講義「情報理論」

第8回　情報源符号化法（2）

情報理工学部門 情報知識ネットワーク研究室
喜田拓也

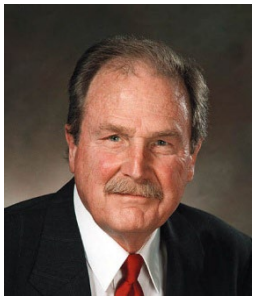
講義資料

2019/7/24

ハフマン符号はなぜ大事か？（おさらい）

ハフマン符号は コンパクト符号 である！

コンパクト符号とは，１記号ずつ符号化する際，その平均符号長を最小とする効率のよい符号のこと



David Albert Huffman
(1925 –1999)

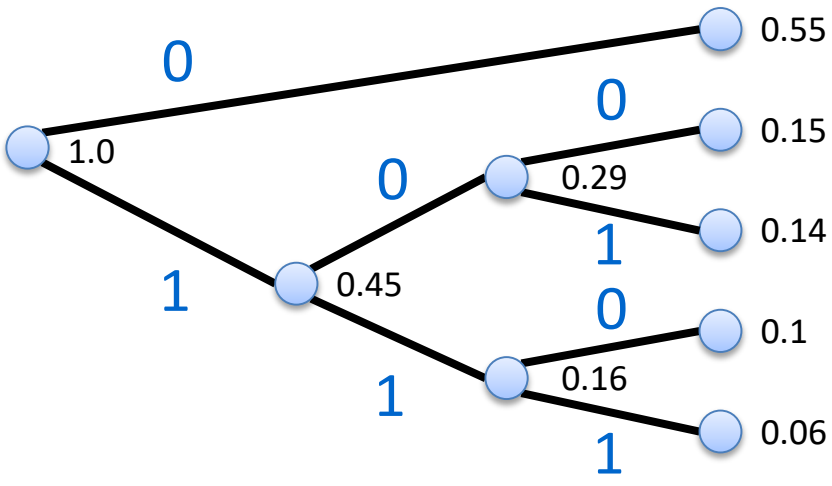
http://www.adeptis.ru/vinci/m_part5_2.html より

ハフマン符号(おさらい)

ハフマン木を符号木として各記号に符号語を割り当てる符号化

また，元の情報源Sに対し，n次拡大情報源S^nを考え，その上でハフマン符号化することで，平均符号長をエントロピー $H(S)$により近づけることができる．これをブロックハフマン符号化と呼ぶ

ハフマン木



情報源記号 x	確率 $P(x)$	符号語
A	0.55	0
D	0.15	100
B	0.14	101
E	0.1	110
C	0.06	111

今日の内容

5.3 非等長情報源系列の符号化

5.4 ひずみが許される場合の情報源符号化

ブロックハフマン符号化の問題点

ブロックハフマン符号化におけるブロック長nを十分大きくすれば，
1情報源記号あたりの平均符号長をいくらでも下限に近づけられる

だがしかし

nを大きくすると，記号の数が急増する！

M元情報源の場合，符号化すべき長さnの情報源系列の数が，
M^n個に増大する！（ハフマン符号化が困難になる！）

$1, 0$の発生確率が$0.01, 0.99$の無記憶定常情報源Sを考える
このSのエントロピーは$H(S)=0.081$

平均符号長Lを，この1割り増しの0.089までに抑えたい！
nを$1/0.008=125$以上にすれば確実である
しかし$n=125$の系列は$2^{125} \fallingdotseq 4\times10^{37}$個もある！

