Nonlinear Dynamics and Neuroscience

Izvestiya Vysshikh Uchebnykh Zavedeniy. Applied Nonlinear Dynamics. 2023;31(4)

Article DOI: 10.18500/0869-6632-003052

Noise influence on recurrent neural network with nonlinear neurons

V. M. Moskvitin, N. I. Semenova✉

Saratov State University, Russia
E-mail: vmmoskvitin@gmail.com, ✉semenovani@info.sgu.ru

Received 6.03.2023, accepted 2.05.2023, available online 12.07.2023, published 31.07.2023

Abstract. The *purpose* of this study is to establish the features of noise propagation and accumulation in a recurrent neural network using a simplified echo network as an example. In this work, we studied the influence of activation function of artificial neurons and the connection matrices between them. *Methods.* We have considered white Gaussian noise sources. We used additive, multiplicative and mixed noise depending on how the noise is introduced into artificial neurons. The noise impact was estimated using the dispersion (variance) of the output signal. *Results.* It is shown that the activation function plays a significant role in noise accumulation. Two nonlinear activation functions have been considered: the hyperbolic tangent and the sigmoid function with range form 0 to 1. It is shown that some types of noise are suppressed in the case of the second function. As a result of considering the influence of coupling matrices, it was found that diagonal coupling matrices with a large blurring coefficient lead to less noise accumulation in the echo network reservoir with an increase in the reservoir memory influence. *Conclusion.* It is shown that activation functions of the form of sigmoid with range from 0 to 1 are suitable for suppressing multiplicative and mixed noise. The accumulation of noise in the reservoir was considered for three types of coupling matrices inside the reservoir: a uniform matrix, a band matrix with a small blurring coefficient, and a band matrix with a large blurring coefficient. It has been found that the band matrix echo networks with a high blurring coefficient accumulates the least noise. This holds for both additive and multiplicative noise.

Keywords: neural networks, recurrent neural networks, echo-state networks, noise influence, white noise, nonlinear activation function.

Acknowledgements. This work was supported by Russian Science Foundation (Project no. 21-72-00002).

For citation: Moskvitin VM, Semenova NI. Noise influence on recurrent neural network with nonlinear neurons. Izvestiya VUZ. Applied Nonlinear Dynamics. 2023;31(4):484–500. DOI: 10.18500/0869-6632-003052

Introduction

At the moment, artificial neural networks (ANNs) are already successfully used in numerous areas of machine learning and solve problems of varying complexity [1]. Such tasks include pattern recognition [2, 3], their classification, improvement of sound recordings, speech recognition [4], prediction of climatic phenomena [5] and much more.

In this paper, recurrent networks are considered using the example of a simplified echo-state network. Such networks are characterized by the fact that some of their neurons have delayed feedback, which allows the network to «remember» its previous states. Such networks are often used to solve forecasting problems [6] or real-time recognition [7].

The increasing complexity of the ANNs topologies and tasks may soon lead to a crisis [8, 9], when the capabilities of modern computers and computing clusters will no longer be enough to meet the growing needs. Here a new direction of ANNs design comes to the rescue - hardware neural networks [10]. According to this approach, neural networks are not created using a computer, but are a real device capable of learning and solving problems. The neurons themselves and the connection between them are implemented at the physical level, that is, the system is not modeled on a computer, but is implemented in hardware according to the corresponding physical principles. In recent years, there has been an exponential increase in work with hardware implementations of ANNs. The greatest efficiency is currently shown by ANNs, which are based on lasers [11], memristors [12], spin-transfer oscillators [13]. The connection between neurons in optical implementations of ANNs is based on the principles of holography [14], diffraction [15, 16], integrated networks of Mach-Zener modulators [17], wavelength division multiplexing [18], optical connections implemented using a 3D printer [19–21].

The basic principle of building an ANNs is the propagation of a signal between neurons using connections with certain coefficients (weights). At the same time, the greatest efficiency and speed can be achieved by parallelizing computing on high-performance computing clusters. However, in this case, the «bottleneck» is the speed of memory access and data processing. Maximum computing performance can be achieved only if the ANNs is implemented completely in hardware. In this case, the problem of accessing memory and mathematical operations on a large amount of data disappears, since each neuron corresponds to a hardware nonlinear component, and each connection is a physical channel. The physical implementation of the ANNs fundamentally changes the features of the influence of noise. In the case of a digital computer implementation of the ANNs, noise can enter the system exclusively with an input signal, whereas in the hardware ANNs there are many internal noise sources with different properties. This article is aimed at investigating the features of the propagation of internal noise in recurrent ANNs, identifying ways to suppress such noise and substantiating the stability of networks to certain types of noise.

In our previous work, we have already considered the effect of noise on deep neural networks [22, 23] and proposed noise suppression strategies in them [24]. This article is a significant complication of the task, since recurrent networks have the property of remembering their previous states, therefore, the accumulation of noise in them should be more difficult than in ANNs that do not depend on the time implementation. A simplified echo-state network will be considered as an example of a recurrent network.

The main research subject of this work is the effect of the activation function and the properties of the echo-state network memory on noise accumulation. For this reason, simplified connection matrix matrices between neurons are considered. The specific type of connection matrices significantly depends on the tasks for which the ANNs was trained, and their statistical features can significantly affect the accumulation of noise. Therefore, the effect of special connection matrix matrices of already trained ANNs on noise accumulation will be the subject of research in subsequent works.

1. The effect of noise on a single neuron

1.1. The basic equations. ANNs consist of artificial neurons, whose role is to transform the input signal linearly or non-linearly. This transformation is implemented using the so-called

activation function. The type of activation function depends on the task at hand and on the effectiveness of the training. For example, piecewise linear functions and a family of functions of the «sigmoids» class are often used in recurrent neural networks. In this article, two functions of this class will be considered: the hyperbolic tangent $f(x) = \tanh \alpha x$ and the sigmoid of the form $f(x) = \frac{1}{1+e^{-\alpha x}}$. In both functions, the multiplier α is responsible for the «steepness» of the nonlinearity. Thus, the output signal of the ith artificial neuron without the influence of noise is determined as follows:

$$x_i^{\text{out}}(t) = f(x_i^{\text{in}}(t)), \tag{1}$$

where $x_i^{\text{out}}(t)$ is the output signal of the ith neuron at time t, and $x_i^{\text{in}}(t)$ is its input signal.

In order to specify which neurons are under noise influence, we will use the noise operator $\hat{\mathbf{N}}$, which will be applied to the output signal without noise influence: $y_i^{\text{out}} = \hat{\mathbf{N}} x_i^{\text{out}}$.

This article discusses additive and multiplicative noise. The peculiarity of additive noise is that it is added to the output signal, and the multiplicative one is multiplied by it:

$$y_i^{\text{out}}(t) = \hat{\mathbf{N}} x_i^{\text{out}}(t) = x_i^{\text{out}}(t) \cdot (1 + \xi_{\text{M}}(t, i)) + \xi_{\text{A}}(t, i), \tag{2}$$

where ξ — these are independent sources of additive (index «A») and multiplicative (index «M») white Gaussian noise with zero mean and corresponding variances σ_{A}^2 and σ_{M}^2. The property of independence of noise sources does not play a role in this section, since only one neuron is considered. However, in the following sections, which study the behavior of the ANNs, this property is important and will influence subsequent conclusions. It is also worth noting that the study of the peculiarities of the influence of correlated noise on the INS and its accumulation is also of great interest, and will be the subject of research in subsequent works.

