Probability of Divergence for the Least-Mean Fourth Algorithm

Vítor H. Nascimento, *Member, IEEE,* and José Carlos M. Bermudez, *Senior Member, IEEE*

***Abstract*—In this paper, it is shown that the least-mean fourth (LMF) adaptive algorithm is not mean-square stable when the regressor input is not strictly bounded (as happens, for example, if the input has a Gaussian distribution). For input distributions with infinite support, even for the Gaussian distribution, the LMF always has a nonzero probability of divergence, no matter how small the step-size is chosen. This result is proven for a slight modification of the Gaussian distribution in a one-tap filter and corroborated with several simulations. In addition, an upper bound is given for the probability of divergence of LMF as a function of the filter length, input power, step-size, and noise variance, for the case of Gaussian regressors. The results reported in this paper provide tools for designers to better understand the behavior of the LMF algorithm and decide on the convenience or not of its use for a given application.**

***Index Terms*—Adaptive filters, stability, stochastic processes.**

I. Introduction

THE least-mean fourth (LMF) algorithm was proposed almost 20 years ago [1] as an alternative to the least-mean-square (LMS) algorithm. The goal was to achieve a lower steady-state misadjustment for a given speed of convergence using a different cost-function. It is not difficult to intuitively understand how this is accomplished if we compare the update laws of both algorithms:

$$\begin{aligned} \text{LMS:}\quad \boldsymbol{W}_2(n+1) &= \boldsymbol{W}_2(n) + \mu e_2(n)\boldsymbol{X}(n) \\ e_2(n) &= d(n) - \boldsymbol{W}_2(n)^T\boldsymbol{X}(n) \qquad (1) \\ \text{LMF:}\quad \boldsymbol{W}(n+1) &= \boldsymbol{W}(n) + \mu e(n)^3\boldsymbol{X}(n) \\ e(n) &= d(n) - \boldsymbol{W}(n)^T\boldsymbol{X}(n) \qquad (2) \end{aligned}$$

where $\boldsymbol{W}_2(n)$ and $\boldsymbol{W}(n) \in \mathbb{R}^M$ are current estimates of a parameter (column) vector $\boldsymbol{W}_o \in \mathbb{R}^M$. $\boldsymbol{X}(n) \in \mathbb{R}^M$ is a known regressor vector, and $d(n)$ is a known scalar sequence, usually called *desired* sequence. Note that we use capital bold letters $\boldsymbol{W}$ and $\boldsymbol{X}$ for the filter parameter and regressor vectors (in the sequence, we also use $\boldsymbol{V}$ for the filter parameter error vector).

Manuscript received August 23, 2004; revised March 12, 2005. This work was partly supported by CNPq under grants no. 303361/2004-2, 308095/2003-0, and 472762/2003-6, and by FAPESP under grant no. 2004/15114-2. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Martin Haardt.

V. H. Nascimento is with the Department of Electronic Systems Engineering, University of São Paulo, São Paulo, SP, 05508-900 Brazil (e-mail: vitor@lps.usp.br).

J. C. M. Bermudez is with the Department of Electrical Engineering, Federal University of Santa Catarina, Florianópolis, SC, 88040-900, Brazil (e-mail:j.bermudez@ieee.org).

Digital Object Identifier 10.1109/TSP.2006.870546

It is well known [2], [3] that, if $\{d(n), \boldsymbol{X}(n)\}$ are zero-mean, jointly wide-sense stationary sequences, one can always model the relationship between $d(n)$ and $\boldsymbol{X}(n)$ as

$$d(n) = \boldsymbol{W}_o^T\boldsymbol{X}(n) + e_0(n), \qquad (3)$$

where $e_0(n)$ is a zero-mean scalar sequence, uncorrelated with $\boldsymbol{X}(n)$ and with variance $\mathrm{E}\{e_0(n)^2\} = \sigma_0^2$ ($\mathrm{E}\{\cdot\}$ is the statistical expectation operator). In this context, $\boldsymbol{W}_o$ is called the *Wiener solution*. The LMS estimate $\boldsymbol{W}_2(n)$ converges in the mean to $\boldsymbol{W}_o$ with a finite covariance matrix, as long as the step-size μ is small enough. It is also known that, for small μ, the LMS steady-state mean-square estimation error (MSE) is approximately given by

$$\lim_{n\to\infty} \mathrm{E}\left\{e_2(n)^2\right\} \approx \sigma_0^2 + \mu\sigma_0^2\frac{\mathrm{Tr}(\boldsymbol{R}_x)}{2} \qquad (4)$$

where $\boldsymbol{R}_x = \mathrm{E}\{\boldsymbol{X}(n)\boldsymbol{X}(n)^T\}$ is the autocorrelation matrix of $\boldsymbol{X}(n)$, and $\mathrm{Tr}(\boldsymbol{R}_x)$ is its trace. The second term in the right-hand side of (4) is the steady-state excess MSE, which is caused by the fluctuations of $\boldsymbol{W}_2(n)$ around $\boldsymbol{W}_o$ after convergence. This term is proportional to μ. It can also be shown that the worst-case rate of convergence of $\mathrm{E}\{e_2(n)^2\}$ is $1 - 2\mu\lambda_{\min}$ for small μ, where $\lambda_{\min}$ is the smallest eigenvalue of $\boldsymbol{R}_x$.

One can see that μ controls the behavior of the algorithm, and that two important goals are competing: for fast convergence, one would use a large step-size μ, but to achieve low steady-state MSE, a smaller step-size would be better. One intuitive way to understand the LMF algorithm is to consider it as a variant to LMS with a variable step-size $\bar{\mu}(n) = e(n)^2\mu$. When the error is large, adaptation is faster; when the error is small, adaptation is slower, resulting in a fast convergence with small steady-state error.

Regarding the LMF algorithm in this way also highlights its main drawback: If the error gets too large, the "equivalent step-size" $\bar{\mu}(n)$ may get large enough for the algorithm to diverge. This happens for inputs with long tail distributions (and even for the Gaussian distribution, as we show in the following sections). Thus, one can expect the convergence properties of the LMF algorithm to be dependent on 1) the initial weight vector estimate $\boldsymbol{W}(0)$ and 2) the probability that the error gets too large at any given algorithm iteration.

Recent literature [4]–[7] studied the behavior of the LMF algorithm for Gaussian noise and regressors, finding approximate mean-square stability conditions. Other literature also studied the stability of LMF. For example, [8] proves that $\boldsymbol{W}(n)$ converges to a ball around $\boldsymbol{W}_o$ when the regressor vector sequence is bounded, i.e., when there is a $B < \infty$ such that $\|\boldsymbol{X}(n)\| < B$ for all n ($\|\cdot\|$ is the Euclidean norm). Deterministic results such

as [8] tend to be very conservative, requiring that the step-size be quite small in order to guarantee stability. However, the fact that the regressor vector attains large values with a small probability will usually not destabilize an algorithm: this is why LMS is mean-square stable for distributions with finite fourth-order moments (this condition is for independent regressors, see [3], [9]) and is one of the reasons why LMS is so robust.

In this paper, we argue that there is always a nonzero probability of divergence in any given realization of the LMF algorithm when the entries of $\boldsymbol{X}(n)$ have a probability density function (pdf) with infinite support, i.e., there is a small (but nonzero) probability that an entry is larger than any $C > 0$. This is what happens, for example, with the Gaussian distribution. The practical consequence of this result is that LMF is not robust to inputs that have small probabilities of large errors. Rare gross errors in the regressor sequence may make the algorithm unstable.

We prove this property for a simple case, when $M = 1$ (scalar filter) and the distribution of $X(n)$ is a slight modification of the normal, with pdf given by

$$p_X(x) = \begin{cases} 0, & \text{if } |x| < \epsilon \\ \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(|x|-\epsilon)^2}{2\sigma_x^2}}, & \text{if } |x| \geq \epsilon \end{cases} \tag{5}$$

for $\epsilon > 0$.

This result means that the LMF algorithm is *not* mean-square stable with these near-Gaussian inputs. In other words, the steady-state MSE is unbounded. Notice that this result does not imply that *every* realization of the LMF algorithm will result in divergence. In fact, the probability of divergence on a single realization of the algorithm decreases as the step-size is decreased to zero, as we show in a few examples further on.

In light of this result, we can better understand the approximations given in the literature for the MSE of the LMF algorithm. For small step-sizes, the probability of divergence is very small and the approximations in the literature are in fact computing $\mathrm{E}\{e(n)^2 | \text{the filter coefficients did not diverge}\}$. Thus, there is not a step-size boundary $\mu = \mu_{\max}$ above which the algorithm starts diverging. What happens is that the probability of divergence increases with μ. This property has a similarity with what happens with LMS, as explained in [10]: The LMS algorithm has a range of step-sizes for which it converges with probability one, but diverges in the mean-square (MS) sense; a range for which the algorithm diverges almost always (and in the MS sense); and a range for which it converges in the MS sense. The LMF properties differ in that MS convergence happens only when both the regressor and the noise are bounded. In practical terms, this means that the step-size should be chosen rather conservatively.

