Information Theory I

Fall Semester 2017

Prof. Dr. A. Lapidoth

Signal and Information Processing Laboratory

Institut für Signal- und Informationsverarbeitung

Model Answers to Exercise 2 of September 27, 2017

http://www.isi.ee.ethz.ch/teaching/courses/it1.html

Problem 1 ***Example of Joint Entropy***

a) $$H(X) = -\sum_x P_X(x) \log P_X(x) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = \log 3 - \frac{2}{3} = 0.918 \text{ bits},$$

$$H(Y) = -\sum_y P_Y(y) \log P_Y(y) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = \log 3 - \frac{2}{3} = 0.918 \text{ bits}.$$

b) We need the conditional probabilities $P_{X|Y}$ and $P_{Y|X}$. With $P_{X|Y}(x|y) = \frac{P_{X,Y}(x,y)}{P_Y(y)}$ we get

$P_{X\|Y}(x\|y)$	$x=0$	$x=1$
$y=0$	1	0
$y=1$	1/2	1/2

$P_{Y\|X}(y\|x)$	$y=0$	$y=1$
$x=0$	1/2	1/2
$x=1$	0	1

Thus, we can calculate

$$\begin{aligned} H(X|Y) &= -\sum_y P_Y(y) \sum_{x\in\text{supp}(P_{X|Y}(\cdot|y))} P_{X|Y}(x|y)\log P_{X|Y}(x|y) \\ &= -\frac{1}{3}(1\log 1) - \frac{2}{3}\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right) = \frac{2}{3} \text{ bits}, \\ H(Y|X) &= -\sum_x P_X(x) \sum_{y\in\text{supp}(P_{Y|X}(\cdot|x))} P_{Y|X}(y|x)\log P_{Y|X}(y|x) \\ &= -\frac{2}{3}\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right) - \frac{1}{3}(1\log 1) = \frac{2}{3} \text{ bits}. \end{aligned}$$

c) $H(X,Y) = 3\cdot\left(-\frac{1}{3}\log\frac{1}{3}\right) = \log 3 = 1.585$ bits.

d) $H(Y) - H(Y|X) = \log 3 - \frac{4}{3} = 0.252$ bits.

e) $I(X;Y) = H(Y) - H(Y|X) = \log 3 - \frac{4}{3} = 0.252$ bits.

Problem 2 ***Zero Conditional Entropy***

Note that $H(Y|X)$ can be written as

$$\begin{aligned} H(Y|X) &= \sum_{x\in\text{supp}(P_X)} P_X(x)\, H(Y|X=x) \\ &= \sum_{x\in\text{supp}(P_X)} P_X(x)\, H(P_{Y|X=x}) \\ &= \sum_{x\in\text{supp}(P_X)} P_X(x) \sum_{y\in\text{supp}(P_{Y|X=x})} P_{Y|X=x}(y)\log\frac{1}{P_{Y|X=x}(y)}. \end{aligned}$$

We know that the entropy of a probability mass function is zero if and only if the corresponding chance variable is deterministic. Consequently, $H(P_{Y|X=x})$ is zero if and only if Y, conditional on $X=x$, is deterministic.
If Y is a function of X, then $P_{Y|X=x}(\cdot)$ is a deterministic distribution for all $x\in\text{supp}(P_X)$, so $H(P_{Y|X=x})=0$ for all $x\in\text{supp}(P_X)$, and thus $H(Y|X)=0$.
Conversely, because entropy is nonnegative, $H(Y|X)=0$ implies $\big(H(P_{Y|X=x})=0\,\forall x\in\text{supp}(P_X)\big)$. So for every $x\in\text{supp}(P_X)$, $P_{Y|X=x}(\cdot)$ must be a deterministic distribution, and there must exist a y such that $P_{Y|X=x}(y)=1$. By setting $g(x)$ to such a y for every $x\in\text{supp}(P_X)$, we obtain a function $g(\cdot)$ such that $\Pr[Y=g(X)]=1$ holds. (The value of $g(x)$ can be chosen arbitrarily for those x with $P_X(x)=0$.) Therefore, $H(Y|X)=0$ implies that Y is a function of X.

Problem 3 ***Entropy of Functions of a Chance Variable***

a) This follows from the *chain rule.*

b) This is a consequence of Problem 2.

c) This also follows from the *chain rule.*

d) This holds because the conditional entropy is nonnegative.

Thus, applying a function to a chance variable never increases the entropy. We have equality if and only if $H(X|g(X))=0$, which is satisfied if and only if X is a function of $g(X)$ with probability one, i.e., if and only if the *restriction* of $g(\cdot)$ to the support of P_X is injective. (The restriction of $g(\cdot)$ to the support of P_X is the function $g|_{\text{supp}(P_X)}\colon \text{supp}(P_X)\to\mathcal{Y};\ x\mapsto g(x)$.)

Problem 4 ***Entropy of a Sum***

a) Observe that

$$\begin{aligned} H(X,Y,Z) &\overset{\text{(i)}}{=} H(X)+H(Y|X)+\underbrace{H(Z|X,Y)}_{=0} = H(X)+H(Y|X), \\ H(X,Y,Z) &\overset{\text{(ii)}}{=} H(X)+H(Z|X)+\underbrace{H(Y|X,Z)}_{=0} = H(X)+H(Z|X), \end{aligned}$$

where (i) and (ii) follow from the chain rule; and the underbraced terms are zero because Z is a function of the pair (X,Y) and Y is a function of the pair (X,Z). Therefore, we conclude that $H(Z|X)=H(Y|X)$.

