

Information Theory Through Toy Examples

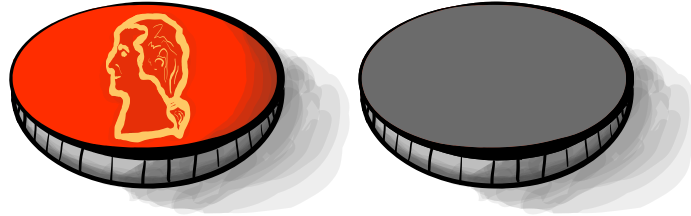
Matthew Andres Moreno

October 25, 2017

1 Introduction

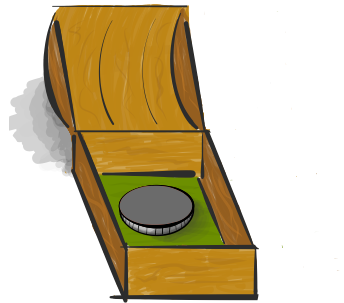
We must start out by defining *entropy*. Entropy is kind of a fancy word for uncertainty. My past encounters with the word entropy have been mostly confusing, verging on mystical. I'm sure many of us have been told, "the entropy of the universe is always increasing." Let's leave entropy's existential baggage behind and treat it just like any other run-of-the-mill word. The best way to get a practical feel for entropy is with an example.

We have a coin. The coin has a red side and a grey side.



The coin is evenly weighted so if we were to throw it in the air, half of the time it would land red side up and the other half of the time it would land grey side up. If we model our coin as a random variable C , we can write this mathematically as $p_{C=\text{red}} = 0.5$ and $p_{C=\text{grey}} = 0.5$.

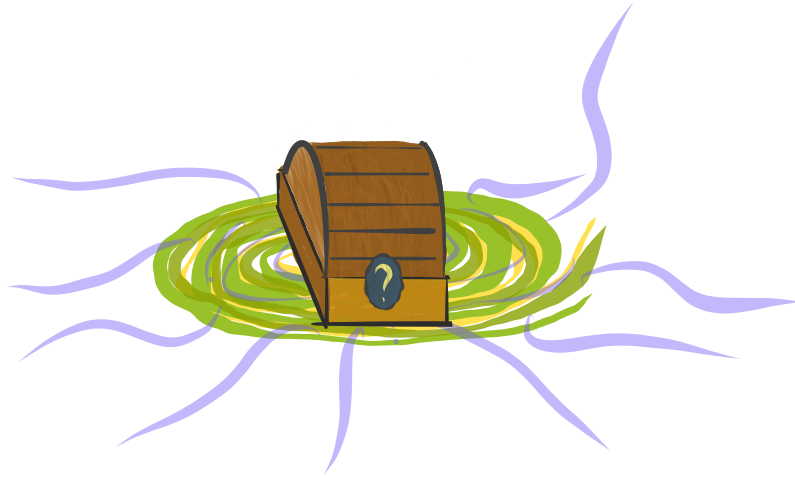
Let's put the coin in a box.



Right now, the grey side of the coin is facing up. How much uncertainty do we have in regard to the position of the coin? None. We know the grey side of the coin is facing up. Thus, entropy is zero. To say this mathematically, we write $S = 0$.

What about the entropy of the molecules that make up the coin? What about the entropy of the universe? We don't have to worry about any of that. Entropy has no intrinsic meaning. It's just a tool that we use to help us concretely understand a particular situation. Right now we are just thinking about which way a coin is facing. So, right now entropy just describes our uncertainty about which way that coin is facing.

Let's close the box and give it a good shake.



We're no longer certain of the coin's position. So, the entropy is greater than zero, or $S > 0$. We can use Shannon's equation to make a more specific mathematical statement about the entropy of our coin.

In general, Shannon's equation allows us to use the set of probabilities for possible outcomes from a random variable $\{p_1, p_2, \dots, p_n\}$ to calculate the entropy of that random variable \mathbf{X} . The equation looks like this,

$$S = H(\mathbf{X}) \tag{1}$$

$$= \sum_{i=1}^n -p_i \log(p_i) \tag{2}$$

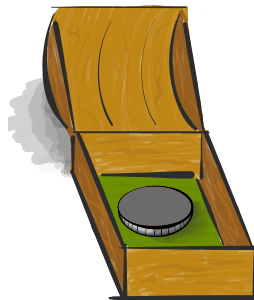
This equation is the main tool we'll use to work through our toy examples. If you're curious about the intuition behind its mathematical form, check out [\[Adami, 2016\]](#).

