情報理論 (5)
情報源符号化定理

人間コミュニケーション学科
梶本　裕之
kajimoto@hc.uec.ac.jp

授業の流れ（予定）

第4週 (11/04)　情報源符号化とデータ圧縮
第5週 (11/11)　出張のため休講
第6週 (11/18)　調布祭のため休講
第7週 (11/25)　ハフマン符号とデータ圧縮
第8週（12/ 2)　情報源符号化定理
第9週 (12/09)　出張のため休講
第10週 (12/16)　マルコフ情報源モデル
第11週 (12/23)　休日
第12週 (1/ 6)　通信路のモデル化
第13週 (1/13)　センター試験準備のため休講　←変更点
第14週 (1/20)　誤り検出・誤り訂正符号
第15週 (1/27)　線形符号
第16週 (2/ 3)　ハミング符号，秘密鍵暗号
第17週 (2/10)　公開鍵暗号　←変更点

前回の小レポート(1)

- 身の周りにある情報源からある程度複雑なもの（事象数25個程度以上，データ量500事象程度以上：コンピュータで処理できるテキストデータなどがよい）を選び，それをハフマン符号によって符号化し，固定長符号による符号化に比較した圧縮率を求めよ．
- （手作業は無理．プログラミングに慣れた人向け）

略

前回の小レポート(2)

- 以下の文字列を動的ハフマン符号を用いて符号化せよ．また静的ハフマン符号を用いた場合との圧縮率の違いを比較せよ．
- 注意：アスキー符号を使ってしまうと，一文字あたり7bitになり，非常に長くなってしまう．まず固定長符号の表を自作し，その後動的ハフマン符号を適用する．

C D E G F E B A E A
C E F A D A E F E A

固定長符号の場合

- AからGまでの7文字⇒3bitで表せる

C D E G F E B A E A
C E F A D A E F E A

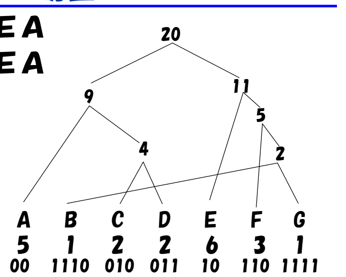
A	000	E	100
B	001	F	101
C	010	G	110
D	011		

静的ハフマンの場合

C D E G F E B A E A
C E F A D A E F E A

A	5	E	6
B	1	F	3
C	2	G	1
D	2		

出現回数



平均符号長=(2×5+ 4×1+ 3×2+ 3×2+ 2×6+ 3×3+ 4×1)/20
=2.55

動的ハフマンの場合

A	000	E	100
B	001	F	101
C	010	G	110
D	011		

CDEGFEBAEA
CEFADAEFEA

符号化された記号列

0 010

符号木 0 ⟷ 1

	A	B	C	D	E	F	G
積算頻度	0	0	1	0	0	0	0
符号			0				

動的ハフマンの場合

A	000	E	100
B	001	F	101
C	010	G	110
D	011		

CDEGFEBAEA
CEFADAEFEA

符号化された記号列

0 010
0 011

符号木 0 ⟷ 1

	A	B	C	D	E	F	G
積算頻度	0	0	1	1	0	0	0
符号			0	1			

動的ハフマンの場合

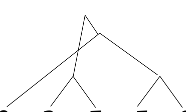
A	000	E	100
B	001	F	101
C	010	G	110
D	011		

CDEGFEBAEA
CEFADAEFEA

符号化された記号列

0 010
0 011
0 100
0 110
0 101

符号木 0 ⟷ 1



	A	B	C	D	E	F	G
積算頻度	0	0	1	1	1	1	1
符号			10	00	01	110	111

動的ハフマンの場合

A	000	E	100
B	001	F	101
C	010	G	110
D	011		

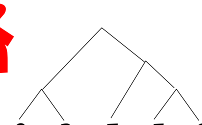
CDEGFEBAEA
CEFADAEFEA

符号化された記号列

0 010
0 011
0 100
0 110
0 101
1 01

符号木 0 ⟷ 1

以下略



	A	B	C	D	E	F	G
積算頻度	0	0	1	1	2	1	1
符号			00	01	10	110	111

この場合の動的ハフマンは

- 文字数が長くなるにつれて，圧縮率は静的ハフマンに近づく(2.55)
- しかし，符号の頭に1,0の記号を入れているので，いくらがんばっても3.55にしかならない
- つまり，実は固定長符号よりも悪い結果に．．．

