Information Theory (5XSE0)
Ch.1: Information Measures

Hamdi Joudeh

TU/e (Q3 2020-2021)

Reading

- Cover & Thomas, Ch. 2 (excluding 2.7, 2.9).
- Supplementary: Gallager, Ch. 2.

Before we start

- Throughout the course, we focus on discrete-valued random variables. Therefore, by a random variable we implicitly mean a *discrete* random variable. Continuous-valued random variables will only be considered towards the end of the course.
- $X \sim p_X(x)$ means that a random variable X has a probability mass function (pmf) $p_X(x)$. X takes values on a finite alphabet $\mathcal{X}$. We have $p_X(x) \geq 0$ for all $x \in \mathcal{X}$, and $\sum_{x \in \mathcal{X}} p_X(x) = 1$.
- For brevity, we will denote $p_X(x)$ by $p(x)$. In the presence of a second random variable $Y \sim p_Y(y)$, which takes values on $\mathcal{Y}$, we use $p(y)$ to denote $p_Y(y)$. Here $p(y)$ and $p(x)$ should not be confused: these are two different pmfs, where $p(y)$ is associated with Y and $p(x)$ is associated with X.
- The Cartesian product of $\mathcal{X}$ and $\mathcal{Y}$ is defined as $\mathcal{X} \times \mathcal{Y} \equiv \big\{(x, y) : x \in \mathcal{X}, y \in \mathcal{Y}\big\}$. Therefore, the joint pmf $p_{XY}(x, y)$ of X and Y, denoted by $p(x, y)$ for brevity, is defined over the alphabet $\mathcal{X} \times \mathcal{Y}$.
- For a function $g(X)$, the expected value of $g(X)$ is defined as $\mathbb{E}g(X) \equiv \sum_{x \in \mathcal{X}} p(x)g(x)$. Sometimes to emphasize the pmf, we write the expectation as $\mathbb{E}_p g(X)$ or $\mathbb{E}_{p(x)} g(X)$. This is especially useful when we have two different pmfs, say $p(x)$ and $q(x)$, on the same alphabet $\mathcal{X}$.

Information is the *resolution of uncertainty*. The more uncertain we are about something, the more information there is to gain. Consider an experiment of tossing a coin and observing its outcome (or asking a yes/no question). If the coin is extremely bent and gives heads 95% of the time, then observing heads does not give us much information (as we could have predicted the outcome). If the coin is fair, we gain more information when observing the outcome as there is no way to predict it beforehand (heads and tails are equally likely). We will see how to formalize this intuition through entropy.

1 Entropy

Consider a random variable X taking values in $\mathcal{X}$ with a pmf of $p(x)$. We wish to quantify the amount of *surprise* in the event of observing an outcome $x \in \mathcal{X}$. Intuitively, an outcome x_1 with a low probability $p(x_1)$ is more surprising than an outcome x_2 with a high probability $p(x_2)$, and therefore the amount of surprise should increase with $1/p(x)$. We quantify surprise using the logarithm as follows:

$$\text{surprise}(x) = \log \frac{1}{p(x)} = -\log p(x). \tag{1}$$

Surprise in intimately related to *information*: a more surprising outcome is more informative. For instance, if it almost never rains in summer, the event of no rain in August is not surprising at all, and a forecast that confirms this does not relay much information. On the contrary, the event of having a rainy week in August is very surprising, and a forecast that predicts this relays much more information.

Why the log function in (1)? There are formal ways to justify this choice of function, but here we will stick with one intuitive justification. Consider the event of observing two independent outcomes x_1 and x_2. This event has a probability of $p(x_1, x_2) = p(x_1)p(x_2)$ and a surprise of $\log \frac{1}{p(x_1)p(x_2)} = \log \frac{1}{p(x_2)} + \log \frac{1}{p(x_1)}$. Accepting that surprise and information are the same thing, the information gained by observing (x_1, x_2) must be equal to the information gained from x_1 plus the information gained from x_2. Laws of logarithms guarantee that this is indeed the case, and we have $\text{surprise}(x_1, x_2) = \text{surprise}(x_1) + \text{surprise}(x_2)$

A common measure of information is entropy, which is precisely the expected amount of surprise.

Definition 1. For $X \sim p(x)$, the (Shannon) entropy $H(X)$ is defined as

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

In the definition of entropy, we use the convention $0 \log 0 = 0$, and adding zero-probability outcomes does not change entropy. Note that entropy only depends on the pmf of X, and does not depend on the values that X takes. Moreover, note the difference between $p(x)$ and $p(X)$: the former is a deterministic function that maps each x to a probability; while the latter is a random variable defined as

$$p(X) = p(x), \text{with probability (w.p.) } p(x).$$

This self-referential expression may take some time to get used to. We often take the base of the logarithm to be 2, and in this case entropy is measured in *bits*. Sometimes, base e is used for which entropy is measured in *nats*. 1 bits is equal to $\ln 2 = 0.693$ nats, while 1 nat is equal to $\log_2 e = 1.443$ bits. When we wish to highlight the base a of the logarithm, we denote the entropy by $H_a(X)$. Changing between arbitrary bases a and b is straightforward as seen in the properties listed further on. First we look at examples.

