F-024

Gaussian-Bernoulli restricted Boltzmann machine に対する平均場近似
Mean-field approximation for Gaussian-Bernoulli restricted Boltzmann machine

高橋　茶子 *　　安田　宗樹 †
Chako Takahashi　　Muneki Yasuda

1　はじめに

深層学習 (deep learning) の登場 [1] で，ここ 10 年来急速に応用を含めた機械学習分野が発展してきている．深層学習は事前学習 (pretraining) を基本戦略としており，事前学習においては制限ボルツマンマシン (restricted Boltzmann machine; RBM) に対する学習が鍵となる [1]．RBM は可視変数の層と隠れ変数の層の 2 層からなるボルツマンマシンであり，通常の RBM では可視変数と隠れ変数共に離散変数として定義される．しかし，画像処理や音声処理など，実用的にはデータは必ずしも離散量ではなく連続量である場合もある．そこで可視変数（かまたは場合によっては隠れ変数）を連続変数にした RBM である Gaussian-Bernoulli RBM (GBRBM) が登場する [3, 4, 5]．GBRBM は連続量のデータを扱うことのできる RBM であり，GBRBM をデータ層に置いた Gaussian-Bernoulli DBM (GBDBM) なる深層学習モデルも提案されている [2]．GBDBM は深層ボルツマンマシン (deep Boltzmann machine; DBM) [6, 7] の一種であり，連続量のデータを扱える確率的深層学習モデルとなっている．

確率的深層学習モデルや RBM 上での推論や学習の中では平均場近似 (mean-field approximation) [8] と呼ばれる統計力学由来の近似計算アルゴリズムが利用されている [9, 6, 7, 2]．したがって，平均場近似の性能を定性的・定量的に調べることは深層学習研究において未だ重要なものとなっている．本論文では GBRBM に対する平均場近似の性能を調べることを目的とする．GBRBM に対する平均場近似は大きく分けて 2 種類の方法が考えられる．一つ目は全ての確率変数をそれぞれ統計的に独立として近似するような従来的な方法である．またそれとは別に，一部の確率変数を周辺化により消去した周辺分布に対して平均場近似を適用する方法も考えられる．本論文ではこれら 2 種類の平均場近似をそれぞれ導出し，両者を定性的視点と定量的視点から比較する．

2 節で本論文の基礎モデルとなる GBRBM を定義し，続く 3 節において GBRBM に対する 2 種類の平均場近似を導出する．4 節では前節で導出した 2 種類の平均場近似の性能を定性的視点から比較し，5 節では数値実験を用いて両者を定量的に比較する．6 節は本論文のまとめである．

* 山形大学大学院理工学研究科; CREST, JST (Yamagata University)
† 山形大学大学院理工学研究科; CREST, JST (Yamagata University)

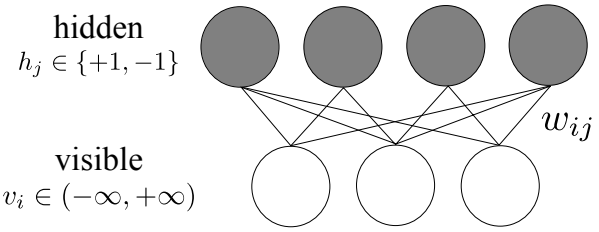

図 1　完全 2 部グラフ上に定義される GBRBM．下層が可視層で，上層が隠れ層である．

2　Gaussian-Bernoulli restricted Boltzmann machine

GBRBM は，図 1 に示すような完全 2 部グラフ上に定義される確率的学習モデルである．下層は可視変数のみから構成される可視層 V で，上層は隠れ変数のみから構成される隠れ層 H である．V と H はそれぞれ可視層と隠れ層のノード番号の集合である．可視変数 $\boldsymbol{v} = \{v_i \in (-\infty, \infty) \mid i \in V\}$ は入出力データと直接関連付けられる変数であり，連続値を取る確率変数である．また，隠れ変数 $\boldsymbol{h} = \{h_j \in \{+1, -1\} \mid j \in H\}$ は入出力データとは直接関連付けられないシステムの内部変数であり，2 値をとる離散確率変数である．GBRBM のエネルギー関数は次のように定義される [5]．

$$E(\boldsymbol{v}, \boldsymbol{h} \mid \boldsymbol{\theta}) := \sum_{i \in V} \frac{(v_i - b_i)^2}{2{\sigma_i}^2} - \sum_{i \in V} \sum_{j \in H} \frac{w_{ij}}{{\sigma_i}^2} v_i h_j - \sum_{j \in H} c_j h_j \tag{1}$$

ここで，$\boldsymbol{b} = \{b_i \mid i \in V\}$ と $\boldsymbol{c} = \{c_j \mid j \in H\}$ はそれぞれ可視変数と隠れ変数に対するバイアスパラメータであり，$\boldsymbol{w} = \{w_{ij} \mid i \in V, j \in H\}$ は可視変数と隠れ変数間の結合パラメータである．$\boldsymbol{\sigma} = \{\sigma_i \mid i \in V\}$ は可視変数の分散に関連するパラメータとなっている．これらのモデルパラメータをまとめて $\boldsymbol{\theta}$ で表すこととする．式 (1) のエネルギー関数を用いて，GBRBM は

$$P(\boldsymbol{v}, \boldsymbol{h} \mid \boldsymbol{\theta}) := \frac{1}{Z(\boldsymbol{\theta})} \exp\big(-E(\boldsymbol{v}, \boldsymbol{h} \mid \boldsymbol{\theta})\big) \tag{2}$$

