

# Adaptively biasing the weights of adaptive filters

Miguel Lázaro-Gredilla, Luis A. Azpicueta-Ruiz, Aníbal R. Figueiras-Vidal, *Senior Member, IEEE*, Jerónimo Arenas-García\*, *Member, IEEE*

## Abstract

It is a well-known result of estimation theory that biased estimators can outperform unbiased ones in terms of expected quadratic error. In steady-state, many adaptive filtering algorithms offer an unbiased estimation of both the reference signal and the unknown true parameter vector. In this correspondence, we propose a simple yet effective scheme for adaptively biasing the weights of adaptive filters using an output multiplicative factor. We give theoretical results that show that the proposed configuration is able to provide a convenient bias vs variance tradeoff, leading to reductions in the filter mean-square error, especially in situations with a low signal-to-noise ratio (SNR). After reinterpreting the biased estimator as the combination of the original filter and a filter with constant output equal to 0, we propose practical schemes to adaptively adjust the multiplicative factor. Experiments are carried out for the normalized least-mean-squares (NLMS) adaptive filter, improving its mean-square performance in stationary situations and during the convergence phase.

## Index Terms

Adaptive filters, biased estimation, bias-variance tradeoff, combination filters

## I. INTRODUCTION

Adaptive filters are nowadays widely used in many signal processing applications, such as system identification or channel equalization, among many others, due to their ability to track changing systems.

Copyright (c) 2010 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to [pubs-permissions@ieee.org](mailto:pubs-permissions@ieee.org).

The authors are with the Dept. of Signal Theory and Communications, Universidad Carlos III de Madrid, 28911 Leganés, Spain. Their work was partly supported by MEC project TEC2008-02473 and CAM project S-0505/TIC/0223.

Among the many existing algorithms for adaptive filtering, in this correspondence we consider those which try to predict the value of a reference signal, such as least-mean-squares (LMS), normalized LMS (NLMS), recursive least-squares (RLS), or those based on affine projections (AP).

No matter how sophisticated adaptive filtering algorithms become, they are always subject to a compromise regarding their speed of convergence and steady-state misadjustment. Using energy conservation arguments [1], it is possible to derive expressions for the steady-state error of many of the most widely-used schemes. It is also well-known that, under some mild assumptions, most adaptive filtering algorithms converge in the mean, i.e., they provide an asymptotically unbiased estimation of both the reference signal and the true unknown parameter vector [1].

In this correspondence, we propose to reduce the steady-state mean-square error (MSE) of adaptive filters by biasing their weights towards zero. In this way, the gain is obtained by exploiting a bias vs variance tradeoff, as it is customary to do in the estimation theory literature (see e.g., [2], [3]). Although more sophisticated approaches are possible and can potentially provide larger error reductions, here we will illustrate the idea with one of the simplest schemes one could think of: We will introduce a multiplicative factor in the range 0 to 1 at the output of the adaptive filter. This setup has been previously studied in the estimation framework [3], and the main contribution of this correspondence is the extension of this idea to the adaptive filtering context. A performance analysis of the proposed configuration will show that it can significantly reduce the error of the original filter, especially in low signal-to-noise ratio (SNR) scenarios. Therefore, the proposed scheme will offer improved robustness to the frequent situations in which the SNR is not known or changes over time.

By reinterpreting the proposed configuration as the adaptive combination of the original filter and a filter with constant output equal to zero, we can use several algorithms currently available in the literature for adaptive filter combination [4]–[9] to adapt the output multiplicative factor. Experiments are carried out to illustrate the validity of the idea, and how it can be used to improve the MSE of adaptive filters for a wide range of signal-to-noise ratios (SNRs) and adaptation speeds of the original filters.

The methods used in this paper are related to the implementation of adaptive amplitudes in the neural networks context [10], [11]. However, the motivation behind them is rather different. Whereas adaptive amplitudes in neural networks are introduced to facilitate the exploration of the error surface and to accelerate the training phase, the methods in this paper emerge as a natural way to adaptively bias the weight estimation of adaptive filters, thus reducing their MSE.

The rest of the correspondence is organized as follows: Next section introduces the configuration for biasing adaptive filters, and Section III carries out its steady-state analysis. Then, we present in Section

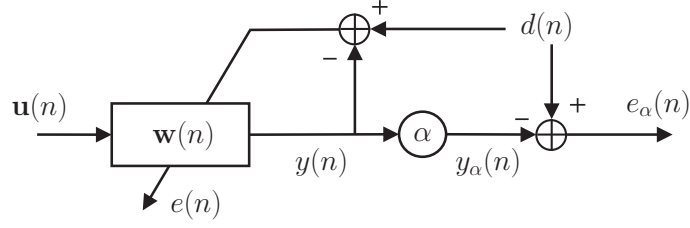


Fig. 1. Block diagram for the proposed configuration of adaptive filters with bias.

IV some realizable schemes for adaptively adjusting the value of the multiplicative factor. Experimental evidence on the effectiveness of the biased configuration is given in Section V. Finally, Section VI summarizes the main conclusions of this work and discusses some lines for further research.

