On the use of analogies in a first course on information theory

Marc A. Armand

National University of Singapore
Singapore

ABSTRACT: Information theory is a subject that is both mathematically intense and abstract. Therefore, it is of no surprise that students reading a first course in information theory at the undergraduate level find this subject particularly difficult. The standard approach to teaching this subject follows the mathematically formal theorem-proof format. Although rigorous, this approach offers little guidance in developing an intuitive feel for the subject. In this article, the author presents some ideas on how simple analogies may be used to help facilitate the development of an intuitive feel for information theory. By invoking familiar analogies to illustrate abstract theorems as part of the theorem-proof approach, students develop an appreciation for the physical meaning behind the abstraction. When used as a supplement to the theorem-proof approach, rather than as a substitute, such a learning outcome may be achieved without sacrificing the mathematical rigour of the subject.

INTRODUCTION

The use of familiar scenes in everyday life as analogous examples to illustrate abstract mathematical theorems is a useful tool in facilitating students developing an intuitive feel for the subject at hand. These ideas are particularly useful for teaching information theory at the undergraduate level, where students generally lack the ability to decipher the physical meaning of equations, definitions, theorems, etc, on their own. Although presented within the context of teaching the subject on information theory, the ideas put forth in this article are general and may be applied to other subjects as well.

A review of some basic concepts on information theory are initially given in order to make this article as self-contained as possible. This is followed with a detailed description on how analogies may be used to illustrate these concepts in the context of transmission through a discrete memoryless channel. The discussion then focuses on the use of analogies to illustrate a fundamental theorem in information theory, namely: the noisy coding theorem. In contrast to the analogies used, concerning the discrete memoryless channel, the analogies used to illustrate that theorem do not capture all of the pertinent physical interpretations that may be derived from the corresponding mathematical relations. Despite this, students generally still find the use of such analogies helpful in developing an intuition for the subject, as was observed in a recent survey conducted with an undergraduate information theory class for which such analogies were heavily used to explain abstract concepts. The results of this survey are summarised at the end of this article.

PRELIMINARIES

Consider the discrete random variable X with possible values x_1, x_2, …. The information content provided by the event X=x_i having probability of occurrence $P(X=x_i)$ is defined as:

$$I(x_i) = -\log P(X = x_i) \qquad (1)$$

This is known as the *self-information* of the event X=x_i. The *average* self-information of X is then the expected value of $I(x_i)$, ie:

$$H(X) = \sum_i I(x_i)P(X = x_i) \qquad (2)$$

This quantity is known as the *entropy* of X. In particular, if the output of a source can be characterised by X, then the average amount of information contained in an output symbol, or more formally, the entropy of the source, is $H(X)$. Such a source is known as a *discrete memoryless source*.

Consider another discrete random variable Y with possible values y_1, y_2, …. The information content provided by the event X=x_i after having observed the event Y=y_j is defined as:

$$I(x_i \mid y_j) = -\log P(X = x_i \mid Y = y_j) \qquad (3)$$

This quantity is known as the *conditional* self-information of the event X=x_i, given that the event Y=y_j has occurred. The *average* conditional self-information of X, given Y, is then the expected value of $I(x_i|y_j)$, ie:

$$H(X \mid Y) \\ = \sum_i \sum_j P(X = x_i, Y = y_j) I(x_i \mid y_j) \qquad (4)$$

This quantity is known as the *conditional entropy* of X, given Y.

The information content provided by the occurrence of the event Y=y_j about the event X=x_i is defined as follows:

$$I(x_i; y_j) \\ = \log P(X = x_i \mid Y = y_j) / P(X = x_i) \qquad (5)$$

This quantity is known as the *mutual information* of the events $X=x_i$ and $Y=y_j$. The *average* mutual information of X and Y is then the expected value of $I(x_i;y_i)$, ie:

$$I(X;Y) \\ = \sum_i \sum_j P(X = x_i, Y = y_j) I(x_i; y_j) \qquad (6)$$

It can be shown that:

$$I(X;Y) = H(X) - H(X \mid Y) \geq 0 \, .$$

Thus, if $H(X|Y)$ is zero, then $I(X;Y)=H(X)$. However, if:

$$0 < H(X \mid Y) \leq H(X) \, ,$$

then

$$0 \leq I(X;Y) < H(X) \, .$$

In particular, $I(X;Y) = 0$ when $H(X|Y) = H(X)$. A more detailed treatment can be found in ref. [1].

