6.574 T

STATISTICAL THEORY OF INFORMATION

C.T. WHITEHEAD

SPRING '60

6.574T : The Statistical Theory of Information.

Prof. R. M. Fano, 26-241 — M.I.T.; Spring 1960

References:

Shannon & Weaver, The Mathematical Theory of Communication
Feinstein, Foundations of Information Theory
Feller,

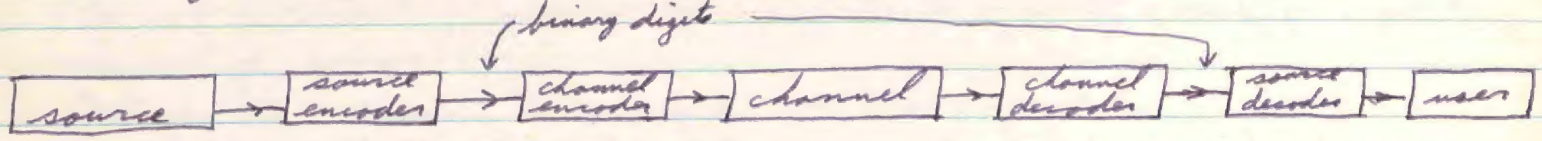
There are two fundamental models for study:

1. Signal not available before it is mixed with noise
   This approach (Wiener) is most useful in control applications.
   Developed for fire control systems.

2. Signal is available before it is mixed with noise.
   This approach (Shannon) is most useful for communications systems.

This course will follow the Shannon model. The two approaches have followed separate paths & don't seem to overlap.

The system model for Shannon's approach is:



3 fixed boxes: source, channel, & user characteristics are assumed fixed & given.

4 arbitrary boxes: encoders & decoders are the boxes we can play with, under the condition that the link between the source & channel coders are binary digital.

Noiseless channel problem is the above system with a lossless short circuit around the channel & channel coders.

Our objective is presumably to "reproduce" the source signal at the output so that the user can use it. The word "reproduce", however, implies a criterion of acceptability; how well must we reproduce so that it is acceptable from the users standpoint. This criterion depends on the situation; e.g., military telephony vs. private.

Given a criterion of acceptability, there are many possible outputs which are equivalent as far as the user is concerned.

This implies that we can divide (conceptually, at least) the inputs into a finite number of classes such that the user cannot (or need not) distinguish ~~which class a given~~ within a class, but can distinguish which class a given signal is in.

> "information" ≡ that which is being transmitted
>
> "message" ≡ class

Assume we have M messages per unit time T that we want to transmit. To each message we assign an integer $0, 1, 2, \ldots, M-1$. We can represent the messages in any form, binary digital being chosen.

If x binary digits are required to express the highest numbered message (i.e., it requires at most x digits to transmit a message), we must transmit x digits for all messages. This is to avoid confusion as to which time period we are in when we are sending a series of T messages.

If M messages are allowed, we need

$$[\log_2 M] \text{ digits/message}$$

where $[x]$ indicates the smallest integer $\geq x$.

e.g., $[6.3] = 7$, $[6.9] = 7$, $[6.0] = 6$

Example:

Message	Code	Message probability = P	Variable length code	$\log_2 1/P$
0	000	1/4	00	2
1	001	1/4	01	2
2	010	1/8	100	3
3	011	1/8	101	3
4	100	1/16	1100	4
5	101	1/16	1101	4
6	110	1/16	1110	4
7	111	1/16	1111	4

On the average, we can save time by using longer words for less probable messages & shorter words for the more probable signals. The variable length code is such that the beginning of no code word is the same as that of a shorter word; so the receiver can tell where a new word starts in a string of digits.

Here, the length/word is 3 for the simple 3 digit code. For the variable length code, the average length is:

$$\bar{n} = \sum_{K=0}^{7} n_K P_K = \tfrac{1}{4}(2) + \tfrac{1}{4}(2) + 2\left(\tfrac{1}{8}\right)(3) + 4\left(\tfrac{1}{16}\right)(4) = 2.75 < 3.$$

This code is optimal in the sense that 2.75 is the smallest possible $\bar{n}$ for this set of probabilities.

Source rate:

Under fairly general conditions, we can define for each source-user pair a rate R = information rate of the source. This rate is relative to the fidelity criterion of the user.

R ≡ greatest lower bound to the average number of binary digits per second which must be transmitted through the system to obtain reproduction at the output which is within the users acceptability criterion.

This reduces the communication problem to two distinct problems:

1. Getting the binary digits
2. Transmitting the binary digit.

We choose binary digits because of simplicity and convention; there is nothing fundamental about a binary system, per se.

Channels (Physically):

The existence of a channel implies:

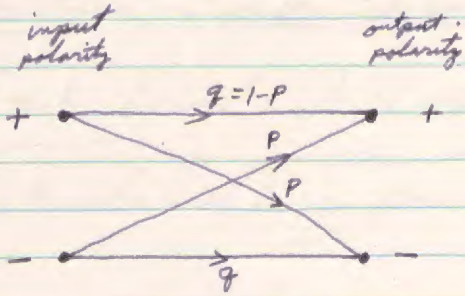
1. A means for generating a physical event (pulses, waving flags, etc.)
2. " " " detecting or observing these event
3. A media for the propagation of energy in some form (elect., mech., ...)

A channel will always have noise present (thermal, at the least), so we are somewhat uncertain as to the received signal.

The only way we can talk about a channel is by constructing and describing a suitable model which approximates a physical channel.

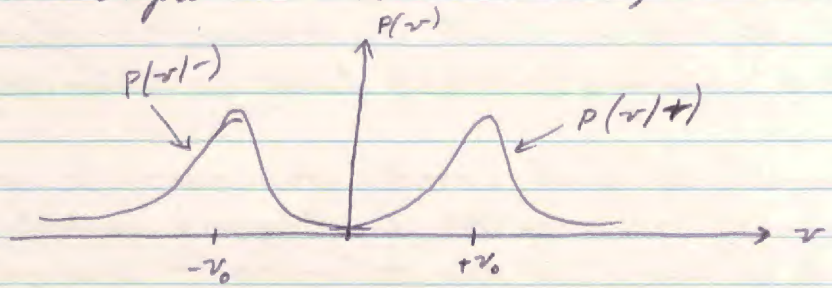
Binary symmetric channel model:

Signals are + or − ▬ pulses of fixed height & width; user detects only sign of signal.



$P \leq \frac{1}{2}$; $P = P_r\{\text{error}\}$

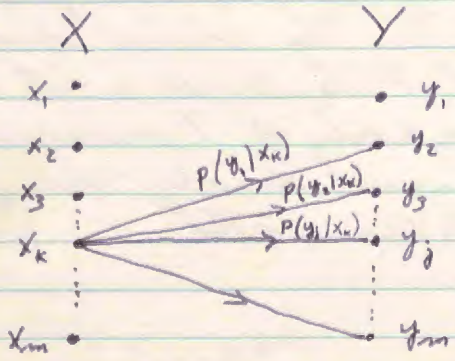
With an amplitude sensitive detector, the distribution of output is:



Or, we could have a continuum of input event, such as pulses of variable amplitude.

General channel model:

Finite number of discrete input & output events.



$\sum_j P(y_j|x_k) = 1$?

$\sum_k P(y_j|x_k) = 1$?

} doubtfully in general.

In general, the channel characteristics, i.e., the P's, will be functions of time and ~~the~~ some of the past states of the system.

Channel with memory ≡ characteristic probabilities depend on past events.

$$P(y_j | x_k) = f(t, x_{k-1}, \ldots, y_{j-1}, \ldots)$$

Shannon's Theorem:

Given a channel model, we want to know:
1. What is the maximum rate R the channel will allow?
2. What is the accuracy (probability of error) we can expect?

Given a fixed channel with finite (possibly zero) memory, we can define the channel capacity = C as the least upper bound to the number of binary digits per second that can be transmitted through the channel.

If $R < C$ (not $R \leq C$), we can design channel codes such that the probability of error is arbitrarily small.

The key to accomplishing this is the channel encoder + decoder. Suppose at any time there are n binary digits in the encoder. The output of the encoder is a function of these n digits; i.e., the output is a function of the state of the box.

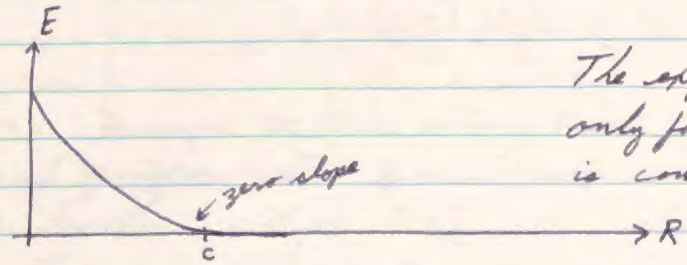
Each digit stays in the encoder a time $T = \frac{n}{R}$ & therefore influences the output of the box for a period T. The channel decoder should store the channel signal for a period T to decide what the digit input to the encoder was. This causes a delay of T seconds in the ~~to~~ overall transmission. The ~~decoder~~ channel decoder then puts out the digit which was most probably transmitted.

The probability of an error in the transmitted digit is given by

$$\lim_{n\to\infty} \frac{\ln P_e}{n} = -E \qquad \text{or} \qquad P_e \approx K e^{-nE}$$

$K = K(n, R, \text{channel})$; a <u>slow</u> variation

$E = E(R, \text{channel})$; generally $\ll 1$.



The exponential e^{-nE} becomes significant only for large n; otherwise, K is controlling.

The <u>law of large numbers</u> gives us that as $n \to \infty$,

$$P_r\{\%\ \text{incorrect pulses} > \text{expected}\ \%\} \to 0 \text{ exponentially as } n \to \infty.$$

Thus, we want to make (nE) as large as possible. But

larger n → more complex terminal equipment
larger E → less efficient usage of channel space.

Wozencraft has shown that complexity goes as (n) for the encoder, and as $(n \log n)$ for the decoder. For very large n, P_e goes as $(\text{complexity})^2$, so 2× complexity : $10^{-6} \to 10^{-12}$.

Signal spaces & probabilities :

We have two fundamental types of probability :

a priori : $P(x)$
a posteriori : $P(x|y)$

For example, consider the 3 digit code on p. 3. $P(x_3) = \frac{1}{8}$, but

$$P(x_3|0) = \frac{1}{6}, \quad P(x_3|01) = \frac{1}{2}, \quad P(x_3|011) = 1.$$

The output changes the probability distribution over the inputs; i.e., it tells us something about what the input was. The output is observed; the input is not.

Let

$X = \{x_k\}$, $k = 1, \ldots, n$, be the set of all possible inputs

$Y = \{y_j\}$, $j = 1, \ldots, m$, be the set of all possible outputs.

$Z = \{z_i\}$, $i = 1, \ldots, v$,

The product space XY is defined as the set of all possible pairs $(x_k y_j)$,

$$XY = \{(x_k, y_j)\}$$

Let $P(x)$ be the probability distribution over X,
$P(x,y)$ " " " " " XY

$$\sum_k P(x_k) = 1 \quad ; \quad \sum_k \sum_j P(x_k, y_j) = 1$$

$$P(x_k) = \sum_j P(x_k, y_j)$$

$$P(y_j) = \sum_k P(x_k, y_j)$$

$$P(x_k | y_j) P(y_j) = P(x_k, y_j)$$

x_k & y_j are statistically independent if & only if

$$P(x_k y_j) = P(x_k) P(y_j)$$

$P(x,y)$ is a measure of the statistical constraint between two events

Information function :

X, Y, and $P(x,y)$ completely describe the system.

We want to know what amount of "information" the knowledge of y gives about X. This process involves the change $P(x_k) \rightarrow P(x_k|y_j)$. Assume the measure of information about x given by y is given by

$$\boxed{I(x;y) \equiv \log_2 \frac{P(x|y)}{P(x)}}$$

The unit of information is a bit for $\log_2$; for $\log_{10}$, the unit is a Hartley ; and for $\log_e$, the nit. When the ratio of the a posteriori probability to the a priori probability changes by a factor of 2, the information changes by one bit.

If we have a triplet of events (x, y, z) and the event z has already occurred & been observed, we want the added information provided by the observation of y :

$$\boxed{I(x;y|z) = \log_2 \frac{P(x|y,z)}{P(x|z)}}$$

Symmetry of $I(x;y)$; Mutual information :

$$I(x;y) = \log_2 \frac{P(x|y)}{P(x)} = \log_2 \frac{P(x|y)\,P(y)}{P(x)\,P(y)} = \frac{P(x,y)}{P(x)\,P(y)}$$

Therefore, $I(x;y)$ is symmetric in x and y. The information provided by the event y about x is the same as the information provided by the event x about y. For this reason,

$$\boxed{\begin{array}{l} I(x;y) \equiv \text{mutual information of } x \text{ and } y. \\ \\ I(x;y) = \dfrac{P(x;y)}{P(x)\,P(y)} = I(y;x) \end{array}}$$

Note: If x & y are statistically independent, $I(x;y) = 0$.

$$I(x;yz) = \log \frac{P(x|yz)}{P(x)} = \log \frac{P(x|yz)\,P(x|y)}{P(x)\,P(x|y)} = \cancel{I(x;y) + I(x|}$$

$$= I(x;y) + I(x;z|y) = I(x;z) + I(x;y|z)$$

$$= I(yz;x)$$

Example: Let $XYUV = \{(x,y,u,v)\}$ ∋

$P(x,y,u,v) = P(x,y)\,P(u,v)$; i.e., XY and UV are stat. indep.

