2003 年情報論的学習理論ワークショップ
2003 Workshop on Information-Based Induction Sciences (IBIS2003)
Kyoto, Japan, Nov 11 - 12, 2003.

オンライン学習理論における非線形パーセプトロンのアンサンブル学習の解析

Analysis of ensemble learning using nonlinear perceptrons based on on-line learning theory

三好 誠司* 原 一之† 岡田 真人‡
Seiji MIYOSHI Kazuyuki HARA Masato OKADA

Abstract: We discuss the ensemble learning using K nonlinear simple perceptrons of which an output function is the sign function based on the on-line learning in the finite K case. First, we derive a macroscopic differential equation describing a dynamics of correlation q between the student weight vectors in a general learning algorithm. Second, we apply the equation to the three well-known learning rules, that is the Hebb rule, the Perceptron rule and the AdaTron rule, and solve those numerically. Third, we obtain the generalization error of these ensemble machines using a majority vote of students. As a result, we show that the correlation between the student weight vectors in the AdaTron rule evolves most slowly, and that the AdaTron rule is the most superior among the three learning rules in the framework of the ensemble learning.

Keyword: ensemble learning, on-line learning, nonlinear perceptron, Perceptron rule, Hebb rule, AdaTron rule, generalization error

1 まえがき

精度の低いルールや学習機械（以後は生徒と呼ぶ）を多数組み合わせることにより精度の高い予測や分類を行おうとすることは一般にアンサンブル学習と呼ばれ，近年注目されている [1, 2, 3, 4, 5, 6]．アンサンブル学習の汎化能力を統計力学的手法によって理論的に解析する研究もさかんに行われている [4, 5, 6].

原と岡田は，生徒が線形パーセプトロンである場合について理論的な解析を行った [4]．一方，符号関数を出力関数とするような非線形パーセプトロンの学習則としてはヘブ学習，パーセプトロン学習，アダトロン学習がよく知られている [7, 8, 9, 10]．Urbanczik[6] は，符号関数を出力関数とする非線形パーセプトロンによるアンサンブル学習をオンライン学習の枠組みで解析したが，生徒の数 K が大きい極限のみを扱っている．また，代表的な学習則であるヘブ学習，パーセプトロン学習，アダトロン学習の三つの学習則をアンサンブル学習に適用した場合の違いはたいへん興味深い課題であるが，この点に着目した解析はこれまで行われていない．

そこで本論文では，これらの先行研究をふまえて，符号関数を出力関数とするような K 個の非線形パーセプトロンによるアンサンブル学習を，オンライン学習の枠組みで，かつ，有限の K で議論する．まず，K 個の生徒が多数決で統合出力を決定する場合の汎化誤差が教師と生徒の類似度と生徒間の類似度という二つの巨視的変数で計算できることを示す．次に，一般の学習則について，これらの巨視的変数のダイナミクスを記述する微分方程式を導出する．さらに，よく知られているヘブ学習，パーセプトロン学習，アダトロン学習の三つの学習則 [7] について，この微分方程式を具体的に導出し，それらを解いた結果を用いて汎化誤差を数値的に求める．その結果，これら三つの学習則は「生徒の多様性維持」というアンサンブル学習との相性という点でそれぞれ異

*神戸市立工業高等専門学校 電子工学科, 651-2194 神戸市西区学園東町 8-3, tel. 078-795-3247, e-mail miyoshi@kobe-kosen.ac.jp,
Department of Electronic Engineering, Kobe City College of Technology, 8-3, Gakuenhigashimachi, Nishi-ku, Kobe 651-2194, Japan
†東京都立工業高等専門学校 電子情報工学科, 140-0011 東京都品川区東大井 1-10-40, tel. 03-3471-6331, e-mail hara@tokyo-tmct.ac.jp,
Devision of electronics and information engineering, Tokyo Metropolitan College of Technology, 1-10-40, Higashi-oi, Shinagawa-ku, Tokyo 140-0011, Japan
‡理化学研究所 脳科学総合研究センター, 科学技術振興事業団 さきがけ, 351–0198 埼玉県和光市広沢 2–1, tel. 048-467-6845, e-mail okada@brain.riken.go.jp,
Laboratory for Mathematical Neuroscience, RIKEN Brain Science Institute and Intelligent Cooperation and Control, PRESTO, Japan Science and Technology Corporation, 2-1, Hirosawa, Wako, Saitama 351-0198, Japan,

なった性質を有しており，汎化誤差の漸近特性の点で最も優れている [7] ことが知られているアダトロン学習が，アンサンブル学習との相性という点でも最も優れているという興味深い事実が明らかになる．

