CROSS CORRELATION P-VECTOR INFLUENCE ON LMS CONVERGENCE

Luis Vicente, Enrique Masgrau
Dept. of Computer Science and Systems Eng. & Dept. of Electronics and Communications Eng.
University of Zaragoza
María de Luna, 3 - E50015 Zaragoza SPAIN
e-mail: lvicente@posta.unizar.es

ABSTRACT

The principal weakness of Least Mean Squares (LMS) algorithm is that adaptation can be sometimes slow. Convergence is known to depend mainly on eigenvalue spread of the input signal, through the time constants of the various convergence modes. However, most LMS convergence analysis do not consider the influence of cross correlation between input and desired output signals, which plays also a significant role on convergence and is the main topic of this paper. The extreme cases of high and low statistical similarity between input and desired output are analysed in detail. Furthermore, an LMS-based adaptive system that seizes the convergence properties explored is also introduced. This system is shown to achieve better performance (that is, faster convergence while maintaining the steady-state error level) than LMS when input and desired output present low or moderately low statistical similarity.

1 INTRODUCTION

The Least Mean Squares (LMS) algorithm is, surely, the most widely used adaptive system, due to its simplicity as well as robustness [1]. The main disadvantage of LMS is that convergence can be sometimes slow. This happens when the eigenvalues of the input signal are disparate and slow modes of adaptation dominate the settling time. The input signal determines by itself whether a mode is fast or slow, since the time constant of a convergence mode depends only on the corresponding eigenvalue of the input signal, and on the adaptation step size, which is the same for every mode. Nevertheless, it is the relative excitation of convergence modes what makes them dominant or non-dominant. In this paper, we show that modes excitation depends on cross correlation between the input and the desired output. Therefore, it is possible to obtain fast adaptation even with high eigenvalue spread, if the excitation of convergence modes and the steady-state mean-square error are such that slow modes are completely unnoticed.

In next section, we analyse LMS convergence in order to reach a formulation for modes excitation as a function of cross correlation between input and desired output. For better text understanding and completeness, some well known results on LMS convergence are reproduced here, following [1] and [2] . Subsequently, convergence for the extreme situations, from the point of view of statistical similarity between input and desired output, is examined. Experimental results are also provided.

In section 3, an LMS-based system, which takes advantage of the results exposed here to speed up convergence, is introduced and compared to single LMS.

2 CONVERGENCE ANALYSIS

Convergence of any adaptive system can be analysed by means of the *learning curve* of the algorithm, which is a plot of the cost function versus time. For LMS, we have $\xi_n \equiv E\{e_n^2\}$, where $e_n = d_n - \mathbf{x}_n^T\mathbf{w}_n$ is the instantaneous error signal. Thus,

$$\xi_n = E\{d_n^2\} - 2E\{d_n\mathbf{x}_n^T\mathbf{w}_n\} + E\{\mathbf{w}_n^T\mathbf{x}_n\mathbf{x}_n^T\mathbf{w}_n\} \quad (1)$$

We can assume that the data signal, x_n, and the LMS weights, $\mathbf{w}_n$, are uncorrelated with each other [1], [2]. This assumption holds when the data signal changes much more rapidly than the mean value of the weights (*slow adaptation* approach). Making use of the previous assumption and the usual definitions for the correlation matrix, $\mathbf{R} = E\{\mathbf{x}_n\mathbf{x}_n^T\}$, and the cross correlation vector, $\mathbf{p} = E\{\mathbf{x}_n d_n\}$, (1) can be rewritten as

$$\xi_n = \sigma_d^2 - 2\mathbf{p}^T E\{\mathbf{w}_n\} + E\{\mathbf{w}_n^T\mathbf{R}\mathbf{w}_n\} \quad (2)$$

where $\sigma_d^2 = E\{d_n^2\}$ is the desired output power.

