5J-04

深層ボルツマンマシンに対する高性能な平均場近似アルゴリズム

高橋　茶子 *
山形大学大学院理工学研究科

安田　宗樹 †
山形大学大学院理工学研究科

1　序論

深層ボルツマンマシン (deep Boltzmann machine; DBM) [1] は，制限ボルツマンマシン (restricted Boltzmann machine; RBM) [2] を多層に積み重ねて構成される確率的深層学習モデルである．

DBM を利用するにあたり困難な問題として，DBM に含まれる変数の期待値の計算がある．変数の期待値は，変数の数が増加するにつれ計算量が指数的に増加するため，実際には近似計算が必要とされる場面がほとんどである．本研究では，統計力学分野の近似法の一つである平均場近似 [3] を用い，DBM の変数の期待値を計算する 3 つのアルゴリズムを導出する．さらに，3 つのアルゴリズムの性能を数学的見地，数値的見地からそれぞれ比較する．

2　深層ボルツマンマシン

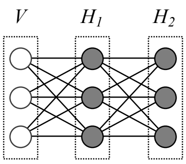

図 1: 本研究で扱う DBM．2 つの完全 2 部グラフからなり，層 V–H_1 は Gaussian-Gaussian RBM，層 H_1–H_2 は Gaussian-Bernoulli RBM であるとみなすことができる．

本研究では，DBM を図 1 に示すような確率的深層学習モデルであると定義する．DBM は 3 つの層を持ち，可視層 V は可視変数 $\boldsymbol{v} = \{v_i \in (-\infty, +\infty) \mid i \in V\}$，中間層 H_1 は隠れ変数 $\boldsymbol{h}^{(1)} = \{h_j^{(1)} \in (-\infty, +\infty) \mid j \in H_1\}$，隠れ層 H_2 は隠れ変数 $\boldsymbol{h}^{(2)} = \{h_k^{(2)} \in \{-1, +1\} \mid k \in H_2\}$ からそれぞれ構成される層である．可視変数は入出力データと直接関連付けられる変数であり，隠れ変数は入出力データとは関連しないシステムの内部変数として扱われる．

DBM のエネルギー関数を次のように定義する．

$$
\begin{aligned}
&E(\boldsymbol{v}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)} \mid \theta) \\
&= \sum_{i \in V} \frac{v_i^2}{2\sigma_i^2} + \sum_{j \in H_1} \frac{(h_j^{(1)})^2}{2 s_j^2} \\
&\quad - \sum_{i \in V} b_i v_i - \sum_{j \in H_1} c_j^{(1)} h_j^{(1)} - \sum_{k \in H_2} c_k^{(2)} h_k^{(2)} \\
&\quad - \sum_{i \in V} \sum_{j \in H_1} w_{ij}^{(1)} v_i h_j^{(1)} - \sum_{j \in H_1} \sum_{k \in H_2} w_{jk}^{(2)} h_j^{(1)} h_k^{(2)}
\end{aligned}
\tag{1}
$$

ここで，$\theta = \{\boldsymbol{b}, \boldsymbol{c}^{(1)}, \boldsymbol{c}^{(2)}, \boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \boldsymbol{\sigma}^2, \boldsymbol{s}^2\}$ はモデルパラメータである．$\boldsymbol{b}$，$\boldsymbol{c}^{(1)}$，$\boldsymbol{c}^{(2)}$ は可視変数 $\boldsymbol{v}$，隠れ変数 $\boldsymbol{h}^{(1)}$，隠れ変数 $\boldsymbol{h}^{(2)}$ に対するバイアスをそれぞれ表し，$\boldsymbol{w}^{(1)}$ は可視変数 $\boldsymbol{v}$ と隠れ変数 $\boldsymbol{h}^{(1)}$ の間の相互作用を，$\boldsymbol{w}^{(2)}$ は隠れ変数 $\boldsymbol{h}^{(1)}$ と隠れ変数 $\boldsymbol{h}^{(2)}$ の間の相互作用をそれぞれ表している．$\boldsymbol{\sigma}^2$ は可視変数 $\boldsymbol{v}$ の分散に関連するパラメータであり，$\boldsymbol{s}^2$ は隠れ変数 $\boldsymbol{h}^{(1)}$ の分散に関連するパラメータである．

式 (1) のエネルギー関数を用いて，DBM は

$$
P(\boldsymbol{v}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)} \mid \theta) = \frac{1}{Z(\theta)} \mathrm{e}^{-E(\boldsymbol{v}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)} \mid \theta)} \tag{2}
$$

