Pergamon

Neural Networks

Neural Networks 11 (1998) 297–313

Contributed article

Evolution and generalization of a single neurone
II. Complexity of statistical classifiers and sample size considerations

Šarūnas Raudys*

Institute of Mathematics and Informatics, Akademijos 4, Vilnius 2600, Lithuania

Received 3 January 1997; accepted 5 July 1997

Abstract

Unlike many other investigations on this topic, the present one does not consider the nonlinear SLP as a single special type of the classification rule. In SLP training we can obtain seven statistical classifiers of differing complexity: (1) the Euclidean distance classifier; (2) the standard Fisher linear discriminant function (DF); (3) the Fisher linear DF with pseudo-inversion of the covariance matrix; (4) regularized linear discriminant analysis; (5) the generalized Fisher DF; (6) the minimum empirical error classifier; and (7) the maximum margin classifier. A survey of earlier and new results, referring to relationships between the complexity of six classifiers, generalization error, and the number of learning examples, is presented. These relationships depend on the complexities of both the classifier and the data. This knowledge indicates how to control the SLP classifier complexity purposefully by determining optimal values of the targets, learning-step and its change in the training process, the number of iterations, and addition or subtraction of a regularization term. A correct initialization of weights, and a simplifying data structure can help to reduce the generalization error.

Keywords: Single-layer perceptron; Statistical classification; Generalization error; Initialization; Overtraining; Dimensionality; Complexity; Sample size; Scissors effect

1. Introduction

In Part I (Raudys, 1998) we demonstrated that in the nonlinear SLP training we can obtain seven statistical classifiers of differing complexity. In theoretical analysis and applications, it is important to know the relationship between the complexities of classifiers, their generalization properties, and the numbers of learning examples. A great deal of research work concerned with this relationship has been performed during the last three decades.

After proposing the first stochastic descent algorithm, Widrow and Hoff (1960) concluded that the sample size required to achieve a given learning quality of the adaline type algorithm should increase in proportion to the number of inputs. Cover (1965) introduced a capacity—a measure of the complexity, and showed that the generalization error decreases in proportion to p/n, dimensionality–learning-set size ratio.

Several approaches have been proposed to study the generalization error in finite learning-set size situations. In a number of research papers beginning with their 1968 paper, Vapnik and Chervonenkis (1968) developed the Cover capacity concept and obtained a number of upper estimates for the generalization error.

In the classical statistical approach, vector $\mathbf{x}$ to be classified into classes π_1, π_2 is assumed to be a random variable with a certain conditional probability density function $f(\mathbf{x}|\pi_i)$. To estimate the structure of the classifier and its weight vector $\mathbf{w}$, one uses assumptions on the probabilistic structure of $f(\mathbf{x}|\pi_i)$, and learning-set observation vectors. To analyse a dependence of the generalization error on the structure of the classifier and the learning-set size, one uses standard statistical methods. This approach is considered in his paper. Among other approaches, the most popular are: a probable almost correct (PAC) framework (Valiant, 1984); the statistical mechanics approach; and *the information-theoretic and statistical approach*, based on statistical models of conditional density $f(o_i|\mathbf{w},\mathbf{x}_i)$ of the output o_i of the network, $f(\mathbf{x}_i)$, an unconditional density, and the standard technique of asymptotic statistical inference, which is valid under regularity conditions such as the existence of the moments of random variables and the existence of the Fisher information (see, e.g., Levin et al., 1990; Amari and Murata, 1993; Amari, 1993).

* Requests for reprints should be sent to Sarunas Raudys. E-mail: raudys@ktl.mii.lt. Fax: (370) 2 729-209.

0893–6080/98/$19.00
PII S0893-6080(97)00136-6

In the latter stream of investigations, Amari et al. (1992) showed that the average generalization error EP_n behaves asymptotically as L/n, when the network is deterministic, the teacher signal is noiseless, and the network giving the correct classification is uniquely specified by the L-dimensional parameter $\mathbf{w}^*$. In the case with an empty zone between the pattern classes, we have much better small sample behaviour $EP_n \sim c/n^2$, where c is an unknown constant. For a unique deterministic network trained by a noisy teacher $EP_n \sim c/n^{1/2}$, and for a stochastic network $EP_n \sim P_\infty + c_1/n$. Amari and Murata (1993) proved fundamental universal convergence theorems for the average generalization and training errors measured as the predictive entropic loss EH_n (stochastic complexity) evaluated by the expectation of $-\log f(o|\mathbf{w},\mathbf{x})$ for an input–output pair $(\mathbf{x}, o)$. For the weights estimated by the maximum likelihood method or by the Bayes posterior distribution, it was proved that an average generalization entropic error of the stochastic network, $EH_n = H_\infty + L^*/(2n)$, where L^* shows the complexity of the network. For the faithful (realizable) network, $L^* = L$, and for the unfaithful (unrealizable) network, $L^* = \mathrm{tr}\mathbf{K}^{-1}\mathbf{G}$, where $\mathbf{K}$ is the Hessian matrix, and $\mathbf{G}$ is the Fisher information matrix. For a deterministic dichotomy network, $EH_n = L/n$ (Amari, 1993).

A characteristic property of *the statistical-mechanics approach* is the so-called ''*thermodynamic limit*'', when one examines the generalization error both as $n \to \infty$ and as $L \to \infty$, but at some fixed rate. This allows us to meaningfully investigate, for instance, an asymptotic generalization error when the number of examples is half the number of parameters, twice the number of parameters, 10 times the number of parameters, and so on (Haussler et al., 1994). This approach uses mathematical methods from statistical mechanics, such as the replica symmetry technique and the annealed approximation. There, a mean value of the ratio of two random variables is substituted by the ratio of mean values of these two random variables. The validity of this approximation is still open. For some specific models the statistical-mechanic approach succeeds in obtaining the average generalization error, and its ''phase transition'' (sudden drops in the generalization error). For the deterministic dichotomy network, for example, a strong rigorous result was proved: $EH_n = 0.62 \times L/n$ (Gyorgyi and Tishby, 1990; Opper and Haussler, 1991). In certain cases, a different power law than $1/n$ or $1/n^{1/2}$ was demonstrated (Haussler et al., 1994; Seung et al., 1992).

An interesting and promising approach is that of *combining statistical physics with VC-bounds*, that allows us to incorporate of some problem specific information. It was demonstrated that the introduction of limited information on the distribution of error patterns to the classical-VC formalism permits much tighter bounds on learning curves. The ''phase transitions'', as well as significant drops in learning errors, can be modelled for low sizes of training samples for which the classical VC-bounds are void (see Kowalczyk, 1996, and references therein).

A great deal has been done on the *analysis of the small sample behaviour of statistical classifiers*. As for the statistical-mechanics approach, classical statistical analysis also requires knowledge of the input signal distribution $f(\mathbf{x}|\pi_i)$. This is the weak point of these approaches. However, assumptions on the probabilistic structure of pattern classes and on the parameters make it possible to obtain narrower error bounds. In some cases, absolutely exact results can be obtained and only one question remains—how to use these results in practice, where true distributions are unknown.

Rao (1949) was the first to emphasize, then, problems when the number of learning examples was close to the number of dimensions. The first numerical estimate of the difference between the generalization and asymptotic errors was obtained by numerical simulation at the Institute for Numerical Analysis of University of California in Los Angeles (see references in Solomon, 1956). Sitgreaves (1961) derived the first exact formula for the expected classification error of the standard Fisher linear discriminant function (DF) in the form of a five-times infinite sum of products of certain hypergeometric functions. Estes (1965) succeeded in calculating this sum, and Pikelis improved the calculation accuracy and presented a table (Pikelis, 1974, see also Raudys and Pikelis, 1980, and references therein). The first asymptotic expansion for the expected classification error of the Fisher linear DF belongs to Okamoto (1963). It is obtained asymptotically, where $n \to \infty$, and often yields inaccurate values, if the dimensionality p is large. John (1961) represented the linear discriminant function with the known covariance matrix as a difference of two independent chi-square variables, and expressed the expected error in a form of infinite sum. Raudys (1967) used this result and derived the first simple asymptotic formula for the expected probability of misclassification (PMC) of an Euclidean distance classifier. Faithful and unfaithful cases were first analysed here, as well as the ''thermodynamic limit'' where both the learning set size $n \to \infty$ and the dimensionality $p \to \infty$.

Deev (1970, 1972) formalized this thermodynamic limit approach in a strictly mathematical way: it was formally required that $n \to \infty$, $p \to \infty$, $p/n \to$ *constant*, and Mahalanobis distance $\delta =$ *const*. Under this approach, several subsequent asymptotic expansions were obtained for Gaussian and non-Gaussian models. Two simple formulae for the expected error for the standard Fisher linear DF were obtained in Deev (1970, 1972), and Raudys (1972). Further analysis (Pikelis, 1976; Wyman et al., 1990) showed that on the ''thermodynamic limit'' based asymptotic expansions give very accurate estimates. This approach was used to obtain the generalization error for the standard quadratic DF, linear and nonlinear classifiers for independent Gaussian variables (Raudys, 1972), a block type and a tree type dependency between the Gaussian variables (Deev, 1974; Zarudskij, 1979), the classifier for independent categorical variables (Meshalkin, 1976), and the regularized DA (Raudys and Skurikhina, 1994). Meshalkin and Serdobolskij

(1978) proved a fundamental limit theorem for arbitrary non-Gaussian classes.

The ''*curse of dimensionality*'' was first described by Lbov (1966) and Hughes (1968), and the ''*scissors effect*'' was discovered in Raudys (1970), and Kanal and Chandrasekaran (1971). See, also, Jain and Chandrasekaran (1982), Raudys and Jain (1991): in small learning-set cases, it is often preferable to use simple-structured classification rules instead of complex ones, and *vice versa*; in large learning-set cases, the complex classifiers can be used more efficiently. In the statistical mechanics approach, this effect was found much later (van Dam et al., 1994; Meir, 1995).

Typically, in the generalization error study, the SLP is analysed as a separate special specimen of the classification algorithm. Most often the activation function (or the pattern error function), it is assumed to be the linear or the threshold function; sometimes it is assumed to be a softlimiting one. As a matter of fact in all connectionist analysis, the authors analyse the asymptotic behaviour of the perceptron for individual models, paying too little attention to very small learning-set situations, where the generalization error is high in comparison with the asymptotic error. Too little attention is paid to different mathematical models of the data.

Unlike many *other investigations, the present paper does not consider the nonlinear SLP as a single classifier*. We analyse the perceptron as a dynamical process, and pay special attention to the type of distribution of the pattern classes and the situations where the learning-set size is small. In Part I we have shown that on the way between the starting point and the minimum of the cost function, the weights of the perceptron gradually increase, and decision boundaries of SLP become identical or close to those of seven statistical classifiers. The aim of Part II is to show how a substantial number of results from standard multivariate statistical analysis can be used in the generalization error analysis of simple artificial neural nets.

The paper has been split into two parts. In Part I, we have shown that the nonlinear SLP is not a single classifier, it is a process. In this, Part II, we analyse the small learning-set properties of several well known statistical classifiers which can be detected in nonlinear SLP training results. A great number of these results were published either in Russian or in conference proceedings, and remained unknown to the connectionist community. We analyse these former and new results from a fresh unique point-of-view, using the terminology popular in the statistical mechanics approach. We demonstrate how theoretical results referring to statistical classifiers can be used for conscious control of the SLP complexity in its learning process.

In Section 3 of this part, we analyse small sample properties of four parametric statistical classifiers based on the class distribution density. In the Section 4 we analyse two nonparametric classifiers, based on the type of decision rule. Section 5 is an experimental one. It shows that the theoretical results presented in the previous two sections are valid for nonlinear SLP classifier analysis. Section 6 analyses overtraining and the dynamics of the SLP training process. Section 7 presents some additional references, and compares the theoretical results with those obtained by other approaches. Section 8 discusses practical aspects of using theoretical knowledge presented in the paper: complexity control; data transformations; weight initialization; and so on.

