8

Dependent Random Variables

In the last three chapters on data compression we concentrated on random vectors $\mathbf{x}$ coming from an extremely simple probability distribution, namely the separable distribution in which each component x_n is independent of the others.

In this chapter, we consider *joint ensembles* in which the random variables are dependent. This material has two motivations. First, data from the real world have interesting correlations, so to do data compression well, we need to know how to work with models that include dependences. Second, a noisy channel with input x and output y defines a joint ensemble in which x and y are dependent – if they were independent, it would be impossible to communicate over the channel – so communication over noisy channels (the topic of chapters 9–11) is described in terms of the entropy of joint ensembles.

8.1 More about entropy

This section gives definitions and exercises to do with entropy, carrying on from section 2.4.

The joint entropy of X, Y is:

$$H(X,Y) = \sum_{xy \in \mathcal{A}_X \mathcal{A}_Y} P(x,y) \log \frac{1}{P(x,y)}. \tag{8.1}$$

Entropy is additive for independent random variables:

$$H(X,Y) = H(X) + H(Y) \text{ iff } P(x,y) = P(x)P(y). \tag{8.2}$$

The conditional entropy of X **given** $y = b_k$ is the entropy of the probability distribution $P(x \mid y = b_k)$.

$$H(X \mid y = b_k) \equiv \sum_{x \in \mathcal{A}_X} P(x \mid y = b_k) \log \frac{1}{P(x \mid y = b_k)}. \tag{8.3}$$

The conditional entropy of X **given** Y is the average, over y, of the conditional entropy of X given y.

$$\begin{aligned} H(X \mid Y) &\equiv \sum_{y \in \mathcal{A}_Y} P(y) \left[\sum_{x \in \mathcal{A}_X} P(x \mid y) \log \frac{1}{P(x \mid y)} \right] \\ &= \sum_{xy \in \mathcal{A}_X \mathcal{A}_Y} P(x,y) \log \frac{1}{P(x \mid y)}. \end{aligned} \tag{8.4}$$

This measures the average uncertainty that remains about x when y is known.

The marginal entropy of X is another name for the entropy of X, $H(X)$, used to contrast it with the conditional entropies listed above.

Chain rule for information content. From the product rule for probabilities, equation (2.6), we obtain:

$$\log \frac{1}{P(x,y)} \;=\; \log \frac{1}{P(x)} + \log \frac{1}{P(y\,|\,x)} \tag{8.5}$$

so

$$h(x,y) = h(x) + h(y\,|\,x). \tag{8.6}$$

In words, this says that the information content of x and y is the information content of x plus the information content of y given x.

Chain rule for entropy. The joint entropy, conditional entropy and marginal entropy are related by:

$$H(X,Y) = H(X) + H(Y\,|\,X) = H(Y) + H(X\,|\,Y). \tag{8.7}$$

In words, this says that the uncertainty of X and Y is the uncertainty of X plus the uncertainty of Y given X.

The mutual information between X **and** Y is

$$I(X;Y) \;\equiv\; H(X) - H(X\,|\,Y), \tag{8.8}$$

and satisfies $I(X;Y) = I(Y;X)$, and $I(X;Y) \geq 0$. It measures the average reduction in uncertainty about x that results from learning the value of y; **or vice versa**, the average amount of information that x conveys about y.

The conditional mutual information between X **and** Y **given** $z = c_k$ is the mutual information between the random variables X and Y in the joint ensemble $P(x, y\,|\,z = c_k)$,

$$I(X;Y\,|\,z=c_k) = H(X\,|\,z=c_k) - H(X\,|\,Y, z=c_k). \tag{8.9}$$

The conditional mutual information between X **and** Y **given** Z is the average over z of the above conditional mutual information.

$$I(X;Y\,|\,Z) = H(X\,|\,Z) - H(X\,|\,Y,Z). \tag{8.10}$$

No other 'three-term entropies' will be defined. For example, expressions such as $I(X;Y;Z)$ and $I(X\,|\,Y;Z)$ are illegal. But you may put conjunctions of arbitrary numbers of variables in each of the three spots in the expression $I(X;Y\,|\,Z)$ – for example, $I(A,B;C,D\,|\,E,F)$ is fine: it measures how much information on average c and d convey about a and b, assuming e and f are known.

