A NORMALIZED MINIMUM ERROR ENTROPY STOCHASTIC ALGORITHM

Seungju Han, Sudhir Rao, Kyu-Hwa Jeong, Jose Principe

CNEL, Department of Electrical and Computer Engineering, University of Florida, Gainesville, USA

ABSTRACT

We propose in this paper the normalized Minimum Error Entropy (NMEE). Following the same rational that lead to the normalized LMS, the weight update adjustment for Minimum Error Entropy (MEE) is constrained by the principle of minimum disturbance. Unexpectedly, we obtained an algorithm that not only is insensitive to the power of the input, but is also faster than the MEE for the same misadjustment, and also that is less sensitive to the kernel size. We explain these results analytically, and through system identification simulations.

1. INTRODUCTION

Information Theoretic Learning has become quite popular in recent years with the introduction of smooth sample estimators of entropy as proposed by Principe and collaborators [1],[2]. The nonparametric entropy estimator is based on Renyi's quadratic entropy, which can be applied directly to data collected from experiments and without imposing any a priori assumptions about the underlying probability density function (PDF). The goal of entropy in supervised learning follows the MSE framework. Given a set of input-desired signal pairs, the entropy of the error over the training dataset is minimized. The procedure is called Minimum Error Entropy (MEE) [3]. Extensive comparisons have already been done between MEE and MSE techniques by Erdogmus [2],[4] and this is not the goal of this paper.

The MEE algorithm requires a priori knowledge of the input process statistics (power and dynamic range) to select the learning rate μ for stability and convergence and the kernel size for good results. Since this knowledge is usually unavailable, the step size is normally estimated prior to beginning the adaptation process for a given misadjustemnt or speed of convergence, and the kernel size is estimated by Silverman's rule [5] or similar heuristics [2]. Changes in stepsize or kernel size if the input power fluctuates should be done for optimal performance, but are seldom performed leading to sub-optimal performance. The purpose of the paper is to propose an enhanced MEE algorithm where the selection of μ will be independent of the input power and where the effect of the kernel size will be minimized. When the MEE algorithm is modified in this manner, we refer to this as the normalized Minimum Error Entropy (NMEE).The paper is organized as follows. Section 2 summarizes the concept of MEE. We introduce NMEE derived from the solution of the constrained optimal solution and stability of NMEE in section 3. Section 4 deals with simulation results and finally we conclude in section 5.

2. SUMMARY OF MINIMUM ERROR ENTROPY

Suppose that the adaptive system is an FIR structure with a weight vector $\mathbf{w}$. The error samples are $e(n) = d(n) - \mathbf{w}(n)^T \mathbf{u}(n)$, where $d(n)$ is the desired response, and $\mathbf{u}(n)$ is the input vector. The error PDF at a point e is estimated using Parzen window estimation with a kernel function as

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

where $k_\sigma(\cdot)$ is kernel function with a kernel size σ. So, Renyi's quadratic entropy estimator for a set of discrete data samples becomes [1]:

$$H_{R2}(e) = -\log \int f^2(d)\,de = -\log V(e) \tag{2}$$

$$V(e) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k_{\sigma\sqrt{2}}(e(j) - e(i)) \le V(0) \tag{3}$$

Minimizing the entropy $H_{R2}(e)$ is equivalent to maximizing the information potential $V(e)$ since the log is a monotonic function.

For online training methods, the information potential can be estimated using Stochastic Information Gradient (SIG) as shown in (4), where the sum is over the most recent L samples at time n [6]. Thus for a filter order of length M, the complexity of MEE is equal to $O(ML)$ per weight update,

$$V(e(n)) \approx \frac{1}{L}\sum_{i=n-L}^{n-1} k_{\sigma\sqrt{2}}(e(n) - e(i)) \tag{4}$$

where $e(i) = d(i) - \mathbf{w}^T(n)\mathbf{u}(i)$, for $n - L \le i \le n$. Thus the MEE update is

$$\text{MEE}: \mathbf{w}(n+1) = \mathbf{w}(n) + \mu \nabla V(e(n)) \tag{5}$$

where assuming a Gaussian kernel the gradient is,

$$\nabla V(e(n)) = \frac{1}{2\sigma^2 L} \sum_{i=n-L}^{n-1} G_{\sigma\sqrt{2}}(e(n)-e(i))\{e(n)-e(i)\}\{\mathbf{u}(n)-\mathbf{u}(i)\} \tag{6}$$

3. NORMALIZED MINIMUM ERROR ENTROPY

3.1. NMEE as the solution to a constrained optimization problem

We will derive NMEE by analogy with NLMS, i.e. as a modification of the ordinary MEE in light of the principle of minimal disturbance [7]. The criterion of NMEE is formulated as constrained optimization:

$$\min \quad \left\|\mathbf{w}(n+1)-\mathbf{w}(n)\right\|^2 \tag{7}$$

$$\text{subject to} \quad V(0) - V(e_p(n)) = 0 \tag{8}$$

where $e_p(n) = d(n) - \mathbf{w}(n+1)^T \mathbf{u}(n)$ is the posterior error, and (8) translates the constraint of optimal performance in terms of the information potential.