This corresponds to two independent systems:

We want to know $I(x,u;y,v)$, which is the information provided about the input pair x,u by the observation of the output pair y,v.
Intuitively, $I(x,u;y,v) = I(x;y) + I(u;v)$

$$I(x,u;y,v) = \log \frac{P(x,u,y,v)}{P(x,u)\,P(y,v)} = \log \frac{P(x,y)\,P(u,v)}{P(x)\,P(u)\,P(y)\,P(v)} = I(x;y) + I(u;v)$$

message (u)	code word (x)	code word (y)	P P(x,y)
0	0	0	1/2
1	0	1	1/4
2	1	0	1/8
3	1	1	1/8

Let $x_0 = 0$
$x_1 = 1$

If we put in message u_1, we first get out the digit 0; the information about u_1 given by the observation of x_0 is

$$I(u_1;x_0) = \log \frac{P(u_1|x_0)}{P(u_1)} = \log \frac{1/3}{1/4} = \log 4/3$$

After the zero has been observed, we see a "1". This gives us an additional amount of information

$\cancel{I(u_1;y_1}$

$$I(u_1; y_1 | x_0) = \log 1/3 = \log 3.$$

$$I(u_1; x_0 y_1) = \log 4$$

$$I(u_1; x_0 y_1) \stackrel{?}{=} I(u_1; y_1 | x_0) + I(u_1; x_0)$$

$$\log 4 = \log 3 + \log 4/3 = \log 4$$

Self-information:

If $P(x|y) = 1$, y specifies x completely, &

$$\boxed{I(x; y) = \log \frac{1}{P(x)} = -\log P(x) \equiv I(x) = \text{self-information of } x}$$

This is the maximum information that can be provided about x and is the ~~amount of~~ information that is necessary to specify x. This is not the information conveyed or associated with x, but rather the information that something external must provide to specify x.

$$I(x; y) \leq \begin{cases} I(x) = \log \frac{1}{P(x)} \\ I(y) = \log \frac{1}{P(y)} \end{cases}$$

Other information functions we can define are:

$$I(x,y) = \log \frac{1}{P(x,y)} = \text{joint self information}$$

$$I(x|y) = \log \frac{1}{P(x|y)} = \text{conditional self information}$$

$I(x;y) = I(x) - I(x|y)$

This simply says that the information given about x by y is the inf. necessary to specify x uniquely, less the inf. needed to specify x if y is given. This can be re-written as

$$I(x;y) = I(x) + I(y) - I(x,y) = I(y;x)$$

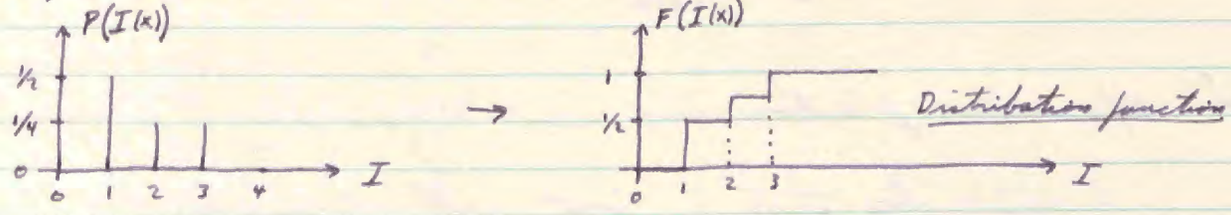
Another way of looking at this is

$$I(x,y) = I(x) + I(y) - I(x;y)$$

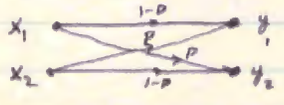
$I(x,y)$: inf nec to spec x, y; $I(x;y)$: inf counted twice.

Information functions as random variables:

The functions $I(x)$, $I(x;y)$, etc. can be regarded as random variables over the product space XY. From the distribution $P(x,y)$, we can find a probability distribution over these functions. For example: in the code table on page 10, we find



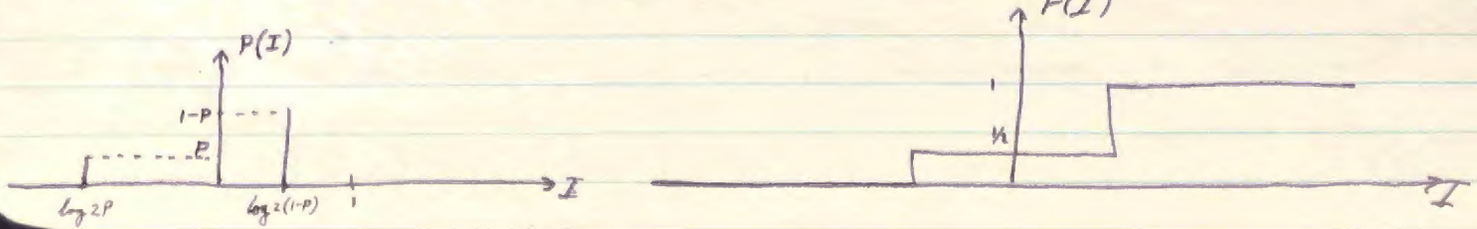
Binary asymmetric channel:



$P(x_1) = P(x_2) = 1/2$

$$I(x_1;y_1) = \log \frac{1/2\,(1-P)}{(1/2)(1/2)} = \log 2(1-P)$$

$$I(x_1;y_2) = \log \frac{1/2\,P}{1/2\,1/2} = \log 2P$$



$$E[I(x)] = \sum_X P(x) I(x)$$

$$E[I(x;y)] = \sum_{XY} P(x,y)\, I(x;y) = \sum_Y P(y)\, E[I(x;y_i)]$$

$$E[I(x;y_i)] = \sum_X P(x|y_i)\, I(x;y_i)$$

To fully specify these quantities, we must specify the space over which the random variable is evaluated.

$$Var[I(x)] = \sum_X P(x) I^2(x) - \left(\sum_X P(x) I(x)\right)^2$$

Entropy of a set:

Let $X = (x_1, \ldots, x_m)$; $P(x)$ over X.

$$\boxed{E[I(x)] = \sum_X P(x) I(x) = -\sum_X P(x) \log P(x) \equiv I(X) \equiv H(X) \equiv \text{entropy of } X}$$

This is the average self information of $x \in X$ and can be considered a measure of the disorder, or entropy.

$$\boxed{0 \le H(X) \le \log m}$$

$$H(X) = 0 \leftrightarrow \left\{ \begin{matrix} P(x_k) = 1 \\ P(x_i) = 0,\ i \neq k \end{matrix} \right\} \ ; \quad H(X) = \log m \leftrightarrow P(x_i) = \frac{1}{m}, \text{ all } i.$$

Proof: $H(X) \le \log m$:

Want to prove $H(X) - \log m \le 0$:

$$H(X) - \log m = \sum_X P(x) \log \frac{1}{m P(x)}$$

But $\ln w \le w - 1$, so $H(X) - \log m \le \sum_X P(x) \left[\frac{1}{m P(x)} - 1\right]$

$$= m \sum_X \frac{1}{m} \frac{P(x)}{P(x)} - \sum P(x) = \frac{m}{m} - 1 = 0. \quad \text{QED.}$$

Maximizing $H(X)$:

If $H(X) = \log m$, we have the maximum uncertainty in each digit. Since $H(X)$ is the average ~~self-information of a~~ info that can be supplied by a digit, we want to maximize $H(X)$ to provide the maximum information per digit. This applies to any alphabet, where we want each letter to occur with equal probability in each digit of the code words in order to max $H(X)$.

Capacity of an encoding alphabet:

This gives us the idea of the capacity of an alphabet. E.g., for binary digits, max $H(X) = \log_2(2) = 1$.

Example:

message	code word	Max H code	$P(x)$
u_0	000	00	1/4
u_1	001	01	1/4
u_2	010	101	1/8
u_3	011	110	1/8
u_4	100	1000	1/16
u_5	101	1101	1/16
u_6	110	1110	1/16
u_7	111	1111	1/16
			1

Here the max H code is arranged so that the digits 0 and 1 occur with equal probability in each digit (first, second, etc.) of the code words. $X = (0, 1) = (x_1, x_2)$

$H(X) = 1$ for max H code; for the other code in the first digit,

$$H(X) = \tfrac{3}{4}\log\tfrac{4}{3} + \tfrac{1}{4}\log 4$$

Conditional entropy:

The conditional entropy $H(Y|X)$ is the average self information of $y \in Y$ when the set X is known to have occured.

$$\boxed{H(Y|X) = -\sum_{XY} P(x,y) \log P(y|x) = \underset{XY}{\text{avg}}\, I(y|x)}$$

$$\boxed{H(Y|X) \leq H(Y)}$$

This says that it doesn't help to condition Y on X. The uncertainty of Y is greatest when X & Y are statistically independent.

Proof:

$$H(Y|X) - H(Y) = -\sum_{XY} P(x,y) \log P(y|x) + \sum_{Y} P(y) \log P(y)$$

$$= -\sum_{XY} P(x,y) \log P(y|x) + \sum_{XY} P(x,y) \log P(y)$$

$$= \sum_{XY} P(x,y) \log \frac{P(y)P(x)}{P(y|x)P(x)} \leq \sum_{XY} P(x,y) \left[\frac{P(x)P(y)}{P(x,y)} - 1 \right] \quad \text{since } \ln w \leq w-1$$

$$= \sum_{XY} P(x)P(y) - \sum_{XY} P(x,y) = 1 - 1 = 0.$$

Entropy of a message set:

$H(U)$ is dependent on the message code words and their probabilities. It is the average amount of information that must be supplied to specify a message.

Other set information functions:

We can talk about the information provided by a particular $y \in Y$ about the entire set X and vice versa for $x \in X$; these are essentially averages over one set of $I(x;y)$:

$$I(X;y) = \sum_{X} P(x|y)\, I(x;y) \geq 0$$

$$I(x;Y) = \sum_{Y} P(y|x)\, I(x;y) \geq 0$$

$$I(X;Y) = \sum_{Y} P(y)\, I(X;y) = \sum_{X} P(x)\, I(x;Y) = \sum_{XY} P(x,y)\, I(x;y) \geq 0$$

These can be shown to be always positive, which means that after we are told something about the events, we can't know less than we did before.

Interpretation of conditional entropies:

$$I(x;y) = I(x) - I(x|y) = I(y) - I(y|x)$$

Similarly,

$$H(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

$H(X;Y)$ is the average information given about X (or Y) by Y (or X).
$H(X)$ and $H(Y)$ are just the average information needed to specify an $x \in X$ or a $y \in Y$.

$H(X|Y) \equiv$ the equivocation = the average information needed to specify an $x \in X$ after the observation of a $y \in Y$.

$H(Y|X) \equiv$ noise entropy = useless information received; it is a measure of the severity of the noise.

Interpretation: change of uncertainty vs. mutual information:

$$H(X) - H(X|Y) = H(X;y) \leftrightarrow I(x;y) = I(x) - I(x|y)$$

$$H(x) - H(x|Y) = \sum_Y P(y)\left[H(X) - \sum_X P(x|y) \log \frac{1}{P(x|y)}\right] = \sum_Y P(y)\left[H(x) - H(X|y)\right]$$

but $\underset{x}{\text{avg}}\, I(x;y) = I(X;y) = \sum_X P(x|y)\left[I(x) - I(x|y)\right]$

$$\boxed{H(X) - H(X|y) \neq I(X;y)}$$

$H(X) - H(X|y)$ = difference of averages = average change of uncertainty

$I(X;y)$ = average of differences = average mutual information

Continuous Spaces:

A brief idea of how the following model can be ~~applied~~ related to a physical channel is to consider a signal to be divided into slots in time of duration T. The continuous signal in each slot can be represented as a point in n-space, where each coordinate is the amplitude of a ~~the~~ Fourier series expansion of the signal over the interval.

The model: Let U be a continuous space $U = \{\underline{u}\} \subset S_n$. Let δU be a small region of U about a point $\underline{u}$. We now define the quantity

$$\boxed{P(\delta U) \equiv P_r\{\underline{u} \in \delta U\}}$$

This is an additive measure for disjoint regions δU_1 and δU_2 ∋

$$\boxed{P(\delta U_1 + \delta U_2) = P(\delta U_1) + P(\delta U_2)}$$

Further, over the entire space U, $\boxed{P(U) = 1.}$

Probability density:

Regarding δU as a "volume" element of the space U, the density of the probability is given by

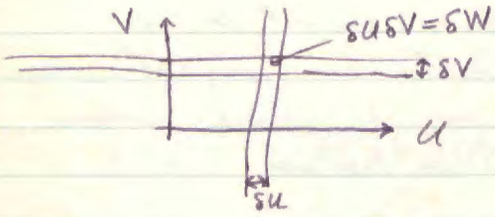
$$\lim_{\delta U \to 0} \frac{P(\delta U)}{\delta U} \equiv P(\underline{u})$$

This function will be considered piecewise continuous over U and can be interpreted as

$$P(\underline{u})\, dU = P_r\{\underline{u} \in dU\}$$

Continuous Product Spaces:

Consider two sets U and V with the product space $W = U V$



$\underline{u} \in U$

$\underline{v} \in V$

$(\underline{u}, \underline{v}) = \underline{w} \in W$

δU is a hypercylinder in the space W.

$$P(\delta W) = P_r\{\underline{w} \in \delta W\} = P_r\{\underline{u} \in \delta U,\ \underline{v} \in \delta V\}$$

We can again define a probability density

$$\lim_{\delta W \to 0} \frac{P(\delta W)}{\delta W} = \lim_{\substack{\delta U \to 0 \\ \delta V \to 0}} \frac{P(\delta U, \delta V)}{\delta U\, \delta V} = P(\underline{u}, \underline{v}) = P(\underline{w})$$

$$P(\underline{u}) = \lim_{\delta U \to 0} \frac{P(\delta U)}{\delta U} = \lim_{\delta U \to 0} \frac{\int_{\delta U} \int_V P(\underline{u}, \underline{v})\, dU\, dV}{\delta U\, \delta V} = \lim_{\delta U \to 0} \frac{\int_{\delta U} P(\underline{u})\, dU}{\delta U} = P(\underline{u})$$

$$P(\underline{u}) = \int_V P(\underline{u}, \underline{v})\, dv \quad \& \quad P(\underline{v} \mid \underline{u}) = \frac{P(\underline{u}, \underline{v})}{P(u)}$$

Mutual information:

$$I(\delta U; \delta V) = \log \frac{P(\delta U, \delta V)}{P(\delta U)\, P(\delta V)}$$

This is the information that the fact $\underline{u} \in \delta U$ gives about $\underline{v} \in \delta V$. In the limit, this goes to

$$I(dU; dV) = \log \frac{P(\underline{u}, \underline{v})\, dU\, dV}{P(\underline{u}) dU\, P(\underline{v}) dV} = \log \frac{P(\underline{u}, \underline{v})}{P(\underline{u})\, P(\underline{v})}$$

Thus the mutual information is independent of the size of the region in the limit and depends only on the ~~location of the~~ points. It has nothing to do with the coordinates and is a characteristic of the system, dependent only on the probabilities of events.

