連載解説「Deep Learning（深層学習）」〔第1回〕

ディープボルツマンマシン入門
—ボルツマンマシン学習の基礎—

Introduction of Deep Boltzmann Machine
—Fundamentals of Boltzmann Machine Learning—

安田 宗樹
Muneki Yasuda
山形大学大学院理工学研究科
Graduate School of Sciences and Engineering, Yamagata University.
muneki@yz.yamagata-u.ac.jp

Keyword: machine learning, deep learning, restricted Boltzmann machine, deep Boltzmann machine.

1. はじめに

ボルツマンマシン（Boltzmann machine）[Ackley 85, MacKay 03, Wainwright 08, 渡邊 01]は1980年代に提案されたニューラルネットワーク由来の相互結合型機械学習モデルであり，マルコフ確率場（Markov random field）[Bishop 06, Koller 09]と呼ばれる応用上非常に重要なモデルクラスに属する統計的機械学習（statistical machine learning）のモデルである．ボルツマンマシンは一般に複数のピークをもつ分布になり得るため，比較的複雑なデータ構造に対応可能である．

ところがボルツマンマシンは一つの重大な問題を抱えているためにこれまで積極的な利用が避けられてきた．その問題とは計算量の問題である．ボルツマンマシンの学習あるいはボルツマンマシンを利用した確率的推論を実行するためには非現実的な計算時間を要してしまい，その面であまり実用的とは考えられてこなかった．

しかし，ここ十数年でその状況はかなり変化してきた．2000年代初頭にリストリクティッドボルツマンマシン（restricted Boltzmann machine: RBM）[Hinton 02, Smolensky 86]と呼ばれるボルツマンマシンモデルの再発掘とそれに対するコントラスティブダイバージェンス（contrastive divergence: CD）法と呼ばれる効果的な近似学習アルゴリズムの提案[Hinton 02]がなされ，それらが最近話題となっている深層学習（deep learning）の概念へとつながった[Hinton 06]．深層学習のモデルは現在さまざまな応用課題において目覚ましい成果をあげてきている．

本稿ではボルツマンマシンの基礎から出発し，次いでRBMについて解説する．最後に統計的機械学習を基礎とした確率的深層学習モデルの一つであるディープボルツマンマシン（deep Boltzmann machine: DBM）[Salakhutdinov 09, Salakhutdinov 12]に触れる．DBMは深層学習の研究の皮切りとなったディープビリーフネットワーク（deep belief network: DBN）[Hinton 06]の拡張であり，ボルツマンマシンを基礎とした最新の確率的深層学習モデルである．

2. 統計的機械学習理論—データ生成モデルの再現

あるn次元の$\{+1, -1\}$の2値の観測データ点があったとしよう[*1]．そのデータ点を$\mathbf{x}=\{\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_n\}=\{\mathrm{x}_i \mid i=1, 2, \cdots, n\}$と表す．データは通常確実なもの（完全に予想できるもの）ではなく，次にどのようなデータを受け取るかわからないといったような不確実性をもつ．ここでいうパターンとは「x_1が$+1$でd_2が-1で…」というように観測したデータ$\mathbf{x}$の値の一つの集まりのことを指す．

不確実性をもつため，データは確率的に出現するものと考える．そしてデータはあるn次元の確率分布$P_g(\boldsymbol{X})$から生成されたものと考える．ここで$\boldsymbol{X}=\{X_i \in \{+1, -1\} \mid i=1, 2, \cdots, n\}$は$n$次元の確率変数である．つまり確率分布$P_g(\boldsymbol{X})$に従ってサンプルされた一つの実現値が$\mathbf{x}$という解釈である（確率変数$X_i$に対応するサンプル点が$\mathrm{x}_i$であると考えている）．$P_g(\boldsymbol{X})$は対象のデータの確率的な生成規則を規定するものであるから，$P_g(\boldsymbol{X})$をデータの生成モデル（generative model）と呼ばれる．データパターンには出現しやすいものやそうでないものがあるであろうから，データパターンごとに出現確率は異なるであろう．そのデータパターンごとの出現確率を記述するのが生成モデルである．

観測データとして図1に示したような白黒の2値画像の例を考えよう．白が$+1$で黒が-1に対応するとして，各画像中の各画素（ピクセル）の色をデータとすると，データは画像の全画素数の次元となり，一つのデータパターン$\mathbf{x}$がある一つの白黒画像となる．その場合，$P_g(\boldsymbol{X})$

*1 普通2値というと0, 1で表現されることが多いが，本稿では$+1, -1$を用いる．

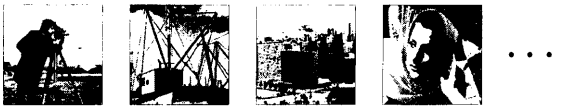


図1　白黒の2値画像の例.
白が+1で黒が−1に対応するとする

は白黒画像の生成確率を表すこととなる（例えば，図1の一番左の画像が出現する確率は○○%など，画像ごとの出現確率を表す）.

データが生成モデルに従い発生しているので，もし生成モデルの詳細がわかれば，そのデータの（確率的）生成メカニズムは完全に掌握できるはずである．しかし残念ながら，（当たり前ではあるが）現実のデータに対する生成モデルは一般には未知の分布である.

データを生成している未知の生成モデルを再現することが統計的機械学習理論の目的であり，受け取った観測データのセットを利用して生成モデルを構築するための枠組みを与えるための理論枠組みである．図2に統計的機械学習理論のスキームを示す．未知の生成モデルからその確率に従ってデータが生成されているとする．我々は生成モデルを知らないので，適当な学習モデル $P(\boldsymbol{X}|\boldsymbol{\Theta})$ を仮定する．そして観測データを利用し，仮定した学習モデル（のパラメータ $\boldsymbol{\Theta}$）を調整することで生成モデルの再現を試みる.

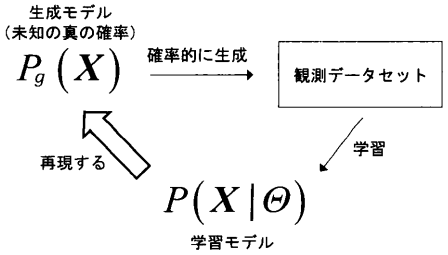


図2　統計的機械学習のスキーム.
$P_g(\boldsymbol{X})$ はデータ生成の確率規則である生成モデルであり，$P(\boldsymbol{X}|\boldsymbol{\Theta})$ がパラメータ $\boldsymbol{\Theta}$ をもった学習モデルである

一度統計的機械学習により観測データを用いて学習モデル $P(\boldsymbol{X}|\boldsymbol{\Theta})$ を学習し，適切な学習モデルを手に入れると，それを利用することにより数多くの有意義な応用が可能となる．したがって統計的機械学習理論による学習モデルからのデータの生成モデルの再現は，その有用性からパターン認識をはじめその他多くの情報科学分野の課題において求められており，現代の情報科学分野での根幹技術の一つとなり始めている．統計的機械学習理論のより詳細や種々の応用については文献 [Bishop 06, Koller 09] などがお勧めである.

以下の章で説明するボルツマンマシンは受け取った観測データから生成モデルを統計的機械学習の方法で構築するための学習モデル[*2]の一つである.

3. ボルツマンマシン

連想記憶のモデルとして知られているホップフィールドモデル（Hopfield model）[MacKay 03] があり，ホップフィールドモデルの決定論的ダイナミクスを確率的ダイナミクスに拡張したモデルとしてボルツマンマシンが導入された [Ackley 85]．ボルツマンマシンは多次元のボルツマン分布（あるいはギブス分布）と呼ばれる確率分布として定義される.

いくつかのノード $\Omega=\{1, 2, \cdots, |\Omega|\}$ といくつかの無向リンク E からなる無向グラフ $G(\Omega, E)$ があるとする．ここでノード i と j の間のリンクを (i, j) で表し，無向リンクなので (i, j) と (j, i) は同一のリンクを示すとする．図3に5ノードの無向グラフの例を示している．i 番目のノードに確率変数 $x_i \in \{+1, -1\}$ を対応させ，グラフ $G(\Omega, E)$ 上に次の確率モデルを定義する.

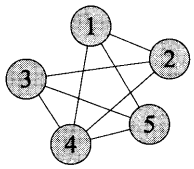


図3　5ノードの無向グラフの例.
黒丸がノードであり，線がリンクを表している．ノード中の数字はそのノードの番号である．$\Omega=\{1,2,3,4,5\}$, $E=\{(1,2),(1,4),(1,5),(2,3),(2,4),(3,4),(3,5),(4,5)\}$ である

$$P_B(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{w}) := \frac{1}{Z_B(\boldsymbol{\theta}, \boldsymbol{w})} \exp(-\Phi(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{w})) \quad (1)$$

ここで，$\Phi(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{w})$ はエネルギー関数（あるいはポテンシャル関数）と呼ばれ

$$\Phi(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{w}) := \underbrace{-\sum_{i\in\Omega}\theta_i X_i}_{\text{バイアス項}} \underbrace{-\sum_{(i,j)\in E} w_{ij} X_i X_j}_{\text{相互作用項}} \quad (2)$$

のようにバイアス項と相互作用項の和により定義される．$Z_B(\boldsymbol{\theta}, \boldsymbol{w})$ は $\boldsymbol{x}$ に関する総和 $\sum_{\boldsymbol{X}} P_B(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{w})$ を1にするための規格化定数（あるいは分配関数）であり

$$Z_B(\boldsymbol{\theta}, \boldsymbol{w}) := \sum_{\boldsymbol{X}} \exp(-\Phi(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{w}))$$

である．ここで式中の和 $\sum_{\boldsymbol{X}}$ は

$$\sum_{\boldsymbol{X}} := \prod_{i\in\Omega}\sum_{X_i=\pm 1} = \sum_{X_1=\pm 1}\sum_{X_2=\pm 1}\sum_{X_3=\pm 1}\cdots\sum_{X_{|\Omega|}=\pm 1}$$

なる確率変数 $\boldsymbol{X}$ の実現可能パターンすべてに関する総和を表している．以降，変数が変わってもこの種の和の意味は同様である．モデルパラメータは $\boldsymbol{\theta}$ と $\boldsymbol{w}$ であり，各 θ_i は各ノード i のバイアスパラメータ，各 w_{ij} は各リンク (i, j) の重み（相互作用）パラメータと呼ばれる．また $w_{ij}=w_{ji}$ であるとする．この確率モデルがボルツマンマシンであり，本稿で中心的な役割を果たすモデルである．このようにグラフ表現と確率モデルを対応させた

*2 ボルツマンマシン学習は統計的機械学習理論の枠組み内ではパラメトリック（parametric）・教師あり（supervised）学習となる.

