INFORMATICA, 2001, Vol. 12, No. 2, 221–238

Generalization Error of Randomized Linear Zero Empirical Error Classifier: Non-Centered Data Case

Valdas DIČIŪNAS[*]

Department of Computer Science, Vilnius University
Naugarduko 24, 2600 Vilnius, Lithuania
e-mail: valdas.diciunas@maf.vu.lt

Received: December 2000

Abstract. One of the main problems in pattern classification and neural network training theory is the generalization performance of learning. This paper extends the results on randomized linear zero empirical error (RLZEE) classifier obtained by Raudys, Dičiūnas and Basalykas for the case of *centered* multivariate spherical normal classes. We derive an exact formula for an expected probability of misclassification (PMC) of RLZEE classifier in a case of *arbitrary* (*centered* or *non-centered*) spherical normal classes. This formula depends on two parameters characterizing the "degree of non-centering" of data. We discuss theoretically and illustrate graphically and numerically the influence of these parameters on the PMC of RLZEE classifier. In particular, we show that in some cases non-centered data has smaller expected PMC than centered data.

Key words: randomized linear classifier, generalization error, Gaussian classes, probability of misclassification, non-centered data.

1. Introduction

One of the main problems in pattern classification and neural network training theory is the generalization performance of learning, i.e., how well a classifier or network will perform in future (for unknown data) trained on a fixed-size training set (of known data). In some cases of simply distributed data (for example, for two normal classes) very accurate estimates of generalization error were obtained for many statistical classifiers (Wyman *et al.*, 1990). However, for neural networks and, in particular, for single layer perceptrons only very pessimistic bounds of the generalization error are known; see, e.g., (Amari and Murata, 1993; Dičiūnas and Raudys, 2000; Haussler *et al.*, 1994).

In this paper we consider a randomized linear zero empirical error (RLZEE) classifier. Since a single layer perceptron is also a linear classifier and its training can be considered as a random process (Raudys and Dičiūnas, 1994), an estimation of RLZEE classifier can be useful in neural network training theory, too (see Section 4).

[*] Part of this work was done while being at the Institute of Mathematics and Informatics, Vilnius, Lithuania.

An RLZEE classifier shortly can be defined as follows. A training set

$$L = \{\mathbf{X}_1^{(1)}, \mathbf{X}_2^{(1)}, \ldots, \mathbf{X}_N^{(1)},\ \mathbf{X}_1^{(2)}, \mathbf{X}_2^{(2)}, \ldots, \mathbf{X}_N^{(2)}\}$$

of p-dimensional ($p \geqslant 2$) observation vectors from two different classes π_1 and π_2 is given. Let $\mathbf{C}_1$ and $\mathbf{C}_2$ be centers (means) of classes π_1 and π_2, respectively. We call the data (i.e., classes π_1 and π_2) *centered* if $\mathbf{C}_1 + \mathbf{C}_2 = \mathbf{0}$ and *non-centered* otherwise. According to some a priori distribution we randomly generate coefficients $a_0, a_1, \ldots, a_p \in \mathbb{R}$ and verify if a hyperplane

$$a_1x_1 + \cdots + a_px_p + a_0 = 0$$

discriminates given vectors without errors, i.e., if it satisfies the following (ZEE) condition:

$$\begin{cases} \mathbf{A}^{\mathrm{T}}\mathbf{X}_j^1 + a_0 > 0 & \forall \mathbf{X}_j^1 \in L \cap \pi_1, \\ \mathbf{A}^{\mathrm{T}}\mathbf{X}_j^2 + a_0 \leqslant 0 & \forall \mathbf{X}_j^1 \in L \cap \pi_2, \end{cases} \tag{ZEE}$$

where $\mathbf{A}^{\mathrm{T}} = (a_1, \ldots, a_p)$. Suppose that we successfully found such a hyperplane $(a_0, \mathbf{A})$ (in opposite case we generate new coefficients $a_0, a_1, \ldots, a_p$ and so on). The question is how well this hyperplane discriminates *unknown* data from classes π_1 and π_2, i.e., we are interested in an expected *probability of misclassification* (PMC) of this classifier. In neural networks literature the expected PMC usually is called *generalization error* because it shows how well the given classifier generalizes. In this paper we will use both these terms in arbitrary way.

RLZEE classifier since 1993 has been intensively studied by Raudys *et al.*; see, for example, (Basalykas *et al.*, 1996; Dičiūnas and Raudys, 2000; Raudys, 1993, 1997; Raudys and Dičiūnas,1996). This classifier is worth for studying by following reasons:

(i) it allows to obtain an exact formula for an expected PMC while for the other investigated linear statistical classifiers only complicated asymptotic formulae are known, see (Basalykas *et al.*, 1996);

(ii) it works (i.e., gives a comparatively low expected PMC) in the cases when $p > N$ while Fisher's and some other classifiers do not work in such cases;

(iii) an expected PMC of RLZEE classifier can be successfully applied as an upper bound for a generalization error of single layer neural network classifiers.

The main results of above mentioned works are:

(1) an exact formula for an expected PMC in a case of centered multivariate spherical normal classes π_1, π_2 (Raudys, 1993);

(2) asymptotics of expected PMC for the same classes as in item (1) (Basalykas *et al.*, 1996); and

(3) simple "thumb" asymptotics of expected PMC constructed heuristically on the ground of the table of the values of exact formula (Raudys, 1997).

This paper generalizes above mentioned results for a non-centered data. In Section 2 we derive an exact formula for an expected PMC in a case of arbitrary (*centered* or *non-centered*) multivariate spherical normal classes π_1, π_2. This formula depends on two parameters m_1 and m_2 characterizing the "degree of non-centering" of data. In Section 3 we investigate the influence of these parameters on the PMC of RLZEE classifier and present the numerical simulations illustrating our theoretical results. Finally, in Section 4 we give some conclusions.

2. Mean Expected PMC of RLZEE Classifier

Let us consider RLZEE classifier defined in Section 1 in a simple case of two spherical multivariate normal classes π_1 and π_2. We will make the following assumptions:

- classes $\pi_1, \pi_2 \subset \mathbb{R}^p$ have equal prior probabilities $q_1 = q_2 = 1/2$ and densities $\mathrm{N}(\mathbf{X}, \mathbf{C}_1, I)$ and $\mathrm{N}(\mathbf{X}, \mathbf{C}_2, I)$, respectively; here $\mathbf{C}_1, \mathbf{C}_2$ are means (or centers) of classes and I is a $p \times p$ identity matrix;
- training set L contains the same number of training vectors from the each class: $N_1 = N_2 = N$ (consequently, $|L| = 2N$);
- the training vectors $\mathbf{X}_1^{(1)}, \mathbf{X}_2^{(1)}, \ldots, \mathbf{X}_N^{(1)}, \mathbf{X}_1^{(2)}, \mathbf{X}_2^{(2)}, \ldots, \mathbf{X}_N^{(2)}$ are independently and identically distributed in their own classes;
- the components of the vector $(\mathbf{A}^{\mathrm{T}}, a_0)$ a priori have normal distribution with zero mean and variance 1: $a_i \sim \mathrm{N}(0, 1)$, $i = 0, \ldots, p$.

