DOI https://doi.org/10.15407/usim.2020.03.015

УДК 004.852

O.G. RUDENKO, Doctor of Technical Sciences, Professor, Head of Information Systems Departament, Kharkiv National University of Radioelectronics,
Nauky ave., 14, Kharkiv, 61166, Ukraine,
oleh.rudenko@hneu.net

O.O. BEZSONOV,Doctor of Technical Sciences, Professor,
Kharkiv National University of Radioelectronics,
Nauky ave., 14, Kharkiv, 61166, Ukraine,
oleksandr.bezsonov@hneu.net

ADALINE ROBUST MULTISTEP TRAINING ALGORITHM

The article considers the multi-step ADALINE training algorithm when using the correntropy information criterion as a learning criterion, determines the conditions for the convergence of the algorithm, and shows that in the steady state the resulting estimate is unbiased. The importance of choosing the width of the Gaussian core, which affects the convergence rate of the estimation algorithms and the error in the steady state, is noted, and the feasibility of developing procedures for adaptive correction of the core width is indicated.

Keywords: ADALINE, correntropy, least squares method, adaptive core width correction, algorithm convergence.

Introduction

Adaptive linear element (ADALINE) was the first linear neural network proposed by Widrow B. and Hoff M., and became an alternative to the perceptron [1]. Subsequently, this element and its learning algorithm are being very commonly used in problems of identification, control, filtering, etc. The learning algorithm of Widrow — Hoff is the Kaczmarz algorithm for solving systems of linear algebraic equations [2]. Properties of this algorithm dealt with the solution of the identification problem is sufficiently described in [3]. In [4], regularized Kaczmarz algorithm (Widrow-Hoff's) was used for training ADALINE in the task of estimating non stationary parameters.

The Problem of ADALINE Learning

ADALINE shown in Figure, is described by the equation:

$$y_{n+1} = c^{*T} x_{n+1} + \xi_{n+1}, \qquad (1)$$

where y_{n+1} — is the observed output signal; $x_{n+1} = (x_{1,n+1}, x_{2,n+1}, ..x_{N,n+1})^T$ — vector of output signals $N \times 1$; $c^* = (c_1^*, c_2^*, ..c_N^*)^T$ — is the vector of desired parameters $N \times 1$, ξ_{n+1} ; — is the obstacle; n — is the discrete time.

The task of its learning consists in the definition (estimation) of the vector of parameters c^* and is reduced to minimize some of the chosen in advance performance functional (identification criterion)

$$F[e_n] = \sum_{i=1}^{n} \rho(e_i), \qquad (2)$$

where $e_i = y_i - \hat{y}_i$; $\hat{y}_i = c_{i-1}^T x_i$ — is the output model signal; c — vector estimation c^*;$\rho(e_i)$— some differential loss function satisfying the conditions:

$$\rho(e_i) \geq 0;$$

$$\rho(0) = 0;$$

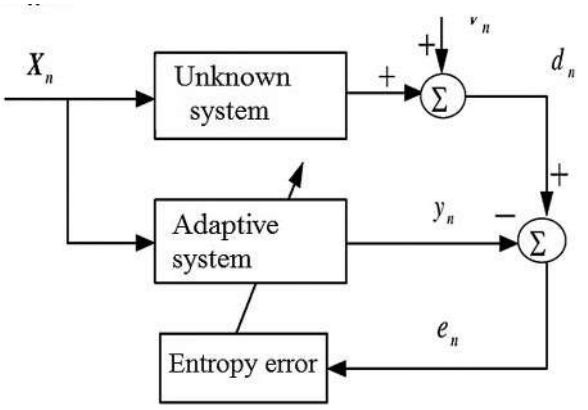

Fig. ADALINE

$$\rho(e_i)=\rho(-e_i);$$

$$\rho(e_i)\geq\rho(e_j) \quad \text{for} \quad |e_i|\geq|e_j|.$$

The learning objective is to search for estimate c defined as the solution of an minimum extreme problem

$$F(c)=\min, \tag{3}$$

or as solving equation system

$$\frac{\partial F(e)}{\partial c_j}=\sum_{i=1}^{n}\rho'(e_i)\frac{\partial e_i}{\partial c_j}=0, \tag{4}$$

where $\rho'(e_i))=\dfrac{\partial\rho(e_i)}{\partial e_i}$ — is the function of influence.

If we introduce the weigh function $\omega(e)=\rho'(e)/e$, the system of equations (4) may be put as following:

$$\sum_{i=1}^{n}\omega(e_i)e_i\frac{\partial e_i}{\partial c_j}=0, \tag{5}$$

while functional minimization (2) will be equivalent to minimizing a weighted quadratic functional, most of ten seen in practice

$$\min\sum_{i=1}^{n}\omega(e_i)e_i^2. \tag{6}$$

A quadratic functional the most widely used in estimating the parameters uses the second order statistics of the error signal and is quite optimal in assuming linearity and Gauss nature of signals. Indeed, when choosing $\rho(e_i)=0,5e_i^2$ the influence function $\rho'(e_i)=e_i$, i.e. grows linearly with the increase of e_i, that explains the volatility of the least squares method valuation to outliers and distortions with big distribution “tails”.

Stable *M*-estimation is also estimation , defined as solving an extremal problem (3) or solving a system of equations (4), however loss function $\rho(e_i)$ is chosen as different from the quadratic one.

There are quite a number of functionals that provide the robust *M*-estimates but the most common are combined functionals proposed by Huber [5] and Hampel [6] consisting of quadratic, that ensures optimal estimates for the Gaussian distribution, and modular, that allows to get an estimate that is more robust to distributions with heavy "tails" (outliers). However, the effectiveness of the resulting robust estimations depends significantly on many parameters used in these criteria and chosen depending on the experience of the researcher.

The practical application of these functionals for solving the identification problem was considered in many works, in [7–9], in particular.

Another approach to obtain robust estimates, devoid of this drawback, is the use of the fourth degree criterion [10], combined criteria using a combination of the quadratic criterion and the criterion of smallest moduli [11–13], the quadratic criterion and the fourth degree criterion [14], the fourth degree criterion and the criterion of smallest moduli [15, 16]. It should be noted that the use of the combined criterion turned out to be very effective and much simpler when implementing the identification procedure.

Another approach that is currently widely used is the approach based on information characteristics of signals, entropy, in particular. The functional used in this case is an explicit functional of the probability density function (PDF) and includes all the higher-order statistical properties defined in PDF. Since entropy measures the mean uncertainty contained in a given PDF, minimizing it provides a reduction in error. In [17, 18], the concept of information theoretic learning (ITL) was introduced, using as a criterion the R nyi quadratic entropy, for which a nonparametric estimate based on Parzen windows with Gauss kernels is determined directly from data samples. In these works, it was proved that when using the Renyi entropy, as

a result of training, the Renyi distance between the conditional probability of the density function of the desired and actual output signals for the given input signals is minimized. In [19], a more general criterion for the entropy of the error was proposed — the (h, φ)-entropy criterion, which covers various definitions of entropy.

The results of numerous studies indicate that in the presence of non-Gaussian, in particular, impulse noise, in measurements, an approach based on information characteristics of signals is very effective, while a criterion that considers all statistics of a higher-order error signal turns out to be more appropriate. Correntropy was introduced in [20, 21] as a generalized measure of similarity, the maximization of which underlies the development of sufficiently simple and efficient robust algorithms.

Correntropy as a Measure of Similarity

Correntropy, defined as a localized measure of similarity, has proven to be very efficient for obtaining robust estimates due to its less sensitivity to outliers. Its name emphasizes the relationship with correlation, and also indicates the fact that its average value over time or measurements is associated with entropy, more precisely, with the argument of the logarithm in the quadratic R nyi entropy, estimated with the help of Parzen windows [22].