ものを（**確率的**）**グラフィカルモデル**（(probabilistic) graphical model）と呼ぶ．

3・1 ボルツマンマシンのモデル解釈

さて，式（1）で定義したボルツマンマシンが一体どのような数理モデルなのかを考えてみよう[*3]．式（1）を見てみると，指数関数の引数がエネルギー関数の負となっている．すなわち，**エネルギーが低い$\boldsymbol{X}$のパターンがより高確率で出現する**．次にパラメータの意味について考えてみよう．式（2）の第1項のみに注目した場合，例えば$\theta_i > 0$であったら，$X_i = +1$のほうがエネルギーを低くする（確率が高くなる）．逆に$\theta_i < 0$ならば$X_i = -1$のほうがエネルギーが低い．つまり，θ_iの正負によって対応するX_iの値の出現確率に偏りが出る．そのため$\boldsymbol{\theta}$がバイアスと呼ばれる．今度は式（2）の第2項のみに注目しよう．$w_{ij} > 0$ならば，X_iとX_jは同じ値を取ったほうがエネルギーが低く，逆に$w_{ij} < 0$ならば，X_iとX_jは互いに異なる値を取ったほうがエネルギーが低い．つまり，$\boldsymbol{w}$はリンクで結ばれた変数同士の値の関連性（そろいやすさ・そろい難さ）を調整するパラメータといえる．

例えば，2章で触れた白黒画像（図1）に対する学習モデルとしてボルツマンマシンを利用することを考える．画像において，ピクセルは碁盤の目上に並べられているので，各確率変数X_iをそれぞれ白（$X_i = +1$）か黒（$X_i = -1$）になる各ピクセルに対応させると，適当なグラフ構造として図4（a）のような画像サイズと同じ正方格子型が一つの選択肢となるだろう．最も近いピクセル同士は強く依存し合うであろうという仮定のもと，リンクは最近接のピクセル間のみに存在するとする．この場合，バイアスパラメータはそのピクセルの色の偏り（白（または黒）になりやすさ）を表し，重みパラメータは隣同士のピクセルの色のそろいやすさを表すこととなる（図4（b））．それゆえ，画像処理の分野では相互作用項のことを平滑化項と呼ぶこともある．これらのパラメータの値を目的の白黒画像の統計性に合うように設定し，ボルツマンマシンにより目的の白黒画像の生成モデルを再現する．ボルツマンマシンをはじめとしたマルコフ確率場モデルと画像処理の関連に関する詳細は，[田中 06]を参照されたい．

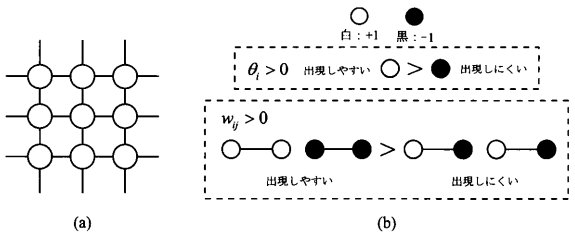


図4 (a) 白黒画像（図1）に対するボルツマンマシンのグラフ構造．(b) バイアス・重みパラメータの意味

以上の説明はあくまでエネルギー関数中のそれぞれの項ごとの個別の解釈であるから，実際はそれらが絡み合ってさらに複雑な確率的な変数間の関係性（これを**確率的相関構造**と本稿では呼ぶ）を生み出す（直接リンクでつながっていない変数同士もリンクを介して関係性をもつ）．この確率的相関構造が各データパターンごとの出現確率の複雑な差異をつくり出すこととなる．

3・2 ボルツマンマシンの学習へ

ここまでで見てきたように，$\boldsymbol{X}$のパターンによって式（2）のエネルギー関数はさまざまな値を取り，それゆえ，式（1）のボルツマンマシンはそのパターンに応じた確率を示す．学習ですることは，実際に観測したデータの出現確率に沿うように，パラメータ$\{\boldsymbol{\theta}, \boldsymbol{w}\}$の値を調整することである．以下の章では観測データからのボルツマンマシンの学習の方法について具体的に見ていく．

ボルツマンマシンの学習には可視変数のみの学習，隠れ変数ありの学習と，大きく分けて二つの種類がある．**可視変数**（visible variable）（あるいは観測変数（observable variable））と呼ばれる変数はデータに対応している変数であり，**隠れ変数**（hidden variable）（あるいは潜在変数（latent variable））はデータに直接対応していない変数のことである．

3章で触れた白黒画像の例では各確率変数は画像の各ピクセルに対応していた．観測データ点として（完全な欠損部位のない）白黒画像が得られるとすると，すべての確率変数に対応したデータが存在することとなる．データに対応する変数が可視変数であるので，この場合は可視変数のみのボルツマンマシンである．対して，カメラの劣化などの原因で画像中のある一部が欠損した画像しか得られない場合，データは一部が欠損しているため，欠損部に対応する確率変数には対応するデータが存在しない．このように対応する観測データが得られない変数が隠れ変数である．

まず4章で最も基本となる可視変数のみの学習について考え，次いで5章で深層学習においても重要となる隠れ変数ありの学習について考えていく．

3章では確率変数として$\boldsymbol{X}$を用いてきたが，変数には可視変数と隠れ変数の2種類あるため，以下では区別するためにノードiに対応する確率変数が可視変数である場合にはX_iの代わりにv_iを用い，隠れ変数である場合にはh_iを用いることとする．ただし，ボルツマンマシンの話に限定されない一般論を述べるときは$\boldsymbol{X}$を用いる．

4. 可視変数のみのボルツマンマシン学習

本章では，可視変数のみの場合のボルツマンマシンの学習の方法を見ていく．すべての確率変数が可視変数で

*3 0, 1の2値でも，あるいは，多値や連続であっても本質的にはここで説明する直感的な解釈が可能である．

あるから X_i の代わりに v_i を用いることに注意されたい．つまりボルツマンマシンは $P_B(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{w}) = P_B(\boldsymbol{v}|\boldsymbol{\theta}, \boldsymbol{w})$ である．

n 次元の観測データ点がそれぞれ統計的に独立に同分布から N 個手に入ったとしよう．μ 番目の観測データ点を $\mathbf{x}^{(\mu)} = \{\mathrm{x}_i^{(\mu)} \in \{+1, -1\} \mid i = 1, 2, \cdots, n\}$ と表す（図5参照）．図1の白黒画像を観測データ点とするならば，各画像一つが各観測データ点 $\mathbf{x}^{(\mu)}$ に対応する．このデータセットを用いてこれから式（1）のボルツマンマシンを学習する．各データ点中の $\mathrm{x}_i^{(\mu)}$ に対してそれぞれ一つの確率変数を対応させるとすると，ボルツマンマシンは観測データ点と同じ n 次元の確率変数 $\boldsymbol{v} = \{v_i \mid i = 1, 2, \cdots, n\}$ をもつ確率モデルとなる．ボルツマンマシンを定義するにはまず土台となるグラフのリンク構造を決定しなくてはならないが，ここでは適当に何らかの構造を決定した下での学習を考える[*4]．

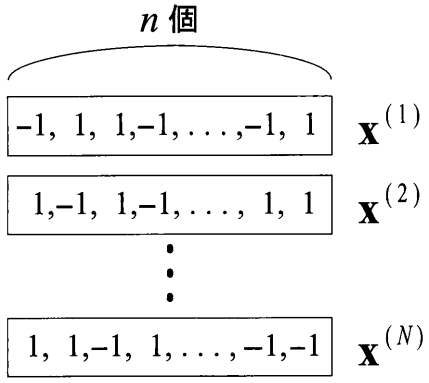


図5　観測データセット $\mathcal{D} = \{\mathbf{x}^{(\mu)} \mid \mu = 1, 2, \cdots, N\}$ の例．四角で括られた行が一点の n 次元観測データ点である．各観測データ点は統計的に独立であるとする

ボルツマンマシンの学習では**最尤法**（maximum likelihood estimation: MLE）を用いる．まず，得られた観測データセット $\mathcal{D} = \{\mathbf{x}^{(\mu)} \mid \mu = 1, 2, \cdots, N\}$ に対して尤度関数

$$\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, \boldsymbol{w}) := \prod_{\mu=1}^{N} P_B(\mathbf{x}^{(\mu)}|\boldsymbol{\theta}, \boldsymbol{w}) \tag{3}$$

を定義する．$P_B(\mathbf{x}^{(\mu)} \mid \boldsymbol{\theta}, \boldsymbol{w})$ はボルツマンマシンが観測データ点 $\mathbf{x}^{(\mu)}$ を生成する確率であり，各観測データ点は独立に発生しているので，それらの積すなわち尤度関数（3）は観測データセット $\mathcal{D}$ をボルツマンマシンが生成する確率であると解釈できる．最尤法とはこの確率（尤度関数）を最大とするパラメータの値（**最尤解**）を求めることであり，得られた観測データセットを生成するに一番尤もらしい分布を求めることと解釈できる．

