隠れマルコフモデル
Hidden Markov Model

山下 滉

2017/04/21

1　マルコフ連鎖

「コイントスして，表なら左へ一歩，裏なら右へ一歩」を延々と繰り返すという最も基本的な確率過程のことを**ランダムウォーク（酔歩）**という．左右の確率を変えたり前後左右へ動くようにしたりと，バリエーションはいろいろ考えられるが，ここでは単純に一次元で左右等確率の設定とする．形式的に書くと，$+1$ か -1 かが半々の確率で出る i.i.d[1]な確率変数たち $Z_1, Z_2, Z_3, ...$ を使って，

$$X_0 = 0, \quad X_t = X_{t-1} + Z_t \ (t = 1, 2, ...) \tag{1.1}$$

と表される X_t のことである [1].

ランダムウォークの場合，

$$P(X_{t+1} = x_{t+1} | X_t = x_t, X_{t-1} = x_{t-1}, ..., X_0 = x_0) = P(X_{t+1} = x_{t+1} | X_t = x_t) \tag{1.2}$$

が成り立つ．つまり，未来の状態は今の状態だけから定まり，過去の履歴（どこからどんな経路をたどって今の状態にたどりついたか）には無関係であった．現在の状態が一時点前の状態に依存して確率的に決まるような特性を**マルコフ性（Markov property）**という．そしてこのような確率過程を**マルコフ過程（Markov process）**という．その中でも特に，X_t がとり得る値が有限とおりなものを**マルコフ連鎖（Markov chain）**と呼ぶ.

マルコフモデルは複数の状態を持ち，ある状態から別の状態（元の状態も含む）へ一定の確率で遷移する．この確率を**遷移確率（transition probability）**という．また，現在の状態に依存した一定の確率で特定の**出力記号（output symbol）**を出力する．この確率を**出力確率（output probability）**という [1].

[1]独立同一分布（independent and indentically distributed; i.i.d）

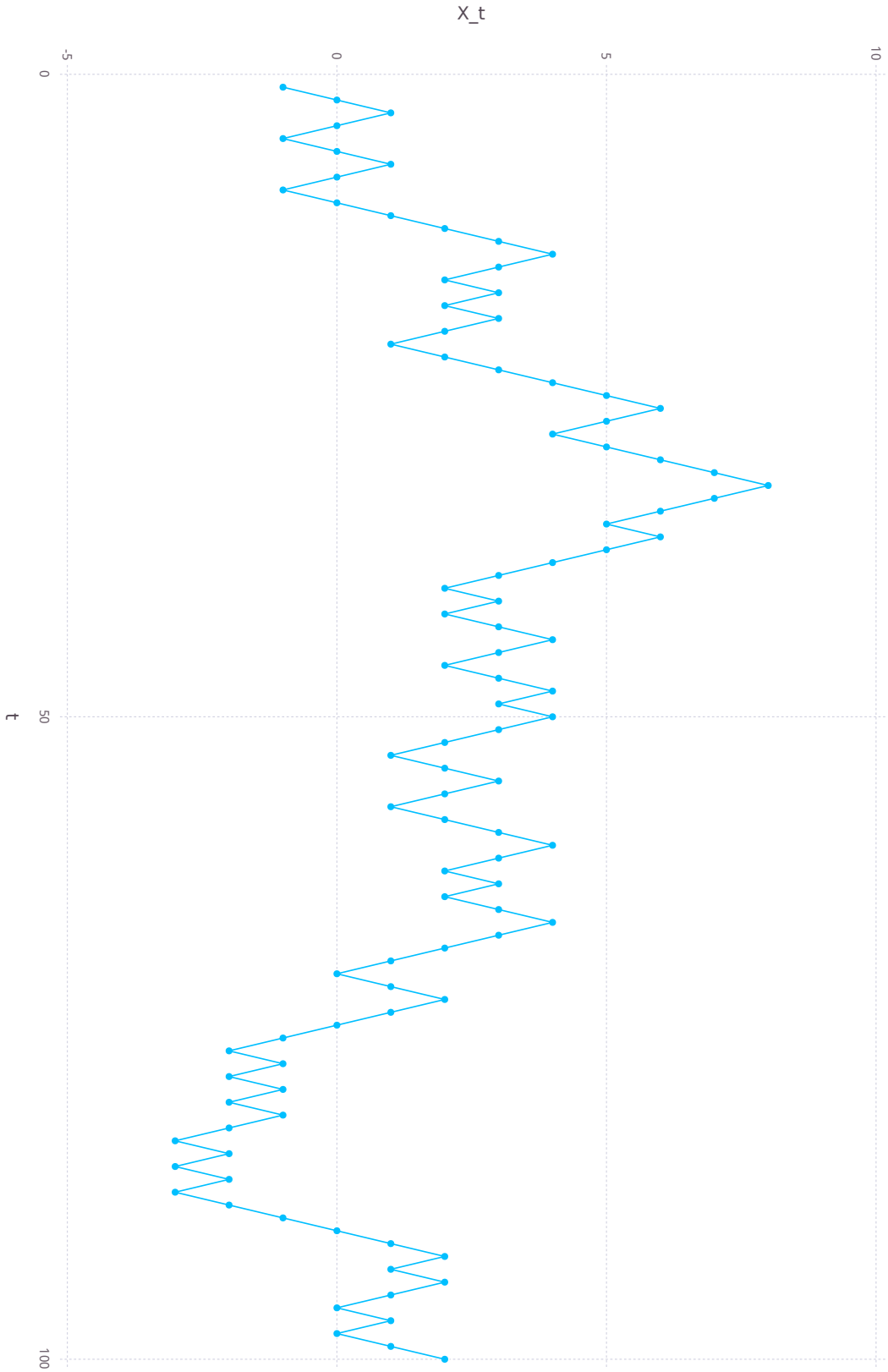


図 1: ランダムウォークの例

2　マルコフモデル

例として，3 種類のサイコロ ω_1,ω_2,ω_3 を投げて出た目を観測する場合を考える．ここではサイコロの目として奇数と偶数の 2 種類を考える．

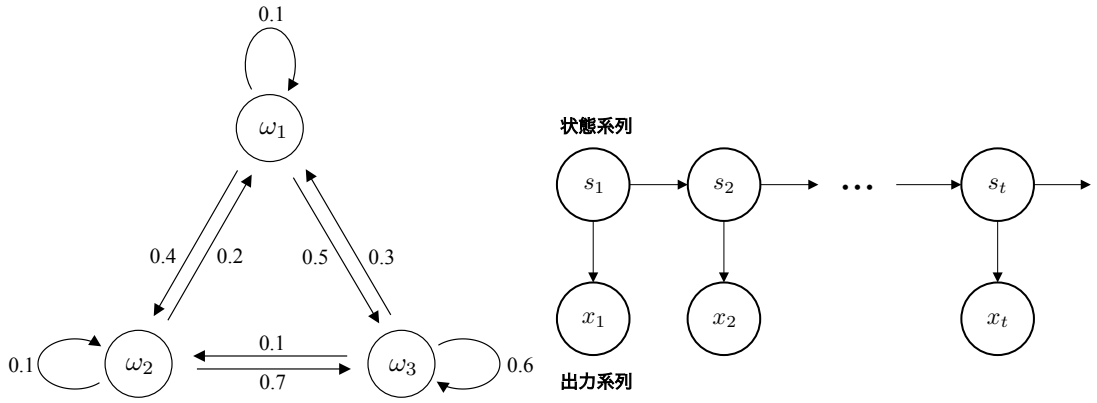
遷移確率行列 $\boldsymbol{A}$，出力確率行列 $\boldsymbol{B}$ が以下のように与えられたとする．

