北海道大学
Hokkaido University

情報エレクトロニクス学科共通科目・2年次・夏ターム〔必修科目〕

講義「情報理論」

第8回

第5章 情報源符号化法

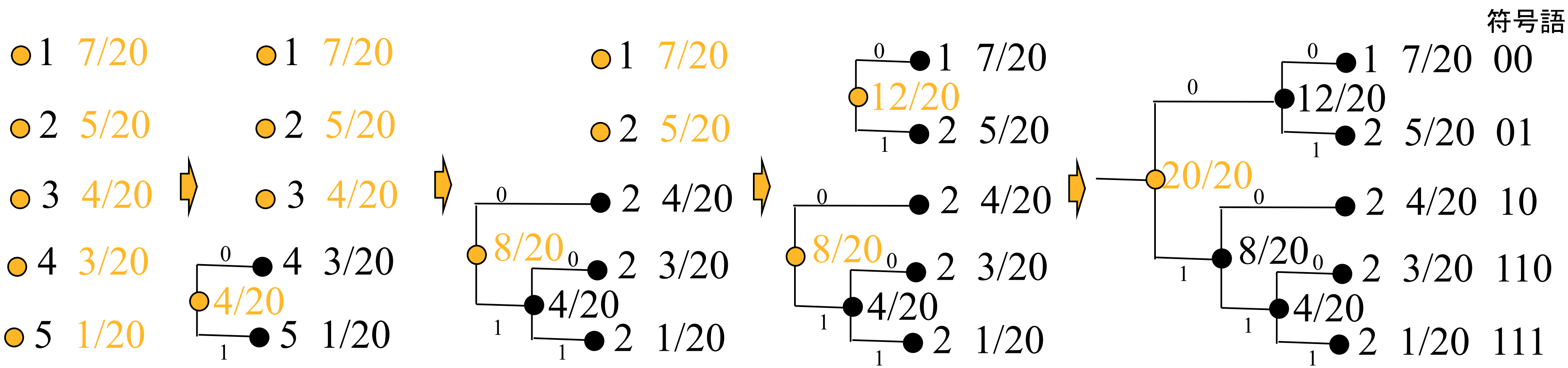
5.3 非等長情報系列の符号化
5.4 ひずみが許される場合の情報源符号化

前回の演習問題

[問] 袋の中に空くじなしの１等から5等までのくじが20本入っている。各等の本数は、下表の通りとする。

１等	２等	３等	４等	５等
1	3	4	5	7

このくじを１本ずつ繰り返し引き、何等が当たったかを情報源と考え、i等が当たったとき情報源記号iを発生させることにより情報源系列をつくる。ただし、引いたくじは次のくじを引く前に袋に戻すものとする。この情報源を、符号アルファベット{0,1}を使って、情報源記号１つ１つを一意に復号可能な2元符号に符号化する場合において、コンパクト符号を1つ求めよ。また、その符号の平均符号長を求めよ。



[復習]コンパクト符号

コンパクト符号

各情報源記号を符号化した一意復号可能な符号で

平均符号長が最小の符号

<u>情報源Sの2元コンパクト符号の平均符号長</u>$\geq H_1(S)$

↑

各情報源記号を2元符号化した場合の本当の限界

代表的なコンパクト符号

ハフマン符号

[復習]ブロックハフマン符号化

情報源Sから発生するn個の情報源記号ごとにブロック2元符号化

n記号ごとに符号化した符号の一情報源記号当たりの平均符号長L_nが以下の式を満たすものが存在

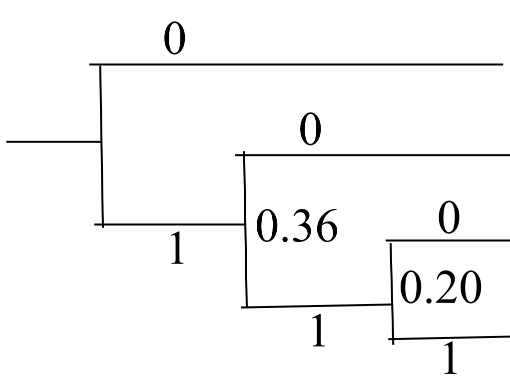
$$H_1(S^n)/n \leqq L_n < H_1(S^n)/n + 1/n \quad \cdots\cdots\cdots ①$$

