神経回路網とEMアルゴリズム

村田 昇，池田 思朗
理化学研究所 脳科学総合研究センター
埼玉県和光市広沢 2-1
{Noboru.Murata@eb.waseda.ac.jp， shiro@ism.ac.jp

2004年3月改訂

1　はじめに

神経回路網の研究の目的は，一つには複雑に変化する環境にうまく適応し生き延びていくことができる生物の脳や神経系の機能を数学的にモデル化し解明していくことであるが，同時にその工学的な応用も様々な観点から検討されている ([1] などを参照).

一般に神経回路網と呼ばれる計算モデルの特徴は，大多数の単純な素子からなる均質な集合体，いわゆる超並列であること，およびそれらの相互関係を局所的な情報処理で変化させていく，すなわち学習を行うことの二点に纏めることができる．現在扱われているモデルは脳の構造を単純化したものであり，生物学的に真に適切かという点で疑問は残るが，数学的に取り扱うための妥当なモデル化だと考えられている．最も多くの場面で利用される多層パーセプトロンは素子数を増やすことによって任意の関数を近似できる万能近似装置であることも知られており (例えば文献 [2])，このことは工学的応用を考える上で非常に大事な性質である．また学習は例題に基づきある規準を最適化していることになり，推定方程式との類似性をみることができる．つまり神経回路網を統計モデル，学習を一種の統計的推定と考えることもできる．これらを始めとして神経回路網は様々な分野における広範な問題を含んだ興味深い問題であると言える.

神経回路網における情報の流れは一般に入力素子から隠れ素子と呼ばれる入出力と直接関連しない素子を経て出力素子に到達する．学習においてはこの外からは見ることのできない隠れ素子の値を入出力の関係から適宜定め，それにしたがってパラメタの更新を行わなくてはならない．この構造はEMアルゴリズムと非常に関係が深く，実際陰に陽に学習の中にEMアルゴリズムは現れてくる.

本稿ではまずEMアルゴリズムを幾何学的に見直したemアルゴリズムの説明を行う．この考え方は損失関数から導かれる学習則を解釈する上で有効

である．次に EM アルゴリズムが陰に現れる特別な神経回路モデルの例を述べる．生物学的な観点から複雑な学習則をを考える場合大きな問題となるのは EM アルゴリズムを生物の構造上妥当な形で実現できるかという点であるが，この問題についてもヘルムホルツマシンというモデルを通じて簡単に触れる．最後に直接 EM アルゴリズムが学習則として用いられる神経回路モデルを例示する．これらは生物のモデルというより工学的応用に重点を置いて提案されたものであり，従来の単純なモデルでは不十分なより複雑なタスク，例えばロボットの制御等に積極的に用いられ始めている．

2 EM アルゴリズムと em アルゴリズム

2.1 情報幾何の考え方

EM アルゴリズムの幾何的解釈を説明する前にこの節では情報幾何の枠組を簡単に説明する．情報幾何学は確率密度関数のつくる空間を微分幾何学を用いて扱うことにより，統計的推定や検定問題を幾何学的に解釈しようとする方法論である．数学的に厳密な話は [3, 4, 5, 6] などの成書に譲るとして，以下では直観的な説明を試みる．

次のような通常のパラメトリックな推定問題を考える．

確率変数 $\boldsymbol{X}$ を考え，その確率密度関数 $p(\boldsymbol{x})$ のなす空間 S を考える．直観的には経験分布を含むあらゆる確率分布を想定しておけばよいが，厳密には定義域での正値性や可微分性などの適当な正則条件を満たすものを考えなくてはいけない．S に含まれる分布の中で特にパラメタ $\boldsymbol{\theta}$ で表現される確率密度関数 $p(\boldsymbol{x};\boldsymbol{\theta})$ を確率モデルとした推定問題を考える．集合 $\{p(\boldsymbol{x};\boldsymbol{\theta})\}$ は空間 S の中で部分多様体を成すが，これをモデル多様体 M と呼ぶことにする．

観測データ $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_T$ が与えられると，データから様々な統計量を計算することができ，それに基づいて観測データを発生する尤もらしい分布を確率分布の空間 S の中に一点定めることができる．直観的には経験分布に相当する点が一点決まると思えばよい．このようにして観測データは空間 S 中では一点として扱われるが，この点は必ずしもモデル多様体 M には属していない．したがって想定した確率モデルの中から一つの候補を選んでやるためには，S の一点から何らかの意味で最も近いモデル多様体 M 上の一点を定めてやらなくてはならない．この操作は幾何学的には空間 S 内の一点からモデル多様体 M 上の一点への射影を求める問題として捉えることができる (図 1)．

考えている空間が線形空間であれば，ある点から一番近い線形部分空間の一点は直交射影によって求めることができるが，空間 S は一般に線形空間ではなく “曲った” 空間である．“曲った” 空間の中で射影を定めるためには直線の概念を拡張する必要があり，また直交射影を定めるためにはさらに接ベクトルの内積を定義して直交性を定義しなくてはならない．

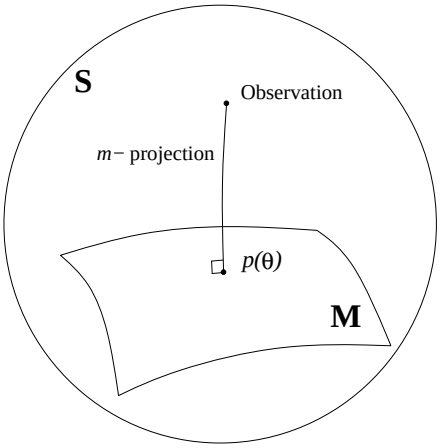

図 1: 統計的推定の幾何学的イメージ

まず直線の概念を拡張して S の中で “まっすぐ” な線を決めてやる．“曲った” 空間において直線に対応するものは測地線と呼ばれ，様々な定義が可能であるが，Kullback-Leibler 情報量を統計的距離として用いる場合には m-測地線と e-測地線という二つの概念が重要な役割を果す．m-測地線は 2 つの確率密度関数 $p(\boldsymbol{x})$ と $q(\boldsymbol{x})$ の内分点の集合

$$r(\boldsymbol{x};t) = (1-t)\cdot p(\boldsymbol{x}) + t\cdot q(\boldsymbol{x}), \quad 0 \le t \le 1 \tag{1}$$

として定義される．一方 e-測地線は 2 つの確率密度関数 $p(\boldsymbol{x})$ と $q(\boldsymbol{x})$ の対数の意味での内分点の集合

$$\log r(\boldsymbol{x};t) = (1-t)\cdot \log p(\boldsymbol{x}) + t\cdot \log q(\boldsymbol{x}) - \phi(t), \quad 0 \le t \le 1 \tag{2}$$

として定義される．ただし $\phi(t)$ は $r(\boldsymbol{x};t)$ を確率密度とするための正規化因子で

$$\phi(t) = \log \int p(\boldsymbol{x})^{1-t} q(\boldsymbol{x})^t d\boldsymbol{x} \tag{3}$$

である．

測地線を用いて “まっすぐ” な線を定義したと同様に，平面の概念も拡張することができる．例えば n 個の確率密度関数 $p_i(\boldsymbol{x})$ からなる混合型分布族の多様体

$$M_m = \left\{ p(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{i=1}^{n} \theta_i p_i(\boldsymbol{x}), \quad \theta_i > 0, \quad \sum_{i=1}^{n} \theta_i = 1 \right\} \tag{4}$$

を考える．この中から任意の 2 つの分布を選び m-測地線を考えると，m-測地線は元の多様体 M_m に含まれていることが容易にわかる．これは多様体 M_m

が “まっすぐ” な線だけで構成されていることを意味し，S の中で M_m は m-測地線の意味において “平ら” な部分集合になっている．同様に

$$M_e = \left\{ p(\boldsymbol{x};\boldsymbol{\theta}) = \exp\left(\sum_{i=1}^{n} \theta_i r_i(\boldsymbol{x}) - \psi(\boldsymbol{\theta}) \right) \right\} \tag{5}$$

のような指数型分布族を考えると，今度は M_e の 2 点を結ぶ任意の e-測地線が M_e に含まれることがわかり，これも “平ら” な部分集合になっている．上記の “平ら” の概念は直観的で数学的には正しくなく，厳密には微分幾何における接続の概念を用いて m-平坦性，e-平坦性として定義されるべきものであるが，そのような取り扱いについては文献 [3, 6] を参照されたい．

さて以上のようにして空間 S 内に “まっすぐ” な線の概念を導入することができたので，次に直交射影を決める．このためにはある方向に沿う接ベクトルを定義し，内積を導入する．天下りではあるが，接ベクトルは確率密度の対数 $\log p(\boldsymbol{x})$ の微少変化によって表し，内積はその相関

$$E_p\left(\partial_\alpha \log p(\boldsymbol{X}) \cdot \partial_\beta \log p(\boldsymbol{X})\right) \tag{6}$$

として定義する．ただし ∂_α は α で示されるある方向に沿った微分を表すものとする．例えば測地線に沿った方向は測地線のパラメタ t での微分で，m-測地線の場合は

$$\begin{aligned} \partial_t \log r(\boldsymbol{x};t) &= \frac{\partial_t r(\boldsymbol{x};t)}{r(\boldsymbol{x};t)} \\ &= \frac{\frac{d}{dt}\left\{(1-t)\cdot p(\boldsymbol{x}) + t\cdot q(\boldsymbol{x})\right\}}{r(\boldsymbol{x};t)} \\ &= \frac{q(\boldsymbol{x}) - p(\boldsymbol{x})}{r(\boldsymbol{x};t)} \end{aligned} \tag{7}$$

e-測地線の場合は

$$\begin{aligned} \partial_t \log r(\boldsymbol{x};t) &= \frac{d}{dt}\left\{(1-t)\cdot \log p(\boldsymbol{x}) + t\cdot \log q(\boldsymbol{x}) - \phi(t)\right\} \\ &= \log q(\boldsymbol{x}) - \log p(\boldsymbol{x}) - \frac{d}{dt}\phi(t) \end{aligned} \tag{8}$$

