講義「情報理論」

第7回　情報源符号化法（1）

情報理工学部門 情報知識ネットワーク研究室
喜田拓也

講義資料

2019/7/24

いろいろな符号と要求条件（おさらい）

一意復号可能な符号 (uniquely decodable code)

一意（できれば瞬時）に復号可能なこと

C5
0
0 1
0 1 1
1 1 1

C3
0
0 1
1 0
0

特異符号

瞬時符号 (Instantaneous code)：前から解読可能な符号

等長符号

C1
0 0
0 1
1 0
1 1

カンマ符号

C2
0
1 0
1 1 0
1 1 1 0

ハフマン符号

C6
0
1 0
1 1 0
1 1 1

仕組みが簡単で装置が高速安価なこと

平均符号長が短いこと

C4
0
0 1
1 0
1 1

一意復号不可能な符号

平均符号長 L の限界(おさらい)

定理 4.3 [情報源符号化定理(シャノンの第一基本定理)]

情報源Sは，任意の一意復号可能なr元符号で
符号化する場合，その平均符号長Lは，

$$\frac{H(S)}{\log_2 r} \leq L$$

を満たす．
また，任意の正数$\varepsilon > 0$について平均符号長Lが，

$$L < \frac{H(S)}{\log_2 r} + \varepsilon$$

となるr元瞬時符合を作ることができる．

Claude Shannon (1916 - 2001)
www.ausbcomp.com/~bbott/wik/mmtimeln.htmより

2元符号の場合，
$\log_2 r = 1$

どんなに工夫しても，平均符号長 L はエントロピー$H(S)$までしか改善できない（でもがんばれば，そこまではできる）

今日の内容

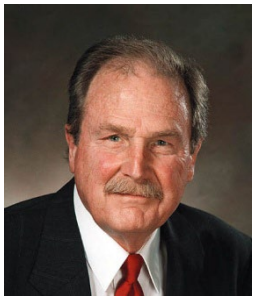
5.1 ハフマン符号

5.3 ブロックハフマン符号

ハフマン符号はなぜ大事か？

ハフマン符号は［　　］である！

［　　］とは，１記号ずつ符号化する際，その平均符号長を最小とする効率のよい符号のこと



David Albert Huffman
(1925 –1999)

http://www.adeptis.ru/vinci/m_part5_2.html より

ハフマン符号の作り方

各記号が下の表で与えられる確率分布で出力されるような，
記憶のない5元定常情報源を考える．
この情報源から出力される系列をハフマン符号化しよう．

情報源記号 x	確率 $P(x)$
A	0.55
B	0.14
C	0.06
D	0.15
E	0.1

ハフマン木の作り方

（STEP 0）まず初めに，確率の高い順に記号を並べ替える

※必須ではないが，しておくと後が楽

情報源記号 x	確率 $P(x)$
A	0.55
B	0.14
C	0.06
D	0.15
E	0.1

並べ替え

情報源記号 x	確率 $P(x)$
A	0.55
D	0.15
B	0.14
E	0.1
C	0.06

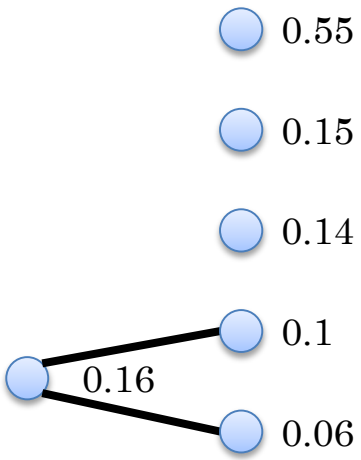
ハフマン木の作り方

（STEP 1） 各記号に対応する符号木の葉を作る
葉には確率を添えて書いておく

	情報源記号 x	確率 $P(x)$
0.55	A	0.55
0.15	D	0.15
0.14	B	0.14
0.1	E	0.1
0.06	C	0.06

ハフマン木の作り方

（STEP 2） 最も確率が小さい葉を二つ選び，それを集約するためのノードを新たに作って枝で結ぶ．
そのノードを新しい葉として扱い，元の二つの葉の確率を足し合わせたものを添える．

