A MODEL-BASED APPROACH FOR THE DEVELOPMENT OF LMS ALGORITHMS

Guang Deng

Department of Electronic Engineering, La Trobe University
Bundoora, Vicotira 3086, Austrailia
d.deng@latrobe.edu.au

Wai-Yin Ng

Department of Information Engineering
The Chinese University of Hong Kong, Shatin, Hong Kong
w.ng@acm.org

ABSTRACT

The LMS algorithm is one of the most popular adaptive filter algorithms. Many variants of the algorithm have been developed for different applications. In this paper, we propose a unified model-based approach for developing LMS algorithms. We use a number of probability density functions to model the filtering error and the filter coefficients. The filter coefficients are determined by maximizing the posterior distribution function. We demonstrate that using this approach, we can not only develop existing LMS algorithms with further insights, we can also explore a number of new algorithms with certain desired properties such as robustness and sparseness.

1. INTRODUCTION

The classical LMS algorithm [1, 2] can be summarized as the following. Given the input data vector $\mathbf{x}$ of M elements, the desired scaler output y, a linear model

$$y = \mathbf{w}^T\mathbf{x} + e \tag{1}$$

and the filter coefficient vector s from the previous iteration, the current coefficient vector is given by

$$\mathbf{w} = \mathbf{s} + \mu\hat{e}\mathbf{x} \tag{2}$$

where μ is an adaptation constant that determines the step-size of the update and $\hat{e} = y - \mathbf{s}^T\mathbf{x}$. We also define the difference vector $\mathbf{r} = \mathbf{w} - \mathbf{s}$. The mth element of the vector $\mathbf{w}$ is denoted by w_m. The same notation is used for other vectors.

In the following, we briefly review optimization-based research and recent developments in LMS algorithm related to this work. A representative example of formulating the development of the LMS algorithm as a constrained optimization problem is that of the normalized LMS algorithm [1], which is generalized in [3]. In a recent paper [4], the sparsity of the coefficient vector is considered and algorithms are developed based on solutions to a number of constrained optimization problems. LMS algorithms with sparse coefficient vectors, which arise from echo-cancellation application [5], have been proposed by researchers using the exponentiated gradient [6], the natural gradient [7] and the approximate natural gradient [8]. On the other hand, an LMS algorithm, which is robust to outliers, is desirable in many applications. Robustness can be achieved, from an algorithm development point of view, by using Huber's M-estimator [9] to measure the filtering error [10, 11], or by using a mixed-norm for the error [12].

This study is motivated by the optimization approach for the LMS algorithm development. We formulate the problem of developing an LMS algorithm as a maximum a posterior estimation (MAP) problem. A distinctive advantage of this formulation is that it is a unified approach by which a wide range of existing LMS algorithms in their generalized form can be derived. These algorithms include the classical LMS, the normalized LMS, the signed LMS [2], the proportionate LMS and the proportionate normalized LMS [4, 5]. This approach also opens new pathways to explore different model settings that lead to LMS algorithms with desired properties such as sparseness and robustness. The sparseness and robustness constraints are imposed naturally by specifying suitable prior distributions for the coefficient vector and the likelihood functions for the data [13–15].

2. PROBLEM FORMULATION

We formulate the problem of determining the filter coefficient vector w as a maximum a posteriori (MAP) estimation problem

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} p(\mathbf{w}|y, \mathbf{s}, \mathcal{H}) \tag{3}$$

where $\mathcal{H}$ represents the assumptions about the statistical model of the modelling errors and the prior model for the filter coefficients. Using Bayes' theorem, we can write

$$p(\mathbf{w}|y, \mathbf{s}, \mathcal{H}) \propto p(y|\mathbf{w}, \mathcal{H})p(\mathbf{s}|\mathbf{w}, \mathcal{H})p(\mathbf{w}|\mathcal{H}) \tag{4}$$

Since it is easier to work with the logarithm of the conditional density function, we define the following cost function

$$\begin{aligned} J(\mathbf{w}) &= -\log p(\mathbf{w}|y, \mathbf{s}, \mathcal{H}) \\ &= -\log p(y|\mathbf{w}, \mathcal{H}) - \log p(\mathbf{s}|\mathbf{w}, \mathcal{H}) - \log p(\mathbf{w}|\mathcal{H}) \end{aligned} \tag{5}$$

Note that we have omitted unrelated constants in the cost function. To obtain a MAP estimation of $\mathbf{w}$, we calculate the gradient of the cost function and set the result to zero.