This can be written as

$$\boxed{I(\underline{u};\underline{v}) = \log \frac{P(\underline{u},\underline{v})}{P(\underline{u})P(\underline{v})} = \log \frac{P(\underline{u}|\underline{v})}{P(\underline{u})}}$$

All the properties of symmetry are maintained in this function as in the discrete case.

Self-information:

$$I(\underline{u}) = \lim_{\delta u \to 0} \frac{1}{P(\delta u)} \to \infty \text{ & is meaningless}$$

Averages:

$$I(U;V) = \int_u \int_v P(\underline{u},\underline{v})\, I(\underline{u};\underline{v})\, du\, dV = \int_v P(\underline{v})\, I(U;\underline{v})\, dV \geq 0$$

$$I(U;\underline{v}) = \int_u P(\underline{u}|\underline{v})\, I(\underline{u};\underline{v})\, du \geq 0$$

$I(U;V)$ is the limit of the discrete function

$$\sum_{k=1}^{n} \sum_{i=1}^{m} P(\delta U_k;\delta V_i)\, I(\delta U_k;\delta V_i)$$

As we further and further sub-divide ~~this~~ the spaces U and V, this function monotonically increases to the maximum $I(U;V)$ as defined by the above integral.

$$I(U;V) = H(U) - H(U|V) = H(V) - H(V|U)$$

so $$H(U) = -\int_u P(\underline{u}) \log P(\underline{u})\, du$$

$$H(U|V) = \int_u \int_v P(\underline{u},\underline{v}) \log P(\underline{u}|\underline{v})\, du\, dV$$

These quantities can not be interpreted as an average self information as self information has no significance in a continuous system.

Binary coding:

Let our message space be $U = (u_1, \ldots, u_M)$ with a probability distribution P over U.

Given an alphabet of D symbols, we want to encode the messages of U into sequences of symbols. $(D < M)$

$H(U)$ = average information necessary to specify a message.

$\log D$ = maximum average information that one symbol can provide.

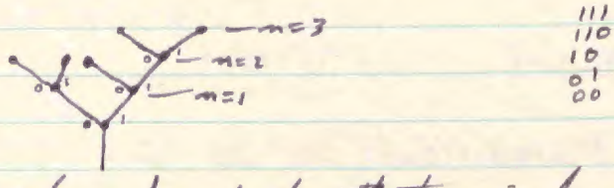
Hence, the average number of symbols per message is necessarily at least

$$\boxed{\bar{n} \geq \frac{H(U)}{\log D}}$$

To realize the equality, each digit in a message should be statistically independent of all orders. Since in general we cannot achieve this, can we find an upper bound to $\bar{n}$ which we can always attain?

Kraft's Inequality:

The requirement that no code word be the same as the beginning of a longer code word is equivalent to the requirement that each code word correspond to a unique branch on a tree.



Let n_K be the order of the K^{th} terminal node. For terminal nodes of order n, there are D^n possible nodes of order n. But if a branch prematurely terminates at n_K, there are D^{n-n_K} nodes not realized. The total number of nodes not realized is

$$D^{n-n_1} + D^{n-n_2} + \cdots + D^{n-n_M} = D^n \sum_{K=1}^{M} D^{-n_K}$$

But the number not realized must be less than or equal to the total number possible, so we have

$$D^{n} \sum_{K=1}^{M} D^{-n_K} \leq D^{n}$$

or, as a condition for the existence of code words of length n_K in a D symbol alphabet to cover M messages

$$\boxed{\sum_{K=1}^{M} D^{-n_K} \leq 1}$$

<u>An upper bound to $\bar{n}$</u> :

Kraft's inequality is tightest as a constraint on the lower bounds to n_K when it becomes an equality. Thus we want to minimize $\bar{n}$ subject to this constraint:

$$\min\left\{ \bar{n} = \sum_{K=1}^{M} P(u_K) n_K \right\} \quad ; \quad \sum_{K=1}^{M} D^{-n_K} = 1$$

Using Lagrange multipliers:

$$\frac{\partial}{\partial n_j}\left[\sum_{K=1}^{M} P(u_K) n_K + \lambda \sum_{K=1}^{M} D^{-n_K} \right] = 0 = P(u_j) - \lambda D^{-n_j} \ln D$$

$$D^{-n_j} = \frac{P(u_j)}{\lambda \ln D}$$

$$\sum_{j=1}^{M} D^{-n_j} = \frac{\sum_{j=1}^{M} P(u_j)}{\lambda \ln D} = \frac{1}{\lambda \ln D} = 1 \implies \lambda \ln D = 1$$

so $P(u_j) = D^{-n_j}$, $\log P(u_j) = -n_j \log D$

$$\boxed{n_j = \frac{-\log P(u_j)}{\log D} = \frac{I(u_j)}{\log D}} \text{ minimum}$$

Let n_K^* be the smallest integer $\geq \frac{I(u_K)}{\log D}$

Since we cannot realize a non-integral number of code symbols, we must use n_i^* symbols for word i. Thus

$$n_i^* < \frac{I(u_i)}{\log D} + 1$$

$$\bar{n} < \frac{H(U)}{\log D} + 1 \qquad \text{averaged.}$$

So $$\boxed{\frac{H(U)}{\log D} \leq \bar{n} < \frac{H(U)}{\log D} + 1}$$

If we have a sequence of L statistically independent messages, the entropy increases to $LH(U)$ for the sequence.

$$\frac{LH(U)}{\log D} \leq \bar{n}_L < \frac{LH(U)}{\log D} + 1$$

or $$\boxed{\frac{H(U)}{\log D} \leq \frac{\bar{n}_L}{L} < \frac{H(U)}{\log D} + \frac{1}{L}}$$

Thus for very long sequences, we can very closely approach the lower bound to $\bar{n}$. This lower bound can be interpreted physically as

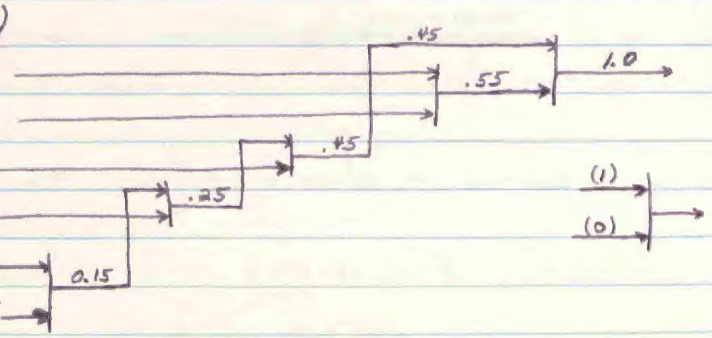
$\frac{H(U)}{\log D}$ = average number of digits required at minimum to transmit a message.

Huffman binary coding procedure

List the messages in order of decreasing probability. Combine the two least probable messages to get a new "group message 1" with a probability which is the sum of its component probabilities. Now repeat this procedure treating each "group message" as a single message. The result is a tree which defines an optimal code.

Example:

u	final code	$P(u)$
u_1	01	0.3
u_2	00	0.25
u_3	10	0.2
u_4	110	0.1
u_5	1111	0.1
u_6	1110	0.05



For an alphabet of base D, the tree splits into D branches at each node. If D and M are such that we cannot split evenly into D branches at every node [$M \neq 0 \pmod{D}$] then branch into fewer nodes for the least probable messages.

Each split of a node adds $(D-1)$ ~~nodes~~ terminal nodes. If the last split is into m_0 nodes,

$$M = 1 + t(D-1) + m_0 - 1 = t(D-1) + m_0 \qquad , 2 \leq m_0 \leq D.$$

or $$M - m_0 = t(D-1) = 0 \pmod{(D-1)}$$

Coding from a source:

$A = (a_1, \ldots, a_n) = \{\text{all source output letters}\}$

$A_i = \{\text{all possible letters in the } i^{th} \text{ position of a word}\} = \{\text{all } \alpha_i\}$

where a word is a sequence $\alpha_1 \alpha_2 \ldots \alpha_i \ldots \alpha_n$

A controlled source gives out a letter whenever we want one

An uncontrolled source puts out letters at a fixed rate or, more generally, in a known pattern in time.

A controlled source is described by the probabilities

$$P(\alpha_1), \; P(\alpha_2 | \alpha_1), \; P(\alpha_3 | \alpha_1, \alpha_2), \; \ldots, \; P(\alpha_K | \alpha_1, \ldots, \alpha_{K-1})$$

From these probabilities we can get any other distributions we want; e.g.,

$$P(\alpha_1, \alpha_2, \alpha_3) = P(\alpha_1) P(\alpha_2 | \alpha_1) P(\alpha_3 | \alpha_1, \alpha_2)$$

$$P(\alpha_1, \alpha_3) = \sum_{A_2} P(\alpha_1, \alpha_2, \alpha_3)$$

$$P(\alpha_3 | \alpha_1) = \frac{P(\alpha_3, \alpha_1)}{P(\alpha_1)} \qquad \text{etc.}$$

Stationarity:

If we assume some regularity in the source output so that

$$\boxed{P(\alpha_K | \alpha_{K-1}, \ldots, \alpha_{K-m}) = P(\alpha | \alpha_{-1}, \ldots, \alpha_{-m}) \text{ independent of } K}$$

then we say that the source output is stationary

We can now form the average self informations

$$H(A_1) = -\sum_{A_1} P(\alpha_1) \log P(\alpha_1) \qquad \geq 0$$

$$H(A_2|A_1) = -\sum_{A_1, A_2} P(\alpha_1, \alpha_2) \log P(\alpha_2|\alpha_1) \qquad \geq 0$$

Recall from page 15 that these entropies H decreases monotonously as more conditions are added:

$$H(A_K) \geq H(A_K|A_{K-1}) \geq H(A_K|A_{K-1}, A_{K-2}) \cdots$$

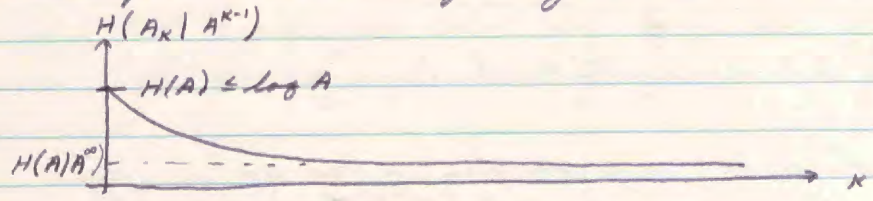
Applying this result to a stationary source, we see that since the subscript K is unnecessary,

$$H(A|A^{K-1}) \geq H(A|A^K)$$

$$\text{or} \quad H(A_m|A^{K-1}) \geq H(A_{m+1}|A^K)$$

$$\geq H(A_m|A^K)$$

This conditional entropy will approach a limit for conditions infinitely far back; i.e., the past has less & less influence the longer ago it occurred.



~~Result~~ Convergence:

$$H(A_1, \ldots, A_m) = H(A_1) + H(A_2|A_1) + \cdots + H(A_m|A_{m-1}, \ldots, A_1)$$

Dividing both sides by n & taking the limit, we see that as $n \to \infty$, ~~each term on the right~~ the first terms on the right contribute little to the total, so

$$\lim_{n\to\infty} \frac{1}{n} H(A_1, \ldots, A_n) = H(A|A^\infty)$$

$H(A_n|A^{n-1})$ = average information necessary to specify event n when the $n-1$ preceding events are known.

In general, $H(A_i|A^{n+1}) \leq H(A_i|A^n) \leq \cdots \leq H(A_i) \leq \log A$

If the source is stationary, we can translate the subscript i as it ~~is meaningless~~ has no effect on the relations involving it as such. Hence, we get for a stationary process

$$\boxed{H(A_1) \geq H(A_2|A_1) \geq H(A_3|A_2A_1) \geq \cdots \geq H(A_n|A^{n-1})}$$

It is only with a stationary process that we can say something about the future probabilities & entropies.

$H(A|A^\infty)$, (cont):

Since $0 \leq H(A_n|A^{n-1}) \leq \log A$ and since this quantity is a monotonously decreasing function of n, there must exist a limit as $n \to \infty$. We define this to be

$$\boxed{\lim_{n\to\infty} H(A_n|A^{n-1}) \equiv H(A|A^\infty)}$$

Now let us consider $\boxed{H_n(A) = \frac{1}{n} H(A_1, \ldots, A_n)}$

$$H_n(A) = \frac{1}{n}\sum_{i=1}^{j-1} H(A_i|A^{i-1}) + \frac{1}{n}\sum_{i=j}^{n} H(A_i|A^{i-1})$$

For any fixed j, the first sum approaches zero as a limit as each H is finite. Further, since $\lim_{i\to\infty} H(A_i|A^{i-1}) \equiv H(A|A^\infty)$ ~~for all~~ for any small $\epsilon > 0$, $\exists$ an i large enough so that

$$H(A_i|A^{i-1}) \leq H(A|A^\infty) + \epsilon$$

so $$\frac{1}{n}\sum_{i=j}^{n} H(A_i|A^{i-1}) \leq \frac{1}{n}\sum_{i=j}^{n}\left(H(A|A^\infty)+\epsilon\right) = \frac{n-j}{n}\left[H(A|A^\infty)+\epsilon\right]$$

for any ϵ there is a j sufficiently large so that this is true.

