables. (X is distributed on finite set $\mathcal{X}$, and Y is distributed on finite set $\mathcal{Y}$.) An $(n, 2^{nL_1}, 2^{nL_2})$-separately-encoded-jointly-decoded source code consists of a pair of encoders e_1, e_2, and a decoder d, where

- $e_1 : \mathcal{X}^n \to \{0,1\}^{nL_1}$,
- $e_2 : \mathcal{Y}^n \to \{0,1\}^{nL_2}$, and
- $d : \{0,1\}^{nL_1} \times \{0,1\}^{nL_2} \to \mathcal{X}^n \times \mathcal{Y}^n$.



The rate pair (R_1, R_2) is said to be achievable for DMS (X, Y) if there exists a sequence of $(n, 2^{nL_1}, 2^{nL_2})$-codes with encoders $e_{1,n}$, $e_{2,n}$ and decoder d_n such that

$$\lim_{n\to\infty} P\{(\hat{X}^n, \hat{Y}^n) \neq (X^n, Y^n)\} = 0$$

where

$$(\hat{X}^n, \hat{Y}^n) = d_n(M_1, M_2),\ M_1 = e_{1,n}(X^n),\ \text{and } M_2 = e_{2,n}(Y^n)$$

are the reconstructed source and codewords respectively.
Prove that, for any (R_1, R_2) achievable, it must hold that

$$R_1 \geq H(X|Y), \tag{1}$$
$$R_2 \geq H(Y|X), \tag{2}$$
$$R_1 + R_2 \geq H(X, Y). \tag{3}$$

Solution: By considering (M_1, M_2) as a single message, (3) follows directly from the proof in the lecture notes. We are only going to show the proof for (1) below. The proof for (2) follows symmetrically.
Suppose (R_1, R_2) is achievable. For any $\epsilon > 0$, there exists some $N \in \mathbb{N}$ such that for all $n > N$

$$P\{\hat{X}^n \neq X^n\} \leq P\{(\hat{X}^n, \hat{Y}^n) \neq (X^n, Y^n)\} < \epsilon.$$

By Fano's inequality, we have

$$I(\hat{X}^n : X^n) = nH(X) - H(\hat{X}^n|X^n) \geq nH(X) - \epsilon n \log|\mathcal{X}| - 1.$$

On the other hand, by how M_1 and M_2 are constructed, we have

$$\begin{aligned} I(\hat{X}^n : X^n) &\overset{\text{(a)}}{\leq} I(M_1, M_2 : X^n) \\ &\overset{\text{(b)}}{=} I(M_1 : X^n) + I(M_2 : X^n) \\ &\overset{\text{(c)}}{\leq} I(M_1 : X^n) + I(X^n : Y^n) \\ &\leq nR_1 + nI(X : Y) \end{aligned}$$

where (a) follows from the Markov chain $X^n - (M_1, M_2) - \hat{X}^n$, (b) follows from the Markov chain $M_1 - X^n - M_2$, and (c) follows from the Markov chain $M_2 - Y^n - X^n$. Thus,

$$nH(X) - \epsilon n \log |\mathcal{X}| - 1 \leq nR_1 + nI(X : Y),$$

which implies

$$R_1 \geq H(X|Y) - \epsilon \log |\mathcal{X}| - \frac{1}{n}.$$

Since above holds from all $\epsilon > 0$, and all $n > N_\epsilon$, one must have $R_1 \geq H(X|Y)$.

EE5139/EE6139
Solution 5

Semester I, AY2021/2022
Marco Tomamichel

Exercise 5.1 Min-entropy and Shannon entropy as Rényi entropies [EE5139]

Both the min-entropy and the Shannon entropy are limiting cases of the following family of Rényi entropies:

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \sum_x P(x)^\alpha, \qquad \alpha \in (0,1) \cup (1,+\infty)\,. \tag{1}$$

a.) To verify this, compute the limit of the above quantities for $\alpha \to \{0_+, 1, +\infty\}$. (Here, by saying $\alpha \to 0_+$, we mean α "approaching 0 from right-hand side".)

Solution:

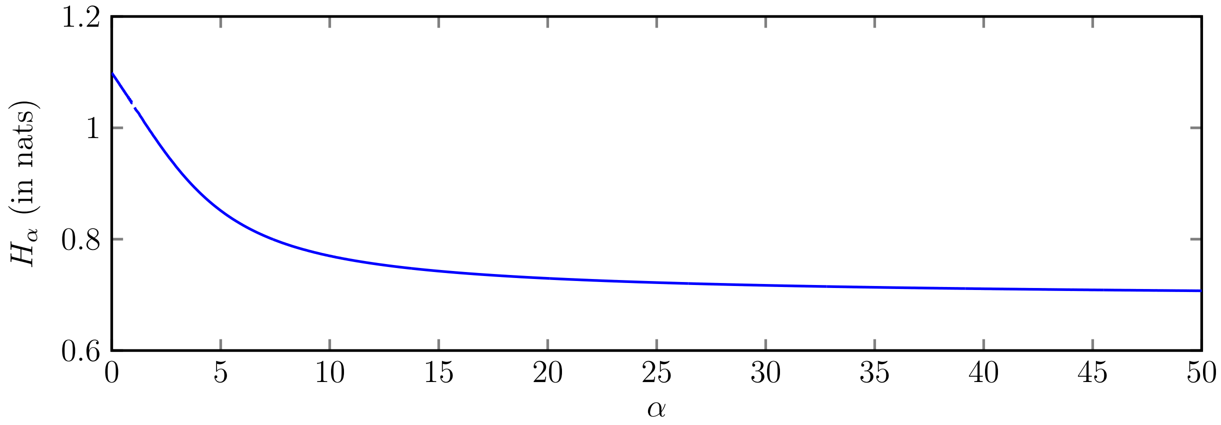
$$\begin{aligned}
\lim_{\alpha\to 0_+} H_\alpha(X) &= \log\left(\sum_x \lim_{\alpha\to 0_+} P(x)^\alpha\right) \\
&= \log|\{x : P(x) > 0\}| = H_{\max}(X). \\
\lim_{\alpha\to 1} H_\alpha(X) &= \frac{\left(\sum_x P(x)\right)^{-1} \cdot \sum_x \log P(x) \cdot P(x)^\alpha}{-1} && \blacktriangleright \text{ L'Hôpital's rule} \\
&= -\sum_x P(x) \cdot \log P(x) = H(X). \\
\lim_{\alpha\to+\infty} H_\alpha(X) &= -\lim_{\alpha\to+\infty} \log\left[\left(\sum_x P_{\max}^\alpha \cdot \left(\frac{P(x)}{P_{\max}}\right)^\alpha\right)^{\frac{1}{\alpha-1}}\right] && \blacktriangleright\ P_{\max} \triangleq \max_x P(x) \\
&= -\lim_{\alpha\to+\infty} \frac{\alpha}{\alpha-1} \cdot \log P_{\max} + \frac{1}{\alpha-1} \cdot \log\left(\sum_x \left(\frac{P(x)}{P_{\max}}\right)^\alpha\right) \\
&= \log P_{\max} = H_{\min}(X). && \blacktriangleright\ 1 \le \sum_x \left(\frac{P(x)}{P_{\max}}\right)^\alpha \le |\mathcal{X}|
\end{aligned}$$

□

b.) Plot the Rényi entropy as a function of α for the random variable X distributed as

x	0	1	2
$P(x)$	1/2	1/4	1/4

Solution: We plot the Rényi entropy of the above random variable in nats.



□

c.) Show that, for any random variable $X \in \mathcal{X}$ and any pmf $P(x)$, the Rényi entropy is monotonically non-increasing in the parameter α. Argue how this yields an alternative proof of the fact that $H_{\min}(X) \leq H(X) \leq \log|\mathcal{X}|$.

