情報理論と符号理論

2.6 情報源の拡大(p.30)

拡大情報源

- 今までは情報源Sのシンボル$s_1, \cdots, s_q$毎に符号を割り当てていた
- 符号を$s_1, \cdots, s_q$をn個つなげた$s_{i_1}, \cdots, s_{i_n}$をアルファベットを情報源とするS^nを用意
- このS^nのアルファベット$s_{i_1}, \cdots, s_{i_n}$に符号を割り当てることを考える
- このS^nをSのn次の**拡大情報源(extended source)**という

拡大情報源の性質

- Sがq個のシンボルからなっているので、S^nはq^n個のアルファベットをもつ
  - $s_{i_1},\cdots,s_{i_n}$のn個の連続するシンボルをアルファベットをもつため
- S^nのアルファベットの生起確率は$p_{i_1},\cdots,p_{i_n}$をかけた$p_{i_1}\cdots p_{i_n}$になる
  - Sのシンボル$s_1,\ldots,s_q$が$i_1,\ldots,i_n$番目に独立に表れるため

拡大情報源の性質

- $p_{i_1}, \cdots, p_{i_n}$をかけた$p_{i_1} \cdots p_{i_n}$は確率分布となる
  - $(p_1 + \cdots + p_q)^n = 1^n = 1$
  - $(p_1 + \cdots + p_q)^n$を展開すると$p_{i_1} \cdots p_{i_n}$は一度しか表れない、つまり

$$(p_1 + \cdots + p_q)^n = \sum_{(i_1,\ldots,i_n)\in\{1,\ldots,q\}^n} p_{i_1} \cdots p_{i_n}$$

    + 多項定理からわかる

拡大情報源の例

- $S = \{s_1, s_2\}$としてs_1の生起確率を${}^2/_3$、s_2の生起確率を${}^1/_3$とする
- $S^2 = \{s_1 s_1, s_1 s_2, s_2 s_1, s_2 s_2\}$
  - $s_1 s_1$の生起確率: ${}^4/_9$
  - $s_1 s_2$の生起確率: ${}^2/_9$
  - $s_2 s_1$の生起確率: ${}^2/_9$
  - $s_2 s_2$の生起確率: ${}^1/_9$

なぜ拡大情報源を考えるのか？

- $p_{\max}$ をSの生起確率で最大のもの、$p_{\min}$ を最小のものとする
  - ただし、$p_{\max} > p_{\min}$(一様分布とならないとする)
- $\frac{p_{\max}}{p_{\min}} > 1$から$\frac{p_{\max}^n}{p_{\min}^n} = \left(\frac{p_{\max}}{p_{\min}}\right)^n \to \infty (n \to \infty)$
- すなわち、nが増大するとS^nの確率の変動が大きくなる

なぜ拡大情報源を考えるのか？

- 確率の変動が大きいと平均符号長が短かくなる
  - 2.2 2元ハフマン符号(p. 26)
- ざっくり言うと出現頻度が高いものに短い符号長を割り当てて低いものに長い符号長を割り当てるため

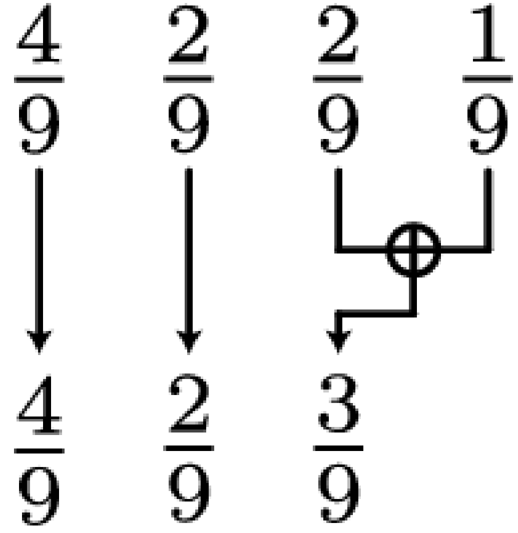
拡大情報源による符号長変化の例

- $S = \{s_1, s_2\}$としてs_1の生起確率を$^2/_3$、s_2の生起確率を$^1/_3$とする
- ハフマン符号を考えると$s_1 \mapsto 0, s_2 \mapsto 1$から平均符号長は$1\left(\frac{2}{3}+\frac{1}{3}\right) = 1$
- $S^2 = \{s_1s_1, s_1s_2, s_2s_1, s_2s_2\}$で確率はそれぞれ$\frac{4}{9}, \frac{2}{9}, \frac{2}{9}, \frac{1}{9}$