の形で定義される．$Z(\theta)$ は分配関数である．

式 (2) の DBM の各変数の期待値は，次のように表される．

$$
\langle v_i \rangle = \iint \sum_{\boldsymbol{h}^{(2)}} v_i P(\boldsymbol{v}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)} \mid \theta) d\boldsymbol{v} d\boldsymbol{h}^{(1)} \tag{3}
$$

$$
\langle h_j^{(1)} \rangle = \iint \sum_{\boldsymbol{h}^{(2)}} h_j^{(1)} P(\boldsymbol{v}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)} \mid \theta) d\boldsymbol{v} d\boldsymbol{h}^{(1)} \tag{4}
$$

$$
\langle h_k^{(2)} \rangle = \iint \sum_{\boldsymbol{h}^{(2)}} h_k^{(2)} P(\boldsymbol{v}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)} \mid \theta) d\boldsymbol{v} d\boldsymbol{h}^{(1)} \tag{5}
$$

式 (3)–(5) の厳密計算には $O(\mathrm{e}^M)$ の計算量が必要であり，システム実装の際は現実的な時間内で計算を行うために近似計算を行う必要がある．ここで，各層に含まれる変数の個数を $|V| = M$，$|H_1| = \alpha M$，$|H_2| = \beta M$ としており，$\alpha > 0$，$\beta > 0$ である．

3　平均場近似アルゴリズム

隠れ層と可視層の 2 層で構成される RBM に対して平均場近似を行う場合，その周辺分布に対して平均場近似を適用すると高い精度で近似を行うことができる [4]．本節では [4] の方針を採用し，DBM に対する 3 つの平均場近似アルゴリズムを導出する．

確率分布 $P(\boldsymbol{x})$ の平均場近似分布は，$P(\boldsymbol{x})$ に含まれる全ての変数が統計的に独立であるとする平均場仮定を付加したテスト分布 $T(\boldsymbol{x})$ と $P(\boldsymbol{x})$ の間の Kullback-Leibler divergence (KLD) と呼ばれる量

$$
K = \sum_{\boldsymbol{x}} T(\boldsymbol{x}) \ln \frac{T(\boldsymbol{x})}{P(\boldsymbol{x})}
$$

を最小化することで得られる．KLD が小さいほど近似精度が高い平均場近似分布であるといえる．

I. DBM の全変数に対する平均場近似

式 (2) の DBM に含まれる全ての変数が統計的に独立であるとする平均場仮定を付加したテスト分布

$$
\begin{aligned}
&T_{\mathrm{I}}(\boldsymbol{v}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)}) \\
&\quad = Q(\boldsymbol{v})\, U^{(1)}(\boldsymbol{h}^{(1)})\, U^{(2)}(\boldsymbol{h}^{(2)}) \\
&\quad = \Big(\prod_{i \in V} q_i(v_i) \Big) \Big(\prod_{j \in H_1} u_j^{(1)}(h_j^{(1)}) \Big) \Big(\prod_{k \in H_2} u_k^{(2)}(h_k^{(2)}) \Big)
\end{aligned}
$$

と式 (2) との間の KLD を最小化すると，平均場近似分布

$$
T_{\mathrm{I}}^{\flat}(\boldsymbol{v}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)}) = Q_{\mathrm{I}}(\boldsymbol{v})\, U_{\mathrm{I}}^{(1)}(\boldsymbol{h}^{(1)})\, U_{\mathrm{I}}^{(2)}(\boldsymbol{h}^{(2)}) \tag{6}
$$

を得る．ここで，

$$
Q_{\mathrm{I}}(\boldsymbol{v}) = \prod_{i \in V} \mathcal{N}(v_i \mid \nu_i^{\flat}, \sigma_i^2), \tag{7}
$$

Better mean-field algorithm for deep Boltzmann machine

*Chako Takahashi; Graduate School of Science and Engineering, Yamagata University
†Muneki Yasuda; Graduate School of Science and Engineering, Yamagata University

$$U_{\mathrm{I}}^{(1)}(\boldsymbol{h}^{(1)}) = \prod_{j\in H_1} \mathcal{N}(h_j^{(1)} \mid m_j^{\flat}, s_j^2), \tag{8}$$

$$U_{\mathrm{I}}^{(2)}(\boldsymbol{h}^{(2)}) = \prod_{k\in H_2} \frac{\exp\left((c_k^{(2)} + \sum_{j\in H_1} w_{jk}^{(2)} m_j^{\flat}) h_k^{(2)}\right)}{2\cosh(c_k^{(2)} + \sum_{j\in H_1} w_{jk}^{(2)} m_j^{\flat})}$$

