# Interference-Normalized Least Mean Square Algorithm

Jean-Marc Valin, *Member, IEEE*, and Iain B. Collings, *Senior Member, IEEE*

***Abstract*—An interference-normalized least mean square (INLMS) algorithm for robust adaptive filtering is proposed. The INLMS algorithm extends the gradient-adaptive learning rate approach to the case where the signals are nonstationary. In particular, we show that the INLMS algorithm can work even for highly nonstationary interference signals, where previous gradient-adaptive learning rate algorithms fail.**

***Index Terms*—Adaptive filtering, gradient-adaptive learning rate, normalized least mean square (NLMS) algorithm.**

## I. Introduction

THE choice of learning rate is one of the most important aspects of least mean square adaptive filtering algorithms as it controls the trade off between convergence speed and divergence in presence of interference.

In this letter, we introduce a new interference-normalized least mean square (INLMS) algorithm. In the same way as the NLMS algorithm introduces normalization against the filter input $x(n)$, our proposed INLMS algorithm extends the normalization to the interference signal $v(n)$. The approach is based on the gradient-adaptive learning rate class of algorithms [1]–[4], but improves upon these algorithms by being robust to nonstationary signals.

We consider the adaptive filter illustrated in Fig. 1, where the input signal $x(n)$ is convolved by an unknown $\mathbf{h}(n)$ filter (to produce $y(n)$) which has an additive interference signal $v(n)$, before being observed as $d(n)$. The adaptive filter attempts to estimate the impulse response $\hat{\mathbf{h}}(n)$ to be as close as possible to the real impulse response $\mathbf{h}(n)$ based only on the observable signals $x(n)$ and $d(n)$. The estimated convolved signal $\hat{y}(n)$ is subtracted from $d(n)$, giving an output signal $e(n)$ containing both the interference $v(n)$ and a residual signal $r(n) = y(n) - \hat{y}(n)$. In many scenarios, such as echo cancellation, the interference $v(n)$ is actually the signal of interest in the system.

The standard normalized least mean squares (NLMS) algorithm is given by

$$e(n) = d(n) - \hat{\mathbf{h}}^H(n-1)\mathbf{x}(n) \tag{1}$$

$$\hat{\mathbf{h}}(n) = \hat{\mathbf{h}}(n-1) + \frac{\mu(n)}{||\mathbf{x}(n)||^2} e^*(n)\mathbf{x}(n) \tag{2}$$

Manuscript received May 18, 2007; revised August 14, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Patrick A. Naylor.

The authors are with CSIRO ICT Centre, Marsfield, NSW, 2122, Australia (e-mail: jean-marc.valin@csiro.au; iain.collings@csiro.au).

Digital Object Identifier 10.1109/LSP.2007.908017

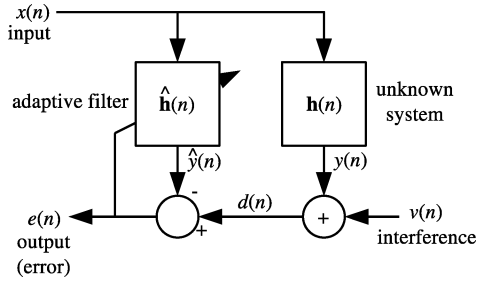


Fig. 1. Block diagram of echo cancellation system.

where $\mathbf{x}(n) = [x(n), x(n-1), \ldots, x(n-L+1)]^T$ and $\mu(n)$ is the learning rate. Here, we propose to extend this algorithm, by adaptively updating $\mu(n)$. By adopting our approach, we develop an algorithm which we call the INLMS algorithm and which works even for highly nonstationary interference signals, where previous gradient-adaptive learning rate algorithms fail.

Section II introduces existing gradient-adaptive learning rate algorithms and their limitations. Section III describes our proposed INLMS algorithm, followed by the results and discussion in Section IV. Section V concludes this letter.

