コンパクト符号

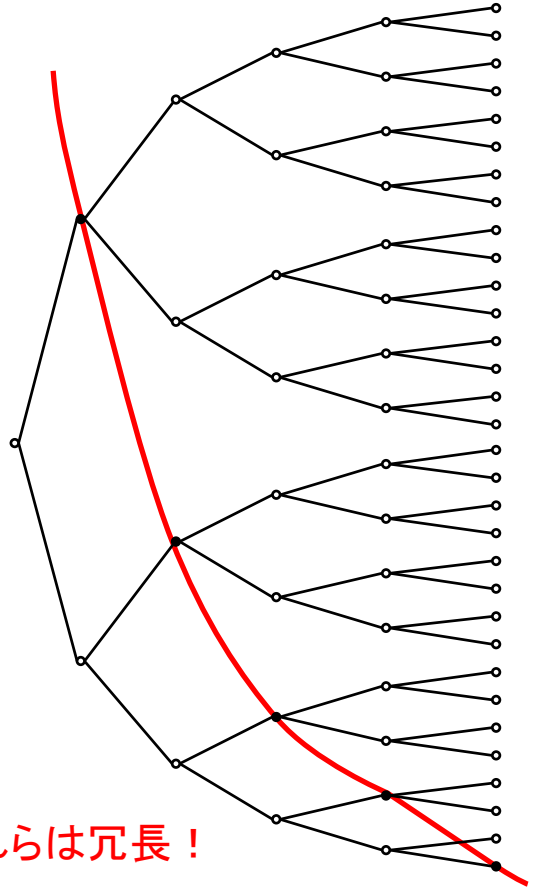
コンパクト符号

- 与えられた情報源Sから発生する情報源記号に一つずつ符号語を割り当てる符号化によってできる一意復号可能な符号のうち，平均符号長が最小となる符号をコンパクト符号という．
- マクミラン不等式とクラフト不等式が同形だから，任意の一意復号可能なコンパクト符号に対して，それと同じ符号語長バッグを持つ瞬時符号が存在する．
- 概念的には，与えられた情報源Sに対するコンパクト符号を見つけるためには，すべての瞬時符号を枚挙し，そのなかの平均符号長最小のものを選べばよい．
- 情報源Sに対するコンパクト符号が複数存在することもある．

