

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

三好 誠司[†] 原 一之^{††} 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

^{††} 東京都立工業高等専門学校 〒140-0011 東京都品川区東大井 1-10-40

^{†††} 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

^{††††} 科学技術振興事業団 さきがけ

E-mail: †miyoshi@kobe-kosen.ac.jp, ††hara@tokyo-tmct.ac.jp, †††okada@brain.riken.go.jp

あらまし 符号関数を出力関数とする K 個の非線形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議論する。まず、一般の学習則について、生徒間の類似度 q のダイナミクスを記述する微分方程式を導出する。次に、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について、この微分方程式を具体的に導出し、それらを数値的に解く。さらに、各生徒の出力の多数決により統合出力を決定する場合の汎化誤差 ϵ_g 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においてそれぞれ異なった性質を有しており、その観点で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る。

キーワード アンサンブル学習、オンライン学習、非線形パーセプトロン、パーセプトロン学習、ヘブ学習、アダトロン学習、汎化誤差

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

Seiji MIYOSHI[†], Kazuyuki HARA^{††}, and Masato OKADA^{†††,††††}

[†] Kobe City College of Technology 8-3 Gakuenhigashimachi, Nishi-ku, Kobe-shi, 651-2194 Japan

^{††} Tokyo Metropolitan College of Technology 1-10-40 Higashi-oi, Shinagawa, Tokyo, 140-0011 Japan

^{†††} RIKEN Brain Science Institute 2-1 Hirosawa, Wako-shi, Saitama, 351-0198 Japan

^{††††} JST PRESTO

E-mail: †miyoshi@kobe-kosen.ac.jp, ††hara@tokyo-tmct.ac.jp, †††okada@brain.riken.go.jp

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.

Key words 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] ~ [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] ~ [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] . 彼らは、生徒が線形な場合には、各生徒の結合荷重の平均を結合荷重とする線形パーセプトロンの出力と各生徒の出力の平均が一致することに着目し、統計力学的手法を用いて解析を行い、生徒の数 K が大きい極限における汎化誤差は $K = 1$ のときの $1/2$ であり、 K を増やしていくとその値に $1/K$ で収束していくことを明らかにした .

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、アダルトン学習がよく知られている [7] . Krogh と Sollich は、学習過程でノイズが混入する場合について、非線形パーセプトロンによるアンサンブル学習をバッチ学習の枠組みで解析した [5] . その結果、生徒の数 K が大きい極限では学習セットのサイズを最適に選ぶことにより汎化能力を最適化できること、 K が有限の場合には、特に学習過程でノイズが混入する場合に各生徒に学習サンプルを分割して学習させることにより汎化能力を改善できることを示した . 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、生徒の数 K が大きい極限についてオンライン学習の枠組みで解析した . 彼は、ヘブ学習とパーセプトロン学習の両方を特殊ケースとして含むような一般化された学習則（彼はこれを、パーセプトロン学習のソフトバージョン、と呼んだ）を考え、汎化誤差の観点から議論を行った . その結果、通常はアンサンブルは単一生徒よりも優れた性能を有するが、パーセプトロン学習のソフトバージョンの枠組み内で最適化された学習則の場合には、アンサンブルのメリットはないことを示し、アンサンブル学習の限界に関する一考察を行った .

Urbanczik の研究は、非線形パーセプトロンのアンサンブル学習をオンライン学習の枠組みで議論した研究であるが、上述のように生徒の数 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、アダルト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 .

