International Journal of Control, Automation, and Systems (2011) 9(6):1049-1055 http://www.springer.com/12555
DOI 10.1007/s12555-011-0604-6

A Variable Step-Size SIG Algorithm for Realizing the Optimal Adaptive FIR Filter

Badong Chen, Yu Zhu, Jinchun Hu, and Jose C. Principe

Abstract: In this paper, we propose an optimal adaptive FIR filter, in which the step-size and error nonlinearity are simultaneously optimized to maximize the decrease of the mean square deviation (MSD) of the weight error vector at each iteration. The optimal step-size and error nonlinearity are derived, and a variable step-size stochastic information gradient (VS-SIG) algorithm is developed to approximately implement the optimal adaptation. Simulation results indicate that this new algorithm achieves faster convergence rate and lower misadjustment error in comparison with other adaptive algorithms.

Keywords: Adaptive FIR filter, optimal error nonlinearity, stochastic information gradient (SIG), variable step-size.

1. INTRODUCTION

Adaptive finite-impulse-response (FIR) filter is one of the core technologies in digital signal processing and finds a number of applications in areas such as channel equalization, system identification, time-series prediction, noise cancellation, and beamforming [1]. The adaptive FIR filter algorithms have attracted research attention for over 50 years, since the late 1950s when the well-known least-mean-square (LMS) algorithm was first developed by Widrow and Hoff [2]. A large family of the tap-weight update-equations for adaptive FIR filter can be expressed as

$$\boldsymbol{W}(k+1)=\boldsymbol{W}(k)+\mu_k f\left(e(k)\right)\boldsymbol{X}^T(k), \tag{1}$$

where $\boldsymbol{W}(k)$ denotes the $M\times 1$ weight vector at iteration k, $e(k)$ is the error signal, $\boldsymbol{X}(k)$ represents the $1\times M$ input (row) regressor, μ_k is the step-size, and $f(.)$ is a scalar (linear or nonlinear) function of the error.

The step-size μ_k and the error function $f(.)$ are two key factors in the adaptation algorithm (1), because they govern the convergence speed as well as the steady-state misadjustment of the algorithm. Up to now, a lot of step-sizes (usually variable step-sizes [3-11]) and error functions (usually error nonlinearities [12-16]) have been proposed to improve the convergence performance. The previous studies, however, focus only on one of the two factors, and to the best of our knowledge, no reports in the literature have attempted to optimize both the step-size and error nonlinearity at the same time. In this work, we propose an optimal adaptive FIR filter, in which the step-size and the error nonlinearity are simultaneously optimized to maximize the decrease of the mean square deviation (MSD) of the weight error vector at each iteration. In particular, we develop a variable step-size stochastic information gradient (SIG) [17] algorithm to approximately realize this optimal adaptive filter. As will be shown in the simulation part, the new algorithm achieves a noticeable performance improvement over some existing algorithms.

2. THE OPTIMAL ADAPTIVE FIR FILTER

Consider the case in which the adaptive FIR filter attempts to identify the $M\times 1$ weight vector $\boldsymbol{W}^*$ of an unknown FIR system, whose output samples $\{d(k)\}$ are related via

$$d(k)=\boldsymbol{X}(k)\boldsymbol{W}^*+v(k), \tag{2}$$

where $v(k)$ is the disturbance noise. In this case, the error signal $e(k)$ is given by

$$e(k)=\boldsymbol{X}(k)\tilde{\boldsymbol{W}}(k)+v(k), \tag{3}$$

where $\tilde{\boldsymbol{W}}(k)=\boldsymbol{W}^*-\boldsymbol{W}(k)$ is the weight error vector. By the energy conservation relation [16], we have

$$\begin{aligned}E\left[\left\|\tilde{\boldsymbol{W}}(k+1)\right\|^2\right]&=E\left[\left\|\tilde{\boldsymbol{W}}(k)\right\|^2\right]-2\mu_k E\left[e_a(k)f\left(e(k)\right)\right]\\&\quad+\mu_k^2 E\left[\left\|\boldsymbol{X}(k)\right\|^2 f^2\left(e(k)\right)\right]\\&=E\left[\left\|\tilde{\boldsymbol{W}}(k)\right\|^2\right]-\Delta_{MSD}\left(\mu_k,f\right),\end{aligned} \tag{4}$$

Manuscript received February 8, 2010; revised March 3, 2011; accepted June 21, 2011. Recommended by Editorial Board member Young Soo Suh under the direction of Editor Young Il Lee.

This work was supported by National Natural Science Foundation of China (No. 60904054), and was partially supported by NSF grant ECCS 0856441, NSF IIS 0964197 and ONR N00014-10-1-0375.

