# 1
# *Information measures*

**Intended learning outcomes:**

- You can compute the entropy and conditional entropy for any discrete random variable and understand the basic properties of these two quantities, e.g. you can apply the chain rule or sub-additivity.
- You can compute mutual information and now how it relates to entropy and conditional entropy. You can apply the data-processing inequality for mutual information.
- You can compute the relative entropy and understand how entropy and mutual information can be expressed in terms of the relative entropy.

**Book reference:** Chapter 2 in Cover & Thomas[1], but we are not following it too closely.

[1] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991. ISBN 9780471748823. DOI: 10.1002/047174882X

## *1.1 Surprisal and entropy*

It is not immediately clear how to model our intuitive notion of "information" in a mathematical language. In this chapter we take a somewhat axiomatic approach to information measures, i.e. we try to build them up from our intuitive understanding of what entropy and information should be. But we will only really be able to justify the choices we make here once we start analysing practical problems in information theory, and see that the quantities we derive and investigate here pop up again and again as solutions.

### *1.1.1 Surprisal*

It turns out to be fruitful to start not by finding an expression for the information contained in a random variable, but rather the lack of information, or uncertainty inherent in a random experiment. Let us consider a discrete random variable $X$ taking values in $\mathcal{X}$ following the pmf $P_X(x) = p_x$. How surprised are we to see a particular outcome $x \in \mathcal{X}$ of this random experiment? Clearly this depends

on the probability $p_x$ and not the value of $x$ itself. In fact, we do not even need to know what $\mathcal{X}$ really is. On the one hand, if $p_x = 1$ we are not surprised at all since we already knew that we would see $x$. On the other hand, the smaller $p_x$ is the more surprised we are to see this particular outcome. If $p_x = 0$ we will never see $x$, so our surprise when seeing it anyway would be literally off the scale. Furthermore — and this turns out to be a very convenient choice — if we do a random experiment twice independently and both times observe $x$, we say that we will be twice as surprised as if we had seen $x$ once in a single random experiment.

The above notions can be formalised, and that is essentially what Shannon did when he introduce the notion of *surprisal*. Let us denote the surprisal of $x$ as $s(p_x)$. We want this function to satisfy the following three conditions:

1. **Monotonicity:** $s(p_x) = 0$ if $p_x = 1$ and $s(p_x)$ increases monotonically as $p_x$ decreases.
2. **Additivity:** The surprisal of seeing a pair of outcomes of independent random experiments is simply the sum of the individual surprisals, i.e. $s(p_x p_y) = s(p_x) + s(p_y)$.
3. **Normalisation:** $s(\frac{1}{2}) = 1$

We do not really need the condition $s(p_x) = 0$ if $p_x = 1$ under Point 1 as it follows directly from additivity. Can you see how?

It turns out that the only positive function that satisfies the first two is the logarithm. To show this one uses a result by Erdös that characterises additive functions, but that is beyond the scope here. We therefore pick

$$s(p_x) = \log \frac{1}{p_x} \,. \tag{1.1}$$

where the logarithm is taken to base 2 (as everywhere in these notes) so that the normalisation requirement is satisfied.

We can see the surprisal as another random variable, say $S$, that takes the value $s(p_x) = \log \frac{1}{p_x}$ with probability $p_x$. Since $S = S(X)$ is a function of $X$ we usually simply write this new random variable as

$$S(X) = \log \frac{1}{P_X(X)} \,. \tag{1.2}$$

### 1.1.2 *Entropy*

Entropy measures how much we can learn by looking at the outcome of a random experiments, or, in other words, how much uncertainty there is about the outcome. It is simply the expected surprisal of $X$.

Given a discrete random variable $X$, the *entropy* of $X$ is defined as

$$H(X) := \mathbb{E}[S(X)] = \mathbb{E}\left[\log \frac{1}{P_X(X)}\right] = \sum_{x\in\mathcal{X}} p_x \log \frac{1}{p_x} \tag{1.3}$$

Here and throughout we use the convention that $0 \log 0 = 0$. This is reasonable since $\lim_{\epsilon\to 0} \epsilon \log \epsilon = 0$, and thus we simply continuously extend the function to the point 0.

Verify that $\epsilon \log \epsilon \to 0$ as $\epsilon \to 0$.