For two random variables X and Y, the correntropy is defined as

$$V(X,Y) = M\{k_\sigma(X,Y)\}, \qquad (7)$$

where $M\{\bullet\}$ — is the expectation symbol; $k_\sigma(\bullet)$ — rotation invariant Mercer kernels; σ — kernel width.

The most widely used in calculating the correntropy are Gaussian ones, defined by the formula

$$k_\sigma(X,Y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{\|X-Y\|^2}{2\sigma^2}\right\}. \qquad (8)$$

When calculating the correntropy, it is necessary to know the joint distribution of random variables X and Y, which, as a rule, is not known. Since in practice there are usually a finite number of samples $\{x_i, y_i\}$, $i = 1,2,...,N$, the most simple estimate of the correntopy is calculated as follows:

$$\hat{V}(X,Y) = \frac{1}{N}\sum_{i=1}^{N} k_\sigma(x_i - y_i). \qquad (9)$$

In tasks of identification, filtering, etc. as a functional, the correntropy between the required output signal d_i and the model output signal (real) y_i is used. When using Gaussian kernels, the optimized functional takes the form

$$J_{corr}(n) = \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{N}\sum_{i=n-N+1}^{N}\exp\left(-\frac{e_i^2}{2\sigma^2}\right), \qquad (10)$$

where $e_i = d_i - y_i$ — is the identification (filtration) error.

The use of the Taylor series expansion for the Gaussian kernel makes it possible to write the correntropy as follows:

$$V(X,Y) = \frac{1}{\sqrt{2\pi}\sigma}\sum_{n=0}^{\infty}\frac{(-1)^n}{2^n\sigma^{2n}n!}M\left\{\|X-Y\|^{2n}\right\}. \qquad (11)$$

Expression (11) includes all moments of an even order of a random variable $\|x_i - y_j\|$. In particular, the term corresponding to $n = 1$ in (11) is proportional to

$$M\left\{\|x_i\|^2\right\} + M\left\{\|y_j\|^2\right\} - 2M\{x_i y_j\} = \sigma_{x_i}^2 + \sigma_{y_j}^2 - 2R_{xy}(i,j).$$

It can be seen from this expression that the information described by the usual autocorrelation function is included in the new function.

It's not hard to see that the choice $\rho(e_i) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{e_i^2}{2\sigma^2}\right)$ meets the above requirements for $\rho(e_i)$. Indeed [17],

$$\min_\theta \sum_{i=1}^{n}\rho(e_i) = \min\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\left(1-\exp\left(-\frac{e_i^2}{2\sigma^2}\right)\right) \Leftrightarrow$$

$$\Leftrightarrow \max_0 \rho(e_i) = \sum_{i=1}^{n}\exp\left(-\frac{e_i^2}{2\sigma^2}\right)\Big/\frac{1}{\sqrt{2\pi}\sigma} = \max_0\sum_{i=1}^{n}k_\sigma(e_i),$$

and the weight function used in (6) looks as follows:

$$\omega(e_i) = \exp\left(-\frac{e_i^2}{2\sigma^2}\right)\Big/\frac{1}{\sqrt{2\pi\sigma^3}}.$$

Thus, the correntropy allows to obtain robust estimates.

Correntropy Maximization Algorithms

The gradient optimization algorithm (10) at N=1 will have the form [23, 24]

$$w_{n+1} = w_n + \gamma \exp\left(-\frac{e_{n+1}^2}{2\sigma^2}\right) e_n x_{n+1}, \quad (12)$$

where γ — parameter affecting the rate of convergence.

A significant drawback of this algorithm is the low convergence rate, which significantly limits the possibility of its use in identifying nonstationary objects. It should be noted that finding the optimal value of the parameter γ, that provides the maximum convergence rate of the algorithm, equal, as it is easy to show,

$$\gamma_{n+1} = \left(\psi_{n+1}\|x_{n+1}\|^2\right)^{-1}, \quad (13)$$

where $\psi_{n+1} = \exp\left(-\frac{e_{n+1}^2}{2\sigma^2}\right)$, leads to an analogue of Kaczmarz algorithm (Widrow–Hoff's).

In [25–28], to combat impulse noise, a recurrent weighted least squares (RWLS) method was proposed, which minimizes the criterion

$$\psi_{n+1} = \exp\left(-\frac{e_{n+1}^2}{2\sigma^2}\right), \quad (14)$$

and having the form

$$c_{n+1} = c_n + \frac{\psi_{n+1} P_n x_{n+1}}{\lambda + \psi_{n+1} x_{n+1}^T P_n x_{n+1}} (y_{n+1} - c_n^T x_{n+1}) \quad (15)$$

$$P_{n+1} = \lambda^{-1}\left(P_n - \frac{\psi_{n+1} P_n x_{n+1} x_{n+1}^T P_n}{\lambda + \psi_{n+1} x_{n+1}^T P_n x_{n+1}}\right). \quad (16)$$

Where $0 \le \lambda < 1$ — is the weighing factor.

Thus, when deriving the formula for calculating P_{n+1} (16), the approximation was used

$$P_{n+1} = \lambda P_n + \psi_{n+1} x_{n+1} x_{n+1}^T. \quad (17)$$

As known, introducing a parameter λ into an algorithm is advisable when identifying nonstationary parameters.

Another approach used to estimate nonstationary parameters is the use of a limited number of measurements in RWLS, that leads to the algorithm of the current regression analysis method [29].

Learning Algorithm based on the Current Regression Analysis Algorithm

The current regression analysis algorithm (CRA), which has the form

$$c_{n+1|L} = (X_{n+1|L}^T X_{n+1|L})^{-1} X_{n+1|L}^T Y_{n+1|L}, \quad (18)$$

was proposed in [29], and in [30] a modification of this algorithm using the mechanism of forgetting the past information (smoothing) was considered. $L = \text{const}(L \ge N)$ here is the memory of the algorithm.

A feature of algorithms with L=const is that the matrices and observation vectors used in constructing estimates at each estimation step are formed as follows: they include information about newly received measurements and exclude information about the oldest ones. Depending on how these matrices and vectors are formed (whether new information is added first, and then outdated information is excluded, or obsolete information is excluded first and then new information is added), two forms of assessment are possible. Let's look at this in greater detail.

Obtaining new information (adding up a new dimension) leads to the calculation of an estimate, which, by analogy with (18), can be written as follows:

$$c_{n+1|L+1} = (X_{n+1|L+1}^T X_{n+1|L+1})^{-1} X_{n+1|L+1}^T Y_{n+1|L+1} \quad (19)$$

where

$$Y_{n+1(L+1} = \begin{pmatrix} Y_{n(L} \\ --- \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} y_{n-L+1} \\ --- \\ Y_{n+1(L} \end{pmatrix} \text{— vector } (L+1)\times 1; \quad (20)$$

$$X_{n+1(L+1} = \begin{pmatrix} X_{n(L} \\ --- \\ x_{n+1}^T \end{pmatrix} = \begin{pmatrix} x_{n-L+1}^T \\ --- \\ X_{n+1(L} \end{pmatrix} \text{— matrix } (L+1)\times N. \quad (21)$$

Modified Recurrent CRA Algorithm to Maximize the Correntropy

Consider a modification of the current regression analysis algorithm used to maximize the correntropy (14) and which will have the form

$$c_{n+1|L} = (X_{n|L}^T X_{n|L} + \psi_{n+1} x_{n+1} x_{n+1}^T - \psi_{n-L+1} x_{n-L+1} x_{n-L+1}^T)^{-1} \times$$
$$\times (x_{n-L+1} X_{n|L}^T \vdots x_{n+1}) \begin{bmatrix} y_{n-L+1} \\ Y_{n|L} \\ --- \\ y_{n+1} \end{bmatrix}. \quad (22)$$

We introduce the following designations:

$$P_{n+1|L+1}^{-1} = X_{n+1|L+1}^T A_{L+1} X_{n+1|L+1} = P_{n|L}^{-1} + \psi_{n+1} x_{n+1} x_{n+1}^T; \quad (23)$$
$$P_{n+1|L}^{-1} = X_{n+1|L}^T A_L X_{n+1|L} = P_{n+1|L+1}^{-1} - \psi_{n-L+1} x_{n-L+1} x_{n-L+1}^T.$$

Then

$$P_{n+1|L}^{-1} = P_{n|L}^{-1} + \psi_{n+1} x_{n+1} x_{n+1}^T - \psi_{n-L+1} x_{n-L+1} x_{n-L+1}^T. \quad (24)$$

Applying the matrix inversion lemma to (23), (24), one can obtain, as already noted, two forms of computation: in one, the accumulation of information is used (the newly received signal x_{n+1} is turned on), and then obsolete information is discarded (the signal x_{n-L+1} is excluded) and vice versa.

