CF-007

ガウシアン・ベルヌーイ制限ボルツマンマシンにおける不完全データ学習の改良
Improvement of Incomplete Data Learning in Gaussian–Bernoulli Restricted Boltzmann Machine

関本 快士[1)]　安田 宗樹[1)]
Kaiji Sekimoto　Muneki Yasuda

1　はじめに

機械学習を行う際，使用可能なデータが一部欠損していることが度々ある．そのため，不完全データに対する機械学習は応用上重要である．不完全データの対処法としては，例えば欠損を含むデータを除去するか，データ前処理として何らかの方法で欠損値補完を行うことが考えられる．しかし，データの除去は扱えるデータ数の減少を引き起こし，また，平均値埋めなどの安易な欠損値補完はデータに不適切な情報を付加してしまい，学習に悪影響を及ぼす恐れがある．よって，データの除去・補完をせずに不完全データを直接扱える機械学習が理想的である．統計的機械学習モデルのガウシアン・ベルヌーイ制限ボルツマンマシン (Gaussian–Bernoulli restricted Boltzmann machine (GBRBM)) [1, 2] は理想的なモデルのひとつであり，欠損部を周辺化消去することで補完を行わずに学習することができる [3]．しかし，GBRBM の学習では計算困難な期待値の近似計算を必要とし，その近似精度が学習性能のボトルネックになる．よって，高精度な期待値近似法に基づく学習法が求められる．

Lossy contrastive divergence (Lossy-CD) [3] と呼ばれる従来の学習法では，サンプリングベースの近似法が用いられる．これは，モデル分布からサンプル点を生成するサンプリング過程と，得られたサンプル点を用いて期待値を近似する近似過程の 2 つの過程から構成され，各過程にはそれぞれブロック化ギブスサンプリング (blocked Gibbs sampling (bGS)) とモンテカルロ積分法 (Monte Carlo integration (MCI)) という標準的な手法が採用されている．bGS はマルコフ連鎖モンテカルロ法のひとつであり，GBRBM では容易に実行できるためよく使われるが，緩和時間が長く，効率性が低いことが報告されている [4, 5]．さらに，近似過程で用いられる MCI は近似精度が低い [6]．よって，従来学習法の期待値近似は粗い近似となっている．本研究では，この近似法を改良することで，不完全データ学習における高性能学習法を提案する．

提案法では，サンプリング過程には周辺分布上でのギブスサンプリング (marginalized-space Gibbs sampling (mGS)) [4] を用い，近似過程には MCI を空間的に拡張した空間 MCI (spatial MCI (SMCI)) [6] を用いる．SMCI は条件付き期待値の標本平均で期待値を近似する方法であり，部分的に和をとることによって，一切の和をとらない MCI よりも高精度となる [6]．提案法では，隠れ変数に対応するサンプル点のみを要する SMCI を設計し，利用する．そして，そのサンプル点は隠れ変数上の mGS によって生成する．隠れ変数の周辺分布はボルツマンマシン (Boltzmann machine (BM)) の形式で得られるため，サンプリングを容易に行うことができる．また，

1)　山形大学大学院理工学研究科

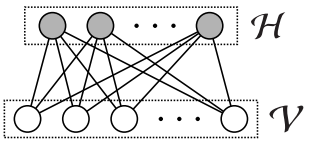


図 1　GBRBM のグラフ構造．可視層 $\mathcal{V}$ と隠れ層 $\mathcal{H}$ の 2 層構造から成り，層内に結合をもたない．

周辺化によって変数空間が削減されるため，mGS は元の分布上での bGS よりも緩和時間が短く，効率的である．よって，提案法は従来法よりも効率的に高精度近似を行えるため，学習性能が高いことが期待される．

2　ガウシアン・ベルヌーイ制限ボルツマンマシン

GBRBM は完全二部グラフ上に定義されるマルコフ確率場であり，データに対応する可視変数から成る可視層と，モデルの表現能力を制御する隠れ変数から成る隠れ層の 2 層構造をもつ [1]．可視変数と隠れ変数をそれぞれ $\boldsymbol{v} := \{v_i \in \mathbb{R} \mid i \in \mathcal{V}\}$ と $\boldsymbol{h} := \{h_j \in \{0,1\} \mid j \in \mathcal{H}\}$ と表す．ここで，$\mathcal{V} := \{1, 2, \ldots, n_v\}$ と $\mathcal{H} := \{n_v + 1, n_v + 2, \ldots, n_v + n_h\}$ はそれぞれ可視層と隠れ層のノード番号の集合である．図 1 に GBRBM のグラフ構造を示す．GBRBM のエネルギー関数は様々な定義が存在するが，本研究では以下で定義される canonicalized GBRBM [4] のエネルギー関数を用いる．

$$E_\theta(\boldsymbol{v}, \boldsymbol{h}) := \sum_{i \in \mathcal{V}} \frac{v_i^2}{2\,\mathrm{sfp}(\sigma_i)} - \sum_{i \in \mathcal{V}} b_i v_i - \sum_{j \in \mathcal{H}} c_j h_j - \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{H}} w_{i,j} v_i h_j \tag{1}$$

ここで，$\mathrm{sfp}(x) := \ln(1 + e^x)$ はソフトプラス関数である．また，$b_i, c_j, w_{i,j}, \sigma_i \in \mathbb{R}$ は学習パラメータであり，まとめて θ と表す．式 (1) のエネルギー関数を用いて，GBRBM の確率分布は次のように定義される．

$$P_\theta(\boldsymbol{v}, \boldsymbol{h}) := \frac{1}{Z_\theta} \exp(-E_\theta(\boldsymbol{v}, \boldsymbol{h})) \tag{2}$$

ここで，Z_θ は規格化定数であり次のように定義される．

$$Z_\theta := \int d\boldsymbol{v} \sum_{\boldsymbol{h}} \exp(-E_\theta(\boldsymbol{v}, \boldsymbol{h}))$$

