情報理論 配布資料 #6

担当：井上 純一 (情報科学研究科棟 8-13)
URL：http://chaosweb.complex.eng.hokudai.ac.jp/~j_inoue/

平成 17 年 5 月 30 日

目 次

演習問題 5 の解答例

1. n 確率変数に関するエントロピー $H_M(X_1, X_2, \cdots, X_n)$ は

$$H_M(X_1, X_2, \cdots, X_n) = -\sum_{x_1}\sum_{x_2}\cdots\sum_{x_n} P(x_1, x_2, \cdots, x_n)\log_M P(x_1, x_2, \cdots, x_n) \quad (190)$$

で定義されるが, 確率変数 $X_1, X_2, \cdots, X_n$ が全て独立で, 同一の分布 P から生成されるならば, 上記のエントロピーの中に現れる同時分布が

$$P(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} P(x_i) \quad (191)$$

と書けることから, これを (190) 式に代入して

$$\begin{aligned}
H_M(X_1, X_2, \cdots, X_n) &= -\sum_{x_1}\sum_{x_2}\cdots\sum_{x_n}\prod_{i=1}^{n} P(x_i)\log_M \prod_{i=1}^{n} P(x_i) \\
&= -\sum_{x_1}\sum_{x_2}\cdots\sum_{x_n}\prod_{i=1}^{n} P(x_i)\left\{\sum_{i=1}^{n}\log_M P(x_i)\right\} \\
&= -\sum_{x_1} P(x_1)\log_M P(x_1)\left\{\sum_{x_2} P(x_2)\sum_{x_3} P(x_3)\cdots\sum_{x_n} P(x_n)\right\} \\
&- \sum_{x_2} P(x_2)\log_M P(x_2)\left\{\sum_{x_1} P(x_1)\sum_{x_3} P(x_3)\cdots\sum_{x_n} P(x_n)\right\}\cdots \\
&- \sum_{x_n} P(x_n)\log_M P(x_n)\left\{\sum_{x_1} P(x_1)\sum_{x_2} P(x_2)\cdots\sum_{x_{n-1}} P(x_{n-1})\right\} \quad (192)
\end{aligned}$$

となるが, 確率の規格化条件から, $\sum_{x_i} P(x_i) = 1$ となることに注意すれば

$$\begin{aligned} H(X_1, X_2, \cdots, X_n) &= -\sum_{x_1} P(x_1) \log_M P(x_1) - \cdots - \sum_{x_n} P(x_n) \log_M P(x_n) \\ &= -n \sum_{X} P(X) \log_M P(X) \\ &= nH(X) \end{aligned} \tag{193}$$

が成立する.

2. まず, $\mathcal{B} = \{\mathsf{a}, \mathsf{b}, \mathsf{c}, \mathsf{d}, \mathsf{e}\}$ であるから $K = 5$ であり, 各符号の符号長は表に与えられているわけであるから, クラフト不等式 : $\sum_{i=1}^{M} K^{-l_i} \leq 1$ は

$$5^{-1} \times 2 + 5^{-2} + 5^{-3} \times 3 + 5^{-5} = \frac{826}{3125} < 1 \tag{194}$$

であるから, 満たされる. よって, 表に与えられた**符号長を実現する一意復号可能な符号は存在する**ことがわかる. しかし, 存在するからといって, 具体的に与えられた符号が一意復号可能であるか, 否かはわからない[1] . ちなみに, 語頭条件は大丈夫であろうか ? 念のため, これを調べるために次のような考察を行う. S_0 をオリジナルな符号からなる集合とし, これとは別の集合 $S_1, S_2, \cdots, S_n$ (n は情報源アルファベットの個数. 今の例の場合には $n = 7$) を次の規則に従って作っていこう. まず, S_1 を S_0 の中の符号 W_i が他の記号 W_j と記号列 A を用いて $W_j = W_i A$ の関係を満たすとき, A を S_1 の要素に入れる. この例では $S_1 = \{\mathsf{d}, \mathsf{ad}\}$ が該当する. 続いて, $n > 1$ の場合には S_0 の中の符号 W とある記号列 B を用いて S_{n-1} の要素 A が $A = WB$ と書ける場合, B を S_n の要素に加えるか, あるいは, S_0 の要素 $W^{'}$ が S_{n-1} の要素 $A^{'}$ とある記号列 $B^{'}$ を用いて $W^{'} = A^{'}B^{'}$ と書けるとき, $B^{'}$ を S_n の要素に入れることと約束する. つまり, S_n $(n \geq 2)$ に対しては