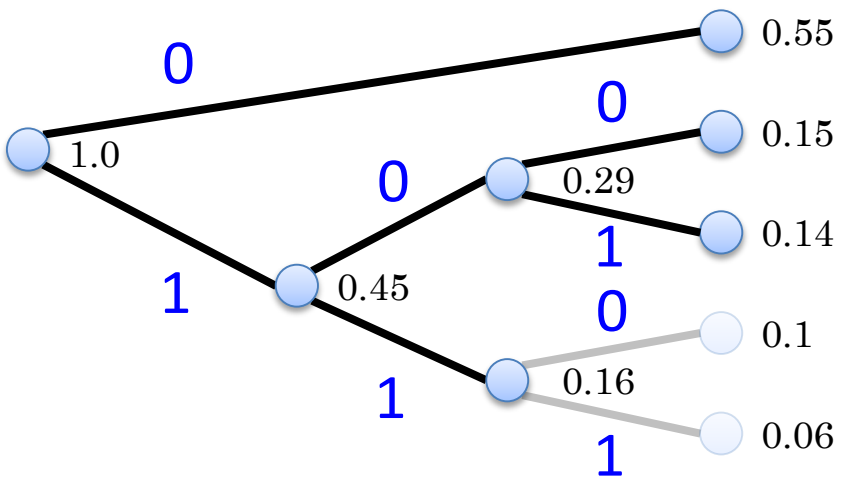


情報源記号 x	確率 $P(x)$
A	0.55
D	0.15
B	0.14
E	0.1
C	0.06

ハフマン木の作り方

（STEP 3） STEP 2を，繰り返して符号木を作る.

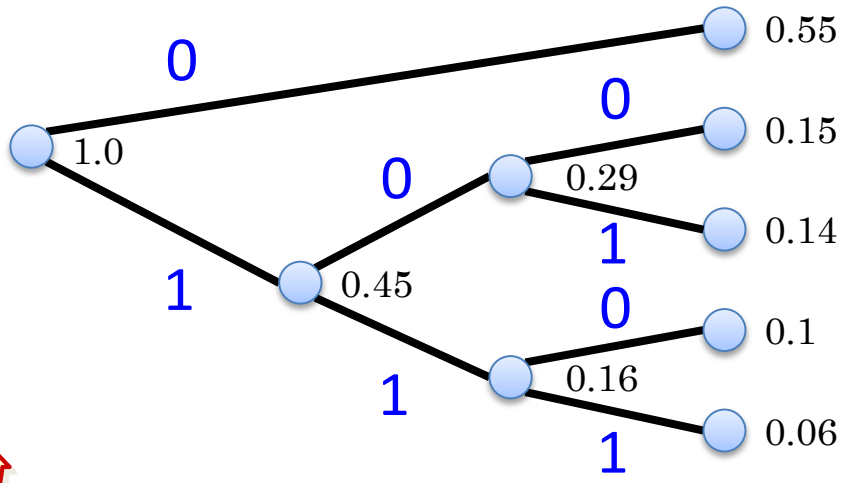
（STEP 4） 各ノードから葉へ向かう方向の2本の枝に，
0と1のラベルを割り当てる.



