情報理論、Claude E. Shannon 再読

Information theory, a revisit to the work of C. E. Shannon

2002／4／17　　市吉　修
2004/04/16 －６／１１OsI

The Bell System Technical Journal, Vol.XXVII July,1948 No.3
“A Mathematical Theory of Communication”

1.　Introduction

先行技術,Predecessors

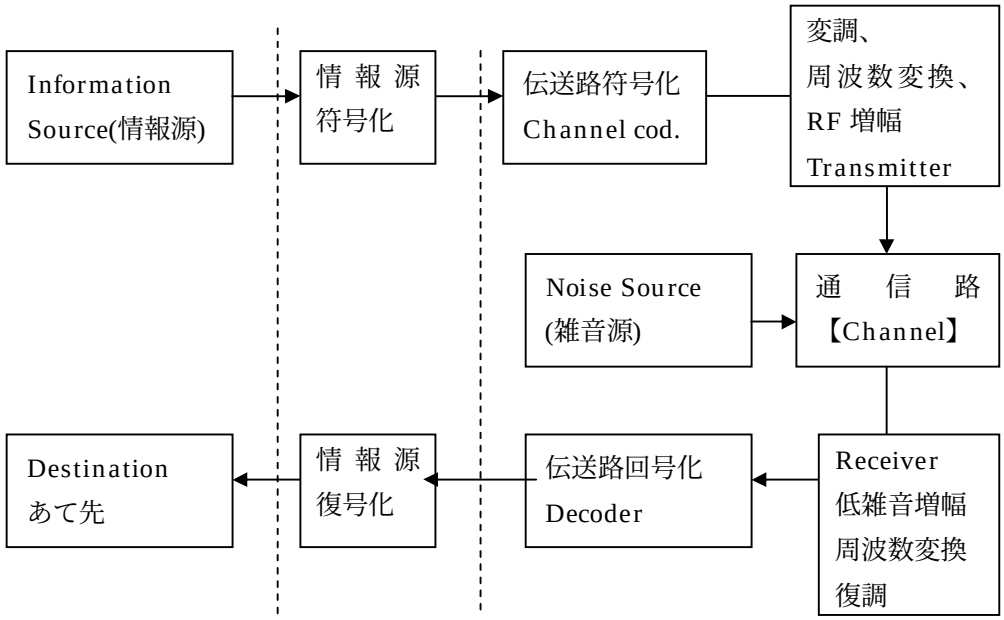
（１）Nyquist（1924）

（２）Hartley【1928】

情報量の定義として

ある通報の運ぶ情報量＝Log｛１／（その通報の生起確率)｝

通信回線の定義



文字、音、絵、動画　　二進符号列
(電報の場合、dot / dash)

情報通信回線の一般構成

2　情報量と情報エントロピー　Information Quantity and Entropy

情報源,Information source

符号(文字)の集合　{Li ;　ｉ＝１，２,,,,　ｎ}

符号 Li の生起確率 ;　　Pi

当然　　［ｉ］ΣPi　＝　１

情報量の定義

ある通報(文)S の運ぶ情報量 I(S)の定義

I(S) = Log(1/P(S)) = -Log(P(S))

P(S) ; 通報 S の出現確率

これは通報の「意味」は全く捨象して事象の生起確率のみで定義される抽象的な定義である。

P(S) =1 ➔ I(S) = 0

P(S) > P(T) ➔ I(S) < I(T)

即ち意外性の高い通報ほど情報量が大である事は直感的にも合理的である。

情報量の単位

上の対数の底は通常 2 に取る。

すると P(S) = 1/2 なる通報の運ぶ情報量は

I(S) = -Log(P(S)) = 1.0 (bit)　　Bit;　Binary Unit

対数の底を自然対数に取ることもできる。この場合の単位を Nat; Natural Unit と呼ぶ。

例　　P(S) = 1/4 なら　　I(S) = 2 (bits)　　＝1.39 (nats)

P(S) = 1/3 なら　　I(S) = 1.6 (bits)　= 1.10 (nats)

長さＮ文字から成る通報(文) S の運ぶ情報量

N が非常に大きい場合にはその文の中に含まれる各文字 Li の数は典型的に N .Pi (i = 1,2,,,n)となる。そうでない非典型的な文の出現確率は N が大きくなると急速に小さくなるので無視できる。