コンパクト符号

例： 情報源記号{A, B, C, D, E}に対する2元コンパクト符号？

瞬時符号P

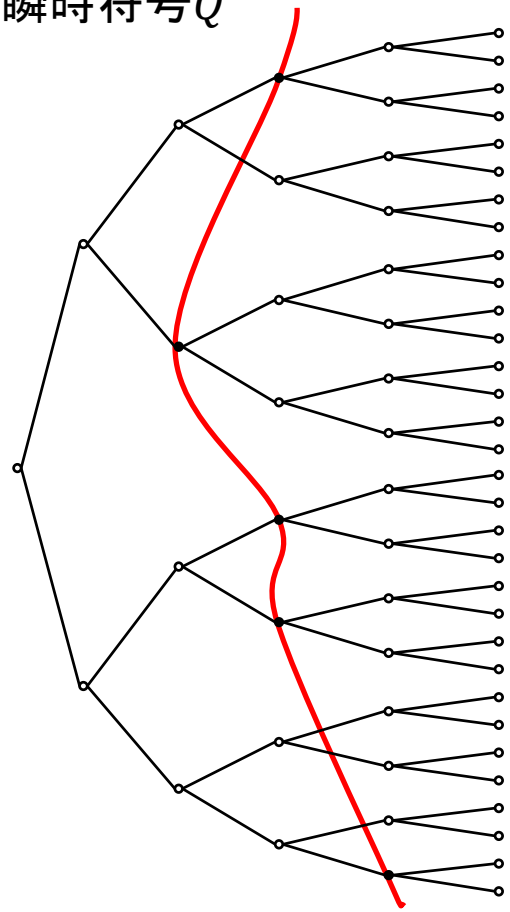


これらは冗長！

符号長バッグ：{1, 2, 3, 4, 5}

→ {1, 2, 3, 4, 4}

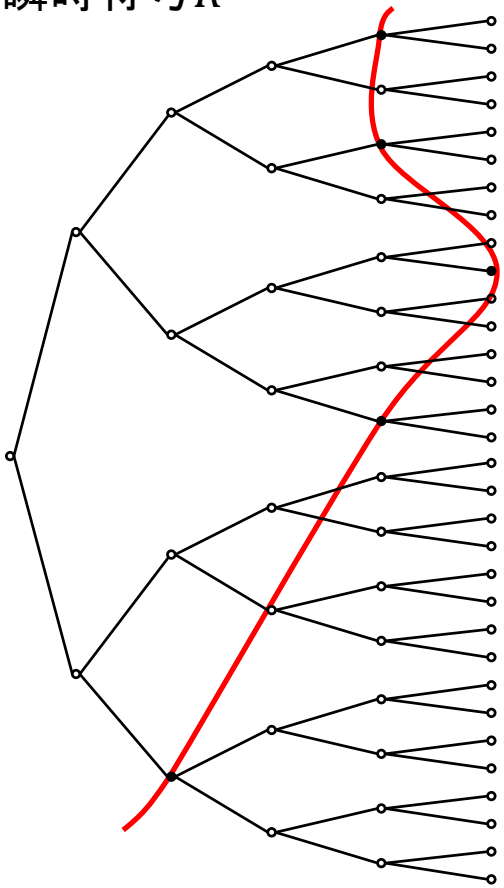
瞬時符号Q



符号長バッグ：{3, 2, 3, 3, 4}

→ {2, 2, 3, 3, 2}

瞬時符号R



符号長バッグ：{4, 4, 5, 4, 2}

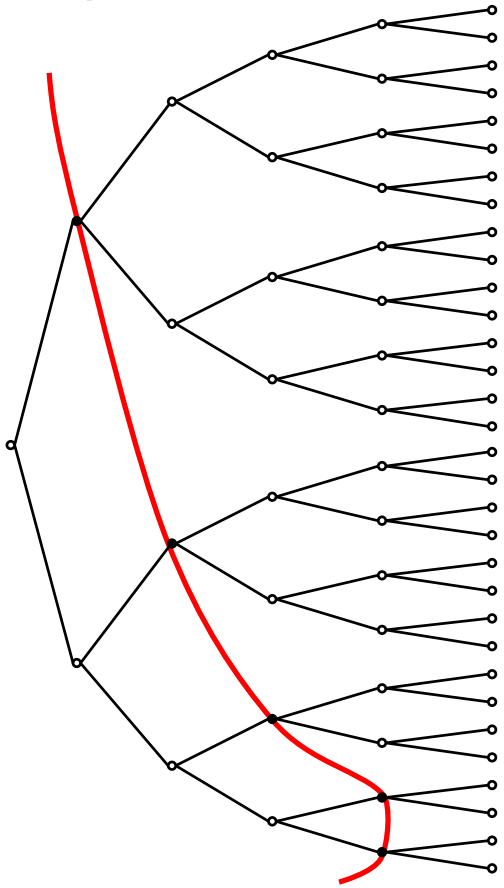
→ {3, 3, 3, 3, 1}

…

コンパクト符号

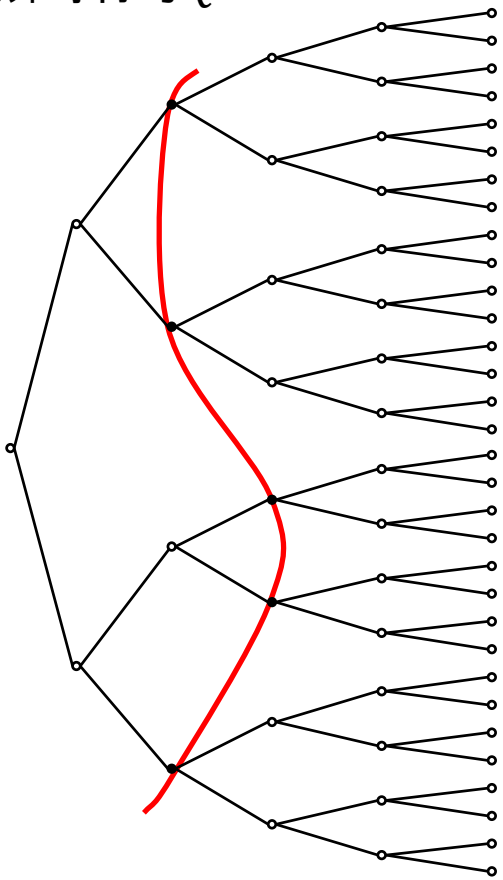
例： 情報源記号{A, B, C, D, E}に対する2元コンパクト符号？

瞬時符号P



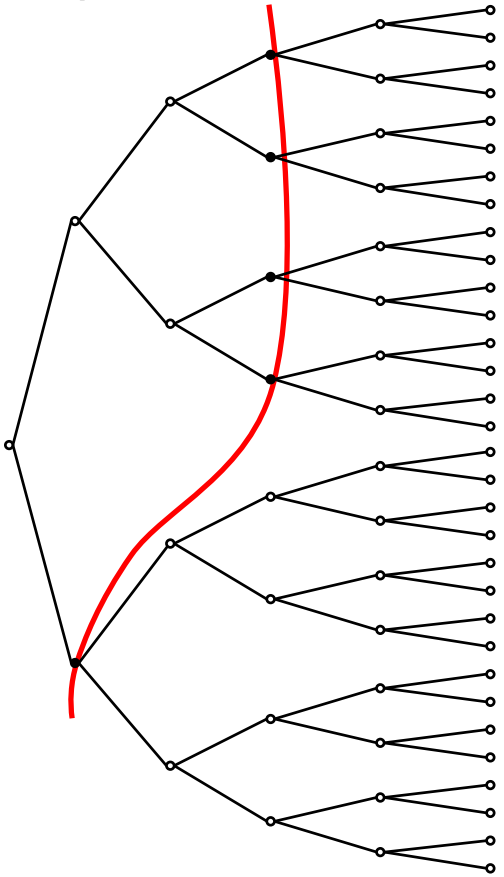
{1, 2, 3, 4, 4}

瞬時符号Q



{2, 2, 3, 3, 2}

瞬時符号R



{3, 3, 3, 3, 1}

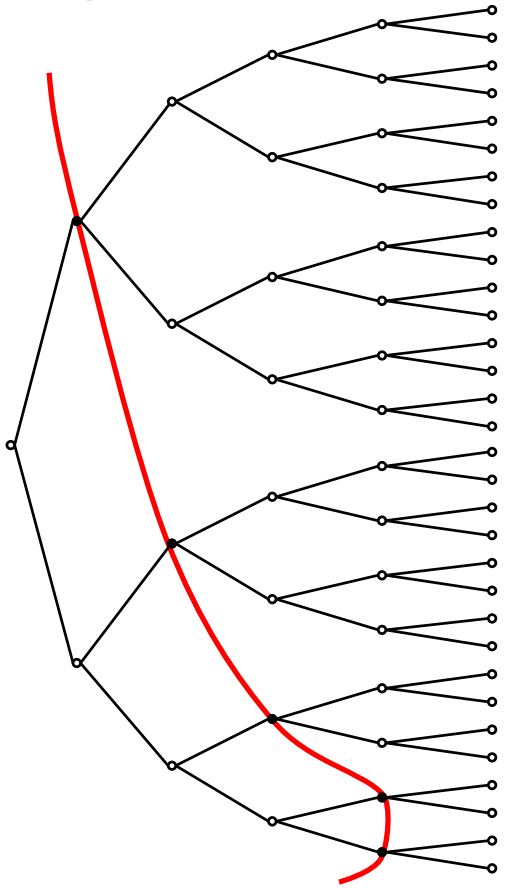
…

どれがいいか？

コンパクト符号

例： 情報源記号発生確率が⟨A: 0.2, B: 0.2, C: 0.2, D: 0.2, E: 0.2⟩の場合

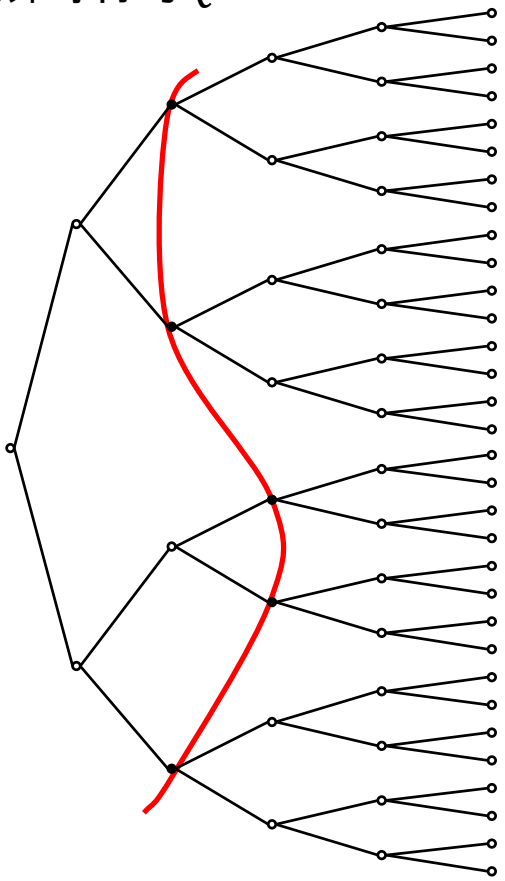
瞬時符号P



符号長バッグ：$\{1, 2, 3, 4, 4\}$

→ 平均符号長2.8

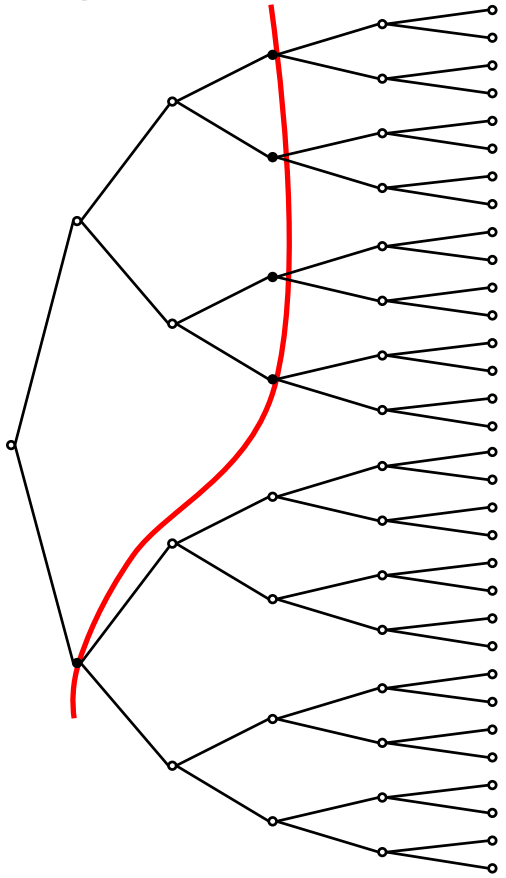
瞬時符号Q



符号長バッグ：$\{2, 2, 3, 3, 2\}$

→ 平均符号長2.4

瞬時符号R



符号長バッグ：$\{3, 3, 3, 3, 1\}$

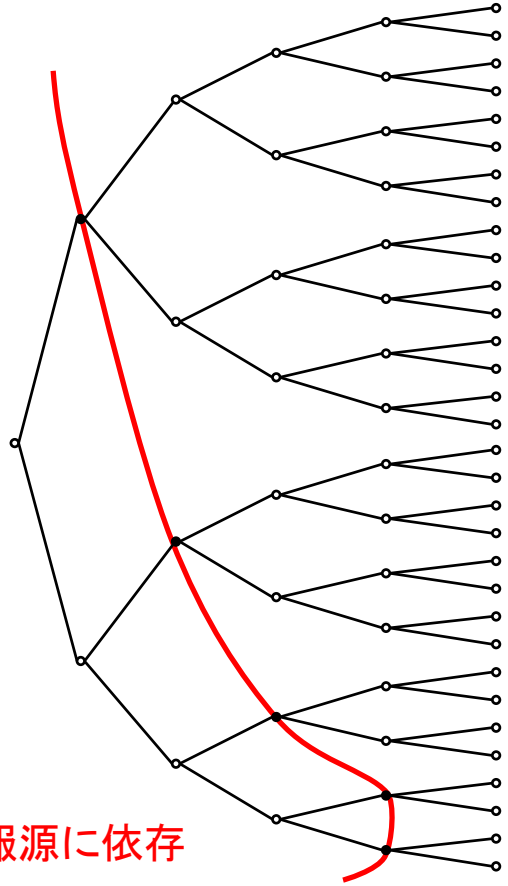
→ 平均符号長2.6

…

コンパクト符号

例： 情報源記号発生確率が⟨A: 0.6, B: 0.2, C: 0.1, D: 0.07, E: 0.03⟩の場合

瞬時符号P

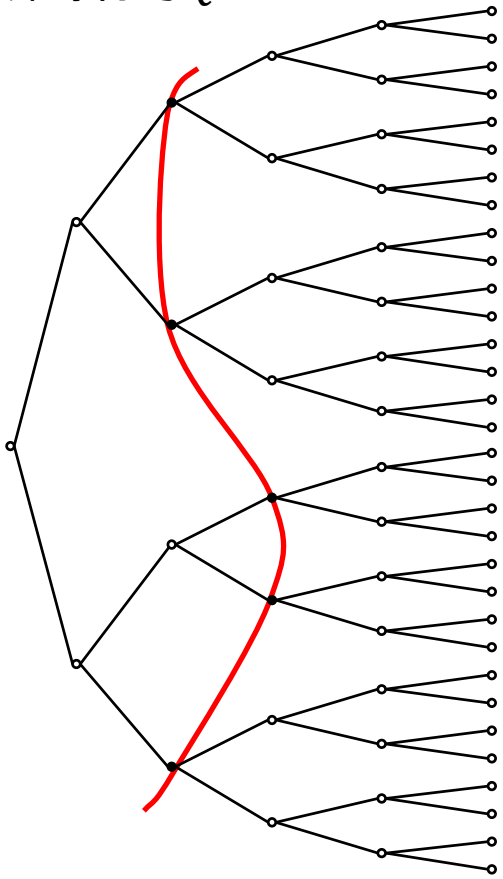


情報源に依存

符号長バッグ：$\{1, 2, 3, 4, 4\}$

→ 平均符号長1.7

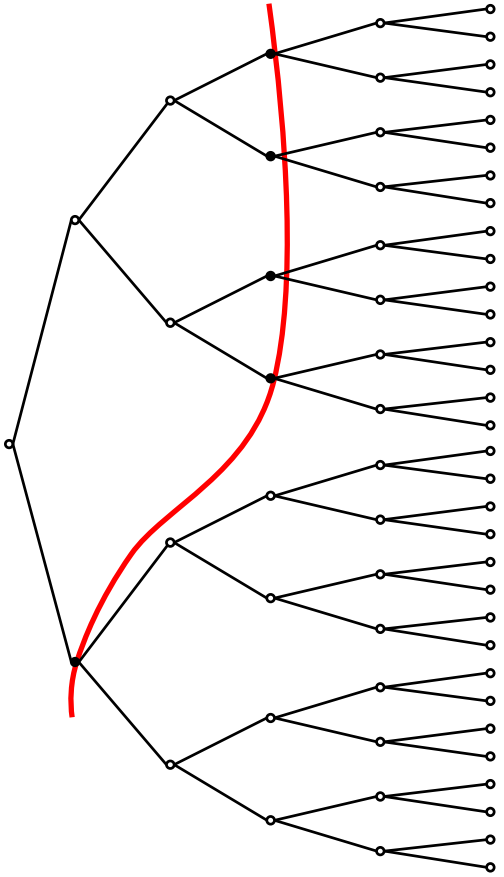
瞬時符号Q



符号長バッグ：$\{2, 2, 3, 3, 2\}$

→ 平均符号長2.17

瞬時符号R



符号長バッグ：$\{3, 3, 3, 3, 1\}$

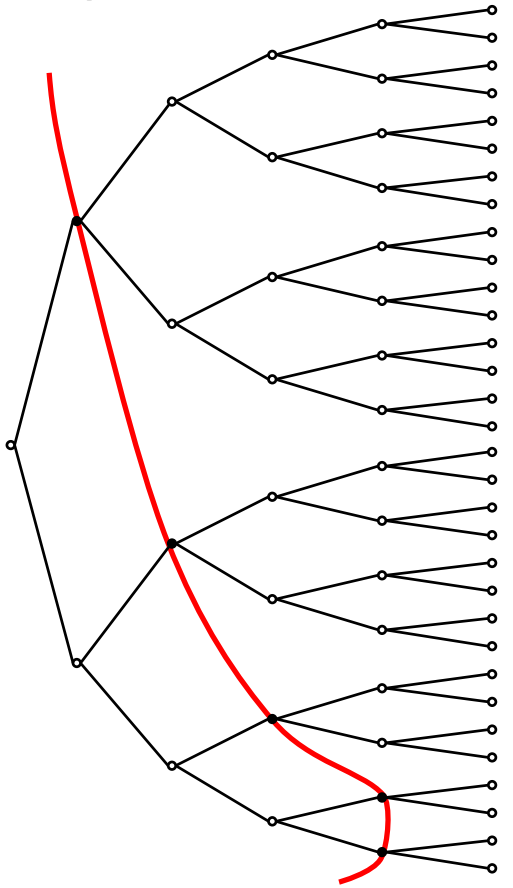
→ 平均符号長2.94

…

コンパクト符号

問題： 情報源記号発生確率⟨A:?, B:?, C:?, D:?, E:?⟩がどのような場合？

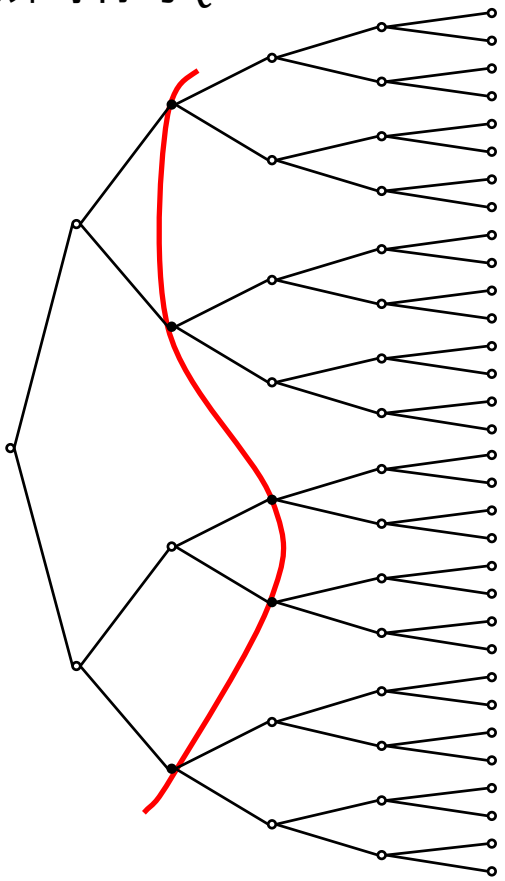
瞬時符号P



符号長バッグ：$\{1, 2, 3, 4, 4\}$

→ 平均符号長?

