Some Rudiments of Information Theory

James L. Massey

Prof.-em. ETH Zurich
Vester Søgade 14, 5TH
DK-1601 Copenhagen V

jamesmassey@compuserve.com
1928 R.V. L. Hartley

To be able to access any one of one million telephones, the telephone numbers must have at least $\log_{10}(1\,000\,000) = 6$ digits.

The information given by the value of a discrete random variable $X$ is

$$ I(X) = \log_b(\#(\text{supp}(P_X))). $$

For $b = 10$, the unit is called the Hartley.
Two different random experiments:

Hartley says $I(X_1) = I(X_2) = \log(2) = 0.300$ Hartleys.

But $P_{X_1}(1) = 1/2$ whereas $P_{X_2}(1) = 1/4$

so it certainly seems that we are getting more information from the first experiment where we cannot predict the result as well.
Claude Elwood Shannon

1962
1961

Claude Elwood Shannon
Claude Elwood Shannon

17 April
1961
Claude Elwood Shannon

Photo by Lotfi Zadeh.
Claude Elwood Shannon

In memoriam.
Shannon introduced a quantity that he called entropy and also called uncertainty, which can be interpreted as the average Hartley information. The value $x$ of $X$ can be considered as one of $1/P_X(x)$ equally likely values and thus the Hartley information for observing $x$ would be $\log (1/P_X(x)) = - \log P_X(x)$. Averaging gives

$$H(X) = - \sum_{x \in \text{supp } P_X} P_X(x) \log P_X(x)$$
When the base $b$ of the logarithm is 2, then Shannon, following a suggestion of J. W. Tukey, called the unit of information the bit.

Note that we can write

$$H(X) = E[-\log P_X(X)].$$

It immediately follows that

$$H(XY) = E[-\log P_{XY}(XY)].$$

An obvious further definition is

$$H(X|Y) = E[-\log P_{X|Y}(X|Y)].$$
Some basic, simply proved, results.

\[ 0 \leq H(X) \leq \log (\#\text{supp}(P_X)) \]
with equality on the left if and only if there is an \( x \) with \( P_X(x) = 1 \),
and with equality on the right if and only if \( P_X(x) = 1/ (\#\text{supp}(P_X)) \) for all \( x \).

\[ H(X|Y) \leq H(X) \]
with equality if and only if \( X \) and \( Y \) are statistically independent.

One more definition (due to Fano):
\[ I(X;Y) = H(X) - H(X|Y) \]
is the \textbf{mutual information} between \( X \) and \( Y \),
How was Shannon’s work greeted by the mathematics community?

J. L. Doob

Kolmogorov

B. McMillan

Inamori
Sources and Channels

A source

\[ U_1, U_2, \ldots \]

A channel

\[ Y_1, Y_2, \ldots \]

\[ X_1, X_2, \ldots \]
Here is a real binary symmetric source (BSS).

(Monkey with fair binary coin.)
What is Random Coding?

James L. Massey*
Prof.-em. ETH Zürich
Vester Søgade 14, 5TH
DK-1601 Copenhagen V
JamesMassey@compuserve.com
In his proof of the famous “noisy coding theorem”, which he called “The Fundamental Theorem for a Discrete Channel with Noise”, Shannon writes:

“The method of proving the first part of this theorem is not by exhibiting a coding method having the desired properties, but by showing that such a code must exist in a certain group of codes. In fact we will average the frequency of errors over this group and show that this average can be made less than $\varepsilon$. If the average of a set of numbers is less than $\varepsilon$ there must exist one in the set which is less than $\varepsilon$.” [1, §13]

But in fact Shannon never seems to do any averaging at all. Was he pulling our leg?
The “communications experiment”
of the
“two random experiments model”

For every code, $x_1, x_2, \ldots, x_M$ there is such a communications experiment with error probability $P_e(x_1, x_2, \ldots, x_M)$. 
The "code selection experiment" of the "two random experiments model"

\[ Q_{x_1, x_2, \ldots, x_M} (\cdot, \cdot, \ldots, \cdot) \]

The probability of selecting the code \( x_1, x_2, \ldots, x_M \) is

\[ Q_{x_1, x_2, \ldots, x_M} (x_1, x_2, \ldots, x_M) . \]
Shannon’s words, quoted on slide 2, sure sound as if he wants us to work with the “two random experiments model” and to compute the following average error probability.

\[
E[P_e(X_1, X_2, \ldots, X_M)] = \sum_{x_1, x_2, \ldots, x_M} Q_{x_1, x_2, \ldots, x_M}(x_1, x_2, \ldots, x_M)P_e(x_1, x_2, \ldots, x_M).
\]

But he doesn’t seem to do this himself!
Shannon instead computed the error probability $P$ in the following “combined random experiment” and he did no averaging whatsoever!

$$P = P[\mathcal{E}]$$ where $\mathcal{E}$ is the event that $\hat{Z} \neq Z$. 

