Example Sheet 1 Solutions: Classical Information Theory and Linear Operators

Classical information theory

Exercise 1. The Typical Sequence Theorem is given below. Prove part (a), and prove part (c) using part (b).

Theorem 1. (Typical Sequence Theorem) *Fix $\varepsilon \in (0,1)$. Then for any $\delta > 0$ there exists an integer $n_0(\delta) > 0$, such that $\forall\ n \geq n_0(\delta)$, the following are true:*

a) If $(u_1, \ldots, u_n) \in T_\varepsilon^{(n)}$, then

$$H(U) - \varepsilon \leq -\frac{1}{n} \log p(u_1, \ldots, u_n) \leq H(U) + \varepsilon.$$

b) $\mathbf{P}\{T_\varepsilon^{(n)}\} > 1 - \delta$.

c) $(1-\delta)2^{n(H(U)-\varepsilon)} < |T_\varepsilon^{(n)}| \leq 2^{n(H(U)+\varepsilon)}$.

Solution

The definition of $T_\varepsilon^{(n)}$ is that

$$2^{-n(H(U)+\varepsilon)} \leq p(u_1, \ldots, u_n) \leq 2^{-n(H(U)-\varepsilon)}.$$

Therefore, taking the logarithm and dividing by $-n$ yields the (a).

Each typical sequence has probability at least $2^{-n(H(U)+\varepsilon)}$ by definition. Therefore,

$$1 \geq \mathbf{P}\{T_\varepsilon^{(n)}\} \geq |T_\varepsilon^{(n)}| 2^{-n(H(U)+\varepsilon)}$$

so $|T_\varepsilon^{(n)}| \leq 2^{n(H(U)+\varepsilon)}$.

On the other hand, each typical sequence has probability at most $2^{-n(H(U)-\varepsilon)}$, so by (b)

$$1 - \delta < \mathbf{P}\{T_\varepsilon^{(n)}\} \leq |T_\varepsilon^{(n)}| 2^{-n(H(U)-\varepsilon)}$$

yielding $2^{-n(H(U)-\varepsilon)}(1-\delta) < |T_\varepsilon^{(n)}|$.

Exercise 2. (Typical set versus set of high probability.)

Please feel free to use a calculator or computer to aid in computation for this exercise. Consider a binary source described by random variables U_1, U_2, U_3 on the alphabet $\{0,1\}$ with common probability mass function given by $p(0) = 0.4$, $p(1) = 0.6$.

1. What is the most probable sequence in $\{0,1\}^3$ emitted by this source?
2. What is the set of typical sequences $T_\varepsilon^{(3)}$, for $\varepsilon = 0.2$?
3. What is the total probability of sequences in the typical set, $\mathbf{P}\{T_{0.2}^{(3)}\}$?
4. What is the smallest set of sequences in $\{0,1\}^3$ with total probability at least $\mathbf{P}\{T_{0.2}^{(3)}\}$?
5. What does this tell you about the typical set? Which is more useful for compression?

Solution

There are 8 sequences in $\{0,1\}^3$. Setting $p := p(1) > q := p(0)$, these may be arranged by probability

	111		$p^3 = 0.216$
011	101	110	$qp^2 = 0.144$
001	010	100	$q^2p = 0.096$
	000		$q^3 = 0.064$

1. Most probable: 111
2. The entropy is $H(U) \approx 0.971$. A sequence is typical for $\varepsilon = 0.2$ if it has probability between $2^{-n(H(U)+\varepsilon)} \approx 0.088$ and $2^{-n(H(U)-\varepsilon)} \approx 0.201$. That's the six sequences in the middle two rows of the table.
3. The total probability of the set of typical sequences is $1 - p^3 - q^3 = 3 * qp^2 3 * q^2p = 0.72$.

4. To find the smallest set of sequences with total probability at least 0.72, we start with the most probable and add sequences until we reach total probability 0.72. The top two rows give us 0.648, so we need to add one sequence from the third row as well, giving total probability 0.744. So, $\{111, 011, 101, 110, 001\}$ for example.

5. It tells us that there can be smaller sets than the typical set with still at least much total probability.

   On the one hand, such sets are useful for compression because they are of small size but have large total probability. Then a compression scheme like the one shown for source coding could be performed, in which an error is made on (and only on) sequences outside of this set. In this case, since the set is smaller and has higher total probability, this would yield a lower probability of error, and more compression (i.e. to a smaller space).

   However, suppose we want to find the smallest set with total probability at least $1 - \delta$, for some $\delta > 0$. We don't have a simple criteria to decide if a given sequence should be in this set or not[1] (which could make such a compression scheme impractical), and we don't know a good way to count how many sequences are necessary to form this set. The typical set has a simple criteria to check if a sequence should be included or not, as well as useful bounds on its size. Moreover, we know that for any $\delta > 0$, we can always choose n large enough such that the total probability of the typical set is at least $1 - \delta$.