Recall that for our coin, $\{p_{\text{red}} = 0.5, p_{\text{grey}} = 0.5\}$. Plugging and chugging with Equation 1, we calculate

$$\begin{aligned} S_{\text{coin}} &= H(\mathbf{C}) \\ &= -0.5 \log(0.5) + -0.5 \log(0.5) \\ &= 1. \end{aligned}$$

We used the base two logarithm to perform this calculation, so the unit that describes our entropy value is the "bit." (For those unfamiliar, bit refers to a binary value, one that may only take one of two possible states: a one or a zero. The bit is the appropriate unit for the entropy value because the base of our logarithm matches the number of possible states of a bit.) For consistency's sake, we will use the base two logarithm and the bit unit for all the rest of the calculations we perform. Thus, we say that the coin closed in the box has one bit of entropy.

Now, let's open the box and observe the state of the coin.



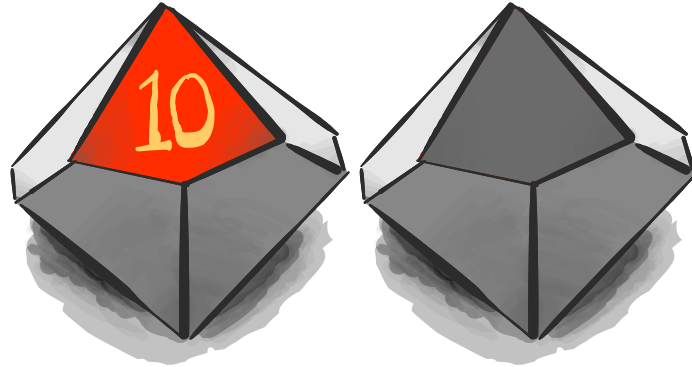
We are again certain about the state of the coin, so $S = 0$. This is where *information* comes into play. *Information* is the difference between two entropies. We can calculate the information that

was gained opening the box by subtracting the ending entropy from the starting entropy,

$$\begin{aligned} I &= S_{\text{before}} - S_{\text{after}} \\ &= 1 - 0 \\ &= 1 \end{aligned}$$

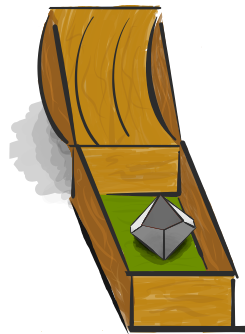
We gained one bit of information. Neat!

We can cement our intuition for what entropy measures by repeating our little thought experiment with a slightly different set up. Now, instead of a coin we have a die. The die has ten faces. Nine of the faces are painted grey. The last face is painted red.

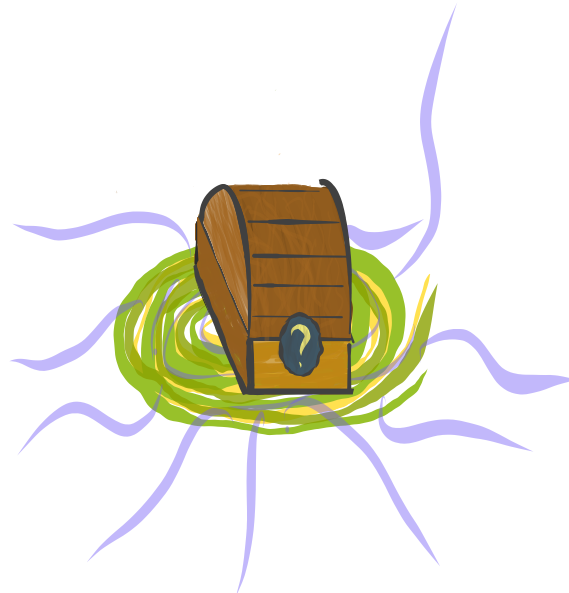


The die is evenly weighted; if we were to throw it in the air, it would land at equal frequency with any of its ten sides up. Thus, nine times out of ten it will land with a grey side up and one time out of ten it will land with the red side up. If we model our die as a random variable D , we can write this mathematically as $p_{D=\text{red}} = 0.1$ and $p_{D=\text{grey}} = 0.9$.

Let's put the die in a box, like so.



We close the box and give it a shake.



We're no longer certain which color is facing up. As before with the coin, we're faced with uncertainty. This raises a key question: are we more uncertain about the color facing up of the die than about the color facing up of the coin?

Let's think through it. Observing both the coin and the die, we recognize two outcomes: red side up or grey side up. The difference is that for the coin, these outcomes have equal probability. For the die, it is more likely that we will observe the grey outcome. Thus, we can make a better guess as to what outcome we will observe on the grey die than we can for the coin. If we always guess grey on the die, we will be correct much more often than if we always guessed grey on the coin. So, we would expect that there is more uncertainty associated with the coin than with the die.

