OPTIMAL ESTIMATION BY USING NEURAL NETWORKS

O.A. Stepanov* and O.S. Amosov**

**State Research Center of Russia - Central Scientific & Research Institute Elektropribor*
30, Malaya Posadskaya Str., St. Petersburg, 197046, Russia
Tel. (812) 232-82-53, fax (812) 232-33-76, e-mail: elprib@online.ru
***Komsomolsk-on-Amur State Technical University*
27, Lenina Str., Komsomolsk-on-Amur, 681013, Russia
Tel. (4217) 55-41-51, fax (4217) 53-61-50, e-mail: aos@kmscom.ru

Abstract: The relation between the traditional (minimum variance) algorithms for estimation of random vectors and the algorithms based on the use of neural networks has been investigated. It is shown that the Bayesian and neural network algorithms provide estimates with similar properties. The results derived are discussed. The examples (in particular problem with a non-Gaussian a posterior probability density function) are considered.

Keywords: minimum variance estimation, non-Gaussian a posteriori probability density function, neural network, comparison.

1. INTRODUCTION

Neural networks (NN) have been extensively investigated in the context of adaptive control and system identification. But only recently, as it is noted in (Parlos, *et al*., 2001), they came to be used for filtering problem. Optimal filtering is known to be widely used in estimation of random processes and sequences (Kalman, 1960, Meditch, 1969; Jazwinski, 1970; Gelb A., 1974; Yarlykov, and Mironov, 1999; Stepanov, 1998). However constructing algorithms requires comprehensive a priori information about the processes estimated and their measurement errors. Besides, serious difficulties emerge in constructing nonlinear filtering algorithms (Jazwinski, 1970; Yarlykov and Mironov, 1999; Stepanov, 1998; Dmitriev and Stepanov, 1998). These disadvantages make the researchers look for new approaches to the construction of algorithms. One of such approaches can be based on neural networks their advantages being training capabilities and the possibility to be applied to the solution of difficult (from the calculation standpoint) problems (Haykin, 1994). Among the publications devoted to the application of NN for the estimation problem, of particular interest for us are the following:

- References (Parisini, *et al*., 1994; Alessandri, *et al*., 1999) reduces the estimation problem to one of nonlinear programming problems and NN is used for the solution of this problem;
- Reference (Haykin and Yee, 1997) discusses the application of NN to the nonlinear filtering problem, in particular, the application of the radial basis functions for NN estimation;
- Reference (Parlos, *et al*., 2001) deals with the application of the recurrent NN for adaptive filtering problems.

However, in our opinion, there is no unambiguous answer about the advantages or disadvantages of the neural approach in comparison with the traditional one. Most attention in the papers has been concentrated on the methods of applying neural networks to filtering and estimation. The publications that do compare the NN and optimal filtering approaches concern, as a rule, some particular examples and are based on simulation. Theory is most substantially dealt with in the publication of Lo (1994), in which it is shown that the estimate generated by the recurrent neural networks considered there tends, under certain conditions, to the minimum variance estimate. However the author does not discuss the relation between the traditional and neural network algorithms. In our opinion, this makes it difficult to use widely an NNs for the solution of applied problems.

There is another line of investigation that also makes use of both the filtering theory methods and the neural network approach. It has emerged in connection with the training problem which is of

great importance and is treated as a nonlinear estimation problem and solved by using various modifications of nonlinear filtering algorithms (Puskorius 1996, Simandl *et al*., 2004). It is this line of investigation that the recent book by Haykin is devoted to (Haykin, 2001). Unfortunately, it does not consider the possibilities of NN themselves to solve estimation and filtering problems, nor is the relation between the traditional and neural network algorithms discussed. For the particular problem of linear estimation such relation is investigated in (Stepanov and Amosov, 2004). It is shown that for linear neural networks and the appropriate choice of the criterion used for its off-line training, the traditional and neural network algorithms are practically identical and they provide estimates with similar properties. The present paper is devoted to a more general nonlinear, non-Gaussian case, for which the problem of linear estimation is particular case.

