B8.4 Information Theory

Sheet 2 — MT25

Section A

1. (a) Let (X_i) be a sequence of independent and identically distributed random variables taking values in a discrete set $\mathcal{X}$ and with pmf p. Let q be another probability mass function on $\mathcal{X}$ such that $q(x_1, \ldots, x_n) = \prod_{i=1}^n q(x_i)$. Show that, in probability,

$$\lim_{n\to\infty} -\frac{1}{n} \log q(X_1, X_2, \ldots, X_n) = D(p||q) + H(X).$$

(b) Let (Y_i) be a sequence of independent and identically distributed uniform random variables on $[0, 1]$. Let $V_n = \prod_{i=1}^n Y_i$ be the volume of an n-dimensional box with side lengths Y_i. Let $l_n = V_n^{1/n}$, the side length of the box with equal edges which has the same volume. Find $l = \lim_{n\to\infty} l_n$ and compare it with $\lim_{n\to\infty}(\mathbb{E}V_n)^{1/n}$.

Solution: (a)Using the form of q we have

$$-\frac{1}{n} \log q(X_1, X_2, \ldots, X_n) = -\frac{1}{n} \sum_{i=1}^n \log q(X_i).$$

Applying the WLLN we have convergence in probability

$$\begin{aligned} -\frac{1}{n} \sum_{i=1}^n \log q(X_i) &\to \mathbb{E}\left(-\log q(X)\right) \\ &= -\sum_{x\in\mathcal{X}} p(x) \log q(x) \\ &= \sum_{x\in\mathcal{X}} \left(p(x) \log \frac{p(x)}{q(x)} - p(x) \log p(x) \right) \\ &= D(p||q) + H(X). \end{aligned}$$

This convergence also holds almost surely by the SLLN.

(b) We just apply the WLLN to find l

$$V_n^{1/n} = \left(\prod_{i=1}^n Y_i \right)^{1/n} = \exp(\frac{1}{n} \sum_{i=1}^n \log Y_i) \to \exp(\mathbb{E} \log(Y)) = \exp(-1) = l.$$

Also $\mathbb{E}V_n = \prod_{i=1}^n \mathbb{E}Y_i$ by independence and so $\lim_{n\to\infty}(\mathbb{E}V_n)^{1/n} = 1/2$. The typical box volume is much smaller than that of the box with mean side lengths.

2. For any $q \in [0,1]$ and $n \in \mathbb{N}$ such that nq is an integer, show that

$$\frac{2^{nH(q)}}{n+1} \leq \binom{n}{nq} \leq 2^{nH(q)}.$$

Hint: Consider the i.i.d. Bernoulli sequence $X_1, X_2, \cdots, X_n$ with probabilities defined by $\mathbb{P}(X=1) = q$, $\mathbb{P}(X=0) = 1-q$.

Solution: As in the hint, construct an i.i.d. sequence $X_1, X_2, \cdots, X_n$ with $\mathbb{P}(X=1) = q$, $\mathbb{P}(X=0) = 1-q$. Let $S_n = \sum_{i=1}^n X_i$, and $\Gamma = \{(x_1, \cdots, x_n) : x_i \in \{0,1\}, \sum_{i=1}^n x_i = nq\}$. Then the number of elements in Γ is

$$|\Gamma| = \binom{n}{nq}.$$

It is easy to see that

$$\begin{aligned}
\mathbb{P}(S_n = np) &= \sum_{(x_1,\cdots,x_n)\in\Gamma} \mathbb{P}\{(X_1, \cdots, X_n) = (x_1, \cdots, x_n)\} \\
&= \sum_{(x_1,\cdots,x_n)\in\Gamma} q^{nq}(1-q)^{n(1-q)} \\
&= |\Gamma| 2^{-nH(q)}.
\end{aligned}$$

On one hand, it is trivial that $\mathbb{P}(S_n = nq) < 1$.

On the other hand, we know S_n follows the binomial distribution with parameter n and q. If we let $p_k = \mathbb{P}(S_n = k) = \binom{n}{nq} q^k (1-q)^{n-k}$, then

$$\frac{p_{k+1}}{p_k} = \frac{n-k}{k+1}\frac{q}{1-q},$$

so

$$\begin{aligned}
p_{k+1} \leq p_k &\Leftrightarrow (n-k)q \leq (k+1)(1-q) \\
&\Leftrightarrow nq \leq kq + (k+1)(1-q) = k + (1-q) \\
&\Leftrightarrow k \geq nq - (1-q).
\end{aligned}$$

