ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Error Entropy Criterion in Echo State Network Training

Levy Boccato[1], Daniel G. Silva[1], Denis Fantinato[1], Kenji Nose Filho[1], Rafael Ferrari[1], Romis Attux[1], Aline Neves[2], Jugurta Montalvão[3] and João Marcos Travassos Romano[1] *

1- School of Electrical and Computer Engineering - University of Campinas
Av. Albert Einstein, 400, 13083-852, Campinas, São Paulo - Brazil

2- CECS - Federal University of ABC
Av. dos Estados, 5001, 09210-170, Santo André, São Paulo - Brazil

3- Department of Electrical Engineering - Federal University of Sergipe
Av. Marechal Rondon, 49100-000, São Cristóvão, Sergipe - Brazil

Abstract.
Echo state networks offer a promising possibility for an effective use of recurrent structures as the presence of feedback is accompanied with a relatively simple training process. However, such simplicity, which is obtained through the use of an adaptive linear readout that minimizes the mean-squared error, limits the capability of exploring the statistical information of the involved signals. In this work, we apply an information-theoretic learning framework, based on the error entropy criterion, to the ESN training, in order to improve the performance of the neural model, whose advantages are analyzed in the context of supervised channel equalization problem.

1 Introduction

From a signal processing standpoint, echo state networks (ESNs) can be seen as attractive possibilities for adaptive filtering, as they ally, to a certain degree, the benefits of classical recurrent neural networks (RNNs) with respect to the emergence of a dynamical memory to an adaptation complexity equivalent to that of linear finite impulse response (FIR) filters [1] [2]. This trade-off is accomplished by using an intermediate recurrent layer, named dynamical reservoir, which is not subject to adaptation. Hence, the network training process is significantly simplified, amounting to the task of determining the coefficients of the linear combiner at the output that minimize the mean-squared error (MSE), which can be solved in an online fashion with the aid of well-known methods like the least mean squares (LMS) and recursive least squares (RLS) algorithms [3].

Nevertheless, canonical ESNs cannot make full use of the statistical information associated with both the reservoir and desired signals. This limitation is due to two main aspects: (1) the linear character of the readout structure and (2) the adoption of the MSE criterion. While the first factor has motivated the proposal of alternative readout structures [4], the possibility of using criteria other than the MSE in ESN training has not been addressed so far.

*This work was sponsored by grants from CAPES, CNPq and FAPESP (2010/51027-8).

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Interestingly, the study of adaptive filtering criteria and algorithms capable of employing in a more extensive manner the statistical information contained in the input and reference signals constitutes the main motivation underlying the research field known as information-theoretic learning (ITL) [5]. Within this framework, the error entropy criterion (EEC) deserves special attention as it brings to fruition the statistical completeness of ITL by taking into account the probability density function (PDF) of the error signal, instead of only the second-order statistics, as occurs with the MSE, along with efficient online learning algorithms, such as the stochastic information gradient for minimum error entropy (MEE-SIG) [5].

In this work, we propose the use of the entropy error criterion in lieu of the MSE for training the ESN readout. This idea shall be analyzed in the context of an emblematic information retrieval problem - supervised channel equalization -, which poses a crucial demand for an effective trade-off between reachable performance and operational tractability, characteristics that suit adequately the spirit of ESNs.

2 Echo State Networks

The basic ESN architecture is composed of three layers of neurons: (i) the input layer, which receives the stimuli $\mathbf{u}(n) \in \mathcal{R}^{K\times 1}$ and transmits them to the internal neurons by means of linear combinations, whose coefficients are specified in matrix $\mathbf{W}^{in} \in \mathcal{R}^{N\times K}$; (ii) the internal layer, called dynamical reservoir, whose states, represented by $\mathbf{x}(n) \in \mathcal{R}^{N\times 1}$, are determined as follows:

$$\mathbf{x}(n+1) = \mathbf{f}\left(\mathbf{W}^{in}\mathbf{u}(n+1) + \mathbf{W}\mathbf{x}(n)\right), \tag{1}$$

where $\mathbf{W} \in \mathcal{R}^{N\times N}$ contains the synaptic weights of the recurrent connections within the reservoir and $\mathbf{f}(\cdot) = (f_1(\cdot), \ldots, f_N(\cdot))$ denotes the activation functions of the internal units, and (iii) the output layer, called readout, which combines the reservoir signals to produce the network outputs according to:

$$\mathbf{y}(n+1) = \mathbf{W}^{out}\mathbf{x}(n+1), \tag{2}$$

where $\mathbf{W}^{out} \in \mathcal{R}^{L\times N}$ brings the coefficients of the output linear combiner [1].

A common strategy for the reservoir design is to randomly create a sparse reservoir weight matrix $\mathbf{W}$, which is then globally scaled with the aim of controlling the spectral radius, i.e., the largest absolute eigenvalue of the weight matrix [1].

