情報科学講座 A·2·5

情報理論 II

—情報の幾何学的理論—

北川敏男編

共立出版株式会社
1968

執筆者

甘利俊一　東京大学工学部

第5章　学習識別の理論

5·1 学習識別系

パターン認識系において，パターンの特徴信号の分布が知られている場合はむしろまれである。分布がわかっていなければ，前章で述べた方法で最適な決定方法を定めることはできない。また，たとえ分布がわかったとしても，それを用いて最適な決定方法を定めることは，実は容易でない。それに，パターンの分布は固定しているとは限らず，刻一刻と変動している場合も多い。このような状況を考えるとき，識別方法を計算で求めて固定しておくやり方は，あまり得策ではないことになる。これに代わるものとして，識別系に学習能力を付しておき，系にはいってくるパターン信号に基づいて学習を行なって，識別方法を逐次定めていく方法が考えられる。これが**学習識別** (learning pattern-classification) の考えである。

学習識別には，おおざっぱにいって，次の二つの方法が考えられる。一つは過去のパターン信号のデータに基づいて，パターンの確率分布そのものを逐次求めていく方法であり，もう一つは，分布を媒介とせず，識別関数を直接求めていく方法である。

まず，前者について考えよう。過去に現われた，各クラスのパターン $\boldsymbol{\xi}$ をすべて記憶しておき，そのヒストグラムをつくれば，確率分布 p_α や $p_\alpha(\boldsymbol{\xi})$ が近似的に求まる。p_α や $p_\alpha(\boldsymbol{\xi})$ が求まれば，計算によって最適な識別ベクトルが求まることになる。しかし，過去に生じたパターンをすべて記憶しておくには，ばくだいな容量の記憶装置が必要であるし，またその後の計算も大変なものである。そこで，通常は，確率分布の形を適当に仮定して，分布のパラメータを推定する方法が用いられる。

たとえば，分布 $p_\alpha(\boldsymbol{\xi})$ は正規分布であると仮定しよう。すると，分布 $p_\alpha(\boldsymbol{\xi})$

は C_α に属する $\boldsymbol{\xi}$ の平均値 $\bar{\boldsymbol{\xi}}_\alpha$ と共分散行列 Σ_α とによって

$$p_\alpha(\boldsymbol{\xi})=\frac{1}{\sqrt{(2\pi)^n \det|\Sigma_\alpha|}}\exp\{-\frac{1}{2}(\boldsymbol{\xi}-\bar{\boldsymbol{\xi}}_\alpha)_t\Sigma_\alpha^{-1}(\boldsymbol{\xi}-\bar{\boldsymbol{\xi}}_\alpha)\} \qquad (5\cdot1)^*$$

と表わせる。それゆえ，過去のデータから正規分布のパラメータ $\bar{\boldsymbol{\xi}}_\alpha$ と Σ_α を求めれば，分布が求まる。この方法は，分布のパラメータを推定することから，**パラメトリックな方法**とよばれる。

例として，$\bar{\boldsymbol{\xi}}_\alpha$ を求めよう。過去に C_α に属するパターンが k 回出たとし，それに基づいた $\bar{\boldsymbol{\xi}}_\alpha$ の推定値を $\bar{\boldsymbol{\xi}}_\alpha{}^k$ と書く。いままた，C_α のパターン $\boldsymbol{\xi}$ が出たとすると，これを修正して

$$\bar{\boldsymbol{\xi}}_\alpha{}^{k+1}=\frac{k\bar{\boldsymbol{\xi}}_\alpha{}^k+\boldsymbol{\xi}}{k+1} \qquad (5\cdot2)$$

とすれば，$\bar{\boldsymbol{\xi}}_\alpha$ の値が逐次求まっていくことになる。

分布自体を求めていくならば，系についての正確な知識が得られようが，この方法には次のような欠点も考えられる。第一に，分布自体を推定するとなると，非常にたくさんのデータをたくわえねばならない。これを避けるためには，分布の形を事前に知らなければならない。第二に，分布がわかったとしても，それに基づいて識別ベクトルを求めることが容易でない。第三に，パターンの確率分布などの系の構造が変動しつつあるときには，この方法では変動に追従しにくいきらいがある。

したがって，この方法は系の構造が変動しない場合に，過去のデータより系に対する精密な知識を得たいときに用いられる。そして，それにより最適な識別ベクトルを求め，これを固定して使用する。系の構造が刻々と不規則に変化していく場合や，識別ベクトルを学習により直接求めたい場合は，この方法は不適当である。

そこで，分布を媒介とせず，識別ベクトルを学習により直接求める方法を考

---

* 行ベクトルと列ベクトルとを区別する必要があるときは $\boldsymbol{\xi}$, $\boldsymbol{\theta}$ は，列ベクトルを表わすものとし，添字 t は転置を意味する．

える。識別ベクトル Θ を，現在系にはいってくるパターン ξ とそれがどのクラスに属するかという情報に基づいて，逐次修正していく方法である。これは，分布の形を知る必要がなく，分布のパラメータの推定が不要であるため，**ノンパラメトリックな学習**といわれる。

ここでは，ノンパラメトリックな学習法を追求することとし，前者の分布の推定を含む学習法は，通常の統計学や他の学習識別の教科書[45]にゆずる。

ノンパラメトリックな方法として，いわゆる**パーセプトロンの学習法**[46]が有名である。しかし，これはパターンの分布が線形分離可能なときにのみ収束する方法であり，線形識別関数を用いた場合に限られている。ここでは，パターンの分布が重なっていてもよく，また一般の識別関数について成立する学習法である，**確率的降下法**について述べる。後節で，学習の収束，収束速度，収束精度，確率構造の変化に対する追従特性，などが明らかにされる。

確率的降下法の問題点として，次の点があげられる。まず，この方法が使えるのは，連続損失の場合に限られる。したがって，連続でない場合に用いたければ，損失を連続な関数で近似しなければならない*。第二に，確率的降下法による識別ベクトルは，$L(\Theta)$ の極小値に収束するが，これが必ずしも最小値とは限らない。したがって，最適な識別ベクトルを得るには，適当な初期ベクトルから出発しなければならないことがある**。

以上の点を考えると，この方法は，最適な識別ベクトルのだいたいの値はわかっているが，系の構造が変動しているために，細かい所までは定められないようなときに，特に有効といえる。音声認識装置は，特定の発声者に合わせて識別方法を固定したのでは，発声者が変わった場合にうまく働かない。識別装置に学習機能をもたせて，発声者の性質に合わせて，識別ベクトルを自動的に調整する場合などが，この例である。そのほか，プラントなどで，原料その他の状況の変動に応じて，制御方法を自動調整する場合なども，この例として考

* 近似を良くすると，必要な情報が境界 $B_{\beta\gamma}$ 上に局在するため，収束の速度が落ちる．4·3D項参照．

** 線形分離可能なパターン分布に対しては，$L(\Theta)$ の極小値はただ一つであることが証明できる．

えられる。

5·2　確率的降下法による学習

A．確率的降下法

識別ベクトル Θ を，与えられたパターン ξ に基づいて逐次修正していく識別系を考える。新しく得られる識別ベクトルを Θ' と書くと

$$\Theta'=\Theta+\delta\Theta \tag{5·3}$$

であり，修正項 $\delta\Theta$ は ξ および ξ の属するクラスに依存する。

C_α のパターン ξ が提示されたときの修正項を

$$\delta\Theta=\delta\Theta(\xi,\ \alpha,\ \Theta) \tag{5·4}$$

と書く。修正項を，$L(\Theta)$ を減少させるように選びたい。しかし，パターンの確率分布が未知であるために，$L(\Theta)$ も未知であり，$L(\Theta)$ を減少させる方向はわからない。そこで，次善の策として，$L(\Theta)$ を平均として減少させるように修正項を選ぶことを考える。「平均」とは，系にはいってくるあらゆるパターン ξ についての未知の分布 $p_\alpha, p_\alpha(\xi)$ に基づいた期待値の意味である。