   So this last question was a bit of a setup (which I only wrote in such a way because these exercises aren't graded): forgetting other concerns, the smallest set with total probability at least $1 - \delta$ (which McKay calls the "δ-sufficient set") does seem better for compression in terms of the probability of error and size of the resulting space. However, the typical set has other very useful properties, and serves us well in our proofs, leaving the δ-sufficient set simply as an interesting consideration rather than a fundamental tool in source compression.

Exercise 3. Given random variables X, Y with probability mass functions $p(x)$ and $p(y)$, prove the following:

[1] Imagine trying to make a table like the one above when n is large– the number of entries is 2^n.

1. $H(X, Y) \leq H(X) + H(Y)$, with equality if and only if X, Y are independent. This property is called **subadditivity**. It implies that $H(Y) \geq H(Y|X)$, i.e., *conditioning reduces the entropy.*

2. The Shannon entropy is **concave**. If $p = \{p(x)\}$ and $q = \{q(x)\}$ are two probability distributions and $1 > \lambda > 0$, then $H(\lambda p + (1 - \lambda)q) \geq \lambda H(p) + (1 - \lambda)H(q)$.

Solution

The easiest way to prove (a) might be to make use of the fact that the mutual information can be expressed in terms of the relative entropy:

$$I(X : Y) = D(\{p(x, y)\}_{x,y} \| \{p(x)p(y)\}_{x,y}) \tag{1}$$

and the fact that the relative entropy is non-negative (as proven in the lecture). Then by the definition of the mutual information,

$$H(X) + H(Y) - H(X, Y) = I(X : Y) \geq 0$$

yielding the result. As we will prove in the next exercise, the relative entropy is zero if and only if the two distributions are equal. That is, $I(X : Y) = 0$ if and only if $p(x, y) = p(x)p(y)$ for each x and y, i.e. if and only if X and Y are independent.

We may prove (1) by expanding the definition

$$\begin{aligned}
D(\{p(x, y)\}_{x,y} \| \{p(x)p(y)\}_{x,y}) &= \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\
&= \sum_{x,y} p(x, y) \log p(x, y) - \sum_{x,y} p(x, y) \log[p(x)p(y)] \\
&= -H(X, Y) - \sum_{x,y} p(x, y) \log p(x) - \sum_{x,y} p(x, y) \log p(y) \\
&= -H(X, Y) - \sum_{x} p(x) \log p(x) - \sum_{y} p(y) \log p(y) \\
&= -H(X, Y) + H(X) + H(Y).
\end{aligned}$$

To prove (b), we will use the concavity of the function $\eta(x) = -x \log x$. We can check this by differentiation:

$$\eta'(x) = -\log x - \frac{1}{\log_e 2}$$
$$\eta''(x) = -\frac{1}{x}\frac{1}{\log_e 2} < 0$$

for $x > 0$, so η is concave. Then,

$$\begin{aligned} H(\lambda p + (1-\lambda)q) &= \sum_x \eta(\lambda p(x) + (1-\lambda)q(x)) \\ &\geq \sum_x \lambda\eta(p(x)) + (1-\lambda)\eta(q(x)) \\ &= \lambda H(p) + (1-\lambda)H(q). \end{aligned}$$

Exercise 4. In the lectures we proved that

- For two probability distributions $p = \{p(x)\}$ and $q = \{q(x)\}$,
$$D(p||q) \geq 0$$
- For a random variable $X \sim p(x)$, $x \in J_X$ (where J_X is a finite alphabet)
$$0 \leq H(X) \leq \log|J_X|$$

When do the equalities in (4) and (4) hold?

Solution

By inspecting the proof of $D(p\|q) \geq 0$, we see the only inequality is Jensen's inequality. Briefly,

$$-D(p\|q) = \sum_x p(x) \log \frac{q(x)}{p(x)} = \mathbb{E}_p[\log \frac{q(X)}{p(X)}]$$

where $p(X)$ is the r.v. that takes on value $p(x)$ when $X = x$, and similarly for q, and the $\mathbb{E}_p$ means the expectation value with respect to the probability distribution p. Since log is strictly concave,

$$\begin{aligned} D(p\|q) &= -\mathbb{E}_p[\log \frac{q(X)}{p(X)}] \\ &\geq \log \mathbb{E}_p[\frac{q(X)}{p(X)}] \\ &= \log \sum_x p(x)\frac{q(x)}{p(x)} \\ &= \log \sum_x q(x) = \log 1 = 0. \end{aligned}$$

Moreover, since log is nonlinear, equality in Jensen's inequality holds if and only if we have $\frac{q(x)}{p(x)} = c$ is constant, independent of x. That is, $q(x) = cp(x)$. Summing over x yields

$$1 = \sum_x q(x) = c\sum_x p(x) = c$$

and therefore $c = 1$. Thus, we have $D(p\|q) = 0$ if and only if the two distributions are the same: $q(x) = p(x)$ for all x.

Now, let us consider equality in $0 \leq H(X) \leq \log|J_X|$. We have that

$$H(X) = \sum_x p(x) \log \frac{1}{p(x)}.$$

Each term is non-negative, so to have $H(X) = 0$, each term must be zero. That is, for each x either $p(x) = 0$ or $\log \frac{1}{p(x)} = 0$. Since only $\log 1 = 0$, we must have either $p(x) = 0$ or $p(x) = 1$ for each x. That is, X is deterministic.