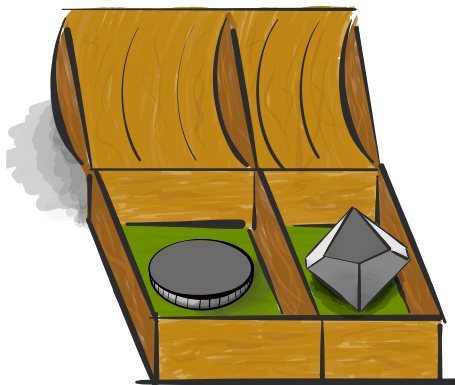
We can use Shannon's equation (1) to formally determine whether the flipped coin or rolled die has more entropy. We will calculate the entropy of the die in the box and then compare it with the entropy of the coin in the box. Plugging and chugging with Equation 1, we calculate

$$\begin{aligned} S_{\text{die}} &= H(\mathbf{D}) \\ &= -0.1 \log(0.1) + -0.9 \log(0.9) \\ &\approx 0.469. \end{aligned}$$

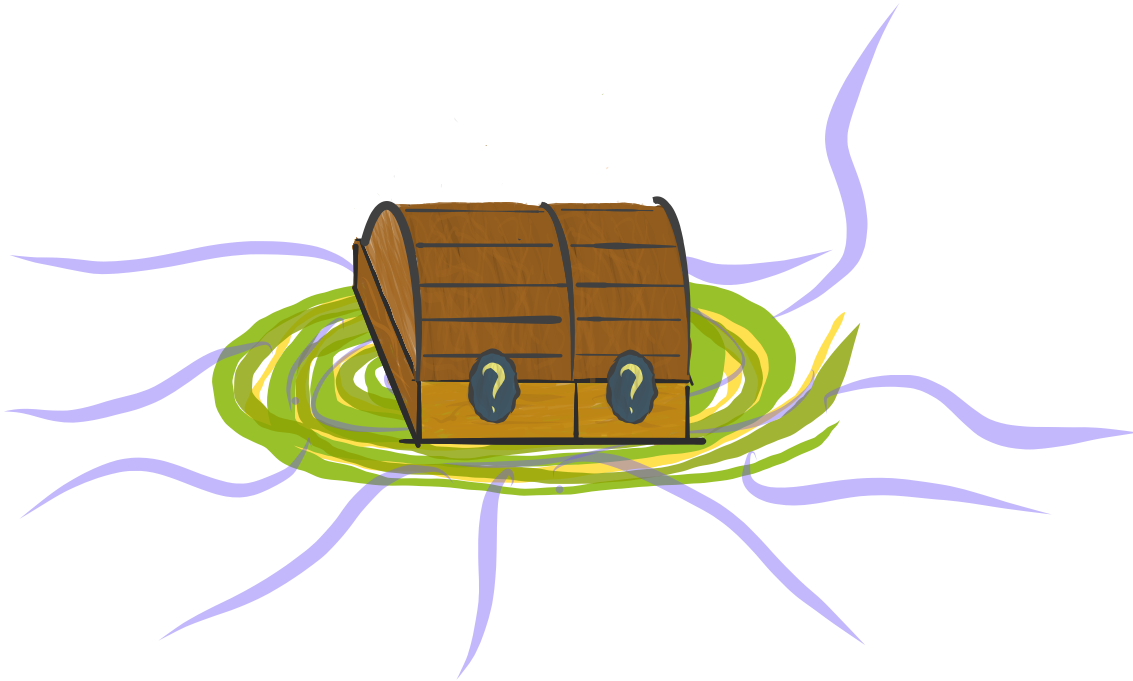
The die in the box has approximately 0.469 bits of entropy. It follows that when we open the box with the die in it (and entropy returns to zero), we gain 0.469 bits of information.

2 Independent Random Variables




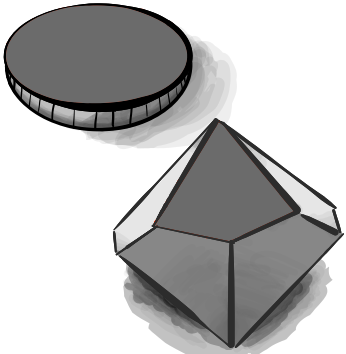
But, wait! There's more fun to be had with the coin and the die. Let's take our two boxes and screw them together.



As before, let's close both lids and give the big box a shake.



We will model the big box as a single random variable A . Now, there are four possible states that our big box could be in. Both the coin and the die could be red side up or grey side up. The possible states of the big box are summarized below.

		C	
		red	grey
D	red	 <p>$A = \text{"red coin, red die"}$</p>	 <p>$A = \text{"grey coin, red die"}$</p>
	grey	 <p>$A = \text{"red coin, grey die"}$</p>	 <p>$A = \text{"grey coin, grey die"}$</p>

Because the die and the coin are independent, it is straightforward to calculate the probability of each of the four possible states of the big box. For example, to compute the probability that

the die and the coin are both grey side up we calculate,

$$\begin{aligned}
 p_{\mathbf{A}=\text{"grey coin, grey die"}} &= p_{\mathbf{C}=\text{grey}} \times p_{\mathbf{D}=\text{grey}} \\
 &= 0.5 \times 0.9 \\
 &= 0.45.
 \end{aligned}$$

The probabilities of all four possible states of the big box are summarized below.

