Information Theory for Intelligent People

Simon DeDeo*

February 13, 2017

Contents

1 Twenty Questions

The story of information theory begins with the children's game usually known as "twenty questions". The first player (the "adult") in this two-player game thinks of something, and by a series of yes-no questions, the other player (the "child") attempts to guess what it is. "Is it bigger than a breadbox?" No. "Does it have fur?" Yes. "Is it a mammal?" No. And so forth.

*Current version available at `http://santafe.edu/~simon/it.pdf`. Article modified from text for the Santa Fe Institute Complex Systems Summer School 2012, and updated for a meeting of the Indiana University Center for 18th Century Studies in 2015, a Colloquium at Tufts University Department of Cognitive Science on Play and String Theory in 2016, and a meeting of the SFI ACtioN Network in 2017. Please send corrections, comments and feedback to `sdedeo@andrew.cmu.edu`; `http://santafe.edu/~simon`.

If you play this game for a while, you learn that some questions work better than others. Children usually learn that it's a good idea to eliminate general categories first before becoming more specific, for example. If you ask on the first round "is it a carburetor?" you are likely wasting time—unless you're playing the game on Car Talk.

If a game is lasting a *very* long time, you might start to wonder if you could improve your strategy beyond this simple rule of thumb. "Could I have gotten the answer sooner, if I had skipped that useless question about the fur?" A moment's reflection shows that, in fact, what counts as a good set of questions depends upon the player: if someone is biased towards material things, you'll tend to focus on questions that split hairs among weights, sizes, and shapes. If someone is biased towards historical figures, you might split hairs about eras of birth.

Strange as it may seem at first, it turns out that for any particular opponent, you can talk not just about strategies being better or worse, but about the possibility of an optimal strategy, one that is better (or at least as good as) any other you might invent. To see how this works, first imagine writing down a script for playing against an opponent, a set of rules that covers all the contingencies: "first ask if it's a person; then if yes, ask if they were born before 1900, if no, ask if it's a country..."; or for another friend: "first ask if it's bigger than a breadbox; if yes, then..."; or for someone else, "first ask if it's Donald Duck. It almost always is." Visually, the script could be represented by a branching tree (Fig. 1), with a question at each branch, and one of two paths to take depending on the opponent's answer.

Once we specify a script, we can try to work out how effective it is. if we describe your opponent by the list of probabilities he has of choosing any particular option, we can then talk about the average number of yes/no questions it takes before the game is finished. Let's say there are N words, and label the words x_i, where the index i runs from one to N.For each thing your opponent could be thinking of, x_i, the game will end in a predetermined number of steps, $L(x_i)$—you can just read this off the tree, if you like. The average number of steps is just the sum of all the $L(x_i)$, weighted by the probability $P(x_i)$,

$$\text{Script Performance} = \sum_{i=1}^{N} P(x_i)L(x_i) \tag{1}$$

Having defined a script, and figured out how to compute how quickly it can finish the game, we can then ask the big question: what's the optimal script, the script with the shortest average time-to-finish for that particular player. Consider a simple case where there are three options—"tree", "car", and "bird"; Fig. 1, which you looked at before. If the adult picks "car" half the time when he plays the game, and the other two options a quarter of the time each, then the optimal script for the child is easy to guess. The child should ask, first, if it is a car, because then, half the time, the game will be over in one question. The other half of the time, she will then go on to ask "is it a tree", and now (knowing that there are only three options) the answer, even if it is "no", definitively picks out one of the two options. The average number of yes/no questions is 1.5, and a little thought (or experimentation) shows that that's the best you can do.

To recap: we just figured out the optimal tree, and then, using the probabilities $P(x)$, worked out exactly how well it did (1.5 questions, on average). Now things get very strange. It so turns out that we could have gotten the answer to the second question without having to solve the first. In particular, if we compute the quantity $H(X)$,

$$H(X) = -\sum_{i=1}^{N} P(x_i) \log_2 P(x_i), \tag{2}$$

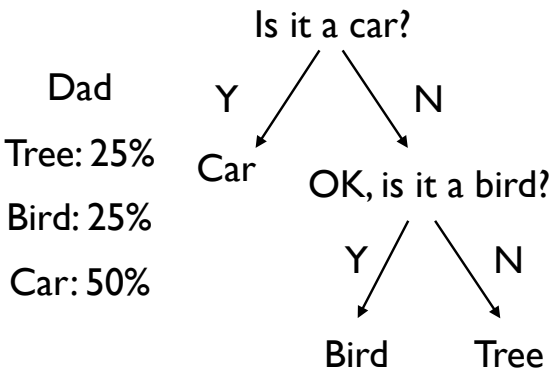

Figure 1: A script for twenty questions, where there are only three options: at each branch-point in the tree, we ask a question, depending on the answer, we take either the left-hand or right-hand branch. Depending on the nature of the opponent, we can then compute the average number of questions required to terminate the game. Given an opponent ("Dad") who picks "car" half the time, and the two other options one-quarter of the time each, the script here will terminate in one step half the time ("Is it a car?" "Yes"); and in two steps the other half of the time ("Is it a car?" "No." "OK, is it a bird?" "Yes."). Since we know that Dad only picks from a set of three, two "no"s suffice to tell us it's a tree. As we'll see later, this particular script turns out to be optimal; there's no set of questions that will end the game sooner (on average).