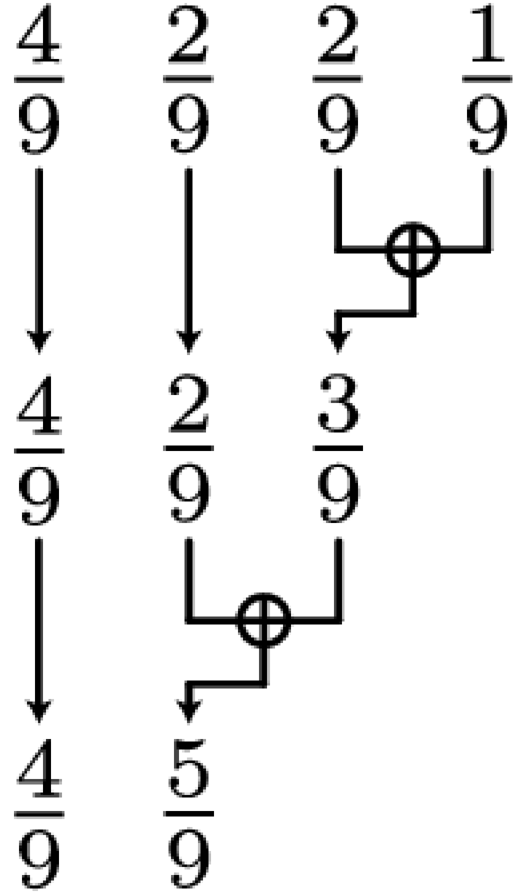
拡大情報源による符号長変化の例

$$\frac{4}{9} \quad \frac{2}{9} \quad \frac{2}{9} \quad \frac{1}{9}$$

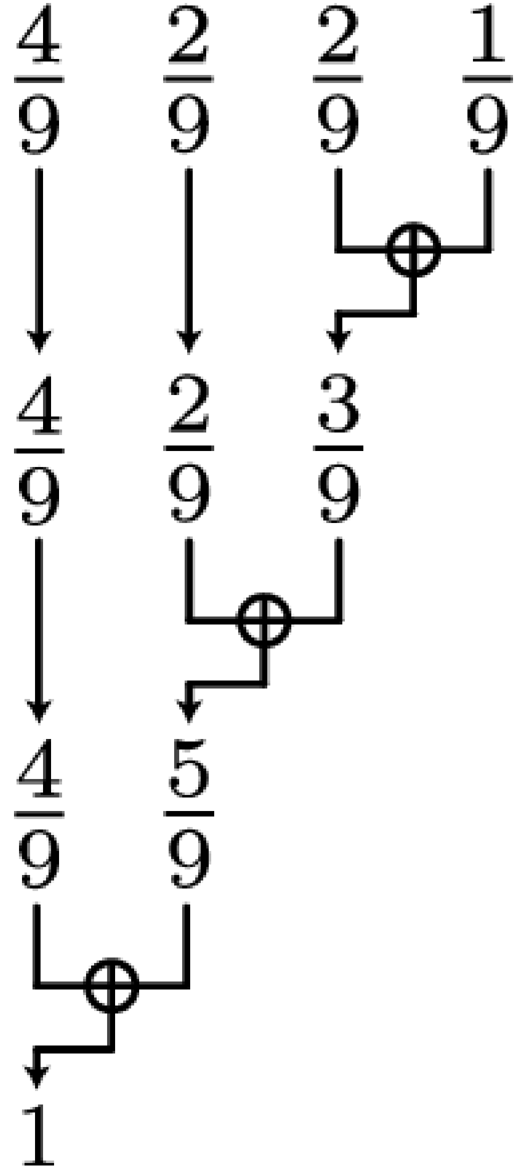
拡大情報源による符号長変化の例



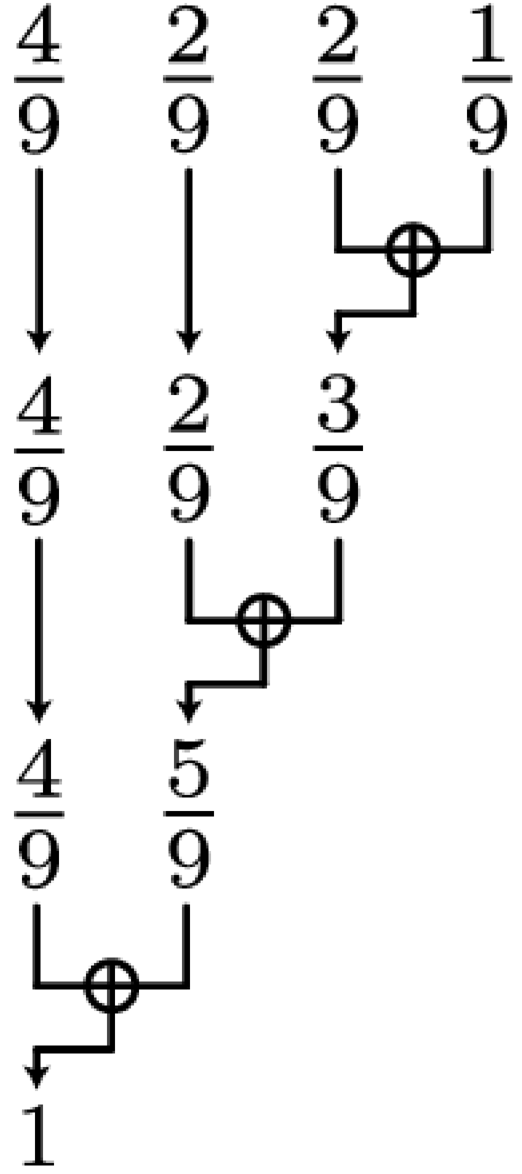