1.2. Noise impact assessment. In our previous works [22–24], the signal-to-noise ratio (SNR) characteristic was used to estimate the noise level, which was calculated as the ratio of the average signal to its standard deviation. The use of such an interpretation of the SNR has some limitations. In particular, it can only be applied to a positive signal. Since in this article one of the activation functions is a hyperbolic tangent, the output values of which are in the range from -1 to 1, such a formula for calculating SNR can no longer be used. For this reason, this article uses a more universal characteristic, the variance $\sigma^2[\cdot]$, which is a measure of the spread of values of a random variable relative to its mathematical expectation and is calculated as follows:

$$\sigma^2[X] = \text{Var}[X] = \text{E}[(X - \text{E}[X])^2], \tag{3}$$

where X is some discrete random variable, $\text{E}[X]$ is its mathematical expectation (average value). To estimate the level of variance, as a set of X, a set of output values is taken, which are the response of a network or one individual neuron to the same input signal. This interpretation allows us to estimate which range of values is most susceptible to various noise influences. Regardless of the sign of the output signal, the variance will always be positive, and the closer its value is to zero, the less noisy the output signal is. Further in the text, the generally accepted mathematical expectation notation $\text{E}[\cdot]$ and variations $\text{Var}[\cdot]$ are used to estimate the noise level.

The variance of the output signal of one noisy neuron for the same input signal can be calculated as follows:

$$\sigma^2[y^{\text{out}}] = \text{Var}[x^{\text{out}} \cdot (1 + \xi_{\text{M}}) + \xi_{\text{A}}] = \Big(\text{E}[y^{\text{out}}]\Big)^2 \cdot \sigma_{\text{M}}^2 + \sigma_{\text{A}}^2. \tag{4}$$

As can be seen from this equation, the activation function does not affect the dependence of the variance on the average output signal. In the case of additive noise, the variance does not depend on the output signal and is determined only by the variance of the additive noise source. In the case of multiplicative noise, the dependence of the variance of the output signal on its average is

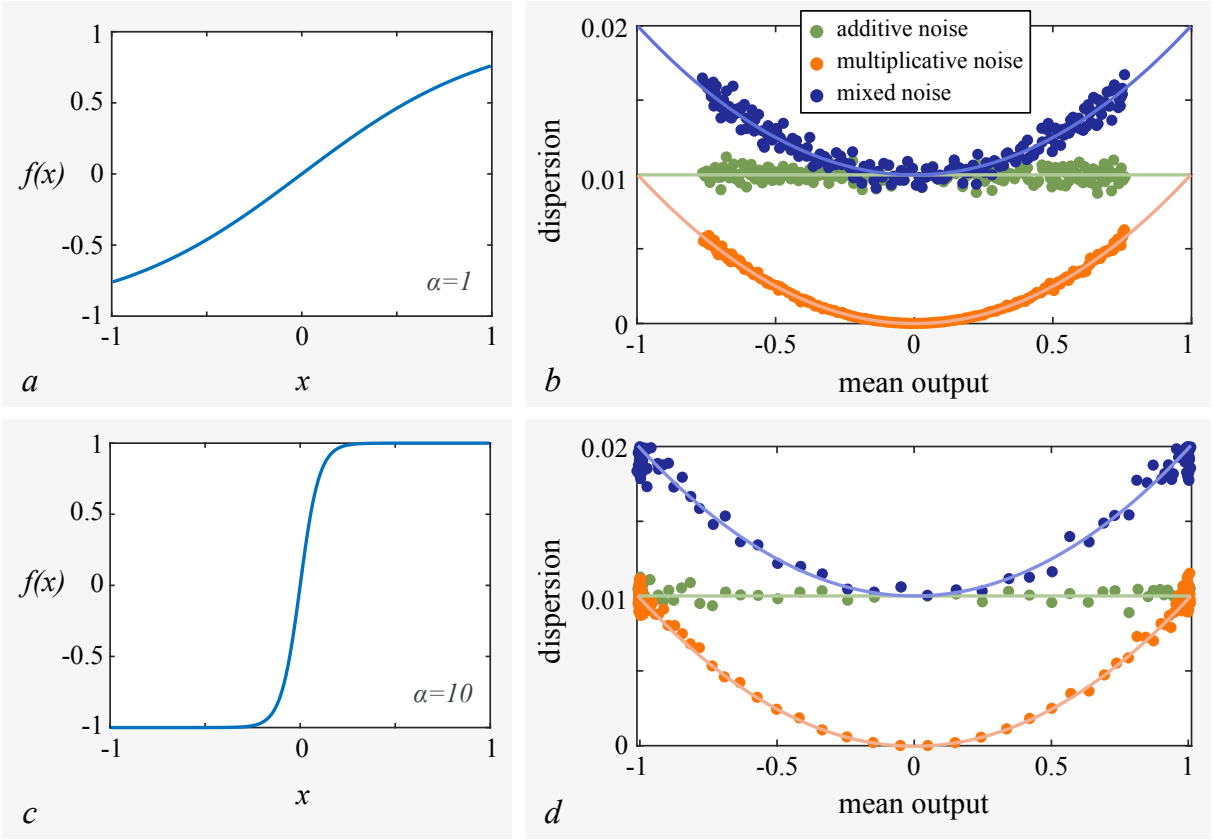

Fig. 1. Noise impact on one neuron with activation function $f(x) = \tanh \alpha x$. Input signal contains random numbers from the range $[-1; 1]$. Nonlinear function with $\alpha = 1$ (a) was used to calculate the dependencies of dispersion on the mean output signal (b) for additive (green points), multiplicative (orange points) and mixed noise (blue points) with dispersions $\sigma_{\text{A}}^2 = \sigma_{\text{M}}^2 = 10^{-2}$. Panels c, d were prepared in the same manner for parameter $\alpha = 10$. Panels b, d contain analytical estimation of the noise level based on Eq. (4) shown by the lines of corresponding colors (color online)

quadratic. To assess the validity of these conclusions, the results of numerical modeling for two activation functions with different steepness of nonlinearity are further considered.

For numerical simulation, the variance of the output signal is calculated as follows. An artificial neuron at each moment of time t receives an input signal $x_i^{\text{in}}(t)$ and outputs an output signal $y_i^{\text{out}}(t)$, including the influence of the activation function and noise. In order to evaluate its statistical characteristics, each input signal $x_i^{\text{in}}(t)$ is repeated $K = 1000$ times. As a result, for each t, we get a set of K values of $y_i^{\text{out}}(k, t)$, which calculates the mean and variance: $\mu[y_i^{\text{out}}(t)]$ and $\sigma^2[y_i^{\text{out}}(t)]$. In the future, the same technique will be used to assess the noise level of the echo-state network output signal.

1.3. One neuron with activation function tanh**.** Let us consider the features of the effect of noise exposure on one neuron with the activation function «hyperbolic tangent». As an input signal, we use $T = 200$ random values from the range from -1 to 1. In Fig. 1, a shows a graph of the function $f(x) = \tanh \alpha x$ when $\alpha = 1$.

In order to evaluate the effect of noise on the neuron, in Fig. 1, b the values of the variance are given depending on the average value of the output signal. The green color shows the dependencies obtained for additive noise with variance $\sigma_{\text{A}}^2 = 10^{-2}$. The orange color shows the dependencies obtained for multiplicative noise with $\sigma_{\text{M}}^2 = 10^{-2}$, and the blue color shows mixed noise combining additive and multiplicative noise. The selected noise characteristics correspond

to the noise observed in the ANN, implemented in the photonic experiment [15]. The lines of the same colors show the dependencies obtained from the equation (4).

Despite the fact that the hyperbolic tangent function can take real values from -1 to 1, in Fig. 1, *b* the range of average output values is much smaller. This is explained by the following feature. If the range of values of the variable x is from -1 to 1 with the value $\alpha = 1$, then this corresponds to the range of values of $f(x)$ from -0.76 to 0.76. Thus, due to the small steepness coefficient α, the range of accepted values can be greatly reduced. Otherwise, the type of variance dependencies themselves, as well as the distribution of points along them, are very similar to what we have already obtained for linear neurons [22]. The dependencies obtained for multiplicative noise combine the levels of variance obtained for additive and multiplicative noise separately.