When $nq = k_0$ is an integer, we can see p_k is increasing over $k \leq k_0$ and decreasing over $k > k_0$, which means nq achieves the maximal value of p_k, and hence

$$\mathbb{P}(S_n = nq) \geq \frac{1}{n+1}.$$

Together with the equality $\mathbb{P}(S_n = nq) = |\Gamma| 2^{-nH(q)}$, we have

$$2^{nH(q)} \geq \binom{n}{nq} \geq \frac{2^{nH(q)}}{n+1}.$$

3. Let X be a random variable taking values in a finite set $\mathcal{X}$ with pmf p. We write $\vec{X} = (X_1, \cdots, X_n)$ for a random variable in $\mathcal{X}^n$. We label elements in $\mathcal{X}$ in non-decreasing order according to $p(x)$, so that $p_i = \mathbb{P}(X = x_i)$ is non-decreasing in i. Using this we can rank the probability $\mathbb{P}(\vec{X} = \vec{x})$ for all $\vec{x} = (x_1, \cdots, x_n) \in \mathcal{X}^n$, and explicitly construct the smallest subset $\mathcal{S}_n^\varepsilon$ of $\mathcal{X}^n$, by greedily including the elements in $\mathcal{X}^n$ with highest probabilities one-by-one, such that $\mathbb{P}(\vec{X} \in \mathcal{S}_n^\varepsilon) \geq 1 - \varepsilon$.

   Show that for any $\varepsilon > 0$, there exists n_0, such that for any $n \geq n_0$, we have

$$(1 - 2\varepsilon)2^{n(H(X)-\varepsilon)} \leq |\mathcal{S}_n^\varepsilon| \leq 2^{n(H(X)+\varepsilon)}.$$

   Use this to complete the proof of Proposition 2.15 in the lecture notes.

   Hint: For any $\varepsilon_1 \in [0, 1), \varepsilon_2 \in [0, 1)$ and events A, B with $\mathbb{P}(A) \geq 1 - \varepsilon_1, \mathbb{P}(B) \geq 1 - \varepsilon_2$, show that $\mathbb{P}(A \cap B) \geq 1 - \varepsilon_1 - \varepsilon_2$. Use this inequality to estimate $\mathbb{P}(\mathcal{S}_n^\varepsilon \cap \mathcal{T}_n^\varepsilon)$.

   Solution: Firstly, if $\mathbb{P}(A_i) \geq 1 - \varepsilon_i$ for $i = 1, 2$, then $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1) - \mathbb{P}(A_1 \cap A_2^c) \geq 1 - \varepsilon_1 - \varepsilon_2$.

   Recall the set of typical sequences $\mathcal{T}_n^\varepsilon$. We know for any $\varepsilon > 0$, any $n > 0$, $\mathbb{P}(\vec{X} = \vec{x}) \in (2^{-n(H(X)+\varepsilon)}, 2^{-n(H(X)-\varepsilon)}]$, and there exists $n_0 > 0$, such that for any $n \geq n_0$,

$$\mathbb{P}(\vec{X} \in \mathcal{T}_n^\varepsilon) \geq 1 - \varepsilon, \text{ and } |\mathcal{T}_n^\varepsilon| \in [(1 - \varepsilon)2^{n(H(X)-\varepsilon)}, 2^{nH(X)+\varepsilon}].$$

   So for any $n \geq n_0$, we have

$$\begin{aligned} 1 - 2\varepsilon &\leq \mathbb{P}(\vec{X} \in \mathcal{S}_n^\varepsilon \cap \mathcal{T}_n^\varepsilon) \\ &= \sum_{\vec{x} \in \mathcal{S}_n^\varepsilon \cap \mathcal{T}_n^\varepsilon} \mathbb{P}(\vec{X} = \vec{x}) \\ &\leq \sum_{\vec{x} \in \mathcal{S}_n^\varepsilon \cap \mathcal{T}_n^\varepsilon} 2^{-n(H(X)-\varepsilon)} \\ &\leq \sum_{\vec{x} \in \mathcal{S}_n^\varepsilon} 2^{-n(H(X)-\varepsilon)} \\ &= |\mathcal{S}_n^\varepsilon| 2^{-n(H(X)-\varepsilon)}, \end{aligned} \tag{1}$$

   hence $|\mathcal{S}_n^\varepsilon| \geq (1 - 2\varepsilon)2^{n(H(X)-\varepsilon)}$.

   On the other hand, $\mathcal{S}_n^\varepsilon$ is the smallest set of probability $1 - \varepsilon$, hence $|\mathcal{S}_n^\varepsilon| \leq |\mathcal{T}_n^\varepsilon| \leq 2^{n(H(X)+\varepsilon)}$.