We also find an approximation for the probability of divergence, assuming that $\{\boldsymbol{X}(n)\}$ is an independent identically distributed (i.i.d.) vector sequence with an M-dimensional Gaussian distribution and covariance $\sigma_x^2 I$. Our approximation is a function of the filter length M, of the initial weight vector $\boldsymbol{W}(0)$, of σ_x^2, and of the distribution of $e_0(n)$. In [11], we present some simulations for LMF and extend these results to the least-mean mixed-norm algorithm (LMMN) [3].

In the next sections, we find an approximate model for this behavior, and provide several simulations corroborating our affirmations.

II. Simple Example of Instability

Our goal here is to give a simple example showing that LMF will have a nonzero probability of divergence for a rather nice distribution of the regressor input, no matter how small we choose the (nonzero) step-size. We believe that this scalar example explains clearly what is the mechanism of divergence, so there is no need to expand the example for longer filters.[1]

A. Proof of Instability for Scalar Filters

Consider the LMF algorithm (2) applied with filter length $M = 1$ to identify a constant W_o, given an i.i.d. sequence $X(n)$ with pdf given by (5). Assume also that there is no noise, so $d(n) = W_o X(n)$. Defining the weight estimation error $V(n) = W_o - W(n)$, the LMF weight-error update equation is written

$$V(n+1) = \left(1 - \mu X(n)^4 V(n)^2\right) V(n). \tag{6}$$

We show first that there is a value $0 < K < \infty$ such that, if $|V(n)| > K$ for any n, then $\lim_{n\to\infty} |V(n)| = \infty$. Later we will show that the probability of $|V(n+1)| > K$ given $|V(n)| = \alpha$ is nonzero for all $\alpha > 0$. Let K be such that ($\delta > 0$ is any positive number)

$$\mu\epsilon^4 K^2 - 1 > 1 + \delta \Leftrightarrow K > \sqrt{\frac{2+\delta}{\mu\epsilon^4}}. \tag{7}$$

Given inequality (7) and since $|X(n)| \geq \epsilon$ by (5), it necessarily holds that

$$\left|1 - \mu X(n)^4 K^2\right| > \left|1 - \mu\epsilon^4 K^2\right| > 1 + \delta. \tag{8}$$

If we now assume that $|V(n)| > K$, (6) yields

$$\begin{aligned} |V(n+1)| &= \left|1 - \mu X(n)^4 V(n)^2\right| |V(n)| \\ &> \left|1 - \mu\epsilon^4 K^2\right| |V(n)| > (1+\delta)\,|V(n)| \end{aligned} \tag{9}$$

and we conclude that $|V(n)| \to \infty$ if at any time instant it happens that $|V(n)| > K$.

We complete our argument by showing that the probability of $|V(n+1)| > K$ given $|V(n)| = \alpha$ is nonzero for any $\alpha > 0$. Define, for a given $\alpha > 0$

$$\beta(\alpha) \triangleq \frac{K}{\mu\alpha^3} + \frac{1}{\mu\alpha^2}. \tag{10}$$

Then, for $|V(n)| = \alpha$, an input $X(n)$ such that $X(n)^4 > \beta(\alpha)$ leads to

$$\mu X(n)^4 V(n)^2 - 1 > \frac{K}{|V(n)|} \tag{11}$$

[1]Part of this section was presented at [12]. The paper may be obtained from http://www.lps.usp.br/~vitor/nascimento_rev.pdf

and thus

$$|V(n+1)| = \left|1 - \mu X(n)^4 V(n)^2\right| |V(n)| > K. \qquad (12)$$

Expressions (9) and (12) show that $|V(n)| \to \infty$ if $X(n)^4 > \beta(\alpha)$ for any given $|V(n)| = \alpha$. Thus, to prove that there is a nonzero probability of divergence, it remains to show that there is a nonzero probability that $X(n)^4 > \beta(\alpha)$, given that $|V(n)| = \alpha$. Using (5), it follows that

$$\begin{aligned} &\Pr\{|V(n+1)| > K \,|\, |V(n)| = \alpha\} \\ &\quad > \Pr\left\{X(n)^4 > \beta(\alpha) \,|\, |V(n)| = \alpha\right\} \\ &\quad = 2 \int_{\beta(\alpha)^{\frac{1}{4}}}^{\infty} p_X(x)dx > 0 \end{aligned} \qquad (13)$$

where $\Pr\{A|B\}$ is the probability of occurrence of A given B. This concludes the proof.

B. Gaussian Regressors

When $X(n)$ is normal ($\epsilon = 0$ in (5)), the simple proof above does not apply. However, we now present simulations showing that the result still holds.

Assume that LMF is applied to the same situation as before, but with $\epsilon = 0$. In our simulations, we evaluated the following:

- the probability of divergence of LMF, measured as follows: We ran $L = 10^6$ realizations of the algorithm, starting from the same initial condition $V(0) = 1$ and with zero noise; we counted a "divergence" every time the absolute error $|V(n)|$ became larger than 10^{100} (choosing this value in a very large range does not affect the results);
- the probability $P_>$ of $|V(1)| > |V(0)|$;
- the value $V_{1/2}$ for which the probability $\Pr\{|V(n+1)| > |V(n)| \,|\, |V(n)| = V_{1/2}\} = 0.5$;
- the probability $P_{>V_{1/2}}$ that $|V(1)| > V_{1/2}$, given the initial condition.

The probability of divergence was obtained experimentally. All other values can be computed as follows. We start by computing the pdf of $V(n+1)$ given $V(n)$. From (6), it is clear that

$$\begin{aligned} &\Pr\{V(n+1) < z | V(n) = Z > 0\} \\ &\quad = \Pr\left\{\left(1 - \mu X(n)^4 Z^2\right) Z < z\right\} \\ &\quad = \Pr\left\{X(n)^4 > \frac{1 - \frac{z}{Z}}{\mu Z^2}\right\}. \end{aligned} \qquad (14)$$

The pdf of $X(n)^4$ is given by

$$\begin{aligned} p_{X^4}(y) &= \frac{d \Pr\{X^4 < y\}}{dy} \\ &= \frac{d\left(2\int_0^{y^{\frac{1}{4}}} \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{x^2}{2\sigma_x^2}} dx\right)}{dy} \\ &= \frac{1}{\sqrt{8\pi}\sigma_x y^{\frac{3}{4}}} e^{-\frac{\sqrt{y}}{2\sigma_x^2}}, \quad y \geq 0. \end{aligned} \qquad (15)$$

TABLE I
OBSERVED PROBABILITY OF DIVERGENCE FOR SEVERAL STEP-SIZES, ALWAYS FOR $M = 1$ AND INITIAL CONDITION $|V(0)| = 1$. COLUMN DEFINITIONS ARE GIVEN IN THE TEXT

μ	$P_>$	$V_{1/2}$	$P_{>V_{1/2}}$	N_{div}/L
0.01	1.7×10^{-4}	31.1	5.2×10^{-14}	7×10^{-6}
0.02	1.6×10^{-3}	22.0	5.8×10^{-9}	3.0×10^{-4}
0.03	4.3×10^{-3}	17.9	5.4×10^{-7}	1.6×10^{-3}
0.04	7.8×10^{-3}	15.5	6.5×10^{-6}	4.4×10^{-3}
0.05	1.2×10^{-2}	13.9	3.3×10^{-5}	8.5×10^{-3}
0.06	1.6×10^{-2}	12.7	1.0×10^{-4}	1.4×10^{-2}
0.07	2.1×10^{-2}	11.7	2.4×10^{-4}	2.0×10^{-2}
0.08	2.5×10^{-2}	11.0	4.7×10^{-4}	2.8×10^{-2}
0.09	3.0×10^{-2}	10.4	8.0×10^{-4}	3.5×10^{-2}
0.10	3.4×10^{-2}	9.8	1.3×10^{-3}	4.3×10^{-2}
0.20	7.5×10^{-2}	7.0	1.2×10^{-2}	1.3×10^{-1}

Thus

$$\begin{aligned} &\Pr\{V(n+1) < z | V(n) = Z > 0\} \\ &\quad = \Pr\left\{X(n)^4 > \frac{1 - \frac{z}{Z}}{\mu Z^2}\right\} \\ &\quad = \int_{\frac{1-\frac{z}{Z}}{\mu Z^2}}^{\infty} \frac{1}{\sqrt{8\pi}\sigma_x y^{\frac{3}{4}}} e^{-\frac{\sqrt{y}}{2\sigma_x^2}} dy. \end{aligned} \qquad (16)$$

Finally, the desired pdf is obtained by differentiating (16) with respect to z

$$\begin{aligned} &p_{V(n+1)|V(n)}(z|V(n) = Z) \\ &\quad = \frac{d \Pr\{V(n+1) < z | V(n) = Z > 0\}}{dz} \\ &\quad = \frac{1}{\sqrt{8\pi}\sigma_x \mu^{\frac{1}{4}} Z^{\frac{3}{4}} (Z - z)^{\frac{3}{4}}} e^{-\frac{\sqrt{Z-z}}{2\sigma_x^2 \sqrt{\mu Z^3}}}. \end{aligned} \qquad (17)$$

Assuming $\sigma_x^2 = 1$, we can use (17) with $V(0) = 1$ (fixed) to determine the probabilities $P_> = \Pr\{|V(1)| > |V(0)|\}$ and $P_{>V_{1/2}} = \Pr\{|V(1)| > |V_{1/2}| \,|\, V(0) = 1\}$, and the point $V_{1/2} > 0$ for which $\Pr\{|V(1)| > V(0) \,|V(0) = V_{1/2}\} = 0.5$. These values are given, for several choices of μ, in Table I. The table also shows N_{div}, the observed number of realizations of the LMF algorithm for which $|V(n)| > 10^{100}$ for some n, as explained above.