## II. Gradient-Adaptive Learning Rate

Gradient-adaptive learning rate algorithms are based on the fact that when the adaptation rate is too small, the gradient tends to keep pointing in the same direction, while if it is too large, the gradient oscillates. Based on the behavior of the stochastic gradient, it is thus possible to infer whether the learning rate must be increased or decreased, and several methods have been proposed in the past to adjust the learning rate based on the gradient.

These methods each have a *control parameter* that is used to determine the learning rate. In the case of [1]–[3], the control parameter is the learning rate itself. In the generalized normalized gradient descent (GNGD) algorithm [4], the (normalized) learning rate is

$$\mu(n) = \frac{\mu_0 ||\mathbf{x}(n)||^2}{||\mathbf{x}(n)||^2 + \epsilon(n)} \tag{3}$$

where $\epsilon(n)$ is the control parameter. Because the control parameter is adapted based on the NLMS stochastic gradient behavior, it can only vary relatively slowly (typically requiring tens or hundreds of samples). For that reason, it is important for the optimal learning rate not to depend on rapid changes of the control parameter. We will show in the next section that none of the

methods cited above can fulfil this condition for nonstationary sources.

### A. Analysis for Nonstationary Signals

Under the assumption that $x(n)$ and $v(n)$ are zero-mean and uncorrelated to each other and that $v(n)$ is i.i.d., the theoretical optimal learning rate is equal to the residual-to-error ratio [5] as follows:

$$\mu_{opt}(n) = \frac{E\left\{r^2(n)\right\}}{E\left\{e^2(n)\right\}} \tag{4}$$

where $r(n) = y(n) - \hat{y}(n)$ is the (unknown) residual echo and $e(n)$ is the error signal. It turns out that although the assumption on $v(n)$ is not verified for speech, (4) nonetheless remains a good approximation. Earlier gradient-adaptive algorithms vary $\mu(n)$ directly as a response to the behavior of the gradient ($\mu(n)$ is the control parameter). It is a sensible thing to do if one assumes that $v(n)$ and $x(n)$ are stationary signals, because it means that both $E\{r^2(n)\}$ and $E\{e^2(n)\}$ vary slowly and, as a consequence, so does $\mu_{opt}(n)$. On the other hand, if the statistics of either $v(n)$ or $x(n)$ change abruptly, then the algorithm is not capable of changing $\mu(n)$ fast enough to prevent the adaptive filter from diverging.

The GNGD algorithm provides more robustness to nonstationarity. If we examine $\epsilon(n)$ more closely, it is reasonable to surmise that (3) eventually converges to the optimal learning rate defined by (4). Assuming steady-state behavior ($\epsilon(n)$ is stable) and $\mu_0 = 1$, we find (by multiplying the left-hand-side numerator and denominator by $\gamma(n)$) that

$$\frac{E\left\{r^2(n)\right\}}{E\left\{r^2(n)\right\} + \gamma(n)\epsilon(n)} = \frac{E\left\{r^2(n)\right\}}{E\left\{e^2(n)\right\}} \tag{5}$$

where $\gamma(n) = E\{r^2(n)\}/\|\mathbf{x}(n)\|^2$ is analogous to the filter misalignment. Assuming that $r(n)$ and $v(n)$ are zero-mean and uncorrelated to each other, we have $E\{e^2(n)\} = E\{r^2(n)\} + E\{v^2(n)\}$, which results in the relation $\epsilon(n) = E\{v^2(n)\}/\gamma(n)$. In other words, the optimal value for the gradient-adaptive parameter $\epsilon(n)$ depends on the filter misalignment and on the variance of the interference signal, but is independent of the variance of the input signal. Because $\epsilon(n)$ can only be adapted slowly over time, there is an implicit assumption in (3) that $E\{v^2(n)\}$ also varies slowly. While this is a reasonable assumption in some applications, it does not hold for scenarios like echo cancellation, where the interference is speech (double-talk) that can start or stop at any time.