$n > j$

If we now pick our i so that this is true and let $n \to \infty$,

we have $\lim_{n\to\infty} \left(\frac{n-i}{n}\right)\left[H(A|A^\infty) + \epsilon\right] = H(A|A^\infty) + \lim_{n\to\infty} \frac{\epsilon}{n}$

$$\boxed{\lim_{n\to\infty} H_n(A) = H(A|A^\infty)}$$

It is obvious that the quantity $H(A|A^\infty)$ has some significance as a characteristic of the source.

$H_n(A) \equiv$ average entropy per event

$H(A|A^\infty) \equiv$ information per event rate of the source

$= \lim_{n\to\infty} \{$average entropy per event$\}$

How well can we encode?

$$H(A^L) \to L H(A|A^\infty) \quad \text{as} \quad L \to \infty$$

Since $\frac{H(A^L)}{L \log D} \leq \bar{n} < \frac{H(A^L)}{L \log D} + \frac{1}{L}$, $\bar{n} = \frac{\bar{n}_L}{L}$

Thus, as $L \to \infty$, we see that for any arbitrary & stationary source, the average number of — symbols per event is bounded by

$$\frac{H(A|A^\infty)}{\log D} \leq \bar{n} < \frac{H(A|A^\infty)}{\log D} + \epsilon \quad , \epsilon > 0$$

Coding efficiency $\equiv \eta = \frac{H(A|A^\infty)}{\bar{n} \log D}$

Averages: Ensemble & Time:

Ensemble average:

Take m identical & independent sources & divide their outputs into sequences of length L.

$$I^{(k)}(\alpha^L) = -\log P(\alpha_i, \ldots, \alpha_{i+L-1}) \quad \text{for } k^{th} \text{ source.}$$

$$J_m = \frac{1}{m} \sum_k I^{(k)} \rightarrow E[J_m] = E[I(\alpha^L)] = \underline{H(A^L)}$$

By the strong law of large numbers,

$$\lim_{j \to \infty} \Pr\{|J_m - H(A^L)| > \epsilon\} = 0 \quad \text{for any } m \geq j$$

The ensemble average is the mean of a large number m of similar sources operating simultaneously.

Time average:

Consider one source, many sequences of length L in sequence.

$$\lim_{m \to \infty} J_m \equiv \underline{\langle I(\alpha^L) \rangle = \text{time average}}$$

$$+ \quad \Pr\{|J_m - \langle I(\alpha^L) \rangle| > \epsilon\} < \delta \quad \text{for } m \geq j$$

The limit of J_m may be different for different sequences from the same source. But if the source is ergodic, then

$$\langle I(\alpha^L) \rangle \rightarrow H(A^L).$$

$J_m \leftrightarrow$ particular sequence

If events of a sequence are statistically independent,

$$Pr\{|J_m - H(A_L)| > \epsilon\} < \delta \quad , \quad m \geq j$$

i.e., the time average approaches the ensemble average.

If the events of a sequence are not statistically independent, we define $\lim J_m = \langle I(\alpha_i^L) \rangle$ and then

$$Pr\{|J_m - \langle I(\alpha^L) \rangle| > \epsilon\} < \delta \quad , \quad m \geq j$$

A source is ergodic if the ensemble of all possible sequences does not contain a subset that is a stationary process.

Any non-ergodic source can be represented as a collection of ergodic sources.

If a source is ergodic, then $H(A^L) = \langle I(\alpha^L) \rangle$.

Message encoding:

Divide a sequence of events from a source into groups of ν events and assign to each message a code word of length n.

A^ν = number of possible messages

D^n = " " " " with code words of n symbols from a alphabet of D symbols.

$$A^\nu = D^n \rightarrow n = \nu \frac{\log A}{\log D}$$

This is just a simple change of alphabets & gains us nothing. The inefficiency is caused by the requirement that all code words have the same ~~message~~ number of symbols.

Message encoding:

Consider a fixed rate ergodic source. Let a sequence of ν events be a message and consider n sequences of n messages. We now try encoding groups of n messages ($n\nu$ events).

$I_i(\alpha^\nu) = -\log P(\alpha_{i\nu+1}, \alpha_{i\nu+2}, \ldots, \alpha_{i\nu+\nu})$ = self information of particular message i.

$J_n = \frac{1}{n}\sum_{i=1}^{n} I_i(\alpha^\nu)$ = mean self information of n particular messages

Let

$$T = \{\text{sequences} : |J_n - H(A^\nu)| \le \epsilon\}$$

$$\overline{T} = \{\text{sequences} : |J_n - H(A^\nu)| > \epsilon\}$$

By the law of large numbers, for sufficiently large n

$$P(\overline{T}) < \delta \quad \text{for any } \epsilon > 0.$$

or $\underline{P(T) > 1-\delta.}$

This says that the probability that the mean of n actual sequences differs from $H(A^\nu)$ by less than ϵ can be made arbitrarily close to unity (within δ for given ϵ). I.e., n may be selected large enough so that (with probability $1-\delta$)

$$n[H(A^\nu) - \epsilon] < n J_n = \sum_{i=1}^{n} I_i(\alpha^\nu) < n[H(A^\nu) + \epsilon]$$

or $$2^{-\sum I_i(\alpha^\nu)} = \prod_{i=1}^{n} P(\alpha_{i\nu+1} \cdots \alpha_{i\nu+\nu}) > 2^{-n[H(A^\nu)+\epsilon]}$$

Now $$\sum_{T+\overline{T}} \prod_{i=1}^{n} P(\alpha_{i\nu+1} \cdots \alpha_{i\nu+\nu}) = 1, \text{ so } 1 > \sum_{T} \prod_{i=1}^{n} P(\alpha_{i\nu+1} \cdots \alpha_{i\nu+\nu})$$

Define $M \equiv$ number of sequences of n messages belonging to T.

If we now disregard all ~~sets~~ sequences of messages in $\overline{T}$, we ~~get~~ loose very little as $P(\overline{T}) < \delta$.

$$P(T) > 2^{-n[H(A^\nu)+\epsilon]}$$ so

$$1 > \sum_T \prod_{i=1}^{n} P(\alpha_{i\nu+1} \cdots \alpha_{i\nu+\nu}) > M\, 2^{-n[H(A^\nu)+\epsilon]}$$

or $$\boxed{M < 2^{n[H(A^\nu)+\epsilon]}}$$

$u \equiv$ length of code word to encode a sequence of n messages.

Then the above result says

$$\frac{u}{n} < \frac{H(A^\nu)+\epsilon}{\log D}$$

$$\lim_{\nu\to\infty} \frac{H(A^\nu)}{\nu} = H(A|A^\infty)$$

i.e., $$\frac{H(A^\nu)}{\nu} < H(A|A^\infty) + \epsilon$$

Since $\frac{u}{n\nu}$ = # of coding symbols per event, we have

$$\boxed{\frac{u}{n\nu} < \frac{H(A|A^\infty) + \epsilon\left(1+\frac{1}{\nu}\right)}{\log D}}$$

For sufficiently large n and u, the coding efficiency can be made arbitrarily close to $H(A|A^\infty)/\log D$. This is for a fixed rate source with some messages not encoded.

Noisy channels:

Channel model: inputs $x \in X$, where X is the input alphabet.
outputs $y \in Y$, " Y " output "

An input sequence is $\xi^n = (\xi_1, \ldots, \xi_n)$, $\xi_i \in X$

An output sequence is $\eta^n = (\eta_1, \ldots, \eta_n)$, $\eta_i \in Y$.

$P(y|x)$ can be discrete or continuous.
The channel performs a transformation of X to give elements of Y.
Discrete – continuous & vice versa transformations are allowed.

In general there will be constraints on the inputs, e.g. the input power must be kept below some level on the average.

Types of channels:

(1) Constant channel: $P(\eta_i | \xi_i)$ = constant for all i.

(2) Constant channel with memory: probabilities depend on past events:

$P(\eta_i | \xi_i, \xi_{i-1}, \ldots, \xi_{i-m})$ e.g. η_i multiple path transmission.

(3) Time varying channel: $P(\eta_i | \xi_i)$ depends on α_i where α_i is generated by a stationary stochastic process.

Two-way channels: $x_1 \longrightarrow y_1$, $y_2 \longleftarrow x_2$

Here, consider $X = X_1 X_2$ & $Y = Y_1 Y_2$

Constant, discrete, one-way channels:

$$P(y|x) = P(\eta_i|\xi_i) \quad \text{, all } i$$

For sequences, $P(\eta^n|\xi^n) = \prod_{i=1}^{n} P(\eta_i|\xi_i)$

Define $X^n = X_1 X_2 \cdots X_n$

$$P(\eta^n) = \sum_{X^n} P(\xi^n) P(\eta^n|\xi^n)$$

Average mutual information $= I(X^n; Y^n) = H(Y^n) - H(Y^n|X^n)$

where $H(Y^n) = H(Y_1) + H(Y_2|Y_1) + \cdots + H(Y^n|Y_{n-1}\cdots Y_1)$

$$H(Y^n|X^n) = -\sum_{Y^n X^n} P(\xi^n) P(\eta^n|\xi^n) \log P(\eta^n|\xi^n)$$

$$= -\sum_{Y^n X^n} P(\xi^n) P(\eta^n|\xi^n) \sum_{i=1}^{n} \log P(\eta_i|\xi_i)$$

$$= -\sum_{i=1}^{n} P(\xi_i) P(\eta_i|\xi_i) \log P(\eta_i|\xi_i) = \sum_{i=1}^{n} H(Y_i|X_i)$$

$$\therefore \quad I(X^n; Y^n) = \sum_{i=1}^{n} [H(Y_i|Y_{i-1}\cdots Y_1) - H(Y_i|X_i)]$$

$$= H(Y^n) - H(Y^n|X^n)$$

$$I(X^n; Y^n) = \sum_{i=1}^{n} [H(Y_i|Y_{i-1}\cdots Y_1) - H(Y_i|X_i)] \leq \sum_{i} [H(Y_i) - H(Y_i|X_i)] = \sum_{i} I(X_i; Y_i)$$

This says that given $P(\xi_i)$ for all i, the quantity $I(X^n; Y^n)$ is a maximum when the events are statistically independent, i.e., when $P(\xi^n) = \prod_{i=1}^{n} P(\xi_i)$.

A constant channel ~~defines~~ provides maximum mutual information when the input event are statistically independent. Under this condition the outputs are also statistically independent.

Now define $C = \max_{P(x)} I(X;Y) = \max_{P(x)} P(x)P(y|x) \log \frac{P(y|x)}{P(y)}$

where $P(y) = \sum_x P(x)P(y|x)$

and $P(y|x)$ is given as a characteristic of the channel.

Convexity of $I(X;Y)$:

If $P_1(x) \leftrightarrow I_1(X;Y)$ and $P_2(x) \leftrightarrow I_2(X;Y)$

and if $P_0 = \lambda P_1 + (1-\lambda) P_2$, ~~then $I_0(X;Y) \geq \lambda I_1(X;$~~

then $I_0(X;Y) \geq \lambda I_1(X;Y) + (1-\lambda) I_2(X;Y)$

See next page →

Probability space:

Represent all possible probability distributions over X as points in M-space. $P_i = P(x_i)$. $\sum_x P(x) = 1$ so the set of points is an $M-1$ dimensional hyper-plane.

Consider m distributions $P_1(x), \ldots, P_m(x)$ and let $\alpha_1, \ldots, \alpha_m$ be a set of real positive numbers such that $\sum \alpha_i = 1$. Then $P_0 = \sum_i \alpha_i P_i$ is still a probability distribution where

$$P_0(x) = \sum_i \alpha_i P_i(x) \quad + \quad \sum_x P_0(x) = \sum_x \sum_i \alpha_i P_i(x) = \sum_i \alpha_i \sum_x P_i(x) = \sum \alpha_i = 1.$$

$$P_i(y) = \sum_x P_i(x) P(y|x)$$

$$P_0(y) = \sum_i \alpha_i P_i(y)$$

Define $I_i(X;Y) = H_i(Y) - H_i(Y|X)$

$$H_i(Y) = -\sum_y P_i(y) \log P_i(y)$$

$$H_i(Y|X) = -\sum_x P_i(x) \sum_y P(y|x) \log P(y|x)$$

In general, $\boxed{I_0(X;Y) \geq \sum_i \alpha_i I_i(X;Y)}$ $\boxed{\rightarrow I(X;Y) \text{ is convex}}$

Proof: define $\delta I(X;Y) = \sum_i \alpha_i I_i - I_0 = \sum_i \alpha_i \left[H_i(Y) - H_i(Y|X)\right] - \left[H_0(Y) - H_0(Y|X)\right]$

But $H_0(Y|X) = -\sum_x \sum_i \alpha_i P_i(x) \sum_y P(y|x) \log P(y|x) = -\sum_i \alpha_i \sum_x P_i(x) \sum_y P(y|x) \log P(y|x) = \sum_i \alpha_i H_i(Y|X)$

so $\delta I(X;Y) = \delta H(Y) = \sum_i \alpha_i H_i(Y) - H_0(Y)$

$$= -\sum_i \alpha_i \sum_y P_i(y) \log P_i(y) + \sum_y \left[\sum_i \alpha_i P_i(y)\right] \log P_0(y) = \sum_i \alpha_i \sum_y P_i(y) \log \frac{P_0(y)}{P_i(y)}$$

$$\leq \sum_i \alpha_i \sum_y P_i(y) \left[\frac{P_0(y)}{P_i(y)} - 1\right] = \sum_i \alpha_i (1-1) = 0. \qquad \text{Q.E.D.}$$

Evaluation of channel capacity:

$C = \max_{P(x)} I(X;Y)$, $P(y|x)$ is given.