（$H_1(S^n)/n$はSの n次エントロピーと呼ばれる量）

ブロックハフマン符号化

情報源Sから発生するn個の情報源記号ごとにハフマン符号化
平均符号長L_nは①式を満たす → nを無限大に近づけると$H(S)$に近づく。

[例：1の出る確率が0.2の2元無記憶定常情報源を2情報源記号ごとにハフマン符号化]



確率	情報源系列	ハフマン符号
0.64	00	0
0.16	01	10
0.16	10	110
0.04	11	111

1情報源記号あたりの平均符号長は0.78

ブロックハフマン符号化法の問題点

- ブロックハフマン符号化法
  - n次の拡大情報源に対してハフマン符号化を行うやり方
  - n を大きくすることにより、1情報源記号あたりの平均符号長をいくらでも下限$H(S)$に近づけられる。

しかし

長さnのすべての情報源系列と符号語の巨大な対応表が必要

無記憶ならば
$H_n(S)=H_1(S)=H(S)$

[例] 1, 0の発生確率が0.01, 0.99の無記憶情報源S

$H(S)=0.081$

目標：平均符号長L を0.089 以下

無記憶だから$L<H(S)+1/n$となり、n を 1/0.008＝125 以上にすれば確実。

1/125=0.008

$n=125$ の系列は $2^{125}\fallingdotseq 4\times 10^{37}$ 個もある！

事実上、そのようなハフマン符号を構成することはできない・・・

非等長情報源系列の符号化

- n記号毎符号化するハフマンブロック符号化法では、符号化すべき情報源系列の数は、M元情報源の場合、M^n 個！
- 符号化すべき情報源系列を非等長にしてはどうだろう？
  - 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、短い符号語を割り当てる
  - 符号化する情報源系列の数を減らし、符号化のために記憶すべき表を削減する

↓

  - 比較的に長いブロックで符号化したときと同じような効果を持たせられないだろうか？

[例]2元無記憶定常情報源の符号化

1の出る確率が0.2の2元無記憶定常情報源を考える

ブロック符号化

情報源系列 0 0 1 1 0 0 1 0 0 0 0 0 0 1

符号語列 0 1 1 1 0 1 1 0 0 0 1 0

確率	情報源系列	符号語
0.64	00	0
0.16	01	10
0.16	10	110
0.04	11	111

1情報源記号あたりの平均符号長は0.78

非等長情報源系列の符号化

情報源系列 0 0 1 1 0 0 1 0 0 0 0 0 0 1

符号語列 1 0 0 1 1 1 0 0 0 0 0 1 1

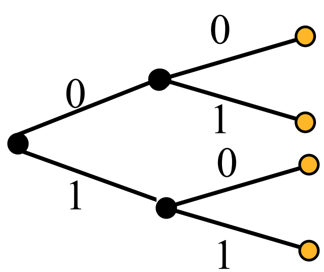
確率	情報源系列	符号語
0.512	000	0
0.128	001	100
0.16	01	101
0.2	1	11