$L < H_n(S) + 1/n$

ここが0.008

40澗（かん）

非等長情報源系列に対する符号化

符号化すべき情報源系列を非等長にしてはどうだろう？
すなわち，長い情報源系列と短い情報源系列を組み合わせ，
長いがよく発生する系列に，より短い符号語を割り当てる

利点：

符号化する情報源系列の数を減らして，符号化のために
記憶すべき表を削減できる

0 0 0 1 0 1 0 0 1 1 0 1 0 1 0 0 0 ・・・

どう区切ればいいの？

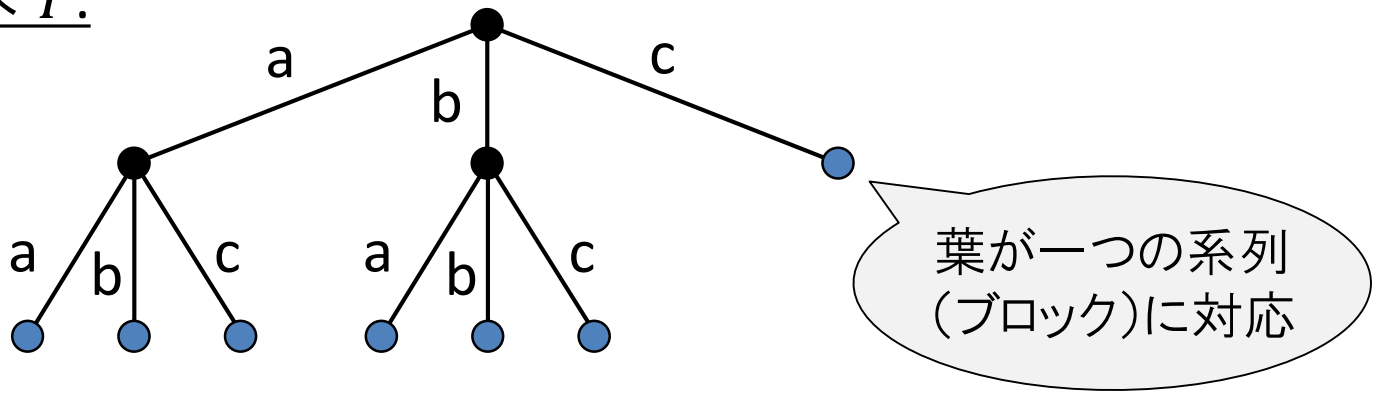
情報源から出力される任意の系列が，
一意に分解できなければならない！！

情報源系列	確率
0 0 0	0.512
0 0 1	0.128
0 1	0.16
1	0.2

分節木を用いた情報源系列の分割

分節木と呼ばれる木構造を用いて，情報源系列を長さの異なる系列（ブロック）に分割し，各ブロックに対して符号化を行う
分節木の各葉ノードはある一つの系列に対応している

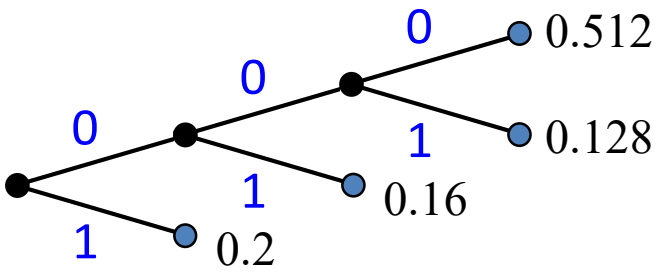
分節木 T:



情報源系列：a b | b a | a b | b a | a a | c | a c | c

非等長情報源系列の符号化の例

1, 0を確率0.2, 0.8で発生する無記憶定常情報源 S を考える.
Sから発生する系列を4つ選び, ハフマン符号化を行う.



情報源系列を分割する分節木

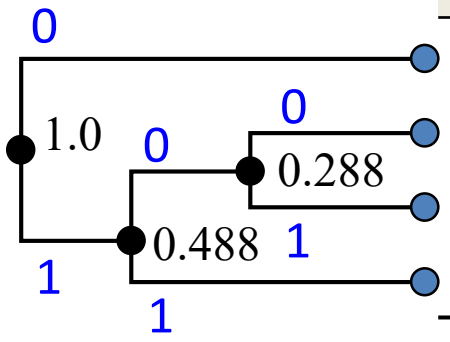
各ブロックの平均長$\overline{n}$は

$$\overline{n} = 1 \times 0.2 + 2 \times 0.16 + 3 \times 0.128 + 3 \times 0.512 = 2.44$$

右の符号の平均符号長$L' = 1.776$

よって1記号あたりの
平均符号長Lは

$$L = \frac{1.776}{2.44} = 0.728$$



情報源系列	確率	ハフマン符号
0 0 0	0.512	0
0 0 1	0.128	100
0 1	0.16	101
1	0.2	11

ランレングス･ハフマン符号化

系列中に同じ記号が連続するとき，その連続する長さを符号化して送る方法を一般に，ランレングス符号化と呼ぶ

ああたたたたたたたた＝ あ2た8

先の例は，系列を長さ3までの0の連続（0のラン）でブロック化している．このように，ランでブロック化してからハフマン符号化する方法を，ランレングス･ハフマン符号化と呼ぶ