Statistical models may be specified according to different considerations. For example, a Gaussian distribution can be used to model the filtering error. If robustness to outliers is required, we could use other distribution functions such as Laplacian and Huber's M-estimator to model the error. The second term in the cost function is mainly responsible for the smoothness constraint for the filter coefficient vector from the previous iteration to the current iteration. This constraint can be imposed by using a Gaussian distribution or a generalized Gaussian. It also affects the convergence rate as well as the stability of the iterative algorithm. Generally speaking, tightening this distribution increases smoothness and improves stability of the iteration at the expense of the convergence rate. The last term in the cost function is related to the prior distribution of the filter coefficients. The simplest choice is

0-7803-8834-8/05/$20.00 ©2005 IEEE.

	$p(y\|\mathbf{w},\mathcal{H})$	$p(\mathbf{s}\|\mathbf{w},\mathcal{H})$	$p(\mathbf{w}\|\mathcal{H})$
Section 3	Gaussian	Gaussian	uniform
Section 4	Laplacian, M-est.	Gaussian	uniform
Section 5	Gaussian	gen. Gaussian	uniform
Section 6	Gaussian	Gaussian	gen. Gaussian

Table 1. Model settings in sections 3 to 6.

the uniform distribution which makes the last term a constant. Using the uniform prior, the MAP estimation problem reduces to a maximum likelihood (ML) estimation problem.

However, interesting algorithms can be derived by setting the prior distribution to Gaussian, Laplacian and generalized Gaussian. Such settings are well justified in terms of controlling the model complexity to avoid over-fitting the data [14]. Due to space limitation, we only present algorithmic development results with certain model settings shown in Table 1. Other combinations of model settings can be studied following similar methods outlined in this paper.

3. ALGORITHMS BASED ON GAUSSIAN MODELS AND UNIFORM PRIOR

The cost function, ignoring the constants, can be expressed as

$$J(\mathbf{w}) = \frac{\beta}{2}(y - \mathbf{w}^T\mathbf{x})^2 + \frac{1}{2}\mathbf{r}^T\mathbf{A}^{-1}\mathbf{r} \tag{6}$$

where $\sigma_e^2 = 1/\beta$, is the variance of the modelling error, and $\mathbf{A}$ is the co-variance matrix.

3.1. Algorithms without error approximation

The MAP estimate of the filter coefficient vector is given by

$$\mathbf{w} = \mathbf{s} + \frac{\beta\hat{e}}{1 + \beta\mathbf{x}^T\mathbf{A}\mathbf{x}}\mathbf{A}\mathbf{x} \tag{7}$$

In a special case where $\mathbf{A}^{-1} = \text{diag}[\alpha_m]$, the filter coefficients are updated according to

$$w_m = s_m + \frac{\mu_m}{1 + \sum_{k=1}^{M}\mu_k x_k^2}\hat{e}x_m \tag{8}$$

where $\mu_m = \beta/\alpha_m$, which is the signal (coefficient) variance to noise variance ratio. In another special case where $\mathbf{A} = \sigma_w^2\mathbf{I}$, we have

$$\mathbf{w} = \mathbf{s} + \frac{\mu}{1 + \mu\sum_{k=1}^{M} x_k^2}\hat{e}\mathbf{x} \tag{9}$$

where $\mu = \sigma_w^2/\sigma_e^2$. Equation (9) represents a normalised LMS (NLMS) algorithm which is a special case of that represented by equation (8). We note that in the above development of the NLMS algorithm, the step-size μ is introduced as a natural result of the MAP optimization process. This is in contrast to the development in [1], by which the step-size μ is not a result of the optimization.

