Bias-Variance Decomposition of the Mean-Square Deviation of the LMS Algorithm: Transient and Steady-State Analysis

Daniel G. Tiglea[1*], Renato Candido[1†] and Magno T.M. Silva[1†]

[1*]Electronic Systems Engineering Department, University of São Paulo,
Avenida Prof. Luciano Gualberto, trav. 3, no. 158, São Paulo,
05508-010, São Paulo, Brazil.

*Corresponding author(s). E-mail(s): dtiglea@lcs.poli.usp.br;
Contributing authors: renatocan@lps.usp.br; magno@lps.usp.br;
†These authors contributed equally to this work.

Abstract

In this paper, we perform the bias-variance decomposition of the mean-square deviation of the least-mean-squares algorithm during both the transient and steady-state phases. Although this solution has been extensively studied, to the best of our knowledge, this type of analysis has not been done before explicitly in this manner. We analyze a wide range of scenarios, including cases in which the filter length is not equal to that of the optimal solution and situations in the presence of impulsive noise. The conclusions thus obtained provide novel insights into the inner workings of the algorithm, and are supported by simulations. Moreover, we conduct experiments with real-world data considering an acoustic echo cancellation application, which show that the theoretical model thus obtained may perform reasonably well even when many of the assumptions made in the analysis do not hold.

Keywords: Adaptive filtering, bias-variance decomposition, least-mean-squares, mean-square deviation, transient analysis

Acknowledgements. This work was supported by the Coordination of Superior Level Staff Improvement (CAPES) under Grant 88887.512247/2020-00 and Finance Code 001, by the National Council for Scientific and Technological Development

(CNPq) under Grants 303826/2022-3 and 404081/2023-1, and by the São Paulo Research Foundation (FAPESP) under Grant 2021/02063-6.

1 Introduction

Adaptive filtering algorithms have been widely applied in several signal processing applications, including channel equalization [44, 55], acoustic echo cancellation [10, 11], active noise control [40, 41], and biomedical engineering [28, 59]. Among these solutions, the least-mean-squares (LMS) algorithm is one of the most popular, due to its relative simplicity [13, 23, 46]. As a result, the LMS algorithm has been extensively studied in the literature (see, e.g., [6, 7, 13, 15, 22, 23, 37–39, 46, 49, 57] and their references).

In order to assess the performance of adaptive filtering algorithms, a commonly adopted metric is the mean-square deviation (MSD) [13, 23, 39, 46]. In this paper, we carry out the bias-variance decomposition [9, 14, 18, 35] of the MSD of the LMS algorithm during both the transient and steady-state phases. Despite the prolonged interest in adaptive filters, to the best of our knowledge, this is the first time in which explicit theoretical expressions are obtained for both terms along the iterations. By doing so, we can gather new qualitative insights into how the LMS algorithm works from the perspective of statistics. We also examine scenarios in which the filter length does not perfectly match that of the optimal solution. The analysis shows that, although there is an evident trade-off between the bias and variance terms in regards to the step size, their behavior in relation to the filter length is not as straightforward, assuming that the length of the optimal solution remains fixed. We also investigate the effects of impulsive noise, and show that the total noise variance is the key factor when it comes to the behavior of the bias and variance terms, regardless of the noise distribution. Furthermore, our analysis agrees with classical results, although obtained via a different route, and is supported by simulation results.

1.1 Relation with Other Works and Contributions

The fact that the MSD can generally be divided in two terms, one related to the bias, and another one to the variance of the estimates, has been noted before. However, the ways in which this notion has been explored in the literature are different from the direction considered in this paper. For example, bias-compensated versions of the LMS algorithm have been proposed in the literature [29, 50, 53], attracting significant attention in recent years and leading to the emergence of several state-of-the-art solutions [12, 26, 36, 45, 56]. In particular, the algorithm of [29] was analyzed in [42, 43]. However, these solutions are mostly concerned with the case in which the adaptive filter input signal is corrupted by additive noise. Under these circumstances, it can be shown that the optimal solution achieved by gradient descent and stochastic gradient descent algorithms with noisy input is biased in relation to the optimal solution that would be obtained with the "clean" input signal. Thus, these bias-compensated algorithms seek to modify the update rule of the conventional adaptive filters, such as the standard LMS, in order to make up for the presence of noise in the input. This differs from the situation considered here, in which we assume that the input signal of the adaptive filter is not corrupted by any type of noise. Instead, our goal is to understand how the bias of the estimates evolves along the iterations, as well as their variance. Moreover, we remark that the analyses presented in [42, 43] focus on the bias-compensated algorithm of [29], whereas ours is aimed at the transient and steady-state phases of the conventional LMS algorithm.

Another example in which the bias-variance trade-off is taken into account in the adaptive filtering literature is the algorithm proposed in [35]. In this work, the authors modify the LMS algorithm in order to deliberately introduce a bias in its estimate so as to obtain an overall smaller steady-state excess mean-square error (EMSE). Simply put, this bias is adapted along the iterations with the goal of minimizing the steady-state EMSE based on an analogy with the convex combination of adaptive filters [2, 4]. Besides proposing a new algorithm, in [35] the authors also analyze the steady-state performance of their solution. However, they do not analyze the evolution of the bias and variance components of the MSD of the conventional LMS algorithm along the iterations, as we do in this paper. Other works with similar approaches to that of [35] have been published [3, 17, 48], but in none of them the analysis presented in this paper is carried out.

Finally, it is worth noting that the notion of bias and variance has also been explored to propose variable step-size (VSS) algorithms. For instance, in [33], the authors propose a VSS algorithm for tracking the optimal solution in nonstationary environments, which results in a weighting vector lag. This technique is based on an analysis presented in [24], in which the steady-state MSD of the LMS algorithm is broken down into three terms in such environments: one due to the bias of the weighting vector lag, another one due to the presence of additive noise, and a third one associated with the lag variance. Thus, the goal of the VSS algorithm of [33] is to seek, in an adaptive manner, the optimal step size that can minimize the overall sum of these three terms.

Despite the differences between the works mentioned and the present paper, they all illustrate the potential gains arising from a deeper knowledge about the behavior of the bias and variance of the estimates of adaptive filters. Thus, we believe that the analysis presented in this work may be of interest to a wide range of researchers in the field. Next, we provide a list of the contributions of the paper.

- **New approaches and perspectives for the theoretical analysis of the LMS algorithm**. By carrying out the bias-variance decomposition of the MSD of the LMS algorithm along the iterations, we seek to offer a different approach than usual to examine its transient performance. The results obtained agree with existing results [13, 23, 38, 46], although obtained in a different manner, and thus enable a different perspective on why the MSD curves typically behave the way they do.
- **Connections with the machine learning field**. The bias-variance decomposition is oftentimes used in the machine learning field [9] due to the "bias-variance dilemma": more flexible models may be capable of capturing interesting and important trends in the data, but are comparatively more susceptible to over-fitting than more rigid models. Thus, when comparing these types of model, the former ones tend to present a low bias and a high variance, whereas the latter ones typically

present a high bias and lower variance [9]. Extending this type of analysis to adaptive filters may be useful to researchers in both fields, since it establishes yet another bridge between them, with many similarities and analogies that can be exploited in the future.

- **A helpful tool for the design of future solutions**. We discuss how the analysis presented may aid in the design of novel solutions, such as VSS algorithms [1, 5, 8, 20, 27, 33, 34, 52, 58], for example. We remark, however, that proposing such a solution is out of the scope of the present paper.
- **Impact of filter length mismatch**. We also investigate what occurs with the bias and variance of the MSD when the length of the adaptive filter is not equal to that of the unknown system that we wish to estimate. It is well known that in such cases the performance is degraded in comparison with the scenario in which they match perfectly. The analysis presented provides an explanation for this phenomenon in terms of the bias and variance of the estimates. Moreover, we believe that this could support the proposal of novel variable tap-length adaptive filters in the future [30–32].
- **Effects of impulsive noise**. The analysis presented also holds for scenarios in the presence of impulsive noise, which oftentimes occur in practical applications.
- **Simulations with real-world data**. We validate the theoretical model in a wide range of scenarios, including one with real-world speech signals as the input to the adaptive filter in an acoustic echo cancellation (AEC) application. We believe that this further adds value to the present paper, given the relevance of AEC in the literature as well as in practical applications [10, 11, 13, 23, 46].

1.2 Organization of the Paper and Notation

The remainder of the paper is organized as follows. In Sec. 2, we present the problem formulation. In Sec. 3, we perform the bias-variance decomposition of the MSD of the LMS algorithm. The simulation results are presented in Sec. 4, and Sec. 5 closes the paper with the main conclusions.

We use normal font letters for scalars, boldface lowercase letters for vectors, and boldface uppercase letters for matrices. Moreover, $(\cdot)^{\mathrm{T}}$ denotes transposition, $\mathbf{I}_L$ the $L \times L$ identity matrix, $\mathbf{0}_L$ an L-length vector of zeros, $\mathbf{0}_{M\times L}$ an $M \times L$ matrix of zeros, $\log(\cdot)$ the natural logarithm, $\lfloor\cdot\rceil$ the rounding to the nearest integer, $\mathrm{E}\{\cdot\}$ the mathematical expectation, $|\cdot|$ the absolute value, $\mathrm{Tr}[\cdot]$ the trace of a matrix, $\mathcal{U}(a, b)$ a uniform distribution in the range $[a, b]$, and $\|\cdot\|$ the Euclidean norm. To simplify the arguments, we assume real data throughout the paper.

2 Problem Formulation

Let us consider an M-tap adaptive filter with a finite impulse response (FIR) structure, input signal $u(n)$, and desired signal $d(n)$ given by

$$d(n) = \mathbf{u}_L^{\mathrm{T}}(n)\mathbf{w}^{\mathrm{o}} + v(n), \tag{1}$$

where $\mathbf{w}^{\mathrm{o}} = [w_1^{\mathrm{o}}\ w_2^{\mathrm{o}}\ \cdots\ w_L^{\mathrm{o}}]$ is an L-length vector that represents an unknown system to be estimated, $\mathbf{u}_L(n) = [u(n)\ u(n-1)\ \cdots\ u(n-L+1)]^{\mathrm{T}}$ is the input regressor vector, and $v(n)$ is the measurement noise [23, 46]. The vector $\mathbf{w}^{\mathrm{o}}$ is oftentimes referred to as the "optimal solution" in the adaptive filtering literature [13, 23, 46]. Let us denote the estimate of $\mathbf{w}^{\mathrm{o}}$ produced by the algorithm at time instant n by the M-length column vector $\mathbf{w}(n)$. Then, the update equation of the LMS algorithm is given by [23, 46]

$$\mathbf{w}(n) = \mathbf{w}(n-1) + \mu \mathbf{u}_M(n) e(n), \tag{2}$$

where $\mathbf{u}_M(n) = [u(n)\ u(n-1)\ \cdots\ u(n-M+1)]^{\mathrm{T}}$,

$$e(n) = d(n) - \mathbf{u}_M^{\mathrm{T}}(n)\mathbf{w}(n-1) \tag{3}$$

is the estimation error, and $\mu > 0$ is a step size [23, 46].

A commonly adopted performance indicator for adaptive filters is the MSD. Assuming that $M = L$, its expression at a certain iteration n is given by

$$\mathrm{MSD}(n) \triangleq \mathrm{E}\{\|\widetilde{\mathbf{w}}(n)\|^2\}, \tag{4}$$

in which

$$\widetilde{\mathbf{w}}(n) \triangleq \mathbf{w}^{\mathrm{o}} - \mathbf{w}(n) \tag{5}$$

is the weight vector error of the algorithm [23, 46].

If $M < L$, Eq. (5) can be adapted by making

$$\widetilde{\mathbf{w}}(n) \triangleq \mathbf{w}^{\mathrm{o}} - \boldsymbol{\omega}(n) \tag{6}$$

where $\boldsymbol{\omega}(n)$ is obtained by applying zero padding to $\mathbf{w}(n)$ so as to obtain an L-length vector, i.e., $\boldsymbol{\omega}(n) = [\mathbf{w}(n)\ \mathbf{0}_{L-M}]^{\mathrm{T}}$. In contrast, if $M > L$, we need to apply zero padding to the vector $\mathbf{w}^{\mathrm{o}}$. Denoting the resulting vector by $\boldsymbol{\omega}^{\mathrm{o}} = [\mathbf{w}^{\mathrm{o}}\ \mathbf{0}_{M-L}]$, (5) can be recast as

$$\widetilde{\mathbf{w}}(n) \triangleq \boldsymbol{\omega}^{\mathrm{o}} - \mathbf{w}(n). \tag{7}$$

3 Bias-Variance Decomposition

Let us denote the covariance matrix of $\widetilde{\mathbf{w}}(n)$ by

$$\mathbf{C}(n) \triangleq \mathrm{E}\left\{[\widetilde{\mathbf{w}}(n) - \mathrm{E}\{\widetilde{\mathbf{w}}(n)\}][\widetilde{\mathbf{w}}(n) - \mathrm{E}\{\widetilde{\mathbf{w}}(n)\}]^{\mathrm{T}}\right\}. \tag{8}$$

Applying the bias-variance decomposition for vectors to the MSD of (4), it can be recast as [35]

$$\boxed{\mathrm{E}\left\{\|\widetilde{\mathbf{w}}(n)\|^2\right\} = \|\underbrace{\mathrm{E}\{\widetilde{\mathbf{w}}(n)\}}_{\text{bias}}\|^2 + \mathrm{Tr}[\underbrace{\mathbf{C}(n)}_{\text{covariance matrix}}].} \tag{9}$$

This result is obtained in the Appendix at the end of this paper. The transient analysis of the MSD of the LMS algorithm has been performed, e.g., in [23, 39, 46]. However, the impact of each term in the right-hand side (rhs) of (9) has not been analyzed individually. Next, we study how each of them evolves separately over time. Throughout this paper we adopt the following assumptions:

A1. $v(n)$ is independent and identically distributed (iid) along the iterations with zero mean and variance σ_v^2, and is independent from any other signal;
A2. $\widetilde{\mathbf{w}}(n-1)$ is statistically independent from $\mathbf{u}_M(n)$;
A3. The input signal $u(n)$ is white Gaussian, which leads to $\mathbf{R}_L \triangleq \mathrm{E}\{\mathbf{u}_L(n)\mathbf{u}_L^{\mathrm{T}}(n)\} = \sigma_u^2\mathbf{I}_L$ and $\mathbf{R}_M \triangleq \mathrm{E}\{\mathbf{u}_M(n)\mathbf{u}_M^{\mathrm{T}}(n)\} = \sigma_u^2\mathbf{I}_M$, where σ_u^2 is the power of the input signal;
A4. $\mathbf{w}(0)$ is initialized as a vector of zeros, i.e., $\mathbf{w}(0) = \mathbf{0}_M$.