The last column in Table I shows that the probability of divergence grows with the step-size. However, even for the largest step-size in the table, the filter coefficients behave rather nicely in most realizations. Fig. 1 shows three realizations for a scalar filter ($M = 1$ coefficient), with Gaussian i.i.d. input $X(n)$ with unit variance, step-size $\mu = 0.03$ (probability of divergence of 0.16% according to Table I), and initial condition $V(0) = 1$. The figure shows two realizations where the algorithm converged, and one for which the algorithm diverged. Note that divergence does not take long to become clear. This has been verified to be a typical algorithm behavior. In addition, we note that the probability of divergence depends on the initial condition: The larger the initial error $V(0)$, the larger the probability of divergence. This behavior is in agreement with the results derived in [7]. The same behavior is observed for filters with $M > 1$.

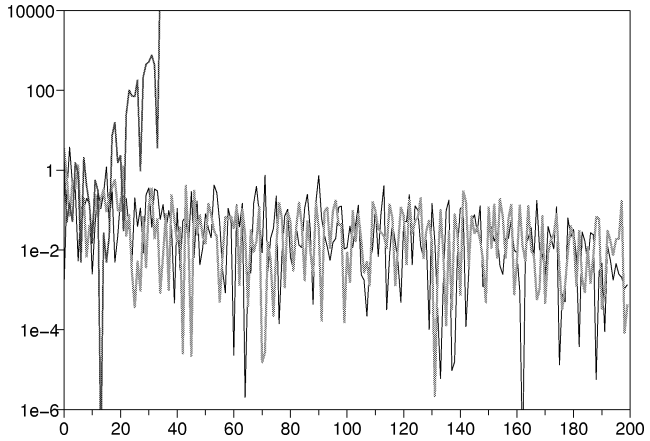

Fig. 1. Three runs of LMF with scalar regressors, true $W_o = 0, V(0) = W(0) = 1, \mu = 0.03, X(n) \sim N(0,1)$.

III. Probability of Divergence

We now turn to the problem of estimating the probability of divergence of LMF for filters of any length. First of all, we need to define clearly what is meant by divergence. Unacceptable behavior could be of many forms: the estimation error could grow to large values before decreasing, or it might stay at reasonable values for most of the time, but with bursts of large errors, or the estimation error could grow unboundedly. The usual definition of stability in adaptive filtering involves the variance of the estimates; i.e., one wishes that the variance of the estimation error remains bounded and not much larger than the noise variance. Given the kind of behavior we saw in the previous section, we shall employ in this paper the following definition for divergence:

Definition 1 (Divergence): In this paper, we say that a realization (a single run) of the LMF recursion diverged if $\lim_{n\to\infty} ||\boldsymbol{W}(n)|| = \infty$. We shall also say that a realization of the algorithm converged if it did not diverge (note that this is not the usual definition of divergence, but is adequate for our study).

We are interested in the following question: Given the initial condition $\boldsymbol{W}(0)$, the step-size μ, the filter length M, and the noise and regressor statistics, what is the probability that a realization of the filter will diverge?

For the scalar filter with the modified Gaussian distribution, if the absolute error $|V(n)|$ becomes as large as a certain value V, the LMF algorithm will necessarily diverge, as we saw in the previous section. The analysis is more complicated for the unmodified Gaussian distribution, since in this case, no matter how large $|V(n)|$ gets, there is a small nonzero probability that the error may return to reasonable values (however, this small probability quickly decreases as $|V(n)|$ increases). This makes the estimation of the probability of divergence a difficult problem. We propose here an approximate model that attempts to capture the essence of the dependence of the probability of divergence on μ, M, the variance of the input sequence σ_x^2, and the distribution of the noise $e_0(n)$.

A. Recursion for $||\boldsymbol{V}(n)||^2$

We start by finding an approximated recursion for $||\boldsymbol{V}(n)||^2$. We assume that $\{\boldsymbol{X}(n)\}$ is i.i.d. and Gaussian, and that the entries of $\boldsymbol{X}(n)$ are uncorrelated. Thus, $\mathrm{E}\{\boldsymbol{X}(n)\boldsymbol{X}(m)^T\} = \boldsymbol{0}$ for $m \neq n$ and $\boldsymbol{R}_x = \mathrm{E}\{\boldsymbol{X}(n)\boldsymbol{X}(n)^T\} = \sigma_x^2\boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix. We further assume that the pdf of $e_0(n)$ in (3) is even, so that $\mathrm{E}\{e_0(n)^k\} = 0$ whenever k is an odd integer.

Substituting $d(n)$ from (3) in (2), we obtain

$$e(n) = \boldsymbol{V}(n)^T\boldsymbol{X}(n) + e_0(n) \tag{18}$$

where $\boldsymbol{V}(n) = \boldsymbol{W}_o - \boldsymbol{W}(n)$. Writing the LMF recursion (2) in terms of $\boldsymbol{V}(n)$, we obtain

$$\boldsymbol{V}(n+1) = \boldsymbol{V}(n) - \mu\left(\boldsymbol{V}(n)^T\boldsymbol{X}(n) + e_0(n)\right)^3\boldsymbol{X}(n). \tag{19}$$

Defining $p(n) = \boldsymbol{V}(n)^T\boldsymbol{X}(n)$ to shorten the notation

$$\begin{aligned}\boldsymbol{V}(n+1) = \boldsymbol{V}(n) - \mu\big[&p(n)^3 + 3p(n)^2e_0(n) \\ &+3p(n)e_0(n)^2 + e_0^3(n)\big]\boldsymbol{X}(n).\end{aligned} \tag{20}$$

Now define $y(n) \triangleq \boldsymbol{V}(n)^T\boldsymbol{V}(n) = ||\boldsymbol{V}(n)||^2$. From (20), we have

$$\begin{aligned}y(n+1) =& y(n) - 2\mu\left[p(n)^4 + 3p(n)^3e_0(n)\right. \\ &\left.+3p(n)^2e_0(n)^2 + p(n)e_0(n)^3\right] \\ &+\mu^2\left[p(n)^6 + 6p(n)^5e_0(n) + 15p(n)^4e_0(n)^2\right. \\ &+20p(n)^3e_0(n)^3 + 15p(n)^2e_0(n)^4 \\ &\left.+6p(n)e_0(n)^5 + e_0(n)^6\right]||\boldsymbol{X}(n)||^2.\end{aligned} \tag{21}$$

Our goal is to estimate the probability that $\lim_{n\to\infty} y(n) = \infty$.

We need to simplify (21) to proceed. Thus, we make the approximation $y(n) \approx \mathrm{E}\{y(n)|\, y(n-1),\, \boldsymbol{X}(n)\}$. This approximation replaces the noise $e_0(n)$ and its powers by their means.

$$\begin{aligned}y(n+1) = y(n) &- 2\mu\left[p(n)^4 + 3p(n)^2\sigma_0^2\right] \\ &+\mu^2\left[p(n)^6 + 15p(n)^4\sigma_0^2 + 15p(n)^2\psi_0^4 + \eta_0^6\right]||\boldsymbol{X}(n)||^2\end{aligned}$$

where we used our assumption that $\mathrm{E}\{e_0(n)\} = \mathrm{E}\{e_0(n)^3\} = \mathrm{E}\{e_0(n)^5\} = 0$, and defined $\psi_0^4 = \mathrm{E}\{e_0(n)^4\}$, $\eta_0^6 = \mathrm{E}\{e_0(n)^6\}$.

To proceed, we need to approximate $(\boldsymbol{V}(n)^T\boldsymbol{X}(n))^{2k}$ in terms of $y(n)^k$ and of $||\boldsymbol{X}(n)||^{2k}$. Recalling our assumption that the entries of $\boldsymbol{X}(n)$ are uncorrelated and Gaussian (and thus also independent), we note that the vector $\boldsymbol{X}(n)$ can point to any direction in $\mathbb{R}^M$ with equal probability. Under our assumption of i.i.d. $\boldsymbol{X}(n)$, the directions of $\boldsymbol{V}(n)$ and of $\boldsymbol{X}(n)$ are independent, and using this observation, we show in Appendix I that the following approximations can be used:

$$\begin{aligned}\frac{\left(\boldsymbol{V}(n)^T\boldsymbol{X}(n)\right)^{2k}}{\left(||\boldsymbol{V}(n)||\,||\boldsymbol{X}(n)||\right)^{2k}} &\approx \mathrm{E}\left\{\frac{\left(\boldsymbol{V}(n)^T\boldsymbol{X}(n)\right)^{2k}}{\left(||\boldsymbol{V}(n)||\,||\boldsymbol{X}(n)||\right)^{2k}}\right\} \\ &\triangleq \alpha_M(k)\end{aligned} \tag{22}$$

with $\alpha_M(1) = 1/M$, $\alpha_M(2) = 3/(M(M+2))$, and $\alpha_M(3) = 15/(M(M+2)(M+4))$.