拡大情報源による符号長変化の例



拡大情報源による符号長変化の例

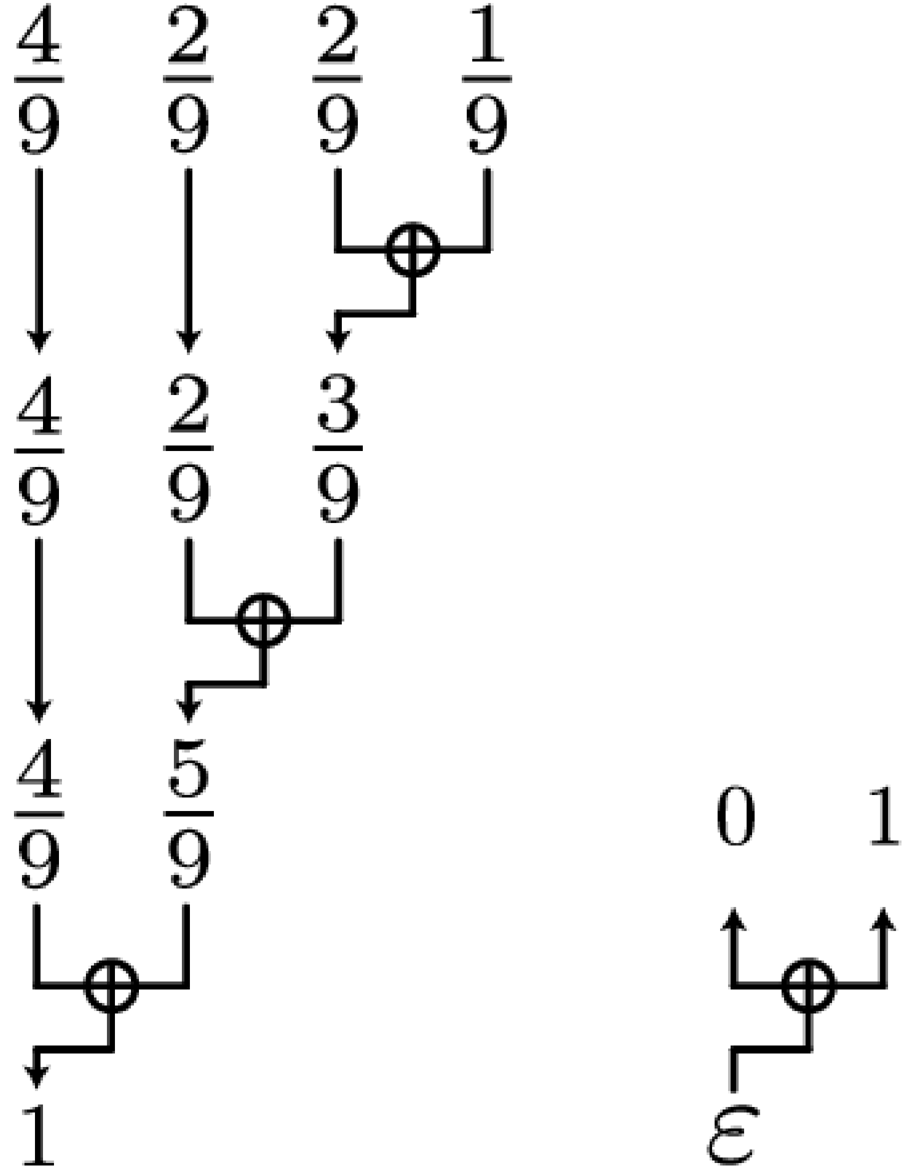


拡大情報源による符号長変化の例

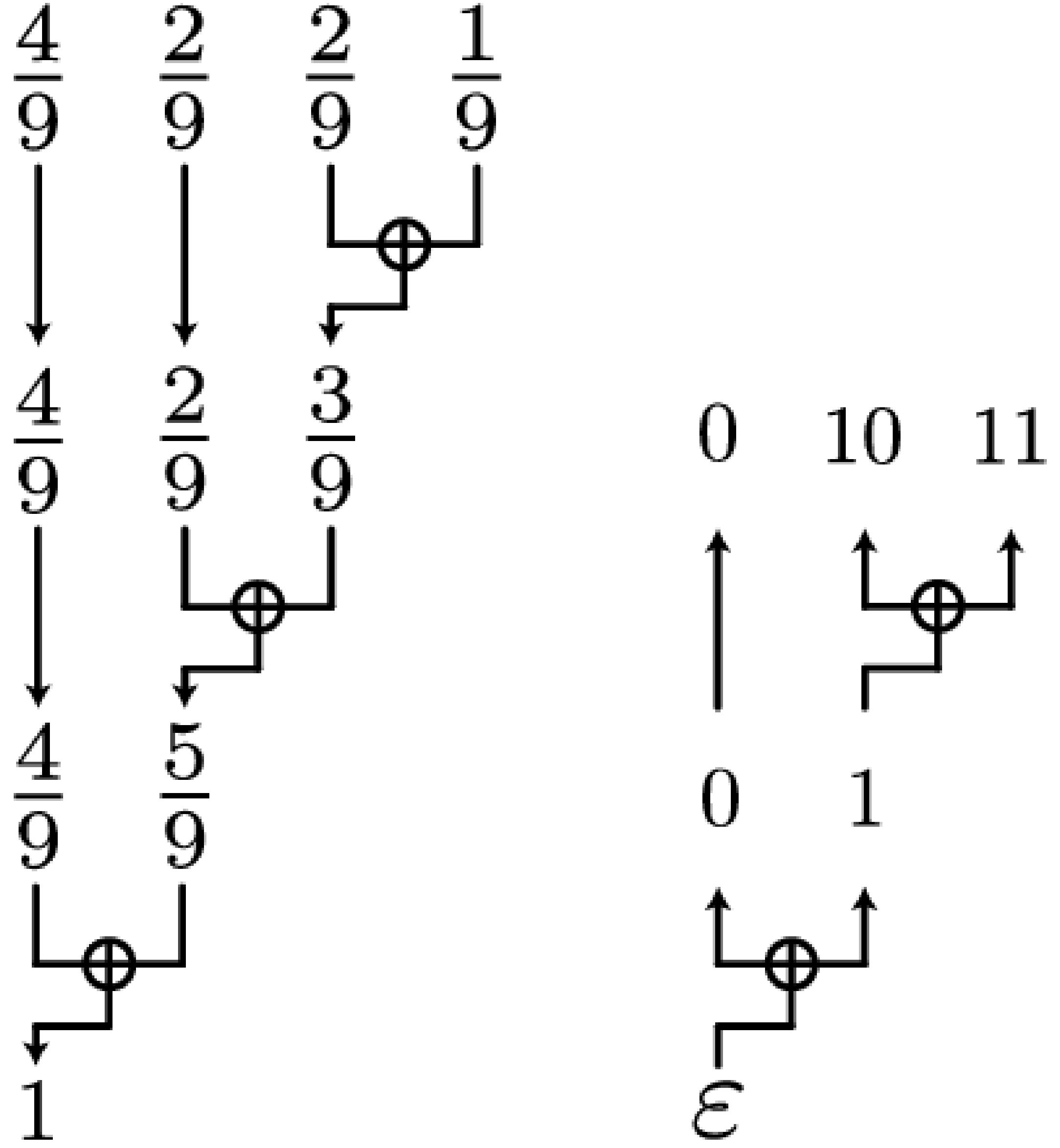


$\mathcal{E}$

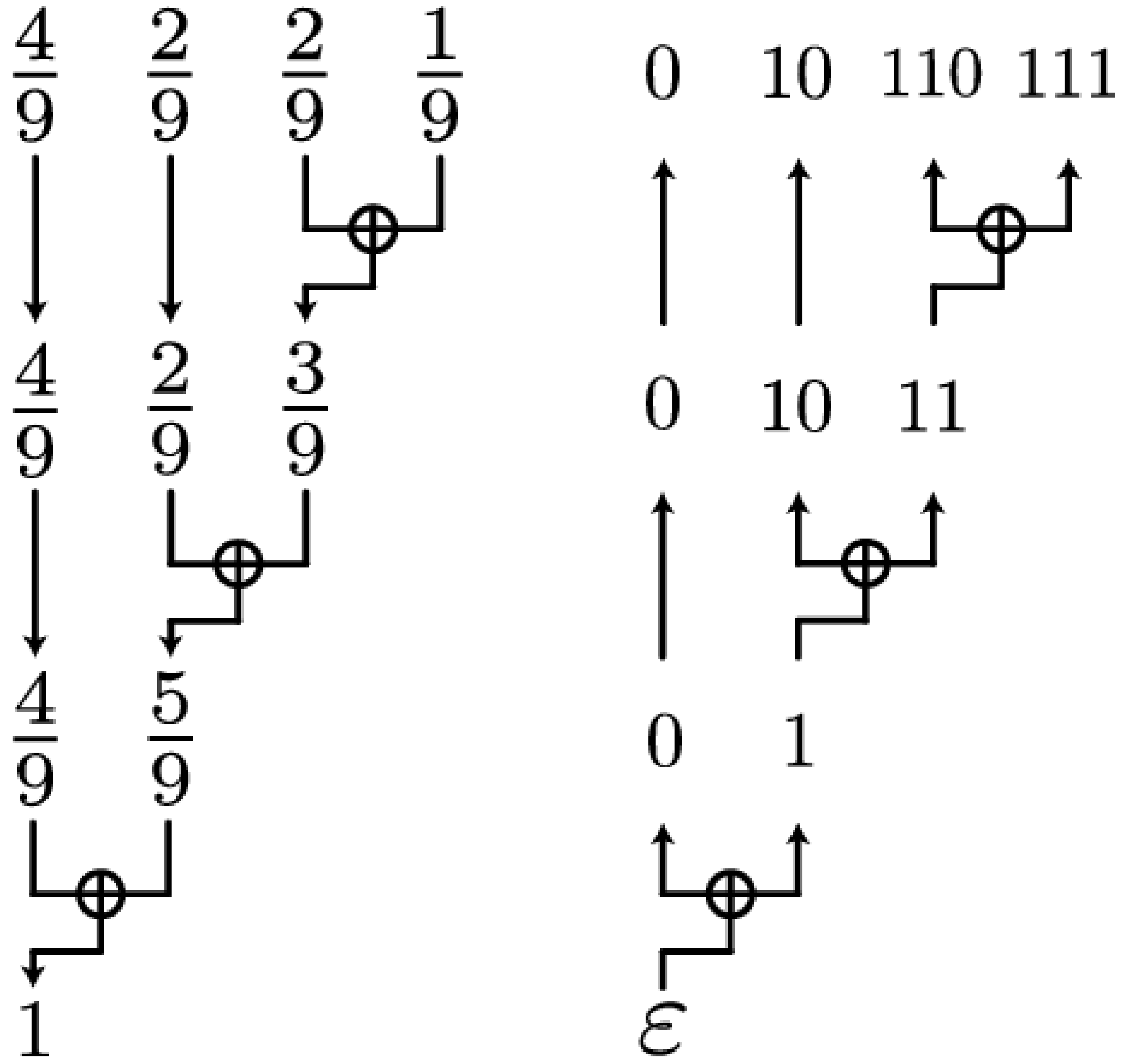
拡大情報源による符号長変化の例



拡大情報源による符号長変化の例



拡大情報源による符号長変化の例



拡大情報源による符号長変化の例

- ハフマン符号は
  - $s_1 s_1 \mapsto 0$
  - $s_1 s_2 \mapsto 10$
  - $s_1 s_2 \mapsto 110$
  - $s_1 s_2 \mapsto 111$
- 平均符号長は

$$\frac{4}{9} + 2 \cdot \frac{2}{9} + 3 \cdot \frac{2}{9} + 3 \cdot \frac{1}{9} = \frac{17}{9}$$

