情報理論

第10回　講義

情報源のエントロピー

2015. 6. 24

植松　芳彦

前回分かったこと(1／2)

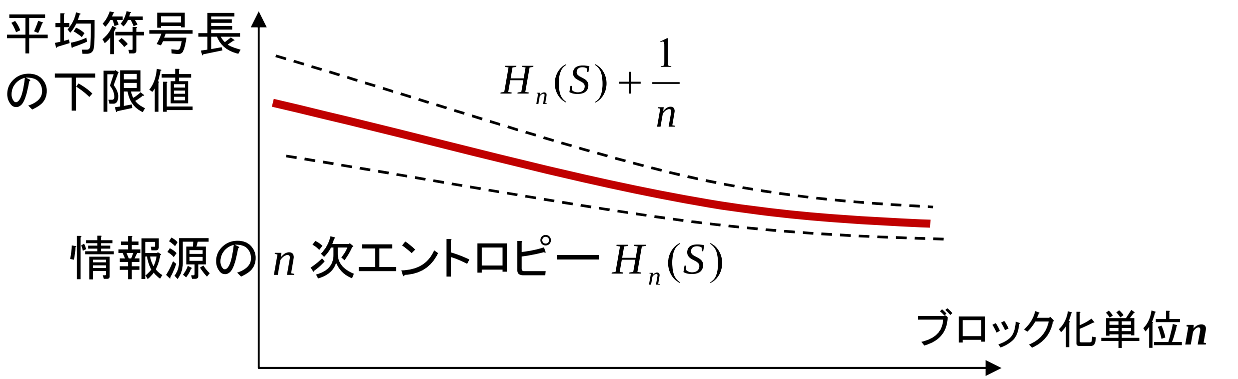
- n 個の情報源記号を纏めて符号化するブロック符号化により，より符号化効率が高められる.

$$\overbrace{X_1}\ \overbrace{X_2}\ \cdots \Rightarrow \overbrace{X_1 \cdots X_n}\ \overbrace{\cdots}$$

*1*記号づつ符号化　　　　n記号づつ纏めて符号化

1情報源記号あたりの平均符号長 $$L = \frac{L_n}{n} \qquad \text{（式4.25）}$$

平均符号長の下限値 $$H_n(S) \le L < H_n(S) + \frac{1}{n} \qquad \text{（式4.26）}$$



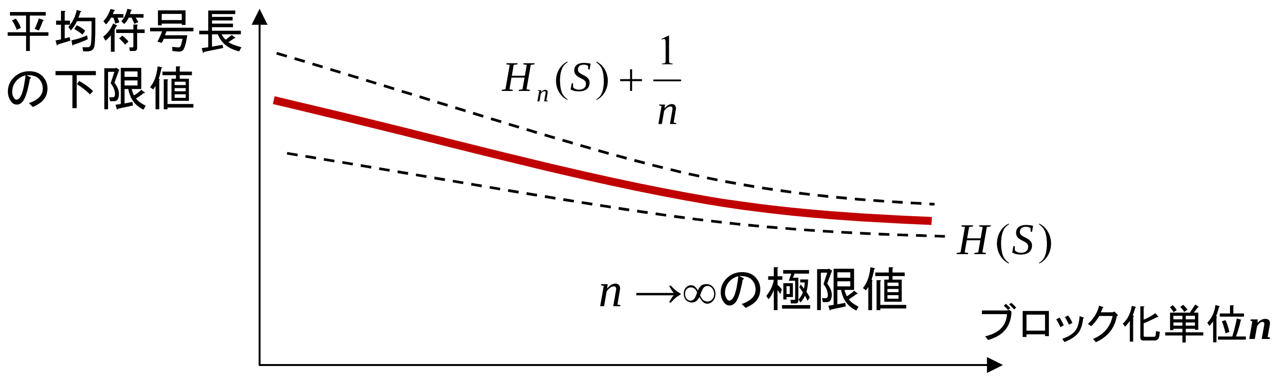
前回分かったこと(2／2)

