in: W.Güttinger, Eikemeier (Eds)
Struktural Stability in Physics
Springer Verlag Berlin 1979

# Catastrophic Effects in Pattern Recognition

R. Brause and M. Dal Cin

Institute for Information Sciences, University of Tübingen
D-7400 Tübingen, Fed. Rep. of Germany

## 1. Introduction

The purpose of this note is to report on an ongoing investigation into the steady state behavior of pattern recognition systems. The system considered in this paper generates the decision boundaries which separate pattern classes on the basis of a stochastic learning algorithm [1]. The inputs of this system, and hence, of the algorithm, are observed patterns drawn from a probability distribution $p(\underline{x})$, $\underline{x}$ a pattern vector. In the one dimensional case considered later on the decision boundaries are points in the pattern space. The decision of the system is based on a risk function R. That is, the system selects the boundaries that minimize R. As the distribution $p(\underline{x})$ of inputs is gradually changing the steady state boundaries will vary continuously most of the time. However, at certain instances we observe abrupt changes of the decision boundaries.

Abrupt changes in decision making were also investigated by E.G. Zeeman [2]. Our purpose is to derive the precise analytical condition for such catastrophic effects and to verify the results by simulation experiments.

## 2. A stochastic learning algorithm

When the number of pattern classes are given, the pattern recognition system tries to find an optimal separation of these classes. This search is controlled by a sequence of observed samples and is based on a risk function (even so the precise form of the risk function is not known to the system).

A typical configuration with two classes $\omega_1$ and $\omega_2$ and a two-dimensional distribution of patterns is shown in Fig. 1. (The boundary was obtained after 1000 iterations of the learning algorithm given in (3) below.)

Is there always one optimal boundary? Fig. 2 shows a situation with two optimal boundaries provided the boundary is a single straight line. It will be shown that in this case gradual changes of certain pattern distributions give rise to abrupt switches from one to two or more stationary states of the boundary.

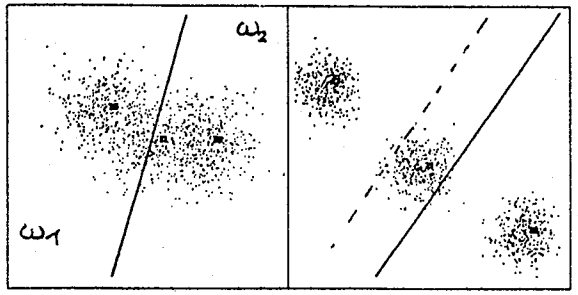


Fig. 2 Two optimal boundaries

Let a pattern be characterized by N features $x_1, x_2, \ldots, x_N$ where $x_i \in \mathbb{R}$. It is represented by a point $\underline{x} = (x_1, x_2, \ldots, x_N)$ of the set $X \subset \mathbb{R}^N$ of all possible patterns. The problem of classifying patterns into M classes $\omega_1, \omega_2, \ldots, \omega_M$ is equivalent to finding a partition of X into M corresponding disjoint subsets $X_1, X_2, \ldots, X_M$ which solve the problem. Let the i'th subset $X_i$ be represented by a reference pattern $\underline{c}_i$ belonging to $X_i$.

Instead of a teacher the system is provided with M loss functions

$$L_1(\underline{x},\underline{c}_1),\ L_2(\underline{x},\underline{c}_2),\ \ldots,\ L_M(\underline{x},\underline{c}_M)$$

where $L_i(\underline{x},\underline{c}_i) \geq 0$ is the loss that will be incurred if the system classifies pattern $\underline{x}$ as belonging to $X_i$. Of course, the loss should be minimal if $\underline{x}$ actually belongs to $\omega_i$. We assume that the system chooses $\underline{c}_i$ such that for patterns of class $X_i$ the expectation value

$$R(\underline{c}_i) = \sum_{j=1}^{M} R_j(\underline{c}_i)\ P(\omega_i)$$

is minimal, where

$$R_j(\underline{c}_i) = \int_{X_i} L_i(\underline{x},\underline{c}_i)\ p(\underline{x}|\omega_j)\ d\underline{x} \tag{1}$$

is the risk due to a misclassification of patterns from $\omega_j$ into $X_i$ $(i \neq j)$ and a bad choice of the reference pattern $\underline{c}_i$ $(i = j)$. $P(\omega_i)$ is the a priori probability of class $\omega_i$. Hence, the system tries to minimize the risk function

$$R(\underline{c}_i) = \int_{X_i} L_i(\underline{x},\underline{c}_i)\ p(\underline{x})\ d\underline{x},\quad i = 1, 2, \ldots, M \tag{2}$$

by choosing the optimal set $c = (\underline{c}_1, \underline{c}_2, \ldots, \underline{c}_M)$ of reference patterns. It can be shown that in this case, also, the total risk

$$\hat{R}(c) = \sum_{i=1}^{M} R(\underline{c}_i)$$

is minimal.

