LMS: PAST, PRESENT AND FUTURE.

Victor Solo

School of Electrical Engineering and Telecommunications
University of New South Wales
Sydney, AUSTRALIA
email: v.solo@unsw.edu.au

ABSTRACT

We sketch some aspects of LMS particularly focussing on some puzzles, problems and potentials. We explain the 'independence heuristic' by means of averaging theory for which we give a simple expose'. We suggest an explanation for the 'urban myth' that white noise only performance formulae can be used as surrogates for the correct performance formulae which involve autocorrelations.

We then extend the discussion to more recent network versions such as diffusion LMS. We show that the sharing aspect of network LMS algorithms induces a two time scale structure (something not yet widely known) and exhibit its consequences. We comment on the recent upsurge of interest in 'online learning' in machine learning and its failure to reference the adaptive signal processing literature. Finally we speculate on future developments of LMS.

Index Terms— LMS, adaptive, online, averaging

1. INTRODUCTION

Adaptive or online estimation algorithms developed in the 1950s in two separate disciplines: control and signal processing. In control the seminal algorithm was the 'MIT rule' [1], while in signal processing it was the Least Mean Square (LMS) algorithm of Widrow and Hoff [2],[3]. Here we concentrate on the signal processing aspects.

Further development has continued in both disciplines to the present time, although without much interaction. This is partly explained by the fact that in signal processing one is mostly in an open loop setting whereas in control one is in a closed loop setting. More recently online algorithms have attracted attention in machine learning from two directions. Firstly from reinforcement learning and so dating from the mid 1980s [4],[5]. Secondly due to the emergence of streaming data applications and so dating from the early 2000s [6].

No aspect of online parameter estimation has escaped attention: algorithm develoment; algorithm convergence; algorithm performance; and applications. Thus variants of LMS such as NLMS [7],[8] soon emerged (and continue to do so e.g. [9]) and more recently network versions are under development [10]. Convergence analysis remains a challenging problem as does performance analysis [11],[12],[13]. Applications - in diverse areas [14],[15] (two notable success areas are Telecommunications and Biomedical Engineering) continue to emerge particularly driven recently in the signal processing community by sensor network applications [10],[16],[17].

It would require a full length journal paper to even attempt to survey just some of these aspects. Instead here we focus on some puzzles and blindspots of algorithm development, convergence and performance.

Before continuing let us note the parallel tradition in online estimation of algorithms with decaying gains e.g. [18],[19]. These algorithms have limited use in practice because they are incapable of tracking time-varying parameters since they lose the ability to adjust to new information [12]. On the other hand they are often easier to analyse than fixed gain algorithms and so can provide insight in complicated cases. Here we discuss only fixed gain algorithms.

The remainder of the paper is organised as follows. In section 2 we discuss the first order behaviour of LMS. In section 3 we discuss performance i.e. second order behaviour. In section 4 we discuss network versions of LMS. In section 5 are brief speculations on the future and conclusions. Lack of space precludes a treatment of tracking: see [20].

2. THE FIRST ORDER BEHAVIOUR OF LMS

LMS deals with the problem of estimating, in real time, a time-varying vector parameter or weight sequence w_t in the regression problem

$$y_t = x_t^T w_t + n_t \tag{2.1}$$

where y_t is an observed scalar sequence; x_t is an observed d-vector regressor sequence and n_t is an unobserved noise sequence. The problem is to track w_t.

Usually the problem is approached into two stages:

static estimation when w_t is static or at equilibrium i.e. $w_t = w_e$ for all t

time-varying estimation when w_t varies with time.

The static case is simpler than the time-varying case but provides a lot of useful insightful for studying the time-varying case. For lack of space we limit our discussion to the static case; see [12],[11],[21] for the time-varying case.