Assumptions **A1** and **A2** are common in the adaptive filtering literature, with the latter being known as the independence theory [23, 46]. In its turn, Assumption **A3** shall greatly simplify the arguments henceforth. Finally, Assumption **A4** corresponds to a usual practice in the adaptive filtering field. It is worth noting that we are only assuming Gaussianity for the input signal $u(n)$, not for the measurement noise $v(n)$. As a result, so long as Assumption **A1** holds, the analysis remains valid regardless of the distribution of $v(n)$, including scenarios with impulsive noise. We remark, however, that even when the input signal is not Gaussian or not wide-sense stationary, our theoretical model can work reasonably well, as will be shown in Sec. 4.

Next, we divide our analysis according to the three possible relations between M and L. In Sec. 3.1, we examine the case in which $M = L$. In Secs. 3.2 and 3.3, we extend the analysis to the case in which $M < L$ and $M > L$, respectively. Finally, in Sec. 3.4, we analyze the effects of the existence of impulsive noise on the theoretical models obtained.

3.1 Case 1: $M = L$

We remark that in this case, the vectors $\mathbf{u}_L(n)$ and $\mathbf{u}_M(n)$ coincide. Thus, in this section we make no distinction between them, and denote both of them by $\mathbf{u}(n)$ to simplify the notation. The same reasoning is applied to the matrices $\mathbf{R}_L = \mathbf{R}_M = \mathbf{R}$.

Let us begin by examining the term related to the bias. Subtracting both sides of (2) from $\mathbf{w}^{\mathrm{o}}$, and replacing (1) and (3) in the resulting equation, after some algebraic manipulations, we can see that the weight vector error of the LMS algorithm evolves according to

$$\widetilde{\mathbf{w}}(n) = [\mathbf{I}_M - \mu\mathbf{u}(n)\mathbf{u}^{\mathrm{T}}(n)]\widetilde{\mathbf{w}}(n-1) - \mu\mathbf{u}(n)v(n). \tag{10}$$

Under Assumptions **A1** – **A3**, taking the expectations from both sides of (10), we obtain

$$\mathrm{E}\{\widetilde{\mathbf{w}}(n)\} = (1-\mu\sigma_u^2)\mathrm{E}\{\widetilde{\mathbf{w}}(n-1)\}. \tag{11}$$

As, from **A4**, $\widetilde{\mathbf{w}}(0)=\mathbf{w}^{\mathrm{o}}$, the recursive application of (11) yields

$$\mathrm{E}\{\widetilde{\mathbf{w}}(n)\} = (1-\mu\sigma_u^2)^n\mathbf{w}^{\mathrm{o}}. \tag{12}$$

Squaring the Euclidean norm of (12), we obtain

$$\boxed{\|\mathrm{E}\{\widetilde{\mathbf{w}}(n)\}\|^2 = (1-\mu\sigma_u^2)^{2n}\|\mathbf{w}^{\mathrm{o}}\|^2,} \tag{13}$$

which determines how the term related to the bias in (9) evolves. Although $\mathbf{w}^{\mathrm{o}}$ appears in (13), we do not need to know it beforehand. Instead, only the knowledge of its norm is required. In the adaptive filtering literature, it is not uncommon to assume that $\|\mathbf{w}^{\mathrm{o}}\|^2 = 1$, which can be achieved by employing automatic gain control (see, e.g., [23]).

Let us now resume the analysis of the the trace of the covariance matrix $\mathbf{C}(n)$. From Eq. (A3), located in the Appendix, we can see that

$$\mathrm{Tr}\,[\mathbf{C}(n)] = \mathrm{Tr}[\overline{\mathbf{C}}(n)] - \|\mathrm{E}\{\widetilde{\mathbf{w}}(n)\}\|^2, \tag{14}$$

where we have introduced

$$\overline{\mathbf{C}}(n) \triangleq \mathrm{E}\{\widetilde{\mathbf{w}}(n)\widetilde{\mathbf{w}}^{\mathrm{T}}(n)\}. \tag{15}$$

From (10), under Assumption **A1**, we can write [39]

$$\begin{aligned}\overline{\mathbf{C}}(n) = \; & \mathrm{E}\{\widetilde{\mathbf{w}}(n-1)\widetilde{\mathbf{w}}^{\mathrm{T}}(n-1)\} - \mu\mathrm{E}\{\mathbf{u}(n)\mathbf{u}^{\mathrm{T}}(n)\widetilde{\mathbf{w}}(n-1)\widetilde{\mathbf{w}}^{\mathrm{T}}(n-1)\}\\ & - \mu\mathrm{E}\{\widetilde{\mathbf{w}}(n-1)\widetilde{\mathbf{w}}^{\mathrm{T}}(n-1)\mathbf{u}(n)\mathbf{u}^{\mathrm{T}}(n)\}\\ & + \mu^2\mathrm{E}\{\mathbf{u}(n)\mathbf{u}^{\mathrm{T}}(n)\widetilde{\mathbf{w}}(n-1)\widetilde{\mathbf{w}}^{\mathrm{T}}(n-1)\mathbf{u}(n)\mathbf{u}^{\mathrm{T}}(n)\}\\ & + \mu^2\mathrm{E}\{\mathbf{u}(n)\mathbf{u}^{\mathrm{T}}(n)v^2(n)\}.\end{aligned} \tag{16}$$

Moreover, under **A1** – **A3**, it can be shown that the fourth-order term in (16) is given by [39]

$$\mathrm{E}\{\mathbf{u}(n)\mathbf{u}^{\mathrm{T}}(n)\widetilde{\mathbf{w}}(n-1)\widetilde{\mathbf{w}}^{\mathrm{T}}(n-1)\mathbf{u}(n)\mathbf{u}^{\mathrm{T}}(n)\} = \sigma_u^4\mathrm{Tr}\{\overline{\mathbf{C}}(n-1)\}\mathbf{I}_M + 2\sigma_u^4\overline{\mathbf{C}}(n-1). \tag{17}$$

Replacing (17) in (16), we arrive at [39]

$$\overline{\mathbf{C}}(n) = \overline{\mathbf{C}}(n-1) - 2\mu\sigma_u^2\overline{\mathbf{C}}(n-1) + \mu^2[\sigma_u^4\mathrm{Tr}\{\overline{\mathbf{C}}(n-1)\}\mathbf{I}_M + 2\sigma_u^4\overline{\mathbf{C}}(n-1) + \sigma_v^2\sigma_u^2\mathbf{I}_M]. \tag{18}$$

Defining $\bar{c}(n) \triangleq \mathrm{Tr}[\overline{\mathbf{C}}(n)]$ and taking the trace of both sides in (18), we obtain

$$\bar{c}(n) = \alpha\bar{c}(n-1) + \mu^2 M\sigma_u^2\sigma_v^2, \tag{19}$$

where we have introduced

$$\alpha \triangleq 1 - 2\mu\sigma_u^2 + \mu^2(M+2)\sigma_u^4 \tag{20}$$

for convenience. Observing that, under **A4**, $\bar{c}(0) = \|\mathbf{w}^{\mathrm{o}}\|^2$, replacing this result in (19), and applying it recursively, after some algebraic manipulations, we arrive at

$$\bar{c}(n) = \alpha^n\|\mathbf{w}^{\mathrm{o}}\|^2 + \mu^2 M\sigma_u^2\sigma_v^2 \cdot \left(\frac{1-\alpha^n}{1-\alpha}\right). \tag{21}$$

Replacing (20) in the denominator of the second term in the rhs of (21), we obtain

$$\bar{c}(n) = \alpha^n\|\mathbf{w}^{\mathrm{o}}\|^2 + \chi(1-\alpha^n), \tag{22}$$

where

$$\chi \triangleq \frac{\mu M \sigma_v^2}{2-\mu(M+2)\sigma_u^2}. \tag{23}$$

Using (22) and (13) in (14), after some algebra, we finally arrive at

$$\boxed{\mathrm{Tr}[\mathbf{C}(n)] = \|\mathbf{w}^{\mathrm{o}}\|^2[\alpha^n - (1-\mu\sigma_u^2)^{2n}] + \chi(1-\alpha^n).} \tag{24}$$

Examining (13) and (24), we arrive at some interesting conclusions. Firstly, at the iteration $n = 0$, the norm of the bias, given by (13), coincides with that of the optimal solution, and then decays exponentially, so long as $0 < \mu < \frac{2}{\sigma_u^2}$. As a result, the contribution of the norm of the bias should be more significant during the first iterations of the transient phase, and vanish in steady state. In contrast, if we replace $n = 0$ in (24), we can see that the trace of the covariance matrix is initially zero. This is reasonable, since, from **A4**, $\widetilde{\mathbf{w}}(0) = \mathbf{w}^{\mathrm{o}}$ is deterministic. Hence, under typical circumstances, the MSD is predominantly determined by the norm of the bias in the first iterations of the transient phase. However, as n grows larger, $\mathrm{Tr}[\mathbf{C}(n)]$ typically begins to increase, up to a certain point. Differentiating (24) with respect to n and setting the resulting equation to zero, we can conclude that the trace of the covariance matrix reaches its peak around the iteration

$$\boxed{n_p = \left\lfloor \frac{\log\left(\dfrac{\|\mathbf{w}^{\mathrm{o}}\|^2}{\|\mathbf{w}^{\mathrm{o}}\|^2-\chi}\right) + \log\left[\dfrac{2\log(1-\mu\sigma_u^2)}{\log(\alpha)}\right]}{\log(\alpha) - 2\log(1-\mu\sigma_u^2)} \right\rfloor}, \tag{25}$$

so long as $\chi < \|\mathbf{w}^{\mathrm{o}}\|^2$. Otherwise, simulations suggest that $\mathrm{Tr}[\mathbf{C}(n)]$ typically increases and stabilizes without any noticeable peaks. We remark that (25) is only valid as an approximation, since by differentiating Eq. (24) with respect to n we are treating n as a continuous variable, which it is not. Regardless, as we shall see in Sec. 4, this approach leads to satisfactory results. For typical values of μ, M, σ_v^2, σ_u^2, and $\|\mathbf{w}^{\mathrm{o}}\|^2$, in the absence of impulsive noise, n_p corresponds to some point during the transient, as will become clear in Sec. 4. Intuitively, the peak of the variance of the estimates produced by the algorithm occurs at this time because they are sufficiently far from the initial guess $\mathbf{w}(0) = \mathbf{0}_M$, but have not fully converged in the mean to $\mathbf{w}^{\mathrm{o}}$ yet. Thus, assuming that $|\alpha| < 1$, which occurs if

$$0 < \mu < \mu_{\max} = \frac{2}{(M+2)\sigma_u^2}, \tag{26}$$

the trace of the covariance matrix begins to decrease for $n > n_p$, until it stabilizes at

$$\lim_{n\to\infty} \mathrm{Tr}[\mathbf{C}(n)] = \chi. \tag{27}$$

This result is best known in the adaptive filtering literature as the steady-state MSD of the LMS algorithm [23, 46], i.e.,

$$\lim_{n\to\infty} \mathrm{MSD}(n) = \chi. \tag{28}$$

Therefore, contrary to what happens in the beginning of the transient, the MSD is dominated during the steady state by the trace of the covariance matrix, as the norm of the bias should be negligible at this point. Evidently, there is a time instant n_s at which the two terms switch places in terms of relevance. Equating the rhs of (13) and (24), we can see that n_s satisfies

$$\boxed{\|\mathbf{w}^{\mathrm{o}}\|^2[2(1-\mu\sigma_u^2)^{2n_s} - \alpha^{n_s}]+\chi\alpha^{n_s}=\chi.} \tag{29}$$

There is no closed-form solution for (29), but the value of n_s may be evaluated numerically from this equation. After that, it must be rounded to the nearest integer.