Example 1. Consider X which is uniformly distributed over $\{0, 1, 2, 3\}$. Here we have an equal surprise of $\log 4$ for all outcomes, and the entropy is given by

$$H(X) = 4 \times 0.25 \times \log 4 = \log 4 = 2 \text{ bits}.$$

This makes sense as we can represent each of the 4 outcomes using 2 bits, e.g. $\{00, 01, 10, 11\}$. We may claim that each outcome relays 2 bits of information, and hence the expected value is 2 bits.

Example 2. Consider Y which takes values in $\{0, 1, 2, 3\}$ according to the following pmf: $p(0) = 0.5$, $p(1) = 0.25$ and $p(2) = p(3) = 0.125$. The surprise for each of these outcomes is gives as follows: $-\log p(0) = 1$, $-\log p(1) = 2$ and $-\log p(2) = -\log p(3) = 3$, from which the entropy is given by

$$H(Y) = 0.5 \times 1 + 0.25 \times 2 + 2 \times 0.125 \times 3 = 1.75 \text{ bits}.$$

The entropy of Y is less than the entropy of X from Example 1: outcomes of Y relay less information than outcomes of X on average. Intuitively, this is because all outcomes of X are equally likely, making it impossible to predict X. On the other hand, we have a 75% chance that Y will be 1 or 2, and this decrease in uncertainty translates to a decrease in information. Another interpretation is obtained by representing each outcome of Y by a binary string of length equal to its surprise, e.g. $\{0, 10, 110, 111\}$. The least surprising outcome relays only 1 bit, while most surprising outcomes relay 3 bits each. These more informative outcomes are less likely to occur, and on average Y conveys 1.75 bits of information.

The assignment of binary strings of different lengths to outcomes of Y in Example 2 may seem arbitrary at first sight. We will learn how to do this systematically in the next chapter, and prove that the above binary representation is the shortest possible representation of Y on average. Therefore it is indeed a good measure of information content, which also matches the entropy. In this chapter, we focus on understanding the mathematical properties of entropy. Here are some basic properties of entropy:

- $H(X) \geq 0$ (non-negativity).
- $H(X) = 0$ if and only if X is a constant (i.e. $p(x)$ has a single mass point).
- $H_b(X) = (\log_b a) \cdot H_a(X)$ (change of base).

Exercise 1. Prove that the the above properties hold.

1.1 Binary Entropy

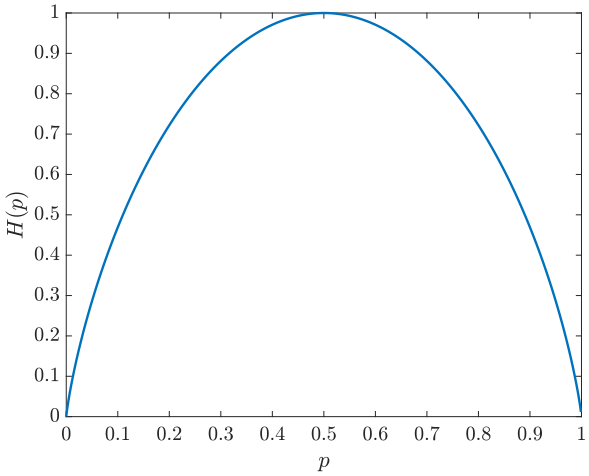
Suppose $X \sim \text{Bern}(p)$ for some $p \in [0, 1]$, i.e. X is a Bernoulli random variable given by

$$X = \begin{cases} 1, & \text{w.p. } p \\ 0, & \text{w.p. } 1 - p. \end{cases}$$

The entropy of X is given by

$$H(X) = -p \log p - (1 - p) \log(1 - p). \tag{2}$$

With a slight abuse of notation, we will use $H(p)$ to denote the entropy in (2) as a function of p. This is known as the *binary entropy function*. An illustration is shown below.



We can see from the graph that when $p = 0$ or 1, we have $H(p) = 0$. This makes sense, as in these cases the outcome is not random at all, and there is no uncertainty (and hence no surprise). On the other hand, $H(p)$ is maximized when $p = \frac{1}{2}$, which gives us an entropy of 1 bit. This also makes sense, as in this case the outcome cannot be predicted a priori. The entropy decreases as we move into the direction $p < 0.5$ or $p > 0.5$. Taking the case $p < 0.5$ for example, we know that the outcome here is more likely to be 0 than 1, and therefor there is less uncertainty on average compared to the case where $p = 0.5$.

1.2 Joint Entropy

Here we consider the (joint) entropy of a pair of random variables X and Y with a joint pmf $p(x, y)$. Note that X and Y can be considered as a single vector-valued random variable $(X, Y) \sim p(x, y)$.

Definition 2. For $(X, Y) \sim p(x, y)$, the joint entropy $H(X, Y)$ is defined as

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

The above extends to n jointly distributed random variables $(X_1, X_2, \ldots, X_n) \sim p(x_1, x_2, \ldots, x_n)$, with joint entropy given by $H(X_1, X_2, \ldots, X_n) = -\mathbb{E}\log p(X_1, X_2, \ldots, X_n)$.

