隠れ変数と階層的学習

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

ここでは，確率分布における隠れ変数の考え方を説明する．隠れ変数とは，外からは直接観測できない確率変数である．

一つ例を考えよう．$0, 1, 2, \ldots, 9$ の数字が書かれた紙が沢山あるとしよう．そのうち一枚をとってきたとする．我々はこれを見て，今書かれている数字が 1 なのか 7 なのか，または 4 なのか，10 のカテゴリーのうちの一つに分類したい．

どの数字が表現されているのかを確率変数として考えてみる．すると，これは直接は観測できない確率変数であることが分る．我々は数字の書かれた紙を通じて何が書かれているのか推定しなければいけない．

図 1: 活字の場合

図 2: あいまいな数字

数字を見る前は，それぞれが等確率で出てくると予測しても構わないだろう．仮りに図 1 のようなデータを観測したのなら，ほとんど迷うことなく数字きた数字を当てることができる．

では，図 2 のような図が現れた場合はどうだろう．これはおそらく 1 だろうと考えられるが，7 である可能性も拭えない．つまり，数字を書いた人がどう思って書いたのかは直接は観測できないのである．ここまで書くと，この問題に隠れ変数を用いる意味が分るだろう．

実際の文字認識装置などを構成する場合，上のような問題をどうやって解決するのかは難しい問題であり，ここでは論じない．その代わりに，このような問題を情報量，確率を使って見るとどうなるのかを説明する．結果として，見通しの良い議論ができるようになると考える．

2 エントロピーと情報量

2.1 エントロピー

ある確率変数 X を考える．この確率変数の密度関数が $p(X)$ で与えられるとする．このとき，エントロピーは，

$$H(X) = -\int p(x) \log p(x) dx \tag{1}$$

と定義される．X が離散値を取る場合には積分記号は単に和となり，

$$H(X) = -\sum_x p(x)\log p(x) \tag{2}$$

と定義される．

数字の例で考えると，10 個の数字が等確率で出現するならば，X をどの数字が出現するかという確率変数と置くと，

$$\begin{aligned} p(x=0) &= \frac{1}{10}, \cdots, p(x=9) = \frac{1}{10} \\ H(X) &= -\sum \frac{1}{10}\log\frac{1}{10} = \log 10 \end{aligned}$$

となる．エントロピーとはある確率変数のもつ「曖昧さ」と考えてみても良い．この場合，どの数字が書かれているかという確率変数 X の持つエントロピー，すなわち曖昧さは $\log 10$ である．

2.2 相互情報量

エントロピーの考え方について，もう少し説明を続ける．

今，図 1 を観測したとする．ここで，新たな確率変数 Y を考える．Y はどのような絵が観測されたかを示す確率変数である．例えば図 1 を観測したのであれば，Y は図 1 となる．Y は様々な図を取り得るので，確率変数として考えても良いだろう．

Y を観測したときの X の分布はどうなるだろうか．これは X の Y に対する条件付き確率 $p(X|Y)$ である．例えば図 1 の場合には書き手の意図としての数字は，確率 1 で 「3」 であろう．

これを確率で書くならば，

$$\begin{aligned} &p(x \in \{0,1,2,4,5,6,7,8,9\} | y = \boxed{3}) = 0, \\ &p(x = 3 | y = \boxed{3}) = 1. \end{aligned} \tag{3}$$

すると，この図を観測したときのエントロピーは，

$$H(X | y = \boxed{3}) = 0 \tag{4}$$

となる (計算中，$0\log 0 = 0$ とした) すなわち，図を観測することで，曖昧さが消えエントロピーが 0 となり，結果として，$\log 10$ であったエントロピーが 0 となった．このとき，$\log 10 - 0$ の “情報量” が得られたことになる．

一方図 2 のような曖昧な図を観測したときは，

$$H(X | y = \boxed{7})$$

は 0 とならないだろう．どのように判断するかは別にして，これが 1 である確率が 90% で，7 である確率は 10%とするならば，

$$H(X | y = \boxed{7}) = -\frac{1}{10}\log\frac{1}{10} - \frac{9}{10}\log\frac{9}{10} = \log 10 - \frac{9}{10}\log 9 \tag{5}$$