So the refinement of the estimates and the calculation of the matrix when dumping outdated information occurs according to the formulas

$$c_{n|L-1} = c_{n|L} - \frac{\psi_{n-L+1} P_{n+1|L+1} x_{n-L+1}}{\psi - x_{n-L+1}^T P_{n+1|L+1} x_{n-L+1}} \times (25)$$
$$\times (y_{n-L+1} - c_{n|L}^T x_{n-L+1});$$

$$P_{n|L-1} = P_{n|L} + \frac{\psi_{n-L+1} P_{n+1|L+1} x_{n-L+1} x_{n-L+1}^T P_{n+1|L+1}}{1 - \psi_{n-L+1} x_{n-L+1}^T P_{n+1|L+1} x_{n-L+1}}, \quad (26)$$

and the ratios describing the accumulation of information will have the form

$$c_{n+1|L} = c_{n|L-1} + \frac{\psi_{n+1} P_{n|L} x_{n+1}}{1 + \psi_{n+1} x_{n+1}^T P_{n|L} x_{n+1}} (y_{n+1} - c_{n|L-1}^T x_{n+1}); \quad (27)$$

$$P_{n+1|L+1} = P_{n|L} - \frac{\psi_{n+1} P_{n|L} x_{n+1} x_{n+1}^T P_{n|L}}{1 + \psi_{n+1} x_{n+1}^T P_{n|L} x_{n+1}}. \quad (28)$$

Thus, the recurrent estimation algorithm obtained by excluding outdated information and then adding up new information is described by ratios (25)—(28).

It is not difficult to obtain ratios for the case, first — the newly received information is added, and then — dump the obsolete information.

It should be noted that both of these procedures implement the same assessment $c_{n+1|L}$ and differ only in the rule for generating an auxiliary assessment $c_{n+1|L+1}$ or $c_{n|L-1}$.

Algorithm Convergence Study

To determine the convergence conditions for algorithm (25)—(28), we introduce the Lyapunov function [31]

$$V_{n+1|L} = \Theta_{n+1|L}^T P_{n+1|L}^{-1} \Theta_{n+1|L} \quad (29)$$

where $\Theta_{n+1|L} = c_{n+1|L} - c^*$ — is the estimation error, at the $(n + 1)$-th step, obtained from L observations.

Subtracting (25) c^* from both parts, write down an algorithm for identification errors

$$\Theta_{n+1|L} = \left(I + \frac{\psi_{n-L+1} P_{n+1|L} x_{n-L+1} x_{n-L+1}^T}{\psi - x_{n-L+1}^T P_{n+1|L} x_{n-L+1}} \right) \Theta_{n+1|L+1}, \quad (30)$$

where I — is the identity matrix NxN.

On the other hand, taking into account (26), the relation (30) can be rewritten as follows:

$$\Theta_{n+1|L} = P_{n+1|L} P_{n+1|L+1}^{-1} \Theta_{n+1|L+1}. \quad (31)$$

Writing (27) in a similar way with respect to estimation errors and taking into account (28), thus obtain

$$\Theta_{n+1|L} = \left(I - \frac{\psi_{n+1} P_{n|L} x_{n+1} x_{n+1}^T}{1 + \psi_{n+1} x_{n+1}^T P_{n|L} x_{n+1}} \right) \Theta_{n|L-1} =$$
$$= P_{n+1|L+1} P_{n|L}^{-1} \Theta_{n|L}. \quad (32)$$

Substitution of (32) into (31) gives

$$\Theta_{n+1|L} = P_{n+1|L} P_{n|L}^{-1} \Theta_{n|L}. \quad (33)$$

Then

$$V_{n+1|L} = \Theta_{n\,L}^T P_{n|L}^{-1} P_{n+1|L} P_{n|L}^{-1} \Theta_{n|L}, \quad (34)$$

and the increment of the Lyapunov function is

$$\Delta V_{n+1|L} = V_{n+1|L} - V_{n|L} =$$
$$= \Theta_{n|L}^T P_{n|L}^{-1} \left[P_{n+1|L} P_{n|L}^{-1} - I \right] \Theta_{n|L}. \quad (35)$$

Using formulas (23), (24), thus define

$$P_{n+1|L}P_{n|L}^{-1}=\left[P_{n|L}+\frac{\psi_{n+1}P_{n|L}x_{n+1}x_{n+1}^TP_{n|L}}{1-\psi_{n+1}x_{n+1}^TP_{n|L}x_{n+1}}\right]-$$
$$-\frac{\psi_{n+1}\left(P_{n|L}+\frac{\psi_{n+1}P_{n|L}x_{n+1}x_{n+1}^TP_{n|L}}{1-\psi_{n+1}x_{n-L+1}^TP_{n|L}x_{n-L+1}}\right)x_{n+1}x_{n+1}^T\times\left(P_{n|L}+\frac{\psi_{n+1}P_{n|L}x_{n+1}x_{n+1}^TP_{n|L}}{1-\psi_{n+1}x_{n-L+1}^TP_{n|L}x_{n-L+1}}\right)}{1-\psi_{n+1}x_{n+1}^T\left(P_{n|L}-\frac{\psi_{n+1}P_{n|L}x_{n+1}x_{n+1}^TP_{n|L}}{1-\psi_{n+1}x_{n+1}^TP_{n+1|L}x_{n-L+1}}\right)x_{n-L+1}}+P_{n|L}^{-1}. \quad (36)$$

We calculate the third term in square brackets. To simplify the notation, we introduce the following designations:

$$\begin{aligned}e_n&=\Theta_{n+1|n-L}^Tx_{n+1};\\ e_L&=\Theta_{n+1|L}^Tx_{n-L+1};\\ \alpha&=x_{n+1}^TP_{n+1|L}x_{n+1};\\ \beta&=x_{n-L+1}^TP_{n+1|L}x_{n-L+1};\\ \gamma&=x_{n+1}^TP_{n+1|L}x_{n-L+1};\\ \psi_L&=\psi_{n-L+1};\quad \psi_n=\psi_{n+1}.\end{aligned} \quad (37)$$

After simple transformations, we find that the numerator and denominator of this term will have the form

the numerator

$$\psi_n\left(1-\psi_L\beta\right)e_n^2+2\psi_n\psi_L\gamma e_ne_L+\frac{\psi_L^2\psi_n\gamma^2e_L^2}{1-\psi_L\beta};$$

the denominator

$$\frac{\left(1-\psi_L\beta\right)+\psi_n\alpha\left(1-\psi_L\beta\right)+\psi_n\psi_L\gamma^2}{1-\psi_L\beta}.$$