Badong Chen and Jose C. Principe are with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 USA (e-mails: chenbd04@mails.tsinghua.edu.cn, principe@cnel.ufl.edu).

Yu Zhu and Jinchun Hu are with the Institute of Manufacturing Engineering, Department of Precision Instruments and Mechanology, Tsinghua University, Beijing 100084, P. R. China (e-mails: {zhuyu, hujinchun}@tsinghua.edu.cn).

© ICROS, KIEE and Springer 2011

where

$$\Delta_{MSD}\left(\mu_k, f\right) \triangleq 2\mu_k E\left[e_a(k) f\left(e(k)\right)\right] - \mu_k^2 E\left[\|\boldsymbol{X}(k)\|^2 f^2\left(e(k)\right)\right]$$

is the decrease of the mean square deviation (MSD) at k iteration ($MSD \triangleq E[\|\tilde{\boldsymbol{W}}(k)\|^2]$), $\|.\|$ denotes the Euclidean norm, $e_a(k) \triangleq \boldsymbol{X}(k)\tilde{\boldsymbol{W}}(k)$ is the so called *a priori* error [16]. The MSD is usually used as the performance measure for the adaptation algorithm (1). To obtain the fast convergence speed and the smallest misadjustment, one should maximize the MSD decrease at each iteration. Therefore, the optimum step-size and error function would be

$$\begin{aligned}
(\mu_k^*, f^*) &= \arg\max_{\mu_k, f} \Delta_{MSD}(\mu_k, f) \\
&= \arg\max_{\mu_k, f} \left\{ \begin{array}{l} 2\mu_k E\left[e_a(k) f\left(e(k)\right)\right] \\ -\mu_k^2 E\left[\|\boldsymbol{X}(k)\|^2 f^2\left(e(k)\right)\right] \end{array} \right\} \\
&= \arg\max_{\mu_k, f} \left\{ \begin{array}{l} -E\left[\|\boldsymbol{X}(k)\|^2 f^2(e(k))\right] \\ \times \left(\mu_k - \dfrac{E\left[e_a(k) f(e(k))\right]}{E\left[\|\boldsymbol{X}(k)\|^2 f^2(e(k))\right]} \right)^2 \\ + \dfrac{\left(E\left[e_a(k) f(e(k))\right]\right)^2}{E\left[\|\boldsymbol{X}(k)\|^2 f^2(e(k))\right]} \end{array} \right\}.
\end{aligned} \tag{5}$$

From (5), we get

$$\mu_k^* = \frac{E\left[e_a(k) f\left(e(k)\right)\right]}{E\left[\|\boldsymbol{X}(k)\|^2 f^2\left(e(k)\right)\right]}. \tag{6}$$

Fig. 1 depicts the curve of the MSD decrease Δ_{MSD} as a function of the step-size, from which we see that the optimal step-size equals $\mu_{\max}/2$. Here $\mu_{\max}$ is the maximum step-size which ensures $\Delta_{MSD} \geq 0$. Of course, the optimal step-size μ_k^* will guarantee the convergence of the recursion, since $\mu_k^* < \mu_{\max}$, and we always have $E\left[\|\tilde{\boldsymbol{W}}(k+1)\|^2\right] \leq E\left[\|\tilde{\boldsymbol{W}}(k)\|^2\right]$.

To derive the optimal error function f^*, we give the following assumptions [16]:

Assumption 1: The noise sequence $\{v(k)\}$ is independent, identically distributed, and independent of the input sequence $\{\boldsymbol{X}(k)\}$;

Assumption 2: The filter is long enough such that $e_a(k)$ is Gaussian distributed;

Assumption 3: [1]The filter is long enough such that

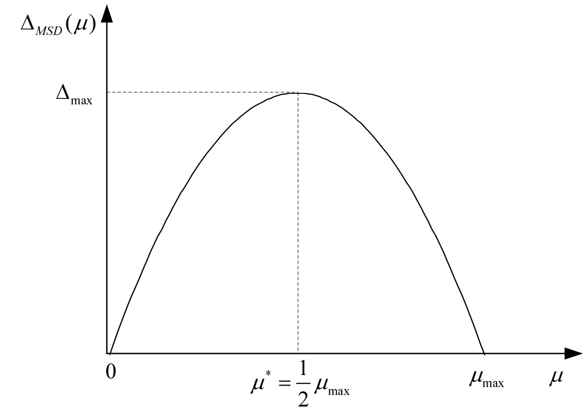

Fig. 1. MSD decrease Δ_{MSD} versus step-size μ.

the random variables $\|\boldsymbol{X}(k)\|^2$ and $f^2\left(e(k)\right)$ are uncorrelated, i.e.,

$$E\left[\|\boldsymbol{X}(k)\|^2 f^2\left(e(k)\right)\right] = E\left[\|\boldsymbol{X}(k)\|^2\right] E\left[f^2\left(e(k)\right)\right] \tag{7}$$

Moreover, we assume the error function $f(.)$ satisfies

$$\lim_{e \to \pm\infty} f(e) p_e(e) = 0 \tag{8}$$

where $p_e(e)$ is the probability density function (PDF) of error $e(x)$. Notice condition (8) is not too restrictive, because for most physical signals, the PDF $p(x)$ decreases rapidly as x goes to infinity.