2.1. Derivation

LMS is simply a steepest descent algorithm for minimizing the instantaneous squared error criterion $J_t(w) = \frac{1}{2}e_t^2$ where $e_t = e_t(w) = y_t - x_t^T w$ is the error signal. It has the classic form of all adaptive algorithms

$$\begin{aligned} w_{new} &= w_{old} + gain \times gradient \times error \\ \hat{w}_t &= \hat{w}_{t-1} + \mu x_t e_t \\ &= \hat{w}_{t-1} + \mu x_t (y_t - x_t^T \hat{w}_{t-1}) \end{aligned} \tag{2.2}$$

We rewrite this in 'difference' form

$$\delta\hat{w}_t = \hat{w}_t - \hat{w}_{t-1} = \mu x_t (y_t - x_t^T \hat{w}_{t-1})$$

2.2. Assumptions

We introduce the following assumptions.

A1. The sequences n_t, x_t are strictly stationary and statistically independent. Further x_t has zero mean with variance matrix R_x and n_t has zero mean with variance σ_n^2. Also x_t, n_t have corresponding autocovariance sequences γ_k^x, γ_k^n.

A2. The matrix $A = I - \mu R_x$ is a stability matrix. This is equivalent to the condition that all eigenvalues λ_i of R_x satisfy $0 < \mu\lambda_i < 2$.

2.3. Error System

To analyse the behaviour of the algorithm we need to form an error system. Introduce the estimation error $\tilde{w}_t = \hat{w}_t - w_e$. Then from the LMS update (2.2) and the regression model (2.1) we find

$$\delta\tilde{w}_t = -\mu x_t x_t^T \tilde{w}_{t-1} + \mu x_t n_t \tag{2.3}$$

This is the LMS error system and is a time-varying stochastic difference equation (sde) whose behaviour is by no means simple to analyse.

2.4. Convergence via the Independence Heuristic

We proceed initially in a traditional manner and introduce the:

Independence Heuristic[1] **(IH)**. $\tilde{w}_{t-1}$ and n_t are treated as being statistically independent.

[1] A better word than assumption

This heuristic gained wide attention following [22] but has an earlier origin: see [8] for further details and references.

Denote $\tilde{m}_t = E(\tilde{w}_t)$. Then using A1 and the IH and taking iterated conditional expectations through (2.3) gives

$$\delta\tilde{m}_t = -\mu R_x \tilde{m}_{t-1} \equiv \tilde{m}_t = (I - \mu R_x)\tilde{m}_{t-1} \tag{2.4}$$

For reasons explained below we call this difference equation the 'averaged system'. It is well known that stability of the averaged system i.e. convergence to 0 occurs iff A2 holds.

While this result has been obtained by an heuristic argument it is borne out remarkably well in practice [14]. The question is why?

This remains a topic of debate [23],[24]. But by far the best explanation is to be found in averaging theory [12].

2.5. Convergence via Averaging

We sketch the argument. Starting at an arbitrary time T we sum the error system over a time interval of length N to get

$$\tilde{w}_{T+N} - \tilde{w}_T = -\mu\Sigma_T^{T+N} x_t x_t^T \tilde{w}_{t-1} + \mu\Sigma_T^{T+N} x_t n_t$$

Noow if μ is small and N is not to large then for $t \in [T, T+N]$ $\tilde{w}_{t-1}$ will not have changed very much from its value $\tilde{w}_{T-1}$ at the start of the interval and so we can replace $\tilde{w}_{t-1}$ by that value to get

$$\tilde{w}_{T+N} - \tilde{w}_T \approx -\mu\Sigma_T^{T+N} x_t x_t^T \tilde{w}_{T-1} + \mu\Sigma_T^{T+N} x_t n_t$$

For fixed T, by A1 and the ergodic theorem, we have $\frac{1}{N}\Sigma_T^{T+N} x_t x_t^T \to R_x$ and $\frac{1}{N}\Sigma_T^{T+N} x_t n_t \to E(x_t n_t) = 0$. So if N is large enough we can approximate the sum by NR_x and the driving term by 0 to get

$$\tilde{w}_{T+N} - \tilde{w}_T \approx -\mu N R_x \tilde{w}_{T-1} + 0$$

Now successive differencing yields the 'averaged system' already found.