\mathbf{A}		\mathbf{C}		sum
		red	grey	
\mathbf{D}	red	0.05	0.05	0.1
	grey	0.45	0.45	0.9
sum		0.5	0.5	1

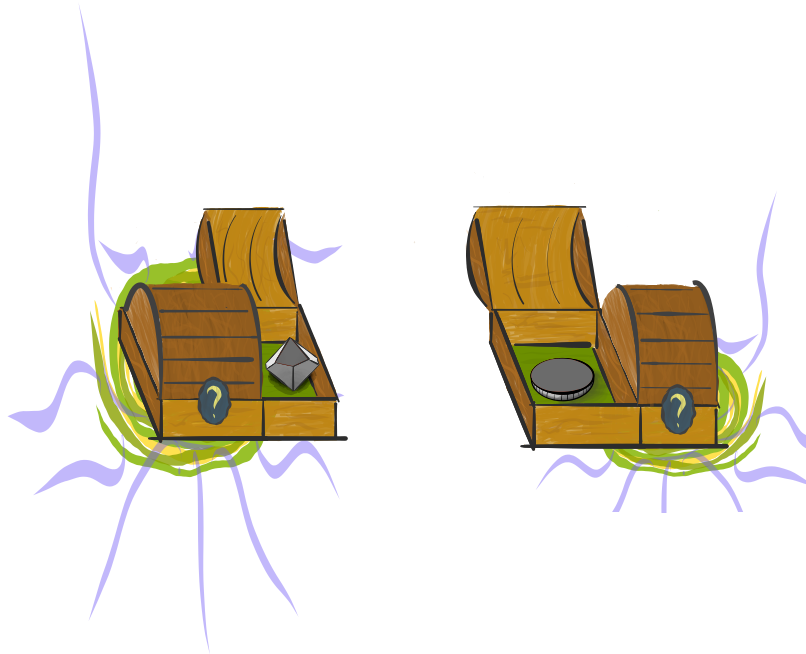
Now that we know the probabilities for all the possible outcomes of \mathbf{A} , we can calculate the entropy associated with the the closed box.

$$\begin{aligned}
 S_{\text{big box}} &= H(\mathbf{A}) \\
 &= -0.05 \times \log_2(0.05) - 0.05 \times \log_2(0.05) - 0.45 \times \log_2(0.45) + -0.45 \times \log_2(0.45) \\
 &\approx 1.469
 \end{aligned}$$

Let's take a moment to notice,

$$S_{\text{big box}} = S_{\text{die}} + S_{\text{coin}}.$$

Intuitively, it makes sense that the entropy associated with the big box is the same as the sum of the entropy associated with the die box and the entropy associated with the coin box. This relation holds true because the coin and the die are independent. Intuitively speaking, the coin and the die are independent because knowing the state of the coin doesn't the probability of outcomes from the die and vice versa.



The coin still has the same probability of taking on the red state or grey state given we know the state of the die. The die still has the same probability of taking on the red state or grey state given we know the state of the coin. Mathematically,

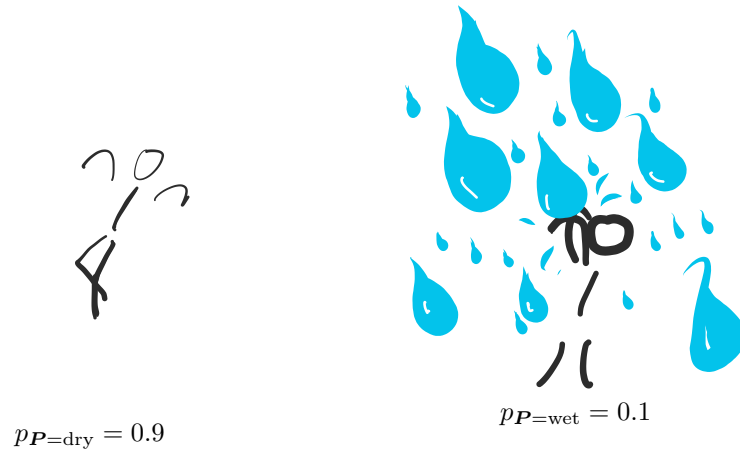
$$\begin{aligned}
 P(\mathbf{C}) &= P(\mathbf{C}|\mathbf{D}) \\
 P(\mathbf{D}) &= P(\mathbf{D}|\mathbf{C}).
 \end{aligned}$$

For those unfamiliar, the vertical bar is read as “given.”

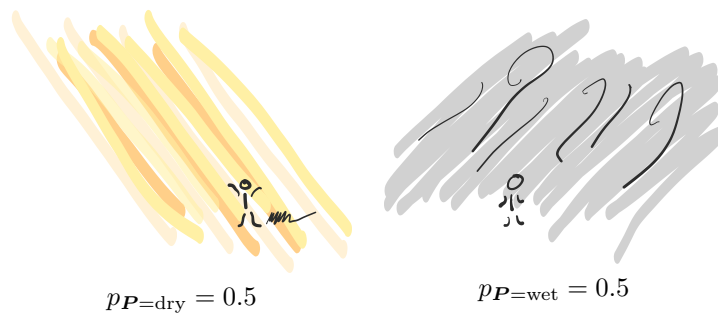
In the next section, we'll look at entropy and information in a situation where independence *doesn't* hold. This is where things get a little more interesting.