With the above assumptions, we derive

$$\begin{aligned}
E\left[e_a(k) f\left(e(k)\right)\right] &= E\left[e_a(k) f\left(e_a(k) + v(k)\right)\right] \\
&\overset{(a)}{=} E\left[e_a^2(k)\right] E\left[f'\left(e(k)\right)\right] \\
&\overset{(b)}{=} -E\left[e_a^2(k)\right] \int_{-\infty}^{+\infty} p_e'(e) f(e) de,
\end{aligned} \tag{9}$$

where (a) follows from the Gaussian assumption and Price theorem [18,19], and (b) follows from the condition (8). Thus the MSD decrease Δ_{MSD} can be expressed as

$$\begin{aligned}
\Delta_{MSD}(\mu_k, f) = &-2\mu_k E\left[e_a^2(k)\right] \int_{-\infty}^{+\infty} p_e'(e) f(e) de \\
&- \mu_k^2 E\left[\|\boldsymbol{X}(k)\|^2\right] \int_{-\infty}^{+\infty} p_e(e) f^2(e) de.
\end{aligned} \tag{10}$$

And then, the Gateaux derivative of Δ_{MSD} with respect to f in the direction of β is given by

$$\begin{aligned}
&\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left\{ \Delta_{MSD}\left(\mu_k, f + \varepsilon\beta\right) - \Delta_{MSD}\left(\mu_k, f\right) \right\} \\
&= \int_{-\infty}^{+\infty} \left(\begin{array}{l} -2\mu_k E\left[e_a^2(k)\right] p_e'(e) \\ -2\mu_k^2 E\left[\|\boldsymbol{X}(k)\|^2\right] f(e) p_e(e) \end{array} \right) \beta(e) de.
\end{aligned} \tag{11}$$

Let $\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left\{ \Delta_{MSD}\left(\mu_k, f + \varepsilon\beta\right) - \Delta_{MSD}\left(\mu_k, f\right) \right\} \equiv 0$ for all β, we have $f^*(e) = -\lambda\, p_e'(e) / p_e(e)$, where $\lambda =$

[1] Similar uncorrelation assumption appears in [16]. This assumption can be justified by the law of large numbers. It becomes more realistic as the filter gets longer.

$E\left[e_a^2(k)\right]\Big/\left(\mu_k E\left[\|\boldsymbol{X}(k)\|^2\right]\right)$. As λ can be absorbed into the step-size μ_k, we choose

$$f^*(e) = -\frac{p_e'(e)}{p_e(e)}. \tag{12}$$

Remark 1: It is interesting to observe that the optimal error function $f^*(.)$ is just the minus *score* [20] of the error variable. Moreover, we can rewrite (12) as

$$f^*(e) = \frac{\partial \phi^*(e)}{\partial e} = \frac{\partial}{\partial e}\left(-\log p_e(e)\right), \tag{13}$$

where $\phi^*(e) = -\log p_e(e)$ is the underlying cost function of the adaptation. Clearly, minimizing $\phi^*(e)$ is equivalent to maximizing the logarithmic likelihood function $\log p_e(e)$. Heuristically, we could say the optimal function (12) gives the maximum likelihood (ML) method, which has the best possible asymptotic properties (e.g. the asymptotic efficiency) one can hope for. Here we should note that the optimal function $f^*(e)$ is time-varying, because the error's PDF $p_e(e)$ always changes across iterations.

Remark 2: The optimal cost can also be regarded as the minimum error entropy (MEE) criterion [21-26]. In fact, the expectation of the cost function $\phi^*(e)$ is

$$\begin{aligned} E\left[\phi^*(e)\right] &= E\left[-\log p_e(e)\right] \\ &= -\int_{-\infty}^{+\infty} p_e(e)\log p_e(e)de. \end{aligned} \tag{14}$$

Thus minimizing $E[\phi^*(e)]$ is equivalent to minimizing the error's entropy $H(e) = -\int_{-\infty}^{+\infty} p_e(e)\log p_e(e)de.$ This gives an interesting interpretation for why the MEE criterion can be successfully used in the areas such as machine learning and adaptive system training [22-26].

