(Preprint of paper to appear in *Proc. 1990 Intl. Symp. on Info. Th. and its Applications*, Waikiki, Hawaii, Nov. 27-30, 1990.)

CAUSALITY, FEEDBACK AND DIRECTED INFORMATION

James L. Massey

Institute for Signal and Information Processing
Swiss Federal Institute of Technology
CH-8092 Zürich, Switzerland

ABSTRACT

It is shown that the "usual definition" of a discrete memoryless channel (DMC) in fact prohibits the use of feedback. The difficulty stems from the confusion of causality and statistical dependence. An adequate definition of a DMC is given, as well as a definition of using a channel without feedback. A definition, closely based on an old idea of Marko, is given for the directed information flowing from one sequence to another. This directed information is used to give a simple proof of the well-known fact that the use of feedback cannot increase the capacity of a DMC. It is shown that, when feedback is present, directed information is a more useful quantity than the traditional mutual information.

INTRODUCTION

Information theory has enjoyed little success in dealing with systems that incorporate feedback. Perhaps it was for this reason that C.E. Shannon chose feedback as the subject of the first Shannon Lecture, which he delivered at the 1973 IEEE International Symposium on Information Theory in Ashkelon, Israel. Unfortunately, few researchers in information theory have followed Shannon's lead. As we shall see, information-theoretic definitions more often than not tacitly assume that no feedback is present (usually without the awareness of the definer). To deal with feedback requires system definitions much more careful than those that we are accustomed to use in information theory. In the sequel, we try to provide some such careful definitions and to investigate their consequences. We shall make important use of an idea introduced almost two decades ago by Marko (Ref. 1), who is one of those rarities among workers in information theory who has understood the importance of bringing feedback explicitly into the theory.

For definitions of the standard information-theoretic quantities used in this paper, we refer the reader to the book of Gallager (Ref. 2), whose notation we will also follow as closely as possible. In particular, X^n will denote the n-tuple $[X_1, X_2, ..., X_n]$, whose components are discrete random variables, and $x^n = [x_1, x_2, ..., x_n]$ will denote a possible value of X^n. The probability distribution for X^n evaluated at x^n will be written simply as $P(x^n)$.

DISCRETE CHANNELS

A discrete channel with finite input alphabet A and finite output alphabet B is the specification of the conditional probability $P(y_n \mid x^n y^{n-1})$ for all $n \geq 1$, all $x^n \in A^n$ and all $y^n \in B^n$. The discrete channel is memoryless if this conditional probability satisfies

$$P(y_n \mid x^n y^{n-1}) = P(y_n \mid x_n). \tag{1}$$

The reader will note that (1) is not the conventional definition of a discrete memoryless channel (DMC), but we will argue that the usual definition is indefensible.

We will say that a discrete channel is used without feedback if

$$P(x_n \mid x^{n-1} y^{n-1}) = P(x_n \mid x^{n-1}) \tag{2}$$

holds for all $n \geq 1$, all $x^n \in A^n$ and all $y^{n-1} \in B^{n-1}$, i.e., when the choice of the next channel input digit, given all previous input digits, is not further related to the previous channel output digits.

The multiplication rule for probabilities gives for the joint probability of a channel input sequence and output sequence

$$P(x^N y^N) = \prod_{n=1}^{N} P(x_n y_n \mid x^{n-1} y^{n-1})$$

$$= \prod_{n=1}^{N} P(x_n \mid x^{n-1} y^{n-1})\, P(y_n \mid x^n y^{n-1}) \tag{3}$$

for all $N \geq 1$. If the channel is used without feedback, (2) can be used in (3) to give

$$P(x^N y^N) = P(x^N) \prod_{n=1}^{N} P(y_n \mid x^n y^{n-1})$$

or, equivalently,

$$P(y^N \mid x^N) = \prod_{n=1}^{N} P(y_n \mid x^n y^{n-1}) \tag{4}$$

for any channel input sequence x^N with $P(x^N) \neq 0$. If the channel is also memoryless, then (1) can be used in (4) to give

$$P(y^N \mid x^N) = \prod_{n=1}^{N} P(y_n \mid x_n), \qquad (5)$$

which the reader will recognize as the "usual definition" of the DMC. In fact, we now see that (5) can be taken as at most the "definition of a DMC used without feedback". If one uses a DMC with noiseless feedback in the manner that the next input digit is chosen as the previous output digit, i.e., so that $x_n = y_{n-1}$ for all $n > 1$, then $P(y^N \mid x^N) = 0$ when $x_n \neq y_{n-1}$ holds for some n satisfying $1 < n \leq N$ so that (5) will not be satisfied.

It is hardly a wonder that information theory has had problems dealing with feedback when our usual definition of our most basic channel, the DMC, explicitly precludes the use of feedback. It is interesting to note that Ash (Ref. 3) in his generally excellent book "correctly" defines a DMC according to (1) but contrives to convert his definition to the "incorrect" definition (5) of a DMC by invoking the relation

$$P(y_n \mid x^N y^{n-1}) = P(y_n \mid x^n y^{n-1}) \qquad (6)$$

for all $1 \leq n \leq N$ that he attributes to "causality", i.e., to the fact that the value of the channel output at time n should not depend on the future inputs x_{n+1}, ..., x_N. What Ash actually has done, however, is to rule out feedback since, via feedback, the value of y_n could indeed influence x_{n+1}, ..., x_N. The lesson to be learned here is that probabilistic dependence is quite distinct from causal dependence. Whether X causes Y or Y causes X, the random variables X and Y will be statistically dependent. Indeed, this phenomenon lies at the heart of the "mutuality" of mutual information: $I(X;Y) = I(Y;X)$. Statistical dependence, unlike causality, has no inherent directivity.

