Technical Report of the 3rd Workshop on Latent Dynamics (Sep 24, 2012, Tokyo, Japan)

量子揺らぎと Latent Dynamics

佐藤一誠*
Issei Sato

Abstract: 確率的潜在変数モデルは，データの背後にある潜在的な情報を確率変数としてモデル化し，データの理解や予測などを可能とする数理モデルである．本稿では，この潜在変数に対して量子揺らぎを導入し変分ベイズ法により学習する手法 [1] を紹介する．変分ベイズ法は，統計的機械学習，特に確率的潜在変数モデルの学習において最も多く用いられている学習アルゴリズムの一つであるが，局所解に陥るという問題がある．量子揺らぎを導入することで変分ベイズ法で陥る局所解からより良い解への状態遷移が期待される．

Keywords: 確率的潜在変数モデル，量子揺らぎ，量子アニーリング，変分ベイズ法

1　はじめに

確率的潜在変数モデルは，潜在変数と呼ばれる確率変数をデータの生成過程のモデル化に導入した数理モデルである．潜在変数を導入することで，データの隠れた性質を抽出し，予測やデータ解析に役立てることができる．確率的潜在変数モデルの学習は，多数の局所解を持つ非線形最適化問題として定式化される．このような場合の１つのアプローチとして， Deterministic Annealing (DA) [2] がある．DA は，Simulated Annealing (SA) [3] に基づき，温度を模したパラメータを導入し，熱揺らぎを制御することで，より最適な解を探索する手法である．近年，量子情報理論では熱揺らぎとは異なる揺らぎとして量子揺らぎを用いた Quantum annealing (QA) が注目を集めている [4, 5, 6]．本稿では，統計的機械学習分野で幅広く用いられている学習アルゴリズムである変分ベイズ法 [7] に対して量子揺らぎを導入した研究 [1] について紹介する．量子揺らぎを導入することで，古典系とは異なる空間での学習が可能となる．これにより古典系では起こることがないような学習プロセス（状態遷移）が起こり効率的な学習が可能になることが期待されている．

2　変分ベイズ法の概要

データ数を N，データを $\boldsymbol{x}(=(x_1,\cdots,x_N))$ と表記する．データの生成過程では，各データ x_i は，各々潜在変数 z_i に依存していると仮定する．本研究では，潜在変数 z_i は離散値 $\{1,2,\cdots,K\}$ を取ると仮定する．例えば，確率的混合モデルの場合には，z_i は，潜在クラスを意味し，K はクラス数もしくはコンポーネント数である．以下，確率的混合モデルを例に話を進める．

N データに対する潜在変数割り当ての総数を $L(=K^N)$ とする．潜在変数によるクラスタリングの場合には，N データに対する潜在変数割り当ての総数とは，クラス割り当ての総数を意味する．次の節で，量子系に拡張するために，$\boldsymbol{z}=(z_1,z_2,\cdots,z_N)$ の状態のとり方に対応する変数を導入する．N データに対する１つの潜在変数割り当てを $\sigma^{(\ell)}(\ell=1,2,\cdots,L)$ とする．$\sigma^{(\ell)}$ は，ℓ 番目の要素が 1 でそれ以外が 0 である L 次元指標ベクトルとする．また $\sigma=\{\sigma^{(\ell)}\}_{l=1}^{L}$ とする．例えば，$N=3$,$K=2$ のとき，(z_1,z_2,z_3) の取りうるパターンは，$(1,1,1)$, $(2,1,1)$, $(1,2,1)$, $(2,1,1)$, $(2,2,1)$, $(2,1,2)$, $(1,2,2)$, $(2,2,2)$ の $L=2^3=8$ 通りある．$\sigma^{(\ell)}$ $(\ell=1,2,\cdots,8)$ は，これらのパターンに対応する変数である．つまり，$\sigma^{(1)}=(1,0,0,0,0,0,0,0)$ が $(z_1,z_2,z_3)=(1,1,1)$，$\sigma^{(2)}=(0,1,0,0,0,0,0,0)$ が $(z_1,z_2,z_3)=(2,1,1)$ を表すなど．$\boldsymbol{\theta}$ をモデルパラメータの集合とする．

