1. Binary random variables $X$, $Y$, and $Z$, where $X$ and $Z$ are independent and are both distributed uniformly and $Y = X \oplus Z$.

$$H(X) = -\frac{1}{2} \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2}$$

$$= 1$$

$$= H(Z).$$

$$P_Y(0) = P_X(1)P_Z(1) + P_X(0)P_Z(0)$$

$$= \frac{1}{2}.$$

$$\therefore H(Y) = 1.$$  

$$H(X,Y) = H(X) + H(Y|X)$$

$$= H(X) + H(X \oplus Z|X)$$

$$= H(X) + H(Z)$$

$$= 2.$$  

$$P(X = 0, Y = 0|Z = 0) = P_X(0) = \frac{1}{2}$$

$$P(X = 1, Y = 1|Z = 0) = P_X(1) = \frac{1}{2}$$

$$P(X = 0, Y = 1|Z = 0) = 0$$

$$P(X = 1, Y = 0|Z = 0) = 0$$

$$\therefore H(X, Y|Z = 0) = 1.$$  

Similarly, $H(X, Y|Z = 1) = 1.$

Therefore, $H(X, Y|Z) = P_Z(0)H(X, Y|Z = 0) + P_Z(1)H(X, Y|Z = 1)$

$$= 1.$$  

From properties of entropy we know that

$$H(X, Y) = H(X) + H(Y) \iff X \perp Y.$$  

Further,  

$$H(X, Y|Z) = H(X|Z) + H(Y|Z) \iff X \text{ conditionally independent of } Y \text{ given } Z.$$  

Therefore, clearly, $X$ and $Y$ are independent but $X$ and $Y$ are not independent given $Z.$
2.  $X \sim Ber(\frac{1}{2})$, $P_{Y|X}(0|0) = 0.5$, $P_{Y|X}(1|1) = 1$.

$\therefore H(Y|X = 0) = 1.$

$H(Y|X = 1) = 0.$

$H(Y|X) = P_X(0)H(Y|X = 0) + P_X(1)H(Y|X = 1)$

$= \frac{1}{2}.$

$P_Y(0) = P_X(0)P_{Y|X}(0|0) + P_X(1)P_{Y|X}(0|1)$

$= \frac{1}{2} \cdot \frac{1}{2}$

$= \frac{1}{4} = 1 - P_Y(1).$

$\therefore H(Y) = \frac{1}{4} \log 4 + \frac{1}{4} (2 - \log 3)$

$= 0.603.$

$H(X, Y) = H(X) + H(Y|X)$

$= 1 + \frac{1}{2} = 1.5.$

$H(X|Y) = H(X, Y) - H(Y)$

$= 1.5 - 0.603 = 0.897.$

3.  (a) \textit{(Conditioning reduces entropy)} Using chain rule, $H(X_1, X_2, X_3)$ can be expanded in the following two ways.

$$2H(X_1, X_2, X_3) = H(X_1, X_2) + H(X_3|X_1, X_2) + H(X_2, X_3) + H(X_1|X_2, X_3)$$

$$= H(X_1, X_2) + H(X_2, X_3) + H(X_3|X_1, X_2) + H(1|X_2, X_3)$$

$$\leq H(X_1, X_2) + H(X_2, X_3) + H(X_3|X_1, X_2) + H(X_1)$$

$$\leq H(X_1, X_2) + H(X_2, X_3) + H(X_3|X_1) + H(X_1)$$

$$= H(X_1, X_2) + H(X_2, X_3) + H(X_3, X_1).$$

(b) \textit{(Sub additivity of entropy)} Add and subtract $H(X_1) + H(X_2) + H(X_3)$.

$$H(X_1, X_2|X_3) + H(X_2, X_3|X_1) + H(X_3, X_1|X_2)$$

$$= H(X_1, X_2|X_3) + H(X_3) + H(X_2, X_3|X_1) + H(X_1) + H(X_3, X_1|X_2) + H(X_2)$$

$$- [H(X_1) + H(X_2) + H(X_3)]$$

$$= 3H(X_1, X_2, X_3) - [H(X_1) + H(X_2) + H(X_3)]$$

$$\leq 3H(X_1, X_2, X_3) - [H(X_1) + H(X_2) + H(X_3)]$$

$$= 2H(X_1, X_2, X_3).$$

4.  (i) $H(X|Y) \leq H(X)$ \hspace{1cm} (Conditioning reduces entropy)

$\leq \log |\mathcal{X}|$. \hspace{1cm} (using the operational definition of entropy)

Observe that equality holds when $Y \perp \!
\!\!\!\!\perp X$ and $X \sim \text{Unif}(\mathcal{X})$.

