**Homework Set #1**
**Properties of Entropy, Mutual Information and Divergence**

1. **Entropy of functions of a random variable.**
   Let $X$ be a discrete random variable. Show that the entropy of a function of $X$ is less than or equal to the entropy of $X$ by justifying the following steps:

$$
\begin{aligned}
H(X, g(X)) &\stackrel{(a)}{=} H(X) + H(g(X)|X) \\
&\stackrel{(b)}{=} H(X). \\
H(X, g(X)) &\stackrel{(c)}{=} H(g(X)) + H(X|g(X)) \\
&\stackrel{(d)}{\geq} H(g(X)).
\end{aligned}
$$

   Thus $H(g(X)) \leq H(X)$.

   **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 $H(X) \geq H(g(X))$.

2. **Example of joint entropy.**
   Let $p(x, y)$ be given by

| $X$ \ $Y$ | 0 | 1 |
|---|---|---|
| 0 | $\frac{1}{3}$ | $\frac{1}{3}$ |
| 1 | 0 | $\frac{1}{3}$ |

   Find

(a) $H(X), H(Y)$.

(b) $H(X|Y), H(Y|X)$.

(c) $H(X, Y)$.

(d) $H(Y) - H(Y|X)$.

(e) $I(X; Y)$.

**Solution: Example of joint entropy**

(a) $H(X) = \frac{2}{3}\log\frac{3}{2} + \frac{1}{3}\log 3 = .918$ bits $= H(Y)$.

(b) $H(X|Y) = \frac{1}{3}H(X|Y=0) + \frac{2}{3}H(X|Y=1) = .667$ bits $= H(Y|X)$.

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

(d) $H(Y) - H(Y|X) = .251$ bits.

(e) $I(X; Y) = H(Y) - H(Y|X) = .251$ bits.

3. **Bytes.**
The entropy, $H_a(X) = -\sum p(x)\log_a p(x)$ is expressed in bits if the logarithm is to the base 2 and in bytes if the logarithm is to the base 256. What is the relationship of $H_2(X)$ to $H_{256}(X)$?

**Solution: Bytes.**

$$
\begin{aligned}
lim_{i=\infty} I(Y_i; Y^{i-1}|Q_i) = 0 H_2(X) &= -\sum p(x)\log_2 p(x) \\
&= -\sum p(x)\frac{\log_2 p(x)\log_{256}(2)}{log_{256}(2)} \\
&\stackrel{(a)}{=} -\sum p(x)\frac{\log_{256} p(x)}{log_{256}(2)} \\
&= \frac{-1}{\log_{256}(2)}\sum p(x)\log_{256} p(x) \\
&\stackrel{(b)}{=} \frac{H_{256}(X)}{\log_{256}(2)},
\end{aligned}
$$

where $(a)$ comes from the property of logarithms and $(b)$ follows from the definition of $H_{256}(X)$. Hence we get

$$H_2(X) = 8H_{256}(X).$$

4. **Two looks.**
Here is a statement about pairwise independence and joint independence. Let $X, Y_1$, and $Y_2$ be binary random variables. If $I(X; Y_1) = 0$ and $I(X; Y_2) = 0$, does it follow that $I(X; Y_1, Y_2) = 0$?

   (a) Yes or no?

   (b) Prove or provide a counterexample.

   (c) If $I(X; Y_1) = 0$ and $I(X; Y_2) = 0$ in the above problem, does it follow that $I(Y_1; Y_2) = 0$? In other words, if $Y_1$ is independent of $X$, and if $Y_2$ is independent of $X$, is it true that $Y_1$ and $Y_2$ are independent?

   **Solution: Two looks.**

   (a) The answer is "no".

   (b) Although at first the conjecture seems reasonable enough–after all, if $Y_1$ gives you no information about $X$, and if $Y_2$ gives you no information about $X$, then why should the two of them together give any information? But remember, it is NOT the case that $I(X; Y_1, Y_2) = I(X; Y_1) + I(X; Y_2)$. The chain rule for information says instead that $I(X; Y_1, Y_2) = I(X; Y_1) + I(X; Y_2|Y_1)$. The chain rule gives us reason to be skeptical about the conjecture.

   This problem is reminiscent of the well-known fact in probability that pair-wise independence of three random variables is not sufficient to guarantee that all three are mutually independent. $I(X; Y_1) = 0$ is equivalent to saying that $X$ and $Y_1$ are independent. Similarly for $X$ and $Y_2$. But just because $X$ is pairwise independent with each of $Y_1$ and $Y_2$, it does not follow that $X$ is independent of the vector $(Y_1, Y_2)$.

   Here is a simple counterexample. Let $Y_1$ and $Y_2$ be independent fair coin flips. And let $X = Y_1 \text{ XOR } Y_2$. $X$ is pairwise independent of both $Y_1$ and $Y_2$, but obviously not independent of the vector $(Y_1, Y_2)$, since $X$ is uniquely determined once you know $(Y_1, Y_2)$.

   (c) Again the answer is "no". $Y_1$ and $Y_2$ can be arbitrarily dependent with each other and both still be independent of $X$. For example, let $Y_1 = Y_2$ be two observations of the same fair coin flip, and

$X$ an independent fair coin flip. Then $I(X;Y_1) = I(X;Y_2) = 0$ because $X$ is independent of both $Y_1$ and $Y_2$. However, $I(Y_1;Y_2) = H(Y_1) - H(Y_1|Y_2) = H(Y_1) = 1$.

5. **A measure of correlation.**
Let $X_1$ and $X_2$ be *identically distributed*, but not necessarily independent. Let

$$\rho = 1 - \frac{H(X_1|X_2)}{H(X_1)}.$$

(a) Show $\rho = \frac{I(X_1;X_2)}{H(X_1)}$.

(b) Show $0 \leq \rho \leq 1$.

(c) When is $\rho = 0$?

(d) When is $\rho = 1$?

**Solution: A measure of correlation.**

$X_1$ and $X_2$ are identically distributed and

$$\rho = 1 - \frac{H(X_2|X_1)}{H(X_1)}$$

(a)

$$\begin{aligned}
\rho &= \frac{H(X_1) - H(X_2|X_1)}{H(X_1)} \\
&= \frac{H(X_2) - H(X_2|X_1)}{H(X_1)} \quad \text{(since } H(X_1) = H(X_2)\text{)} \\
&= \frac{I(X_1;X_2)}{H(X_1)}.
\end{aligned}$$

(b) Since $0 \leq H(X_2|X_1) \leq H(X_2) = H(X_1)$, we have

$$0 \leq \frac{H(X_2|X_1)}{H(X_1)} \leq 1$$

$$0 \leq \rho \leq 1.$$

(c) $\rho = 0$ iff $I(X_1;X_2) = 0$ iff $X_1$ and $X_2$ are independent.

