平均場近似の情報幾何

田中 利幸

東京都立大学大学院工学研究科

1 はじめに

学習の問題を統計学的に捉えると，パラメトリックなデータ生成モデルから生成されたデータ集合が与えられたとき，データを要約する統計量からモデルパラメータを推定する問題だということができる．ボルツマンマシンの学習は，そのような問題の一例である．このような問題は，データ生成過程を「順過程」とみなすならばいわゆる「逆問題」だということができ，Ackley, Hinton, Sejnowski のボルツマンマシン学習則 [1] は，順過程を模擬するモデルを利用した誤差フィードバックによってこの問題を解く方法だということができる．学習の問題に対してこうしたアプローチをとる際には，「順問題」すなわちモデルパラメータからデータに関する統計量がどのように定まるのか，という問題を詳しく調べることが重要である．

順問題を解くことも一般には簡単ではなく，ボルツマンマシンの場合には，変数の数 N の指数オーダーの計算量が必要となる．この困難を回避するアプローチは大きく分けて 2 通りある．ひとつは，ギブスサンプラーなどを使ってデータ生成過程を模擬する動的モンテカルロ法のアプローチである．けれどもこのアプローチも，データに関する統計量を精度よく求めるためには多量のデータを必要とし，やはり計算時間の問題が残る．もうひとつは，何らかの近似をほどこすことによって順問題を解析的に解こうとするものである．本稿で扱う平均場近似法 [2] は，そのようなアプローチのひとつである．

ニューラルネットワークの分野では，これまでにいわゆるナイーブな平均場近似がおもに使われてきており，それについては，Kullback ダイバージェンスの最小化という形での近似の情報理論的な意味づけもすでになされている [3, 4]．いっぽうで，TAP のアプローチ [5, 6] や線形応答定理 [7] などのより進んだ統計物理学の手法も応用が試みられているが，情報理論の立場からのそれらの解釈はなされていない．

本稿では，情報幾何学 [8, 9] を使って，情報理論の立場からの平均場近似法の定式化を示す．また，TAP のアプローチや線形応答定理も，ここでの定式化から自然に導かれることを示す．

2 情報幾何による平均場近似の定式化

状態変数 $\boldsymbol{s} = (s_1, \ldots, s_N) \in \{-1, 1\}^N$，パラメータ h^i, w^{ij} $(i, j = 1, \ldots, N)$ をもつボルツマンマシンは，ボルツマン・ギブス分布

$$p(\boldsymbol{s}) = \exp[\sum_i h^i s_i + \sum_{(ij)} w^{ij} s_i s_j - \psi] \qquad (1)$$

を与えるパラメトリックなデータ生成モデルだとみなすことができる．式 (1) の形で書かれるすべての分布の集合 $\mathcal{M}$ を，ここではモデルとよぶ．モデル $\mathcal{M}$ は $\{\theta^i \equiv h^i\}$, $\{\theta^{ij} \equiv w^{ij}\}$ を正準パラメータとする指数分布族であり，$\{\eta_i \equiv \langle s_i \rangle\}$, $\{\eta_{ij} \equiv \langle s_i s_j \rangle\}$ はそれと双対な期待値パラメータを構成する．

モデル $\mathcal{M}$ に対して，以下のような直交双対な葉層構造 $\mathcal{F} = \{\mathcal{F}(\boldsymbol{w})\}$, $\mathcal{A} = \{\mathcal{A}(\boldsymbol{m})\}$ を導入する (図 1)．

$$\mathcal{M} = \bigcup_{\boldsymbol{w}} \mathcal{F}(\boldsymbol{w}) = \bigcup_{\boldsymbol{m}} \mathcal{A}(\boldsymbol{m}) \qquad (2)$$

$$\mathcal{F}(\boldsymbol{w}) = \{p \mid \theta^{ij}(p) = w^{ij}\} \qquad (3)$$

$$\mathcal{A}(\boldsymbol{m}) = \{p \mid \eta_i(p) = m_i\} \qquad (4)$$

$q \in \mathcal{M}$ の正準パラメータが与えられているものとし，q の期待値パラメータを求めるために，

> q が属する葉 $\mathcal{F}(\boldsymbol{w})$ の上で，q にもっとも近い分布 $p \in \mathcal{F}(\boldsymbol{w})$ を求める．