$$\boldsymbol{A} = \begin{array}{c} \\ \omega_1 \\ \omega_2 \\ \omega_3 \end{array} \begin{array}{c} \begin{array}{ccc} \omega_1 & \omega_2 & \omega_3 \end{array} \\ \begin{pmatrix} 0.1 & 0.4 & 0.5 \\ 0.2 & 0.1 & 0.7 \\ 0.3 & 0.1 & 0.6 \end{pmatrix} \end{array} \tag{2.1}$$

$$\boldsymbol{B} = \begin{array}{c} \\ \omega_1 \\ \omega_2 \\ \omega_3 \end{array} \begin{array}{c} \begin{array}{cc} \text{奇数} & \text{偶数} \end{array} \\ \begin{pmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \\ 0.3 & 0.7 \end{pmatrix} \end{array} \tag{2.2}$$

行列 $\boldsymbol{A}$ により，遷移確率 a_{ij} を反映した状態遷移系列が得られる．状態遷移の様子を図 2(a) に示す．行列 $\boldsymbol{B}$ は，サイコロ ω_1,ω_2,ω_3 を投げて奇数の目が出る確率がそれぞれ $0.8, 0.6, 0.3$，偶数の目が出る確率がそれぞれ $0.2, 0.4, 0.7$ であることを示している．

時刻 t における状態を s_t，出力記号を x_t として，マルコフモデルをグラフィカルモデルで表したものを図 2(b) に示す．マルコフモデルでは状態系列と出力系列の両方を観測することができる．この観測値から $\boldsymbol{A}$, $\boldsymbol{B}$ などのパラメータを推定するのがマルコフモデルの問題である．状態系列が与えられていることから，マルコフモデルは教師あり学習であるといえる．



(a) 状態遷移図　　(b) 状態空間モデル

図 2: マルコフモデル

3　隠れマルコフモデル

例題

箱の中に c 種類のサイコロ $\omega_1, \omega_2, ..., \omega_c$, があり，そのいずれかを取り出して投げ，出た目を観測した後，サイコロを元の箱に戻すという操作を n 回繰り返す．ここで，

1. 最初にサイコロ ω_i を取り出す確率は π_i である．
2. サイコロ ω_i を取り出した後にサイコロ ω_j を取り出す確率は a_{ij} である．
3. サイコロ ω_j を投げて出た目が v_k となる確率は b_{jk} である．

とする．ただし，$i, j = 1, 2, ..., c,\ k = 1, 2, ..., m$ である．その結果，サイコロの目の系列として $\mathbf{x} = x_1 x_2 \cdots x_t \cdots x_n$ が得られた．ただし，取り出したサイコロの種類については知ることができないものとする．

(1) パラメータ $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}$ が既知のとき，このような観測結果が得られる確率 $P(\mathbf{x})$ を求めよ．

(2) パラメータ $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}$ が未知のとき，観測結果より，これらのパラメータを最尤推定により推定せよ．

パラメータ

c	: 状態数
m	: 出力記号の数
$s_t \in \{\omega_1, \omega_2, ..., \omega_c\}$	: 時点 t での状態
$x_t \in \{v_1, v_2, ..., v_m\}$	: 時点 t での観測結果 (出力記号)
$\mathbf{s} = s_1 s_2 \cdots s_t \cdots s_n$	: 状態系列
$\mathbf{x} = x_1 x_2 \cdots x_t \cdots x_n$	: 観測記号系列
a_{ij}, $a(\omega_i, \omega_j)$	: 状態 ω_i から状態 ω_j への遷移確率 $P(\omega_j \mid \omega_i)$
b_{jk}, $b(\omega_j, v_k)$	: 状態 ω_j で v_k を出力する確率 $P(v_k \mid \omega_j)$
π_i	: 初期状態 $(t = 1)$ が ω_i である確率 $P(s_1 = \omega_i)$
$\boldsymbol{A}$	: a_{ij} を (i, j) 成分としてもつ $c \times c$ の行列
$\boldsymbol{B}$	: b_{jk} を (j, k) 成分としてもつ $c \times m$ の行列
$\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_c)$	: π_i を成分としてもつ c 次元のベクトル

この例題では，観測者はサイコロの投げた結果（出力記号) は知ることができるものの，サイコロの種類（状態）の系列を知ることができない．このようなモデルを**隠れマルコフモデル（Hidden Markov Model）**という．

例題 (1) は，パラメータを既知としたとき，出力記号系列としての観測結果 $\mathbf{x}$ が得られる確率を求める**評価**の問題である．例題 (2) は，パラメータを未知としたとき，観測結果から未知パラメータを求める**推定**の問題である．

4 Forward アルゴリズム

例題 (1) は，サイコロを投げた観測結果 $\mathbf{x}$ の生起確率 $P(\mathbf{x})$ を求める問題である．観測できるのは $\mathbf{x}$ のみであり，サイコロの種類の系列 $\mathbf{s}$ については知ることができないので，観測結果 $\mathbf{x}$ が得られたとき，$\mathbf{s}$ としては可能な全ての系列を想定しなくてはならない．すなわち，このような観測結果が得られる確率 $P(\mathbf{x})$ は，

$$P(\mathbf{x}) = \sum_{\mathbf{s}} P(\mathbf{x}, \mathbf{s}) \tag{4.1}$$

である．ここで $\sum_{\mathbf{s}}$ は，

$$\sum_{s_1} \sum_{s_2} \cdots \sum_{s_n} \tag{4.2}$$

を表しており，可能な全ての系列 $\mathbf{s}$ についての和をとるという網羅的な演算である．式 (4.1) の $P(\mathbf{x}, \mathbf{s})$ は確率の乗法定理より，

