社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

補助関数法に基づく制約付きボルツマンマシンの学習アルゴリズム

高宗　典玄†　　亀岡　弘和†,††

† 東京大学大学院情報理工学系研究科
〒 113–8656 東京都文京区本郷 7–3–1
†† 日本電信電話株式会社ＮＴＴコミュニケーション科学基礎研究所
〒 243–0198 神奈川県厚木市森の里若宮 3–1
E-mail: †{takamune,kameoka}@hil.t.u-tokyo.ac.jp

あらまし　深層学習の重要な一要素として，レイヤーワイズの pre-training がある．レイヤーワイズの pre-training の一つとして制約つき Boltzmann マシン (RBM) が有名である．RBM には Bernoulli-Bernoulli 型と Gaussian-Bernoulli 型があり，従来法の学習アルゴリズムとして Contrastive Divergence 法が有名である．本発表では Bernoulli-Bernoulli 型，Gaussian-Bernoulli 型の両方の最尤学習アルゴリズムと，最適化規準として新たに最大再構築確率を導入し，その学習アルゴリズムに焦点を当て，経験的に高速で安定に収束する補助関数法による新たな更新アルゴリズムの導出を行う．そして，人工データによる収束性能の比較実験を行い，その挙動に対して議論する．
キーワード　深層学習，制約付き Boltzmann マシン，補助関数法，最大再構築確率学習

Training Algorithm for Restricted Boltzmann Machines Using Auxiliary Function Approach

Norihiro TAKAMUNE† and Hirokazu KAMEOKA†,††

† Graduate School of Information Science and Technology, The University of Tokyo
Hongo 7–3–1, Bunkyo-ku, Tokyo, 113–8656 Japan
†† NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation
Morinosato Wakamiya 3–1, Atsugi, Kanagawa, 243–0198, Japan
E-mail: †{takamune,kameoka}@hil.t.u-tokyo.ac.jp

Abstract　Layerwise pre-training is one of important elements for deep learning, and Restricted Boltzmann Machines (RBMs) is popular layerwise pre-training method. At present, the most popular training algorithm for RBMs is the Contrastive Divergence (CD) learning algorithm. We propose deriving a new training algorithm based on an auxiliary function approach for RBMs using the likelihood and the reconstruction probability of observations as the optimization criterion. Through an experiment on parameter training of few RBMs, we confirmed that the present algorithm outperformed the CD algorithm in terms of the convergence speed and the reconstruction error when used as an autoencoder.
Key words　Deep Learning, Restricted Boltzmann Machine, Auxiliary Function Approach, Maximum Recosntruction Probability Training

1. はじめに

近年，深層学習 (Deep learning) の有効性は画像認識や音声認識をはじめ，様々な分野で示されている．深層学習において，大量のデータの下で学習をいかに効率的に行えるかは重要な課題の一つである．その中で，Hinton らによってレイヤーワイズの pre-training が最終的な Deep Neural Network(DNN) を学習するときのパラメータの初期値として高い効果が得られることが分かってきた [1], [2].

レイヤーワイズの pre-training として制約付き Boltzmann マシン (Restricted Boltzmann Machine; RBM) [3] が有名であり，DNN の最下層である実数の観測データを 0 または 1 のバイナリの特徴量に変換する層の学習モデルとして Gaussian-Bernoulli 型 RBM，DNN の中間層としてバイナリからバイナリに変換

する層の学習モデルとして Bernoulli-Bernoulli 型 RBM が存在する．これらを学習するアルゴリズムは Contrastive Divergence(CD) 法 [1], [4] が非常に有名であるが，pre-training にかかる時間と最終的に DNN を学習する時間は pre-training のほうが大きくなる場合があり，pre-training をいかに効率的に行えるかが DNN の全体の学習にかかる計算時間に直結する．

我々の研究室では，これまで様々な音響信号処理問題における最適化問題に対し，補助関数法と呼ぶ原理に基づく最適化アルゴリズムを導出し，その効果を示してきた（例えば [5]）．RBM の学習においても補助関数法に基づく学習則を導出することができれば，DBN の初期学習方式として高い効果を発揮する可能性がある．以上の動機のもと，本発表では Bernoulli-Bernoulli 型 RBM と Gaussian-Bernoulli 型 RBM の学習問題に焦点を当て，補助関数法に基づく新しい学習則を提案する．

2. 制約付き Boltzmann マシン [3]

2.1 RBM 学習における目的関数

RBM は，Fig. 1 で示されるように完全 2 部グラフの無向グラフの構造を持ち，観測される状態を可視層，背後にある状態を隠れ層と呼ぶ．このとき，可視層同士や隠れ層同士には結合が無いため “制約付き” と呼ばれる．RBM には，可視層の状態と隠れ層の状態が 0 または 1 のバイナリ値をとる Bernoulli-Bernoulli 型，可視層の状態が実数値を取り，隠れ層の状態が 0 または 1 のバイナリ値をとる Gaussian-Bernoulli 型と呼ばれるものがある．

そこで，まず Bernoulli-Bernoulli 型 RBM について説明する．可視層の状態数を I，可視層の状態を $\boldsymbol{v} = \{v_i\} \in \{0,1\}^I$，隠れ層の状態数を J，隠れ層の状態を $\boldsymbol{h} = \{h_j\} \in \{0,1\}^J$ とすると可視層の状態と隠れ層の状態を確率変数とした Bernoulli-Bernoulli 型 RBM の同時確率は

$$p(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\Theta}) = \frac{\exp(-E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\Theta}))}{Z(\boldsymbol{\Theta})} \tag{1}$$

で定義される．ここで，

$$E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\Theta}) = -\sum_i b_i^{\mathrm{V}} v_i - \sum_j b_j^{\mathrm{H}} h_j - \sum_{i,j} W_{ij} v_i h_j, \tag{2}$$

$$Z(\boldsymbol{\Theta}) = \sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} \exp(-E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\Theta})) \tag{3}$$

であり，$\boldsymbol{\Theta} = \{b_i^{\mathrm{V}}, b_j^{\mathrm{H}}, W_{ij}\}$ は分布パラメータである．また，$\sum_{\boldsymbol{v}}$ や $\sum_{\boldsymbol{h}}$ は $\boldsymbol{v}$ や $\boldsymbol{h}$ のすべてのバイナリパターンに関しての和を表す．

ここで，可視層同士，隠れ層同士の依存関係が無いことから，

$$p(\boldsymbol{v}|\boldsymbol{h}, \boldsymbol{\Theta}) = \prod_i p(v_i|\boldsymbol{h}, \boldsymbol{\Theta}), \tag{4}$$

$$p(\boldsymbol{h}|\boldsymbol{v}, \boldsymbol{\Theta}) = \prod_j p(h_j|\boldsymbol{v}, \boldsymbol{\Theta}) \tag{5}$$

という関係があり，式 (1) から

$$p(v_i = 1|\boldsymbol{h}, \boldsymbol{\Theta}) = \frac{1}{1 + \exp(-b_i^{\mathrm{V}} - \sum_j W_{ij} h_j)}, \tag{6}$$

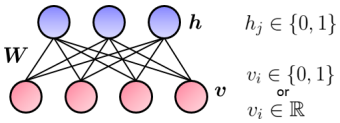


図 1　RBM のグラフ表現

$$p(h_j = 1|\boldsymbol{v}, \boldsymbol{\Theta}) = \frac{1}{1 + \exp(-b_j^{\mathrm{H}} - \sum_i W_{ij} v_i)} \tag{7}$$

となる．このように可視層の条件付分布が Bernoulli 分布，隠れ層の条件付分布が Bernoulli 分布となるため，Bernoulli-Bernoulli 型と呼ばれる．

次に Gaussian-Bernoulli 型 RBM について説明する．可視層の状態数，隠れ層の状態数，隠れ層の状態は Bernoulli-Bernoulli 型と同様に I，J，$\boldsymbol{h} = \{h_j\} \in \{0,1\}^J$ とし，可視層の状態は実数値をとるので，$\boldsymbol{v} = \{v_i\} \in (-\infty, \infty)^I$ とおくと可視層の状態と隠れ層の状態を確率変数とした Gaussian-Bernoulli 型 RBM の同時確率分布は

$$p(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\Theta}) = \frac{\exp(-E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\Theta}))}{Z(\boldsymbol{\Theta})} \tag{8}$$

