隠れ状態最尤推定と反復解法
–EM アルゴリズムと Wake-Sleep アルゴリズム–

池田 思朗
科学技術振興事業団 さきがけ研究 21
埼玉県和光市広沢 2-1 理化学研究所 脳科学総合研究センター
Shiro.Ikeda@brain.riken.go.jp

1 はじめに

脳での情報処理は目や耳，皮膚を通じての入力，目や口や手足を動かす，声を出すなど，外部への出力がある．神経細胞においても感覚器からの入力を与えるものと外部への出力を与えるものが存在し，それらとは別に外とは直接は繋がっていない細胞がある．

情報処理の立場では，このような外部に直接繋がっている部分と外部からは直接観測できない部分が共存するシステムをどのように扱えるのだろうか．この質問への一つの回答は確率モデルにおける隠れ変数 (Latent Variable) の考え方である．本稿ではこの隠れ変数の定義と今までの神経回路網モデルにおけるその用いられ方，学習法を紹介する．

2 隠れ変数

ここでは，隠れ変数とはどういうものかを例を交えて説明する．

2.1 混合正規分布

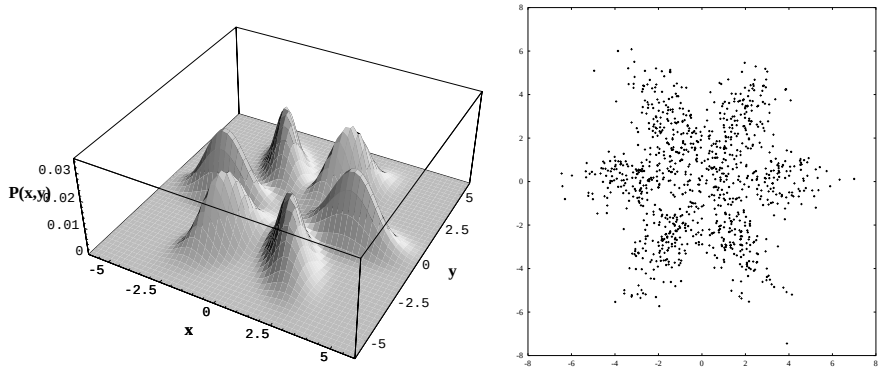

図 1: 正規混合分布

まずは，混合正規分布を考えよう．確率変数 $\boldsymbol{y}$ の平均が $\boldsymbol{\mu}$ で共分散行列が Σ の正規分布であるとき，その密度関数を $G(\boldsymbol{y};\boldsymbol{\mu},\Sigma)$ と書くことにする．混合正規分布はこの正規分布の混合分布として表わされる分布である．その密度関数 $p(y;\boldsymbol{\theta})$ (ただし $\boldsymbol{\theta}$ はモデルのパラメータ) は，ある π_i ($\sum_i \pi_i = 1$) を重み係数

として，

$$p(y;\boldsymbol{\theta}) = \sum_i \pi_i G(\boldsymbol{y};\boldsymbol{\mu}_i,\Sigma_i)$$

となる．このモデルは隠れ変数を持つ．ではその隠れ変数はなんだろうか，

2 次元の正規分布が 6 つ重なった正規混合分布を例に考えよう．確率分布の形を図 1 に示す．この確率分布からデータが得られているとする．この場合，出力されるデータのみからはどの正規分布によって出力されたかが分らない．すなわち，どの正規分布から出力されたかという情報は観測できない．この「どの正規分布から発生したか」という情報が隠れ変数となる．正規混合分布の場合は隠れている確率変数は離散的な確率変数である．