Substituting $(\boldsymbol{V}(n)^T\boldsymbol{X}(n))^{2k}$ by $\alpha_M(k)y(n)^k\|\boldsymbol{X}(n)\|^{2k}$, we obtain

$$y(n+1) \approx \left[1-\mu\left(6\sigma_0^2-15\mu\psi_0^4\|\boldsymbol{X}(n)\|^2\right)\frac{\|\boldsymbol{X}(n)\|^2}{M} - 3\mu\left(2-15\mu\sigma_0^2\|\boldsymbol{X}(n)\|^2\right)\frac{\|\boldsymbol{X}(n)\|^4}{M(M+2)}y(n) +15\mu^2\frac{\|\boldsymbol{X}(n)\|^8}{M(M+2)(M+4)}y(n)^2\right]y(n) + \mu^2\eta_0^6\|\boldsymbol{X}(n)\|^2. \quad (23)$$

B. Estimating the Probability of Divergence

What we want to evaluate now is the probability that the $y(n)$ given by recursion (23) grows unboundedly, given the initial condition $y(0) = \boldsymbol{V}(0)^T\boldsymbol{V}(0)$, and μ, σ_x^2, σ_0^2, ψ_0^4, η_0^6, and M.

Denote by $D(n)$ the factor between brackets multiplying $y(n)$ in (23), i.e.,

$$D(n) = 1-\mu\left(6\sigma_0^2-15\mu\psi_0^4\|\boldsymbol{X}(n)\|^2\right)\frac{\|\boldsymbol{X}(n)\|^2}{M} - 3\mu\left(2-15\mu\sigma_0^2\|\boldsymbol{X}(n)\|^2\right)\frac{\|\boldsymbol{X}(n)\|^4}{M(M+2)}y(n) + 15\mu^2\frac{\|\boldsymbol{X}(n)\|^8}{M(M+2)(M+4)}y(n)^2. \quad (24)$$

We show in Appendix II that $D(n)$ is always nonnegative.

We find an approximation P_d for the probability of divergence as follows. Recursion (23) converges if there is a fixed D_0 such that $0 \le D(n) < D_0 < 1$ for all n, since in this case $y(n+1) \le D_0 y(n) + \mu\eta_0^6\|\boldsymbol{X}(n)\|^2$, and thus

$$y(n) \le D_0^n y(0) + \mu\eta_0^6\sum_{k=0}^{n-1} D_0^{n-1-k}\|\boldsymbol{X}(k)\|^2. \quad (25)$$

If $D(n) \le D_0 < 1$ for all n, $\|\boldsymbol{X}(n)\|^2$ must necessarily be bounded (since $D(n) \to \infty$ when $\|\boldsymbol{X}(n)\| \to \infty$). This boundedness of $\|\boldsymbol{X}(n)\|$ together with (25) imply that $y(n)$ remains bounded (and therefore, according to our definition, the algorithm converges).

However, it may happen that the algorithm converges even if a few of the $D(n)$ are larger than 1. Thus, $P_c = \Pr\{0 < D(n) < 1 \text{ for all } n \ge 0\}$ is a lower bound for the probability of convergence, and $P_d = 1 - P_c$ is an upper bound for the probability of divergence. To evaluate P_c, we do the following.

1) Find the probabilities $\Pr\{D(n) < 1 | y(n) = \hat{y}(n)\}$, for $0 \le n \le N$, starting from a given $y(0)$.
2) A problem in the previous expression is that $y(n)$, for $n > 0$, varies for each realization of the filter. In order to proceed, we must find an approximation $\hat{y}(n)$ for $y(n)$.[2]
3) Make

$$P_c \approx \prod_{n=0}^{N}\Pr\{D(n) < 1 | y(n) = \hat{y}(n)\}. \quad (26)$$

Assuming for now that we have an approximation $\hat{y}(n)$, let us find the probability in Step 1. From (24), we see that $D(n) < 1$ means

$$1-\mu\left(6\sigma_0^2-15\mu\psi_0^4\|\boldsymbol{X}(n)\|^2\right)\frac{\|\boldsymbol{X}(n)\|^2}{M} - 3\mu\left(2-15\mu\sigma_0^2\|\boldsymbol{X}(n)\|^2\right)\frac{\|\boldsymbol{X}(n)\|^4}{M(M+2)}y(n) + 15\mu^2\frac{\|\boldsymbol{X}(n)\|^8}{M(M+2)(M+4)}y(n)^2 < 1,$$

which, if we let $z \triangleq \|\boldsymbol{X}(n)\|^2$, reduces to

$$Q(z) \triangleq 6\sigma_0^2 + \left(\frac{6}{M+2}y(n) - 15\mu\psi_0^4\right)z - \frac{45\mu\sigma_0^2}{M+2}y(n)z^2 - 15\frac{\mu}{(M+2)(M+4)}y(n)^2z^3 > 0.$$

$Q(z)$ has no positive zero if $y(n) = \sigma_0^2 = \psi_0^4 = 0$; otherwise $Q(z)$ has one and only one positive real zero $z_0(n)$ for $y(n) \ge 0$, as we show now:

1) If the noise is identically zero (i.e., $\sigma_0^2 = \psi_0^4 = 0$), then $Q(z)$ reduces to

$$Q(z)|_{\text{zero noise}} = z\left(\frac{6}{M+2}y(n) - \frac{15\mu}{(M+2)(M+4)}y(n)^2z^2\right)$$

and the only positive zero is trivially

$$z_0(n)|_{\text{zero noise}} = \sqrt{\frac{2(M+4)}{5\mu y(n)}}.$$

2) If $y(n) = 0$ and $\sigma_0^2 > 0$, $Q(z)$ has degree one, and the single positive zero is

$$z_0(n)|_{y(n)=0} = \frac{2\sigma_0^2}{5\mu\psi_0^4}.$$

3) For nonzero noise and $y(n) > 0$, note that $Q(z)$ is of the form $a + bz - cz^2 - dz^3$, where $a, c, d > 0$, but b is indefinite. Since $a > 0$, $Q(0) > 0$, and since $d > 0$, $\lim_{z\to\infty} Q(z) = -\infty$, so there is at least one positive zero for $Q(z)$. We argue that this zero is unique as follows. The first- and second-order derivatives of $Q(z)$ are $Q'(z) = b - 2cz - 3dz^2$, $Q''(z) = -2c - 6dz$. Since $Q''(z) < 0$ for $z > 0$, $Q'(z)$ is strictly decreasing for $z > 0$. $Q'(z)$ will therefore have only one positive zero if $b > 0$, and none if $b \le 0$. Since $Q'(z)$ is strictly decreasing, it may have at most one positive zero $z_1 > 0$.

[2]The use of $\hat{y}(n)$, of (22), and of the expected values of the powers of $e_0(n)$, makes our analysis approximate. We validate our approximations in Section IV.

This implies that $Q(z)$ may have at most one finite extremum point $z_1 > 0$. $Q(z)$ may therefore follow one of the two following paths for z increasing from 0 to ∞:

a) if $b > 0$, $Q(z)$ starts from $Q(0) > 0$, increases for $0 < z < z_1$, reaches a maximum at z_1, then strictly decreases to $-\infty$;
b) if $b \leq 0$, $Q(z)$ starts at $Q(0) > 0$ and strictly decreases to $-\infty$ for $z > 0$.

In both cases, $Q(z)$ crosses the z axis once and only once for $z > 0$, and the zero $z_0(n) > 0$ must be unique. The sets $\{z \in \mathbb{R} |\ z > 0,\ Q(z) > 0\}$ and $\{z \in \mathbb{R} |\ 0 < z < z_0(n)\}$ are therefore equal, and our probability is given by (recall that $z_0(n)$ depends on $y(n)$)

$$\Pr\{D(n) < 1 | y(n) = \hat{y}(n)\} = \Pr\left\{\|\boldsymbol{X}(n)\|^2 \leq z_0(n) | y(n) = \hat{y}(n)\right\}. \quad (27)$$

Since the entries of $\boldsymbol{X}(n)$ are Gaussian and i.i.d., $\|\boldsymbol{X}(n)\|^2/\sigma_x^2$ follows a χ^2 distribution with M degrees of freedom, and the last probability in (27) can be easily evaluated.

We still need approximations $\hat{y}(n)$ to $y(n) = \|\boldsymbol{V}(n)\|^2$ at every time instant n. The most reasonable choice would be the median of the distribution of $\|\boldsymbol{V}(n)\|^2$, but this quantity is not easily computable. Another approximation would be to use $\hat{y}(n) = \mathrm{E}\{\|\boldsymbol{V}(n)\|^2\}$. However, this choice turns out to be inconvenient, since $\mathrm{E}\{\|\boldsymbol{V}(n)\|^2\}$ in fact diverges, as we noted in Section II. We propose in the next section an alternative choice for $\hat{y}(n)$ that gives reasonable results.