上の典型的な文の数 M(N)は

M(N) = N! / {(N .P1)! .(N .P2)! ....　(N .Pn)!}

これらの文は等確率で生起するので、通報 S の運ぶ情報量 I(S)は

$$I(S) = \mathrm{Log}(1/M(N))$$
$$= N \cdot H$$

ここで

$$H = -\sum_{i=1,n} Pi \cdot \mathrm{Log}(Pi)$$

これは次のように情報量の性質にかなっている。

(1) ある文の伝える情報量はその文の長さに比例する。

(2) H は上の定義式から分かるように一文字当たりの平均情報量と解釈できる。従って N 文字から成る文の運ぶ情報量が N .H で与えられるのは極めて自然。

情報源の Entropy

上の H は情報源から生起する符号が運ぶ平均情報量である。

$$H = -\sum_{i=1,n} Pi \cdot \mathrm{Log}(Pi)$$
$$\sum_{i=1,n} Pi = 1$$

Shannon は上の H が統計熱力学の Entropy の定義式と同じ形である事から情報エントロピーと名づけた。

これは非常に示唆に富む名称であるが拡大解釈も生じやすい。
誤解を生じないためには情報源の符号当たりの平均情報量と呼ぶ事を薦めたい。

情報 Entropy の性質；

$$H[P1, P2,,,,Pn] = \langle \log(1/Pi) \rangle = [i] \sum Pi \cdot \mathrm{Log}(1/Pi)$$

Properties

（１） Ｈ＝０となるのはＰｉ＝１，　Ｐｊ＝０（ｊ=/= ｉ）の場合のみ

（２） Hが最大になるのはすべての符号の生起確率が等しくなる場合

Pi =1/n (i=1,2,,,n)である。この場合の情報 Entropy は

$$H=\mathrm{Log}(n)$$

(3)　相加法則　,Additive law;

$$PA = P[1] + P[2] +,,,+ P[k]$$
$$PB = P[k+1] + P[k+2] + ,,,,, P[n]$$
$$pa[i]=Pi / PA \qquad (i=1,2,,,,,k)$$
$$pb[i]=Pi / PB \quad (i=k+1, k+2,,,,,n)$$

この時

$$H[P1,P2,,,,Pn] = H[PA,PB] + PA \cdot H[\, pa[1], pa[2],,,,pa[k] \,]$$
$$PB \cdot H[\, pb[k+1], pb[k+2],,,,,pb[n] \,]$$

同時確率　Simultaneous Probability

Two probabilistic variables; x, y

P(i, j) = Probably{ x = i, y = j}

＝ P(i) .Pi(j)

Pi(j) ;条件付確率(Conditional Probability)

; x = i なる条件の下で y = j となる確率。

P(i) = [j]∑ P(i,j) = [j]∑ P(j) .Pj(i)

独立事象 Statistical Independence

x と y が独立な確率変数であれば

P(i,j) = P(i) .P(j)

同時確率分布の各種情報 Entropy;

$$H(x) = -[i, j]\Sigma P(i, j).\log(P(i))$$

$$Hx(y) = -[i, j]\Sigma P(i, j).\log(Pi(j))$$

$$H(x, y) = -[i, j]\Sigma P(i, j).\log(P(i, j))$$

この時

$$H(x, y) = H(x) + Hx(y)$$

$$= H(y) + Hy(x)$$

また

$$H(x, y) =< H(x) + H(y)$$

$$H(y) >= Hx(y)$$

上の不等式が等式になるのは確率変数 x と y が独立な場合だけ。

$H(y) >= Hx(y)$

この不等式は右辺が確率変数 x からの予備知識の上に、左辺は何ら予備知識無しに確率変数 y が運ぶ平均情報であると解釈されるので極めて合理的。

3 情報源符号化　Information Source Coding

情報源 S

Symbols; {Si　;　i=1,2,,,,n }

Associated Probabilities; {Pi ;　[i=1,n]∑ Pi = 1}

各符号 Si を二進符号に変換する。

情報源符号化の目的; 生起確率の大きな情報源符号には短い伝送符号を割り当てることによって情報の伝送効率をできるだけ大きくすること。The objective of source coding is to maximize the transmission efficiency by assigning the shorter transmission codes for the source symbols with the higher frequency probabilities.

