再帰的学習による EM アルゴリズムの加速

池田 思朗

理化学研究所国際フロンティア情報表現研究チーム

〒 351-01　埼玉県和光市広沢 2-1　Email: shiro@irl.riken.go.jp

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 やその他隠れた構造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，統計学の分野でスコアリング法と呼ばれる手法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．これは収束は速いが計算量が多く実現が難しい．本研究では 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[7] や Meng and Rubin[10] 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．

EM アルゴリズム，スコアリング法，最尤推定，Louis turbo

Acceleration of the EM algorithm

Shiro Ikeda

Lab. For Info. Representation, Frontier Research Program, RIKEN

2-1 Hirosawa, Wako, Saitama 351-01　Email: shiro@irl.riken.go.jp

The EM algorithm is widely used for many applications including Boltzmann machine, mixture of expert networks and HMM. This algorithm gives an iterating procedure for calculating the MLE of stochastic models which has some hidden random variables. The calculation of the algorithm is simple, but usually the convergence speed is slow. We also have another algorithm called “the scoring method” in statistics. This method can also be applied to those models and the convergence speed is faster, but the calculation is usually very complicated. We show that by using the EM algorithm in a recursive way, we can connect these two method and accelerate the EM algorithm. Also Louis[7], Meng and Rubin[10] showed they can accelerate the EM algorithm, but our algorithm is simpler and easier. We show some results of the simulations using our algorithm.

EM algorithm, scoring method, maximum likelihood estimate, Louis turbo

1 はじめに

EM (Expectation Maximization) アルゴリズム [8] 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 [3] によって提案された．このアルゴリズムは現在では様々なモデルに応用されている．神経回路網モデルではボルツマンマシン [2] や確率的パーセプトロン，Mixture of Expert networks[4][5][6] に用いられ，また音声認識で広く使われている HMM (Hidden Markov Model)[9] にも適用され，大きな成功を納めている．

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り，通常，その演算は非常に簡単である．しかし，収束は一般に遅い．一方，統計学の分野では，同様に直接観測できない確率変数を持つ確率モデルの最尤推定を求める手法としてスコアリング法と呼ばれるアルゴリズムがある．スコアリング法は EM アルゴリズムより収束が速いが，その計算は複雑であり，大規模な神経回路網モデルや HMM に適用するのは難しい．スコアリング法とは別に，EM アルゴリズムの加速を行うアルゴリズムはいくつか提案されている [7][10] が，具体的な定式化はされていない．また，計算量が多く，適用できるモデルは少ないと考えられる．

本技術報告では，再帰的に EM アルゴリズムを用い，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タを作りだせる．本技術報告ではスコアリング法と EM アルゴリズムの関係を示しながら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

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，$p(x|\boldsymbol{\theta})$ と表せる．このようなモデルのパラメタ推定を行う際，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．

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

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

と表される．従って対数尤度は，

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

と表され，最尤推定ではこの対数尤度 $L(Y^N|\boldsymbol{\theta})$ を最大にするパラメタ $\hat{\boldsymbol{\theta}}$ を求めることになる．

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

本技術報告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 アルゴリズムが適用できる．

本技術報告では，$p(x|\boldsymbol{\theta})$ は指数分布属であるとして扱う．指数分布属とは，その確率分布関数が次のように表せるものをいう．

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

ここで，$\boldsymbol{\theta} = (\theta^1, \cdots, \theta^p)$ は自然母数と呼ばれる．また，$\boldsymbol{r}(x) = (r_1(x), \cdots, r_p(x))$ である．様々なモデルが指数分布属に含まれる．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 も指数分布属に属する [1][2]．ただし，たとえ $p(x|\boldsymbol{\theta})$ が指数分布属であったとしても，y に関する周辺分布 $p(y|\boldsymbol{\theta})$ は必ずしも指数分布には属さない．

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る．ある初期パラメタ $\boldsymbol{\theta}_0$ から $\{\boldsymbol{\theta}_t\}$ $(t = 1, 2, 3, \cdots)$ とパラメタを更新していく．それぞれ新しいパラメタを求める際には，次の 2 つの手続きを行う．

