B8.4 Information Theory

Sheet 2 — MT23

Section A

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

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

with 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 $\mathbb{P}(Z=0)=\mathbb{P}(Z=1)=\frac{1}{2}$. Denote $U=\sum_{i=1}^{+\infty} Z_i 2^{-i}$, 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) Denote I as the minimal number of n such that we can tell $U<p$ based on $Z_1,\cdots Z_n$. Calculate $\mathbb{E}[I]$ and show that $\mathbb{E}[I]\le 2$.

Solution:

(a) For any $q \in [0,1)$, denote its binary expansion as $q=0.a_1a_2\cdots$, i.e. $q=\sum_{i=1}^{+\infty} a_i 2^{-i}$ with $a_i \in \{0,1\}$ (with the convention that $1111\cdots$ is not allowed), and define

$$I=\min\{i: Z_i \neq a_i\}.$$

Then $\mathbb{P}(U=p)=\mathbb{P}(I=\infty)=0$, and

$$\begin{aligned}
\mathbb{P}(U<p) &= \mathbb{P}(I<+\infty, Z_I<a_i) \\
&= \sum_{n=1}^{+\infty} \mathbb{P}(I=n, Z_n<a_n) \\
&= \sum_{n=1}^{+\infty} \mathbb{P}\{Z_1=a_1,\cdots,Z_{n=1}=a_{n-1},\ \text{and } Z_n<a_n\} \\
&= \sum_{n=1}^{+\infty} 2^{-(n-1)}2^{-1}a_n \\
&= p.
\end{aligned}$$

Sam Cohen: cohens@maths.ox.ac.uk

Since $\mathbb{P}(X = 1) = \mathbb{P}(U < p) = p$, we know the distribution of X is $(p, 1-p)$.

(b) With a little abuse of notation, suppose $p = 0.a_1a_2\cdots$ and I defined as above. Then we can tell $U < p$ at time I.

Since

$$\begin{aligned}
\mathbb{P}(I = n) &= \mathbb{P}(Z_1 = a_1, \cdots Z_{n-1} = a_{n-1}, Z_n \neq a_n) \\
&= 2^{-n}, \\
\mathbb{E}[I] &= \sum_{n=1}^{+\infty} n2^{-n} \\
&= 2.
\end{aligned}$$

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 $\mathbb{P}(X = 1) = q$, $\mathbb{P}(X = 0) = 1 - q$. Denote $S = \sum_i X_i$, and $\Gamma = \{(x_1, \cdots, x_n) : x_i \in \{0,1\}, \sum_n x_i = nq\}$. Then the number of elements in Γ is

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

It is easy to see that

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

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

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

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

so

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

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

$$\mathbb{P}(S = nq) \ge \frac{1}{n+1}.$$

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

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

Sam Cohen: cohens@maths.ox.ac.uk

Section B

3. Let X_1 be a random variable valued in $\mathcal{X}_1 = \{1, 2, \cdots, m\}$ and X_2 be a random variable valued 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 to 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?

4. Let X be a random variable with pmf p over the image space $\mathcal{X}$ with finite elements $k = |\mathcal{X}|$, $\vec{X} = (X_1, \cdots, X_n)$. We label elements in $\mathcal{X}$ by a non-decreasing order of $p(x)$, such that $p_i = \mathbb{P}(X = x_i)$ is non-decreasing in i. By this labelling, we can easily rank the probability $\mathbb{P}(\vec{X} = \vec{x})$ for all $\vec{c} \in \mathcal{X}^n$, and explicitly construct the smallest set $\mathcal{S}_n^\varepsilon$ by greedily including the element in $\mathcal{X}^n$ with highest probabilities one-by-one until we have $\mathbb{P}(\vec{X} \in \mathcal{S}_n^\varepsilon) \geq 1 - \varepsilon$.

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

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

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 $\mathbb{P}(\mathcal{S}_n^\varepsilon \cap \mathcal{T}_n^\varepsilon)$.

Sam Cohen: cohens@maths.ox.ac.uk

5. 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).

   Remark: To account for Morse code being a ternary code, multiply the average length of a message by $\log(3)$, for a fair comparison with binary codes.

Sam Cohen: cohens@maths.ox.ac.uk

Section C

6. 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

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

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

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

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

Sam Cohen: cohens@maths.ox.ac.uk

7. Consider the space of random variables $\mathcal{X}$ on a discrete space.

   (a) Show that the function $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by $(X, Y) \mapsto H(X|Y) + H(Y|X)$ is a pseudo-metric (that is, it is positive, symmetric and satisfies the triangle inequality).

   (b) Show that $\rho(X, Y) = 0$ if and only if there exists a function f such that $f(X) = Y$ with probability one, and hence ρ is a metric on the corresponding equivalence class (where $X \sim Y$ iff $f(X) = Y$ for some Y)

Solution:

(a) Clearly $\rho(X, Y) \geq 0$ and $\rho(X, Y) = \rho(Y, X)$. For any three random variables, we have

$$\begin{aligned} H(X|Y) + H(Y|Z) &\geq H(X|Y, Z) + H(Y|Z) \\ &= H(X, Y|Z) \\ &= H(X|Z) + H(Y|X, Z) \\ &\geq H(X|Z) \end{aligned}$$

Therefore,

$$\begin{aligned} \rho(X, Y) + \rho(Y, Z) &= H(X|Y) + H(Y|X) + H(Y|Z) + H(Z|Y) \\ &\geq H(X|Z) + H(Z|X) = \rho(X, Z). \end{aligned}$$

(b) If such an f exists, then it is easy to see that $H(X|Y) = H(Y|X) = 0$, as the conditional probabilities are trivial. Conversely, by positivity if $\rho(X, Y) = 0$ then $H(Y|X) = 0$, and so the conditional probability must be trivial. We can then define the map f to be the selector: $f(x) = y$ if $\mathbb{P}(Y = y|X = x) = 1$. The result follows.

Sam Cohen: cohens@maths.ox.ac.uk