今日の話

ハフマン符号で達成される
実際の最小平均符号長 Lmin

情報源のエントロピーH(X)

この二つの関係

復習：平均符号長とエントロピー

Lmin:平均符号長

X：整数制約を外した場合の符号長

- L min ≧ min Σi pi xi
- = − Σi pi log pi　　情報源が持つエントロピー

情報源のエントロピーは，平均符号長の下界であった．

Lmin ≧H(X)

13

最小平均符号長とエントロピー

- 情報源の確率分布 $\{p_i\}$ が2のべき乗であるような特殊な場合を除くと，

最小平均符号長(ハフマン符号化で実現)は 一般に情報源のエントロピーより大きい

⬇

その差に上限はあるか？

14

−log p を切り上げた符号長を持つ符号

- 各記号が $-\log_2 p_i$ に近い符号長をもつ符号

$$L_i = \lceil -\log_2 p_i \rceil$$

を考える（⌈ ⌉は切り上げ）．つまり，

$$-\log_2 p_i \leqq L_i < -\log_2 p_i + 1$$

このような符号長をもつ符号は作成可能か？

15

（復習）Kraft の不等式

- 事象数：n
- 各事象に対する符号長：L_i (i=1～n)

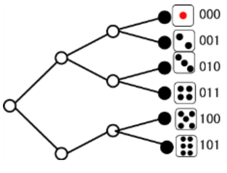
プレフィックス符号（あるいは復号可能な符号）が存在する必要十分条件は

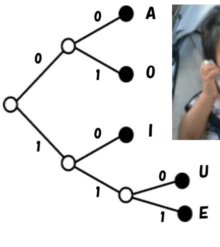
$$\Sigma_{i=1\sim n}\, 2^{-L_i} \leqq 1$$

16

（復習） Kraft の不等式(例)

- $\Sigma_i\, 2^{-L_i} \leqq 1$ が成立することを確かめる．



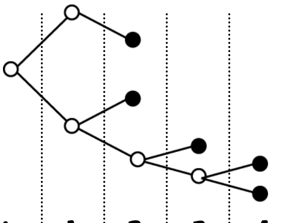


- $2^{-3} \times 6 = 6/8 < 1$
- $2^{-2}+2^{-2}+2^{-2}+2^{-3}+2^{-3} = 3/4+2/8 = 1$

17

（復習） Kraftの不等式の意味(1)

Kraft の不等式　　$\Sigma_i\, 2^{-L_i} \leqq 1$

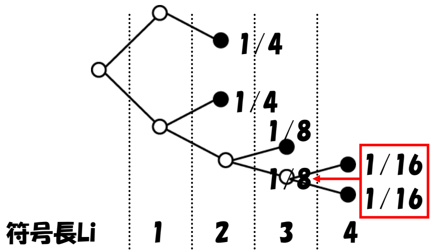


符号長Li	1	2	3	4
2^{-Li}	1/2	1/4	1/8	1/16

18

（復習） Kraftの不等式の意味(2)

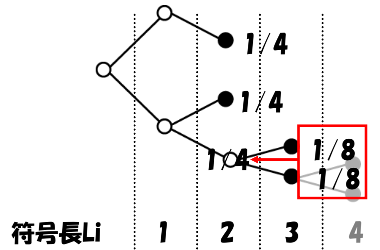
Kraft の不等式　$\Sigma_i\, 2^{-Li} \leqq 1$



右から順に足していく

（復習） Kraftの不等式の意味(3)

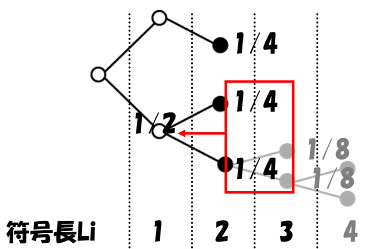
Kraft の不等式　$\Sigma_i\, 2^{-Li} \leqq 1$



右から順に足していく

（復習） Kraftの不等式の意味(4)