そこで本研究では、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、一般の学習則について、生徒間の類似度 q のダイナミクスを記述する微分方程式を導出する . 次に、よく知られているヘブ学習、パーセプトロン学習、アダルト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、それらを数値的に解く . さらに、 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 ϵ_g を q を用いて求める

式を示し、その値を数値的に求める . 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 . アダルトン学習が、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

本研究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 J_1, J_2, \dots, J_K とする . J_k と入力 x は N 次元ベクトルであり、 x の各要素は平均 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、 J_k の初期値 J_k^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るものとする . 各生徒の出力は $\text{sgn}(u_1 l), \text{sgn}(u_2 l), \dots, \text{sgn}(u_K l)$ である . ここで、

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

$$u_k l = J_k \cdot x, \quad (2)$$

である . l は生徒の長さであり、これは本研究で扱うモデルの巨視的変数のひとつであるが詳しくは後で述べる . 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 B とする . 本研究では B は不変とする . ここで、生徒の初期値同様に教師 B は N 次元ベクトルであり、 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るものとする . 教師の出力は $\text{sgn}(v)$ である . ここで、

$$v = B \cdot x, \quad (3)$$

である . v を教師の内部状態と呼ぶことにする .

いま、 $N \rightarrow \infty$ の熱力学的極限を考えると、

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

となる . 生徒の大きさ $|J_k|$ 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 とし、これを生徒の長さと呼ぶことにする . すなわち、

$$|J_k| = l\sqrt{N} \quad (5)$$

である . l は本研究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 x が同じ順序で与えられる . 個々の生徒は入力 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、アダルトン学習などがよく知られている [7] . 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 x とそれに対する教師の出力 $\text{sgn}(v)$ だけであるから、学習は一般に以下のように表せる .

$$J_k^{m+1} = J_k^m + f_k^m x^m, \quad (6)$$

$$f_k^m = f(\text{sgn}(v^m), u_k^m), \quad (7)$$

ここで, m は時刻ステップを表す.

3. 理 論

3.1 教師と生徒の類似度 R と生徒間の類似度 q

本研究はオンライン学習の枠組みで議論を行う. オンライン学習では, 入力 x が与えられると, 生徒は学習則 (6) にしたがって直ちに結合荷重 J_k を修正する. 学習に用いた入力 x は破棄され, 以降の修正では用いられない. この場合, 教師 B , 各生徒 J_k は新たな入力パターン x とは独立になる. このことから, 簡単な計算より教師の内部状態 v , 各生徒の規格化内部状態 u_k は平均 0, 分散 1 のガウス分布に従うことがわかる.

本研究における一番目の巨視的変数は, 前節で述べたように生徒の長さ l である. ここでは二番目の巨視的変数として, 以下の式で定義される教師と生徒の類似度 R を導入する.

$$R = \frac{B \cdot J}{|B||J|}. \quad (8)$$

教師 B と生徒 J に相関がないとき $R = 0$ であり, 両者の方向が同じときには $R = 1$ である. 簡単な計算より教師の内部状態 v と生徒の規格化内部状態 u_k の共分散が R であることがわかる.

本研究ではアンサンブル学習を考えるので, 三番目の巨視的変数として以下の式で定義される生徒間の類似度 q を導入する.

$$q = \frac{J_\alpha \cdot J_\beta}{|J_\alpha||J_\beta|}, \quad (9)$$

ここで, $\alpha \neq \beta$ である. 生徒 J_α と生徒 J_β に相関がないとき $q = 0$ であり, 両者の方向が同じときには $q = 1$ である. 簡単な計算より生徒 J_α , 生徒 J_β の規格化内部状態 u_α, u_β の共分散が q であることがわかる.

すでに述べたように, 本研究では教師 B , 生徒 J の初期値は独立な確率変数であるとし, さらに $N \rightarrow \infty$ の熱力学的極限を考えるので, 学習の更新式の対称性から, すべての時刻において, すべての生徒に関して R は等しく, 生徒 α, β のすべての組み合わせ (ただし $\alpha \neq \beta$) に関して q は等しい. よって以後も R, q に生徒識別の添字は付けない.

