RECURSIVE RENYI'S ENTROPY ESTIMATOR FOR ADAPTIVE FILTERING

Jian-Wu Xu, Deniz Erdogmus, Mustafa C. Ozturk, Jose C. Principe
Computational NeuroEngineering Laboratory,
Electrical & Computer Engineering Department
University of Florida, Gainesville, FL 32611.
[jianwu,deniz,can,principe]@cnel.ufl.edu

ABSTRACT

Recently we have proposed a recursive estimator for Renyi's quadratic entropy. This estimator can converge to accurate results for stationary signals or track the changing entropy of nonstationary signals. In this paper, we demonstrate the application of the recursive entropy estimator to supervised and unsupervised training of linear and nonlinear adaptive systems. The simulations suggest a smooth and fast convergence to the optimal solution with a reduced complexity in the algorithm compared to a batch training approach using the same entropy-based criteria. The presented approach also allows on-line information theoretic adaptation of model parameters.

1. INTRODUCTION

Mean square error (MSE) is the fundamental performance measure in training adaptive linear filters and neural networks. For the linear cilter case, the Wiener-Hopf equation yields the analytical solution for the optimal filter coefficients [1]. Similar analytical solutions exist for unsupervised training problems involving linear filters and second-order statistics (e.g., principal components analysis). Second-order statistics are able to extract all information under the assumption of Gaussianity. However, in practice neither data distributions are Gaussian, nor we always use linear adaptive filters. These more realistic situations involving nonlinear models and non-Gaussian data distributions require the consideration of higher order-statistics for optimal information processing. Thus, in this respect, second-order statistics become suboptimal.

Therefore, for optimal information processing, it is necessary to consider criteria that emerge from information theory. These criteria (e.g., entropy and mutual information) not only deal with the second-order statistics, but also naturally take into account the higher-order statistics in adaptive filter training.

Entropy, defined by Shannon [2], is a measure of average information contained in random variable with a certain probability distribution function. Thus, for both supervised and unsupervised learning, it becomes a suitable criterion: in supervised training, minimizing the entropy of the error corresponds to minimizing the information content of this signal, whereas in unsupervised training maximizing the entropy of the filter output will guarantee that maximum amount of information is transferred from the filter input to its output (i.e., maximum mutual information between the filter's input and output is achieved).

Adaptation in nonlinear and non-Gaussian domains needs to be tackled using nonparametric approaches in general, since parametric families of the data distributions involved are simply unknown or very difficult to obtain or guess. Hence, a nonparametric entropy estimator is essential for information theoretic learning. Although many entropy estimators exist in the literature [10], most are unsuitable for on-line entropy manipulation. This process requires a recursive estimator that is able to update the estimate on a sample-by-sample basis as new data arrives to the input of the filter. The recursive entropy estimator we have proposed earlier is, therefore extremely suitable for this task [5]. An alternative to recursive estimates is to use stochastic gradients. We have previously proposed a stochastic gradient rule for entropy manipulation in training adaptive systems [9]. We will demonstrate that the stochastic information gradient presented in this earlier publication remains as a special case of the recursive information gradient (RIG) that we present here. The main advantage of recursive updates over stochastic ones is the reduction of misadjustment, which is the fluctuation of the weight vectors in the vicinity of the optimal solution [3,4].

In the following sections, we describe the recursive estimator for entropy and its recursive gradient (called RIG). In addition, we demonstrate the performance of the proposed algorithm in supervised and unsupervised adaptation scenarios, namely, linear system identification, nonlinear time-series prediction, and projection pursuit.

2. RECURSIVE RENYI'S ENTROPY ESTIMATOR

For a random variable X with probability distribution function (pdf) $f_X(x)$, Renyi's entropy of order-α is [6]

$$H_\alpha(X) = \frac{1}{1-\alpha}\log\int_{-\infty}^{\infty} f_X^\alpha(x)dx = \frac{1}{1-\alpha}\log E_X[f_X^{\alpha-1}(X)] \qquad (1)$$

Using Parzen window with kernel function $\kappa_\sigma(.)$ to estimate the pdf from its samples $\{x_1,\ldots,x_N\}$, and approximating the expectation operator with sample mean, we obtain the following estimator for Renyi's entropy [7]

$$H_\alpha(X) = \frac{1}{1-\alpha}\log\left[\frac{1}{N^\alpha}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\kappa_\sigma(x_j - x_i)\right)^{\alpha-1}\right] \qquad (2)$$

The quadratic entropy, for α =2, is given and estimated by

$$H_2(X) = -\log E_X\left[f_X(X)\right] = -\log\frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N}\kappa_\sigma(x_j - x_i) \qquad (3)$$

The argument of the 'log' is named as the quadratic information potential due to the similarity between this quantity and the physical potential energy of an ensemble of particles. Investigating the structure of the nonparametric estimator for quadratic information potential in (3), we obtained a recursive formula to update the information potential estimate when a new sample is acquired.