二分木 Binary tree

二進符号は０か１の系列であるので二分木 Binary tree で表す事ができる。

0	0	0	0	００００
			1	０００１
		1		００１
	1			０１
	0			１０
	1	0	0	１１００
			1	１１０１
		1	0	１１１０
			1	１１１１

上の下線部は葉(Leaves)という。

各葉に符号を割り当てれば自然に符号の区別がつく。すなわち区切りが不要である。日本語の「言葉」に一脈通ずるものを感じる。

If the symbols are coded to the leaves, then comma-free codes are achieved.

クラフトの不等式

各符号 Si の符号長を li(bits)とすると　Let li be length(bits) of symbol Si, then;

[i=1,n]∑ 2^-li　= < 1

符号の平均長　Average number of length of Symbols

L = [i=1,n]∑ Pi .li

情報源符号化定理(個々の符号) Source Coding Theorem of each symbol

L > = H[**P**] = H[P1,P2,,,,Pn]

証明 Proof

自然対数の不等式; Ln(x) =< x – 1

において x = qi / pi

と置く。但し qi = 2^-li

すると Log (qi / pi) =< (qi / pi – 1) .Log(e)

両辺に Pi をかけて i について和をとると

－[i=1,n]∑Pi .li ― [i=1,n]∑Pi .Log(Pi) =< ([i=1,n]∑2^-li – 1) .Log(e)

クラフトの不等式より

－L + H[**P**] =< 0

故に L >= H[**P**]

Shannon の符号化法 Shannon's Source Coding Method

符号の順番を生起確率の大きい順番に並べる。

P1>=P2>=,,,,>=Pn

積算確率 Accumulated Probability AP(j)

AP(i) = [j=1, i-1]∑Pj

そこで次の方法によって i 番目の符号 Si は長さ li(bits)の符号を割り当てる。

(1) li の決定 ; Log（1/Pi） =< li < Log(1/Pi) +1

即ち 2^-li =< Pi < 2 .2^-li

(2)符号の割り当て法

AP(i)の二進数展開の小数点下 li ビットを Si の伝送符号とする

上の方法で Comma-free 符号が生成できる事は二進展開で Pi=0.0,,01xyz,,の形をしており最初の 1 が li 桁にあるので AP(i)はそれ以外の符号とは少なくとも li 番目のビットが異なる事と AP(1)=0.000,,,から AP(n)=0.11111,,,まで AP(i)は i について単調増加関数である事から Binary tree の leaves に符号が割り当てられることから分かる。

Shannon の符号化法の性能 Efficiency of Shannon's Source Coding Method

上の(1)に Pi をかけて i について総和を取る(平均を計算)と

[i=1,n]∑Pi .Log(1/Pi) =< [i=1,n]∑Pi .li <[i=1,n]∑{ Pi .Log(1/Pi) + Pi }

定義より

H =< L < H + 1

情報源符号化定理(一括符号) Source Coding Theorem on Block of symbols

情報源 S➔情報源 S^m ;新たな情報源を作る。

S = {S(i) ; i=1,2,,,n } 符号 S(i)の生起確率;P(i); ; [i=1,n]∑P(i) = 1

➔ S^m={S(i1,i2,,,,im)} = {S(i1)-S(i2)---S(im)}

符号　S(i1,i2,,,,im)の生起確率; P(i1,i2,,,,,im)　; [i=1,n ;j=1,m] ∑ P(ij) =1

S^m に対して Shannon の符号化法を適用するとその平均符号長 Lm は

H(S^m) =< Lm < H(S^m) +1

情報源 S の符号は独立に生起するものとすると

P(i1,i2,,,,im) = P(i1) .P(i2) …..P(im)