We will find a mean expected probability of misclassification of pattern vectors from π_1 and π_2 which did not participate in the training. An expectation is taken both with respect to all possible random training sets of length $2N$ and with respect to the random character of generating $(p + 1)$-variate weight vector $(\mathbf{A}^{\mathrm{T}}, a_0)$. Our derivation of mean expected probability of misclassification MEP_N follows the derivation of analogical formula in *centered data* case presented in (Basalykas *et al.*, 1996). This method was proposed by Raudys (1993). By definition,

$$\mathrm{MEP}_N = \iint \mathrm{P}(\mathrm{MC} \mid \mathbf{A}, a_0) f_{\text{apost}}(\mathbf{A}, a_0 \mid \mathrm{ZEE})\, \mathrm{d}\mathbf{A}\, \mathrm{d}a_0, \tag{1}$$

where $\mathrm{P}(\mathrm{MC} \mid \mathbf{A}, a_0)$ is a conditional PMC given the vector of weights $(\mathbf{A}^{\mathrm{T}}, a_0)$:

$$\begin{aligned}\mathrm{P}(\mathrm{MC} \mid \mathbf{A}, a_0) = {} & \tfrac{1}{2}\mathrm{P}\{\mathbf{A}^{\mathrm{T}}\mathbf{X} + a_0 < 0 \mid \mathbf{X} \in \pi_1\} \\ & + \tfrac{1}{2}\mathrm{P}\{\mathbf{A}^{\mathrm{T}}\mathbf{X} + a_0 \geqslant 0 \mid \mathbf{X} \in \pi_2\}\end{aligned} \tag{2}$$

and $f_{\text{apost}}(\mathbf{A}, a_0 \mid \mathrm{ZEE})$ is a posterior probability density function of vector $(\mathbf{A}^{\mathrm{T}}, a_0)$ if the training was successful, i.e., condition (ZEE) was satisfied. According to Bayes' rule,

$$\begin{aligned}f_{\text{apost}}(\mathbf{A}, a_0 \mid \mathrm{ZEE}) &= \frac{\mathrm{P}(\mathrm{ZEE}, \mathbf{A}, a_0)}{\mathrm{P}(\mathrm{ZEE})} \\ &= \frac{\mathrm{P}(\mathrm{ZEE} \mid \mathbf{A}, a_0) f_{\text{aprior}}(\mathbf{A}, a_0)}{\iint \mathrm{P}(\mathrm{ZEE} \mid \mathbf{A}, a_0) f_{\text{aprior}}(\mathbf{A}, a_0)\, \mathrm{d}\mathbf{A}\, \mathrm{d}a_0},\end{aligned} \tag{3}$$

where the conditional probability of event (ZEE) is

$$\begin{aligned}\mathrm{P}(\mathrm{ZEE} \mid \mathbf{A}, a_0) \;=\; & \big[\mathrm{P}\big(\mathbf{A}^{\mathrm{T}}\mathbf{X} + a_0 \geqslant 0 \mid \mathbf{X} \in \pi_1\big)\big]^N \\ & \times\big[\mathrm{P}\big(\mathbf{A}^{\mathrm{T}}\mathbf{X} + a_0 < 0 \mid \mathbf{X} \in \pi_2\big)\big]^N. \end{aligned} \tag{4}$$

According to (Anderson, 1963, p. 39) if patterns belong to spherical normal classes the values of discrimination function for these patterns also have normal distribution:

$$\mathbf{X} \sim \mathrm{N}(\mathbf{C}_i, \mathrm{I}) \Rightarrow g(\mathbf{X}) = \mathbf{A}^{\mathrm{T}}\mathbf{X} + a_0 \sim \mathrm{N}\big(\mathbf{A}^{\mathrm{T}}\mathbf{C}_i + a_0, \mathbf{A}^{\mathrm{T}}\mathbf{A}\big), \quad i = 1, 2.$$

Then for $i = 1, 2$,

$$\mathrm{P}\big\{g(\mathbf{X}) < 0 \mid \mathbf{X} \in \pi_i\big\} = \Phi\bigg(-\frac{\mathbf{A}^{\mathrm{T}}\mathbf{C}_i + a_0}{\sqrt{\mathbf{A}^{\mathrm{T}}\mathbf{A}}}\bigg) \tag{5}$$

and

$$\begin{aligned}\mathrm{P}\big\{g(\mathbf{X}) > 0 \mid \mathbf{X} \in \pi_i\big\} \;&=\; 1 - \mathrm{P}\big\{g(\mathbf{X}) < 0 \mid \mathbf{X} \in \pi_i\big\} \\ &= 1 - \Phi\bigg(-\frac{\mathbf{A}^{\mathrm{T}}\mathbf{C}_i + a_0}{\sqrt{\mathbf{A}^{\mathrm{T}}\mathbf{A}}}\bigg) \\ &= \Phi\bigg(\frac{\mathbf{A}^{\mathrm{T}}\mathbf{C}_i + a_0}{\sqrt{\mathbf{A}^{\mathrm{T}}\mathbf{A}}}\bigg) \end{aligned} \tag{6}$$

since $1 - \Phi(-u) = \Phi(u)$, where Φ is normal distribution function

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \mathrm{e}^{-t^2/2}\,\mathrm{d}t.$$

Inserting (5) and (6) into (2) and (4) we obtain

$$\mathrm{P}(\mathrm{MC} \mid \mathbf{A}, a_0) = \frac{1}{2}\Phi\bigg(-\frac{\mathbf{A}^{\mathrm{T}}\mathbf{C}_1 + a_0}{\sqrt{\mathbf{A}^{\mathrm{T}}\mathbf{A}}}\bigg) + \frac{1}{2}\Phi\bigg(\frac{\mathbf{A}^{\mathrm{T}}\mathbf{C}_2 + a_0}{\sqrt{\mathbf{A}^{\mathrm{T}}\mathbf{A}}}\bigg) \tag{7}$$

and

$$\mathrm{P}(\mathrm{ZEE} \mid \mathbf{A}, a_0) \;=\; \bigg[\Phi\bigg(\frac{\mathbf{A}^{\mathrm{T}}\mathbf{C}_1 + a_0}{\sqrt{\mathbf{A}^{\mathrm{T}}\mathbf{A}}}\bigg)\bigg]^N \bigg[\Phi\bigg(-\frac{\mathbf{A}^{\mathrm{T}}\mathbf{C}_2 + a_0}{\sqrt{\mathbf{A}^{\mathrm{T}}\mathbf{A}}}\bigg)\bigg]^N. \tag{8}$$

Conditional probabilities (7) and (8) depend on $(p + 1)$-variate vector $(\mathbf{A}^{\mathrm{T}}, a_0)$. Following (Raudys, 1993) we will perform a special orthogonal transformation T reducing the number of variables. This transformation is a superposition of the following orthogonal transformations:

(i) first transformation R_1 is a rotation in the space $\mathbb{R}^p$ around the origin of coordinates which transfers the plane through the points $\mathbf{C}_1$, $\mathbf{C}_2$ and $\mathbf{0}$ into the plane $x_3 = \cdots = x_p = 0$;