ランレングス･ハフマン符号化の平均符号長を考察してみよう

1, 0の出現確率が，それぞれ $p, 1-p$ $(p<1-p)$ の無記憶定常情報源Sについて，$N-1$個までの0のランを符号化する

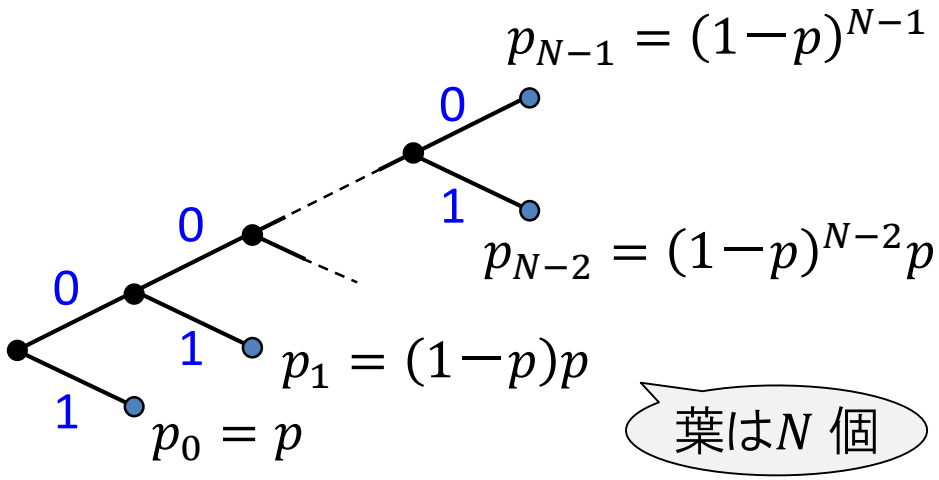


図5.10 ランランレングス符号化のための分節木

ランレングス・ハフマン符号化(続き)

これらN個の系列の平均長$\overline{n}$は

$$\overline{n} = \sum_{i=0}^{N-2}(i+1)(1-p)^i p + (N-1)(1-p)^{N-1}$$

0^{N-1}の系列

$$= \frac{1-(1-p)^{N-1}}{p} .$$

$0^i 1$の系列

これらの系列をハフマン符号化したときの平均符号長L_Nは次を満たす.

