The Joint and Conditional Entropies

PUBLISHED

30 July 2020

This post is part of a series of notes on topics in information theory. The notes assume background knowledge equivalent to an introductory probability course and at some points will require knowledge of univariate calculus. I'd also recommend a refresher on logarithm rules for those who may have forgotten them. For those new to the series the first post is here.

Last time we introduced the concepts of *self-information* and *entropy* as ways of quantifying uncertainty. However, we often want to understand not just the uncertainty in a single event X, but the joint uncertainty of two events X and Y (and potentially a third event Z, etc.) For example, suppose we want to extend our weather reporting service from the previous set of notes. Instead of just reporting the overall weather conditions for the day X (e.g. sunny/overcast) we may also want to report a simplified description of the day's temperature Y (say, above/below 70 degrees F). We can describe the total uncertainty in our new forecast using the *joint entropy*.

Also, often we're also interested in how knowing the outcome of one event Y changes the amount of uncertainty we have about X. Going back to our weather example, suppose that we're back in Seattle during the winter and everyone knows the weather will be overcast all the time - how does this affect the amount of information/surprise contained in our temperature report? We'll learn how to quantify this with our second concept for the day, *conditional entropy*.

We'll start off with the joint entropy. Let's say we have a pair of random variables $(X, Y) \sim p(x, y)$ with X taking on some value $x \in \mathcal{X}$ and Y some value $y \in \mathcal{Y}$. We then have

Definition: *The* ***joint entropy*** $H(X, Y)$ *of two discrete random variables* $(X, Y) \sim p(x, y)$ *is defined as*

$$\begin{aligned} H(X, Y) &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) I(x, y) \\ &= E[I(X, Y)] \end{aligned}$$

where we define $I(x, y) = \log(1/p(x, y))$ as the joint self-information/surprisal of a pair of outcomes $(x, y) \in \mathcal{X} \times \mathcal{Y}$. While this definition may look more complicated than our original definition of entropy $H(X)$ at first glance, we aren't really doing anything new here; we're still weighting the information in each possible outcome (pair) by how often that outcome occurs.

We can generalize the definition of joint entropy to an arbitrary number of random variables $(X_1, X_2, \ldots, X_n) \sim p(x_1, x_2, \ldots, x_n)$, with

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &= -\sum_{x_1, x_2, \ldots, x_n} p(x_1, x_2, \ldots, x_n) \log p(x_1, x_2, \ldots, x_n) \\ &= \sum_{x_1, x_2, \ldots, x_n} p(x_1, x_2, \ldots, x_n) I(x_1, x_2, \ldots, x_n) \\ &= E[I(X_1, X_2, \ldots, X_n)] \end{aligned}$$

Note that the joint entropy depends crucially on the relationship between values of X and Y (i.e., $p(x, y)$). The joint entropy for independent variables X and Y can look *very* different from that when X and Y have some wonky dependency between them. Some examples will help illustrate this idea.

Example 1: *Let* (X, Y) *be a pair of independent discrete random variables with* $X \sim Bern(p_1)$ *and* $Y \sim Bern(p_2)$*. In other words, we have*

$$X = \begin{cases} 1 & \text{with probability } p_1 \\ 0 & \text{with probability } 1 - p_1 \end{cases} \qquad Y = \begin{cases} 1 & \text{with probability } p_2 \\ 0 & \text{with probability } 1 - p_2 \end{cases}$$

Starting from our definition of joint entropy, we have

$$\begin{aligned} H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x)p(y)\log\big(p(x)p(y)\big) \qquad \text{(Independence)} \end{aligned}$$

And then plugging in we have

$$\begin{aligned} H(X,Y) = -\Big[&p_1p_2\log(p_1p_2) + p_1(1-p_2)\log\big(p_1(1-p_2)\big) + \\ &(1-p_1)p_2\log\big((1-p_1)p_2\big) + (1-p_1)(1-p_2)\log\big((1-p_1)(1-p_2)\big)\Big] \\ = -\Big[&p_1p_2(\log p_1 + \log p_2) + p_1(1-p_2)\big(\log p_1 + \log(1-p_2)\big) + \\ &(1-p_1)p_2\big(\log(1-p_1) + \log p_2\big) + (1-p_1)(1-p_2)\big(\log(1-p_1) + \log(1-p_2)\big)\Big] \end{aligned}$$

We can then simplify this equation by grouping terms and remembering $p_i + (1-p_i) = 1$ to eliminate some of the clutter, giving us

$$\begin{aligned} H(X,Y) &= -\Big[p_1\log p_1 + (1-p_1)\log(1-p_1) + p_2\log p_2 + (1-p_2)\log(1-p_2)\Big] \\ &= H(X) + H(Y) \end{aligned}$$