We can also analyze the impact of the step size μ on the norm of the bias and on the trace of the covariance matrix individually. In the adaptive filtering literature, it is well known that large values of μ lead to a faster convergence rate, but deteriorate the steady-state MSD [13, 23, 39, 46]. In contrast, if μ is small, the steady-state MSD decreases, but the convergence rate slows down. From our analysis, we can see that this trade-off is related to evolution of the bias and the variance. It is clear from (13) that the smaller the step size, the slower the exponential decay of the norm of the bias. However, we notice from (24) that the adoption of smaller step sizes also decreases the variance of the estimates in steady state, which is reasonable, and also its peak value at $n = n_p$. Conversely, higher values for μ lead to a fast reduction in the norm of the bias, but overall increase the variance of the estimates in steady state and at $n = n_p$. This is in accordance with observations made in, e.g., [51], where it was pointed out that oftentimes the bias and the variance display opposite behaviors with respect to parameters in statistical and adaptive signal processing. As for the impact of the filter length, we can see that, for a fixed step size μ, and maintaining the assumption that $M = L$, the value of M does not impact the rate at which the norm of the bias decays in (13), so long as (26) holds. In contrast, the trace of the covariance matrix increases with M. This implies that n_s should be affected by both the step size and by the filter length. Overall, n_s decreases with the increase of μ and M, and can be significantly affected by them. This is illustrated in Tab. 1, in which we show theoretical values obtained for n_s from (29), considering $\sigma_u^2 = 1$, $\sigma_v^2 = 0.01$, and different values for μ and M. Although out of the scope for this paper, we believe that this information can aid in the future design of VSS algorithms [1, 5, 8, 20, 27, 34, 52, 58]. For instance, in [8], a VSS-normalized LMS algorithm was proposed that switches the step size from a greater value to a smaller one after a certain number of iterations, based on the predicted MSD. Using (29) and the analysis presented, a similar idea could be employed, in order to gradually reduce the step size by switching it when the norm of the bias is surpassed by the trace of the covariance matrix. Lastly, in Secs. 3.2 and 3.3, it will be shown that, if $M \neq L$, the behavior of the variance and of the bias with respect to M may not be so obvious, and depends on the characteristics of the optimal solution.

Table 1: Values of n_s obtained from (29) for $\sigma_u^2 = 1$ and $\sigma_v^2 = 0.01$, and different values for μ and M.

M μ	10	20	30	40	50
0.01	344	266	203	159	130
0.005	796	694	614	541	473
0.001	4917	4545	4315	4143	4003

3.2 Case 2: $M < L$

To enable the comparison with $\mathbf{w}^{\mathrm{o}}$, in this section we shall use the L-length weight vector $\boldsymbol{\omega}(n) = [\mathbf{w}(n)\ \mathbf{0}_{L-M}]^{\mathrm{T}}$ described in Sec. 2. In this case, the update equation of the LMS algorithm may be recast as

$$\boldsymbol{\omega}(n) = \boldsymbol{\omega}(n-1) + \boldsymbol{\mathcal{S}}\mathbf{u}_L(n)\left[d(n) - \mathbf{u}_L^{\mathrm{T}}(n)\boldsymbol{\omega}(n-1)\right], \tag{30}$$

where $\boldsymbol{\mathcal{S}}$ is a block diagonal matrix given by

$$\boldsymbol{\mathcal{S}} = \begin{bmatrix} \mu\mathbf{I}_M & \mathbf{0}_{M\times\Delta_L} \\ \mathbf{0}_{\Delta_L\times M} & \mathbf{0}_{\Delta_L\times\Delta_L} \end{bmatrix} \tag{31}$$

with $\Delta_L \triangleq L - M$.

Subtracting both sides of (30) from $\mathbf{w}^{\mathrm{o}}$, we obtain after some manipulations

$$\widetilde{\mathbf{w}}(n) = [\mathbf{I}_L - \boldsymbol{\mathcal{S}}\mathbf{u}_L(n)\mathbf{u}_L^{\mathrm{T}}(n)]\widetilde{\mathbf{w}}(n-1) - \boldsymbol{\mathcal{S}}\mathbf{u}_L(n)v(n). \tag{32}$$

We begin by investigating the bias term. Under Assumptions **A1**–**A3**, by taking the expectations from both sides of (32) we obtain

$$\mathrm{E}\{\widetilde{\mathbf{w}}(n)\} = [\mathbf{I}_L - \sigma_u^2\boldsymbol{\mathcal{S}}]\widetilde{\mathbf{w}}(n-1). \tag{33}$$

Defining

$$\boldsymbol{\mathcal{A}} \triangleq \mathbf{I}_L - \sigma_u^2\boldsymbol{\mathcal{S}} = \begin{bmatrix} (1-\mu\sigma_u^2)\mathbf{I}_M & \mathbf{0}_{M\times\Delta_L} \\ \mathbf{0}_{\Delta_L\times M} & \mathbf{I}_{\Delta_L} \end{bmatrix}, \tag{34}$$

under Assumption **A4** we may obtain from (33)

$$\mathrm{E}\{\widetilde{\mathbf{w}}(n)\} = \boldsymbol{\mathcal{A}}^n\mathbf{w}^{\mathrm{o}} = \begin{bmatrix} (1-\mu\sigma_u^2)^n\mathbf{I}_M & \mathbf{0}_{M\times\Delta_L} \\ \mathbf{0}_{\Delta_L\times M} & \mathbf{I}_{\Delta_L} \end{bmatrix}\mathbf{w}^{\mathrm{o}}, \tag{35}$$

where we took advantage of the fact that $\boldsymbol{\mathcal{A}}$ is a diagonal matrix.

Introducing $\mathbf{w}_1^{\mathrm{o}} \triangleq [w_1^{\mathrm{o}} \; w_2^{\mathrm{o}} \; \cdots \; w_M^{\mathrm{o}}]^{\mathrm{T}}$ and $\mathbf{w}_2^{\mathrm{o}} \triangleq [w_{M+1}^{\mathrm{o}} \; w_{M+2}^{\mathrm{o}} \; \cdots \; w_L^{\mathrm{o}}]^{\mathrm{T}}$, we can see from (35) that

$$\boxed{\|\mathrm{E}\{\widetilde{\mathbf{w}}(n)\}\|^2 = (1-\mu\sigma_u^2)^{2n}\|\mathbf{w}_1^{\mathrm{o}}\|^2 + \|\mathbf{w}_2^{\mathrm{o}}\|^2.} \tag{36}$$

Thus, different from the case in which the length of the filter matches that of the optimal solution, if $M < L$ the bias does not vanish as $n \to \infty$, but rather converges to $\|\mathbf{w}_2^{\mathrm{o}}\|^2$. This is reasonable, as the filter cannot properly identify the elements of $\mathbf{w}_2^{\mathrm{o}}$.

As for the trace of the covariance matrix, given by (14), we can follow an analogous line of thought to the one adopted in Sec. 3.1. Thus, we obtain from (15)

$$\begin{aligned}\overline{\mathbf{C}}(n) = {} & \mathrm{E}\{\widetilde{\mathbf{w}}(n-1)\widetilde{\mathbf{w}}^{\mathrm{T}}(n-1)\} - \boldsymbol{\mathcal{S}}\mathrm{E}\{\mathbf{u}_L(n)\mathbf{u}_L^{\mathrm{T}}(n)\widetilde{\mathbf{w}}(n-1)\widetilde{\mathbf{w}}^{\mathrm{T}}(n-1)\} \\ & - \mathrm{E}\{\widetilde{\mathbf{w}}(n-1)\widetilde{\mathbf{w}}^{\mathrm{T}}(n-1)\mathbf{u}_L(n)\mathbf{u}_L^{\mathrm{T}}(n)\}\boldsymbol{\mathcal{S}}^{\mathrm{T}} \\ & + \boldsymbol{\mathcal{S}}\mathrm{E}\{\mathbf{u}_L(n)\mathbf{u}_L^{\mathrm{T}}(n)\widetilde{\mathbf{w}}(n-1)\widetilde{\mathbf{w}}^{\mathrm{T}}(n-1)\mathbf{u}_L(n)\mathbf{u}_L^{\mathrm{T}}(n)\}\boldsymbol{\mathcal{S}}^{\mathrm{T}} \\ & + \boldsymbol{\mathcal{S}}\mathrm{E}\{\mathbf{u}_L(n)\mathbf{u}_L^{\mathrm{T}}(n)v^2(n)\}\boldsymbol{\mathcal{S}}^{\mathrm{T}},\end{aligned} \tag{37}$$

where we used the fact that $\boldsymbol{\mathcal{S}}$ is deterministic. Analogously to (17), in this case we obtain

$$\mathrm{E}\{\mathbf{u}_L(n)\mathbf{u}_L^{\mathrm{T}}(n)\widetilde{\mathbf{w}}(n-1)\widetilde{\mathbf{w}}^{\mathrm{T}}(n-1)\mathbf{u}_L(n)\mathbf{u}_L^{\mathrm{T}}(n)\} = \sigma_u^4\mathrm{Tr}\{\overline{\mathbf{C}}(n-1)\}\mathbf{I}_L + 2\sigma_u^4\overline{\mathbf{C}}(n-1). \tag{38}$$

Replacing this result in (37), and using Assumptions **A1**–**A3**, we obtain

$$\begin{aligned}\overline{\mathbf{C}}(n) = {} & \overline{\mathbf{C}}(n-1) - \sigma_u^2\boldsymbol{\mathcal{S}}\overline{\mathbf{C}}(n-1) - \sigma_u^2\overline{\mathbf{C}}(n-1)\boldsymbol{\mathcal{S}}^{\mathrm{T}} \\ & + \sigma_u^4\mathrm{Tr}\{\overline{\mathbf{C}}(n-1)\}\boldsymbol{\mathcal{S}}^2 + 2\sigma_u^4\boldsymbol{\mathcal{S}}\overline{\mathbf{C}}(n-1)\boldsymbol{\mathcal{S}}^{\mathrm{T}} + \boldsymbol{\mathcal{S}}^2\sigma_u^2\sigma_v^2,\end{aligned} \tag{39}$$

where we used the fact that $\boldsymbol{\mathcal{S}}\boldsymbol{\mathcal{S}}^{\mathrm{T}} = \boldsymbol{\mathcal{S}}^2$.

Since $\mathrm{Tr}\{\mathbf{AB}\} = \mathrm{Tr}\{\mathbf{BA}\}$ for any arbitrary matrices $\mathbf{A}$ and $\mathbf{B}$ of appropriate dimensions, we notice that $\mathrm{Tr}\{\overline{\mathbf{C}}(n-1)\boldsymbol{\mathcal{S}}^{\mathrm{T}}\} = \mathrm{Tr}\{\boldsymbol{\mathcal{S}}^{\mathrm{T}}\overline{\mathbf{C}}(n-1)\} = \mathrm{Tr}\{\boldsymbol{\mathcal{S}}\overline{\mathbf{C}}(n-1)\}$, where we took advantage of the fact that $\boldsymbol{\mathcal{S}}$ is symmetric. Thus, taking the trace of both sides of (39), we obtain after some algebraic manipulations:

$$\begin{aligned}\mathrm{Tr}\{\overline{\mathbf{C}}(n)\} = {} & \mathrm{Tr}\left\{\overline{\mathbf{C}}(n-1)\right\} - 2\sigma_u^2\mathrm{Tr}\left\{\boldsymbol{\mathcal{S}}\overline{\mathbf{C}}(n-1)\right\} + 2\sigma_u^4\mathrm{Tr}\left\{\boldsymbol{\mathcal{S}}^2\overline{\mathbf{C}}(n-1)\right\} \\ & + \mu^2M\sigma_u^4\mathrm{Tr}\{\overline{\mathbf{C}}(n-1)\} + \mu^2M\sigma_u^2\sigma_v^2.\end{aligned} \tag{40}$$

Let us introduce $\bar{c}_1(n)$ as the sum of the first M elements in the main diagonal of $\overline{\mathbf{C}}(n)$, i.e.,

$$c_1(n) \triangleq \sum_{k=1}^{M}[\overline{\mathbf{C}}(n)]_{k,k} \tag{41}$$

and $\bar{c}_2(n)$ as the sum of the last $L-M$ elements, i.e.,

$$c_2(n) \triangleq \sum_{k=M+1}^{L}[\overline{\mathbf{C}}(n)]_{k,k}. \tag{42}$$

Clearly, we can see that

$$\mathrm{Tr}\{\overline{\mathbf{C}}(n)\} = c_1(n) + c_2(n). \tag{43}$$

Furthermore, since $\boldsymbol{\mathcal{S}}$ is diagonal, we notice that

$$\mathrm{Tr}\left\{\boldsymbol{\mathcal{S}}\overline{\mathbf{C}}(n-1)\right\} = \mu c_1(n) \tag{44}$$

and

$$\mathrm{Tr}\left\{\boldsymbol{\mathcal{S}}^2\overline{\mathbf{C}}(n-1)\right\} = \mu^2 c_1(n). \tag{45}$$

Thus, (40) can be recast as

$$\mathrm{Tr}\{\overline{\mathbf{C}}(n)\} = \alpha c_1(n-1) + c_2(n-1) + \mu^2 M\sigma_u^4 c_2(n-1) + \mu^2 M\sigma_u^2\sigma_v^2, \tag{46}$$

with α given by (20).

