CONVERGENCE ANALYSIS OF THE INFORMATION POTENTIAL CRITERION IN ADALINE TRAINING

Deniz Erdogmus, Jose C. Principe
Computational NeuroEngineering Lab, Electrical & Computer Engineering Dept.
University of Florida, Gainesville, FL 32611
[deniz,principe]@cnel.ufl.edu

Abstract. In our recent studies we have proposed the use of minimum error entropy criterion as an alternative to minimum square error (MSE) in supervised adaptive system training. We have formulated a nonparametric estimator for Renyi's entropy with the help of Parzen windowing. This formulation revealed interesting insights about the process of information theoretical learning. We have applied this new criterion to the training of linear and nonlinear adaptive topologies under the problems of blind source separation, channel equalization, and time-series prediction with superb results. In this paper, we analyze the structure of the entropy criterion performance surface around the optimal solution and we derive the upper bound for the step size in Adaline training with the steepest descent algorithm. We also investigate the effects on adaptation of the kernel size in the Parzen windowing, and order of Renyi's entropy.

INTRODUCTION

Mean square error (MSE) has been the fundamental performance criterion in function approximation and optimal filtering. Starting with the pioneering work of Wiener [1] and Kolmogorov that instituted the viewpoint of regarding the adaptive filters as statistical function approximators, MSE has become the workhorse of adaptive filtering theory. Combined with the basic adaptive FIR filter structure, MSE yields a simple optimization problem, whose solution is given by the Wiener-Hopf equation [2]. After the development of this analytical solution, the equivalent procedure of using the steepest descent algorithm to minimize the tap weights of the FIR filter was proposed and analyzed [2, 3]. The least square algorithm (LMS) proposed by Widrow is the most widely recognized variant of this algorithm [4]. The two main thrusts in these analyses were the issues of stability and convergence speed. Due to the quadratic form of the cost function in terms of the weight vector, the analysis of the algorithms could be simply and accurately done [4].

Previously, we proposed the use of quadratic Renyi's entropy of the error signal in supervised adaptive system training [5], and used a nonparametric estimator based on Parzen windowing with Gaussian kernels due to analytical reasons. Later, we have proved that a system trained with the error entropy criterion minimized an information theoretic distance measure between the probability density functions (pdf) of the desired and the actual outputs, and this was demonstrated experimentally for chaotic time series prediction using neural networks [6]. Recently, we have formulated a new nonparametric estimator for Renyi's entropy that allows us to compute any order of entropy using any suitable kernel function [7]. This generalized estimator reduced to the previous estimator for quadratic entropy for the appropriate choices of order and kernel function. In that work, we were also able to generalize the concepts of information potential and information force to any order α. These quantities were previously defined by Principe et.al. in the context of blind source separation for quadratic entropy [8].

In the adaptation process, the steepest ascent was extensively used due to its speed and accuracy. The step size was arbitrarily chosen small, but how to choose suitable values for different kernel sizes and entropy orders was unknown. In this paper, we seek to establish the rules governing this association, and devise a way to determine a proper value for step size to guarantee stability of the algorithm.

The organization of this paper is as follows. In the next section, we give a brief overview of the information potential criterion, and its nonparametric estimator. We derive the steepest ascent algorithm for Adaline, and find the linearized dynamic equations that govern the dynamics of weights. An upper bound for the step size for stability, and an approximate time constant expression for the dynamics of the weights are also derived. Next, the effects of the entropy order, and the kernel size on performance surface, hence on the dynamics of the modes are investigated. A section is devoted to a case study very helpful in visualizing the results of the preceding sections. Finally, a summary of the main results and observations that have been made is given in the conclusions.

INFORMATION POTENTIAL CRITERION

Consider the supervised training scheme depicted in Fig. 1. We have previously showed that training the adaptive system to minimize the entropy of the error distribution is equivalent to minimizing the α-divergence [9] between the joint pdfs of the input-desired and input-output signal pairs [6]. Therefore, as the cost function, we utilize Renyi's entropy of the error distribution.