3 Dependent Random Variables

Let's put our die and coin away to step outside. To look at dependent random variables, we will consider a hypothetical weather scenario. We will consider two aspects of the weather: precipitation and light. In our hypothetical scenario, precipitation conditions may either be dry or wet. We will model the situation with a random variable P that can take on the value "dry" or "wet." The probability that conditions are dry is 0.9 and the probability that conditions are wet is 0.1.





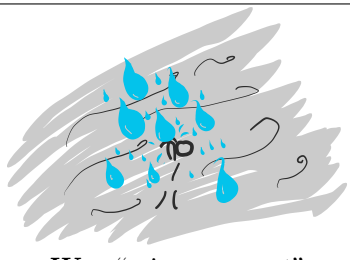

Similarly, we model light conditions with a random variable L that can take on the value "sunny" or "overcast." Sunny means that clouds are not currently occluding the sun (but may be present elsewhere in the sky). Overcast means that the sun is currently occluded by cloud cover. The probability that conditions are sunny is 0.5 and the probability that conditions are overcast is 0.5.



This setup looks familiar. In the coin and die scenario, we had a pair of random variables that could each take on two values. One random variable takes on each of its values with probability $\frac{1}{2}$. the other random variable takes on one of its values with probability $\frac{9}{10}$ and the other with probability $\frac{1}{10}$.

The key difference is that precipitation and light are not independent. Intuitively, we expect the wet and overcast conditions to be correlated and the dry and sunny conditions to be correlated. For example, if it's raining we would be very surprised if it was sunny!

Let's look at how precipitation and light interact in our hypothetical scenario. We will model the weather — the combination of precipitation and light — as a single random variable W . The weather can be in four possible states,

		P	
		rainy	dry
L	sunny		
		$W = \text{"rainy, sunny"}$	$W = \text{"dry, sunny"}$
overcast			
		$W = \text{"rainy, overcast"}$	$W = \text{"dry, overcast"}$

Because we are no longer operating under the assumption of independence, the probabilities of the four states of W can no longer be derived from the probability distributions of P (precipitation) and L (light). Just as we took it at face value that in our scenario the probability of the sunny and cloudy condition are both $\frac{1}{2}$, we take the probability of each of the four possible states of W at face value as describing the particular scenario we are imagining. (Of course, this is under the constraint that the marginal probabilities of P and L are consistent with what we stated before). Let's define the probability distribution of W as follows,

W		L		sum
		sunny	overcast	
P	wet	0.005	0.095	0.1
	dry	0.495	0.405	0.9
sum		0.5	0.5	1

The best way to get a feel for the distribution of W is to look at conditional probabilities. For example, here's the probability that we observe the sunny weather given we know that it's wet outside,

$$P(L = \text{sunny} | P = \text{wet}) = 0.05.$$

Although we won't elaborate on it, calculating such conditional probabilities from the table above is very straightforward. Let's look through the rest of the rest of the conditional probabilities we can calculate.

Consider

$$P(L = \text{sunny} | P = \text{dry}) = 0.55$$

$$P(L = \text{sunny}) = 0.5$$

$$P(L = \text{sunny} | P = \text{wet}) = 0.05.$$

Here, we see that the probability of observing sunny weather is greatest if we know that it's dry outside and very small if we know that it's wet outside.

Likewise,

$$P(L = \text{overcast} | P = \text{wet}) = 0.95$$

$$P(L = \text{overcast}) = 0.5$$

$$P(L = \text{overcast} | P = \text{dry}) = 0.45.$$

The probability of observing overcast weather is greatest if we know that it's wet outside and less if we know that it's dry outside.

Similarly,

$$\begin{aligned} P(\mathbf{P} = \text{wet} | \mathbf{L} = \text{overcast}) &= 0.19 \\ P(\mathbf{P} = \text{wet}) &= 0.1 \\ P(\mathbf{P} = \text{wet} | \mathbf{L} = \text{sunny}) &= 0.01. \end{aligned}$$

The probability of observing wet weather is greatest if we know it's overcast outside and least if we know it's sunny outside.

Finally,

$$\begin{aligned} P(\mathbf{P} = \text{dry} | \mathbf{L} = \text{sunny}) &= 0.99 \\ P(\mathbf{P} = \text{dry}) &= 0.9 \\ P(\mathbf{P} = \text{dry} | \mathbf{L} = \text{overcast}) &= 0.81. \end{aligned}$$

The probability of observing dry weather is greatest if we know it's sunny outside and least if we know it's overcast.