The method of stochastic gradient search proposed by Robbins and Monro [3] provides us with the following algorithm for finding the minimum of $\hat{R}(c)$.

Let $\underline{c}_i[n]$ be the reference pattern representing $X_i$ after the n'th learning step and let $\underline{x}[n]$ be the next pattern shown to the system. Then a new set of reference patterns will be generated according to the following algorithm (or stability mechanisms).

$$\begin{aligned} \underline{c}_i[n+1] &= \underline{c}_i[n] - \gamma[n]\ \nabla_{\underline{c}_i} L(\underline{x}[n], \underline{c}_i)\Big|_{\underline{c}_i[n]} \\ \underline{c}_j[n+1] &= \underline{c}_j[n],\quad i \neq j, \end{aligned} \tag{3}$$

where index i is such that

$$L_i(\underline{x},[n], \underline{c}_i[n]) = \min_k \{L_k(\underline{x}[n], \underline{c}_k[n])\} \tag{4}$$

and $\gamma[n] \in \mathbb{R}$. Observe that knowledge of $p(\underline{x})$ is not necessary for this algorithm.

A steady state $c^*$ of this algorithm is never reached by finitely many iteration steps but the convergence is guaranteed with

$$\lim_{n\to\infty} P(c[n] = c^*) = 1$$

and[+] $E(c^* - c[n]) = 0$,

[+] –

if the following conditions for $\gamma[n]$ hold [3]:

(a) $\lim_{n\to\infty} \gamma[n] = 0$; (b) $\lim_{n\to\infty} \sum_{i=1}^{n} \gamma[i] = \infty$; and (c) $\lim_{n\to\infty} \sum_{i=1}^{n} \gamma[i]^2 = s < \infty$.

Next we derive a condition which tells us when switches of decision can occur.

## 3. A criterion for instabilities

A simple example of a loss function which will be used in the following is

$$L_i(\underline{x},\underline{c}_i) = \frac{1}{2}(\underline{x} - \underline{c}_i)^2. \tag{5}$$

Then, $\hat{R}(c^*) = \min_c \hat{R}(c)$ if

$$\begin{aligned} &\nabla_{\underline{c}_i} R(\underline{c}_i)\Big|_{\underline{c}_i^*} = 0 \\ &= \int_{X_i} \nabla_{\underline{c}} L_i(\underline{x},\underline{c}_i^*)\ p(\underline{x})\ d\underline{x} \qquad \text{(The variation of } X_i \text{ vanishes.)} \\ &= \int_{X_i} (\underline{x} - \underline{c}_i^*)\ p(\underline{x})\ d\underline{x}, \qquad i = 1, 2, \ldots, M. \end{aligned}$$

Hence, the critical points are $c^* = (\underline{c}_1^*, \ldots, \underline{c}_M^*)$, where

$$\underline{c}_i^* = E(\underline{x}|X_i) = \int_{X_i} \underline{x}p(\underline{x})\ d\underline{x} \Big/ \int_{X_i} p(\underline{x})\ d\underline{x}. \tag{6}$$

Now, the boundary $\underline{d}_{ij}$ between two classes $\omega_i$ and $\omega_j$ is given by points $\underline{x}$ for which

$$L_i(\underline{x},\underline{c}_i) = L_j(\underline{x},\underline{c}_j).$$

Hence, the boundaries chosen by the system after the n'th learning step are determined by

$$\underline{d}_{ij}[n] = \frac{1}{2}(\underline{c}_i[n] + \underline{c}_j[n]) \tag{7}$$

with the steady states

$$\underline{d}_{ij}^* = \frac{1}{2}(\underline{c}_i^* + \underline{c}_j^*) = \frac{1}{2}(E(\underline{x}|X_i) + E(\underline{x}|X_j)).$$

In the case of N = 1, M = 2 this reduces to

$$\begin{aligned} d^* &= \frac{1}{2}(E(x|x > d^*) + E(x|x \leq d^*)) \\ &:= \chi(d^*) \end{aligned} \tag{8}$$

where now $X_1 = (-\infty, d)$ and $X_2 = [d, +\infty)$.

If $p(x)$ is symmetric (i.e. $p(x) = p(-x)$), the following relations hold (see Appendix):

$$\begin{aligned} &\text{(a)} \quad \lim_{d\to\infty} [d/2 - \chi(d)] = 0, \\ &\text{(b)} \quad \chi(d) = -\chi(-d), \text{ hence, } \chi(0) = 0, \\ &\text{(c)} \quad \frac{\partial}{\partial d}\chi(d)\Big|_{d=0} = 2p(0)\ E(x|x > 0). \end{aligned} \tag{9}$$

That is, $\chi(d)$ approaches d/2 and d = 0 is a steady state of the algorithm.