Assuming that the kernel function is chosen to be an even-symmetric pdf, the information potential estimate at time k, denoted by V_k, is given by

$$V_k(X) = \frac{1}{k^2}\sum_{j=1}^{k}\sum_{i=1}^{k}\kappa_\sigma(x_j - x_i) \qquad (4)$$

When a new sample arrives, V_k is modified by using the new sample x_{k+1} as

$$V_{k+1}(X) = \frac{1}{(k+1)^2}\sum_{j=1}^{k+1}\sum_{i=1}^{k+1}\kappa_\sigma(x_j - x_i)$$
$$= \frac{k^2}{(k+1)^2}V_k + \frac{1}{(k+1)^2}\left[2\sum_{i=1}^{k}\kappa_\sigma(x_{k+1} - x_i) + \kappa_\sigma(0)\right] \qquad (5)$$

When the information potential estimate is updated, the new entropy estimate can be obtained by calculating $H_{k+1}(X) = -\log V_{k+1}(X)$. This exact recursive algorithm is useful for estimating the entropy of stationary signals, however it is not suitable for nonstationary signals due to its increasing memory depth. Therefore, a *forgetting recursive entropy estimator* is necessary to serve satisfactorily in such situations. The forgetting recursive entropy estimator updates the quadratic information potential according to [5]

$$V_{k+1}(X) = (1-\lambda)V_k(X) + \frac{\lambda}{L}\sum_{i=k-L+1}^{k}\kappa_\sigma(x_i - x_{k+1}) \qquad (6)$$

The parameters, λ, L, and σ are called the forgetting factor, window length, and kernel size, respectively, and they all affect the convergence properties of this recursive entropy estimator. Increasing the forgetting factor results in faster convergence and larger estimation variance, increasing the window length results in smaller estimation variance and has no effect on convergence time, and finally, increasing the kernel size results in smaller variance and larger estimation bias.

The forgetting recursive entropy estimator reduces the computational complexity from $O(N^2)$ to $O(L)$. This is a dramatic reduction in the computation requirements. These properties make the forgetting recursive entropy estimator appealing for training adaptive systems. Hence, in the following demonstrations using entropy criteria, we will employ the forgetting recursive entropy estimator.

3. SUPERVISED LEARNING

Consider the on-line supervised training of an adaptive filter $g(\mathbf{x}_k;\mathbf{w}_k)$ where $\mathbf{w}_k = [w_1(k), w_2(k), \ldots w_M(k)]^T$ is the weight vector and $\mathbf{x}_k$ is the input vector and d_k is the desired signal at time k. The instantaneous error is [8]

$$e_k = d_k - g(\mathbf{x}_k;\mathbf{w}_k) \qquad (7)$$

The weights of the filter are adapted to minimize the error entropy using gradient descent. Since minimizing quadratic entropy is equivalent to maximizing the quadratic information potential, the following update rule can be employed:

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \mu\frac{\partial V_{k+1}}{\partial \mathbf{w}} \qquad (8)$$

The gradient will be updated according to the RIG formulation, which, in this case corresponds to

$$\frac{\partial V_{k+1}}{\partial \mathbf{w}} = (1-\lambda)\frac{\partial V_k}{\partial \mathbf{w}} + \frac{\lambda}{L}\sum_{i=k-L+1}^{k}\kappa'_\sigma(e_i - e_{i+1})\left[\frac{\partial e_i}{\partial \mathbf{w}_k} - \frac{\partial e_{i+1}}{\mathbf{w}_k}\right] \qquad (9)$$

Notice that if $\lambda = 1$, then this recursive gradient reduces to the stochastic information gradient (SIG) [9].