C. Evaluation of $\hat{y}(n)$

The expected value of (19) for $\boldsymbol{X}(n)$ Gaussian with covariance $\sigma_x^2 I$ has been evaluated in [6] as

$$E\{\boldsymbol{V}(n+1)\} = E\{\boldsymbol{V}(n)\} - 3\mu\left[\sigma_0^2 + \sigma_x^2 E\left\{\boldsymbol{V}(n)^T\boldsymbol{V}(n)\right\}\right]\sigma_x^2 E\{\boldsymbol{V}(n)\}. \quad (28)$$

To proceed with the determination of the estimate $\hat{y}(n)$, we use the approximations $\boldsymbol{V}(n) \approx E\{\boldsymbol{V}(n)\}$, $\hat{y}(n) \approx \mathrm{E}\{\boldsymbol{V}(n)\}^T\mathrm{E}\{\boldsymbol{V}(n)\}$. These approximations are good in the beginning of the adaptation phase, when $\boldsymbol{V}(n)$ is dominated by its mean value (divergence is most likely to occur in the initial iterations, as our simulations show). Multiplying (28) by its transpose leads to a recursion for $\mathrm{E}\{\boldsymbol{V}(n)\}^T\mathrm{E}\{\boldsymbol{V}(n)\}$, which we use to define $\hat{y}(n)$ as

$$\hat{y}(n+1) = \left[1 - 3\mu\sigma_x^2\left(\sigma_0^2 + \sigma_x^2\hat{y}(n)\right)\right]^2\hat{y}(n)$$
$$\hat{y}(0) = y(0) = \mathrm{E}\left\{\boldsymbol{V}(0)^T\boldsymbol{V}(0)\right\} = \boldsymbol{V}(0)^T\boldsymbol{V}(0). \quad (29)$$

Note that this choice for $\hat{y}(n)$ is only reasonable if μ is small. If the step-size is so large that $1 - 3\mu\sigma_x^2(\sigma_0^2 + \sigma_x^2\hat{y}(0)) = 0$, then $\hat{y}(n) = 0$ for $n \geq 1$, which would be far from the actual behavior of the algorithm.

We now have all the necessary elements for our estimate of the probability of divergence for LMF.

Algorithm 1: Computation of the probability of divergence

Probability of divergence

$$P_d \approx 1 - P_c, \quad (30)$$

where

$$P_c = \prod_{n=0}^{N} \Pr\left\{\|\boldsymbol{X}(n)\|^2 \leq z_0(n) | y(n) = \hat{y}(n)\right\}$$

where $\|\boldsymbol{X}(n)\|^2/\sigma_x^2$ follows a χ^2 distribution with M degrees of freedom, and $z_0(n)$ is the only positive root of (if $\hat{y}(n) = 0$ and $\sigma_0^2 = 0$, $P_c = 1$)

$$Q(z) = 6\sigma_0^2 + \left(\frac{6}{M+2}\hat{y}(n) - 15\mu\psi_0^4\right)z - \frac{45\mu\sigma_0^2}{M+2}\hat{y}(n)z^2 - \frac{15\mu}{(M+2)(M+4)}\hat{y}(n)^2z^3 \quad (31)$$

with $\sigma_0^2 = \mathrm{E}(e_0(n)^2)$, $\psi_0^4 = \mathrm{E}(e_0(n)^4)$, and $\hat{y}(n)$ is computed from the recursion

$$\hat{y}(n+1) = \left[1 - 3\mu\sigma_x^2\left(\sigma_0^2 + \sigma_x^2\hat{y}(n)\right)\right]^2\hat{y}(n)$$
$$\hat{y}(0) = y(0) = \mathrm{E}\|\boldsymbol{V}(0)\|^2$$

for $\mu < [3\sigma_x^2(\sigma_0^2 + \sigma_x^2\hat{y}(0))]^{-1}$.

IV. Simulations

In this section, we compare our estimates for the probability of divergence of LMF with the results of several experiments. We tested our estimates for the probability of divergence for Gaussian measurement noise and for Gaussian regressors of two types: truly independent vectors $\boldsymbol{X}(n)$ (referred to as "IND regressors" hereafter) and vectors $\boldsymbol{X}(n)$ formed from a tap-delay line (referred to as "TDL regressors"), as usual in adaptive filtering. The covariance matrix of $\boldsymbol{X}(n)$ was always $\sigma_x^2\boldsymbol{I}$ (i.e., in the independent case, $\mathrm{E}\{\boldsymbol{X}(n)\boldsymbol{X}(n-k)^T\} = 0$ for all $k \neq 0$, and in the TDL case, $\mathrm{E}\{\boldsymbol{X}(n)\boldsymbol{X}(n-k)^T\}$ is nonzero and constant on the kth lower diagonal and zero elsewhere.)

Our first example validates the procedure used for the simulations. We compare the probabilities of divergence as a function of the step-size obtained under different simulation conditions. Each curve in Fig. 2 was obtained running L independent realizations of LMF for N_{it} time-steps each. For this example, we used $M = 100$, $\sigma_x^2 = 1$, $\sigma_z^2 = 0.01$, and initial condition $y(0) = \|\boldsymbol{V}(0)\|^2 = 1$. We labeled a curve as "diverging" if $\|\boldsymbol{V}(N_{\mathrm{it}})\| \geq 10^{100}$, and computed the observed probability of divergence as $P_{\mathrm{d,o}} = (\text{Number of curves diverging})/L$. There are seven curves in Fig. 2. The two curves to the left were obtained using our theoretical model for $N_{\mathrm{it}} = 100$ (o) and 1000 (solid). The five curves to the right were obtained by simulation (observed probabilities $P_{\mathrm{d,o}}$). The rightmost curve $(+)$ was

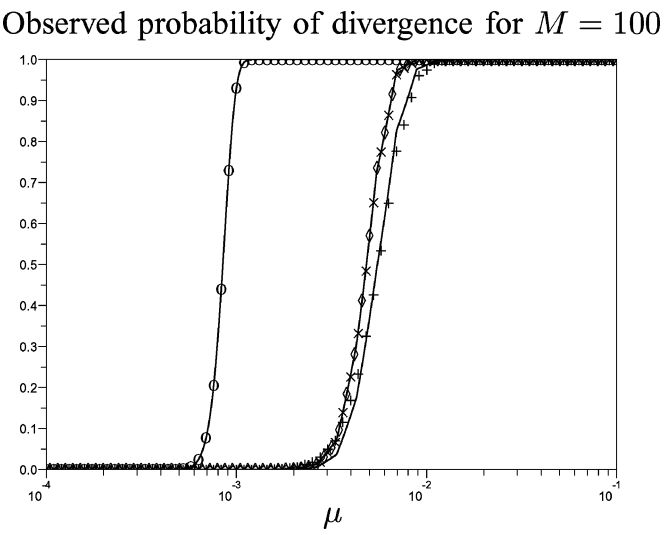

Fig. 2. Probability of divergence, Gaussian noise, $M = 100$, $\sigma_x^2 = 1$, $y(0) = 1$, $\sigma_0^2 = 0.01$. To the left are two superimposed curves obtained from our theoretical model, computed with 10^2 (o) and 10^3 (solid line) iterations. The rightmost curve, marked by $+$, was obtained from simulations, using TDL regressors, $N_{\text{it}} = L = 10^4$. The solid curve next to it was obtained with IND regressors, $N_{\text{it}} = 10^2$ and $L = 10^4$. In the middle are three superimposed curves, obtained with IND regressors and (a) $N_{\text{it}} = 10^3$ and $L = 10^4$ (solid line); (b) $N_{\text{it}} = L = 10^4$ ($\diamondsuit$); and (c) $N_{\text{it}} = 2.5 \times 10^4$ and $L = 10^3$ ($\times$).

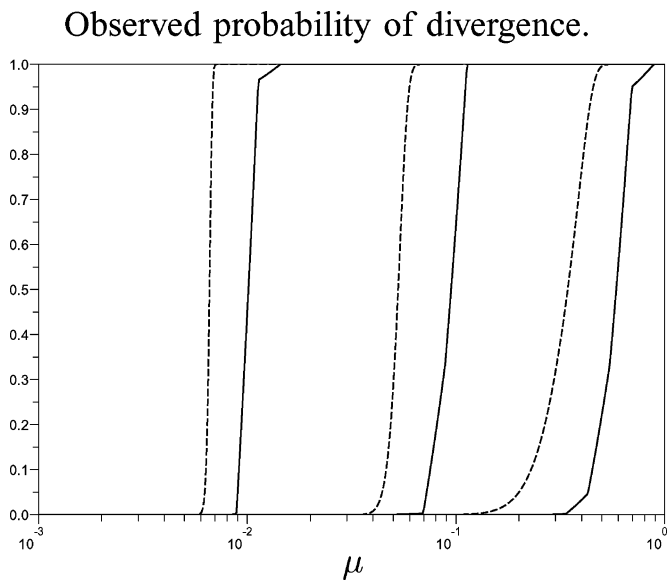

Fig. 3. Probability of divergence, TDL regressors, Gaussian noise, $\sigma_x^2 = 1$, $y(0) = 0.01$, $\sigma_0^2 = 0.01$, and $M = 10$, 100, and 1000. There is a pair of curves for each value of M. The solid curves are simulations, and the broken, the theoretical approximations. The leftmost pair is for $M = 10^3$, the center pair is for $M = 10^2$, and the rightmost is for $M = 10$. In all cases, $N_{\text{it}} = 10^4$ and $L = 10^3$.

obtained using TDL regressors. The remaining four simulation curves were obtained using IND regressors.

