情報論的学習理論テクニカルレポート 2009
Technical Report on Information-Based Induction Sciences 2009 (IBIS2009)

Kullback-Leibler Importance Estimation Procedureを用いたRestricted Boltzmann Machineの学習アルゴリズム
Learning Algorithm in Restricted Boltzmann Machines using Kullback-Leibler Importance Estimation Procedure

桜井　哲治* 安田　宗樹 * 田中　和之 *
Tetsuharu Sakurai Muneki Yasuda Kazuyuki Tanaka

Abstract: Deep Belief Networks (DBN) are generative neural network models with many layers of hidden units which were recently introduced along with a greedy layer-wise learning algorithm by Hinton et al. The main building block of a DBN is a bipartite undirected graphical model called "restricted Boltzmann machine ". In the present paper, we propose a new and less greedy learning algorithm for restricted Boltzmann machines within DBNs using Kullback-Leibler Importance Estimation Procedure. We also show its validity by comparing our proposed algorithm with the exactly calculated $\mathrm{KL}(P_D||P_V^D)$ learning algorithm using numerical experiments based on artificial data.

Keywords: deep belief network, variational bound, restricted Boltzmann machine, learning algorithm, Kullback-Leibler Importance Estimation Procedure

1　はじめに

Deep Belief Network (DBN) は Hinton らによって，その学習アルゴリズムとともに提案された階層構造をもつベイジアンネットワークモデルである [1]．Greedy な学習アルゴリズムの存在やそこでの推論が容易であることを背景に次元圧縮器やパータン認識問題などへの応用が期待されている [2]．

DBN の学習は各層間の結合確率を Restricted Boltzmann Machine (RBM)[3] とよばれる特別な構造をもつボルツマンマシンと考え，RBM の学習アルゴリズムを用いて各層間のパラメータを学習する．RBM は可視素子層と隠れ素子層の 2 層からなる 2 部グラフの構造をもっており，各層内での結合はない．DBN の学習はその構成要素である RBM の学習を逐次的に進めていくことによりおこなわれる [1, 4]．したがって DBN のよりよい学習を得るためには各層をなしている RBM のよりよい学習アルゴリズムを設計する必要がある．

しかしながら RBM の厳密な学習アルゴリズムは素子数に対して指数的に増加する計算量をもっているため一般には NP-hard のクラスに属し，計算が困難である．そこで実装においては何らかの近似的手段に頼らざるを得ないこととなる．Contrastive divergence 法 [5, 6] とよばれるギブスサンプリング法を基礎とした確率近似アルゴリズムが RBM の近似学習アルゴリズムとして近年よく知られた方法の一つである．

Roux and Bengio は DBN の学習に対して有効な方法（Variational Bound の最適化法）として従来とは異なる RBM の学習基準を提案した [7]．彼等の基準による学習アルゴリズムを用いると 3 層の DBN の場合，システムが観測データ点の経験分布を十分に表現し得るものであるならば，Greedy 学習の際にこの基準による RBM 学習を用いることで最適な DBN を学習できることが示されている [7]．しかしながら彼らの学習基準による RBM の学習アルゴリズムもやはり計算コストの問題を抱えており，それを解決する近似アルゴリズムの開発が必要とされている．

本論文では Sugiyama 等 [8] によって提案された Kullback-Leibler Importance Estimation Procedure (KLIEP) と呼ばれる近似手法を Roux and Bengio[7] の学習基準による学習アルゴリズムに適用することにより，実装に耐え得る性能のよい新しい RBM の近似学習アルゴリズム

*東北大学大学院情報科学研究科，〒 980-8573 仙台市青葉区荒巻字青葉 6-3-09，E-mail { tsakurai,muneki,kazu }@smapip.is.tohoku.ac.jp,
Graduate School of Information Sciences, Tohoku University.

を提案する．

第 2 節では Deep Belief Network に対する Greedy 学習アルゴリズムの概要について要約して説明する．第 3 節では KLIEP にもとづく RBM の新しい近似学習アルゴリズムを提案する．第 4 節では提案した手法における人工データに対する数値実験を行い，第 5 節ではまとめをあたえる．

2 Deep Belief Network

本節では DBN に対する Greedy 学習アルゴリズムとその基本単位となる RBM について概説する．

2.1 Restricted Boltzmann Machine

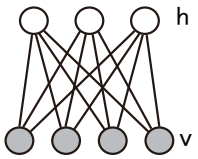

図 1: RBM の例．上段が隠れ素子の層であり，下段が可視素子の層である（$|V|=4, |H|=3$）．

RBM は可視素子の層と隠れ素子の層からなる無向 2 部グラフで表現されるボルツマンマシンであり，各層内での結合はない．可視素子はデータの入出力をおこなう素子であり，隠れ素子はモデルの内部自由度を増加させる．各素子は $\{0,1\}$ の値を確率的にとる．

可視素子 $\mathbf{v}=\{0,1\}^{|V|}$，隠れ素子 $\mathbf{h}=\{0,1\}^{|H|}$ からなる RBM を考える．ここで $V=\{1,\cdots,|V|\}$ を可視素子のラベルの集合，$H=\{1,\cdots,|H|\}$ を隠れ素子のラベルの集合としている．このとき，RBM の状態は次のボルツマン分布であらわされる．

$$P_{\mathrm{RBM}}(\mathbf{v},\mathbf{h}\mid\mathbf{\Theta})=\frac{1}{Z_{\mathrm{RBM}}(\mathbf{\Theta})}\exp\Big(-E(\mathbf{v},\mathbf{h}\mid\mathbf{\Theta})\Big) \tag{1}$$

