パターン計測論 講義資料　2015. 5. 20

第5章　情報のエントロピー

篠田 裕之

http://www.hapis.k.u-tokyo.ac.jp/

hiroyuki_shinoda@k.u-tokyo.ac.jp

情報のエントロピー

N 次元ベクトル $(a_1, a_2, \ldots, a_N)$ のパターンの数を考える

エントロピー = ある制約のもとで有意に生じる場合の数（の対数）

制約の例

・総エネルギー 一定 ⟶ 物理におけるエントロピー
・発生の確率 ⟶ 今回の話

2. 情報のエントロピー　準備　1

シャノンの発見

「事象の数」には
① 生じうる全ての場合の数を数え上げたもの
② 有意な確率*で生ずる場合の数
の２種類がある。

＊「有意な確率」の意味は徐々に明らかになる。

準備

n 回のコイン投げの結果を記録するのに何ビット必要か？

起こり得る結果　----　2^n 通り

記録に必要なビット数 ＝ n

2. 情報のエントロピー　準備　１

（前スライドつづき）

表の出る確率が１の場合、結果を記録するのに何ビット必要か？

n 回の試行で起こりうる結果 ---- １通り

結果の記録に必要なビット数 ---- 0

したがって、想定される場合の数が 2^n であったとしても、その記録に必要なビット数の最小値は n とは限らない。

では、表の出る確率が 0.1 の場合は？

２. 情報のエントロピー　準備 ２

準備 --- 半端なビット数

１～６までの目が出るサイコロを振る
結果を記録するのに何ビット必要か？

１回の結果の記録に必要なビット数 - - - 3ビット

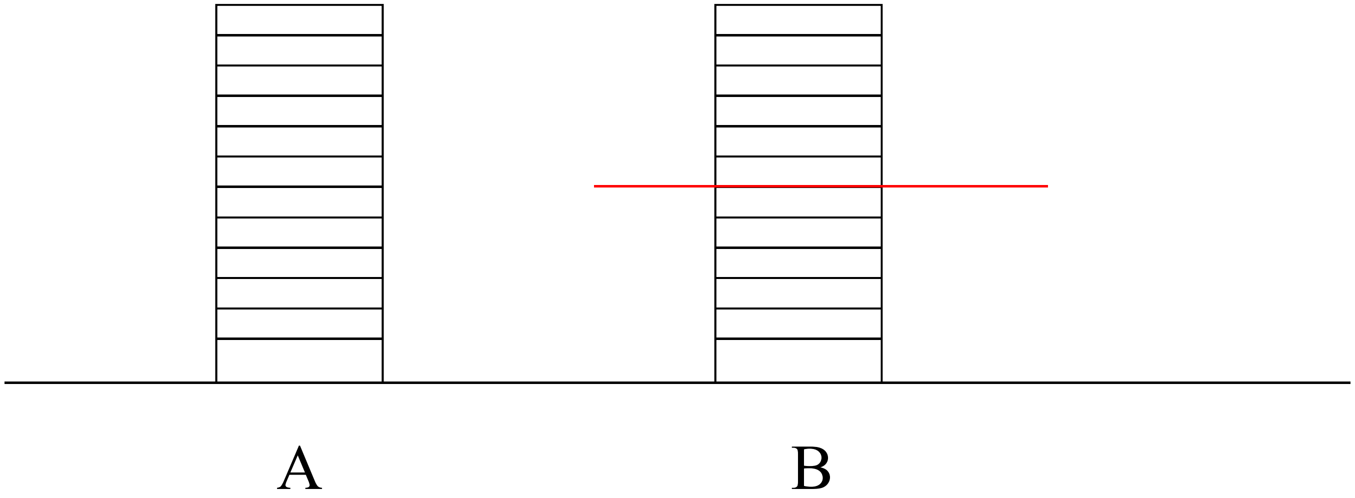
100 回の結果の記録に必要なビット数

$$\log_2 6^{100} = 100 \log_2 6 = 258.5$$ より 259 ビット

⇒ $\boxed{\text{１回あたり 2.585 ビット}}$

＊多数回の試行に対し１回あたりの平均値を考える。そうすれば１回あたりの場合の数が 2^n ちょうどでない場合にも「１回あたりの正味の情報量」を考えることができる。