となる．得られた情報量は $\log 10$ との差であるから，$\frac{9}{10}\log 9$ となり，依然として残っている曖昧さは $\log 10 - \frac{9}{10}\log 9$ である．

このように観測される図 Y と隠れた確率変数 X の間には相互に関係している．このような 2 つの確率変数の間に，相互情報量を考えることができる．Y の密度関数を $q(Y)$ とすると，相互情報

量は Y を観測したときに，平均としてどのくらい情報量を得られるかとして定義できる．

$$\begin{aligned} I(X,Y) &= H(X) - \int q(y) H(X|y) dy \\ &= H(X) - H_Y(X) \end{aligned} \tag{6}$$

ここで，

$$H_Y(X) = \int q(y) \left[-\int p(x|y) \log p(x|y) dx \right] dy \tag{7}$$

と定義した．$p(X)$ と $q(Y)$ が定義され，さらに $p(X|Y)$ が定義されれば X と Y の同時分布の密度関数を $r(X,Y) = p(X|Y)q(Y)$ として定義できる．これを用いると，

$$\begin{aligned} I(X,Y) &= H(X) - H_Y(X) \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(Y) - H_X(Y) \end{aligned} \tag{8}$$

と書ける．ここで，

$$H(X,Y) = -\int r(x,y) \log r(x,y) dx dy \tag{9}$$

とした．この結果から相互情報量は X, Y に関して対称であることがわかる．

エントロピーや相互情報量を短い紙面で考えてみたが，ここに書いてあることのみでは十分ではない．情報理論の教科書を参考にして欲しい [1, 2]．

3 最尤推定

今までは，確率分布が定義されている上で，いくつかの事柄について説明をした．ここでは，ある程度確率分布が分っているが，そのパラメータが未知の場合，パラメータをどうやって求めるかについて，特に統計的な推定法である最尤推定を説明する．

例えば，データが正規分布に従っているということは既知だとする．しかし，平均値と分散が幾つかは分らないとする．このときに，データからパラメータをどうやって推定するかといったことにあたる．

$\boldsymbol{\theta}$ をパラメータとする確率分布 $p(y;\boldsymbol{\theta})$ を考える．データが同じ確率分布から，独立に T 個 $\{y_1, y_2, \ldots, y_T\}$ として得られたとする．このとき，$\boldsymbol{\theta}$ を推定したい．最尤推定では，$p(y;\boldsymbol{\theta})$ 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．

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_s p(y_s;\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_s \log p(y_s;\boldsymbol{\theta}) \tag{10}$$

正規分布であれば，$\boldsymbol{\theta} = (\mu, \sigma^2)$ (平均と分散) であり，

$$p(y;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{(-\frac{(y-\mu)^2}{\sigma^2})}$$

である．対数尤度は

$$\sum_s \log p(y;\boldsymbol{\theta}) = -\sum_s \frac{(y_s-\mu)^2}{\sigma^2} - \frac{T}{2} \log 2\pi\sigma^2$$

となり，これを最大にする μ,σ^2 は

$$\mu^* = \frac{1}{T}\sum_s y_s, {\sigma^2}^* = \frac{1}{T}\sum_s (y_s - \mu^*)^2$$

となる．

ここで，便利のために経験分布を $q(y)$ と定義する．例えば y が離散値を取るのであれば，

$$q(y) = \frac{1}{T}\sum_{i=1}^{T} \delta_{y_i}(y)$$

とすれば良い．$\delta_{y_i}(y)$ は $y = y_i$ のときに 1 を取る関数である．

2 つの確率分布の間の Kullback-Leibler Divergence は(11) のように定義される．Kullback-Leibler Divergence はお互いの確率分布が一致したときのみ 0 になり，それ以外は正の値を取る．したがって分布間の違いを表わす．最尤推定は経験分布 $q(y)$ と $p(y;\theta)$ との間の Kullback-Leibler Divergence を最小にするようにパラメータを求めるのだと考えられる．

$$\begin{aligned} D(q, p(\theta)) &= \int q(y) \log \frac{q(y)}{p(y;\boldsymbol{\theta})} dy \\ &= \int q(y) \log q(y) dy - \int q(y) \log p(y;\boldsymbol{\theta}) dy \end{aligned} \tag{11}$$

右側の式の第 2 項は(10) 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(11) 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 [3] を用いて解釈する．図 3 はこのイメージを示したものである．図中の S は 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 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 $q(y)$ 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 $D(q(y), p(y;\boldsymbol{\theta}))$ を最小とする点を求めることと等しい [3].