- Expectation-ステップ:
  $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t) = E_{\hat{q}(y)p(z|y,\boldsymbol{\theta}_t)}\left[l(y, z|\boldsymbol{\theta})\right]$ を計算する．
- Maximization-ステップ:
  $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t)$ を最大にするパラメタを求める．

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

この 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，この新たなパラメタに関して尤度の値が大きくなっていることを示せる [3]，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タは収束し，これが最尤推定であると考えられる．ここで，E- M- ステッ

プを通じて得られるパラメタについて，次の近似が得られる．証明に関しては付録 A.1 を参照のこと．

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

ここで $\partial = (\partial_1, \cdots, \partial_p)^T = (\partial/\partial\theta^1, \cdots, \partial/\partial\theta^p)^T$ 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，

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

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タを更新していることが分かる．ただし，この近似は指数分布属の自然母数 $\boldsymbol{\theta}$ に対してのみ成り立ち，曲指数分布属に付いては成り立たない (付録 A.2)．

次に，統計学において，スコアリング法と呼ばれる手法について述べる．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，

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

と表される．最尤推定を求める手法としては，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，それぞれで用いられる係数行列，$G_X(\boldsymbol{\theta})$ と $G_Y(\boldsymbol{\theta})$ 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，$p(y|\boldsymbol{\theta})$ の情報行列である．すなわち，見えない確率変数 z はここでは消えている．

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

$G_X(\boldsymbol{\theta})$ と $G_Y(\boldsymbol{\theta})$ とは異なる情報量行列であり，次の関係式が成り立つ．

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

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

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

これら G_Y，G_X，$G_{Z|Y}$ は一般に正定値対称行列である．

さて，スコアリングで用いる ${G_Y}^{-1}$ は EM アルゴリズムの対象となる確率分布では，直接求めるのが難しい．我々が提案するアルゴリズムでは，${G_Y}^{-1}$ を直接求めるのではなく，EM アルゴリズムを用いて間接的に求める．その際，次の関係式が重要になる．

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

(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 (付録 A.3)．同時対角化の結果，行列 ${G_{Z|Y}G_X}^{-1}$ の固有値が全て実数で，0 より大きく 1 よりも小さいことがわかる．従って $(I-{G_{Z|Y}G_X}^{-1})^{-1} = I+{G_{Z|Y}G_X}^{-1}+\cdots+({G_{Z|Y}G_X}^{-1})^i+\cdots$ と展開でき，(6) 式が得られる．

この結果を用いると，スコアリング法の (4) 式は，

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

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の 1 次近似だとみなせる．本技術報告で提案するアルゴリズムは，さらに (7) 式の高次の項を EM アルゴリズムを再帰的に用いて近似しようというものである．

3 提案するアルゴリズム

前章で述べた通り，スコアリング法では，パラメタ $\boldsymbol{\theta}$ 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 EM アルゴリズムの対象となるモデルでは通常簡単ではない．本技術報告では，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，ある一つの確率分布 $p(y|\boldsymbol{\theta}_{t+1})$ を与える．この確率分布 $p(y|\boldsymbol{\theta}_{t+1})$ からデータを生成したとしよう．この新たに作ったデータ $\{\bar{y}_1, \cdots, \bar{y}_{N'}\}$ を用いて，パラメタ $\boldsymbol{\theta}_t$ を EM ステップで更新してみる．EM ステップ を 1 回行った後で得られたパラメタを $\bar{\boldsymbol{\theta}}_{t+1}$ とすると，この新たに得られたパラメタは $\boldsymbol{\theta}_t$ とも $\boldsymbol{\theta}_{t+1}$ とも異なる．本アルゴリズムで提案する手法では，$\boldsymbol{\theta}_t$，$\boldsymbol{\theta}_{t+1}$，$\bar{\boldsymbol{\theta}}_{t+1}$ の 3 つのパラメタから，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 1 $p(y|\boldsymbol{\theta}_{t+1})$ を真の分布として，$\boldsymbol{\theta}_t$ から一度 *EM* ステップを行い，得られたパラメタを $\bar{\boldsymbol{\theta}}_{t+1}$ とする．このとき，$\bar{\boldsymbol{\theta}}_{t+1}$ には次の性質がある．