のようなボルツマン分布の形で表される．$Z(\boldsymbol{\theta})$ は規格化定数（分配関数と呼ばれることもある）であり

$$Z(\boldsymbol{\theta}) := \int_{-\infty}^{\infty} \sum_{\boldsymbol{h}} \exp\big(-E(\boldsymbol{v}, \boldsymbol{h} \mid \boldsymbol{\theta})\big) \mathrm{d}\boldsymbol{v}$$

のように定義されている．ここで $\int_{-\infty}^{\infty}(\cdots)\mathrm{d}\boldsymbol{v}$ は $\boldsymbol{v}$ に関する多重積分を表しており，$\sum_{\boldsymbol{h}}$ は $\boldsymbol{h}$ の可能な全ての組み合わせに関する多重和を表している．

式 (2) において，片方の層を条件とした場合のもう片方の層の条件付き分布はそれぞれ次のようになる．

$$P(\boldsymbol{v} \mid \boldsymbol{h}, \boldsymbol{\theta}) = \prod_{i \in V} \mathcal{N}\left(v_i \mid b_i + \sum\nolimits_{j \in H} w_{ij} h_j, \sigma_i^2\right) \quad (3)$$

$$P(\boldsymbol{h} \mid \boldsymbol{v}, \boldsymbol{\theta}) = \frac{\exp\left\{\left(c_j + \sum_{i \in V} w_{ij} v_i\right) h_j\right\}}{2\cosh\left(c_j + \sum_{i \in V} w_{ij} v_i\right)} \quad (4)$$

ここで，$\mathcal{N}(v \mid \mu, \sigma^2)$ は平均 μ，分散 σ^2 の 1 次元ガウス分布を表している．式 (3) と式 (4) は片方の層の変数の値が条件として与えられると，もう片方の層の変数がそれぞれ統計的に独立になるということを表しており，この性質は条件付き独立性と呼ばれる．GBRBM は 2 部グラフ構造のおかげでこのようのな層ごとの条件付き独立性をもつ．式 (3) と式 (4) はそれぞれガウス分布とベルヌーイ分布となっており，これが GBRBM の名前の由来となっている．

GBRBM の可視変数の期待値 $\langle v_i \rangle$ と隠れ変数の期待値 $\langle h_j \rangle$ はそれぞれ

$$\langle v_i \rangle := \int_{-\infty}^{\infty} \sum_{\boldsymbol{h}} v_i P(\boldsymbol{v}, \boldsymbol{h} \mid \boldsymbol{\theta}) \mathrm{d}\boldsymbol{v} \quad (5)$$

$$\langle h_j \rangle := \int_{-\infty}^{\infty} \sum_{\boldsymbol{h}} h_j P(\boldsymbol{v}, \boldsymbol{h} \mid \boldsymbol{\theta}) \mathrm{d}\boldsymbol{v} \quad (6)$$

により得られる．確率モデル上での推論や学習においてはこれらの期待値計算が非常に重要となる．しかしながら，これらの期待値計算はノード数の増加に対して指数的に増加する計算量を必要としてしまうため，実用的な場面において現実的な時間内に厳密な期待値計算を終えるのはほとんど不可能である．そこで，実装上は何らかの近似的計算手法を採用することとなる．近似的計算手法の代表的なものの一つが平均場近似である．以下の節では GBRBM に対する平均場近似について議論していく．

3 GBRBM に対する平均場近似

平均場近似の基本的な手続きは以下のような変分近似 (variational approximation) である．まず，全ての確率変数について因数分解された形（つまり，全ての確率変数が統計的に独立であるような形）のテスト分布を用意する．そして，元の真の分布と用意したテスト分布の間のカルバック・ライブラー (Kullback-Leibler; KL) 情報量を作り，その KL 情報量を最小とするテスト分布を元の真の分布の平均場近似分布であるとするのである [8]．つまり，全ての確率変数について因数分解された形の分布の中で，KL 情報量の尺度においてもっとも元の真の分布に近いものを近似分布とするということになる．

本節では，GBRBM に対する 2 通りの平均場近似法を考える．一つは可視変数と隠れ変数の両変数に対して平均場近似を適用する方法であり，もっともナイーブな平均場近似の適用法である（3.1 節）．例えば文献 [9] では欠損データの推定アルゴリズムにおいてこの一つ目の方法に基づく平均場近似が採用されている．もう一つの方法は，隠れ変数のみに平均場近似を適用する方法，つまり周辺化により可視変数を消去した隠れ変数のみの分布に対する平均場近似である（3.2 節）．周辺化した分布に対して平均場近似を適用することはどのようなモデルに対しても原理的には可能であるが，簡単な形で結果が定式化できるかどうかはモデルの構造に大きく依存する．GBRBM は特殊なモデル構造のおかげでこの二つ目の平均場近似法が容易に定式化可能となっている．