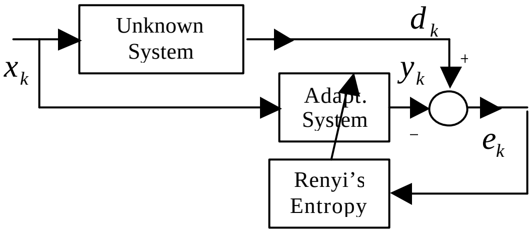

Figure 1. Supervised adaptive system training

Renyi's entropy for a random variable e is given in terms of its pdf as [10]

$$H_a(e) = \frac{1}{1-a} \log \int_{-\infty}^{\infty} f_e^a(e) de \qquad (1)$$

where α >0 is the family parameter of the entropy, sometimes called the order of entropy. In the limit, Renyi's entropy approaches Shannon's entropy as α→1. We have defined the argument of the log to be the (order-α) information potential [7, 8]. We can write the information potential in a different form using the expected value operator as

$$V_a(e) = \int f_e^a(e) de = E\left[f_e^{a-1}(e)\right] \approx \frac{1}{N} \sum_i f_e^{a-1}(e_i) \qquad (2)$$

By substituting the Parzen window estimator [11] for the pdf of e in (2), we get

$$V_a(e) = \frac{1}{N^a} \sum_j \left(\sum_i k_s(e_j - e_i) \right)^{a-1} \qquad (3)$$

where the kernel function in Parzen windowing is k_s with σ denoting the width of the window in terms of a predetermined unit-width kernel κ as given below.

$$k_s(x) = \frac{1}{s} k\left(\frac{x}{s}\right) \qquad (4)$$

Since log is a monotonic function, minimization of Renyi's entropy corresponds to maximization of the information potential for α >1. Therefore, the information potential can replace the entropy criterion resulting in computational savings.

We have previously investigated the relationship of this new criterion with the convolution smoothing method and the link between the quadratic entropy and the MSE criterion [7]. Specifically, we have shown that MSE is a special case of the quadratic entropy minimization criterion, corresponding to a very large kernel size. As for its connection to convolution smoothing, we have shown that increasing the kernel size causes a dilation of the performance surface in the weight space, which in the limit as the number of samples go to infinity, turns into an equivalence with smoothing the cost function (eliminating local maxima) by convolving with a function, where the kernel size controls the amount of smoothing applied.

STEEPEST DESCENT TRAINING OF ADALINE

Suppose the adaptive system under consideration in Fig. 1 is an Adaline structure with a weight vector w. The error samples can then be represented by $e_k = d_k - w^T x_k$, where x_k is the input vector, formed by feeding the input signal to a tapped delay line for the special case of FIR filter. Then the gradient of the information potential estimator in (3) with respect to the weights is simply

$$\frac{\partial V_a}{\partial w} = \frac{(a-1)}{N^a} \sum_j \left(\sum_i k_s(e_j - e_i) \right)^{a-2} \cdot \left(\sum_i k_s'(e_j - e_i)(x_i - x_j)^T \right) \qquad (5)$$

In this expression, further simplifications are possible through the use of (4) and the following identity between the derivatives of a width-σ kernel and a unit-width kernel.

$$k'_s(x) = \frac{1}{s^2} k'\left(\frac{x}{s}\right) \tag{6}$$

With these substitutions, the explicit expression for the gradient becomes

$$\frac{\partial V_a}{\partial w} = \frac{(a-1)}{s^a N^a} \sum_j \left(\sum_i k\left(\Delta e_w^{ji}\right) \right)^{a-2} \cdot \left(\sum_i k'\left(\Delta e_w^{ji}\right) \cdot (x_i - x_j)^T \right) \tag{7}$$

From here on we will use the following notation.

$$\Delta e_w^{ji} = \left(\frac{\left(d_j - d_i\right) - w^T\left(x_j - x_i\right)}{s} \right) \tag{8}$$

In order to maximize the information potential, we update the weights along the gradient direction with a certain step size h.

$$w(n+1) = w(n) + \eta \ \nabla V_\alpha(w(n)) \tag{9}$$

where $\nabla V_\alpha(w(n))$ denotes the gradient of V_a given in (7) evaluated at $w(n)$. To continue with our analysis, we consider the Taylor series expansion truncated to the linear term of the gradient around the optimal weight vector w_*.

$$\nabla V_a(w) = \nabla V_a(w_*) + \frac{\partial \nabla V_a(w_*)}{\partial w}\left(w - w_*\right) \tag{10}$$

