A Theory of Adaptive Pattern Classifiers

SHUNICHI AMARI

Abstract—**This paper describes error-correction adjustment procedures for determining the weight vector of linear pattern classifiers under general pattern distribution. It is mainly aimed at clarifying theoretically the performance of adaptive pattern classifiers. In the case where the loss depends on the distance between a pattern vector and a decision boundary and where the average risk function is unimodal, it is proved that, by the procedures proposed here, the weight vector converges to the optimal one even under nonseparable pattern distributions. The speed and the accuracy of convergence are analyzed, and it is shown that there is an important tradeoff between speed and accuracy of convergence. Dynamical behaviors, when the probability distributions of patterns are changing, are also shown. The theory is generalized and made applicable to the case with general discriminant functions, including piecewise-linear discriminant functions.**

Index Terms—**Accuracy of learning, adaptive pattern classifier, convergence of learning, learning under nonseparable pattern distribution, linear decision function, piecewise-linear decision function, rapidity of learning.**

I. Introduction

An ADAPTIVE pattern classifier system is one of the most typical learning or self-organizing systems. We shall first consider a simple classifier categorizing given patterns into two classes by a linear discriminant function which is automatically modified whenever a pattern is misclassified. Such a classifier has been investigated as the perceptron [1] or in the theory of threshold logic [2]. For the case where the patterns of the two classes are finite and linearly separable, various learning rules are known, and the discriminant function converges to the optimal one within a finite number of learning steps [1], [3], [4]. However, if the patterns are not linearly separable, it is not clear what is obtained using these rules. We shall treat the classifier in the general nonseparable case, assuming that the loss caused by misclassification is a monotonically increasing function of the distance between a pattern vector and the decision boundary. The loss which is some constant for an incorrect decision, can be approximated by choosing an appropriate function of the distance.

If we could use the knowledge of the probability distributions of the patterns of the two classes, the optimal linear discriminant function could be obtained by calculation. In the case of nonseparable patterns, most of the learning rules proposed so far are based on the estimation of the probability distributions. However, this needs a parametric treatment, that is, the distributions must be limited to those of a certain known kind whose distributions can be specified by a finite number of parameters. Moreover, the discriminant functions thus obtained depend directly on all of the past patterns so that they are not able to quickly follow the sudden change of the distributions. In order to avoid these shortcomings, we shall propose nonparametric learning procedures, by which the present discriminant function is modified according only to the present misclassified pattern.

The steepest-descent method is often used in order to minimize a known function. However, in our learning situation, we cannot obtain the descending directions of the average risk which we intend to minimize, because the probability distributions of the patterns are unknown. What we can utilize is the present pattern only, which obeys the unknown probability distribution. We shall associate a correction vector to each pattern in such a manner that the average of the correction vectors is in a descending direction. By the above correction, it is guaranteed that the discriminant function becomes better on the average, but in any given trial it may happen that the discriminant function becomes worse. This method may be called the probabilistic-descent method.

We shall prove that the discriminant function approaches a minimal one (this is the optimal if there is only one minimum) as near as desired, even if the distributions are overlapping. However, there is an important tradeoff between speed and accuracy of convergence. The speed and the accuracy of the classifier are explicitly obtained, and the performance of the classifier is theoretically clarified.

The learning rule of the simple classifier mentioned above can be generalized, and the learning rules of more complex classifiers are obtained. We first treat the classifier with multicategory or many-pattern classes. Next, we treat the classifier having piecewise-linear discriminant functions. We, then, generalize the theory and make it applicable to the general pattern classifiers having nonlinear discriminant functions. Finally, we discuss the adaptive determination of the constants contained in the learning rule, i.e., the learning of learning rule.

II. Optimal Linear Discriminant Function

Let C_1 and C_2 be the pattern classes or categories into which the given patterns are to be classified, and let a pattern be represented by an n-dimensional column

Manuscript received October 4, 1966; revised January 18, 1967. This work was supported by the Sakkokai Foundation, and was developed in RAAG (Research Association of Applied Geometry) activities. It is part of an investigation of information sciences using geometrical methods, which is called the theory of information spaces [5]–[7].

The author is with the Dept. Commun. Engrg., Kyushu University, Fukuoka, Japan.

Authorized licensed use limited to: KAUST. Downloaded on November 22,2021 at 11:04:14 UTC from IEEE Xplore. Restrictions apply.

vector $x=(x_1, x_2, \cdots, x_n)^t$, where t denotes the transposition. A linear equation

$$g(x) = \sum_{i=1}^{n} w_i x_i + w_0, \tag{1}$$

where w_i and w_0 are constants, divides the space into the following two regions:

$$V_1 = \{x \mid g(x) > 0\}, \quad V_2 = \{x \mid g(x) < 0\}. \tag{2}$$

When a pattern classifier decides that x belongs to C_1 or C_2 according as $x \in V_1$ or $x \in V_2$, respectively,[1] it is said to be linear and the function $g(x)$ is called the linear discriminant function. The boundary of V_1 and V_2 is a hyperplane D determined by $g(x)=0$, and it is called the decision surface. For simplicity's sake, let us augment x by adding 1 as the $(n+1)$st component and denote the $(n+1)$-dimensional vector $(x^t, 1)^t$ by $\boldsymbol{X}$. We also define an $(n+1)$-dimensional vector by $\boldsymbol{W}=(w_1, w_2, \cdots, w_n, w_0)$ and call it the weight vector. Then, the linear discriminant function is specified by $\boldsymbol{W}$ as

$$g(x) = \boldsymbol{W}^t \boldsymbol{X}. \tag{3}$$

Let the a priori probability of receiving a pattern which belongs to $C_\alpha (\alpha=1, 2)$ be p_α, and let the probability density function of the patterns of C_α be $p_\alpha(x)$.[2] Assuming that these quantities are known, we shall find the optimal linear discriminant function. The word "optimal" means to minimize the average risk, which is the expected value of the loss caused by misclassification. Let us denote by $l_{\alpha\beta}(x, \boldsymbol{W})$ the loss which we suffer when a pattern x belonging to C_α is mistakenly decided to belong to C_β $(\beta=\alpha)$by using the discriminant function of the weight vector $\boldsymbol{W}$. We call $l_{\alpha\beta}(x, \boldsymbol{W})$ the loss function. Since a pattern $x \in C_\alpha$ is misclassified when it is contained in $V_\beta(\beta \neq \alpha)$, and since the probability density of such a pattern is $p_\alpha p_\alpha(x)$, the average risk R is expressed by

$$R(\boldsymbol{W}) = \int_{V_1} f_1(x, \boldsymbol{W}) dX + \int_{V_2} f_2(x, \boldsymbol{W}) dX, \tag{4}$$

where we put

$$\begin{aligned} f_1(x, \boldsymbol{W}) &= p_2 p_2(x) l_{21}(x, \boldsymbol{W}) \\ f_2(x, \boldsymbol{W}) &= p_1 p_1(x) l_{12}(x, \boldsymbol{W}) \end{aligned} \tag{5}$$

and $dX = dx_1 dx_2 \cdots dx_n$. R is a function of $\boldsymbol{W}$, and the optimal weight vector is one which minimizes R.

We assume that $R(\boldsymbol{W})$ is differentiable, and that it has no local minima but the global minimum. In this case, the optimal $\boldsymbol{W}$ is given by

$$\nabla R(\boldsymbol{W}) = 0, \tag{6}$$

where ∇ is the gradient operator

$$\nabla = \left(\frac{\partial}{\partial w_1}, \frac{\partial}{\partial w_2}, \cdots, \frac{\partial}{\partial w_n}, \frac{\partial}{\partial w_0} \right). \tag{7}$$

Since $\boldsymbol{W}$ is contained in both f_α and V_α, ∇R consists of two terms: one concerning the gradient of the integrands f_α and the other concerning the integration over the boundary of V_α. By calculation, the following theorem is obtained [8].