2.2 パーセプトロン

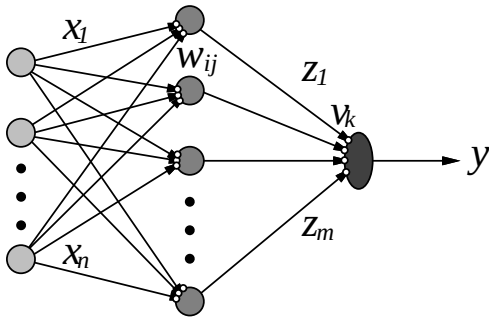

図 2: パーセプトロン

隠れ変数は離散な数値を取るとは限らない．図 2 のような 1 出力の 3 層パーセプトロンを考える．入力を $\boldsymbol{x}$ (n 次元実数ベクトル)，出力を y (実数) とする．中間素子の出力を $\boldsymbol{z} = (z_1,\ldots,z_m)^T$ として，中間素子の出力は sigmoid 関数だとし，出力は中間素子の出力の重み付きの和とする．これは良く用いられる神経回路網の一つである．神経回路網全体の関数を $y = g(\boldsymbol{x})$ であらわすと，

$$g(\boldsymbol{x}) = \boldsymbol{v}\cdot\boldsymbol{z} = \sum_j v_j z_j \tag{1}$$

$$z_i = f\left(\textstyle\sum_j w_{ij}x_j\right), \qquad f(u) = \dfrac{1}{1+e^{-u}} \tag{2}$$

となる．ここで z_i と y の定義を変え，

$$z_i = f\left(\textstyle\sum_j w_{ij}x_j\right) + n_i \tag{3}$$

$$y = \boldsymbol{v}\cdot\boldsymbol{z} + n, \qquad n_1,\ldots,n_m,n \sim \mathcal{N}(0,\sigma^2)$$

とすると，中間層の出力 $z_1,\ldots,z_m$ は外からは直接観測できず，隠れた確率変数となる．確率分布の形はパラメータを $\boldsymbol{\theta} = (W,\boldsymbol{v},\sigma^2)$ と置くと，

$$p(y,\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}^{m+1}}\exp\left\{-\frac{1}{2\sigma^2}(y-\boldsymbol{v}\cdot\boldsymbol{z})^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{m}\left(z_i - f\left(\textstyle\sum_j w_{ij}x_j\right)\right)^2\right\} \tag{4}$$

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \int p(y,\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{z} = \frac{1}{\sqrt{2\pi(1+|\boldsymbol{v}|^2)\sigma^2}}\exp\left\{-\frac{1}{2(1+|\boldsymbol{v}|^2)\sigma^2}(y-g(\boldsymbol{x}))^2\right\} \tag{5}$$

である．この場合隠れ変数 $\boldsymbol{z}$ は連続な変数となる．

2.3 ボルツマンマシン

ボルツマンマシン (Boltzmann machine, 図 3) は互いの細胞が全て結合された n 個の細胞からなる．各細胞は確率的に 0 または 1 の二値をとる．各細胞は同じ振舞いをするが，外部から入力が与えられる入力細胞，外部に出力を出す出力細胞，そのどちらでもない隠れ細胞の 3 種類を定義する．具体的な動作は以下のようになる．

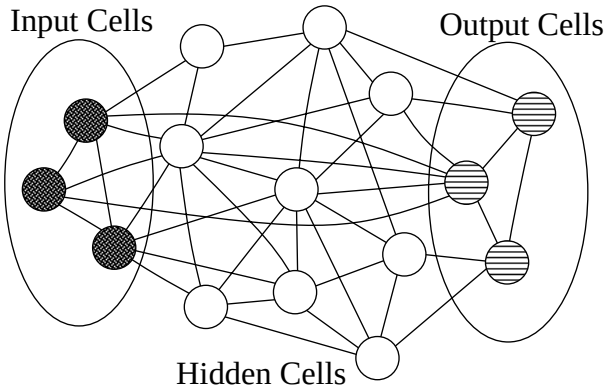

図 3: ボルツマンマシン

第 i 番目の素子の内部状態を u_i とし，出力を x_i とする．第 i 素子と第 j 素子との相互結合の強さを w_{ij} で表わす．ただし自己結合はないとする．内部状態は

