Asymptotic Analysis of the LMS Algorithm with Momentum

Citation for published version:
Gerencser, L, Csaji, BC & Sabanis, S 2019, 'Asymptotic Analysis of the LMS Algorithm with Momentum' Paper presented at 2018 IEEE Conference on Decision and Control (CDC), 17/12/18 - 19/12/18, pp. 3062-3067. DOI: 10.1109/CDC.2018.8619576

Digital Object Identifier (DOI):
10.1109/CDC.2018.8619576

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Asymptotic Analysis of the LMS Algorithm with Momentum

László Gerencsér, *Member, IEEE*, Balázs Csanád Csáji, *Member, IEEE*, Sotirios Sabanis

***Abstract*—A widely studied filtering algorithm in signal processing is the least mean square (LMS) method, due to B. Widrow and T. Hoff, 1960. A popular extension of the LMS algorithm, which is also important in deep learning, is the LMS method with momentum, originated by S. Roy and J.J. Shynk back in 1988. This is a fixed gain (or constant step-size) version of the LMS method modified by an additional momentum term that is proportional to the last correction term. Recently, a certain equivalence of the two methods has been rigorously established by K. Yuan, B. Ying and A.H. Sayed, assuming martingale difference gradient noise. The purpose of this paper is to present the outline of a significantly simpler and more transparent asymptotic analysis of the LMS algorithm with momentum under the assumption of stationary, ergodic and mixing signals.**

***Index Terms*—least mean square methods, statistical analysis, recursive estimation, gradient methods, machine learning**

I. Introduction

A classical, widely studied recursive estimation method for determining the mean-square optimal linear filter is the *least mean square* (LMS) method, due to B. Widrow and T. Hoff [1], devised for pattern recognition problems. The algorithm can be seen as a stochastic gradient (SG) method with fixed gain. The fine structure of the estimation error process for small adaptation gain has been studied in a number of works.

A general class of fixed gain recursive estimation methods, under mild ergodicity assumptions, with applications to variants of the LMS algorithm, including the sign-error and sign-sign algorithms, was studied by J. A. Bucklew, T. G. Kurtz and W. A. Sethares [2, Theorem 2], leading to a result establishing the *weak convergence* of the (piecewise constant extension of the) rescaled estimation error process to the solution of a linear stochastic differential equation on the semi-infinite interval $[0,\infty)$ with a concise and transparent proof.

An alternative general class of fixed gain recursive estimation methods defined in a Markovian framework was studied by A. Benveniste, M. Metivier and P. Priouret, see [3]. They formulate a similar weak convergence result for fixed finite time intervals, see Theorem 7 of Part II, Section 4.4.1 [3]. The advantage of their approach is that their framework allows recursive algorithms with feedback effects, which is typical, e.g., for recursive estimation of linear stochastic systems.

A refined characterization of the (piecewise constant extesion of the) of LMS in a different direction was given by A. Heunis and L.A. Joslin [4], providing a limit theorem in the form of a functional law of the iterated logarithm.

Higher order moments of the estimation error of LMS were estimated in [5] for bounded signals satisfying a certain mixing condition, showing that the L_p-norms of these errors are proportional to the square root of the gain. A similar result was established under much weaker conditions for general stochastic approximation (SA) methods, allowing discontinuous correction terms, satisfying a relaxed mixing condition by H. N. Chau, Ch. Kumar, M. Rásonyi and S. Sabanis [6].

A common experimental finding with stochastic gradient methods is that they tend to be slow in the initial phase, especially if the number of parameters is huge, as is the case with problems in deep learning. A currently widely applied modification of standard stochastic gradient methods, resulting in the acceleration of the early stages of the algorithms, is the use of a *momentum* term, a device that has proven to be succesful in determimistic optimization, see Polyak [7]. The original method is also known as the *heavy-ball* method referring to the fact that the dynamics of the minimization method can be described as the motion of a heavy-ball along a hilly terrain trying to find its way to the absolute minimum by trying to avoid undesirable local minima.

Theoretical justification of the superiority of SG methods with momentum, in the early stages, are not available in the literature, however the "steady-state" behavior of the estimator process generated by SG methods with momentum have been known to be inferior to that of the standard SG methods since the works of Polyak [8]. In a paper of 2016 K. Yuan, B. Ying and A.H. Sayed established a remarkable equivalence of SG methods with momentum to the standard SG methods with a rescaled gain [9]. Their result is obtained among others under the condition that what is called the gradient noise is a martingale difference. In case of LMS, paper [9] assumes an *independent* sequence of observations to ensure this.

The objective of the present paper is to significantly relax the assumptions on the "gradient noise", and to provide an accurate characterization of the relationship between the two estimator processes in an asymptotic sense, relying on weak convergence results developed in [2], leading to a transparent proof. In particular, we show that the asymptotic distribution of

This research was partially supported by the Royal Society International Exchange Program, UK, Grant no. IE150128 and the National Research, Development and Innovation Office (NKFIH), Hungary, Grant no. 2018-1.2.1-NKP-00008. B. Cs. Csáji was supported by NKFIH, Grant no. KH_17 125698, and the János Bolyai Research Fellowship, Grant no. BO/00217/16/6. The third author is a Turing fellow and was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1 (Turing award number TU/B/000026).

L. Gerencsér and B. Cs. Csáji are with MTA SZTAKI: The Institute for Computer Science and Control of the Hungarian Academy of Sciences, Kende utca 13–17, Budapest, Hungary, 1111; email: {laszlo.gerencser, balazs.csaji}@sztaki.mta.hu

S. Sabanis is with School of Mathematics, University of Edinburgh, UK, and Alan Turing Institute, London, UK; email: s.sabanis@ed.ac.uk

the two estimator processes are *identical modulo scaling*, and the effect of the various scaling factors is precisely explored.

