Adaptive FIR Filtering under Minimum Error/Input Information Criterion ⋆

Badong Chen* Jinchun Hu* Hongbo Li* Zengqi Sun*

** State Key Laboratory of Intelligent Technology and Systems,*
Department of Computer Science and Technology,
Tsinghua University, Beijing, 100084, P.R. China
(Tel: 86-10-62777704; e-mail: chenbd04@mails.tsinghua.edu.cn).

Abstract: In this paper, we use the mutual information between error/input as the cost function for adaptive filtering. For the finite-impulse response (FIR) filter, the connections between the minimum error/input information (MEII) criterion and traditional mean-square error (MSE) criterion are investigated. We show that, for Gaussian case, the MEII criterion is equivalent to the well-known orthogonality condition. Based on the MEII criterion and kernel density estimation, we derive a stochastic gradient algorithm. Simulation results emphasize the effectiveness of this new algorithm.

1. INTRODUCTION

In traditional estimation and filtering theory, for reasons of mathematic convenience, the minimum mean square error (MMSE) criterion is widely used (Kailath [2000], Haykin [1996]). However, in most situations, the system would be nonlinear and non-Gaussian, thus the MMSE criterion fails to extract all the information in the error signals. Recently, a novel criterion has been proposed to solve the optimal filtering problems, in which the entropy of the error is minimized (Erdogmus [2003], Wolsztynski [2005], Guo [2006]). The entropy, which measures the average information contained in a random variable, is related to various statistical behaviors (Cover [1991], Ihara [1993]). Numerical examples suggest that the minimum error entropy (MEE) criterion could be able to achieve a better error distribution (Erdogmus [2000]). However, in the case of continuous random variable, the differential entropy is just a relative quantity (Ihara [1993]). This leads to the fact that the differential entropy is not always positive and its minima may approach to $-\infty$. Unlike the entropy measure, the mutual information between two continuous random variables is identical to the limit of the mutual information between their quantized versions. This makes mutual information positive definite and reach its global minima (zero) if and only if the two random variables are independent (Cover [1991], Ihara [1993]). Further, the mutual information is invariant under monotonous distortion of random variables. Thus, the application of mutual information shall yield a more robust performance. In the present paper, we use the mutual information between error/input as the cost function for the adaptive filtering. First, we give a brief introduction to the entropy and mutual information (see section 2). And then, we investigate the connections between the minimum error/input information (MEII) criterion and traditional mean-square error (MSE) criterion, some interesting results are obtained (see section 3). Further, under the MEII criterion, we derive a novel adaptation algorithm (see section 4). And finally, the performance of this new algorithm is demonstrated by Monte Carlo simulations, in comparison with the well-known LMS algorithm (see section 5).

Notation: The superscript $(.)^T$ denotes the transpose. $\det(.)$ denotes the determinant of a matrix. $\|X\|$ denotes the Euclidean norm of vector X , i.e. $\|X\| = \sqrt{X^TX}$. $E(.)$ denotes the expectation operator. Throughout this paper nats are used as the information units and log denotes the natural logarithm.

2. ENTROPY AND MUTUAL INFORMATION

Consider a d-dimensional continuous random variable $X \in R^d$ with probability density function (PDF) $p_X(x)$, the differential entropy is defined by (Cover [1991])

$$h(X) = -\int\limits_{R^d} p_X(x)\log p_X(x)\,dx \qquad (1)$$

The differential entropy $h(X)$ measures the dispersion of the random vector X . The smaller the entropy $h(X)$, the more concentrated the density function $p_X(x)$. If X is a Gaussian random vector with PDF $p_X(x) = \left((2\pi)^{d/2}|\Sigma|^{1/2}\right)^{-1}\exp\left(-\frac{1}{2}\times(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$, where μ is the mean and Σ is the covariance matrix, the differential entropy is

$$h(X) = \frac{1}{2}\log\left((2\pi e)^d|\Sigma|\right) \qquad (2)$$

For random variables X and Y, with joint probability density function $p_{XY}(x,y)$, the conditional differential entropy is given by

$$h(X|Y) = -\int p_{XY}(x,y)\log p_{X|Y}(x|y)\,dxdy \qquad (3)$$

Here the quantity $h(X|Y)$ can be explained as the remained uncertainty of X after we are informed of the