$$P(\mathbf{x}, \mathbf{s}) = P(\mathbf{x}|\mathbf{s})P(\mathbf{s}) \tag{4.3}$$

で表される．ここで，$P(\mathbf{x}|\mathbf{s})$ は，$\mathbf{s}$ が与えられている前提があるため，

$$P(\mathbf{x}|\mathbf{s}) = P(x_1 x_2 \cdots x_n | s_1 s_2 \cdots s_n) \tag{4.4}$$

$$= P(x_1|s_1)P(x_2|s_2) \cdots P(x_n|s_n) \tag{4.5}$$

が成り立つので，

$$b(s_t, x_t) = P(x_t|s_t) \tag{4.6}$$

であることを用いて，

$$P(\mathbf{x}|\mathbf{s}) = b(s_1, x_1)b(s_2, x_2) \cdots b(s_n, x_n) \tag{4.7}$$

$$= \prod_{t=1}^{n} b(s_t, x_t) \tag{4.8}$$

と書ける．一方，$P(\mathbf{s})$ は上記のような独立性は成り立たないが，

$$P(\mathbf{s}) = P(s_1 s_2 \cdots s_n) \tag{4.9}$$

$$= P(s_1)a(s_1, s_2)a(s_2, s_3) \cdots a(s_{n-1}, s_n) \tag{4.10}$$

$$= \prod_{t=1}^{n} a(s_{t-1}, s_t) \tag{4.11}$$

が成り立つ．ただし，

$$a(s_0, s_1) \stackrel{\text{def}}{=} P(s_1) \tag{4.12}$$

である．式 (4.8) と式 (4.11) を式 (4.3) に代入することによって，

$$P(\mathbf{x}, \mathbf{s}) = \prod_{t=1}^{n} a(s_{t-1}, s_t)b(s_t, x_t) \tag{4.13}$$

が得られる．よって式 (4.13) より，式 (4.1) は，

$$P(\mathbf{x}) = \sum_{\mathbf{s}} \prod_{t=1}^{n} a(s_{t-1}, s_t)b(s_t, x_t) \tag{4.14}$$

と表せる．ここで，式 (4.14) の計算量について求めると，$O(nc^n)$ となる．これは膨大な計算が必要になり，この方法は非現実的である．

そこで，$P(\mathbf{x})$ の効率的な計算方法について述べる．まず，隠れマルコフモデルには以下の様な条件付き独立性が成立する [2].

隠れマルコフモデルの条件付き独立性

(1) $P(\mathbf{x}|s_t) = P(x_1 \cdots x_t|s_t)P(x_{t+1} \cdots x_n|s_t)$

(2) $P(x_1 \cdots x_{t-1}|x_t, s_t) = P(x_1 \cdots x_{t-1}|s_t)$

(3) $P(x_1 \cdots x_{t-1}|s_{t-1}, s_t) = P(x_1 \cdots x_{t-1}|s_{t-1})$

(4) $P(x_{t+1} \cdots x_n|s_t, s_{t+1}) = P(x_{t+1} \cdots x_n|s_{t+1})$

(5) $P(x_{t+2} \cdots x_n|s_{t+1}, x_{t+1}) = P(x_{t+2} \cdots x_n|s_{t+1})$

(6) $P(\mathbf{x}|s_{t-1}, st) = P(x_1 \cdots x_{t-1}|s_{t-1})P(x_t|s_t)P(x_{t+1} \cdots x_n|s_t)$

(7) $P(x_{n+1}|\mathbf{x}, s_{n+1}) = P(x_{n+1}|s_{n+1})$

(8) $P(s_{n+1}|s_n, \mathbf{x}) = P(s_{n+1}|s_n)$

観測結果 $\mathbf{x} = x_1x_2 \cdots x_n$ が得られ，かつ t 回目にサイコロ ω_i を取り出している確率 $P(\mathbf{x}, s_t = \omega_i)$ は，確率の乗法定理と隠れマルコフモデルの条件付き独立性 (1) より，

$$\begin{aligned} P(\mathbf{x}, s_t = \omega_i) &= P(s_t = \omega_i)P(\mathbf{x}|s_t = \omega_i) \\ &= P(s_t = \omega_i)P(x_1x_2 \cdots x_t|s_t = \omega_i)P(x_{t+1}x_{t+2} \cdots x_n|s_t = \omega_i) \\ &= P(x_1x_2 \cdots x_t, s_t = \omega_i)P(x_{t+1}x_{t+2} \cdots x_n|s_t = \omega_i) \\ &= \alpha_t(i)\beta_t(i) \end{aligned} \tag{4.15}$$

と書ける．ここで，$\alpha_t(i)$,$\beta_t(i)$ は

$$\alpha_t(i) \stackrel{\text{def}}{=} P(x_1x_2 \cdots x_t, s_t = \omega_i) \tag{4.16}$$

$$\beta_t(i) \stackrel{\text{def}}{=} P(x_{t+1}x_{t+2} \cdots x_n|s_t = \omega_i) \tag{4.17}$$

とした．$\alpha_t(i)$ は，$x_1x_2 \cdots x_t$ という観測結果が得られ，かつ t 回目にサイコロ ω_i を取り出している確率である．一方，$\beta_t(i)$ は，t 回目にサイコロ ω_i を取り出したという条件のもとで，それ以降の観測結果が $x_{t+1}x_{t+2} \cdots x_n$ となる確率である．すなわち，式 (4.15) では，t 回目までの観測結果までと，それより後の観測結果とに分けて考えている．

ここで，$\alpha_t(i)$ は，次式で示すように再帰的な計算方法で求めることができる．

$$\alpha_t(j) = P(x_1x_2 \cdots x_{t-1}x_t, s_t = \omega_j) \tag{4.18}$$

$$= P(x_1x_2 \cdots x_t|s_t = \omega_j)P(s_t = \omega_j) \tag{4.19}$$

$$= P(x_t|s_t = \omega_j)P(x_1x_2 \cdots x_{t-1}|s_t = \omega_j)P(s_t = \omega_j) \tag{4.20}$$

$$= P(x_t|s_t = \omega_j)P(x_1x_2 \cdots x_{t-1}, s_t = \omega_j) \tag{4.21}$$

$P(x_t|s_t = \omega_j) = b(\omega_j, x_t)$ より，

$$= b(\omega_j, x_t) \sum_{s_{t-1}} P(x_1 x_2 \cdots x_{t-1}, s_{t-1}, s_t = \omega_j) \tag{4.22}$$

$$= b(\omega_j, x_t) \sum_{s_{t-1}} P(x_1 x_2 \cdots x_{t-1}, s_t = \omega_j | s_{t-1}) P(s_{t-1}) \tag{4.23}$$

$$= b(\omega_j, x_t) \sum_{s_{t-1}} P(x_1 x_2 \cdots x_{t-1} | s_{t-1}) P(s_t = \omega_j | s_{t-1}) P(s_{t-1}) \tag{4.24}$$

$$= b(\omega_j, x_t) \sum_{s_{t-1}} P(x_1 x_2 \cdots x_{t-1}, s_{t-1}) P(s_t = \omega_j | s_{t-1}) \tag{4.25}$$

$$= b(\omega_j, x_t) \sum_{i=1}^{c} P(x_1 x_2 \cdots x_{t-1}, s_{t-1} = \omega_i) P(s_t = \omega_j | s_{t-1} = \omega_i) \tag{4.26}$$

