Learner's Space 2025
Information Theory and Coding
Week 2, Handout 1

Suketu Patni

June 18, 2025

Abstract

In this document, we will see some jugglery that we can do when considering the entropies of multiple random variables together.

1 Joint and Conditional Entropy

Let X and Y be two discrete random variables with finite image $\mathcal{X}$ and $\mathcal{Y}$ respectively.

Definition 1.1. *The **joint entropy** of X and Y is denoted as $H(X,Y)$ and is defined as*

$$H(X,Y) = -\sum_{(x,y)\in(\mathcal{X}\times\mathcal{Y})} p(x,y)\log_2(p(x,y))$$

where $p(x,y) = P(X=x, Y=y)$ i.e. the probability that $X=x$ and $Y=y$.

This definition is as one would expect. We might as well define a new random variable Z with $P(Z=(x,y)) = P(X=x, Y=y)$ and say that the joint entropy of X and Y is the entropy of Z.

Definition 1.2. *The **conditional entropy** of Y given X is denoted as $H(Y|X)$ and is defined as*

$$H(Y|X) = \sum_{x\,\in\,\mathrm{supp}\,X} p(x)H(Y|X=x)$$

One should think of "$Y|X=x$" as a random variable of its own right, where the probabilities $p(y) = P(Y=y)$ are replaced with $p(y|x) = P(Y=y|X=x)$ i.e. $\frac{P(X=x,Y=y)}{P(X=x)}$. We

can thus write down another expression for $H(Y|X)$, namely

$$\begin{aligned} H(Y|X) &= \sum_{x \in \operatorname{supp} X} p(x) H(Y|X=x) \\ &= \sum_{x \in \operatorname{supp} X} p(x) \left(- \sum_{y \in \mathcal{Y}} p(y|x) \log_2(p(y|x)) \right) \\ &= - \sum_{(x,y) \in ((\operatorname{supp} X) \times \mathcal{Y})} p(x) p(y|x) \log_2(p(y|x)) \\ &= - \sum_{(x,y) \in ((\operatorname{supp} X) \times \mathcal{Y})} p(x,y) \log_2(p(y|x)) \end{aligned}$$

Theorem 1.1. *For X and Y as defined in the beginning of the section, $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$.*

This is known as the chain rule for entropy, very similar to the way we have a chain rule for probability.

Proof. Straightforward algebra.

$$\begin{aligned} H(X,Y) &= - \sum_{(x,y) \in (\mathcal{X} \times \mathcal{Y})} p(x,y) \log_2(p(x,y)) \\ &= - \sum_{(x,y) \in ((\operatorname{supp} X) \times \mathcal{Y})} p(x,y) (\log_2(p(x)) + \log_2(p(y|x)) \\ &= - \sum_{(x,y) \in ((\operatorname{supp} X) \times \mathcal{Y})} p(x,y) \log_2(p(x)) - \sum_{(x,y) \in ((\operatorname{supp} X) \times \mathcal{Y})} p(x,y) \log_2(p(y|x)) \\ &= - \sum_{x \in \operatorname{supp} X} \sum_{y \in \mathcal{Y}} p(x,y) \log_2(p(x)) + H(Y|X) \\ &= - \sum_{x \in \operatorname{supp} X} \log_2(p(x)) \sum_{y \in \mathcal{Y}} p(x,y) + H(Y|X) \\ &= - \sum_{x \in \operatorname{supp} X} (\log_2(p(x))) p(x) + H(Y|X) \\ &= H(X) + H(Y|X) \end{aligned}$$

□

It can identically be shown that $H(X,Y) = H(Y) + H(X|Y)$. The qualitative meaning of this rule is that "the information supplied by both X and Y is the sum of the information supplied by X and that supplied by Y given X". Further, using induction, we can show that for discrete random variables X_i with finite images $\mathcal{X}_i$ (for $i \in \{1, 2, \ldots, n\}$,

$$H(X_1, X_2, \ldots, X_n) = H(X_1) + \sum_{k=2}^{n} H(X_k | X_{k-1}, \ldots, X_1)$$

Exercise 1.1. *For X and Y as defined in the beginning of the section, if further X and Y are independent, show that $H(Y|X) = H(Y)$ and similarly $H(X|Y) = H(X)$ (easily done using straightforward algebra).*

So for independent discrete random variables X and Y with finite images, we get $H(X, Y) = H(X) + H(Y)$.