(ii) We know that for $Z \sim Geo(p), \mathbb{E}[Z] = \frac{1}{p}$.

Let $X$ be a r.v. such that $X = i$ with probability $P_X(i), i \in \mathbb{N}$ and $\mathbb{E}[X] = \mu$. 
Now let $Y \sim P_Y \equiv Geo(\frac{1}{\mu})$. Therefore, $\mathbb{E}[Y] = \mu$.

Now consider

$$D(P_X||P_Y) = \sum_{i=1}^{\infty} P_X(i) \log \frac{P_X(i)}{P_Y(i)} = \sum_{i=1}^{\infty} P_X(i) \log P_X(i) - P_Y(i) \log P_Y(i) + P_Y(i) \log P_Y(i) - P_X(i) \log P_Y(i)$$

$$= H(Y) - H(X) + \sum_{i=1}^{\infty} \left[ P_Y(i) \log P_Y(i) - P_X(i) \log P_Y(i) \right]. \quad (1)$$

Since $P_Y(i) = \left(1 - \frac{1}{\mu}\right)^{i-1} \left(\frac{1}{\mu}\right)$,

$$\sum_{i=1}^{\infty} P_X(i) \log P_Y(i) = \sum_{i=1}^{\infty} P_X(i) \cdot (i-1) \log(\mu - 1) - \sum_{i=1}^{\infty} P_X(i) \cdot i \cdot \log \mu$$

$$= (\mu - 1) \log(\mu - 1) - \mu \log \mu. \quad (2)$$

From the entropy calculation of a Geometric r.v., we know that,

$$\sum_{i=1}^{\infty} P_Y(i) \log P_Y(i) = \left(1 - \frac{1}{\mu}\right) \log \left(\frac{1}{\mu}\right) + \frac{1}{\mu} \log \left(\frac{1}{\mu}\right)$$

$$= (\mu - 1) \log(\mu - 1) - \mu \log \mu. \quad (3)$$

Substituting (2) and (3) in (1), we see that,

$$H(Y) - H(X) = D(P_X||P_Y) \geq 0.$$ 

Therefore, for any r.v. $X \in \mathbb{N}$ with $\mathbb{E}[X] = \mu$, $H(X) \leq H(Y)$ where $Y \sim Geo(\frac{1}{\mu})$.

5. (i) For a Markov chain, $X \rightarrow Y \rightarrow Z$, by data processing inequality,

$$I(X \wedge Y) \geq I(X \wedge Z) \Rightarrow H(X) - I(X \wedge Y) \leq H(X) - I(X \wedge Z)$$

i.e., $H(X|Y) \leq H(X|Z)$.

(ii) Let $\mathcal{X} = S_{52}$ denote the set of all permutations of numbers $\{1, 2, ..., 52\}$, namely each element $x \in \mathcal{X}$ represents an order of shuffled cards. A random shuffle of cards can be represented by a channel $W : \mathcal{X} \rightarrow \mathcal{X}$. We say that a shuffle is good if it increases the “randomness” in ordering of cards. Formally, a shuffle $W$ is said to be good if for every pmf $P$ on $\mathcal{X},$

$$H(P) \leq H(PW).$$

Claim: A shuffle is good if and only if it does not decrease the entropy of randomly ordered cards (here randomly ordered stand for $X$ distributed uniformly over $\mathcal{X}$).

Proof. Clearly, if a shuffle is good then for a uniform distribution $P_{\text{unif}}(\mathcal{X})$ on $\mathcal{X}$

$$\log |\mathcal{X}| = H(P_{\text{unif}}(\mathcal{X})) \leq H(P_{\text{unif}}(\mathcal{X})W) \leq \log |\mathcal{X}|,$$
which can only happen if \( H(\mathcal{P}_{\text{uni}}(\mathcal{X}))W) = \log |\mathcal{X}| \). Thus, the shuffle does not decrease the entropy of randomly ordered cards, and \( \mathcal{P}_{\text{uni}}(\mathcal{X})W \) is uniform as well.