$$u_i = \sum_{j \neq i} w_{ij} x_j - h_i \tag{6}$$

と計算される．h_i は内部状態の正負の範囲を規定するしきい値である．以下記法を簡単にするため $w_{ii} = 0$, $w_{i0} = h_i$ として恒等的に 1 を出力する素子 $x_0 = 1$ を付け加えて

$$u_i = \sum_{j=0}^{n} w_{ij} x_j \tag{7}$$

により内部状態を表記する．

各素子は非同期，つまり一回の更新に際しただ一つの素子の出力が次の確率にしたがって更新される．

$$\begin{aligned} p(x_i = 1|u_i) &= \frac{1}{1 + \exp(-\frac{u_i}{T})} \\ p(x_i = 0|u_i) &= \frac{\exp(-\frac{u_i}{T})}{1 + \exp(-\frac{u_i}{T})} \end{aligned} \tag{8}$$

T は温度パラメタと呼ばれ，素子の確率的動作の度合を制御する役割をする．$T = 0$ の極限では素子は確定的に動作，すなわち内部状態の正負によって出力は 1 または 0 に確定される．

素子数 n のボルツマンマシンは状態数 2^n の有限状態のマルコフ連鎖とみなせる．

$$E(\boldsymbol{x}) = -\sum_{i,j} w_{ij} x_i x_j \tag{9}$$

とすると，定常分布において出力 $\boldsymbol{x}$ を取る確率はボルツマン分布に従う．これは $\boldsymbol{\theta} = \{w_{ij}\}$ とおき，

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{Z} \exp\left(-\frac{E(\boldsymbol{x})}{T}\right) \tag{10}$$

で表される．$\boldsymbol{x}$ には隠れている細胞と外から観測できる細胞がある．$\boldsymbol{x} = (\boldsymbol{x}_v, \boldsymbol{x}_h)$ とし，$\boldsymbol{x}_v$ が観測できる細胞，$\boldsymbol{x}_h$ が観測できないとすると，

$$p(\boldsymbol{x}_v; \boldsymbol{\theta}) = \sum_{\boldsymbol{x}_h} p(\boldsymbol{x}_v, \boldsymbol{x}_h; \boldsymbol{\theta})$$

であることから，これは隠れ変数を持つ確率モデルである．Boltzmann マシンの幾何的な解釈については [1] に示されている．

2.4 Helmholtz マシン

ヘルムホルツマシン (Helmholtz machine) は生成モデル (generative model) と認識モデル (recognition model) の 2 つのモジュールからなるモデルである (詳しくは [2] を参照)．図 4 に示すのが Helmholtz マシンの模式図である．Helmholtz マシンは外部からの入力となる Visible variable と外部からは直接観測できない Hidden factor を持つ．これを脳に対応させ，visible variable を低次の神経細胞，hidden factor を高次の細胞と呼ぶこともある．

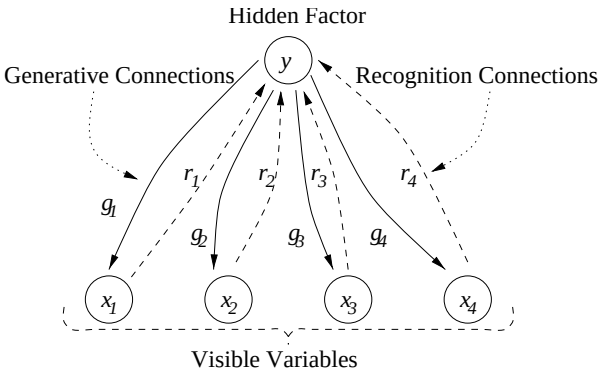

図 4: Helmholtz マシン

この 2 つの変数の組に対し，Helmholtz マシンでは 2 組のパラメータを与える．1 つは visible variable を得たときに hidden factor の確率分布を与える認識モデル (recognition model) であり，もう 1 つは visible variable と hidden factor の両方の分布を与える 生成モデル (generative model) である．この 2 つのモデルを 一組の変数に重ねて定義することが Helmholtz マシンの特徴である．脳に対応させると，低次から高次への結合と高次から低次への結合が共存していることになる．