## III. Proposed Algorithm for Nonstationary Signals

In previous work [5], we proposed to use (4) directly to adapt the learning rate. While $E\{e^2(n)\}$ can easily be estimated, the estimation of the residual echo $E\{r^2(n)\}$ is difficult because one does not have access to the real filter coefficients. One reasonable assumption we can make is that

$$\begin{aligned} E\left\{r^2(n)\right\} &= \eta(n) E\left\{y^2(n)\right\} \\ &\approx \eta(n-1) E\left\{\hat{y}^2(n)\right\} \end{aligned} \tag{6}$$

where $\eta(n)$ is a form of normalized filter misalignment, and is easier to estimate than $E\{r^2(n)\}$ directly, because it is assumed to vary slowly as a function of time. Although it is possible to estimate $\eta(n)$ directly through linear regression, the estimation remains a difficult problem.

In this letter, we propose to apply a gradient adaptive approach using $\eta(n)$ as the control parameter. By substituting (6) into (4), we obtain the learning rate as follows:

$$\mu(n) = \min\left(\eta(n-1)\frac{\widehat{\sigma_{\hat{y}}^2}(n)}{\widehat{\sigma_e^2}(n)}, 1\right) \tag{7}$$

where $\widehat{\sigma_{\hat{y}}^2}(n)$ and $\widehat{\sigma_e^2}(n)$ are, respectively, the estimates for $E\{\hat{y}^2(n)\}$ and $E\{e^2(n)\}$ and the upper bound imposed by the $\min(\cdot)$ reflects the fact that the optimal learning rate can never exceed unity.

### A. Adaptation

In this letter, we bypass the difficulty of estimating $\eta(n)$ directly and instead propose a closed-loop gradient adaptive estimation of $\eta(n)$. The parameter $\eta(n)$ is no longer an estimate of the normalized misalignment, but is instead adapted in closed-loop in such a way as to achieve a fast convergence of the adaptive filter.

As with other gradient-adaptive methods, we update the control parameter $\eta(n)$ by computing the derivative of the squared error $\mathcal{E}(n) = (1/2)e^*(n)e(n)$, this time with respect to $\eta(n-1)$, using the chain derivation rule without the independence assumption [1], [6]

$$\begin{aligned} \frac{\partial \mathcal{E}(n)}{\partial \eta(n-1)} &= \frac{1}{2}\left(\frac{\partial e^*(n)}{\partial \eta(n-1)}e(n) + e^*(n)\frac{\partial e(n)}{\partial \eta(n-1)}\right) \\ &= -\Re\left\{\frac{e(n)\mathbf{x}^H(n)\boldsymbol{\psi}(n-1)}{\|\mathbf{x}(n)\|^2}\right\}\frac{\widehat{\sigma_{\hat{y}}^2}(n)}{\widehat{\sigma_e^2}(n)} \end{aligned} \tag{8}$$

where

$$\boldsymbol{\psi}(n) = \left[\mathbf{I} - \frac{\mu(n)\mathbf{x}(n)\mathbf{x}^H(n)}{\|\mathbf{x}(n)\|^2}\right]\boldsymbol{\psi}(n-1) + \mathbf{x}(n)e^*(n) \tag{9}$$

is a smoothed version of the gradient. We further rewrite the update of $\boldsymbol{\psi}(n)$ in (9) as

$$\boldsymbol{\psi}(n) = \boldsymbol{\psi}(n-1) - \frac{\mu(n)}{\|\mathbf{x}(n)\|^2}\mathbf{x}(n)\left[\mathbf{x}^H(n)\boldsymbol{\psi}(n-1)\right] + \mathbf{x}(n)e^*(n) \tag{10}$$

so that it does not require a matrix-by-vector multiplication.