3.1. Linear system identification using FIR filters

Consider the case where the unknown system is an FIR filter. For simplicity, we assume that our adaptive FIR filter is sufficiently long. In practice, both the input and the desired signals are generally contaminated by independent additive white Gaussian (AWGN). In this example, we further assume that they have equal variances for simplicity.

We performed Monte Carlo simulations using a forgetting factor of 0.3. A good rule of thumb we use to select the kernel size (for Gaussian kernels) is to set it equal to 0.2 times the estimated standard deviation of the variable of interest. The performance results are

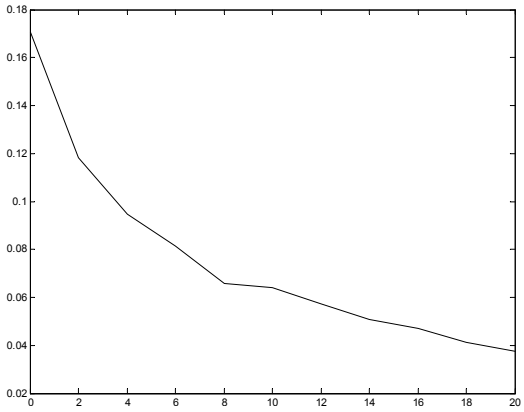

Fig 1. Average model error versus SNR.

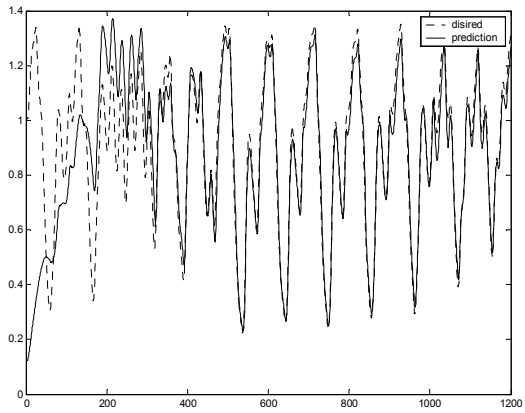

Fig 2. Mackey-Glass series prediction by TDNN.

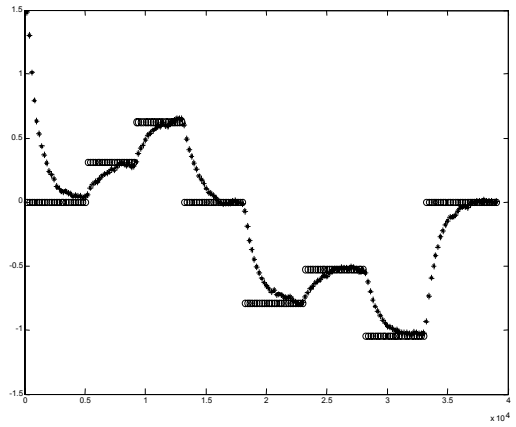

Figure 3. Nonstationary projection pursuit: tracking the desired direction

summarized in figure 1 in the form $E\left\|\mathbf{w}_{true}-\mathbf{w}_{final}\right\|$ versus input SNR, where $\mathbf{w}_{true}$ and $\mathbf{w}_{final}$ are the weight vectors of the reference model and the adaptive filter after convergence (1000 samples and iterations).

3.2. Time-series prediction using TDNN

Next we use recursive Renyi's entropy estimator to train a 4:4:1 (4-tap delay line input, 4 tanh-neurons in hidden layer and 1 linear output neuron) TDNN [11] to perform single-step forward prediction of the Mackey-Glass chaotic time-series, which often serves as a benchmark data set in testing prediction algorithms. This signal has a delay-based chaotic behavior and an attractor associated with the given delay amount. A time-series is generated by sampling the MG30 signal at T=1s intervals whose continuous time dynamics are defined by

$$\dot{x}(t) = -0.1x(t) + \frac{0.2x(t-30)}{1+x^{10}(t-30)} \tag{10}$$

The integration is performed using the Runge-Kutta4 technique with a time-step of 0.1s. A total of 10000 samples are generated using a random initial condition. Figure 2 shows the predictor output converging to the desired signal as the weights of the TDNN are updated using RIG and the minimum error entropy criterion.

In this experiment the kernel size, window length and forgetting factor are σ=0.1, L=5, λ=0.3, respectively.