クイズ



3割が表
7割が裏

半分から上が表、下が裏

コインを見ないでA, B間でコインを入れ替え、
A, Bともに表と裏が半分ずつ入るように
するにはどうすればよいか？

２．情報のエントロピー　準備　３

表と裏の回数がちょうど同じである場合の数と ２ⁿ の比較

$$ {}_{n}\mathrm{C}_{n/2} = \frac{n!}{\{(n/2)!\}^{2}} $$

この数の対数をスターリング公式で近似計算

$$ \ln {}_{n}\mathrm{C}_{n/2} = n\ln n - n - 2\left(\frac{n}{2}\ln\frac{n}{2} - \frac{n}{2}\right) + O(\ln n) $$

$$ = n\ln 2 + O(\ln n) $$

（対数の底はｅ）

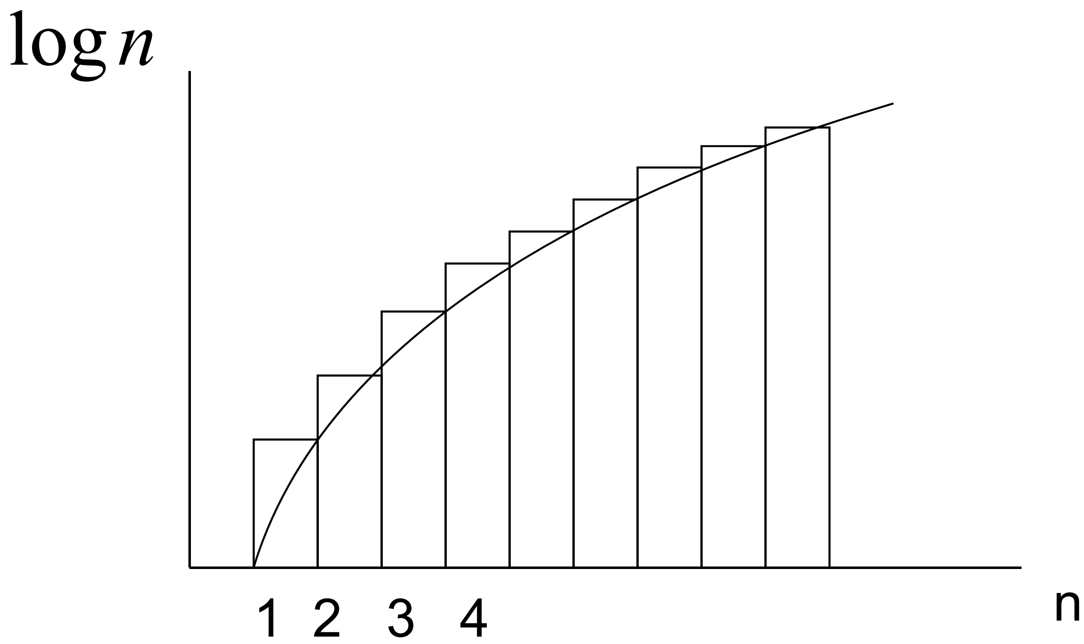