It turns out that the averaging approach has a huge advantage over the IH because the argument above can be made rigorous [12]. Further averaging methods can be used to analyse the behaviour of any adpative algorithm.

2.6. What Kind of Convergence?

Athough both heuristic arguments have produced the same result, a careful perusal of the two approaches shows that they are actually delivering different kinds of convergence. The IH is delivering convergence in mean whereas the averaging heuristic is based on a realization-wise analysis and so is delivering some kind of converge with probability 1 (wp1).

In fact because we are dealing with online parameter estimation we only have one realization and so wp1 convergence is what we need to address. Convergence in mean applies to averages over independent repeat simulations and so is of lesser interest. Let us then consider a formal wp1 analysis.

2.7. wp1 Analysis

The first result is a surprise.

Result I. LMS does not converge.

Proof. In order for the sequence $\tilde{w}_t$ to converge it must have at least one equilibrium point i.e. a value where motion ceases. Let w_* be such a point. Then we must have

$$0 = \delta w_* = \mu x_t x_t^T w_* + \mu x_t n_t$$

There are two cases. If $w_* \neq 0$ then cancelling μx_t we see that n_t = a linear combination of the regressors. But this contradicts A1. If $w_* = 0$ then we get $x_t n_t = 0$ which is a contradiction of A1. The result follows.

Where then does this leave the 'convergence' analyses above? What happens is that the LMS trajectory fluctuates in the vicinity of the trajectory of the averaged system [12]. The magnitude of these fluctuations can be captured in a 'hovering theorem' [12] which, for any given $T > 0$, provides a wp1 bound on $\sup_{0<t<\frac{T}{\mu}} \parallel \tilde{w}_t - m_t \parallel$ and is controlled by the size of μ.

Averaging methods, which may be regarded as a generalization of perturbation methods, have a long history (see chapter notes in [12]). Averaging for deterministic systems was first put on a rigorous footing in the 1930s and averaging for stochastic systems emerged in the 1960s in the Russian literature. Significant developments followed in the American literature in the 1970s. Averaging methods were introduced into adaptive control by Ljung (stochastic) [25] and Kokotovoic (deterministic) [26]. Competitors to averaging are weak convergence methods ,[21],[27] and the so-called ODE method [25]. Both weak convergence and the ODE method gain some mathematical advantages by moving to continuous time. But they lose the ability to provide specific stability conditions such as A2 (they can get the lower bound but not the upper bound). Further they cannot deliver crucial results such as the hovering theorems. Weak convergence also requires a sophisticated mathematical development which averaging does not.

3. LMS PERFORMANCE

The issue here is to get more detailed information on the size of the variance of the steady state fluctuations of the weight error

$$P(\mu) = E(\tilde{w}_t \tilde{w}_t^T)$$

as well as the mean squared error (MSE)

$$\mathcal{E}(\mu) = E(e_t^2) = E(y_t - x_t^T \hat{w}_{t-1})^2$$

These classic performance measures were introduced and approximated in the seminal paper [22].

The calculation of $P(\mu)$ and $\mathcal{E}(\mu)$ are very challenging problems which however have been carried out in earlier work.

Result II. Under A1,A2 we have $P(\mu) = P_o \mu + o(\mu)$ where P_o obeys the Lyapunov equation

$$R_x P_o + P_o R_x = F_{xn}(0) = \Sigma_{-\infty}^{\infty} \gamma_k^x \gamma_k^n$$

Proof. This is a special case of results in [20].
Note that the term on the right side is the spectral matrix of $x_t n_t$ at zero frequency.

Result III. Under A1,A2 we have $\mathcal{E}(\mu) = \mu tr(F_{xn}(0)) + o(\mu)$.

Proof. This is a special case of results in [20].

The crucial feature of these performance formulae is that they involve autocovariances of the signals x_t, n_t. However they have the unusual property that if either x_t or n_t is a white noise then they reduce to expressions that depend only on the signal variances i.e. they reduce to the formuae one gets by assuming the signals are both white noise i.e. the 'white noise only' formulae. This helps explain the 'urban myth' that the 'white noise only' formulae can be used as surrogates for the autocorrelated formulae above.