Let us consider what happens to the output signal of a noisy neuron when the steepness of the nonlinearity increases. In Fig. 1, *c* shows a graph of the hyperbolic tangent at $\alpha = 10$, which is accompanied by a significant change in the steepness of the nonlinearity compared to the graph in Fig. 1, *a*. As can be seen from the graphs in Fig. 1, *d*, the overall level of variance remains approximately the same as for $\alpha = 1$, their slight change is caused by the fact that now $f(x)$ already extends over the entire range of values from -1 to 1. Strictly speaking, it should be the same dependence described by the equation (4). The main feature of Fig. 1, *d* is the focus of the points. If at $\alpha = 1$ the points were uniformly distributed along the entire range of accepted values, now at $\alpha = 10$ the points are focused near the average values of the output signal-1 и 1.

The property of focusing points can be explained as follows. Returning to the graph of Fig. 1, *c*, it can be noted that due to the pronounced nonlinearity and features of the sigmoid class of functions, most negative values of x correspond to the values of $f(x) \approx -1$, and most positive values of x are — values of $f(x) \approx 1$. For this reason, for pronounced nonlinearity, the points begin to focus around two values. Next, we will call them *focusing points*.

1.4. One neuron with the activation function «sigmoid». Let us consider what happens with another nonlinearity -the sigmoid. The corresponding graphs are shown in Fig. 2. Just as for the hyperbolic tangent, two values of the steepness coefficient of nonlinearity are considered here $\alpha = 1$ (Fig. 2, *a, b*) and $\alpha = 10$ (Fig. 2, *c, d*). Comparing the activation functions themselves in Fig. 1 and 2, you can see that in the case of a sigmoid, the range of accepted values changes to $(0; 1)$. This explains that the variance now has a slightly different appearance. However, if we compare these dependencies with what was obtained for the hyperbolic tangent, it can be seen that in the case of the sigmoid we see only the right part of the dependencies obtained for the hyperbolic tangent. The overall noise level remains about the same, and at high α the focusing points appear again. If the focusing points for the hyperbolic tangent were -1 and 1, then for the sigmoid they are 0 and 1.

It is worth noting that if one of the focusing points is zero, then half of the dispersion points will be focused near the value 0. In the case of multiplicative noise (orange dots), it turns out that half of the output values are deprived of noise exposure due to multiplication by 0. Thus, in the case of such noise exposure, it is better to use the activation function of the sigmoid type, since half of the output signal ceases to experience noise exposure. This does not work for additive noise, but for multiplicative and mixed noise, such a strategy can be successfully applied. This is confirmed by the graphs in Fig. 2, *b, d*.

All the above results are confirmed by an analytical assessment of the noise level of the output signal (4).

Let us consider how the variance behaves when changing the steepness parameter of the nonlinearity α and the input signal x^{in} (Fig.3). The color in the figure shows the variation depending on the parameter α and the input signal. In the case of both activation functions, the level of additive noise does not depend on either the input signal or the steepness of the nonlinearity (Fig. 3 *a, d*).

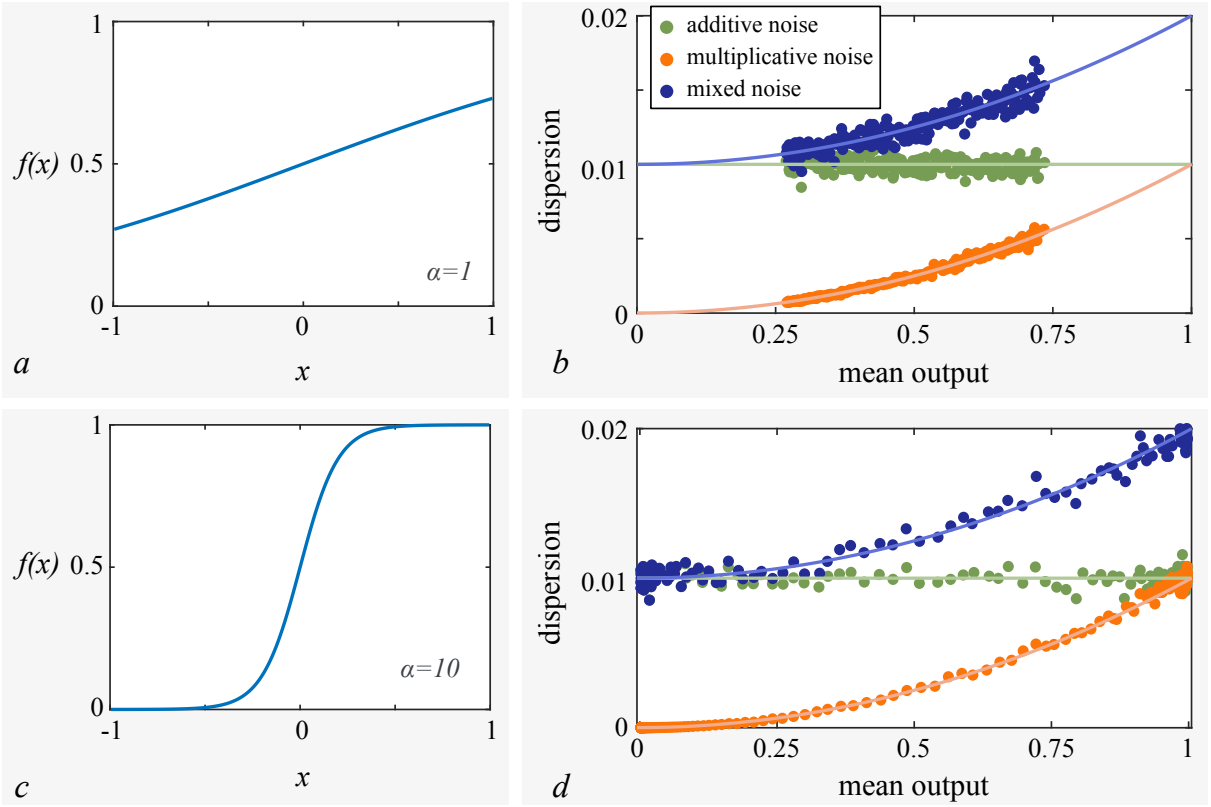

Fig. 2. Noise impact on one neuron with activation function $f(x) = 1/(1 + e^{-\alpha x})$. Input signal contains random numbers from the range $[-1; 1]$. Nonlinear function with $\alpha = 1$ (a) was used to calculate the dependencies of dispersion on the mean output signal (b) for additive (green points), multiplicative (orange points) and mixed noise (blue points) with dispersions $\sigma_A^2 = \sigma_M^2 = 10^{-2}$. Panels c, d were prepared in the same manner for parameter $\alpha = 10$. Panels b, d contain analytical estimation of the noise level based on Eq. (4) shown by the lines of corresponding colors (color online)

As can be seen from Fig. 3,b, in the case of a hyperbolic tangent, the smallest variance can be achieved when working with small values of $|\alpha|$ or small $|x^{\text{in}}|$. However, the activation function is close to linear, and this is not applicable for some tasks. If you increase $|\alpha|$, then the overall level of variance increases. As can be seen from Fig. 3, e, the most effective activation function in terms of multiplicative noise suppression is a nonlinearity of the form $f(x) = 1/(1 + e^{-\alpha x})$. The noise reduction property for small $|\alpha|$ and $|x^{\text{in}}|$ is also valid for this activation function, however, the property of large $|\alpha|$ is added to this, which is that with a high steepness of the nonlinearity, half of the points focuses around 0, while zeroing out the noise effect. This nonlinearity is optimal for multiplicative and mixed noise.