## II. ADAPTIVE FILTERS WITH BIASED ESTIMATIONS

In this section we propose a configuration to bias the weights of adaptive filters. For general transversal schemes, we shall assume the following adaptation rule:

$$\mathbf{w}(n+1) = \mathbf{f}[\mathbf{w}(n), \mathbf{u}(n), d(n), \mathbf{q}(n)], \quad (1)$$

where  $\mathbf{w}(n)$  are the filter weights at iteration  $n$ ,  $\mathbf{u}(n)$  stands for the input regressor,  $d(n)$  is the reference signal,  $\mathbf{q}(n)$  is a state vector, and  $\mathbf{f}[\cdot]$  refers to the function which characterizes each particular adaptation algorithm. Let us also define the output and error of the filter as  $y(n) = \mathbf{w}^T(n)\mathbf{u}(n)$  and  $e(n) = d(n) - y(n)$ , respectively, where superscript  $T$  denotes vector transposition.

Consider now the configuration of Fig. 1, in which the output of the filter is multiplied by a constant  $\alpha \in [0, 1]$  to produce a modified estimator of the reference signal,  $y_\alpha(n) = \alpha y(n)$ . The resulting scheme can be considered as a new filter with weights  $\mathbf{w}_\alpha(n) = \alpha\mathbf{w}(n)$  and output error  $e_\alpha(n) = d(n) - y_\alpha(n)$ .

The introduction of the multiplicative factor sets a tradeoff regarding the bias and the variance of the filter weight error. To see this, let us denote the weight errors of the unbiased and proposed configurations by  $\tilde{\mathbf{w}}(n) = \mathbf{w}_o - \mathbf{w}(n)$  and  $\tilde{\mathbf{w}}_\alpha(n) = \mathbf{w}_o - \mathbf{w}_\alpha(n)$ , respectively,  $\mathbf{w}_o$  being the optimal stationary solution (the Wiener solution) [1]. The Mean Square Deviation (MSD) of the configuration in Fig. 1 is defined as the expected squared norm of  $\tilde{\mathbf{w}}_\alpha(n)$ , and it can be decomposed as:

$$E\{\|\tilde{\mathbf{w}}_\alpha(n)\|_2^2\} = E\{\tilde{\mathbf{w}}_\alpha^T(n)\}E\{\tilde{\mathbf{w}}_\alpha(n)\} + E\{[\tilde{\mathbf{w}}_\alpha(n) - E\{\tilde{\mathbf{w}}_\alpha(n)\}]^T[\tilde{\mathbf{w}}_\alpha(n) - E\{\tilde{\mathbf{w}}_\alpha(n)\}]\}, \quad (2)$$

where the first and second terms are associated, respectively, to the bias and the variance of the weight estimation. Assuming by now that, for sufficiently large  $n$ ,  $\mathbf{w}(n)$  provides an unbiased estimation of the optimal solution (i.e.,  $\lim_{n \rightarrow \infty} E\{\mathbf{w}(n)\} = \mathbf{w}_o$ ), it is clear that

$$\lim_{n \rightarrow \infty} E\{\tilde{\mathbf{w}}_\alpha(n)\} = (1 - \alpha)\mathbf{w}_o,$$

so that the proposed filter configuration is asymptotically biased for any  $\alpha \neq 1$ . Introducing this result in (2) we obtain

$$\lim_{n \rightarrow \infty} E\{\|\tilde{\mathbf{w}}_\alpha(n)\|_2^2\} = \underbrace{(1 - \alpha)^2 \|\mathbf{w}_o\|_2^2}_{\text{Bias Term}} + \alpha^2 \underbrace{\lim_{n \rightarrow \infty} E\{\|\tilde{\mathbf{w}}(n)\|_2^2\}}_{\text{Variance Term}}. \quad (3)$$

In the light of this expression, it is clear that the variance of the estimator can be reduced by decreasing  $\alpha$  at the cost of a larger bias. Thus, it makes sense to tackle the problem of finding the optimal  $\alpha$  to provide an estimator with minimum square error.

In the following section we analyze the MSE of the configuration in Fig. 1, and provide expressions for the optimal value of  $\alpha$ . We should mention here that the scheme we have just reviewed constitutes just one possibility (and maybe the simplest one) to establish a bias vs variance tradeoff in adaptive filters, and that more sophisticated approaches can be proposed, for instance to benefit from structured filter inputs.

### III. MEAN-SQUARE PERFORMANCE ANALYSIS

#### A. Stationary data model and definitions

In the sequel we adopt the following assumptions:

- AI.  $d(n)$  and  $\mathbf{u}(n)$  are related via a linear regression model,  $d(n) = \mathbf{w}_o^T \mathbf{u}(n) + e_0(n)$ , for some unknown weight vector  $\mathbf{w}_o$  of length  $M$ , and where  $e_0(n)$  is an independent and identically distributed (i.i.d.) noise with zero mean and variance  $\sigma_0^2$ , and independent of  $\mathbf{u}(m)$  for any  $n$  and  $m$ .
- AII. First and second order moments of the input regressors are  $E\{\mathbf{u}(n)\} = \mathbf{0}$  and  $E\{\mathbf{u}(n)\mathbf{u}^T(n)\} = \mathbf{R}$ .

It is also convenient to introduce some notation and additional variables. We define the signal-to-noise ratio in the reference signal as the quotient of the powers of the signal and noise components in  $d(n)$ ,

$$\text{SNR} = \frac{E\{[\mathbf{w}_o^T \mathbf{u}(n)]^2\}}{E\{e_0^2(n)\}} = \frac{\mathbf{w}_o^T \mathbf{R} \mathbf{w}_o}{\sigma_0^2}.$$

To measure filter performance it is customary to use the excess MSE (EMSE), which is defined as the excess over the minimum MSE that can be achieved by any filter of length  $M$ , namely  $\sigma_0^2$ . Rewriting the filter error as

$$e(n) = d(n) - y(n) = [\mathbf{w}_o - \mathbf{w}(n)]^T \mathbf{u}(n) + e_0(n) = e_a(n) + e_0(n), \quad (4)$$

where  $e_a(n)$  is the so-called *a priori* error of the filter, it can be easily seen that the EMSE of the original filter is given by  $J_{\text{ex}}(n) = E\{e_a^2(n)\}$ . The limiting value of the EMSE as  $n \rightarrow \infty$  will be

denoted as  $J_{\text{ex}}(\infty)$ . Expressions for the stationary steady-state EMSE of several well-known adaptive filtering algorithms, derived using energy conservation arguments, can be found in [1, p.269].