2. TRADITIONAL BAYESIAN ESTIMATION

It is not uncommon that the applied estimation problems (in particular, the problems connected with navigation) can be reduced to a rather simple problem of estimating n-dimensional vector $\mathbf{x} = [x_1 \quad ... \quad x_n]^T$ by m-dimensional measurements $\mathbf{y} = [y_1 \quad ... \quad y_m]^T$, which can be written as follows (Stepanov, 1998):

$$\mathbf{y} = \mathbf{s}(\mathbf{x}) + \mathbf{v}, \tag{1}$$

where $\mathbf{s}(\mathbf{x}) = [s_1(\mathbf{x}) \quad ... \quad s_m(\mathbf{x})]^T$ is an m - dimensional, in a general case, nonlinear vector-function, which is usually assumed to be known; and $\mathbf{v} = [v_1 \quad ... \quad v_m]^T$ is a random vector of measurement errors.

Suppose that the joint probability density function (p.d.f.) $f(\mathbf{x}, \mathbf{v})$ is known. This allows deriving the joint $f(\mathbf{x}, \mathbf{y})$ for the vectors $\mathbf{x}$ and $\mathbf{y}$. In other words, suppose that $f(\mathbf{x}, \mathbf{v})$ or $f(\mathbf{x}, \mathbf{y})$ is the a priori information. In this case the estimation problem can be formulated in the framework of the Bayesian approach as follows: using the vector $\mathbf{y}$, find the estimate $\hat{\mathbf{x}}(\mathbf{y})$ that minimizes the criterion:

$$J = E[(\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y}))^T (\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y}))] = E\left\|(\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y}))\right\|^2$$
$$= \int\int \left\|(\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y}))\right\|^2 f(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}, \tag{2}$$

where E is the mathematical expectation corresponding to the p.d.f. $f(\mathbf{x}, \mathbf{y})$. In this discussion integrals are regarded as multiple integrals with infinite limits. The estimate that minimizes this criterion is called a minimum variance (optimal) estimate. It is well known that this estimate is determined as in (Jazwinski, 1970; Yarlykov and Mironov, 1999)

$$\hat{\mathbf{x}}(\mathbf{y}) = \int \mathbf{x} f(\mathbf{x} / \mathbf{y}) d\mathbf{x}, \tag{3}$$

where $f(\mathbf{x} / \mathbf{y})$ is the conditional (a posteriori) p.d.f. for the vector $\mathbf{x}$, for which the following relation holds true:

$$f(\mathbf{x} / \mathbf{y}) = \frac{f(\mathbf{x}, \mathbf{y})}{f(\mathbf{y})}, \tag{4}$$

where $f(\mathbf{y}) = \int f(\mathbf{x}, \mathbf{y}) d\mathbf{x}$.

It is known that if the p.d.f. $f(\mathbf{x}, \mathbf{y})$ is Gaussian, then the a posterior p.d.f. $f(\mathbf{x} / \mathbf{y})$ is Gaussian too.

In this case the estimate $\hat{\mathbf{x}}(\mathbf{y})$, which minimizes (2), and the covariance matrix of the estimation errors $\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}}(\mathbf{y})$ are determined as in (Meditch, 1969):

$$\hat{\mathbf{x}}(\mathbf{y}) = \overline{\mathbf{x}} + \mathbf{P}_{\mathbf{xy}} \mathbf{P}_{\mathbf{yy}}^{-1} [\mathbf{y} - \overline{\mathbf{y}}], \tag{5}$$

$$\mathbf{P}_{\mathbf{e}} = E[(\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y}))(\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y}))^{\mathrm{T}}] = \mathbf{P}_{\mathbf{xx}} - \mathbf{P}_{\mathbf{xy}} \mathbf{P}_{\mathbf{yy}}^{-1} \mathbf{P}_{\mathbf{yx}}, \tag{6}$$

where $\overline{\mathbf{x}}$, $\overline{\mathbf{y}}$, $\mathbf{P}_{\mathbf{xx}}, \mathbf{P}_{\mathbf{yy}}, \mathbf{P}_{\mathbf{xy}}$ are the mathematical expectations and covariance matrices for $\mathbf{x}$ and $\mathbf{y}$.