という問題を考える．近さの尺度として Kullback ダイバージェンス

$$D(p\|q) = \mathrm{Tr}_{\boldsymbol{s}}\, p(\boldsymbol{s}) \log \frac{p(\boldsymbol{s})}{q(\boldsymbol{s})} \qquad (5)$$

をとれば，与えられた q に対して $D(p\|q)$ を最小にする $p \in \mathcal{F}(\boldsymbol{w})$ を求める問題となる．$p, q \in \mathcal{F}(\boldsymbol{w})$

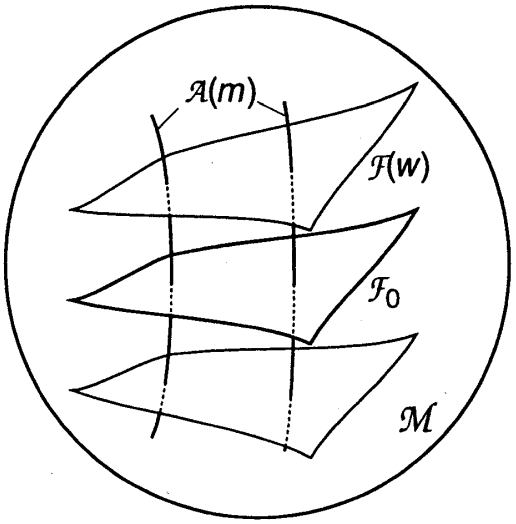

図 1　直交双対な葉層構造

に対しての $D(p\|q)$ は，$\mathcal{F}(\boldsymbol{w})$ 上の双対なポテンシャル関数

$$\tilde{\psi} \equiv \psi, \quad \tilde{\phi} \equiv \sum_i \theta^i \eta_i - \tilde{\psi} \tag{6}$$

を使って

$$D(p\|q) = \tilde{\psi}(q) + \tilde{\phi}(p) - \sum_i \theta^i(q)\eta_i(p) \tag{7}$$

とあらわされるが，$\tilde{\psi}(q)$ は p によらないので，問題は

$$G(p) = \tilde{\phi}(p) - \sum_i \theta^i(q)\eta_i(p) \tag{8}$$

の最小化問題に帰着する．式 (8) の右辺第 1 項，第 2 項はそれぞれ，p のギブス自由エネルギー，Zeeman エネルギーに相当する．

$p = q$ がこの最小化問題の自明な解だが，この最小化問題を $\{m_i = \eta_i(p)\}$ を独立変数として解くことによって，$p = q$ の期待値パラメータを得ることができる．これを実際に実行するためには，$G(p)$ の第 1 項 $\tilde{\phi}(p)$ が $\{\eta_i(p)\}$ の関数として陽に与えられている必要があるが，多くの場合そのような陽な表式は得られない．そこで，部分モデル $\mathcal{F}(\mathbf{0})$ 上では各変数が統計的に互いに独立になり，問題がしばしば簡単になることに注目して，$\tilde{\phi}(p)$ を $\boldsymbol{w} = \mathbf{0}$ に関してテイラー展開することを考える．この展開は Plefka 展開 [10] とよばれており，$\mathcal{F}$ と双対な葉層構造 $\mathcal{A}$ を利用して組織的に展開の係数を求めることができる．

