相関のある入力を用いた学習に関する統計力学的解析

積　千洋 *　櫻井 信吾 *　松野 雅文 †　三好 誠司 ‡

1　はじめに

学習とは観測データを用いてその背後にあるデータの生成過程を推定することである．教師つき学習においては観測データは教師の入出力であり，これは例題とも呼ばれる．学習はバッチ学習とオンライン学習 [1, 3] に大別できる．バッチ学習においては与えられたいくつかの例題を繰り返し使用する．この場合，生徒が適切な自由度を持っていればすべての例題に正しく答えられるようになるが，それまでに長い時間が必要である．また，多くの例題を蓄えておくためのメモリが必要である．これに対してオンライン学習では一度使った例題は捨ててしまう．この場合，過去に使った例題に対して生徒が必ず正しく答えられるとは限らないが，多くの例題を蓄えておくためのメモリが不要であり，また時間的に変化する教師にも追随できるなどの利点がある．

本稿では相関のある入力を使う学習の解析をする．手書き文字認識を例に取ると，文字を認識するためには，まず例題を集めなければならない．そのとき，ある程度の人数を集めてひとりひとりにいろいろな字を書いてもらうわけだが，同一人物が書く文字はよく似ている．すなわち，相関がある．このように実際のパターン認識への応用を考えた場合，相関がある入力に対するふるまいを理論的に解析しておくことは有意義である．

本稿では理論的なとり扱いを容易にするため学習機械は線形パーセプトロンとし，入力が相関を持つような場合の学習についてオンライン学習の枠組みで統計力学的手法を用いた解析 [2, 3, 4, 5, 6] を行う．

2　モデル

本稿では教師，生徒の二個の線形パーセプトロンを考え，それぞれの結合荷重を $\boldsymbol{B}$，$\boldsymbol{J}$ とする．なお，本稿では簡単のため，教師の結合荷重，生徒の結合荷重のことをそれぞれ単に教師，生徒と呼ぶことにする．教師 $\boldsymbol{B}$=$(B_1, \cdots, B_N)$，生徒 $\boldsymbol{J}$=$(J_1, \cdots, J_N)$ および入力 $\boldsymbol{x}$=$(x_1, \cdots, x_N)$ は N 次元ベクトルであり，教師 $\boldsymbol{B}$ の各要素 B_i は平均 0，分散 1 のガウス分布にしたがい独立に生成され，不変であるとする．$\boldsymbol{J}$ の初期値 $\boldsymbol{J}^0$ の各要素 $\boldsymbol{J}_i^0$ は平均 0，分散 1 のガウス分布に従い独立に生成されるものとする．すなわち，

$$\langle \boldsymbol{B}_i \rangle = 0, \quad \langle (\boldsymbol{B}_i)^2 \rangle = 1, \tag{1}$$

$$\langle \boldsymbol{J}_i^0 \rangle = 0, \quad \langle (\boldsymbol{J}_i^0)^2 \rangle = 1, \tag{2}$$

ここで，$\langle \cdot \rangle$ は平均をあらわす．

また，$\boldsymbol{x}$ の要素 x_i は以下のように生成される．

$$\boldsymbol{\xi}^m = (\xi_1^m, \cdots, \xi_N^m)^T \tag{3}$$

$$\boldsymbol{x}_j^m = (x_{j1}^m, \cdots, x_{jN}^m)^T (j = 1, \cdots K) \tag{4}$$

$$P(\xi_i^m = \pm 1) = \frac{1}{2} \tag{5}$$

$$P(x_{ji}^m = \pm \frac{1}{\sqrt{N}}) = \frac{1 \pm a\xi_i^m}{2} \tag{6}$$

ここで m は時刻ステップであり，a は$\boldsymbol{\xi}$と $\boldsymbol{x}_j$ の方向余弦である．式 (3)〜(6) は，時刻 m でまず$\boldsymbol{\xi}$がランダムに生成され，$\boldsymbol{\xi}$との方向余弦が a である入力パターン $\boldsymbol{x}_j$ を K 個生成することを意味している．すなわち，中心$\boldsymbol{\xi}$の周りに入力$\boldsymbol{x_j}$を K 個用意し，それをひとつのセットとして学習に用いる．なお，各中心$\boldsymbol{\xi}$は互いに相関が無いものとする．

