Classical Information Theory

Robert B. Griffiths
Version of 12 January 2010

Contents

References:

CT = Cover and Thomas, *Elements of Information Theory*, 2d edition (Wiley, 2006)

QCQI = *Quantum Computation and Quantum Information* by Nielsen and Chuang (Cambridge, 2000), Secs. 11.1, 11.2

1 Introduction

★ Classical information theory is a well-developed subject—see CT for a very thorough presentation—which provides some of the motivation and many of the tools and concepts used in quantum information. Consequently it is useful to know something about classical information theory before studying how it needs to be modified in situations where quantum effects are important.

• These notes are intended to provide a quick survey with an emphasis on intuitive ideas. Proofs and lots of details are left to CT and the much more compact treatment in QCQI.

2 Shannon Entropy

★ Suppose we have a certain message we want to transmit from location A to location B. What resources are needed to get it from here to there? How long will it take if we have a channel with a capacity of c bits per second? If transmission introduces errors, what do we do about them? These are the sorts of questions which are addressed by information theory as developed by Shannon and his successors.

• There are, clearly, $N = 2^n$ messages which can be represented by a string of n bits; e.g., 8 messages 000, 001,... 111 for $n = 3$. Hence if we are trying to transmit one of N distinct messages it seems sensible to define the amount of information carried by a single message to be $\log_2 N$ bits, which we hereafter abbreviate to $\log N$ (in contrast to $\ln N$ for the natural logarithm).

◦ Substituting some other base for the logarithm merely means multiplying $\log_2 N$ by a constant factor, which is, in effect, changing the units in which we measure information. E.g., $\ln N$ nits in place of $\log N$ bits.

★ A key observation is that if we are in the business of *repeatedly* transmitting messages from a collection of N messages, and if the messages can be assigned a non-uniform *probability distribution*, then it is, on average, possible to use fewer than $\log N$ bits per message in order to transmit them, or to store them.

• Efficiently storing messages is known as *data compression*, and exemplified by gzip. The trick is to *encode* them in some clever way so that, for example, more common messages are represented using short strings of bits, and longer strings are used for less common messages.

★ The first of Shannon's logarithmic information measures which we will consider is called *entropy*, or *Shannon entropy*. We denote it by $H(X)$.

• Think of X as a collection of labels, each denoted by a lower-case letter x, for a set of N messages. E.g., $X = \{1, 2, 3, 4, 5, 6\}$, $x = 2$ or $x = 3$, etc. Let $p(x)$ be the probability for message x, with $\sum_x p(x) = 1$.

◦ More generally, X is a *random variable*, a numerical function on some sample space. Each x may correspond to several different points in the sample space, and $p(x)$ is the sum of the probabilities associated with these points.

★ Then define

$$H(X) = H(p) = -\sum_x p(x) \log p(x), \tag{1}$$

where log, as previously noted, will be understood as $\log_2$, so H is in bits.

◦ Using the same symbol H for $H(X)$ and $H(p)$ is convenient and will not lead to confusion.

• Since $p(x) \leq 1$, the minus sign in (1) ensures that $H(X) \geq 0$. If $p(x) = 0$, understand $p(x) \log p(x)$ as 0, which is the limit of this quantity as $p(x) > 0$ tends to 0.

□ Exercise. Show the following:
(i) $H(X) = 0$ if and only if there is some x_0 such that $p(x_0) = 1$ and $p(x) = 0$ for $x \neq x_0$.
(ii) If x can take on only d values, then $H(X) \leq \log d$, with equality achieved if and only if all the probabilities are equal, $p(x) = 1/x$.

□ Exercise. How can you express a thermodynamic entropy given in J/K, Joules per Kelvin, in units of bits? How much does the entropy of 1 cc of water increase when its temperature rises by 1°C, as measured in bits? [Hint: in statistical mechanics one writes $S = k \ln W$, where W is the number of microscopic quantum states corresponding to the thermodynamic state of interest, and k is Boltzmann's constant.]

★ An intuitive interpretation of $H(X)$ is the amount of information *on average* conveyed by an observation of x. If x is a message, then $H(X)$ is the *average missing information* about the message before it is received, and thus the average information conveyed by the message, since after a message is received (and read), the missing information (about this particular message) is 0.

◦ Comment. We are assuming, and will continue to assume, that messages are drawn repeatedly from the collection X, with the chance of drawing message x always equal to $p(x)$, no matter which messages have been drawn earlier or will be drawn later. The sequence of messages $x_1, x_2, \ldots x_n$ obtained in this way is a sequence of n *independent* and *identically distributed*, or i.i.d, random variables.