となる．またモデル多様体に沿った方向はモデルのパラメタ $\boldsymbol{\theta}$ での微分を考えればよく，$\boldsymbol{\theta}$ の第 i 成分を θ_i とすれば

$$\begin{aligned} \partial_{\theta_i} \log p(\boldsymbol{x};\boldsymbol{\theta}) &= \frac{\partial}{\partial\theta_i} \log p(\boldsymbol{x};\boldsymbol{\theta}) \\ &= \frac{\partial_{\theta_i} p(\boldsymbol{x};\boldsymbol{\theta})}{p(\boldsymbol{x};\boldsymbol{\theta})} \end{aligned} \tag{9}$$

となる．以降混乱のない限り簡単のため ∂ によって微分演算子を表すことにする．

こうした準備のもとで m-射影と e-射影の二種類の射影が決められる．m-射影とは，空間 S の一点 q からモデル多様体 M に降ろした m-測地線がモデル

多様体 M と直交するように m-測地線の足 $p(\hat{\boldsymbol{\theta}})$ を決めたものである．これは m-測地線の意味で観測点からモデルへの一番近い点になり，実は Kullback-Leibler 情報量

$$\begin{aligned} D(q, p(\boldsymbol{\theta})) &= \int q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x}; \boldsymbol{\theta})} d\boldsymbol{x} \\ &= E_q \left[\log q(\boldsymbol{x}) - \log p(\boldsymbol{x}; \boldsymbol{\theta}) \right] \end{aligned} \tag{10}$$

を最小にするパラメタ $\hat{\boldsymbol{\theta}}$ を求めることと同値である．実際 Kullback-Leibler 情報量を最小にする点 $\hat{\boldsymbol{\theta}}$ においてはパラメタのどの方向に関する微分も 0，すなわち

$$\begin{aligned} \partial_{\theta_i} D(q, p(\boldsymbol{\theta}))\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} &= -E_q \left[\partial_{\theta_i} \log p(\boldsymbol{X}; \hat{\boldsymbol{\theta}}) \right] \\ &= 0 \end{aligned} \tag{11}$$

が成り立っており，$p(\hat{\boldsymbol{\theta}})$ における m-測地線に沿う接ベクトル

$$\begin{aligned} \partial_t \log r(\boldsymbol{x}; t)\Big|_{t=0} &= \frac{q(\boldsymbol{x}) - p(\boldsymbol{x}; \hat{\boldsymbol{\theta}})}{r(\boldsymbol{x}; t)}\Big|_{t=0} \\ &= \frac{q(\boldsymbol{x}) - p(\boldsymbol{x}; \hat{\boldsymbol{\theta}})}{p(\boldsymbol{x}; \hat{\boldsymbol{\theta}})} \end{aligned} \tag{12}$$

とモデル多様体に沿う接ベクトル

$$\partial_{\theta_i} \log p(\boldsymbol{x}; \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{13}$$

の内積を考えると

$$\begin{aligned} & E_{p(\hat{\boldsymbol{\theta}})} \left[\partial_t \log r(\boldsymbol{X}; 0) \cdot \partial_{\theta_i} \log p(\boldsymbol{X}; \hat{\boldsymbol{\theta}}) \right] \\ &= \int \left(\frac{q(\boldsymbol{x}) - p(\boldsymbol{x}; \hat{\boldsymbol{\theta}})}{p(\boldsymbol{x}; \hat{\boldsymbol{\theta}})} \right) \partial_{\theta_i} \log p(\boldsymbol{x}; \hat{\boldsymbol{\theta}}) p(\boldsymbol{x}; \hat{\boldsymbol{\theta}}) d\boldsymbol{x} \\ &= E_q \left[\partial_{\theta_i} \log p(\boldsymbol{X}; \hat{\boldsymbol{\theta}}) \right] - E_{p(\hat{\boldsymbol{\theta}})} \left[\partial_{\theta_i} \log p(\boldsymbol{X}; \hat{\boldsymbol{\theta}}) \right] \\ &= 0 \end{aligned} \tag{14}$$

となり，確かに q と $p(\boldsymbol{\theta})$ を結ぶ m-測地線がモデル多様体と直角に交わっていることがわかる．ただしモデルは課せられた正則条件のもとで微積の順序が可換であるとし，自明な恒等式

$$\begin{aligned} E_{p(\boldsymbol{\theta})} \left[\partial_{\theta_i} \log p(\boldsymbol{X}; \boldsymbol{\theta}) \right] &= \partial_{\theta_i} \int p(\boldsymbol{x}; \boldsymbol{\theta}) d\boldsymbol{x} \\ &= 0 \end{aligned} \tag{15}$$

を用いた．このことから q として経験分布を用いる場合には m-射影は最尤推定に対応することがわかる (図 1)．一方 e-測地線の意味で一番近い点への射

影を求めるのが e-射影であり，これは

$$D(p(\boldsymbol{\theta}), q) = \int p(\boldsymbol{x}; \boldsymbol{\theta}) \log \frac{p(\boldsymbol{x}; \boldsymbol{\theta})}{q(\boldsymbol{x})} d\boldsymbol{x} \tag{16}$$

を最小にするパラメタ $\hat{\boldsymbol{\theta}}$ を求めていることに相当する．m-射影と同様にこの場合も $p(\boldsymbol{x}; \hat{\boldsymbol{\theta}})$ において e-測地線とモデル多様体が直交している．e-測地線に沿う接ベクトルが

$$\begin{aligned} \partial_t \log r(\boldsymbol{x}; t)\Big|_{t=0} &= \log q(\boldsymbol{x}) - \log p(\boldsymbol{x}; \hat{\boldsymbol{\theta}}) - \frac{d}{dt}\phi(t)\Big|_{t=0} \\ &= \log q(\boldsymbol{x}) - \log p(\boldsymbol{x}; \hat{\boldsymbol{\theta}}) - E_{p(\hat{\boldsymbol{\theta}})}\left[\log q(\boldsymbol{X}) - \log p(\boldsymbol{X}; \hat{\boldsymbol{\theta}})\right] \end{aligned} \tag{17}$$

となることに注意すると，e-測地線の接ベクトルとモデル多様体の接ベクトルの内積は

$$\begin{aligned} & E_{p(\hat{\boldsymbol{\theta}})}\left[\partial_t \log r(\boldsymbol{X}; 0) \cdot \partial_{\theta_i} \log p(\boldsymbol{X}; \hat{\boldsymbol{\theta}})\right] \\ &= \int \partial_{\theta_i} p(\boldsymbol{x}; \hat{\boldsymbol{\theta}}) \left\{\log q(\boldsymbol{x}) - \log p(\boldsymbol{x}; \hat{\boldsymbol{\theta}}) - E_{p(\hat{\boldsymbol{\theta}})}\left[\log q(\boldsymbol{x}) - \log p(\boldsymbol{x}; \hat{\boldsymbol{\theta}})\right]\right\} d\boldsymbol{x} \\ &= \int \partial_{\theta_i} p(\boldsymbol{x}; \hat{\boldsymbol{\theta}}) \left\{\log q(\boldsymbol{x}) - \log p(\boldsymbol{x}; \hat{\boldsymbol{\theta}})\right\} d\boldsymbol{x} \end{aligned} \tag{18}$$

となるが，

$$\begin{aligned} \partial_{\theta_i} D(p(\boldsymbol{\theta}), q)\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} &= \int \partial_{\theta_i} p(\boldsymbol{x}; \hat{\boldsymbol{\theta}}) \log p(\boldsymbol{x}; \hat{\boldsymbol{\theta}}) d\boldsymbol{x} + \int \partial_{\theta_i} p(\boldsymbol{x}; \hat{\boldsymbol{\theta}}) d\boldsymbol{x} \\ &\quad - \int \partial_{\theta_i} p(\boldsymbol{x}; \hat{\boldsymbol{\theta}}) \log q(\boldsymbol{x}) d\boldsymbol{x} \\ &= \int \partial_{\theta_i} p(\boldsymbol{x}; \hat{\boldsymbol{\theta}}) \left(\log p(\boldsymbol{x}; \hat{\boldsymbol{\theta}}) - \log q(\boldsymbol{x})\right) d\boldsymbol{x} \\ &= 0 \end{aligned} \tag{19}$$

であり，直交していることが容易に確かめられる．

射影が一点に決まるかどうかはモデル多様体の曲り具合によるが，例えば典型的な十分条件としては，モデル多様体が e-平坦のとき m-射影は一点に決まることがわかっている．これはちょうどユークリッド空間においてある一点から平面の直交射影が一意に決まることと対応すると考えればよい．

2.2 EM アルゴリズムの幾何学的解釈

ある確率変数 $\boldsymbol{X}$ があり，その一部のみ観測でき，残りは観測することができない状況を考える．観測できる確率変数を $\boldsymbol{X}_V$，観測できない確率変数 (潜在変数，隠れ変数) を $\boldsymbol{X}_H$ とし，$\boldsymbol{X} = (\boldsymbol{X}_V, \boldsymbol{X}_H)$ と書く．観測データ $\{\boldsymbol{x}_{V,1}, \boldsymbol{x}_{V,2}, \ldots, \boldsymbol{x}_{V,T}\}$ が得られたときに，確率モデル $p(\boldsymbol{x}; \boldsymbol{\theta}) = p(\boldsymbol{x}_V, \boldsymbol{x}_H; \boldsymbol{\theta})$