Figure 8.1 shows how the total entropy $H(X,Y)$ of a joint ensemble can be broken down. **This figure is important.** ∗

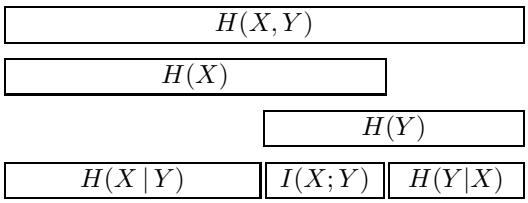


Figure 8.1. The relationship between joint information, marginal entropy, conditional entropy and mutual entropy.

8.2 Exercises

▷ Exercise 8.1.[1] Consider three independent random variables u, v, w with entropies H_u, H_v, H_w. Let $X \equiv (U, V)$ and $Y \equiv (V, W)$. What is $H(X,Y)$? What is $H(X \mid Y)$? What is $I(X;Y)$?

▷ Exercise 8.2.[3, p.142] Referring to the definitions of conditional entropy (8.3–8.4), confirm (with an example) that it is possible for $H(X \mid y=b_k)$ to exceed $H(X)$, but that the average, $H(X \mid Y)$, is less than $H(X)$. So data are helpful – they do not increase uncertainty, on average.

▷ Exercise 8.3.[2, p.143] Prove the chain rule for entropy, equation (8.7). $[H(X,Y) = H(X) + H(Y \mid X)]$.

Exercise 8.4.[2, p.143] Prove that the mutual information $I(X;Y) \equiv H(X) - H(X \mid Y)$ satisfies $I(X;Y) = I(Y;X)$ and $I(X;Y) \geq 0$.

[Hint: see exercise 2.26 (p.37) and note that

$$I(X;Y) = D_{\mathrm{KL}}(P(x,y) || P(x)P(y)).] \tag{8.11}$$

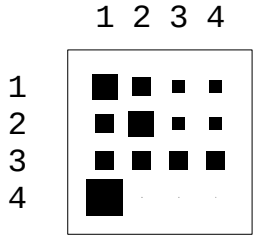
Exercise 8.5.[4] The 'entropy distance' between two random variables can be defined to be the difference between their joint entropy and their mutual information:

$$D_H(X,Y) \equiv H(X,Y) - I(X;Y). \tag{8.12}$$

Prove that the entropy distance satisfies the axioms for a distance – $D_H(X,Y) \geq 0$, $D_H(X,X) = 0$, $D_H(X,Y) = D_H(Y,X)$, and $D_H(X,Z) \leq D_H(X,Y) + D_H(Y,Z)$. [Incidentally, we are unlikely to see $D_H(X,Y)$ again but it is a good function on which to practise inequality-proving.]

Exercise 8.6.[2, p.147] A joint ensemble XY has the following joint distribution.

$P(x,y)$		x = 1	2	3	4
	1	1/8	1/16	1/32	1/32
y	2	1/16	1/8	1/32	1/32
	3	1/16	1/16	1/16	1/16
	4	1/4	0	0	0



What is the joint entropy $H(X,Y)$? What are the marginal entropies $H(X)$ and $H(Y)$? For each value of y, what is the conditional entropy $H(X \mid y)$? What is the conditional entropy $H(X \mid Y)$? What is the conditional entropy of Y given X? What is the mutual information between X and Y?

