EE326 Introduction to Information Theory and Coding

- TA in charge of Homework #1: Chaedam Son (scd5929@kaist.ac.kr)

2.14

(a) We have $H(Z|X) = H(Y|X)$ because

$$\begin{aligned}H(Z|X) &= H(X+Y|X) \\ &= \sum p(x)H(x+Y|X=x) \\ &\overset{(a)}{=} \sum p(x)H(Y|X=x) \\ &= H(Y|X),\end{aligned}$$

where (a) is because $x+Y$ and Y have a one-to-one relationship. If X and Y are independent, we have

$$H(Z) \overset{(a)}{\geq} H(Z|X) \overset{(b)}{=} H(Y|X) \overset{(c)}{=} H(Y),$$

where (a) follows from the fact that conditioning reduces entropy, (b) is since $H(Z|X) = H(Y|X)$ as we just proved, and (c) follows from the independence between X and Y. Similarly, we have $H(X) \leq H(Z)$ if X and Y are independent.

(b) Let X uniformly distributed over $\{-1, +1\}$ and set $Y = -X$. Then $Z = X + Y = 0$ with probability 1. Then

$$H(Z) = 0 \quad \text{while} \quad H(X) = H(Y) = 1 \text{ bit},$$

hence $H(X) > H(Z)$ and $H(Y) > H(Z)$. One might expect that the entropy increases since $Z = X+Y$ is the sum of two random variables, but this example shows that it may not be the case if X and Y are not independent.

(c) In general, the following inequalities always hold, as long as Z is a function of X and Y.

$$H(Z) \overset{(a)}{\leq} H(X,Y,Z) = H(X,Y) + H(Z|X,Y) = H(X,Y) \overset{(b)}{\leq} H(X) + H(Y).$$

We have $H(Z) = H(X) + H(Y)$ when both (a) and (b) are tight. Note that (a) becomes tight when $H(X,Y|Z) = 0$, i.e., (X,Y) can be exactly inferred from Z, and (b) is tight when X and Y are independent. An example that satisfies these two conditions is as follows: X and Y are uniformly distributed over $\{0,2\}$ and $\{0,1\}$, respectively, and they are independent.

2.29

(a) We can express $H(X, Y \mid Z)$ through the following relation:

$$H(X, Y \mid Z) \overset{(a)}{=} H(X \mid Z) + H(Y \mid X, Z) \overset{(b)}{\geq} H(X \mid Z),$$

where (a) is because Chain rule for entropy, (b) is since non-negativity of entropy. That is, the inclusion of an additional random variable is non decreasing the entropy. This is make sense, since adding the information of Y can only increase (or maintain) the total information.
Equality iff $H(Y \mid X, Z) = 0$. That is, Y carries no additional information once X and Z are given. (there exists a situation where $Y = f(X, Z)$).

(b) We can express $I(X, Y; Z)$ through the following relation:

$$I(X, Y; Z) \overset{(a)}{=} I(X; Z) + I(Y; Z \mid X) \overset{(b)}{\geq} I(X; Z),$$

where (a) is because Chain rule for mutual information, (b) follows nonnegativity of mutual information.
Similar to 2.29 (a), when considering the mutual information of random variables, the introduction of an additional variable never reduces the mutual information. This also makes sense, since the newly introduced variable simply adds further information to be considered.
Equality iff $I(Y; Z \mid X) = 0$ (i.e., $Y \perp Z \mid X$). For example, in a Markov chain $Y - X - Z$, once X is given, Y and Z become independent.

(c) We can express $H(X, Y, Z) - H(X, Y)$ through the following relation:

$$H(X, Y, Z) - H(X, Y) = {}^{(a)} H(Z \mid X, Y) \leq {}^{(b)} H(Z \mid X) = {}^{(a)} H(X, Z) - H(X),$$

where (a) is since chain rule for entropy, (b) follows conditioning reduces entropy.
Equality iff $I(Z; Y \mid X) = 0$ (i.e., $Z \perp Y \mid X$). For example, in a Markov chain $Y - X - Z$, once X is given, Y and Z become independent.

(d) We can express $I(X; Z \mid Y)$ through the following relation:

$$I(X; Z \mid Y) + I(Z; Y) =^{(a)} I(X, Y; Z) =^{(a)} I(Z; Y \mid X) + I(X; Z)$$

where (a) is Chain rule for mutual information.

Hence the displayed "$\geq$" actually holds with **equality for all** (X, Y, Z) (no extra assumptions needed).