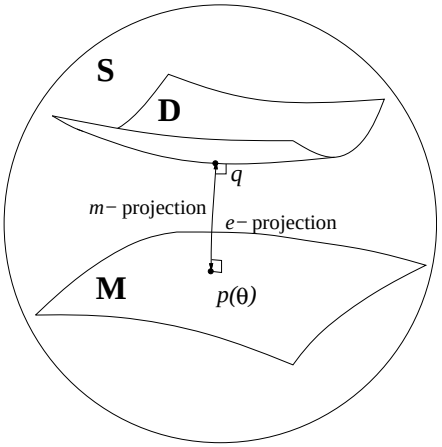

図 2: 観測多様体とモデル多様体

のパラメタ $\boldsymbol{\theta}$ を決定したい．このような隠れた確率変数がある場合のパラメタ推定の問題を前節と同様に幾何学的な枠組で考えてみる．

潜在変数がある問題では，観測されるデータだけからでは空間 S の中の一点を決めるために必要な統計量を全て計算できない．そこで観測できない確率変数は含めず，観測できる確率変数による周辺分布を考える．直感的には観測できるデータの周辺分布がその経験分布に一致するような分布を集め，これ全体を観測データを表す候補と考える (図 2)．このように周辺分布が条件付けられた S の部分集合を観測多様体 D と呼ぶ．ここで観測多様体の各点を表すためにパラメタ $\boldsymbol{\eta}$ を導入する．D 上の全ての点は $\boldsymbol{x}_V$ について同一の周辺分布を持つので，これを $q(\boldsymbol{x}_V)$ と書くことにすると，D の各点は

$$q(\boldsymbol{x}_V, \boldsymbol{x}_H; \boldsymbol{\eta}) = q(\boldsymbol{x}_V) q(\boldsymbol{x}_H | \boldsymbol{x}_V; \boldsymbol{\eta}) \tag{20}$$

と表され，$\boldsymbol{\eta}$ は条件付き確率密度関数 $q(\boldsymbol{x}_H|\boldsymbol{x}_V; \boldsymbol{\eta})$ を定めるパラメタとみなせる．

モデル多様体の中に一点を決めるためには，観測多様体 D とモデル多様体 M が最も近くなる点を用いるのが一つの自然な方法であると考えられる．すなわち観測多様体 D の点 $q(\boldsymbol{\eta})$ とモデル多様体 M の点 $p(\boldsymbol{\theta})$ の間の距離を Kullback-Leibler 情報量

$$D(q(\boldsymbol{\eta}), p(\boldsymbol{\theta})) = \int q(\boldsymbol{x}_V, \boldsymbol{x}_H; \boldsymbol{\eta}) \log \frac{q(\boldsymbol{x}_V, \boldsymbol{x}_H; \boldsymbol{\eta})}{p(\boldsymbol{x}_V, \boldsymbol{x}_H; \boldsymbol{\theta})} d\boldsymbol{x}_V d\boldsymbol{x}_H \tag{21}$$

によって測り，これを最小にする $\hat{\boldsymbol{\eta}}$ と $\hat{\boldsymbol{\theta}}$ を求めることになる．この推定問題を e-射影と m-射影を交互に繰り返すことにより解くのが em アルゴリズムである (図 3)．

具体的な手続きは

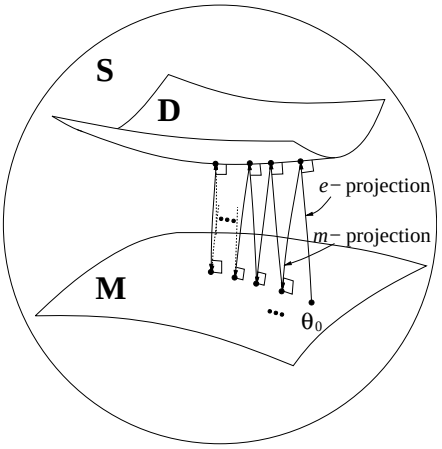

図 3: em アルゴリズム

- **e-step**

  $\boldsymbol{\theta}$ の推定量 $\boldsymbol{\theta}_t$ から D 上に e-射影を行い，$\boldsymbol{\eta}$ の推定量 $\boldsymbol{\eta}_{t+1}$ を得る．すなわち

  $$\boldsymbol{\eta}_{t+1} = \operatorname*{argmin}_{\boldsymbol{\eta}} D(q(\boldsymbol{\eta}), p(\boldsymbol{\theta}_t)) \tag{22}$$

  を計算する．

- **m-step**

  $\boldsymbol{\eta}_{t+1}$ から M 上に m-射影を行い，新たな $\boldsymbol{\theta}$ の推定量 $\boldsymbol{\theta}_{t+1}$ を得る．すなわち

  $$\boldsymbol{\theta}_{t+1} = \operatorname*{argmin}_{\boldsymbol{\theta}} D(q(\boldsymbol{\eta}_{t+1}), p(\boldsymbol{\theta})) \tag{23}$$

  を計算する．

の 2 つの部分からなる．これを適当な初期値 $\boldsymbol{\theta}_0$ から始めて，十分な繰り返しを行えば最適値に収束することが期待される．

実際の e-step は

$$\begin{aligned} D(q(\boldsymbol{\eta}), p(\boldsymbol{\theta}_t)) &= \int q(\boldsymbol{x}_V) q(\boldsymbol{x}_H|\boldsymbol{x}_V;\boldsymbol{\eta}) \log \frac{q(\boldsymbol{x}_V) q(\boldsymbol{x}_H|\boldsymbol{x}_V;\boldsymbol{\eta})}{p(\boldsymbol{x}_V;\boldsymbol{\theta}_t) p(\boldsymbol{x}_H|\boldsymbol{x}_V;\boldsymbol{\theta}_t)} d\boldsymbol{x}_V d\boldsymbol{x}_H \\ &= \int q(\boldsymbol{x}_V) \log \frac{q(\boldsymbol{x}_V)}{p(\boldsymbol{x}_V;\boldsymbol{\theta}_t)} d\boldsymbol{x}_V \\ &\quad + \int q(\boldsymbol{x}_V) q(\boldsymbol{x}_H|\boldsymbol{x}_V;\boldsymbol{\eta}) \log \frac{q(\boldsymbol{x}_H|\boldsymbol{x}_V;\boldsymbol{\eta})}{p(\boldsymbol{x}_H|\boldsymbol{x}_V;\boldsymbol{\theta}_t)} d\boldsymbol{x}_V d\boldsymbol{x}_H \\ &= \int q(\boldsymbol{x}_V) \left(\log \frac{q(\boldsymbol{x}_V)}{p(\boldsymbol{x}_V;\boldsymbol{\theta}_t)} + D(q(\boldsymbol{x}_V, \boldsymbol{\eta}), p(\boldsymbol{x}_V, \boldsymbol{\theta}_t)) \right) d\boldsymbol{x}_V \end{aligned} \tag{24}$$

を最小とする $\boldsymbol{\eta}$ を求めることになる．これは結局条件付き Kullback-Leibler 情報量 $D(q(\boldsymbol{x}_V, \boldsymbol{\eta}), p(\boldsymbol{x}_V, \boldsymbol{\theta}_t))$ を最小とする $\boldsymbol{\eta}$ を求めることに帰着され，Kullback-Leibler 情報量の正値性より

$$q(\boldsymbol{x}_H|\boldsymbol{x}_V, \boldsymbol{\eta}_{t+1}) = p(\boldsymbol{x}_H|\boldsymbol{x}_V, \boldsymbol{\theta}_t) \tag{25}$$

とすればよいことがわかる．

一方 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ものである．

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．

  $$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t) = \frac{1}{T}\sum_{k=1}^{T}\left\{\int p(\boldsymbol{x}_H|\boldsymbol{x}_{V,k}; \boldsymbol{\theta}_t)\log p(\boldsymbol{x}_{V,k}, \boldsymbol{x}_H; \boldsymbol{\theta})d\boldsymbol{x}_H\right\} \tag{26}$$

- **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．

  $$\boldsymbol{\theta}_{t+1} = \operatorname*{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t) \tag{27}$$

EM アルゴリズムを観測多様体とモデル多様体の間の動きとして説明すると，M-step では観測多様体上の一点からモデル多様体上の一点を m-射影によって求めていることになり，これは m-step と同等である．一方 E-step では条件付き期待値を計算しているが，これは e-step とは微妙に異る操作となる (図 4)．$q(\boldsymbol{x}_V)$ を観測できるデータの経験分布とし，$q(\boldsymbol{x}_H|\boldsymbol{x}_V, \boldsymbol{\eta}_{t+1}) = p(\boldsymbol{x}_H|\boldsymbol{x}_V, \boldsymbol{\theta}_t)$ を代入すれば，E-step および e-step により計算される次の M-step または m-step のための評価関数は形式的には

$$\begin{aligned}
&D(q(\boldsymbol{\eta}_{t+1}), p(\boldsymbol{\theta})) \\
&= \int q(\boldsymbol{x}_V)p(\boldsymbol{x}_H|\boldsymbol{x}_V; \boldsymbol{\theta}_t)\log\frac{q(\boldsymbol{x}_V)p(\boldsymbol{x}_H|\boldsymbol{x}_V; \boldsymbol{\theta}_t)}{p(\boldsymbol{x}_V, \boldsymbol{x}_H; \boldsymbol{\theta})}d\boldsymbol{x}_V d\boldsymbol{x}_H \\
&= \int q(\boldsymbol{x}_V)p(\boldsymbol{x}_H|\boldsymbol{x}_V; \boldsymbol{\theta}_t)\log q(\boldsymbol{x}_V)p(\boldsymbol{x}_H|\boldsymbol{x}_V; \boldsymbol{\theta}_t)d\boldsymbol{x}_V d\boldsymbol{x}_H - Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t)
\end{aligned} \tag{28}$$