All conclusions drawn for multiplicative noise are also valid for mixed noise (Fig. 3,c, f). The qualitative form of the variance and its dependence on the input signal and the parameter α are the same for both types of noise. Quantitatively, the level of variance is the sum of the variances obtained for additive and multiplicative noise separately.

2. Recurrent neural networks

2.1. The main definitions and components of the network. There is a large number of different types of neural networks. Their topology and the type of artificial neurons significantly

depend on the task being solved. Recurrent networks are usually used to solve forecasting or recognition problems over time. In this paper, as an example of such networks, we will consider the echo state network, the scheme of which is shown in Fig.4.

The echo-state network is characterized by the presence of an input layer responsible for receiving and transforming the input signal. Next, this signal is transmitted to a hidden layer called *reservoir*. The reservoir consists of N neurons, which are connected both to the input neurons that transmit the input signal at time t, and to the same reservoir, receiving the signal of the reservoir at the previous time $(t-1)$. connection matrix with the input layer is carried out using the connection matrix $\mathbf{W}^{\text{in}}$. If the input layer consists of a single neuron, then the size of this matrix is $(1 \times N)$. The coupling matrix inside the reservoir $\mathbf{W}^{\text{res}}$ has a size of $(N \times N)$. Then the state of the neurons at time t can be set using the following equation:

$$\mathbf{x}_t^{\text{res}} = f\left(\beta x_t^{\text{in}} \cdot \mathbf{W}^{\text{in}} + \gamma \mathbf{y}_{t-1}^{\text{res}} \cdot \mathbf{W}^{\text{res}}\right); \qquad \mathbf{y}_t^{\text{res}} = \hat{\mathbf{N}}\mathbf{x}_t^{\text{res}}, \tag{5}$$

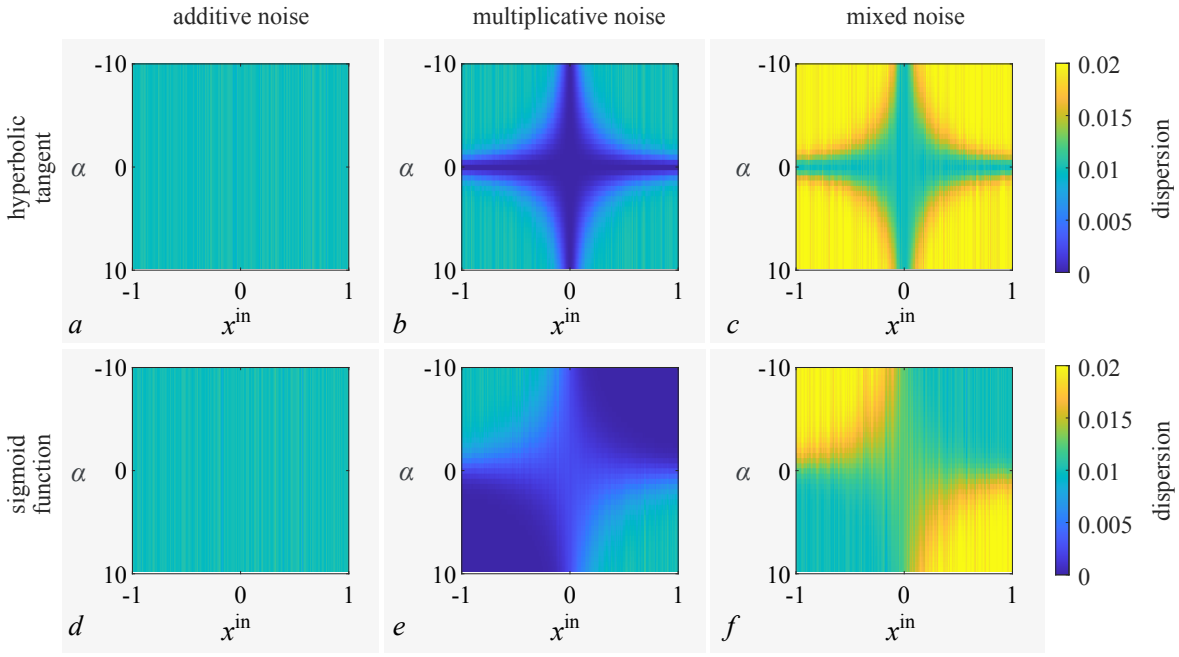

Fig. 3. The influence of input signal x^{in} and parameter α on dispersion of the output signal of one neuron with activation function $f(x) = \tanh \alpha x$ (*a*–*c*) and activation function $f(x) = 1/(1+e^{-\alpha x})$ (*d*–*f*). Noise types: *a*, *d* — additive noise, *b*, *e* — multiplicative noise, *c*, *f* — mixed noise. Other parameters: $\sigma_{\text{A}}^2 = \sigma_{\text{M}}^2 = 10^{-2}$ (color online)

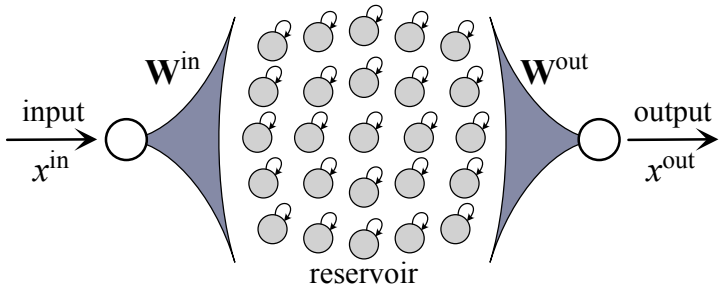

Fig. 4. Schematic representation of considered echo state network (ESN). Input and output neurons without noise impact are shown by white color. The neurons of reservoir with nonlinear activation function receiving noise influence are colored in gray

where the indices t and $t-1$ denote points in time, the function $f(\cdot)$ is an activation function. Bold font indicates $\mathbf{x}^{\text{res}}$ and $\mathbf{y}^{\text{res}}$, which are vectors $(1 \times N)$ describing the state of N neurons of the reservoir. In the future, the value of $N = 100$ will not change. It has been shown that the echo-state network described above makes it possible to predict quasi-periodic and chaotic dynamics [25, 26].

The parameters β and γ are responsible for the contribution of the input signal and the reservoir, respectively. In an already trained network, it is not possible to change the connection matrices $\mathbf{W}^{\text{in}}$ and $\mathbf{W}^{\text{res}}$, and therefore it is impossible to control which has the greatest effect on the state of the reservoir: the input signal or previous reservoir's state . In this paper, the main task is to establish the general features of the noise effect, so it is planned to consider how strongly the memory property affects the output signal by introducing an additional parameter γ. However, if you change only the parameter γ, this will increase or decrease the amplitude of the output signal. In order to preserve the range of output values, the parameter β was added, controlling the input signal contribution, with the condition $\beta + \gamma = 1$. A similar approach was used in the work [27].

In the case of an echo-state network, the number of neurons in the output layer usually coincides with the number of neurons in the input layer. This is done so that the neural network can be closed to itself to solve forecasting problems when the output of the neural network is fed to its own input. Thus, as can be seen from Fig. 4, the output layer consists of a single neuron. The output signal of this neuron, as well as the output signal of the entire network, is given as

$$\mathbf{x}_t^{\text{out}} = \mathbf{y}_t^{\text{res}} \cdot \mathbf{W}^{\text{out}}, \tag{6}$$

where the matrix $\mathbf{W}^{\text{out}}$ of size $(N \times 1)$ is responsible for the connection of the reservoir with the output layer.

