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】(p. 618-620)
9.2

$$I\left(s_k\right)=\log_2\left(\frac{1}{p}\right)\text{bits}$$

s_k	s_1	s_2	s_3	s_4
p_k	0.4	0.3	0.2	0.1
$I\left(s_k\right)$ bits	1.322	1.737	2.322	2.233

9.3

$$\begin{aligned}H\left(S\right)&=\sum_{k=0}^{3}p_k\log_2\left(\frac{1}{p_k}\right)\\&=\frac{1}{3}\log_2 3+\frac{1}{6}\log_2 6+\frac{1}{4}\log_2 4+\frac{1}{4}\log_2 4\\&=0.528+0.431+0.5+0.5\\&=1.959\text{ bits}\end{aligned}$$

9.8
文字列と対応する音声を比較すると，後者は文字列が表わす情報の他に発話者の性質（声質，音の高さ，発話の速さなど）を含むため，必要とされる情報量は大きくなる．

【講義の要点】
<u>復習</u>
情報量 (amount of information) (9.4)
エントロピー (entropy): 情報量の期待値 (9.9)　　情報源の性質を表わす
情報源符号化定理: 離散無記憶情報源のエントロピー≦平均符号長 (9.20)

<u>データ圧縮 (data compression, p. 575-581)</u>
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, …, s_{K-1}}，確率{p_0, …, 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)

9.9 Consider a discrete memoryless source whose alphabet consists of *K* equiprobable symbols.
(a) Explain why the use of a fixed-length code for the representation of such a source is about as efficient as any code can be.
(b) What conditions have to be satisfied by K and the code-word length for the coding efficiency to be 100%?

9.11 Consider a sequence of letters of the English alphabet with their probabilities of occurrence as given here:

Letter	a	i	l	m	n	o	p	y
Probability	0.1	0.1	0.2	0.1	0.1	0.2	0.1	0.1

Compute two different Huffman codes for this alphabet. In one case, move a combined symbol in the coding procedure as high as possible, and in the second case move it as low as possible. Hence, for each of the two codes, find the average code-word length over the ensemble of letters.

9.16 Consider the following binary sequence
11101001100010110100...
Use the Lempel-Ziv algorithm to encode this sequence. Assume that the binary symbols 0 and 1 are already in the codebook.

- 提出締切 10/29（月）: 南 6 号館 1F メールボックス S6-4

通信伝送工学 演習 2007 年 10 月 23 日

学籍番号 _____ 氏名 _____

演習 9.10
Consider the four codes listed below:

Symbol	*Code I*	*Code II*	*Code III*	*Code IV*
s_0	0	0	0	00
s_1	10	01	01	01
s_2	110	001	011	10
s_3	1110	0010	110	110
s_4	1111	0011	111	111

(a) Two of these four codes are prefix codes. Identify them, and construct their individual decision trees.
(b) Apply the Kraft-McMillan inequality to codes I, II, III, and IV. Discuss your results in light of those obtained in part (a).