拡大情報源の例のまとめ

- S^2はSの2個のシンボルからなるブロックで構成されている+S^2の平均符号長は$\frac{17}{9}$
- Sの各シンボルに必要な符号化長は平均$\frac{17}{18}$

拡大情報源の例のまとめ

- Sに対するハフマン符号の平均符号長は1
- S^2に対するハフマン符号の平均符号長は$\frac{17}{9}$で、S^2が2個のSのシンボルからなるのでSの平均符号長は$\frac{17}{18} = 0.944 \cdots < 1$
- Sに対するハフマン符号よりも短い符号(?)を達成できた

拡大情報源の例のまとめ

- 厳密に言えばS^2から求めたものは符号<u>ではない</u>
  - Sに対して符号を割り当てているわけではない
- しかし、ここではS^2に対する符号化もSの符号化と呼ぶ
- S^2に対するハフマン符号は一意なので任意の符号列をさらに分解してSに対する一意な符号をつくれる
  - 連続な組によって復号しなければならないため、瞬時に復号できるとは限らない(符号語が2つ以上割り当てられる可能性がある)

拡大情報源のまとめ

SからS^3をつくると更に符号長を短くできる
(実際に確かめると$\frac{76}{81} = 0.938\cdots < \frac{17}{18} = 0.944\cdots$)

拡大情報源のまとめ

S^3は27個の要素を持つので手で確かめる

拡大情報源のまとめ

SからS^3をつくると更に符号長を短くできる
(実際に確かめると$\frac{76}{81} = 0.938\cdots < \frac{17}{18} = 0.944\cdots$)

$\Downarrow$

一般化してS^nに対しても
成立すると予想できる

拡大情報源のまとめ

- 一般化するにあたって、2つの疑問がでてくる
  - $n \to \infty$としたときのSの平均符号長$\frac{L_n}{n}$はどうなる?
    - L_n: S^nの平均符号長
  - 同じ手法で他の情報源に対しても
    同様のことが言える?
- これは次に出てくるエントロピーを
  用いると議論ができる

演習問題2.9

$S=\{s_1,s_2\}$とし、s_1とs_2の生起確率はそれぞれ${}^2/_3$, ${}^1/_3$とする。
S^3に対する確率分布を求めよ。また、S^3に対する2元ハフマン符号C^3の平均符号長が${}^{76}/_{27}$であることを確認する

先程求めたので割愛

演習問題2.10

Sを生起確率が0.3,0.3,0.2,0.2の情報源とする。
Sの最適な2元符号はいくつ存在するか？
それらはハフマン符号か？

演習問題2.10

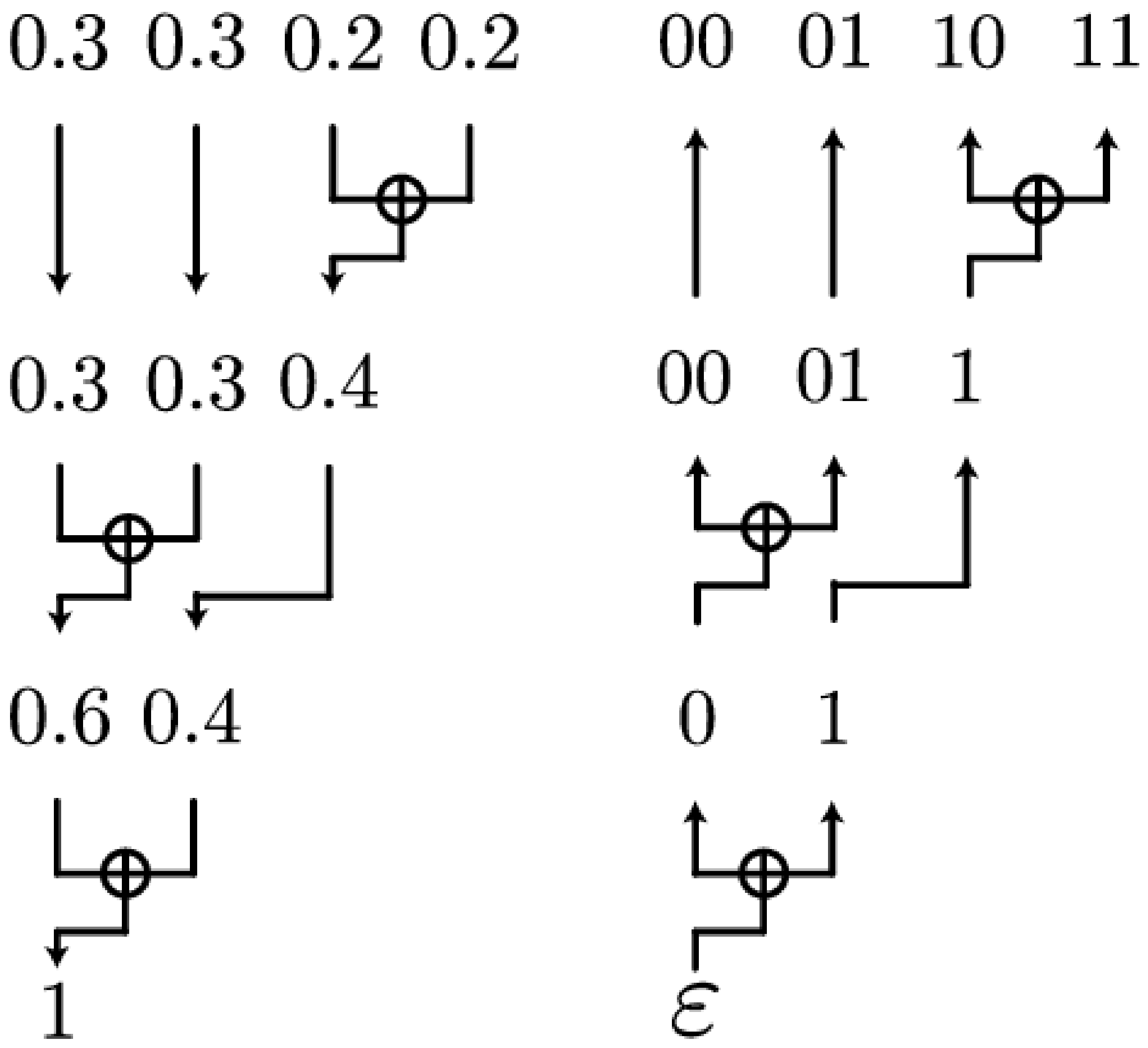
Sを生起確率が0.3,0.3,0.2,0.2の情報源とする。
Sの最適な2元符号はいくつ存在するか？
それらはハフマン符号か？