この時

H(S^m) = m .H

となる事が示せる。Lm=m .L であるから上の式より

H　=<　L　<　H + 1/m

となる。

誤りの生じない通信路に対する情報源符号化定理

毎秒 C(bits)の速度で誤り無しにデータ伝送できる通信路に上の情報源符号化回路を接続すると毎秒Ｒ＝Ｃ／Ｌ（符号）の速度で信号伝送ができる。

Ｃ／Ｈ　＝＜　Ｒ　＜　Ｃ／（Ｈ　＋　１／ｍ）

m→∞　に対して　Ｒ→Ｈ

ＦＡＮＯの符号化法

これは二分枝法で次のように符号化を行う。

（１）符号の順番を生起確率の大きい順番に並べる。

P1>=P2>=,,,,>=Pn

（２）上の符号を二つのグループに分け一方に０，他方に１をつける。

このとき二つのグループの確率の総計ができるだけ等しくなるように分ける。

境目がｋ番目の符号とすると［ｉ＝１，ｋ］∑Ｐｉと［ｉ＝ｋ＋１，ｎ］∑Ｐｉができるだけ等しくなるようにｋを選ぶ。

（３）　それぞれのグループについて同様な事を繰り返す。

（４）　上の動作を各グループについて含まれる符号が一個になるまで繰り返す。各段階でつけられた０，１の列を各符号の二進伝送路符号とする。

例

源符号	Ｐｉ	Step1	Step2	Step3	Step4	結果としての二進符号
Ｓ１	0.2	０（0.57）	０（0.2）			00
Ｓ２	0.19		１（0.37）	０		010
Ｓ３	0.18			１		011
Ｓ４	0.17	１（0.43）	０（0.17）			10
Ｓ５	0.15		１（0.26）	０（0.15）		110
Ｓ６	0.10			１	０	1110
Ｓ７	0.01			0.11	１	1111

Huffman の符号化法

これは上の二分枝法とは反対に二合枝法で次の方法で符号化を行う。

Ｓｔｅｐ（1）

符号の順番を生起確率の大きい順番に並べる。

P1>=P2>=,,,,>=Pn

これらは皆符号Ｔｒｅｅの葉である。

Ｓｔｅｐ（2）

上の葉の一番確率の小さな二つに０，１の符号を貼り付ける。

同時に二つを合わせて新たな葉をつくる。その葉に付随する確率はもとの確率を加えたものとする。

Ｓｔｅｐ（3）

Ｓｔｅｐ2で作った新たな葉の集合を確率の大きい順に並べる。

Ｓｔｅｐ（1),（2),（3）を繰返し最後に葉が一枚になるまで続ける。

Ｓｔｅｐ（4）

上の過程を逆にたどりもとの源符号にもどる。この際上のＳｔｅｐ（2）で貼り付けた０，１の系列が各源符号の二進符号となる。

問題 Exercise

（1）上の例に Huffman の符号を適用して符号化を行い，Fano の符号化と比較せよ。

Apply Huffman's method to the previous example and compare the results with that by Fano's method.

（3）一般に Fano 法よりも Huffman 符号化のほうが効率が良い。その理由を述べよ。

State the reason Huffman's method is generally more efficient than Fano's method.

Ｒｕｎ－Ｌｅｎｇｔｈ符号

白い紙の上に文字や絵を描いた白黒図形を掃引符号化するＦＡＸ等においては引き続く白の系列が非常に長い。パターン黒白白白、、、白白黒の中で白の数がｎ個あるものを長さｎの Run と呼ぶ。黒の生起確率をｐとすると長さｎのＲｕｎの生起確率Ｐｎは(1-p)^n .p となる。Ｒｕｎの最大の長さをNに制限すればＲｕｎの平均長<n>={1-(1-p)^(N-1)}/p となる。ｎ＝１，２，３,,,，NなるＲｕｎの集合をＦａｎｏの符号化法で符号化すると最大の長さの符号長でもＬｏｇ（N）となる。ｐが小さい場合にはＲｕｎの平均長＜ｎ＞はNに近いのでNを大きくとると大幅な情報圧縮が可能となる。

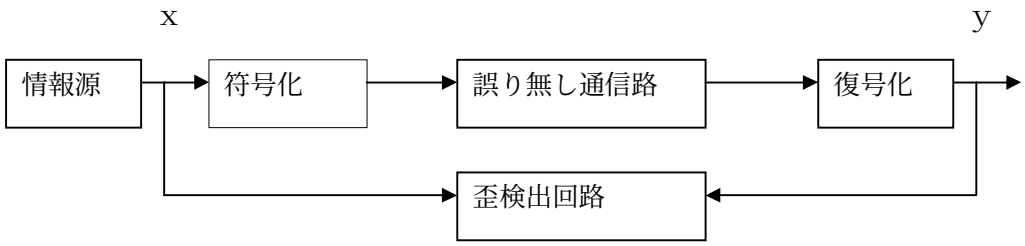
実際のＦＡＸでは Huffman 符号化を用いたものが Run-length 符号とてして採用されている。

4 Rate-Distortion theory

情報源符号化定理の意味はある情報源の出力符号を歪み無しに二進符号に変換する場合に二進符号の平均長には下限があり、それが情報エントロピーであるということである。The entropy of an information source is defined as the lower limit in the average length of the binary codes mapped from the source symbols by any method with the constraint of no distortion introdeced by the coding.