Theorem 1: The optimal weight vector is given by

$$\begin{aligned} \nabla R = \frac{1}{w} \int_D \boldsymbol{X}(f_1 - f_2) dX + \int_{V_1} \nabla f_1 dX \\ + \int_{V_2} \nabla f_2 dX = 0, ^3 \end{aligned} \tag{8}$$

where

$$w = \left(\sum_{i=1}^{n} w_i \right)^{1/2}.$$

In the special case where the loss function does not depend on $\boldsymbol{W}$, ∇f_1 and ∇f_2 vanish identically. Hence, we have the following corollary.

Corollary 1: In the case where the loss function does not depend on $\boldsymbol{W}$, the optimal $\boldsymbol{W}$ is given by

$$\int_D \boldsymbol{X}(f_1 - f_2) dX = 0. \tag{9}$$

This is the same result as was obtained by Highleyman [9]. On the other hand, when the loss function identically vanishes on the decision surface D, the surface integral over D vanishes identically. Hence, we have the following corollary.

Corollary 2: When the loss function satisfies $l_{\alpha\beta}(x, \boldsymbol{W}) = 0$ on D, the optimal weight vector is given by

$$\int_{V_1} \nabla f_1 dX + \int_{V_2} \nabla f_2 dX = 0. \tag{10}$$

Let $d = |g(x)|/w$ be the distance from x to D, and let $l(d)$ be a monotonically increasing function satisfying $l(0)=0$. The loss function defined by

$$l_{12}(x, \boldsymbol{W}) = l_{21}(x, \boldsymbol{W}) = l(d)$$

satisfies the condition of Corollary 2. We call it the distance loss function. Denoting the $(n+1)$-dimensional vector $(w_1, w_2, \cdots, w_n, 0)$ by $\boldsymbol{w}$, we obtain

$$\nabla \left(\frac{g(x)}{w} \right) = \frac{1}{w^3} (w^2 \boldsymbol{X} - g(x) \boldsymbol{w}) = T\boldsymbol{X},$$

where T is the matrix defined by

$$T(\boldsymbol{W}) = \frac{1}{w^2} (w^2 E - \boldsymbol{w} \boldsymbol{W}^t), \tag{11}$$

[1] When x satisfies $g(x)=0$, any decision will do.

[2] Here we assume that x is a continuous variable. If x is discrete, replace integration with summation.

[3] Here $\int_D dX$ means an integration over D, i.e., an $(n-1)$-dimensional integration. Hence, when $n=1$, special treatment is required. In this case, D is a point and $\int_D dX$ denotes the value of the integrand on that point.

Authorized licensed use limited to: KAUST. Downloaded on November 22,2021 at 11:04:14 UTC from IEEE Xplore. Restrictions apply.

and E is the unit matrix. Noting that $l_{21}=l(g/w)$ and $l_{12}=l(-g/w)$, we can write ∇R in the form

$$\nabla R=\int_{V_1} p_2 p_2(\mathbf{x}) l'(d) T\mathbf{X} dX-\int_{V_2} p_1 p_1(\mathbf{x}) l'(d) T\mathbf{X} dX, \quad (12)$$

where the prime denotes the differentiation.

We treat hereafter the distance loss functions only. As an example, let us consider the family of loss functions $l(d)=d^k$. If we put $k=2$, the criterion is to minimize the sum of the squared distances of misclassified patterns, i.e., the least-square criterion. The criterion with $k\rightarrow\infty$ is to minimize the maximum of the distances of misclassified patterns, i.e., the minimax criterion. The criterion with $k\rightarrow 0$ is to minimize the percentage of misclassified patterns.

In classification problems, the most important criterion is to minimize the percentage of misclassified patterns. In this case, the loss is some constant for an incorrect decision and zero for a correct decision. Such a loss is not expressed by a distance loss function. Hence, we need to approximate it by a distance loss function.[4] For this purpose, we may adopt

$$l(d) = \arctan d/d_0,$$

$$l(d) = \begin{cases} 1 & d \geqq d_0, \\ d/d_0 & d < d_0, \end{cases}$$

etc., where d_0 is a sufficiently small constant. When the patterns are linearly separable, the optimal decisions based on a distance loss and a constant loss are exactly identical.

III. Learning Rule and Convergence Theorem

We have derived the equation of the optimal weight vector, assuming that the probability structures p_α and $p_\alpha(\mathbf{x})$ are known. In many practical cases, however, they are unknown and varying with time. Moreover, even if they are known, it is usually difficult to solve the equation. This fact suggests that the weight vector is determined step by step utilizing the information of the input patterns. We propose a learning rule by which the weight vector $\mathbf{W}_i$ at time i is modified to $\mathbf{W}_{i+1}$ by referring to the input pattern $\mathbf{x}_i$ at time i only. The computation by this rule is very simple and there is no need of storing the information of the past input data, nor assuming the type of the distributions.

Let the correction vector of $\mathbf{W}_i$, be $\delta\mathbf{W}_i$, which depends on the present input pattern $\mathbf{x}_i$, and the new weight vector at time $i+1$ be $\mathbf{W}_{i+1}=\mathbf{W}_i+\delta\mathbf{W}_i$. We put

$$\delta\mathbf{W}_i = \epsilon C\mathbf{H}(\mathbf{x}_i, \mathbf{W}_i), \quad (13)$$

where ϵ is a small positive constant and C is a positive-definite matrix. We call ϵ the learning constant. Assuming that the correction takes place only when $\mathbf{x}_i$ is mistakenly classified, we can put

$$\mathbf{H}(\mathbf{x}, \mathbf{W}) = \begin{cases} \mathbf{H}_1(\mathbf{x}, \mathbf{W}), & \text{when } \mathbf{W}^t\mathbf{X}<0 \text{ and } \mathbf{x}\in C_1, \\ \mathbf{H}_2(\mathbf{x}, \mathbf{W}), & \text{when } \mathbf{W}^t\mathbf{X}>0 \text{ and } \mathbf{x}\in C_2, \\ 0, & \text{when } \mathbf{x} \text{ is correctly classified.} \end{cases} \quad (14)$$

We call $\mathbf{H}_1(\mathbf{x}, \mathbf{W})$ and $\mathbf{H}_2(\mathbf{x}, \mathbf{W})$ the learning functions, and they will be determined in the following.

When ϵ is sufficiently small, the increment of the average risk is $\delta R=\delta\mathbf{W}^t\nabla R(\mathbf{W})$ for one step of learning, neglecting higher order terms of ϵ. In order to design an effective learning system, it is suggested that $\delta\mathbf{W}$ should be chosen so as to make δR always negative [10], e.g., $\delta\mathbf{W}=-\nabla R$, like the steepest descent method in nonlinear-programming problems. However, it is impossible to make $\delta\mathbf{W}$ equal to $-\nabla R$, since ∇R depends on the unknown quantities p_α, $p_\alpha(\mathbf{x})$. Therefore, we try to make negative the average of δR over all possible $\mathbf{x}$, i.e., $\overline{\delta R}=\overline{\delta\mathbf{W}}^t R\nabla<0$, where the bar denotes the averaging over all $\mathbf{x}\in C_1$, C_2. Since δR is negative only as the average, this method may be called the probabilistic-descent method.

Lemma: For the following learning functions

$$\mathbf{H}_1 = -\mathbf{H}_2 = l'(d)T(\mathbf{W})\mathbf{X}, \quad (15)$$

the relation $\overline{\delta R}\leqq 0$ holds, and the equality holds when and only when $\mathbf{W}$ is the optimal weight vector.