it is either equal to, or slightly less than, the average length for the optimal script. When it is slightly less than, the true length is always less than or equal to $H(X)$ rounded up to the nearest integer. Here $\log_2 x$ is the logarithm of x, in base two; so, for example, $\log_2 1/8$ is equal to -3 because $1/8$ is equal to 2^{-3}. You can check by explicit computation that $H(X)$ for the tree/car/bird problem gets the answer exactly; this is true when the probabilities are all integer powers of $1/2$.[1]

Eq. 2 is quite a remarkable result: we're able to compute the average length for a script without explicitly constructing it. Indeed, there are often a number of different, equivalent optimal scripts (in Fig. 1, for example, the second question could be "is it a tree"), and in many cases, particularly when there are a large number of options, script construction is non-trivial. We won't prove it in this brief introduction, but if you compare $H(X)$ to the way you'd compute the average length of an arbitrary tree, Eq. 1, you'll notice that it suggests that the number of questions to get to choice x, $L(x)$, is equal to $-\log_2 P(x)$—*i.e.*, the less likely possibilities are buried deeper in the question tree.

$H(X)$ is the basic quantity in information theory. In an important sense, all the other quantities we'll compute are variant on it. $H(X)$ goes by a number of different names: "uncertainty", "information", even "entropy" (a term from the physical sciences, which we'll return to later). It has units—bits—a name which comes from the phrase "binary digit", which we'll understand a bit better below. For any probability distribution, we can now talk about how uncertain we are about the outcome, how much information is in the process, or "how much entropy the process has", and

[1]If this little inaccuracy in the relationship between $H(X)$ and the optimal script bothers you, here's a way to solve it. Imagine that instead of building a script for a single round of the game, you build it for N rounds simultaneously. You're now asking questions to resolve what the last N guesses were, all at once, and the questions might be weird—"was your fourth guess an animal and your twelfth guess larger than a breadbox?"—but no matter. The question tree will take longer, but the average number of questions for the simultaneous game, divided by N, will get closer and closer to $H(X)$ as N gets large.

even (if we are told, or find out, what the probabilities actually are) measure it, in bits.

2 Sidebar: Information on Ice

$H(X)$ is a fundamentally epistemic quantity. It quantifies how an actual agent in the world goes about gathering information about what is going on. The adult knows with certainty what he has in mind; it is the child, asking yes/no questions, to whom we attach the uncertainty. (To learn more about the modern interpretation of probabilities as describing mental states, rather than facts about the world, see the companion article, Bayesian Reasoning for Intelligent People.)

Because of this, we're tempted to say that whatever information theory measures is a subjective thing, a fact not about the thing, but rather the mind of the beholder. Since we're usually quantifying information in terms of what goes on (literally) in someone's head, this interpretation works very well. But if you want to go a little deeper, you'll discover that information in this subjective sense corresponds directly to a basic quantity in the physical sciences, entropy. Entropy is a fact about the world; you can look up, for example, the "entropy released when ice melts into water". How is this possible?

Let's take the water-ice case. Ice is a crystal, which means that, in any little patch of ice, the water molecules are arranged in a repeating pattern. The nature of that crystal is such that for every position that a water molecule occupies, there's a gap right next door. When ice melts, there are (roughly) the same number of water molecules in that patch. But now they're jumbled about, no longer constrained to stay on the crystal positions, and (in particular) they can now float freely between where they "should" be, and that gap next door. Put in an epistemic fashion, once the ice melts, you become more uncertain about each of the water molecules. You have to ask not only, say, "what's the orientation of the molecule?", but also "is it on the crystal lattice, or off?" That's, literally, one extra yes/no question, one extra bit, and in fact if you do the conversion to the units the chemists use, you get pretty close to the measured value if you say that "the change in entropy when ice melts to water is one bit/molecule"

This is exceptional. A basic quantity in industrial chemistry is directly related to a statement about how physical changes affect what we (or, rather, an ideal angelic agent who can operate on the molecular level) can know. Chemists don't usually talk this way. I learned this way of computing the entropy of the ice-water transition from a 1970 paper published in the Soviet Union, by Yu. V. Gurikov of the All-Union Scientific-Research Institute of Petrochemical Processes, Leningrad. But it is true nonetheless.

So something strange is going on: information is "real." It's up to you to decide, perhaps on the sofa one evening, whether that means that mind is just matter, or (conversely) that matter is an illusion of mind. To give a sense of current exchange rates between mind and matter: the number of bits released when one gram of ice melts is about 100 billion Terabytes, just a little bit more than all the long-term storage humans attach to their computers. I can't tell if it's more amazing that (1) the information in a cube of ice exceeds all of our hard-drives, or (2) humans are starting to get within striking distance of basic thermodynamic limits. Think of that the next time you have a Scotch on the rocks.