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

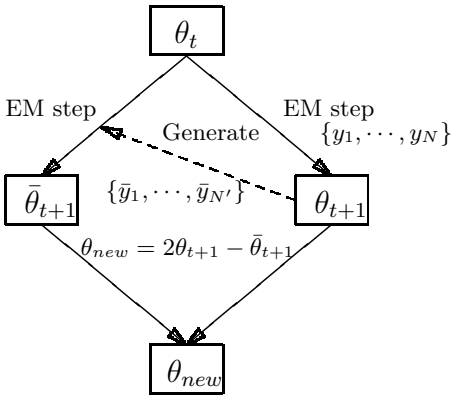

図 1: アルゴリズムの概要

証明　付録 A.4 に詳しく述べるが，概要は (3) 式の導出と同様の手続きをとればよい．教師分布 $\hat{q}(y)$ を $p(y|\boldsymbol{\theta}_{t+1})$ に変更するには，(3) 式の $L(Y^N|\boldsymbol{\theta}_t)$ を $E_{p(y|\boldsymbol{\theta}_{t+1})}[l(y|\boldsymbol{\theta}_t)]$ に変えると，

$$\bar{\boldsymbol{\theta}}_{t+1} \simeq \boldsymbol{\theta}_t + {G_X}^{-1} \int \partial l(y|\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t} p(y|\boldsymbol{\theta}_{t+1})\mu(y),$$

が得られる．ここで，

$$p(y|\boldsymbol{\theta}_{t+1}) \simeq p(y|\boldsymbol{\theta}_t) + p(y|\boldsymbol{\theta}_t)(\partial l(y|\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t})^T(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t),$$

と近似すると，証明が完了する．∎

(5) 式と (8) 式から，

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

が得られる．スコアリング法の 2 次の項の近似は，(3) 式と (9) 式から，

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

となる．従ってスコアリング法の 2 次までの近似は，

$$\begin{aligned}\boldsymbol{\theta}' &= 2\boldsymbol{\theta}_{t+1} - \bar{\boldsymbol{\theta}}_{t+1} \\ &= \boldsymbol{\theta}_t + (\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) + (\boldsymbol{\theta}_{t+1} - \bar{\boldsymbol{\theta}}_{t+1}) \\ &\simeq \boldsymbol{\theta}_t + {G_X}^{-1}(I + G_{Z|Y}{G_X}^{-1})\partial L(Y^N|\boldsymbol{\theta}_t),\end{aligned}$$

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次のようにして近似できる．

$p(y|\bar{\boldsymbol{\theta}}_{t+i-1})$ を真の分布 (教師) として，N' 個のサンプルを作る．そのデータを用いて $\boldsymbol{\theta}_t$ から EM ステップを一回行い，得られたパラメタを $\bar{\boldsymbol{\theta}}_{t+i}$ とする ($i = 0, 1, \cdots$, 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．

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

$\bar{\boldsymbol{\theta}}_t, \cdots, \bar{\boldsymbol{\theta}}_{t+i}$, 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，$i = p$ であれば，それ以上の次数に付いては線形演算で計算できる．これは [10] に示されている通りである．

提案するアルゴリズムが示すのは，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，より良いパラメタを求められるということである．

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．モデルに基づいてこの周辺分布を示すと $P(A_i, B_j) = \sum_k P_{i|k}P_{j|k}P_k$ 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，$I = J = 5$,

