2J-02

深層ボルツマンマシンにおける学習アルゴリズムの改良

勝亦 利宗 †　　安田 宗樹 ‡
山形大学大学院理工学研究科 †　山形大学大学院理工学研究科 ‡

1　はじめに

深層ボルツマンマシン (deep Boltzmann machine (DBM))[1] は確率的に信号が伝わる深層ニューラルネットワークモデルであり, その学習は事前学習と最尤学習 (ファインチューニング) の二段階のプロセスによって行われる. DBM の最尤学習はその対数尤度関数の最大化によって達成するが, そのアルゴリズムの中には指数時間の計算量を必要とするモデルの期待値計算を含むため, 単純には学習を行うことができない. DBM の提案論文 [1] ではこの問題に対し, モデル期待値を近似した値で学習する方法を提案している. しかし, より良い期待値近似とそれを利用した学習性能に関する研究は少なく, さらなる研究が必要とされている.

そこで本稿では, マルコフ確率場における期待値近似の手法である空間モンテカルロ積分 (spatial Montecarlo integration (SMCI))[2, 3] 法を DBM の最尤学習に適用し, 新たな学習法を提案する. 提案法ではより厳密値に近い期待値近似が可能となり, 性能の高い学習を行うことができる.

2　深層ボルツマンマシン

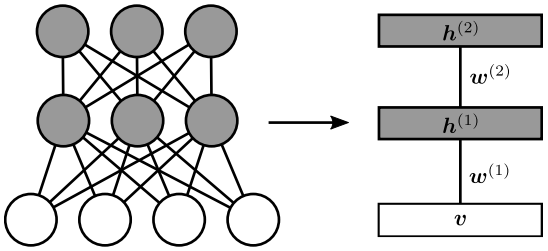

図 1　DBM のグラフ構造.

DBM のグラフ構造を図 1 に示す. DBM は V 個の素子を持つ可視層 $\boldsymbol{v} = \{v_i \in \{-1, +1\} \mid 1 \leq i \leq V\}$ と, r 番目の層が H_r 個の素子を持つ L 個の隠れ層 $\boldsymbol{h} = \{\boldsymbol{h}^{(r)} \mid 1 \leq r \leq L\}$, $\boldsymbol{h}^{(r)} = \{h_j^{(r)} \in \{-1, +1\} \mid 1 \leq j \leq H_r\}$ からなる多層構造を持っている. 可視層はデータの入力に使う層であるため, 素子数 V は入力データの次元数に対応する. 一方で, 隠れ層はデータから学習する特徴量の値を取る層であり, H_r は入力データに対して適切な値を任意に設定する.

DBM の確率分布関数は, エネルギー関数

$$E(\boldsymbol{v}, \boldsymbol{h}; \theta) = -\sum_{i=1}^{V} b_i v_i - \sum_{r=1}^{L}\sum_{j=1}^{H_r} c_j^{(r)} h_j^{(r)} - \sum_{i=1}^{V}\sum_{j=1}^{H_1} w_{ij}^{(1)} v_i h_j^{(1)} - \sum_{r=2}^{L}\sum_{j=1}^{H_{r-1}}\sum_{k=1}^{H_r} w_{jk}^{(r)} h_j^{(r-1)} h_k^{(r)} \quad (1)$$

を使って,

$$P(\boldsymbol{v}, \boldsymbol{h} \mid \theta) = \frac{1}{Z(\theta)} \exp\left(-E(\boldsymbol{v}, \boldsymbol{h}; \theta)\right) \quad (2)$$

と定義される. ここで, $\theta = \{\boldsymbol{b}, \boldsymbol{c}^{(1)}, \ldots, \boldsymbol{c}^{(L)}, \boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(L)}\}$ はモデルのパラメータである. それぞれ, $\boldsymbol{b}$ と $\boldsymbol{c}$ はバイアスパラメータ, $\boldsymbol{w}$ は層間の結合パラメータを表す. また, $Z(\theta)$ は分配関数である.

