⑱ 20の扉について

高野金作

戦後アメリカから流行してきた20の扉（*twenty questions*）は数学的には次の様に考えられる。

有限集合Aとそれに属する未知の確定要素xとがあって，Aの任意の部分集合に対し，それがxを含むか含まないかは確めることができる時に，未知要素xを求めること。そのためにまずAを互に共通要素のない二つの空でない部分集合A_0とA_1とに分けて，xを含むのがどちらであるかを確める（第1問）。

その結果例えばxはA_0に含まれることが分ったとすれば，A_0を再び前の様に二つの部分集合A_{00}とA_{01}とに分けて，xを含むのがどちらであるかを確める（第2問）．その結果例えばxはA_{01}に含まれることが分ったとすれば，A_{01}を再び前の様に二つの部分集合A_{010}とA_{011}とに分けて，xを含むのがどちらであるかを確める（第3問）．この様にしてゆくと，Aは有限集合であるから，いつかは唯一つの要素を含む部分集合に到達してxを求めることができる。

実際の場合にはxを含む方の部分集合を次々に二つに分けてゆけばよいのであるが，理論的には起り得る部分集合の系列，

(1) $A_0, A_1, A_{00}, A_{01}, A_{10}, A_{11}, A_{000}, A_{001}, A_{010}, A_{011}, \cdots\cdots\cdots$

をきめてかゝることもできる。　するとAのどの要素をとっても

それだけからなる集合 $A_{\nu_1\nu_2\cdots\nu_i}$ が存在する。

このときその要素を0と1との有限系列 $\nu_1\nu_2\cdots\nu_i$ 又は二進小数 $0.\nu_1\nu_2\cdots\nu_i$ で表はすことができる。　即ち，集合列(1)をきめておけば，Aの任意の要素を0と1との有限系列又は二進小数で表はすことができる。この二進小数表示は次の性質をもつ。

C_1　一つの要素がある二進小数xで表はされたとすれば，その小数点下の桁数を l とするとき，小数点下 l 桁までxと一致する様な二進小数は，他の異なる要素を表はすのには用いられない。

逆に，有限小数Aの各要素に二進小数を対応させ，條件 C_1 が成立する様にしておくと，(1)の集合の系列を作ることができる。

Aの要素をその小数第1位が0又は1であるに従って A_0 又は A_1 に入れ，A_0 の要素をその小数第2位が0又は1であるに従って A_{00} 又は A_{01} に入れ，以下同様にやればよい。

二進小数の小数点下の桁数（系列で言えば項数）をその長さということにする。　20の扉では，解答に到達するまでの質問数はなるべく少い方がよいから，二進小数の長さはなるべく短い方がよい。　解答者には本当の解は分っていないわけであるが，解に対する確率分布は分っているとしよう。　前に未知の確定要素として考えたxを，Aの要素の値をとる確率変数の実現値として考えるわけである。　さうすると

C_2　xの値を表はす二進小数の長さの平均値を最小にする様な表はし方が望ましい。　そのためにはAの各要素をどんな方法で二進小数で表はしたらよいであろうか。

問題を整理して述べると次の様になる。有限集合 $A=\{a_1,a_2,\cdots,a_n\}$ を値域とする確率変数xがあつて，xは $a_1, a_2, \cdots, a_n$ なる値をそれぞれ確率 $p_1, p_2, \cdots, p_n$ でとるとき（$p_1+p_2+\cdots+p_n=1$），Aの要素 $a_1, a_2, \cdots, a_n$ をそれぞれ二進小数 x_1, x_2

$\cdots, x_n$ で表はし，條件 C_1 及び C_2 を成立させるには，$x_1, x_2, \cdots, x_n$ をどの様にとったらよいか。　換言すれば，(1) の集合列をどの様にとってゆけばよいか。そのためには，A の要素を確率の減少する順に並べ，適当に番号をつけかえて，それを $a_1, a_2, \cdots, a_n$ ($p_1 \geq p_2 \geq \cdots \geq p_n$) とし，確率がほぼ半分になる様に集合を分けてゆけばよいことが知られている。

詳言すれば，$p_1 + \cdots + p_k \fallingdotseq 1/2$，$p_{k+1} + \cdots + p_n \fallingdotseq 1/2$ なる k をとって，$A_0 = \{a_1, \cdots, a_k\}$ と $A_1 = \{a_{k+1}, \cdots, a_n\}$ とに分け，A_0 は $p_1 + \cdots + p_i \fallingdotseq 1/4$，$p_{i+1} + \cdots + p_k \fallingdotseq 1/4$ なる i をとって，$A_{00} = \{a_1, \cdots, a_i\}$ と $A_{01} = \{a_{i+1}, \cdots, a_k\}$ とに分け，以下同様にしてやってゆけばよい。

問題が動物であるとき，それが人間である確率が $1/2$ に近い様であれば，動物を人間とそれ以外の動物とに分ければよいし，求めるべきものが人間の部分であるとき，特別な理由がなければ，上半身と下半身とに分けて聞けばよい。　これらのことは20の扉の解答者達がいつもやっていることである。