2. Definitions and notation

We analyse a nonlinear SLP dichotomy classifier that has p inputs, and one output $output = o(\mathbf{w}'\mathbf{x} + w_o)$, where w_o, $\mathbf{w} = (w_1, w_2,...,w_p)'$ are weights, $\mathbf{x} = (x_1,x_2,...,x_p)'$ is the input vector, $o(g)$ is the nonlinear ''tanh'' activation function. To find the perceptron weights we minimize the cost function

$$cost_l = \frac{1}{2N_1 + N_2} \sum_{i=1}^{2} \sum_{j=1}^{N_i} (t_j^{(i)} - o(\mathbf{w}'\mathbf{x}_j^{(i)} + w_o))^2. \quad (1)$$

In the above formula, $t_j^{(i)}$ is the desired output (a target) of $\mathbf{x}_j^{(i)}$, the jth learning-set observation vector from π_i, the ith class, N_i is the number of learning vectors from π_i. Usually we use $t_j^{(1)} = 1$ and $t_j^{(2)} = -1$ for the tanh(g) activation function. We call these values *limiting* ones. In simulations with the sigmoid function, we use $t_j^{(1)} = 0$ and $t_j^{(2)} = 1$ (limiting values), or $t_j^{(1)} = 0.1$ and $t_j^{(2)} = 0.9$. We analyse the standard total gradient delta learning rule (back-propagation, BP) where the weight vector is adapted according to the iterative rule $\mathbf{w}_{(t+1)} = \mathbf{w}_{(t)} - \eta \partial cost_l / \partial \mathbf{w}$, where η is called a learning-step.

2.1. Data types used in numerical calculations and simulation studies

GCCM (Gaussian with common variance matrices) are two p-variate Gaussian classes $N(\boldsymbol{\mu}_1,\boldsymbol{\Sigma})$, $N(\boldsymbol{\mu}_2,\boldsymbol{\Sigma})$ with different mean vectors $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and a common $p \times p$ covariance matrix $\boldsymbol{\Sigma}$.

SGC are two multivariate spherical Gaussian classes $N(\boldsymbol{\mu}_1,\mathbf{I})$, $N(\boldsymbol{\mu}_2,\mathbf{I})$.

$EP_N^{(\alpha)}$ stands for the expected probability of misclassification (PMC)—the mean generalization error—of the classifier trained by method α, $P_\infty^{(\alpha)}$ is the asymptotic PMC: $EP_N^{(\alpha)} \rightarrow P_\infty^{(\alpha)}$ as the learning-set sizes N_1, $N_1 \rightarrow \infty$, and P_B is the Bayes error.

C are two 20-variate ($p = 20$) GCCM classes; unit variances of all the variables, correlations between all the variables $\rho = 0.213$, $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' = (-1.7040,0.0326,0.0599,0.0872,...,0.4970,0.5244)$. $P_\infty^{(E)} = 0.03$, $P_\infty^{(F)} = 0.01$, and the effective dimensionality for EDC $p^* = p$.

D1 are two 100-variate ($p = 100$) GCCM classes; unit variances; correlations between all the variables $\rho = 0.3$, $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2 = 1.042 \times (1,1,...,1)$. $P_\infty^{(E)} = P_\infty^{(F)} = 0.03$, and $p^* \approx 1.05$ (definition of p^* in Eq. (5)).

D2 are two 100-variate ($p = 100$) GCCM classes; unit variances; correlations between all the variables $\rho = -0.0101$, $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2 = 0.0018805 \times (1,1,...,1)$. $P_\infty^{(E)} = P_\infty^{(F)} = 0.03$, and $p^* \approx 10^{10}$.

3. Generalization errors of parametric classifiers

3.1. The Euclidean distance classifier (EDC)

In Raudys (1967), the generalization error was first considered asymptotically, when the dimensionality p and the learning-set sizes N_1, N_2 are large and increasing simultaneously. In statistical mechanics, this is called the ''*thermodynamic limit*''. An increase in p implies that the conditional distribution of the discriminant function—the random variable $g(\mathbf{X},\bar{\mathbf{X}}^{(1)},\bar{\mathbf{X}}^{(2)}|\mathbf{X} \in \pi_i)$—asymptotically tends to the Gaussian distribution, and allows us to obtain very simple, but accurate estimates. The result for EDC is unknown, and in fact is unavailable for Western researchers. Therefore we repeat the main steps of its derivation.

We consider $N_1 + N_2$ learning set vectors $\mathbf{X}_1^{(1)},\mathbf{X}_2^{(1)},\cdots,\mathbf{X}_{N_1}^{(1)},\mathbf{X}_1^{(2)},\cdots,\mathbf{X}_{N_2}^{(2)}$ as random vectors. Then we have to consider the discriminant function $\mathrm{g}^{\mathrm{E}}(\mathbf{X},\bar{\mathbf{X}}^{(1)},\bar{\mathbf{X}}^{(2)}|\mathbf{X} \in \pi_i)$ as a random variable that depends on three independent p-variate random vectors $\mathbf{X}$, $\bar{\mathbf{X}}^{(1)}$ and $\bar{\mathbf{X}}^{(2)}$ (in order to stress that the variables are considered as random ones, it is common in statistics to denote them by capital letters). Then the expected PMC (mean generalization error) can be written as a sum of two conditional probabilities:

$$EP_N^{(\mathrm{E})} = q_1 Prob\{\mathrm{g}^{\mathrm{E}}(\mathbf{X},\bar{\mathbf{X}}^{(1)},\bar{\mathbf{X}}^{(2)}) < 0|\mathbf{X} \in \pi_1\} + q_2 Prob\{\mathrm{g}^{\mathrm{E}}(\mathbf{X},\bar{\mathbf{X}}^{(1)},\bar{\mathbf{X}}^{(2)}) \geq 0|\mathbf{X} \in \pi_2\}, \quad (2)$$

where q_1 denotes *a priori* probabilities of class π_1, and $q_2 = 1 - q_1$.

Asymptotically, when p and N_1, N_2 are increasing, the expected probability of misclassification (the generalization error) is

$$EP_N^{(\mathrm{E})} = q_1\Phi\left\{-\frac{\mathrm{E}[\mathrm{g}(\mathbf{X},\mathbf{X}^{(1)},\mathbf{X}^{(2)})|\mathbf{X} \in \pi_1]}{\mathrm{V}[\mathrm{g}(\mathbf{X},\mathbf{X}^{(1)},\mathbf{X}^{(2)})|\mathbf{X} \in \pi_1]}\right\} + q_2\Phi\left\{\frac{\mathrm{E}[\mathrm{g}(\mathbf{X},\mathbf{X}^{(1)},\mathbf{X}^{(2)})|\mathbf{X} \in \pi_2]}{\mathrm{V}[\mathrm{g}(\mathbf{X},\mathbf{X}^{(1)},\mathbf{X}^{(2)})|\mathbf{X} \in \pi_2]}\right\}, \quad (3)$$

where E denotes the expectation, and V the variance, with respect to three independent random vectors $\mathbf{X}$,$\bar{\mathbf{X}}^{(1)}$, $\bar{\mathbf{X}}^{(2)}$.

Let $N_2 = N_1 = N$, $q_2 = q_1 = 0.5$, and the classes be multivariate Gaussian with different means and a common covariance matrix: $\mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $\mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$—GCCM model.

Note that while designing EDC one assumes the covariance matrix $\boldsymbol{\Sigma} = \mathbf{I}\sigma^2$, and in the analysis of the generalization error, we consider the case where the probabilistic model of the pattern classes is *different*, i.e., $\boldsymbol{\Sigma} \neq \mathbf{I}\sigma^2$. In statistical mechanics the difference in mathematical descriptions is called an ''*unfaithful*'' (unrealizable) model. In our statistical approach, the term ''unrealizable'' is not exact, since there exist models where $\boldsymbol{\Sigma} \neq \mathbf{I}\sigma^2$, but the asymptotic PMC of the Euclidean distance classifier coincides with the Bayes error.

For the GCCM model and $N_2 = N_1$, the distribution of DF $g^{\mathrm{E}}(\mathbf{X}, \bar{\mathbf{X}}^{(1)}, \bar{\mathbf{X}}^{(2)})$ can be analysed as the distribution of a vector product of two independent random vectors, $\mathbf{Z}$ and $\mathbf{Y}$, i.e., $\mathrm{g}^{\mathrm{E}}(\mathbf{X}, \bar{\mathbf{X}}^{(1)}, \bar{\mathbf{X}}^{(2)}) = \mathbf{Z}'\mathbf{Y}$ is the difference of two quadratic forms of Gaussian random variable. We have denoted here $\mathbf{Z} = \mathbf{X} - \frac{1}{2}(\bar{\mathbf{X}}^{(1)} + \bar{\mathbf{X}}^{(2)})$, $\mathbf{Z} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}(1 + \frac{1}{2N}))$, $\mathbf{Y} = \bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(2)}$, $\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}\frac{2}{N})$, $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\mu}$. Taking into account that $(\mathbf{Z}'\mathbf{Y})^2 = \mathrm{tr}(\mathbf{Z}'\mathbf{Y}\mathbf{Z}'\mathbf{Y}) = \mathrm{tr}(\mathbf{Y}\mathbf{Y}'\mathbf{Z}\mathbf{Z}')$, we get

$$E[g(\mathbf{X},\bar{\mathbf{X}}^{(1)},\bar{\mathbf{X}}^{(2)})|\mathbf{X} \in \pi_i] = (-1)^{i-1}\frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\mu}, \quad (4a)$$

$$V[g(\mathbf{X},\bar{\mathbf{X}}^{(1)},\bar{\mathbf{X}}^{(2)})|\mathbf{X} \in \pi_i] = \mu'\Sigma\mu\left(1 + \frac{1}{N}\right) + tr(\Sigma^2)\frac{2}{N}\left(1 + \frac{1}{2N}\right)., \quad (4b)$$

An expression for the expected PMC follows directly from Eqs. (3), (4a) and (4b):

$$EP_N^{(\mathrm{E})} \approx \Phi\left\{-\frac{\frac{1}{2}\mu'\mu}{\sqrt{\boldsymbol{\mu}'\boldsymbol{\Sigma}\boldsymbol{\mu}\left(1 + \frac{1}{N}\right) + tr\Sigma^2\left(1 + \frac{1}{2N}\right)}}\right\}.$$

In the thermodynamic limit, for $\delta_1^* = const.$, and large p and N, ignoring the terms of order $\frac{1}{N^2}$ and $\frac{p}{N^2}$, one obtains a very simple expression

$$EP_N^{(\mathrm{E})} \approx \Phi\left\{-\frac{\delta^*}{2}\frac{1}{\sqrt{T_\mu^*}}\right\}, \quad (5)$$

where $\delta^* = \frac{\mu'\mu}{\sqrt{\mu'\Sigma\mu}}$, $T_\mu^* = 1 + \frac{2p^*}{\delta^{*2}N}$, $p^* = \frac{(\mu'\mu)^2(tr\Sigma^2)}{(\mu'\Sigma\mu)^2}$.

Asymptotically, as $N \to \infty$ we obtain the asymptotic PMC of the Euclidean distance classifier: $P_\infty^{(\mathrm{E})} = \Phi\{-\delta^*/2\}$. For the spherical Gaussian case we have $\boldsymbol{\Sigma} = \mathbf{I}\sigma^2$. Then $\delta^* = \delta$, where $\delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is the squared Mahalanobis distance. In a more general case (when $\boldsymbol{\Sigma} \neq \mathbf{I}\sigma^2$), $\delta^* \leq \delta$. Then the asymptotic error $P_\infty^{(\mathrm{E})}$ can be larger than the asymptotic PMC obtained for the standard Fisher DF $P_\infty^{(\mathrm{F})} = \Phi\{-\delta/2\}$. For example, for 20-variate Gaussian data $\mathbf{C}$, $\delta^* = 3.7616$, $\delta = 4.65$, $P_\infty^{(\mathrm{E})} = 0.03$ and $P_\infty^{(\mathrm{F})} = 0.01$. There exist situations where the features are correlated, but $P_\infty^{(\mathrm{E})} = P_\infty^{(\mathrm{F})}$. Two such examples are presented in Fig. 1 (pairs of the classes π_3 and π_4, π_3 and π_5). Two other examples are the pattern classes **D1** and **D2**, with $\delta^* = \delta = 3.76$, and $P_\infty^{(\mathrm{E})} = P_\infty^{(\mathrm{F})} = 0.03$. Since the parameter δ^* controls the asymptotic PMC, we call it *a modified* (*effective*) *Mahalanobis distance*.

In the spherical Gaussian case, $p^* = p$. The EDC can be trained with comparatively small learning-sets in this case.