Our final result isn't too surprising here. If we think of our two Bernoulli variables as independent coin flips, we might expect intuitively that the total uncertainty in the two flips would be the sum of that of the individual flips. Later in these notes we'll prove that this result holds for *any* two independent random variables, regardless of their distributions. For now, to further build on our understanding we can graph $H(X, Y)$ as a function of p_1 and p_2, yielding

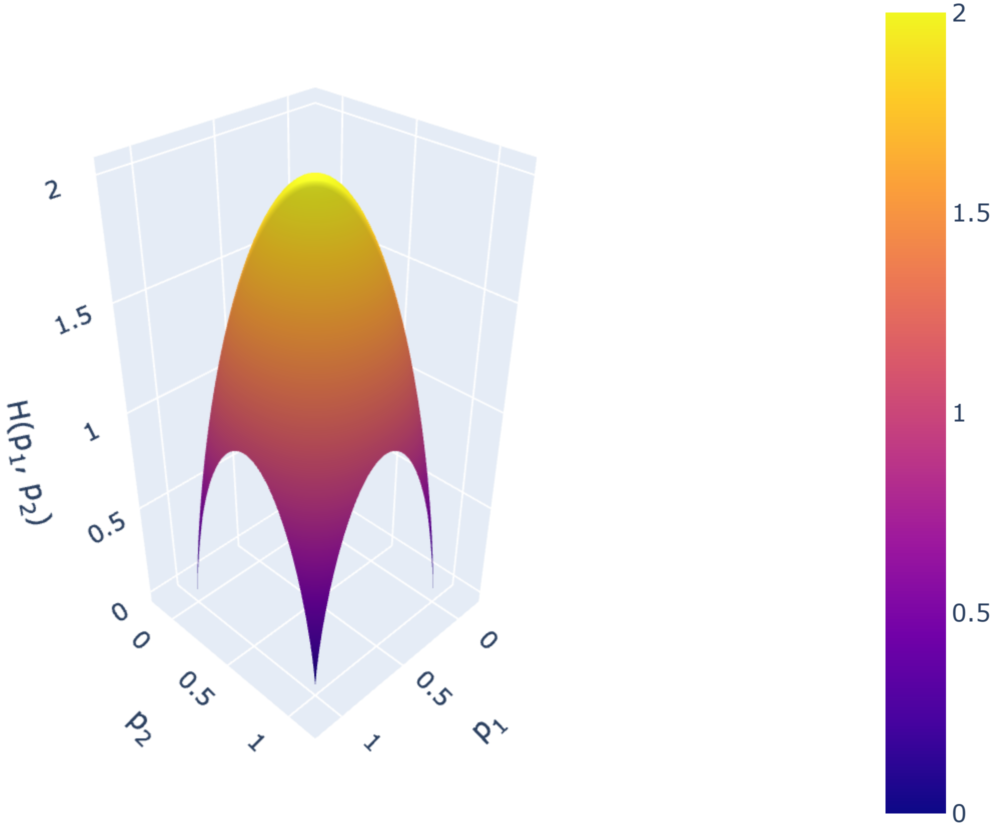
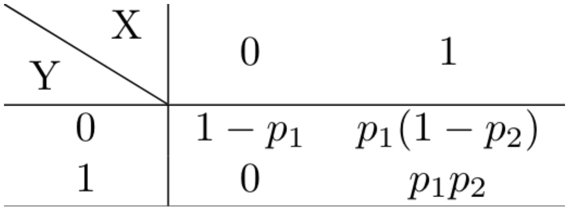


Figure 1. Joint entropy of $X \sim Bern(p_1)$ and $Y \sim Bern(p_2)$ (mouse over to see precise values and drag to rotate). Note that entropy is maximized when $p_1 = p_2 = 1/2$, i.e., when the individual uncertainties of our variables are maximized. At the same time, the joint entropy is minimized with $H(X, Y) = 0$ when X and Y are both deterministic (both p_1 and p_2 are 0 or 1).

As we would expect, our joint entropy is maximized when the individual uncertainties of both variables are maximized. We can also see that our joint entropy is minimized when both variables are deterministic. Finally, if we rotate our plot to face the p_1 or p_2 axis head on, we should see a familiar shape in the outline of our curve - the individual entropy of a Bernoulli variable. This also should make sense, as being fully on the p_1 or p_2 axis in our visualization implies that one of our variables is deterministic and so all the uncertainty is due to the other variable. Now let's consider a more complicated example with non-independent X and Y.