Armed with a good understanding of the probability distribution of our weather \mathbf{W} , we're ready to talk entropy and information. We can calculate entropy of a random variable whose distribution we know using Equation 1. Let's do that for \mathbf{W} ,

$$\begin{aligned} S_{\text{weather}} &= H(\mathbf{W}) \\ &= -0.005 \times \log_2(0.005) + -0.095 \times \log_2(0.095) - 0.495 \times \log_2(0.495) + -0.405 \times \log_2(0.405) \\ &\approx 1.391 \end{aligned}$$

for \mathbf{L} ,

$$\begin{aligned} S_{\text{light}} &= H(\mathbf{L}) \\ &= -0.5 \times \log_2(0.5) + -0.5 \times \log_2(0.5) \\ &= 1 \end{aligned}$$

and, finally, for \mathbf{P} ,

$$\begin{aligned} S_{\text{precipitation}} &= H(\mathbf{P}) \\ &= -0.1 \times \log_2(0.1) + -0.9 \times \log_2(0.9) \\ &\approx 0.469. \end{aligned}$$

Just like for the die and the coin, $S_{\text{light}} > S_{\text{precipitation}}$. Like before, this makes sense because we can make a better guess about precipitation (it is wet infrequently) than about light conditions (it is overcast or sunny with equal probability).

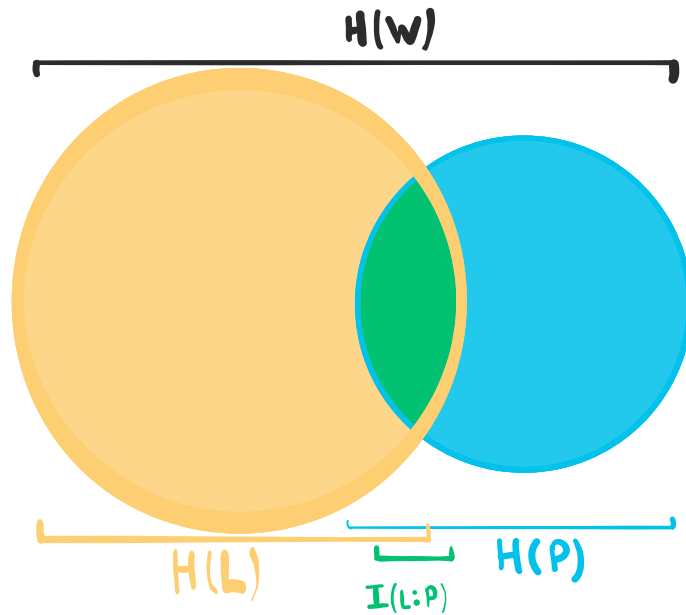
Here's the shocker,

$$S_{\text{weather}} \neq S_{\text{light}} + S_{\text{precipitation}}$$

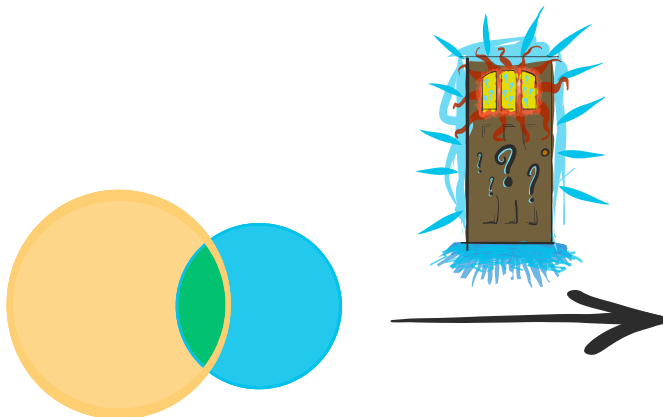
In fact, $1.391 < 1.469$, so

$$S_{\text{weather}} < S_{\text{light}} + S_{\text{precipitation}}.$$

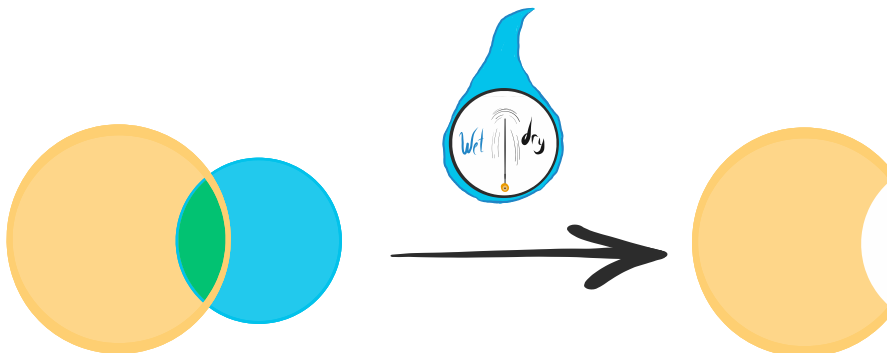
How can our uncertainty about the weather as an entire system be less than our uncertainty about its constituent parts? The answer is that some entropy is shared between \mathbf{L} and \mathbf{P} . The situation looks like this.



