補助関数法によるGaussian-Bernoulli RBMの学習アルゴリズムの検討

高宗 典玄[1,a)]　亀岡 弘和[1,2,b)]

概要：近年，深層学習 (Deep learning) の有効性は音声認識をはじめ様々な分野で示されており，その重要な一要素として，制約付きボルツマンマシン (RBM) による pre-training がある．実数の観測データを取り扱うための Gaussian-Bernoulli RBM というモデルがあり，その学習アルゴリズムとして，最急降下法を基とした Contrastive Divergence 法が提案されてきた．そこで，本発表ではその学習問題に対して，経験的に高速で安定に収束する補助関数法による更新アルゴリズムを提案する．小規模な人工データによる実験を行い，その挙動に対して提案法と従来法を比較し議論する．

1. はじめに

近年，深層学習 (Deep learning) の有効性は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，大量データの下で学習をいかに効率的に行えるかは重要課題の一つである．Deep Neural Network (DNN) の一種である Deep Belief Network (DBN)[1], [2] は制約付きボルツマンマシン (Restricted Boltzmann Machine; RBM)[3] を多層に積み上げたものと見なせ，各層の RBM の教師なし学習を順次行っていくことにより初期学習を行う方式が DBN の学習において効果的であることが知られている．RBM の学習アルゴリズムとして，Contrastive Divergence (CD) 法 [1], [4] が非常に有名であるが，RBM の学習をいかに効率的に行えるかが DBN 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])．そこで，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と考え，Bernoulli-Bernoulli 型の RBM の補助関数法に基づく学習則を導出してきた [6]．

Bernoulli-Bernoulli 型の RBM はバイナリデータしか取り扱えないが，実際の音声や画像といったデータに用いるためには実数を取り扱う必要がある．そこで，本発表では実数のデータに適用するために Gaussian-Bernoulli 型の 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．

1 東京大学大学院情報理工学系研究科
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan
2 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
NTT Communication Science Laboratories
3-1, Morinosato Wakamiya Atsugi-shi, Kanagawa, 243-0198, Japan
a) takamune@hil.t.u-tokyo.ac.jp
b) kameoka@hil.t.u-tokyo.ac.jp

2. Gaussian-Bernoulli 型制約付きボルツマンマシン

2.1 Gaussian-Bernoulli 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．このとき，可視層の状態数を I，可視層の状態を $\boldsymbol{v} = \{v_i\} \in (-\infty, \infty)^I$，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

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

で定義される．ここで，

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

$$Z(\boldsymbol{\Theta}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \sum_{\boldsymbol{h}} \exp(-E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\Theta}))\, \mathrm{d}v_1 \cdots \mathrm{d}v_I \tag{3}$$

であり，$\boldsymbol{\Theta} = \left(b_i^{\mathrm{V}}, b_j^{\mathrm{H}}, W_{ij}\right)$ や (σ_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 (σ_i) は定数として取り扱う．

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

このとき 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．

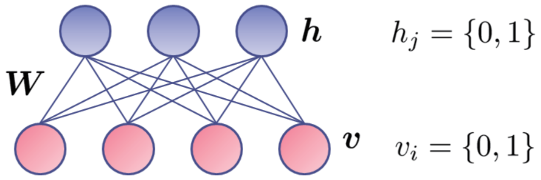

図 1　RBM のグラフ表現

$$\begin{aligned} J(\boldsymbol{\Theta}) &= \frac{1}{N}\sum_n \log p\left(\boldsymbol{v}^{(n)}|\boldsymbol{\Theta}\right) \\ &= \frac{1}{N}\sum_n \log \sum_{\boldsymbol{h}} p\left(\boldsymbol{v}^{(n)}, \boldsymbol{h}|\boldsymbol{\Theta}\right). \end{aligned} \tag{4}$$

2.2　Contrastive Divergence 法 [1], [4]

最急降下法により，式 (4) 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．$J(\boldsymbol{\Theta})$ を Θ に関して微分すると，

$$\begin{aligned} &\frac{\partial J}{\partial \boldsymbol{\Theta}}(\boldsymbol{\Theta}) = \\ &\quad -\frac{1}{N}\sum_n \sum_{\boldsymbol{h}} p\left(\boldsymbol{h}|\boldsymbol{v}^{(n)}, \boldsymbol{\Theta}\right) \frac{\partial E}{\partial \boldsymbol{\Theta}}\left(\boldsymbol{v}^{(n)}, \boldsymbol{h}|\boldsymbol{\Theta}\right) \\ &\quad + \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \sum_{\boldsymbol{h}} p(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\Theta}) \frac{\partial E}{\partial \boldsymbol{\Theta}}(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\Theta})\, \mathrm{d}v_1 \cdots \mathrm{d}v_I \end{aligned} \tag{5}$$