• Given that (1) represents an average, the information carried by message x must be $-\log p(x) = \log\big(1/p(x)\big)$. Thus if we use this information measure, unlikely messages carry more information than likely messages. Is that reasonable?

• It seems reasonable if "information" means information which is missing before the actual message appears. Thus suppose that your colleague pulls messages out of a hat, at random but with probability $p(x)$, and sends them to you over a classical channel. You know what $p(x)$ is, but you do not know what x is. Imagine that there are only two messages, 0 and 1, and $p(0) = 0.99$, $p(1) = 0.01$. What will the next message be? It is very likely that it will be a 0, so when a 0 arrives you are not surprised. But 1 is much less expected, so when it does come you learn more than in the case when you receive 0.

• One can also think of $H(X)$ as the difference, on average, of the information possessed by someone who knows what the actual message is, over against someone who knows the probability distribution but does not know the message.

★ If a large number M of messages, all drawn from the same collection X, are sent one after another, one expects the total information conveyed by all the messages to be $MH(X)$

• Equally, one can interpret $MH(X)$ as the *minimum number of bits required to transmit M messages* when M is large. There are, as noted previously, tricks for sending messages drawn from some collection using fewer bits, on average, than one might naively expect, as long as the probability distribution $p(x)$ is nonuniform.

• The same applies if one is not trying to transmit messages, but simply store them: $MH(X)$ is the *minimum number of bits required to store M messages* when M is large. This puts a limit on the possible amount of data compression.

◦ Making the preceding statements precise and proving them is nontrivial. See the relevant sections of CT and QCQI.

3 Two Random Variables

★ New ideas appear when we consider two random variables, X and Y, and their joint probability distribution. The concepts and formulas introduced below hold for general choices of X and Y, though for convenience we shall assume that each random variable takes on only a finite number of values (which need not be the same for Y as for X). The simplest situation is one in which $x = 0$ or 1 represents the value of a bit sent into a one-bit channel, and $y = 0$ or 1 the value which emerges from the channel. It will aid our discussion if we imagine that Alice feeds messages into the channel, and Bob, on the receiving end, reads what emerges, as in Fig. 1.

Figure 1: Channel transmitting X to Y.

• It will be convenient to employ the notation $p(x)$, $p(y)$, $p(y|x)$ and the like for $\Pr(x)$, etc. As long as the arguments are denoted by special symbols the meaning will be clear.

★ For a noisy channel of a simple sort, without memory, the output is related to the input by *conditional probabilities*: given an input x, the probability that y emerges is given by $p(y|x)$. We shall assume that the $p(y|x)$ are characteristic of the channel, determined by its physical structure, and not something which we can control, except by replacing the channel with some other channel with different characteristics.

◦ As a function of its first argument, and for a fixed value of its second argument (the condition), a conditional probability has all the properties of an ordinary probability; in particular,

$$p(y|x) \geq 0, \quad \sum_y p(y|x) = 1. \tag{2}$$

• The probability $p(x)$ that a message x is sent into the channel is determined by the ensemble of messages we are considering, and not by any property of the channel. Once $p(x)$ is given, the joint probability $p(x, y)$ that x enters the channel and y emerges, and the (marginal) probability $p(y)$ that y emerges are given by

$$p(x, y) = p(y|x)p(x), \quad p(y) = \sum_x p(x, y), \tag{3}$$

and thus depend on $p(x)$ as well as on properties of the channel.

★ Given the two random variables X and Y one can define various information entropies, of which the simplest are

$$H(X) = H(p(x)), \quad H(Y) = H(p(y)), \quad H(X, Y) = H(p(x, y)), \tag{4}$$

where each H is related to the corresponding probability distribution in the manner indicated in (1). In particular, note that

$$H(X, Y) = -\sum_{x,y} p(x, y) \log p(x, y) \tag{5}$$

involves a double sum, since the random variable denoted by X, Y consists of all pairs (x, y) of input and output values.

• The interpretation in terms of our channel model of the entropies in (4) is as follows. To begin with, $H(X)$ is determined by the ensemble of input messages. If these messages are produced automatically by a machine, then $H(X)$ is the average information which Alice lacks about what message the machine will produce next, and which she learns when the message appears and she feeds it into the channel. Likewise, $H(Y)$ is the average information which Bob is lacking about what will emerge next from the channel, and what he learns when it actually comes out. Obviously, $H(X,Y)$ is what Alice and Bob learn jointly, on average, from a particular transmission, when they compare their notes, so it is what they did not (collectively) know in advance.