Information Source $\rightarrow Z \rightarrow X \rightarrow Y \rightarrow \hat{Z}$

$P_Z(\cdot)$

$1 \leq Z \leq M$

$Q_{X_1, X_2, \ldots, X_M}(\cdot, \cdot, \ldots, \cdot)$
Comparing the combined experiment with the communications experiment, we see that

$$P[\mathcal{E} \mid \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M = \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M] = P_e(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M).$$

Comparing the code selection experiment with the communications experiment, we see that

$$P[\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M = \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M] = Q_{x_1 x_2 \ldots x_M}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M).$$

Thus the “theorem on total probability of an event” implies

$$P = E[P_e(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M)]$$

which is nothing short of miraculous!
Directed Information
-from where to where?

James L. Massey
Prof.-em. ETH Zurich
Vester Søgade 14, 5TH
DK-1601 Copenhagen V

jamesmasseym@compuserve.com
In the following two papers, **Hans Marko**
gave the first convincing demonstration that
**one can quantitatively measure the
information transmitted in particular
directions in a communication system.**

Kommunikation and ihre Anwendung auf die
Informationsübermittlung zwischen Menschen

Theory—A Generalization of Information Theory,”
*IEEE Trans. Commun.*, vol. COM-21, pp. 1345-1351,
This is a Figure from [2] in which Marko indicates the kind of communications scenario with which he will deal.

Here, $S, C, D, N$ and $\text{Ch}$ denote source, coder, decoder, noise, and channel, respectively.

Fig. 2. Block schematic of a bidirectional communication between $M_1$ and $M_2$. Here, $S, C, D, N$ and $\text{Ch}$ denote source, coder, decoder, noise, and channel, respectively.
Here are the results of two sets of measurements as reported in [2] when $M_1$ and $M_2$ are two monkeys.

Fig. 9. Information flow diagram for bidirectional communication as group behavior between two monkeys. (All figures given by bit/action.)
The previous slides should make it clear that Marko’s research was more in the spirit of Norbert Wiener’s cybernetics (“the science of communication and control in the animals and man”) rather than in the spirit of Claude Shannon’s information theory (“the mathematical theory of communication”).

Marko’s mode of thinking is very broad and often motivated by the consideration of communications among animals. His arguments are often philosophical and intuitive, rather than mathematical.

To appreciate Marko’s work in the light of information theory, we need to have a mathematical model consistent with bidirectional communications.
Early steps in defining and using directed information
1977/78 - while visiting that year at M.I.T., I had many discussions with Sanjoy Mitter, Demos Teneketzis, and other control theorists about the need to be able to distinguish directions of information flows. This led me to formulate basic definitions for directed information. Our goal was to formulated real-time information theory, but we never got very far with this.
1990 - I presented the paper “Causality, Feedback and Directed Information” at the ISITA meeting in Hawaii. It contained essentially the work I had done at M.I.T.
What I tried to formulate in 1990:

Channel definitions

Information measures
(undirected & directed)

Causality
What is the probability law of a Discrete Memoryless Channel (DMC)?

$$P(y^N \mid x^N) = \prod_{n=1}^{N} p_{y|x}(y_n \mid x_n), \text{ all } n \geq 1 \quad (???)$$

Does it matter what is in the gray area?
The natural and least restrictive assumption to make about the gray area is that it contains no negative delays and no closed paths through boxes for which the path delay is zero.

This is the kind of assumption made in the study of discrete-time systems.
Let $Y_0 = 0$ be the initial condition in the delay.