$$L_N < -\sum_{i=0}^{N-1} p_i \log_2 p_i + 1$$
$$= H(S)\overline{n} + 1 .$$

よって1記号あたりの平均符号長L_rは

$$L_r = \frac{L_N}{\overline{n}} < H(S) + \frac{1}{\overline{n}} .$$

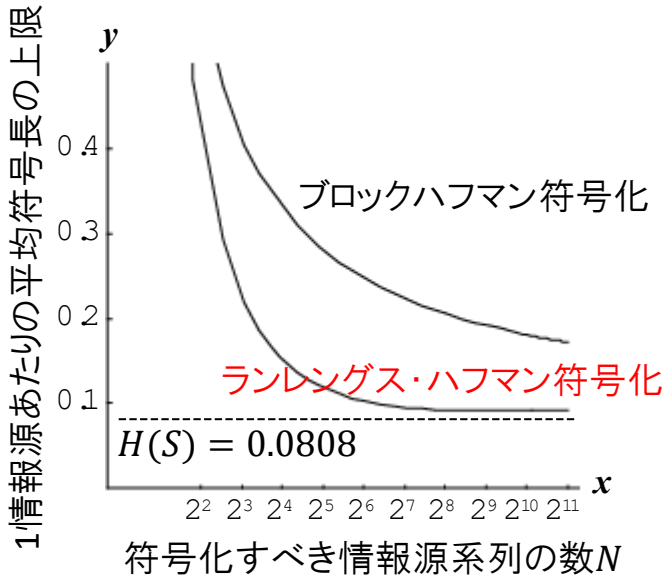


図5.11 ブロックハフマン符号化とランレングス・ハフマン符号化の上限

ブロックハフマン符号化の場合: $L_h < H(S) + \frac{1}{n} = H(S) + \frac{1}{\log_2 N}$

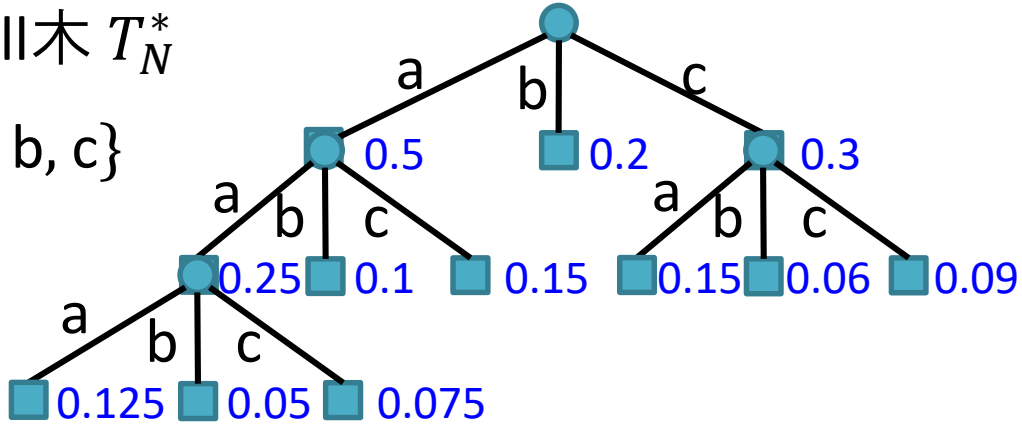
最適な分節木の構築

与えられた$N \geq 1$と記号の生起確率$p_i = P(a_i)$ $(a_i \in \{a_0, a_1, \ldots, a_M\})$に対し，次のステップにより最適な分節木（Tunstall木）T_N^*を構築する

1. 根から各情報源記号に対応するM本の枝を伸ばし，葉に確率p_iをつける．これを初期木T_1^*とする．
2. 現在の分節木の葉の数に$M-1$を足したものがNを超えないのであれば，次のステップを実行する．そうでなければ終了．
3. 葉の中から，最大の確率$\hat{p}$を持つ葉v_i^*を選ぶ．葉v_i^*の下に各情報源記号に対応するk本の枝を伸ばしてT_i^*を作る．新しくできた葉には確率$\hat{p}p_i$をつける．そしてステップ2に戻る．