添字 ij を I などと略記し，$\partial_I \equiv \partial/\partial w^I$ などと表記する．

定理　$p \in \mathcal{A}(\boldsymbol{m})$ とし，$p_0 \in \mathcal{A}(\boldsymbol{m}) \cap \mathcal{F}(\mathbf{0})$ とする．$\tilde{\phi}(p)$ のテイラー展開の 1 次，2 次，3 次の係数は，符号を除いてそれぞれ対応する次数のキュムラントテンソルに等しい．

$$\begin{aligned}
\partial_I \tilde{\phi}(p_0) &= -\eta_I(p_0) \\
\partial_I \partial_J \tilde{\phi}(p_0) &= -\big\langle (\partial_I \ell)(\partial_J \ell) \big\rangle_{p_0} \\
\partial_I \partial_J \partial_K \tilde{\phi}(p_0) &= -\big\langle (\partial_I \ell)(\partial_J \ell)(\partial_K \ell) \big\rangle_{p_0}
\end{aligned} \tag{9}$$

ただし，$\ell \equiv \log p_0$ である．

テイラー展開の 4 次以上の項の係数は，一般には対応する次数のキュムラントテンソルとは一致しない．

Plefka 展開を n 次項までで打ち切ることによって，級数の収束性を問題にしなければ，G に対する n 次近似 G_n が得られる．G_1 は Weiss 自由エネルギーであり，G_2 は SK モデルに対する TAP 自由エネルギーである．SK モデルでは Plefka 展開の 3 次以上の項は熱力学的極限 $N \to \infty$ で消えるため G_2 は正確な自由エネルギーを与える [10] が，より一般のモデルの場合，とくに応用上重要な N が有限の場合には高次項は消えず，それを打ち切ることによって実際に近似をおこなっていることになる．

G の最小化問題の解は，停留条件 $\partial G/\partial m_i = 0$ を $\boldsymbol{m}$ を変数として解くことで求められる．平均場近似では，G のかわりにその n 次近似 G_n の停留条件から $\boldsymbol{m}$ を求める．G_n の停留条件を与える式 $\partial G_n/\partial m_i = 0$ を，n 次の平均場近似における平均場方程式とよぶ．

また，線形応答定理は，葉 $\mathcal{F}(\boldsymbol{w})$ での Fisher 計量に関する自明な恒等式として得られる．葉 $\mathcal{F}(\boldsymbol{w})$ での Fisher 計量 (g_{ij}) は，

$$\begin{aligned}
g_{ij} &= \partial_i \partial_j \tilde{\psi}(p) \\
&= \eta_{ij}(p) - \eta_i(p)\eta_j(p)
\end{aligned} \tag{10}$$

で与えられる．その逆行列 $(g^{ij}) \equiv (g_{ij})^{-1}$ は

$$g^{ij} = \partial^i \partial^j \tilde{\phi}(p) \tag{11}$$

で与えられる．ただし，$\partial_i \equiv \partial/\partial\theta^i$, $\partial^i \equiv \partial/\partial m_i$ である．式 (8) から，

$$g_{ij} = \partial^i \partial^j G \tag{12}$$

である．統計物理学でいえば，$(g_{ij}), (g^{ij})$ はそれぞれ感受性行列，安定性行列であり，線形応答定理はこれらが互いに逆行列の関係にあることを主張している．

すでに述べたように，G が $\{m_i\}$ の関数として陽に与えられることは一般には期待できず，したがって式 (12) の微分を解析的に計算することはできない．平均場近似では G をその n 次近似 G_n で代用することによって近似的にこれを計算する．これが線形応答定理の n 次近似を構成する．