Remark 3: It should be noted that the authors of [16] have proposed to optimize the error function by minimizing the steady-state excess mean-square error (EMSE), and obtained the same optimal function. Their approach is based on the Cramer-Rao lower bound (CRLB). Further, in an earlier work [14], the error nonlinearity is optimized to minimize the EMSE at each iteration by using the constrained optimization and calculus of variations method, with which the optimal error function is derived as

$$f^*(e) = -\frac{p_e'(e)}{p_e(e) + \mu\lambda p_e''(e)}, \tag{15}$$

where λ is the input signal power. In the case of slow adaptation, the step-size μ will be chosen small such that the optimal nonlinearity (15) is approximately given by (12).

Combining (9) and (12), we have

$$\begin{aligned} &E\left[e_a(k)f^*\left(e(k)\right)\right] \\ &= -E\left[e_a^2(k)\right]\int_{-\infty}^{+\infty} p_e'(e)f^*(e)de \\ &= E\left[e_a^2(k)\right]\int_{-\infty}^{+\infty}\left(\frac{p_e'(e)}{p_e(e)}\right)^2 p_e(e)de \\ &= E\left[e_a^2(k)\right]J_F(e), \end{aligned} \tag{16}$$

where $J_F(e) \triangleq \int_{-\infty}^{+\infty}\left(p_e'(e)/p_e(e)\right)^2 p_e(e)de$ is the Fisher information with respect to location parameter [20]. In addition, combining (7) and (12) yields

$$E\left[\|\boldsymbol{X}(k)\|^2 f^{*2}\left(e(k)\right)\right] = E\left[\|\boldsymbol{X}(k)\|^2\right]J_F(e). \tag{17}$$

Substituting (16) and (17) into (6), we obtain the optimal step-size with respect to the optimal error nonlinearity f^*, that is

$$\mu_k^* = \frac{E\left[e_a^2(k)\right]}{E\left[\|\boldsymbol{X}(k)\|^2\right]}. \tag{18}$$

Now we have derived the optimal adaptive FIR filter, whose error nonlinearity and step-size are given by (12) and (18), respectively.

3. A VARIABLE STEP-SIZE SIG ALGORITHM

There are two obstacles to the practical implementation of the optimal adaptive FIR filter: (1) the error's PDF $p_e(e)$ is not available during adaptation, and (2) the *a priori* error $e_a(k)$ depends on $\boldsymbol{W}^*$, which is unknown. One approach to deal with the first obstacle is the online density estimation, which estimates the error distribution from the latest error samples available. This method has been widely used in the areas of information theoretic learning (ITL) [22-26]. By this approach, the optimal adaptation algorithm becomes

$$\boldsymbol{W}(k+1) = \boldsymbol{W}(k) - \mu_k^*\frac{\partial}{\partial \boldsymbol{W}}\left(-\log \hat{p}_e\left(e(k)\right)\right), \tag{19}$$

where $\hat{p}_e(.)$ denotes the estimated PDF of the error. In [17], the gradient $\frac{\partial}{\partial \boldsymbol{W}}\left(-\log \hat{p}_e\left(e(k)\right)\right)$ is called the stochastic information gradient (SIG), since it can be viewed as the stochastic gradient of the error's entropy. By kernel density estimation (KDE) [27], the PDF estimate of the error evaluated at $e(k)$ is

$$\hat{p}_e\left(e(k)\right) = \frac{1}{L}\sum_{i=k-L+1}^{k} K_\sigma\left(e(k) - e(i)\right), \tag{20}$$

where L is the sliding error samples length, $K_\sigma(.)$ is the kernel function with width σ [27]. Then the stochastic information gradient can be calculated as [17]

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{W}}\left(-\log \hat{p}_e\left(e(k)\right)\right) \\
&=\frac{\partial}{\partial \boldsymbol{W}}\left\{-\log\left(\frac{1}{L}\sum_{i=k-L+1}^{k} K_\sigma\left(e(k)-e(i)\right)\right)\right\} \\
&=\frac{\sum_{i=k-L+1}^{k} K'_\sigma\left(e(k)-e(i)\right)\left(\boldsymbol{X}(k)-\boldsymbol{X}(i)\right)}{\sum_{i=k-L+1}^{k} K_\sigma\left(e(k)-e(i)\right)}.
\end{aligned}
\tag{21}
$$

We now calculate the optimal step-size μ_k^* of (18). The problem here is of course that the expectations are not computable since the underlying distributions are unknown. A simple estimate of the expectation is to replace it by the sample mean, thus we have

$$
\mu_k^* \approx \left(\sum_{i=k-L+1}^{k} e_a^2(i)\right) \Big/ \left(\sum_{i=k-L+1}^{k} \|\boldsymbol{X}(i)\|^2\right). \tag{22}
$$

