Solutions for Homework 2

1. (a) False.

   Conditional independence does not imply independence. Since given that

$$f_{X,Y|Z}(x,y|z) = f_{X|Z}(x|z)f_{Y|Z}(y|z)$$

   we can only have

$$\frac{f_{X,Y,Z}(x,y,z)}{f_Z(z)} = \frac{f_{X,Z}(x,z)}{f_Z(z)}\frac{f_{Y,Z}(y,z)}{f_Z(z)}$$

   which does not imply that

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

   hence we cannot conclude that X and Y are independent.

   (b) False.

   In general, $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$ is not sufficient for showing that X, Y are independent. You must show that $f_{XY}(x,y) = f_X(x)f_Y(y)$, i.e. that the joint pdf can be written as the product of the marginal pdf's.

   (c) True.

   If X and Y are independent, then $\mathrm{cov}(X,Y) = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] = 0$.
   Therefore, $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$.

   (d) True.

   Uncorrelated jointly Gaussian random variables are independent.

   (e) True.

$$\begin{aligned}
\mathrm{VAR}(X+Y) &= \mathrm{E}[((X+Y) - \mathrm{E}(X+Y))^2] \\
&= \mathrm{E}[X^2 + 2XY + Y^2 - 2(X+Y)\mathrm{E}(X+Y) + \mathrm{E}^2(X+Y)] \\
&= \mathrm{E}[X^2] + 2\mathrm{E}[XY] + \mathrm{E}[Y^2] - \mathrm{E}^2[X+Y] \\
&= \mathrm{E}[X^2] + 2\mathrm{E}[XY] + \mathrm{E}[Y^2] - (\mathrm{E}[X] + \mathrm{E}[Y])^2 \\
&= \mathrm{E}[X^2] + 2\mathrm{E}[XY] + \mathrm{E}[Y^2] - \mathrm{E}^2[X] - 2\mathrm{E}[X]\mathrm{E}[Y] - \mathrm{E}^2[Y] \\
&= (\mathrm{E}[X^2] - \mathrm{E}^2[X]) + (\mathrm{E}[Y^2] - \mathrm{E}^2[Y]) + 2(\mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]) \\
&= \mathrm{VAR}(X) + \mathrm{VAR}(Y) + 2\mathrm{COV}(X,Y).
\end{aligned}$$

2. First, note from the covariance matrix we have

$$\begin{aligned}
\sigma_x^2 &= E[(X-\mu_x)^2] \\
&= E[X^2] - \mu_x^2 \\
\sigma_y^2 &= E[(Y-\mu_y)^2] \\
&= E[Y^2] - \mu_y^2 \\
\rho &= E[(X-\mu_x)(Y-\mu_y)] \\
&= E[XY] - \mu_x\mu_y.
\end{aligned}$$

(a)

$$\begin{aligned} E[(X-Y)^2] &= E[X^2 - 2XY + Y^2] \\ &= E[X^2] - 2E[XY] + E[Y^2] \\ &= (\sigma_x^2 + \mu_x^2) - 2(\rho + \mu_x\mu_y) + (\sigma_y^2 + \mu_y^2). \end{aligned}$$

(b)

$$\begin{aligned} E[5X^2 + 2] &= 5E[X^2] + 2 \\ &= 5(\sigma_x^2 + \mu_x^2) + 2. \end{aligned}$$

(c) This can be solved a variety of ways (e.g. by calculating the pdf explicitly, or using characteristic functions). Here, we will use the known property that for jointly Gaussian random variables, the conditional expectation is affine functions of the observed variables. Hence, we know $E[X|Y] = a_0 + a_1 Y$ for some a_0, a_1. To find a_0, a_1, we use successive conditioning to give

$$\begin{aligned} \mu_x &= E[X] \\ &= E[E[X|Y]] \\ &= E[a_0 + a_1 Y] \\ &= a_0 + a_1 \mu_y. \qquad (1) \end{aligned}$$