Proof: Since $\delta\mathbf{W}=\epsilon C\mathbf{H}_\alpha$ when a pattern $\mathbf{x}\in C_\alpha$ is misclassified, the average of the correction vectors is

$$\overline{\delta\mathbf{W}} = \epsilon C\left\{\int_{V_2} p_1 p_1(\mathbf{x})\mathbf{H}_1 dX + \int_{V_1} p_2 p_2(\mathbf{x})\mathbf{H}_2 dX\right\}.$$

Substituting (15), we can derive

$$\overline{\delta W} = -\epsilon C\nabla R. \quad (16)$$

Since C is positive-definite, we get

$$\overline{\delta R} = -\epsilon\nabla R^t C\nabla R \leqq 0. \quad (17)$$

The equality holds only when $\nabla R=0$, which is satisfied by the optimal weight vector only.

We shall consider a classifier with the above-mentioned learning rule. Let the classifier start with an initial weight vector $\mathbf{W}_1$ at time l, under the condition that the probability distributions are fixed. Since the weight vector $\mathbf{W}_i$ depends on the sequence of the input patterns $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{i-1}$ randomly selected from the distributions p_α, $p_\alpha(\mathbf{x})$, it is also a random variable vector. Let its density function be $q_i(\mathbf{W})$. Then the expected value of the average risk at time i, i.e., after $i-1$ steps of learning, is

$$\overline{R}_i = \int q_i(\mathbf{W})R(\mathbf{W})d\mathbf{W}, \quad (18)$$

[4] In the case where the loss is some constant for an incorrect decision, there is no exact learning rule of the nonparametric type. In this case, the optimal decision boundary is one satisfying (9) of Corollary 1. However, the probability of the appearance of the patterns on a hyperplane D is 0, because the measure of D is 0. Hence, we are obliged to obtain the information about the distribution of the patterns on D from the patterns around D. For this purpose, we use a distance loss, and approximate the constant loss by it. In the parametric case where the type of the distributions is known, the distribution on D can be estimated using the patterns of the whole space.

Authorized licensed use limited to: KAUST. Downloaded on November 22,2021 at 11:04:14 UTC from IEEE Xplore. Restrictions apply.

where $dW = dw_1 dw_2 \cdots dw_n dw_0$. The increment of $\overline{R}_i$ by a step of learning is

$$\delta\overline{R}_i = \overline{R}_{i+1} - \overline{R}_i \tag{19}$$

which is the expectation of $\delta R(\boldsymbol{x}_i, \boldsymbol{W}_i)$ with respect to both $\boldsymbol{x}_i$ and $\boldsymbol{W}_i$. Hence

$$\delta\overline{R}_i = -\epsilon \int \boldsymbol{\nabla} R^t C \boldsymbol{\nabla} R q_i(\boldsymbol{W}) dW \leqq 0. \tag{20}$$

Consequently, $\overline{R}_i$ is proved to be monotonically nonincreasing. Obviously, $\overline{R}_i \geqq 0$, since the loss function is nonnegative. Therefore, the sequence $\overline{R}_i$ converges, and it follows that $\lim_{i\to\infty} \delta\overline{R}_i = 0$. However, we have already proved that $\boldsymbol{\nabla} R^t C \boldsymbol{\nabla} R > 0$ holds for all but the optimal weight vector $\boldsymbol{W}_{\text{op}}$. Therefore, roughly speaking, it is expected that

$$\lim_{i\to\infty} q_i(\boldsymbol{W}) = \delta(\boldsymbol{W} - \boldsymbol{W}_{\text{op}}) \tag{21}$$

holds, where $\delta(\boldsymbol{W})$ is the delta function.

Let us prove the convergence theorem more exactly.[5]

Theorem 2: For any μ, the probability that $|\boldsymbol{W}_i - \boldsymbol{W}_{\text{op}}| \geqq \mu$ can be made as small as desired for sufficiently large i, by choosing a sufficiently small learning constant ϵ.

Proof: Let $M_\mu{}^i(\epsilon)$ be the probability that the weight vector at time i is still apart from the optimal one further than μ, i.e.,

$$M_\mu{}^i(\epsilon) = \Pr\left\{ |\boldsymbol{W}_i - \boldsymbol{W}_{\text{op}}| \geqq \mu \right\}, \tag{22}$$

and let $M_\mu(\epsilon) = \lim_{i\to\infty} M_\mu{}^i(\epsilon)$.[6] Then we need only to prove

$$\lim_{\epsilon\to 0} M_\mu(\epsilon) = 0. \tag{23}$$

Expanding $\delta R(\delta R(\boldsymbol{x}, \boldsymbol{W})$, we obtain

$$\delta R(\boldsymbol{x}, \boldsymbol{W}) = \delta \boldsymbol{W}^t \boldsymbol{\nabla} R + \tfrac{1}{2}(\delta \boldsymbol{W}^t \boldsymbol{\nabla}{}^t R \delta \boldsymbol{W}) + \cdots.$$

Averaging it over all $\boldsymbol{x}$, we get

$$\overline{\delta R} = -\epsilon \boldsymbol{\nabla} R^t C \boldsymbol{\nabla} R + \frac{\epsilon^2}{2} \operatorname{tr}\left\{ \overline{\delta \boldsymbol{W} \delta \boldsymbol{W}^t} \boldsymbol{\nabla}^t R \right\} + 0(\epsilon^3),$$

where tr denotes the trace of a matrix. Hence, for sufficiently small ϵ, there exists a positive constant K, for which the inequality

$$\overline{\delta R} \leqq -\epsilon r + K\epsilon^2 \tag{24}$$

holds, where we put

$$r(\boldsymbol{W}) = \boldsymbol{\nabla} R^t C \boldsymbol{\nabla} R. \tag{25}$$

Let U_μ be the set of the $\boldsymbol{W}$'s defined by

$$U_\mu = \left\{ \boldsymbol{W} \,\middle|\, |\boldsymbol{W} - \boldsymbol{W}_{\text{op}}| \geqq \mu \right\}. \tag{26}$$

Its complement is a neighborhood of $\boldsymbol{W}_{\text{op}}$. $M_\mu{}^i$ can be written as

$$M_\mu{}^i = \int_{U_\mu} q_i(\boldsymbol{W}) dW. \tag{27}$$

Next, we define another set U_λ' by

$$U_\lambda' = \left\{ \boldsymbol{W} \,\middle|\, r(\boldsymbol{W}) \geqq \lambda \right\}. \tag{28}$$

Since r is a continuous function of $\boldsymbol{W}$ and is equal to 0 when and only when $\boldsymbol{W} = \boldsymbol{W}_{\text{op}}$, for any U_μ, there exists a positive constant $\lambda(\mu)$ for which $U_\lambda' \supset U_\mu$ holds. Putting

$$M_\lambda'{}^i = \int_{U_\lambda'} q_i(\boldsymbol{W}) dW, \tag{29}$$

we obtain the inequality $M_\mu{}^i \leqq M_\lambda'{}^i$. By averaging (24) with respect to $q_i(\boldsymbol{W})$ and taking account of the relation

$$\begin{aligned} \int r(\boldsymbol{W}) q_i(\boldsymbol{W}) dW &\geqq \int_{U_\lambda'} r(\boldsymbol{W}) q_i(\boldsymbol{W}) dW \\ &\geqq \lambda \int_{U_\lambda'} q_i(\boldsymbol{W}) dW = \lambda M_\lambda'{}^i \geqq \lambda M_\mu{}^i, \end{aligned}$$

we can prove the inequality

$$\delta\overline{R}_i \leqq -\epsilon\lambda M_\mu{}^i + K\epsilon^2. \tag{30}$$

Summing up the both sides of the above relation over i from 1 to N, dividing them by N, and taking the limit $N\to\infty$, we derive the relation $0 \leqq -\epsilon\lambda M_\mu + K\epsilon^2$. Consequently, we get the required relation $\lim_{\epsilon\to 0} M_\mu = 0$.