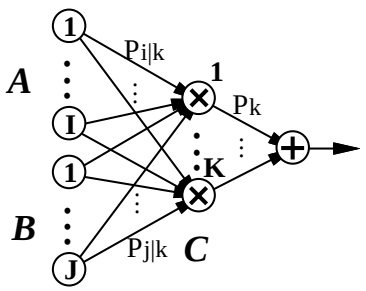

図 2: モデルの定義

$K = 2$ とした．すなわち，求めたい周辺分布は $p(A_i, B_j)$

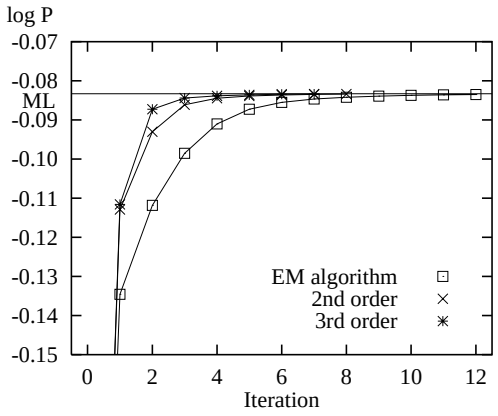

図 3: 対数尤度の増加の様子

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 $(K-1)+K(I-1)+K(J-1)=17$ のパラメタしか持っておらず，完全に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

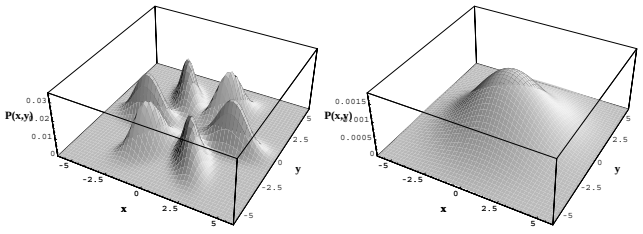

図 4: 教師モデルと学習の初期分布

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 $\{\bar{y}_1,\cdots,\bar{y}_{N'}\}$ を本当に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，ここでは正規混合分布 [11] を用いて実験を行った．

図 4 に教師モデルと学習する際の初期分布を示す．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，初期分布のではその分散が大きいので，それぞれの正規分布は全体の分布から区別できない．

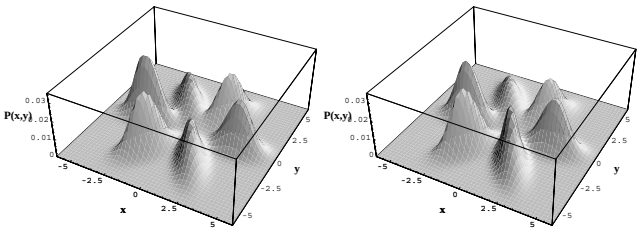

図 5: 学習の結果

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タ $\boldsymbol{\theta}_0$ を定める．
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．

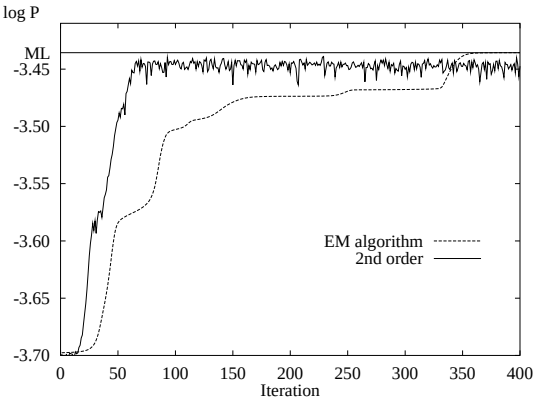

図 6: 対数尤度の変化

EM アルゴリズムにって，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確率分布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，我々は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．

5 まとめ

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，これが計算量の意味で加速になっているかとなると，それには問題がある．スコアリング法の 2 次の近似を得るため必要な計算量は，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，はたしてそうなるかはデータとモデルの形によって，一概には分からない．正規混合分布の実験では 2 倍以上収束が速かったが，対数線形モデルではほとんど同じ程度の結果であった．