   To complete the proof of Proposition 2.15 we know that by the minimality of $|\mathcal{S}_n^\varepsilon|$ we have

$$\frac{|\mathcal{S}_n^\varepsilon|}{|\mathcal{T}_n^\varepsilon|} \leq 1.$$

Using that $|\mathcal{T}_n^\varepsilon| \leq 2^{n(H(X)+\epsilon)}$ we have from (1) that

$$\frac{|\mathcal{S}_n^\varepsilon|}{|\mathcal{T}_n^\varepsilon|} \geq (1-2\epsilon)2^{-2\epsilon n}.$$

Thus taking logs we have for each n and $\epsilon > 0$ that

$$\log(1-2\epsilon) - 2\epsilon n \leq \log \frac{|\mathcal{S}_n^\varepsilon|}{|\mathcal{T}_n^\varepsilon|} \leq 0.$$

Hence dividing by n, taking $n \to \infty$ gives

$$-2\epsilon \leq \lim_{n\to\infty} \frac{1}{n} \log \frac{|\mathcal{S}_n^\varepsilon|}{|\mathcal{T}_n^\varepsilon|} \leq 0.$$

Now take ϵ to 0 to get the result.

Section B

4. Let (X_i) be a sequence of independent and identically distributed random variables taking values in a discrete set $\mathcal{X}$ and with pmf p.

   Let $\hat{p}_n$ be the empirical measure obtained from the first n samples; that is for $x \in \mathcal{X}$ we set

   $$\hat{p}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i=x\}}.$$

   (I_A is the indicator of the event A).

   (a) Show that for each x we have $\hat{p}_n(x) \to p(x)$ in probability as $n \to \infty$.

   (b) Show that

   $$\mathbb{E}D(\hat{p}_{2n}||p) \leq \mathbb{E}D(\hat{p}_n||p).$$

   Hint: D is convex

   (c) Show that the mean divergence from the true pmf decreases monotonically in the sample size used to generate the empirical measure:

   $$\mathbb{E}D(\hat{p}_n||p) \leq \mathbb{E}D(\hat{p}_{n-1}||p).$$

   Hint: Write the empirical measure from n samples as an average of empirical measures with one sample deleted

5. We are given a fair coin, and want to generate a random variable X, by i.i.d. sampling from tossing the coin, such that X follows the distribution

   $$\mathbb{P}(X = 1) = p,\ \mathbb{P}(X = 0) = 1 - p$$

   for any given constant $p \in (0, 1)$.

   Suppose $Z_1, Z_2, \cdots$ are the results of independent tossing of the coin, i.e., $\{Z_i\}$ is an i.i.d. sequence of random variables with the distribution $\mathbb{P}(Z = 0) = \mathbb{P}(Z = 1) = \frac{1}{2}$. Let $U = \sum_{i=1}^{+\infty} Z_i 2^{-i}$, and define

   $$X = \begin{cases} 1 & \text{if } U < p \\ 0 & \text{otherwise} \end{cases}.$$

   (a) Show that U follows a uniform distribution over $[0, 1)$, and hence show that $\mathbb{P}(X = 1) = p,\ \mathbb{P}(X = 0) = 1 - p$.

   (b) Let I be the minimal number n such that we can tell $U < p$ based on $Z_1, \cdots Z_n$. Calculate $\mathbb{E}[I]$ and show that $\mathbb{E}[I] \leq 2$.

6. Let X_1 be a random variable taking values in $\mathcal{X}_1 = \{1, 2, \cdots, m\}$ and X_2 be a random variable taking values in $\mathcal{X}_2 = \{m+1, \cdots, n\}$ for integers $n > m$. Let θ be a random variable with $\mathbb{P}(\theta = 1) = \alpha$, $\mathbb{P}(\theta = 2) = 1 - \alpha$ for some $\alpha \in [0, 1]$. Define a new random variable

$$X = X_\theta.$$

Furthermore, suppose θ, X_1, X_2 are independent of each other.

(a) Express $H(X)$ in terms of $H(X_1), H(X_2)$ and $H(\theta)$.

(b) Show that $2^{H(X)} \leq 2^{H(X_1)} + 2^{H(X_2)}$. Can the equality hold in this inequality?

7. (a) Let X be a random variable taking 6 values $\{A, B, C, D, E, F\}$ with probabilities 0.5, 0.25, 0.1, 0.05, 0.05, 0.05 respectively.

(i) Construct a binary Huffman code for this random variable and compute its expected length.