式 (4.16) と $P(s_t = \omega_j | s_{t-1} = \omega_i) = a_{ij}$ より，

$$= b(\omega_j, x_t) \sum_{i=1}^{c} \alpha_{t-1}(i) a_{ij} \tag{4.27}$$

よって，

$$\alpha_t(j) = b(\omega_j, x_t) \sum_{i=1}^{c} \alpha_{t-1}(i) a_{ij} \tag{4.28}$$
$$(t = 2, 3, ..., n) \quad (j = 1, 2, ..., c)$$

ただし，

$$\begin{aligned} \alpha_1(i) &= P(x_1, s_1 = \omega_i) \\ &= P(s_1 = \omega_i) P(x_1 | s_1 = \omega_i) \\ &= \pi_i b(\omega_i, x_1) \quad (i = 1, 2, ..., c) \end{aligned} \tag{4.29}$$

である．

観測順と同じ方向に向かって順次 $\alpha_t(j)$ を求めていることから，このような計算方法**前向きアルゴリズム（Forward algorithm）**と呼ぶ [3]．

ここで，式 (4.16) において，$t = n$ とすることにより，

$$\alpha_n(i) = P(x_1 x_2 \cdots x_n, s_n = \omega_i) \tag{4.30}$$

が得られ，式 (4.30) の同時確率を s_n について周辺化することにより，求める $P(\mathbf{x})$ は

$$P(\mathbf{x}) = \sum_{i=1}^{c} \alpha_n(i) \tag{4.31}$$

となる．この計算量は $O(c^2 n)$ であり，先ほどと比較すればその効率性は明らかである．

Forward アルゴリズムの手順を Algorithm 1 に示す．

Algorithm 1 Forward algorithm

Step 1 初期化

$\alpha_1(i) = \pi_i b(\omega_i, x_1) \quad (i = 1, 2, ..., c)$

Step 2 再帰的計算

$\alpha_t(j) = b(\omega_j, x_t) \sum_{i=1}^{c} \alpha_{t-1}(i) a_{ij} \qquad (t = 2, 3, ..., n) \quad (j = 1, 2, ..., c)$

Step 3 確率の計算

$P(\mathbf{x}) = \sum_{i=1}^{c} \alpha_n(i)$

5 Backward アルゴリズム

Forward アルゴリズムは $\alpha_t(i)$ から $P(\mathbf{x})$ を求めたが，$\beta_t(i)$ を用いても同様に $P(\mathbf{x})$ を求めることができる．これを**後向きアルゴリズム（Backward algorithm）**と呼ぶ.

$\alpha_t(i)$ と同様に，$\beta_t(i)$ も次式で示すように再帰的な計算方法で求めることができる.

$$\beta_t(i) = P(x_{t+1}x_{t+2}\cdots x_n | s_t = \omega_i) \tag{5.1}$$

$$= \sum_{s_{t+1}} P(x_{t+1}x_{t+2}\cdots x_n, s_{t+1} | s_t = \omega_i) \tag{5.2}$$

$$= \sum_{s_{t+1}} P(x_{t+1}x_{t+2}\cdots x_n | s_{t+1}, s_t = \omega_i) P(s_{t+1} | s_t = \omega_i) \tag{5.3}$$

隠れマルコフモデルの条件付き独立性 (4) より，

$$= \sum_{s_{t+1}} P(x_{t+1}x_{t+2}\cdots x_n | s_{t+1}) P(s_{t+1} | s_t = \omega_i) \tag{5.4}$$

$$= \sum_{s_{t+1}} P(x_{t+2}x_{t+3}\cdots x_n | s_{t+1}) P(x_{t+1} | s_{t+1}) P(s_{t+1} | s_t = \omega_i) \tag{5.5}$$

$$= \sum_{j=1}^{c} P(x_{t+2}x_{t+3}\cdots x_n | s_{t+1} = \omega_j) P(x_{t+1} | s_{t+1} = \omega_j) P(s_{t+1} = \omega_j | s_t = \omega_i) \tag{5.6}$$

式 (4.17) と $P(s_t = \omega_j | s_{t-1} = \omega_i) = a_{ij}$，$P(x_t | s_t = \omega_j) = b(\omega_j, x_t)$ より，

$$= \sum_{j=1}^{c} a_{ij} b(\omega_j, x_{t+1}) \beta_{t+1}(j) \tag{5.7}$$

よって，

$$\beta_t(i) = \sum_{j=1}^{c} a_{ij} b(\omega_j, x_{t+1}) \beta_{t+1}(j) \tag{5.8}$$

$$(t = 2, 3, ..., n) \quad (j = 1, 2, ..., c)$$

ただし，

$$\beta_n(i) = 1 \quad (i = 1, 2, ..., c) \tag{5.9}$$

である.

$\beta_t(i)$ を用いて $P(\mathbf{x})$ を求める方法を考えてみる．まず，式 (4.17) において，$t = 1$ とすることにより，

$$\beta_1(i) = P(x_2 x_3 \cdots x_n | s_1 = \omega_i) \tag{5.10}$$

が得られる．

$$P(x_1 x_2 \cdots x_n | s_1 = \omega_i) = P(x_1 | s_1 = \omega_i) P(x_2 x_3 \cdots x_n | s_1 = \omega_i) \tag{5.11}$$

$$= b(\omega_i, x_1)\beta_1(i) \tag{5.12}$$

であるから，求めたい $P(\mathbf{x})$ は

$$P(\mathbf{x}) = \sum_{i=1}^{c} P(x_1 x_2 \cdots x_n, s_1 = \omega_i) \tag{5.13}$$

$$= \sum_{i=1}^{c} P(s_1 = \omega_i) P(x_1 x_2 \cdots x_n | s_1 = \omega_i) \tag{5.14}$$

$$= \sum_{i=1}^{c} \pi_i b(\omega_i, x_1)\beta_1(i) \tag{5.15}$$

となる．

Backward アルゴリズムの手順を Algorithm 2 に示す．

Algorithm 2 Backward algorithm

Step 1 初期化

$\beta_n(i) = 1 \quad (i = 1, 2, ..., c)$

Step 2 再帰的計算

$$\beta_t(i) = \sum_{j=1}^{c} a_{ij} b(\omega_j, x_{t+1})\beta_{t+1}(j) \qquad (t = (n-1), ..., 2, 1) \quad (i = 1, 2, ..., c)$$