1情報源記号あたりの平均符号長は0.728

同じ数の符号語でより効率的な符号を実現

分節木

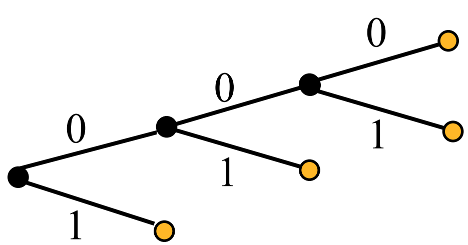
分節木



符号化する情報源系列
（橙の節点）を表す木

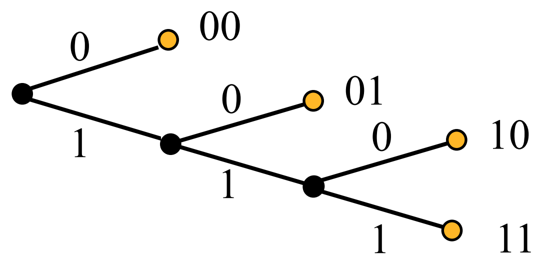
情報源系列	符号語
00	0
01	10
10	110
11	111

情報源系列	符号語
000	0
001	100
01	101
1	11

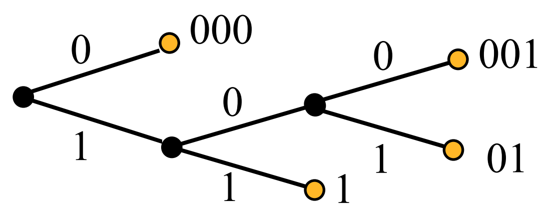


平均系列長が長いほど効率的

符号の木



符号語（橙の節点）を表す木



平均符号長が短いほど効率的

ランレングス符号化法(1)

情報源系列	符号語
000	0
001	100
01	101
1	11

- ランレングス符号化法

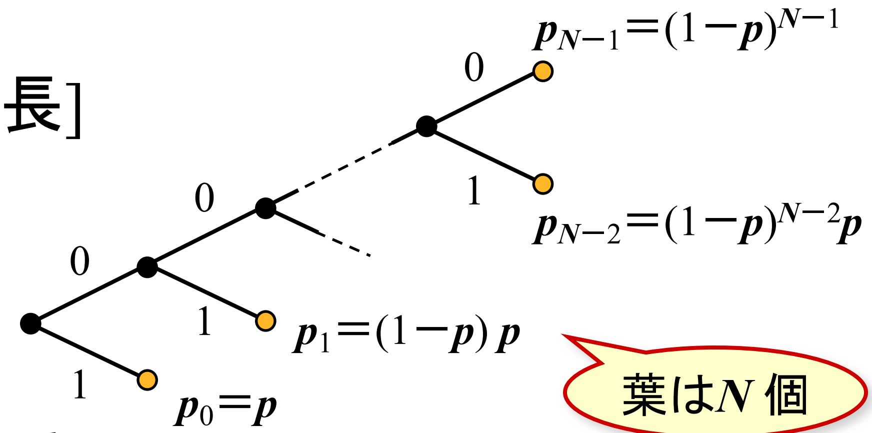
  情報源系列において同じ記号が連続する長さ(ランレングス)を符号化して送る方法

  - 右上の符号は長さ3までの 0の連続(ラン)に対するランレングス符号である。情報源系列 1, 01, 001, 000 はそれぞれ 0のランが 0, 1, 2, 3 である。
  - ランレングスをさらに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法と呼ぶ。

[ランレングス符号化法の平均符号長]

1, 0の発生確率が $p, 1-p$ ($p<1-p$)の無記憶情報源の、$N-1$までの 0 のランレングスを符号化

右の分節木で示されるようなN個の情報源系列に対して符号化



ランラングス符号化のための分節木

ランレングス符号化法(2)

これらN個の系列の平均系列長は

$$\bar{n} = \sum_{i=0}^{N-2}(i+1)\,p_i + (N-1)\,p_{N-1} = \underbrace{\sum_{i=0}^{N-2}(i+1)(1-p)^i p}_{0^i1\text{の系列}} + \underbrace{(N-1)(1-p)^{N-1}}_{0^{N-1}\text{の系列}}$$

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

となる。一方、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は定理4.2から

$H(S)=-p\log_2 p-(1-p)\log_2(1-p)$

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

を満たす。よって、1情報源記号あたりの平均符号長L_rは

$$L_r = \frac{L_N}{\bar{n}} < H(S) + \frac{1}{\bar{n}} = H(S) + \frac{p}{1-(1-p)^{N-1}}$$