ここで，$\int d\boldsymbol{v} := \prod_{i \in \mathcal{V}} \int_{-\infty}^{+\infty} dv_i$ は $\boldsymbol{v}$ についての多重積分であり，$\sum_{\boldsymbol{h}} := \prod_{j \in \mathcal{H}} \sum_{h_j \in \{0,1\}}$ は $\boldsymbol{h}$ についての多重和である．片方の層が与えられたときのもう片方の層の条件付き分布は，それぞれ以下の形で表される．

$$P_\theta(\boldsymbol{v} \mid \boldsymbol{h}) = \prod_{i \in \mathcal{V}} \mathcal{N}(v_i \mid \lambda_i(\boldsymbol{h}), \mathrm{sfp}(\sigma_i)), \tag{3}$$

$$P_\theta(\boldsymbol{h} \mid \boldsymbol{v}) = \prod_{j \in \mathcal{H}} \frac{\exp(\tau_j(\boldsymbol{v}) h_j)}{1 + \exp(\tau_j(\boldsymbol{v}))} \tag{4}$$

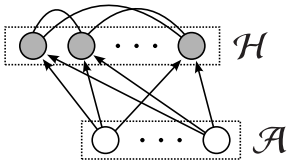


図 2　式 (5) の隠れ変数の分布のグラフ構造．可視層の一部 $\mathcal{A}$ ($\subset \mathcal{V}$) が条件として与えられ，隠れ層 $\mathcal{H}$ 内は全結合をもつ．

ここで，$\lambda_i(\boldsymbol{h}) := \mathrm{sfp}(\sigma_i)(b_i + \sum_{j\in\mathcal{H}} w_{i,j}h_j)$，$\tau_j(\boldsymbol{v}) := c_j + \sum_{i\in\mathcal{V}} w_{i,j}v_i$ である．また，$\mathcal{N}(x \mid \mu, \sigma^2)$ は平均 μ，分散 σ^2 の正規分布を表す．式 (3) と (4) より，片方の層が与えられたときのもう片方の層の変数は統計的に独立であることがわかる．この性質は条件付き独立性と呼ばれ，式 (3) と (4) の条件付き分布を用いて，可視層と隠れ層の間の bGS を効率的に行うことができる．

可視層の部分集合 $\mathcal{A}$ ($\subset \mathcal{V}$) 上の可視変数 $\boldsymbol{v}_\mathcal{A} := \{v_i \mid i \in \mathcal{A}\}$ が与えられたときの隠れ変数の分布は，完全グラフ上の BM となる．

$$\begin{aligned} P_\theta(\boldsymbol{h} \mid \boldsymbol{v}_\mathcal{A}) &= \int d\boldsymbol{v}_{\mathcal{V}\setminus\mathcal{A}} P_\theta(\boldsymbol{v}_{\mathcal{V}\setminus\mathcal{A}}, \boldsymbol{h} \mid \boldsymbol{v}_\mathcal{A}) \\ &= \frac{1}{Z_\theta^{(\mathrm{h})}(\boldsymbol{v}_\mathcal{A})} \exp\Bigg(\sum_{i\in\mathcal{H}} \beta_i(\boldsymbol{v}_\mathcal{A}) h_j \\ &\qquad + \sum_{i\in\mathcal{H}} \sum_{j\in\mathcal{H}\setminus\{i\}} J_{i,j}(\mathcal{A}) h_i h_j \Bigg) \end{aligned} \tag{5}$$

図 2 にこの分布のグラフ構造を示す．ここで，$\beta_i(\boldsymbol{v}_\mathcal{A})$ と $J_{i,j}(\mathcal{A})$ はそれぞれバイアスパラメータと結合パラメータであり，それぞれ以下で定義される．

$$\begin{aligned} \beta_i(\boldsymbol{v}_\mathcal{A}) &:= c_i + \sum_{k\in\mathcal{A}} w_{k,i} v_k \\ &\qquad + \sum_{k\in\mathcal{V}\setminus\mathcal{A}} \mathrm{sfp}(\sigma_k) \Big(w_{k,i} b_k + \frac{1}{2} w_{k,i}^2 \Big) \\ J_{i,j}(\mathcal{A}) &:= \sum_{k\in\mathcal{V}\setminus\mathcal{A}} w_{k,i} w_{k,j} \,\mathrm{sfp}(\sigma_k) \end{aligned}$$

また $Z_\theta^{(\mathrm{h})}(\boldsymbol{v}_\mathcal{A})$ は，式 (5) の分布の規格化定数であり，次の関係式が成り立つ．

$$P_\theta^\dagger(\boldsymbol{v}_\mathcal{A}) = Z_\theta^{(\mathrm{h})}(\boldsymbol{v}_\mathcal{A}) \exp\Bigg(\frac{1}{2} \sum_{i\in\mathcal{V}\setminus\mathcal{A}} \Big[\ln(2\pi\,\mathrm{sfp}(\sigma_i)) \tag{6}$$

$$+ b_i^2 \,\mathrm{sfp}(\sigma_i) \Big] + \sum_{i\in\mathcal{A}} \Big[-\frac{v_i^2}{2\,\mathrm{sfp}(\sigma_i)} + b_i v_i \Big] \Bigg) \tag{7}$$

ここで，$P_\theta^\dagger(\boldsymbol{v}_\mathcal{A}) := \int d\boldsymbol{v}_{\mathcal{V}\setminus\mathcal{A}} \sum_{\boldsymbol{h}} \exp(-E_\theta(\boldsymbol{v}, \boldsymbol{h}))$ は周辺分布 $P_\theta(\boldsymbol{v}_\mathcal{A})$ の非規格化確率である．$\mathcal{A} = \varnothing$ のとき，式 (5) の分布は隠れ変数の周辺分布 $P_\theta(\boldsymbol{h})$ ($= P_\theta(\boldsymbol{h} \mid \boldsymbol{v}_\varnothing)$) と一致する．

3　GBRBM における不完全データ学習

各データ点の一部要素がランダムに欠損した不完全な N 個のデータ点の集合

$$\begin{aligned} \mathbb{D} &:= \{\mathbf{d}^{(\mu)} \mid \mu = 1, 2, \ldots, N\} \\ \mathbf{d}^{(\mu)} &:= \{\mathrm{d}_i^{(\mu)} \mid i \in O^{(\mu)} \subseteq \mathcal{V}\} \end{aligned}$$