For the sake of simplicity, our results will be presented for the LMS method, but they can be adapted directly to general recursive estimation methods discussed in [2].

II. Preliminaries

Let $(x_n, y_n), \infty < n < +\infty$ be a jointly wide sense stationary stochastic process, where (x_n) is $\mathbb{R}^p$-valued and (y_n) is real-valued. The best linear mean-square estimator of y_n in terms of the instantenous signal x_n is defined as the solution of the following minimization problem

$$\min_{\theta} \mathbb{E}\left[(y_n - x_n^{\mathrm{T}}\theta)^2\right], \tag{1}$$

the solution of which will be denoted by θ^*. Thus, θ^* is the solution of the linear algebraic equation

$$\mathbb{E}\left[x_n x_n^{\mathrm{T}}\right]\ \theta = R^*\theta = \mathbb{E}\left[x_n y_n\right] \quad \text{with} \quad R^* := \mathbb{E}\left[x_n x_n^{\mathrm{T}}\right]. \tag{2}$$

[C0] We assume that matrix R^* is non-singular, so that θ^* is uniquely defined as $\theta^* = (R^*)^{-1}\ \mathbb{E}\left[x_0 y_0\right]$.

Then, the LMS method is described by the algorithm

$$\theta_{n+1} = \theta_n + \mu\, x_{n+1}(y_{n+1} - x_{n+1}^{\mathrm{T}}\theta_n), \quad n \geq 0, \tag{3}$$

with some non-random initial condition θ_0. Here $\mu > 0$ is a fixed gain or constant step-size, also called learning rate. Introducing an artificial observation error v_n, and the (filter coefficient) estimation error Δ_n as

$$v_n := y_n - x_n^{\mathrm{T}}\theta^* \quad \text{and} \quad \Delta_n := \theta_n - \theta^*, \tag{4}$$

the estimation error process (Δ_n) follows the dynamics:

$$\Delta_{n+1} = \Delta_n - \mu\, x_{n+1}x_{n+1}^{\mathrm{T}}\Delta_n + \mu\, x_{n+1}v_{n+1}, \quad n \geq 0, \tag{5}$$

with $\Delta_0 = \theta_0 - \theta^*$. Note that $\mathbb{E}\left[x_n v_n\right] = 0$ for any $n \geq 0$, i.e., the observation error v_n is orthogonal to data x_n for any $n \geq 0$.

Henceforth, we shall strengthen our initial condition by assuming the following:

[C1] The joint process $(x_n, y_n), \infty < n < +\infty$ is a *strictly stationary* and *ergodic* stochastic process.

The above algorithm is a special case of the more general *stochastic approximation* (SA) method defined by

$$\theta_{n+1} = \theta_n + \mu H(\theta_n, X_{n+1}), \quad n \geq 0, \tag{6}$$

with some non-random initial condition θ_0, where (X_n) is a strictly stationary, ergodic stochastic process and $H(\theta, X)$ is integrable w.r.t. the law of X_0. In the case of the LMS method,

$$H(\theta, X_n) = x_n(y_n - x_n^{\mathrm{T}}\theta) =: H_n(\theta), \tag{7}$$

with $X_n = (x_n^{\mathrm{T}},\ y_n)^{\mathrm{T}}$.

A standard tool for the analysis of stochastic approximation methods is the associated ODE, two early, scholarly references for which are [2], [3]. The ODE in our case takes the form, with the notation $h(\theta) := \mathbb{E}\left[x_{n+1}(y_{n+1} - x_{n+1}^{\mathrm{T}}\theta)\right]$,

$$\frac{d}{dt}\bar{\theta}_t = h(\bar{\theta}(t)) = b - R^*\bar{\theta}_t, \qquad t \geq 0, \tag{8}$$

where $b := \mathbb{E}\left[x_n y_n\right]$. For the sake of convenience in formulating the relevant results, we set $\bar{\theta}_0 = \theta_0$.

One of the benefits of the ODE method is that it provides quantified bounds or even characterization of the estimation error. To describe the magnitude of the estimator error process (θ_n) let us first consider its piecewise constant extension defined by $\theta_t^c = \theta_n$ for $n \leq t < n+1$. Equivalently, we may write $\theta_t^c = \theta_{[t]}$, where $[t]$ denotes the integer part of t. Then, an early result along the lines of applying the ODE method is that, assuming bounded signals, satisfying certain mixing conditions, we have for any fixed $T > 0$, and k being a non-negative integer, that the following holds:

$$\sup_{kT \leq t \leq (k+1)T} |\theta_t^c - \bar{\theta}_t| = O_M((\mu T)^{1/2}), \tag{9}$$

assuming the initial condition $\bar{\theta}_{kT} = \theta_{kT}^c$, see [5].

The assumption on the boundedness of the signals would ensure that the estimator process itself stay bounded w.p.1, and thus a common problem in recursive estimation, namely the need to enforce the boundedness of the estimator process, does not arise. In the general case of possibly unbounded signals we resort to a standard device, which is the use of truncation. This is in fact applied in our prime reference, [2]. Thus the original LMS algorithm is modified by taking a truncation domain D, where D is the interior of a compact set, and we stop the estimator process (θ_n) if it leaves D. In technical terms,

$$\tau := \inf\{t : \theta_t^c \notin D\}. \tag{10}$$

[C2] We assume that the truncation domain is such that the solution of the ODE (8), with $\bar{\theta}_0 = \theta_0$, does not leave D.