When we solve the above constrained optimization problem using the method of Lagrange multipliers and applying the stochastic nature of the estimator, we finally get

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \eta \frac{e_a(n)\nabla V(e_p(n))}{\left(\nabla V(e_p(n))\right)^T \mathbf{u}(n)} \tag{9}$$

where $e_a(n) = d(n) - \mathbf{w}(n)^T \mathbf{u}(n)$ is the a priori error, and η is the normalized stepsize which can be proven to be between 0 and 2 for stability. In this update, there is an added difficulty because estimating $\mathbf{w}(n+1)$ requires $e_p(i)$. We propose to substitute $e_a(n)$ by $e_p(i)$ because we aim to minimize $\left\|\mathbf{w}(n+1)-\mathbf{w}(n)\right\|^2$. Therefore, we obtain the following weight update for NMEE,

$$\text{NMEE}: \mathbf{w}(n+1) = \mathbf{w}(n) + \eta \frac{e_a(n)\nabla V(e_a(n))}{\left(\nabla V(e_a(n))\right)^T \mathbf{u}(n)}. \tag{10}$$

3.2. Discussion of the NMEE solution

It is instructive to examine the Lagrangian and write it as

$$\lambda = \frac{4\sigma^2 L e_a(n)}{\sum_{i=n-L}^{n-1} \kappa(e_p(n)-e_p(i))\{e_p(n)-e_p(i)\}\{\left\|\mathbf{u}(n)\right\|^2 - \mathbf{u}^T(i)\mathbf{u}(n)\}}. \tag{11}$$

First, recall that in the LMS, the Lagrangian only depends upon the error and the norm of the input. For MEE we see that it depends not only on the error and the norm of the input, but also on the kernel size through the information force [1] given by

$$\frac{1}{2\sigma^2 L} \sum_{i=n-L}^{n-1} \kappa(e_p(n)-e_p(i))\{e_p(n)-e_p(i)\}. \tag{12}$$

This means that we can expect NMEE to be less dependent, when compared to the MEE algorithm, not only to the input power, but also to the kernel size because of the normalization by the information force.

Furthermore, we see that an extra error term appears in the numerator of the update which indicates that the speed of convergence will likely change. All these aspects follow from the particular constraint the MEE solution requires and also from the nonlinear nature of the error gradient.

4. SIMULATION AND DISCUSSION

In the simulations, we will experimentally verify the different adaptation performance of the newly introduced NMEE with respect to MEE. Simulations are particular useful in this case because the differences between the two algorithms are not as easy to translate to practical measures of convergence rate and misadjustment as in the LMS case. We consider two kinds of plant identification models, a moving-average model (MA), and an autoregressive model (AR). We analyze these problems using the stochastic information gradient (SIG). The adaptive weights were initialized to zero at each instance. Further, in order to make the result independent of the initial conditions, we performed Monte-Carlo simulations with 100 random inputs.

4.1. Example 1: Dependence on Input Power

We consider a moving-average model with transfer function given by (order = 9) [8]

$$H(z) = 0.1 + 0.2z^{-1} + 0.3z^{-2} + 0.4z^{-3} + 0.5z^{-4} + 0.4z^{-5} + 0.3z^{-6} + 0.2z^{-7} + 0.1z^{-8} \tag{13}$$

The FIR adaptive filter is selected with equal order. A standard method of comparing the performance in system identification is to plot the weight error norm since this is directly related to misadjustment [8]. In each case the power of the weight noise (averaged over 125 samples) was plotted versus the number of iterations performed. The input to both the plant and the adaptive filter is white Gaussian noise. In the first experiment, an unit power input was used (1P), whereas in the second experiment input with 10 x power (P) was selected. We choose a proper kernel size by using Silverman's rule in MEE ($\sigma_{mee} = 0.7$) and NMEE ($\sigma_{nmee} = 1$), respectively.