$$ \ln 2^{n} = n\ln 2 $$

２．情報のエントロピー　準備 ３

スターリング公式

$$\ln n! = \ln n + \ln(n-1) + \cdots + \ln 1$$

$$= n \ln n - n + O(\ln n)$$

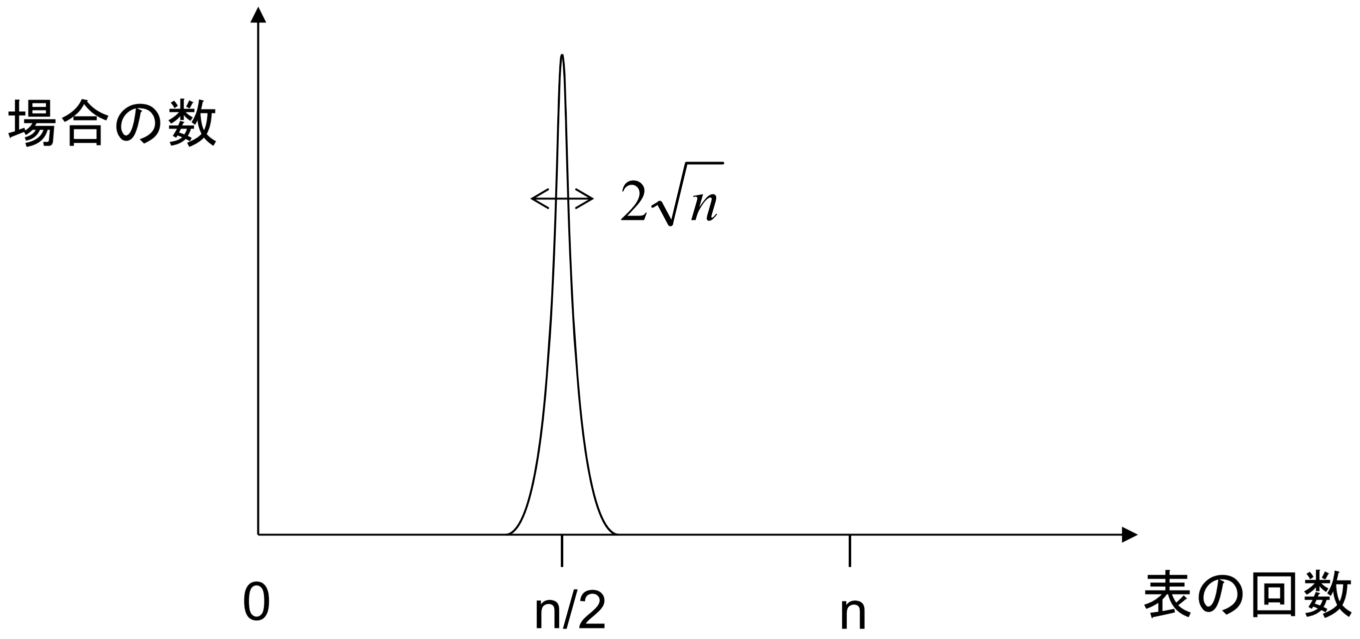


２．情報のエントロピー　準備 ３

表と裏の回数がちょうど同じである場合の数と ２ⁿ の比較

$$\frac{1}{n}\ln\left({}_n\mathrm{C}_{n/2}\right)\to\ln 2\qquad\left(=\frac{1}{n}\ln 2^n\right)$$

問）上式を、下図の分布から説明せよ



前スライドの分布の横幅に関する考察

n 回のランダムウォークで原点から d 離れる確率 p(d)