In this article, we are primarily interested in the issue of noise accumulation inside the reservoir. Therefore, the noise effect is applied only to the neurons of the reservoir (see (5)). Thus, the input and output neurons are linear without noise exposure, so in Fig. 4 they are indicated in white, and the neurons of the reservoir are gray. The input matrix consists of N elements equal to one, it means that, each neuron of the reservoir receives the same input signal. In order to exclude the influence of connection matrices on statistics, the output matrix will consist of values $1/N$.

In echo-state networks, the connection matrix inside the reservoir is set once and does not change during the learning process. As a rule, only the matrix connecting the reservoir to the output layer [28] changes during the training process. The following section discusses and describes the main types of matrices $\mathbf{W}^{\text{res}}$, which are often used in echo-state networks.

2.2. Connection matrices in reservoir. This article discusses two main types of connection matrices $\mathbf{W}^{\text{res}}$.

One of them is a uniform connection matrix, in which the elements are equal to $1/N$. As already shown in our previous papers [23, 24], from the point of view of noise propagation, this is similar to a matrix consisting of random values with an average value $1/N$.

On the other hand, in recurrent neural networks, the connection matrices inside the reservoir sometimes have a special appearance when all elements are set to zero, except for the elements of the main diagonal and some parallel ones [15, 28]. Such matrices are often called band matrices. With this configuration, the elements of the main diagonal turn out to be larger than the rest of the elements, that is, the largest contribution to the state of the i th neuron of the reservoir is made by the same ith element at the previous time. In order to recreate a similar situation, we will introduce an additional *scattering coefficient* ζ, with which we will set the bandwidth, that is, how many elements to the left and right of the main diagonal are not equal to zero. To compare the noise impact, it is necessary to keep the total range of values

$y_{t-1}^{\text{res}} \cdot W^{\text{res}}$, that is, so that the sums of the elements of each row and column are equal to one. To set a matrix satisfying the conditions described above, we use a Gaussian function. Then the nonzero elements of the diagonal matrix with the scattering coefficient ζ will be given as follows:

$$W_{k,i}^{\text{res}} = \frac{e^{-(k/\zeta^2)}}{\sum_{j=i-\zeta}^{i+\zeta} e^{-(j/\zeta^2)}}, \quad k \in [i-\zeta; i+\zeta]. \tag{7}$$

Examples of matrices defined in this way are shown in Fig. 5 for the scattering coefficients $\zeta = 2$ (a) and $\zeta = 20$ (b) corresponding to the band width of the ribbon matrix 4 и 40.

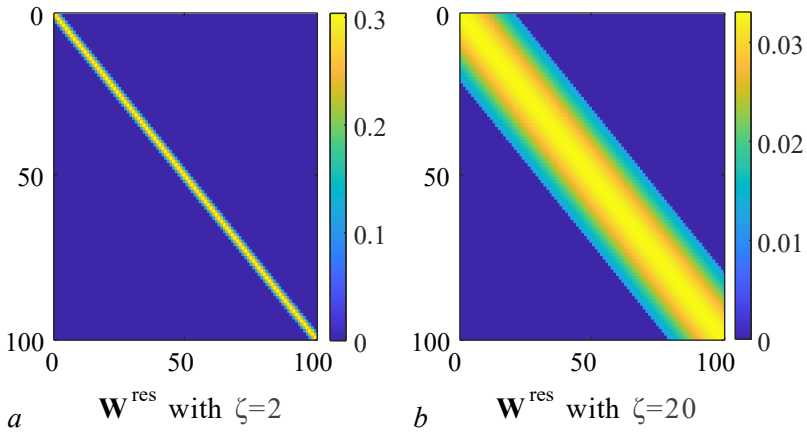

Fig. 5. Examples of considered band matrices obtained using (7) with blurring coefficient $\zeta = 2$ (a) and $\zeta = 20$ (b). These values correspond to bandwidth of a band matrix equal to 4 and 40 (color online)

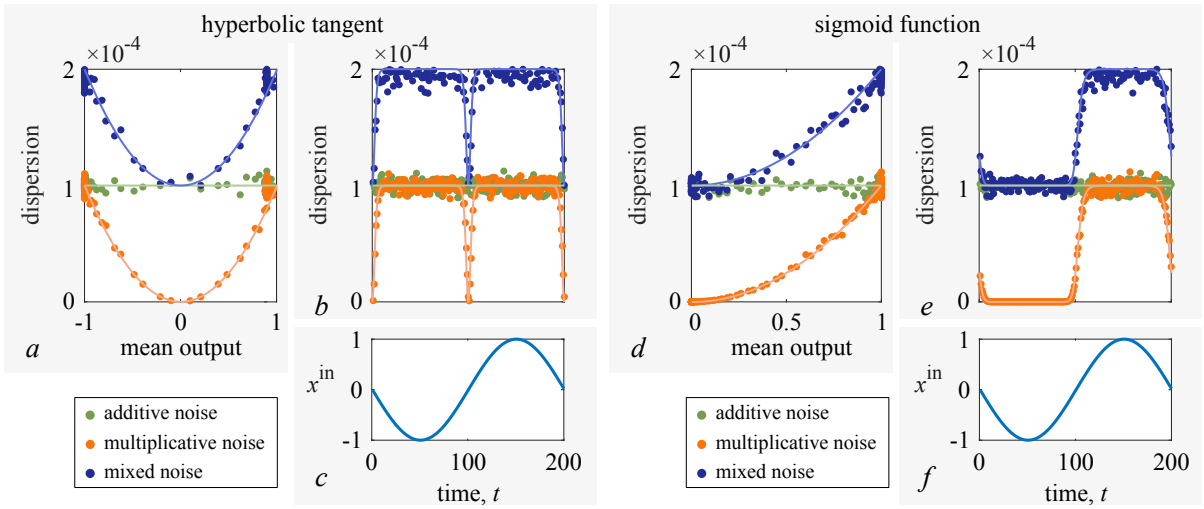

Fig. 6. Dependencies of dispersion of the output ESN signal on mean output signal (a, d), and how this dispersion (b, e) and input signal (c, f) changes in time in ESN with uniform connection matrix $\mathbf{W}^{\text{res}}$. Panels a–c were prepared for ESN with activation function $f(x) = \tanh \alpha x$, while panels d–f correspond to function $f(x) = 1/(1+e^{-\alpha x})$. Parameters: $\gamma = 0$, $\alpha = 10$, $\sigma_{\text{A}}^2 = \sigma_{\text{M}}^2 = 10^{-2}$. Panels a, b, d, e contain analytical estimation of the noise level based on Eq. (4) shown by the lines of corresponding colors (color online)

3. The influence of the connection matrix $\mathbf{W}^{\text{out}}$

Let's consider how the noise effect inside the reservoir affects the statistical characteristics of the echo-state network output signal. Previously, one neuron was considered, and variance was used to evaluate its output signal.

For example, in Fig. 6, *a* shows the dependence of the variance of the output signal of the echo-state network x_t^{out} on the average output value. At the same time, using the same data, a graph of the dependence of variance on time was constructed (Fig. 6, *b*) for an echo-state network in which the input signal changes over time according to the graph Fig. 6, *c*. These results were obtained for a reservoir with neurons in which the hyperbolic tangent acts as an activation function. The graphs of Fig. 6, *d–f* for the activation function $f(x) = 1/(1 + e^{-\alpha x})$.