- ブロック化の単位 $\boldsymbol{n}$ を無限大とした極限状態において，平均符号長の下限が情報源のエントロピーで与えられる．

情報源の n 次エントロピー $$H_n(S) = \frac{H_1(S^n)}{n} \qquad \text{(式4.27)}$$

ただし $$H_1(S^n) = -\sum_{x_0=a_1}^{a_M} \cdots \sum_{x_{n-1}=a_1}^{a_M} P(x_0,\cdots,x_{n-1}) \bullet \log_2 P(x_0,\cdots,x_{n-1}) \qquad \text{(式4.24)}$$

情報源のエントロピー
（n→∞の極限値） $$H(S) = \lim_{n\to\infty} H_n(S) \qquad \text{(式4.28)}$$



本日の講義内容

- いくつかの基本的な情報源のエントロピーを分析

1．記憶のない情報源のエントロピー

2．記憶のある情報源のエントロピー
（マルコフ情報源）

記憶のない情報源のエントロピー

- 記憶のない情報源 S では，相続く n 個の記号の結合確率分布は単純化できる.

　情報源　　S

　情報源記号 a_i $(i = 1, \cdot\cdot, \mathrm{M})$

　発生確率　p_i

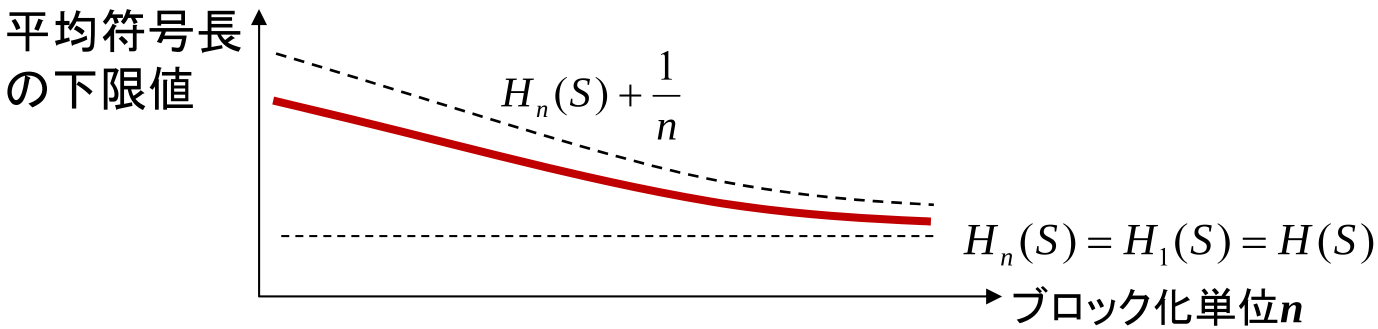
（式4.33）

　n個の記号の結合確率分布 $P(x_0, \cdots, x_{n-1}) = P(x_0)P(x_1)\cdots P(x_{n-1})$

- この時情報源のエントロピーには以下が成り立つ

$$H_n(S) = \frac{H_1(S^n)}{n} = H_1(S) = H(S)$$

（式4.34）



【証明】

- ブロック化単位 $\boldsymbol{n}=2$ の場合を考える.

$$H_1(S^2) = -\sum_{x_0=a_1}^{a_M}\sum_{x_1=a_1}^{a_M} P(x_0,x_1)\bullet\log_2 P(x_0,x_1)$$