2 モデル

本論文で対象とする生徒は，符号関数を出力関数とするパーセプトロンである．K 個の生徒からなるアンサンブルを考え，各生徒の結合荷重を $\boldsymbol{J}_1, \boldsymbol{J}_2, \cdots, \boldsymbol{J}_K$ とする．$\boldsymbol{J}_k = (J_{k1}, \cdots, J_{kN}), k = 1, 2, \cdots, K$ と入力 $\boldsymbol{x} = (x_1, \cdots, x_N)$ は N 次元ベクトルであり，$\boldsymbol{x}$ の各要素 x_i は平均 0，分散 $1/N$ のガウス分布に従う独立な確率変数であるとする．また，$\boldsymbol{J}_k$ の初期値 $\boldsymbol{J}_k^0$ の各要素 J_{ki}^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．すなわち，

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

ここで，$\langle \cdot \rangle$ は平均を表す．各生徒の出力は $\mathrm{sgn}(u_1 l_1)$, $\mathrm{sgn}(u_2 l_2)$, $\cdots$, $\mathrm{sgn}(u_K l_K)$ である．ここで，

$$\mathrm{sgn}(ul) = \begin{cases} +1, & ul \geq 0, \\ -1, & ul < 0, \end{cases} \quad (2)$$

$$u_k l_k = \boldsymbol{J}_k \cdot \boldsymbol{x}, \quad (3)$$

である．l_k については後で述べる．また，u_k を各生徒の規格化内部状態と呼ぶことにする．

教師機械も符号関数を出力関数とするパーセプトロンであるとし，その結合荷重を $\boldsymbol{B}$ とする．本論文では $\boldsymbol{B}$ は不変とする．ここで，生徒の初期値同様に教師 $\boldsymbol{B} = (B_1, \cdots, B_N)$ は N 次元ベクトルであり，$\boldsymbol{B}$ 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るものとする．すなわち，$\langle B_i \rangle = 0$, $\left\langle (B_i)^2 \right\rangle = 1$ である．教師の出力は $\mathrm{sgn}(v)$ である．ここで，$v = \boldsymbol{B} \cdot \boldsymbol{x}$ である．v を教師の内部状態と呼ぶことにする．

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，

$$|\boldsymbol{x}| = 1,\ \ |\boldsymbol{B}| = \sqrt{N},\ \ |\boldsymbol{J}_k^0| = \sqrt{N}, \quad (4)$$

となる．生徒の大きさ $|\boldsymbol{J}_k|$ は一般には時間の経過とともに変化するが，$\sqrt{N}$ に対する比を l_k とし，これを生徒 $\boldsymbol{J}_k$ の長さと呼ぶことにする．すなわち，$|\boldsymbol{J}_k| = l_k \sqrt{N}$ である．l_k は本論文で扱う巨視的変数のひとつである．

教師と個々の生徒には共通の入力 $\boldsymbol{x}$ が同じ順序で与えられる．個々の生徒は入力 $\boldsymbol{x}$ に対する教師の出力と自分の出力を比べ，教師と同じ出力を出す確率が上がるように，必要に応じて自分の結合荷重を修正していく．この手続きを学習と呼ぶ．修正の方法は学習則と呼ばれ，ヘブ学習，パーセプトロン学習，アダトロン学習がよく知られている [7, 8, 9, 10]．自分自身に関する情報以外に生徒が修正のために使える情報は，入力 $\boldsymbol{x}$ とそれに対する教師の出力 $\mathrm{sgn}(v)$ だけであるから，学習は一般に以下のように表せる．ここで，m は時刻ステップを表す．

$$\boldsymbol{J}_k^{m+1} = \boldsymbol{J}_k^m + f(\mathrm{sgn}(v^m), u_k^m) \boldsymbol{x}^m. \quad (5)$$

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．本論文では，K 個の非線形単純パーセプトロンが単純多数決でアンサンブルとしての出力を決定するものとする．このとき，誤差 ϵ として，

$$\epsilon = \Theta\left(-\mathrm{sgn}\left(\boldsymbol{B} \cdot \boldsymbol{x} \right) \sum_{k=1}^{K} \mathrm{sgn}\left(\boldsymbol{J}_k \cdot \boldsymbol{x} \right) \right) \quad (6)$$

を用いることにする．ここで，$\Theta(\cdot)$ は以下のようなステップ関数である．

$$\Theta(z) = \begin{cases} +1, & z \geq 0, \\ 0, & z < 0. \end{cases} \quad (7)$$

汎化誤差 ϵ_g は式 (6) を入力 $\boldsymbol{x}$ の確率分布 $p(\boldsymbol{x})$ で平均したものと定義する．すなわち，汎化誤差 ϵ_g は新たな入力 $\boldsymbol{x}$ に対するアンサンブルの出力が教師の出力と異なる確率と言うこともできる．誤差 ϵ は，教師の内部状態 v と生徒の規格化内部状態 u_k を用いて，$\epsilon = \epsilon(\{u_k\}, v)$ と書くことができるので，汎化誤差 ϵ_g も u_k，v の確率分布 $p(\{u_k\}, v)$ を用いて，

$$\epsilon_g = \int d\boldsymbol{x} p(\boldsymbol{x}) \epsilon = \int \prod_{k=1}^{K} du_k dv p(\{u_k\}, v) \epsilon, \quad (8)$$