2 KL Divergence

For a finite subset of real numbers $\mathcal{X}$, let p and q be two probability mass functions defined on it. Just to clarify, this means that $p : \mathcal{X} \to [0, 1]$ and $q : \mathcal{X} \to [0, 1]$ and that there exist random variables X and Y with image $\mathcal{X}$ such that $p(x) = P(X = x)$ and $q(y) = P(Y = y)$.

Definition 2.1. *The **Kullback-Leibler divergence** between p and q, is defined as*

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log_2 \left(\frac{p(x)}{q(x)} \right)$$

A few points to note on this:-

1. Notice how there isn't a restriction on $p(x)$ or $q(x)$ being nonzero in the summation. That is because of the following "convention": if for some $x \in \mathcal{X}$, $p(x) = 0$, then the corresponding summand is taken to be 0, and if $\exists\, x \in \mathcal{X}$ for which $p(x) \neq 0$ and $q(x) = 0$, then $D(p||q)$ is said to be "equal" to ∞.

2. It is sometimes also called the "relative entropy" between the distributions p and q.

3. It is also sometimes called the "KL distance", which is misleading because it is not a metric on the set of distributions on $\mathcal{X}$. One reason is obviously that $D(p||q) \neq D(q||p)$ in general.

4. $D(p||q)$ measures the "inefficiency" of assuming the distribution to be $q(x)$ when the true distribution is $p(x)$. This gives some sense to the convention stated in the first point; in a loose sense, if there is an outcome for which we assign 0 probability, but it has a nonzero probability in reality, we are highly mistaken.

Exercise 2.1. *Add further salt to the wound and show that $D(\cdot||\cdot)$ also does not satisfy the triangle inequality, which is a requisite for any metric.*

However, as it turns out, the KL divergence does satisfy one of the conditions of a metric, namely, positivity.

Theorem 2.1. *For two probability mass functions p and q on a finite $\mathcal{X} \subset \mathbb{R}$, $D(p||q) \geq 0$ with equality iff $p = q$.*

This key result (which we'll use later also) is known as **Gibbs' inequality**.

Proof. In case $\exists\, x \in \mathcal{X}$ for which $p(x) \neq 0$ and $q(x) = 0$, then $D(p||q) = \infty > 0$ and we are done. So assume that $q(x) = 0 \Rightarrow p(x) = 0$. Consider now the set $S = \{x \in \mathcal{X} : p(x) \neq 0\}$ which by assumption is equal to the set $\{x \in \mathcal{X} : p(x) \neq 0, q(x) \neq 0\}$. Define a random variable X with support S such that

$$P\left(X = \frac{q(x)}{p(x)}\right) = \sum_{y \in S, \frac{q(x)}{p(x)} = \frac{q(y)}{p(y)}} p(y)$$

and apply Jensen's inequality on X with the function $f(x) = -\log_2(x)$, which is convex on $\mathbb{R}^+$. Be careful and do not get confused by the apparent self-referential-ness. We get

$$\begin{aligned}
\mathbb{E}[-\log_2(X)] &\geq -\log_2(\mathbb{E}[X]) \\
\Rightarrow -\sum_{x \in S} p(x) \log_2\left(\frac{q(x)}{p(x)}\right) &\geq -\log_2\left(\sum_{x \in S} p(x)\frac{q(x)}{p(x)}\right) = \log_2\left(\sum_{x \in S} q(x)\right) \\
&\geq \log_2\left(\sum_{x \in \mathcal{X}} q(x)\right) = \log_2(1) = 0 \\
\Rightarrow \sum_{x \in S} p(x) \log_2\left(\frac{p(x)}{q(x)}\right) \geq 0 &\Rightarrow \sum_{x \in \mathcal{X}} p(x) \log_2\left(\frac{p(x)}{q(x)}\right) \geq 0
\end{aligned}$$

since the set $\mathcal{X} \setminus S$ contributes nothing to the KL divergence. The last statement is what we wanted to show. □

Exercise 2.2. *Prove the condition for equality for Gibbs' inequality, using the condition for equality in Jensen's inequality.*

3 Mutual Information

Definition 3.1. *The **mutual information** between discrete random variables X and Y with finite images $\mathcal{X}$ and $\mathcal{Y}$, written as $I(X;Y)$ is the KL divergence between the joint distribution $p(x,y)$ and the product distribution $p(x)p(y)$.*

We can talk about this since both the probability mass functions are defined on the set $\mathcal{X} \times \mathcal{Y}$ (it is easy to verify that $p(x)p(y)$ is a valid pmf).