⇦ 単なる置換

$$= -\sum_{x_0=a_1}^{a_M}\sum_{x_1=a_1}^{a_M} P(x_0)\bullet P(x_1)\bullet\log_2 P(x_0)\bullet P(x_1)$$

$$\log_2 P(x_0)\bullet P(x_1) = \log_2 P(x_0)+\log_2 P(x_1)$$

⇦

$$= \underbrace{-\sum_{x_0=a_1}^{a_M}\sum_{x_1=a_1}^{a_M} P(x_0)\bullet P(x_1)\bullet\log_2 P(x_0)}_{= -\sum_{x_0=a_1}^{a_M} P(x_0)\bullet\log_2 P(x_0)\bullet\sum_{x_1=a_1}^{a_M} P(x_1)} - \sum_{x_0=a_1}^{a_M}\sum_{x_1=a_1}^{a_M} P(x_0)\bullet P(x_1)\bullet\log_2 P(x_1)$$

$$= -\sum_{x_0=a_1}^{a_M} P(x_0)\bullet\log_2 P(x_0)\bullet\sum_{x_1=a_1}^{a_M} P(x_1)$$

各情報源記号を
出力する確率の総和は1 ⇨

$$= -\sum_{x_0=a_1}^{a_M} P(x_0)\bullet\log_2 P(x_0)$$

$P(x_i)=p_i$ ⇨

$$= -\sum_{i=1}^{M} p_i\bullet\log_2 p_i = H_1(S)$$

【証明】

- $n = 2$ の場合について言えたこと

$$H_1(S^2) = 2H_1(S)$$

- 同様に任意の整数 n に対して以下が成り立つことを証明可能.

$$H_1(S^n) = n \bullet H_1(S) = -n \bullet \sum_{i=1}^{M} p_i \bullet \log_2 p_i$$

- よって以下が成り立つ.

$$H_n(S) = \frac{H_1(S^n)}{n} = H_1(S) = \sum_{i=1}^{M} p_i \bullet \log_2 p_i = H(S)$$

***n* がいくつであっても**
***H_n (S)* の値は変わりようがない**

記憶のない情報源のエントロピー

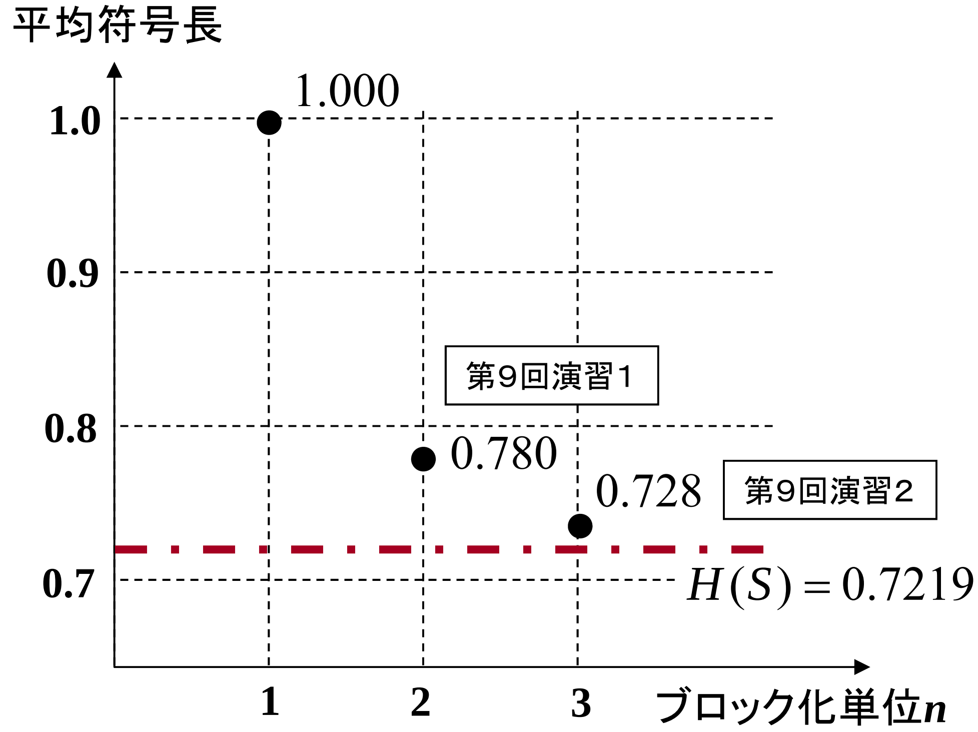
- 前回分析した記憶のない2元情報源モデルに立ち返る
- 3記号纏めた符号化により下限値に近い平均符号長を達成している.