(d) $\rho = 1$ iff $H(X_2|X_1) = 0$ iff $X_2$ is a function of $X_1$. By symmetry, $X_1$ is a function of $X_2$, i.e., $X_1$ and $X_2$ have a one-to-one correspondence. For example, if $X_1 = X_2$ with probability 1 then $\rho = 1$. Similarly, if the distribution of $X_i$ is symmetric then $X_1 = -X_2$ with probability 1 would also give $\rho = 1$.

6. **The value of a question.**
Let $X \sim p(x), \quad x = 1, 2, \ldots, m$.

We are given a set $S \subseteq \{1, 2, \ldots, m\}$. We ask whether $X \in S$ and receive the answer

$$Y = \begin{cases} 1, & \text{if } X \in S \\ 0, & \text{if } X \notin S. \end{cases}$$

Suppose $\Pr\{X \in S\} = \alpha$.

(a) Find the decrease in uncertainty $H(X) - H(X|Y)$.

(b) Is it true that any set $S$ with a given probability $\alpha$ is as good as any other.

**Solution: The value of a question.**

(a) Consider

$$H(X) - H(X|Y) = H(Y) - H(Y|X) = H(Y) = H_b(\alpha) \qquad (1)$$

(b) Yes, since the answer depends only on $\alpha$.

7. **Relative entropy is not symmetric**
Let the random variable $X$ have three possible outcomes $\{a, b, c\}$. Consider two distributions on this random variable

| Symbol | $p(x)$ | $q(x)$ |
|---|---|---|
| a | 1/2 | 1/3 |
| b | 1/4 | 1/3 |
| c | 1/4 | 1/3 |

Calculate $H(p), H(q), D(p \parallel q)$ and $D(q \parallel p)$.

Verify that in this case $D(p \parallel q) \neq D(q \parallel p)$.

**Solution: Relative entropy is not symmetric.**

(a) $H(p) = 1/2 \log 2 + 2 \times 1/4 \log 4 = 1.5$ bits.

(b) $H(q) = 3 \times 1/3 \log 3 = \log 3 = 1.585$ bits.

(c) $D(p\|q) = 1/2 \log 3/2 + 2 \times 1/4 \log 3/4 = \log 3 - 3/2 = 0.0850$ bits.

(d) $D(q\|p) = 1/3 \log 2/3 + 2 \times 1/3 \log 4/3 = 5/3 - \log 3 = 0.0817$ bits.

$D(p\|q) \neq D(q\|p)$ as expected.

8. **"True or False" questions**

Copy each relation and write **true** or **false**. Then, if it's true, prove it. If it is false give a counterexample or prove that the opposite is true.

(a) $H(X) \geq H(X|Y)$

(b) $H(X) + H(Y) \leq H(X, Y)$

(c) Let $X, Y$ be two independent random variables. Then

$$H(X - Y) \geq H(X).$$

(d) Let $X, Y, Z$ be three random variables that satisfies $H(X, Y) = H(X) + H(Y)$ and $H(Y, Z) = H(Z) + H(Y)$. Then the following holds

$$H(X, Y, Z) = H(X) + H(Y) + H(Z).$$

(e) For any $X, Y, Z$ and the deterministic function $f, g$ $I(X; Y|Z) = I(X, f(X, Y); Y, g(Y, Z)|Z)$.

**Solution to "True or False" questions** e.

(a) $H(X) \geq H(X|Y)$ is **true**. Proof: In the class we showed that $I(X; Y) > 0$, hence $H(X) - H(X|Y) > 0$.

(b) $H(X) + H(Y) \leq H(X, Y)$ is **false**. Actually the opposite is true, i.e., $H(X) + H(Y) \geq H(X, Y)$ since $I(X; Y) = H(X) + H(Y) - H(X, Y) \geq 0$.

(c) Let $X, Y$ be two independent random variables. Then

$$H(X - Y) \geq H(X).$$

**True**

$$H(X - Y) \overset{(a)}{\geq} H(X - Y|Y)) \overset{(b)}{\geq} H(X)$$

(a) follows from the fact that conditioning reduces entropy.
(b) Follows from the fact that given $Y$, $X - Y$ is a Bijective Function.

(d) Let $X, Y, Z$ be three random variables that satisfies $H(X,Y) = H(X) + H(Y)$ and $H(Y,Z) = H(Z) + H(Y)$. Then the following holds $H(X,Y,Z) = H(X) + H(Y) + H(Z)$. This is **false**. Consider the following derivations

$$\begin{aligned} H(X,Y,Z) &= H(X,Y) + H(Z|X,Y) && (2)\\ &= H(X) + H(Y) + H(Z) - I(Z;X,Y) && (3)\\ &= H(X) + H(Y) + H(Z) - I(Z;X|Y) && (4)\\ &\leq H(X) + H(Y) + H(Z) && (5) \end{aligned}$$

since $I(Z;X|Y)$ can be greater than 0. For example, $X, Y$ are two independent RV distributed uniformly over $\{0,1\}$ and $Z = X \bigoplus_2 Y$. In this case, $X$ is independent of $Y$ and $Y$ is independent of $Z$ buy $Z$ is dependent on $(X,Y)$.

(e) For any $X, Y, Z$ and the deterministic function $f, g$, $I(X;Y|Z) = I(X, f(X,Y); Y, g(Y,Z)|Z)$ is **false** since adding the function $f(X,Y)$ to the left hand side increases the mutual information.

$$\begin{aligned} I(X, f(X,Y); Y, g(Y,Z)|Z) &= I(X, f(X,Y); Y|Z) && (6)\\ &= I(X;Y|Z) + I(f(X,Y);Y|Z,X) && (7)\\ &= I(X;Y|Z) + H(f(X,Y)|Z,X) && (8)\\ &\geq I(X;Y|Z) && (9) \end{aligned}$$

since $H(f(X,Y)|Z,X) \geq$.

9. **Entropy of 3 pairwise independent random variables**:

Let $W, X, Y$ be 3 random variables distributed each Bernoulli (0.5) that are pairwise independent, i.e., $I(W;X) = I(X;Y) = I(W;Y) = 0$.

