ENTROPY MINIMIZATION ALGORITHM FOR MULTILAYER PERCEPTRONS

Deniz Erdogmus, Jose C. Principe
Computational NeuroEngineering Lab (CNEL), University of Florida, Gainesville, FL 32611
[deniz,principe]@cnel.ufl.edu

ABSTRACT

We have previously proposed the use of quadratic Renyi's error entropy with a Parzen density estimator with Gaussian kernels as an alternative optimality criterion for supervised neural network training, and showed that it produces better performance on the test data compared to the MSE. The error entropy criterion imposes the minimization of average information content in the error signal rather than simply minimizing the energy as MSE does. Recently, we developed a nonparametric entropy estimator for Renyi's definition that makes possible the use of any entropy order and any suitable kernel function in Parzen density estimation. The new estimator reduces to the previously used estimator for the special choice of Gaussian kernels and quadratic entropy. In this paper, we briefly present the new criterion and how to apply it to MLP training. We also address the issue of global optimization by the control of the kernel size in the Parzen window estimation.

1. INTRODUCTION

Mean square error (MSE) has been the most widely utilized performance criterion in supervised learning due to its mathematical simplicity that allows theoretical analysis simple. It also provided sufficient means for exploiting the second order statistics of signals in a world of Gaussian distributions and linear systems [Haykin]. Recently, the concept of information filtering started to develop in the signal processing community, and we had presented an entropy minimization algorithm, which uses the unconventional Renyi's quadratic entropy with Parzen windowing, for supervised training of an MLP [Deniz ICA]. Also we had shown that the entropy minimization algorithm outperformed the MSE criterion in terms of acquiring more 'information' about the probability distribution of the desired signal. In a later work, recently submitted we proved that this process was equivalent to maximizing the mutual information between the desired signal and the MLP output [Deniz SP]. In this paper, we present a generalized entropy minimization algorithm for MLPs.

In this generalized entropy criterion framework, our previous algorithm is preserved as a special case corresponding to Renyi's quadratic entropy. The equivalence arguments about entropy minimization and mutual information maximization are still valid. We also show that Parzen windowing, which is used to estimate the error probability density function, preserves the global minimum provided that certain constraints on the kernel function are satisfied. Furthermore, we analyze the effect of kernel size on the cost function and show that there is a link with the convolution smoothing method of global optimization.

The entropy minimization algorithm gives rise to some interesting analogies between information theoretical learning and physics. With this criterion, it becomes possible to talk about quantities like information potential in a set of samples from a random variable and the information forces that these information particles exert on each other during learning. These concepts were previously introduced by Principe et.al. for Renyi's quadratic entropy in the context of blind source separation and SAR image pose estimation problems [Principe et al]. Here we extend the definition of these quantities to any choice of the parameter for Renyi's entropy and investigate their role in supervised learning.

Finally, the algorithm is tested on the prediction of Mackey-Glass time series. Results show that the entropy minimization algorithm is an effective tool for supervised training of adaptive systems. Our analyses on the effect of kernel size show that this parameter can be utilized to modify the performance surface to our advantage in avoiding local minima. Such modifications are not possible with the MSE criterion.

The organization of this paper is as follows. First we derive the estimator for Renyi's entropy in Section II. Next, we define the order-α information potential and information forces, study their relationship with their quadratic counterparts, and demonstrate their role in the training process in Section III. This investigation is followed by the presentation of the supervised steepest descent training algorithm for adaptive systems using the entropy as the performance measure. Following the algorithm, we start analyzing the

mathematical details of the criterion in more detail. In Section V, we demonstrate the link between our estimation method and the convolution smoothing method of global optimization. We also look into the question of finding the relationship between the entropy criterion and the MSE criterion, and we show in Section VI that MSE is a special case of the quadratic entropy criterion under quite restrictive conditions. Finally, we present a case study where we present result from a MLP training example in Section VII, followed by a discussion and conclusion section.