本稿では，$N \to \infty$ の熱力学的極限を考えることにする．このとき，

$$\|\boldsymbol{B}\| = \sqrt{N}, \qquad \|\boldsymbol{J}^0\| = \sqrt{N} \tag{7}$$

である．生徒の大きさ $\|\boldsymbol{J}\|$ は一般には時間の経過とともに変化するが，初期値 $\sqrt{N}$ に対する比を l とし，これを生徒の長さと呼ぶことにする．すなわち，$\|\boldsymbol{J}\| = l\sqrt{N}$ である．

教師の出力 v，生徒の出力 ul は，それぞれ以下の通りであり，このとき，v, u は平均 0，分散 1 のガウス分布にしたがう確率変数となる．

$$v = \boldsymbol{B}\cdot\boldsymbol{x}, \tag{8}$$

$$ul = \boldsymbol{J}\cdot\boldsymbol{x} \tag{9}$$

本稿で扱うモデルにおいては，生徒 $\boldsymbol{J}$ は入力 $\boldsymbol{x}$ とそれに対する教師 $\boldsymbol{B}$ の出力を用いて結合荷重の更新を行う．ただし，教師の出力 v，生徒の出力 ul にはそれぞれ分散

*神戸高専 専攻科
†富士通
‡神戸高専 電子工学科教員 miyoshi@kobe-kosen.ac.jp

σ_B^2, σ_J^2 の互いに独立なガウス雑音 n_B, n_J が重畳されるものとする．すなわち，

$$n_B \sim \mathcal{N}(0, \sigma_B^2), \quad n_J \sim \mathcal{N}(0, \sigma_J^2) \tag{10}$$

ここで，$\mathcal{N}(0, \sigma^2)$ は平均 0，分散 σ^2 のガウス分布を表す．

いま，教師と生徒の誤差 ϵ を両者の二乗誤差で定義する．すなわち

$$\epsilon \equiv \frac{1}{2}\left(v^m + n_B^m - u^m l^m - n_J^m\right)^2 \tag{11}$$

また，生徒は学習に勾配法を用いるものとする．このとき，K 個の例題を同時に用いるので

$$\boldsymbol{J}^{m+1} = \boldsymbol{J}^m + \sum_{j=1}^{K} \eta \frac{\partial \epsilon^m}{\partial \boldsymbol{J}^m} \tag{12}$$

$$= \boldsymbol{J}^m + \sum_{j=1}^{K} f_j^m \boldsymbol{x}_j^m \tag{13}$$

$$f_j^m = \eta\left(v_j^m + n_{Bj}^m - (u^m l^m)_j - n_{Jj}^m\right) \tag{14}$$

ここで，η は生徒の学習係数であり，定数とする．

3 理論

3.1 汎化誤差

統計的学習理論の目的のひとつは汎化誤差 ϵ_g を理論的に求めることである [3, 4, 5, 6]．汎化誤差は未知の入力，雑音に関する誤差の平均である．よって生徒の汎化誤差 ϵ_g は以下のように計算できる．

$$\epsilon_g = \int d\boldsymbol{x} dn_B dn_J P(\boldsymbol{x}, n_B, n_J)\epsilon \tag{15}$$

$$= \int du dv dn_B dn_J P(u, v, n_B, n_J) \times \frac{1}{2}(v + n_B - ul - n_J)^2 \tag{16}$$

$$= \frac{1}{2}\langle v^2 + u^2 l^2 - 2vul + n_B^2 + n_J^2 + 2n_B(v - ul) - 2n_J(v - ul)\rangle \tag{17}$$

$$= \frac{1}{2}(1 + l^2 - 2Rl + \sigma_B^2 + \sigma_J^2) \tag{18}$$

ここで，R は $\boldsymbol{B}$ と $\boldsymbol{J}$ の方向余弦である．

3.2 巨視的変数の微分方程式とその解

l と R のダイナミクスを記述する連立微分方程式は $N \to \infty$ の熱力学的極限における自己平均性に基づき以下のように決定論的な形で導出できる [2, 3, 4, 5, 6].