Notice that truncating the gradient at the linear term corresponds to approximating the cost function around the optimal point by a quadratic curve. The Hessian matrix of this quadratic performance surface is $R/2$, where R is given in (11).

$$R = \frac{\partial \nabla V_a(w_*)}{\partial w} = \frac{\partial^2 V_a(w_*)}{\partial w^2} = \frac{(a-1)}{s^a N^a} \sum_j \left[\sum_i k(\Delta e_{w_*}^{ji}) \right]^{a-3} \cdot$$
$$\left\{ \begin{array}{l} (a-2)\left[\sum_i k'(\Delta e_{w_*}^{ji}) \cdot (x_i - x_j) \right] \cdot \left[\sum_i k'(\Delta e_{w_*}^{ji}) \cdot (x_i - x_j)^T \right] \\ + \left[\sum_i k(\Delta e_{w_*}^{ji}) \right] \cdot \left[\sum_i k''(\Delta e_{w_*}^{ji}) \cdot (x_i - x_j)(x_i - x_j)^T \right] \end{array} \right\} \tag{11}$$

Now, defining a new weight vector space $\overline{w} = w - w_*$ whose origin is translated to the optimal solution w_*, we can rewrite the linearized dynamics of the weight equations in the vicinity of the solution in terms of the step size and the Hessian matrix as given in (12). These are coupled equations for the translated weights.

$$\overline{w}(n+1) = \left[I + \eta \ R \right] \overline{w}(n) \tag{12}$$

In order to obtain decoupled equations, we rotate the vector space by defining $v = Q^T \overline{w}$, Q being the orthonormal matrix consisting of the eigenvectors of R. Thus, the uncoupled dynamics for the translated and rotated weights are

$$v(n+1) = \left[I + \eta \ \Lambda \right] v(n) \tag{13}$$

where Λ is the diagonal eigenvalue matrix with entries ordered in correspondence with the ordering in Q. From this set of equations, we can isolate the dynamics of the weight vector along each mode of the matrix R. Specifically, for the i^th^ mode, the dynamic equation will only depend on the i^th^ eigenvalue of R by

$$v_i(n+1) = \left[1 + \eta \ \lambda_i \right] v_i(n), \qquad i = 1, \ldots, l \tag{14}$$

Note that, since R is the Hessian of the performance surface evaluated at a maximum point, its eigenvalues are negative. For a stable dynamics, all of the coefficients in the n equations of (14) must be inside the unit circle, that is $\left| 1 + h \ \ l_i \right| < 1$. This, results in the following bound for the step size for stability.

$$0 < h \ \ < \frac{1}{\max_i \left| l_i \right|} \tag{15}$$

This condition is very similar (as it should be) to what we obtain for the MSE criterion [2, 4]; except, we consider the eigenvalues of the Hessian of information potential instead of those of the autocorrelation matrix of the input.

At this point, it also becomes possible to talk about time constants of the modes in the neighborhood of the optimum point. We can determine an approximate time constant for each individual mode whose dynamic equations are governed by (14). Specifically, for the k^th^ mode, we write

$$(1 + h \ \ l_k) = e^{-1/t_k} \tag{16}$$

from which the time constant is evaluated as

$$t_k = \frac{-1}{\ln(1 + h l_k)} \approx \frac{-1}{h l_k} = \frac{1}{h \left| l_k \right|} \tag{17}$$

The time constants allow us to compare the convergence times of different modes. In order to evaluate the overall convergence speed, one must consider the slowest mode, which corresponds to the largest time constant, i.e. the one that corresponds to the largest (smallest absolute value) eigenvalue.

EFFECT OF KERNEL SIZE AND ENTROPY ORDER ON EIGENVALUES

Understanding the relationship between the eigenvalues and the kernel size and α is crucial to maintain the convergence of the algorithm under changes in these parameters. One practical case where this relationship becomes important is when we adapt the kernel size during the training. Motivated by the link between the information potential estimator in (3) and the convolution smoothing method of global optimization [7, 12], we suggested starting from a large kernel size and decreasing it to a nominal value during adaptation. It is then possible to use steepest ascent to maximize the information potential, but still guarantee convergence to the global maximum by smoothing of the cost function by convolution by a suitable functional. Since in this approach, the kernel size is decreased, we need to now how to adapt the step size to achieve faster learning in the initial phase of adaptation (by using a larger step size) and stable convergence in the final phase (by using a smaller step size).