2. ENTROPY ESTIMATOR

Renyi's entropy (order α) for a random variable with probability density function (pdf) $f_e(.)$ is given by [11]

$$H_\alpha(e) = \frac{1}{1-\alpha}\log\int_{-\infty}^{\infty} f_e^\alpha(e)de \tag{1}$$

Renyi's entropy shares the same extreme points with Shannon's definition for all values of α, i.e. its minimum value occurs at the δ-distribution, and the maximum occurs when at a uniform pdf. In practice, usually it is necessary to nonparametrically estimate the density from the samples. The pdf estimate of a random variable e with the samples $\{e_1,\ldots,e_N\}$, is obtained with the following expression using a kernel $\kappa_\sigma(.)$, where σ specifies the size of the window.

$$\hat{f}_e(e) = \frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma(e-e_i) \tag{2}$$

By observing that the integral of $f_e^\alpha(e)$ in (1) is the expected value of $f_e^{\alpha-1}(e)$, we replace the expectation operator by the sample mean, and then we replace the actual pdf with the Parzen window estimate [12] to obtain the nonparametric entropy estimator.

$$H_\alpha(e) = \frac{1}{1-\alpha}\log\left[\frac{1}{N^\alpha}\sum_j\left(\sum_i\kappa_\sigma(e_j-e_i)\right)^{\alpha-1}\right] \tag{3}$$

We call the argument of the log in (1) the information potential [Principe et al]. Minimizing the entropy is equivalent to maximizing the information potential for $\alpha>1$, or minimizing the information potential for $\alpha<1$, since the log is a monotonous function.

It can be shown that the minimum value of the entropy will be achieved when the error samples are $e_1=\ldots=e_N=0$. To prove that this point is a minimum, we evaluate the gradient and the Hessian of the estimator in (3) at these error samples. When these quantities are evaluated, we find that

$$\left.\frac{\partial\hat{H}_\alpha}{\partial e_k}\right|_{e=\bar{0}} = \frac{1}{1-\alpha}\left.\frac{\partial\hat{V}_\alpha/\partial e_k}{\hat{V}_\alpha}\right|_{e=\bar{0}} = 0 \tag{4}$$

and the eigenvalue-vector pairs of the Hessian are

$$\begin{aligned}&\{0,[1,\ldots,1]^T\},\{aN/(N-1),[1,-1,0,\ldots,0]^T\},\\&\{aN/(N-1),[1,0,-1,0,\ldots,0]^T\},\ldots\end{aligned} \tag{5}$$

where $a\stackrel{\Delta}{=}-(N-1)\kappa_\sigma^{\alpha-1}(0)\left[(\alpha-2)\kappa'^2(0)+2\kappa(0)\kappa''(0)\right]/N^2$

The zero eigenvalue and the corresponding eigenvector are due to the fact that the entropy is invariant with respect to changes in the mean of the random variable. Thus, the entropy has a minimum line instead of a single point along the direction where only the mean of the error samples changes. For this reason, entropy training is unable to match the mean of the MLP output to that of the desired when the otput layer consists of linear neurons with bias terms. These weights must be assigned proper values to set the error sample mean to zero after the training converges for the other weights. The eigenvalues given in (5) show how to choose kernel functions to guarantee a minimum point. Provided that $\kappa'(0)=0$ (which is the case for symmetric and differentiable kernels), the nonzero eigenvalue with multiplicity (N-1) at $\bar{e}=0$ is positive iff N>1, $\kappa(0)>0$, and $\kappa''(0)<0$.

3. COMPENSATING FOR VARIANCE OF THE SAMPLE MEAN

We know that the sample mean is an unbiased and asymptotically consistent estimator for the expected value operator. However, it still introduces extra variance in the estimator, in addition to that of the Parzen windowing. Consider the original quadratic information potential estimator, which uses Gaussian kernels [Deniz ICA].

$$\begin{aligned}V_2(e) &= \int_{-\infty}^{\infty} f_e^2(e)dy^\rho = \int_{-\infty}^{\infty}\left(\frac{1}{N}\sum_i G_\sigma(e-e_i)\right)^2 de\\&= \frac{1}{N^2}\sum_i\sum_j\int_{-\infty}^{\infty} G_\sigma(e-e_i)G_\sigma(e-e_j)de\\&= \frac{1}{N^2}\sum_i\sum_j G_{\sigma\sqrt{2}}(e_j-e_i)\end{aligned} \tag{6}$$