3 考察

$p, q \in \mathcal{F}(\boldsymbol{w})$ に対する Kullback ダイバージェンス $D(p\|q)$ は，p と同じ葉 $\mathcal{A}(\boldsymbol{m})$ に属する $p_0 \in \mathcal{F}(\boldsymbol{0})$ によって

$$D(p\|q) = D(p_0\|q) - D(p_0\|p) \qquad (13)$$

とあらわされることが，情報幾何における拡張されたピタゴラスの定理 [8, 9] から示される．$D(p_0\|p)$ は $\boldsymbol{w}$ に関して 2 次のオーダーの量であり，1 次の平均場近似を考えるときには無視されるべきものである．その結果，$D(p\|q)$ を最小にする $p \in \mathcal{F}(\boldsymbol{w})$ を求める問題は，$D(p_0\|q)$ を最小にする $p_0 \in \mathcal{F}(\boldsymbol{0})$ を求める問題に帰着するが，これがナイーブな平均場近似に対して情報理論の立場からなされてきた解釈 [3, 4] である．逆に，不規則系に対する平均場方程式にあらわれる反跳場補正は，部分モデル $\mathcal{F}(\boldsymbol{w})$ と $\mathcal{F}(\boldsymbol{0})$ とのずれ $D(p_0\|p)$ からくる効果を補正している，という解釈が成立する．

本稿では，ボルツマンマシンをとりあげてその平均場近似の定式化を示した．けれどもここでの議論は，支配測度が各変数に関する測度の直積測度であるような指数型分布族がモデルである場合に対してもそのまま適用できる．たとえば，より高次の相互作用を含む高次ボルツマンマシンへの平均場近似の適用が考えられる．興味深いのは，高次ボルツマンマシンに対して線形応答定理を自然に拡張することが可能である，ということである．この場合には，3 階以上のより高階の微分係数についての恒等式が線形応答定理の拡張を与える．べつの拡張として，Potts スピン系への適用が考えられる．Potts スピン系のパターン認識課題への応用に対しても TAP のアプローチの適用がすでに試みられている [6] が，ここでの定式化はそのような試みに対しても理論的な基礎を提供するものだと考えている．

最後に，実用的な側面について考察しておく．平均場近似の基礎となる Plefka 展開は本質的にはテイラー展開だから，その精度については一般的に，以下のような予想が可能である．

予想 摂動パラメータ $\boldsymbol{w}$ の大きさが小さいときには精度がよく，より高次の近似のほうが精度も高い．逆に，摂動パラメータ $\boldsymbol{w}$ の大きさが大きくなると精度は低下し，高次の近似をするほうがかえって精度は悪くなる．

ボルツマンマシンに対する数値実験の結果，上記の予想が妥当であることを確認することができた．

参考文献

[1] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "A learning algorithm for Boltzmann machines," *Cognitive Science*, vol. 9, pp. 147–169, 1985.

[2] C. Peterson and J. R. Anderson, "A mean field theory learning algorithm for neural networks," *Complex Systems*, vol. 1, pp. 995–1019, 1987.

[3] T. Tanaka, "Information geometry of mean field theory," *IEICE Trans. Fundamentals*, vol. E79-A, no. 5, pp. 709–715, May 1996.

[4] L. K. Saul and M. I. Jordan, "Exploiting tractable substructures in intractable networks," in D. S. Touretzky et al. (eds.), *Advances in Neural Information Processing Systems*, vol. 8, pp. 486–492, The MIT Press, 1996.

[5] C. C. Galland, "The limitations of deterministic Boltzmann machine learning," *Network: Comput. Neural Syst.*, vol. 4, no. 3, pp. 355–379, Aug. 1993.

[6] T. Hofmann and J. M. Buhmann, "Pairwise data clustering by deterministic annealing," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 19, no. 1, pp. 1–14, Jan. 1997; Errata, *ibid.*, vol. 19, no. 2, p. 197, Feb. 1997.

[7] H. J. Kappen and F. B. Rodríguez, "Efficient learning in Boltzmann machines using linear response theory," to appear in *Neural Computation*, 1998.

[8] S.-I. Amari, *Differential-Geometrical Method in Statistics*, Lecture Notes in Statistics, vol. 28, Springer, 1985.

[9] 甘利，長岡，情報幾何の方法，岩波講座応用数学，岩波書店，1993.

[10] T. Plefka, "Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model," *J. Phys. A: Math. Gen.*, vol. 15, no. 6, pp. 1971–1978, Jun. 1982.