Exercise 8.7.[2, p.143] Consider the ensemble XYZ in which $\mathcal{A}_X = \mathcal{A}_Y = \mathcal{A}_Z = \{0, 1\}$, x and y are independent with $\mathcal{P}_X = \{p, 1-p\}$ and $\mathcal{P}_Y = \{q, 1-q\}$ and

$$z = (x + y) \bmod 2. \tag{8.13}$$

(a) If $q = 1/2$, what is $\mathcal{P}_Z$? What is $I(Z; X)$?

(b) For general p and q, what is $\mathcal{P}_Z$? What is $I(Z; X)$? Notice that this ensemble is related to the binary symmetric channel, with $x =$ input, $y =$ noise, and $z =$ output.

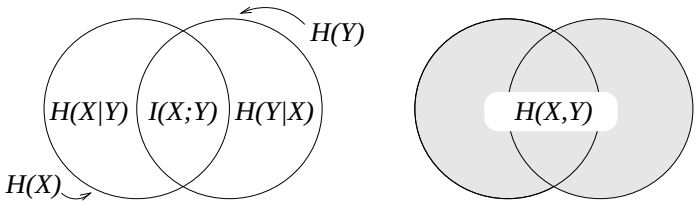


Figure 8.2. A misleading representation of entropies (contrast with figure 8.1).

Three term entropies

Exercise 8.8.[3, p.143] Many texts draw figure 8.1 in the form of a Venn diagram (figure 8.2). Discuss why this diagram is a misleading representation of entropies. Hint: consider the three-variable ensemble XYZ in which $x \in \{0, 1\}$ and $y \in \{0, 1\}$ are independent binary variables and $z \in \{0, 1\}$ is defined to be $z = x + y \bmod 2$.

▶ 8.3 Further exercises

The data-processing theorem

The data processing theorem states that data processing can only destroy information.

Exercise 8.9.[3, p.144] Prove this theorem by considering an ensemble WDR in which w is the state of the world, d is data gathered, and r is the processed data, so that these three variables form a *Markov chain*

$$w \to d \to r, \tag{8.14}$$

that is, the probability $P(w, d, r)$ can be written as

$$P(w, d, r) = P(w)P(d \mid w)P(r \mid d). \tag{8.15}$$

Show that the average information that R conveys about W, $I(W; R)$, is less than or equal to the average information that D conveys about W, $I(W; D)$.

This theorem is as much a caution about our definition of 'information' as it is a caution about data processing!

Inference and information measures

Exercise 8.10.[2] The three cards.

(a) One card is white on both faces; one is black on both faces; and one is white on one side and black on the other. The three cards are shuffled and their orientations randomized. One card is drawn and placed on the table. The upper face is black. What is the colour of its lower face? (Solve the inference problem.)

(b) Does seeing the top face convey *information* about the colour of the bottom face? Discuss the *information contents* and *entropies* in this situation. Let the value of the upper face's colour be u and the value of the lower face's colour be l. Imagine that we draw a random card and learn both u and l. What is the entropy of u, $H(U)$? What is the entropy of l, $H(L)$? What is the mutual information between U and L, $I(U;L)$?

Entropies of Markov processes

▷ Exercise 8.11.[3] In the guessing game, we imagined predicting the next letter in a document starting from the beginning and working towards the end. Consider the task of predicting the *reversed* text, that is, predicting the letter that precedes those already known. Most people find this a harder task. Assuming that we model the language using an N-gram model (which says the probability of the next character depends only on the $N-1$ preceding characters), is there any difference between the average information contents of the reversed language and the forward language?

▶ 8.4 Solutions

Solution to exercise 8.2 (p.140). See exercise 8.6 (p.140) for an example where $H(X\,|\,y)$ exceeds $H(X)$ (set $y=3$).