ハフマンブロック符号化の場合と比較すると

$$L_h < H(S) + \frac{1}{n} = H(S) + \frac{1}{\log_2 N}$$

N:ハフマンブロック符号化において符号化すべき情報源系列の数

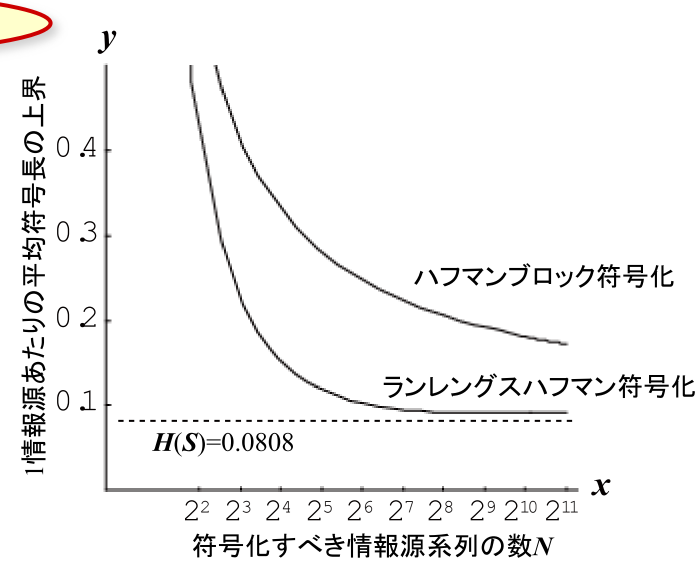


図5.11 ランレングス符号化とハフマンブロック符号化の比較(p=0.01)

[発展]タンストール木による符号化(1)

情報源系列を瞬時に(符号語を割り当てる情報源系列の境目で)分解可能

⇔分節木の葉のみに符号化すべき情報源系列が割り当たっている

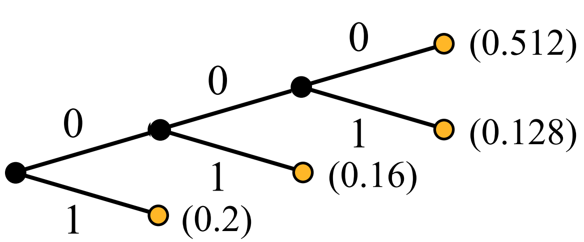
N個の符号化すべき情報源系列で平均長が最大なもの

⇔<u>N個の葉をもつ分節木において葉の深さの平均が最大なもの</u>

(平均は葉に対応する記号列の出現確率に関してとる)

タンストール木

M元情報源の場合、「最大確率が割り当てられている葉を、M個の葉を子としてもつノードに置き換える」ことを繰り返すことにより最適な分節木が得られる

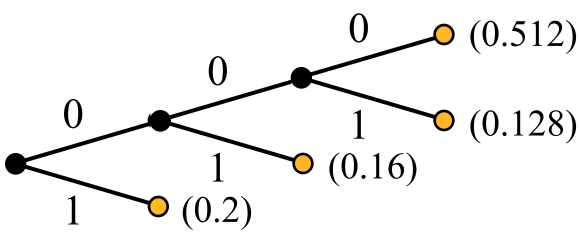


1の発生確率が0.2の2元無記憶定常情報源の4枚の葉をもつタンストール木

平均長n は

$$n \fallingdotseq 1\times 0.2+2\times 0.16+3\times 0.128+3\times 0.512$$
$$= 2.44$$

[発展]タンストール木による符号化(2)



