Questions

1. Consider two discrete probability distributions $p(x)$ and $q(x)$ over the same set of four values $\{x\}$ of a random variable:

$p(x)$	1/8	1/8	1/4	1/2
$p(x)$	1/4	1/4	1/4	1/4

   (a) Calculate the cross-entropy $H(p, q)$ between $p(x)$ and $q(x)$.

$$\begin{aligned} H(p,q) &= \sum_x p(x) \cdot \lg \frac{1}{q(x)} \\ &= \frac{1}{8} \cdot \lg 4 + \frac{1}{8} \cdot \lg 4 + \frac{1}{4} \cdot \lg 4 + \frac{1}{2} \cdot \lg 4 \\ &= \frac{1}{4} \lg 2 + \frac{1}{4} \lg 2 + \frac{1}{2} \lg 2 + \lg 2 \\ &= 2 \lg 2 \end{aligned}$$

   (b) Calculate their Kullback-Leibler divergence $D_{\mathrm{KL}}(p \parallel q)$

$$\begin{aligned} D_{\mathrm{KL}}(p \parallel q) &= H(p,q) - H(p) \\ &= 2 \lg 2 - \sum_x p(x) \cdot \lg \frac{1}{p(x)} \\ &= 2 \lg 2 - \frac{1}{8} \lg 8 - \frac{1}{8} \lg 8 - \frac{1}{4} \lg 4 - \frac{1}{2} \lg 2 \\ &= 2 \lg 2 - \frac{3}{8} \lg 2 - \frac{3}{8} \lg 2 - \frac{1}{2} \lg 2 - \frac{1}{2} \lg 2 \\ &= \frac{1}{4} \lg 2 \end{aligned}$$

   (c) Comment on the use of the metrics $H(p, q)$ and $D_{\mathrm{KL}}(p \parallel q)$ in machine learning and for calculating the efficiency of codes.

   $H(p, q)$ is the metric often minimised in classification tasks. In this context p would be the true probabilities (or class) and q would be the predicted probabilities.

   $D_{\mathrm{KL}}(p \parallel q)$ is often maximised in the context of generative AI – for example in Autoencoders, where the model learns to minimise the KL-divergence between the latent space representation of images encoded and the standard multivariate normal distribution. KL-divergence also occurs as a loss function in other ML disciplines – such as Bayesian ML.

2. Continuous random variables X and Y both have uniform probability density distributions on some interval. For X, $p(x) = 1/2$ if $x \in [0, 2]$, otherwise 0, while for Y, $p(y) = 1/8$ if $y \in [0, 8]$, otherwise 0. Calculate the differential entropies $h(X)$ and $h(Y)$.

$$\begin{aligned} h(X) &= \int_0^2 p(x) \cdot \ln \frac{1}{p(x)} \mathrm{d}x \\ &= \int_0^2 \frac{1}{2} \cdot \ln \frac{1}{2} \mathrm{d}x \\ &= \left[x \cdot \frac{1}{2} \cdot \ln 2 \right]_0^2 \\ &= \ln 2 \end{aligned}$$

$$
\begin{aligned}
h(Y) &= \int_0^8 p(y) \cdot \ln \frac{1}{p(y)} \mathrm{d}x \\
&= \int_0^8 \frac{3}{8} \cdot \ln 2 \mathrm{d}x \\
&= \left[x \cdot \frac{3}{8} \cdot \ln 2 \right]_0^8 \\
&= 3 \cdot \ln 2
\end{aligned}
$$

3. (a) Find the optimal input distribution for the Binary Symmetric Channel and show that its capacity is given by $C = 1 - H_2(f)$.

The optimal input distribution for the Binary Symmetric Channel is the distribution which maximises the mutual information $I(X;Y)$ between the output and the input. We can find this by differentiating the expression for the mutual information $I(X;Y)$ with respect to p.

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X \mid Y) \\
&= H_2(X) - H_2(f) \\
\frac{\partial I(X;Y)}{\partial p} &= \frac{\partial H_2(X)}{\partial p}
\end{aligned}
$$

Thus, the optimal input distribution is that which maximises the entropy of the input. We know that entropy is maximised when events are equiprobable. Thus, the optimal input distribution occurs when $p = 1 - p = \frac{1}{2}$.