We can prove the inequality $H(X\,|\,Y) \leq H(X)$ by turning the expression into a relative entropy (using Bayes' theorem) and invoking Gibbs' inequality (exercise 2.26 (p.37)):

$$\begin{aligned} H(X\,|\,Y) &\equiv \sum_{y\in\mathcal{A}_Y} P(y) \left[\sum_{x\in\mathcal{A}_X} P(x\,|\,y) \log \frac{1}{P(x\,|\,y)} \right] \\ &= \sum_{xy\in\mathcal{A}_X\mathcal{A}_Y} P(x,y) \log \frac{1}{P(x\,|\,y)} \qquad (8.16) \\ &= \sum_{xy} P(x)P(y\,|\,x) \log \frac{P(y)}{P(y\,|\,x)P(x)} \qquad (8.17) \\ &= \sum_{x} P(x) \log \frac{1}{P(x)} + \sum_{x} P(x) \sum_{y} P(y\,|\,x) \log \frac{P(y)}{P(y\,|\,x)}. \qquad (8.18) \end{aligned}$$

The last expression is a sum of relative entropies between the distributions $P(y\,|\,x)$ and $P(y)$. So

$$H(X\,|\,Y) \leq H(X) + 0, \qquad (8.19)$$

with equality only if $P(y\,|\,x) = P(y)$ for all x and y (that is, only if X and Y are independent).

Solution to exercise 8.3 (p.140). The chain rule for entropy follows from the decomposition of a joint probability:

$$\begin{aligned} H(X,Y) &= \sum_{xy} P(x,y) \log \frac{1}{P(x,y)} && (8.20)\\ &= \sum_{xy} P(x)P(y\,|\,x) \left[\log \frac{1}{P(x)} + \log \frac{1}{P(y\,|\,x)} \right] && (8.21)\\ &= \sum_{x} P(x) \log \frac{1}{P(x)} + \sum_{x} P(x) \sum_{y} P(y\,|\,x) \log \frac{1}{P(y\,|\,x)} && (8.22)\\ &= H(X) + H(Y\,|\,X). && (8.23) \end{aligned}$$

Solution to exercise 8.4 (p.140). Symmetry of mutual information:

$$\begin{aligned} I(X;Y) &= H(X) - H(X\,|\,Y) && (8.24)\\ &= \sum_{x} P(x) \log \frac{1}{P(x)} - \sum_{xy} P(x,y) \log \frac{1}{P(x\,|\,y)} && (8.25)\\ &= \sum_{xy} P(x,y) \log \frac{P(x\,|\,y)}{P(x)} && (8.26)\\ &= \sum_{xy} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}. && (8.27) \end{aligned}$$

This expression is symmetric in x and y so

$$I(X;Y) = H(X) - H(X\,|\,Y) = H(Y) - H(Y\,|\,X). \quad (8.28)$$

We can prove that mutual information is positive in two ways. One is to continue from

$$I(X;Y) = \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \quad (8.29)$$

which is a relative entropy and use Gibbs' inequality (proved on p.44), which asserts that this relative entropy is ≥ 0, with equality only if $P(x,y) = P(x)P(y)$, that is, if X and Y are independent.

The other is to use Jensen's inequality on

$$-\sum_{x,y} P(x,y) \log \frac{P(x)P(y)}{P(x,y)} \geq -\log \sum_{x,y} \frac{P(x,y)}{P(x,y)} P(x)P(y) = \log 1 = 0. \quad (8.30)$$

Solution to exercise 8.7 (p.141). $z = x + y \bmod 2$.

(a) If $q = 1/2$, $\mathcal{P}_Z = \{1/2, 1/2\}$ and $I(Z;X) = H(Z) - H(Z\,|\,X) = 1 - 1 = 0$.

(b) For general q and p, $\mathcal{P}_Z = \{pq+(1–p)(1–q),\ p(1–q)+q(1–p)\}$. The mutual information is $I(Z;X) = H(Z) - H(Z\,|\,X) = H_2(pq+(1–p)(1–q)) - H_2(q)$.

Three term entropies

Solution to exercise 8.8 (p.141). The depiction of entropies in terms of Venn diagrams is misleading for at least two reasons.

