2K-02

適応 TAP 近似を用いたガウス型制限ボルツマンマシンの学習

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　はじめに

restricted Boltzmann machine (RBM) は深層学習 (deep learning) の重要な要素であり，そのため，RBM の学習法や RBM を用いた推論アルゴリズムなどが盛んに研究されている．現実のアプリケーションはしばしば連続値のデータを扱うため，連続値の入出力データを扱える形に拡張された RBM として Gaussian RBM (GRBM) が知られている [1, 2].

RBM に対する厳密な学習アルゴリズムは，組み合わせ爆発の問題を抱えた期待値計算を必要とするため，実装においては何らかの近似学習アルゴリズムが必要となる．RBM に対する近似学習法として，contrastive divergence (CD) 法 [3] が広く知られている．CD 法は計算困難な期待値計算をデータとギブスサンプリングを利用して近似するため，期待値計算の計算量がデータの数に依存して大きくなってしまうという問題がある．そのため，大規模データを扱う際には，データ数に依存しない計算量で行うことのできるアルゴリズムが必要である．本稿では，統計力学における平均場法の一つである適応 TAP 近似 [4, 5] を用いた学習法を導出し，CD 法と性能を比較する．

2　ガウス型制限ボルツマンマシン

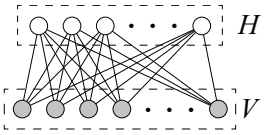

図 1　GRBM．可視層 V と隠れ層 H の 2 層からなる完全 2 部グラフ上に定義される．

GRBM は，図 1 のように完全 2 部グラフ上に定義される確率的学習モデルである．可視層は可視変数 $\boldsymbol{v} = \{v_i \in (-\infty, +\infty) \mid i \in V\}$ から構成される層であり，隠れ層は隠れ変数 $\boldsymbol{h} = \{h_j \in \mathcal{X} \mid j \in H\}$ から構成される層である．ここで，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．

本稿では，GRBM の一つである Gaussian-Bernoulli restricted Boltzmann machine (GBRBM) [1] に注目する．GBRBM では，隠れ変数 $\boldsymbol{h}$ が 2 値をとる離散確率変数となる．本稿では $h_j \in \mathcal{X} = \{-1, +1\}$ の場合について考える．

文献 [1] で提案されているエネルギー関数を変形して，次のような新しいエネルギー関数を定義する．

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

ここで，$\theta = \{\boldsymbol{b}, \boldsymbol{c}, \boldsymbol{w}, \boldsymbol{\sigma}^2\}$ はモデルパラメータである．$\boldsymbol{b}$ と $\boldsymbol{c}$ はそれぞれ可視変数と隠れ変数に対するバイアス項，$\boldsymbol{w}$ は可視変数と隠れ変数の間の相互作用項である．$\boldsymbol{\sigma}^2$ は可視変数の分散に関連する項である．式 (1) のエネルギー関数は，本質的には文献 [1] のエネルギー関数と同等である．式 (1) を用いると，定義される GBRBM が指数分布族となる．

式 (1) のエネルギー関数を用いて，GBRBM は

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

の形で定義される．$Z(\theta)$ は分配関数である．

N 個のデータのセット $\mathcal{D} := \{\boldsymbol{d}^{(\mu)} \in (-\infty, +\infty)^{|V|} \mid \mu = 1, 2, \cdots, N\}$ を用いた GBRBM の学習は，対数尤度関数

$$l_{\mathcal{D}}(\theta) = \big\langle \ln P(\boldsymbol{v} \mid \theta) \big\rangle_{\mathcal{D}} \tag{3}$$

