社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

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

三好　誠司†　　岡田　真人††,†††,††††

† 神戸市立工業高等専門学校　〒 651–2194 神戸市西区学園東町 8–3
†† 東京大学大学院 新領域創成科学研究科 複雑理工学専攻　〒 277–8561 千葉県柏市柏の葉 5–1–5
††† 理化学研究所 脳科学総合研究センター　〒 351–0198 埼玉県和光市広沢 2–1
†††† 科学技術振興機構 さきがけ
E-mail: †miyoshi@kobe-kosen.ac.jp, ††okada@k.u-tokyo.ac.jp

あらまし　オンライン学習において，教師と生徒の構造や出力特性の相違，雑音の影響などにより汎化誤差がゼロにならないモデルでは，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解析した．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モデルについて，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，真の教師と動く教師の汎化誤差よりも真の教師と生徒の汎化誤差の方が小さくなりうることが明らかになった．

キーワード　オンライン学習，汎化誤差，動く教師，真の教師，学習不能な場合

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

Seiji MIYOSHI† and Masato OKADA††,†††,††††

† Kobe City College of Technology　8–3 Gakuenhigashimachi, Nishi-ku, Kobe-shi, 651–2194 Japan
†† Division of Transdisciplinary Sciences, Graduate School of Frontier Sciences, The University of Tokyo
5–1–5 Kashiwanoha, Kashiwa-shi, Chiba, 277–8561 Japan
††† RIKEN Brain Science Institute　2–1 Hirosawa, Wako-shi, Saitama, 351–0198 Japan
†††† JST PRESTO
E-mail: †miyoshi@kobe-kosen.ac.jp, ††okada@k.u-tokyo.ac.jp

Abstract　In the framework of on-line learning, the learning machine might move around the teacher due to the difference of structures or output functions between the teacher and the learning machine or the noises. We analyzed the generalization performance of a new student supervised by the moving machine. Calculating generalization errors of a model composed by the true teacher, the moving teacher and the student that are all linear perceptrons with noises analytically by using statistical mechanics, it is proved that the generalization error between the true teacher and the student can be smaller than that between the true teacher and the moving teacher, though the student uses examples from only the moving teacher.

Key words　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．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] 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．

これまでに我々はオンライン学習の枠組みで，特にアンサンブル学習 [2]〜[6] 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 [7]〜[10]．その過程で副次的に以下のことが明らかになった．生徒が単純パーセプトロンであるのに対し，教師がコミティマシンであったり，あるいは教師の出力特性が非単調であるような場合には汎化誤差がゼロにならない [15]〜[19]．よってこのようなモデルを学習不能な場合 [13], [14] と呼ぶことができるが，このときの生徒のふるまいは学習則によって異なる．すなわち，ヘブ学習を用いる場合は生徒は一方向に収束する．これに対し，パーセプトロン学習やアダトロン学習を用いる場合には生徒は一方向に収束せず，動き続ける．教師の出力特性が非単調な場合に限定して表現すると，生徒は教師と一定の方向余弦を保ったまま教師のまわりを動き続ける．

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．また，上で述べたように，学習不能な場合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．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．すなわち，この動き続ける教師の入出力を例題として学習を行う生徒を考え（ここで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する），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．現実の人間社会においても，生徒が入出力を観測できる教師は必ずしも正しい解答を示すとは限らず，また，教師自身も学習しており，変わり続ける存在である場合が多いことから，このようなモデルの解析は統計的学習理論と現実社会とのアナロジーを考える上でも興味深い．

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 [7] 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た．その結果，上に述べたように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，真の教師と動く教師の汎化誤差よりも真の教師と生徒の汎化誤差の方が小さくなりうること，すなわち，動く教師よりも生徒の方が賢くなりうることが明らかになった．

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，それぞれの結合荷重を $\hat{\boldsymbol{B}}, \boldsymbol{B}, \boldsymbol{J}$ とする．なお，本論文では簡単のため真の教師の結合荷重，動く教師の結合荷重，生徒の結合荷重のことをそれぞれ単に真の教師，動く教師，生徒と呼ぶことにする．真の教師 $\hat{\boldsymbol{B}} = (\hat{B}_1, \ldots, \hat{B}_N)$，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 $\hat{\boldsymbol{B}}$ の各要素 $\hat{B}_i$ 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．$\boldsymbol{B}, \boldsymbol{J}$ 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．すなわち，

$$\left\langle \hat{B}_i \right\rangle = 0,\ \left\langle (\hat{B}_i)^2 \right\rangle = 1, \tag{1}$$

$$\left\langle B_i^0 \right\rangle = 0,\ \left\langle (B_i^0)^2 \right\rangle = 1, \tag{2}$$

$$\left\langle J_i^0 \right\rangle = 0,\ \left\langle (J_i^0)^2 \right\rangle = 1, \tag{3}$$

$$\langle x_i \rangle = 0,\ \left\langle (x_i)^2 \right\rangle = \frac{1}{N}. \tag{4}$$