Substitution of these expressions into (36) and simple transformations considering the introduced notation (37) give

$$\frac{\psi_Le_L^2}{\left(1-\psi_L\beta\right)}+\frac{\psi_n\left(1-\psi_L\beta\right)e_n^2-2\psi_n\psi_L\gamma e_ne_L+\frac{\psi_L^2\psi_n\gamma^2e_L^2}{\left(1-\psi_L\beta\right)}}{\left[\left(1-\psi_L\beta\right)\left(1+\psi_n\beta\right)+\psi_n\psi_L\gamma^2\right]}=$$
$$=\frac{\psi_L\left(e_L^2+\psi_n\psi_L\alpha e_L^2-\psi_n\left(1-\psi_L\beta\right)e_n^2-2\psi_n\psi_L\gamma e_ne_L\right)}{\left[\left(1-\psi_L\beta\right)\left(1+\psi_n\beta\right)+\psi_n\psi_L\gamma^2\right]}=$$
$$=\frac{\psi_L\left[e_L^2+\psi_n\alpha e_L^2+\psi_n\beta e_n^2 2\psi_n\gamma e_ne_L\right]-\psi_ne_n^2}{\left[\psi_L\beta-\psi_n\psi_L\gamma^2+\psi_n\psi_L\alpha\beta\right]-1-\psi_n\alpha}. \quad (38)$$

For the convergence of the algorithm, the condition $V_{n+1|L}-V_{n|L}=$

$$=\frac{\psi_L\left[e_L^2+\psi_n\alpha e_L^2+\psi_n\beta e_n^2-2\psi_n\gamma e_ne_L\right]-\psi_ne_n^2}{\left[\psi_L\beta-\psi_n\psi_L\gamma^2+\psi_n\psi_L\alpha\beta\right]-1-\psi_n\alpha}\le 0, \quad (39)$$

i.e. the expression on the right side of (38) must be negative.

Consider the second term. For this term to be negative, the condition needs to be met

$$\left[\psi_L\left(e_L^2+\psi_n\psi_L\alpha e_L^2+\psi_n\beta e_n^2-2\psi_n\psi_L\gamma e_ne_L\right)-\psi_ne_n^2\right]\times$$
$$\times\left[\left(\psi_L\beta-\psi_n\psi_L\gamma^2+\psi_n\psi_L\alpha\beta\right)-1-\psi_n\alpha\right]>0. \quad (40)$$

Consider the first factor of this inequality. Taking into account the introduced notation, thus have

$$\psi_L\left(e_L^2+\psi_n\psi_L\alpha e_L^2+\psi_n\beta e_n^2-2\psi_n\gamma e_ne_L\right)-\psi_ne_n^2=$$
$$=\psi_L\left(e_L^2+\psi_n\left(x_{n+1}e_L-x_{n-L+1}e_n\right)^TP_{n+1|L}\left(x_{n+1}e_L-x_{n-L+1}e_n\right)\right)-$$
$$-\psi_ne_n^2.$$

As $\psi_L\ge 0$,

$$\psi_L\left(e_L^2+\psi_n\psi_L\alpha e_L^2+\psi_n\beta e_n^2-2\psi_n\gamma e_ne_L\right)= \quad (41)$$
$$=\psi_L\left(e_L^2+\psi_n\left(x_{n+1}e_L-x_{n-L+1}e_n\right)^TP_{n+1|L}\left(x_{n+1}e_L-x_{n-L+1}e_n\right)\right)\ge 0.$$

Denote

$$A=\frac{\psi_ne_n^2}{\psi_L\left(\lambda e_L^2+\psi_n\left(\alpha e_L^2-2\gamma e_ne_L+\beta e_n^2\right)\right)}. \quad (42)$$

For the second factor, we can write

$$\psi_n\psi_L\left(\alpha\beta-\gamma^2\right)+\psi_L\beta-1-\psi_n\alpha=$$
$$=\psi_L\left[\psi_n\left(\alpha\beta-\gamma^2\right)+\beta\right]-1-\psi_n\alpha.$$

Note that due to the fulfillment of the Cauchy–Bunyakovskyine quality [32, 33], it is true that

$$\alpha\beta-\gamma^2\ge 0$$

or

$$x_{n+1}^TP_{n|L}x_{n+1}x_{n-L+1}^TP_{n|L}x_{n-L+1}\ge$$
$$\ge x_{n+1}^TP_{n|L}x_{n-L+1}x_{n+1}^TP_{n|L}x_{n-L+1}.$$

Thus, given that $\psi_L\ge 0$,

$$\psi_L\left[\psi_n\left(\alpha\beta-\gamma^2\right)+\beta\right]\ge 0. \quad (43)$$

$$\psi_{n+1}e_{n+1}^2 - A\psi_{n-L+1}e_{n-L+1}^2 =$$
$$= \psi_{n+1}e_{n+1}^2\left[\frac{\psi_{n+1}(e_{n+1}x_{n-L+1} - e_{n-L+1}x_{n+1})^T P_{n|L}(e_{n+1}x_{n-L+1} - e_{n-L+1}x_{n+1})}{e_{n-L+1}^2 + \psi_{n+1}(e_{n+1}x_{n-L+1} - e_{n-L+1}x_{n+1})^T P_{n|L}(e_{n+1}x_{n-L+1} - e_{n-L+1}x_{n+1}) + e_{n-L+1}}\right] \geq 0. \quad (*)$$

Denote

$$B = \frac{1+\psi_n\alpha}{\psi_L\left[\psi_n\left(\alpha\beta-\gamma^2\right)+\beta\right]}. \quad (44)$$

It should be noted that due to the performance of (41) and (43) $A \geq 0$ and $B \geq 0$.

Taking into account the introduced notation, the condition (40) can be written as follows:

$$(1-A)(1-B) > 0. \quad (45)$$

Thus, to ensure the convergence of the algorithm, it is necessary that A and B be greater than one.

On the other hand, consider the difference $(B-A)$.

Substituting expressions for A and B (42) and (43), after simple calculations we have

$$B - A = \frac{\left(e_L + \psi_n\alpha e_L - \psi_n\gamma e_n\right)^2}{\psi_L\left[e_L + \psi_n\alpha e_L^2 + y_n\beta e_n^2 - 2\psi_n\gamma e_n e_L\right]\left[\psi_n\left(\alpha\beta-\gamma^2\right)+\beta\right]}.$$

And since both the numerator and the denominator of this expression are not negative, the difference $B - A \geq 0$, i.e. $B \geq A$. Thus, to satisfy the convergence conditions of the algorithm, i.e.

$$\Theta_{n+1|L}^T P_{n+1|L}^{-1}\Theta_{n+1|L} \leq \Theta_{n|L}^T P_{n|L}^{-1}\Theta_{n|L} \quad (46)$$

or inequality (45), it is necessary that $A \geq 1$.

Let us dwell on the properties of the matrix included in the Lyapunov function P^{-1}. For the algorithm to converge, this matrix must be positively definite. Consider a step-by-step change P^{-1} in accordance with (24).

If at the n-th step the matrix $P_{n|L}^{-1}$ is positively definite, then in the case of positive definiteness of the matrix $\psi_{n+1}x_{n+1}x_{n+1}^T - \psi_{n-L+1}x_{n-L+1}x_{n-L+1}^T$ the matrix $P_{n+1|L}^{-1}$ will also be positively definite. Multiplying this matrix on the left by $\Theta_{n|L}^T$, and on the right by $\Theta_{n|L}$ and taking into account the notation (40), we consider the scalar quantity

$$\Theta_{n|L}^T\left[\psi_{n+1}x_{n+1}x_{n+1}^T - \psi_{n-L+1}x_{n-L+1}x_{n-L+1}^T\right]\Theta_{n|L} =$$
$$= \psi_{n+1}e_{n+1}^2 - \psi_{n-L+1}e_{n-L+1}^2 \quad (47)$$

As

$$\psi_{n+1}e_{n+1}^2 - \psi_{n-L+1}e_{n-L+1}^2 \geq \psi_{n+1}e_{n+1}^2 - A\psi_{n-L+1}e_{n-L+1}^2,$$

then substituting the expression for A (42) into this inequality, we obtain (*).

Hence, it is clear that a matrix $P_{n+1|L}^{-1}$ will be nonnegatively definite if $P_{n|L}$ is nonnegatively definite (and hence $P_{n+1|L}^{-1}$ — is also a nonnegative definite matrix). This can be achieved by choosing an initial matrix $P_{0|L}$ that is no negatively definite, for example, as in ordinary RWLS, $P_{0|L} = \alpha I$, where α — is a positive number.