For example, the p.d.f. $f(\mathbf{x}, \mathbf{y})$ is Gaussian, when the joint p.d.f. of the vector $\mathbf{x}$ and measurement errors $\mathbf{v}$ is Gaussian and the relation between them is linear, i.e.,

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{v}, \tag{7}$$

where $\mathbf{H}$ is the known $m \times n$ matrix. In this case, if the Gaussian vectors $\mathbf{x}$ and $\mathbf{v}$ are statistically independent of each other, then

$$\mathbf{P}_{\mathbf{xy}} = \mathbf{P}_{\mathbf{xx}} \mathbf{H}^T, \tag{8}$$

$$\mathbf{P}_{\mathbf{yy}} = \mathbf{H}\mathbf{P}_{\mathbf{xx}} \mathbf{H}^T + \mathbf{P}_{\mathbf{v}}, \tag{9}$$

where $\mathbf{P}_{\mathbf{v}}$ is the covariance matrix of the vector $\mathbf{v}$. Of vital importance is the following circumstance known from the estimation theory (Kalman, 1960, Meditch, 1969). If $f(\mathbf{x} / \mathbf{y})$ is non-Gaussian, the estimate (5) is optimal in the class of linear estimates. This means that the value of the criterion (2) sought for this estimate will be lower or equal to the value of the criterion corresponding to any other linear (relative measurements) estimate. In a general case some numerical procedures have been specially developed to derive optimal estimates (Jazwinski, 1970; Stepanov, 1998; Yarlykov and Mironov, 1999).

3. BAYESIAN ESTIMATION IN THE PRESENCE OF A TRAINING SET

NN training suggests the presence of a training set. For the problem considered it means that there is a set of data

$$\{(\mathbf{y}^{(j)}, \mathbf{x}^{(j)})\}, \; j = \overline{1.n_o}, \tag{10}$$

in which the pairs $\mathbf{y}^{(j)}$, $\mathbf{x}^{(j)}$, $j = \overline{1.n_o}$ are the independent-of-each-other realizations of the random vector $\mathbf{z} = [\mathbf{x}^T \quad \mathbf{y}^T]^T$, with the p.d.f. $f(\mathbf{x},\mathbf{y})$.

Let us consider a possible statement of the estimation problem for the case when, instead of the $f(\mathbf{x},\mathbf{v})$ or $f(\mathbf{x},\mathbf{y})$, the set of data (10) is known. In other words, assume that the a priori information is given in the form of (10) and it is necessary, having this set and the measurement $\mathbf{y}$, to find the estimate $\tilde{\mathbf{x}}(\mathbf{y})$ that minimizes the following criterion:

$$\tilde{J} = \frac{1}{n_o}\sum_{j=1}^{n_o}\left\|\mathbf{x}^{(j)} - \tilde{\mathbf{x}}(\mathbf{y}^{(j)})\right\|^2 . \qquad (11)$$

As $E\left\|\mathbf{x} - \hat{\mathbf{x}}\ (\mathbf{y})\right\|^2 = \int\int\left\|\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y})\right\|^2 f(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y}$, then, in accordance with the Monte Carlo method, it is possible to write (Zaritsky, *et. al.,* 1975):

$$\lim_{n_o\to\infty}\frac{1}{n_o}\sum_{j=1}^{n_o}\left\|\mathbf{x}^{(j)} - \tilde{\mathbf{x}}(\mathbf{y}^{(j)})\right\|^2 = E\left\|\mathbf{x} - \tilde{\mathbf{x}}(\mathbf{y})\right\|^2 ,$$

i.e. criterion (11) tends to (2) as the n_o increases. It is evident that in these conditions the estimation algorithm, optimal in the sense of criterion (11), will be similar to the traditional Bayesian algorithm (3), optimal in the minimum variance sense.

The approximate solution to this problem can be found by introducing a class of parameter-dependent functions used for the calculation of the estimate. Then the criterion (11) can be written as:

$$\tilde{J}^*(\tilde{\mathbf{W}}) = \frac{1}{n_o}\sum_{j=1}^{n_o}\left\|\mathbf{x}^{(j)} - \tilde{\mathbf{x}}(\mathbf{y}^{(j)},\tilde{\mathbf{W}})\right\|^2 , \qquad (12)$$

where $\tilde{\mathbf{W}}$ is the vector or matrix determining a set of free parameters that define the function $\tilde{\mathbf{x}}(\mathbf{y},\tilde{\mathbf{W}})$.

Hence it follows that the problem of deriving the estimation algorithm is reduced to finding the parameters $\tilde{\mathbf{W}}$ determined by the minimization of the criterion formed with the use of the data of the training set (10). The algorithms based on the minimization of the criterion of the types (11), (12) are widely used in pattern recognition. They are usually called the algorithms of empirical risk minimization (Vapnik, 1982; Haykin, 1994).