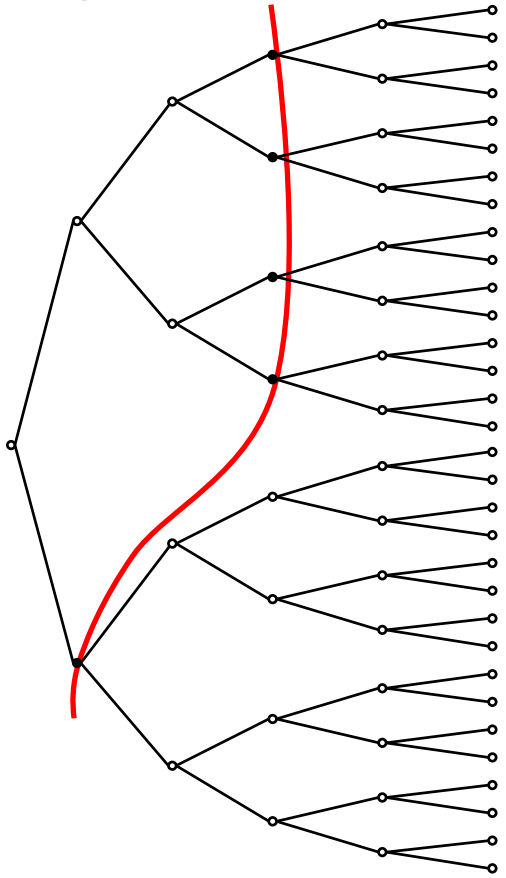
瞬時符号Q



符号長バッグ：$\{2, 2, 3, 3, 2\}$

→ 平均符号長?

瞬時符号R



符号長バッグ：$\{3, 3, 3, 3, 1\}$

→ 平均符号長?

…

コンパクト符号

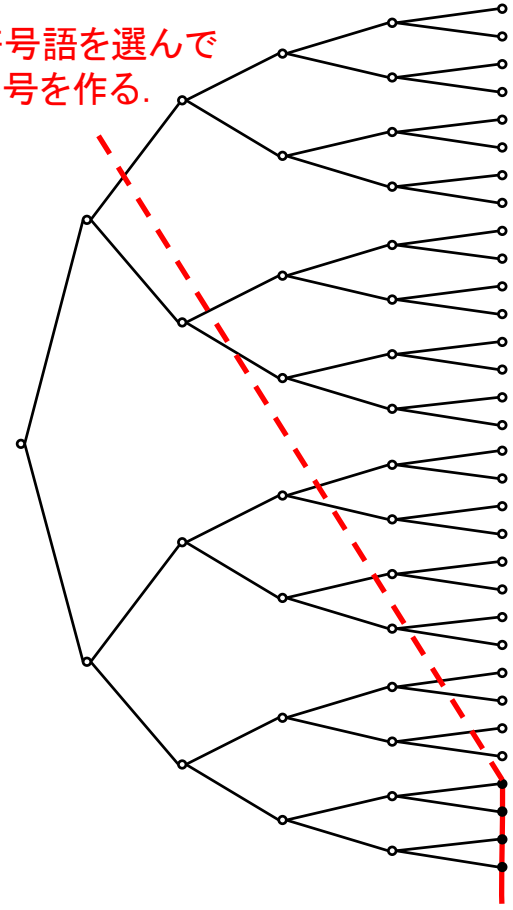
練習問題3-1

情報源記号$\{a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8\}$をもつ情報源$S$に対して，1情報源記号ごとに2元符号によるコンパクト符号化を行った結果，符号
$\{c(a_1), c(a_2), c(a_3), c(a_4), c(a_5), c(a_6), c(a_7), c(a_8)\}$
が得られ，

$$1 \leq |c(a_1)| \leq |c(a_2)| \leq |c(a_3)| \leq |c(a_4)| \leq |c(a_5)| = |c(a_6)| = |c(a_7)| = |c(a_8)| = 5$$

になったという．ただし，$c(a_i)$は情報源記号a_iに対する符号語，$|c(a_i)|$は符号語$c(a_i)$の長さを表す．

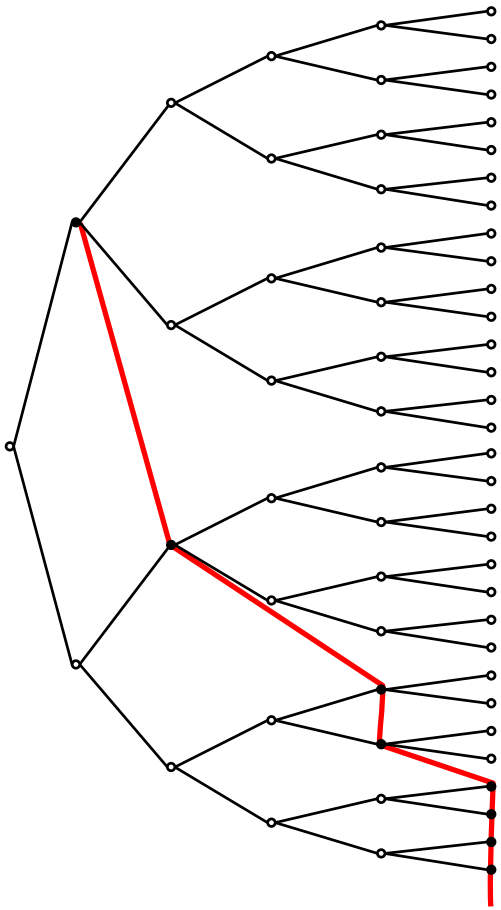
このとき，$|c(a_1)|, |c(a_2)|, |c(a_3)|, |c(a_4)|$の可能な組み合わせをすべて書き出せ．また，それぞれの組み合わせは，符号の木，および各符号に対応づけられる各情報源記号の生起確率がどのようになっている場合に生じるか，例示せよ．



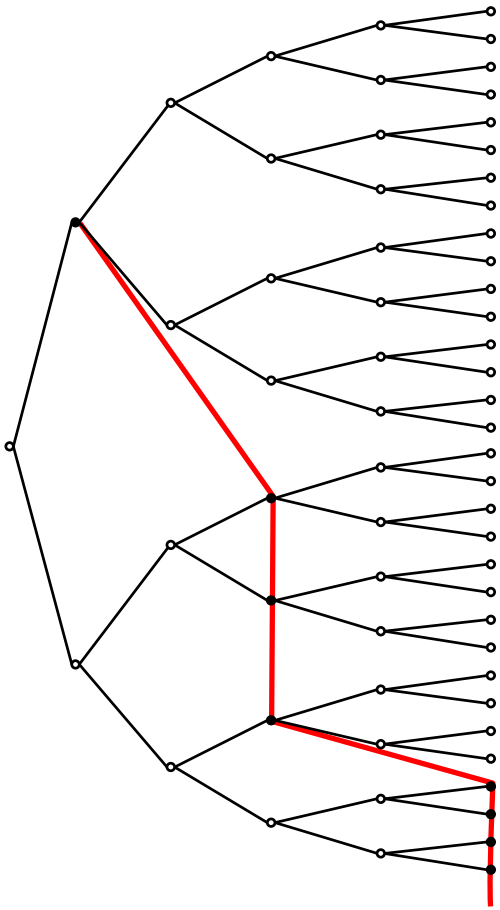
符号長バッグ: {?, ?, ?, ?, 5, 5, 5, 5}

コンパクト符号

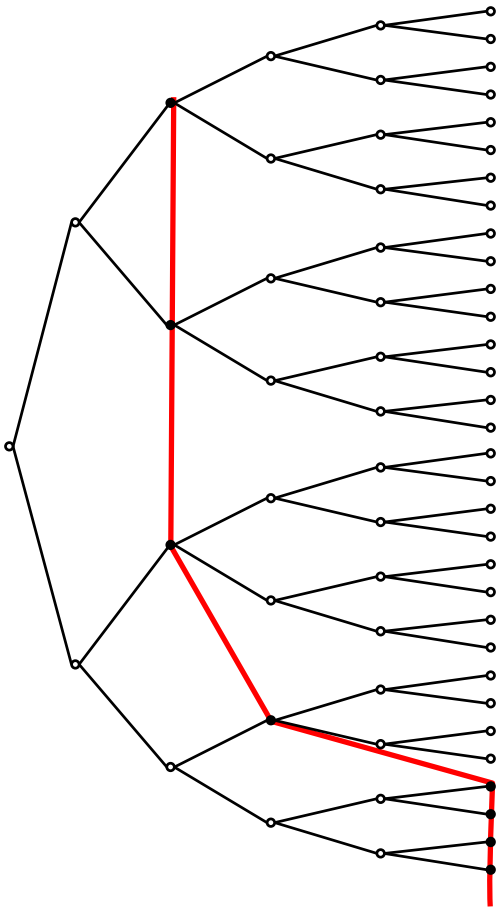
練習問題3-1



符号長バッグ：{1, 2, 4, 4, 5, 5, 5, 5}



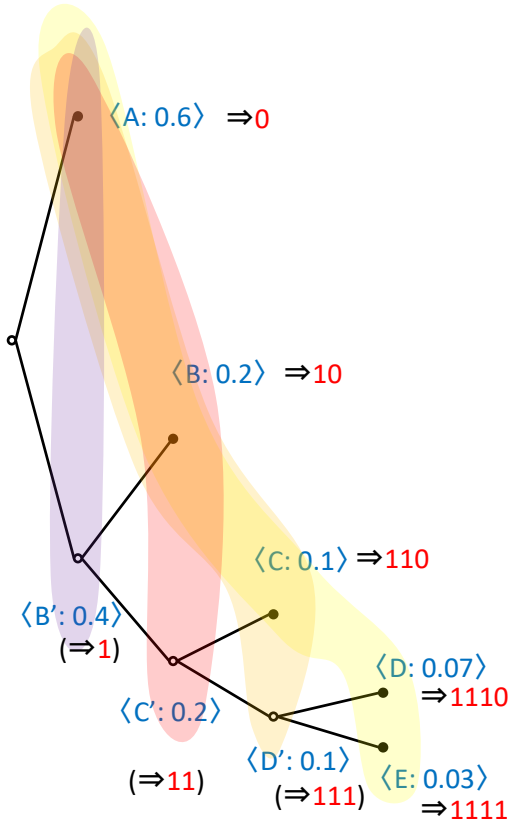
符号長バッグ：{1, 3, 3, 3, 5, 5, 5, 5}



符号長バッグ：{2, 2, 2, 3, 5, 5, 5, 5}

ハフマン符号化：コンパクト符号を構成する

例： 情報源記号発生確率が⟨A: 0.6, B: 0.2, C: 0.1, D: 0.07, E: 0.03⟩の場合



符号長バッグ：{1, 2, 3, 4, 4}

→ 平均符号長1.7

情報源S_5：⟨A: 0.6, B: 0.2, C: 0.1, D: 0.07, E: 0.03⟩のコンパクト符号構成法

① （すぐに答えが分からないので），情報源S_5の発生確率の最も小さな2つの情報源記号D: 0.07, E: 0.03に着目し，それらを一つにまとめてD′ ← D＋Eとした情報源