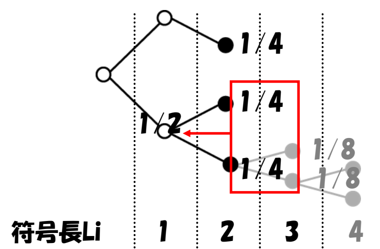
Kraft の不等式　$\Sigma_i\, 2^{-Li} \leqq 1$



右から順に足していく

（復習） Kraftの不等式の意味(5)

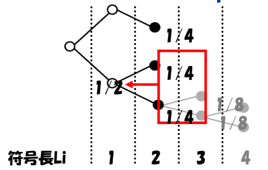
Kraft の不等式　$\Sigma_i\, 2^{-Li} \leqq 1$



右から順に足していく

（復習） Kraftの不等式の意味(6)

Kraft の不等式　$\Sigma_i\, 2^{-Li} \leqq 1$



結局,
・プレフィックス符号であれば成り立つ.
・完全プレフィックス符号では等号が成立.
・ハフマン符号は完全プレフィックス符号.

(再) −log p を切り上げた符号長を持つ符号

- 各記号が $-\log_2 p_i$ に近い符号長をもつ符号

$$L_i = \lceil -\log_2 p_i \rceil$$

を考える（⌈ ⌉は切り上げ）．つまり,

$$-\log_2 p_i \leqq L_i < -\log_2 p_i + 1$$

このような符号長をもつ符号は
作成可能か？

$-\log_2 p_i \leqq L_i < -\log_2 p_i + 1$

- 左辺から

$$-L_i \leqq \log_2 p_i$$
$$2^{-L_i} \leqq 2^{\log_2 p_i}$$
$$\Sigma\, 2^{-L_i} \leqq \Sigma\, 2^{\log_2 p_i}$$
$$= \Sigma\, p_i = 1$$

よって，このような符号長をもつ符号は
Kraftの不等式を満たす．
つまり，この符号は

つくれる．

ではそのときの平均符長は？

平均符号長 $\Sigma\, p_i L_i$

$$-\log_2 p_i \leqq L_i < -\log_2 p_i + 1$$

- この符号の平均符号は．．．

$$\Sigma\, p_i L_i \geqq -\Sigma\, p_i \log_2 p_i$$
$$= H(X)$$
$$\Sigma\, p_i L_i < \Sigma\, p_i\,(-\log_2 p_i + 1)$$
$$= -\Sigma\, p_i \log_2 p_i + \Sigma\, p_i$$
$$= H(X) + 1$$

最小平均符号長の上限

$$\Sigma\, p_i L_i < H(X) + 1$$

結論：

情報源符号化において
その最小平均符号長が
情報源のエントロピーよりも
1以上大きくなることはない．

最小平均符号長の範囲

- 下限と上限をまとめると，

$$H(X) \leqq L_{min} < H(X) + 1$$

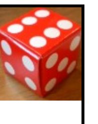
2進符号化する場合，ハフマン符号化すれば
情報源エントロピーに比べたムダは
1事象あたり平均1ビット未満に必ず収まる！

例題

- 以下の確率分布をもつ情報源をハフマン符号した際の平均符号長が情報源のエントロピー＋1ビットの範囲内に収まっていることを確認せよ．

  - {1/6, 1/6, 1/6, 1/6, 1/6, 1/6}
  - {0.4, 0.3, 0.1, 0.05, 0.05, 0.05, 0.05}

例題



– {1/6, 1/6, 1/6, 1/6, 1/6, 1/6}

H =

= }

1/6　1/6　1/6　1/6　1/6　1/6

L =

$\Sigma\, p_i L_i$ =

例題

– {0.4, 0.3, 0.1, 0.05, 0.05, 0.05, 0.05}

H =

=

0.4 0.3 0.1 0.05 0.05 0.05 0.05

L =

$\Sigma\, p_i L_i$ =

小レポート(1)

- 身の周りにある情報源から適当なものを選び,
  - 情報源エントロピー
  - ハフマン符号による最小平均符号長
  - $-\log_2 p_i$ の小数点以下を切り上げたものを符号長としたときの平均符号長

をそれぞれ求め,比較せよ.