となる．このため，最急降下法によるパラメータの更新は

$$\boldsymbol{\Theta} \leftarrow \boldsymbol{\Theta}^{\mathrm{old}} + \epsilon \Delta_{\mathrm{cd}} \boldsymbol{\Theta} \tag{6}$$

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，$\boldsymbol{h}$ についての和に関しては厳密に計算しようとすると $O(2^J)$ の計算量となるため，現実的ではない．しかし，RBM の特徴として可視層同士，隠れ層同士の依存関係が無いため，

$$p(\boldsymbol{v}|\boldsymbol{h}, \boldsymbol{\Theta}) = \prod_i p(v_i|\boldsymbol{h}, \boldsymbol{\Theta}), \tag{7}$$

$$p(\boldsymbol{h}|\boldsymbol{v}, \boldsymbol{\Theta}) = \prod_j p(h_j|\boldsymbol{v}, \boldsymbol{\Theta}) \tag{8}$$

という関係がある．ただし，

$$\begin{aligned} &p(v_i|\boldsymbol{h}, \boldsymbol{\Theta}) \\ &= \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{\left(v_i - b_i^{\mathrm{V}} - \sigma_i \sum_j W_{ij} h_j\right)^2}{2\sigma_i^2}\right), \end{aligned} \tag{9}$$

$$p(h_j = 1|\boldsymbol{v}, \boldsymbol{\Theta}) = \frac{1}{1 + \exp\left(-b_j^{\mathrm{H}} - \sum_i W_{ij} v_i\right)} \tag{10}$$

である．そこで，式 (5) の第 1 項に関しては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 (5) 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 (5) 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 Gibbs 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．

$$p(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\Theta}) \approx \frac{1}{M}\sum_m \delta\left(\boldsymbol{v} - \boldsymbol{v}^{(m)}\right) p\left(\boldsymbol{h}|\boldsymbol{v}^{(m)}, \boldsymbol{\Theta}\right) \tag{11}$$

ここで，$\boldsymbol{v}^{(m)}$ は Gibbs サンプリングによりサンプリングされた値であり，M はその個数である．本稿では，$M = N$ とし，各 $\boldsymbol{v}^{(m)}$ は対応する $\boldsymbol{v}^{(n)}$ を初期値としてサンプリングされたものを用いる．

ここで，式 (7)，(8) から Gibbs サンプリングは

$$h_j^{d-1} \sim p\left(h_j|\boldsymbol{v}^{d-1}, \boldsymbol{\Theta}\right), \tag{12}$$

$$v_i^d \sim p\left(v_i|\boldsymbol{h}^{d-1}, \boldsymbol{\Theta}\right) \tag{13}$$

と容易に行うことが可能である．

式 (6) による更新を式 (11) の Gibbs サンプリングによる近似で行う手法を CD 法という．

3.　補助関数法による RBM の学習アルゴリズム

3.1　補助関数法 1

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 $\boldsymbol{y}$ を導入して，任意の $\boldsymbol{x}$，$\boldsymbol{y}$ で $F(\boldsymbol{x}) \geq F^{+}(\boldsymbol{x}, \boldsymbol{y})$ となり，$F(\boldsymbol{x}) = \min_{\boldsymbol{y}} F^{+}(\boldsymbol{x}, \boldsymbol{y})$ となるような下限関数 $F^{+}(\boldsymbol{x}, \boldsymbol{y})$ を設計して，$F^{+}(\boldsymbol{x}, \boldsymbol{y})$ を $\boldsymbol{x}$ についての最小化と $\boldsymbol{y}$ 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，$F^{+}(\boldsymbol{x}, \boldsymbol{y})$ を $\boldsymbol{x}$ についての最小化が容易に行え，かつ $F^{+}(\boldsymbol{x}, \boldsymbol{y})$ がが $F(\boldsymbol{x})$ 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 $F^{+}(\boldsymbol{x}, \boldsymbol{y})$ を設計することを目指す．

そこで，式 (4) による目的関数を考えたとき，Jensen の不等式から

$$\begin{aligned} J(\boldsymbol{\Theta}) &= \frac{1}{N}\sum_n \log \sum_{\boldsymbol{h}} p\left(\boldsymbol{v}^{(n)}, \boldsymbol{h}|\boldsymbol{\Theta}\right) \\ &\geq \frac{1}{N}\sum_n \sum_{\boldsymbol{h}} \lambda_{n,\boldsymbol{h}} \log p\left(\boldsymbol{v}^{(n)}, \boldsymbol{h}|\boldsymbol{\Theta}\right) \\ &\quad - \sum_{\boldsymbol{h}} \lambda_{n,\boldsymbol{h}} \log \lambda_{n,\boldsymbol{h}}. \end{aligned} \tag{14}$$