が得られたとする．ここで，$O^{(\mu)}$ は μ 番目のデータ点の観測部を表し，対応する欠損部を $M^{(\mu)} := \mathcal{V}\setminus O^{(\mu)}$ と表す．不完全データ集合に対する学習は，隠れ変数に加えて各データ点の欠損部も周辺化消去した対数尤度関数

$$\begin{aligned} \ell(\theta) &:= \frac{1}{N} \sum_{\mu=1}^{N} \ln \Bigg[\int d\boldsymbol{v}_{M^{(\mu)}} \sum_{\boldsymbol{h}} P_\theta(\mathbf{d}^{(\mu)}, \boldsymbol{v}_{M^{(\mu)}}, \boldsymbol{h}) \Bigg] \\ &= \frac{1}{N} \sum_{\mu=1}^{N} \ln P_\theta^\dagger(\mathbf{d}^{(\mu)}) - \ln Z_\theta \end{aligned} \tag{8}$$

を最大化する学習パラメータを求めることで行われる [3]．式 (8) の対数尤度は勾配上昇法によって最大化される．学習パラメータに関する対数尤度の勾配は以下の形で得られる．

$$\begin{aligned} \frac{\partial\,\ell(\theta)}{\partial\, b_i} &= \frac{1}{N} \sum_{\mu=1}^{N} \Big[I_{O^{(\mu)}}(i)\, \mathrm{d}_i^{(\mu)} \\ &\qquad + (1 - I_{O^{(\mu)}}(i))\, \mathbb{E}[v_i \mid \mathbf{d}^{(\mu)}] \Big] - \mathbb{E}[v_i] \end{aligned} \tag{9}$$

$$\frac{\partial\,\ell(\theta)}{\partial\, c_j} = \frac{1}{N} \sum_{\mu=1}^{N} \mathbb{E}[h_j \mid \mathbf{d}^{(\mu)}] - \mathbb{E}[h_j] \tag{10}$$

$$\begin{aligned} \frac{\partial\,\ell(\theta)}{\partial\, w_{i,j}} &= \frac{1}{N} \sum_{\mu=1}^{N} \Big[I_{O^{(\mu)}}(i)\, \mathrm{d}_i^{(\mu)}\, \mathbb{E}[h_j \mid \mathbf{d}^{(\mu)}] \\ &\qquad + (1 - I_{O^{(\mu)}}(i))\, \mathbb{E}[v_i h_j \mid \mathbf{d}^{(\mu)}] \Big] - \mathbb{E}[v_i h_j] \end{aligned} \tag{11}$$

$$\begin{aligned} \frac{\partial\,\ell(\theta)}{\partial\, \sigma_i} &= \frac{\mathrm{sig}(\sigma_i)}{2\,\mathrm{sfp}^2(\sigma_i)} \Bigg(\frac{1}{N} \sum_{\mu=1}^{N} \Big[I_{O^{(\mu)}}(i) \left(\mathrm{d}_i^{(\mu)}\right)^2 \\ &\qquad + (1 - I_{O^{(\mu)}}(i))\, \mathbb{E}[v_i^2 \mid \mathbf{d}^{(\mu)}] \Big] - \mathbb{E}[v_i^2] \Bigg) \end{aligned} \tag{12}$$

ここで，$I_{O^{(\mu)}}(i)$ は $i \in O^{(\mu)}$ のときに 1 を，$i \notin O^{(\mu)}$ のときに 0 を返す指示関数であり，$\mathrm{sig}(x) := 1/(1 + e^{-x})$ はシグモイド関数である．勾配には，データ点 $\mathbf{d}^{(\mu)}$ が与えられた下での条件付き期待値

$$\begin{aligned} \mathbb{E}_\theta[f(\boldsymbol{v}_{M^{(\mu)}}, \boldsymbol{h}) \mid \mathbf{d}^{(\mu)}] &:= \int d\boldsymbol{v}_{M^{(\mu)}} \sum_{\boldsymbol{h}} \\ &\qquad \times f(\boldsymbol{v}_{M^{(\mu)}}, \boldsymbol{h}) P_\theta(\boldsymbol{v}_{M^{(\mu)}}, \boldsymbol{h} \mid \mathbf{d}^{(\mu)}) \end{aligned} \tag{13}$$

と通常の期待値

$$\mathbb{E}_\theta[f(\boldsymbol{v}, \boldsymbol{h})] := \int d\boldsymbol{v} \sum_{\boldsymbol{h}} f(\boldsymbol{v}, \boldsymbol{h}) P_\theta(\boldsymbol{v}, \boldsymbol{h}) \tag{14}$$

が含まれており，これらは組み合わせ爆発を引き起こす多重和や多重積分を含むため，勾配の厳密計算は一般的に計算困難である．そのため，式 (13) と (14) の期待値を何らかの手法を用いて近似しなければならない．そして，期待値の近似精度は学習性能に影響を与えることが知られている [7, 8] ため，高精度な近似法が求められる．

従来法 (Lossy-CD) [3] は，式 (13) と (14) の期待値を標準的な MCI によって近似する．例えば式 (14) の期待値を近似する場合，$P_\theta(\boldsymbol{v}, \boldsymbol{h})$ から生成された K 個のサンプル点の集合 $\{\mathbf{v}^{(\nu)}, \mathbf{h}^{(\nu)} \mid \nu = 1, 2, \ldots, K\}$ を用いて次のよ

うに近似する．

$$\mathbb{E}_\theta[f(\boldsymbol{v},\boldsymbol{h})] \approx \frac{1}{K}\sum_{\nu=1}^{K} f(\mathbf{v}^{(\nu)},\mathbf{h}^{(\nu)})$$

サンプル点の集合は，ランダムな値を初期値とした bGS によって生成される．bGS や MCI はシンプルなアルゴリズムであるため実装が容易であるが，bGS は緩和時間が長く，MCI は近似精度が低いという欠点をもつ [4, 6].