ここで $Z_{\mathrm{RBM}}(\mathbf{\Theta})$ は規格化定数であり，

$$E(\mathbf{v},\mathbf{h}|\mathbf{\Theta})=-\sum_{i\in V}a_iv_i-\sum_{j\in H}b_jh_j-\sum_{i\in V}\sum_{j\in H}w_{ij}v_ih_j \tag{2}$$

と定義されている．$\mathbf{a}=\{a_i\mid i\in V\}$，$\mathbf{b}=\{b_j\mid j\in H\}$ はそれぞれ可視素子，隠れ素子のバイアスパラメータであり，$\mathbf{w}=\{w_{ij}\mid i\in V, j\in H\}$ は可視素子と隠れ素子との間の結合パラメータである．表記の簡単のためパラメータの集合を $\mathbf{\Theta}=\{\mathbf{a},\mathbf{b},\mathbf{w}\}$ であらわす．

可視素子 $\mathbf{v}$，隠れ素子 $\mathbf{h}$ に関する周辺分布はそれぞれ以下のようになる．

$$\begin{aligned}P_V(\mathbf{v}\mid\mathbf{\Theta})&\equiv\sum_{\mathbf{h}}P_{\mathrm{RBM}}(\mathbf{v},\mathbf{h}\mid\mathbf{\Theta})\\&=\frac{G_V(\mathbf{v},\mathbf{b},\mathbf{w})}{Z_{\mathrm{RBM}}(\mathbf{\Theta})}\exp\left(\sum_{i\in V}a_iv_i\right)\end{aligned} \tag{3}$$

$$\begin{aligned}P_H(\mathbf{h}\mid\mathbf{\Theta})&\equiv\sum_{\mathbf{v}}P_{\mathrm{RBM}}(\mathbf{v},\mathbf{h}\mid\mathbf{\Theta})\\&=\frac{G_H(\mathbf{h},\mathbf{a},\mathbf{w})}{Z_{\mathrm{RBM}}(\mathbf{\Theta})}\exp\left(\sum_{j\in H}b_jh_j\right)\end{aligned} \tag{4}$$

ただし

$$G_V(\mathbf{v},\mathbf{b},\mathbf{w})\equiv\prod_{j\in H}\left\{1+\exp\left(b_j+\sum_{i\in V}w_{ij}v_i\right)\right\} \tag{5}$$

$$G_H(\mathbf{h},\mathbf{a},\mathbf{w})\equiv\prod_{i\in V}\left\{1+\exp\left(a_i+\sum_{j\in H}w_{ij}h_j\right)\right\} \tag{6}$$

としている．可視素子で条件付けられた隠れ素子の条件付き確率 $P_{H|V}(\mathbf{h}\mid\mathbf{v},\mathbf{b},\mathbf{w})$ はベイズの定理より

$$P_{H|V}(\mathbf{h}\mid\mathbf{v},\mathbf{b},\mathbf{w})=\prod_{j\in H}\frac{\exp\left\{\left(b_j+\sum_{i\in V}w_{ij}v_i\right)h_j\right\}}{1+\exp\left(b_j+\sum_{i\in V}w_{ij}v_i\right)} \tag{7}$$

となる．同様にして隠れ素子で条件付けられた可視素子の条件付き確率 $P_{V|H}(\mathbf{v}\mid\mathbf{h},\mathbf{a},\mathbf{w})$ は

$$P_{V|H}(\mathbf{v}\mid\mathbf{h},\mathbf{a},\mathbf{w})=\prod_{i\in V}\frac{\exp\left\{\left(a_i+\sum_{j\in H}w_{ij}h_j\right)v_i\right\}}{1+\exp\left(a_i+\sum_{j\in H}w_{ij}h_j\right)} \tag{8}$$

となる．したがって $\mathbf{v}$ と $\mathbf{h}$ は互いに条件付き独立である．

M 個の観測データ点 $\{\mathbf{d}^\mu\in\{0,1\}^{|V|}\mid\mu=1,\cdots,M\}$ を得たとすると，観測データの経験分布 $P_D(\mathbf{v})$ は

$$P_D(\mathbf{v})\equiv\frac{1}{M}\sum_{\mu=1}^{M}\prod_{i\in V}\delta(v_i,d_i^\mu) \tag{9}$$

で定められる．ここで $\delta(x,y)$ はクロネッガーのデルタ関数である．この経験分布に対して RBM の学習は Kullback-Leibler (KL) 情報量

$$\mathrm{KL}(P_D||P_V)=\sum_{\mathbf{v}}P_D(\mathbf{v})\ln\frac{P_D(\mathbf{v})}{P_V(\mathbf{v}\mid\mathbf{\Theta})} \tag{10}$$

を用いて次のように表現される．

$$\widehat{\mathbf{\Theta}}=\arg\min_{\mathbf{\Theta}}\mathrm{KL}(P_D||P_V) \tag{11}$$

式 (11) の厳密な遂行には一般に $O(e^{|V|+|H|})$ の計算量が必要となるので現実的ではない．したがって実装の際には平均場近似や CD 法などの近似的手段が必要となる．