となる．ここで，等号の成立は

$$\lambda_{n,\boldsymbol{h}} = p\left(\boldsymbol{h}|\boldsymbol{v}^{(n)}, \boldsymbol{\Theta}\right) \tag{15}$$

である．式 (14) を整理すると，

$$\begin{aligned} J(\boldsymbol{\Theta}) \geq &-\frac{1}{N}\sum_n \sum_{\boldsymbol{h}} \lambda_{n,\boldsymbol{h}} E\left(\boldsymbol{v}^{(n)}, \boldsymbol{h}|\boldsymbol{\Theta}\right) \\ &- \log Z(\boldsymbol{\Theta}) - \sum_{\boldsymbol{h}} \lambda_{n,\boldsymbol{h}} \log \lambda_{n,\boldsymbol{h}} \end{aligned} \tag{16}$$

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．

$$-\log Z(\boldsymbol{\Theta}) \geq -\frac{Z(\boldsymbol{\Theta})}{\zeta} - \log \zeta + 1. \tag{17}$$

ここで，等号の成立は接点となるので，

$$\zeta = Z(\boldsymbol{\Theta}) \tag{18}$$

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

$$a_k(\boldsymbol{v},\boldsymbol{h}) = \begin{cases} \frac{v_i}{\sigma_i^2} & (k=i) \\ h_j & (k=I+j) \\ \frac{v_i h_j}{\sigma_i} & (k=I+J+J\times(i-1)+j) \end{cases}, \tag{19}$$

$$\theta_k = \begin{cases} b_i^{\mathrm{V}} & (k=i) \\ b_j^{\mathrm{H}} & (k=I+j) \\ W_{ij} & (k=I+J+J\times(i-1)+j) \end{cases} \tag{20}$$

とおくと，式 (2) は

$$E(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta}) = \sum_i \frac{v_i^2}{2\sigma_i^2} - \sum_k a_k(\boldsymbol{v},\boldsymbol{h})\,\theta_k \tag{21}$$

と表すことができる．

ここで，$-\exp\left(\sum_k a_k(\boldsymbol{v},\boldsymbol{h})\,\theta_k\right)$ は凹関数であるので，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，

$$\begin{aligned} &-\exp\left(\sum_k a_k(\boldsymbol{v},\boldsymbol{h})\,\theta_k\right) \\ &\geq -\sum_k \beta_k(\boldsymbol{v},\boldsymbol{h}) \\ &\times \exp\left(\frac{a_k(\boldsymbol{v},\boldsymbol{h})\,\theta_k - \alpha_k(\boldsymbol{v},\boldsymbol{h})}{\beta_k(\boldsymbol{v},\boldsymbol{h})}\right) \end{aligned} \tag{22}$$

となり，等号の成立は

$$\begin{aligned} &{}^\forall\beta_k(\boldsymbol{v},\boldsymbol{h}) \in [0,1], \\ &\sum_k \beta_k(\boldsymbol{v},\boldsymbol{h}) = 1, \end{aligned} \tag{23}$$

$$\begin{aligned} \alpha_k(\boldsymbol{v},\boldsymbol{h}) =& a_k(\boldsymbol{v},\boldsymbol{h})\,\theta_k \\ &- \beta_k(\boldsymbol{v},\boldsymbol{h}) \sum_l a_l(\boldsymbol{v},\boldsymbol{h})\,\theta_l \end{aligned} \tag{24}$$

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 とできる．更に，新たに補助変数として一反復前の値を表す θ_k^{old} を導入し，補助変数の更新式 (24) を式 (22) に代入し，式 (17) を式 (3), (21) 用いて整理すると，

$$\begin{aligned} &-\log Z(\boldsymbol{\Theta}) \\ &\geq -\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\sum_{\boldsymbol{h}} \frac{\exp\left(-\sum_i \frac{v_i^2}{2\sigma_i^2} + \sum_l a_l(\boldsymbol{v},\boldsymbol{h})\,\theta_l^{\mathrm{old}}\right)}{\zeta} \\ &\times \sum_k \beta_k \exp\left(\frac{a_k(\boldsymbol{v},\boldsymbol{h})\left(\theta_k-\theta_k^{\mathrm{old}}\right)}{\beta_k}\right) \mathrm{d}v_1\cdots\mathrm{d}v_I \\ &-\log\zeta + 1 \end{aligned} \tag{25}$$

となる．ここで，式 (1)，(18) から式 (25) は

$$\begin{aligned} &-\log Z(\boldsymbol{\Theta}) \\ &\quad\geq -\sum_k \beta_k \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\sum_{\boldsymbol{h}} p\left(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta}^{\mathrm{old}}\right) \\ &\quad\times \exp\left(\frac{a_k(\boldsymbol{v},\boldsymbol{h})\left(\theta_k-\theta_k^{\mathrm{old}}\right)}{\beta_k}\right) \mathrm{d}v_1\cdots\mathrm{d}v_I \\ &\quad-\log\zeta + 1 \end{aligned} \tag{26}$$

