B8.4 Information Theory

Sheet 1 — MT23

Section A

1. (Polling inequalities) Let $a \geq 0, b \geq 0$ are given with $a+b>0$. Show that

$$-(a+b)\log(a+b) \leq -a\log(a) - b\log(b) \leq -(a+b)\log(\frac{a+b}{2})$$

and that the first inequality becomes an equality iff $ab = 0$, the second inequality becomes an equality iff $a = b$.

Solution: Denote $p = \frac{a}{a+b}$. Divide by $a+b$ and then add $\log(a+b)$ on all three terms, the equalities are equivalent to

$$0 \leq -p\log(p) - (1-p)\log(1-p)) \leq -\log(\frac{1}{2}),$$

which is obvious according to the first basic property of entropy.

2. Let X, Y, Z be discrete random variables. Prove or provide a counterexample to the following statements:

(a) $H(X) = H(-42X)$;

(b) $H(X|Y) \geq H(X|Y,Z)$;

(c) $H(X,Y) = H(X) + H(Y)$.

Solution: The first one is true : $f(x) = -42x$ is a bijective.

The second is true: $H(X|Y) - H(X|Y,Z) = I(X,Z|Y) > 0$, which motivates the interpretation of information as a 'surprise'.

The third is wrong: By the chain rule, $H(X,Y) = H(Y|X) + H(X)$, and $H(Y|X) = H(Y)$ if and only if X, Y are independent. An easy counter example is when $Y = X$ and $H(X) > 0$, we have $H(X,Y) = H(X,X) = H(X) < H(Y) + H(X)$.

Section B

3. Does there exist a discrete random variable X with a distribution such that $H(X) = +\infty$? If so, describe it as explicitly as possible.

4. An urn contains r red, w white and b black balls. Compute the entropy of the outcome of drawing two balls from this urn with replacement, and determine whether this entropy is higher or lower than when the balls are drawn without replacement.

5. Let $\mathcal{X}$ be a finite set, f a real-valued function $f : \mathcal{X} \mapsto \mathbb{R}$ and fix $\alpha \in \mathbb{R}$. We want to maximise the entropy $H(X)$ of a random variable X taking values in $\mathcal{X}$ subject to the constraint

$$\mathbb{E}[f(X)] \leq \alpha. \tag{1}$$

Denote by U a uniformly distributed random variable over $\mathcal{X}$. Prove the following optimal solutions for the maximisation.

(a) If $\alpha \in [\mathbb{E}[f(U)],\ \max_{x\in\mathcal{X}} f(x)\,]$, then the entropy is maximised subject to (1) by the uniformly distributed random variable U.

(b) If f is non-constant and $\alpha \in [\min_{x\in\mathcal{X}} f(x),\ \mathbb{E}[f(U)]\,]$, then the entropy is maximised subject to (1) by the random variable Z given by

$$\mathbb{P}(Z = x) = \frac{e^{\lambda f(x)}}{\sum_{y\in\mathcal{X}} e^{\lambda f(y)}} \qquad \text{for } x \in \mathcal{X}.$$

where $\lambda < 0$ is chosen such that $\mathbb{E}[f(Z)] = \alpha$.

(c) Prove that, under the assumptions of (b), the choice for λ is unique and we have $\lambda < 0$.

6. The attached file `letter_counts.zip` contains the counts of how many times every combination of up to four letters (from an alphabet of A-Z and an underscore for space) occurs in a small corpus of writing (the collected works of Shakespeare, Michel de Montaigne (translated), Mark Twain and Walt Whitman, almost all in English), with all punctuation removed. (For interest, the file `corpus.zip` contains the original files and the python script used to extract the counts.)

Suppose X_1, X_2, X_3, X_4 are four consecutive letters chosen at random in an English text. Using this count data, estimate the conditional entropies $H(X_1)$, $H(X_2|X_1)$, $H(X_3|X_2, X_1)$, $H(X_4|X_3, X_2, X_1)$ along with the entropy if letters were uniformly chosen from the 27-letter alphabet.