⋆ This work was supported in part by the National Science Foundation (NSF) of P.R. China under grants 60674053 and 60604010.

output of Y. Based on the entropy measures, the mutual information between X and Y is then defined by

$$I\,(X;Y) = h\,(X) - h\,(X\,|Y\,) \tag{4}$$

The mutual information $I\,(X;Y)$ describes the amount of information about X contained in Y. It is a general measure of statistical dependence between random variables. In addition, the conditional mutual information between X and Y given Z is

$$I\,(X;Y\,|Z\,) = \int p(x,y,z)\log\left(\frac{p(x,y\,|z\,)}{p(x\,|z\,)p(y\,|z\,)}\right)dxdydz$$

Some important properties of mutual information are listed as follows (Ihara [1993]).

Lemma 1. $I\,(X;Y) \geq 0$, with equality if and only if X and Y are independent.

Lemma 2. If $f\,(.)$ and $g\,(.)$ are continuous and strictly monotonic function, then $I\,(X;Y) = I\,(f(X);g(Y))$.

Lemma 3. $I\,(X;Y_1,\cdots,Y_n) = \sum_{i=1}^{n} I\,(X;Y_i\,|Y_1,\cdots,Y_{i-1})$.

Lemma 4. Let X, Y and (X,Y) be respectively m, n, $m+n$ dimensional Gaussian vectors, with covariance matrices A,B, C , then

$$I\,(X;Y) = \frac{1}{2}\log\frac{\det A\det B}{\det C} \tag{5}$$

3. MINIMUM ERROR/INPUT INFORMATION CRITERION

For an adaptive finite-impulse response (FIR) filter (Kailath [2000], Haykin [1996]), we can define the input data vector $X\,(k)$ and the filter parameter (or weight) vector $W\,(k)$ as

$$\begin{cases} X(k) = [x(k), x(k-1), \cdots, x(k-m+1)]^T \in R^m \\ W(k) = [w_1(k), w_2(k), \cdots, w_m(k)]^T \in R^m \end{cases} \tag{6}$$

where $m-1$ is called the number of order. The output of the filter is given by

$$y(k) = W^T(k)X(k) \tag{7}$$

Denote $d(k)$ the desired signal, the error signal $e(k)$ is

$$e(k) = d(k) - y(k) = d(k) - W^T(k)X(k) \tag{8}$$

Conventional filtering algorithm is usually designed to minimize the following mean-square error (MSE)

$$J_{mse} = E\left[e^2\,(k)\right] \tag{9}$$

If the input and the desired response are stationary signals, J_{mse} is a quadratic function of the weight W:

$$J_{mse} = W^T R_X W - 2P^T W + \sigma_d^2 \tag{10}$$

where correlation matrix $R_X = E\left(X(k)X^T(k)\right)$, cross-correlation vector $P = E\,(X(k)d(k))$, and $\sigma_d^2 = E\left(d^2(k)\right)$. Then we get the following gradient vector

$$\nabla_{mse} \triangleq \frac{\partial}{\partial W}J_{mse} = 2R_X W - 2P \tag{11}$$

Let $\nabla_{mse} = 0$, the unique optimum weight vector is

$$W_{mse}^{*} = R_X^{-1}P \tag{12}$$

Under MSE criterion, the well-known LMS (least mean square) algorithm is developed (see Kailath [2000]), which is expressed as

$$W(k+1) = W(k) + 2\mu e(k)X(k) \tag{13}$$

In this paper, the filter parameters are adjusted to minimize the following error/input information (EII) instead of the MSE criterion.

$$J_{meii} = I\,(e(k);X(k)) \tag{14}$$

Remark 1: The above cost function is very natural, and shall provide us a more profound interpretation of the filtering problems. The filter makes the mutual information $I\,(e(k);X(k))$ minimized, hence the error signal $e(k)$ contains little "information" about the input data $X(k)$, and we can hardly learn anything from $e(k)$. This criterion is similar to the minimum error/observation information criterion, which is proposed for the state estimation (see Feng [1997] for details).

Theorem 1. Under the minimum error/input information (MEII) criterion, the optimum FIR filter minimizes the error's entropy $H\,(e(k))$.