This estimator exploits the fact that the integral of the product of two Gaussian functions is another Gaussian function with twice the variance. Now consider the new estimator with quadratic entropy and the same choice of kernel function. We get the results by direct substitution of α=2 and $\kappa = G_\sigma$ in (3).

$$V_2(e) = \frac{1}{N^2}\sum_i \sum_j G_\sigma(e_j - e_i) \tag{7}$$

The estimator in (7) has exactly the same form as the one in (6), but a larger variance since the effective kernel size used in Parzen windowing is smaller. This extra variance is introduced by the sample mean approximation. In this special case, this can be compensated by simply choosing a larger kernel size in (7), specifically $\sqrt{2}$ times the original kernel size. In general, for quadratic entropy, there exists a kernel function for (7) that corresponds to any choice of kernel in (6), which renders the exact same estimation for the information potential. The relationship between these two kernel functions are given by

$$\kappa_{new}(x_j - x_i) = \int_{-\infty}^{\infty} \kappa_{old}(x - x_i)\kappa_{old}(x - x_j)dx \tag{8}$$

Gaussian kernels are very special, because a simple rescaling of the kernel function achieves this equality.

4. INFORMATION POTENTIAL AND INFORMATION FORCES

The use of kernels results in a formulation of the entropy that brings an interesting interpretation to the training process. The concept of information potential fields, and information forces were defined and investigated for the quadratic entropy with Gaussian kernels before [Principe et al]. The training samples under this interpretation become information particles. With the new nonparametric entropy estimator, it becomes possible to define the order-α potentials and information forces. The information potential estimator is

$$\hat{V}_{\alpha,\sigma}(e) = \frac{1}{N^\alpha}\sum_j \left(\sum_i \kappa_\sigma(e_j - e_i)\right)^{\alpha-1} \tag{9}$$

where the n-dimensional size-σ kernel is expressed in terms of the unit-size kernel as

$$\kappa_\sigma(x) = \frac{1}{\sigma^n}\kappa(x/\sigma) \tag{10}$$

From (9), potential energy of an information particle e_j can be immediately deduced.

$$\hat{V}_{\alpha,\sigma}(e_j) = \frac{1}{N^\alpha}\left(\sum_i \kappa_\sigma(e_j - e_i)\right)^{\alpha-1} \tag{11}$$

From (11) the information force on e_j can be evaluated.

$$\begin{aligned} F_\alpha(e_j) &= \frac{\partial \hat{V}_\alpha(e_j)}{\partial e_j} \\ &= \frac{(\alpha-1)}{N^\alpha}\left(\sum_i \kappa_\sigma(e_j - e_i)\right)^{\alpha-2}\left(\sum_{i\neq j} \kappa'_\sigma(e_j - e_i)\right) \end{aligned} \tag{12}$$

It can be shown that this is equivalent to

$$F_\alpha(e_j) = (\alpha-1)\hat{f}_e^{\alpha-2}(e_j)F_2(e_j) \tag{13}$$

where the quadratic force is defined as

$$F_2(e_j) = \frac{1}{N^2}\left(\sum_{i\neq j} \kappa'_\sigma(e_j - e_i)\right) \tag{14}$$

This clarifies the relationship between the α-force and the quadratic force. Obviously, the quadratic force expression in (14) reduces to the exact same definition in [Principe et al] when Gaussian kernels are assumed. Now from (13) and (14), it is possible to define the force on e_j due to e_i.

$$F_\alpha(e_j; e_i) = (\alpha-1)\hat{f}_e^{\alpha-2}(e_j)F_2(e_j; e_i) \tag{15}$$

where

$$F_2(e_j; e_i) = \frac{1}{N^2}\kappa'_\sigma(e_j - e_i) \tag{16}$$

This formulation indicates that the quadratic force can be regarded as the foundation for all other information forces. Forces of any order are the scaled version of the quadratic force, with the scaling factor being a power of the probability density of the particle that the force acts upon. For $\alpha > 2$, the force on a particle increases with increased probability density, while it decreases for $\alpha < 2$.