各情報源記号の発生確率

情報源記号	発生確率
A	0.8
B	0.2

情報源記号列のエントロピー

$$H_n(S) = H_1(S) = H(S)$$
$$= -p_A \bullet \log_2 p_A - p_B \bullet \log_2 p_B$$
$$= 0.7219$$



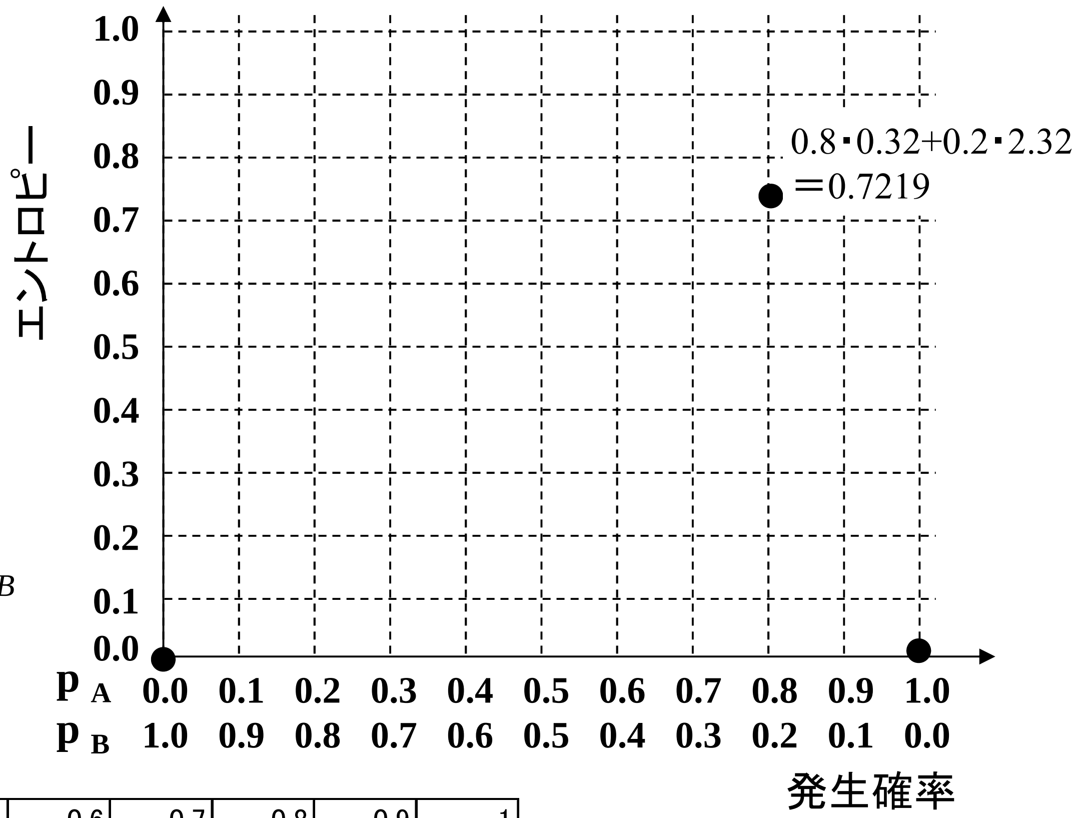
【演習1】記憶のない情報源のエントロピー

- 各情報源記号の発生確率に対するエントロピーの依存性を評価する.