Looking at the simulation curves, note that there is basically no difference between the curves obtained with $L = 10^3$ and $L = 10^4$ [cases (a) and (b) in the figure]. Regarding the number of iterations, we notice a shift of the estimated curve (divergence for smaller μ) when N_{it} is increased from 10^2 to 10^3, but practically no variation for $L \geq 10^3$. Regarding the type of regressor, comparison of both curves obtained for $N_{\text{it}} = L = 10^4$ [the curve marked $(+)$ and the curve marked $\diamondsuit$ for case (b)], it is clear that the difference between the simulation results with IND and TDL regressors have little impact on the theoretical model's accuracy. The theoretical model indeed upper bounds the observed probability of divergence in both cases.

In Fig. 3, we used only TDL regressors. The solid (unbroken) curves are the results of simulations, with $y(0) = 0.01$ and $\sigma_x^2 = 1$, $\sigma_0^2 = 0.01$, $N_{\text{it}} = 10^4$, and $L = 10^3$; and the broken curves give our approximation P_d, using 10^3 steps for the iterations (the same theoretical results were obtained using 10^4 steps). Note that for smaller initial error $y(0)$ (smaller than $y(0) = 1$ used previously), our approximations are closer to the simulations.

Fig. 4 shows again results for TDL regressors, $M = 10^2$, $\sigma_x^2 = 1$, $N_{\text{it}} = 10^4$, and $L = 10^3$, but now with Gaussian noise with $\sigma_0^2 = 0.1$. The top curves are for $y(0) = 0.1$, and the bottom curves, for $y(0) = 1$. Again, the theoretical approximations are better for smaller $y(0)$.

Fig. 5 again shows simulations for TDL regressors and $M = 100$, now with different values for σ_x^2. In all cases, the solid lines are simulations made with $N_{\text{it}} = 10^4$, $L = 10^3$, and $\sigma_z^2 = 0.01$, and the broken lines are from the theoretical model. In (a), $y(0) = 0.1$ and $\sigma_x^2 = 2$, in (b) $y(0) = 0.1$ and $\sigma_x^2 = 4$, in (c) $y(0) = 1$ and $\sigma_x^2 = 0.5$, and in (d) $y(0) = 1$ and $\sigma_x^2 = 4$.

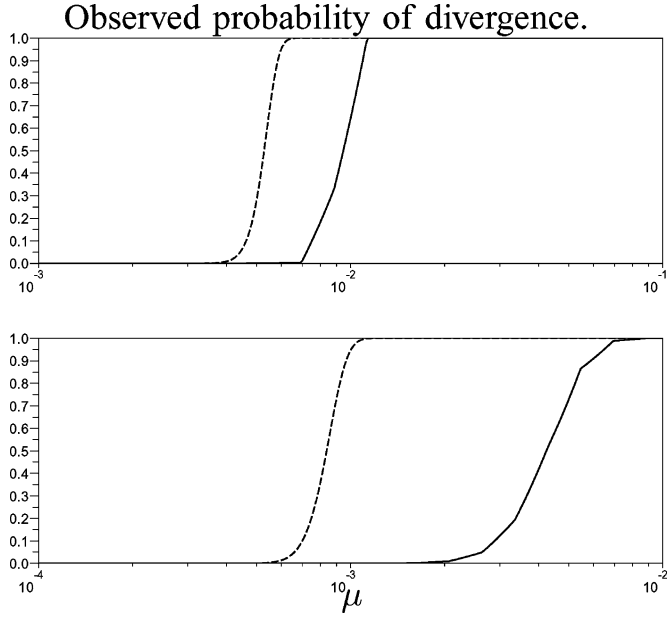

Fig. 4. Probability of divergence, TDL regressors, Gaussian noise, $\sigma_0^2 = 0.1$, $\sigma_x^2 = 1$, and $M = 100$. Top: $y(0) = 0.1$; bottom: $y(0) = 1$. In all cases, $N_{\text{it}} = 10^4$ and $L = 10^3$. The solid curves are simulations, the broken curves are from our theoretical model using 10^3 iterations.

The figures show that our estimates indeed upper bound the probability of divergence of LMF. We noticed that our approximation is closer to the observed probability of divergence for smaller values of the initial condition $y(0)$, and if the noise variance is not larger than the initial condition. We ran many other simulations, varying all possible parameters, always with similar results. Simulations for which $\sigma_0^2 \geq y(0)$ are more difficult

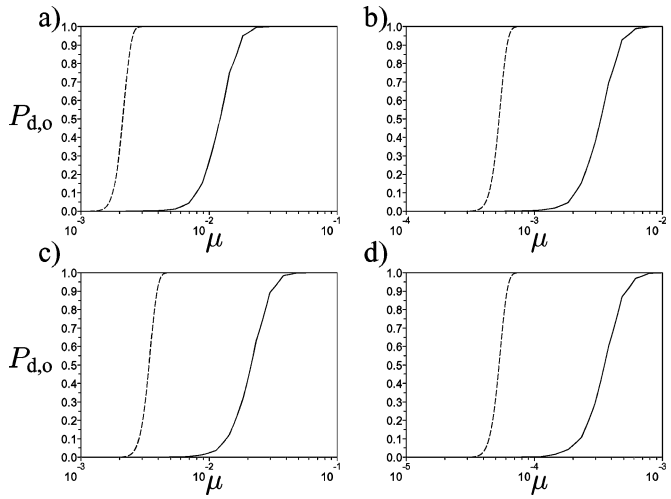

Fig. 5. Probability of divergence as a function of signal power, TDL regressors, Gaussian noise, $M = 100$. Solid lines: Simulations, $N_{\mathrm{it}} = 10^4$, $L = 10^3$, and $\sigma_z^2 = 0.01$. Broken lines: Theoretical model, computed with 10^3 iterations. (a) $y(0) = 0.1$ and $\sigma_x^2 = 2$; (b) $y(0) = 0.1$ and $\sigma_x^2 = 4$; (c) $y(0) = 1$ and $\sigma_x^2 = 0.5$; and (d) $y(0) = 1$ and $\sigma_x^2 = 4$.

TABLE II
MAXIMUM STEP-SIZES $\mu_{\max}$ FOR DIFFERENT NOISE VARIANCES AND FILTER LENGTHS, FOR GAUSSIAN REGRESSORS

σ_0^2	M	$\mu_{\max}$	$y(0)$
0.01	100	0.00370	1
0.01	10	0.7692	0.01
0.01	100	0.097087	0.01
0.01	1000	0.00997008	0.01
0.1	100	0.00313	1
0.1	100	0.009709	0.1

to perform, since in this case the number of time-steps necessary for stabilization of $P_{\mathrm{d,o}}$ tends to be too large, and the simulations, too lengthy. The same happens if σ_x^2 is decreased too much.

Our results may be compared with the bounds for the step-size given in [7]. These results are shown in Table II. Comparing the maximum step-sizes predicted by Table II with our simulations and the observed probability of divergence in Figs. 2–4, one can see that the maximum step-sizes given in [7] fall in the region where the observed probability of divergence is increasing rapidly.

V. CONCLUSION

In this paper we argued that the least-mean fourth algorithm cannot be mean-square stable when the regressor sequence is not strictly bounded, and provided an upper bound for the probability of divergence of the algorithm. In practice (since all actual regressor sequences are bounded), our result means that the algorithm is very sensitive to large values of the regressor sequence, even if they occur very rarely, as is the case for the Gaussian distribution. Other conclusion is that the step-size for the LMF algorithm should be chosen rather conservatively, mainly when a good initial guess for the Wiener solution is not available.

The behavior of the LMF algorithm in this respect is very different from that of LMS. If the weight error vector $\boldsymbol{V}(n)$ is taken by chance to a large value in a particular realization of the LMS algorithm, it tends to return quickly to reasonable behavior [10]. The LMF algorithm, on the other hand, may become completely unstable if the weight error vector becomes too large. This behavior is due to its cubic nonlinearity.

Our upper bound for the probability of divergence provides designers with tools to decide whether using the LMF algorithm in a particular situation is a sensible choice. Since step-sizes too close to the stability margin often lead to slow convergence and poor performance, the fact that the bound is not tight is not a major hindrance. Our results also open a new way of looking at adaptive filter behavior, which may lead to better ways of increasing algorithm robustness and performance.