情報源記号 x	確率 $P(x)$
A	0.55
D	0.15
B	0.14
E	0.1
C	0.06

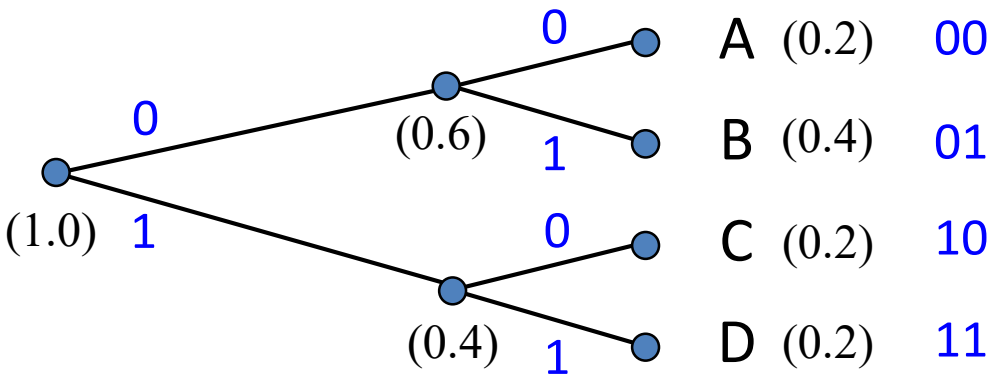
ハフマン符号

（符号の構成）構築した符号木を用いて，根から各々の葉へ至るパスをなぞりながら，ラベルの列を符号語として記号に割り当てる



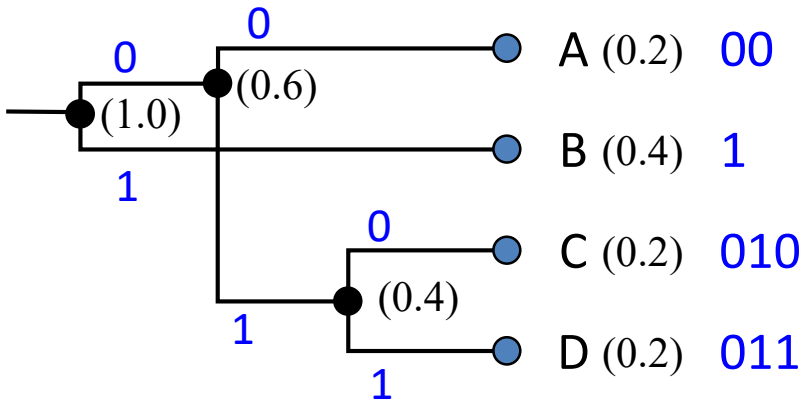
情報源記号 x	確率 $P(x)$	符号語
A	0.55	0
D	0.15	100
B	0.14	101
E	0.1	110
C	0.06	111

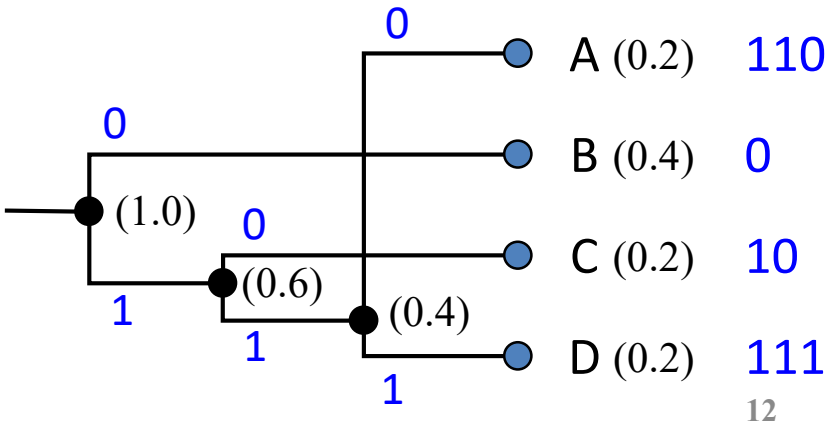
ハフマン符号は数通り作成できる



情報源記号	確率
A	0.2
B	0.4
C	0.2
D	0.2

他にも，複数通りの符号化が可能（平均符号長はすべて同じ）





ハフマン符号の構築アルゴリズム(まとめ)

2元ハフマン符号構成法

1. 各情報源記号に対応する葉を作る．各々の葉には，情報源記号の発生確率(葉の確率)を記しておく．
2. 確率の最も小さい2枚の葉に対して一つ節点を作り，その節点と2枚の葉を枝で結ぶ．2本の枝の一方に0を，他方に1を割り当てる．さらにこの節点に，2枚の葉の確率の和を記し，この節点を新たな葉と考える．
3. 葉が1枚になったら，符号木の構成を終了する．
4. そうでなければ(2)に戻り処理を繰り返す．

手順
(アルゴリズム)