First, one is used to thinking of Venn diagrams as depicting sets; but what are the 'sets' $H(X)$ and $H(Y)$ depicted in figure 8.2, and what are the objects that are members of those sets? I think this diagram encourages the novice student to make inappropriate analogies. For example, some students imagine

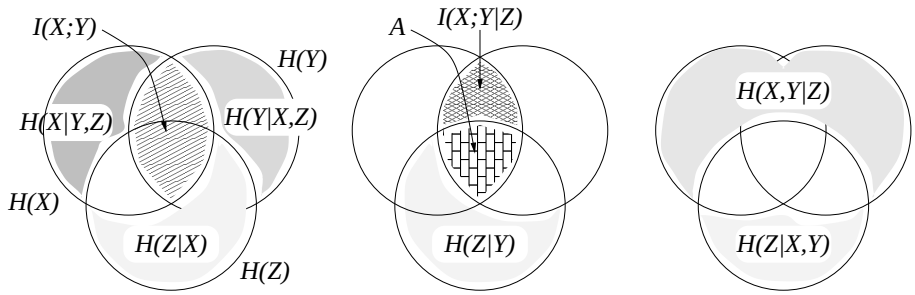


Figure 8.3. A misleading representation of entropies, continued.

that the random outcome (x, y) might correspond to a point in the diagram, and thus confuse entropies with probabilities.

Secondly, the depiction in terms of Venn diagrams encourages one to believe that all the areas correspond to positive quantities. In the special case of two random variables it is indeed true that $H(X \mid Y)$, $I(X;Y)$ and $H(Y \mid X)$ are positive quantities. But as soon as we progress to three-variable ensembles, we obtain a diagram with positive-looking areas that may actually correspond to negative quantities. Figure 8.3 correctly shows relationships such as

$$H(X) + H(Z \mid X) + H(Y \mid X, Z) = H(X, Y, Z). \tag{8.31}$$

But it gives the misleading impression that the conditional mutual information $I(X;Y \mid Z)$ is *less than* the mutual information $I(X;Y)$. In fact the area labelled A can correspond to a *negative* quantity. Consider the joint ensemble (X, Y, Z) in which $x \in \{0, 1\}$ and $y \in \{0, 1\}$ are independent binary variables and $z \in \{0, 1\}$ is defined to be $z = x + y \bmod 2$. Then clearly $H(X) = H(Y) = 1$ bit. Also $H(Z) = 1$ bit. And $H(Y \mid X) = H(Y) = 1$ since the two variables are independent. So the mutual information between X and Y is zero. $I(X;Y) = 0$. However, if z is observed, X and Y become dependent — knowing x, given z, tells you what y is: $y = z - x \bmod 2$. So $I(X;Y \mid Z) = 1$ bit. Thus the area labelled A must correspond to -1 bits for the figure to give the correct answers.

The above example is not at all a capricious or exceptional illustration. The binary symmetric channel with input X, noise Y, and output Z is a situation in which $I(X;Y) = 0$ (input and noise are independent) but $I(X;Y \mid Z) > 0$ (once you see the output, the unknown input and the unknown noise are intimately related!).

The Venn diagram representation is therefore valid only if one is aware that positive areas may represent negative quantities. With this proviso kept in mind, the interpretation of entropies in terms of sets can be helpful (Yeung, 1991).

Solution to exercise 8.9 (p.141). For any joint ensemble XYZ, the following chain rule for mutual information holds.

$$I(X;Y,Z) = I(X;Y) + I(X;Z \mid Y). \tag{8.32}$$

Now, in the case $w \to d \to r$, w and r are independent given d, so $I(W;R \mid D) = 0$. Using the chain rule twice, we have:

$$I(W;D,R) = I(W;D) \tag{8.33}$$

and

$$I(W;D,R) = I(W;R) + I(W;D \mid R), \tag{8.34}$$

so

$$I(W;R) - I(W;D) \leq 0. \tag{8.35}$$