Note again that the entropy $X$ is really only a function of the pmf of $X$, and in particular independent of the alphabet $\mathcal{X}$, in contrast to potential alternative uncertainty measures like the variance of $X$.

Sometimes we are interested in more than just the expected surprisal. The minimum surprisal, or min-entropy, for example, has applications in cryptography (see Chapter 3) and the variance of $S(X)$ has itself operational meaning in many information-theoretic problems when we go beyond first order asymptotics.

Can you find an expression for $\mathrm{Var}[S(X)]$ in terms of the probabilities $p_x$?

Now let us explore the entropy a bit. First we want to show the following basic property.

**Proposition 1.1.** *Let $X$ be a discrete random variable taking values in $\mathcal{X}$. We have*

$$H(X) \geq 0, \tag{1.4}$$

*with equality if and only if $X$ is deterministic.*

*Proof.* Since $p_x \leq 1$, we have $\log \frac{1}{p_x} \geq 0$ for every $x \in \mathcal{X}$, so the expectation of this quantity over $x$ must be non-negative too. In fact, $\log \frac{1}{p_x}$ equals 0 if and only if $p_x = 1$ and hence $H(X) = 0$ only if there exists an $x \in \mathcal{X}$ for which $p_x = 1$, which is the hallmark of a deterministic rv. □

The entropy is a strictly concave function of the probability mass function $P_X$. To see this, we first verify that $f(t) = t \log \frac{1}{t} = -t \log t$ is concave on $(0, 1)$ by taking its second derivative:

$$f'(t) = -\log t - \log e, \qquad f''(t) = -\frac{\log e}{t}. \tag{1.5}$$

Since the latter is always negative for $t \in (0, 1)$, the function is indeed strictly concave according to Lemma 0.12. Now since the entropy is simply the sum $\sum_{x\in\mathcal{X}} f(p_x)$ it is indeed a strictly concave function of the pmf. This simple property, together with Jensen's inequality, has profound implications. The first one is that the entropy has a unique maximum. Intuitively we would want that entropy is maximal when uncertainty about the outcome is greatest, namely when the rv is uniformly distributed. And this is indeed the case.

**Proposition 1.2.** *Let $X$ be a discrete random variable taking values in $\mathcal{X}$. We have*

$$H(X) \leq \log |\mathcal{X}|, \tag{1.6}$$

*with equality if and only if $X$ is uniformly distributed.*

The general case will be covered in the homework but here we give a proof for the case when the set $\mathcal{X}$ is a bit, i.e. when the random variable is binary.

*Proof for $\mathcal{X} = \{0,1\}$.* It is easy to verify by a simple computation that $H(X) = 1$ for a uniformly distributed random variable, so the difficulty is only in showing that this is the maximum and only achieved for the uniform distribution.

Let now $\{p, 1-p\}$ for $p \in [0,1]$ be a general pmf for the random variable $X$. We use the function $f(t) = -t \log t$ to simplify notation. Then we can write

$$\begin{aligned} H(X) &= f(p) + f(1-p) && (1.7)\\ &= \frac{1}{2}\left(f(p) + f(1-p)\right) + \frac{1}{2}\left(f(p) + f(1-p)\right) && (1.8)\\ &\leq f\left(\frac{1}{2}p + \frac{1}{2}(1-p)\right) + f\left(\frac{1}{2}p + \frac{1}{2}(1-p)\right) && (1.9)\\ &= f\left(\frac{1}{2}\right) + f\left(\frac{1}{2}\right) && (1.10)\\ &= 1\,. && (1.11) \end{aligned}$$

The inequality is due to Jensen's inequality and the strict concavity of $f$, and equality holds only if $p = 1 - p = \frac{1}{2}$, i.e. when the random variable $X$ follows the uniform distribution. □

Concavity in fact has even stronger consequences, and we will show a few additional properties of entropy later on using it.

**Example.** *The simplest example of a random variable is the Bernoulli random variable $X \sim \mathrm{Bern}(p)$ with $\mathcal{X} = \{0,1\}$ and $P_X(0) = p$ for $p \in [0,1]$. The entropy of the Bernoulli random variable is called the* binary entropy*,*

$$H(X) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p} =: h(p)\,. \tag{1.12}$$

*From the plot we can easily verify all the properties we discussed above.*

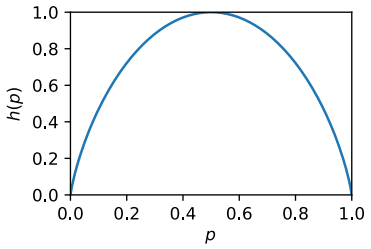


Figure 1.1: The binary entropy function.