THE DISCRETE MEMORYLESS CHANNEL

A discrete memoryless channel is a channel wherein the current output symbol depends only on the current input symbol, the input and output symbols being characterised as discrete random variables. If X and Y are used to model the input and output, respectively, of such a channel, then $H(X|Y)$ can be interpreted as the average amount of information about X that does not get through the channel. That is, $H(X|Y)$ represents the average information loss about X. On the one hand, $I(X;Y)$ represents the average amount of information about X that gets through the channel. Thus, the channel is *noiseless* when $H(X|Y)=0$, and *noisy* when:

$$0 < H(X \mid Y) \leq H(X) \, .$$

In particular, the channel is *useless* when $H(X|Y)=H(X)$.

These are typical statements found in textbooks on information theory. Analogies are now brought in. First, the above statements are summarised pictorially in Figures 1a to 1c.

An analogy for the channel is a water pipe. The quantity $H(X)$ is then the amount of water put through the pipe, $I(X;Y)$ is the amount of water that goes out the other end of the pipe, and so $H(X|Y)$ is the amount of water lost due to water leaking through perforations in the pipe, if any. A pipe with no perforations is analogous to the noiseless channel, while a pipe with perforations is analogous to a noisy one. In particular, a useless channel is analogous to a pipe where perforations are so severe that all the water that enters the pipe leaks out before reaching the other end. These analogies may further be used to give interesting intuitive justifications for relations, such as:

$$0 \leq H(X \mid Y) \leq H(X) \qquad (7)$$

and

$$0 \leq I(X;Y) \leq H(X) \qquad (8)$$

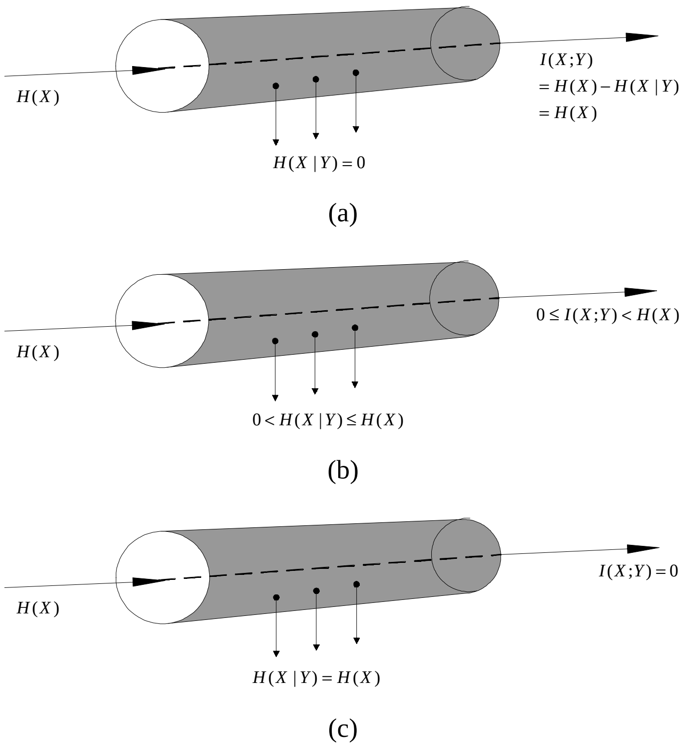


Figure 1: a) Noiseless channel (top); b): Noisy channel (middle); and c): Useless channel.

An intuitive argument for the validity of the upper-bound in the first of the two relations above is that more water cannot be lost than what is put into the water pipe in the first place. Similarly, one cannot get out from one end of the pipe more water than what one puts in at the other end, thus giving an intuitive explanation for the upper-bound in the second relation. The lower-bounds in the above two relations are even more intuitive and so any further elaboration can be omitted.

Undeniably, it is not always easy to find good analogies. In some cases, even the best analogy does not paint a complete picture. Nevertheless, even in such circumstances, an analogy can still shed intuitive light on the subject matter at hand. This is illustrated in the following section.

THE NOISY CODING THEOREM

Let the output of a discrete memoryless source be modelled by X. Suppose the source outputs a symbol every T_s seconds. Then the *average information rate* of the source is $H(X)/T_s$. Recall that if X and Y characterise, respectively, the input to and output from a discrete memoryless channel, then $I(X;Y)$ is the average amount of information about X that gets through the channel. The maximum value of $I(X;Y)$, denoted by C, is taken over all of the probability density functions of X. Assuming that the channel is used every T_c seconds, the *channel capacity per unit time* is the quantity C/T_c.

Suppose that the source symbols are encoded using a code A_n of length n prior to transmission. Reliable communications can then be achieved if the probability that the decoder at the receiver makes an erroneous decision on what was transmitted, $p_e(n)$, can be made arbitrarily small.