計算上は尤度関数（3）の自然対数を取った対数尤度関数

$$\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, \boldsymbol{w}) := \ln \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, \boldsymbol{w}) = \sum_{\mu=1}^{N} \ln P_B(\mathbf{x}^{(\mu)}|\boldsymbol{\theta}, \boldsymbol{w}) \tag{4}$$

を最大化したほうが便利な場合が多い．対数関数は単調増加関数なので，最尤解は対数尤度関数（4）を最大化するパラメータの値と一致する．

最大点を求めるために対数尤度関数（4）のパラメータに関する勾配を計算すると，それぞれ

$$\frac{\partial \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, \boldsymbol{w})}{\partial \theta_i} = \sum_{\mu=1}^{N} \mathrm{x}_i^{(\mu)} - N\mathbb{E}_B[v_i|\boldsymbol{\theta}, \boldsymbol{w}] \tag{5}$$

$$\frac{\partial \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, \boldsymbol{w})}{\partial w_{ij}} = \sum_{\mu=1}^{N} \mathrm{x}_i^{(\mu)} \mathrm{x}_j^{(\mu)} - N\mathbb{E}_B[v_i v_j|\boldsymbol{\theta}, \boldsymbol{w}] \tag{6}$$

となる．ここで $\mathbb{E}_B[\cdots|\boldsymbol{\theta}, \boldsymbol{w}]$ はボルツマンマシンに関する期待値を表しており，

$$\begin{aligned}\mathbb{E}_B[\cdots|\boldsymbol{\theta}, \boldsymbol{w}] &= \sum_{\boldsymbol{v}} (\cdots) P_B(\boldsymbol{v}|\boldsymbol{\theta}, \boldsymbol{w}) \\ &= \sum_{v_1=\pm 1} \sum_{v_2=\pm 1} \cdots \sum_{v_n=\pm 1} (\cdots) P_B(\boldsymbol{v}|\boldsymbol{\theta}, \boldsymbol{w})\end{aligned} \tag{7}$$

なる確率変数 $\boldsymbol{v}$ のすべての実現パターンに関する総和を実行することにより得られる．したがって最大点では勾配（5, 6）を0とする条件，すなわち，$\partial \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, \boldsymbol{w})/\partial \theta_i = \partial \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, \boldsymbol{w})/\partial w_{ij} = 0$ から

$$\frac{1}{N} \sum_{\mu=1}^{N} \mathrm{x}_i^{(\mu)} = \mathbb{E}_B[v_i|\boldsymbol{\theta}, \boldsymbol{w}] \tag{8}$$

$$\frac{1}{N} \sum_{\mu=1}^{N} \mathrm{x}_i^{(\mu)} \mathrm{x}_j^{(\mu)} = \mathbb{E}_B[v_i v_j|\boldsymbol{\theta}, \boldsymbol{w}] \tag{9}$$

なる連立方程式が成り立つ．この方程式のことを**ボルツマンマシンの学習方程式**と呼び，最尤解はこの方程式の解である．学習方程式（8），（9）を見ていると，左辺は観測データセットから得られる標本期待値であり，右辺はそれに対応したボルツマンマシンの期待値となっている．つまり，ボルツマンマシンの学習解は観測データセットの一次・二次の標本期待値と学習モデルの一次・二次の期待値を一致させるものと解釈できる．このことから，学習方程式は別名**モーメントマッチング**（moment matching）とも呼ばれ，モーメントマッチングはボルツマンマシンをはじめ多くの確率モデルの学習の中で現れる[*5]．

ボルツマンマシンの学習方程式は一般に数値的に解くこととなる．解き方の方針については付録Aを参照されたい．

4・1　カルバック・ライブラー情報量からの学習方程式の導出

本節では4章で述べた最尤法の別の重要な見方について考える．**カルバック・ライブラー**（**Kullback-Leibler: KL**）情報量は確率変数 $\boldsymbol{X}$ に対する二つの確率分布 $P_0(\boldsymbol{X})$

*4 実際には構造の決定はデータの性質に合わせて適切に行われなければならず，非常に重要な問題である．

*5 指数分布族に属する確率モデルの最尤法による学習は十分統計量同士のモーメントマッチングに帰着される．ボルツマンマシンは指数分布族に属する．

と $P_1(\boldsymbol{X})$ の間の近さの尺度を与える重要な量として知られており，次のように定義される．

$$D(P_0\|P_1):=\sum_{\boldsymbol{X}} P_0(\boldsymbol{X})\ln\frac{P_0(\boldsymbol{X})}{P_1(\boldsymbol{X})} \tag{10}$$

KL情報量は $D(P_0\|P_1)\geqq 0$ であり，$P_0(\boldsymbol{X})=P_1(\boldsymbol{X})$ のときのみ $D(P_0\|P_1)=0$ となる量である．この性質からKL情報量は二つの確率分布 $P_0(\boldsymbol{X})$ と $P_1(\boldsymbol{X})$ の間の距離のようなものとして扱われる*6．

さてここで観測データセット $\mathcal{D}$ に対する経験分布（empirical distribution）を定義しよう．経験分布とは観測データ点の頻度分布のことである．観測データセット $\mathcal{D}$ の経験分布は

$$Q_{\mathcal{D}}(\boldsymbol{v}):=\frac{1}{N}\sum_{\mu=1}^{N}\delta(\boldsymbol{v},\mathbf{x}^{(\mu)}),\ \delta(\boldsymbol{x},\boldsymbol{y})=\begin{cases}1 & \boldsymbol{x}=\boldsymbol{y}\\ 0 & \boldsymbol{x}\neq\boldsymbol{y}\end{cases} \tag{11}$$

のように表すことができる．$Q_{\mathcal{D}}(\boldsymbol{v})\geqq 0$ であり，規格化条件 $\sum_{\boldsymbol{v}}Q_{\mathcal{D}}(\boldsymbol{v})=1$ を満たすので，確率分布になっていることは容易に確認できる．この経験分布と式（1）のボルツマンマシンの間のKL情報量

$$D(Q_{\mathcal{D}}\|P_B)=\sum_{\boldsymbol{v}}Q_{\mathcal{D}}(\boldsymbol{v})\ln\frac{Q_{\mathcal{D}}(\boldsymbol{v})}{P_B(\boldsymbol{v}|\boldsymbol{\theta},\boldsymbol{w})} \tag{12}$$

を考え，$D(Q_{\mathcal{D}}\|P_B)$ を最小とするパラメータの値を求める．微分が0となる条件から，式（12）のKL情報量の最小点では最尤法と同様にボルツマンマシンの学習方程式（8），（9）が成り立つことが示される．したがって，最尤解と式（12）のKL情報量を最小化する解は同じであり，最尤法はKL情報量の観点で見れば，**観測データセットの経験分布とボルツマンマシンを最も近づける方法**であると解釈できる．このことは以下のようにして簡単に確かめられる．

$$\begin{aligned}D(Q_{\mathcal{D}}\|P_B)&=-\sum_{\boldsymbol{v}}Q_{\mathcal{D}}(\boldsymbol{v})\ln P_B(\boldsymbol{v}|\boldsymbol{\theta},\boldsymbol{w})-S_{\mathcal{D}}\\&=-\frac{1}{N}\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta},\boldsymbol{w})-S_{\mathcal{D}}\end{aligned}$$

である．ここで $S_{\mathcal{D}}:=-\sum_{\boldsymbol{x}}Q_{\mathcal{D}}(\boldsymbol{v})\ln Q_{\mathcal{D}}(\boldsymbol{v})$ は観測データセットのエントロピーであり，パラメータ $\{\boldsymbol{\theta},\boldsymbol{w}\}$ に依存しない項である．これより

$$\arg\max_{\boldsymbol{\theta},\boldsymbol{w}}\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta},\boldsymbol{w})=\arg\min_{\boldsymbol{\theta},\boldsymbol{w}}D(Q_{\mathcal{D}}\|P_B)$$

であることがわかる．

4·2 ボルツマンマシン学習の計算困難性

ボルツマンマシンの学習は学習方程式（8），（9）を数値的に解くことで達成される．しかしながら，ボルツマンマシン学習の厳密な実行は実際のところ現実的ではない．学習方程式（8），（9）を見てもわかるとおり，ボルツマンマシンの学習は期待値の計算を要求する（後述の隠れ変数のある場合でも事情は同様である）．式（7）を見てもわかるとおり，この期待値計算は確率変数の全実現可能パターンに関する和であるため，n 次元のボルツマンマシンにおける期待値計算は 2^n 個の多重和の演算となる．

表1 2^n 回の演算に必要な計算時間

n	time
10	約0.00001秒
30	約0.18分
50	約130日
70	約37万4000年
100	約4000億年

仮に1秒間に1億回の演算が可能な計算機があったとして，その計算機を使って 2^n 回の演算するとする．表1に 2^n 回の演算に必要なおおよその計算時間を示している．n の増加に伴い計算時間が爆発的に増加していることがわかる．このような問題は**計算量爆発**などと呼ばれたりする．したがってボルツマンマシンの学習はこの計算量爆発の問題を抱えており，これがこれまでボルツマンマシンがあまり注目されてこなかった背景の一つとなっている．

しかし，近年になってさまざまな有用な近似学習法が開発・適用されはじめ，ボルツマンマシンの学習が近似的にではあるが，現実的な時間内で実行することが可能になってきており，それにより，ボルツマンマシンが最近見直され始めてきた．確率伝搬法（loopy belief propagation）を基礎としたもの [Yasuda 09] や擬似最尤法（maximum pseudo-likelihood estimation: MPLE）[Besag 75, Hyvärinen 06]・複合最尤法（maximum composite likelihood estimation）[Lindsay 88]，スコアマッチング（score matching）[Hyvärinen 05] やCD法 [Hinton 02] などさまざまな近似学習法が現在までに知られている．特に，近似学習法の一つである擬似最尤法は実装の簡便さと比較的高い学習性能をもつことから，最近よく利用されているのを目にする．付録Bに擬似最尤法についての簡単な解説を示す．

