Solutions for Homework 3

1. Let X be the r.v. indicating if it did not rain. Then X is a Bernoulli r.v. with parameter 3/16+10/16=13/16. Let W be the r.v. indicating whether the weatherperson predicts no rain; W is Bernoulli with parameter 1/16+10/16=11/16. Also note that the following conditional probabilities can be inferred from the table provided in the problem statement:

$$\begin{aligned} P(X = \text{rain}|W = \text{rain}) &= 2/5 \\ P(X = \text{no rain}|W = \text{rain}) &= 3/5 \\ P(X = \text{rain}|W = \text{no rain}) &= 1/11 \\ P(X = \text{no rain}|W = \text{no rain}) &= 10/11 \end{aligned}$$

The uncertainty of the true weather given the weatherperson's prediction is given by:

$$\begin{aligned} H(X|W) &= 5/16\, H(X|W = \text{rain}) + 11/16\, H(X|W = \text{no rain}) \\ &= 5/16\, h(3/5) + 11/16\, h(10/11) \\ &= 0.6056 \text{ bits.} \end{aligned}$$

If, on the other hand, you were always to declare that there would be no rain (i.e. acting as a "deterministic" weatherperson), then the uncertainty of the true weather from your "clever" prediction is:

$$\begin{aligned} H(X|W) &= H(X) \\ &= h(13/16) \\ &= 0.6962 \text{ bits.} \end{aligned}$$

Hence, there is more uncertainty about the true weather under the strategy of always declaring it not to rain than there is in listening to the WBZ-TV weatherperson. The station manager is right.

Identical arguments can be made by calculating the mutual information, or reduction in uncertainty of one variable due to knowledge of another. We can compute this for the weatherperson as

$$I(X;W) = h(13/16) - 5/16\, h(3/5) - 11/16\, h(10/11) = 0.0906 \text{ bits}$$

but under your strategy of always declaring no rain, the mutual information between X and your prediction is zero. So, your clever strategy does not reduce the uncertainty of the forecast at all, while the WBZ weatherperson reduces the uncertainty by 0.0906 bits.

2. **Solution:** *Data Processing.* By the chain rule for mutual information,

$$I(X_1; X_2, \ldots, X_n) = I(X_1; X_2) + I(X_1; X_3|X_2) + \cdots + I(X_1; X_n|X_2, \ldots, X_{n-2}). \quad (2.20)$$

By the Markov property, the past and the future are conditionally independent given the present and hence all terms except the first are zero. Therefore

$$I(X_1; X_2, \ldots, X_n) = I(X_1; X_2). \quad (2.21)$$

3. **Solution:** *Conditional mutual information.*

Consider a sequence of n binary random variables $X_1, X_2, \ldots, X_n$. Each sequence of length n with an even number of 1's is equally likely and has probability $2^{-(n-1)}$.

Any $n-1$ or fewer of these are independent. Thus, for $k \leq n-1$,

$$I(X_{k-1}; X_k|X_1, X_2, \ldots, X_{k-2}) = 0.$$

However, given $X_1, X_2, \ldots, X_{n-2}$, we know that once we know either X_{n-1} or X_n we know the other.

$$\begin{aligned} I(X_{n-1}; X_n|X_1, X_2, \ldots, X_{n-2}) &= H(X_n|X_1, X_2, \ldots, X_{n-2}) - H(X_n|X_1, X_2, \ldots, X_{n-1}) \\ &= 1 - 0 = 1 \text{ bit.} \end{aligned}$$

Note that

$$I(X_1; X_2) = \begin{cases} 1 & \text{when n=2} \\ 0 & \text{when n} > 2 \end{cases}$$

4. **Solution:** *Inequalities.*

(a) Using the chain rule for conditional entropy,

$$H(X,Y|Z) = H(X|Z) + H(Y|X,Z) \geq H(X|Z),$$

with equality iff $H(Y|X,Z) = 0$, that is, when Y is a function of X and Z.

(b) Using the chain rule for mutual information,

$$I(X,Y;Z) = I(X;Z) + I(Y;Z|X) \geq I(X;Z),$$

with equality iff $I(Y;Z|X) = 0$, that is, when Y and Z are conditionally independent given X.

(c) Using first the chain rule for entropy and then the definition of conditional mutual information,

$$\begin{aligned} H(X,Y,Z) - H(X,Y) &= H(Z|X,Y) = H(Z|X) - I(Y;Z|X) \\ &\leq H(Z|X) = H(X,Z) - H(X), \end{aligned}$$

with equality iff $I(Y;Z|X) = 0$, that is, when Y and Z are conditionally independent given X.