3.1 全変数に対する平均場近似

本節では一つ目の平均場近似法，即ち，可視変数と隠れ変数の両変数に対して平均場近似を適用する方法を導出する．テスト分布を次のように設定する．

$$T_1(\boldsymbol{v}, \boldsymbol{h}) := Q(\boldsymbol{v}) U(\boldsymbol{h}) \quad (7)$$

ただし，$Q(\boldsymbol{v})$ と $U(\boldsymbol{h})$ は

$$Q(\boldsymbol{v}) := \prod_{i \in V} q_i(v_i), \quad U(\boldsymbol{h}) := \prod_{j \in H} u_j(h_j) \quad (8)$$

により定義されている．ここで $q_i(v_i)$ は可視変数 v_i に対する分布であり，$u_j(h_j)$ は隠れ変数 h_j の分布である．平均場近似の処方箋に従い，真の分布である式 (2) の GBRBM と式 (7) のテスト分布との間の KL 情報量

$$K_1[Q, U] := \int_{-\infty}^{\infty} \sum_{\boldsymbol{h}} T_1(\boldsymbol{v}, \boldsymbol{h}) \ln \frac{T_1(\boldsymbol{v}, \boldsymbol{h})}{P(\boldsymbol{v}, \boldsymbol{h} \mid \boldsymbol{\theta})} \mathrm{d}\boldsymbol{v} \quad (9)$$

を定義し，この KL 情報量を $Q(\boldsymbol{v})$ と $U(\boldsymbol{h})$ に関して最小化する．式 (9) は

$$K_1[Q, U] = \mathcal{F}_1[Q, U] + \ln Z(\boldsymbol{\theta})$$

と変形することができる．ここで

$$\begin{aligned} \mathcal{F}_1[Q, U] := & \int_{-\infty}^{\infty} \sum_{\boldsymbol{h}} T_1(\boldsymbol{v}, \boldsymbol{h}) E(\boldsymbol{v}, \boldsymbol{h} \mid \boldsymbol{\theta}) \mathrm{d}\boldsymbol{v} \\ & + \int_{-\infty}^{\infty} \sum_{\boldsymbol{h}} T_1(\boldsymbol{v}, \boldsymbol{h}) \ln T_1(\boldsymbol{v}, \boldsymbol{h}) \mathrm{d}\boldsymbol{v} \end{aligned} \quad (10)$$

は変分自由エネルギーと呼ばれ，この変分自由エネルギーを最小とする $Q(\boldsymbol{v})$ と $U(\boldsymbol{h})$ は式 (9) の KL 情報量を最小化する $Q(\boldsymbol{v})$ と $U(\boldsymbol{h})$ と一致する．式 (10) は

$$\begin{aligned} \mathcal{F}_1[Q, U] = & \sum_{i \in V} \int_{-\infty}^{\infty} \frac{(v_i - b_i)^2}{2\sigma_i^2} q_i(v_i) \mathrm{d}v_i \\ & - \sum_{i \in V} \sum_{j \in H} \frac{w_{ij}}{\sigma_i^2} \int_{-\infty}^{\infty} v_i q_i(v_i) \mathrm{d}v_i \sum_{h_j} h_j u_j(h_j) \\ & - \sum_{j \in H} c_j \sum_{h_j} h_j u_j(h_j) + \sum_{i \in V} \int_{-\infty}^{\infty} q_i(v_i) \ln q_i(v_i) \mathrm{d}v_i \\ & + \sum_{j \in H} \sum_{h_j} u_j(h_j) \ln u_j(h_j) \end{aligned} \quad (11)$$

と変形される．$q_i(v_i)$ と $u_j(h_j)$ に関する規格化条件

$$\int_{-\infty}^{\infty} q_i(v_i) \mathrm{d}v_i = 1, \quad \sum_{h_j} u_j(h_j) = 1$$

を条件としたラグランジュ未定乗数法により式 (11) を $q_i(v_i)$ と $u_j(h_j)$ に関して最小化すると，その極致条件より以下の結果を得る．

$$Q(\boldsymbol{v}) = \prod_{i\in V} \mathcal{N}\big(v_i \mid b_i + \textstyle\sum_{j\in H} w_{ij}\gamma_j, \sigma_i^2\big)$$
$$U(\boldsymbol{h}) = \prod_{j\in H} \frac{\exp\big\{\big(c_j + \sum_{i\in V}(w_{ij}/\sigma_i^2)\mu_i\big)h_j\big\}}{2\cosh\big(c_j + \sum_{i\in V}(w_{ij}/\sigma_i^2)\mu_i\big)} \tag{12}$$

ここで μ_i は

$$\mu_i := \int_{-\infty}^{\infty} \sum_{\boldsymbol{h}} v_i T_1(\boldsymbol{v},\boldsymbol{h})\mathrm{d}\boldsymbol{v} = b_i + \sum_{j\in H} w_{ij}\gamma_j \tag{13}$$

により定義されており，本節の平均場近似下での可視変数 v_i の近似期待値となっている．また γ_j は

$$\gamma_j := \int_{-\infty}^{\infty} \sum_{\boldsymbol{h}} h_j T_1(\boldsymbol{v},\boldsymbol{h})\mathrm{d}\boldsymbol{v} \tag{14}$$

