

# Electrical Engineering 229A Lecture 3 Notes

Daniel Raban

September 2, 2021

## 1 Entropy Over Countable Alphabets and Features of Conditional Entropy

### 1.1 Entropy of distributions over countable sets

Let's adjust our definitions to allow for distributions over countable sets. Let  $X$  be a random variable taking values in  $\mathcal{X}$ , a finite or countably infinite set, and let  $(p(x), x \in \mathcal{X})$  be its probability distribution. Its **entropy** is

$$H(X) = H((p(x), x \in \mathcal{X})) = - \sum_x p(x) \log p(x).$$

This is well-defined, even if  $\mathcal{X}$  is countably infinite, because all the terms have the same sign.

**Remark 1.1.** In general, to define  $\sum_{x \in \mathcal{X}} a(x)$ , where  $\mathcal{X}$  is countably infinite, define it to be  $(\sum_{x \in \mathcal{X}} a^+(x)) - (\sum_{x \in \mathcal{X}} a^-(x))$ , where  $a^+(x) := \max(a(x), 0)$  and  $a^-(x) := \max(-a(x), 0)$ . This definition makes sense when at least one of  $\sum_{x \in \mathcal{X}} a^+(x)$ ,  $\sum_{x \in \mathcal{X}} a^-(x)$  is finite.

To avoid subtracting infinities when dealing with entropies over countable sets, proceed as follows: Given a pair of random variables  $X, Y$  taking values taking values in (finite or countably infinite)  $\mathcal{X}, \mathcal{Y}$ , respectively, for each  $y \in \mathcal{Y}$ , define  $H(X | Y = y)$  to be the entropy of the conditional distribution of  $X$  given  $Y = y$ :

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

We can alternatively express

$$H(X) = \mathbb{E} \left[ \log \frac{1}{p(X)} \right], \quad \mathbb{E} \left[ \log \frac{1}{p(X | Y)} \mid Y = y \right],$$

as before.

Define the **conditional entropy** of  $X$  given  $Y$  to be  $\sum_y p(y)H(X | Y = y)$ , denoted  $H(X | Y)$ . So

$$H(X | Y) = \mathbb{E} \left[ \log \frac{1}{p(X | Y)} \right].$$

Now  $H(X, Y) = H(Y) + H(X | Y)$  becomes a theorem, called the chain rule for entropy.

**Theorem 1.1** (Chain rule).

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

*Proof.*

$$\mathbb{E} \left[ \log \frac{1}{p(X, Y)} \right] = \mathbb{E} \left[ \log \frac{1}{p(Y)} \right] + \mathbb{E} \left[ \log \frac{1}{p(X | Y)} \right]. \quad \square$$

We define  $D(p || q)$  for  $(p(x), x \in \mathcal{X})$ ,  $(q(x), x \in \mathcal{X})$  as

$$D(p || q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

To see that this is well-defined, observe that

$$= \sum_x q(x) \frac{p(x)}{q(x)} \log \frac{p(x)}{q(x)}.$$

Then this is well-defined because the function  $u \mapsto u \log u$  defined on  $\mathbb{R}^+$  is bounded below.

Then, we can define  $I(X; Y) := D(p(x, y) || p(x)p(y))$ , and our previous definition for mutual information becomes a theorem:

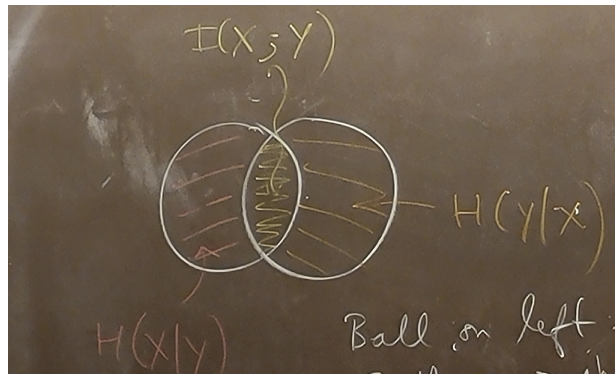
**Theorem 1.2.**

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

*Proof.*

$$\mathbb{E} \left[ \log \frac{1}{p(X)} \right] = \mathbb{E} \left[ \log \frac{p(X, Y)}{p(X)p(Y)} \right] + \mathbb{E} \left[ \log \frac{1}{p(X | Y)} \right]. \quad \square$$

These “theorems” or  $(X, Y)$  can be schematically visualized via a Venn diagram.