各々のモデルの形については Helmholtz マシン自体は規定していない．つまり，どのようなモデルを定義するかによって，Helmholtz マシンは様々な分布となる．例えば最も簡単な線型なモデルの場合，これは古典的な因子分析のモデルと一致する．また，混合正規分布や HMM(隠れマルコフモデル)，など隠れ変数を持つモデルは全てこの形で表現することができる．ここでは線型の場合のモデルについて定義をしておく [3].

$\boldsymbol{x}$ を n 次元の visible variable とし，y を 1 次元の hidden factor とする．

- **生成モデル**：$\boldsymbol{x}$ を n 次元の信号が標準正規分布 $N(0,1)$ にしたがう確率変数 y によって

$$\boldsymbol{x} = \boldsymbol{g}y + \boldsymbol{\varepsilon} \tag{11}$$

  により生成されるとする．$\boldsymbol{\varepsilon}$ は対角行列 $\Sigma = \mathrm{diag}(\sigma_i^2)$ を分散行列とする正規分布 $N(0,\Sigma)$ にしたがう雑音である．

- **認識モデル**：観測された信号 $\boldsymbol{x}$ から対応する y が

$$y = \boldsymbol{r}^T\boldsymbol{x} + \delta \tag{12}$$

  のように分布するとする．ただし δ は $N(0,s^2)$ にしたがう雑音である．

3　隠れ変数を持つモデルのパラメータ推定

統計モデルでは，データを基にそれをうまく表現するようにモデルのパラメータを推定する．脳においても学習を通じてその結合や結合強度を学習していると考えられる．情報の立場から隠れ変数を持つモデルの学習法を学ぶことは脳の研究にも役立つと考えられる．

3.1　最尤推定

まず，統計的な推定法の 1 つである最尤推定について説明する．

$\boldsymbol{\theta}$ をパラメータとする確率分布 $p(\boldsymbol{y};\boldsymbol{\theta})$ を考える．今データが i.i.d. で $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T\}$ として得られたとき，$\boldsymbol{\theta}$ を推定したい．最尤推定では，$p(\boldsymbol{y};\boldsymbol{\theta})$ がそのデータを受けとる確率 (尤度) を最大にするパラメータを推定量 $\boldsymbol{\theta}^*$ とする．

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_s p(\boldsymbol{y}_s;\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_s \log p(\boldsymbol{y}_s;\boldsymbol{\theta}) \tag{13}$$

これは，見方を変えると，経験分布 $q(\boldsymbol{y})$ とモデルの間の Kullback-Leibler Divergence を最小にしているとも考えられる．Kullback-Leibler Divergence は次のように定義される．

$$D(q, p(\boldsymbol{\theta})) = \int q(\boldsymbol{y}) \log \frac{q(\boldsymbol{y})}{p(\boldsymbol{y};\boldsymbol{\theta})} d\boldsymbol{y} = \int q(\boldsymbol{y}) \log q(\boldsymbol{y}) d\boldsymbol{y} - \int q(\boldsymbol{y}) \log p(\boldsymbol{y};\boldsymbol{\theta}) d\boldsymbol{y} \tag{14}$$

右側の式の第 2 項は(13) 式の対数尤度と等しい．パラメータに関する部分はこの項だけなので，対数尤度を最大にすることは(14) 式の量を最小にしていることと同値となる．

この結果を情報幾何的に解釈する．図 5 はこのイメージを示したものである．図中の S は $\boldsymbol{y}$ の確率分布の空間を考えたものである．この空間中の各点は $\boldsymbol{y}$ の確率分布となる．モデルは $\boldsymbol{\theta}$ というパラメータを持つ集合であるので，この空間中では多様体 M として表わさていれる．経験分布を得たとき，ここからパラメータを最尤推定するとは，経験分布の 1 点 $q(\boldsymbol{y})$ からモデル多様体 M への一種の射影だとみなせる．この場合の射影は $D(q(\boldsymbol{y}), p(\boldsymbol{y};\boldsymbol{\theta}))$ を最小とする点を求めることと等しい [4].