Step 3 確率の計算

$$P(\mathbf{x}) = \sum_{i=1}^{c} \pi_i b(\omega_i, x_1)\beta_1(i)$$

6 Baum-Welch アルゴリズム

ここではまず準備として，次のような $\gamma_t(i), \xi_t(i,j)$ を定義する．

$$\gamma_t(i) \stackrel{\text{def}}{=} P(s_t = \omega_i|\mathbf{x}) \tag{6.1}$$

$$\xi_t(i,j) \stackrel{\text{def}}{=} P(s_t = \omega_i, s_{t+1} = \omega_j|\mathbf{x}) \tag{6.2}$$

上式の $\gamma_t(i)$ は，観測結果 $\mathbf{x}$ が得られたという条件で，t 回目にサイコロ ω_i を取り出している確率である．よって，

$$\sum_{i=1}^{c} \gamma_t(i) = 1 \tag{6.3}$$

が成り立つ．

式 (6.1) に式 (4.15)，(4.31) を適用することにより，$\gamma_t(i)$ は

$$\begin{aligned}\gamma_t(i) &= P(s_t = \omega_i|\mathbf{x}) \\ &= \frac{P(\mathbf{x}, s_t = \omega_i)}{P(\mathbf{x})} \\ &= \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{c} \alpha_n(i)}\end{aligned} \tag{6.4}$$

と表される．

式 (6.2) は，観測結果 $\mathbf{x} = x_1 x_2 \cdots x_n$ が得られたという条件で，t 回目にサイコロ ω_i を，$(t+1)$ 回目にサイコロ ω_j を取り出している確率を表している．

また，以下の式が成り立つ．

$$\sum_{j=1}^{c} \xi_t(i,j) = \sum_{j=1}^{c} P(s_t = \omega_i, s_{t+1} = \omega_j|\mathbf{x}) \tag{6.5}$$

$$= P(s_t = \omega_i|\mathbf{x}) \tag{6.6}$$

$$= \gamma_t(i) \tag{6.7}$$

ここで，$\mathbf{x} = x_1 x_2 \cdots x_n$ が観測され，かつ t 回目にサイコロ ω_i を，$(t+1)$ 回目にサイコロ ω_j を取り出している確率 $P(\mathbf{x}, s_t = \omega_i, s_{t+1} = \omega_j)$ を求めると，

$$\begin{aligned}&P(\mathbf{x}, s_t = \omega_i, s_{t+1} = \omega_j) \\ &\quad = P(\mathbf{x}|s_t = \omega_i, s_{t+1} = \omega_j) \cdot P(s_t = \omega_i, s_{t+1} = \omega_j)\end{aligned}$$

隠れマルコフモデルの条件付き独立性 (6) より，

$$\begin{aligned}&= P(x_1 x_2 \cdots x_t|s_t = \omega_i) \cdot P(x_{t+1}|s_{t+1} = \omega_j) \cdot P(x_{t+2} \cdots x_n|s_{t+1} = \omega_j) \cdot P(s_{t+1} = \omega_j|s_t = \omega_i) \cdot P(s_t = \omega_i) \\ &= P(x_1 x_2 \cdots x_t, s_t = \omega_i) \cdot P(s_{t+1} = \omega_j|s_t = \omega_i) \cdot P(x_{t+1}|s_{t+1} = \omega_j) \cdot P(x_{t+2} \cdots x_n|s_{t+1} = \omega_j) \\ &= \alpha_t(i) a_{ij} b(\omega_j, x_{t+1}) \beta_{t+1}(j)\end{aligned} \tag{6.8}$$

となる．これより

$$\xi_t(i,j) = P(s_t = \omega_i, s_{t+1} = \omega_j|\mathbf{x}) \tag{6.9}$$

$$= \frac{P(\mathbf{x}, s_t = \omega_i, s_{t+1} = \omega_j)}{P(\mathbf{x})} \tag{6.10}$$

$$= \frac{\alpha_t(i)a_{ij}b(\omega_j, x_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{c}\alpha_n(i)} \tag{6.11}$$

が得られる．

これらの結果をもとに，最尤推定を用いてパラメータ推定を行う．推定すべきパラメータ $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{\pi}$ をまとめて

$$\boldsymbol{\theta} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}) \tag{6.12}$$

とする．

最尤推定により最適なパラメータ $\boldsymbol{\theta}$ を求めることは，$P(\mathbf{x};\boldsymbol{\theta})$ を $\boldsymbol{\theta}$ に関して最大化することである．そのためには EM アルゴリズムを適用し，Q 関数を最大化すれば良い．

$$Q(\boldsymbol{\theta}^0, \boldsymbol{\theta}) = \sum_{\mathbf{s}} P(\mathbf{s}|\mathbf{x};\boldsymbol{\theta}^0)\log P(\mathbf{x},\mathbf{s};\boldsymbol{\theta}) \tag{6.13}$$

ここで，$\boldsymbol{\theta}$ は $\boldsymbol{\theta}^0$ を更新した結果得られる新しいパラメータである．Q 関数を最大化するには，パラメータを繰り返し更新すればよい．

式 (6.13) の右辺の $P(\mathbf{x},\mathbf{s};\boldsymbol{\theta})$ は確率の乗法定理より，

$$P(\mathbf{x},\mathbf{s};\boldsymbol{\theta}) = P(\mathbf{s};\boldsymbol{\theta})P(\mathbf{x}|\mathbf{s};\boldsymbol{\theta}) \tag{6.14}$$

と表せる．両辺の対数をとると，

$$\log P(\mathbf{x},\mathbf{s};\boldsymbol{\theta}) = \log P(\mathbf{s};\boldsymbol{\theta}) + \log P(\mathbf{x}|\mathbf{s};\boldsymbol{\theta}) \tag{6.15}$$

ここで，式 (4.8) と式 (4.11) を用いると

$$= \log P(s_1) + \sum_{t=1}^{n-1}\log a(s_t, s_{t+1}) + \sum_{t=1}^{n}\log b(s_t, x_t) \tag{6.16}$$