2.2 Deep Belief Networkに対するGreedy学習アルゴリズム

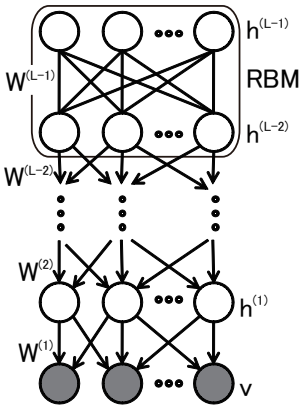

図 2: L 層の DBN．図において最下層がデータの入出力をおこなう可視素子の層であり，最深の 2 層間は RBM を構成している．最深の 2 層間以外の結合には方向性がある．

DBN は確率的素子からなる階層構造をもつ確率推論モデルである（図 2）．最深の 2 層間は RBM を構成しているが，それ以外では有効非巡回グラフになっている．以下で Hinton らにより提案された DBN の Greedy な学習アルゴリズムを概説する．

L 層の DBN を考える．層 l は確率変数 $\mathbf{h}^{(l)} \in \{0,1\}^{\nu(l)}$ をもつとする．ここで $\nu(l)$ を層 l に含まれている素子数とした．層 $l=0$ は可視素子（データの入出力の素子）の層となるため，特別に $\mathbf{h}^{(0)} = \mathbf{v}$ とする．素子 $\mathbf{h}^{(l)}$ の状態確率は遷移確率

$$P_{\mathrm{tr}}(\mathbf{h}^{(l)} \mid \mathbf{h}^{(l+1)}, \mathbf{W}^{(l+1)}) = \prod_{i \in \Omega_l} \frac{\exp\left\{\left(\theta_i^{(l)} + \sum_{j\in\Omega_{l+1}} w_{ij}^{(l+1)} h_j^{(l+1)}\right) h_i^{(l)}\right\}}{1 + \exp\left(\theta_i^{(l)} + \sum_{j\in\Omega_{l+1}} w_{ij}^{(l+1)} h_j^{(l+1)}\right)} \tag{12}$$

によって決定される．つまり $\boldsymbol{\theta}^{(l)}$ は素子 $\mathbf{h}^{(l)}$ のバイアスパラメータであり，$\mathbf{w}^{(l+1)}$ は素子 $\mathbf{h}^{(l)}$ と素子 $\mathbf{h}^{(l+1)}$ の間の結合パラメータである．ここで Ω_l は層 l に含まれる素子の集合である．また $\mathbf{W}^{(l)} = \{\boldsymbol{\theta}^{(l-1)}, \mathbf{w}^{(l)}\}$ としている．このとき，DBN に対する Greedy 学習アルゴリズムは

STEP 1. $l \leftarrow 0$

STEP 2. $\mathbf{h}^{(l)}$ と $\mathbf{h}^{(l+1)}$ の結合確率を $P_{\mathrm{RBM}}(\mathbf{h}^{(l)}, \mathbf{h}^{(l+1)} \mid \boldsymbol{\theta}^{(l)}, \boldsymbol{\theta}^{(l+1)}, \mathbf{w}^{(l+1)})$ と考え，RBM の学習アルゴリズムを用いてパラメータ $\{\boldsymbol{\theta}^{(l)}, \boldsymbol{\theta}^{(l+1)}, \mathbf{w}^{(l+1)}\}$ を学習する（$\mathbf{h}^{(l)}$ を可視素子の層，$\mathbf{h}^{(l+1)}$ を隠れ素子の層であると考える）．

STEP 3. $\mathbf{W}^{(l+1)}$ を固定する．この $\mathbf{W}^{(l+1)}$ を用いて $\mathbf{h}^{(l)}$ から $\mathbf{h}^{(l+1)}$ を式 (7) の確率に従って発生させ，これを新しい “データ” とみなす．

STEP 4. $l \leftarrow l+1$
if($l \leq L-2$) → **STEP 2.** に戻る．
if($l = L-1$) → 終了．

となる．Greedy 学習アルゴリズムが終了した段階のパラメータの値を有効な初期値と考え，Wake-Sleep アルゴリズム [9] などの少々コストの高い学習アルゴリズムを実行してパラメータを再度調整することも可能である．

Hinton 等は上記の Greedy 学習アルゴリズムは“ Variational Bound” と呼ばれる DBN におけるデータの尤度の下限を上昇させるアルゴリズムになっていると指摘している [1]．Roux and Bengio はこの Variational Bound 最適化の観点から式 (11) の基準とは異なる RBM に対する新しい学習の基準を提案した [7].

Roux and Bengio は 2.1 節で議論した観測データの経験分布 $P_D(\mathbf{v})$ と RBM の周辺分布 $P_V(\mathbf{v} \mid \boldsymbol{\Theta})$ との間の KL 情報量最小化の基準の変わりにデータの経験分布 $P_D(\mathbf{v})$ と分布

$$P_V^D(\mathbf{v} \mid \boldsymbol{\Theta}) \equiv \sum_{\mathbf{h},\mathbf{v}^0} P_{V|H}(\mathbf{v} \mid \mathbf{h}, \mathbf{a}, \mathbf{w}) P_{H|V}(\mathbf{h} \mid \mathbf{v}^0, \mathbf{b}, \mathbf{w}) P_D(\mathbf{v}^0) \tag{13}$$

との間の KL 情報量最小化を RBM の学習の基準とした：

$$\widehat{\boldsymbol{\Theta}} = \arg\min_{\boldsymbol{\Theta}} \mathrm{KL}(P_D || P_V^D) \tag{14}$$