Now, if $\chi(d)$ crosses the diagonal $f_1(d) = d$ below and above the d-axis, then the cross-points are also steady states, cf. Fig. 3. This certainly occurs if the derivative of $\chi(d)$ at $d = 0$ is greater than 1. Thus, if a symmetric probability distribution satisfies

$$m := 2p(0)\, E(x|x > 0) > 1, \tag{10}$$

then there are at least three possible steady states of the learning algorithm (3); $d^* = 0$ is unstable in this case.

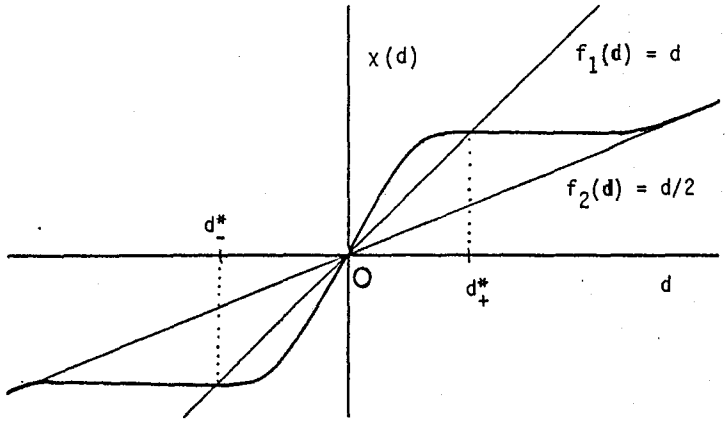


Fig. 3 Graph of $\chi$

In the next section we show the performance of our learning algorithm. To this end, we choose the following family of pattern distributions:

$$P_{B,z}(x) = \frac{1-B}{2}\, N(-z,\sigma) + B\, N(0,\sigma) + \left(\frac{1-B}{2}\right) N(z,\sigma) \tag{11}$$

where

$$N(z,\sigma) = (\sqrt{2\pi\sigma})^{-1} \exp(-(x-z)^2/2\sigma)$$

and $0 \le B \le 1$. Thus, Eq. (10) is now

$$m(z,B) > 1. \tag{12}$$

We compare the performance of the stochastic learning algorithm with that of the following two learning algorithms

$$\bar{d}[n+1] = 1/2\, (1/n_1 \sum_{i=1}^{n_1} x_i + 1/n_2 \sum_{i=1}^{n_2} x_i),\ n = n_1 + n_2. \tag{13}$$

where the patterns are drawn sequentially from pattern distribution (11) and

$$\bar{d}[n+1] = \chi(\bar{d}[n]) = 1/2\, (E(x|x < \bar{d}[n]) + E(x|x > \bar{d}[n])), \tag{14}$$

The second algorithm utilizes the maximum amount of information available. Its steady states are the same as that for (3).

## 4. Simulation experiments

The computer simulations [4] of the stochastic learning algorithm (3) and its averaged versions (13) and (14) confirm the theoretical results. Diagram (4.1) shows the bifurcational splitting of the steady states $d^*$ of the boundary $\bar{d}[n]$ for algorithm (14), when the control parameter $m \ge 1$ is linearly varied. (B=const =.5)

For every value of m(z), 20 iterations were initialized with 3 different starting values $\bar{d}[0] = -1, 0, +1$ and after a fixed number $N_{max}$ of iterations the state $\bar{d}[N_{max}]$ of the boundary was recorded. State $d^* = 0$ is a stable solution for the determinate algorithm (14). Of course, this is no longer true for the stochastic algorithms (3) and (13), cf. Figs. 4.2 and 4.3, respectively.
After the $N_{max}$'th iteration the state of the boundary is stochastically distributed around the two stable states $d^*_{+-}$. (It can be shown analytically that $d^*(m)$ is linear in m if there are at least two patterns $x_+$, $x_-$ between $\pm z$ and 0 where $p_{B,z}(x_\pm) = 0$.)

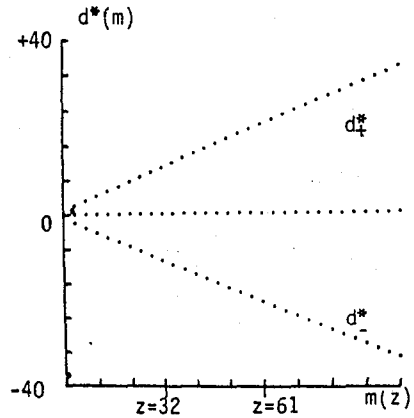


Fig. 4.1

The histograms of Figs. 5.1 - 5.3 show the abrupt switch of the stable solutions for the three algorithms when the control parameter m(z) is gradually changed from $m(z) > 1$ to $m(z) < 1$. Here too, the boundaries of the stochastic algorithms are distributed around the stable states. However, the variance is too great for us to see the switching.