の θ に関する最大化で達成される．ここで，$P(\boldsymbol{v} \mid \theta)$ は式 (2) の周辺分布であり，$\langle \cdots \rangle_{\mathcal{D}}$ はデータセット $\mathcal{D}$ に対する標本平均を表している．式 (3) の対数尤度関数の各パラメータに対する勾配は

$$\nabla_{b_i} l_{\mathcal{D}}(\theta) = \langle v_i \rangle_{\mathcal{D}} - \langle v_i \rangle \tag{4}$$

$$\nabla_{c_j} l_{\mathcal{D}}(\theta) = \big\langle \tanh \lambda_j(\boldsymbol{v}) \big\rangle_{\mathcal{D}} - \langle h_j \rangle \tag{5}$$

$$\nabla_{w_{ij}} l_{\mathcal{D}}(\theta) = \big\langle v_i \tanh \lambda_j(\boldsymbol{v}) \big\rangle_{\mathcal{D}} - \langle v_i h_j \rangle \tag{6}$$

$$\nabla_{\sigma_i^2} l_{\mathcal{D}}(\theta) = \frac{\langle v_i^2 \rangle_{\mathcal{D}}}{2\sigma_i^4} - \frac{\langle v_i^2 \rangle}{2\sigma_i^4} \tag{7}$$

となる．式 (5), (6) において，$\lambda_j(\boldsymbol{v}) = c_j + \sum_{i\in V} w_{ij} v_i$ である．また，$\langle \cdots \rangle$ は式 (2) の GBRBM の対応する変数の期待値を表している．式 (4)–(7) の勾配を用いた勾配上昇法により，GBRBM の学習は達成される．しかしながら GBRBM の期待値計算の計算量は変数の数に対して指数的に増加してしまうため，近似無しでの実装はほとんど不可能である．

Learning algorithm for Gaussian restricted Boltzmann machine using adaptive TAP approximation

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

CD 法では，データ $\mathcal{D}$ を初期値とした 1 回の層間ギブスサンプリングにより得られたサンプル点の標本平均でこの計算困難の期待値 (式 (4)–(7) の勾配の第 2 項) を近似する．CD 法における勾配の第 2 項の計算量は $O(NM^2)$ である．ここで，ある定数 α に対して $|V| = M$，$|H| = \alpha M$ としている．

3 適応 TAP 近似を用いたパラメータ学習法

本節では，平均場近似の一種である適応 TAP 近似 [4] を用いた新しい学習アルゴリズムを提案する．GBRBM に対して平均場近似を行う場合，GBRBM の隠れ層のみからなる周辺分布 $P(\boldsymbol{h} \mid \theta)$ 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ことが知られている [2] ため，本稿でもその方針を採用することにする．

式 (2) の隠れ層の周辺分布は

$$P(\boldsymbol{h} \mid \theta) \propto \exp\Big(\sum_{j\in H} \beta_j h_j + \sum_{j<k\in H} \omega_{jk} h_j h_k\Big) \tag{8}$$

と表される．ここで，$\boldsymbol{\beta}$, $\boldsymbol{\omega}$ は θ によって定義されるパラメータである．式 (4)，(6)，(7) の勾配は，一般的に以下のように書き換えることができる．

$$\nabla_{b_i} l_{\mathcal{D}}(\theta) = \langle v_i \rangle_{\mathcal{D}} - \sigma_i^2 \Big(b_i + \sum_{j\in H} w_{ij} \langle h_j \rangle\Big) \tag{9}$$

$$\nabla_{w_{ij}} l_{\mathcal{D}}(\theta) = \big\langle v_i \tanh \lambda_j(\boldsymbol{v}) \big\rangle_{\mathcal{D}} - \sigma_i^2 \Big(b_i \langle h_j \rangle + \sum_{l\in H} w_{il} \langle h_j h_l \rangle\Big) \tag{10}$$

$$\nabla_{\sigma_i^2} l_{\mathcal{D}}(\theta) = \frac{\langle v_i^2 \rangle_{\mathcal{D}}}{2\sigma_i^4} - \frac{1}{2\sigma_i^2} - \frac{b_i^2}{2} - b_i \sum_{j\in H} w_{ij} \langle h_j \rangle - \frac{1}{2} \sum_{j\in H} \sum_{l\in H} w_{ij} w_{il} \langle h_j h_k \rangle \tag{11}$$

この表式を用いると，対数尤度関数の勾配の第 2 項を隠れ変数の 1 次，2 次の期待値のみで表すことができる．

適応 TAP 近似 [4, 5] を用いると，隠れ変数の 1 次，2 次の期待値の近似値は

$$\langle h_j \rangle \approx \tanh\Big(\beta_j + \sum_{k\in H\setminus\{j\}} \omega_{jk} m_k - \Lambda_j m_j\Big) \tag{12}$$

$$\langle h_j h_k \rangle \approx \chi_{jk} + \langle h_j \rangle \langle h_k \rangle \tag{13}$$