学習が進むにつれ, R と q は 0 から 1 に近づいていく. このとき, R と q は全く無関係な値をとるわけではない. たとえば, $R=1$ のときに $q=0$ とはなり得ないことは R と q の意味合いを考えても明らかである. このように R と q は相互に何らかの拘束関係にあるのであるが, R と比較して q 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. すなわち, 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. すでに, 線形パーセプトロンについては, この関係が非常に明快な形で定量的に明らかにされている [4] ので, 以下では非線形パーセプトロンについて三つの学習則に関する R と q の関係を解析的に調べることにする.

3.2 一般の学習則の場合

一般の学習則の l, R 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].

$$\frac{dl}{dt} = \langle fu \rangle + \frac{\langle f^2 \rangle}{2l}, \quad (10)$$

$$\frac{dR}{dt} = \frac{\langle fv \rangle - \langle fu \rangle R}{l} - \frac{R}{2l^2} \langle f^2 \rangle. \quad (11)$$

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

次に, 一般の学習則の q に関する微分方程式を導出する. いま, 生徒 J_α と生徒 J_β を考える. 本研究では, 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. この自己平均性を仮定し, $l^{m+1} \rightarrow l + dl$, $q^{m+1} \rightarrow q + dq$, $1/N \rightarrow dt$ とおくと式 (6), (9) より以下の式が得られる.

$$\frac{dq}{dt} = \frac{\langle f_\alpha u_\beta \rangle + \langle f_\beta u_\alpha \rangle}{l} + \frac{\langle f_\alpha f_\beta \rangle}{l^2} - \frac{2q}{l} \frac{dl}{dt}. \quad (12)$$

式 (10) を式 (12) に代入し, さらに, 生徒の対称性より $\langle f_\alpha u_\beta \rangle = \langle f_\beta u_\alpha \rangle$ であることを考慮すると, q に関する微分方程式が以下のように得られる.

$$\frac{dq}{dt} = \frac{2(\langle f_\alpha u_\beta \rangle - q \langle fu \rangle)}{l} + \frac{\langle f_\alpha f_\beta \rangle - q \langle f^2 \rangle}{l^2}, \quad (13)$$

ここで, $\langle f_\alpha u_\beta \rangle$ は異なる生徒の f と u の積のサンプル平均, $\langle f_\alpha f_\beta \rangle$ は異なる生徒の f と f の積のサンプル平均, $\langle fu \rangle$ は同一の生徒の f と u の積のサンプル平均を表す.

3.3 汎化誤差

K 個の非線形単純パーセプトロンを統合しアンサンブルとしての出力を決定するための方法として, ここでは多数決を考えることにする. このときの汎化誤差 ϵ_g は以下のように計算される.

$$\epsilon_g = \int \prod_{k=1}^K du_k dv P(\mathbf{u}) \Theta \left(-\text{sgn}(v) \sum_{k=1}^K \text{sgn}(u_k) \right), \quad (14)$$

$$P(\mathbf{u}) = \frac{1}{(2\pi)^{\frac{K+1}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left(-\frac{\mathbf{u}^T \Sigma \mathbf{u}}{2} \right), \quad (15)$$

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_K \\ v \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & q & \dots & q & R \\ q & 1 & \dots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & q \\ q & \dots & q & 1 & R \\ R & \dots & \dots & R & 1 \end{pmatrix}, \quad (16)$$

ここで, Σ は \mathbf{u} の分散共分散行列で $(K+1)$ 行 $(K+1)$ 列の正方行列である.

このように, R と q がわかれば汎化誤差 ϵ_g が計算できる. よって, 次節では代表的な三つの学習則について式 (10), (11), (13) を計算するために必要な五つのサンプル平均 $\langle fu \rangle$, $\langle fv \rangle$, $\langle f^2 \rangle$, $\langle f_\alpha u_\beta \rangle$, $\langle f_\alpha f_\beta \rangle$ を具体的に計算する.