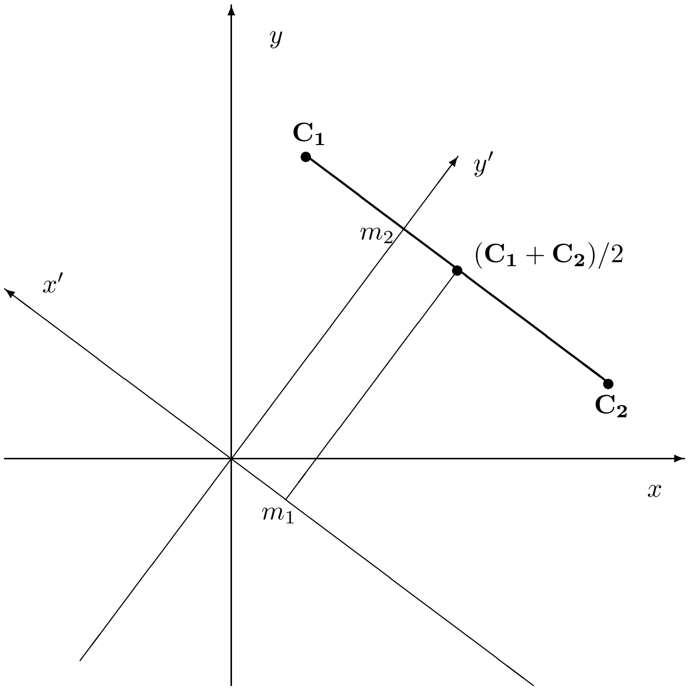

Fig. 1. Transformation T in $\mathbb{R}^2$.

(ii) second transformation, rotation R_2, transfers the vector $R_1(\mathbf{C}_1 - \mathbf{C}_2)/2$ lying on the plane $x_3 = \cdots = x_p = 0$ into the vector $(\delta/2, 0, \ldots, 0)^{\mathrm{T}}$ of the same plane, where $\delta = |\mathbf{C}_1 - \mathbf{C}_2|$ is a Euclidean distance between $\mathbf{C}_1$ and $\mathbf{C}_2$; if $R_2R_1((\mathbf{C}_1 + \mathbf{C}_2)/2) < 0$ we additionally make a reflection with respect to the line $\mathrm{O}x_1'$ (see below).

Let $T = R_2 \circ R_1$. Then T is an orthogonal $p \times p$ matrix satisfying $T^{\mathrm{T}}T = I$ and

$$T\left(\frac{\mathbf{C}_1 - \mathbf{C}_2}{2}\right) = \begin{bmatrix} \delta/2 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad T\left(\frac{\mathbf{C}_1 + \mathbf{C}_2}{2}\right) = \begin{bmatrix} m_1 \\ m_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

where $m_1, m_2 \in \mathbb{R}$ are the new coordinates of vector $(\mathbf{C}_1 + \mathbf{C}_2)/2$ after transformation T (see Fig. 1).

Two first rows $\mathbf{T}_1^{\mathrm{T}}$ and $\mathbf{T}_2^{\mathrm{T}}$ of T are the following:

$$\mathbf{T}_1 = \frac{\mathbf{C}_1 - \mathbf{C}_2}{\delta}, \qquad \mathbf{T}_2 = \pm\frac{\delta}{2}\,\frac{\mathbf{C}_1 + \mathbf{C}_2 - \frac{\mathbf{C}_1^2 - \mathbf{C}_2^2}{\delta^2}(\mathbf{C}_1 - \mathbf{C}_2)}{\sqrt{\mathbf{C}_1^2\mathbf{C}_2^2 - (\mathbf{C}_1\mathbf{C}_2)^2}},$$

where we choose the suitable sign of $\mathbf{T}_2$ to make $m_2 \geqslant 0$. Then after simple algebra we have

$$m_1 = \frac{\mathbf{C}_1^2 - \mathbf{C}_2^2}{2\delta} \quad \text{and} \quad m_2 = \frac{1}{\delta}\sqrt{\mathbf{C}_1^2\mathbf{C}_2^2 - (\mathbf{C}_1\mathbf{C}_2)^2}. \tag{9}$$

The parameters $m_1 \in \mathbb{R}$ and $m_2 \in \mathbb{R}^+$ characterize a "degree of non-centering" of data. Obviously,

$$m_1 = m_2 = 0 \quad \Longleftrightarrow \quad \mathbf{C}_1 = -\mathbf{C}_2,$$

i.e., data is centered.

Now we will derive the formula for MEP$_N$ for arbitrary m_1 and m_2. An influence of parameters m_1 and m_2 on generalization error will be investigated in Section 3.

Let us denote

$$V = TA = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_p \end{bmatrix}.$$

Then using the property $T^{\mathrm{T}}T = I$ we have

$$\begin{aligned}\frac{\mathbf{A}^{\mathrm{T}}\mathbf{C}_1 + a_0}{\sqrt{\mathbf{A}^{\mathrm{T}}\mathbf{A}}} &= \frac{\mathbf{A}^{\mathrm{T}}(\frac{\mathbf{C}_1-\mathbf{C}_2}{2} + \frac{\mathbf{C}_1+\mathbf{C}_2}{2}) + a_0}{\sqrt{\mathbf{A}^{\mathrm{T}}\mathbf{A}}} \\ &= \frac{(T\mathbf{A})^{\mathrm{T}}(T\frac{\mathbf{C}_1-\mathbf{C}_2}{2} + T\frac{\mathbf{C}_1+\mathbf{C}_2}{2}) + a_0}{\sqrt{\mathbf{A}^{\mathrm{T}}T^{\mathrm{T}}T\mathbf{A}}} \\ &= \frac{v_1\delta/2 + v_1m_1 + v_2m_2 + a_0}{\sqrt{\sum_{i=1}^p v_i^2}} = u\frac{\delta}{2} + w,\end{aligned} \tag{10}$$

and, analogously,

$$\frac{\mathbf{A}^{\mathrm{T}}\mathbf{C}_2 + a_0}{\sqrt{\mathbf{A}^{\mathrm{T}}\mathbf{A}}} = -u\frac{\delta}{2} + w, \tag{11}$$

where

$$u = \frac{v_1}{\sqrt{\sum_{i=1}^p v_i^2}} \quad \text{and} \quad w = \frac{v_1m_1 + v_2m_2 + a_0}{\sqrt{\sum_{i=1}^p v_i^2}}.$$

Now conditional probabilities (7) and (8) depend only on two scalar variables u and w. Let us find their distribution. We have to consider cases $p \geqslant 3$ and $p = 2$ separately.

2.1. *Distribution of u and w for $p \geqslant 3$*

Denoting $y = \sum_{i=3}^{p} v_i^2$ and adding two new variables s and t we have four random variables

$$\begin{aligned} u &= \frac{v_1}{\sqrt{v_1^2+v_2^2+y}}, \\ w &= \frac{v_1 m_1 + v_2 m_2 + a_0}{\sqrt{v_1^2+v_2^2+y}}, \\ s &= \frac{v_2}{\sqrt{v_1^2+v_2^2+y}}, \\ t &= y, \end{aligned}$$

depending on four random variables v_1, v_2, a_0 and y.