4. NETWORK LMS

While Electrical Engineering was an early entrant into the study of networks e.g. telephone exchanges in the 1910s, network science has a complex history spread across a number of disciplines [28]. The interest in adaptive signal processing on networks emerged around 2005 in conjunction with a related interest in the control community; both driven partly by the 'consensus' problem. Distributed versions of LMS date from this period e.g. [29]. A survey and extensive reference list can be found in [10].

In the static network setting we have one regression equation at each node k of N nodes but with a common weight vector

$$y_{k,t} = x_{k,t}^T w_e + n_{k,t}, k == 1, \cdots, N$$

There are a number of generalizations of LMS to the network case [10]. But they each use information from network neighbours to improve the quality of the updates. But what was missed until our recent work [30],[31] is that this sharing induces a mixed time scale structure in these network versions of LMS.

The general diffusion LMS algorithm has the form [30]

$$\hat{w}_t = \mathcal{A}_2^T (\mathcal{A}_o^T - \mu \mathcal{R}_t) \mathcal{A}_1^T \hat{w}_{t-1} + \mu \mathcal{A}_1^T \sigma_t^{xy}$$

where $\tilde{w}_t$ has weight entries from each node; $\mathcal{A}_i = A_i \otimes I_d$ where A_i are $N \times N$ adjacency matrices; $\mathcal{R}_t$ has blockdiago-

nal entries $x_{k,t}x_{k,t}^T$; σ_t^{xy} has entries $x_{k,t}y_{k,t}$. The corresponding error system is [30]

$$\tilde{w}_t = \mathcal{A}_2^T\mathcal{A}_o^T\mathcal{A}_1^T\tilde{w}_{t-1} - \mu\mathcal{A}_2^T\mathcal{R}_t\mathcal{A}_1^T\tilde{w}_{t-1} + \mu\mathcal{A}_2^T\nu_t \quad (4.1)$$

where ν_t has entries $x_{k,t}n_{k,t}$. This has some similarities with the single node case but also significant differences. In the single node case $\delta\tilde{w}_t = \tilde{w}_t - \tilde{w}_{t-1} = O(\mu)$. That is not true here. Instead due to the eigen properties of the adajacency matrices (see below) there is a two time scale structure whereby a special linear combination of $\delta\tilde{w}_t$ is $O(1)$ while all others are $O(\mu)$. This is crucial to the analysis of the error system and was revealed for the first time in [30].

The error system in (4.1) is an example of a single time scale system where all states have a rate of change (i.e. $\delta\tilde{w}_t$) which is $O(\mu)$. A two time scale system has the following structure [12],[31]

$$\begin{aligned} \delta z_t &= \mu f(t, z_{t-1}, y_{t-1}) \\ y_t &= Sy_{t-1} + \mu g(t, z_{t-1}, y_{t-1}) \end{aligned}$$

where S is a stability matrix. So some states, the 'slow' states, have a rate of change $O(\mu)$ but the other states , the 'fast' states, have a rate of change $O(1)$. Averaging analysis can handle these kinds of systems.

Using averaging analysis (for two time scale systems) the following results were obtained in [30],[31] for the first time. To state them we need some assumptions.

N1 $M = A_1A_0A_2$ is primitive.
A sufficient condition for this is that the network graph is strongly connected with at least one self loop.

N2 $x_{k,t}, n_{k,t}$ are zero mean strictly stationary.
The nodal variances are $R_{x,k}, \sigma_{n,k}^2$. The nodal autocovariances are $\gamma_r^{x_k}, \gamma_r^{n_k}$.

A crucial consequence of N1 is that M is then left stochastic i.e. $1^TM = 1^T$ and has a right eigenvector with unit eigenvalue, called the Perron eigenvector which is a mass function i.e. its entries are ≥ 0 and sum to 1. The left stochastic property of M induces the two time scale structure.

Introduce $A = \Sigma_1^N\alpha_kR_{x,k}$ where α_k are coefficients computed from the Perron eigenvector of M. Also denote the spectral radius of A by $\rho(A)$.