Under Assumption **A4**, we have that $\mathrm{Tr}\{\overline{\mathbf{C}}(0)\} = \|\mathbf{w}^{\mathrm{o}}\|^2$. Since $[\boldsymbol{\mathcal{S}}]_{k,k} = 0$ for $M+1 \leq k \leq L$, as can be attested from (31), we notice from (39) that, for these indices k, $[\overline{\mathbf{C}}(n)]_{k,k} = (w_k^{\mathrm{o}})^2$ at every iteration n. Thus, we conclude that

$$c_2(n) = \|\mathbf{w}_2^{\mathrm{o}}\|^2 \tag{47}$$

for every n. In contrast, for $1 \leq k \leq M$, $[\boldsymbol{\mathcal{S}}]_{k,k} = \mu$ and therefore $[\overline{\mathbf{C}}(n)]_{k,k}$ varies along the iterations. Consequently, so does $c_1(n)$, with

$$c_1(0) = \|\mathbf{w}_1^{\mathrm{o}}\|^2. \tag{48}$$

Using (43), (47), and (48), by recursively applying (46) we obtain, in a similar fashion to (22), that

$$c_1(n) = \alpha^n\|\mathbf{w}_1^{\mathrm{o}}\|^2 + \chi'(1-\alpha^n), \tag{49}$$

where we introduced

$$\chi' \triangleq \frac{\mu M(\sigma_v^2 + \sigma_u^2\|\mathbf{w}_2^{\mathrm{o}}\|^2)}{2 - \mu(M+2)\sigma_u^2}. \tag{50}$$

Thus, from (43), (47), and (49), we get

$$\mathrm{Tr}\{\overline{\mathbf{C}}(n)\} = \alpha^n\|\mathbf{w}_1^{\mathrm{o}}\|^2 + \chi'(1-\alpha^n) + \|\mathbf{w}_2^{\mathrm{o}}\|^2. \tag{51}$$

From (14), we see that we can obtain the trace of the covariance matrix by subtracting (36) from (51). Doing so, we finally get

$$\boxed{\mathrm{Tr}[\mathbf{C}(n)] = \|\mathbf{w}_1^{\mathrm{o}}\|^2[\alpha^n - (1-\mu\sigma_u^2)^{2n}] + \chi'(1-\alpha^n).} \tag{52}$$

Taking the limit of (52), we can see that, in this case, the trace of the covariance matrix stabilizes in the steady state at

$$\lim_{n\to\infty} \mathrm{Tr}[\mathbf{C}(n)] = \chi'. \tag{53}$$

Finally, we notice from (9), (36), and (53) that the steady-state MSD is given by

$$\lim_{n\to\infty} \mathrm{MSD}(n) = \|\mathbf{w}_2^{\mathrm{o}}\|^2 + \chi'. \tag{54}$$

Thus, we remark that χ' does not equal the steady-state MSD, unlike what was observed in Sec. 3.1 for the case in which $M = L$, where the steady-state MSD was equal to χ. This is in accordance with results presented in [38]. In that work, it was shown that the steady-state EMSE of a deficient length LMS filter had both a bias and a variance component, unlike the case in which the length of the filter is sufficient in comparison with the optimal solution. However, this conclusion was reached through a different path in that reference, and the analysis followed a more traditional approach without relying on an explicit bias-variance decomposition.

The results above allow us to make some interesting comparisons. For instance, suppose that we have two scenarios. In both of them, the adaptive filter has M coefficients, but in the first case we have that $M < L$, whereas in the other the filter length matches that of the optimal solution, i.e., $M = L$. In this situation, if we compare (50) with (23), we can easily notice that $\chi' > \chi$, since the value of M is the same in either scenario. Thus, from (53), we can see that the variance of the elements of the vector $\widetilde{\mathbf{w}}(n)$ is greater in the former case in comparison with the latter. This can be interpreted as follows. Since in the first case the LMS algorithm cannot appropriately estimate $L - M$ coefficients of the optimal solution, its error is typically larger in magnitude in comparison with the case in which $M = L$. As a result, the elements of the vector $\mathbf{w}(n)$ tend to vary more from one iteration to the other, even in the steady state. From (50), we can see that, from the perspective of the variance of the estimates, everything occurs as if the filter were subject to a noise of variance $\sigma_v^2 + \sigma_u^2\|\mathbf{w}_2^{\mathrm{o}}\|^2$ in the case in which $M < L$, in contrast to the actual noise variance of σ_v^2 that affects the algorithm when $M = L$.

Let us now consider a different comparison. Suppose that the value of L is the same in both cases, but that the length of the adaptive filter is different. Once again, in the first case, we have that $M < L$, whereas in the other we have $M = L$. By comparing the performance obtained in both cases, we can evaluate the impact of an inadequate choice for M given a certain optimal solution. However, in this situation, we have to be careful when comparing (50) with (23), since the value of M is not the same in both cases. Subtracting χ' with $M < L$ from χ with $M = L$, we obtain

$$\Delta\chi \triangleq \chi - \chi' = \frac{\mu L \sigma_v^2}{2 - \mu(L+2)\sigma_u^2} - \frac{\mu M(\sigma_v^2 + \sigma_u^2\|\mathbf{w}_2^{\mathrm{o}}\|^2)}{2 - \mu(M+2)\sigma_u^2}, \tag{55}$$

which after some algebra leads to

$$\Delta\chi = \mu \cdot \left\{ \frac{2\sigma_v^2(1 - \mu\sigma_u^2)(L - M) - M\sigma_u^2\|\mathbf{w}_2^{\mathrm{o}}\|^2[2 - \mu(L+2)\sigma_u^2])}{[2 - \mu(L+2)\sigma_u^2][2 - \mu(M+2)\sigma_u^2]} \right\}. \tag{56}$$

Depending on the values of M, L, σ_v^2, μ, σ_u^2, and $\|\mathbf{w}^{\mathrm{o}}\|^2$, (56) may yield a positive or a negative number, meaning that the variance of the estimates may increase or decrease when $M < L$ in comparison with the case in which $M = L$. Moreover, the fact that the sign of $\Delta\chi$ depends on $\|\mathbf{w}_2^{\rm o}\|^2$ shows that the distribution of the coefficients in the optimal solution $\mathbf{w}^{\rm o}$ also influences whether the variance of the estimates increases or decreases when $M < L$ in comparison with the case in which $M = L$. Intuitively, the greater the value of $\|\mathbf{w}_2^{\rm o}\|^2$, the more $\Delta\chi$ tends to become negative, assuming that all other parameters remain fixed. Thus, the variance of the estimates tends to increase in comparison with the case in which $M = L$. Conversely, the greater the value of $\|\mathbf{w}_2^{\rm o}\|^2$, the more the rhs of (56) tends to be positive, indicating a smaller variance in comparison with the situation in which $M = L$. This can be attributed to the fact that, on the one hand, the variance associated with the $L - M$ that are not estimated is zero, but on the other hand the variance of the first M coefficients may rise due to the increased error, as discussed before. If $\|\mathbf{w}_2^{\rm o}\|^2$ is small, the impact of the choice of M on the estimation error is limited, and is thus outweighed by the zero variance associated with the $L - M$ coefficients that are not estimated. If $\|\mathbf{w}_2^{\rm o}\|^2$ is large, the opposite occurs.

Let us now examine the difference in steady-state MSD between the case in which $M = L$ and another one in which $M < L$, which we shall denote by $\Delta\text{MSD}(\infty)$. Comparing Eqs. (28) and (54), we notice that, for a fixed L, $\Delta\text{MSD}(\infty)$ is given by

$$\Delta\text{MSD}(\infty)=\mu\cdot\left\{\frac{2\sigma_v^2(1-\mu\sigma_u^2)(L-M)-\|\mathbf{w}_2^{\rm o}\|^2(1+M\sigma_u^2)[2-\mu(L+2)\sigma_u^2])}{[2-\mu(L+2)\sigma_u^2][2-\mu(M+2)\sigma_u^2]}\right\}. \quad (57)$$

Once again, the term in the rhs can be positive or negative depending on the values of the parameters. In order for $\Delta\text{MSD}(\infty)$ to be positive, the trace of the covariance matrix must decrease when $M < L$ is adopted in comparison with the case in which $M = L$. Furthermore, this decrease must compensate for the addition of the norm of the bias in (54). As a result, $\Delta\text{MSD}(\infty)$ is typically positive only when $\|\mathbf{w}_2^{\rm o}\|^2$ is very small. Essentially, in this case it is not worth it to try to estimate these coefficients, since the variance of the estimates outweighs the benefits of identifying them.

3.3 Case 3: $M > L$

In this case, we must consider the vector $\widetilde{\mathbf{w}}(n)$ given by (7). Thus, if we subtract both sides of (2) from $\boldsymbol{\omega}^{\rm o}$, and replace (1) and (3) in the resulting equation, we obtain after some algebra

$$\widetilde{\mathbf{w}}(n) = [\mathbf{I}_M - \mu\mathbf{u}_M(n)\mathbf{u}_M^{\rm T}(n)]\widetilde{\mathbf{w}}(n-1) - \mu\mathbf{u}_M(n)v(n), \quad (58)$$

which is the expression as (10), with the only difference that in this case we must highlight that we are considering the M-length regressor vector $\mathbf{u}_M(n)$, since $M \neq L$. Thus, following a similar line of thought to the one adopted in Sec. 3.1, it is straightforward to conclude that the squared norm of the bias evolves according to

$$\|\text{E}\{\widetilde{\mathbf{w}}(n)\}\|^2 = (1-\mu\sigma_u^2)^{2n}\|\boldsymbol{\omega}^{\rm o}\|^2, \quad (59)$$

which is clearly analogous to (13). However, since $\boldsymbol{\omega}^{\mathrm{o}}$ is obtained by applying zero padding to the vector $\mathbf{w}^{\mathrm{o}}$, we notice that $\|\boldsymbol{\omega}^{\mathrm{o}}\|^2 = \|\mathbf{w}^{\mathrm{o}}\|^2$. Consequently, (59) coincides with (13), and we conclude that the latter is valid for $M \geq L$ as a whole.

In regards to the trace of the covariance matrix, as in the previous sections, we use (14) with $\overline{\mathbf{C}}(n)$ given by (15). Thus, starting from (58), we still arrive at Eq. (16), which under Assumptions **A1**–**A4** leads to (18). Consequently, the latter expression still holds when $M > L$. The only difference is that in this case the matrix $\overline{\mathbf{C}}$ must be initialized as

$$\mathbf{C}(0) = \boldsymbol{\omega}^{\mathrm{o}}\boldsymbol{\omega}^{\mathrm{o}^{\mathrm{T}}} = \begin{bmatrix} \mathbf{w}^{\mathrm{o}}\mathbf{w}^{\mathrm{o}^{\mathrm{T}}} & \mathbf{0}_{L\times\Delta_M} \\ \mathbf{0}_{\Delta_M\times L} & \mathbf{0}_{\Delta_M\times\Delta_M} \end{bmatrix},$$

where $\Delta_M \triangleq M - L$. We remark, however, that $\mathrm{Tr}\{\mathbf{C}(0)\} = \|\mathbf{w}^{\mathrm{o}}\|^2$, as was the case in Sec. 3.1. As a result, we conclude that the trace of the covariance matrix, in the case in which $M > L$, is still given by (24). Moreover, (23) and (27) also hold in this scenario.

The discussion above shows that, if $M > L$, the norm of the bias vanishes as $n \to \infty$, just as in the case in which $M = L$. This is reasonable, since in this case there is no reason why the filter would not be able to estimate each coefficient of the optimal solution. Furthermore, both the norm of the bias and the trace of the covariance matrix are governed by the same expressions as in the situation in which the length of the filter perfectly matches that of $\mathbf{w}^{\mathrm{o}}$. However, since χ depends on M, we can see from (27) that the trace of the covariance matrix stabilizes at a greater value than the one that would be achieved if $M = L$. More specifically, if we compare the value of χ in the situation in which $M > L$ with the one achieved when $M = L$, which we shall denote by χ_M and χ_L, respectively, we notice that

$$\Delta\chi_{ML} \triangleq \chi_M - \chi_L = \frac{2\mu\sigma_v^2(1-\mu\sigma_u^2)(M-L)}{[2-\mu(M+2)\sigma_u^2][2-\mu(L+2)\sigma_u^2]}, \tag{60}$$

which is always positive assuming that (26) holds. In other words, the variance of the estimates increases in this scenario in comparison with the case in which $M = L$. Once again, this makes sense, since the $M - L$ coefficients of the filter that seek to estimate the zero elements of $\boldsymbol{\omega}^{\mathrm{o}}$ should fluctuate with the variations in the estimation error and thus contribute to the overall variance of the estimates.

3.4 Impulsive Noise

In this section, we discuss the effects of the existence of impulsive noise on our models. For the sake of simplicity, we focus our discussion on the results of Sec. 3.1, but the conclusions extend to the scenarios of Secs. 3.2 and 3.3 as well. We begin by noticing that, impulsive or not, the measurement noise does not affect the bias of the estimates, as can be seen from Eqs. (11)–(13). As a result, the presence or absence of impulsive noise does not affect the evolution of the term related to the bias in (9) whatsoever. The measurement noise only affects the trace of the covariance matrix along the iterations, as we can see from Eqs. (16)–(24). More specifically, we can see that the influence of the measurement noise is simply due to its effect on the parameter χ given by (23). This is reasonable, as we should expect noisier environments to lead to a greater variance in the estimates of the optimal solution.

As mentioned previously, so long as Assumption **A1** holds, all of the theoretical results obtained thus far remain valid, regardless of the distribution of $v(n)$. This may extend, for instance, to scenarios in which $v(n)$ represents some sort of impulsive noise. The only point of attention is that in this case the parameter σ_v^2 that appears in our models must represent the total variance of the measurement noise, considering both the impulsive and non-impulsive terms, if present. To illustrate this, let us consider the Bernoulli-Gaussian model [16, 19, 47] for impulsive noise, for example. In this case, we can write [19]

$$v(n) = v_0(n) + b(n)v_1(n), \tag{61}$$

where $v_0(n)$ and $v_0(n)$ are two Gaussian noises with zero mean and variances $\sigma_{v_0}^2$ and $\sigma_{v_1}^2$, respectively, and $b(n)$ is a Bernoulli random variable such that $b(n) = 1$ with probability p and $b(n) = 0$ with probability $1 - p$ at every iteration n. Typically, one considers $\sigma_{v_1}^2 \gg \sigma_{v_0}^2$. Thus, $v_0(n)$ represents the background noise, whereas the term $b(n)v_1(n)$ represents the impulsive interference from some source, with the incidence of impulses represented by a binary Bernoulli distribution, and their amplitudes by a zero-mean Gaussian distribution [47]. In this case, it can be shown that the total variance of $v(n)$ is given by [16]

$$\sigma_v^2 = \sigma_{v_0}^2 + p\sigma_{v_1}^2. \tag{62}$$

By replacing (62) in (23) and then in Eq. (24), we can directly apply the theoretical model to this type of scenario.