(I)

$$A = WB \in S_{n-1}, W \in S_0 \quad \Rightarrow \quad A \in S_n$$

(II)

$$W^{'} = A^{'}B^{'} \in S_0, A^{'} \in S_{n-1} \quad \Rightarrow \quad B^{'} \in S_n$$

のように各操作 (I)(II) を定義すれば, S_2 は (II) が該当し, $W^{'} = \mathsf{deb}$ に対し, $A^{'} = \mathsf{d} \in S_1, B^{'} = \mathsf{eb} \in S_2$ となり, また $W^{'} = \mathsf{bbcde}$ に対し, $A^{'} = \mathsf{bb} \in S_1, B^{'} = \mathsf{cde} \in S_2$ となるので, 集合 S_2 の要素は $S_2 = \{\mathsf{eb}, \mathsf{cde}\}$ となる. また, S_3 に関しては (I) が該当し, $A = \mathsf{cde} \in S_2$ に対し, $W = \mathsf{c} \in S_0, B = \mathsf{de} \in S_3$ となるので, 集合 S_3 は $S_3 = \{\mathsf{de}\}$ となる. 次に, S_4 に関しては (II) が該当し, $W^{'} = \mathsf{deb} \in S_0$ に対し, $A^{'} = \mathsf{de} \in S_3, B^{'} = \mathsf{b} \in S_4$ となるので, 集合 S_4 は $S_4 = \{\mathsf{b}\}$ となる. S_5 に関しては, (II) が該当し, $W^{'} = \mathsf{bad} \in S_0$ に対し, $A^{'} = \mathsf{b} \in S_4, B^{'} = \mathsf{ad} \in S_5$, または, $W^{'} = \mathsf{bbcde} \in S_0$ に対し, $A^{'} = \mathsf{b} \in S_4, B^{'} = \mathsf{bcde} \in S_5$ となるので, 集合 S_5 としては $S_5 = \{\mathsf{ad}, \mathsf{bcde}\}$ となる. S_6 に関しては (I) が該当し, $A = \mathsf{ad} \in S_5$ に対して, $W = \mathsf{a} \in S_0, B = \mathsf{d} \in S_6$ となるので, 集合 S_6 は $S_6 = \{\mathsf{d}\}$ となる. 最後に S_7 に関しては (II) が該当し, $W^{'} = \mathsf{deb}$ に対し, $A^{'} = \mathsf{d} \in S_0, B^{'} = \mathsf{eb} \in S_7$ となるので, 集合 S_7 は $S_7 = \{\mathsf{eb}\}$ となる.
以上を表にまとめると, 今考えている例では具体的に

[1] 前回見たクラフト不等式の証明では, 符号の長さを使ってはいるが, 具体的な符号, 及び, 語頭条件に関しては何ら言及しておらず, 証明の過程で全く使われていないことに注意しよう.

集合	要素
S_0	a,c,ad,abb,bad,deb,bbcde
S_1	d,bb
S_2	eb,cde
S_3	de
S_4	b
S_5	ad,bcde
S_6	d
S_7	eb

となる. このとき, 集合 $S_1, \cdots, S_n$ が S_0 の要素を含まなければ語頭条件が満たされていることになり, その符号は一意復号可能であることになる. 今の場合には $S_1, \cdots, S_7$ の中に S_0 の要素は ad が含まれているので語頭条件は満たされていない. 従って, 問題の符号は一意復号不可能である. しかし, クラフトの不等式を満たしているので, 問題に与えれた符号長で各記号を符号化すること (そのような符号を見つけること) 自体は可能である.
これを定理にしておこう.

定理
ある符号が一意復号可能であるのは上記の作り方で得られる集合 $S_1, \cdots, S_n$ が S_0 の要素を含まないときで, かつ, このときに限られる.

