教師が真の教師のまわりをまわる場合のオンライン学習

Analysis of on-line learning when a moving teacher goes around a true teacher

三好 誠司 (P)†, 岡田 真人 ‡
Seiji Miyoshi(P) and Masato Okada
† 神戸高専電子工学科 miyoshi@kobe-kosen.ac.jp, ‡ 東大新領域, 理研脳総研, 科技機構さきがけ

Abstract— A learning machine might move around a teacher due to the differences in structures or output functions between the teacher and the learning machine or due to noises. We analyzed the generalization performance of a new student supervised by the moving machine. It is proven that the generalization error of the student can be smaller than that of the moving teacher.

Keywords— **on-line learning, generalization error, moving teacher, true teacher, unlearnable case**

1 まえがき

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．これまでに我々はオンライン学習の枠組み [1] で，特にアンサンブル学習 [2, 3, 4] の汎化能力について統計力学的手法を用いた解析を行ってきた [5, 6, 7]．その過程で，学習不能な場合 [8, 9] には学習機械が真の教師のまわりを動き続ける場合があることが明らかになった [10, 11]．現実の問題には学習不能な場合も多いと考えられることから，統計的学習理論の応用を考えた場合，学習不能な場合の系のふるまいを調べることはきわめて重要である．

さて，ここでこの動き続ける学習機械を教師とするような新たな生徒を考えることにする．この場合，生徒が学習に用いる例題は動き続ける教師の入出力だけであり，生徒は直接には真の教師の入出力を観測できないことに注意しておく．この生徒が真の教師に対してどれほどの汎化能力を持つことができるかを考えることにする．

今回，教師が真の教師のまわりを動き続けるもっとも単純なモデルとして，真の教師，動く教師，生徒のいずれもが雑音の重畳された線形なパーセプトロンである場合を考え，オンライン学習の枠組みで統計力学的手法を用いることによりいくつかの巨視的変数や汎化誤差を解析的に求める [12].

2 モデル

本論文では真の教師，動く教師，生徒の三個の線形パーセプトロンを考える．真の教師 $\boldsymbol{A} = (A_1, \ldots, A_N)$，動く教師 $\boldsymbol{B} = (B_1, \ldots, B_N)$，生徒 $\boldsymbol{J} = (J_1, \ldots, J_N)$ および入力 $\boldsymbol{x} = (x_1, \ldots, x_N)$ は N 次元ベクトルであり，真の教師 $\boldsymbol{A}$ の各要素 A_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．$\boldsymbol{B}, \boldsymbol{J}$ の初期値 $\boldsymbol{B}^0, \boldsymbol{J}^0$ の各要素 B_i^0, J_i^0 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るものとする．また，$\boldsymbol{x}$ の各要素 x_i は平均 0，分散 $1/N$ のガウス分布にしたがい独立に生成されるものとする．本論文では，$N \to \infty$ の熱力学的極限を考えることにする．このとき，$\|\boldsymbol{A}\| = \sqrt{N}$, $\|\boldsymbol{B}^0\| = \sqrt{N}$, $\|\boldsymbol{J}^0\| = \sqrt{N}$, $\|\boldsymbol{x}\| = 1$ となる．動く教師の大きさ $\|\boldsymbol{B}\|$，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_B, l_J とし，これらをそれぞれ動く教師の長さ，生徒の長さと呼ぶことにする．

真の教師の出力は $y + n_A = \boldsymbol{A} \cdot \boldsymbol{x} + n_A$，動く教師の出力は $vl_B + n_B = \boldsymbol{B} \cdot \boldsymbol{x} + n_B$，生徒の出力は $ul_J + n_J = \boldsymbol{J} \cdot \boldsymbol{x} + n_J$ である．ここで，n_A, n_B, n_J はそれぞれ分散 $\sigma_A^2, \sigma_B^2, \sigma_J^2$ の独立なガウス雑音であり，y, v, u は平均 0，分散 1 のガウス分布にしたがう確率変数である．

動く教師 $\boldsymbol{B}$ は入力 $\boldsymbol{x}$ とそれに対する真の教師 $\boldsymbol{A}$ の出力を用いて結合荷重の更新を行う．また，生徒 $\boldsymbol{J}$ は入力 $\boldsymbol{x}$ とそれに対する動く教師 $\boldsymbol{B}$ の出力を用いて結合荷重の更新を行う．いま，真の教師と動く教師の誤差 ϵ_B を両者の出力の二乗誤差で定義し，教師は学習に勾配法を用いるものとする．すなわち，

$$\boldsymbol{B}^{m+1} = \boldsymbol{B}^m + \eta_B \left(y^m + n_A^m - v^m l_B^m - n_B^m\right) \boldsymbol{x}^m, \quad (1)$$