For the other inequality, we may expand

$$\begin{aligned} D\left(p\|\{\tfrac{1}{|J_X|}\}\right) &= \sum_x p(x)[\log p(x) - \log \tfrac{1}{|J_X|}] \\ &= -H(X) + \log \frac{1}{|J_X|} = \log|J_X| - H(X). \end{aligned}$$

Since $D\left(p\|\{\frac{1}{|J_X|}\}\right) = 0$ if and only if $p(x) = \frac{1}{|J_X|}$ for each x, we find $H(X) = \log|J_X|$ only for the uniform distribution.

Exercise 5. Given two discrete random variable $X \sim p_X(x)$, $x \in J$, $Y \sim p_Y(y)$, $x \in J$, with joint distribution $\{p_{X,Y}(x,y)\}_{x,y\in J}$, express the mutual information $I(X:Y)$ and the conditional entropy $H(Y|X)$ in terms of the relative entropy.

Solution

We actually already proved the first part below equation (1) in the solution to Exercise 3.

Defining the function $1(y) = 1$, we expand

$$\begin{aligned} D(\{p(x,y)\}_{x,y}\|\{p(x)1(y)\}_{x,y}) &= \sum_{x,y} p(x,y)[\log p(x,y) - \log p(x)] \\ &= -H(X,Y) - \sum_{x,y} p(x,y)\log p(x) \\ &= -H(X,Y) - \sum_{x} p(x)\log p(x) \\ &= -H(X,Y) + H(X). \end{aligned}$$

Therefore,

$$H(Y|X) := H(X,Y) - H(X) = -D(\{p(x,y)\}_{x,y}\|\{p(x)1(y)\}_{x,y})$$

Exercise 6. Consider two random variables X and Y with joint probability mass function $p(x,y)$, for $x, y \in J$.

1. Prove that their mutual information $I(X:Y)$ can be expressed as
$$I(X:Y) = H(X) - H(X|Y).$$
2. Show if X and Y are independent, $I(X:Y) = 0$.

Solution

1. The right-hand side is
$$H(X) - H(X|Y) = H(X) - H(X,Y) + H(Y) = I(X:Y).$$

2. Since the mutual information is symmetric, we therefore also have
$$I(X:Y) = H(Y) - H(Y|X).$$
In Exercise 3, we showed that $H(Y) = H(Y|X)$ if and only if X, Y are independent, so the mutual information is zero if and only if X, Y are independent.

Exercise 7. Prove the following:

1. If two random variables X and Y are equal, then their mutual information is equal to the Shannon entropy of X or Y.
2. If X is a uniformly random bit (i.e., $X \sim p(x)$, $x \in \{0,1\}$ with $p(0) = 1/2 = p(1)$) and Y is a random variable defined as follows:
$$\begin{aligned} Y &= X \quad \text{with probability } p \\ Y &= 1 - X \quad \text{with probability } (1-p), \end{aligned}$$
for some $p \geq 1/2$, then
$$I(X:Y) = 1 - h(p),$$
where $h(p) = -p\log p - (1-p)\log(1-p)$ (binary entropy).

Solution

1. Let $X = Y \sim p$. Then their joint distribution is $p(x,y) = \delta_{x,y}p(x)$ where $\delta_{x,y} = 0$ if $x \neq y$ and one otherwise. Then
$$H(X,Y) = -\sum_x \sum_y p(x,y)\log p(x,y) = -\sum_x p(x)\log p(x) = H(X).$$
So
$$I(X:Y) = H(X) + H(Y) - H(X,Y) = H(Y) = H(X).$$
2. We use Exercise 6 to write
$$I(X:Y) = H(Y) - H(Y|X).$$

Let us denote $p(x, y)$ for the joint distribution of X and Y, and $p(y|x)$ as the conditional distribution of Y given X. Let us consider $p(y|x)$.

$$\mathbf{P}(Y = 0|X = 0) = p, \qquad \mathbf{P}(Y = 0|X = 1) = 1 - p,$$
$$\mathbf{P}(Y = 1|X = 0) = 1 - p, \qquad \mathbf{P}(Y = 1|X = 1) = p.$$

Then

$$p(x, y) = p(x)p(y|x) = \frac{1}{2}p(y|x).$$

Thus,

$$H(Y|X) = -\sum_{x,y} p(x, y) \log p(y|x) = -\frac{1}{2}\sum_{x,y} p(y|x) \log p(y|x).$$

We see the four values of $p(y|x)$ are $p, p, 1 - p, 1 - p$. So we have

$$H(Y|X) = -p \log(p) - (1 - p) \log(1 - p) = h(p).$$

Next, we have

$$p(y) = p(0, y) + p(1, y) = \frac{1}{2}(p(y|0) + p(y|1)).$$

So in either case $y = 0$ or $y = 1$, we are summing p and $1 - p$, so we have $\mathbf{P}(Y = y) = \frac{1}{2}$ for each $y = 0, 1$. Since Y has a uniform marginal distribution, we have $H(Y) = \log 2 = 1$.