変分ベイズ法では，N データの対数尤度 $\log p(\boldsymbol{x})$ の $\sigma,\boldsymbol{\theta}$ に関する周辺対数尤度の下限を考える．

$$
\begin{aligned}
\log p(\boldsymbol{x}) &= \int \sum_{\ell=1}^{L} \log p(\boldsymbol{x},\sigma^{(\ell)},\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&\geq \int \sum_{\ell=1}^{L} q(\sigma^{(\ell)},\boldsymbol{\theta}) \log \frac{p(\boldsymbol{x},\sigma^{(\ell)},\boldsymbol{\theta})}{q(\sigma^{(\ell)},\boldsymbol{\theta})} d\boldsymbol{\theta} \qquad (1)
\end{aligned}
$$

ここで

$$
F_c[q] = \int \sum_{\ell=1}^{L} q(\sigma^{(\ell)},\boldsymbol{\theta}) \log \frac{p(\boldsymbol{x},\sigma^{(\ell)},\boldsymbol{\theta})}{q(\sigma^{(\ell)},\boldsymbol{\theta})} d\boldsymbol{\theta} \qquad (2)
$$

*東京大学　情報基盤センター，〒 113-0033 東京都文京区本郷 7-3-1, e-mail sato@r.dl.itc.u-tokyo.ac.jp,
Academic Information Science Research Division Information Technology Center of University of Tokyo 113-0033 7-3-1 Hongo, Bunkyo-ku, Tokyo

とする．この変分下限 $F_c[q]$ を最大にする $q(\sigma, \boldsymbol{\theta})$ を求めるのが変分ベイズ法である．

3 量子揺らぎを導入した変分ベイズ法

量子揺らぎを導入するために，周辺対数尤度をハミルトニアンを用いて再定式化する．

エネルギ関数を $E[\sigma^{(\ell)}] = -\log p(\boldsymbol{x}, \sigma^{(\ell)})$ とし，ハミルトニアン $\mathcal{H}_c$ を以下のように定義する．

$$\mathcal{H}_c = \begin{pmatrix} E[\sigma^{(1)}] & & & \\ & E[\sigma^{(2)}] & & 0 \\ 0 & & \ddots & \\ & & & E[\sigma^{(L)}] \end{pmatrix}. \quad (3)$$

データ $\boldsymbol{x}$ は，ハミルトニアン $\mathcal{H}_c$ を用いて以下のように表すことができる．

$$\log p(\boldsymbol{x}) = \sum_{l=1}^{L} \log p(\boldsymbol{x}, \sigma^{(\ell)}) = \log \mathrm{Tr}\{e^{-\mathcal{H}_c})\} \quad (4)$$

したがって，変分ベイズ法は，$\log \mathrm{Tr}\{e^{-\mathcal{H}_c})\})$ の $\sigma, \boldsymbol{\theta}$ に関する周辺対数尤度の下限を最大にする $q(\sigma, \boldsymbol{\theta})$ を推定する手法であると考えられる．

ハミルトニアン $\mathcal{H}_c$ に対して量子揺らぎを制御するパラメータ Γ を導入したハミルトニアン $\mathcal{H}$ を以下のようにして定式化する．

$$\mathcal{H} = \mathcal{H}_c + \mathcal{H}_q \quad (5)$$

ここで

$$\mathcal{H}_q = \sum_{i=1}^{N} \sigma_{xi} \quad (6)$$

$$\sigma_{xi} = \left(\bigotimes_{j=1}^{i-1} \mathbb{E}_k \right) \otimes \sigma_x \otimes \left(\bigotimes_{l=i+1}^{N} \mathbb{E}_k \right) \quad (7)$$

$$\sigma_x = \Gamma(\mathbb{E}_k - 1_k) \quad (8)$$

$$1_k = \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{pmatrix} \qquad (k \text{ by } k \text{ matrix}) \quad (9)$$

$$\mathbb{E}_k = k \times k \text{ 単位行列} \quad (10)$$

$\bigotimes$ はクロネッカー積を示す．$\mathcal{H}_q$ の導入は，対角行列である $\mathcal{H}_c$ の非対角項に対して Γ を導入することを意味する．$\Gamma = 0$ の場合, $\mathcal{H} = \mathcal{H}_c$ で古典系のハミルトニアンとなり，$\Gamma \neq 0$ の場合は，非対角要素の古典状態 ($\mathcal{H}_c$ の固有ベクトル) 間の量子力学的遷移が引き起こされる．この状態遷移をパラメータ Γ によって制御することにより量子揺らぎが導入される．