と書ける．v と u_k は入力 $\boldsymbol{x}$ とそれとは無関係な変数 $\boldsymbol{B}$, $\boldsymbol{J}_k$ で書けるので，$p(\{u_k\}, v)$ は平均 0 の多重ガウス分布である．ここで，u_k と v は平均 0 分散 1 のガウス分布にしたがうので，$p(\{u_k\}, v)$ の共分散行列 $\boldsymbol{\Sigma}$ の対角要素は 1 である．次に，この行列の非対角要素を求めるために，結合荷重間の方向余弦を議論する．まず，教師 $\boldsymbol{B}$ と生徒 $\boldsymbol{J}_k$ の方向余弦として R_k を定義する．すなわち，

$$R_k \equiv \frac{\boldsymbol{B} \cdot \boldsymbol{J}_k}{|\boldsymbol{B}||\boldsymbol{J}_k|} = \frac{1}{l_k N} \sum_{i=1}^{N} B_i J_{ki}. \quad (9)$$

教師 $\boldsymbol{B}$ と生徒 $\boldsymbol{J}_k$ に相関がなければ $R_k = 0$ であり，両者の方向が同じであれば $R_k = 1$ であるから，以後は R_k のことを教師と生徒の類似度と呼ぶことにする．R_k は本研究で扱う二番目の巨視的変数である．

また，生徒 $\boldsymbol{J}_k$ と生徒 $\boldsymbol{J}_{k'}$ の方向余弦として $q_{kk'}$ を定義する．すなわち，

$$q_{kk'} \equiv \frac{\boldsymbol{J}_k \cdot \boldsymbol{J}_{k'}}{|\boldsymbol{J}_k||\boldsymbol{J}_{k'}|} = \frac{1}{l_k l_{k'} N} \sum_{i=1}^{N} J_{ki} J_{k'i}, \quad (10)$$

ここで，$k \neq k'$ である．

生徒 $\boldsymbol{J}_k$ と生徒 $\boldsymbol{J}_{k'}$ に相関がなければ $q_{kk'}=0$ であり，両者の方向が同じであれば $q_{kk'}=1$ であるから，以後は $q_{kk'}$ のことを生徒間の類似度と呼ぶことにする．$q_{kk'}$ は本研究で扱う三番目の巨視的変数である．

教師 $\boldsymbol{B}$ の内部状態 v と生徒 $\boldsymbol{J}_k$ の規格化内部状態 u_k の共分散は教師 $\boldsymbol{B}$ と生徒 $\boldsymbol{J}_k$ の類似度 R_k に等しい．また，生徒 $\boldsymbol{J}_k$ の規格化内部状態 u_k と生徒 $\boldsymbol{J}_{k'}$ の規格化内部状態 $u_{k'}$ の共分散は生徒間の類似度 $q_{kk'}$ に等しい．よって，式 (6)，式 (8) より汎化誤差 ϵ_g は R_k と $q_{kk'}$ を用いて以下のように書ける．

$$\epsilon_g = \int \prod_{k=1}^{K} du_k dv p(\{u_k\},v)\Theta\left(-\mathrm{sgn}(v)\sum_{k=1}^{K}\mathrm{sgn}(u_k)\right), \tag{11}$$

$$\begin{aligned} p(\{u_k\},v) &= \frac{1}{(2\pi)^{\frac{K+1}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \\ &\times \exp\left(-\frac{(\{u_k\},v)\boldsymbol{\Sigma}^{-1}(\{u_k\},v)^T}{2}\right), \end{aligned} \tag{12}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & q_{12} & \dots & q_{1K} & R_1 \\ q_{21} & 1 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & q_{K-1,K} & \vdots \\ q_{K1} & \dots & q_{K,K-1} & 1 & R_K \\ R_1 & \dots & \dots & R_K & 1 \end{pmatrix}. \tag{13}$$

3.2 巨視的変数の微分方程式

式 (11)–(13) より，汎化誤差 ϵ_g は教師と生徒の類似度 R_k と生徒間の類似度 $q_{kk'}$ がすべてわかれば計算できる．よってここではこれらの巨視的変数のダイナミクスを記述する微分方程式について議論する．

本論文では，入力，教師，生徒の大きさを式 (4) のように設定しているので，N が大きい極限では入力の影響は入力に関する平均（サンプル平均）で置き換えることができる．この考え方を統計物理の分野では自己平均性と呼ぶ．自己平均性に基づく一般の学習則の l_k，R_k に関する微分方程式はすでに求められており，以下の通りである [7].