From the above it follows that the statement of the problem under consideration is in full agreement with the statement for the solution of problems with the use of NN. Thus in order to find the estimate $\tilde{\mathbf{x}}(\mathbf{y},\tilde{\mathbf{W}})$, it is possible to use an NN, i.e.

$$\hat{\mathbf{x}}^{NN}(\mathbf{y}) = \mathbf{K}^{NN}(\mathbf{y},\tilde{\mathbf{W}}) , \qquad (13)$$

where $\mathbf{K}^{NN}(\mathbf{y},\tilde{\mathbf{W}})$ is the NN; $\tilde{\mathbf{W}}$ is the matrix that specifies the free parameters (biases and weighting coefficients) and $\mathbf{y}$ is the input of the NN. The matrix $\tilde{\mathbf{W}}$ is determined when the NN is trained according with the criterion (12), where $\tilde{\mathbf{x}}(\mathbf{y},\tilde{\mathbf{W}}) = \hat{\mathbf{x}}^{NN(j)}(\mathbf{y}^{(j)},\tilde{\mathbf{W}})$ is the estimate generated by the NN by the measurements $\mathbf{y}^{(j)}$ corresponding to the realization of $\mathbf{x}^{(j)}$.

The following discussion concerns the solution of the estimation problem with the use of the so-called linear NN.

4. SOLUTION OF THE ESTIMATION PROBLEM WITH THE USE OF A LINEAR NEURAL NETWORK

Let us solve the estimation problem by using a linear NN under the assumption that the training set (10) has been specified. Taking into consideration the dimensions of the vector to be estimated, the linear NN $\hat{\mathbf{x}}^{NN}(\mathbf{y},\tilde{\mathbf{W}})$ can be written as follows:

$$\hat{\mathbf{x}}^{NN}(\mathbf{y},\tilde{\mathbf{W}}) = \mathbf{w}_0 + \mathbf{W}\mathbf{y} , \qquad (14)$$

where $\tilde{\mathbf{W}} = [\mathbf{w}_0 \mid \mathbf{W}]$ is an $n\times(m+1)$ - dimensional matrix that includes an n - dimensional biases vector $\mathbf{w}_0 = [w_{10} \quad ... \quad w_{n0}]^T$ and an $n\times m$ - dimensional matrix of weighing coefficients $\mathbf{W} = [\mathbf{w}_1 \mid ... \mid \mathbf{w}_l \mid ... \mid \mathbf{w}_n]^T$, in which $\mathbf{w}_l = [w_{l1} \quad ... \quad w_{lm}]^T$ are m - dimensional vectors $l = \overline{1.n}$. This NN has a single neuron layer. The number of neurons is the same as the dimensions of the estimated vector $\mathbf{x}$, and their activation function that depends on the scalar argument s represents identical transformation (linear activation function), i.e. $\psi(s) = s$, $-\infty < s < \infty$ (Haykin, 1994). Using (14), the criterion (12) can be represented in the following form

$$\tilde{J}^*(\tilde{\mathbf{W}}) = \frac{1}{n_o}\sum_{j=1}^{n_o}\left\|\mathbf{x}^{(j)} - (\mathbf{w}_0 + \mathbf{W}\mathbf{y}^{(j)})\right\|^2 . \qquad (15)$$

From the previous part it follows that the estimate (14) determined with the use of NN trained in accordance with the criterion (15) will tend to the optimal estimate (5) as the number of realizations n_o increases. To do this, one should, similarly to the way it was done in Reference (Stepanov and Amosov, 2004) calculate partial derivatives with respect to $\mathbf{w}_0$ and $\mathbf{W}$, and put them to zero. After some not complicated but tiresome transformations the derived equations can be resolved with respect to $\mathbf{w}_0$ and $\mathbf{W}$. As the result, the estimate $\hat{\mathbf{x}}^{NN}(\mathbf{y},\tilde{\mathbf{W}})$ derived by the measurements $\mathbf{y}$ with the use of NN (14) trained in accordance with (15) can be given as:

$$\hat{\mathbf{x}}^{NN}(\mathbf{y},\widetilde{\mathbf{W}}) = \bar{\mathbf{x}}^{*} + \mathbf{P}^{*}_{\mathbf{xy}}\left(\mathbf{P}^{*}_{\mathbf{yy}}\right)^{-1}[\mathbf{y}-\bar{\mathbf{y}}^{*}], \qquad (16)$$