次に, DBM の最尤学習について説明する. 前述の通り, DBM の最尤学習は対数尤度関数の最大化によって達成される. 学習データを $D = \{\mathbf{v}_i^{(\mu)} \in \{-1, +1\}^V \mid \mu = 1, \ldots, N\}$ としたときの DBM の対数尤度関数 $l_D(\theta)$ は,

$$l_D(\theta) = \frac{1}{N}\sum_{\mu=1}^{N} \ln \sum_{\boldsymbol{h}} P(\mathbf{v}^{(\mu)}, \boldsymbol{h} \mid \theta) \quad (3)$$

となる. ここで, $\sum_{\boldsymbol{h}}$ は隠れ層 $\boldsymbol{h}$ に関する全ての実現値の組み合わせの総和を表す. 式 (3) を勾配法によって最大化するため, パラメータ θ に関するそれぞれの勾配が必要となる. 例として, パラメータ $w_{jk}^{(r)}$ に関する $l_D(\theta)$ の勾配を求めると,

$$\frac{\partial l_D(\theta)}{\partial w_{jk}^{(r)}} = \frac{1}{N}\sum_{\mu=1}^{N} \mathbb{E}_{\mathrm{cl}}[h_j^{(r-1)} h_k^{(r)} \mid \mathbf{v}^{(\mu)}, \theta] - \mathbb{E}_{\mathrm{fr}}[h_j^{(r-1)} h_k^{(r)} \mid \theta] \quad (4)$$

となる. ここで, $\mathbb{E}_{\mathrm{cl}}[\cdot]$ は可視層をデータで固定した DBM の期待値, $\mathbb{E}_{\mathrm{fr}}[\cdot]$ はモデルの期待値である. 他のパラメータに対する勾配も式 (4) と同様に, $\mathbb{E}_{\mathrm{cl}}[\cdot]$ と $\mathbb{E}_{\mathrm{fr}}[\cdot]$ の差がパラメータの勾配となる. 式 (4) に含まれる期待値項は

$$\mathbb{E}_{\mathrm{fr}}[h_j^{(r-1)} h_k^{(r)} \mid \theta] = \sum_{\boldsymbol{v}}\sum_{\boldsymbol{h}} h_j^{(r-1)} h_k^{(r)} P(\boldsymbol{v}, \boldsymbol{h} \mid \theta) \quad (5)$$

$$\mathbb{E}_{\mathrm{cl}}[h_j^{(r-1)} h_k^{(r)} \mid \boldsymbol{v}^{(\mu)}, \theta] = \sum_{\boldsymbol{h}} h_j^{(r-1)} h_k^{(r)} P(\boldsymbol{h} \mid \boldsymbol{v}^{(\mu)}, \theta) \quad (6)$$

である. $\sum_{\boldsymbol{v}}$ も $\sum_{\boldsymbol{h}}$ と同じく, 可視層 $\boldsymbol{v}$ に関する全ての実現値の総和を表す. これら総和の計算には指数時間の計算量を要するため, 提案論文 [1] では, 式 (5) をモンテカルロ積分, 式

Effective Learning Algorithm for Deep Boltzmann Machine

† Tomu Katsumata, Graduate School of Science and Engineering, Yamagata University
‡ Muneki Yasuda, Graduate School of Science and Engineering, Yamagata University

(6) を平均場近似によって近似し学習を行う手法が提案されている.

3 空間モンテカルロ積分法

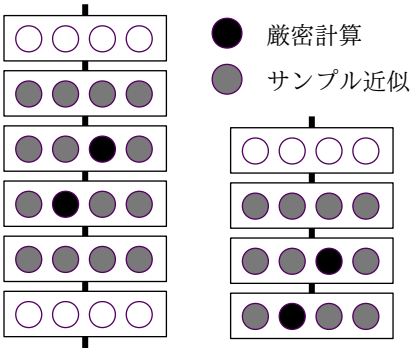

図 2 DBM における 2 変数の 1-SMCI 法. 左の図はモデル内部の層, 右の図はモデル末端の層に対する変数の割り当てを表す.