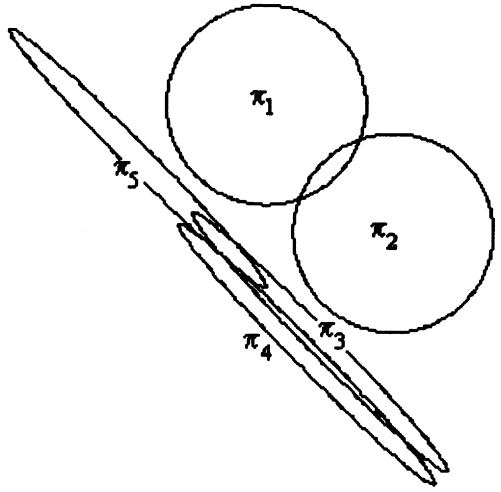

Fig. 1. Effect of a configuration of GCCM model on the effective dimensionality p^*: classes p_1 and p_2—$p^* = p$, classes p_3 and p_4—$p^* \gg p$, classes p_4 and p_5—$p^* \gg p$.

For example, for $p = 20$, $\delta = 4.65$ from Eq. (5) we calculate: $EP_N^{(E)} \approx 0.0469$ for $N = 15$; $EP_N^{(E)} \approx 0.0400$ for $N = 25$; $EP_N^{(E)} \approx 0.0362$ for $N = 40$, and $EP_N^{(E)} \approx 0.033$ for $N = 80$ (graph 1 in Fig. 2). It is important to stress that in special cases, where $\mathbf{\Sigma} \neq \mathbf{I}\sigma^2$, theoretically $1 \leq p^* \leq \infty$. It means that hypothetically there exist situations where the EDC is either very insensitive to the learning-set size or, on the contrary, very sensitive to the learning-set size.

When $p^* = 1$ we call the model *the most favourable distributions* of Gaussian pattern classes for the EDC. An example of densities of such a type is presented in Fig. 1 (pairs of the classes π_3 and π_5). Another example is the 100-variate data **D1** with $p^* \approx 1.05$. Because of the small

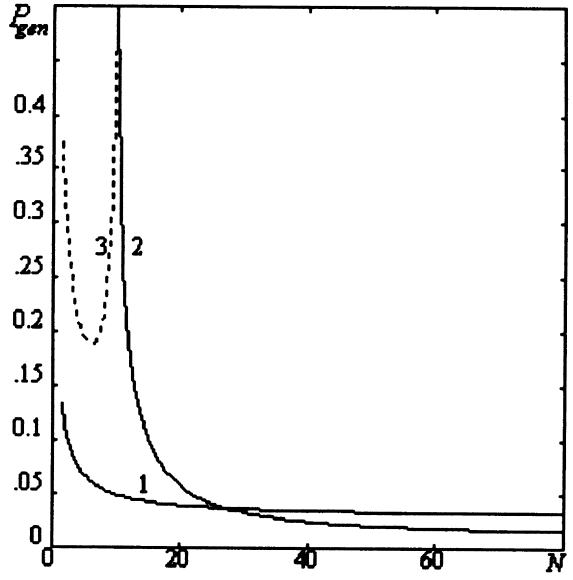

Fig. 2. The ''scissors effect''. The generalization error versus N, the learning set size: (1) Euclidean distance classifier—Hebb algorithm—for $p^* = 20$; (2) Fisher classifier—adaline—(graphs from Raudys, 1970); (3) Fisher classifier with pseudoinversion.

effective dimensionality for this specific choice of parameters, we can train the SLP on very small learning-sets. It follows from Eq. (5) that $EP_N^{(E)} = 0.0318$ for $N = 5$, and $\delta^* = 3.76$. We see that for this very favourable case, in spite of the high formal number of variables ($p = 100$), only five vectors per class are sufficient to train the EDC perfectly.

Hypothetically, there exist models where $p^* \to \infty$. We call such a *model the least favourable distributions* of pattern classes for EDC. An example of densities of such a type is presented in Fig. 1 (pairs of the classes π_3 and π_4). Another example is 100-variate data **D2** with $p^* \approx 10^{10}$. Even an insignificant deviation in sample means $\bar{\mathbf{x}}^{(1)}$, $\bar{\mathbf{x}}^{(2)}$ causes a critical rotation of the decision boundary and a distressing increase in the generalisation error. From Eq. (5) we calculate $EP_N^{(E)} = 0.4997$ for $N = 200$, $p^* = 10^{10}$, and $\delta^* = 3.76$. In theory, p^* can be close to infinity. Thus, for $q_2 = q_1 = 0.5$, and any number of learning observations the generalization error of the EDC is close to 0.5. The parameter p^* controls the sensitivity of the EDC to the learning-set size. Therefore, we have called it *a modified* (*effective*) *dimensionality* (Raudys, 1967).

Consider the model GCCM^r in which the covariance matrix $\mathbf{\Sigma}_r$ can be represented as $\mathbf{\Sigma}_r = \mathbf{G}'\mathbf{D}\mathbf{G}$, where $\mathbf{G}$ is a $p \times p$ orthonormal matrix of eigenvalues of $\mathbf{\Sigma}_r$, and $\mathbf{D}$ is a $p \times p$ diagonal matrix of the eigenvectors, such that

$$\mathbf{D} = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \epsilon \mathbf{I}_{p-r} \end{bmatrix},$$

and $\mathbf{I}_r$ is $r \times r$ identity matrix, $\mathbf{I}_{p-r}$ is a $(p - r) \times (p - r)$ identity matrix, and ϵ is a small positive constant, such that $(p - r)\epsilon \ll 1$. Let

$$\boldsymbol{\mu}'\mathbf{G}' = \boldsymbol{\mu}' \begin{bmatrix} \mathbf{g} \\ \mathbf{g}_2 \end{bmatrix}' = (\boldsymbol{m}', \boldsymbol{m}_2'),$$

where absolute values of components m_{2j} of the $(p-r)$-variate vector $\boldsymbol{m}_2$ are very small: $m_{2j} \ll \epsilon$, and can be ignored. This model implies that the distribution of the vector $\mathbf{X}$ lies in a subspace of dimensionality r. We say that such data are of *an intrinsic dimensionality*, equal to r. The effective dimensionality of such data

$$p* = \frac{(\mu'\mu)^2(tr\Sigma_r^2)}{(\mu'\Sigma_r\mu)^2} = \frac{(\boldsymbol{\mu}'\mathbf{G}'\mathbf{G}\mu)^2(tr\mathbf{G}\mathbf{\Sigma}_r\mathbf{G}'\mathbf{G}\mathbf{\Sigma}_r\mathbf{G}')}{(\boldsymbol{\mu}'\mathbf{G}'\mathbf{G}\mathbf{\Sigma}_r\mathbf{G}'\mathbf{G}\mu)^2}$$

$$= \frac{(\boldsymbol{m}'\boldsymbol{m})^2(tr\mathbf{D}^2)}{(\boldsymbol{m}'\mathbf{D}\boldsymbol{m})^2} = r$$

The intrinsic dimensionality of the multivariate data model with $p^* = 1$, discussed above, is equal to 1. We see that an increase in the generalization error of the EDC depends not on the formal, but on the intrinsic dimensionality of the data. In practice, most often the pattern vectors lie in a nonlinear subspace of lower dimensionality, but the variability of the other $p - r$ dimensions is not extremely small, i.e., the condition $(p - r)\epsilon \ll 1$ is not fulfilled. Then we have intermediate cases.

We need to know only the difference in sample means in order to find the weight vector, **w**, of EDC. It is, in fact, the same weight vector found by the Hebb training rule (see, e.g., Barkai et al., 1993). For spherical Gaussian patterns, Eq. (5) is in fact identical to that derived by the statistical mechanics approach (Barkai et al., 1993; Meir, 1995).

As a general conclusion, we can say that *the sensitivity of EDC to the learning-set size strongly depends on the data*. In principle, the sensitivity can be very low, but it can also be extremely high. In practical problems, however, we seldom have cases similar to the least favourable or to the least unfavourable ones just discussed.

3.2. The standard Fisher linear discriminant function

The number of parameters to be estimated from the learning-set is much larger than for the EDC: we need to estimate $2p$ components of the mean vectors for EDC, and we have to estimate $p(p+1)/2$ components of the covariance matrix for the Fisher linear DF, in addition. For the GCCM model, when $N_2 = N_1 = N$, $q_2 = q_1$, the generalization error of the above classifier (for $\delta^2 = const.$, and large N and large p) can be asymptotically expressed as (Deev, 1970, 1972); Raudys, 1972)

$$EP_{\mathrm{N}}^{(\mathrm{F})} \approx \Phi\left\{-\frac{\delta}{2}\frac{1}{\sqrt{T_\mu T_\Sigma}}\right\}, \tag{6}$$

where δ^2 is the squared Mahalanobis distance, the term $\mathrm{T}_\mu = 1 + \dfrac{2p}{\delta^2 N}$ arises from inexact sample estimation of the mean vectors of the classes, and the term $\mathrm{T}_\Sigma = 1 + \dfrac{p}{2N-p}$ arises from inexact sample estimation of the covariance matrix. In spite of its simplicity, Eq. (6) yields very exact values for the GCCM classes (Pikelis, 1976; Wyman et al., 1990).

If $p \to 2N$, the estimate of the covariance matrix becomes very inexact, and the term T_Σ increases without limit. Then the expected PMC tends to 0.5 (when $q_2 = q_1 = 0.5$). When N increases, and p remains constant, the expected error tends to its asymptotic value $P_\infty^{(F)}$. For example, for the 20-variate Gaussian model GCCM with $\delta = 4.653$, $P_\infty^{(F)} = 0.01$ from Eq. (6) we calculate: $EP_N^{(\mathrm{F})} \approx 0.1094$ for sample size $N = 15$; $EP_N^{(\mathrm{F})} \approx 0.0441$ for $N = 25$; $EP_N^{(\mathrm{F})} \approx 0.0259$ for $N = 40$, and $EP_N^{(\mathrm{F})} \approx 0.0163$ for $N = 80$ (graph 2 in Fig. 2).

By the example given in Fig. 1, we see that for small learning sets (up to $N \approx 30$) it is preferable to use a simple structured Euclidean distance classifier. Furthermore, for large learning sets (over $N \approx 30$) it is preferable to use a complex structured Fisher classifier. It is the ''*scissors effect*'' known in Statistical pattern recognition already for 25 years: in small learning-set cases, it is often preferable to use simple structured classification rules instead of complex ones, and, vice versa, in large learning-set cases, complex classifiers can be used more efficiently. The learning curves $EP_N^{(1)} = f_1(N)$ and $EP_N^{(2)} = f_2(N)$ of two classifiers intersect and resemble scissors, see Fig. 2, where for small learning-set sizes the graph for the EDC (1) is significantly lower than graph for the Fisher DF (2) and graph for the ''pseudo Fisher'' classifier (3). Note that the Hebb training rule is, in fact, the EDC, and the adaline is the Fisher linear DF. Thus, for small learning-sets it is preferable to use the Hebb rule, and for large ones the adaline.

3.3. The Fisher classifier with the pseudo-inverse covariance matrix

The generalization error can be understood on considering that in the pseudo-inverse approach, the feature space is rotated by means of a certain orthogonal transformation $\mathbf{Y} = \mathbf{TX}$ and afterwards classified by a ''diagonal'' classifier in an r-variate space of new directions corresponding to r non-zero eigen-values of the sample covariance matrix $\mathbf{S}$ ($r = N_1 + N_2 - 2$ is the rank of the sample covariance matrix $\mathbf{S}$). In this classifier design model, it is assumed that a covariance matrix of the vector $\mathbf{Y} = \mathbf{TX}$ is a diagonal matrix $\mathbf{d}$, composed of variances of the vector $\mathbf{Y}$—the r non-zero eigenvalues $d_1, d_2, \ldots, d_r$ of the matrix $\mathbf{S}$. This is not the optimal way to design a classifier in the very small learning-set case. The expected error of the ''diagonal'' classifier is expressed by the equation

$$EP_{\mathrm{N}}^{(\mathrm{PF})} \approx \Phi\left\{\frac{\delta\sqrt{r/p}}{2}\frac{1}{\sqrt{(1+\gamma^2)T_\mu + \gamma^2\dfrac{3\delta^2}{4Np}}}\right\}, \tag{7}$$

where

$\gamma = \sqrt{V_d}/E_d$; E_d, V_d are respectively mean and variance of $1/d$, and d is a randomly chosen eigenvalue of the matrix $\mathbf{S}$ having Wishart $\mathrm{W}(\mathbf{I}_p, n-2)$ distribution. Eq. (7) has some similarity with Eq. (5). To find γ we have to calculate moments of the inversion of eigenvalues of the random Wishart $\mathrm{W}(\mathbf{I}_p, n-2)$ matrix (Raudys and Duin, 1998). With an increase in the learning-set size n from 1 up to p, the terms r/p and γ are increasing, e.g., for $p = 20$ we have: $\gamma = 0.3247$ for $n = 3$, $\gamma = 1.03$ for $n = 11$, and $\gamma = 9.75$ for $n = 21$. In the nominator of Eq. (7), the term r/p tends to decrease the classification error. In the denominator of Eq. (7), the term γ tends to increase the generalization error. Numerical calculations by using Eq. (7) show an interesting and unexpected behaviour of the classification error: with an increase in the learning set size N, the generalization error decreases at first, reaches the minimum, and afterwards begins increasing (see graph 3 in Fig. 2). The minimal error is obtained for $N = p/4$ ($n = p/2$) and the maximal errors are obtained for $N = p/2$ ($n = p$). It is a consequence of non-optimality of the plug-in pseudo Fisher classifier. If $n > p$, we obtain the Fisher linear DF, and the expected error regularly decreases with an increase in N.