Since the LMS weight vector fluctuates, it can be modelled as the sum of the mean weight vector, $\overline{\mathbf{w}}_n = E\{\mathbf{w}_n\}$, plus a stochastic noisy component

$$\mathbf{w}_n = \overline{\mathbf{w}}_n + \widetilde{\mathbf{w}}_n \quad (3)$$

From the previous definition it follows that $E\{\widetilde{\mathbf{w}}_n\} = 0$ and $E\{\overline{\mathbf{w}}_n\} = \overline{\mathbf{w}}_n$. Using (3), the last term of the sum

in (2) is given by

$$\begin{aligned} E\left\{\mathbf{w}_n^T\mathbf{R}\mathbf{w}_n\right\} &= E\left\{\left(\overline{\mathbf{w}}_n^T+\widetilde{\mathbf{w}}_n^T\right)\mathbf{R}\left(\overline{\mathbf{w}}_n+\widetilde{\mathbf{w}}_n\right)\right\} \\ &= E\left\{\overline{\mathbf{w}}_n^T\mathbf{R}\overline{\mathbf{w}}_n\right\}+E\left\{\overline{\mathbf{w}}_n^T\mathbf{R}\widetilde{\mathbf{w}}_n\right\} \\ &\quad +E\left\{\widetilde{\mathbf{w}}_n^T\mathbf{R}\overline{\mathbf{w}}_n\right\}+E\left\{\widetilde{\mathbf{w}}_n^T\mathbf{R}\widetilde{\mathbf{w}}_n\right\} \\ &= \overline{\mathbf{w}}_n^T\mathbf{R}\overline{\mathbf{w}}_n+E\left\{\widetilde{\mathbf{w}}_n^T\mathbf{R}\widetilde{\mathbf{w}}_n\right\} \end{aligned} \quad (4)$$

The minimum mean-square error, obtained when the weight vector is set at its optimal value, $\mathbf{w}_{opt}=\mathbf{R}^{-1}\mathbf{p}$, is given by $\xi_{min}=\sigma_d^2-\mathbf{w}_{opt}^T\mathbf{R}\mathbf{w}_{opt}$. Using this expression and (4) and rearranging, eq. (2) becomes

$$\begin{aligned} \xi_n &= \xi_{min}+\mathbf{w}_{opt}^T\mathbf{R}\mathbf{w}_{opt}-2\mathbf{w}_{opt}^T\mathbf{R}\overline{\mathbf{w}}_n+\overline{\mathbf{w}}_n^T\mathbf{R}\overline{\mathbf{w}}_n \\ &\quad +E\left\{\widetilde{\mathbf{w}}_n^T\mathbf{R}\widetilde{\mathbf{w}}_n\right\} \\ &= \xi_{min}+\left(\overline{\mathbf{w}}_n^T-\mathbf{w}_{opt}^T\right)\mathbf{R}\left(\overline{\mathbf{w}}_n-\mathbf{w}_{opt}\right) \\ &\quad +E\left\{\widetilde{\mathbf{w}}_n^T\mathbf{R}\widetilde{\mathbf{w}}_n\right\} \end{aligned} \quad (5)$$

This expression is equal to the one obtained for the steepest descent algorithm, except for the last term in the sum. According to [2], the following expression is derived assuming the weight noise, $\widetilde{\mathbf{w}}_n$, is uncorrelated with the data, x_n,

$$\begin{aligned} E\left\{\widetilde{\mathbf{w}}_n^T\mathbf{R}\widetilde{\mathbf{w}}_n\right\} &= \sigma_x^2 Tr\left\{E\left\{\widetilde{\mathbf{w}}_n^T\widetilde{\mathbf{w}}_n\right\}\right\} \\ &\approx \sigma_x^2\sum_{l=1}^{L}\frac{\mu\xi_{min}}{(1-\mu\lambda_l)} \end{aligned} \quad (6)$$

where μ is the LMS adaptation step size and λ_l are the eigenvalues of the correlation matrix. Consequently, this term is only due to the excess mean-square error. That is, $\xi_\infty=\xi_{min}+E\left\{\widetilde{\mathbf{w}}_n^T\mathbf{R}\widetilde{\mathbf{w}}_n\right\}$, and

$$\xi_n-\xi_\infty=\left(\overline{\mathbf{w}}_n^T-\mathbf{w}_{opt}^T\right)\mathbf{R}\left(\overline{\mathbf{w}}_n-\mathbf{w}_{opt}\right) \quad (7)$$