Consequently, the Lyapunov function under the indicated conditions will be nonnegative and bounded [31]

$$\Theta_{n+1|L}^T P_{n+1|L}^{-1}\Theta_{n+1|L} \leq \Theta_{n|L}^T P_{n|L}^{-1}\Theta_{n|L} \ldots \leq \theta_{0|L}^T P_{0|L}^{-1}\theta_{0|L} \quad (48)$$

i.e. it is limited by $V_{0|L} = \theta_{0|L}^T P_{0|L}^{-1}\theta_{0|L}$.

It follows from (35) that

$$V_{n+1|L} - V_{n|L} = \quad (49)$$
$$= \frac{\psi_L\left[e_L^2 + \psi_n\alpha e_L^2 + \psi_n\beta e_n^2 - 2\psi_n\gamma e_n e_L\right] - \psi_n e_n^2}{1+\psi_n\alpha - \left[\psi_L\beta - \psi_n\psi_L\gamma^2 + \psi_n\psi_L\alpha\beta\right]} \leq 0,$$

whence follows

$$\lim_{N\to\infty}\sum_{n=1}^{N}\frac{\psi_L\left[e_L^2 + \psi_n\alpha e_L^2 + \psi_n\beta e_n^2 - 2\psi_n\gamma e_n e_L\right] - \psi_n e_n^2}{1+\psi_n\alpha - \left[\psi_L\beta - \psi_n\psi_L\gamma^2 + \psi_n\psi_L\alpha\beta\right]} < 0. \quad (50)$$

In turn, it follows from (49) that

$$\lim_{n\to\infty}\frac{\psi_L\left[e_L^2 + \psi_n\alpha e_L^2 + \psi_n\beta e_n^2 - 2\psi_n\gamma e_n e_L\right] - \psi_n e_n^2}{1+\psi_n\alpha - \left[\psi_L\beta - \psi_n\psi_L\gamma^2 + \psi_n\psi_L\alpha\beta\right]} = 0,$$

i.e. the identification error decreases over time.

Let us denote the eigenvalues of the matrix $\lambda[\bullet]$. For the eigenvalues of the matrices, taking into account (23), it is true

$$\lambda_{\min}\left[P_{n+1|L}^{-1}\right] \geq \lambda_{\min}\left[P_{n|L}^{-1}\right] \geq \lambda_{\min}\left[P_0^{-1}\right],$$

and from the other hand

$$\leq \theta_{n+1}^T P_{n+1|L}^{-1}\theta_{n+1} \leq \theta_0^T P_0^{-1}\theta_0 \leq \lambda_{\max}\left[P_0^{-1}\right]\|\theta_0\|^2,$$

therefore, the value of the estimation error $\|\theta_{n+1}\|^2$ satisfies the inequality

$$\|\theta_{n+1}\|^2 \le \frac{\lambda_{\max}\left[P_0^{-1}\right]}{\lambda_{\min}\left[P_0^{-1}\right]}\|\theta_0\|^2. \qquad (51)$$

Finally, let us consider the steady-state regime, when the estimate is no longer corrected, i.e. $c_{n+1|L} = c_{n|L}$. Relation (33) for this case can bere written as follows:

$$\Theta_{n|L} = P_{n+1nL}P_{n|L}^{-1}\Theta_{n|L}$$

or

$$(I - P_{n+1|L}P_{n|L}^{-1})\Theta_{n|L} = 0.$$

The convergence of the algorithm means that this equality holds only at $\Theta_{n|L} = 0$, i.e. $c_{n|L} = c^*$, and the matrix $I - P_{n+1|L}P_{n|L}^{-1}$ is not null. The latter is true if the eigenvalues of this matrix are nonzero. Let's dwell on this in more detail.

Let us denote the eigenvalues of the matrix $\lambda[\bullet]$. It is necessary to show that

$$\lambda\left[I - P_{n+1|L}P_{n|L}^{-1}\right] \ne 0 \qquad (52)$$

or

$$1 \ne \lambda\left[P_{n+1|L}P_{n|L}^{-1}\right].$$

As known, for the maximum and minimum values of singular numbers $\sigma_{\max}, \sigma_{\min}$ and eigenvalues λ of a square matrix A the following relation is valid [34]

$$\sigma_{\min}[A] \le |\lambda[A]| \le \sigma_{\max}[A].$$

Since both matrices $P_{n+1|L}$ and $P_{n|L}^{-1}$ are square and their maximum singular values $\lambda_{\max}$ are not less than the maximum eigenvalues , we have

$$\sigma_{\max}[P_{n+1|L}P_{n|L}^{-1}] \ge \lambda_{\max}[P_{n+1|L}P_{n|L}^{-1}]. \qquad (53)$$

Using the Cauchy-Bunyakovskyine quality, we obtain

$$\sigma_{\max}[P_{n+1|L}P_{n|L}^{-1}] \le \sigma_{\max}[P_{n+1|L}] \le \sigma_{\max}[P_{n|L}^{-1}]. \qquad (54)$$

Due to the fact that, as shown above, the matrix $\psi_{n+1}x_{n+1}x_{n+1}^T - \psi_{n-L+1}x_{n-L+1}x_{n-L+1}^T$ is square and nonnegatively definite, it follows from (23) that

$$\sigma_{\max}[P_{n+1|L}^{-1}] \ge \sigma_{\max}[P_{n|L}^{-1}].$$

Therefore, inequality (54) can be written as follows:

$$\sigma_{\max}\left[P_{n+1|L}P_{n|L}^{-1}\right] \le \sigma_{\max}[P_{n+1|L}]\sigma_{\max}[P_{n+1|L}^{-1}] =$$

$$= \frac{\sigma_{\max}[P_{n+1|L}^{-1}]}{\sigma_{\min}[P_{n+1|L}^{-1}]} \le 1,$$

where $\sigma_{\min}[P_{n+1|L}]$ — is the minimal singular value of the matrix $I - P_{n+1|L}P_{n|L}^{-1}$. Taking this into account, it follows from (54) that

$$\lambda\left[P_{n+1|L}P_{n|L}^{-1}\right] < 1,$$

i.e. (52) is fair.

Thus, in the steady state $\Theta_{n|L} = 0$ or $c_{n|L} = c^*$, i.e. the estimate obtained using the considered algorithm is unbiased.

σ-parameter Selection

Gaussian variance (also called kernel width or kernel size) is a free, user-defined parameter. Therefore, when estimating the correntropy, the result depends on the chosen kernel size. In addition, th correntropy kernel size affects the character of the criterion surface, the presence of local optima, the rate of convergence, and resistance to impulse noise during adaptation [20]. If the amount of training data is not large enough, the kernel size should be chosen regarding the tradeoffs between outlier bias and efficiency estimation [35].

Various methods can be used to determine the size of the kernel, for example, the statistical method [36], Silverman's rule [37], cross-validation methods [38–40], etc.

One of the most commonly used methods for choosing an appropriate kernel width in machine learning is cross validation. Other simpler approaches include Silverman's rule of thumb [37]

$$\sigma = 0{,}9AN^{-1/5},$$

where A — the smallest value between the standard deviation of the data sample and the interquartile range of the data, scaled by 1,34; N — number of data samples.

Despite its ease of use, Silverman's rule often provides a good choice of kernel size.

The width of the Gaussian kernel significantly affects the rate of convergence of estimation algorithms. For fixed other parameters, a smaller kernel width can increase the convergence rate, but the error in the steady state may be large. If a larger kernel width is chosen, the algorithm converges slowly, but has a smaller steady-state error. All this testifies to the advisability of developing

procedures for adaptive correction of the kernel width σ [4].