となる．式 (15)，(16)，(18)，(26) より，$\bar{\boldsymbol{\Theta}} = \left(\boldsymbol{\Theta}^{\mathrm{old}}, \beta_k\right)$ とおいたとき，$J(\boldsymbol{\Theta})$ の下限関数 $J^+\left(\boldsymbol{\Theta},\bar{\boldsymbol{\Theta}}\right)$ は

$$\begin{aligned} &J^+\left(\boldsymbol{\Theta},\bar{\boldsymbol{\Theta}}\right) \\ &= \frac{1}{N}\sum_n \sum_{\boldsymbol{h}} p\left(\boldsymbol{h}|\boldsymbol{v}^{(n)},\boldsymbol{\Theta}^{\mathrm{old}}\right) \\ &\qquad\times \sum_k a_k\left(\boldsymbol{v}^{(n)},\boldsymbol{h}\right)\theta_k \\ &\quad -\sum_k \beta_k \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\sum_{\boldsymbol{h}} p\left(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta}^{\mathrm{old}}\right) \\ &\qquad\times \exp\left(\frac{a_k(\boldsymbol{v},\boldsymbol{h})\left(\theta_k-\theta_k^{\mathrm{old}}\right)}{\beta_k}\right) \mathrm{d}v_1\cdots\mathrm{d}v_I \\ &\qquad + C\left(\bar{\boldsymbol{\Theta}}\right) \end{aligned} \tag{27}$$

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

ここで，式 (27) を θ_k について微分すると，

$$\begin{aligned} &\frac{\partial J^+}{\partial\theta_k}\left(\boldsymbol{\Theta},\bar{\boldsymbol{\Theta}}\right) \\ &\quad= \frac{1}{N}\sum_n \sum_{\boldsymbol{h}} p\left(\boldsymbol{h}|\boldsymbol{v}^{(n)},\boldsymbol{\Theta}^{\mathrm{old}}\right) a_k\left(\boldsymbol{v}^{(n)},\boldsymbol{h}\right) \\ &\qquad -\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\sum_{\boldsymbol{h}} p\left(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta}^{\mathrm{old}}\right) a_k(\boldsymbol{v},\boldsymbol{h}) \\ &\qquad\quad \times \exp\left(\frac{a_k(\boldsymbol{v},\boldsymbol{h})\left(\theta_k-\theta_k^{\mathrm{old}}\right)}{\beta_k}\right) \mathrm{d}v_1\cdots\mathrm{d}v_I \end{aligned} \tag{28}$$

となる．ここで，式 (28) 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 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 (11) で近似すると式 (28) は

$$\begin{aligned} &\frac{\partial J^+}{\partial\theta_k}\left(\boldsymbol{\Theta},\bar{\boldsymbol{\Theta}}\right) \\ &\quad= \frac{1}{N}\sum_n \sum_{\boldsymbol{h}} p\left(\boldsymbol{h}|\boldsymbol{v}^{(n)},\boldsymbol{\Theta}^{\mathrm{old}}\right) a_k\left(\boldsymbol{v}^{(n)},\boldsymbol{h}\right) \\ &\qquad -\frac{1}{M}\sum_m \sum_{\boldsymbol{h}} p\left(\boldsymbol{h}|\boldsymbol{v}^{(m)},\boldsymbol{\Theta}^{\mathrm{old}}\right) a_k\left(\boldsymbol{v}^{(m)},\boldsymbol{h}\right) \\ &\qquad\quad \times \exp\left(\frac{a_k\left(\boldsymbol{v}^{(m)},\boldsymbol{h}\right)\left(\theta_k-\theta_k^{\mathrm{old}}\right)}{\beta_k}\right) \end{aligned} \tag{29}$$

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 [6] 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，$\partial^2 J^+/\partial\theta_k^2$ 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，[6] と同型の補助関数を用いているので，収束を速くするために以下の近似を考える．

$$\beta_k' \leftarrow \beta_k^{\gamma}. \tag{30}$$

ただし，β_k' は式 (23) 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．

以上より補助関数法による 1 つ目の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 $p\left(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta}^{\mathrm{old}}\right)$ 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 補助関数法 2

次に，補助関数法を用いた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．ここで用いる目的関数は可視層に観測データが来たときに，Gibbs 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尤度としたもので,

$$J_r\left(\boldsymbol{\Theta}\right)=\frac{1}{N}\sum_n\log\sum_{\boldsymbol{h}}p\left(\boldsymbol{v}^{(n)}|\boldsymbol{h},\boldsymbol{\Theta}\right)p\left(\boldsymbol{h}|\boldsymbol{v}^{(n)},\boldsymbol{\Theta}\right) \tag{31}$$