In order to understand how the memory property of the reservoir affects the final result, first consider the case when $\gamma = 0$. This corresponds to a situation where the neurons of the reservoir are deprived of the memory property, and receive only an input signal at time t. These results are shown in Fig. 6 for both activation functions. Comparing this graph with the results obtained for a single neuron, it can be seen that the forms of dependencies coincide, but quantitatively the order of variance for the network changes from 10^{-2} to 10^{-4}. The drop in the total level of variance to 10^{-4} is caused by the output matrix of the connection with the elements $1/N$. The output signal is set as follows:

$$x^{\text{out}} = \sum_{i=1}^{N} y_i^{\text{res}} \cdot W_i^{\text{out}} = \frac{1}{N} \sum_{i=1}^{N} y_i^{\text{res}}. \tag{8}$$

According to the rules for calculating variation, the variation of the sum of uncorrelated independent random variables is calculated as the sum of the variations of these quantities: $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$. If a random variable is multiplied by some constant, then the resulting variation is calculated as the product of the square of this constant and the variation of the random variable: $\text{Var}[c \cdot X] = c^2 \cdot \text{Var}[X]$. Then if additive noise acts on the reservoir, the variation of the output signal is calculated as

$$\text{Var}[x^{\text{out}}] = \frac{1}{N^2} \sum_{i=1}^{N} \text{Var}[y_i^{\text{res}}] = \frac{1}{N} \cdot \sigma_{\text{A}}^2. \tag{9}$$

For one isolated neuron, the variation was σ_{A}^2. Thus, the variation of the output signal decreases by a factor of $N = 100$. Similar conclusions can be drawn for multiplicative noise.

The specific type of the connection matrix $\mathbf{W}^{\text{out}}$ directly depends on the problem being solved and the learning process. However, from the point of view of noise propagation, this matrix can be considered as a set of random values with some mean value and a mean squared value. Despite the simplification, we successfully applied this approach to the deep network [23] and showed its effectiveness for a trained network. In general, the multiplier $1/N$ cannot be extended beyond the sum in the formula (9), and then the variance of the output signal can be calculated as follows:

$$\begin{gathered} \text{Var}[x^{\text{out}}] = \sum_{i=1}^{N} \text{Var}[y_i^{\text{res}} \cdot W_i^{\text{out}}] = \sum_{i=1}^{N} (W_i^{\text{out}})^2 \cdot \text{Var}[y_i^{\text{res}}] = \\ \sum_{i=1}^{N} (W_i^{\text{out}})^2 \cdot \left(\sigma_{\text{A}}^2 + \sigma_{\text{M}}^2 \cdot (\text{E}[y_i^{\text{res}}])^2\right) = \sigma_{\text{A}}^2 \cdot N\eta^2(\mathbf{W}^{\text{out}}) + \sigma_{\text{M}}^2 \cdot \sum_{i=1}^{N} (W_i^{\text{out}})^2 \left(\text{E}[y_i^{\text{res}}]\right)^2, \end{gathered} \tag{10}$$

where $\eta^2(\mathbf{W}^{\text{out}})$ — the average square value of the matrix $\mathbf{W}^{\text{out}}$, which is calculated as $\eta^2(\mathbf{W}^{\text{out}}) = \frac{1}{N} \sum_{i=1}^{N} (W_i^{\text{out}})^2$. Thus, if only additive noise acts on the reservoir, the variance of the output signal generally looks like $N \cdot \eta^2(\mathbf{W}^{\text{out}}) \cdot \sigma_{\text{A}}^2$. If the matrix $\mathbf{W}^{\text{out}}$ consists of elements $1/N$, then this value can be reduced to the formula (9).

In the following sections, the influence of the *inside* coupling matrices on the obtained results is mainly considered. In order for the output coupling matrix $\mathbf{W}^{\text{out}}$ not to distort the results obtained, we will set it homogeneous and consisting of elements $1/N$.

4. Noise propagation in an echo-state network with a homogeneous matrix $\mathbf{W}^{\text{res}}$

This section also discusses the effect of the parameter γ on noise accumulation. This γ parameter is responsible for the contribution of the previous reservoir states to the new output signal. In fact, this parameter is responsible for the «memory» of the echo-state network. This suggests that it is no longer sufficient to use a random sequence of numbers as an input signal, since the shape of the input signal can also be of great importance. For this reason, we will now show the dependence of the variance not only on the average output signal, but also on time t.

Let's consider what happens to the variance if the neurons of the reservoir receive both the signal from the input neuron and the previous signal of the reservoir at the same time. In Fig. 7 the results of the variance calculation for the case $\gamma = 0.5$ are given.for both types of activation function. From the comparison of Fig. 6 and 7 it can be seen that the general appearance of dependencies is changing.

In the case of $\gamma = 0$, the type of dependence of the variance on the average output signal coincides with the corresponding graphs for one neuron. As for the dependencies of variance on time, they can be characterized as follows. Additive noise leads to a constant level of dispersion regardless of time for both activation functions, and the effect of multiplicative noise depends significantly on the input signal. The variance is at the same level as for additive noise if the input signal corresponds to an output signal close to 1 or -1. For example, in the case of a hyperbolic tangent (see Fig. 6, *b*) the left half of the orange dependence corresponds to a negative input signal (see Fig. 6, *c*), and the right one is positive. Similarly for sigmoids: positive values of the input signal lead to the same level of variance (see Fig. 6, *e*). If the values of the input signal lead to a zero output signal, then the level of dispersion of the multiplicative noise drops to zero. For the selected input waveform in combination with the activation function tanh (see Fig. 6, *b*) this leads to a sharp drop in the level of variance near the values of $t = 0$, 100, 200, when the input signal is zero. In the case of a sigmoid, the zero output signal corresponds not to one

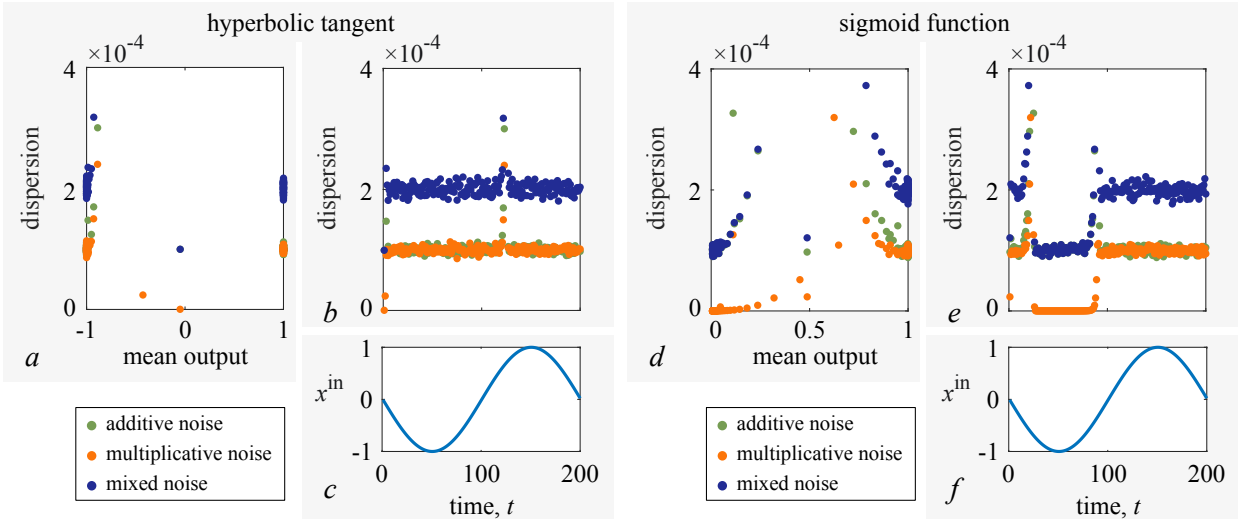

Fig. 7. Dispersion of the output ESN signal depending on its mean and time. This figure was prepared in the same manner as Fig. 6 but for parameter $\gamma = 0.5$ (color online)

value of the input signal, but to a whole range of negative values of the input signal, therefore, the variance for multiplicative noise in this case becomes close to zero for the time $t \in [0; 100]$ (see Fig. 6, *e*). In addition, both nonlinearities, as well as for one neuron, are characterized by focusing points.