ここで，$\langle \cdot \rangle$ は平均を表す．

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，

$$|\hat{\boldsymbol{B}}| = \sqrt{N},\ \ |\boldsymbol{B}^0| = \sqrt{N},\ \ |\boldsymbol{J}^0| = \sqrt{N},\ \ |\boldsymbol{x}| = 1. \tag{5}$$

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．すなわち，$|\boldsymbol{B}| = l_B\sqrt{N}$，$|\boldsymbol{J}| = l_J\sqrt{N}$ である．

真の教師の出力 $\hat{v}$，動く教師の出力 vl_B，生徒の出力 ul_J はそれぞれ以下の通りであり，このとき，$\hat{v}$，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．

$$\hat{v} = \hat{\boldsymbol{B}} \cdot \boldsymbol{x}, \tag{6}$$

$$vl_B = \boldsymbol{B} \cdot \boldsymbol{x}, \tag{7}$$

$$ul_J = \boldsymbol{J} \cdot \boldsymbol{x}. \tag{8}$$

本研究で扱うモデルにおいては，動く教師 $\boldsymbol{B}$ は入力 $\boldsymbol{x}$ とそれに対する真の教師 $\hat{\boldsymbol{B}}$ 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．真の教師の出力 $\hat{v}$，動く教師の出力 vl_B，生徒の出力 ul_J にはそれぞれ分散 $\sigma_{\hat{B}}^2, \sigma_B^2, \sigma_J^2$ の互いに独立なガウス雑音が重畳されるものとする．すなわち，本研究で扱うモデルの学習は以下のように表せる．

$$\boldsymbol{B}^{m+1} = \boldsymbol{B}^m + g\left(\hat{v}^m + n_{\hat{B}}^m, v^m l_B^m + n_B^m\right)\boldsymbol{x}^m, \tag{9}$$

$$\boldsymbol{J}^{m+1} = \boldsymbol{J}^m + f\left(v^m l_B^m + n_B^m, u^m l_J^m + n_J^m\right)\boldsymbol{x}^m, \tag{10}$$

$$n_{\hat{B}}^m \sim \mathcal{N}(0, \sigma_{\hat{B}}^2), \tag{11}$$

$$n_B^m \sim \mathcal{N}(0, \sigma_B^2), \tag{12}$$

$$n_J^m \sim \mathcal{N}(0, \sigma_J^2). \tag{13}$$

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

いま，真の教師と動く教師の誤差 $\hat{\epsilon}$ を両者の出力の二乗誤差で定義する．すなわち，

$$\hat{\epsilon}^m \equiv \frac{1}{2}\left(\hat{v}^m + n_{\hat{B}}^m - v^m l_B^m - n_B^m\right)^2. \tag{14}$$

また，教師は学習に勾配法を用いるものとする．すなわち，

$$\boldsymbol{B}^{m+1} = \boldsymbol{B}^m - \eta_B \frac{\partial \hat{\epsilon}^m}{\partial \boldsymbol{B}^m} \tag{15}$$

$$= \boldsymbol{B}^m + \eta_B\left(\hat{v}^m + n_{\hat{B}}^m - v^m l_B^m - n_B^m\right)\boldsymbol{x}^m. \tag{16}$$

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 を両者の出力の二乗誤差で

定義する．すなわち，

$$\epsilon^m \equiv \frac{1}{2}\left(v^m l_B^m + n_B^m - u^m l_J^m - n_J^m\right)^2. \tag{17}$$

生徒も学習に勾配法を用いるものとする．すなわち，

$$\boldsymbol{J}^{m+1} = \boldsymbol{J}^m - \eta_J \frac{\partial \epsilon^m}{\partial \boldsymbol{J}^m} \tag{18}$$

$$= \boldsymbol{J}^m + \eta_J \left(v^m l_B^m + n_B^m - u^m l_J^m - n_J^m\right) \boldsymbol{x}^m. \tag{19}$$

ここで，η_J は生徒の学習係数であり定数とする．

よって，式 (9)，(10) において

$$g = \eta_B \left(\hat{v}^m + n_{\hat{B}}^m - v^m l_B^m - n_B^m\right), \tag{20}$$

$$f = \eta_J \left(v^m l_B^m + n_B^m - u^m l_J^m - n_J^m\right). \tag{21}$$

また，真の教師と生徒の誤差 $\bar{\epsilon}$ も両者の二乗誤差で定義しておく．すなわち，

$$\bar{\epsilon}^m \equiv \frac{1}{2}\left(\hat{v}^m + n_{\hat{B}}^m - u^m l_J^m - n_J^m\right)^2. \tag{22}$$