Show that $h(p) = h(1-p)$.

## 1.2 Conditional entropy, and mutual information

### 1.2.1 Joint entropy

For two discrete random variables $X$ and $Y$ with joint pmf $P_{XY}(x,y) = p_{xy}$ we can simply consider $(X,Y)$ as one single random variable and use the same construction to define the surprisal of a tuple $(X,Y)$ as $S(X,Y) = -\log P_{XY}(X,Y)$. Its expectation is the *joint entropy*, $H(XY)$, given by

Find an example for which $H(XY) = H(X) = H(Y) = 1$.

$$H(XY) := \mathbb{E}[S(X,Y)] = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{xy}\log\frac{1}{p_{xy}} \tag{1.13}$$

The first thing to note is that — if $X$ and $Y$ are independent — then $p_{xy} = p_x \cdot p_y$ and thus the expression simplifies to

$$\begin{aligned} H(XY) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{xy}\log\frac{1}{p_x p_y} && (1.14)\\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{xy}\log\frac{1}{p_x} + \log\frac{1}{p_y} && (1.15)\\ &= \sum_{x\in\mathcal{X}} p_x\log\frac{1}{p_x} + \sum_{y\in\mathcal{Y}} p_y\log\frac{1}{p_y} && (1.16)\\ &= H(X) + H(Y)\,. && (1.17) \end{aligned}$$

This is not true in general though if the two random variables are correlated.

### 1.2.2 Conditional entropy

So why do these entropies not just add up? Fundamentally, this is because once we learn $X$ we might not be so surprised seeing some particular outcomes of the random variable $Y$ anymore. In fact, in the most extreme case, we have $Y = f(X)$ for some function $f$; hence, once we know that $X$ takes on the value $x$, we can immediately deduce that $Y$ will take on the value $f(x)$ with probability one, and thus there is no surprisal anymore! We model this "conditional suprisal" using the conditional pmfs, $P_{Y|X}(y|x) = p_{y|x}$, which leads us to conditional entropy.

The *conditional entropy* of $Y$ given $X$ is defined as

$$H(Y|X) = \mathbb{E}\left[\log\frac{1}{P_{Y|X}(Y|X)}\right] = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{xy}\log\frac{1}{p_{y|x}}\,. \tag{1.18}$$

This can be interpreted as the expectation of the entropy of $Y$ over all outcomes $X$. We sometimes use the notation $H(Y|X=x) = H(Y_x)$

to denote the entropy of the random variable $Y_x$ that follows the pmf $\{p_{y|x}\}_{y\in\mathcal{Y}}$, i.e., the pmf of $Y$ when we already know that $X = x$. Using this and the expression in (1.18) we can write the conditional entropy as

$$H(Y|X) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p_{xy} \log \frac{1}{p_{y|x}} \tag{1.19}$$

$$= \sum_{x\in\mathcal{X}} p_x \sum_{y\in\mathcal{Y}} p_{y|x} \log \frac{1}{p_{y|x}} \tag{1.20}$$

$$= \sum_{x} p_x \, H(Y|X = x) \,. \tag{1.21}$$

The last line which expresses the conditional entropy in terms of an average of (unconditional) entropies is particularly useful since it allows us to immediately conclude that the conditional entropy is also bounded from below and above, like the entropy. We thus have

$$0 \leq H(Y|X) \leq \log |\mathcal{Y}| \,. \tag{1.22}$$

Moreover, our definition of conditional entropy also allows us to establish a *chain rule* for the conditional entropy, which sometimes is in fact used as the definition of conditional entropy itself. This rule is very useful because it allows us to write the joint entropy as a sum of its parts, even if the two random variables are not independent.