と表される．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,

$$\boldsymbol{h}^{(n)}\sim p\left(\boldsymbol{h}|\boldsymbol{v}^{(n)},\boldsymbol{\Theta}\right) \tag{32}$$

でサンプリングした値を元に

$$\begin{aligned}&J_r\left(\boldsymbol{\Theta}\right)\\&\approx\frac{1}{N}\sum_n\log p\left(\boldsymbol{v}^{(n)}|\boldsymbol{h}^{(n)},\boldsymbol{\Theta}\right)p\left(\boldsymbol{h}^{(n)}|\boldsymbol{v}^{(n)},\boldsymbol{\Theta}\right)\end{aligned} \tag{33}$$

と近似することを考える．この右辺を $\tilde{J}_r\left(\boldsymbol{\Theta}\right)$ とおくと，式 (7)∼(10) より,

$$\begin{aligned}\tilde{J}_r\left(\boldsymbol{\Theta}\right)=-\frac{1}{N}\sum_n\Bigg\{&\sum_i\frac{\left(v_i^{(n)}-f_{ni}^{\mathrm{V}}\right)^2}{2\sigma_i^2}\\&+\sum_i\log\sigma_i+\frac{I}{2}\log\left(2\pi\right)\\&-\sum_j\left(h_j^{(n)}-\frac{1}{2}\right)f_{nj}^{\mathrm{H}}\\&+\sum_j\log\left(\exp\left(f_{nj}^{\mathrm{H}}/2\right)+\exp\left(-f_{nj}^{\mathrm{H}}/2\right)\right)\Bigg\}\end{aligned} \tag{34}$$

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

ここで,

$$\begin{aligned}&\log\left(\exp\left(x\right)+\exp\left(-x\right)\right)\\&\quad\leq\frac{\tanh\left(\kappa\right)}{2\kappa}x^2-\frac{\kappa\tanh\left(\kappa\right)}{2}+\log\left(2\cosh\left(\kappa\right)\right)\end{aligned} \tag{35}$$

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

$$\kappa=\pm x \tag{36}$$

である．この不等式を用いると式 (34) は

$$\begin{aligned}\tilde{J}_r\left(\boldsymbol{\Theta}\right)\geq-\frac{1}{N}\sum_n\Bigg\{&\sum_i\frac{\left(f_{ni}^{\mathrm{V}}\right)^2}{2\sigma_i^2}\\&+\sum_j\frac{\tanh\left(\kappa_{nj}\right)}{8\kappa_{nj}}\left(f_{nj}^{\mathrm{H}}\right)^2\\&-\sum_i\frac{v_i^{(n)}}{\sigma_i^2}f_{ni}^{\mathrm{V}}\\&-\sum_j\left(h_j^{(n)}-\frac{1}{2}\right)f_{nj}^{\mathrm{H}}\\&+\sum_i\left(\frac{\left(v_i^{(n)}\right)^2}{2\sigma_i^2}+\log\sigma_i\right)+\frac{I}{2}\log\left(2\pi\right)\\&+\sum_j\left(\log\left(\cosh\left(\kappa_{nj}\right)\right)-\frac{\kappa_{nj}\tanh\left(\kappa_{nj}\right)}{8}\right)\Bigg\}\end{aligned} \tag{37}$$

と下から押さえられる．

ここで，$f_{ni}^{\mathrm{V}}=b_i^{\mathrm{V}}+\sigma_i\sum_j W_{ij}h_j^{(n)}$，$f_{nj}^{\mathrm{H}}=b_j^{\mathrm{H}}+\sum_i W_{ij}v_i^{(n)}/\sigma_i$ であるので，$-\left(f_{ni}^{\mathrm{V}}\right)^2$ と $-\left(f_{nj}^{\mathrm{H}}\right)$ 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

$$\begin{aligned}-\left(f_{ni}^{\mathrm{V}}\right)^2\geq&-\frac{\left(b_i^{\mathrm{V}}-\alpha_{ni0}^{\mathrm{V}}\right)^2}{\beta_{ni0}^{\mathrm{V}}}\\&-\sum_j\frac{\left(\sigma_iW_{ij}h_j^{(n)}-\alpha_{nij}^{\mathrm{V}}\right)^2}{\beta_{nij}^{\mathrm{V}}},\end{aligned} \tag{38}$$

$$\begin{aligned}-\left(f_{nj}^{\mathrm{H}}\right)^2\geq&-\frac{\left(b_j^{\mathrm{H}}-\alpha_{n0j}^{\mathrm{H}}\right)^2}{\beta_{n0j}^{\mathrm{H}}}\\&-\sum_i\frac{\left(W_{ij}v_i^{(n)}/\sigma_i-\alpha_{nij}^{\mathrm{H}}\right)^2}{\beta_{nij}^{\mathrm{H}}}\end{aligned} \tag{39}$$