Conclusion

The work considered a multistep learning algorithm for ADALINE when using the information criterion of correntropy as a learning criterion, the conditions for the convergence of the algorithm are determined and it is shown that in the steady state the resulting estimate is unbiased. The importance of choosing the width of the Gaussian kernel, which affects the rate of convergence of estimation algorithms and the error in the steady state, is noted, and the expediency of developing procedures for adaptive correction of the kernel width is indicated.

REFERENCES

1. *Widrow B., Hoff M.,* 1960. "Adaptive switching circuits", 1960 IRE WESCON Convention Record, Part 4, Institute of Radio Engineers, New York, pp. 96–104.
2. *Kaczmarz S.* Approximate solution of systems of linear equations, Int. J. of Control, 1993, 57, pp. 1269-1271. DOI: https://doi.org/10.1080/00207179308934446.
3. *Liberol B. D., Rudenko O. G., Bessonov O. O.*, 2018. "Issledovaniye skhodimosti odnoshagovykh adaptivnykh algoritmov identifikatsii", Problemy upravleniya i informatiki. 5, pp. 19–32. (In Russian).
4. *Rudenko O. G., Bessonov O. O.*, 2019. "Regulyarizovannyy algoritm obucheniya adaliny v zadache otsenivaniya nestatsionarnykh parametrov", Control Systems and Computers, 1, pp. 22-30. DOI: https://doi.org/10.15407/usim.2019.01.022. (In Russian).
5. *KHyuber P.*, 1984. Robastnost v statistike, Mir, Moscow, 304 p. (In Russian).
6. *Hampel F. R., Ronchetti E. M., Rousseeuw P. J., Stahel W. A.*, 1986. Robust Statistics. The Approach Based on Influence Functions, John Wiley and Sons, N.-Y., 526 p.
7. *Rudenko O. G., Bessonov O. O.,* 2010. "Robastnoye obucheniye veyvlet-neyrosetey", Problemy upravleniya i informatiki, 5, pp. 66–79. (In Russian).
8. *Rudenko O., Bezsonov O.*, 2011. "Function approximation using robust radial basis function networks", J. of Intelligent Learning Systems and Applications, 3, pp. 17–25.
9. *Rudenko O. G., Bessonov O. O.*, 2012. "M-obucheniye radialno-bazisnykh setey s ispolzovaniyem asimmetrichnykh funktsiy vliyaniya", Problemy upravleniya i informatiki, 1, pp. 79–93. (In Russian).
10. *Walach E., Widrow D.*, 1984. "The least mean fourth (LMF) adaptive algorithm and its family", IEEE Transactions on Information Theory, 30 (2), pp. 275–283.
11. *Chambers J., Avlonitis A.*, 1997. "A Robust Mixed-Norm Adaptive Filter Algorithm", IEEE Signal Processing Letters, 4 (2), pp. 46–48.
12. *Papoulis E. V., Stathaki T.*, 2004. "A Normalized Robust Mixed-Norm Adaptive Algorithm for System Identification", IEEE Signal Processing Letters, 11 (1), pp. 56–59.
13. *Chambers J., Tanrikulu O., Constantinides A. G.*, 1984. "Least mean mixed-norm adaptive filtering", Electronics letters, 30 (19), pp. 1574–1575.
14. *Zerguine A.,* 2012. "A variable-parameter normalized mixed-norm (VPNMN) adaptive algorithm", EURASIP Journal on Advances in Signal Processing, 55, 13 p.
15. *Rudenko O. G., Bezsonov O. O., Serdyuk N. M., Oliynyk K. O. Romanyuk O. S.*, 2019. "Robastna identyfikatsiya obyektiv za nayavnostyu nehausivskykh zavad", Bionika intellekta, 2 (93), pp. 7–12. (In Ukrainian).
16. *Rudenko O. G., Bezsonov O. O., Serdyuk N. M., Oliynyk K. O. Romanyuk O. S.*, 2020. "Robastna identyfikatsiya ob'yektiv na osnovi minimizatsiyi kombinovanoho funktsionalu", Systemy obrobky informatsiyi, 1 (160), pp. 80–88. (In Ukrainian).
17. *Principe J. C., Xu D., Zhao Q., Fisher J. W.,* 2000. "Learning from examples with information theoretic criteria", J. VLSI Signal Process. Syst., 26 (1–2), pp. 61–77.
18. *Principe J. C., Xu D., Fisher J.*, 2000. "Information Theoretic Learning", S. Haykin (Ed.), Unsupervised Adaptive Filtering, Wiley, New York, pp. 265–319.
19. *Chen B., Hu J., Pu L., Sun Z.,* 2007. "Stochastic gradient algorithm under (h, φ)-entropy criterion", Circuits Syst. Signal Process, 26, pp. 941–960.
20. *Santamar a I., Pokharel P. P, Jose C. Principe J. C.,* 2006. "Generalized Correlation Function: Definition, Properties, and Application to Blind Equalization", IEEE Trans. on Signal Processing, 54 (6), pp. 2187–2197.

21. *Liu W., Pokharel P. P, Principe J. C.*, 2007. "Correntropy: Properties and Applications in Non-Gaussian Signal Processing", IEEE Trans. on Signal Processing, 1, pp. 5286-5298.
22. *Wang W., Zhao J., Qu H., Chen B., Principe J. C.*, 2015. "An adaptive kernel width update method of correntropy for channel estimation", IEEE International Conference on Digital Signal Processing (DSP), pp. 916–920.
23. *Chen B., Xing L., Liang J., Zheng N., Principe J. C.,* 2014. "Steady-state mean-square error analysis for adaptive filtering under the maximum correntropy criterion", Signal Process. Lett. IEEE, 21 (7), pp. 880–884.
24. *Ma W., Qua H., Guib G., Li Xu L., Zhao J., Chen B.*, 2015. "Maximum correntropy criterion based sparse adaptive filtering algorithms for robust channel estimation under non-Gaussian environments", J. of the Franklin Institute, 352 (2), pp. 2708–2727.
25. *Guo Y., Ma B, Li Y.*, 2016. "Kernel-Width Adaption Diffusion Maximum Correntropy Algorithm", IEEE Acces, 4, pp. 1–14.
26. *Lu L., Zhao H.*, 2017. "Active impulsive noise control using maximum correntropy with adaptive kernel size", Mechanical Systems and Signal Processing, 87, pp. 180–191.
27. *Qi Y., Wang Y., Zhang J., Zhu J., Zheng X.*, 2014. "Robust Deep Network with Maximum Correntropy Criterion for Seizure Detection", BioMed Research International, Article ID 703816, 10 p.
28. *Huang F., Zhang J., Zhang S.,* 2017. "Adaptive filtering under a variable kernel width maximum correntropy criterion", IEEE Trans. on Circuits and Systems II: Express Briefs, 64 (10), pp. 1247–1251.
29. *Perelman I. I.*, 1982. Operativnaya identifikatsiya obyektov upravleniya, Energoizdat, Moscow, 272 p. (In Russian).
30. *Rudenko O. G., Terenkovskiy I. D., Shtefan A., Oda G. A.*, 1998. "Modifitsirovannyy algoritm tekushchego regressionnogo analiza v zadachakh identifikatsii i prognozirovaniya", Radioelektronika i informatika, 4 (05), pp. 58–61. (In Russian).
31. *Goodwin G., Sin K. S.*, 2014. Adaptive filtering prediction and control. Dover Publications, N.-Y., 560 p.
32. *Streng G.,* 1980. Lineynaya algebra i yeye primeneniya, Mir, Moscow, 454 p. (In Russian).
33. *Karchevskiy Ye. M., Karchevskiy M. M.*, 2018. Lektsii po lineynoy algebre i analiticheskoy geometrii, uchebnoye posobiye, Izdatelstvo Kazanskogo universiteta, Kazan, 426 p. (In Russian).
34. *Doyle J. C., Stein G.*, 1981. "Multivariable Feedback Design : Concepts for a Classical. Modern Synthesis", IEEE Transactions on Automatic Control, 26 (1), pp. 4–16. DOI: https://doi.org/10.1109/TAC.1981.1102555.
35. *Zhao S., Chen B., Principe J. C.*, 2012. "An adaptive kernel width update for correntropy", Proc. of the International Joint Conference on Neural Networks (IJCNN '12), Brisbane, Australia, pp. 1–5.
36. *Jones M. C., Marron J. S., Sheather S. J.*, 1996. "A brief survey of bandwidth selection for density estimation", Journal of the American Statistical Association, 91 (433), pp. 401–407.
37. *Silverman B. W.*, 1986. Density Estimation for Statistics and Data Analysis, 3, CRC Press, New York, NY, USA, 176 p.
38. *Bowman A. W.*, 1984. "An alternative method of cross-validation for the smoothing of density estimates", Biometrika, 71 (2), pp. 353–360.
39. *Scot D. W., Terrell G. R.*, 1987. "Biased and unbiased crossvalidation in density estimation", Journal of the American Statistical Association, 82 (400), pp. 1131–1146.40. *Shi L., Zhao H., Zakharov Y.*, 2018. "An Improved Variable Kernel Width for Maximum Correntropy Criterion Algorithm", IEEE Trans. on Circuits and Systems II: Express Briefs, 5 p.