Similarly, we will denote in the sequel the *a priori* error and the EMSE of the biased scheme as  $e_{a,\alpha}(n) = e_\alpha(n) - e_0(n)$ , and  $J_{\text{ex},\alpha}(n) = E\{e_{a,\alpha}^2(n)\}$ , respectively.

### B. Steady-state performance

We study next the behavior of the proposed configuration. To keep derivations as simple as possible, while still illustrating the advantages of biasing adaptive filters, we restrict the analysis to the stationary case in which  $\mathbf{w}_o$  is kept unchanged. However, the analysis can be extended with some minor modifications to tracking scenarios using e.g. the random-walk model of [1, Eq. (20.13)].

To obtain an analytical expression for the EMSE of the configuration, we start by rewriting its error as

$$\begin{aligned} e_\alpha(n) &= d(n) - \alpha y(n) = \alpha[d(n) - y(n)] + (1 - \alpha)d(n) \\ &= \alpha[e_a(n) + e_0(n)] + (1 - \alpha)[\mathbf{w}_o^T \mathbf{u}(n) + e_0(n)] \\ &= [\alpha e_a(n) + (1 - \alpha)\mathbf{w}_o^T \mathbf{u}(n)] + e_0(n). \end{aligned} \quad (5)$$

Therefore, the term inside square brackets in the last line of the previous expression can be identified as the *a priori* error of the overall scheme,  $e_{a,\alpha}(n)$ . Squaring  $e_{a,\alpha}(n)$ , and taking expectations, we obtain

$$J_{\text{ex},\alpha}(n) = \alpha^2 J_{\text{ex}}(n) + (1 - \alpha)^2 \mathbf{w}_o^T \mathbf{R} \mathbf{w}_o + 2\alpha(1 - \alpha) \mathbf{w}_o^T E\{\mathbf{u}(n) \mathbf{u}^T(n) [\mathbf{w}_o - \mathbf{w}(n)]\}. \quad (6)$$

Note that this expression has been obtained without recurring to any assumptions, and is therefore valid both in steady-state and transient situations.

Next, we will consider how (6) simplifies in steady-state. To proceed further, we will use the standard assumption that, after the filter has completely converged,  $\mathbf{w}(n)$  and  $\mathbf{u}(n)$  are independent. Assuming also that the original filter provides an unbiased estimation of the optimal solution<sup>1</sup>, we have that

$$\lim_{n \rightarrow \infty} E\{\mathbf{u}(n) \mathbf{u}^T(n) [\mathbf{w}(n) - \mathbf{w}_o]\} = \mathbf{R} \left[ \lim_{n \rightarrow \infty} E\{\mathbf{w}(n)\} - \mathbf{w}_o \right] = \mathbf{0},$$

and the steady-state EMSE of the proposed scheme for biased adaptive filtering simplifies to

$$\boxed{J_{\text{ex},\alpha}(\infty) = \lim_{n \rightarrow \infty} J_{\text{ex},\alpha}(n) = \alpha^2 J_{\text{ex}}(\infty) + (1 - \alpha)^2 \mathbf{w}_o^T \mathbf{R} \mathbf{w}_o.} \quad (7)$$

<sup>1</sup>This asymptotic convergence in the mean condition holds for many of the most important adaptive filtering algorithms, when they operate under the conditions given in Subsec. III-A, and provided that the length of the adaptive filter  $\mathbf{w}(n)$  is at least  $M$ .

Comparing this expression to (3), we can match each of the two terms that compose the EMSE  $J_{\text{ex},\alpha}(\infty)$  to the bias and the variance incurred by the adaptive filter in the estimation of the optimal solution.

Finally, setting the derivative of (7) with respect to  $\alpha$  equal to zero, we find the optimal steady-state value of the multiplicative factor:

$$\alpha^*(\infty) = \frac{1}{1 + \frac{J_{\text{ex}}(\infty)}{\mathbf{w}_0^T \mathbf{R} \mathbf{w}_0}}. \quad (8)$$

Introducing this result back in (7) gives the minimum steady-state EMSE of the proposed configuration,  $J_{\text{ex},\alpha}^*(\infty)$ . We should emphasize that although (7) and (8) hold true only for adaptive filters providing unbiased estimations in steady-state, parameter  $\alpha$  can potentially provide advantages in other situations, since in the light of (6) it is clear that, for the optimal  $\alpha$ ,  $J_{\text{ex},\alpha}(n) \leq J_{\text{ex}}(n)$  will always hold.

