Problem 1: Suppose P is a probability measure defined on $S = \{1, 2, 3, 4, 5\}$ with mass function

s	$p(s)$
1	0.1
2	0.3
3	0.2
4	0.3
5	0.1

Compute the information content $I(s)$ of each sample point and the entropy $H(P)$.

We compute

s	$p(s)$	$I(s)$
1	0.1	3.322
2	0.3	1.737
3	0.2	2.322
4	0.3	1.737
5	0.1	3.322

where the surprisals are rounded to three places after the decimal point. The entropy is $H(P) \approx 2.171$.

Problem 2: Let $X \sim \mathcal{B}er(\theta)$ for $\theta \in [0, 1]$. Compute a formula for $H(X)$ in terms of θ.

We compute:

$$H(X) = -\sum_{x=0}^{1} p(x) \log_2(p(x)) = -(1-\theta)\log_2(1-\theta) - \theta \log_2(\theta).$$

Problem 3: Suppose P and Q are probability measures defined on $S = \{1, 2, 3, 4, 5\}$ with mass functions

s	$p(s)$	$q(s)$
1	0.1	0.05
2	0.3	0.15
3	0.2	0.7
4	0.3	0.03
5	0.1	0.07

Compute the cross entropies $H_P(Q)$ and $H_Q(P)$.

$$H_P(Q) = -\sum_{s=1}^{5} p(s)\log_2(q(s)) \approx 3.258, \quad H_Q(P) = -\sum_{s=1}^{5} q(s)\log_2(p(s)) \approx 2.337.$$

Problem 4: Suppose P and Q are probability measures defined on $S = \{1, 2, 3, 4, 5\}$ with mass functions

s	$p(s)$	$q(s)$
1	0.1	0.05
2	0.3	0.15
3	0.2	0.7
4	0.3	0.03
5	0.1	0.07

Compute $D(P \parallel Q)$ and $D(Q \parallel P)$.

$$D(P \parallel Q) = \sum_{s=1}^{5} p(s) \log_2 \left(\frac{p(s)}{q(s)} \right) \approx 1.087 \quad \text{and} \quad D(Q \parallel P) = \sum_{s=1}^{5} q(s) \log_2 \left(\frac{q(s)}{p(s)} \right) \approx 0.929.$$

Problem 5: While it is **not** true that $D(P \parallel Q) = D(Q \parallel P)$ for **all** P and Q, there are examples of special distributions P and Q for which equality does hold. Find examples.

For $\theta \in [0, 1]$, define P_θ and Q_θ on the sample space $S = \{0, 1\}$ by

s	$p_\theta(s)$	$q_\theta(s)$
0	θ	$1-\theta$
1	$1-\theta$	θ

One then easily proves $D(P_\theta \parallel Q_\theta) = D(Q_\theta \parallel P_\theta)$.

Problem 6: For each $\phi \in [0, 1]$, consider the communication channel with transition matrix

$$\mathbf{K} = \begin{bmatrix} 1-\phi & \phi \\ \phi & 1-\phi \end{bmatrix}$$

These are called *binary symmetric channels*. Suppose that $X \sim \mathcal{B}er(\alpha)$ for some $\alpha \in [0, 1]$, with the range of X enumerated as $x_0 = 0$ and $x_1 = 1$. Show that X and the communication channel determine a random variable Y with range $y_0 = 0$ and $y_1 = 1$. Determine its distribution.

Since the range of Y is $\{0, 1\}$, it must be Bernoulli, with $Y \sim \mathcal{B}er(\beta)$ for some $\beta \in [0, 1]$. We need to determine the parameter β. But notice that the probability vectors encoding the mass functions of X and Y have the form

$$\boldsymbol{\pi}(X)^\intercal = \begin{bmatrix} 1-\alpha & \alpha \end{bmatrix} \quad \text{and} \quad \boldsymbol{\pi}(Y)^\intercal = \begin{bmatrix} 1-\beta & \beta \end{bmatrix}.$$

So, if we conceptualize the entries in the transition matrix as the conditional probabilities, then by the Law of Total Probability we must have

$$\boldsymbol{\pi}(Y)^\intercal = \boldsymbol{\pi}(X)^\intercal \mathbf{K} = \begin{bmatrix} 1-\alpha & \alpha \end{bmatrix} \begin{bmatrix} 1-\phi & \phi \\ \phi & 1-\phi \end{bmatrix} = \begin{bmatrix} (1-\alpha)(1-\phi) + \alpha\phi & (1-\alpha)\phi + \alpha(1-\phi) \end{bmatrix}.$$

Thus, $\beta = \phi + \alpha - 2\phi\alpha$.

Problem 7: Suppose X and Y are random variables with range $\{0, 1\}$ and joint mass function given by

$p(x, y)$	$y = 0$	$y = 1$
$x = 0$	1/3	1/3
$x = 1$	0	1/3

(a) Compute the transition matrix of the communication channel induced from X and Y.

This just means that we need to compute the conditional mass function $p(y|x)$. We begin by computing the marginals:

x	$p(x)$
0	2/3
1	1/3

and

y	$p(y)$
0	1/3
1	2/3

Then

$p(y\|x)$	$y = 0$	$y = 1$
$x = 0$	1/2	1/2
$x = 1$	0	1

Thus, the transition matrix is given by

$$\mathbf{K} = \begin{bmatrix} 1/2 & 1/2 \\ 0 & 1 \end{bmatrix}$$

with the ranges of X and Y given by $x_0 = y_0 = 0$ and $x_1 = y_1 = 1$.

(b) Compute the mutual information $I(X, Y)$.

From the marginals in part (a), we get

$$H(X) = -(2/3)\log_2(2/3) - (1/3)\log_2(1/3) = -2/3 + \log_2(3)$$

and

$$H(Y) = -(1/3)\log_2(1/3) - (2/3)\log_2(2/3) = -2/3 + \log_2(3).$$

But

$$H(X, Y) = -\log_2(1/3) = \log_2(3),$$

and so

$$I(X, Y) = H(X) + H(Y) - H(X, Y) = -4/3 + \log_2(3) \approx 0.252.$$