1.3 Conditional Entropy

Now suppose that we observe the value of Y, and it turns out that $Y = y$. Conditioned on the event $\{Y = y\}$, the random variable X has an updated (*posterior*) pmf of $p_{X|Y}(x|y)$, denoted by $p(x|y)$ for brevity. If X and Y are independent, we will have $p(x|y) = p(x)$, however this is not the case in general. Our uncertainty about X generally changes given that $Y = y$, and the new uncertainty is captured by

$$H(X|Y=y) \equiv -\sum_{x\in\mathcal{X}} p(x|y)\log p(x|y).$$

The conditional entropy of X given Y is the average of the quantity $H(X|Y = y)$ over the pmf $p(y)$.

Definition 3. For $(X, Y) \sim p(x, y)$, the conditional entropy $H(X|Y)$ is defined as

$$\begin{aligned} H(X|Y) &\equiv \sum_{y\in\mathcal{Y}} p(y)H(X|Y=y) \\ &= -\sum_{y\in\mathcal{Y}} p(y)\sum_{x\in\mathcal{X}} p(x|y)\log p(x|y) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x|y) \\ &= -\mathbb{E}_{p(x,y)}\log p(X|Y) = \mathbb{E}_{p(x,y)}\log\frac{1}{p(X|Y)}. \end{aligned}$$

Intuition. $H(X|Y)$ can be understood as the average remaining uncertainty in X after observing Y.

Example 3. Consider binary random variables X and Y with a joint distribution given by

	$Y=0$	$Y=1$
$X=0$	0.5	0.2
$X=1$	0.1	0.2

Here $H(X) = 0.8813$, $H(X|Y = 0) = 0.65$, $H(X|Y = 1) = 1$ and $H(X|Y) = 0.79$.

Note that $H(X|X) = H(Y|Y) = 0$, i.e. there is no uncertainty about a random variable once its value has been revealed. This observation is generalized through the following exercise.

Exercise 2. Suppose that $Y = f(X)$, where $f : \mathcal{X} \to \mathcal{Y}$ is a bijective function (look it up if necessary!). Show that $H(Y|X) = H(X|Y) = 0$. In this case, knowing X is as good as knowing Y, and vice-versa.

1.4 Chain Rule

We now revisit the joint entropy in light of the conditional entropy. It turns out that the joint entropy of (X, Y) is equal to the entropy of X plus the conditional entropy of Y given X.

Theorem 1. (Chain rule for entropy: 2 random variables). *We have*

$$\begin{aligned} H(X,Y) &= H(X) + H(Y|X) \\ &= H(Y) + H(X|Y). \end{aligned}$$

Proof. Starting from the definition of the joint entropy, we write:

$$\begin{aligned} H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x)p(y|x) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\ &= -\sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\ &= H(X) + H(Y|X). \end{aligned}$$

The second equality in Theorem 1, i.e. $H(X,Y) = H(Y) + H(X|Y)$, is shown similarly by using the expansion $p(x,y) = p(y)p(x|y)$ in the above steps instead of $p(x,y) = p(x)p(y|x)$. □

Intuition. Recall that information is the *resolution of uncertainty*: information is gained by observing (or learning) the outcome of a random variable. The information gained by learning (X,Y) is equal to the information gained by first learning X, plus the remaining information gained by learning Y given that we already know X. By symmetry, we reach the same conclusion by learning Y first, and then learning X given that we already know Y.

Exercise 3. Use the fact that $p(X,Y) = p(X)p(Y|X)$, and hence $\log p(X,Y) = \log p(X) + \log p(Y|X)$, to rewrite the proof of Theorem 1 in a *tidier* fashion. *Hint: use the* $\mathbb{E}(\cdot)$ *notation.*

We now present a conditional version of Theorem 1.

Corollary 1. *For* $(X,Y,Z) \sim p(x,y,z)$*, we have*

$$\begin{aligned} H(X,Y|Z) &= H(X|Z) + H(Y|X,Z) \\ &= H(Y|Z) + H(X|Y,Z). \end{aligned}$$

Proof.

$$\begin{aligned} H(X,Y|Z) &= \sum_{z\in\mathcal{Z}} p(z) H(X,Y|Z=z) \\ &= \sum_{z\in\mathcal{Z}} p(z)\Big[H(X|Z=z) + H(Y|X,Z=z)\Big] \\ &= H(X|Z) + H(Y|X,Z). \end{aligned}$$

In the above, we used the fact that $H(X,Y|Z=z)$ is just the joint entropy of (X,Y) with respect to the distribution $p(x,y|z)$ instead of $p(x,y)$, and the applied the chain rule. □

We now extend the chain rule to n jointly distributed random variables $(X_1, X_2, \ldots, X_n)$, which have a joint pmf $p(x_1, x_2, \ldots, x_n)$.

Theorem 2. (Chain rule for entropy). *We have*

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1).$$

Note that the set of random variable $\{X_{i-1}, \ldots, X_1\}$ is empty for $i = 1$, and the chain rule in Theorem 2 is equivalently written as

$$H(X_1, X_2, \ldots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) + \cdots + H(X_n|X_{n-1}, X_{n-2}, \ldots, X_1).$$

Proof. We can prove Theorem 2 by extending the proof of Theorem 1 to n random variables (try this!). To avoid repetition, we present an alternative proof using *induction*. This goes as follows:

- We prove the statement in Theorem 2 for $n = 2$ (given already in Theorem 1).
- We then assume that the statement in Theorem 2 holds for any $n - 1$.
- We use this assumption to show that the statement must also hold for n.
- The above steps imply that the statement must hold for any n (i.e if it holds for $n = 2$, then it must hold for $n = 3$. If it holds for $n = 3$, then it must hold for $n = 4$, and so on).