The green silver in the middle of this venn diagram is the entropy shared between L and P . I denoted that sliver $I(L : P)$. (Shared entropy is symmetric, so it would be just as valid to denote it $I(P : L)$). In total, the shaded area represents the entropy of the entire system $H(W)$. If we stepped out the door to observe the state of the weather all at once, all of that entropy would go away.



This idea of sharing entropy has an interesting implication. Say we had a rain meter that tells us the precipitation conditions. Knowing the precipitation conditions removes all the blue uncertainty ($H(P)$) from our diagram, taking some of the yellow uncertainty ($H(L)$) with it.



Thus, we would expect our uncertainty about light conditions to be reduced by knowing the precipitation conditions. This is indeed the case. We can calculate this directly!

To calculate the entropy after peeking at the rain meter, we need to calculate our uncertainty as to the light condition after we observe the rain meter. There are two outcomes from looking at

the rain meter — we learn it's wet outside or we learn it's dry outside. Remember the conditional probabilities we reviewed? To calculate entropy given a particular precipitation state, we proceed as usual with Shannon's equation but substitute probabilities conditional on the precipitation state we observed for unconditional probabilities.

If we find it's wet outside, we calculate

$$\begin{aligned} H(\mathbf{L}|\mathbf{P} = \text{wet}) &= -P(\mathbf{L} = \text{sunny}|\mathbf{P} = \text{wet}) \times \log_2(P(\mathbf{L} = \text{sunny}|\mathbf{P} = \text{wet})) \\ &\quad + -P(\mathbf{L} = \text{overcast}|\mathbf{P} = \text{wet}) \times \log_2(P(\mathbf{L} = \text{overcast}|\mathbf{P} = \text{wet})) \\ &\approx -0.05 \times \log_2(0.05) + -0.95 \times \log_2(0.95) \\ &\approx 0.286. \end{aligned}$$

If we find it's dry outside, we calculate

$$\begin{aligned} H(\mathbf{L}|\mathbf{P} = \text{dry}) &= -P(\mathbf{L} = \text{sunny}|\mathbf{P} = \text{dry}) \times \log_2(P(\mathbf{L} = \text{sunny}|\mathbf{P} = \text{dry})) \\ &\quad + -P(\mathbf{L} = \text{overcast}|\mathbf{P} = \text{dry}) \times \log_2(P(\mathbf{L} = \text{overcast}|\mathbf{P} = \text{dry})) \\ &= -0.55 \times \log_2(0.55) + -0.45 \times \log_2(0.45) \\ &\approx 0.993. \end{aligned}$$

From these results, we can calculate the expected entropy of \mathbf{L} given observation of \mathbf{P} . We denote this quantity as $H(\mathbf{L}|\mathbf{P})$. We calculate $H(\mathbf{L}|\mathbf{P})$ by taking an average of $H(\mathbf{L}|\mathbf{P} = \text{dry})$ and $H(\mathbf{L}|\mathbf{P} = \text{wet})$, weighted by the probabilities $P(\mathbf{P} = \text{wet})$ and $P(\mathbf{P} = \text{dry})$,

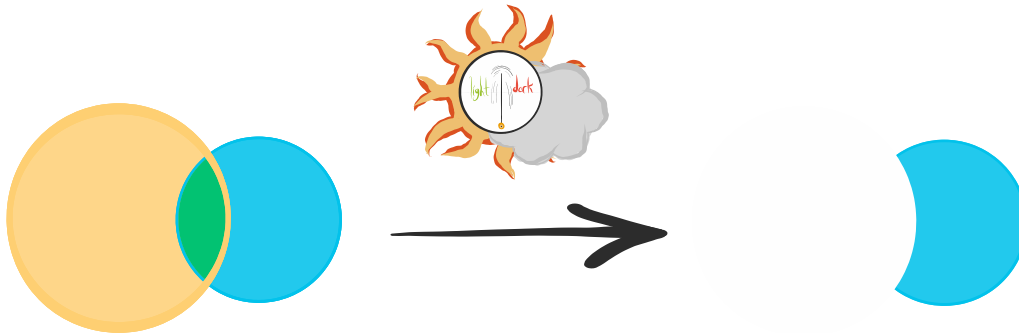
$$\begin{aligned} H(\mathbf{L}|\mathbf{P}) &= P(\mathbf{P} = \text{wet}) \times H(\mathbf{L}|\mathbf{P} = \text{wet}) \\ &\quad + P(\mathbf{P} = \text{dry}) \times H(\mathbf{L}|\mathbf{P} = \text{dry}) \\ &\approx 0.1 \times [-0.05 \times \log_2(0.05) + -0.95 \times \log_2(0.95)] \\ &\quad + 0.9 \times [-0.55 \times \log_2(0.55) + -0.45 \times \log_2(0.45)] \\ &\approx 0.922. \end{aligned}$$

Recall that information is the difference between entropies. Thus, we can think of the shared entropy between \mathbf{P} and \mathbf{L} as information gained with respect to \mathbf{L} when we know the state of \mathbf{P} ,

$$\begin{aligned} I(\mathbf{L} : \mathbf{P}) &= H(\mathbf{L}) - H(\mathbf{L}|\mathbf{P}) \\ &\approx 1 - 0.922 \\ &\approx 0.078 \end{aligned}$$

(This is why shared entropy is denoted with the capital I !)