$$\frac{dr}{dt} = K\langle f_i v_i\rangle \tag{19}$$

$$2l\frac{dl}{dt} = 2Kl\langle f_i u_i\rangle + K\langle f_i^2\rangle + K(K-1)\langle f_i f_j\rangle\langle x_i x_j\rangle \tag{20}$$

ここで，r は解析を容易にするために導入した補助的な巨視的変数で，$r = Rl$ である．

この連立微分方程式に現れるサンプル平均は v, u が平均 0，分散 1，共分散 R の二重ガウス分布に従う確率変数であることを用いて以下のように計算できる．

$$\langle f_i v_i\rangle = \eta(1 - r) \tag{21}$$

$$\langle f_i u_i\rangle = \eta(r - l^2) \tag{22}$$

$$\langle f_i^2\rangle = \eta^2(1 + \sigma_J^2 + \sigma_B^2 + l^2 - 2r) \tag{23}$$

$$\langle f_i f_j\rangle = \eta^2 a^2(1 - 2r + l^2) \tag{24}$$

$$\langle x_i x_j\rangle = a^2 \tag{25}$$

本稿では教師 $\boldsymbol{B}$，生徒 $\boldsymbol{J}$ の初期値 $\boldsymbol{J}^0$ の各要素は平均 0，分散 1 のガウス分布にしたがい独立に生成され，また，$N \to \infty$ の熱力学的極限を考えているので，初期状態においてこれらはすべて直交しており，$R^0 = 0$ である．また，$l^0 = 1$ である．これらを用いて連立微分方程式 (19)〜(25) は以下のように解析的に解くことができる．

$$r = 1 - e^{-K\eta t} \tag{26}$$

$$l = \sqrt{-2e^{-K\eta t} + \frac{\eta}{A}\sigma^2 + 1 + \left(2 + \frac{\eta}{A}\sigma^2\right)e^{K\eta A t}} \tag{27}$$

ここで，

$$A = \eta(a^4(K-1) + 1) - 2 \tag{28}$$

$$\sigma^2 = \sigma_B^2 + \sigma_J^2 \tag{29}$$

である．

4 結果と議論

式 (18) を用いて理論的に計算される汎化誤差 ϵ_g のダイナミクスを計算機シミュレーションの結果と重ねて図 1，図 2 に示す．計算機実験は $N = 1000$ で実行し，その際の汎化誤差は各時点で 10000 個のランダム入力に対する教師と生徒の二乗誤差の平均を計算することにより求めた．また，このときの R, l を図 3 から図 6 に示す．

これらの図において曲線は理論値を，$+, \times$ 等の印は計算機実験の値を表す．また，σ_B^2, σ_J^2 以外の条件は共通で，$K = 3, \eta = 1.0$ である．図 1,3,5 は $\sigma_B^2 = 0, \sigma_J^2 = 0$ の場合の結果であり，図 2,4,6 は $\sigma_B^2 = 0.1, \sigma_J^2 = 0.2$ の場合の

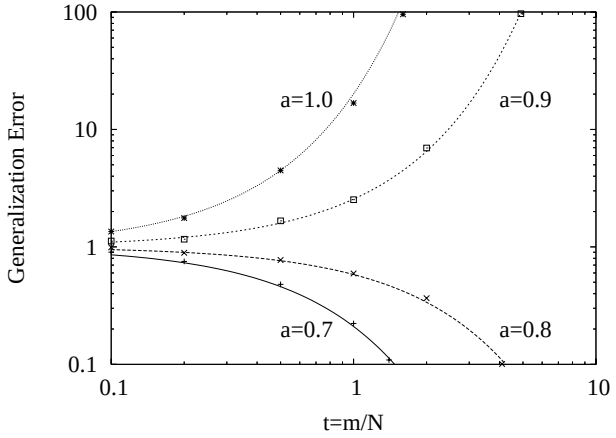

図 1: $\eta = 1.0$ の場合の汎化誤差 ϵ_g. 理論と計算機実験. η 以外の条件は $K = 3, \sigma_B^2 = 0, \sigma_J^2 = 0$