であり，本節の平均場近似下での隠れ変数 h_j の近似期待値を表している．式 (12) を式 (14) に代入することで

$$\gamma_j = \tanh\Big(c_j + \sum_{i\in V} \frac{w_{ij}}{\sigma_i^2}\mu_i\Big) \tag{15}$$

を得る．

式 (13)，式 (15) が可視変数と隠れ変数変数の両方に対して平均場近似を施した場合の平均場方程式である．式 (13)，式 (15) を逐次代入法で数値的に解くことにより式 (5)，式 (6) で示されている期待値の近似値を得ることができる．式 (13)，式 (15) の平均場近似を解くのに必要な計算量は $O(|V||H|)$ である．ここで $|V|$，$|H|$ はそれぞれ可視変数と隠れ変数の個数を表している．

3.2 隠れ変数のみ周辺分布に対する平均場近似

本節では二つ目の平均場近似法，即ち，周辺化により可視変数を消去した隠れ変数のみの周辺分布に対する平均場近似法を導出する．確率の積法則より式 (2) の GBRBM は次のように表すことができる．

$$P(\boldsymbol{v},\boldsymbol{h} \mid \boldsymbol{\theta}) = P(\boldsymbol{v} \mid \boldsymbol{h},\boldsymbol{\theta})P(\boldsymbol{h} \mid \boldsymbol{\theta}) \tag{16}$$

ここで $P(\boldsymbol{v} \mid \boldsymbol{h},\boldsymbol{\theta})$ は式 (3) で示されている条件付き分布である．また，$P(\boldsymbol{h} \mid \boldsymbol{\theta})$ は隠れ変数 $\boldsymbol{h}$ に関する周辺分布であり，

$$\begin{aligned} P(\boldsymbol{h} \mid \boldsymbol{\theta}) &= \int_{-\infty}^{\infty} P(\boldsymbol{v},\boldsymbol{h} \mid \boldsymbol{\theta})\mathrm{d}\boldsymbol{v} \\ &= \frac{1}{\mathcal{Z}(\boldsymbol{\theta})}\exp\Big(\sum_{j\in H}\beta_j h_j + \sum_{j<k\in H}\omega_{jk}h_j h_k\Big) \end{aligned} \tag{17}$$

と表される．ここで

$$\beta_j := c_j + \sum_{i\in V}\frac{b_i}{{\sigma_i}^2}w_{ij}, \quad \omega_{jk} := \sum_{i\in V}\frac{w_{ij}w_{ik}}{{\sigma_i}^2}$$

であり，式 (17) 中の $\sum_{j<k\in H}$ は隠れ変数の異なる全てのペアに関する和を表している．$\mathcal{Z}(\boldsymbol{\theta})$ は規格化定数である．式 (17) は完全結合グラフ上に定義された隠れ変数のみで構成されたボルツマンマシン [10] とみなすことができる．図 2 に図 1 の GBRBM に対する周辺分布 $P(\boldsymbol{h} \mid \boldsymbol{\theta})$ のグラフ構造を示す．

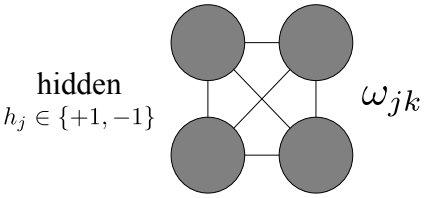

図 2　図 1 の GBRBM に対する周辺分布 $P(\boldsymbol{h} \mid \boldsymbol{\theta})$ のグラフ構造．隠れ変数が完全結合しているボルツマンマシンになっている．

テスト分布

$$T_2(\boldsymbol{v},\boldsymbol{h}) := P(\boldsymbol{v} \mid \boldsymbol{h},\boldsymbol{\theta})U(\boldsymbol{h}) \tag{18}$$

のように定義する．ここで $P(\boldsymbol{v} \mid \boldsymbol{h},\boldsymbol{\theta})$ は式 (3) で示されている条件付き分布であり，$U(\boldsymbol{h})$ は式 (8) で定義されている隠れ変数に関して因数分解された形の分布である．式 (18) のテスト分布は式 (16) の右辺の $P(\boldsymbol{h} \mid \boldsymbol{\theta})$ を $U(\boldsymbol{h})$ により近似した分布となっている．再び平均場近似の処方箋に従い，真の分布である式 (2) の GBRBM と式 (18) のテスト分布との間の KL 情報量

$$K_2[U] := \int_{-\infty}^{\infty}\sum_{\boldsymbol{h}} T_2(\boldsymbol{v},\boldsymbol{h})\ln\frac{T_2(\boldsymbol{v},\boldsymbol{h})}{P(\boldsymbol{v},\boldsymbol{h} \mid \boldsymbol{\theta})}\mathrm{d}\boldsymbol{v} \tag{19}$$

を定義し，この KL 情報量を $U(\boldsymbol{h})$ に関して最小化する．式 (19) は次のように変形することができる．

$$K_2[U] = \mathcal{F}_2[U] + \ln\mathcal{Z}(\boldsymbol{\theta}) \tag{20}$$

ここで

$$\begin{aligned} \mathcal{F}_2[U] := &-\sum_{j\in H}\beta_j\sum_{h_j} h_j u_j(h_j) + \sum_{j\in H}\sum_{h_j} u_j(h_j)\ln u_j(h_j) \\ &-\sum_{j<k\in H}\omega_{jk}\sum_{h_j,h_k} h_j h_k u_j(h_j)u_k(h_k) \end{aligned} \tag{21}$$