そこで条件を緩めてある程度の歪の発生を許せば平均符号長を小さくできることが期待される。それを以下に検討する。

通信路モデル



歪の定義

上のｘに対して復号された信号をｙとすると符号化歪 d(x,y)は種々の方式で定義される。

例

計数誤差

d(x,y) = 0 (x=y)

= 1 (x=/=y)

<d>は通常のビット誤り率になる。

２乗誤差

d(x,y) = (x - y)^2

相互情報量

情報源Ｘのエントロピー H(X)と復号情報の集合Ｙのエントロピー H(Y)及びＸとＹの同時確率 P(x,y)集合のエントロピー H(X,Y)の関係は

H(X,Y) = H(X) + Hx(Y)

= H(Y) + Hy(X)

相互情報量 I(X:Y)は次式によって定義される。

I(X:Y) = H(X) – Hy(X)

= H(Y) – Hx(Y)

歪が無い可逆的な符号化においては

Py(x)=1(x=y),0(x=/=y)であるから Hy(X)=0 となる。即ちｙを知ればｘに関する曖昧度は０となる。

I(X:Y)の意味；
歪がある場合には Hy(X)は０とはならない。Hy(X)は受信部で受信信号を復号した上でなお残る情報源の符号当りの曖昧度と考えられる。他方 H(X)は情報源Ｘの符号当りの平均曖昧度である。これから I(X:Y)の意味は送信部から受信部に伝達される符号当りの曖昧度、即ち情報伝送量であると考えられる。

歪関数 Distortion Function

平均歪　＜ｄ＞ ＝[i,j]∑ ∑d(i,j).P(i,j)

与えられた歪の上限Ｄに対して平均歪<d> ＜Ｄなる条件を満たすあらゆる符号化法に対応する相互情報量 I(X:Y)の下限を歪関数 Distortion Function と呼び R(D)で表す。

二進誤り符号化の場合の歪関数 Distortion Function
情報源符号化回路が 0,1 を生起確率 p，1-p で発生し復号後には確率ｅでランダム誤りが生ずる符号化について考える。

Py(x) ;　P0(0) = P1(1) = 1-e,　P0(1)=P1(0) = e

これより Hy(X) = h(e) = -e.Log(e) – (1-e).Log(1-e)
平均歪＜ｄ＞＝ｅであるから

R(D) = h(p) – h(D)

となる。

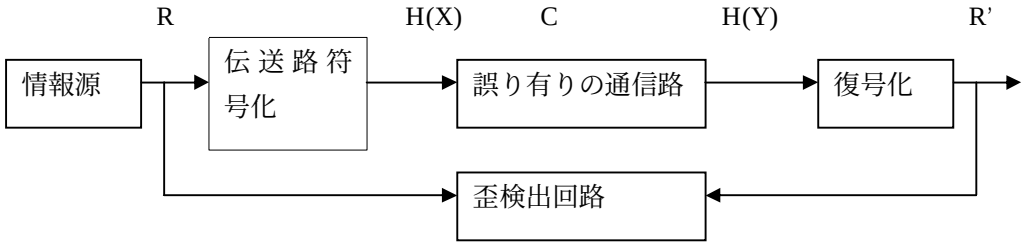
歪と符号化長
ここで情報エントロピーの意味は符号当りのビット単位情報量であり、具体的には二進符号の長さに相当する。許容歪Ｄに対して歪み関数

R(D) =Min { H(X)-Hy(X)} = H(X) – Max Hy(X)　(<d> =< D)

の意味は歪を許容すれば符号化後の二進符号長を短くできる事である。即ち歪みを許さなければ平均符号長の下限が H(X)であるが、歪を許せば R(D)となる。

５　誤りのある通信路を通じた通信　Communication through Channel with Errors

通信路モデル



定義

Ｒ　　；情報源の情報発生速度

Ｈ（Ｘ）；符号化回路出力の情報発生速度

Ｃ　　；通信容量、Communication Capacity

H(Y)　；受信信号の情報受信速度

単位はいずれも Bits/sec である。

相互情報量

前述の相互情報量 I(X:Y)は次式によって定義される。

I(X:Y)　= H(X) – Hy(X)

= H(Y) – Hx(Y)