The noisy coding theorem states that if:

$$H(X)/T_s \leq C/T_c \, , \qquad (9)$$

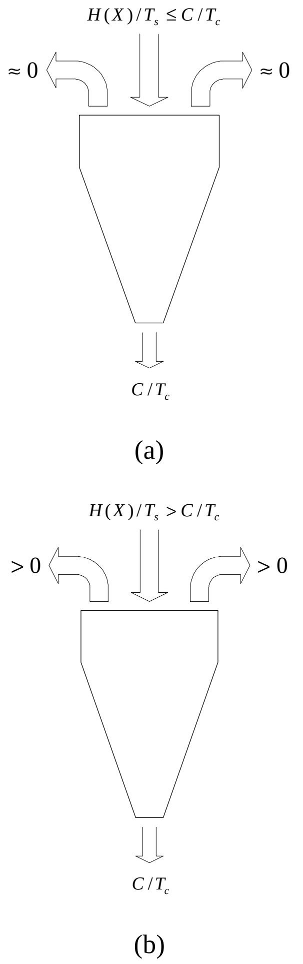
then, there exists a sequence of codes A_n and corresponding decoding schemes with an associated probability of decoder error $p_e(n)$, such that $p_e(n)$ tends to zero as n tends to infinity. In other words, the loss of information about X tends to zero as n tends to infinity.

On the other hand, if:

$$H(X)/T_s > C/T_c,$$

then $p_e(n)$ is bounded away from zero for all n. To put it differently, the loss of information about X is always greater than zero for all n.

Analogously, $H(X)/T_s$ could be considered as being the rate that one pours water down a funnel, and C/T_c denotes the maximum rate that it can be done, beyond which water starts to overflow from the top of the funnel, leading to a loss of water. This is depicted in Figures 2a and 2b.



(a)

(b)

Figure 2: a) Average information rate less than, or equal to, channel capacity per unit time (top); and b): Average information rate exceeding channel capacity per unit time (bottom).

It is obvious that this water funnel analogy is somewhat lacking as it does not take n into account. Nevertheless, it still captures the essence of the theorem at hand (ie if the average information rate does not exceed the channel capacity per unit time, then an arbitrarily small loss of information about what is being transmitted is achievable; otherwise, this loss will forever be bounded away from zero). For this reason, it is still useful in helping students attain an intuitive feel for this fundamental result.

Understandably, finding a suitable analogy to explain a concept may not always be easy. So when any analogy seems illusive, what next? This is a question that is addressed in the next section.

INTUITIVE EXPLANATIONS WITHOUT ANALOGIES

When a suitable analogy to explain a concept or theorem does not seem to exist, one could, with a bit of imagination, still come up with intuitive explanations. This is illustrated by invoking the noiseless source coding theorem.

Suppose that the output of a discrete memoryless source, characterised by X, is encoded symbol-by-symbol, using a variable-length code A. Variable length code denotes an ensemble of variable-length vectors over some fixed alphabet, while symbol-by-symbol encoding means that by some functional mapping, indicated by, say, f, each distinct source symbol is mapped to a distinct element, ie codeword, of A. Furthermore, suppose that A is *uniquely decodable*. That is, by f^{-1}, each distinct string of codewords is inversely mapped to a distinct string of source symbols. The abovementioned theorem states that there exists a uniquely decodable code, whose average codeword length L satisfies:

$$H(X) \le L < H(X) + 1 \qquad (10)$$

and that L is minimal among the average codeword lengths of all other uniquely decodable codes. An immediate consequence of this result is that there does not exist a uniquely decodable code of average codeword length less than the entropy of the source.

How can this last statement be explained intuitively? One way is to argue that if a given source output contains x symbols of information, then at least x symbols are needed (over the same alphabet to represent that output).

For example, consider the phrase *United States of America* as a source output. That output contains 24 letters (if the inter-word spaces are counted as well). Clearly, there are many redundant letters in that output, since only the three letters, *USA*, are needed to denote the United States of America. In fact, only two are needed, ie *US*. If each letter can be represented by one symbol of some fixed alphabet, then of the 24 corresponding symbols that can be used to represent that phrase, 22 are redundant, and so it can be said that the phrase only contains two symbols of information. Accordingly, only two symbols are required (from the same alphabet), *but no less*, in order to represent the source output – one symbol for the letter *U*, the other for *S*. One symbol alone can only capture *U* or *S*, but not both, and there is no way of telling that *U* or *S* alone actually refers to the United States of America.