となり，同一の意味を持つことがわかる．ただし数学的には積分内の経験分布 (デルタ関数) の取り扱いに問題があり，上の形式論を正当化できない場合もある．例えば S を指数型分布族とし，モデル多様体 M をその中に埋め込まれた曲指数型分布族とする場合，観測できない変数に対する平均と観測できる変数で条件付けた平均が観測多様体上で一致しない

$$E_{q(\boldsymbol{\eta})}(\boldsymbol{X}_H) \neq E_{q(\boldsymbol{\eta})}(\boldsymbol{X}_H|\boldsymbol{x}_V = E_{q(\boldsymbol{\eta})}(\boldsymbol{X}_V)) \tag{29}$$

場合に E-step と e-step が異る例が文献 [7] に述べられている．しかしながら殆どの場合 E-step と e-step は一致すると考えてよい．

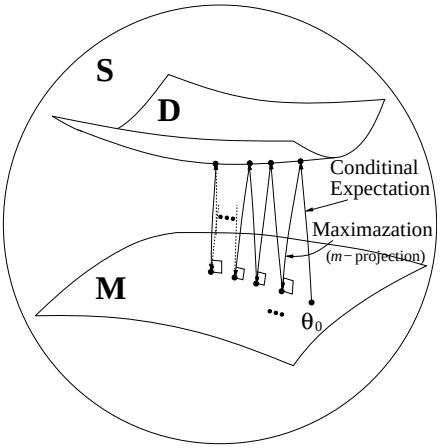

図 4: EM アルゴリズム

3　学習則としての EM アルゴリズム

神経回路網における学習は与えられた入出力信号に対して，回路網全体が適当な振舞いをするように素子間の結合状態を変化させていくことである．これは入出力の組にもとづいて回路網のパラメタを決定する統計的推定の問題としても定式化できる．通常神経回路モデルには外からは直接見えない隠れ素子が存在し，これらの素子の出力は潜在変数に対応すると考えられる．また学習に際し必要となる教師信号はこの隠れ素子に対しては陽には与えられない．神経回路網のパラメタは主に二つの素子を結ぶ結合係数であるが，その更新則は両端の素子の状態と誤差により決められるのが普通である．このため学習時には隠れ素子が潜在的に持つ誤差を計算する必要があり，この計算過程は EM アルゴリズムと密接に関係する．本節では特定の構造をもつ神経回路網に付随した学習則について EM アルゴリズムとの関係を考えてみることにする．

3.1　誤差逆伝搬学習法

代表的な神経回路網の一つに多層パーセプトロンと呼ばれる構造がある．ここでは図 5 のような n 次元入力，1 出力の 3 層パーセプトロンを例に具体的な学習法を導出してみる．入力を $\boldsymbol{x} \in R^n$，出力を $y \in R$，中間素子の出力を $\boldsymbol{z} \in R^m$ とする．神経回路網全体の入出力関係を $y = g(\boldsymbol{x})$ で表すこと

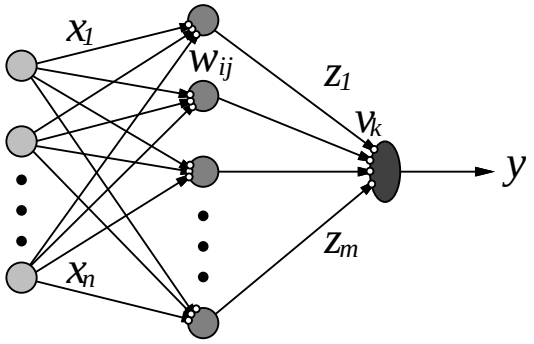

図 5: パーセプトロン

にすると，入力，中間素子の出力，および出力は次のような関数関係で結ばれる．

$$z_i = f\left(\sum_j w_{ij}x_j\right) \tag{30}$$

$$f(u) = \frac{1}{1+e^{-u}}$$

$$g(\boldsymbol{x}) = \boldsymbol{v}\cdot\boldsymbol{z} = \sum_j v_j z_j \tag{31}$$

入出力の例として $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_T, y_T)$ が与えられたとし，この入出力関係をできるだけ良く近似するよう回路網のパラメタ $W = \{w_{ij}\}, \boldsymbol{v}$ を決定することを考える．通常二乗誤差を用いた評価関数

$$E = \frac{1}{T}\sum_{k=1}^{T}(y_k - g(\boldsymbol{x}_k))^2$$

を考え，これを最小にする $W, \boldsymbol{v}$ を求めることになる．この問題は一般に非線形最適化問題になり直接法で解くことはできないので，何らかの最適化手法を用いることになるが，最も単純な勾配法を用いたものは誤差逆伝搬学習法 (error-backpropagation, バックプロパゲーション) と呼ばれる．具体的な更新則は以下のようになる．

$$v_{i,t+1} = v_{i,t} + \Delta v_i \tag{32}$$

$$w_{ij,t+1} = w_{ij,t} + \Delta v_i \tag{33}$$

$$\Delta v_i \propto -\frac{\partial E}{\partial v_i} = -2\frac{1}{T}\sum_{k=1}^{T}(y_k - g(\boldsymbol{x}_k))z_i \tag{34}$$

$$\Delta w_{ij} \propto -\frac{\partial E}{\partial w_{ij}}$$

$$= -2\frac{1}{T}\sum_{k=1}^{T}(y_k - g(\boldsymbol{x}_k))v_i f'\left(\sum_{j'} w_{ij'}x_{j',k}\right)x_{j,k} \tag{35}$$

上の式で入力・中間素子間の結合係数の更新において

$$(y_k - g(\boldsymbol{x}_k))v_i \tag{36}$$

により計算される部分が出力誤差を中間素子に分配しており，これが誤差逆伝搬と呼ばれる所以である．3 層以上の多段にしてもこの事情は同様で，基本的に入力から出力への計算手順とは逆に，誤差は出力から入力に向けて計算される．

さてここではこの問題を統計的に焼き直し，EM アルゴリズムを用いて誤差逆伝播法と同様の学習則が導かれることを示す．まず z_i と y を確率変数として取り扱うため定義を変え，

$$z_i = f\left(\textstyle\sum_j w_{ij}x_j\right) + n_i \tag{37}$$

$$y = \boldsymbol{v}\cdot\boldsymbol{z} + n \tag{38}$$

$$n_1, \ldots, n_m, n \sim N(0, \sigma^2)$$

とし，$z_1, \ldots, z_m$ を潜在変数として EM アルゴリズムを適用する．$\boldsymbol{\theta} = (W, \boldsymbol{v})$ と置くと，$y, \boldsymbol{z}$ の同時，および y の確率分布は

$$\begin{aligned} p(y, \boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}) = &\frac{1}{\sqrt{2\pi\sigma^2}^{m+1}} \\ &\times \exp\left\{-\frac{1}{2\sigma^2}(y - \boldsymbol{v}\cdot\boldsymbol{z})^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{m}\left(z_i - f\left(\textstyle\sum_j w_{ij}x_j\right)\right)^2\right\} \end{aligned} \tag{39}$$

$$\begin{aligned} p(y|\boldsymbol{x};\boldsymbol{\theta}) &= \int p(y, \boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})d\boldsymbol{z} \\ &= \frac{1}{\sqrt{2\pi(1+|\boldsymbol{v}|^2)\sigma^2}}\exp\left\{-\frac{1}{2(1+|\boldsymbol{v}|^2)\sigma^2}(y - g(\boldsymbol{x}))^2\right\} \end{aligned} \tag{40}$$

である．

この統計モデルに基づいて EM アルゴリズムを適用する．ただし M-step において $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t)$ を最大化するパラメタは直接解くことができないので，Q の勾配に沿ってパラメタを更新する方法を用いる．あるいは Q の厳密な最大化は行わずに，E-step に戻っても良い．後者は Generalized EM (GEM) アルゴリズム [8] と呼ばれる方法である．

$$\begin{aligned} p(\boldsymbol{z}|y, \boldsymbol{x};\boldsymbol{\theta}) &= \frac{p(y, \boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta})}{p(y|\boldsymbol{x};\boldsymbol{\theta})} \\ &= \frac{\sqrt{1+|\boldsymbol{v}|^2}}{(2\pi\sigma^2)^{m/2}}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{z}-\boldsymbol{r})^T(I + \boldsymbol{v}\boldsymbol{v}^T)(\boldsymbol{z}-\boldsymbol{r})\right) \end{aligned} \tag{41}$$

$$\boldsymbol{r} = \left(I - \frac{\boldsymbol{v}\boldsymbol{v}^T}{1+|\boldsymbol{v}|^2}\right)(y\boldsymbol{v} - \boldsymbol{f}), \tag{42}$$

$$\boldsymbol{f} = \left(f\left(\textstyle\sum_j w_{1j}x_j\right), \ldots, f\left(\textstyle\sum_j w_{mj}x_j\right)\right)^T \tag{43}$$