ここで，η_B は動く教師の学習係数であり定数とする．

同様に，動く教師と生徒の誤差 ϵ_{BJ} を両者の出力の二乗誤差で定義し，生徒も学習に勾配法を用いるものとする．すなわち，

$$\boldsymbol{J}^{m+1} = \boldsymbol{J}^m + \eta_J \left(v^m l_B^m + n_B^m - u^m l_J^m - n_J^m\right) \boldsymbol{x}^m, \quad (2)$$

ここで，η_J は生徒の学習係数であり定数とする．

また，真の教師と生徒の誤差 ϵ_J も両者の二乗誤差で定義しておく．

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を理論的に求めることである．汎化誤差は真の教師に対する誤差を未知の入力に関して平均したものであるから，動く教師の汎化誤差 ϵ_{Bg}，生徒の汎化誤差 ϵ_{Jg}，および，動く教師と生徒の誤差の平均 ϵ_{BJ_g} はそれぞれ以下のように計算される [12].

$$\epsilon_{Bg} = \frac{1}{2}\left(-2R_B l_B + (l_B)^2 + 1 + \sigma_A^2 + \sigma_B^2\right), \quad (3)$$

$$\epsilon_{Jg} = \frac{1}{2}\left(-2R_J l_J + (l_J)^2 + 1 + \sigma_A^2 + \sigma_J^2\right), \quad (4)$$

$$\epsilon_{BJ_g} = \frac{1}{2}\left(-2R_{BJ} l_B l_J + (l_J)^2 + (l_B)^2 + \sigma_B^2 + \sigma_J^2\right) \quad (5)$$

ここで $R_B \equiv \boldsymbol{A} \cdot \boldsymbol{B}/\|\boldsymbol{A}\|\|\boldsymbol{B}\|$, $R_J \equiv \boldsymbol{A} \cdot \boldsymbol{J}/\|\boldsymbol{A}\|\|\boldsymbol{J}\|$, $R_{BJ} \equiv \boldsymbol{B} \cdot \boldsymbol{J}/\|\boldsymbol{B}\|\|\boldsymbol{J}\|$ である．

3.2 巨視的変数の微分方程式とその解

巨視的変数のダイナミクスを記述する連立微分方程式を熱力学的極限における自己平均性 [5, 13, 14] に基づき以下のような決定論的な形で導出した [12].

$$\frac{dr_B}{dt} = \langle gy \rangle, \quad \frac{dr_J}{dt} = \langle fy \rangle, \tag{6}$$

$$\frac{dr_{BJ}}{dt} = l_J \langle gu \rangle + l_B \langle fv \rangle + \langle gf \rangle, \tag{7}$$

$$\frac{dl_B}{dt} = \langle gv \rangle + \frac{\langle g^2 \rangle}{2l_B}, \quad \frac{dl_J}{dt} = \langle fu \rangle + \frac{\langle f^2 \rangle}{2l_J}. \tag{8}$$

ここで r_B, r_J, r_{BJ} は解析を容易にするため導入した補助的巨視的変数であり，$r_B \equiv R_B l_B$, $r_J \equiv R_J l_J$, $r_{BJ} \equiv R_{BJ} l_B l_J$ である．また，$\langle \cdot \rangle$ はサンプル平均を表す．

本論文では線形なパーセプトロンを考えているので，これらの連立微分方程式に現れるサンプル平均は容易に計算することができる [12]．それらを用いて連立微分方程式 (6)–(8) を解析的に解いた [12].

4 結果と議論

理論的に求められた汎化誤差 $\epsilon_{Bg}, \epsilon_{Jg}$ と ϵ_{BJg} のダイナミクスを計算機シミュレーション（$N = 10^3$）の結果と重ねて図 1 に示す．図中，B は動く教師の汎化誤差 ϵ_{Bg} を，J は生徒の汎化誤差 ϵ_{Jg} を，$B - J$ は動く教師と生徒の誤差の平均 ϵ_{BJg} をそれぞれ表す．

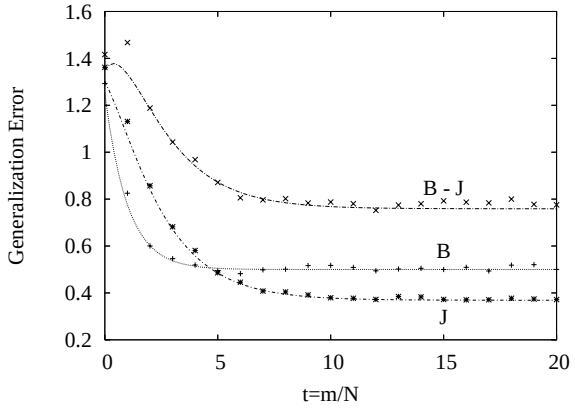

図 1: $\epsilon_{Jg}, \epsilon_{Bg}, \epsilon_{BJg}$ のダイナミクス．曲線は理論，ドットは計算シミュレーション．$\eta_J = 0.3, \eta_B = 1.0, \sigma_A^2 = 0.2, \sigma_B^2 = 0.3, \sigma_J^2 = 0.4$.