3 Encoding and Memory

Let's return to the question script, understood as a binary yes/no tree. Since the answer is arrived at by a specific string of yes/no answers in the script, that string can be understood as an encoding

of the answers themselves. In the script of Fig. 1, for example, we can encode "car" as Y, "tree" as NY, and "bird" as NN, a series of "binary digits" (hence, bits) that could be written 0/1 so they look more computery. Each string is unique, and so it really is a code. If two strings match exactly, it means that the script was unable to distinguish them, and a new question needs to be added. You can then transmit to a friend the guess by flashing the Y/N code—long-flash for Y, say, short-flash for N.

The code has some very nice properties. In particular, you don't need a special "space" or "pause" character to transmit multiple symbols. If you encode a series of choices in an unbroken fashion, they can be uniquely decoded. Consider the transmission YNYNNNYYNYNYNNY. Because the code is variable length, you can't figure out how many choices are in this list from the length alone. Because a valid codeword is never the prefix to another valid codeword, however, as you read along you know when you've finished reading a word and are beginning a new one (try it). Reading left to right gives an unambiguous decomposition. (It's a nice little puzzle to try to prove this to yourself. Can you see why?)

When the tree is optimal, the encoding is efficient. Computer scientists recognize the construction of the optimal tree as *Huffman Encoding*. Huffman Encoding is the most basic form of lossless compression, when your information units come in chunks—the choices of the adult—and you want to transmit them efficiently in a binary code down, say, a telephone wire. $H(X)$ then gives you the number of "bits per symbol" your system requires on average; if the underlying process you wish to encode has $H(X)$ equal to, say, 55, then if the symbols are thrown up at a rate of one per second, you'll need transmission system that can handle (on average) 55 bits per second. Sometimes you'll do much better (if you get a common symbol, it will tend to have a short code, and your telegraph operator can take a break) and sometimes much worse (in rare cases, you're standing around asking questions much longer than you expect).

Huffman Encoding is an incredibly simple way to achieve optimal transmission and storage, and it's therefore widespread in the computer world. When you "zip" something, the algorithm under the hood is very often some variant of Huffman Encoding. It's a lossless encoding, meaning that even though it generally finds a much faster way to transmit the information, nothing is lost. It just takes advantage of the biases in the source, giving nicknames or shortcuts to the most common things you say.[2]

It's more than just an engineering trick, however. The information encoding story enlarges the epistemic "script" interpretation by presenting a story about optimal storage, or communication. The outcome of a process with a lot of uncertainty is harder to remember, or transmit; conversely, an agent with a little experience can develop a codebook that allows them to efficiently gather and store information about the world. We'll return to these ideas further down when we talk about expectations and learning.

4 Coarse-graining

Our question-tree presentation is a nice way to introduce the story of information theory, but it's not the only way, and, historically, it wasn't the way it was introduced by the founder of information theory, Claude Shannon, when he was working at AT&T (see Ref. [1]).

Shannon built up his story axiomatically, by saying that he wanted to measure the uncertainty in a process. We want a function, in other words, call it $H(\vec{p})$, that takes a list of probabilities and

[2]Note that unless you have a pre-arrangement with the other end, you'll need to transmit the tree itself—the question script—down the wire as well.

spits out a single number, uncertainty. Let's require the function to obey four simple axioms. The first two are so simple that they almost seem trivial. The third is intuitive. The fourth is deep.

1. Continuity (if I only change the probabilities a little, the information of the process should change only a little).

2. Symmetry (if I reorder the list of probabilities I gave you, you should get the same answer).

3. Condition of Maximum Information: $H(\vec{p})$ is at its maximum value when all the p_i are equal.

4. Coarse-Graining (discussed below).

Continuity is simple, but Symmetry is a bit loaded: it says that the information is the same whether the probability of events I feed it is (for example) $\{p_A = 0.2, p_B = 0.8\}$ or $\{p_A = 0.8, p_B = 0.2\}$. Of course, if the two probabilities represent the dispositions of a jury, option A is "guilty" and option B is "not guilty," the prisoner will care a great deal! But H does not care. For this reason, information is often called a "syntactic" theory, concerned only with the properties of symbols in abstraction from their meanings.

The Maximum condition fits with our folk concept: if every possible outcome (from the list of possible outcomes) is equally likely, the process has the maximum information.[3]

The Coarse-Graining axiom takes the idea of uncertainty step further. It's really about how we group things together and throw out distinctions. When we say "he was driving a car" rather than "he was driving a red Chevrolet with Massachusetts plates", we're coarse-graining, ignoring, or refusing to transmit, a bunch of information that would distinguish very different events. The information you discard may or may not be useful, depending on the use you expect to put the information to, but at the very least it's more efficient to drop it. Imagine, for example, an indulgent father who allows his child to get "close enough" in the game of twenty questions; depending on the child's script, it could have a significant effect on the game.