Proof. By information theory, we have

$$\begin{aligned} I\,(e(k);X(k)) &= H\,(e(k)) - H\,(e(k)\,|X(k)\,) \\ &= H\,(e(k)) - H\left(d(k) - W^T(k)X(k)\,|X(k)\,\right) \\ &= H\,(e(k)) - H\,(d(k)\,|X(k)\,) \end{aligned}$$

It follows that

$$\begin{aligned} &\arg\min_{W\in R^m} J_{meii} = \arg\min_{W\in R^m} I\,(e(k);X(k)) \\ &= \arg\min_{W\in R^m}\{H\,(e(k)) - H\,(d(k)\,|X(k)\,)\} \\ &\overset{(a)}{=} \arg\min_{W\in R^m}\{H\,(e(k))\} \end{aligned}$$

where (a) follows from the fact that W has no effects on the conditional entropy $H\,(d(k)\,|X(k)\,)$.

Theorem 2. Assume $d(k)$, $X(k)$ are zero-mean and jointly Gaussian, $\det R_X \neq 0$, then the optimum weight vector W_{meii}^{*} under MEII criterion is equal to that under MSE criterion, and $I\,(e(k);X(k)\,|W = W_{meii}^{*}\,) = 0$.

Proof.

$$\begin{aligned} \nabla_{meii} &\triangleq \frac{\partial}{\partial W}I\,(e(k);X(k)) \overset{(a)}{=} \frac{\partial}{\partial W}\{H\,(e(k))\} \\ &\overset{(b)}{=} \frac{\partial}{\partial W}\left\{\frac{1}{2}\log\left((2\pi e)\,\mathrm{E}\left[e^2(k)\right]\right)\right\} \\ &= \frac{1}{2}\frac{\partial}{\partial W}\log\left\{W^T R_X W - 2P^T W + \sigma_d^2\right\} \\ &= \frac{R_X W - P}{W^T R_X W - 2P^T W + \sigma_d^2} \end{aligned}$$

where (a) follows from the fact that W has no effects on the conditional entropy $H\,(d(k)\,|X(k)\,)$, and (b) follows from (2). Let $\nabla_{meii} = 0$, we get the optimum weight vector

$$W_{meii}^{*} = R_X^{-1}P$$

From (12), we get $W_{meii}^{*} = W_{mse}^{*}$. Further, in this case, we have

$$E\left(e(k)X^T(k)\right) = E\left\{\left(d(k) - \left(R_X^{-1}P\right)^T X(k)\right)X^T(k)\right\} = 0$$

By *Lemma 4*, the mutual information $I\,(e(k);X(k))$ is

$$
\begin{aligned}
& I\left(e(k); X(k)\left|W=W_{meii}^{*}\right.\right) \\
& =\frac{1}{2}\log\left\{\frac{E\left(e^{2}(k)\right)\det E\left(X(k)X^{T}(k)\right)}{\det\begin{pmatrix} E\left(e^{2}(k)\right) & E\left(e(k)X^{T}(k)\right) \\ E\left(e(k)X(k)\right) & E\left(X(k)X^{T}(k)\right)\end{pmatrix}}\right\} \\
& =\frac{1}{2}\log\left\{\frac{E\left(e^{2}(k)\right)\det E\left(X(k)X^{T}(k)\right)}{\det\begin{pmatrix} E\left(e^{2}(k)\right) & 0 \\ 0 & E\left(X(k)X^{T}(k)\right)\end{pmatrix}}\right\}=0
\end{aligned}
$$

Remark 2: Above derivation suggests that, for zero-mean Gaussian case, $I\left(e(k); X(k)\right)=0$ is equivalent to the well-known orthogonality condition ($E\left[e(k)X(k)\right]=0$) (Kailath [2000]). However, it should be noted that, for most situations, such as the non-Gaussian cases, $\min\limits_{W\in R^{m}} I\left(e(k); X(k)\right)=0$ does not always hold. In fact, the mutual information $I\left(e(k); X(k)\right)$ between error/input takes into accounts both the second and higher order dependencies, while the orthogonality conditions only consider the second order dependencies.