As another example, let us consider the Middleton Class-A model [47, 54], for instance. It seeks to represent a situation in which there is a superposition of statistically independent sources of impulsive noise. The amplitude of the impulse produced by the k-th source is modeled as a Gaussian distribution with zero mean and variance

$$\sigma_k^2 = \sigma_I^2 \left(\frac{k}{A}\right) + \sigma_G^2, \tag{63}$$

where σ_G^2 denotes the variance of the background noise and σ_I^2 is the average variance of the impulsive noise. In its turn, the probability of occurrence of an impulse from the k-th source is modeled according to a Poisson distribution, i.e.,

$$P_k = e^{-A}\left(\frac{A^k}{k!}\right) \tag{64}$$

for $k = 0, 1, 2, \cdots$, where A is a parameter of the Poisson distribution related to the average number of impulses per time unit and to the duration of a typical interfering signal [54]. The greater the value of A, the more common the impulsive events are. The probability density function (pdf) $f[v(n)]$ of the noise is modeled as [54]

$$f[v(n)] = \sum_{k=0}^{\infty} P_k f_G(v; \sigma_k^2), \tag{65}$$

where $f_G(v;\sigma_k^2)$ denotes the zero-mean Gaussian pdf with variance σ_k^2. Furthermore, a parameter of interest when dealing with the Middleton Class-A model is the ratio Γ between the power of the background noise and the power of the impulsive noise, i.e.,

$$\Gamma = \frac{\sigma_G^2}{\sigma_I^2}.$$

It can be shown that the total noise variance is given by

$$\sigma_v^2 = \sigma_G^2 + \sigma_I^2 \tag{66}$$

in this case. Thus, our theoretical models can be straightforwardly applied to a scenario with impulsive noise by replacing (66) in (23).

Evidently, the previous arguments apply to different models for impulsive noise, so long as we can calculate the total noise variance σ_v^2 and replace it in (23). Interestingly, the discussion so far reveals that if we compare two scenarios, one in which $v(n)$ has a Gaussian distribution, and another one in which it is impulsive, we should not see any difference between them in terms of both the bias and the variance terms so long as σ_v^2 is the same in both cases.

4 Simulation Results

In this section we present simulation results to validate the analyses conducted thus far. Unless stated otherwise, these results were obtained over an average of 10^3 independent realizations, considering different values for μ and M, in a system identification setup. In each case, the coefficients of $\mathbf{w}^{\mathrm{o}}$ are generated randomly following a uniform distribution in the range $[0,1]$, and later normalized so that $\mathbf{w}^{\mathrm{o}}$ has unit norm. In Secs. 4.1–4.3, we consider a Gaussian distribution for $u(n)$ with zero mean, and variance $\sigma_u^2 = 1$. In these cases, we already initialize the regressor vectors $\mathbf{u}_L(0)$ and $\mathbf{u}_M(0)$ with L and M samples, respectively, drawn from a Gaussian distribution with the aforementioned parameters.

Next, we divide the current section in different subsections, one for each type of scenario considered. In Secs. 4.1, we explore the case in which $M = L$, whereas in Sec. 4.2, we examine scenarios in which $M \neq L$. In both of these sections, we consider a Gaussian distribution for $v(n)$. However, in Sec. 4.3, we study scenarios in the presence of impulsive noise. Lastly, in Sec. 4.4 we investigate a scenario involving AEC in which real-world speech signals as the input of the adaptive filter.

4.1 Gaussian Noise, $M = L$

In this section, we consider a Gaussian distribution for $v(n)$, with zero mean and variance $\sigma_v^2 = 0.01$. In Fig. 1 we present a comparison between the simulations and the theoretical results of Eqs. (9), (13), and (24), considering $\mu = 10^{-3}$ and $M = 10$. In Fig. 1(a), we show the norm of the bias, in Fig. 1(b) the trace of the covariance matrix, and in Fig. 1(c) the overall MSD. In addition to the theoretical curves, we also indicate the values of χ, n_p, and n_s given by (23), (25), and (29), respectively, by dashed lines. In Fig. 1(a), we can see that the theoretical results match the simulations very closely during the first half of each realization. Interestingly, the norm of the bias eventually stabilizes at a level of approximately -70 dB, unlike the theoretical curve, which proceeds with the exponential decay of (13). This discrepancy may be related to Assumption **A2**, which does not hold in practice. However, we can see that this difference between the simulation results and the theoretical model occurs only when the norm of the bias is negligible and practically ceases to affect the MSD. This can be attested from Fig. 1(c), in which the theoretical curve practically overlaps with the simulation results. Furthermore, in Fig. 1(b), we can see that the simulations match (24) closely, and that (25) correctly predicts the iteration at which the peak of the trace of the covariance matrix occurs. At $n = n_p$, the difference between the simulation results and the theoretical model is less than 1 dB. Finally, by comparing Figs. 1(a), (b), and (c), we can see that the iteration n_s occurs when the MSD is approximately only 3 dB higher than its steady-state value, which corresponds to χ. Up until this point, the norm of the bias is the predominant term in the composition of the MSD in (9). Thus, in this case, it is the preponderant factor for the performance during most of the transient phase. In contrast, in the steady state, the MSD is dominated by the trace of the covariance matrix, which converges to χ, as we should expect. This can be seen from Figs. 1(b) and (c), and agrees with (27).

In Fig. 2, we repeat the simulations of Fig. 1, but considering $\mu = 10^{-2}$. In this case, we can draw similar conclusions to those obtained from Fig. 1, with a few differences. Firstly, we observe that the norm of the bias and the trace of the covariance matrix change places in terms of relevance earlier on, when the MSD is approximately 6.2 dB above its steady-state value. Moreover, comparing Fig. 2(a) with Fig. 1(a), it is clear that the norm of the bias decays at a faster rate in the former case, considering the difference in the time scale. Analyzing Figs. 2(b) and 1(b), we notice that, for $\mu = 10^{-2}$, the trace of the covariance matrix reaches a higher value at steady state and when the peak occurs, in comparison with the case in which $\mu = 10^{-3}$. Combined, these observations support the idea that, as μ increases, the norm of the bias decays at a faster rate, but the trace of the covariance matrix increases.

Lastly, in Fig. 3, we present the results for $\mu = 10^{-2}$ and $M = 50$. Comparing Figs. 2(a) and 3(a), we can see that the change in the value of M did not alter the decay rate of the norm of the bias, which is in accordance with our expectations. From Fig. 3(b), we can see that, in this case, the theoretical model slightly underestimates the value of n_p. However, for later iterations, the simulations match the theoretical curve closely. Finally, in this scenario, we can see that the norm of the bias and the trace of the covariance matrix change places in terms of relevance in the middle of the transient, while the MSD is far from its steady-state value χ, unlike what we observed in the simulations of Figs. 1 and 2. This difference can be attributed to the fact that both μ and M are greater in this case, in comparison with the other scenarios considered.

Finally, in order to enable a better understanding of the effects of μ and M on the trace of the covariance matrix, in Fig. 4, we present the simulation results for $\text{Tr}[\mathbf{C}(n)]$ of Figs. 1(b), 2(b), and 3(b). For $M = 10$, the adoption of a larger μ increases the

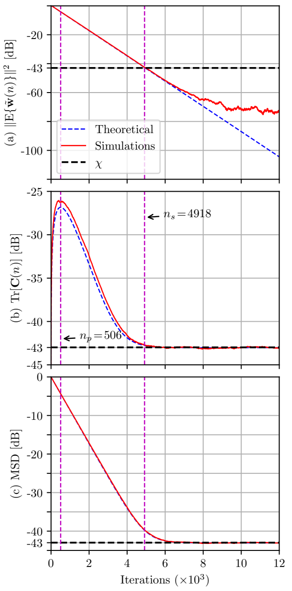

Fig. 1: Theoretical and simulation results for $\mu = 10^{-3}$ and $L = M = 10$.

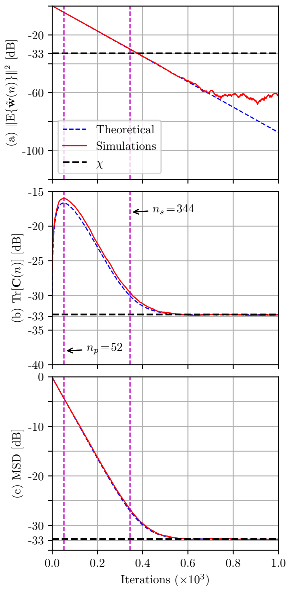

Fig. 2: Theoretical and simulation results for $\mu = 10^{-2}$ and $L = M = 10$.

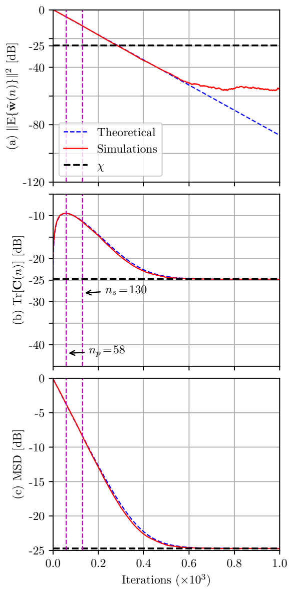

Fig. 3: Theoretical and simulation results for $\mu = 10^{-2}$ and $L = M = 50$.

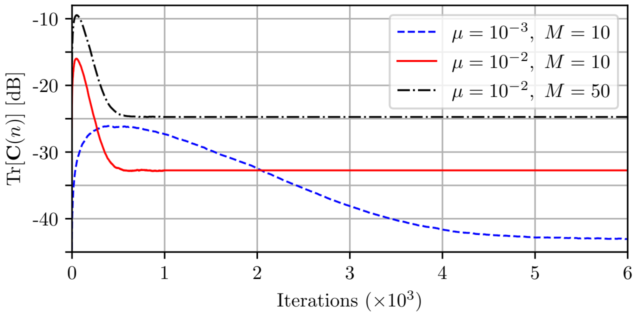

Fig. 4: Simulation results for the trace of the covariance matrix in the scenarios of Figs. 1, 2, and 3.

steady-state and peak values of $\mathrm{Tr}[\mathbf{C}(n)]$. For $\mu = 10^{-2}$, the increase from $M = 10$ to $M = 50$ seems to practically shift the curve upwards by approximately 8 dB.

4.2 Gaussian Noise, $M \neq L$

In this section, we consider the case in which the optimal solution has $L = 50$ coefficients, as was the case in the simulations of Fig. 3. However, we vary the length of

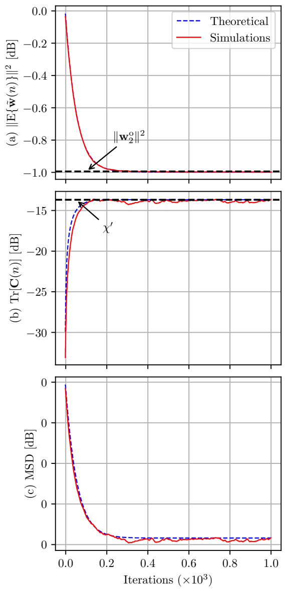

Fig. 5: Theoretical and simulation results for $\mu = 10^{-2}$ and $L = 50$, with $M = 10$.

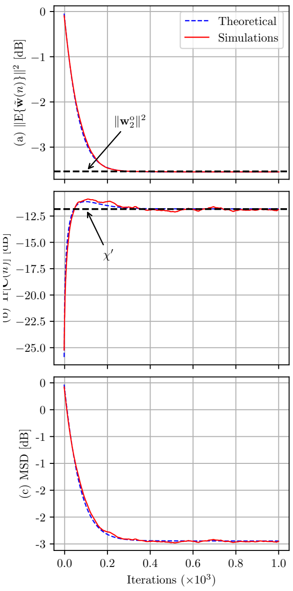

Fig. 6: Theoretical and simulation results for $\mu = 10^{-2}$ and $L = 50$, with $M = 25$.

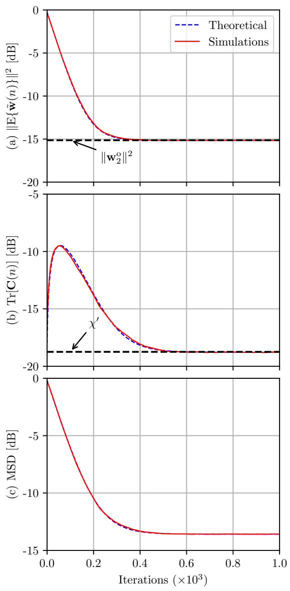

Fig. 7: Theoretical and simulation results for $\mu = 10^{-2}$ and $L = 50$, with $M = 49$.

the adaptive filter in order to verify the results of Secs. 3.2 and 3.3. Once again, $v(n)$ is Gaussian with zero mean and variance $\sigma_v^2 = 0.01$ and we adopt $\mu = 10^{-2}$.

In the simulations of Figs. 5, 6, and 7, we show the results obtained with $M = 10$, $M = 25$, and $M = 49$, respectively. In Figs. 5(a), 6(a), and 7(a) we present the norm of the bias along the iterations, in Figs. 5(b), 6(b), and 7(b) the trace of the covariance matrix, and in Figs. 5(c), 6(c), and 7(c) the total MSD.