で定義される．ここで，

$$E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\Theta}) = \sum_i \frac{v_i^2}{2\sigma_i^2} - \sum_i \frac{b_i^{\mathrm{V}} v_i}{\sigma_i^2} - \sum_j b_j^{\mathrm{H}} h_j - \sum_{i,j} \frac{W_{ij} v_i h_j}{\sigma_i}, \tag{9}$$

$$Z(\boldsymbol{\Theta}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \sum_{\boldsymbol{h}} \exp(-E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\Theta})) \mathrm{d}v_1 \cdots \mathrm{d}v_I \tag{10}$$

であり，$\boldsymbol{\Theta} = \{b_i^{\mathrm{V}}, b_j^{\mathrm{H}}, W_{ij}\}$ や $\{\sigma_i\}$ は分布パラメータである．ここで，問題の簡単化のために本論文では $\{\sigma_i\}$ は定数として取り扱う．

Bernoulli-Bernoulli 型と同様に式 (8) から条件付分布を求めると，

$$p(v_i|\boldsymbol{h}, \boldsymbol{\Theta}) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(v_i - b_i^{\mathrm{V}} - \sigma_i \sum_j W_{ij} h_j)^2}{2\sigma_i^2} \right), \tag{11}$$

$$p(h_j = 1|\boldsymbol{v}, \boldsymbol{\Theta}) = \frac{1}{1 + \exp(-b_j^{\mathrm{H}} - \sum_i W_{ij} v_i)} \tag{12}$$

となり，可視層の条件付分布が Gauss 分布，隠れ層の条件付分布が Bernoulli 分布となるため，Gaussian-Bernoulli 型と呼ばれる．

RBM の学習問題とは観測される N 個の可視層のデータ $\boldsymbol{v}^{(1)}, \cdots, \boldsymbol{v}^{(N)}$ からこれらのパラメータ $\boldsymbol{\Theta}$ を推定することである．

このとき，Bernoulli-Bernoulli 型，Gaussian-Bernoulli 型の RBM の学習問題に対するよく用いられる目的関数として，次の周辺分布の対数尤度関数が挙げられる．

$$J(\boldsymbol{\Theta}) = \frac{1}{N} \sum_n \log p(\boldsymbol{v}^{(n)}|\boldsymbol{\Theta}) = \frac{1}{N} \sum_n \log \sum_{\boldsymbol{h}} p(\boldsymbol{v}^{(n)}, \boldsymbol{h}|\boldsymbol{\Theta}). \tag{13}$$

2.2 Contrastive Divergence 法 [1], [4]

最急降下法により，式 (13) で表される周辺分布の対数尤度関

数 $J(\boldsymbol{\Theta})$ の最大化を考える．

まず，Bernoulli-Bernoulli 型 RBM について，$J(\boldsymbol{\Theta})$ を $\boldsymbol{\Theta}$ に関して微分すると，

$$\begin{aligned}\frac{\partial J}{\partial \boldsymbol{\Theta}}(\boldsymbol{\Theta}) = &-\frac{1}{N}\sum_n \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}^{(n)},\boldsymbol{\Theta})\frac{\partial E}{\partial \boldsymbol{\Theta}}(\boldsymbol{v}^{(n)},\boldsymbol{h}|\boldsymbol{\Theta}) \\ &+\sum_{\boldsymbol{v}}\sum_{\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta})\frac{\partial E}{\partial \boldsymbol{\Theta}}(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta})\end{aligned} \tag{14}$$

となる．このため，最急降下法によるパラメータの更新は

$$\boldsymbol{\Theta} \leftarrow \boldsymbol{\Theta}^{\mathrm{old}} + \epsilon \Delta_{\mathrm{cd}}\boldsymbol{\Theta} \tag{15}$$

となる．ただし，ϵ は学習率，$\Delta_{\mathrm{cd}}\boldsymbol{\Theta} = \partial J/\partial \boldsymbol{\Theta}$ である．

ここで，$\boldsymbol{v}$ や $\boldsymbol{h}$ についての和に関しては厳密に計算しようとすると $O(2^I)$ や $O(2^J)$ の計算量となるため，現実的ではない．しかし，式 (14) の第 1 項に関しては式 (5) を用いて周辺化を行うことにより，$\sum_{\boldsymbol{h}}$ を $\sum_j$ にすることが出来る．

しかし，式 (14) の第 2 項に関してはどうしても計算が困難である．そこで，式 (14) の第 2 項は同時確率による期待値計算であることに注目すると，次式のようなギブスサンプリングによる同時確率の近似を用いることで計算量を削減することが考えられる．

$$p(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta}) \approx \frac{1}{M}\sum_m \delta(\boldsymbol{v}-\boldsymbol{v}^{(m)})p(\boldsymbol{h}|\boldsymbol{v}^{(m)},\boldsymbol{\Theta}) \tag{16}$$

ここで，式 (4)，(5) からギブスサンプリングは

$$h_j^{d-1} \sim p(h_j|\boldsymbol{v}^{d-1},\boldsymbol{\Theta}), \tag{17}$$

$$v_i^{d} \sim p(v_i|\boldsymbol{h}^{d-1},\boldsymbol{\Theta}) \tag{18}$$

と容易に行うことが可能である．

式 (15) による更新を式 (16) のギブスサンプリングによる近似で行う手法を CD 法という．

次に Gaussian-Bernoulli 型 RBM について，$J(\boldsymbol{\Theta})$ を $\boldsymbol{\Theta}$ に関して微分すると，

$$\begin{aligned}\frac{\partial J}{\partial \boldsymbol{\Theta}}(\boldsymbol{\Theta}) = &-\frac{1}{N}\sum_n \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}^{(n)},\boldsymbol{\Theta})\frac{\partial E}{\partial \boldsymbol{\Theta}}(\boldsymbol{v}^{(n)},\boldsymbol{h}|\boldsymbol{\Theta}) \\ &+\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\sum_{\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta})\frac{\partial E}{\partial \boldsymbol{\Theta}}(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta})\mathrm{d}v_1\cdots\mathrm{d}v_I\end{aligned} \tag{19}$$

となる．これを計算するには，第 1 項に関しては Bernoulli-Bernoulli 型と同様に式 (5) を用いて周辺化を行えばよい．また，第 2 項に関しても Bernoulli-Bernoulli 型と同様に同時確率による期待値計算であるので，同様にギブスサンプリングによる近似を行うことで近似的に計算が可能となる．よって，Gaussian-Bernoulli 型でも Bernoulli-Bernoulli 型と同様に CD 法によりパラメータ学習を行うことができる．

3. 補助関数法による RBM の学習アルゴリズム

補助関数法による目的関数 $F(\boldsymbol{x})$ の最大化問題の最適化アルゴリズムは，補助変数 $\bar{\boldsymbol{x}}$ を導入して，任意の $\boldsymbol{x}$，$\bar{\boldsymbol{x}}$ で $F(\boldsymbol{x}) \geqq F^+(\boldsymbol{x},\bar{\boldsymbol{x}})$ となり，$F(\boldsymbol{x}) = \max_{\bar{\boldsymbol{x}}} F^+(\boldsymbol{x},\bar{\boldsymbol{x}})$ となるような下限関数 $F^+(\boldsymbol{x},\bar{\boldsymbol{x}})$ を設計して，$F^+(\boldsymbol{x},\bar{\boldsymbol{x}})$ を $\boldsymbol{x}$ についての最小化と $\bar{\boldsymbol{x}}$ についての最小化を交互に行うことである．ここで重要なのは，$F^+(\boldsymbol{x},\bar{\boldsymbol{x}})$ を $\boldsymbol{x}$ についての最小化が容易に行え，かつ $F^+(\boldsymbol{x},\bar{\boldsymbol{x}})$ が $F(\boldsymbol{x})$ によくフィットするように設計できるかであるので，本研究では $\boldsymbol{x}$ の各変数の相互依存をなくすような $F^+(\boldsymbol{x},\bar{\boldsymbol{x}})$ を設計することを目指す．

3.1 最尤学習

ここでは最尤学習，つまり，目的関数として式 (13) で表される対数尤度を用い，これを最大化することを考える．

まず，Bernoulli-Bernoulli 型 RBM について考えると，Jensen の不等式から

$$\begin{aligned}J(\boldsymbol{\Theta}) &= \frac{1}{N}\sum_n \log\sum_{\boldsymbol{h}} p(\boldsymbol{v}^{(n)},\boldsymbol{h}|\boldsymbol{\Theta}) \\ &\geqq \frac{1}{N}\sum_n\sum_{\boldsymbol{h}} \lambda_{n,\boldsymbol{h}}\log p(\boldsymbol{v}^{(n)},\boldsymbol{h}|\boldsymbol{\Theta}) \\ &\quad -\sum_{\boldsymbol{h}}\lambda_{n,\boldsymbol{h}}\log\lambda_{n,\boldsymbol{h}}.\end{aligned} \tag{20}$$