修正項の期待値は

$$\overline{\delta\Theta}=\sum_\alpha\int p_\alpha p_\alpha(\xi)\delta\Theta(\xi,\alpha,\Theta)d\xi \tag{5·5}$$

と表わせる。$\delta\Theta$ は微小と考えると，1回の修正による $L(\Theta)$ の変化は

$$\delta L(\Theta)=\delta\Theta\cdot\nabla L(\Theta) \tag{5·6}$$

と書ける。$\delta L(\Theta)$ の期待値は

$$\overline{\delta L(\Theta)}=\overline{\delta\Theta}\cdot\nabla L(\Theta)$$

であるから，これを負にするには，∇L と $\overline{\delta\Theta}$ とが鈍角をなすようにする必要がある。このためには，ε を正の定数，Cを正の定符号をもつ行列* として

$$\overline{\delta\Theta}=-\varepsilon C\nabla L(\Theta) \tag{5·7}$$

* 任意のベクトル $a\neq0$ に対して，常に $a_tCa>0$ が成立する行列を，正の定符号をもつ行列という．

ならばよい。関数

$$\boldsymbol{a}_\alpha(\boldsymbol{\xi},\Theta)=-\nabla l_\alpha(\boldsymbol{\xi},\Theta) \tag{5·8}$$

の期待値は，(4·88) からわかるように

$$\overline{\boldsymbol{a}_\alpha(\boldsymbol{\xi},\Theta)}=\sum_\alpha\int p_\alpha p_\alpha(\boldsymbol{\xi})\boldsymbol{a}_\alpha(\boldsymbol{\xi},\Theta)d\boldsymbol{\xi}=-\nabla L(\Theta) \tag{5·9}$$

であるから*，

$$\delta\Theta(\boldsymbol{\xi},\alpha,\Theta)=\varepsilon C\boldsymbol{a}_\alpha(\boldsymbol{\xi},\Theta) \tag{5·10}$$

とおけば，(5·7) が成立する。$\boldsymbol{a}_\alpha(\boldsymbol{\xi},\Theta)$ を**学習関数**と名づけ，この学習法を**確率的降下法**という**。

この学習法では，より好ましい識別ベクトルが次々に得られる，というわけではない。修正の結果，識別ベクトルがまえよりも悪くなる場合も生ずる。しかし，(5·7) に示されるように，修正ベクトルの平均方向はより良くなる方向を向いている。

この方法を，$L(\Theta)$ の最小点を求めるのによく用いる，**最急降下法**と比較してみよう。最急降下法では，$L(\Theta)$ や $\nabla L(\Theta)$ がわかっているために，$L(\Theta)$ を減少させる方向，すなわち山を下る方向がわかる。したがって，$L(\Theta)$ が必ず減るように修正を行なうことができる。これに反して学習問題においてはパターン分布が未知なために，$L(\Theta)$ や $\nabla L(\Theta)$ がわからず，山を下る方向がわからない。しかし，識別装置にはいってくるパターンは，未知のある確率分布に従って発生している。この情報を利用するために，パターンの関数 $\boldsymbol{a}_\alpha(\boldsymbol{\xi},\Theta)$ で，その平均値が山を下る方向を向いているものをうまく探し出して，その方向へ確率的に下っていこうとするものである。

これは，坂の途中に立たされた「よっぱらい」が酔歩を行なう状況に似ている。彼は，あるときは山を登る方向によろめき，また次には山を下る方向によろける。登る確率よりは下る確率のほうが大きいため，彼は平均としては山をずり落ちて行き，ついには谷底へ落ち込む。そして谷底の近傍で，よろめき（微小振動）を続けるであろう。

確率的降下法により学習を行なわせると，同様の状況で，識別ベクトルが最適値の近傍に落ち着くことが予想される。収束の証明や，収束の精度（微小振動の状況），収束の速度などは後に示される。

B．　種々の識別関数に対する学習法

a．　一次識別関数　初めに二分割の場合について述べる。この場合，識別関数は　$g(\boldsymbol{\xi})=g_1(\boldsymbol{\xi})-g_2(\boldsymbol{\xi})$ を用いると

$$g(\boldsymbol{\xi},\Theta)=\Theta\cdot\hat{\boldsymbol{\xi}}$$

と書けて，g の正負に応じてパターンは C_1 もしくは C_2 と決定される。損失を

$$\left.\begin{aligned} l_1(\boldsymbol{\xi},\Theta)&=l(-g)\\ l_2(\boldsymbol{\xi},\Theta)&=l(g)\end{aligned}\right\} \tag{5·11}$$

とおこう*。学習関数は

$$\left.\begin{aligned} \boldsymbol{a}_1(\boldsymbol{\xi},\Theta)&=-\nabla l_1=l'(-g)\hat{\boldsymbol{\xi}}\\ \boldsymbol{a}_2(\boldsymbol{\xi},\Theta)&=-\nabla l_2=-l'(g)\hat{\boldsymbol{\xi}}\end{aligned}\right\} \tag{5·12}$$

となるから，次の学習規則が得られる。

$\delta\Theta=\varepsilon l'(-g)C\hat{\boldsymbol{\xi}}$,　　$\boldsymbol{\xi}\in C_1$ のパターンを識別したとき

$\delta\Theta=-\varepsilon l'(g)C\hat{\boldsymbol{\xi}}$,　　$\boldsymbol{\xi}\in C_2$ のパターンを識別したとき

特別な場合として，

$$\left.\begin{aligned} l(d)&=\overset{d}{0}, \quad d\geqq 0\\ C&=\mathrm{E}\ (\text{単位行列})\end{aligned}\right\} \tag{5·13}$$

を考えると

$\delta\Theta=\pm\varepsilon\hat{\boldsymbol{\xi}}$，$\boldsymbol{\xi}$ を誤識別したとき（＋は C_1，－は C_2 のパターン）となる。

---

* 本章では連続損失を考える．

** これは，確率的近似法（stochastic approximation method）の応用と考えられる[47,48]．

* これは (4·89)，(4·90)，(4·91) で，$g_1\equiv g$，$g_2\equiv 0$ とし，$s_\beta=1$ とおいたものである．

これは，パーセプトロン学習法として知られるものである[46]。

損失（5·11）の幾何学的意味はあまり明確でない。損失として，識別面Bから誤識別されるパターン $\boldsymbol{\xi}$ までの距離 d を用いたほうが意味が明確である。パターン $\boldsymbol{\xi}$ から B までの距離は

$$d=\frac{|g|}{\theta} \tag{5·14}$$

である。ただし

$$\theta=\sqrt{\sum_{i=1}^{n}\Theta_i^2} \tag{5·15}$$

これを用いて，損失を

$$l_1=l\left(\frac{-g}{\theta}\right),\quad l_2=l\left(\frac{g}{\theta}\right) \tag{5·16}$$