Conversely, suppose a shuffle does not decrease the entropy of randomly ordered cards. Then, \( H(\mathcal{P}_{\text{uni}}(\mathcal{X}))W) = H(\mathcal{P}_{\text{uni}}(\mathcal{X})) = \log |\mathcal{X}| \), and so,

\[
\mathcal{P}_{\text{uni}}(\mathcal{X})W = \mathcal{P}_{\text{uni}}(\mathcal{X}).
\]

Note that for any pmf \( \mathcal{P} \) on \( \mathcal{X} \)

\[
D(\mathcal{P}||\mathcal{P}_{\text{uni}}(\mathcal{X})) = \log |\mathcal{X}| - H(\mathcal{P}).
\]

Also, since \( \mathcal{P}_{\text{uni}}(\mathcal{X})W = \mathcal{P}_{\text{uni}}(\mathcal{X}) \),

\[
D(\mathcal{P}W||\mathcal{P}_{\text{uni}}(\mathcal{X})W) = \log |\mathcal{X}| - H(\mathcal{P}W).
\]

Thus, by the data processing inequality we get

\[
\log |\mathcal{X}| - H(\mathcal{P}W) = D(\mathcal{P}W||\mathcal{P}_{\text{uni}}(\mathcal{X})W) \\
\leq D(\mathcal{P}||\mathcal{P}_{\text{uni}}(\mathcal{X})) \\
= \log |\mathcal{X}| - H(\mathcal{P}),
\]

which gives

\[
H(\mathcal{P}) \leq H(\mathcal{P}W).
\]

Therefore, the shuffle is good if it does not reduce the entropy of randomly ordered cards. \( \Box \)

Now, to answer the question asked in the homework, we show that the shuffle mentioned in homework is good. To that end, by the claim above, it suffices to show that it does not reduce the entropy of randomly ordered cards. Specifically, the shuffle in the homework can be described as

\[
W(y|x) = \begin{cases} 
\frac{1}{52}, & \text{if } y = (x_i, x_1, ..., x_{i-1}, x_{i+1}, ..., x_{52}) \text{ for some } 1 \leq i \leq 52, \\
0, & \text{otherwise}.
\end{cases}
\]

For \( W \) above, it is easy to see that

\[
\mathcal{P}_{\text{uni}}(\mathcal{X})W = \mathcal{P}_{\text{uni}}(\mathcal{X}),
\]

which by the claim above shows that shuffle \( W \) is good, \( i.e. \),

\[
H(Y) = H(\mathcal{P}W) \geq H(\mathcal{P}) = H(\mathcal{X}).
\]

6. Let \( p \) denote the maximum probability of guessing \( X \) correctly. Clearly \( p = \max_x P_X(x) \). Let \( \hat{X} \) be the random variable denoting the guess for \( X \). The probability of correctness of the guess could be maximized by always guessing the \( \text{argmax}_x P_X(x) \). Then, by Fano’s inequality,

\[
H(X|\hat{X}) \leq h(1 - p) + (1 - p) \log(|\mathcal{X}|-1)
\]

Since the guess is independent of the random variable itself, \( H(X|\hat{X}) = H(X) \).

For conditions of equality, verify from the proof of Fano’s inequality in the class that equality holds in Fano’s inequality if and only if

\[
H(X|\hat{X}, 1(X \neq \hat{X}) = 1) = \log(|\mathcal{X}|-1).
\]
For the specific case here, this equality holds if and only if
\[
P(X = x | X \neq x_{\text{max}}) = \frac{1}{|X| - 1}, \quad x \in X \setminus \{x_{\text{max}}\},
\]
where \( x_{\text{max}} = \arg\max_x P_X(x) \). This in turn holds if and only if \( P_X(x) \) is the same for every \( x \neq x_0 \).

