# Renyi 二次熵在適應性無限脈波響應濾波器的全域最佳化
# Global Optimization of Adaptive IIR Filters Using Renyi's Quadratic Entropy

賴慶安
Ching-An Lai

*Abstract*

In this paper, we propose an adaptive IIR filter training algorithm, referred to as the ITL algorithm, which is based on minimizing Renyi's quadratic entropy by utilizing a non-parametric pdf estimator, Parzen windowing. By exploiting the kernel size used in the Parzen window estimator, we force the proposed algorithm to converge to the global minimum of the performance surface. We compare the performance of the ITL algorithm with that of the LMS-SAS and NLMS algorithms with decreasing step size capable of finding the global optimum and conclude in simulations that the ITL algorithm is superior.

*Keywords*: adaptive IIR filter, Renyi's quadratic entropy, MSE, global optimization.

*摘要*

這篇文章建議使用資料理論學習演算法則到適應性無限脈波響應濾波器訓練演算法則上。此演算奠基於極小化 Renyi 的二次熵。此二次熵是由 Parzen 視窗(一種非參數型機率密度函數估測器)所估算。借著運用視窗估測器中核心大小的特性而能強迫演算法則收斂到表現面的絕對極佳點。我們比較資料理論學習演算法則和二種最小均方根演算法則(機率估算迴旋平滑法、標準化法)的表現。模擬結果顯示資料理論學習演算法則比較優良。

*關鍵詞*：適應性無限脈波響應濾波器，Renyi 二次熵，最小均方差，全域最佳化

## I. INTRODUCTION

Adaptive infinite impulse response (IIR) filters have the advantage of approximating pole-zero models more accurately than the equivalent order finite impulse response (FIR) filters, thereby reducing the computational cost in terms of the number of coefficients to be estimated. Unfortunately, adaptive IIR filtering has some drawbacks, such as instability during the adaptation process, slow convergence and local minimum in the cost function [1], [2], [3], [4]. Therefore, conventional gradient-based algorithms, e.g., the least mean square (LMS) algorithm [1], might converge to one of these local minimum, resulting in an unacceptable suboptimal solution.

Several methods have been proposed for the global optimization of adaptive IIR filters [5], [6], [7]. Srinivasan et al. [8] have used a stochastic approximation for the convolution smoothing (SAS) technique in order to obtain a global optimization algorithm [7], [9], [10]. They showed that smoothing can be approximated by the addition of a variable perturbing noise source to the LMS algorithm. We modify this perturbing noise by multiplying it with its cost function. The modified algorithm, which is referred to as the LMS-SAS algorithm in this paper, results in better performance when compare to the original algorithm by Srinivasan et al. Since we use the instantaneous (stochastic) gradient instead of the expected value of the gradient, error in estimating the gradient naturally occurs. This gradient estimation error can also be utilized to act as the perturbing noise. Consequently, another approach for global IIR filter optimization is the normalized LMS (NLMS) algorithm. The behavior of the NLMS algorithm with decreasing step size is similar to that of the LMS-SAS algorithm from a global optimization perspective.

The mean square error criterion has been extensively used in the theory of adaptive systems [11]. This is due to its analytical simplicity and the common assumption of Gaussian distributed signals. However, the assumption of Gaussian distribution is not always sufficiently accurate. Therefore, a criterion that considers higher-order statistics is necessary for the training of adaptive systems, in general.

樹德科技大學資訊工程系
[*] Corresponding author. E-mail: chingan@mail.stu.edu.tw
Department of Computer Science and Information Engineering, Shu Te University, Kaohsiung, Taiwan, 82445 R.O.C.