とおく。これは（4·90）で $s_\beta=1/\theta$ とおいたものにほかならない。$\nabla(g/\theta)$ を計算すると，

$$\nabla\frac{g}{\theta}=\mathrm{T}\hat{\boldsymbol{\xi}} \tag{5·17}$$

が得られる。ただし，Tは行列で

$$\mathrm{T}(\boldsymbol{\xi},\boldsymbol{\Theta})=\frac{1}{\theta^3}(\theta^2E-\check{\boldsymbol{\Theta}}_t\boldsymbol{\Theta}) \tag{5·18}$$

$\check{\boldsymbol{\Theta}}$ は $\boldsymbol{\Theta}$ の第 $n+1$ 成分を0とおいたもの

$$\check{\Theta}=(\Theta_1,\Theta_2,\cdots,\Theta_n,0)$$

である。

これより次の学習規則を得る。

$\delta\boldsymbol{\Theta}=\pm\varepsilon l'(|g|/\theta)\mathrm{C}\mathrm{T}\hat{\boldsymbol{\xi}}$，$\boldsymbol{\xi}$ を誤識別したとき（＋は C_1，－は C_2 のパターン）

パターンクラスの数が多数個ある場合にも，同様のやり方で学習規則が得られる。

識別関数を　$g_\alpha(\boldsymbol{\xi},\boldsymbol{\Theta})=\boldsymbol{\theta}_\alpha\cdot\hat{\boldsymbol{\xi}},\ \alpha=1,2,\cdots,k$

損失を

$$l_\alpha(\boldsymbol{\xi},\boldsymbol{\Theta})=l(d_\alpha)$$

$$d_\alpha=\sum_{\beta\in N_\alpha}\frac{1}{n_\alpha}(g_\beta-g_\alpha)$$

とする*。ここに，n_α は N_α に含まれているβの個数である。$\boldsymbol{\Theta}$ は k 個のベクトル $\boldsymbol{\theta}_\alpha$ から成っているので，各成分ベクトルに分けて考えよう。すると，次の学習規則が得られる**。

一次識別関数の学習規則：$\boldsymbol{\xi}\in C_\alpha$ なるパターンが誤識別されたときは

$$\begin{cases}\delta\boldsymbol{\theta}_\alpha=\varepsilon l'(d_\alpha)\mathrm{C}\hat{\boldsymbol{\xi}}\\ \delta\boldsymbol{\theta}_\beta=-\dfrac{\varepsilon}{n_\alpha}l'(d_\alpha)\mathrm{C}\hat{\boldsymbol{\xi}},\quad \beta\in N_\alpha\\ \delta\boldsymbol{\theta}_\gamma=0,\quad \gamma\neq\alpha,\quad \gamma\notin N_\alpha\end{cases}$$

とする。

b．線形識別関数

線形識別関数

$$g_\alpha(\boldsymbol{\xi},\boldsymbol{\Theta})=\boldsymbol{\theta}_\alpha\cdot\boldsymbol{\varphi}(\boldsymbol{\xi})$$

の場合には，一次識別関数と同様の取り扱いができる[42,43]。

線形識別関数の学習規則：C_α に属するパターン $\boldsymbol{\xi}$ が誤識別されたときは

$$\begin{cases}\delta\boldsymbol{\Theta}_\alpha=\varepsilon l'(d_\alpha)\mathrm{C}\boldsymbol{\varphi}\\ \delta\boldsymbol{\Theta}_\beta=-\varepsilon\dfrac{1}{n_\alpha}l'(d_\alpha)\mathrm{C}\boldsymbol{\varphi},\quad \beta\in N_\alpha\\ \delta\boldsymbol{\theta}_\gamma=0,\quad \gamma\neq\alpha,\quad \gamma\notin N_\alpha\end{cases}$$

---

* $s_\beta=\dfrac{1}{n_\alpha}$ とおいた．　$s_\beta=\begin{cases}1,\ \max\limits_\gamma g_\gamma=g_\beta\ \text{のとき}\\ 0,\ \text{その他のとき}\end{cases}$ とおくこともできる．

** 損失を識別面からの距離の関数にとるときは，二分割の場合に行なったと同様の修正を加えればよい．

とする。

c．区分的線形識別関数

簡単のため，二分割の場合のみを取り扱う。この場合，区分的線形識別関数は

$$g(\boldsymbol{\xi},\Theta)=\max_{i=1,\cdots,p}\boldsymbol{\theta}_1^{(i)}\cdot\hat{\boldsymbol{\xi}}+\min_{j=1,2,\cdots,q}\boldsymbol{\theta}_2^{(j)}\cdot\hat{\boldsymbol{\xi}} \qquad (5\cdot19)$$

と書ける。Θ は $p+q$ 個のベクトル，$\boldsymbol{\theta}_1^{(i)}$, $\boldsymbol{\theta}_2^{(j)}$ より成る。p と q とは異なった数でよく，どちらか一方が0であってもよい。損失として

$$l_1(\boldsymbol{\xi},\Theta)=l(-g)$$
$$l_2(\boldsymbol{\xi},\Theta)=l(g)$$

をとる。

Θ に対応して，学習関数 $\boldsymbol{\alpha}_\alpha(\boldsymbol{\xi},\Theta)$ も，各成分ベクトルに分けて考えよう。$\boldsymbol{\theta}_\alpha{}^i$ に対応する学習関数を $\boldsymbol{\alpha}_\alpha{}^{(i)}(\boldsymbol{\xi},\Theta)$ と書き，識別ベクトルの修正を

$$\delta\boldsymbol{\theta}_\alpha{}^{(i)}=\varepsilon C\boldsymbol{\alpha}_\alpha{}^{(i)} \qquad (5\cdot20)$$

とする。すると，C_β のパターンに対する学習関数は

$$\boldsymbol{\alpha}_\alpha{}^{(i)}=-\nabla_\alpha{}^{(i)}l_\beta(\boldsymbol{\xi},\Theta) \qquad (5\cdot21)$$

である。ただし

$$\nabla_\alpha{}^{(i)}=\frac{\partial}{\partial\boldsymbol{\theta}_\alpha{}^{(i)}}$$

さて，$\boldsymbol{\xi}$ の領域 $W_1{}^{(i)}$, $W_2{}^{(j)}$ をそれぞれ

$$\left.\begin{aligned}W_1{}^{(i)}&=\{\boldsymbol{\xi}\quad|\max_k\boldsymbol{\theta}_1{}^{(k)}\cdot\hat{\boldsymbol{\xi}}=\boldsymbol{\theta}_1{}^{(i)}\cdot\hat{\boldsymbol{\xi}}\}\\W_2{}^{(j)}&=\{\boldsymbol{\xi}\quad|\min_k\boldsymbol{\theta}_2{}^{(k)}\cdot\hat{\boldsymbol{\xi}}=\boldsymbol{\theta}_2{}^{(j)}\cdot\hat{\boldsymbol{\xi}}\}\end{aligned}\right\} \qquad (5\cdot22)$$

で定義し，

$$W_{ij}=W_1{}^{(i)}\cap W_2{}^{(j)}$$

とする。W_{ij} においては

$$g(\boldsymbol{\xi},\Theta)=(\boldsymbol{\theta}_1{}^{(i)}+\boldsymbol{\theta}_2{}^{(j)})\cdot\hat{\boldsymbol{\xi}}$$