となる．ここで，等号の成立は

$$\lambda_{n,\boldsymbol{h}} = p(\boldsymbol{h}|\boldsymbol{v}^{(n)},\boldsymbol{\Theta}) \tag{21}$$

のときである．式 (20) を整理すると，

$$\begin{aligned}J(\boldsymbol{\Theta}) \geqq &-\frac{1}{N}\sum_n\sum_{\boldsymbol{h}}\lambda_{n,\boldsymbol{h}}E(\boldsymbol{v}^{(n)},\boldsymbol{h}|\boldsymbol{\Theta}) \\ &-\log Z(\boldsymbol{\Theta}) - \sum_{\boldsymbol{h}}\lambda_{n,\boldsymbol{h}}\log\lambda_{n,\boldsymbol{h}}\end{aligned} \tag{22}$$

となる．ここで，この式の第 2 項について考えると，負の対数関数は凸関数であるので，接線の方程式を用いて下から抑えることが出来る．

$$-\log Z(\boldsymbol{\Theta}) \geqq -\frac{Z(\boldsymbol{\Theta})}{\zeta} - \log\zeta + 1. \tag{23}$$

ここで，等号の成立は接点となるので，

$$\zeta = Z(\boldsymbol{\Theta}) \tag{24}$$

となる．ここで，式 (2) から $E(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta})$ はパラメータ $\boldsymbol{\Theta}$ に対し線形であるので，

$$a_k(\boldsymbol{v},\boldsymbol{h}) = \begin{cases} v_i & (k=i) \\ h_j & (k=I+j) \\ v_i h_j & (k=I+J+J\times(i-1)+j)\end{cases}, \tag{25}$$

$$\theta_k = \begin{cases} b_i^{\mathrm{V}} & (k=i) \\ b_j^{\mathrm{H}} & (k=I+j) \\ W_{ij} & (k=I+J+J\times(i-1)+j)\end{cases} \tag{26}$$

とおくと，式 (2) は

$$E(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta}) = -\sum_k a_k(\boldsymbol{v},\boldsymbol{h})\theta_k \tag{27}$$

と表すことができる．このとき，$-\exp(\sum_k a_k(\boldsymbol{v},\boldsymbol{h})\theta_k)$ に対して，複素 NMF [5] で用いられている Jensen の不等式を用いた補助関数の設計法を用いると，

$$
\begin{aligned}
&- \exp\Big(\sum_k a_k(\boldsymbol{v},\boldsymbol{h})\theta_k\Big) \\
&\geqq - \sum_k \beta_k(\boldsymbol{v},\boldsymbol{h}) \exp\Big(\frac{a_k(\boldsymbol{v},\boldsymbol{h})\theta_k - \alpha_k(\boldsymbol{v},\boldsymbol{h})}{\beta_k(\boldsymbol{v},\boldsymbol{h})}\Big)
\end{aligned} \tag{28}
$$

となり，等号の成立は

$$
\forall \beta_k(\boldsymbol{v},\boldsymbol{h}) \in [0,1],\ \sum_k \beta_k(\boldsymbol{v},\boldsymbol{h}) = 1, \tag{29}
$$

$$
\alpha_k(\boldsymbol{v},\boldsymbol{h}) = a_k(\boldsymbol{v},\boldsymbol{h})\theta_k - \beta_k(\boldsymbol{v},\boldsymbol{h}) \sum_l a_l(\boldsymbol{v},\boldsymbol{h})\theta_l \tag{30}
$$

となる．ここで，$\beta_k(\boldsymbol{v},\boldsymbol{h})$ は任意に設計できるので，$\boldsymbol{v}$，$\boldsymbol{h}$ に依存しない定数 β_k とする．また，補助関数法は補助変数と従来変数を交互に更新する手法であるので，一反復前の従来変数 θ^{old} を新たな補助変数として導入すると，補助変数の更新式の従来変数 θ を θ^{old} に置き換えて代入することができるので，式 (28) に補助変数の更新式 (30) を代入し，式 (23) を式 (3), (27) 用いて整理すると，

$$
\begin{aligned}
-\log Z(\boldsymbol{\Theta}) \geqq &- \sum_{\boldsymbol{v}}\sum_{\boldsymbol{h}} \frac{\exp(\sum_l a_l(\boldsymbol{v},\boldsymbol{h})\theta_l^{\mathrm{old}})}{\zeta} \\
&\times \sum_k \beta_k \exp\Big(\frac{a_k(\boldsymbol{v},\boldsymbol{h})(\theta_k - \theta_k^{\mathrm{old}})}{\beta_k}\Big) \\
&- \log\zeta + 1
\end{aligned} \tag{31}
$$

となる．ここで，式 (1)，(24) から式 (31) は

$$
\begin{aligned}
-\log Z(\boldsymbol{\Theta}) \geqq &- \sum_k \beta_k \sum_{\boldsymbol{v}}\sum_{\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta}^{\mathrm{old}}) \\
&\times \exp\Big(\frac{a_k(\boldsymbol{v},\boldsymbol{h})(\theta_k - \theta_k^{\mathrm{old}})}{\beta_k}\Big) \\
&- \log\zeta + 1
\end{aligned} \tag{32}
$$

となる．式 (21)，(22)，(24)，(32) より，$\bar{\boldsymbol{\Theta}} = \{\boldsymbol{\Theta}^{\mathrm{old}}, \beta_k\}$ とおいたとき，$J(\boldsymbol{\Theta})$ の下限関数 $J^+(\boldsymbol{\Theta},\bar{\boldsymbol{\Theta}})$ は

$$
\begin{aligned}
J^+(\boldsymbol{\Theta},\bar{\boldsymbol{\Theta}}) = &\frac{1}{N}\sum_n \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}^{(n)},\boldsymbol{\Theta}^{\mathrm{old}}) \sum_k a_k(\boldsymbol{v}^{(n)},\boldsymbol{h})\theta_k \\
&- \sum_k \beta_k \sum_{\boldsymbol{v}}\sum_{\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta}^{\mathrm{old}}) \\
&\quad \times \exp\Big(\frac{a_k(\boldsymbol{v},\boldsymbol{h})(\theta_k - \theta_k^{\mathrm{old}})}{\beta_k}\Big) \\
&+ C(\bar{\boldsymbol{\Theta}})
\end{aligned} \tag{33}
$$

と定義できる．ただし $C(\bar{\boldsymbol{\Theta}})$ は $\boldsymbol{\Theta}$ に対して定数の項である．

ここで，式 (33) を θ_k について微分すると，

$$
\begin{aligned}
\frac{\partial J^+}{\partial \theta_k}(\boldsymbol{\Theta},\bar{\boldsymbol{\Theta}}) = &\frac{1}{N}\sum_n \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}^{(n)},\boldsymbol{\Theta}^{\mathrm{old}}) a_k(\boldsymbol{v}^{(n)},\boldsymbol{h}) \\
&- \sum_{\boldsymbol{v}}\sum_{\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta}^{\mathrm{old}}) a_k(\boldsymbol{v},\boldsymbol{h}) \\
&\quad \times \exp\Big(\frac{a_k(\boldsymbol{v},\boldsymbol{h})(\theta_k - \theta_k^{\mathrm{old}})}{\beta_k}\Big)
\end{aligned} \tag{34}
$$

となる．ここで，式 (25) から Bernoulli-Bernoulli 型 RBM においては $a_k(\boldsymbol{v},\boldsymbol{h}) \in \{0,1\}$ であるため，式 (34) の第二項は

$$
\begin{aligned}
&\sum_{\boldsymbol{v}}\sum_{\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta}^{\mathrm{old}}) a_k(\boldsymbol{v},\boldsymbol{h}) \exp\Big(\frac{a_k(\boldsymbol{v},\boldsymbol{h})(\theta_k - \theta_k^{\mathrm{old}})}{\beta_k}\Big) \\
&= \exp\Big(\frac{\theta_k - \theta_k^{\mathrm{old}}}{\beta_k}\Big) \sum_{\boldsymbol{v}}\sum_{\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta}^{\mathrm{old}}) a_k(\boldsymbol{v},\boldsymbol{h})
\end{aligned} \tag{35}
$$