$$\log_e p(d) = \log_e \frac{{}_n\mathrm{C}_{d/2+n/2}}{2^n}$$

$$\approx -\frac{d^2}{2n} + O(\log n, \log d)$$

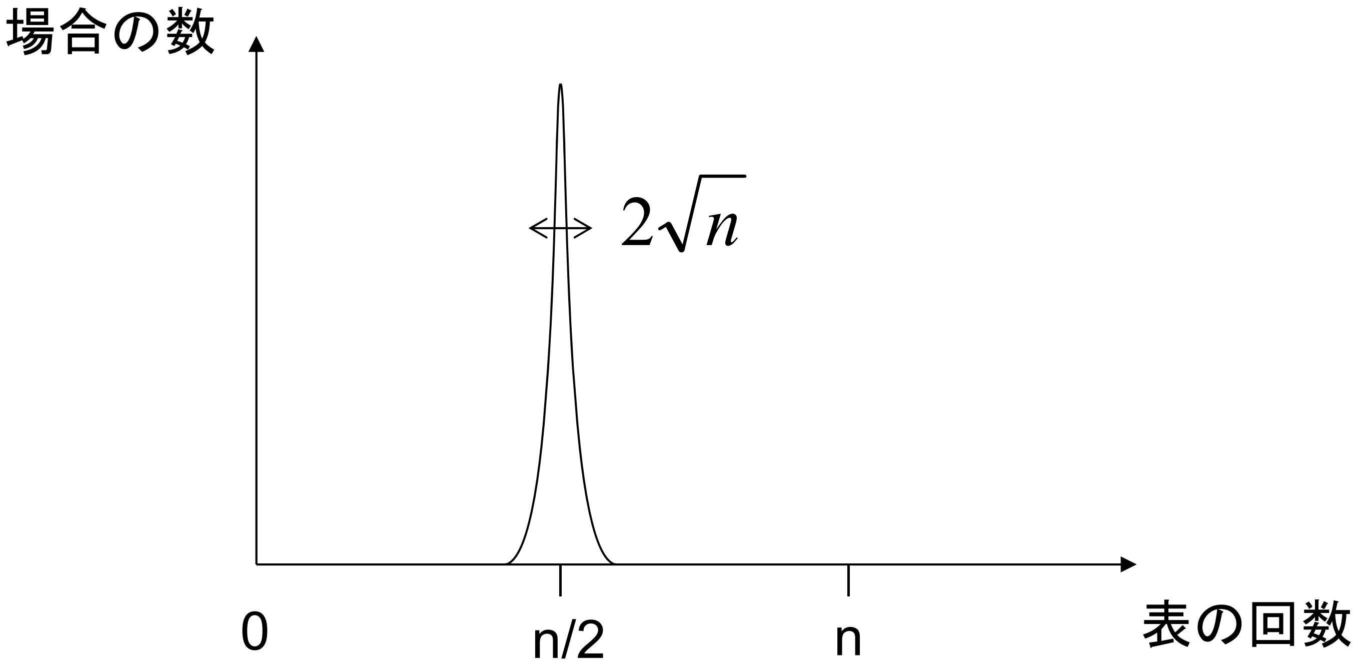
（2項分布を n >> d として
スターリング近似）

n = 10,000　d = 1,000　　とすると $\dfrac{1}{5.2\times 10^{21}}$

２．情報のエントロピー　準備　３

問の答え

$$\log(\text{有意に生じる場合の数}) \approx \log(\text{丁度半分が表の場合}) + \log\sqrt{n}$$

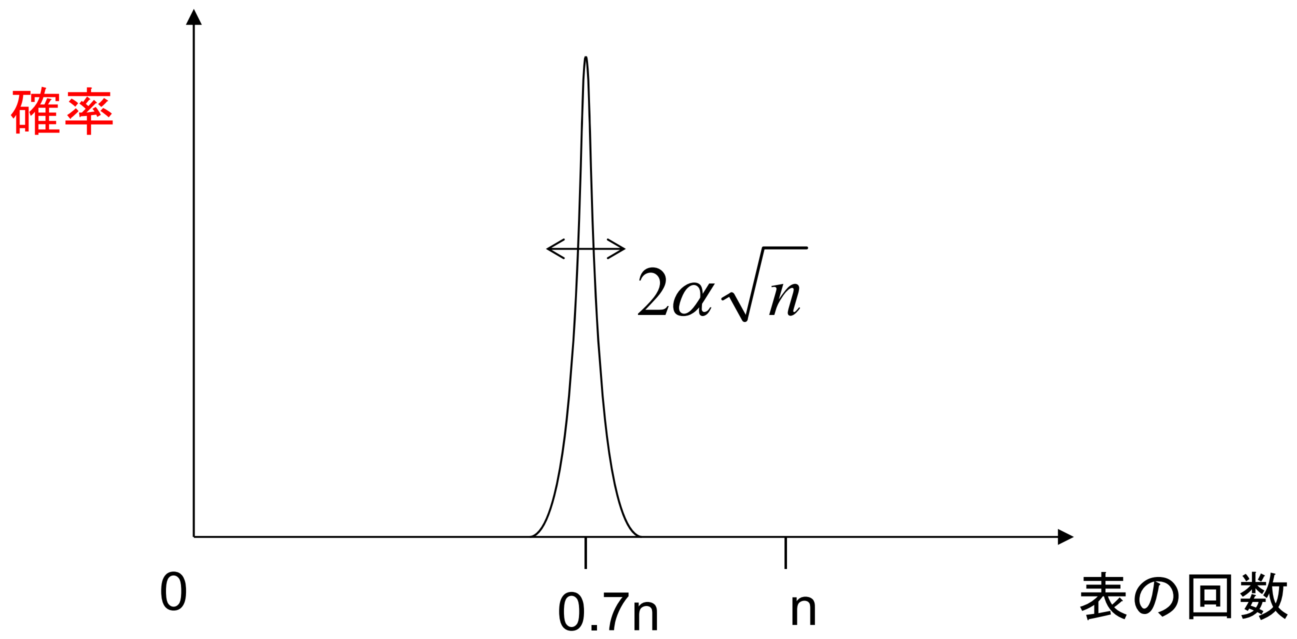


2. 情報のエントロピー　準備 4

問題

表の出る確率が 0.7 のコインがある。
このコインを n 回投げたときの結果を記録するのに何ビット必要か？

→ ちょうど 0.7 n 回表が出る場合の数を数えて対数をとればよい　??