Based on this derivative, we propose the following exponential update of $\eta(n)$. We propose to use an exponential update in place of a more standard additive update since the misalignment has a large dynamic range; and we want the step size to scale with the value of $\eta(n)$. The exponential update is given as follows:

$$\eta(n) = \eta(n-1)\exp\left(\frac{\rho}{\widehat{\sigma_e^2}(n)}\frac{\partial \mathcal{E}(n)}{\partial \eta(n-1)}\right) \tag{11}$$

$$
\begin{aligned}
\hat{y}(n) &= = \hat{\mathbf{h}}(n-1)\,\mathbf{x}(n) \\
e(n) &= d(n) - \hat{y}(n) \\
\widehat{\sigma_{\hat{y}}^2}(n) &= \min\left(\hat{E}_3\left\{|\hat{y}(n)|^2\right\}, \hat{E}_{10}\left\{|\hat{y}(n)|^2\right\}\right) \\
\widehat{\sigma_{e}^2}(n) &= \max\left(\hat{E}_1\left\{|e(n)|^2\right\}, \hat{E}_3\left\{|e(n)|^2\right\}, \hat{E}_{10}\left\{|e(n)|^2\right\}\right) \\
\mu(n) &= \min\left(\nu(n-1)\,\frac{\widehat{\sigma_{\hat{y}}^2}(n)}{\widehat{\sigma_{e}^2}(n)}, 1\right) \\
\hat{\mathbf{h}}(n) &= \hat{\mathbf{h}}(n-1) + \frac{\mu(n)}{\|\mathbf{x}(n)\|^2} e^*(n)\,\mathbf{x}(n) \\
\eta(n) &= \eta(n-1)\exp\left[\frac{\rho\widehat{\sigma_{\hat{y}}^2}(n)\Re\left\{e(n)\mathbf{x}^H(n)\boldsymbol{\psi}(n-1)\right\}}{\widehat{\sigma_{e}^2}(n)\|\mathbf{x}(n)\|^2\widehat{\sigma_{e}^2}(n)}\right] \\
\boldsymbol{\psi}(n) &= \boldsymbol{\psi}(n-1) - \frac{\mu(n)}{\|\mathbf{x}(n)\|^2}\mathbf{x}(n)\left[\mathbf{x}^H(n)\,\boldsymbol{\psi}(n-1)\right] \\
&\quad + e^*(n)\,\mathbf{x}(n)
\end{aligned}
$$

Fig. 2. Summary of the INLMS algorithm.

where $\rho$ is a step size and we have normalized the gradient $\partial\mathcal{E}(n)/\partial\eta(n-1)$ by $\widehat{\sigma_e^2}(n)$ to obtain a nondimensional value.

It remains to estimate $\widehat{\sigma_{\hat{y}}^2}(n)$ and $\widehat{\sigma_e^2}(n)$. For $\widehat{\sigma_e^2}(n)$, we have the following recursive estimator with time constant $N$ ($\widehat{\sigma_{\hat{y}}^2}(n)$ is estimated similarly)

$$\hat{E}_N\left\{|e(n)|^2\right\} = \left(1-\frac{1}{N}\right)\hat{E}_N\left\{|e(n-1)|^2\right\} + \frac{1}{N}|e(n)|^2. \tag{12}$$

The question then becomes what value of $N$ to use. To maximize stability, a conservative approach is to err on the side of picking the smallest $\widehat{\sigma_{\hat{y}}^2}(n)$ and the biggest $\widehat{\sigma_e^2}(n)$ out of the set of estimated obtained by varying $N$. For efficiency, we have chosen a subset of all possible $N$ values. The values $N = 3$ and $N = 10$ provide good short- to medium-term term estimation, though the algorithm is not very sensitive to the exact choice of $N$. For the estimation of $\widehat{\sigma_e^2}(n)$, we also include $N = 1$ to make sure that even an instantaneous onset of interference cannot cause the filter to diverge. The complete algorithm is summarized in Fig. 2.