Putting it together,

$$I(X : Y) = H(Y) - H(Y|X) = 1 - h(p).$$

Exercise 8. In the lectures it was mentioned that the Shannon entropy $H(X)$ of a random variable X is a measure of its uncertainty. This would imply that if X is nearly a constant then $H(X)$ is very small. Prove the rigorous statement of this claim, which reads as follows:

Suppose a random variable X takes $m \geq 2$ values and one of these values has a probability $(1 - \varepsilon)$, then

$$H(X) \leq h(\varepsilon) + \varepsilon \log(m - 1),$$

where $h(\varepsilon)$ denotes the binary entropy. This is known as Fano's inequality.

Solution

Let x_1 be the value with probability $p_1 \geq 1 - \varepsilon$. Let us define again $\eta(x) = -x \log x$. Then we can write

$$\begin{aligned} H(X) &= -\sum_{i=1}^{m} p_i \log p_i \\ &= \eta(p_1) + \eta(1 - p_1) - \eta(1 - p_1) + \sum_{i=2}^{m} \eta(p_i) \end{aligned}$$

Then $h(p_1) = \eta(p_1) + \eta(1 - p_1)$. Additionally, $1 - p_1 = \sum_{i=2}^{m} p_i$. So,

$$\begin{aligned} &= h(p_1) + (1 - p_1) \log(1 - p_1) - \sum_{i=2}^{m} p_i \log(p_i) \\ &= h(p_1) + \sum_{i=2}^{m} p_i \log(1 - p_1) - \sum_{i=2}^{m} p_i \log(p_i) \\ &= h(p_1) - \sum_{i=2}^{m} p_i \log \frac{p_i}{1 - p_1} \\ &= h(p_1) - (1 - p_1) \sum_{i=2}^{m} \frac{p_i}{1 - p_1} \log \frac{p_i}{1 - p_1} \\ &= h(p_1) + (1 - p_1) H\left(\{\frac{p_i}{1 - p_1}\}_{i=2}^{m} \right) \\ &\leq h(\varepsilon) + \varepsilon \log(m - 1) \end{aligned}$$

since $h(p_1) = h(1 - p_1) = h(\varepsilon)$, and $H\left(\{\frac{p_i}{1-p_1}\}_{i=2}^{m} \right) \leq \log(m - 1)$ since $\{\frac{p_i}{1-p_1}\}_{i=2}^{m}$ is a probability distribution.

Exercise 9. The interpretation of entropy as a measure of uncertainty would also imply that if a random variable X is nearly a function of another random variable Y, then the entropy of $X|Y$ is very small. A rigorous statement of this claim is given by the generalized Fano inequality:

For a pair of random variables X and Y, if we can rearrange the values so that they pair up with $x_1, \ldots, x_m$ and $y_1, \ldots, y_m$, such that

$$\sum_{j=1}^{m} P(X = x_j, Y = y_j) = 1 - \epsilon,$$

then $\quad H(X|Y) \leq h(\varepsilon) + \varepsilon \log(m-1)$.

Prove (9). {*Hint: Use the result of Ex. 4 and the concavity of the entropy.*}

Solution

We have that

$$\sum_{j=1}^{m} P(X = x_j, Y = y_j) = 1 - \epsilon.$$

Then

$$\begin{aligned}
\varepsilon &= 1 - \sum_{j=1}^{m} P(X = x_j, Y = y_j) \\
&= 1 - \sum_{j=1}^{m} P(X = x_j | Y = y_j) P(Y = y_j) \\
&= \sum_{j=1}^{m} (1 - P(X = x_j | Y = y_j)) P(Y = y_j) \\
&= \sum_{j=1}^{m} P(X \neq x_j | Y = y_j) P(Y = y_j) \\
&= \sum_{j=1}^{m} P(X \neq x_j, Y = y_j)
\end{aligned}$$

Define $\varepsilon_j = P(X \neq x_j | Y = y_j)$. Then $\sum_j \varepsilon_j P(Y = y_j) = \varepsilon$. Additionally, $\Pr(X = x_j | Y = y_j) = 1 - \varepsilon_j$. So we can apply Fano's to the distribution $\Pr(X | Y = y_j)$:

$$H(X|Y = y_j) \leq h(\varepsilon_j) + \varepsilon_j \log(m-1).$$

Then

$$\begin{aligned}
H(X|Y) &= \sum_j p(y_j) H(X|Y = y_j) \\
&\leq \sum_j p(y_j) h(\varepsilon_j) + \sum_j p(y_j) \varepsilon_j \log(m-1) \\
&\leq h\left(\sum_j p(y_j) \varepsilon_j \right) + \varepsilon \log(m-1) \\
&= h(\varepsilon) + \varepsilon \log(m-1)
\end{aligned}$$

using $\sum_j p(y_j)\varepsilon_j = \varepsilon$, and the concavity of the binary entropy.

Exercise 10. An important inequality satisfied by the mutual information is the so-called data-processing inequality. We will also study its quantum analogue later. It states that one can never retrieve lost information by any manipulation.