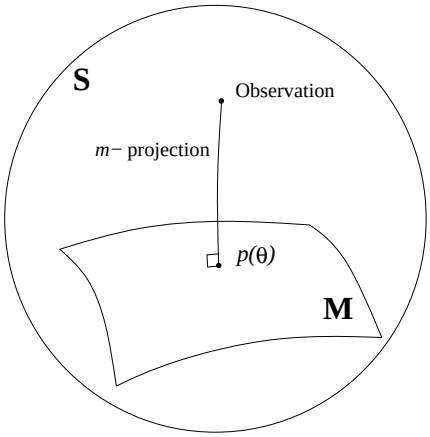

図 5: 統計的推定の幾何学的イメージ

3.2　EM アルゴリズム

では，隠れ変数のあるモデルにおける最尤推定はどうであろう．ある確率変数 $\boldsymbol{X} = \{\boldsymbol{Y}, \boldsymbol{Z}\}$ があり，その一部 $\boldsymbol{Y}$ のみが観測でき，残り $\boldsymbol{Z}$ は観測できない状況を考える．観測データ $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T\}$ が得られた

ときに，確率モデル $p(\boldsymbol{x};\boldsymbol{\theta}) = p(\boldsymbol{y},\boldsymbol{z};\boldsymbol{\theta})$ のパラメタ $\boldsymbol{\theta}$ を推定したいとする．

$$p(\boldsymbol{y};\boldsymbol{\theta}) = \int p(\boldsymbol{y},\boldsymbol{z};\boldsymbol{\theta})d\boldsymbol{z}$$

と定義されるが，この形は通常単純ではなく，(13) 式を直接解くのは難しい．このような場合に用いられる手法の 1 つに EM アルゴリズムがある．

EM アルゴリズムは E-step (Expectation step) と M-step (Maximization step) の二つの部分からなり，これらを交互に繰り返してパラメタを更新することにより，最尤推定量あるいは尤度関数の極大点を得ることができる．

適当な初期値 $\boldsymbol{\theta}_0$ から始めて t 回更新した後のパラメタを $\boldsymbol{\theta}_t$ として，E-step と M-step の具体的な手続きは以下のように定義される．

- **E-step**

  次式で定義される $Q(\boldsymbol{\theta},\boldsymbol{\theta}_t)$ を求める．

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}_t) = \frac{1}{T}\sum_{k=1}^{T}\left\{\int p(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta}_t)\log p(\boldsymbol{y},\boldsymbol{z};\boldsymbol{\theta})d\boldsymbol{z}\right\} \tag{15}$$

- **M-step**

  $Q(\boldsymbol{\theta},\boldsymbol{\theta}_t)$ を最大にする $\boldsymbol{\theta}$ を求め，それを $\boldsymbol{\theta}_{t+1}$ にする．

$$\boldsymbol{\theta}_{t+1} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q(\boldsymbol{\theta},\boldsymbol{\theta}_t) \tag{16}$$

この結果得られた $\boldsymbol{\theta}_t$ と $\boldsymbol{\theta}_{t+1}$ との間には $\sum_t \log p(\boldsymbol{y}_s;\boldsymbol{\theta}_t) \leq \sum_s \log p(\boldsymbol{y}_s;\boldsymbol{\theta}_{t+1})$ という関係がある．

EM アルゴリズムを情報幾何的に解釈する．単純な問題では図 5 のように最尤推定は点から多様体への射影として捉えられる．一方，隠れ変数を持つモデルでは観測できる確率変数 $\boldsymbol{y}$ の確率分布の空間ではなく，確率変数 $\boldsymbol{x} = \{\boldsymbol{y},\boldsymbol{z}\}$ の確率分布の空間を考えた方が分り易い場合が多い．この空間を考えよう．