により得られる．ここで，

$$m_j = \tanh\Big(\beta_j + \sum_{k\in H\setminus\{j\}} \omega_{jk} m_k - \Lambda_j m_j\Big),$$

$$\Lambda_j = \frac{1}{1-m_j^2} \sum_{k\in H\setminus\{j\}} \omega_{jk} \chi_{kj},$$

$$\chi_{jk} = \frac{1-m_j^2}{1+\Lambda_j(1-m_j^2)} \Big(\delta_{j,k} + \sum_{l\in H\setminus\{j\}} \omega_{jl} \chi_{lk}\Big)$$

であり，$\delta_{j,k}$ はクロネッカーのデルタである．本手法による勾配の第 2 項の計算量は $O(M^3)$ である．したがって，$N \gg M$ の場合，すなわちデータ数が膨大である場合は $O(NM^2) \gg O(M^3)$ となる．

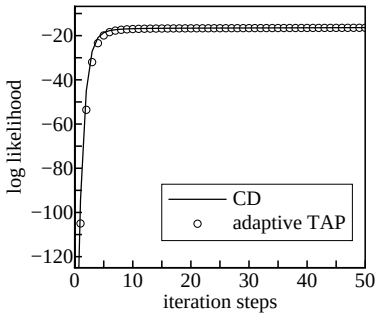

図 2　データ生成モデルの相互作用項 $\boldsymbol{w}$ を $\mathcal{N}(\boldsymbol{w} \mid 0, \frac{5}{\sqrt{|H|}})$ で発生させた場合の対数尤度関数の上昇．

4 数値実験

式 (2) の GBRBM において可視変数の個数を $|V| = 12$，隠れ変数の個数を $|H| = 20$ とする．図 2 はデータ生成モデルのパラメータ $\boldsymbol{b}$, $\boldsymbol{c}$, $\boldsymbol{\sigma}^2$ を平均 0，標準偏差 1 のガウス分布 $\mathcal{N}(\boldsymbol{x} \mid 0, 1)$ からそれぞれ独立に生成し，相互作用項 $\boldsymbol{w}$ を $\mathcal{N}(\boldsymbol{w} \mid 0, 5/\sqrt{|H|})$ から生成した場合の対数尤度関数の上昇を，勾配上昇法の反復回数に対しそれぞれ示している．観測データセットは 10000 セット，学習率は 0.01 に設定した．グラフ中の “CD” は CD 法による結果を，“adaptive TAP” は 3 節で導出した提案手法による結果を表している．図 2 より，提案手法は CD 法と同等の性能であるといえる．

5 まとめ

本稿では GBRBM の学習に適応 TAP 近似を用いた新しい学習法を提案し，CD 法との比較を行った．その結果，提案手法と CD 法はほとんど同等の性能を示した．提案手法は計算量がデータ数に依らないため，大規模データを扱う際に非常に有用な方法である．実際の大規模データを用いた情報処理システムの構築が今後の課題の一つとして挙げられる．

謝辞

本研究の一部は文部科学省科学研究費補助金 (15K00330, 25280089, 15H03699) と CREST, JST の補助を得て行われたものである．

参考文献

[1] K. Cho, A. Ilin and T. Raiko: Improved Learning of Gaussian-Bernoulli Restricted Boltzmann Machines, ICANN2011, pp.10–17, 2011.
[2] C. Takahashi and M. Yasuda: Mean-Field Inference on Gaussian Restricted Boltzmann Machine, http://arxiv.org/abs/1512.00927, 2015.
[3] G. E. Hinton: Training products of experts by minimizing contrastive divergence, Neural Computation, vol.14, pp.1771–1800, 2002.
[4] M. Opper and O. Winther: Adaptive and self-averaging Thouless-Anderson-Palmer mean-field theory for probabilistic modeling, Physical Review E, vol.64, 056131, 2001.
[5] M. Yasuda and K. Tanaka: Susceptibility propagation by using diagonal consistency, Physical Review E, vol. 87, 012134, 2013.