3.4. Regularized linear discriminant analysis

When calculating the weights of the linear discriminant function one uses $\mathbf{S}^{\mathrm{RDA}} = \mathbf{S} + \lambda\mathbf{I}$ instead of the conventional sample estimate $\mathbf{S}$. Positive terms λ added to each diagonal element of the covariance matrix help to invert the covariance matrix and act as regularizers. Asymptotically, as p and N increase, the distribution of the random variable $g^{\mathrm{RDA}}(\mathbf{X}, \bar{\mathbf{X}}^{(1)}, \bar{\mathbf{X}}^{(2)}, \mathbf{S}|\mathbf{X} \in \pi_i)$ tends to a Gaussian. To obtain an analytical expression for the expected PMC, we have used the first two terms of the Taylor series expansion of $(\mathbf{S} + \lambda\mathbf{I})^{-1} = \mathbf{S}^{-1} + \lambda^2\mathbf{S}^{-2} + ...$, calculated mixed second- and higher-order moments of an inverse covariance matrix $\mathbf{S}^{-1}$. After some simple but tedious algebra we obtained (for details see Raudys and Skurikhina, 1994)

$$EP_N^{(\mathrm{RDA})} \approx \Phi\left\{ -\frac{\delta_\lambda}{2}\frac{\sqrt{1+\lambda T_\lambda}}{\sqrt{T_\mu \Sigma}} \right\}, \tag{8}$$

where

$$\delta_\lambda^2 = \frac{(\boldsymbol{\mu}'(\boldsymbol{\Sigma}+\lambda\mathbf{I})^{-1}\boldsymbol{\mu})^2}{\boldsymbol{\mu}'(\boldsymbol{\Sigma}+\lambda\mathbf{I})^{-1}\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda\mathbf{I})^{-1}\boldsymbol{\mu}},$$

the term T_λ is a certain function of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.

As

$$N \to \infty, EP_N{}^{RDA} \to P_\infty^{(\mathrm{RDA})} = \Phi\left\{ -\frac{\delta_\lambda}{2} \right\}.$$

An increase in the regularization parameter λ increases the asymptotic error $P_\infty^{(\mathrm{RDA})}$. The term T_λ, is trying to reduce the negative influence of T_Σ, the term responsible for estimation of the covariance matrix. Thus, the regularization can improve the small sample properties of the classifier. Therefore with an increase in λ, the generalization error decreases at first, and afterwards begins increasing. The optimal value of λ decreases with an increase in the learning-set size (Raudys and Skurikhina, 1994). For the GCCM model, after optimization with respect to λ, the resulting generalization error is smaller than both the generalization error of the Euclidean distance classifier and that of Fisher and pseudo Fisher.

3.5. Generalized discriminant analysis

No theoretical results have been obtained yet on the generalization error of the generalized robust Fisher linear classifier. Papers and a monograph of Kharin (1992) analyse robust statistical classifiers where the learning-set of each class is contaminated by vectors of the opposite pattern class.

4. Generalization of nonparametric classifiers

4.1. The minimum empirical error classifier

In the analysis of the Euclidean distance classifier, we have seen that there exist *favourable and unfavourable distributions* of the random vector $\mathbf{X}$. A similar situation arises in the analysis of small sample properties of nonparametric algorithms.

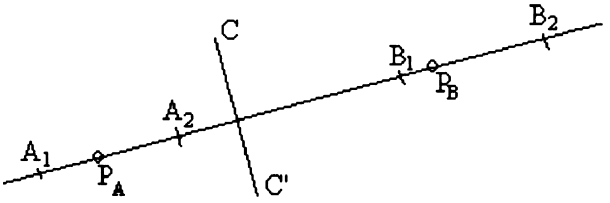

Fig. 3. The ''most favourable'' distribution of two pattern classes in the multivariate space.

Let us consider the following model of real multivariate distributions as *the favourable case*. Suppose, the pattern vectors are distributed on a straight line in multivariate feature space. Let the first class vectors be distributed in the interval (A_1,A_2) on this line, and the second class vectors be distributed in the interval (B_1,B_2). The classes do not overlap, and Euclidean distances $|A_1,A_2| < |A_2,B_1|$, $|B_1,B_2| < |A_2,B_1|$ (see Fig. 3).

Suppose now, that only one observation per class is available for training—P_A and P_B. Let us design a linear classifier with zero empirical error and the maximum margin between the discriminant hyperplane and both training vectors. Obviously, the linear decision boundary C–C′ $\mathbf{w'X} + w_o = 0$ will intersect the interval (A_2,B_1), and for this model of the pattern classes the generalization error will be zero. This provides a very *favourable* distribution of the pattern classes. Possibly, it *is the most favourable case*. It has a configuration very similar to the distribution of the classes π_3 and π_5 in Fig. 1, the most favourable Gaussian distribution for the Euclidean distance classifier.

In order to obtain low generalization errors one needs many more training examples *in unfavourable cases* of distributions of the pattern classes. The classes π_3 and π_4 in Fig. 1 represent the unfavourable case for the EDC. Such a configuration of the class conditional densities is also unfavourable for the minimum empirical error classifier design. The upper bounds for the true and estimated classification errors of the minimum empirical error classifier indicate that, in theory, very ''bad'' distributions of pattern classes can occur. Therefore in practice it is important to obtain results for intermediate cases.

In Raudys (1993) an analytical expression for the mean generalization error of the zero empirical error (ZEE) linear classifier (a particular case of the minimum empirical error classifier) was obtained for an intermediate case—''a more realistic situation''—a model of two spherical Gaussian distributions. A hypothetical ''random search'' (Monte-Carlo) training procedure was analysed theoretically. Here, one repeatedly generates many random discriminant hyperplanes $w_o + \mathbf{w}'\mathbf{x} = 0$ according to a particular prior distribution of the weights w_o, $\mathbf{w}$, defined by the particular *a priori* density of the weights $q_{\mathrm{prior}}(w_o,\mathbf{w})$. One selects only

those hyperplanes that classify all learning-set vectors without error, and the margin (the Euclidean distance between the discriminant hyperplane and the learning vector closest to it) exceeds Δ. In the statistical mechanics approach, a randomized training procedure of such type is called ''the Gibbs algorithm''. It is one possible training method from a variety of optimization techniques that can be used to find the weight vector. It is not the best choice in practice; however, it is very convenient for analytical investigations.

When $\Delta = 0$, we have the zero empirical error (ZEE) classifier. When $\Delta > 0$, we have the margin classifier. In Raudys (1993) *a mean expected probability of misclassification* EP_N of the pattern vectors that did not participate in the training was considered. The expectation was taken both with respect to $2N$ random training vectors and to the random character of generating $p + 1$ weights. When the *a priori* distribution $q_{\text{prior}}(w_o,\mathbf{w})$ of the $(p + 1)$-variate weight vector w_o, $\mathbf{w}$ is spherical Gaussian, only vague *a priori* information on the weights is used to design the classification rule. Thus, the classification rule is designed only on the information contained in the learning-set data.

Suppose, now, that *additional information* on the weights w_o, $\mathbf{w}$ *is available*. Let this vector be generated not at random, but found from *an additional data-set* by using the Euclidean distance classifier. In Part I of this paper, it was shown that such a weight vector can be obtained in batch-mode training the nonlinear SLP after the first iteration. Then the *prior* distribution $q_{\text{prior}}(w_o,\mathbf{w})$ will be narrower than the distribution obtained in the case of the random weight generation.

The above model allows us to calculate the mean expected classification error using the technique of *numerical integration*. The analysis of numerical results obtained for the spherical Gaussian model indicates that an increase in the expected classification error of the linear classifiers is in fact a function of the ratio p/N and distance δ, only. It depends on the prior distribution $q_{\text{prior}}(w_o,\mathbf{w})$ and the distance between the pattern classes δ. In Table 1, for 50-variate spherical Gaussian centred classes we present a relative increase in the mean expected classification error, the learning quantity, the ratio $\kappa = EP_N/P_\infty$, the zero empirical error classifier with: a) random; and b) EDC priors (the data from Raudys and Diciunas, 1996). For comparison in the same Table, we present κ values for two parametric classifiers (the Euclidean distance and the Fisher one; data from Raudys and Pikelis, 1980).

Clearly, the learning quantity κ depends on the type of classifier, the ratio N/p, and δ, the distance between the pattern classes. The Euclidean distance classifier enables us to design a classification rule in cases where the number of learning vectors $n = 2N$ is smaller than the number of features. However, this classifier makes assumptions that the components of the feature vector $\mathbf{X}$ are mutually independent. As a result, this classifier will not work well in certain applications. The Fisher DF allows us to evaluate the dependencies between the features but requires many more learning-set vectors. The zero empirical error classifier allows us to take into account statistical dependencies between the features and, at the same time, can be used in cases where the number of dimensions is higher than the number of learning examples.

Comparison of the last six columns of Table 1 with the previous five columns indicates that *the favourable (tight) prior distribution of the weights can reduce the generalization error dramatically*. Recall that this conclusion was obtained for the spherical Gaussian model of the pattern classes.

In order to analyse the character of the learning curves we used the data in Table 1 to plot the generalization error *versus* $(p/N)^S$ for different values of parameter S. For large sample sizes (when $2N \gg p$) we found that an increase in the generalization error $EP_N - P_\infty$ of the Fisher classifier (adaline rule) is proportional to p/N. This agrees with the asymptotical universal learning curves derived by Amari and Murata (1993). For very small p/N, however (when $n = N/2$ approaches p), an increase in the generalization error of the Fisher classifier is proportional to $(p/N)^2$, and only for very large N we have the linear relationship.

We found that the increase in the mean expected

Table 1
Learning quantity, ratio $\kappa = MEP_N/MEP_N/P_\infty$ of the Euclidean distance E, Fisher F and the zero empirical error (with random and ''Euclidean'' prior weights) classifiers *versus* N/p, learning set size/dimensionality ratio

E					F					ZEE with Gaussian priors					ZEE with Euclidean priors					N/p
1.82	2.34	3.09	3.66	4.22						2.16	3.76	10.0	25.1	71.2	1.63	1.99	2.70	3.47	4.42	0.16
1.70	2.03	2.41	2.65	2.87						2.04	3.43	8.58	20.7	56.9	1.48	1.69	2.12	2.57	3.08	0.24
1.54	1.70	1.84	1.92	1.99						1.88	2.93	6.77	15.3	41.7	1.29	1.40	1.66	1.91	2.16	0.40
1.43	1.50	1.55	1.58	1.61	2.05	3.39	8.40	19.7	52.0	1.74	2.57	5.58	12.3	31.7	1.17	1.25	1.43	1.60	1.77	0.60
1.30	1.32	1.33	1.34	1.35	1.62	2.15	3.61	5.95	10.6	1.56	2.16	4.34	9.13	22.5	1.08	1.13	1.26	1.37	1.48	1.0
1.18	1.17	1.16	1.16	1.17	1.33	1.51	1.93	2.47	3.27	1.35	1.73	3.09	6.04	14.1	1.03	1.06	1.13	1.21	1.27	2.0
1.08	1.07	1.06	1.06	1.06	1.14	1.19	1.31	1.44	1.61	1.16	1.32	2.06	3.59	7.68	1.01	1.02	1.07	1.10	1.14	5.0
1.04	1.03	1.03	1.03	1.03	1.07	1.09	1.15	1.20	1.27	1.08	1.19	1.59	2.53	4.98	1.01	1.01	1.03	1.06	1.09	10.0
1.02	1.02	1.02	1.02	1.02	1.04	1.05	1.07	1.10	1.13	1.04	1.10	1.34	1.86	3.35	1.01	1.01	1.02	1.04	1.06	20.0
1.01	1.01	1.01	1.01	1.01	1.01	1.02	1.03	1.04	1.05	1.02	1.04	1.15	1.39	2.11	1.01	1.01	1.02	1.03	1.04	50.0
1.68	2.56	3.76	4.65	5.50	1.68	2.56	3.76	4.65	5.50	1.68	2.56	3.76	4.65	5.50	1.68	2.56	3.76	4.65	5.50	δ
0.2	0.1	0.03	0.01	0.003	0.2	0.1	0.03	0.01	0.003	0.2	0.1	0.03	0.01	0.003	0.2	0.1	0.03	0.01	0.003	P_∞

generalization error of the minimum empirical error classifier is proportional to $(p/N)^S$. The order parameter S depends on p/N and δ. We have $S \approx 0.1$ for $\delta = 2.56$, and $S \approx 0.4$ for $\delta = 4.56$, when p/N is large. Parameter S increases with N: we have $S \approx 0.7$ for $\delta = 4.56$, and small p/N. According to the asymptotical universal learning curve theory, for very large N the parameter S approaches 1. Note, the results for the ZEE classifier with Gaussian priors in Table 1 are obtainable from the approximate formula:

$$MEP_N^{(\mathrm{ZEE})} \approx \Phi\left\{ -\frac{\delta}{2}\frac{1}{\sqrt{1+(1.6+0.18\delta)\left(\frac{p}{N}\right)^{1.8-\delta/5}}} \right\}, \tag{9}$$

that can be compared with analogical formulae derived for the parametric classifiers.