N.B.: $X_n = Y_{n-1}$, all $n \geq 1$

$$P(Y_1 Y_2 = 1 0 \mid X_1 X_2 = 0 0) = 0$$

But $P_{Y \mid X}(1 \mid 0)P_{Y \mid X}(0 \mid 0) = 1/4$

**What is wrong** with our definition (???) of a DMC?
Here is the "correct" probability law of a DMC!

\[ P(y_n|x^n y^{n-1}) = P_{Y|X}(y_n|x_n), \text{ all } n \geq 1 \]

This means that \( X_n \) is a sufficient statistic for reasoning about \( Y_n \) given the observation \((X^n, Y^{n-1})\).
In his book [R. Ash, *Information Theory*. New York: Wiley Interscience, 1965], Bob Ash correctly gave the probability law for the DMC but “spoiled” his definition by further requiring the "causality condition" that, for $1 \leq n \leq N$,

$$P(y_n | x^N y^{n-1}) = P(y_n | x^n y^{n-1}).$$

Adding this condition gives

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

Ash’s “causality condition” has nothing to do with causality! What it does do is to prohibit feedback!
Note that Ash's "causality condition"
\[
P(y_n | x^n y^{n-1}) = P(y_n | x^n y^{n-1}) \text{ for } 1 \leq n \leq N
\]
can equivalently be written as
\[
H(Y_n | X^N Y^{n-1}) = H(Y_n | X^n Y^{n-1}) \text{ for } 1 \leq n \leq N.
\]
It is often convenient to work with uncertainties (discrete entropies) rather than with the probability distributions themselves.

A more natural definition for prohibiting feedback is to say that a channel is used without feedback if
\[
P(x_n | x^{n-1} y^{n-1}) = P(x_n | x^{n-1}) \text{ for all } n \geq 1.
\]
To show that these two conditions for prohibiting the use of feedback are equivalent, it suffices to show that both give the same \( P(x^N y^N) \).

Using the “natural condition,” we have

\[
P(x_n y_n \mid x_{n-1}^n y_{n-1}^n) = P(x_n \mid x_{n-1}^n y_{n-1}^n) P(y_n \mid x_n^n y_{n-1}^n)
= P(x_n \mid x_{n-1}^n) P(y_n \mid x_n^n y_{n-1}^n)
\]

\[
\Rightarrow P(x^N y^N) = P(x^N) \prod_{n=1}^{N} P(y_n \mid x_n^n y_{n-1}^n).
\]

Using Ash’s condition, we have

\[
P(x^N y^N) = P(x^N) P(y^N \mid x^N) = P(x^N) \prod_{n=1}^{N} P(y_n \mid x^n y_{n-1}^n)
= P(x^N) \prod_{n=1}^{N} P(y_n \mid x^n y_{n-1}^n).
\]
Unlike causality, probabilistic dependence has no direction. Whether A causes B or B causes A, A and B will be statistically dependent.

Essentially this is why \( I(X_N; Y_N) = I(Y_N; X_N) \)
or, equivalently,
\[
H(Y_N) - H(Y_N | X_N) = H(X_N) - H(X_N | Y_N).
\]
My 1990 definition of directed information:

\[ I(X^N \rightarrow Y^N) = \sum_{n=1}^{N} I(X^n; Y_n | Y_{n-1}). \]

(By ignoring information that \( Y_n \) may be giving about future \( X \) digits, we are considering only the information flowing from \( X \) digits to \( Y \) digits.)

In terms of uncertainties (discrete entropies)

\[ I(X^N \rightarrow Y^N) = \sum_{n=1}^{N} [H(Y_n | Y_{n-1}) - H(Y_n | X^n Y_{n-1})]. \]

We recall that

\[ I(X^N; Y^N) = \sum_{n=1}^{N} [H(Y_n | Y_{n-1}) - H(Y_n | X^N Y_{n-1})]. \]
Equivalently, we can write

\[ I(X^N \rightarrow Y^N) = \sum_{n=1}^{N} [H(X^n | Y^{n-1}) - H(X^n | Y^n)]. \]

Because \( H(Y_n | X^N Y^{n-1}) \leq H(Y_n | X^n Y^{n-1}) \), it follows that

\[ I(X^N \rightarrow Y^N) \leq I(X^N ; Y^N) \]

with equality if and only if there is no feedback, i.e., if and only if (natural condition)