$$\frac{dl_k}{dt} = \langle f_k u_k\rangle + \frac{\langle f_k^2\rangle}{2l_k}, \tag{14}$$

$$\frac{dR_k}{dt} = \frac{\langle f_k v\rangle - \langle f_k u_k\rangle R_k}{l_k} - \frac{R_k}{2l_k^2}\langle f_k^2\rangle. \tag{15}$$

ここで，$\langle\cdot\rangle$ はサンプル平均を表す．すなわち，

$$\langle f_k u_k\rangle = \int du_k dv p_2(u_k,v) f(\mathrm{sgn}(v),u_k)u_k, \tag{16}$$

$$\langle f_k v\rangle = \int du_k dv p_2(u_k,v) f(\mathrm{sgn}(v),u_k)v, \tag{17}$$

$$\langle f_k^2\rangle = \int du dv p_2(u_k,v)\left(f(\mathrm{sgn}(v),u_k)\right)^2, \tag{18}$$

$$p_2(u_k,v) = \frac{1}{2\pi|\boldsymbol{\Sigma}_2|^{\frac{1}{2}}}\exp\left(-\frac{(u_k,v)\boldsymbol{\Sigma}_2^{-1}(u_k,v)^T}{2}\right) \tag{19}$$

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & R_k \\ R_k & 1 \end{pmatrix}. \tag{20}$$

次に，一般の学習則の $q_{kk'}$ に関する微分方程式を導出する．いま，生徒 $\boldsymbol{J}_k$ と生徒 $\boldsymbol{J}_{k'}$ を考える．$l_k^m \to l_k$，$l_k^{m+1} \to l_k + dl_k$，$q_{kk'}^m \to q_{kk'}$，$q_{kk'}^{m+1} \to q_{kk'} + dq_{kk'}$，$1/N \to dt$ とおき，自己平均性を仮定すると式 (5)，(10)，(14) より q に関する微分方程式が以下のように得られる．

$$\begin{aligned} \frac{dq_{kk'}}{dt} &= \frac{\langle f_{k'}u_k\rangle - q_{kk'}\langle f_{k'}u_{k'}\rangle}{l_{k'}} + \frac{\langle f_k u_{k'}\rangle - q_{kk'}\langle f_k u_k\rangle}{l_k} \\ &+ \frac{\langle f_k f_{k'}\rangle}{l_k l_{k'}} - \frac{q_{kk'}}{2}\left(\frac{\langle f_k^2\rangle}{l_k^2} + \frac{\langle f_{k'}^2\rangle}{l_{k'}^2}\right). \end{aligned} \tag{21}$$

ここで，

$$\langle f_k u_{k'}\rangle = \int du_k du_{k'} dv p_3(u_k,u_{k'},v) f(\mathrm{sgn}(v),u_k)u_{k'}, \tag{22}$$

$$\langle f_{k'} u_k\rangle = \int du_k du_{k'} dv p_3(u_k,u_{k'},v) f(\mathrm{sgn}(v),u_{k'})u_k, \tag{23}$$

$$\begin{aligned} \langle f_k f_{k'}\rangle = \int du_k du_{k'} dv p_3(u_k,u_{k'},v) \\ \times f(\mathrm{sgn}(v),u_k) f(\mathrm{sgn}(v),u_{k'}), \end{aligned} \tag{24}$$

$$\begin{aligned} p_3(u_k,u_{k'},v) &= \frac{1}{(2\pi)^{\frac{3}{2}}|\boldsymbol{\Sigma}_3|^{\frac{1}{2}}} \\ &\times \exp\left(-\frac{(u_k,u_{k'},v)\boldsymbol{\Sigma}_3^{-1}(u_k,u_{k'},v)^T}{2}\right) \end{aligned} \tag{25}$$

$$\boldsymbol{\Sigma}_3 = \begin{pmatrix} 1 & q_{kk'} & R_k \\ q_{k'k} & 1 & R_{k'} \\ R_k & R_{k'} & 1 \end{pmatrix}. \tag{26}$$

4 ダイナミクス

4.1 計算の条件

学習が進むにつれ，教師と生徒の類似度 R_k と生徒間の類似度 $q_{kk'}$ は 0 から 1 に近づいていく．このとき，R_k と $q_{kk'}$ は相互に何らかの拘束関係にあるのであるが，R_k と比較して $q_{kk'}$ が小さいほど生徒の多様性が維持されていることになるわけだから，アンサンブル学習の効果が大きいと言える．逆に，$q_{kk'}=1$ になってしまえば生徒 $\boldsymbol{J}_k$ と生徒 $\boldsymbol{J}_{k'}$ の出力は同一であるのでこれらを組み合わせる意味はない．このように，アンサンブル学習においては R_k と $q_{kk'}$ の関係が本質的である．

