ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 451-456

Recursive Least Squares for an Entropy Regularized MSE Cost Function

Deniz Erdogmus[1], Yadunandana N. Rao[1], Jose C. Principe[1]
Oscar Fontenla-Romero[2], Amparo Alonso-Betanzos[2]

[1] Electrical Eng. Dept., University of Florida, Gainesville, FL 32611, USA
[2] Dept. Computer Science, University of A Coruna, 15071 A Coruna, Spain

Abstract. Minimum MSE plays an indispensable role in learning and adaptation of neural systems. Nevertheless, the instantaneous value of the modeling error alone does not convey sufficient information about the accuracy of the estimated model in representing the underlying structure of the data. In this paper, we propose an extension to the traditional MSE cost function, a regularization term based on the incremental errors in model output. We demonstrate the stochastic equivalence between the proposed regularization term and the error entropy. Finally, we derive an RLS-type algorithm for the proposed cost function, which we call recursive least squares with entropy regularization (RLSER) algorithm. The performance of RLSER is shown to be better than RLS in supervised training with noisy data.

1. Introduction

The interest in MSE roots from the analytical simplicities that arise when the adaptive structure is a simple linear neuron (ADALINE) [1]. In this case, the optimal MSE solution for the weight vector is given by the Wiener-Hopf equation [2]. In on-line adaptation, the weights are updated by LMS or RLS [2]. Since RLS uses a second-order weight update, it converges much faster than LMS, which uses a stochastic gradient update. The success of RLS lies in its low-complexity updating of the inverse of the input covariance matrix. This avoids the requirement of a matrix inversion. Nevertheless, RLS also has shortcomings. It is susceptible to noise in the training data, which results in a biased estimate of the covariance matrix, and its inverse.

The MSE and associated algorithms fail to take into consideration the behavior of the modeling error over time. For an adaptive system to successfully extract the model behind the data, it is necessary to account for the behavior of error in consecutive updates. This could allow robust noise rejection and facilitate continuity of the success of the estimated model. One way to achieve such performance improvements through simple modifications to the traditional MSE cost function, which is given by $E[e_k^2]$, is to introduce a regularization term to modify the cost function to

$$J = E[e_k^2] + \lambda E[(e_k - e_{k-1})^2] \quad (1)$$

The proposed cost function tries to minimize the MSE, while it pays attention to maintaining the variation between consecutive errors small. Especially near the

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 451-456

optimal solution, we expect this behavior to help provide additional robustness to present noise in the training data. In this paper, we will show that there still exists an analytic solution for the optimal weights of the combined cost.

The regularization term in (1) could be also obtained from Shannon's entropy estimated using Parzen windows. We will demonstrate how the instantaneous increments in error are related stochastically to the entropy of the error. Recently, we have proposed the use of minimum error entropy (MEE) as an alternative information theoretic learning criterion [3]. We have shown that minimizing the error entropy in a supervised training scenario is equivalent to minimizing information theoretic divergence measures between the joint densities of the input-output and input-desired signal pairs (for Shannon's entropy, this is the Kullback-Leibler divergence) [4]. A stochastic gradient for entropy, called the stochastic information gradient (SIG) was derived and applied successfully to learning problems [5]. For ADALINE, in the special case of Gaussian kernels in Parzen windowing, it could be shown that SIG becomes a stochastic gradient for the regularization term in (1), as well [5]. SIG has a very interesting structure that exploits relations between sequential errors in time. Hence, when applied in conjunction with LMS, can be thought as a regularization of the jaggedness of the trajectory in the weight space.

The convergence speed and the misadjustment of these gradient-based algorithms are susceptible to the selected step size. In this paper, motivated by this regularization property, we derive an entropy regularized RLS algorithm for the proposed regularized MSE cost function. In the following, we will present the entropy estimator and the steps that lead to SIG and the RLSER (recursive least squares with entropy regularization) algorithm. The performance of RLSER is compared with that of RLS in ADALINE training where both input and desired samples are contaminated with white noise.