S_4：⟨A: 0.6, B: 0.2, C: 0.1, D′: 0.1⟩

に対するコンパクト符号構成を試みる．

② 前項と同じ：C′ ← C＋D′として，S_3：⟨A: 0.6, B: 0.2, C′: 0.2⟩に対してトライ

③ 前項と同じ：B′ ← B＋C′として，S_2：⟨A: 0.6, B′: 0.4⟩に対してトライ

④ S_2に対するコンパクト符号⟨A ← 0, B′ ← 1⟩

⑤ B′ ← B＋C′であったので，⟨A ← 0, B ← 10, C′← 11⟩

⑥ C′ ← C＋D′であったので，⟨A ← 0, B ← 10, C ← 110, D′ ← 111⟩

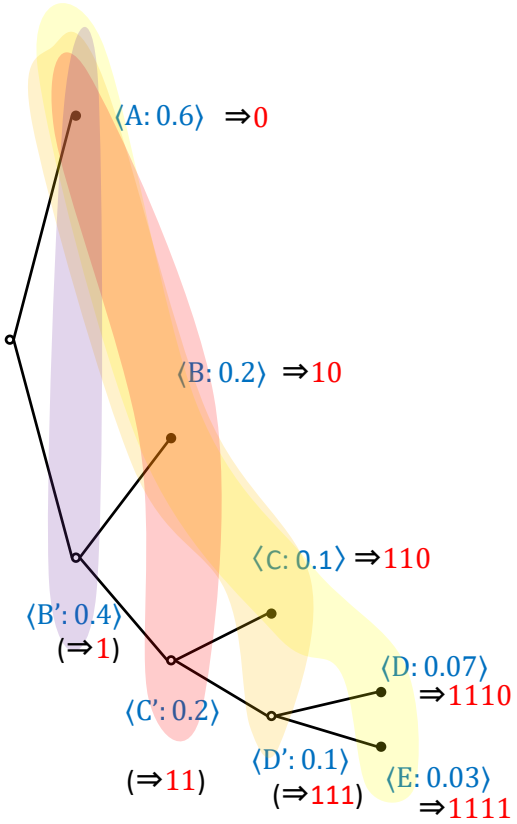
⑦ D′ ← D＋Eであったので，⟨A ← 0, B ← 10, C ← 110, D ← 1110, E ← 1111⟩

なぜこれでいいのか？

情報源S_nの発生確率の最も小さな2つの情報源記号をa_n, a_{n-1}とする．
$a'_{n-1} \leftarrow a_n + a_{n-1}$によってできる情報源$S_{n-1}$のコンパクト符号を$C_{n-1}$とする．
C_{n-1}のa'_{n-1}に対する符号語c'_{n-1}とする．C_{n-1}の$a'_{n-1} \leftarrow c'_{n-1}$の代わりに
$a_{n-1} \leftarrow c'_{n-1}0, a_n \leftarrow c'_{n-1}1$としてできる符号$C_n$はコンパクト符号である．

コンパクト符号

例： 情報源記号発生確率が⟨A:0.6, B:0.2, C:0.1, D:0.07, E:0.03⟩の場合



符号長バッグ：$\{1, 2, 3, 4, 4\}$

→ 平均符号長1.7

情報源S_nの発生確率の最も小さな2つの情報源記号をa_n, a_{n-1}とする．
$a'_{n-1} \leftarrow a_n + a_{n-1}$によってできる情報源$S_{n-1}$のコンパクト符号を$C_{n-1}$とする．
C_{n-1}のa'_{n-1}に対する符号語c'_{n-1}とする．C_{n-1}の$a'_{n-1} \leftarrow c'_{n-1}$の代わりに
$a_{n-1} \leftarrow c'_{n-1}0, a_n \leftarrow c'_{n-1}1$としてできる符号$C_n$はコンパクト符号である．

【性質1（S_5の場合）】
S_5に対する任意のコンパクト符号C_5において，発生確率の最も小さな2つの情報源記号D, Eには最も長い符号語が割り当てられる．

【証明】
仮にCに対応づける符号語の長さをs，Eに対応づける符号の長さをt，
$s > t$とすると，

平均符号長 $L = \cdots + s \times 0.1 + \cdots + t \times 0.03$

ここで，Cに対する符号語とEに対する符号語を交換すると

平均符号長 $L' = \cdots + t \times 0.1 + \cdots + s \times 0.03$

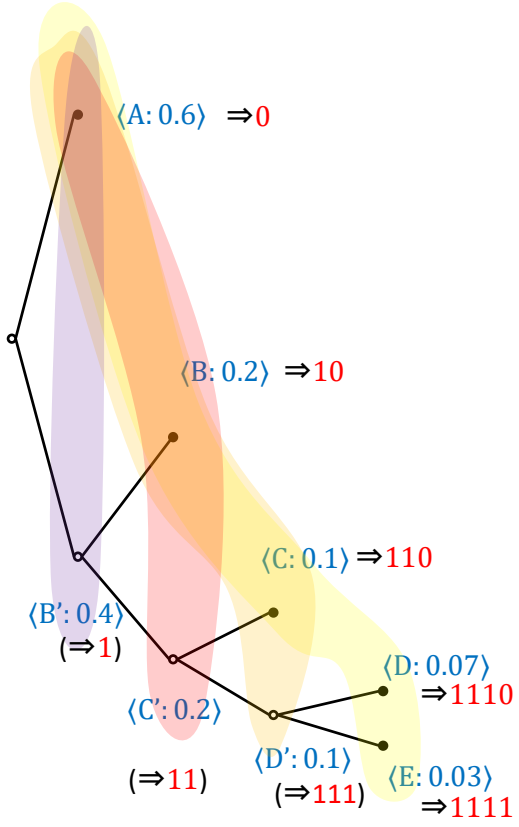
次のように$L' < L$となり，S_5がコンパクト符号であるという前提に矛盾する．

$$L' - L = (t - s) \times (0.1 - 0.03) < 0$$

【性質1'】
S_5に対して，D, Eに対する符号語が符号の木において兄弟ノードとなっているコンパクト符号C'_5が存在する．定義によりC_5との平均符号長は等しいが，C'_5はC_5と同じであるとは限らない．（証明は後述の通り）

コンパクト符号

例： 情報源記号発生確率が⟨A:0.6, B:0.2, C:0.1, D:0.07, E:0.03⟩の場合



符号長バッグ：{1, 2, 3, 4, 4}

→ 平均符号長1.7

情報源S_nの発生確率の最も小さな2つの情報源記号をa_n, a_{n-1}とする．
$a'_{n-1} \leftarrow a_n + a_{n-1}$によってできる情報源$S_{n-1}$のコンパクト符号を$C_{n-1}$とする．
C_{n-1}のa'_{n-1}に対する符号語c'_{n-1}とする．C_{n-1}の$a'_{n-1} \leftarrow c'_{n-1}$の代わりに
$a_{n-1} \leftarrow c'_{n-1}0, a_n \leftarrow c'_{n-1}1$としてできる符号$C_n$はコンパクト符号である．

【性質2（S_5の場合）】
情報源S_5の発生確率の最も小さな2つの情報源記号D:0.07, E:0.03に着目し，それらを一つにまとめてD′ ← D＋Eとした情報源

S_4: ⟨A:0.6, B:0.2, C:0.1, D′:0.1⟩

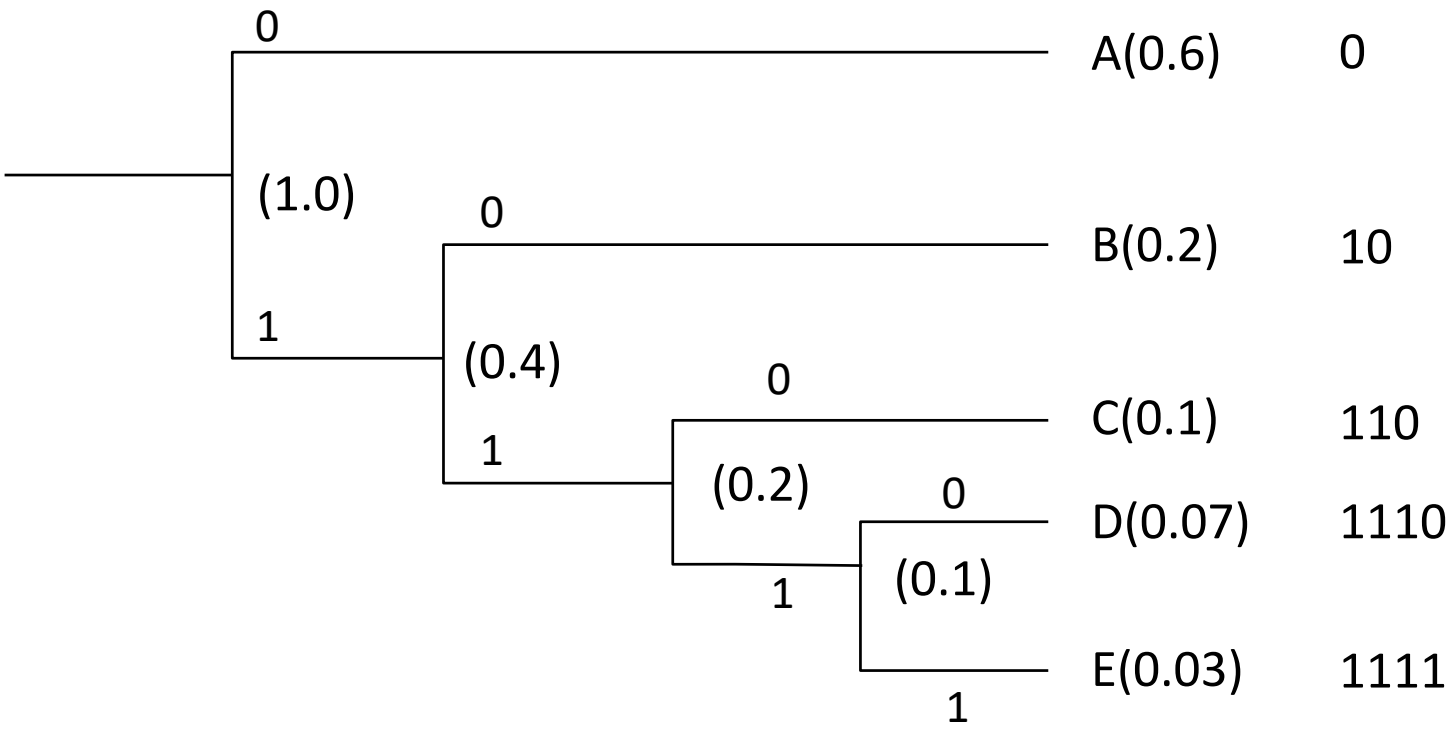
に対するコンパクト符号をC_4とする．C_4におけるD′に対する符号語c_4とする．C_4においてD′ ← c_4の代わりにD ← $c_4$0, D ← $c_4$1としてできる符号C_5はコンパクト符号である．

【証明骨子】
C_4はコンパクト符号であるが，C_5はコンパクト符号でないと仮定する．S_5に対するコンパクト符号C'_5が存在し，C'_5の平均符号長はC_5の平均符号長より短い．性質1′により，S_5に対して，D, Eに対する符号語が符号の木において兄弟ノードとなっているコンパクト符号C''_5が存在し，その平均符号長はC'_5の平均符号長に等しい．C''_5に対する符号木においてDとEを子とするノードをxとする．C''_5において，DとEに対する符号語割り当ての代わりにD′に符号語xを割り当てる符号C''は，S_4に対する瞬時符号であり，その平均符号長はC_4より短い．これは，C_4がコンパクト符号であることに矛盾．

ハフマン符号化（作業のみ）

- ハフマン符号化はコンパクト符号の一つの構成法である.