Next, for jointly Gaussian rv's we know that $COV(X - E[X|Y], Y) = 0$, or

$$\begin{aligned} 0 &= COV(X - E[X|Y], Y) \\ &= E[(X - a_0 - a_1 Y)(Y - \mu_y)] \\ &= \rho - a_1 \sigma_y^2. \qquad (2) \end{aligned}$$

Solving (1) and (2) gives $a_0 = \mu_x - \mu_y \frac{\rho}{\sigma_y^2}$, $a_1 = \frac{\rho}{\sigma_y^2}$. Hence,

$$\begin{aligned} E[X|Y] &= a_0 + a_1 Y \\ &= \mu_x + \frac{\rho}{\sigma_y^2}\left(Y - \mu_y\right). \end{aligned}$$

3.

(a) The number X of tosses till the first head appears has the geometric distribution with parameter $p = 1/2$, where $P(X = n) = pq^{n-1}$, $n \in \{1, 2, \ldots\}$. Hence the entropy of X is

$$\begin{aligned}
H(X) &= -\sum_{n=1}^{\infty} pq^{n-1} \log(pq^{n-1}) \\
&= -\left[\sum_{n=0}^{\infty} pq^{n} \log p + \sum_{n=0}^{\infty} npq^{n} \log q\right] \\
&= \frac{-p \log p}{1-q} - \frac{pq \log q}{p^2} \\
&= \frac{-p \log p - q \log q}{p} \\
&= H(p)/p \text{ bits.}
\end{aligned}$$

If $p = 1/2$, then $H(X) = 2$ bits.

(b) Intuitively, it seems clear that the best questions are those that have equally likely chances of receiving a yes or a no answer. Consequently, one possible guess is that the most "efficient" series of questions is: Is $X = 1$? If not, is $X = 2$? If not, is $X = 3$? ...with a resulting expected number of questions equal to $\sum_{n=1}^{\infty} n(1/2^n) = 2$. This should reinforce the intuition that $H(X)$ is a measure of the uncertainty of X. Indeed in this case, the entropy is exactly the same as the average number of questions needed to define X, and in general $E(\#\text{ of questions}) \geq H(X)$. This problem has an interpretation as a source coding problem. Let $0 =$no, $1 =$yes, $X =$Source, and $Y =$Encoded Source. Then the set of questions in the above procedure can be written as a collection of (X, Y) pairs: $(1, 1)$, $(2, 01)$, $(3, 001)$, etc. . In fact, this intuitively derived code is the optimal (Huffman) code minimizing the expected number of questions.

4.

Solution: *Entropy of functions of a random variable.*

(a) $H(X, g(X)) = H(X) + H(g(X)|X)$ by the chain rule for entropies.

(b) $H(g(X)|X) = 0$ since for any particular value of X, g(X) is fixed, and hence $H(g(X)|X) = \sum_x p(x) H(g(X)|X = x) = \sum_x 0 = 0$.