は変分自由エネルギーであり，この変分自由エネルギーを最小とする $U(\boldsymbol{h})$ が式 (19) の KL 情報量を最小とする $U(\boldsymbol{h})$ となる．$u_j(h_j)$ に関する規格化条件 $\sum_{h_j} u_j(h_j) = 1$ を条件としたラグランジュ未定乗数法により式 (21) を $u_j(h_j)$ に関して最小化すると，その極致条件より

$$u_j(h_j) = \prod_{j\in H}\frac{\exp\big\{\big(\beta_j + \sum_{k\in H_\partial(j)}\omega_{jk}m_j\big)h_j\big\}}{2\cosh\big(\beta_j + \sum_{k\in H_\partial(j)}\omega_{jk}m_j\big)} \tag{22}$$

を得る．ここで m_j は

$$m_j := \int_{-\infty}^{\infty}\sum_{\boldsymbol{h}} h_j T_2(\boldsymbol{v},\boldsymbol{h})\mathrm{d}\boldsymbol{v} \tag{23}$$

により定義されており，本節の平均場近似下での隠れ変数 h_j の近似期待値を表している．また，$H_\partial(j)$ は集合 H から要素 j を除いた集合である．式 (22) を式 (23) に代入することにより

$$m_j = \tanh\Big(\beta_j + \sum_{k\in H_\partial(j)}\omega_{jk}m_k\Big) \tag{24}$$

を得る．式 (24) は隠れ変数のみの周辺分布に対する平均場方程式である．式 (24) を逐次代入法により数値的に解くことにより，式 (6) で示されている隠れ変数の期待値の平均場近似による近似値を得ることができる．

式 (24) の平均場方程式から得られるのは隠れ変数の近似期待値のみであるので，可視変数については別途に計算する必要がある．可視変数に対する近似期待値は以下のようにして得ることができる．

$$\int_{-\infty}^{\infty} \sum_{\boldsymbol{h}} v_i T_2(\boldsymbol{v}, \boldsymbol{h}) \mathrm{d}\boldsymbol{v} = b_i + \sum_{j \in H} w_{ij} m_j$$

であるから，式 (5) で示されている可視変数の期待値は

$$\langle v_i \rangle \approx \tilde{\mu}_i := b_i + \sum_{j \in H} w_{ij} m_j \tag{25}$$

により近似される．式 (25) から分かるように，可視変数の近似期待値は平均場方程式 (24) の解の線形結合より直接求めることができる．

式 (24) と式 (25) を組み合わせることにより，式 (24) を

$$m_j = \tanh\Big(c_j + \sum_{i \in V} \frac{w_{ij}}{\sigma_i^2} \tilde{\mu}_i \ - \sum_{i \in V} \frac{w_{ij}^2}{{\sigma_i}^2} m_j\Big) \tag{26}$$

のように書き換えることが可能である．つまり，式 (24) と式 (25) の代わりに，式 (26) と式 (25) を連立して可視変数と隠れ変数の近似期待値を同時に解いても同じ解が得られる．式 (26) と式 (25) を解くのに必要な計算量は $O(|V||H|)$ であり，3.1 節で導出した平均場近似の計算量と同等である．

4 KL 情報量の観点での 2 手法の定性的比較

3.1 節と 3.2 節で GBRBM に対する異なる 2 通りの平均場近似を導出した．どちらも共に可視変数と隠れ変数の近似期待値を計算することができる手法となっており，計算量も同等である．では，どちらの手法の方がより精度の高い近似になっているのだろうか？直感的には平均場近似が施される変数の個数が少ない 3.2 節の方法の方がより良い手法になっていると期待される．本節では KL 情報量の観点から定性的に両者を比較し，この直感を後押しする結果を示す．

いま，n 個の確率変数 $\boldsymbol{x} = \{x_i \mid i \in \Omega = \{1, 2, \ldots, n\}\}$ に対する任意の結合分布 $P(\boldsymbol{x})$ を考える．ここで x_i は i ごとに離散・連続を含めて定義域が異なっていても構わない．変数番号の全体の集合 Ω を A と B に分割する: $A \subset \Omega$，$B = \Omega \setminus A$. そして A に割り当てられた変数を $\boldsymbol{x}_A$ で表し，同様に B に割り当てられた変数を $\boldsymbol{x}_B$ で表すこととする．

ここでテスト分布

$$T_{\mathrm{all}}(\boldsymbol{x}) := T_A(\boldsymbol{x}_A) T_B(\boldsymbol{x}_B) \tag{27}$$

を用意する．$T_A(\boldsymbol{x}_A)$ と $T_B(\boldsymbol{x}_B)$ はそれぞれ $\boldsymbol{x}_A$ と $\boldsymbol{x}_B$ に関するテスト分布である[*1]．式 (27) のテスト分布と結合分布 $P(\boldsymbol{x})$ の間の KL 情報量

$$K_{\mathrm{all}}[T_A, T_B] := \sum_{\boldsymbol{x}} T_{\mathrm{all}}(\boldsymbol{x}) \ln \frac{T_{\mathrm{all}}(\boldsymbol{x})}{P(\boldsymbol{x})} \tag{28}$$