この定理の証明は比較的長いので, ここでは省略するが, 各自, 自分で証明を試みられるか, あるいは次に挙げる文献 :

Information Theory : By R. B. Ash, Dover (1990) Chapter 2 : Noiseless Coding, pp. 29-33

を読んでみることを薦める (大学図書館等にない場合には申し出て頂ければその部分をコピーさせて差し上げます).

また, この例以外に $\phi(\mathsf{aa}) = 00, \phi(\mathsf{ab}) = 10, \phi(\mathsf{ba}) = 01, \phi(\mathsf{bb}) = 011$ で与えられる符号 ϕ を考えると, 01 の部分で語頭条件を満たしていないが, 復号の際に符号を逆から読むことと約束すれば語頭条件には引っかからない. **受け取った符号を順に読んでいった場合**に語頭条件に引っかからずに復号できることを**瞬時復号可能である**という. この符号 ϕ のように, ある長さの符号を受け取った後にそれを逆から読んで復号化すれば元の記号列を復元できる場合は, 「瞬時」には復号できない[2] ことになるが, 瞬時復号は不可能であっても, つまり, 語等条件を満たしていなくても, 一意復号可能な符号も存在し, ここで取り上げた ϕ はその一例である.

5　ハフマン符号

ここでは, 最適な平均符号長をもつ符号として知られているハフマン符号の構成法とその性質について学ぶ.

[2] 「瞬時」というからには復号に時間的な意味合いが入ってくることになるが, ここでの例のように「逆向き」に復号化していく際には, どうしてもある必要な長さの符号を受け取るまで待たなければならず, この部分でのタイム・ラグが無視できなくなる. よって「瞬時」ではないのである.

5.1　ハフマン符号の構成法

ハフマン符号の構成法を簡単な例を取り上げて見ていく．まず, 情報源のアルファベットとして $\mathcal{A} = \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}, \mathrm{E}, \mathrm{F}\}$ の 6 つをとり, この各々を $\mathcal{B} = \{0, 1\}$, すなわち, 0, 1 を並べたもので符号化する．ただし, ここでは何らかの測定により, 情報源の中に $\mathrm{A}, \cdots, \mathrm{F}$ の各々の記号が現れる確率がわかっており, 記号 A の出現確率を p_{A} で表記することに約束すれば, $\{p_{\mathrm{A}}, p_{\mathrm{B}}, p_{\mathrm{C}}, p_{\mathrm{D}}, p_{\mathrm{E}}, p_{\mathrm{F}}\} = \{0.4, 0.3, 0.11, 0.09, 0.08, 0.02\}$ であることがわかっているものとしよう．

すると, 具体的なハフマン符号の構成法は次のようになる．

(1) まずは図 17 (左) のように, 記号 $\{\mathrm{A}, \cdots, \mathrm{F}\}$ をその出現確率が高い順に並べ, その横にその出現確率を記入しておく (この状態を I としょう).

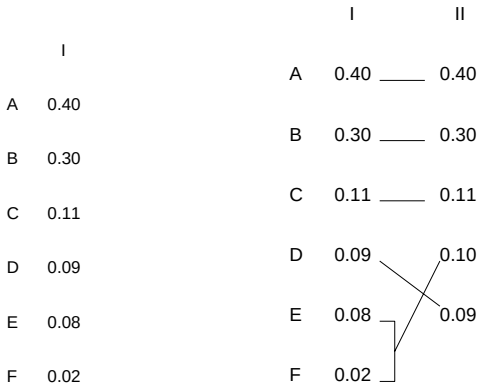

図 17: 状態 I (左) と状態 II (右).

(2) 確率が最も低い 2 つを選び (今の場合には E の 0.08 と F の 0.02 が該当する), その 2 つの確率の和を算出し (今の場合には $0.08 + 0.02 = 0.1$ となる), その値と残りの確率を図 17 (右) のようにその値が大きい順に並べる (この状態を II としょう).

(3) 状態 II において, 最も小さな 2 つの確率 (今の場合で言えば, 0.09 と 0.1 が該当する) を足し (この場合には $0.09 + 0.1 = 0.19$ となる), これと残りの 3 つの確率を図 18 (左) のようにその値が大きい順に並べる (この状態を III としよう).