Example 2: *Let (X, Y) be a pair of discrete random variables distributed as follows:*



Y \ X	0	1
0	$1 - p_1$	$p_1(1 - p_2)$
1	0	$p_1 p_2$

where p_1 and p_2 are both between 0 and 1.

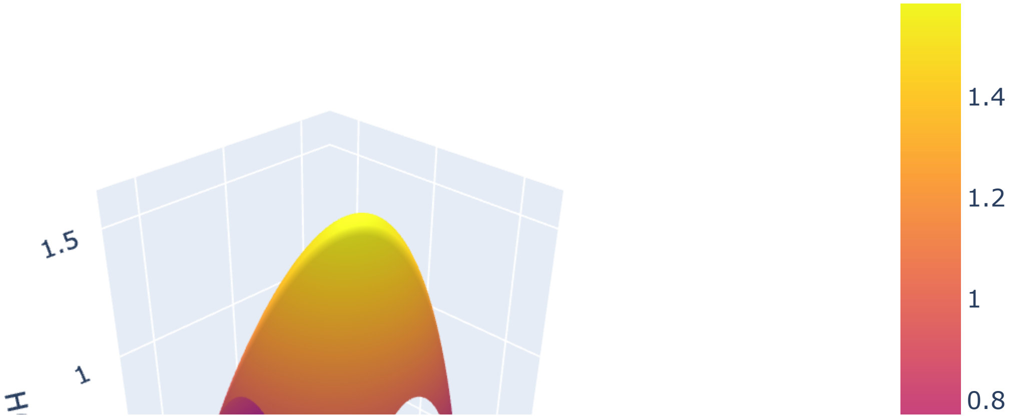
From our table we can see that X has a marginal distribution of $\big(P(X = 0), P(X = 1)\big) = (1 - p_1, p_1)$. In other words $X \sim Bern(p_1)$ with no dependence on the value of Y. On the other hand, $Y \sim Bern(p_2)$ given $X = 1$, otherwise Y is deterministic.

When might such a situation come up? Let's go back to our weather example: If X represents the overall forecast (say $1 =$ sunny, $0 =$ overcast), then Y being 1 or 0 could represent the temperature being above/below 70 degrees respectively. In a simplified model of the weather, we might have that sunny days can potentially have either high or low temperatures while cloudy days are always cool.

Now plugging in to our formula for joint entropy we have

$$\begin{aligned}
H(X,Y) &= -\Big[(1-p_1)\log\big(1-p_1\big) + 0\log 0 + p_1(1-p_2)\log\big(p_1(1-p_2)\big) + p_1p_2\log(p_1p_2)\Big] \\
&= -\Big[(1-p_1)\log\big(1-p_1\big) + p_1(1-p_2)\log p_1 + p_1(1-p_2)\log(1-p_2) + \\
&\qquad p_1p_2\log p_1 + p_1p_2\log p_2\Big] \\
&= -\Big[p_1\log p_1 + (1-p_1)\log(1-p_1) + p_1\big(p_2\log p_2 + (1-p_2)\log(1-p_2)\big)\Big] \\
&= H(X) + \underbrace{\big[-p_1\big(p_2\log p_2 + (1-p_2)\log(1-p_2)\big)\big]}_{\textit{Almost}\text{ the entropy of a Bernoulli RV}}
\end{aligned}$$

Which we can then visualize as



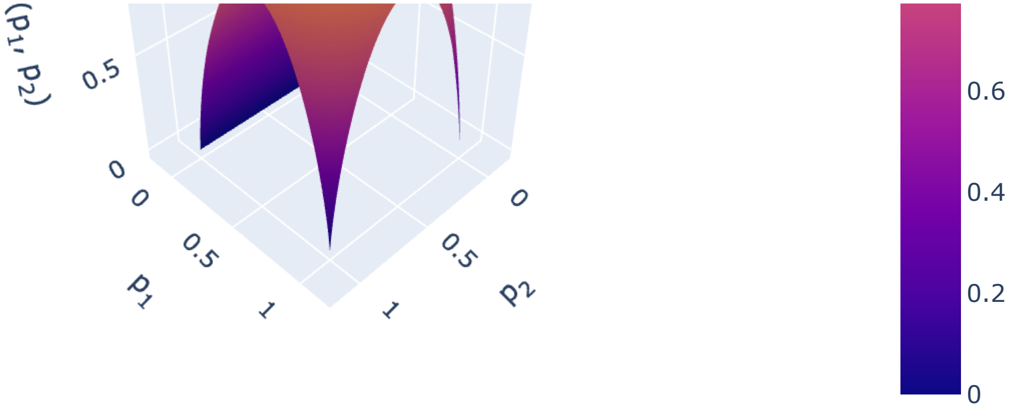


Figure 2. Joint entropy of X and Y as per Example 2. Note that our maximium possible entropy is less than in the independent case. Also note that $H(X,Y) \to 0$ as $p_1 \to 0$, no matter our value of p_2.