であり，$\mathcal{N}(x \mid \mu, \sigma)$ は平均 μ，分散 σ^2 のガウス分布を表す．また，$\boldsymbol{\nu}^{\flat}$ は式 (7) の近似分布における可視変数 $\boldsymbol{v}$ の期待値を，$\boldsymbol{m}^{\flat}$ は式 (8) の近似分布における隠れ変数 $\boldsymbol{h}^{(1)}$ の期待値を表している．

Ⅱ. DBM の隠れ変数 $\boldsymbol{h}^{(1)}$ と隠れ変数 $\boldsymbol{h}^{(2)}$ に対する平均場近似

確率の乗法定理を用い，式 (2) の DBM を $P(\boldsymbol{v}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)} \mid \theta) = P(\boldsymbol{v} \mid \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)}, \theta) P(\boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)} \mid \theta)$ と表すことができる．周辺分布 $P(\boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)} \mid \theta)$ の隠れ変数 $\boldsymbol{h}^{(1)}$，$\boldsymbol{h}^{(2)}$ に平均場仮定を付加したテスト分布

$$T_{\mathrm{II}}(\boldsymbol{v}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)}) = P(\boldsymbol{v} \mid \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)}, \theta)\, U^{(1)}(\boldsymbol{h}^{(1)})\, U^{(2)}(\boldsymbol{h}^{(2)})$$

と式 (2) との間の KLD を最小化すると，平均場近似分布

$$T_{\mathrm{II}}^{\natural}(\boldsymbol{v}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)}) = P(\boldsymbol{v} \mid \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)}, \theta)\, U_{\mathrm{II}}^{(1)}(\boldsymbol{h}^{(1)})\, U_{\mathrm{II}}^{(2)}(\boldsymbol{h}^{(2)}) \tag{9}$$

を得る．ここで，

$$U_{\mathrm{II}}^{(1)}(\boldsymbol{h}^{(1)}) = \prod_{j\in H_1} \mathcal{N}(h_j^{(1)} \mid m_j^{\natural}, s_j^2), \tag{10}$$

$$U_{\mathrm{II}}^{(2)}(\boldsymbol{h}^{(2)}) = \prod_{k\in H_2} \frac{\exp\left((c_k^{(2)} + \sum_{j\in H_1} w_{jk}^{(2)} m_j^{\natural}) h_k^{(2)}\right)}{2\cosh(c_k^{(2)} + \sum_{j\in H_1} w_{jk}^{(2)} m_j^{\natural})}$$

であり，$\boldsymbol{m}^{\natural}$ は式 (10) の近似分布における隠れ変数 $\boldsymbol{h}^{(1)}$ の期待値を表している．

Ⅲ. DBM の隠れ変数 $\boldsymbol{h}^{(2)}$ に対する平均場近似

前項と同様に，確率の乗法定理を用いると，式 (2) の DBM を $P(\boldsymbol{v}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)} \mid \theta) = P(\boldsymbol{v}, \boldsymbol{h}^{(1)} \mid \boldsymbol{h}^{(2)}, \theta) P(\boldsymbol{h}^{(2)} \mid \theta)$ のように表すことができる．周辺分布 $P(\boldsymbol{h}^{(2)})$ の隠れ変数 $\boldsymbol{h}^{(2)}$ のみに平均場仮定を付加したテスト分布

$$T_{\mathrm{III}}(\boldsymbol{v}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)}) = P(\boldsymbol{v}, \boldsymbol{h}^{(1)} \mid \boldsymbol{h}^{(2)}, \theta)\, U^{(2)}(\boldsymbol{h}^{(2)})$$

と式 (2) の間の KLD を最小化すると，平均場近似分布

$$T_{\mathrm{III}}^{\sharp}(\boldsymbol{v}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)}) = P(\boldsymbol{v}, \boldsymbol{h}^{(1)} \mid \boldsymbol{h}^{(2)}, \theta)\, U_{\mathrm{III}}^{(2)}(\boldsymbol{h}^{(2)}) \tag{11}$$

を得る．ここで，

$$U_{\mathrm{III}}^{(2)}(\boldsymbol{h}^{(2)}) = \prod_{k\in H_2} \frac{\exp\left((B_k^{(2)} + \sum_{l\in\partial(k)} \Omega_{kl} M_l^{\sharp}) h_k^{(2)}\right)}{2\cosh(B_k^{(2)} + \sum_{l\in\partial(k)} \Omega_{kl} M_l^{\sharp})} \tag{12}$$