Received 14.06.2020

ЛІТЕРАТУРА

1. *Widrow B., Hoff M.* Adaptive switching circuits. 1960 IRE WESCON Convention Record. Part 4. New York: Institute of Radio Engineers. 1960. P. 96–104.
2. *Kaczmarz S.* Angen herle Aufl sung von Systemen linearer Gleichungen, Bull. Int. Acad. Polon. Sci. Lett., C 1, Sci. Math. Nat., Ser. A, 1937. S. 355–357.
3. *Либероль Б. Д., Руденко О. Г., Бессонов А. А.* Исследование сходимости одношаговых адаптивных алгоритмов идентификации. Проблемы управления и информатики. 2018. № 5. С. 19–32.
4. *Руденко О. Г., Бессонов А. А.* Регуляризованный алгоритм обучения адалины в задаче оценивания нестационарных параметров. Управляющие системы и машины. 2019. № 1. С. 22-30. DOI: 10.15407/usim.2019.01.022.
5. *Хьюбер П.* Робастность в статистике. М. : Мир, 1984. 304 с.
6. *Hampel F. R., Ronchetti E. M., Rousseeuw P. J., Stahel W. A.* Robust Statistics. The Approach Based on Influence Functions. N.-Y. : John Wiley and Sons, 1986. 526 p.
7. *Руденко О. Г., Бессонов А. А.* Робастное обучение вейвлет-нейросетей, Проблемы управления и информатики. 2010. № 5. С. 66–79.

8. *Rudenko O., Bezsonov O.* Function approximation using robust radial basis function networks, J. of Intelligent Learning Systems and Applications. 2011. № 3. P. 17–25.
9. *Руденко О. Г., Бессонов А. А.* М-обучение радиально-базисных сетей с использованием асимметричных функций влияния. Проблемы управления и информатики. 2012. № 1. С. 79–93.
10. *Walach E., Widrow D.* The least mean fourth (LMF) adaptive algorithm and its family. IEEE Transactions on Information Theory. 1984. № 30 (2). P. 275–283.
11. *Chambers J., Avlonitis A. A* Robust Mixed-Norm Adaptive Filter Algorithm. IEEE Signal Processing Letters. 1997. № 4 (2). P. 46–48.
12. P*apoulis E. V., Stathaki T.* A Normalized Robust Mixed-Norm Adaptive Algorithm for System Identification. IEEE Signal Processing Letters. 2004. № 11 (1). P. 56–59.
13. *Chambers J., Tanrikulu O., Constantinides A. G.* Least mean mixed-norm adaptive filtering. Electronics letters. 1984. № 30 (19). P. 1574–1575.
14. *Zerguine A.* A variable-parameter normalized mixed-norm (VPNMN) adaptive algorithm. EURASIP Journal on Advances in Signal Processing. 2012. № 55. 13 p.
15. *Руденко О. Г., Безсонов О. О., Сердюк Н. М., Олійник К. О. Романюк О. С.* Робастна ідентифікація об'єктів за наявностю негаусівських завад. Бионика интеллекта. 2019. № 2 (93). С. 7–12.
16. *Руденко О. Г., Безсонов О. О., Сердюк Н. М., Олійник К. О. Романюк О. С.* Робастна ідентифікація об'єктів на основі мінімізації комбінованого функціоналу. Системи обробки інформації. 2020. № 1 (160). С. 80–88.
17. *Principe J. C., Xu D., Zhao Q., Fisher J. W.* Learning from examples with information theoretic criteria. J. VLSI Signal Process. Syst. 2000. № 26 (1–2). P. 61–77.
18. *Principe J. C., Xu D., Fisher J.* Information Theoretic Learning / In: S. Haykin (Ed.). Unsupervised Adaptive Filtering. New York: Wiley, 2000. P. 265–319.
19. *Chen B., Hu J., Pu L., Sun Z.* Stochastic gradient algorithm under (h, φ)-entropy criterion. Circuits Syst. Signal Process. 2007. № 26. P. 941–960.
20. *Santamar a I., Pokharel P. P, Jose C. Principe J. C.* Generalized Correlation Function: Definition, Properties, and Application to Blind Equalization. IEEE Trans. on Signal Processing. 2006. № 54 (6). P. 2187–2197.
21. *Liu W., Pokharel P. P, Principe J. C.* Correntropy: Properties and Applications in Non-Gaussian Signal Processing. IEEE Trans. on Signal Processing. 2007. № 1. P. 5286–5298.
22. *Wang W., Zhao J., Qu H., Chen B., Principe J. C.* An adaptive kernel width update method of correntropy for channel estimation. IEEE International Conference on Digital Signal Processing (DSP). 2015. P. 916–920.
23. *Chen B., Xing L., Liang J., Zheng N., Principe J. C.* Steady-state mean-square error analysis for adaptive filtering under the maximum correntropy criterion. Signal Process. Lett. IEEE. 2014. № 21 (7). P. 880–884.
24. *Ma W., Qua H., Guib G., Li Xu L., Zhao J., Chen B.* Maximum correntropy criterion based sparse adaptive filtering algorithms for robust channel estimation under non-Gaussian environments. J. of the Franklin Institute. 2015. № 352 (2). P. 2708–2727.
25. *Guo Y., Ma B, Li Y.* Kernel-Width Adaption Diffusion Maximum Correntropy Algorithm. IEEE Acces. 2016. № 4. P. 1–14.
26. *Lu L., Zhao H.* Active impulsive noise control using maximum correntropy with adaptive kernel size. Mechanical Systems and Signal Processing. 2017. № 87. P. 180–191.
27. *Qi Y., Wang Y., Zhang J., Zhu J., Zheng X.* Robust Deep Network with Maximum Correntropy Criterion for Seizure Detection. BioMed Research International. 2014. Article ID 703816. 10 p.
28. *Huang F., Zhang J., Zhang S.* Adaptive filtering under a variable kernel width maximum correntropy criterion. IEEE Trans. on Circuits and Systems II: Express Briefs. 2017. № 64 (10). P. 1247–1251.
29. *Перельман И. И.* Оперативная идентификация объектов управления. М. : Энергоиздат, 1982. 272 с.
30. *Руденко О. Г., Теренковский И. Д., Штефан А., Ода Г. А.* Модифицированный алгоритм текущего регрессионного анализа в задачах идентификации и прогнозирования. Радиоэлектроника и информатика. 1998. № 4 (05). С. 58–61.
31. *Goodwin G., Sin K. S.* Adaptive filtering prediction and control. N.-Y. : Dover Publications, 2014. 560 p.
32. *Стренг Г.* Линейная алгебра и ее применения. М. : Мир, 1980. 454 с.
33. *Карчевский Е. М., Карчевский М. М.* Лекции по линейной алгебре и аналитической геометрии : учебное пособие. Казань: Изд-во Казанского ун-та, 2018. 426 с.
34. *Doyle J. C., Stein G.* Multivariable Feedback Design : Concepts for a Classical. Modern Synthesis. IEEE Transactions on Automatic Control. 1981. № 26 (1). P. 4–16. DOI: 10.1109/TAC.1981.1102555.