APPENDIX I
AN APROXIMATION TO $(\boldsymbol{V}(n)^T\boldsymbol{X}(n))^{2k}$

In Section III-A, we approximated

$$\left(\boldsymbol{V}(n)^T\boldsymbol{X}(n)\right)^{2k} \approx \alpha_M(k)\,||\boldsymbol{V}(n)||^{2k}\,||\boldsymbol{X}(n)||^{2k}, \quad k=1,2,3. \tag{32}$$

Our choice of $\alpha_M(k)$ is as given in (22), $\alpha_M(k) = \mathrm{E}\{((\boldsymbol{V}(n)^T\boldsymbol{X}(n))/(||\boldsymbol{V}(n)||||\boldsymbol{X}(n)||))^{2k}\}$.

We now derive expressions for $\alpha_M(k)$ for $k = 1, 2, 3$, under the following assumptions.

1) The *vector* sequence $\{\boldsymbol{X}(n)\}$ is i.i.d.
2) The entries of $\boldsymbol{X}(n)$ are independent (and thus $\mathrm{E}\{\boldsymbol{X}(n)\boldsymbol{X}(n)^T\} = \sigma_x^2\boldsymbol{I}$).

From these assumptions and from (22), the following can be found.

1) The weight error vector $\boldsymbol{V}(n)$ is independent of $\boldsymbol{X}(n)$.
2) If $\boldsymbol{V}(n) = \boldsymbol{0}$ or $\boldsymbol{X}(n) = \boldsymbol{0}$, (32) holds for any finite $\alpha_M(k)$.
3) We can rewrite $\alpha_M(k)$ as

$$\alpha_M(k) = \mathrm{E}\left\{\mathrm{E}\left[\left(\frac{\boldsymbol{V}(n)^T\boldsymbol{X}(n)}{||\boldsymbol{V}(n)||\,||\boldsymbol{X}(n)||}\right)^{2k}\middle|\,\boldsymbol{V}(n)\right]\right\}.$$

4) Note that in the inner expectation $\boldsymbol{V}(n)$ is fixed. Since $\boldsymbol{X}(n)$ has independent entries, $\boldsymbol{X}(n)$ may point to any direction in $\mathbb{R}^M$ with equal probability, so the inner probability is independent of the particular direction taken by $\boldsymbol{V}(n)$. Therefore, the outer expectation is over a constant. Recalling assumption 2) and defining $\boldsymbol{e}_1 = [1\ 0 \ldots 0]^T \in \mathbb{R}^M$, $\bar{\boldsymbol{X}} \triangleq \boldsymbol{X}(n)/||\boldsymbol{X}(n)||$, we are left with

$$\alpha_M(k) = \mathrm{E}\left[\left(\boldsymbol{e}_1^T\bar{\boldsymbol{X}}\right)^{2k}\right] \tag{33}$$

where both $\boldsymbol{e}_1$ and $\bar{\boldsymbol{X}}$ have unit Euclidean length.

Before tackling the general solution for (33), let us consider the three-dimensional case. $\bar{\boldsymbol{X}}$ is a vector whose tip lies in the unit sphere, and the scalar product $\boldsymbol{e}_1^T\bar{\boldsymbol{X}}$ equals the cosine of the angle θ between $\bar{\boldsymbol{X}}$ and $\boldsymbol{e}_1$ (see Fig. 6).

All vectors $\bar{\boldsymbol{X}}$ with tips in the circle indicated in the figure have the same angle θ with respect to $\boldsymbol{e}_1$. The distance of any point in the circle to the $\boldsymbol{e}_1$ axis is $\sin(\theta)$ (since the length of $\bar{\boldsymbol{X}}$ is one, and $\sin(\theta) = \sqrt{1-\cos(\theta)^2}$ for $0 \le \theta \le \pi$). The

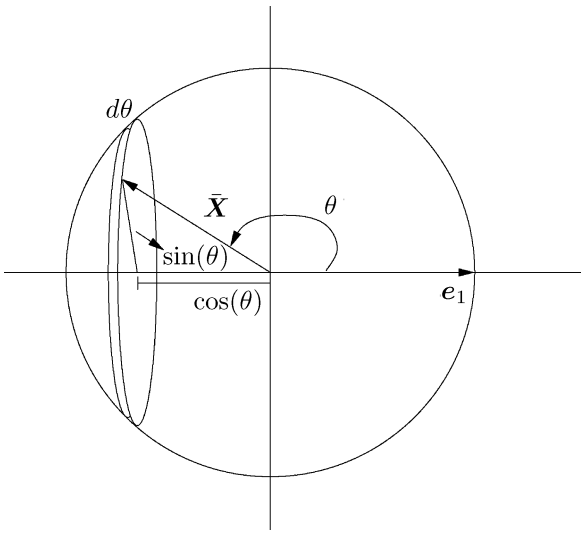

Fig. 6. Evaluation of $\alpha_3(k)$.

element of area for all $\bar{\boldsymbol{X}}$ with angle θ will then be $2\pi\sin(\theta)d\theta$. Thus, we may evaluate $\alpha_3(k)$ by the expression

$$\alpha_3(k) = \frac{\int_0^\pi \cos(\theta)^{2k} 2\pi \sin(\theta) d\theta}{\text{Area of the unit sphere}}.$$

The area of the unit sphere is $A_3 = 4\pi$. It is given by

$$A_3 = \int_0^\pi 2\pi \sin(\theta) d\theta.$$

The general M-dimensional case is similar. Fixing the angle θ between $\bar{\boldsymbol{X}}$ and $\boldsymbol{e}_1$, the unit length vector $\bar{\boldsymbol{X}}$ is constrained to a $M-1$-dimensional hyper-sphere of radius $\sin(\theta)$ (since we fixed the projection of $\bar{\boldsymbol{X}}$ with respect to $\boldsymbol{e}_1$ to $\cos(\theta)$, the other $M-1$ coordinates of $\bar{\boldsymbol{X}}$ must have total length $\sin(\theta) = \sqrt{1-\cos(\theta)^2}$, for $0 \le \theta \le \pi$). The expression for $\alpha_M(k)$ is then

$$\alpha_M(k) = \frac{\int_0^\pi \cos(\theta)^{2k} A_{M-1} \sin(\theta)^{M-2} d\theta}{\int_0^\pi A_{M-1} \sin(\theta)^{M-2} d\theta}$$

where A_{M-1} is the "area" of the surface of an $M-1$-dimensional hyper-sphere of radius 1, $A_{M-1}\sin(\theta)^{M-2}$ is the surface area for an $M-1$-dimensional hyper-sphere of radius $\sin(\theta)$, and the denominator is the area of the surface of the M-dimensional hyper-sphere of radius 1 (i.e., the denominator is A_M). Note that A_{M-1} may be canceled, so

$$\alpha_M(k) = \frac{\int_0^\pi \cos(\theta)^{2k} \sin(\theta)^{M-2} d\theta}{\int_0^\pi \sin(\theta)^{M-2} d\theta}. \tag{34}$$

We can further simplify this expression integrating the numerator by parts. Taking the case $k=1$, let $u = \cos(\theta)$ and $dv = \cos(\theta)\sin(\theta)^{M-2}d\theta$. We obtain $du = -\sin(\theta)d\theta$, $v = 1/(M-1)\sin(\theta)^{M-1}$, and, since $\cos(\theta)\sin(\theta)^{M-1}|_0^\pi = 0$

$$\alpha_M(1) = \frac{1}{M-1}\frac{\int_0^\pi \sin(\theta)^M d\theta}{\int_0^\pi \sin(\theta)^{M-2} d\theta}. \tag{35}$$

For $k = 2$ and 3 the integration by parts must be repeated twice and thrice, respectively, with the final result

$$\alpha_M(2) = \frac{3}{(M+1)(M-1)}\frac{\int_0^\pi \sin(\theta)^{M+2} d\theta}{\int_0^\pi \sin(\theta)^{M-2} d\theta} \tag{36}$$

$$\alpha_M(3) = \frac{15}{(M+3)(M+1)(M-1)}\frac{\int_0^\pi \sin(\theta)^{M+4} d\theta}{\int_0^\pi \sin(\theta)^{M-2} d\theta}. \tag{37}$$

We now need to evaluate $\int_0^\pi \sin(\theta)^n d\theta$ for any integer $n > 1$. This integral can be computed directly, or with the help of a book of tables, such as [13]. The result is

$$\int_0^\pi \sin(\theta)^n d\theta = \frac{1}{2^n}\binom{n}{\frac{n}{2}}\pi, \quad n \text{ even} \tag{38a}$$

$$\int_0^\pi \sin(\theta)^n d\theta = \sqrt{\pi}\cdot\frac{\Gamma\left(\frac{(n+1)}{2}\right)}{\Gamma\left(\frac{(n+2)}{2}\right)}, \quad n \text{ odd} \tag{38b}$$

where $\binom{n}{k} = (n!/k!(n-k)!)$. Using (38a) and (38b) we obtain (22) (which holds for all n.) Since $\alpha_1(k) = 1$ for all k, (22) holds for $M \ge 1$.