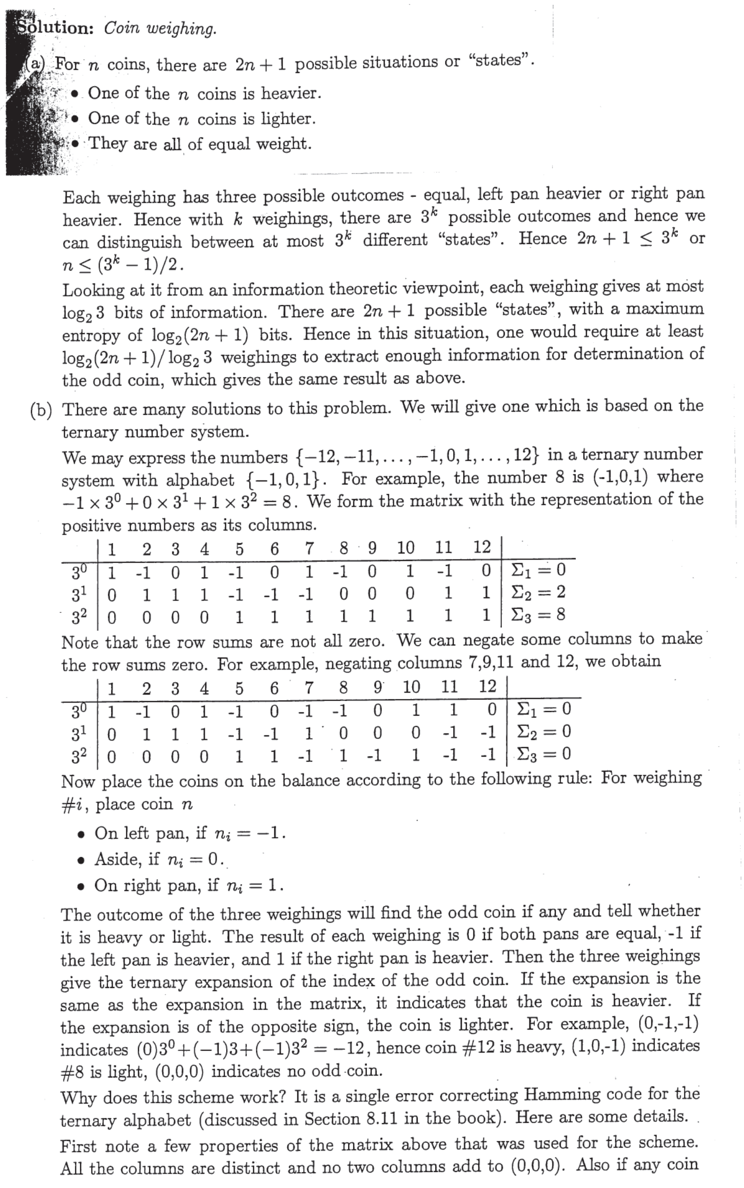
(c) $H(X, g(X)) = H(g(X)) + H(X|g(X))$ again by the chain rule.

(d) $H(X|g(X)) \geq 0$, with equality iff X is a function of $g(X)$, i.e., $g(.)$ is one-to-one. Hence $H(X, g(X)) \geq H(g(X))$.

Combining parts (b) and (d), we obtain $H(X) \geq H(g(X))$.

5. Aside for part(a):

For n coins, there are $2n + 1$ possible situations or "states", for example $n = 3$ coins, then you know that the possibilities are one of the following: coin 1 is the heavy coin.
coin 1 is the light coin.
coin 2 is the heavy coin.
coin 2 is the light coin.
coin 3 is the heavy coin.
coin 3 is the light coin.
all coins are the same weight.
So there are 7 states.



Solution: *Coin weighing.*

(a) For n coins, there are $2n+1$ possible situations or "states".

- One of the n coins is heavier.
- One of the n coins is lighter.
- They are all of equal weight.

Each weighing has three possible outcomes - equal, left pan heavier or right pan heavier. Hence with k weighings, there are 3^k possible outcomes and hence we can distinguish between at most 3^k different "states". Hence $2n+1 \leq 3^k$ or $n \leq (3^k - 1)/2$.

Looking at it from an information theoretic viewpoint, each weighing gives at most $\log_2 3$ bits of information. There are $2n+1$ possible "states", with a maximum entropy of $\log_2(2n+1)$ bits. Hence in this situation, one would require at least $\log_2(2n+1)/\log_2 3$ weighings to extract enough information for determination of the odd coin, which gives the same result as above.

(b) There are many solutions to this problem. We will give one which is based on the ternary number system.

We may express the numbers $\{-12, -11, \ldots, -1, 0, 1, \ldots, 12\}$ in a ternary number system with alphabet $\{-1, 0, 1\}$. For example, the number 8 is (-1,0,1) where $-1 \times 3^0 + 0 \times 3^1 + 1 \times 3^2 = 8$. We form the matrix with the representation of the positive numbers as its columns.