(a) What is the **maximum** possible value of $H(W,X,Y)$?
From the chain rule for entropy and the fact that $W$ and $X$ are independent,

$$H(W,X,Y) = H(W) + H(X) + H(Y|W,X) \leq H(W) + H(X) + H(Y),$$

where the inequality follows since conditioning reduces entropy. $H(W) = H(X) = H(Y) = 1$, thus $H(W, X, Y) \leq 3$.

(b) What is the condition under which this **maximum** is achieved?
Notice that the maximum is achieved if $H(Y|X, W) = H(Y)$, i.e., $Y$ is independent of the pair $(X, W)$ (or similarly $W$ is independent of $(X, Y)$, or $X$ is independent of $(W, Y)$).

(c) What is the **minimum** possible value of $H(W, X, Y)$?
On the other hand,

$$H(W, X, Y) = H(W) + H(X) + H(Y|W, X) \geq H(W) + H(X), \tag{10}$$

where the inequality follows from the non-negativity of the conditional entropy $H(Y|W, X)$. Thus, $H(W, X, Y) \geq 2$.

(d) Give a specific example achieving this **minimum**.
If $Y$ is a deterministic function of both $(W, X)$, the inequality is achieved. Notice that it cannot be a deterministic function of just one of them since it contradicts the assumption of the question. Let, for instance, $Y = W \oplus X$, where $\oplus$ denotes the addition modulo 2 (i.e., XOR).

10. **Joint Entropy** Consider $n$ different discrete random variables, named $X_1, X_2, ..., X_n$. Each random variable separately has an entropy $H(X_i)$, for $1 \leq i \leq n$.

(a) What is the upper bound on the joint entropy $H(X_1, X_2, ..., X_n)$ of all these random variables $X_1, X_2, ..., X_n$ given that $H(X_i)$, for $1 \leq i \leq n$ are fixed?

(b) Under what conditions will this upper bound be reached?

(c) What is the lower bound on the joint entropy $H(X_1, X_2, ..., X_n)$ of all these random variables?

(d) Under what condition will this upper bound be reached?

**Solution:**

(a) The upper bound is $\sum_{i=1}^{n} H(X_i)$.

$$\begin{aligned} H(X^n) &= \sum_{i=1}^{n} H(X_i|X^{i-1}) \\ &\leq \sum_{i=1}^{n} H(X_i) \end{aligned} \tag{11}$$

(please explain each step of the equation above)

(b) It can be achieved if all $\{X_i\}_{i=1}^{n}$ are independent, since for this case $H(X^n) = \sum_{i=1}^{n} H(X_i)$.

(c) The lower bound is $H(X_i)$, where $X_i$ has the largest entropy.

$$H(X^n) \geq H(X_i) \; \forall i = 1, 2, ..., n. \tag{12}$$

(d) It can be achieved if for all $j \neq i$: $X_j = f_j(X_i)$ for some deterministic function $f_j$.

11. **True or False**

Let $X, Y, Z$ be discrete random variable. Copy each relation and write **true** or **false**. If it's true, prove it. If it is false give a counterexample or prove that the opposite is true.

For instance:

- $H(X) \geq H(X|Y)$ is **true**. Proof: In the class we showed that $I(X;Y) > 0$, hence $H(X) - H(X|Y) > 0$.
- $H(X) + H(Y) \leq H(X, Y)$ is **false**. Actually the opposite is true, i.e., $H(X) + H(Y) \geq H(X, Y)$ since $I(X;Y) = H(X) + H(Y) - H(X, Y) \geq 0$.

(a) If $H(X|Y) = H(X)$ then $X$ and $Y$ are independent.

(b) For any two probability mass functions (pmf) $P, Q$,

$$D\left(\frac{P+Q}{2}||Q\right) \leq \frac{1}{2}D(P||Q),$$

where $D(||)$ is a divergence between two pmfs.

(c) Let $X$ and $Y$ be two independent random variables. Then

$$H(X+Y) \geq H(X).$$

(d) $I(X;Y) - I(X;Y|Z) \leq H(Z)$

(e) If $f(x,y)$ is a convex function in the pair $(x,y)$, then for a fixed $y$, $f(x,y)$ is convex in $x$, and for a fixed $x$, $f(x,y)$ is convex in $y$.

(f) If for a fixed $y$ the function $f(x,y)$ is a convex function in $x$ , and for a fixed $x$, $f(x,y)$ is convex function in $y$, then $f(x,y)$ is convex in the pair $(x,y)$. (Examples of such functions are $f(x,y) = f_1(x) + f_2(y)$ or $f(x,y) = f_1(x)f_2(y)$ where $f_1(x)$ and $f_2(y)$ are convex.)

(g) Let $X, Y, Z, W$ satisfy the Markov chain $X-Y-Z$ and $Y-Z-W$. Does the Markov $X-Y-Z-W$ hold? (The Markov $X-Y-Z-W$ means that $P(x|y,z,w) = P(x|y)$ and $P(x,y|z,w) = P(x,y|z)$.)

(h) $H(X|Z)$ is concave in $P_{X|Z}$ for fixed $P_Z$.

**Solution to True or False**

(a) If $H(X|Y) = H(X)$ then $X$ and $Y$ are independent.
**True**:

$$I(X;Y) = H(X) - H(X|Y)$$

If $I(X;Y) = 0$ then $H(X) = H(X|Y)$. We can write:

$$I(X;Y) = D(\ P_{X,Y} \ || \ P_X P_Y \ ) = 0$$

$D(Q||P) = 0$ iff $P(x) = Q(x)\ \forall x$, therefore $P_{X,Y}(x,y) = P_X(x)P_Y(y)$ for every $x, y$ and as result $X \perp Y$.

(b) For any two probability mass functions (pmf) $P, Q$,

$$D\left(\frac{P+Q}{2}||Q\right) \leq \frac{1}{2}D(P||Q),$$

where $D(||)$ is a divergence between two pmfs.
**True**:
Using the concave property of the divergence function:

$$D\left(\lambda P + (1-\lambda)Q \ || \ Q\right) \leq \lambda D(P \ || \ Q) + (1-\lambda)D(Q \ || \ Q)$$

Assigning $\lambda = \frac{1}{2}$, and since $D(Q||Q) = 0$:

$$D\left(\frac{1}{2}P + \frac{1}{2}Q \mid\mid Q\right) \leq \frac{1}{2}D(P||Q)$$

(c) Let $X$ and $Y$ be two independent random variables. Then

$$H(X+Y) \geq H(X).$$

**True**:

$$H(X+Y) \geq H(X+Y|Y) \stackrel{(a)}{=} H(X)$$

(a) - since X is independent of Y.