\( P(x_n | x^{n-1} y^{n-1}) = P(x_n | x^{n-1}) \) [or, equivalently, (Ash's condition) \( P(y_n | x^N y^{n-1}) = P(y_n | x^n y^{n-1}) \)]

for \( 1 \leq n \leq N \).
If $X^N$ and $Y^N$ are the input and output sequences of a DMC, then

$$I(X^N \rightarrow Y^N) = \sum_{n=1}^{N} [H(Y_n | Y_{n-1}) - H(Y_n | X^n Y_{n-1})]$$

$$= \sum_{n=1}^{N} [H(Y_n | Y_{n-1}) - H(Y_n | X_n)]$$

$$\leq \sum_{n=1}^{N} [H(Y_n) - H(Y_n | X_n)] =$$

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

with equality if and only if $Y_1, Y_2, \ldots, Y_N$ are statistically independent.
**Conservation Law** for Directed Information

\[ I(X^N \rightarrow Y^N) + I(0 \ast Y^{N-1} \rightarrow X^N) = I(X^N ; Y^N). \]

Proof by induction:

I(\(X^1 \rightarrow Y^1\)) = I(\(X_1 ; Y_1\)) and I(\(0 \rightarrow X^1\)) = 0.

Note that

I(\(X^{n+1} \rightarrow Y^{n+1}\)) = I(\(X^n \rightarrow Y^n\)) + I(\(X^{n+1} ; Y_{n+1} | Y^n\))

and similarly that

I(\(0 \ast Y^n \rightarrow X^{n+1}\)) = I(\(0 \ast Y^{n-1} \rightarrow X^n\)) + I(\(0 \ast Y^n ; X_{n+1} | X^n\))

= I(\(0 \ast Y^{n-1} \rightarrow X^n\)) + I(\(Y^n ; X_{n+1} | X^n\)).

Thus, by the induction hypothesis,

I(\(X^{n+1} \rightarrow Y^{n+1}\)) + I(\(0 \ast Y^n \rightarrow X^{n+1}\)) = I(\(X^n ; Y^n\)) + I(\(X^{n+1} ; Y_{n+1} | Y^n\))

+ I(\(Y^n ; X_{n+1} | X^n\))

= I(\(X^{n+1} ; Y^n\)) + I(\(X^{n+1} ; Y_{n+1} | Y^n\))

= I(\(X^{n+1} ; Y^{n+1}\)).
Gerhard Kramer's definition of causal conditioning:

\[ H(Y^n \mid X^n) = \sum_{n=1}^{N} H(Y_n \mid X_n Y_{n-1}) \]

In terms of causal conditioning, one can write

\[ I(X^n \rightarrow Y^n) = H(Y^n) - H(Y^n \mid X^n) \]

Gerhard also defined causally conditioned directed information in the following manner:

\[ I(X^n \rightarrow Y^n \mid Z^n) = H(Y^n \mid Z^n) - H(Y^n \mid X^n Z^n) \]
We want to consider a synchronized network in which all devices are governed by the same notion of time.
Three arbitrary clocked sequences

\[
\begin{align*}
X_1, X_2, X_3, X_4, X_5, \ldots, X_N \\
Y_1, Y_2, Y_3, Y_4, Y_5, \ldots, Y_N \\
Z_1, Z_2, Z_3, Z_4, Z_5, \ldots, Z_N
\end{align*}
\]

The meaning is that the \(n\)th variable in each sequence, i.e., \(X_n, Y_n, \) and \(Z_n\), take on their values at the same time instant. Moreover, the \(n\)th variable takes on its value before the \((n+1)\)st variable. Note also that \(Y_{n-1}\) is the \(n\)th variable in the concatenated sequence \(0 \ast Y^{N-1}\).

Which kinds of sequences should be clocked?
**Channels**: Input and output sequences are clocked.

**Delays**: Input and output sequences are clocked.

**Sources**: Not clocked - output present at creation!

**Channel Encoders**: Only output sequence and input feedback sequence (if present) are clocked.

**Channel Decoders**: Only input sequence and output feedback sequence (if present) are clocked.

**Source Encoders**: Not clocked - input and output present at creation!

**Source Reconstructors**: Not clocked - input and output present at end of world!
Clocked inputs to network components can come only from clocked outputs of other network components.

⇒ you cannot connect a source directly to a channel! You must use a channel encoder.

Michael Gastpar’s direct transmission of a source over a channel has to be understood as using a parallel-to-series converter as the channel encoder.
The **causality assumption** for synchronized networks:

For any channel (whether or not memoryless), the input-output sequences \((X^n, Y^{n-1})\) are a **sufficient statistic** for reasoning about the output \(Y_n\) given the observation \((X^n, Y^{n-1}, W)\), where \(W\) is any random quantity composed of inputs and outputs of other network elements at time \(n\) or earlier, \(1 \leq n \leq N\).