Solution: Consider the derivative of H_α w.r.t α

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} H_\alpha(X) = \frac{1}{(1-\alpha)^2} \cdot \left\{ (1-\alpha) \cdot \frac{\sum_x \log P(x) \cdot P(x)^\alpha}{\sum_x P(x)^\alpha} + \log \sum_x P(x)^\alpha \right\}.$$

Note that $\lim_{\alpha \to 1} H_\alpha$ exists from both sides. It suffices to show $\frac{\mathrm{d}}{\mathrm{d}\alpha} H_\alpha(X)$ to be non-positive for all $\alpha \in (0,1) \cup (1,\infty)$. By letting

$$f(\alpha) \triangleq (1-\alpha) \cdot \frac{\sum_x \log P(x) \cdot P(x)^\alpha}{\sum_x P(x)^\alpha} + \log \sum_x P(x)^\alpha$$

it suffices to show f to be non-positive for $\alpha > 0$.

However, noticing that function $g : t \mapsto t \cdot \log t$ is convex for $t > 0$, we have (for each α)

$$\begin{aligned} f(\alpha) &= \frac{\left(\sum_x P(x) \cdot P(x)^{\alpha-1}\right) \log \left(\sum_x P(x) \cdot P(x)^{\alpha-1}\right) - \sum_x P(x) \cdot \left(P(x)^{\alpha-1} \log P(x)^{\alpha-1}\right)}{\sum_x P(x)^\alpha} \\ &= \frac{g(\sum_x P(x) \cdot t_x) - \sum_x P(x) \cdot g(t_x)}{\sum_x P(x)^\alpha} \leq 0, \end{aligned}$$

where $t_x \triangleq P(x)^{\alpha-1}$. Thus, we have finished the proof. □

d.) Compute the min-entropy $H_{\min}(X|Y)$ of the joint random variables (X,Y) distributed as

$P(x,y)$		X		
		0	1	2
Y	0	1/6	1/12	1/12
	1	1/12	1/6	1/12
	2	1/12	1/12	1/6

Solution: $H_{\min}(X|Y) = 1$. □

Exercise 5.2 Distributions with a large entropy gap [all]

It is possible to construct distributions that have a large gap between min-entropy and Shannon entropy. This shows that controlling the Shannon entropy or the mutual information is not sufficient for most cryptographic tasks.

a.) Given $\epsilon \in (0,1)$, construct a sequence of random variables $(X_2, X_3, \ldots, X_n, \ldots)$ where $X_n \in \{0,1,\ldots,n-1\}$, such that

$$\left.\begin{aligned} H(X_n) &\geq (1-\epsilon)\log n \\ H_{\min}(X_n) &= C, \end{aligned}\right\} \forall n \geq N$$

for some $N \in \mathbb{N}$ and some constant $C > 0$.

Solution: For each $n = 2, 3, 4, \ldots$, we consider the following distribution

$$P_{X_n}(x) = \begin{cases} \epsilon & \text{if } x = 0, \\ \frac{1-\epsilon}{n-1} & \text{otherwise.} \end{cases}$$

In this case,

$$\begin{aligned} H(X_n) &= H(\epsilon) + (1-\epsilon)\cdot\log(n-1) \geq (1-\epsilon)\cdot\log n && \forall n \geq \frac{1-\epsilon}{H(\epsilon)} + 1, \\ H_{\min}(X_n) &= -\log\epsilon && \forall n \geq \frac{1-\epsilon}{\epsilon} + 1. \end{aligned}$$

Thus the construction satisfies the requirements by letting $N = \left\lceil \max\{\frac{1-\epsilon}{H(\epsilon)} + 1, \frac{1-\epsilon}{\epsilon} + 1\} \right\rceil$, and $C = -\log\epsilon$. □

b.) Given $\epsilon \in (0,1)$, construct a sequence of random variables $((X_2, Y_2), (X_3, Y_3), \ldots, (X_n, Y_n), \ldots)$, where X_n, $Y_n \in \{0, 1, \ldots, n-1\}$, such that

$$\begin{aligned} H(X_n) = H_{\min}(X_n) &= \log n && \forall n \\ \left.\begin{aligned} H(X_n|Y_n) &\geq (1-\epsilon)\log n \\ H_{\min}(X_n|Y_n) &= C \end{aligned}\right\} && && \forall n \geq N \end{aligned}$$

for some $N \in \mathbb{N}$ and some constant $C > 0$.

Solution: For each $n = 2, 3, 4, \ldots$, we consider the following distribution

$$\begin{aligned} P_{Y_n}(y) &= \frac{1}{n} \\ P_{X_n|Y_n}(x|y) &= \begin{cases} \epsilon & \text{if } x = y, \\ \frac{1-\epsilon}{n-1} & \text{otherwise.} \end{cases} \end{aligned}$$

In this case, $P_{X_n}(x) = \sum_y P_{Y_n}(y) \cdot P_{X_n|Y_n}(x|y) = 1/n$. Thus, $H(X_n|Y_n) \geq (1-\epsilon)\log n$. Additionally,

$$\begin{aligned} H(X_n|Y_n) &= H(\epsilon) + (1-\epsilon)\cdot\log(n-1) \geq (1-\epsilon)\cdot\log n && \forall n \geq \frac{1-\epsilon}{H(\epsilon)} + 1, \\ H_{\min}(X_n|Y_n) &= -\log\epsilon && \forall n \geq \frac{1-\epsilon}{\epsilon} + 1. \end{aligned}$$

Thus the construction satisfies the requirements by letting $N = \left\lceil \max\{\frac{1-\epsilon}{H(\epsilon)} + 1, \frac{1-\epsilon}{\epsilon} + 1\} \right\rceil$, and $C = -\log\epsilon$. □

Exercise 5.3 Typical sets [all]

Consider a DMS with a two symbol alphabet $\{a, b\}$ where $p_X(a) = 2/3$ and $p_X(b) = 1/3$. Let $X^n = (X_1, \ldots, X_n)$ be a string of symbols emitted by the source with $n = 100,000$. Let $W(X_j)$ be the suprisal for the j-th source output, i.e., $W(X_j) = -\log 2/3$ for $X_j = a$ and $-\log 1/3$ for $X_j = b$. Define $W(X^n) = \sum_{j=1}^{n} W(X_j)$.

a.) Find the variance of $W(X_j)$. For $\epsilon = 0.01$, evaluate a bound on the probability of the typical set $\mathcal{A}_\epsilon^{(n)}$ using Chebyshev's inequality.

Solution: For notational convenience, we will denote the log pmf random variable by W. Now, note that W takes on values $-\log 2/3$ with probability $2/3$ and $-\log 1/3$ with probability $1/3$. Hence,

$$\operatorname{Var}(W) = \mathbb{E}[W^2] - \mathbb{E}[W]^2 = \frac{2}{9}.$$

The bound on the typical set, as derived using Chebyshev's inequality is

$$\Pr(X^n \in A_\epsilon^{(n)}) \geq 1 - \frac{\sigma_W^2}{n\epsilon^2}.$$

Substituting the values of $n = 10^5$ and $\epsilon = 0.01$, we obtain

$$\Pr(X^n \in A_\epsilon^{(n)}) \geq 1 - \frac{1}{45} = \frac{44}{45}$$

Loosely speaking this means that if we were to look at sequences of length 100,000 generated from our DMS, more than 97% of the time the sequence will be typical. □

b.) Let N_a be the number of a's in the string $X^n = (X_1, \ldots, X_n)$. The random variable (rv) N_a is the sum of n iid rv's. Show what these rv's are.

Solution: The rv N_a is the sum of n iid rv's Y_i , $N_a = \sum_{i=1}^{n} Y_i$ where Y_i's are Bernoulli with $\Pr(Y_i = 1) = 2/3$. □

c.) Express the rv $W(X^n)$ as a function of the rv N_a. Note how this depends on n.

Solution: The probability of a particular sequence X^n with N_a number of a's $(2/3)^{N_a}(1/3)^{n-N_a}$. Hence,

$$W(X^n) = -\log p_{X^n}(x^n) = -\log[(2/3)^{N_a}(1/3)^{n-N_a}] = n \log 3 - N_a.$$

□

d.) Express the typical set in terms of bounds on N_a. Use Chebyshev's inequality to derive bounds on the probability of the typical set, using properties of N_a instead of $W(X_j)$.
Hint: You may write $\mathcal{A}_\epsilon^{(n)} = \{x^n : \alpha < N_a < \beta\}$ and calculate α and β.