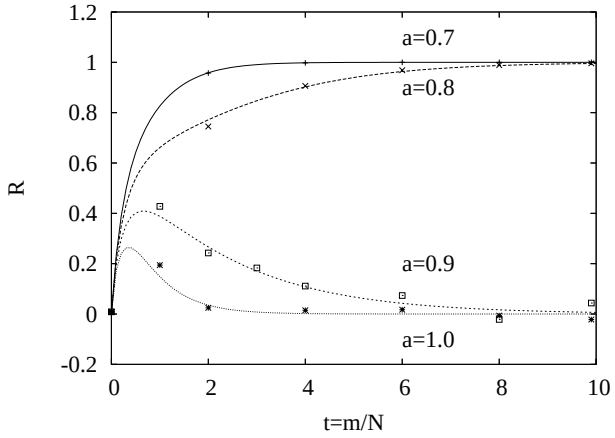

図 3: $\eta = 1.0$ の場合の R. 理論と計算機実験. η 以外の条件は $K = 3, \sigma_B^2 = 0, \sigma_J^2 = 0$

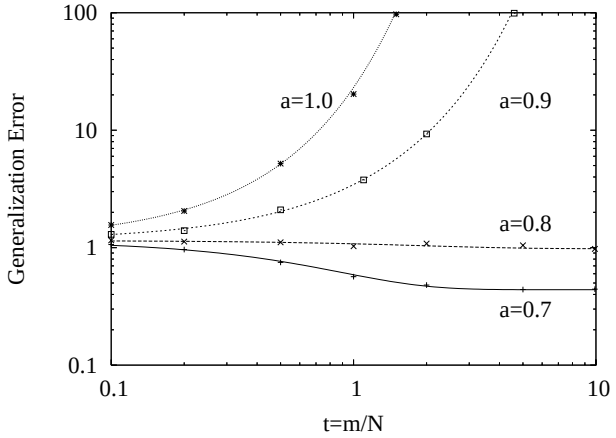

図 2: $\eta = 1.0$ の場合の汎化誤差 ϵ_g. 理論と計算機実験. η 以外の条件は $K = 3, \sigma_B^2 = 0.1, \sigma_J^2 = 0.2$

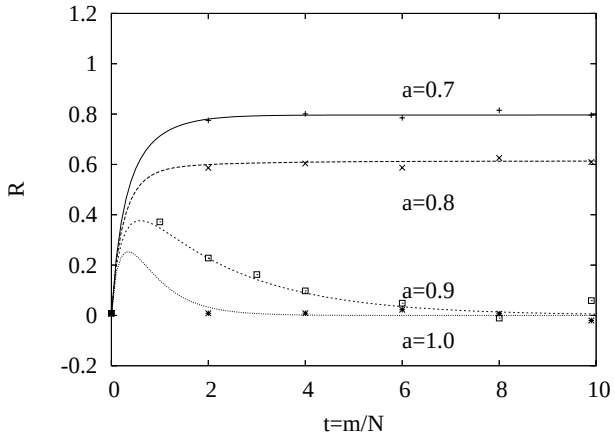

図 4: $\eta = 1.0$ の場合の R. 理論と計算機実験. η 以外の条件は $K = 3, \sigma_B^2 = 0.1, \sigma_J^2 = 0.2$

結果である．図 1–6 において理論と計算機実験の値はよく一致しており，理論が正しく導出されたといえる．

図 1 と図 3 から，$a = 0.8$ と $a = 0.9$ では学習のようすが質的に違うことがわかる．すなわち，$a = 0.8$ の場合には汎化誤差はゼロに漸近し，方向余弦 R は 1 に漸近するように見える．一方，$a = 0.9$ の場合には汎化誤差は発散し，R は 0 に漸近している．今回，汎化誤差や巨視的変数 R, l が時間の関数として解析的に得られているので，$t \to \infty$ におけるこれらのふるまいについても理論的な洞察が可能である．$K > 0, \eta > 0$ であるから，式 (26) より

$$r \to 1 \tag{30}$$

である．式 (27) の指数関数のべきの符号に着目することにより

$$K\eta A > 0 \tag{31}$$

の場合には生徒の長さ l は発散する．このことと式 (30) より $R \to 0$ であることがわかる．またこのとき式 (18) より汎化誤差は発散する．

一方，

$$K\eta A < 0 \tag{32}$$

の場合には

$$l \to \sqrt{\frac{\eta}{A}\sigma^2 + 1 + (2 + \frac{\eta}{A}\sigma^2)e^{K\eta A t}} \tag{33}$$