Appendix II
Proof That $D(n) \ge 0$

From (24), we have (for simplicity we omit the dependence on n in this section)

$$\begin{aligned} D &= D(y) \\ &= 1 - \mu\left(6\sigma_0^2 - 15\mu\psi_0^4\|\boldsymbol{X}\|^2\right)\frac{\|\boldsymbol{X}\|^2}{M} \\ &\quad - \mu\left(6 - 45\mu\sigma_0^2\|\boldsymbol{X}\|^2\right)\frac{\|\boldsymbol{X}\|^4}{M(M+2)}y \\ &\quad + 15\mu^2\frac{\|\boldsymbol{X}\|^8}{M(M+2)(M+4)}y^2. \end{aligned}$$

We prove now that $D(y) \ge 0$ for all $y \ge 0$. Our proof is as follows.

1) First, we show that $D(0) \ge 0$ always.
2) Next, we find the minimum $D_{\min}$ of $D(y)$, and find conditions for $D_{\min} < 0$.
3) We then show that when $D_{\min} < 0$, the value $y_{\min}$ at which the minimum is achieved is necessarily negative.
4) We conclude noticing that, since the coefficient of y^2 in $D(y)$ is positive, $D(y)$ is strictly growing for $y > y_{\min}$, so 1) and 3) imply that $D(y) \ge 0$ for $y \ge 0$ if $D_{\min} < 0$ (if $D_{\min} \ge 0$, there is nothing to prove).

We now prove 1)–3). Recalling that for any random variable x it holds that $0 \le \mathrm{E}(x^2 - \mathrm{E}(x^2))^2 = \mathrm{E}(x^4) - (\mathrm{E}(x^2))^2$, we have $\psi_0^4 \ge \sigma_0^4$; thus

$$\begin{aligned} D(0) &= 1 + 15\mu^2\psi_0^4\frac{\|\boldsymbol{X}\|^4}{M} - 6\mu\sigma_0^2\frac{\|\boldsymbol{X}\|^2}{M} \\ &> 1 + 15\mu^2\sigma_0^4\frac{\|\boldsymbol{X}\|^4}{M} - 6\mu\sigma_0^2\frac{\|\boldsymbol{X}\|^2}{M}. \end{aligned}$$

This last quantity may be rewritten as

$$D(0) \geq \left(1-3\mu\sigma_0^2\frac{\|\boldsymbol{X}\|^2}{M}\right)^2+15\mu^2\sigma_0^4\frac{\|\boldsymbol{X}\|^4}{M}-9\mu^2\sigma_0^4\frac{\|\boldsymbol{X}\|^4}{M^2}\geq 0.$$

This proves 1). For 2), consider first a generic second-degree polynomial $c+by+ay^2$, with $a>0$. Its minimum is achieved at

$$y_{\min}=\frac{-b}{2a}$$

and the minimum value is

$$c+b\frac{-b}{2a}+a\left(\frac{-b}{2a}\right)^2=c-\frac{b^2}{4a}.$$

Applying this result to $D(y)$, recalling that $\psi_0^4 \geq \sigma_0^4$, and defining $\beta=\mu\sigma_0^2\|\boldsymbol{X}\|^2$, we have

$$\begin{aligned}
y_{\min} &= \frac{2-15\mu\sigma_0^2\|\boldsymbol{X}\|^2}{10\mu\|\boldsymbol{X}\|^4}(M+4),\\
D_{\min} &= 1+15\mu^2\psi_0^4\frac{\|\boldsymbol{X}\|^4}{M}-6\mu\sigma_0^2\frac{\|\boldsymbol{X}\|^2}{M}\\
&\quad -3(M+4)\frac{4-60\mu\sigma_0^2\|\boldsymbol{X}\|^2+225\mu^2\sigma_0^4\|\boldsymbol{X}\|^4}{20M(M+2)}\\
&= 1-\frac{3}{5}\frac{M+4}{M(M+2)}\\
&\quad -\frac{60\left(2-\frac{5\beta\psi_0^4}{\sigma_0^4}\right)\beta(M+2)+(M+4)\beta(675\beta-180)}{20M(M+2)}\\
&\geq 1-\frac{3}{5}\frac{M+4}{M(M+2)}-\frac{(75M+420)\beta-(12M+96)}{4M(M+2)}\beta.
\end{aligned}$$

We now prove 3). $D_{\min}$ may be negative only if $(75\,\mathrm{M}+420)\mu\sigma_0^2\|\boldsymbol{X}\|^2-(12\,\mathrm{M}+96)>0$. Under this condition, $y_{\min}$ is bounded by

$$\begin{aligned}
y_{\min} &< \frac{M+4}{10\mu\|\boldsymbol{X}\|^4}\left(2-15\frac{12M+96}{75M+420}\right)\\
&= -\frac{M+4}{10\mu\|\boldsymbol{X}\|^4}\frac{2M+40}{5M+28}<0.
\end{aligned}$$

This concludes our proof.

REFERENCES

[1] E. Walach and B. Widrow, "The least mean fourth (LMF) adaptive algorithm and its family," *IEEE Trans. Inf. Theory*, vol. IT-30, no. 2, pp. 275–283, Mar. 1984.

[2] P. S. R. Diniz, *Adaptive Filtering: Algorithms and Practical Implementation*, 2nd ed. Norwell, MA: Kluwer, 2002.

[3] A. H. Sayed, *Fundamentals of Adaptive Filtering*. New York: Wiley-Interscience, 2003.

[4] S. Koike, "Stability conditions for adaptive algorithms with non-quadratic error criteria," in *Proc. EUSIPCO 2000*, Tampere, Finland, Sep. 2000, pp. 131–134.

[5] ——, "Analysis of the least mean fourth algorithm based on Gaussian distributed tap weights," in *Proc. IEEE Int. Symp. Intelligent Signal Processing Communication Systems (ISPACS 2001)*, Nashville, TN, Nov. 2001, pp. 20–25.

[6] P. I. Hübscher and J. C. M. Bermudez, "An improved statistical analysis of the least mean fourth (LMF) adaptive algorithm," *IEEE Trans. Signal Process.*, vol. 51, no. 3, pp. 664–671, Mar. 2003.

[7] P. I. Hbscher, V. H. Nascimento, and J. C. M. Bermudez, "New results on the stability analysis of the LMF (least mean fourth) adaptive algorithm," in *Proc. 2003 IEEE Int. Conf. Acoustics, Speech Signal Processing (ICASSP)*, vol. VI, 2003, pp. VI-369–VI-372.

[8] W. A. Sethares, "Adaptive algorithms with nonlinear data and error functions," *IEEE Trans. Signal Process.*, vol. 40, no. 9, pp. 2199–2206, Sep. 1992.

[9] O. Macchi and E. Eweda, "Second-order convergence analysis of stochastic adaptive linear filtering," *IEEE Trans. Autom. Control*, vol. AC-28, no. 1, pp. 76–85, Jan. 1983.

[10] V. H. Nascimento and A. H. Sayed, "On the learning mechanism of adaptive filters," *IEEE Trans. Signal Process.*, vol. 48, no. 6, pp. 1609–1625, Jun. 2000.

[11] V. H. Nascimento and J. C. M. Bermudez, "When is the least-mean fourth algorithm mean-square stable?," in *Proc. 2005 IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. IV, Mar. 2005, pp. 341–344.

[12] ——, "On the stability of the least-mean fourth (LMF) algorithm," in *Proc. Anais do XXI Simpósio Brasileiro de Telecomunicações*, 2004, pp. 1–5.

[13] I. S. Gradshteyn, I. M. Ryzhik, and A. Jeffrey, *Table of Integrals, Series, and Products—Corrected and Enlarged Edition*. New York: Academic, 1980.

Vítor H. Nascimento (M'90) was born in São Paulo, Brazil. He received the B.Sc. and M.Sc. degrees in electrical engineering from Escola Politécnica, University of São Paulo, São Paulo, Brazil, in 1989 and 1992, respectively, and the Ph.D. degree from the University of California, Los Angeles, in 1999.

From 1990 to 1994, he was a Lecturer at Escola Politécnica where he has been a faculty member since 1999. His research interests include adaptive filtering theory and applications, linear and nonlinear estimation, and applied linear algebra.

Dr. Nascimento is a recipient of the IEEE Signal Processing Society 2002 Best Paper Award and served as an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS from 2003 to 2005. He is currently an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING.

José Carlos M. Bermudez (M'85–SM'02) received the B.E.E. degree from the Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil, in 1978, the M.Sc. degree from the Electrical Engineering Graduate Program Office (COPPE)/UFRJ, in 1981, and the Ph.D. degree from Concordia University, Montreal, QC, Canada, in 1985, both in electrical engineering.

In 1985, he joined the Department of Electrical Engineering, Federal University of Santa Catarina (UFSC), Florianópolis, SC, Brazil, where he is currently a Professor of electrical engineering. In winter 1992, he was a Visiting Researcher with the Department of Electrical Engineering, Concordia University. In 1994, he was a Visiting Researcher with the Department of Electrical Engineering and Computer Science, University of California, Irvine (UCI). His research interest have involved analog signal processing using continuous-time and sampled-data systems. His recent research interests are in digital signal processing, including linear and nonlinear adaptive filtering, active noise and vibration control, echo cancellation, image processing, and speech processing.

Prof. Bermudez was a member of the Signal Processing Theory and Methods technical committee of the IEEE Signal Processing Society from 1998 to 2004. He served as an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING in the area of adaptive filtering from 1994 to 1996 and from 1999 to 2001.