In this perfect identification, we can exactly track the output of the plant. Fig.1 shows the plot of the weight error

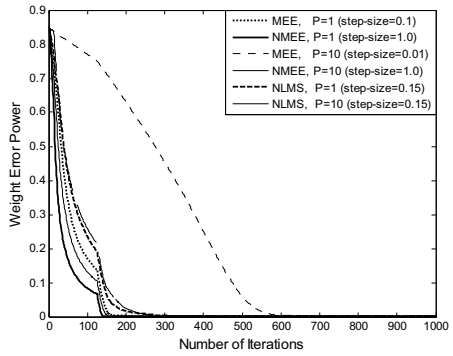

Fig 1. Average weight error power of three algorithms (MEE, NMEE and NLMS) for MA(9) with different input powers (P=1, 10)

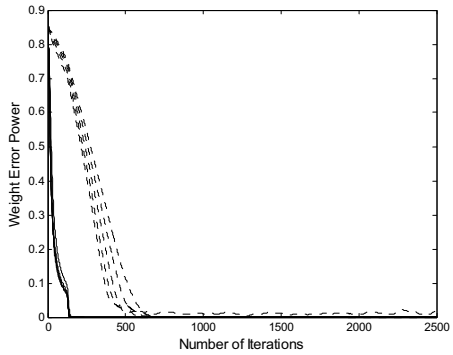

Fig 2. Average weight error power for MA(9) with different kernel sizes, solid lines – NMEE, dashed line-MEE, in the case of white Gaussian input with variance 10, eigenspread S=1. (Kernel Sizes of left to right dashed lines: 0.7, 1, 1.3, 1.6, 2)

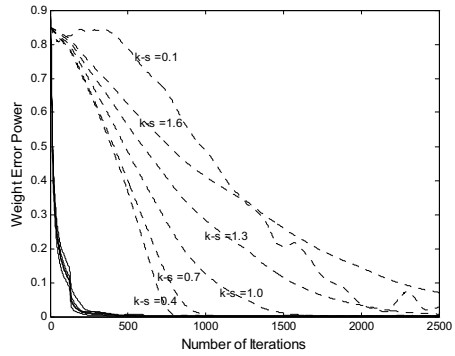

Fig 3. Average weight error power for MA(9) with different kernel sizes, solid lines – NMEE, dashed line-MEE, in the case of colored Gaussian input with variance 15, eigenspread S=550.

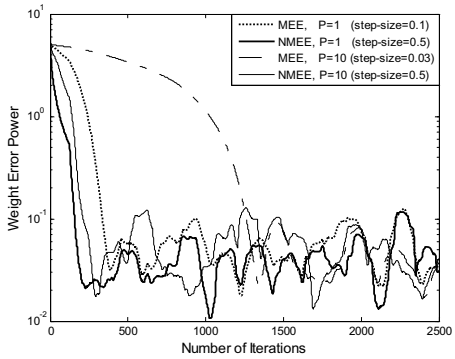

Fig 4. Average weight error power of two algorithms (MEE and NMEE) for AR(1) in case of best results

norm for a moving average model. We choose as large a step size as possible within the range of MEE stability.

For the unit input power case, both MEE and NMEE converge in 190 iterations with basically the same misadjustment (10^{-16} or below). To guarantee the stability of MEE adaptation for 10x input power, the step size is chosen almost 10 times smaller, while it remains at the same value for NMEE. NMEE just takes 190 iterations to converge with a misadjustment of 1.562×10^{-10} as compared to MEE which takes nearly 700 iterations with a misadjustment of 1.243×10^{-8}. We can therefore conclude that the NMEE is insensitive to the input power, as expected. We also present in Fig 1 the convergence of the NLMS for comparison purposes (misadjustment of 10^{-16}), and conclude that the NMEE is faster than the NLMS.

4.2. Dependence on Kernel Size and Speed of Adaptation

Selecting a proper kernel size is one important factor because it influences the performance of the MEE algorithm. A proper kernel size depends on the error statistics [2]. The effect of kernel sizes on both MEE and NMEE is shown in Fig. 2 and 3. Fig. 2 shows the weight error power curves with different kernel sizes when the input data is white Gaussian with zero mean, 10x unit variance and with an eigenvalue spread ratio (S) of 1. Fig. 3 shows the results in the case of colored Gaussian input, whose variance is 6 and eigenvalue spread ratio (S) is 550. We observe that the performance of MEE is sensitive to the kernel size and this sensitivity increases when the eigenvalue spread of the input signal increases. However, NMEE shows a much more uniform performance with different kernel sizes even when the input has a large dynamic range. The misadjustment of NMEE in the worst case is almost the same as that of MEE in the best case. Furthermore, the kernel size of 0.7 and 1 for MEE and NMEE respectively are found to be optimal, giving the lowest misadjustment of 6.711×10^{-10} and 8.299×10^{-14} in the case when S=1 and 3.865×10^{-8} and 1.472×10^{-12} when S=550.