In practical situations, the *a priori* error samples $\{e_a(i)\}$ are usually unknown. However, in the initial stage of the adaptation, the algorithm is far from the optimum solution such that $e_a(i) \approx e(i)$. And hence, the optimal step-size in the initial stage can be estimated by

$$
\mu_k^* \approx \left(\sum_{i=k-L+1}^{k} e^2(i)\right) \Big/ \left(\sum_{i=k-L+1}^{k} \|\boldsymbol{X}(i)\|^2\right). \tag{23}
$$

Suppose now the algorithm is near the optimum solution when $k \geq k_0$. In this case, we consider the following staircase optimal step-sizes:

$$
\mu_k = \mu_{k_0+i\tau}^*,\ \text{if}\ k_0 + i\tau \leq k < k_0 + (i+1)\tau \quad i = 0, 1, 2, \cdots, \tag{24}
$$

where $\tau \in \mathbb{N}$ is large enough such that at iteration k_0+ $(i$+1$)\tau$, $E[e_a^2(k)] \approx S(\mu_{k_0+i\tau}^*, f^*)$. Here $S(\mu_{k_0+i\tau}^*, f^*)$ denotes the steady-state EMSE $\left(\lim_{k\to\infty} E\left[e_a^2(k)\right]\right)$ with step-size $\mu = \mu_{k_0+i\tau}^*$ and optimal error nonlinearity f^*. According to [16], the steady-state EMSE $S(\mu_{k_0+i\tau}^*, f^*)$ can be expressed as

$$
\begin{aligned}
S\left(\mu_{k_0+i\tau}^*, f^*\right) &= \frac{\mu_{k_0+i\tau}^*}{2} E\left[\|\boldsymbol{X}(k)\|^2\right] \frac{E\left[f^{*2}(e)\right]}{E\left[f^{*\prime}(e)\right]} \\
&\overset{(a)}{=} \frac{\mu_{k_0+i\tau}^*}{2} E\left[\|\boldsymbol{X}(k)\|^2\right],
\end{aligned}
\tag{25}
$$

where (a) follows from the fact that $E[f^{*\prime}(e)] = E[f^{*2}(e)] = J_F(e)$. Thus we have

$$
E\left[e_a^2(k_0 + (i+1)\tau)\right] \approx \frac{\mu_{k_0+i\tau}^*}{2} E\left[\|\boldsymbol{X}(k)\|^2\right]. \tag{26}
$$

Combining (18) and (26) yields

$$
\mu_{k_0+(i+1)\tau}^* \approx \mu_{k_0+i\tau}^* / 2. \tag{27}
$$

Therefore, the staircase optimal step-sizes can be approximately given by

$$
\mu_k = \mu_{k_0}^* / 2^i,\ \text{if}\ k_0 + i\tau \leq k < k_0 + (i+1)\tau. \tag{28}
$$

One drawback of the staircase optimal step-sizes is that, the step-sizes are frozen between any two successive turning points. To deal with this problem, we give the following smoothed optimal step-sizes:

$$
\mu_{k+1}^* = \mu_k^* / 2^{1/\tau},\ \text{for}\ k \geq k_0. \tag{29}
$$

Now we have derived the computable optimal step-sizes for both the initial and final stages of the adaptation. In order to ultimately implement the algorithm, we need identify the dividing point k_0 between the two stages. To this end, we introduce the autocorrelation $\rho(k) \triangleq E[e(k)e(k-1)]$ between $e(k)$ and $e(k$–1$)$ to measure how far the algorithm is from the optimum solution. As argued in [4], the error autocorrelation $\rho(k)$ will be large in the early stage of adaptation and will approach zero as the algorithm approaches the optimum even in the presence of noises $\{v(k)\}$. The time-average estimate of $\rho(k)$ can be expressed as

$$
\hat{\rho}(k) = \alpha\hat{\rho}(k-1) + (1-\alpha)e(k)e(k-1), \tag{30}
$$

where $0 < \alpha < 1$ is the exponential weighting parameter. Based on $\hat{\rho}(k)$, the proposed step-size is given by

$$
\mu_k^* = \begin{cases} \dfrac{\left(\sum_{i=k-L+1}^{k} e^2(i)\right)}{\left(\sum_{i=k-L+1}^{k} \|\boldsymbol{X}(i)\|^2\right)} & \text{if}\ \ |\hat{\rho}(k)| \geq \varepsilon \\ \mu_{k-1}^* / 2^{1/\tau} & \text{if}\ \ |\hat{\rho}(k)| < \varepsilon, \end{cases} \tag{31}
$$

where $\varepsilon > 0$ is a small positive number used for the threshold. Sometimes we need choose a minimum step-size $\mu_{\min}$ to provide a minimum level of tracking ability, i.e., $\mu_k = \max\{\mu_k^*, \mu_{\min}\}$.