となる．よって，$\partial J^+/\partial\theta_k = 0$ を解析的に求めることができ，$\Delta_{\mathrm{af}}\theta_k$ を

$$
\begin{aligned}
\Delta_{\mathrm{af}}\theta_k = &\log\frac{1}{N}\sum_n\sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}^{(n)},\boldsymbol{\Theta}^{\mathrm{old}}) a_k(\boldsymbol{v}^{(n)},\boldsymbol{h}) \\
&- \log\sum_{\boldsymbol{v}}\sum_{\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta}^{\mathrm{old}}) a_k(\boldsymbol{v},\boldsymbol{h})
\end{aligned} \tag{36}
$$

と定義すると，θ_k の更新式は，

$$
\theta_k \leftarrow \theta_k^{\mathrm{old}} + \beta_k \Delta_{\mathrm{af}}\theta_k \tag{37}
$$

となる．ここで，式 (36) の右辺の第 2 項は 2.2 節で言及したように厳密に計算することは困難である．そこで，CD 法と同様にギプスサンプリングによる $\boldsymbol{v}$ の周辺確率の近似を行うことで，式 (37) の計算が可能となる．

さて，式 (15)，(37) を比較すると，式 (37) の更新式はあたかも学習率が β_k であるかのような形となっている．ところが，式 (29) よりパラメータ数が多くなると β_k の平均的な値は小さくなるため，収束が遅くなることが予想される．そこで，収束を速くするために以下の近似を考える．

$$
\beta_k' \leftarrow \beta_k^{\gamma}. \tag{38}
$$

このとき，$\gamma \in [0,1]$ ならば，$\beta_k \in [0,1]$ を満たすので β_k' は β_k よりも大きくなる．そのため，式 (37) による更新が速くなることが期待できる．ただし，β_k' は式 (29) を満たさないため，補助関数法における収束性は保証されないことには注意されたい．

以上より Bernoulli-Bernoulli 型 RBM の補助関数法による最尤学習アルゴリズムは，次の 1)〜3) を反復することである．1) 補助変数 $\bar{\boldsymbol{\Theta}}$ を求める．2) ギプスサンプリングで $p(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta}^{\mathrm{old}})$ の近似値を求める．3) 式 (37) の更新式でパラメータを更新．

次に，Gaussian-Bernoulli 型 RBM について考える．式 (1) と式 (8) は同じ構造をしており，式 (2) と式 (9) を比較すると，$\{\sigma_i\}$ を定数だとしているので，学習すべきパラメータは Bernoulli-Bernoulli 型と同じ $\boldsymbol{\Theta} = \{b_i^{\mathrm{V}}, b_j^{\mathrm{H}}, W_{ij}\}$ であり，Bernoulli-Bernoulli 型と同じく $E(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta})$ において線形和で表現されているため，Bernoulli-Bernoulli 型と同じ手順で補助関数が設計可能である．よって，$J(\boldsymbol{\Theta})$ の下限関数 $J^+(\boldsymbol{\Theta},\bar{\boldsymbol{\Theta}})$ は

$$
\begin{aligned}
J^+(\boldsymbol{\Theta},\bar{\boldsymbol{\Theta}}) = &\frac{1}{N}\sum_n \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}^{(n)},\boldsymbol{\Theta}^{\mathrm{old}}) \sum_k a_k(\boldsymbol{v}^{(n)},\boldsymbol{h})\theta_k \\
&- \sum_k \beta_k \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \sum_{\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta}^{\mathrm{old}}) \\
&\quad \times \exp\Big(\frac{a_k(\boldsymbol{v},\boldsymbol{h})(\theta_k - \theta_k^{\mathrm{old}})}{\beta_k}\Big) \mathrm{d}v_1 \cdots \mathrm{d}v_I \\
&+ C(\bar{\boldsymbol{\Theta}})
\end{aligned} \tag{39}
$$

と設計できる．ただし $C(\bar{\boldsymbol{\Theta}})$ は $\boldsymbol{\Theta}$ に対して定数の項であり，

$$a_k(\boldsymbol{v},\boldsymbol{h})=\begin{cases}\frac{v_i}{\sigma_i^2} & (k=i)\\ h_j & (k=I+j)\\ \frac{v_i h_j}{\sigma_i} & (k=I+J+J\times(i-1)+j)\end{cases}, \tag{40}$$

$$\theta_k=\begin{cases}b_i^{\mathrm{V}} & (k=i)\\ b_j^{\mathrm{H}} & (k=I+j)\\ W_{ij} & (k=I+J+J\times(i-1)+j)\end{cases}, \tag{41}$$

また，β_k は $\beta_k\in[0,1]$，$\sum_k\beta_k=1$ を満たす任意定数である．

ここで，式 (39) を θ_k について微分すると，

$$\begin{aligned}\frac{\partial J^+}{\partial\theta_k}(\boldsymbol{\Theta},\bar{\boldsymbol{\Theta}})=&\frac{1}{N}\sum_n\sum_{\boldsymbol{h}}p(\boldsymbol{h}|\boldsymbol{v}^{(n)},\boldsymbol{\Theta}^{\mathrm{old}})a_k(\boldsymbol{v}^{(n)},\boldsymbol{h})\\ &-\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\sum_{\boldsymbol{h}}p(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta}^{\mathrm{old}})a_k(\boldsymbol{v},\boldsymbol{h})\\ &\times\exp\Big(\frac{a_k(\boldsymbol{v},\boldsymbol{h})(\theta_k-\theta_k^{\mathrm{old}})}{\beta_k}\Big)\mathrm{d}v_1\cdots\mathrm{d}v_I\end{aligned} \tag{42}$$

となる．ここで，式 (42) の右辺の第 2 項は 2.2 節で言及し，また Bernoulli-Bernoulli 型に対する補助関数法によるアルゴリズムと同様に厳密に計算することは困難である．そこで，CD 法などと同様に Gibbs サンプリングによる $\boldsymbol{v}$ の周辺確率の近似を行い，同時確率を式 (16) で近似すると式 (42) は

$$\begin{aligned}\frac{\partial J^+}{\partial\theta_k}(\boldsymbol{\Theta},\bar{\boldsymbol{\Theta}})=&\frac{1}{N}\sum_n\sum_{\boldsymbol{h}}p(\boldsymbol{h}|\boldsymbol{v}^{(n)},\boldsymbol{\Theta}^{\mathrm{old}})a_k(\boldsymbol{v}^{(n)},\boldsymbol{h})\\ &-\frac{1}{M}\sum_m\sum_{\boldsymbol{h}}p(\boldsymbol{h}|\boldsymbol{v}^{(m)},\boldsymbol{\Theta}^{\mathrm{old}})a_k(\boldsymbol{v}^{(m)},\boldsymbol{h})\\ &\times\exp\Big(\frac{a_k(\boldsymbol{v}^{(m)},\boldsymbol{h})(\theta_k-\theta_k^{\mathrm{old}})}{\beta_k}\Big)\end{aligned} \tag{43}$$

となる．ここで，$a_k(\boldsymbol{v},\boldsymbol{h})\in\{0,1\}$ になる場合を除いて $\partial J^+/\partial\theta_k=0$ は解析的に解くことは困難である．Bernoulli-Bernoulli 型の RBM では常に $a_k(\boldsymbol{v},\boldsymbol{h})\in\{0,1\}$ となっているが，Gaussian-Bernoulli 型の RBM では θ_k が b_j^{H} のみ $a_k(\boldsymbol{v},\boldsymbol{h})\in\{0,1\}$ となる．しかし，$\partial^2J^+/\partial\theta_k^2$ は常に負となるため，Newton 法により安定して最大点を求めることが期待できる．また，Bernoulli-Bernoulli 型に対する補助関数と同型の補助関数を用いているので，β_k は同様の役割を担っていると考えられる．そのため，Bernoulli-Bernoulli 型で議論したように，収束を速くするために式 (38) による近似を考える．

以上より Gaussian-Bernoulli 型 RBM の補助関数法による最尤学習アルゴリズムは，次の 1)〜3) を反復することである．1) 補助変数 $\bar{\boldsymbol{\Theta}}$ を求める．2)Gibbs サンプリングで $p(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta}^{\mathrm{old}})$ の近似値を求める．3)$\partial J^+/\partial\theta_k=0$ となるようにパラメータを更新 (一部 Newton 法を用いる)．