To state the inequality we need to consider a Markov chain: it is a sequence $X_1 \to X_2 \to \cdots$ of random variables such that X_{n+1} is independent of $X_1, \cdots X_{n-1}$, given X_n. That is,

$$P(X_{n+1} = x_{n+1} | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_{n+1} = x_{n+1} | X_n = x_n).$$

Then the data processing inequality is stated as follows:

Data-processing inequality: If $X \to Y \to Z$ then

$$I(X : Y) \geq I(X : Z)$$

Prove the above inequality.

Solution

We have

$$P(Z = z | X = x, Y = y) = P(Z = z | Y = y).$$

Then

$$\begin{aligned} H(Z|XY) &= \sum_{xy} p(x,y) H(Z|X=x, Y=y) \\ &= \sum_{xy} p(x,y) H(Z|Y=y) \\ &= \sum_{y} p(y) H(Z|Y=y) \\ &= H(Z|Y). \end{aligned}$$

That is,

$$\begin{aligned} H(ZXY) - H(XY) &= H(ZY) - H(Y) \\ H(XY) - H(Y) &= H(ZXY) - H(ZY) \\ H(X|Y) &= H(X|ZY). \end{aligned}$$

So,

$$\begin{aligned} I(X:Y) &= H(X) - H(X|Y) \\ I(X:Z) &= H(X) - H(X|Z) \\ I(X:Y) - I(X:Z) &= H(X|Z) - H(X|Y) \\ &= H(X|Z) - H(X|ZY). \end{aligned}$$

It remains to show that the quantity

$$H(X|Z) - H(X|ZY) \geq 0.$$

This quantity is also called the conditional mutual information, $I(X:Y|Z)$, and is in fact non-negative for any X, Y, Z (without any Markov chain assumption)[2]. We can see this as follows:

$$\begin{aligned} H(X|Z) - H(X|ZY) &= H(XZ) - H(Z) - H(XYZ) + H(ZY) \\ &= \sum_{x,y,z} p(x,y,z) \log \frac{p(z)p(x,y,z)}{p(x,z)p(z,y)} \end{aligned}$$

[2]The intuition is again that conditioning reduces entropy.

Grouping all the logarithms,

$$\begin{aligned} &= \sum_{x,y,z} p(x,y,z) \log \frac{p(x|y,z)}{p(x|z)} \\ &= \sum_{x,y,z} p(y,z) p(x|y,z) \log \frac{p(x|y,z)}{p(x|z)} \\ &= \sum_{y,z} p(y,z) \sum_{x} p(x|y,z) \log \frac{p(x|y,z)}{p(x|z)} \\ &= \sum_{y,z} p(y,z) D\left(\{p(x|y,z)\}_x \,\|\, \{p(x|z\}_x\right) \\ &\geq 0 \end{aligned}$$

as an average of non-negative quantities. We used that for each fixed y and z, both $\{p(x|y,z)\}_x$ and $\{p(x|z\}_x$ are probability distributions over the alphabet for X, since each element is non-negative and they sum to 1.

Exercise 11. A binary erasure channel has binary input alphabet $I = \{0, 1\}$. With probability p a bit is erased and with probability $(1-p)$ it remains unchanged. The channel has two inputs 0 and 1 and three outputs 0, 1 and e (the latter denoting an erased bit). Calculate the capacity of this channel for $p = 1/3$.

Solution

The capacity is defined as

$$C = \max_{p(x)} I(X:Y) = \max_{p(x)} H(Y) - H(Y|X)$$

where X is the input to the channel and Y is the output.

We can work out the channel matrix as

$$p(y|x) = \begin{pmatrix} 1-p & p & 0 \\ 0 & p & 1-p \end{pmatrix}$$

where the columns are 0, e, 1, and the rows are 0, 1. Note the rows are permutations of each other.

Then

$$\begin{aligned} H(Y|X) &= \sum_{x=0}^{1} p(x) H(Y|X=x) \\ &= \sum_{x} p(x) \underbrace{(-1) \sum_{y\in\{0,1,e\}} p(y|x)\log p(y|x)}_{\eta(p)+\eta(1-p)+\eta(0)=h(p)} \\ &= h(p). \end{aligned}$$

Since $H(Y) \leq \log 3$, we have $C \leq 3 - h(p)$.

Can we achieve this value? We need Y to be uniformly distributed. So we need

$$\begin{aligned} P(Y=1) &= P(Y=1|X=0)P(X=0) + P(Y=1|X=1)P(X=1) = 0 + (1-p)P(X=1) \\ &= \frac{2}{3}P(X=1) \\ P(Y=e) &= P(Y=e|X=0)P(X=0) + P(Y=e|X=1)P(X=1) = pP(X=0) + pP(X=1) \\ &= \frac{1}{3} \\ P(Y=0) &= P(Y=0|X=0)P(X=0) + P(Y=0|X=1)P(X=1) = (1-p)P(X=0) + 0 \\ &= \frac{2}{3}P(X=0). \end{aligned}$$

to each be $\frac{1}{3}$ in order to have a uniform distribution on Y. We see choosing $P(X=1) = P(X=0) = \frac{1}{2}$ yields this. Thus, the capacity is $\log 3 - h(\frac{1}{3}) = \frac{2}{3}$.