Combining (19), (21) and (31), we obtain the variable stochastic information gradient (VS-SIG) algorithm.

4. SIMULATION RESULTS

In this section, we perform simulation experiments to illustrate the favorable behavior of the VS-SIG algorithm in comparison to other adaptive algorithms. The system to be identified is an FIR channel with 14-dimentional normalized weight vector ($\|\boldsymbol{W}^*\|$=1). The input signal is a zero-mean white Gaussian process with unit power. The noise sequence $\{v(k)\}$ is zero-mean Laplace distributed with variance 0.01 such that *SNR*=20*dB*. To calculate the stochastic information gradient, we set the sliding error samples length L=20, and choose Gaussian function as the kernel, whose kernel width is determined by the

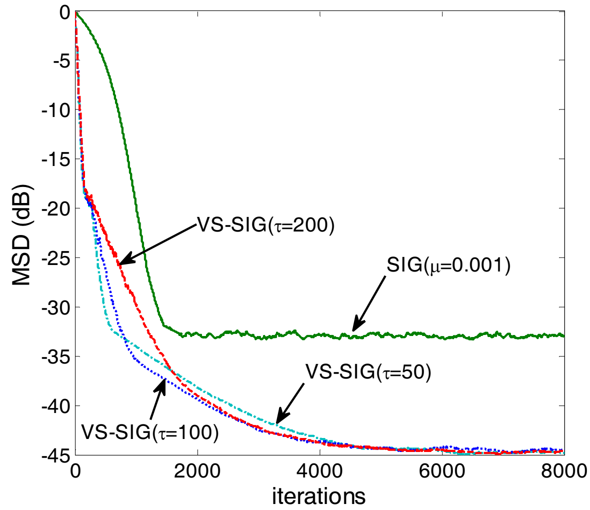

Fig. 2. Convergence curves of the VS-SIG and SIG algorithms.

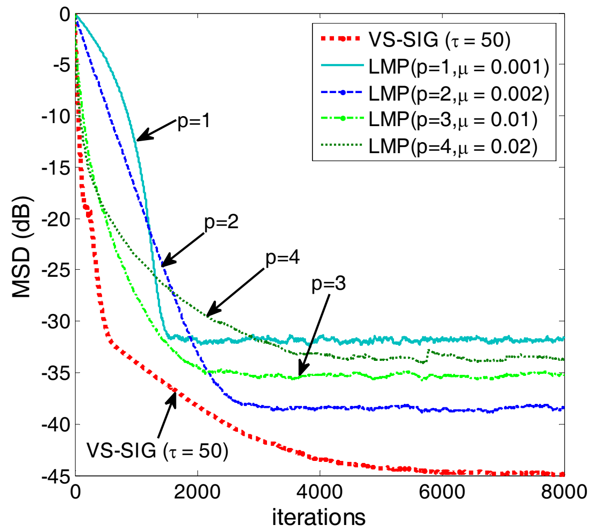

Fig. 3. Convergence curves of the VS-SIG and LMP-family algorithms.

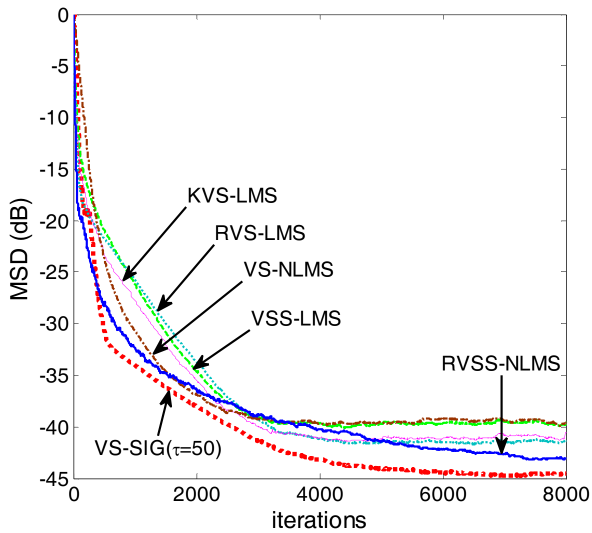

Fig. 4. Convergence curves of the VS-SIG and several variable step-size LMS algorithms.

Table 1. Parameters setting for variable step-size LMS algorithms.