分布 $P_V^D(\mathbf{v} \mid \boldsymbol{\Theta})$ は観測データ点の経験分布から分布 $P_{H|V}(\mathbf{h} \mid \mathbf{v}, \mathbf{b}, \mathbf{w})$ にしたがい隠れ素子の状態をつくり，それを用いて分布 $P_{V|H}(\mathbf{v} \mid \mathbf{h}, \mathbf{a}, \mathbf{w})$ にしたがい可視素子の状態分布を再構成した分布であると解釈できる．式 (14) の学習基準は DBN の Variational Bound の最適化を達成する基準であり，とくに $L=3$ で DBN が観測データ点の経験分布を十分に表現し得るものであるならば，Greedy 学習の際にこの基準による RBM 学習を用いることで最適な DBN を学習できることが示されている [7].

3 Restricted Boltzmann Machineの近似学習アルゴリズム

本節では KLIEP を用いて Roux and Bengio[7] により提案された新しい基準による RBM の学習アルゴリズムの近似アルゴリズムを提案する．

3.1 Roux and Bengioの基準による学習アルゴリズム

Roux and Bengio の基準 (14) にしたがい再構成された分布 $P_V^D(\mathbf{v} \mid \mathbf{\Theta})$ とデータの経験分布との間の KL 情報量

$$\mathrm{KL}(P_D||P_V^D) = \sum_{\mathbf{v}} P_D(\mathbf{v}) \ln \frac{P_D(\mathbf{v})}{P_V^D(\mathbf{v} \mid \mathbf{\Theta})} \tag{15}$$

をパラメータ $\mathbf{\Theta}$ に関して最小化する．ただし $P_D(\mathbf{v})$ は観測データ点の経験分布であり式 (9) で定義されている．

パラメータ $\mathbf{\Theta} = \{\mathbf{a}, \mathbf{b}, \mathbf{w}\}$ に関する勾配はそれぞれ

$$\frac{\partial \mathrm{KL}(P_D||P_V^D)}{\partial a_i} = -\frac{1}{M}\sum_{\mu=1}^{M} d_i^\mu + \frac{1}{M}\sum_{\mu=1}^{M}\sum_{\mathbf{h}} f_i(\mathbf{h}, a_i, \mathbf{w}) W(\mathbf{h}, \mathbf{\Theta}) P_{H|V}(\mathbf{h} \mid \mathbf{d}^\mu, \mathbf{b}, \mathbf{w}) \tag{16}$$

$$\frac{\partial \mathrm{KL}(P_D||P_V^D)}{\partial b_j} = -\frac{1}{M}\sum_{\mu=1}^{M}\sum_{\mathbf{h}} \Big(h_j - g_j(\mathbf{d}^\mu, b_j, \mathbf{w})\Big) \times W(\mathbf{h}, \mathbf{\Theta}) P_{H|V}(\mathbf{h} \mid \mathbf{d}^\mu, \mathbf{b}, \mathbf{w}) \tag{17}$$

$$\frac{\partial \mathrm{KL}(P_D||P_V^D)}{\partial w_{ij}} = -\frac{1}{M}\sum_{\mu=1}^{M}\sum_{\mathbf{h}} \Big\{ h_j W_i(\mathbf{h}, \mathbf{\Theta}) + \Big(d_i^\mu h_j - h_j f_i(\mathbf{h}, a_i, \mathbf{w}) - d_i^\mu g_j(\mathbf{d}^\mu, b_j, \mathbf{w})\Big) W(\mathbf{h}, \mathbf{\Theta}) \Big\} \times P_{H|V}(\mathbf{h} \mid \mathbf{d}^\mu, \mathbf{b}, \mathbf{w}) \tag{18}$$

となる．ここで $f_i(\mathbf{h}, a_i, \mathbf{w})$ と $g_j(\mathbf{v}, b_j, \mathbf{w})$ はそれぞれ

$$f_i(\mathbf{h}, a_i, \mathbf{w}) \equiv \frac{\exp\left(a_i + \sum_{j\in H} w_{ij} h_j\right)}{1 + \exp\left(a_i + \sum_{j\in H} w_{ij} h_j\right)} \tag{19}$$

$$g_j(\mathbf{v}, b_j, \mathbf{w}) \equiv \frac{\exp\left(b_j + \sum_{i\in V} w_{ij} v_i\right)}{1 + \exp\left(b_j + \sum_{i\in V} w_{ij} v_i\right)} \tag{20}$$

と定義しており，$W(\mathbf{h}, \mathbf{\Theta})$，$W_i(\mathbf{h}, \mathbf{\Theta})$ はそれぞれ

$$W(\mathbf{h}, \mathbf{\Theta}) \equiv \sum_{\mathbf{v}} \frac{P_D(\mathbf{v}) P_{V|H}(\mathbf{v} \mid \mathbf{h}, \mathbf{a}, \mathbf{w})}{P_V^D(\mathbf{v} \mid \mathbf{\Theta})} = \frac{1}{M}\sum_{\mu=1}^{M} \frac{P_{V|H}(\mathbf{d}^\mu \mid \mathbf{h}, \mathbf{a}, \mathbf{w})}{P_V^D(\mathbf{d}^\mu \mid \mathbf{\Theta})} \tag{21}$$

$$W_i(\mathbf{h}, \mathbf{\Theta}) \equiv \sum_{\mathbf{v}} v_i \frac{P_D(\mathbf{v}) P_{V|H}(\mathbf{v} \mid \mathbf{h}, \mathbf{a}, \mathbf{w})}{P_V^D(\mathbf{v} \mid \mathbf{\Theta})} = \frac{1}{M}\sum_{\mu=1}^{M} d_i^\mu \frac{P_{V|H}(\mathbf{d}^\mu \mid \mathbf{h}, \mathbf{a}, \mathbf{w})}{P_V^D(\mathbf{d}^\mu \mid \mathbf{\Theta})} \tag{22}$$