4. ダイナミクス

4.1 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.

$$f(\text{sgn}(v), u) = \text{sgn}(v) \quad (17)$$

ヘブ学習の $\langle fu \rangle$, $\langle fv \rangle$, $\langle f^2 \rangle$ はすでに求められており、以下の通りである [7].

$$\langle fu \rangle = \frac{2R}{\sqrt{2\pi}}, \quad \langle fv \rangle = \sqrt{\frac{2}{\pi}}, \quad \langle f^2 \rangle = 1. \quad (18)$$

以下では新たに $\langle f_\alpha u_\beta \rangle$ と $\langle f_\alpha f_\beta \rangle$ を導出する. 式 (17) は u に依存しない形なので, ヘブ学習においては $f_\alpha = f_\beta = f$ である. よって,

$$\langle f_\alpha u_\beta \rangle = \langle \text{sgn}(v) u_\beta \rangle = \langle fu \rangle = \frac{2R}{\sqrt{2\pi}}, \quad (19)$$

$$\langle f_\alpha f_\beta \rangle = \langle (\text{sgn}(v))^2 \rangle = 1. \quad (20)$$

式 (10), (11), (13), (18)–(20) 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. 理論と計算機実験の結果はほとんど重なっており, 今回導いた理論が計算機実験の結果をよく説明することがわかる. また, 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.

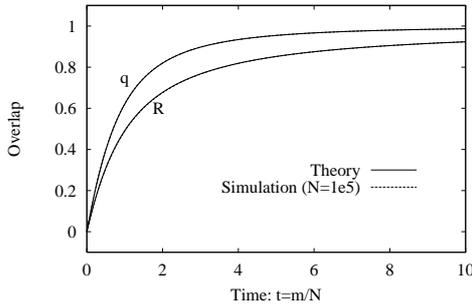


図 1 R と q のダイナミクス (ヘブ学習)

4.2 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.

$$f(\text{sgn}(v), u) = \Theta(-u \text{sgn}(v)) \text{sgn}(v) \quad (21)$$

ここで, $\Theta(\cdot)$ は以下のようなステップ関数である.

$$\Theta(z) = \begin{cases} +1, & z \geq 0, \\ 0, & z < 0. \end{cases} \quad (22)$$

パーセプトロン学習の $\langle fu \rangle$, $\langle fv \rangle$, $\langle f^2 \rangle$ はすでに求められており, 以下の通りである [7].

$$\langle fu \rangle = \frac{R-1}{\sqrt{2\pi}}, \quad \langle fv \rangle = \frac{1-R}{\sqrt{2\pi}}, \quad (23)$$

$$\langle f^2 \rangle = 2 \int_0^\infty Dv H\left(\frac{Rv}{\sqrt{1-R^2}}\right) = \frac{1}{\pi} \tan^{-1} \frac{\sqrt{1-R^2}}{R} \quad (24)$$

以下では新たに $\langle f_\alpha u_\beta \rangle$ と $\langle f_\alpha f_\beta \rangle$ を導出する. R と q の定義より, 確率変数 u_α, u_β, v は以下のガウス分布に従う. こ

で T は転置を表す.

$$P(u_\alpha, u_\beta, v) = \frac{1}{(2\pi)^{\frac{3}{2}} |\mathbf{\Lambda}|^{\frac{1}{2}}} \times \exp\left(-\frac{(u_\alpha, u_\beta, v) \mathbf{\Lambda}^{-1} (u_\alpha, u_\beta, v)^T}{2}\right) \quad (25)$$

$$\mathbf{\Lambda} = \begin{pmatrix} 1 & q & R \\ q & 1 & R \\ R & R & 1 \end{pmatrix} \quad (26)$$

これを用いてガウス積分を解析的に実行することにより, $\langle f_\alpha u_\beta \rangle$ と $\langle f_\alpha f_\beta \rangle$ を以下のように導出することができる. なお, すでに述べたように, 式 (13) の導出においては, 生徒の対称性より $\langle f_\alpha u_\beta \rangle = \langle f_\beta u_\alpha \rangle$ が成り立つことを用いているので, ここでも $\langle f_\beta u_\alpha \rangle$ を求める必要はなく, $\langle f_\alpha u_\beta \rangle$ を求めれば十分である.