STUDENT FEEDBACK

Thus far, it has been illustrated how analogies and intuitive explanations may be used to facilitate learning in an information theory class. It is natural to ask at this juncture how effective this method is. So as to get some feel for it, students in an undergraduate information theory class were asked to complete a feedback form that contained the following three statements:

- *Statement* 1: I find the need to develop an intuitive feel for the subject necessary, as understanding the subject from a mathematical point of view alone is insufficient.
- *Statement* 2: I find that the analogies and intuitive explanations offered in class to illustrate difficult concepts helped me a lot in developing an intuitive feel for the subject.
- *Statement* 3: I prefer more time be spent constructing proofs to theorems as opposed to analogies or intuitive explanations.

Students were asked to respond based on a scale of 1 to 5, where

- 1 means that they strongly disagree;
- 2 means that they somewhat disagree;
- 3 means that they neither agree nor disagree;
- 4 means that they somewhat agree;
- 5 means that they strongly agree.

The responses of the 108 students who participated in this survey are summarised in Figures 3 to 5.

As can be seen from the three charts, the vast majority of students who participated in the survey value having an intuitive feel for the subject at hand.

Furthermore, the large majority of these students, although to a lesser degree, are of the opinion that the use of analogies and intuitive explanations help them greatly in developing their intuition for the subject. In fact, the majority are not in favour of spending more time on formulating rigorous mathematical proofs and less on analogies and intuitive explanations.

It should be further highlighted that some of the analogies presented in this class did not capture the entire picture painted by the corresponding mathematical relations, such as the water funnel example described above. As such, it is interesting to note that the majority of students in that class still felt that the use of analogies facilitated the development of an intuition for the subject. This supports the author’s claim that analogies that do not capture all of the physical interpretations of the corresponding mathematical relations, but rather shed intuitive light on the concepts at hand.

CLOSING REMARKS

In this article, the author has discussed the use of analogies with the objective of helping students develop an intuitive feel for the subject being studied. Based on the results of the survey conducted, students find this method effective in achieving such a learning outcome.

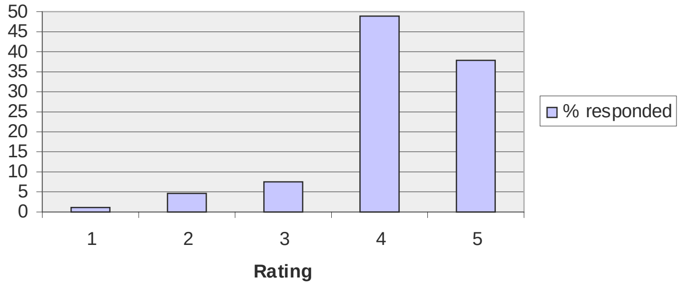


Figure 3: Responses to Statement 1.

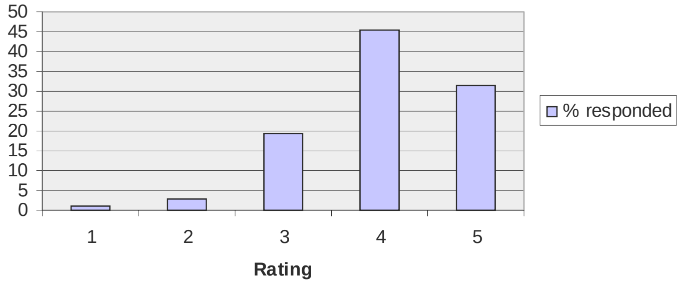


Figure 4: Responses to Statement 2.

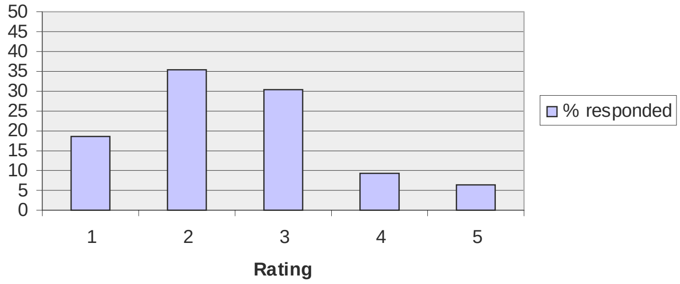


Figure 5: Responses to Statement 3.

However, it is important to note that the mathematical rigour of the subject need not be compromised, so long as this technique is employed to supplement, rather than replace, the standard theorem-proof teaching approach.

The proposed method of teaching is particularly well suited for a course in information theory offered as a cross-faculty module; as such, a class would typically contain students of diverse academic background, eg students from the arts or science faculties, who may struggle with the abstract and mathematical nature of the subject, and hence benefit from the use of more familiar analogies.

REFERENCES

1. Proakis, J.G., *Digital Communications*. New York: McGraw-Hill (2001).