Substituting this back into the original expression gives:

$$C = \max_p I(X;Y) = \max_p H_2(p) - H_2(f) = \frac{1}{2} \lg 2 + \frac{1}{2} \lg 2 - H_2(f) = 1 - H_2(f)$$

(b) Consider the Z channel:)

$$
\begin{aligned}
P(y = 0 \mid x = 0) &= 1 \\
P(y = 1 \mid x = 0) &= 0 \\
P(y = 0 \mid x = 1) &= f \\
P(y = 1 \mid x = 1) &= 1 - f
\end{aligned}
$$

Show that the mutual information given by (p_1 is $P(X = 1)$ for source X):

$$I(X;Y) = H_2(p_1(1 - f)) - p_1 H_2(f)$$

$$
\begin{aligned}
H(Y \mid X) &= \sum_x p(x) \sum_y h(p(y \mid x)) \\
&= p_1 \cdot (h(f) + h(1 - f)) + p_2 \cdot (h(1) + h(0)) \\
&= p_1 H_2(f) + p_2 0 \\
&= p_1 H_2(f)
\end{aligned}
$$

$$
\begin{aligned}
H(Y) &= \sum_y p(y) \lg \frac{1}{p(y)} \\
&= \sum_y \sum_x p(x,y) \lg \frac{1}{\sum_x p(x,y)} \\
&= (p_1 f + 1 - p_1) \lg \frac{1}{p_1 f + 1 - p_1} + (p_1(1-f)) \lg \frac{1}{p_1(1-f)} \\
&= (1 - p_1(1-f)) \lg \frac{1}{1 - p_1(1-f)} + (p_1(1-f)) \lg \frac{1}{p_1(1-f)} \\
&= H_2(p_1(1-f))
\end{aligned}
$$

$$
\begin{aligned}
I(X;Y) &= H(Y) - H(Y \mid X) \\
&= H_2(p_1(1-f)) - p_1 H_2(f)
\end{aligned}
$$

Hence show that the optimal input distribution is given by

$$
p_1^\star = \frac{1/(1-f)}{1 + 2^{H_2(f)/(1-f)}}
$$

We can find the optimal input distribution by differentiating the mutual information with respect to p_1 and finding the maxima.

$$
\begin{aligned}
I(X;Y) &= H_2(p_1(1-f)) - p_1 H_2(f) \\
\frac{\partial I(X;Y)}{\partial p_1} &= \frac{\partial}{\partial p_1(1-f)}(H_2(p_1(1-f))) \cdot \frac{\partial}{\partial p_1}(p_1(1-f)) - H_2(f) \\
&= -\lg \frac{p_1(1-f)}{1 - p_1(1-f)} \cdot (1-f) - H_2(f) \\
0 &= -(1-f) \lg \frac{p_1^\star(1-f)}{1 - p_1^\star(1-f)} - H_2(f) \\
2^{\frac{H_2(f)}{1-f}} &= \frac{1 - p_1^\star(1-f)}{p_1^\star(1-f)} \\
p_1^\star(1-f) &= \frac{1}{1 + 2^{\frac{H_2(f)}{1-f}}} \\
p_1^\star &= \frac{1/(1-f)}{1 + 2^{H_2(f)/(1-f)}}
\end{aligned}
$$

(c) Plot a graph of p_1^* and investigate what happens as the noise level f becomes close to 1. Observe that $p_1^* < 0.5$, for all values of f.

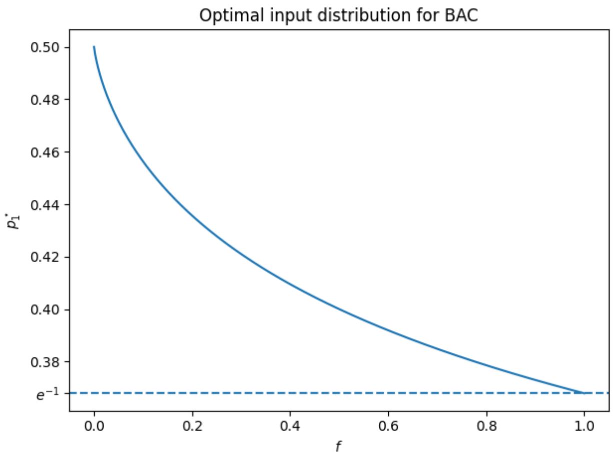

Figure 1: A graph of the optimal input distribution $p_1^\star$ of a BAC against the error probability

Notice that $\lim_{f\to 1} p_1^\star = e^{-1}$.

