# 学習を反映した重みパラメータを持つ深層ボルツマンマシンの統計力学的解析

市川佑馬（東京大学大学院 総合文化研究科 広域科学専攻）
指導教員 福島孝治（東京大学大学院 総合文化研究科 広域科学専攻, 先進科学研究機構

## 導入：深層ボルツマンマシンついて

### 統計的機械学習

- データ$D = \{\boldsymbol{\sigma}^{(0),\mu}\}_{\mu=1}^{P}$の背後に確率分布$q(\boldsymbol{\sigma}^{(0)})$の存在を仮定し, 確率分布$p(\boldsymbol{\sigma}^{(0)})$により, $q(\boldsymbol{\sigma}^{(0)})$を模倣する枠組み
  - 模倣するプロセスを**学習**と呼び, $p(\boldsymbol{\sigma}^{(0)})$が$q(\boldsymbol{\sigma}^{(0)})$を良く模倣するとき, $p(\boldsymbol{\sigma}^{(0)})$は**生成機として機能する**と呼ぶ

### 深層ボルツマンマシン[1]（deep Boltzmann machine; DBM）

- 学習モデルの一つであり, 以下のように定義される
  - エネルギー関数と呼ばれる関数を定義

$$E(\boldsymbol{\sigma};\boldsymbol{\theta}) := -\sum_{l=1}^{L} (\boldsymbol{\sigma}^{l-1})^{\top} \boldsymbol{W}^{(l)} \boldsymbol{\sigma}^{(l)}, \quad \boldsymbol{\theta} := \{\boldsymbol{W}^{(l)}\}_{l=1}^{L}$$

    - $\boldsymbol{\sigma} := \{\boldsymbol{\sigma}^{(l)}\}_{l=1}^{L}$, $\boldsymbol{\sigma}^{(l)} := \{\sigma_j^{(l)} \mid \sigma_j^{(l)} \in \{\pm 1\}\}_{j=1}^{N_l}$, $\boldsymbol{W}^{(l)} \in \mathbb{R}^{N_{l-1} \times N_l}$
    - $\boldsymbol{\sigma}^{(0)}$：可視層（要素は可視ユニット）, $\boldsymbol{\sigma}^{(l)}$：第$l$隠れ層（要素は第$l$隠れユニット）$\boldsymbol{W}^{(l)}$：重みパラメータと呼ばれる
  - 以下の確率分布を定義

$$p(\boldsymbol{\sigma} \mid \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta},\beta)} e^{-\beta E(\boldsymbol{\sigma};\boldsymbol{\theta})}, \quad Z(\boldsymbol{\theta},\beta) = \sum_{\boldsymbol{\sigma}} e^{-\beta E(\boldsymbol{\sigma};\boldsymbol{\theta})}$$

  - $p(\boldsymbol{\sigma} \mid \boldsymbol{\theta})$を$\boldsymbol{\sigma} \backslash \boldsymbol{\sigma}^{(0)}$について周辺化した分布がDBM

$$p(\boldsymbol{\sigma}^{(0)} \mid \boldsymbol{\theta}) = \sum_{\boldsymbol{\sigma} \backslash \boldsymbol{\sigma}^{(0)}} p(\boldsymbol{\sigma} \mid \boldsymbol{\theta})$$

## DBMの問題点・研究目的

### 背景

- 既存の学習手法に関する理論的な妥当性が存在しない
- 与えられたデータに対して最適なネットワーク構造を決定する方法論が存在しない
- 層状にネットワークを拡張することが, 学習や表現能力に与える影響は明らかになっていない

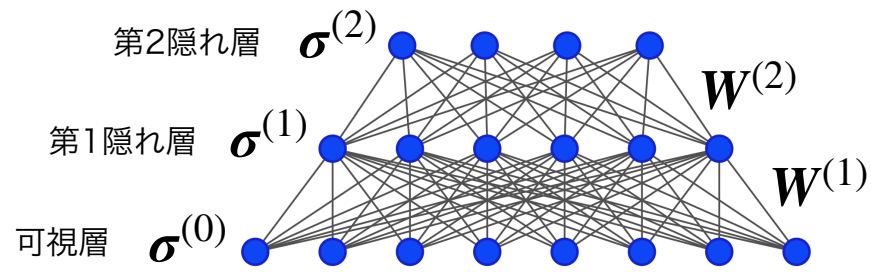
### 研究目的

課題解決には, **DBMが生成機として機能する際に要求される典型的な性質**や**層状のネットワーク構造が学習や表現能力に与える影響**を明らかにすることは重要であり, 本研究ではこれらを明らかにする