For a more formal definition of coarse graining, let's say we have a set, X, with three options, $\{a, b, c\}$. I can talk about the uncertainty of a process that spits out one of these three symbols. But what if I don't care about (or can't tell) the difference between b and c? Instead of making a finer distinction between b and c, I just lump them together into some super-symbol S. If I coarse-grain X in this way, I can talk about the uncertainty of probability distribution over the reduced set, X', $\{a, S\}$, where S refers to "b or c, I don't care", and p_{bc} is just $p_b + p_c$.

When Shannon thought about this process, he realized it could be a very nice way to constrain the function form of entropy. In general, we'd like $H(X')$ to be less than (or at worst, equal to) $H(X)$—if you have an indulgent father, it will help you win faster. Even better, Shannon decided he wanted the following "tree like" property to hold:

$$H(X) = H(X') + p_{bc}H(G), \tag{3}$$

where $H(G)$ is the uncertainty of the choice between the group G containing b and c; more explicitly, it's the uncertainty of the distribution $\{p_b/p_{bc}, p_c/p_{bc}\}$.

[3]Note! Do not get confused here. The process itself in this case is very random, and we often associate randomness with lack of information (*e.g.*, if a student starts talking randomly in class, we say he is telling us nothing). But information is concerned with specifying the outcome of a process; imagine having to describe the behavior of a student to a doctor. If the student's behavior is very random, you have to have a longer conversation ("on Monday he was taking about cats, but then Tuesday he was on to bus schedules, and Wednesday it was about the filling in his teeth...") as opposed to a less random, and thus lower information, process ("every day, he says hello and that's it.")

People reading Shannon's original paper realized that Eq. 3 was a seriously elegant move, and to see why it's worth looking at it in different ways. If, for example, you consider a question script, then $H(X')$ is the average length of the coarse-grained script, and $H(G)$ the average length of the subscript necessary to make the fine-grained distinctions. The average length of the full script is equal to the coarse-grained script except in the case you have to split within the G, which happens with probability p_{bc}. So the coarse-graining axiom contains within itself that branching, tree-like logic.

Once we demand this mathematical property hold, something magical happens. There is now only (up to a constant factor) one (one!) possible mathematical form. It is

$$H(\{p_1, p_2, \ldots, p_n\}) = -\sum_{i=1}^{n} p_i \log p_i \tag{4}$$

where the choice of the base of the logarithm is the one freedom. If you chose the base-2 logarithm, H has units of bits—the same bits as in the opening question-script story. Any other choice of function, beyond picking different bases, will violate at least one of the four conditions (and usually more than one). In fact, as is easy to show, with this choice, the coarse-graining property of Eq. 3 holds for an arbitrary coarse-graining operation where a fine-grained description X is transformed into a coarse-grained description X' with a bunch of different groups G_1, G_2, and so on.

5 Alternatives to Entropy?

$H(X)$ can also be considered a measure of diversity; indeed, it's a much better version than simply "counting up appearances of unique types". In many fields where people have tried to come up with a diversity measure, they've ended up inventing a function that, while not quite equal to $H(X)$, approximates it over some interval while not having any of its nice properties. One common alternative choice is to work from the idea that while probabilities must sum to unity, if you have a reasonably even distribution, with lots of small probabilities, the sum of the squares will be lower. Thus, one can define a diversity index

$$D(\{p_1, p_2, \ldots, p_n\}) = 1 - \sum_{i=1}^{n} p_i^2, \tag{5}$$

which is equal to zero when only one option is possible, and gets larger as other options come in to play; another justification for the choice is that it's equal to the probability that two samples taken in a row have different types. It can be shown that $D(X)$ is proportional to a second-order Taylor series approximation to $H(X)$; see Fig. 2 for the binary cases where there are two probabilities. $D(X)$ satisfies three of the conditions, but fails to obey the coarse-graining principle.

While $D(X)$ is nice because you don't have to take logarithms, and so can sometimes be used in algebraic solutions, I recommend not reinventing the wheel! $H(X)$ has so many beautiful and useful properties—we'll see more later—it's a pity to spoil it. Economists are particularly bad at doing this, and the Herfindahl index (Economics), as well as the Simpson index (Ecology) are both equal to $1 - D(X)$; there's also the Inverse Simpson index, $1/D(X)$, which has the same property as $1 - D(X)$ of being large when probabilities are more "evenly spread".

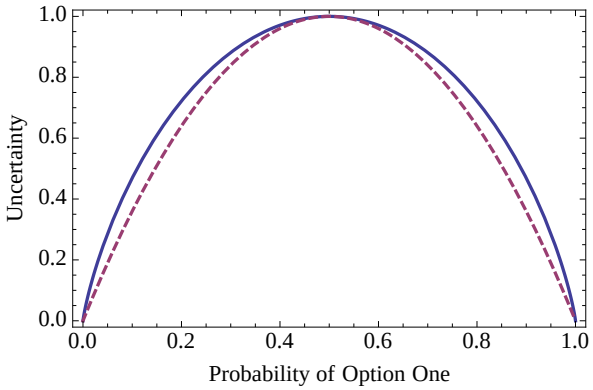

Figure 2: Alternatives to entropy? Blue solid line: the function $H(X)$ for a binary choice, as a function of the probability of heads. Red dashed line: a common approximation to $H(X)$ that people sometimes invent, based on the fact that more "uncertain" or "diverse" distributions tend to have lots of small probabilities. They're close, but not quite overlapping.