量子系ハミルトニアン $\mathcal{H}$ のダイナミクスは，シュレディンガー方程式

$$i\hbar \frac{\partial}{\partial t} |\sigma(t)\rangle = \mathcal{H} |\sigma(t)\rangle \quad (11)$$

に従う．ここで $|\sigma(t)\rangle$ は時刻 t での状態ケットベクトル，i は虚数で $\hbar$ は換算プランク定数．もしくは，微小時間 Δt での時間発展演算子 $e^{-i\mathcal{H}\Delta t/\hbar}$ を用いて

$$|\sigma(t + \Delta t)\rangle = e^{-i\mathcal{H}\Delta t/\hbar} |\sigma(t)\rangle \quad (12)$$

により状態ベクトルの時間発展を追うことができる．しかしながら，機械学習での応用を考えた場合，N は非常に大きく，それに伴い K も大きく取る必要があり，$\mathcal{H}$ は高次元な行列となり現実的には実行不可能である．また，統計的機械学習では状態ベクトル (潜在変数) の事後分布に興味があることが多いため $q(\sigma)$ を求めたい．したがって，変分ベイズ法の枠組みを用いる．

逆温度 β を導入して，$\log \mathrm{Tr}\{e^{-\mathcal{H}_c}\}$ を $\log \mathrm{Tr}\{e^{-\beta\mathcal{H}}\}$ と拡張する．この $\log \mathrm{Tr}\{e^{-\beta\mathcal{H}})\}$ の $\sigma, \boldsymbol{\theta}$ に関する周辺対数尤度の下限を最大にする $q(\sigma, \boldsymbol{\theta})$ を推定することによって，量子効果を考慮した変分ベイズ法を導出する．すなわち，

$$\log p(\boldsymbol{x}) = \log \mathrm{Tr}\{e^{-\beta\mathcal{H}}\} \geq F[q] \quad (13)$$

となる $F[q]$ を最大にする変分事後分布 $q(\sigma, \boldsymbol{\theta})$ を求める．$F[q]$ を求める 1 つの方法として，$\mathcal{H}$ の固有値分解を伴いることが考えられるが，シュレディンガー方程式を解く場合と同様に計算量的には現実的ではない．したがって，$\mathrm{Tr}\{e^{-\beta\mathcal{H}}\}$ を鈴木トロッター展開 [8, 9] に基づき経路積分表示することで変分下限を導出する．

経路積分表示を行うと実時間とは別にトロッター (虚時間) 軸と呼ばれる軸が導入される．j をトロッター軸上で異なるトロッター平面 (虚時刻) を識別する添字とする．実際に変分ベイズ法を古典計算機上で実行する場合は，複数の変分ベイズ法を動かし，それぞれのシミュレーションがトロッター平面に対応する．つまり j は並列実行する際のプロセスの識別番号である．今トロッター数 (プロセスの並列数) を m とする．また σ_j $(j = 1, 2, \cdots, m)$ をプロセス j での潜在変数の割り当て状態とする．$\sigma_{j,i}$ を，プロセス j におけるデータ z_i に対応する K 次元指標ベクトルとし，z_i 番目の要素が 1 で，それ以外を 0 とする．$\sigma_j = \bigotimes_{i=1}^{N} \sigma_{j,i}$ とする．

導出された変分下限は以下のようになる (詳細は [1] を

参照).

$$F[q] = \frac{\beta}{m}\sum_{j=1}^{m} F_c[q(\sigma_j,\theta_j)] + f(\beta,\Gamma)R[q(\sigma_1,\sigma_2,\cdots,\sigma_m)] \tag{14}$$

$$R[q(\sigma_1,\sigma_2,\cdots,\sigma_m)] = \sum_{j=1}^{m}\sum_{\sigma_j}\sum_{\sigma_{j+1}} q(\sigma_j)q(\sigma_{j+1})s(\sigma_j,\sigma_{j+1}) \tag{15}$$

$$s(\sigma_j,\sigma_{j+1}) = \sum_{i=1}^{N}\delta(\sigma_{j,i},\sigma_{j+1,i}) \tag{16}$$

$$f(\beta,\Gamma) = \log\left(\frac{e^{-\frac{\beta\Gamma}{m}} + \frac{1}{k}e^{-\frac{\beta\Gamma}{m}(1-K)} - \frac{1}{K}e^{-\frac{\beta\Gamma}{m}}}{\frac{1}{k}e^{-\frac{\beta\Gamma}{m}(1-K)} - \frac{1}{K}e^{-\frac{\beta\Gamma}{m}}}\right) \tag{17}$$