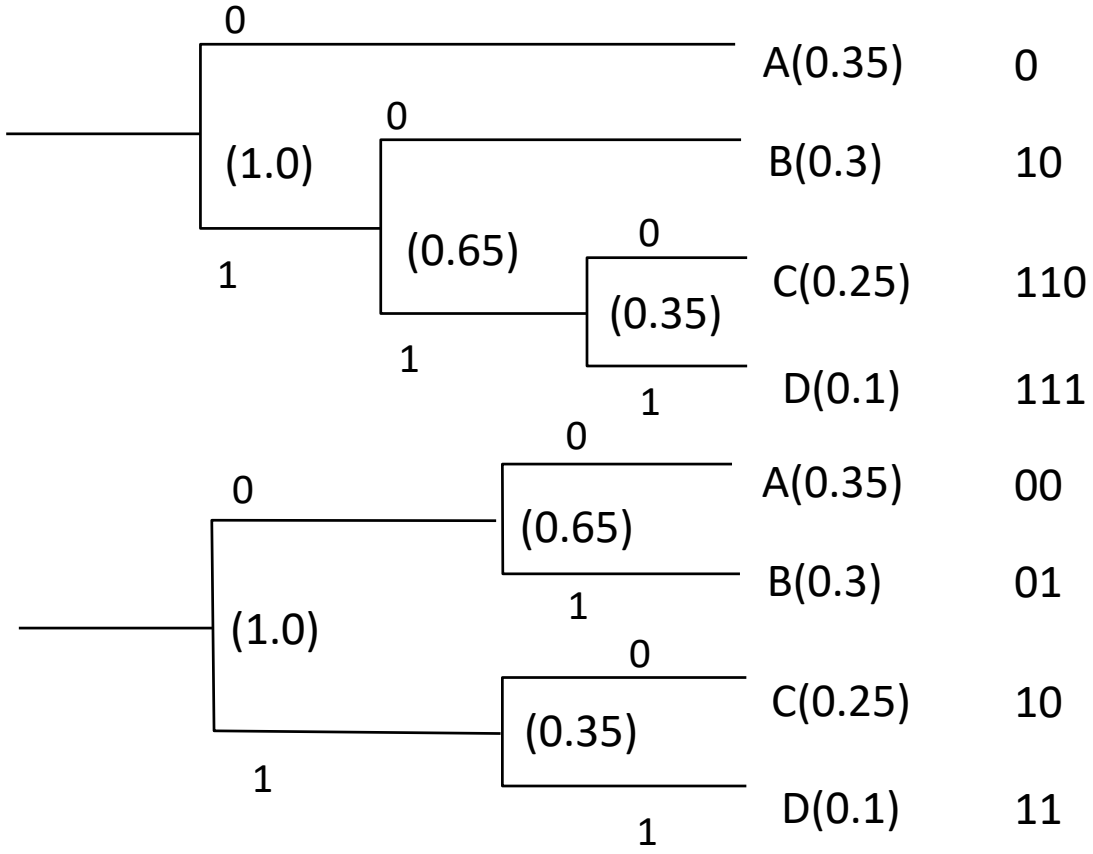
例：情報源記号発生確率を⟨A:0.6, B:0.2, C:0.1, D:0.07, E:0.03⟩とすれば，
ハフマン符号化は次のように行われる.



ハフマン符号

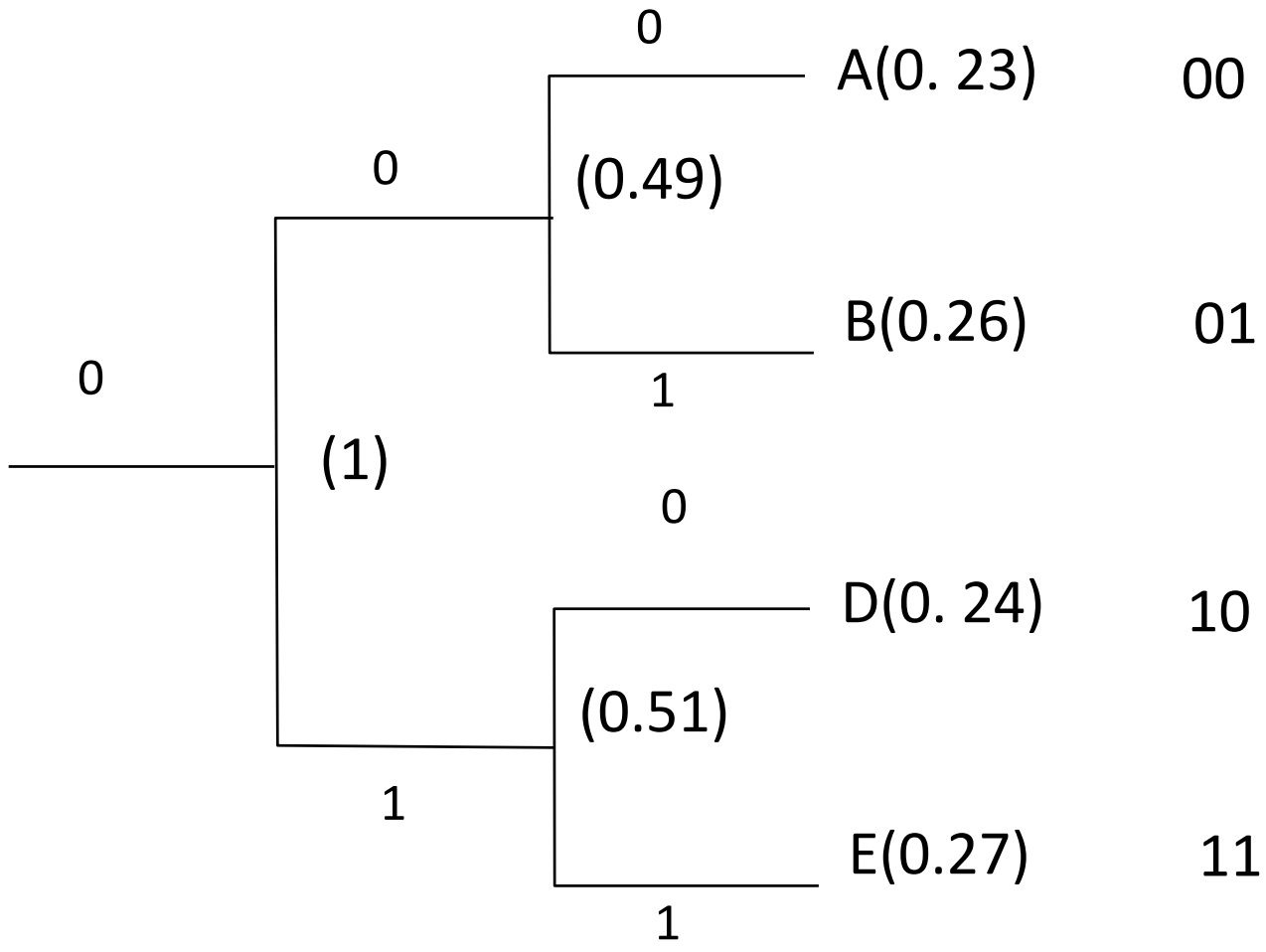
複数のハフマン符号化が可能なケース

情報源記号発生確率が⟨A:0.35, B:0.3, C:0.25, D:0.1⟩となっているとき



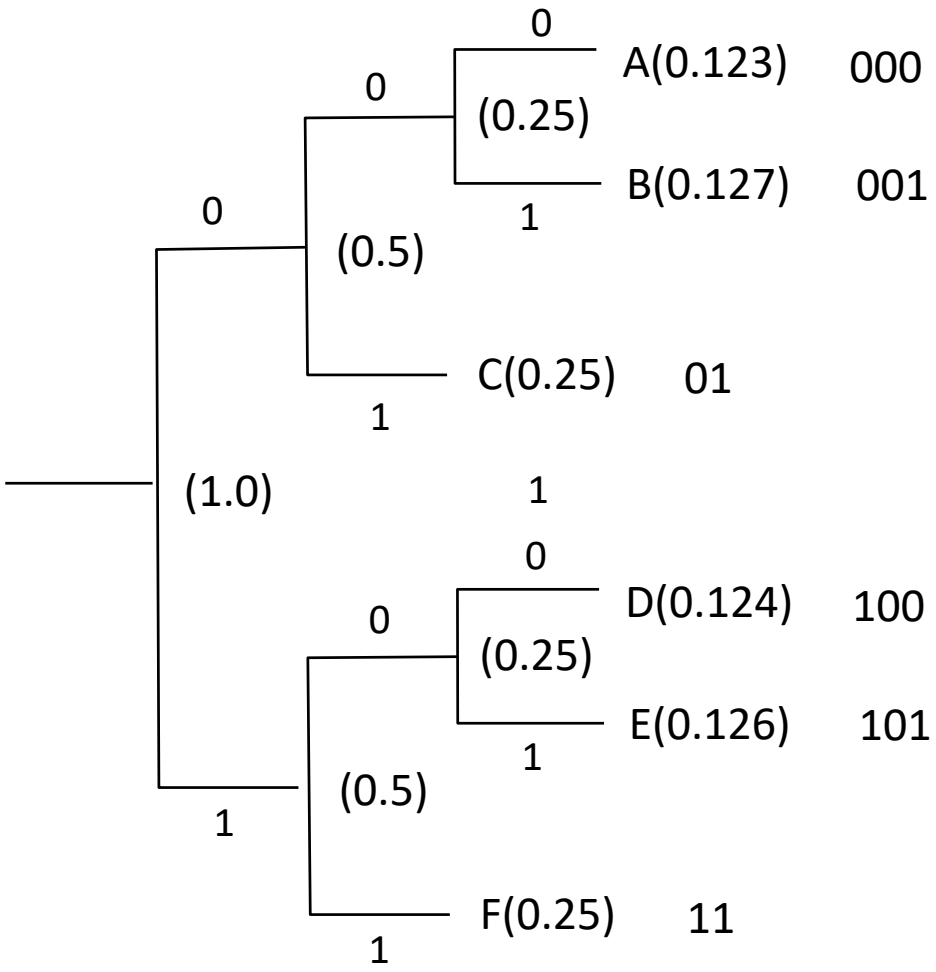
ハフマン符号

ハフマン符号化によって導き出されないコンパクト符号　その1



ハフマン符号

ハフマン符号化によって導き出されないコンパクト符号　その2



コンパクト符号

【補助定理1】（瞬時符号Cがコンパクトであるための必要条件）

2個以上の情報源記号をもつ情報源Sが与えられたとする．Sに対する瞬時符号Cがコンパクトであるならば，Cに対する符号の木Tにおいて，

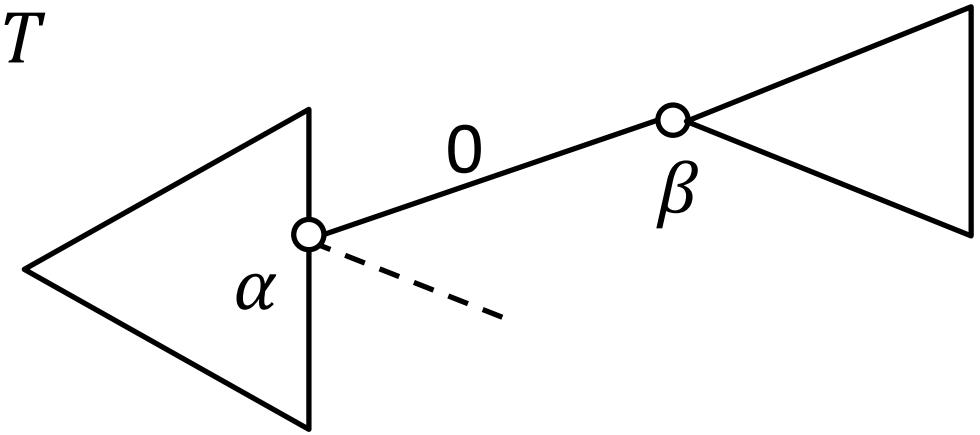
(1) 葉以外の節点は必ず2個の子節点を持つ．

(2) 情報源記号α, βの出現確率をそれぞれp_α, p_βとする．$p_\alpha < p_\beta$ならば情報源記号α, βに割り付けられる情報源記号をそれぞれ，c_α, c_βとすると，$|c_\alpha| \geq |c_\beta|$である．