Solution: For a sequence X^n to be typical, it must satisfy

$$\left| -\frac{1}{n} \log p_{X^n}(x^n) - H(X) \right| < \epsilon$$

From (a) the source entropy is $H(X) = \mathbb{E}[W(X)] = \log 3 - 2/3$ and substituting in ϵ and $W(X^n)$ from part (d), we get

$$\left| \frac{N_a}{n} - \frac{2}{3} \right| \leq 0.01$$

Note the intuitive appeal of this condition! It says that for a sequence to be typical, the proportion of a's in that sequence will be very close to the probability that the DMS generates an a. Plugging in the value of n in the above equation, we get the bounds on

$$65,667 \leq N_a \leq 67,666.$$

□

e.) Find $\Pr(N_a = i)$ for $i = 0, 1, 2$. Find the probability of each individual string x^n for those values of i. Find the particular string x^n that has maximum probability over all sample values of X^n. What are the next most probable n-strings. Give a brief discussion of why the most probable n-strings are not regarded as typical strings.

Solution:

$$\Pr(N_a = 0) = \left(\frac{1}{3}\right)^n, \Pr(x^n = b^n) = \left(\frac{1}{3}\right)^n.$$

$$\Pr(N_a = 1) = n \left(\frac{2}{3}\right)\left(\frac{1}{3}\right)^{n-1}, \text{ the probability of each } x^n \text{ is } \left(\frac{2}{3}\right)\left(\frac{1}{3}\right)^{n-1}.$$

$$\Pr(N_a = 2) = \frac{n(n-1)}{2}\left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right)^{n-2}, \text{ the probability of each } x^n \text{ is } \left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right)^{n-2}.$$

The particular string x^n with maximum probability is a^n:

$$\Pr(x^n = a^n) = \left(\frac{2}{3}\right)^n .$$

The next most probable n-strings are x^n with 1 symbol b and $(n-1)$ symbol a's.

The typical strings should have around $\frac{2}{3}n$ symbol a's and around $\frac{1}{3}n$ symbol b's. The most probable n-strings are usually far from this situation. □

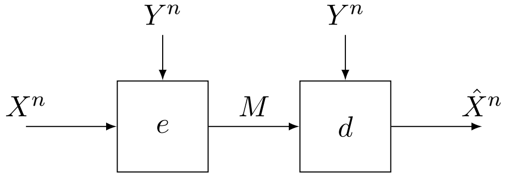
Exercise 5.4 Source coding with side information [EE5139]

Consider a memoryless source $(\boldsymbol{X}, \boldsymbol{Y})$ that produces in each iteration two random variables, X_i and Y_i, where X_i is private information and Y_i is public information. The pairs (X_i, Y_i) follow a joint distribution P_{XY} and are i.i.d.. We are looking for a fixed-length block code that compresses the private information $X^n = (X_1, X_2, \ldots, X_n)$ using the public information $Y^n = (Y_1, Y_2, \ldots, Y_n)$ such that the code can be decoded asymptotically error-free with help of the public information.

An $(n, 2^L)$-code for such a source is given by an encoder, $e : (X^n, Y^n) \to M$, and decoder, $d : (M, Y^n) \to \hat{X}^n$, as illustrated in the figure below. The codeword $M \in \{0,1\}^L$ is a binary string of length L. We define $R^*(\boldsymbol{X}|\boldsymbol{Y})$ as the infimum over all rates R such that there exists a sequence of $(n, 2^{nR})$-codes satisfying

$$\lim_{n\to\infty} \Pr\left[X^n \neq \hat{X}^n\right] = 0, \quad \text{where} \quad \hat{X}^n = d_n(e_n(X^n, Y^n), Y^n) \tag{2}$$

is a function of both X^n and Y^n. We want to establish that $R^*(\boldsymbol{X}|\boldsymbol{Y}) = H(X|Y)$.



a.) Determine $R^*(\boldsymbol{X}|\boldsymbol{Y})$, by intuitive or formal arguments, for the simple cases where

i.) X and Y are independent,

Solution: $R^*(\boldsymbol{X}|\boldsymbol{Y}) = H(X)$. If X and Y are independent, it means Y does not provide useful side information about X. □

ii.) $X = Y$,

Solution: $R^*(\boldsymbol{X}|\boldsymbol{Y}) = 0$. $X = Y$ means that the decoder can exactly recover X^n simply from the side information Y^n. □

b.) By explicitly constructing a code for the source (X, Y) using codes for the sources Y and X (with side information Y), show that $R^*(\boldsymbol{X}, \boldsymbol{Y}) \leq R^*(\boldsymbol{X}|\boldsymbol{Y}) + R^*(\boldsymbol{Y})$.

Solution: A code for the source (X, Y) can be constructed by concatenating the codes for sources Y and X (with side information Y). The error probability is given by

$$\begin{aligned}\Pr\left[(\hat{X}^n, \hat{Y}^n) \neq (X^n, Y^n)\right] &= \Pr\left[\hat{Y}^n \neq Y^n \vee \hat{X}^n \neq X^n\right] \\ &\leq \Pr\left[\hat{Y}^n \neq Y^n\right] + \Pr\left[\hat{X}^n \neq X^n\right] \to 0, \text{ as } n \to \infty.\end{aligned}$$

Let L_Y and $L_{X|Y}$ denote the number of bits of the optimal (shortest) codes for the sources Y and X with side information Y, respectively. Then we have

$$R^*(\boldsymbol{X}, \boldsymbol{Y}) \leq \frac{L_{X|Y} + L_Y}{n} = R^*(\boldsymbol{X}|\boldsymbol{Y}) + R^*(\boldsymbol{Y}).$$

□

c.) Show that the converse, $R^*(\boldsymbol{X}|\boldsymbol{Y}) \geq H(X|Y)$ using Fano's inequality. **Hint:** You will also need the following sequence of inequalities, which needs to be verified.

$$\begin{aligned} H(X^n|\hat{X}^n) &\geq H(X^n|Y^n M) && (3)\\ &= H(X^n M|Y^n) - H(M|Y^n) && (4)\\ &\geq H(X^n M|Y^n) - L && (5)\\ &\geq H(X^n|Y^n) - L\,. && (6) \end{aligned}$$

Solution: We first verify the sequence of inequalities. Eq. (3) is the data-processing inequality applied to the fact that $\hat{X}^n$ is computed from Y^n and M. Eq. (4) is the chain rule for conditional entropy. Eq. (5) follows from the dimension bound for $|M| \leq 2^L$. Finally, Eq. (6) uses the chain rule and the fact that $H(M|X^nY^n) \geq 0$.

Consider now any sequence of $(n, 2^L)$-codes that satisfy $\epsilon_n = \Pr[\hat{X}^n \neq X^n] \to 0$ in the limit $n \to \infty$.

By Fano's inequality and using the given sequence of inequalities, we have

$$H(\epsilon_n) + \epsilon_n n \log|\mathcal{X}| \geq H(X^n|\hat{X}^n) \geq H(X^n|Y^n) - L.$$

Hence, as $n \to \infty$,

$$\begin{aligned} \frac{L}{n} &\geq \frac{1}{n}(H(X^n|Y^n) - H(\epsilon_n) - \epsilon_n n \log|\mathcal{X}|)\\ &\geq \frac{1}{n}H(X^n|Y^n) - \frac{1}{n} - \epsilon_n \log|\mathcal{X}|\\ &= H(X|Y) - \frac{1}{n} - \epsilon_n \log|\mathcal{X}|\\ &\to H(X|Y). \end{aligned}$$

Since this holds for any sequence of codes, we conclude that $R^*(\boldsymbol{X}|\boldsymbol{Y}) \geq H(X|Y)$. □

d.) Give a formal proof or a sketch of a proof that $R^*(\boldsymbol{X}|\boldsymbol{Y}) \leq H(X|Y)$. **Hint:** Consider the typical set

$$\mathcal{A}_\epsilon^{(n)}(\boldsymbol{X}|\boldsymbol{Y}) := \left\{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \left|\frac{1}{n}\log\frac{1}{P_{X^n|Y^n}(x^n|y^n)} - H(X|Y)\right| \leq \epsilon\right\}. \quad (7)$$

Solution: For any $\epsilon > 0$, define a typical set for $(\boldsymbol{X}, \boldsymbol{Y})$,

$$\mathcal{A}_\epsilon^{(n)}(\boldsymbol{X}|\boldsymbol{Y}) \triangleq \left\{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \left|\frac{1}{n}\log\frac{1}{P_{X^n|Y^n}(x^n|y^n)} - H(X|Y)\right| \leq \epsilon\right\},$$

where $P_{X^n|Y^n}(x^n|y^n) = \prod_{i=1}^n P_{X|Y}(x_i|y_i)$ for all $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$.