We can use in (7) the usual weight vector transformations, $\mathbf{v}_n=\mathbf{w}_n-\mathbf{w}_{opt}$ and $\mathbf{v}'_n=\mathbf{Q}^T\mathbf{v}_n$, where $\mathbf{R}=\mathbf{Q}\Lambda\mathbf{Q}^T$is the normal form of the correlation matrix, being $\mathbf{Q}$ the eigenvector matrix and Λ the eigenvalue matrix. Thus,

$$\xi_n-\xi_\infty=\overline{\mathbf{v}}_n^T\mathbf{R}\overline{\mathbf{v}}_n=\overline{\mathbf{v}}_n'^T\Lambda\overline{\mathbf{v}}'_n \quad (8)$$

According to [1], we have

$$\overline{\mathbf{v}}'_n=\left(\mathbf{I}-2\mu\Lambda\right)^n\mathbf{v}'_0 \quad (9)$$

Taking into account that

$$\mathbf{Q}^T\mathbf{w}_{opt}=\mathbf{Q}^T\mathbf{R}^{-1}\mathbf{p}=\mathbf{Q}^T\mathbf{Q}\Lambda^{-1}\mathbf{Q}^T\mathbf{p}=\Lambda^{-1}\mathbf{Q}^T\mathbf{p} \quad (10)$$

and after some computations, (9) becomes

$$\overline{\mathbf{v}}'_n=\left(\mathbf{I}-2\mu\Lambda\right)^n\left(\mathbf{Q}^T\mathbf{w}_0-\Lambda^{-1}\mathbf{Q}^T\mathbf{p}\right) \quad (11)$$

Eventually, using (11) and considering that $\left(\mathbf{I}-2\mu\Lambda\right)^n\Lambda\left(\mathbf{I}-2\mu\Lambda\right)^n$ is a diagonal matrix, we derive the final expression for the mean-square error

$$\begin{aligned} \xi_n-\xi_\infty &= \sum_{l=1}^{L}\left(1-2\mu\lambda_l\right)^{2n}\lambda_l\left(\mathbf{q}_l^T\mathbf{w}_0-\frac{\mathbf{q}_l^T\mathbf{p}}{\lambda_l}\right)^2 \\ &= \sum_{l=1}^{L}K_l\exp\left(-\frac{n}{\tau_l}\right) \end{aligned} \quad (12)$$

where $\mathbf{q}_l$ is the eigenvector associated with the eigenvalue λ_l and L is the order of the adaptive filter.

It is clear from (12) that the learning curve is the sum of L decaying exponential functions when the adaptation step size, μ, is properly chosen. Each of these exponential functions is called a convergence *mode*, which is defined by its *time constant,*

$$\tau_l\approx\frac{1}{4\mu\lambda_l} \quad (13)$$

and its *excitation,*

$$K_l=\lambda_l\left(\mathbf{q}_l^T\mathbf{w}_0-\frac{\mathbf{q}_l^T\mathbf{p}}{\lambda_l}\right)^2 \quad (14)$$

Thus, (14) relates the excitation of an adaptation mode with its corresponding eigenvalue, which also determines the mode time constant, and with the input-desired output cross correlation, through the cross correlation vector $\mathbf{p}$.

In the *white input* case, there is just one convergence mode, $K\exp\left(-n/\tau\right)$, since there is only one eigenvalue, λ, with multiplicity L. The excitation K is the sum of every mode excitations in (14). Therefore, convergence speed can be easily optimised in this case.

2.1 High statistical similarity between input and desired output

The maximum statistical similarity is obtained when the desired output is simply a delayed version of the input signal, $d_n=Ax_{n-\delta}$. When the delay is inside the filter span, that is, when $0\leq\delta<L$, the cross correlation vector is proportional to one of the columns in the input correlation matrix

$$\mathbf{p}=\gamma_{xd}\mathbf{r}_\delta=\gamma_{xd}\sum_{j=1}^{L}\lambda_j q_{j,\delta}\mathbf{q}_j \quad (15)$$

being $q_{j,\delta}$ the δ-th component of the j-th eigenvector. Using the orthogonality property between eigenvectors, we find that

$$\mathbf{q}_l^T\mathbf{p}=\mathbf{q}_l^T\gamma_{xd}\sum_{j=1}^{L}\lambda_j q_{j,\delta}\mathbf{q}_j=\lambda_l q_{l,\delta}\gamma_{xd} \quad (16)$$

Therefore, using (16) in (14) we find the excitation of convergence modes for this case to be

$$K_l=\lambda_l\left(\mathbf{q}_l^T\mathbf{w}_0-q_{l,\delta}\gamma_{xd}\right)^2 \quad (17)$$