3.2 最大再構築確率学習

次に，RBM の学習アルゴリズムとして，今度は目的関数が他と異なるものを導出する．RBM を学習する際は，可視層に入力されたデータが隠れ層で変換された特徴量を更に可視層に変換するときに元の入力データに近くなるように学習が進むという特徴がある．これは，変換された特徴量が入力データの情報をなるべく保持しようと学習されていると考えられる．そこで，本研究ではこの特徴を陽に記述した目的関数を導入することを考え，次のような可視層に観測データが来たときに，ギプスサンプリングを 1 回行ったときに元の観測データが再現される確率の対数を考える．

$$J_r(\boldsymbol{\Theta})=\frac{1}{N}\sum_n\log\sum_{\boldsymbol{h}}p(\boldsymbol{v}^{(n)}|\boldsymbol{h},\boldsymbol{\Theta})p(\boldsymbol{h}|\boldsymbol{v}^{(n)},\boldsymbol{\Theta}). \tag{44}$$

この元の観測データが再現される確率を再構築確率と呼び，これを最大化する最大再構築確率学習について補助関数法を用いて学習アルゴリズムを設計することを目指す．

この式を直接最大化するのは困難であるので，この式に対して，

$$\boldsymbol{h}^{(n)}\sim p(\boldsymbol{h}|\boldsymbol{v}^{(n)},\boldsymbol{\Theta}) \tag{45}$$

でサンプリングした値を元に

$$J_r(\boldsymbol{\Theta})\approx\frac{1}{N}\sum_n\log p(\boldsymbol{v}^{(n)}|\boldsymbol{h}^{(n)},\boldsymbol{\Theta})p(\boldsymbol{h}^{(n)}|\boldsymbol{v}^{(n)},\boldsymbol{\Theta}) \tag{46}$$

と近似することを考える．この右辺を $\tilde{J}_r$ とおく．

まず，Bernoulli-Bernoulli 型 RBM について $\tilde{J}_r$ の補助関数の設計を考える．式 (4)，(5)，Bernoulli-Bernoulli 型 RBM の条件付分布が Bernoulli 分布となることより，

$$\begin{aligned}\tilde{J}_r=&\frac{1}{N}\sum_n\Big\{\sum_i v_i^{(n)}\log q_{ni}^{\mathrm{V}}+\sum_i(1-v_i^{(n)})\log(1-q_{ni}^{\mathrm{V}})\\ &+\sum_j h_j^{(n)}\log q_{nj}^{\mathrm{H}}+\sum_j(1-h_j^{(n)})\log(1-q_{nj}^{\mathrm{H}})\Big\}\end{aligned} \tag{47}$$

となる．ただし，$q_{ni}^{\mathrm{V}}=p(v_i^{(n)}=1|\boldsymbol{h}^{(n)},\boldsymbol{\Theta})$，$q_{nj}^{\mathrm{H}}=p(h_j^{(n)}=1|\boldsymbol{v}^{(n)},\boldsymbol{\Theta})$ である．よって，式 (6)，(7) より式 (47) を整理すると，

$$\begin{aligned}\tilde{J}_r=-\frac{1}{N}\sum_n\Big\{&\sum_i v_i^{(n)}\log(1+\exp(-f_{ni}^{\mathrm{V}}))\\ &+\sum_i(1-v_i^{(n)})\log(1+\exp f_{ni}^{\mathrm{V}})\\ &+\sum_j h_j^{(n)}\log(1+\exp(-f_{nj}^{\mathrm{H}}))\\ &+\sum_j(1-h_j^{(n)})\log(1+\exp f_{nj}^{\mathrm{H}})\Big\}\end{aligned} \tag{48}$$

となる．ただし，$f_{ni}^{\mathrm{V}}=b_i^{\mathrm{V}}+\sum_j W_{ij}h_j^{(n)}$，$f_{nj}^{\mathrm{H}}=b_j^{\mathrm{H}}+\sum_i W_{ij}v_i^{(n)}$ である．ここで，式 (48) のそれぞれの項は負の対数関数であり，その引数の項をみると，式 (23)〜(32) と同様の論理で下限関数が設計できることが分かる．

よって，下限関数 $\tilde{J}_r^+(\boldsymbol{\Theta},\bar{\boldsymbol{\Theta}})$ は

$$\begin{aligned}&\tilde{J}_r^+(\boldsymbol{\Theta},\bar{\boldsymbol{\Theta}})\\ &=-\frac{1}{N}\Big\{\sum_i v_i^{(n)}(1-\hat{q}_{ni}^{\mathrm{V}})\xi_{ni}^{\mathrm{V}}+\sum_i(1-v_i^{(n)})\hat{q}_{ni}^{\mathrm{V}}\eta_{ni}^{\mathrm{V}}\\ &\quad+\sum_j h_j^{(n)}(1-\hat{q}_{nj}^{\mathrm{H}})\xi_{nj}^{\mathrm{H}}+\sum_j(1-h_j^{(n)})\hat{q}_{nj}^{\mathrm{H}}\eta_{nj}^{\mathrm{H}}\Big\}\\ &\quad+C(\bar{\boldsymbol{\Theta}})\end{aligned} \tag{49}$$

となる．ただし，$\hat{q}_{ni}^{\mathrm{V}}=p(v_i^{(n)}=1|\boldsymbol{h}^{(n)},\boldsymbol{\Theta}^{\mathrm{old}})$，$\hat{q}_{nj}^{\mathrm{H}}=$

$p(h_j^{(n)}=1|\boldsymbol{v}^{(n)},\boldsymbol{\Theta}^{\rm old})$, $\xi_{ni}^{\rm V}=\beta_{i0}^{\rm V}e^{-\hat{b}_i^{\rm V}}+\sum_j\beta_{ij}^{\rm V}e^{-\hat{W}_{nij}^{\rm V}}$, $\eta_{ni}^{\rm V}=\beta_{i0}^{\rm V}e^{\hat{b}_i^{\rm V}}+\sum_j\beta_{ij}^{\rm V}e^{\hat{W}_{nij}^{\rm V}}$, $\xi_{nj}^{\rm H}=\beta_{0j}^{\rm H}e^{-\hat{b}_j^{\rm H}}+\sum_i\beta_{ij}^{\rm H}e^{-\hat{W}_{nij}^{\rm H}}$, $\eta_{nj}^{\rm H}=\beta_{0j}^{\rm H}e^{\hat{b}_j^{\rm H}}+\sum_i\beta_{ij}^{\rm H}e^{\hat{W}_{nij}^{\rm H}}$, $\hat{b}_i^{\rm V}=\frac{b_i^{\rm V}-b_i^{\rm V,old}}{\beta_{i0}^{\rm V}}$, $\hat{b}_j^{\rm H}=\frac{b_j^{\rm H}-b_j^{\rm H,old}}{\beta_{0j}^{\rm H}}$, $\hat{W}_{nij}^{\rm V}=\frac{W_{ij}-W_{ij}^{\rm old}}{\beta_{ij}^{\rm V}}h_j^{(n)}$, $\hat{W}_{nij}^{\rm H}=\frac{W_{ij}-W_{ij}^{\rm old}}{\beta_{ij}^{\rm H}}v_i^{(n)}$, であり，$\beta_{i0}^{\rm V}$, $\beta_{ij}^{\rm V}$, $\beta_{0j}^{\rm H}$, $\beta_{ij}^{\rm H}$ は

$$\begin{aligned}&\beta_{i0}^{\rm V},\beta_{ij}^{\rm V},\beta_{0j}^{\rm H},\beta_{ij}^{\rm H}\in[0,1],\\&\beta_{i0}^{\rm V}+\sum_j\beta_{ij}^{\rm V}=1,\quad\forall i,\ \beta_{0j}^{\rm H}+\sum_i\beta_{ij}^{\rm H}=1,\quad\forall j,\end{aligned}\tag{50}$$

を満たす任意定数である．また，$\bar{\boldsymbol{\Theta}}=\{\boldsymbol{\Theta}^{\rm old},\beta_{i0}^{\rm V},\beta_{ij}^{\rm V},\beta_{0j}^{\rm H},\beta_{ij}^{\rm H}\}$ であり，$C(\bar{\boldsymbol{\Theta}})$ は $\boldsymbol{\Theta}$ に対して定数の項である．

ここで，$\partial\tilde{J}_r^+/\partial\boldsymbol{\Theta}=0$ を解くことについて考えると，$b_i^{\rm V}$, $b_j^{\rm H}$ については解析的に解くことが出来るが，W_{ij} については解析的に解くことが出来ない．しかし，これは，$\partial^2\tilde{J}_r^+/\partial W_{ij}^2<0$ であるので，Newton 法を用いて効率的に求めることが出来る．

また，$\beta_{i0}^{\rm V}$, $\beta_{ij}^{\rm V}$, $\beta_{0j}^{\rm H}$, $\beta_{ij}^{\rm H}$ が最尤学習における補助関数法を用いた学習アルゴリズムと同様に学習率を担っていると考えられるため，これらについても γ 乗を行う近似を考える．

以上より Bernoulli-Bernoulli 型 RBM の補助関数法による最大再構築確率学習アルゴリズムは，次の 1)〜3) を反復することである．1) サンプリングにより $h_j^{(n)}$ を求める．2) 補助変数 $\bar{\boldsymbol{\Theta}}$ を求める．3)$\partial\tilde{J}_r^+/\partial\boldsymbol{\Theta}=0$ から求まる更新式でパラメータを更新.

