再帰的学習によるEMアルゴリズムの加速

正　員　池田　思朗†

Acceleration of the EM algorithm

Shiro IKEDA,*Member* †

あらまし EM アルゴリズムはボルツマンマシンや確率的パーセプトロンなどの学習を始め，HMM やその他隠れた確率変数を持つ確率分布の学習に対して広く持ちいられている．このアルゴリズムは繰り返し演算により最尤推定を求めるものであり，計算量が少なく実現が容易だが，一般に収束が遅い．一方，統計学の分野で Fisher のスコアリング法と呼ばれる手法があり，これも同様のモデルに対して適用できる繰り返し演算である．スコアリング法は収束は速いが計算量が多く実現が難しい．本論文では EM アルゴリズムを再帰的に用いてスコアリング法を近似し，EM アルゴリズムを加速できることを示す．Louis や Meng and Rubin も同様のアプローチを行なっているが，本手法はそれらに比べ，計算量が少なく実現が容易である．計算機実験を交えて結果を示す．

キーワード EM アルゴリズム，Fisher のスコアリング法，最尤推定，Louis turbo

1　はじめに

EM(Expectation Maximization) アルゴリズム [9] は，直接観測できない確率変数をもつ確率モデルの最尤推定 (MLE: Maximum Likelihood Estimate) のために，Dempster ら [4] によって提案された．例えば，隠れた細胞があるボルツマンマシン [2] や確率的パーセプトロン，Mixture of Expert networks[5][6][7] などは直接観測できない確率変数を持ち，EM アルゴリズムを適用できる．また音声認識で広く使われている HMM (Hidden Markov Model)[10] にも適用され，大きな成功を納めている．

具体的な計算はモデル毎に異なるが，EM アルゴリズムは繰り返し演算で最尤推定を求める手法である．各繰り返しで行う演算は通常簡単であるが，収束は一般に遅い．EM アルゴリズムの加速を行うアルゴリズムは Louis による Louis Turbo [8] と呼ばれるものがあるが，理論的な導出に留まっている．また，具体的な計算法が Meng and Rubin によって提案されている [11] が，計算量が多く，適用できるモデルは少ないと考えられる．

一方，統計学の分野では，このような確率モデルの最尤推定を求める手法として Fisher のスコアリング法 (スコアリング法) [9] と呼ばれるアルゴリズムがある．スコアリング法も EM アルゴリズムと同様に繰り返し演算で最尤推定を行い，収束は EM アルゴリズムよりも速い．ただし，各繰り返し演算の計算は複雑であり，前にあげた神経回路網モデルや HMM などパラメータの数が多い場合，適用するのは難しい．

本論文では，再帰的に EM アルゴリズムを用い，EM アルゴリズムとスコアリング法を結びつけることができ，結果的に EM アルゴリズムを加速できることを示す．本アルゴリズムは 2 つの段階からなる．まず，与えられたデータを用いて通常の EM アルゴリズムを行う．次の段階では，与えられたデータではなく，モデル自身がデータを作り出し，そのデータを用いて EM アルゴリズムを行う．この 2 つの段階を通じて得たパラメータを用いると，単に EM アルゴリズムを行うよりも良いパラメータを作りだせる．以下，スコアリング法と EM アルゴリズムの関係を示し，提案するアルゴリズムの理論的導出を示す．さらに計算機実験の結果を示し，実際に本アルゴリズムにより EM アルゴリズムが加速できることを示す．

2　EM アルゴリズムとスコアリング法

2.1　EM アルゴリズム

ボルツマンマシン [2] や，確率的パーセプトロン [1] のパラメータを推定する場合を考えよう．これらのモデルは，確率変数を $x = (y, z)$ とし，y は観測できる確率変数 (出力細胞)，z は観測できない隠れた確率変数 (中間層の細胞の出力) と定義すれば，$p(x|\boldsymbol{\theta}) = p(y, z|\boldsymbol{\theta})$ と表せる．このようなモデルのパラメータ推定を行う際，我々が教師から得られるデータは観測できる確率変数 y についてのサンプル $\{y_1, \cdots, y_N\}$ のみである．この y についての経験分布を $\hat{q}(y) = \sum_{s=1}^{N} \delta(y_s)/N$ と定める．我々は $\hat{q}(y)$ から $\boldsymbol{\theta}$ を推定しなければならない．

†理化学研究所脳科学総合研究センター，埼玉県
Brain Science Institute, Institute for Phisical and Chemical Research (RIKEN)

$p(x|\boldsymbol{\theta})$ の y についての周辺分布は,

$$p(y|\boldsymbol{\theta}) = E_{p(z|\theta)}\left[p(x|\boldsymbol{\theta})\right] = \int p(x|\boldsymbol{\theta})d\mu(z)$$

と表される．$l(y|\boldsymbol{\theta}) = \log p(y|\boldsymbol{\theta})$ とすると対数尤度は,

$$\begin{aligned} L(Y^N|\boldsymbol{\theta}) &\stackrel{\text{def}}{=} \frac{1}{N}\sum_{s=1}^{N} l(y_s|\boldsymbol{\theta}) \\ &= \int \hat{q}(y) l(y|\boldsymbol{\theta}) d\mu(y) = E_{\hat{q}(y)}\left[l(y|\boldsymbol{\theta})\right], \end{aligned}$$

となる．最尤推定では尤度あるいは対数尤度 $L(Y^N|\boldsymbol{\theta})$ を最大にするパラメータ $\hat{\boldsymbol{\theta}}$ を求める．

$$\hat{\boldsymbol{\theta}} = \underset{\theta}{\operatorname{argmax}}\, L(Y^N|\boldsymbol{\theta}). \tag{1}$$

本論文で扱うモデルのように，観測できない確率変数 z がある場合，最尤推定を直接 (1) 式から求めるのは難しい．この場合に EM アルゴリズムを適用できる．

ここでは，$p(x|\boldsymbol{\theta})$ は指数型分布族であるとして扱う．指数型分布族とは，その確率分布関数が次のように表せるものをいう．

$$p(x|\boldsymbol{\theta}) = \exp\left(\sum_{i=1}^{n} \theta^i r_i(x) - k(\boldsymbol{r}(x)) - \psi(\boldsymbol{\theta})\right). \tag{2}$$

$\boldsymbol{\theta} = (\theta^1, \cdots, \theta^n)^T$ は自然母数と呼ばれ $\psi(\boldsymbol{\theta})$ はその関数である．また $\boldsymbol{r}(x) = (r_1(x), \cdots, r_n(x))^T$ であり $k(\boldsymbol{r}(x))$ はその関数である．様々なモデルが指数型分布族に含まれる．先に述べたボルツマンマシンや，確率的パーセプトロン，HMM も指数型分布族に属する [1][2]．ただし，たとえ $p(x|\boldsymbol{\theta})$ が指数型分布族であっても，y に関する周辺分布 $p(y|\boldsymbol{\theta})$ は必ずしも指数型分布族には属さない．