である．特にノイズが無い場合（$\sigma^2 = 0$）については $l \to 1$ となる．このことと式 (18),(30) より，このとき汎化誤差 $\epsilon_g \to 0$，方向余弦 $R \to 1$ であることがわかる．

式 (28),(32) より，汎化誤差が収束するために学習係数 η が満たすべき条件は

$$0 < \eta < \frac{2}{a^4(K-1)+1} \tag{34}$$

であることがわかる．

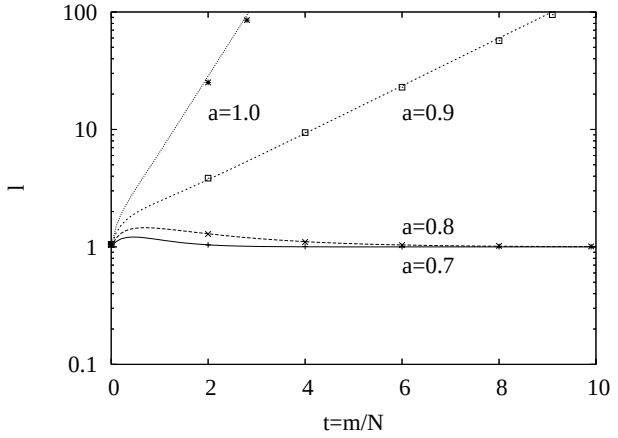

図 5: $\eta = 1.0$ の場合の汎化誤差 l. 理論と計算機実験．η 以外の条件は $K = 3, \sigma_B^2 = 0, \sigma_J^2 = 0$

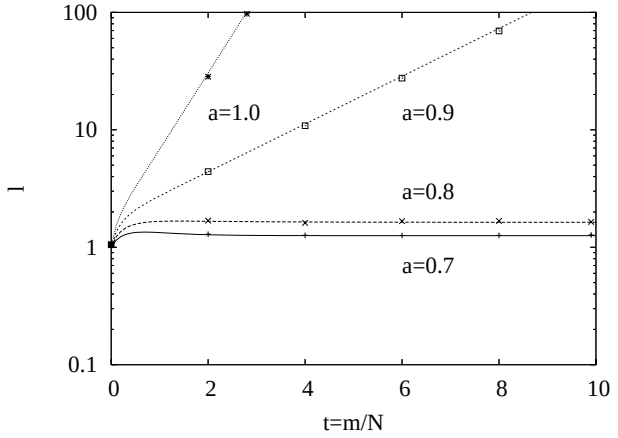

図 6: $\eta = 1.0$ の場合の l. 理論と計算機実験．η 以外の条件は $K = 3, \sigma_B^2 = 0.1, \sigma_J^2 = 0.2$

5　今後の予定

式 (34) や図 1，図 2 から，同時に使う入力間の相関 a が大きいほど，また，同時に使う入力の数 K が多いほど，学習係数 η が満たすべき条件は厳しくなる．また，学習が収束する場合でもその速度が遅くなる．今後は相関 a の影響を受けない学習則の提案とその解析をおこなう予定である．

参考文献

[1] Saad,D. (ed.), “On-Line Learning in Neural Networks”, Cambridge University Press, 1998.

[2] 西森秀稔,“ スピングラス理論と情報統計力学 ”, 岩波書店, 1999.

[3] Miyoshi,S., Hara,K. and Okada,M., “Analysis of ensemble learning using simple perceptrons based on online learning theory”, Physical Review E, 71, 036116. March 2005.

[4] Miyoshi,S. and Okada,M., “Analysis of on-line learning when a moving teacher goes around a true teacher”, J. Phys. Soc. Jpn., Vol.75, No.2, 024003, Feb. 2006.

[5] Miyoshi,S. and Okada,M., “Statistical mechanics of online learning for ensemble teachers”, J. Phys. Soc. Jpn., Vol.75, No.4, 044002, Apr. 2006.

[6] Miyoshi,S., Uezu,T. and Okada,M., “Statistical mechanics of time-domain ensemble learning”, J. Phys. Soc. Jpn.,, Vol.75, No.8, 084007, Aug. 2006.

【情報通信に役立つ】