補足 0.5を挟んで左右対称
小数点以下2桁で充分

記憶のない2元情報源
（各情報源記号の発生確率）

情報源記号	発生確率
A	p_A
B	p_B



記憶のない2元情報源のエントロピー

$$H(S) = -p_A \bullet \log_2 p_A - p_B \bullet \log_2 p_B$$

数値データ

p	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
−log2p	∞	3.322	2.322	1.737	1.322	1.000	0.737	0.515	0.322	0.152	0.000

記憶のない情報源のエントロピー

- 記憶のない2元情報源のエントロピー式では以下が成立

  $$p_A + p_B = 1$$

- $p_A = x$, $p_B = 1 - x$ としたxの関数をエントロピー関数と呼ぶ

記憶のない2元情報源のエントロピー

$$H(S) = -p_A \bullet \log_2 p_A - p_B \bullet \log_2 p_B$$

$$p_A = x$$

$$p_B = 1 - x$$

$$\mathscr{H}(x) = -x \bullet \log_2 x - (1-x) \bullet \log_2 (1-x)$$

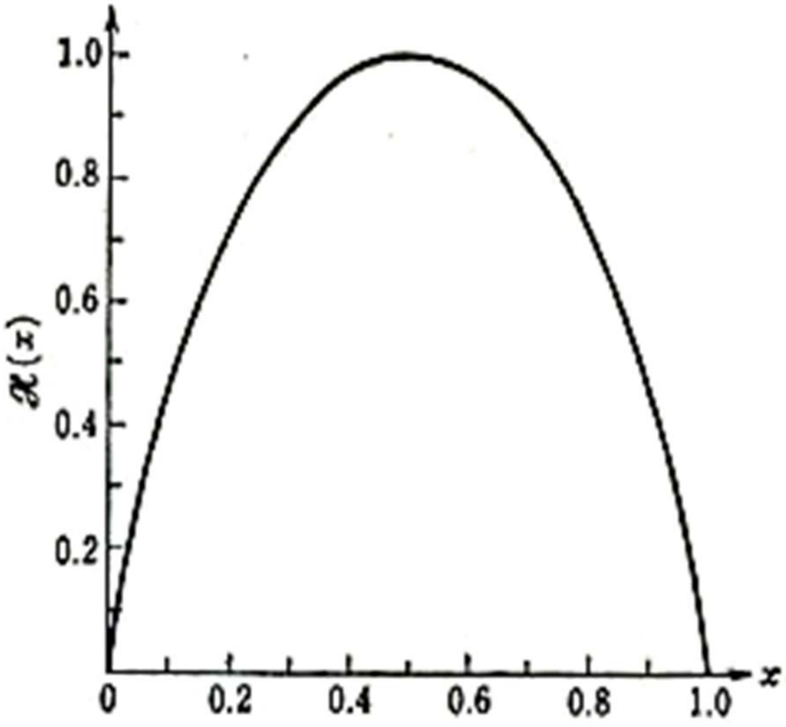
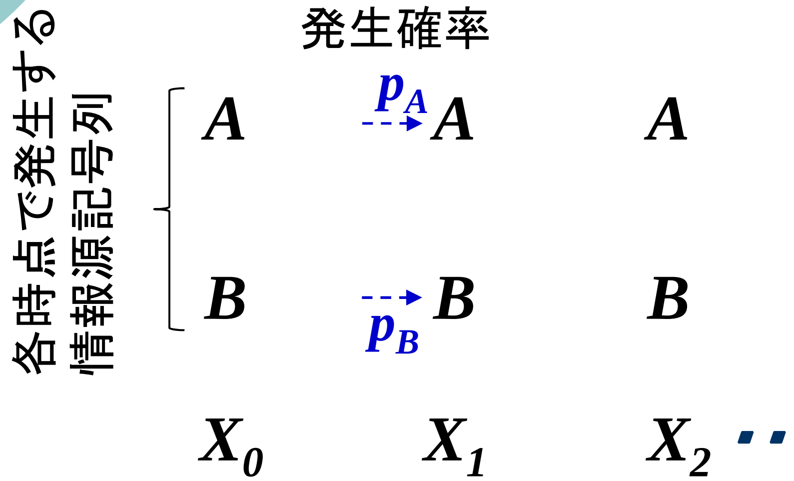
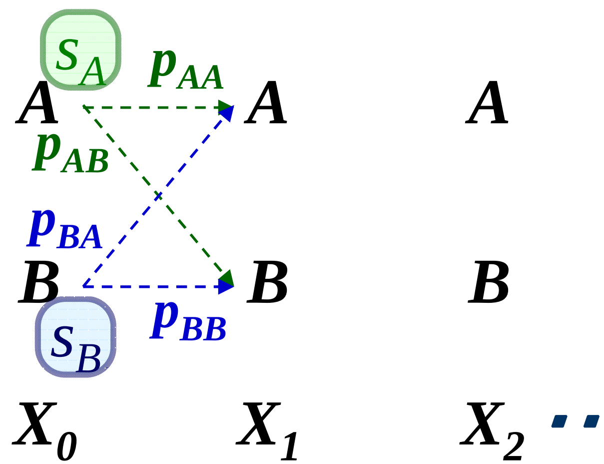