となり，等号の成立は

$$\begin{aligned}&{}^{\forall}\beta_{ni0}^{\mathrm{V}},\beta_{nij}^{\mathrm{V}}\in[0,1],\\&\beta_{ni0}^{\mathrm{V}}+\sum_j\beta_{nij}^{\mathrm{V}}=1,\end{aligned} \tag{40}$$

$$\begin{aligned}&{}^{\forall}\beta_{n0j}^{\mathrm{H}},\beta_{nij}^{\mathrm{H}}\in[0,1],\\&\beta_{n0j}^{\mathrm{H}}+\sum_i\beta_{nij}^{\mathrm{H}}=1,\end{aligned} \tag{41}$$

$$\begin{aligned}\alpha_{ni0}^{\mathrm{V}}=&b_i^{\mathrm{V}}-\beta_{ni0}^{\mathrm{V}}f_{ni}^{\mathrm{V}},\\\alpha_{nij}^{\mathrm{V}}=&\sigma_iW_{ij}h_j^{(n)}-\beta_{nij}^{\mathrm{V}}f_{ni}^{\mathrm{V}},\end{aligned} \tag{42}$$

$$\begin{aligned}\alpha_{n0i}^{\mathrm{H}}=&b_j^{\mathrm{H}}-\beta_{ni0}^{\mathrm{H}}f_{nj}^{\mathrm{H}},\\\alpha_{nij}^{\mathrm{H}}=&\frac{W_{ij}v_i^{(n)}}{\sigma_i}-\beta_{nij}^{\mathrm{H}}f_{nj}^{\mathrm{H}}\end{aligned} \tag{43}$$

である．

よって, 式 (37), 式 (38), 式 (39) から下限関数 $\tilde{J}_r^{+}\left(\boldsymbol{\Theta},\bar{\boldsymbol{\Theta}}\right)$ は

$$\begin{aligned}\tilde{J}_r^+\left(\mathbf{\Theta},\bar{\mathbf{\Theta}}\right)=&-\frac{1}{N}\sum_n\left\{\sum_i\frac{\left(b_i^{\mathrm{V}}-\alpha_{ni0}^{\mathrm{V}}\right)^2}{2\beta_{ni0}^{\mathrm{V}}\sigma_i^2}\right.\\&+\sum_i\sum_j\frac{\left(\sigma_iW_{ij}h_j^{(n)}-\alpha_{nij}^{\mathrm{V}}\right)^2}{2\beta_{nij}^{\mathrm{V}}\sigma_i^2}\\&+\sum_j\frac{\tanh\left(\kappa_{nj}\right)}{8\kappa_{nj}\beta_{n0j}^{\mathrm{H}}}\left(b_j^{\mathrm{H}}-\alpha_{n0j}^{\mathrm{H}}\right)^2\\&+\sum_j\sum_i\frac{\tanh\left(\kappa_{nj}\right)}{8\kappa_{nj}\beta_{nij}^{\mathrm{H}}}\left(W_{ij}v_i^{(n)}/\sigma_i-\alpha_{nij}^{\mathrm{H}}\right)^2\\&-\sum_i\frac{v_i^{(n)}}{\sigma_i^2}\left(b_i^{\mathrm{V}}+\sigma_i\sum_jW_{ij}h_j^{(n)}\right)\\&\left.-\sum_j\left(h_j^{(n)}-\frac{1}{2}\right)\left(b_j^{\mathrm{H}}+\sum_iW_{ij}v_i^{(n)}/\sigma_i\right)\right\}\\&+C\left(\bar{\mathbf{\Theta}}\right)\end{aligned}\tag{44}$$

となる．ただし，

$$\bar{\mathbf{\Theta}}=\left(\kappa_{nj},\alpha_{ni0}^{\mathrm{V}},\alpha_{nij}^{\mathrm{V}},\alpha_{n0j}^{\mathrm{H}},\alpha_{nij}^{\mathrm{H}},\beta_{ni0}^{\mathrm{V}},\beta_{nij}^{\mathrm{V}},\beta_{n0j}^{\mathrm{H}},\beta_{nij}^{\mathrm{H}}\right)\tag{45}$$