２. 情報のエントロピー　準備 ５

全体の試行 n に対し np 回表が出る場合の数（の対数）

$$\ln {}_n\mathrm{C}_{pn} = \ln \frac{n!}{(n-pn)!(pn)!}$$

$$= n\ln n - n - \{(n-pn)\ln(n-pn) - n + pn + pn\ln pn - pn\} + O(\ln n)$$

$$= np\ln\frac{1}{p} + n(1-p)\ln\frac{1}{1-p} + O(\ln n)$$

$$\boxed{\frac{1}{n}\ln {}_n\mathrm{C}_{pn} = p\ln\frac{1}{p} + (1-p)\ln\frac{1}{1-p}}$$

2. 情報のエントロピー

1 回の試行に対し、m 個の事象 e_i が存在し、それぞれが
確率 $p_1, p_2, \ldots p_m$ で生ずる $(p_1 + p_2 + \ldots + p_m = 1)$

（注意）「１回の試行」が多数回の行為をまとめたものであってもよい

n 回の試行で各事象 e_i が np_i 回ずつ生じる場合の数 N は以下のとおり

$$N = \frac{n!}{(n \cdot p_1)!(n \cdot p_2)! \cdots (n \cdot p_m)!}$$

$$H \equiv \frac{1}{n}\log_2 N = \frac{\ln n - 1 - \sum_{i=1}^{m}\left(p_i \ln np_i - p_i\right)}{\ln 2} = \sum_{i=1}^{m} p_i \log_2 \frac{1}{p_i}$$

H を「試行１回あたりのエントロピー」と考えることができる

（注意：ただし試行が多数回ないと意味がない）

m 個の排他的事象 e_i が確率　$p_1, p_2, \ldots p_m$ で発生するとき

$$H \equiv \sum_{i=0}^{m} p_i \log_2 \frac{1}{p_i}$$

とすると、n 回の試行において有意な確率で起こりうる
場合の数の対数

$$E \approx nH$$

m 通りの排他的な結果が確率 $p_1, p_2, \ldots p_m$ で発生する。
この試行を n 回繰り返した結果を記録するのに必要な記録媒体の容量は以下の通り。

① 起こりうる全ての結果を記録するには

$$\log_2 m^n$$ ビット必要

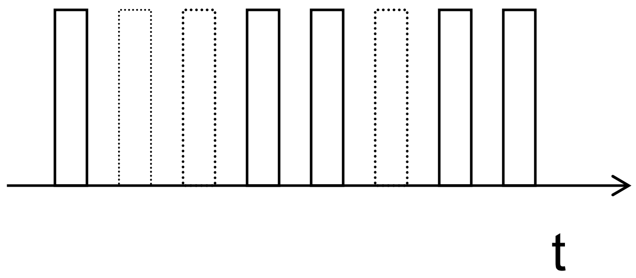
---

② 十分 n が大きい場合には

$$n \times (H + \delta)$$ ビット

でほとんどの場合を網羅できる符号化方法が存在する。
どんな小さな δ に対しても、前記符号化方法で網羅されない結果が生じる確率は、n を大きくすればいくらでも小さくなる。

通信容量



単位時間に C_0 ビット伝送可能な通信路

ノイズが存在しなければ

通信路容量 $C = C_0$

1 ビットの伝送あたり 確率 p でビットの反転が生じるとすると
単位時間に（誤りなく）伝送可能な情報量はどのように与えられるか？

３. 通信容量

> 100 回に１回誤りを起こす通信路で 1,000,000 個のデータを誤りなく伝送するにはどうすればよいか？

十分大きな数 n ビット分のデータを送る

誤りが起こる率を p とする

全 n ビット中 np ビットに誤りが生じる

3. 通信容量

誤り率 p で n ビットを伝送するとき、誤り無く伝送可能な状態数の上限

全体：n ビット

$$\overbrace{1\,0\,0\,0\,1\,1\,0\,1\,1\,1\,1\,0\,0\,1\,0\,1\,0\,0\,1\,0\,0\,1\,0\,0\,1\,0\,0}^{}$$