Exercise 12. Find the capacity of the following memoryless channel, where an additive noise Z takes values 0 and a with probability $1/2$, a is a given real number. The input alphabet is $\{0,1\}$ and Z is independent of X. How does the capacity depend on a? [*Hint*: Use the operational definition of the channel capacity]

Solution

Note if $a = 0$, the channel is noiseless and has capacity 1. So let us consider $a \neq 0$.

We input a binary r.v. X, and get out $Y = X + Z$. The output alphabet is $J_Y = \{0, a, 1, 1+a\}$. If $a \neq 1$ and $a \neq -1$, then the four outputs of the channel are distinct. Given output 0 or a, we know the input was 0. Given output 1 or $1+a$, we know the input was 1.

Therefore, to send 1 bit through the channel, we can encode it by exactly 1 bit, so the capacity (the max number of bits of messaged transmitted per use of the channel) is 1.

Let us consider the case $a = -1$, then if we send 0, we get either 0 or -1 with probability $\frac{1}{2}$. If we send 1, we get either 1 or 0 with probability $\frac{1}{2}$.

The channel matrix is

$$p(y|x) = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

where the columns are $0, 1, -1$ and the rows are $0, 1$. The rows are permutation invariant, as before we have $H(Y|X) = h(\frac{1}{2}) = 1$.

Thus, the capacity is

$$C = \max_{p(x)} H(Y) - H(Y|X) = \max_{p(x)} H(Y) - 1.$$

Then like before, we can calculate the distribution of Y.

$$\begin{aligned} P(Y=0) &= \frac{1}{2}P(X=0) + \frac{1}{2}P(X=1) = \frac{1}{2} \\ P(Y=-1) &= \frac{1}{2}P(X=0) \\ P(Y=1) &= \frac{1}{2}P(X=1). \end{aligned}$$

So

$$H(Y) = \eta(\frac{1}{2}) + \eta(\frac{1}{2}P(X=0)) + \eta(\frac{1}{2}(1 - P(X=1))).$$

we can check that a uniform distribution has the most uncertainty here. So $H(Y) = \eta(\frac{1}{2}) + 2\eta(\frac{1}{4}) = \frac{3}{2}$.

In this case then, the capacity is $C = \frac{3}{2} - 1 = \frac{1}{2}$. The case $a = 1$ is similar.

Vector spaces and linear operators

Note: the question below only specified that V was a vector space. However, in class we defined the adjoint $\dagger$ with respect to an inner product. Therefore, we'll take V to be a Hilbert space instead.

Exercise 13. Hilbert Schmidt inner product (adapted from *Nielsen and Chuang*) Let V be a finite–dimensional Hilbert space and L_V be the set of all linear operators acting on V. It is easy to prove that L_V is a vector space.

1. Prove that the function $(\cdot,\cdot)$ on $L_V \times L_V$ defined by

$$(A, B) = \mathrm{Tr}(A^\dagger B),$$

   defines an inner product. This is known as the *Hilbert Schmidt inner product*. Equipping L_V with this inner product, makes it a Hilbert Space.

2. Using the *outer product representation for operators*[3] (or otherwise), show that if V has dimension d then L_V has dimension d^2.

3. Write the Cauchy-Schwarz inequality for the Hilbert Schmidt inner product.

Solution

1. Inner product:

   (a) $(B, A) = \mathrm{tr}[B^\dagger A] = \overline{\mathrm{tr}[(B^\dagger A)^\dagger]} = \overline{\mathrm{tr}[A^\dagger B]} = \overline{(A, B)}$. Here, we used that $(AB)^\dagger = B^\dagger A^\dagger$, that $(A^\dagger)^\dagger = A$, and that $\mathrm{tr}[A^\dagger] = \overline{\mathrm{tr}[A]}$.

[3]Outer product representation: Let A be a self-adjoint operator acting on a finite-dimensional Hilbert space $\mathcal{H}$ with $\dim\mathcal{H} = d$, and let $\{|i\rangle\}_{i=1}^d$ be an orthonormal basis in $\mathcal{H}$. Then the outer product representation of A in this basis is given by

$$A = \sum_{i,j=1}^{d} a_{ij}|i\rangle\langle j|, \text{with } a^*_{ji} = a_{ij}.$$

   (b) Linear in the second argument: $(A, B_1 + xB_2) = \mathrm{tr}[A^\dagger(B_1 + xB_2)] = \mathrm{tr}[A^\dagger B_1] + x\mathrm{tr}[A^\dagger B_2]$.

   (c) Positive definite: $(A, A) = \mathrm{tr}[A^\dagger A]$. Now,

$$\langle\psi|A^\dagger A|\psi\rangle = \langle A\psi|A\psi\rangle = \|A\psi\|^2 \geq 0.$$

   We may write

$$\mathrm{tr}[A^\dagger A] = \sum_i \langle i|A^\dagger A|i\rangle \geq 0$$

   where $|i\rangle$ is some basis of V.