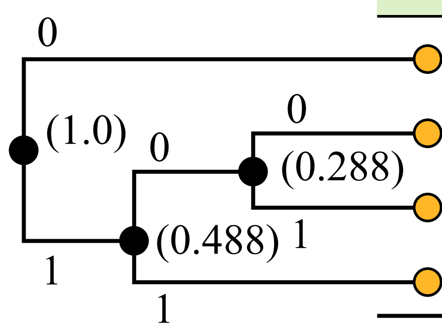
平均長　$\bar{n}$ ＝2.44

1の発生確率が0.2の2元無記憶定常情報源の
４枚の葉をもつタンストール木

右の符号の平均符号長＝1.776

よって1記号あたりの
平均符号長Lは

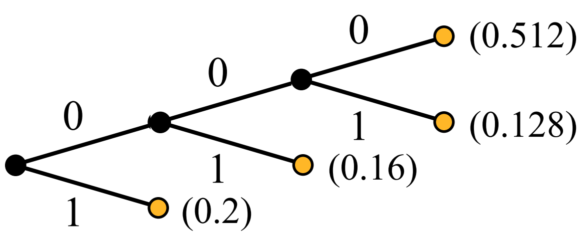
$$L = \frac{1.776}{2.44} = 0.728$$

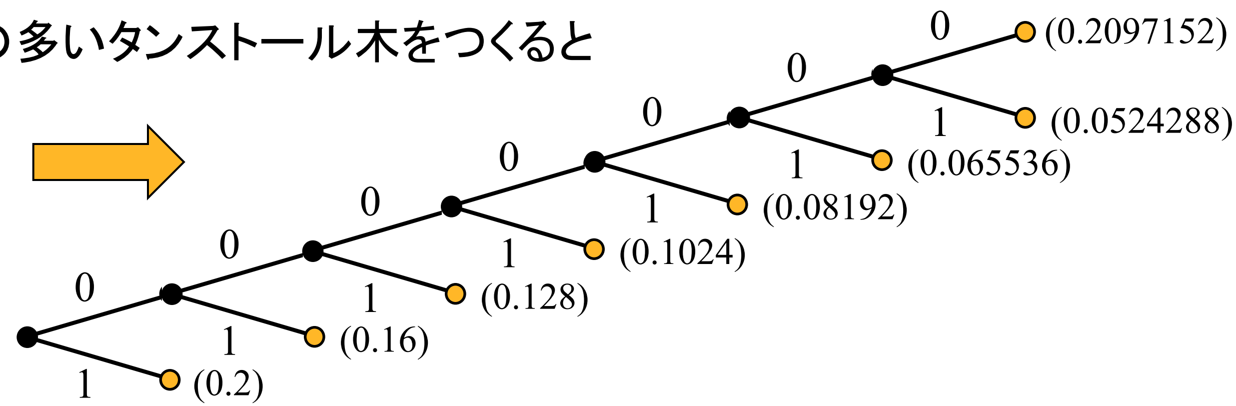


情報源系列	確率	ハフマン符号
0 0 0	0.512	0
0 0 1	0.128	100
0 1	0.16	101
1	0.2	11

[発展]タンストール符号

さらに葉の数の多いタンストール木をつくると





最大確率 0.512
最小確率 0.128] 0.384 ——→ 0.1572864 [0.2097152 最大確率 / 0.0524288 最小確率

差が縮まっていく

コンパクト符号と等長符号の平均符号長の差がなくなっていく

タンストール符号

① N枚の葉をもつタンストール木を構築 → ② 等長符号化

可変長系列-固定長符号(Variable-length-to-Fixed-length code: VF符号)

N→∞とすれば1情報源記号あたりの平均符号長はH(S)に収束

ひずみが許される場合とは？

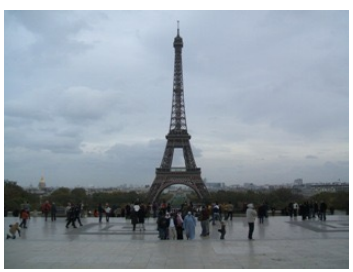
- ある程度ひずみを許しても、1情報源記号あたりの平均符号長を小さくしたい！（つまり、より小さくデータを圧縮したい）