$$\langle f_\alpha u_\beta \rangle = \langle \Theta(-u_\alpha v) \text{sgn}(v) u_\beta \rangle \quad (27)$$

$$= \int du_\alpha du_\beta dv P(u_\alpha, u_\beta, v) \Theta(-u_\alpha v) \text{sgn}(v) u_\beta \quad (28)$$

$$= \frac{R-q}{\sqrt{2\pi}} \quad (29)$$

$$\langle f_\alpha f_\beta \rangle = \langle \Theta(-u_\alpha v) \text{sgn}(v) \Theta(-u_\beta v) \text{sgn}(v) \rangle \quad (30)$$

$$= \langle \Theta(-u_\alpha v) \Theta(-u_\beta v) \rangle \quad (31)$$

$$= 2 \int_0^\infty Dv \int_{\frac{Rv}{\sqrt{1-R^2}}}^\infty Dx H(z) \quad (32)$$

ここで,

$$z \equiv \frac{-(q-R^2)x + R\sqrt{1-R^2}v}{\sqrt{(1-q)(1+q-2R^2)}} \quad (33)$$

であり, $H(u)$ と Dx の定義は以下の通りである.

$$H(u) \equiv \int_u^\infty Dx, \quad Dx \equiv \frac{dx}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right). \quad (34)$$

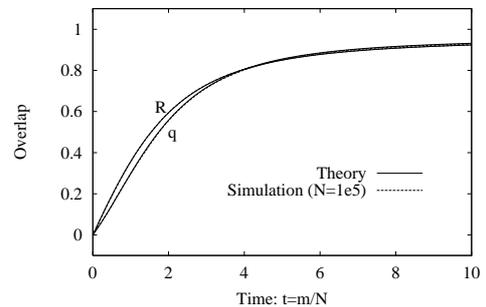


図 2 R と q のダイナミクス (パーセプトロン学習)

式 (10), (11), (13), (23), (24), (29), (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 2 に示す. 理論と計算機実験の結果はほとんど重なっており, 今回導いた理論が計算機実験の結果をよく説明することがわかる. また, パーセプトロン学習ではヘブ学習と異なり, 学習の初期 ($t < 4$) 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維持されていると言える。

4.3 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。

$$f(\text{sgn}(v), u) = -u\Theta(-u \text{sgn}(v)) \quad (35)$$

アダトロン学習の $\langle fu \rangle$, $\langle fv \rangle$, $\langle f^2 \rangle$ はすでに求められており、以下の通りである [7]。

$$\langle fu \rangle = -2 \int_0^\infty Du u^2 H\left(\frac{Ru}{\sqrt{1-R^2}}\right) \quad (36)$$

$$= -\frac{1}{\pi} \cot^{-1}\left(\frac{R}{\sqrt{1-R^2}}\right) + \frac{1}{\pi} R \sqrt{1-R^2} \quad (37)$$

$$\langle fv \rangle = \frac{(1-R^2)^{\frac{3}{2}}}{\pi} + R \langle fu \rangle, \quad \langle f^2 \rangle = -\langle fu \rangle \quad (38)$$