Manuscript received 24 March 2006; revised 24 May 2006; accepted 3 August 2006

Shannon [12] first introduced the entropy of a given probability distribution function, which provides a measure of the average information in that distribution. By utilizing the Parzen window estimator [13], we can estimate the pdf directly from a set of samples. It is quite straightforward to apply the entropy criterion to the system identification framework [14], [15]. The pdf of the error signal between the desired signal and the output signal of adaptive filters must be as close as possible to a delta distribution, δ(.). Hence, the supervised training problem becomes an entropy minimization problem, as suggested by Erdogmus and Principe [14]. The kernel size of the Parzen window estimator is an important parameter in the global optimization procedure that we are about to discuss. It was conjectured in [14] that for a sufficiently large kernel size, the local minimum of the error entropy criterion can be eliminated. It was suggested that starting with a large kernel size, and then slowly decreasing this parameter to a predetermined suitable value, the training algorithm can converge to the global minimum of the cost function. The error entropy criterion considered in [14], however, does not consider the mean of the error signal, since entropy is invariant to translation. A modification to the error entropy criterion, in order to take this point into account, has been successfully applied to echo cancellation by global optimization [25]. The proposed criterion is then shown to exhibit the conjectured global optimization behavior in the training of kautz filters. In this paper, we discuss theoretically, in detail, the global optimization behavior.

This paper is organized as follows: Section II reviews the LMS-SAS and NLMS algorithms. Section III derives the ITL algorithm. Monte Carlo simulation results from the training IIR filter model are given in Section IV. Section V discusses the global minimum searching capability of the ITL algorithm for the special case studied in the previous section. We conclude the paper and summarize our results in Section VI.

## II. LMS-SAS AND NLMS ALGORITHM

We have modified the algorithm by Srinivasan et al. [8] to obtain global optimization algorithms. This class of algorithms involve minimizing the MSE between the desired output, *d(n)*, and the output of the adaptive filter, *y(n)* (as depicted in Fig. 1). The MSE objective function can be written as:

$$\xi(n,\theta)=\frac{1}{2}E\left\{e^{2}(n,\theta)\right\}=\frac{1}{2}E\left\{\left[d(n)-y(n)\right]^{2}\right\} \tag{1}$$

Where $E$ denotes statistical expectation and the output can be described as:

$$y(n)=\sum_{i=0}^{M}a_{i}x(n-i)+\sum_{j=0}^{N}b_{i}y(n-i) \tag{2}$$

Let the parameter vector

$$\theta=\left[\mathrm{a}_{0},\cdots,a_{M,}b_{1},\cdots,b_{N}\right]^{\mathrm{T}} \tag{3}$$

and data vector $\Phi(n)$ be defined as:

$$\Phi(n)=\left[\mathrm{x}(n),\cdots,x(n-M),y(n-1),\cdots,y(n-N)\right]^{\mathrm{T}} \tag{4}$$

From the approximated relation

$$\nabla_{\theta}\xi(\mathrm{n},\theta-\eta(\mathrm{n}))\triangleq\nabla_{\theta}\xi(\mathrm{n},\theta)-\eta(\mathrm{n})\ [8],$$

we can derive the LMS-SAS algorithm as:

$$\theta(n+1)=\theta(n)+\mu(n)e(n)\nabla_{\theta}y(n)+\mu(n)e(n)\eta(n) \tag{5}$$

Where $\eta(n)$represents the added random number, $\mu(n)$is the learning rate, which decreases over the iterations, and *e(n)* is the error between the desired output and the adaptive filter output.

The gradient $\nabla_{\theta}y(n)$ is defined as:

$$\nabla_{\theta}y(\mathrm{n})=\Phi(\mathrm{n})+\sum_{\mathrm{j}=1}^{\mathrm{N}}\mathrm{b}_{\mathrm{j}}\nabla_{\theta}y(n-j) \tag{6}$$

The advantage of the LMS-SAS algorithm Eq.(5) is that a single error source needs to be added to the desired response, unlike a different noise source to each weight as in previous work. Alternatively, the NLMS algorithm is given by:

$$\theta(n+1)=\theta(n)+\frac{\mu(n)}{\left\|\nabla_{\theta}y(n)\right\|^{2}}\nabla_{\theta}y(n)e(n) \tag{7}$$

Due to the noisy gradient estimate, the expected behavior of the NLMS algorithm is similar to that of the LMS-SAS algorithm. Hence, the NLMS algorithm with decreased step size has the capacity to converge to the global minimum. For LMS-SAS and NLMS algorithms, interested readers can refer to [26].