### 研究手法

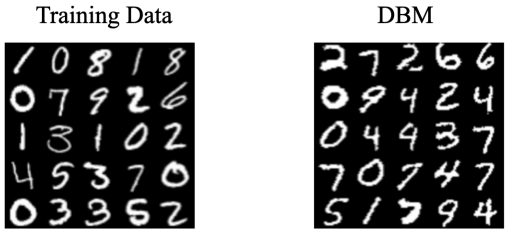
(1) 数値実験によりDBMの性質を明らかにする
(2) 数値実験の結果を反映したモデル化を行い, 統計力学の観点からDBMの典型的な性質を調べる

## DBMのネットワーク図（3層DBMの具体例）



## 数値実験の設定

- 使用したデータ：$6 \times 10^4$枚, $28 \times 28$ピクセルの数字データ
- 3層DBM：$(N_0, N_1, N_2) = (784, 700, 700)$
- 学習手法：事前学習を行い, 数値的に対数尤度を最大化[2]




## 数値実験で観測した量

- 学習後の重み$\{\boldsymbol{W}^{(l)}\}_{l=1}^{L}$の特異モードの性質に着目

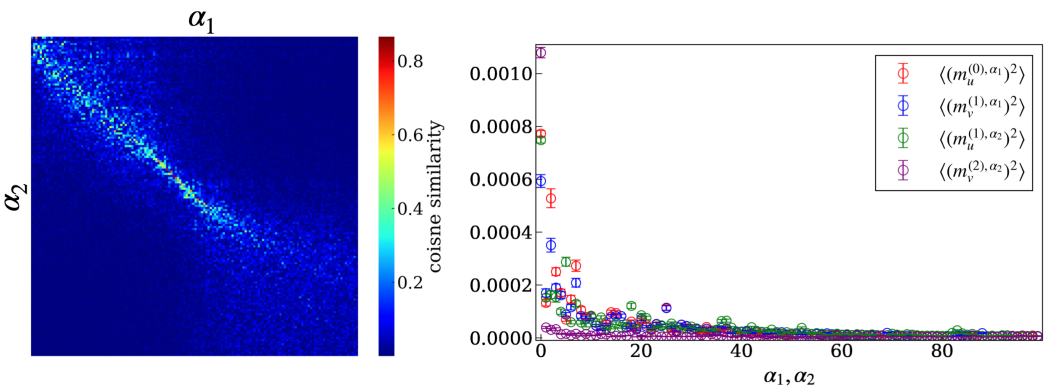
$$\boldsymbol{W}^{(l)} = \boldsymbol{U}^{(l)}\boldsymbol{\Sigma}^{(l)}(\boldsymbol{V}^{(l)})^{\mathsf{T}}$$

  - 特異値は降順に$w_1 \geq w_2 \geq \cdots$とし, $\alpha_l$は特異モードを指定
  - $\boldsymbol{U}^{(l)}$, $\boldsymbol{V}^{(l)}$：各列が左, 右特異ベクトル$\boldsymbol{u}_{\alpha_l}^{(l)}$, $\boldsymbol{v}_{\alpha_l}^{(l)}$の直交行列
  - $\boldsymbol{\Sigma}^{(l)}$：対角要素が特異値$w_{\alpha_l}^{(l)}$の対角行列
- 数値実験で観測した量
  - **Interlayer correlation**：$\boldsymbol{W}^{(l)}$, $\boldsymbol{W}^{(l+1)}$の右・左特異ベクトルの内積: $\rho_{\alpha_l,\alpha_{l+1}}^{(l)} := \langle \boldsymbol{v}_{\alpha_l}^{(l)}, \boldsymbol{u}_{\alpha_{l+1}}^{(l+1)} \rangle$
  - **Overlap**：$\boldsymbol{W}^{(l)}$の特異ベクトルに対して定義される以下の量:

$$m_u^{(l-1),\alpha_l} := \frac{(\boldsymbol{\sigma}^{(l-1)})^{\mathsf{T}}\boldsymbol{u}_{\alpha_l}^{(l)}}{\sqrt{M_l}},\ m_v^{(l),\alpha_l} := \frac{(\boldsymbol{\sigma}^{(l)})^{\mathsf{T}}\boldsymbol{v}_{\alpha_l}^{(l)}}{\sqrt{M_l}},\ M_l := \sqrt{N_{l-1}N_l}$$

## 数値実験の結果

- **$\mathcal{O}(1)$個の特異値が大きい特異モードにより, DBMは生成機として機能する**
- **Interlayer correlation $\langle \boldsymbol{v}_{\alpha_1}^{(1)}, \boldsymbol{u}_{\alpha_2}^{(2)} \rangle$の性質**
  - 最初の120個程度の特異モードのInterlayer correlationは, $\alpha_1 \approx \alpha_2$のとき大きな値を持つ
  - それ以降の特異モードに関してはほとんど0となる
- **overlapの性質**
  - 生成機として機能するDBMは, 複数の特異ベクトルのoverlapが同時に非ゼロとなる
  - 生成機として機能しないDBMは, overlapがほとんど0 or 最大特異値のみoverlapが非ゼロとなる