例）Tunstall木 T_N^*

$\sum = \{a, b, c\}$
$M = 3,$
$N = 9$



$P(a) = 0.5$
$P(b) = 0.2$
$P(c) = 0.3$

Try 練習問題5.3

ちょっと休憩

ひずみが許される場合の情報源符号化

ひずみが許される場合とは？

例えば画像データの通信とか

大丈夫だ．問題ない

Lena Söderberg
（レナ･ソーダバーグ）

BMPファイル
（775 KB）

ひずみを許す
符号化

0011010
1110111
0001･･･

復号

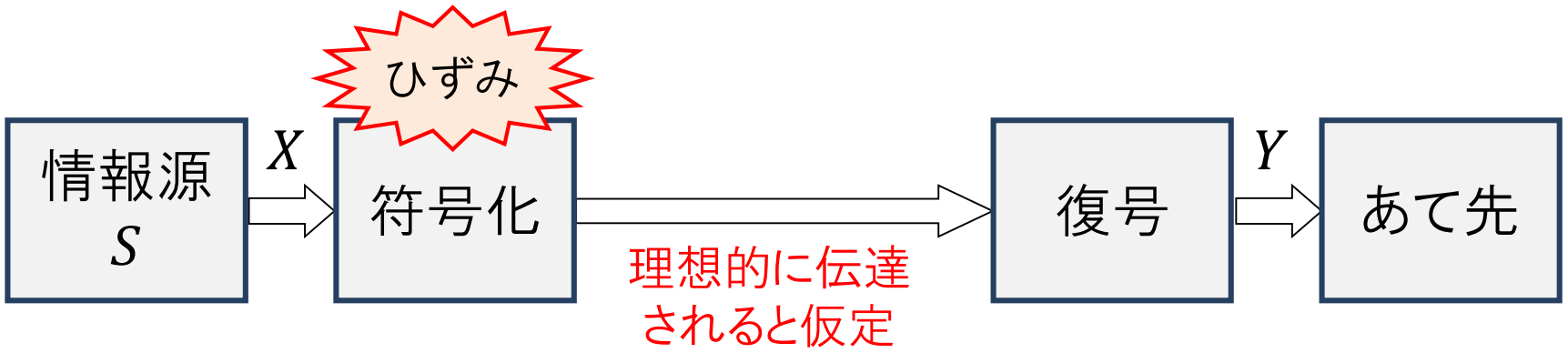
JPEGファイル
（中画質 59.5 KB）

ひずみを入れた情報源符号化

通信路でひずみが入るのではなく，符号化時に（わざと）ひずみを入れる

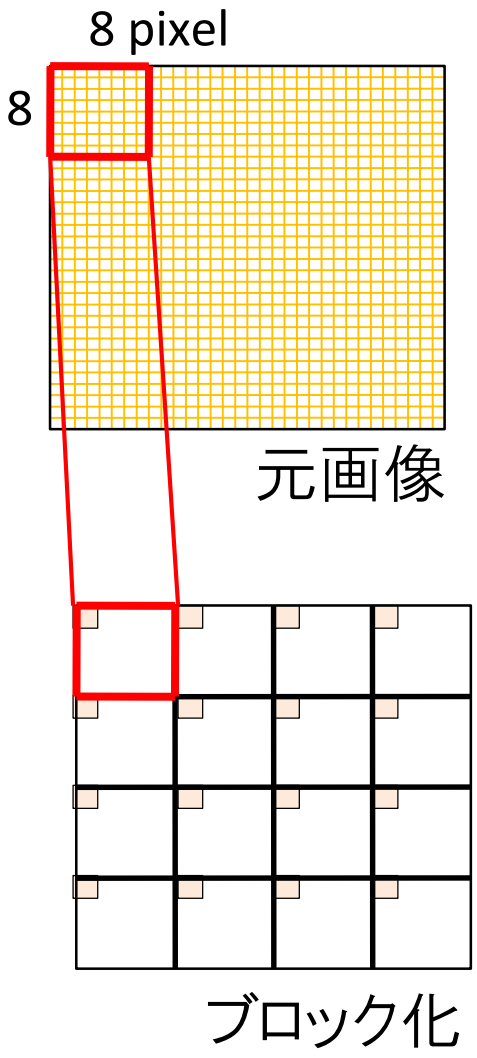
元の情報（量）を削って通信することに相当する

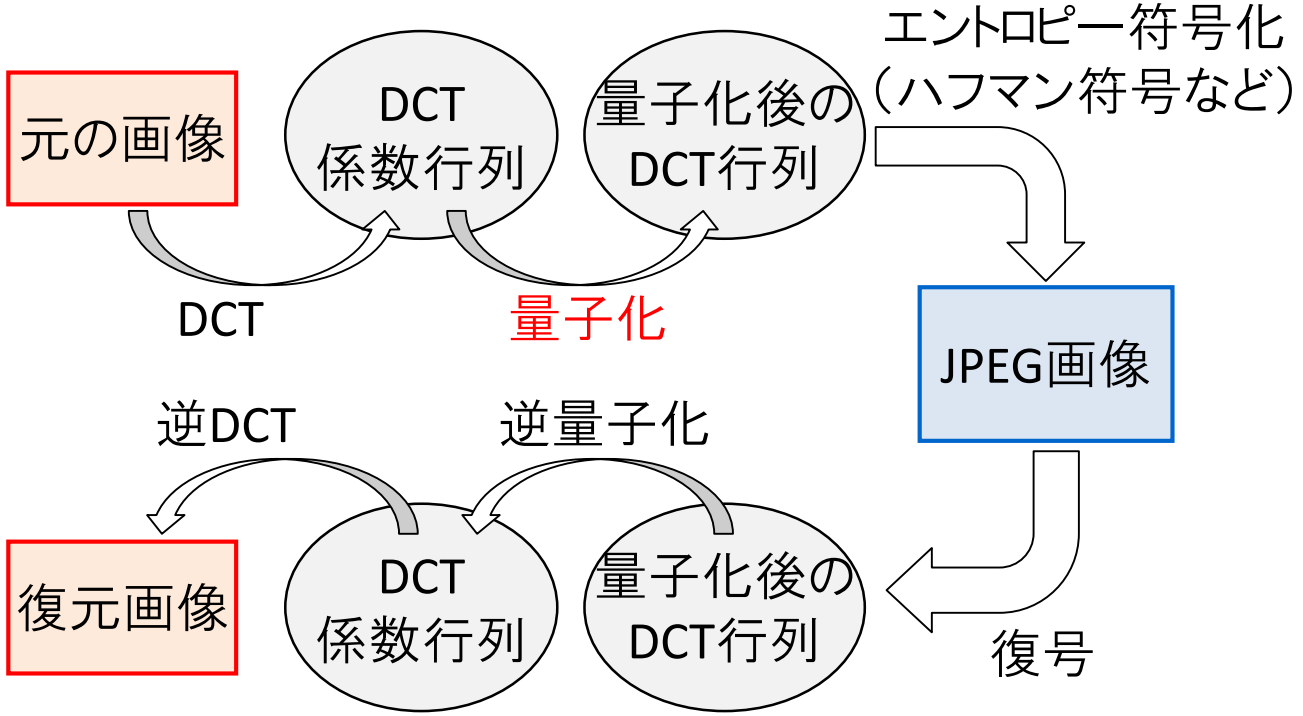
そうすることで圧縮率を向上させられる

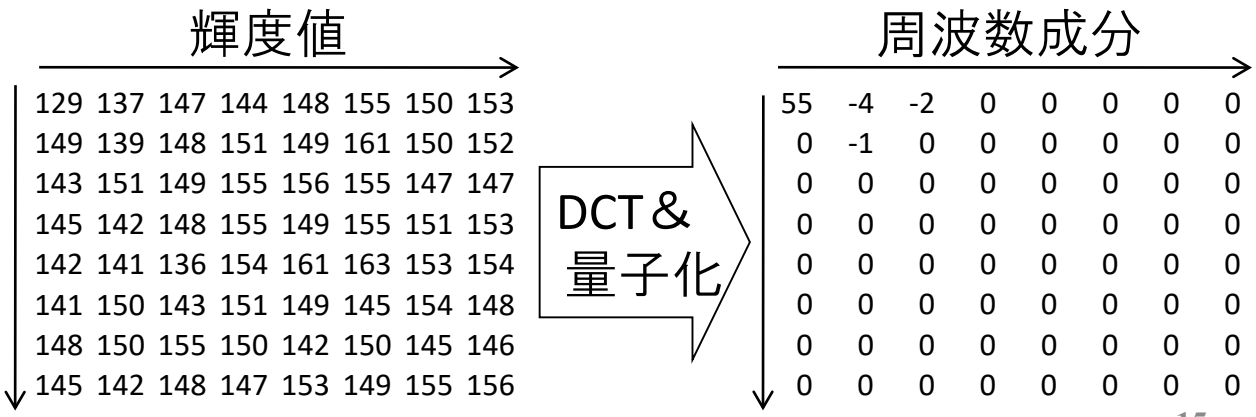


情報源出力 X と復号結果 Y

例）JPEG画像の仕組みの概要







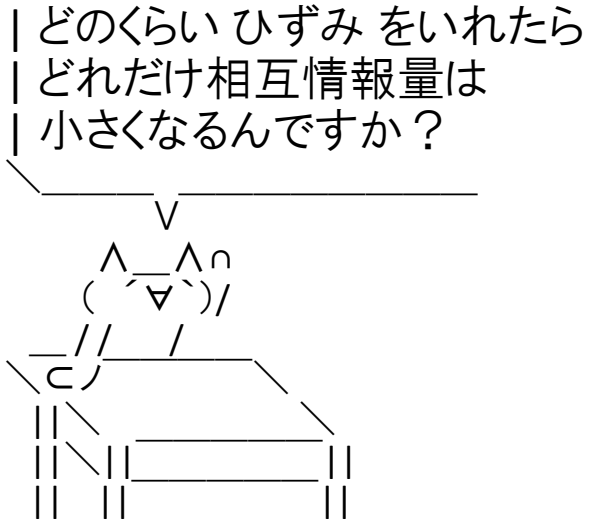
どのくらい平均符号長の限界を下げられるか？

1情報源記号あたりの平均符号長の下限 ＝ エントロピー$H(S)$

だがしかし！

ひずみを許した場合，出力Yの値を知っても，
元の入力Xに関してなお平均して $H(X|Y)$ のあいまいさが残る