In the special case where the patterns C_1 and C_2 are linearly separable, we can prove that a separating hyperplane is obtained with probability one by using the above learning rule.

IV. Convergence Rate and Accuracy of Learning

Let $f(\boldsymbol{W})$ be a function of $\boldsymbol{W}$. When $\boldsymbol{W}$ is determined by learning, the expected value of $f(\boldsymbol{W})$ at time i is defined by

$$\overline{f(\boldsymbol{W})_i} = \int f(\boldsymbol{W}) q_i(\boldsymbol{W}) dW. \tag{31}$$

The aspect of the learning process will be clarified by studying how $\overline{f(\boldsymbol{W})_i}$ changes as i increases. We put

$$p(\mathbf{x}) = \begin{cases} p_1 p_1(\mathbf{x}), & \mathbf{x} \in V_2, \\ p_2 p_2(\mathbf{x}), & \mathbf{x} \in V_1, \end{cases} \tag{32}$$

and

$$B(\boldsymbol{W}) = \overline{\boldsymbol{H}\boldsymbol{H}^t} = \int \boldsymbol{H}\boldsymbol{H}^t p(x) dX, \quad B_0 = B(\boldsymbol{W}_{\text{op}}). \tag{33}$$

Lemma: The increment of $\bar{f}_i$ due to a step of learning is given by

$$\bar{f}_{i+1} - \bar{f}_i = -\epsilon(\overline{\boldsymbol{\nabla} f^t C \boldsymbol{\nabla} R)_i} + \epsilon^2 \operatorname{tr}(\overline{CBC^t \boldsymbol{\nabla}\boldsymbol{\nabla}^t f)_i} + 0(\epsilon^3). \tag{34}$$

[5] In the case where there are many $\boldsymbol{W}$'s satisfying $\boldsymbol{\nabla} R = 0$, $\boldsymbol{W}_{\text{op}}$ in the following theorem should be regarded as the set of such weight vectors and the theorem guarantees only that the weight vector converges to one of such vectors.

[6] For the convergence of $M_\mu{}^i(\epsilon)$, see Doob [1].

Authorized licensed use limited to: KAUST. Downloaded on November 22,2021 at 11:04:14 UTC from IEEE Xplore. Restrictions apply.

Proof: Let $\boldsymbol{W}_i$ be the weight vector at time i. If a pattern $\mathbf{x}$ is presented, $\boldsymbol{W}_i$ changes to $\boldsymbol{W}=\boldsymbol{W}_i+\delta\boldsymbol{W}_i(\mathbf{x}, \boldsymbol{W}_i)$. Since $\mathbf{x}$ obeys the probability distributions p_α, $p_\alpha(\mathbf{x})$, $\boldsymbol{W}$ is also a random variable and its density function $q(\boldsymbol{W})$ is related to that of $\mathbf{x}$ by $q(\boldsymbol{W})dW=p(\mathbf{x})dX$. Since the probability distribution of $\boldsymbol{W}_i$ is $q_i(\boldsymbol{W})$, $q_{i+1}(\boldsymbol{W})$ is obtained by averaging it with respect to $\boldsymbol{W}_i$, i.e.,

$$q_{i+1}(\boldsymbol{W})dW = dX\int q_i(\boldsymbol{W}_i)p(\mathbf{x})dW_i, \tag{35}$$

where $\mathbf{x}$ is considered a function of $\boldsymbol{W}$ and $\boldsymbol{W}_i$. Using (35), we obtain $\bar{f}_{i+1}$ as

$$\begin{aligned}\bar{f}_{i+1} &= \int q_{i+1}(\boldsymbol{W})f(\boldsymbol{W})dW \\ &= \int q_i(\boldsymbol{W}_i)f(\boldsymbol{W}_i+\delta\boldsymbol{W}_i)p(x)dXdW_i.\end{aligned} \tag{36}$$

Expanding $f(\boldsymbol{W}_i+\delta\boldsymbol{W}_i)$, integrating with respect to dX, and taking (16) and (33) into account, we derive (34).

In the lemma, the function f may be a vector-valued or matrix-valued function. Hence, if we put $f(\boldsymbol{W})=\boldsymbol{W}$, the expected value of the weight vector $\overline{\boldsymbol{W}}_i=\bar{f}_i$ is derived from (34). As has been proved, $\overline{\boldsymbol{W}}_i$ converges to the optimal vector $\boldsymbol{W}_{\text{op}}$. Now we can examine the manner in which it converges to $\boldsymbol{W}_{\text{op}}$. In this case we can expand R around $\boldsymbol{W}_{\text{op}}$,

$$\begin{aligned}R(\boldsymbol{W}) = R(\boldsymbol{W}_{\text{op}}) &+ \tfrac{1}{2}(\boldsymbol{W}-\boldsymbol{W}_{\text{op}})^tA(\boldsymbol{W}-\boldsymbol{W}_{\text{op}}) \\ &+ 0(|\,\boldsymbol{W}-\boldsymbol{W}_{\text{op}}|^3),\end{aligned} \tag{37}$$

where

$$A = \boldsymbol{\nabla\nabla}^tR|_{\overline{W}_{\text{op}}}.$$

We shall consider the neighborhood of $\boldsymbol{W}_{\text{op}}$, neglecting the last term.

Theorem 3: The expected value of the weight vector $\boldsymbol{W}_i$ is given by

$$\overline{\boldsymbol{W}}_i = \boldsymbol{W}_{\text{op}} + (E-\epsilon CA)^{i-1}(\boldsymbol{W}_1-\boldsymbol{W}_{\text{op}}). \tag{38}$$

Proof: For $f(\boldsymbol{W})=\boldsymbol{W}$, it is easily shown that $\boldsymbol{\nabla}f=E$ and $\boldsymbol{\nabla\nabla}^tf=0$. By applying the lemma to this case, we obtain

$$\overline{\boldsymbol{W}}_{i+1} = \overline{\boldsymbol{W}}_i - \epsilon C\overline{\boldsymbol{\nabla}R}_i, \tag{39}$$

where the term $0(\epsilon^3)$ is neglected. By using (37), this reduces to the linear difference equation

$$\overline{\boldsymbol{W}}_{i+1} = (E-\epsilon CA)\overline{\boldsymbol{W}}_i + \epsilon CA\boldsymbol{W}_{\text{op}}. \tag{40}$$

This can easily be solved, giving (38) as the solution.

Let $\lambda_0>0$ be the minimum eigenvalue of the matrix CA. Then the corresponding eigenvector shows the direction of the slowest convergence, and the time constant in that direction is $\epsilon\lambda_0$. We have thus obtained the expected weight vector $\overline{\boldsymbol{W}}_i$. However, the actual weight vector is not necessarily identical with it. The difference between the actual vector and the expected is evaluated by the covariance matrix Σ_i:

$$\Sigma_i = \{\overline{(\boldsymbol{W}-\overline{\boldsymbol{W}}_i)(\boldsymbol{W}-\overline{\boldsymbol{W}}_i)^t}\}_i = (\overline{\boldsymbol{WW}^t})_i - \overline{\boldsymbol{W}}_i\overline{\boldsymbol{W}}_i^t. \tag{41}$$

Since $\overline{\boldsymbol{W}}_i$ converges to $\boldsymbol{W}_{\text{op}}$, Σ_i can be considered to represent the degree of the accuracy of learning.

Theorem 4: The covariance matrix Σ_i of the weight vector at time i is

$$\Sigma_i = 2\epsilon\{\tilde{E}-(\tilde{E}-\epsilon\tilde{S})^{i-1}\}(\tilde{S})^{-1}CB_0C^t, \tag{42}$$

where $\tilde{E}$ is the identity operator and $\tilde{S}$ is the linear operator transforming an arbitrary matrix M by

$$\tilde{S}M = 2(CAM)^s,[7] \tag{43}$$

the superscript s denoting the symmetric part of a matrix.