where $\bar{\mathbf{x}}^{*} = \mathbf{m}^{*}_{\mathbf{x}}$; $\bar{\mathbf{y}}^{*} = \mathbf{m}^{*}_{\mathbf{y}}$; $\mathbf{P}^{*}_{\mathbf{yy}}$, $\mathbf{P}^{*}_{\mathbf{xy}}$ are the sample values of the mathematical expectations and corresponding covariance matrices:

$$\mathbf{m}^{*}_{\mathbf{x}} = \frac{1}{n_o}\sum_{j=1}^{n_o}\mathbf{x}^{(j)}; \qquad \mathbf{m}^{*}_{\mathbf{y}} = \frac{1}{n_o}\sum_{j=1}^{n_o}\mathbf{y}^{(j)}, \qquad (17)$$

$$\mathbf{P}^{*}_{\mathbf{yy}} = \alpha^{*}_{2}[\mathbf{y}] - \mathbf{m}^{*}_{\mathbf{y}}(\mathbf{m}^{*}_{\mathbf{y}})^{T}; \qquad (18)$$

$$\mathbf{P}^{*}_{\mathbf{xy}} = \alpha^{*}_{1,1}[\mathbf{x},\mathbf{y}] - \mathbf{m}^{*}_{\mathbf{x}}(\mathbf{m}^{*}_{\mathbf{y}})^{T}. \qquad (19)$$

$$\alpha^{*}_{1,1}[\mathbf{x},\mathbf{y}] = \frac{1}{n_o}\sum_{j=1}^{n_o}\mathbf{x}^{(j)}(\mathbf{y}^{(j)})^{T};$$

$$\alpha^{*}_{2}[\mathbf{y}] = \frac{1}{n_o}\sum_{j=1}^{n_o}\mathbf{y}^{(j)}(\mathbf{y}^{(j)})^{T}.$$

From Expressions (16) and (5) it is follows that NN, after some adequate training under the conditions when the specified sample values of mathematical expectations and covariance matrices are close to their true values, provide the determination of the estimate close to optimal in the linear class. Thus, the optimal linear algorithm can be treated as a neural network of the simplest kind trained in accordance with (15).

5. EXAMPLES

Example 1. It is necessary to estimate the random variable x, uniformly distributed on the interval $[0, b]$, from the noisy measurements of the form

$$y_l = x + v_l,\ l = \overline{1.i}, \qquad (20)$$

in which the measurement errors v_l, $l = \overline{1.i}$ are the random zero-mean Gaussian variables independent of each other and of x with the covariance r^2. In this example $\mathbf{x} \equiv x$, $\mathbf{y} \equiv [y_1 \ \ldots \ y_i]^T$, $\mathbf{H} = [1 \ \ldots \ 1]^{\mathrm{T}}$, $\mathbf{v} = [v_1 \ \ldots \ v_i]^{\mathrm{T}}$. It should be noted that the a posteriori p.d.f. $f(\mathbf{x}/\mathbf{y})$ is non Gaussian here, as x is a uniformly distributed random variable.

It is possible to find the linear optimal estimate $x^{*}(\mathbf{y})$ and the corresponding error covariance P^{*}_{e} by using (5), (6), i.e.

$$x^{*}(\mathbf{y}) = \bar{x} + \mathbf{P}_{x\mathbf{y}}\mathbf{P}^{-1}_{\mathbf{yy}}[\mathbf{y}-\bar{\mathbf{y}}], \qquad (21)$$

$$P^{*}_{e} = P_{xx} - \mathbf{P}_{x\mathbf{y}}\mathbf{P}^{-1}_{\mathbf{yy}}\mathbf{P}_{\mathbf{y}x}, \qquad (22)$$

where $\bar{x} = \frac{b}{2}$; $\bar{\mathbf{y}} = \frac{1}{2}[b \ \ \ldots \ \ b]^{T}$;

$$P_{xx} = \sigma^2_x = b^2/12;\ \mathbf{P}_{x\mathbf{y}} = \sigma^2_x\mathbf{H}^T;$$

$$\mathbf{P}_{\mathbf{yy}} = \sigma^2_x\mathbf{I}_i + r^2\mathbf{E}_i. \qquad (23)$$

Here $\mathbf{I}_i$ is a square matrix composed of 1; $\mathbf{E}_i$ – a unit matrix, $\sigma^2_x = b^2/12$.