We require for a maximum:

$$\frac{\partial}{\partial P(x)}\left[I(X;Y) + \lambda \sum_X P(x)\right] = 0 \quad , \text{ each } x \in X.$$

$$I(X;Y) = H(Y) - H(Y|X)$$

Using natural logs, we have $H(Y) = -\sum_Y P(y) \ln P(y)$; $P(y) = \sum_X P(y|x) P(x)$

$$\frac{\partial H(Y)}{\partial P(x)} = \sum_Y \frac{\partial H(Y)}{\partial P(y)} \frac{\partial P(y)}{\partial P(x)} = -\sum_Y \left[1 + \ln P(y)\right] P(y|x) = -1 - \sum_Y P(y|x) \ln P(y)$$

$$H(Y|X) = -\sum_{X,Y} P(x) P(y|x) \ln P(y|x)$$

$$\frac{\partial H(Y|X)}{\partial P(x)} = -\sum_Y P(y|x) \ln P(y|x) \qquad * \quad \frac{\partial}{\partial P(x)} \sum_X P(x) = 1$$

Then our condition for a maximum becomes

$$0 = \sum_Y P(y|x) \ln \frac{P(y|x)}{P(y)} - 1 + \lambda = I(x;Y) - u \quad ; \quad \boxed{u \equiv 1 - \lambda}$$

This gives N equations. If we multiply each by $P(x)$ and add, we get

$$\sum_X P(x) I(x;Y) - \sum_X P(x)\, u = I(X;Y) - u = 0 \implies \boxed{u = C = \text{capacity}}$$

Thus the condition is that

$$\boxed{I(x;Y) = C \text{ , all } x \in X \text{; then } C = \text{channel capacity}}$$

Or, we can solve for $P(x)$ and C from the equations

$$(M) \quad \sum_y P(y|x) \ln P(y) = -[C + H(Y|x)]$$

$$(N) \quad \sum_x P(x) P(y|x) = P(y)$$

$$(1) \quad \sum P(x) = 1$$

$M+N+1$ equations $\Leftrightarrow$ $M+N+1$ unknowns $[P(x), P(y), C]$

We also require that $P(x) \geq 0$. If our solution of the equations does not conform to ~~the equations~~ this restriction, we know the solution lies on the boundary of the region, i.e., $P(x) = 0$ for at least one $x \in X$. Solution is by trial & error.

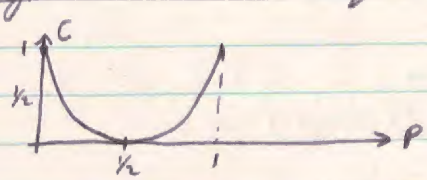
$M = N$, we can solve the first two sets of equations in sequence.
$M > N$, set $P(x) = 0$ for $M-N$ x's.
$M < N$, must solve all equations simultaneously.

Uniformity:

If all rows of the conditional probability matrix are circular permutations of one another, the channel is said to be uniform from the input. If the columns have this property, the channel is said to be uniform from the output.

If both conditions hold, the max of $I(X;Y)$ occurs when the inputs are equally probable; & the outputs will also be equiprobable.

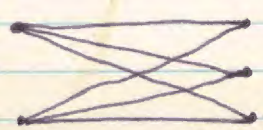
Binary symmetric channel capacity:



Binary channel with erasure:

$C \leftrightarrow P(x_1) = P(x_2)$

Binary channel with erasure & crossover:



$C \leftrightarrow P(x_1) = P(x_2)$

Capacity of continuously distributed channels:

We now consider sequences of discrete events which are distributed according to a continuous distribution.

$$I(X;Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(y|x)\,p(x) \log \frac{p(y|x)}{p(y)}\,dy\,dx = H(Y) - H(Y|X)$$

$$H(Y) = -\int_{-\infty}^{\infty} p(y) \log p(y)\,dy$$

$$H(Y|X) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(y|x)\,p(x) \log p(y|x)$$

We want to find $C = \max_{p(x)} I(X;Y)$. First, assume $y = x + z$; x and z are independent; $\overline{z} = 0$;

$$\int_{-\infty}^{\infty} x^2 p(x)\,dx \leq S \quad ; \quad \int_{-\infty}^{\infty} z^2 p(z)\,dz = N$$

Since $H(Y|X)$ is independent of x, we will realize C when $H(Y)$ is maximum. The $p(y)$ must satisfy the relation $p(y) = \int_{-\infty}^{\infty} p(x)\,p(z)\,dx$: It is intuitively apparent that we will get maximum mutual information when $\int x^2 p(x)\,dx = S$. Also $\int p(y)\,dy = 1$

We now max $H(Y)$ according to these constraints by using generalized Lagrange multipliers. First though, we must find the constraint on y due to the constraints on $\overline{z^2}$ and $\overline{x^2}$. This is:

$$\overline{y^2} = \int_{-\infty}^{\infty} (x^2 + z^2 + 2xz) \int_{-\infty}^{\infty} p(x)\,p(z)\,dx\,dy = S + N.$$

so $$\delta\left[\int_{-\infty}^{\infty} p(y) \ln p(y)\,dy + \lambda \int_{-\infty}^{\infty} y^2 p(y)\,dy + \mu \int_{-\infty}^{\infty} p(y)\,dy\right] = 0$$

or $$\delta \int_{-\infty}^{\infty} \left[p(y) \ln y + \lambda y^2 p(y) + \mu\, p(y)\right] dy = 0$$

so $\frac{\partial}{\partial P(y)}\left[P(y)\ln P(y) + \lambda y^2 P(y) + \mu P(y)\right] = 0$, all y

Since $H(Y)$ is convex this will always be a maximum.

Then $0 = -(1+\ln P(y)) + \lambda^2 y^2 + \mu$

or $\ln P(y) = \lambda y^2 + \mu - 1 \Rightarrow P(y) = e^{\mu-1} e^{\lambda y^2} \leftrightarrow$ Gaussian.

We further find that $\lambda = -\frac{1}{2(S+N)}$; $e^{\mu-1} = \frac{1}{\sqrt{2\pi(S+N)}}$

Thus, if we fix $\overline{X^2}$, Gaussian $P(y)$ yields the maximum mutual information for this channel model if we can realize that distribution.

Continuing with this result, we find $H(Y) = \frac{1}{2}\log 2\pi e (S+N)$

& $C = \frac{1}{2}\log 2\pi e (S+N) - H(Y|X)$

so if S decreases, C decreases as $H(Y|X)$ is independent of S.

If the noise z is gaussian (e.g., thermal noise),

$P(z) = \frac{1}{\sqrt{2\pi N}} e^{-\frac{z^2}{2N}}$

We can now make $P(y)$ Gaussian by making $P(x)$ Gaussian.

$P(x) = \frac{1}{\sqrt{2\pi S}} e^{-\frac{x^2}{2S}}$

$H(Y|X) = H(z) = \frac{1}{2}\log 2\pi e N$

$C = \frac{1}{2}\log\left(1 + \frac{S}{N}\right)$

Effective variance & bounds on C.

Consider a Gaussian distribution which has the same entropy as the true distribution $P(x)$. We call the variance of this Gaussian $\bar{\sigma}_x^2$. It can be shown that

$$\boxed{\log \bar{\sigma}_x^2 = 2H(X) - \log 2\pi e} \qquad \text{(Shannon)}$$

$\boxed{\bar{\sigma}_x^2 \text{ is called the entropy power of the input.}}$

Shannon has shown that $\boxed{\bar{\sigma}_x^2 + \bar{\sigma}_z^2 \leq \bar{\sigma}_y^2 \leq \sigma_x^2 + \sigma_z^2 \quad , \quad y = x + z}$

Now $C \geq I(X;Y) = H(Y) - H(Y|X)$

$$= \tfrac{1}{2}\log 2\pi e \bar{\sigma}_y^2 - \tfrac{1}{2}\log 2\pi e \bar{\sigma}_z^2 = \tfrac{1}{2}\log \frac{\bar{\sigma}_y^2}{\bar{\sigma}_z^2} \geq \tfrac{1}{2}\log \frac{\bar{\sigma}_x^2 + \bar{\sigma}_z^2}{\bar{\sigma}_z^2}$$

$$= \tfrac{1}{2}\log\left(1 + \frac{\bar{\sigma}_x^2}{\bar{\sigma}_z^2}\right)$$

so $\boxed{\tfrac{1}{2}\log\left(1 + \frac{\bar{\sigma}_x^2}{\bar{\sigma}_z^2}\right) \leq C \leq \tfrac{1}{2}\log \frac{S+N}{\bar{\sigma}_z^2}}$ since $H(Y|X) = \tfrac{1}{2}\log 2\pi e \sigma_z^2$

If we make $P(x)$ Gaussian, $\bar{\sigma}_x^2 = S$
" " " $P(z)$ " , $\bar{\sigma}_z^2 = N$ $\Big\}$ $C = \tfrac{1}{2}\log\left(1 + \frac{S}{N}\right)$.

$\bar{\sigma}_z^2 \leq N$ so $C \geq \tfrac{1}{2}\log\left(1 + \frac{S}{\bar{\sigma}_z^2}\right) \geq \tfrac{1}{2}\log\left(1 + \frac{S}{N}\right)$

If we specify S and N, gaussian noise yields the lowest capacity. Thus in this sense we can say " " is the worst noise.

Continuous Channels:

Consider a continuous function $u(t)$. We can represent $u(t)$ over the range $0 \le t \le T$ by

$$u(t) = \sum_{m=-\infty}^{\infty} X_m \left[\cos m\varphi t + \sin m\varphi t\right] \quad , \varphi = \frac{2\pi}{T}$$

$$X_m = \frac{1}{T}\int_0^T u(t)\left[\cos m\varphi t + \sin m\varphi t\right] dt.$$

Let $u(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} g(\omega)\, e^{j\omega t}\, d\omega$

$$g(\omega) = \int_0^T u(t)\, e^{-j\omega t}\, dt$$

Define $E = \int_0^T u^2(t)\,dt = T\sum_{-\infty}^{\infty} X_m^2$

$$E = \frac{1}{2\pi}\int_{-\infty}^{\infty} |g(\omega)|^2\, d\omega$$

We can expand $g(\omega)$ as

$$g(\omega) = \sum_{m=-\infty}^{\infty} g(m\varphi)\left[\frac{\sin\frac{T}{2}(\omega - m\varphi)}{\frac{T}{2}(\omega - m\varphi)}\right]$$

↳ orthogonal fns, parameter m.

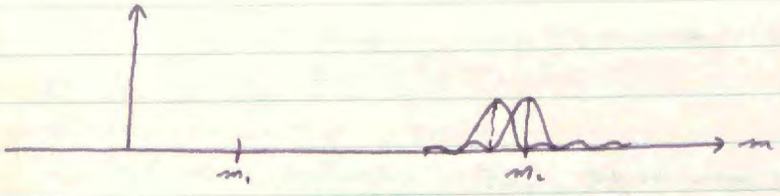
$$E = \frac{1}{2\pi}\sum_{-\infty}^{\infty} |g(\omega)|^2 \int_{-\infty}^{\infty}\left[\frac{\sin T/2(\omega - m\varphi)}{T/2\,(\omega - m\varphi)}\right]^2 d\omega = \frac{1}{T}\sum_{-\infty}^{\infty} |g(\overset{m\varphi}{\downarrow})|^2$$

$$X_0^2 = \frac{|g(0)|^2}{T^2} \quad ; \quad X_m^2 + X_{-m}^2 = \frac{2\,|g(m\varphi)|^2}{T^2}$$

Thus, we need talk only about $m \ge 0$.

Band-limited signals

Suppose in the ~~previous~~ formulation on the last page that $m = 0$ for all m outside the region $0 \leq m_1 < m \leq m_2$:



The $\left(\frac{\sin x}{x}\right)^2$ terms die out rapidly beyond the boundaries m_1, m_2. Thus, nearly all the energy is contained in the band

$$W = \frac{(m_2 - m_1)\,\theta}{2\pi} = \frac{m_2 - m_1}{T}$$

~~Strictly~~ time limited signals for which $u(t) = 0$ outside $(0, T)$ ~~are not bandlimited~~ are called band-limited in this approximate (F.C.C.) sense if all $m = 0$ outside some band.

Thus any such band-limited function can be completely specified over a range T by $2TW$ numbers which represent the amplitudes of the $\frac{\sin x}{x}$ functions.

Continuous time channels:

Let $u(t)$ = input; $n(t)$ = noise (assumed gaussian); $v(t)$ = output

$$v(t) = u(t) + n(t)$$

Definition of continuous gaussian noise:

Take M samples of $n(t)$ at different times. Let these values of $n(t)$ be $z_1, \ldots, z_M$.

Let $E(z_i z_k) = a_{ik} = a_{ki}$; $A = \det(a_{ik})$; A_{ik} = cofactor (a_{ik}).

$n(t)$ is gaussian if

$$\boxed{P(z_1, \ldots, z_M) = \frac{1}{(2\pi)^{M/2} A^{1/2}} \exp\left[-\frac{1}{2A}\sum_k \sum_i A_{ik} z_i z_k\right]}$$

If $a_{ik} = \begin{cases} 0, & k \neq i \\ \sigma_k^2, & k = i \end{cases}$, then $A = \prod_k \sigma_k^2$; $A_{kk} = \frac{A}{\sigma_k^2}$

and $P(z_1, \ldots, z_M) = \frac{1}{(2\pi)^{M/2} A^{1/2}} \exp\left[-\frac{1}{2A}\sum_k \frac{A}{\sigma_k^2} z_k^2\right] = \prod_k \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{z_k^2}{2\sigma_k^2}}$

Then the samples are mutually independent. It is both necessary & sufficient that the cross-correlation coefficients a_{ij} be zero for the samples to independently gaussian distributed.

A linear transformation on the z's will always put the distribution in this independent form. A linear filter might be such a transformation.

Stationary processes in continuous channels:

Let $n_t \equiv n(t)$

The noise is the result of a stationary process if the cross-correlation between two samples is dependent only on the time difference between samples. I.e.,

$$E(n_t n_\theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} n_t n_\theta P(n_t n_\theta)\, dn_t\, dn_\theta = R(t-\theta) = R(\tau), \quad \tau \equiv t-\theta.$$

n_t is a random variable over many identical processes.