4.2. Intrinsic dimensionality

Consider the GCCM^r model $\mathrm{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_r)$, $\mathrm{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_r)$ with r nonzero eigenvalues of $\boldsymbol{\Sigma}_r$, which has already been considered in the previous section. This model implies that the distribution of the vector $\mathbf{X}$ lies in the subspace of dimensionality r. The effective dimensionality of such data $p^* = r$, and for this model with *the intrinsic dimensionality* equal to $r < p$, the small learning-set properties of the zero empirical error classifier can be analysed in the r-variate space. In this space, the r-variate vector $\mathbf{Y} = \mathbf{gX}$ is spherical Gaussian, and all the above conclusions derived for the p-variate spherical Gaussian model are valid. For the GCCM^r model, the small sample properties of the ZEE classifier are determined by the ratio r/N, and not by the formal dimensionality/sample size ratio p/N (for details see Raudys, 1993).

4.3. Maximal margin classifier

The numerical calculations performed according analytical equations derived for the SGC model indicate that an increase in Δ, the value of a bound for the margin, diminishes the mean expected classification error. We present six graphs: the mean expected error MEP_n *versus* Δ in Fig. 4. The graphs are calculated for a random Gaussian prior distribution of the weights, different p/N, and two values of p. The theoretical results indicate that an increase in the margin width can diminish the mean generalization error (two to three times in the given example).

The mean expected classification error is derived as a mean value averaged over those parts of learning-sets for which it is possible to obtain margins larger than Δ. Therefore this estimate is valid only for certain learning-sets. In the next section, we report experiments with the nonlinear SLP, that show that in spite of the fact that, on average, the mean generalization error decreases with Δ, for particular Gaussian learning-sets the generalization error has a peaking behaviour, i.e., it begins to increase when the margin becomes too large.

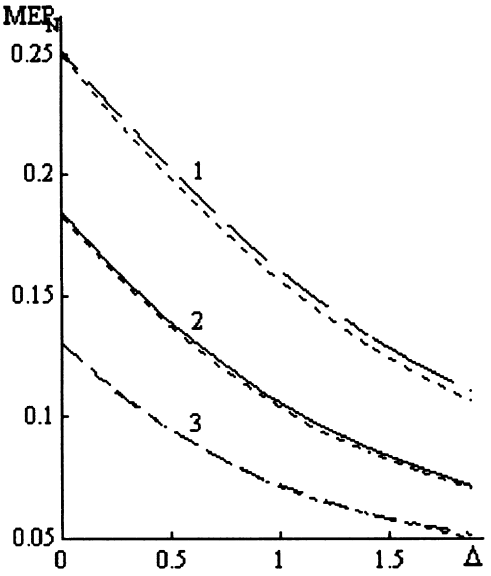

Fig. 4. The mean generalization error of the maximum margin classifier versus Δ, the bound for the margin: (1) $N = p/4$; (2) $N = p/2$; (3) $N = p$. Spherical Gaussian classes $\delta = 3.76$; $p = 20\&200$ (dots).

Clearly, in the spherical Gaussian case, the generalization error of the Euclidean distance classifier is much lower than that of the Fisher DF and ZEE classifier. This can be explained by the fact that while designing EDC, one estimates only sample mean vectors and ignores covariances. In order to design the Fisher DF one also needs to estimate the $p \times p$ covariance matrix. For a small number of features (when $p \ll 2N$) the generalization error of the Fisher classifier is lower than that of the zero empirical error classifier. However, for a large number of features (when p is close to $2N$ or exceeds $2N$), the minimum empirical error classifier nearly to outperforms the Fisher classifier.

As a *general conclusion* we can state that the nonparametric approach for designing the linear classifier generates reliable rules even in cases where the number of features is significantly larger than the number of training vectors. We do not need to estimate the class global parameters $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, the covariance matrix, and the means, when we reject the assumption that the classes are Gaussian The estimation of these parameters in a high-dimensional case is not favoured in classifier design. Additional information supplied as a prior distribution $q_{\mathrm{prior}}(w_o,\mathbf{w})$ can reduce the generalization error dramatically.

5. Learning-set size and generalization error of single-layer perceptrons. Simulation study

In Part I, we have demonstrated that there is no unique nonlinear SLP classifier. The SLP appears equivalent to a sequence of statistical classifiers. Which particular type of

classifier will be obtained depends on: the data, the cost function to be minimized, the optimization technique and its parameters, the stopping criteria. An objective of this section is to show that the theoretical results presented in Part I, and in the previous two sections, are valid for the nonlinear SLP classifier. In our simulations, we have used the GCCM and SGC data, the batch-mode BP training algorithm, the cost function of the sum of squares with the *sigmoid* activation function. At different moments of the training process, we calculated the generalization error analytically

$$P_N = \frac{1}{2}\Phi\left\{\frac{w_0 + \mathbf{w}'\mu^1}{\sqrt{\mathbf{w}'\Sigma^{-1}\mathbf{w}}}\right\} + \frac{1}{2}\Phi\left\{\frac{w_0 + \mathbf{w}'\mu_2}{\sqrt{\mathbf{w}'\Sigma_{-1}\mathbf{w}}}\right\}. \quad (10)$$

In most experiments, except when stated otherwise, the conditions ***E*** were fulfilled. Most often we used the target values $t_1 = 0$ and $t_2 = 1$. Close targets, e.g., $t_1 = 0.45$ and $t_2 = 0.55$, make the sigmoid activation function act as a linear function. Thus after minimizing the cost function we obtain the standard Fisher DF. Therefore in experiments with the non-limiting targets, we used: $t_1 = 0.1$ and $t_2 = 0.9$, the target values recommended by Rumelhart et al. (1986).

5.1. The SLP and parametric classifiers

Target values essentially influence the learning process when the empirical classification error is small. In Fig. 5 we plot the dependence of the generalization error on the number of iterations t for the 20-variate GCCM data **C**. Both graphs were obtained for one learning set with $N = 14$ vectors from each class; however different target values were used. To make the learning process faster we used a slightly increasing learning-step: $\eta = 10{*}1.0005^t$. After the first iteration we got EDC with $P_{\text{gen}} = 0.058$ in both cases. At the beginning of training, the generalization error decreases: there we have the regularized DA. The different target values, however, lead to different classification rules later: with targets ''0.1&0.9'' we are approaching the standard Fisher DF with $P_{\text{gen}} = 0.093$; and with targets ''0&1'' we are approaching the maximal margin classifier with a significantly smaller generalization error.

More information concerning the generalization error can be obtained from average values of the generalization error. We used the 20-variate GCCM data **C** again, and compared experimental learning curves of the SLP with those of the

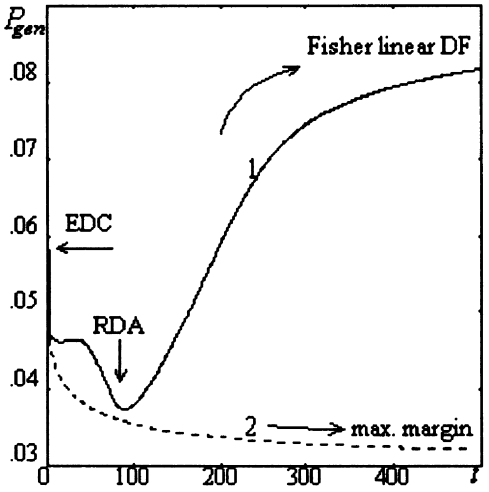

Fig. 5. Effect of targets on the BP training process: the generalization error *versus* t, the number of iterations. (1) targets $t_1 = 0.1$, $t_2 = 0.9$; (2) $t_1 = 0$, $t_2 = 1$. 20–variate GCCM data **C**; $N = 14$.

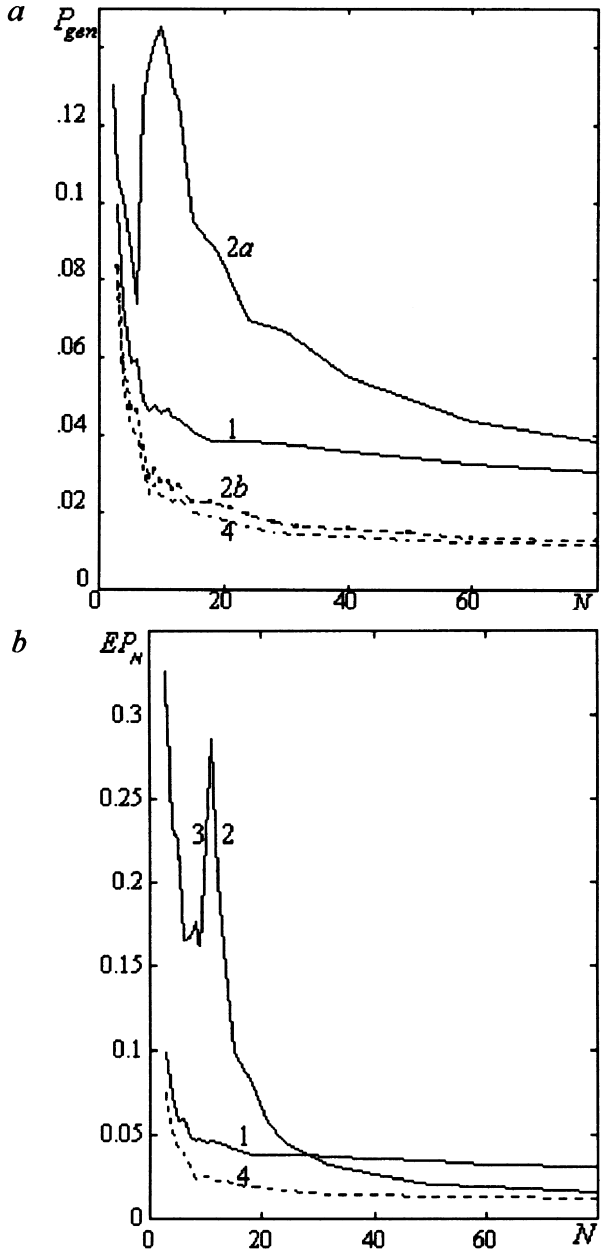

Fig. 6. ''Scissors effect'' in practice: the average generalization error *versus* N. (a) The nonlinear SLP: (1) after the first iteration; (2a) after 500 iterations (targets ''0.1&0.9''); (2b) after 100 iterations (targets ''0&1''); (4) after the optimal number of iterations. (b) The statistical classifiers: (1) EDC; (2) Fisher DF; (3) pseudo Fisher DF; (4) regularized DA for optimal l. 20-variate GCCM data **C**.

EDC, the Fisher linear classifier and the regularized DA with the optimal value of l (Fig. 6a and b). Each curve is an average value obtained from the same 50 randomly selected learning-sets.

In spite of the fact that the targets ''0.1&0.9'' are far from being close, and force the sigmoid activation function to act as a linear function, we see that the learning curve of the nonlinear SLP exhibits a clear *peaking behaviour* in the interval ($1 < n < p$). The shape of the SLP learning curve with targets ''0.1&0.9'' (2a in Fig. 6a) resembles the experimental and theoretical curves for the standard Fisher DF and the Fisher DF with pseudo-inverse (2 and 3 in Fig. 6b—simulation, and 2 and 3 in Fig. 2—theory). Both combinations of the target values yield EDC after the first iteration—the learning curves numbered by 1 in Fig. 6a and b are identical, and both experimental curves coincide with the theoretical one (1 in Fig. 2).