2. Entropy and Regularization

In general, the parametric error pdf in supervised learning is not known. In such circumstances, non-parametric approaches are used. Parzen windowing is a non-parametric density estimation method that is simple and smooth pdf estimates can be obtained [6]. Given the *iid* samples $\{x_1,\ldots,x_N\}$, the Parzen window estimate for the underlying pdf $f_X(.)$ is obtained by $\hat{f}_X(x) = (1/N)\sum_{i=1}^{N} \kappa_\sigma(x - x_i)$, where $\kappa_\sigma(.)$ is the kernel function and σ is the kernel size. Typically, Gaussian kernels are preferred. Shannon's entropy for a random variable X with pdf $f_X(.)$ is defined as [7]

$$H(X) = -\int_{-\infty}^{\infty} f_X(x) \log f_X(x) dx \tag{2}$$

Given *iid* samples, the entropy of X can be estimated using [5]

$$\hat{H}(X) = -(1/N)\sum_{j=1}^{N} \log\left((1/N)\sum_{i=1}^{N} \kappa_\sigma(x_j - x_i) \right) \tag{3}$$

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 451-456

In order to arrive at this estimator, we write (2) as an expectation and substitute the Parzen estimate. In on-line applications of this entropy cost function, it is desirable to utilize a stochastic approximation to the entropy, and to its gradient if steepest descent is used. Viola proposed a similar entropy estimator, however, he suggested dividing the samples into two subsets: one for estimating the pdf, and one for evaluating the expectation in the entropy [8]. To obtain a stochastic estimator, we approximate the expectation by evaluating the argument at the most recent sample [5]. Then, estimating the pdf using the most recent L samples, the $O(L)$ complexity stochastic entropy estimator becomes $\overline{H}_k(X) = -\log(1/L)\sum_{i=1}^{L}\kappa_\sigma(x_k - x_{k-i})$.

For supervised training of an ADALINE with weight vector $w \in \Re^n$, given the input-desired training sequence (u_k,d_k), where $u_k \in \Re^n$ and $d_k \in \Re$, the instantaneous error is given by $e_k = d_k - w_k^T u_k$. Then, for the specific choice of L=1 and a Gaussian kernel function, the stochastic gradient of the error entropy (SIG) with respect to the weights becomes ($e_{k-i} = d_{k-i} - w_k^T u_{k-i}$)

$$\partial\overline{H}(X)/\partial w = -(e_k - e_{k-1})(u_k - u_{k-1})/\sigma^2 \tag{4}$$

Note that the simplified SIG in (5) is also the stochastic gradient for the cost function $J = E[(e_k - e_{k-1})^2]/(2\sigma^2)$. Taking the derivative after dropping the expectation from the definition of J and evaluating the argument using the most recent two samples, we arrive at (4). The approximate equivalence between the minimizations of J and H makes sense, because when the entropy of the error samples is small, then they are close to each other, thus the kernel evaluations in (3) can be approximated using the quadratic Taylor approximation to the kernel. Hence, the cost function becomes simply the squared instantaneous increments in the error. We have shown that when an ADALINE is trained under noisy data using the MSE or MEE with batch learning methods, it learns the underlying weights more accurately in the finite-sample case, when trained using MEE rather than MSE [5]. Motivated by this, we aim to incorporate the noise rejection capability of entropy into the RLS algorithm by combining the two cost functions for ADALINE training as described in (1).

3. RLS with Entropy Regularization

In practice, there is noise present in the data that leads to biased solutions when MSE is the optimality criterion. This might cause large differences between consecutive error samples. In order to counter this problem, the regularized MSE function in (1) could be used, where $0 < \lambda$. It can be shown that (1) is equivalently written as

$$J = \gamma_d(0) + w^T R w - 2w^T P + \lambda\left(\gamma_{\dot{d}}(0) + w^T R_1 w - 2w^T P_1\right) \tag{5}$$

where $\gamma_d(0) = \mathrm{var}(d_k)$, $\gamma_{\dot{d}}(0) = \mathrm{var}(d_k - d_{k-1})$, $R = E[u_k u_k^T]$, $P = E[d_k u_k]$, $R_1 = E[(u_k - u_{k-1})(u_k - u_{k-1})^T]$, $P_1 = E[(d_k - d_{k-1})(u_k - u_{k-1})]$. Taking the gradient of (5) and equating to zero yields a Wiener-type solution.