※この処理過程において，符号語が符号の木の葉にだけ割り当てられているので，ハフマン符号は瞬時符号である．

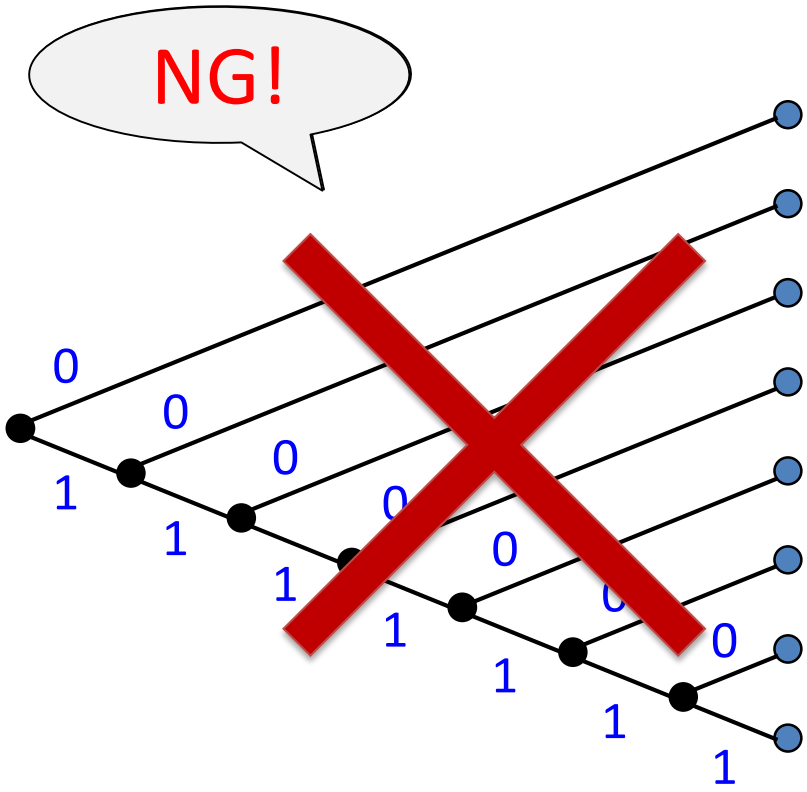
ハフマン符号化してみよう！

各情報源が右の表のような確率分布を持つ記憶のない定常情報源Sが与えられたとする.
この定常情報源Sに対するハフマン符号を構築せよ.

情報源記号 x	確率 $P(x)$
A	0.363
B	0.174
C	0.143
D	0.098
E	0.087
F	0.069
G	0.045
H	0.021

答え？

$H(S) \fallingdotseq 2.59$



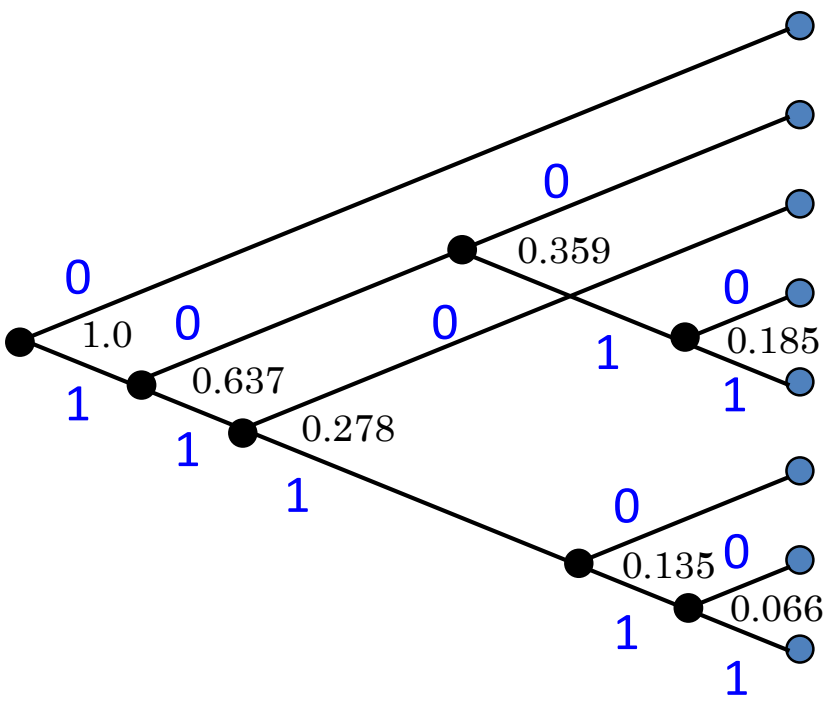
情報源記号	確率	符号語
A	0.363	0
B	0.174	10
C	0.143	110
D	0.098	1110
E	0.087	11110
F	0.069	111110
G	0.045	1111110
H	0.021	1111111

平均符号長 L を求めてみよう！

$$L = 0.363 \times 1 + 0.174 \times 2 + 0.143 \times 3 + 0.098 \times 4 + 0.087 \times 5 + 0.069 \times 6 + 0.045 \times 7 + 0.021 \times 7 = 2.843$$

正しい答え

$H(S) \fallingdotseq 2.59$



情報源記号	確率	符号語
A	0.363	0
B	0.174	100
C	0.143	110
D	0.098	1010
E	0.087	1011
F	0.069	1110
G	0.045	11110
H	0.021	11111