を最小とする $T_{\mathrm{all}}(\boldsymbol{x})$ が結合分布 $P(\boldsymbol{x})$ の近似分布となる．ここで $\sum_{\boldsymbol{x}}$ は $\boldsymbol{x}$ の可能な全ての組み合わせに関する多重和を表しており，連続変数の場合は積分となる．他方，$\boldsymbol{x}_B$ についてのみ近似するためのテスト分布

$$T_{\mathrm{part}}(\boldsymbol{x}) := P(\boldsymbol{x}_A \mid \boldsymbol{x}_B) T_B(\boldsymbol{x}_B) \tag{29}$$

を用意する．ここで $P(\boldsymbol{x}_A \mid \boldsymbol{x}_B)$ は結合分布 $P(\boldsymbol{x})$ から得られる真の条件付き分布である．式 (29) のテスト分布と結合分布 $P(\boldsymbol{x})$ の間の KL 情報量

$$\begin{aligned} K_{\mathrm{part}}[T_B] &:= \sum_{\boldsymbol{x}} T_{\mathrm{part}}(\boldsymbol{x}) \ln \frac{T_{\mathrm{part}}(\boldsymbol{x})}{P(\boldsymbol{x})} \\ &= \sum_{\boldsymbol{x}_B} T_B(\boldsymbol{x}_B) \ln \frac{T_B(\boldsymbol{x}_B)}{P(\boldsymbol{x}_B)} \end{aligned} \tag{30}$$

を最小とする $T_{\mathrm{part}}(\boldsymbol{x})$ も結合分布 $P(\boldsymbol{x})$ に対する近似分布である．ここで $P(\boldsymbol{x}_B)$ は $P(\boldsymbol{x})$ の周辺分布である．

式 (28) と式 (30) の KL 情報量について次の定理が成り立つ．

> **定理**　式 (28) と式 (30) の KL 情報量に関して，任意の結合分布 $P(\boldsymbol{x})$ と任意の $\boldsymbol{x}_A$ と $\boldsymbol{x}_B$ の分割に対して不等式 $\min_{T_A, T_B} K_{\mathrm{all}}[T_A, T_B] \geq \min_{T_B} K_{\mathrm{part}}[T_B]$ が成り立つ．

証明　式 (28) の KL 情報量の最小値は

$$\begin{aligned} &\min_{T_A, T_B} K_{\mathrm{all}}[T_A, T_B] \\ &= \min_{T_B} \Big\{ \min_{T_A} \sum_{\boldsymbol{x}} T_A(\boldsymbol{x}_A) T_B(\boldsymbol{x}_B) \ln \frac{T_A(\boldsymbol{x}_A)}{P(\boldsymbol{x}_A \mid \boldsymbol{x}_B)} \\ &\quad + K_{\mathrm{part}}[T_B] \Big\} \end{aligned} \tag{31}$$

と変形できる．ここで式 (31) を最小とする $T_B(\boldsymbol{x}_B)$ を $T_B^\dagger(\boldsymbol{x}_B)$ とする．明らかに

$$K_{\mathrm{part}}[T_B^\dagger] \geq \min_{T_B} K_{\mathrm{part}}[T_B] \tag{32}$$

であるから，式 (31) と式 (32) より不等式

$$\begin{aligned} &\min_{T_A, T_B} K_{\mathrm{all}}[T_A, T_B] \\ &\geq \min_{T_A} \sum_{\boldsymbol{x}} T_A(\boldsymbol{x}_A) T_B^\dagger(\boldsymbol{x}_B) \ln \frac{T_A(\boldsymbol{x}_A)}{P(\boldsymbol{x}_A \mid \boldsymbol{x}_B)} + \min_{T_B} K_{\mathrm{part}}[T_B] \end{aligned} \tag{33}$$

が成り立つ．次に式 (33) の右辺第 1 項に注目する．$Y > 0$ に対する不等式 $\ln Y \leq Y - 1$ を用いると

$$\begin{aligned} &- \sum_{\boldsymbol{x}} T_A(\boldsymbol{x}_A) T_B^\dagger(\boldsymbol{x}_B) \ln \frac{P(\boldsymbol{x}_A \mid \boldsymbol{x}_B)}{T_A(\boldsymbol{x}_A)} \\ &\geq \sum_{\boldsymbol{x}} T_A(\boldsymbol{x}_A) T_B^\dagger(\boldsymbol{x}_B) \Big(1 - \frac{P(\boldsymbol{x}_A \mid \boldsymbol{x}_B)}{T_A(\boldsymbol{x}_A)}\Big) = 0 \end{aligned} \tag{34}$$

が成り立つ．式 (33) と式 (34) より $\min_{T_A, T_B} K_{\mathrm{all}}[T_A, T_B] \geq \min_{T_B} K_{\mathrm{part}}[T_B]$ が導かれる．□

[*1] $T_A(\boldsymbol{x}_A)$ と $T_B(\boldsymbol{x}_B)$ を変数ごとの因数分解の形で定義した場合が平均場近似のテスト分布である．

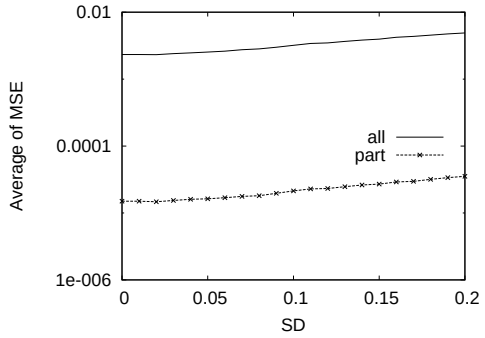

図 3　b_i の標準偏差 SD を変化させた場合の隠れ変数 h_j の期待値の MSE.

