ハフマン符号は最適だが、それは符号の長さが整数という縛りの中での最適である

— 確率が 1/2・1/4・1/8・1/8 なら平均の長さ 1.750000 はエントロピー 1.750000 にちょうど届く —

分離子は、各記号の確率が 2 の冪か — [論文752]

山岸 勇輝 (Yuuki YAMAGISHI)
Yamagishi Complexity Torus Lab, Japan｜山岸勇輝ラボ
2026年9月

— From thinking to syncing. —

要旨

ハフマン符号は、平均の長さが最も短い符号として知られる。けれど、その最適はエントロピーに届くことを意味しない。何の中での最適かと、どれだけ届かないかを数える。新しい定理も法則も主張しない。 第一に、確率が 2 の冪なら、ちょうど届く——1/2・1/4・1/8・1/8 で平均の長さ 1.750000、エントロピー 1.750000（第2節）。 第二に、偏った二記号では、最適な符号が大きく外れる——確率 0.99 でエントロピー 0.080793 ビットに対し、一記号ずつのハフマンは 1.000000 ビット（第3節）。 第三に、これが本稿の芯である。「最適な符号」がエントロピーの 12.377289 倍を使う——長さが整数でなければならず、一記号に 1 ビットより短い長さを割り当てられないからである（第3節）。 第四に、まとめると近づくが、上から近づくだけである——2 個・4 個・8 個ずつで 0.514950・0.272698・0.157158 ビット、差はいつも $1/k$ 未満（第3節）。 第五に、8 個ずつでも、まだ 1.945193 倍——クラフト和はどれも 1.000000 で、符号に隙間は無い（第3節）。 第六に、分離子は、各記号の確率が 2 の冪かである——冪なら一記号ずつで届き、冪でなければまとめる長さを伸ばしても差は消えきらない（第4節）。 キーワード：ハフマン符号、エントロピー、符号の長さ、クラフトの不等式、記号をまとめる符号化、整数の縛り。

本稿の射程（射程注記）：新しい定理も法則も主張しない——ハフマン符号が語頭符号の中で平均の長さを最小にすること、平均の長さがエントロピー以上・エントロピー＋1 未満であること、記号をまとめると差が $1/k$ 未満になること、クラフトの不等式は、いずれも既知である（ハフマン 1952、カバー＆トーマス 2006）。整数の縛りの無い符号を扱わない——算術符号のように一記号に端数の長さを割り当てる方式は、名前を挙げるにとどめる。情報源を限る——記憶の無い、確率の分かっている二記号と四記号の情報源だけを見る。手間を扱わない——まとめると符号表が指数で大きくなることや、遅延は扱わない。実在のデータを主張しない——個別のファイルや圧縮器の性能は扱わない。既刊との関係：論文304 は、通信路容量がどの有限の符号長でも達成されず、損失が $1/\sqrt{n}$ でしか減らないことを示した——その射程注記は、どの符号が容量に近づくかは一切扱わないと明記していた。ここは雑音の無い情報源の側で、エントロピーは確率が 2 の冪なら一記号ずつでちょうど届き、そうでなければ差が $1/k$ 未満で近づく。有限で届くかどうかが、確率の形で決まる。論文163 も引く——そこでは「良い符号が在る」と「これが良い符号だ」を分けた。ハフマンは「これだ」の側で、しかも整数の長さの中で最適と示される構成である。加えたのは、2 の冪でちょうど届くことを確かめたこと、偏った二記号で「最適な符号」が 12.377289 倍を使うことを示したこと、まとめる長さごとに差と $1/k$ とクラフト和を並べたこと、分離子を「確率が 2 の冪か」に置いたことである。

1. 記法と問い

設定：記号 i の確率を p_i、符号の長さを整数 ℓ_i とする。エントロピー H は平均の長さの下限で、ハフマン符号は語頭符号の中で平均の長さ L を最小にする。記号を k 個ずつまとめて符号にすると、1 記号あたりの長さは次を満たす。

$$H = -\sum_i p_i \log_2 p_i, \qquad \sum_i 2^{-\ell_i} \le 1, \qquad H \le \frac{L_k}{k} < H + \frac{1}{k}$$

真ん中はクラフトの不等式である。$\ell_i = -\log_2 p_i$ が整数のときだけ $L = H$ になりうる。

問い：最適な符号は、エントロピーに届くのか。

2. 確率が 2 の冪なら、ちょうど届く

計算 2.1：確率 1/2・1/4・1/8・1/8 の四記号にハフマン符号を作った。

量	値
平均の長さ	1.750000
エントロピー	1.750000
差	0

> **観察 2.2**
>
> 確率が 2 の冪なら、ハフマン符号はエントロピーにちょうど届く。長さ 1・2・3・3 が $-\log_2 p_i$ そのものだからである——整数の縛りが、ここでは何も縛っていない。