誤りが無い通信路においては

Py(x)=1(x=y),0(x=/=y)であるから Hy(X)=0 となる。即ちｙを知ればｘに関する曖昧度は０となる。同様に Hx(Y)=0 となる。

現実には通信路で誤りが生ずるので Hx(Y)も Hy(X)も０にはならない。Hx(Y)はｘなる信号を送信した時に受信側で受けた信号がｘになるとは限らずそれに誤りが加わるために生じる散布度(Dissemination)（情報エントロピー）である。Hy(X)はｙなる信号を受信しても通信路で生じる誤りのために送信側の符号を決定できない曖昧(Equivocation)（情報エントロピー）である。

このことから相互情報量 I(X:Y)は通信路を通じて伝送される単位時間あたりの情報量であると考えられる。

通信容量 Communication Capacity

通信容量＝伝送路符号化法を工夫して通信路を通じて伝送できる最大の情報量

Ｃ　＝　Ｍａｘ　Ｉ（Ｘ：Ｙ）

以下ではＨ（Ｘ）がそのような最適符号化法の出力における情報エントロピーであるものとする。

伝送路符号化定理 Channel Coding Theorem

上図の通信系において情報源の情報発生速度ＲがＲ＜Ｃであれば任意の値よりも小さな誤り率で通信が可能である。

証明 Proof

伝送路では一定の率の誤りが生じるので対策としては符号化回路で時間Ｔにわたる信号系列について符号化を行う。時間Ｔの間に情報源から生じ得る符号列の数は 2^(T.R) 通りとなる。符号化回路の出力に生じうる符号列の数は 2^(T.H(X))通りである。他方受信側において信号の数は 2^(T.H(Y))通りある。各受信信号に対応する送信信号の曖昧度は通りの送信符号列よりなる。

ランダム符号化方法

情報源の出力の 2^(T.R)通りの符号を符号化回路の出力の 2^(T.H(X))通りの符号の中から選んで通信するのが符号化である。具体的には無限に多くの方法があると考えられるがここではランダムに対応付ける方法を取るものとする。

伝送路符号の確率

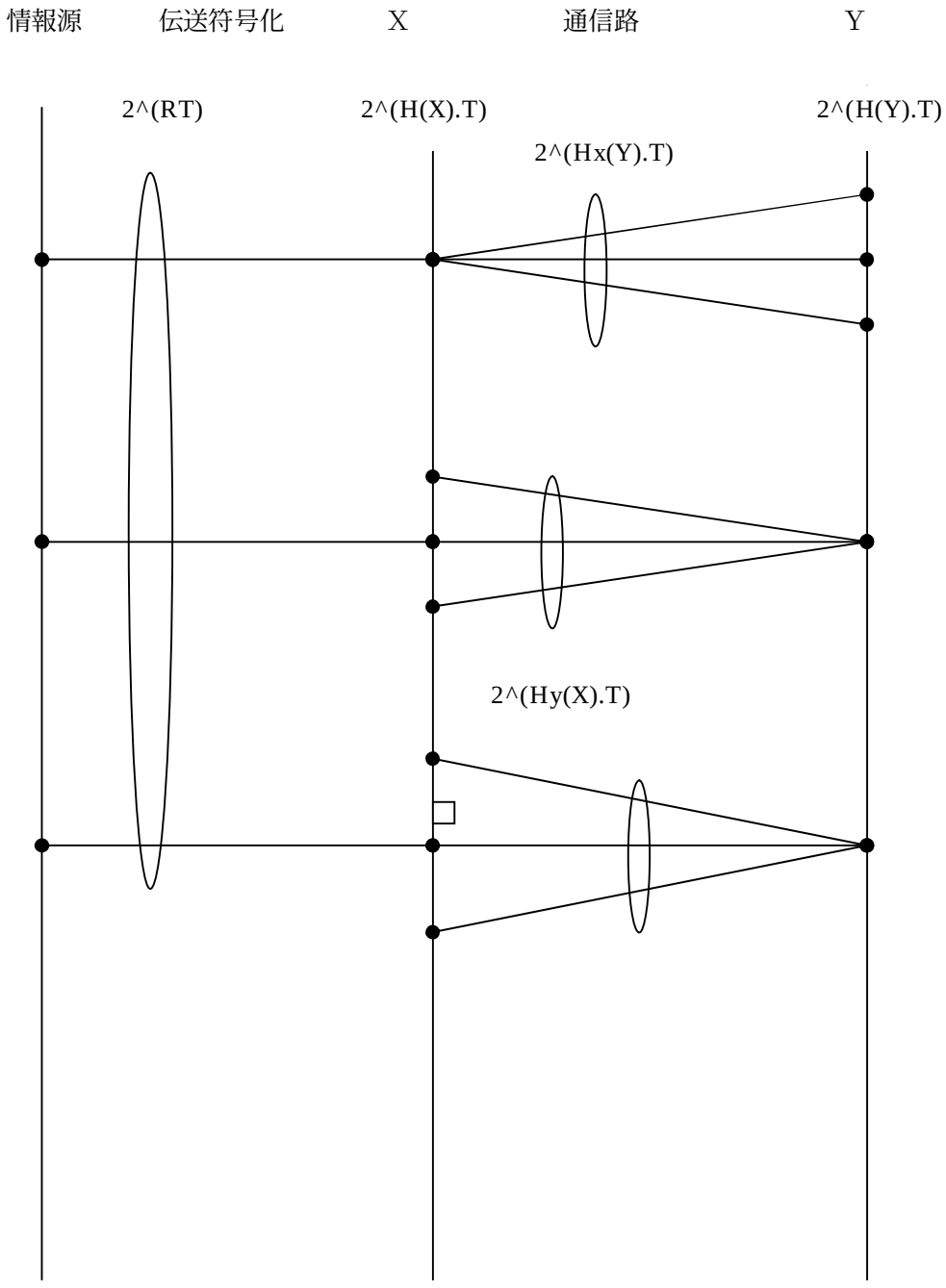
伝送路符号化回路の出力において特定の符号が送信符号に選ばれる確率は 2^(R-H(X).T)となる。これが前述の曖昧度集合 2^(T.Hy(X))の中に入らなければ誤は起こらない。