5. QUADRATIC ENTROPY AND MSE

In this section, we will show that under quite restrictive conditions MSE is a special case of the quadratic entropy in the limit. Since minimizing the quadratic entropy is equivalent to maximizing the quadratic information potential, we focus on this

quantity. The quadratic information potential estimated with Gaussian kernels is given by (7). Assuming 1-dim error samples for simplicity, if the kernel size, which corresponds to the standard deviation in this case, is chosen to be very large, then the Gaussian kernel can be approximated by its first order Taylor expansion at zero.

$$G_\sigma(x) = ce^{-x^2/2\sigma^2} \approx c(1 - x^2/2\sigma^2) \tag{17}$$

Substituting this approximation in (7), a large-kernel-size approximation to the quadratic information potential is obtained, whose terms can be rearranged to yield the equivalence with MSE.

$$\begin{aligned}
\max \hat{V}_{2,\sigma}(e) &\approx \frac{1}{N^2}\sum_i\sum_j c(1-(e_i-e_j)^2/2\sigma^2) \\
&= c - \frac{c}{2\sigma^2 N^2}\sum_i\sum_j (e_i-e_j)^2 \\
&= c - \frac{c}{2\sigma^2 N^2}\sum_i\sum_j (e_i-e_j)^2 \\
\equiv \min \sum_i\sum_j (e_i-e_j)^2 &= 2N\sum_i e_i^2 - 2\left(\sum_i e_i\right)^2 \\
&= 2N \cdot MSE(e) - 2N^2\bar{\mu}_e^2
\end{aligned} \tag{18}$$

Thus, we conclude that under the following listed conditions the minimization of quadrtic entropy is equivalent to the minimization of MSE.

i. The first order Taylor expansion around zero of the kernel must be quadratic in its argument with a negative coefficient.
ii. The kernel size must be chosen large enough to have a valid first order approximation.
iii. Sample mean of the error must be zero. (Introducing a bias term at the output of the function approximator and setting it to yield zero error mean at each step can achieve this.)

The equivalence is also valid if the variance/entropy of the error is small, so that the difference between the error samples is much smaller than the kernel size used. This occurs if the topology is sufficient to approximate the desired function very accurately. An example is the equalization of an FIR channel, where a sufficiently long FIR equalizer can practically eliminate the inter-symbol interference. We have shown experimentally that both criteria provide very close solutions for the FIR equalizer [Ignacio].

6. KERNEL AND SMOOTHING

The kernel size σ is a very important parameter if efficiently exploited. Parzen windowing is a biased estimator of the pdf and in determining the kernel size, a trade-off has to be made between low bias and low variance. Once a suitable value is set, training can be carried out using that fixed kernel size. However, it turns out that there is a way to utilize the kernel size as a means of avoiding local optimum solutions in training. The following is the relation between the information potential estimates using an arbitrary kernel size and a unit kernel, in n-dim error space.

$$\hat{V}_{\alpha,\sigma}(e) = \frac{1}{\sigma^{n(\alpha-1)}}\hat{V}_{\alpha,1}(e/\sigma) \tag{19}$$

The change in kernel size causes a scaling of the cost function accompanied by dilation. Thus, all the points, except for the origin, including all local extremes move radially away from the origin as the kernel size increases. This leads to a global optimization procedure for the training process. Starting with a large kernel size, and then slowly decreasing it towards a predetermined suitable value, the local solutions may be avoided. Hence, global optimization will be achieved using a gradient descent approach.

In addition to the dilation property in the finite sample case discussed above, there is another interesting property that the Parzen windowing brings about. We have shown that in the limit as the number of samples approach to infinity, the kernel pdf estimation converges to a convolution operation, which in turn is strongly related to the well-known convolution smoothing method of global optimization [Rubinstein].

7. GRADIENT BASED ADAPTATION ALGORITHM USING ENTROPY

Suppose we are training an adaptive system to approximate an unknown function using the information in the training data, as depicted in Fig. 1. We define the error as the difference between the desired output and the output of the mapper to a corresponding input. The optimization criterion will be the minimization of error entropy.