4 空間モンテカルロ積分法

SMCI は，MCI を部分的に和をとるように拡張した手法であり，MCI より高い近似精度をもつ [6]．説明のため，確率変数 $\boldsymbol{x} := \{x_i \in X_i \mid i \in \mathcal{X}\}$ に対するマルコフ確率場 $P(\boldsymbol{x})$ を考える．ここで，X_i は x_i の標本空間であり，本節では X_i を離散標本空間として数式を記述する．

関数 $f(\boldsymbol{x}_\mathcal{T})$ についての期待値

$$\mathbb{E}[f(\boldsymbol{x}_\mathcal{T})] := \sum_{\boldsymbol{x}} f(\boldsymbol{x}_\mathcal{T})P(\boldsymbol{x})$$

を近似したいとする．$P(\boldsymbol{x})$ から生成された K 個のサンプル点の集合 $\{\mathbf{x}^{(1)},\mathbf{x}^{(2)},\dots,\mathbf{x}^{(K)}\}$ を用いて，SMCI では次の形で期待値を近似する．

$$\mathbb{E}[f(\boldsymbol{x}_\mathcal{T})] \approx \frac{1}{K}\sum_{\nu=1}^{K}\sum_{\boldsymbol{x}_\mathcal{S}} f(\boldsymbol{x}_\mathcal{T})P\big(\boldsymbol{x}_\mathcal{S} \mid \mathbf{x}_{\partial\mathcal{S}}^{(\nu)}\big) \tag{15}$$

ここで，$\mathcal{T} \subseteq \mathcal{S} \subseteq \mathcal{X}$ である．また，$\partial\mathcal{S}$ は $\mathcal{S}$ の隣接領域であり，$\mathbf{x}_{\partial\mathcal{S}}^{(\nu)}$ は $\boldsymbol{x}_{\partial\mathcal{S}}$ に対応する ν 番目のサンプル点を表す．3 つの領域 $\mathcal{T},\mathcal{S},\partial\mathcal{S}$ は，それぞれ目標領域，和領域，サンプル領域と呼ばれる．式 (15) の近似式は，条件付き期待値 $\mathbb{E}[f(\boldsymbol{x}_\mathcal{T}) \mid \boldsymbol{x}_{\partial\mathcal{S}}]$ の標本平均の形となっているため，SMCI は Rao-Blackwellization [9] をマルコフ確率場へ応用した手法と見なせる．SMCI では，2 つの和領域 $\mathcal{S}_1,\mathcal{S}_2$ の間に包含関係 $\mathcal{S}_1 \supseteq \mathcal{S}_2$ が成り立つとき，$\mathcal{S}_2$ を用いた推定量よりも $\mathcal{S}_1$ を用いた推定量の方が統計的に高精度である [6]．そのため，高精度近似のためには和領域を可能な限り広くとるべきだが，無闇な和領域の拡大は期待値計算と同様の組合わせ爆発を引き起こしてしまうため，多重和 $\sum_{\boldsymbol{x}_\mathcal{S}}$ を解析的に取れる範囲で和領域を拡大していくことが望ましい．

本節の以降では，変数 $\boldsymbol{v}_\mathcal{A} = \{v_i \mid i \in \mathcal{A} \subset \mathcal{V}\}$ が与えられたときの条件付き期待値

$$\mathbb{E}_\theta[f(\boldsymbol{v}_{\mathcal{V}\backslash\mathcal{A}},\boldsymbol{h}) \mid \boldsymbol{v}_\mathcal{A}] = \int d\boldsymbol{v}_{\mathcal{V}\backslash\mathcal{A}} \sum_{\boldsymbol{h}} \times f(\boldsymbol{v}_{\mathcal{V}\backslash\mathcal{A}},\boldsymbol{h})P_\theta(\boldsymbol{v}_{\mathcal{V}\backslash\mathcal{A}},\boldsymbol{h} \mid \boldsymbol{v}_\mathcal{A}) \tag{16}$$

の近似を考える．式 (16) は，$\mathcal{A} = O^{(\mu)}$ のときに式 (13) の条件付き期待値に一致し，$\mathcal{A} = \varnothing$ のときに式 (14) の期待値に一致する．したがって，本節で議論する SMCI は，勾配に含まれる式 (13) と式 (14) の双方の期待値に対して直接適用することができる．

式 (16) の条件付き期待値を，隠れ層 $\mathcal{H}$ がサンプル領域となるように和領域を設定した SMCI に基づいて近似する．式 (5) の分布 $P_\theta(\boldsymbol{h} \mid \boldsymbol{v}_\mathcal{A})$ から生成された K 個のサンプル点の集合 $\mathbb{S} := \{\mathbf{h}^{(1)},\mathbf{h}^{(2)},\dots,\mathbf{h}^{(K)}\}$ が得られたとする．条件付き期待値 $\mathbb{E}_\theta[v_i \mid \boldsymbol{v}_\mathcal{A}]$ と $\mathbb{E}_\theta[v_i^2 \mid \boldsymbol{v}_\mathcal{A}]$ $(i \in \mathcal{V}\backslash\mathcal{A})$ は，和領域を $\mathcal{S} = \{i\}$ と設定して以下のように近似される．

$$\begin{aligned}\mathbb{E}_\theta[v_i \mid \boldsymbol{v}_\mathcal{A}] \approx \mathcal{E}_i^{(\nu)}(\mathbb{S}) &:= \frac{1}{K}\sum_{\nu=1}^{K}\int_{-\infty}^{+\infty} dv_i v_i P_\theta(v_i \mid \mathbf{h}^{(\nu)}) \\ &= \frac{1}{K}\sum_{\nu=1}^{K}\lambda_i\big(\mathbf{h}^{(\nu)}\big)\end{aligned} \tag{17}$$

$$\begin{aligned}\mathbb{E}_\theta[v_i^2 \mid \boldsymbol{v}_\mathcal{A}] \approx \mathcal{E}_i^{(\nu^2)}(\mathbb{S}) &:= \frac{1}{K}\sum_{\nu=1}^{K}\int_{-\infty}^{+\infty} dv_i v_i^2 P_\theta(v_i \mid \mathbf{h}^{(\nu)}) \\ &= \mathrm{sfp}(\sigma_i) + \big(\mathcal{E}_i^{(\nu)}(\mathbb{S})\big)^2\end{aligned} \tag{18}$$

$\mathbb{E}_\theta[h_j \mid \boldsymbol{v}_\mathcal{A}]$ は，和領域を $\mathcal{S} = (\mathcal{V}\backslash\mathcal{A}) \cup \{j\}$ と設定して次のように近似される．

$$\mathbb{E}_\theta[h_j \mid \boldsymbol{v}_\mathcal{A}] \approx \mathcal{E}_j(\mathbb{S},\boldsymbol{v}_\mathcal{A}) := \frac{1}{K}\sum_{\nu=1}^{K}\int d\boldsymbol{v}_{\mathcal{V}\backslash\mathcal{A}} \sum_{h_j\in\{0,1\}} \times P_\theta(\boldsymbol{v}_{\mathcal{V}\backslash\mathcal{A}},h_j \mid \boldsymbol{v}_\mathcal{A},\mathbf{h}_{-j}^{(\nu)}) \tag{19}$$