(d) $I(X;Y) - I(X;Y|Z) \leq H(Z)$

**True**:

$$\begin{aligned} I(X;Y) - I(X;Y|Z) &= H(X) - H(X|Y) - [H(X|Z) - H(X|Y,Z)] \\ &= \underbrace{H(X) - H(X|Z)}_{I(X;Z)} - \underbrace{[H(X|Y) - H(X|Y,Z)]}_{\geq 0} \\ &\leq I(X;Z) \\ &= H(Z) - \underbrace{H(Z|X)}_{\geq 0} \\ &\leq H(Z) \end{aligned}$$

(e) If $f(x,y)$ is a convex function in the pair $(x,y)$, then for a fixed $y$, $f(x,y)$ is convex in $x$, and for a fixed $x$, $f(x,y)$ is convex in $y$.

**True** If the function is Convex for every combination of $(x,y)$ it is necessarily Convex for Affine Function of the pair.

(f) If for a fixed $y$ the function $f(x,y)$ is a convex function in $x$ , and for a fixed $x$, $f(x,y)$ is convex function in $y$, then $f(x,y)$ is convex in the pair $(x,y)$.

**False**

Consider the function $f(x,y) = xy$. Its linear in $x$ for fixed $y$ and vice versa but the function its neither convex nor concave. The second derivative matrix is not semi-definite positive.

(g) **False** Let us assume that

$$\begin{aligned} Z &\sim Bern(0.5), && (13)\\ W &\sim Bern(0.5), && (14)\\ X &= Z \oplus W, && (15)\\ Y &= X \oplus A, && (16) \end{aligned}$$

where $A \sim Bern(0.1)$. The Markov $X - Y - Z$ holds since $X$ and $Z$ are independent and the relation $Y - Z - W$ holds from the fact that $Y$ is independent of $(Z, W)$. However, by knowing $Z$ and $W$ we know $X$ and therefore $p(x, y|z, w) = p(x, y|z)$ does not hold in general.

(h) **True** We know that,

$$H(X|Z) = \sum_{z \in \mathcal{Z}} p(z) H(X|Z = z). \quad (17)$$

For a fixed $p(z)$, $H(X|Z)$ is formed as a linear combination of concave functions ($H(X|Z = z)$ is concave), thus, $H(X|Z)$ is concave in $P_{X|Z}$.

12. **Random questions.**
One wishes to identify a random object $X \sim p(x)$. A question $Q \sim r(q)$ is asked at random according to $r(q)$. This results in a deterministic answer $A = A(x, q) \in \{a_1, a_2, \ldots\}$. Suppose the object $X$ and the question $Q$ are independent. Then $I(X; Q, A)$ is the uncertainty in $X$ removed by the question-answer $(Q, A)$.

(a) Show $I(X; Q, A) = H(A|Q)$. Interpret.

(b) Now suppose that two i.i.d. questions $Q_1, Q_2 \sim r(q)$ are asked, eliciting answers $A_1$ and $A_2$. Show that two questions are less valuable than twice the value of a single question in the sense that $I(X; Q_1, A_1, Q_2, A_2) \leq 2I(X; Q_1, A_1)$.

**Solution: Random questions.**

(a) Since $A$ is a deterministic function of $(Q, X)$, $H(A|Q, X) = 0$.

Also since $X$ and $Q$ are independent, $H(Q|X) = H(Q)$. Hence,

$$\begin{aligned}
I(X;Q,A) &= H(Q,A) - H(Q,A,|X) \\
&= H(Q) + H(A|Q) - H(Q|X) - H(A|Q,X) \\
&= H(Q) + H(A|Q) - H(Q) \\
&= H(A|Q).
\end{aligned}$$

The interpretation is as follows. The uncertainty removed in $X$ given $(Q,A)$ is the same as the uncertainty in the answer given the question.

(b) Using the result from part (a) and the fact that questions are independent, we can easily obtain the desired relationship.

$$\begin{aligned}
I(X;Q_1,A_1,Q_2,A_2) &\stackrel{(a)}{=} I(X;Q_1) + I(X;A_1|Q_1) + I(X;Q_2|A_1,Q_1) \\
&\quad + I(X;A_2|A_1,Q_1,Q_2) \\
&\stackrel{(b)}{=} I(X;A_1|Q_1) + H(Q_2|A_1,Q_1) - H(Q_2|X,A_1,Q_1) \\
&\quad + I(X;A_2|A_1,Q_1,Q_2) \\
&\stackrel{(c)}{=} I(X;A_1|Q_1) + I(X;A_2|A_1,Q_1,Q_2) \\
&= I(X;A_1|Q_1) + H(A_2|A_1,Q_1,Q_2) - H(A_2|X,A_1,Q_1,Q_2) \\
&\stackrel{(d)}{=} I(X;A_1|Q_1) + H(A_2|A_1,Q_1,Q_2) \\
&\stackrel{(e)}{\leq} I(X;A_1|Q_1) + H(A_2|Q_2) \\
&\stackrel{(f)}{=} 2I(X;A_1|Q_1)
\end{aligned}$$

(a) Chain rule.
(b) $X$ and $Q_1$ are independent.
(c) $Q_2$ are independent of $X$, $Q_1$, and $A_1$.
(d) $A_2$ is completely determined given $Q_2$ and $X$.
(e) Conditioning decreases entropy.
(f) Result from part (a).

13. **Entropy bounds.**
Let $X \sim p(x)$, where $x$ takes values in an alphabet $\mathcal{X}$ of size $m$. The entropy $H(X)$ is given by

$$\begin{aligned}
H(X) &\equiv -\sum_{x\in\mathcal{X}} p(x)\log p(x) \\
&= E_p \log \frac{1}{p(X)}.
\end{aligned}$$

Use Jensen's inequality ($Ef(X) \le f(EX)$, if $f$ is concave) to show

(a) $H(X) \le \log E_p \frac{1}{p(X)}$
$= \log m$.

(b) $-H(X) \le \log(\sum_{x \in \mathcal{X}} p^2(x))$, thus establishing a lower bound on $H(X)$.