Now we assume that the statement in Theorem 2 holds for $n - 1$, that is

$$H(X_1, X_2, \ldots, X_{n-1}) = \sum_{i=1}^{n-1} H(X_i|X_{i-1}, \ldots, X_1). \tag{3}$$

Proceeding to the case with n random variables, we write

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &= H(X_n|X_1, X_2, \ldots, X_{n-1}) + H(X_1, X_2, \ldots, X_{n-1}) && (4)\\ &= H(X_n|X_1, X_2, \ldots, X_{n-1}) + \sum_{i=1}^{n-1} H(X_i|X_{i-1}, \ldots, X_1) && (5)\\ &= \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1). \end{aligned}$$

In (4), we used the chain rule for 2 random variable (Theorem 1), while (5) is obtained from the assumption in (3). It follows that the statement must hold for n random variables given that it holds for $n-1$ random variables. This completes the proof. □

As a final remark on the chain rule in Theorem 2, we have chosen a certain order of expansion: X_1, then X_2 given X_1, then X_3 given (X_2, X_1), and so on. This expansion can take any order in general. Take $\pi(1), \pi(2), \ldots, \pi(n)$ to be a permutation of $1, 2, \ldots, n$. A more general chain rule is given by

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_{\pi(i)}|X_{\pi(i-1)}, \ldots, X_{\pi(1)}).$$

2 Mutual Information

The mutual information is a measure of the amount of information that one random variable contains about another random variable. Equivalently, it is the reduction in the uncertainty of one random variable due to the knowledge of another random variable.

Definition 4. For $(X, Y) \sim p(x, y)$, the mutual information $I(X; Y)$ is defined as

$$\begin{aligned} I(X; Y) &\equiv \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\ &= \mathbb{E}_{p(x,y)} \log \frac{p(X, Y)}{p(X)p(Y)}. \end{aligned}$$

In the definition of the mutual information, two different distributions over the same alphabet $\mathcal{X} \times \mathcal{Y}$ appear: the *joint distribution* $p(x, y)$, and the *product distribution* $p(x)p(y)$. Note that the two distributions are equal if X and Y are independent, and in this case we get $I(X;Y) = \mathbb{E}_{p(x,y)} \log 1 = 0$. This makes sense, as a pair of independent random variables contain no information about each other.

2.1 Relationship between mutual information and entropy

Theorem 3. (Mutual information and entropy). *We have*

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{6}$$

Proof. Starting from the definition of the mutual information, we write:

$$\begin{aligned} I(X;Y) &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log p(x,y) - \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log p(x)p(y) \\ &= -H(X,Y) - \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log p(x) \; - \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log p(y) \\ &= -H(X,Y) + H(X) + H(Y). \end{aligned}$$

□

Exercise 4. Write a tidier proof for Theorem 3 using the $\mathbb{E}(\cdot)$ notation (see Exercise 3).

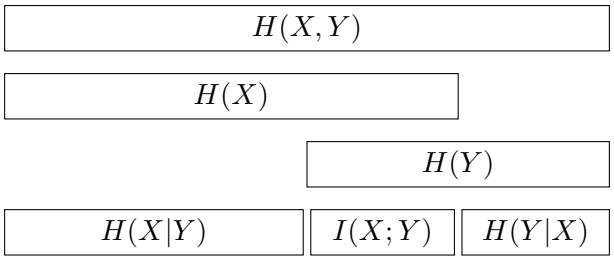
From Theorem 3, a few results follow (show that these results hold).

- $I(X;Y) = H(X) - H(X|Y)$ (reduction in entropy).
- $I(X;Y) = H(Y) - H(Y|X)$ (reduction in entropy).
- $I(X;Y) = I(Y;X)$ (symmetry).
- $I(X;X) = H(X)$ (self-information).

Intuition. The *prior* uncertainty in X is given by $H(X)$. On the other hand, the *posterior* uncertainty in X after observing Y is given by $H(X|Y)$. The mutual information $I(X;Y) = H(X) - H(X|Y)$ is the amount of information we *learn* about X by observing Y. By symmetry, this also applies in the other direction: $I(X;Y) = H(Y) - H(Y|X)$ is the amount of information we *learn* about Y by observing X.

Building on this intuition, the amount of information you learn about X by observing X itself is the amount of information contained in X (i.e. self-information), and therefore $I(X;X) = H(X)$.

The relationship between $H(X)$, $H(Y)$, $H(X,Y)$, $H(X|Y)$, $H(Y|X)$ and $I(X;Y)$ is illustrated in the following diagram. Note that $H(X) + H(Y)$ is in general larger (or no smaller) than $H(X,Y)$.



2.2 Conditional Mutual Information and Chain Rule

Now suppose that we observe a third random variable Z, and it turns out that $Z = z$. Conditioned on this event, the joint pmf of (X, Y) is given by $p(x, y|z)$, and their product pmf is given by $p(x|z)p(y|z)$. Using these posterior pmfs, the mutual information between X and Y, given that $Z = z$, is defined as

$$I(X;Y|Z=z) \equiv \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}$$

The conditional mutual information between X and Y given Z is the average of $I(X;Y|Z=z)$ over $p(z)$.