ボルツマンマシン上の確率推論を含むボルツマンマシンのより数学的な詳細については [Wainwright 08] が詳しい．

5. 隠れ変数ありのボルツマンマシン学習

本章ではデータには直接対応しない隠れ変数がある場合のボルツマンマシン学習について考える．4章のときと同様に n 次元の観測データ点がそれぞれ独立に同分布から N 個手に入ったとしよう．この観測データセットに対して $(n+m)$ 次元の確率変数をもつボルツマンマシンを用いて学習を行うことが本章の目的である．観測データ点の次元より学習モデルの次元のほうが高いので，すべての変数が各観測データ点に対応するわけでは

*6 距離の公理の満たさないのでKL情報量は厳密には距離とは呼べない．

ない．

$(n+m)$ 次元の変数のうちノード番号（確率変数の添字の番号）の若い順に並べ，最初の n 次元を各観測データ点に対応させ（可視変数），残りの m 次元の変数は観測データとは関係ない変数とする（隠れ変数）．観測データ点に対応するノード番号の集合を $V=\{1,\cdots,n\}$ とし，対応しないノード番号の集合を $H=\{n+1,\cdots,n+m\}$ とする．ノード全体の集合は Ω であるので $\Omega=V+H$ である．図6に例を示す．三次元の観測データセットに対して図3のグラフ上に定義された五次元のボルツマンマシンを用いて学習する場合，$V=\{1,2,3\}$，$H=\{4,5\}$ となり，$\boldsymbol{X}=\{X_1,X_2,X_3,X_4,X_5\}=\{v_1,v_2,v_3,h_4,h_5\}$ である．つまりノード1～3は可視変数として扱われ，ノード4と5は隠れ変数として扱われる．

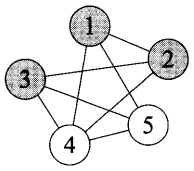


図6　三次元の観測データセットを図3のグラフ上に定義された五次元のボルツマンマシンを用いて学習する場合の例．
黒丸が可視変数で，白丸が隠れ変数である

以下では式（1）の確率変数 $\boldsymbol{X}=\{X_i \mid i\in\Omega\}$ に対するボルツマンマシン $P_B(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{w})$ を可視変数 $\boldsymbol{v}=\{v_i \mid i\in V\}$ と隠れ変数 $\boldsymbol{h}=\{h_i|i\in H\}$ の結合分布として表現する：$P_B(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{w})=P_B(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\theta},\boldsymbol{w})\propto\exp(-\Phi(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\theta},\boldsymbol{w}))$．確率変数の表記法が変化しただけで，定義そのものは変わらないことに注意されたい．

隠れ変数がある場合の学習もやはり最尤法によって行われる．しかし4章のときとは異なり，データに対応していない隠れ変数が含まれているため，少々異なる定式化が必要である．隠れ変数がある場合は，隠れ変数に関して周辺化した可視変数 $\boldsymbol{v}$ のみの n 次元の確率分布

$$P_V(\boldsymbol{v}|\boldsymbol{\theta},\boldsymbol{w}):=\sum_{\boldsymbol{h}}P_B(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\theta},\boldsymbol{w}) \tag{13}$$

を用いる．この周辺分布は可視変数のみの確率分布であるため，すべての変数にデータが対応しており，4章のときと同様な方法で尤度関数をつくることができる．この場合の対数尤度関数は

$$\mathcal{L}_{\mathcal{D}}^{H}(\boldsymbol{\theta},\boldsymbol{w}):=\sum_{\mu=1}^{N}\ln P_V(\mathbf{x}^{(\mu)}|\boldsymbol{\theta},\boldsymbol{w}) \tag{14}$$

で表され，最尤解はこの対数尤度関数を最大化するパラメータの値となる．KL情報量最小化の観点では次のKL情報量を最小化することで学習は達成される．

$$D(Q_{\mathcal{D}}\|P_V)=\sum_{\boldsymbol{v}}Q_{\mathcal{D}}(\boldsymbol{v})\ln\frac{Q_{\mathcal{D}}(\boldsymbol{v})}{P_V(\boldsymbol{v}|\boldsymbol{\theta},\boldsymbol{w})} \tag{15}$$

ここで $Q_{\mathcal{D}}(\boldsymbol{v})$ は式（11）で定義された観測データセットの経験分布である．

式（14）の対数尤度関数の最大化の条件，もしくは，式（15）のKL情報量の最小化の条件より，隠れ変数がある場合のボルツマンマシンの学習方程式は

$$\sum_{\boldsymbol{v},\boldsymbol{h}}z_iP_{H|V}(\boldsymbol{h}|\boldsymbol{v},\boldsymbol{\theta},\boldsymbol{w})Q_{\mathcal{D}}(\boldsymbol{v})=\mathbb{E}_B[z_i|\boldsymbol{\theta},\boldsymbol{w}] \tag{16}$$

$$\sum_{\boldsymbol{v},\boldsymbol{h}}z_iz_jP_{H|V}(\boldsymbol{h}|\boldsymbol{v},\boldsymbol{\theta},\boldsymbol{w})Q_{\mathcal{D}}(\boldsymbol{v})=\mathbb{E}_B[z_iz_j|\boldsymbol{\theta},\boldsymbol{w}] \tag{17}$$

となる．ここで z_i は

$$z_i=\begin{cases}v_i & i\in V\subset\Omega\\ h_i & i\in H\subset\Omega\end{cases}$$

のようにノード i が可視変数ノードか隠れ変数ノードかに応じて変換される変数である．式（16）は θ_i に関する勾配が0となる極値条件から，式（17）は w_{ij} に関する勾配が0となる極値条件からそれぞれ導出される．また，$P_{H|V}(\boldsymbol{h}\,|\,\boldsymbol{v},\boldsymbol{\theta},\boldsymbol{w})$ は可視変数が与えられたもとでの隠れ変数の確率であり，ベイズの公式より

$$P_{H|V}(\boldsymbol{h}|\boldsymbol{v},\boldsymbol{\theta},\boldsymbol{w})=\frac{P_B(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\theta},\boldsymbol{w})}{P_V(\boldsymbol{v}|\boldsymbol{\theta},\boldsymbol{w})}$$

により与えられる．学習方程式（16），（17）中の $\mathbb{E}_B[\cdots\,|\boldsymbol{\theta},\boldsymbol{w}]$ はこれまでと同様ボルツマンマシンの期待値を表す記号であり，この場合は $\mathbb{E}_B[\cdots\,|\boldsymbol{\theta},\boldsymbol{w}]=\sum_{\boldsymbol{v},\boldsymbol{h}}(\cdots)P_B(\boldsymbol{v},\boldsymbol{h}\,|\,\boldsymbol{\theta},\boldsymbol{w})$ である．隠れ変数がある場合のボルツマンマシンの学習は学習方程式（16），（17）を解くことで達成される．

以下では隠れ変数の意味について少し詳しく考えていく．

3·2節で述べた白黒画像の例と同様に，何らかの原因で一部のデータが得られない場合に，その得られないデータに対応する変数を隠れ変数として扱うというのが隠れ変数導入の一つの重要な動機である．

重要なもう一つが以下で述べる学習モデルの表現能力の向上のための導入である．本稿でいう分布の表現能力とはパラメータの値を変化させることにより再現できる確率分布の種類の多さである．学習モデルは人間が勝手に仮定したモデルであるので，仮定したモデルが本当に知りたい未知の生成モデル（データを生成する真の確率分布）$P_g(\boldsymbol{X})$ を含んでいるという保証はない．理想的な観測データセットが得られ，最尤法により生成モデルに近づけることができたとしても，学習モデルの表現能力が低ければ，生成モデルとの間には一般には誤差が生まれる．これは学習モデルを仮定したことにより生まれる本質的な誤差であり，この誤差のことを**モデル誤差**と呼ぶ．図7にモデル誤差のイメージを示す．実線の楕円は

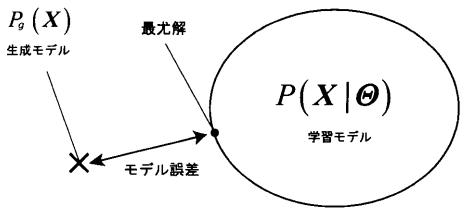


図7　学習モデルの表現能力とモデル誤差のイメージ．
実線の楕円が学習モデルのパラメータ $\boldsymbol{\Theta}$ を変化させることにより再現できる確率分布の範囲を表す

学習モデル $P(\boldsymbol{X}|\boldsymbol{\Theta})$ のパラメータ $\boldsymbol{\Theta}$ を変化させることにより再現できる確率分布の空間を表している．しかし真の生成モデル $P_g(\boldsymbol{X})$ は図中のバツ印の所にある．最尤法により実線の楕円の範囲の中で最も近いところに行き着いたとしても，真の生成モデルとの間には誤差がある．モデル誤差がある場合は，より複雑で表現能力の高い学習モデルを利用して学習しなくてはならない．

式（1）のボルツマンマシンを複雑化する方法としてまず単純に思い付くのが，エネルギー関数の関数形をより複雑なものにするというアプローチであろう．実際，エネルギー関数に三次以上の相互作用の項を加えた高次ボルツマンマシン [Sejnowski 87] なるモデルが提案されており，それは式（1）で提示されている二次の相互作用のみのボルツマンマシンよりも高い表現能力をもつ．

学習モデルを複雑化するもう一つのアプローチが隠れ変数の導入である．こちらはエネルギー関数の関数形を変えることなくモデルを複雑化することが可能となる．隠れ変数がある場合は隠れ変数について周辺化した可視変数のみの分布 $P_V(\boldsymbol{v}|\boldsymbol{\theta},\boldsymbol{w})$ に注目し，最尤法を適用した（式（14）参照）．$P_V(\boldsymbol{v}|\boldsymbol{\theta},\boldsymbol{w})$ は式（13）より

$$P_V(\boldsymbol{v}|\boldsymbol{\theta},\boldsymbol{w}) \propto \exp(-\Phi_V(\boldsymbol{v}|\boldsymbol{\theta},\boldsymbol{w}))$$

$$\Phi_V(\boldsymbol{v}|\boldsymbol{\theta},\boldsymbol{w}) := -\ln\sum_{\boldsymbol{h}}\exp(-\Phi(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\theta},\boldsymbol{w}))$$