## III. ITL ALGORITHM

In this section, we derive the information-theoretic learning (ITL) algorithm based on Renyi's quadratic entropy. If the criterion for adaptation is minimum error entropy as proposed in [14] the steepest descent training algorithm for adapting the IIR filter weights becomes:

$$\theta(n+1)=\theta(n)-\mu(n)\frac{\partial H(e)}{\partial\theta} \tag{8}$$

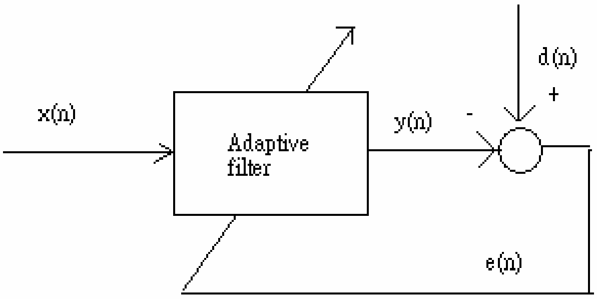


Fig. 1 Adaptive filter model

where Shannon's entropy [12] given by

$$H_s(\varepsilon) = -\int_{-\infty}^{\infty} f_\varepsilon(\xi)\log\ f_\varepsilon(\xi)d\xi \tag{9}$$

could be utilized. In practice, an analytical expression of the error pdf is not available and nonparametric estimators for Shannon's entropy require heavy computations. If we utilize the (quadratic) entropy definition by Renyi, which is

$$H_{R2}(\varepsilon) = -\log \int_{-\infty}^{\infty} f_\varepsilon^2(\xi)d\xi \tag{10}$$

as an alternative, it is possible to derive a low-complexity estimator using Parzen windowing with Gaussian kernels [15]. In the system identification scheme using adaptive IIR filters, clearly, our goal is to adjust the coefficients of the adaptive IIR filter such that the error pdf, $f_e$ is close to a delta distribution, δ(.). Hence, we consider minimizing the integrated square error $I_{ED}$ between the error pdf and the targeted delta distribution.

$$\begin{aligned} I_{ED}(f_e) &= \int_{-\infty}^{\infty}(f_e(\varepsilon)-\delta(\varepsilon))^2 d\varepsilon \\ &= \int_{-\infty}^{\infty} f_e(\varepsilon)^2 d\varepsilon - 2\mathrm{f_e}(0) + c \end{aligned} \tag{11}$$

were c stands for the portions of this Euclidean distance measure that do not depend on the weights of the adaptive system. Notice that, the integral of the square of the error pdf appears exactly as in the definition of Renyi's quadratic entropy. Therefore, it can be estimated directly from its

$$f_e(\varepsilon) = \frac{1}{N}\sum_{i=1}^{N}\kappa(\varepsilon - e_i, \sigma^2) \tag{12}$$