Result IV. [30] Under N1,N2 the averaged system associated with the error system (4.1) is exponentially stable if $0 < \mu\rho(A) < 2$.

The associated hovering thoerem is given in [30].

Result V. [31] Under N1,N2 the network weight error variance matrix $P(\mu) = P_o\mu + o(\mu)$ where P_o satisfies a Lyapunov equation

$$\begin{aligned} AP + PA &= F_{xn}(0) = \Sigma_1^N\alpha_k^2F_{xnk}(0)) \\ F_{xnk}(0) &= \Sigma_{-\infty}^{\infty}\gamma_r^{n_k}\gamma_r^{x_k} \end{aligned}$$

Result VI. [31] Under N1,N2 the network mse (which totals mse over all nodes) is given by $\mathcal{E}(\mu) = \mu tr(F_{xn}(0) + o(\mu)$.

These results are remarkable extensions of those in section 2. Again we need to correct the 'urban myth' that white noise only performance formulae are 'good' surrogates for the correct 'autocorrelated' formulae above. The requirements for this are much more stringent than in the single node case. It is necessary that for every node either $x_{k,t}$ or $n_{k,t}$ (or both) are white noises.

5. CONCLUSIONS AND THE FUTURE

In this paper we began by sketching first and second order analysis of the single node LMS algorithm. We showed that the success of the independence heuristic in first order analysis can be largely explained by an averaging analysis. Under stationarity assumptions we then recalled formulae for weight error variance and mean squared error. It turns out these formulae reduce to the white noise only formulae when either the regressors or the measurement noise are white. This likely explains the 'urban myth' that white noise only formulae are adequate surrogates even when the regressors and measurement noise are autocorrelated.

We then sketched extensions to the network case. We pointed out a fundamental two-time scale property that is as yet not widely understood. We then sketched extensions of the single node performance analyses based on very recent results.

Adaptive signal processing and LMS in particular have a very rich history of successful applications across a range of disciplines. However the upsurge in interest in online 'learning' from the machine learning community has raised significant problems. So far the machine learning community shows little awareness of the rich six decade literature on adaptive signal processing and adaptive control and is spending considerable energy working on problems that are already solved. How then to address this? Currently just a handful of signal processing researchers present their work at machine learning conferences. That needs to change.

Because adaptive algorithms have short memory they can extract patterns effectively from data streams while needing little storage. In a 'big data'[2] world, this gives adaptive algorithms a powerful comparative advantage even in offline or batch settings. For this simple reason adaptive algorithms (and the simplest of them all, LMS) have a guaranteed future. The internet of things won't be possible without them.

[2]You have big data if it can't be fitted on one computer

6. REFERENCES

[1] HP. Whitaker, "An adaptive system for control of the dynamics performance of aircraft and spacecraft," *Inst. Aeronautical Sciences*, vol. Paper 59-100, pp. –, 1959.

[2] B. Widrow and ME. Hoff, "Adaptive switching circuits," in *IRE Wescon Convention Record Part IV*. IRE, 1960, pp. 96–104.

[3] B. Widrow, "Thinking about thinking: The discovery of the lms algorithm," *IEEE Signal Proc. Mag.*, pp. 100–106, 2005.

[4] A. Barto, R. Sutton, and C. Anderson, "Neuronlike adaptive elements that can solve difficult learning control problems," *EEE Trans. Syst. Man Cybern*, vol. 13, pp. 834–846, 1983.

[5] R. Sutton and A. Barto, *Introduction to reinforcement learning*, MIT Press, Cambridge, MA, USA, 1998.

[6] Leon Bottou, "Stochastic learning," in *Machine Learning 2003, LNAI 3176, O. Bousquet et.al. (eds.)*, 2004, pp. 146–168.

[7] JI. Nagumo and A. Noda, "A learning method for system identification," *IEEE Trans. Autom. Contr.*, vol. 12, pp. 282–287, 1967.