であり，$C\left(\bar{\mathbf{\Theta}}\right)$ は $\mathbf{\Theta}$ 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 $\partial\tilde{J}_r^+/\partial\mathbf{\Theta}=0$ より

$$b_i^{\mathrm{V}}\leftarrow\frac{\sum_n\left(\alpha_{ni0}^{\mathrm{V}}/\beta_{ni0}^{\mathrm{V}}+v_i^{(n)}\right)}{\sum_n1/\beta_{ni0}^{\mathrm{V}}},\tag{46}$$

$$b_j^{\mathrm{H}}\leftarrow\frac{\sum_n\left(\tanh\left(\kappa_{nj}\right)\alpha_{n0j}^{\mathrm{H}}/4\kappa_{nj}\beta_{n0j}^{\mathrm{H}}+h_j^{(n)}-1/2\right)}{\sum_n\tanh\left(\kappa_{nj}\right)/4\kappa_{nj}\beta_{n0j}^{\mathrm{H}}},\tag{47}$$

$$W_{ij}\leftarrow\frac{\sum_n\left(\frac{\alpha_{nij}^{\mathrm{V}}}{\beta_{nij}^{\mathrm{V}}\sigma_i}+\frac{\tanh(\kappa_{nj})\alpha_{nij}^{\mathrm{H}}}{4\kappa_{nj}\beta_{nij}^{\mathrm{H}}}+\frac{v_i^{(n)}\left(2h_j^{(n)}-1/2\right)}{\sigma_i}\right)}{\sum_n\left(\frac{h_j^{(n)}}{\beta_{nij}^{\mathrm{V}}\sigma_i}+\frac{\tanh(\kappa_{nj})v_i^{(n)}}{4\kappa_{nj}\beta_{nij}^{\mathrm{H}}\sigma_i}\right)}\tag{48}$$

となる．ここで，β_{ni0}^{V}，β_{nij}^{V}，β_{n0j}^{H}，β_{nij}^{H} は式 (40)，(41) 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 (36)，(42)，(43) を代入して整理すると

$$b_i^{\mathrm{V}}\leftarrow b_i^{\mathrm{V,old}}+\frac{\beta_{i0}^{\mathrm{V}}}{N}\sum_n\left(v_i^{(n)}-f_{ni}^{\mathrm{V,old}}\right),\tag{49}$$

$$b_j^{\mathrm{H}}\leftarrow b_j^{\mathrm{H,old}}+\beta_{0j}^{\mathrm{H}}\frac{\sum_n\left(h_j^{(n)}-q_{nj}\right)}{\sum_n\left(q_{nj}-1/2\right)/f_{nj}^{\mathrm{H,old}}},\tag{50}$$

$$\begin{aligned}W_{ij}\leftarrow&W_{ij}^{\mathrm{old}}\\&+\frac{\sum_n\left(\left(v_i^{(n)}-f_{ni}^{\mathrm{V,old}}\right)h_j^{(n)}+v_i^{(n)}\left(h_j^{(n)}-q_{nj}\right)\right)}{\sum_n\left(\frac{h_j^{(n)}\sigma_i}{\beta_{nij}^{\mathrm{V}}}+\frac{\left(v_i^{(n)}\right)^2(q_{nj}-1/2)}{\beta_{nij}^{\mathrm{H}}\sigma_if_{nj}^{\mathrm{H,old}}}\right)}\end{aligned}\tag{51}$$

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

$$\bar{\mathbf{\Theta}}=\left(b_i^{\mathrm{V,old}},b_j^{\mathrm{H,old}},W_{ij}^{\mathrm{old}},\beta_{ni0}^{\mathrm{V}},\beta_{nij}^{\mathrm{V}},\beta_{n0j}^{\mathrm{H}},\beta_{nij}^{\mathrm{H}}\right)\tag{52}$$

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 1 つ目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．

以上より補助関数法による 3 つ目の学習アルゴリズムは，次の 1) 3) を反復することとなる．1) サンプリングにより $h_j^{(n)}$ を求める．2) 補助変数 $\bar{\mathbf{\Theta}}$ を求める．3) 式 (49)〜(51) の更新式でパラメータを更新．

4. 動作確認実験

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，可視層の状態数 $I=20$，隠れ層の状態数 $J=15$ という非常に小さな系で実験を行った．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 $N=5000$ となる．また，σ_i は一様に 1 とし，学習の反復回数 T は 20000 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，

$$\beta_k=\frac{1}{I+J+IJ},\tag{53}$$

$$\beta_{ni0}^{\mathrm{V}}=\beta_{nij}^{\mathrm{V}}=\frac{1}{1+J},\tag{54}$$

$$\beta_{n0j}^{\mathrm{H}}=\beta_{nij}^{\mathrm{H}}=\frac{1}{1+I}\tag{55}$$

とし，CD 法の学習率 ϵ や式 (30) に示す γ のスケジューリングを t 回目の反復のとき

$$\epsilon\left(t\right)=\epsilon_{\mathrm{init}}\left(\frac{\epsilon_{\mathrm{end}}}{\epsilon_{\mathrm{init}}}\right)^{\frac{t-1}{T-1}},\tag{56}$$

$$\gamma\left(t\right)=\gamma_{\mathrm{init}}\left(\frac{\gamma_{\mathrm{end}}}{\gamma_{\mathrm{init}}}\right)^{\frac{t-1}{T-1}}\tag{57}$$