EM アルゴリズムは繰り返し演算で最尤推定を求めるアルゴリズムであり，ある初期パラメータ $\boldsymbol{\theta}_0$ からパラメータを更新していく．新しいパラメータ $\{\boldsymbol{\theta}_t\}$ $(t = 1, 2, 3, \cdots)$ を求める際には，次の 2 つの手続きを行う．

- Expectation-ステップ: $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t)$ を計算する

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t) &= E_{\hat{q}(y)p(z|y,\theta_t)}\left[l(y, z|\boldsymbol{\theta})\right] \\ &= \sum_{s=1}^{N} p(z|y, \boldsymbol{\theta}_t) l(y, z|\boldsymbol{\theta}) \end{aligned}$$

- Maximization-ステップ: $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t)$ を最大にするパラメータを求める．

$$\boldsymbol{\theta}_{t+1} = \underset{\theta}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t)$$

E-ステップ と M-ステップ の手続きを通じ，$\boldsymbol{\theta}_t$ から $\boldsymbol{\theta}_{t+1}$ を得るが，この新たなパラメータに関して尤度の値が大きくなっていることを示せる [4]，すなわち次式が成り立つ,

$$L(Y^N|\boldsymbol{\theta}_{t+1}) \geq L(Y^N|\boldsymbol{\theta}_t).$$

E- と M-ステップを繰り返すとパラメータは収束し，これが最尤推定であると考えられる．一回の E- M-ステップを通じて得られるパラメータについて，次の近似が得られる．証明に関しては [9](3.76),[12] を参照のこと．なお，この近似が成り立つのは，指数型分布族の場合のみで，曲指数型分布族では成り立たない (付録 A).

$$\boldsymbol{\theta}_{t+1} \simeq \boldsymbol{\theta}_t + {G_X}^{-1}(\boldsymbol{\theta}_t)\partial L(Y^N|\boldsymbol{\theta}_t). \tag{3}$$

ここで $\partial = (\partial_1, \cdots, \partial_n)^T = (\partial/\partial\theta^1, \cdots, \partial/\partial\theta^n)^T$ であり，$G_X(\boldsymbol{\theta}) = (g_{X\,ij}(\boldsymbol{\theta}))$ は確率分布 $p(x|\boldsymbol{\theta})$ の Fisher 情報行列である．定義は,

$$\begin{aligned} g_{X\,ij}(\boldsymbol{\theta}) &= E_{p(x|\theta)}\left[\partial_i l(x|\boldsymbol{\theta})\partial_j l(x|\boldsymbol{\theta})\right] \\ &= -E_{p(x|\theta)}\left[\partial_i\partial_j l(x|\boldsymbol{\theta})\right]. \end{aligned}$$

である．(3) 式から，EM アルゴリズムが G_X で定められる計量に基づき，その最急降下の方向にパラメータを更新していることが分かる．

2.2 スコアリング法との関係

次に，統計学において Fisher のスコアリング法と呼ばれる手法と EM アルゴリズムとの関係について述べる．スコアリング法も繰り返し演算によってパラメータを更新するが，その更新ルールは,

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + {G_Y}^{-1}(\boldsymbol{\theta}_t)\partial L(Y^N|\boldsymbol{\theta}_t), \tag{4}$$

と表される．スコアリング法は EM アルゴリズムよりも収束が速いことが知られている．これは (3) 式と (4) 式の係数行列 $G_X(\boldsymbol{\theta})^{-1}$ と $G_Y(\boldsymbol{\theta})^{-1}$ の差によって生じる．$G_Y(\boldsymbol{\theta}) = (g_{Y\,ij}(\boldsymbol{\theta}))$ も $G_X(\boldsymbol{\theta})$ と同様に Fisher 情報量行列であるが周辺分布 $p(y|\boldsymbol{\theta})$ の情報量行列である．

$$\begin{aligned} g_{Y\,ij}(\boldsymbol{\theta}) &= E_{p(y|\theta)}\left[\partial_i l(y|\boldsymbol{\theta})\partial_j l(y|\boldsymbol{\theta})\right] \\ &= -E_{p(y|\theta)}\left[\partial_i\partial_j l(y|\boldsymbol{\theta})\right]. \end{aligned}$$

$G_X(\boldsymbol{\theta})$ と $G_Y(\boldsymbol{\theta})$ との間には次の関係式が成り立つ．

$$\begin{aligned} -l(y|\boldsymbol{\theta}) &= -l(x|\boldsymbol{\theta}) + l(z|y, \boldsymbol{\theta}) \\ -E_{p(y|\theta)}\left[\partial_i\partial_j l(y|\boldsymbol{\theta})\right] &= -E_{p(x|\theta)}\left[\partial_i\partial_j l(x|\boldsymbol{\theta})\right] \\ &\quad + E_{p(x|\theta)}\left[\partial_i\partial_j l(z|y, \boldsymbol{\theta})\right] \\ G_Y(\boldsymbol{\theta}) &= G_X(\boldsymbol{\theta}) - G_{Z|Y}(\boldsymbol{\theta}) \end{aligned} \tag{5}$$

$G_{Z|Y} = (g_{Z|Y}{}_{ij}(\boldsymbol{\theta}))$ は次のように定まる条件付き Fisher 情報量行列である．

$$\begin{aligned} g_{Z|Y}{}_{ij}(\boldsymbol{\theta}) &= -E_{p(y|\theta)}\left[E_{p(z|y,\theta)}\left[\partial_i\partial_j l(z|y,\boldsymbol{\theta})\right]\right] \\ &= E_{p(y|\theta)}\left[g_{Z|y}{}_{ij}(\boldsymbol{\theta})\right]. \end{aligned}$$

G_Y，G_X，$G_{Z|Y}$ は一般に正定値対称行列である．

スコアリング法で用いる Fisher 情報量行列 ${G_Y}^{-1}$ は EM アルゴリズムの対象となる確率分布では直接求めることが難しい．本論文で提案するアルゴリズムは，EM アルゴリズムを用いてスコアリング法を近似しようという物である．理論的導出には次の定理が重要となる．

定理 1 ${G_Y}^{-1}$ は次のように $G_X, G_{Z|Y}$ によって展開できる．

$$ {G_Y}^{-1} = \left(I + \sum_{i=1}^{\infty}({G_X}^{-1}G_{Z|Y})^i\right){G_X}^{-1} \tag{6}$$

証明　(6) 式は，G_Y，G_X，$G_{Z|Y}$ の同時対角化により簡単に導かれる [9]．

この結果を用いると (4) 式は，

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t + {G_Y}^{-1}\partial L(Y^N|\boldsymbol{\theta}_t) \\ &= \boldsymbol{\theta}_t + {G_X}^{-1}\partial L(Y^N|\boldsymbol{\theta}_t) \\ &\quad + {G_X}^{-1}G_{Z|Y}{G_X}^{-1}\partial L(Y^N|\boldsymbol{\theta}_t) \\ &\quad + ({G_X}^{-1}G_{Z|Y})^2{G_X}^{-1}\partial L(Y^N|\boldsymbol{\theta}_t) \\ &\quad + \cdots \end{aligned} \tag{7}$$