ここで，$\mathbf{h}_{-j}^{(\nu)}$ は $\boldsymbol{h}_{-j} := \{h_k \mid k \in \mathcal{H}\backslash\{j\}\}$ に対応する ν 番目のサンプル点である．式 (19) に含まれる条件付き分布は，式 (5) から次のように表される．

$$\begin{aligned}&\int d\boldsymbol{v}_{\mathcal{V}\backslash\mathcal{A}}P_\theta(\boldsymbol{v}_{\mathcal{V}\backslash\mathcal{A}},h_j \mid \boldsymbol{v}_\mathcal{A},\boldsymbol{h}_{-j}) \\ &\quad = P_\theta(h_j \mid \boldsymbol{v}_\mathcal{A},\boldsymbol{h}_{-j}) \propto \exp\big[\gamma_j(\boldsymbol{h}_{-j},\boldsymbol{v}_\mathcal{A})h_j\big]\end{aligned} \tag{20}$$

ここで，$\gamma_j(\boldsymbol{h}_{-j},\boldsymbol{v}_\mathcal{A}) := \beta_j(\boldsymbol{v}_\mathcal{A}) + \sum_{k\in\mathcal{H}\backslash\{j\}} J_{j,k}(\mathcal{A})h_k$ である．式 (19) と (20) より，

$$\mathcal{E}_j(\mathbb{S},\boldsymbol{v}_\mathcal{A}) = \frac{1}{K}\sum_{\nu=1}^{K}\mathrm{sig}\big[\gamma_j(\mathbf{h}_{-j}^{(\nu)},\boldsymbol{v}_\mathcal{A})\big] \tag{21}$$

が得られる．$\mathbb{E}_\theta[v_i h_j \mid \boldsymbol{v}_\mathcal{A}]$ $(i \in \mathcal{V}\backslash\mathcal{A})$ は，和領域を $\mathcal{S} = (\mathcal{V}\backslash\mathcal{A}) \cup \{j\}$ と設定して次のように近似される．

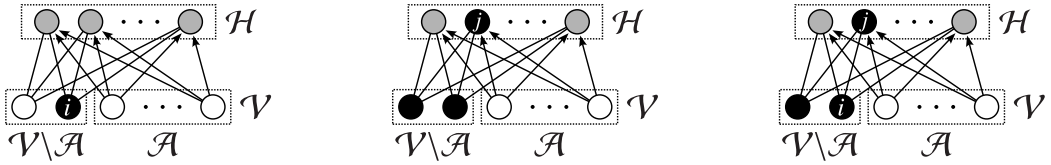
$$\mathbb{E}_\theta[v_i h_j \mid \boldsymbol{v}_\mathcal{A}] \approx \mathcal{E}_{i,j}(\mathbb{S},\boldsymbol{v}_\mathcal{A}) := \frac{1}{K}\sum_{\nu=1}^{K}\int d\boldsymbol{v}_{\mathcal{V}\backslash\mathcal{A}} \sum_{h_j\in\{0,1\}} \times v_i h_j P_\theta(\boldsymbol{v}_{\mathcal{V}\backslash\mathcal{A}},h_j \mid \boldsymbol{v}_\mathcal{A},\mathbf{h}_{-j}^{(\nu)}) \tag{22}$$

上式に含まれる条件付き分布は次のように表される．

$$\begin{aligned}&\int d\boldsymbol{v}_{\mathcal{V}\backslash(\mathcal{A}\backslash\{i\})}P_\theta(\boldsymbol{v}_{\mathcal{V}\backslash\mathcal{A}},h_j \mid \boldsymbol{v}_\mathcal{A},\boldsymbol{h}_{-j}) \\ &\quad = P_\theta(v_i,h_j \mid \boldsymbol{v}_\mathcal{A},\boldsymbol{h}_{-j}) = P_\theta(v_i \mid \boldsymbol{h})P_\theta(h_j \mid \boldsymbol{v}_\mathcal{A},\boldsymbol{h}_{-j})\end{aligned} \tag{23}$$

式 (22) と (23) より，

$$\begin{aligned}\mathcal{E}_{i,j}(\mathbb{S},\boldsymbol{v}_\mathcal{A}) = \frac{1}{K}\sum_{\nu=1}^{K}&\big[\lambda_{j,i}(\mathbf{h}^{(\nu)}) \\ &+ \mathrm{sfp}(\sigma_i)w_{i,j}\big]\,\mathrm{sig}\big[\gamma_j(\mathbf{h}_{-j}^{(\nu)},\boldsymbol{v}_\mathcal{A})\big]\end{aligned} \tag{24}$$



(a) (b) (c)

図 3 SMCI 推定量 (a) $\mathcal{E}_i^{(\nu)}(\mathbb{S})$ と $\mathcal{E}_i^{(\nu^2)}(\mathbb{S})$，(b) $\mathcal{E}_j(\mathbb{S}, \boldsymbol{v}_{\mathcal{A}})$，(c) $\mathcal{E}_{i,j}(\mathbb{S}, \boldsymbol{v}_{\mathcal{A}})$ の和領域とサンプル領域．黒色に塗られたノードが和領域に対応し，灰色に塗られた隠れ層のノードがサンプル領域に対応する．

Algorithm 1 : $\bar{\mathbb{S}}^{(\mu)}$ のサンプリング

Input: サンプル数 $\bar{K}$，ステップ数 $\bar{T}$，欠損データ点 $\mathbf{d}^{(\mu)}$
  for $\nu = 1, 2, \ldots, \bar{K}$ **do**
    式 (26) の自己無撞着方程式に基づいて，近似分布 $Q(\boldsymbol{h} \mid \mathbf{d}^{(\mu)})$ を求める．
    $\bar{\mathbf{h}}_0^{(\mu,\nu)} \sim Q(\boldsymbol{h} \mid \mathbf{d}^{(\mu)})$
    for $t = 1, 2, \ldots, \bar{T}$ **do** ▷ $\bar{T}$-step Gibbs sampling
      for $j \in \mathcal{H}$ **do**
        $\bar{\mathrm{h}}_{t,j}^{(\mu,\nu)} \sim P_\theta(h_j \mid \mathbf{d}^{(\mu)}, \bar{\mathbf{h}}_{t-1,-j}^{(\mu,\nu)})$
      end for
    end for
  end for
Output: $\{\bar{\mathbf{h}}_T^{(\mu,\nu)} \mid \nu = 1, 2, \ldots, \bar{K}\}$