It is well known that a linear combination of normal variables

$$v_i = \sum_{j=1}^{p} t_{ij} a_j, \quad \text{where } a_j \sim \mathrm{N}(0,1),$$

also has normal distribution $\mathrm{N}(0, \sum_{j=1}^{p} t_{ij}^2) = \mathrm{N}(0,1)$; see, for example, (Kruopis, 1993). Then $y \sim \mathrm{C}(p-2)$, i.e., variable y has χ^2-distribution. Now we have the following densities:

$$\begin{aligned} f_1(v_1) &= \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-v_1^2/2}, \\ f_2(v_2) &= \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-v_2^2/2}, \\ f_3(a_0) &= \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-a_0^2/2}, \\ f_4(y) &= \frac{y^{(p-2)/2-1}\mathrm{e}^{-y/2}}{2^{(p-2)/2}\Gamma((p-2)/2)}, \quad y > 0. \end{aligned} \tag{12}$$

It is easy to obtain the inverse transformation of random variables:

$$\begin{aligned} v_1 &= \frac{u\sqrt{t}}{\sqrt{1-u^2-s^2}} = g_1(u,w,s,t), \\ v_2 &= \frac{s\sqrt{t}}{\sqrt{1-u^2-s^2}} = g_2(u,w,s,t), \\ a_0 &= (w - u m_1 - s m_2)\frac{\sqrt{t}}{\sqrt{1-u^2-s^2}} = g_3(u,w,s,t), \\ y &= t = g_4(u,w,s,t). \end{aligned}$$

According to the formula for the density of functions of random variables we have

$$f(u,w,s,t) = f_1(g_1) f_2(g_2) f_3(g_3) f_4(g_4) |J(u,w,s,t)|, \tag{13}$$

where J is the Jacobian of functions g_1, g_2, g_3, g_4. Denoting by $\ast$ the terms which will be multiplied by zero below, we have

$$J = \begin{vmatrix} \frac{\sqrt{t(1-u^2-s^2)}+\frac{u^2\sqrt{t}}{\sqrt{1-u^2-s^2}}}{1-u^2-s^2} & 0 & \frac{su\sqrt{t}}{(1-u^2-s^2)^{3/2}} & \ast \\ \frac{su\sqrt{t}}{(1-u^2-s^2)^{3/2}} & 0 & \frac{\sqrt{t(1-u^2-s^2)}+\frac{s^2\sqrt{t}}{\sqrt{1-u^2-s^2}}}{1-u^2-s^2} & \ast \\ \ast & \frac{\sqrt{t}}{\sqrt{1-u^2-s^2}} & \ast & \ast \\ 0 & 0 & 0 & 1 \end{vmatrix}$$

$$= -\frac{\sqrt{t}}{\sqrt{1-u^2-s^2}}\left[\frac{t(1-s^2)(1-u^2)}{(1-u^2-s^2)^3} - \frac{s^2u^2t}{(1-u^2-s^2)^3}\right]$$

$$= -\frac{\sqrt{t}}{\sqrt{1-u^2-s^2}}\frac{t}{(1-u^2-s^2)^2} = -\frac{t^{3/2}}{(1-u^2-s^2)^{5/2}}. \tag{14}$$

Inserting (12) and (14) into (13) we obtain

$$f(u,w,s,t) = c_1 \cdot \exp\left\{-\frac{t}{2}\frac{1+(w-um_1-sm_2)^2}{1-u^2-s^2}\right\} \cdot \frac{t^{(p-1)/2}}{(1-u^2-s^2)^{5/2}},$$

where

$$c_1 = \frac{1}{2^{(p+1)/2}\pi^{3/2}\Gamma((p-2)/2)}.$$

Now we are ready to calculate

$$f(u,w) = \int_{-\sqrt{1-u^2}}^{\sqrt{1-u^2}} \left(\int_0^\infty f(u,w,s,t)\,\mathrm{d}t\right)\mathrm{d}s. \tag{15}$$

Let

$$\alpha = \frac{1+(w-um_1-sm_2)^2}{2(1-u^2-s^2)}.$$

By integrating by parts is easy to show that (Prudnikov *et al.*, formula 2.3.3.1, p. 322)

$$\int_0^\infty \mathrm{e}^{-\alpha t}t^{(p-1)/2}\,\mathrm{d}t = \frac{\Gamma((p+1)/2)}{\alpha^{(p+1)/2}}. \tag{16}$$

Inserting (16) into (15) we finally obtain

$$f(u,w) = \frac{\Gamma((p+1)/2)}{\pi^{3/2}\Gamma((p-2)/2)}\int_{-\sqrt{1-u^2}}^{\sqrt{1-u^2}} \frac{(1-u^2-s^2)^{(p-4)/2}\,\mathrm{d}s}{[1+(w-um_1-sm_2)^2]^{(p+1)/2}}. \tag{17}$$

It is easy to verify that for the particular case of centered data ($m_1 = m_2 = 0$) the density $f(u,w)$ coincides with the density obtained by Basalykas *et al.* (1996).

2.2. *Distribution of u and w for $p = 2$*

We have

$$\begin{aligned} u &= \frac{v_1}{\sqrt{v_1^2 + v_2^2}}, \\ w &= \frac{v_1 m_1 + v_2 m_2 + a_0}{\sqrt{v_1^2 + v_2^2}}, \\ t &= v_2, \end{aligned}$$

where $v_1, v_2, a_0 \sim \mathrm{N}(0,1)$ (see Section 2.1). Inverse transformation is the following:

$$\begin{aligned} v_1 &= \frac{u|t|}{\sqrt{1-u^2}}, \\ v_2 &= t, \\ a_0 &= \frac{(w - um_1)|t|}{\sqrt{1-u^2}} - tm_2 \end{aligned}$$

with Jacobian

$$J = \begin{vmatrix} \dfrac{|t|\sqrt{1-u^2} + \frac{u^2|t|}{\sqrt{1-u^2}}}{1-u^2} & 0 & * \\ 0 & 0 & 1 \\ * & \dfrac{|t|}{\sqrt{1-u^2}} & * \end{vmatrix} = -\frac{t^2}{(1-u^2)^2}.$$

For $t > 0$, in the same way as in Section 2.1 we obtain

$$\begin{aligned} f_1(u,w,t) &= \frac{1}{(2\pi)^{3/2}} \exp\left\{ -\frac{u^2 t^2}{2(1-u^2)} - \frac{t^2}{2} - \frac{t^2 (w - um_1 - \sqrt{1-u^2} m_2)^2}{2(1-u^2)} \right\} \\ &\quad \times \frac{t^2}{(1-u^2)^2} \\ &= \frac{t^2}{(2\pi)^{3/2}(1-u^2)^2} \\ &\quad \times \exp\left\{ -\frac{t^2}{2(1-u^2)} \left(1 + \left(w - um_1 - \sqrt{1-u^2}\, m_2\right)^2\right) \right\}. \end{aligned}$$