By definition, for any $(x^n, y^n) \in \mathcal{A}_\epsilon^{(n)}(\boldsymbol{X}|\boldsymbol{Y})$, we can establish that

$$\begin{aligned} & P_{X^n|Y^n}(x^n|y^n) \geq 2^{-n(H(X|Y)+\epsilon)}\\ \implies\ & 1 \geq \sum_{x^n, y^n \in \mathcal{A}_\epsilon^{(n)}(\boldsymbol{X}|\boldsymbol{Y})} P_{X^n|Y^n}(x^n|y^n) \geq |\mathcal{A}_\epsilon^{(n)}(\boldsymbol{X}|\boldsymbol{Y})| 2^{-n(H(X|Y)+\epsilon)}\\ \implies\ & |\mathcal{A}_\epsilon^{(n)}(\boldsymbol{X}|\boldsymbol{Y})| \leq 2^{nH(X|Y)+\epsilon}. \end{aligned}$$

Furthermore, let $Z_i = \log \frac{1}{P_{X|Y}(X_i|Y_i)} - H(X|Y)$ and we have

$$\begin{aligned}
\Pr\left[(X^n, Y^n) \in \mathcal{A}_\epsilon^{(n)}(\boldsymbol{X}|\boldsymbol{Y})\right] &= \Pr\left[\left|\frac{1}{n}\log\frac{1}{P_{X^n|Y^n}(X^n|Y^n)} - H(X|Y)\right| \le \epsilon\right] \\
&= \Pr\left[\left|\frac{1}{n}\log\frac{1}{\prod_{i=1}^n P_{X|Y}(X_i|Y_i)} - H(X|Y)\right| \le \epsilon\right] \\
&= \Pr\left[\left|\frac{1}{n}\sum_{i=1}^n \log\frac{1}{P_{X|Y}(X_i|Y_i)} - H(X|Y)\right| \le \epsilon\right] \\
&= 1 - \Pr\left[\left|\frac{1}{n}\sum_{i=1}^n \log\frac{1}{P_{X|Y}(X_i|Y_i)} - H(X|Y)\right| > \epsilon\right] \\
&= 1 - \Pr\left[\left|\frac{1}{n}\sum_{i=1}^n Z_i\right| > \epsilon\right].
\end{aligned}$$

Since Z_i are i.i.d and zero mean, by the weak law of large numbers, we have

$$\lim_{n\to\infty} \Pr\left[(X^n, Y^n) \in \mathcal{A}_\epsilon^{(n)}(\boldsymbol{X}|\boldsymbol{Y})\right] = 1 - \lim_{n\to\infty} \Pr\left[\left|\frac{1}{n}\sum_{i=1}^n Z_i\right| > \epsilon\right] = 1.$$

Encoder e:

$$e(x^n|y^n) = \begin{cases} m(x^n|y^n) & (x^n, y^n) \in \mathcal{A}_\epsilon^{(n)}(\boldsymbol{X}|\boldsymbol{Y}), \\ 0^L & (x^n, y^n) \notin \mathcal{A}_\epsilon^{(n)}(\boldsymbol{X}|\boldsymbol{Y}). \end{cases}$$

Decoder d: Given y^n, m, output any x^n such that $m = m(x^n|y^n)$.

Then the error probability is given by

$$\Pr\left[\hat{X}^n \neq X^n\right] = 1 - \Pr\left[(X^n, Y^n) \in \mathcal{A}_\epsilon^{(n)}(\boldsymbol{X}|\boldsymbol{Y})\right] \to 0 \ \text{ for } n \to \infty.$$

This implies that $R = \frac{L}{n} \le H(X|Y) + \epsilon$ is achievable and thus

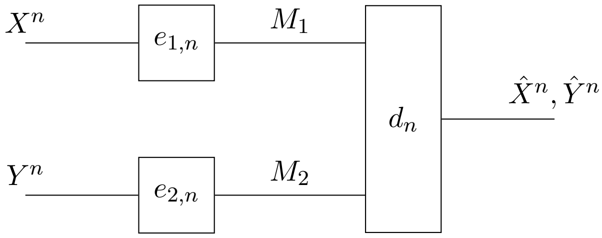
$$R^*(\boldsymbol{X}|\boldsymbol{Y}) \le H(X|Y).$$

□

Exercise 5.5 Achievability for the Slepian–Wolf coding problem [EE6139]

We return to the Exercise 4.4 from the last homework. Let X and Y be a pair of jointly distributed random variables. (X is distributed on finite set $\mathcal{X}$, and Y is distributed on finite set $\mathcal{Y}$.) An $(n, 2^{nL_1}, 2^{nL_2})$-separately-encoded-jointly-decoded source code consists of a pair of encoders e_1, e_2, and a decoder d, where

- $e_1 : \mathcal{X}^n \to \{0,1\}^{nL_1}$,
- $e_2 : \mathcal{Y}^n \to \{0,1\}^{nL_2}$, and
- $d : \{0,1\}^{nL_1} \times \{0,1\}^{nL_2} \to \mathcal{X}^n \times \mathcal{Y}^n$.



The rate pair (R_1, R_2) is said to be achievable for DMS (X, Y) if there exists a sequence of $(n, 2^{nL_1}, 2^{nL_2})$-codes with encoders $e_{1,n}$, $e_{2,n}$ and decoder d_n such that

$$\lim_{n\to\infty} P\{(\hat{X}^n, \hat{Y}^n) \neq (X^n, Y^n)\} = 0$$

where

$$(\hat{X}^n, \hat{Y}^n) = d_n(M_1, M_2),\ M_1 = e_{1,n}(X^n),\ \text{and } M_2 = e_{2,n}(Y^n)$$

are the reconstructed source and codewords respectively.
This time, we are interested in the achievability of the problem.

a.) **An alternative for typical sequences** Given $n \in \mathbb{N}$ and $\epsilon \in (0,1)$, we define the set of Y-sequences as

$$\mathcal{T}_\epsilon^{(n)}(Y) \triangleq \left\{ \boldsymbol{y} \in \mathcal{Y}^n : \left| \frac{\sum_{i=1}^n \delta_{y,y_i}}{n} - p_Y(y) \right| < \left\lceil \sqrt{\frac{|\mathcal{Y}|}{\epsilon}} \right\rceil \sqrt{\frac{p_Y(y)(1-p_Y(y))}{n}} \quad \forall y \in \mathcal{Y} \right\}.$$