と書き直せる．(3) 式と (7) 式を比べると，EM アルゴリズムはスコアリング法を G_X で展開したときの 1 次近似だとみなせる．

3　提案するアルゴリズム

前章で述べた通り，スコアリング法はパラメータ $\boldsymbol{\theta}$ を計量 G_Y に基づいて最急降下の方向に更新していく．これは通常 EM アルゴリズムよりも収束が速い．しかしながら，${G_Y}^{-1}$ の計算は HMM や混合分布など EM アルゴリズムの対象となるモデルでは簡単ではない．本論文では，EM アルゴリズムを再帰的に用いてスコアリング法を近似する手法を提案する．

ある $\boldsymbol{\theta}_t$ から一度 EM ステップを行い，パラメータを一度更新したとしよう．このとき得られた $\boldsymbol{\theta}_{t+1}$ は，一つの確率分布 $p(y|\boldsymbol{\theta}_{t+1})$ を与える．そこで，経験分布の $\hat{q}(y)$ の代わりに $p(y|\boldsymbol{\theta}_{t+1})$ を真の分布としてパラメータ $\boldsymbol{\theta}_t$ を EM ステップで更新する．もし $p(y|\boldsymbol{\theta}_{t+1})$ が連続分布の場合には $p(y|\boldsymbol{\theta}_{t+1})$ にしたがってデータを生成して，そのデータを用いて学習を行う．離散分布の場合には $p(y|\boldsymbol{\theta}_{t+1})$ そのものを真の分布として学習を行えば良い．EM ステップを 1 回行ったあとで得られたパラメータを $\bar{\boldsymbol{\theta}}_{t+1}$ とすると，この新たに得られたパラメータは $\boldsymbol{\theta}_t$ とも $\boldsymbol{\theta}_{t+1}$ とも異なる．提案するアルゴリズムでは，$\boldsymbol{\theta}_t$，$\boldsymbol{\theta}_{t+1}$，$\bar{\boldsymbol{\theta}}_{t+1}$ の 3 つのパラメータから，より良い推定量を作り出す (図 1)．これが提案するアルゴリズムの概要である．理論的な導出を示すため，まず $\bar{\boldsymbol{\theta}}_{t+1}$ の持つ性質を示す．

定理 2 $p(y|\boldsymbol{\theta}_{t+1})$ を真の分布とし，$\boldsymbol{\theta}_t$ から一度 *EM* ステップを行い，得られたパラメータを $\bar{\boldsymbol{\theta}}_{t+1}$ とする．このとき，$\bar{\boldsymbol{\theta}}_{t+1}$ には次の性質がある．

$$\bar{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}_t \simeq {G_X}^{-1}G_Y{G_X}^{-1}\partial L(Y^N|\boldsymbol{\theta}_t). \tag{8}$$

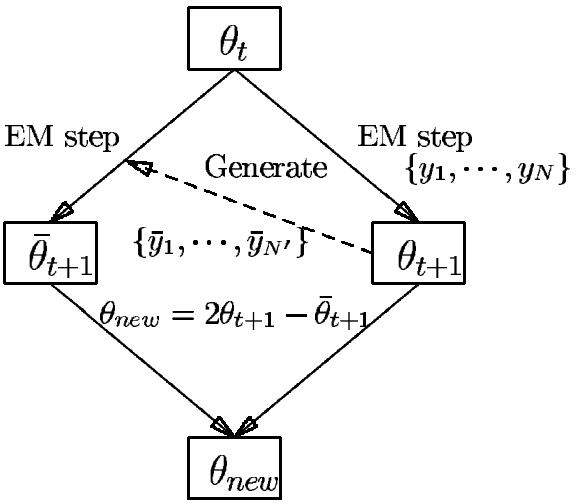


図 1: アルゴリズムの概要
Fig 1. Flowchart of the proposed algorithm

証明　付録 B を参照のこと．

(3) 式，(5) 式と (8) 式から，

$$\begin{aligned} & \bar{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}_t \\ \simeq\ & {G_X}^{-1}(G_X - G_{Z|Y}){G_X}^{-1}\partial L(Y^N|\boldsymbol{\theta}_t) \\ \simeq\ & (\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) - {G_X}^{-1}G_{Z|Y}{G_X}^{-1}\partial L(Y^N|\boldsymbol{\theta}_t) \end{aligned} \tag{9}$$

が得られる．スコアリング法の 2 次の項の近似は．

$$\begin{aligned} & {G_X}^{-1}G_{Z|Y}{G_X}^{-1}\partial L(Y^N|\boldsymbol{\theta}_t) \\ \simeq\ & (\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) - (\bar{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}_t) = \boldsymbol{\theta}_{t+1} - \bar{\boldsymbol{\theta}}_{t+1}, \end{aligned}$$

となる．スコアリング法の 2 次までの近似は，

$$\boldsymbol{\theta}' = 2\boldsymbol{\theta}_{t+1} - \bar{\boldsymbol{\theta}}_{t+1}$$

$$
\begin{aligned}
&= \boldsymbol{\theta}_t + (\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) + (\boldsymbol{\theta}_{t+1} - \bar{\boldsymbol{\theta}}_{t+1}) \\
&\simeq \boldsymbol{\theta}_t + {G_X}^{-1}(I + G_{Z|Y}{G_X}^{-1})\partial L(Y^N|\boldsymbol{\theta}_t),
\end{aligned}
$$

とすればよい．また，同様の手法を用いて更に高次までスコアリング法を近似できる．

系 1 $p(y|\bar{\boldsymbol{\theta}}_{t+i-1})$ を真の分布 *(*教師*)* として $\boldsymbol{\theta}_t$ から *EM* ステップを一回行い，得られたパラメータを $\bar{\boldsymbol{\theta}}_{t+i}$ とする *(*$i = 1, 2, \cdots,$ であり，$\bar{\boldsymbol{\theta}}_t = \boldsymbol{\theta}_{t+1}$ と定める*)*．$\bar{\boldsymbol{\theta}}_{t+i}$ は次の性質を持つ．

$$
\begin{aligned}
\bar{\boldsymbol{\theta}}_{t+i} - \boldsymbol{\theta}_t &\simeq ({G_X}^{-1}G_Y)^i {G_X}^{-1}\partial L(Y^N|\boldsymbol{\theta}_t) \\
&= (I - {G_X}^{-1}G_{Z|Y})^i {G_X}^{-1}\partial L(Y^N|\boldsymbol{\theta}_t)
\end{aligned}
$$