35. *Zhao S., Chen B., Principe J. C.* An adaptive kernel width update for correntropy. Proc. of the International Joint Conference on Neural Networks (IJCNN '12). Brisbane, Australia, 2012. P. 1–5.
36. *Jones M. C., Marron J. S., Sheather S. J.* A brief survey of bandwidth selection for density estimation. Journal of the American Statistical Association. 1996. № 91 (433). P. 401–407.
37. *Silverman B. W.* Density Estimation for Statistics and Data Analysis. № 3. CRC Press, New York, NY, USA, 1986. 176 p.
38. *Bowman A. W.* An alternative method of cross-validation for the smoothing of density estimates. Biometrika. 1984. № 71 (2). P. 353–360.
39. *Scot D. W., Terrell G. R.* Biased and unbiased crossvalidation in density estimation. Journal of the American Statistical Association. 1987. № 82 (400). P. 1131–1146.
40. *Shi L., Zhao H., Zakharov Y.* An Improved Variable Kernel Width for Maximum Correntropy Criterion Algorithm. IEEE Trans. on Circuits and Systems II: Express Briefs. 2018. 5 p.

Надійшла 14.06.2020

О.Г. Руденко, доктор технічних наук, професор, зав. кафедри,
Харківський національний університет радіоелектроніки,
61166, м. Харків, просп. Науки, 14, Україна,
oleh.rudenko@hneu.net
О.О. Безсонов, доктор технічних наук, професор,
Харківський національний університет радіоелектроніки,
61166, м. Харків, пр. Науки, 14, Україна,
oleksandr.bezsonov@hneu.net

РОБАСТНИЙ БАГАТОКРОКОВИЙ АЛГОРИТМ НАВЧАННЯ АДАЛІНИ

Вступ. Адаптивний лінійний елемент (АДАЛІНА) — перша лінійна нейронна мережа, запропонована Уїдроу Б. і Хоффом М.Є., є альтернативою персептрону. Навчання АДАЛІНИ здійснюється за допомогою алгоритму Качмажа рішення систем лінійних алгебраїчних рівнянь. Цей алгоритм є оптимальним в сенсі швидкості збіжності однокроковим алгоритмом в припущеннях про лінійність і гауссовість сигналів, однак при порушенні цих припущень він стає нестійким. Для забезпечення його робастності необхідно використовувати неквадратичні критерії, найбільш поширеними серед яких є комбіновані функціонали, запропоновані Хьюбером і Хемпелем. Однак ефективність їх застосування істотно залежить від численних параметрів, використовуваних в цих умовах і обираємих на основі досвіду дослідника. Результати численних досліджень свідчать про те, що при наявності в вимірах негауссівського, зокрема, імпульсного шуму, досить ефективним є підхід, в основі якого лежать інформаційні характеристики сигналів, а більш відповідним виявляється критерій, що враховує всі статистики сигналу помилки вищого порядку. Таким критерієм є критерій максимуму корентропії. У статті розглянуто багатокроковий алгоритм навчання АДАЛІНИ, що дає більш високу швидкіст збіжності при використанні в якості критерію навчання інформаційного критерію корентропіі, що забезпечує робастність одержуваних оцінок.

Метою статті є дослідження властивостей багатокрокового алгоритму навчання АДАЛІНИ при виборі в якості критерію — критерію максимуму корентропії і розробка рекомендацій щодо його практичного застосування.

Методи дослідження базуються на теорії ідентифікації. На їх основі були досліджені властивості модифікованого багатокрокового алгоритму Качмажа.

Результати. Визначено умови збіжності алгоритму і показано, що в сталому режимі одержувана оцінка є незміщеною. Отримані неасимптотичні і асимптотичні оцінки є досить загальними і залежать від статистичних характеристик сигналів і перешкод.

Висновки. Як показали результати досліджень, використання багатокрокового алгоритму навчання, прискорює процес побудови нейромережевої моделі. Визначено умови збіжності алгоритму при виборі критерію максимуму корентропіі. Показано, що в сталому режимі одержувана оцінка є незміщеною. Відзначено важливість вибору ширини Гауссова ядра, що впливає на швидкість збіжності алгоритмів оцінювання та помилку в сталому режимі, і вказано на доцільність розробки процедур адаптивної корекції ширини ядра.

Ключові слова: *АДАЛІНА, корентропія, метод найменших квадратів, адаптивна корекція ширини ядра, збіжність алгоритму.*

О.Г. Руденко, доктор технических наук, профессор, зав. кафедрой,
Харьковский национальный университет радиоэлектроники,
Харьков, 61166, пр. Науки, 9-А, Украина,
oleg.rudenko@hneu.net
А.А. Бессонов, доктор технических наук, профессор,
Харьковский национальный университет радиоэлектроники,
Харьков, 61166, пр. Науки, 9-А, Украина,
oleksandr.bezsonov@hneu.net

РОБАСТНЫЙ МНОГОШАГОВЫЙ АЛГОРИТМ ОБУЧЕНИЯ АДАЛИНЫ

Введение. Адаптивный линейный элемент (АДАЛИНА) — первая линейная нейронная сеть, предложенная Уидроу Б. и Хоффом М.Е. и являющаяся альтернативой персептрону. Обучение АДАЛИНЫ осуществляется с помощью алгоритма Качмажа решения систем линейных алгебраических уравнений. Этот алгоритм является оптимальным в смысле скорости сходимости одношаговым алгоритмом в предположениях о линейности и гауссовости сигналов, однако при нарушении этих предположений он становится неустойчивым. Для обеспечения его робастности необходимо использовать неквадратичные критерии, наиболее распространенными среди которых являются комбинированные функционалы, предложенные Хьюбером и Хемпелем. Однако эффективность их применения существенно зависит от многочисленных параметров, используемых в этих критериях и выбираемых на основе опыта исследователя. Результаты многочисленных исследований свидетельствуют о том, что при наличии в измерениях негауссовского, в частности, импульсного шума, весьма эффективным является подход, в основе которого лежат информационные характеристики сигналов, а более подходящим оказывается критерий, учиты-вающий все статистики сигнала ошибки более высокого порядка. Таким критерием является критерий максимума коррэнтропии. В статье рассмотрен многошаговый алгоритм обучения АДАЛИНЫ, обладающий существенно более высокой скоростью сходимости при использовании в качестве критерия обучения информационного критерия коррэнтропии, обеспечивающего робастность получаемых оценок.

Целью статьи является исследование свойств многошагового алгоритма обучения АДАЛИНЫ при выборе в качестве критерия — критерий максимума коррэнтропии и разработка рекомендаций по его практическому применению.

Методы исследования базируются на теории идентификации. На их основе были исследованы свойства модифицированного многошагового алгоритма Качмажа.

Результаты. Определены условия сходимости алгоритма и показано, что в установившемся режиме получаемая оценка является несмещенной. Полученные неасимптотические и асимптотические оценки являются достаточно общими и зависят от статистических характеристик сигналов и помех.

Выводы. Как показали результаты исследований, использование многошагового алгоритма обучения, ускоряет процесс построения нейросетевой модели. Определены условия сходимости алгоритма при выборе критерия максимума коррэнтропии. Показано, что в установившемся режиме получаемая оценка является несмещенной. Отмечена важность выбора ширины Гауссова ядра, влияющей на скорость сходимости алгоритмов оценивания и ошибку в установившемся режиме, и показана целесообразность разработки процедур адаптивной коррекции ширины ядра.

Ключевые слова: *АДАЛИНА, коррэнтропия, метод наименьших квадратов, адаптивная коррекция ширины ядра, сходимость алгоритма.*