最小平均符号長と
情報源のエントロピーは
どこまで近づきうるか

最小平均符号長とエントロピー

- 情報源の確率分布 $\{p_i\}$ が2のべき乗であるような特殊な場合を除くと,

最小平均符号長は 一般に
情報源のエントロピーより大きい

⬇

その差を縮めることはできないか?

最小平均符号長とエントロピー

その差を縮めることはできないか?

できる.

シャノンの情報源符号化定理

シャノンの情報源符号化定理

情報源の中の連続した事象をK個まとめて各々に新しい記号を与えたとき(K次の拡大情報源を作ったとき),Kを限りなく大きくすることによって,情報源のエントロピーに限りなく近い最小平均符号長を達成することができる.

拡大情報源

- 事象系 $A=\{a_i\}$ （i=1〜n）
- 確率分布 $\{p_i\}$
- Aで連続して生じるk個の事象を１つの組み合わせ事象として考えると，次のような情報源を新たにつくることができる

$A^k=\{a_{i_1}a_{i_2}\dots a_{i_k}\}$，確率分布 $\{p_{i_1 i_2 \dots i_k}\}$

Aのk次の拡大情報源

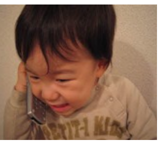
37

例　コイントス

- 情報源
  A＝{ 表，裏 }
- A^2＝{ 表表，表裏，裏表，裏裏 }
- A^3＝{ 表表表，表表裏，表裏表，表裏裏，裏表表，裏表裏，裏裏表，裏裏裏 }

38

例　あいうえお世界



- 情報源
  A＝{ あ，い，う，え，お }
- A^2＝{ ああ，あい，あう，あえ，あお，いあ，いい，いう，いえ，いお，うあ，うい，うう，うえ，うお，えあ，えい，えう，ええ，えお，おあ，おい，おう，おえ，おお}

39

例　サイコロ

- 情報源
  A＝{ 1，2，3，4，5，6 }
- A^2＝{ 11,12,13,14,15,16, 21,22,23,24,25,26, 31,32,33,34,35,36, 41,42,43,44,45,46, 51,52,53,54,55,56, 61,62,63,64,65,66}

40

拡大情報源の持つ性質

- $|A^k| = |A|^k$
  - | | は事象数．例えばコイントスは|A|=2
- $\Sigma\ p_{i_1 i_2 \dots i_k} = 1$
- Aが独立生起情報源ならば（連続して起こる事象が互いに独立ならば），

$p_{i_1 i_2 \dots i_k} = p_{i_1}\ p_{i_2} \dots p_{i_k}$
$H(A^k) = k\ H(A)$

41

小レポート(2)

- Aが独立生起情報源であるとき，

$p_{i_1 i_2 \dots i_k} = p_{i_1}\ p_{i_2} \dots p_{i_k}$

を用いて

$H(A^k) = k\ H(A)$

を証明せよ．

42

拡大情報源の最小平均符号長

- k次の拡大情報源について，それぞれハフマン符号化を行うことができる

→kが増加するにつれて，その最小平均符号長 $L_{min}(k)$ はどうなるか？

→もとの事象１個あたりの最小平均符号長 $L_{min}(k)$ / k はどうなるか？

43

例：変なコイン

- 表，裏の出る確率が{1/4, 3/4} という確率分布を持つ独立生起情報源の2, 3, 4次の拡大情報源を作り，ハフマン符号化してみる．

44

1事象の場合

元の情報源のエントロピー

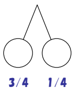
$H = -\Sigma pilogpi$

$= -1/4 \log_2 1/4 - 3/4 \log_2 3/4$

≒ 0.811

(当然1より小さい！)