伝えられる情報の量は $H(X)-H(X|Y) = I(X;Y)$

ひずみを許した場合の限界は
相互情報量で表される！



ひずみ測度

ひずみ測度：x と y の相違を評価する関数 $d(x,y)$

関数 $d(x,y)$ が大きいほど，ひずみが大きい．
また次の性質を持つ．

$$d(x,y) \geq 0$$

$x=y$ のとき $d(x,y)=0$

ひずみ測度の平均値を平均ひずみと呼び，$\overline{d}$ で表す

$$\overline{d} = \sum_x \sum_y d(x,y) P_{XY}(x,y)$$

ひずみ測度の例

例1） 情報源アルファベットを$\Sigma = \{0, 1\}$とし，ひずみ測度を

$$d(x, y) = \begin{cases} 0; & x = y \\ 1; & x \neq y \end{cases}$$

とする．このとき，平均ひずみは

$$\overline{d} = \sum_x \sum_y d(x, y) P_{XY}(x, y)$$
$$= P(1, 0) + P(0, 1) .$$

$P(1,0)$：入力 1 → 出力 0
$P(0,1)$：入力 0 → 出力 1

これは要するに，符号器の出力が元の情報源の出力と異なる確率であり，通常ビット誤り率と呼ばれる．

例2） 情報源アルファベットを有限個の整数または実数の集合としよう．このとき，ひずみ測度を$d(x, y) = |x - y|^2$とすれば，平均ひずみは2乗平均誤差（mean square error）と呼ばれる量となる．ひずみの評価量として非常によく用いられる．