Definition 5. For $(X, Y, Z) \sim p(x, y, z)$, the conditional mutual information $I(X;Y|Z)$ is defined as

$$\begin{aligned} I(X;Y|Z) &\equiv \sum_{z\in\mathcal{Z}} p(z) I(X;Y|Z=z) \\ &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \sum_{z\in\mathcal{Z}} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} \\ &= \mathbb{E}_{p(x,y,z)} \log \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}. \end{aligned}$$

From the above definition, it follows that the conditional mutual information is given by

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

The mutual information also satisfies a chain rule, as seen next.

Theorem 4. (Chain rule for mutual information: 3 random variables). *We have*

$$\begin{aligned} I(X;Y,Z) &= I(X;Y) + I(X;Z|Y) \\ &= I(X;Z) + I(X;Y|Z). \end{aligned}$$

Proof. Starting by expressing the mutual information as a reduction in entropy, we write:

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

The second equality in Theorem 4 can be shown using similar steps. □

As for the entropy, the mutual information has a general chain rule given as follows.

Theorem 5. (Chain rule for mutual information). *We have*

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, \ldots, X_1)$$

Proof.

$$I(X_1, X_2, \ldots, X_n; Y) = H(X_1, X_2, \ldots, X_n) - H(X_1, X_2, \ldots, X_n|Y)$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} H(X_i|X_{i-1},\ldots,X_1) - \sum_{i=1}^{n} H(X_i|X_{i-1},\ldots,X_1,Y) \\
&= \sum_{i=1}^{n} \Big(H(X_i|X_{i-1},\ldots,X_1) - H(X_i|X_{i-1},\ldots,X_1,Y) \Big) \\
&= \sum_{i=1}^{n} I(X_i;Y|X_{i-1},\ldots,X_1).
\end{aligned}
$$

□

3 Relative Entropy (KL Divergence)

Consider the two pmfs $p(x)$ and $q(x)$ defined over the same alphabet $\mathcal{X}$. These can represent two potential distributions for some random variable X. The relative entropy, also called the Kullback–Leibler (KL) divergence, is a measure of closeness between the two pmfs $p(x)$ and $q(x)$. This is defined as follows.

Definition 6. The relative entropy between the two pmfs $p(x)$ and $q(x)$ is defined as

$$
\begin{aligned}
D(p\|q) &\equiv \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\
&= \mathbb{E}_p \log \frac{p(X)}{q(X)}.
\end{aligned}
$$

In the definition of relative entropy, the following conventions are used: $0 \log \frac{0}{0} = 0$, $0 \log \frac{0}{q} = 0$, and $p \log \frac{p}{0} = \infty$. This implies that for a pair of pmfs $p(x)$ and $q(x)$, if there exists $x \in \mathcal{X}$ such that $p(x) > 0$ and $q(x) = 0$, then $D(p\|q) = \infty$. Here the two distributions are far apart in the KL divergence sense.

Intuition. It is useful to think of the relative entropy $D(p\|q)$ as a measure of distance between the distributions $p(x)$ and $q(x)$. For instance, we will see later that $D(p\|q) = 0$ if and only if $p(x)$ and $q(x)$ for all $x \in \mathcal{X}$ (note that the "if" direction can be easily verified). Nevertheless, $D(p\|q)$ it is not a distance in the strict sense: it is not symmetric and does not satisfy the triangular inequality.

3.1 Conditional relative entropy and chain rule*

In this part, we will revert to highlighting the subscripts of pmfs for clarity, e.g. p_{XY} and $p_{Y|X}$.

Definition 7. For joint pmfs p_{XY} and q_{XY}, the conditional relative entropy $D\big(p_{Y|X}\|q_{Y|X}\big)$ is defined as

$$
\begin{aligned}
D\big(p_{Y|X}\|q_{Y|X}\big) &\equiv \sum_{x\in\mathcal{X}} p_X(x) \sum_{y\in\mathcal{Y}} p_{Y|X}(y|x) \log \frac{p_{Y|X}(y|x)}{q_{Y|X}(y|x)} \\
&= \mathbb{E}_{p_{XY}} \log \frac{p_{Y|X}(Y|X)}{q_{Y|X}(Y|X)}.
\end{aligned}
$$

Note that the conditional relative entropy $D\big(p_{Y|X}\|q_{Y|X}\big)$ is the average of $D\big(p_{Y|X}(y|x)\|q_{Y|X}(y|x)\big)$ over the pmf $p(x)$. We now present a chain rule for relative entropy.

Theorem 6. (Chain rule for relative entropy). *We have*

$$
D\big(p_{XY}\|q_{XY}\big) = D\big(p_X\|q_X\big) + D\big(p_{Y|X}\|q_{Y|X}\big).
$$

Proof. This is left as an exercise. □

3.2 Mutual information as a relative entropy

Recall that for a pair of jointly distributed random variable $(X, Y) \sim p(x, y)$, the product distribution is given by $p(x)p(y)$, which is a valid pmf over $\mathcal{X} \times \mathcal{Y}$. As it turns out, the mutual information $I(X;Y)$ is in fact the relative entropy between the joint distribution of (X, Y) and their product distribution.