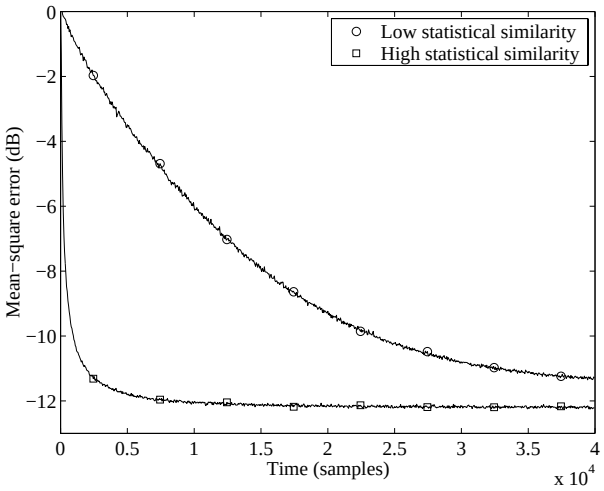

Figure 1: Learning curves comparison for high and low statistical similarity between input and desired output.

The term $\left(\mathbf{q}_l^T \mathbf{w}_0 - q_{l,\delta}\gamma_{xd}\right)^2$ is a positive value completely independent from the eigenvalue λ_l. Therefore, the excitation of each mode is, in the case of high statistical similarity, directly proportional to its own eigenvalue.

2.2 Low statistical similarity between input and desired output

In this case the desired output is correlated with the input signal, to make possible some mean-square error reduction (we do not consider here the trivial case of no cross correlation), but the cross correlation function is just an impulse. Thus, cross correlation vector components are zero except for one of them

$$\mathbf{p} = [0, \cdots, 0, \gamma_{xd}, 0, \cdots, 0]^T \tag{18}$$

Consequently, we have

$$\mathbf{q}_l^T \mathbf{p} = q_{l,\delta}\gamma_{xd} \tag{19}$$

Substituting (19) in (14), the modes excitation is given by

$$K_l = \lambda_l \left(\mathbf{q}_l^T \mathbf{w}_0 - \frac{q_{l,\delta}\gamma_{xd}}{\lambda_l}\right)^2 \tag{20}$$

Supposing that $\mathbf{q}_l^T \mathbf{w}_0 \neq 0$, when λ_l is very high the excitation can be approximated by

$$K_l \approx \lambda_l \left(\mathbf{q}_l^T \mathbf{w}_0\right)^2 \tag{21}$$

that is to say, excitation is again directly proportional to the corresponding eigenvalue. Nevertheless, when λ_l is very low, the excitation will be given by

$$K_l \approx \frac{\left(q_{l,\delta}\gamma_{xd}\right)^2}{\lambda_l} \tag{22}$$

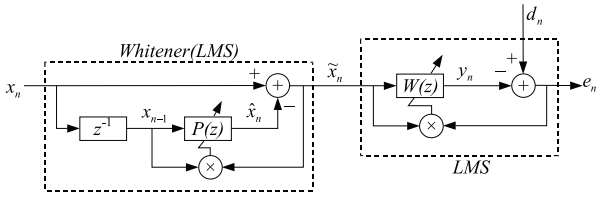

Figure 2: Whitener+LMS system block diagram.

So, the excitation is in this case inversely proportional to the eigenvalue of the convergence mode.

Note that $\mathbf{q}_l^T \mathbf{w}_0 = 0$ is a very common case, since usually $\mathbf{w}_0 = 0$ to avoid an increase in the mean-square error level at the beginning of the algorithm. In this case the excitation is also given by eq. (22), independently of the eigenvalue. Therefore, every convergence mode will be excited in a way inversely proportional to its eigenvalue.

To sum up, in the low statistical similarity case, slow modes are very excited whereas fast modes will be very or little excited depending on the term $\mathbf{q}_l^T \mathbf{w}_0$. In any case, slow modes can dominate convergence.

Figure 1 shows the learning curves for the same input signal (with eigenvalue spread $\lambda_{max}/\lambda_{min} \approx 240$) and two different desired outputs obtained by computer simulation. In both cases there was an uncorrelated noise added to the desired output, limiting the achievable cancellation to the same level, and the same filter lengths and adaptation step sizes were used. It is clearly appreciated how the low similarity case converges much more slowly than the high similarity one, in accordance with our previous discussion.