と書けることから次のようになる．ただし，^{T} は転置を表すものとする．

- **E-step**

$$
\begin{aligned}
Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t) =& \frac{1}{T} \sum_{k=1}^{T} \left\{ \int p(\boldsymbol{z}|y_k, \boldsymbol{x}_k; \boldsymbol{\theta}_t) (\log p(\boldsymbol{z}|y_k, \boldsymbol{x}_k; \boldsymbol{\theta})) d\boldsymbol{z} \right\} \\
&+ \frac{1}{T} \sum_{k=1}^{T} \log p(y_k|\boldsymbol{x}_k; \boldsymbol{\theta}) \qquad (44) \\
=& \frac{1}{T\sigma^2} \sum_{k=1}^{T} \Big\{ (\boldsymbol{r}_t - \boldsymbol{r})^T (I + \boldsymbol{v}\boldsymbol{v}^T)(\boldsymbol{r}_t - \boldsymbol{r}) \\
&+ \mathrm{tr} \left(I - \frac{\boldsymbol{v}_t \boldsymbol{v}_t^T}{1 + |\boldsymbol{v}_t|^2} \right) (I + \boldsymbol{v}\boldsymbol{v}^T) \Big\} \\
&- \frac{1}{(2\pi(1 + |\boldsymbol{v}|^2)\sigma^2)^{1/2}} E - \frac{1+m}{2} \log(2\pi\sigma^2) \qquad (45)
\end{aligned}
$$

- **M-step**

  勾配に沿ってパラメタを更新する．

$$
\Delta v_i \propto \frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t)}{\partial v_i} = -\frac{1}{T\sigma^2} \sum_{k=1}^{T} (y_k - g(\boldsymbol{x}_k)) f\left(\textstyle\sum_j w_{ij} x_{j,s} \right) \qquad (46)
$$

$$
\Delta w_{ij} \propto \frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t)}{\partial w_{ij}} = -\frac{1}{T\sigma^2} \sum_{k=1}^{T} (y_k - g(\boldsymbol{x}_k)) v_i f' \left(\sum_{j'} w_{ij'} x_{j',s} \right) x_{j,s} \qquad (47)
$$

これは定数倍を除いて誤差逆伝搬法と一致することがわかる．

3.2 ボルツマンマシン

ボルツマンマシン (Boltzmann machine) [9, 10] は与えられた信号に内在する確率構造を抽出することを目的として考えられた非常に単純な構造の神経回路網である (図 6)．モデルは全結合された n 個の素子からなり，各素子は確率的に動作し 0 または 1 の二値をとる．ボルツマンマシンの構造は均質で，各素子は動作上違いはないが，通常は環境との相互作用の違いによって外部から直接入力を受け取る入力素子，外部に出力を出す出力素子，そのどちらでもない隠れ素子の 3 種類に分けて考えられる．

具体的な動作は以下のようになる．第 i 番目の素子の内部状態を u_i とし，その出力を 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} で表わし，また自己結合はないとして，内部状態は

$$
u_i = \sum_{j \neq i} w_{ij} x_j - h_i \qquad (48)
$$

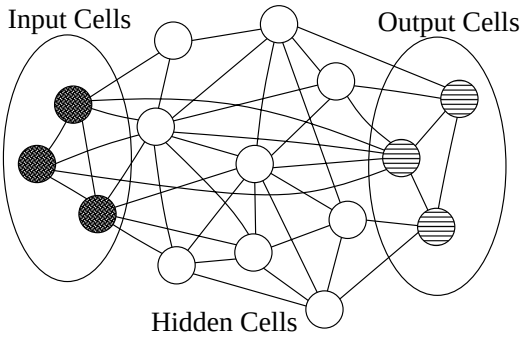

図 6: ボルツマンマシン

により計算されるものとする．すなわち自分以外の素子の出力を荷重平均して内部状態を決定する．ここで h_i は内部状態の正負の範囲を規定する閾値である．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

$$u_i = \sum_{j=0}^{n} w_{ij} x_j \tag{49}$$

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一つの素子の出力のみが x_i から x_i' に確率的に更新される．

$$\begin{aligned} p(x_i' = 1|u_i) &= \frac{1}{1 + \exp(-u_i/T)} \\ p(x_i' = 0|u_i) &= \frac{\exp(-u_i/T)}{1 + \exp(-u_i/T)} \end{aligned} \tag{50}$$

ここで 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．

容易にわかるように素子数 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看做すことができる．定義から任意の状態間の遷移確率が 0 にならないので既約であり，出力 $\boldsymbol{x}$ を取る系のエネルギーを

$$E(\boldsymbol{x}) = -\sum_{i,j} w_{ij} x_i x_j \tag{51}$$

とすると，定常分布はボルツマン分布

$$p(\boldsymbol{x}) = \frac{1}{Z} \exp\left(-\frac{E(\boldsymbol{x})}{T}\right) \tag{52}$$

で表される．これが唯一の定常分布になっていて，ボルツマンマシンの名前の由来となっている．上記のボルツマン分布が定常分布になっていることは

以下のようにして確かめられる．ある時刻において第 i 番目の素子が更新の対象になり，式 (50) にしたがい出力が x_i から x_i' に変化したと仮定する．この場合他の素子の出力は変化しないことに注意する．出力がボルツマン分布にしたがっていると仮定しているので

$$\begin{aligned}\frac{p(x_i=0)}{p(x_i=1)} &= \exp\left(-\frac{E(\ldots,x_i=0,\ldots)-E(\ldots,x_i=1,\ldots)}{T}\right)\\ &= \exp\left(-\frac{\sum_{j\neq i}w_{ij}x_j}{T}\right)\\ &= \exp\left(-u_i/T\right)\end{aligned}$$

が成り立つことに注意すると

$$\begin{aligned}p(x_i'=1) &= p(x_i'=1,x_i=1)+p(x_i'=1,x_i=0)\\ &= p(x_i'=1|x_i=1)p(x_i=1)+p(x_i'=1|x_i=0)p(x_i=0)\\ &= p(x_i'=1|u_i)p(x_i=1)+p(x_i'=1|u_i)p(x_i=0)\\ &= p(x_i'=1|u_i)\left\{1+\frac{p(x_i=0)}{p(x_i=1)}\right\}p(x_i=1)\\ &= \frac{1}{1+\exp\left(-\frac{u_i}{T}\right)}\left\{1+\exp\left(-\tfrac{u_i}{T}\right)\right\}p(x_i=1)\\ &= p(x_i=1)\end{aligned}$$

となり，確かにボルツマン分布が定常分布になっていることがわかる．ただし

$$p(x_i'=1|x_i=1)=p(x_i'=1|x_i=0)=p(x_i'=1|u_i) \tag{53}$$

はボルツマンマシンが自己結合を持たないことによる．

ボルツマンマシンの大きな特徴はその学習機構にある．以下に述べるような学習を通じて，与えられた例題を用いて外界の確率構造を回路網の中に取り込むことができる．以下では入力，出力，隠れ素子群の状態を $\boldsymbol{X}=(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma})$ にそれぞれ分けて表わすことにする．

- **Phase I**

  与えられた例題にしたがい，入出力素子を固定してボルツマンマシンを動作させる．平衡状態において素子 i と j が同時に 1 となっている確率

  $$p_{ij}=\sum_{\boldsymbol{\alpha},\boldsymbol{\beta}}q(\boldsymbol{\alpha},\boldsymbol{\beta})E(x_ix_j|\boldsymbol{\alpha},\boldsymbol{\beta}) \tag{54}$$

  を計算する．ただし $q(\boldsymbol{\alpha},\boldsymbol{\beta})$ は例題に基づく経験分布である．ε を適当な正の定数として，p_{ij} を用いて素子 i,j 間の結合係数 w_{ij} を

  $$\Delta w_{ij}=\varepsilon p_{ij} \tag{55}$$

  だけ強化する．

- **Phase II**

  入力素子は与えられた例題にしたがい固定するが，出力素子は自由に動作させることとし，平衡状態において

$$p'_{ij} = \sum_{\boldsymbol{\alpha}} q(\boldsymbol{\alpha}) E(x_i x_j | \boldsymbol{\alpha}) \tag{56}$$

  を計算する．ただし $q(\boldsymbol{\alpha})$ は例題の入力のみの経験分布である．p'_{ij} に基づいて素子 i,j 間の結合係数 w_{ij} を

$$\Delta w_{ij} = -\varepsilon p'_{ij} \tag{57}$$

  だけ弱める．

上記の二つの相を適当な初期結合から出発して繰り返すことにより，回路網は与えられた入出力を発現する確率構造を獲得することができる．この学習則は例題の入出力の確率分布とボルツマンマシンの入出力の周辺分布間の Kullback-Leibler 情報量の最小化を行う勾配法から導出されている．すなわち与えられた入出力のしたがう確率分布を $q(\boldsymbol{\alpha}, \boldsymbol{\beta})$ で，ボルツマンマシンの表す確率分布を $p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ とし，Kullback-Leibler 情報量を

$$\begin{aligned} D(\boldsymbol{w}) &= \sum_{\boldsymbol{\alpha},\boldsymbol{\beta}} q(\boldsymbol{\alpha}, \boldsymbol{\beta}) \log \frac{q(\boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\boldsymbol{\alpha}, \boldsymbol{\beta})} \\ &= \sum_{\boldsymbol{\alpha},\boldsymbol{\beta}} q(\boldsymbol{\alpha}, \boldsymbol{\beta}) \log \frac{q(\boldsymbol{\beta}|\boldsymbol{\alpha})}{p(\boldsymbol{\beta}|\boldsymbol{\alpha})} \end{aligned} \tag{58}$$