	1	2	3	4	5	6	7	8	9	10	11	12	
3^0	1	-1	0	1	-1	0	1	-1	0	1	-1	0	$\Sigma_1 = 0$
3^1	0	1	1	1	-1	-1	-1	0	0	0	1	1	$\Sigma_2 = 2$
3^2	0	0	0	0	1	1	1	1	1	1	1	1	$\Sigma_3 = 8$

Note that the row sums are not all zero. We can negate some columns to make the row sums zero. For example, negating columns 7,9,11 and 12, we obtain

	1	2	3	4	5	6	7	8	9	10	11	12	
3^0	1	-1	0	1	-1	0	-1	-1	0	1	1	0	$\Sigma_1 = 0$
3^1	0	1	1	1	-1	-1	1	0	0	0	-1	-1	$\Sigma_2 = 0$
3^2	0	0	0	0	1	1	-1	1	-1	1	-1	-1	$\Sigma_3 = 0$

Now place the coins on the balance according to the following rule: For weighing $\#i$, place coin n

- On left pan, if $n_i = -1$.
- Aside, if $n_i = 0$.
- On right pan, if $n_i = 1$.

The outcome of the three weighings will find the odd coin if any and tell whether it is heavy or light. The result of each weighing is 0 if both pans are equal, -1 if the left pan is heavier, and 1 if the right pan is heavier. Then the three weighings give the ternary expansion of the index of the odd coin. If the expansion is the same as the expansion in the matrix, it indicates that the coin is heavier. If the expansion is of the opposite sign, the coin is lighter. For example, (0,-1,-1) indicates $(0)3^0+(-1)3+(-1)3^2 = -12$, hence coin #12 is heavy, (1,0,-1) indicates #8 is light, (0,0,0) indicates no odd coin.

Why does this scheme work? It is a single error correcting Hamming code for the ternary alphabet (discussed in Section 8.11 in the book). Here are some details.

First note a few properties of the matrix above that was used for the scheme. All the columns are distinct and no two columns add to (0,0,0). Also if any coin is heavier, it will produce the sequence of weighings that matches its column in the matrix. If it is lighter, it produces the negative of its column as a sequence of weighings. Combining all these facts, we can see that any single odd coin will produce a unique sequence of weighings, and that the coin can be determined from the sequence.

One of the questions that many of you had whether the bound derived in part (a) was actually achievable. For example, can one distinguish 13 coins in 3 weighings? No, not with a scheme like the one above. Yes, under the assumptions under which the bound was derived. The bound did not prohibit the division of coins into halves, neither did it disallow the existence of another coin known to be normal. Under both these conditions, it is possible to find the odd coin of 13 coins in 3 weighings. You could try modifying the above scheme to these cases.

6.

Solution: *Drawing with and without replacement.* Intuitively, it is clear that if the balls are drawn with replacement, the number of possible choices for the i-th ball is larger, and therefore the conditional entropy is larger. But computing the conditional distributions is slightly involved. It is easier to compute the unconditional entropy.

- With replacement. In this case the conditional distribution of each draw is the same for every draw. Thus

$$X_i = \begin{cases} \text{red} & \text{with prob.}\frac{r}{r+w+b} \\ \text{white} & \text{with prob.}\frac{w}{r+w+b} \\ \text{black} & \text{with prob.}\frac{b}{r+w+b} \end{cases} \qquad (2.8)$$

and therefore

$$H(X_i|X_{i-1},\ldots,X_1) = H(X_i) \qquad (2.9)$$

$$= \log(r+w+b) - \frac{r}{r+w+b}\log r - \frac{w}{r+w+b}\log w - \frac{b}{r+w+b}\log b \qquad (2.10)$$

- Without replacement. The unconditional probability of the i-th ball being red is still $r/(r+w+b)$, etc. Thus the unconditional entropy $H(X_i)$ is still the same as with replacement. The conditional entropy $H(X_i|X_{i-1},\ldots,X_1)$ is less than the unconditional entropy, and therefore the entropy of drawing without replacement is lower.