としている．

勾配 (16)〜(18) は素子数の指数オーダーの計算量を必要とするため，大きいシステムでは現実的な時間内で厳密に計算することはできない．そこで平均場近似やモンテカルロ積分等のなんらかの近似法が必要となるのであるが，これら勾配の計算にはもう一つ問題点が存在する．式 (21)，(22) からわかるように $W(\mathbf{h}, \mathbf{\Theta})$，$W_i(\mathbf{h}, \mathbf{\Theta})$ の計算には観測データ点の個数の和を実行する必要がある．よってそれぞれの勾配は観測データ点の個数に関する 2 重和をもつこととなる．

観測データ点の個数が大きい場合はこの計算にかなりのコストがかかってしまうため，まず何らかの近似を施す前にこの 2 重和を回避する方法を考える必要がある．

3.2 Kullback-Leibler Importance Estimation Procedureによる近似

式 (21)，(22) の $W(\mathbf{h}, \mathbf{\Theta})$，$W_i(\mathbf{h}, \mathbf{\Theta})$ における密度比 $P_D(\mathbf{v})/P_V^D(\mathbf{v}|\mathbf{\Theta})$ を Sugiyama 等によって提案された KLIEP[8] の近似法を用い近似的に評価することで，勾配 (16)〜(18) における観測データ点の個数に関する 2 重和を回避する．以下では，密度比 $P_D(\mathbf{v})/P_V^D(\mathbf{v}|\mathbf{\Theta})$ に対して KLIEP を適用し，$W(\mathbf{h}, \mathbf{\Theta})$，$W_i(\mathbf{h}, \mathbf{\Theta})$ を近似的に評価する方法について述べる．

密度比 $P_D(\mathbf{v})/P_V^D(\mathbf{v}|\mathbf{\Theta})$ に対して

$$P_D(\mathbf{v})/P_V^D(\mathbf{v} \mid \mathbf{\Theta}) \propto \exp\left(\sum_{i\in V} c_i v_i\right) \tag{23}$$

なる近似的モデル化を試みる．これは規格化定数を考慮すると

$$P_D(\mathbf{v}) \approx \frac{1}{\mathcal{Z}(\mathbf{c}, \mathbf{\Theta})} \exp\left(\sum_{i\in V} c_i v_i\right) P_V^D(\mathbf{v}|\mathbf{\Theta}) \equiv Q_V(\mathbf{v} \mid \mathbf{c}, \mathbf{\Theta}) \tag{24}$$

と近似することに対応する．ただし

$$\mathcal{Z}(\mathbf{c}, \mathbf{\Theta}) \equiv \sum_{\mathbf{v}} \exp\left(\sum_{i\in V} c_i v_i\right) P_V^D(\mathbf{v}|\mathbf{\Theta}) \tag{25}$$

は規格化定数である．KLIEP の文脈によると KL 情報量 $\mathrm{KL}(P_D||Q_V)$ を最小にする $\mathbf{c}$ を最適な $\mathbf{c}$ と考える．したがって KLIEP の文脈で最適な $\mathbf{c}$ は方程式

$$\sum_{\mu=1}^{M} d_i^\mu = \sum_{\mu=1}^{M}\sum_{\mathbf{h}} f_i(\mathbf{h}, a_i + c_i, \mathbf{w}) P_{H|V}(\mathbf{h} \mid \mathbf{d}^\mu, \mathbf{b}, \mathbf{w}) \tag{26}$$

の解であることが導ける．密度比推定式 (24) を式 (21)，(22) に代入することにより

$$W(\mathbf{h}, \mathbf{\Theta}) \approx \frac{1}{\mathcal{Z}(\mathbf{c}, \mathbf{\Theta})} \frac{G_H(\mathbf{h}, \mathbf{a} + \mathbf{c}, \mathbf{w})}{G_H(\mathbf{h}, \mathbf{a}, \mathbf{w})} \tag{27}$$

$$W_i(\mathbf{h},\boldsymbol{\Theta}) \approx \frac{f_i(\mathbf{h},a_i+c_i,\mathbf{w})}{\mathcal{Z}(\mathbf{c},\boldsymbol{\Theta})}\frac{G_H(\mathbf{h},\mathbf{a}+\mathbf{c},\mathbf{w})}{G_H(\mathbf{h},\mathbf{a},\mathbf{w})} \quad (28)$$