EM アルゴリズムの加速に関しては，Louis による Louis Turbo という手法が有名である [7]．Louis も本技術報告にあるのと同様，スコアリング法の近似を行列を展開するような近似として与えている．ただし，$\boldsymbol{\theta}_{t+1} = EM(\boldsymbol{\theta})$ という関数を定義し，このヤコビアン J と $\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t$ を用いて，スコアリング法の展開をし，EM アルゴリズムの加速提案している．J は，本技術報告中の $J = (G_{Z|Y}{G_X}^{-1})$ と同値である．しかし，J を計算するのは通常容易でない．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が，彼らの方法では EM ステップをパラメタの数だけ行わなければならない．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タ数分の EM ステップ [10] を行う必要がある．一方，我々が提案した手法では，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，低い次数の近似を得たいのならば計算量は少く，パラメタ数よりも大きい次数の近似には同じ計算量が必要となる．

この提案するアルゴリズムは特に on-line での学習に役にたつと考えられる．on-line で，各時刻にデータが来ない場合を考えよう．もし，新たなデータを得たならば，EM algorithm アルゴリズムによってパラメタを更新すれば良い．しかし，しばらくの間データが無い時には，提案するアルゴリズムを用いて，データを作りながら，学習を継続すれば良いのである．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 on-line 学習に用いていくつもりである．

参考文献

[1] Shun-ichi Amari. Dualistic geometry of the manifold of higher-order neurons. *Neural Networks*, 4(4):443–451, 1991.

[2] Shun-ichi Amari, Koji Kurata, and Hiroshi Nagaoka. Information geometry of Boltzmann machines. *IEEE Transactions on Neural Networks*, 3(2):260–271, March 1992.

[3] 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*, 39:1–38, 1977.

[4] Robert A. Jacobs and Michael I. Jordan. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

[5] Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, March 1994.

[6] Michael I. Jordan and Lei Xu. Convergence results for the EM approach to mixture of experts architectures. *Neural Networks*, 8(9):1409–1431, 1995.

[7] Thomas A. Louis. Finding the observed information matrix when using the EM algorithm. *J. R. Statistical Society, Series B*, 44(2):226–233, 1982.

[8] Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions*. Wiley series in probability and statistics. John Wiley & Sons, Inc., 1997.

[9] 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*, 62(4):1075–1105, April 1983.

[10] Martin A. Tanner. *Tools for Statistical Inference – Observed Data and Data Augmentation Methods*, volume 67 of *Lecture Notes in Statistics*. Springer-Verlag, 1991.

[11] Lei Xu and Michael 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 付録

A.1 EM ステップの近似

EM の 1 ステップで得られた $\boldsymbol{\theta}_{t+1}$ は M ステップの定義から，

$$\partial Q(\boldsymbol{\theta},\boldsymbol{\theta}_t)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{t+1}}=0, \tag{10}$$

を満たす．もし $\boldsymbol{\theta}_{t+1}$ が $\boldsymbol{\theta}_t$ と近いならば，(10) 式を展開して，

$$\partial Q(\boldsymbol{\theta},\boldsymbol{\theta}_t)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t}\simeq-\partial\partial Q(\boldsymbol{\theta},\boldsymbol{\theta}_t)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t}(\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t), \tag{11}$$

が得られる．ここで，$\partial\partial=(\partial_i\partial_j)$ と定める．

さて，尤度関数 $L(Y^N|\boldsymbol{\theta})$ は，全ての $\boldsymbol{\theta}$ と $\boldsymbol{\theta}'$ に対して，

$$\begin{aligned}L(Y^N|\boldsymbol{\theta}) &= E_{\hat{q}(y)}\left[l(y|\boldsymbol{\theta})\right]\\ &= E_{\hat{q}(y)p(z|y,\boldsymbol{\theta}')}\left[l(y|\boldsymbol{\theta})\right]\\ &= E_{\hat{q}(y)p(z|y,\boldsymbol{\theta}')}\left[l(y,z|\boldsymbol{\theta})\right]\\ &\quad -E_{\hat{q}(y)p(z|y,\boldsymbol{\theta}')}\left[l(z|y,\boldsymbol{\theta})\right]\\ &= Q(\boldsymbol{\theta},\boldsymbol{\theta}')-H(\boldsymbol{\theta},\boldsymbol{\theta}').\end{aligned} \tag{12}$$