Now, we consider the case in which the desired response $d(k)$ is expressed as

$$
d(k)=W_{0}^{T}X(k)+n(k) \tag{15}
$$

where the disturbance noise $\{n(k)\}$ is independent with $\{x(k)\}$. In this case, the error signal $e\left(k\right)$ is formed as

$$
\begin{aligned}
e\left(k\right) &= d\left(k\right)-y\left(k\right) \\
&= W_{0}^{T}X\left(k\right)+n\left(k\right)-W^{T}\left(k\right)X\left(k\right) \\
&= V^{T}\left(k\right)X\left(k\right)+n\left(k\right)
\end{aligned} \tag{16}
$$

where $V\left(k\right)=W_{0}-W(k)$ is the weight error vector. *Theorem 3*. For the desired response (15), we have $W_{meii}^{*}=W_{0}$.

Proof. By *Lemma 1*, we have $I\left(n(k); X(k)\right)=0$, and hence

$$
\begin{aligned}
& I\left(e(k); X(k)\left|W=W_{0}\right.\right) \\
& = I\left(e(k); X(k)\left|V=0\right.\right)=I\left(n(k); X(k)\right)=0
\end{aligned}
$$

It follows that

$$
I\left(e(k); X(k)\left|W=W_{0}\right.\right)=\min_{W\in R^{m}} I\left(e(k); X(k)\right)
$$

which means $W_{meii}^{*}=W_{0}$.

Denote $\hat{X}$ the conditional mean estimation of X based on the error signal $e\left(k\right)$, that is

$$
\hat{X}(k)=E\left\{X(k)\left|e(k); V(k)\right.\right\} \tag{17}
$$

Let $MMSE\left(\|V(k)\|\right)\triangleq E\left\{\left\|V_{0}^{T}(k)\left(X(k)-\hat{X}(k)\right)\right\|^{2}\right\}$, where $V_{0}\left(k\right)=\frac{V(k)}{\|V(k)\|}$,then the following theorem holds.

Theorem 4. For the desired response (15), assume $\{x\left(k\right)\}$ and $\{n\left(k\right)\}$ are both unit-power white Gaussian noise (WGN), then we have

$$
I\left(e(k); X(k)\left|V\left(k\right)\right.\right)=-\frac{1}{2}\log\left(MMSE\left(\|V(k)\|\right)\right) \tag{18}
$$

Proof. Clearly, we have $\|V_{0}(k)\|=1$, and

$$
e\left(k\right)=\|V\left(k\right)\|V_{0}^{T}\left(k\right)X\left(k\right)+n\left(k\right)
$$

According to the minimum mean-square error (MMSE) estimation theory (Kailath [2000], see pp. 95-96), we have

$$
\begin{aligned}
& E\left[\left(X(k)-\hat{X}(k)\right)\left(X(k)-\hat{X}(k)\right)^{T}\right] \\
& = I-\left\|V\left(k\right)\right\|^{2}V_{0}\left[1+\left\|V\left(k\right)\right\|^{2}V_{0}^{T}V_{0}\right]^{-1}V_{0}^{T} \\
& = I-\frac{\left\|V\left(k\right)\right\|^{2}}{1+\left\|V\left(k\right)\right\|^{2}}V_{0}V_{0}^{T}
\end{aligned}
$$

where I is the identity matrix. And hence

$$
\begin{aligned}
& MMSE\left(\|V(k)\|\right)=E\left\{\left\|V_{0}^{T}(k)\left(X(k)-\hat{X}(k)\right)\right\|^{2}\right\} \\
& = V_{0}^{T}(k)E\left[\left(X(k)-\hat{X}(k)\right)\left(X(k)-\hat{X}(k)\right)^{T}\right]V_{0}(k) \\
& = V_{0}^{T}(k)\left\{I-\frac{\left\|V\left(k\right)\right\|^{2}}{1+\left\|V\left(k\right)\right\|^{2}}V_{0}V_{0}^{T}\right\}V_{0}(k) \\
& = \frac{1}{1+\left\|V\left(k\right)\right\|^{2}}
\end{aligned} \tag{19}
$$

On the other hand, by *Lemma 4*, we can calculate the mutual information $I\left(e(k); X(k)\left|V\left(k\right)\right.\right)$ as follows

$$
\begin{aligned}
& I\left(e(k); X(k)\left|V\left(k\right)\right.\right) \\
& =\frac{1}{2}\log\left\{\left(1+\left\|V\left(k\right)\right\|^{2}\right)\det\begin{bmatrix} I & V\left(k\right) \\ V^{T}(k) & \left\|V\left(k\right)\right\|^{2}+1\end{bmatrix}^{-1}\right\} \\
& =\frac{1}{2}\log\left(1+\left\|V\left(k\right)\right\|^{2}\right)
\end{aligned} \tag{20}
$$

Combining (19) and (20), we arrive the result.