Proof: Applying the lemma to $f(\boldsymbol{W})=\boldsymbol{WW}^t$, we obtain

$$\begin{aligned}(\overline{\boldsymbol{WW}^t})_{i+1} = (\overline{\boldsymbol{WW}^t})_i &- 2\epsilon\{\overline{CA(\boldsymbol{W}-\boldsymbol{W}_{\text{op}})\boldsymbol{W}^t}\}_i^s \\ &+ 2\epsilon^2(CB_0C^t).\end{aligned}$$

Subtracting

$$\overline{\boldsymbol{W}}_{i+1}\overline{\boldsymbol{W}}_{i+1}^t = \{(E-2\epsilon CA)\overline{\boldsymbol{W}}_i\overline{\boldsymbol{W}}_i^t\}^s + 2\epsilon(CA\boldsymbol{W}_{\text{op}}\overline{\boldsymbol{W}}_i^t)^s,$$

we obtain the difference equation

$$\Sigma_{i+1} = (\tilde{E}-\epsilon\tilde{S})\Sigma_i + 2\epsilon^2CB_0C^t. \tag{44}$$

Since the classifier started with a fixed initial vector $\boldsymbol{W}_1$, the initial covariance matrix Σ_1 is equal to 0. The corresponding solution of (44) is given by (42). The final accuracy of learning is represented by

$$\lim_{i\to\infty}\Sigma_i = 2\epsilon(\tilde{S})^{-1}CB_0C^t. \tag{45}$$

By the above two theorems, it has been shown that the convergence rate of learning is represented by the matrix ϵCA, while the accuracy is given by the matrix $2\epsilon(\tilde{S})^{-1}CB_0C^t$. The constants ϵ and C of the adaptive classifier should be determined by taking these relations into account. If we can put $C=A^{-1}$, the convergence rate of $\overline{\boldsymbol{W}}_i$ is uniform for all directions. On the other hand, if we can determine C in such a way that $(\tilde{S})^{-1}CB_0C^t=E$ holds, the deviation of $\boldsymbol{W}_i$ from $\boldsymbol{W}_{\text{op}}$ becomes isotropic. The larger ϵ we choose, the faster the convergence becomes, and the worse the accuracy. On the contrary, the smaller ϵ we choose, the more accurate the learning becomes, and the slower the convergence.

[7] $\tilde{S}$ can be considered a tensor having four indexes. Using the tensorial notation, $\tilde{S}M$ is represented by $(C_i{}^jA_j{}^k\delta_m{}^n+C_m{}^jA_j{}^k\delta_i{}^n)M_{nk}$, where $\delta_m{}^n$ is the Kronecker delta and Einstein's summation convention is used. Hence, $\tilde{S}$ is a tensor whose components are

$$S_i{}^{kn}{}_m = C_i{}^jA_j{}^k\delta_m{}^n + C_m{}^jA_j{}^k\delta_i{}^n.$$

Authorized licensed use limited to: KAUST. Downloaded on November 22,2021 at 11:04:14 UTC from IEEE Xplore. Restrictions apply.

V. Dynamical Behavior of Adaptive Classifiers

The probability structures of the input patterns are not necessarily fixed but may vary from time to time. The optimal vector will vary according to these disturbances. We shall briefly analyze the manner how the weight vector follows the moving optimal vector under our learning rule.

Let $\boldsymbol{W}_0+\boldsymbol{D}_i$ be the optimal weight vector at time i, $\boldsymbol{D}_i$ denoting the fluctuation. In this case the matrix A also depends on i. We denote A at time i by A_i. Then

$$\overline{\boldsymbol{W}}_{i+1} = \overline{\boldsymbol{W}}_i - \epsilon C A_i \overline{\boldsymbol{W}}_i + \epsilon C A_i(\boldsymbol{W}_0 + \boldsymbol{D}_i) \tag{46}$$

is derived instead of (40). Although we can solve (46) explicitly, we shall assume $A_i = A$ for all i for simplicity's sake. In this case, the solution is

$$\begin{aligned}\overline{\boldsymbol{W}}_i = \boldsymbol{W}_0 &+ (E - \epsilon CA)^{i-1}(\boldsymbol{W}_1 - \boldsymbol{W}_0) \\ &+ \epsilon \sum_{k=1}^{i-1} (E - \epsilon CA)^{i-k-1} CA\boldsymbol{D}_k.\end{aligned} \tag{47}$$

The second term, depending on the initial weight vector $\boldsymbol{W}_1$, is transient. The third term depends on the deviation $\boldsymbol{D}_i$. From this, we see that the present deviation $\boldsymbol{D}$ causes the deviation $\epsilon(E-\epsilon CA)^{i-1}\ CA\boldsymbol{D}$ of the weight vector of i times later. Hence, the matrices

$$I_i = \epsilon(E - \epsilon CA)^{i-1}CA \tag{48}$$

are considered to represent the impulse response of the classifier. The step response of the classifier is given by the matrices

$$S_i = E - (E - \epsilon CA)^{i-1}. \tag{49}$$

As an example, let us consider the case in which the optimal vector changes periodically. We put

$$\boldsymbol{D}_i = \boldsymbol{D} \sin \omega i, \tag{50}$$

where the period is $2\pi/\omega$ and assumed to be large. We need only to solve the case where $\boldsymbol{D}$ is an eigenvector of CA, $CA\boldsymbol{D} = \lambda \boldsymbol{D}$, because the solution of general cases are obtained by superposition.

A particular solution of the difference equation is written as

$$\overline{\boldsymbol{W}}_i = \boldsymbol{W}_0 + a\boldsymbol{D} \sin(\omega i + \theta),$$

where the transient term is put equal to 0. Substituting this in (47), we derive the following equations:

$$\begin{aligned} a\{\cos(\omega + \theta) - (1 - \epsilon\lambda)\cos\theta\} - \epsilon\lambda &= 0, \\ \sin(\omega + \theta) - (1 - \epsilon\lambda)\sin\theta &= 0, \end{aligned} \tag{51}$$

from which the unknown parameters a and θ can be determined. Neglecting the higher order terms of ϵ and ω, we obtain

$$a = 1/\sqrt{1 + \alpha^2}, \qquad \tan\theta = -\alpha, \tag{52}$$

where we put $\alpha = \omega/\epsilon\lambda$. Accordingly, the stationary solution is

$$\overline{\boldsymbol{W}}_i = \boldsymbol{W}_0 + \frac{1}{\sqrt{1 + \alpha^2}} \boldsymbol{D} \sin(\omega i - \theta), \tag{53}$$

where α is considered small. This shows the frequency response of the system.

From this, we see that, when the optimal vector changes sinusoidally with frequency $\omega/2\pi$, the weight vector follows it with the amplitude divided by $\sqrt{1+\alpha^2}$ and with the phase shifted by $-\alpha$. Therefore, for $\omega \ll \epsilon\lambda$, we may say that the classifier is able to trace the change well.

VI. Generalization

Multicategory Classifiers

We have so far assumed that there are only two categories C_1 and C_2. Our theory can easily be generalized to the case with many categories or pattern classes $C_1, C_2, \cdots, C_m$. In this case, we use m discriminant functions

$$g_\alpha(\mathbf{x}) = \hat{\boldsymbol{W}}_\alpha{}^t \boldsymbol{X}, \qquad \alpha = 1, 2, \cdots, m \tag{54}$$

and decide that a pattern $\mathbf{x}$ belongs to C_α when and only when $g_\alpha(\mathbf{x}) > g_\beta(\mathbf{x})$ for all $\beta(\neq\alpha)$. We need to obtain a set of m weight vectors $\hat{\boldsymbol{W}}_\alpha$ by learning.