と表され，一般に元のボルツマンマシンに比べてより複雑な関数形のエネルギー関数をもつ．つまり，隠れ変数がある場合の学習は式（1）で定義される通常のボルツマンマシンよりもより複雑なエネルギー関数をもった確率モデル $P_V(\boldsymbol{v}|\boldsymbol{\theta},\boldsymbol{w})$ を学習モデルとして利用した可視変数のみの学習であるとみなすこともできる．このことは後述の6·2節 §2にて特別なケースを用いて再び議論する．

6. リストリクティッドボルツマンマシン

リストリクティッドボルツマンマシン（restricted Boltzmann machine: RBM）は**完全2部グラフ**（complete bipartite graph）上に定義された隠れ変数ありのボルツマンマシンである [Smolensky 86]．片方の層は可視変数のみの層（可視層）であり，もう片方は隠れ変数のみの層（隠れ層）である．図8に下層が可視層で上層が隠れ層であるRBMを示す．同層内のリンク結合はなく，異層間のリンク結合のみ存在する．5章の説明に合わせるため，可視変数は n 個，隠れ変数は m 個とする．つまりこのRBMは n 次元の観測データセットを用いて学習するモデルである．

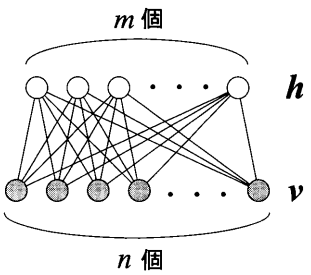


図8 完全2部グラフ上に定義されたRBM. 下層が可視層で上層が隠れ層である

5章のときと同様に，可視変数と隠れ変数の番号の集合をそれぞれ $V=\{1,\cdots,n\}$ と $H=\{n+1,\cdots,n+m\}$ とし，可視変数と隠れ変数をそれぞれ $\boldsymbol{v}$ と $\boldsymbol{h}$ と表すと，RBMのエネルギー関数は

$$\Phi(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{w}) = \Phi(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\theta}^v,\boldsymbol{\theta}^h,\boldsymbol{w}) = -\sum_{i\in V}\theta_i^v v_i - \sum_{j\in H}\theta_j^h h_j - \sum_{i\in V}\sum_{j\in H} w_{ij} v_i h_j \quad (18)$$

のように表される．ここで，可視変数に対するバイアスと隠れ変数に対するバイアスを明確に区別するために $\boldsymbol{\theta}$ を $\boldsymbol{\theta}^v$ と $\boldsymbol{\theta}^h$ に分けている．したがって式（1）に従うと，RBMは次のような確率モデルとして表現することができる．

$$P_B(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\theta}^v,\boldsymbol{\theta}^h,\boldsymbol{w}) \propto \exp\left(\sum_{i\in V}\theta_i^v v_i + \sum_{j\in H}\theta_j^v h_j + \sum_{i\in V}\sum_{j\in H} w_{ij} v_i h_j\right) \quad (19)$$

以下，表記の簡単化のためパラメータをまとめて $\boldsymbol{\Theta}$ で表す：$\boldsymbol{\Theta}=\{\boldsymbol{\theta}^v,\boldsymbol{\theta}^h,\boldsymbol{w}\}$.

6·1 RBMの学習方程式

RBMは5章で見てきた隠れ変数ありのボルツマンマシンの一つの特別なケースなので，学習の解は学習方程式（16, 17）の解に準ずる．

これまでと同様に，N 個の n 次元観測データセット $\mathcal{D}$ を得たとすると，RBMの場合の学習方程式（16）は

$$\frac{1}{N}\sum_{\mu=1}^{N}\mathrm{x}_i^{(\mu)} = \mathbb{E}_B[v_i|\boldsymbol{\Theta}] \quad (i\in V) \quad (20a)$$

$$\frac{1}{N}\sum_{\mu=1}^{N}\tanh\left(\theta_i^h + \sum_{j\in V} w_{ji}\mathrm{x}_j^{(\mu)}\right) = \mathbb{E}_B[h_i|\boldsymbol{\Theta}] \quad (i\in H) \quad (20b)$$

となり，学習方程式（17）は

$$\frac{1}{N}\sum_{\mu=1}^{N}\mathrm{x}_i^{(\mu)}\tanh\left(\theta_j^h + \sum_{k\in V} w_{kj}\mathrm{x}_k^{(\mu)}\right) = \mathbb{E}_B[v_i h_j|\boldsymbol{\Theta}] \quad (i\in V,\ j\in H) \quad (21)$$

となる．ここで $\mathbb{E}_B[\cdots|\boldsymbol{\Theta}]$ は式（19）のRBMに関する期待値を表している．式（16）が式（20a），（20b）の2式になっているのは，それぞれ θ_i^v と θ_i^h に関する極値条件が異なることに由来する．

RBMの学習方程式（20a），（20b），（21）の左辺は観測データの値を用いて簡単に計算できる形になっている*7が，右辺はRBMの期待値の計算となっており，やはり

*7 これも後述する条件付き独立の性質のご利益の一つである．一般の場合（式（16），（17）の左辺）はRBMと異なりそう簡単に計算することができず，一般に計算量爆発の問題を抱える．

簡単には計算できず計算量爆発の問題を抱えている．2部グラフという一見すると単純そうな構造であるにもかかわらず，RBMも一般のボルツマンマシンと同様に何らかの近似的アプローチを必要とする．

6·2 RBMの性質

RBMは2部グラフという特殊なグラフ構造をもつおかげで，通常のボルツマンマシンにはないいくつかの有用な性質をもつ．

§1 条件付き独立の性質

RBMの重要な性質の一つが条件付き独立性である．可視層を固定したもとでの隠れ層の条件付き確率は

$$P_{H|V}(\boldsymbol{h}|\boldsymbol{v},\boldsymbol{\Theta})=\frac{P_B(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta})}{P_V(\boldsymbol{v}|\boldsymbol{\Theta})}=\prod_{j\in H}\frac{\exp(\lambda_j^H h_j)}{2\cosh(\lambda_j^H h_j)}$$
$$\lambda_j^H:=\theta_j^h+\sum_{i\in V}w_{ij}v_i \tag{22}$$

であり，同様に隠れ層を固定したもとでの可視層での条件付き確率は

$$P_{V|H}(\boldsymbol{v}|\boldsymbol{h},\boldsymbol{\Theta})=\prod_{i\in V}\frac{\exp(\lambda_i^V v_i)}{2\cosh(\lambda_i^V v_i)}$$
$$\lambda_i^V:=\theta_i^v+\sum_{j\in H}w_{ij}h_j \tag{23}$$

となる．条件付き確率（22），（23）はともに確率変数同士の積の形となっている．これはすなわち，片方の層の確率変数が何らかの値に固定されると，もう片方の層の確率変数は互いに統計的に独立となるということを示している．この性質はサンプリングにおいて大いに役に立つ．例えば可視変数を観測データの値で固定したとして，そのときの隠れ変数からのサンプリングはこの条件付き独立の性質から容易に行える．逆もまた然りである．これには層内結合がないという点が効いており，RBM特有の性質である．層内結合がある一般の場合は，マルコフ連鎖モンテカルロ法（Markov chain Monte Carlo method: MCMC）などを用いてかなりの手間をかけないとサンプリングできない．

この条件付き独立性を意識し，考案されたのがコントラスティブダイバージェンス（contrastive divergence: CD）法である [Hinton 02]．CD法は観測データセットの経験分布を可視変数の初期分布とし，そこから交互に両層のサンプリングを実行する．そして得られたサンプリング点の標本期待値をRBMの期待値$\mathbb{E}_B[\cdots|\boldsymbol{\Theta}]$の近似値として確率的近似学習を行う方法である．通常のCD法を拡張し，より性能を向上させたpersistent CDと呼ばれる学習アルゴリズムも提案されており，これは通常のCD法や擬似最尤法などを超える学習性能をもつことが知られている [Tieleman 08]．

ボルツマンマシンが現在においてある一定の市民権を得るに至った大きな要因の一つがCD法の成功と普及であるといえよう．また，後述するディープボルツマンマシンの学習（7章参照）においてもこの条件付き独立の性質は重要な役割りを担うこととなる．

§2 RBMの周辺確率

RBMの構造によりもたらされるもう一つの利点は可視変数に関する周辺確率$P_V(\boldsymbol{v}|\boldsymbol{\Theta})$を簡単に計算できることである．具体的には

$$P_V(\boldsymbol{v}|\boldsymbol{\Theta})\propto\exp(-\Phi_V(\boldsymbol{v}|\boldsymbol{\Theta})) \tag{24}$$
$$\Phi_V(\boldsymbol{v}|\boldsymbol{\Theta})=-\sum_{i\in V}\theta_i^v v_i-\sum_{j\in H}\ln 2\cosh\lambda_j^H \tag{25}$$

である．これは式（13）の周辺化の計算を式（19）で表されるRBMに直接適用することにより得られる．この周辺化の計算は隠れ変数間の結合が存在しないから可能となっている．周辺確率$P_V(\boldsymbol{v}|\boldsymbol{\Theta})$が具体的に記述できるので，擬似最尤法などの隠れ変数がない場合に対して考案されている近似学習法をRBMの学習に適用することが可能となり [Marlin 10]，最近，擬似最尤法を拡張した複合最尤法を利用したRBMの近似学習アルゴリズムも提案されている [Yasuda 12a]．