6 Coding Failure, Cognitive Surprise, and Kullback-Leibler Divergence

An optimal script is, in general, only optimal for a particular probability distribution over symbols. What happens when the distribution changes, but you keep the same script? Let's say that the first adult, who was biased towards "car", was your father (he's an Uber driver). Your mother, by contrast, is a professor of ecology and so she is biased towards trees, choosing "tree", and not "car", with probability one-half (and the other two options with probability one-quarter each).

If we use the Dad script on your mother, we get the situation on the right-hand panel of Fig. 3. Even though your mother's choice has the same uncertainty (and so, therefore, there's an equally optimal script), the dad script is no longer optimal, and instead of taking 1.5 questions on average, it wastes time by splitting on a less-likely choice, and takes 1.75—an extra quarter-question—on average. That inefficiency can be thought of as due to a failure of expectations, a violation (or refinement) of a prior, or a change in the nature of the information source itself.

It so turns out that this coding failure can also be quantified without explicitly constructing script pairs. It is equal to what is called the Kullback-Leibler divergence (KL), which is defined as

$$KL(p|q) = \sum_{i=1}^{N} q(x_i) \log_2 \frac{q(x_i)}{p(x_i)}, \tag{6}$$

where $p(x)$ is the distribution you trained on (built the question script for; your dad), but $q(x)$ is the new distribution you're now encountering (the mother). The "Kullback-Leibler from p to q", or $KL(p|q)$, tells you how many additional questions you will ask, on average, over and above what you'd have to ask if you were using an optimal script. (As in the relationship between uncertainty and tree depth, there are caveats about cases where this can be off because the question tree is short; but if one lumps together a bunch of outputs and encodes that, these differences go to zero.)

We can describe KL in terms of question script inefficiencies, or coding failures, and like uncertainty itself, there are many interpretations beyond that. We can talk, for example, about the "surprise" (or "Bayesian surprise") of one distribution given that you're expecting another. Fitting in with an epistemic interpretation about agents testing beliefs, It's also equal to the rate at which

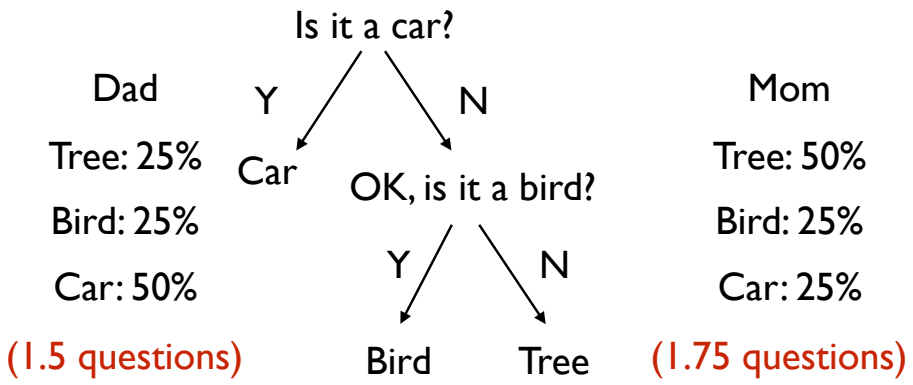

Figure 3: Optimal scripts for twenty questions. Given an opponent ("Dad") who picks "car" half the time, and the two other options one-quarter of the time each, the game can be solved in an average of 1.5 questions when the script is optimal. But scripts are optimal usually only for the probability distribution they were built for; when an optimal script is applied to a new distribution, it is usually no longer optimal, and inefficiencies in representation—"coding failures"—emerge.

evidence in favor of the events being drawn from distribution q (rather than p) accumulate over time—where evidence is defined as the log of the ratio of probability between the q hypothesis and the p hypothesis, averaged over lots of trials, and the rate is per sample; see Ref. [2].

For these latter reasons, minimizing the KL of a distribution p from the "true" distribution q can be understood as making p as epistemically close to truth as possible; hence its use in model selection, where it is part of the Akaike Information Criterion [3], as well as its role in our new proposal for correcting algorithmic bias when machine learning is used for criminal justice and policy-making [4].

One of the most exciting things I discovered in my work is that the surprise that KL divergence measures has real mental effects. The first use of KL divergence to quantify surprise, or coding failure, in a cognitive system appears to be Itti & Baldi's work on eye-tracking [6]. They did something very clever: they took a movie clip, and measured the KL divergence spatially over the screen. Formally, they divided up the screen into little patches. In each patch, they computed the probability distribution over pixel colors at time t (that's the p distribution) and then again at time t plus one second (that's the q distribution). Then they could compute the KL divergence between p and q in each patch, and (since the movie clip lasts more than a few seconds), actually see how it evolves over time.