すでに述べたように，本論文では生徒 $\boldsymbol{J}_k$ の初期値 $\boldsymbol{J}_k^0$，教師 $\boldsymbol{B}$ 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，

$$R_k^0 = 0, \quad q_{kk'}^0 = 0 \tag{27}$$

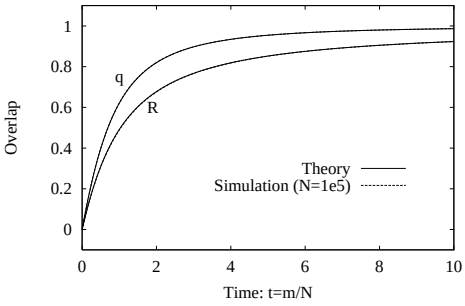

図 1: R と q（ヘブ学習）

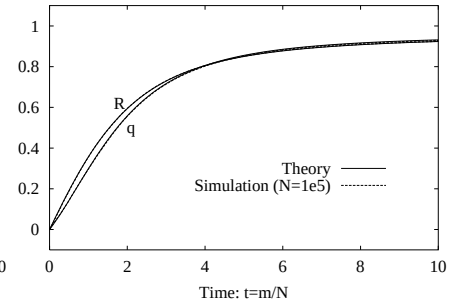

図 2: R と q（パーセプトロン学習）

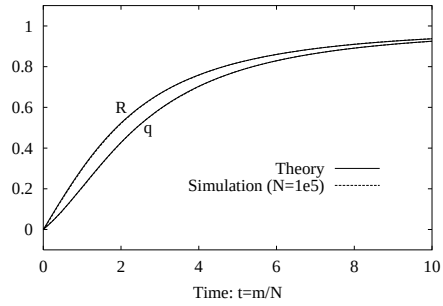

図 3: R と q（アダトロン学習）

である．式 (27) と生徒の対称性より，式 (21) において，

$$\langle f_k u_{k'} \rangle = \langle f_{k'} u_k \rangle\,, \quad \langle f_k f_{k'} \rangle = \langle f_{k'} f_k \rangle \tag{28}$$

が成り立つ．また，式 (27) と生徒の対称性より，式 (14)–(21) の巨視的変数 $l_k, R_k, q_{kk'}$ から添え字 k, k' を落としてそれぞれを l, R, q と書くことにする．次節以降では代表的な三つの学習則について式 (27)–(28) の条件下で式 (14)，(15)，(21) を計算するために必要な五つのサンプル平均 $\langle f_k u_k \rangle$，$\langle f_k v \rangle$，$\langle f_k^2 \rangle$，$\langle f_k u_{k'} \rangle$，$\langle f_k f_{k'} \rangle$ を具体的に議論する．

4.2 ヘブ学習

ヘブ学習は以下の式で更新を行う学習則である．

$$f(\mathrm{sgn}(v), u) = \mathrm{sgn}(v) \tag{29}$$

ヘブ学習の $\langle f_k u_k \rangle$，$\langle f_k v \rangle$，$\langle f_k^2 \rangle$ は式 (29) を用いて式 (16)–(18) を解析的に実行することにより，以下のように求められる [7].

$$\langle f_k u_k \rangle = \frac{2R}{\sqrt{2\pi}}, \quad \langle f_k v \rangle = \sqrt{\frac{2}{\pi}}, \quad \langle f_k^2 \rangle = 1. \tag{30}$$

ここでは新たに $\langle f_k u_{k'} \rangle$ と $\langle f_k f_{k'} \rangle$ を導出する．式 (29) は u に依存しない形なので，

$$\langle f_k u_{k'} \rangle = \langle f_k u_k \rangle = \frac{2R}{\sqrt{2\pi}}, \tag{31}$$

$$\langle f_k f_{k'} \rangle = \left\langle (\mathrm{sgn}(v))^2 \right\rangle = 1. \tag{32}$$

式 (14)，(15)，(21)，(27)，(28)，(30)–(32) を数値的に解いて得られた R と q のダイナミクスに関する理論計算の結果と計算機実験 ($N = 10^5$) の比較を図 1 に示す．ヘブ学習においては，q の立ち上がりが R の立ち上がりよりも速いことがわかる．すなわち，ヘブ学習においては R と比較して q が大きい．このことは，ヘブ学習においては生徒の多様性が急速に失われることを表している．

4.3 パーセプトロン学習

パーセプトロン学習は以下の式で更新を行う学習則である．

$$f(\mathrm{sgn}(v), u) = \Theta(-uv)\,\mathrm{sgn}(v). \tag{33}$$

パーセプトロン学習の $\langle f_k u_k \rangle$，$\langle f_k v \rangle$，$\langle f_k^2 \rangle$ は式 (33) を用いて式 (16)–(18) を解析的に実行することにより，以下のように求められる [7].