k = 1



$L_{min}(k)$	1
$L_{min}(k)/k$	1

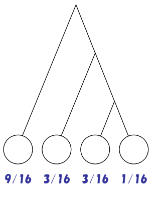
ハフマン符号化⇒平均符号長は 1

45

2事象の場合

p表表=3/4*3/4=9/16
p表裏=3/4*1/4=3/16
p裏表=1/4*3/4=3/16
p裏裏=1/4*1/4=1/16

k = 2



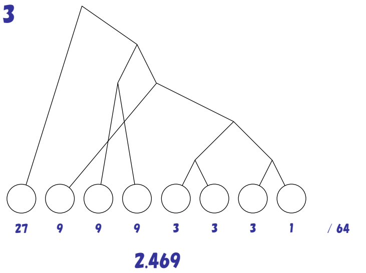
$L_{min}(k)$	1.688
$L_{min}(k)/k$	0.844

ハフマン符号化⇒
元の一文字あたりの平均符号長は0.844

46

3事象の場合

k = 3



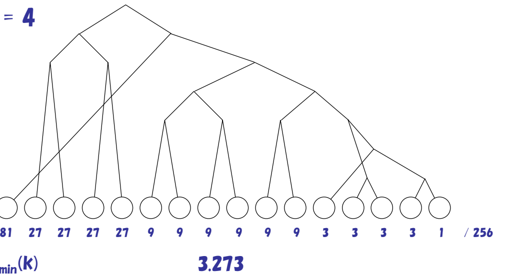
$L_{min}(k)$	2.469
$L_{min}(k)/k$	0.823

ハフマン符号化⇒
元の一文字あたりの平均符号長は0.823

47

元の１事象あたりの符号長が減少！

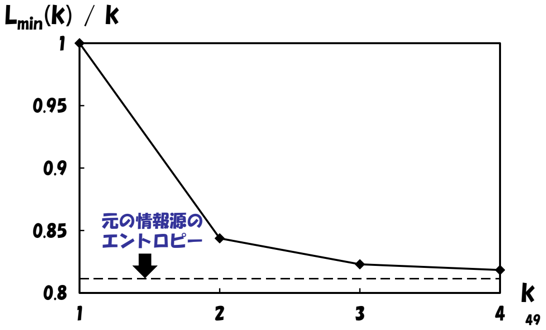
k = 4



$L_{min}(k)$	3.273
$L_{min}(k)/k$	0.818

48

元の１事象あたりの符号長が減少！



49

なぜか？

最小平均符号長の上限と下限

$$H(X) \leqq L_{min} < H(X) + 1$$

のｋ次拡大情報源バージョンを考えると，

$$H(X^k) \leqq L_{min}(k) < H(X^k) + 1$$

となる．

50

なぜか？

$$H(X^k) \leqq L_{min}(k) < H(X^k) + 1$$

全体をｋで割ると，

$$H(X^k)/k \leqq L_{min}(k)/k < H(X^k)/k + \underline{1/k}$$

ｋ→∞のとき，１事象あたりの最小平均符号長は $\lim_{k\to\infty} H(X^k)/k$ に収束する．

51

なぜか？

ｋ→∞のとき，１事象あたりの最小平均符号長は $\lim_{k\to\infty} H(X^k)/k$ に収束する．

ここで，もとの情報源が独立生起情報源であるならば $H(X^k) = k\,H(X)$ であるから，上記の値はｋに関わりなく $H(X)$ である．

ｋを限りなく大きくすることによって１事象あたりの最小平均符号長は $H(X)$ に収束する！

52

シャノンの情報源符号化定理

情報源の中の連続した事象をｋ個まとめて各々に新しい記号を与えたとき（ｋ次の**拡大情報源**を作ったとき），**ｋを限りなく大きくすることによって，情報源のエントロピーに限りなく近い最小平均符号長を達成することができる．**