平均ひずみと相互情報量の関係

相互情報量$I(X;Y)$ が同じでも，平均ひずみ$\overline{d}$ は同じとは限らない

⇔ 平均ひずみ$\overline{d}$ が同じでも，$I(X;Y)$ は符号化の仕方で異なる

ある与えられた値D に対し，平均ひずみ$\overline{d}$ が

$$\overline{d} \leq D$$

D は平均ひずみのしきい値

を満たす条件の下で，あらゆる情報源符号化法を考えたときの相互情報量$I(X;Y)$ の最小値を考え，これを$R(D)$ と表す．

すなわち，

$$R(D) = \min_{\overline{d} \leq D} \{I(X;Y)\} .$$

これを情報源 S の速度･ひずみ関数（rate-distortion function）と呼ぶ

つまり，これが平均符号長の下限！

ひずみが許される場合の情報源符号化定理

定理［ひずみが許される場合の情報源符号化定理］

平均ひずみ $\overline{d}$ を D 以下に抑えるという条件の下で，任意の正数εに対して，情報源S を1情報源記号あたりの平均符号長L が

$$R(D) \leq L < R(D) + \varepsilon$$

となるような2元符号へ符号化できる．しかし，どのような符号化を行っても，$\overline{d} \leq D$ である限り，L をこの式の左辺より小さくすることはできない．

この定理は，1情報源記号あたりの平均符号長を，速度･ひずみ関数$R(D)$ にいくらでも近づく符号化法の存在を示している

具体的な符号化方法はあるのか？

ひずみのない場合に比べてはるかに難しい！

教科書【例5.8】参照

今日のまとめ

基本的な情報源符号化法

- ハフマンブロック符号化法の問題点
- 非等長情報源系列の符号化
- ランレングスハフマン符号化

ひずみが許される場合の情報源符号化

- 情報源符号化におけるひずみ
- ひずみが許される場合の情報源符号化定理
- 速度･ひずみ関数

次回

- 通信路符号化の基礎概念

【例5.8】ひずみ速度関数の例

$1, 0$ を確率 $p, 1-p$ で発生する記憶のない2元情報源を考える.
また, ひずみ測度としては先の例1と同じ

$$d(x,y) = \begin{cases} 0; & x = y \\ 1; & x \neq y \end{cases}$$

を用いるものとする. このとき, 平均ひずみ $\overline{d}$ はビット誤り率となる.
この情報源に対して, $0 \leq D \leq 0.5$ が与えられたとき,
$\overline{d} \leq D$ の元での速度･ひずみ関数 $R(D)$ を求めよう.
相互情報量は, $I(X;Y) = H(X) - H(X|Y)$.
$H(X) = \mathcal{H}(p)$ なので,
$H(X|Y)$ を最大化すればよい.
ここで, Y は右図のように,
1の発生確率が $\overline{d}$ である
ような誤り源の出力 E と X
の排他的論理和で表せる.

$D > 0.5$ の場合は
信号よりもひずみが
大きいことを意味する

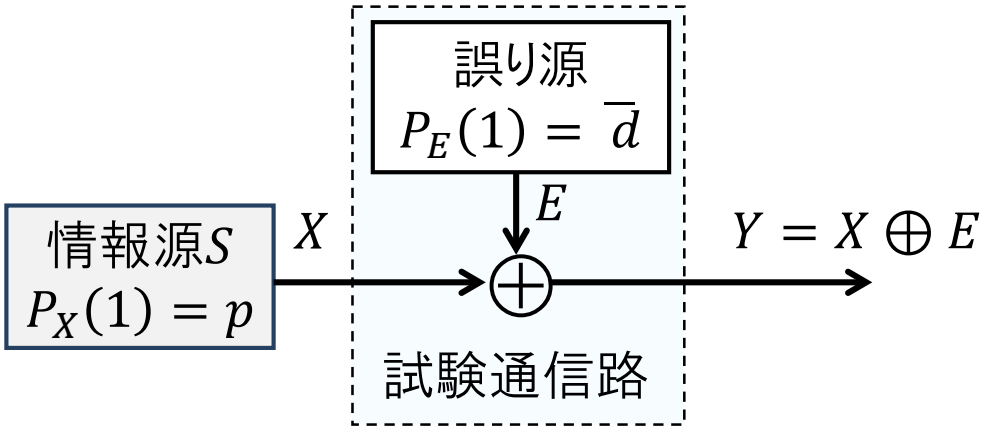


図. 2元情報源に対する試験通信路

【例5.8】ひずみ速度関数の例（続き）

$Y = X \oplus E$ であるから，$X = Y \oplus E$ となる．したがって，

$$H(X \mid Y) = H(Y \oplus E \mid Y) = H(E \mid Y) \ .$$

$H(E|Y)$ は Y の値を知ったときの E のあいまいさであるから，何も知らないときの E のあいまいさ $H(E)$ より大きくなることはない．さらに，誤り源に記憶がなく定常であれば，$H(E) = \mathcal{H}(\overline{d})$ であるが，そうでなければ，$H(E) < \mathcal{H}(\overline{d})$ であるから，