Algorithm 2 : $\hat{\mathbb{S}}$ のサンプリング

Input: サンプル数 $\hat{K}$，ステップ数 $\hat{T}$
  for $\nu = 1, 2, \ldots, \hat{K}$ **do**
    ランダム初期化：$\hat{\mathbf{h}}_0$
    for $t = 1, 2, \ldots, \hat{T}$ **do** ▷ $\hat{T}$-step Gibbs sampling
      for $j \in \mathcal{H}$ **do**
        $\hat{\mathrm{h}}_{t,j}^{(\nu)} \sim P_\theta(h_j \mid \hat{\mathbf{h}}_{t-1,-j}^{(\nu)})$
      end for
    end for
  end for
Output: $\{\hat{\mathbf{h}}_T^{(\nu)} \mid \nu = 1, 2, \ldots, \hat{K}\}$

が得られる．ここで，$\lambda_{j,i}(\boldsymbol{h}) := \lambda_j(\boldsymbol{h}) - w_{i,j}\,\mathrm{sfp}(\sigma_i)h_j$ である．式 (17), (18), (21), (24) の SMCI 推定量の和領域とサンプル領域を図 3 に示す．

5 提案学習法

3 節で述べたように，式 (9)–(12) の勾配に計算困難な期待値が含まれるため，期待値近似に基づく学習法が必要となる．本節では，mGS と 4 節で議論した SMCI に基づく学習法を提案する．

式 (13) の条件付き期待値の近似を考える．まず，周辺分布 $P_\theta(\boldsymbol{h} \mid \mathbf{d}^{(\mu)})$ から $\bar{K}$ 個のサンプル点の集合を生成する．

$$\bar{\mathbb{S}}^{(\mu)} := \left\{\bar{\mathbf{h}}^{(\mu,\nu)} \mid \nu = 1, 2, \ldots, \bar{K}\right\} \tag{25}$$

式 (5) より，$P_\theta(\boldsymbol{h} \mid \mathbf{d}^{(\mu)})$ は完全グラフ上の BM の形で与えられ，この分布上でギブスサンプリング (mGS) を行うことで $\bar{\mathbb{S}}^{(\mu)}$ を生成する．$\bar{\mathbb{S}}^{(\mu)}$ の生成は各データ点毎に行われるため，可能な限り少ないステップ数で良質なサンプリングをすることが理想であり，これを達成するためにはランダムな初期値ではなく，ある程度分布に従う初期値が望ましい．提案法では以下の手順でそのような初期値を得る．まず，平均場近似 [10] を用いて周辺分布 $P_\theta(\boldsymbol{h} \mid \mathbf{d}^{(\mu)})$ の粗い近似を求める．$P_\theta(\boldsymbol{h} \mid \mathbf{d}^{(\mu)})$ の近似分布 $Q(\boldsymbol{h} \mid \mathbf{d}^{(\mu)})$ は自己無撞着な平均場方程式

$$\begin{aligned} \mathrm{r}_j^{(\mu)} &= \mathrm{sig}\left[\tau_j\left(\mathbf{v}_{\mathrm{mf}}^{(\mu)}\right)\right] \qquad (j \in \mathcal{H}) \\ \mathrm{q}_i^{(\mu)} &= \lambda_i\left(\mathbf{r}^{(\mu)}\right) \qquad (i \in \mathcal{M}^{(\mu)}) \end{aligned} \tag{26}$$

の解を用いて，$Q(\boldsymbol{h} \mid \mathbf{d}^{(\mu)}) \propto \prod_{j\in\mathcal{H}} \exp\left[\tau_j\left(\mathbf{v}_{\mathrm{mf}}^{(\mu)}\right)h_j\right]$ により得られる．ここで，$\mathbf{v}_{\mathrm{mf}}^{(\mu)} := \left\{\mathbf{d}^{(\mu)}, \mathbf{q}^{(\mu)}\right\}$ である．平均場近似は精度は低いが高速に分布を近似できる手法として知られており，その近似分布から生成された値の確率値は，ランダム値のそれよりも高いことが期待される．よって，$Q(\boldsymbol{h} \mid \mathbf{d}^{(\mu)})$ から生成した値を mGS の初期値として採用する．その後，得られた初期値を用いて mGS を行う．$\bar{\mathbb{S}}^{(\mu)}$ の生成手順をアルゴリズム 1 に示す．得られた $\bar{\mathbb{S}}^{(\mu)}$ を用いて，式 (9)–(12) の勾配に現れる条件付き期待値は，以下のように近似される．

$$\mathbb{E}_\theta[v_i \mid \mathbf{d}^{(\mu)}] \approx \mathcal{E}_i^{(\nu)}(\bar{\mathbb{S}}^{(\mu)}) \tag{27}$$
$$\mathbb{E}_\theta[v_i^2 \mid \mathbf{d}^{(\mu)}] \approx \mathcal{E}_i^{(\nu^2)}(\bar{\mathbb{S}}^{(\mu)}) \tag{28}$$
$$\mathbb{E}_\theta[h_j \mid \mathbf{d}^{(\mu)}] \approx \mathcal{E}_j(\bar{\mathbb{S}}^{(\mu)}, \mathbf{d}^{(\mu)}) \tag{29}$$
$$\mathbb{E}_\theta[v_i h_j \mid \mathbf{d}^{(\mu)}] \approx \mathcal{E}_{i,j}(\bar{\mathbb{S}}^{(\mu)}, \mathbf{d}^{(\mu)}) \tag{30}$$

次に，式 (14) の期待値の近似を考える．まず，周辺分布 $P_\theta(\boldsymbol{h})$ から $\hat{K}$ 個のサンプル点の集合を生成する．

$$\hat{\mathbb{S}} := \left\{\hat{\mathbf{h}}^{(\nu)} \mid \nu = 1, 2, \ldots, \hat{K}\right\} \tag{31}$$