註：2元符号と仮定

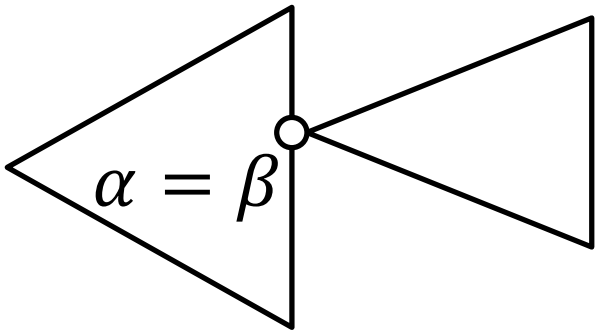
コンパクト符号

符号の木T

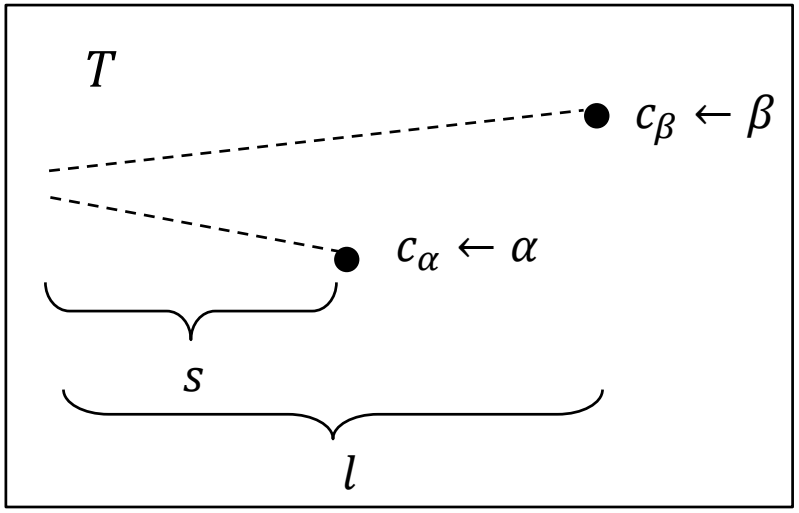


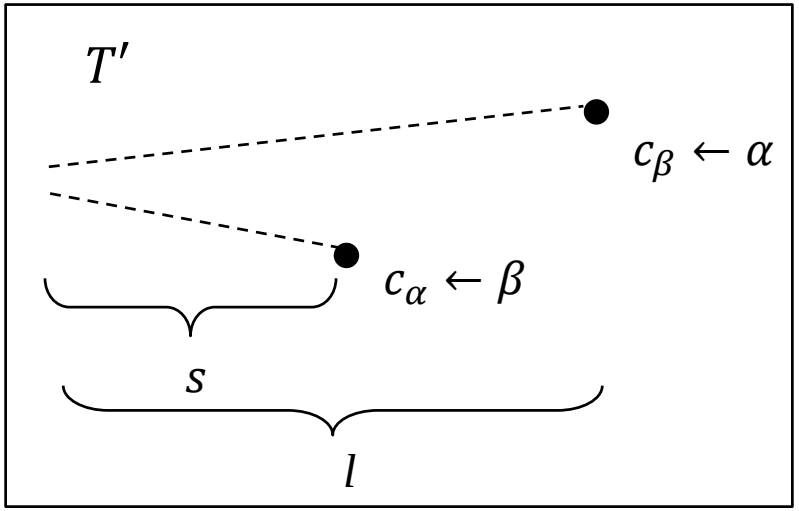
ここで，ノードαは分岐0に対応する子節点βしか持っていないとしたら，下のようにするほうが平均符号長が短くなることは明らか．

符号の木T'



コンパクト符号





コンパクト符号に対する符号Cの木Tにおいて$p_\alpha < p_\beta$なる出現確率を持つ情報源記号α, βに割り付けられる符号語c_α, c_βの長さ，つまり，深さを$s = |c_\alpha|$, $l = |c_\beta|$とすると，$s \geq l$となる.

なぜならば，$s < l$と仮定すると，T において情報源記号α, βへの符号語の割り付け方だけを入れ替えた符号の木 T'に対応する符号C'の平均符号長

$$L' = \cdots + sp_\beta + lp_\alpha$$

を

$$L = \cdots + sp_\alpha + lp_\beta$$

と比較すると，$s < l, p_\alpha < p_\beta$から

$$L' - L = (sp_\beta + lp_\alpha) - (sp_\alpha + lp_\beta)$$
$$= (l - s)(p_\alpha - p_\beta < 0$$

つまり，$L' < L$ となって，Cがコンパクトであるという前提に反するからである.

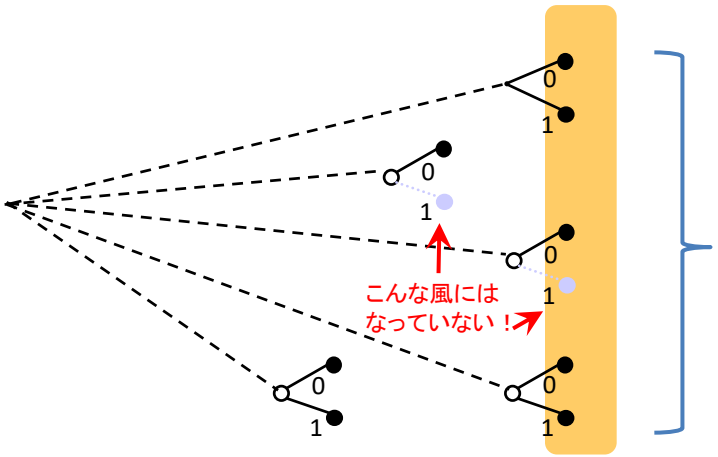
註：2元符号と仮定

コンパクト符号

【補助定理1】（瞬時符号Cがコンパクトであるための必要条件）

2個以上の情報源記号をもつ情報源Sが与えられたとする．Sに対する瞬時符号Cがコンパクトであるならば，Cに対する符号 の木Tにおいて，

(1) 深さdが最大の葉（最高次の葉）は少なくとも2枚あり，そのうちの一つは出現確率の最も小さい情報源記号に割り付けられている．
(2) 深さdの葉に割り付けられている情報源記号のなかには，Sの情報源記号からαを除いた情報源記号のなかで最も小さな出現確率をもつものがある．



深さ最大の節点たち

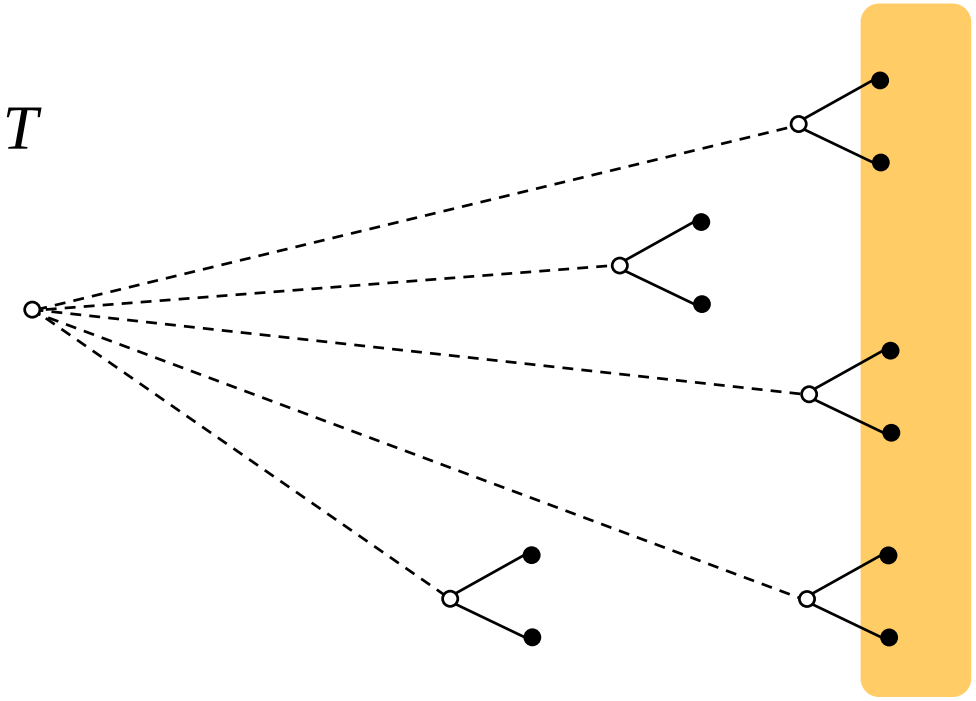
このなかのどれかはSの情報源記号のなかで出現確率の最小のものαに割り付けられている．

それ以外の葉のどれかは，Sの情報源記号からαを除いた情報源記号のなかで最も小さな出現確率をもつものに割り付けられている．

註：2元符号と仮定

コンパクト符号

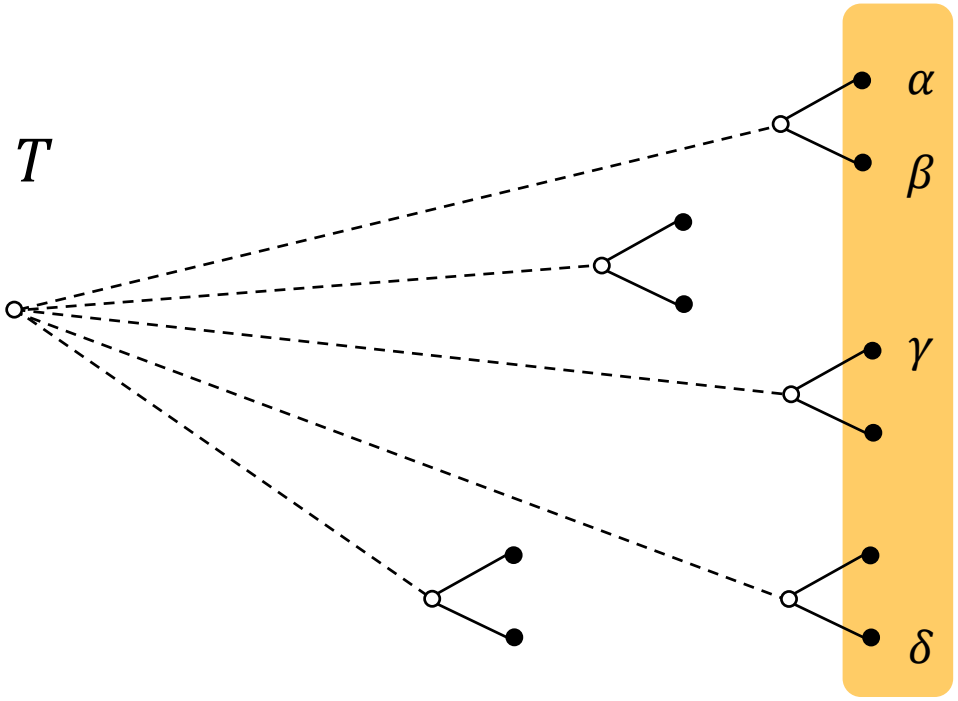
3個以上の情報源記号をもつ情報源Sの情報源記号を$A = \{a_1, \cdots, a_n\}$，各情報源記号a_iの出現確率をp_iとする．また，すべての$1 \leq j \leq n-2$に対して$p_j \geq p_{n-1} \geq p_n$とする．$S$のコンパクトな瞬時符号$T$において，$a_{n-1}$と$a_n$はどの節点に対応づけられているか？