証明　定理 2 の証明と同じ方法で行えば良い (付録 B)．

この結果を用いると，$\bar{\boldsymbol{\theta}}_t$, $\cdots$, $\bar{\boldsymbol{\theta}}_{t+i}$, と $\boldsymbol{\theta}_t$ から，$({G_X}^{-1}G_{Z|Y})^i {G_X}^{-1}\partial L(Y^N|\boldsymbol{\theta}_t)$ が近似でき，スコアリング法を i 次まで近似できる．ただし，対象とするモデルが連続分布の場合，次章のシミュレーションのように Monte Carlo 的な手法を用いる必要があるため，2 次以上の近似は誤差が大きくなり，用いられない．

また，離散分布に対しては $i = n$ であれば，以下の理由からそれ以上の次数については線形演算で計算できる．$\boldsymbol{g}_i$ を次のように定める，

$$
\begin{aligned}
\boldsymbol{g}_0 &= {G_X}^{-1}\partial L(Y^N|\boldsymbol{\theta}) \\
&\vdots \\
\boldsymbol{g}_n &= ({G_X}^{-1}G_{Z|Y})^n {G_X}^{-1}\partial L(Y^N|\boldsymbol{\theta}).
\end{aligned}
$$

$\boldsymbol{\theta}$ は n 次元なので，$\boldsymbol{g}_1, \cdots, \boldsymbol{g}_n$ は線形従属となる．そこで，

$$
\boldsymbol{g}_n = a_1\boldsymbol{g}_1 + \cdots + a_{n-1}\boldsymbol{g}_{n-1}
$$

となる $a_1, \cdots, a_n$ を求めれば，

$$
\begin{aligned}
\boldsymbol{g}_{n+1} &= ({G_X}^{-1}G_{Z|Y})^{n+1} {G_X}^{-1}\partial L(Y^N|\boldsymbol{\theta}) \\
&= a_1\boldsymbol{g}_2 + \cdots + a_{n-1}\boldsymbol{g}_n
\end{aligned}
$$

となり，実際に EM ステップを行うまでもなく，線形演算でより高次の近似を順次求めることができる．このように，与えられたデータを用いて EM step を行った後，与えられたデータではなくデータを作り出しそれを学習すれば，より良いパラメータを求められる．以上が提案するアルゴリズムである．

4　シミュレーション

4.1　対数線形モデル

まず，対数線形モデルを用いた計算機実験の結果を示す．モデルは (図 2) (A, B, C) 3 つの確率変数を持っており，A,B,C はそれぞれ $\{A_i\}$, $\{B_j\}$, $\{C_k\}$ $(i = 1, \cdots, I,$ $j = 1, \cdots, J,$ $k = 1, \cdots, K)$ の値のどれかをとる．我々はそのうち，A, B の値を観測できるが，C (潜在変数) は観測できない．モデルの確率分布は，$P(A, B, C) = P(A_i|C_k)P(B_j|C_k)P(C_k)$ と定める．つまり観測できない変数 C の条件つきで A と B は独立だと仮定し，それぞれは多項分布に従うとする．

我々は，データから A, B についての周辺分布のみしか得られない．すなわち，$m_{ij} = n_{ij}/\sum_{i'j'} n_{i'j'}$ を得るだけである．ここで，n_{ij} は $(A = A_i, B = B_j)$ を観測した個数である．この周辺分布は，パラメータを用いて $P(A_i, B_j) = \sum_k P_{i|k}P_{j|k}P_k$ となる．得られた観測データ $m_{ij} = n_{ij}/\sum_{i'j'} n_{i'j'}$ から，潜在変数 C も含めてパラメータを推定しなければならない．ここで，EM アルゴリズムが適用できる．シミュレーションでは，$I = J = 5$,

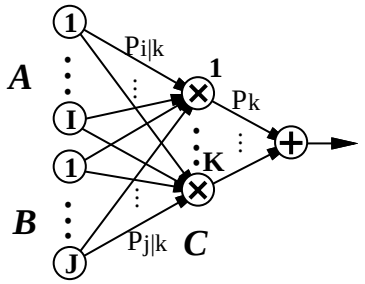


図 2: モデルの定義
Fig 2. Definition of the model

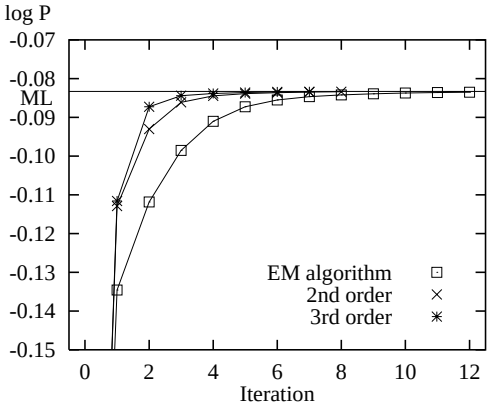


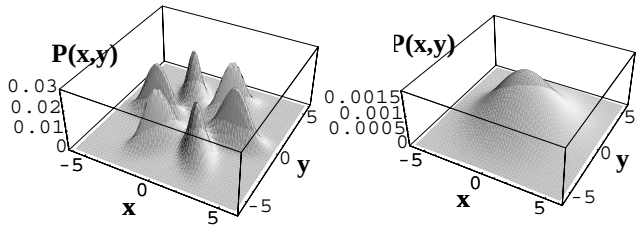
図 3: 対数尤度の増加の様子．計算量は考慮せず，EM アルゴリズム，及び提案するアルゴリズムの 1 ステップをそれぞれの横軸とした
Fig 3. The increase of the Log-likelihood

$K = 2$ とした．すなわち，求めたい周辺分布は $p(A_i, B_j)$ であり，これは要素が 25 の多項分布となる．もし，24 のパラメータを持っているとすれば，この分布を正確に

表現できるが，モデルは $(K-1)+K(I-1)+K(J-1)=17$ のパラメータしか持っておらず，完全には分布を表現できない．教師分布は乱数で作った多項分布を用い，モデルのパラメータをこの教師分布に合うように推定する．

図 3 は学習を通じての尤度の変化の結果である．提案した手法を用いて，スコアリング法を 2 次，そして 3 次まで近似し，学習を行った．図から，EM アルゴリズムに比べ，高次の近似を行った方が収束が速いことがわかる．

4.2 正規混合分布



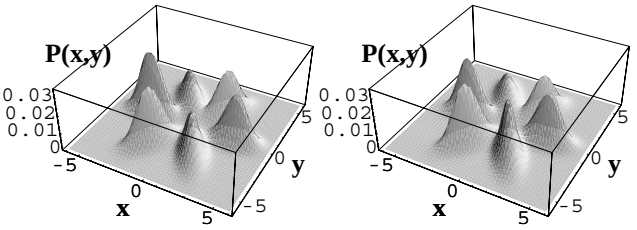
教師モデル (teacher model)
学習の初期分布 (initial model for learning)