2 節で述べたように，$P_\theta(\boldsymbol{h})$ は完全グラフ上の BM の形で与えられ，この分布上の mGS により $\hat{\mathbb{S}}$ を生成する．$\hat{\mathbb{S}}$ の生成手順をアルゴリズム 2 に示す．式 (9)–(12) の勾配に現れる式 (14) の期待値は，式 (27)–(30) と同様に以下のように近似される．

$$\mathbb{E}_\theta[v_i] \approx \mathcal{E}_i^{(\nu)}(\hat{\mathbb{S}}) \tag{32}$$
$$\mathbb{E}_\theta[v_i^2] \approx \mathcal{E}_i^{(\nu^2)}(\hat{\mathbb{S}}) \tag{33}$$
$$\mathbb{E}_\theta[h_j] \approx \mathcal{E}_j(\hat{\mathbb{S}}, \boldsymbol{v}_\varnothing) \tag{34}$$
$$\mathbb{E}_\theta[v_i h_j] \approx \mathcal{E}_{i,j}(\hat{\mathbb{S}}, \boldsymbol{v}_\varnothing) \tag{35}$$

以上より，提案学習法では式 (9)–(12) の勾配は式

(27)–(35) に基づいて以下のように近似される．

$$\frac{\partial \ell(\theta)}{\partial b_i} \approx \frac{1}{N}\sum_{\mu=1}^{N}\Big[I_{O^{(\mu)}}(i)\,\mathrm{d}_i^{(\mu)} + (1-I_{O^{(\mu)}}(i))\,\mathcal{E}_i^{(\nu)}(\bar{\mathbb{S}}^{(\mu)})\Big] - \mathcal{E}_i^{(\nu)}(\hat{\mathbb{S}})$$

$$\frac{\partial \ell(\theta)}{\partial c_j} \approx \frac{1}{N}\sum_{\mu=1}^{N}\mathcal{E}_j(\bar{\mathbb{S}}^{(\mu)},\mathbf{d}^{(\mu)}) - \mathcal{E}_j(\hat{\mathbb{S}},\boldsymbol{v}_\varnothing)$$

$$\frac{\partial \ell(\theta)}{\partial w_{i,j}} \approx \frac{1}{N}\sum_{\mu=1}^{N}\Big[I_{O^{(\mu)}}(i)\,\mathrm{d}_i^{(\mu)}\,\mathcal{E}_j(\bar{\mathbb{S}}^{(\mu)},\mathbf{d}^{(\mu)}) + (1-I_{O^{(\mu)}}(i))\,\mathcal{E}_{i,j}(\bar{\mathbb{S}}^{(\mu)},\mathbf{d}^{(\mu)})\Big] - \mathcal{E}_{i,j}(\hat{\mathbb{S}},\boldsymbol{v}_\varnothing)$$

$$\frac{\partial \ell(\theta)}{\partial \sigma_i} \approx \frac{\mathrm{sig}(\sigma_i)}{2\,\mathrm{sfp}^2(\sigma_i)}\left(\frac{1}{N}\sum_{\mu=1}^{N}\Big[I_{O^{(\mu)}}(i)\,(\mathrm{d}_i^{(\mu)})^2 + (1-I_{O^{(\mu)}}(i))\,\mathcal{E}_i^{(\nu^2)}(\bar{\mathbb{S}}^{(\mu)})\Big] - \mathcal{E}_i^{(\nu^2)}(\hat{\mathbb{S}})\right)$$

周辺化によって変数空間が削減されるため，mGS は元の分布上での bGS よりも緩和時間が短く，効率的である．さらに，4 節で述べたように，SMCI は標準的な MCI よりも近似精度が高い．よって，mGS と SMCI に基づく提案学習法は，bGS と MCI に基づく従来学習法よりも効率的に良質なサンプル点を生成し，高精度に期待値を近似することで高い学習性能をもつことが期待される．

6 数値実験

本節では，提案学習法の有効性を数値実験を通して検証する．本実験では学習データとして，手書き数字画像データセットの MNIST とファッション商品画像データセットの Fashion MNIST (F-MNIST) を用いた．MNIST と F-MNIST は，それぞれ 10 種類の数字画像（“0” から “9”）と 10 種類のファッション画像（T シャツやスニーカーなど）から構成される．画像はどちらも 28×28 ピクセルのグレースケールである．本実験では，各種類に対応する画像を 100 枚ずつランダムに選んだ合計 1000 枚のデータセットを使用した．画像データ点 $\mathbf{d}$ は，$\mathrm{d}_i \leftarrow 2(\mathrm{d}_i/255)-1$ により画素値が区間 $[-1,+1]$ 内に収まるように正規化した後に，各ピクセルを確率 p でランダムに欠損させた．そして，前処理をした MNIST または F-MNIST データセットを用いて，$n_v = 784$ 個の可視変数と $n_m = 300$ 個の隠れ変数をもつ GBRBM を学習した．学習 RBM のパラメータ b_i, c_j は 0 で，$w_{i,j}$ は Xavier の初期値 [11] で，σ_i は $\mathrm{sfp}(\sigma_i)=1$ となるように初期化した．学習は最適化アルゴリズムの AdaMax [12] を用いて，ミニバッチサイズが 64 のミニバッチ学習により行った．式 (13) の条件付き期待値の近似では，従来法 [3] と提案法どちらもサンプル数 $\bar{K}$ とステップ数 $\bar{T}$ を 8 に設定した．また，式 (14) の期待値の近似では，両手法ともサンプル数 $\hat{K}=64$，ステップ数 $\hat{T}=128$ とした．本実験では，学習法の性能を 3 つの対数尤度で評価する．1 つ目は式 (8) の対数尤度であり，これは学習の目的関数であるため，大きいほど学習性能が高いといえる．2 つ目は欠損処理を施す前の学習データ集合に対する対数尤度である．これは完全データ点に対する学習（標準的な最尤推定）の目的関数であり，この値が大きいほど，欠損の影響を受けにくい理想的な学習が可能であることを意味する．3 つ目は欠損のない完全テストデータ集合に対する対数尤度であり，この値が大きいほど学習の汎化性能が高いといえる．本実験では，10000 個の MNIST と F-MNIST のテストデータ集合を用意し，学習データと同様の前処理を施した．これらの対数尤度の評価には，式 (7) と (8) より，2 つの計算困難な分配関数 $Z_\theta^{(\mathrm{h})}(\boldsymbol{v}_\varnothing)$ と $Z_\theta^{(\mathrm{h})}(\mathbf{d}^{(\mu)})$ の近似計算が必要となる．そこで，周辺分布 $P_\theta(\boldsymbol{h})$ と $P_\theta(\boldsymbol{h}\mid\mathbf{d}^{(\mu)})$ に対する周辺化焼なまし重点サンプリング [13, 14] に基づいてこれらの分配関数を近似計算することで，3 つの対数尤度を評価した．本実験では，周辺化焼なまし重点サンプリングを 100 回実行して得られた推定量の平均値を用いて分配関数を近似した．