If $N \to \infty$, then $\hat{f}_e(\varepsilon) = f_e(\varepsilon) * \kappa(\varepsilon, \sigma^2)$ [14], where * denotes the convolution operator. Thus, utilizing a Parzen window estimator for the error pdf is equivalent to add an independent random noise with the pdf $\kappa(\varepsilon, \sigma^2)$ to the error. The error, with the additive noise, becomes *d-y+n = (d+n)-y*. This is similar to injecting a random noise to the desired signal as suggested by Wang et al. in [16]. The advantage of our approach is that we do not explicitly generate noise samples. We simple take advantage of the estimation noise produced by the Parzen estimator, which as demonstrated above, works as an additive, independent noise source. The kernel size, which controls the variance of the hypothetical noise term, should be annealed during the adaptation, just like the variance of the injected noise in [16]. From the injected noise point of view, the algorithm behaves similar to the well-known stochastic annealing algorithm; the noise which is added to the desired signal backpropagates through the error gradient, resulting in perturbations in the weight updates in a random manner. However, since our algorithm does not explicitly use a noise signal, its operation is more similar to convolution smoothing. For a sufficiently large kernel size, the local minimum of the ITL criterion are eliminated by smoothening of the performance surface. Thus, by starting with a large kernel size, the algorithm can approach to the global minimum, avoiding any local minimum that would have existed if the kernel size was to be small. Since the global minimum of the error entropy criterion with large kernel size does not, in general, coincide with the true global minimum, annealing the kernel size is required. This is equivalent to gradually reducing the amount of the noise injected to the desired signal to a small suitable value. At the end, the algorithm with the small kernel size can converge to the true global minimum.

By substituting the Parzen window estimator for the error pdf in the integral of Eq.(11), and recognizing that the convolution of two Gaussian functions is also a Gaussian, we obtain the ITL criterion as (after dropping all the terms that are independent of the weights):

$$I_{ED}(f_\varepsilon) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\kappa(e_i - e_j, 2\sigma^2) - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\kappa(e_i, \sigma^2) \tag{13}$$

The gradient vector $\nabla_\theta I_{ED}(f_e)$ to be used in the steepest descent algorithm is obtained as

$$\begin{aligned} &\nabla_\theta I_{ED}(f_\varepsilon) = \frac{1}{2N^2\sigma^2}\sum_{i=1}^{N}[(e_i - e_j)\kappa(e_i - e_j, 2\sigma^2) \\ &(\nabla_\theta y(n-i) - \nabla_\theta y(n-j) - 2e_i\kappa(e_i, \sigma^2)\nabla_\theta \mathrm{y(n-i)}] \end{aligned} \tag{14}$$

## IV. SIMULATION RESULTS

In this section, we adapt all the parameters and compare the performances of the LMS, LMS-SAS,
NLMS, and ITL algorithms in terms of their capability to obtain the global optimum in a system identification framework using a simple one section IIR.

In this example, we will identify the following unknown system.

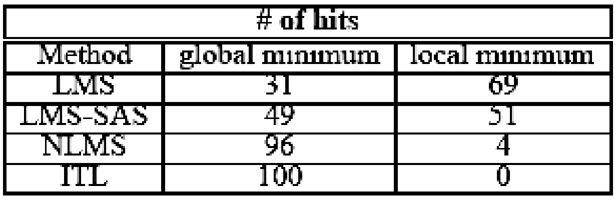
$$\mathrm{H(z)} = \frac{1 - 1.1z^{-1}}{1 - 1.1314\ z^{-1} + 0.25\ z^{-2}} \tag{15}$$

by a reduced order pole-zero adaptive filter of the form

$$\mathrm{H(z)} = \frac{1 + \mathrm{bz}^{-1}}{1 - az^{-1}} \tag{16}$$

The goal is to determine the values of the coefficients {*a*, *b*} of the above equation, such that the proper cost function (MSE or ITL) is minimized. The reference IIR filter is excited with a white Gaussian input signal with zero mean and unit variance. The cost function has two minimum, and we would like to avoid the local solution (see Fig. 2). The step size is chosen to be $\mu(n) = 0.01$ for LMS, $\mu(n) = 0.001$ for the ITL algorithm, and a linearly decreasing step size of $\mu(n) = 0.1(1 - 0.5\times10^{-5}n)$ is used for the LMS-SAS and NLMS algorithms. The kernel size of the Parzen window estimator is also a linearly decreasing function of iterations, $\mu(n) = 3(1 - 0.5\times10^{-5}n) + 0.25$. Table 1 shows the number of times the global and local minimum are hit by various algorithms. The results are given for a set of 100 Monte Carlo simulations where random initial conditions of are used in each run. For demonstration purposes, a single weight-track for each algorithm is given in