まずは最適な2元符号の数を数える

演習問題2.10

- 2元ハフマン符号を1つ書き出してみる

演習問題2.10



演習問題2.10

- 2元ハフマン符号を1つ書き出してみる
- すると符号$\{00,01,10,11\}$を得る

演習問題2.10

- 2元ハフマン符号を1つ書き出してみる
- すると符号{00,01,10,11}を得る
- つまり、Sの最適な符号は00,01,10,11の4つから構成される

演習問題2.10

Sを生起確率が0.3,0.3,0.2,0.2の情報源とする。
Sの最適な2元符号はいくつ存在するか？
それらはハフマン符号か？

次にハフマン符号の数を数える

演習問題2.10

- $00, 01, 10, 11$の置換は$4! = 24$個で、最適な2元符号は24個

演習問題2.10

- $00,01,10,11$の置換は$4! = 24$個で、最適な2元符号は24個
- うちハフマン符号になるものは分岐の組み合わせ分$2^3 = 8$個

演習問題2.11

$S = \{s_1, \cdots s_q\}$とし、生起確率は各々$p_1 \geq \cdots \geq p_q$とし、
$p_i > p_{i+2} + \cdots + p_q$を満たす。$S$に対する任意の
2元ハフマン符号の符号長が$1, 2, \cdots, q-1, q-1$
となることを示せ。また、Sに対する異なる
2元ハフマン符号はいくつあるか？さらに各$q \geq 1$に対し、
上の不等式を満たす確率分布の例を構成せよ

演習問題2.11

$S = \{s_1, \cdots s_q\}$とし、生起確率は各々$p_1 \geq \cdots \geq p_q$とし、
$p_i > p_{i+2} + \cdots + p_q$を満たす。$S$に対する任意の
2元ハフマン符号の符号長が$1,2,\cdots,q-1,q-1$
となることを示せ。また、Sに対する異なる
2元ハフマン符号はいくつあるか？さらに各$q \geq 1$に対し、
上の不等式を満たす確率分布の例を構成せよ

まずは2元ハフマン符号の符号長が
$1,2,\cdots,q-1,q-1$となることを示す

演習問題2.11

- $p_i > p_{i+2} + \cdots + p_q$という条件に着目してハフマン符号のプロセスを実行すると、$p_{q-3} > p_{q-2} + p_{q-1} + p_q$から$s_{q-1}$と$s_q$を縮退した$s^{(1)}$はさらに$s_{q-2}$と縮退され、$s^{(2)}$をつくる

演習問題2.11

- $p_i > p_{i+2} + \cdots + p_q$という条件に着目してハフマン符号のプロセスを実行すると、$p_{q-3} > p_{q-2} + p_{q-1} + p_q$から$s_{q-1}$と$s_q$を縮退した$s^{(1)}$はさらに$s_{q-2}$と縮退され、$s^{(2)}$をつくる
- 繰り返せば$s^{(k)} = s_{q-k} \vee s^{(k-1)} = s_{q-k} \vee \cdots \vee s_q$

演習問題2.11

- $p_i > p_{i+2} + \cdots + p_q$という条件に着目してハフマン符号のプロセスを実行すると、$p_{q-3} > p_{q-2} + p_{q-1} + p_q$から$s_{q-1}$と$s_q$を縮退した$s^{(1)}$はさらに$s_{q-2}$と縮退され、$s^{(2)}$をつくる
- 繰り返せば$s^{(k)} = s_{q-k} \vee s^{(k-1)} = s_{q-k} \vee \cdots \vee s_q$
- よって$i \leq q-1$に対してi回縮退されていることがわかる

演習問題2.11

- $p_i > p_{i+2} + \cdots + p_q$という条件に着目してハフマン符号のプロセスを実行すると、$p_{q-3} > p_{q-2} + p_{q-1} + p_q$から$s_{q-1}$と$s_q$を縮退した$s^{(1)}$はさらに$s_{q-2}$と縮退され、$s^{(2)}$をつくる
- 繰り返せば$s^{(k)} = s_{q-k} \vee s^{(k-1)} = s_{q-k} \vee \cdots \vee s_q$
- よって$i \leq q-1$に対してi回縮退されていることがわかる
- 最後にqは$q-1$と兄弟になるため符号長は$q-1$