であり，$\boldsymbol{B}^{(2)}$ は $P(\boldsymbol{h}^{(2)})$ における隠れ変数 $\boldsymbol{h}^{(2)}$ のバイアスを，$\boldsymbol{\Omega}$ は隠れ変数 $\boldsymbol{h}^{(2)}$ 同士の相互作用を表す．また，$\boldsymbol{M}^{\sharp}$ は式 (12) の近似分布における隠れ変数 $\boldsymbol{h}^{(2)}$ の期待値を表している．

4　アルゴリズムの比較

数学的比較

KLD に関する定理 [4] を拡張し，次の系を導く．

系　アルゴリズム I，Ⅱ，Ⅲ で得られる式 (6)，(9)，(11) の平均場近似分布と式 (2) の間の KLD をそれぞれ $K_{\mathrm{I}}, K_{\mathrm{II}}, K_{\mathrm{III}}$ とする．$K_{\mathrm{I}}, K_{\mathrm{II}}, K_{\mathrm{III}}$ に対して，不等式

$$K_{\mathrm{I}} \geq K_{\mathrm{II}} \geq K_{\mathrm{III}} \tag{13}$$

が常に成り立つ．

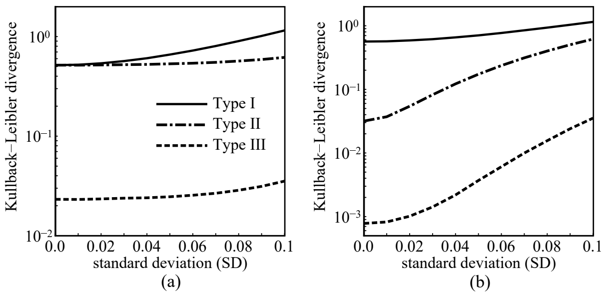

図 2: 各アルゴリズムで得る平均場近似分布と式 (2) の DBM の間の KLD の値．

不等式 (13) は，式 (11)，式 (9)，式 (6) の順に近似精度が優れていることを示している．すなわち，DBM に対して可能な限りの周辺化を行っているアルゴリズム Ⅲ が，3 つの中では最も高精度な近似手法であるといえる．

数値的比較

式 (2) の各変数の個数を $|V| = |H_1| = |H_2| = 10$ とする．図 2 (a) は相互作用 $\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}$ をそれぞれガウス乱数 $\mathcal{N}(0, \mathrm{SD}^2), \mathcal{N}(0, 0.1^2)$ から生成した場合の，図 2 (b) は $\mathcal{N}(0, 0.1^2), \mathcal{N}(0, \mathrm{SD}^2)$ から生成した場合の各アルゴリズムの KLD の値を示している．その他のモデルパラメータについて，$\boldsymbol{b}, \boldsymbol{c}^{(1)}, \boldsymbol{c}^{(2)}$ は $\mathcal{N}(0, 0.1^2)$ から生成しており，$\sigma_i^2 = s_j^2 = 1$ としている．グラフ中の “Type I” は K_{I} の値を，“Type Ⅱ” は K_{II} の値を，“Type Ⅲ” は K_{III} の値をそれぞれ表す．図 2 (a)，図 2 (b) の結果はどちらも数学的比較における不等式 (13) を満たしており，アルゴリズム Ⅲ で得られる式 (11) の平均場近似分布が最も近似精度の高い方法であることが確認できる．

5　結論

本研究では，3 つの層を持つ DBM に対する高精度な平均場近似アルゴリズムを導出した．近似の対象とする変数の数を周辺化によって可能な限り少なくすることで，近似性能の向上を実現することができた．本研究で取り扱った DBM は，実際のシステムへの応用には適用範囲が限られる．そのため，より汎用的なモデルに対する高性能な近似手法の開発が今後の課題の一つである．

謝辞

本研究の一部は，文部科学省科学研究費補助金 (15K00330, 15H03699) と JST-CREST の補助を得て行われたものである．

参考文献

[1] R. Salakhutdinov and G. E. Hinton: Deep Boltzmann Machines, In Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS2009), pp.448–455, 2009.

[2] G. E. Hinton: Training Products of Experts by Minimizing Contrastive Divergence, Neural Computation, Vol.14, pp.1771–1800, 2002.

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

[4] C. Takahashi and M. Yasuda: Mean-Field Inference on Gaussian Restricted Boltzmann Machine, Journal of the Physical Society of Japan, Vol.85, No.3, Article ID: 034001, pp.1–6, 2016.