For each pattern $\mathbf{x} \in C_\alpha$, we can define a set N_α of integers by

$$N_\alpha = \{\beta \mid g_\beta(\mathbf{x}) > g_\alpha(\mathbf{x})\}.$$

For a correctly classified $\mathbf{x} \in C_\alpha$, N_α is the null set. It is natural to define the loss caused by misclassification of a signal $\mathbf{x} \in C_\alpha$ by

$$l_\alpha(\mathbf{x}) = \max_{\beta \in N_\alpha} l(d_{\alpha\beta}), \tag{55}$$

where $d_{\alpha\beta}$ denotes the distance from $\mathbf{x}$ to the hyperplane defined by $g_\alpha(\mathbf{x}) = g_\beta(\mathbf{x})$. We can write

$$d_{\alpha\beta} = \frac{g_\alpha(\mathbf{x}) - g_\beta(\mathbf{x})}{\hat{w}_{\alpha\beta}} \tag{56}$$

where $\hat{w}_{\alpha\beta}$ is the length of the vector $\hat{\boldsymbol{w}}_\alpha - \hat{\boldsymbol{w}}_\beta$ and $\hat{\boldsymbol{w}}_\alpha = (\hat{w}_{\alpha 1}, \hat{w}_{\alpha 2}, \cdots, \hat{w}_{\alpha n}, 0)^t$. Obviously, when a pattern is correctly classified, the corresponding N_α is null and $l_\alpha(\mathbf{x}) = 0$. The average risk accompanied with the set of m weight vectors $\hat{\boldsymbol{W}}_1, \cdots, \hat{\boldsymbol{W}}_m$ is expressed as

$$R(\hat{\boldsymbol{W}}_1, \cdots, \hat{\boldsymbol{W}}_m) = \sum_{\alpha=1}^{m} \int p_\alpha p_\alpha(\mathbf{x}) l_\alpha(\mathbf{x}) dX. \tag{57}$$

Denoting the gradient operator with respect to $\hat{\boldsymbol{W}}_\alpha$ by $\boldsymbol{\nabla}_\alpha$, we can obtain the following relation:

$$\begin{aligned}\boldsymbol{\nabla}_\beta R = \sum_{\alpha \neq \beta} \Big\{ &\int_{V_{\beta\alpha}} p_\alpha p_\alpha(\mathbf{x}) l'(d_{\alpha\beta}) T_{\alpha\beta} \boldsymbol{X} dX \\ &- \int_{V_{\alpha\beta}} p_\beta p_\beta(\mathbf{x}) l'(d_{\alpha\beta}) T_{\alpha\beta} \boldsymbol{X} dX,\end{aligned} \tag{58}$$

Authorized licensed use limited to: KAUST. Downloaded on November 22,2021 at 11:04:14 UTC from IEEE Xplore. Restrictions apply.

where

$$V_{\beta\alpha} = \{\mathbf{x} \mid \max_{\gamma \in N_\alpha} d_{\gamma\alpha} = d_{\beta\alpha}, \tag{59}$$

and

$$T_{\alpha\beta} = \frac{1}{\hat{w}_{\alpha\beta}{}^3} \{\hat{w}_{\alpha\beta}{}^2 E - (\hat{\boldsymbol{w}}_\alpha - \hat{\boldsymbol{w}}_\beta)(\hat{\boldsymbol{W}}_\alpha - \hat{\boldsymbol{W}}_\beta)^t\}. \tag{60}$$

Let us modify the weight vectors by

$$\delta\hat{\boldsymbol{W}}_\alpha = \epsilon C \hat{\boldsymbol{H}}_\alpha(\mathbf{x}; \hat{\boldsymbol{W}}_1, \cdots, \hat{\boldsymbol{W}}_m), \tag{61}$$

when $\mathbf{x}$ is mistakenly classified. Then we can prove the following convergence theorem.

Theorem 5: By using the following learning functions,

$$\hat{\boldsymbol{H}}_\gamma = \begin{cases} -l'(d_{\gamma\alpha}) T_{\gamma\alpha} \boldsymbol{X}, & \text{when } \mathbf{x} \in C_\alpha \text{ is contained in } V_{\gamma\alpha}, \\ l'(d_{\gamma\alpha}) T_{\gamma\alpha} \boldsymbol{X}, & \text{when } \mathbf{x} \in C_\gamma \text{ is contained in } V_{\gamma\alpha}, \\ 0, & \text{when } \mathbf{x} \text{ is correctly classified,} \end{cases} \tag{62}$$

the probability that the set of the weight vectors approaches the optimal one as near as desired, can be made as near to l as desired by choosing a sufficiently small learning constant ϵ.

Classifiers with Piecewise-Linear Discriminant Functions

Although the linear discriminant function is realizable with technical ease, it is a very restricted one. Hence, we consider the piecewise-linear discriminant functions [3], [12], which are much more general but also realizable with technical ease. We shall generalize the learning rule and make it applicable to the classifiers whose discriminant functions are convex piecewise-linear.

We treat the case with two pattern classes. Let us consider m linear functions

$$g_i(\mathbf{x}) = \hat{\boldsymbol{W}}_i{}^t \boldsymbol{X}, \qquad i = 1, 2, \cdots, m$$

and decide that $\mathbf{x}$ belongs to C_1 when $\max_i g_i(\mathbf{x}) > 0$ and to C_2 when $g_i(\mathbf{x}) < 0$ for all i. Such a decision is known as a convex piecewise-linear decision [12] or a threshold-or decision [13].

When a pattern $\mathbf{x} \in C_1$ is misclassified into C_2, all of $g_i(\mathbf{x})$ are negative. It is natural to define the loss by

$$l(\mathbf{x}) = \min_i l(d_i), \tag{63}$$

where $d_i = |g_i(\mathbf{x})| / \hat{w}_i$ is the shortest distance between $\mathbf{x}$ and the boundary of $V_2 = \{\mathbf{x} | g_i < 0 \text{ for all } i\}$ and $\hat{w}_i$ is the length of $\hat{w}_i$, because $\mathbf{x}$ is correctly classified if any one of $g_i(\mathbf{x})$ is positive. On the other hand, when a pattern $\mathbf{x} \in C_2$ is misclassified, we can define a nonempty set N of integers by $N = \{i | g_i(\mathbf{x}) > 0\}$. In this case, we define the loss by

$$l(\mathbf{x}) = \max_{i \in N} l(d_i),^8 \tag{64}$$

considering that $\mathbf{x}$ cannot be correctly classified unless all of $g_i(\mathbf{x})$, $i \in N$ are negative. Defining the sets V_{1i} and V_{2i} by

$$V_{1i} = \{\mathbf{x} \mid \max_{j \in N} d_j = d_i\},$$

$$V_{2i} = \{\mathbf{x} \mid \min_j d_j = d_i \quad \text{and} \quad g_j < 0 \text{ for all } j\}, \tag{65}$$

the average risk can be written as

$$R = \sum_i \left(\int_{V_{2i}} p_1 p_1(\mathbf{x}) l(d_i) dX + \int_{V_{1i}} p_2 p_2(\mathbf{x}) l(d_i) dX \right). \tag{66}$$

The gradient of the average risk is expressed as follows:

$$\nabla_i R = -\sum_i \left\{ \int_{V_{2i}} p_1 p_1(\mathbf{x}) l'(d_i) T_i \boldsymbol{X} dX - \int_{V_{1i}} p_2 p_2(\mathbf{x}) l'(d_i) T_i \boldsymbol{X} dX \right\}, \tag{67}$$

where

$$T_i = \frac{1}{\hat{w}_i{}^3} (\hat{w}_i{}^2 E - \hat{\boldsymbol{w}}_i \hat{\boldsymbol{W}}_i{}^t), \tag{68}$$

and the fact that the term concerning the integration over the boundaries of the V_{1i}'s and the V_{2i}'s vanishes as a whole is taken into account.