Consider the time interval $-T/2 \le t \le T/2$. We can expand $n(t)$ over this interval as a series of functions which are mutually orthogonal over this interval.

$$n(t) = \sum_{k=-\infty}^{\infty} z_k \varphi_k(t) \quad ; \quad z_k = \frac{1}{T}\int_{-T/2}^{T/2} n(t)\varphi_k(t)\, dt$$

$$\frac{1}{T}\int_{-T/2}^{T/2} \varphi_k(t)\varphi_i(t) = \begin{cases} 0, & k \neq i \\ 1, & k = i \end{cases}$$

If $n(t)$ is gaussian, then the z_k are gaussianly distributed also.

$$E(z_i z_k) = \frac{1}{T^2} E\left[\int_{-T/2}^{T/2}\int_{-T/2}^{T/2} n_t n_\theta \varphi_i(\theta)\varphi_k(t)\, dt\, d\theta\right]$$

$$= \frac{1}{T^2}\int_{-T/2}^{T/2}\int_{-T/2}^{T/2} E(n_t n_\theta)\varphi_k(t)\varphi_i(\theta)\, dt\, d\theta = \frac{1}{T^2}\int_{-T/2}^{T/2}\int_{-T/2}^{T/2} R(t-\theta)\varphi_k(t)\varphi_i(\theta)\, dt\, d\theta$$

Again, z_k and z_i are independent if $E(z_k z_i) = 0$, $k \neq i$

Onfath, $\int_{-T/2}^{T/2} R(t-\theta)\varphi_k(\theta)\, d\theta = \sigma_k^2 \varphi_k(t)$

so $E(z_k z_i) = \frac{1}{T^2}\int_{-T/2}^{T/2} \sigma_k^2 \varphi_k(t)\varphi_i(t)\, dt = \begin{cases} 0, & k \neq i \\ \frac{\sigma_n^2}{T}, & k = i \end{cases}$

Henceforth, we assume that this relation is satisfied, and z_i and z_k are independent random variables with gaussian distributions.

Evaluation of capacity of continuous time channels

Let $u(t) = \sum_{-\infty}^{\infty} x_k \varphi_k(t) \rightarrow \underline{u} = (x_0, x_1, x_{-1}, \ldots)$

$v(t) = \sum y_k \varphi_k(t) \rightarrow \underline{v}$

$n(t) = \sum z_k \varphi_k(t) \rightarrow \underline{n}$

$$I(u;v) = H(v) - H(v|u)$$

$$H(v) = -\int_v p(\underline{v}) \log p(\underline{v})\, dv$$

$$H(v|u) = -\int_u \int_v p(\underline{u})\, p(\underline{v}|\underline{u}) \log p(\underline{v}|\underline{u})\, du\, dv$$

$$p(\underline{v}) = \int_u p(\underline{v}|\underline{u})\, p(\underline{u})\, du = \int_u p(\underline{u})\, q(\underline{v}-\underline{u})\, du$$

$$\underline{u}\cdot\underline{u} = |\underline{u}|^2 = \sum_{k=-\infty}^{\infty} |x_k|^2 = \frac{1}{T}\int_{-T/2}^{T/2} u^2(t)\, dt$$

Constraint: $\int_u |\underline{u}|^2 p(\underline{u})\, du = S = \int |x|^2 p(x)\, dx = \sum_k S_k$

$$I(u;v) \le \sum_{k=-\infty}^{\infty} I(X_k;Y_k) \le \sum_{k=-\infty}^{\infty} \frac{1}{2}\log\left(1+\frac{S_k}{N_k}\right)$$

This is since the best we can do is to make each coordinate statistically independent. Once we have this, we can solve each coordinate separately. Now we max this over S_k:

$$\frac{\partial}{\partial S_k}\left[\sum_k \frac{1}{2}\log\left(1+\frac{S_k}{N_k}\right) + \lambda\sum_k S_k\right] = 0 = \frac{\frac{1}{2}\log e}{S_k+N_k} + \lambda$$

Lumping all the constants, $S_k + N_k = A$ so

$$S = \sum_k S_k = \sum_k A - N_k$$

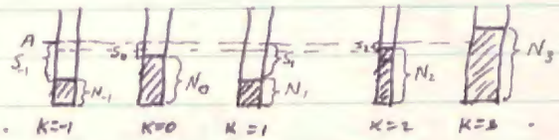
If the solution gives some $S_j < 0$, set $S_j = 0$ & resolve. Eventually we will get a solution such that all $S_k \ge 0$. Then we will have for

the channel capacity

$$C_T = \frac{1}{2} \sum_{K=-\infty}^{\infty} \log\left(1 + \frac{S_K}{N_K}\right)$$

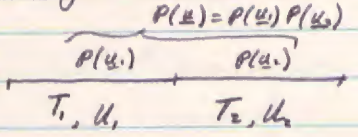
where the S_K are the values after all constraints have been satisfied.

Analogy:



Consider A to be the volume of a water glass, the contents of which we pour into the various containers labeled as channels; these cylinders are already filled with ~~sand~~ solid to level N_K. We pour water first into the cylinders that are lowest in N_K & continue, keeping the water level constant in the cylinders where $N_K <$ water level. When we have used up all A of the water, the resulting height of water in each column is S_K. I.e., we use the lowest noise channels first.

Capacity variation with interval T:



$u = u_1, u_2$

Assuming successive intervals independent, we can write:

$$I(u;v) = H(u) - H(u|v)$$

$$H(u) = H(u_1, u_2) = H(u_1) + H(u_2|u_1) = H(u_1) + H(u_2)$$

$$H(u|v) = H(u_1, u_2 | v_1, v_2) = H(u_1|v_1, v_2) + H(u_2|v_1, v_2, u_2)$$

$$\leq H(u_1|v_1) + H(u_2|v_2)$$

or $$H(u) - H(u|v) \geq H(u_1) + H(u_2) - H(u_1|v_1) - H(u_2|v_2)$$

so $$\boxed{I(u;v) \geq I(u_1;v_1) + I(u_2;v_2)}$$

If $T_1 = T_2 = T$, $P(u_1) = P(u_2)$, $C_{2T} \geq I(u;v) \geq 2C_T$

or $$\boxed{\frac{C_{2T}}{2T} \geq \frac{C_T}{T}}$$

Thus (C_T/T) is a monotonically increasing function of T. We now want to find the limit as $T \to \infty$ of $\frac{C_T}{T}$

Recall $\psi_k(t) = \cos k\varphi t + \sin k\varphi t$, $\varphi = \frac{2\pi}{T}$

$$\frac{1}{T}\int_{-T/2}^{T/2} R(T-\theta)\,\psi_k(\theta)\,d\theta = \frac{\lambda}{T}\psi_k(t)$$

If we now go to the frequency domain, we see

$$N(f)\,\delta\left(f = \tfrac{k}{T}\right) = \lambda\,\delta\left(f = \tfrac{k}{T}\right) \Rightarrow \lambda = N(f), \quad \delta(x) \equiv u_0(x).$$

$$N_k = \frac{\lambda}{T} = \left(\frac{N(f)}{T}\right)_{f=\frac{k}{T}}; \quad S_k = \left.\frac{S(f)}{T}\right|_{f=\frac{k}{T}}$$

Now $\lim_{T\to\infty} \frac{C_T}{T} = \lim_{T\to\infty} \frac{1}{T} \sum_{K=-\infty}^{\infty} \frac{1}{2} \log\left(1 + \frac{S(f)}{N(f)}\right)_{f=\frac{K}{T}}$

Let $V = \frac{K}{T}$, $\frac{1}{T} \to dV$ as $T \to \infty$

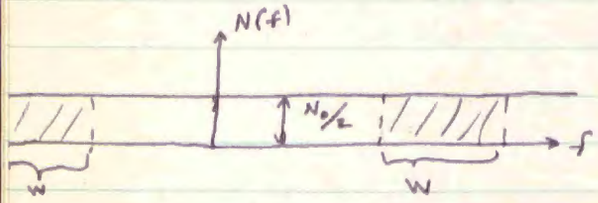
$$\lim_{T\to\infty} \frac{C_T}{T} = \int_{-\infty}^{\infty} \frac{1}{2} \log\left(1 + \frac{S(V)}{N(V)}\right) dV = \int_{0}^{\infty} \log\left(1 + \frac{S(f)}{N(f)}\right) df \equiv C = \text{capacity in bits/sec.}$$

if S is selected in the optimum manner.

$\lim \sum_K S_K = \int_{-\infty}^{\infty} S(f)\,df$ & $S(f) + N(f) = A$,

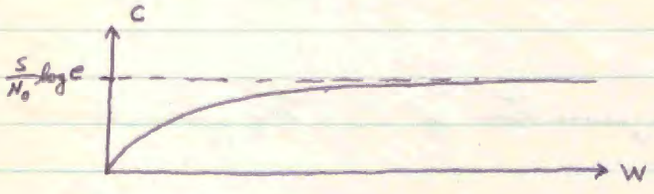
so $S = \int_{-\infty}^{\infty} (A - N(f))\,df$

Channel capacity for white noise:



$$C = W \log\left(1 + \frac{S}{N_0 W}\right)$$

$$\lim_{W\to\infty} C = W(\log e)\frac{S}{N_0 W} = \frac{S}{N_0} \log e \text{ bit/sec} = \frac{S}{N_0} \text{ nit/sec.}$$



Time to transmit one nit $\Rightarrow 1 = \frac{S}{N_0} T \to ST = N_0 \neq E = KT^\circ$ for effective thermal noise temperature.

Coding in the presence of noise:

Sequential encoding: binary digits into encoder one at a time.; $t = \frac{1}{R}$

Block encoding: N digits all go into the encoder at the same time. $t = T = \frac{N}{R}$

Block encoding: $M = 2^N$ possible messages of N binary digits: $m_1, \ldots, m_M$. at the coder input.

Channel input: u possible input symbols
n symbols per message per time T.

Hence, $u^n \geq 2^N = M$

$(m_1, \ldots, m_M)$ = set of coder inputs

$U = (u_1, \ldots, u_{u^n})$ = set of channel inputs $= X^n$

$V = (v_1, \ldots\ \ )$ = set of channel outputs $= Y^n$

We will assign to each $v \in V$ a message m. The set of all outputs v assigned to m_i will form a subset w_i of V. $w_i \in W$; $w_i \subset V$; if $v \in w_i$, we guess that m_i was sent.

We can talk about an over-all channel from m's to w's, which includes the coder, channel, and decoder. Once we decide how to relate $m \leftrightarrow u$ and $v \leftrightarrow w$, we can, however, talk only about the channel $u \to v$.

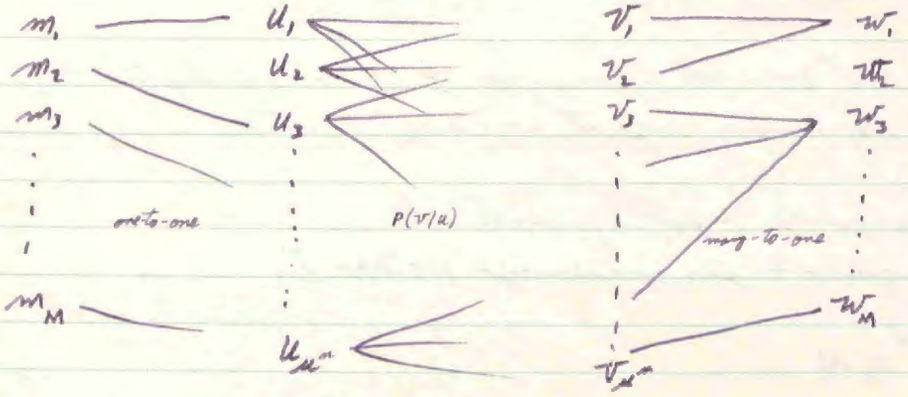
Now, $I(U;V) \leq \begin{cases} nC & \text{for discrete channel} \\ TC & \text{for continuous channel.} \end{cases}$

$\Rightarrow \frac{\log M}{n} \leq C$ bits/symbol (discrete)

$\frac{\log M}{T} \leq C$ bits/sec. (continuous)

n^{th}-power channel model:

L binary digits in the coder at any time $\Leftrightarrow M = 2^L$
Channel input has n symbols per message; u symbols in alphabet



$$P(v/u) = \prod_{i=1}^{n} P(\eta_i / \xi_i) \quad ; \quad v = \eta_1 \eta_2 \eta_3 \cdots \eta_{u^n} \quad ; \quad u = \xi_1 \xi_2 \cdots \xi_{u^n}$$

$$\eta_i \in Y, \quad \xi_i \in X \quad ; \quad u \in U = X^n, \quad v \in V = Y^n$$

Decisions: What m ~~caused~~ caused the v we have received?

(1) $P(m)$ unknown: pick the m which maximizes $P(v/m)$ and guess that this was transmitted.

$$\max_m P(v/m) \Leftrightarrow \text{maximum likelihood}$$

(2) $P(m)$ is known: calculate the a posteriori probabilities

$$P(m/v) = \frac{P(v/m)\,P(m)}{\sum_V P(v/m)\,P(m)}$$

Now select the m which maximizes $P(m/v)$

$$\max_m P(m/v) \Leftrightarrow \text{maximum a posteriori probability} \Leftrightarrow \text{minimum error}$$

Minimum error since $P(e/v) = 1 - P(m/v)$

Using one of these decision rules, we divide V into subsets $W_i \subset V$, such that if $v \in W_i$, we will guess that m_i was transmitted.

Now we can regard the entire process as a single channel:

$$P(w|m) = \sum_{v \in W} P(v|m) = \sum_{v \in W} P(v|u) = \sum_{v \in W} \prod_{i=1}^{n} P(\eta_i | \xi_i)$$

Entropies: $H(U) = H(M)$

~~$I(M;W)$~~ $I(M;W) = H(M) - H(M|W) = H(M) - H(M|VW)$

$I(M;V) = H(M) - H(M|V)$

so $\boxed{I(M;W) \leq I(M;V)}$

Error: $P(e) = \sum_k P(m_k) \sum_{j \neq k} P(w_j|m_k) = \sum_j P(w_j) \sum_{k \neq j} P(m_k|w_j)$

$= 1 - \sum_i P(m_i|w_i) P(m_i) = \sum_j P(w_j) P(e|w_j)$

where $P(e|w_j) = \sum_{k \neq j} P(m_k|w_j)$

$H(e) \equiv -P(e) \log P(e) - [1 - P(e)] \log [1 - P(e)]$

$H(e) \geq H(e|W) = -\sum_W P(w) \left[+ P(e|w) \log P(e|w) + (1 - P(e|w)) \log (1 - P(e|w)) \right]$

$H(e|w) = P(e|w) \log P(e|w) + [1 - P(e|w)] \log [1 - P(e|w)]$

Theorem: $H(M|W) \le H(e|W) + P(e)\log(M-1) \le H(e) + P(e)\log(M-1)$

$H(e|W)$ = average information necessary to tell the decoder that he has erred

$P(e)\log(M-1)$ = average information needed to specify a correct symbol from $(M-1)$ alternatives.