（証明されました）

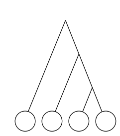
53

情報源符号化定理の直観的理解

ハフマン符号による最小平均符号長は多くの場合情報エントロピーに到達できない

⬇ その訳は

符号長 L_i が **離散的**な値しか取れないから



$$L_i \in \{1, 2, 3\}$$

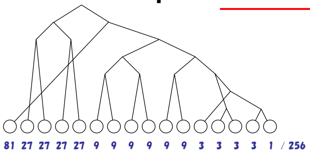
54

情報源符号化定理の直観的理解

しかし，

高次の拡大情報源をつくることによって符号長 L_i の取りうる値の選択肢が広がる！

$L_i \in \{1, 2, \dots, 15\}$



より連続変数っぽくなる

⬇

L_i を連続変数 x_i に置き換えて求めた理論値に近づく！

エントロピーの一般化

- 連続する事象の間に**何らかの関連がある**（独立生起情報源でない）場合は，情報源のエントロピーは $L_{min}(k)/k$ の**収束値**によって逆に定義される

$$\overline{H}(X) = \lim_{k\to\infty} L_{min}(k)/k$$

$-\Sigma\, p_i \log p_i$ で計算される通常の情報エントロピーではないので区別して $\overline{H}$ と表記するが，意味するものは同じ

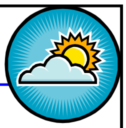
1事象の表現に最低限必要なビット数

従って．．．

情報源の中の連続した事象をk個まとめて各々に新しい記号を与えたとき（k次の拡大情報源を作ったとき），kを限りなく大きくすることによって，情報源のエントロピーに限りなく近い最小平均符号長を達成することができる．

（は，任意の情報源について成り立つと言える）

例　天気



- 天気は典型的な非独立生起情報源．

（前日が晴れなら，次の日も晴れの確率は高い）

  - P雨=8/30, P晴=22/30
  - P雨雨=3/15, P雨晴=1/15, P晴雨=1/15, P晴晴=10/15
  - H(X)=−8/30log(8/30)−22/30log(22/30)=**0.8366**
  - H(X²)=(−3/15*log2(3/15)−1/15*log2(1/15)*2−10/15*log2(10/15))/2=**0.6876**

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
雨	雨	晴	晴	雨	雨	晴	晴	晴	晴	晴	晴	晴	晴	晴	晴
17	18	19	20	21	22	23	24	25	26	27	28	29	30		
晴	晴	晴	晴	晴	雨	雨	雨	晴	晴	晴	晴	雨	晴		

10月の天気

小レポート（3）

- サイコロの目は典型的な独立生起情報源であり，各事象の生起確率は一律に1/6 である．これをk次に拡大してハフマン符号化した場合の1事象あたりの最小平均符号長を，kの関数で表せ．
- 元の一事象あたりの最小平均符号長の収束値を求めよ．

（ヒント：ハフマン符号化しているので完全プレフィックス符号となり，Kraft の不等式が等式になる．これを利用して符号木の形を求めればよい）

付録：底が2の対数表(1/2)