The last aspect that needs to be addressed is the initial condition. When the filter is initialized, all the weights are set to zero ($\hat{\mathbf{h}}(0) = \mathbf{0}$), which means that $\hat{y}(n) = 0$ and no adaptation can take place in (7) and (8). In order to bootstrap the adaptation process, the learning rate $\mu(n)$ is set to a fixed constant (we use $\mu = 0.25$) for a short time [until (7) gives $\mu(n) > 0.1$]. This *ad hoc* procedure is only necessary when the filter is initialized and is not required in case of echo path change. In practice, any method that provides a small initial convergence can be used.

### B. Analysis

The adaptive learning rate described above is able to deal with both double-talk and echo path change without explicit modelling. From (7), we can see that when the interference changes abruptly, the denominator $\widehat{\sigma_e^2}(n)$ rapidly increases, causing an instantaneous decrease in the learning rate. In the case of a stationary interference, the learning rate depends on both the presence of an input signal and on the misalignment estimate. As the filter misalignment becomes smaller, the learning rate also becomes smaller. When the echo path changes, the gradient starts pointing steadily in the same direction, thus significantly increasing $\eta(n)$, which is a clear sign that the filter is no longer properly adapted.

In gradient adaptive methods [1]–[3], the implicit assumption is that both the near-end and the far-end signals are nearly stationary. We have shown that the GNGD algorithm [4] only requires the near-end signal to be nearly stationary. In the proposed INLMS method, both signals can be nonstationary, which is a requirement for double-talk robustness.

It should be noted that the per-sample complexity of the proposed algorithm only differs from the complexity of the "classic" algorithm in [1] by a constant ($O(1)$). For example, the total increase in complexity for the real-valued case is due to the increased cost of computing $\eta(n)$ and amounts to only 23 multiplications, five additions, two divisions, and one exponential. Considering that the algorithms have an $O(L)$ complexity ($L$ is the filter length), the difference is negligible for any reasonable filter length.

## IV. Results and Discussion

We compare three algorithms:

- **Direct** learning rate adaptation [1]
- Generalized normalized gradient descent (**GNGD**) [4]
- **INLMS** algorithm (proposed)

In each case, we use a 32-s test sequence sampled at 8 kHz with an abrupt change in the unknown system $\mathbf{h}(n)$ at 16 s. The impulse responses are taken from ITU-T recommendation G.168 (impulse responses D.7 and D.9) and the filter length is 128 samples (16 ms). We choose $\rho = 0.005$ since it gave good results over a wide range of operating conditions ($\rho$ should be inversely proportional to the filter length). To make the comparison fair, we also used the exponential update for the Direct method ($\rho = 0.0005$ gave the best results) and the GNGD method ($\rho = 0.005$ gave the best results).

We test the algorithms for three scenarios:

1) Both the input $x(n)$ and the interference $v(n)$ are white Gaussian noise (see Fig. 3).
2) The input $x(n)$ is speech and the interference $v(n)$ is white Gaussian noise (see Fig. 4).
3) Both the input $x(n)$ and interference $v(n)$ are speech (see Fig. 5) with frequent overlap (double-talk).

The normalized misalignment is defined as

$$\Lambda(n) = \left\|\hat{\mathbf{h}}(n) - \mathbf{h}(n)\right\|^2 / \|\mathbf{h}(n)\|^2 \tag{13}$$

In Fig. 3, we can see that all three algorithms successfully converge, albeit with differing convergence rates. In Fig. 4, it can be observed that the direct algorithm fails to converge for scenario 2 where the input is nonstationary, but GNGD and INLMS perform well. Finally, in Fig. 5, we see that when the interference is nonstationary, only the proposed INLMS algorithm performs well.