ある病気の治療にある薬草が特効があることが分っているが，どの成分が主として効いているのか分らないときにも，同様に考えることができる。　全成分を二つの組に分けてどちらの組が効くかを実験によって確め，有効な方を又二つの組に分けて同様にしてやってゆけばよいが，その分け方が問題である。

薬学や医学等の知識又は経験によって，どの成分が効くか大凡その見当はつけ得るであろう。　確定成分である筈の未知の有効成分を確率変数の実現値とみて，その豫想確率によって計算したときの実験回数の平均値を最小にするには，上記の方法を適用することができる。

これらのことは通信工学に於ける encoding の問題と全く同じであるので，それを紹介して証明もそれについて行うことにす

る。

通信に於ては送ることになっている message を，具体的な通信方法で実際送ることのできる符号（ symbol ）の系列で，一対一の対応がつく様に表はさなければならない。

数学的に言えば，起り得る message の集合Mと符号の系列の集合Sとがあつて，MからSの中えの一対一の対応 f を求めることが問題になる。 これを encoding（符号化）という。

Mに属する各 message に対する f の値を求める装置は encoder 又は送信機といはれる。

受信地に於てこの逆の対応 f^{-1} を求めるものを decoder 又は受信機という。 簡単のため，通信路に起る雑音や歪みは考えないことにする。

こゝではMが有限集合で符号の種類も有限個である場合についてだけ考える。 符号が k 種類あるとすれば，それらを 0, 1, 2, ---, $(k-1)$ で表はすことができる。 これらの符号から成る l 項系列は，0と1との間のある実数の k 進法による小数表示に於ける最初の l 個の数字の系列に対応する。

例えば $k=10$ ならば，14159 なる系列は十進小数 $x=0.14159\cdots$ を小数第5位までとって6位以下を切捨てたものと見ることができる。 従って

$$0.14159 \leqq x < 0.14160.$$

k 種類の符号を用いたときの encoding の問題は，可能な message の各々に0と1との間のある実数を対応させ，それを k 進法によってある桁まで展開したもので表はす問題である。 このとき 1415 及び 14159 の様な二つの系列で二つの異る message を表はすわけにはいかないことに注意しなければならない。 何となれば，f^{-1} を求める受信機は意味のある系列（ある message に対応する系列）を受け取る度に

それに対応する message を再現するからである。

故に若し 1415 なる系列が意味をもっているときに，系列 14159 が送られると，1415 に対応する message が作られ，9は次の message に属することになる。

M に属する messages を実数 $x_1, x_2, \cdots, x_n$ で表はしたとき，それらの小数点下の桁数 $l_1, l_2, \cdots, l_n$ の満たすべき必要充分条件は，任意の x_i, x_j $(i \neq j)$ が，それらを $l = \min(l_i, l_j)$ 桁までとったとき，一致しないことである。

さて M に属する message の中には比較的頻繁に送られるものと稀にしか送られないものとがあって，その確率分布を考えることができる。各符号を送るに要する時間は同一であるものとしておく。さうすると，もっと頻繁に起るものは短い系列で表はすべきであり，稀にしか起らないものは長い系列で表はしてもよいことになる。このためにM内の確率分布が與えられたときに，系列の長さの平均値を最小にする様な encoding が望ましい。それは次の様にして得られる。

M が n 個の message を含むものとし，それを確率の小さい方から並べ，i 番目の message の確率を $p(i)$ $(i = 1, 2, \cdots n)$ とする。$p(1) \leqq p(2) \leqq \cdots \leqq p(n)$, $\Sigma p(i) = 1$,
さて i 番目の message を実数

$$x_i = \sum_{r=1}^{i} p(r) \tag{2}$$

で表はす。

$p(i)$ の有効数字が小数第 l_i 位から始まるとすれば，x_i は小数第 l_i 位まで（l_i 位未満切捨）求めておけばよいことは明らかである。l_i は次の不等式

$$k^{-l_i} \leqq p(i) < k^{-l_i+1}$$

従って

(3) $-\log_k p(i) \leqq l_i < -\log_k p(i) + 1$

を満たす。 message を表はすに用いられる符号の平均数は，$\sum p(i) l_i$ で，これは (3) により次の不等式をみたす。

(4) $-\sum_i p(i) \log_k p(i) \leqq \sum_i p(i) l_i < -\sum_i p(i) \log_k p(i) + 1$.