Table 1 System identification of adaptive filter



| # of hits | | |
|---|---|---|
| Method | global minimum | local minimum |
| LMS | 31 | 69 |
| LMS-SAS | 49 | 51 |
| NLMS | 96 | 4 |
| ITL | 100 | 0 |

Fig. 2, 3, 4, and 5, where is initialized to a point near the local minimum. Notice that the ITL algorithm achieves the global minimum 100 % of the time, while the other algorithms are susceptible to the local minimum.

## V. DISCUSSION

In the previous example, if the input signal is set to have a Gaussian distribution, $N(\mu_x,\sigma_x^2)$, then the desired signal will also be Gaussian, $N(\mu_d,\sigma_d^2)$. The output signal of the adaptive filter will be a Gaussian as well, $N(\mu_y,\sigma_y^2)$. Under these assumptions, the analytical expression for the ITL criterion for the example studied in the previous section can be determined to be:

$$\mathrm{I}_{ED}(f_e)=\frac{1}{\sqrt{4\pi\sigma_e^2}}-\frac{2}{\sqrt{2\pi\sigma_e^2}}e^{-\frac{\mu_e^2}{2\sigma_e^2}} \qquad (17)$$

Here,

$$\mu_e=\mu_d-\mu_y$$
$$\sigma_e^2=R_d^2(0)+R_y^2(0)-2R_{dy}(0)+\sigma^2 \qquad (18)$$

where $R_d(t)$ and $R_x(t)$ are the variance of desired output signal and input signal, respectively, $R_{dy}(t)$ is the covariance of desired output signal and input signal, and $\sigma_e^2$ increases by $\sigma^2$, which is corresponding to the Gaussian kernel function of the Parzen window estimator. Fig. 6 shows the equilevel contours of the analytical expression in Eq. (17) for the ITL criterion. Notice that for large the performance surface is smoothen and there is a single minimum (top left). When θ is decreased the local minimum at the bottom right shows up, and so any initial condition on the top half will not converge to the global minimum. The convergence characteristics of the adaptation process for the filter coefficients towards the global optimum are shown in Fig. 6. In the beginning of the adaptation process, the estimated error variance $\sigma_e^2$. is large due to the significantly large value of the kernel size, $\sigma^2$, in the Gaussian kernel function of the Parzen window estimator. Therefore, the first term of the right hand side of Eq. (17) is considerably smaller than the second term. Thus it can be neglected in the beginning stage of the adaptation process. We observe that the second term concentrates more tightly around $\mu_e=\mu_d-\mu_y=0$ associated with the increasing $\sigma_e^2$, i.e., the increasing $\sigma^2$. The straight line in Fig. 6 (b) is the line of $\mu_e=\mu_d-\mu_y=0$. It is clear from Fig. 6,