テキストの場合

6時にスタバでね。ちゃんと来いよ。 → ひずみを許す符号化 → 0011010 1110111 0001・・・ → 復号 → 6時にスタバでねーちゃんと来いよ。

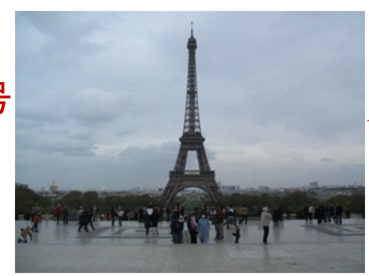
非常にまずい！

画像の場合



ビットマップファイル
（.bmp; 288 KB）

ひずみを許す符号 →



jpegファイル
（.jpg; 17.5 KB）

ほとんど問題ない

情報源符号化におけるひずみ

- 通信路でひずみが入るのではなく、符号化時にひずみを入れる

図: 情報源出力 x と情報源復号結果 y

- 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。

    $d(x, y) \geqq 0$

    $x = y$ 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。

    $$\overline{d} = \sum_{x}\sum_{y} d(x, y)P(x, y)$$

ひずみ測度の例

- 例1　情報源アルファベットを $A=\{0, 1\}$ 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

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

$$\bar{d}=\sum_x\sum_y d(x, y)P(x, y)=P(1, 0)+P(0, 1)$$

となる。これは、要するに、符号器の出力が元の情報源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。

ひずみが許される場合、情報源符号化定理はどうなるか？

［情報源符号化定理］

情報源S は、任意の正数εに対して、1情報源記号あたりの平均符号長Lが

$$H(S) \leqq L < H(S) + \varepsilon$$

となるような2元瞬時符号に符号化できる。

しかし、どのような一意復号可能な2元符号を用いても、

平均符号長が$H(S)$より小さくなるような符号化はできない。

$H(S)$が限界！

- ひずみを許せば、どのくらい平均符号長の限界を下げられるか？

S: 記憶の無い定常情報源、X:任意の時点においてSから出力される情報源記号

– 1情報源記号あたりの平均符号長の下限 ＝ 情報量$H(S)=H(X)$

– 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 $H(X|Y)$ のあいまいさが残る。

– 伝えられる情報の量は $I(X;Y) = H(X) - H(X|Y)$

ひずみを許した場合の限界！

ひずみが許される場合の情報源符号化定理

- 相互情報量$I(X;Y)$ が同じでも、平均ひずみ $\bar{d}$ は同じとは限らない
  ⇔ 平均ひずみ $\bar{d}$ が同じであっても、$I(X;Y)$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 $\bar{d}$ が

  $$\bar{d} \leqq D$$

  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、

  $$R(D) = \min_{\bar{d} \leqq D} \{I(X;Y)\}$$

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

[ひずみが許される場合の情報源符号化定理]

平均ひずみ $\bar{d}$ 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

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

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、
$\bar{d} \leqq D$ である限り、Lを$R(D)$より小さくすることはできない。

速度-ひずみ関数(1)

- $R(D)$ は、$\overline{d} \leqq D$ を満たす、あらゆる情報源符号化法を考えたときの $I(X;Y)$ の最小値。 では実際、これをどのようにして求めるのか？
- 情報源符号化法を変えると、条件付確率 $P(y|x)$ が変わる！
  → $I(X;Y)$ の最小化は、$\overline{d} \leqq D$ の条件の下で $P(y|x)$ を変えることで行う*
  → 下図のような通信路を考え、この通信路の特性を変えることで $I(X;Y)$ の最小化を計ると解釈できる。この仮想的通信路を試験通信路と呼ぶ。

