Information Theory **DCC/ICEx/UFMG**
Prof. Mário S. Alvim **2017/02**

SOLUTION OF HOMEWORK

Probability, Entropy, and Inference / The Source Coding Theorem
(MacKay - Chapter 2 / Chapter 4)

Necessary reading for this assignment:

- *Information Theory, Inference, and Learning Algorithms* (MacKay):

  Chapter 2

  - Chapter 2.4: *Definition of entropy and related functions*
  - Chapter 2.7: *Jensen's inequality for convex functions*

  Chapter 4

  - Chapter 4.1: *How to measure the information content of a random variable?*
  - Chapter 4.2: *Data compression*
  - Chapter 4.3: *Information content defined in terms of lossy compression*
  - Chapter 4.4: *Typicality*
  - Chapter 4.6: *Comments*

Note: The exercises are labeled according to their level of difficulty: `[Easy]`, `[Medium]` or `[Hard]`. This labeling, however, is subjective: different people may disagree on the perceived level of difficulty of any given exercise. Don't be discouraged when facing a hard exercise, you may find a solution that is simpler than the one the instructor had in mind!

Review questions.

1. Answer formally the following questions:

   (a) Define the Shannon information content $h(x)$ of the outcome x of a random experiment. Explain what the value $h(x)$ means.

   Instructor's solution: The Shannon information content $h(x)$ of the outcome of a random experiment x is defined as $h(x) = \log {}^1\!/_{p(x)}$, where $p(x)$ is the probability of outcome x.
   The value $h(x)$ represents a measure of how surprising or informative the happening of outcome x is. Intuitively, the less probable an outcome x, the more information the happening of x conveys, and this is reflected by $h(x)$ being larger.

   (b) Define the entropy $H(X)$ of an ensemble X. Explain what the value $H(X)$ means.

   Instructor's solution: The entropy $H(X)$ of an ensemble $X = (x, \mathcal{A}_X, \mathcal{P}_X)$ is defined as $H(X) = \sum_{x \in \mathcal{A}_X} p(x) \log 1/p(x)$, with the convention that $0 \cdot \log 1/0 = 0$.
   The value $H(X)$ represents the expected value of how surprising or informative outcomes drawn according to the probabilities of the ensemble X are. Intuitively, the higher the entropy of the ensemble, the more information is contained in it (since the expected surprise of outcomes in the ensemble is higher).

   (c) Define what is a convex $\smile$ function. Give at least two examples of functions that are convex $\smile$, and at least two of functions that are not.

Instructor's solution: A function $f(x)$ is a convex $\smile$ function if, for every interval (a, b) in the function's domain, every chord of the function lies above the function. That is, for all $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$, $f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2$.
Examples of convex $\smile$ functions include $f(x) = \log x$, $f(x) = -x^2$, $f(x) = x$, and $f(x) = k$ (for some $k \in \mathbb{R}$). Examples of functions that are not convex include $f(x) = x^2$, $f(x) = x^3$, and $f(x) = \sin x$.

(d) State *Jensen's inequality.*

Instructor's solution: Jensen's inequality states that if f is a *simle* convex function and x is a random variable, then $E[f(x)] \geq f(E[x])$, where $E[\cdot]$ denotes expected value.

(e) What is the formula for the raw bit content of an ensemble X? What does it mean?

Instructor's solution: The raw bit content of an ensemble X is defined as $H_0(X) = \log |\mathcal{A}_X|$, and it is a lower bound on the number of binary questions that are always guaranteed to identify an outcome of the ensemble X.

(f) Given an ensemble X, what is its smallest δ-sufficient subset S_δ?

Instructor's solution: S_δ is the smallest subset of $\mathcal{A}_X$ having probability of at least $1 - \delta$. That is, $p(x \in S_\delta) \geq 1 - \delta$.

(g) Given an ensemble X and a value $0 < \delta < 1$, what is the essential bit content $H_\delta(X)$? What does it mean?

Instructor's solution: The essential bit content of X is $H_\delta(X) = \log |S_\delta|$, and it represents the minimum number of bits we need to encode every symbol of source X if we are willing to accept at most a probability δ of error in decoding.

(h) Shannon's source coding theorem can be stated as follows: If X is an ensemble with entropy $H(X) = H$ bits, then given any $\epsilon > 0$ and $0 < \delta < 1$, there exists a positive integer N_0 such that for $N > N_0$,

$$\left| \frac{1}{N} H_\delta(X^N) - H \right| < \epsilon.$$

Explain what it means for data compression.

Instructor's solution: The theorem implies that any source with entropy H can be compressed to $H_\delta(X)$ bits per symbol with probability of error at most δ.

Problems (Chapter 2).

2. (Lower bound for Shannon entropy) `[Easy]` Show that for every ensemble $X = (x, \mathcal{A}_X, \mathcal{P}_X)$, it is the case that $H(X) \geq 0$.

Instructor's solution: The probability of any outcome x is, by definition, always non-negative, so it is always the case that $\log 1/p(x) \geq 0$. Therefore every term in the sum $H(X) = \sum_x p(x) \log \frac{1}{p(x)}$ is non-negative and we must have $H(X) \geq 0$.

3. (Upper bound for Shannon entropy) The following exercises are designed so you can prove an upper bound for Shannon entropy.

   (a) (MacKay 2.21) `[Easy]`

   Instructor's solution: Given in the textbook.

   (b) (MacKay 2.22) `[Easy]`

   Instructor's solution: Given in the textbook.

   (c) (MacKay 2.25) `[Hard]` (Hint: Use Jensen's inequality!)