Remark 3: The MMSE and the mutual information (in nats) are plotted in Fig. 1. It is evident that, as the weight vector $W(k)$ approaches to the optimum solution W_{0} ($\|V(k)\|\rightarrow 0$), the mutual information $I\left(e(k); X(k)\left|V\left(k\right)\right.\right)\rightarrow 0$, and $MMSE\left(\|V(k)\|\right)\rightarrow\max\limits_{V} MMSE\left(\|V(k)\|\right)$. This implies that, if the error/input information is minimized, we can hardly estimate the input data $X(k)$ by using the error signal $e(k)$, or in other word, the input data $X(k)$ are utilized completely, such that the error signal $e(k)$ becomes "useless" (all the information contained in $e(k)$ is extracted).

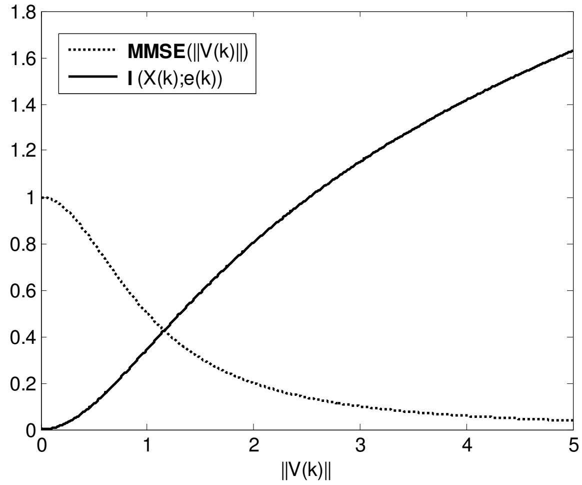

Fig. 1. The MMSE and the mutual information (m=5)

4. STOCHASTIC GRADIENT ALGORITHM

Under the minimum error/input information (MEII) criterion, the weights of the filter are adjusted to minimize the following mutual information.

$$I\left(e(k);X(k)\right)=E\left\{\log\left(\frac{p_{Xe}\left(X(k),e(k)\right)}{p_X\left(X(k)\right)p_e\left(e(k)\right)}\right)\right\} \tag{21}$$

where $p_X\left(.\right)$, $p_e\left(.\right)$ and $p_{Xe}\left(.\right)$ denote the PDF of $X(k)$, $e(k)$, and $(X(k),e(k))$, respectively. In practical situations, these PDFs are usually unknown; hence we have to estimate them from sample data. However, if $m \gg 1$, $X(k)\in R^m$ will be a high-dimensional random variable. In this case, it is impossible to estimate the PDF with finite sample data. In the following, we derive an approximate expression for the mutual information $I\left(e(k);X(k)\right)$, which contains only low dimensional random variables.

By the chain rule(see *Lemma 3*), we have

$$\begin{aligned}&I\left(e(k);X(k)\right)\\&=I\left(e(k);\left[\,x(k)\;\;x(k-1)\;\cdots\;x(k-m+1)\,\right]^T\right)\\&=\sum\nolimits_{i=1}^{m}I\left(e(k);x(k-i+1)\left|x(k)x(k-1)\cdots x(k-i+2)\right.\right)\end{aligned}$$

Assume the input signal $\{x(k)\}$ is an independent stochastic process, i.e. $I\left(x_i;x_j\right)=0$, $\forall i\neq j$, then $(\forall i, 2\leq i\leq m)$

$$\begin{aligned}&I\left(e(k);x(k-i+1)\right)\\&\quad-I\left(e(k);x(k-i+1)\left|x(k)x(k-1)\cdots x(k-i+2)\right.\right)\\&=I\left(x(k)x(k-1)\cdots x(k-i+2);x(k-i+1)\right)\\&\quad-I\left(x(k)x(k-1)\cdots x(k-i+2);x(k-i+1)\left|e(k)\right.\right)\\&=-I\left(x(k)x(k-1)\cdots x(k-i+2);x(k-i+1)\left|e(k)\right.\right)\end{aligned}$$