次に，Gaussian-Bernoulli 型 RBM について補助関数の設計を考える．式 (46) に式 (4)，(5)，(11)，(12) を代入すると，

$$\begin{aligned}\tilde{J}_r(\boldsymbol{\Theta})=-\frac{1}{N}\sum_n\Big\{&\sum_i\frac{(v_i^{(n)}-f_{ni}^{\rm V})^2}{2\sigma_i^2}+\sum_i\log\sigma_i\\&+\frac{I}{2}\log(2\pi)-\sum_j\Big(h_j^{(n)}-\frac{1}{2}\Big)f_{nj}^{\rm H}\\&+\sum_j\log(\exp(f_{nj}^{\rm H}/2)+\exp(-f_{nj}^{\rm H}/2))\Big\}\end{aligned}\tag{51}$$

となる．ただし，$f_{ni}^{\rm V}=b_i^{\rm V}+\sigma_i\sum_j W_{ij}h_j^{(n)}$, $f_{nj}^{\rm H}=b_j^{\rm H}+\sum_i W_{ij}v_i^{(n)}/\sigma_i$ である．

ここで，

$$\begin{aligned}&\log(\exp(x)+\exp(-x))\\&\quad\leqq\frac{\tanh(\kappa)}{2\kappa}x^2-\frac{\kappa\tanh(\kappa)}{2}+\log(2\cosh(\kappa))\end{aligned}\tag{52}$$

という不等式を考える．(右辺) − (左辺) が連続関数で，その極小値が $x=\pm\kappa$ で 0 になり，かつ $x\to\pm\infty$ で正に発散することからこの不等式が成立することが分かる．また，等号の成立は

$$\kappa=\pm x\tag{53}$$

のときである．この不等式を用いると式 (51) は

$$\begin{aligned}\tilde{J}_r(\boldsymbol{\Theta})\geqq-\frac{1}{N}\sum_n\Big\{&\sum_i\frac{(f_{ni}^{\rm V})^2}{2\sigma_i^2}+\sum_j\frac{\tanh(\kappa_{nj})}{8\kappa_{nj}}(f_{nj}^{\rm H})^2\\&-\sum_i\frac{v_i^{(n)}}{\sigma_i^2}f_{ni}^{\rm V}-\sum_j\Big(h_j^{(n)}-\frac{1}{2}\Big)f_{nj}^{\rm H}\\&+\sum_i\Big(\frac{(v_i^{(n)})^2}{2\sigma_i^2}+\log\sigma_i\Big)+\frac{I}{2}\log(2\pi)\\&+\sum_j(\log(\cosh(\kappa_{nj}))-\frac{\kappa_{nj}\tanh(\kappa_{nj})}{8})\Big\}\end{aligned}\tag{54}$$

と下から押さえられる．

ここで，$f_{ni}^{\rm V}=b_i^{\rm V}+\sigma_i\sum_j W_{ij}h_j^{(n)}$, $f_{nj}^{\rm H}=b_j^{\rm H}+\sum_i W_{ij}v_i^{(n)}/\sigma_i$ であるので，$-(f_{ni}^{\rm V})^2$ と $-(f_{nj}^{\rm H})^2$ に対して複素 NMF [5] で用いられている Jensen の不等式を用いた補助関数の設計法を用いると

$$-(f_{ni}^{\rm V})^2\geqq-\frac{(b_i^{\rm V}-\alpha_{ni0}^{\rm V})^2}{\beta_{ni0}^{\rm V}}-\sum_j\frac{(\sigma_i W_{ij}h_j^{(n)}-\alpha_{nij}^{\rm V})^2}{\beta_{nij}^{\rm V}},\tag{55}$$

$$-(f_{nj}^{\rm H})^2\geqq-\frac{(b_j^{\rm H}-\alpha_{n0j}^{\rm H})^2}{\beta_{n0j}^{\rm H}}-\sum_i\frac{(W_{ij}v_i^{(n)}/\sigma_i-\alpha_{nij}^{\rm H})^2}{\beta_{nij}^{\rm H}}\tag{56}$$

となり，等号の成立は

$$^\forall\beta_{ni0}^{\rm V},\beta_{nij}^{\rm V}\in[0,1],\ \beta_{ni0}^{\rm V}+\sum_j\beta_{nij}^{\rm V}=1,\tag{57}$$

$$^\forall\beta_{n0j}^{\rm H},\beta_{nij}^{\rm H}\in[0,1],\ \beta_{n0j}^{\rm H}+\sum_i\beta_{nij}^{\rm H}=1,\tag{58}$$

$$\alpha_{ni0}^{\rm V}=b_i^{\rm V}-\beta_{ni0}^{\rm V}f_{ni}^{\rm V},\ \alpha_{nij}^{\rm V}=\sigma_i W_{ij}h_j^{(n)}-\beta_{nij}^{\rm V}f_{ni}^{\rm V},\tag{59}$$

$$\alpha_{n0i}^{\rm H}=b_j^{\rm H}-\beta_{ni0}^{\rm H}f_{nj}^{\rm H},\ \alpha_{nij}^{\rm H}=\frac{W_{ij}v_i^{(n)}}{\sigma_i}-\beta_{nij}^{\rm H}f_{nj}^{\rm H}\tag{60}$$

のときである．

よって，式 (54)，式 (55)，式 (56) から下限関数 $\tilde{J}_r^+(\boldsymbol{\Theta},\bar{\boldsymbol{\Theta}})$ は

$$\begin{aligned}\tilde{J}_r^+(\boldsymbol{\Theta},\bar{\boldsymbol{\Theta}})=-\frac{1}{N}\sum_n\Big\{&\sum_i\frac{(b_i^{\rm V}-\alpha_{ni0}^{\rm V})^2}{2\beta_{ni0}^{\rm V}\sigma_i^2}\\&+\sum_i\sum_j\frac{(\sigma_i W_{ij}h_j^{(n)}-\alpha_{nij}^{\rm V})^2}{2\beta_{nij}^{\rm V}\sigma_i^2}\\&+\sum_j\frac{\tanh(\kappa_{nj})}{8\kappa_{nj}\beta_{n0j}^{\rm H}}(b_j^{\rm H}-\alpha_{n0j}^{\rm H})^2\\&+\sum_j\sum_i\frac{\tanh(\kappa_{nj})}{8\kappa_{nj}\beta_{nij}^{\rm H}}(W_{ij}v_i^{(n)}/\sigma_i-\alpha_{nij}^{\rm H})^2\\&-\sum_i\frac{v_i^{(n)}}{\sigma_i^2}\Big(b_i^{\rm V}+\sigma_i\sum_j W_{ij}h_j^{(n)}\Big)\\&-\sum_j\Big(h_j^{(n)}-\frac{1}{2}\Big)\Big(b_j^{\rm H}+\sum_i W_{ij}v_i^{(n)}/\sigma_i\Big)\Big\}+C(\bar{\boldsymbol{\Theta}})\end{aligned}\tag{61}$$

となる．ただし，

$$\bar{\boldsymbol{\Theta}}=\{\kappa_{nj},\alpha_{ni0}^{\rm V},\alpha_{nij}^{\rm V},\alpha_{n0j}^{\rm H},\alpha_{nij}^{\rm H},\beta_{ni0}^{\rm V},\beta_{nij}^{\rm V},\beta_{n0j}^{\rm H},\beta_{nij}^{\rm H}\}\tag{62}$$