Show that

i.) $P\left[Y^n \in \mathcal{T}_\epsilon^{(n)}(Y)\right] \geq 1 - \epsilon$
Solution:

$$\begin{aligned} P\left[Y^n \notin \mathcal{T}_\epsilon^{(n)}(Y)\right] &= P\left[\left| \frac{\sum_{i=1}^n \delta_{y,Y_i}}{n} - p_Y(y) \right| \geq \left\lceil \sqrt{\frac{|\mathcal{Y}|}{\epsilon}} \right\rceil \sqrt{\frac{p_Y(y)(1-p_Y(y))}{n}} \quad \exists y \in \mathcal{Y}\right] \\ &\leq \sum_{y\in\mathcal{Y}} P\left[\left| \frac{\sum_{i=1}^n \delta_{y,Y_i}}{n} - p_Y(y) \right| \geq \left\lceil \sqrt{\frac{|\mathcal{Y}|}{\epsilon}} \right\rceil \sqrt{\frac{p_Y(y)(1-p_Y(y))}{n}}\right] \\ &\leq \sum_{y\in\mathcal{Y}} \left(\sqrt{\frac{|\mathcal{Y}|}{\epsilon}}\right)^{-2} = \epsilon, \end{aligned}$$

where we have used Chebyshev's inequality in the last line. □

ii.) There exists some $A > 0$ independent from n and ϵ such that

$$2^{-nH(Y) - A\sqrt{n/\epsilon}} < p_{Y^n}(\boldsymbol{y}) < 2^{-nH(Y) + A\sqrt{n/\epsilon}}$$

for all $\boldsymbol{y} \in \mathcal{T}_\epsilon^{(n)}(Y)$.
Solution: Note that for any n-length vector $\boldsymbol{y}$, we have

$$\log p_{Y^n}(\boldsymbol{y}) = \log \prod_{y\in\mathcal{Y}} p_Y(y)^{\sum_{i=1}^n \delta_{y,y_i}} = \sum_{y\in\mathcal{Y}} \left(\sum_{i=1}^n \delta_{y,y_i}\right) \cdot \log p_Y(y)$$

For $\boldsymbol{y} \in \mathcal{T}_\epsilon^{(n)}(Y)$, note that

$$\begin{aligned} \sum_{i=1}^n \delta_{y,y_i} \in &\left(np_Y(y) - n \left\lceil \sqrt{\frac{|\mathcal{Y}|}{\epsilon}} \right\rceil \sqrt{\frac{p_Y(y)(1-p_Y(y))}{n}}, \right. \\ &\left. np_Y(y) + n \left\lceil \sqrt{\frac{|\mathcal{Y}|}{\epsilon}} \right\rceil \sqrt{\frac{p_Y(y)(1-p_Y(y))}{n}} \right). \\ &\subset \left(np_Y(y) - f(y)\sqrt{\frac{n}{\epsilon}}, np_Y(y) + f(y)\sqrt{\frac{n}{\epsilon}} \right) \end{aligned}$$

where $f(y) \triangleq 2\sqrt{|\mathcal{Y}| \cdot p_Y(y)(1-p_Y(y))}$. Thus, by defining

$$A \triangleq -\sum_{y\in\mathcal{Y}} f(y) \cdot \log p_Y(y),$$

we have

$$-nH(Y) - A\sqrt{\frac{n}{\epsilon}} < \log p_{Y^n}(\boldsymbol{y}) < -nH(Y) + A\sqrt{\frac{n}{\epsilon}},$$

which are equivalent to the to-be-proven inequalities. □

iii.) $\lim_{n\to\infty} \frac{1}{n} \log_2 \left|\mathcal{T}_\epsilon^{(n)}(Y)\right| = H(Y)$.

Solution: Note that

$$\left|\mathcal{T}_\epsilon^{(n)}(Y)\right| \cdot \min_{\boldsymbol{y}\in\mathcal{T}_\epsilon^{(n)}(Y)} p_{Y^n}(\boldsymbol{y}) \leq P\left[Y^n \in \mathcal{T}_\epsilon^{(n)}(Y)\right] \leq 1,$$
$$\left|\mathcal{T}_\epsilon^{(n)}(Y)\right| \cdot \max_{\boldsymbol{y}\in\mathcal{T}_\epsilon^{(n)}(Y)} p_{Y^n}(\boldsymbol{y}) \geq P\left[Y^n \in \mathcal{T}_\epsilon^{(n)}(Y)\right] \geq 1-\epsilon.$$

Combining with the results from ii.), we have

$$(1-\epsilon)\cdot 2^{nH(Y)-A\sqrt{n/\epsilon}} \leq \left|\mathcal{T}_\epsilon^{(n)}(Y)\right| \leq 2^{nH(Y)+A\sqrt{n/\epsilon}},$$

or, equivalently,

$$\log(1-\epsilon) + nH(Y) - A\sqrt{n/\epsilon} \leq \log\left|\mathcal{T}_\epsilon^{(n)}(Y)\right| \leq nH(Y) + A\sqrt{n/\epsilon}.$$

Thus,

$$\left|\frac{1}{n}\log\left|\mathcal{T}_\epsilon^{(n)}(Y)\right| - H(Y)\right| \leq \frac{|\log(1-\epsilon)|}{n} + \frac{A}{\sqrt{\epsilon}\sqrt{n}} \to 0$$

as $n \to \infty$. □

b.) **Position-based coding** Given positive integer n and $\epsilon > 0$, let $M_X \triangleq \left\lfloor 2^{n(H(Y|X)+\epsilon)} \right\rfloor$, and let M be another positive integer. Let $\{\boldsymbol{X}_{i,j}\}_{i,j}$ be a set of i.i.d. random variables on $\mathcal{X}^n$, where $i \in \{1,\ldots,M_X\}$, $j \in \{1,\ldots,M\}$, and

$$p_{\boldsymbol{X}_{i,j}}(\boldsymbol{x}) = \prod_{k=1}^n p_X(x_k)$$

for each (i,j).

i.) Suppose $I(X,Y) > \frac{1}{2}\epsilon$, and let $M = \left\lfloor 2^{n(I(X,Y)-\frac{1}{2}\epsilon)} \right\rfloor$. Prove that, for n large enough,

$$P\left[X^n \neq \boldsymbol{X}_{i,j}\ \forall(i,j)\right] < 2\epsilon.$$

Solution: Considering the set $\mathcal{T}_\epsilon^n(X) \subset \mathcal{X}^n$, and that $\{\boldsymbol{X}_{i,j}\}_{i,j}$ are i.i.d., we have

$$\begin{aligned}
&P\left[X^n \neq \boldsymbol{X}_{i,j}\ \forall(i,j)\right] = P\left[X^n \neq \boldsymbol{X}_{i,j}\ \forall(i,j)|X^n = \boldsymbol{x}\right]\cdot P[X^n = \boldsymbol{x}] \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}^n} p_{X^n}(\boldsymbol{x}) \cdot \prod_{i,j}(1 - p_{\boldsymbol{X}_{i,j}}(\boldsymbol{x})) \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}^n} p_{X^n}(\boldsymbol{x}) \cdot (1 - p_{X^n}(\boldsymbol{x}))^{MM_X} \\
&= \sum_{\boldsymbol{x}\in\mathcal{T}_\epsilon^n(X)} p_{X^n}(\boldsymbol{x}) \cdot (1 - p_{X^n}(\boldsymbol{x}))^{MM_X} + \sum_{\boldsymbol{x}\in\mathcal{X}^n\setminus\mathcal{T}_\epsilon^n(X)} p_{X^n}(\boldsymbol{x}) \cdot (1 - p_{X^n}(\boldsymbol{x}))^{MM_X} \\
&\leq \sum_{\boldsymbol{x}\in\mathcal{T}_\epsilon^n(X)} p_{X^n}(\boldsymbol{x}) \cdot (1 - p_{X^n}(\boldsymbol{x}))^{MM_X} + \epsilon.
\end{aligned}$$

Note that for any $\boldsymbol{x} \in \mathcal{T}_\epsilon^n(X)$, we have $p_{X^n}(\boldsymbol{x}) \geq 2^{-nH(X)-A\sqrt{n/\epsilon}}$ for some A independent from n and ϵ. In this case,

$$\begin{aligned}\sum_{\boldsymbol{x}\in\mathcal{T}_\epsilon^n(X)} p_{X^n}(\boldsymbol{x}) \cdot (1-p_{X^n}(\boldsymbol{x}))^{MM_X} &\leq \left[1-2^{-nH(X)-A\sqrt{n/\epsilon}}\right]^{MM_X} \cdot \sum_{\boldsymbol{x}\in\mathcal{T}_\epsilon^n(X)} p_{X^n}(\boldsymbol{x}) \\ &\leq \left[1-2^{-nH(X)-A\sqrt{n/\epsilon}}\right]^{MM_X}.\end{aligned}$$