Under the independent assumption, we have

$$I\left(x(k)x(k-1)\cdots x(k-i+2);x(k-i+1)\left|e(k)\right.\right)\approx 0$$

It follows that

$$\begin{aligned}&I\left(e(k);x(k-i+1)\left|x(k)x(k-1)\cdots x(k-i+2)\right.\right)\\&\approx I\left(e(k);x(k-i+1)\right)\end{aligned}$$

and hence

$$I\left(e(k);X(k)\right)\approx\sum\nolimits_{i=1}^{m}I\left(x(k-i+1);e(k)\right) \tag{22}$$

Now, we design the adaptive algorithm to minimize the following cost function

$$\begin{aligned}J&=\sum_{i=1}^{m}I\left(x(k-i+1);e(k)\right)\\&=\sum_{i=0}^{m-1}E\left\{\log\left(\frac{p_{xe}\left(x(k-i),e(k)\right)}{p_x\left(x(k-i)\right)p_e\left(e(k)\right)}\right)\right\}\end{aligned} \tag{23}$$

We adopt the following popular form

$$W(k+1)=W(k)+\eta\hat{\nabla}_W\left\{J\right\} \tag{24}$$

where η is the adaptation gain (or step-size), $\hat{\nabla}_W\left\{J\right\}$ is the instantaneous estimate of the gradient of J evaluated as the current value of the parameter vector $W(k)$. Obviously, the key problem of the recursive equation (24) is how to calculate the instantaneous gradient $\hat{\nabla}_w\left\{J\right\}$. We start with the calculation of the gradient (not the instantaneous gradient) of J.

$$\begin{aligned}\nabla_W\left\{J\right\}&=\frac{\partial}{\partial W}\left\{\sum\nolimits_{i=1}^{m}E\left\{\log\left(\frac{p_{xe}\left(x(k-i),e(k)\right)}{p_x\left(x(k-i)\right)p_e\left(e(k)\right)}\right)\right\}\right\}\\&=\sum\nolimits_{i=1}^{m}E\left\{\frac{\partial}{\partial W}\left(\log\left\{p_{xe}\left(x(k-i),e(k)\right)\right\}-\log\left\{p_e\left(e(k)\right)\right\}\right)\right\}\\&=\sum\nolimits_{i=1}^{m}E\left\{\frac{\frac{\partial}{\partial W}p_{xe}\left(x(k-i),e(k)\right)}{p_{xe}\left(x(k-i),e(k)\right)}-\frac{\frac{\partial}{\partial W}p_e\left(e(k)\right)}{p_e\left(e(k)\right)}\right\}\end{aligned}$$

By dropping the expectation operator and estimating the PDFs, we obtain the instantaneous gradient $\hat{\nabla}_w\left\{J\right\}$.

$$\hat{\nabla}_w\left\{J\right\}=\sum\nolimits_{i=1}^{m}\left\{\frac{\frac{\partial\hat{p}_{xe}(x(k-i),e(k))}{\partial W}}{\hat{p}_{xe}\left(x(k-i),e(k)\right)}-\frac{\frac{\partial\hat{p}_e(e(k))}{\partial W}}{\hat{p}_e\left(e(k)\right)}\right\} \tag{25}$$

We choose the kernel based approach (Silverman [1986]) for the estimation of PDFs. By kernel method, the estimated PDFs are differentiable. This is very important for the calculation of the gradient. The d-dimensional Gaussian kernel is defined as

$$K\left(x\right)=\left(2\pi\right)^{-d}\exp\left(-\frac{1}{2}x^Tx\right), x\in R^d \tag{26}$$