誤りが起こる確率

(1-2^(R-H(X).T)^(2^(T.Hy(X)) (=) 2^{(R-C).T}　　(C=H(X)-Hy(X))

故にＲ－Ｃ＜０であればＴを大きくすれば誤り率はいくらでも小さくなる。

以上を下図に示す。

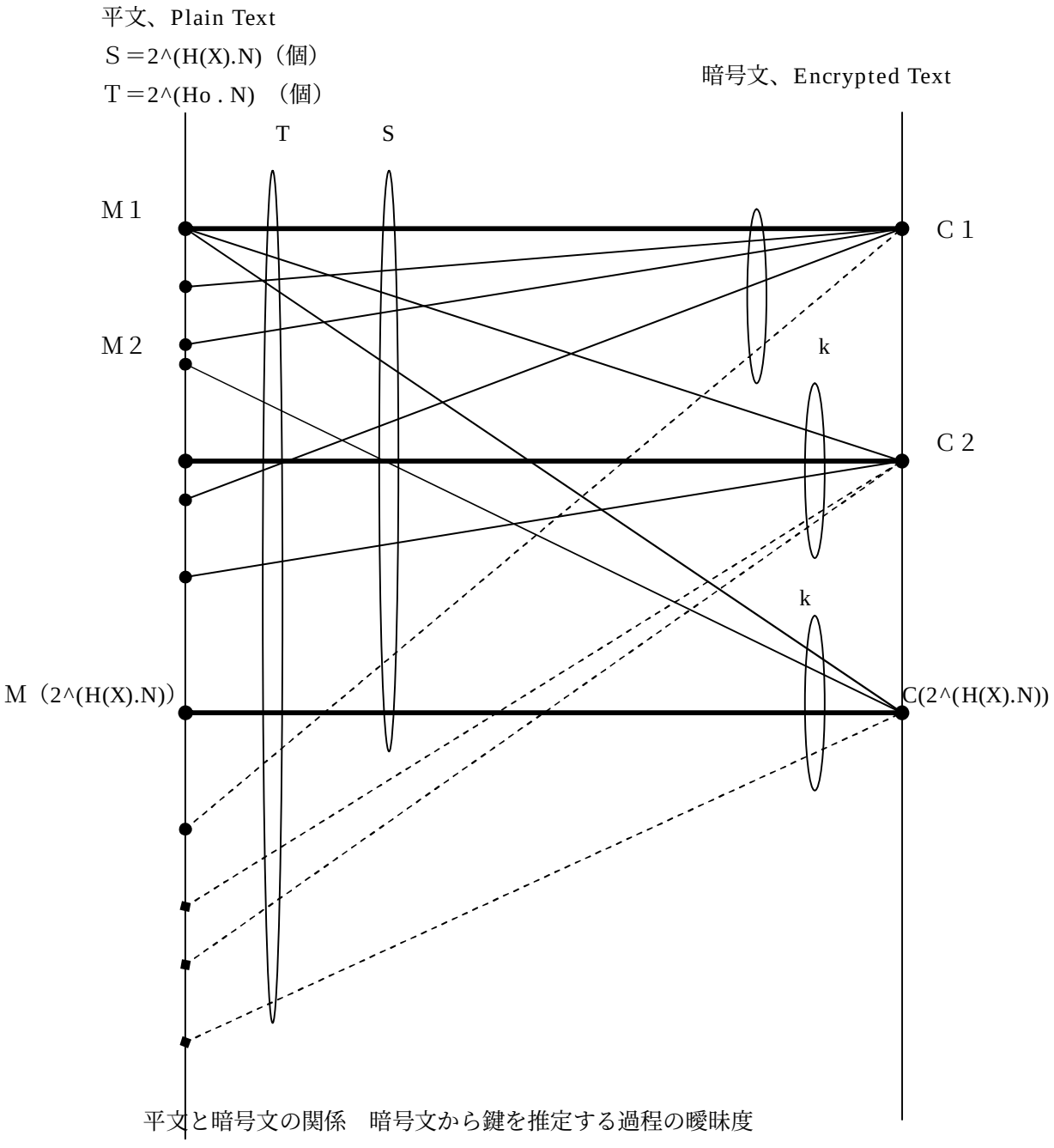


誤りのある通信路の模型、散布度と曖昧度

Model of Communication Path with Errors; Dissemination and Equivocation

６．暗号の基礎理論 Basic Theory of Safe Communication; Encription

シャノンは前述の伝送路符号化と似たモデルで暗号化の基本的な性質を明らかにした。
長さＮ（符号）から成る平文を k 個の内から無作為に選んだ鍵で暗号化して暗号文を作るものとする。



平文と暗号文の関係　暗号文から鍵を推定する過程の曖昧度

Relations between Plain and Enciphered Texts; Equivocation in Crypto-Analysis

情報源の符号の種類をｎ個とする。
仮にｎ個の符号の生起確率がすべて等しい場合の情報エントロピーをＨｏとすると

Ｈｏ＝－Ｌｏｇ（ｎ）

実際はH（X）である。

Ｄ＝Ho－H(X)　　(bits/Symbol)

を冗長量（redundancy）と呼ぶ。