To describe the finer structure of the estimator error process (θ_n) let us define the error processes

$$\tilde{\theta}_n := (\theta_n - \bar{\theta}_n), \tag{11}$$

for $n \geq 0$, and similarly, set $\tilde{\theta}_t^c := (\theta_t^c - \bar{\theta}_t)$. The key object of study for the weak convergence theory of the LMS, and in fact for more general class of SA processes is the normalized and time-scaled process $(V_t(\mu))$ defined by

$$V_t(\mu) := \mu^{-1/2}\tilde{\theta}_{[(t\wedge\tau)/\mu]} = \mu^{-1/2}\tilde{\theta}^c_{(t\wedge\tau)/\mu}. \tag{12}$$

In describing the weak limit of the stopped SA process a crucial role is played by the asymptotic covariance matrices of the empirical means of the centered correction terms $(H_n(\theta) - h(\theta)$, which can be expressed, under reasonable conditions, as

$$S(\theta) := \sum_{k=-\infty}^{+\infty} \mathbb{E}\left[(H_k(\theta) - h(\theta)(H_0(\theta) - h(\theta))^{\mathrm{T}}\right], \tag{13}$$

which series converges, e.g., under various mixing conditions. This is ensured by [C3] bellow (cf. [10, Theorem 19.1]).

For $\theta = \theta^*$, in the case of the LMS method, we get

$$S := S(\theta^*) = \sum_{k=-\infty}^{+\infty} \mathbb{E}\left[x_k w_k w_0 x_0^{\mathrm{T}}\right]. \tag{14}$$

[C3] We assume that the process defined by

$$L_t(\mu) = \sum_{n=0}^{[t/\mu]-1} \left(H_n(\bar{\theta}_{\mu n}) - h(\bar{\theta}_{\mu n})\right)\sqrt{\mu}, \tag{15}$$

converges weakly, as $\mu \to 0$, to a time-inhomogeneous zero-mean Brownian motion (L_t) with local covariances $(S(\bar{\theta}_t))$.

We conjecture that for the verification of the above condition, it is sufficient to check that for any *fixed* $\bar{\theta}$ the process

$$L_t(\mu) = \sum_{n=0}^{[t/\mu]-1} \left(H_n(\bar{\theta}) - h(\bar{\theta})\right)\sqrt{\mu}, \tag{16}$$

converges weakly, as $\mu \to 0$, to a time-homogeneous zero-mean Brownian motion $L_t(\bar{\theta})$ with covariance matrix $S(\bar{\theta})$.

We note that there is a wide range of results ensuring a Donsker-type theorem as stated above, including stochastic processes with various mixing conditions, or martingales, see [10]. A prominent example is given in [10, Theorem 19.1].

We can conclude, using Theorem 2 of [2], that the following weak convergence result holds:

Theorem 1. *Under conditions C0, C1, C2 and C3, process $(V_t(\mu))$ converges weakly, as $\mu \to 0$, to a process (Z_t) satisfying the linear stochastic differential equation (SDE),*

$$dZ_t = -R^* Z_t dt + S^{1/2}(\bar{\theta}_t) dW_t, \tag{17}$$

for $t \geq 0$, with initial condition $Z_0 = 0$, where (W_t) is a standard Brownian motion in $\mathbb{R}^p$.

Let us denote the asymptotic covariance matrix of process (Z_t) by P_0, It is known that matrix P_0 is the unique solution of the algebraic Lyapunov equation

$$-R^* P_0 - P_0 R^* + S = 0, \tag{18}$$

where matrix $S := S(\theta^*)$ is given by equation (14). Although the weak convergence of $(V_t(\mu))$ does not imply directly that the distribution of $\mu^{-1/2}\tilde{\theta}_{[(t\wedge\tau)/\mu]}$ converges weakly to $\mathcal{N}(0, P_0)$, when $\mu \to 0$ and $t \to \infty$, the corresponding claim for general SA processes in a Markovian framework has been established in [3, Part II, Chapter 4, Theorem 15]. Surprisingly, the covariance matrix P_0 will pop up also in the asymptotic analysis of the LMS method with momentum.

III. LMS with Momentum

A widely studied modification of the fixed gain LMS method is the LMS method with momentum, using a device that has proven to be succesful in determimistic optimization [7]. The original method is also known as the heavy-ball method, since the dynamics of the minimization method can be described as the motion of a heavy-ball along a hilly terrain:

$$\theta_{n+1} = \theta_n + \mu\, x_{n+1}(y_{n+1} - x_{n+1}^{\mathrm{T}}\theta_n) + \gamma(\theta_n - \theta_{n-1}), \tag{19}$$

where $0 < \gamma < 1$ and $n \geq 0$, with some non-random initial condition θ_0, and $\theta_{-1} = \theta_0$. The momentum term intruduces some kind of memory into the dynamics, and it is hoped that it has a smoothing effect on the estimator process. Note that the LMS with momentum is driven by a second order dynamics.

The parameter-error process, (Δ_n), is then defined by the following second order dynamics

$$\Delta_{n+1} = \Delta_n - \mu x_{n+1} x_{n+1}^{\mathrm{T}}\Delta_n + \gamma(\Delta_n - \Delta_{n-1}) + \mu x_{n+1} v_{n+1}, \tag{20}$$

for $n \geq 0$, with $\Delta_{-1} = \Delta_0$.