**Proposition 1.3** (Chain Rule). *We have* $H(XY) = H(X) + H(Y|X)$.

*Proof.* We take advantage of $p_{xy} = p_x p_{y|x}$ to write

$$H(XY) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p_{xy} \log \frac{1}{p_{xy}} \tag{1.23}$$

$$= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p_{xy} \log \frac{1}{p_x} + \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p_{xy} \log \frac{1}{p_{y|x}} \tag{1.24}$$

$$= \sum_{x\in\mathcal{X}} p_x \log \frac{1}{p_x} + \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p_{xy} \log \frac{1}{p_{y|x}} \tag{1.25}$$

$$= H(X) + H(Y|X) \,. \tag{1.26}$$

□

Show that $H(Y|X) = H(Y)$ for independent random variables. Using the chain rule, find a different proof that $H(XY) = H(X) + H(Y)$ in this case.

Now we have put everything in place to show our first entropic inequality, which relates the entropy of two random variables with their joint entropy. This result shows the *sub-additivity* of entropy.

**Proposition 1.4** (Sub-Additivity). *Let $X$ and $Y$ be two discrete random variables. Then*

$$H(XY) \leq H(X) + H(Y)\,, \tag{1.27}$$

*or, equivalently, $H(X|Y) \leq H(X)$. Equality holds in either statement if and only if $X$ and $Y$ are independent.*

*Proof.* The equivalence of the two relations follows directly from the chain rule, we thus only need to show the second statement. We already know from Eq. (1.17) that equality hold if $X$ and $Y$ are independent. It remains to show that the inequality holds, and that it holds with equality only if $X$ and $Y$ are independent.

We start with Eq. (1.21), which states that

$$\begin{aligned} H(Y|X) &= \sum_x p_x H(Y|X=x) && (1.28)\\ &= \sum_x p_x \sum_y p_{y|x} \log \frac{1}{p_{y|x}} && (1.29)\\ &= \mathbb{E}\left[\sum_y p_{y|X} \log \frac{1}{p_{y|X}}\right] && (1.30) \end{aligned}$$

Note that sum inside the expectation is simply another expectation, as in the definition of entropy — but since we only want to apply Jensen's inequality on the outer expectation we spell this one out explicitly. Moreover, by definition of the conditional pmf we have $\mathbb{E}[p_{y|X}] = \sum_x p_x p_{y|x} = \sum_x p_{xy} = p_y$. Hence, using concavity of the entropy as a function of the pmf and Jensen's inequality for the outer expectation, we find

$$\begin{aligned} H(Y|X) &= \mathbb{E}\left[\sum_y p_{y|X} \log \frac{1}{p_{y|X}}\right] && (1.31)\\ &\leq \sum_y \Big(\mathbb{E}[p_{y|X}]\Big) \log \frac{1}{\mathbb{E}[p_{y|X}]} && (1.32)\\ &= \sum_y p_y \log \frac{1}{p_y} && (1.33)\\ &= H(Y)\,. && (1.34) \end{aligned}$$

Equality in Jensen's inequality only holds if either $X$ is deterministic or if $p_{y|x} = p_y$ for all $x$ and $y$, but this only holds if $X$ and $Y$ are in fact independent. □

The second relation in Eq. (1.27) can be strengthened by considering three random variables $X$, $Y$ and $Z$. In that case, we have

$$H(X|YZ) \leq H(X|Z)\,. \tag{1.35}$$

This is sometimes referred to as *strong sub-additivity*. The proof follows from (regular) sub-additivity, applied to the entropies $H(X|Y, Z = z)$ and $H(X|Z = z)$, and averaging the resulting inequalities.

Complete this argument to a formal proof.

### 1.2.3 Mutual information

We have already established that $H(XY) \neq H(X) + H(Y)$ in general, and hence also $H(Y|X) \neq H(Y)$ by the chain rule. The difference between these two quantities clearly tells us something about how much the uncertainty about $Y$ changes when we learn $X$, or in other words, about how much information $X$ contains about $Y$. This leads us to the definition of mutual information,

The *mutual information* between $X$ and $Y$ is defined as

$$I(X : Y) := H(Y) - H(Y|X) \tag{1.36}$$

It is not evident immediately from the way we defined it here but this expression is symmetric between $X$ and $Y$. Namely, using the chain rule for conditional entropy (recall Proposition 1.3) twice, we can write

$$\begin{aligned} I(X : Y) = H(Y) - H(Y|X) &= H(Y) + H(X) - H(XY) \\ &= H(X) - H(X|Y)\,. \end{aligned} \tag{1.37}$$

The mutual information is thus a symmetric measure of the correlation between the two random variables.