7. If \( X \perp Y \perp Z \),
\[
I(X \land Y | Z) = H(X|Z) - H(X|Y, Z)
= H(X|Z) - H(X|Y) \quad (\text{by Markovity})
= H(X|Z) - H(X|Y) + H(X) - H(X)
= H(X|Z) + I(X \land Y) - H(X)
\leq I(X \land Y). \quad (\text{since } H(X|Z) \leq H(X))
\]

It is easy to observe that equality is attained \textit{iff} \( X \perp Y \perp Z \).

- \textit{An example where } \( I(X \land Y | Z) > I(X \land Y) \): Let \( X \sim \text{Ber}(\frac{1}{2}), Y \sim \text{Ber}(\frac{1}{2}) \) and \( X \perp Y \). Then, clearly, \( I(X \land Y) = 0 \). Let \( Z = X \oplus Y \). Verify that \( I(X \land Y | Z) = 1 \).

8. Shape of \( D(W||Q|P) \)

(a) To show that \( D(W||Q|P) \) is convex in \( Q \) for a fixed \( P \), let \( Q = \alpha Q_1 + (1 - \alpha)Q_2, 0 < \alpha < 1 \).
\[
D(W||Q|P) = \sum (P \circ W)(x, y) \log \frac{W(y|x)}{\alpha Q_1(y) + (1 - \alpha)Q_2(y)}
= \sum (P \circ W)(x, y) \log W(y|x) - \sum (P \circ W)(x, y) \log [\alpha Q_1(y) + (1 - \alpha)Q_2(y)]
\leq \sum \left[ \alpha (P \circ W)(x, y) \log W(y|x) + (1 - \alpha)(P \circ W)(x, y) \log W(y|x) \right]
- \sum \left[ \alpha (P \circ W)(x, y) \log Q_1(y) - (1 - \alpha)(P \circ W)(x, y) \log Q_2(y) \right]
= \alpha \sum (P \circ W)(x, y) \log \frac{W(y|x)}{Q_1(y)} + (1 - \alpha) \sum (P \circ W)(x, y) \log \frac{W(y|x)}{Q_2(y)}
= \alpha D(W||Q_1|P) + (1 - \alpha)D(W||Q_2|P).
\]
where the inequality follows by applying Jensen’s inequality to the convex function
\(- \log(.)\).

(b) To show that \( D(W||Q|P) \) is linear in \( P \) for a fixed \( Q \), let \( P = aP_1 + bP_2, a, b \neq 0 \).
\[
D(W||Q|P) = \sum P(x) \log \frac{W(y|x)}{Q(y)}
= \sum [aP_1(x) + bP_2(x)] \log \frac{W(y|x)}{Q(y)}
= \sum aP_1(x) \log \frac{W(y|x)}{Q(y)} + \sum bP_2(x) \log \frac{W(y|x)}{Q(y)}
= aD(W||Q|P_1) + bD(W||Q|P_2).
\]
Therefore, \( D(W||Q|P) \) is linear in \( P \) for a fixed \( Q \).
9. Let $X$ be distributed uniformly over all subsets of $\{1, \ldots, n\}$ of size less than or equal to $np$. Let $X_i = 1(i \in X)$, i.e., $X_i$ is 1 if the random set contains the element $i$ and 0 otherwise. Thus, $X = (X_1, \ldots, X_n)$.

Note that the number of values taken by the random variable is $\sum_{i \leq np} \binom{n}{i}$.

Thus,

$$\log \sum_{i \leq np} \binom{n}{i} = H(X)$$

$$= H(X_1, \ldots, X_n)$$

$$\leq \sum_{i=1}^{n} H(X_i)$$

$$= nH(X_1),$$

(4)

where in the last step we used the fact that random variables $X_i$s have the same distribution.

To bound $H(X_1)$, we bound $P(X_1 = 1)$. To that end, note that for every $\delta \leq p$

$$P(X_1 = 1| |X| = n\delta) = \frac{\text{number of subsets of } [n] \text{ of size } n\delta \text{ containing 1}}{\text{number of subsets of } [n] \text{ of size } n\delta}$$

$$= \frac{(n-1)}{(n-\delta)}$$

$$= \delta$$

$$\leq p.$$

Therefore, $P(X_1 = 1) \leq p$. Since $h(\delta)$ is increasing in $\delta$ for $0 \leq \delta \leq 1/2$,

$$H(X_1) = h(P(X_1 = 1)) \leq h(p),$$

which together with (4) completes the proof.