$$\langle f_k u_k \rangle = \frac{R-1}{\sqrt{2\pi}}, \quad \langle f_k v \rangle = \frac{1-R}{\sqrt{2\pi}}, \tag{34}$$

$$\langle f_k^2 \rangle = \frac{1}{\pi} \tan^{-1} \frac{\sqrt{1-R^2}}{R}. \tag{35}$$

ここでは新たに $\langle f_k u_{k'} \rangle$ と $\langle f_k f_{k'} \rangle$ を導出する．式 (33) を用いて式 (22)，(24) を解析的に実行することにより，パーセプトロン学習の $\langle f_k u_{k'} \rangle$，$\langle f_k f_{k'} \rangle$ は以下のように求められる．

$$\langle f_k u_{k'} \rangle = \frac{R-q}{\sqrt{2\pi}} \tag{36}$$

$$\langle f_k f_{k'} \rangle = 2 \int_0^\infty Dv \int_{\frac{Rv}{\sqrt{1-R^2}}}^\infty DxH(z) \tag{37}$$

ここで，

$$z \equiv \frac{-(q-R^2)x + R\sqrt{1-R^2}v}{\sqrt{(1-q)(1+q-2R^2)}} \tag{38}$$

であり，$H(u)$ と Dx の定義は以下の通りである．

$$H(u) \equiv \int_u^\infty Dx, \quad Dx \equiv \frac{dx}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \tag{39}$$

式 (14)，(15)，(21)，(27)，(28)，(34)–(37) を数値的に解いて得られた R と q のダイナミクスに関する理論計算の結果と計算機実験 ($N = 10^5$) の比較を図 2 に示す．パーセプトロン学習ではヘブ学習と異なり，学習の初期 ($t < 4.0$) においては q の方が R よりも小さいことがわかる．すなわち，パーセプトロン学習ではヘブ学習よりも生徒の多様性が長時間維持されていると言える．

4.4 アダトロン学習

アダトロン学習は以下の式で更新を行う学習則である．

$$f(\mathrm{sgn}(v), u) = -u\Theta(-uv). \tag{40}$$

アダトロン学習の $\langle f_k u_k \rangle$，$\langle f_k v \rangle$，$\langle f_k^2 \rangle$ は式 (40) を用いて式 (16)–(18) を解析的に実行することにより，以

下のように求められる [7].

$$\langle f_k u_k \rangle = -\frac{1}{\pi}\cot^{-1}\left(\frac{R}{\sqrt{1-R^2}}\right) + \frac{R\sqrt{1-R^2}}{\pi} \quad (41)$$

$$\langle f_k v \rangle = \frac{\left(1-R^2\right)^{\frac{3}{2}}}{\pi} + R\langle f_k u_k \rangle \quad (42)$$

$$\left\langle f_k^2 \right\rangle = -\langle f_k u_k \rangle \quad (43)$$

ここでは新たに $\langle f_k u_{k'} \rangle$ と $\langle f_k f_{k'} \rangle$ を導出する．式 (40) を用いて式 (22)，(24) を解析的に実行することにより，アダトロン学習の $\langle f_k u_{k'} \rangle$，$\langle f_k f_{k'} \rangle$ は以下のように求められる．

$$\langle f_k u_{k'} \rangle = \frac{1+q}{\pi} R\sqrt{1-R^2} - 2q\int_0^\infty Dv \int_{\frac{Rv}{\sqrt{1-R^2}}}^\infty Dx x^2 \quad (44)$$

$$\begin{aligned}\langle f_k f_{k'} \rangle =& \frac{(1-q)^2\left(1+q-2R^2\right)}{2\pi\left(1-R^2\right)^{\frac{3}{2}}}\left(\sqrt{\frac{(1+q)\left(1-R^2\right)}{1-q}} - R\right) \\ &+ 2(q-R^2)\int_0^\infty Dv \int_{\frac{Rv}{\sqrt{1-R^2}}}^\infty Dx x^2 H(z) \\ &- \frac{2R\left(1+q-R^2\right)}{\sqrt{1-R^2}}\int_0^\infty Dv v \int_{\frac{Rv}{\sqrt{1-R^2}}}^\infty Dx x H(z) \\ &+ 2R^2\int_0^\infty Dv v^2 \int_{\frac{Rv}{\sqrt{1-R^2}}}^\infty Dx H(z) \end{aligned} \quad (45)$$

ここで，z，$H(u)$，Dx の定義は式 (38)，(39) である．

式 (14)，(15)，(21)，(27)，(28)，(41)–(45) を数値的に解いて得られた R と q のダイナミクスに関する理論計算の結果と計算機実験 ($N = 10^5$) の比較を図 3 に示す．アダトロン学習においてはヘブ学習やパーセプトロン学習よりも R と比較した q が小さいことがわかる．すなわち，三つの学習則で生徒の多様性がもっとも維持されるのはアダトロン学習である．