[8] DTM. Slock, "On the convergence behaviour of the lms and normalized lms algorithms," *IEEE trans. Sig. Proc.*, vol. 41, pp. 2811–2825, 1993.

[9] DP. Mandic and SL. Goh, *Complex Valued Nonlinear Adaptive Filters: Noncircularity, Widely Linear and Neural Models*, J. Wiley, New York, 2009.

[10] A.H Sayed, "Adaptive networks," *Proc. IEEE*, vol. 102, pp. 460–497, 2014.

[11] O. Macchi, *Adaptive Signal Processing, The Least Mean Squares Approach*, J. Wiley, New York, 1995.

[12] V. Solo and X. Kong, *Adaptive Signal Processing Algorithms: Stability and Performance*, Prentice Hall, New Yrok, 1995.

[13] A.H. Sayed, *Adaptive Filters*, J. Wiley, New York, 2008.

[14] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, New York, 2002.

[15] B. Widrow and SD. Stearns, *Adaptive Signal Processing*, Prentice Hall, Englewood Cliffs, NJ, 1985.

[16] ID. Schizas, G. Mateos, and GB. Giannakis, "Distributed lms for consensus-based in-network adaptive processing," *IEEE Trans. Sig. Proc.*, vol. 57, pp. 2365–2382, 2009.

[17] S. Kar and JMF. Moura, "Consensus + innovations distributed inference over networks: Cooperation and sensing in networked systems," *IEEE Signal Process. Mag*, vol. 30, pp. 99–109, 2013.

[18] SS. Stankovic, MS. Stankovic, and DM. Stipanovic, "Decentralized parameter estimation by consensus based stochastic approximation," *IEEE Trans. Autom. Control*, vol. 56, pp. 531–543, 2011.

[19] S. Kar and J.M. F. Moura, "Convergence rate analysis of distributed gossip (linear parameter) estimation: Fundamental limits and tradeoffs," *IEEE J. Sel. Topics in Sig. Proc.*, vol. 5, pp. 674–690, 2011.

[20] V. Solo, "The error variance of LMS with time varying weights," *IEEE Trans. Acoustics Speech Sig. Proc.*, vol. 40, pp. 803–813, 1992.

[21] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, Berlin, 1990.

[22] B. Widrow, JM McCool, MG. Larimore, and CR Johnson Jr., "Stationary and nonstationary learning characteristics of the lms adaptive filter," *Proc IEEE*, vol. 64, pp. 1151–1162, 1976.

[23] H.J. Butterweck, "The independence assumption: A dispensable tool in adaptive filter theory," *Signal Processing*, vol. 57, pp. 305–310, 1997.

[24] M. Rupp and H-J. Butterweck, "Overcoming the independence assumption in lms filtering," in *Proc. 37th Asilomar Conf. Circ. Sys. Comp.* IEEE, 2003, p. 5pp.

[25] L. Ljung, "Analysis of recursive stochastic algorithms," *IEEE Trans. Autom. Contr.*, vol. 22, pp. 551–575, 1977.

[26] P. Ioannou and P. Kokotovic, *Adaptive Systems with reduced Models*, Springer, Berlin, 1983.

[27] H.J. Kushner and G.C. Yin, *Stochastic Approximation Algorithms and Applications*, Springer, New York, 1997.

[28] M. Newman, *Networks: An Introduction, 2nd. edn.*, Oxford Univ. Press, Oxford, UK, 2010.

[29] CG. Lopes and AH. Sayed, "Distributed processing over adaptive networks," in *Proc. Adaptive Sensor Array Proc. Workshop*. MIT Lincoln Laboratory, Cambridge, MA, USA,, 2006, pp. 1–5.

[30] M. Piggott and V. Solo, "Diffusion lms with correlated regressors i: Realization-wise stability," *IEEE Trans. Sig. Proc.*, vol. 64, pp. 5473–5484, 2016.

[31] M. Piggott and V. Solo, "Diffusion lms with correlated regressors ii: Performance," *IEEE Trans. Sig. Proc.*, vol. 65, pp. 3934–3947, 2017.