Proof:

$$H(M|W) = \sum_W P(w)\, H(M|w)$$

$$H(M|w_j) = -\sum_K P(m_K|w_j)\log P(m_K|w_j)$$

$$= -P(m_j|w_j)\log P(m_j|w_j) - \sum_{K\neq j} P(m_K|w_j)\log P(m_K|w_j)$$

$$= -[1-P(e|w_j)]\log[1-P(e|w_j)] - \sum_{K\neq j} P(m_K|w_j)\log P(m_K|w_j)$$

~~$= \log[1-P(e|w_j)] - \left[\sum_{K\neq j} P(m_K|w_j)\log \frac{P(m_K|w_j)}{1-P(e|w_j)}\right]$~~

~~since $P(e|w_j) = \sum_{K\neq j} P(m_K|w_j)$~~

$$= H(e|w_j) - P(e|w_j)\sum_{K\neq j}\frac{P(m_K|w_j)}{P(e|w_j)}\log\frac{P(m_K|w_j)}{P(e|w_j)}$$

$$= H(e|w_j) + P(e|w_j)\sum_{K\neq j}\frac{P(m_K|w_j)}{P(e|w_j)}\log\frac{P(e|w_j)}{P(m_K|w_j)}$$

$$\le H(e|w_j) + P(e|w_j)\log(M-1)$$

so $$\boxed{H(M|W) \le H(e|W) + P(e|W)\log(M-1) \le H(e) + P(e)\log(M-1)}$$

Now $nC \ge H(M)$ so $I(M;W) \le I(M;V) \le nC$
& $H(M|W) = H(M) - I(M;W) \ge H(M) - nC$

Combining this & the result of the last theorem gives us

$$0 \le H(e) + P(e)\log(M-1)$$
$$H(M) - nC \le \cdots$$

so if $nC \le H(M)$, then $P(e) \ge 0$

Converse of coding theorem.

Channel with white gaussian noise :

$$N(f) = \frac{N_0}{2} \quad \Rightarrow R(\tau) = \frac{N_0}{2} u_0(\tau)$$

$$\int_{-T/2}^{T/2} R(t-\theta)\,\varphi(\theta)\,d\theta = \lambda \varphi(t) \quad \Rightarrow \quad \lambda = \frac{N_0}{2}$$

$E(z_i^2) = \sigma^2 = \frac{\lambda}{T} = \frac{N_0}{2T}$ = variation of distribution of coefficients in the series expansion of $n(t)$, for any set of orthogonal functions.

Let $u(t) + n(t) = v(t) \leftrightarrow \underline{u} + \underline{n} = \underline{v}$, where the components of these vectors are the coefficients in the series expansion of the time functions.

$$P(\underline{v}|\underline{u}) = P(\underline{n}) = \prod_{i=-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{z_i^2}{2\sigma^2}}$$

$$= A \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=-\infty}^{\infty} z_i^2\right\} = A e^{-\frac{|n|^2}{2\sigma^2}} = A e^{-\frac{|\underline{v}-\underline{u}|^2}{2\sigma^2}}$$

Assume a distribution $P(\underline{u})$ over the set $(\underline{u}_1, \ldots, \underline{u}_M)$. Then

$$P(\underline{v}) = \sum_{K=1}^{M} P(\underline{u}_K)\, P(\underline{v}|\underline{u}_K) = \sum_{K=1}^{M} P(\underline{u}_K)\, A\, e^{-\frac{|\underline{v}-\underline{u}_K|^2}{2\sigma^2}}$$

$$P(\underline{u}_j|\underline{v}) = \frac{P(\underline{u}_j)\,P(\underline{v}|\underline{u}_j)}{P(\underline{v})} = P(\underline{u}_j)\,\frac{e^{-\frac{|\underline{v}-\underline{u}_j|^2}{2\sigma^2}}}{\sum_{K=1}^{M} P(\underline{u}_K)\, e^{-|\underline{v}-\underline{u}_j|^2/2\sigma^2}}$$

Decoding in presence of white noise:

$$(\underline{v}-\underline{u}_i)\cdot(\underline{v}-\underline{u}_i) = |\underline{v}|^2 + |\underline{u}_j|^2 - 2\,\underline{v}\cdot\underline{u}_j$$

$$P(\underline{u}_i|v) = P(\underline{u}_i)\,\frac{e^{\frac{2P_j - E_{jj}}{N_0}}}{\sum_{K=1}^{M} P(\underline{u}_K)\, e^{\frac{2P_K - E_{KK}}{N_0}}}$$

where

$$P_j = (\underline{v}\cdot\underline{u}_j)\,T = \int_{-T/2}^{T/2} v(t)\,u(t)\,dt \quad = \text{cross-correlation of input + output}$$

$$E_{jj} = |\underline{u}_j|^2\,T = \int_{-T/2}^{T/2} u_j^2(t)\,dt$$

Maximum likelihood decoding:

Select subsets w_i by criterion that $\underline{v} \in w_i$ if $\underline{v}$ and $\underline{u}_i$ differ in mean square some less that does $\underline{v}$ and $\underline{u}_j$, $j \neq i$, etc, since this corresponds to a maximum of $P(\underline{v}|\underline{u})$. Or,

$$w_i = \{\underline{v} \mid |\underline{v}-\underline{u}_i|^2 \leq |\underline{v}-\underline{u}_j|^2 \text{ for all } j \neq i\}$$

~~Maximum~~

Minimum error decoding:

$$\ln P(\underline{u}_j|v) = \frac{2P_j - E_{jj}}{N_0} + \ln P(\underline{u}_j) + F(\underline{v})$$

Define $$d_j(\underline{v}) = \frac{2P_j - E_{jj}}{N_0} + \ln P(\underline{u}_j)$$

Then $$\max_j P(\underline{u}_j|\underline{v}) \leftrightarrow \max_j d_j(\underline{v})$$

so

$$w_i = \{\underline{v} \mid d_i(\underline{v}) > d_j(\underline{v}),\ \text{all } j \neq i\}$$

In the special case where $P(u) = \frac{1}{M}$, we have that maximum likelihood and minimum error decoding are the same and

$$E_{jj} = E = ST.$$

Now, if in addition, the $u_j(t)$ are orthogonal over $(-T/2, T/2)$,

$$\int_{-T/2}^{T/2} u_j(t)\, u_k(t)\, dt = \begin{cases} E, & j = k \\ 0, & j \neq k \end{cases}$$

Now let

$$\rho_j(t) = \sqrt{\frac{T}{E}}\, u_j(t) \qquad , 1 \leq j \leq M.$$

Then only one component of the series expansion of $u_j(t)$ is non-zero:

$$x_j = \sqrt{\frac{E}{T}} = \sqrt{S}$$

[$E_{jj} = E$, all j, so $\max_j d_j(\tau) \Leftrightarrow \max_j \rho_j$
$\Rightarrow$ minimum error and maximum likelihood coincide]

$$\rho_j = T\, y_j\, x_j = T\sqrt{S}\, y_j = y_j \sqrt{ET} = z_j \sqrt{ET} + \begin{cases} E, & j = k \\ 0, & j \neq k \end{cases}$$

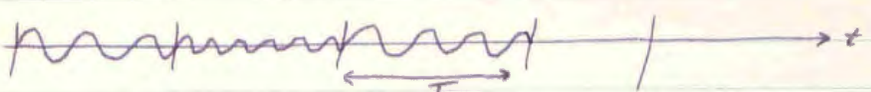
~~$= \sqrt{ET}\, z_j + E$, $j = k$; $= \sqrt{ET}\, z_j$, $j \neq k$; actually transmitted $u_k(t)$~~

$$\rho_j = \sqrt{ET}\, z_j + \begin{cases} E, & j = k \\ 0, & j \neq k \end{cases} \qquad , u_k(t) \text{ actually transmitted}$$

$$y_j = \frac{\rho_j}{\sqrt{ET}} = z_j + \begin{cases} \sqrt{S}, & j = k \\ 0, & j \neq k \end{cases}$$

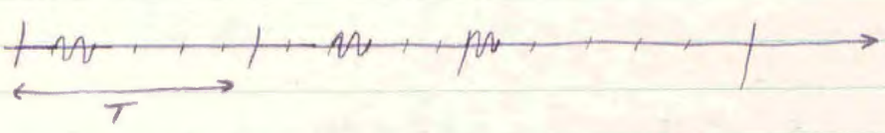
$$y_j = \frac{1}{\sqrt{ET}} \int_{-T/2}^{T/2} v(t)\, u_j(t)\, dt$$

Assume that in the previous example, the $u(t)$ are sinusoids at frequencies $\frac{\nu}{T}$, $\frac{\nu+1}{T}$, ..., $\frac{\nu+M-1}{T}$

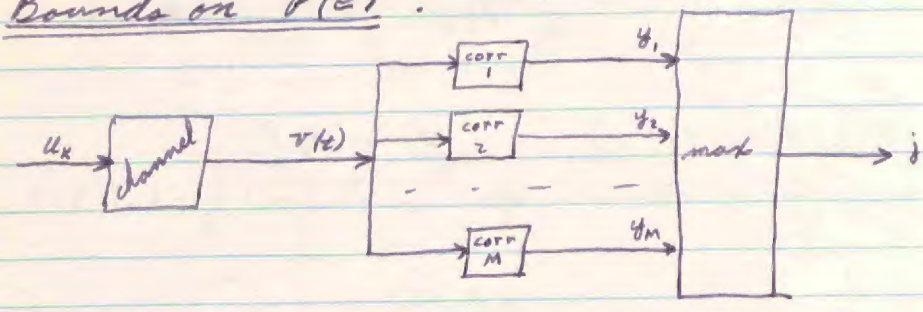


This gives us a crude form of FM.

Or, we could have pulse position modulation of sinusoidal bursts:



Bounds on $P(e)$:



$$\boxed{P(e) = 1 - P\{ y_j < y_k \text{ for all } j \neq k, \text{ all } k \}}$$

$$P(y_j < y_k) = 1 - P(y_j \geq y_k) = 1 - \int_{y_k}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}} dy$$

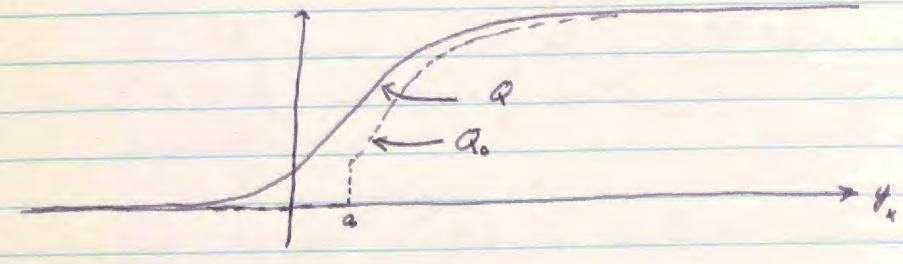
$$Q(y_k) = \prod_{j \neq k} P(y_j < y_k) = \left[1 - P(y_j \geq y_k)\right]^{M-1} = P(\text{all } y_j < y_k)$$

Then $$\boxed{P(e) = 1 - \int_{-\infty}^{\infty} p(y_k)\, Q(y_k)\, dy_k}$$

where $$p(y_k) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_k - \sqrt{s})^2}{2\sigma^2}}$$

$$Q(y_k) = \left[1 - P(y_j \geq y_k)\right]^{M-1} \geq \begin{cases} 1 - (M-1) P(y_j \geq y_k) & \text{if } y_k > a \\ 0, & \text{if } y_k \leq a \end{cases}$$ good if $P(y \geq y_k)$ is small.

or $$Q(y_k) \geq \begin{cases} 0, & y \leq a \\ 1 - M P(y > y_k), & y > a \end{cases} = Q_0(y_k)$$



This is a widely used technique for computing a bound for $P(e)$.

$$P(e) < 1 - \int_{-\infty}^{\infty} P(y_k)\left[1 - MP(y \geq y_k)\right] dy_k = 1 - \int_{a}^{\infty} P(y_k)\left[1 - MP(y \geq y_k)\right] dy_k$$

$$P(e) < \int_{-\infty}^{a} P(y_k)\, dy_k + M \int_{a}^{\infty} P(y_k)\, P(y \geq y_k)\, dy_k = P_1 + M P_2$$

$$P(e) < P_1 + M P_2$$

$P_1 \leftrightarrow y_k < a$: give up and call it an error

$P_2 \leftrightarrow y_k \geq a$: we may or may not have an error

Computation of P_1 and P_2 :

$$P_1 = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(y_k - \sqrt{5})^2}{2\sigma^2}} dy_k = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{a-\sqrt{5}}{\sigma}} e^{-\frac{t^2}{2}} dt \quad , \quad t = \frac{y_k - \sqrt{5}}{\sigma}$$

We are interested in the case where $a < \sqrt{5}$, or

$$P_1 = \frac{1}{\sqrt{2\pi}} \int_{\frac{\sqrt{5}-a}{\sigma}}^{\infty} e^{-\frac{t^2}{2}} dt$$