5 議論

前節では，三つの学習則について R と q のダイナミクスを理論的に導いた．図 1，図 2，図 3 より，三つの学習則のうち，R と比較して q がもっとも小さい学習則はアダトロン学習であることがわかる．すでに述べたように，アンサンブル学習においては R と q の関係が本質的である．そこで，両者の関係をより明確にするため，R と q を軸に取って描き直したグラフを図 4 に示す．この図を見ると，アダトロン学習の特性がもっとも下になっている．すなわち，R と比較して q が最も小さい学習則はアダトロン学習であることが明確になった．言い換えるならば，q の立ち上がりが最も遅く，生徒の多様性が長時間維持される学習則はアダトロン学習である．

ここまでの議論で，三つの学習則のうち，アンサンブル学習を行うメリットがもっとも大きい学習則はアダトロン学習であることが期待される．この予想を確認するため，前節までで得られた各学習則の R，q の値と式 (8)–(13) から $K = 3$ における汎化誤差 ϵ_g を数値的に求め，計算機実験の結果と比較するとともに，三つの学習則を比較した．結果を図 5〜図 7 に示す．計算機実験においては $N = 10^4$ とし，各時点で 10^5 個のランダム入力によるテストにより汎化誤差 ϵ_g を計算した．

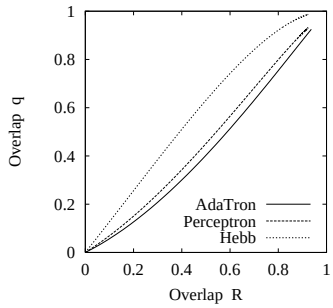

図 4: R と q の関係（理論）

これら三つの図より，K を 1 から 3 にすることにより汎化誤差が最も大きく改善される学習則はアダトロン学習であることがわかる．すなわち，図 4 の R と q の関係から予想されたとおり，アンサンブルの効果が最も大きいのはアダトロン学習である．もともとアダトロン学習は三つの学習則の中でもっとも高速な漸近特性を示す [7] という長所を持つが，学習の初期段階では遅い（$t < 6$ 程度では他の二つより汎化誤差が大きい）という短所も持つ．今回，アダトロン学習が「生徒の多様性維持」という点でアンサンブル学習との相性の良さを有することが明らかにされ，学習初期段階での短所もアンサンブル学習と組み合わせることにより大きく改善できることが明らかになった．

ここまでは K 個の生徒が出力を統合する方法として単純多数決を考えてきた．出力を統合する方法は単純多数決以外にも考えられる．結合荷重の平均 $\bar{\boldsymbol{J}} = \sum_{k=1}^K \boldsymbol{J}_k$ を結合荷重とするような新たな生徒を考え，その出力を統合出力とする方法，規格化内部状態の絶対値 $|u|$ が最大である生徒の出力を統合出力とする方法については式 (11) の代わりにそれぞれ式 (46)，式 (47) を用いることで汎化誤差を計算できる．

$$\epsilon_{g(2)} = \int\prod_{k=1}^K du_k dv p(v,\{u_k\})\Theta\left(-\mathrm{sgn}(v)\mathrm{sgn}\left(\sum_{k=1}^K u_k\right)\right), \quad (46)$$

$$\epsilon_{g(3)} = \int\prod_{k=1}^K du_k dv p(v,\{u_k\})\Theta\left(-\mathrm{sgn}(v)\arg\max_{u_k}\{|u_k|\}\right). \quad (47)$$

アダトロン学習の場合の結果を図 8 に示す．結合荷重の平均，$|u|$ 最大，多数決の順にアンサンブルの効果が大きいことがわかる．また，$|u|$ 最大は結合荷重の平均