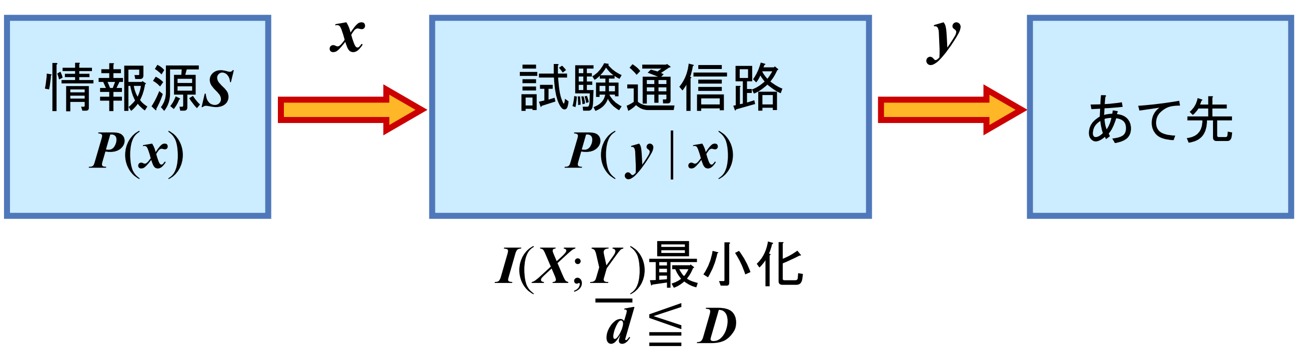


図: 試験通信路による速度・ひずみ関数$R(D)$ の解釈

* 任意の条件付確率 $P(y|x)$ を与えるような情報源符号化・復号法が存在する

速度-ひずみ関数(2)

- 条件付確率 $P(y|x)$ を用いれば、相互情報量 $I(X;Y)$ は

$$I(X;Y)=\sum_x\sum_y P(x,y)\log_2\frac{P(x,y)}{P(x)P(y)}$$

$$=\sum_x P(x)\sum_y P(y|x)\log_2\frac{P(y|x)}{P(y)} \quad \cdots\cdots \text{①}$$

と表せる。ここに $P(y)$ は

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

により計算できる。また、$\overline{d}\leqq D$ という条件は

$$\overline{d}=\sum_x P(x)\sum_y P(y|x)\,d(x,y)\leqq D \quad \cdots\cdots \text{②}$$

と表せる。また、条件付確率は

$$P(y|x)\geqq 0 \quad \cdots\cdots \text{③}$$

$$\sum_y P(y|x)=1 \quad \text{for each } x \quad \cdots\cdots \text{④}$$

を満たさなければならない。

- 式②～④の条件の下に、式①の $I(X;Y)$ を最小化すればよい！

$P(x,y)=P(x)P(y|x)$

$P(y|x)$ が変数
他は既知

一般には難しい

[例]記憶のない2元情報源の場合(1/3)

- 1, 0 を確率 p, $1-p$ で発生する記憶のない2元情報源Sを考える。
また、ひずみ測度としては

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

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、$0\leqq D\leqq 0.5$ が与えられたとき、$\overline{d}\leqq D$ の元での速度・ひずみ関数 $R(D)$を求めよう。

$D>0.5$ の場合は
誤る確率の方が
大きいことを意味する

- 相互情報量$I(X;Y)=H(X)-H(X|Y)$
$H(X)=\mathcal{H}(p)$ なので、
$H(X|Y)$ を最大化すればよい。
- ここで、Yは左図のように、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。

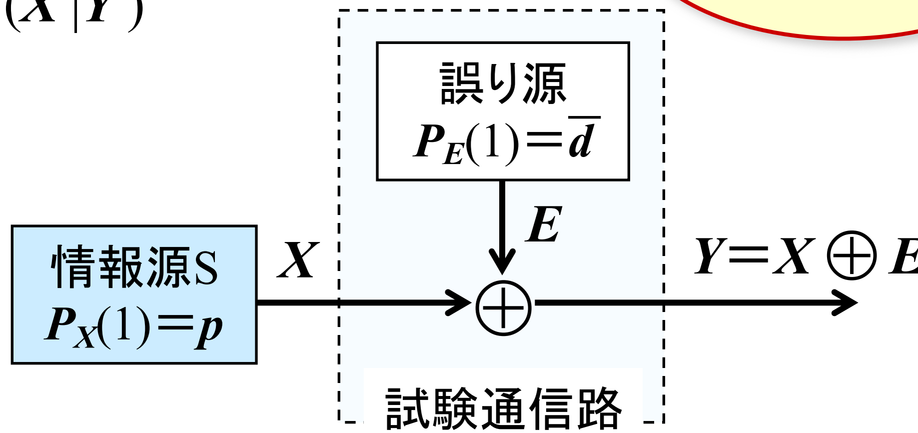


図: 2元情報源に対する試験通信路

[例]記憶のない2元情報源の場合(2/3)