以下では新たに $\langle f_\alpha u_\beta \rangle$ と $\langle f_\alpha f_\beta \rangle$ を導出する。確率変数 u_α , u_β , v は式 (25), (26) に従う。これを用いてガウス積分を解析的に実行することにより $\langle f_\alpha u_\beta \rangle$ と $\langle f_\alpha f_\beta \rangle$ を以下のように導出することができる。

$$\langle f_\alpha u_\beta \rangle = \langle -u_\alpha \Theta(-u_\alpha v) u_\beta \rangle \quad (39)$$

$$= - \int du_\alpha du_\beta dv P(u_\alpha, u_\beta, v) \Theta(-u_\alpha v) u_\alpha u_\beta \quad (40)$$

$$= \frac{1+q}{\pi} R \sqrt{1-R^2} - 2q \int_0^\infty Dv \int_{\frac{Rv}{\sqrt{1-R^2}}}^\infty Dx x^2 \quad (41)$$

$$\langle f_\alpha f_\beta \rangle = \langle u_\alpha u_\beta \Theta(-u_\alpha v) \Theta(-u_\beta v) \rangle \quad (42)$$

$$= \int dv du_\alpha u_\alpha du_\beta u_\beta P(u_\alpha, u_\beta, v) \Theta(-u_\alpha v) \Theta(-u_\beta v) \quad (43)$$

$$= \frac{(1-q)^2 (1+q-2R^2)}{2\pi (1-R^2)^{\frac{3}{2}}} \left(\sqrt{\frac{(1+q)(1-R^2)}{1-q}} - R \right)$$

$$+ 2(q-R^2) \int_0^\infty Dv \int_{\frac{Rv}{\sqrt{1-R^2}}}^\infty Dx x^2 H(z)$$

$$- \frac{2R(1+q-R^2)}{\sqrt{1-R^2}} \int_0^\infty Dvv \int_{\frac{Rv}{\sqrt{1-R^2}}}^\infty Dx x H(z)$$

$$+ 2R^2 \int_0^\infty Dvv^2 \int_{\frac{Rv}{\sqrt{1-R^2}}}^\infty Dx H(z) \quad (44)$$

ここで、 z , $H(u)$, Dx の定義は式 (33), (34) である。

式 (10), (11), (13), (37), (38), (41), (44) 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。理論と計算機実験の結果はほとんど重なっており、今回導いた理論が計算機実験の結果をよく説明することがわかる。また、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。

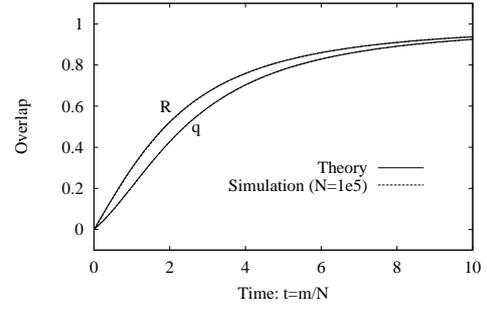


図 3 R と q のダイナミクス (アダトロン学習)

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。よってこの段階で、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。

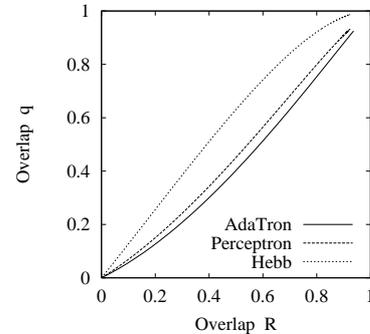


図 4 R と q の関係 (理論)

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 (14) ~ 式 (16) 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。理論計算における数値積分にはニュートン・コーツの 6 点公式を用いた。また、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。なお、各図においてはアンサンブル学習の効果を明らかにするための比較対象として $K = 1$ の場合の理論計算結果もプロットしている。

これら三つの図より、まず、今回理論計算により求めた汎化誤差が計算機実験の結果をよく説明することがわかる。また、 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、

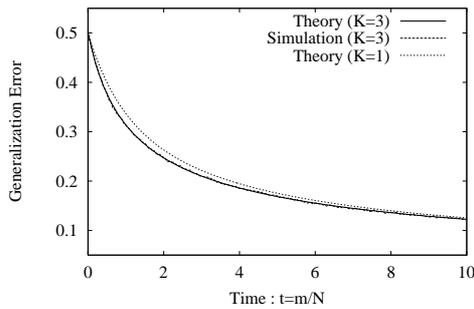


図 5 汎化誤差 (ヘブ学習, $K = 3$)