Corollary 2. *For $(X, Y) \sim p(x, y)$, we have*

$$I(X;Y) = D\big(p(x,y) \| p(x)p(y)\big).$$

If X and Y are independent, we will have $p(x, y) = p(x)p(y)$, and in this case

$$I(X;Y) = D\big(p(x,y) \| p(x)p(y)\big) = D\big(p(x)p(y) \| p(x)p(y)\big) = 0.$$

This matches our intuition about the mutual information, being a measure of information that one random variables gives about another: if the random variables are independent, they contain no information about each other. We will next show that the relative entropy, and hence the mutual information, are non-negative, which is inline with relative entropy being a measure of distance, and the mutual information being a measure of information.

4 Information Inequality and its Consequences

We now present a property of the relative entropy which is of fundamental importance.

Theorem 7. (Information inequality). *For any two pmfs $p(x)$ and $q(x)$ defined over $\mathcal{X}$, we have*

$$D(p\|q) \geq 0 \tag{7}$$

with equality if and only if $p(x) = q(x)$ for all $x \in \mathcal{X}$.

Proof. For the proof, we need to recall the following:

- Jensen's inequality: $\mathbb{E} f\big(g(X)\big) \leq f\big(\mathbb{E} g(X)\big)$ for concave $f(t)$. For strictly concave $f(t)$, then equality $\mathbb{E} f\big(g(X)\big) = f\big(\mathbb{E} g(X)\big)$ implies that $g(X) = \mathbb{E} g(X)$ almost surely (i.e. $g(X)$ is a constant).
- $f(t) = \log(t)$ is strictly concave.

Let $\mathcal{A} = \{x : p(x) > 0\}$ be the support set of $p(x)$, i.e. a subset of the alphabet $\mathcal{X}$ over which the pmf $p(x)$ is strictly positive. Starting from the definition of the relative entropy, we write:

$$\begin{aligned}
-D(p\|q) &= -\sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\
&= -\sum_{x \in \mathcal{A}} p(x) \log \frac{p(x)}{q(x)} \\
&= \sum_{x \in \mathcal{A}} p(x) \log \frac{q(x)}{p(x)} \\
&= \mathbb{E}_p \log \frac{q(X)}{p(X)} \\
\text{(Jensen)} \quad &\leq \log \mathbb{E}_p \frac{q(X)}{p(X)} \\
&= \log \sum_{x \in \mathcal{A}} p(x) \frac{q(x)}{p(x)}
\end{aligned} \tag{8}$$

$$
\begin{aligned}
&= \log \sum_{x \in \mathcal{A}} q(x) \\
&\leq \log \sum_{x \in \mathcal{X}} q(x) \qquad (9) \\
&= \log 1 \\
&= 0.
\end{aligned}
$$

Next, we prove the following statement:

$$D(p\|q) = 0 \iff p(x) = q(x) \text{ for all } x \in \mathcal{X}.$$

- "$\Longleftarrow$" (the "if" direction): By setting $p(x) = q(x)$ for all $x \in \mathcal{X}$, we get $D(p\|q) = 0$.
- "$\Longrightarrow$" (the "only if" direction): Suppose that $D(p\|q) = 0$. Then equality must hold in (8) and (9).
  - Starting with (9), equality holds here only if $\sum_{x \in \mathcal{A}} q(x) = \sum_{x \in \mathcal{X}} q(x) = 1$. Therefore, the support of $q(x)$ must also be $\mathcal{A}$ to have equality in (9).
  - Since $\log t$ is strictly concave, equality in (8) implies that $\frac{q(X)}{p(X)}$ is a constant, that is $\frac{q(x)}{p(x)} = c$ for all $x \in \mathcal{A}$, and therefore $\sum_{x \in \mathcal{A}} q(x) = c \sum_{x \in \mathcal{A}} p(x) = c$.

  It follows that $c = \sum_{x \in \mathcal{A}} q(x) = 1$, and hence $p(x) = q(x)$ for all $x \in \mathcal{A}$ whenever $D(p\|q) = 0$.

□

The information inequality is also known as Gibbs' inequality.

4.1 Consequences of the information inequality

We now present a few important consequences of Theorem 7.

Theorem 8. (Non-negativity of mutual information). *For $(X, Y) \sim p(x, y)$, we have*

$$I(X;Y) \geq 0$$

with equality if and only if X and Y are independent.

Proof. From Corollary 2 and Theorem 7, we have $I(X;Y) = D\big(p(x,y)\|p(x)p(y)\big) \geq 0$, with equality if and only if $p(x, y) = p(x)p(y)$, that is X and Y are independent. □

Intuition. An information measure should be non-negative. Moreover, when X and Y are independent, they contain no information about each other. Conversely, if we can gain no information about X by observing Y (or about Y by observing X), then X and Y must be independent.

Theorem 9. (Uniform distribution maximizes entropy). *Let X be a random variable with alphabet $\mathcal{X}$, and let $|\mathcal{X}|$ denote the cardinality (i.e. number of elements) of $\mathcal{X}$. We have*

$$H(X) \leq \log |\mathcal{X}|$$

with equality if and only if X is uniformly distributed over $\mathcal{X}$.