Finally, it is interesting to have a deeper look at the situations that make  $\alpha$  approach its limiting values 0 and 1. From (8), it is clear that  $\alpha \rightarrow 1$  when  $J_{\text{ex}}(\infty) \ll \mathbf{w}_0^T \mathbf{R} \mathbf{w}_0$ . The opposite situation occurs for  $J_{\text{ex}}(\infty) \gg \mathbf{w}_0^T \mathbf{R} \mathbf{w}_0$ , with  $\alpha \rightarrow 0$ . For comparable  $J_{\text{ex}}(\infty)$  and  $\mathbf{w}_0^T \mathbf{R} \mathbf{w}_0$ ,  $\alpha$  will take intermediate values in the interval  $[0, 1]$ . Since for most adaptive filters  $J_{\text{ex}}(\infty)$  is proportional to  $\sigma_0^2$  [1], the situations where  $\alpha \rightarrow 1$  or  $\alpha \rightarrow 0$  correspond, respectively, to the cases with large and small SNRs in the reference signal.

#### IV. ADAPTIVE ADJUSTMENT OF THE MULTIPLICATIVE FACTOR

In this section we present two practical algorithms for obtaining the value of the multiplicative factor  $\alpha$ . It should be clear that the optimal value for  $\alpha$  can change over time, which makes evident the need for adaptive learning rules that adjust a time-varying parameter  $\alpha(n)$ . By doing so, the biased filter will show an improved performance when the SNR is unknown or time-varying.

The proposed scheme can be seen as the combination of the output of the original filter and a second filter with constant output equal to 0,

$$y_\alpha(n) = \alpha(n)y(n) + [1 - \alpha(n)] 0. \quad (9)$$

During the last years such adaptive combinations have received a lot of attention, and several algorithms have been proposed for learning  $\alpha(n)$  which, in this context, is usually referred to as a mixing parameter [4]–[9]. Here, we pay attention to the normalized rules proposed in [7] and [8], which are easier to adjust than their non-normalized counterparts from [5] and [9], especially for unknown or time-varying SNRs.

The normalized stochastic gradient rule proposed in [8] is given by<sup>2</sup>

$$\alpha(n+1) = \alpha(n) - \frac{\mu_\alpha}{p(n)} \frac{\partial e_\alpha^2(n)}{\partial \alpha(n)} = \alpha(n) + \frac{\mu_\alpha}{p(n)} e_\alpha(n) y(n), \quad (10)$$

<sup>2</sup>Note that this adaptation rule could also have been directly derived as a normalized LMS update to minimize  $e_\alpha^2(n)$  [4].

where  $p(n)$  is a low-pass filtered estimation of the power of  $y(n)$  given by  $p(n) = \beta p(n-1) + (1-\beta)y^2(n)$ . Selection of  $\beta$  is not critical for the appropriate performance of the algorithm, and we will simply set it to 0.9, as it was done in [7], [8].

Adaptation rule (10) allows  $\alpha(n)$  to lie outside the interval  $[0,1]$ , something which can be corrected, if necessary, by direct truncation. A more critical aspect is that the gradient noise introduced due to the adaptation of  $\alpha(n)$  will increase the final MSE of the overall structure with respect to the optimal value studied in the previous section, what can eventually result in a degradation of the performance of the original (asymptotically unbiased) adaptive filter  $\mathbf{w}(n)$ .

To keep  $\alpha(n)$  in the interval of interest and, more importantly, to reduce the gradient noise near  $\alpha(n) = 0$  and  $\alpha(n) = 1$ , we can proceed as in [7] and define  $\alpha(n)$  as the output of a sigmoid activation function,

$$\alpha(n) = \text{sgm}[a(n)] = \frac{1}{1 + \exp[-a(n)]}. \quad (11)$$

Then, at each iteration  $a(n)$  is adapted according to

$$a(n+1) = a(n) - \frac{\mu_a}{p(n)} \frac{\partial e_\alpha^2(n)}{\partial a(n)} = a(n) + \frac{\mu_a}{p(n)} e_\alpha(n) y(n) \frac{\partial \alpha(n)}{\partial a(n)}, \quad (12)$$

from which  $\alpha(n+1)$  is recovered using (11).

In practice,  $a(n)$  is truncated to an interval  $[-a^+, a^+]$ , so that adaptation rule (12) never gets stuck because of the derivative of the sigmoid being too close to 0. To guarantee that  $\alpha(n)$  can still reach all values within the interval of interest, in this correspondence we propose to use a slightly different activation for  $\alpha(n)$ ,

$$\alpha(n) = \frac{\text{sgm}[a(n)] - \text{sgm}[-a^+]}{\text{sgm}[a^+] - \text{sgm}[-a^+]}, \quad (13)$$

which is just a scaled and shifted version of the sigmoid. In this way,  $\alpha(n)$  attains values 0 and 1 for  $a(n) = -a^+$  and  $a(n) = a^+$ , respectively.

Fig. 2 depicts the above activation function as a function of  $a(n)$  for  $a^+ = 4$ . Since the derivative of the activation function enters (12), it can be seen that the speed of adaptation of  $\alpha(n)$  decreases when it approaches its limiting values. As a consequence, the gradient noise which propagates to  $y_\alpha(n)$  is almost negligible for  $\alpha(n) \approx 1$ , thus avoiding that the adaptation of the multiplicative factor degrades the performance of the original filter in this important case. Note that, in order to remove most of the gradient noise in this situation, the exact shape of the activation function is not critical as long as it keeps a small derivative near the extremes of  $\alpha(n)$ . In the experiments section we will illustrate this advantage of the non-linear adaptation rule with respect to the direct use of (10).