(c) Evaluate the upper and lower bounds on $H(X)$ when $p(x)$ is uniform.

(d) Let $X_1, X_2$ be two independent drawings of $X$. Find $\Pr\{X_1 = X_2\}$ and show $\Pr\{X_1 = X_2\} \ge 2^{-H}$.

**Solution: Entropy Bounds.**
To prove (a) observe that

$$\begin{aligned}
H(X) &= E_p \log \frac{1}{p(X)} \\
&\le \log E_p \frac{1}{p(X)} \\
&= \log \sum_{x \in \mathcal{X}} p(x) \frac{1}{p(x)} \\
&= \log m
\end{aligned}$$

where the the first inequality follows from Jensen's, and the last step follows since the size of $\mathcal{X}$ is $m$.

To prove (b) proceed

$$\begin{aligned}
-H(X) &= E_p \log p(X) \\
&\le \log E_p p(X) \\
&= \log \left( \sum_{x \in \mathcal{X}} p^2(x) \right)
\end{aligned}$$

where the second step again follows from Jensen's and the third step is just the definition of $E_p(p(X)$. Thus, we have the lower bound

$$H(X) \ge -\log \left( \sum_{x \in \mathcal{X}} p^2(x) \right) \quad .$$

The upper bound is $m$ irrespective of the distribution. Now, $p(x) = 1/m$ for the uniform distribution, and therefore

$$\begin{aligned} -\log \sum_{x \in \mathcal{X}} p^2(x) &= -\log \sum_{x \in \mathcal{X}} \frac{1}{m^2} \\ &= -\log \frac{1}{m} \end{aligned}$$

and therefore the upper and lower bounds agree, and are $\log m$. A direct calculation of the entropy yields the same result immediately.

The derivation of $(d)$ follows from

$$\begin{aligned} \Pr\{X_1 = X_2\} &= \sum_{x,y \in \mathcal{X}} \Pr\{X_1 = x, X_2 = y\} \delta_{xy} \\ &= \sum_{x \in \mathcal{X}} p^2(x) \end{aligned}$$

where the second step follows from the independence of $X_1, X_2$, and the fact that they are identically distributed $X_1, X_2 \sim p(x)$. Here $\delta_{xy}$ is Kronecker's delta function.

14. **Bottleneck.**
Suppose a (non-stationary) Markov chain starts in one of $n$ states, necks down to $k < n$ states, and then fans back to $m > k$ states. Thus $X_1 \to X_2 \to X_3$, $X_1 \in \{1, 2, \ldots, n\}$, $X_2 \in \{1, 2, \ldots, k\}$, $X_3 \in \{1, 2, \ldots, m\}$, and $p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_2)$.

    (a) Show that the dependence of $X_1$ and $X_3$ is limited by the bottleneck by proving that $I(X_1; X_3) \le \log k$.

    (b) Evaluate $I(X_1; X_3)$ for $k = 1$, and conclude that no dependence can survive such a bottleneck.

    **Solution: Bottleneck.**

    (a) From the data processing inequality, and the fact that entropy is maximum for a uniform distribution, we get

$$\begin{aligned} I(X_1; X_3) &\le I(X_1; X_2) \\ &= H(X_2) - H(X_2 \mid X_1) \\ &\le H(X_2) \\ &\le \log k. \end{aligned}$$

Thus, the dependence between $X_1$ and $X_3$ is limited by the size of the bottleneck. That is $I(X_1; X_3) \leq \log k$.

(b) For $k = 1$, $0 \leq I(X_1; X_3) \leq \log 1 = 0$ so that $I(X_1, X_3) = 0$. Thus, for $k = 1$, $X_1$ and $X_3$ are independent.

15. **Convexity of Halfspaces, hyperplanes and polyhedron**

Let $\mathbf{x}$ be a real vector of finite dimension $n$, i.e., $x \in \mathbb{R}^n$. A *halfspace* is the set of all $x \in \mathbb{R}^n$ that satisfies $a^T x \leq b$, where $a \neq 0$. In other words a halfspace is the set

$$\{\mathbf{x} \in \mathbb{R}^{\mathbf{n}} : \mathbf{a^T x} \leq \mathbf{b}\}.$$

A hyperplan is the set of the form

$$\{\mathbf{x} \in \mathbb{R}^{\mathbf{n}} : \mathbf{a^T x} = \mathbf{b}\}.$$

(a) Show that a halfspace and a hyperplan are convex sets.

(b) show that for any two sets $\mathcal{A}$ and $\mathcal{B}$ that are convex the intersection $\mathcal{A} \bigcap \mathcal{B}$ is also convex.

(c) A *polyhedron* is an intersection of halfspaces and a hyperplans. Deduce that a polyhedron is a convex set.

(d) A probability vector $\mathbf{x}$ is such that each element is positive and it sums to 1. Is the set of all vector probabilities of dimension $n$ (called the probability simplex) a halfspace, hyperplan or polyhedron?

**Solution:**

(a) Hyperplane : Let $x_1$ and $x_2$ be vectors that belong to the hyperplane. Since they belong to the hyperplane,$a^T x_1 = b$ and $a^T x_2 = b$ (where $a$ is the scalar vector).

$$\begin{aligned} a^T(\lambda x_1 + (1-\lambda)x_2) &= \lambda a^T x_1 + (1-\lambda) a^T x_2 \\ &= \lambda b + (1-\lambda) b = b. \end{aligned} \tag{18}$$

So the set is indeed convex.

Now consider a Halfspace :Let $x_1$ and $x_2$ be vectors that belong to the halfspace. Since they belong to the hyperplane,$a^T x_1 \leq b$ and $a^T x_2 \leq b$.

$$\begin{aligned} a^T(\lambda x_1 + (1-\lambda)x_2) &= \lambda a^T x_1 + (1-\lambda)a^T x_2 \\ &= \lambda a^T x_1 + (1-\lambda)a^T x_2 \leq \lambda b + (1-\lambda)b \\ &= b. \end{aligned} \tag{19}$$

So the set is indeed convex.

(b) Let $\mathcal{A}$ and $\mathcal{B}$ be convex sets. We want to show that $\mathcal{A} \bigcap \mathcal{B}$ is also convex. Take $x_1, x_2 \in \mathcal{A} \bigcap \mathcal{B}$, and let x lie on the line segment between these two points. Then $x \in \mathcal{A}$ because A is convex, and similarly, $x \in \mathcal{B}$ because $\mathcal{B}$ is convex. Therefore $x \in \mathcal{A} \bigcap \mathcal{B}$, as desired.