We can see that the simulation results match the theoretical curves in Figs. 5–7, which validates the results of Sec. 3.2. Analyzing Figs. 5(a), 6(a), and 7(a), we can clearly see that the theoretical curve for the bias does not decrease continuously, unlike what was observed in Fig. 3(a). Furthermore, we can see in these figures that the norm of the bias stabilizes at the value of $\|\mathbf{w}_2^{\mathrm{o}}\|^2$, as was expected. This leads to a steady-state bias norm of roughly −1 dB in Fig. 5(a), −3.5 dB in Fig. 6(a), and −15 dB in Fig. 7(a), in contrast to the value of −70 dB observed in Fig. 3(a). Evidently, the impact of varying the value of M, with $M < L$, may change from one scenario to the other depending on how the coefficients of the vector $\mathbf{w}^{\mathrm{o}}$ are distributed, which influences the value of $\|\mathbf{w}_2^{\mathrm{o}}\|^2$. For example, if $\mathbf{w}^{\mathrm{o}}$ is sparse, gradually increasing M may not immediately reduce the value of $\|\mathbf{w}_2^{\mathrm{o}}\|^2$ due to the presence of a significant number of zero or near-zero elements in $\mathbf{w}^{\mathrm{o}}$. Consequently, the norm of the bias in the steady state might not change significantly for a certain range of possible choices for $M < L$. However, as can be attested from the results obtained, for a fixed L and $M < L$, the norm of the bias either decreases or remains unchanged as we increase M, up until the point when they coincide. At this point, the norm of the bias theoretically tends to zero in the steady state, and increasing M even further does not change this.

Analyzing Figs. 5(b), 6(b), and 7(b), we can also see that, the trace of the covariance matrix increases in the steady state in these scenarios when compared to the case in which $M = L$, as can be attested from Fig. 3. However, we notice that, as we decrease the value of M, the variance of the estimates does not increase monotonically in the steady state. Comparing Figs. 7(b) and 6(b), we can see that decreasing the value of M in this case led to an increase in the trace of the covariance matrix during the steady state. However, comparing Figs. 6(b) and 5(b), we notice that the variance of the estimates decreases even though we reduce the length of the filter from $M = 25$ to $M = 10$.

Finally, comparing Figs. 5(c), 6(c) and 7(c) with Fig. 3(c), we notice that in this scenario the selection of $M < L$ leads to a deterioration in the performance in comparison with the case in which $M = L$. However, as discussed in Sec. 3.2, this is not necessarily always the case. To illustrate this, in the simulations of Figs. 8, 9, and 10 we repeat the experiments of Figs. 5, 6, and 3, respectively, but with a different optimal system. In this scenario, the k-th element of $\mathbf{w}^{\text{o}}$ is generated randomly following a uniform distribution in the range $[0, 1/k^2]$ for $k = 1, \cdots, 50$. Then, the coefficients were normalized so that $\mathbf{w}^{\text{o}}$ has unit norm. To facilitate comparisons between the three scenarios, we adopt the same scale for the y axis in all three figures. Comparing Figs. 8(a), 9(a), and 10(a), we can see that the smaller the M, the greater the steady-state norm of the bias, as before. In Figs. 8(b), 9(b), and 10(b), we notice that, for the three values adopted for M, a shorter filter length leads to a smaller trace of the covariance matrix in the steady state. More interestingly, we notice from Figs. 8(c), 9(c), and 10(c) that, in this case, adopting $M < L$ leads to an overall decrease in the steady-state MSD, rather than an increase. This is due to the fact that, in this case, the rise in the norm of the bias is more than compensated by the reduction in the variance of the estimates. This can be attributed to the smaller values of $\|\mathbf{w}_2^{\text{o}}\|^2$ in comparison with the scenarios considered in the simulations of Figs. 5 and 6, as discussed in Sec. 3.2.

In the simulations of Figs. 11, 12, and 13, we once again consider the same vector $\mathbf{w}^{\text{o}}$ used in the simulations of Figs. 1–7, but the filter length M is set to 51, 60, and 75, respectively. Thus, we consider the scenario in which $M > L$, which we analyzed in Sec. 3.3. We can see that once again the simulation results match the theoretical curves well. There is only a slight discrepancy between the theory and simulation results during the transient phase in Figs. 13(b) and (c). This may be due to the fact that the independence theory, i.e., Assumption **A2**, tends to be more realistic for relatively small values of M [39]. Regardless, we can see from Figs. 11(a), 12(a), and 13(a) that the squared norm of the bias becomes negligible in the steady state, similarly to what was observed in Fig. 3(a). Furthermore, from Figs. 11(b), 12(b), and 13(b), we can see that the value of the trace of the covariance matrix in steady state rises as we increase M. This results in a slight deterioration in the steady-state MSD, as can be perceived from Figs.. 11(c), 12(c), and 13(c). All of this is in accordance with the observations made in Sec. 3.3.

In order to better understand the possible effects of the choice of M on the steady-state performance, in Figs. 14(a), (b), and (c) we present the theoretical and simulation

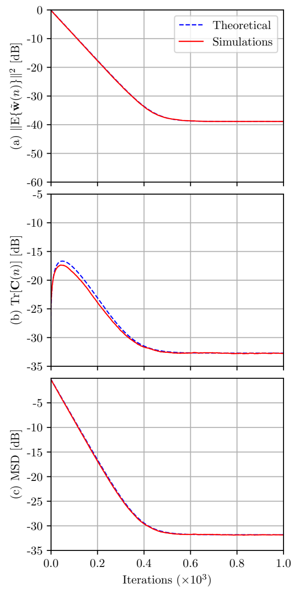

Fig. 8: Theoretical and simulation results for $\mu = 10^{-2}$ and $L = 50$, with $M = 10$, in a scenario in which each coefficient w_k^{o} is drawn from $\mathcal{U}(0, 1/k^2)$ and then normalized so that $\|\mathbf{w}^{\mathrm{o}}\| = 1$.

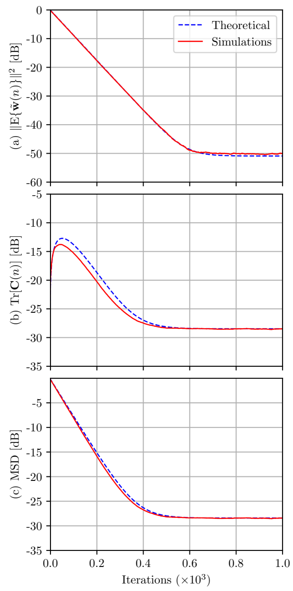

Fig. 9: Theoretical and simulation results for $\mu = 10^{-2}$ and $L = 50$, with $M = 25$, in a scenario in which each coefficient w_k^{o} is drawn from $\mathcal{U}(0, 1/k^2)$ and then normalized so that $\|\mathbf{w}^{\mathrm{o}}\| = 1$.

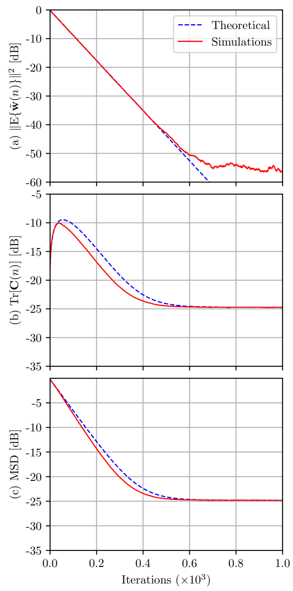

Fig. 10: Theoretical and simulation results for $\mu = 10^{-2}$ and $M = L = 50$, in a scenario in which each coefficient w_k^{o} is drawn from $\mathcal{U}(0, 1/k^2)$ and then normalized so that $\|\mathbf{w}^{\mathrm{o}}\| = 1$.

results for the steady-state values of the squared norm of the bias, the trace of the covariance matrix, and the MSD, respectively, for $1 \leq M \leq 100$. The step size was set to $\mu = 10^{-2}$. In order to obtain these results, we increased the number of iterations from the usual 1000 to 2000 so as to ensure that the algorithm achieved the steady state for every value of M considered. Then, we calculated the average of the variables of interest along the course of the last 400 time instants of each realization. We can see that overall the simulation results match the theoretical values well. From Fig. 14(a), we notice that the squared norm of the bias gradually decreases as increase M, up until the point in which $M = L = 50$. From this point forward, the bias should vanish according to our theoretical model, but in practice its squared norm stabilizes at around -55 dB for $M = L = 50$ and slightly increases to roughly -50 dB for $M = 100$. From Fig. 14(b), we can see that the trace of the covariance matrix displays an interesting behavior in regards to M. As we increase the filter length, its steady-state value gradually increases, until it reaches a peak and starts to decrease as we approach the scenario in which $M = L$. At this particular point, the trace of the covariance matrix reaches its minimum, and then gradually increase once again as M becomes larger in comparison with L. As a result, we can observe from Fig. 14(c)

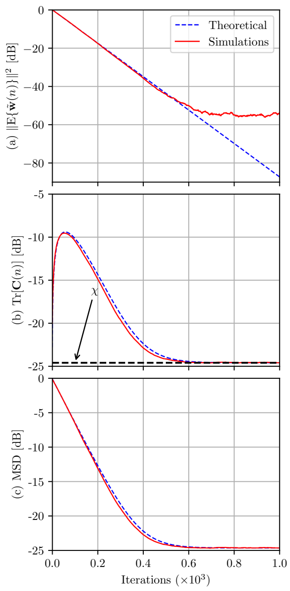

Fig. 11: Theoretical and simulation results for $\mu = 10^{-2}$ and $L = 50$, with $M = 51$.

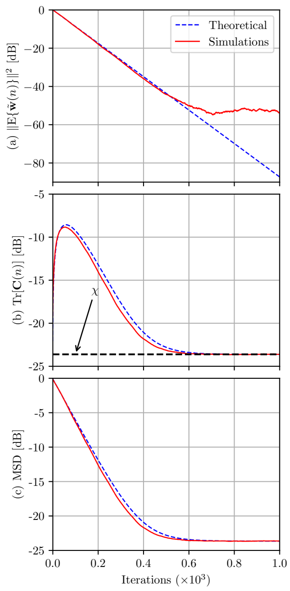

Fig. 12: Theoretical and simulation results for $\mu = 10^{-2}$ and $L = 50$, with $M = 50$.

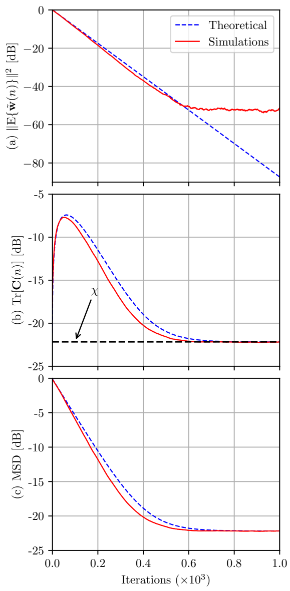

Fig. 13: Theoretical and simulation results for $\mu = 10^{-2}$ and $L = 50$, with $M = 75$.

that the steady-state MSD is at its minimum when $M = L$. Comparing Fig. 14(c) with Figs. 14(a) and (b), for low values of M, we can notice that the steady-state MSD is dominated by the bias for low values of M, and by the trace of the covariance matrix when M increases. This is in accordance with our expectations, based on the discussion presented so far. We remark, however, that the results obtained in Fig. 14 could change depending on the distribution of the coefficients in the optimal solution $\mathbf{w}^{\mathrm{o}}$. This is exemplified by the simulations of Figs. 8–10, in which the steady-state MSD is smaller for $M < L$ than for $M = L$. Thus, some care must be taken when interpreting the results of Fig. 14, but it exemplifies what may occur as we vary the value of M, and shows that our theoretical model accurately predicts the impact of each term on the steady-state MSD.

4.3 Effects of Impulsive Noise

In the simulations of this section, we consider $M = L = 50$ and $\mu = 10^{-2}$. We begin by comparing three scenarios, each one with a different type of noise. In all three cases, however, the total noise variance σ_v^2 is the same. Thus, from Eqs. (13), (23), and (24), we should expect to see the same results in all three scenarios. We consider: i) a Gaussian distribution for $v(n)$ with zero mean and variance $\sigma_v^2 = 0.11$, ii) a Bernoulli-Gaussian model for $v(n)$ with $\sigma_{v_0}^2 = 0.01$, $\sigma_{v_1}^2 = 1$ and $p = 0.1$, resulting in $\sigma_v^2 = 0.11$ in (62), and iii) the Middleton Class-A model with $A = 1$, $\sigma_G^2 = 0.0011$,

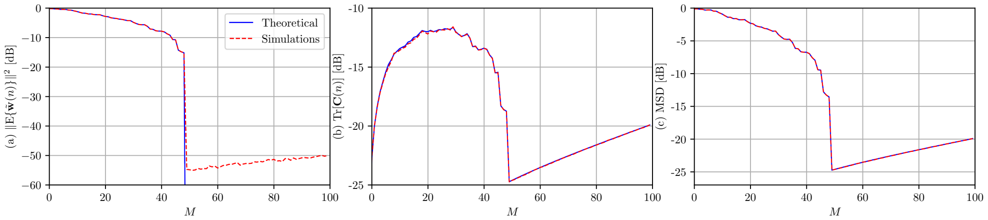

Fig. 14: Steady-state values of (a) the squared norm of the bias, (b) the trace of the covariance matrix, and (c) the MSD for different values of M, with $\mu = 10^{-2}$ and $L = 50$.

and $\sigma_I^2 = 0.1089$, which results in $\sigma_v^2 = 0.11$ in Eq. (66) and corresponds to $\Gamma = 0.01$. In Figs. 15(a), (b), and (c) we show the normalized histograms of the measurement noise for the scenarios i), ii), and iii), respectively, considering all 1000 realizations of $v(n)$. For reference, in each plot we also represent the Gaussian distribution with zero mean and variance $\sigma_v^2 = 0.11$ by a solid red line. We can see that the distributions resulting from the Bernoulli-Gaussian and Middleton models are clearly distinct from the normal distribution in the scenarios considered.