なる近似を得る．

式 (27)，(28) を勾配 (16)〜(18) に代入することによりそれぞれ

$$\begin{aligned}&\frac{\partial \mathrm{KL}(P_D||P_V^D)}{\partial a_i}\\ &\approx -\frac{1}{M}\sum_{\mu=1}^{M} d_i^\mu + \frac{1}{M\mathcal{Z}(\mathbf{c},\boldsymbol{\Theta})}\sum_{\mu=1}^{M}\sum_{\mathbf{h}} f_i(\mathbf{h},a_i,\mathbf{w})\\ &\quad\times \frac{G_H(\mathbf{h},\mathbf{a}+\mathbf{c},\mathbf{w})}{G_H(\mathbf{h},\mathbf{a},\mathbf{w})}P_{H|V}(\mathbf{h}\mid\mathbf{d}^\mu,\mathbf{b},\mathbf{w}) \quad (29)\\ &\frac{\partial \mathrm{KL}(P_D||P_V^D)}{\partial b_j} \approx -\frac{1}{M\mathcal{Z}(\mathbf{c},\boldsymbol{\Theta})}\sum_{\mu=1}^{M}\sum_{\mathbf{h}}\Big(h_j\\ &\quad - g_j(\mathbf{d}^\mu,b_j,\mathbf{w})\Big)\frac{G_H(\mathbf{h},\mathbf{a}+\mathbf{c},\mathbf{w})}{G_H(\mathbf{h},\mathbf{a},\mathbf{w})}P_{H|V}(\mathbf{h}\mid\mathbf{d}^\mu,\mathbf{b},\mathbf{w}) \quad (30)\\ &\frac{\partial \mathrm{KL}(P_D||P_V^D)}{\partial w_{ij}}\\ &\approx -\frac{1}{M\mathcal{Z}(\mathbf{c},\boldsymbol{\Theta})}\sum_{\mu=1}^{M}\sum_{\mathbf{h}}\Big(d_i^\mu h_j + h_j f_i(\mathbf{h},a_i+c_i,\mathbf{w})\\ &\quad - h_j f_i(\mathbf{h},a_i,\mathbf{w}) - d_i^\mu g_j(\mathbf{d}^\mu,b_j,\mathbf{w})\Big)\frac{G_H(\mathbf{h},\mathbf{a}+\mathbf{c},\mathbf{w})}{G_H(\mathbf{h},\mathbf{a},\mathbf{w})}\\ &\quad\times P_{H|V}(\mathbf{h}\mid\mathbf{d}^\mu,\mathbf{b},\mathbf{w}) \quad (31)\end{aligned}$$

なる近似を得ることができる．さらに式 (26) は

$$\begin{aligned}\frac{1}{M}\sum_{\mu=1}^{M} d_i^\mu &= \frac{1}{M\mathcal{Z}(\mathbf{c},\boldsymbol{\Theta})}\sum_{\mu=1}^{M}\sum_{\mathbf{h}} f_i(\mathbf{h},a_i+c_i,\mathbf{w})\\ &\quad\times\frac{G_H(\mathbf{h},\mathbf{a}+\mathbf{c},\mathbf{w})}{G_H(\mathbf{h},\mathbf{a},\mathbf{w})}P_{H|V}(\mathbf{h}\mid\mathbf{d}^\mu,\mathbf{b},\mathbf{w}) \quad (32)\end{aligned}$$

と変形可能であるので式 (29) は式 (26) の解である最適な $\mathbf{c}$ を用いて

$$\begin{aligned}&\frac{\partial \mathrm{KL}(P_D||P_V^D)}{\partial a_i}\\ &\approx -\frac{1}{M\mathcal{Z}(\mathbf{c},\boldsymbol{\Theta})}\sum_{\mu=1}^{M}\sum_{\mathbf{h}}\Big(f_i(\mathbf{h},a_i+c_i,\mathbf{w})\\ &\quad - f_i(\mathbf{h},a_i,\mathbf{w})\Big)\frac{G_H(\mathbf{h},\mathbf{a}+\mathbf{c},\mathbf{w})}{G_H(\mathbf{h},\mathbf{a},\mathbf{w})}P_{H|V}(\mathbf{h}\mid\mathbf{d}^\mu,\mathbf{b},\mathbf{w}) \quad (33)\end{aligned}$$

と表現することができる．

KLIEP の適用により近似された勾配 (29)〜(31) は観測データ点の個数に関する和は一つしかもたない．したがって 3.1 節で議論した観測データ点の個数に関する 2 重和の問題は回避されていることがわかる．勾配 (30)，(31)，(33) に共通する係数 $\mathcal{Z}(\mathbf{c},\boldsymbol{\Theta})^{-1}$ は有限の $\mathbf{c}$ に対して常に正なので，この係数を無視しても勾配の方向は変わらない．したがってこの係数の寄与を無視するという近似をさらに施す．

以上の議論より $\mathrm{KL}(P_D||P_V^D)$ のパラメータ a_i，b_j，w_{ij} に対する勾配 Δa_i，Δb_j，Δw_{ij} は KLIEP によりそれぞれ以下のように近似される．