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 451-456

$$\frac{\partial J}{\partial w} = 2Rw - 2P + \lambda\left(2R_1 w - 2P_1\right) = 0 \Rightarrow w^* = (R + \lambda R_1)^{-1}(P + \lambda P_1) \quad (6)$$

Letting $Q = (R + \lambda R_1)$ and $V = (P + \lambda P_1)$, we obtain the following recursions.

$$Q(k) = Q(k-1) + (2\lambda u_k - \lambda u_{k-1})u_k^T + u_k(u_k - \lambda u_{k-1})^T \quad (7)$$

At this point, we employ the Sherman-Morrison-Woodbury identity, which is

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + D^T A^{-1} B)^{-1} D^T A^{-1} \quad (8)$$

Substituting A=$Q(k$-1), $B = [(2\lambda u_k - \lambda u_{k-1}) \quad u_k]$, C=I_{2x2}, $D = [u_k \quad (u_k - \lambda u_{k-1})]$, we obtain $BCD^T = \ldots = (2\lambda u_k - \lambda u_{k-1})u_k^T + u_k(u_k - \lambda u_{k-1})^T$. Therefore, the recursion for the inverse of Q becomes

$$Q^{-1}(k) = Q^{-1}(k-1) - Q^{-1}(k-1)B(I_{2x2} + D^T Q^{-1}(k-1)B)^{-1} D^T Q^{-1}(k-1) \quad (9)$$

Notice that the update of Q^{-1} requires a matrix inversion of only the 2x2 matrix $(I_{2x2} + D^T Q^{-1}(k-1)B)$. The recursion for V, is much simpler to obtain.

$$V(k) = (1 - 1/k)V(k-1) + (1/k)[(1 + 2\lambda)d_k u_k - \lambda d_k u_{k-1} - \lambda d_{k-1} u_k] \quad (10)$$

The overall complexity of this algorithm is $O(n^2)$, the same as the classical RLS algorithm. The optimal weights are estimated at each step by $w_k = Q^{-1}(k)V(k)$. We call this algorithm recursive least squares with entropy regularization (RLSER), since it minimizes a regularized MSE cost function, where the importance given to the regularization portion is determined by the parameter λ.

4. Performance of RLSER and Comparison with RLS

In the preceding section, we developed the RLSER algorithm, which has the same computational complexity as RLS, yet it also minimizes the square of the derivative of the error, leading to smoother transition in the weight space in noisy situations. In this section, we will perform Monte Carlo simulations to compare RLS and RLSER in supervised training of ADALINEs with noisy training data. For simplicity, the length of the adaptive system was set to that of the reference model. Linear structures with length L = 5, 10, 15, and 20 were used. For each filter length, noisy (white Gaussian) data samples with signal-to noise ratio (SNR) of SNR = 0, 5, 10, 15, 20, 25, and 30 dB, with noise on both the inputs and the desired output were generated. With the noisy training data, we created training sets of lengths N = 25, 50, 75, and 100. For each combination (L,N,SNR), 100 Monte Carlo simulations were performed using RLS and RLSER. In each simulation, the number of epochs is selected such that the total number of updates is 10000. For RLSER, we selected λ=1.

The results are summarized in Fig. 1 and Fig. 2. The performance of these two algorithms are compared based on their ability to yield ADALINE parameters that are close to the reference model parameters. In order to measure the divergence of the produced weight estimates from their actual values, we use the angle between the estimated and actual vectors (ideal would be zero) and the norm ratios of the

estimated weights and the actual weights (ideal would be one). Observing the results in Fig. 1, we notice that almost for all combinations of L, N, and SNR, RLSER outperforms RLS in acquiring the direction of the actual weight vector from noisy data. The exception occurs only for large L, small N, and small SNR, but even then, the performances are very similar. Similarly, from Fig. 2, we deduce that RLSER is better in all cases except for the same situation as above. Although, we do not present simulations regarding the effect of λ, due to lack of space, we believe that by modifying it, RLSER could be tuned to outperform RLS for even these situations.

5. Conclusions

Motivated by the noise rejection capabilities of the previously proposed minimum error entropy criterion in supervised training, in this paper, we proposed an extension to the traditional MSE criterion, which we called the regularized MSE. For this new cost function, we derived an RLS-type algorithm; called recursive regularized least squares (RLSER). This extended cost function for supervised ADALINE training includes the squares of the differences of consecutive modeling errors. We have demonstrated the superiority of the new criterion and the associated RLSER algorithm over RLS in noise rejection, when training data is corrupted with white noise, through extensive Monte Carlo simulations. RLSER is designed to have the same computational complexity as RLS, and is based on the same principles. Therefore, their convergence speed and computational requirements are identical.

Acknowledgments: This work was partially supported by grant NSF-ECS-9900394 and Xunta de Galicia project PGIDT-01PXI10503PR.