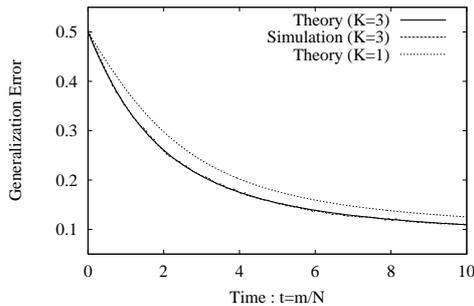


図 6 汎化誤差 (パーセプトロン学習, $K = 3$)

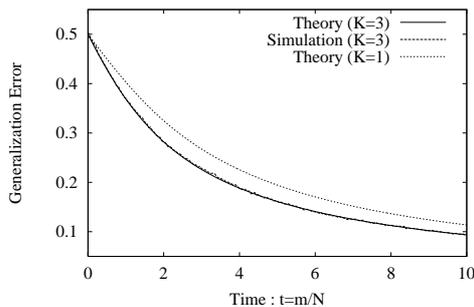


図 7 汎化誤差 (アダトロン学習, $K = 3$)

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) という短所を持つ. 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.

ここまでは生徒の長さ l の初期値を 1 とした場合について述べてきたが, アダトロン学習で l の初期値を変化させた場合の理論計算結果を図 8, 図 9 に示す. 図 8 より, l 初期値を小さくすると R と比較した q が大きくなることわかる. また, 式 (10), (38) よりアダトロン学習では $l = 0.5$ は不動点であるが, 図 8 より, このときちょうど $q = R$ となることわかる. 図 9 より, l 初期値が小さいほど学習初期段階の汎化誤差が改善されることがわかる. この理由は l 初期値が小さいと相対的に更新量が大きくなるためであると理解することができる. ところが, 上で述べたように l 初期値を小さくすると R と比較した q が大きくなるので, アンサンブル学習の効果は小さくなってし

まう. 図 9 から l 初期値が小さいと漸近的な汎化誤差は大きくなってしまふことがわかる. 逆に言うと, l 初期値を大きくすることにより汎化誤差の漸近特性を改善できる可能性がある.

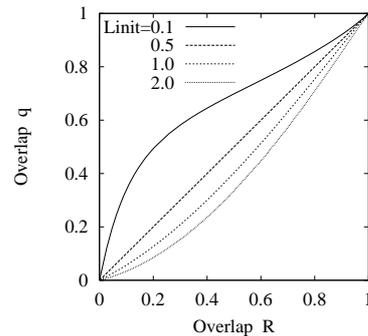


図 8 l 初期値による $R - q$ の変化 (アダトロン学習, 理論)

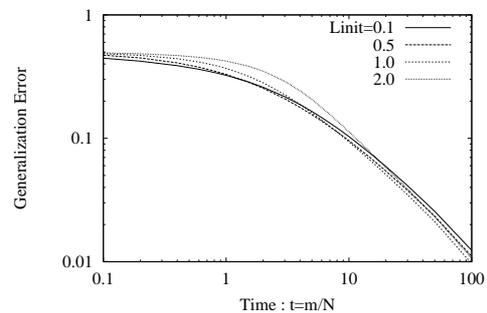


図 9 l 初期値による汎化誤差の変化 (アダトロン学習, $K = 3$, 理論)

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. また, アダトロン学習において, 生徒の長さ l の初期値が学習のダイナミクスに与える影響について興味深い結果が得られたので, これを解析することが今後の課題である.

謝 辞

本研究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.

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, 情報論的学習理論ワークショップ予稿集, 113 (2002).
- [5] Krogh, A. and Sollich, P., Phys. Rev. E, 55(1), 811 (1997).
- [6] Urbanczik, R., 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.