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．汎化誤差は未知の入力に関する誤差の平均であるから，真の教師に対する動く教師の汎化誤差 $\hat{\epsilon}_g$，動く教師に対する生徒の汎化誤差 ϵ_g，真の教師に対する生徒の汎化誤差 $\bar{\epsilon}_g$ はそれぞれ以下のように計算される．なお，以後は時刻ステップを表す添字 m は簡単のために省略する．

$$\hat{\epsilon}_g = \int d\boldsymbol{x} dn_{\hat{B}} dn_B P(\boldsymbol{x}, n_{\hat{B}}, n_B) \hat{\epsilon} \tag{23}$$

$$= \int d\hat{v} dv dn_{\hat{B}} dn_B P(\hat{v}, v, n_{\hat{B}}, n_B) \times \frac{1}{2}\left(\hat{v} + n_{\hat{B}} - v l_B - n_B\right)^2 \tag{24}$$

$$= \frac{1}{2}\left(-2\hat{R} l_B + (l_B)^2 + 1 + \sigma_{\hat{B}}^2 + \sigma_B^2\right), \tag{25}$$

$$\epsilon_g = \int d\boldsymbol{x} dn_B dn_J P(\boldsymbol{x}, n_B, n_J) \epsilon \tag{26}$$

$$= \int dv du dn_B dn_J P(v, u, n_B, n_J) \times \frac{1}{2}\left(v l_B + n_B - u l_J - n_J\right)^2 \tag{27}$$

$$= \frac{1}{2}\left(-2 R l_B l_J + (l_J)^2 + (l_B)^2 + \sigma_B^2 + \sigma_J^2\right), \tag{28}$$

$$\bar{\epsilon}_g = \int d\boldsymbol{x} dn_{\hat{B}} dn_J P(\boldsymbol{x}, n_{\hat{B}}, n_J) \bar{\epsilon} \tag{29}$$

$$= \int d\hat{v} du dn_{\hat{B}} dn_J P(\hat{v}, u, n_{\hat{B}}, n_J) \times \frac{1}{2}\left(\hat{v} + n_{\hat{B}} - u l_J - n_J\right)^2 \tag{30}$$

$$= \frac{1}{2}\left(-2\bar{R} l_J + (l_J)^2 + 1 + \sigma_{\hat{B}}^2 + \sigma_J^2\right). \tag{31}$$

ここで積分の実行には，$\hat{v}$,v,u が平均 0，分散 1 のガウス分布に従うこと，真の教師と動く教師，真の教師と生徒，動く教師と生徒の方向余弦をそれぞれ

$$\hat{R} \equiv \frac{\hat{\boldsymbol{B}} \cdot \boldsymbol{B}}{|\hat{\boldsymbol{B}}||\boldsymbol{B}|}, \quad \bar{R} \equiv \frac{\hat{\boldsymbol{B}} \cdot \boldsymbol{J}}{|\hat{\boldsymbol{B}}||\boldsymbol{J}|}, \quad R \equiv \frac{\boldsymbol{B} \cdot \boldsymbol{J}}{|\boldsymbol{B}||\boldsymbol{J}|}. \tag{32}$$

とすると，$\hat{v}$ と v の共分散が $\hat{R}$，v と u の共分散が R，$\hat{v}$ と u の共分散が $\bar{R}$ であること，および，$n_{\hat{B}}$，n_B，n_J はいずれも他の確率変数とは独立であることを利用した．真の教師 $\hat{\boldsymbol{B}}$，動く教師 $\boldsymbol{B}$，生徒 $\boldsymbol{J}$ および $\hat{R}, \bar{R}, R$ の関係を図 1 に示す．

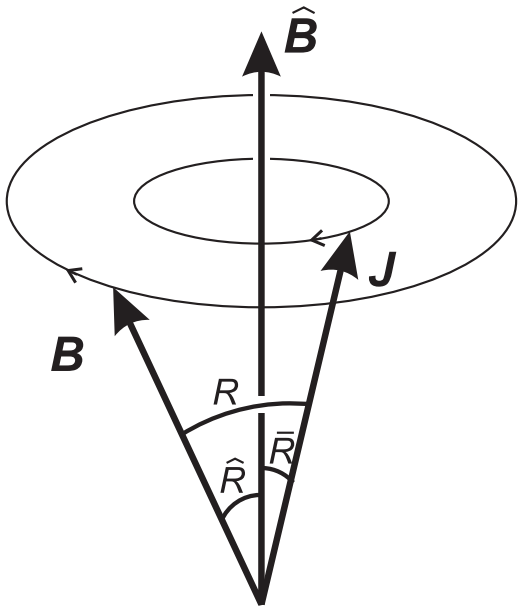

図 1　真の教師 $\hat{\boldsymbol{B}}$，動く教師 $\boldsymbol{B}$，生徒 $\boldsymbol{J}$．$\hat{R}, \bar{R}, R$ は方向余弦である．