が得られる．したがって，式 (6.13) より，

$$Q(\boldsymbol{\theta}^0, \boldsymbol{\theta}) = \sum_{\mathbf{s}} P(\mathbf{s}|\mathbf{x};\boldsymbol{\theta}^0)\log P(s_1) + \sum_{\mathbf{s}} P(\mathbf{s}|\mathbf{x};\boldsymbol{\theta}^0)\sum_{t=1}^{n-1}\log a(s_t, s_{t+1}) + \sum_{\mathbf{s}} P(\mathbf{s}|\mathbf{x};\boldsymbol{\theta}^0)\sum_{t=1}^{n}\log b(s_t, x_t) \tag{6.17}$$

$$= Q(\boldsymbol{\theta}^0, \boldsymbol{\pi}) + Q(\boldsymbol{\theta}^0, \boldsymbol{A}) + Q(\boldsymbol{\theta}^0, \boldsymbol{B}) \tag{6.18}$$

となる．ここで，

$$Q(\boldsymbol{\theta}^0, \boldsymbol{\pi}) \stackrel{\text{def}}{=} \sum_{\mathbf{s}} P(\mathbf{s}|\mathbf{x};\boldsymbol{\theta}^0)\log P(s_1) \tag{6.19}$$

$$Q(\boldsymbol{\theta}^0, \boldsymbol{A}) \stackrel{\text{def}}{=} \sum_{\mathbf{s}} P(\mathbf{s}|\mathbf{x};\boldsymbol{\theta}^0)\sum_{t=1}^{n-1}\log a(s_t, s_{t+1}) \tag{6.20}$$

$$Q(\boldsymbol{\theta}^0, \boldsymbol{B}) \stackrel{\text{def}}{=} \sum_{\mathbf{s}} P(\mathbf{s}|\mathbf{x};\boldsymbol{\theta}^0)\sum_{t=1}^{n}\log b(s_t, x_t) \tag{6.21}$$

と定義した．上の 3 式は，それぞれパラメータ $\boldsymbol{\pi}$, $\boldsymbol{A}$, $\boldsymbol{B}$ のみを含む．したがって，$Q(\boldsymbol{\theta}^0, \boldsymbol{\theta})$ を $\boldsymbol{\theta}$ について最大化するには，$Q(\boldsymbol{\theta}^0, \boldsymbol{\pi})$, $Q(\boldsymbol{\theta}^0, \boldsymbol{A})$, $Q(\boldsymbol{\theta}^0, \boldsymbol{B})$ を，それぞれパラメータ $\boldsymbol{\pi}$, $\boldsymbol{A}$, $\boldsymbol{B}$ について最大化すればよい．ここで，$\boldsymbol{\theta}^0$ は更新前のパラメータであるので，最大化に当たっては定数とみなしてよい．

$\boldsymbol{Q(\theta^0, A)}$ の最大化

$Q(\boldsymbol{\theta}^0, \boldsymbol{A})$ は，

$$Q(\boldsymbol{\theta}^0, \boldsymbol{A}) = \sum_{s_1} \cdots \sum_{s_n} P(s_1 \cdots s_n | \mathbf{x}; \boldsymbol{\theta}^0) \sum_{t=1}^{n-1} \log a(s_t, s_{t+1}) \tag{6.22}$$

と書ける．上式の加算部分に現れるペア (s_t, s_{t+1}) の中には，t が異なっていても同じ内容 (ω_i, ω_j) を持つものが含まれている．そこで．上式を $\log a_{ij} (= \log a(\omega_i, \omega_j))$ でくくりだして整理すると，

$$Q(\boldsymbol{\theta}^0, \boldsymbol{A}) = \sum_{i=1}^{c} \sum_{j=1}^{c} \left(\sum_{t=1}^{n-1} \sum_{\substack{\mathbf{s} \\ (s_t=\omega_i) \\ (s_{t+1}=\omega_j)}} P(s_1 \cdots s_n | \mathbf{x}; \boldsymbol{\theta}^0) \right) \log a_{ij} \tag{6.23}$$

となる．ただし，上式で $\mathbf{s}$ についての加算は，

$$\mathbf{s} = s_1 s_2 \cdots \overset{(t)}{\omega_i} \overset{(t+1)}{\omega_j} \cdots s_n \tag{6.24}$$

のように，$s_t = \omega_i$，$s_{t+1} = \omega_j$ を満たす全ての $\mathbf{s}$ について実行することを示している．ここで，式 (6.2) を用いると，式 (6.23) において，

$$\sum_{\substack{\mathbf{s} \\ (s_t=\omega_i) \\ (s_{t+1}=\omega_j)}} P(s_1 \cdots s_n | \mathbf{x}; \boldsymbol{\theta}^0) = P(s_t = \omega_i, s_{t+1} = \omega_j | \mathbf{x}; \boldsymbol{\theta}) \tag{6.25}$$

$$= \xi_t(i, j) \tag{6.26}$$

が成り立つ．この結果より，

$$Q(\boldsymbol{\theta}^0, \boldsymbol{A}) = \sum_{i=1}^{c} \sum_{j=1}^{c} \left(\sum_{t=1}^{n-1} \xi_t(i, j) \right) \log a_{ij} \tag{6.27}$$

$$= \sum_{i=1}^{c} \left(\sum_{j=1}^{c} c_{ij} \log a_{ij} \right) \tag{6.28}$$

となる．ここで，

$$c_{ij} \overset{\text{def}}{=} \sum_{t=1}^{n-1} \xi_t(i, j) \tag{6.29}$$

と定義した．上記 c_{ij} は，更新前のパラメータ $\boldsymbol{\theta}^0$ に基づいて計算されるので，$\boldsymbol{A}$ に関する最大化に際しては定数として扱える．

式 (6.28) を最大化するには，各 i ごとに，

$$\sum_{j=1}^{c} c_{ij} \log a_{ij} \qquad (i = 1, 2, ..., c) \tag{6.30}$$

を最大化すればよい．ここで，

$$\sum_{j=1}^{c} a_{ij} = 1 \qquad (i = 1, 2, ..., c) \tag{6.31}$$

が成立するので，式 (6.31) の条件の下で式 (6.30) を最大にする a_{ij} を推定値 $\hat{a}_{ij}$ とすると，以下の定理が使える．

定理 1

いま，n 個の正の定数 $w_1, w_2, ..., w_n$ がある．
ここで，n 個の変数 $x_1, x_2, ..., x_n$ $(0 < x_i < 1)$ が，拘束条件

$$\sum_{i=1}^{n} x_i = 1 \tag{6.32}$$

を満たしているものとする．このとき，

$$f(x_1, x_2, ..., x_n) = \sum_{i=1}^{n} w_i \log x_i \tag{6.33}$$

を最大にする x_i は次式で与えられる．

$$x_i = \frac{w_i}{\sum_{k=1}^{n} w_k} \qquad (i = 1, 2, ..., n) \tag{6.34}$$

この定理を用いると，

$$\hat{a}_{ij} = \frac{c_{ij}}{\sum_{i=1}^{c} c_{ij}} \tag{6.35}$$

である．この式の分母は，式 (6.7) を用いることにより，

$$\sum_{j=1}^{c} c_{ij} = \sum_{j=1}^{c} \sum_{t=1}^{n-1} \xi_t(i, j) = \sum_{t=1}^{n-1} \gamma_t(i) \tag{6.36}$$