(c) Let $x_1$ and $x_2$ be vectors that belong to the halfspace or the hyperplan sets. then as was shown in (b) $x_1 \bigcap x_2$ is also a convex set. Therefore polyhedron is indeed a convex set. definition of polyhedron: $[x|Ax \leq b; Cx = d]$.

(d) The probability simplex $\sum_{i=1}^{n} x_i = 1$ and $x_i \geq 0$ is a special case of a polyhedron.

16. **Some sets of probability distributions.**

Let X be a real-valued random variable with $\Pr(X = a_i) = p_i, i = 1, ..., n$, where $a_1 < a_2 < \ldots < a_n$. Let $\mathbf{p}$ denote the vector $p_1, p_2, ..., p_n$. Of course $\mathbf{p} \in \mathbb{R}^{\mathbf{n}}$ lies in the standard probablility simplex. Which of the follwing conditions are convex in $\mathbf{p}$? (That is, for which of the following conditions is the set of $\mathbf{p} \in \mathbf{P}$ that satisfy the condition convex?)

(a) $\alpha \leq E[f(X)] \leq \beta$,where $E[f(X)]$ is the expected value of $f(X)$, i.e. $E[f(x)] = \sum_{i=1}^{n} p_i f(a_i)$ (The function $f : \mathbb{R} \mapsto \mathbb{R}$ is given.)

(b) $\Pr(X > \alpha) \leq \beta$

(c) $E[|X^3|] \leq \alpha E[|X|]$.

(d) $\mathbf{var}(X) \leq \alpha$, where $\mathbf{var}(X) = E(X - EX)^2$ is the variance of $X$.

(e) $E[X^2] \leq \alpha$

(f) $E[X^2] \geq \alpha$

**Solution :** First we note that P is a polyhedron because $p_i, i = 1, ..., n$ defines halfspaces and $\sum_{i=1}^n p_i = 1$ defines a hyperplane.

(a) $\alpha \leq \sum_{i=1}^n p_i f(a_i) \leq \beta$,so the constraint is equivalent to two linear inequalities in the probabilities $p_i$ -**convex set**.

(b) $\Pr(X > \alpha) \leq \beta$ is equivalent to a linear inequality: $\sum_{i:ai \geq \alpha}^n p_i \leq \beta$ - **convex set**.

(c) The constraint is equivalent to a linear inequality: $\sum_{i=1}^n p_i(|a_i^3| - \alpha|a_i|) \leq 0$ - **convex set**.

(d) $\mathbf{var}(X) = \sum_{i=1}^n p_i a_i^2 - (\sum_{i=1}^n p_i a_i)^2 \leq \alpha$ is not convex in general. As a counterexample, we can take n = 2, $a_1 = 1$, $a_2 = 0$, and $\alpha = 1/8$. p = (0,1) are two points that satisfy var(x)=0 $\leq \alpha$ , but if we take the convex combination p = (1/2, 1/2) then var(x)=1/4 - **not a convex set**.

(e) The constraint is equivalent to a linear inequality: $\sum_{i=1}^n p_i a_i^2 \leq \alpha$ - **convex set**.

(f) The constraint is equivalent to a linear inequality: $\sum_{i=1}^n p_i a_i^2 \geq \alpha$ - **convex set**.

17. **Perspective transformation preserve convexity** Let $f(x)$, $f : \mathbb{R} \to \mathbb{R}$, be a convex function.

(a) Show that the function

$$tf(\frac{x}{t}), \tag{20}$$

is a convex function in the pair $(x, t)$ for $t > 0$. (The function $tf(\frac{x}{t})$ is called perspective transformation of $f(x)$.)

(b) Is the preservation true for concave functions too?

(c) Use this property to prove that $D(P||Q)$ is a convex function in $(P, Q)$.

**Solution:**

(a) Let $f(x)$, $f : \mathbb{R} \to \mathbb{R}$, be a convex function. Lets define $g(x,t) = tf(\frac{x}{t})$.

$$\begin{aligned} g(\lambda(x_1,t_1)+\bar{\lambda}(x_2,t_2)) &= (\lambda t_1+\bar{\lambda}t_2) f\left(\frac{\lambda t_1(\frac{x_1}{t_1})+\bar{\lambda}\frac{x_2}{t_2}}{\lambda t_1+\bar{\lambda}t_2}\right) \\ &\leq (\lambda t_1+\bar{\lambda}t_2)\frac{\lambda t_1}{\lambda t_1+\bar{\lambda}t_2} f(\frac{x_1}{t_1}) + \frac{\bar{\lambda}t_2}{\lambda t_1+\bar{\lambda}t_2} f(\frac{x_2}{t_2}) \\ &= \lambda t_1 f(\frac{x_1}{t_1}) + \bar{\lambda}t_2 f(\frac{x_2}{t_2}) \\ &= \lambda g(x_1,t_1)+\bar{\lambda}g(x_2,t_2) \end{aligned} \tag{21}$$

So g is indeed a convex function.

Another way to solve, is to assume that $f()$ has a second derivative and show that the Hessian is semi-definite positive. However, the first proof is more general since its true for any convex function even if the derivative does not exist.

(b) Now let $f(x)$, $f : \mathbb{R} \to \mathbb{R}$, be a concave function.
-f(x) is convex function and by the same way of (a) we got that g is a concave function. therefore the preservation is true for concave functions too.

(c) $D(P||Q) = \sum_x P(x)\log\frac{P(x)}{Q(x)} = -\sum_x P(x)\log\frac{Q(x)}{P(x)}$. If we consider $Q = (q_1, ..., q_k)$ and $P = (p_1, ..., p_k)$ and choose $p_1 = t$ and $q_1 = x$ , and $f(x) = -log(x)$ (convex function) then we conclude from (a) that $p_1 \log\frac{p_1}{q_1}$ is convex in $(p_1, q_1)$ and since $D(P||Q)$ a summation of convex functions then it is convex.

18. **Coin Tosses**

Consider the next joint distribution: $X$ is the number of coin tosses until the first head appears and $Y$ is the number of coin tosses until the second head appears. The probability for a head is $q$, and the tosses are independent.

a. Compute the distribution of $X$, $p(x)$, the distribution of $Y$, $p(y)$, and the conditional distributions $p(y|x)$ and $p(x|y)$.

b. Compute $H(X)$, $H(Y|X)$, $H(X, Y)$. Each term should not include a series. Hint: Is $H(Y|X) = H(Y - X|X)$?

c. Compute $H(Y)$, $H(X|Y)$, and $I(X; Y)$. If necessary, answers may include a series.