(ii) Construct a quaternary Huffman code for this random variable and compute its expected length. You may find it helpful to use $\{a, b, c, d\}$ for the symbols in the quaternary code.

(iii) Construct a binary code for the random variable by converting the symbols in the quaternary code to binary by setting $a \to 00, b \to 01, c \to 10, d \to 11$. What is the expected length of this code?

(b) We now consider any random variable X

(i) Let L_H be the expected length of the binary Huffman code for X and L_{QB} be the expected length of the binary code obtained by constructing the quaternary code and then converting to binary. Show that

$$L_H \leq L_{QB} < L_H + 2.$$

Hint: You may want to compare with entropy

(ii) Give an example where the optimal quaternary code gives the optimal binary code, so $L_H = L_{QB}$.

(iii) In fact the upper bound can be reduced to $L_{QB} \leq L_H + 1$. Can you find an example where this bound is tight?

8. International Morse code is a ternary encoding of the Latin alphabet, traditionally represented as dots and dashes. A version of the encoding (written in terms of digits 0,1) is given in the file `IMC.csv`. Here we represent a dot as '10', a dash as '1110' and the pause between letters as '0000000' (representing the typical length of the dot-dash-pause).

   (a) Explain why Morse code is a prefix code, but is not a uniquely decodable code if the ending pauses are excluded.

   (b) Using the single letter counts and the Huffman algorithm, determine a binary code which encodes each single character as a single block.

   (c) Using the single letter counts and the Huffman algorithm, determine a binary code which encodes each pair of characters as a single block, assuming characters are sampled independently.

   (d) Using the double letter counts and the Huffman algorithm, determine a binary code which encodes each pair of consecutive letters as a single block.

   (e) Using the double letter counts, evaluate the average message lengths of each of the codes above (including International Morse code), when used on pairs of consecutive English characters.

   Remark: You only need to submit solutions to (a,e).

Section C

9. The *differential entropy* of a $\mathbb{R}^n$-valued random variable X with density function $f(\cdot)$ is defined as

$$h(X) := -\int_{\mathbb{R}^n} f(x)\log(f(x))\mathrm{d}x$$

with the convention $0\log(0) = 0$.

(a) Calculate $h(X)$ for the following cases with $n = 1$.

(1) X is uniformly distributed on an interval $[a, b] \subset \mathbb{R}$;

(2) X is a standard normal distribution;

(3) X is exponential distributed with parameter $\lambda > 0$.

(b) For general n-dimensional case, if $\mathbb{E}[X] = 0$, and $\mathrm{Var}(X) = K$, (K is the variance-covariance matrix). Show that

$$h(X) \leq n\log(\sqrt{2\pi e}) + \log(\sqrt{|K|})$$

with the equality hold iff X is multivariable normal.

Hint: you can firstly prove the continuous version of Gibbs' inequality: For any two density functions $f(\cdot)$ and $g(\cdot)$,

$$-\int f(x)\log(f(x))\mathrm{d}x \leq -\int f(x)\log(g(x))\mathrm{d}x.$$

Also, you can try to prove (or use it without proof) the following property of the variance-covariance matrix: If $X = (X_1, \cdots, X_n)^\top$ has expectation 0 and variance-covariance matrix $\mathrm{Var}(X) = K$, then

$$\mathbb{E}[X^\top K^{-1} X] = n.$$

Solution:

(a) $h(X) = -\mathbb{E}[\log(f(X)] = \mathbb{E}[\log(1/f(X))]$.

(a.1) $f(x) = \frac{1}{b-a}$ for any $x \in [a, b]$, and $f(x) = 0$ otherwise. So $h(X) = \mathbb{E}[\log(b - a)] = \log(b - a)$.

(a.2) $f(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$, so

$$\begin{aligned} h(X) = \mathbb{E}[\log(\sqrt{2\pi}e^{X^2/2})] &= \log(\sqrt{2\pi}) + \mathbb{E}[\frac{X^2}{2}\log(e)] \\ &= \log(\sqrt{2\pi}) + \frac{1}{2}\log(e) = \log(\sqrt{2\pi e}). \end{aligned}$$

(a.3) $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ and $f(x) = 0$ for $x < 0$. So

$$h(X) = \mathbb{E}[-\log(\lambda) + \lambda X \log(e)] = -\log(\lambda) + \lambda \log(e)\frac{1}{\lambda} = \log(e) - \log(\lambda).$$

(b) Let $X = (X_1, \cdots, X_n)^\top$ be a normal random vector with mean $\mathbb{E}[X] = 0$ and variance $\mathbb{E}[X^\top X] = K$. Let g be its density function, i.e.