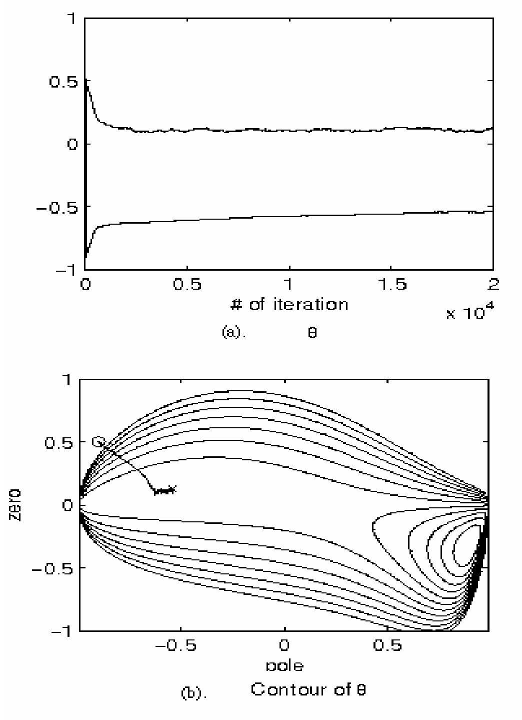


Fig. 2 Convergence characteristic of $\theta$ in LMS algorithm

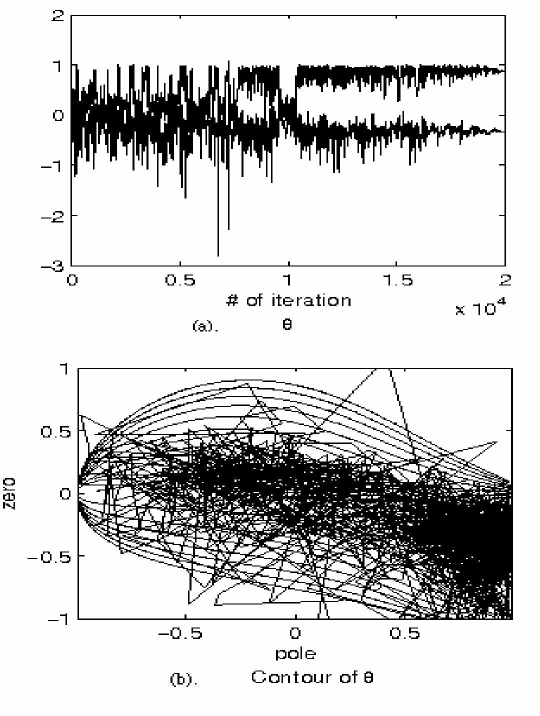


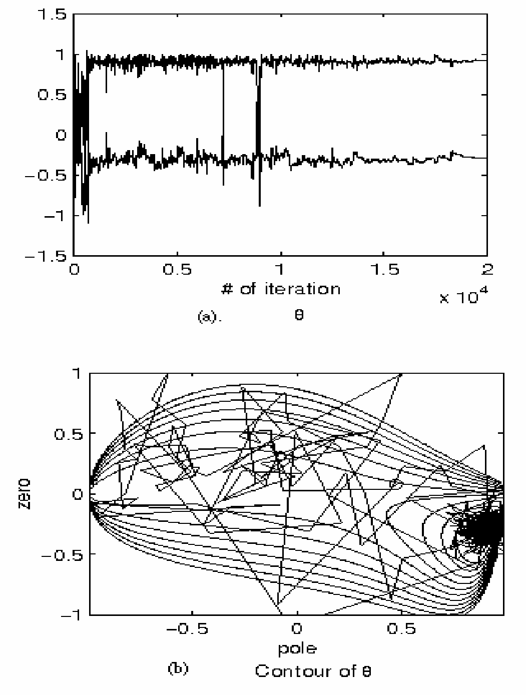
Fig. 3 Convergence characteristic of $\theta$ in LMS-SAS algorithm

that the weight-track of the ITL algorithm converges towards the line of $\mu_e = \mu_d - \mu_y = 0$ as we predicted in the theoretical analysis given above. When the size, $\sigma^2$, of the Gaussian kernel function slowly decreases during adaptation, the ITL cost function will gradually converge back to the original one, which might exhibit local minimum.

## VI. CONCLUSION

Global optimization of adaptive IIR filters is a major problem that prevents the wide use of these filters in adaptive signal processing tasks. The main difficulty in training IIR filters is the existence of local optima when the poles are adapted to minimize the commonly used MSE criterion. In this paper, we proposed the information theoretic adaptation criterion based on Renyi's quadratic entropy. The proposed ITL criterion and kernel annealing approach allowed stable adaptation of the poles to their global optimal values.