   Instructor's solution: Given in the textbook.
   However, the following may help.
   Recall that Jensen's inequality states that if f is a $\smile$ convex function and x is a random variable, then $E[f(x)] \geq f(E[x])$, where $E[\cdot]$ denotes expected value.
   If we want to use Jensen's inequality to bound Shannon entropy, it may be tempting to note that $f(x) = \log {}^1\!/_x$ is a convex function and write Shannon entropy as the expectation $H(x) = E[f(p(x))] = \sum_x p(x) \log {}^1\!/_{p(x)}$. Applying Jensen's inequality to this expectation would give us

$$\begin{aligned} H(x) &= E[f(p(x))] \\ &\geq f(E[p(x)]) && \text{(by Jensen's inequality)} \\ &= \log {}^1\!/_{(E[p(x)])}, \end{aligned}$$

   which is a valid lower bound, but not an upper bound, as we wanted.
   To show what we want, we can combine Jensen's inequality with the result of the Exercise (MacKay 2.22). Noting that $-\log x$ is a convex function, we can derive:

$$\begin{aligned} -\log |\mathcal{A}_X| &= -\log \left(E[{}^1\!/_{p(x)}]\right) && \text{(by Exercise MacKay 2.22)} \\ &\leq -E[\log_2 {}^1\!/_{p(x)}] && \text{(by Jensen's inequality)} \\ &= -H(X) && \text{(by the definition of entropy),} \end{aligned}$$

   which implies that $H(X) \leq \log |\mathcal{A}_X|$.

4. (Thomas&Cover 2.1) (The entropy of a countably infinite probability distribution) `[Medium]` A fair coin is flipped until the first head occurs. Let X denote the number of flips required. Find the entropy $H(X)$ in bits. (The following expressions may be useful: $\sum_{n=0}^{\infty} r^n = 1/(1-r)$, and $\sum_{n=0}^{\infty} nr^n = r/(1-r)^2$.)

Instructor's solution: The value $p(X = k)$ is the probability that the first heads occurs on the k^{th} flip of the coin. Note that to have $X = k$ we must have a sequence of $k - 1$ tails (each with probability 1/2) followed by a flip in which we get heads (with probability 1/2). Hence we can calculate

$$\begin{aligned} p(X = k) &= \left(\frac{1}{2}\right)^{k-1} \cdot \left(\frac{1}{2}\right) \\ &= \left(\frac{1}{2}\right)^{k}. \end{aligned}$$

We can then use the definition of entropy directly to calculate $H(X)$:

$$\begin{aligned}
H(X) &= \sum_{k=1}^{\infty} p(X=k) \log \frac{1}{p(X=k)} && \text{(by def. of entropy)} \\
&= \sum_{k=1}^{\infty} \left(\frac{1}{2}\right)^k \log \frac{1}{(1/2)^k} \\
&= \sum_{k=1}^{\infty} \left(\frac{1}{2}\right)^k k \\
&= \sum_{k=0}^{\infty} \left(\frac{1}{2}\right)^k k && \text{(making the sum start from } k=0 \text{ instead of } k=1) \\
&= \frac{1/2}{(1-1/2)^2} && \text{(using } \textstyle\sum_{n=0}^{\infty} nr^n = r/(1-r)^2 \text{ for } n=k \text{ and } r={}^1\!/_2) \\
&= 2.
\end{aligned}$$

Problems (Chapter 4).

5. (MacKay 4.2) `[Easy]`

   Instructor's solution: Given in the textbook.

6. (MacKay 4.5) `[Medium]`

   Instructor's solution: Given in the textbook.

7. (MacKay 4.9) `[Easy]`

   Instructor's solution: In the weighing problem with 12 balls and the three outcome balance, before any weighing is done there are 24 possible states of the world:

   $$\{1^+, 2^+, 3^+, 4^+, 5^+, 6^+, 7^+, 8^+, 9^+, 10^+, 11^+, 12^+, 1^-, 2^-, 3^-, 4^-, 5^-, 6^-, 7^-, 8^-, 9^-, 10^-, 11^-, 12^-\},$$

   and at this point we have no information of which is the odd ball, and no information of whether the odd ball is heavier or lighter than the other ones.

   When we weigh six balls against six balls, say, $1, 2, 3, 4, 5, 6$ on the left side of the balance against $7, 8, 9, 10, 11, 12$ on the right side of the balance (the exact choice of what six balls will be on each side of the balance is irrelevant), each of the possible outcomes will rule out some states of the world.

   If the left side is heavier, we will have as remaining possible states of the world

   $$\{1^+, 2^+, 3^+, 4^+, 5^+, 6^+, 7^-, 8^-, 9^-, 10^-, 11^-, 12^-\},$$

   and if the right side is heavier, we will have as remaining possible states of the world

   $$\{1^-, 2^-, 3^-, 4^-, 5^-, 6^-, 7^+, 8^+, 9^+, 10^+, 11^+, 12^+\}.$$

   If we are interested in *which is the odd ball*, neither of the outcomes is informative because none of the 12 balls id ruled out as the odd ball after the weighing process. In other words, the process conveys no information about which ball is the odd one.

On the other hand, if we are interested in *which is the odd ball and whether it is heavy or light*, both outcomes of the weighing are informative: each rules out half of the possibilities. If the outcome is that the left side is heavier, we rule out the possibility of balls $1-6$ being lighter and of balls $7-12$ being heavier, and if the outcome is that the right side is heavier, we rule out the possibility of balls $1-6$ being heavier and of balls $7-12$ being lighter. In other words, even if neither outcome of the weighing process rules out any ball the odd one, each outcome rules out for each ball the possibility of it being heavier or lighter, in case it is the odd one. That is, the weighing process gives information on whether a ball is heavier or lighter, if it is indeed the odd ball.