$$g(x) = \frac{1}{\sqrt{(2\pi)^n |K|}} e^{-\frac{1}{2}x^\top K^{-1} x} \quad \forall\, x \in \mathbb{R}^n.$$

We first calculate $h(g)$.

$$\begin{aligned} h(g) &= -\mathbb{E}[\log(g(X))] \\ &= \frac{1}{2}\log((2\pi)^n |K|) + \frac{1}{2}\log(e)\mathbb{E}[X^\top K^{-1} X] \\ &= \frac{n}{2}\log(2\pi) + \frac{1}{2}\log|K| + \frac{1}{2}\log(e)n \\ &= \frac{n}{2}\log(2\pi e) + \frac{1}{2}\log|K| \\ &= n\log(\sqrt{2\pi e}) + \log(\sqrt{|K|}). \end{aligned}$$

Then we prove that $h(f) \leq h(g)$ for any f with mean 0 and variance-covariance matrix K. For any random vector Y with the density f, we have

$$\begin{aligned} h(f) &= -\mathbb{E}[\log(f(Y)] \\ &= -\mathbb{E}[\log(g(Y))] + \mathbb{E}[\log(g(Y)/f(Y))]. \end{aligned}$$

For the first term

$$\begin{aligned} -\mathbb{E}[\log(g(Y))] &= \frac{1}{2}\log((2\pi)^n |K|) + \frac{1}{2}\log(e)\mathbb{E}[Y^\top K^{-1} Y] \\ &= -\mathbb{E}[\log(g(X))] = h(g). \end{aligned}$$

For the second term, by Jensen's inequality,

$$\begin{aligned} \mathbb{E}[\log(g(Y)/f(Y))] &\leq \log(\mathbb{E}[g(Y)/f(Y)]) \\ &= \log(1) = 0. \end{aligned}$$

So we get $h(f) \leq h(g)$, and the equality hold iff $g(Y) \equiv f(Y)$.

10. Prove the following weaker version of the Kraft-McMillan theorem (called Kraft's theorem) using rooted trees

    (a) Let $c : \mathcal{X} \mapsto \{0, \cdots, d-1\}^*$ be a prefix code. Consider its code-tree and argue that $\sum_{x\in\mathcal{X}} d^{-|c(x)|} \leq 1$. [Note that the assumption that c is a prefix code is crucial here, otherwise the code-tree cannot be defined to begin with. In the Kraft-McMillan theorem from the lecture we only require c to be uniquely decodable].

    (b) Assume that $\sum_{x\in\mathcal{X}} d^{-l_x} \leq 1$ with $l_x \in \mathbb{N}$. Show that there exists a prefix code c with codeword lengths $|c(x)| = l_x$ for $x \in \mathcal{X}$ by constructing a rooted tree.

Solution: A prefix code is equivalent to a rooted tree, where each codeword corresponds to a path from a leave to the root.

(a) We call a d-ary tree semi-complete if every non-leaf vertex has d direct descendants. In a semi-complete d-ary tree for any leaf x, denote $h(x)$ as the height from the root to the tree with $h(root) = 0$. It is easy to check that $\sum_{\text{every leaf } x} d^{-h(x)} = 1$.

For the code-tree of a prefix code, it can be expanded to a semi-complete tree by adding some leaves to a non-leave vertex. Hence $\sum_{\text{every leaf } x} d^{-h(x)} \leq 1$.

(b) We call a d-ary tree complete with height h if it is semi-complete, the distance from each leaf to the root is h.

Given l_x satisfies the condition, denote $h = \max_x l_x$, then we can construct a d-ary complete tree with maximal height h.

Suppose $l_1 \leq l_2 \leq \cdots \leq l_m$. We mark nodes and cut branches of a complete tree as follows:

(1) Take $i = 1$.

(2) Find the first non-marked node on the left of the tree with height l_i, cut off its descendant vertices, and mark all ancestral vertices (including itself) and their edges down to the root.

(3) Set $i = i + 1$ and repeat (2) until $i = m + 1$.

For each x, we find a vertex with height l_x, cut off its descendant vertices and mark it the leaf of x, and mark all ancestral vertices and edges between the leaf x and the root.

By the assumption $\sum_{i=1}^{m} d^{-l_i} \leq 1$, we know we can run this construction for all $k \leq m$ (otherwise, if we cannot find a node with height l_k at some $k \leq m$, then it must happen that $\sum_{i=1}^{k} d^{-l_x} > 1$).

The labels of all marked vertices and the i^{th} leaf in the algorithm corresponds to the codeword i.