であるので，式 (6.35) は下式のように書ける．

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{n-1} \xi_t(i, j)}{\sum_{t=1}^{n-1} \gamma_t(i)} \tag{6.37}$$

ここで，上式の妥当性について考察する．式 (6.1)，(6.2) で示した $\gamma_t(i)$ および $\xi_t(i, j)$ の定義より，上式の分母，分子は，観測結果 $\mathbf{x}$ が得られたという条件の下でそれぞれ以下の値を計算してい

ることがわかる．

$$\begin{aligned}\sum_{t=1}^{n-1}\gamma_t(i) &= \text{サイコロ}\omega_i\text{を取り出した回数の期待値}\\ &= \text{状態}\omega_i\text{となった回数の期待値}\\ \sum_{t=1}^{n-1}\xi_t(i,j) &= \text{サイコロ}\omega_i\text{の次に}\omega_j\text{を取り出した回数の期待値}\\ &= \text{状態}\omega_i\text{から状態}\omega_j\text{へ遷移した回数の期待値}\end{aligned}$$

したがって，a_{ij} の推定値として，式 (6.35) を用いるのは妥当である．

$\boldsymbol{Q(\theta^0, B)}$ の最大化

上で述べた方法と同様に，式 (6.21) についても，(s_t, x_t) のペアの中で，同じ内容 (ω_j, v_k) であるものをまとめるため，$\log b_{jk}$ でくくりだして整理すると，$Q(\boldsymbol{\theta}^0, \boldsymbol{B})$ は

$$Q(\boldsymbol{\theta}^0, \boldsymbol{B}) = \sum_{\boldsymbol{s}} P(\boldsymbol{s}|\mathbf{x};\boldsymbol{\theta}^0)\sum_{t=1}^{n}\log b(s_t, x_t) \tag{6.38}$$

$$= \sum_{s_1}\cdots\sum_{s_n} P(s_1\cdots s_n|x_1\cdots x_n;\boldsymbol{\theta}^0)\sum_{t\ 1}^{n}\log b(s_t, x_t) \tag{6.39}$$

$$= \sum_{j=1}^{c}\sum_{k=1}^{m}\left(\sum_{t=1}^{n}\delta(x_t, v_k)\sum_{\substack{\mathbf{s}\\(s_t=\omega_i)}} P(s_1\cdots s_n|x_1\cdots x_n;\boldsymbol{\theta}^0)\right)\log b_{jk} \tag{6.40}$$

と表される．ただし，上式で $\boldsymbol{s}$ についての加算は，$s_t = \omega_j$ を満たすすべての $\boldsymbol{s}$ について実行することを示している．また，$\delta(x_t, v_k)$ は，$P(s_1\cdots s_n|x_1\cdots x_n)$ の中から $x_t = v_k$ を満たすもののみを抽出するための項であり，

$$\delta(x_t, v_k) = \begin{cases} 1 & (x_t = v_k)\\ 0 & (otherwise)\end{cases} \tag{6.41}$$

である．式 (6.1) を用いると，式 (6.40) において，

$$\sum_{\substack{\mathbf{s}\\(s_t=\omega_i)}} P(s_1\cdots s_n|x_1\cdots x_n;\boldsymbol{\theta}^0) = P(s_t = \omega_j|\mathbf{x};\boldsymbol{\theta}^0) \tag{6.42}$$

$$= \gamma_t(j) \tag{6.43}$$

が成り立つ．この結果より，式 (6.40) は

$$Q(\boldsymbol{\theta}^0, \boldsymbol{B}) = \sum_{j=1}^{c}\sum_{k=1}^{m}\left(\sum_{t=1}^{n}\delta(x_t, v_k)\gamma_t(j)\right)\log b_{jk} \tag{6.44}$$

$$= \sum_{j=1}^{c}\left(\sum_{k=1}^{m} d_{jk}\log b_{jk}\right) \tag{6.45}$$

と表される．ここで，

$$d_{jk} \stackrel{\text{def}}{=} \sum_{t=1}^{n}\delta(x_t, v_k)\gamma_t(j) \tag{6.46}$$

と定義した．これまで同様，d_{jk} はパラメータ $\boldsymbol{\theta}^0$ にのみ依存するので，定数として扱える．式 (6.45) を最大化するためには，各 j ごとに

$$\sum_{k=1}^{m} d_{jk} \log b_{jk} \tag{6.47}$$

を最大化すればよい．出力確率 b_{jk} は

$$\sum_{k=1}^{m} b_{jk} = 1 \qquad (j = 1, 2, ..., c) \tag{6.48}$$

が成り立つので，式 (6.48) の条件のもとで式 (6.47) を最大にする b_{jk} を推定値 $\hat{b_{jk}}$ とすると，定理 1 より

$$\hat{b}_{jk} = \frac{d_{jk}}{\sum_{l=1}^{m} d_{jl}} \tag{6.49}$$

と求めることができる．式 (6.49) は

$$\sum_{l=1}^{m} d_{jl} = \sum_{l=1}^{m} \sum_{t=1}^{n} \delta(x_t, v_l)\gamma_t(j) \tag{6.50}$$

$$= \sum_{t=1}^{n} \gamma_t(j) \sum_{l=1}^{m} \delta(x_t, v_l) \tag{6.51}$$

$$= \sum_{t=1}^{n} \gamma_t(j) \tag{6.52}$$

である．よって推定値 $\hat{b}_{jk}$ は次のようになる．

$$\hat{b}_{jk} = \frac{\sum_{t=1}^{n} \delta(x_t, v_k)\gamma_t(j)}{\sum_{t=1}^{n} \gamma_t(j)} \tag{6.53}$$

$\boldsymbol{Q(\theta^0, \pi)}$ の最大化

これまでと同様に

$$Q(\boldsymbol{\theta}^0, \boldsymbol{\pi}) = \sum_{\mathbf{s}} P(\mathbf{s}|\mathbf{x}; \boldsymbol{\theta}^0) \log P(s_1) \tag{6.54}$$

$$= \sum_{s_1} \cdots \sum_{s_n} P(s_1 \cdots s_n|\mathbf{x}; \boldsymbol{\theta}^0) \log P(s_1) \tag{6.55}$$

$$= \sum_{i=1}^{c} \left(\sum_{\substack{\mathbf{s} \\ (s_1=\omega_i)}} P(s_1 \cdots s_n|\mathbf{x}; \boldsymbol{\theta}^0) \right) \log P(s_1 = \omega_i) \tag{6.56}$$