$$\begin{aligned}\Delta a_i &\propto -\sum_{\mu=1}^{M}\sum_{\mathbf{h}}\Big(f_i(\mathbf{h},a_i+c_i,\mathbf{w}) - f_i(\mathbf{h},a_i,\mathbf{w})\Big)\\ &\quad\times\frac{G_H(\mathbf{h},\mathbf{a}+\mathbf{c},\mathbf{w})}{G_H(\mathbf{h},\mathbf{a},\mathbf{w})}P_{H|V}(\mathbf{h}\mid\mathbf{d}^\mu,\mathbf{b},\mathbf{w}) \quad (34)\\ \Delta b_j &\propto -\sum_{\mu=1}^{M}\sum_{\mathbf{h}}\Big(h_j - g_j(\mathbf{d}^\mu,b_j,\mathbf{w})\Big)\frac{G_H(\mathbf{h},\mathbf{a}+\mathbf{c},\mathbf{w})}{G_H(\mathbf{h},\mathbf{a},\mathbf{w})}\\ &\quad\times P_{H|V}(\mathbf{h}\mid\mathbf{d}^\mu,\mathbf{b},\mathbf{w}) \quad (35)\\ \Delta w_{ij} &\propto -\sum_{\mu=1}^{M}\sum_{\mathbf{h}}\Big(d_i^\mu h_j + h_j f_i(\mathbf{h},a_i+c_i,\mathbf{w})\\ &\quad - h_j f_i(\mathbf{h},a_i,\mathbf{w}) - d_i^\mu g_j(\mathbf{d}^\mu,b_j,\mathbf{w})\Big)\\ &\quad\times\frac{G_H(\mathbf{h},\mathbf{a}+\mathbf{c},\mathbf{w})}{G_H(\mathbf{h},\mathbf{a},\mathbf{w})}P_{H|V}(\mathbf{h}\mid\mathbf{d}^\mu,\mathbf{b},\mathbf{w}) \quad (36)\end{aligned}$$

$\mathbf{c}$ を求める方程式 (26) と式 (34)〜(36) における $\mathbf{h}$ に関する和は $O(e^{|H|})$ 個の項の和なので大規模なシステムでは厳密にこれらの勾配を計算することが難しい．したがって実装の際にはここから更に何らかの近似をする必要がある．

そこで本論文では $\mathbf{h}$ に関する和，すなわち分布 $P_{H|V}(\mathbf{h}\mid\mathbf{d}^\mu,\mathbf{b},\mathbf{w})$ に関する平均を $P_{H|V}(\mathbf{h}\mid\mathbf{d}^\mu,\mathbf{b},\mathbf{w})$ からのサンプリングによるモンテカルロ積分で近似的に評価する．2.1 節で議論したように分布 $P_{H|V}(\mathbf{h}\mid\mathbf{d}^\mu,\mathbf{b},\mathbf{w})$ は条件付き独立であるので，ここからサンプルを得ることは容易である．モンテカルロ積分のためのサンプリングとして K 点のサンプリングをおこなうとすると，$\mathbf{c}$ の計算，あるいは勾配 (34)〜(36) を一回計算するのに必要な計算量は $O(KM|V||H|)$ 程度である．

提案近似アルゴリズムは以下のようにまとめられる．

STEP 1. パラメータ $\boldsymbol{\Theta}$ を初期化する．

STEP 2. 方程式 (26) を解き $\mathbf{c}$ を求める．

STEP 3. **STEP 2.** で求めた $\mathbf{c}$ を用いて勾配 (34)〜(36) を計算する．

STEP 4. 勾配降下法によりパラメータ $\boldsymbol{\Theta}$ を更新する．

STEP 5. 収束条件を満たさなければ **STEP 2.** に戻る．

4　数値実験

本節では提案学習アルゴリズムを用いて数値実験を行う．まず，Roux and Bengio (RB) の基準 (14) にしたがい勾配 (16)〜(18) を用いて RBM を学習するアルゴリズムを RB アルゴリズムとする．これに対して KLIEP の近似法を適用したアルゴリズムを RB-K アルゴリズムとする．RB-K の中で式 (26) を解いて $\mathbf{c}$ を求め，式 (29)〜(31) によりパラメータの勾配を計算するアルゴリズムを RB-KZ，式 (26) を解いて $\mathbf{c}$ を求め，式 (34)〜(36) を用いるアルゴリズムを RB-KE，そして，RB-KE の $\mathbf{c}$ を求める計算と勾配計算における分布 $P_{H|V}(\mathbf{h} \mid \mathbf{d}^\mu, \mathbf{b}, \mathbf{w})$ に関する平均を $P_{H|V}(\mathbf{h} \mid \mathbf{d}^\mu, \mathbf{b}, \mathbf{w})$ からのサンプリングによるモンテカルロ積分で近似的に評価し計算するアルゴリズムを RB-KS と呼ぶこととする．RB アルゴリズムの近似アルゴリズムである RB-K の近似性能を同一のデータを RB アルゴリズムで学習した分布 $P_{\mathrm{RB}}(\mathbf{v})$ と RB-K アルゴリズムで学習した分布 $P_{\mathrm{RB-K}}(\mathbf{v})$ の KL 情報量

$$\mathrm{KL}(P_{\mathrm{RB}}||P_{\mathrm{RB-K}}) \equiv \frac{1}{|V|} \sum_{\mathbf{v}} P_{\mathrm{RB}}(\mathbf{v}) \ln \frac{P_{\mathrm{RB}}(\mathbf{v})}{P_{\mathrm{RB-K}}(\mathbf{v})} \tag{37}$$