図 4.8 エントロピー関数 $\mathscr{H}(x)$

マルコフ情報源のエントロピー

- 記憶のある2元情報源では，以前に発生した情報源記号に応じて次の情報源記号の発生確率が異なる．
- マルコフ情報源を例に，記憶のなし／ありでのエントロピーの違いを分析する．





記憶のない2元情報源
発生確率が前に発生した記号に依存しない

記憶のある2元情報源（マルコフ情報源）
発生確率が，直前に発生した記号に依存する（直前に何を発生したかで「状態」が異なる）

マルコフ情報源のエントロピー

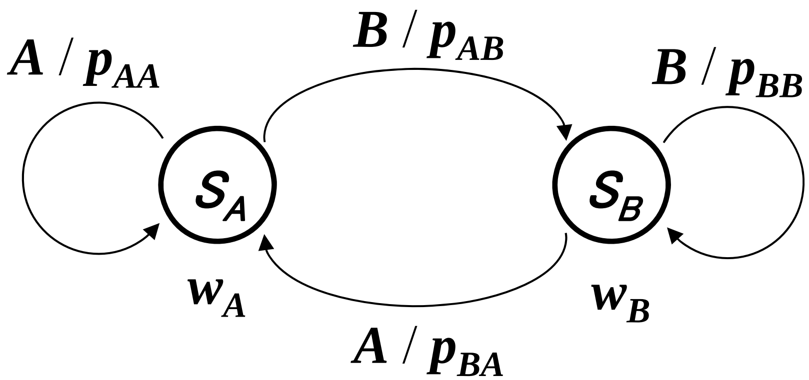
- 各状態$\boldsymbol{S_A}$，$\boldsymbol{S_B}$にある場合のエントロピーを考える．例えば状態$\boldsymbol{S_A}$にある時，記号$\boldsymbol{A}$，$\boldsymbol{B}$を$\boldsymbol{p_{AA}}$，$\boldsymbol{p_{AB}}$で発生する記憶のない情報源と振る舞いは変わらない．

$$H_{SA}(S) = -p_{AA} \bullet \log_2 p_{AA} - p_{AB} \bullet \log_2 p_{AB}$$

$$H_{SB}(S) = -p_{BA} \bullet \log_2 p_{BA} - p_{BB} \bullet \log_2 p_{BB}$$

- エントロピーは各状態の存在確率に各状態にいる時のエントロピーを掛け合わせることで求められる．

$$H(S) = w_A \bullet H_{SA}(S) + w_B \bullet H_{SB}(S)$$



【演習2】マルコフ情報源のエントロピー

- 教科書図4.10のマルコフ情報源のエントロピーを考える.
- まず状態遷移図を解いて，各状態に存在する確率分布を求めよう.