x	log2x	x	log2x	x	log2x	x	log2x	x	log2x	x	log2x	x	log2x	x	log2x
0.01	−6.64	0.26	−1.94	0.51	−0.97	0.76	−0.40	1.1	0.14	3.6	1.85	6.1	2.61	8.6	3.10
0.02	−5.64	0.27	−1.89	0.52	−0.94	0.77	−0.38	1.2	0.26	3.7	1.89	6.2	2.63	8.7	3.12
0.03	−5.06	0.28	−1.84	0.53	−0.92	0.78	−0.36	1.3	0.38	3.8	1.93	6.3	2.66	8.8	3.14
0.04	−4.64	0.29	−1.79	0.54	−0.89	0.79	−0.34	1.4	0.49	3.9	1.96	6.4	2.68	8.9	3.15
0.05	−4.32	0.30	−1.74	0.55	−0.86	0.80	−0.32	1.5	0.58	4.0	2.00	6.5	2.70	9.0	3.17
0.06	−4.06	0.31	−1.69	0.56	−0.84	0.81	−0.30	1.6	0.68	4.1	2.04	6.6	2.72	9.1	3.19
0.07	−3.84	0.32	−1.64	0.57	−0.81	0.82	−0.29	1.7	0.77	4.2	2.07	6.7	2.74	9.2	3.20
0.08	−3.64	0.33	−1.60	0.58	−0.79	0.83	−0.27	1.8	0.85	4.3	2.10	6.8	2.77	9.3	3.22
0.09	−3.47	0.34	−1.56	0.59	−0.76	0.84	−0.25	1.9	0.93	4.4	2.14	6.9	2.79	9.4	3.23
0.10	−3.32	0.35	−1.51	0.60	−0.74	0.85	−0.23	2.0	1.00	4.5	2.17	7.0	2.81	9.5	3.25
0.11	−3.18	0.36	−1.47	0.61	−0.71	0.86	−0.22	2.1	1.07	4.6	2.20	7.1	2.83	9.6	3.26
0.12	−3.06	0.37	−1.43	0.62	−0.69	0.87	−0.20	2.2	1.14	4.7	2.23	7.2	2.85	9.7	3.28
0.13	−2.94	0.38	−1.40	0.63	−0.67	0.88	−0.18	2.3	1.20	4.8	2.26	7.3	2.87	9.8	3.29
0.14	−2.84	0.39	−1.36	0.64	−0.64	0.89	−0.17	2.4	1.26	4.9	2.29	7.4	2.89	9.9	3.31
0.15	−2.74	0.40	−1.32	0.65	−0.62	0.90	−0.15	2.5	1.32	5.0	2.32	7.5	2.91	10.0	3.32
0.16	−2.64	0.41	−1.29	0.66	−0.60	0.91	−0.14	2.6	1.38	5.1	2.35	7.6	2.93		
0.17	−2.56	0.42	−1.25	0.67	−0.58	0.92	−0.12	2.7	1.43	5.2	2.38	7.7	2.94		
0.18	−2.47	0.43	−1.22	0.68	−0.56	0.93	−0.10	2.8	1.49	5.3	2.41	7.8	2.96		
0.19	−2.40	0.44	−1.18	0.69	−0.54	0.94	−0.09	2.9	1.54	5.4	2.43	7.9	2.98		
0.20	−2.32	0.45	−1.15	0.70	−0.51	0.95	−0.07	3.0	1.58	5.5	2.46	8.0	3.00		
0.21	−2.25	0.46	−1.12	0.71	−0.49	0.96	−0.06	3.1	1.63	5.6	2.49	8.1	3.02		
0.22	−2.18	0.47	−1.09	0.72	−0.47	0.97	−0.04	3.2	1.68	5.7	2.51	8.2	3.04		
0.23	−2.12	0.48	−1.06	0.73	−0.45	0.98	−0.03	3.3	1.72	5.8	2.54	8.3	3.05		
0.24	−2.06	0.49	−1.03	0.74	−0.43	0.99	−0.01	3.4	1.77	5.9	2.56	8.4	3.07		60
0.25	−2.00	0.50	−1.00	0.75	−0.42	1.00	0.00	3.5	1.81	6.0	2.58	8.5	3.09		

付録：底が２の対数表(2/2)

x	$\log_2 x$	x	$\log_2 x$	x	$\log_2 x$	x	$\log_2 x$
11	3.46	36	5.17	61	5.93	86	6.43
12	3.58	37	5.21	62	5.95	87	6.44
13	3.70	38	5.25	63	5.98	88	6.46
14	3.81	39	5.29	64	6.00	89	6.48
15	3.91	40	5.32	65	6.02	90	6.49
16	4.00	41	5.36	66	6.04	91	6.51
17	4.09	42	5.39	67	6.07	92	6.52
18	4.17	43	5.43	68	6.09	93	6.54
19	4.25	44	5.46	69	6.11	94	6.55
20	4.32	45	5.49	70	6.13	95	6.57
21	4.39	46	5.52	71	6.15	96	6.58
22	4.46	47	5.55	72	6.17	97	6.60
23	4.52	48	5.58	73	6.19	98	6.61
24	4.58	49	5.61	74	6.21	99	6.63
25	4.64	50	5.64	75	6.23	100	6.64
26	4.70	51	5.67	76	6.25		
27	4.75	52	5.70	77	6.27		
28	4.81	53	5.73	78	6.29		
29	4.86	54	5.75	79	6.30		
30	4.91	55	5.78	80	6.32		
31	4.95	56	5.81	81	6.34		
32	5.00	57	5.83	82	6.36		
33	5.04	58	5.86	83	6.38		
34	5.09	59	5.88	84	6.39		
35	5.13	60	5.91	85	6.41		