3.2 巨視的変数の微分方程式とその解

解析を容易にするため，以下の補助的な巨視的変数を導入する．

$$\hat{r} \equiv \hat{R} l_B, \tag{33}$$

$$\bar{r} \equiv \bar{R} l_J, \tag{34}$$

$$r \equiv R l_B l_J, \tag{35}$$

$$L_B \equiv l_B^2, \tag{36}$$

$$L_J \equiv l_J^2. \tag{37}$$

今回，式 (33)–(37) のダイナミクスを記述する連立微分方程式 [11], [12] を熱力学的極限における自己平均性に基づき以下のような決定論的な形で導出した [8].

$$\frac{d\hat{r}}{dt} = \langle g\hat{v} \rangle, \tag{38}$$

$$\frac{d\bar{r}}{dt} = \langle f\hat{v} \rangle, \tag{39}$$

$$\frac{dr}{dt} = l_J \langle gu \rangle + l_B \langle fv \rangle + \langle gf \rangle, \tag{40}$$

$$\frac{dl_B}{dt} = \langle gv \rangle + \frac{\langle g^2 \rangle}{2l_B}, \tag{41}$$

$$\frac{dl_J}{dt} = \langle fu \rangle + \frac{\langle f^2 \rangle}{2l_J}. \tag{42}$$

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以下のように容易に計算することができる．

$$\langle gu \rangle = \eta_B(\bar{r} - r)/l_J, \tag{43}$$
$$\langle fv \rangle = \eta_J(l_B - r/l_B), \tag{44}$$
$$\langle gf \rangle = \eta_B\eta_J(\hat{r} - \bar{r} - l_B^2 + r - \sigma_B^2), \tag{45}$$
$$\langle f\hat{v} \rangle = \eta_J(\hat{r} - \bar{r}), \tag{46}$$
$$\langle g\hat{v} \rangle = \eta_B(1 - \hat{r}), \tag{47}$$
$$\langle gv \rangle = \eta_B(\hat{r}/l_B - l_B), \tag{48}$$
$$\langle g^2 \rangle = \eta_B^2(1 + \sigma_{\hat{B}}^2 + \sigma_B^2 + l_B^2 - 2\hat{r}), \tag{49}$$
$$\langle fu \rangle = \eta_J(r/l_J - l_J), \tag{50}$$
$$\langle f^2 \rangle = \eta_J^2(l_B^2 + l_J^2 + \sigma_B^2 + \sigma_J^2 - 2r). \tag{51}$$

本論文では真の教師 $\hat{\boldsymbol{B}}$，動く教師 $\boldsymbol{B}$ の初期値 $\boldsymbol{B}^0$，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，

$$\hat{R}^0 = \bar{R}^0 = R^0 = 0 \tag{52}$$

である．また，

$$l_B^0 = l_J^0 = 1 \tag{53}$$

である．式 (43)–(53) を用いて連立微分方程式 (38)–(42) は以下のように解析的に解ける．

$$\hat{r} = 1 - e^{-\eta_B t}, \tag{54}$$
$$\bar{r} = 1 + \frac{\eta_B}{\eta_J - \eta_B}e^{-\eta_J t} - \frac{\eta_J}{\eta_J - \eta_B}e^{-\eta_B t}, \tag{55}$$
$$\begin{aligned} r = & -\frac{C}{\eta_B\eta_J - \eta_B - \eta_J} \\ & + \frac{2\eta_J - \eta_B}{\eta_B - \eta_J}e^{-\eta_B t} + \frac{\eta_B}{\eta_J - \eta_B}e^{-\eta_J t} \\ & + \frac{\eta_J}{\eta_J - \eta_B}Ae^{\eta_B(\eta_B - 2)t} + De^{(\eta_B\eta_J - \eta_B - \eta_J)t}, \end{aligned} \tag{56}$$
$$L_B = 3 - A - 2e^{-\eta_B t} + Ae^{\eta_B(\eta_B - 2)t}, \tag{57}$$
$$\begin{aligned} L_J = & -\frac{F}{\eta_J(\eta_J - 2)} \\ & + \frac{E}{\eta_B(\eta_B - 2) - \eta_J(\eta_J - 2)}e^{\eta_B(\eta_B - 2)t} \\ & + \frac{2\eta_B}{\eta_J - \eta_B}e^{-\eta_J t} - \frac{2\eta_J}{\eta_J - \eta_B}e^{-\eta_B t} \\ & - \frac{2\eta_J D}{\eta_B - \eta_J}e^{(\eta_B\eta_J - \eta_B - \eta_J)t} + Ge^{\eta_J(\eta_J - 2)t}. \end{aligned} \tag{58}$$