DIRECTED INFORMATION

We now attempt to give a meaningful notion of directivity to the information flow through a channel. Our basic definition is a slight modification of that introduced by Marko (Ref. 1) seventeen years ago. The directed information $I(X^N \rightarrow Y^N)$ from a sequence X^N to a sequence Y^N will be defined by

$$I(X^N \rightarrow Y^N) = \sum_{n=1}^{N} I(X^n; Y_n \mid Y^{n-1}). \qquad (7)$$

The reader can easily satisfy himself that $I(X^N \rightarrow Y^N) \neq I(Y_N \rightarrow X_N)$ in general. The usefulness of directed information is indicated in the following theorems.

Theorem 1: If X^N and Y^N are the input and output sequence, respectively, of a discrete channel, then $I(X^N \rightarrow Y^N) \leq I(X^N; Y^N)$ with equality if the channel is used without feedback.

Proof: Replacing X^n by X^N on the right in (7) can only increase the sum, which is then $I(X^N;Y^N)$ so the claimed inequality is established. If the discrete channel is used without feedback, (4) gives

$$H(Y^N \mid X^N) = \sum_{n=1}^{N} H(Y_n \mid X^n Y^{n-1})$$

and hence

$$\begin{aligned} I(X^N;Y^N) &= H(Y^N) - H(Y^N \mid X^N) \\ &= \sum_{n=1}^{N} [H(Y_n \mid Y^{n-1}) - H(Y_n \mid X^n Y^{n-1})] \\ &= \sum_{n=1}^{N} I(X^n;Y_n \mid Y^{n-1}) \\ &= I(X^N \rightarrow Y^N) \end{aligned}$$

as was to be shown.

Theorem 2: If X^N and Y^N are the input and output sequences, respectively, of a DMC, then

$$I(X^N \rightarrow Y^N) \leq \sum_{n=1}^{N} I(X_n;Y_n)$$

with equality if and only if $Y_1, Y_2, \ldots, Y_n$ are statistically independent.

Proof: $I(X^n;Y_n \mid Y^{n-1}) = H(Y_n \mid Y^{n-1}) - H(Y_n \mid X^n Y^{n-1})$

$$= H(Y_n \mid Y^{n-1}) - H(Y_n \mid X_n)$$

where we have made use of the definition (1) of a DMC. Using this relation in (7) gives

$$I(X^N \rightarrow Y^N) = H(Y^N) - \sum_{n=1}^{N} H(Y_n \mid X_n)$$

from which the theorem now follows trivially.

CAUSAL SYSTEMS

We next venture a definition of causality for a discrete-time system composed of sources, discrete channels, and various encoding and decoding devices. We will say such a system is causal if, for every source output sequence u^k and every input sequence x^n and corresponding output sequence y^n for some channel,

$$P(y_n \mid x^n y^{n-1} u^k) = P(y_n \mid x^n y^{n-1}). \qquad (8)$$

The idea of this definition is that source output sequences should be thought of as specified prior to the process of sending sequences over channels and the channel should be aware of such sequences only via its past inputs and outputs and its current input.

Theorem 3: If X^N and Y^N are the input and output sequences, respectively, of a discrete channel in a causal system and U^K is a source output sequence, then $I(U^K;Y^N) \leq I(X^N \rightarrow Y^N)$.

Proof:
$$H(Y^N \mid U^K) = \sum_{n=1}^{N} H(Y_n \mid Y^{n-1} U^K)$$

$$\geq \sum_{n=1}^{N} H(Y_n \mid X^n Y^{n-1} U^K) =$$

$$= \sum_{n=1}^{N} H(Y_n \mid X^n Y^{n-1})$$

from which the theorem follows directly.

Note that Theorems 2 and 3 directly imply the well-known fact that (in a causal system) the capacity of a DMC is not increased by the use of feedback. It is more interesting to note that this "well-known result" is logically meaningless if one adopts the "usual definition" (5) of a DMC, since this "definition" does not permit feedback to be used!

From Theorems 1 and 3, one sees in general that when feedback is present the directed information $I(X^N \rightarrow Y^N)$ gives a better upper bound on the information that the channel output sequence Y^N gives about the source sequence U^K than does the conventional mutual information $I(X^N;Y^N)$. This means that, when feedback is present, we will obtain stronger results if we work with directed mutual information rather than with mutual information. Whether directed information suffices as a basis for a genuine information theory of feedback systems remains, however, to be seen.

REMARKS

We first pointed out the inadequacy of the "usual definition" of a DMC to deal with feedback in our keynote lecture given at the International Conference on Information Theory and Systems, Berlin, September 18-20, 1978. The other results in this paper were presented orally at the Information Theory Meeting held at the Mathematical Research Center, Oberwolfach, Germany, May 14-19, 1989.

REFERENCES

l. H. Marko, The Bidirectional Communication Theory - A Generalization of Information Theory, IEEE Trans. Comm., vol. COM-21, pp. 1345-1351, December 1973.

2. R.G. Gallager, Information Theory and Reliable Communication. New York: Wiley 1968.

3. R. Ash, Information Theory. New York: Wiley Interscience, 1965.