Analogically, for $t < 0$,

$$\begin{aligned} &f_2(u,w,t) \\ &\quad = \frac{t^2}{(2\pi)^{3/2}(1-u^2)^2} \exp\left\{ -\frac{t^2}{2(1-u^2)} \left(1 + \left(w - um_1 + \sqrt{1-u^2}\, m_2\right)^2\right) \right\}. \end{aligned}$$

Then

$$\begin{aligned} f(u,w) &= \int_0^\infty f_1(u,w,t)\,\mathrm{d}t + \int_{-\infty}^0 f_2(u,w,t)\,\mathrm{d}t \\ &= \frac{1}{(2\pi^{3/2}(1-u^2)^2}\biggl(\int_0^\infty t^2\mathrm{e}^{-\alpha t^2/2}\,\mathrm{d}t + \int_0^\infty t^2\mathrm{e}^{-\beta t^2/2}\,\mathrm{d}t\biggr), \end{aligned}$$

where

$$\begin{aligned} \alpha &= \frac{1+(w-um_1-\sqrt{1-u^2}m_2)^2}{1-u^2}, \\ \beta &= \frac{1+(w-um_1+\sqrt{1-u^2}m_2)^2}{1-u^2}. \end{aligned}$$

It is easy to show that

$$\begin{aligned} \int_0^\infty t^2\mathrm{e}^{-\alpha t^2/2}\,\mathrm{d}t &= \frac{1}{\alpha^{3/2}}\int_0^\infty z^2\mathrm{e}^{-z^2/2}\,\mathrm{d}z = -\frac{1}{\alpha^{3/2}}\int_0^\infty z\,\mathrm{d}\mathrm{e}^{-z^2/2} \\ &= -\frac{1}{\alpha^{3/2}}\biggl(z\mathrm{e}^{-z^2/2}\Big|_{z=0}^{\infty} - \int_0^\infty \mathrm{e}^{-z^2/2}\,\mathrm{d}z\biggr) = \frac{1}{\alpha^{3/2}}\sqrt{\frac{\pi}{2}}. \end{aligned}$$

We obtain

$$\begin{aligned} f(u,w) &= \frac{1}{4\pi\sqrt{1-u^2}}\biggl\{\frac{1}{[1+(w-um_1-\sqrt{1-u^2}\,m_2)^2]^{3/2}} \\ &\qquad + \frac{1}{[1+(w-um_1+\sqrt{1-u^2}\,m_2)^2]^{3/2}}\biggr\}. \end{aligned} \tag{18}$$

For $m_1 = m_2 = 0$ this density also coincides with the density given by Basalykas *et al.* (1996).

Inserting (3), (7), (8), (10) and (11) into (1) we finally obtain the following exact formula for an expected PMC of RLZEE classifier for arbitrary (centered or non-centered) spherical normal classes:

$$\mathrm{MEP}_N = \frac{\int_{-\infty}^\infty\int_{-1}^1 \mathrm{P}(\mathrm{MC}\mid u,w)\mathrm{P}(\mathrm{ZEE}\mid u,w)f(u,w)\,\mathrm{d}u\,\mathrm{d}w}{\int_{-\infty}^\infty\int_{-1}^1 \mathrm{P}(\mathrm{ZEE}\mid u,w)f(u,w)\,\mathrm{d}u\,\mathrm{d}w} \tag{19}$$

with

$$\begin{aligned} \mathrm{P}(\mathrm{MC}\mid u,w) &= \frac{1}{2}\biggl[\Phi\Bigl(-\frac{u\delta}{2}+w\Bigr) + \Phi\Bigl(-\frac{u\delta}{2}-w\Bigr)\biggr], \\ \mathrm{P}(\mathrm{ZEE}\mid u,w) &= \biggl[\Phi\Bigl(\frac{u\delta}{2}+w\Bigr)\biggr]^N\biggl[\Phi\Bigl(\frac{u\delta}{2}-w\Bigr)\biggr]^N, \end{aligned}$$

and $f(u,v)$ given by (17) ($p \geqslant 3$) or (18) ($p = 2$).

3. On Influence of Parameters m_1 and m_2 on the Generalization Error

In this section we discuss the dependence of generalization error MEP$_N$ on parameters m_1 and m_2. Let us consider only the case $p = 2$, for simplicity. However, most of the arguments presented below are also valid in general case (see the end of the section).

The analytical investigation is very complicated. Let

$$E(m_1, m_2) = \text{MEP}_N(p, N, \delta, m_1, m_2),$$

where p, N, δ are fixed, $m_1 \in \mathbb{R}$ and $m_2 \in \mathbb{R}^+$. Since $E(m_1, m_2) = E(-m_1, m_2)$ below we consider only the case $m_1 \geqslant 0$ (i.e., $|\mathbf{C}_1| \geqslant |\mathbf{C}_2|$, see (9)).

We only succeeded to show that $\partial E/\partial m_1, \partial E/\partial m_2 \to 0$ for fixed m_1 and $m_2 \to \infty$ and for fixed m_2 and $m_1 \to \infty$. Indeed, let us denote

$$g(u, w) \coloneqq \text{P(MC} \mid u, w)\text{P(ZEE} \mid u, w) \quad \text{and} \quad h(u, w) \coloneqq \text{P(ZEE} \mid u, w).$$

Let also f be given by (18) and let us denote

$$f_1(u, w) \coloneqq \frac{\partial f(u, w)}{\partial m_1} = \frac{3u}{4\pi\sqrt{1-u^2}}\Bigg(\frac{w - um_1 - \sqrt{1-u^2}\, m_2}{[1 + (w - um_1 - \sqrt{1-u^2}\, m_2)^2]^{5/2}} + \frac{w - um_1 + \sqrt{1-u^2}\, m_2}{[1 + (w - um_1 + \sqrt{1-u^2}\, m_2)^2]^{5/2}}\Bigg).$$

and

$$f_2(u, w) \coloneqq \frac{\partial f(u, w)}{\partial m_2} = \frac{3}{4\pi}\Bigg(\frac{w - um_1 - \sqrt{1-u^2}\, m_2}{[1 + (w - um_1 - \sqrt{1-u^2}\, m_2)^2]^{5/2}} - \frac{w - um_1 + \sqrt{1-u^2}\, m_2}{[1 + (w - um_1 + \sqrt{1-u^2}\, m_2)^2]^{5/2}}\Bigg).$$

Then we have

$$\begin{aligned}\frac{\partial \text{MEP}_N}{\partial m_i} = {} & \Bigg[\int_{-\infty}^{\infty}\int_{-1}^{1} g(u, w) f_i(u, w)\, \mathrm{d}u\, \mathrm{d}w \cdot \int_{-\infty}^{\infty}\int_{-1}^{1} h(u, w) f(u, w)\, \mathrm{d}u\, \mathrm{d}w \\ & - \int_{-\infty}^{\infty}\int_{-1}^{1} g(u, w) f(u, w)\, \mathrm{d}u\, \mathrm{d}w \cdot \int_{-\infty}^{\infty}\int_{-1}^{1} h(u, w) f_i(u, w)\, \mathrm{d}u\, \mathrm{d}w\Bigg] \\ & \times \frac{1}{(\int_{-\infty}^{\infty}\int_{-1}^{1} h(u, w) f(u, w)\, \mathrm{d}u\, \mathrm{d}w)^2}, \quad i = 1, 2.\end{aligned}$$