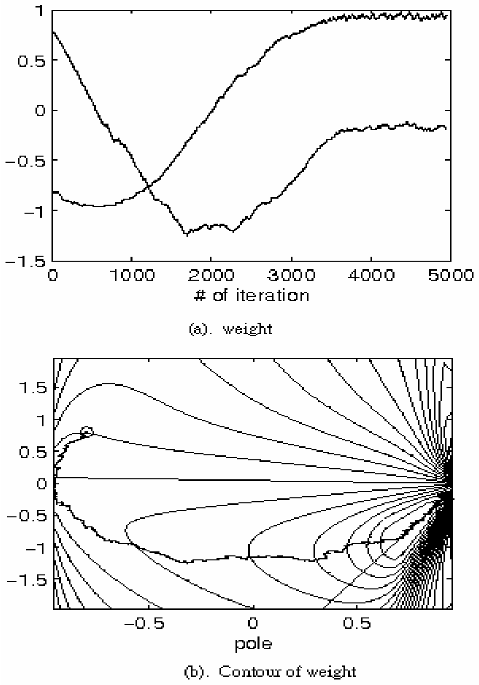
We have investigated the performance of the proposed criterion and the associated steepest descent algorithm in IIR filter adaptation. Based on a previous conjecture that proposed annealing the kernel size in the non-parametric estimator of Renyi's entropy to achieve global optimization, we have designed the proposed information theoretic learning algorithm, which is shown to converge to the global minimum of the performance surface for various adaptive filter topologies. The proposed algorithm successfully adapted the filter poles avoiding local minimum 100 % of the time. This behavior has been found in many other cases.



(a). θ

(b) Contour of θ

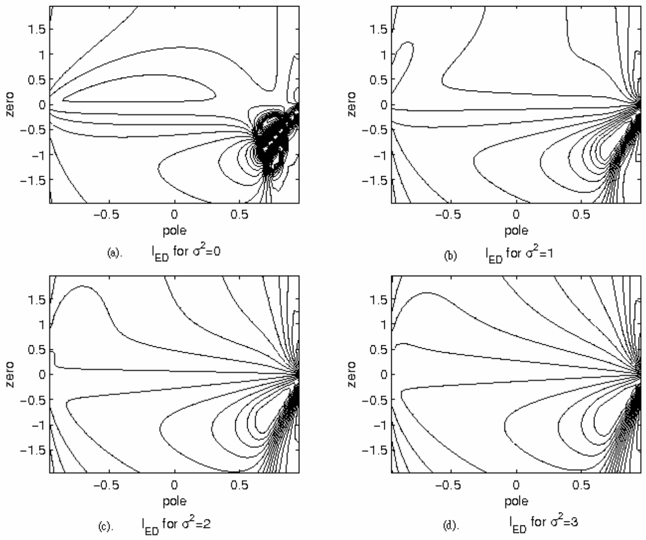
Fig. 4 Convergence characteristic of $\theta$ in NLMS algorithm



(a). weight

(b). Contour of weight

Fig. 5 Convergence characteristic of $\theta$ in ITL algorithm



(a). $I_{ED}$ for $\sigma^2$=0

(b) $I_{ED}$ for $\sigma^2$=1

(c). $I_{ED}$ for $\sigma^2$=2

(d). $I_{ED}$ for $\sigma^2$=3

Fig. 6 ITL contour plots for different $\sigma$

## REFERENCES

[1] B. Widrow and S. D. Stearns, *Adaptive Singal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1985.

[2] J. J. Shynk, "Adaptive IIR filtering," *IEEE Acoust., Speech, Signal Processing Mag.*, pp. 4-21, 1989.

[3] C. R. Johnson Jr., "Adaptive IIR filtering: current results and open issues," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 237-250, March 1984.

[4] P. A. Regalia, *Adaptive IIR filtering in signal processing and control*, Mercel Dekker Inc., 1995.

[5] G. A. Williamson and S. Zimmermann, "Globally convergent adaptive IIR filter based on fixed pole locations," *IEEE Trans. Signal Processing*, vol. 44, pp. 1418-1427, June 1996.