(d) Sketch graphs of the capacity of the binary symmetric channel and the Z channel as a function of f.

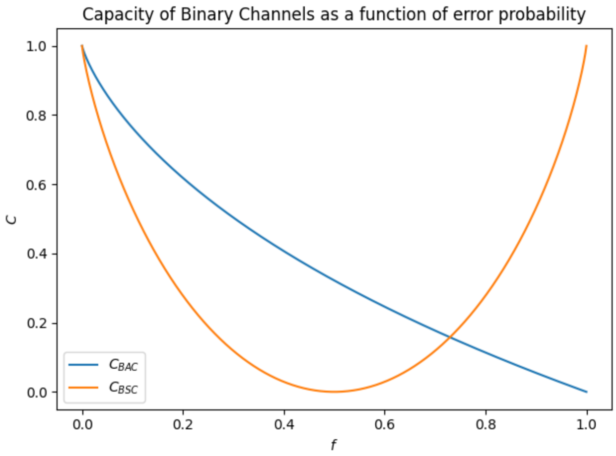

Figure 2: A graph of the capacity of BSC's and BAC's against the error probability

4. A continuous communication channel adds Gaussian white noise to signals transmitted through it. The ratio of signal power to noise power is 30 decibels, and the frequency bandwidth of this channel is 10MHz. roughly what is the information capacity C of this channel, in bits/second?

$$\begin{aligned} C &= B \cdot \lg\left(1 + \frac{S}{N}\right) \\ &= 10 \times 10^6 \cdot \lg(1 + 30) \\ &= 49.5\ldots \times 10^6 \end{aligned}$$

5. (a) In lectures, you have seen correlation. related concept, the autocorrelation of a signal (correlation of a signal with itself after some time lag) is defined as

$$R_{yy}(\ell) = \sum_{n\in\mathbb{Z}} y(n)\overline{y(n-\ell)}$$

The following CSV file contains a sine wave ($\sin(2\pi f_0 t_i)$) embedded in additivee white Gaussian Noise: `http://ireland.cx/teaching/2122/infoth/autocorrelation.txt`. The sampling frequency is 2 MHz.

Plot the autocorrelation of the signal in the CSV file ($R_{yy}(\ell)$ against ℓ). Hence find the frequency f_0 of the sine wave.

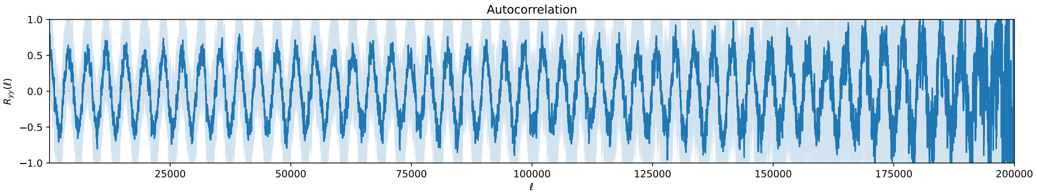

Figure 3: Graph of ℓ against $R_{yy}(\ell)$

As we see in Figure 3, the autocorrelation mirrors a wave. When ℓ is close to the true offset, we have that the autocorrelation is high. When it is half the period, we have low autocorrelation.

Note that the true autocorrelation was too noisy. So, I plot the mean over 50 adjacent values on either side (and `fill_between` standard deviation). From this. Observe that the autocorrelation "cycles" every 4000 readings. This implies that the frequency of the sine wave f_0 is roughly $\frac{1}{4000} \cdot 2000000 \approx 500Hz$.

(b) Research and explain how GPS works at the physical and data-link layers. Explain how GPS implements the spreading of the spectrum of the transmitted signal over a very large bandwidth, and how the wide-bandwidth signal is decoded in a GPS receiver.

Do this!

6. Explain the concept of entropy in:

(a) Classical thermodynamics ("entropy", developed by Clausius and others)

Classical Entropy is a measure of disorder and represents the energy of a system which is *not* available for work.

(b) Statistical mechanics ("statistical entropy", developed by Boltzmann and others – presented in lecture 12)

A measure of the number of possible states that a system in thermodynamic equilibrium could be in.

(c) Information theeory ("information entropy", developed by Shannon)

The average amound of information required to completely describe a state.

To what extent do these different notions of entropy relate to each other?

These notions of entropy are similar at an intuitive level. They are all measures of how much we *don't* know about a state. This means they're *related*.