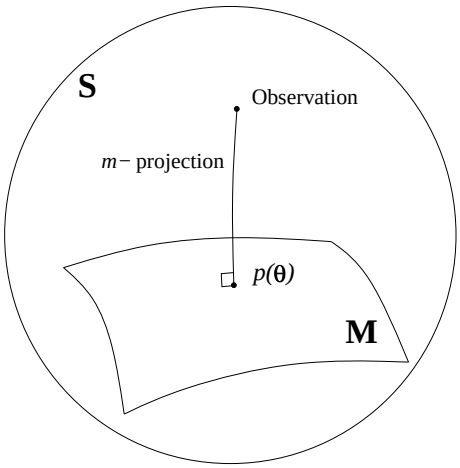

図 3: 統計的推定の幾何学的イメージ

4　隠れ変数を持つモデルとパラメータ推定

4.1　隠れ変数を持つモデル

はじめに隠れ変数について触れたが，ここでは幾つかの具体例を考える．まず，数式での扱い方を示し，例を交えて説明する．また，このようなモデルの最尤推定に関しても説明する．

Y として，観測できる確率変数があったとする．一方，変数 Z は観測できないとする．Y, Z を両方含むモデルのパラメータを $\boldsymbol{\theta}$ と書けば，全ての確率変数を含む確率密度関数は $p(y, z; \boldsymbol{\theta})$ と表わせる．しかし，本当に観測できるのは観測データ $(y_1, y_1, \ldots, y_n)$ のみであり，我々が観測データから得られる密度関数は，

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

という確率変数 Y についての周辺分布に関するものだけである．

例えば，例えば，上のような $\{0, 1\}$ からなる図があり，Fig.4 の 5 つの図のどれかが確率 $1/5$ で

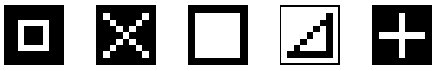

図 4: True images

取られるものとする．しかし，実際に提示される場合にはノイズが乗り，ビットがある確率 p_z で反転するとする． 例えば，のような図が与えられるとする．我々が目にするのはこのようにノイズの

図 5: Sample images

乗った図のみであるとする．

この問題では，

隠れ変数 z ：実際に選択されている図は  のうちのどれかという確率変数．

観測される変数 $\boldsymbol{y}$ ：実際に観測されるノイズに汚れた図 5() のような図．

となる．モデルでのパラメータは $\boldsymbol{\theta} = \{$, $p_z\}$ である．このパラメータを用いて尤度がどのように定義されるかを見よう．図 5 のような図が得られたのならば，

$$p(y; \boldsymbol{\theta}) = \sum_z p(y, z; \boldsymbol{\theta}) = \sum_z p(z; \boldsymbol{\theta}) p(y|z; \boldsymbol{\theta})$$

から，

$$p(\text{}\ \boldsymbol{\theta}) = \frac{1}{5} p(\text{}\ \text{};\ \boldsymbol{\theta}) + \frac{1}{5} p(\text{}\ \text{};\ \boldsymbol{\theta}) + \cdots + \frac{1}{5} p(\text{}\ \text{};\ \boldsymbol{\theta})$$

となる．

このように例え $p(y; \boldsymbol{\theta})$ を知ったとしても，Y のみからでは z を直接は知ることができない．そこで Z は隠れ変数 (Latent Variable) と呼ばれる．統計学で用いられる確率分布のなかには，混合分布や因子分析のモデルのように隠れ変数を持つものが多くある．また，脳のモデルとして提案されているニューラルネットワークにおいても隠れ変数の考え方を用いるものがある．次節以降でいくつかの例を示す．