## 1.2 Relationship between mutual information and independence

It is important to recognize that the condition for  $I(X; Y) = 0$  is  $p(x, y) = p(x)p(y)$  for all  $x, y$ , i.e.  $X, Y$  are independent (denoted  $X \amalg Y$ ). Since  $I(X; Y) = H(X) + H(Y) - H(X, Y)$  (inclusion-exclusion),

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

## 1.3 General form of the chain rule

If we apply the chain rule twice, we get

$$\begin{aligned} H(X_1, X_2, X_3) &= H(X_1 | X_2, X_3) + H(X_2, X_3) \\ &= H(X_1 | X_2, X_3) + H(X_2 | X_3) + H(X_3). \end{aligned}$$

Similarly, using the notation  $X_1^n$  to denote  $(X_1, \dots, X_n)$ , we get the general chain rule:

**Theorem 1.3** (Chain rule, general form).

$$H(X_1, \dots, X_n) = H(X_1) + H(X_2 | X_1) + H(X_3 | X_1, X_2) + \dots + H(X_n | X_1^{n-1}).$$

**Example 1.1.** Consider an urn<sup>1</sup> with 3 balls, two white and 1 red. Pull out all 3 balls in a random order. Let  $X_1$  be the color of the first ball, let  $X_2$  be the color of the second ball, and let  $X_3$  be the color of the third ball. Then

$$H(X_1) = H(X_2) = H(X_3) = \frac{1}{3} \log 3 + \frac{2}{3} \log \frac{3}{2} = \log 3 - \frac{2}{3}.$$

We can also calculate the conditional entropies:

$$\begin{aligned} H(X_2 | X_1) &= \mathbb{P}(X_1 = \text{red})H(X_2 | X_1 = \text{red}) + \mathbb{P}(X_1 = \text{white})H(X_2 | X_1 = \text{white}) \\ &= \frac{2}{3} \log 2 \\ &= \frac{2}{3}. \end{aligned}$$

On the other hand,  $H(X_3 | X_1, X_2) = 0$  because  $X_3$  is determined by  $X_1, X_2$ . So the chain rule gives

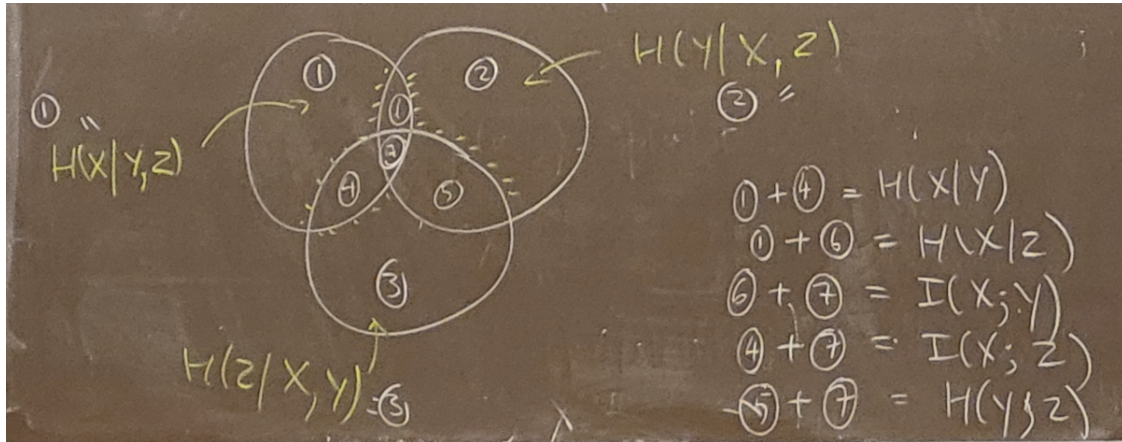
$$\begin{aligned} H(X_1, X_2, X_3) &= H(X_1) + H(X_2 | X_1) + H(X_3 | X_1, X_2) \\ &= \log 3 - \frac{2}{3} + \frac{2}{3} + 0 \\ &= \log 3. \end{aligned}$$

---

<sup>1</sup>No one in the 21st century has ever seen an urn.