である。これより $l(g)$ の微分を計算すると

$$\nabla_\alpha{}^{(i)}l(g)=\begin{cases}l'(g)\hat{\boldsymbol{\xi}}, & \boldsymbol{\xi}\in W_\alpha{}^{(i)}\\0, & \boldsymbol{\xi}\notin W_\alpha{}^{(i)}\end{cases}$$

である。したがって，次の学習規則を得る。

区分的線形識別関数の学習規則：誤識別されたパターン $\boldsymbol{\xi}$ が W_{ij} にはいっているときは

$$\begin{cases}\delta\boldsymbol{\theta}_1{}^{(i)}=\pm\varepsilon l'(|g|)C\hat{\boldsymbol{\xi}}\\\delta\boldsymbol{\theta}_2{}^{(j)}=\pm\varepsilon l'(|g|)C\hat{\boldsymbol{\xi}}\\\delta\boldsymbol{\theta}_1{}^{(k)}=\delta\boldsymbol{\theta}_2{}^{(k')}=0,\quad k\neq i,\ k'\neq j\end{cases}$$

（＋は C_1，－は C_2 のパターンに対して）

とする。

多分割の場合も同様にして学習規則が得られる。

区分的学習関数を用いた簡単な実験例を示す*。パターンの特徴ベクトル $\boldsymbol{\xi}$ は二次元で，C_1, C_2 のパターンはそれぞれ，図5·1（a）に示した W 字型の領域内で一様に分布している。パターンは計算機を用いて発生させる。識別関数は，4本の直線より成る区分的線形関数

$$g(\boldsymbol{\xi},\Theta)=\max(\boldsymbol{\theta}_1{}^{(1)}\cdot\hat{\boldsymbol{\xi}},\ \boldsymbol{\theta}_1{}^{(2)}\cdot\hat{\boldsymbol{\xi}})+\min(\boldsymbol{\theta}_2{}^{(1)}\cdot\boldsymbol{\xi},\ \boldsymbol{\theta}_2{}^{(2)}\cdot\hat{\boldsymbol{\xi}})$$

とし，図5·1（b）に示される識別領域より出発して学習を行なう。識別領域の修正は，パターンを誤識別したときにのみ行なう。1回の修正で，識別領域は図（c）のように変化し，以下2回めの修正で図（d）のようになる。25回の修正で図（e）のようになり，ほとんど正しい識別が行なえるようになる。

---

* 斉藤庄司氏（電々公社）の九州大学大学院工学研究科（通信工学）修士論文（昭和42年）による。

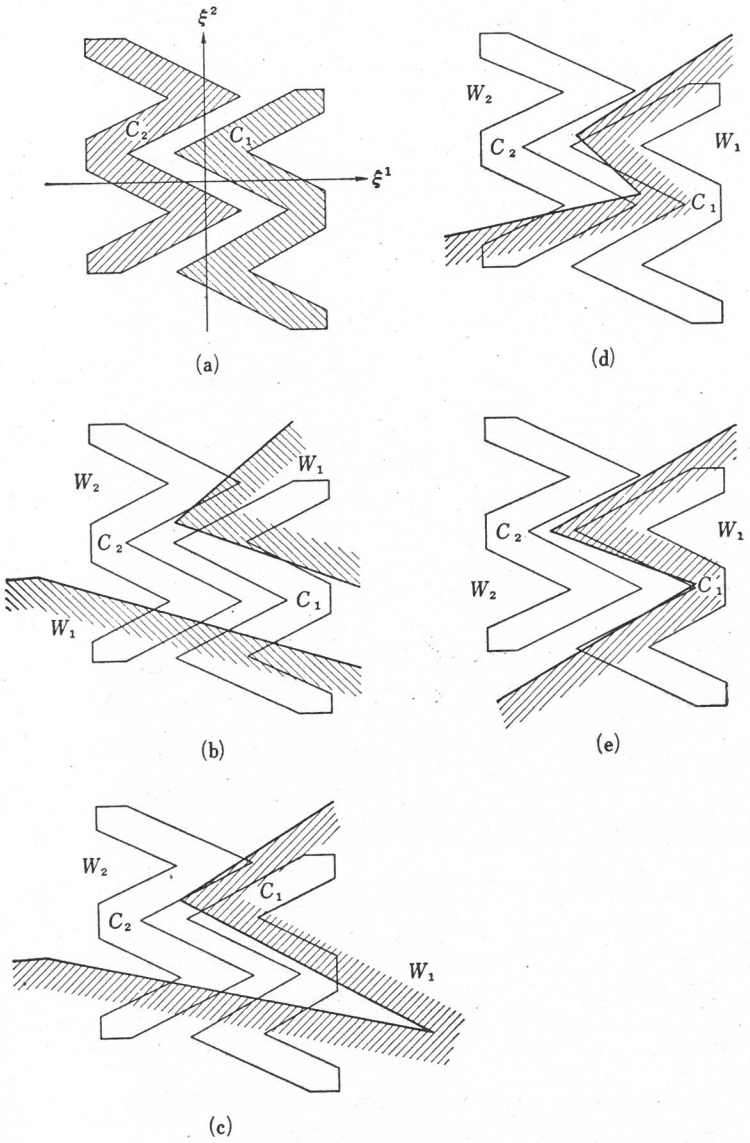

図 5·1

これは，分布が分離可能な例であるが，この学習法は分離可能でない場合でも使える。学習の収束については次節で扱う。なお，この場合でも，$L(\boldsymbol{\Theta})$ に最小値以外の極値が存在することに注意を要する。初めの識別領域の選び方によっては，正しい識別領域が得られないことがある。

この例では，四つの一次関数を用いているが，実際は三つの関数でパターンを識別することができる。事実，3回めの修正で，一つは不要になってしまう。しかし，一次関数の数を必要とするものより多めに選んでおくことは，極小値に落ち込むのを避ける点で効果がある。

C．学習の収束

初期値 $\boldsymbol{\Theta}_1$ から出発して，逐次学習を続けていく系を考える。時刻 $i(i=1, 2, \cdots)$ における識別ベクトルを $\boldsymbol{\Theta}_i$ で表わす。これは，時刻 i に系に提示されるパターン $\boldsymbol{\xi}_i$ に基づいて修正され，

$$\boldsymbol{\Theta}_{i+1}=\boldsymbol{\Theta}_i+\delta\boldsymbol{\Theta}_i$$

になる。時刻 i における L の値は

$$L_i=L(\boldsymbol{\Theta}_i) \tag{5·23}$$

であり，これが学習によって

$$L_{i+1}=L_i+\delta L_i$$

に変わる。変化分 δL_i は時刻 i に提示されるパターンに依存している。いま，ε を微小として，ε^2 の項を省略すると，δL_i の期待値は (5·6)，(5·7) より

$$\overline{\delta L_i}=-\varepsilon(\nabla L_i)_t C(\nabla L_i)\leqq 0 \tag{5·24}$$