In this case the second term in our equation doesn't simplify as nicely as in Example 1 - one can verify that it does *not* equal $H(Y)$. This begs some questions: Are we always guaranteed for arbitrary (X,Y) that $H(X,Y) = H(X) + \text{mystery term}$? Does this second term have a name? The answer to both questions is yes, and this brings us to our next concept, *conditional entropy*.

Definition: *The **conditional entropy** $H(Y|X)$ for discrete random variables $(X,Y) \sim p(x,y)$ is defined as*

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

where we define the conditional self-information $I(y|x) = \log(1/p(y|x))$. We can now calculate $H(Y|X)$ for the variables in our previous example, yielding

$$\begin{aligned} H(Y|X) &= -\Big[\underbrace{p(X=0)H(Y=y|X=0)}_{=0 \text{ since } Y \text{ deterministic when } X=0} + p(X=1)H(Y=y|X=1)\Big] \\ &= -p_1\left(p_2\log p_2 + (1-p_2)\log(1-p_2)\right) \end{aligned}$$

and so we have shown for our previous example that $H(X,Y) = H(X) + H(Y|X)$. This finding should feel natural in some sense; it says that we can compute the joint uncertainty of our variables by first considering the uncertainty of X, and then adding in whatever additional uncertainty Y introduces that isn't already accounted for by knowing X. It turns out that this relationship isn't exclusive to our toy example, and indeed holds in the general case. This brings us to our next theorem:

Theorem: The Chain Rule of Entropy *For two discrete random variables* $(X,Y)\sim p(x,y)$*, we have*

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

Proof:

We start from the definition of the joint entropy

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

This theorem can also be proven in a simpler way using the linearity of expectation and the definitions of joint/individual/conditional entropy. The reader is encouraged to attempt this proof as an exercise and then confirm their answer.[1]

Remark: *By symmetry, we also have*

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

The reader may recall that we stated earlier in the notes $H(X,Y) = H(X) + H(Y)$ for independent X and Y. This follows immediately from the chain rule, if we can show that $H(Y|X) = H(Y)$ for independent X and Y. We do so now.

Lemma : *For two discrete random variables X and Y with $X \perp\!\!\!\perp Y$ (i.e., X and Y are independent) we have*

$$H(Y|X) = H(Y)$$

Proof:

$$\begin{aligned}
H(Y|X) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y) \quad \text{(Independence)} \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x)p(y)\log p(y) \quad \text{(Independence again)} \\
&= -\sum_{x\in\mathcal{X}} p(x)\sum_{y\in\mathcal{Y}} p(y)\log p(y) \\
&= H(Y)\square
\end{aligned}$$

where in the last step we use the definition of entropy and the fact that $\sum_{x\in\mathcal{X}} p(x) = 1$

So far we've defined three(!) different entropies: individual, joint and conditional. Remembering the relationships between all of them can be confusing at first, so the following visual may be helpful:

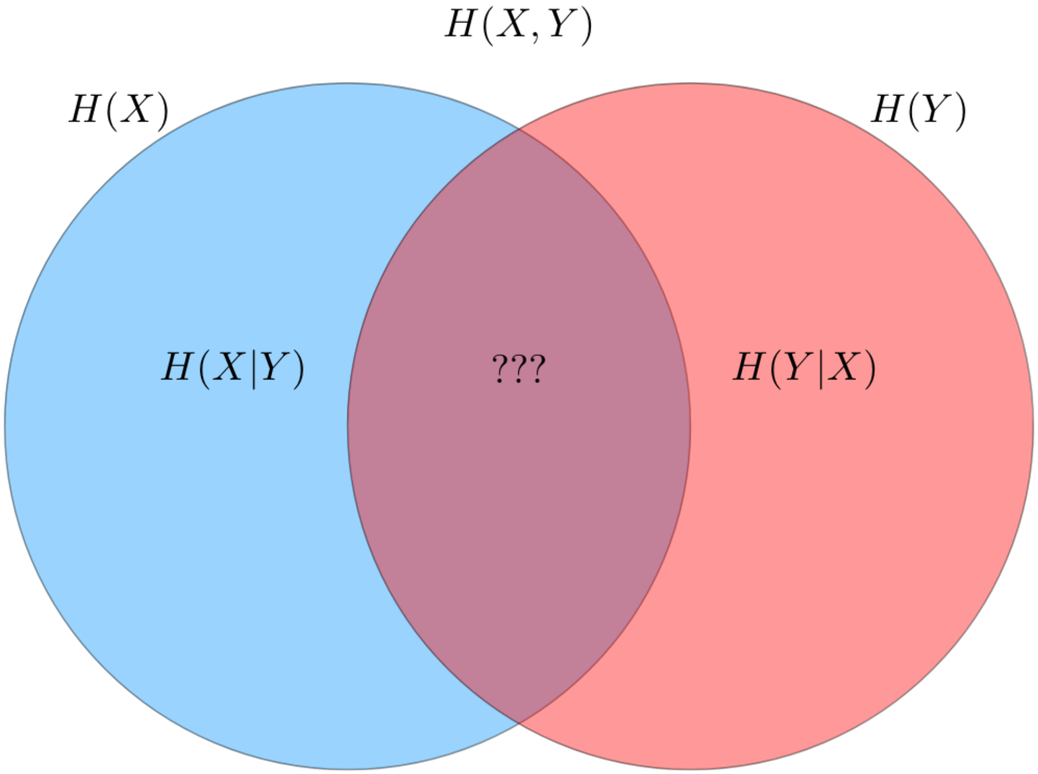


Figure 3. A visual representation of the relationship between individual, joint, and conditional entropies for two variables X and Y. The area contained by both circles is the joint entropy $H(X,Y)$. The left circle is $H(X)$, with the blue (non-intersecting) portion being the conditional entropy $H(X|Y)$. The right circle is $H(Y)$, with the red portion being $H(Y|X)$. The purple intersection of our circles is another interesting quantity called the $\mathit{mutual\ information}$, which we'll explore in the next set of notes.

Our chain rule doesn't just apply to the two variable case. In fact, we can show that it generalizes to any collection of n random variables for our last proof of the day.

Theorem: The Chain Rule (general case) *For a collection of discrete random variables* $(X_1, X_2, \dots, X_n) \sim p(x_1, x_2, \dots, x_n)$ *we have*

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

Proof:

Recall the chain rule of probability; i.e., for a collection of variables $(X_1, X_2, \dots, X_n) \sim p(x_1, x_2, \dots, x_n)$, we have $p(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} p(x_i | x_{i-1}, \dots, x_1)$. We'll use this fact soon.

Now starting from the definition of joint entropy we have

$$\begin{aligned}
H(X_1, X_2, \dots, X_n) &= -\sum_{x_1, x_2, \dots, x_n} p(x_1, x_2, \dots, x_n) \log p(x_1, x_2, \dots, x_n) \\
&= -\sum_{x_1, x_2, \dots, x_n} p(x_1, x_2, \dots, x_n) \log \left(\prod_{i=1}^{n} p(x_i | x_{i-1}, \dots, x_1) \right) \\
&= -\sum_{x_1, x_2, \dots, x_n} p(x_1, x_2, \dots, x_n) \left(\sum_{i=1}^{n} \log p(x_i | x_{i-1}, \dots, x_1) \right) \\
&= -\sum_{x_1, x_2, \dots, x_n} \sum_{i=1}^{n} p(x_1, x_2, \dots, x_n) \log p(x_i | x_{i-1}, \dots, x_1) \\
&= -\sum_{i=1}^{n} \sum_{x_1, x_2, \dots, x_n} p(x_1, x_2, \dots, x_n) \log p(x_i | x_{i-1}, \dots, x_1)
\end{aligned}$$

From here we marginalize out any variables x_j in $p(x_1, x_2, \dots, x_n)$ for which $j > i$. This gives us

$$\begin{aligned}
H(X_1, X_2, \dots, X_n) &= -\sum_{i=1}^{n} \sum_{x_1, x_2, \dots, x_i} p(x_1, x_2, \dots, x_i) \log p(x_i | x_{i-1}, \dots, x_1) \\
&= \sum_{i=1}^{n} H(X_i | X_{i-1}, \dots, X_1) \square
\end{aligned}$$

which concludes this set of notes.

Footnotes:

[1] $-\log p(x,y) = -\Big[\log p(x) + \log p(y|x)\Big] \implies I(x,y) = I(x) + I(y|x)$. Taking the expected value of both sides completes the proof. ↩