The optimal Bayesian (nonlinear optimal) estimate (3) can be determined as:

$$\hat{x}(\mathbf{y}) = \frac{1}{f(\mathbf{y})}\int_0^b xf(\mathbf{y}/x)dx, \qquad (24)$$

where

$$f(\mathbf{y}) = \int_0^b f(\mathbf{y}/x)dx, \qquad (25)$$

$$f(\mathbf{y}/x) = \frac{1}{(2\pi)^{i/2}r^{i}}\exp\left(-\frac{1}{2r^2}\sum_{l=1}^{i}(y_l - x)^2\right). \qquad (26)$$

Assume that the a priori information is represented by a set of pairs $x^{(j)}$, $\mathbf{y}^{(j)}$, $j = \overline{1.n_o}$. Then the estimation problem can be solved by using NN, in particular, using a linear single-layer NN with one neuron, with the identity activation function $\psi(s) = s$ and i inputs. Below are the results obtained by the simulation, corresponding to the linear and nonlinear optimal and NN algorithms derived for different number of measurements i. For the simulation it was assumed that $i = \overline{1.k}$, $k = 1,2,...,10$.

The calculation of the optimal estimate and the normalizing factor in accordance with (24)–(26) involved numerical integration. Training was performed in accordance with the criterion (15), based on the iterative Widrow-Hoff algorithm (Haykin, 1994). To provide training, the realizations $x^{(j)}$, $\mathbf{y}^{(j)}$, $j = \overline{1.n_o}$, $n_o = 3000$ were simulated in accordance with (20). Training was followed by testing. For this purpose $n_\omega = 1000$ pairs of the realizations $x^{(j)}$, $\mathbf{y}^{(j)}$, $j = \overline{1.1000}$ were simulated for various $i = \overline{1.k}$, $k = 10$.

Fig. 1 shows the sample r.m.s. estimation errors for the nonlinear optimal ($\widetilde{\sigma}_i$) and NN ($\widetilde{\sigma}^{NN}_i$) estimates. It should be noted that from the simulation results it also follows that $\widetilde{\sigma}^{NN}_i \approx \sqrt{P^{*}_{e}}$, in other words, the NN estimates are close to the linear optimal estimates.

The values $\widetilde{\sigma}_i$, $\widetilde{\sigma}^{NN}_i$ were calculated as

$$\widetilde{\sigma}_i \approx \sqrt{\frac{1}{n_\omega}\sum_{j=1}^{n_\omega}(e^{(j)}_i)^2},$$

$$e_i^{(j)} = x^{(j)} - \hat{x}^{(j)}(\mathbf{y}^{(j)});$$

$$\widetilde{\sigma}_i^{NN} \approx \sqrt{\frac{1}{n_\omega}\sum_{j=1}^{n_\omega}(e^{NN}{}_i^{(j)})^2},$$

$$e^{NN}{}_i^{(j)} = x^{(j)} - \hat{x}^{NN(j)}(\mathbf{y}^{(j)}, \widetilde{\mathbf{W}}).$$

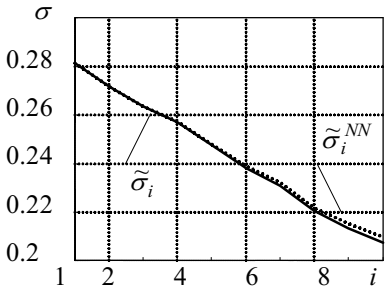

Fig. 1. The r.m.s. estimation errors x_i.

As seen from Fig. 1, the estimate of the trained NN is close to the optimal estimate. Therefore it follows that in this example there is no need for using a more complicated NN.

It is of importance to note that the derivation of the optimal estimate involved a priori information presented in the form of the analytic dependence (20) and the known joint probability distribution density function $f(x,\mathbf{y}) = f(x)f(\mathbf{y}/x)$. At the same time the derivation of estimates with the use of NN only involved a set of realizations $x^{(j)}$, $\mathbf{y}^{(j)}$, $j = \overline{1.n_o}$.

Example 2. This example is similar to Example 1, but the measurement errors v_l, $l = \overline{1.i}$ are assumed to be zero-mean random variables independent of each other and of x uniformly distributed on the interval $[-a/2, a/2]$.