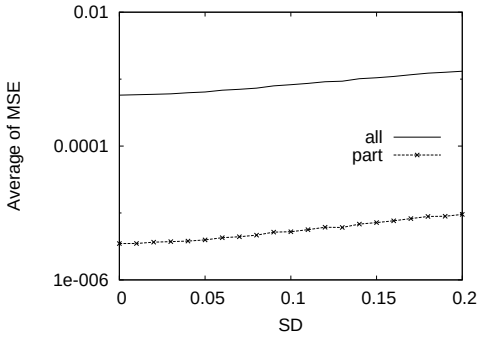

図 4　b_i の標準偏差 SD を変化させた場合の可視変数 v_i の期待値の MSE.

結合分布 $P(\boldsymbol{x})$ を式 (2)，$\boldsymbol{x}_A$ と $\boldsymbol{x}_B$ をそれぞれ可視変数と隠れ変数と読み替えることにより，上述の定理より直ちに次が導かれる．

> **系**　式 (9) と式 (19) の KL 情報量に関して，不等式 $\min_{Q,U} K_1[Q,U] \geq \min_U K_2[U]$ が成り立つ．

この系により 3.1 節の平均場近似より 3.2 節の平均場近似の方が KL 情報量をより小さくすることが分かる．これはつまり，3.2 節の平均場近似により求まる近似分布の方が KL 情報量の尺度で真の分布により近いということを意味している．

5　数値実験

本節では人工データに対する数値実験を用いて，3.1 節と 3.2 節で導出した 2 種類の平均場近似を定量的に比較する．

式 (2) の GBRBM において可視変数の個数を 24 個，隠れ変数の個数を 12 個とする．この GBRBM は比較的小さいサイズであるので，可視変数と隠れ変数の期待値を式 (5) と式 (6) に従い厳密に計算可能である．この GBRBM 上で 3.1 節と 3.2 節の 2 種類の平均場近似を実行し，厳密な期待値とその近似値との間の平均二乗誤差 (mean square error; MSE) により定量的な近似性能を測ることとする．以下の数値実験では全ての $i \in V$ で $\sigma_i^2 = 1$ としている．

図 3 と図 4 は $\boldsymbol{c}$，$\boldsymbol{w}$ を $\mathcal{N}(x \mid 0, 0.1^2)$ のガウス分布からそれぞれ独立に生成し，可視変数のバイアス $\boldsymbol{b}$ を $\mathcal{N}(x \mid 0, \mathrm{SD}^2)$ から生成した場合の隠れ変数と可視変数の MSE を SD に対してそれぞれ示している．プロットは 10000 回の試行の平均である．グラフ中の “all” は 3.1 節の平均場近似による結果であり，“part” は 3.2 節の平均場近似による結果を表している． 図 3 と図 4 より，3.2 節の方法の方が高精度の近似値となっていることが分かる．

図 5 と図 6 は $\boldsymbol{b}$，$\boldsymbol{w}$ を $\mathcal{N}(x \mid 0, 0.1^2)$ のガウス分布からそれぞれ独立に生成し，隠れ変数のバイアス $\boldsymbol{c}$ を $\mathcal{N}(x \mid 0, \mathrm{SD}^2)$ から生成した場合の隠れ変数と可視変数の MSE を SD に対してそれぞれ示している． 図 5 と図 6 も図 3 と図 4 と同様の傾向で 3.2 節の方法の方が高精度の近似値となっていることが確認できる．

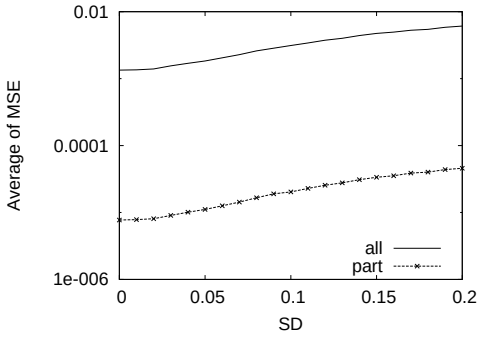

図 5　c_j の標準偏差 SD を変化させた場合の隠れ変数 h_j の期待値の MSE.

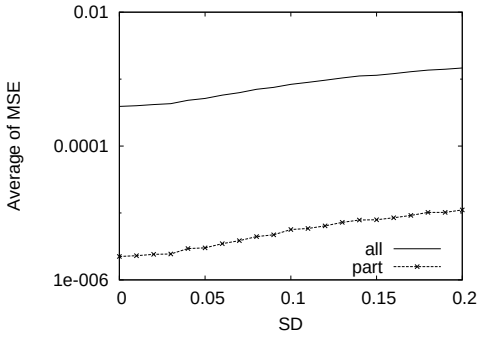

図 6　c_j の標準偏差 SD を変化させた場合の可視変数 v_i の期待値の MSE.