平均符号長 L を求めてみよう！

Try 練習問題5.1

$$L = 0.363 \times 1 + 0.174 \times 3 + 0.143 \times 3 + 0.098 \times 4 + 0.087 \times 4 + 0.069 \times 4 + 0.045 \times 5 + 0.021 \times 5 = 2.66$$

ちょっと休憩

ブロック符号化

情報源の一記号ごとに符号化すると・・・

（2元情報源） 0, 1 → 0, 1 （2元符号化）

まったく効率が上がらない！

連続する何個かの情報源記号をまとめて符号化しよう！

一定個数の情報源記号ごとにまとめて符号化する方法を
ブロック符号化（block coding）と呼ぶ

特に，もとの情報源Sに対し，n次拡大情報源S^n を考え，
その上の記号に対してハフマン符号化を行う方法を，
ブロックハフマン符号化（block Huffman coding）と呼ぶ

ブロック符号化の例

1, 0をそれぞれ確率0.2，0.8で発生する記憶のない2元定常情報源を考え，これが発生する系列を2個ずつまとめて符号化する

ブロックごとの平均符号長L'は

$$L' = 1 \times 0.64 + 2 \times 0.16 + 3 \times 0.16 + 3 \times 0.04 = 1.56$$

情報源系列	確率	ハフマン符号
0 0	0.64	0
0 1	0.16	10
1 0	0.16	110
1 1	0.04	111

一記号あたりの平均符号長Lは

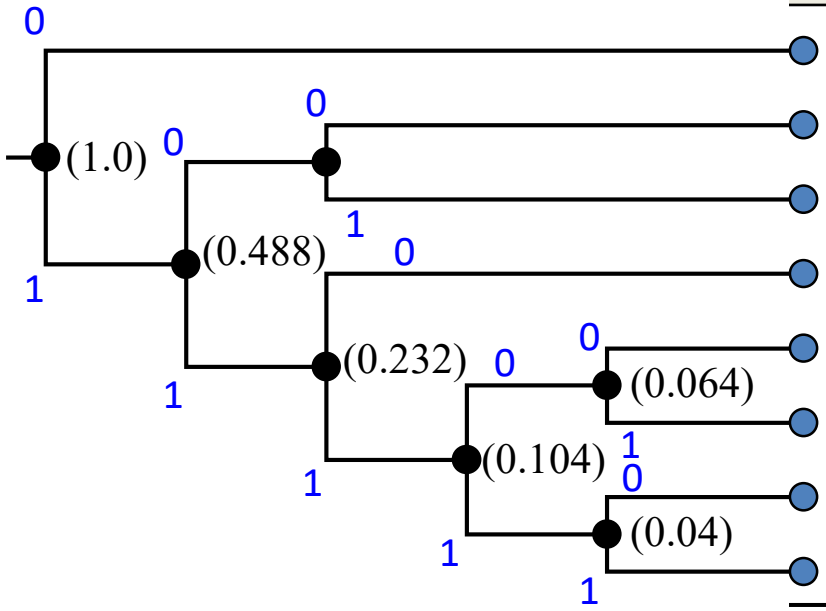
$$L = L'/2 = 0.78$$

22%の
効率アップ

3個まとめて符号化すると・・・

先の例では，平均符号長が 0.78 だった

エントロピーは $H(S) = \mathcal{H}(0.2) \fallingdotseq 0.7219$ なので，まだ差がある



情報源系列	確率	ハフマン符号
0 0 0	0.512	0
1 0 0	0.128	100
0 1 0	0.128	101
0 0 1	0.128	110
1 1 0	0.032	11100
0 1 1	0.032	11101
1 0 1	0.03	
1 1 1	0.008	

より近くなった！

$$L = \frac{1 \times 0.512 + 3 \times (0.128 + 0.128 + 0.128) + 5 \times (0.032 + 0.032 + 0.032 + 0.008)}{3} = 0.728$$

Try 練習問題5.2

今日のまとめ

ハフマン符号

- ハフマン符号はコンパクト符号
- ハフマン符号の構築アルゴリズム

ブロック符号化

- ブロックハフマン符号： n次拡大情報源に対するハフマン符号

次回：

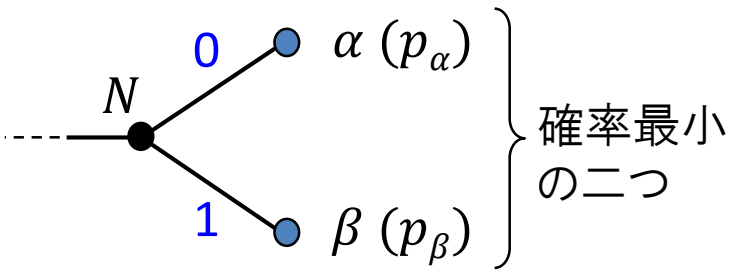
- 情報源符号化法つづき

補足資料：補助定理5.1

補助定理5.1

コンパクトな瞬時符号の符号木において，最も長い符号語に対応する葉は少なくとも二つあり，それらのどの葉に対しても共通の親を持つもう一つの葉が存在する.