For $n > \frac{1}{2}p$, the experimental learning curve 2 in Fig. 6b becomes very close to the theoretical curve 2 for the Fisher DF in Fig. 2. The same can be said about the experimental learning curve 4 in Fig. 6a of the SLP after the optimal number of iterations t_{opt} (targets ''0 and 1''), and curve 4 in Fig. 6b for the regularized DA with the optimal value of the regularization parameter λ_{opt}. To find the optimal values t_{opt}, and λ_{opt} we used Eq. (9) to calculate the generalization error. The learning curve 2b in Fig. 6a corresponds to targets ''0&1''. For this type of the data, the targets ''0&1'' allow us to obtain an essentially smaller generalization error, and to confirm our theoretical considerations as to the importance of choosing proper target values.

Fig. 6b demonstrates a clear ''scissors effect'': for small learning-sets up to $N \approx 30$ it is preferable to use the simple structured EDC than the complex Fisher linear DF, and, *vice versa*; in large learning-set cases, the Fisher linear DF can be used more efficiently. The same conclusion is valid for the SLP: for small learning-sets it is preferable to train the SLP for a short time, and, *vice versa*, in large learning-set cases, one needs to use more iterations. Other regularizing factors, such as target and learning-step values, operate simultaneously, and the problem to find optimal values of all these parameters is not easy.

Theoretical considerations on the effective dimensionality p^* of the EDC indicate situations where the SLP can be trained perfectly on very small learning-sets. In Section 2 we have analysed an extreme case: the 100-variate GCCM data model **D1** with the effective dimensionality p^* close to 1: $p^* \approx 1.05$. Theoretical calculation gives the generalization error of EDC $EP_N^{(\text{E})} = 0.0318$. In a series of 10 experiments with learning-sets containing *five* 100-variate vectors from each class, we have obtained a very small generalization error: the EDC yielded, on average, 0.039 error with standard deviation 0.009. The same result was obtained for the SLP after the first iteration.

The same theoretical considerations on the effective dimensionality p^* indicate situations where is difficult to train the SLP classifier. The 100-variate GCCM data model **D2** with $p^* \approx 10^{10}$ is a perfect example. Theoretical calculation gives the generalization error of EDC $EP_N^{(\text{E})} = 0.4997$. In a series of 10 experiments with learning-sets of size $N = 200$, we have obtained a very high generalization error—the EDC yielded, on average, 0.4997 of the error. After the first iteration, the SLP gave the same result. The Fisher DF, however, yielded a ''reasonable'' error 0.058, i.e., 1.93 times higher than the asymptotic error $P_\infty^{(\text{E})} = 0.03$. This corresponds to Eq. (6), and Table 1 for $\delta = 3.76$ and $N = 2p$. Note this type of almost singular data is a very hard problem for BP training. In such a situation, a ''decorrelating'' transformation

$$\mathbf{Y} = \mathbf{TX} \tag{11}$$

is very helpful. In Eq. (11) $\mathbf{T} = \mathbf{D}^{-1/2}\mathbf{G}$, and $\mathbf{G}$ is an orthonormal $p \times p$ matrix such that $\mathbf{GSG}' = \mathbf{D}$ (diagonal matrix of the eigenvalues). Then, in a new space Ω_{Y}, again we obtain the EDC after the first iteration; however, this classification rule is equivalent to Fisher's rule in the original Ω_{X} space.

5.2. The SLP and nonparametric classifiers

The weights of the nonlinear SLP are increasing when we use limiting target values and have small empirical error. The SLP then becomes similar to the minimum empirical error and maximum margin classifiers. We demonstrate that the theoretical results of the previous sections are consistent with simulation studies.

In experiments, we use exactly the same type of data as in the previous analytical study—the multivariate spherical Gaussian data. In order to have possibility of increasing the margin and analysing the influence of the margin width on the generalization error, we have chosen a relatively large Mahalanobis distance $\delta = 3.76$ (the asymptotic and Bayes error $P_{\text{B}} = 0.03$), a small learning-set size ($N = 100$) and have used an exponentially increasing learning-step $\eta = \eta_0 \times 1.1^t$.

In the previous section we have seen that in the random search optimization (the Gibbs algorithm), the generalization error essentially depends on the prior distribution of the weight vector. Thus, one of the objectives of this subsection is to verify whether the starting position $\text{W}_{0(0)}$, $\mathbf{W}_{(0)}$ of the weight vector can also be important for the accuracy of the final position of the weight vector in the gradient search procedure. In all our experiments so far, we fulfilled the learning conditions ***E***, i.e., just after the first iteration we used to obtain EDC. There are many theoretical arguments that EDC is the best sample-based classification rule for spherical Gaussian patterns. In order to overcome this complication, nearly in all of the experiments, we initialized the weights randomly: starting weights $\text{W}_{0(0)}$, $\mathbf{W}_{(0)}$ were chosen from the Gaussian distribution: $(\text{W}_{0(0)}, \mathbf{W}_{(0)}')' \sim \text{N}(\mathbf{0}, \sigma^2\mathbf{I})$.

Low and moderate values of σ^2 lead to EDC, while very high ones lead to an immediate saturation of the activation function and can stop the training just after the first iteration.

In order to choose σ^2 we calculated the variance V of a double random variable $W_{0(0)} + \mathbf{W}_{(0)}'\mathbf{X}$, where $(W_{0(0)},\mathbf{W}_{(0)}')' \sim N(0,\mathbf{I}\sigma^2)$, and $\mathbf{X} \sim N(\boldsymbol{\mu}_i,\mathbf{I})$. For symmetrically situated classes ($\boldsymbol{\mu} = -\boldsymbol{\mu}_1 = \boldsymbol{\mu}/2$), variance $V\{W_{0(0)} + \mathbf{W}_{(0)}'\mathbf{X}\} = \sigma^2(p + 1 + \delta^2)$. This variance influences the saturation of the activation function $o(W_{0(0)} + \mathbf{W}_{(0)}'\mathbf{X})$ and, consequently, affects the training process. Hence, when analysing the influence of weight initialization on the generalization error, we defined the initialization variance σ^2 as a function of dimensionality:

$$\sigma^2 = (\sigma_\alpha)^2 \frac{1}{p+1+\delta^2}. \tag{12}$$

Thus, the coefficient σ_α controls width of the weight initialization interval.

Two graphs in Fig. 7 are typical of this type of experiment with spherical Gaussian data. When starting from *zero* initial weights, we obtain the best sample-based classifier (EDC) just after the first iteration. Therefore, we have a constant increase in the generalization error later (graph 2). After starting from a distant, *inexact* initial weight vector, the training process ''corrects'' the weight vector, and therefore reduces the generalization error at first; however, later on, it leads to the maximal margin classifier. This classifier is far from being the best one for spherical Gaussian patterns. Therefore, after reaching the minimum, we obtain an increase and approach the learning curve for the zero weight initialization: graph 1 (dots, 25^{th} to 150^{th} iterations only). The minimum of graph 1 with the random initialization is notably higher than the minimum of generalization with zero weight initialization. This experiment indicates that random initialization and a search for the *maximal* margin is not always the best strategy in classifier design.

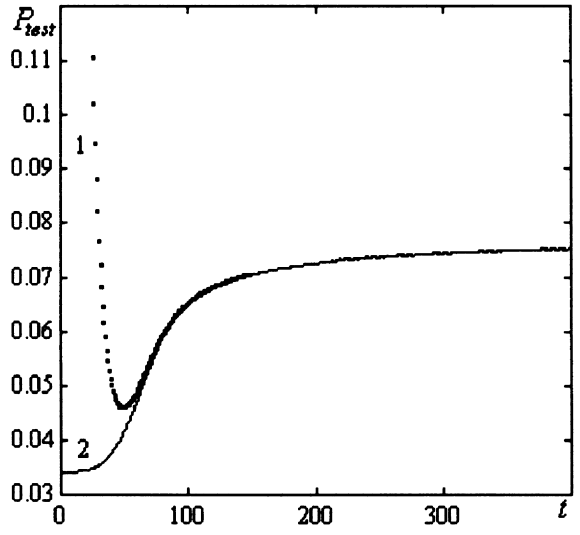

Fig. 7. Effect of weights initialization on SLP training: the generalization error *versus t*, the number of iterations. Spherical Gaussian classes, $\delta = 3.76$, $p = 100$, $N = 200$. Varying learning-step $n = 1.1^t$, random $N(0,\sigma^2)$ weight initialization: (1) $\sigma^2 = 0.1$; (2) $\sigma^2 = 0.0$.

Comparisons of empirical and theoretical results are summarized in Fig. 8. For spherical Gaussian data, with a decrease in the value of σ_α, the generalization error decreases, and approaches values obtained for EDC (graph 9—theory and simulation). High values of σ_α often cause an immediate saturation of the activation function and a small gradient of the cost function. Then the perceptron does not learn or learns extremely slowly. In Fig. 8, for dimensionality $p = 200$ and a bound for the margin $\Delta = 0$, by ''x'' we denoted average generalization errors for three different initialization intervals defined by σ_α: $\sigma_\alpha = 5$ (utmost upper point), $\sigma_\alpha = 2$ (a point on graph 4) and $\sigma_\alpha = 0$ (a point on graph 8). All experimental graphs are average values obtained in 50 independent learning experiments. We have chosen the initialization with $\sigma_\alpha = 2$ as sufficiently wide, yielding a generalization error close to the theoretical values calculated for *random* prior distribution of the weights.

Graphs 1, 2 and 3 for the zero empirical error classifier are *theoretical* ones. They are calculated for a random Gaussian prior distribution of weights, and the bound for the margin $\Delta = 0$, 0.4 and 0.8, respectively. Graphs 4, 5 and 6 are *experimental* ones (initialization interval $\sigma_\alpha = 2$). These graphs are average values found only from those learning sets whose the margin values are higher than $\Delta = 0$, 0.4 and 0.8. Note that for $\delta = 3.76$, $N = 100$, and $p > 100$ we succeeded in obtaining zero empirical error and large margins—$M \geq \Delta = 0.8$—in almost all the experiments.

Both the theoretical and the simulation experiments indicate that with an increase in margin width *on average* the generalization error decreases. In *all* the *individual* training experiments performed with different learning-sets, however, we noticed the *overtraining effect*—an excessive growth of margin width increases the generalization error. This is no surprise, since the maximum margin classifier is not the optimal classification rule for spherical Gaussian classes. The Euclidean distance classifier (SLP after the first iteration) is the optimal sample-based classification rule for this model of the pattern classes. Another explanation of inconsistency of the theoretical and simulation results is embodied in the fact that we calculated the expectation of the generalization error over random Gaussian prior distribution $q_{\text{prior}}(w_o,\mathbf{w})$ in the previous section's ''random search'' optimization procedure (the Gibbs algorithm). During gradient training, however, the weight vector $(w_o,\mathbf{w})$ is not random: it moves according a certain trajectory, where a distribution of values $(w_o,\mathbf{w})$ differs from the Gaussian model. This means that the theoretical estimates for the margin classifiers trained by the Gibbs algorithm should be considered with a certain prudence.