と書き直せる．$H(\boldsymbol{\theta},\boldsymbol{\theta}')$ は

$$\begin{aligned}H(\boldsymbol{\theta},\boldsymbol{\theta}') &= E_{\hat{q}(y)p(z|y,\boldsymbol{\theta}')}\left[l(z|y,\boldsymbol{\theta})\right]\\ &= \frac{1}{N}\sum_{s=1}^{N}\int l(z|y_s,\boldsymbol{\theta})p(z|y_s,\boldsymbol{\theta}')d\mu(z),\end{aligned}$$

と定義した．(12) 式の両辺を $\boldsymbol{\theta}$ で変微分し，$\boldsymbol{\theta}$ を $\boldsymbol{\theta}_t$ と等しくおくと，

$$\partial L(Y^N|\boldsymbol{\theta}_t)=\partial Q(\boldsymbol{\theta},\boldsymbol{\theta}_t)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t}-\partial H(\boldsymbol{\theta},\boldsymbol{\theta}_t)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t}, \tag{13}$$

となる．$H(\boldsymbol{\theta},\boldsymbol{\theta}_t)$ の変分は次式から明らかなように 0 となるので，

$$\begin{aligned}\partial H(\boldsymbol{\theta},\boldsymbol{\theta}_t)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t} &= E_{\hat{q}(y)p(z|y,\boldsymbol{\theta}_t)}\left[\partial l(z|y,\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_t}\right]\\ &= E_{\hat{q}(y)}\left[E_{p(z|y,\boldsymbol{\theta}_t)}\left[\partial l(z|y,\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_t}\right]\right]\\ &= E_{\hat{q}(y)}\left[\int\partial p(z|y_s,\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t}d\mu(z)\right]\\ &= 0,\end{aligned}$$

(13) 式から直ちに，

$$\partial L(Y^N|\boldsymbol{\theta}_t)=\partial Q(\boldsymbol{\theta},\boldsymbol{\theta}_t)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t}, \tag{14}$$

が得られる．(11) 式の左辺を (14) 式で置き換えると，

$$\partial L(Y^N|\boldsymbol{\theta}_t)\simeq-\partial\partial Q(\boldsymbol{\theta},\boldsymbol{\theta}_t)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t}(\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t),$$

となる．$p(x|\boldsymbol{\theta})$ が指数分布属であるので，(2) 式からも明らかなように $Q(\boldsymbol{\theta},\boldsymbol{\theta}_t)$ の 2 階微分は簡単な形に書き下せる，

$$\begin{aligned}\partial\partial Q(\boldsymbol{\theta},\boldsymbol{\theta}_t)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t} &= E_{\hat{q}(y)p(z|y,\boldsymbol{\theta}_t)}\left[\partial\partial l(y,z|\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t}\right]\\ &= E_{\hat{q}(y)p(z|y,\boldsymbol{\theta}_t)}\left[-\partial\partial\psi(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t}\right]\\ &= -\partial\partial\psi(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t}\\ &= -G_X(\boldsymbol{\theta}_t).\end{aligned} \tag{15}$$

これは $\boldsymbol{\theta}$ のみの関数であり，データとは関係ない．この結果を用いると (11) 式は

$$\partial L(Y^N|\boldsymbol{\theta}_t)\simeq G_X(\boldsymbol{\theta}_t)(\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t) \tag{16}$$