3 Information-Theoretic Learning

The research field known as information-theoretic learning (ITL) provides a set of alternative adaptation criteria capable of reaching a more effective extraction of the statistical information available by resorting to fundamental concepts borrowed from information theory, like entropy and mutual information [5]. Particularly, the error entropy criterion (EEC), which aims to minimize the uncertainty

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

associated with the error signal, emerges as an interesting alternative due to its capability of dealing with non-Gaussian, fat-tail distributions and with the occurrence of outliers, as well as to the availability of online algorithms, such as the stochastic information gradient for minimum error entropy (MEE-SIG) [5].

In this context, Rényi's definition of entropy, especially its quadratic version $\hat{H}_2(\cdot)$, is particularly useful in view of the possibility of obtaining nonparametric estimators with the aid of kernel density estimation methods, such as Parzen windowing, to approximate the probability density function (PDF) of the error signal [5]. The online algorithm MEE-SIG computes the instantaneous stochastic gradient similarly to the LMS in the case of the MSE criterion. Hence, by employing Gaussian kernels, the cost function associated with the EEC can be stated as follows:

$$\min_{\mathbf{w}} \hat{H}_2(e(n)) \approx \min_{\mathbf{w}} \left\{ -\log \left(\frac{1}{L} \sum_{i=n-L}^{n-1} G_\sigma(e_n - e_i) \right) \right\}, \tag{3}$$

where $e(n)$ denotes the error signal, e_i is the error sample at time instant i, $G_\sigma(\cdot)$ is the Gaussian kernel function, σ denotes the kernel size and L is the number of samples (time window) used to estimate the error PDF. This method shall be used to train the ESN readouts in the context of the channel equalization problem, which is briefly described in the following section.

4 Channel Equalization

Fundamentally, the problem of channel equalization corresponds to the task of recovering an information signal from distorted measurements resulting from the action of a noisy linear / nonlinear system (channel). In the case of digital communications, which is the focus of this work, an important effect is the temporal superposition between transmitted samples, called intersymbol interference (ISI).

A common strategy to counterbalance the undesirable effects of the channel is to use a specially-tailored filter, called equalizer, at the receiver. A block diagram illustrating the channel equalization problem is displayed in Figure 1, in which $s(n)$ corresponds to the source signal, $r(n)$ is the observed signal, $\eta(n)$ represents the additive noise, $y(n)$ is the equalizer output and $d(n)$ is the desired signal, i.e., $s(n)$ or a delayed version thereof ($s(n-d)$).

The classical formulation of the equalization problem consists in selecting the filter parameters that minimize the mean-squared error between the desired signal and the equalizer output, i.e., MSE $= E\left\{[d(n) - y(n)]^2\right\}$. The solution evokes the conceptual framework of optimum (Wiener) filtering, as well as well-known online algorithms such as LMS and RLS [3]. In this work, the ESNs, described in Section 2, will play the role of nonlinear and recurrent equalizers, and will be adapted according to the error entropy criterion, aiming at a more effective use of the statistical information of the signals, and, ultimately, a better equalization performance.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

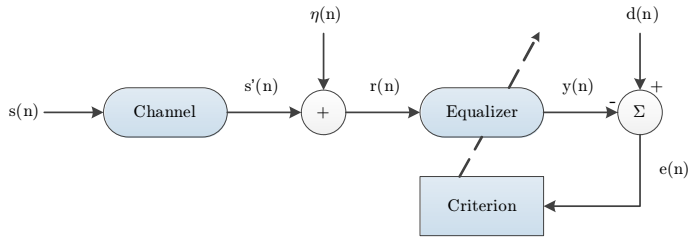


Fig. 1: Block diagram of the channel equalization problem.

4.1 Methodology

In all simulations, the source signals $s(n)$ are composed of i.i.d samples belonging to a binary alphabet $\{+1, -1\}$ (BPSK modulation). We considered two different performance measures: (1) the bit error rate (BER) and (2) the error probability density function (PDF), which offers an overview of how close to the delta function, i.e., the ideal error PDF, is the actual error PDF associated with each possible ESN and adaptation criterion.

In all experiments, for each pair ESN architecture / adaptation criterion, and for each value of signal-to-noise ratio (SNR), the BER value was obtained by transmitting symbols until 600 errors were detected or a total of 10^6 symbols was achieved. The resulting curve of BER versus SNR actually displays an average computed over 100 independent experiments for the sake of reliability.

The elements of the reservoir weight matrix $\mathbf{W}$ were set to -0.4, 0.4 and 0 with probabilities of 0.025, 0.025 and 0.95, respectively [1], while the input weights ($\mathbf{W}_{ij}^{in}$) were set to -1 or $+1$ with equal probability. The ESNs were trained using 50000 samples for both LMS and MEE-SIG algorithms. Based on preliminary tests, the step size (μ) employed for the LMS was 0.005, while, in the case of MEE-SIG, μ, σ and L assumed the values 5, 5 and 10, respectively. Finally, the number of reservoir units was equal to $N = 100$.