In order to analyze the behaviour of (Δ_n) we follow standard recipes of the theory of linear systems and introduce the state-vector having twice the dimension of that of Δ_n,

$$U_n := \begin{bmatrix} \Delta_n \\ \Delta_{n-1} \end{bmatrix}. \tag{21}$$

Then, the state-space dynamics will become:

$$U_{n+1} = U_n + A_{n+1} U_n + \mu W_{n+1}, \tag{22}$$

where

$$A_{n+1} = \begin{bmatrix} \gamma I - \mu \cdot x_{n+1} x_{n+1}^{\mathrm{T}} & -\gamma I \\ I & -I \end{bmatrix}, \tag{23}$$

$$W_{n+1} = \begin{bmatrix} x_{n+1} v_{n+1} \\ 0 \end{bmatrix}. \tag{24}$$

It is not obvious if and how the above dynamics can be interpreted as a SA method. Note that for small μ and γ close to 1 the matrix A_{n+1} is close to the singular matrix

$$T_1^+ = \begin{bmatrix} I & -I \\ I & -I \end{bmatrix}. \tag{25}$$

for which we have $(T_1^+)^2 = 0$.

Linear transformation of the state-space. In order to capture the effect and the interaction of the small parameters μ and $1-\gamma$ on the dynamics (22), following [9], we introduce a linear state-space transformation $\bar{U} := TU$ with

$$T := T(\gamma) = \frac{1}{1-\gamma}\begin{bmatrix} I & -\gamma I \\ I & -I \end{bmatrix}, \tag{26}$$

$$T^{-1} := T^{-1}(\gamma) = \begin{bmatrix} I & -\gamma I \\ I & -I \end{bmatrix}. \tag{27}$$

We decompose $\bar{A}_n$ into two parts $\bar{A}_n = \bar{A}^{(1)} + \bar{A}_n^{(2)}$, where

$$A^{(1)} = \begin{bmatrix} \gamma I & -\gamma I \\ I & -I \end{bmatrix} \qquad \text{and} \qquad A_n^{(2)} = \begin{bmatrix} -\mu x_n x_n^{\mathrm{T}} & 0 \\ 0 & 0 \end{bmatrix}.$$

Then, multiplying (22) by T from the left, and substituting $U = T^{-1}\bar{U}$ we get that the new state-transition matrix $\bar{A}_n$ can be written as the sum $\bar{A}_n = \bar{A}^{(1)} + \bar{A}_n^{(2)}$, where

$$\begin{aligned} \bar{A}^{(1)} = TA^{(1)}T^{-1} &= \frac{1}{1-\gamma}\begin{bmatrix} I & -\gamma I \\ I & -I \end{bmatrix}\begin{bmatrix} \gamma I & -\gamma I \\ I & -I \end{bmatrix} T^{-1} \\ &= \frac{1}{1-\gamma}\begin{bmatrix} 0 & 0 \\ (\gamma-1)I & (-\gamma+1)I \end{bmatrix}\begin{bmatrix} I & -\gamma I \\ I & -I \end{bmatrix} \\ &= (1-\gamma)\begin{bmatrix} 0 & 0 \\ 0 & -I \end{bmatrix}, \end{aligned} \tag{28}$$

and for $\bar{A}_n^{(2)} = TA_n^{(2)}T^{-1}$ we have

$$
\begin{aligned}
\bar{A}_n^{(2)} &= \frac{1}{1-\gamma}\begin{bmatrix} I & -\gamma I \\ I & -I \end{bmatrix}\begin{bmatrix} -\mu x_n x_n^T & 0 \\ 0 & 0 \end{bmatrix} T^{-1} \\
&= \frac{1}{1-\gamma}\begin{bmatrix} -\mu x_n x_n^T & 0 \\ -\mu x_n x_n^T & 0 \end{bmatrix}\begin{bmatrix} I & -\gamma I \\ I & -I \end{bmatrix} \\
&= \frac{1}{1-\gamma}\begin{bmatrix} -\mu x_n x_n^T & \mu\gamma x_n x_n^T \\ -\mu x_n x_n^T & \mu\gamma x_n x_n^T \end{bmatrix} \\
&= \frac{\mu}{1-\gamma}\begin{bmatrix} -1 & \gamma \\ -1 & \gamma \end{bmatrix} \otimes x_n x_n^T. \qquad (29)
\end{aligned}
$$

After multiplication by T, the stochastic input becomes

$$
\begin{aligned}
\bar{W}_n &= T\mu \begin{bmatrix} x_n v_n \\ 0 \end{bmatrix} = \frac{1}{1-\gamma}\begin{bmatrix} I & -\gamma I \\ I & -I \end{bmatrix} \cdot \mu \begin{bmatrix} x_n v_n \\ 0 \end{bmatrix} \\
&= \frac{\mu}{1-\gamma}\begin{bmatrix} x_n v_n \\ x_n v_n \end{bmatrix}. \qquad (30)
\end{aligned}
$$

The transformed dynamics. A shorthand description for the dynamics of the transformed state process is

$$\bar{U}_{n+1} = \bar{U}_n + \bar{A}_{n+1}\bar{U}_n + \bar{W}_{n+1}. \qquad (31)$$

For the initial condition we have

$$
\begin{aligned}
\bar{U}_0 = T\bar{\Delta}_0 &= \frac{1}{1-\gamma}\begin{bmatrix} I & -\gamma I \\ I & -I \end{bmatrix}\begin{bmatrix} \Delta_0 \\ \Delta_0 \end{bmatrix} \\
&= \frac{1}{1-\gamma}\begin{bmatrix} (1-\gamma)\Delta_0 \\ 0 \end{bmatrix} = \begin{bmatrix} \Delta_0 \\ 0 \end{bmatrix}, \qquad (32)
\end{aligned}
$$

thus the initial condition is independent of μ and γ !