VSS-LMS	RVS-LMS	KVS-LMS	VS-NLMS [7]	RVSS-NLMS [8]
$\mu_{\max}$ =0.1	$\mu_{\max}$ =0.1	$\mu_{\max}$ =0.1	$\mu_{\max}$ =0.1	δ_0 =0.2
$\mu_{\min}$ =0.001	$\mu_{\min}$ =0.001	α =10	α =0.99	α =0.97
α =0.9	α =0.97	β =0.98	C =0.0001	ε =0.0001
γ =0.005	β =0.99	-	-	-
-	γ =0.01	-	-	-

Silverman's rule [27]. Further, we set α=0.99, ε=0.1 and $\mu_{\min}$= 0.000005 for the VS-SIG algorithm. All the simulation results below are obtained by ensemble averaging over 100 independent trials.

Fig. 2 shows the convergence curves of the MSD for the VS-SIG algorithm with different τ values and the SIG algorithm with step-size $\mu = 0.001$. It is clear that the VS-SIG algorithm converges faster and has lower misadjustment error. The excellent performance of the VS-SIG algorithm can also be observed from Fig. 3 and Fig. 4, in which the learning curves of VS-SIG ($\tau = 50$) are compared, respectively, with those of the least mean *p*-power (LMP) [13] algorithms and several variable step-size LMS algorithms. Except the RVSS-NLMS algorithm [8], the used variable step-size LMS algorithms are summarized in [7]. Table 1 lists the parameters setting for different variable step-size algorithms. These parameters are experimentally chosen such that the algorithms achieve a good tradeoff between the convergence speed and the final misadjustment.

5. CONCLUSION

A variable step-size stochastic information gradient (VS-SIG) algorithm has been developed to approximately realize the optimal adaptive FIR filter designed by optimizing both the step-size and error nonlinearity such that the MSD decrease at each iteration is maximized. Simulation experiments have shown that the proposed algorithm is highly effective in improving the convergence speed and misadjustment error.

REFERENCES

[1] S. Haykin, *Adaptive Filtering Theory*, 3rd ed., Prentice Hall, NY, 1996.

[2] B. Widrow and M. E. Hoff, "Adaptive switching circuits," *IRE WESCON Convention Record*, pp. 96-104, 1960.

[3] R. H. Kwong and E. W. Johnston, "A variable step-size LMS algorithm," *IEEE Trans. on Signal Processing*, vol. 40, pp. 1633-1642, 1992.

[4] T. Aboulnasr and K. Mayyas, "A robust variable step-size LMS-type algorithm: analysis and simulations," *IEEE Trans. on Signal Processing*, vol. 45, pp. 631-639, 1997.

[5] D. I. Pazaitis and A. G. Constantinides, "A novel kurtosis driven variable step-size adaptive algorithm," *IEEE Trans. on Signal Processing*, vol.