Furthermore, denoting the quantity on the right-hand side of the above line by Z, we have

$$\begin{aligned}\log Z &= MM_X \log\left[1-2^{-nH(X)-A\sqrt{n/\epsilon}}\right]^{MM_X} \\ &\leq 2^{n(I(X,Y)-\frac{1}{2}\epsilon)} \cdot 2^{n(H(Y|X)+\epsilon)} \cdot \log\left[1-2^{-nH(X)-A\sqrt{n/\epsilon}}\right]^{MM_X} \\ &\leq -2^{n(I(X,Y)-\frac{1}{2}\epsilon)} \cdot 2^{n(H(Y|X)+\epsilon)} \cdot 2^{-nH(X)-A\sqrt{n/\epsilon}} \\ &= -2^{n\epsilon/2-A\sqrt{n/\epsilon}}.\end{aligned}$$

Notice that for fixed ϵ, above tends to $-\infty$ as $n \to \infty$. It must hold that $-2^{n\epsilon/2-A\sqrt{n/\epsilon}} < \log\epsilon$ (and thus $Z < \epsilon$) for n large enough. Therefore,

$$P\left[X^n \neq \boldsymbol{X}_{i,j}\ \forall(i,j)\right] \leq Z + \epsilon < 2\epsilon$$

for n large enough. □

ii.) Let A and B be a pair of random variable denoting the "smallest" indices a, b such that that $X^n = \boldsymbol{X}_{a,b}$. Namely,

$$p_{A,B|X^n,\{\boldsymbol{X}_{i,j}\}}(a,b|\boldsymbol{x},\{\boldsymbol{x}_{i,j}\}) = \begin{cases} & \boldsymbol{x} = \boldsymbol{x}_{a,b} \\ 1 \quad \text{if} & \boldsymbol{x} \neq \boldsymbol{x}_{i,j} \quad \forall i < a \\ & \boldsymbol{x} \neq \boldsymbol{x}_{a,j} \quad \forall j < b \\ 0 \quad \text{otherwise} & \end{cases}.$$

We take the convention that $(A,B) = (\infty,\infty)$ if $X^n \neq \boldsymbol{X}_{i,j}$ for all i,j. Prove that

$$P\left[A<\infty,\, B<\infty,\, p_{Y^n|X^n}(Y^n|\boldsymbol{X}_{A,j}) \geq p_{Y^n|X^n}(Y^n|X^n)\ \exists j \neq B\right] < \epsilon$$

for n large enough.

Solution: Firstly, we can rewrite above probability into

$$\begin{aligned}&P\left[A<\infty,\, B<\infty,\, p_{Y^n|X^n}(Y^n|\boldsymbol{X}_{A,j}) \geq p_{Y^n|X^n}(Y^n|X^n)\ \exists j \neq B\right] \\ =&\sum_{\boldsymbol{x},\boldsymbol{y}} p_{X^nY^n}(\boldsymbol{x},\boldsymbol{y}) \sum_{a=1}^{M_X} p_{A|X^n,Y^n}(a|\boldsymbol{x},\boldsymbol{y}) \sum_{b=1}^{M} p_{B|A,X^n,Y^n}(b|a,\boldsymbol{x},\boldsymbol{y}) \cdot \\ &P\left[p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{X}_{a,j}) \geq p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{x})\ \exists j \neq b | X^n = \boldsymbol{x}, Y^n = \boldsymbol{y}, A = a, B = b\right] \\ =&\sum_{\boldsymbol{x},\boldsymbol{y}} p_{X^nY^n}(\boldsymbol{x},\boldsymbol{y}) \cdot \sum_{a=1}^{M_X} (1-p_{X^n}(\boldsymbol{x}))^{(a-1)M} \cdot \sum_{b=1}^{M} (1-p_{X^n}(\boldsymbol{x}))^{b-1} p_{X^n}(\boldsymbol{x}) \cdot C_{a,b}(\boldsymbol{x},\boldsymbol{y})\end{aligned}$$

where

$$C_{a,b}(\boldsymbol{x},\boldsymbol{y}) \triangleq P\left[\begin{array}{c|ll} p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{X}_{a,j}) \geq p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{x}) & X^n = \boldsymbol{x},\, Y^n = \boldsymbol{y} & \\ \exists j \neq b & \boldsymbol{X}_{i,j} \neq \boldsymbol{x} & \forall i < a \\ & \boldsymbol{X}_{a,j} \neq \boldsymbol{x} & \forall j < b \end{array}\right].$$

For any $\rho \in (0,1)$, we can bound $C_{a,b}(\boldsymbol{x},\boldsymbol{y})$ as

$$C_{a,b}(\boldsymbol{x},\boldsymbol{y}) \leq \left\{ \sum_{j\neq b} P\left[p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{X}_{a,j}) \geq p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{x})|\cdots\right] \right\}^{\rho}.$$

Note that, for $j > b$,

$$\begin{aligned}
&P\left[p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{X}_{a,j}) \geq p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{x})|\cdots\right] \\
=&P\left[p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{X}_{a,j}) \geq p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{x})|X^n=\boldsymbol{x}, Y^n=\boldsymbol{y}\right] \\
=&\sum_{\tilde{\boldsymbol{x}}\in\mathcal{X}^n: p_{Y^n|X^n}(\boldsymbol{y}|\tilde{\boldsymbol{x}})\geq p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{x})} p_{X^n}(\tilde{\boldsymbol{x}}).
\end{aligned}$$

Whereas, for $j < b$,

$$\begin{aligned}
&P\left[p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{X}_{a,j}) \geq p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{x})|\cdots\right] \\
=&P\left[p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{X}_{a,j}) \geq p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{x})|X^n=\boldsymbol{x}, Y^n=\boldsymbol{y}, \boldsymbol{X}_{a,j}\neq\boldsymbol{x}\right] \\
=&\frac{P\left[p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{X}_{a,j}) \geq p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{x}), \boldsymbol{X}_{a,j}\neq\boldsymbol{x}|X^n=\boldsymbol{x}, Y^n=\boldsymbol{y}\right]}{P\left[\boldsymbol{X}_{a,j}\neq\boldsymbol{x}|X^n=\boldsymbol{x}, Y^n=\boldsymbol{y}\right]} \\
=&\frac{P\left[p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{X}_{a,j}) \geq p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{x}), \boldsymbol{X}_{a,j}\neq\boldsymbol{x}|X^n=\boldsymbol{x}, Y^n=\boldsymbol{y}\right]}{P\left[\boldsymbol{X}_{a,j}\neq\boldsymbol{x}\right]} \\
\leq&\frac{P\left[p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{X}_{a,j}) \geq p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{x})|X^n=\boldsymbol{x}, Y^n=\boldsymbol{y}\right]}{P\left[\boldsymbol{X}_{a,j}\neq\boldsymbol{x}\right]} \\
=&\left(1-p_{X^n}(\boldsymbol{x})\right)^{-1}\cdot\sum_{\tilde{\boldsymbol{x}}\in\mathcal{X}^n: p_{Y^n|X^n}(\boldsymbol{y}|\tilde{\boldsymbol{x}})\geq p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{x})} p_{X^n}(\tilde{\boldsymbol{x}}).
\end{aligned}$$