3.2. Algorithms with error approximation

The classical LMS algorithm can be derived when we use the approximation $e \approx \hat{e}$

$$\begin{aligned}\nabla J(\mathbf{w}) &= -\beta e\mathbf{x} + \mathbf{A}^{-1}\mathbf{r} \\ &\approx -\beta\hat{e}\mathbf{x} + \mathbf{A}^{-1}\mathbf{r}\end{aligned} \tag{10}$$

Therefore we have

$$\mathbf{w} = \mathbf{s} + \beta\hat{e}\mathbf{A}\mathbf{x} \tag{11}$$

In a special case where $\mathbf{A}^{-1} = \text{diag}[\alpha_m]$, the filter coefficients are updated according to

$$w_m = s_m + \mu_m\hat{e}x_m \tag{12}$$

where $\mu_m = \beta/\alpha_m$. This is the classical LMS algorithm with an individual adaptation constant for each filter coefficient. In another special case where $\mathbf{A} = \sigma_w^2\mathbf{I}$, we have the classical LMS algorithm

$$\mathbf{w} = \mathbf{s} + \mu\hat{e}\mathbf{x} \tag{13}$$

where $\mu = \sigma_w^2/\sigma_e^2$.

To summarize, we can see that the classical LMS and the NLMS algorithms can be easily derived from the MAP approach. In both algorithms, the adaptation step-size μ is expressed as the ratio of the coefficient variance to that of the error. As such, when we know the noise variance σ_e^2, μ is determined by σ_w^2. On the other hand, when σ_w^2 is fixed, μ can be estimated in each iteration through the estimation of the noise variance.

4. ROBUST LMS ALGORITHMS

In this section, we study LMS algorithms that are robust to a small number of large errors. We consider two models for the error: Laplacian and Huber's M-estimator. A Gaussian model is assumed for the filter coefficient vector, and to simplify our discussion in sub-section (4.2) , we assume that $\mathbf{A} = \sigma_w^2\mathbf{I}$ (see equation (14)).

4.1. LMS algorithms based on Laplacian distribution

The cost function is given by

$$J(\mathbf{w}) = \alpha|y - \mathbf{w}^T\mathbf{x}| + \frac{1}{2}\mathbf{r}^T\mathbf{A}^{-1}\mathbf{r} \tag{14}$$

We consider two cases. In the first case where $e = 0$, the problem becomes the following constrained optimization problem

$$\begin{array}{ll}\text{mimimize} & \frac{1}{2}\mathbf{r}^T\mathbf{A}^{-1}\mathbf{r} \\ \text{subject to} & \mathbf{w}^T\mathbf{x} = y\end{array} \tag{15}$$

This is the problem used in the original development of the NLMS algorithm.

In the second case, we consider $e \neq 0$. We have the following results

$$\nabla J(\mathbf{w}) = -\beta\text{sign}(e)\mathbf{x} + \mathbf{A}^{-1}\mathbf{r} = 0 \tag{16}$$

This equation is equivalent to the following equation

$$e = \hat{e} - \beta\text{sign}(e)\mathbf{x}^T\mathbf{A}\mathbf{x} \tag{17}$$

It can be easily shown that for the above equation to have a solution, the following conditions must be satisfied

$$|\hat{e}| > \beta\mathbf{x}^T\mathbf{A}\mathbf{x} \tag{18}$$

and

$$\text{sign}(e) = \text{sign}(\hat{e}) \tag{19}$$

Therefore, we have the MAP estimate for $\mathbf{w}$

$$\mathbf{w} = \mathbf{s} + \beta\text{sign}(\hat{e})\mathbf{A}\mathbf{x} \tag{20}$$

When $\mathbf{A} = \sigma_w^2\mathbf{I}$, we have the so-called signed-LMS algorithm

$$\mathbf{w} = \mathbf{s} + \mu\text{sign}(\hat{e})\mathbf{x} \tag{21}$$

where $\mu = \sigma_w^2/\sigma_e^2$.