[6] A. Luk, S. C. Ng, S. H. Leung, C. Y. Chung and W. H. Lau, "The genetic search approach-a new learning algorithm for adaptive IIR filtering," *IEEE Signal Processing magazine*, pp. 38-46, Nov. 1996.

[7] P. M. Pardalos and R. Horst, *Introduction to global optimization*, Norwood, MA: Kluwer, 1989.

[8] K. Srinivasan, W. Edmonson, J. Principe and C. Wang, "A global least square algorithm for adaptive IIR filtering," *Proc.IEEE Trans. Circuit and Sys.*, vol. 45, pp. 379-383, March 1998.

[9] M. A. Styblinski and T. S. Tang, "Experiments in nonconvex optimization: stochastic approximation with function smoothing and stimulated annealing," *Neural Networks*, vol. 3, pp. 467-4833, 1990.

[10] H.Robins and S.Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, pp. 400-407, 1951.

[11] S. Haykin, *Introduction to adaptive filters*, MacMillan, NY, 1984.

[12] C. E. Shannon, "A mathematical theory of communication," *Bell System Tech. J.*, vol. 27, pp. 379-423, 623-653, 1984.

[13] E. Parzen, "On the estimation of a probability density function and the mode," *Ann. Math. Stat.*, vol. 33, pp. 1065, 1962.

[14] D. Erdogmus and J. C. Principe, "Generalized information potential criterion for adaptive system training," *IEEE Transactions on Neural Networks*, (to appear) Sep. 2002.

[15] K. Hild, D. Erdogmus and J.C. Principe, "Blind source separation using renyi's mutual information," *IEEE Signal Processing Letters*, vol. 8, no. 6, pp. 174-176, June 2001.

[16] C. Wang and J. C. Principe, "Training neural networks with additive noise in the desired signal," *IEEE Trans. Neural Networks*, vol. 10, no. 6, pp. 1511-1517, Nov. 1999.

[17] R. Roberts and C. Mullis, *Digital Signal Processing*, Addison-Wesley, 1987.

[18] W. H. Kautz, "Transient synthesis in the time domain," *IRE Trans. Circuit Theory*, vol. 1, pp. 22-39, Sep. 1954.

[19] P. W. Broome, "Discrete orthonormal sequences," *J. Assoc. Comput. Machinery*, vol. 12, no. 2, pp. 151-168, Dec. 1965.

[20] B. deVries, J. Principe and P. Oliveira, "The gamma filter: a new class of adaptive IIR filters with restricted feedback," *IEEE Trans. Signal Processing*, vol. 41, no. 2, pp. 649-656, Feb. 1993.

[21] T. O. Silva, "Optimality conditions for truncated kautz networks with two periodically repeating complex conjagates poles," *IEEE Trans. Automatic Contr.*, vol. 40, pp. 342-346, Feb. 1995.

[22] B. Wahlberg, "System identification using Kautz models," *IEEE Trans. Automatic Control*, vol. 39, no. 6, pp. 1276-1282, June 1994.

[23] D. Erdogmus and J. C. Principe, "An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems," *IEEE Trans. Signal Processing*, vol. 50, no. 7, pp. 1780-1786, July 2002.

[24] D. Erdogmus and J. C. Principe, "An on-line adaptation algorithm for adaptive system training with minimum error entropy: stochastic information gradient," in *Intl. Conf. on ICA and Signal Separation*, San Diego, CA, pp. 7-12, Dec. 2001.

[25] C. A. Lai, D. Erdogmus and J. C. Principe, "Echo cancellation by global optimization of kautz filter using an information theoretic criterion," *IEEE ICASSP*, 2003.

[26] C. A. Lai, "NLMS algorithm with decreasing step size for adaptive IIR filters," *Signal Processing,* vol. 82, pp. 1305-1316, 2002.