ここで，

$$A = 2 - \frac{\eta_B}{2 - \eta_B}(\sigma_{\hat{B}}^2 + \sigma_B^2), \tag{59}$$
$$C = \eta_B(1 - \eta_J\sigma_B^2) + \eta_J(1 - \eta_B)(3 - A), \tag{60}$$
$$\begin{aligned} D = & \frac{-\eta_B^2\eta_J}{(\eta_J - \eta_B)(\eta_B\eta_J - \eta_B - \eta_J)}(\sigma_{\hat{B}}^2 + \sigma_B^2) \\ & - \frac{2\eta_B}{\eta_J - \eta_B} + \frac{\eta_B(1 - \eta_J\sigma_B^2)}{\eta_B\eta_J - \eta_B - \eta_J}, \end{aligned} \tag{61}$$
$$E = \eta_J^2\frac{\eta_B + \eta_J - 2}{\eta_B - \eta_J}A, \tag{62}$$
$$F = \eta_J^2\left(3 + \sigma_B^2 + \sigma_J^2 - A\right) - \frac{2\eta_J(1 - \eta_J)C}{\eta_B\eta_J - \eta_B - \eta_J}, \tag{63}$$
$$\begin{aligned} G = & 3 - \frac{E}{\eta_B(\eta_B - 2) - \eta_J(\eta_J - 2)} \\ & + \frac{F}{\eta_J(\eta_J - 2)} + \frac{2\eta_J}{\eta_B - \eta_J}D, \end{aligned} \tag{64}$$

である．

4. 結果と議論

式 (25)，(28)，(31) ，(33)–(37) ，(54)–(64) を用いて理論的に計算される三個の汎化誤差 $\hat{\epsilon}_g, \epsilon_g, \bar{\epsilon}_g$ 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 2，図 3 に示す．計算機シミュレーションは $N = 10^3$ で実行し，汎化誤差は各時点で 10^4 個のランダム入力に対する誤差の平均を計算することにより計算した．また，このときの $\hat{R}, R, \bar{R}, l_B, l_J$ を図 4，図 5 に示す．

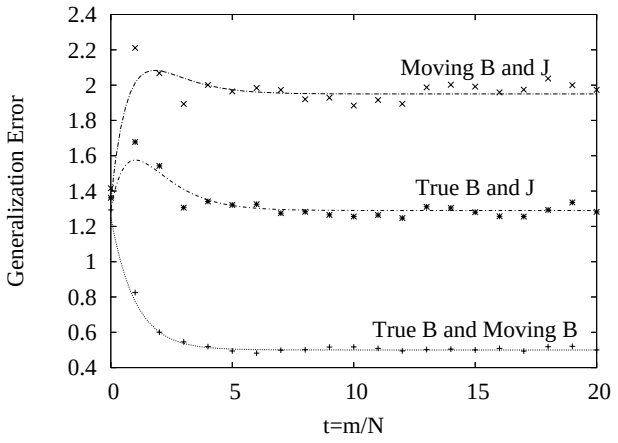

図 2　$\eta_J = 1.2$ の場合の汎化誤差 $\epsilon, \bar{\epsilon}_g, \hat{\epsilon}_g$．理論と計算機シミュレーション．$\eta_J$ 以外の条件は $\eta_B = 1.0, \sigma_{\hat{B}}^2 = 0.2, \sigma_B^2 = 0.3, \sigma_J^2 = 0.4$ である．

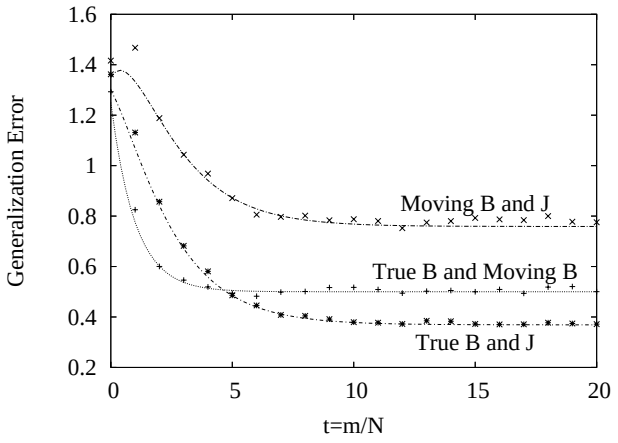

図 3　$\eta_J = 0.3$ の場合の汎化誤差 $\epsilon, \bar{\epsilon}_g, \hat{\epsilon}_g$．理論と計算機シミュレーション．$\eta_J$ 以外の条件は $\eta_B = 1.0, \sigma_{\hat{B}}^2 = 0.2, \sigma_B^2 = 0.3, \sigma_J^2 = 0.4$ である．