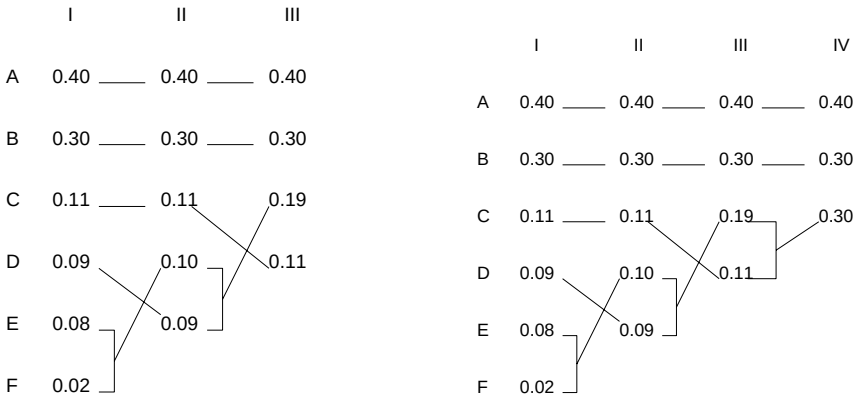

図 18: 状態 III (左) と状態 IV (右).

(4) 状態 III において, 最も小さな 2 つの確率 (今の場合で言えば, 0.19 と 0.11 が該当する) を足し ($0.19 + 0.11 = 0.3$ となる), これと残りの 2 つを図 18 (右) のように, その値が大きな順に並べる (この状態を IV としょう).

(5) 状態 IV において最も確率の小さな 2 つ (この場合には 0.3 と 0.3 が該当する) を選び, その両者を足し ($0.3 + 0.3 = 0.6$ となる), これと残りの確率をその値が大きな順に並べる (図 19 (左) 参照. この状態を V とする).

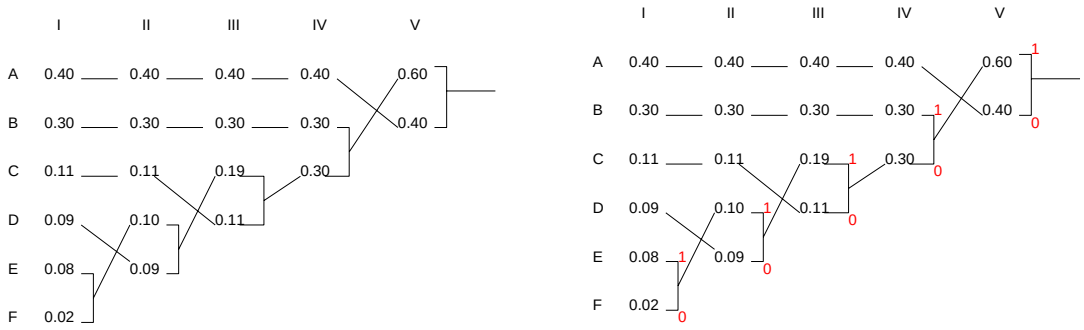

図 19: 状態 V (左) と最終状態 (右).

(6) 上記の (1)-(6) のステップで作られた大小関係の評価に対し, 図 19 (右) のように大きな方 (図 19 (右) の各枝で言えば上に付いた枝) に 1 を小さな方に 0 を割り当てる.

(7) 一番上位にの枝から, 状態 I へさかのぼって該当する記号に到達するまでに並べられた 0, 1 の列がその記号に割り当てられる符号になる. 例えば, A の場合には 0 であり, B の場合には 11 であり, C の場合には 100 であり, D の場合には 1010 であり,E の場合には 10111 であり, 最後に F は 10110 となる. これをまとめたものを表にしておこう.

記号	出現確率	ハフマン符号
A	0.40	0
B	0.30	11
C	0.11	100
D	0.09	1010
E	0.08	10111
F	0.02	10110

そこで, この符号の平均符号長 L を計算してみると

$$L = 0.4 \times 1 + 0.3 \times 2 + 0.11 \times 3 + 0.09 \times 4 + 0.08 \times 5 + 0.02 \times 5 = 2.19 \tag{195}$$

となる.

5.2 ハフマン符号の最適性

ここでは, 上で具体的に構成されたハフマン符号の最適性について詳しく見ていく.