であることがわかる。ただし

$$\nabla L_i=\nabla L(\boldsymbol{\Theta}_i)$$

である。等号は

$$\nabla L_i=0$$

すなわち，$\boldsymbol{\Theta}_i$ が最適識別ベクトルであるときにのみ成立する。

ここで，学習の収束を調べる。識別ベクトル $\boldsymbol{\Theta}_i$ は，これまでに発生したパターン $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_{i-1}$ の系列に依存している。パターンは確率分布 p_α, $p_\alpha(\boldsymbol{\xi})$に基づいて発生する確率変数であるから，$\boldsymbol{\Theta}_i$ もまた確率変数になる。$\boldsymbol{\Theta}_i$ の確率密度関数を $q_i(\boldsymbol{\Theta})$ と書こう。時刻 i における L の期待値を $\hat{L}_i$ と書くと*，

$$\hat{L}_i=\int q_i(\boldsymbol{\Theta})L(\boldsymbol{\Theta})d\boldsymbol{\Theta} \tag{5·25}$$

である。1回の学習による $\hat{L}_i$ の変化分 $\delta\hat{L}_i$ を調べると

$$\begin{aligned}\delta\hat{L}_i&=\hat{L}_{i+1}-\hat{L}_i\\&=\sum_\alpha\int q_i(\boldsymbol{\Theta})p_\alpha p_\alpha(\boldsymbol{\xi})\{L(\boldsymbol{\Theta}+\delta\boldsymbol{\Theta}(\boldsymbol{\xi}_i,\ \boldsymbol{\Theta}))-L(\boldsymbol{\Theta})\}d\boldsymbol{\Theta}d\boldsymbol{\xi}_i\\&=\int q_i(\boldsymbol{\Theta})\overline{\delta L}(\boldsymbol{\Theta})d\boldsymbol{\Theta}\leqq 0\end{aligned}$$

すなわち，$\delta\hat{L}_i$ は非負であることがわかる。したがって，$\hat{L}_i$ は単調減少する。明らかに

$$\hat{L}_i\geqq 0$$

であるから，数列 $\hat{L}_i$ は収束し，

$$\lim_{i\to\infty}\delta\hat{L}_i=0$$

が成立する。ところが

$$\delta\hat{L}_i=\int q_i(\boldsymbol{\Theta})\overline{\delta L(\boldsymbol{\Theta})}d\boldsymbol{\Theta}$$

であり，$\overline{\delta L}$ は最適値以外の $\boldsymbol{\Theta}$ に対しては常に正であるから，$\delta\hat{L}_i$ が0になるためには，$q_i(\boldsymbol{\Theta})$ が $\boldsymbol{\Theta}$ の最適値以外のところでは，0に収束しなければならない。これは $\boldsymbol{\Theta}_i$ が最適値に収束することを意味する。

以上の議論は ε^2 の項を省略したおおざっぱなものであった。収束を厳密に保証するには，ε を定数とせず，時刻 i における値を ε_i とおいて，これをだ

* $\boldsymbol{\Theta}_i$ はパターン列 $\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_{i-1}$ に依存している．$\hat{L}_i$ は $L(\boldsymbol{\Theta}_i)$ をすべての可能なパターン列について平均したものである．

んだん小さくしていけばよい。この場合，修正方法は

$$\delta\boldsymbol{\Theta}_i=\varepsilon_i C\boldsymbol{a}_\alpha(\boldsymbol{\xi}_i,\boldsymbol{\Theta}_i)$$

となる。

[定理 5·1]　条件

$$\left.\begin{aligned}&\lim_{i\to\infty}\varepsilon_i=0\\&\sum_{i=1}^{\infty}\varepsilon_i=\infty\end{aligned}\right\} \tag{5·26}$$

を満たすように ε_i を選ぶならば，$\boldsymbol{\Theta}_i$ は確率1で最適な識別ベクトルに収束する。

[証明]　確率的近似法を用いて行なえばよい[47]。定理の条件を満たす ε_i としては，たとえば

$$\varepsilon_i=\frac{1}{i}$$

がある。しかし，ε_i をこのように選ぶと，収束がきわめて遅くなり，学習によって系の変動を追従する場合にはほとんど意味のないことになる。

そこで，ε をある小さい定数に選んでおくと，どうなるかを調べていくことにする。なお，ε の大きさを自動的に調整していく系についても，後節でふれる。

[定理 5·2]　$\boldsymbol{\Theta}_{OP}$ を最適な識別ベクトルとし，$\boldsymbol{\Theta}_i$ が $\boldsymbol{\Theta}_{OP}$ から μ 以上離れている確率を $M_\mu{}^i(\varepsilon)$ とする*。このとき，任意の μ に対して，i が十分に大きいときは，ε を十分に小さく選ぶことにより，$M_\mu{}^i(\varepsilon)$ をいくらでも小さくできる。すなわち

$$\lim_{\varepsilon\to 0}\left(\lim_{i\to\infty}M_\mu{}^i(\varepsilon)\right)=0 \tag{5·27}$$

[証明]　$M_\mu{}^i(\varepsilon)$ を式で表わすと

$$M_\mu{}^i(\varepsilon)=\mathrm{Prob}\{|\boldsymbol{\Theta}_i-\boldsymbol{\Theta}_{OP}|\geqq\mu\} \tag{5·28}$$

* 最適な識別ベクトルが一つではないときは，$\boldsymbol{\Theta}_{OP}$ の集合を S_{OP}とし，$M\mu^i(\varepsilon)$ は $\boldsymbol{\Theta}_i$ から S_{OP} までの距離が μ 以上離れている確率とすればよい．

となる。

$$M_\mu(\varepsilon)=\lim_{i\to\infty} M_\mu{}^i(\varepsilon) \qquad (5\cdot29)^*$$

とおけば，

$$\lim_{\varepsilon\to 0} M_\mu(\varepsilon)=0$$

を証明すればよい。まず，$\overline{\delta L}$ の厳密な式を求めておこう。δL は

$$\delta L=-\varepsilon(\nabla L)_t C\boldsymbol{a}_\alpha(\boldsymbol{\xi})+\frac{\varepsilon^2}{2}\mathrm{tr}(C\boldsymbol{a}_\alpha)_t\nabla\nabla L(C\boldsymbol{a}_\alpha)+O(\varepsilon^3)$$

と書ける。$\boldsymbol{\xi}$ について平均をとることにすると，$\boldsymbol{a}_\alpha(\boldsymbol{\xi})$ の平均は，明らかに $-\nabla L$ である。$\boldsymbol{a}_\alpha(\boldsymbol{a}_\alpha)_t$ の平均を行列

$$Q=\Sigma\int p_\alpha p_\alpha(\boldsymbol{\xi})\boldsymbol{a}_\alpha(\boldsymbol{a}_\alpha)_t d\boldsymbol{\xi} \qquad (5\cdot30)$$

で表わそう。行列の関係式

$$\mathrm{tr}(AB)=\mathrm{tr}(BA)$$

を用いて，第2項を整理すれば

$$\overline{\delta L}=-\varepsilon(\nabla L)_t C(\nabla L)+\frac{\varepsilon^2}{2}\mathrm{tr}(CQC_t\nabla\nabla L)+O(\varepsilon^3) \qquad (5\cdot31)$$

が得られる。したがって，十分小さい ε に対しては，定数 c が存在して

$$\overline{\delta L}\leqq-\varepsilon(\nabla L)_t C\nabla L+c\varepsilon^2 \qquad (5\cdot32)$$

と書ける。

ここで

$$U_\mu=\{\boldsymbol{\Theta}\mid|\boldsymbol{\Theta}-\boldsymbol{\Theta}_{OP}|\geqq\mu\}$$
$$U'_\lambda=\{\boldsymbol{\Theta}\mid(\nabla L)_t C\nabla L\geqq\lambda\}$$

なる2種類の集合を考えよう。任意の $\mu>0$ に対して，

$$U_\lambda'\supset U_\mu$$

* $\lim_{i\to\infty} M_\mu{}^i(\varepsilon)$ の収束の条件については，確率過程の本，たとえば文献[49]を参照.