References

[1] B. Widrow, S.D. Stearns: *Adaptive Signal Processing*, Prentice-Hall, (1985).
[2] S. Haykin: *Introduction to Adaptive Filters*, MacMillan, (1984).
[3] D. Erdogmus, J.C. Principe: An Entropy Minimization Algorithm for Supervised Training of Nonlinear Systems, IEEE Trans. Sig. Proc., 50, 1780-1786, (2002).
[4] D. Erdogmus, J.C. Principe: Generalized Information Potential Criterion for Adaptive System Training, IEEE Trans. Neural Net., 13, 1035-1044, (2002).
[5] D. Erdogmus: *Information Theoretic Learning: Renyi's Entropy and its Applications to Adaptive System Training*, PhD Diss., Univ. Florida, (2002).
[6] E. Parzen: On Estimation of a Probability Density Function and Mode, In *Time Series Analysis Papers*, Holden-Day, (1967).
[7] C.E. Shannon, W. Weaver: *The Mathematical Theory of Communication*, University of Illinois Press, (1964).
[8] P. Viola, N.N. Schraudolph, T.J. Sejnowski: Empirical Entropy Manipulation for Real-World Problems, Proc. NIPS 8, 851-857, (1995).

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 451-456

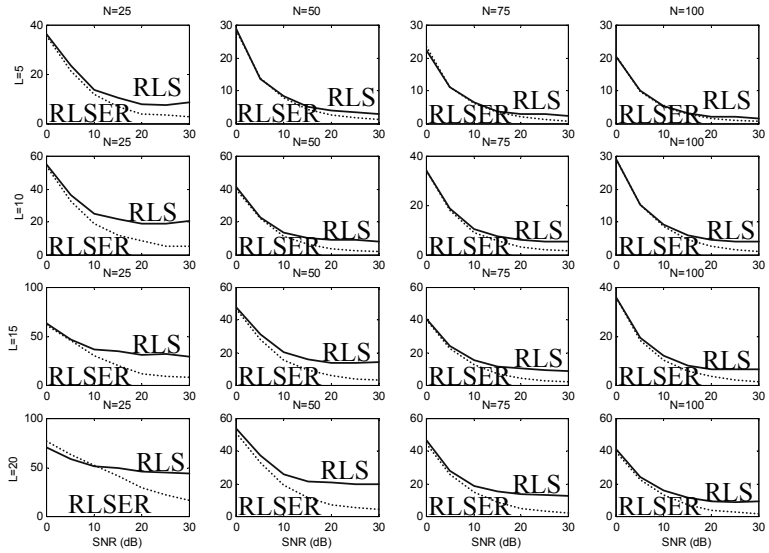

Figure 1. RMS angle (degrees) between the estimated and actual weight vectors for various combinations of filter length (L) and sample size (N) versus SNR (dB) using RLS (solid) and RLSER (dashed).

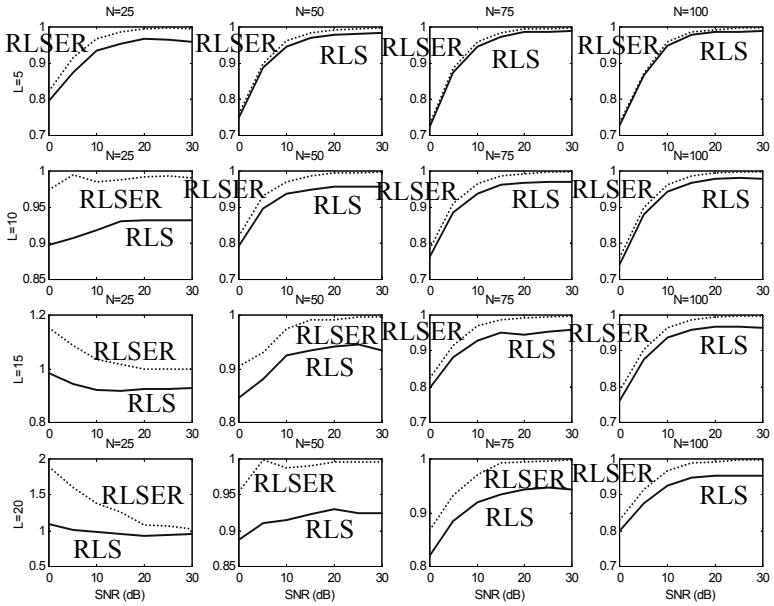

Figure 2. RMS norm ratio of the estimated weight vector to actual weight vector for various combinations of filter length (L) and sample size (N) versus SNR (dB) using RLS (solid) and RLSER (dashed).