により評価する．

学習には可視素子数 $|V| = 4$，隠れ素子数 $|H| = 4$ で各可視素子 (隠れ素子) がすべての隠れ素子 (可視素子) と結合をもつ RBM を用いる．データは学習に用いるものと同じ構造をもつ RBM により生成する．データ生成モデルにおけるパラメータ $\mathbf{\Theta} = \{\mathbf{a}, \mathbf{b}, \mathbf{w}\}$ はそれぞれ平均 0，分散 $(0.2)^2$ のガウス分布 $\mathcal{N}(0, 0.2)$ から独立にサンプルした値を用いる．このデータ生成モデルから $\{0,1\}^{|V|}$ の 2 値の人工データを $M = 100$ 個生成し，これを観測データとする．まず，サンプリングにより得られた観測データ点 $\{\mathbf{d}^\mu \in \{0,1\}^{|V|} \mid \mu = 1, \cdots, M\}$ に対して RB アルゴリズムを用いてパラメータが収束するまで RBM の学習を行う．次に，パラメータ更新回数を 500 に固定して同一のデータで RB-K アルゴリズムにより RBM の学習を行い，パラメータ更新ごとに KL 情報量 37 を計算する．これを，異なるデータに関して 100 回試行しその平均値を図 3 に示す．ただし，RB-K による学習におけるステップサイズはすべて 0.05 で一定とし，RB-KS におけるサンプル点数は $K = 5$ としている．

図 3 より，RB-KZ，RB-KE，RB-KS はそれぞれ RB アルゴリズムのよい近似を与えているといえる．RB-KZ の収束が速いのは，ステップサイズ一定の下では RB-KE，RB-KS と比較すると勾配の値が大きくなるためである．これは，パラメータ更新回数が少ないうちは

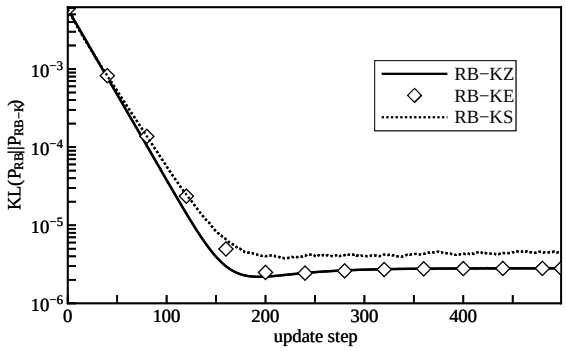

図 3: $\mathrm{KL}(P_{\mathrm{RB}}||P_{\mathrm{RB-K}})$ の収束性．異なるデータに関して 100 回試行し，その平均値を与えている．式 (29)〜(31) によりパラメータの勾配を計算するアルゴリズムを RB-KZ，式 (34)〜(36) を用いるアルゴリズムを RB-KE，RB-KE の勾配計算における一部の統計量をモンテカルロ積分で近似的に評価し計算するアルゴリズムを RB-KS とそれぞれ表している．

$\mathcal{Z}(\mathbf{c}, \mathbf{\Theta}) \leq 1$ であることが多いため起こる．3.2 節で示したように RB-K アルゴリズムは RB アルゴリズムにおけるデータに関する 2 重和を回避できている．したがって，提案手法は RB アルゴリズムの性能を低下させることなくより大規模なネットワークに適用できるアルゴリズムであるといえる．

5　まとめ

本論文では，RBM の学習アルゴリズムを Roux and Bengio の基準 (14) による RBM の学習法に KLIEP を組み合わせることによって提案した．また，人口データに対する数値実験において，提案アルゴリズムは Roux and Bengio の基準 (14) による学習法に対するよい近似となっていることを確認した．

本論文では厳密計算との比較を行うためにノード数の少ない場合の数値実験のみを与えているが，より多くのノード数をもつ大規模な体系に対する数値実験を行う準備を進めている．また，提案法の近似精度だけでなく，提案法を採用することによりどの程度計算時間が短縮されるか評価することも必要である．観測データ点数 M，サンプル点数 K をそれぞれ変化させたときの計算時間と近似精度の関係を調べる実験の準備も進めている．さらに，手書き文字認識などの具体的な情報処理の応用における問題に適用することも検討している．

謝辞

本研究の一部は文部科学省科学研究費補助金 (No.21700247 and No.18079002) およびグローバル COE プログラム

“Center of Education and Research for Information Electronics Systems” の補助を得て行われたものである.

参考文献

[1] G. E. Hinton, S. Osindero, and Y. W. Teh: A fast learning algorithm for deep belief nets. *Neural Computation*, Vol.18, No.7, pp.1527-1554, 2006.

[2] G. E. Hinton and R. R. Salakhutdinov: Reducing the dimensionality of data with neural networks. *Science*, 313, pp.504-507, 2006.

[3] G. E. Hinton: Products of experts. In *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN)* , Vol.1, pp.1-6, 1999.

[4] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle: Greedy layer-wise training of deep networks. I*Advances in Neural Information Processing Systems* 19, pp.153-160, 2007.

[5] G. E. Hinton: Training products of experts by minimizing contrastive divergence. *Neural Computation*, Vol.14, No.8, pp.1771-1800, 2002.

[6] M. A. Carreira-Perpinan and G. E. Hinton: On contrastive divergence learning. In *Artificial Intelligence and Statistics* , 2005.

[7] N. Le Roux and Y. Bengio: Representational Power of Restricted Boltzmann Machines and Deep Belief Networks. *Neural Computation*, Vol.20, No.6, pp.1631-1649, 2008.

[8] M. Sugiyama, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe: Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation. In J. C. Platt, D. Koller, Y. Singer, and S. Roweis (Eds.), *Advances in Neural Information Processing Systems* 20, pp.1433-1440, 2008.

[9] R. Neal and P. Dayan: Factor Analysis Using Delta-Rule Wake-Sleep Learning. *Neural Computation*, Vol.9, pp.1781-1803, 1996.