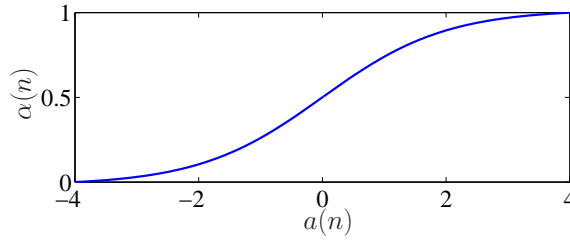


Fig. 2. Activation function for  $\alpha(n)$  with  $a^+ = 4$ .

The additional operations required for obtaining  $y_\alpha(n)$  and adapting  $\alpha(n)$  are 5 products, 1 division and 3 sums, if (10) is used. Therefore, the order of complexity of the algorithm is similar to that of the original (unbiased) filter. If (12) is used instead, it is also necessary to compute the output of the activation function and its derivative; alternatively, these values could also be read from a look-up table.

## V. EXPERIMENTS

In this section we will investigate the behavior of the biased version of NLMS as compared with its standard implementation. We have verified that analogous results are observed for other filtering schemes such as LMS or RLS.

### A. Steady-state behavior in stationary scenarios

First, we will analyze the behavior of the EMSE and multiplicative factor  $\alpha(n)$  in steady state. The experimental setup is as follows: The coefficients of a plant  $\mathbf{w}_o$  of length  $M = 16$  are randomly generated and scaled so that  $\mathbf{w}_o^T \mathbf{R} \mathbf{w}_o = \sigma_s^2$ , where  $\sigma_s^2$  is the power of the desired signal excluding the noise component. Input regressors  $\mathbf{u}(n)$  are obtained from an i.i.d. Gaussian random process  $u(n)$ , so that  $\mathbf{u}(n) = [u(n), u(n-1), \dots, u(n-M+1)]^T$ . The variance of  $u(n)$  is adjusted to get  $\text{tr}(\mathbf{R}) = 0.1$ . We will consider SNRs in the  $[-25, 15]$  range, which can be obtained by adjusting either the noise power  $\sigma_0^2$  or the scaling of  $\mathbf{w}_o$  (i.e., modifying  $\sigma_s^2$ ). We explore four logarithmically spaced step sizes for the NLMS algorithm:  $\mu = \{0.03, 0.1, 0.3, 1\}$ . Step sizes for the adaptation of the multiplicative factor using (10) and (12) have been fixed to get an appropriate behavior from each learning rule, selecting  $\mu_\alpha = 0.005$  and  $\mu_\alpha = 0.1$ , respectively. Each point in the following figures has been obtained as an average over 100 independent realizations, and over 20 000 iterations after the filters have completely converged.

Fig. 3 shows the steady state behavior of the standard NLMS filter and its biased counterpart when we fix the signal power to  $\sigma_s^2 = \text{tr}(\mathbf{R})$  and vary  $\sigma_0^2$  so that  $\text{SNR} \in [-25, 15]$ . The top right panel shows quite good agreement between the theoretically optimal value  $\alpha^*(\infty)$  from (8) and the average value that