が成り立つような $\lambda>0$ を必ず求めることができる。

U_μ を用いて $M_\mu{}^i(\varepsilon)$ を表わすと

$$M_\mu{}^i(\varepsilon)=\int_{U_\mu} q_i(\boldsymbol{\Theta})d\boldsymbol{\Theta}$$

となる。したがって

$$\int(\nabla L)_t C\nabla L q_i(\boldsymbol{\Theta})d\boldsymbol{\Theta}\geqq\int_{U_\lambda'}(\nabla L)_t C\nabla L q_i(\boldsymbol{\Theta})d\boldsymbol{\Theta}$$
$$\geqq\lambda\int_{U_\lambda'} q_i(\boldsymbol{\Theta})d(\boldsymbol{\Theta})\geqq\lambda\int_{U_\mu} q_i(\boldsymbol{\Theta})d\boldsymbol{\Theta}=\lambda M_\mu{}^i(\varepsilon)$$

なる不等式が得られる。(5·32) の両辺を $q_i(\boldsymbol{\Theta})$ を用いて $\boldsymbol{\Theta}$ について平均し，この不等式を用いると

$$\delta\hat{L}_i\leqq-\varepsilon\lambda M_\mu{}^i(\varepsilon)+c\varepsilon^2$$

が得られる。これを i について1から N まで加え合わせ，N で割れば

$$\frac{\hat{L}_N}{N}\leqq-\varepsilon\lambda\frac{1}{N}\sum_{i=1}^{N} M_\mu{}^i(\varepsilon)+c\varepsilon^2$$

さらに，$N\to\infty$ の極限をとると，$\hat{L}_N$ は有限であるから，

$$0\leqq-\varepsilon\lambda M_\mu(\varepsilon)+c\varepsilon^2$$

が得られる。したがって

$$M_\mu(\varepsilon)\leqq\frac{c}{\lambda}\varepsilon$$

すなわち

$$\lim_{\varepsilon\to 0} M_\mu(\varepsilon)=0$$

が証明される。

なお，パターンの分布が分離可能であり，しかもパターンの個数が有限である場合には，有限回の学習回数で収束することがわかる。

5·3　学習の速度，精度，動特性

A．学習の特性

本節では，最適な識別ベクトル Θ_{OP} の近傍における，学習系の諸特性を調べるのに有効な補題を証明する。Θ の関数 $f(\Theta)$ を考え，その値が学習の進展とともにどう変化していくかを，一般的に考えることにする。Θ_i はパターン列 $\xi_1, \cdots, \xi_{i-1}$ に依存しているから，$f(\Theta_i)$ の値もまたいかなるパターン列が過去に発生したかに関係している。時刻 i における，$f(\Theta)$ の期待値を $\widehat{f(\Theta)}_i$ で表わす。

$$\widehat{f(\Theta)}_i=\int f(\Theta)q_i(\Theta)d\Theta \tag{5·33}$$

以下，$q_i(\Theta)$ による期待値を $\hat{\ }_i$ で表わす。

［補題］　1回の学習により，$\hat{f}_i$ は

$$\hat{f}_{i+1}=\hat{f}_i-\varepsilon\widehat{\{(\nabla f)_t C\nabla L\}}_i+\frac{\varepsilon^2}{2}\mathrm{tr}\widehat{(C\nabla\nabla f)}_i+O(\varepsilon^3) \tag{5·34}$$

に変化する。

［証明］　時刻 i における識別ベクトルを Θ とし，このとき提示されるパターンを $\xi\in C_\alpha$ とする。時刻 $i+1$ には，識別ベクトルは

$$\Theta'=\Theta+\delta\Theta(\xi,\alpha,\Theta)$$

になる。ここで Θ' の確率密度関数 $q(\Theta')$ を求めよう。

提示されるパターンが ξ と $\xi+d\xi$ の間にある確率は*，$p_\alpha(\xi)d\xi$ である。$d\xi$ が

$$d\Theta'=\frac{\partial\Theta'(\xi,\alpha,\Theta)}{\partial\xi}d\xi$$

を満足するときは，修正された識別ベクトルは $\Theta'+d\Theta'$ の間にある。したがって，修正された識別ベクトルが Θ' と $\Theta'+d\Theta'$ との間にある確率は

* 正確には，$\xi_i\leqq\eta_i\leqq\xi_i+d\xi$ を満たす η_i を成分にもつパターンが発生する確率．

$$q_\alpha(\Theta')d\Theta'=p_\alpha(\xi)d\xi$$

である。C_α のパターンの生ずる確率は p_α であるから，この両辺に p_α を掛けて α について加えれば，Θ' の確率密度関数 $q(\Theta')$ が

$$q(\Theta')d\Theta'=\sum_\alpha p_\alpha p_\alpha(\xi)d\xi \tag{5·35}$$

として求まる。

時刻 i における識別ベクトルを Θ としたが，これは実は $q_i(\Theta)$ なる分布に従う確率変数である。したがって，(5·35) の両辺の，$q_i(\Theta)$ による期待値をとろう。するとこれが，時刻 $i+1$ における識別ベクトルの確率密度を表わすことになる。したがって

$$q_{i+1}(\Theta')d\Theta'=\sum_\alpha d\xi\int q_i(\Theta)p_\alpha p_\alpha(\xi)d\Theta$$

ここで右辺の ξ は，Θ と Θ' との関数と考えている。

$$\hat{f}_{i+1}=\int f(\Theta')q_{i+1}(\Theta')d\Theta'$$

であるから，

$$\hat{f}_{i+1}=\sum_\alpha\iint f(\Theta')q_i(\Theta)p_\alpha p_\alpha(\xi)d\Theta d\xi$$

$$=\int\overline{f(\Theta')}q_i(\Theta)d\Theta$$

が得られる。ここに $\overline{\ }$ はパターン ξ についての平均を意味する。

$$\overline{f(\Theta')}=\overline{f(\Theta+\delta\Theta)}=f(\Theta)-\varepsilon(\nabla f)_t C\nabla L+\frac{\varepsilon^2}{2}\mathrm{tr}C\mathrm{Q}C_t\nabla\nabla f+O(\varepsilon^3)$$

を考慮に入れれば，(5·34) が得られる。

この補題において，関数 f はベクトル値や行列値をとるものであってもよいことに注意されたい。

B．学習の速度と精度

前節で求めた補題を用いると，学習の速度と精度とを求めることができる。

収束の速度は，識別ベクトルの期待値 $\hat{\Theta}_i$ が Θ_{OP} に近づく近づき方を示すものである。

$$f(\Theta)=\Theta$$

とおくと

$$\hat{f}_i=\hat{\Theta}_i$$

である。したがって，補題を用いれば $\hat{\Theta}_i$ に対する漸化式が得られる。

$L(\Theta)$ を

$$L(\Theta)=L(\Theta_{OP})+\frac{1}{2}(\Theta-\Theta_{OP})_t A(\Theta-\Theta_{OP})+O(|\Theta-\Theta_{OP}|^3) \qquad (5\cdot36)$$