Another important aspect of these experiments is the different speed of convergence between the MEE and the

NMEE. Fig. 2 clearly shows that there are two sets of curves, one for each algorithm. We could interpret this by hypothesizing that the learning rates are not compatible with the same misadjustment. However, a closer look shows that the NMEE is faster than the MEE and provides a smaller misadjustment (8.299×10^{-14} versus 6.711×10^{-10}). Therefore the NMEE converges faster than the MEE.

4.3. Example 2: AR(1)

As a second case study, we select for the unknown plant a first order autoregressive model,

$$H_2(z) = \frac{1}{1-0.9z^{-1}}. \qquad (14)$$

The unknown system $H_2(z)$ is approximated by a moving average model with 16 taps. Two sets of white Gaussian noise inputs with unit variance and 10 x unit variance are applied to both the plant and the adaptive filter. We choose a proper kernel size for MEE ($\sigma_{mee}=0.7$) and NMEE ($\sigma_{nmee}=1$), respectively.

Fig. 4 shows the best results for MEE and NMEE of the average weight error power to identify the first order autoregressive model. The 10 x input variance produces the larger misadjustment of 0.1 , while the unit input variance has a misadjustment of 0.05 . We can observe that the MEE with 10 x input variance converges after 1300 iterations. The NMEE results are very different. Indeed, the curves for the unit and 10 x power input are very close to each other, and converge within 400 iterations. These results show that even when the identification is done with a residual error, the NMEE converges faster and is basically independent of the input power.

5. CONCLUSIONS

We have presented a new algorithm, the normalized Minimum Error Entropy (NMEE) by minimizing the weight change subject to the constraint of optimal information potential. Despite the similarity of the derivation with the LMS, NMEE is quite different from MEE, and performs better with respect to two major points: it is less sensitive to the kernel size, and converges faster.

The goal of achieving independence of the stepsize on the input power is corroborated in all the experiments performed with the NMEE. This is a convenience and may be important when handling nonstationary signals with large dynamic range. But the advantage of the NMEE over the MEE does not stop here. Indeed, NMEE is less sensitive to the kernel size, which is known to affect the performance of the MEE. This is a major advantage for ITL algorithms because it helps the selection of the extra parameter introduced by this class of cost functions. The second aspect is the faster convergence of the NMEE versus the MEE. We claim this feature because we controlled the misadjustment of each algorithm and found out the NMEE converges faster for the same misadjustment when compared with the MEE. The reason why NMEE rate of convergence outperforms the MEE should be thoroughly investigated theoretically, because it is unexpected.

For the problems tested (linear models), NMEE also outperforms the NLMS convergence rate as briefly presented in Fig 1. One might wonder why use another criterion when the optimal weight solution is the same as MSE. Our previous results show that the MEE optimal solution is superior in the identification of nonlinear plants when the model system is nonlinear [3], but it defaults to the MSE solution when the model is linear [2]. However, since the goal of this paper is to compare the dynamics of learning, we thought it was more appropriate to restrict the analysis to liner models since they converge to the same optimal weight vector.

Acknowledgments: This work was partially supported by NSF grant ECS-0300340. We thank the anonymous reviewers for their constructive suggestions which will be useful in further investigating this novel algorithm.

6. REFERENCES

[1] J.C. Principe, D. Xu and J. Fisher, "Information Theoretic Learning," in S. Haykin, *Unsupervised Adaptive Filtering*, Wiley, Newyork, vol I, pp. 265-319, 2000.

[2] D. Erdogmus, "Information Theorectic Learning: Renyi's Entropy and its Applications to Adaptive System Training," Ph.D Dissertation, University of Florida, Gainesville, FL, 2002.

[3] D. Erdogmus, J.C. Principe, "An Entropy Minimization algorithm for Supervised Training of Nonlinear Systems," *IEEE trans. of Signal Processing*, vol.50, no.7, pp. 1780-1786, July 2002.

[4] D. Erdogmus and J.C. Principe, "Generalized Information Potential Criterion for Adaptive System Training," *IEEE Trans. of Neural Networks*, vol.13, no.5, pp. 1035-1044, Sep. 2002.

[5] B.W.Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, 1986.

[6] D. Erdogmus, J.C. Principe, K.E.Hild II, "Online entropy manipulation: Stochastic Information Gradient," *IEEE Signal Processing Letters*, vol. 10, no. 8, pp. 242-245, Aug. 2003.

[7] Simon Haykin, *Adaptive Filter Theory*, Prentice Hall, Upper Saddle River, 4th edition, 2001.

[8] E. Walach and B. Widrow, "The Least Mean Fourth (LMF) Adaptive Algorithm and its Family," *IEEE trans. of Inf. Theory*, vol. IT 30, no.2, pp. 275-283, March 1984.