となる．ここで，

$$\sum_{\substack{\mathbf{s} \\ (s_1=\omega_i)}} P(s_1 \cdots s_n|\mathbf{x}; \boldsymbol{\theta}^0) = P(s_1 = \omega_i|\mathbf{x}; \boldsymbol{\theta}^0) \tag{6.57}$$

であり，

$$P(s_1 = \omega_i) = \pi_i \tag{6.58}$$

が成り立つので，これらを式 (6.56) に代入することにより

$$Q(\boldsymbol{\theta}^0, \boldsymbol{\pi}) = \sum_{i=1}^{c} P(s_1 = \omega_i | \mathbf{x}; \boldsymbol{\theta}^0) \cdot \log \pi_i \tag{6.59}$$

$$= \sum_{i=1}^{c} e_i \log \pi_i \tag{6.60}$$

を得る．ここで

$$e_i \stackrel{\text{def}}{=} P(s_1 = \omega_i | \mathbf{x}; \boldsymbol{\theta}^0) \tag{6.61}$$

と定義した．これも更新前のパラメータ $\boldsymbol{\theta}^0$ のみに依存し，定数として扱える．

初期確率 π_i は，

$$\sum_{i=1}^{c} \pi_i = 1 \tag{6.62}$$

である．式 (6.62) のもとで式 (6.60) を最大にする π_i を推定値 $\hat{\pi}_i$ とすると，ここでも定理 1 が使えるので，

$$\hat{\pi}_i = \frac{e_i}{\sum_{i=1}^{c} e_i} \tag{6.63}$$

として求められる．一方，式 (6.1) より

$$\gamma_1(i) = P(s_1 = \omega_i | \mathbf{x}; \boldsymbol{\theta}^0) \qquad (i = 1, 2, ..., c) \tag{6.64}$$

であるので，式 (6.61) は

$$e_i = \gamma_1(i) \tag{6.65}$$

と書け，これらより

$$\hat{\pi}_i = \frac{\gamma_1(i)}{\sum_{i=1}^{c} \gamma_1(i)} \tag{6.66}$$

$$= \gamma_1(i) \tag{6.67}$$

が得られる．ここで，

$$\sum_{i=1}^{c} \gamma_1(i) = 1 \tag{6.68}$$

を用いている．

以下，最尤推定により求めたパラメータの推定式である．

$$\begin{aligned}\hat{a}_{ij} &= \frac{\sum_{t=1}^{n-1}\xi_t(i,j)}{\sum_{t=1}^{n-1}\gamma_t(i)} \\ &= \frac{\sum_{t=1}^{n-1}\alpha_t(i)a_{ij}b(\omega_j, x_{t+1})\beta_{t+1}(j)}{\sum_{t=1}^{n-1}\alpha_t(i)\beta_t(i)} \end{aligned} \tag{6.69}$$

$$\begin{aligned}\hat{b}_{jk} &= \frac{\sum_{t=1}^{n}\delta(x_t, v_k)\gamma_t(j)}{\sum_{t=1}^{n}\gamma_t(j)} \\ &= \frac{\sum_{t=1}^{n}\delta(x_t, v_k)\alpha_t(j)\beta_t(j)}{\sum_{t=1}^{n}\alpha_t(j)\beta_t(j)} \end{aligned} \tag{6.70}$$

$$\begin{aligned}\hat{\pi}_i &= \gamma_1(i) \\ &= \frac{\alpha_1(i)\beta_1(i)}{\sum_{j=1}^{c}\alpha_n(j)} \end{aligned} \tag{6.71}$$

なお，

$$\delta(x_t, v_k) = \begin{cases} 1 & (x_t = v_k) \\ 0 & (otherwise) \end{cases}$$

である．

この式に含まれている $\alpha_t(i)$,$\beta_t(i)$ は a_{ij}, b_{jk} を用いて表される．すなわち，このパラメータ推定式の右辺には，推定すべき a_{ij}, b_{jk} が含まれている．したがって，上式は a_{ij}, b_{jk} を陽に求める形にはなっていない．しかし，本手法は EM アルゴリズムに則っているので，パラメータ $\mathbf{A}$, $\mathbf{B}$,$\boldsymbol{\pi}$ を適当な初期値に設定し，式 (6.69)〜(6.71) を反復的に計算することにより，より良い推定値が得られることが保証されている．この計算方法は**バウム・ウェルチアルゴリズム (Baum-Welch algorithm)** と呼ばれている [3, 4]．ただし，得られる解は最適解である保証はなく，一般的には局所的最適解である．

Baum-Welch アルゴリズムの手順を Algorithm 3 に示す．

Algorithm 3 Baum-Welch algorithm

Step 1 初期化

パラメータ a_{ij}, b_{jk},π_i に適当な初期値を与える．

Step 2 再帰的計算

式 (6.69)，(6.70)，(6.71) を用いて，$\hat{a}_{ij}$,$\hat{b}_{jk}$,$\hat{\pi}_i$ を計算する．

Step 3 パラメータの更新

パラメータを $a_{ij} = \hat{a}_{ij}$, $b_{jk} = \hat{b}_{jk}$,$\pi_i = \hat{\pi}_i$ により更新する．

Step 4 判定

式 (4.31) により対数尤度 $\log P(\mathbf{x})$ を計算する．パラメータ更新前の対数尤度と比べ，その増分が予め定めたしきい値より小さければ，収束したとみなし終了する．さもなければ，**Step 2** に戻って処理を続行する．

7　スケーリング

Baum-Welch アルゴリズムを実装する際には，アンダーフローを防ぐための措置が必要になる．観測回数 n が大きくなった場合には，$\alpha_t(i)$,$\beta_t(i)$ は極端に小さな値になるので，アンダーフローの傾向が顕著である．アンダーフローを防ぐための手段の一つが**スケーリング（scaling）**である [3]．スケーリングは $\alpha_t(i)$,$\beta_t(i)$ に適当な値を掛け，それらの値が 1 のオーダーに留まるようにする．スケーリングに関しては，参考文献 [2,5] を参照されたい．

参考文献

[1] 平岡和幸堀玄. プログラミングのための確率統計. オーム社, 2009.

[2] ビショップ CM. パターン認識と機械学習下:ベイズ理論による統計的予測. 丸善出版, 2008.

[3] 石井健一郎上田修功. 続・わかりやすいパターン認識: 教師なし学習入門. オーム社, 2014.

[4] Stephen E Levinson, Lawrence R Rabiner, and Man Mohan Sondhi. An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition. *Bell System Technical Journal, The*, Vol. 62, No. 4, pp. 1035–1074, 1983.

[5] Dawei Shen. Some mathematics for hmm, 2008.