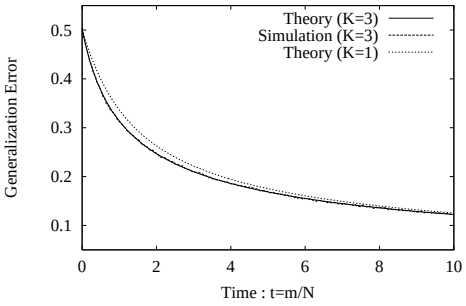

図 5: 汎化誤差（ヘブ学習）

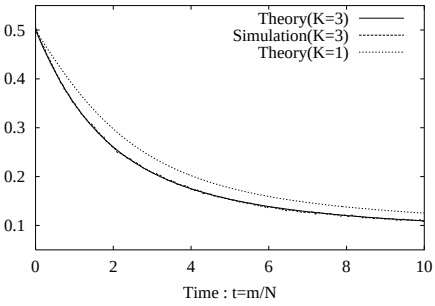

図 6: 汎化誤差（パーセプトロン学習）

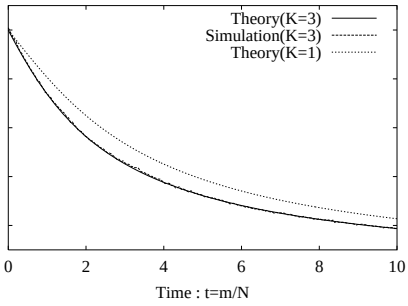

図 7: 汎化誤差（アダトロン学習）

にかなり近い性能を有することがわかる．これは「各生徒にアナログ値を出してもらって足すのがもちろん一番いいけれど，強い意見を主張している生徒の言うことだけを尊重するという方法もそれよりちょっと悪いだけで，多数決よりはずっといい」ということを表しており興味深い結果である．

さらに，$N = 1000$，$K = 1, 3, 11, 31$ で $t = 10000$ まで計算機実験を行った結果を図 9–図 11 に示す．これらより，アダトロン学習，パーセプトロン学習のアンサンブルの効果は漸近的にも維持されることがわかる．

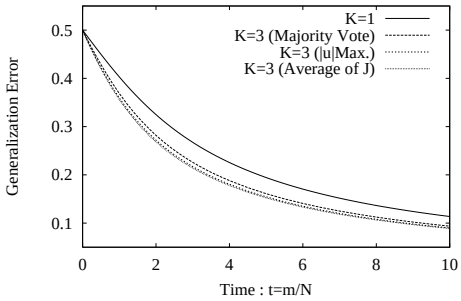

図 8: 統合方法による汎化誤差の違い（アダトロン学習）

6 むすび

符号関数を出力関数とする K 個の単純パーセプトロンによるアンサンブル学習をオンライン学習の枠組みで，かつ，有限の K で議論した．その結果，ヘブ学習，パーセプトロン学習，アダトロン学習の三つの学習則は「生徒の多様性維持」というアンサンブル学習との相性においてそれぞれ異なった性質を有しており，アダトロン学習がもっとも優れているという興味深い事実が明らかになった．

謝辞 本論文の一部は科学研究費補助金（課題番号 13780313, 14084212, 14580438, 15500151）によるものである．

参考文献

[1] 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).

[2] http://www.boosting.org/

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

[4] 原, 岡田, “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，” IBIS 予稿集，113–118 (2002).

[5] Krogh, A. and Sollich, P., “Statistical mechanics of ensemble learning,” Phys. Rev. E, **55**(1), 811 (1997).

[6] Urbanczik, R., “Online learning with ensembles,” Phys. Rev. E, **62**(1), 1448 (2000).

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

[8] Anlauf, J.K. and Biehl, M., “The AdaTron: an adaptive perceptron algorithm,” Europhys. Lett., **10**(7), 687 (1989).

[9] Biehl, M. and Riegler, P., “On-line learning with a perceptron,” Europhys. Lett., **28**(7), 525 (1994).

[10] Inoue, J. and Nishimori, H., “On-line AdaTron learning of a unlearnable rules,” Phys. Rev. E, **55**(4), 4544 (1997).

[11] Engel, A. and Broeck, C.V., “Statistical Mechanics of Learning,” Cambridge University Press, (2001)

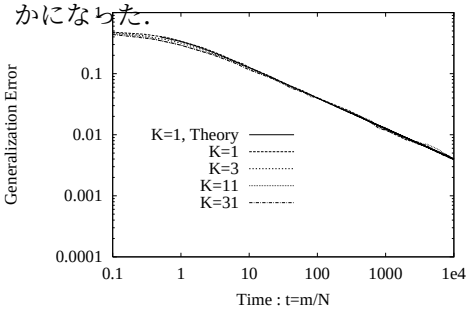

図 9: 漸近特性（ヘブ学習）

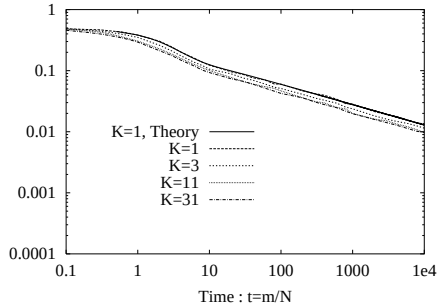

図 10: 漸近特性（パーセプトロン学習）

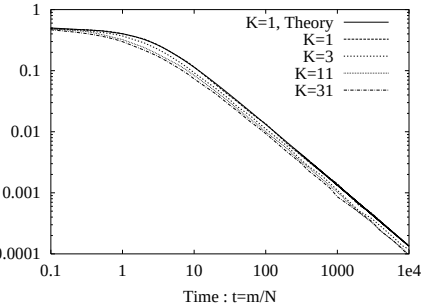

図 11: 漸近特性（アダトロン学習）