Using these various equivalent expressions it is then easy to derive some bounds on the mutual information. First, sub-additivity of the entropy directly implies that $I(X : Y) \geq 0$, so the mutual information is non-negative, and it vanishes if and only if the two random variables are independent (a consequence of Proposition 1.4). This is consistent with our intuitive notion of information — we cannot know less than nothing after all! We also cannot know more than everything, i.e. the mutual information can never exceed the entropy of any of its constituent parts.

Using the bounds on entropies established in the previous sections, show that

$$I(X : Y) \leq \log \min\{|\mathcal{X}|, |\mathcal{Y}|\}.$$

Give an example that saturates the bound.

**Example.** *Consider two binary random variables $X$ and $Y$ with joint pmf*

$$\begin{aligned} P_{XY}(0,0) = P_{XY}(1,1) &= \frac{1}{4}(1+r), \\ P_{XY}(0,1) = P_{XY}(1,0) &= \frac{1}{4}(1-r) \end{aligned} \tag{1.38}$$

*for $r \in [-1, 1]$. We can compute the mutual information between $X$ and $Y$ as follows:*

$$I(X : Y) = H(X) - H(X|Y) = 1 - h\left(\frac{1+r}{2}\right) \tag{1.39}$$

You might have heard of the correlation coefficient in statistics:

$$\rho = \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

Can you determine $\rho$ as a function of $r$?

*So this function takes its maximum value at $r = -1$ and $r = 1$ and drops to zero for $r = 0$.*

If we have three random variables $X$, $Y$ and $Z$ we can ask for the mutual information between $X$ and $Y$ conditioned on knowing $Z$, the *conditional mutual information*. It is defined as

$$I(X : Y|Z) := \sum_z P_Z(z) I(X : Y|Z = z)\,. \tag{1.40}$$

Various equivalent expressions can then be readily derived, e.g.,

$$I(X : Y|Z) = H(Y|Z) - H(Y|XZ) = H(X|Z) - H(X|YZ)\,. \tag{1.41}$$

Moreover, the *chain rule* for the mutual information states that

$$I(X : YZ) = I(X : Y) + I(X : Z|Y)\,, \tag{1.42}$$

Verify Eqs. (1.41) and (1.42) using the definition in Eq. (1.40).

which can be verified by a close inspection of the definition of both conditional and unconditional mutual information.

On the one hand, consider the case where $X - Z - Y$ form a Markov chain. In this case $P_{X|YZ} = P_{X|Z}$ and thus $H(X|YZ) = H(X|Z)$. As a consequence, the conditional mutual information $I(X : Y|Z)$ as written in (1.41) vanishes. On the other hand, if $I(X : Y|Z) = 0$ then we must have that $I(X : Y|Z = z) = 0$ for every $z$ with $P_Z(z) > 0$ according to our definition in (1.40). Hence, $P_{XY|Z=z} = P_{X|Z=z} P_{Y|Z=z}$ is independent according to Prop. 1.4. This means that $X - Z - Y$ must be a Markov chain. We summarise this in the following proposition:

**Proposition 1.5.** *The following two conditions are equivalent: a) $X - Z - Y$ form a Markov chain, and b) $I(X : Y|Z) = 0$.*

One of the most intriguing properties of the mutual information is the *data-processing inequality* (DPI) for mutual information. It states that the mutual information can never increase when we apply an operation that only acts on one of the parts. Intuitively this tells us that by manipulating one of the random variables without looking at the other we cannot increase the correlations between the pair.

We can formalise this using the notion of Markov chains.

Recall that for any random variables $X$ and $Y$ and any channel $W$ from $Y$ to $Z$, the resulting random variables form a Markov chain $X - Y - Z$.

**Proposition 1.6** (DPI for Mutual Information)**.** *Let $X - Y - Z$ be a Markov chain. Then, $I(X : Y) \geq I(X : Z)$.*

*Proof.* Since $I(X : Z|Y) = 0$, the chain rule for mutual information implies that $I(X : Y) = I(X : YZ)$. It thus remains to show that

$$I(X : Z) \leq I(X : YZ)\,. \tag{1.43}$$

Can you also show that

$$I(Y : Z) \geq I(X : Z)$$

under the same assumption?