とする．ボルツマンマシンへの入力は例題に合わせて固定されるので，$q(\boldsymbol{\alpha}) = p(\boldsymbol{\alpha})$ である．定常分布がボルツマン分布で表されることに注意して結合に関する微分を求めると

$$\begin{aligned} \frac{\partial}{\partial w_{ij}} D(\boldsymbol{w}) &= \sum_{\boldsymbol{\alpha},\boldsymbol{\beta}} \frac{q(\boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\boldsymbol{\beta}|\boldsymbol{\alpha})} \frac{\partial p(\boldsymbol{\beta}|\boldsymbol{\alpha})}{\partial w_{ij}} \\ &= \sum_{\boldsymbol{\alpha},\boldsymbol{\beta}} \frac{q(\boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\boldsymbol{\beta}|\boldsymbol{\alpha})} \frac{\partial}{\partial w_{ij}} \left(\sum_{\boldsymbol{\gamma}} p(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{\alpha}) \right) \\ &= \sum_{\boldsymbol{\alpha},\boldsymbol{\beta}} \frac{q(\boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\boldsymbol{\beta}|\boldsymbol{\alpha})} \frac{\partial}{\partial w_{ij}} \left(\sum_{\boldsymbol{\gamma}} \frac{\exp\left(-E(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{\alpha})/T\right)}{\sum_{\boldsymbol{\beta}',\boldsymbol{\gamma}'} \exp\left(-E(\boldsymbol{\beta}', \boldsymbol{\gamma}'|\boldsymbol{\alpha})/T\right)} \right) \\ &= -\frac{1}{T} \sum_{\boldsymbol{\alpha},\boldsymbol{\beta}} \frac{q(\boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\boldsymbol{\beta}|\boldsymbol{\alpha})} \left\{ \sum_{\boldsymbol{\gamma}} p(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{\alpha}) x_i x_j - p(\boldsymbol{\beta}|\boldsymbol{\alpha}) \sum_{\boldsymbol{\beta}',\boldsymbol{\gamma}'} p(\boldsymbol{\beta}', \boldsymbol{\gamma}'|\boldsymbol{\alpha}) x_i x_j \right\} \\ &= -\frac{1}{T} \left\{ \sum_{\boldsymbol{\alpha},\boldsymbol{\beta}} q(\boldsymbol{\alpha}, \boldsymbol{\beta}) \sum_{\boldsymbol{\gamma}} p(\boldsymbol{\gamma}|\boldsymbol{\alpha}, \boldsymbol{\beta}) x_i x_j - \sum_{\boldsymbol{\alpha}} q(\boldsymbol{\alpha}) \sum_{\boldsymbol{\beta},\boldsymbol{\gamma}} p(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{\alpha}) x_i x_j \right\} \\ &= -\frac{1}{T} \left\{ p_{ij} - p'_{ij} \right\} \end{aligned}$$

となり

$$\Delta w_{ij} \propto -\frac{\partial}{\partial w_{ij}} D(\boldsymbol{w}) \propto p_{ij} - p'_{ij} \tag{59}$$

にしたがい，w_{ij} を変化させることにより，Kullback-Leibler 情報量は減少する．実際の計算ではボルツマンマシンを動作させることにより一種の Monte-Carlo 法を行い p_{ij},p'_{ij} を求めていることになる．

EM アルゴリズムと比較すると Phase I は条件付き確率 $p(\boldsymbol{\gamma}|\boldsymbol{\alpha}, \boldsymbol{\beta})$ を用いた統計量を計算しており，E-step に相当する計算を行っていると看倣せる．また Phase II により最終的に最急降下方向が計算され，Kullback-Leibler Diverngence の最小化が行われるので，これが M-step に対応している．この対応関係は em アルゴリズムと比較しても同様である．

ボルツマンマシンは人間の認識機構の非常に単純なモデルとして考案された．例えば自然画像は近傍の画素間にある種の関係が存在し，良く見知った画像であれば雑音により汚されていても我々は容易に元の画像を推測することができる．このような確率的構造を取り込む機能を説明するモデルとして考えられたものである．統計的モデルとしては非常に単純ではあるが，特殊な構造を入れることにより局所的な情報で効率のよい学習を行っていることがわかる．ボルツマンマシンをそのまま脳のモデルとして捉えることにはもちろん難はあるが，記銘 (Phase I) と忘却 (Phase II) を交互に繰り返すことによって信号に潜む確率構造を取得し記憶するモデルとして，人間の認識・記憶のメカニズムを考える上で示唆するところが大きい．

3.3 ヘルムホルツマシン

ヘルムホルツマシン (Helmholtz machine) [11, 12, 13] 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，一種の認知モデルである．統計的には因子分析に相当すると考えられるが，一般には非線型回帰による分析を行っており，また近傍の素子間での相互関係のみによる局所的な学習を行っているため，実は統計的には少々無理のあるモデルとなっている．学習則は Wake-Sleep アルゴリズムと呼ばれる生成モデルと認識モデル間での交代的な繰り返し演算を用いる．以下では生成・認識モデルとして単純な線形モデルを用いた場合の Wake-Sleep アルゴリズムを文献 [14] に従い説明する．まず回路網の構造を概説する．この場合ヘルムホルツマシンは因子分析のモデルと一致する．

- **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

  $$\boldsymbol{x} = \boldsymbol{g}y + \boldsymbol{\varepsilon} \tag{60}$$

  により生成されるとする．ここで $\boldsymbol{\varepsilon}$ 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．また $\boldsymbol{g}$ は因子負荷と等しい．

- **認識モデル**

  観測された信号 $\boldsymbol{x}$ から対応する y を

$$y = \boldsymbol{r}^T \boldsymbol{x} + \delta \tag{61}$$

  にしたがい推測するとする．ただし δ は $N(0, s^2)$ にしたがう雑音である．

このモデルの目的は観測データ $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ が与えられたとき，このデータを記述する最適な生成・認識モデルを求めることである．Wake-Sleep アルゴリズムでは以下の 2 つの phase によってパラメタ $\boldsymbol{g}, \Sigma, \boldsymbol{r}, s$ の学習が行われる．

- **Wake-phase**

  観測データ $\{\boldsymbol{x}_i\}$ より $\boldsymbol{x}$ をいくつかランダムに抽出し (全部選んでもよいし，同じものを何度か選んでもよい)，各 $\boldsymbol{x}$ に対して認識モデル

$$y = \boldsymbol{r}_t^T \boldsymbol{x} + \delta, \quad \delta \sim N(0, s_t^2) \tag{62}$$

  にしたがい y を生成する．こうして得られる $(\boldsymbol{x}, y)$ の組を用いて以下のように $\boldsymbol{g}$ および Σ を更新する．

$$\boldsymbol{g}_{t+1} = \boldsymbol{g}_t + \alpha \left\langle (\boldsymbol{x} - \boldsymbol{g}_t y) y \right\rangle \tag{63}$$

$$\sigma_{i,t+1}^2 = \beta \sigma_{i,t}^2 + (1 - \beta) \left\langle (x_i - g_{i,t} y)^2 \right\rangle \tag{64}$$

  ただし α は正の定数，β は 1 より小さい正の定数である．また $\langle \cdot \rangle$ は抽出された観測データ $\boldsymbol{x}$ とそれに対応して生成された y に関する平均を表す．

- **Sleep-phase**

  標準正規分布にしたがい y を発生し，生成モデル

$$\boldsymbol{x} = \boldsymbol{g}_t y + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \Sigma_t) \tag{65}$$

  によって擬似的な観測データ $\boldsymbol{x}$ を作る．こうして得られる $(\boldsymbol{x}, y)$ を用いて生成モデルのパラメタ $\boldsymbol{r}$, s^2 を更新する．

$$\boldsymbol{r}_{t+1} = \boldsymbol{r}_t + \alpha' \left\langle \boldsymbol{x} (y - \boldsymbol{r}_t^T \boldsymbol{x}) \right\rangle \tag{66}$$

$$s_{t+1}^2 = \beta' s_t^2 + (1 - \beta') \left\langle (y - \boldsymbol{r}_t^T \boldsymbol{x})^2 \right\rangle \tag{67}$$

$$\tag{68}$$

  ただし $\langle \cdot \rangle$ は生成したデータに関する平均を表す．

端的に言えば bootstrap 法と Monte-Carlo 法の組み合わせによりパラメタ推定を行っているとも考えられるが，この学習則の特徴は局所的な情報で構成されていることである．例えば $\boldsymbol{g}$ の第 i 成分 g_i は $\boldsymbol{x}$ の第 i 成分 x_i と y 間の結合を表しているが，学習における更新量は

$$g_{i,t+1} = g_{i,t} + \alpha \langle (x_i - g_{i,t}y)y \rangle \tag{69}$$

となり結合の両端の値だけで計算される．こうした制約は計算上はあまり意味のないものであるが，生体の情報処理のモデルとしての妥当性を考える場合は非常に重要となる．何故なら学習則に大域的な情報が現れる場合，その情報を取得するための生物学・解剖学的な観点からの妥当な機構が必要となるからである．そうした情報伝達機構が生体内に存在する場合は問題ないが，存在しない場合には生体の計算モデルとしては不適切なものとなってしまう．

Wake-Sleep アルゴリズムは幾何学的には次の二つの Kullback-Leibler 情報量を交互に小さくしていると解釈できる．まず生成モデルにおける $\boldsymbol{x}$ と y の同時確率密度関数を $\boldsymbol{\theta} = (\boldsymbol{g}, \Sigma)$ として

$$\begin{aligned} p(y, \boldsymbol{x}; \boldsymbol{\theta}) &= \exp\left(-\frac{1}{2} \begin{pmatrix} y & \boldsymbol{x}^T \end{pmatrix} A \begin{pmatrix} y \\ \boldsymbol{x} \end{pmatrix} - \psi(\boldsymbol{\theta}) \right) \\ A &= \left(\begin{array}{c|c} 1 + \boldsymbol{g}^T \Sigma^{-1} \boldsymbol{g} & -\boldsymbol{g}^T \Sigma^{-1} \\ \hline -\Sigma^{-1} \boldsymbol{g} & \Sigma^{-1} \end{array} \right) \\ \psi(\boldsymbol{\theta}) &= \frac{1}{2} \left(\sum \log \sigma_i^2 + (n+1) \log 2\pi \right) \end{aligned} \tag{70}$$