Let us consider a learning rule, by which the set of the weight vectors are modified by $\delta\hat{\boldsymbol{W}}_i = \epsilon C \hat{\boldsymbol{H}}_i(\mathbf{x}; \hat{\boldsymbol{W}}_j)$ when a pattern $\mathbf{x}$ is misclassified. In this case, we can prove the following convergence theorem.

Theorem 6: In the case of the convex piecewise-linear decision, by using the following learning functions:

$$\hat{\boldsymbol{H}}_i = \begin{cases} l'(d_i) T_i \boldsymbol{X}, & \text{when } \mathbf{x} \in C_1 \text{ is contained in } V_{2i}, \\ -l'(d_i) T_i \boldsymbol{X}, & \text{when } \mathbf{x} \in C_2 \text{ is contained in } V_{1i}, \\ 0, & \text{when } \mathbf{x} \text{ is correctly classified,} \end{cases} \tag{69}$$

the probability that the set of the weight vectors approaches the optimal one as near as desired, can be made as near to l as desired by choosing a sufficiently small learning constant ϵ.

General Adaptive Classifiers

Here we consider a general adaptive classifier which classifies a given pattern into m classes C_α ($\alpha = 1, \cdots, m$) and whose discriminant function $g_\alpha(\mathbf{x})$ is specified by a set of parameters $\boldsymbol{\theta}_{\alpha 1}, \cdots, \boldsymbol{\theta}_{\alpha k}$. We represent the parameters by a vector $\boldsymbol{\theta}_\alpha = (\boldsymbol{\theta}_{\alpha 1}, \cdots, \boldsymbol{\theta}_{\alpha k})^t$, and denote by $g_\alpha(\mathbf{x}, \boldsymbol{\theta}_\alpha)$ the discriminant function specified by $\boldsymbol{\theta}_\alpha$. $g_\alpha(\mathbf{x}, \boldsymbol{\theta}_\alpha)$ need not be linear nor piecewise-linear with respect to $\boldsymbol{\theta}_\alpha$. For simplicity's sake, we unite the m vectors $\boldsymbol{\theta}_\alpha$ and denote it by an mk-dimensional vector $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1{}^t, \cdots, \boldsymbol{\theta}^{mt})^t$. We call it the decision vector. By specifying a decision vector $\boldsymbol{\Theta}$, the decision is completely determined, that is, when a pattern $\mathbf{x}$ is contained in V_α,

$$V_\alpha = \{\mathbf{x} \mid \max_\beta g(\mathbf{x}, \boldsymbol{\theta}_\beta) = g_\alpha(\mathbf{x}, \boldsymbol{\theta}_\alpha)\}, \tag{70}$$

it is considered to belong to C_α.

[8] We may adopt $l(\mathbf{x}) = \sum_{i \in N} l(d_i)$. All of the following discussions are valid, if we replace the definition of V_{1i} by $V_{1i} = \{\mathbf{x} | g_i(\mathbf{x}) > 0\}$. Generally speaking, we may adopt $l(\mathbf{x}) = \sum_{i \in N} s_i l(d_i)$. See the following subsection.

Authorized licensed use limited to: KAUST. Downloaded on November 22,2021 at 11:04:14 UTC from IEEE Xplore. Restrictions apply.

Let a pattern $\mathbf{x}$ belonging to C_α be presented. Then we can define a set of integers $N_\alpha(\mathbf{x})$ associated with the pattern by

$$N_\alpha = \{\beta \mid g_\beta(\mathbf{x}, \boldsymbol{\theta}_\beta) > g_\alpha(\mathbf{x}, \boldsymbol{\theta}_\alpha)\}. \tag{71}$$

N_α obviously depends on $\boldsymbol{\Theta}$, and when the pattern is correctly classified, it is the null set. Let us consider a linear combination of $g_\beta(\mathbf{x}) - g_\alpha(\mathbf{x})$, $\beta \in N$,

$$d_\alpha(\mathbf{x}, \boldsymbol{\Theta}) = \sum_{\beta \in N_\alpha} s_\beta{}^\alpha(g_\beta(\mathbf{x}, \boldsymbol{\theta}_\beta) - g_\alpha(\mathbf{x}, \boldsymbol{\theta}_\alpha)), \tag{72}$$

where $s_\beta{}^\alpha$'s are weights and they may depend on $\mathbf{x}$ and $\boldsymbol{\Theta}$. When N_α is null, $d_\alpha(\mathbf{x}, \boldsymbol{\Theta})$ vanishes.

Let us define the loss caused by misclassification of $\mathbf{x} \in C_\alpha$ by

$$l_\alpha(\mathbf{x}, \boldsymbol{\Theta}) = l(d_\alpha(\mathbf{x}, \boldsymbol{\Theta})). \tag{73}$$

Obviously, the loss is 0 only when the pattern is correctly classified. The average risk can be written

$$R(\boldsymbol{\Theta}) = \sum_\alpha \int p_\alpha p_\alpha(\mathbf{x}) l_\alpha(\mathbf{x}, \boldsymbol{\Theta}) dX. \tag{74}$$

We call the vector which minimizes $R(\boldsymbol{\Theta})$ the optimal decision vector. It satisfies $\nabla R(\boldsymbol{\Theta}) = 0$, where ∇ is the gradient operator with respect to $\boldsymbol{\Theta}$.

Let us consider a learning rule, by which the present decision vector $\boldsymbol{\Theta}$ is modified by

$$\delta\boldsymbol{\Theta} = \epsilon C \boldsymbol{H}_\alpha(\mathbf{x}, \boldsymbol{\Theta}), \tag{75}$$

when a pattern $\mathbf{x} \in C_\alpha$ is presented. If we choose

$$\boldsymbol{H}_\alpha = -\nabla l_\alpha(\mathbf{x}, \boldsymbol{\Theta}), \tag{76}$$

we can easily obtain

$$\overline{\delta\boldsymbol{\Theta}} = -\epsilon C \nabla R(\boldsymbol{\Theta}). \tag{77}$$

Thus the probabilistic-descent method is obtained for the general classifier. We can prove the following convergence theorem.

Theorem 7: By using the learning functions

$$\boldsymbol{H}_\alpha = -\nabla l_\alpha(\mathbf{x}, \boldsymbol{\Theta}),$$

the probability that the decision vector approaches the optimum as near as desired, can be made as near to l as desired by choosing a sufficiently small learning constant ϵ.

The theorems concerning the convergence rate and the accuracy can also be obtained by using discussions similar to those given in Section IV.

The linear classifier is obtained as a special case of the general classifier. In this case, the parameter $\boldsymbol{\theta}_\alpha$ is identified with the weight vector $\hat{\boldsymbol{W}}_\alpha$, and the discriminant function is $g_\alpha(\mathbf{x}, \hat{\boldsymbol{W}}_\alpha) = \hat{\boldsymbol{W}}_\alpha{}^t X$. By putting

$$s_\beta{}^\alpha = \begin{cases} 1/\hat{w}_{\alpha\beta}, & \text{when } \max\limits_{\gamma \in N_\alpha} g_\gamma(\mathbf{x}, \hat{\boldsymbol{W}}_\gamma) = g_\beta(\mathbf{x}, \hat{\boldsymbol{W}}_\beta) \\ 0, & \text{otherwise,} \end{cases} \tag{78}$$

we obtain the learning functions of (62). If we put

$$s_\beta{}^\alpha = \begin{cases} 1, & \text{when } \max\limits_{\gamma \in N_\alpha} g_\gamma = g_\beta, \\ 0, & \text{otherwise,} \end{cases} \tag{79}$$

and $l(d) = d$, we obtain simpler learning functions

$$\boldsymbol{H}_\gamma = \begin{cases} -\boldsymbol{X}, & \text{when } \mathbf{x} \in C_\alpha \text{ is contained in } V_{\gamma\alpha}, \\ \boldsymbol{X}, & \text{when } \mathbf{x} \in C_\gamma \text{ is contained in } V_{\gamma\alpha}, \\ 0, & \text{when } \mathbf{x} \text{ is correctly classified.} \end{cases} \tag{80}$$

This gives the perceptron learning procedures.