今，モデルのほうは前節と同様に 1 つの多様体 M を構成する．一方データのほうは $\boldsymbol{y}$ に関する経験分布 $q(\boldsymbol{y})$ しか与えない．このままでは $\boldsymbol{x} = \{\boldsymbol{y},\boldsymbol{z}\}$ の確率分布の空間中の点とはならないので，$\boldsymbol{z}$ に関する任意の分布を付け加え D (図 6) という多様体を構成する (より詳しくは [4][5] を参照されたい)．EM アルゴリズムはこの 2 つの多様体の間のそれぞれの点で D と M とを最も近くする点を求めることに対応している．これを元に EM アルゴリズムを書き換えると甘利によって提案された em アルゴリズムとなる [5]．

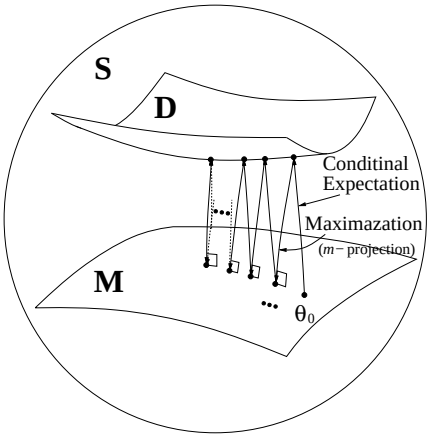

図 6: EM アルゴリズム

- e**-step** 多様体 D 上で $D(q(\boldsymbol{x};\boldsymbol{\eta}), p(\boldsymbol{x};\boldsymbol{\theta}_t))$ を最小にする $\boldsymbol{\eta}_{t+1}$ を求める．
- m**-step** 多様体 M 上で $D(q(\boldsymbol{x};\boldsymbol{\eta}_{t+1}), p(\boldsymbol{x};\boldsymbol{\theta}))$ を最小にする $\boldsymbol{\theta}_{t+1}$ を求める．

3.3 Wake-Sleep アルゴリズム

EM アルゴリズムの用いられるような神経回路網モデルの学習則として，G. Hinton らは EM アルゴリズムとは別に Wake-Sleep アルゴリズムと呼ばれる学習則を提案した [6]．これは元来 Helmholtz マシンに対して提案されたが，特に Helmholtz マシンに限らず，用いることができる．

ここでまず Helmholtz マシンについて見直してみる．Helmholtz マシンは認識モデルと生成モデルを定義することで成り立っている．これは図 6 の M 多様体を生成モデルとして，そして D 多様体を認識モデルとして提案したことに他ならない．今，認識モデルを $q(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\eta})$ とし，生成モデルを $p(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})$ とする．この 2 つのモデルの間で Wake-Sleep アルゴリズムは次のように定義される．

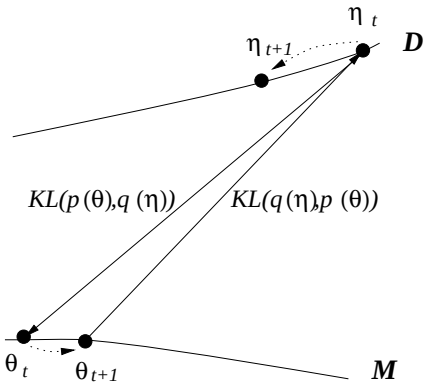

図 7: Wake Sleep アルゴリズム

- **Wake-phase**

  Visible variable に関する得られた経験分布 $q(\boldsymbol{y})$ と認識モデル $q(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\eta}_t)$ を基に $\{\boldsymbol{x},\boldsymbol{y}\}$ のサンプルを生成し，そのサンプルに対し，生成モデルを近づけるように $\boldsymbol{\theta}$ を更新し，$\boldsymbol{\theta}_{t+1}$ とする．これは

  $$D(q(\boldsymbol{y})q(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\eta}_t), p(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})) \tag{17}$$