Amazingly, their eye-tracking devices showed that people tended to look not (say) at the bright patches, or the uncertain patches, but at the high KL-divergence patches. We tend, in other words, to look towards those parts of a video that violate our expectations (rather than, for example, those parts that just have high uncertainty—we get bored of television fuzz). With some colleagues and students, we went to town on these results, extending them to the cultural and economic domains. Our work on Darwin's reading patterns [7] is (to our knowledge) the first use of this in the study of natural language (or, indeed, historical archives).

7 Small Towns, Big Cities, and Cromwell's Rule

People sometimes use KL as a distance measure—"how different are two distributions"—but this is incorrect for a very interesting reason: KL is not necessarily symmetric! The surprise on encountering q given that you're expecting p may not be the surprise on encountering p given that you're expecting q. An organism that grows up in a rich sensory environment will, in general, be less surprised on encountering an improvised one than an organism who has grown up in an impoverished environment and encounters a rich one.

If you'd like an explicit example, consider the difference between me and Einstein. We both sleep and eat a lot, but if you're watching Einstein he has a small chance p of inventing a new theory of physics. Say my distribution over sleeping, eating, and inventing new physics is $\{0.7, 0.3, 0\}$ while Einstein's distribution is $\{0.7, 0.2, 0.1\}$. If you're expecting Einstein, but you get me, you're not that surprised; we both spend a lot of time sleeping and eating. But if you expect me, and you get Einstein, you will be very surprised when you see him invent a new theory of physics.

(Because Kullback-Leibler divergence is not symmetric, some people like to invent a distance measure for the "distance" between p and q by symmetrizing it: $KL(p|q) + KL(q|p)$. This is a suboptimal solution because, while it does make the measure symmetric, it still fails as a measure of distance because it violates the triangle inequality, which requires that the distance from p to q must be less than or at worst equal to the distance from p to r plus the distance from r to q. Below, we'll see a better metric to use in its place.)

Returning to the Simon-Einstein case, if you think the probability of me inventing a new theory of physics is precisely zero, you will be infinitely surprised—KL divergence will blow up because of the $0.1 \log (0.1/0)$ term. It's as if the question script never terminates because the possibility wasn't written in, or the transmission mechanisms stalls because it encounters a symbol not in its look-up table.

This is a good place to be reminded of how strange it is to attribute zero probability to an event; as you can learn in greater detail in the accompanying article on Bayesian reasoning, if an agent attributes probability zero to an event it means that no evidence whatsoever, no evidence of any form, can convince you that the event took place. This violates what is sometimes referred to as the "Cromwell Rule", after the 17th Century English military dictator, who famously urged the Church of Scotland "I beseech you, in the bowels of Christ, think it possible that you may be mistaken" (*i.e.*, do not attribute probability zero to something). Practically speaking, if you're estimating probabilities using a Bayesian method (such as a topic model), you'll never get a zero value; if you're using a naive estimator of probabilities that relies on frequencies, you might, and I recommend adding a small regularizing psuedo-count, $1/k$, to each bin. See Ref. [5] for more on why this makes sense.[4]

8 Mutual Information

Once you can measure $H(X)$, the uncertainty in a process, you can talk about how that uncertainty changes when you acquire information from somewhere else. For example, while you might be

[4]The Simon-Einstein case also provides a nice example of how KL is an asymptotic measure of inefficiency. If my probability of revolutionizing physics is non-zero (call it ϵ), then because an observer needs to code for all three possibilities, they will end up using a Y/NY/NN code for both of us, with Y coding sleep, and the inefficiency will be zero even though KL grows as ϵ becomes small. It's only when we start encoding multiple moments at once that the differences in coding efficiency per event approach the KL value—the Einstein code will start to make efficient codes for combinations of moments that include revolutionizing physics well before the Simon code will.

maximally uncertain about the card on the top of a shuffled deck, your uncertainty goes down a great deal once I tell you the suit. Instead of having an equal choice over 52 cards (an entropy of $\log_2 52$, about 5.7 bits), you have an equal choice over the thirteen cards of that suit ($\log_2 13$, or 3.7, bits). Put another way, telling you the suit was worth two bits of information.

In general, if two variables, X and Y, are correlated with each other, this means that the probability distribution over one of the variables is affected by the value of the other variable. If it's raining, then I'm more likely to be carrying an umbrella; more formally, $P(X|y)$ will be different depending on the value of y, and the uncertainty of the distribution over X given that Y is equal to y, which we write as $H(X|y)$, will differ from the uncertainty over X before you got this information, $H(X)$.

If we average over all the possible values of Y, weighted by their probability, we get what is called the conditional entropy of X given Y, or $H(X|Y)$,

$$H(X|Y) = \sum_{j \in M} H(X|y_j)P(y_j) \tag{7}$$

where we assume that there are M possible values of Y. in some cases, $H(X|y)$ can be larger than $H(X)$; if, for example, X gets more random under unusual conditions. However, it can be shown (using something called Jensen's inequality) that on average, *information never hurts*. In other words, $H(X|Y)$ is always less than, or (in the case that Y and X are independent) equal to $H(X)$.