図 7 と図 8 は $\boldsymbol{b}$，$\boldsymbol{c}$ を $\mathcal{N}(x \mid 0, 0.1^2)$ のガウス分布からそれぞれ独立に生成し，可視変数と隠れ変数の間の結合 $\boldsymbol{w}$ を $\mathcal{N}(x \mid 0, \mathrm{SD}^2)$ から生成した場合の隠れ変数と可視変数の MSE を SD に対してそれぞれ示している． SD の増加に伴い性能が急速に落ちていくが，やはりここでも 3.2 節の方法の方が 3.1 節の方法を上回る近似精度を示している．

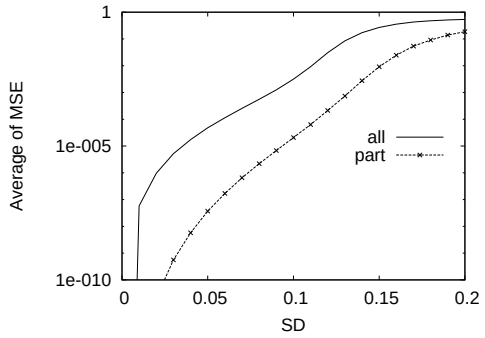

図 7　w_{ij} の標準偏差 SD を変化させた場合の隠れ変数 h_j の期待値の MSE 平均

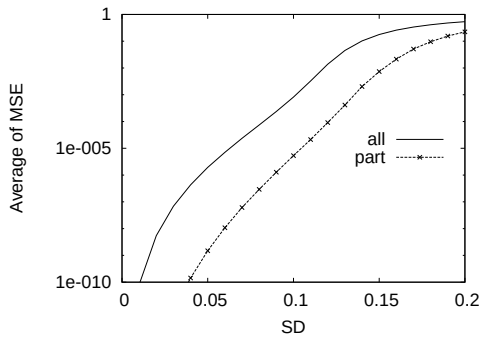

図 8　w_{ij} の標準偏差 SD を変化させた場合の可視変数 v_i の期待値の MSE 平均

6　まとめ

本論文では GBRBM に対する 2 種類の平均場近似を考え，両者を定性的視点・定量的視点の双方から比較した．その結果，3.2 節で導出した隠れ変数のみの周辺分布に対して平均場近似を適用する方法が，3.1 節の平均場近似よりも，定性的・定量的比較の両方で優れた近似計算アルゴリズムであることが分かった．特に，4 節で示した議論は KL 情報量を基礎とした定性的比較であるが，平均場近似の性能について数理的に考察できる例はほとんどないため，その意味では非常に貴重な例であると考えられる．

今後は TAP 近似 [11] などの平均場近似を超えたより高次の近似の GBRBM に対する適用が課題である．

謝辞

本研究の一部は文部科学省科学研究費補助金 (No.24700220, No.25280089) と CREST, JST の補助を得て行われたものである．

参考文献

[1] G. E. Hinton: Training products of experts by minimizing contrastive divergence, *Neural Computation*, Vol.14, pp.1771–1800, 2002.

[2] K. Cho, T. Raiko and A. Ilin: Gaussian-Bernoulli deep Boltzmann machine, *In Proceedings of the 2013 International Joint Conference on Neural Networks (IJCNN2013)*, pp.1–7, 2013.

[3] G. E. Hinton and R. Salakhutdinov: Reducing the Dimensionality of Data with Neural Networks, *Science*, vol.313, pp. 504–507, 2006.

[4] R. Salakhutdinov, A. Mnih and G. E. Hinton: Restricted Boltzmann machines for collaborative filtering, *In Proceedings of the 24th International Conference on Machine Learning (ICML2007)*, pp.791–798, 2007.

[5] K. Cho, A. Ilin and T. Raiko: Improved Learning of Gaussian-Bernoulli Restricted Boltzmann Machines, *In Proceedings of the 12th International Conference on Artificial Neural Networks (ICANN2011)*, pp.10–17, 2011.

[6] R. Salakhutdinov and G. E. Hinton: Deep Boltzmann machines, *In Proceedings of the Internation Conference on Artificial Intelligence and Statistics (AISTATS 2009)*, pp. 448-455, 2009.

[7] R. Salakhutdinov and G. E. Hinton: An efficient learning procedure for deep boltzmann machines, *Neural Computation*, vol.24, pp. 1967-2006, 2012.

[8] M. Opper and D. Saad: *Advanced Mean Field Methods — Theory and Practice —*, MIT Press, 2001.

[9] T. Tran, D. Phung and S. Venkatesh: Mixed-variate restricted Boltzmann machines, *In proceedings of the 3rd Asian Conference on Machine Learning (ACML2011)*, pp.213–229, 2011.

[10] D. H. Ackley, G. E. Hinton and T. J. Sejnowski: A Learning Algorithm for Boltzmann Machines, *Cognitive Science*, vol.9, pp.147–169, 1985.

[11] T. Plefka: Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model, *Journal of Physics A: Mathematical and General*, vol.15, pp. 1971–1978, 1982.