さて，エネルギー関数（25）を$\boldsymbol{w}$についてテイラー展開してみよう．すると

$$\Phi_V(\boldsymbol{v}|\boldsymbol{\Theta})=c\underbrace{-\sum_{i\in V}\theta_i^v v_i-\sum_{j\in H}\tanh\theta_j^h\sum_{i\in V}w_{ij}v_i}_{\text{バイアス項}}$$
$$\underbrace{-\frac{1}{2}\sum_{j\in H}\operatorname{sech}^2\theta_j^h\sum_{i,k\in V}w_{ij}w_{kj}v_iv_k}_{\text{二次相互作用項}}$$
$$\underbrace{-\frac{1}{3}\sum_{j\in H}\operatorname{sech}^2\theta_j^h\tanh\theta_j^h\sum_{i,k,l\in V}w_{ij}w_{kj}w_{lj}v_iv_kv_l}_{\text{三次相互作用項}}$$
$$+O(w^4)$$

となる．ここでcは確率変数$\boldsymbol{v}$に無関係な定数項である．これを見ると，二次以上の高次の相互作用項の効果がエネルギー関数内に入っていることが確認でき，5章で述べたように隠れ変数を導入することでボルツマンマシンのエネルギー関数を変更することなく確かにモデルが複雑化できていることがわかる．

RBMはその構造上，系統的な方法でいくらでも隠れ変数の数を増やすことができる．隠れ変数の数を増やすごとにモデルの複雑さは上がっていき，$|H|=m\to\infty$の極限では任意の確率分布を表現可能であることが示されている [Roux 08]．

7. ディープボルツマンマシン

本章では深層学習の一つのモデルであるディープボルツマンマシン（deep Boltzmann machine: DBM）[Salakhutdinov 09, Salakhutdinov 12] について考えていく．深層学習研究の皮切りとなったディープビリーフネットワーク（deep belief network: DBN）[Hinton 06] を拡張したものがDBMである．

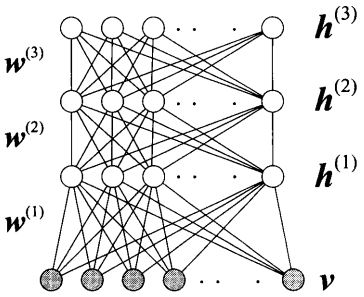


図9 DBM.
最下層が可視層で，隠れ層が3層積まれている例である

DBMは隠れ層を図9のように階層的に積み上げていくことで構成される隠れ素子ありのボルツマンマシンの一種である．

可視層のノードの番号の集合をVで表し，第r番目の隠れ層のノードの集合をH_rとする．また可視変数を$\boldsymbol{v}=\{v_i \in \{+1, -1\} \mid i \in V\}$で表し，第$r$番目の隠れ層内の隠れ変数を$\boldsymbol{h}=\{h_i^{(r)} \in \{+1, -1\} \mid i \in H_r\}$で表すと$R$層の隠れ素子からなるDBMのエネルギー関数は

$$\Phi(\boldsymbol{v}, \boldsymbol{h} | \boldsymbol{W}) = -\sum_{i \in V} \sum_{j \in H_1} w_{ij}^{(1)} v_i h_j^{(1)} - \sum_{r=2}^{R} \sum_{i \in H_{r-1}} \sum_{j \in H_r} w_{ij}^{(r)} h_i^{(r-1)} h_j^{(r)} \quad (26)$$

と表される．ここで$\boldsymbol{h}^{(1)}, \cdots, \boldsymbol{h}^{(R)}$をまとめて$\boldsymbol{h}$で表し，パラメータ$\boldsymbol{w}^{(1)}, \cdots, \boldsymbol{w}^{(R)}$をまとめて$\boldsymbol{W}$で表している．ここで$\boldsymbol{w}^{(r)}$は第$r-1$層と第$r$との間の結合を表している（この場合，可視層は第0層と数える）．DBMの定義においては式(26)バイアス項を無視することが多いが，必要ならばバイアス項を加えてもよい．R=1のときは(バイアスなしの)RBMと等価である．RBMと違いDBMは隠れ変数間の結合をもつため，RBMとは質的に異なる表現能力をもつことが期待される．

DBMの学習

DBMもボルツマンマシンの一種なので，その学習は原理的にはボルツマンマシンの学習方程式(16)，(17)に準ずるが，計算量爆発の問題から，厳密な学習は望めない．そこで近似学習のアプローチが必要となるわけだが，隠れ層が階層的に積まれているため6·2節で説明してきたRBMのときのような性質を利用することができない．

そこでよく利用されるのがRBMを基本とした**貪欲学習**（greedy learning）である．以下ではDBMに対する貪欲学習の概要について見ていく．

図10に図9に示したDBMに対する貪欲学習の概要図を示す．まず可視層と第1層目の隠れ層に注目し，第2層目以上の隠れ層は無視する．すると，可視層と第1層目の隠れ層は$\boldsymbol{v}$と$\boldsymbol{h}^{(1)}$からなるRBMとしてみなせる(図10(a))．6·1節で見てきたRBMの学習法に則って$\boldsymbol{v}$と$\boldsymbol{h}^{(1)}$との間の結合$\boldsymbol{w}^{(1)}$を学習する．

次に$\boldsymbol{h}^{(1)}$と$\boldsymbol{h}^{(2)}$に注目し，その他の層は無視して$\boldsymbol{h}^{(1)}$と$\boldsymbol{h}^{(2)}$を再びRBMとみなす（図10(b)）．その際，$\boldsymbol{h}^{(1)}$を擬似的な可視層とし$\boldsymbol{h}^{(2)}$を隠れ層と考える．このRBMの学習には条件付き確率$P_{H|V}(\boldsymbol{h}^{(1)} | \boldsymbol{v}, \boldsymbol{w}^{(1)})$（式(22)参照）を利用して観測データセットからサンプリング*8したサンプル点をデータとして利用する（このデータを実際の観測データと区別するために特徴点(feature)と呼ぶことがある）．6·2節で説明した条件付き独立の性質よりこのサンプリングは容易である．このRBM学習により結合$\boldsymbol{w}^{(2)}$が学習される．

同様な方法で，最後は$\boldsymbol{h}^{(2)}$と$\boldsymbol{h}^{(3)}$を$\boldsymbol{h}^{(2)}$を擬似的な可視層，$\boldsymbol{h}^{(3)}$を隠れ層としたRBMとみなし，観測データセットから$\boldsymbol{h}^{(1)}$の特徴点をサンプルしたときと同様な流れで$\boldsymbol{h}^{(1)}$の特徴点から条件付き確率

$$P_{H_r|H_{r-1}}(\boldsymbol{h}^{(r)} | \boldsymbol{h}^{(r-1)}, \boldsymbol{w}^{(r)}) = \prod_{j \in H_r} \frac{\exp(\lambda_j^{H_{r-1}} h_j^{(r)})}{2\cosh(\lambda_j^{H_{r-1}} h_j^{(r)})}$$
$$\lambda_j^{H_{r-1}} := \sum_{k \in H_{r-1}} w_{jk}^{(r)} h_j^{(r-1)} \quad (27)$$

に従いサンプルした$\boldsymbol{h}^{(2)}$のサンプル点をデータとし，RBMの学習により結合$\boldsymbol{w}^{(3)}$を学習する．もちろん条件付き確率(27)に従うサンプルも条件付き独立の性質より容易である．さらに層が増えても基本的には同様で，層ごとにRBMとみなし，RBMの学習により結合を順次学習していくという手続きを繰り返すこととなる．DBMのみならずその前身となったDBNもほぼ同様な手続きで学習される．

本章で解説した貪欲学習は本来のボルツマンマシン学習の目的である尤度最大化とは異なるもので，あくまで近似的アプローチであり数理的な根拠もまだまだ十分でない方法であるが，経験的に比較的良い学習解を与えることが知られており，DBMやDBNの学習の基本戦略として現在用いられている．DBMの学習において本節の貪欲学習は事前学習（pre-training）と呼ばれ，パラメータの適切な初期値決定に用いられることがある．その場合はその初期値をもとに，より正確で計算コストの高い学習アルゴリズム（例えばDBM全体に関するMCMCを利用した学習法など）でパラメータの値を調整することになるが，貪欲学習による事前学習の性能がやはりキーとなるようである．

以上がDBMの貪欲学習（または事前学習）の概略である．本章の説明は理解の助けになるよう本質的な部分のみを抽出した解説であり，実際に提案されているDBMの学習法 [Salakhutdinov 09, Salakhutdinov 12] は実は若干ではあるが異なる．より正確な貪欲学習についての補足は付録Cで述べる．

RBMやDBNなどの深層学習の最近の動向について

*8 可視変数の値を観測データ点に固定し，最初のRBM学習により学習された$\boldsymbol{w}^{(1)}$を用いた条件付き確率$P_{H_1|V}(\boldsymbol{h}^{(1)} | \boldsymbol{v}, \boldsymbol{w}^{(1)})$に従い$\boldsymbol{h}^{(1)}$の値をサンプルする．

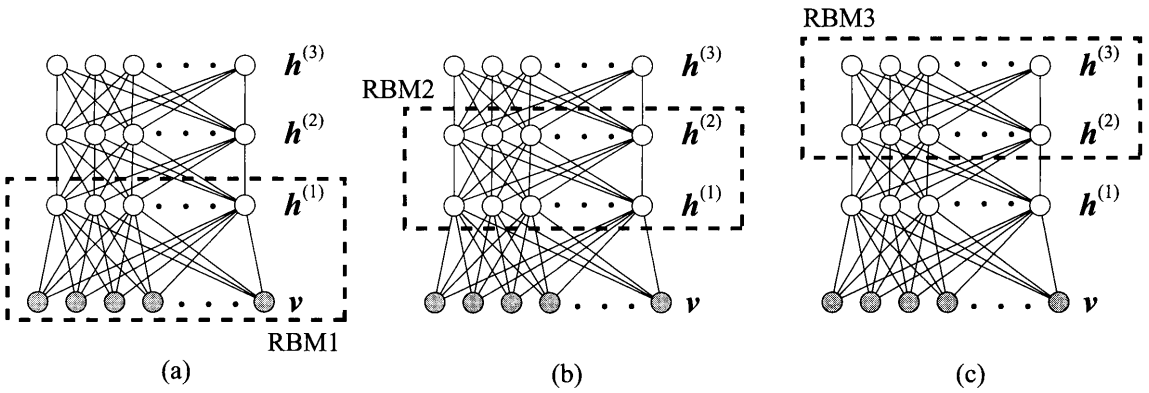


図 10 図 9 に示した DBM に対する貪欲学習の概要図.
各 2 層を独立の RBM とみなし，それぞれについて RBM 学習を適用する.
(a) 1 段階目の RBM 学習. (b) 2 段階目の RBM 学習. (c) 3 段階目の RBM 学習

は [Bengio 09] が詳しい.