Thus, for any **source output** \(U^K\),

\[
H(Y_n | X^n Y^{n-1} U^K) = H(Y_n | X^n Y^{n-1}).
\]

\[
\Rightarrow H(Y^N | U^K) = \sum_{n=1}^{N} H(Y_n | U^K Y^{n-1}) \geq \sum_{n=1}^{N} H(Y_n | U^K X^n Y^{n-1}) = \sum_{n=1}^{N} H(Y_n | X^n Y^{n-1}) = H(Y^N || X^N)
\]

Thus, \(I(U^K ; Y^N) \leq I(X^N \rightarrow Y^N)\).
An immediate consequence of the fact that
\[ I(U^K; Y^N) \leq I(X^N \rightarrow Y^N) \leq \sum_{n=1}^{N} I(X_n; Y_n) \leq N C_{DMC} \]
is that feedback does not increase the capacity of a discrete memoryless channel (DMC).

How can one logically prove this result if one uses the “usual definition” of a DMC, which is given on slide 8 and which in fact prohibits the use of feedback?

How can we reason correctly about complicated networks of sources, channel, encoders, decoders and delays without some sort of careful statement of our assumptions similar to that given here?
Some references:


Zero Error

James L. Massey

Prof. em. ETH Zürich
Vester Søgade 14. 5TH
DK-1601 Copenhagen V

JamesMassey@compuserve.com
Since its founding by Shannon in 1948, information theory has mostly dealt with the **small-error capacity** $C$ of channels.

The **small-error capacity** $C$ is the largest number such that for every given $\varepsilon > 0$ and every given $\delta > 0$, $K$ information bits from a Binary Symmetric Source (BSS) can, with the use of a block code of sufficiently large length $N$, be sent over the channel at a **rate** $R = K/N > C - \delta$ bits/channel-use and with block error probability $P_B < \varepsilon$.

BSS = Device that on demand produces digits from a binary coin-tossing sequence.

It is tacitly assumed that there is **no feedback** from the receiver.
Block Coding for a Discrete Memoryless Channel

DMC with small-error capacity $C$

Block of $N$ DMC inputs

Block of $K$ information bits

rate $R = K/N$ bits/channel-use
Here and hereafter, we consider only Discrete Memoryless Channels (DMCs).

Example: The Binary Erasure Channel (BEC)

As (almost) everybody knows,

\[ C = 1 - p \] bits/use
Peter Elias devoted his Shannon Lecture on 13 October 1978 in Ithaca, NY, to showing how the Binary Erasure Channel (BEC) incorporates all the essential complexity of a general noisy channel.
In his 1956 paper, “The zero error capacity of a noisy channel”, Shannon introduced another kind of capacity.

**Zero-error capacity** $C_o$ is the largest number such that for every given $\delta > 0$, $K$ information bits from a BSS can, with the use of a block code of length $N$, be sent over the channel at a rate $R = K/N > C_o - \delta$ bits/channel-use and with bit error probability $P_b = 0$.

N.B. When one deals with zero error, bit error probability, $P_b$, and block error probability, $P_B$, coincide--both must be zero.
First consider the special case of the BEC with $p = 0$.

(We do not, and will not, show impossible transitions.)

The output letter $\Delta$ is unreachable for this BEC.

Obviously, $C_o = C = 1$ bit/use.

It is obviously true that, for every DMC, $C_o \leq C$.

**Question #1:**
For the BEC in the nontrivial case where $p > 0$, does $C_o = 0$ or does $C_o = C = 1 - p$ bits/use?
Trivial Lemma:
If you can’t send even one information bit over the channel with zero error no matter how large the block length $N$, then $C_o = 0$ bits/use.

To send one information bit, we need a code with only two codewords of length $N$, e.g., 0 0 … 0 and 1 1 … 1. But no matter which of the two codewords we send, $\Delta \Delta \ldots \Delta$ will be received with nonzero probability and the decoder cannot then make a zero-error decision.

Therefore, for a BEC with $p > 0$,

$$C_o = 0 \text{ bits/use.}$$
Shannon liked ideas more than equations!

Claude Elwood Shannon (1916–2001)  
(photograph 17 April 1961 by Göran Einarsson)
Shannon called two input letters $x$ and $x'$ **non-adjacent** if there is no output letter $y$ reachable from both. I will say “**non-confusable**” instead of “non-adjacent”.