であり，$C(\bar{\boldsymbol{\Theta}})$ は $\boldsymbol{\Theta}$ に対して定数の項である．

よって，b_i^{V}，b_j^{H}，W_{ij} の更新式は $\partial \tilde{J}_r^{+}/\partial \boldsymbol{\Theta} = 0$ より

$$b_i^{\mathrm{V}} \leftarrow \frac{\sum_n (\alpha_{ni0}^{\mathrm{V}}/\beta_{ni0}^{\mathrm{V}} + v_i^{(n)})}{\sum_n 1/\beta_{ni0}^{\mathrm{V}}}, \tag{63}$$

$$b_j^{\mathrm{H}} \leftarrow \frac{\sum_n (\tanh(\kappa_{nj})\alpha_{n0j}^{\mathrm{H}}/4\kappa_{nj}\beta_{n0j}^{\mathrm{H}} + h_j^{(n)} - 1/2)}{\sum_n \tanh(\kappa_{nj})/4\kappa_{nj}\beta_{n0j}^{\mathrm{H}}}, \tag{64}$$

$$W_{ij} \leftarrow \frac{\sum_n \left(\frac{\alpha_{nij}^{\mathrm{V}}}{\beta_{nij}^{\mathrm{V}}\sigma_i} + \frac{\tanh(\kappa_{nj})\alpha_{nij}^{\mathrm{H}}}{4\kappa_{nj}\beta_{nij}^{\mathrm{H}}} + \frac{v_i^{(n)}(2h_j^{(n)}-1/2)}{\sigma_i} \right)}{\sum_n \left(\frac{h_j^{(n)}}{\beta_{nij}^{\mathrm{V}}\sigma_i} + \frac{\tanh(\kappa_{nj})v_i^{(n)}}{4\kappa_{nj}\beta_{nij}^{\mathrm{H}}\sigma_i} \right)} \tag{65}$$

となる．ここで，β_{ni0}^{V}，β_{nij}^{V}，β_{n0j}^{H}，β_{nij}^{H} は式 (57)，(58) を満たす任意定数であるので，n によらない定数 β_{i0}^{V}，β_{ij}^{V}，β_{0j}^{H}，β_{ij}^{H} とする．そして，新たに補助変数として一反復前の値を表す $b_i^{\mathrm{V,old}}$，$b_j^{\mathrm{H,old}}$，W_{ij}^{old} を導入し，補助変数の更新式 (53)，(59)，(60) を代入して整理すると

$$b_i^{\mathrm{V}} \leftarrow b_i^{\mathrm{V,old}} + \frac{\beta_{i0}^{\mathrm{V}}}{N} \sum_n (v_i^{(n)} - f_{ni}^{\mathrm{V,old}}), \tag{66}$$

$$b_j^{\mathrm{H}} \leftarrow b_j^{\mathrm{H,old}} + \beta_{0j}^{\mathrm{H}} \frac{\sum_n (h_j^{(n)} - q_{nj})}{\sum_n (q_{nj} - 1/2)/f_{nj}^{\mathrm{H,old}}}, \tag{67}$$

$$W_{ij} \leftarrow W_{ij}^{\mathrm{old}} + \frac{\sum_n ((v_i^{(n)} - f_{ni}^{\mathrm{V,old}})h_j^{(n)} + v_i^{(n)}(h_j^{(n)} - q_{nj}))}{\sum_n \left(\frac{h_j^{(n)}\sigma_i}{\beta_{ij}^{\mathrm{V}}} + \frac{(v_i^{(n)})^2(q_{nj}-1/2)}{\beta_{ij}^{\mathrm{H}}\sigma_i f_{nj}^{\mathrm{H,old}}} \right)} \tag{68}$$

となる．ただし，$f_{ni}^{\mathrm{V,old}} = b_i^{\mathrm{V,old}} + \sigma_i \sum_j W_{ij}^{\mathrm{old}} h_j^{(n)}$，$f_{nj}^{\mathrm{H,old}} = b_j^{\mathrm{H,old}} + \sum_i W_{ij}^{\mathrm{old}} v_i^{(n)}/\sigma_i$，$q_{nj} = 1/(1+\exp(-f_{nj}^{\mathrm{H,old}}))$ である．また，このときの補助変数は

$$\bar{\boldsymbol{\Theta}} = \{b_i^{\mathrm{V,old}}, b_j^{\mathrm{H,old}}, W_{ij}^{\mathrm{old}}, \beta_{i0}^{\mathrm{V}}, \beta_{ij}^{\mathrm{V}}, \beta_{0j}^{\mathrm{H}}, \beta_{ij}^{\mathrm{H}}\} \tag{69}$$

となる．

この更新式を見てみると，β_{i0}^{V}，β_{ij}^{V}，β_{0j}^{H}，β_{ij}^{H} は他の補助関数法を用いた学習アルゴリズムと同様に学習率を担っていると考えられるため，これらについても γ 乗にする近似を考える．

以上より Gaussian-Bernoulli 型 RBM の補助関数法による最大再構築確率学習アルゴリズムは，次の 1)〜3) を反復することである．1) サンプリングにより $h_j^{(n)}$ を求める．2) 補助変数 $\bar{\boldsymbol{\Theta}}$ を求める．3) 式 (66)〜(68) の更新式でパラメータを更新．

4. 収束性能の比較実験

2.，3. 章で説明した各学習アルゴリズムがどのような挙動を示すかについて，まず，Bernoulli-Bernoulli 型 RBM について実験を行った．また，これ以降，提案した補助関数法に基づく最尤学習アルゴリズムを提案法 1，補助関数法に基づく最大再構築確率学習アルゴリズムを提案法 2 と呼ぶこととする．最初に可視層の状態数 $I = 15$，隠れ層の状態数 $J = 10$ という非常に小さな系で実験を行い，次に可視層の状態数 $I = 1000$，隠れ層の状態数 $J = 800$ で実験を行った．小さな状態数で行うのは，尤度を厳密に計算するには状態数に対して指数的な計算量が必要となるため，尤度による収束の比較を行いたいからである．このとき，可視層に入力するバイナリデータは乱数で 100 個生成し，生成したそれぞれに対し，80 %のノイズをかけたものを 100 個づつ用意した．つまり，入力するデータ数は $N = 10000$ となる．また，学習の反復回数 T は 3000 回とし，各パラメータの初期値は $[-1, 1]$ の一様乱数から生成し，すべてのアルゴリズムで共通の初期値とした．

次に，β_k，β_{i0}^{V}，β_{ij}^{V}，β_{0j}^{H}，β_{ij}^{H} は一様，つまり，

$$\beta_k = \frac{1}{I+J+IJ},\ \beta_{i0}^{\mathrm{V}} = \beta_{ij}^{\mathrm{V}} = \frac{1}{1+J},\ \beta_{0j}^{\mathrm{H}} = \beta_{ij}^{\mathrm{H}} = \frac{1}{1+I} \tag{70}$$

とし，CD 法の学習率 ϵ や式 (38) に示す γ のスケジューリングを t 回目の反復のとき

$$\epsilon(t) = \epsilon_{\mathrm{init}} \left(\frac{\epsilon_{\mathrm{end}}}{\epsilon_{\mathrm{init}}} \right)^{\frac{t-1}{T-1}},\ \gamma(t) = \gamma_{\mathrm{init}} \left(\frac{\gamma_{\mathrm{end}}}{\gamma_{\mathrm{init}}} \right)^{\frac{t-1}{T-1}} \tag{71}$$

とした．このとき，$J = 15$ かつ $I = 10$ とした実験のときは $\epsilon_{\mathrm{init}} = 0.1$，$\epsilon_{\mathrm{end}} = 0.01$，提案法 1 に関して $\gamma_{\mathrm{init}} = 0.1$，$\gamma_{\mathrm{init}} = 1$，提案法 2 に関して $\gamma_{\mathrm{init}} = 0.1$，$\gamma_{\mathrm{init}} = 1$ とし，$J = 1000$ かつ $I = 800$ とした実験のときは $\epsilon_{\mathrm{init}} = 0.1$，$\epsilon_{\mathrm{end}} = 0.01$，提案法 1 に関しては $\gamma_{\mathrm{init}} = 0.3$，$\gamma_{\mathrm{init}} = 0.5$，提案法 2 に関しては $\gamma_{\mathrm{init}} = 0.5$，$\gamma_{\mathrm{init}} = 1.0$ とした．また，ギプスサンプリングの回数を 1 回，Newton 法を用いる場合はその反復数を 1 回とした．