47, pp. 864-872, 1999.
[6] A. Mader, H. Puder, and G. U. Schmidt, "Step-size control for acoustic echo cancellation filters - an overview," *Signal Processing*, vol. 80, pp. 1697-1719, 2000.
[7] H. C. Shin, A. H. Sayed, and W. J. Song, "Variable step-size NLMS and affine projection algorithms," *IEEE Signal Processing Letters*, vol. 11, pp. 132-135, 2004.
[8] L. R. Vega, H. Rey, J. Benesty, and S. Tressens, "A new robust variable step-size NLMS algorithm," *IEEE Trans. Signal Processing*, vol. 56, pp. 1878-1893, 2008.
[9] T. Shao and Y. R. Zheng, "A new variable step-size fractional lower order moment algorithm for non-Gaussian interference environments," *Proc. IEEE ISCAS*, pp. 2065-2068, May 2009.
[10] Y. R. Zheng and T. Shao, "A variable step-size LMP algorithm for heavy-tailed interference suppression in phased array radar," *Proc. IEEE Aerospace Conf.*, pp. 1-6, Mar. 2009.
[11] L. R. Vega, H. Rey, and J. Benesty, "A robust variable step-size affine projection algorithm," *Signal Processing*, vol. 90, pp. 2806-2810, 2010.
[12] E. Walach and B. Widrow, "The least mean fourth (LMF) adaptive algorithm and its family," *IEEE Trans. Inform. Theory*, vol. IT-30, no. 2, pp. 275-283, 1984.
[13] S. C. Pei and C. C. Tseng, "Least mean p-power error criterion for adaptive FIR filter," *IEEE Journal on Selected Areas in Communications*, vol. 12, no. 9, pp. 1540-1547, 1994.
[14] S. C. Douglas and H. Y. Meng, "Stochastic gradient adaptation under general error criteria," *IEEE Trans. Signal Processing*, vol. 42, pp. 1335-1351, 1994.
[15] P. Petrus, "Robust Huber adaptive filter," *IEEE Trans. on Signal Processing*, vol. 47, pp. 1129-1133, 1999.
[16] T. Y. Al-Naffouri and A. H. Sayed, "Adaptive filters with error nonlinearities: mean-square analysis and optimum design," *EURASIP Journal on Applied Signal Processing*, vol. 4, 2001, 192-205.
[17] D. Erdogmus, E. H. Kenneth, and J. C. Principe, "Online entropy manipulation: stochastic information gradient," *IEEE Signal Processing Letters*, vol. 10, pp. 242-245, 2003.
[18] N. Bershad and M. Bonnet, "Saturation effects in LMS adaptive echo cancellation for binary data," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 38, pp. 1687-1696, 1990.
[19] R. Price, "A useful theorem for nonlinear devices having Gaussian inputs," *IEEE Trans. Inform. Theory*, vol. IT-4, no. 6, pp. 69-72, 1958.
[20] T. M. Cover and J. A. Thomas, *Element of Information Theory*, Wiley & Son, Inc., Chichester, 1991.
[21] M. Janzura, T. Koski, and A. Otahal, "Minimum entropy of error principle in estimation," *Information Sciences*, vol. 79, pp. 123-144, 1994.
[22] J. C. Principe, D. Xu, Q. Zhao, and J. W. Fisher III, "Learning from examples with information theoretic criteria," *Journal of VLSI Signal Processing Systems*, vol. 26, pp. 61-77, 2000.
[23] D. Erdogmus and J. C. Principe, "An error entropy minimization algorithm for supervised training of nonlinear adaptive systems," *IEEE Trans. on Signal Processing*, vol. 50, pp. 1780-1786, 2002.
[24] D. Erdogmus and J. C. Principe, "From linear adaptive filtering to nonlinear information processing," *IEEE Signal Processing Magazine*, vol. 23, no. 6, pp. 15-33, 2006.
[25] B. Chen, J. Hu, L. Pu, and Z. Sun, "Stochastic gradient algorithm under (h, ϕ)-entropy criterion," *Circuits Systems Signal Processing*, vol. 26, pp. 941-960, 2007.
[26] B. Chen, Y. Zhu, and J. Hu, "Mean-square convergence analysis of ADALINE training with minimum error entropy criterion," *IEEE Trans. on Neural Networks*, vol. 21, no. 7, pp. 1168-1179, 2010.
[27] B. W. Silverman, *Density Estimation for Statistic and Data Analysis*, Chapman & Hall, NY, 1986.

Badong Chen received his Ph.D. in Computer Science and Technology from Tsinghua University, Beijing, China, in 2008. He is currently a post-doctoral associate with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, USA. His research interests are in statistical signal processing, information theoretic learning (ITL), kernel adaptive filters, and applications in neuroscience, brain-machine interface, bioinformatics, sensor and social networks.

Yu Zhu received his B.S. of Radio Electronics in 1983 at Beijing Normal University, and an M.S. of Computer Applications in 1993, and a Ph.D. of Mechanical Design and Theory in 2001 at China University of Mining & Technology. He is now a professor with the Institute of Manufacturing Engineering, Department of Precision and Mechanology, Tsinghua University, Beijing, China. His current research interests are parallel machanism and theory, two photon micro-fabrication, ultra-precision motion system and motion control.

Jinchun Hu received his Ph.D. in Control Science and Engineering from Nanjing University of Science and Technology, Nanjing, China, in 1998. Since then, he has been a postdoctoral researcher in Nanjing University of Aeronautics and Astronautics in 1999 and Tsinghua University in 2002, respectively. His research interests are in system modeling, robotics and intelligent control. Dr. Hu is currently an associate professor in the Department of Precision Instruments and Mechanology, Tsinghua University, Beijing, China.

Jose C. Principe is currently the Distinguished Professor of Electrical and Biomedical Engineering at the University of Florida, Gainesville, where he teaches advanced signal processing and artificial neural networks (ANNs) modeling. He is BellSouth Professor and Founder and Director of the University of Florida Computational Neuro-Engineering Laboratory (CNEL). He is involved in biomedical signal processing, in particular, the electroencephalogram (EEG) and the modeling and applications of adaptive systems. He has more than 150 publications in refereed journals, 15 book chapters, and over 400 conference papers. He has directed over 60 Ph.D. dissertations and 61 Master's degree theses.

Dr. Príncipe is the past Editor-in-Chief of the IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, past President of the International Neural Network Society, and former Secretary of the Technical Committee on Neural Networks of the IEEE Signal Processing Society. He is an IEEE Fellow and an AIMBE Fellow and a recipient of the IEEE Engineering in Medicine and Biology Society Career Service Award. He is also a former member of the Scientific Board of the Food and Drug Administration, and a member of the Advisory Board of the McKnight Brain Institute at the University of Florida.