The point of this transformation is to get a fixed gain SA procedure for $\bar{U}_n$ in its standard form. This is achived by synchronizing the parameters μ and γ. Note that $\bar{A}^{(1)}$ is scaled by $1-\gamma$, while $\bar{A}_n^{(2)}$ and the input noise is scaled by $\mu/(1-\gamma)$. Therefore, a natural way of synchronizing them is to set

$$\frac{\mu}{1-\gamma} = c(1-\gamma) \quad \text{leading to} \quad \mu = c(1-\gamma)^2. \qquad (33)$$

with some fixed constant $c > 0$. Thus (31) can be rewritten as a SA recursion with the fixed gain $\lambda := 1-\gamma$ as follows:

$$\bar{U}_{n+1} = \bar{U}_n + \lambda\,\bar{B}_{n+1}\bar{U}_n + \lambda^2\bar{D}_{n+1}\bar{U}_n + \lambda\,\bar{W}_{n+1}, \qquad (34)$$

for $n \geq 0$, where

$$\bar{B}_n := \begin{bmatrix} 0 & 0 \\ 0 & -I \end{bmatrix} + c\begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix} \otimes x_n x_n^{\mathrm{T}}, \qquad (35)$$

$$\bar{D}_n := c\begin{bmatrix} 0 & -1 \\ 0 & -1 \end{bmatrix} \otimes x_n x_n^{\mathrm{T}}, \qquad (36)$$

$$\bar{W}_n = c\begin{bmatrix} x_n v_n \\ x_n v_n \end{bmatrix}. \qquad (37)$$

Let us approximate (34) by a standard SA recursion where the term with step-size λ^2 has been removed, that is

$$\bar{U}^*_{n+1} = \bar{U}^*_n + \lambda\,\bar{B}_{n+1}\bar{U}^*_n + \lambda\,\bar{W}_{n+1}, \quad \text{with} \quad \bar{U}^*_0 = \bar{U}_0. \qquad (38)$$

Using the linearity of the dynamics and under some technical conditions it can be shown for the difference process,

$$\Delta\bar{U}_n := \bar{U}_n - \bar{U}^*_n, \qquad (39)$$

that $\|\Delta\bar{U}_n\| \leq C_n\lambda^2$, where (C_n) is a strictly stationary process.

The associated ODE. Let us define the random field $\mathbb{R}^{2p} \to \mathbb{R}^{2p}$, and introduce the notations

$$\bar{H}_n(\bar{U}) := (\bar{B}_n + \lambda\bar{D}_n)\bar{U} + \bar{W}_n \qquad (40)$$

$$h(\bar{U}) := \mathbb{E}[\bar{H}_n(\bar{U})] = \bar{B}_\lambda\,\bar{U}, \qquad (41)$$

where

$$\bar{B}_\lambda := \mathbb{E}[\bar{B}_n + \lambda\bar{D}_n] = \begin{bmatrix} 0 & 0 \\ 0 & -I \end{bmatrix} + c\begin{bmatrix} -1 & 1-\lambda \\ -1 & 1-\lambda \end{bmatrix} \otimes R^*. \qquad (42)$$

Then, the associated ODE takes the form

$$\frac{d}{dt}\bar{\bar{U}}_t = \bar{h}(\bar{\bar{U}}_t) = \bar{B}_\lambda\bar{\bar{U}}_t, \qquad t \geq 0. \qquad (43)$$

For the sake of convenience, we set $\bar{\bar{U}}_0 = \bar{U}_0$. The solution for the limit when $\lambda \downarrow 0$, corresponding to (38), is denoted by $\bar{\bar{U}}^*_t$.

Lemma 1. *If λ is sufficiently small, then $\bar{B}_\lambda$ is stable.*

The proof of Lemma 1 can be found in Appendix A. It is straightforward to show that

$$\|\,\bar{\bar{U}}_t - \bar{\bar{U}}^*_t\,\| \leq \bar{\bar{c}}\lambda, \qquad (44)$$

for all $t \geq 0$, where $\bar{\bar{c}}$ is a deterministic constant.

As in the plain LMS case, the assumption on the boundedness of the signals x_n, v_n would ensure that the estimator process itself stay bounded w.p.1. In the general case of possibly unbounded signals we resort to a (virtual) truncation in order to analyze $\bar{U}_n$. Thus transformed estimator process is modified by taking a truncation domain $\bar{D}$, where $\bar{D}$ is the interior of a compact set, such that $\bar{U}^* := 0 \in \bar{D}$, and we stop the process $(\bar{U}_n)$ if it leaves $\bar{D}$.

[C2'] We assume that the truncation domain is such that the solution of the ODE (43), with $\bar{\bar{U}}_0 = \bar{U}_0$, does not leave $\bar{D}$.

We set

$$\bar{\tau} := \inf\{\, n : \bar{U}_n \notin \bar{D}\,\}. \qquad (45)$$

Let us define the error process, for $n \geq 0$, as

$$\tilde{\bar{U}}_n := (\bar{U}_n - \bar{\bar{U}}_n). \qquad (46)$$

and define the normalized and time-scaled error process as

$$\bar{V}_t(\lambda) := \lambda^{-1/2}\tilde{\bar{U}}_{[(t\wedge\bar{\tau})/\lambda]}. \qquad (47)$$

Analogously for the process $(\bar{U}^*_n)$ we take a truncation domain $\bar{D}^*$ such that $\bar{D} \subseteq \mathrm{int}(\bar{D}^*)$ and define $\bar{\tau}^*$ as in (45). Repeating the above procedure we get

$$\bar{V}^*_t(\lambda) := \lambda^{-1/2}\tilde{\bar{U}}^*_{[(t\wedge\bar{\tau}^*)/\lambda]}. \qquad (48)$$

It can be shown under suitable and reasonable technical conditions that the following assumption is satisfied

[CW] $\bar{V}_t(\lambda) - \bar{V}^*_t(\lambda)$ converges weakly to zero, as $\lambda \to 0$.