With Gaussian kernel, the nonparametric estimation of $p_e\left(e(k)\right)$ and $p_{xe}\left(x(k-i),e(k)\right)$ can be expressed as

$$\left\{\begin{aligned}&\hat{p}_e\left(e(k)\right)=\frac{1}{\sqrt{2\pi}Lh_1}\sum_{j=1}^{L}\exp\left(-\frac{\left(\varepsilon_j(k)\right)^2}{2h_1^2}\right)\\&\hat{p}_{xe}\left(x(k-i),e(k)\right)\\&=\frac{1}{2\pi Lh_2^2}\sum_{j=1}^{L}\exp\left(-\frac{\left(\xi_{ij}(k)\right)^2+\left(\varepsilon_j(k)\right)^2}{2h_2^2}\right)\end{aligned}\right. \tag{27}$$

where $\varepsilon_j(k)=e(k)-e(k-j)$, $\xi_{ij}(k)=x(k-i)-x(k-i-j)$, L is the length of sample data, h_1 and h_2 are the kernel widths. Then it follows that

$$\left\{\begin{aligned}&\frac{\partial}{\partial W}\hat{p}_e\left(e(k)\right)=\frac{1}{\sqrt{2\pi}Lh_1^3}\sum_{j=1}^{L}\left\{\exp\left(-\frac{\left(\varepsilon_j(k)\right)^2}{2h_1^2}\right)\pi_{kj}\right\}\\&\frac{\partial}{\partial W}\hat{p}_{xe}\left(x(k-i),e(k)\right)\\&=\frac{1}{2\pi Lh_2^4}\sum_{j=1}^{L}\left\{\exp\left(-\frac{\left(\xi_{ij}(k)\right)^2+\left(\varepsilon_j(k)\right)^2}{2h_2^2}\right)\pi_{kj}\right\}\end{aligned}\right. \tag{28}$$

where $\pi_{kj}=\left(e(k)-e(k-j)\right)\left(X(k)-X(k-j)\right)$. Combine (24), (25), (27) and (28), we obtain the adaptation algorithm under MEII criterion.

5. SIMULATION RESULTS

We now perform Monte-Carlo simulations for system identification to demonstrate the performance of the MEII

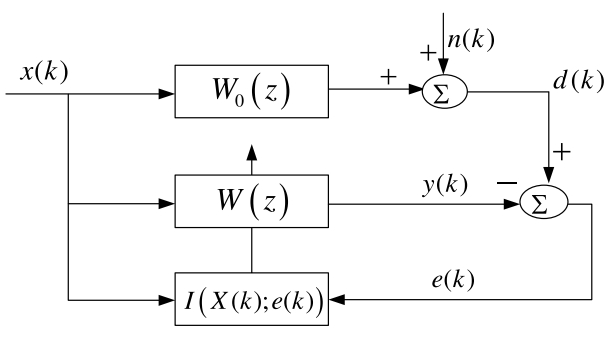

Fig. 2. Scheme of FIR identification under the minimum error/input information (MEII) criterion

algorithm, in comparison with the well-known LMS algorithm. Consider the system identification scheme of Fig. 2, in which the transfer functions of the plant and the adaptive filter are both represented in the FIR form by $W(z) = \sum_{i=1}^{m} w(i)z^{-i+1}$. Let $m = 5$, and the parameter vector of plant be $W_0 = [0.1, 0.3, 0.5, 0.3, 0.1]^T$. The input signal $x(k)$ is chosen as unity-power white Gaussian noise (WGN), and the initial parameters of the adaptive filter are set to be zero. Further, the Gaussian kernels are used, and the kernel sizes h_1 and h_2 are kept fixed at 0.5 and 1.0, respectively. For the disturbance noise $n(k)$, we consider two cases, one for which $n(k)$ is of uniform distribution ($n(k) \sim U[-2, 2]$) and the other for which $n(k)$ is of exponent distribution ($n(k) \sim \exp(1)$). For each case, 100 Monte-Carlo simulations were run and the results averaged.

The average convergence curves of MEII algorithm and LMS algorithm for each case are shown in Fig. 3 and Fig. 4, respectively. In the experiments, the step-sizes are chosen so that the initial convergence rates of the two algorithms were visually identical. From Fig. 3, the MEII algorithm achieves a faster convergence speed during the transient stage. And from Fig. 4, it is evident that, the MEII algorithm achieves a smaller residual error. Both simulation results confirm that, the MEII algorithm may outperform the conventional LMS algorithm.