3 FAST CONVERGENCE SYSTEM

In figure 2 it is depicted a system that seizes the convergence properties discussed in the previous section by means of a "divide & conquer" approach. The aim of this system is to obtain faster convergence than LMS without affecting the steady-state mean-square error performance. This system has two cascaded stages, the first of them being a *whitener*, that is, a prediction error filter, that pre-processes or conditions the input signal for the second one, which is an LMS.

The prediction filter in the whitener, $P(z)$, is adaptive, and uses the LMS as adaptation algorithm. In this case the desired output is the input signal delayed by one sample. According to our previous discussion, convergence of this stage is expected to be fast, since there is high statistical similarity between input, x_{n-1}, and desired output, x_n.

The second stage is just an LMS system with whitened input, $\widetilde{x}_n$. As we said before, a white input signal means that there is only one convergence mode,

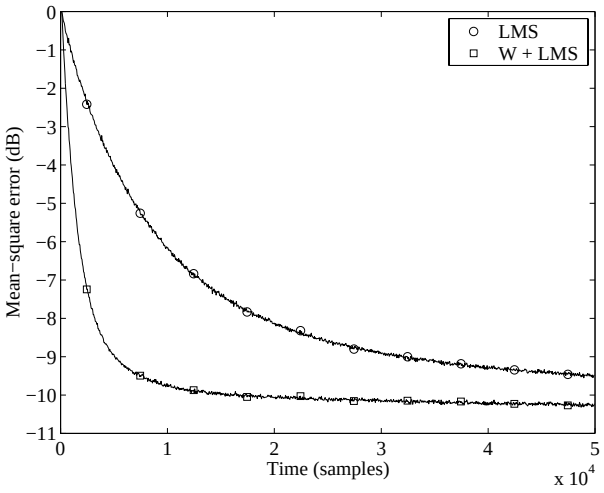

Figure 3: Learning curves comparison: Whitener + LMS vs. single LMS.

and so, there are no fast nor slow modes. The efficiency of LMS approaches in this case its theoretical limit [1]. However, $\widetilde{x}_n$ whiteness will depend on the whitening capability of the first system. In any case, $\widetilde{x}_n$ will be whiter than the original input, x_n and eigenvalue spread will be at least reduced.

Experimental results were obtained with the same signals for the *Whitener + LMS* system and single LMS. The learning curves for both systems, in a case of moderately low statistical similarity between x_n and d_n, are shown in figure 3. It can be seen that the Whitener + LMS system achieves much better performance than single LMS, since convergence is faster while maintaining the steady-state error level.

Another system, called ALE + FxLMS (also Whitener + FxLMS), that exploits the convergence properties analysed here is exposed in references [3] and [4]. This system is similar to the Whitener + LMS system, with the second stage being a Filtered-x LMS system instead of an LMS. It was derived in the context of active control of sound and vibration.

4 CONCLUSIONS

In this paper we have analysed convergence properties of LMS systems depending on the statistical similarity between the input and the desired output signals. It has been seen that high statistical similarity is wished in order to achieve fast convergence, even when the input signal presents great eigenvalue spread. Since the input signal and the desired output are not normally a choice, the Whitener + LMS system, introduced here, divides the adaptive problem in two different stages that are expected to converge faster than the one-stage or single LMS. Experimental results confirm that the Whitener + LMS system can perform much better than single LMS.

Acknowledgements

This work was supported by the Spanish Education and Culture Ministry through its National Program on Environment (research project: AMB99-1095-C02-02).

References

[1] Widrow, B., Stearns, S. D., *Adaptive Signal Processing*, New Jersey: Prentice-Hall, 1985, ch. 6.

[2] Alexander, S. T., *Adaptive Signal Processing: Theory and Applications*, New York: Springer-Verlag, 1986, ch. 5.

[3] Vicente, L., Elliott, S. J., Masgrau, E., "Fast Active Noise Control for Robust Speech Acquisition" in *Proceedings of Eurospeech*, Budapest, pp. 2403-2406, 1999.

[4] Vicente, L., Masgrau, E., "Performance Comparison of Two Fast Algorithms for Active Control" in *Proceedings of Active 99*, Fort Lauderdale, Florida, pp. 1089-1100, 1999.