We note in passing that $\mathbb{P}(\bar{\tau}^* \geq \bar{\tau})$ tends to 1 as $\lambda \to 0$. Due to assumption [CW] we can work with the asymptotic properties of $(\bar{U}^*_n)$ and thus henceforth we will focus on this process.

The asymptotic covariance matrices of the empirical means of the centered correction terms $(\bar{H}_n^*(\bar{U}) - \bar{h}^*(\bar{U}))$, can be expressed, under reasonable conditions (e.g., [10]) as

$$\bar{S}(\bar{U}) := \sum_{k=-\infty}^{+\infty} \mathbb{E}\,[(\bar{H}_k^*(\bar{U}) - \bar{h}^*(\bar{U})(\bar{H}_0^*(\bar{U}) - \bar{h}^*(\bar{U}))^{\mathrm{T}}, \quad (49)$$

where H_k^* and h^* denote the limit of H_k and h as $\lambda \downarrow 0$.

It can be easily seen that, in the case of the approximate LMS method with momentum (38), for $\bar{U} = \bar{U}^* = 0$, we get

$$\bar{S} := \bar{S}(0) = c^2 \begin{bmatrix} S & S \\ S & S \end{bmatrix}. \quad (50)$$

In analogy with Condition 2 of [2], we have:

[C3'] We assume that the process defined by

$$\bar{L}_t(\lambda) = \sum_{n=0}^{[t/\lambda]-1} \left(\bar{H}_n^*(\bar{\bar{U}}_{\lambda n}^*) - \bar{h}^*(\bar{\bar{U}}_{\lambda n}^*) \right) \sqrt{\lambda}, \quad (51)$$

converges weakly, as $\lambda \to 0$, to a time-inhomogeneous zero-mean Brownian motion $(\bar{L}_t)$ with local covariances $(\bar{S}(\bar{\bar{U}}_t^*))$. Then, analogously to Theorem 1, also using Theorem 2 of [2], the following weak convergence result:

Theorem 2. *Under conditions C0, C1, C2', C3' and CW, process $(\bar{V}_t(\lambda))$ converges weakly, as $\lambda \to 0$, to a process $(\bar{Z}_t)$ satisfying the linear stochastic differential equation (SDE),*

$$d\bar{Z}_t = \bar{B}_* \bar{Z}_t dt + \bar{S}^{1/2}(\bar{\bar{U}}_t^*) d\bar{W}_t, \quad (52)$$

for $t \geq 0$, with initial condition $\bar{Z}_0 = 0$, where $(\bar{W}_t)$ is a standard Brownian motion in $\mathbb{R}^{2p}$ and $\bar{B}_*$ is

$$\bar{B}_* := \lim_{\lambda \downarrow 0} \bar{B}_\lambda = \begin{bmatrix} 0 & 0 \\ 0 & -I \end{bmatrix} + c \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix} \otimes R^*. \quad (53)$$

Let us denote the asymptotic covariance matrix of the process $(\bar{Z}_t)$ by $\bar{P}$. Then, matrix $\bar{P}$ is the unique solution of the algebraic Lyapunov equation

$$\bar{B}_* \bar{P} + \bar{P} \bar{B}_*^{\mathrm{T}} + \bar{S} = 0, \quad (54)$$

where matrix $\bar{S} := \bar{S}(0)$ is given by equation (50).

The relationship between $\bar{P}$ and P_0 will be given in Lemma 2. Assuming that the weak convergence of $(\bar{V}_t(\lambda))$ to $\mathcal{N}(0, \bar{P})$, when $\lambda \to 0$ and $t \to \infty$, can be established, we will be able to infer a weak convergence result for the original error process.

IV. Comparing LMS with and without Momentum

The main aim of this section is to compute the asymptotic covariance of the weak limit process associated with momentum LMS and compare it to that of plain LMS. We do this in two steps. First, we compute the asymptotic covariance of the transformed process, then, we map it to the original space.

The asymptotic covariance matrix of process $(\bar{Z}_t)$, namely, the one obtained from the extended and transformed filter coefficient estimaton error process of LMS with momentum, is denoted by $\bar{P}$. Matrix $\bar{P}$ satisfies the Lyapunov equation

$$\bar{B}_* \bar{P} + \bar{P} \bar{B}_*^{\mathrm{T}} + \bar{S} = 0, \quad (55)$$

where $\bar{S}$ and $\bar{B}_*$ are defined by (50) and (53), respectively.

Lemma 2. *The solution of the Lyapunov equation* (55) *is*

$$\bar{P} = \frac{c}{2} \begin{bmatrix} cS + 2P_0 & cS \\ cS & cS \end{bmatrix}. \quad (56)$$

The proof of Lemma 2 can be found in Appendix B.

With Theorem 2 and matrix $\bar{P}$ at hand, we aim at establishing a weak convergence result and a corresponding covariance matrix for the LMS method with momentum.

Recall that the linear transformation introduced for the state space recursion, $\bar{U}_n = T U_n$, implies that $U_n = T^{-1} \bar{U}_n$. However, matrix $T^{-1} = T^{-1}(\gamma)$ depends on γ, and $T^{-1}(1)$ is singular.