捕題 3.1
与えられた情報源に対し, 以下の 2 つの条件を満たし, かつ, 平均符号長が最短の語頭符号が存在する.

(1) 確率が最も小さい 2 つの符号は同じ節点から出ている 2 つの葉に割り当てられる.
(2) その節点のレベルは, 木の中で最高である.

(証明):
まずは準備として, 次のように約束する.
情報源アルファベットを $\{a_1, a_2, \cdots, a_M\}$ とし, そのそれぞれの出現確率を $\{p_1, p_2, \cdots, p_M\}$ であるとする. このとき

(1) 最高レベルの節点には 1 つの葉しかない, と仮定すると, その葉に対応する符号語の長さを 1 つ減らした語頭符号を作ることができる (図 20 参照). しかも, 出来上がる語頭符号は平均符号長が T よりも短

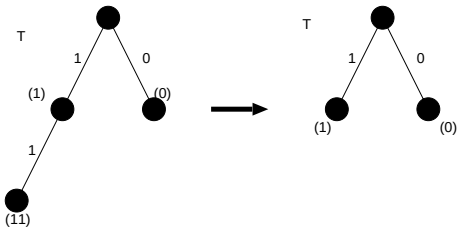

図 20: 最高レベルの節点にある 1 つの葉を減らした語等符号を作る.

い. すなわち, T の「平均符号長最短性」に反する. 従って, 最高レベルの節点には 2 つの葉が割り当てられる. (2 つの葉があれば, その 2 つの符号の長さを減らすことはできない. ⇒語頭条件を満たさなくなる).

(2) a_M に対する符号が最高レベルの節ではない, とする. このとき, $a_i\,(i \neq M, M-1)$ と a_M の符号を変換した新しい符号の木 $T^{'}$ を作ると, $L(T), L(T^{'})$ の差は

$$\begin{aligned} L(T) - L(T^{'}) &= p_i l_i + p_M l_M - \{p_i l_M + p_M l_i\} \\ &= (p_i - p_M)(l_i - l_M) \geq 0 \end{aligned} \tag{196}$$

となる. ただし, 最後の部分での不等号は $p_i \geq p_M, l_i - l_M > 0$ であることから来ていることに注意しよう.
従って, T の最適性より, $L(T) \leq L(T^{'})$ であるが, 上の結果と合わせると, $L(T) = L(T^{'})$, すなわち, 上記の変形により, 平均符号長は変化せず, $T^{'}$ も最適な符号の木となる.

定理 3.1
ハフマン符号は平均符号長が最も短い符号である.

(証明)
2 つの記号を 1 つにして得られる新しい情報源 (記号数は 3 以上とする) を**縮退した情報源**とする. このとき, T：平均符号長最短な木 (情報源：$\mathcal{X}$) とすれば

$$L(T) = p_1 l_1 + p_2 l_2 + \cdots + p_{M-1} l_{M-1} + p_M l_{M-1} \tag{197}$$

である. 一方, a_M, a_{M-1} を 1 つにまとめた $a_{M'-1}$ に対して (情報源：$\mathcal{X}'$), T も縮退させて T' を作る (図 21 参照). このとき, $L(T')$ は

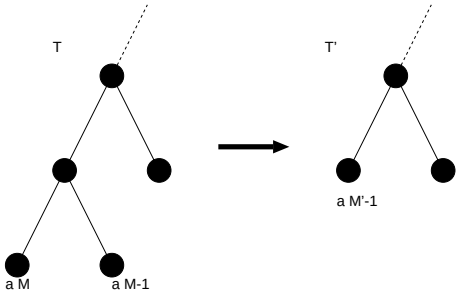

図 21: a_M, a_{M-1} を 1 つにまとめた $a_{M'-1}$ に対して (情報源：$\mathcal{X}'$), T も縮退させて T' を作る.

$$L(T') = p_1 l_1 + p_2 l_2 + \cdots + p_{M-2} l_{M-2} + (p_M + p_{M-1}) l_{M-2} \tag{198}$$