4.2 混合正規分布

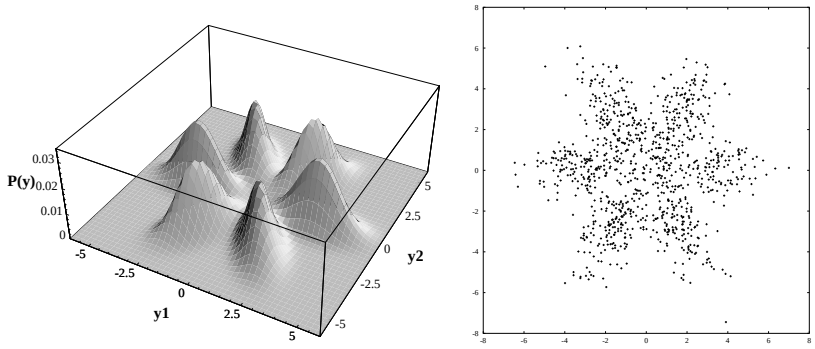

図 6: 正規混合分布

まずは，混合正規分布を考えよう．確率変数 Y の平均が $\boldsymbol{\mu}$ で共分散行列が Σ の多次元正規分布にしたがうとき，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．密度関数 $p(\boldsymbol{y};\boldsymbol{\theta})$ (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，

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

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 6 左に示す．この確率分布からデータが得られているとする．データを 2 次元平面に表示したものを図 6 右に示す．この場合，出力されるデータのみからでは，そのデータがどの正規分布によるものかははっきりとは分らない．すなわち，どの正規分布からのデータ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．この例であれば，z を $1,\cdots,k$ までを取る離散の隠れ変数として，$p(z,\boldsymbol{y};\boldsymbol{\theta})$ は，

$$p(\boldsymbol{y},z;\boldsymbol{\theta}) = \sum_{i=1}^{k} \pi_i \delta_i(z) G(\boldsymbol{y};\boldsymbol{\mu}_i,\Sigma_i),$$

と書ける．ここで，$\delta_i(z)$ は $z=i$ のときにのみ 1 を取る関数である．

4.3 Helmholtz マシン

脳の神経細胞には，神経細胞の集合から成る複数のモジュールがあり，それらが相互に結合している．このモジュール間の結合を用い，学習を行なうモデルとして Helmholtz マシンが提案された．Helmholtz マシンは生成モデル (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 [4]．図 7 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．これを脳のモジュール間の相互の結合に対応させ，visible variable を持つ細胞を低次の神経細胞，hidden factor を持つ細胞を高次の細胞と呼ぶこともある．

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

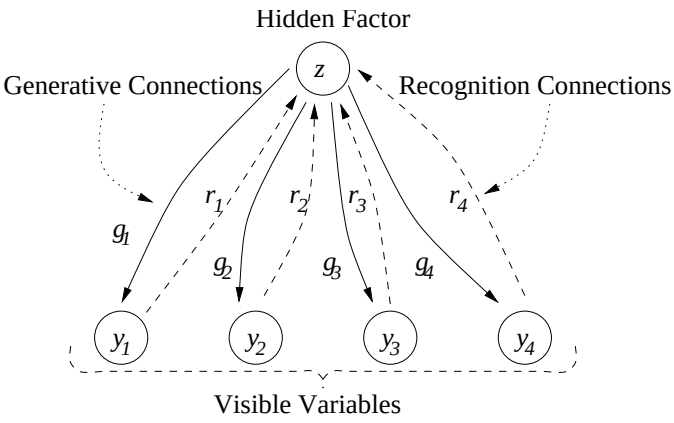

図 7: Helmholtz マシン

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．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 [5].