となり，(3) 式が得られる． ∎

A.2 曲指数分布属

曲指数分布属では，一般に A.2 の近似は正しくない．その理由は (15) 式が成り立たないからである．

$\boldsymbol{\theta}$ が $\boldsymbol{u}=(u^1,\cdots,u^q)$, の関数であり $\boldsymbol{\theta}=\boldsymbol{\theta}(\boldsymbol{u})$, $q<p$ だとする．

$$\begin{aligned}&\frac{\partial^2 Q(\boldsymbol{u},\boldsymbol{u}_t)}{\partial u^k\partial u^l}\Big|_{\boldsymbol{u}=\boldsymbol{u}_t}\\ &= E_{\hat{q}(y)p(z|y,\boldsymbol{u}_t)}\left[\frac{\partial^2 l(y,z|\boldsymbol{u})}{\partial u^k\partial u^l}\Big|_{\boldsymbol{u}=\boldsymbol{u}_t}\right]\\ &= \sum_i\frac{\partial^2\theta^i(\boldsymbol{u})}{\partial u^k\partial u^l}E_{\hat{q}(y)p(z|y,\boldsymbol{u}_t)}\left[r_i(x)-\partial_i\psi(\boldsymbol{\theta}(\boldsymbol{u}_t))\right]\\ &\qquad -\frac{\partial^2\psi(\boldsymbol{\theta}(\boldsymbol{u}))}{\partial u^k\partial u^l}\Big|_{\boldsymbol{u}=\boldsymbol{u}_t}.\end{aligned} \tag{17}$$

(17) 式の第 1 項は一般には 0 とならず, (15) 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, (17) 式の第 1 項は 0 となり，17 の近似は正しい．

A.3 スコアリング法の展開

まず，$G_Y, G_X, G_{Z|Y}$ が同時対角化可能なことを示す．

これらのは実対象正定値行列である．$\{\boldsymbol{e}_{X1},\cdots,\boldsymbol{e}_{Xp}\}$ を行列 G_X の固有ベクトルとし，それらを列ベクトルとする行列 $E_X=(\boldsymbol{e}_{X1},\cdots,\boldsymbol{e}_{Xp})$ を定義する．E_X を使って G_X を対角化する．

$$\begin{aligned}G_Y &= G_X-G_{Z|Y}\\ E_X^TG_YE_X &= \Lambda_X-E_X^TG_{Z|Y}E_X.\end{aligned} \tag{18}$$

Λ_X は，

$$\Lambda_X=\begin{pmatrix}\lambda_{X1} & & O\\ & \ddots & \\ O & & \lambda_{Xp}\end{pmatrix}.$$

である．$\Lambda_X^{-\frac{1}{2}}$ を次のように定め，

$$\Lambda_X^{-\frac{1}{2}} = \begin{pmatrix} \lambda_{X1}^{-\frac{1}{2}} & & O \\ & \ddots & \\ O & & \lambda_{X1}^{-\frac{1}{2}} \end{pmatrix},$$

この $\Lambda_X^{-\frac{1}{2}}$ を用いて，

$$\Lambda_X^{-\frac{1}{2}} E_X^T G_Y E_X \Lambda_X^{-\frac{1}{2}} = I - \Lambda_X^{-\frac{1}{2}} E_X^T G_{Z|Y} E_X \Lambda_X^{-\frac{1}{2}}.$$

とする．ここで I は p 次元の単位行列である．さて，$\Lambda_X^{-\frac{1}{2}} E_X^T G_{Z|Y} E_X \Lambda_X^{-\frac{1}{2}}$ もまた実対称正定値行列である．$E_{Z|Y}$ を $\Lambda_X^{-\frac{1}{2}} E_X^T G_{Z|Y} E_X \Lambda_X^{-\frac{1}{2}}$ の固有ベクトルでできた行列とする．これを用いて，3 つ全ての行列の対角化ができる．

$$\begin{aligned} & E_{Z|Y}^T \Lambda_X^{-\frac{1}{2}} E_X^T G_Y E_X \Lambda_X^{-\frac{1}{2}} E_{Z|Y} \\ & \quad = E_{Z|Y}^T E_{Z|Y} \\ & \qquad - E_{Z|Y}^T \Lambda_X^{-\frac{1}{2}} E_X^T G_{Z|Y} E_X \Lambda_X^{-\frac{1}{2}} E_{Z|Y} \\ & D_Y = D_X - D_{Z|Y}. \end{aligned} \tag{19}$$