註：2元符号と仮定

コンパクト符号

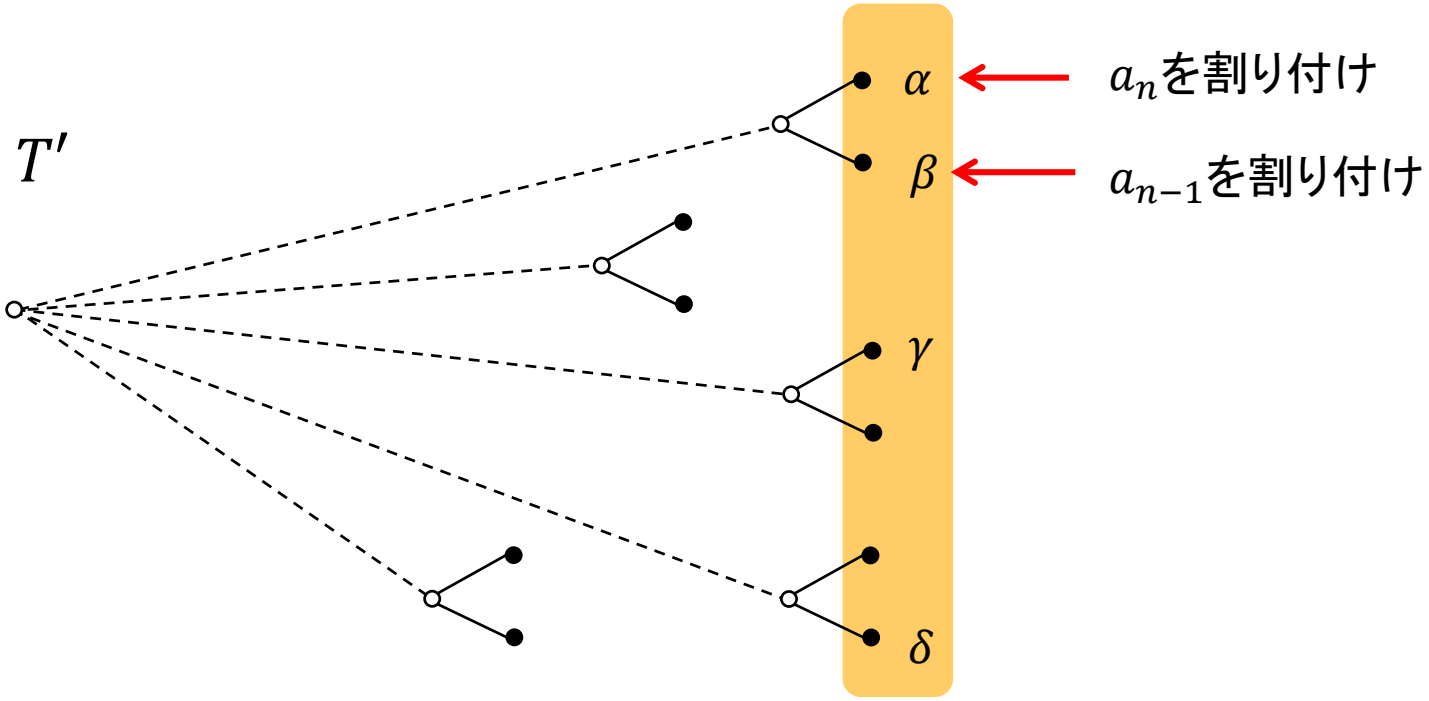
典型的には，a_{n-1}とa_nは最高次の節点（例えば，下図α, β）に割り付けられる．しかし，$1 \leq \{i, j\} \leq n-2$に対して$p_i = p_j = p_{n-1} = p_n$となっているときは，下図節点$\alpha$や$\beta$には$a_i$や$a_j$が割り付けられ，$a_{n-1}$と$a_n$は隣り合っていないかもしれない（例えば，$\gamma$と$\delta$）に割り付けられることもある．



註：2元符号と仮定

コンパクト符号

いずれの場合でも，平均符号長を変えることのないように，節点への符号語の割り付けを変えることにより，下図のT'に対応するコンパクト符号を構成できる.



註：2元符号と仮定

コンパクト符号

【補助定理2】　（コンパクトな瞬時符号Cを構成するための十分条件）

3個以上の情報源記号をもつ情報源Sの情報源記号を$A = \{a_1, \cdots, a_n\}$，各情報源記号a_iの出現確率をp_iとする．また，すべての$1 \leq j \leq n-2$に対して$p_j \geq p_{n-1} \geq p_n$とする．

このとき，a_nとa_{n-1}を一つの記号b_{n-1}に統合して，情報源記号の集まり$A' = \{a_1, \cdots, a_{n-2}, b_{n-1}\}$，各情報源記号の出現確率を$p_1, \cdots, p_{n-1} + p_n$とする情報源$S'$のコンパクトな瞬時符号$C'$が得られたとする．

すると，C'において，情報源記号b_{n-1}に割り付けられた符号語c_bのかわりに，$a_{n-1} \leftarrow c_b 0, a_n \leftarrow c_b 1$という割り付けを加えた符号$C$もまたコンパクトな瞬時符号である．

註：2元符号と仮定

コンパクト符号

【補助定理2】の使い方

情報源S: 情報源記号発生確率 $\langle A: 0.6, B: 0.2, C: 0.1, D: 0.07, E: 0.03\rangle$

情報源S_1: 情報源記号発生確率 $\langle A: 0.6, B: 0.2, C: 0.1, D: 0.1\rangle$

情報源S_2: 情報源記号発生確率 $\langle A: 0.6, B: 0.2, C: 0.2\rangle$

情報源S_3: 情報源記号発生確率 $\langle A: 0.6, B: 0.4\rangle$

$\langle A \leftarrow 0, B \leftarrow 1\rangle$は情報源$S_3$のコンパクト符号

$\langle A \leftarrow 0, B \leftarrow 10, C \leftarrow 11\rangle$は，情報源$S_2$のコンパクト符号

$\langle A \leftarrow 0, B \leftarrow 10, C \leftarrow 110, D \leftarrow 111\rangle$は，情報源$S_1$のコンパクト符号

$\langle A \leftarrow 0, B \leftarrow 10, C \leftarrow 110, D \leftarrow 1110, E \leftarrow 1111\rangle$は，情報源Sのコンパクト符号

註：2元符号と仮定

コンパクト符号

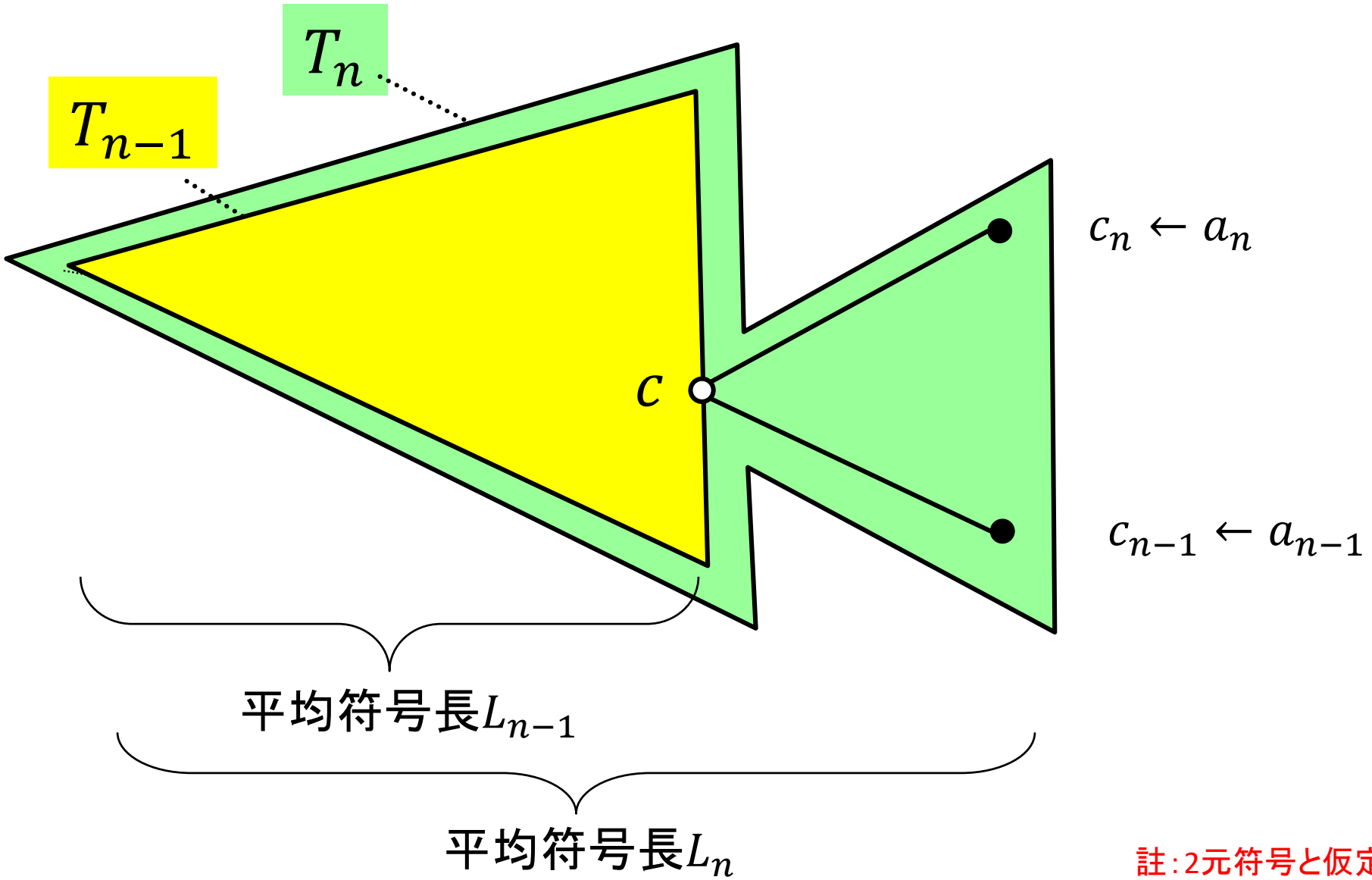
【補助定理2】の証明骨子

情報源S_n: 情報源記号発生確率　$\langle a_1{:}p_1, \cdots, a_{n-1}{:}p_{n-1}, a_n{:}p_n \rangle$