For convenience, we repeat here how to observe the dilation in weight space. Consider the information potential expression in (3) evaluated in terms of the unit-size kernel by the help of (4). It is clear that the introduction of a kernel size other than unity causes the error samples to be treated as if they are divided by σ. Thus in the error-space the location of the global optimum is scaled along a radial direction from the origin. The exception is the case of zero error because then the global optimum is the origin in the error space and the optimal solution does not change with kernel size. Since the adaptive topologies used in practice are mainly contractive or volume-preserving structures, the dilation/stretching effect is directly translated to the weight-space. This property will be observed in the behavior eigenvalues under changing kernel size.

As an example consider the case where we evaluate the quadratic information potential using Gaussian kernels. In this case, the Hessian matrix simplifies to

$$R=\frac{1}{s^2N^2}\sum_j\left[\sum_i k''(\Delta e_{w*}^{ji})(x_i-x_j)(x_i-x_j)^T\right] \tag{18}$$

Observe from (8) that as σ increases, $\Delta e_{w*}^{ji} \to 0^-$, therefore, $k''(\Delta e_{w*}^{ji}) \to 0^-$ with speed $O(s^{-6})$. This is faster than the reduction rate of the denominator, which is $O(s^{-2})$, hence overall, the eigenvalues of R approach 0^-. This means, one can use a larger step size in steepest ascent, and still get stable convergence to the global maximum. In fact, this result can be generalized to any kernel function and any α. The dilation is a direct cause of the increase in eigenvalues towards zero.

The analysis of the eigenvalues for varying α is more complicated. In order to estimate the behavior of the eigenvalues under changing α, we will exploit the following well-known result from linear algebra relating the eigenvalues of a matrix to its trace. For any matrix R, whose eigenvalues are given by the set $\{l_i\}$, the following identity holds.

$$\sum_i l_i = tr(R) \tag{19}$$

Now consider the general expression of R given in (11). The trace of R is easily computed to be as given below in (20). The eigenvalues of R are negative and the dominant component, which introduces this negativity, is the term in the last line of (20). The negativity arises naturally since we use a differentiable symmetric kernel, and since at w_* the entropy is small, the error samples are close to each other and the second derivative evaluates as a negative coefficient. Now let's focus on the term which involves the $(a-3)$-power in the first line of (20). Since all other terms vary linearly with α, this term will dominantly affect the behavior of the trace when α is varied. Consider the case where σ is large enough such that the small entropy causes the kernel evaluations in the brackets to be close to their maximum possible values and the sum therefore exceeds one. In that case, the power of the quantity in the brackets will increase exponentially with increasing α, thus regardless of the terms affected linearly by a, the overall trace value will decrease (increase in absolute value). As a result, a narrower valley towards the maximum will appear. Consequently, the upper bound on the step size for stable convergence will be reduced.

$$tr(R) = \frac{(a-1)}{s^{a} N^{a}} \sum_{j} \left[\sum_{i} k\,(\Delta e^{ji}_{w_*}) \right]^{a-3} \cdot \left\{ \begin{aligned} &(a-2)\sum_{k} \left[\sum_{i} k'(\Delta e^{ji}_{w_*}) \cdot (x_{ik} - x_{jk}) \right]^{2} \\ &+ \left[\sum_{i} k\,(\Delta e^{ji}_{w_*}) \right] \cdot \left[\sum_{i} k'(\Delta e^{ji}_{w_*}) \cdot \left(\sum_{k} (x_{ik} - x_{jk})^{2} \right) \right] \end{aligned} \right\} \tag{20}$$

On the other hand, if the kernel size is (very) small such that the sum in the brackets is less than one, then the $(a-3)$-power of this quantity will decrease, thus result in a wider valley towards the maximum in contrast to the previous case. However, in practice we do not want to use a very small kernel size, as it will increase the variance of the Parzen pdf estimation [11].