4. UNSUPERVISED LEARNING

In this section, we discuss the raining of a linear filter in a projection pursuit context. Projection pursuit is the problem of determining *interesting* linear projections of vector signals. For example, principal components are *interesting* in the maximum variance sense. In this example, we take the minimum entropy direction as the desired projection of the data. A synthetic 2-dimensional vector signal using zero-mean, unit-variance Gaussian and uniform random variables is generated and these are mixed by a rotation matrix.

$$\mathbf{x}_k = \begin{bmatrix} Cos(\theta_1) & -Sin(\theta_1) \\ Sin(\theta_2) & Cos(\theta_2) \end{bmatrix} \begin{bmatrix} u_{1k} \\ u_{2k} \end{bmatrix} \tag{13}$$

Since the Gaussian distribution has maximum entropy among all-fixed-variance distributions, we expect the RIG-based projection pursuit algorithm to determine the uniformly distributed projection and track this direction should it change in time. After every update, the weight vector is normalized to unit norm to prevent it from shrinking to zero.

In this experiment, we used 0.3, 100 and 0.1 for the forgetting factor, window length, and kernel size, respectively. Figure 3 shows the variation of the actual angle of the desired direction (piecewise stationary) and the RIG-projection pursuit algorithm estimates with respect to time.

5. CONCLUSIONS

We are currently at a point where second-order statistical signal processing is not sufficient for our purpose anymore. Information theory provides a natural extension of many familiar ideas such as variance and correlation to nonlinear and non-Gaussian situations in the form of entropy and mutual information. Information theoretic signal processing requires the knowledge or the estimation of signal probability distributions so that the necessary information theoretic statistics can be calculated and manipulated by adaptation algorithms. In this paper, we demonstrated how to use the Parzen window based recursive estimator for Renyi's quadratic entropy and its gradient (RIG) for supervised and unsupervised adaptive signal processing. In previous studies, the data efficiency and the robustness of this Parzen window based nonparametric entropy estimator have been demonstrated [12]. The recursive information gradient presented here eliminates the high computational complexity drawback of the training algorithms associated with this entropy estimator. It also allows entropy-based on-line adaptation on a sample-by-sample basis without suffering much from misadjustment. Simulation results using linear system identification, chaotic time-series prediction, and projection pursuit suggest the usefulness of the proposed RIG-based learning rules.

Acknowledgments: This work was supported by the NSF grant ECS-0300340.

REFERENCES

[1] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Application,* MIT Press, MA, 1966.

[2] C.E. Shannon, "A Mathematical Theory of Communication," Bell Sys. Tech. J., vol. 27, pp. 379-423, 623-653, 1948.

[3] E. Parzen, "On Estimation of a Probability Density Function and Mode," in *Time Series Analysis Papers*, Holden-Day, CA, 1967.

[4] J.F. Bercher, C. Vignat, "Estimating the Entropy of a Signal with Applications," IEEE Trans. Signal Processing, vol. 48, no. 6, 2000.

[5] D. Erdogmus, J.C. Principe, S.P. Kim, J.C. Sanchez, "A Recursive Renyi's Entropy Estimator," *Proc. NNSP'02*, pp. 209-217, Martigny, Switzerland, 2002.

[6] A. Renyi, *Probability Theory,* Elsevier, NY, 1970.

[7] D. Erdogmus, J.C. Principe, "Generalized Information Potential Criterion for Adaptive System Training," IEEE Trans. Neural Networks, vol. 13, no. 5, 2002.

[8] S. Haykin, *Adaptive Filter Theory*, 4th ed., Prentice Hall, NJ, 2002.

[9] D. Erdogmus, J.C. Principe, K.E. Hild II, "On-Line Entropy Manipulation: Stochastic Information Gradient," IEEE Signal Processing Letters, vol. 10, no. 8, pp. 242-245, 2003.

[10] J. Beirlant, E.J. Dudewicz, L. Gyorfi, E.C. van der Meulen, "Nonparametric Entropy Estimation: An Overview," Int. J. Math. Stat. Sci., vol. 6, pp. 17-39, 1997.

[11] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed., Prentice Hall, NJ, 1999.

[12] K.E. Hild II, D. Erdogmus, J.C. Principe, "Blind Source Separation Using Renyi's Mutual Information," IEEE Signal Processing Letters, vol. 8, no. 6, pp. 174-176, 2001.