Theorem 3.1. $I(X;Y) = H(X) - H(X|Y)$

Proof. Again, straightforward algebra. It is left as an exercise to the reader. □

Some notes on the above definition and theorem:-

1. Qualitatively, it is the reduction in the uncertainty of X that Y provides.

2. See that $H(X,X) = H(X)$ so $H(X|X) = 0$. This means $I(X;X) = H(X)$. This is why entropy is also sometimes referred to as **self-information**.

3. $I(X;Y) = I(Y;X)$. So (on average), X gives as much information about Y as Y does about X.

4. $I(X;Y) = H(X) - H(X|Y) = H(X) - (H(X,Y) - H(Y)) = H(X) + H(Y) - H(X,Y)$.

5. $I(X;Y) = 0 \Rightarrow$ the joint and the product distributions are the same. This happens iff X and Y are independent. This makes sense since if they are, we expect Y to give no information about X.

6. $I(X;Y) \geq 0 \Rightarrow H(X|Y) \leq H(X)$ since it is a KL divergence. So, on average, knowing another random variable Y can only reduce the uncertainty in X.

7. We may define the **conditional mutual information** between X and Y given Z (a discrete random variable with finite image $\mathcal{Z}$) similarly as the expected value over Z of the KL divergence between the joint and product distributions of $X|Z = z$ and $Y|Z = z$. It is the information Y gives about X, given Z.

Exercise 3.1. *Prove that, according to the above definition,* $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$.

Theorem 3.2.

$$I(X_1, X_2, \ldots, X_n; Y) = I(X_1; Y) + \sum_{i=2}^{n} I(X_i; Y | X_{i-1}, \ldots, X_1)$$

The above statement is called the chain rule for mutual information.

Proof.

$$\begin{aligned}
I(X_1, X_2, \ldots, X_n; Y) &= H(X_1, X_2, \ldots, X_n) - H(X_1, X_2, \ldots, X_n | Y) \\
&= H(X_1) + \sum_{i=2}^{n} H(X_i | X_{i-1}, \ldots, X_1) - \left(H(X_1|Y) + \sum_{i=2}^{n} H(X_i | X_{i-1}, \ldots, X_1, Y) \right) \\
&= I(X_1; Y) + \sum_{i=2}^{n} (H(X_i | X_{i-1}, \ldots, X_1) - H(X_i | X_{i-1}, \ldots, X_1, Y)) \\
&= I(X_1; Y) + \sum_{i=2}^{n} I(X_i; Y | X_{i-1}, \ldots, X_1)
\end{aligned}$$

□

4 Data Processing Inequality

To understand this inequality, we will need to know what a discrete-time Markov chain is first.

Definition 4.1. *A sequence of discrete random variables $\{X_i\}_{i\in\mathbb{N}}$, each with the same countable or finite image S which satisfies*

$$P(X_{n+1} = x_{n+1} \mid X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1) = P(X_{n+1} = x_{n+1} \mid X_n = x_n)$$

for all $n \geq 1$ and all $x_1, \ldots, x_{n+1} \in S$ such that $P(X_1 = x_1, \ldots, X_n = x_n) > 0$ is called a ***discrete-time Markov chain****.*

Additionally, we will enforce the finiteness of the image set S. The property that "the future is independent of the past given the present" is called the Markov property. Loosely speaking, the next "state" does not depend on the past "states" if we are given the present state.

Theorem 4.1. *For a Markov chain $\{X_i\}_{i\in\mathbb{N}}$, $I(X_j, X_k) \geq I(X_j, X_{k+1})\ \forall\, j, k \in \mathbb{N}, j \leq k$.*

Proof. Rename $X_j = X$, $X_k = Y$, $X_{k+1} = Z$. Consider

$$\begin{aligned} I(X; Y, Z) &= I(Y, Z; X) \\ &= I(Y; X) + I(Z; X|Y) && \text{(chain rule for mutual information)} \\ &= I(X; Y) + I(Z; X|Y) \end{aligned}$$

Similarly,

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

So,

$$\begin{aligned} &I(X; Y) + I(Z; X|Y) = I(X; Z) + I(Y; X|Z) \\ &\Rightarrow I(X; Y) - I(X; Z) = I(Y; X|Z) - I(Z; X|Y) \\ &\qquad\qquad \geq -I(Z; X|Y) && \text{(mutual information is always non-negative)} \end{aligned}$$

But since Z and X are conditionally independent given Y, $I(Z; X|Y) = 0$, and we are done. □

But why is this called the "data processing inequality"? It is because it states that no **processing** of Y can yield more information about X than Y itself. No manipulation of the data can ever yield new information. This is a deep result, and one should be astonished that in such a short time, we have been able to prove this in a mathematical sense using the tools of information theory.