式 (8) の対数尤度の結果を図 4 に示す．また，欠損のない学習・テストデータ集合に対する対数尤度を表 1 に示す．MNIST と F-MNIST の双方のデータセットにおける 3 つの対数尤度は，従来法よりも提案法の方が安定的に上昇し，高い値となっている．これら結果から，提案法は従来法よりも高性能であることが確認できた．

7 まとめ

GBRBM における不完全データ学習では，勾配計算に要する期待値の近似精度が学習性能のボトルネックとなるため，高精度な近似に基づく学習法が求められる．そこで本研究では，bGS と MCI に基づく従来学習法を mGS と SMCI を用いて改良した新たな学習法を提案した．従来法よりも効率的に高精度な近似を行える提案学習法により，不完全データに対する学習性能が向上することを数値的に示した．GBRBM を深層化したガウシアン・ベルヌーイ深層ボルツマンマシン [15] への提案学習法の応用は今後の課題である．

謝辞

本研究は JSPS 科研費 JP24KJ0452, JP21K11778 の助成を受けたものである．

参考文献

[1] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[2] K. Cho, A. Ilin, and T. Raiko, “Improved learning of gaussian-bernoulli restricted boltzmann machines,” in *Artificial Neural Networks and Machine Learning – ICANN 2011* (T. Honkela, W. Duch, M. Girolami, and S. Kaski, eds.), (Berlin, Heidelberg), pp. 10–17, Springer Berlin Heidelberg, 2011.

[3] G. Fissore, A. Decelle, C. Furtlehner, and Y. Han, “Robust multi-output learning with highly incomplete data via restricted boltzmann machines,” *arXiv preprint arXiv:1912.09382*, 2019.

[4] M. Yasuda, “Effective sampling on gaussian-bernoulli restricted boltzmann machines,” *Nonlinear Theory and Its Applications, IEICE*, vol. 15, no. 2, pp. 217–225, 2024.

[5] R. Liao, S. Kornblith, M. Ren, D. J. Fleet, and G. Hinton, “Gaussian-bernoulli rbms without tears,” *arXiv preprint arXiv:2210.10318*, 2022.

[6] M. Yasuda and K. Uchizawa, “A generalization of spatial monte

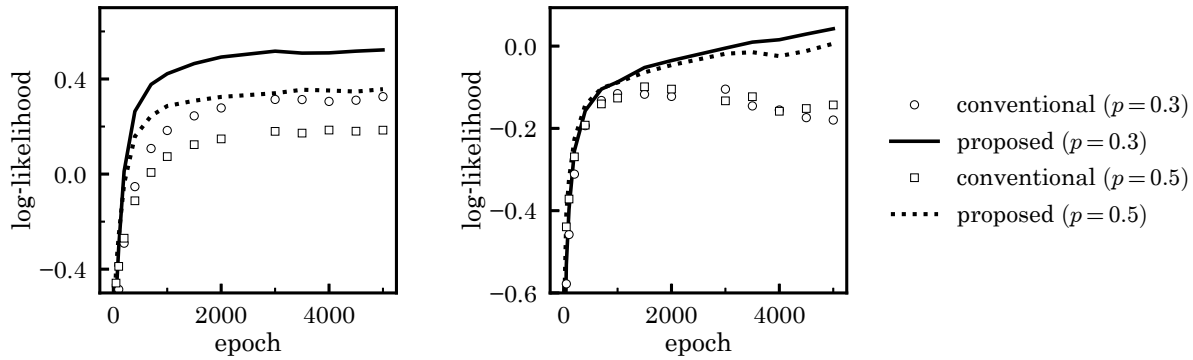


図 4　(a) MNIST，(b) F-MNIST を用いて学習したときの対数尤度．各プロットは 3 回の実験の平均値である．

表 1　欠損のない学習・テストデータ集合に対する対数尤度．各値は 3 回の実験の平均値である．

	完全データ点の集合に対する対数尤度							
	MNIST (p = 0.3)		MNIST (p = 0.5)		F-MNIST (p = 0.3)		F-MNIST (p = 0.5)	
	Train	Test	Train	Test	Train	Test	Train	Test
従来学習法 [3]	0.47	0.33	0.39	0.28	−0.21	−0.38	−0.21	−0.34
提案学習法	0.79	0.50	0.65	0.44	0.02	−0.13	−0.05	−0.17

carlo integration," *Neural Computation*, vol. 33, no. 4, pp. 1037–1062, 2021.

[7] M. Yasuda, "Learning algorithm of boltzmann machine based on spatial monte carlo integration method," *Algorithms*, vol. 11, no. 4, p. 42, 2018.

[8] K. Sekimoto and M. Yasuda, "Effective learning algorithm for restricted boltzmann machines via spatial monte carlo integration," *Nonlinear Theory and Its Applications, IEICE*, vol. 14, no. 2, pp. 228–241, 2023.

[9] J. S. Liu, *Monte Carlo strategies in scientific computing*. Springer, 2001.

[10] C. Takahashi and M. Yasuda, "Mean-field inference in gaussian restricted boltzmann machine," *Journal of the Physical Society of Japan*, vol. 85, no. 3, p. 034001, 2016.

[11] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *In Proc. of the 13th International Conference on Artificial Intelligence and Statistics*, vol. 9, pp. 249–256, 2010.

[12] D. P. Kingma and L. J. Ba, "Adam: A method for stochastic optimization," *In Proc. of the 3rd International Conference on Learning Representations*, pp. 1–13, 2015.

[13] M. Yasuda and C. Takahashi, "Free energy evaluation using marginalized annealed importance sampling," *Phys. Rev. E*, vol. 106, p. 024127, 2022.

[14] R. M. Neal, "Annealed importance sampling," *Statistics and computing*, vol. 11, no. 2, pp. 125–139, 2001.

[15] K. H. Cho, T. Raiko, and A. Ilin, "Gaussian-bernoulli deep boltzmann machine," in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, 2013.