## 1.4 Problems with intuiting mutual information

Here is the Venn diagram for  $(X_1, X_2, X_3)$ :



What does region 6 represent? This could be  $I(X;Y | Z)$ , the conditional relative entropy between the joint distribution  $(X, Y)$ , conditioned on  $Z$  and the product distribution with the corresponding marginals, conditioned on  $Z$ . That is, region 6 is

$$H(X | Z) - H(X | Y, Z).$$

What does region 7 represent? This region is

$$I(X;Y) - I(X;Y | Z).$$

Here is a big problem, not for the math but for any hope of intuition: This can be *negative*. In particular, this says that in the presence of  $Z$ ,  $Y$  can tell you more about  $X$  than it does alone.

**Example 1.2.** Let  $X \amalg Y$ , with  $X \in \{1, -1\}$ ,  $Y \in \{1, -1\}$ ,  $\mathbb{P}(X = 1) = 1/2$ , and  $\mathbb{P}(Y = 1) = 1/2$ . Let  $Z = XY$  so  $Z \in \{1, -1\}$  with  $\mathbb{P}(Z = 1) = 1/2$ . Then  $Y \amalg Z$  and  $X \amalg Z$ , but  $X, Y, Z$  are not mutually independent. Since  $X \amalg Y$ , we have  $I(X;Y) = 0$ . However,

$$\begin{aligned} I(X;Y | Z) &= \mathbb{P}(Z = 1)I(X;Y | Z = 1) + \mathbb{P}(Z = -1)I(X;Y | Z = -1) \\ &= \mathbb{P}(Z = 1)(H(X | Z = 1) - H(X | Y, Z = 1)) \\ &\quad + \mathbb{P}(Z = -1)(H(X | Z = -1) - H(X | Y, Z = -1)) \end{aligned}$$

Since  $X \amalg Z$ ,  $H(X | Z = 1) = H(X | Z = -1) = H(X) = \log 2 = 1$ . Also,  $H(X | Y, Z = 1) = 0$  because  $X = Y$  when  $Z = 1$  and  $H(X | Y, Z = -1) = 0$  because  $X = -Y$  when  $Z = -1$ . So

$$= \frac{1}{2}(1 - 0) + \frac{1}{2}(1 - 0)$$

$$= 1.$$

This is strictly bigger than  $I(X; Y)$ .

Let's define  $I(X; Y | Z)$  in a way that works for a countably infinite alphabet. We first define, given  $p(x, y, z)$ ,

$$\sum_z p(z) D(p(x | z) || p(y | z)),$$

denoted  $D(p(x | z) || p(y | z) | p(z))$  to be the conditional relative entropy of  $p(x, z)$  with respect to  $p(y, z)$  given  $z$ . Then  $D(p(x, y | z) || p(x | z)p(y | z) | p(z))$  would then be  $I(X; Y | Z)$ . That is,

$$\begin{aligned} I(X; Y | Z) &:= \sum_z p(z) \sum_{x,y} p(x, y | z) \log \frac{p(x, y | z)}{p(x | z)p(y | z)} \\ &= \mathbb{E} \left[ \log \frac{p(X, Y | Z)}{p(X | Z)p(Y | Z)} \right] \\ &= H(X | Z) + H(Y | Z) - H(X, Y | Z). \end{aligned}$$

Then the chain rule gives

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

## 1.5 The chain rule for relative entropy

**Theorem 1.4** (Chain rule for relative entropy).

$$D(p(x, y) || q(x, y)) = D(p(x) || q(x)) + D(p(y | x) || q(y | x) | p(x)).$$

*Proof.*

$$\begin{aligned} D(p(x, y) || q(x, y)) &= \sum_{x,y} p(x, y) \log \frac{p(x, y)}{q(x, y)} \\ &= \mathbb{E}_p \left[ \log \frac{p(X, Y)}{q(X, Y)} \right] \\ &= \mathbb{E}_p \left[ \log \frac{p(X)}{q(X)} \right] + \mathbb{E}_p \left[ \log \frac{p(Y | X)}{q(Y | X)} \right] \\ &= D(p(x) || q(x)) + D(p(y | x) || q(y | x) | p(x)). \quad \square \end{aligned}$$

Similarly, there is a chain rule for mutual information

**Theorem 1.5** (Chain rule for mutual information).

$$I(X; Y_1, \dots, Y_n) = I(X; Y_1) + I(X; Y_2 | Y_1) + \dots + I(X; Y_n | Y_1^{n-1}).$$