[You do not need to provide the details of your computation, the numerical answer is sufficient]

Section C

7. Let X be a real-valued random variable.

   (a) Assuming additionally that X is non-negative, show that for every $x > 0$, we have

   $$\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}[X]}{x}.$$

   (b) Let X be a random variable of mean μ and variance σ^2. Show that

   $$\mathbb{P}(|X - \mu| > \varepsilon) \leq \frac{\sigma^2}{\epsilon^2}.$$

   (c) Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d random variables with mean μ and variance σ^2. Show that $\frac{1}{m}\sum_{n=1}^{m} X_n$ converges to μ in probability, i.e. for every $\varepsilon > 0$,

   $$\lim_{m\to+\infty} \mathbb{P}\left(\left|\frac{1}{m}\sum_{n=1}^{m} X_n - \mu\right| \leq \epsilon\right) = 1.$$

   (This is the *weak law of large numbers*)

   Remark: Using the Borel–Cantelli lemma (and the existence of a fourth moment for X), or using martingale theory (see B8.1), we can show that the *strong law of large numbers* also holds, that is

   $$\mathbb{P}\left(\lim_{m\to+\infty} \frac{1}{m}\sum_{n=1}^{m} X_n = \mu\right) = 1.$$

   Solution:

   (a) $\mathbb{E}[X] = \mathbb{E}[X\mathbf{1}_{X\geq x}] + \mathbb{E}[X\mathbf{1}_{X<x}] \geq \mathbb{E}[X\mathbf{1}_{X\geq x}] \geq \mathbb{E}[x\mathbf{1}_{X\geq x}] = x\mathbb{P}(X \geq x)$, so we have the inequality.

   (b) Similar to part (a), for any random variable Y and constant $\varepsilon > 0$, $\mathbb{P}(|Y| > \varepsilon) \leq \frac{\mathbb{E}[Y^2]}{\varepsilon^2}$. Apply $Y = X - \mu$ in this inequality, we get the one in the question.

   (c) For any integer m, denote $Y_m = \frac{1}{m}\sum_{n=1}^{m} X_n - \mu$, then $\mathbb{E}[Y_m] = 0, \mathrm{Var}(Y_m) = \frac{\sigma^2}{m}$. Hence $\mathbb{P}(|Y_m| > \varepsilon) \leq \frac{\sigma^2}{m\varepsilon} \overset{m\to+\infty}{\longrightarrow} 0$.

8. 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.

9. Partition the interval $[0, 1]$ into n disjoint sub-intervals of length $p_1, \cdots, p_n$. Let $X_1, X_2, \cdots$ be i.i.d. random variables, uniformly distributed on $[0, 1]$, and $Z_m(i)$ be the number of the $X_1, \cdots, X_m$ that lie in the i^{th} interval of the partition. Show that the random variables

$$R_m = \Pi_{i=1}^n p_i^{Z_m(i)} \text{ satisfy } \frac{1}{m} \log(R_m) \overset{m\to+\infty}{\longrightarrow} \sum_{i=1}^n p_i \log(p_i) \text{ with probability 1.}$$

Solution: Denote I_i as the i^{th} subinterval. By the definition of $Z_m(i)$, we have $Z_m(i) = \sum_{j=1}^m \mathbf{1}_{X_j \in I_i}$, and by the law of large numbers,

$$\mathbb{P}\left(\lim_{m\to+\infty} \frac{\sum_{j=1}^m \mathbf{1}_{X_j\in I_i}}{m} = p_i\right) = 1.$$

It is easy to see that

$$\frac{1}{m}\log(R_m) = \frac{1}{m}\sum_{i=1}^n Z_m(i)\log(p_i) = \sum_{i=1}^n \log(p_i)\frac{\sum_{j=1}^m \mathbf{1}_{X_j\in I_i}}{m} \overset{m\to+\infty}{\longrightarrow} \sum_{i=1}^n p_i \log(p_i).$$