これらの図において曲線は理論計算の結果を，$\times, +, *, \square, \blacksquare$ などの印は計算機シミュレーションの結果を表す．また，η_J 以外の条件は共通で $\eta_B = 1.0, \sigma_{\hat{B}}^2 = 0.2, \sigma_B^2 = 0.3, \sigma_J^2 = 0.4$ である．図 2 と図 4 は $\eta_J = 1.2$ の場合の結果であり，図 3 と図

5 は $\eta_J = 0.3$ の場合の結果である．これらを見ると以下のことがわかる．

図 2 より，生徒の学習係数が $\eta_J = 1.2$ と比較的大きい場合には，真の教師と生徒の汎化誤差 $\bar{\epsilon}_g$ は真の教師と動く教師の汎化誤差 $\hat{\epsilon}_g$ よりも常に大きくなっていることがわかる．また，動く教師と生徒の汎化誤差 ϵ_g は $\bar{\epsilon}_g$ よりもさらに大きい．図 4 より，このとき真の教師と生徒の方向余弦 $\bar{R}$ は真の教師と動く教師の方向余弦 $\hat{R}$ よりも常に小さいことがわかる．

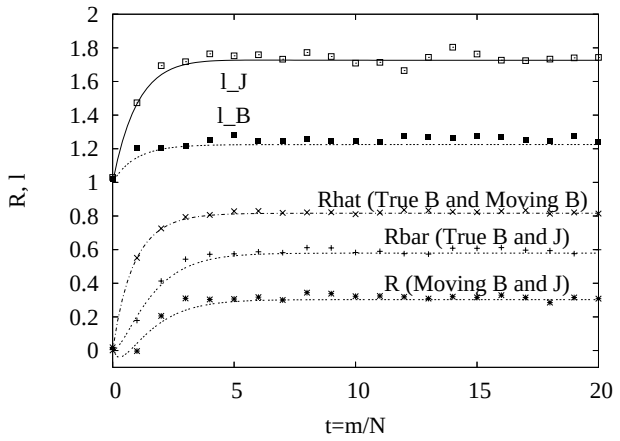

図 4　$\eta_J = 1.2$ の場合の R と l．理論と計算機シミュレーション．η_J 以外の条件は $\eta_B = 1.0, \sigma_B^2 = 0.2, \sigma_B^2 = 0.3, \sigma_J^2 = 0.4$ である．

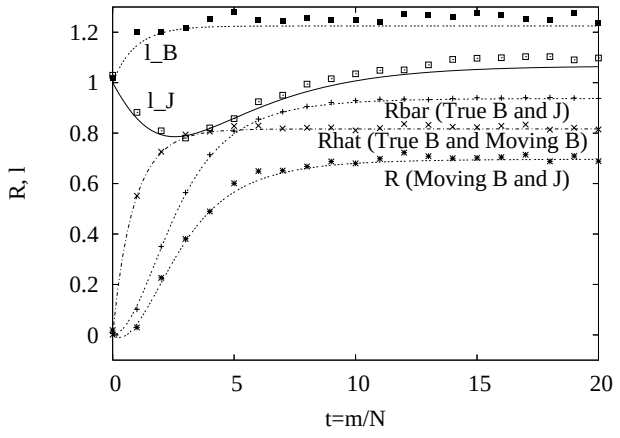

図 5　$\eta_J = 0.3$ の場合の R と l．理論と計算機シミュレーション．η_J 以外の条件は $\eta_B = 1.0, \sigma_B^2 = 0.2, \sigma_B^2 = 0.3, \sigma_J^2 = 0.4$ である．

これに対して，図 3 より，生徒の学習係数が $\eta_J = 0.3$ と比較的小さい場合，学習の初期においては $\eta_J = 1.2$ の場合と同様に真の教師と生徒の汎化誤差 $\bar{\epsilon}_g$ は真の教師と動く教師の汎化誤差 $\hat{\epsilon}_g$ 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 $\bar{\epsilon}_g$ が $\hat{\epsilon}_g$ 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 5 より，このとき，学習の初期においては真の教師と生徒の方向余弦 $\bar{R}$ は真の教師と動く教師の方向余弦 $\hat{R}$ よりも小さいが，$t = 5.2$ でその大小関係が逆転し，それ以後は $\bar{R}$ が $\hat{R}$ よりも大きくなっている．すなわち，生徒は動く教師の入出力しか見ていないにもかかわらず，動く教師よりも真の教師の方向に近づいていることになる．