5 Simulation Results

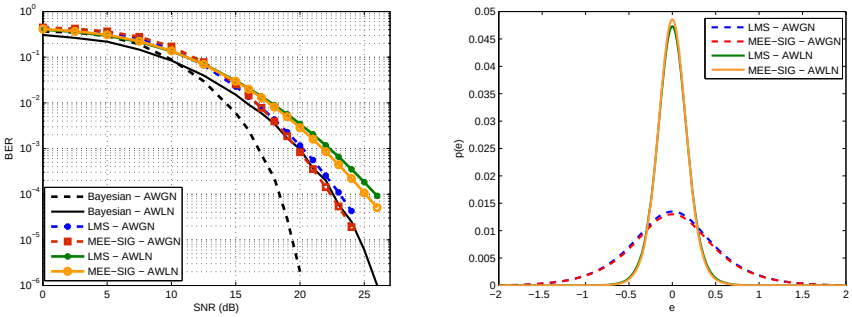
5.1 First Scenario

In this scenario, the channel is a maximum-phase system with transfer function $H(z) = 0.5 + z^{-1}$, being a nonlinear equalizer absolutely necessary when the equalization delay is zero, as considered here. With respect to the channel noise distortion, we analyzed two situations: additive white Gaussian and Laplacian noise (AWGN and AWLN, respectively). The corresponding BER versus SNR curves are shown in Figure 2(a).

Some interesting remarks can be drawn from Figure 2(a). On the one hand, for low values of SNR, the performance associated with each criterion is quite similar, having the LMS presented a slightly better BER value. This may be due to parameter misadjustements in the PDF estimation (e.g., σ) involved in

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

EEC, especially when the noise power becomes dominant. On the other hand, for higher SNRs, the EEC is effectively capable of extracting more statistical information of the reservoir signals, so that the equalization performance was improved when compared with MSE. It is also possible to observe that both approaches remained distant from the BER values related to the maximum a posteriori (MAP) equalizer with two inputs, which, in a certain sense, was expected, since the MAP equalizer has complete knowledge about the source, channel and noise characteristics and is explicitly formulated as a decision-error minimizer.



(a) BER versus SNR (b) Error PDF

Fig. 2: Equalization results for EEC and MSE considering $H(z) = 0.5 + z^{-1}$.

Figure 2(b) displays the error PDF associated with each criterion for two specific SNR values and for each noise model: 12.5 dB and 26 dB for Gaussian and Laplacian noises, respectively. In the former case, the error PDF achieved with MSE is narrower than that of EEC, which may be related to a slight σ misadjustment in the MEE-SIG algorithm. In the latter case, the error PDF associated with the EEC is more peaky and less tailed, being a better approximation of the delta function, which indicates that this criterion can better employ the statistical information of the signal for equalization.

5.2 Second Scenario

The second channel is described by the transfer function $H(z) = 1 + z^{-1}$. The peculiar characteristic of this channel is the existence of coincident states, which means that it cannot be equalized by means of feedforward structures, either linear or nonlinear. In this case, the presence of feedback connections can be decisive from the standpoint of performance improvement. Figure 3 exhibits the BER versus SNR curves obtained with each criterion for AWGN.

The results displayed in Figure 3 confirmed that the presence of recurrent connections within the reservoir allowed the ESNs to distinguish between the coincident channel states: the BER values obtained with ESNs are significantly smaller than those associated with the Bayesian approach, using two inputs, which cannot transcend an error rate of 12.5%. Additionally, we can observe

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

that the EEC led to a pronounced performance improvement, especially for high SNRs, which indicates the relevance of higher-order statistical information.

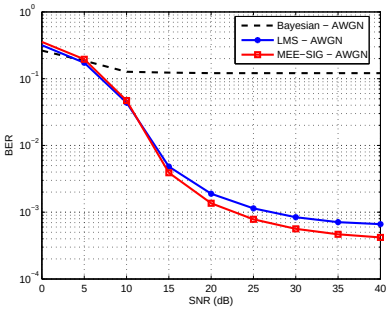


Fig. 3: BER versus SNR - MSE and EEC.

For the sake of brevity, we do not display the BER curves for the Laplacian noise nor the error PDFs. It suffices to say that the change in the noise model did not significantly affect the BER versus SNR curve, and the behavior of the error PDFs were similar to those observed in the first scenario.

6 Conclusions

In this work, we proposed the adoption of the error entropy criterion instead of the classical mean-squared error for the adaptation of the readout parameters of echo state networks. The main motivation is the possibility of achieving a more effective use of the statistical information associated with the reservoir dynamics. In the context of supervised channel equalization problem, it has been shown that the use of EEC can bring relevant performance improvements when compared with MSE, which encourages further investigations involving alternative adaptation criteria.

As future perspectives, we highlight the possibility of employing other criteria, like those based on different error norms, as well as the use of ITL criteria to train other unorganized neural networks, like extreme learning machines (ELMs).

References

[1] H. Jaeger. The echo state approach to analyzing and training recurrent neural networks. Technical Report 148, German National Research Center for Information Technology, 2001.

[2] H. Jaeger and H. Hass. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, 304(5667):78–80, 2004.

[3] S. Haykin. *Adaptive filter theory*. NJ: Prentice Hall, 3rd edition, 1996.

[4] L. Boccato, A. Lopes, R. Attux, and F. J. Von Zuben. An extended echo state network using volterra filtering and principal component analysis. *Neural Networks*, 32:292–302, 2012.

[5] J. C. Principe. *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer, 2010.