In fact, there is another approach that directly demonstrates how the eigenvalues of R will decrease with increasing α and vice versa. Consider expression (11) again. Since at the operating point the error entropy is small and the difference between error samples is close to zero, the sums involving the derivative of the kernel function are approximately zero. Under the conditions mentioned in the previous paragraph, all the terms involving α remain as scalar coefficients that multiply a matrix, whose eigenvalues are negative. With the same arguments on how increasing α increases these coefficients, we conclude that the eigenvalues of the matrix R will increase in absolute value for a large kernel size and decrease for a small kernel size.

In this section, we have investigated the effect of the entropy order α and the kernel size σ on the eigenvalues of the Hessian matrix of the information potential criterion around the optimal solution. We have seen that the entropy order can have differing effects depending on the kernel size. As for the effect of kernel size, we have observed that as it increases, the quadratic approximation to the cost function has larger eigenvalues. This points out a wider region of validity for the approximation. We remark that our conclusions in this section do not only apply to the eigenvalues of R, but they generalize to how these two parameters affect the volume of the region where our quadratic approximation is valid. Indeed, as we will demonstrate with an example in the following section, the absolute values of the eigenvalues and this volume are inversely proportional. This result is imperative from a practical point of view, because it explains how the structure of the performance surface can be manipulated by adjusting these parameters.

CASE STUDY

In the previous section, we had analyzed the effect of the kernel size and the entropy order on the location of the global optimum and the eigenstructure of the performance surface around this point. In this section, we present a case study to visualize the conclusions derived above. Consider a time series prediction example, where the training set consisting of input and corresponding desired output pairs is constructed from the impulse response of a single pole (at 0.9) transfer function with a unit gain. The FIR filter length is chosen to be two so that we can show equilevel contours in the weight space. With this configuration, the prediction task becomes finding the best weights according to the information potential criterion to predict the next sample in the time series as a linear combination of the previous two values. The training set consists of 20 input-output pairs. We evaluate the information potential expression given in (3) for two values of α (specifically 2 and 4) and two values of σ (specifically 1 and 2) on a grid in the weight space. Fig. 2 summarizes these results.

Upon observation of Fig. 2a, we notice that along both rows (upper row uses α=2, lower row uses α=4) as we increase the kernel size σ, the equilevel ellipses expand showing us that the eigenvalues of R have decreased in absolute value and we can use a larger step size and still converge stably. Along the columns we observe the behavior of the performance surface under changing α. Notice that increasing the order α resulted in reduction of the ellipses (larger magnitude eigenvalues). This case study demonstrates one other interesting property of the information potential cost function. We have previously proved that MSE is a special case of the information potential criterion under proper choices of the kernel function (Gaussian kernels satisfies those requirements). This example clearly shows that if we indefinitely increase the kernel size, the performance equilevel contours will merely consist of ellipses, hence the equivalence with MSE in the limit, however, this process introduces a higher bias in Parzen window estimate of the pdf. We are interested in the volume of the region where the contours are ellipses, as all the theory we have derived in the previous sections, with the linear approximation to the gradient in (10), is accurate there.

Besides the above stated reasons, there are practical issues for which the volume of the region with elliptical contours is crucial. The kernel size is an active parameter that controls the amount of dilation/smoothing effect as proven by Erdogmus and Principe [7]. Thus, we study the effect of the kernel size on the structure of the cost function, especially in the vicinity of the global maximum, in more detail in this case study. Since in this toy problem the solution achieves zero error by setting the weight vector to [0.9,0], the dilation occurs around this point when the kernel size is varied.

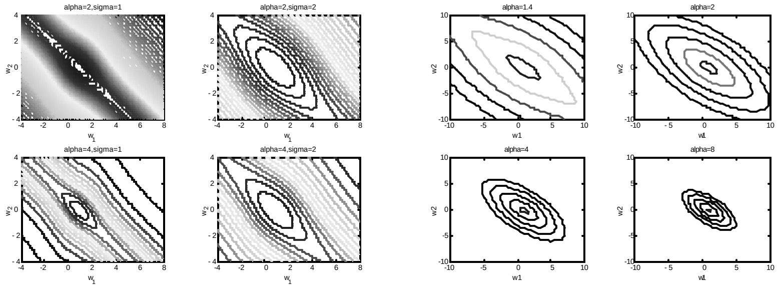

Figure 2. a) Information potential contour plots for different α and σ values b) Information potential contours at 95% of max possible value for σ=1,3,5,7,9