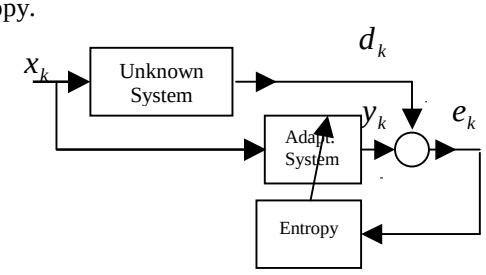

Figure 1. Supervised Adaptive System Training

The training process will be carried out by maximizing the information potential for $\alpha>1$. We encounter the information forces in the computation of

the gradient of the information potential. The gradient consists of the information force and the sensitivity terms as seen below in (20).

$$\frac{\partial \hat{V}_\alpha(e)}{\partial w} = \sum_j \frac{\partial \hat{V}_\alpha(e_j)}{\partial e_j}\frac{\partial e_j}{\partial w} = \sum_j F_\alpha(e_j) S_w(e_j) \qquad (20)$$

Explicitely, written, the gradient of the information potential with respect to the weights is

$$\frac{\partial \hat{V}_\alpha}{\partial w} = \frac{(\alpha-1)}{N^\alpha}\sum_j \left(\sum_i \kappa_\sigma(e_j - e_i)\right)^{\alpha-2} \cdot \left[\sum_i \kappa'_\sigma(e_j - e_i)\left(\frac{\partial y_i}{\partial w} - \frac{\partial y_j}{\partial w}\right)\right] \qquad (21)$$

where the sensitivity of the output for an MLP can be computed by backpropogation [Rumelhart].

Investigating the relationship between the gradient of the order-α information potential and the quadratic information potential is of theoretical interest. The expression in (21) can be rearranged to yield

$$\frac{\partial \hat{V}_\alpha(e)}{\partial w} = (\alpha-1)\sum_j \hat{f}_e^{\alpha-2}(e_j)\frac{\partial \hat{V}_2(e_j)}{\partial w} \qquad (22)$$

Clearly, in the order-α case, the total gradient is a weighted mixture of the individual gradients created by each particle where the mixture coefficients are the powers of the pdf estimate of the corresponding particle. This property directly translates from what was observed for the information forces.

Gradient adaptation, although not the only possibility, is preferred due to its simplicity and efficient convergence characteristics [Haykin]. Alternative optimization approaches may also be used, global or otherwise [Aarts,Morejon].

8. MLP TRAINING EXAMPLE

In this section, we present results for an MLP training example which uses the entropy minimization to achieve single-step prediction of Mackey-Glass (MG) series [Kaplan]. The time series is generated with delay parameter τ=30. MLP input vector consists of 6 consecutive samples of the MG time series, thus a TDNN. A training set of 200 input-output pairs is prepared. There are 6 neurons in the only hidden layer with biases and *tanh* nonlinearities, and a single, linear output neuron. The bias value of the output neuron is set to match the means of the desired and the actual outputs. In our simulations we used fixed values of α and σ, in order to investigate their effect on the performance. Each MLP is trained starting from a set of predetermined 100 initial weights generated by a uniform distribution in the interval [-1,1]. Then, the best solution among the 100 candidates was selected.

In the testing process, the errors of the MLPs for each value of α and σ were evaluated on an independently generated 10,000-point test set. Parzen windowing with Gaussian kernels (σ=0.001) is applied to estimate the error pdfs.