4.2. LMS algorithms based on the M-estimator

Using Huber's formulation of M-estimator [9], the cost function is given by

$$J(\mathbf{w}) = \rho\left(\frac{\mathbf{y}-\mathbf{w}^T\mathbf{x}}{\sigma_e}\right) + \frac{1}{2\sigma_w^2}\mathbf{r}^T\mathbf{r} \tag{22}$$

where

$$\rho(t) = \begin{cases} \frac{1}{2}t^2, & \text{if } |t| \le \gamma \\ \gamma|t| - \frac{1}{2}\gamma^2, & \text{if } |t| > \gamma \end{cases} \tag{23}$$

Following the same procedure, we have the following results

$$\mathbf{w} = \begin{cases} \mathbf{s} + \mu\dfrac{\hat{e}}{1+\mu\mathbf{x}^T\mathbf{x}}\mathbf{x}, & \text{if } |\hat{e}| \le \delta \\ \mathbf{s} + \mu\gamma\sigma_e \text{sign}(\hat{e})\mathbf{x}, & \text{if } |\hat{e}| > \delta \end{cases} \tag{24}$$

where $\mu = \sigma_w^2/\sigma_e^2$ and $\delta = \gamma\sigma_e(1+\mu\mathbf{x}^T\mathbf{x})$. We can see that the above LMS algorithm switches between the two update options: the normalized LMS and signed-LMS. The switch is controlled by previous modelling error and the three parameters σ_e^2, σ_w^2 and γ. If the first two parameters are fixed, then the behaviour of the algorithm is mainly controlled by γ. If the value of γ is sufficiently large, then the algorithm is mainly a normalized LMS algorithm. On the other hand, if it is sufficiently small, then the algorithm is mainly a signed LMS algorithm.

5. PROPORTIONATE LMS ALGORITHMS

In this section, we consider a Gaussian model for the modelling error and a generalized Gaussian model for the filter coefficients

$$p(\mathbf{w}|\mathbf{s},\mathcal{H}) = c\exp\left(-\alpha\sum_{m=1}^{M}|r_m|^p\right) \tag{25}$$

where c is a normalization constant, α and p are two parameters. The Gaussian and Laplacian distributions are special cases where $p = 2$ and $p = 1$. When $0 \le p \le 1$, it has been demonstrated that sparse solutions are possible [15–17]. The cost function and its gradient are given by

$$J(\mathbf{w}) = \frac{\beta}{2}e^2 + \alpha\sum_{m=1}^{M}|r_m|^p \tag{26}$$

and

$$\nabla J(\mathbf{w}) = -\beta e\mathbf{x} + p\alpha\mathbf{D}^{-1}\mathbf{r} \tag{27}$$

where $\mathbf{D} = \text{diag}[|r_m|^{2-p}]$. Determining a vector $\mathbf{w}$ that minimizes $J(\mathbf{w})$ for $p \neq 2$ is not a trivial problem. For $p = 1$, the problem can be casted as a second order cone program [18]. An alternative way is to use the idea of iterative re-weighted least squares [19] which iteratively solves equation (27) for $\mathbf{w}$ by assuming a fixed matrix $\mathbf{D}$. However, both methods require substantial amount of computation when compared to that required by the classical LMS algorithm.

In order to simplify the algorithm, we follow a similar idea as that of the EM algorithm [20] and replace the elements of $\mathbf{D}$ with their respective minimum mean square error estimate

$$\hat{a}_m = \int |r_m|^{2-p} p(w_m|y,s_m,\mathcal{H})dw_m. \tag{28}$$

As such, we have $\hat{\mathbf{D}} = \text{diag}[\hat{a}_m]$. In order to obtain a closed-form solution, we use Taylor series as an approximation

$$|w_m - s_m|^{2-p} \approx |s_m|^{2-p} + (2-p)\frac{|s_m|^{2-p}}{s_m}w_m \tag{29}$$

With this approximation, we can show that

$$\hat{a}_m = (3-p)|s_m|^{2-p} \tag{30}$$

Therefore, the MAP estimate is given by

$$\mathbf{w} = \mathbf{s} + \frac{\hat{e}}{\frac{p\alpha}{\beta} + \mathbf{x}^T\hat{\mathbf{D}}\mathbf{x}}\hat{\mathbf{D}}\mathbf{x} \tag{31}$$

It is interesting to note that when we use error approximation $e \approx \hat{e}$ in equation (27), we have

$$\mathbf{w} = \mathbf{s} + \frac{\beta}{p\alpha}\hat{e}\hat{\mathbf{D}}\mathbf{x} \tag{32}$$