と書く．また認識モデルにおいては，$\boldsymbol{x}$ で条件付けた y の確率密度関数は正規分布 $N(\boldsymbol{r}^T \boldsymbol{x}, s^2)$ にしたがい，$\boldsymbol{x}$ の確率密度は観測データ $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ から計算される共分散行列

$$C = \frac{1}{N} \sum_{s=1}^{N} \boldsymbol{x}_s \boldsymbol{x}_s^T, \tag{71}$$

によって定義される正規分布 $N(0, C)$ にしたがうので，その同時確率密度は $\boldsymbol{\eta} = (\boldsymbol{r}, s^2)$ として

$$\begin{aligned} q(y, \boldsymbol{x}; \boldsymbol{\eta}) &= \exp\left(-\frac{1}{2} \begin{pmatrix} y & \boldsymbol{x}^T \end{pmatrix} B \begin{pmatrix} y \\ \boldsymbol{x} \end{pmatrix} - \psi(\boldsymbol{\eta}) \right) \\ B &= \frac{1}{s^2} \left(\begin{array}{c|c} 1 & -\boldsymbol{r}^T \\ \hline -\boldsymbol{r} & s^2 C^{-1} + \boldsymbol{r}\boldsymbol{r}^T \end{array} \right) \\ \psi(\boldsymbol{\eta}) &= \frac{1}{2} \left(\log s^2 + \log |C| + (n+1) \log 2\pi \right) \end{aligned} \tag{72}$$

と書ける．この二つの分布を用いて Wake-phase は

$$D(q(\boldsymbol{\eta}), p(\boldsymbol{\theta})) = E_{q(\boldsymbol{\eta})} \left(\log \frac{q(y, \boldsymbol{x}; \boldsymbol{\eta})}{p(y, \boldsymbol{x}; \boldsymbol{\theta})} \right) \tag{73}$$

を $\boldsymbol{\theta}$ に関して減少させ，Sleep-phase は

$$D(p(\boldsymbol{\theta}), q(\boldsymbol{\eta})) = E_{p(\boldsymbol{\theta})}\left(\log \frac{p(y, \boldsymbol{x}; \boldsymbol{\theta})}{q(y, \boldsymbol{x}; \boldsymbol{\eta})}\right) \tag{74}$$

を $\boldsymbol{\eta}$ に関して減少させていると看做すことができるため，幾何学的には e-射影を繰り返し用いていることになる．

ここで扱った単純な線形モデルの場合は学習のダイナミクスは EM および em アルゴリズムとは異るが，収束点は最尤推定量になる．しかしながら，例えば

$$\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{g}y) + \boldsymbol{\varepsilon} \tag{75}$$

$$y = h(\boldsymbol{r}^T \boldsymbol{x}) + \delta \tag{76}$$

で表されるような非線形変換をともなうモデルが用いられた場合，Wake-Sleep アルゴリズムは一般に最尤推定量には収束しない．

ヘルムホルツマシンは構造としてはボルツマンマシンより複雑なモデルとなっているが，信号の認識と生成・想起というはっきりした役割を異る部位に割り当て，連関されることにより認知モデルとしては非常に魅力的なものになっている．また覚醒・睡眠状態で異った学習を行うものとしても興味深いモデルであるが，数学上は学習の安定性，収束点の性質等解決すべき様々な問題がある．もちろん大域的な情報を用いれば，最尤推定量と一致する形に Wake-Sleep アルゴリズムを変更できるが，生体における認識モデルとしての立場からはこれはあまり望ましくない．

4　EM アルゴリズムを用いるモジュール型モデル

本節では陽に EM アルゴリズムを用いて学習を行う神経回路モデルの例を概説する．前節の例ではモデルの中で一部の素子の取る値が見えないという状況を想定したものであったが，これとは異り神経回路網内が幾つかのモジュールに分けられ，入力信号に応じて担当するモジュールが異るといった状況が考えられている．この場合どのモジュールが担当すべきかということが隠れ変数の役割を果すことになる．

4.1　Mixtures of Experts

Mixtures of Experts [15, 16] はいくつかのモジュールを協調・競合させて情報処理を行うための階層型の神経回路モデルである．

モデルは expert network と gating network の二つからなる．さらに多段にして用いる場合 (hierarchical mixtures of experts) もあるが，ここでは最も単純な場合を説明する．

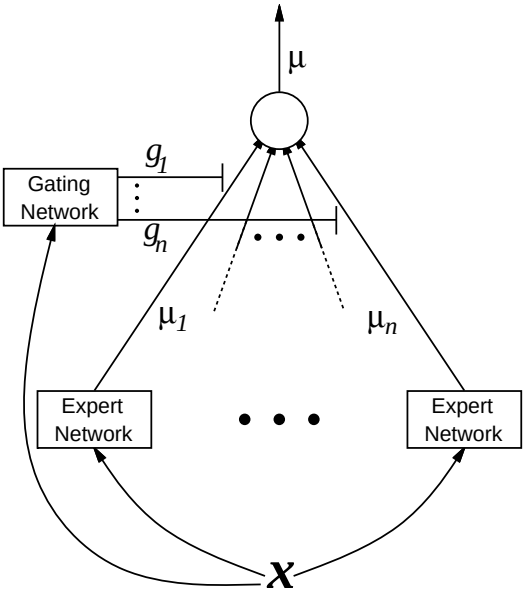

図 7: Mixtures of Experts

- **expert network**

  各 network は入力 $\boldsymbol{x} \in R^m$ を受けて $\boldsymbol{\mu}_i \in R^d$ を出力するものとする．

$$\boldsymbol{\mu}_i = \boldsymbol{f}_i(\boldsymbol{x}; \boldsymbol{\theta}_i), \quad i = 1, \ldots, K \tag{77}$$

  特に構造は規定しないが，応用上は単純な線形回帰モデルや多層パーセプトロンが用いられる．前者の場合パラメタ $\boldsymbol{\theta}_i$ は回帰係数に対応し，また後者の場合パラメタは結合係数に対応する．

  通常実際に観測される出力には正規雑音が加わっていると考え，network の出力はある正規分布の平均を表していると考える．すなわち $\boldsymbol{n}$ を正規雑音として

$$\boldsymbol{y} = \boldsymbol{\mu}_i + \boldsymbol{n} \tag{78}$$

  と考える．この場合 $\boldsymbol{y}$ の条件付き確率は

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_i) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{f}_i(\boldsymbol{x}; \boldsymbol{\theta}_i))^T \Sigma_i^{-1} (\boldsymbol{y} - \boldsymbol{f}_i(\boldsymbol{x}; \boldsymbol{\theta}_i))\right) \tag{79}$$

  で表される．

- **gating network**

  適当な関数 $s_i(\boldsymbol{X}; \boldsymbol{\theta}_0)$ を用いて

$$g_i(\boldsymbol{x}; \boldsymbol{\theta}_0) = \frac{\exp\left(s_i(\boldsymbol{x}; \boldsymbol{\theta}_0)\right)}{\sum_{j=1}^{K} \exp\left(s_j(\boldsymbol{x}; \boldsymbol{\theta}_0)\right)} \tag{80}$$

なる競合系を構成する．ここでも s_i として良く用いられるのは多層パーセプトロンである．g_i は正で，その和は 1 になっているので，(g_i) は確率ベクトルと考えることができ，回路網の出力はこの確率にしたがい一つの expert network を選ぶか，重み付け平均を取るかして決定される．

学習は回路網への入出力の組 $\{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_T, \boldsymbol{y}_T)\}$ が例題として与えられたとき，g_i が観測できない確率変数であるとして，EM アルゴリズムを用いる．以下に応用上良く用いられる expert network を線形回帰モデルとした場合の具体的な更新則を記述しておく [17].

- **E-step**

  各入出力の例題 $(\boldsymbol{x}_k, \boldsymbol{y}_k)$ について条件付き確率

  $$p(i|\boldsymbol{x}, \boldsymbol{y}) = \frac{g_i(\boldsymbol{x}, \boldsymbol{\theta}_{0,t}) p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_{i,t})}{\sum_{j=1}^{n} g_j(\boldsymbol{x}, \boldsymbol{\theta}_{0,t}) p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_{j,t})} \tag{81}$$

  を計算する

- **M-step**

$$\Sigma_{i,t+1} = \frac{\sum_{k=1}^{T} p(i|\boldsymbol{x}_k, \boldsymbol{y}_k)(\boldsymbol{y}_k - \boldsymbol{f}_i(\boldsymbol{x}_k; \boldsymbol{\theta}_{i,t}))(\boldsymbol{y}_k - \boldsymbol{f}_i(\boldsymbol{x}_k; \boldsymbol{\theta}_{i,t}))^T}{\sum_{k=1}^{T} p(i|\boldsymbol{x}_k, \boldsymbol{y}_k)} \tag{82}$$

$$\boldsymbol{\theta}_{i,t+1} = R_{i,t}^{-1} \boldsymbol{e}_{i,t} \tag{83}$$

$$\boldsymbol{e}_{i,t} = \sum_{k=1}^{T} p(i|\boldsymbol{x}_k, \boldsymbol{y}_k) X_k \Sigma_{i,t}^{-1} \boldsymbol{y}_k$$

$$R_{i,t} = \sum_{k=1}^{T} p(i|\boldsymbol{x}_k, \boldsymbol{y}_k) X_k \Sigma_{i,t}^{-1} X_k^T$$

$$X_k^T = \left(\begin{array}{cccc|cccc} \boldsymbol{x}_k^T & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & & \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & \cdots & \boldsymbol{x}_k^T & 0 & 0 & \cdots & 1 \end{array} \right)$$

$$\boldsymbol{\theta}_{0,t+1} = \boldsymbol{\theta}_{0,t} + \delta R_{0,t}^{-1} \boldsymbol{e}_{0,t} \tag{84}$$

$$\boldsymbol{e}_{0,t} = \sum_{k=1}^{T} \sum_{i=1}^{K} \left(p(i|\boldsymbol{x}_k, \boldsymbol{y}_k) - g_i(\boldsymbol{x}_k; \boldsymbol{\theta}_{0,t}) \right) \frac{\partial s_i(\boldsymbol{x}_k; \boldsymbol{\theta}_{0,t})}{\partial \boldsymbol{\theta}_0}$$

$$R_{0,t} = \sum_{k=1}^{T} \sum_{i=1}^{K} g_i(\boldsymbol{x}_k; \boldsymbol{\theta}_{0,t}) \left(1 - g_i(\boldsymbol{x}_k; \boldsymbol{\theta}_{0,t}) \right) \frac{\partial s_i(\boldsymbol{x}_k; \boldsymbol{\theta}_{0,t})}{\partial \boldsymbol{\theta}_0} \frac{\partial s_i(\boldsymbol{x}_k; \boldsymbol{\theta}_{0,t})^T}{\partial \boldsymbol{\theta}_0}$$