鍵の数をｋとしランダムに使用されるものとする。
鍵の使い方の愛毎度は

Ｈ（Ｋ）＝－［Ｋ＝１，ｋ］∑Ｐ（Ｋ）．Ｌｏｇ（Ｐ（Ｋ））
＝＜　Ｌｏｇ（ｋ）

いま長さＮの平文Mi を鍵Ｋで暗号化して暗号文Ｃi を作成して伝送する。
これを解読するためには鍵Ｋを知らなければならない。
暗号文から鍵を推定する場合の愛毎度は

Hc(K)=-∑P(C).Pc(K).Log(Pc(K))

これが大きいものが強い暗号である。

今暗号文Ｃを受けてこれをｋ個の鍵で復号すれば平文集合Ｔ（要素数２＾（Ｈｏ．Ｎ））にランダムに分布する。実際の平文は集合Ｓ（要素数２＾（Ｈ（Ｘ）．Ｎ））から出ている。
上のランダム復号信号が平文集合に入る数をｍとするとその確率は

Ｐｍ＝ｋＣｍ.（Ｓ／Ｔ）＾ｍ．（１－Ｓ／Ｔ）＾（ｋ－ｍ）

上の図において平文集合Ｓからでる Ｓ．ｋ本の暗号文のうち上のｍ本に一致する確率は
ｍ／Ｓｋである。これより

Hc(K) = -<Log(m)> = -[m]∑m/Sk.Pm.T.Log(m) = -[m] ∑1/k.Pm.Log(m)

＜ｍ＞が非常に大きい時、＜Ｌｏｇ（ｍ）＞　（＝）　Ｌｏｇ＜ｍ＞，故に概算で

Hc(K)（=）Log(<m>) = Log(S) – Log(T) + Log(k)
= H(K) - Ｄ．Ｎ　　(注　<m> = [m]∑m .Pm)

Hc(K)＝０となるＮ＝Ｈ（Ｋ）／Ｄを判別距離（Unicity Distance）とよぶ。

即ち平文の長さが大きいと暗号解読の危険性が高くなる。