図 4: 教師モデルと学習の初期分布
Fig 4. The teacher model and the initial model for learning

前節の実験で扱った分布は離散分布であり，EM ステップを行う際，分布からサンプルを得る必要は無く，その分布自体を用いれば良かった．しかし，分布が連続分布である場合には，実際にサンプルを作り出し，そのサンプルに対して EM ステップを行う，すなわち $p(y|\boldsymbol{\theta}_{t+1})$ から $\bar{\boldsymbol{\theta}}_{t+1}$ を求める際には，データ $\{\bar{y}_1,\cdots,\bar{y}_{N'}\}$ をサンプリングによって作り，それを用いて EM ステップを行う必要がある．このようなサンプリングを行う場合，実際にアルゴリズムがどのように働くかを知るために，ここでは 2 次元の正規混合分布 [13] を用いて実験を行った．図 4 に教師モデルと学習する際に用いた初期分布の密度関数を示す．両方とも 6 つの正規分布の重ね合わせで定義されている．ただし，初期分布ではそれぞれの分散が大きいので，個々の正規分布は全体の分布から区別できない．

EM アルゴリズムの具体的な形はここでは示さないが，混合正規分布に対する EM アルゴリズムは簡単で，その計算量は少ない．提案するアルゴリズムの有用性を示すために，次のように実験を行った．

1. 教師分布から y について 1000 個のサンプルを作る．モデルの初期分布のパラメータ $\boldsymbol{\theta}_0$ を定める．



EM アルゴリズム (EM algorithm)
提案するアルゴリズム (proposed algorithm)

図 5: 学習の結果
Fig 5. Results of learning

2. 教師分布から得られたデータを用いて，EM ステップを一回行い，$\boldsymbol{\theta}_t$ から $\boldsymbol{\theta}_{t+1}$ を得る．
3. 1000 個の新しいデータを $p(y|\boldsymbol{\theta}_{t+1})$ から生成する．
4. 新しく作られたデータを用いて，EM ステップを一回行い，$\boldsymbol{\theta}_t$ から $\bar{\boldsymbol{\theta}}_{t+1}$ を求める．
5. 新しいパラメータを $\boldsymbol{\theta}_{new}=2\boldsymbol{\theta}_{t+1}-\bar{\boldsymbol{\theta}}_{t+1}$ とし，$\boldsymbol{\theta}_t=\boldsymbol{\theta}_{new}$ と定めて，2 へもどる．

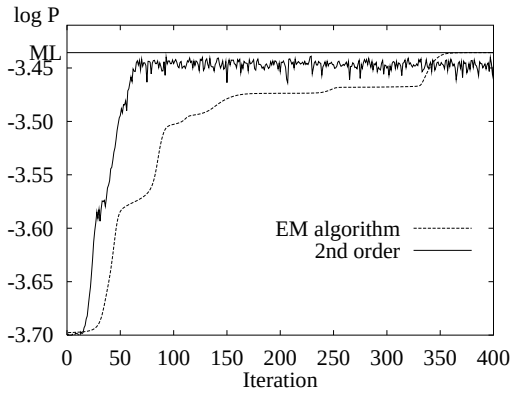


図 6: 対数尤度の変化
Fig 6. The transition of the Log-likelihood according to the iteration

EM アルゴリズムによって，そして提案するアルゴリズムによって学習の結果得られた密度関数がどのようになったかを図 5 に示す．さらに，学習の際に尤度がどのように変化していくかを図 6 に示す．提案するアルゴリズムはある種の Monte Carlo 法を用いているため，完全に収束すせず，絶えずふらついてる．このことから，より高次のスコアリング法の近似は行わなかった．図 6 の結果から，提案するアルゴリズムが EM を加速しているのがわかる．

ここで，もう一度計算量の点から提案するアルゴリズムの 1 ステップを見直してみる．提案するアルゴリズムの 1 ステップは実際には 2 ステップの EM アルゴリズムを含んでいる．もし，計算量も含めて EM アルゴリズムと速さを比べるのであれば，提案するアルゴリズムの横軸を変えて比べる方が適切であろう．図 7 に結果を示す．

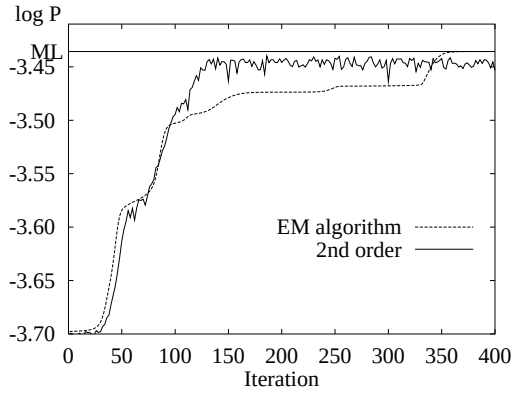


図 7: 計算量を考慮した対数尤度の変化

Fig 7. The transition of the Log-likelihood according to the iteration considering the amount of the calculation

この結果から分かるように，やはり提案するアルゴリズムの方が速く最尤推定に近付いている．

次に，後半のふらつきをなくし，収束させるために，提案するアルゴリズムを途中から通常の EM アルゴリズムに変えることを考える．ここでは，切替えるタイミングを決めるため，

$$\begin{aligned} \lambda(t) &= \eta\lambda(t-1) + (1-\eta)L(Y^N|\boldsymbol{\theta}_t), t = 1, \cdots, \\ \lambda(0) &= L(Y^N|\boldsymbol{\theta}_0) \end{aligned} \tag{10}$$

という関数を用いて $\lambda(t)$ の値が下がったら，通常の EM アルゴリズムに切替えることにした．なお，η は 0.7 とした．結果を図 8 に示す．この結果をみると，ほぼ 3 倍程度収束が速いことがわかる．提案したアルゴリズムと EM アルゴリズムを組み合わせることで，かなり速く収束するアルゴリズムを構成できる．

5 考察

実験を通じ，提案するアルゴリズムによって EM アルゴリズムを加速できることが示せた．ただし，計算量の意味で加速になっているかは問題による．提案するアルゴリズムでスコアリング法の 2 次の近似を得るため必要

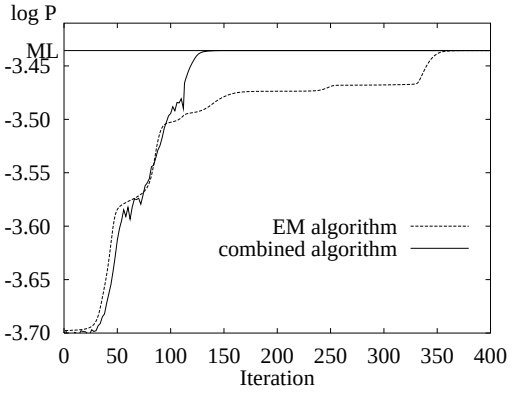


図 8: 提案するアルゴリズムと EM アルゴリズムとを組合せた場合

Fig 8. Comnbining the proposed algorithm and the EM algorithm

な計算量は，EM ステップを 2 回行うのと同じ計算量である．したがって，もとの EM アルゴリズムの 2 倍以上速く収束することが期待されるがこれは一概には分からない．正規混合分布の実験では 2 倍以上速かったが，対数線形モデルではほとんど同じ程度の結果であった．