演習問題2.11

- $p_i > p_{i+2} + \cdots + p_q$という条件に着目してハフマン符号のプロセスを実行すると、$p_{q-3} > p_{q-2} + p_{q-1} + p_q$から$s_{q-1}$と$s_q$を縮退した$s^{(1)}$はさらに$s_{q-2}$と縮退され、$s^{(2)}$をつくる
- 繰り返せば$s^{(k)} = s_{q-k} \vee s^{(k-1)} = s_{q-k} \vee \cdots \vee s_q$
- よって$i \leq q-1$に対してi回縮退されていることがわかる
- 最後にqは$q-1$と兄弟になるため符号長は$q-1$
- 以上から2元ハフマン符号の符号長が$1,2,\cdots,q-1,q-1$となることが示せた

演習問題2.11

$S=\{s_1,\cdots s_q\}$とし、生起確率は各々$p_1 \geq \cdots \geq p_q$とし、
$p_i > p_{i+2}+\cdots+p_q$を満たす。Sに対する任意の
2元ハフマン符号の符号長が$1,2,\cdots,q-1,q-1$
となることを示せ。また、Sに対する異なる
2元ハフマン符号はいくつあるか？さらに各$q \geq 1$に対し、
上の不等式を満たす確率分布の例を構成せよ

次にSに対する異なる2元ハフマン符号の数を求める

演習問題2.11

- 2元ハフマン符号は構成上、分岐の部分に
  2通りの選択肢がある

演習問題2.11

- 2元ハフマン符号は構成上、分岐の部分に
  2通りの選択肢がある
- 分岐の数は$q-1$個ある

演習問題2.11

- 2元ハフマン符号は構成上、分岐の部分に2通りの選択肢がある
- 分岐の数は$q-1$個ある
- よってハフマン符号の数は2^{q-1}個

演習問題2.11

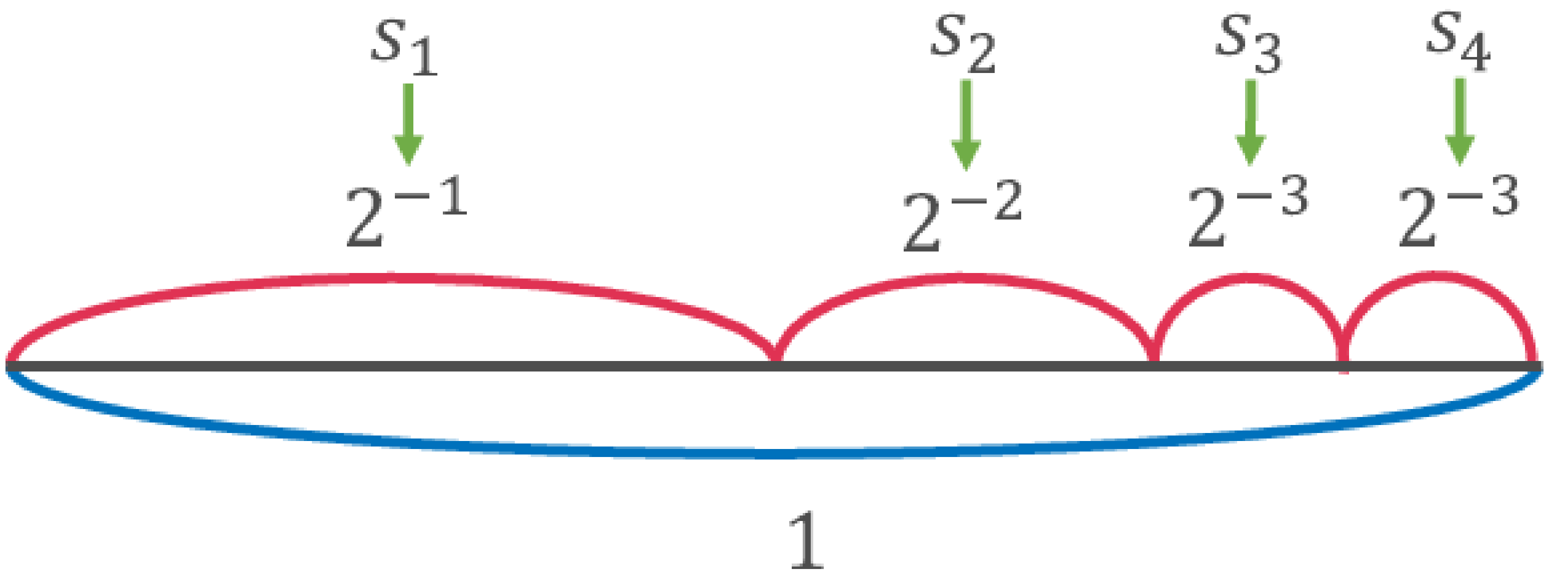
$S=\{s_1,\cdots s_q\}$とし、生起確率は各々$p_1\geq\cdots\geq p_q$とし、
$p_i>p_{i+2}+\cdots+p_q$を満たす。Sに対する任意の
2元ハフマン符号の符号長が$1,2,\cdots,q-1,q-1$
となることを示せ。また、Sに対する異なる
2元ハフマン符号はいくつあるか？さらに各$q\geq 1$に対し、
上の不等式を満たす確率分布の例を構成せよ

最後に具体的な確率分布を構成する

演習問題2.11

- これは次のアルゴリズムによって構成できる
  1. 長さ1の線分を用意してそれをlとし、$i=1$とする
  2. もし$i=q-1$ならばlをs_qに割り当て、p_qをlの長さとする
  3. 割り当てられていない線分を半分にし、半分にした一方をs_iに割り当てて、p_iをその線分の長さとする
  4. あまった線分をlとし、$i \coloneqq i+1$とする
  5. 2.へ