Hence, following holds for all $s \geq 0$

$$\begin{aligned}
C_{a,b}(\boldsymbol{x},\boldsymbol{y}) &\leq \left\{ \left[(b-1)(1-p_{X^n}(\boldsymbol{x}))^{-1}+M-b\right]\cdot\sum_{\substack{\tilde{\boldsymbol{x}}\in\mathcal{X}^n:\\ p_{Y^n|X^n}(\boldsymbol{y}|\tilde{\boldsymbol{x}})\geq p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{x})}} p_{X^n}(\tilde{\boldsymbol{x}}) \right\}^{\rho} \\
&\leq \left[(b-1)(1-p_{X^n}(\boldsymbol{x}))^{-1}+M-b\right]^{\rho}\cdot\left[\sum_{\tilde{\boldsymbol{x}}\in\mathcal{X}^n} p_{X^n}(\tilde{\boldsymbol{x}})\left(\frac{p_{Y^n|X^n}(\boldsymbol{y}|\tilde{\boldsymbol{x}})}{p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{x})}\right)^{s}\right]^{\rho},
\end{aligned}$$

since $\frac{p_{Y^n|X^n}(\boldsymbol{y}|\tilde{\boldsymbol{x}})}{p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{x})} \geq 1$ for all $\tilde{\boldsymbol{x}} \in \mathcal{X}^n$ such that $p_{Y^n|X^n}(\boldsymbol{y}|\tilde{\boldsymbol{x}}) \geq p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{x})$. Substituting above bound on $C_{a,b}(\boldsymbol{x},\boldsymbol{y})$ into the expression for the targeting probability, we have

$$\begin{aligned}
&P\left[A<\infty,\, B<\infty,\, p_{Y^n|X^n}(Y^n|\boldsymbol{X}_{A,j}) \geq p_{Y^n|X^n}(Y^n|X^n)\ \exists j\neq B\right] \\
\leq&\sum_{\boldsymbol{x},\boldsymbol{y}} p_{X^nY^n}(\boldsymbol{x},\boldsymbol{y})\cdot\sum_{a=1}^{M_X}(1-p_{X^n}(\boldsymbol{x}))^{(a-1)M}\cdot\sum_{b=1}^{M}(1-p_{X^n}(\boldsymbol{x}))^{b-1}p_{X^n}(\boldsymbol{x})\cdot \\
&\left[(b-1)(1-p_{X^n}(\boldsymbol{x}))^{-1}+M-b\right]^{\rho}\cdot\left[\sum_{\tilde{\boldsymbol{x}}\in\mathcal{X}^n} p_{X^n}(\tilde{\boldsymbol{x}})\left(\frac{p_{Y^n|X^n}(\boldsymbol{y}|\tilde{\boldsymbol{x}})}{p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{x})}\right)^{s}\right]^{\rho}.
\end{aligned}$$

We make the following claim, and defer its proof to the very end.
Claim. For $\alpha \in [0,1)$, $\beta \in [0,1]$ and m, n being positive integers, it holds that

$$\sum_{i=1}^{m}\alpha^{(i-1)n}\cdot\sum_{j=1}^{n}\alpha^{j-1}\cdot\left(\frac{j-1}{\alpha}+n-j\right)^{\beta} \leq \frac{n^{\beta}}{1-\alpha}.$$

Using above claim, we have

$$\begin{aligned}
&P\left[A<\infty,\, B<\infty,\, p_{Y^n|X^n}(Y^n|\boldsymbol{X}_{A,j}) \geq p_{Y^n|X^n}(Y^n|X^n)\ \exists j \neq B\right] \\
\leq& \sum_{\boldsymbol{x},\boldsymbol{y}} p_{X^nY^n}(\boldsymbol{x},\boldsymbol{y}) \cdot M^\rho \cdot \left[\sum_{\tilde{\boldsymbol{x}}\in\mathcal{X}^n} p_{X^n}(\tilde{\boldsymbol{x}}) \left(\frac{p_{Y^n|X^n}(\boldsymbol{y}|\tilde{\boldsymbol{x}})}{p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{x})}\right)^s\right]^\rho \\
=& \sum_{\boldsymbol{x},\boldsymbol{y}} p_{X^n}(\boldsymbol{x}) \cdot \left[p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{x})\right]^{1-\rho s} \cdot M^\rho \cdot \left[\sum_{\tilde{\boldsymbol{x}}\in\mathcal{X}^n} p_{X^n}(\tilde{\boldsymbol{x}}) \left(p_{Y^n|X^n}(\boldsymbol{y}|\tilde{\boldsymbol{x}})\right)^s\right]^\rho .
\end{aligned}$$

By picking $s = 1/(1+\rho)$, above can be rewritten as

$$\begin{aligned}
P[\cdots] &\leq M^\rho \sum_{\boldsymbol{y}} \left(\sum_{\boldsymbol{x}} p_{X^n}(\boldsymbol{x}) p_{Y^n|X^n}(\boldsymbol{y}|\boldsymbol{x})^{\frac{1}{1+\rho}}\right)^{1+\rho} \\
&= M^\rho \cdot \left[\sum_{y} \left(\sum_{x} p_X(x) p_{Y|X}(y|x)^{\frac{1}{1+\rho}}\right)^{1+\rho}\right]^n \\
&\leq 2^{n\rho(I(X,Y)-\frac{1}{2}\epsilon)} \cdot \left[\sum_{y} \left(\sum_{x} p_X(x) p_{Y|X}(y|x)^{\frac{1}{1+\rho}}\right)^{1+\rho}\right]^n \\
&= 2^{-n\rho(V(\rho)/\rho - I(X,Y)+\frac{1}{2}\epsilon)},
\end{aligned}$$

where

$$V(\rho) \triangleq -\log_2 \left\{\sum_{y} \left(\sum_{x} p_X(x) p_{Y|X}(y|x)^{\frac{1}{1+\rho}}\right)^{1+\rho}\right\}.$$

Now, notice that V is differentiable in an open neighborhood around $\rho = 0$ and that $V(0) = 0$. Define function E in this neighborhood excluding the point $\rho = 0$ as $E(\rho) \triangleq V(\rho)/\rho$. Function E must be continuous in this deleted neighborhood, and is continuous in the neighborhood around $\rho = 0$ by extending to point 0 via limit. Namely,

$$E(0) \triangleq \lim_{\rho\to 0} E(\rho) = \lim_{\rho\to 0} \frac{V(\rho)}{\rho} = \left.\frac{\mathrm{d}}{\mathrm{d}\rho}\right|_{\rho=0} V(\rho) = I(X,Y).$$

Thus, we can pick some $\rho_0 > 0$ such that $V(\rho_0)/\rho_0 - I(X,Y) > -\epsilon/4$. In this case,

$$P[\cdots] \leq 2^{-n\rho_0(V(\rho_0)/\rho_0 - I(X,Y)+\frac{1}{2}\epsilon)} < 2^{-\frac{1}{4}n\rho\epsilon} \to 0$$

as $n \to \infty$. Therefore, for n large enough, $P[\cdots] < \epsilon$.

Proof of the claim. Firstly, note that the function $x \mapsto x^\beta$ is concave. By Jensen's inequality, we have

$$\begin{aligned}
\frac{\alpha-1}{\alpha^n-1} \cdot \sum_{j=1}^n \alpha^{j-1} \cdot \left(\frac{j-1}{\alpha} + n - j\right)^\beta &\leq \left(\frac{\alpha-1}{\alpha^n-1} \cdot \sum_{j=1}^n \alpha^{j-1} \cdot \left(\frac{j-1}{\alpha} + n - j\right)\right)^\beta \\
&= \left(\frac{\alpha^{n-1}-1}{\alpha^n-1} \cdot n\right)^\beta .
\end{aligned}$$

Since $\alpha < 1$, we further have

$$\sum_{j=1}^n \alpha^{j-1} \cdot \left(\frac{j-1}{\alpha} + n - j\right)^\beta \leq \frac{\alpha^n-1}{\alpha-1} \cdot n^\beta \cdot \left(\frac{\alpha^{n-1}-1}{\alpha^n-1}\right)^\beta \leq \frac{\alpha^n-1}{\alpha-1} \cdot n^\beta .$$

Also note that $\sum_{i=1}^{m} \alpha^{(i-1)n} = \frac{\alpha^{nm}-1}{\alpha^{n}-1}$. Therefore,

$$\sum_{i=1}^{m} \alpha^{(i-1)n} \cdot \sum_{j=1}^{n} \alpha^{j-1} \cdot \left(\frac{j-1}{\alpha} + n - j \right)^{\beta} \leq \frac{\alpha^{nm}-1}{\alpha^{n}-1} \cdot \frac{\alpha^{n}-1}{\alpha - 1} \cdot n^{\beta} \leq \frac{n^{\beta}}{1-\alpha}.$$

□

c.) Based on the arguments in a.) and b.), show that, for any δ_1, $\delta_2 > 0$, the following rates are achievable

$$R_1 = H(X|Y) + \delta_1, \tag{8}$$

$$R_2 = H(Y) + \delta_2. \tag{9}$$

Solution: Let $\epsilon \in (0, \min\{1, \delta_1\})$ be arbitrarily picked.