提案するアルゴリズムでスコアリング法を 2 次まで近似して得たパラメータを $\boldsymbol{\theta}_{new}$ とし，EM アルゴリズムを 2 回行って得たパラメータを $\boldsymbol{\theta}_{t+2}$ とする．一般に $\boldsymbol{\theta}_{new} \neq \boldsymbol{\theta}_{t+2}$ である．ここで $L(Y^N|\boldsymbol{\theta}_{t+2})$ と $L(Y^N|\boldsymbol{\theta}_{new})$ の大きさを比較しておこう．仮りに $\boldsymbol{\theta}_t$ $\bar{\boldsymbol{\theta}}_t$, $\boldsymbol{\theta}_{t+1}$, $\boldsymbol{\theta}_{t+2}$ が十分近いならば 3 節の結果から $L(Y^N|\boldsymbol{\theta}_{new})$ を $\boldsymbol{\theta}_t$ の周りで展開して，

$$\begin{aligned} & L(Y^N|\boldsymbol{\theta}_{new}) = L(Y^N|2\boldsymbol{\theta}_{t+1} - \hat{\boldsymbol{\theta}}_{t+1}) \\ \simeq\ & L(Y^N|\boldsymbol{\theta}_t) + \partial {L_t}^T G_X(\boldsymbol{\theta}_t)^{-1} \partial L_t \\ & + \partial {L_t}^T G_X(\boldsymbol{\theta}_t)^{-1} G_{Z|Y}(\boldsymbol{\theta}_t) G_X(\boldsymbol{\theta}_t)^{-1} \partial L_t. \end{aligned} \tag{11}$$

ここで，$L_t = L(Y^N|\boldsymbol{\theta}_t)$ とした．同様に，

$$\begin{aligned} & L(Y^N|\boldsymbol{\theta}_{t+2}) = L(Y^N|\boldsymbol{\theta}_{t+2} - \boldsymbol{\theta}_{t+1} + \boldsymbol{\theta}_{t+1}) \\ \simeq\ & L(Y^N|\boldsymbol{\theta}_t) + \partial {L_t}^T G_X(\boldsymbol{\theta}_t)^{-1} \partial L_t \\ & + \partial {L_t}^T G_X(\boldsymbol{\theta}_{t+1})^{-1} \partial L_t \\ & - \partial {L_t}^T G_X(\boldsymbol{\theta}_{t+1})^{-1} \mathcal{G}_Y(\boldsymbol{\theta}_t) G_X(\boldsymbol{\theta}_t)^{-1} \partial L_t \\ =\ & L(Y^N|\boldsymbol{\theta}_t) + \partial {L_t}^T G_X(\boldsymbol{\theta}_t)^{-1} \partial L_t \\ & + \partial {L_t}^T G_X(\boldsymbol{\theta}_{t+1})^{-1} \mathcal{G}_{Z|Y}(\boldsymbol{\theta}_t)^{-1} G_X(\boldsymbol{\theta}_t)^{-1} \partial L_t \end{aligned} \tag{12}$$

と書ける．ただし $\mathcal{G}_Y(\boldsymbol{\theta}_t) = -\sum_{i=1}^N \partial^2 l(y_i|\boldsymbol{\theta}_t)$, $\mathcal{G}_{Z|Y}(\boldsymbol{\theta}_t) = -\sum_{i=1}^N E_{p(z|y_i,\theta_t)}\left[\partial^2 l(z|y_i, \boldsymbol{\theta}_t)\right]$ とした．(11) 式と

(12) 式のどちらが大きいかは，一概には言えない．定性的には，収束点に近く真の分布とモデルが近くなれば $G_X(\boldsymbol{\theta}_{t+1})$ と $G_X(\boldsymbol{\theta}_t), \mathcal{G}_{Z|Y}(\boldsymbol{\theta}_t)$ と $G_{Z|Y}(\boldsymbol{\theta}_t)$ は近くなり，EM アルゴリズムの 2 ステップと提案したアルゴリズムはほとんど変わらないことが予想される．また，$\mathcal{G}_{Z|Y}(\boldsymbol{\theta}_t)$ と $G_{Z|Y}(\boldsymbol{\theta}_t)$ が $\boldsymbol{O}$ に近い場合，ほとんど差が無くなることが予想できる．これは y を観測することで z に関する情報がほとんど全て分ってしまう場合にあたる．例えば混合正規分布で，お互いの正規分布の平均の値が分散に比べ充分遠い場合，データを観測しただけでどの正規分布からの出力かがほとんど明らかに分る．このような場合，EM アルゴリズムとスコアリング法との差が無くなってしまい，提案するアルゴリズムの 1 ステップは EM アルゴリズムの 2 ステップとほとんど同じになる．

一方，提案するアルゴリズムの方が EM アルゴリズムの 2 ステップよりも尤度の上昇が大きくなる条件を考えてみる．$G_X(\boldsymbol{\theta}_t)$ と $G_X(\boldsymbol{\theta}_{t+1})$ とがさほど変わらないのであれば，$\mathcal{G}_{Z|Y}(\boldsymbol{\theta}_t)$ の計量よりも $G_{Z|Y}(\boldsymbol{\theta}_t)$ の計量の方が大きい場合提案するアルゴリズムのほうが尤度を上昇させる．これは，観測データ $\{y_i\}$ と $p(y|\boldsymbol{\theta}_t)$ にしたがって生成された y では，$\{y_i\}$ に対して のほうが z についての情報が明らかな場合にあたる．このように，計算量も含め提案したアルゴリズムが EM アルゴリズムよりも常に速く収束するかは一概には明らかではない．これはより高次の近似を行う場合でも同様である．

連続の分布の場合，Monte Carlo 的な操作のため，尤度 $L(Y^N|\boldsymbol{\theta}_{new})$ が耐えずふらついてしまう．付録 B の結果を用いてこの分散を推定する．Monte Carlo 法で N' 個サンプルを発生させた場合，$\boldsymbol{\theta}_{new}$ は，正しい推定値 $\boldsymbol{\theta}^*_{new}$ を平均に ${G_X}^{-1}G_Y(\boldsymbol{\theta}_{t+1}){G_X}^{-1}/N'$ 程度の分散を持つ．$L_{new} = L(Y^N|\boldsymbol{\theta}_{new}), L^*_{new} = L(Y^N|\boldsymbol{\theta}^*_{new})$ と書き，高次の項を無視して，L_{new} を $\boldsymbol{\theta}^*_{new}$ のまわりで展開する．充分 N' が大きい場合，$\boldsymbol{\theta}_{new} - \boldsymbol{\theta}^*_{new}$ は充分小さいと考えられることから，このような近似を用いる．

$$L_{new} \simeq L^*_{new} + \partial {L^*_{new}}^T(\boldsymbol{\theta}_{new} - \boldsymbol{\theta}^*_{new}) \quad (13)$$