と展開し，Θ_{OP} の近傍のみを考えることにして，$O(|\Theta-\Theta_{OP}|^3)$ の項を省略しよう。すると，学習速度の定理が得られる。

[定理 5·3]（学習速度の定理）　時刻 i における識別ベクトルの期待値は

$$\hat{\Theta}_i=\Theta_{OP}+(E-\varepsilon CA)^{i-1}(\Theta_1-\Theta_{OP}) \qquad (5\cdot37)$$

である。

[証明]

$$f(\Theta)=\Theta$$

に対しては

$$\nabla f=E$$
$$\nabla\nabla f=0$$

が成立する。これらを補題に代入すれば

$$\hat{\Theta}_{i+1}=\hat{\Theta}_i-\varepsilon C\widetilde{(\nabla L)_i}$$

となる。ところが

$$\nabla L=A(\Theta-\Theta_{OP})$$

であるから，これを代入すると

$$\hat{\Theta}_{i+1}=(E-\varepsilon CA)\hat{\Theta}_i+\varepsilon CA\Theta_{OP}$$

が得られる。これは $\hat{\Theta}_i$ に関する差分方程式である。初期値を Θ_1 としてこれを解くと，(5·37) が得られる。

これで，$\hat{\Theta}_i$ は指数的に Θ_{OP} に近づくことがわかった。しかし，実際の Θ_i は期待値 $\hat{\Theta}_i$ に必ずしも一致するわけではない。現実の Θ_i の $\hat{\Theta}_i$ からのずれは，共分散行列

$$V_i=\widetilde{\{(\Theta-\hat{\Theta}_i)(\Theta-\hat{\Theta}_i)_t\}_i}$$
$$=\widetilde{(\Theta\Theta_t)_i}-\hat{\Theta}_i\hat{\Theta}_{it} \qquad (5\cdot38)$$

で評価できる。すなわち，V_i は時刻 i における学習の精度を表わすものとみてよい。V_i もまた，補題に基づいて計算することができる。

[定理 5·4]（学習精度の定理）　時刻 i における識別ベクトルの共分散行列は

$$V_i=2\varepsilon\{E-(E-\varepsilon\widetilde{CA})^{i-1}\}(\widetilde{CA})^{-1}CQC_t \qquad (5\cdot39)$$

である。特に，終局の精度は

$$\lim_{i\to\infty}V_i=2\varepsilon(\widetilde{CA})^{-1}CQC_t \qquad (5\cdot40)$$

となる。ただし，$\widetilde{CA}$ は行列に対する線形演算子で，行列 M に対して

$$\widetilde{CA}M=CAM+(CAM)_t \qquad (5\cdot41)$$

で定義される。

[証明]　ここでは，細かく式を追うことはせず，大まかな方針を示すだけに止める*。

$$f(\Theta)=\Theta\Theta_t$$

---

* 証明に用いる計算に，本質的にむずかしい点はないが，相当にめんどうである．めんどうになる理由は，f を行列とすると，∇f や $\nabla\nabla f$ が，3階ないし4階のテンソル量になるためである．

とおいて，補題を用いると

$$\widehat{(\Theta\Theta_t)}_{i+1}=\widehat{(\Theta\Theta_t)}_i-2\varepsilon\{\widehat{CA(\Theta-\Theta_{OP})\Theta_t}\}^s+2\varepsilon^2CQC_t \qquad (5\cdot42)$$

が得られる。ここに添字 s は，行列の対称部分をとることを意味する。一方，(5·37) より

$$\hat{\Theta}_{i+1}\hat{\Theta}_{i+1t}=\{(E-2\varepsilon CA)\hat{\Theta}_i\hat{\Theta}_{it}\}^s+2\varepsilon(CA\Theta_{OP}\hat{\Theta}_{it})^s$$

が得られる。(5·42) からこれを引けば，差分方程式

$$V_{i+1}=(E-\varepsilon\widehat{CA})V_i+2\varepsilon^2CQC_t \qquad (5\cdot43)$$

が得られる。初期値を $V_1=0$ として，これを解くと (5·39) が証明される。

こうして学習の速度と精度とが求まった。いま，CA の最小固有値を λ_0 としよう。すると，[定理 5·3] により，これに対応する固有ベクトルは収束が最も遅い方向を示し，収束の度合が時定数 $\varepsilon\lambda_0$ で示されることがわかる。一方，学習の終局の精度は $2\varepsilon(\widehat{CA})^{-1}CQC_t$ で示される。それゆえ，ε を大きくすれば，収束は速くなるが精度は悪くなるし，ε を小さくすれば精度は良くなるが収束は遅くなることがわかる。A や Q についての，なんらかの知識が得られている場合には，速度と精度とのかねあいをみながら，定数 ε と行列 C を定めればよい。

C．学習系の動特性

識別系の確率構造は時間とともに変動することが多い。この場合，最適な識別ベクトル Θ_{OP} もまた変動する。学習識別系がこの変化にどう追従できるかを調べる。

時刻 i における最適ベクトルを $\Theta_0+\Delta_i$ としよう。Δ_i は最適識別ベクトルの変動を示している。この場合，$\hat{\Theta}_{i+1}$ の漸化式は

$$\hat{\Theta}_{i+1}=(E-\varepsilon CA)\hat{\Theta}_i+\varepsilon CA(\Theta_0+\Delta_i) \qquad (5\cdot44)$$

となる。行列Aも，一般には時刻とともに変動するが，ここでは簡単のため，定数行列とした。これを解くと

$$\hat{\Theta}_i=\Theta_0+(E-\varepsilon CA)^i(\Theta_1-\Theta_0)+\varepsilon\sum_{k=0}^{i-1}(E-\varepsilon CA)^{i-k-1}CA\Delta_k \qquad (5\cdot45)$$

が得られる。最後の項が，確率構造の変動の影響を示すものである。

最適識別ベクトルが，急に Δ だけずれたとしよう。すると，i 時刻後には，識別ベクトルの期待値は

$$\varepsilon\sum_{k=0}^{i-1}(E-\varepsilon CA)^{i-k-1}CA\Delta=\{E-(E-\varepsilon CA)^{i-1}\}\Delta$$

だけ変動して，このずれを追従していく。したがって

$$S_i=E-(E-\varepsilon CA)^{i-1} \qquad (5\cdot46)$$

とおけば，S_i は学習系の**階段応答** (step-response) を示すものといえる。

より直観的に応答をみるために，Θ_{OP} が周期的に変動する場合に，系がこれをどう追従するかをみよう。いま，最適識別ベクトルが $2\pi/\omega$ の周期で変動しているとして

$$\Delta_i=\Delta\sin\omega i \qquad (5\cdot47)$$

とおく。変動の周期は十分大きい，すなわち ω は微小であると仮定しておく。また，Δ は CA の固有ベクトルであり，その固有値を λ とする（一般の場合は，固有解を重ね合わせればよい）

$$CA\Delta=\lambda\Delta$$

(5·44) に (5·47) を代入すれば，差分方程式

$$\hat{\Theta}_{i+1}=(E-\varepsilon CA)\hat{\Theta}_i+\varepsilon CA(\Theta_0+\Delta\sin\omega i)$$