4 Conditional Entropies and Mutual Information

★ Next we define the *conditional* entropies $H(Y|x)$ and $H(Y|X)$ in the following way:

$$H(Y|x) = -\sum_y p(y|x)\log p(y|x), \quad H(Y|X) = \sum_x p(x)H(Y|x). \tag{6}$$

Similarly, $H(X|y)$ and $H(X|Y)$ are defined by the analogous equations in which the roles of x and y are interchanged, with $p(y|x)$ replaced by $p(x|y) = p(x,y)/p(y)$, and $p(x)$ by $p(y)$.

◦ Note that while $p(x|y)$ is in an obvious sense a counterpart of $p(y|x)$, the latter depends only on the channel, whereas the former depends on $p(x)$ as well as properties of the channel.

• Intuitive interpretation of (6). Given that the channel is noisy, when Alice puts an x into the channel, she cannot be sure what y will emerge. Then on the average (i.e., over the different occasions when she sends the same x), her ignorance about y is given by $H(Y|x)$. Averaging this over all possible input messages x gives the overall average $H(Y|X)$ of the information which Alice lacks about outputs when she knows (of course) the inputs. To put it another way, if the channel is used a large number of times M, then $MH(Y|X)$ is the total information which Alice lacks about the outputs, and thus the minimum number of additional bits of information she would have to receive from Bob in order for her to know exactly what output occurred in each case. Of course, $H(X|y)$ and $H(X|Y)$ can be given similar interpretations with the roles of Alice and Bob interchanged.

◦ In this and in all our discussions we assume that both Alice and Bob know the joint probability distribution $p(x,y)$, and hence all the marginals and conditionals. What they don't know until they see it is what actually occurs in a particular case.

• Alternative formulas. Given the definitions in (6) and their analogs with x and y interchanged, it is easy to show, using the definitions, that

$$H(Y|X) = H(X,Y) - H(X), \quad H(X|Y) = H(X,Y) - H(Y). \tag{7}$$

□ Exercise. Go ahead and show it.

• Intuitive interpretation of (7). The H quantities can always be thought of as missing information. At the outset, before she knows what x will go through the channel, Alice's (average) ignorance about both x and y is measured quantitatively by $H(X,Y)$. When message x actually appears and she sees it, her ignorance is reduced on average by $H(X)$, so $H(X,Y) - H(X)$ is the information about the pair (x,y) that she is still lacking, and which, since she now knows x, is missing information about y. This means the first expression in (7) is consistent with the interpretation of $H(Y|X)$ given previously.

◦ Mnemonic. Think of $H(Y|X) = H(X,Y) - H(X)$ as analogous to the formal expression $\Pr(Y|X) = \Pr(X,Y)/\Pr(X)$, which you already know. Division in the latter has become a minus sign in the former, as expected given that H is a *logarithmic* measure.

★ The *mutual information* $I(X{:}Y)$ is the average amount of information which Alice, knowing x, has about the output y resulting from this x. It is not the same as $H(Y|X)$, which measures the information that Alice still lacks when she knows x. Instead, $I(X{:}Y) = H(Y) - H(Y|X)$ is Alice's (average) ignorance about

y before knowing x, minus her (lesser) ignorance about y when she knows x, and therefore the (positive) amount that she learns about y (on average) from observing x.

• With the help of (7), one can show that the three formulas for $I(X{:}Y)$,

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

are equivalent, so any one of them could serve as the definition. The first two expressions have a fairly simple intuitive interpretation, see above. The last expression tells us that mutual information is symmetrical: $I(X{:}Y) = I(Y{:}X)$. That is, the average amount which Bob learns about x by observing y is the same as the average amount which Alice knows about y when she sends x. This is an important symmetry, one which is not intuitively obvious.

• Mnemonic. The following diagram, Fig. 2, is helpful for relating the different entropies and the mutual information for two random variables.