演習問題2.11



$q = 4$の場合

演習問題2.11

- このようにして構成される$p_1, \cdots, p_q$は必ず$\sum_{i=1,\cdots,q} p_i = 1$を満たす

演習問題2.11

- このようにして構成される$p_1, \cdots, p_q$は必ず$\sum_{i=1,\cdots,q} p_i = 1$を満たす
- 実際、$p_{q-1} + p_q = 2^{-(q-1)} + 2^{-(q-1)} = 2^{-(q-2)}$を繰り返すと$\sum_{i=1,\cdots,q} p_i = p_1 + \sum_{i=2,\cdots,q} p_i = 2^{-1} + 2^{-1} = 1$

演習問題2.11

符号長の総和$\sigma(C) = \sum_i l_i$を最小にするように
ハフマン符号を構成できるか？

演習問題2.12

- r元ハフマン符号を構成する際に縮退された
  情報源に対する符号C'から符号Cを構成するとき、
  $w' \in C'$から符号語を構成すると符号長は$|w'|$から
  $|w'|+1$になる

演習問題2.12

- r元ハフマン符号を構成する際に縮退された情報源に対する符号C'から符号Cを構成するとき、$w' \in C'$から符号語を構成すると符号長は$|w'|$から$|w'| + 1$になる
- C'とCの符号長の総和の差は、増えた符号長$l + 1$なr個の符号の総和$r(l + 1)$から増やす前の符号長lを引いた$r(l + 1) - l = l(r - 1) + r$になる

演習問題2.12

- r元ハフマン符号を構成する際に縮退された情報源に対する符号C'から符号Cを構成するとき、$w' \in C'$から符号語を構成すると符号長は$|w'|$から$|w'| + 1$になる
- C'とCの符号長の総和の差は、増えた符号長$l + 1$なr個の符号の総和$r(l + 1)$から増やす前の符号長lを引いた$r(l + 1) - l = l(r - 1) + r$になる
- rは定数なので、符号を構成する各段階のlを最小化するだけでハフマン符号の最適性から達成される

演習問題2.13

$S=\{s_1,\cdots,s_q\}$から1つ要素を選ぶが、どの要素が
選ばれたかはわからないものとする。
質問を繰り返して選ばれた要素$s\in S$を当てる
ただし、各s_iが選ばれる確率p_iは全て既知で、
質問は「Sの部分集合$T\subseteq S$に$s\in S$が
含まれるか？」のみできる。このとき、
質問回数の期待値を最小にするにはどのように
質問すればよいだろうか？

Sに対して0を「$s\notin T$」、1を「$s\in T$」として
2元ハフマン符号を構成すれば良い

演習問題2.14

Cを生起確率が等しいq個のシンボルからなる情報源Sの2元ハフマン符号とする。$L(C^2) < L(C)$とできるか？
全てのnに対し、${}^{L_n}/_{n} = L(C)$となるqの値をいくつか示せ

演習問題2.14

Cを生起確率が等しいq個のシンボルからなる情報源Sの2元ハフマン符号とする。$L(C^2) < L(C)$とできるか？
全てのnに対し、${}^{L_n}/_{n} = L(C)$となるqの値をいくつか示せ

まず、$L(C^2) < L(C)$とできることを示す

演習問題2.14

- $q = 3$のときを考える

演習問題2.14

ここでSとS^2のハフマン符号を構成する

演習問題2.14

- $q = 3$のときを考える
- $L(C)$と$L(C^2)$を計算すると、

$$L(C) = \frac{5}{3}$$

$$L(C^2) = \frac{29}{9}$$

演習問題2.14

- $q = 3$のときを考える
- $L(C)$と$L(C^2)$を計算すると、

$$L(C) = \frac{5}{3}$$

$$L(C^2) = \frac{29}{9}$$

- よって$L(C) = \frac{5}{3} = 1.66 \cdots > \frac{L(C^2)}{2} = \frac{29}{18} = 1.61 \cdots$

演習問題2.14

Cを生起確率が等しいq個のシンボルからなる情報源Sの2元ハフマン符号とする。$L(C^2) < L(C)$とできるか？
全てのnに対し、${}^{L_n}/_{n} = L(C)$となるqの値をいくつか示せ

最後に全てのnに対し、${}^{L_n}/_{n} = L(C)$となるqの値を探す

演習問題2.14

- $q = 2^l$を考える

演習問題2.14

- $q = 2^l$を考える
- 全てのシンボルの生起確率は等しいので、Sに対する2元ハフマン符号は、その構成法から全ての符号長は等しくなる

演習問題2.14

- $q = 2^l$を考える
- 全てのシンボルの生起確率は等しいので、Sに対する2元ハフマン符号は、その構成法から全ての符号長は等しくなる
- 各分岐点毎に2つの符号ができるので$L(C) = \log 2^l = l$

演習問題2.14

- $q = 2^l$を考える
- 全てのシンボルの生起確率は等しいので、Sに対する2元ハフマン符号は、その構成法から全ての符号長は等しくなる
- 各分岐点毎に2つの符号ができるので$L(C) = \log 2^l = l$
- S^nを考えるとSの全てのシンボルの生起確率が等しいことから$\left(2^l\right)^n = 2^{ln}$個の生起確率が等しいシンボルを持つ

演習問題2.14

- $q = 2^l$を考える
- 全てのシンボルの生起確率は等しいので、Sに対する2元ハフマン符号は、その構成法から全ての符号長は等しくなる
- 各分岐点毎に2つの符号ができるので$L(C) = \log 2^l = l$
- S^nを考えるとSの全てのシンボルの生起確率が等しいことから$\left(2^l\right)^n = 2^{ln}$個の生起確率が等しいシンボルを持つ
- よって同様の操作で$\frac{L(C^n)}{n} = \frac{ln}{n} = l = L(C)$