The linear optimal estimate can be calculated by using (21), but in (23) r^2 must be determined as $r^2 = a^2/12$. It is essential that the optimal nonlinear estimate can be determined exactly for this example. To explain it, let us introduce the domain Ω that represents the crossing of all the intervals $[y_l - a/2, y_l + a/2], l = \overline{1.i}$, i.e.

$$\Omega \equiv [d_1, d_2] = \bigcap_{l=1}^{i}[y_l - a/2, y_l + a/2]. \qquad (27)$$

It can be shown that the a posteriori density in the example considered is uniform on the interval $[c_1, c_2]$, which represents the crossing of the a priori domain $[0, b]$ and the domain Ω so that $c_1 = \max\{0, d_1\}$, $c_2 = \min\{b, d_2\}$. Then it follows that

$$\hat{x}(\mathbf{y}) = \frac{(c_2 + c_1)}{2}. \qquad (28)$$

Assuming, just as in Example 1, that the a priori information is represented by a set of pairs $x^{(j)}$, $\mathbf{y}^{(j)}$, $j = \overline{1.n_o}$, the estimation problem can be solved by using NN. Let us use both a linear single-layer NN with one neuron with the identity activation function and i inputs and a nonlinear NN2 – a two-layer NN with i inputs, q neurons in the hidden layer and one neuron in the output layer. The NN2 output can be written as:

$$\hat{x}^{NN2}(\mathbf{y}) = \psi\left(\sum_{\mu=1}^{q}\left(w_{1\mu}^2\varphi\left(\sum_{l=1}^{i}(w_{\mu l}^1 y_l) + w_{\mu 0}^1\right)\right) + w_{10}^2\right),$$

where $w_{\mu 0}^1$, $w_{\mu l}^1$, $\mu = \overline{1.q}$, $l = \overline{1.i}$ – the bias and the weights of the hidden layer neurons; w_{10}^2, $w_{1\mu}^2$, $\mu = \overline{1.q}$ – the bias and the weights of the NN2 output layer neuron; $\varphi(s) = th\, s = \dfrac{e^s - e^{-s}}{e^s + e^{-s}}$ – the activation function for the neurons of the hidden layer; $\psi(s) = s$ – the activation function for the output neuron.

Below are the simulation results corresponding to the linear and nonlinear optimal estimates and linear and nonlinear NN estimates derived for different number of measurements i. The simulation was performed under the assumption that $b = 1$, $a = 1$, $q = 20$, $i = \overline{1.k}$, $k = 1,2,...,10$. The calculation of the optimal nonlinear estimate was carried out in accordance with (28).

Training of the linear and nonlinear NN in accordance with (20) required simulation of the realizations $x^{(j)}$, $\mathbf{y}^{(j)}$, $j = \overline{1.n_o}$, $n_o = 3000$. Training was followed by checking, for which $n_\omega = 1000$ pairs of realizations $x^{(j)}$, $\mathbf{y}^{(j)}$ were additionally simulated for various $i = \overline{1.k}$, $k = 10$.

Figure 2 shows the sample r.m.s. errors: $\widetilde{\sigma}_i^* \approx \sqrt{P_e^*}$ - for the linear optimal estimates; $\widetilde{\sigma}_i$ – for the nonlinear optimal estimates; $\widetilde{\sigma}_i^{NN1}$ – for the linear NN estimates; $\widetilde{\sigma}_i^{NN2}$ – for the nonlinear NN estimates.

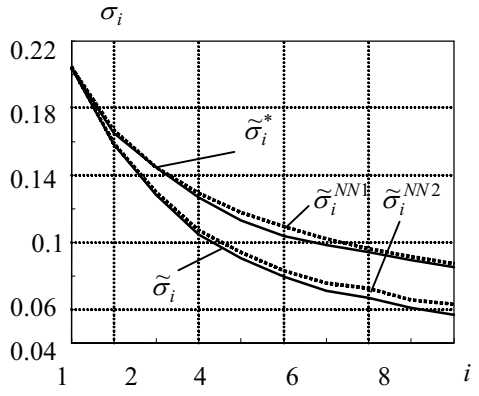

Fig. 2. The r.m.s. estimation errors x_i.

As seen from Fig. 2 the estimate of the trained linear NN and the optimal linear estimate are identical, but they differ very much from the optimal nonlinear estimate. At the same time the estimate of the nonlinear NN2 is close to the optimal nonlinear estimate.

6. CONCLUSIONS

It is shown that after some adequate training the suggested linear NN provides determination of the estimate close to optimal in the linear class. From the practical point of view the result derived may be of advantage in the situations when, on the one hand, it is known that the use of linear estimation algorithms allows achievement of acceptable accuracy, on the other hand, the a priori information about the properties of the vectors being estimated and measured is presented by a set of pairs $\mathbf{y}^{(j)}$, $\mathbf{x}^{(j)}$, $j = \overline{1.n_o}$. In this case the use of a linear NN (14) trained by using this set of data allows deriving estimates that are close in their properties to optimal linear estimates.