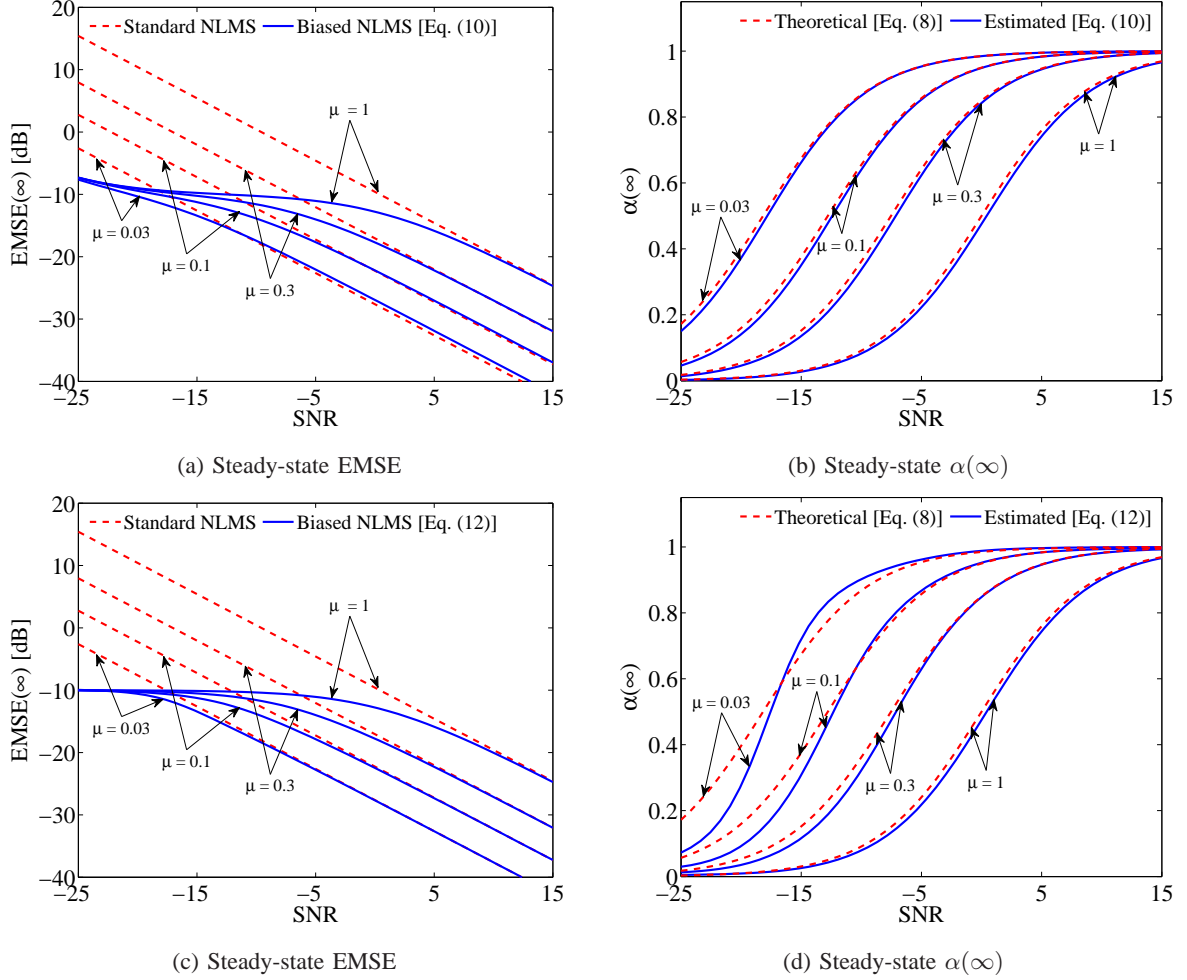


Fig. 3. Steady-state behavior of the standard and biased NLMS filters for different SNRs, obtained by fixing  $\sigma_s^2 = \text{tr}(\mathbf{R})$  and varying  $\sigma_0^2$ . Adaptive biasing is performed with rules (10) (top row) and (12) (bottom row). EMSEs are displayed in subplots (a) and (c), whereas theoretically optimal and averaged estimations of  $\alpha(\infty)$  are shown in subplots (b) and (d).

results from (10) after filter convergence. The top left panel shows that biasing the NLMS filter using (10) can indeed be very beneficial in terms of steady-state EMSE. This advantage appears consistently across a wide range of SNRs and step sizes, and is especially remarkable for low SNRs. Note that this results especially useful in the frequent situations in which the SNR is not known or changes over time. In spite of this general good performance, we observe that for large SNR and  $\mu = 0.03$  the unbiased scheme slightly degrades the performance of the original filter. This might be surprising, since the biased version could select  $\alpha(n) = 1$  and revert to standard NLMS. However, though we know from the top right panel that  $\alpha(\infty)$  is accurately determined *on average*, gradient noise is introduced by the adaptive estimation of  $\alpha(n)$ , thus degrading the EMSE of the biased scheme.

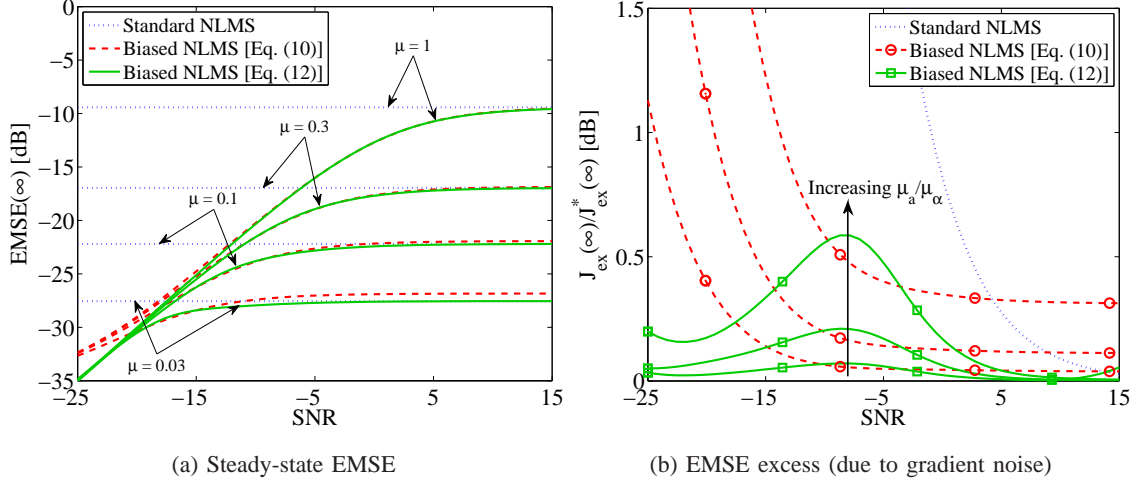


Fig. 4. Steady-state behavior of standard and biased NLMS filters for different SNRs, obtained by fixing  $\sigma_0^2 = 0.1$  and varying  $\sigma_s^2$ . Adaptive biasing is performed using rules (10) and (12). (a) EMSE of biased and unbiased NLMS. (b) Excess over the EMSE of a biased filter with optimal multiplicative factor  $\alpha^*(\infty)$ , for  $\mu = 0.3$  and different  $\mu_a = \{0.03, 0.1, 0.3\}$  and  $\mu_\alpha = \mu_a/0.05$ . The arrow indicates the direction of increasing step sizes  $\mu_a$  and  $\mu_\alpha$ .

Following our previous discussion in Section IV, we can circumvent this problem by regarding  $\alpha(n)$  as the output of the non-linear activation function (13), and adapting it according to (12). The superiority of this scheme is shown in Fig. 3(c), where the biased version of NLMS always performs better than the standard one, achieving large EMSE gains for low SNRs [even better than those obtained using (10)], and performing just like the original filter for high SNRs. The reduction of gradient noise in this case is critical, as we can conclude when comparing the results in Figs. 3 (a) and (c). Note that the improved behavior of (12) is obtained even though  $\alpha^*(\infty)$  estimation is not as accurate as in the previous case [compare Figs. 3 (b) and (d)] due to the asymmetric gradient noise introduced by the sigmoid.