が得られる。この方程式の，過渡状態を示す項を除いた定常振動解は

$$\hat{\Theta}_i=\Theta_0+a\Delta\sin(\omega i+\varphi)$$

とおける。これを代入すると，振幅の倍率 a と位相遅れ φ とを求める方程式

$$\left.\begin{aligned}a\{\cos(\omega+\varphi)-(1-\varepsilon\lambda)\cos\varphi\}-\varepsilon\lambda=0\\ \sin(\omega+\varphi)-(1-\varepsilon\lambda)\sin\varphi=0\end{aligned}\right\}$$

が出てくる。

$$\alpha=\frac{\omega}{\varepsilon\lambda} \tag{5·48}$$

とおき，αを微小とみて高次の項を省略すれば，

$$a=\frac{1}{\sqrt{1+\alpha^2}}$$

$$\tan\varphi=-\alpha$$

が求まる。それゆえ，定常解は

$$\hat{\Theta}_i=\Theta_0+\frac{1}{\sqrt{1+\alpha^2}}\Delta\sin(\omega i-\alpha) \tag{5·49}$$

である。

これは，系の**周波数応答**（frequency response）を示すものである。すなわち，最適な識別ベクトルが $2\pi/\omega$ の周期で変動するとき，学習により得られる識別ベクトルは，位相がα遅れ，振幅が $1/\sqrt{1+\alpha^2}$ 倍になって，これを追従する。したがってαが小，すなわち

$$\omega\ll\varepsilon\lambda$$

ならば，学習系はこの変動を十分によく追従できるといえる。

D．学習法の学習

学習系の特性，特に収束速度と精度とは定数 ε と行列 C とに強く依存している。ところで，まえに示したように，εC を調整して速度を上げるようにすると精度が落ち，逆に精度を上げれば速度が落ちる。両者の間には相反関係があって，両方を同時に良くするわけにはいかない。

しかし，よく考えてみると，速度と精度とは必ずしも同時に要求されるのではない。識別ベクトルが最適値から遠く離れているときは収束速度が速いことが望まれるが，精度すなわち平均値のまわりのばらつきはさしあたりたいした問題にはならない。ところが識別ベクトルが最適値に近づいたときは，ばらつきの小さいこと，すなわち精度が重要になってくる。したがって，識別ベクトルが最適値から遠いときは速度を速めるように εC を選び，最適値に近いときは精度を良くするように εC を選ぶことができれば，きわめて望ましい系が設計できる。

ところが，最適識別ベクトルは未知であるから，現在の識別ベクトルがこれに近いか遠いかの判断はできず，εC をこう都合よく選ぶことはできない。そこで，εC を過去のデータに基づいて自動的に修正していき，その結果，識別ベクトルが最適点から遠いときは速度が速まり，近いときは精度が高まるようにすることを考える。εC は具体的な学習方法を指定する定数であるから，これは**学習法の学習**といえる。

識別ベクトルが最適点から離れているときは，学習により生ずる修正ベクトルは，同方向を向いているものが多く出るだろう。これに反して，最適点に近い場合には，いちど修正すると Θ が最適点を飛びこすなどして，修正ベクトルの方向はまちまちになってくる。この情報を，εC の学習に用いる。次のような εC の学習規則を考える。

識別ベクトル Θ がパターン $\xi\in C_\alpha$ によって

$$\Theta'=\Theta+\delta\Theta(\xi,\alpha,\Theta)$$

に修正され，これがさらにパターン $\xi'\in C_{\alpha'}$ によって

$$\Theta''=\Theta+\delta\Theta+\delta\Theta'(\xi',\alpha',\Theta')$$

に修正されたとしよう。ここで

$$\delta\Theta\neq0,\quad\delta\Theta'\neq0$$

とする（修正ベクトルが 0 である場合は，それを飛ばして，0 でないものだけを考えればよい）。このとき，次の式を用いて，εC を εC+ΔC に変える

$$\Delta C=\gamma\boldsymbol{a}_\alpha(\xi,\Theta)\boldsymbol{a}_{\alpha'}(\xi',\Theta')_t \tag{5·50}$$

ここにγは正の定数である。

この修正の効果をみるために，ΔC の期待値を求める。ΔC をパターン ξ, ξ' について平均すると

$$\overline{\Delta C}=\sum_{\alpha,\alpha'}\int \gamma p_\alpha p_\alpha(\xi)p_{\alpha'}p_{\alpha'}(\xi')\boldsymbol{a}_\alpha(\xi,\Theta)\boldsymbol{a}_{\alpha'}(\xi',\Theta')_t d\xi d\xi'$$

が得られる。これを ξ' について積分すると，(5·9) を用いて，

$$\overline{\Delta C}=-\sum_{\alpha}\int \gamma p_\alpha p_\alpha(\xi)\boldsymbol{a}_\alpha(\xi,\Theta)(\nabla L(\Theta'))_t d\xi$$

となる。

$$\nabla L(\Theta')=\nabla L(\Theta)+\nabla\nabla L(\Theta)\cdot\delta\Theta(\xi,\alpha,\Theta)$$

を代入して，ξ についての積分を行なうと

$$\overline{\Delta C}=\gamma\{\nabla L(\nabla L)_t-\varepsilon Q(\Theta)C_tA(\Theta)\}\qquad(5\cdot51)$$

が求まる。ただし

$$Q(\Theta)=\sum_{\alpha}\int p_\alpha p_\alpha(\xi)\boldsymbol{a}_\alpha\boldsymbol{a}_{\alpha t}d\xi$$

$$A(\Theta)=\nabla\nabla L(\Theta)$$

である。

識別ベクトルが最適点から遠いときは，(5·51) の第2項は第1項に比べて小さい。それゆえ

$$\overline{\Delta C}=\gamma\nabla L(\nabla L)_t\qquad(5\cdot52)$$

となる。いま εC を $\overline{\Delta C}$ だけ変えたとすると，修正ベクトルは

$$\overline{\Delta C}\boldsymbol{a}_\alpha=\gamma\nabla L(\nabla L)_t\boldsymbol{a}_\alpha\qquad(5\cdot53)$$

だけ変わる。

$$\overline{\nabla L\cdot\boldsymbol{a}_\alpha}=-\nabla L\cdot\nabla L\leqq 0$$

であるから，修正 ΔC は，平均として $-\nabla L$ 方向を強め，収束の速度を速めるように働く。

一方，識別ベクトルが最適点に近いときは

$$\nabla L\approx 0$$

であるので，(5·51) の第1項は小さくなり

$$\overline{\Delta C}=-\gamma\varepsilon QC_tA\qquad(5\cdot54)$$

と書ける。これは，修正ベクトルを

$$\overline{\Delta C}\boldsymbol{a}_\alpha=-\gamma\varepsilon QC_tA\boldsymbol{a}_\alpha\qquad(5\cdot55)$$

だけ変える。QC_tA は正の定符号をもつ行列であるから，この変化は，修正ベクトル $\varepsilon C\boldsymbol{a}_\alpha$ を弱める方向に働く。すなわち，修正ベクトル $\delta\Theta$ の絶対値が減少し，精度が上がる。

こうして，最適点から遠いときは収束速度を自動的に速め，最適点に近いときは精度を自動的に高めるような，高次の学習識別系が得られた。

一次元のパターンで，分布が重なり合っている場合について，ε を自動的に変える識別実験を行なった結果では，「学習法の学習」は系の動特性をかなり高める。分布を途中で変動させた例では，人間が行なった同様の学習識別の実験結果よりも，よい結果が得られる。

この方向での学習識別理論の今後の発展が待たれる。