Of more interest is the study of the efficiency of using NN for the solution of the so-called essentially nonlinear estimation problems, for which the algorithms optimal in the linear class cannot provide the acceptable accuracy of solution (Stepanov, 1998; Dmitriev *et.al.* 1998). In this case the a posterior p.d.f. is non-Gaussian and the problem can be solved in a similar way, but then NN has to be chosen a nonlinear function.

It is clear that the efficiency of the such solution will depend largely on how good the selection of parametrized functions was from the accuracy viewpoint, i.e. proximity of the minimum value of the criterion $\widetilde{J}^*$ for the specified class of functions as compared to the minimum value of the criterion $\widetilde{J}$ without introduction of any restrictions on the class of functions. Besides, of no small importance is the complexity of both the training algorithm – finding the parameters that provide the minimum of the criterion (12), and the algorithm for the calculation of the estimate itself. The first circumstance seems to be of vital importance as after the substantiated selection of the NN topology (structure) its training can be carried out with the methods developed for training NN used for the solution of other problems. It is in this direction that further investigations in application of NN for the solution of applied essentially nonlinear estimation problems will be conducted.

REFERENCES

Alessandri A., T. Parisini, and R. Zoppoli. Sliding-window neural state estimation in a power plant heater line. *Proceedings of the American Control Conference San Diego, California, June 1999* pp. 880–884.

Gelb A. (1974) *Applied Optimal Estimation* M.I.T. Press, Cambridge, MA

Dmitriev S.P. and O.A. Stepanov (1998). Nonlinear filtering and navigation. *Proceedings of 5th International Conference on Integrated Navigation Systems. Russia, CSRI Elektropribor, Saint Petersburg* pp. 138–149.

Haykin, S. (1994). *Neural networks: A comprehensive foundation.* MacMillan College, New York.

Haykin S. and P. Yee (1997). Optimum nonlinear filtering. *IEEE Trans. On Signal Processing,* **45**(11), 2774–2786.

Haykin S. (2001). *Kalman filtering and neural networks*. Jonh Wiley & Sons, Inc., New York.

Jazwinski, A.H. (1970). *Stochastic processes and filtering theory*. Academic Press, New York.

Kalman R.E. (1960) New Approach to Linear Filtering and Prediction Problems. *J. Basic Eng. March*, 35-46.

Lo J. T. H. (1994). Synthetic approach to optimal filtering. *IEEE Trans. Neural Networks*, **5**(5), 803–811.

Meditch, J.S. (1969). Stochastic optimal linear estimation and control. Mc. Graw Hill, New York.

Parlos A. G., S.K. Menon, and A.F. Atiya (2001). An algorithmic approach to adaptive state filtering using recurrent neural networks. *IEEE Trans. Neural Networks*, **12**(6), 1411–1432.

Puskorius G.V., L.A Feldkamp, and L.I. Davis (1996). Dynamic neural network methods applied to on-vehicle idle speed control. *Proc. IEEE*, **84**, 1407–1419.

Simandle M, P. Hering, L. Kral (2004). Identification of Nonlinear Non-Gaussian Systems by Neural Network. *Proceeding of NOLCOS-04.* Sept. 1–4, Stuttgart, Germany, pp. 919–924.

Stepanov, O.A. (1998). *Nonlinear filtering and its application in navigation.* Russia, CSRI Elektropribor, Saint Petersburg. In Russian.

Stepanov O.A. and O.S. Amosov (2004). Nonrecurrent linear estimation and neural networks. *IFAC Workshop on Adaptation and Learning in Control and Signal Processing, and IFAC Workshop on Periodic Control Systems.* Yokohama, Japan, August 30 – September 1, pp. 213–218.

Yarlykov, M.S. and M.A. Mironov (1999). The Markov theory of estimating random processes. In: *Telecommunications and Radioengineering.* **50**(2–12). Published by Begell House, Inc., New York.

Zaritsky V.S., V.B. Svetnik, and L.I. Shimelevich (1975). The Monte-Carlo techniques in problems of optimal information processing. *Automaion and Remote Control*, **36**, 2015–2022.

Vapnik V.N. (1982). Estimation of dependences based on empirical data. New York: Springer-Verlag.