符号語間ハミング距離 $> np$ となるようにしておけば伝送誤りは生じない

① ノイズが無いとき

2^n 通りのパターンが送れる

② ノイズがあるときの上限

$$K = \frac{2^n}{W} \text{ 通り}$$

W: 有意な確率で起こりうる反転のパターンの数

３．通信容量　　（前スライドつづき）

$$\log K = \log\frac{2^n}{W} = n - \log W$$

W：「反転する」が np 回起こる場合の数

$$\frac{1}{n}\log W = p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p}$$

$\longrightarrow$ １ビット伝送の度に失われるエントロピー

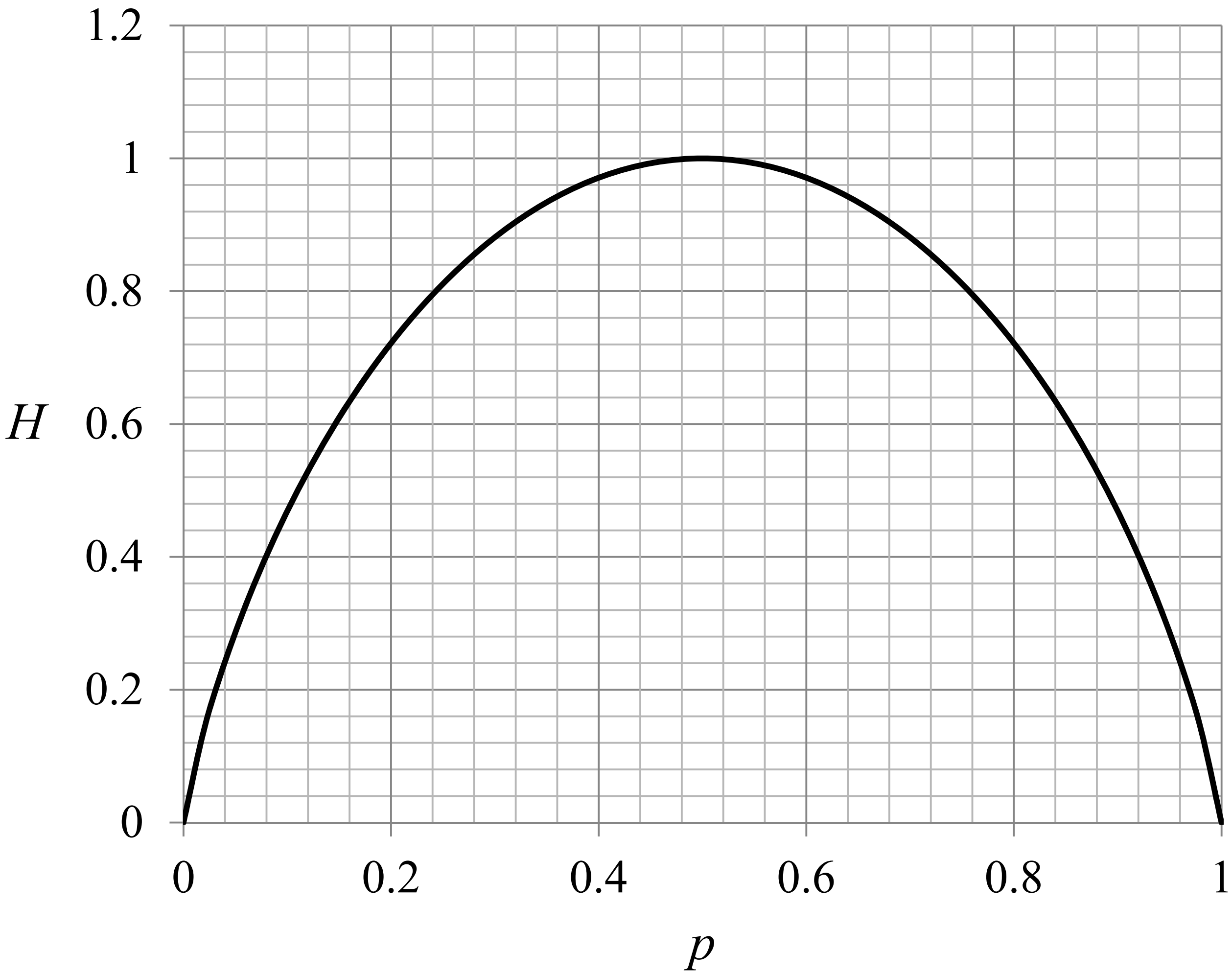
ノイズが存在するときの通信路容量（誤りなく送れる情報量の理論限界）

$$C = C_0 - C_0\left(\underline{p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p}}\right)$$