  を小さくするように $\boldsymbol{\theta}$ を更新することに対応する．

- **Sleep-phase**

  生成モデル $p(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta}_t)$ を基に $\{\boldsymbol{x},\boldsymbol{y}\}$ のサンプルを生成し，そのサンプルに対し，認識モデルを近づけるように $\boldsymbol{\eta}$ を更新し，学習する．$\boldsymbol{\eta}_{t+1}$ とする．これは

  $$D(p(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta}_{t+1}), q(\boldsymbol{y})q(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\eta})) \tag{18}$$

  を小さくするように $\boldsymbol{\eta}$ を更新することに対応する．

Wake-Sleep アルゴリズムは幾何学的には(17),(18) 式の 2 つの Kullback-Leibler Divergence を交互に小さくしていると解釈できる．しかし，この 2 つの Kullback-Leibler Divergence は対称ではなく，(17) を最小とする点が (18) 式を最小にするとは限らない．

Wake-Sleep アルゴリズムは名前に示されている通り，脳の覚醒状態と睡眠中のパラメータの更新になぞらえて提案された．覚醒中は高次の細胞が学習し，睡眠中は低次の細胞が学習するのである．これは大変興味深い提案であるが，脳と直接結びつけるためにはいくつか乗り越えなければならない．

まず，学習の収束の問題である．一般に $q(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\eta}) = p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})$ となる θ が常に存在すれば，Wake-Sleep アルゴリズムは Sleep phase を十分長く取ることで最尤推定点に収束する．しかし，その他の場合には必ずしも収束しない．また，脳の学習モデルとして意味のあるものとなるためには，ある程度 local で単純な計算で学習が進む方が分り易い．しかし local な学習則となるかどうかは Helmholtz マシンとして定義し

た各モデルの形によってしまう．ただし，これは Wake-Sleep アルゴリズムの考え方を否定したものではない．Wake-Sleep アルゴリズムの提案する考え方は興味深く，実際の脳との対応の中でモデルを考えていくヒントとなると思う．

4 まとめ

本稿では隠れ変数を持つモデルとその学習則について説明を行なった．隠れ変数はその定義の方法によって様々なモデルを定義することができる．近年話題になっている ICA のモデルや HMM(隠れマルコフモデル)，因子分析モデルなども含まれる．様々なモデルが含まれるということは，必ずしも隠れ変数を含むこのような表現でなくてもモデルを表現できるということだが，隠れ変数を持つモデルは確率モデルの表現に見通しの良さを与えるものだと考える．

脳の情報処理も隠れ変数を基にした理解により，少なくとも神経回路網モデルの分野では格段に進歩した．実際の脳のモデルにおいても，データ処理では隠れ変数を導入することで新たな解析結果が得られてきているとの報告もある．また Sparse Coding に関しても隠れ変数を用いた説明によってい大変分りやすい結果が得られている．このように隠れ変数を考えることが脳の理解の上でも役立つと考える．

参考文献

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

[2] Peter Dayan, Geoffrey E. Hinton, and Radford M. Neal. The Helmholtz machine. *Neural Computation*, 7(5):889–904, 1995.

[3] Shiro Ikeda, Shun-ichi Amari, and Hiroyuki Nakahara. Convergence of the Wake-Sleep algorithm. to appear In Advances in Neural Information Processing Systems 11, http://www.islab.brain.riken.go.jp/ shiro/publications-e.html.

[4] 甘利 俊一，長岡 浩司. **情報幾何の方法**. 岩波講座 応用数学. 岩波書店, 1993.

[5] Shun-ichi Amari. Information geometry of the EM and em algorithms for neural networks. *Neural Networks*, 8(9):1379–1408, 1995.

[6] Geoffrey E. Hinton, Peter Dayan, B. J. Frey, and Radford M. Neal. The “sake-sleep” algorithm for unsupervised neural networks. *Science*, 268:1158–1160, 1995.