Nevertheless, since $(\bar{V}_t(\lambda)) \Rightarrow (\bar{Z}_t)$, as $\lambda \to 0$, where "$\Rightarrow$" denotes weak convergence; and $T^{-1}(\gamma) \to T_1^+$, as $\gamma \to 1$, where $T^{-1}(\gamma)$ and T_1^+ are constant matrices; we can apply Slutsky's theorem for Polish spaces to conclude that $(T^{-1}(\gamma)\bar{V}_t(\lambda)) \Rightarrow (T_1^+ \bar{Z}_t)$, as $\gamma \to 1$ (or, equivalently, $\lambda \to 0$, since $\lambda = 1 - \gamma$).

In other words, we essentially established that, as $\lambda \to 0$,

$$\lambda^{-1/2} \left(U_{[t/\lambda]} - T^{-1} \bar{\bar{U}}_{[t/\lambda]} \right) \Rightarrow (T_1^+ \bar{Z}_t). \quad (57)$$

Let us denote the asymptotic covariance matrix of process $(T_1^+ \bar{Z}_t)$ by P. Matrix P can be computed from $\bar{P}$ by

$$P = T_1^+ \bar{P} (T_1^+)^{\mathrm{T}} = c \begin{bmatrix} P_0 & P_0 \\ P_0 & P_0 \end{bmatrix}, \quad (58)$$

using the special structure of matrix T_1^+, see (25). As this matrix was obtained from a "doubled" process, cf. (21), its submatrices provide the corresponding covariance in the original space. Now we can state the following theorem:

Theorem 3. *Assume C0, C1, C2, C2', C3, C3', CW and that the weak convergences carry over to $\mathcal{N}(0, P_0)$ and $\mathcal{N}(0, P)$, as $t \to \infty$, in case of plain and momentum LMS, respectively. Then, the covariance (sub)matrix of the asymptotic distribution associated with LMS with momentum is $c \cdot P_0$, where P_0 is the corresponding covariance of plain LMS and $c = \mu/(1-\gamma)^2$.*

Recall that constants μ and γ are the gains of the correction and momentum terms, respectively. Then, for any μ and γ the asymptotic covariances of the associated processes of plain and momentum LMS methods differ only by a constant factor.

If we set $c = 1$, then the two asymptotic covariances are the same, and in this sense the two algorithms are equivalent.

However, while the weak convergence of standard LMS was obtained by normalizing with $\mu^{-1/2}$, in case of LMS with momentum, we need to normalize with $\lambda^{-1/2}$, where $\lambda = \sqrt{\mu}$, which implies a slower convergence to the limiting process; in fact there is an order of magnitude difference.

We can decrease the covariance of the asymptotic distribution for the momentum LMS by decreasing c, however, since $\lambda = \sqrt{\mu/c}$, this will further slow the convergence down.

If, on the contrary, we want a smaller normalization factor for the case of LMS with momentum by setting c large enough, it will obviously increase the covariance of the asymptotic distribution. Therefore, there is a trade-off between achieving a small asymptotic covariance and having a fast rate (i.e., smaller normalization factors for the weak convergence).

V. Conclusions

In this paper we have presented the outline of a transparent proof related to a recent result [9]. We studied the asymptotic behavior of the LMS method with momentum, under different, but significantly more realistic conditions. The key technical tool of our analysis was a beautiful and powerful weak convergence result of [2]. We slightly extended the setup of [9] by allowing the correction and momentum gains to be independently chosen, resulting in a trade-off between the rate and the covariance of the asymptotic distribution.

References

[1] B. Widrow and M. E. Hoff, “Adaptive switching circuits,” tech. rep., Standford Electrodics Lab, Standford University, California, 1960.

[2] J. A. Bucklew, T. G. Kurtz, and W. A. Sethares, “Weak convergence and local stability properties of fixed step size recursive algorithms,” *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 966–978, 1993.

[3] A. Benveniste, M. Métivier, and P. Priouret, *Adaptive algorithms and stochastic approximations*. Springer Science & Business Media, 1990.

[4] J. A. Joslin and A. J. Heunis, “Law of the iterated logarithm for a constant-gain linear stochastic gradient algorithm,” *SIAM Journal on Control and Optimization*, vol. 39, no. 2, pp. 533–570, 2000.

[5] L. Gerencsér, “Rate of convergence of the LMS method,” *Systems & Control Letters*, vol. 24, no. 5, pp. 385–388, 1995.

[6] H. N. Chau, C. Kumar, M. Rásonyi, and S. Sabanis, “On fixed gain recursive estimators with discontinuity in the parameters,” *arXiv preprint arXiv:1609.05166*, 2016.

[7] B. T. Polyak, “Some methods of speeding up the convergence of iteration methods,” *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964.

[8] B. T. Polyak, *Introduction to optimization*. Optimization Software, 1987.

[9] K. Yuan, B. Ying, and A. H. Sayed, “On the influence of momentum acceleration on online learning,” *Journal of Machine Learning Research*, vol. 17, no. 192, pp. 1–66, 2016.

[10] P. Billingsley, *Convergence of Probability Measures*. John Wiley & Sons, 2nd ed., 1999.