**Solution:**

(a) Since $X$ represents the number of coin tosses until the first head appears, it is Geometrically distributed, i.e., $X \sim G(q)$.

$$p(x = k) = \begin{cases} (1-q)^{k-1}q & \text{if } k > 0; \\ 0 & \text{if } k \leq 0. \end{cases}$$

Similarly, $Y$ is Negative Binomial distributed, i.e., $Y \sim NB(2, 1-q)$.

$$p(y = n) = \begin{cases} (n-1)(1-q)^{n-2}q^2 & \text{if } n > 1; \\ 0 & \text{if } n \leq 0. \end{cases}$$

Since the coin tosses are independent, by knowing $X$, the distribution of $Y$ is Geometric distributed with an initial value at $X$, i.e.,

$$p(y = n|x = k) = \begin{cases} (1-q)^{n-k-1}q & \text{if } n > k; \\ 0 & \text{if } n \leq k. \end{cases}$$

Assuming the second head toss was at $n$, the distribution of $X$ is uniform over all values between 1 and $n - 1$, i.e.,

$$p(x = k|y = n) = \begin{cases} \frac{1}{n-1} & \text{if } 1 \leq k \leq n-1; \\ 0 & \text{else.} \end{cases}$$

(b) The computation of $H(X), H(Y|X)$ is immediate by definition,

$$\begin{aligned} H(X) &= \frac{H_b(q)}{q}, && (22) \\ H(Y|X) &= H(Y-X|X) && (23) \\ &= H(Y-X) && (24) \\ &= \frac{H_b(q)}{q}. && (25) \\ & && (26) \end{aligned}$$

$H(X), H(Y|X)$ are equal since $X$ and $Y-X$ are both geometrically distributed with the same success probability. From the properties of joint entropy, we have that

$$H(X,Y) = H(X) + H(Y|X) = \frac{2H_b(q)}{q}, \tag{27}$$

(c) From the definition of entropy,

$$\begin{aligned} H(X|Y) &= \sum_{y\in\mathcal{Y}} \Pr(Y=y) H(X|Y=y) \\ &= \sum_{y\in\mathcal{Y}} \Pr(Y=y)\log(y-1) \\ &= \sum_{y=2}^{\infty} (y-1)(1-q)^{y-2} q^2 \log(y-1). \\ H(Y) &= H(X,Y) - H(X|Y) \\ &= \frac{2H_b(q)}{q} - \sum_{y=2}^{\infty}(y-1)(1-q)^{y-2}q^2\log(y-1). \\ I(X;Y) &= H(X) - H(X|Y) \\ &= \frac{H_b(q)}{q} - \sum_{y=2}^{\infty}(y-1)(1-q)^{y-2}q^2\log(y-1). \end{aligned} \tag{28}$$

19. **Inequalities** Copy each relation to your notebook and write $\leq$, $\geq$ or $=$, prove it.

(a) Let $X$ be a discrete random variable. Compare $\frac{1}{2^{H(X)}}$ vs. $\max_x p(x)$.

(b) Let $H_b(a)$ denote the binary entropy for $a \in [0,1]$ and $H_{ter}$ is the ternary entropy i.e. $H_{ter}(a,b,c) = -a\log a - b\log b - c\log c$, where $p_1, p_2, p_3 \in [0,1]$, and $p_1 + p_2 + p_3 = 1$.
Compare $H_{ter}(ab, a\bar{b}, \bar{a})$ vs $H_b(a) + \bar{a}H_b(b)$.

**Solution:**

(a) Let us show that $\frac{1}{2^{H(X)}} \le \max_x p(x)$.

$$\begin{aligned}
\frac{1}{2^{H(X)}} &= 2^{\mathbb{E}_X[\log p(X)]} \\
&\overset{(a)}{\le} 2^{\log \mathbb{E}_X[p(X)]} \\
&= \mathbb{E}_X[p(X)] \\
&\le \max_x p^2(x) \\
&\le \max_x p(x),
\end{aligned}$$

where $(a)$ follows fron Jensen's inrquality.

(b) We show that $H_{ter}(ab, a\bar{b}, \bar{a}) = H_b(a) + aH_b(b)$.

$$\begin{aligned}
H_{ter}(ab, a\bar{b}, \bar{a}) &= -ab\log(ab) - a\bar{b}\log(a\bar{b}) - \bar{a}\log\bar{a} \\
&= -ab\log a - ab\log b - a\bar{b}\log a - a\bar{b}\log\bar{b} - \bar{a}\log\bar{a} \\
&= -(ab + a\bar{b})\log a - ab\log b - a\bar{b}\log\bar{b} - \bar{a}\log\bar{a} \\
&= -a\log a + a(-b\log b - \bar{b}\log\bar{b}) - \bar{a}\log\bar{a} \\
&= H_b(a) + aH_b(b)
\end{aligned}$$

20. **True or False of a constrained inequality (21 Points)**:

Given are three discrete random variables $X, Y, Z$ that satisfy $H(Y|X,Z) = 0$.

(a) Copy the next relation to your notebook and write **true** or **false**.

$$I(X;Y) \ge H(Y) - H(Z)$$

(b) What are the conditions for which the equality $I(X;Y) = H(Y) - H(Z)$ holds.

(c) Assume that the conditions for $I(X;Y) = H(Y) - H(Z)$ are satisfied. Is it true that there exists a function such that $Z = g(Y)$?

**Solution:**

(a) True. Consider,

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - H(Y|X) + H(Y|X,Z) \\ &= H(Y) - H(Z|X) + H(Z|X,Y) \\ &\overset{(a)}{\geq} H(Y) - H(Z|X) \\ &\overset{(b)}{\geq} H(Y) - H(Z), \end{aligned}$$

where $(a)$ follows from $H(Z|X,Y) \geq 0$ and $(b)$ follows from $H(Z) \geq H(Z|X)$ (conditioning reduces entropy).

(b) We used two inequalities; the first becomes equality if $Z$ is a deterministic function of $(X,Y)$, and the second becomes equality if $Z$ is independent of $X$.

(c) False. For example, $X \sim Bern(\alpha)$, $Z \sim Bern(0.5)$, $Y = X \oplus Z$ and $X$ is independent of $Z$. All conditions are satisfied, and there is no such function.