(i) (ii)

(i) 最初の150個の特異モードのInterlayer correlation（左上が$\langle \boldsymbol{v}_1, \boldsymbol{u}_1 \rangle$に対応）

(ii) 最初の100個の特異モードのoverlapの二次モーメントの$\alpha_1$, $\alpha_2$依存性

## 数値実験を反映したモデル化

- DBMの重みパラメータを以下の形式で表す

$$\boldsymbol{W}^{(l)} = \sum_{\alpha_l=1}^{K_l} w_{\alpha_l}^{(l)} \boldsymbol{u}_{\alpha_l}^{(l)} (\boldsymbol{v}_{\alpha_l}^{(l)}) + \boldsymbol{r}^{(l)}$$

- 第一項：DBMの性能を支配する特異モードからなる項
  - $K_l = \mathcal{O}(1)$：$\mathcal{O}(1)$**個の特異モードでDBMが生成機として機能することを反映**
  - $\boldsymbol{u} \equiv \{\boldsymbol{u}^{(l)}\}_{l=1}^{L}$, $\boldsymbol{v} \equiv \{\boldsymbol{v}^{(l)}\}_{l=1}^{L}$, $\boldsymbol{u}^{(l)} \equiv \{\boldsymbol{u}_{\alpha_l}^{(l)}\}_{\alpha_l=1}^{K_l}$, $\boldsymbol{v}^{(l)} \equiv \{\boldsymbol{v}_{\alpha_l}^{(l)}\}_{\alpha_l=1}^{K_l}$
- 第二項：それ以外の特異モードからなる項

**特異モードの性質は問題に依存するため,**
**ランダム化して典型的な性質を調べる**

以下の確率分布を仮定する

- **第一項：特異ベクトルの確率分布**$p(\boldsymbol{u}, \boldsymbol{v})$**は以下の性質を持つ**
  - $\mathbb{E}[v_{j,\alpha_l}^{(l)} v_{j,\alpha_l}^{(l)}] = (s_{v,\alpha_l}^{(l)})^2 / N_l$, $\mathbb{E}[u_{i,\alpha_l}^{(l)} u_{i,\alpha_l}^{(l)}] = (s_{u,\alpha_l}^{(l)})^2 / N_{l-1}$
  - $\forall j \in \{1,\ldots,N_l\}$, $\mathbb{E}[v_{j,\alpha_l}^{(l)} u_{j,\alpha_{l+1}}^{(l+1)}] = \begin{cases} \rho_{\alpha_l,\alpha_{l+1}}^{(l)} / N_{l-1} & \alpha_l = \alpha_{l+1}, \\ 0 & \alpha_l \neq \alpha_{l+1} \end{cases}$

    $\alpha_l \approx \alpha_{l+1}$**で大きなInterlayer correlationを反映**
- **第二項：**$\boldsymbol{r}^{(l)}$**の確率分布**
  - $\boldsymbol{r}^{(l)}$：$r_{ij}^{(l)} \sim_{\text{iid}} \mathcal{N}\left(0, \sigma_l^2 / M_l\right)$, $M_l := \sqrt{N_{l-1} N_l}$

- **数値実験より, 複数の特異ベクトルのoverlapが同時に非ゼロとなることが重要である**
- 統計力学的な解析手法(3)により, 数値実験を反映したモデルを解析して, 以下の相を調べた.
  - **常磁性相（P)**：無秩序なサンプルを生成する相
  - **スピングラス相（SG）**：特異ベクトルと関係のないサンプルを生成する相
  - **強磁性相（F）**：少なくとも一つの特異ベクトルと関係のあるサンプルを生成する相
  - **合成相（CM）**：少なくとも二つ以上の特異ベクトルと関係のあるサンプルを生成する相
- 強磁性相, 合成相は数値実験で得られた結果を再現するための必要条件となる

## 3層DBMの結果

- 強磁性相では最大特異値のoverlapが有限となる
- Interlayer correlationの増加に伴い, 強磁性相拡大
- 中間層のユニット数を減らす場合が最も強磁性相が大きい
- 最終層のユニット数を減らす場合, 中間層のユニット数を減らす方が良い→**パラメータの増加が必ずしも強磁性相の拡大に直結しない**