ノイズの１サンプルあたりのエントロピー

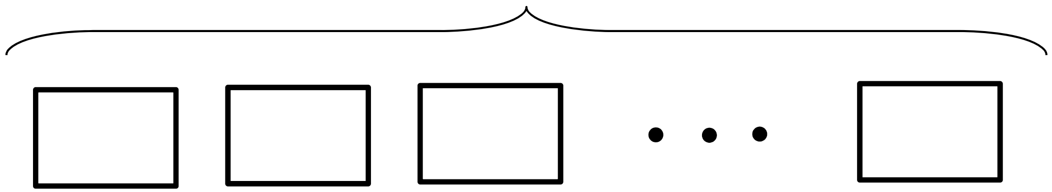
３．通信容量

$$H = p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p}$$ のグラフ



より一般の場合

全体で u ユニットを伝送



１ユニットあたり m ビットを伝送

① 第２章「戦略１」で符号を伝送している場合
１ユニットN 点あたり m ビットの情報を伝送しているのであれば、各ビットあたりの誤り率 p のもとで mu ビットを伝送する問題に帰着される。

３. 通信容量

②第２章「戦略２」で符号を伝送している場合
１ユニットN 点の各基底成分の振幅に対し q ビットが割り当てられ、隣接する振幅値に誤って読み取られる確率を α とする。(ここでは２段階誤る確率は０とする）このようなユニットを u ユニット伝送することは、上記誤り特性を有する q ビット１組の符号を $M=Nu$ 回伝送することと同じである。この誤りのパターンを全て数え上げたときの総数をW とすると、M 回中 αM 回誤りが生じることに注意して

$$W = 2\ {}_M\mathrm{C}_{\alpha M}$$

である。

$$\frac{1}{M}\log {}_M\mathrm{C}_{\alpha M} = \alpha\log\frac{1}{\alpha} + (1-\alpha)\log\frac{1}{1-\alpha}$$

であるから、ノイズのエントロピーは伝送量１ビットあたり

$$\frac{1}{qM}\log W = \frac{1}{q}\left(\alpha\log\frac{1}{\alpha} + (1-\alpha)\log\frac{1}{1-\alpha}\right)$$

まとめ

1. 正味の状態数（の対数）をエントロピーとよぶ。

m 個の排他的事象 e_i が確率　$p_1, p_2, \dots p_m$　で発生するとき

$$H = \sum_{i=0}^{m} p_i \log \frac{1}{p_i}$$

2. 一定確率で誤りが発生する伝送路においても事実上誤り無く情報を伝送することができる。

単位時間に C_0 ビットを伝送する伝送路において各ビットに確率 p で誤りが発生するとき、単位時間に誤りなく伝送可能な情報量は

$$C = C_0 - C_0 \left(p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p} \right)$$

例題1

1) 等確率に表裏が出るコインをふる。
この情報源のエントロピーは１回投げるあたりいくらか？

2) １から６まで等確率に目が出るサイコロをふる。
この情報源のエントロピーは１回投げるあたりいくらか？

$$\log_2 6 = 2.58$$

3) １から６までそれぞれ確率 0.3, 0.3, 0.3, 0.05, 0.05, 0 で出るサイコロがある。
この情報源のエントロピーは１回投げるあたりいくらか？

（ $\log \dfrac{1}{0.3} = 1.74$ ， $\log 20 = 4.32$ を用いよ ）

例題２

A C

2桁のスロットマシーンにA, B および C の３通りの記号が現れる。

1) A, B および C の出現確率が等しく、各窓に現れる記号に相関が無い場合、一回の試行あたりのエントロピーはいくらか？

2) A, B, C の出現確率が下表の場合、一回あたりのエントロピーはいくらか？

		A	B	C	右の窓
	A	1/27	4/27	4/27	
左の窓	B	4/27	1/27	4/27	
	C	4/27	4/27	1/27	

例題２　解答

1)　$2\log 3 = 3.17$

2)　$\dfrac{3}{27}\log 27 + \dfrac{4\times 6}{27}\log\dfrac{27}{4} = 3\log 3 - 16/9 = 2.98$