For $m_1 = \text{const}$, $m_2 \to \infty$ we have

$$\lim_{m_2\to\infty} \frac{\partial \text{MEP}_N}{\partial m_i}(m_1, m_2) = \lim_{m_2\to\infty} \frac{(1/m_2^4)\cdot(1/m_2^3)}{(1/m_2^3)^2} = \lim_{m_2\to\infty} \frac{1}{m_2} = 0.$$

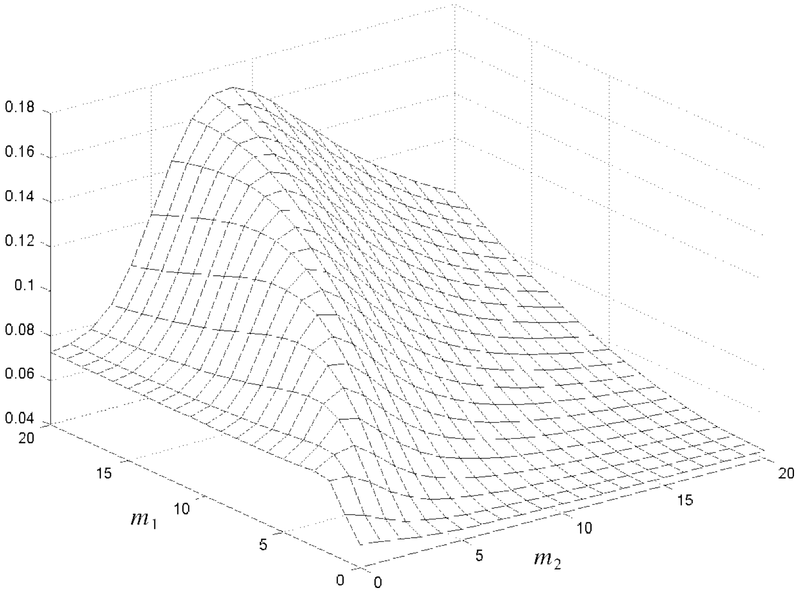

Fig. 2. Surface of the generalization error (19). $p = 2$, $N = 10$, $\delta = 4$, $m_1, m_2 \in [0, 20]$.

Table 1
Generalization error (19) values for $p = 2$, $N = 10$, $\delta = 4$, $m_1, m_2 \in [0, 20]$.

$m_1 \setminus m_2$	0	2	4	6	8	10	20
0	0.0503	0.0478	0.0457	0.0447	0.0442	0.0439	0.0434
2	0.0726	0.0671	0.0554	0.0499	0.0473	0.0459	0.0439
4	0.0731	0.0889	0.0788	0.0645	0.0565	0.0521	0.0455
6	0.0725	0.0930	0.1029	0.0852	0.0709	0.0622	0.0482
8	0.0724	0.0900	0.1181	0.1073	0.0889	0.0755	0.0521
10	0.0724	0.0863	0.1230	0.1263	0.1081	0.0912	0.0570
20	0.0726	0.0781	0.1017	0.1439	0.1689	0.1653	0.0953

Similarly,

$$\lim_{m_1 \to \infty} \frac{\partial \mathrm{MEP}_N}{\partial m_i}(m_1, m_2) = 0.$$

The results of numerical simulations are presented in Fig. 2. The figure shows the surface of the function $E(m_1, m_2)$ for the case $p = 2$, $N = 10$, $\delta = 4$ and $m_1, m_2 \in [0, 20]$. The exact values of $E(m_1, m_2)$ for this particular case are given in Table 1. The numerical simulations show that $E(m_1, m_2)$ does not have finite local minima or maxima points. We conclude that $E(m_1, m_2)$ tends to its minimum value at the infinite point $m_1 = 0$, $m_2 = \infty$ and $E(m_1, m_2)$ tends to its maximum value along some curve $h(m_1, m_2)$

which depends on p, N and δ. In our particular case (see above) this curve approximately can be presented as $m_2 = m_1^{0.75}$.

A bit surprising result we obtained is that *the centered data case is not the best one* (for the normal prior distribution of the coefficients of random hyperplanes). However, this can be easily explained from geometrical point of view. Let us find the distribution of random hyperplanes in $\mathbb{R}^2$ (i.e., lines) with normal coefficients.

The line $a_1x_1 + a_2x_2 + a_0 = 0$ can be represented by its normal equation $x_1 \cos\alpha + x_2 \sin\alpha - n = 0$ with parameters n and α, where

$$n = |\vec{n}| = \frac{|a_0|}{\sqrt{a_1^2 + a_2^2}}$$

is the distance between the line and the origin of coordinate system, and

$$\alpha = \text{arcctg}\left(\frac{a_1}{a_2}\right)$$

is the angle between the line's normal vector $\vec{n}$ and the axis Ox_1. Here $\alpha \in [0, \pi]$. Strictly speaking, for $a_0/a_2 > 0$ (i.e., for a_0 and a_2 of the same sign) we have $\alpha = \text{arcctg}(a_1/a_2) + \pi$ but this does not change the matter.

Let us find distributions of parameters n and α for normally distributed a_1, a_2, a_0.

Distribution of the distance n. Let $a_1, a_2, a_0 \sim \text{N}(0, 1)$ with the density

$$f(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}.$$

Then $n = |x|/\sqrt{2}$, where random variable

$$x = \frac{a_0}{\sqrt{(a_1^2 + a_2^2)/2}} \sim \text{S}(2)$$

has Student's distribution (Kruopis, 1993, p. 75) with the density

$$f(x) = \frac{\Gamma(3/2)}{\sqrt{2\pi}\Gamma(1)}\left(1 + \frac{x^2}{2}\right)^{-3/2}.$$

We obtain (Kruopis, 1993, p. 29)

$$\begin{aligned} g(n) &= \sqrt{2}\big(f\big(\sqrt{2}n\big) + f\big(-\sqrt{2}n\big)\big) = 2\sqrt{2} \cdot \frac{\frac{1}{2}\sqrt{\pi}}{\sqrt{2\pi} \cdot 1}\big(1 + n^2\big)^{-3/2} \\ &= \frac{1}{(1 + n^2)^{3/2}}. \end{aligned} \tag{20}$$

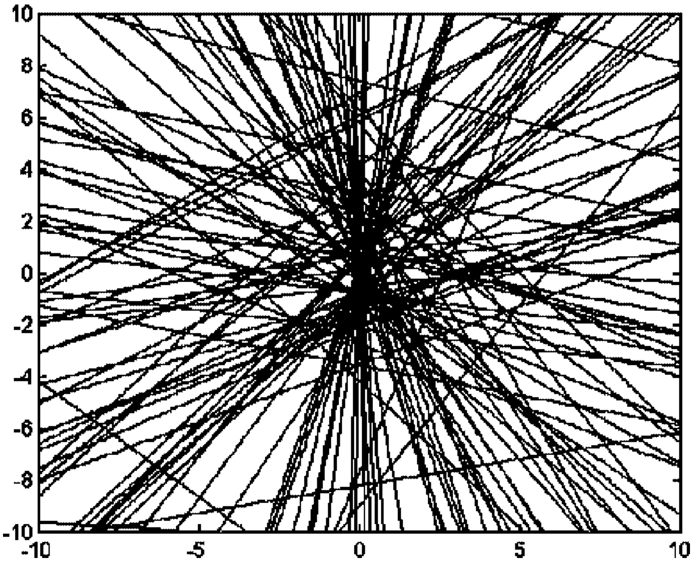

Fig. 3. Distribution of 100 lines with random normal weights in $\mathbb{R}^2$.