$T = E_X \Lambda_X^{-\frac{1}{2}} E_{Z|Y}$ と定めると，

$$T^T G_Y T = D_Y, T^T G_X T = D_X, T^T G_{Z|Y} T = D_{Z|Y}.$$

$$D_Y = \begin{pmatrix} d_{Y1} & & O \\ & \ddots & \\ O & & d_{Yp} \end{pmatrix}. \quad (D_X,\ D_{Z|Y} \text{ も同様})$$

$G_Y, G_X, G_{Z|Y}$ が正定値であるので，$D_Y, D_X, D_{Z|Y}$ も全て正定値である．

さらに，${G_{Z|Y}G_X}^{-1}$ の固有値が全て 0 より大きく，1 より小さいことを示す．T の定義から，

$$\begin{aligned} T^T G_X T &= D_X \\ {G_X}^{-1} &= T D_X^{-1} T^T \\ G_{Z|Y} &= {T^{-1}}^T D_{Z|Y} T^{-1}. \end{aligned}$$

また，${G_{Z|Y}G_X}^{-1}$ の特性多項式から，

$$\begin{aligned} & \det\left|\lambda I - {G_{Z|Y}G_X}^{-1}\right| \\ = \ & \det\left|\lambda I - {T^{-1}}^T D_{Z|Y} T^{-1} T D_X^{-1} T^T\right| \\ = \ & \det\left|\lambda I - {T^{-1}}^T D_{Z|Y} D_X^{-1} T^T\right| \\ = \ & \det\left|{T^{-1}}^T (\lambda I - D_{Z|Y} D_X^{-1}) T^T\right| \\ = \ & \det\left|\lambda I - D_{Z|Y} D_X^{-1}\right|, \end{aligned}$$

${G_{Z|Y}G_X}^{-1}$ の固有値は全て $D_{Z|Y}D_X^{-1}$ の固有値と等しい．$D_{Z|Y}D_X^{-1}$ は，

$$D_{Z|Y} D_X^{-1} = \begin{pmatrix} d_{Y|Z1}/d_{X1} & & O \\ & \ddots & \\ O & & d_{Y|Zp}/d_{Xp} \end{pmatrix},$$

である．よって $d_{Z|Yi}/d_{Xi}$ $(i = 1, \cdots, p)$ が ${G_{Z|Y}G_X}^{-1}$ の固有値となる．$D_Y, D_X, D_{Z|Y}$ が正定値であることと (19) 式から，

$$\begin{aligned} d_{Yi}, d_{Xi}, d_{Z|Yi} &> 0, \quad d_{Yi} = d_{Xi} - d_{Z|Yi} > 0 \\ d_{Xi} &> d_{Z|Yi} > 0 \\ 1 &> \frac{d_{Z|Yi}}{d_{Xi}} > 0, \quad i = 1, \cdots, p. \end{aligned} \tag{20}$$

となり，全ての固有値が 0 より大きく 1 より小さいことを示せる．ある行列 A の最大固有値の大きさが 1 よりも小さい場合，$(I-A)^{-1} = I + A + \cdots + A^i + \cdots$ と展開できる．したがって次の結果を得る，

$$(I - G_{Z|Y}{G_X}^{-1})^{-1} = 1 + \sum_{i=1}^{\infty} (G_{Z|Y}{G_X}^{-1})^i.$$

A.4 定理 1 の証明

(3) 式より，

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

と書ける．$\hat{q}(y)$ を $p(y|\boldsymbol{\theta}_{t+1})$ と置き換え，A.1 と同じ手続きを行うと，

$$\begin{aligned} \bar{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}_t &\simeq {G_X}^{-1} \partial \left(E_{p(y|\boldsymbol{\theta}_{t+1})} \left[l(y|\boldsymbol{\theta}) \right] \right) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t} \\ &= {G_X}^{-1} \int \partial l(y|\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t} p(y|\boldsymbol{\theta}_{t+1}) \mu(y). \end{aligned} \tag{21}$$

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

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

この結果を用いると (21) 式は次のように近似できる．

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

ゆえに (8) 式を得る．ここでは次の結果を用いた，

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