For the second set of experiments we explore the same SNR values, but modifying the scaling of  $\mathbf{w}_0$  (thus  $\sigma_s^2$ ) instead, while keeping the noise power  $\sigma_0^2$  fixed to 0.1. Fig. 4(a) depicts the steady-state EMSE of biased and unbiased NLMS schemes in this case. Since the noise power is fixed, the EMSE of the standard NLMS remains constant, regardless of variations in the SNR. The performance of the biased schemes can be discussed in very similar terms to those used in the previous case.

To get a deeper insight into the reasons that justify the better behavior of learning rule (12) over (10), Fig. 4(b) plots the excess EMSE that results from the application of these schemes over the optimal EMSE that would be obtained with  $\alpha^*(\infty)$ . Thus, a level of 0 dB corresponds to an optimally biased filter, and any excess on top of that is due to gradient noise in the adaptation of  $\alpha(n)$ . This figure explains

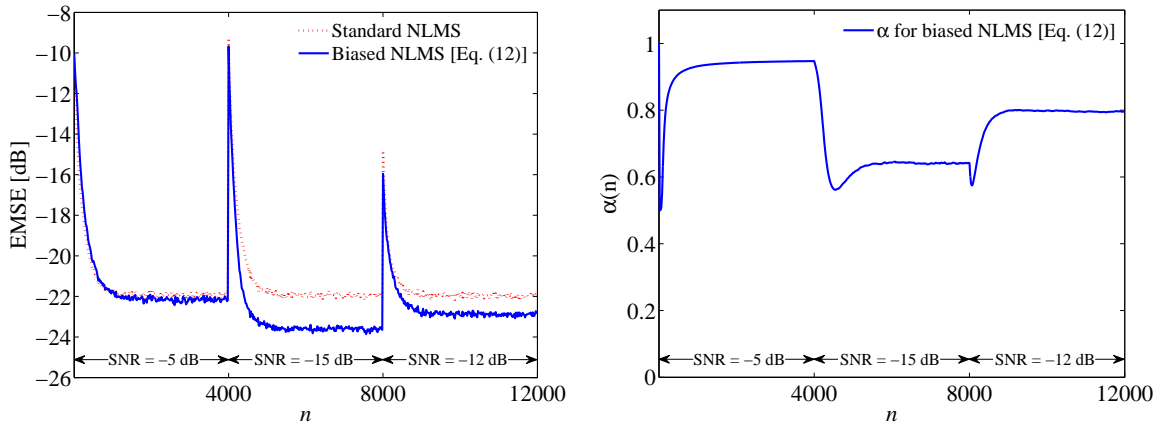


Fig. 5. EMSE and  $\alpha(n)$  time evolution for the biased and standard NLMS implementations for  $\mu = 0.1$ . Multiplicative factor  $\alpha(n)$  is adapted according to (12), with non-linear activation ( $\mu_a = 0.1$ ).

the positive effect of the sigmoid function at reducing the gradient noise when  $\alpha(n)$  is close to 0 or 1 (for small and large SNR, respectively), and shows that (12) keeps the EMSE below that of the standard NLMS for all SNRs. The figure illustrates also that, as expected, the gradient noise introduced by the adaptation of  $\alpha(n)$  grows for increasing  $\mu_a$  and  $\mu_\alpha$ .

The bottom line of this subsection is that biased NLMS is able to outperform standard NLMS in steady state in most situations, regardless of changes in the signal or noise powers. *This is especially useful in situations where the SNR is not known a priori or changes unpredictably.*

### B. Convergence behavior

Next, we will analyze how the NLMS filter converges when  $\mathbf{w}_o$  suffers an instantaneous change, both in the biased and standard configurations. The setup is as follows:  $u(n)$  is a colored sequence obtained from a first-order autoregressive model with transfer function  $\sqrt{1-h^2}/(1-hz^{-1})$ ,  $h = 0.6$ , fed with an i.i.d. Gaussian random process, whose variance is selected so that  $\text{tr}(\mathbf{R}) = 0.1$ . Plant coefficients are initially selected as in the previous subsection, and then changed at  $n = 4000$  and  $n = 8000$ . The three sets of coefficients have been scaled to obtain different SNRs in the reference signal (see Fig. 5), keeping  $\sigma_0^2 = 0.1$  constant during the simulation. NLMS weights are adapted with  $\mu = 0.1$ , while the output factor  $\alpha(n)$  follows (12) and (13), with  $\mu_a = 0.1$ . Ensemble-average curves are obtained from 5000 independent realizations.

The EMSE evolution depicted in Fig. 5(a) is a good example of the advantages that can be obtained with biased filters in scenarios where the SNR is time-varying. As illustrated in Fig. 5(a), both the standard and biased versions of NLMS re-converge at similar speeds after every perturbation in  $\mathbf{w}_o$ , but

the biased version achieves lower EMSE during all the simulation. This shows that for this value of  $\mu_a$  the steady-state EMSE reduction of the biased scheme is not obtained at the cost of a slower convergence. In principle, the speed of convergence could be degraded if a smaller  $\mu_a$  was used to reduce the gradient noise introduced by the estimation of  $\alpha(n)$ . Note, however, that according to Fig. 4(b) the gradient noise that appears for  $\mu_a = 0.1$  is already very small. In the right panel of Fig. 5,  $\alpha(n)$  evolution is plotted. Interestingly,  $\alpha(n)$  decreases towards zero at the beginning of each re-convergence, trying to cancel out the initial meaningless predictions of NLMS. It is only as NLMS starts recovering track of the filter weights that  $\alpha(n)$  heads towards its steady-state value.