We usually talk in terms of the extent to which Y reduces uncertainty about X; the mutual information, I, between X and Y is thus defined as

$$I(X, Y) = H(X) - H(X|Y), \tag{8}$$

and a little algebra suffices to show that, in fact, this is symmetric—the information that Y gives you about X is equal to the information that X gives you about Y; or $H(X) - H(X|Y)$ is equal to $H(Y) - H(Y|X)$. How could it not be?—if you tell me something and I learn about the world, I can reverse the process, look at the world, and figure out what you'll say about it.

Mutual Information is at the heart of a number of problems where people try to figure out how much information, or influence, is propagating between systems, from one part of the country to another, all the way down to one word to the next in a text. You can think of it as the "most general" form of correlation coefficient that you can measure. In an analogous fashion to how uncertainty talks about the optimal question script (without needing to specify it in details), mutual information talks about the optimal method for learning about one thing using information about another.

9 Jensen-Shannon Distance

Let's return to Mom, Dad, and twenty questions one last time. Say you're playing over a computer with one of them, but you don't know which one. Over time, as you play, you will start to accumulate information in favor of one or the other—you'll notice a preponderance of "car" choices, indicating your father, or "tree" choices, indicating your mother.

How much information about the identity of the player do you get from playing a single round? From the symmetry of mutual of information, this is equivalent to how much information about the identity of the word you get from learning who the player is, or how much your uncertainty is reduced, on average, by gaining this. This is called the Jensen-Shannon Distance, or JSD, between P and Q:

$$JSD(P, Q) = H(M) - \frac{1}{2}(H(P) + H(Q)), \tag{9}$$

where M is the mixture distribution of the two adults, m_i is equal to $\frac{1}{2}(p_i + q_i)$. In words, the JSD tells you how much one sample, on average serves to distinguish between the two possibilities.

A simple computation shows that this is equal to the KL from M to P plus the KL from M to Q,

$$JSD(P, Q) = \frac{1}{2}(KL(M|P) + KL(M|Q)) \tag{10}$$

A very nice feature of the JSD is that it (or, rather its square-root) is not just symmetric—it's also a metric that obeys the triangle inequality [8], making it a good tool for using dimensionality reduction tools such as multidimensional scaling (MDS, such as the `cmdscale` in `R`) that work best with sensible distances. Because m_i is zero if and only if both p_i and q_i are zero, it's also the case that JSD never blows up.

For these reasons, JSD serves well as a method for understanding distinctiveness a "distance" between two probability distributions. It plays the same role for probabilities as Euclidean distance does for ordinary "spatial" vectors. Ref. [9] was this author's introduction to its use in empirical work, where it quantifies the distinguishability of two "genres" of court case in the 18th and 19th Centuries.

10 A Note on Measuring Information

Talking about information, and proving theorems about it, is one thing. Measuring the creation, flow, and extinction of information in the real world, on data gathered in the field, is a separate task.

In many cases, the "naive", or "plug-in" approach to measuring information-theoretic quantities in the real world works very well. You begin by estimating the probability of events from frequency counts; given these estimated probabilities, it's a simple matter to plug them in to the formula you care about—whether it be uncertainty, Kullback-Leibler divergence, mutual information, or Jensen-Shannon Distance.

If you have enough data, and not too many categories, this will usually work very well. A good rule of thumb is that you should have at least ten times as many samples as you do classes or categories that you're measuring [10]. So, twenty tosses of a coin are sufficient to get a good estimate of the uncertainty of the distribution; forty samples of two stocks will be enough to determine the mutual information between their "up" or "down" ticks (a 2x2 grid of possibilities—stock one up while stock two down, and so on).

When you don't have enough data—or you want to do more sophisticated things, like determine error bars on estimates—you have to be a bit smarter. The estimation of information theoretic quantities from data has a big history (often independently rediscovered in different fields). Ref. [10] describes some of the issues that arise, and THOTH (`http://thoth-python.org`) implements some of the most common tools in python. When you are measuring the uncertainty of distributions over very large cardinalities, consider the NSB estimator [5] `http://nsb-entropy.sourceforge.net`; over continuous spaces, the NPEET toolkit `http://www.isi.edu/~gregv/npeet.html` which implements a few [11, 12, 13].

11 Minds and Information

Information theory is fundamentally about signals, not the meaning they carry. What we measure thus requires interpretation; that I am uncertain about different options may not be as important

as the fact that I'm uncertain about these particular two possibilities. Information theory tells you what you know, but it doesn't tell you what matters; you need something like a utility function from the decision theorists or game theories to tell you about that. Even though it is about signals and signal transmission, it can't tell you how those signals emerged; you'll need to supplement a story about agents exchanging information with, for example, a Brian Skyrms-like story about the emergence of signaling systems [14].