In the case when $g_\alpha(\mathbf{x}, \boldsymbol{\theta}_\alpha)$ is linear with respect to $\boldsymbol{\theta}_\alpha$, i.e.,

$$g_\alpha(\mathbf{x}, \boldsymbol{\theta}_\alpha) = \sum_i \theta_{\alpha i} \phi_i(\mathbf{x}), \tag{81}$$

where $\phi_i(\mathbf{x})$ is a nonlinear function of $\mathbf{x}$, we obtain the so-called Φ machines [14]. The general piecewise-linear classifier can also obtained as a special case.

Learning of Learning Rules

As has already been shown in Section IV, the performance of the classifier depends on the constant ϵ and the components of C. Here, we shall try to determine the constants adaptively in such a manner that the convergence rate becomes fast when the weight vector is far from the optimal, and the degree of accuracy becomes high when it is nearly optimal. When the weight vector is far from the optimal, it is probable that the two successive nonzero correction vectors are in almost the same direction. On the contrary, when it is nearly optimal, it occurs with relatively large probability that the two successive nonzero correction vectors have opposite directions. It is desirable to increase the length of the correction vector in the former situation and to decrease it in the latter.

Let the present weight vector be $\boldsymbol{W}$ and let $\delta\boldsymbol{W}(\mathbf{x}, \boldsymbol{W})$ be the present nonzero correction vector. When a pattern $\mathbf{x}'$ is again misclassified by the modified weight vector $\boldsymbol{W}' = \boldsymbol{W} + \delta\boldsymbol{W}$, the nonzero correction vector $\delta\boldsymbol{W}'(\mathbf{x}', \boldsymbol{W}')$ will be produced. Let us adopt the following modification rule of ϵC. We change ϵC to $\epsilon C + \Delta C$, where

$$\Delta C = \gamma \boldsymbol{H}(\mathbf{x}, \boldsymbol{W}) \boldsymbol{H}'(\mathbf{x}', \boldsymbol{W}')^t, \tag{82}$$

when a pattern $\mathbf{x}'$ is misclassified, where $\mathbf{x}$ is the previously misclassified pattern and γ is a positive constant.

In order to study the effect of the above modification of ϵC, let us calculate the expected value of ΔC. It is written as

$$\overline{\Delta C} = \int \gamma \boldsymbol{H}(\mathbf{x}, \boldsymbol{W}) \boldsymbol{H}'(\mathbf{x}', \boldsymbol{W}')^t p(\mathbf{x}) p(\mathbf{x}') dX dX'. \tag{83}$$

By integration with respect to dX', it is transformed to

$$\begin{aligned} \overline{\Delta C} &= -\gamma \int \boldsymbol{H}(\mathbf{x}, \boldsymbol{W}) \{\nabla R(\boldsymbol{W} + \delta\boldsymbol{W})\}^t p(\mathbf{x}) dX \\ &= \gamma\{\nabla R \nabla R^t - 2\epsilon B_0 C^t A\}. \end{aligned} \tag{84}$$

Authorized licensed use limited to: KAUST. Downloaded on November 22,2021 at 11:04:14 UTC from IEEE Xplore. Restrictions apply.

When $\boldsymbol{W}$ is far from the optimal, the second term may be neglected, and we obtain

$$\overline{\Delta C} = \gamma \nabla R \nabla R^t. \tag{85}$$

This term acts to emphasize ∇R direction, accelerating the convergence. On the contrary, when $\boldsymbol{W}$ is nearly optimal, the first term may be neglected, and we obtain

$$\overline{\Delta C} = -2\epsilon B_0 C^t A. \tag{86}$$

Hence, we see that the absolute value of $\delta \boldsymbol{W}$ becomes smaller and the degree of accuracy larger. These are the properties we are looking for. Thus we have obtained the learning system equipped with the ability to learn the learning rule. In this system, the rate of convergence automatically increases or the degree of accuracy automatically increases according to whether the weight vector is far from the optimal or nearly optimal.

VII. Conclusion

We have proposed a learning rule for a linear pattern classifier, by which the weight vector converges to the optimal one, even if the patterns are not linearly separable. We have also studied the behavior of the classifier with the proposed learning rule, and clarified the rapidity, the accuracy, and the dynamical behavior of learning. Our theory has been generalized to the multicategory classifiers, to the classifiers having piecewise-linear discriminant functions, and to more general classifiers.

There remain several problems to be studied further. One of them concerns the loss function. Our theory is valid only when the distance loss function is adopted. However, the minimum-error-rate criterion is not included in this class, and we need to approximate it by an appropriate loss function. The problem is how to obtain an effective approximation, by which rapid convergence and good accuracy are guaranteed. Another problem concerns the existence of local minima. We have assumed that the average risk function has only one local minimum, which is the global minimum. This assumption surely holds in the cases where the patterns are linearly separable or nearly so. However, for some more general pattern distributions, the assumption will not hold, and we are not yet certain for what kind of distribution this holds or does not. When the assumption does not hold, we can say merely that the weight vector converges to one of the local minima. The learning of learning rules is also a problem to be studied further in more detail.

Acknowledgment

The author wishes to thank Prof. K. Kondo for his able guidance in applied geometry, as well as Prof. Y. Oono for his helpful comments on the manuscript.

References

[1] F. Rosenblatt, *Principles of Neurodynamics.* Washington, D.C.: Spartan, 1962.

[2] S. T. Hu, *Threshold Logic.* Berkeley, Calif.: University of California Press, 1965.

[3] N. J. Nilsson, *Learning Machines.* New York: McGraw-Hill, 1965.

[4] A. Albert, "A mathematical theory of pattern recognition," *Ann. Math. Stat.*, vol. 34, pp. 284–299, March 1963.

[5] S. Amari, "Diakoptics of information spaces, I, II and III," *RAAG Research Notes*, nos. 56, 60, and 74, October 1962, March and December 1963.

[6] ——, "Theory of normalization in pattern recognition systems," *RAAG Research Notes*, no. 85, January 1965.

[7] ——, "Theory of information spaces—transformations of metric signal spaces," *Electronics and Commun. in Japan*, vol. 48, pp. 35–47, February 1965.

[8] ——, "On learning linear decision systems," *RAAG Research Notes*, no. 107, June 1966.

[9] W. H. Highleyman, "Linear decision functions, with application to pattern recognition," *Proc. IRE*, vol. 50, pp. 1501–1514, June 1962.

[10] J. S. Koford and G. F. Groner, "The use of an adaptive threshold element to design a linear optimal pattern classifier," *IEEE Trans. on Information Theory*, vol. IT-12, pp. 42–50, January 1966.

[11] J. L. Doob, *Stochastic Processes.* New York: Wiley, 1953.

[12] R. O. Duda and H. Fossum, "Pattern classification by iteratively determined linear and piecewise linear discriminant functions," *IEEE Trans. on Electronic Computers*, vol. EC-15, pp. 220–232, April 1966.

[13] M. L. Dertouzos, *Threshold Logic.* Cambridge, Mass.: M.I.T. Press, 1965.

[14] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Trans. on Electronic Computers*, vol. EC-14, pp. 326–334, June 1965.

Authorized licensed use limited to: KAUST. Downloaded on November 22,2021 at 11:04:14 UTC from IEEE Xplore. Restrictions apply.