通信伝送工学第 3 回　情報理論の基礎２：データ圧縮（教科書 p. 575-581）

【課題の解答例】(p. 619-620)

9.5

図 P9.5 の量子化器出力のエントロピーHは，次の式で表される．

$$H=\sum_{k=1}^{4}P\left(X_k\right)\log_2\frac{1}{P\left(X_k\right)}$$

ただし，

$$X_1=-1.5,\ X_2=-0.5,\ X_3=0.5,\ X_4=1.5$$

は量子化後の出力レベルを表す．

各標本は独立の標準ガウス分布（平均 0・分散 1）

$$f\left(x\right)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$$

に従う．ガウス分布の累積確率関数は誤差関数(error function，数表は p. 766)

$$\mathrm{erf}\left(u\right)=\frac{2}{\sqrt{\pi}}\int_0^u\exp\left(-t^2\right)\mathrm{d}t$$

を用いて

$$P\left(x<a\right)=\int_{-\infty}^{a}f\left(x\right)\mathrm{d}x=\begin{cases}\dfrac{1}{2}\left\{1+\mathrm{erf}\left(\dfrac{a}{\sqrt{2}}\right)\right\}, & \left(a\ge 0\right)\\ \dfrac{1}{2}\left\{1-\mathrm{erf}\left(-\dfrac{a}{\sqrt{2}}\right)\right\}, & \left(a<0\right)\end{cases}$$

と表すことができる．これより，

$$P\left(X_1\right)=P\left(X_4\right)=\frac{1}{2}\left\{1-\mathrm{erf}\left(\frac{1}{\sqrt{2}}\right)\right\}=0.1587,\quad P\left(X_2\right)=P\left(X_3\right)=\frac{1}{2}\mathrm{erf}\left(\frac{1}{\sqrt{2}}\right)=0.3413$$

を得る（表計算ソフト等で計算可，あるいは p. 766 の数表から近似可）．エントロピーは，

$$H=2\times\left(0.1587\log_2\frac{1}{0.1587}+0.3413\log_2\frac{1}{0.3413}\right)=1.902\left[\mathrm{bit}\right]$$

となる．

9.7

(a) 情報源のエントロピーは次のようになる．

$$H=\left(0.7\log_2\frac{1}{0.7}+0.15\log_2\frac{1}{0.15}+0.15\log_2\frac{1}{0.15}\right)=1.181\left[\mathrm{bit}\right]$$

(b) 二次拡張情報源のエントロピーを求めるために，アルファベット 2 字の結合確率を求める．

$$p\left(s_0s_0\right)=0.7\times0.7=0.49$$

$$p\left(s_0s_1\right)=p\left(s_0s_2\right)=p\left(s_1s_0\right)=p\left(s_2s_0\right)=0.7\times0.15=0.105$$

$$p\left(s_1s_1\right)=p\left(s_1s_2\right)=p\left(s_2s_1\right)=p\left(s_2s_2\right)=0.15\times0.15=0.0225$$

これより二次拡張情報源のエントロピーは次のようになる．

$$H=0.49\log_2\frac{1}{0.49}+4\times0.105\log_2\frac{1}{0.105}+4\times0.0225\log_2\frac{1}{0.0225}=2.362\equiv1.181\times2\left[\mathrm{bit}\right]$$

【講義の要点】

データ圧縮 (data compression, p. 575-581)

物理的な情報源の冗長性 (redundancy): 伝送の前に取り除かれるべき

無損失データ圧縮 (lossless data compression): 元の情報が完全に復元可能

語頭符号 (prefix coding, p. 575-578)

離散無記憶情報源: アルファベット$\{s_0, \ldots, s_{K-1}\}$，確率$\{p_0, \ldots, p_{K-1}\}$

一意に復号可能 (uniquely decodable): 異なるアルファベットに対し符号語 (code word) が異なる

シンボル s_k ⇒ 符号語$\left(m_{k_1}, m_{k_2}, \ldots, m_{k_n}\right)$， m_k : 0 or 1，n: 符号長

符号語の語頭 (prefix): $m_{k_1}, m_{k_2}, \ldots, m_{k_i}$， $i \le n$

語頭符号: どの符号語も他の符号語の語頭とならない符号化

例: 表 2 の三つの符号化のうち符号 II だけが語頭符号

語頭符号の復号器 (decoder): 判定木 (decision tree)

～ 図 9.4: 初期状態 (initial state) から終端状態 (terminal states) への状態遷移

例: 表 2 の符号 II で 1011111000 を復号

語頭符号は一意に復号可能: 逆は必ずしも成り立たない

語頭符号の符号長: クラフト・マクミランの不等式 (Kraft-McMillan Inequality) (9.22)

語頭符号は瞬時復号可能 (instantaneous code)

語頭符号の符号長とエントロピー: 情報源符号化定理＋クラフト・マクミランの不等式 (9.23)

補足: 右側不等号は$l_k = \lfloor -\log_2 p_k \rfloor$で成立

語頭符号が整合 (matched) する条件 (9.24)

拡張符号 (extended code):シンボル列を符号化 (9.28) ⇒ エントロピーに収束 (9.29)

ハフマン符号 (Huffman coding, p. 578-580)

各シンボルの符号長をその情報量とほぼ同じ長さとする

符号化アルゴリズム (encoding algorithm): 図 9.5 ハフマン木

1. シンボルを出現確率の高い順に並べる．もっとも確率の低い二つのシンボルに 0/1 を割り当てる．
2. ステップ 1.の二つのシンボルを一つの新しいシンボルと見なし，その出現確率を旧シンボルの出現確率の和とする．
3. シンボルが二個となるまで 1.と 2.を繰り返す．

ハフマン符号は一意でない: 0/1 の割り当て法，出現確率が同一の場合の扱い

符号長の分散 (variance) の最小化: 合成シンボルをハフマン木の上に置く

レンペル・ジブ符号 (Lempel-Ziv coding, p. 580-581)

ハフマン符号の問題点: 情報源の確率的性質は事前 (a priori) には不明，文字間・語間の相関無視

レンペル・ジブ符号の考え方:

1. データ列を解析 (parse) し，初めて出現する最短部分列に分割．符号表 (code book) 作成．
2. 符号表の各部分列を，それより過去に出現した部分列の並びで表わす．
3. 各部分列の最終シンボル: イノベーションシンボル (innovation symbol) ～他の部分列と区別．
4. 符号語は長さ一様．最終ビットをイノベーションシンボルとし，残りのビットは過去に出現した部分列へのポインタ．

レンペル・ジブ符号の特徴:

- 可変長情報源を固定長符号化～同期送信 (synchronous transmission)
- 適応的 (adaptive)～高次相関も自動的に考慮 ⇒ ハフマン符号を駆逐

本日の課題 (p. 620-621)

9.10, 9.12

- 提出締切 11/10（月）: 南 6 号館 1F メールボックス S6-4