Conversely, in its universality, information theory applies just as much to the written and spoken words of humans as to the electronic machines for which it was first developed. And it allows us to compare distant worlds—no more, and no less, exciting than, say, comparing the real income of an English bricklayer in 1350 to one in 1780, the hours worked by a French housewife in 1810 and 1950, or the life expectancy of a hunter-gatherer of the Maasai to that of a child in a Manchester factory of 1840.

That we can *quantify* information is both intriguing and mysterious. Intriguing, because information is one of the fundamental features of our minds and our social worlds. History, psychology, economics, cognitive science, economics—all would grind to a halt were their practitioners forbidden from using the concept of information at will. Mysterious, because information is a fundamentally epistemic property: it is about what one knows, and is, as such, relative to that observer in a way that one's (real or nominal) salary, height, daily caloric intake, or place and date of birth are not.

Subject-relative facts, of course, abound—facts about trust, say, or allegiance, virtue, belief, love—and they make up a core part of the worlds we want to understand. What we learned in the twentieth century is that at least one such fact, the information one has, *can* be quantified. The economists have tried to quantify another set of subject-relative facts, one's desires, through utility theory, with somewhat less empirical success. Facts on the edge between reality and perception include those concerning inequality, and we have made a great deal of progress in figuring out both how to measure inequality, and what its implications are [15].

Many years ago, Shannon—the originator of information theory—wrote a somewhat dispirited article titled "the Bandwagon" [16] that worried about the indiscriminate use of information theory in other fields. In the modern era, it appears that Shannon's objections have, by and large, been answered.

In particular, the idealized nature of information theoretic concepts is now well understood. We realize that an information theoretic quantity provides a limit—sometimes, but not always reached—on how well a real system can perform.

Meanwhile, progress towards testing the extent to which information theoretic quantities do, and do not, apply to real cognitive and social phenomena, though at first slow in coming, is now beginning to build. New, often very large, data-sets give us the sample sizes we need to test sophisticated hypotheses about what underlying mechanisms might be tracked by quantities such as KL, JSD, and Mutual Information. Meanwhile, the Bayesian turn in cognitive science [17, 18] has found new uses for the optimal descriptions of reasoning provided by probability theory and Bayes' rule, providing new theoretical backing for the use of information theory to quantify the resulting distributions.

Taken together, it seems clear that, while the study of information in the social world is in its infancy, but not without some recent successes under its belt. This author has worked on a few of them, and expects to see more in the cognitive science to come.

References

[1] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

[2] Jonathon Shlens. Notes on Kullback-Leibler divergence and likelihood. *arXiv preprint*, arXiv:1404.2000, 2014. `https://arxiv.org/abs/1512.04177`.

[3] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[4] Simon DeDeo. Wrong side of the tracks: Big data and protected categories. In Cassidy R. Sugimoto, Hamid Ekbia, and Michael Mattioli, editors, *Big Data is Not a Monolith*. MIT Press, Cambridge, MA, USA, 2016. arXiv:1412.4643; `https://arxiv.org/abs/1412.4643`.

[5] Ilya Nemenman, William Bialek, and Rob de Ruyter van Steveninck. Entropy and information in neural spike trains: Progress on the sampling problem. *Physical Review E*, 69(5):056111, 2004.

[6] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306, 2009.

[7] Jaimie Murdock, Colin Allen, and Simon DeDeo. Exploration and exploitation of victorian science in darwin's reading notebooks. *arXiv preprint*, arXiv:1509.07175, 2015. `https://arxiv.org/abs/1509.07175`.

[8] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, Jan 1991.

[9] Sara Klingenstein, Tim Hitchcock, and Simon DeDeo. The civilizing process in London's Old Bailey. *Proceedings of the National Academy of Sciences*, 111(26):9419–9424, 2014.

[10] Simon DeDeo, Robert X. D. Hawkins, Sara Klingenstein, and Tim Hitchcock. Bootstrap methods for the empirical study of decision-making and information flows in social systems. *Entropy*, 15(6):2246–2276, 2013.

[11] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, Jun 2004.

[12] Greg Ver Steeg and Aram Galstyan. Information transfer in social media. *CoRR*, abs/1110.2724, 2011. `http://arxiv.org/abs/1110.2724`.

[13] Greg Ver Steeg and Aram Galstyan. Inferring predictive links in social media using content transfer. *CoRR*, abs/1208.4475, 2012. `http://arxiv.org/abs/1208.4475`.

[14] Brian Skyrms. *Signals: Evolution, learning, and information*. Oxford University Press, Oxford, UK, 2010.

[15] Thomas Piketty. *Capital in the twenty-first century*. Belknap Press, 2014. Translated by Arthur Goldhammer.

[16] C. Shannon. The bandwagon. *IRE Transactions on Information Theory*, 2(1):3–3, March 1956.

[17] Thomas L. Griffiths, Charles Kemp, and Joshua B. Tenenbaum. Bayesian models of cognition. In Ron Sun, editor, *The Cambridge Handbook of Computational Psychology*. Cambridge University Press, Cambridge, UK, 2008.

[18] Nick Chater, Mike Oaksford, Ulrike Hahn, and Evan Heit. Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(6):811–823, 2010.