## VI. CONCLUSIONS

Biasing the weights of adaptive filters can be an interesting way of reducing their MSE. In this correspondence, we have illustrated this idea with a very simple yet effective configuration, consisting in multiplying the filter output by a constant factor. Realizable schemes for adaptively learning this multiplicative factor are proposed, providing both theoretical and experimental evidence about the benefits of this approach.

We are currently working on more sophisticated biased schemes which exploit the structure of the input regressors, as well as on the extension of these ideas to filters operating in tracking situations.

## REFERENCES

- [1] A. H. Sayed, *Adaptive Filters*, Hoboken, NJ: Wiley, 2008.
- [2] Y. C. Eldar, "Uniformly improving the Cramér-Rao bound and maximumlikelihood estimation," *IEEE Trans. Signal Process.*, vol. 54, pp. 2943–2956, 2006.
- [3] S. Kay and Y. C. Eldar, "Rethinking biased estimation," *IEEE Signal Process. Mag.*, pp. 133–136, 2008.
- [4] S. S. Kozat and A. C. Singer, "Multi-stage adaptive signal processing algorithms," in *Proc. of SAM Signal Process. Workshop*, 2000, pp. 380–384.
- [5] J. Arenas-García, A. R. Figueiras-Vidal, and A. H. Sayed, "Mean-square performance of a convex combination of two adaptive filters," *IEEE Trans. Signal Process.*, vol. 54, pp. 1078–1090, 2006.
- [6] J. Arenas-García, M. Martínez-Ramón, A. Navia-Vázquez, and A. R. Figueiras-Vidal, "Plant Identification via Adaptive Combination of Transversal Filters", *Signal Processing*, vol. 86, pp. 2430-2438, 2006.
- [7] L. A. Azpicueta-Ruiz, A. R. Figueiras-Vidal, and J. Arenas-García, "A normalized adaptation scheme for the convex combination of two adaptive filters," in *Proc. ICASSP'08*, Las Vegas, NV, 2008, pp. 3301–3304.
- [8] R. Candido, M. T. M. Silva, V. H. Nascimento, "Affine combination of adaptive filters," in *Proc. 42nd Asilomar Conf. on Signal, Systems and Computers*, Oct. 2008, pp. 236–240.
- [9] N. J. Bershad, J. C. M. Bermudez, and J.-Y. Tourneret, "An affine combination of two LMS adaptive filters – transient mean-square analysis," *IEEE Trans. Signal Process.*, vol. 56, pp. 1853–1864, 2008.
- [10] E. Trentin, "Networks with trainable amplitude of activation functions," *Neural Networks*, vol. 14, pp. 471–493, 2001.
- [11] S. L. Goh and D. P. Mandic, "Recurrent neural networks with trainable amplitude of activation functions," *Neural Networks*, vol. 16, pp. 1095–1100, 2003.



**Miguel Lázaro-Gredilla** received the Telecommunication Engineer degree (with honors) from Universidad de Cantabria, Spain, in 2004, and his Diploma of Advanced Studies in the area of Signal Processing from Universidad Carlos III de Madrid, Spain, in 2007. After a short stay at the University of Cambridge, UK, he returned to Universidad Carlos III de Madrid, where he is currently a PhD candidate and a Teaching Assistant for the course of digital signal processing. His current research interests include Bayesian models, Gaussian processes, supervised large-scale learning, and the application of machine learning algorithms to adaptive signal processing.



**Luis A. Azpicueta-Ruiz** was born in Guadalajara, Spain, in 1978. He received the Telecommunication Engineer degree in 2004 from Universidad Politécnica de Madrid, Madrid, Spain. He is currently working towards the Ph.D. degree at the Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Madrid, where he is an Assistant Professor. His present research interests are focused in the fields of adaptive signal processing and their applications, mainly in audio and acoustic processing.



**Aníbal R. Figueiras-Vidal** (S'74-M'76-SM'84) received the Telecommunication Engineer degree (honors) from Universidad Politécnica de Madrid, Madrid, Spain, in 1973; and the Doctor degree (honors) from Universidad Politécnica de Barcelona, Barcelona, Spain, in 1976.

He is a Professor of Signal Theory and Communications at Universidad Carlos III, Madrid. His research interests are digital signal processing, digital communications, neural networks, and learning theory. He has (co)authored more than 300 journal and conference papers in these areas.

Dr. Figueiras-Vidal received an "Honoris Causa" Doctor degree from Universidad de Vigo, Vigo, Spain, in 1999. He is currently the President of the Royal Academy of Engineering of Spain.



**Jerónimo Arenas-García** (S'00-M'04) received the Telecommunication Engineer degree (with honors) from Universidad Politécnica de Madrid, Spain, in 2000, and the Ph.D. degree in Telecommunication Technologies (with honors) from Universidad Carlos III de Madrid, Leganés, Spain, in 2004. After a postdoc at the Technical University of Denmark, Lyngby, he returned to Universidad Carlos III de Madrid, where he is currently an Associate Professor of digital signal and information processing with the department of Signal Theory and Communications.

His current research interests include statistical learning, particularly in adaptive algorithms, advanced machine learning techniques, and MVA methods for feature extraction, and their applications, for instance in remote sensing data and multimedia information retrieval. Dr. Arenas-García became a member of the IEEE Machine Learning for Signal Processing TC in January 2009.