Proof. Let $p(x)$ be the pmf of X, and let $u(x) = 1/|\mathcal{X}|$ be the uniform pmf over $\mathcal{X}$. We have the following

$$
\begin{aligned}
D(p\|u) &= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{u(x)} \\
&= -\sum_{x \in \mathcal{X}} p(x) \log u(x) + \sum_{x \in \mathcal{X}} p(x) \log p(x) \\
&= \log |\mathcal{X}| - H(X).
\end{aligned}
$$

Therefore, we have $0 \leq D(p\|u) = \log |\mathcal{X}| - H(X)$. □

Intuition. With a uniform distribution, the outcome cannot be predicted, and entropy is maximized. As a special case, we have the binary entropy (i.e. $|\mathcal{X}| = 2$) seen earlier, where $H(p) \leq \log 2 = 1$.

Theorem 10. (Conditioning reduces entropy). *For $(X, Y) \sim p(x, y)$, we have*

$$H(X|Y) \leq H(X)$$

with equality if and only if X and Y are independent.

Proof. We have $I(X;Y) = H(X) - H(X|Y) \geq 0$. □

Intuition. Information cannot hurt: the average uncertainty in X may only decrease or remain the same (but never increase) after observing a possibly correlated random variable Y. This may not hold for each realization of Y, i.e. one may have $H(X|Y = y) > H(X)$ for some y. However, on average we have $H(X|Y) \leq H(X)$. Check Example 3 again and verify these statements.

Theorem 11. (Independence bound on entropy). *For $(X_1, X_2, \ldots, X_n) \sim p(x_1, x_2, \ldots, x_n)$, we have*

$$H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$$

with equality if and only if the X_is are independent.

Proof. Starting from the chain rule (Theorem 2), we write:

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &= \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1) \\ &\leq \sum_{i=1}^{n} H(X_i) \end{aligned} \tag{10}$$

where the inequality follows from the fact that conditioning reduces entropy. It can also be verified (using Theorem 10) that equality holds in (10) if and only if all n random variables are mutually independent. □

5 Data Processing Inequality

Here we present another fundamental inequality which shows that data processing can only *destroy* information. Before stating this, we present some useful definitions.

Definition 8. The three random variables $(X, Y, Z) \sim p(x, y, z)$ form a *Markov chain* $X \to Y \to Z$ if

$$p(x, y, z) = p(x)p(y|x)p(z|y).$$

The above definition extends to n random variables $(X_1, X_2, \ldots, X_n) \sim p(x_1, x_2, \ldots, x_n)$, which are said to form a a Markov chain $X_1 \to X_2 \to \cdots \to X_n$ if

$$p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_2) \cdots p(x_n|x_{n-1}).$$

Intuition. Let i denote a time index and X_i be the state of some system at time i. If this state evolves according to a Markov chain, then a future state X_{i+1} is independent of past states $X_1, \ldots, X_{i-1}$ given the present state X_i. This is reflected in $p(x_{i+1}|x_i, x_{i-1}, \ldots, x_1) = p(x_{i+1}|x_i)$.

From Definition 8, we can derive the following properties:

- We have a Markov chain $X \to Y \to Z$ if and only if X and Z are conditionally independent given Y, that is $p(x, z|y) = p(x|y)p(z|y)$.

- The Markov chain $X \to Y \to Z$ implies $Z \to Y \to X$.
- If $Z = g(Y)$, i.e. Z is a function of Y, then $X \to Y \to Z$.

Exercise 5. Prove the above properties.

We are now ready to present the data processing inequality.

Theorem 12. (Data processing inequality). *If $X \to Y \to Z$, then*

$$I(X;Y) \geq I(X;Z).$$

Proof. Consider the mutual information term $I(X;Y,Z)$, with X on one side and (Y,Z) on the other. Using the chain rule for mutual information (Theorem 4), we expand $I(X;Y,Z)$ in two ways:

$$I(X;Y,Z) = I(X;Y) + I(X;Z|Y) \tag{11}$$
$$= I(X;Z) + I(X;Y|Z). \tag{12}$$

Next, we argue that $I(X;Z|Y) = 0$. Recall that due to the Markov chain $X \to Y \to Z$, the variables X and Z are conditionally independent given Y, i.e. $p(x,z|y) = p(x|y)p(z|y)$. It hence follows from Theorem 8 that $I(X;Z|Y=y) = 0$ for all $y \in \mathcal{Y}$, and therefore $I(X;Z|Y) = \sum_{y\in\mathcal{Y}} p(y)I(X;Z|Y=y) = 0$.

Given that $I(X;Z|Y) = 0$, we obtain $I(X;Y) = I(X;Z) + I(X;Y|Z)$ from (11) and (12). This implies

$$I(X;Y) \geq I(X;Z) \tag{13}$$

because $I(X;Y|Z) \geq 0$ (mutual information is non-negative). Note that equality in (13) holds if and only if $I(X;Y|Z) = 0$, i.e. the three random variables also form another Markov chain $X \to Z \to Y$. □

Intuition. Suppose that X is some latent (concealed) variable that we wish to learn the value of, Y is some data related to X which we have access to (noisy observations of X), and Z is a processed version of this data (see Corollary 3 below). Theorem 12 tells us that no processing of Y, deterministic or random, can increase the information it contains about X. Or equivalently, processing Y can only destroy (or at best retain) information contained about X. This may seem counter-intuitive at first sight: there are many cases where processing of raw data can help improve our inference. What Theorem 12 says, however, is that while data processing can help *extract* information, it cannot *increase* the amount of information.