The two input letters of the **BEC** with $p > 0$ are confusable because the erasure output symbol $\Delta$ is reachable from both.

**Theorem (Shannon, 1956):**
The zero-error-capacity $C_0$ of a discrete memoryless channel is zero unless and only unless it has **two non-confusable** input letters.

N.B. If $C_0 \neq 0$, then $C_0 \geq 1$ bit/use; in fact $R = 1$ bits/use with zero error can be achieved with a block code of length $N = 1$. 
As usual, **Shannon was putting his finger on the right concept**! When one is considering zero-error capacity, the only thing that counts about a transition probability is whether it is zero or not.

This means that computing zero-error capacity is a **combinatorial problem** rather than an analytical probabilistic problem, to the joy of a few and to the unhappiness of the many.

If the maximum number of non-confusable input letters is $M$, then the zero-error capacity of the channel satisfies

$$C_o \geq \log_2 M \text{ bits/use}.$$
An example from Shannon:

0 and 2 are non-confusable inputs.  
0 and 3 are non-confusable inputs.  
1 and 3 are non-confusable inputs.  
1 and 4 are non-confusable inputs.  
2 and 4 are non-confusable inputs.  

No three input letters are non-confusable!

**Question #2:**
Does $C_o = 1$ bit/use ???
Two or more channel input sequences of the same length $N$ ($N > 0$) are non-confusable if there is no output sequence of length reachable from more than one of these input sequences.

If the maximum number of non-confusable input sequences of length $N$ is $M$, then the zero-error capacity of the channel satisfies

$$C_o \geq \frac{1}{N} \log_2 M \text{ bits/use.}$$
Shannon pointed out that if we use only the following pairs of inputs: 00, 12, 24, 31, 43. Then the channel equivalently becomes

These five inputs for the compound channel are non-confusable so we know that

\[ C_o \geq \frac{1}{2} \left( \log_2 5 \right) \approx 1.16 \text{ bits/use.} \]

**Question #3:** Does \( C_o = \frac{1}{2} \left( \log_2 5 \right) \) bits/use ???
YES !!!

In a celebrated 1979 paper* that won the Information Theory Society’s annual paper award, Lovász proved that $C_o = \frac{1}{2} \log_2 5 \approx 1.16$ bits/use for Shannon’s channel.

This was 23 years after Shannon’s paper was published!

I am interested in zero-error capacity only when there is a feedback channel available.

Shannon defined complete feedback to mean that “there exists a return channel sending back from the receiving point to the transmitting point, without error, the letters actually received. It is assumed that this information is received at the transmitting point before the next letter is transmitted, and can be used, therefore, if desired, in choosing the next transmitted letter.”

Shannon wrote $C_{OF}$ to denote the zero-error capacity with complete feedback.
Example of the BEC:

![Diagram of BEC](image)

(There is always assumed to be a slight delay in the complete feedback line.)

The encoder is using the rule: **transmit each information bit repeatedly until it is received unerased**.

This coding scheme clearly gives **zero error**! Moreover, the **rate** of transmission is $R = 1 - p$ bits/use (which is the probability of success on each transmission) so that the BEC is used $1/(1 - p)$ times on average for each info. bit.

**Question #4:**
Does $C_{OF} = 1 - p = C$ bits/use ???
Shannon says NO!

Shannon proved in 1956 that if $C_o = 0$ then $C_oF = 0$.

How can we reconcile this statement with the example of the BEC where we saw we could send with zero error at rate $C = 1 - p$ when complete feedback was available???
Here’s how Shannon defined $C_{OF}$:

The zero-error capacity $C_{OF}$ with complete feedback is the largest number such that for every given $\delta > 0$, there is an $N$ such that $K$ information bits from a Binary Symmetric Source can be sent with $N$ uses of the DMC with complete feedback at a rate $R = \frac{K}{N} > C_{OF} - \delta$ bits/channel-use and with zero error.

Shannon insisted that we use (adaptive) block coding!

If $C_O = 0$, then one cannot send one information bit with zero error over a BEC, even with an adaptive code of length $N$ and no matter how large the specified $N$ may be, because the received sequence can be $\Delta \Delta \ldots \Delta$ for both values of the information bit. The same argument applies to any channel with $C_O = 0$. 