であるから, $L(T)$ と $L(T')$ の差を計算すれば

$$\begin{aligned} L(T) - L(T') &= p_{M-1} l_{M-1} + p_M l_{M-1} - (p_M + p_{M-1}) l_{M-2} \\ &= p_M (l_{M-1} - l_{M-2}) + p_{M-1} (l_{M-1} - l_{M-2}) = p_M + p_{M-1} \end{aligned} \tag{199}$$

となる.
一方, 逆に $\mathcal{X}'$ に対する符号の木を S' とし, S' を展開して木 S を作る (情報源：$\mathcal{X}$). このとき

$$L(S') = p_1 l_1 + p_2 l_2 + \cdots + p_{M-2} l_{M-2} + p_{M'-1} l_{M-2} \tag{200}$$

$$L(S) = p_1 l_1 + p_2 l_2 + \cdots + p_{M-2} l_{M-2} + p_{M-1} l_{M-1} + p_M l_{M-1} \tag{201}$$

となるから, 両者の差をとると

$$\begin{aligned} L(S) - L(S') &= p_{M-1} l_{M-1} + p_M l_{M-1} - p_{M'-1} l_{M-2} \\ &= p_{M-1} l_{M-1} + p_M l_{M-1} - p_{M-1} l_{M-2} - p_M l_{M-2} \\ &= p_{M-1} + p_M \end{aligned} \tag{202}$$

となるが, 仮定である $L(T) \leq L(S)$, $L(S') \leq L(T')$ より, (199)-(202) を作ると

$$0 \leq L(T') - L(S') = L(T) - L(S) \leq 0 \tag{203}$$

つまり

$$L(T) = L(S) \tag{204}$$

となる. 従って, 記号数が $n-1$ のとき, 定理が正しいと仮定し, 記号数 n の情報源に対し, その縮退した情報源を考えると, 仮定より, この縮退した情報源に対してはハフマン符号の木が最適である.

次に, 縮退した記号を元に戻し, その符号の木から元の情報源に対する符号の木を構成すると, その符号は最適なものとなっており (n のときの最適性), その構成法はハフマン符号の作り方そのものである.

ハフマン符号の作り方に慣れるために, 次の**例題 9** を見ておこう.

例題 9
次の表 2 に与えた情報源アルファベット, 及び, その生成確率に対してハフマン符号を構成せよ. また, 得られるハフマン符号の平均符号長を求めよ.

情報源アルファベット	生成確率
x_1	0.2
x_2	0.18
x_3	0.10
x_4	0.10
x_5	0.10
x_6	0.061
x_7	0.059
x_8	0.04
x_9	0.04
x_{10}	0.04
x_{11}	0.04
x_{12}	0.03
x_{13}	0.01

表 2： ハフマン符号を考えるアルファベットとその生成確率の表.

(解答例)
ハフマン符号を作成するアルゴリズムに従って符号の木を作ると図のようになる. 従って, 求める符号語, 生成確率, 符号長等を表にまとめれば次のようになる.