L_{new} の平均はほぼ L^*_{new} である．一方，分散は $\partial {L^*_{new}}^T({G_X}^{-1}G_Y(\boldsymbol{\theta}_{t+1}){G_X}^{-1})\partial L^*_{new}/N'$ である． この結果から，分散は ∂L^*_{new} を ${G_X}^{-1}G_Y(\boldsymbol{\theta}_{t+1}){G_X}^{-1}/N'$ を計量として計った長さに等しいことがわかる．これは N' に反比例する．問題によってどの程度のサンプル数が必要かを考える必要がある．

本論文では η を用い，学習法を切り替える手法を提案した．$L(Y^N|\theta_t)$ の分散が一定 で t が大きければ $\lambda(t)$ の分散の期待値は最終的に元の分散の $(1-\eta)^2/(1-\eta^2)$ 倍となる．一回のパラメータ更新での尤度の増加分は (11) 式の通りである．この増加分が分散の $(1-\eta)^2/(1-\eta^2)$ 倍程度になったときに，アルゴリズムを切り替えるというのが提案した手法である．例えば $\eta = 0.7$ とすると，$(1-\eta)^2/(1-\eta^2) = 9/51$ であり，尤度の増加分がおおよそ分散の 1/5 程度になったときにアルゴリズムが確率的に切りかわることになる．

6 まとめ

EM アルゴリズムの加速に関しては様々な方法が提案されている．基本的には本論文と同様のスコアリング法の近似を用いているが，表記が異なる．多くの場合 EM ステップを $\boldsymbol{\theta}_{t+1} = EM(\boldsymbol{\theta}_t)$ という関数であると定義し，このヤコビアン J と $(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t)$ を用いて，スコアリング法の最急降下の方向を展開し，EM アルゴリズムの加速としている．この J は，本論文での定義では ${G_X}^{-1}G_{Z|Y}$ と近似的に等しい．この J を求める単純なものとしては Aitken 加速がある．これは関数 $\boldsymbol{\theta}_{t+1} = EM(\boldsymbol{\theta}_t)$ から直接そのヤコビアンとして J を求める手法である [9]．ヤコビアンを求めるのに必要な計算量は EM アルゴリズムの 1 ステップと同じ程度であるとすれば，2 次の近似を行うのに対し，用いる計算量は同じ程度である．ただし，このようにして求めた J の固有値は必ずしも 0 と 1 の間に存在しない．

これを改良するものとして Louis による Louis Turbo という手法が有名である [8]．Louis Tourbo では J の具体的な計算手順は与えられていない．これに対し Meng と Rubin は EM アルゴリズムを使って J を計算する方法を提案している [11]．彼らの方法では J を求めるために，EM ステップをパラメータの数だけ行う．一度 J を求めてしまえば，スコアリング法を何次までででも近似できるが，2 次の近似を求めるためにもパラメータ数分の EM ステップを行う必要がある．一方，本論文で提案した手法では，2 次の近似を求めるためには，EM アルゴリズムを 2 回行えば良く，高次の場合でも，それがパラメータ数以下ならば次数と同じ回数の EM を行えば良いだけである．それ以上の場合には単なる線形演算を行えば良く，Meng と Rubin の手法と比べ，パラメータ数よりも低い次数の近似を得たいのならば計算量は少く，大きい次数の近似には同じ計算量が必要となる．したがって Meng と Rubin の手法と比べても本手法の方が計算量が少なくてすむのである．

今後の課題として，このアルゴリズムをニューラルネットワークの学習や HMM，さらに on-line 学習に用いていくつもりである．

謝辞

本研究に対して有益な助言を下さった理化学研究所脳科学総合研究センターの甘利俊一ディレクター，村田昇氏に感謝致します．

参考文献

[1] S. Amari. “Dualistic geometry of the manifold of higher-order neurons.” *Neural Networks*, Vol. 4, No. 4, pp. 443–451, 1991.

[2] S. Amari, K. Kurata, and H. Nagaoka. “Information geometry of Boltzmann machines.” *IEEE Trans. Neural Networks*, Vol. 3, No. 2, pp. 260–271, March, 1992.

[3] S. Amari. “Information Geometry of the EM and em Algorithms for Neural Networks.” *Neural Networks*, Vol. 8, No. 9, pp. 1379–1408, 1995.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. “Maximum likelihood from incomplete data via the EM algorithm.” *J. R. Statistical Society, Series B*, Vol. 39, pp. 1–38, 1977.

[5] R. A. Jacobs and M. I. Jordan. “Adaptive mixtures of local experts.” *Neural Computation*, Vol. 3, No. 1, pp. 79–87, Spring 1991.

[6] M. I. Jordan and R. A. Jacobs. “Hierarchical mixtures of experts and the EM algorithm.” *Neural Computation*, Vol. 6, No. 2, pp. 181–214, March 1994.

[7] M. I. Jordan and L. Xu. “Convergence results for the EM approach to mixture of experts architectures.” *Neural Networks*, Vol. 8, No. 9, pp. 1409–1431, 1995.

[8] T. A. Louis. “Finding the observed information matrix when using the EM algorithm.” *J. R. Statistical Society, Series B*, Vol. 44, No. 2, pp. 226–233, 1982.

[9] G. J. McLachlan and T. Krishnan. “*The EM Algorithm and Extensions.*” Wiley series in probability and statistics. John Wiley & Sons, Inc., 1997.

[10] L. R. Rabiner, S. E. Levinson, and M. M. Sondhi. “On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition.” *The Bell System Technical Journal*, Vol. 62, No. 4, pp. 1075–1105, April 1983.

[11] M. A. Tanner. “*Tools for Statistical Inference – Observed Data and Data Augmentation Methods*,” Vol. 67 of *Lecture Notes in Statistics*. Springer-Verlag, 1991.

[12] D. M. Titterington. “*Recursive Parameter Estimation using Incomplete Data*,” *J. R. Statist. Soc. B*, Vol. 46, No. 2, pp. 257–267, 1984.

[13] L. Xu and M. I. Jordan. “On convergence properties of the EM algorithm for Gaussian mixture.” A.I.Memo No.1520, C.B.C.L. Paper No.111, 1995.