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

- 生成モデル：$\boldsymbol{y}$ 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 z によって

$$\boldsymbol{y} = \boldsymbol{g}z + \boldsymbol{\varepsilon} \tag{12}$$

  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．このとき，$p(\boldsymbol{y},z;\boldsymbol{g},\Sigma)$ は，

$$p(\boldsymbol{y},z;\boldsymbol{g},\Sigma) = G\left(\begin{pmatrix} z \\ \boldsymbol{y} \end{pmatrix}; \mathbf{o}, \left(\begin{array}{c|c} 1 & \boldsymbol{g}^T \\ \hline \boldsymbol{g} & \Sigma \end{array}\right)\right).$$

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

$$z = \boldsymbol{r}^T\boldsymbol{y} + \delta \tag{13}$$

  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．認識モデルでは $\boldsymbol{y}$ が観測されたときの z の条件付き分布を定義する．$q(z|\boldsymbol{y};\boldsymbol{r},\sigma^2)$ とすると，

$$q(z|\boldsymbol{y};\boldsymbol{r},\sigma^2) = G\left(z;\boldsymbol{r}^T\boldsymbol{y},\sigma^2\right).$$

以上のように，Helmholtz マシンは隠れ変数を持つモデルを表現する一つの手法であり，生成モデルと同時に認識モデルを同じ確率変数の組みに対し定義するところに特徴がある．

4.4 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 $X = \{Y,Z\}$ があり，その一部 Y のみが観測でき，残り Z は観測できない状況を考える．観測データ $\{y_1,y_2,\ldots,y_T\}$

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

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

と定義されるが，この形は必ずしも単純ではなく，(10) 式を直接解くのは難しいことが多い．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．

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}_t) = \frac{1}{T}\sum_{s=1}^{T}\left\{\int p(z|y_s;\boldsymbol{\theta}_t)\log p(y_s,z;\boldsymbol{\theta})dz\right\} \tag{14}$$

- **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．

$$\boldsymbol{\theta}_{t+1} = \operatorname*{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta},\boldsymbol{\theta}_t) \tag{15}$$

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

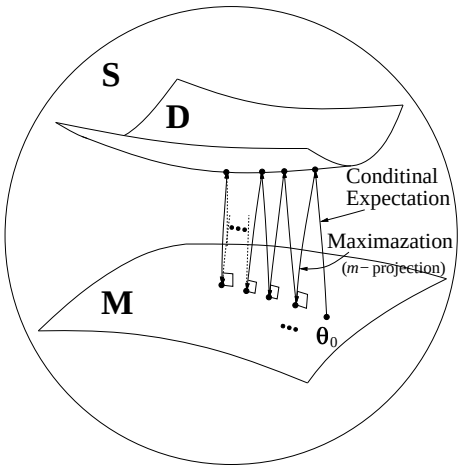

図 8: EM アルゴリズム

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 3 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 y の確率分布の空間ではなく，確率変数 $x=\{y,z\}$ 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 y に関する経験分布 $q(y)$ しか与えない．このままでは $x=\{y,z\}$ の確率分布の空間中の点とはならないので，

z に関する任意の分布を付け加え D (図 8) という多様体を構成する (より詳しくは [3, 6] 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 [6].

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

5 まとめ

本稿ではエントロピーや情報量の考え方から, 隠れ変数, その学習則までを簡単に説明をした. 脳の情報処理を考える際に, 隠れ変数や情報量の考え方が役に立つと思う.

例えば, 高次の情報と低次の情報を考えるときに, 高次の処理では恐らく情報の集約が行なわれているが, そのとき, 低次の神経細胞における情報のなかで, 意味を持つ情報が保存されていることが望ましい. したがって, 低次の神経活動を確率変数を X で表し, 高次の活動を確率変数を Y で表すならば, $I(X,Y)$ は小さい方が望ましいと考えられる. 確率変数, 情報量の考え方は見通しを良くする一つの指針を与えると思う.

参考文献

[1] 甘利 俊一. 情報理論. ダイヤモンド社, 1970.

[2] 有本 卓. 確率 · 情報 · エントロピー. 森北出版, 1980.

[3] 甘利 俊一 and 長岡 浩司. 情報幾何の方法. 岩波講座 応用数学 [対象 12]. 岩波書店, 1993.

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

[5] Shiro Ikeda, Shun-ichi Amari, and Hiroyuki Nakahara. Convergence of the wake-sleep algorithm. In Michael S. Kearns, Sara A. Solla, and David A. Cohn, editors, *Advances in Neural Information Processing Systems*, volume 11, pages 239–245. MIT Press, Cambridge, MA, 1999.

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