そして，これらの二つの葉は，生起確率が最も小さい二つの情報源記号に対応している.

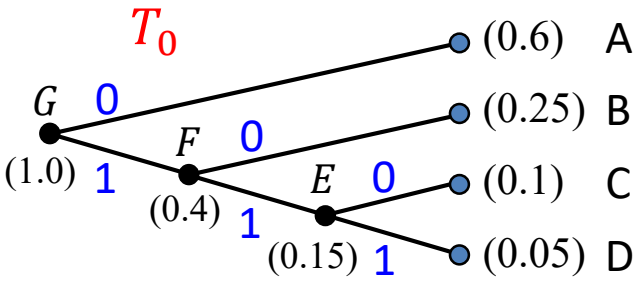


コンパクト符号の符号木における最高次の葉

※証明は教科書を参照のこと

ハフマン符号がコンパクト符号である証明(1)

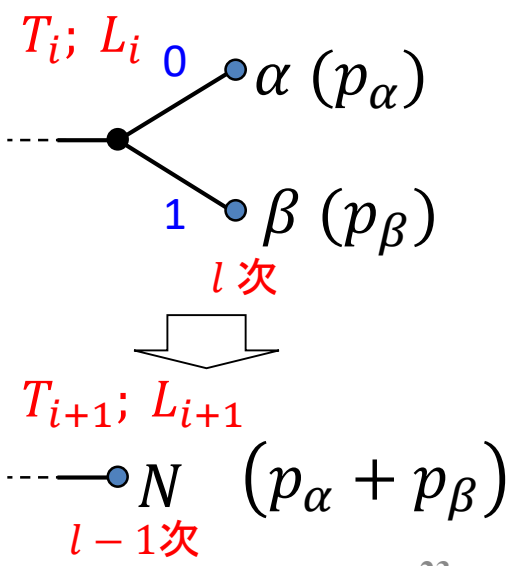
【証明】 ハフマン符号の木をT_0とし，構成法の(STEP 2)によって葉がつぶれていくと見る．このとき，i ステップ目の木を T_iとすると，最終段階の木はただ二つの葉からなる木であるので，コンパクト符号の木である．そこで，T_{i+1}がコンパクト符号の木であると仮定して，T_iもコンパクト符号の木であることが証明できれば帰納法によりT_0 もコンパクト符号の木であるといえる．



T_{i+1}とT_iの平均符号長 L_{i+1}とL_iの関係について考えてみよう．T_iの確率最小の二つの葉の確率をp_α, p_βとすると，T_{i+1}ではこれらが1つの葉にまとめられ，枝一本分短くなるから，それらの葉が l 次の葉であるとすると，

$$L_{i+1} = L_i - l\,p_\alpha - l\,p_\beta + (l-1)(p_\alpha + p_\beta)$$
$$= L_i - p_\alpha - p_\beta$$

である．



ハフマン符号がコンパクト符号である証明(2)

ここで，T_{i+1}がコンパクト符号の木であるのに，T_iがそうでないとする．すると，T_iと同じ葉（および同じ確率）を持ち，平均符号長がより短いコンパクトな瞬時符号の木が存在するはずである．そのような木をT_i'とし，その平均符号長をL_i'とする．仮定により，$L_i' < L_i$である．

補助定理5.1より，T_i'には確率最小の二つの葉が存在する．そこで，これらをまとめて節点N'を葉とした新たな符号の木T_{i+1}'を作る．この木はT_{i+1}と全く同じ葉を持ち，その平均符号長は

$$L_{i+1}' = L_i' - p_\alpha - p_\beta$$
$$< L_i - p_\alpha - p_\beta = L_{i+1}$$

となる．これは，T_{i+1}がコンパクト符号の木であるという前提に矛盾する．よってT_iもコンパクト符号でなければならない．【証明終】