$F[q]$ は，各プロセス j における古典系の変分下限 $F_c[q(\sigma_j,\theta_j)]$ の和と，各々のプロセスにおける潜在変数の事後分布に関する制約項 $R(\cdot)$ によって定式化される．$f(\beta,\Gamma)$ は，制約項 $R(\cdot)$ の重みパラメータでこの関数を制御することで量子揺らぎを制御する．

このF$[q]$ を最大にする事後分布 $q(\sigma_j,\theta_j)$ $(j=1,2,\cdots,m)$ を求めることで量子揺らぎを導入した変分ベイズ法が可能となる．ここで重要なことは，$R(\cdot)$ 及び $f(\beta,\Gamma)$ はモデルに独立である．つまり，古典系の変分下限 $F_c[q(\sigma_j,\theta_j)]$ さえ導出できていれば，$R(\cdot)$ 及び $f(\beta,\Gamma)$ を追加して最適化するだけで，これまで提案されてきた様々なモデルに量子揺らぎを導入して学習することが可能である．

4 Quantum Latent Dynamics

本稿で説明した量子揺らぎを導入した変分ベイズ法では，1つのデータに対して複数の潜在変数の事後分布 $q(\sigma_1,\sigma_2,\cdots,\sigma_m) = \prod_{j=1}^{m} q(\sigma_j)$ を求めることになる．この複数の確率分布 $\{q(\sigma_j)\}_{j=1}^{m}$ は虚時間軸方向での潜在変数のダイナミクスを表現しており，量子揺らぎにより虚時間軸方向でどのような現象が起こっているのかを解析する1つの手段になると考えられる．

5 おわりに

本稿では，変分ベイズ法 [7] に対して量子揺らぎを試みた研究 [1] について簡単に紹介した．変分ベイズ法は，機械学習分野において様々なモデルで適用されている．したがって，学習アルゴリズムの研究としては，変分ベイズ法で学習可能なモデル全般に適用可能であるようにアルゴリズムを構成することは重要であり，今回紹介したアルゴリズムはまさにそのような構成になっている．このことは，機械学習分野に対して量子揺らぎの制御による学習という新しい領域を開き，さらに量子揺らぎを導入可能なモデルの幅を広げたという2つの貢献があると考えらる．最後に，ここで提案されているアルゴリズムは並列計算と相性が良いため今後は並列化技術との組み合わせもまた重要になってくると考えられる．

参考文献

[1] I. Sato, K. Kurihara, S. Tanaka, H. Nakagawa, and S. Miyashita. Quantum Annealing for Variational Bayes Inference. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, 2009.

[2] N. Ueda and R. Nakano. Deterministic Annealing EM Algorithm. *Neural Networks*, 11(2):271–282, 1998.

[3] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, 1983.

[4] T. Kadowaki and H. Nishimori. Quantum Annealing in the Transverse Ising Model. *Physical Review E*, 58:5355–5363, 1998.

[5] E. Farhi, J. Goldstone, S. Gutmann, A. L. J. Lapan, and D. Preda. A Quantum Adiabatic Evolution Algorithm Applied to Random Instances of an NP -complete Problem. *Science*, 292:472–476, 2001.

[6] G. E. Santoro, R. Martoňák, E. Tosatti, and R. Car. Theory of Quantum Annealing of an Ising Spin Glass. *Science*, 295:2427–2430, 2002.

[7] H. Attias. Inferring Parameters and Structure of Latent Variable Models by Variational Bayes. In K. B. Laskey and H. Prade, editors, *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pages 21–30, 1999.

[8] H. F. Trotter. On the Product of Semi-Groups of Operators. *Proceedings of the American Mathematical Society*, 10(4):545–551, 1959.

[9] M. Suzuki. Relationship between d-Dimensional Quantal Spin Systems and $(d+1)$-Dimensional Ising Systems – Equivalence , Critical Exponents and Systematic Approximants of the Partition Function and Spin Correlations –. *Progress of Theoretical Physics*, 56(5):1454–1469, 1976.