Now for a gaussian, we have for a bound on the area under the tail :

$$\frac{1}{\sqrt{2\pi}} \int_{x}^{\infty} e^{-t^2/2} dt \leq \frac{1}{\sqrt{2\pi}\, x}\, e^{-\frac{x^2}{2}} \quad , \quad x > 0$$

so

$$P_1 \leq \frac{\sigma\, e^{-\frac{(\sqrt{5}-a)^2}{2\sigma^2}}}{\sqrt{2\pi}\,(\sqrt{5} - a)} \quad , \quad a < \sqrt{5}$$

$$P_2 = \int_a^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_x-\sqrt{5})^2}{2\sigma^2}} P(y>y_x)\,dy_x$$

$$P(y>y_x) = \int_{y_x}^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}} dy \le \frac{\sigma}{\sqrt{2\pi}\,y_x} e^{-\frac{y_x^2}{2\sigma^2}} \le \frac{\sigma}{\sqrt{2\pi}\,a} e^{-\frac{y_x^2}{2\sigma^2}} \quad \text{if } y_x \ge a$$

$$\text{or} \quad P_2 < \int_a^\infty \frac{1}{2\pi a} \exp\left\{-\frac{(2y_x-\sqrt{5})^2+5}{4\sigma^2}\right\} dy_x$$

$$P_2 < \frac{\sigma}{\sqrt{2\pi}\,a} e^{-\frac{5}{4\sigma^2}} \int_a^\infty \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{\frac{1}{2}(2y_x-\sqrt{5})^2}{2\sigma^2}} dy_x$$

Now let $t = \dfrac{\frac{1}{\sqrt{2}}(2y_x-\sqrt{5})}{\sigma} \rightarrow dt = \dfrac{\sqrt{2}}{\sigma} dy_x$

Then $$P_2 < \frac{\sigma}{\sqrt{4\pi}\,a} e^{-\frac{5}{4\sigma^2}} \int_{\frac{2a-\sqrt{5}}{\sigma\sqrt{2}}}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

or

$$MP_2 < \begin{cases} \dfrac{\sigma^2}{2\pi a(2a-\sqrt{5})} e^{RT-\frac{(2a-\sqrt{5})^2+5}{4\sigma^2}}, & \frac{\sqrt{5}}{2} < a \\ \dfrac{\sigma}{\sqrt{4\pi}\,a} e^{RT-\frac{5^2}{4\sigma^2}}, & 0 < a \le \frac{\sqrt{5}}{2} \end{cases}$$

where $RT \equiv \ln M$ or $R \equiv \dfrac{\ln M}{T}$

We now want to select a so as to minimize $P_1 + MP_2$ and thereby obtain as low an upper bound as possible.

To minimize the sum of two exponential terms, we can equate the two for an approximation. We are justified in doing this since the coefficients vary so slowly compared to the exponentials.

Equating the exponents of P_1 and MP_2 gives

$$-\frac{(\sqrt{5}-a)^2}{2\sigma^2} = RT - \frac{(2a-\sqrt{5})^2+5}{4\sigma^2} \quad , \quad \frac{\sqrt{5}}{2} < a < \sqrt{5}$$

$$-\frac{(\sqrt{5}-a)^2}{2\sigma^2} = RT - \frac{5}{4\sigma^2} \quad , \quad 0 < a \le \frac{\sqrt{5}}{2}$$

Now we substitute $\boxed{\sigma^2 = \frac{N_0}{2T} \leftrightarrow \text{white noise}}$

The dependence on T cancels out, and we get

$$a = \begin{cases} \sqrt{RN_0} = \sqrt{5}\sqrt{\frac{R}{C}} & , \ \frac{\sqrt{5}}{2} < a < \sqrt{5} \text{ or } \frac{1}{4} < \frac{R}{C} < 1 \\ \sqrt{5}\left[1 - \frac{1}{\sqrt{2}}\sqrt{1-2\frac{R}{C}}\right] & , \ 0 < a \le \frac{\sqrt{5}}{2} \text{ or } 0 \le \frac{R}{C} \le \frac{1}{4} \end{cases}$$

Then

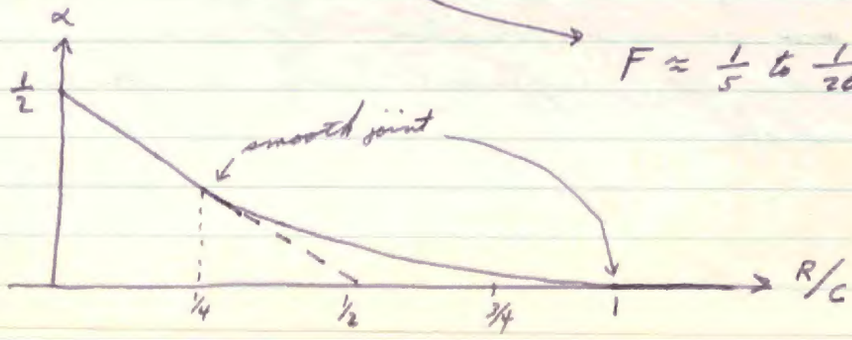
$$P(e) < \frac{1}{\sqrt{4\pi CT}} \begin{cases} \left[\frac{1}{1-\sqrt{R/C}} + \frac{1}{\sqrt{4\pi RT}\left(2\sqrt{\frac{R}{C}}-1\right)}\right] e^{-TC\left(1-\sqrt{\frac{R}{C}}\right)^2} & , \ \frac{1}{4} < \frac{R}{C} < 1 \\ \left[\sqrt{\frac{2}{1-2\frac{R}{C}}} + \frac{1}{\sqrt{2}-\sqrt{1-2\frac{R}{C}}}\right] e^{-\frac{TC}{2}\left(1-2\frac{R}{C}\right)} & , \ 0 < \frac{R}{C} \le \frac{1}{4} \end{cases}$$

Given R, C, the coefficients vary so slowly that the exponent govern the behavior of $P(e)$ with T.

We can always find a T sufficiently large that $P(e)$ can be made arbitrarily small for any C, R. $(\frac{R}{C} < 1)$

Let $P(e) < F(R,C,T)\, e^{-\alpha CT}$, $\alpha = \begin{cases} \left(1-\sqrt{\frac{R}{C}}\right)^2 & , \ \frac{1}{4} < \frac{R}{C} < 1 \\ \frac{1}{2}\left(1-2\frac{R}{C}\right) & , \ 0 < \frac{R}{C} \le \frac{1}{4} \end{cases}$

$F \approx \frac{1}{5}$ to $\frac{1}{20}$



To make $P(e)$ small, we make (αTC) large by doing:

(1) make αTC large $\Rightarrow \frac{R}{C}$ large $\Rightarrow$ time allotted to one message is large.

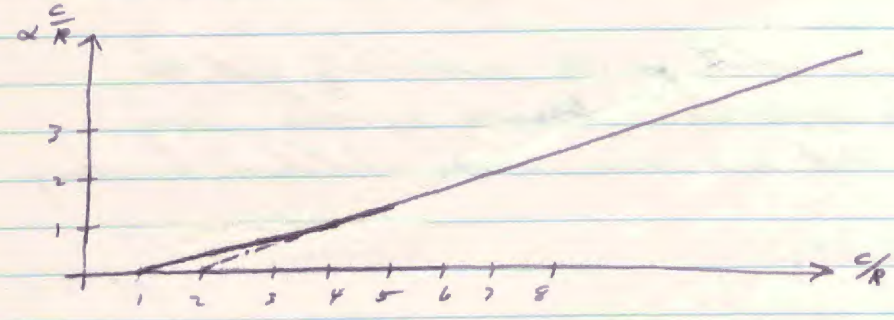
(2) make α large $\Rightarrow \frac{R}{C}$ small $\Rightarrow$ inefficient use of channel.

Assume each message has L binary digits: $M = 2^L$ or $\ln M = L \ln 2$

$$\alpha CT = \left(\alpha \frac{C}{R}\right) L \ln 2$$

$$e^{-\alpha CT} = 2^{-\left(\alpha \frac{C}{R}\right) L}$$

We want $\left(\alpha \frac{C}{R}\right) L$ to be large:



For high efficiency in channel, L is very large & 2^L very large.
Here, a low $P(e)$ is bought at the cost of terminal complexity which goes as 2^L.
Low complexity will cost inefficient usage of channel.

Example: FSK: $L = 1$; assume $\alpha \frac{C}{R} = 10 \Leftrightarrow$ extremely inefficient use of channel.

If $L = 4$, $\frac{C}{R} = 7$ → still inefficient

$L = 10$, $\frac{C}{R} = 4$ → still not good.

It would be very desirable to show how complexity might go linearly with L rather than exponentially as it does here with correlation detection.

Random coding:

Since the number of possible code words $u \in \mathcal{U}$ is larger than or equal to the number of messages $m \in M$ we are going to send, we need some way of assigning to each m_i a particular u_j. No way has yet been found to optimize this mapping, so we use a random mechanism to select the code words.

$$u = \xi_1, \ldots, \xi_n \quad ; \xi_i \in X \quad , \quad P(x) \text{ over } X$$

$$P(u) = P(\xi_1) P(\xi_2) \cdots P(\xi_n)$$

The resulting code probably won't be optimal, but won't be too bad either as the difference from code to code in $P(e)$ is not great.

Given a coding scheme, what is the probability that the decoder will decode correctly? Assume equally likely messages $\left[P(m_u) = \frac{1}{M}\right]$, and use maximum likelihood decoding.

$$P(v/u) = \prod_{i=1}^{n} P(\eta_i / \xi_i)$$

We make an error when $P(v/u) \leq P(v/u')$, $u' \neq u$, and u was transmitted. Or, when

$$I_{uv}(u') \equiv \log \frac{P(v/u')}{P(v/u)} \geq 0$$

$$P_r\{P(v/u') \geq P(v/u)\} = P_r\{I_{uv}(u') \geq 0\}$$

Define $$Q(u,v) = \left[1 - P(I_{uv}(u') \geq 0)\right]^{M-1}$$

This is the probability that we have correct decoding, given a u and v, evaluated over the ensemble of mappings.

$$\overline{P(e)} = 1 - \sum_{u} \sum_{v} P(u) P(v/u) Q(u,v).$$

$\therefore$ There must be a mapping such that $P(e) \leq \overline{P(e)}$

Theorem: $P_r\{P(e) \geq K\overline{P(e)}\} \leq \frac{1}{K}$, K = integer ≥ 1.

Proof: Define:

$$A_K = \{a \mid P(e|a) \geq K\overline{P(e)}\}$$

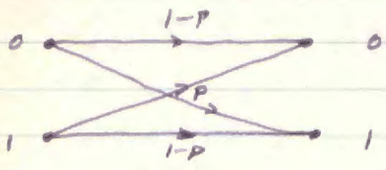
$$\overline{P(e)} = \sum_A P(e|a)P(a) \quad , \quad a \leftrightarrow \text{a particular map } M \to U.$$

Now, $\overline{P(e)} \geq \sum_{A_K} P(e|a)P(a) \geq K\overline{P(e)} \sum_{A_K} P(a) = K\overline{P(e)}\, P_r\{P(e) \geq K\overline{P(e)}\}$

so $1 \geq K P_r\{P(e) \geq K\overline{P(e)}\}$

or $\boxed{P_r\{P(e) \geq K\overline{P(e)}\} \leq \frac{1}{K}}$ or in general, $\boxed{P(x \geq K\bar{x}) \leq \frac{1}{K}}$

Calculation of $\bar{P}(e)$ for binary symmetric channel:



$$C = 1 + P \log P + (1-P) \log (1-P)$$

$$P \le \tfrac{1}{2}$$

$$P(u) = \left(\tfrac{1}{2}\right)^n \quad ; \quad P(v/u) = P^k (1-P)^{n-k}$$

$i \equiv$ # of digits in which v differs from u'

$K = \cdots \; - \; - \; - \; - \; - \; - \; - \; - \; - \; - \; . \; u$

$$I_{uv}(u') = \log \frac{P^k (1-P)^{n-k}}{P^i (1-P)^{n-i}} = (i-k) \log \frac{1-P}{P}$$

$$P_r\{I_{uv}(u') \ge 0\} = P_r\{i-k \ge 0\} = \left(\tfrac{1}{2}\right)^n \sum_{i=0}^{k} \frac{n!}{i!(n-i)!} = \left(\tfrac{1}{2}\right)^n V_k$$

Define $\quad Q_k \equiv \begin{cases} 1 - M\left(\tfrac{1}{2}\right)^n V_k, & k < r \\ 0, & k \ge r \end{cases}$

Then $Q_k \le Q(u, v)$

$$\overline{P(e)} \le 1 - \sum_{k=0}^{n} P^k (1-P)^{n-k} \frac{n!}{k!\,(n-k)!} Q_k$$

$$\le 1 - \sum_{k=0}^{r-1} P^k (1-P)^{n-k} \frac{n!}{k!(n-k)!} \left[1 - M\left(\tfrac{1}{2}\right)^n V_k\right]$$

$$\le \sum_{k=r}^{n} P^k (1-P)^{n-k} \frac{n!}{k!(n-k)!} + M\left(\tfrac{1}{2}\right)^n \sum_{k=0}^{r-1} V_k P^k (1-P)^{n-k} \frac{n!}{k!(n-k)!}$$

$$\le P_1 + \left(\tfrac{1}{2}\right)^n M P_2$$

We can bound this with a power series

$$\overline{P(e)} \leq \begin{cases} K_1 e^{-n\alpha} \\ K_2 e^{-n\alpha} \end{cases}, \quad K_1, K_2 < 2$$

where $$\alpha = \begin{cases} \rho \ln \frac{\rho}{\theta} + (1-\rho) \ln \frac{(1-\rho)}{1-\theta} & , \rho = \frac{r}{n} < \rho_c \\ \rho_c \ln \frac{\rho_c}{\theta} + (1-\rho_c) \ln \frac{1-\rho_c}{1-\theta} + [C(\rho_c) - R] \end{cases}$$

where $$\rho_c = \frac{\sqrt{\theta}}{\sqrt{\theta} + \sqrt{1-\theta}},$$

$$C(\rho) = 1 + \rho \ln \rho + (1-\rho) \ln (1-\rho)$$

$$R = \frac{\log M}{n}, \quad R = C(\rho) \text{ for } \rho < \rho_c$$