Distribution of the angle α. Let $x, y \sim \mathrm{N}(0,1)$ and $z = x/y$. Then (Kruopis, 1993, p. 29)

$$\begin{aligned} g(z) &= \int_{-\infty}^{\infty} |y| \cdot \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2y^2/2} \cdot \mathrm{e}^{-y^2/2}\,\mathrm{d}y \\ &= \frac{1}{\pi} \int_0^{\infty} y\mathrm{e}^{-y^2(z^2+1)/2}\,\mathrm{d}y \\ &= \frac{1}{\pi(z^2+1)} \int_0^{\infty} \mathrm{e}^{-y^2(z^2+1)/2}\,\mathrm{d}\Big(\frac{y^2(z^2+1)}{2}\Big) \\ &= -\frac{1}{\pi(z^2+1)} \mathrm{e}^{-t}\Big|_{t=0}^{\infty} = \frac{1}{\pi(z^2+1)}. \end{aligned}$$

Let $\alpha = \operatorname{arcctg} z$, then $z = h^{-1}(\alpha) = \operatorname{ctg}\alpha$; therefore (Kruopis, 1993, p. 27)

$$\begin{aligned} f(\alpha) &= g(\operatorname{ctg}\alpha)\frac{1}{\sin^2\alpha} = \frac{1}{\pi(\frac{\cos^2\alpha}{\sin^2\alpha}+1)} \cdot \frac{1}{\sin^2\alpha} = \frac{\sin^2\alpha}{\pi} \cdot \frac{1}{\sin^2\alpha} \\ &= \frac{1}{\pi}, \quad \alpha \in (0,\pi). \end{aligned} \tag{21}$$

Equations (20) and (21) show that the number of random lines decreases moving away from the origin of coordinates and the angle between the line and any axis of coordinates is distributed uniformly. In Fig. 3 we see 100 random lines $a_1x_1 + a_2x_2 + a_0 = 0$ with $a_1, a_2, a_0 \sim \mathrm{N}(0,1)$. The generalization error becomes smaller when the data is far from the origin and $|\mathbf{C}_1| = |\mathbf{C}_2|$, i.e., $m_1 = 0$ and $m_2 \to \infty$ (see Fig. 5) because in

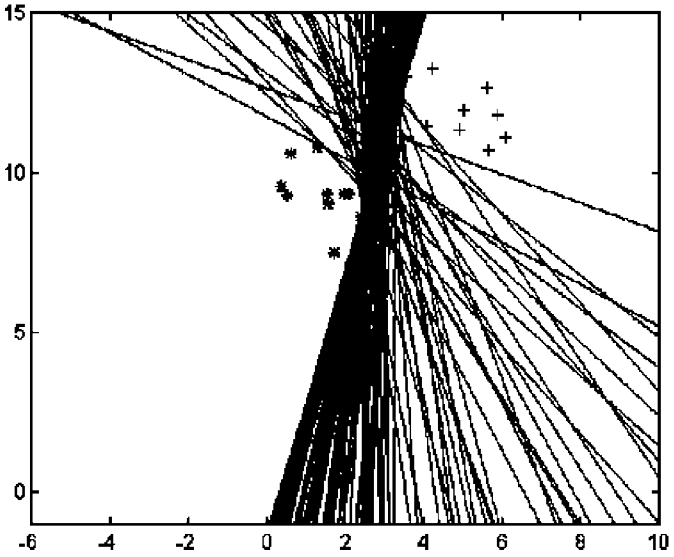

Fig. 4. Distribution of the lines with zero empirical error in "very bad" case.

such situation we have much more lines with zero empirical error and small test error and significantly less lines with zero empirical error and big test error. The following example gives more detailed explanation of the behaviour of MEP$_N$.

EXAMPLE 1. In our numerical simulation we used the same data located, however, in different places of the plane $\mathbb{R}^2$ ($p = 2$). Let classes $\pi_i \sim \mathrm{N}(\mathbf{C}_i, I)$, where $|\mathbf{C}_1 - \mathbf{C}_2| = \delta = 3\sqrt{2}$. Using $N = 10$ random training vectors from the each class π_i, we choose one by one random lines with normal weights and verify if these lines separate training vectors without errors. For any random line with zero empirical error we calculate its test error. Finally, we obtain an experimental generalization error MEP$_{\text{exp}}$ as the mean of these test errors. Instead of calculating the test error by means of the test set, we used its theoretical value (7).

We considered 3 cases: "very bad", "very good" and "good". Case "good" was the case of the centered data. The following results were obtained:

Case 1, very bad: $\mathbf{C}_1 = (5, 12), \mathbf{C}_2 = (2, 9), m_1 = 7\sqrt{2}, m_2 = 3.5\sqrt{2}\,(\sqrt{m_1^2 + m_2^2} = |(\mathbf{C}_1 + \mathbf{C}_2)/2| \approx 11.068)$ (see Fig. 4). Theoretical generalization error MEP$_N = 0.1207$, experimental generalization error MEP$_{\text{exp}} = 0.1242$ (the mean is taken for 134 lines with zero empirical error of total number 50000 random lines).

Case 2, very good: $\mathbf{C}_1 = (1.5\sqrt{2}, 11.068)$, $\mathbf{C}_2 = (-1.5\sqrt{2}, 11.068)$, $m_1 = 0$, $m_2 = 11.068$ (see Fig. 5). Theoretical generalization error MEP$_N = 0.0376$, experimental generalization error MEP$_{\text{exp}} = 0.0373$ (the mean is taken for 51 lines with zero empirical error of total number 2000 random lines).

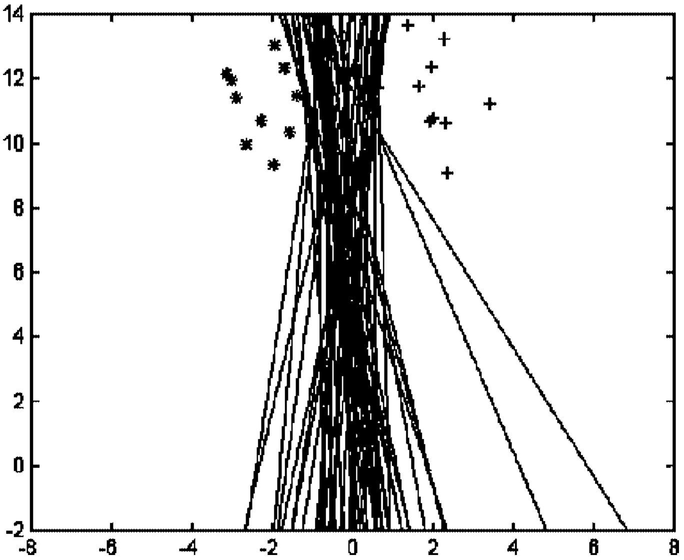

Fig. 5. Distribution of the lines with zero empirical error in “very good” case.