本節では提案手法である SMCI[2, 3] 法について説明する. DBM における SMCI 法の適用を図 2 に示す. SMCI 法は厳密和と, サンプリング点の標本平均を使ったサンプル近似を組み合わせた近似手法である. 図 2 では黒丸が厳密和を行う変数, グレーがサンプリング点に対応する変数である. SMCI 法では, サンプリング点を条件とした条件付き確率分布に対し, 期待値を近似したい変数のグラフ構造の周辺で限定的な厳密和を行う. 厳密和を行う範囲はグラフ構造の許す限り拡大することができ, 拡大するほど近似値は厳密値に近くなることが証明されている [2]. また, 厳密和を行う変数が期待値を近似したい変数のみの場合を 1-SMCI と呼ぶ. 1-SMCI は SMCI 法における最小構成の近似法である.

4 数値実験

本節では提案法の妥当性を数値実験によって検証する. この実験では, パラメータをランダムに決定した生成 DBM を生成モデルとし, そこから生成した $N = 500$ の人工データを学習 DBM が学習する. 今回の実験では, 従来法, 1-SMCI による期待値近似を使い学習する提案法, 厳密値による学習, の 3 つの方法で学習を行う. その学習の結果得られた DBM が, 生成 DBM にどれだけ近いものであるかをカルバックライブラー情報量 (Kullback-Leibler divergence (KLD)) の値で比較し, さらにデータの学習度を対数尤度の値で比較する. KLD は 2 つの確率分布間の差を表す値であるため小さいほど良く, 一方で, 対数尤度はモデルがデータにどれだけ則しているかの度合いを表す値であるため, 高いほど良い結果であると言える. また, 厳密値による学習と KLD・対数尤度の計算を可能とするため, 生成・学習 DBM ともに素子数がすべて 5 個の層を可視層含み 3 つ持つ DBM を実験に使用した. さらに, 学習にあたって勾配法には Adamax [5] を使用し, 学習 DBM の初期パラメータは Xavier の初期値 [4] で初期化を行った.

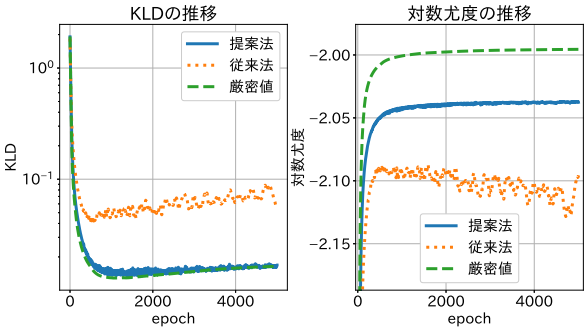

図 3 提案法 (実線), 従来法 (点線)[1], 厳密値による学習 (破線) の KLD の推移. プロットは 1,000 回の試行の平均である.

実験の結果を図 3 に示す. 図 3 より, 提案法による学習は, 従来法よりも KLD が低減し, 対数尤度は上昇していることが分かる. さらに, KLD において提案法は厳密値による学習に近い推移を見せている. これらのことから, 提案法は従来法よりも精度の高い学習を行っているといえる.

5 まとめ

本稿では DBM の期待値近似に SMCI 法を用いた新たな手法を提案し, 従来法との比較を行った. 数値実験の結果, SMCI 法を用いた学習法は従来法よりも厳密値に近い学習を達成できることが示された.

本稿は DBM の最尤学習のみに焦点を当てた研究であるため, 事前学習のアルゴリズムに提案法を組み合わせたときも同様に, 学習精度の向上に繋がるかの検証が今後の課題として挙げられる.

6 謝辞

本研究は科研費 (18K11459, 18H03303), JST-CREST (JPMJCE1312) 及び JST COI プログラム (JPMJCE1312) の助成を受けたものである。

参考文献

[1] R. Salakhutdinov and G. Hinton, "Deep Boltzmann Machines", Artificial intelligence and statistics, pp.448-455, 2009.

[2] M. Yasuda, "Monte Carlo Integration Using Spatial Structure of Markov Random Field", Journal of the Physical Society of Japan, vol. 84, no. 3, p. 034001, 2015.

[3] M. Yasuda and K. Uchizawa, "A Generalization of Spatial Monte Carlo Integration", arXiv:2009.02165, 2020.

[4] X. Glorot and Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, pp.249–256, 2010.

[5] D. Kingma and J. Ba: Adam: a method for stochastic optimization, Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), 2015.