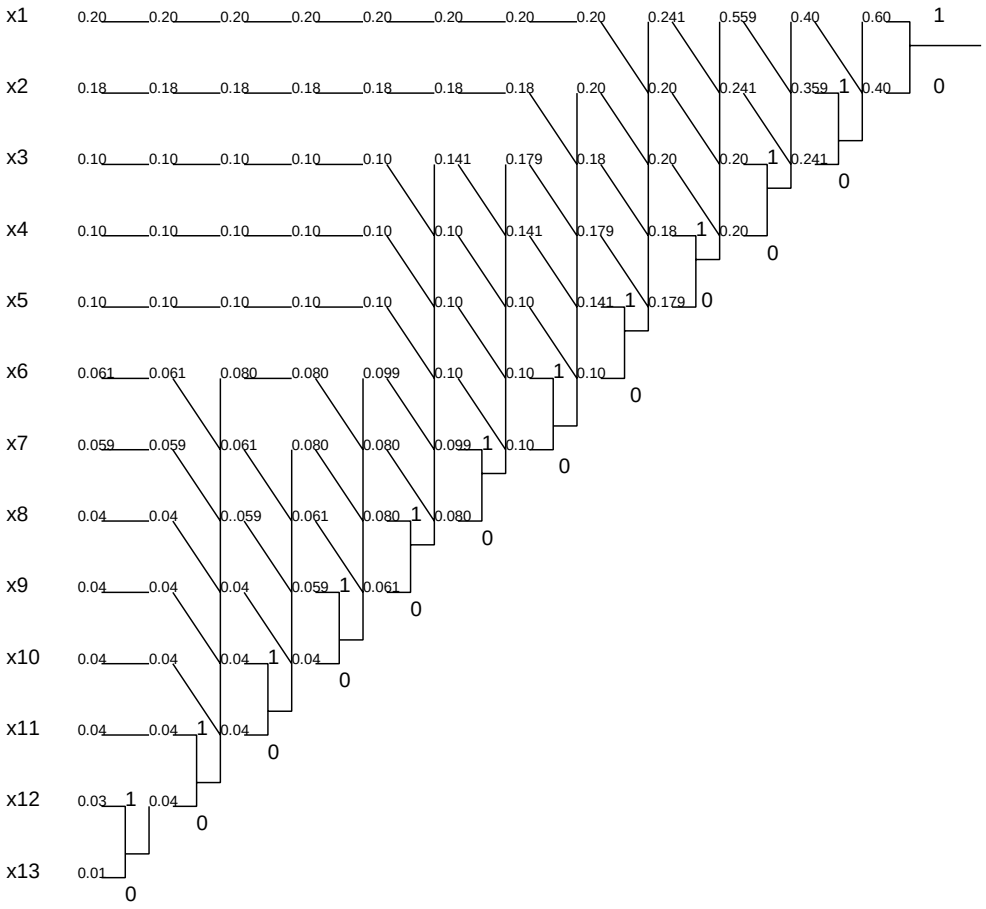

図 22: 情報源アルファベット $x_1 \sim x_{13}$ に対するハフマン符号の符号の木.

アルファベット	符号語	符号長 (l_i)	生成確率 (p_i)	$p_i \times l_i$
x_1	01	2	0.20	0.40
x_2	111	3	0.18	0.54
x_3	100	3	0.10	0.30
x_4	001	3	0.10	0.30
x_5	000	3	0.10	0.30
x_6	1010	4	0.061	0.244
x_7	11011	5	0.059	0.295
x_8	11010	5	0.04	0.20
x_9	10111	5	0.04	0.20
x_{10}	10110	5	0.04	0.20
x_{11}	11001	5	0.004	0.20
x_{12}	110001	6	0.03	0.18
x_{13}	110000	6	0.01	0.06

平均符号長はこの表から直ちに

$$L = \sum_{i=1}^{13} p_i\, l_i = \underline{3.419} \tag{205}$$

と求まる.

5.3 ハフマン符号の問題点

情報源アルファベットの生成確率についての情報が事前に必要である.

※ユニバーサル符号に関しては, ここでは省略し, 時間に余裕ができれば改めてここに戻って説明します.

5.4 雑音が無い状況下での情報源符号化 (圧縮) のまとめ

2 回にわたり情報源符号化について見てきた. 結局, 我々がここでわかったことと言えば

出現確率が小さい情報源アルファベットには長い符号語を, 逆に,
出現確率が大きいアルファベットには短い符号語を割り当てればよい

という, 割と当たり前の事実であった. ただし, 各符号語の長さに関してはもう少し定量的なことまでが言えて, 「出現確率が p_i であるアルファベットの符号語の長さは $-\log_K p_i$ 以上の最小な整数に選べば良い」ことまでがわかった. この指針により, 我々は情報源の冗長性を除去し, 無駄なく情報を伝達することができる. ただし, 以上は伝送時に雑音 (ノイズ) が無い場合に限った話あり, ノイズがある際には逆に情報に冗長性を持たせて多少の間違いがあってもそれを発見し, 修正できるような方策を立てる必要がでてくるのだが, 次回からはそれについて詳しくみて行くことにしょう.

演習問題 6

記号 $x_1, x_2, \cdots, x_6$ の出現確率が $P(x_1) = 0.3, P(x_2) = 0.25, P(x_3) = 0.2, P(x_4) = P(x_5) = 0.1, P(x_6) = 0.05$ で与えられるとき

(1) この情報源のエントロピーを求めよ.
(2) ハフマン符号を構成し, その平均符号長を求めよ.