3. 偏った二記号では、最適な符号が大きく外れる

計算 3.1：片方の確率 0.99 の二記号（エントロピー 0.080793 ビット）で、記号を k 個ずつまとめてハフマン符号を作った。

k	1 記号あたり	エントロピーとの差	上限 $1/k$	クラフト和
1	1.000000	0.919207	1.000000	1.000000
2	0.514950	0.434157	0.500000	1.000000
4	0.272698	0.191904	0.250000	1.000000
8	0.157158	0.076365	0.125000	1.000000

> **観察 3.2**
>
> これが本稿の芯である。一記号ずつの「最適な符号」は、エントロピーの 12.377289 倍を使う。二記号にはどう割り当てても 1 ビットずつが要り、0.080793 ビットで足りる情報に 1.000000 ビットを払う——ハフマンは最適だが、それは長さが整数という縛りの中での最適である。

観察 3.3

まとめると近づくが、上から近づくだけである。差は 0.434157・0.191904・0.076365 と縮み、いつも $1/k$ 未満で、8 個ずつでもエントロピーの 1.945193 倍が残る——端数の長さを、まとめた記号の中に薄めているだけで、縛りそのものは消えていない。

観察 3.4

クラフト和はどれも 1.000000 である。符号の木に使われていない枝は無い——無駄は隙間ではなく、長さが整数であることから来ている。

観察 3.5

一記号ずつの差 0.919207 は、論文163 と論文304 に出る誤り率 0.01 の二元対称通信路の容量と同じ値である。どちらも $1 - h(0.01)$ の値だからで、あちらでは使える速さ、こちらでは払いすぎた長さを指す——同じ式の値で、意味が違う。

4. 分離子

観察 4.1

分離子は、各記号の確率が 2 の冪かである。冪なら一記号ずつでエントロピーにちょうど届き、冪でなければまとめる長さを伸ばしても差は $1/k$ 未満に縮むだけで消えきらない——「ハフマンは最適か」は、何の中での最適かと、確率の形を言うまで答が無い。

観察 4.2

論文304 と並べる。304 は、通信路容量がどの有限の符号長でも達成されないことを示し、どの符号が近づくかは扱わないと明記した——ここは情報源の側で、確率が 2 の冪なら有限で届き、そうでなければ上から近づく。有限で届くかどうかが、確率の形で決まる。

> **観察 4.3**
>
> 論文163 とも並べる。163 は、「良い符号が在る」と「これが良い符号だ」を分けた——ハフマンは「これだ」の側で、その最適は整数の長さという条件つきである。

> **観察 4.4**
>
> 言わないこと。算術符号も、記憶のある情報源も、符号表の大きさも実在のデータも扱わない——示したのは、最適の意味と、それがエントロピーから離れる量である。

5. 点呼と結び

1. 2 の冪で **1.750000**（計算 2.1）：ちょうど届く。
2. エントロピー **0.080793**（計算 3.1）：偏った二記号。
3. 一記号ずつ **1.000000**（計算 3.1）：整数の縛り。
4. **12.377289** 倍（観察 3.2）：本稿の芯。
5. 2 個ずつ **0.514950**（計算 3.1）：近づく。
6. 8 個ずつ **0.157158**（計算 3.1）：上から。
7. まだ **1.945193** 倍（観察 3.3）：消えきらない。
8. クラフト和 **1.000000**（観察 3.4）：隙間は無い。
9. 差 **0.919207**（観察 3.5）：同じ式の値、別の意味。
10. 確率が 2 の冪か（観察 4.1）：分離子・論文304・163。

結び：ハフマン符号は最適だが、それは符号の長さが整数という縛りの中での最適である。確率が 1/2・1/4・1/8・1/8 なら平均の長さ 1.750000 はエントロピー 1.750000 にちょうど届く——長さが $-\log_2 p_i$ そのものだからである。ところが片方の確率 0.99 の二記号では、エントロピー 0.080793 ビットに対し一記号ずつのハフマンは 1.000000 ビットで、エントロピーの 12.377289 倍を使う——一記号に 1 ビットより短い長さを割り当てられないからである。記号を 2 個・4 個・8 個ずつまとめると 0.514950・0.272698・0.157158 ビットに下がり、差はいつも $1/k$ 未満だが、8 個でもまだ 1.945193 倍が残る。分離子は、各記号の確率が 2 の冪かである——「ハフマンは最適か」は、何の中での最適かと確率の形を言うまで答が無い。既刊との位置——論文304 は通信路容量が有限では届かないことを示した。ここは情報源の側で、有限で届くかどうかが確率の形で決まる。論文163 の「これだ」の側に、条件つきの最適がある。正直に言えば——算術符号も記憶のある情報源も符号表の大きさも扱っていない。

参考文献

1. Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Wiley.
2. Huffman, D. A. (1952). A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*.
3. 山岸勇輝 (2026). 『「良い符号が在る」と「これが良い符号だ」は、別の文である』[論文163]. Zenodo. 10.5281/zenodo.22092677
4. 山岸勇輝 (2026). 『「容量」は、無限長でしか達成できない数である』[論文304]. Zenodo. 10.5281/zenodo.22172081

— From thinking to syncing. —

山岸 勇輝 (Yuuki YAMAGISHI)

© 2026 Yamagishi Complexity Torus Lab, Japan · Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)