T_{n-1}がコンパクトであるにも関わらず，T_nがコンパクトでないと仮定する．すると，$A = \{a_1, \cdots, a_n\}$に対するコンパクト符号C'_nが存在し，その平均符号長L'_nはT_nの平均符号長L_nより小さいことになる．補助定理1に関わる議論から，C'_nと平均符号長の等しいコンパクト符号C''_nが存在し，C''_nにおいて，最も深い節点には出現確率p_nとp_{n-1}をもつ符号語が対応づけられている．C''_nにおいて，a_nとa_{n-1}に対する符号語割り付けの代わりに，出現確率$p_n + p_{n-1}$の情報源記号b_{n-1}に対して符号語cを割り付ける符号C'_{n-1}を構成する．すると，C'_{n-1}は，その作り方からC_{n-1}と同じ情報源に対する瞬時符号であり，C'_{n-1}の平均符号長$L'_{n-1} = L'_n - p_n - p_{n-1}$が$T_{n-1}$の平均符号長$L_{n-1} = L_n - p_n - p_{n-1}$より短いことになり，$T_{n-1}$がコンパクトであるという前提に矛盾する．

註：2元符号と仮定

コンパクト符号

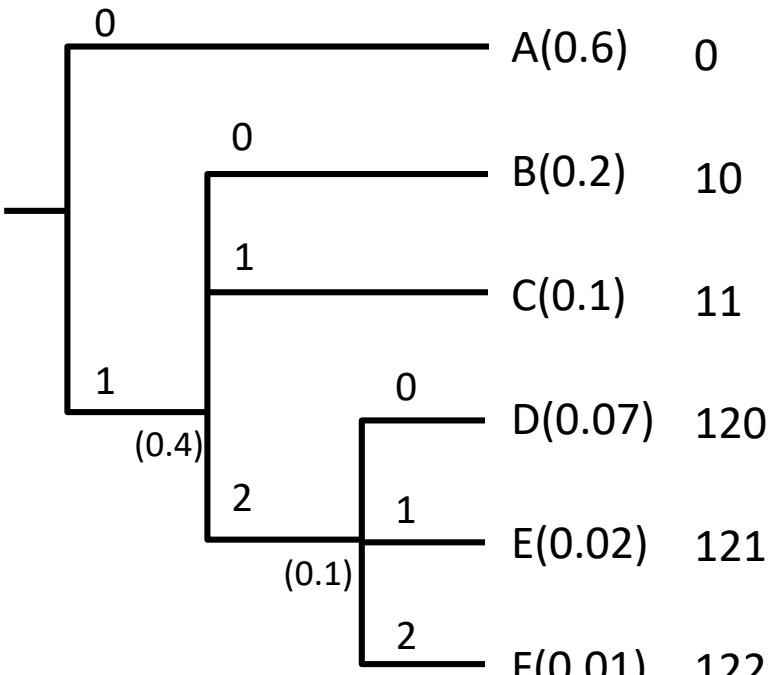


註：2元符号と仮定

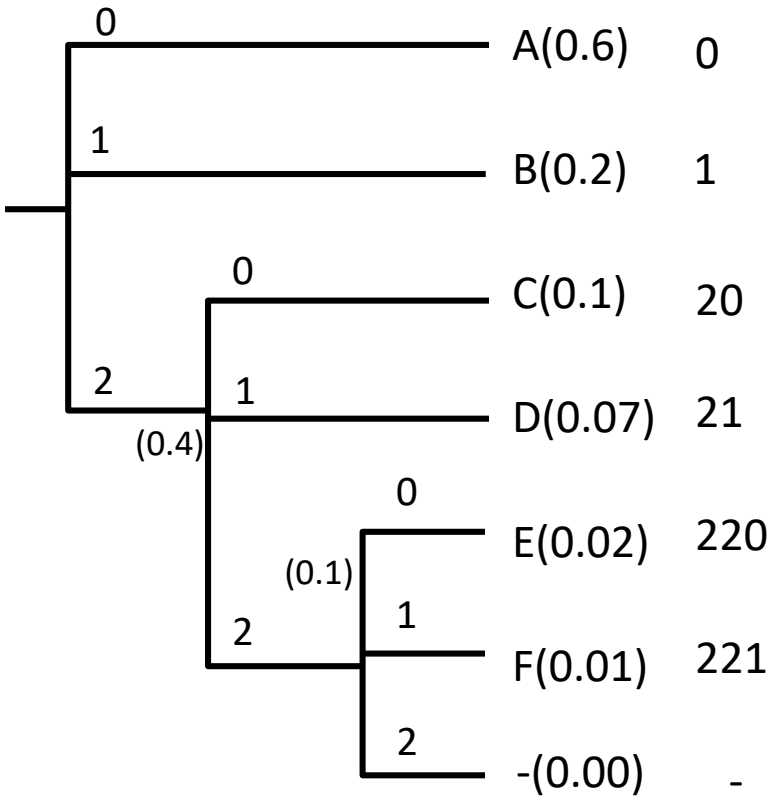
ハフマン符号

- 3元の場合

例：情報源記号発生確率を〈A: 0.6, B : 0.2, C: 0.1, D: 0.07 , E: 0.02, F: 0.01〉



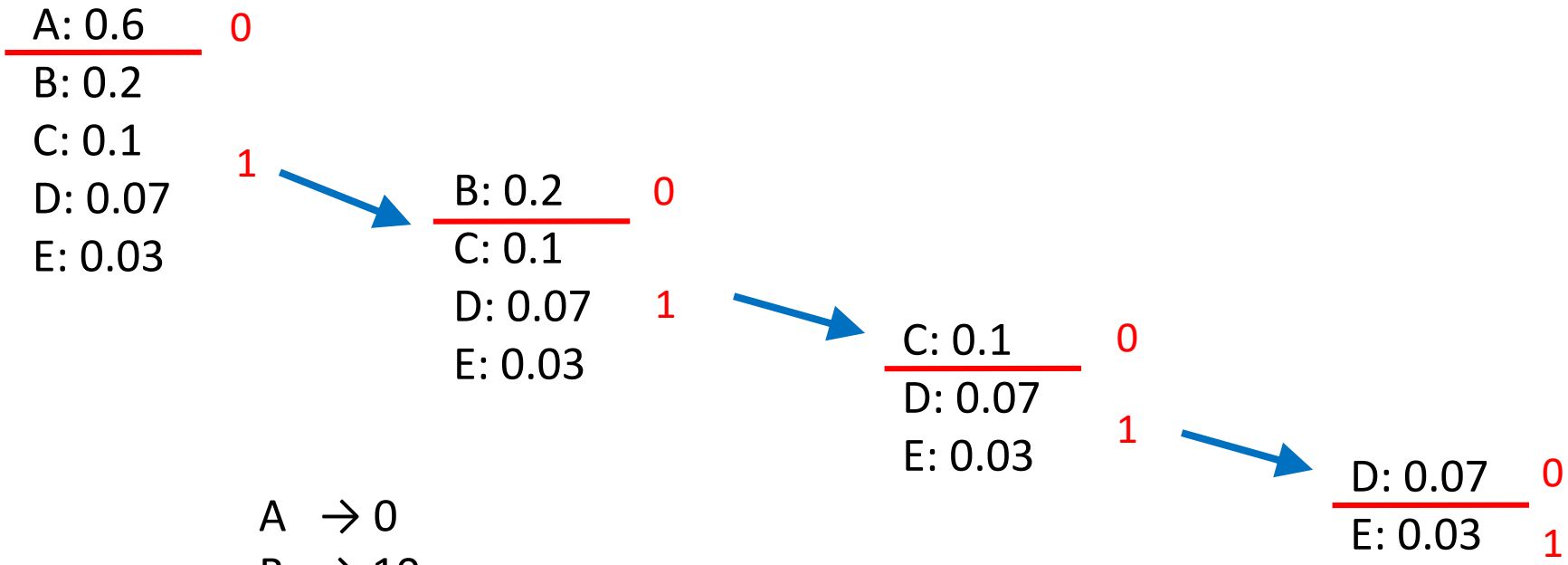
平均符号長　1.5



平均符号長　1.23

シャノン-ファノ符号

例：情報源記号発生確率を〈A: 0.6, B : 0.2, C: 0.1, D: 0.07 , E: 0.03〉

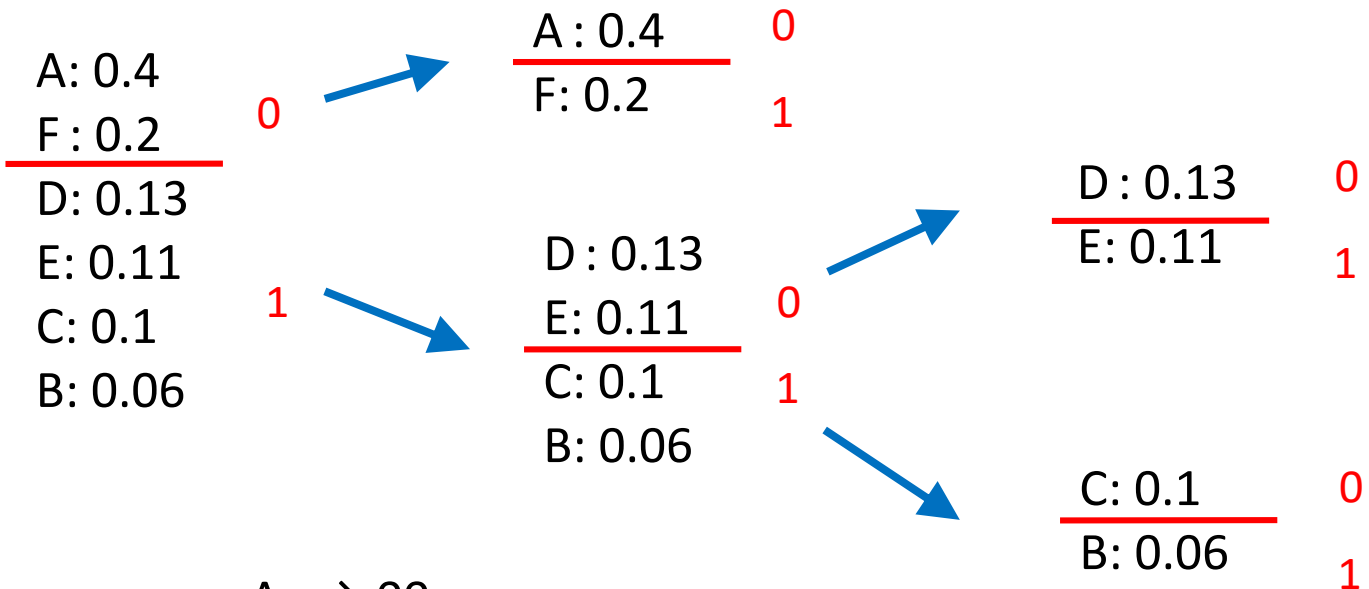


A → 0
B → 10
C → 110
D → 1110
E → 1111

平均符号長　1.7

シャノン-ファノ符号

例：情報源記号発生確率を〈A: 0.4, B : 0.06, C: 0.1, D: 0.13, E: 0.11, F:0.2〉



A → 00
F → 01
D → 100
E → 101
C → 110
B → 111

平均符号長　2.4

課題

- シャノン-ファノ符号化は必ずしもコンパクト符号を生成するとは限らない．どのような場合か？

まとめ

- コンパクト符号：平均符号長の最も短い瞬時符号
- ハフマン符号：コンパクト符号構成法
- どうしてハフマン符号がコンパクト符号になるのか？
- 補助定理１：瞬時符号がコンパクト符号であるための必要条件
- 補助定理２：コンパクトな瞬時符号の十分条件
- コンパクト符号は唯一とは限らない
- ハフマン符号で構成されないコンパクト符号
- 3元符号の場合
- シャノン-ファノ符号