2. We can write

$$\begin{aligned} A = \mathbb{1}A\mathbb{1} &= \sum_{ij} |i\rangle\langle i|A|j\rangle\langle j| \\ &= \sum_{ij} a_{ij}|i\rangle\langle j| \end{aligned}$$

   for $a_{ij} = \langle i|A|j\rangle$. This shows that $\{|i\rangle\langle j|\}_{i,j=1}^d$ spans L_V. We can check that they are linearly independent, and thus form a basis.

Exercise 14. Polar and Singular Value Decompositions The following two decompositions of linear operators will be useful in this course.

1. **Polar Decomposition:** A linear operator A can be expressed as

$$A = U\sqrt{A^\dagger A} = \sqrt{AA^\dagger}\,U, \tag{3}$$

   where U is a unitary operator. Moreover if A is invertible, then U is unique.

2. **Singular Value Decomposition:** If A is a square matrix then there exisits unitary matrices U and V, and a diagonal matrix D with non-negative entries such that

$$A = UDV. \tag{4}$$

   The diagonal entries of D are known as *singular values* of A.

(a) Prove that $|A| := \sqrt{A^\dagger A}$ is a positive semi-definite operator.
(b) Prove (4) using (3).
(c) Use (3) to prove that for any unitary operator U

$$|\mathrm{tr}(AU)| \leq \mathrm{tr}|A|.$$

Solution

1. We saw that $B := A^\dagger A \geq 0$, i.e. was positive semi-definite, in the solution to the previous exercise. Then all of its eigenvalues are non-negative. Then since $B = B^\dagger$ (to be proven on the next example sheet), it has a spectral decomposition

   $$B = \sum_b \lambda_b \left|\phi_b\right\rangle \left\langle\phi_b\right|$$

   so

   $$\sqrt{B} = \sum_b \sqrt{\lambda_b} \left|\phi_b\right\rangle \left\langle\phi_b\right|.$$

   Then the eigenvalues of $\sqrt{B}$ are $\sqrt{\lambda_b}$ which are therefore non-negative. Thus, $\sqrt{B} \geq 0$.

2. We may write $A = U_1 J$ for $J = \sqrt{A^\dagger A} \geq 0$. Then we may write $J = UDU^\dagger$ for D a diagonal matrix with non-negative entries. Then

   $$A = U_1 U D U^\dagger = V D U^\dagger$$

   for $V = U_1 U$, which is unitary since the product of two unitaries is unitary.

3. By the cyclicity of the trace,

   $$|\mathrm{tr}[AU]| = |\mathrm{tr}[UA]|$$

   Substituting the polar decomposition $A = U_1|A|$ for some unitary U_1,

   $$= |\mathrm{tr}[UU_1|A|]|$$

Using $|A| = |A|^{1/2}A^{1/2}$ and the cyclicity of the trace,

$$= \left|\mathrm{tr}[|A|^{1/2}UU_1|A|^{1/2}]\right|$$

Using the Cauchy-Schwarz inequality $|(X,Y)| \leq \sqrt{(X,X)\cdot(Y,Y)}$ for $X = |A|^{1/2}$ and $Y = UU_1|A|^{1/2}$, as well as the fact that $|A|^{1/2}$ is Hermitian,

$$\leq \sqrt{\mathrm{tr}[|A|]\mathrm{tr}[|A|^{1/2}U_1UU^\dagger U_1^\dagger|A|^{1/2}]}$$

Since $U_1UU^\dagger U_1^\dagger = U_1 \mathbb{1} U_1^\dagger = \mathbb{1}$,

$$\leq \sqrt{\mathrm{tr}[|A|]\mathrm{tr}[|A|^{1/2}|A|^{1/2}]}$$
$$= \mathrm{tr}|A|.$$

Further problems in classical information theory (non-examinable)

Please first read the appendix before starting this problem.

Exercise 15. [Proof of the AEP and (b) of Theorem 1]

Consider a sequence of i.i.d. random variables $U_1, U_2, \ldots, U_n$ with common probability mass function $U \sim p(u)$, $u \in J$.

a) What is the expectation value of $-\log p(U_j)$?

b) What is the expectation value of $-\log p(U_1, \ldots, U_n)$?

c) Can you write $-\log p(U_1, \ldots, U_n)$ as a sum of n i.i.d. random variables?

d) Prove equation (8) using the weak law of large numbers.

e) Show that (8) implies

$$\lim_{n\to\infty} \mathbf{P}\left(2^{-n(H(U)+\epsilon)} \leq p(U_1, \ldots, U_n) \leq 2^{-n(H(U)-\epsilon)}\right) = 1. \quad (5)$$

f) Do you see that (5) implies (b) of Theorem 1?