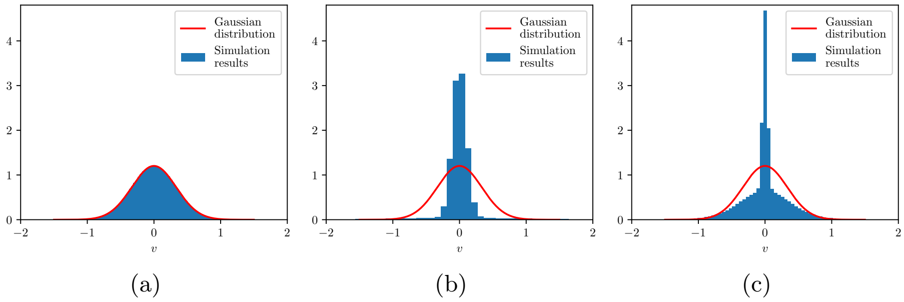

Fig. 15: Normalized histograms for the measurement noise, considering the results from the 1000 realizations. In each plot, we also depict the Gaussian pdf with zero mean and variance $\sigma_v^2 = 0.11$. (a) scenario i), considering a Gaussian distribution with zero mean and variance $\sigma_v^2 = 0.11$. (b) Bernoulli-Gaussian model with $\sigma_{v_0}^2 = 0.01$, $\sigma_{v_1}^2 = 1$ and $p = 0.1$. (c) Middleton Class-A model with $A = 1$, $\sigma_G^2 = 0.0011$, and $\sigma_I^2 = 0.1089$

The simulation results for the scenarios i), ii) and iii) described previously are presented in Figs. 16, 17, and 18, respectively. As was the case in Figs. 1–3, in Figs. 16(a), 17(a) and 18(a) we present the norm of the bias along the iterations, in Figs. 16(b), 17(b) and 18(b) the trace of the covariance matrix, and in Figs. 16(c), 17(c) and 18(c) the total MSD.

Comparing Figs. 16, 17, and 18, we can see that the theoretical curves are the same for all three scenarios, and that the simulation results obtained are very similar

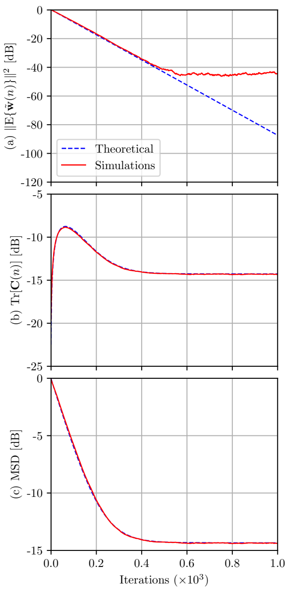

Fig. 16: Theoretical and simulation results for $\mu = 10^{-2}$ and $M = 50$, considering a Gaussian distribution for $v(n)$ with zero mean and $\sigma_v^2 = 0.11$.

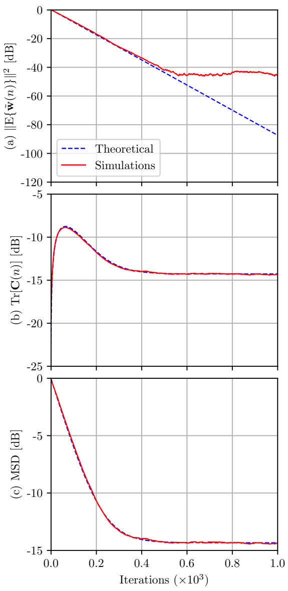

Fig. 17: Theoretical and simulation results for $\mu = 10^{-2}$ and $M = 50$, considering the Bernoulli-Gaussian model for $v(n)$ with $\sigma_{v_0}^2 = 0.01$, $p = 0.1$ and $\sigma_{v_1}^2 = 1$.

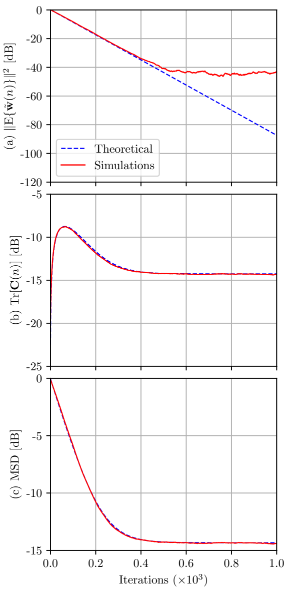

Fig. 18: Theoretical and simulation results for $\mu = 10^{-2}$ and $M = 50$, considering the Middleton Class-A model for $v(n)$ with $A = 1$, $\sigma_G^2 = 0.0011$, and $\sigma_I^2 = 0.1089$.

to one another. Moreover, we notice that the theoretical curves match the simulation results closely in all three cases. This supports the arguments made in Sec. 3.4 that, as long as Assumption **A1** holds, the variance component of the MSD of the LMS algorithm only depends on σ_v^2, regardless of the distribution of $v(n)$. The results also show that our theoretical models can be applied to scenarios with impulsive noise, so long as $v(n)$ is iid along the iterations with zero mean, variance σ_v^2 and is statistically independent from any other variable. Finally, comparing Figs. 3 and 16, we can see that the evolution of the bias is the same in both scenarios, but curves corresponding to the trace of the covariance matrix are different. This is reasonable, since the only difference between both scenarios lies in the value of σ_v^2. As argued in Sec. 3.4, the measurement noise influences the value of χ in (23), and therefore has an effect on the trace of the covariance matrix. However, it does not affect the bias term in Eq. (9) whatsoever, as we notice from (13).

4.4 Acoustic Echo Cancellation

In this section, we test our theoretical model in an AEC setting. For this purpose, we adopt real-world speech signals as the input of the adaptive filter. These signals were downloaded from a publicly available dataset of 40,000 audio samples with both male and female speakers [21, 25]. Since each audio sample is typically around 4 seconds long, we concatenated the data from six samples and then truncated them to obtain speech signals that were exactly 20 seconds long. Whenever the sample resulting from this process was shorter than this, we extended the signal by including a sufficient number of zeros so as to obtain the desired duration. Moreover, we made sure to avoid any overlap between the audio signals thus obtained. The room impulse response has $L = 256$ coefficients and was determined experimentally inside an automobile considering a sampling frequency of 8 kHz. It is worth noting that the signals were downsampled to match the sampling frequency of the experimental setup. We remark that in this case the optimal solution does not have unit norm, and realistically cannot be normalized. This is due to the fact that the input signal is not white Gaussian noise. Hence, depending on the power spectral density of the speech signal, the normalization of the optimal solution might result in an echo with more power than the input signal, which is not physically possible. Thus, in our theoretical model, we assume that we have a perfect knowledge of $\|\mathbf{w}^{\mathrm{o}}\|^2$, but in practice we would need to estimate the norm of the optimal solution. Since the input signal is not wide-sense stationary in this situation, we estimate the variance of the input signal at each time instant using a sliding rectangular window of length 160, which corresponds to a 20 ms period. This value was selected due to the fact that speech signals can be considered approximately stationary within a 20–40 ms time frame. Then, we run the theoretical model at the end of every realization considering the estimated variance along the iterations for that particular experiment. After running all realizations, we take the ensemble average of the results thus obtained to form the theoretical curves. In this section, we consider 100 independent realizations. We remark that in this case we use Eqs. (12), (18), (33), and (39), instead of directly applying Eqs. (13), (24), (36), and (52). This modification was made due to the fact that (13), (24), (36), and (52) are obtained considering that the variance of the input signal is constant along the iterations, which is not the case in this scenario. Then, we take the squared norm of the bias vector and the trace of the covariance matrix, as before. We set $\mu = 10^{-2}$ and $M = L = 256$.

We consider two scenarios, which are depicted in Figs. 19 and 20, respectively. The difference between them lies in the value of σ_v^2. In the former, we consider $\sigma_v^2 = 10^{-3}$, whereas in the latter we set $\sigma_v^2 = 10^{-4}$. These values correspond to a signal-to-noise ratio (SNR) of approximately -2.8 dB and 7.2 dB, respectively, when comparing the average power of the measurement noise with that of the echo signal. Analyzing Fig 19, we can see that in the first scenario the theoretical model predicts the behavior of the algorithm reasonably well. In the case of Fig. 20, we can see that for a higher SNR, the simulation results do not match the theoretical model as closely. This was expected, however, given the fact that the Assumption **A2** clearly does not hold for speech signals. Even in this scenario, we can see that the theoretical curves somehow reflect the tendencies observed in the simulation results from a qualitative perspective. Furthermore, the model accurately predicts the steady-state value of the trace of the covariance matrix in Fig. 20(b). Overall, taking into consideration the fact that the assumptions made do not hold in this case, the theoretical model obtained provides qualitative insights into the inner workings of the LMS algorithm in an AEC setting, which was our main goal after all.

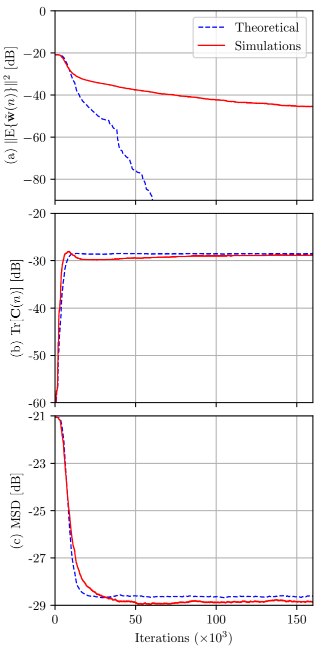

Fig. 19: Theoretical and simulation results for $\mu = 10^{-2}$ and $M = 50$, considering an AEC problem with $\sigma_v^2 = 10^{-3}$.

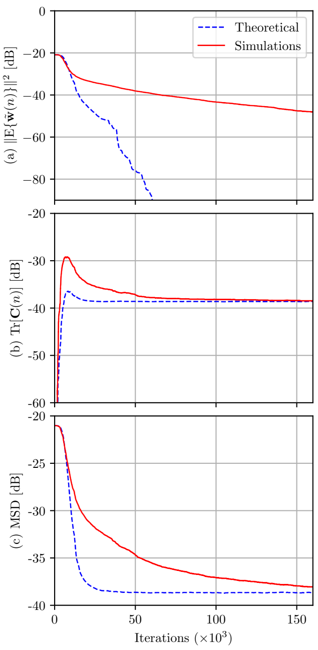

Fig. 20: Theoretical and simulation results for $\mu = 10^{-2}$ and $M = 50$, considering an AEC problem with $\sigma_v^2 = 10^{-4}$.

5 Conclusions

In this paper, we carried out the bias-variance decomposition of the MSD of the LMS algorithm. Our analysis has shown that, during the transient phase, if the filter length M matches that of the optimal solution, L, the norm of the bias is typically larger than the trace of the covariance matrix, and is more determinant to the performance of the algorithm. As the norm of the bias decays exponentially, its relevance is gradually reduced as the algorithm converges, whereas the trace of the covariance matrix increases, reaches a peak, and stabilizes. As a result, it is the latter that determines the steady-state performance of the algorithm, which agrees with our expectations. We also showed that there is a clear trade-off between the bias and variance terms with respect to the step size. Increasing μ causes the bias term to decrease faster, but increases the variance of the estimates in the steady state. In contrast, the adoption of lower values for μ lead to a smaller variance in the steady state, but causes the bias term to decrease more slowly. These results agree with the existing literature [13, 23, 46, 51]. Moreover, we extended the analysis to the cases in which $M \neq L$. It was shown that, if $M < L$, the bias term does not vanish in steady state, but rather stabilizes at a certain value that depends on the optimal solution $\mathbf{w}^{\mathrm{o}}$. This is also in accordance with existing results [38], and differs from what is observed for $M \geq L$. In contrast, the term related to the variance of the estimates may increase or decrease as we reduce the value of M, starting at $M = L$. For $M > L$, the variance rises as we increase M. However, the overall effect of these tendencies on the MSD depends on $\mathbf{w}^{\mathrm{o}}$, and under certain circumstances it is possible to obtain a smaller MSD with $M < L$ in comparison with the case in which $M = L$. In other words, while the bias and variance exhibit a clear trade-off with respect to μ, for a fixed value of L, their behavior in regards to M can be less obvious and depends on the system being identified. Finally, the impact of impulsive noise was analyzed. It was shown that the presence or absence of impulsive noise is indifferent to the bias and variance of the estimates. Instead, only the total variance of the measurement noise determines the performance of the algorithm. Simulation results support the main findings of our analysis, including a scenario with real-world speech data as the input signal. This indicates that the proposed model can perform fairly well under a wide range of circumstances, including those closer to the conditions typically found in practical applications.

For future works, we intend to investigate how the analysis presented in this paper can be extended to VSS algorithms [1, 5, 8, 20, 27, 34, 52, 58]. Moreover, it may be especially interesting to study the extension of the analysis presented in this paper to solutions that rely on the bias-variance trade-off of adaptive signal processing techniques, such as the algorithms proposed in [12, 26, 29, 33, 35, 36, 45, 50, 53, 56]. Given the similarities between the themes of these works and that of this paper, we believe that this would be a natural fit. Thus, we expect that by expanding our analysis to these solutions, we could shed more light into the way they operate, which in its turn could open up opportunities for the proposal of new techniques in the same vein.