8. おわりに

本稿ではボルツマンマシンの基礎から出発し，深層学習の基礎モデルとなる RBM を経て，深層学習モデルの一つである DBM の話題に触れてきた.

深層学習の研究はまだ始まったばかりであり，解決していかなければならない問題はいまだ数多く存在する. 以下では深層学習研究の今後に対する身勝手な私見を述べさせていただく. まず，はじめに考えなければならないのは学習アルゴリズムである. 7 章で見てきたように，DBM は層ごとに RBM とみなし独立に学習していく. この学習方法の理論的根拠は少しずつ説明され始めてはきているが，やはり元々の枠組みである尤度関数最大化を直接目指すものではないため，学習結果がどのようなものになっているかよくわからないということになりかねない. もちろん，計算量の問題により厳密な学習アルゴリズムを実行することは不可能なので，DBM に適したより良い近似学習アルゴリズムを創出していかなければならないであろう.

また同時にモデルの構造の意味についてもより考察を深めていく必要があるであろう. DBM は隠れ変数を階層的に積み上げていくことで構築されている. その意味として，単に隠れ変数の数を増やしてモデルの表現能力を向上させるためのものだと考えるのであれば，DBM は必要なく RBM で十分であるという結論に至ってしまうかもしれない（6·2 節で見てきたように，隠れ変数の数を増やすことで RBM の表現能力はいくらでも増やすことが可能である）.

そうではなく，深層学習モデルの特長は実はその階層性にあると考えている. 潜在的に階層構造をもつ（またはもつと思われる）ような問題は少なくないであろう（例えばコンピュータビジョンの分野で言えば，低レベルのビジョンから高レベルのビジョンへと至るまで間にはある種の階層的なつながりがあると考えられている）. 深層学習モデルのもつ階層構造が問題の潜在的階層性に合致し，問題に内在する本質的な構造の抽出に成功しているからこそ今日の目覚ましい成果に一つとしてつながっているのではないだろうか. よって，深層学習モデルは潜在的な階層性をもつような問題にこそ適用されるべきで，隠れ層もその階層性を強く意識して組み立てられていかなければならないだろうと考えている.

◇ 参考文献 ◇

[Ackley 85] Ackley, D. H., Hinton, G. E. and Sejnowski, T. J.: A learning algorithm for Boltzmann machines, *Cognitive Science*, Vol. 9, pp. 147-169 (1985)
[Bengio 09] Bengio, Y.: Learning deep architectures for AI, *Foundations and Trends in Machine Learning*, Vol. 2, No. 1, pp. 1-127 (2009)
[Besag 75] Besag, J.: Statistical analysis of non-lattice data, *J. Royal Statistical Society D (The Statistician)*, Vol. 24, No. 3, pp. 179-195 (1975)
[Bishop 06] Bishop, C. M.: *Pattern Recognition and Machine Learning*, Springer (2006)
[Hinton 02] Hinton, G. E.: Training products of experts by minimizing contrastive divergence, *Neural Computation*, Vol. 8, No. 14, pp. 1771-1800 (2002)
[Hinton 06] Hinton, G. E., Osindero, S., and Teh, Y. W.: A fast learning algorithm for deep belief net, *Neural Computation*, Vol. 18, No. 7, pp. 1527-1554 (2006)
[Hyvärinen 05] Hyvärinen, A.: Estimation of non-normalized statistical models using score matching, *J. Machine Learning Research*, Vol. 6, pp. 695-709 (2005)
[Hyvärinen 06] Hyvärinen, A.: Consistency of pseudo likelihood estimation of fully visible Boltzmann machines, *Neural Computation*, Vol. 18, No. 10, pp. 2283-2292 (2006)
[Koller 09] Koller, D. and Friedman, N.: *Probabilistic Graphical Models: Principles and Techniques*, MIT Press (2009)
[Lindsay 88] Lindsay, B. G.: Composite likelihood methods, *Contemporary Mathematics*, Vol. 80, No. 1, pp. 221-239 (1988)
[MacKay 03] MacKay, D. J.: *Information Theory, Inference, and Learning Algorithm*, Cambridge University Press (2003)
[Marlin 10] Marlin, B., Swersky, K., Chen, B., and Freitas, de N.: Inductive principles for restricted Boltzmann machine learning, *Proc. 13th Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, Vol. 9, pp. 509-516 (2010)
[Roux 08] Roux, N. L. and Bengio, Y.: Representational power of restricted Boltzmann machines and deep belief networks, *Neural Computation*, Vol. 20, pp. 1631-1649 (2008)

[Salakhutdinov 09] Salakhutdinov, R. and Hinton, G. E.: Deep Boltzmann machines, *Proc. 12th Int. Conf. on Artificial Intelligence and Statistics (AISTATS 2009)*, pp. 448-455 (2009)
[Salakhutdinov 12] Salakhutdinov, R. and Hinton, G. E.: An efficient learning procedure for deep boltzmann machines, *Neural Computation*, Vol. 24, No. 8, pp. 1967-2006 (2012)
[Sejnowski 87] Sejnowski, T. J.: Higher-order Boltzmann machines, *AIP Conference Proceedings 151, Neural Netwowks for Computing*, pp. 398-403 (1987)
[Smolensky 86] Smolensky, P.: *Information Processing in Dynamical Systems: Foundations of Harmony Theory*, Vol. 1. Parallel distributed processing: explorations in the microstructure of cognition (1986)
[田中 06] 田中和之：確率モデルによる画像処理技術入門，森北出版 (2006)
[Tieleman 08] Tieleman, T.: Training restricted Boltzmann machines using approximations to the likelihood gradient, *Proc. 25th Int. Conf. on Machine Learning (ICML)*, pp. 1064-1071 (2008)
[Wainwright 08] Wainwright, M. J. and Jordan, M. I.: Graphical models, exponential families, and variational inference, *Foundations and Trends in Machine Learning*, Vol. 1, No. 1.2, pp. 1-305 (2008)
[渡邊 01] 渡邊澄夫：データ学習アルゴリズム，共立出版 (2001)
[Yasuda 09] Yasuda, M. and Tanaka, K.: Approximate learning algorithm in Boltzmann machines, *Neural Computation*, Vol. 21, No. 11, pp. 3130-3178 (2009)
[Yasuda 12a] Yasuda, M., Kataoka, S., Waizumi, Y. and Tanaka, K.: Composite likelihood estimation for restricted Boltzmann machines, *Proc. 21st Int. Conf. on Pattern Recognition (ICPR2012)*, pp. 2234-2237 (2012)
[Yasuda 12b] Yasuda, M., Tannai, J. and Tanaka, K.: Learning algorithm for Boltzmann machines using max-product algorithm and pseudo-likelihood, *Interdisciplinary Information Sciences*, Vol. 18, No. 1, pp. 55-63 (2012)

2013年3月7日 受理

◇ 付　　録 ◇

A. ボルツマンマシンの学習方程式の解き方

ボルツマンマシンの学習方程式（可視変数のみの場合は式(8)，(9)，隠れ変数がある一般の場合は式（16），（17），RBMの場合は式（20a），（20b），（21））の解がボルツマンマシン学習の解となる．したがって，学習解を得るためにはボルツマンマシンの学習方程式を数値的に解かなければならない．

ボルツマンマシンの学習方程式は対数尤度最大化の最尤法の枠組みから導出されているため，実際には対数尤度関数のパラメータに関する勾配（例えば式（5），（6））を計算し，その勾配を用いて勾配上昇法により方程式を解くこととなる．

可視変数のみの場合の対数尤度関数（4）は実はパラメータに関して凹関数となっているので，勾配上昇法を用いることで対数尤度関数を最大とするパラメータ（すなわち学習解）が（計算量爆発の問題に目をつぶれば）原理的には得られるので話は単純である．

しかしながら，隠れ変数が加わった場合の対数尤度関数（14）はパラメータに関して一般に凹関数となっていないため，多数の極大点をもち得てしまう．したがって，勾配法を用いても大域的最大点に到達せずに，どこかの局所的極大点に捕まってしまうということが起こる恐れがある．これが隠れ変数を含むボルツマンマシン学習の難しい点の一つとなっており，この問題はもちろんRBMやDBMの学習の中にも存在する．

B. 擬似最尤法

ここでは，MLEの統計的近似解法として現在広く用いられている擬似最尤法（maximum pseudo-likelihood estimation: MPLE）[Besag 75]について説明する．

n 次元の確率変数 $\boldsymbol{X}=\{X_i \mid i=1, 2, \cdots, n\}$ に対するある確率モデル $P(\boldsymbol{X}\mid\boldsymbol{\Theta})$ があるとする．ここで $\boldsymbol{\Theta}$ はモデルパラメータである．また，確率変数と同次元の n 次元観測データ点を N 個得ているとする：$\mathcal{D}=\{\mathbf{x}^{(\mu)} \mid \mu=1,2,\cdots,N\}$．この観測データセットの経験分布は式（11）のとおりである．この確率モデルと観測データセット$\mathcal{D}$より，対数尤度関数は

$$\mathcal{L}_{\mathrm{ML}}(\boldsymbol{\Theta})=\sum_{\mu=1}^{N}\ln P(\mathbf{x}^{(\mu)}|\boldsymbol{\Theta}) \tag{B.1}$$