図 3 と図 5 で大小関係が逆転する時刻にずれがあるのは，本研究では線形パーセプトロンを対象とし，誤差として二乗誤差を用いているので，汎化誤差は式 (25)，(28)，(31) に示すように，方向余弦 $\hat{R}, R, \bar{R}$ だけではなく長さ l_B, l_J にも依存するためである．

いずれにせよ，学習係数の値によっては，動く教師よりも生徒の方が高性能になりうることを示しており，この結果は非常に興味深い．

また，図 4，図 5 のいずれにおいても，動く教師と生徒の方向余弦 R が学習の初期にわずかながらいったん負になっている．すなわち，動く教師と生徒のなす角度が初期状態よりもいったん大きくなる．これは，学習開始時に生徒がいったん“ 置いて行かれる ”ことを表しており，興味深い現象である．

図 2〜図 5 を見ると，汎化誤差や R, l は $t = 20$ でほぼ定常値に達しているように見えるが，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，式 (54)–(58) の指数関数のべきの符号に着目することにより $0 < \eta_B < 2$ でなければ $\hat{\epsilon}_g$,ϵ_g は発散し，$0 < \eta_J < 2$ でなければ ϵ_g,$\bar{\epsilon}_g$ は発散することがわかる．$0 < \eta_B, \eta_J < 2$ の場合については，式 (54)–(58) において $t \to \infty$ とすることにより汎化誤差や R,l の定常値は容易に得られる．このようにして得られた生徒の学習係数 η_J と汎化誤差,R,l の定常値の関係を図 6，図 7，図 8 に示す．η_J 以外の条件は図 2〜図 5 と同様 $\eta_B = 1.0, \sigma_B^2 = 0.2, \sigma_B^2 = 0.3, \sigma_J^2 = 0.4$ である．また，計算機シミュレーションでは十分定常に達したと判断される $t = 50$ の値をプロットしている．

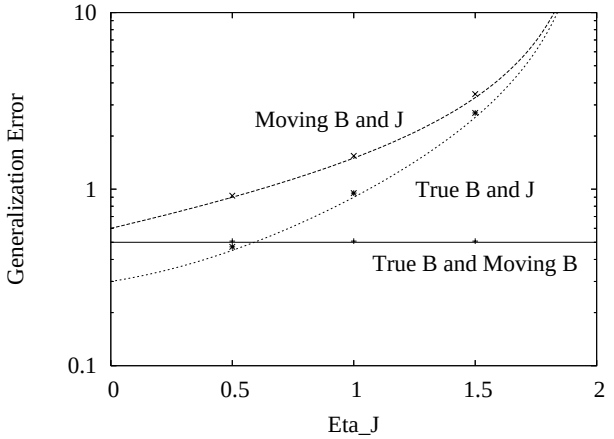

図 6　汎化誤差の定常値．理論と計算機シミュレーション．η_J 以外の条件は $\eta_B = 1.0, \sigma_B^2 = 0.2, \sigma_B^2 = 0.3, \sigma_J^2 = 0.4$

これらの図から以下のことがわかる．生徒の学習係数 η_J が 0.58 よりも大きいときには真の教師と生徒の定常汎化誤差は真の教師と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．定常 R と定常 l については $\eta_J = 0.70$ で定常汎化誤差と同

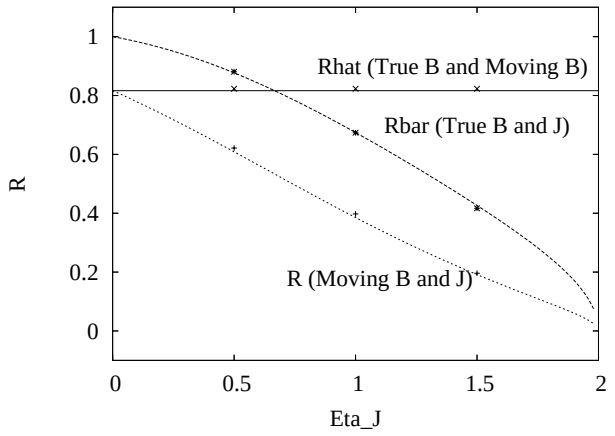

図 7　R の定常値．理論と計算機シミュレーション．η_J 以外の条件は $\eta_B = 1.0, \sigma_{\hat{B}}^2 = 0.2, \sigma_B^2 = 0.3, \sigma_J^2 = 0.4$