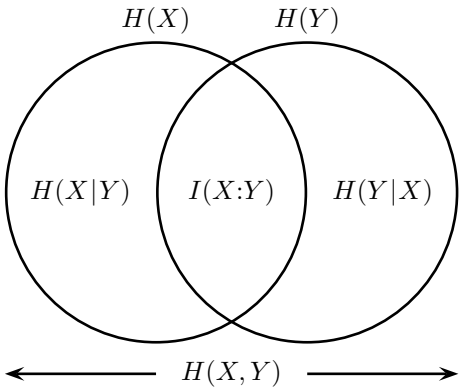

Figure 2: Diagram showing relationship of various entropies, mutual information

□ Exercise. Explain in words how to use this diagram to recall the first and last expressions for $I(X{:}Y)$ in (8)

• One can also express $I(X{:}Y)$ directly in terms of probabilities as

$$I(X{:}Y) = -\sum_{x,y} p(x,y) \log \frac{p(x)p(y)}{p(x,y)} \tag{9}$$

★ Basically, $I(X{:}Y)$ is a measure of *correlation* in the sense of *statistical dependence*. Recall that X and Y are *statistically independent* if $p(x,y) = p(x)p(y)$. Intuitively speaking, knowledge of one of them tells you (in a statistical sense) nothing about the other. In this case, and only in this case, $I(X{:}Y) = 0$. Otherwise, $I(X{:}Y) > 0$, with the interpretation given previously: a quantitative measure of how much you learn about Y when you observe X. Note that this is "on the average", so it only has operational significance if a given experiment is repeated a number of times.

□ Exercise. Show that $I(X{:}Y)$ has the following properties:
i) $I(X{:}Y) \geq 0$
ii) $I(X{:}Y) = 0$ if and only if X and Y are statistically independent.

◦ In cases in which $p(x,y)$ is not known ahead of time, one can imagine first carrying out a number of experiments to determine (approximately, at least) what it is—such is the strategy of opinion polls. Then one can use this to figure out how much the mutual information is when further experiments are carried out. In the case of a channel, one will want to first test it to find experimental values for $p(y|x)$, and then employ these to figure out how to make good use of it.

★ Although we have motivated the discussion of $I(X{:}Y)$ with reference to the transmission of information through a channel, it is worth noting that one can define the same measure for *any* situation in which two random variables X and Y have a joint probability distribution $p(x, y)$, since from this one can deduce the marginals $p(x)$ and $p(y)$, and then compute $H(X, Y)$, $H(X)$, $H(Y)$, and $I(X{:}Y)$ from (8), or directly from (9).

5 Channel Capacity

• Given that $I(X{:}Y)$ is the average amount of information about x that Bob gets when he observes y, it is not surprising that $I(X{:}Y)$ can be identified with the *average rate* at which information is being transmitted through the channel, so that if the channel is used M times, where M is large, the information passing through it is (approximately) $MI(X{:}Y)$ bits.

★ Channel capacity I. Note that $I(X{:}Y)$ as defined in (8) depends not only on the conditional probabilities $p(y|x)$ that define the channel, but also on the $p(x)$ associated with the source of messages. The maximum of $I(X{:}Y)$ over all possible probability distributions consistent with *fixed* conditional probabilities $p(y|x)$ is called the *channel capacity*,

$$C = \max_{\{p(x)\}} I(X{:}Y). \tag{10}$$

It is a function of the $p(y|x)$ alone, and therefore is something characteristic of the channel. CT, p. 184, refers to this as the "information channel capacity."

★ Channel capacity II. What CT call the "operational channel capacity" is defined in the following way. In order to transmit information in a reliable way through a noisy channel it is necessary to *encode* it in a redundant way, e.g., instead of sending simply 0 or 1 through the channel send 000 and 111, three 0's in a row or three 1's. The resulting *error correcting code* allows the receiver by means of a suitable *decoding* procedure to reduce the probability of error, but at the cost of using the channel more times than would be necessary in the absence of noise. By using appropriate encoding and decoding procedures the probability that a message of a given length can be transmitted without error (after decoding has been applied) can be made very close to 1. The ratio of the amount of reliable information transmitted to the number of uses of the channel, assuming the error correction procedure is optimized as this number tends to infinity, is the operational channel capacity. For a precise definition with epsilons in the proper places see CT Ch. 7.

• The good news is the Channel Coding Theorem due to Shannon, which shows that the information and the operational channel capacities are identical. For the proof see CT Ch. 7.

• To summarize, the capacity C of a channel is the maximum possible rate at which information can be reliably (using appropriate error correction) transmitted through a noisy channel, measured in bits of information (logarithms to base 2) per uses of the channel. Formula (10) allows one to calculate it, though generally not in closed form, for the simplest model of a noisy channel in which the conditional probability $p(y|x)$ is the same every time the channel is used, and does not depend upon what was previously sent through the channel. (This is called a "memoryless channel.")