MATLAB による実装により計算時間を計ったところ，CD 法と比べ，提案法 1 による計算時間も提案法 2 による計算時間もおおよそ同じくらいとなった．また，各学習アルゴリズムにより式 (13) に示す対数尤度がどのように遷移したかを Fig. 2(a) に示し，以下のように定義した再構築誤差 e_{recosnt} が各学習アルゴリズムによりどのように遷移したのかを Fig. 2(b)，3 に示す．

$$\bar{\boldsymbol{h}}^{(n)} = \arg\max_{\boldsymbol{h}} p\left(\boldsymbol{h}|\boldsymbol{v}^{(n)}, \boldsymbol{\Theta}\right), \tag{72}$$

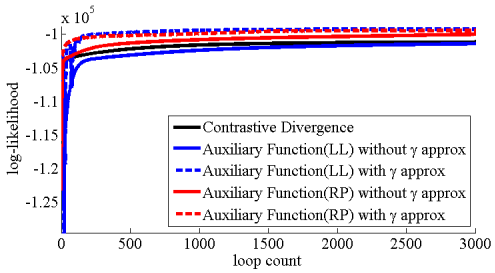
$$\bar{\boldsymbol{v}}^{(n)} = \arg\max_{\boldsymbol{v}} p\left(\boldsymbol{v}|\bar{\boldsymbol{h}}^{(n)}, \boldsymbol{\Theta}\right), \tag{73}$$

$$e_{\mathrm{reconst}} = \frac{1}{NI} \sum_n \sum_i \left(v_i^{(n)} - \bar{v}_i^{(n)} \right). \tag{74}$$

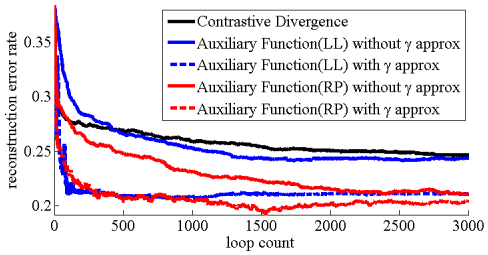
予想したように γ による学習率の近似をしなかったアルゴリズムは収束が遅く，それに対し，γ による学習率の近似をしたアルゴリズムは非常に速く収束するという挙動が観測された．また，γ の加速を行うことにより，CD 法よりも速くなる可能性を示す結果が得られた．

本来，補助関数法は設計パラメータが少ないという利点があるが，このときは γ という設計パラメータが生じてしまう．しかし，補助関数法の原理より，ギプスサンプリングの近似が十分ならば $\gamma = 1$ のときは安定して収束するので，$\gamma = 1$ に近づくようなスケジューリングをすれば良いことから，CD 法の学習率のスケジューリングより設計の指針がはっきりしていると考えられる．

さらに，興味深いことは，本来，目的関数が式 (13) とは異なるものから出発した学習アルゴリズムである最大再構築確率



(a) 対数尤度



(b) 再構築誤差

図 2　Bernoulli-Bernoulli 型 RBM で $J = 15$，$I = 10$ における各学習アルゴリズムの反復毎の対数尤度 (a) と再構築誤差 (b)．黒い実線は CD 法を表し，青，赤の実線と破線はそれぞれ提案手法の γ による学習率の近似をしなかったものとしたものを表す．

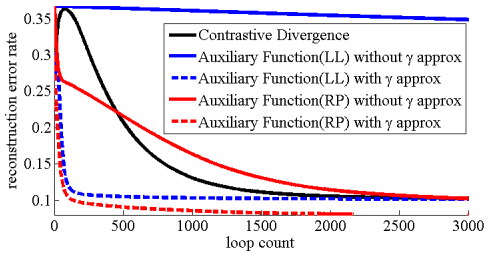


図 3　Bernoulli-Bernoulli 型 RBM で $J = 1000$，$I = 800$ における各学習アルゴリズムの反復毎の再構築誤差．黒い実線は CD 法を表し，青，赤の実線と破線はそれぞれ提案手法の γ による学習率の近似をしなかったものとしたものを表す．

学習アルゴリズムが対数尤度に対して増加傾向にあることである．また，このアルゴリズムは γ による学習の加速を行うと，始めの数反復において他のアルゴリズムより対数尤度の上昇が見られた．

次に Gaussian-Bernoulli 型 RBM について，各学習どのような挙動をするのか，可視層の状態数 $I = 1000$，隠れ層の状態数 $J = 800$ で実験を行った．このとき，可視層に入力するデータは平均 0，標準偏差 100 の正規分布し従う乱数で 50 個生成し，生成したそれぞれに対し，平均 0，標準偏差 0.1 の正規分布に従うノイズを足し合わせたものを 100 個づつ用意した．つまり，入力するデータ数は $N = 10000$ となる．また，σ_i は一様に 1 とし，学習の反復回数 T は 3000 回とし，各パラメータの初期値は $[-1, 1]$ の一様乱数から生成し，すべてのアルゴリズムで共通の初期値とした．また，Bernoulli-Bernoulli 型と同様に任意定数の各 β をそれぞれ一様にし，CD 法の学習率 ϵ や式 (38) に示す γ のスケジューリングを式 (71) に従うよ

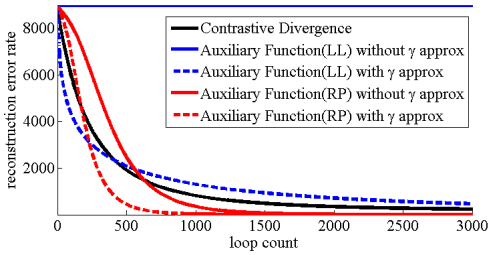


図 4　Gaussian-Bernoulli 型 RBM で $J = 1000$，$I = 800$ における各学習アルゴリズムの反復毎の再構築誤差．黒い実線は CD 法を表し，青，赤の実線と破線はそれぞれ提案手法の γ による学習率の近似をしなかったものとしたものを表す．

うにし，それぞれ，$\epsilon_{\text{init}} = 0.001$，$\epsilon_{\text{end}} = 0.0001$ とし，提案法 1 に関しては $\gamma_{\text{init}} = 0.03$，$\gamma_{\text{init}} = 0.05$ とし，提案法 2 に関しては $\gamma_{\text{init}} = 0.9$，$\gamma_{\text{init}} = 1$ とし，Gibbs サンプリングの回数，Newton 法を用いる場合はその反復数もそれぞれ 1 回とした．

MATLAB による実装により計算時間を計ったところ，CD 法と比べ，提案法 1 による計算時間も提案法 2 による計算時間もおおよそ同じくらいとなった．また，各学習アルゴリズムにより式 (72) に示す再構築誤差がどのように遷移したのかを Fig. 4 に示す．

Bernoulli-Bernoulli 型と同様に，γ による学習率の近似をしなかったアルゴリズムは収束が遅く，それに対し，γ による学習率の近似をしたアルゴリズムは非常に速く収束するという挙動が観測された．また，γ の加速を行うことにより，CD 法よりも速くなる可能性を示す結果が得られた．

5. ま と め

本稿では，Bernoulli-Bernoulli 型 RBM，Gaussian-Bernoulli 型 RBM の学習アルゴリズムとして，それぞれ補助関数法を用いて新たに最尤学習アルゴリズムを導出した．また，それぞれのモデルに対して，新たに最大再構築確率学習を提案し，その学習アルゴリズムを導出した．そして，動作確認実験を通して，既存手法と同等以上の性能を見込めることが確認できた．今後の課題として，実データに用いた場合や多層に重ねて Deep learning を行ったときにどのような挙動を示すかの観察が挙げられる．

謝辞　本研究は JSPS 科研費 26730100 の助成を受けたものです．

文　献

[1] Hinton, G. E., et al. *"A fast learning algorithm for deep belief nets,"* Neural Computation, 2006, 18.7, pp. 1527–1554.
[2] Bengio, Y., et al. *"Greedy layer-wise training of deep networks,"* NIPS, 2007, 19: 153.
[3] Smolensky, P. *"Information processing in dynamical systems: Foundations of harmony theory,"* MIT Press, 1986, pp. 194–281.
[4] Hinton, G. E. *"Training Products of Experts by Minimizing Contrastive Divergence,"* Neural Computation, 2002, 14.8, pp. 1771-1800.
[5] Kameoka, H., et al. *"Complex NMF: A new sparse representation for acoustic signals,"* In Proc. of ICASSP, 2009, p. 3437-3440.