7.

(a) $H(X) = \frac{2}{3}\log\frac{3}{2} + \frac{1}{3}\log 3 = 0.918$ bits $= H(Y)$.

(b) $H(X|Y) = \frac{1}{3}H(X|Y=0) + \frac{2}{3}H(X|Y=1) = 0.667$ bits $= H(Y|X)$.

(c) $H(X,Y) = 3 \times \frac{1}{3}\log 3 = 1.585$ bits.

(d) $H(Y) - H(Y|X) = 0.251$ bits.

(e) $I(X;Y) = H(Y) - H(Y|X) = 0.251$ bits.

8. (a)

For $n = 3$,

$$\begin{aligned} P(X_3 = 1) &= P(X_1 = X_2) \\ &= (P(X_1 = 1) \cap P(X_2 = 1)) \cup (P(X_1 = 0) \cap P(X_2 = 0)) \\ &= \frac{1}{4} + \frac{1}{4} \\ &= \frac{1}{2} \\ P(X_3 = 0) &= P(X_1 \neq X_2) \\ &= 1 - P(X_1 = X_2) \\ &= \frac{1}{2} \end{aligned}$$

Hence

$$\begin{aligned} H(X_3) &= -P(X_3=1)\log_2 P(X_3=1) - P(X_3=0)\log_2 P(X_3=0) \\ &= 1 \end{aligned}$$

Similarly, we have $P(X_2 = 1) = P(X_3 = 1) = \frac{1}{2}, H(X_4) = 1, P(X_3 = 1) = P(X_4 = 1) = \frac{1}{2}, H(X_5) = 1, \cdots$. According to the symmetry we can conclude that $X_n \sim$ Bernoulli($\frac{1}{2}$). Thus, $H(X_n) = 1$.

(b)

Since X_n is a function of X_{n-1}, X_{n-2} and from part(a) we know that X_{n-1}, X_{n-2} are functions of X_1, X_2. Therefore $X_n = g(X_1, X_2)$. Since $g(.)$ is a one-to-one function, X_n is determined when X_1, X_2 are known. Hence there is no uncertainty in X_n given X_1, X_2, i.e., $H(X_n|X_1, X_2) = 0$.

9. (a)

$$\begin{aligned} H(X|XY) &\overset{(a)}{=} H(X, XY) - H(XY) \\ &\overset{(b)}{=} H(X, Y, XY) - H(Y|X, XY) - H(XY) \\ &\overset{(c)}{=} H(X, Y) - H(Y|X, XY) - H(XY) \\ &\overset{(d)}{=} H(X, Y) - H(XY) - Pr(X = 0)H(Y|X = 0, XY) - Pr(X \neq 0)H(Y|X \neq 0, XY) \\ &\overset{(e)}{=} H(X, Y) - H(XY) - Pr(X = 0)H(Y|X = 0) - Pr(X \neq 0)H(Y|X \neq 0, XY) \\ &\overset{(f)}{=} H(X, Y) - H(XY) - Pr(X = 0)H(Y|X = 0) \end{aligned}$$

where each step is justified as follows:

(a) chain rule
(b) chain rule
(c) XY is a function of X and Y
(d) conditioning on whether $X = 0$ or $X \neq 0$
(e) if $X = 0$, then $XY = 0$ regardless of Y, so the XY term can be dropped from the conditioning in $H(Y|X = 0, XY)$
(f) if $X \neq 0$, then we can determine Y given X and XY, so $H(Y|X \neq 0, XY) = 0$.

Thus we conclude that $H(X|XY) \leq H(X, Y) - H(XY)$.

(b)

We see that equality holds if and only if $Pr(X = 0)H(Y|X = 0)$ equals zero. Thus, either $Pr(X = 0) = 0$, or $H(Y|X = 0) = 0$.