Fig. 2b demonstrates the increase of volume where the quadratic approximation is valid as the kernel size increased, for various values of α. Notice also that as α increases, the ellipses corresponding to matching kernel sizes decrease in size in accordance with our conclusions about the effect of α for large σ. Each ellipse in the subplots corresponds to an equilevel contour for a specific kernel size, where the value of the information potential is 95% of its maximum possible value attained when the weights are optimum. For a fixed α, the ellipses become larger as σ increases. As a conclusion, this example displays both the dilation property, thus the effect on the eigenvalues of R, and the proportionality between the volume of the quadratic region of the information potential and the kernel size.

CONCLUSIONS

Earlier we have demonstrated the superiority of the error entropy over the mean square error as the performance criterion in a variety of applications including equalization, and chaotic time series prediction in our previous studies. This paper was however, was dedicated to the investigation of the convergence properties of the steepest descent adaptation in FIR training, when the minimization of the error entropy, equivalently maximization of the information potential for α>1, is utilized as the performance criterion. We have derived the difference equations, which govern the dynamics of the modes of the weight vector in a neighborhood (whose volume can also be controlled) of the global maximum of the information potential surface. These equations gave rise to an upper bound for the step size for stability and a time constant expression to approximate the convergence speed of the in this neighborhood of the optimal solution. We examined the effects of the two characteristic parameters, namely the entropy order α and the kernel size σ, on the structure of the performance surface around the global maximum and on the size of the region where the quadratic approximation, hence all the theoretical results, are accurate. It was found that, increasing α could both increase (for a small kernel size) or decrease (for a large kernel size) this volume, whereas increasing σ, in consistency with our previous observations on the dilation of the cost surface in the weight space, induced smaller eigenvalues in magnitude and resulted in a wider region of validity. Lastly, the case study we have presented demonstrates all these properties as well the equivalence between the information potential criterion and MSE in the limit, a result that was proven in a previous study. It should be noted, however, that the optimal solutions of the minimum entropy criterion and the MSE criterion are not the same except when we can achieve zero error over the complete training data set. Nevertheless, when the filter topology used is sufficient to approximate the target function accurately enough such that the errors become small, the kernel size we choose may appear as ‘large’ and in this case it is possible to obtain very close solutions from the two criteria.

Acknowledgments: This work was supported by the NSF grant ECS-9900394.

REFERENCES

[1] Wiener N., Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications, Cambridge, MA: MIT Press, 1949.
[2] Haykin, S., Adaptive Filter Theory, 3rd ed., NJ: Prentice Hall, Inc., 1996.
[3] Farhang-Boroujeny, B., Adaptive Filters: Theory and Applications, NY: John Wiley & Son Ltd., 1998.
[4] Widrow, B., S.D.Stearns, Adaptive Signal Processing, NJ: Prentice Hall, 1985.
[5] D. Erdogmus, J.C.Principe, “Comparison Of Entropy And Mean Square Error Criteria In Adaptive System Training Using Higher Order Statistics”, Proceedings of 2nd International Workshop on Independent Component Analysis (ICA 2000), Helsinki, Finland, June 2000.
[6] 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.
[7] D. Erdogmus, J.C.Principe, “Generalized Information Potential Criterion for Adaptive System Training,” submitted to IEEE Transactions on Neural Networks, Feb. 2001.
[8] J.C. Principe, D.Xu, J.Fisher, “Information Theoretic Learning,” in Unsupervised Adaptive Filtering, vol I, Simon Haykin Editor, NY: Wiley, 2000, pp 265-319.
[9] S. Amari, Differential–Geometrical Methods in Statistics, Berlin: Springer-Verlag, 1985.
[10] A. Renyi, Probability Theory, NY: American Elsevier Publishing Company Inc., 1970.
[11] E. Parzen, “On Estimation of a Probability Density Function and Mode”, in Time Series Analysis Papers, CA: Holden-Day, Inc., 1967.
[12] R.Y. Rubinstein, Simulation and the Monte Carlo Method, NY: John Wiley & Sons, 1981.