Appendix A Deriving Eq. (9)

Let us expand the term inside the expectations in (8). Doing so, we observe that we can write

$$\begin{aligned}\mathbf{C}(n) = {} & \mathrm{E}\{\widetilde{\mathbf{w}}(n)\widetilde{\mathbf{w}}^{\mathrm{T}}(n)\} - \mathrm{E}\left\{\widetilde{\mathbf{w}}(n)\mathrm{E}\{\widetilde{\mathbf{w}}(n)\}^{\mathrm{T}}\right\} \\ & - \mathrm{E}\left\{\mathrm{E}\{\widetilde{\mathbf{w}}(n)\}\widetilde{\mathbf{w}}^{\mathrm{T}}(n)\right\} + \mathrm{E}\left\{\mathrm{E}\{\widetilde{\mathbf{w}}(n)\}\mathrm{E}\{\widetilde{\mathbf{w}}(n)\}^{\mathrm{T}}\right\}.\end{aligned} \tag{A1}$$

Since $\mathrm{E}\{\widetilde{\mathbf{w}}(n)\}$ is deterministic, we can take it out of the external expectations in the second, third, and fourth terms of the rhs of (A1), which leads to

$$\mathbf{C}(n) = \mathrm{E}\{\widetilde{\mathbf{w}}(n)\widetilde{\mathbf{w}}^{\mathrm{T}}(n)\} - \mathrm{E}\{\widetilde{\mathbf{w}}(n)\}\mathrm{E}\{\widetilde{\mathbf{w}}(n)\}^{\mathrm{T}}. \tag{A2}$$

Taking the trace of both sides in (A2), we obtain

$$\mathrm{Tr}\,[\mathbf{C}(n)] = \mathrm{Tr}\left[\mathrm{E}\{\widetilde{\mathbf{w}}(n)\widetilde{\mathbf{w}}^{\mathrm{T}}(n)\}\right] - \|\mathrm{E}\{\widetilde{\mathbf{w}}(n)\}\|^2, \tag{A3}$$

where we used the fact that

$$\mathrm{Tr}\left[\mathrm{E}\{\widetilde{\mathbf{w}}(n)\}\mathrm{E}\{\widetilde{\mathbf{w}}(n)\}^{\mathrm{T}}\right] = \|\mathrm{E}\{\widetilde{\mathbf{w}}(n)\}\|^2. \tag{A4}$$

Furthermore, since

$$\mathrm{Tr}[\mathrm{E}\{\widetilde{\mathbf{w}}(n)\widetilde{\mathbf{w}}^{\mathrm{T}}(n)\}] = \mathrm{E}\{\|\widetilde{\mathbf{w}}(n)\|^2\}, \tag{A5}$$

we can rewrite (A3) as

$$\mathrm{Tr}\,[\mathbf{C}(n)] = \mathrm{E}\{\|\widetilde{\mathbf{w}}(n)\|^2\} - \|\mathrm{E}\{\widetilde{\mathbf{w}}(n)\}\|^2, \tag{A6}$$

which straightforwardly leads to (9).

Data and Code Availability. Data and/or code generated during the development of the current study are available from the corresponding author on request

References

[1] Aboulnasr, T., Mayyas, K.: A robust variable step-size LMS-type algorithm: analysis and simulations. IEEE Transactions on Signal Processing **45**(3), 631–639 (1997)

[2] Arenas-García, J., Figueiras-Vidal, A.R., Sayed, A.H.: Mean-square performance of a convex combination of two adaptive filters. IEEE Transactions on Signal Processing **54**(3), 1078–1090 (2006)

[3] Arenas-García, J., Lázaro-Gredilla, M.: Tracking performance of adaptively biased adaptive filters. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 4128–4131 (2011)

[4] Azpicueta-Ruiz, L.A., Figueiras-Vidal, A.R., Arenas-García, J.: A normalized adaptation scheme for the convex combination of two adaptive filters. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 3301–3304 (2008)

[5] Benesty, J., Rey, H., Vega, L.R., Tressens, S.: A nonparametric VSS NLMS algorithm. IEEE Signal Processing Letters **13**(10), 581–584 (2006)

[6] Bermudez, J.C.M., Bershad, N. Eweda, E.: Stochastic Analysis of the LMS Algorithm for Cyclostationary Colored Gaussian Inputs. Signal Processing **160**, 127–136 (2019)

[7] Bershad, N.J.: Analysis of the normalized LMS algorithm with Gaussian inputs. IEEE Transactions on Acoustics, Speech, and Signal Processing **34**(4), 793–806 (1986)

[8] Bershad, N.J., Bermudez, J.C.: A switched variable step size NLMS adaptive filter. Digital Signal Processing **101**, 102730 (2020)

[9] Bishop, C.M., Bishop, H.: Deep learning: Foundations and concepts. Springer Nature, Cham, Switzerland (2023)

[10] Burra, S., Kar, A., Christensen, M.G.: An Improved Functional Link Architecture for Nonlinear AEC, In: Proceedings of the European Signal Processing Conference (EUSIPCO), 75–79 (2022)

[11] Burra, S., Kar, A., Christensen, M.G.: Conjugate Gradient Based Adaptive Algorithm for Nonlinear AEC, In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 556–560 (2024)

[12] Danaee, D., de Lamare, R.C., Nascimento, V.H.: Distributed Quantization-Aware RLS Learning With Bias Compensation and Coarsely Quantized Signals. IEEE Transactions on Signal Processing **70**, 3441–3455 (2022)

[13] Diniz, P.S.R.: Adaptive Filtering: Algorithms and Practical Implementation, 5th edn. Springer, Cham, Switzerland (2019)

[14] Domingos, P.: A unified bias-variance decomposition. In: Proceedings of 17th International Conference on Machine Learning, pp. 231–238 (2000). Morgan Kaufmann Stanford

[15] Feuer, A., Weinstein, E.: Convergence analysis of LMS filters with uncorrelated Gaussian data. IEEE Transactions on Acoustics, Speech, and Signal Processing **33**(1), 222–230 (1985)

[16] Finamore, W.A., Pinho, M.S., Sharma, M., Ribeiro, M.V: Modeling Noise as a Bernoulli-Gaussian Process. Journal of Communication and Information Systems **38**, 175–186 (2023)

[17] Fuster, L., de Diego, M., Ferrer, M., Gonzalez, A.: Steady-state analysis of biased filtered-x algorithms for adaptive room equalization. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 6647–6651 (2014)

[18] Geman, S., Bienenstock, E., Doursat, R.: Neural networks and the bias/variance dilemma. Neural computation **4**(1), 1–58 (1992)

[19] Ghosh, M.: Analysis of the Effect of Impulse Noise on Multicarrier and Single Carrier QAM Systems . IEEE Transactions on Communications **44**, 145–147 (1996)

[20] Hamidia, M., Amrouche, A.: Improved variable step-size NLMS adaptive filtering algorithm for acoustic echo cancellation. Digital Signal Processing **49**, 44–55 (2016)

[21] Harwath, D., Glass, H.: Deep Multimodal Semantic Embeddings for Speech and Images. In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop. Scottsdale, AZ, pp. 237–244 (2015)

[22] Hassibi, B., Sayed, A.H., Kailath, T.: H^{∞} optimality of the LS algorithm. IEEE Transactions on Signal Processing **44**(2), 267–280 (1996)

[23] Haykin, S.: Adaptive Filter Theory, 5th edn. Pearson, Upper Saddle River, NJ

(2014)

[24] Ho, K.C.: A Study of Two Adaptive Filters in Tandem. IEEE Transactions on Signal Processing **48**(6), 1626–1636 (2000)

[25] Hodosh, M., Young, P., Hockenmaier , J.: Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. Journal of Artificial Intelligence Research **47**, 853–899 (2013)

[26] Huang, F., Song, F., Zhang, S., So, H.C., Yang, S.: Robust Bias-Compensated LMS Algorithm: Design, Performance Analysis and Applications. Journal of the Franklin Institute **72**(10), 13214–13228 (2023)

[27] Huang, H.-C., Lee, J.: A new variable step-size NLMS algorithm and its performance analysis. IEEE Transactions on Signal Processing **60**(4), 2055–2060 (2011)

[28] Jiao, Y., Cheung, R.Y.P., Mok, M.P.C.: Modified Log-LMS Adaptive Filter with Low Signal Distortion for Biomedical Applications, In: Proceedings of the 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 5210–5213 (2012)

[29] Kang, B., Yoo, J., Park, P.: Bias-compensated normalised LMS algorithm with noisy input. Electronics Letters, **49**(8), 538–539 (2013)

[30] Kar, A., Padhi, T., Majhi, B., Swamy, M.N.S.: Analysing the impact of system dimension on the performance of a variable-tap-length adaptive algorithm. Applied Acoustics **150**, 207–215 (2019)

[31] Kar, A., Swamy, M.N.S.: Convergence and steady state analysis of a tap-length optimization algorithm for linear adaptive filters. International Journal of Electronics and Communications (AEÜ) **70**(9), 1114–1121 (2016)

[32] Kar, A., Swamy, M.N.S.: Tap-length optimization of adaptive filters used in stereophonic acoustic echo cancellation. Signal Processing **131**, 422–433 (2017)

[33] Krstajic, B., Stankovic, L.J., Uskokovic, Z.: An Approach to Variable Step-Size LMS Algorithm. Electronics Letters **38**(16), 885–887 (2002)

[34] Kwong, R.H., Johnston, E.W.: A variable step size LMS algorithm **40**, 1633–1642 (1992)

[35] Lázaro-Gredilla, M., Azpicueta-Ruiz, L.A., Figueiras-Vidal, A.R., Arenas-García, J.: Adaptively biasing the weights of adaptive filters. IEEE Transactions on Signal Processing **58**(7), 3890–3895 (2010)

[36] Liu, D., Zhao, H.: Affine Projection Sign Subband Adaptive Filter Algorithm With Unbiased Estimation Under System Identification, IEEE Transactions on

Circuits and Systems – II: Express Briefs **70**(3), 1209–1213 (2023)

[37] Maruo, M.H., Bermudez, J.C.M.: A linearly constrained framework for the analysis of the deficient length least-mean square algorithm. Digital Signal Processing **155**, 104747 (2024)

[38] Mayyas, K.: Performance Analysis of the Deficient Length LMS Adaptive Algorithm. IEEE Transactions on Signal Processing **53**(8), 2727–2734 (2005)

[39] Nascimento, V.H., Silva, M.T.M.: Adaptive filters. In: Chellapa, R., Theodoridis, S. (eds.) Academic Press Library in Signal Processing: Signal Processing Theory and Machine Learning vol. 1, pp. 619–761. Academic Press, Chennai, India (2014). Chap. 12

[40] Padhi, T., Chandra, M., Kar, A.: A New Hybrid Active Noise Control System with Convex Combination of Time and Frequency Domain Filtered-X LMS Algorithms. Circuits, Systems, and Signal Processing **37**, 3275–3294 (2018)

[41] Padhi, T., Chandra, M., Kar, A., Swamy, M.N.S.: Design and Analysis of an Improved Hybrid Active Noise Control System. Applied Acoustics **127**, 260–269 (2017)

[42] Pimenta, R., Petraglia, M.R., Haddad, D.B.: Stability analysis of the bias compensated LMS algorithm. Digital Signal Processing **147**, 104395 (2024)

[43] Pimenta, R., Resende, L., Petraglia, M.R., Haddad, D.B.: On the steady-state performance of bias-compensated LMS algorithm. Electronics Letters **57**(2), 85–88 (2021)

[44] Qureshi, S.U.: Adaptive equalization, Proceedings of the IEEE **73**(9), 1349–1387 (1985)

[45] Rosalin, Patnaik, A., Nanda, S., Rout, D.K.: A bias-compensated NLMS algorithm based on arctangent framework for system identification. Signal, Image and Video Processing **18**(4), 3595–3601 (2024)

[46] Sayed, A.H.: Adaptive Filters. John Wiley & Sons, Hoboken, NJ (2008)

[47] Selim, B., Alam, M.S., Evangelista, J.V.C., Kaddoum, G., Agba, B.L.: NOMA-Based IoT Networks: Impulsive Noise Effects and Mitigation. IEEE Communications Magazine **58**, 69–75 (2020)

[48] Silva, M.T., Candido, R., Arenas-García, J., Azpicueta-Ruiz, L.A.: Improving multikernel adaptive filtering with selective bias. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 4529–4533 (2018)

[49] Silva, T.T.P., Igreja, F., Lara, P., Tarrataca, L., Kar, A., Haddad, D.B.,: On

the Skewness of the LMS Adaptive Weights. IEEE Transactions on Circuits and Systems – II: Express Briefs **68**(8), 853–899 (2021)

[50] So, H.C.: Unbiased Impulse Response Estimation in Nonstationary Noise. Electronics Letters, **35**(10), 791–792 (1999)

[51] Stankovic, L.J.: Performance Analysis of the Adaptive Algorithm for Bias-to-Variance Tradeoff. IEEE Transactions on Signal Processing **52**(5), 1228–1234 (2004)

[52] Tiglea, D.G., Candido, R., Silva, M.T.: A variable step size adaptive algorithm with simple parameter selection. IEEE Signal Processing Letters **29**, 1774–1778 (2022)

[53] Tingting, X., Lijuan, J., Shunshoku, K.: Bias-Compensated LMS Estimation for Adaptive Noisy FIR Filtering. In: Proceedings of 54th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE), 81–85 (2015)

[54] Vastola, K.: Threshold Detection in Narrow-Band Non-Gaussian Noise. IEEE Transactions on Communications **32**, 134–139 (1984)

[55] Wang, Y.: Channel Equalization Using a Robust Recursive Least-Squares Adaptive-Filtering Algorithm, In: Proceedings of the IEEE 12th International Conference on Computer and Information Technology, 135–138 (2012)

[56] Wen, P., Wang, B., Zhang, S., Qu, B., Song, X., Sun, J., Mu, X.: Bias-compensated augmented complex-valued NSAF algorithm and its low-complexity implementation. Signal Processing **204**, 108812 (2023)

[57] Zeidler, J.R.: Performance analysis of LMS adaptive prediction filters. Proceedings of the IEEE **78**(12), 1781–1806 (1990)

[58] Zhu, Y.-G., Li, Y.-G., Guan, S.-Y., Chen, Q.-S.: A novel variable step-size NLMS algorithm and its analysis. Procedia Engineering **29**, 1181–1185 (2012)

[59] Zode, P., Veena, M.B., Zode, P.: Design of Low Power High-Speed LMS Adaptive Filter For Biomedical Applications, In: Proceedings of the IEEE 9th International Conference for Convergence in Technology (I2CT), 1–4 (2024)