Appendix A
Proof of Lemma 1

Proof. It is sufficient to prove the lemma for $\lambda = 0$. We may also assume $c = 1$, simply replacing R^* by cR^* in the proof below. Then, using the Schur complement corresponding to the $(1,1)$ block, the characteristic polynomial of $\bar{B}$ is

$$\det(\bar{B} - \rho I) = \begin{bmatrix} -R^* - \rho I & R^* \\ -R^* & R^* - I - \rho I \end{bmatrix} = \tag{59}$$
$$\det(-R^* - \rho I)\ \det\left(R^* - I - \rho I + R^*\left(-R^* - \rho I\right)^{-1} R^*\right).$$

The matrix in the second term can be written, using the commutativity of $\left(-R^* - \rho I\right)^{-1}$ and R^*, as

$$\left(-R^* - \rho I\right)^{-1}\left(\left(-R^* - \rho I\right)\left(R^* - I - \rho I\right) + (R^*)^2\right) \tag{60}$$

Since R^* was assumed to be positive definite, it is sufficient to show that the roots of

$$\det\left(\rho^2 I + \rho I + R^*\right) = 0. \tag{61}$$

Performing a diagonalization of R^* via an oprthonormal coordinate transformation, and denoting the eigenvalues of R^* by σ_k, the left hand side can be written

$$\prod_{k=1}^{p}\left(\rho^2 + \rho + \sigma_k\right). \tag{62}$$

Now $\sigma_k > 0$ for all k implies the claim of the lemma by well-known, elementary calculations. □

Appendix B
Proof of Lemma 2

Proof. First, we can observe that

$$\bar{B}_*\bar{P} = \begin{bmatrix} 0 & 0 \\ 0 & -I \end{bmatrix}\begin{bmatrix} \bar{P}_{11} & \bar{P}_{12} \\ \bar{P}_{21} & \bar{P}_{22} \end{bmatrix} + c\begin{bmatrix} -R^* & R^* \\ -R^* & R^* \end{bmatrix}\begin{bmatrix} \bar{P}_{11} & \bar{P}_{12} \\ \bar{P}_{21} & \bar{P}_{22} \end{bmatrix}$$
$$= \begin{bmatrix} 0 & 0 \\ -\bar{P}_{21} & -\bar{P}_{22} \end{bmatrix} + c\begin{bmatrix} -R^*(\bar{P}_{11} - \bar{P}_{21}) & -R^*(\bar{P}_{12} - \bar{P}_{22}) \\ -R^*(\bar{P}_{11} - \bar{P}_{21}) & -R^*(\bar{P}_{12} - \bar{P}_{22}) \end{bmatrix}$$

and thus

$$\bar{P}^{\mathrm{T}}\bar{B}_*^{\mathrm{T}} = \begin{bmatrix} 0 & -\bar{P}_{21} \\ 0 & -\bar{P}_{22} \end{bmatrix} + c\begin{bmatrix} -(\bar{P}_{11} - \bar{P}_{21})R^* & -(\bar{P}_{11} - \bar{P}_{21})R^* \\ -(\bar{P}_{12} - \bar{P}_{22})R^* & -(\bar{P}_{12} - \bar{P}_{22})R^* \end{bmatrix}.$$

One then observes that the (1,1) element of the (block) matrix $\bar{B}_*\bar{P} + \bar{P}^{\mathrm{T}}\bar{B}_*^{\mathrm{T}} + \bar{S}$ satisfies the equation

$$-cR^*(\bar{P}_{11} - \bar{P}_{21}) - (\bar{P}_{11} - \bar{P}_{21})cR^* + c^2 S = 0. \tag{63}$$

It follows from the uniqueness of the solution of the Lyapunov equation associated with the standard LMS, i.e., (18), that

$$\bar{P}_{11} - \bar{P}_{21} = cP_0. \tag{64}$$

The latter also implies (by using transposition) that

$$\bar{P}_{11} - \bar{P}_{12} = cP_0. \tag{65}$$

Summing the last two equations yields

$$2\bar{P}_{11} - \bar{P}_{12} - \bar{P}_{21} = 2cP_0. \tag{66}$$

Moreover, the elements (1,2), (2,1) and (2,2) of the (block) matrix $\bar{B}_*\bar{P} + \bar{P}^{\mathrm{T}}\bar{B}_*^{\mathrm{T}} + \bar{S}$ satisfy the following equations:

$$-cR^*(\bar{P}_{12} - \bar{P}_{22}) - (\bar{P}_{11} - \bar{P}_{12})cR^* - \bar{P}_{12} + c^2 S = 0 \tag{67}$$
$$-cR^*(\bar{P}_{11} - \bar{P}_{21}) - (\bar{P}_{21} - \bar{P}_{22})cR^* - \bar{P}_{21} + c^2 S = 0 \tag{68}$$
$$-cR^*(\bar{P}_{12} - \bar{P}_{22}) - (\bar{P}_{21} - \bar{P}_{22})cR^* - 2\bar{P}_{22} + c^2 S = 0 \tag{69}$$

and recall the equation for the element (1,1), i.e. (63),

$$-cR^*(\bar{P}_{11} - \bar{P}_{21}) - (\bar{P}_{11} - \bar{P}_{21})cR^* + c^2 S = 0. \tag{70}$$

When adding (70) and (69) together and subtracting from them (67) and (68), one concludes that the overall sum of terms having cR^* as a multiplier vanishes. Consequently, due to (66)

$$\bar{P}_{22} = \bar{P}_{11} - cP_0 \tag{71}$$

which yields, also using (65) and (64), that

$$\bar{P} = \begin{bmatrix} \bar{P}_{11} & \bar{P}_{11} - cP_0 \\ \bar{P}_{11} - cP_0 & \bar{P}_{11} - cP_0 \end{bmatrix}. \tag{72}$$

Thus, equation (69) is reduced to $2\bar{P}_{22} = S$, which yields $\bar{P}_{22} = c^2 S/2$, and consequently due to (71), one obtains $\bar{P}_{11} = S/2 + P_0$ and the solution to the Lyapunov equation is (56). □