In the case of $\gamma = 0.5$, this focusing property is more pronounced (Fig. 7, *a*, *d*). An increase in the echo-state network leads to an increase in the dispersion of intermediate points. For this reason, despite the fact that most of the dependency points are at the same level as for $\gamma = 0$, the average level of variance increases with an increase in γ. In the case of the activation function «sigmoid», the number of points corresponding to the zero level of variance decreases with an increase in the parameter γ, and at $\gamma \to 1$ all points of dependence of the variance become one level, as it was for the hyperbolic tangent. The drop in variance in the case of multiplicative noise to almost zero is explained by the fact that for part of the negative input signal, the output signal is zero, as it was for both one neuron and $\gamma = 0$.

5. Noise propagation in a band-matrix echo-state network $\mathbf{W}^{\text{res}}$

If an echo-state network with a diagonal matrix (band matrix) is considered, the variance dependences on the average output signal and time remain qualitatively the same as for a homogeneous matrix. There are only minor quantitative differences.

In the case of a hyperbolic tangent, the variance is at approximately the same level 10^{-4} regardless of the input signal for additive and multiplicative noise. In the case of sigmoid and additive noise, the variance is approximately at the same level. If the activation function is set as a sigmoid and the reservoir neurons are affected by multiplicative or mixed noise, then the dispersion values are characterized by a drop to almost zero with a negative input signal. As the parameter γ increases, the number of such points decreases, and at $\gamma \to 1$, the points of dependence of the variance also tend to the same level. Thus, all conclusions about the qualitative type of dependencies coincide with the conclusions described in the section 4. Therefore, next we will consider the quantitative aspects.

Let's consider what happens to the average level of variance when the parameter γ is changed. In Fig. 8 shows the dependence of the variance of the output signal on the degree of influence of the delayed feedback of the reservoir γ for three coupling matrices $\mathbf{W}^{\text{res}}$: a homogeneous coupling matrix considered in the previous section (Fig. 8, *a*, *b*), and diagonal matrices with scattering coefficients $\zeta = 2$ (Fig. 8, *c*, *d*) and $\zeta = 20$ (Fig.8, *e*, *f*). These dependencies were constructed for both types of activation function: hyperbolic tangent (upper fragments) and sigmoid (lower fragments).

The parameter γ is responsible for the contribution of the reservoir's memory. At $\gamma > 0.5$, the input signal contribution decreases ($\beta < 0.5$), and a larger contribution comes from the reservoir with noisy neurons at the previous time. Since the noise sources are independent, this leads to a significant accumulation of noise. For this reason, at $\gamma < 0.5$, the total noise level remains approximately at the same level, and at $\gamma > 0.5$ it starts to grow sharply.

As can be seen from Fig. 8, at $\gamma \to 0.6$, there is an increase in all dependencies. This is due to the fact that there are some outliers on the time dependence of the variance (see Fig. 7). Their number and amplitude increase with the growth of γ, which affects the calculation of the average level of variance. Strictly speaking, regardless of the parameter γ, most points of dependence of the variance on time belong to the 10^{-4} level for additive and multiplicative noise with an activation function defined as a hyperbolic tangent. The same situation is observed if the activation function is set as a sigmoid and only additive noises are present in the reservoir. If there are multiplicative noises in the system, then it is more correct to use the sigmoid as an activation function, since the average level of dispersion of the output signal for it is lower (blue

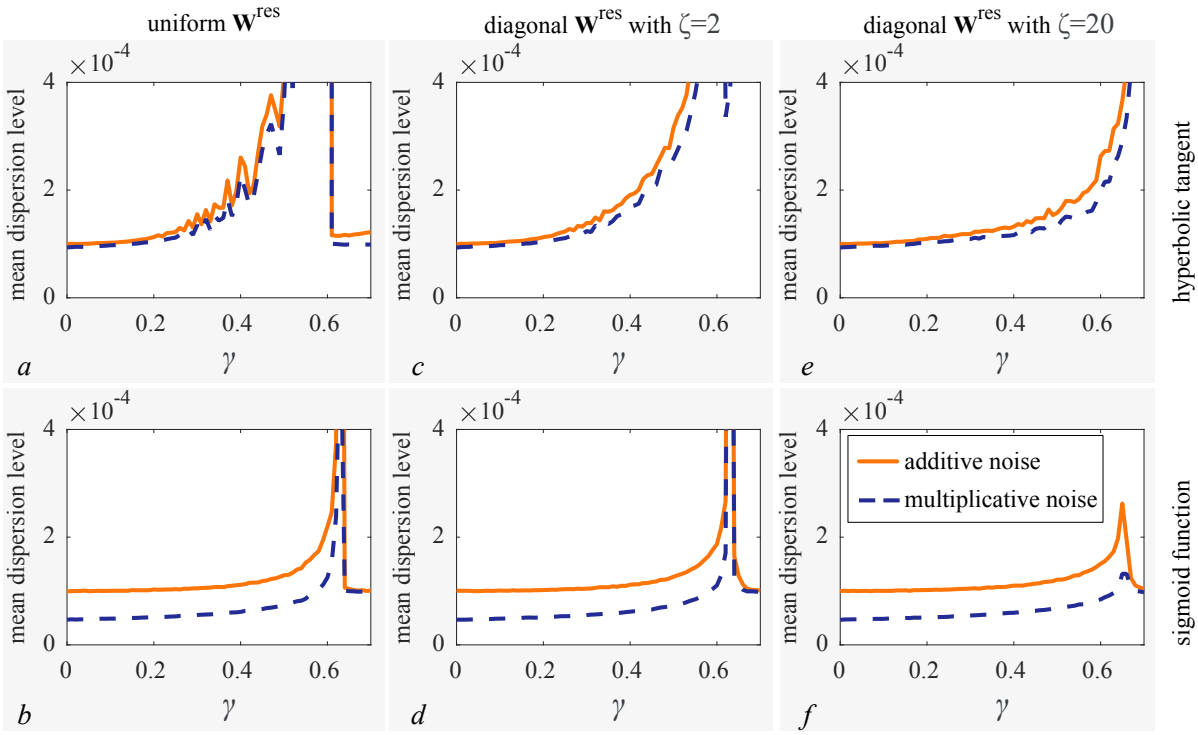

Fig. 8. Dependencies of mean dispersion level on parameter γ. Dispersion is calculated by the output signal of ESN with uniform matrix $\mathbf{W}^{res}$ (*a, b*), band matrix $\mathbf{W}^{res}$ with $\zeta = 2$ (*c, d*), band matrix $\mathbf{W}^{res}$ with $\zeta = 20$ (*e, f*). Top panels were prepared for activation function $f(x) = \tanh \alpha x$, while bottom panels correspond to function $f(x) = 1/(1 + e^{-\alpha x})$. Parameters: $\alpha = 10$, $\sigma_A^2 = \sigma_M^2 = 10^{-2}$ (color online)

graphs in Fig. 8, *b, d, f*). Comparing the dependencies shown in Fig. 8, it can also be concluded that the accumulation of outliers and an increase in the average level of dispersion occurs more slowly if the reservoir matrix $\mathbf{W}^{res}$ is set as a diagonal matrix with a higher scattering coefficient $\zeta = 20$ than with ζ =2 or for a homogeneous matrix.

For some fragments, Fig. 8 is characterized by a sharp drop in the average level of variance at $\gamma > 0.6$. Because of this, you may get the wrong impression that the optimal choice is $\gamma > 0.6$. However, this effect is caused by the fact that the input signal is almost completely replaced in the system. In this regard, such a coefficient γ is rarely used.

Conclusion

In this paper, the influence of noise on recurrent networks is considered using the example of an echo-state network, as well as the influence of the nonlinearity of the activation function and types of connection matrices on the propagation and accumulation of additive, multiplicative and mixed noise.