Solution

1. The expectation value is $\mathbb{E}_p[-\log p(U_j)] = -\sum_{u\in J} p(u)\log p(u)$.

2. Since the $U_1,\ldots,U_n$ are independent, $p(U_1,\ldots,U_n) = p(U_1)\cdots p(U_n)$. Then

$$-\log p(U_1,\ldots,U_n) = -\log[p(U_1)\cdots p(U_n)] = -\sum_{i=1}^{n}\log[p(U_i)]. \quad (6)$$

   Therefore,

$$\mathbb{E}[-\log p(U_1,\ldots,U_n)] = \sum_{i=1}^{n} H(U_i) = nH(U).$$

3. Yes, by (6).

4. The weak law of large numbers says that

$$-\frac{1}{n}\log p(U_1,\ldots,U_n) = -\frac{1}{n}\sum_{i=1}^{n}\log[p(U_i)] \xrightarrow{\mathbf{P}} H(U). \quad (7)$$

5. Equation (7) is equivalent to

$$\lim_{n\to\infty}\mathbf{P}(|-\frac{1}{n}\log p(U_1,\ldots,U_n) - H(U)| \le \varepsilon) = 1.$$

   For real numbers x and y, we have $|x| \le y$ if and only if $-y \le x \le y$. Using this, we have

$$\lim_{n\to\infty}\mathbf{P}(-\varepsilon \le -\frac{1}{n}\log p(U_1,\ldots,U_n) - H(U) \le \varepsilon) = 1$$
$$\lim_{n\to\infty}\mathbf{P}(H(U)-\varepsilon \le -\frac{1}{n}\log p(U_1,\ldots,U_n) \le H(U)+\varepsilon) = 1$$
$$\lim_{n\to\infty}\mathbf{P}(-n(H(U)-\varepsilon) \ge \log p(U_1,\ldots,U_n) \ge -n(H(U)+\varepsilon)) = 1$$
$$\lim_{n\to\infty}\mathbf{P}(2^{-n(H(U)-\varepsilon)} \ge p(U_1,\ldots,U_n) \ge 2^{-n(H(U)+\varepsilon)}) = 1$$

   as desired.

6. Since a sequence $(u_1,\ldots,u_n)$ is typical if $2^{-n(H(U)-\varepsilon)} \ge p(u_1,\ldots,u_n) \ge 2^{-n(H(U)+\varepsilon}$, the previous equation gives

$$\lim_{n\to\infty}\mathbf{P}((U_1,\ldots,U_n)\in T_\varepsilon^{(n)})) = 1$$
$$\lim_{n\to\infty}\mathbf{P}(T_\varepsilon^{(n)})) = 1.$$

   By the definition of convergence for sequences of real numbers, this is equivalent to that for any $\delta > 0$ there exists $n_0(\delta) > 0$ such that for all $n \ge n_0(\delta)$, we have

$$|\mathbf{P}(T_\varepsilon^{(n)})) - 1| < \delta$$

   which implies $\mathbf{P}(T_\varepsilon^{(n)})) > 1-\delta$.

Appendix. Asymptotic Equipartition Property (AEP)

The AEP tells us that some outputs of a classical i.i.d. source occur more frequently than others. It is a direct consequence of the Weak Law of Large Numbers (WLLN). Let us first recall the WLLN and the definition of convergence in probability that it uses.

Definition. *A sequence of random variables $\{R_n\}$ converges in probability to a constant r if for all $\epsilon > 0$,*

$$\lim_{n\to\infty}\mathbf{P}\left(|R_n - r| \le \epsilon\right) = 1,$$

which we denote as $R_n \xrightarrow{\mathbf{P}} r$.

Theorem 2. *(WLLN) : If $X_1, X_2, \ldots X_n$ is a sequence of i.i.d. random variables, with partial sums $S_n = \sum_{i=1}^{n} X_i$ and finite mean μ, then*

$$\frac{1}{n}S_n \xrightarrow{\mathbf{P}} \mu.$$

If U_j is a random variable with probability mass function $p(u) \equiv P(U_j = u)$, $u \in J$, then $X_j := p(U_j)$ is a random variable which takes the value $p(u)$ with probability $p(u)$.

Similarly, if $p(u_1, \ldots, u_n)$ denotes the joint probability mass function of the random variables $U_1, U_2, \ldots, U_n$, then $X^{(n)} := p(U_1, \ldots, U_n)$ denotes a random variable which takes the value $p(u_1, \ldots, u_n)$ with probability $p(u_1, \ldots, u_n)$. Let us consider an i.i.d. information source described by random variables $U_1, U_2, \ldots, U_n$ with common probability mass function $p(u)$, $u \in J$. For such a source

$$p(u_1, \ldots, u_n) = \prod_{i=1}^{n} p(u_i),$$

and we can write the random variable $X^{(n)}$ as follows:

$$X^{(n)} := p(U_1, \ldots, U_n) = \prod_{i=1}^{n} p(U_i).$$

Theorem 3 (Asymptotic Equiprtition Theorem (AEP))**.** *Consider n uses of a memoryless information source modelled by a sequence of i.i.d. random variables $U_1, U_2, \ldots, U_n$ with common probability mass function $U \sim p(u)$, $u \in J$. The AEP states that for such a source*

$$-\frac{1}{n} \log p(U_1, \ldots, U_n) \xrightarrow{\mathbf{P}} H(U) \quad as \quad n \to \infty, \tag{8}$$

where $H(U) \equiv H(U_1) = H(U_2) = \ldots = H(U_n)$ is the Shannon entropy of the source.