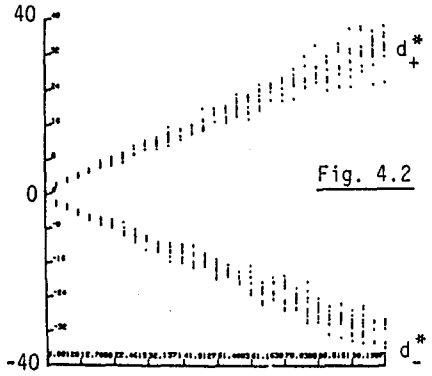


Fig. 4.2

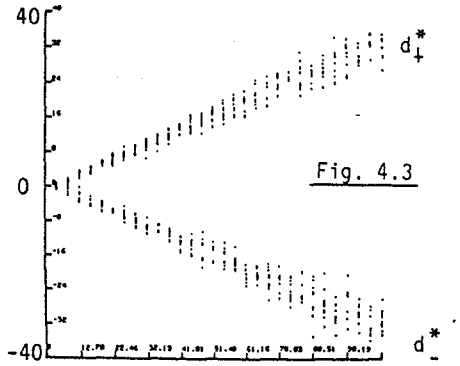


Fig. 4.3

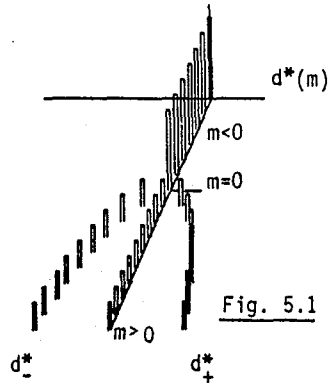


Fig. 5.1

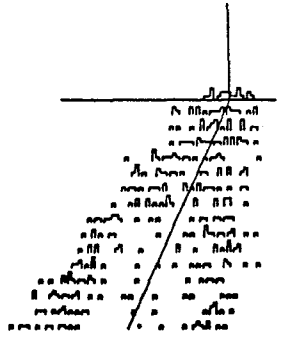

Fig. 5.2

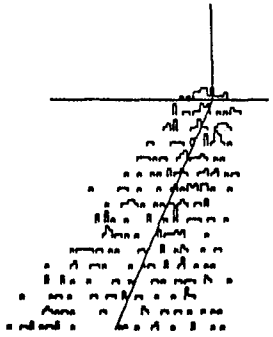

Fig. 5.3

Appendix: Proof of Eq. (9):

Let $I(a,b) = \int_a^b x\,p(x)\,dx \Big/ \int_a^b p(x)\,dx$, $I(-\infty,+\infty) = 0$, since $E(x) = 0$.

(a) $$\lim_{d\to\infty} (d/2 - \chi(d)) = \lim_{d\to\infty} \frac{1}{2}\,(d - I(d,\infty))$$

$$\overset{\text{l'Hopital}}{=} \lim_{d\to\infty} \frac{1}{2}\,[d - d\cdot p(d)/p(d)] = 0,$$

(b) $$2\chi(-d) = I(-\infty,-d) + I(-d,\infty) =$$

$$= -I(\infty,d) - I(d,-\infty) = -2\chi(d), \text{ since } p(x) = p(-x)$$

(c) $$2\chi'(d)\Big|_{d=0} = \Big[\int_0^\infty p(x)\,dx\Big]^{-2} \Big[-dp(d)\cdot\int_d^{-\infty} p(x)\,dx +$$

$$+\int_d^{-\infty} x\,p(x)\,dx\,p(d) - dp(d)\cdot\int_d^\infty p(x)\,dx +$$

$$+\int_d^\infty xp(x)\,dx\,p(d)\Big]\Big|_{d=0} = 8p(0)\int_0^\infty xp(x)\,dx$$

$$= 4p(0)\,E(x \mid x \geq 0).$$

Acknowledgments

The authors are indebted to the late E. Pfaffelhuber, who helped provide the impetus for this investigation. The assistance of E. Dilger is also acknowledged.

References

1. Tou, J.T. and R.C. Gonzalez, Pattern Recognition Principles, Addison-Wesley Publishing Co., 1974
2. E.C. Zeeman, private communication (1978)
3. Robbins, H. and S. Monro, A stochastic approximation method, Ann. Math. Stat. 22, 400-407 (1951)
4. Brause, R., Mustererkennung mit stochastischem Lernalgorithmus, preprint, Institut für Informationsverarbeitung, Tübingen, 1978