Graph 7 in Fig. 8 is obtained for the case of random weight initialization ($\sigma_\alpha = 2$) too, however, it represents the mean generalization error calculated using *optimal margin values* M_{opt} evaluated for each particular learning-set. To find M_{opt}, the optimal number of iterations t_{opt} was evaluated from the minimum of the generalization error calculated analytically after each iteration. There we have used an

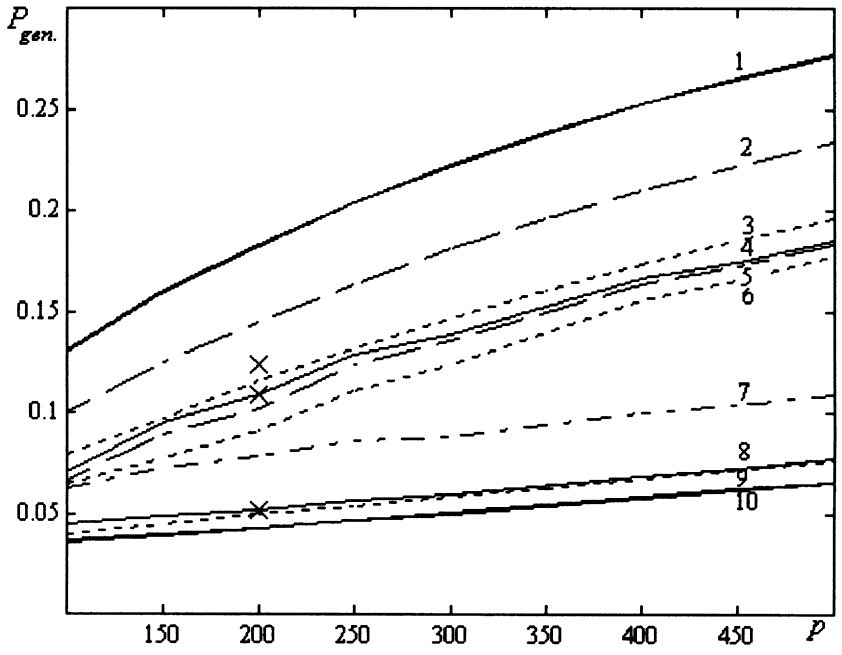

Fig. 8. Generalization error of the zero empirical error and (maximum) margin classifiers *versus* dimensionality. Theoretical (1,2,3,9,10) and simulation (4,5,6,7,8) results (average values from 50 independent experiments). Graphs 1 and 4, margin $M \geq 0$; graphs 2 and 5, margin $M \geq 0.4$; graphs 3 and 6, margin $M \geq 0.8$; 7, the optimal margin M obtained for the optimal number of iterations $t_{optimal}$; 8, $\sigma_\alpha = 0.0$ and $M \geq 0$ (simulation); 9, EDC (theory and simulation); 10, the ZEE classifier, ''initialization by EDC'', and the additional learning set (theory).

additional information. Therefore graph 7 is much lower than any other graph depicted for the fixed value of $\sigma_\alpha = 2$. It demonstrates a definite overtraining. We will return to this effect in the following section.

Graphs 8, 9 and 10 in Fig. 8 are presented in order to illustrate *a positive influence of more exact (non-random) weight initialization* on SLP training. Graph 8 corresponds to SLP, trained from zero initial weights, and the margin $M \geq 0$. In this situation, after the first training iteration we obtain EDC—the optimal sample-based classification rule for spherical Gaussian classes. Hence, roughly speaking, SLP is initialized by the weights of the optimal classifier. Graph 9 corresponds to the generalization error of EDC. Graph 10 (the ZEE classifier) is calculated from the theoretical equations of the previous section for the case where the prior distribution of weights was determined by EDC calculated from an additional learning-set. A comparison of graph 8 with 4 indicates that a *''correct initialization'' reduces the generalization error dramatically*. This means that *the perceptron weights can store a large amount of useful information*.

The graphs in Fig. 8 indicate that the theoretical estimates are rather close to the experimental ones for spherical Gaussian classes. It is desirable to discuss the case of more general distributions. The analysis of EDC has shown: dependent on p^*, the effective dimensionality, this classifier can be trained even with very small learning-sets. In other extreme cases, *any number of learning vectors is insufficient to train the EDC*. Similar considerations are valid for other parametric classifiers (e.g., the Fisher linear DF—adaline, or the SLP with non-limit, close target values). In this sense, the nonparametric minimum empirical error classifier is more favourable: in principle, error bounds exist that give lower estimates than that for EDC for extremely large p^*. However, extreme, unfavourable cases in both approaches are not frequently met in real world problems.

6. Dynamics of the generalization error. Overtraining

The above theoretical results throw new light upon the overtraining problem. The overtraining effect is caused by two factors. First of all, it is a difference between the cost function surfaces, obtained from the learning-set data, and that obtained from the test-set data (a general population). On the way from the starting vector $\mathbf{w}_{(0)}$ to the minimum of the cost function $\hat{\mathbf{w}}$, we can pass $\mathbf{w}^*$, the minimum of the cost function surface of the general population. On the whole, the larger the difference, the larger the overtraining effect can be expected to be. This factor, however, explains only a proportion of the cases where the overtraining effect is observed: dependent on the configuration of the triangle $(\mathbf{w}_{(0)},\hat{\mathbf{w}},\mathbf{w}^*)$ we can either observe or not observe the overtraining.

In many simulation studies, we observe overtraining in all training experiments. This can be explained by another factor: a change in the type of statistical classifier that occurs with an increase in the number of iterations. One of these classifiers appears to be the best one in the finite learning-set size situation. Overtraining can appear when the weights are

small and the activation function acts as a linear one. Then we move from the EDC towards the regularized DA and the Fisher DF—see, for example, curve 1 in Fig. 5.

If the covariance matrices of the pattern classes are different and/or the data is not Gaussian, then the Fisher linear classifier is no longer an asymptotically optimal classification rule. In order to obtain the best linear classifier, one needs to evaluate higher order statistical moments than the mean vectors and the common covariance matrix. The generalized DA and the minimum empirical error classifiers can be the best choice in that case. Therefore the overtraining can occur later, when the weights are large and the activation function acts as a nonlinear one. Then we move from the Fisher classifier towards the generalized Fisher, and further, towards the minimum empirical error classifier. To do this, sometimes it is necessary to add a supplementary anti-regularization term. In Part I we have demonstrated such an example.

For special types of data sets, the best classification rule is the maximal margin classifier. An example of such type of data was presented in Fig. 2 in part I (data **B**). In this case we cannot get any overtraining at all. This model with ''the sharp edges'', however, is not characteristic of the real world problems. Most often we have ''fuzzy'' boundaries of the pattern classes, and obtain the minimum earlier before the maximal margin classifier is reached in SLP training (see Fig. 7).

7. A few additional bibliographical remarks

Different fields (conventional multivariate statistics, neural nets, computational learning theory, AI, machine learning) address the supervised learning problem. All these fields have their own jargon, their own mathematical models, their own concerns, and their own results. And for the most part they don't interact (see the preface in the book edited by Wolpert, 1995a). Hence, it is very difficult to compare the results obtained by different approaches. An attempt to do this was made by Wolpert in his paper (Wolpert, 1995b), and in a dozen papers in the book mentioned. In addition to the remarks in Section 1, we try to compare the results discussed in this paper with those obtained by other approaches.

In multivariate analysis, there are two types of asymptotic investigation of the accuracy of prediction and classification procedures. In one of them, the dimensionality p is kept constant, and the sample size n is increased. In another one, n and p increase simultaneously. The analysis shows that the second approach is much more accurate. In particular, *the difference can be noticed for small values of the classification error and the ratio n/p* (Pikelis, 1976; Wyman et al., 1990; Takeshita and Toriwaki, 1995). Our analysis and that using statistical mechanics (Haussler et al., 1994) showed a qualitative difference between the learning curves of the ZEE classifier (see, e.g., Section 4.1) and the VC error bounds, as well as universal learning curves obtained asymptotically when only $n \to \infty$.

A majority of the results obtained for the classification problem (categorical ''0–1'' loss) agree with the results for a continuous loss obtained by the statistical mechanics approach. Sjoberg and Ljung (1992) have indicated that regularized linear regression can be obtained as a consequence of an increase in the number of iterations while training the linear SLP. For the linear prediction model $y = \mathbf{w}'\mathbf{x} + \xi$, Hansen (1993) showed that the generalization error

$$\bar{\epsilon}_G = \sigma_\xi^2 \left(1 + \eta\sigma_{\mathbf{x}}^2 \frac{2p}{n-p} + \frac{p-1}{n-p} \right), \tag{13}$$

where η is the learning-step parameter in the BP training, σ_ξ^2 is the variance of noise ξ, $\sigma_{\mathbf{x}}^2$ is the variance of an input vector $\mathbf{x}$ in the model $\mathbf{x} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma_{\mathbf{x}}^2)$. The term $\frac{p-1}{n-p}$ agrees with the term T_Σ in Eq. (6) and much earlier result of Davisson (1965). The term $\eta\sigma_{\mathbf{x}}^2 \frac{2p}{n-p}$ arises due afinite value of the learning-step η. A trade-off between the efficiency of learning and the minimization of the classification error was firstly analysed in Amari (1967), where the following fundamental result was obtained: the weight vector $\hat{\mathbf{w}}_t$ of the linear perceptron trained by the gradient descent training asymptotically is a Gaussian random variable. Its mean value is $\hat{\mathbf{w}}$, the minimum of the cost function of the sum of squares, and the covariance matrix of $\hat{\mathbf{w}}_t$ is proportional to η. A similar result to Eq. (13) was also obtained by Bös (1996) who analysed the accuracy of on-line and off-line training. For the linear model it was shown that the optimal selection of η can help to obtain practically the same accuracy in both the types of learning. Analogously to the generalization error of the pseudo Fisher linear DF, the presence of the minimum in the learning curve ''generalization *versus* the learning-set size'' was noticed in the linear prediction by Krogh and Hertz (1992), and Bös (1996).

An important conclusion of this paper is that the characters of the learning curves depend on the data. For the statistical models this was first noticed in Raudys (1967). For some models the character is determined by the intrinsic dimensionality of the data, however, the definition ''intrinsic dimensionality'' is not unique. For example, for the GCCM^r model $\mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_r)$, $\mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_r)$ the effective dimensionality for the EDC $p^* = r$. However, the expected error of the standard Fisher linear DF is determined by Eq. (6), and the ''intrinsic dimensionality'' r here plays no role. The conclusion that the data type strongly affects the character of the learning curve is confirmed by modern approaches, too (Kowalczyk, 1996). Possibly, it agrees with the universal convergence theorem of Amari and Murata (1993) $EH_n = H\infty + L^*/(2n)$, where for the unfaithful (unrealizable) network: $L^* = \mathrm{tr}\mathbf{K}^{-1}\mathbf{G}$, where $\mathbf{K}$ is the Hessian matrix, and $\mathbf{G}$ is the Fisher information matrix. In this approach, unfortunately, no explicit and/or numerical estimates have yet been obtained for the GCCM model.

In principle, the multivariate statistical analysis allows us to obtain the generalization error for a wide variety of multivariate statistical models. For particular models, one obtains absolutely exact results. However, there are two fundamental difficulties. First of all, true distribution densities of the pattern classes are usually unknown. As we have seen, in certain GCCM configurations the EDC can be extremely sensitive or extremely insensitive to the learning-set size. A similar situation arises with the standard Fisher linear DF, where one can invent very unfaithful data models, where the Fisher DF is extremely sensitive to the learning-set size. One can argue that theoretical estimates for these classifiers are useless. Nevertheless, simulations with real world data-sets (see, e.g., Raudys and Pikelis, 1980) show the theoretical estimates derived for the Gaussian data to be fairly close to the experimental results. The dependence of statistical inference on unknown parameters of the models is a general drawback of all statistical techniques, and will not be discussed here.

The second drawback of statistical analysis is that the analytical formulae are often very complex. Sitgreaves (1961) derived the exact formula for the expected classification error of the standard Fisher linear discriminant function in the form of a five times infinite sum of products of certain hypergeometric functions. We had similar problems with the EDC, the quadratic DF, and the ZEE classifier. For the ZEE classifier, after obtaining the exact formula we obtained a simple asymptotic expansion (Raudys, 1993), but its accuracy was low. A more exact expansion (Basalykas et al., 1996) was accurate enough but required solution of a certain differential equation. A similar numerical difficulty was met by Meir (1995) who used the statistical physics approach, and derived the generalization error equations for three versions of SLP.

We have reviewed only a proportion of the known results. In MLP classifier design, the weights of the hidden layer affect all outputs. Thus they are common for all outputs of the network. In statistics, it has been demonstrated that in the thermodynamic limit, under certain conditions, parameters of probabilistic models that are common for all classes asymptotically do not affect the increase in the generalization error (Raudys, 1972; Deev, 1974; Meshalkin and Serdobolskij, 1978). This result is very important for MLP analysis, where the weights of the hidden layer affect all the outputs.

8. Concluding remarks

The analysis of the SLP as a dynamical process allowed us to follow the neurone's evolution: with an increase in the number of iterations the weights gradually increase, the actual slope of the activation function changes, and gradually, step-by-step, seven known statistical classifiers can be obtained. We stopped training at different moments of time, used the multivariate statistical analysis techniques, and found that small learning-set properties of SLP change in time. Our simulations confirmed the theoretical conclusions.

The main corollary of this paper is that, in order to obtain the best generalization in the training process, one needs to control the complexity of the SLP classifier. Recall that the best type of the classifier depends on the learning-data: its size and configuration. For simple-structured data (e.g., spherical Gaussian) and small learning-sets one needs to use simple classifiers (e.g., EDC). For complex (e.g., non-Gaussian data with statistically dependent variables, different covariance matrices, multi-modal classes, etc.) and large data-sets, one can and sometimes must use complex classifiers such as the minimum empirical error one.