- $Y = X \oplus E$ であるから、$X = Y \oplus E$ となる。したがって、

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

が得られる。

- 条件付きエントロピーの性質から$H(E \mid Y) \leqq H(E)$が成り立つ。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、

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

となる。それゆえ

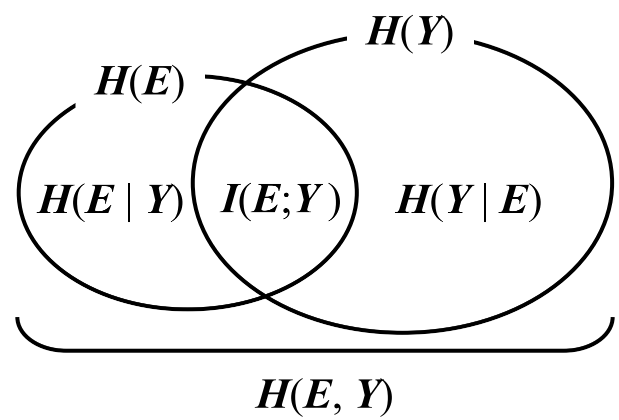
$$H(X \mid Y) \leqq \mathcal{H}(\overline{d})$$

を得る。$\overline{d} \leqq D \leqq 1/2$ なので、さらに

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

となる。したがって、相互情報量 $I(X;Y)$ は、

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



[例]記憶のない2元情報源の場合(3/3)

- $\bar{d}=D$で誤り源が無記憶定常で情報源Sと独立であるとき、等号が成立 ($I(X;Y)=\mathcal{H}(p)-\mathcal{H}(D)$) するので、
  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。
- 左図で分かるように、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、
  $$I(X;Y)=\lim_{n\to\infty} I(X_n;Y_n)/n$$
  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。

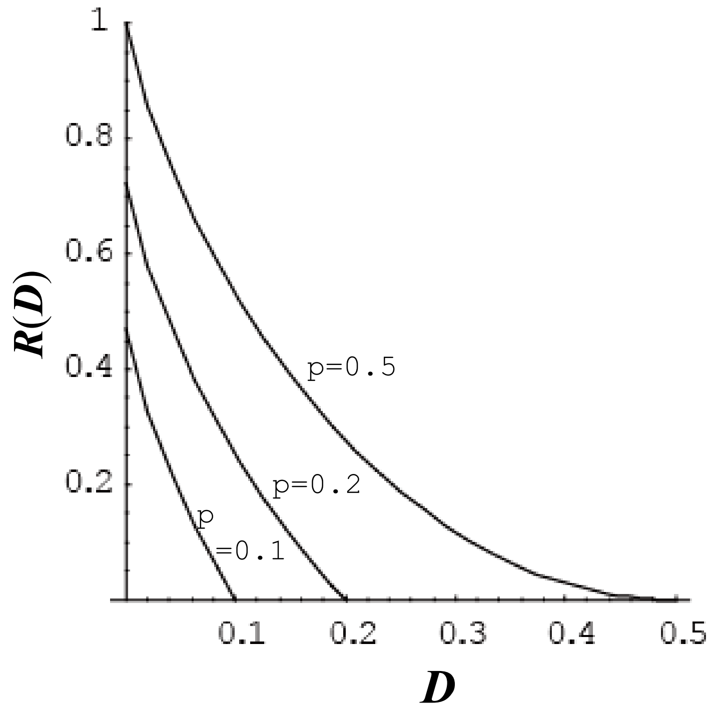


図: 2元情報源の速度・ひずみ関数

ひずみが許される場合の情報源符号化法

［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

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

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、$\overline{d} \leqq D$ である限り、Lを$R(D)$より小さくすることはできない。

- ひずみが許される場合の情報源符号化定理は、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。では、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！
    - ベクトル量子化
    - 変換符号化
    - 予測符号化
    - etc.

参考図書
「情報・符号・暗号の理論」
今井秀樹 著、電子情報通信学会