If X and Y are independent, we have $H(Y) = H(Y|X)$, so

$$H(Y) = H(Y|X) = H(Z|X) \overset{\text{(iii)}}{\leq} H(Z),$$

where (iii) holds because conditioning does not increase entropy. Likewise, one can show that $H(X) \leq H(Z)$ if X and Y are independent.

b) Let X and Y be fair coin flips that are influenced by each other in such way that whenever X equals one, Y equals zero and the other way round, i.e., $P_{Y|X}(0|0) = 0$, $P_{Y|X}(1|0) = 1$, $P_{Y|X}(0|1) = 1$, and $P_{Y|X}(1|1) = 0$. In this case, $Z = 1$ with probability 1. Thus, $H(Z) = 0$, however, $H(X) = H(Y) = 1$ bit. Note that Y is a function of X.

c) Note that Z is a function of the pair (X, Y), so $H(Z) \leq H(X, Y)$ and

$$\begin{aligned} H(Z) &\leq H(X, Y) \\ &\overset{\text{(i)}}{=} H(X) + H(Y) - I(X; Y) \\ &\overset{\text{(ii)}}{\leq} H(X) + H(Y), \end{aligned}$$

where (i) follows from the definition of the mutual information and (ii) holds because mutual information is nonnegative. The first inequality holds with equality if and only if the pair (X, Y) can be recovered from Z with probability one. The second inequality holds with equality if and only if $I(X; Y) = 0$, i.e., if and only if X and Y are independent. Therefore, $H(Z) = H(X) + H(Y)$ if and only if X and Y are independent and the pair (X, Y) can be recovered from Z with probability one.

An example of a situation where the pair (X, Y) can be recovered from Z is $\mathcal{X} = \{0, 10\}$ and $\mathcal{Y} = \{0, \ldots, 9\}$. In this case, the mapping $\mathcal{X} \times \mathcal{Y} \to \{0, \ldots, 19\}$, $(x, y) \mapsto x + y$ is injective.

Problem 5 — *Jensen's Inequality*

Remember what Jensen's inequality states:

Lemma 1. *If f is a concave function and X is a random variable, then*

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

Moreover, if f is strictly concave, then (1) holds with equality if and only if X is deterministic. Similarly, if g is a convex function and X is a random variable, then

$$\mathsf{E}[g(X)] \geq g(\mathsf{E}[X]). \tag{2}$$

Moreover, if g is strictly convex, then (2) holds with equality if and only if X is deterministic.

Let A be a uniformly distributed random variable over the set $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$. Then,

$$\mathsf{E}[A] = \sum_{k=1}^{n} \frac{1}{n} \cdot a_k = \frac{1}{n} \sum_{k=1}^{n} a_k.$$

a) Let $f\colon \mathbb{R}^+ \to \mathbb{R}$, $x \mapsto \log x$. The function f is strictly *concave*, so

$$\begin{aligned}
\log\left(\prod_{k=1}^{n} a_k\right)^{\frac{1}{n}} &= \frac{1}{n}\log\left(\prod_{k=1}^{n} a_k\right) \\
&= \frac{1}{n}\sum_{k=1}^{n}\log a_k \\
&= \mathsf{E}[\log A] \\
&\overset{\text{(i)}}{\leq} \log \mathsf{E}[A] \\
&= \log\left(\frac{1}{n}\sum_{k=1}^{n} a_k\right).
\end{aligned}$$

Because f is strictly increasing, we have

$$\left(\prod_{k=1}^{n} a_k\right)^{\frac{1}{n}} \leq \frac{1}{n}\sum_{k=1}^{n} a_k$$

with equality if and only if (i) holds with equality. Since f is strictly concave, (i) holds with equality if and only if A is deterministic, i.e., if and only if $a_1 = a_2 = \ldots = a_n$.

b) If $\beta \geq 1$, then $f\colon \mathbb{R}^+ \to \mathbb{R}$, $x \mapsto x^\beta$ is a *convex* function. Again using the random variable A,

$$\frac{1}{n}\sum_{k=1}^{n} a_k^\beta = \mathsf{E}\left[A^\beta\right] \geq \left(\mathsf{E}[A]\right)^\beta = \left(\frac{1}{n}\sum_{k=1}^{n} a_k\right)^\beta,$$

which proves the claim.

For $0 < \beta \leq 1$, the function f is *concave*. In this case,

$$\frac{1}{n}\sum_{k=1}^{n} a_k^\beta = \mathsf{E}\left[A^\beta\right] \leq \left(\mathsf{E}[A]\right)^\beta = \left(\frac{1}{n}\sum_{k=1}^{n} a_k\right)^\beta.$$

c) Considering Part b) for $\beta = 2$, we see that $\sqrt{\frac{1}{n}\sum_{k=1}^{n} a_k^2}$ is always at least as large as $\frac{1}{n}\sum_{k=1}^{n} a_k$. For example, if your scores in six exams are 1, 2, 3, 4, 5 and 6, respectively, then $\frac{1}{n}\sum_{k=1}^{n} a_k = 3.5$, while $\sqrt{\frac{1}{n}\sum_{k=1}^{n} a_k^2} = 3.89$.