In Part I, we enumerated a number of *means which can help to control the complexity* of the SLP classifier. The means are associated either with the type of cost function or with the optimization procedure. In addition to the known complexity control techniques, a few new ones were proposed. We stressed the importance of the target values, a gradual increase of the learning-step, and the antiregularization term, if we wish to obtain the minimum empirical error or the maximum margin classifiers. Another complexity control technique is to move the centre of the learning-data into the origin of coordinates. More suggestions for *transforming the data* can be given *to make the data simple (spherical)*, in which case the best classifier is the EDC. The use of data transformations with a view to simplifying the data structure is, in fact, an *introduction of new additional information into the training process*. If the information is correct, it can help reduce the generalization error.

We became convinced that *successful initialization* of the perceptron leads to the best results at the very beginning of the training process, and helps to reduce the training time and the generalization error. Therefore, in practice, it is important to utilize this favourable peculiarity of the perceptron. This can be done by introducing any prior information into the weight initialization process. There are a number of practical means to use the additional information for the weight initialization and the data transformations. These means will be discussed in subsequent publications.

Among other questions not discussed here are the MLP and on-line training. No doubt, in on-line training, where the learning vectors are presented in a sequence, we obtain the simplest classifiers at first, and the most complex ones last. Possibly, a proportion of the results discussed here can be useful during on-line learning analysis. In the output layer, the MLP classifier includes a nonlinear SLP. We believe that the peculiarities of the nonlinear single-layer perceptron will introduce themselves there. In training, after the nonlinear target transformation in the output layer, the hidden layer neurones are fed by non limiting values of the targets. In addition, the target values are different for each learning vector. Our analysis has shown that the target values play an especially important role in perceptron training. Moreover, we have already mentioned that the weights of the hidden

layer affect all outputs. Investigations in statistics have indicated that the common parameters are less important. These two pecularities of the MLP deserves a more attentive investigation.

Acknowledgements

The author thanks Valdas Diciunas and an anonymous referee for useful remarks, and E.R. Davies for his aid in preparing a final version of the paper.

References

Amari, S. (1967). A theory of adaptive pattern classifiers. *IEEE Trans. Electron. Comput., EC-16*, 299–307.

Amari, S. (1993). A universal theorem on learning curves. *Neural Networks, 6*, 161–166.

Amari, S., Fujita, N., & Shinomoto, S. (1992). Four types of learning curves. *Neural Computation, 4*, 605–618.

Amari, S., & Murata, N. (1993). Statistical theory of learning curves under entropy loss criterion. *Neural Computation, 5*, 140–153.

Barkai, N., Seung, H. S., & Sompolinsky, H. (1993). Scaling laws in learning classification tasks. *Phys. Rev. Letters, 70 (20)*, 3167–3170.

Basalykas, A., Diciunas V., & Raudys S. (1996,1997). On expected probability of misclassification of zero empirical error classifier. Vilnius, **7**(2), 137–154; **8**(2), 310–311.

Bös, S. (1996), Learning curves of *on-line* and *off-line* training, *Proc. ICANN'96*, Bohum.

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Tran. Elec. Comp., EC-14*, 325–334.

Davisson, L. O. (1965). The prediction error of stationary Gaussian time series of unknown covariance. *IEEE Trans. on Information Theory*, IT-11, 527–532.

Deev, A. D. (1970). Representation of statistics of discriminant analysis and asymptotic expansions in dimensionalities comparable with sample size. *Reports of Ac. of Sci. of The USSR, 195 (4)*, 756–762. in Russian.

Deev, A. D. (1972). Asymptotic expansions for distributions of statistics W, M, W* in discriminant analysis. In *Statistical Methods of Classification, Vol. 31*. Moscow: Moscow University Press, pp. 6–57 (in Russian).

Deev, A. D. (1974). Discriminant function designed on independent blocks of variables. *Engineering Cybernetics, USSR J., 12*, 153–156.

Estes, S. E. (1965). Measurement selection for linear discriminant used in pattern classification. Ph. D. dissertation, Stanford University, Stanford CA.

Gyorgyi, G., & Tishby N. (1990). Statistical theory of learning a rule. In K. Thuemann & R. H. Koeberle (Eds.), *Neural Networks and Spin Glasses*, Singapore: World Scientific, pp. 31–36.

Hansen, L. K. (1993). Stochastic linear learning: exact test and training error averages. *Neural Networks, 4*, 393–396.

Haussler, D., Kearns, M., Seung, H. S., & Tishby, N. (1994). Rigorous learning curves from statistical mechanics. In Proc. *7th Annual ACM Conf. on Comput. Learning Theory*, pp. 76-87.

Hughes, G. F. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE Trans. on Information Theory, IT-14*, 55–63.

Jain, A., & Chandrasekaran, B. (1982). Dimensionality and sample size considerations in pattern recognition practice. In *Handbook of statistics, Vol. 2*. North Holland, pp. 835–855.

John, S. (1961). Errors in discrimination. *Ann. Math. Statistics, 32*, 1125–1144.

Kanal, L., & Chandrasekaran, B. (1971). On dimensionality and sample size in statistical pattern classification. *Pattern Recognition, 3*, 238–255.

Kharin, Yu. S. (1992). *Robustness in Statistical Pattern Recognition*. Minsk: Universitetskoe Publishing House. English translation: Kluwer, Dordrecht, 1996

Kowalczyk, A. (1996). Model of generalisation error in learning systems with training error selection. In S. I. Amari, L. Xu, L. W. Chan, I. King, and K. S. Leung (Eds.), *Progress in neural information processing, Proc. ICONIP'96*, Hong Kong. Springer, pp. 180–187.

Krogh, A., & Hertz, J. A. (1992), A simple weight decay can improve generalization. In J. Moody, S. J. Hanson, & R. Lippmann, (Eds.), *Advances in neural information processing Vol. 4*, pp. 950–957.

Levin, E., Tishby, N., & Solla, S. A. (1990). A statistical approach to generalization in layered neural networks. *Proceedings of the IEEE, 78*, 2133–2150.

Lbov, G. S. (1966). On representativeness of the sample size while choosing the effective measurement system. In N.G. Zagoruiko, (Ed.), *Computing systems, Issue 22*. Novosibirsk: Inst. of Math. Press, pp. 39-58 (in Russian).

Meir, R. (1995). Empirical risk minimization versus maximum-likelihood estimation: a case study. *Neural Computation, 1*, 144–157.

Meshalkin, L. D. (1976). Assignment of numerical values to nominal variables. In S. Raudys, & L. Meshalkin (Eds.), *Statistical methods of control, Issue 14*. Vilnius: Inst. of Physics and Math. Press, pp. 49–56 (in Russian).

Meshalkin, L. D., & Serdobolskij, V. I. (1978). Errors in classifying multivariate observations. *Theory of Probabilities and Applications, 23 (4)*, 772–781. in Russian.

Okamoto, M. (1963). An asymptotic expansion for the distribution of linear discriminant function. *Ann. Math. Statistics*, **34**, 1286–1301; **39**, 1358–1359.

Opper, M., & Haussler, D. (1991). Calculation of the learning curve of Bayes optimal classification algorithm for learning perceptron with noise. In *Proc. 4th Annual ACM Conf. on Comput. Learning Theory*, pp. 75–87.

Pikelis, V. (1974). Analysis of learning speed of three linear classifiers. Ph.D. dissertation, Inst. of Physics and Mathematics, Vilnius, pp. 1–136 (in Russian).

Pikelis, V. (1976). Comparison of methods of computing the expected classification errors. *Engineering Cybernetics (USSR J.), N5*, 59–63. in Russian.

Rao, C. R. (1949). On some problems arising of discrimination with multiple characters. *Sankya, 9*, 343–365.

Raudys, S. (1967). On determining the training sample size of a linear classifier. In N. Zagoruiko (Ed.) *Computing Systems, Vol. 28*. Novosibirsk: Nauka, Institute of Mathematics, Academy of Sciences USSR, pp. 79–87 (in Russian).

Raudys, S. (1970). On the problems of sample size in pattern recognition. In *Proc., 2nd All-Union Conf. Statistical Methods in Control Theory, Vol. 2*. Moscow: Nauka, pp. 64–67 (in Russian).

Raudys, S. (1972). On the amount of a priori information in designing the classification algorithm. *Proc. Acad. of Sci. of USSR Technical. Cybernetics, 4*, 168–174.

Raudys, S. (1993). On shape of pattern error function, initializations and intrinsic dimensionality in ANN classifier design. *Informatica, Vilnius, 4 (3-4)*, 360–383.

Raudys, S. (1998). Evolution and generalization of single neurone. I. Single-layer perceptron as seven statistical classifier. *Neural Networks, 11*, 283–296.

Raudys, S., & Diciunas V. (1996). Expected error of minimum empirical error and maximal margin classifiers. In *Proceedings 13th ICPR*, (Vienna, Austria, 25–29 August 1996) Vol. 2, Track B. Los Alamitos: IEEE Computer Society Press, pp. 875–879.

Raudys, S., & Duin R.P.W. (1998). On expected classification error of the Fisher classifier with pseudo-inverse covariance matrix. *Pattern Recognition Letters*, accepted.

Raudys, S., & Jain, A. K. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Trans. on Pattern Analysis and Machine Intelligence, PAMI-13*, 252–264.

Raudys, S., & Pikelis, V. (1980). On dimensionality, sample size, classification error and complexity of the classification algorithm in pattern recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence, PAMI-2 (3)*, 242–252.

Raudys, S., & Skurikhina M. (1994). Small sample properties of ridge estimate of the covariance matrix in statistical and neural net classification. In *New Trends in Probability and Statistics, Multivariate Statistics and Matrices in Statistics, Proc. of the 5-th Tartu Conference, Vol. 3*, Tartu-Puhajarve, Estonia, 23–28 May 1994, pp. 237–245.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. I*. Cambridge MA: Bradford Books, pp. 318–362.

Sjoberg, J., & Ljung, L. (1992). Overtraining, regularization, and searching for minimum in neural networks. Technical report, Department of Electrical Engineering, Linkoping University, S-581 83 Linkoping, Sweden, Feb. 1992.

Seung, H. S., Sompolinsky, H., & Tishby, N. (1992). Statistical mechanics from examples. *Physical Review, A45 (8)*, 6056–6091.

Sitgreaves, R., (1961). Some results on the distribution of the W-classification statistics. In *Studies in Item Selection and Prediction*. Stanford CA: Stanford University Press., pp. 241–261.

Solomon, H. (1956). Probability and statistics in psychometric research. In I.J. Neyman (Ed.), *Proc. of 3rd Berkley Symposium on Mathematical statistics and Probability*. Berkley, CA: University of California Press, pp. 169–184.

Takeshita, T., & Toriwaki, J. (1995). Experimental study of performance of pattern classifiers and the size of design samples. *Pattern Recognition Letters, 16*, 307–312.

Valiant, L. G. (1984). A theory of the learnable. *Comm. ACM, 27*, 1134–1142.

van Dam, J. W. M., Krose, B. J. A, & Froen, F. C. A. (1994). Optimal local Hebbian learning: use the d-rule. In M. Marinaro and P.G. Morasso (Eds.), *Artificial Neural Networks, Proc. ICANN'94*, Sorrento. Springer, pp. 631–634.

Vapnik, V. N., & Chervonenkis, D. Ya. (1968). Algorithms with full memory and recurrence algorithms in the problem of training pattern recognition. *Engineering Cybernetics (USSR J.), N4*, 95–106. in Russian.

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. *WESCON Convention Record, 4*, 96–104.

Wolpert, D. H. (Ed.) (1995a). The mathematics of generalization. In *Proceedings of the SFI/CNLS workshop on formal approach to supervised learning*, Nov. 1994. Addisson-Wesley.

Wolpert, D. H. (1995b). On relation between PAC, the statistical physics framework, the Bayesian framework, and the VC framework. In *Proceedings of the SFI/CNLS workshop on formal approach to supervised learning*, Nov. 1994. Addisson-Wesley.

Wyman, F., Young, D., & Turner, D. (1990). A comparison of asymptotic error rate expansions for the sample linear discriminant function. *Pattern Recognition, 23*, 775–783.

Zarudskij, V. I. (1979). The use of models of simple dependence problems of classification. In S. Raudys (Ed.), *Statistical Problems of Control, Issue 38*. Vilnius: Inst. of Math. and Cyb. Press, pp. 33–75 (in Russian).