$$[w_A \quad w_B] = [w_A \quad w_B] \cdot \begin{bmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{bmatrix}$$

$$w_A + w_B = 1$$

［計算］

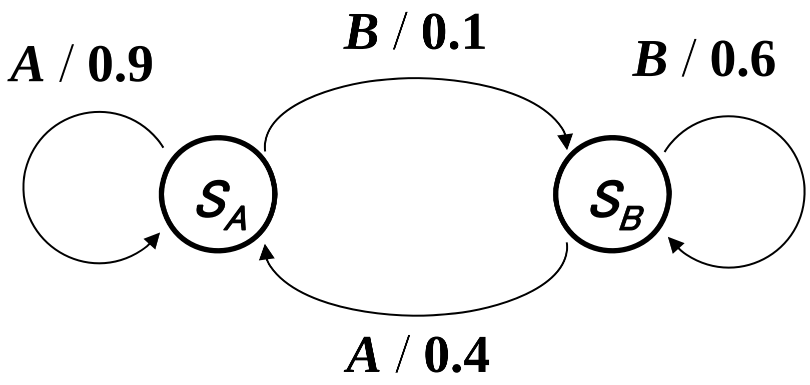


図4.10　マルコフ情報源の例
（発生符号0,1　→　A, B）

【演習2】マルコフ情報源のエントロピー

- エントロピーを求めよう.

状態A

$$H_{SA}(S) = -0.9 \bullet \log_2 0.9 - 0.1 \bullet \log_2 0.1$$
$$=$$

状態B

$$H_{SB}(S) = -\boxed{} \bullet \log_2 \boxed{} - \boxed{} \bullet \log_2 \boxed{}$$
$$=$$

全状態

$$H(S) = w_A \bullet H_{SA}(S) + w_B \bullet H_{SB}(S)$$
$$=$$

数値データ

p	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
-log2p	∞	3.322	2.322	1.737	1.322	1.000	0.737	0.515	0.322	0.152	0.000

マルコフ情報源のエントロピー

- このマルコフ情報源が情報源記号A, Bを発生する確率を求める（レポート問題2／2参照）

記号Aを発生する確率 $=S_A$に存在する確率$\times A$を出す確率 $+S_B$に存在する確率$\times A$を出す確率 $=0.8\times0.9+0.2\times0.4=0.8$

記号Bを発生する確率 $=S_A$に存在する確率$\times B$を出す確率 $+S_B$に存在する確率$\times B$を出す確率 $=0.8\times0.1+0.2\times0.6=0.2$