A　曲指数型分布族

曲指数型分布族では，一般に (3) 式の近似は正しくない．(3) 式の証明には，E ステップで定義された $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t)$ について，

$$\begin{aligned}
\partial\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t)\Big|_{\theta=\theta_t} &= E_{\hat{q}(y)p(z|y,\theta_t)}\left[\partial\partial l(y, z|\boldsymbol{\theta}_t)\right] \\
&= E_{\hat{q}(y)p(z|y,\theta_t)}\left[-\partial\partial \psi(\boldsymbol{\theta}_t)\right] \\
&= -\partial\partial\psi(\boldsymbol{\theta}_t) \\
&= -G_X(\boldsymbol{\theta}_t). \qquad (14)
\end{aligned}$$

という指数型分布族に対する事実を用いる．しかし，これは曲指数型分布族では一般に正しくない．n 個のパラメータからなる $\boldsymbol{\theta}$ が $\boldsymbol{u} = (u^1, \cdots, u^m)$, の関数であり $(\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{u}))$, $m < n$ だとする．

$$\begin{aligned}
&\frac{\partial^2 Q(\boldsymbol{u}, \boldsymbol{u}_t)}{\partial u^k \partial u^l}\Big|_{u=u_t} \\
= \; & E_{\hat{q}(y)p(z|y,u_t)}\left[\frac{\partial^2 l(y, z|\boldsymbol{u})}{\partial u^k \partial u^l}\Big|_{u=u_t}\right] \\
= \; & \sum_i \frac{\partial^2 \theta^i(\boldsymbol{u})}{\partial u^k \partial u^l} E_{\hat{q}(y)p(z|y,u_t)}\left[r_i(x) - \partial_i \psi(\boldsymbol{\theta}(\boldsymbol{u_t}))\right] \\
& - \frac{\partial^2 \psi(\boldsymbol{\theta}(\boldsymbol{u}))}{\partial u^k \partial u^l}\Big|_{u=u_t}. \qquad (15)
\end{aligned}$$

(15) 式の第 1 項は一般には 0 とならず, (14) 式のように Fisher 情報量行列とは等しくならない．ただし，$\boldsymbol{\theta}$ が $\boldsymbol{u}$ の線形関数の場合には, (15) 式の第 1 項は 0 となり，(15) 式の近似は正しい．

B　定理 2 の証明

(3) 式より,

$$\begin{aligned}\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t &\simeq {G_X}^{-1}\partial L(Y^N|\boldsymbol{\theta}_t)\\ &= {G_X}^{-1}\partial\left(E_{\hat{q}(y)}\left[l(y|\boldsymbol{\theta})\right]\right)\Big|_{\theta=\theta_t}\end{aligned}$$

と書ける．$\hat{q}(y)$ を $p(y|\boldsymbol{\theta}_{t+1})$ と置き換え，(3) 式の導出と同様の手続きを行うと,

$$\begin{aligned}&\bar{\boldsymbol{\theta}}_{t+1}-\boldsymbol{\theta}_t \simeq {G_X}^{-1}\partial\left(E_{p(y|\theta_{t+1})}\left[l(y|\boldsymbol{\theta})\right]\right)\Big|_{\theta=\theta_t}\\ =\ & {G_X}^{-1}\int p(y|\boldsymbol{\theta}_{t+1})\partial l(y|\boldsymbol{\theta})\Big|_{\theta=\theta_t}d\mu(y).\end{aligned}\tag{16}$$

ここで $p(y|\boldsymbol{\theta}_{t+1})$ を次のように展開する,

$$\begin{aligned}p(y|\boldsymbol{\theta}_{t+1}) &\simeq p(y|\boldsymbol{\theta}_t)\\ &\quad +p(y|\boldsymbol{\theta}_t)\left(\partial l(y|\boldsymbol{\theta}_t)\right)^T(\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t)\end{aligned}$$

この結果を用いると (16) 式は次のように近似できる．

$$\begin{aligned}&\bar{\boldsymbol{\theta}}_{t+1}-\boldsymbol{\theta}_t\\ \simeq\ & {G_X}^{-1}\int\Big(p(y|\boldsymbol{\theta}_t)\partial l(y|\boldsymbol{\theta}_t)\\ &+p(y|\boldsymbol{\theta}_t)\partial l(y|\boldsymbol{\theta}_t)\partial l(y|\boldsymbol{\theta}_t)^T(\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t)\Big)d\mu(y)\\ =\ & {G_X}^{-1}\left(\int p(y|\boldsymbol{\theta}_t)\partial l(y|\boldsymbol{\theta}_t)\partial l(y|\boldsymbol{\theta}_t)^T d\mu(y)\right)\\ &\cdot(\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t)\\ =\ & {G_X}^{-1}G_Y(\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t)\\ \simeq\ & {G_X}^{-1}G_Y{G_X}^{-1}\partial L(Y^N|\boldsymbol{\theta}_t).\end{aligned}$$

ゆえに (8) 式を得る．ここでは次の結果を用いた,

$$\int p(y|\boldsymbol{\theta}_t)\partial l(y|\boldsymbol{\theta}_t)d\mu(y)=0.$$

また，連続の分布の場合，提案するアルゴリズムでは Monte Carlo 法を用いたが $\bar{\boldsymbol{\theta}}_{t+1}$ が Monte Carlo 法の影響で一点に定まらない．漸近的な $\bar{\boldsymbol{\theta}}_{t+1}$ の分布を示しておく．今 $p(y|\boldsymbol{\theta}_{t+1})$ にしたがって，サンプルを N' 個生成したとする $\{\bar{y}_1,\cdots,\bar{y}_{N'}\}$．$\hat{p}(y|\boldsymbol{\theta}_{t+1})$ を次のように定める．

$$\hat{p}(y|\boldsymbol{\theta}_{t+1})=\frac{1}{N'}\sum_{i=1}^{N'}\delta(y-\bar{y}_i)$$

また $\boldsymbol{\theta}^*_{t+1}$ を $\hat{p}(y|\boldsymbol{\theta}_{t+1})$ に対する最尤推定点とする．これらを用いて 2 次まで (16) 式を展開する．

$$\begin{aligned}&\int\hat{p}(y|\boldsymbol{\theta}_{t+1})\partial l(y|\boldsymbol{\theta})\Big|_{\theta=\theta_t}d\mu(y).\\ =\ & E_{\hat{p}(y|\theta_{t+1})}\left[\partial l(y|\boldsymbol{\theta}^*_{t+1})\right] &(17)\\ & -E_{\hat{p}(y|\theta_{t+1})}\left[\partial^2 l(y|\boldsymbol{\theta}^*_{t+1})\right](\boldsymbol{\theta}^*_{t+1}-\boldsymbol{\theta}_t) &(18)\end{aligned}$$

(17) 式は 0 であり $E_{\hat{p}(y|\theta_{t+1})}\left[\partial^2 l(y|\boldsymbol{\theta}^*_{t+1})\right]$ は漸近的に $-G_Y(\boldsymbol{\theta}_{t+1})$ と等しく $\boldsymbol{\theta}^*_{t+1}$ は $\boldsymbol{\theta}_{t+1}$ を中心に，分散行列が $G_Y(\boldsymbol{\theta}_{t+1})^{-1}/N'$ の正規分布に従う．したがって，$\bar{\boldsymbol{\theta}}_{t+1}$ の分散行列は，${G_X}^{-1}G_Y(\boldsymbol{\theta}_{t+1}){G_X}^{-1}/N'$ 程度である．

池田　思朗

1997 東大大学院工学系研究科計数工学専攻博士課程了．同年理化学研究所入所．現在同研究所脳科学総合研究センターに所属．主に確率モデルの学習，構造選択の研究に従事．