とした．このとき，$\epsilon_{\mathrm{init}}=0.001$，$\epsilon_{\mathrm{end}}=0.0001$ とし，補助関数法 1 に関しては $\gamma_{\mathrm{init}}=0.03$，$\gamma_{\mathrm{init}}=0.01$ とし，補助関数法 2 に関しては $\gamma_{\mathrm{init}}=1$，$\gamma_{\mathrm{init}}=1$ とした．また，Gibbs 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，補助関数法 1 による計算時間も補助関数法 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 (4) に示す対数尤度がどのように遷移したかを Fig. 2 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. 3 に示す．

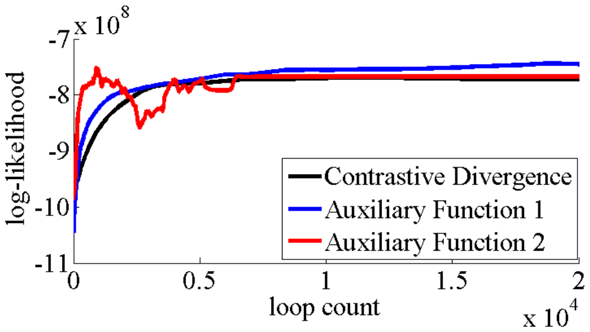

図 2 各学習アルゴリズムにおける反復毎の対数尤度．黒い実線はCD 法を表し，青，赤の実線と破線はそれぞれ提案手法を表す．

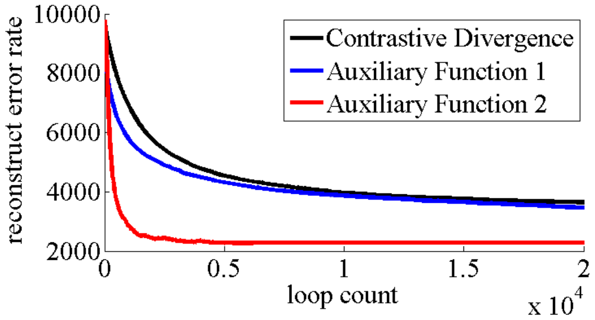

図 3 各学習アルゴリズムにおける反復毎の再構築誤差．黒い実線は CD 法を表し，青，赤の実線と破線はそれぞれ提案手法を表す．

$$\bar{\boldsymbol{h}}^{(n)} = \arg\max_{\boldsymbol{h}} p\left(\boldsymbol{h}|\boldsymbol{v}^{(n)}, \boldsymbol{\Theta}\right), \tag{58}$$

$$\bar{\boldsymbol{v}}^{(n)} = \arg\max_{\boldsymbol{v}} p\left(\boldsymbol{v}|\bar{\boldsymbol{h}}^{(n)}, \boldsymbol{\Theta}\right), \tag{59}$$

$$e_{\text{reconst}} = \frac{1}{NI}\sum_{n}\sum_{i}\left(v_i^{(n)} - \bar{v}_i^{(n)}\right). \tag{60}$$

対数尤度による比較をみると，補助関数法 1 は CD 法よりも速く尤度が増大するが，補助関数法 2 は目的関数が対数尤度でないことが起因しているのか初期の反復において尤度の上下が激しくなるといった結果になった．また，再構築誤差による比較を見ると，補助関数法 1 は CD 法よりも若干速く誤差は小さくなり，補助関数法 2 にいたっては大幅に速く誤差は小さくなり，収束先も誤差がより小さいところに収束しそうな結果が得られた．

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，Gibbs 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．

5. まとめ

本稿では，Gaussian-Bernoulli 型の RBM の学習アルゴリズムとして，補助関数法を用いて新たに 2 つの学習アルゴリズムを導出した．そして，小規模の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，可視層と隠れ層の状態数が多くなったときや多層に重ねて深層学習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 議論に参加してくださった東大院・情報理工の石原達馬氏に深く感謝する．また，本研究は JSPS 科研費 26730100 の助成を受けたものです．

参考文献

[1] Hinton, G. E., Osindero, S. and Teh, Y. W.: *"A fast learning algorithm for deep belief nets,"* Neural Computation, 2006, 18.7, pp. 1527–1554.

[2] Bengio, Y., Lamblin, P., Popovici, D. and Larochelle, H.: *"Greedy layer-wise training of deep networks,"* Advances in neural information processing systems, 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., Ono, N., Kashino, K. and Sagayama, S.: *"Complex NMF: A new sparse representation for acoustic signals,"* Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, 2009, p. 3437-3440.

[6] 高宗 典玄，石原 達馬，亀岡 弘和: "補助関数法による制約付きボルツマンマシンの学習アルゴリズムの検討，" 日本音響学会 2014 年春季研究発表会講演論文集，2014，No. 1-5-4.