Corollary 3. *If $Z = g(Y)$, then $I(X;Y) \geq I(X;Z)$. This holds as $X \to Y \to g(Y)$ is a Markov chain.*

Corollary 4. *Since $X \to Y \to Z$ implies $Z \to Y \to X$, then we also have $I(Y;Z) \geq I(X;Z)$.*

Corollary 5. *If $X \to Y \to Z$, then $I(X;Y|Z) \leq I(X;Y)$. (proof is left as an exercise!).*

Corollary 5 tells us that the dependency between X (concealed variable) and Y (observed data) is decreased, or remains unchanged, by observing Z (processed data). This make sense as $I(X;Y)$ is the amount of information we could learn about X by observing Y. As Z is related to Y, observing Z could possibly tell us something about X. This may only decrease the amount of information gained about X by observing Y.

6 Fano's Inequality

Consider a random variable X taking values in $\mathcal{X}$. Suppose that we do not know the value of X, however, we have access to a correlated random variable Y from which we wish to estimate the value of X. We calculate $\hat{X} = g(Y)$, where $\hat{X}$ is an estimate of X taking values in $\hat{\mathcal{X}}$, while $g : \mathcal{Y} \to \hat{\mathcal{X}}$ is some estimation function. In many scenarios, we have $\hat{\mathcal{X}} = \mathcal{X}$, but this need not be the case. For instance, X could be

binary with $\mathcal{X} = \{0, 1\}$, while $\hat{X}$ takes values in $\hat{\mathcal{X}} = \{0, 1, \mathrm{e}\}$, where $\hat{X} = g(Y) = \mathrm{e}$ indicates the event where X cannot be estimated and instead the estimator declares an error. Note that $X \to Y \to \hat{X}$.

An error event occurs whenever $\hat{X} \neq X$. The probability of this is

$$P_e \equiv \mathbb{P}\left\{\hat{X} \neq X\right\}.$$

The probability of error is an important figure of merit, as it tells us how well our estimator is doing. Fano's inequality finds a lower bound for P_e as follows.

Theorem 13. (Fano's inequality). *Let $\hat{X}$ be an estimate of X such that $X \to Y \to \hat{X}$ form a Markov chain, and let $P_e = \mathbb{P}\{\hat{X} \neq X\}$ be the corresponding probability of error. Then we have*

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y)$$

where $H(P_e) = -P_e \log(P_e) - (1 - P_e) \log(1 - P_e)$ is the binary entropy function.

Proof. We start by proving the first inequality $H(P_e) + P_e \log |\mathcal{X}| \geq H(X|\hat{X})$. We define an error random variable E as follows

$$E = \begin{cases} 1, & \text{if } X \neq \hat{X} \\ 0, & \text{if } X = \hat{X}. \end{cases}$$

Note that $E \sim \mathrm{Bern}(P_e)$, and hence $H(E) = H(P_e)$. Moreover, E is fully determined by X and $\hat{X}$ and hence $H(E|X, \hat{X}) = 0$. Now we expand the entropy term $H(E, X|\hat{X})$ in two ways as follows

$$\begin{aligned} H(E, X|\hat{X}) &= H(X|\hat{X}) + \overbrace{H(E|X, \hat{X})}^{=0} \\ &= \underbrace{H(E|\hat{X})}_{\substack{(a) \\ \leq H(P_e)}} + \underbrace{H(X|E, \hat{X})}_{\substack{(b) \\ \leq P_e \log |\mathcal{X}|}}. \end{aligned}$$

In the above, inequality (a) follows from $H(E|\hat{X}) \leq H(E) = H(P_e)$, i.e. conditioning cannot increase entropy. On the other hand, inequality (b) is obtained as follows

$$\begin{aligned} H(X|E, \hat{X}) &= \mathbb{P}\{E = 1\} H(X|E = 1, \hat{X}) + \mathbb{P}\{E = 0\} H(X|E = 0, \hat{X}) \\ &= P_e H(X|E = 1, \hat{X}) + (1 - P_e) H(X|E = 0, \hat{X}) \\ &= P_e H(X|E = 1, \hat{X}) \\ &\leq P_e \log |\mathcal{X}|. \end{aligned}$$

In the above, $H(X|E = 0, \hat{X}) = 0$ since given that $E = 0$, we have $X = \hat{X}$. The last inequality holds due to $H(X|E = 1, \hat{X}) \leq H(X) \leq \log |\mathcal{X}|$. Putting everything together, we obtain

$$H(X|\hat{X}) \leq H(P_e) + P_e \log |\mathcal{X}|.$$

To complete the proof of Theorem 13, it remains to show that $H(X|Y) \leq H(X|\hat{X})$. This follows from the data processing inequality: since $X \to Y \to \hat{X}$ is a Markov chain, we have $I(X; Y) \geq I(X; \hat{X})$, which is rewritten as $H(X) - H(X|Y) \geq H(X) - H(X|\hat{X})$. This immediately implies $H(X|Y) \leq H(X|\hat{X})$. □

Corollary 6. *Since $H(P_e)$ is a binary entropy term, we have $H(P_e) \leq 1$. Therefore, we can obtain a weaker (yet simpler) form of Fano's inequality as follows:*

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

The weaker form in the above corollary will be used later on in the course.