マルコフ情報源のエントロピー

- 記憶のない／ある情報源では，各記号の発生確率が同じでもエントロピーが異なる

元々の2元情報源

情報源記号	発生確率
A	0.8
B	0.2

記憶なしの場合
エントロピー＝0.7219

記憶ありの場合
（マルコフ情報源の場合）
エントロピー＝0.5694

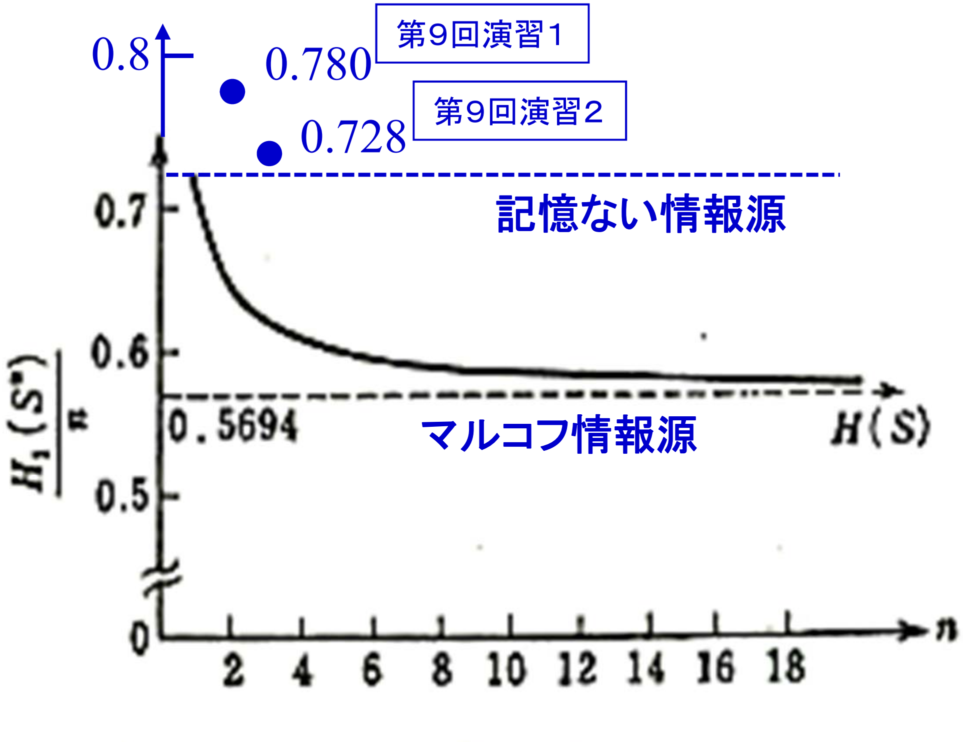
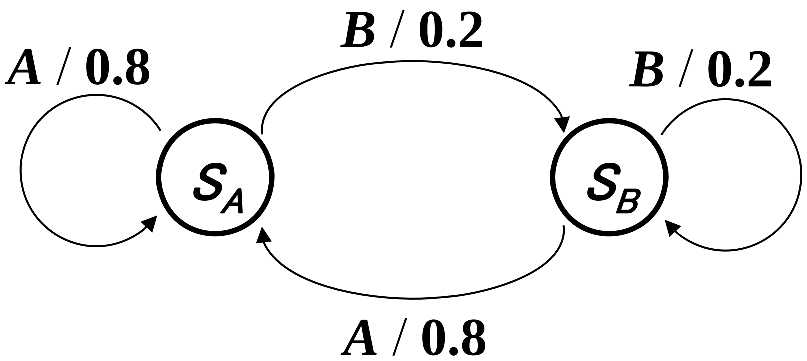


図 4.11 マルコフ情報源のエントロピー

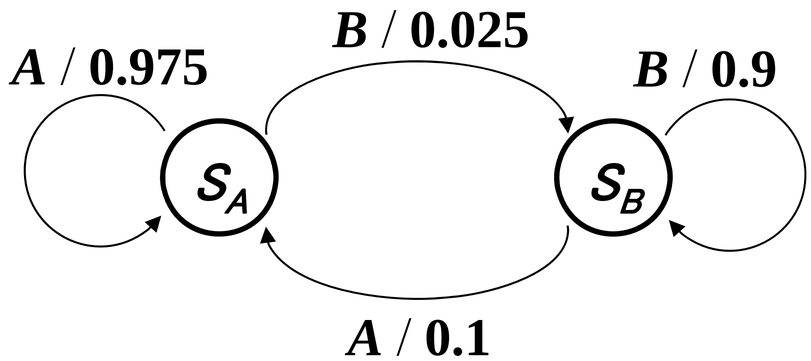
マルコフ情報源のエントロピー

- マルコフ情報源の場合，記号***A***や***B***を発生する平均的な確率が同じでも，状態遷移確率により記号の出方やエントロピーは大きく異なる．

両方の場合とも記号A/Bの平均的な発生確率＝0.8/0.2



・同じ記号が連続しにくい
・エントロピー = 0.7220
（記憶ない場合に近い）



・同じ記号が連続しやすい
A が出始めたら ***A*** ばかり
B が出始めたら ***B*** ばかり
・エントロピー = 0.2287

本日のまとめ

- 記憶のない情報源，記憶のある情報源（代表例としてマルコフ情報源）のエントロピーの傾向を分析．
- 記憶のない／ある情報源のエントロピーの違いの要因を考察．