Part III, Lent 2021; Quantum Information Theory

Example Sheet 1: Classical Information Theory and Linear Operators

Exercises prefaced by $\star$ are eligible for marking.

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

$\star$ **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?

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

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 $H(\lambda p + (1-\lambda)q) \geq \lambda H(p) + (1-\lambda)H(q)$.

Exercise 4. In the lectures we proved that

- For two probability distributions $p = \{p(x)\}$ and $q = \{q(x)\}$,
$$D(p||q) \geq 0 \tag{1}$$

- For a random variable $X \sim p(x)$, $x \in J_X$ (where J_X is a finite alphabet)
$$0 \leq H(X) \leq \log |J_X| \tag{2}$$

When do the equalities in (1) and (2) hold?

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.

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
$$I(X:Y) = H(X) - H(X|Y). \tag{3}$$

2. Show if X and Y are independent, $I(X:Y) = 0$.

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:

$$\begin{aligned} Y &= X \quad \text{with probability } p \\ Y &= 1 - X \quad \text{with probability } (1-p), \end{aligned} \tag{4}$$

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).

⋆ **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.

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

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

$$\text{then} \quad H(X|Y) \leq h(\varepsilon) + \varepsilon \log(m-1). \tag{5}$$

Prove (5). {*Hint: Use the result of Ex. 4 and the concavity of the 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

$$I(X : Y) \geq I(X : Z) \tag{6}$$

Prove the above inequality.

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 $p = 1/3$.

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]

Vector spaces and linear operators

Exercise 13. Hilbert Schmidt inner product (adapted from *Nielsen and Chuang*) Let V be a finite–dimensional vector 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*[1] (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.

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

$$A = U\sqrt{A^\dagger A} = \sqrt{AA^\dagger}\,U, \tag{7}$$

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

$$A = UDV. \tag{8}$$

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 (8) using (7).
(c) Use (7) to prove that for any unitary operator U

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

[1]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

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

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 (10) using the weak law of large numbers.

e) Show that (10) implies

$$\lim_{n\to\infty} \mathbf{P}\left(2^{-n(H(U)+\epsilon)} \leq p(U_1, \ldots, U_n) \leq 2^{-n(H(U)-\epsilon)}\right) = 1. \quad (9)$$

f) Do you see that (9) implies (b) of Theorem 1?

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 $\epsilon > 0$,*

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

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*

$$-\frac{1}{n} \log p(U_1, \ldots, U_n) \xrightarrow{\mathbf{P}} H(U) \quad \text{as} \quad n \to \infty, \tag{10}$$

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