と定義される．この対数尤度関数を $\boldsymbol{\Theta}$ に関して最大化することが最尤推定である．

MPLEでは対数尤度関数（B.1）を次の擬似尤度（pseudo-likelihood）で近似する．

$$\mathcal{L}_{\mathrm{PL}}(\boldsymbol{\Theta})=\sum_{i=1}^{n}\sum_{\mu=1}^{N}\ln P(\mathrm{x}_i^{(\mu)}|\mathbf{x}_{-i}^{(\mu)},\boldsymbol{\Theta}) \tag{B.2}$$

ここで $\mathbf{x}_{-i}^{(\mu)}$ は $\mathbf{x}^{(\mu)}$ から $\mathrm{x}_i^{(\mu)}$ を除いたものである．また右辺の条件付き確率は一つの確率変数のみに注目した1変数の条件付き確率であり，ベイズの公式から

$$P(X_i|\boldsymbol{X}_{-i},\boldsymbol{\Theta})=\frac{P(\boldsymbol{X}|\boldsymbol{\Theta})}{\sum_{X_i}P(\boldsymbol{X}|\boldsymbol{\Theta})}$$

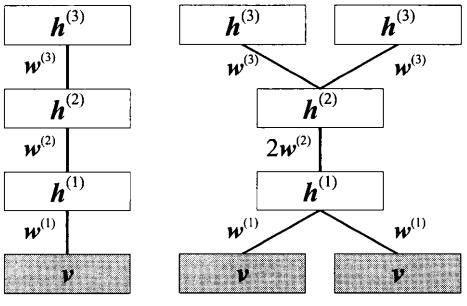
により計算される．MPLEは真の対数尤度関数の代わりに擬似尤度（B.2）を最大化する．MPLEでは規格化定数の計算が必要ないため，（エネルギー関数の計算が可能であれば）通常のMLEと違い計算量爆発の問題を回避できる．MPLEにより得られる学習解は一般に近似解であるが，その解は次の**漸近的一致性**（asymptotic consistency）の性質をもつ：モデル誤差がなく，観測データの数が極めて多い場合 $N\to\infty$，MPLEにより得られる学習解は真の対数尤度関数により得られる最尤解と一致する．

MPLEは実装の簡便さと比較的高い学習性能から広く用いられているが，枠組みとしては隠れ変数のない場合に対応しており，隠れ変数がある場合の学習への適用は工夫なしには難しい．RBMはMPLEが簡単に適用できる例外的な一例である[Marlin 10]．隠れ変数があり，さらに隠れ変数間に結合があるモデルへのMPLEの適用については[Yasuda 12b]にて一つの方法が提案されている．

C. DBMに対する貪欲学習の補足

DBMの貪欲学習の基礎的なことは7章で説明したとおりであるが，[Salakhutdinov 09]で実際に提案されている方法ではもともとのDBMを若干拡張した新しいDBMをつくり，それに対して7章で述べたような逐次的学習を行うことになる．図C.1（a）は図9に示したDBMを簡略化した図である．まずは図C.1（a）を図C.1（b）のようなDBMに拡張する．最下層である可視層と最上の隠れ層をコピーして，さらに中間の層間結合を2倍にしている．図C.1（b）で示されたDBMに対して図C.2に示したように7章で述べた貪欲学習の方法を適用することで図C.1（a）のもともとのDBMの学習は達成される．

図C.1（b）の拡張の（全く形式的ではない）直感的な解釈は以下のようになる[Salakhutdinov 09]．図C.1（a）で表されるもともとのDBMにおいて，ほかのすべての確率変数の値を固定したもとでの v_i，$h_j^{(1)}$，$h_j^{(2)}$，$h_j^{(3)}$ の条件付き確率はそれぞれ



図C.1 （a）図9に示したDBMの簡略化表現．長方形は層を表し，長方形同士をつなぐ線は層間の完全結合を表している．
（b）貪欲学習のために（a）を拡張したDBM．最下層である可視層と最上の隠れ層をコピーして，中間の層間結合を2倍にしている

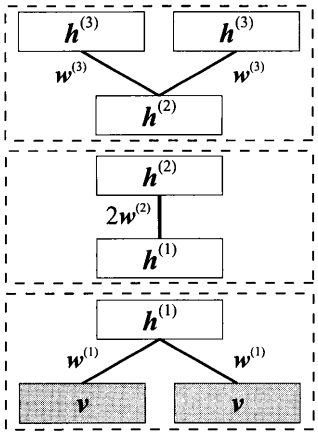


図 C.2 図 C.1（b）の DBM に対する貪欲学習．
最下層からそれぞれ独立の RBM とみなし
下から逐次的に学習していく

$$P(v_i \mid \boldsymbol{h}^{(1)}, \boldsymbol{w}^{(1)}) \propto \exp(\omega_i^V v_i) \tag{C.3}$$
$$P(h_j^{(1)} \mid \boldsymbol{v}, \boldsymbol{h}^{(2)}, \boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}) \propto \exp(\omega_j^{H_1} h_j^{(1)}) \tag{C.4}$$
$$P(h_j^{(2)} \mid \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(3)}, \boldsymbol{w}^{(2)}, \boldsymbol{w}^{(3)}) \propto \exp(\omega_j^{H_2} h_j^{(2)}) \tag{C.5}$$
$$P(h_j^{(3)} \mid \boldsymbol{h}^{(2)}, \boldsymbol{w}^{(3)}) \propto \exp(\omega_j^{H_3} h_j^{(3)}) \tag{C.6}$$

となる．ここで

$$\omega_i^V := \sum_{j \in H_1} w_{ij}^{(1)} h_j^{(1)}$$
$$\omega_j^{H_1} := \sum_{i \in V} w_{ij}^{(1)} v_i + \sum_{k \in H_2} w_{jk}^{(2)} h_k^{(2)}$$
$$\omega_j^{H_2} := \sum_{k \in H_1} w_{kj}^{(2)} h_k^{(1)} + \sum_{k \in H_3} w_{jk}^{(3)} h_k^{(3)}$$
$$\omega_j^{H_3} := \sum_{k \in H_2} w_{kj}^{(3)} h_k^{(2)}$$

と定義されている．

次に拡張された DBM に対する学習について考える．まずは図 C.2 における H_1 の層に注目する．RBM1（すなわち，貪欲学習における最初の RBM）の中では $h_j^{(1)}$ の条件付き確率は

$$P_{\mathrm{RBM1}}(h_j^{(1)} \mid \boldsymbol{v}, \boldsymbol{w}^{(1)}) \propto \exp\left\{\left(2\sum_{i \in V} w_{ij}^{(1)} v_i\right) h_j^{(1)}\right\}$$

であり，一方 RBM2 の中では

$$P_{\mathrm{RBM2}}(h_j^{(1)} \mid \boldsymbol{h}^{(2)}, \boldsymbol{w}^{(2)}) \propto \exp\left\{\left(2\sum_{k \in H_2} w_{jk}^{(2)} h_k^{(2)}\right) h_j^{(1)}\right\}$$

となる．つまり，DBM を層ごとの RBM に分割したことで，H_1 の層の変数に対して 2 通りの異なる近似的見方をすることとなる．RBM1 と RBM2 を統合して一つの DBM とした際に，この二つの異なる条件付き確率の近似の幾何平均で H_1 の層の変数に対する条件付き確率が記述できるとしたら，

$$\begin{aligned} &P_{\mathrm{RBM1+2}}(h_j^{(1)} \mid \boldsymbol{v}, \boldsymbol{h}^{(2)}, \boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}) \\ &\propto \sqrt{P_{\mathrm{RBM1}}(h_j^{(1)} \mid \boldsymbol{v}, \boldsymbol{w}^{(1)}) P_{\mathrm{RBM1}}(h_j^{(1)} \mid \boldsymbol{h}^{(2)}, \boldsymbol{w}^{(2)})} \\ &\propto \exp(\omega_j^{H_1} h_j^{(1)}) \end{aligned}$$

となり，式（C.4）に一致する．H_2 についても事情は同じで，RBM2 と RBM3 のそれぞれにおける条件付き確率は

$$P_{\mathrm{RBM2}}(h_j^{(2)} \mid \boldsymbol{h}^{(1)}, \boldsymbol{w}^{(2)}) \propto \exp\left\{\left(2\sum_{k \in H_1} w_{kj}^{(2)} h_k^{(1)}\right) h_j^{(2)}\right\}$$
$$P_{\mathrm{RBM3}}(h_j^{(2)} \mid \boldsymbol{h}^{(3)}, \boldsymbol{w}^{(3)}) \propto \exp\left\{\left(2\sum_{k \in H_1} w_{jk}^{(3)} h_k^{(3)}\right) h_j^{(2)}\right\}$$

であるから，両者の幾何平均は式（C.5）に一致する．なお，最下層 V と最上層 H_3 に関してはそれぞれ（C.3）と（C.6）にもともと一致している．以上，要するに，図 C.1（b）の拡張は分割した異なる RBM 内での条件付き確率の異なる近似表現が幾何平均により統合された場合，もともとの DBM の条件付き確率に一致するようにするための拡張であると解釈できる．

この拡張の数理的意味付けについては [Salakhutdinov 12] で少し詳しく述べられている．

著 者 紹 介

安田 宗樹

2003 年東北大学工学部通信工学科卒業，2008 年同大学院情報科学研究科応用情報科学専攻博士課程修了．同年，東北大学大学院情報科学研究科応用情報科学にて助教として着任．2013 年山形大学大学院理工学研究科に移り，現在に至る（准教授）．博士（情報科学）．確率的情報処理，統計的機械学習理論，情報統計力学の研究に従事．電子情報通信学会，日本物理学会各会員．