6. CONCLUSION

The minimum error/input information (MEII) criterion is proposed for the optimal estimation and filtering. For the finite-impulse response (FIR) filter, relationships between the MEII criterion and conventional MSE criterion are studied, and a new adaptation algorithm is derived. Simulation examples illustrate the effectiveness and the better performances of this new algorithm. There are some related problems that need to be studied in the future. Examples include the convergence analysis and the extension to the infinite-impulse response (IIR) or nonlinear filtering.

REFERENCES

T. Kailath, B. Hassibi. *Linear Estimation.* NJ: Prentice Hall, 2000.

S. Haykin. *Adaptive Filtering Theory.* NY: Prentice Hall, 3rd edition, 1996.

D. Erdogmus, J.C. Principe. Convergence properties and data efficiency of the minimum error entropy criterion in Adaline training. *IEEE Trans. Signal Processing*, 51: 1966–1978, 2003.

E. Wolsztynski, E. Thierry, L. Pronzato. Minimum entropy estimation in semi-parametric models. *Signal Processing*, 85:937–949, 2005.

L. Guo, H. Wang. Minimum entropy filtering for multivariate stochastic systems with non-Gaussian noises. *IEEE Transactions on Automatic Control*, 51:695–700, 2006.

D. Erdogmus, J.C. Principe. Comparison of entropy and mean square error criteria in adaptive system training using higher order statistics. *Second International Workshop on Independent Component Analysis and Blind Signal Separation*, 75–80, 2000.

T. M. Cover and J. A. Thomas. *Element of Information Theory.* Chichester: Wiley & Son, Inc., 1991.

S. Ihara. *Information Theory for Continuous Systems.* World Scientific Publishing Co. Pte. Ltd., 1993.

X. Feng, K. A. Loparo, Y. Fang. Optimal state estimation for stochastic systems: an information theoretic approach. *IEEE Transactions on Automatic Control*, 42:771–785, 1997.

B. W. Silverman. *Density Estimation for Statistic and Data Analysis.* NY: Chapman & Hall, 1986.

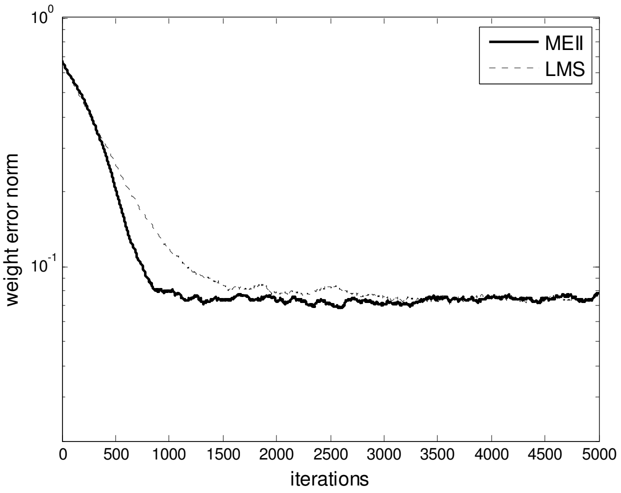

Fig. 3. Average convergence curves of MEII algorithm and LMS algorithm ($n(k) \sim U[-2, 2]$)

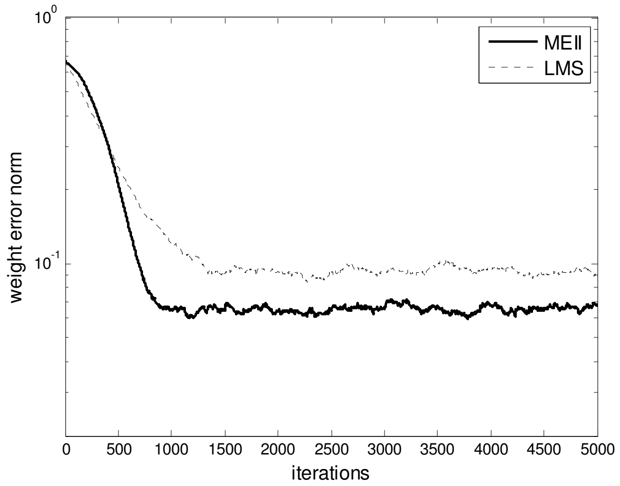

Fig. 4. Average convergence curves of MEII algorithm and LMS algorithm ($n(k) \sim \exp(1)$)