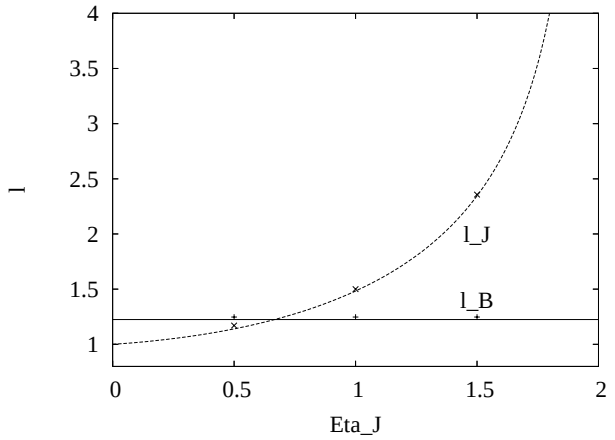

図 8　l の定常値．理論と計算機シミュレーション．η_J 以外の条件は $\eta_B = 1.0, \sigma_{\hat{B}}^2 = 0.2, \sigma_B^2 = 0.3, \sigma_J^2 = 0.4$

様に大小関係の逆転が起こる．また，$\eta_J \to 0$ で l_J は 1 に，R は $\hat{R}$ に，$\bar{R}$ は 1 にそれぞれ漸近する．すなわち，$\eta_J \to 0$ で生徒 $\boldsymbol{J}$ は真の教師 $\hat{\boldsymbol{B}}$ に方向，長さともに一致する．なお，このときに図 6 において真の教師と生徒の汎化誤差 $\bar{\epsilon}_g$ がゼロになっていないのは両者に独立な雑音が重畳されているからである．また，図 6〜図 8 から $\eta_J = 2$ で $\bar{R}$，R が 0 になり，l_J，ϵ_g，$\bar{\epsilon}_g$ が発散する相転移現象が起こることが確認される．

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，真の教師と動く教師の汎化誤差よりも真の教師と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　特定領域研究（課題番号 14084212），同　基盤研究 (C)（課題番号 14580438, 15500151 16500093 ）によるものであり，ここに感謝いたします．

文　献

[1] Saad, D. (ed.), On-line Learning in Neural Networks, Cambridge University Press, (1998)
[2] Freund, Y. and Shapire, R.E.,（安倍直樹訳），“ブースティング入門，” 人工知能学会誌，**14**(5), 771–780 (1999).
[3] http://www.boosting.org/
[4] 麻生 英樹，津田 宏治，村田 昇，“パターン認識と学習の統計学，” 岩波書店, 東京,2003.
[5] Krogh, A. and Sollich, P., “Statistical mechanics of ensemble learning,” Phys. Rev. E, **55**(1), 811–825 (1997).
[6] Urbanczik, R., “Online learning with ensembles,” Phys. Rev. E, **62**(1), 1448–1451 (2000).
[7] 原 一之，岡田 真人，“線形ウィークラーナーによるアンサンブル学習の汎化誤差の解析，” 情報論的学習理論ワークショップ予稿集, 113–118 (2002).
[8] 岡田 真人，原 一之，三好 誠司,“［チュートリアル講演］アンサンブル学習 ”，信学技報，NC2003-35, pp.7-12, 2003.7
[9] Miyoshi,S., Hara,K. and Okada,M., “Analysis of ensemble learning using simple perceptrons based on online learning theory”, Phys. Rev. E, 71, 036116. March 2005.
[10] 三好 誠司，原 一之，岡田 真人，“オンライン学習理論に基づく単純パーセプトロンのアンサンブル学習の解析”，信学論 DII, **J87-D-II**(7), pp.1391–1401 (2004).
[11] 西森 秀俊，“スピングラス理論と情報統計力学，” 岩波書店, 東京,1999.
[12] Nishimori, H., “Statistical Physics of Spin Glasses and Information Processing: An Introduction,” Oxford University Press, (2001)
[13] Inoue, J. and Nishimori, H., “On-line AdaTron learning of a unlearnable rules,” Phys. Rev. E, **55**(4), 4544–4551 (1997).
[14] Inoue, J., Nishimori, H. and Kabashima, Y., “A simple perceptron that learns non-monotonic rules,” cond-mat/9708096 (1997).
[15] 三好 誠司，原 一之，岡田 真人,“ 教師がコミティマシンの場合のアンサンブル学習 ”，日本神経回路学会第 14 回全国大会 (JNNS2004) 講演論文集，pp.36-37，2004.9
[16] 三好 誠司，原 一之，岡田 真人,“ 教師がコミティマシンの場合のアンサンブル学習 ”，信学技報，NC2004-79, pp.63-68, 2004.10
[17] 三好 誠司，原 一之，岡田 真人,“ 教師がコミティマシンの場合のアンサンブル学習 ”，情報論的学習理論ワークショップ (IBIS2004) 予稿集，pp.178-185, 2004.11
[18] 三好 誠司，原 一之，岡田 真人,“ 教師が非単調な場合のアンサンブル学習 ”，信学技報，NC2004-214, pp.123-128, 2005.3
[19] 三好 誠司，原 一之，岡田 真人,“ 教師が非単調な場合のアンサンブル学習 ”，日本物理学会年次大会，24aYB-6, 2005.3