1K-03

スパース正則化した隠れ変数を持つ制限ボルツマンマシン

横山 悠貴 †　　安田 宗樹 †
† 山形大学大学院 理工学研究科

1 はじめに

確率的ニューラルネットワークの一つである制限ボルツマンマシン [2] は最尤法に基づき学習を行うが, 隠れ変数が多いほど過学習を起こしてしまい汎化能力が低下してしまう. 特に学習データが少ない時に過学習を起こすことで機械学習本来の目的である未知のデータの予想精度が悪くなってしまう. 先行研究 [1] では隠れ変数にスパース正則化を行うことで Contrastive Divergence 法 [3] による近似計算を用いた学習において過学習を抑制をすることを確認した. 本稿では厳密計算による制限ボルツマンマシン学習においても過学習を抑制することを示す.

2 制限ボルツマンマシン

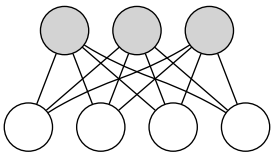


図 1: 制限ボルツマンマシンのグラフ構造

制限ボルツマンマシンは, 図 1 の完全二部グラフ上で定義される確率モデルである. グラフ上のノードにはそれぞれ確率変数が対応しており, 下層のノードに対応する変数を可視変数 $\boldsymbol{v}$, 上層のノードに対応する変数を隠れ変数 $\boldsymbol{h}$ と呼ぶ. また, 可視変数のインデックス集合を V, 隠れ変数のインデックス集合を H と定義する. 確率変数について, 可視変数を $\boldsymbol{v} = \{v_i \in \{-1,+1\} \mid i \in V\}$, 隠れ変数を $\boldsymbol{h} = \{h_j \in \mathcal{X}(s)\} \mid j \in H\}$ と定義し, $\mathcal{X}(s)$ は

$$\mathcal{X}\left(\mathrm{S}\right) = \left\{ \frac{2\left(k-1\right)}{\mathrm{S}-1} - 1 \middle| k = 1,2,...,\mathrm{S} \right\}$$

と定義した集合である. これは区間 $[-1,1]$ を S 値に量子化した点の集合を表す. 具体的には

$$\mathcal{X}(2) = \{-1,1\}, \qquad \mathcal{X}(3) = \{-1,0,1\},$$
$$\mathcal{X}(4) = \left\{-1,-\frac{1}{3},\frac{1}{3},1\right\}, \quad \mathcal{X}(5) = \left\{-1,-\frac{1}{2},0,\frac{1}{2},1\right\},$$
$$\mathcal{X}(\infty) = [-1,+1]$$

のようになり, S が奇数のとき隠れ変数は中間値の 0 を含む. また, 分割区間を無限区間に分割することで $\mathcal{X}(\infty)]$ を連続値とみなすことができる.

隠れ変数が $\mathcal{X}(\mathrm{S})$ の値を取る制限ボルツマンマシン P_{S} は, 式 (2) のエネルギー関数 $E(\boldsymbol{\theta})$ を用いることで

$$P_{\mathrm{s}}(\boldsymbol{v},\boldsymbol{h} \mid \boldsymbol{\theta}) = \frac{1}{Z_s(\boldsymbol{\theta})} \exp\left(-E\left(\boldsymbol{\theta}\right)\right) \tag{1}$$

と定義する. ここで, $Z_s(\boldsymbol{\theta})$ は規格化定数, $\boldsymbol{\theta}$ は制限ボルツマンマシンのパラメータ集合である.

制限ボルツマンマシンのパラメータ集合を $\boldsymbol{\theta} = \{\boldsymbol{b},\boldsymbol{c},\boldsymbol{w}\}$ とし, これらのパラメータを用いて制限ボルツマンマシンのエネルギー関数を

$$E(\boldsymbol{\theta}) = \underbrace{-\sum_{i\in V} b_i v_i}_{\substack{\text{バイアス項}\\(\text{可視変数})}} \underbrace{-\sum_{j\in H} c_j h_j}_{\substack{\text{バイアス項}\\(\text{隠れ変数})}} \underbrace{-\sum_{i\in V}\sum_{j\in H} w_{ij} v_i h_j}_{\text{ウェイト項}} \tag{2}$$

と定義する.