2.32

(a) To find MAP estimator and $P_e^{\min}$, for each y,

$$P(X\!=\!1 \mid Y\!=\!a) = \tfrac{(1/6)}{(1/3)} = \tfrac{1}{2}, \quad P(X\!=\!2 \mid Y\!=\!a) = P(X\!=\!3 \mid Y\!=\!a) = \tfrac{1}{4}.$$

Similarly, $P(X=2 \mid Y=b) = \frac{1}{2}$ and $P(X=3 \mid Y=c) = \frac{1}{2}$ (others $= \frac{1}{4}$). The MAP estimator chooses, for each observed y, the value of x that maximizes the posterior probability:

$$\hat{X}(y) \;=\; \arg\max_{x \in \mathcal{X}} \; P(X = x \mid Y = y).$$

Therefore,

$$\hat{X}(a) = 1, \qquad \hat{X}(b) = 2, \qquad \hat{X}(c) = 3.$$

The minimal error probability is

$$P_e^{\min} = \sum_y p_Y(y)\big(1 - \max_x P(X = x \mid Y = y)\big) = 3 \cdot \Big(\tfrac{1}{3} \cdot (1 - \tfrac{1}{2})\Big) = \tfrac{1}{2}.$$

(b) By the Fano's inequality, probablity of error can be bounded as

$$P_e \geq \frac{H(X|Y) - 1}{\log|\mathcal{X}|}.$$

Here,

$$\begin{aligned} H(X|Y) &= H(X|Y=a)p_Y(a) + H(X|Y=b)p_Y(b) + H(X|Y=c)p_Y(c) \\ &= H\big(\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{4}\big)\, p_Y(a) + H\big(\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{4}\big)\, p_Y(b) + H\big(\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{4}\big)\, p_Y(c) \\ &= H\big(\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{4}\big)\, \big(p_Y(a) + p_Y(b) + p_Y(c)\big) \\ &= H\big(\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{4}\big) \\ &= 1.5 \text{ bits.} \end{aligned}$$

Hence

$$P_e \geq \frac{1.5 - 1}{\log 3} = 0.316.$$

Therefore, our estimator $\hat{X}(Y)$ is not very close to Fano's bound in this form. If $\hat{X} \in \mathcal{X}$, as it does here, we can use the stronger form of Fano's inequality to get (By selecting one element, the overall size can be reduced.)

$$P_e \geq \frac{H(X|Y) - 1}{\log(|\mathcal{X}| - 1)}.$$

and

$$P_e \geq \frac{1.5 - 1}{\log 2} = \tfrac{1}{2}.$$

which matches the MAP error in (a). Hence Fano's bound is *tight* in this problem.

2.43

(a) Let X denote the outcome on the *top* side of a fair coin, Y the outcome on the *bottom* side. If $X = \mathrm{H}$ then $Y = \mathrm{T}$ deterministically, and vice versa. Thus Y is a deterministic function of X.

$$I(X;Y) = H(Y) - H(Y \mid X).$$

Here $H(Y) = 1$ bit (fair coin), and $H(Y \mid X) = 0$ since Y is determined by X. So

$$I(X;Y) = 1 \text{ bit}.$$

Intuition: knowing the top outcome tells you the bottom outcome with certainty, so mutual information equals the entropy of one coin face.

(b) Let X = top side, Z = front face. On a standard cube, once the top side X is fixed, the front face Z is uniformly distributed among 4 possibilities (the sides adjacent to the top). Thus

$$P(Z = z \mid X = x) = \tfrac{1}{4} \quad \text{for 4 values of } z, \quad 0 \text{ otherwise}.$$

So $H(Z \mid X) = \log_2 4 = 2$ bits. Since marginally Z is uniform over 6 faces, $H(Z) = \log_2 6$ bits.

Hence

$$I(X;Z) = H(Z) - H(Z \mid X) = \log_2 6 - 2 \approx 0.585 \text{ bits}.$$

Intuition: knowing the top side rules out 2 of the 6 faces for the front (the top and bottom), leaving 4 possibilities. So your uncertainty drops from $\log_2 6$ to $\log_2 4$, yielding about 0.585 bits of information.

3.4

Setup. $X_i \overset{\text{iid}}{\sim} p(x)$ on alphabet $\mathcal{X} = \{1, \ldots, m\}$. Define

$$A^n = \left\{ x^n : \left| -\tfrac{1}{n} \log p(x^n) - H \right| \le \epsilon \right\}, \qquad B^n = \left\{ x^n : \left| \tfrac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| \le \epsilon \right\}.$$

(a) Does. By the *Asymptotic Equipartition Property (AEP)* ,

$$-\tfrac{1}{n} \log p(X^n) \xrightarrow{a.s.} H.$$

Thus for any $\epsilon > 0$,

$$P\{X^n \in A^n\} \to 1.$$

(b) Does. By 3.4 (a), and the Strong Law of Large Numbers $P\{X^n \in B^n\} \to 1$. So there exists $\epsilon > 0$ and N_1 such that

$$P\{X^n \in A^n\} > 1 - \tfrac{\epsilon}{2} \quad \text{for all } n > N_1,$$

and there exists N_2 such that

$$P\{X^n \in B^n\} > 1 - \tfrac{\epsilon}{2} \quad \text{for all } n > N_2.$$

So for all $n > \max(N_1, N_2)$:

$$\begin{aligned} P\{X^n \in A^n \cap B^n\} &= P\{X^n \in A^n\} + P\{X^n \in B^n\} - P\{X^n \in A^n \cup B^n\} \\ &> 1 - \tfrac{\epsilon}{2} + 1 - \tfrac{\epsilon}{2} - 1 \\ &= 1 - \epsilon. \end{aligned}$$