In his 1956 paper, Shannon proved that the **small-error capacity** of a discrete memoryless channel (the only kind of channel that we have been, and will be, talking about) is not increased by the availability of complete feedback. (Shannon credits **Elias** with pointing out to him that zero-error capacity, when it is non-zero, can sometimes be increased by complete feedback.) It may have seemed natural to Shannon to assume that zero-error capacity, when it is zero, cannot be increased by the availability of complete feedback. Thus, he may not have reflected much over the generality of his restriction to “block” codes in his definition of $C_{OF}$.

There are very few places among his very many contributions to information theory that Shannon adopted an unnecessarily restrictive approach — this appears to be one instance.
Claude E. Shannon devoted his Shannon Lecture on 29 June 1973 in Ashkelon, Israel, entirely to questions of feedback.

I believe he was laying down a challenge to us lesser mortals.

Photograph from IEEE Spectrum, March 2007
The zero-error capacity $C_{oFa}$ with complete feedback is the largest number such that for every $\delta > 0$, information bits from a BSS can, with the use of some adaptive coding scheme, be sent with zero error over the channel at rate $R > C_{oFa} - \delta$ bits/channel-use and with average coding delay at most $D_a$, $D_a < \infty$, for every information bit.

(The “a” in $C_{oFa}$ is intended as a reminder that we are considering “average coding delay”.)
By the coding delay for an information bit, we mean the number of channel uses starting when the information bit enters the encoder and ending when the last symbol is received that is used by the decoder in assigning a value to this information bit.

Let $D_i$ be the decoding delay for the $i$th information bit. For a block code of length $N$, $D_i \leq N$ for all $i$.

Our previous example of zero-error adaptive coding for the BEC gives $D_i \leq 1/(1-p)$ for all $i$ and hence

$$C_{oFa} = 1 - p = C.$$ 

We now consider the general condition for $C_{oFa} > 0$. 


We will say that an output letter $y$ of a discrete memoryless channel is a **disprover for the input letter** $x$ if $y$ cannot be reached from $x$, but $y$ can be reached from at least one input letter.

For the BEC with $0 \leq p < 1$, the output letter 0 is a disprover for the input letter 1, and the output letter 1 is a disprover for the input letter 0.
The idea behind the definition of a “disprover”: if y is a disprover for x, then the appearance of y in the received sequence proves that the corresponding transmitted symbol was not x.

**Question #5:**
Is the Binary Erasure Channel the simplest discrete memoryless channel with $C_o = 0$ whose output letters include a disprover???
NO! It is the Z-channel!

The output letter 1 is a disprover for the input letter 0.

Question #6: Does the Z-channel have $C_{oFa} > 0$???
YES!

Suppose we create a new channel by using the Z-channel for the input pairs 01 and 10 only.

Thus, by our previous result for the BEC, we know that for the Z-channel $C_{oFa} \geq \frac{1}{2} (1 - p)$ bits/use.
We can do this same “trick” for any discrete memoryless channel whose output alphabet contains at least one disprover $y$ for some input $x$. 

$(0 \leq p < 1)$ 

(x’ is any input letter that can reach $y$.) 

It follows that any such channel has 

$C_{oFa} \geq \frac{1}{2} (1 - p)$ bits/use. 

28
We have proved:

**Theorem 1:** For a discrete memoryless channel, 
\( C_{oFa} > 0 \) if and only if the output alphabet contains at least one disprover \( y \) for some input letter \( x \).

**Question #7:** When \( C_{oFa} > 0 \), does \( C_{oFa} = C \) ???
Theorem 2: If a discrete memoryless channel has $C_{oFa} > 0$, i.e., if the output alphabet contains at least one disprover $y$ for some input letter $x$, then $C_{oFa} = C$ (the small-error capacity).

Proof:
We can use the disprover $y$ for the input letter $x$ to create a Z-channel with probability $p$ ($0 \leq p < 1$) to the receiver. We can use two uses of this Z-channel to create a Binary Erasure Channel (BEC) with erasure probability $p$ as shown on slide 27. We will use this BEC to send a one-bit ACK or NAK to the receiver in the following way each time a coded block is sent on the forward channel:

YES!
• We use a block code for the DMC of long length $N$ and rate $R_B$ slightly less than $C$ that achieves very small block error probability $P_B$ on the DMC with maximum-likelihood (ML) decoding.

• From the complete feedback of the block, the encoder knows whether the ML decoding was correct in which case the encoder sets $b$ to 1 (ACK), or was incorrect in which case the encoder sets $b$ to 0 (NAK).