But, since $I(X:Z) = H(X) - H(X|Z)$ and $I(X:YZ) = H(X) - H(X|YZ)$, the relation in Eq. (1.43) is equivalent to the condition $H(X|Z) \geq H(X|YZ)$, which is in turn ensured by the strong subadditivity of entropy. □

## 1.3 *Relative entropy*

The relative entropy, often referred to as Kullback-Leibler divergence, appears when we want to compare two different probability distributions. We define it here only for discrete random variables (or rather the respective pmfs). This can be generalised to general probability measures using the notion of Radon-Nikodyn derivatives, but this is outside our scope.

Let $P$ and $Q$ be two pmfs on an alphabet $\mathcal{X}$. The *relative entropy* of $P$ with regards to $Q$ is defined as

$$D(P\|Q) := \sum_{\substack{x\in\mathcal{X}\\ P(x)>0}} P(x)\log\frac{P(x)}{Q(x)}\,. \tag{1.44}$$

if $P(x) > 0 \implies Q(x) > 0$ for all $x \in \mathcal{X}$, and as $D(P\|Q) = +\infty$ otherwise.

In the following, instead of restricting the sum, we will use the convention that $0\log\frac{0}{0} = 0$.

We can alternatively see the relative entropy as the expectation value of the *log-likelihood ratio*, namely we can write

$$D(P\|Q) = \mathbb{E}\left[Z(X)\right], \qquad \text{where} \qquad Z(X) = \log\frac{P(X)}{Q(X)} \tag{1.45}$$

and $X$ is distributed according to $P$. The random variable $Z(X)$ is called the log-likelihood ratio. It takes on the role of the surprisal in the definition of entropy. We will explore this random variable and its distribution much more when we discuss the information spectrum method and hypothesis testing later on.

**Example.** Consider two Bernoulli distributions $P = \mathrm{Bern}(p)$ and $Q = \mathrm{Bern}(q)$. Then the relative entropy evaluates to

$$D(P\|Q) = p\log\frac{p}{q} + (1-p)\frac{1-p}{1-q}.$$

What values does the random variable $Z$ take in the above example, and with what probability?

Just by manipulating the definition, we are able to show the following equivalences.

**Proposition 1.7.** *Let $X$ and $Y$ be random variables on alphabets $\mathcal{X}$ and $\mathcal{Y}$. Moreover, let $U$ be a uniform random variable on $\mathcal{X}$. Then the follow-*

*ing relations are true:*

$$H(X) = \log|\mathcal{X}| - D(P_X\|U_X) \tag{1.46}$$
$$H(X|Y) = \log|\mathcal{X}| - D(P_{XY}\|U_X \times P_Y) \tag{1.47}$$
$$I(X:Y) = D(P_{XY}\|P_X \times P_Y)\,. \tag{1.48}$$

You will prove these equivalences in the homework. They turn out to be very useful because they essentially tell us that once we established properties of the relative entropy this has immediate consequences also for the derived quantities.

Can you express the conditional mutual information $I(X{:}Z|Y)$ in terms of the relative entropy?

We will need two important properties of the relative entropy. The first proposition establishes that the relative entropy is always positive.

**Proposition 1.8.** *For any two pmfs $P$ and $Q$, we have $D(P\|Q) \geq 0$ with equality if and only if $P = Q$.*

*Proof.* We can assume without loss of generality that the quantity is finite, as otherwise the statement is trivially true. We first note that $x \mapsto -\log x$ is strictly convex. Hence,

$$D(P\|Q) = \sum_{x:P(x)>0} P(x)\log\frac{P(x)}{Q(x)} \tag{1.49}$$
$$= \sum_{x:P(x)>0} P(x)\left(-\log\frac{Q(x)}{P(x)}\right) \tag{1.50}$$
$$\geq -\log\left(\sum_{x:P(x)>0} P(x)\frac{Q(x)}{P(x)}\right) \tag{1.51}$$
$$= -\log\left(\sum_{x:P(x)>0} Q(x)\right) \tag{1.52}$$
$$\geq -\log\left(\sum_{x} Q(x)\right) = 0\,. \tag{1.53}$$

Equality in the second inequality only holds if $P$ and $Q$ have the same support. Moreover, equality in the first inequality holds if $\frac{Q(x)}{P(x)}$ is independent of $x$ for any $x$ in the support of $P$. These two statements are both true only if $P(x) = Q(x)$ for all $x \in \mathcal{X}$, and thus $P = Q$. □

An immediate corollary of Propositions 1.7 and 1.8 is that $I(X:Y)$ is positive and zero if and only if $X$ and $Y$ are independent.