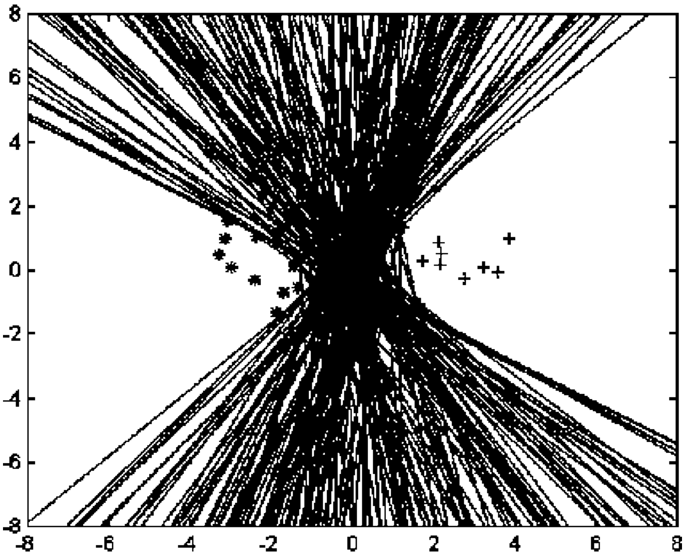

Fig. 6. Distribution of the lines with zero empirical error in “good” case.

Case 3, good: $\mathbf{C}_1 = (1.5\sqrt{2}, 0)$, $\mathbf{C}_2 = (-1.5\sqrt{2}, 0)$, $m_1 = m_2 = 0$ (see Fig. 6). Theoretical generalization error $\text{MEP}_N = 0.0431$, experimental generalization error $\text{MEP}_{\text{exp}} = 0.0458$ (the mean is taken for 173 lines with zero empirical error of total number 1000 random lines).

Figures 4–6 show that in "very bad" case there are much more lines with zero empirical error and big test error and significantly less lines with zero empirical error and small test error while in "very good" case almost all lines with zero empirical error have small test error; in centered data ("good") case we have many lines with a big test error as well as with a small test error.

The numerical simulations and intuition show that similar situation is characteristic also for general case $p \geqslant 3$. Namely, in space $\mathbb{R}^p$ ($p \geqslant 3$) most part of randomly generated $(p-1)$-dimensional hyperplanes with normal prior distribution of there coefficients will lay "near" the origin of the coordinates. Consequently, for some non-centered normally distributed data RLZEE classifier will have smaller generalization error than for the same but centered data.

4. Summary and Conclusions

In this paper we derived an exact formula for an expected PMC in a case of arbitrary multivariate spherical normal classes. We also found the distribution in $\mathbb{R}^2$ of random lines with normally distributed their coefficients. Our analytical investigation and numerical simulations show that the centered data case is not the best one for the RLZEE classifier with normal prior distribution of the coefficients of random hyperplanes.

Our investigation confirms the well-known result that single layer perceptron (SLP) learns better in centered data case due to a random procedure of generating its initial weights. Indeed, according to results of Section 3, the initial separating hyperplane of SLP with higher probability has smaller empirical error for centered data case than for non-centered one (because this hyperplane with higher probability lays near the origin of the coordinates as well as centered data does). Consequently, in centered data case the initialization is better and the training ends sooner.

On the other hand, if we want to use the generalization error of RLZEE classifier as an upper bound of generalization error of the well-trained (i.e., with zero empirical error) SLP classifier we must consider centered data because for the non-centered data RLZEE classifier in some cases can outperform the SLP classifier.

Acknowledgements

The author acknowledges Professor Šarūnas Raudys for a great number of valuable comments and permanent stimulation of this work.

References

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

Anderson, T. (1963). *An Introduction to Multivariate Statistical Analysis*. Fizmatgiz, Moscow (in Russian).

Basalykas, A., V. Dičiūnas, Š. Raudys (1996). On expected probability of misclassification of linear zero empirical error classifier. *Informatica*, **7**(2), 137–154; (1997). Letter to Editors. *Informatica*, **8**(2), 310–311.

Dičiūnas, V., Š. Raudys (2000). Generalization error of randomized linear zero empirical error classifier: Simple asymptotics for centered data case. *Informatica*, **11**(4), 381–396.

Haussler, D., M. Kearns, H.S. Seung, N. Tishby (1994). Rigorous learning curves from statistical mechanics. In *Proceedings of the 7th Annual ACM Workshop on Computational Learning Theory*, Morgan Kaufmann, San Mateo, CA, pp. 76–87.

Kruopis, J. (1993). *Mathematical Statistics*. Publishing House of Science and Encyclopedias, Vilnius, 2nd edn (in Lithuanian).

Prudnikov, A.P., Yu.A. Brychkov, O.I. Marichev (1981). *Integrals and Series*. Nauka, Moscow (in Russian).

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

Raudys, Š. (1997). On dimensionality, sample size, and classification error of nonparametric linear classification algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**(6), 667–671.

Raudys Š., V. Dičiūnas (1994). Generalization error of linear margin classifier. In *Proc. of 6th Microcomputer School on Neural Networks and Applications*, Sedmihorky, Czech Republic, pp. 159–164.

Raudys Š., V. Dičiūnas (1996). Expected error of minimum empirical error and maximal margin classifiers. In *Proc. of 13th Internat. Conf. on Pattern Recognition*, Vienna, Austria, August 25–29, Vol. 2, IEEE Computer Society Press, Los Alamitos, CA, pp. 875–879.

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

V. Dičiūnas received a mathematician's diploma from Moscow State University in 1982. He is a Senior Assistant Professor at the Department of Computer Science of Vilnius University. His research interests include artificial neural networks, statistical classification methods, and complexity theory.

Atsitiktinio nulinės empirinės klaidos tiesinio klasifikatoriaus vidutinė tikėtina klaida: necentruotų duomenų atvejis

Valdas DIČIŪNAS

Šis straipsnis praplečia Raudžio ir kitų autorių rezultatus, gautus nagrinėjant atsitiktinį nulinės empirinės klaidos tiesinį (ANEKT) klasifikatorių centruotų duomenų atveju. Straipsnyje išvesta tiksli ANEKT klasifikatoriaus vidutinės tikėtinos klaidos formulė, tinkanti tiek centruotiems, tiek ir necentruotiems duomenims. Ši formulė priklauso nuo dviejų parametrų, charakterizuojančių klasifikuojamų duomenų "necentruotumo laipsnį". Pateikiama teorinė ir skaitinė šių parametrų įtakos vidutinei tikėtinai klaidai analizė. Tame tarpe buvo parodyta, kad tam tikro pavidalo necentruotiems duomenims gaunama mažesnė klaida negu centruotiems duomenims. Šiais atvejais ANEKT klasifikatoriaus vidutinė tikėtina klaida gali būti mažesnė už be klaidų apmokyto netiesinio perceptrono vidutinę tikėtiną klaidą.