制限ボルツマンマシン学習は最尤法により行い, N 個の観測データ集合 $\mathcal{D} = \{\boldsymbol{d}^{(n)} \in \{-1,+1\}^{|V|} \mid n = 1,2,3,...,N\}$ 用いた対数尤度関数

$$l_{\mathcal{D}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} \ln P_s(\boldsymbol{v} = \boldsymbol{d}^{(n)} \mid \boldsymbol{\theta}) \tag{3}$$

を最大化するパラメータ $\boldsymbol{\theta}$ を決める. ここで, 式 (3) のパラメータの勾配は

$$\frac{\partial l_{\mathcal{D}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \underbrace{-\frac{1}{N}\sum_{n=1}^{N} \frac{\partial E(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}}_{\text{データの標本平均項}} \underbrace{+\sum_{\boldsymbol{v}}\sum_{\boldsymbol{h}} \frac{\partial E(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} P_s(\boldsymbol{v},\boldsymbol{h} \mid \boldsymbol{\theta})}_{\text{制限ボルツマンマシンの期待値項}} \tag{4}$$

である. 式 (4) の第 1 項は観測データの標本平均から容易に計算可能であるが, 第 2 項は制限ボルツマンマシンによる期待値であるためノード数が多いほど組合せ爆発により計算困難である. そのため厳密計算が困難な場合は Contrastive Divergence(CD) 法 [3] などで近似計算を行う.

3 スパース正則化制限ボルツマンマシン

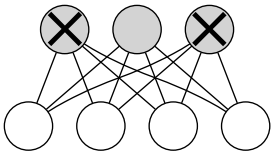


図 2: スパースな構造を持った制限ボルツマンマシン.

制限ボルツマンマシンは隠れ変数が多いと過学習を起こしやすい. そのため不要な隠れ変数を減らすような振る舞いを持たせることで過学習を抑制することが出来る. 式 (2) の制限ボルツマンマシンのエネルギー関数の役割について説明すると, 隠れ変数のバイアスパラメータ $\boldsymbol{c}$ は隠れ変数が大きい値を取りやすいか小さい値を取りやすいかを調整する役割をもつ. 隠れ変数がスパースとなるようにするには隠れ変数の値が 0 に近い値となる制約が必要となる. その制約を満たすためにスパース正則化項を式 (2) のエネルギー関数に追加した, 新たな制限ボルツマンマシンを定義する.

制限ボルツマンマシンのパラメータ集合 $\boldsymbol{\theta}$ にスパース正則化パラメータ $\boldsymbol{\lambda}$ を追加した新たなパラメータ集合を $\boldsymbol{\theta}^{\dagger} = \boldsymbol{\theta} \cup \boldsymbol{\lambda}$

Restricted Boltzmann machine with sparse hidden variables
†Yuki YOKOYAMA, Muneki YASUDA; Graduate School of Science and Engineering, Yamagata University.

とする．また，式 (2) のエネルギー関数に対しスパース正則化項を加えた新たなエネルギー関数

$$E^{\dagger}(\boldsymbol{\theta}^{\dagger}) = E(\boldsymbol{\theta}) + \underbrace{\sum_{j \in H} \exp(\lambda_j)\,|h_j|}_{\text{スパース正則化項}}$$

を定義して用いることによりスパース正則化制限ボルツマンマシン $P_s^{\dagger}$ を以下のように表す．

$$P_s^{\dagger}(\boldsymbol{v}, \boldsymbol{h} \mid \boldsymbol{\theta}^{\dagger}) = \frac{1}{Z_s^{\dagger}(\boldsymbol{\theta}^{\dagger})}(-E^{\dagger}(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta}^{\dagger}))$$

スパース正則化パラメータ $\boldsymbol{\lambda}$ を調整することで隠れ変数 $\boldsymbol{h}$ が 0 に近い値を取る確率が変動し，$\lambda_j > 0$ のとき h_j は 0 に近い値を取りやすくなり，$\lambda_j < 0$ のとき h_j は 0 に近い値を取りにくくなる．このモデルを学習することで不要な隠れ変数に対するスパース正則化パラメータ $\boldsymbol{\lambda}$ が調整され，不要な隠れ変数のみ確率的に 0 を取りやすくなるように学習することができる．それにより図 2 のようなスパースな構造のような振る舞いを持ち過学習を抑制することが出来る．

4　数値実験

制限ボルツマンマシンと，スパース正則化制限ボルツマンマシンの学習において汎化誤差を低減し過学習を抑制できているかどうかを実験的に示す．

この数値実験の枠組みはデータから制限ボルツマンマシン学習を行い，学習モデルと真の分布である生成モデルとの汎化誤差をカルバックライブラー情報量を用いて計算し，各モデルについて比較を行うことである．