When we substitute $\hat{a}_m$ into equations (31) and (32), we have the following updating equations

$$w_m = s_m + \frac{\hat{e}}{\frac{p\alpha}{(3-p)\beta} + \sum_{k=1}^{M}|s_k|^{2-p}x_k^2}|s_m|^{2-p}x_m \tag{33}$$

and

$$w_m = s_m - \frac{(3-p)\beta}{p\alpha}|s_m|^{2-p}\hat{e}x_m \tag{34}$$

We see that equations (33) and (34) represent a family of the proportionate normalized LMS (PNLMS) and the proportionate LMS (PLMS) algorithms, respectively. We can also see that the two equations (31) and (7) are similar. In fact, when $p = 2$, equation (31) is in the same form as equation (9). However, there is a distinctive difference between the two. In equation (7), the matrix A is a co-variance matrix with free parameters that serve as adaptation constants. On the other hand, in equation (31), the matrix $\hat{\mathbf{D}}$ is a diagonal matrix whose elements are given by equation (30). This leads to different results for different settings of p. Therefore, equation (7) represents a general proportionate normalized LMS algorithm (PNLMS) with matrix $\boldsymbol{A}$ to be specified, while equation (31) represents a specific PNLMS algorithm.

6. LMS ALGORITHM WITH SPARSITY CONSTRAINT

In this section, we consider Gaussian models for the filtering error and the smoothness constraint, and a generalized Gaussian distribution for the prior density. We have the general form of the cost function

$$J(\mathbf{w}) = \frac{1}{2}\beta e^2 + \frac{1}{2}\alpha\mathbf{r}^T\mathbf{r} + \frac{1}{p}\gamma||\mathbf{w}||_p \tag{35}$$

where $||\mathbf{w}||_p = \sum_{m=1}^{M}|w_m|^p$, $0 \le p \le 1$. The gradient is given by

$$\nabla J(\mathbf{w}) = -\beta e\mathbf{x} + \alpha(\mathbf{w}-\mathbf{s}) + \gamma\mathbf{B}\mathbf{w} \tag{36}$$

where $\mathbf{B} = \text{diag}[|w_m|^{p-2}]$.

It can be seen that setting equation (36) to zero results in a set of nonlinear equations. A closed form solution for $\mathbf{w}$ is not possible. We study a solution based on two approximations: $e \approx \hat{e}$ and $\mathbf{B} \approx \text{diag}[|s_m|^{p-2}]$. With these approximations, we can derive an update-equation for each element of the coefficient vector

$$w_m = \eta_m(s_m + \frac{\beta}{\alpha}\hat{e}x_m) \tag{37}$$

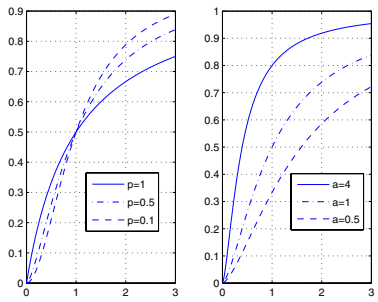

Fig. 1. The nonlinear relationship between s_m (horizontal axis) and η_m (vertical axis). In the left panel, we fix $a = \alpha/\gamma = 1$ and plot three cases $p = 1, 0.5, 0.1$. In the right panel, we fix $p = 0.5$ and plot three cases where $\alpha/\gamma = 4, 1, 0.5$.

where

$$\eta_m = 1 - \frac{1}{\frac{\alpha}{\gamma}|s_m|^{2-p} + 1} \tag{38}$$

We can see that this is a shrinkage-version of the classical LMS algorithm. The shrinking factor is η_m. The ratio α/γ controls the relative weighting we impose on the two constraints. The shrinking factor η_m reduces towards 0 as $|s_m|$ decreases. Noting that w_m and s_m are the current and previous results from iterative algorithm, $w_m = 0$ is a fixed point of the nonlinear update equation. Furthermore, $\eta_m \approx \frac{\alpha}{\gamma}|s_m|^{2-p}$ when $|s_m|$ is sufficiently small, whence $w_m = 0$ is a stable attractor as long as $\frac{\alpha}{\gamma} < 1$. Consequently, robust sparseness becomes a built-in property of the iteration.