そこで

(5) $H = -\sum p(i) \log_k p(i)$

とおけば，(4) により

(6) $H \leqq \sum p(i) l_i < H + 1$

が成立つ。 (5)で定義されるHはMのエントロピーと呼ばれ，通信工学では重要な役割を演ずる。 対数の底 k を変えることは，Hに定数を掛けることになるので，エントロピーの単位を変えることに相当する。 エントロピーの定義としては普通 $k = 2$ の場合をとっている。 そのときの単位を binary digit 略して bit という。

この encoding の方法から分る様に，messageの数 n が k に比較して充分大きく，$\max_i \{p(i)\}$ が充分小さいならば，message を表はす符号系列の初の符号（確率変数である）は，$0, 1, \cdots, (k-1)$ なる符号を一様な確率(M内の確率即ち各 message の出現確率で考えて)でとる。 第2の符号についても同様である。 第1の符号と第2の符号とは独立である。

更に $\max\{p(i)\}$ が $k^{-3}, k^{-4}, \cdots$ に比較して小さければ，第3第4 $\cdots$ の符号についても同様である。

勿論，符号の長さが 3, 4, $\cdots$ 以上であるという条件の下での

語である。　従って，messageを表はす系列はk進法における random sequenceを構成する。　但し messageの出現確率で考えたときのことで，x_1, x_2, ----, x_nに一様な確率を與えたときの議論ではない。

さてわれわれのencodingが実用上符号の平均数を最小ならしめることを証明しよう。　先づ $n=k^S$（Sは正整数）ですべての $p(i)$ が k^{-S} に等しい場合を考える。　$-\log_k p(i)=S$ となるから，どのmessageも丁度S個の符号で表はされる。

從って各messageを表はす符号系列の平均の長さはSである。

符号系列の平均の長さがSより小さい様なencodingは存在しない。　これを証明しよう。　任意のencodingを考え，符号系列の最小の長さを l とする。　$l<S$ ならば，長さの平均を減少させて最小の長さを $l+1$ にすることができる。

何となれば，長さ l の系列が m 個あったとすれば，これらの系列に更にもう一つの符号をつけ加えて出来る長さ $l+1$ の mk 個の系列で，長さ l の m 個の系列及び長い方の $m(k-1)$ 個をおきかえればよいからである。　これを逐次行ってゆけば，長さの平均を減少させて最小の長さをSにすることができる。　最小の長さがSの場合にはわれわれのencodingが平均の長さを最小にすることは明らかである。

次に一般の場合には長さの平均がH−1より小さいencodingは存在しないことを帰謬法によって証明する。（Mのエントロピー HはMの確率分布だけできまる量で，encodingとは無関係である。）存在したとし，そのときの i 番目のmessageに対応する系列の長さを l_i（前のものとは異る）とすれば，仮定により

$$(7)\qquad \sum p(i)l_i < -\sum p(i)\log_k p(i) - 1$$

こゝで l_1, l_2, ----, l_n を固定しておけば，両辺とも $p(1)$, $p(2)$,

---, $p(n)$の連続函数であるから，$p(i)$を適当な

(8) $$p_1(i)=\frac{n_i}{k^S} \quad (S, n_i \text{は整数}, \Sigma n_i = k^S)$$

でおきかえても，(7)は成立つ．　即ち

(9) $$\Sigma p_1(i) l_i < -\Sigma p_1(i) \log_k p_1(i) - 1 = -\Sigma p_1(i) \log_k n_i + S - 1.$$

故に

(10) $$\Sigma p_1(i) \left(l_i + \log_k n_i + 1 \right) < S$$

$[a]$でa以上の最小の整数を表はすことにすれば$(a \leqq [a] < a+1)$

(11) $$\Sigma p_1(i) \left(l_i + [\log_k n_i] \right) < S.$$

第i番目の系列に更に$[\log_k n_i]$個の符号をつけ加えて長さを$l_i + [\log_k n_i]$にすることにより，少くともn_i個の系列を得るから，それらの中から適当にn_i個とり，これをすべてのiについて考え，この様にしてできる$\Sigma n_i = k^S$個の各系列に改めて確率k^{-S}を與えることにすれば，前に証明した結果と矛盾する。

以上により，われわれの encoding では符号系列の長さの平均値$\Sigma p(i) l_i$はHと$H+1$との間（Hを含む）にあり，任意の encoding におけるそれは$H-1$以上であることが分った，こゝにHはMのエントロピーである。

さてある$p\ (0<p<1)$と整数kとが存在し

$$\sum_1^{k-1} p(i) < p \leqq \sum_1^{k} p(i) \quad \text{で且つ } p(k) \text{は充分小さい}$$

ならば，Hは充分大きくなり得るから，この場合には近似的に$\Sigma p(i) l_i \fallingdotseq H \fallingdotseq H-1$としてよい。

従って，われわれの encoding は実際的見地からは最適のものである。

これで証明は終る。

前に述べた 20 の扉の場合の符号は yes と no との二つで，$k=2$ の場合である。

あとがき．

通信工学における encoding の語は G.A. Barnard, The theory of information, J.R.S.S. Vol. 13, No. 1, 1951 による。

但し，証明に不充分な点がある様に思はれるので，部分的な改更及び補足をつけ加えた。

こゝで述べた encoding の方法は $k=2$ の場合は，C.E. Shannon 及び R.M. Fano によって独立に発見されたものである。

（C. E. Shannon and W. Weaver, The mathematical theory of communication, 1949, pp. 29-30 参照）

なお，国沢清典，通信工学に於ける coding の統計的改良，確率過程第1回シンポジューム講演集，1952 年，を参照せよ。

（1953 年 2 月 21 日）