Can you prove this?

Finally, there is one property of the relative entropy that, in conjunction with Prop. **??**, implies most other properties of entropy, conditional entropy and mutual information. It states that applying

a noisy operation, i.e. a stochastic map or channel, on both arguments of the relative entropy will never increase it. Together with the positivity of relative entropy this justifies that we think of it as a measure of similarity or distinguishability. If the relative entropy is small the two pmfs are similar and hard to distinguish by observing the outcomes of a random experiment. Observing the outcomes after further noise has been applied should make distinguishing them even harder, and that is exactly what the *data-processing inequality* for relative entropy tells us.

**Proposition 1.9** (DPI for Relative Entropy). *Let $P_X$ and $Q_X$ be two pmfs on an alphabet $\mathcal{X}$ (the input distributions), and let $W_{Y|X}$ be a channel (a conditional pmf). Define the marginals (the output distributiions)*

$$P_Y(y) = \sum_{x\in\mathcal{X}} W_{Y|X}(y|x)P_X(x) \qquad \text{and}$$
$$Q_Y(y) = \sum_{x\in\mathcal{X}} W_{Y|X}(y|x)Q_X(x)\,. \tag{1.54}$$

*Then, the* data-processing inequality *(DPI) states that*

$$D(P_X\|Q_X) \geq D(P_Y\|Q_Y)\,. \tag{1.55}$$

*Proof.* Consider the joint distributions $P_{XY}(x,y) = W_{Y|X}(y|x)P_X(x)$ and $Q_{XY}(x,y) = W_{Y|X}(y|x)Q_X(x)$, using the usual shorthand notation for conditional and marginal distributions. We first show that

$$D(P_{XY}\|Q_{XY}) - D(P_Y\|Q_Y) = \left(\sum_{x,y} p_{xy}\log\frac{p_{xy}}{q_{xy}}\right) - \left(\sum_y p_y\log\frac{p_y}{q_y}\right) \tag{1.56}$$
$$= \sum_{x,y} p_{xy}\left(\log\frac{p_{xy}}{q_{xy}} - \log\frac{p_y}{q_y}\right) \tag{1.57}$$
$$= \sum_y p_y \sum_x p_{x|y}\log\frac{p_{x|y}}{q_{x|y}} \tag{1.58}$$
$$= \sum_y p_y\, D(P_{X|Y=y}\|Q_{X|Y=y}) \geq 0\,, \tag{1.59}$$

where we have used the positivity of relative entropy in the last step. Similarly, we have

$$D(P_{XY}\|Q_{XY}) - D(P_X\|Q_X) = \sum_x p_x\, D(P_{Y|X=x}\|Q_{Y|X=x}) = 0 \tag{1.60}$$

since $Q_{Y|X} = P_{Y|X} = W_{Y|X}$ by construction of the joint distribution. Combining Eqs. (1.56)–(1.59) and (1.60) yields the desired inequality. □

It turns out that all the properties of entropy, conditional entropy and mutual information we discussed previously can be derived

form the DPI. As an example we give here a strengthening of the strong sub-additivity, which we call the data-processing inequality for conditional entropy. It intuitively states that any processing of the side information can at most increase the conditional entropy.

**Corollary 1.10** (DPI for conditional entropy)**.** *Let $P_{XY}$ be a joint pmf and $W_{Z|Y}$ a channel. Define*

$$P_{XZ}(x,z) = \sum_y P_{XY}(x,y) W_{Z|Y}(z|y)\,. \tag{1.61}$$

*Then, we have $H(X|Y) \leq H(X|Z)$.*

*Proof.* Let us express the inequality in terms of relative entropies using Proposition 1.7. This reads

$$\log|\mathcal{X}| - D(P_{XY}\|U_X \times P_Y) \leq \log|\mathcal{X}| - D(P_{XZ}\|U_X \times P_Z)\,. \tag{1.62}$$

or simply $D(P_{XY}\|U_X \times P_Y) \geq D(P_{XZ}\|U_X \times P_Z)$. But this is imply the DPI applied to the channel $W_{Z|Y}$ that happens to leave $X$ untouched. □