To understand the role of the shrinking factor, we plot it as a function of s_m under different conditions. In the left panel of Figure 1, we show η_m as a function of s_m. We fix $\alpha/\gamma = 1$ and plot three cases where $p = 1, 0.5, 0.1$. In the right panel, we fix $p = 0.5$ and plot three cases where $\alpha/\gamma = 4, 1, 0.5$. We can see that there is nonlinear relationship between η_m and s_m. A smaller value of $|s_m|$ leads to a smaller η_m, which makes a larger shrinkage.

7. CONCLUSION

In this paper, we propose a unified approach for developing the LMS algorithms. We formulate the problem as an MAP estimation problem which permits us to explore different model settings based on practical considerations. In particular, we have developed several well known LMS algorithms as well as algorithms that are results of MAP optimization with desirable constraints such as robustness and sparseness.

8. REFERENCES

[1] S. Haykin, *Adaptive Filter Theory*, 3rd ed. Englewood Cliffs, New Jersey, USA: Prentice Hall Inc., 1996.

[2] W. A. Sethares, "The least mean square family," in *Adaptive system identification and signal processing algorithms*, N. Kalouptsidis and S. Theodoridis, Eds. Prentice Hall, 1993, pp. 84–122.

[3] S. C. Douglas, "A family of normalized LMS algorithms," *IEEE Signal Processing Lett.*, vol. 1, no. 3, pp. 49–51, March 1994.

[4] B. D. Rao and B. Y. Song, "Adaptive filtering algorithms for promoting sparsity," in *Proc. IEEE ICASSP*, vol. VI, 2003, pp. 361–364.

[5] D. L. Duttweiler, "Proportionate normalized least-mean-squares adaptation in echo cancelers," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 8, no. 5, pp. 508–518, September 2000.

[6] J. Benesty, Y. Huang, and D. R. Morgan, "On a class of exponentiated adaptive algorithms for the identification of sparse impulse responses," in *Adaptive Signal Processing Applications to Real-World Problems*, J. Benesty and Y. Huang, Eds. Springer, 2003, ch. 1, pp. 1–22.

[7] S. L. Gay and S. C. Douglas, "Normalized natural gradient adaptive filtering for sparse and non-sparse systems," in *Proc. IEEE ICASSP*, vol. II, 2002, pp. 1405–1408.

[8] R. K. Martin, W. A. Sethares, R. C. Williamson, and J. C. R. Johnson, "Exploiting sparsity in adaptive filters," *IEEE Trans. Signal Processing*, vol. 50, no. 8, pp. 1883–1894, August 2002.

[9] P. J. Huber, *Robust Statistics*. New York: John Wiely, 1981.

[10] P. Petrus, "Robust huber adaptive filter," *IEEE Trans. Signal Processing*, vol. 47, no. 4, pp. 1129–1133, April 1999.

[11] S. C. Bang and S. Ann, "A robust algorithm for adaptive FIR filtering and its performance analysis with additive contaminated-gaussian noise," *IEEE Trans. Circuits Syst. I*, vol. 43, no. 5, pp. 361–369, May 1996.

[12] J. Chambers and A. Avlonitis, "A robust mixed norm adaptive filter algorithm," *IEEE Signal Processing Lett.*, vol. 4, no. 2, pp. 46–48, February 1997.

[13] M. Nikolova, "Regularization functions and estimators," in *Proc. IEEE International Conference on Image Processing*, vol. 2, Sept. 1996, pp. 457–460.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag, 2001.

[15] J. Karvanen and A. Cichocki, "Measureing sparseness of noisy signals," in *4th International Symposium on Independent Component Analysis and Blind Signal Separation*, April 2003, pp. 125–130.

[16] R. Gonin and A. H. Money, *Nonlinear Lp-norm estimation*. New York and Basel: Marcel Dekker, Inc., 1989.

[17] A. Antoniadis and J. Q. Fan, "Regularization of wavelet approximations," *Journal of the American Statistical Association*, vol. 96, pp. 939–955, Sept. 2001.

[18] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge, U.K.: Cambridge University Press, 2004.

[19] E. E. Osborne, *Finite algorithms in optimization and data analysis*. New York: John Wiely, 1985.

[20] A. Gelman, H. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*, 2nd ed. Chapman & Hall/CRC, 2004.