• The encoder now transmits $b$ repeatedly over the BEC with erasure probability $p$ formed from the DMC until $b$ is received unerased. This takes only $1/(1 - p)$ transmissions of $b$ on average, which is negligibly small compared to the block length $N$.

• This whole process needs to be performed for only $1/(1 - P_B)$ block transmissions on average.
Consider now **noiseless**, but **not necessarily complete**, feedback.

How much noiseless feedback is really needed?

**Question #8:**
Is **one binary digit of noiseless (but not complete) feedback per transmitted information bit** essentially enough to approach the zero-error capacity with feedback, $C_{\text{ofa}}$?
YES!

In our ACK/NAK block coding scheme for approaching $C_{oFa}$, $K = NR_B$ binary digits of noiseless feedback suffice for each block of length $N$ transmitted on the forward channel.

[The receiver can simply send back the $K$ decoded information bits over the noiseless channel. The repetitions of “NAK”ed blocks will cause the average number of binary digits sent on the noiseless feedback channel to slightly exceed $K$ for each block of $K$ information bits, but one can make this average approach $K$ arbitrarily closely by choosing $N$ sufficiently large.]
Do we really need this much noiseless feedback?

**Question #9:**
Are there DMCs with $C_{oFa} > 0$ and $C_o = 0$ for which one can send with zero error at a rate approaching $C_{oFa}$ using a number of binary digits of noiseless feedback for each transmitted bit of information that approaches zero?
YES!

Proposition 1:
There are coding schemes using noiseless feedback for sending with zero error at any rate less than $C_{oFa} = C = 1 - p$ over a BEC with $0 < p < 1$, in which the number of binary digits of noiseless feedback for each bit of information can be made to approach 0 arbitrarily closely.
For block coding without feedback on a DMC, we will say a decoding decision is **unambiguous** if the decoder knows that this decoding decision is certain to be correct.

For the **BEC**, if there is only one codeword that agrees with the received word in all unerased positions, then the decoding is unambiguous.

Let $P_{UA}$ be the probability of unambiguous decoding and let $P_A = 1 - P_{UA}$ be the probability of ambiguous decoding.

Let $P_{ML}$ be the error probability for **maximum-likelihood decoding** of the block code (without feedback).
Proposition 2: For block coding on a BEC with $0 < p < 1$, 
$P_A \leq 2 P_{ML}$.

Suppose $y$ is the received block and $x$ is a codeword. 
Then $P(y|x) = 0$ unless $x$ agrees with $y$ in all unerased positions, in which case 
$P(y|x) = (1-p)^{N-e} p^e$ where $e$ is the number of erasures.

$\Rightarrow$ every “all-agreeing” codeword is a valid choice for the maximum-likelihood decoding decision.

The correct codeword must agree with $y$ in all unerased positions. Thus the conditional probability of error for 
the ML decoder must be 0 when the decoding is unambiguous and is at least $\frac{1}{2}$ when decoding is ambiguous. It follows that $P_{ML} \geq \frac{1}{2} P_A$. 
• We use a block code for the DMC of long length $N$ and rate $R_B$ slightly less than $C$ that achieves very small block error probability $P_{ML} \geq \frac{1}{2} P_A$ on the BEC with maximum-likelihood (ML) decoding.

• If the decoding is unambiguous, the decoder sends $b = 1$ (ACK) over the noiseless feedback channel. If the decoding is ambiguous, the decoder sends $b = 0$ (NAK) over the noiseless feedback channel.
Is a noiseless feedback channel really needed?

**Question #10:**

*Do coding schemes* for sending with zero error at rates approaching $C_{oFa} = C = 1 - p$ over a *BEC* with $0 < p < 1$ actually *require noiseless feedback* ???
NO!

We can use a BEC with $0 < p' < 1$ as the feedback channel!

$\lambda$ ACK/NAK bits

- If decoding is unambiguous and the decoded $n$ is different from that in most recent unambiguous decoding, set $b = 1$ (ACK). Otherwise, set $b = 0$ (NAK).
- If encoder receives $b = 1$ on feedback channel, then the encoder encodes a new block and complements $n$.
- Use block length $N >> \lambda >> 1$. 

($n$ is a one-bit block sequence number, i.e., 0 or 1.)
Acknowledgement

The new results in this talk were obtained in joint research by

James L. Massey

and

Peter C. Massey