The reverse holds true, as well. Our uncertainty about light conditions is reduced by knowing the precipitation conditions. Checking our light meter (instead of the rain meter) removes all the yellow uncertainty ($H(\mathbf{L})$) from our diagram, taking some of the blue uncertainty ($H(\mathbf{P})$) with it.



Similar calculations give the expected entropy of \mathbf{P} given observation of \mathbf{L} ,

$$\begin{aligned}
H(\mathbf{P}|\mathbf{L}) &= P(\mathbf{L} = \text{overcast}) \times H(\mathbf{P}|\mathbf{L} = \text{overcast}) + P(\mathbf{L} = \text{sunny}) \times H(\mathbf{P}|\mathbf{L} = \text{sunny}) \\
&= P(\mathbf{L} = \text{overcast}) \times \left[-P(\mathbf{P} = \text{wet}|\mathbf{L} = \text{overcast}) \times \log_2(P(\mathbf{P} = \text{wet}|\mathbf{L} = \text{overcast})) \right. \\
&\quad \left. + -P(\mathbf{P} = \text{dry}|\mathbf{L} = \text{overcast}) \times \log_2(P(\mathbf{P} = \text{dry}|\mathbf{L} = \text{overcast})) \right] \\
&\quad + P(\mathbf{L} = \text{sunny}) \times \left[-P(\mathbf{P} = \text{wet}|\mathbf{L} = \text{sunny}) \times \log_2(P(\mathbf{P} = \text{wet}|\mathbf{L} = \text{sunny})) \right. \\
&\quad \left. + -P(\mathbf{P} = \text{dry}|\mathbf{L} = \text{sunny}) \times \log_2(P(\mathbf{P} = \text{dry}|\mathbf{L} = \text{sunny})) \right] \\
&= 0.5 \times [-0.19 \times \log_2(0.19) + -0.81 \times \log_2(0.81)] \\
&\quad + 0.5 \times [-0.01 \times \log_2(0.01) + -0.99 \times \log_2(0.99)] \\
&\approx 0.391
\end{aligned}$$

Thus, the information gained with respect to \mathbf{P} when we know the state of \mathbf{L} is

$$\begin{aligned}
I(\mathbf{P} : \mathbf{L}) &= H(\mathbf{P}) - H(\mathbf{P}|\mathbf{L}) \\
&\approx 0.469 - 0.391 \\
&\approx 0.078.
\end{aligned}$$

We have computationally confirmed the symmetric nature of shared entropy,

$$\begin{aligned}
I(\mathbf{L} : \mathbf{P}) &= I(\mathbf{P} : \mathbf{L}) \\
&0.078 \checkmark \approx 0.078
\end{aligned}$$

Also, as we would expect from visual inspection of our Venn diagram of entropy,

$$\begin{aligned}
H(\mathbf{W}) &= H(\mathbf{P}) + H(\mathbf{L}) - I(\mathbf{L} : \mathbf{P}) \\
&1.391 \checkmark \approx 0.469 + 1 - 0.078
\end{aligned}$$

4 Conclusion

There you have it. Information is the difference between entropies. Entropy is a quantification of uncertainty. If a discrete random variable models a situation, calculate entropy can be calculated as a function of the set of probabilities associated with the possible outcomes of the situation. Just plug that set of probabilities into Shannon's equation to calculate entropy.

I hope these illustrated examples have helped you get a hands-on feel for information and entropy. Now, go check out [Adami, 2012] for some nicely described applications of information theory to biology and evolution. If you want to further firm up your footing with information theory itself, give [Adami, 2016] a read. Don't be afraid to make your own toy problems and analyze them yourself! It's the best way to get comfortable with any new math topic.

What's the entropy associated with a fair six-sided die? What's the information content of a three letter code if all three letters are drawn independently from a uniform distribution over all 26 letters? What if the three letters are drawn without replacement? If you're looking for a challenge, maybe check out the [Monty Hall problem](#). What's the entropy when all three doors are closed? What's the entropy after the host opens a goat door? How much information is gained when the host opens a goat door? This one's tricky. Have fun!

References

- [Adami, 2012] Adami, C. (2012). The use of information theory in evolutionary biology. *Annals of the New York Academy of Sciences*, 1256(1):49–65.
- [Adami, 2016] Adami, C. (2016). What is information? *Phil. Trans. R. Soc. A*, 374(2063):20150230.