図 1 より，学習の初期においては生徒の汎化誤差 ϵ_{Jg} は動く教師の汎化誤差 ϵ_{Bg} よりも大きいが $t = 4.4$ でその大小関係が逆転し，それ以後は ϵ_{Jg} が ϵ_{Bg} よりも小さくなっている．すなわち，動く教師よりも生徒の方が性能が高くなっている．

図 1 を見ると，$\epsilon_{Bg}, \epsilon_{Jg}, \epsilon_{BJg}$ は $t = 20$ でほぼ定常値に達しているように見える．今回巨視的変数が解析的に得られているのでこれらの $t \to \infty$ におけるふるまいについては理論的な洞察が可能である．すなわち，$0 < \eta_B < 2$ でなければ ϵ_{Bg},ϵ_g は発散し，$0 < \eta_J < 2$ でなければ ϵ_g,ϵ_{Jg} は発散する [12]．$0 < \eta_B, \eta_J < 2$ の場合については，汎化誤差は収束する [12]．このときの生徒の学習係数 η_J と $\epsilon_{Bg}, \epsilon_{Jg}, \epsilon_{BJg}$ の定常値の関係を図 2 に示す．生徒の学習係数 η_J が 0.58 よりも大きいときには生徒の定常汎化誤差は動く教師の定常汎化誤差よりも大きいが，η_J が 0.58 より小さくなるとその大小関係は逆転する．すなわち，η_J が 0.58 より小さい場合には動く教師よりも生徒の方が高性能になる．

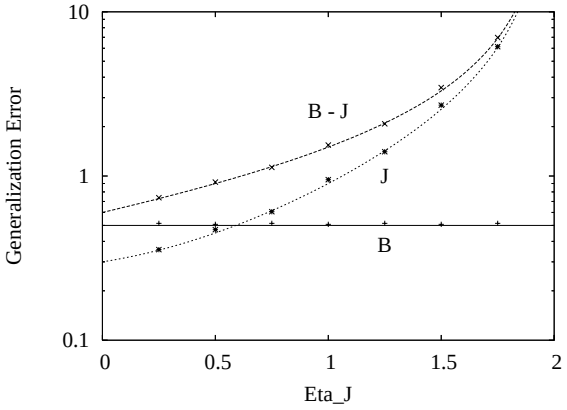

図 2: $\epsilon_{Jg}, \epsilon_{Bg}, \epsilon_{BJg}$ の定常値．曲線は理論，ドットは計算機シミュレーション．η_J 以外の条件は $\eta_B = 1.0, \sigma_A^2 = 0.2, \sigma_B^2 = 0.3, \sigma_J^2 = 0.4$

5 むすび

真の教師，動く教師，生徒のいずれもが雑音の重畳された線形なパーセプトロンである場合を考え，統計力学的手法により汎化誤差を解析的に求めた結果，生徒が真の教師の入出力ではなく，動く教師の入出力だけを例題として使用するにもかかわらず，動く教師の汎化誤差よりも生徒の汎化誤差の方が小さくなりうるという興味深い結果が明らかになった．

謝辞

本論文の一部は科学研究費補助金（課題番号 14084212, 14580438, 15500151, 16500093 ）によるものである．

参考文献

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

[2] 麻生, 津田, 村田, “パターン認識と学習の統計学,” 岩波, 東京,2003.

[3] Krogh, A. & Sollich, P., Phys. Rev. E, **55**(1), 811 (1997).

[4] Urbanczik, R., Phys. Rev. E, **62**(1), 1448 (2000).

[5] 岡田，原，三好，信学技報，NC2003-35, pp.7-12 (2003)

[6] Miyoshi,S., Hara,K. and Okada,M., Phys. Rev. E, 71, 036116 (2005)

[7] 三好，原，岡田，信学論,**J87-D-II**(7),1391 (2004)

[8] Inoue, J. and Nishimori, H., Phys. Rev. E, **55**(4), 4544 (1997)

[9] Inoue, J., Nishimori, H. and Kabashima, Y., cond-mat/9708096 (1997).

[10] 三好，原，岡田，IBIS2004 予稿集，178 (2004)

[11] 三好，原，岡田，信学技報，NC2004-214, 123 (2005)

[12] 岡田，三好，信学技報，NC2005-10, pp.19-24 (2005)

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

[14] Nishimori, H., “Statistical Physics of Spin Glasses and Information Processing: An Introduction,” Oxford University Press, (2001)