To encode $\boldsymbol{y} \in \mathcal{Y}^n$: We firstly prepare the set of Y-sequences $\mathcal{T}_{\epsilon}^{(n)}(Y)$ and index the elements in this set by $\{1, \dots, M_Y\}$, where $M_Y = \left|\mathcal{T}_{\epsilon}^{(n)}(Y)\right|$. Namely, $\mathcal{T}_{\epsilon}^{(n)}(Y) = \{\hat{\boldsymbol{y}}_1, \dots, \hat{\boldsymbol{y}}_{M_Y}\}$. If $\boldsymbol{y}$ is in $\mathcal{T}_{\epsilon}^{(n)}(Y)$, we encode it by its index in $\mathcal{T}_{\epsilon}^{(n)}(Y)$; otherwise we encode it to something fixed. More precisely,

$$e_Y^{(n)} : \boldsymbol{y}^n \mapsto \begin{cases} \mathsf{index}_{\mathcal{T}_{\epsilon}^{(n)}(Y)}(\boldsymbol{y}) & \text{if } \boldsymbol{y} \in \mathcal{T}_{\epsilon}^{(n)}(Y) \\ 1 & \text{otherwise} \end{cases}.$$

By a.)iii.), for n large enough, $M_Y \leq 2^{n(H(Y)+\delta_2)} = 2^{nR_2}$. Thus, to transmit the encoded message, we need at most nR_2 bits. Upon receiving $m_2 \in \{1, \dots, M_Y\}$, we pick the m_2-th element in $\mathcal{T}_{\epsilon}^{(n)}(Y)$ as the decoded message. Namely,

$$d_Y^{(n)} : m_2 \mapsto \hat{\boldsymbol{y}}_{m_2}.$$

Denoting the output by random variable $\hat{Y}^{(n)}$, we have $\hat{Y}^{(n)} = d_Y^{(n)}(e_Y^{(n)}(Y^n))$. By a.)i.), we know

$$P\left[\hat{Y}^{(n)} \neq Y^n\right] \leq P\left[Y^n \notin \mathcal{T}_{\epsilon}^{(n)}(Y)\right] \leq \epsilon$$

for n large enough.

To transmit X^n, we construct following *randomized* encoders and decoders based on auxiliary random variables $\{\boldsymbol{X}_{i,j}\}_{i,j}$ where $i \in \{1, \dots, M_X\}$, $j \in \{1, \dots, M\}$, and $\boldsymbol{X}_{i,j} \in \mathcal{X}^n$ are i.i.d. random variables for each (i,j) and have the same distribution as X^n. Here, we pick $M_X = \lfloor 2^{n(H(Y|X)+\epsilon)} \rfloor$ and $M = \left\lfloor 2^{n(I(X,Y)-\frac{1}{2}\epsilon)} \right\rfloor$.

To encode $\boldsymbol{x} \in \mathcal{X}^n$: We *try* to find the "smallest" (a,b) such that $\boldsymbol{X}_{a,b} = \boldsymbol{x}$. If such (a,b) exists, we use a as the encoded message; otherwise we encode $\boldsymbol{x}$ to something fixed. Namely, given a realization of $\{\boldsymbol{X}_{i,j}\}_{i,j}$ as $\{\tilde{\boldsymbol{x}}_{i,j}\}_{i,j}$,

$$e_X^{(n)}(\{\tilde{\boldsymbol{x}}_{i,j}\}_{i=1,\dots,M_X;j=1,\dots,M}) : \boldsymbol{x}^n \mapsto \begin{cases} & \exists b \text{ s.t. } \boldsymbol{x}_{a,b} = \boldsymbol{x} \\ a & \text{if} \quad \forall i < a,\, \forall j,\, \boldsymbol{x}_{i,j} \neq \boldsymbol{x} \\ & \forall j < b,\, \boldsymbol{x}_{a,j} \neq \boldsymbol{x} \\ 1 & \text{otherwise} \end{cases}.$$

Since $\epsilon \leq \delta_1$, we have $M_X \leq 2^{nR_1}$, i.e., transmitting above encoded message requires at most nR_1 bits. Upon receiving $m_1 \in \{1, \dots, M_X\}$ *and* $m_2 \in \{1, \dots, M_Y\}$, we pick one of the M sequences in $\{\boldsymbol{X}_{m_1,j}\}_{j=1}^{M}$ that maximizes $p_{Y^n|X^n}(\hat{\boldsymbol{y}}|\boldsymbol{X}_{m_1,j})$ as the decoded message, where $\hat{\boldsymbol{y}}$ is the decoded message for $\boldsymbol{y}$ from m_2. Nemely,

$$d_X^{(n)}(\{\tilde{\boldsymbol{x}}_{i,j}\}_{i=1,\dots,M_X;j=1,\dots,M}) : (m_1, m_2) \mapsto \underset{\tilde{\boldsymbol{x}} \in \{\tilde{\boldsymbol{x}}_{m_1,j} | j=1,\dots,M\}}{\operatorname{argmax}} p_{Y^n|X^n}(d_Y^{(n)}(m_2)|\tilde{\boldsymbol{x}})$$

We denote the output by random variable $\hat{X}^{(n)}$. Conditioning on $\hat{Y}^{(n)} = Y^n$, it is clear that $\hat{X}^{(n)} = X^n$ as long as the encoder managed to find some $\boldsymbol{X}_{M_1,\tilde{M}_1}$ equal to X^n, and $p_{Y^n|X^n}(Y^n|X^n) > p_{Y^n|X^n}(Y^n|\boldsymbol{X}_{M_2,j})$ for all $j \neq \tilde{M}_1$. Combining this observation with b.)i.) and b.)ii.), we have, for n large enough,

$$\begin{aligned}
&P[\hat{X}^{(n)} \neq X^n|\hat{Y}^{(n)} = Y^n] \\
\leq &P\left[X^n \neq \boldsymbol{X}_{i,j}\ \forall(i,j)\right] + P\left[M_1, \tilde{M}_1 < \infty, p_{Y^n|X^n}(Y^n|\boldsymbol{X}_{M_1,j}) \geq p_{Y^n|X^n}(Y^n|X^n)\ \exists j \neq \tilde{M}_1\right] \\
< &3\epsilon
\end{aligned}$$

In summary, we have

$$P\left[\hat{X}^{(n)} \neq X^n \text{ or } \hat{Y}^{(n)} \neq Y^n\right] = P\left[\hat{Y}^{(n)} \neq Y^n\right] + P\left[\hat{X}^{(n)} \neq X^n \middle| \hat{Y}^{(n)} = Y^n\right] < 4\epsilon$$

for n large enough. Notice that ϵ can be picked to be arbitrarily small, we have shown (R_1, R_2) to be achievable. □

d.) Show that any (R_1, R_2) satisfying the following inequalities are achievable

$$R_1 > H(X|Y), \tag{10}$$

$$R_2 > H(Y|X), \tag{11}$$

$$R_1 + R_2 > H(X,Y). \tag{12}$$

Solution: *We sketch the proof here and omit the technical details.*
By symmetry, we know both $(H(X|Y) + \delta_1, H(Y) + \delta_2)$ and $(H(Y) + \delta_3, H(Y|X) + \delta_4)$ are achievable for all δ_1, δ_2, δ_3, $\delta_4 > 0$. By time multiplexing, all *rational* convex combinations of two achievable rates are achievable. Suppose (R_1, R_2) is a point in the domain described by (10), (11) and (12). These must exists some δ_1, δ_2, δ_3, $\delta_4 > 0$ such that (R_1, R_2) is some convex combination of $(H(X|Y) + \delta_1, H(Y) + \delta_2)$ and $(H(Y) + \delta_3, H(Y|X) + \delta_4)$. If such convex combination is rational, (R_1, R_2) is achievable. Otherwise, we claim that these must exist a point (R_1', R_2') as a rational covvex combination of $(H(X|Y) + \delta_1/2, H(Y) + \delta_2/2)$ and $(H(Y) + \delta_3/2, H(Y|X) + \delta_4/2)$ such that $R_1' < R_1$ and $R_2' < R_2$. Since (R_1', R_2') is achievable, so is (R_1, R_2). □