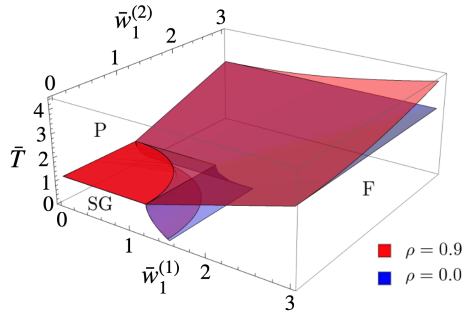



Fig. 3層DBMの相図（$\bar{a} \equiv a/\sigma_1\sigma_2$, $N_0 = N_1 = N_2$）

## 相図の層数依存性

- 解析のため以下を仮定する

$$w_{\alpha_1}^{(1)} = w_{\alpha_2}^{(2)} = \cdots = w_{\alpha_L}^{(L)} \equiv w_\alpha,\ \sigma_1 = \sigma_2 = \cdots = \sigma_L \equiv \sigma$$

- 結果
  - Interlayer correlationが閾値 $\rho^* = \sqrt{2} - 1$より小さいと強磁性相は縮小, 大きいと拡大
  - 層数無限大極限で相境界は収束する
  - 閾値$\rho^*$よりもInterlayer correlationが大きい場合でも, 層数増加に伴い, 強磁性相の拡大は緩やかになる

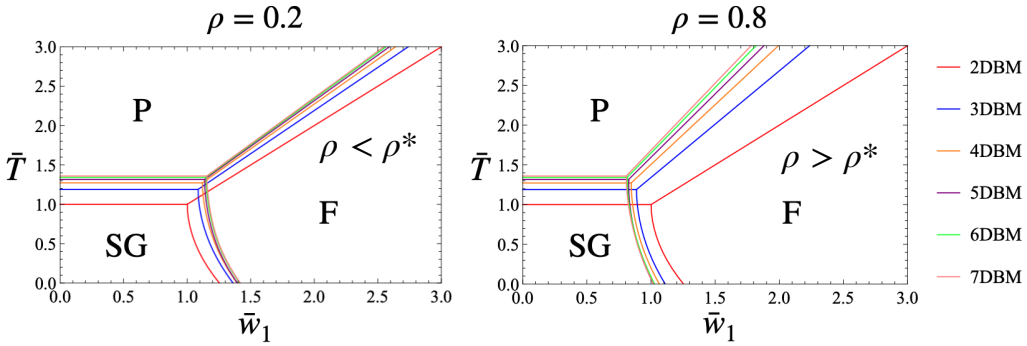



Fig. 相図の層数依存性（$\bar{a} \equiv a/\sigma$）

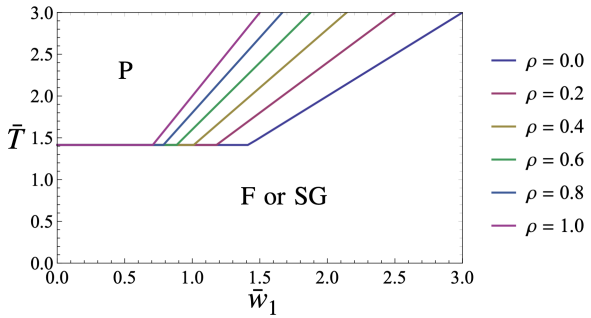



Fig. 層数無限大極限のDBM（$\bar{a} \equiv a/\sigma$）
Interlayer correlation 依存性

## 主結果

- **数値実験により明らかになったこと**
  - 隠れ層間の結合により, Interlayer-correlationが出現
  - DBMが生成機として機能するためには複数のoverlapが同時に非ゼロとなることが重要
- **統計力学的な解析から明らかになったこと**
  - Interlayer correlationの増加に伴い, 強磁性相拡大
  - 各層間のInterlayer correlationが閾値$\rho^*$よりも小さい場合, 層数増加に伴い強磁性相縮小
    - **Interlayer-correlation次第で層数増加は悪影響**
  - 層数増加に伴い強磁性相の拡大は緩やかになる
    - **層数増加が計算量の増加を補うほどの表現力があるとは限らない**
  - 最大特異値と二番目の特異値が同程度の値を持つ場合, 合成層が出現する
    - **DBMが生成機として機能するためには, 特異値が近い値を持つ必要がある**

## 参考文献


(1) R. Salakhutdinov and G. Hinton, *Deep boltzmann machines*, Artificial intelligence and statistics, PMLR, 2009.

(2) G. Hinton, E. Geoffrey and R. Salakhutdinov, *A better way to pretrain deep boltzmann machines,* Advances in Neural Information Processing Systems 25 (2012): 2447-2455.

(3) A. Decelle, E. Giancarlo and F. Cyril, *Thermodynamics of restricted Boltzmann machines and related learning dynamics,* Journal of Statistical Physics 172.6 (2018): 1576-1608.