Using the example of the function sigmoids of the form $f(x) = 1/(1 + e^{-\alpha x})$ with one of the focusing points near zero, we have shown that such activation functions are well suited for suppressing multiplicative and mixed noise. The effect of additive noise and the dispersion of the output signal are independent of the input signal and the type of activation function.

In addition, the accumulation of noise in the echo-state network reservoir was considered for three types of connection matrices inside the reservoir: a homogeneous matrix, a band matrix with a low scattering coefficient and a band matrix with a high scattering coefficient. In order to show the effect of noise in its pure form, simplified models of the main types of connection

matrices inside the reservoir were considered.

The considered homogeneous matrix corresponds to a matrix given randomly with some average value. Another type of matrices that are often used in recurrent neural networks are ribbon matrices in which all elements except the elements of the main diagonal and some elements parallel to them are zero. Similar matrices were recreated in this article when the elements of the matrix were set using a Gaussian function with a scattering coefficient ζ, which specifies how many nonzero elements to the left and right of the main diagonal are not zero. We found that echo-state networks with a diagonal connection matrix with a high scattering coefficient accumulate less noise with an increase in the influence of network memory controlled by the parameter γ.

References

1. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–444. DOI: 10.1038/nature14539.
2. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun. ACM. 2017;60(6):84–90. DOI: 10.1145/3065386.
3. Maturana D, Scherer S. VoxNet: A 3D Convolutional Neural Network for real-time object recognition. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 28 September 2015 – 02 October 2015, Hamburg, Germany. New York: IEEE; 2015. P. 922–928. DOI: 10.1109/IROS.2015.7353481.
4. Graves A, Mohamed AR, Hinton G. Speech recognition with deep recurrent neural networks. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. 26–31 May 2013, Vancouver, BC, Canada. New York: IEEE; 2013. P. 6645–6649. DOI: 10.1109/ICASSP.2013.6638947.
5. Kar S, Moura JMF. Distributed consensus algorithms in sensor networks with imperfect communication: Link failures and channel noise. IEEE Transactions on Signal Processing. 2009;57(1):355–369. DOI: 10.1109/TSP.2008.2007111.
6. Mandic DP, Chambers JA. Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures and Stability. New York: Wiley; 2001. 304 p. DOI: 10.1002/047084535X.
7. Bailador G, Roggen D, Tröster G, Triviño G. Real time gesture recognition using continuous time recurrent neural networks. In: 2nd International ICST Conference on Body Area Networks. 11th–13th Jun 2007, Florence, Italy. ICST; 2007. 8 p. DOI: 10.4108/bodynets.2007.149.
8. Hasler J, Marr H. Finding a roadmap to achieve large neuromorphic hardware systems. Frontiers in Neuroscience. 2013;7:118. DOI: 10.3389/fnins.2013.00118.
9. Gupta S, Agrawal A, Gopalakrishnan K, Narayanan P. Deep learning with limited numerical precision. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37. 6-11 July 2015, Lille, France. JMLR; 2015. P. 1737–1746.
10. Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. Physics-informed machine learning. Nature Reviews Physics. 2021;3(6):422–440. DOI: 10.1038/s42254-021-00314-5.
11. Brunner D, Soriano MC, Mirasso CR, Fischer I. Parallel photonic information processing at gigabyte per second data rates using transient states. Nature connection matrixs. 2013;4(1):1364. DOI: 10.1038/ncomms2368.
12. Tuma T, Pantazi A, Le Gallo M, Sebastian A, Eleftheriou E. Stochastic phase-change neurons. Nature Nanotechnology. 2016;11(8):693–699. DOI: 10.1038/nnano.2016.70.
13. Torrejon J, Riou M, Araujo FA, Tsunegi S, Khalsa G, Querlioz D, Bortolotti P, Cros V, Yakushiji K, Fukushima A, Kubota H, Yuasa S, Stiles MD, Grollier J. Neuromorphic computing with nanoscale spintronic oscillators. Nature. 2017;547(7664):428–431. DOI: 10.1038/nature23011.

14. Psaltis D, Brady D, Gu XG, Lin S. Holography in artificial neural networks. Nature. 1990;343 (6256):325–330. DOI: 10.1038/343325a0.
15. Bueno J, Maktoobi S, Froehly L, Fischer I, Jacquot M, Larger L, Brunner D. Reinforcement learning in a large-scale photonic recurrent neural network. Optica. 2018;5(6):756–760. DOI: 10.1364/ OPTICA.5.000756.
16. Lin X, Rivenson Y, Yardimci NT, Veli M, Luo Y, Jarrahi M, Ozcan A. All-optical machine learning using diffractive deep neural networks. Science. 2018;361(6406):1004–1008. DOI: 10.1126/science.aat8084.
17. Shen Y, Harris NC, Skirlo S, Prabhu M, Baehr-Jones T, Hochberg M, Sun X, Zhao S, Larochelle H, Englund D, Soljačić M. Deep learning with coherent nanophotonic circuits. Nature Photonics. 2017;11(93):441–446. DOI: 10.1038/nphoton.2017.93.
18. Tait AN, de Lima TF, Zhou E, Wu AX, Nahmias MA, Shastri BJ, Prucnal PR. Neuromorphic photonic networks using silicon photonic weight banks. Scientific Reports. 2017;7(1):7430. DOI: 10.1038/s41598-017-07754-z.
19. Moughames J, Porte X, Thiel M, Ulliac G, Larger L, Jacquot M, Kadic M, Brunner D. Three-dimensional waveguide interconnects for scalable integration of photonic neural networks. Optica. 2020;7(6):640–646. DOI: 10.1364/OPTICA.388205.
20. Dinc NU, Psaltis D, Brunner D. Optical neural networks: The 3D connection. Photoniques. 2020;(104):34–38. DOI: 10.1051/photon/202010434.
21. Moughames J, Porte X, Larger L, Jacquot M, Kadic M, Brunner D. 3D printed multimode-splitters for photonic interconnects. Opt. Mater. Express. 2020;10(11):2952–2961. DOI: 10.1364/ OME.402974.
22. Semenova N, Porte X, Andreoli L, Jacquot M, Larger L, Brunner D. Fundamental aspects of noise in analog-hardware neural networks. Chaos: An Interdisciplinary Journal of Nonlinear Science. 2019;29(10):103128. DOI: 10.1063/1.5120824.
23. Semenova N, Larger L, Brunner D. Understanding and mitigating noise in trained deep neural networks. Neural Networks. 2022;146:151–160. DOI: 10.1016/j.neunet.2021.11.008.
24. Semenova N, Brunner D. Noise-mitigation strategies in physical feedforward neural networks. Chaos: An Interdisciplinary Journal of Nonlinear Science. 2022;32(6):061106. DOI: 10.1063/ 5.0096637.
25. Jaeger H. Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the "echo state network" approach. GMD-Report 159. Bonn: German National Research Center for Information Technology; 2002. 48 p.
26. Prokhorov D. Echo state networks: appeal and challenges. In: Proceedings. 2005 IEEE International Joint Conference on Neural Networks. Vol. 3. 31 July 2005 – 04 August 2005, Montreal, QC, Canada. New York: IEEE; 2005. P. 1463–1466. DOI: 10.1109/IJCNN.2005.1556091.
27. Cerina L, Santambrogio MD, Franco G, Gallicchio C, Micheli A. EchoBay: Design and optimization of echo state networks under memory and time constraints. ACM Transactions on Architecture and Code Optimization. 2020;17(3):1–24. DOI: 10.1145/3404993.
28. Lukoševičius M, Jaeger H. Reservoir computing approaches to recurrent neural network training. Computer Science Review. 2009;3(3):127–149. DOI: 10.1016/j.cosrev.2009.03.005.