So for any $\epsilon > 0$ there exists $N = \max(N_1, N_2)$ such that

$$P\{X^n \in A^n \cap B^n\} > 1 - \epsilon \quad \text{for all } n > N,$$

therefore $P\{X^n \in A^n \cap B^n\} \to 1$. In conclusion, the probability that a sequence X^n satisfies both the law of large numbers and the AEP in the large n regime is almost 1, i.e., essentially certain.

(c) For any $x^n \in A^n$, we have

$$2^{-n(H+\epsilon)} \le p(x^n) \le 2^{-n(H-\epsilon)}.$$

Also $\sum_{x^n \in A^n \cap B^n} p(x^n) \le 1$. Hence

$$|A^n \cap B^n| \cdot 2^{-n(H+\epsilon)} \le 1,$$

so

$$|A^n \cap B^n| \le 2^{n(H+\epsilon)}.$$

That is, we can see that it is sufficient to consider only $2^{n(H+\epsilon)}$ sequences, rather than all possible sequences in $\mathcal{X}^n$, to evaluate $|A^n \cap B^n|$. This implies that, for a sufficiently large n, the number of feasible sequences of such a random process is limited to $2^{n(H+\epsilon)}$, and the others can be neglected.

(d) From 3.4 (b), $P\{X^n \in A^n \cap B^n\} \to 1$. So for sufficiently large n,

$$P\{X^n \in A^n \cap B^n\} \ge \tfrac{1}{2}.$$

For $x^n \in A^n$, $p(x^n) \le 2^{-n(H-\epsilon)}$. Hence

$$P\{X^n \in A^n \cap B^n\} = \sum_{x^n \in A^n \cap B^n} p(x^n) \le |A^n \cap B^n| \cdot 2^{-n(H-\epsilon)}.$$

Therefore

$$\tfrac{1}{2} \le |A^n \cap B^n| \cdot 2^{-n(H-\epsilon)} \quad \Rightarrow \quad |A^n \cap B^n| \ge \tfrac{1}{2}\, 2^{n(H-\epsilon)}.$$

Combining the conclusion of (c), $|A^n \cap B^n| \le 2^{n(H+\epsilon)}$, we obtain

$$2^{n(H-\epsilon)} \;\le\; |A^n \cap B^n| \;\le\; 2^{n(H+\epsilon)}.$$

This means that, when n is sufficiently large, it is asymptotically enough to consider about 2^{nH} sequences of a random process. Since this provides the key intuition and idea for the subsequent discussions, it is worth making a remark here.

3.7

(a) The number of 100-bit sequences with ≤ 3 ones is

$$N = \sum_{k=0}^{3} \binom{100}{k} = 166{,}751.$$

With equal-length codewords we need at least

$$\ell_{\min} = \lceil \log_2 N \rceil = \lceil \log_2 166{,}751 \rceil = 18 \text{ bits}.$$

(b) Let $X \sim \text{Binomial}(n = 100, p = 0.005)$ be the number of 1's. No codeword $\iff X \ge 4$. Hence

$$P(\text{no codeword}) = P(X \ge 4) = 1 - \sum_{k=0}^{3} \binom{100}{k} (0.005)^k (0.995)^{100-k}.$$

Numerically,

$$P(X \ge 4) \approx 0.0020 \quad (\text{about } 2.0 \times 10^{-3}).$$

(Using the Poisson approximation with $\lambda = np = 0.5$ gives $1 - e^{-\lambda} \sum_{k=0}^{3} \lambda^k / k! \approx 0.00175$, very close.).
In conclusion, by combining the result of 3.7 (a), we find that only 18 bits are sufficient to construct codewords for all sequences that lie within the probability of 0.998. This demonstrates a significant advantage compared to the 100 bits required when constructing codewords for the entire set of sequences.

(c) Here $\mu = E[X] = np = 0.5$, $\sigma^2 = \text{VAR}(X) = np(1-p) = 0.4975$. Since $X \ge 4 \iff X - \mu \ge 3.5$, Chebyshev's inequality (two-sided) gives

$$P(X \ge 4) \le P\big(|X - \mu| \ge 3.5\big) \le \frac{\sigma^2}{3.5^2} = \frac{0.4975}{12.25} \approx 0.0406.$$

Comparison: Chebyshev's bound $\approx 4.06\%$ is much looser than the true probability $\approx 0.20\%$ from part 3.7 (b).