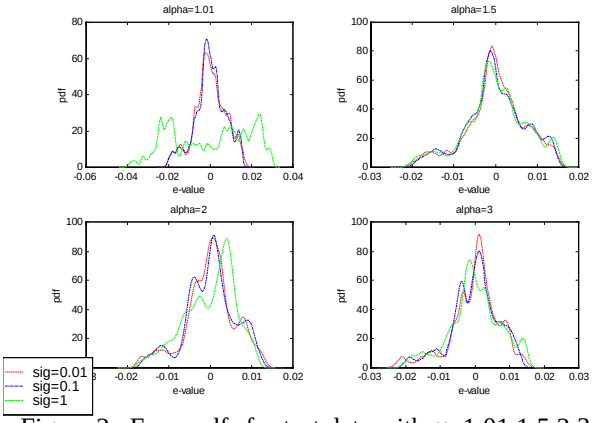

Figure 2. Error pdfs for test data with α=1.01,1.5,2,3

From Fig. 2, we observe that MLPs trained with smaller kernel sizes produce practically the same relatively more spiky solution, whereas the MLPs trained with the largest kernel size yield a widely distributed error pdf, especially for the case α=1.01. In addition, the error pdf becomes closer to the desired δ-function when α is larger. Notice that the peak value achieved by the error pdf of the MLP trained with α=3, and σ=0.01 is the highest among all error pdfs.

9. CONCLUSIONS

Renyi's entropy of the error was previously proposed as an alternative to MSE, and was shown to be advantageous. At that time, the main focus was on the special case of quadratic entropy with Gaussian kernels due to analytical difficulties encountered otherwise. Recently, we proposed an alternative nonparametric estimator for Renyi's entropy that allows us to use any suitable kernel function and entropy order. In this paper, we have demonstrated the equivalence between the old and new estimators for Gaussian kernels and quadratic entropy. In addition, we have shown that MSE is in fact a special case of the proposed quadratic entropy criterion in the limit as the kernel size goes to infinity. We gave the definitions for order-α information force and potential, and illustrated that they are closely linked to their quadratic counterparts. Interestingly, quadratic entropy and all related quantities are the most advantageous in terms of computational

savings. Another very important aspect of the proposed entropy criterion is its relationship with the convolution smoothing method. We have demonstrated here the effect of the kernel size on the criterion. It was noted that by starting with a large kernel size and properly decreasing it may help avoid local-optimum solutions even with gradient based methods.

Finally, we have applied the criterion to the problem of MLP training in a short-term chaotic time series prediction problem. In this, we investigated the performance of the solutions generated by MLPs that are trained using different orders of entropy and different kernel sizes. Simulation results suggested that, smaller kernel sizes produced better solutions in terms of generalizability, however, there is not enough evidence yet to be certain about how the entropy order affects the performance of the resulting system.

Acknowledgements: This work is partially supported by NSF grant ECS-9900394.

REFERENCES

[1] S. Haykin, *Introduction to Adaptive Filters*, MacMillan, NY, 1984.

[2] D. Erdogmus, J.C. Principe, "Comparison of Entropy and Mean Square Error Criteria in Adaptive System Training Using Higher Order Statistics", Proc. ICA 2000, Helsinki.

[3] D. Erdogmus, J.C. Principe, "An Entropy Minimization Algorithm For Short-Term Prediction of Chaotic Time Series," submitted to IEEE Transactions on Signal Processing, Sept. 2000.

[4] J.C.Principe, D.Xu, J.Fisher, "Information Theoretic Learning," in *Unsupervised Adaptive Filtering*, vol I, Simon Haykin Editor, pp265-319, Wiley, 2000.

[5] I.Santamaria, D.Erdogmus, J.C.Principe, "Entropy Minimization for Digital Communications Channel Equalization," submitted to IEEE Transactions on Signal Processing, Dec. 2000.

[6] R.Y. Rubinstein, *Simulation and the Monte Carlo Method*, John Wiley & Sons, NY, 1981.

[7] D. Rumelhart, G. Hinton, R. Williams, "Learning internal representations by error backpropagation," *Nature*, vol 323, pp.533-536, 1986.

[8] E. Aarts, J. Korst, *Simulated annealing and Boltzmann machines: a stochastic approach to combinatorial optimization and neural computing*, Wiley, NY, 1989.

[9] R. Morejon, J.C. Principe, "Application of Levenberg-Marquardt Training to Information Theoretic Learning Systems," submitted to IJCNN 2001.

[10] D. Kaplan, L. Glass, *Understanding Nonlinear Dynamics*, Springer-Verlag, NY, 1995.

[11] Renyi, A., *Probability Theory*, American Elsevier Publishing Company Inc., New York, 1970.

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