$$H(E \mid Y) \leq H(E) \leq \mathcal{H}(\overline{d})$$

となる．それゆえ

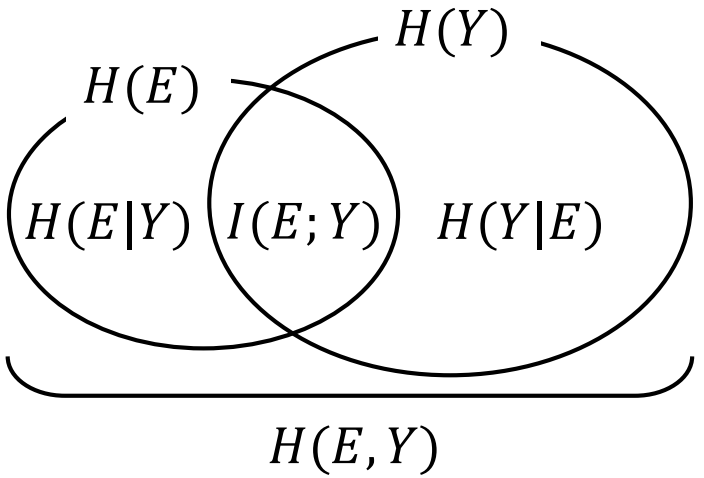
$$H(E \mid Y) \leq \mathcal{H}(\overline{d})$$

を得る．$\overline{d} \leq D$ なので，さらに

$$\mathcal{H}(\overline{d}) \leq \mathcal{H}(D)$$

となる．したがって，相互情報量 $I(X;Y)$ は，

$$I(X;Y) = H(X) - H(X \mid Y) \geq \mathcal{H}(p) - \mathcal{H}(D)$$



【例5.8】ひずみ速度関数の例(続き)

このように $I(X;Y) \geq \mathcal{H}(p) - \mathcal{H}(D)$ となるので,
記憶のない定常2元情報源S の速度･ひずみ関数は

$$R(D) = \mathcal{H}(p) - \mathcal{H}(D)$$

で与えられることが導けた.

右図で分かるように, 速度･ひずみ関数は, Dに関して単調減少であり, 下に凸な関数である. 一般の速度･ひずみ関数も同様な性質を持つことが証明されている.

記憶のある情報源の場合にも,

$$I(X;Y) = \lim_{n\to\infty} I(X_n;Yn)/n$$

の最小値として, 速度･ひずみ関数を定義することができる.

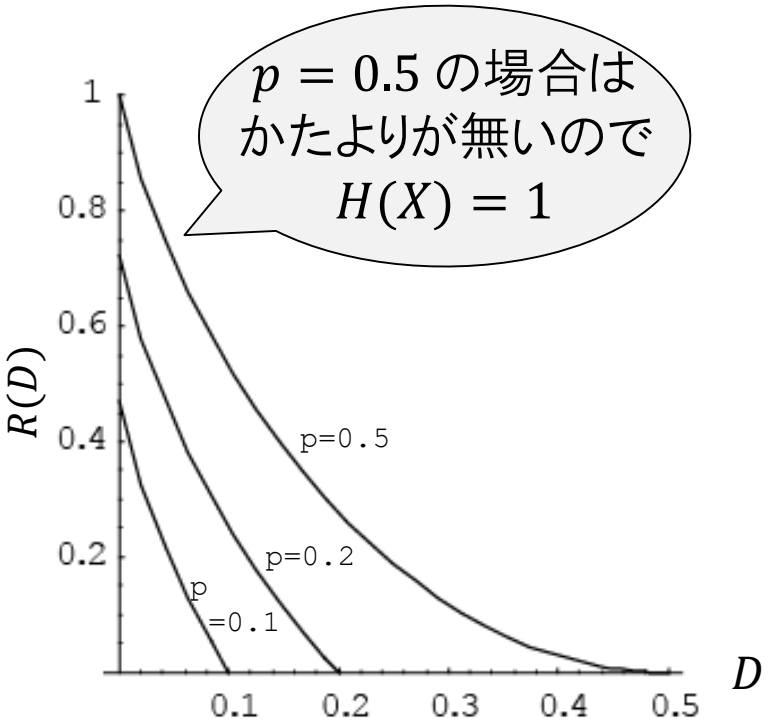


図5.17 記憶のない2元情報源の速度･ひずみ関数