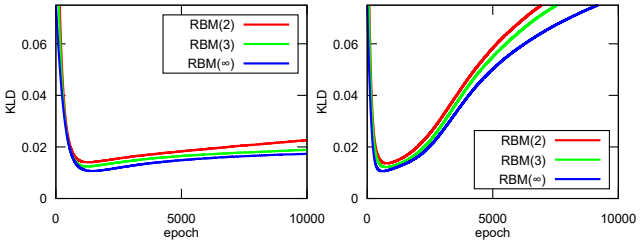
実験ではデータと生成モデルが必要となるため，2 値の制限ボルツマンマシン P_2 を生成モデルとして乱数的に生成したデータを訓練データとして用いる．訓練モデルとしては表 1 で示す制限ボルツマンマシン，スパース正則化制限ボルツマンマシンを用いてそれぞれ学習させる．

表 1: 実験モデルと学習の設定

	生成モデル	訓練モデル
RBM	P_2	P_2, P_3, P_4, P_5, P_∞, $P_3^{\dagger}$, $P_4^{\dagger}$, $P_5^{\dagger}$, $P_\infty^{\dagger}$
$\lvert V\rvert$	8	
$\lvert H\rvert$	4	$4+5$
パラメータ初期値	Xavier の初期値	
学習法	Contrastive Divergence 法，厳密計算	
最適化	AdaMax[4]	
試行回数	200 回	

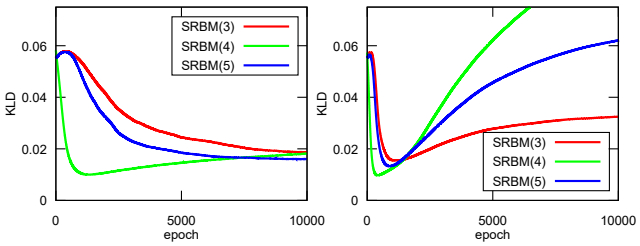
表 1 の補足として，スパース正則化パラメータ $\boldsymbol{\lambda}$ の初期値は $[2, 4]$ の一様乱数で初期化している．実験の制約として訓練モデルの隠れ変数を意図的に多くしている．これは，本研究の目的が過学習の抑制であるため意図的に過学習を起こしやすい学習モデルを用いることで過学習の抑制をわかりやすくするためである．

表 1 の条件の数値実験のうち，制限ボルツマンマシン P_2, P_3, P_∞ の実験結果を図 3，スパース正則化制限ボルツマンマシン $P_3^{\dagger}$, $P_4^{\dagger}$, $P_5^{\dagger}$ の実験結果を図 4 に示す．なお，図示しない他の制限ボルツマンマシンモデルについての実験結果は本講演にて発表する．



(a) CD 法による近似計算.　　(b) 厳密計算.

図 3: 制限ボルツマンマシン学習の実験結果．隠れ変数を多値に分割するほど汎化誤差が大きくなりにくくなる．



(a) CD 法による近似計算.　　(b) 厳密計算.

図 4: スパース正則化制限ボルツマンマシン学習の実験結果．厳密計算において特に汎化誤差が小さくなっている．

図 3 では CD 法，厳密計算のどちらの制限ボルツマンマシン学習であっても隠れ変数が多値を取るほど汎化誤差が小さくなっていることを表している．

一方，図 4 のスパース正則化制限ボルツマンマシン学習においては学習を重ねるうちに一度汎化誤差が大きくなった後汎化誤差が小さくなり，その後過学習により汎化誤差が大きくなっている．これはスパース正則化パラメータの学習の影響からである．また，図 3b と図 4b とを比較するとスパース正則化を入れることで顕著に過学習を抑制できていることを表している．

なお，図 4b を見るとスパース正則化制限ボルツマンマシンは多値制限ボルツマンマシンのように隠れ変数をより細かく多値分割するよりも，3 値，5 値のほうが汎化誤差が小さくなっている．隠れ変数の値として 0 をとるようなスパース正則化制限ボルツマンマシンのほうが傾向的には汎化誤差が小さくなる．この仕組みを本講演にて単純な構造の制限ボルツマンマシンを用いて発表する予定である．

5　まとめ

制限ボルツマンマシン学習において先行研究 [1] では Contrastive Divergence 法による学習においてスパース正則化を導入することで過学習を抑制出来ることを確認したが，本研究ではさらに厳密計算でも過学習を抑制できることを確認した．

謝辞

本研究の一部は，JSPS 科研費（15K00330, 15H03699）及び，JST CREST（JPMJCR1402）の補助を受けて行われたものである．

参考文献

[1] 横山 悠貴，安田 宗樹，多値の隠れ変数を持つ制限ボルツマンマシン，IPSJ. (2017)
[2] G.Hinton, A Practical Guide to Training Restricted Boltzmann Machines, Technical report, UTML TR. (2010)
[3] G.Hinton,Training Products of Experts by Minimizing Contrastive Divergence.Neural Computation, 14, pp 1771-1800.(2002)
[4] D.Kingma, J.Ba, Adam: A Method for Stochastic Optimization.(2015)