21. **True or False of**: Copy each relation to your notebook and write **true** or **false**. If true, prove the statement, and if not provide a counterexample.

(a) Let $X - Y - Z - W$ be a Markov chain, then the following holds:

$$I(X;W) \leq I(Y;Z).$$

(b) For two probability distributions, $p_{XY}$ and $q_{XY}$, that are defined on $\mathcal{X} \times \mathcal{Y}$, the following holds:

$$D(p_{XY}||q_{XY}) \geq D(p_X||q_X).$$

(c) If $X$ and $Y$ are dependent and also $Y$ and $Z$ are dependent, then $X$ and $Z$ are dependent.

**Solution:**

(a) True. By the given Markov, we have that $I(X, Y; W) \leq I(X, Y; Z)$. By the facts that $I(X; Z|Y) = 0$ and $I(Y; W|X) \geq 0$, we get the desired inequality.

(b) True. Consider:

$$\begin{aligned} D(p_{XY}||q_{XY}) &= \sum_{x,y} p(x, y) \log \frac{p(x)}{q(x)} + \sum_{x,y} p(x, y) \log \frac{p(y|x)}{q(y|x)} \\ &= D(p_X||q_X) + \sum_{x} p(x) D(p_{Y|X=x}||q_{Y|X=x}) \\ &\geq D(p_X||q_X), \end{aligned}$$

where the inequality follows from the non-negativity of KL divergence.

(c) False. For any two independent random variables $X$ and $Z$, we can take $Y$ as the pair $(X, Z)$ which results a contradiction.

22. **Cross entropy**:
Often in Machine learning, cross entropy is used to measure performance of a classifier model such as neural network. Cross entropy is defined for two PMFs $P_X$ and $Q_X$ as

$$H(P_X, Q_X) \triangleq -\sum_{x \in \mathcal{X}} P_X(x) \log Q_X(x).$$

In a shorter notation we write as

$$H(P, Q) \triangleq -\sum_{x \in \mathcal{X}} P(x) \log Q(x).$$

Copy each of the following relations to your notebook and write **true** or **false** and provide a proof/disproof.

(a) $0 \leq H(P, Q) \leq \log |\mathcal{X}|$ for all $P, Q$.

(b) $\min_Q H(P, Q) = H(P, P)$ for all $P$.

(c) $H(P, Q)$ is concave in the pair $(P, Q)$.

(d) $H(P, Q)$ is convex in the pair $(P, Q)$.

**Solution:**

(a) **False.**
First, note that $H(P,Q)$ can be rewritten as

$$\begin{aligned} H(P,Q) &= -\sum_{x\in\mathcal{X}} P(x)\log Q(x) \\ &= \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)} - \sum_{x\in\mathcal{X}} P(x)\log P(x) \\ &= D(P||Q) + H_P(X). \end{aligned} \tag{29}$$

Thus, it obvious that $H(P,Q) \geq 0$. However, if we let $P_{\text{unif}}$ be the uniform measure on $\mathcal{X}$, then

$$\begin{aligned} H(P_{\text{unif}},Q) &= D(P_{\text{unif}}||Q) + H_{P_{\text{unif}}}(X) \\ &= D(P_{\text{unif}}||Q) + \log|\mathcal{X}| \\ &\geq \log|\mathcal{X}|, \end{aligned} \tag{30}$$

due to the fact that $D(P_{\text{unif}}||Q) \geq 0$. Now, because $D(P_{\text{unif}}||Q) = 0$ if and only if $Q = P_{\text{unif}}$, by taking any $Q \neq P_{\text{unif}}$, we will get that $D(P_{\text{unif}}||Q) > 0$, which means that $H(P_{\text{unif}},Q) > \log|\mathcal{X}|$ for any $Q \neq P_{\text{unif}}$, contradicting the claim that $H(P,Q) \leq \log|\mathcal{X}|$ for all $P,Q$.

(b) **True.**
This follows from the simple observation that $D(P||Q) \geq 0$ for all $(P,Q)$, and thus

$$\begin{aligned} H(P,Q) &= D(P||Q) + H_P(X) \\ &\geq H_P(X), \end{aligned} \tag{31}$$

with equality if and only if $Q = P$.

(c) **False.**
If $H(P,Q)$ is concave in the pair $(P,Q)$ then it must be concave in $P$ and $Q$ separately. However, it easy to see that $H(P,Q)$ is convex function in $Q$ (for fixed $P$) because $-\log(\cdot)$ is convex.

(d) **False.**
If $P = Q$, then $H(P,Q) = H_P(X)$, which is a concave function of $P$.

23. **Properties of mutual information:** A joint distribution is given by $P(x, \theta, y) = P(x)P(\theta)P(y|x, \theta)$. Answer the following three questions:

(a) **True/False**: Is it true that there is a Markov chain $X - Y - \theta$? Prove or provide a counter example.
**Solution:** False. Counterexample, let $X$ and $\theta$ be two independent random variables, each distributed according to Bernoulli(0.5). Also, let $Y = X \oplus \theta$. One can check that $H(X|Y) \neq H(X|Y, \theta)$.

(b) **Inequalities:** Fill (and prove) one of the relations $\leq, =, \geq$ between the following expressions :

$$I(X;Y) \;\; ??? \;\; I(X;Y|\theta).$$

**Solution:** Consider the following chain of inequalities:

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

where $(a)$ follows from the independence of $X$ and $\theta$, and $(b)$ follows from conditioning reduces entropy. Therefore, $I(X;Y) \leq I(X;Y|\theta)$.

(c) **Convex/Concave:** Determine whether the mutual information, $I(X_1; X_2)$ is convex OR concave function of $P(x_2|x_1)$ for a fixed $P(x_1)$. **Hint: You can use your answers from the previous questions.** You can not use the results we showed in class!
**Solution:** We showed in class that mutual information is convex. Define:

$$\begin{aligned} P(\theta) &\sim Bern(\lambda) \\ P_{Y|X,\theta=0} &= P^1_{Y|X} \\ P_{Y|X,\theta=1} &= P^2_{Y|X} \end{aligned} \tag{32}$$

, where $\lambda \in [0, 1]$, and $P^i_{Y|X}$ are two conditional distributions. From the previous question, we have $I(X;Y) \leq I(X;Y|\theta)$ and substituting (32) into this result shows the desired convexity.