この他オンライン学習や収束性，安定性などを考慮に入れたいくつかの変種も提案されている．

4.2 Normalized Gaussian network

Normalized Gaussian nerwork[18] は m 次元入力を d 次元出力に変換する 1 次のスプライン関数と radial basis function network の双方の性質を持つ神経回路網である．生体の，特に感覚機能を司る神経細胞は，ある特定の信号にのみ強く反応しそれから離れた信号には反応しないといった特性を持つことが多いが，これは受容野と呼ばれる．Radial basis function network はこれを非常に単純な形で実現したものである．

Normalized Gaussian nerwork の構造は

$$
\begin{aligned}
y &= \sum_{i=1}^{M} (W_i \boldsymbol{x} + \boldsymbol{b}_i) N_i(\boldsymbol{x}) \qquad (85)\\
N_i(\boldsymbol{x}) &= \frac{G_i(\boldsymbol{x})}{\sum_{j=1}^{M} G_j(\boldsymbol{x})}\\
G_i(\boldsymbol{x}) &= \frac{1}{\sqrt{(2\pi)^m |\Sigma_i|}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i) \right\}
\end{aligned}
$$

で規定されるが，N_i が Radial basis function network にあたる．ここに M は回路網を構成する素子の個数である．

一つの入出力に対して必ず一つの素子が選ばれるものと仮定し，これを $(\boldsymbol{x}, \boldsymbol{y}, i)$ で表すことにする．この場合 $(\boldsymbol{x}, \boldsymbol{y})$ が観測値に i が隠れ変数に対応する．完全データ $(\boldsymbol{x}, \boldsymbol{y}, i)$ の出現確率を

$$
\begin{aligned}
p(\boldsymbol{x}, \boldsymbol{y}, i; \boldsymbol{\theta}) = \frac{1}{M\sqrt{(2\pi)^{(m+d)} \sigma_i^{2d} |\Sigma_i|}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i) \right\}\\
\times \exp\left\{ -\frac{1}{2\sigma_i^2} (\boldsymbol{y} - W_i \boldsymbol{x} - \boldsymbol{b}_i)^2 \right\} \qquad (86)
\end{aligned}
$$

で表す．但し $\boldsymbol{\theta} = \{\boldsymbol{\mu}_i, \Sigma_i, \sigma_i^2, W_i, \boldsymbol{b}_i; i = 1, \ldots, M\}$，$\sigma_i^2$ は素子 i が選ばれた際に出力 $\boldsymbol{y}$ が平均 $W_i \boldsymbol{x} + \boldsymbol{b}_i$ のまわりでばらつく大きさを表す．この時

$$
\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{x}, i; \boldsymbol{\theta}) &= \frac{1}{\sqrt{2\pi\sigma_i^{2d}}} \exp\left\{ \frac{1}{2\sigma_i^2} (\boldsymbol{y} - W_i \boldsymbol{x} - \boldsymbol{b}_i)^2 \right\} \qquad (87)\\
p(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta}) &= \sum_{i=1}^{M} N_i(\boldsymbol{x}) p(\boldsymbol{y}|\boldsymbol{x}, i, \boldsymbol{\theta}) \qquad (88)
\end{aligned}
$$

である．この条件確率を用いて，入力が $\boldsymbol{x}$ である時の出力 $\boldsymbol{y}$ の期待値は

$$
E(\boldsymbol{y}|\boldsymbol{x}) = \int \boldsymbol{y} p(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta}) d\boldsymbol{y} = \sum_{i=1}^{M} (W_i \boldsymbol{x} + \boldsymbol{b}_i) N_i(\boldsymbol{x}) \qquad (89)
$$

で与えられ，これは回路網の出力と一致する．したがって回路網は出力の期待値を計算していると考えられる．

Normalized Gaussian nerwork を確率モデルと考えると，入出力の観測データの組から以下の EM アルゴリズムに基づいた逐次的方法を用いて最尤推定量を求めることができる [19].

- **E-step**

  パラメタ $\boldsymbol{\theta}_t$ の下で，観測データ $(\boldsymbol{x}, \boldsymbol{y})$ に対して素子 i が選ばれる事後確率を

  $$p(i|\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}_t) = \frac{p(\boldsymbol{x}, \boldsymbol{y}, i; \boldsymbol{\theta}_t)}{\sum_{j=1}^{M} p(\boldsymbol{x}, \boldsymbol{y}, j; \boldsymbol{\theta}_t)} \tag{90}$$

  によって計算する.

- **M-step**

  事後確率を用いた完全データに関する期待対数尤度を

  $$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t) = \sum_{k=1}^{T} \sum_{i=1}^{M} p(i|\boldsymbol{x}_k, \boldsymbol{y}_k; \boldsymbol{\theta}_t) \log p(i|\boldsymbol{x}_k, \boldsymbol{y}_k; \boldsymbol{\theta}) \tag{91}$$

  により定義し，これを $\boldsymbol{\theta}$ に関して最大化する．具体的には観測データの下での事後確率による重み付き平均

  $$E_i(f(\boldsymbol{x}, \boldsymbol{y})) = \frac{1}{T} \sum_{k=1}^{T} f(\boldsymbol{x}_k, \boldsymbol{y}_k) p(i|\boldsymbol{x}_k, \boldsymbol{y}_k; \boldsymbol{\theta}_t) \tag{92}$$

  と書くことにして

  $$\boldsymbol{\mu}_{t+1} = \frac{E_i(\boldsymbol{x})}{E_i(1)} \tag{93}$$

  $$\Sigma_{i,t+1} = \frac{E_i((\boldsymbol{x} - \boldsymbol{\mu}_{i,t})(\boldsymbol{x} - \boldsymbol{\mu}_{i,t})^T)}{E_i(1)} \tag{94}$$

  $$\tilde{W}_{i,t+1} = E_i(\boldsymbol{y}\tilde{\boldsymbol{x}}^T) E_i(\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T)^{-1} \tag{95}$$

  $$\sigma^2_{i,t+1} = \frac{1}{d} \frac{E_i(|\boldsymbol{y} - \tilde{W}_{i,t}\tilde{\boldsymbol{x}}|^2)}{E_i(1)} \tag{96}$$

  で与えられる．ただし

  $$\tilde{W}_i = (W_i, \boldsymbol{b}_i), \quad \tilde{\boldsymbol{x}}^T = (\boldsymbol{x}^T, 1)$$

  である.

実際の場面ではデータの観測とパラメタの更新は交互に行われる，いわゆるオンライン学習を用いる場合が多い．これは変化する環境に追従するような場合を想定しており，その場合 E_i は low-pass filter などで置き換えられる．なお具体的な応用例としては，ロボットの制御に用いた文献 [20] などを参照されたい.

参考文献

[1] D. Rumelhart, J. L McClelland, and the PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. The MIT Press, 1986.

[2] G. Cybenko. Approximation by superpositions of a sigmoid function. *Mathematics of Control, Signals and Systems*, 2:303–314, 1989.

[3] Shun-ichi Amari. *Differential-Geometrical Methods in Statistics*, volume 28 of *Lecture Notes in Statistics*. Springer-Verlag, 1985.

[4] O. Barndorff-Nielsen. *Parametric Statistical Models and Likelihood*, volume 50 of *Lecture Notes in Statistics*. Springer-Verlag, 1988.

[5] M. K. Murrey and J. W. Rice. *Differential Geometry and Statistics*. Chapman, 1993.

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

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

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

[9] G. E. Hinton and T. J. Sejnowski. Optimal perceptual inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 448–453, 1983.

[10] D. H. Ackley, Geoffrey E. Hinton, and T. J. Sejnowski. A learning algorithm for Bopltzmann machines. *Cognitive Science*, 9:147–169, 1985.

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

[12] 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.

[13] Radford M. Neal and Peter Dayan. Factor analysis using delta-rule wake-sleep learning. *Neural Computation*, 9(8):1781–1803, November 1997.

[14] Shiro Ikeda, Shun-ichi Amari, and Hiroyuki Nakahara. Convergence of the Wake-Sleep algorithm. submitted NIPS98.

[15] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991.

[16] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214, 1994.

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

[18] John Moody and Christian Darken. Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1:289–303, 1989.

[19] 石井 信 and 佐藤 雅昭. オンライン EM アルゴリズムによる動的な関数近似. 信学技報 NLP97-142,NC97-94, 電子情報通信学会, 1998.

[20] Kenji Doya. Efficient nonlinear control with actor-tutor architecture. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 9, pages 1012–1018, Cambridge, MA, 1997. The MIT Press.