(d) Using the chain rule for mutual information,

$$I(X;Z|Y) + I(Z;Y) = I(X,Y;Z) = I(Z;Y|X) + I(X;Z),$$

and therefore

$$I(X;Z|Y) = I(Z;Y|X) - I(Z;Y) + I(X;Z).$$

We see that this inequality is actually an equality in all cases.

5.

(a) From inspection we see that

$$\hat{X}(y) = \begin{cases} 1 & y = a \\ 2 & y = b \\ 3 & y = c \end{cases}$$

Hence the associated P_e is the sum of $P(1,b)$, $P(1,c)$, $P(2,a)$, $P(2,c)$, $P(3,a)$ and $P(3,b)$. Therefore, $P_e = 1/2$.

(b) From Fano's inequality we know

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

Here,

$$\begin{aligned} H(X|Y) &= H(X|Y=a)\Pr\{y=a\} + H(X|Y=b)\Pr\{y=b\} + H(X|Y=c)\Pr\{y=c\} \\ &= H\left(\frac{1}{2},\frac{1}{4},\frac{1}{4}\right)\Pr\{y=a\} + H\left(\frac{1}{2},\frac{1}{4},\frac{1}{4}\right)\Pr\{y=b\} + H\left(\frac{1}{2},\frac{1}{4},\frac{1}{4}\right)\Pr\{y=c\} \\ &= H\left(\frac{1}{2},\frac{1}{4},\frac{1}{4}\right)(\Pr\{y=a\} + \Pr\{y=b\} + \Pr\{y=c\}) \\ &= H\left(\frac{1}{2},\frac{1}{4},\frac{1}{4}\right) \\ &= 1.5 \text{ bits.} \end{aligned}$$

Hence

$$P_e \geq \frac{1.5 - 1}{\log 3} = .316.$$

Hence 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

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

and

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

Therefore our estimator $\hat{X}(Y)$ is actually quite good.

6.

Solution:

$$\begin{aligned}
\frac{1}{n}\log\frac{p(X^n)p(Y^n)}{p(X^n,Y^n)} &= \frac{1}{n}\log\prod_{i=1}^{n}\frac{p(X_i)p(Y_i)}{p(X_i,Y_i)} \\
&= \frac{1}{n}\sum_{i=i}^{n}\log\frac{p(X_i)p(Y_i)}{p(X_i,Y_i)} \\
&\to E(\log\frac{p(X_i)p(Y_i)}{p(X_i,Y_i)}) \\
&= -I(X;Y)
\end{aligned}$$

Thus, $\frac{p(X^n)p(Y^n)}{p(X^n,Y^n)} \to 2^{-nI(X;Y)}$, which will converge to 1 if X and Y are indeed independent.

7.

Solution: *An AEP-like limit.* $X_1, X_2, \ldots$, i.i.d. $\sim p(x)$. Hence $\log(X_i)$ are also i.i.d. and

$$\begin{aligned}
\lim(p(X_1,X_2,\ldots,X_n))^{\frac{1}{n}} &= \lim 2^{\log(p(X_1,X_2,\ldots,X_n))^{\frac{1}{n}}} \\
&= 2^{\lim\frac{1}{n}\sum\log p(X_i)} \text{ a.e.} \\
&= 2^{E(\log(p(X)))} \text{ a.e.} \\
&= 2^{-H(X)} \text{ a.e.}
\end{aligned}$$

by the strong law of large numbers (assuming of course that H(X) exists).

8.

Solution: *Time's arrow.* By the chain rule for entropy,

$$\begin{aligned} H(X_0|X_{-1},\ldots,X_{-n}) &= H(X_0,X_{-1},\ldots,X_{-n}) - H(X_{-1},\ldots,X_{-n}) && (4.8)\\ &= H(X_0,X_1,X_2,\ldots,X_n) - H(X_1,X_2,\ldots,X_n) && (4.9)\\ &= H(X_0|X_1,X_2,\ldots,X_n), && (4.10) \end{aligned}$$

where (4.9) follows from stationarity.

9.

Solution: *Initial conditions.* For a Markov chain, by the data processing theorem, we have

$$I(X_0;X_{n-1}) \geq I(X_0;X_n). \tag{4.62}$$

Therefore

$$H(X_0) - H(X_0|X_{n-1}) \geq H(X_0) - H(X_0|X_n) \tag{4.63}$$

or $H(X_0|X_n)$ increases with n.