## V. Conclusion

We have proposed a new interference-normalized least mean square (INLMS) algorithm, based on the gradient-adaptive

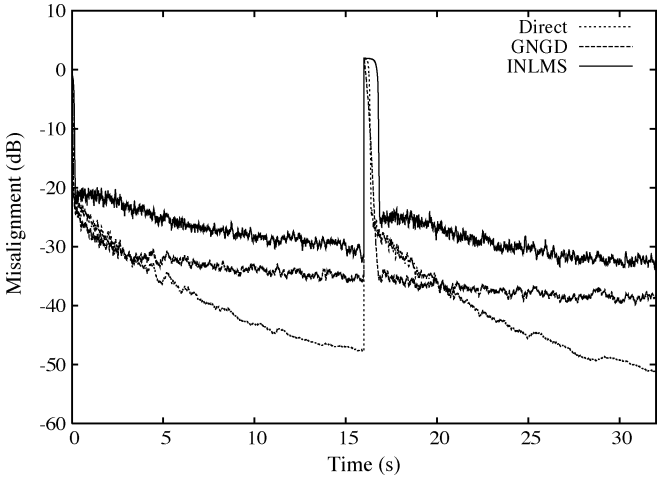


Fig. 3. Normalized misalignment for white Gaussian input and interference (scenario 1) with an abrupt change in the unknown system $\mathbf{h}(n)$ at 16 s. All algorithms converge.

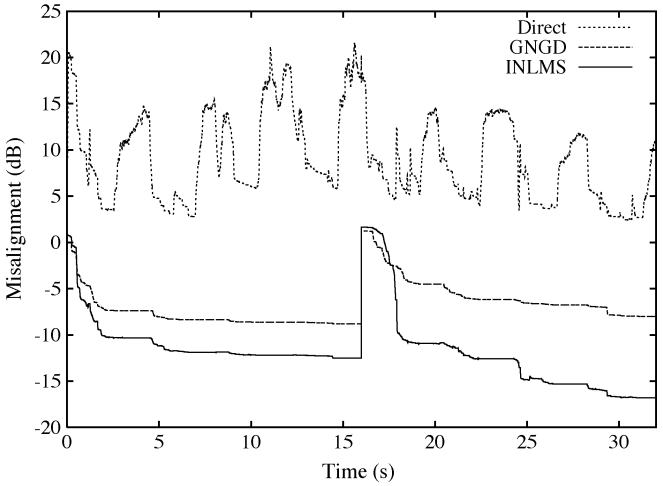


Fig. 4. Normalized misalignment for typical speech input and white Gaussian interference (scenario 2) with an abrupt change in the unknown system $\mathbf{h}(n)$ at 16 s. The direct algorithm diverges.

learning rate class of algorithms. We have demonstrated that unlike other gradient-adaptive methods, it is robust to nonstationarity of both the input and interference signals. This

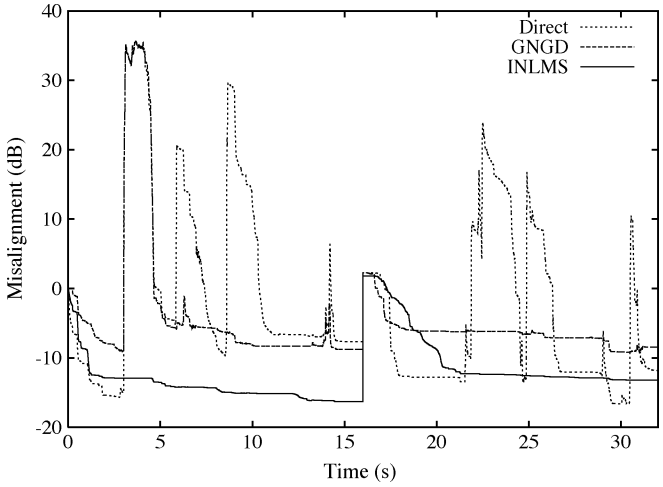


Fig. 5. Normalized misalignment for typical speech input and interference (scenario 3) with an abrupt change in the unknown system $\mathbf{h}(n)$ at 16 s. Note that the direct method frequently diverges, and the GNGD method diverges less often, but still significantly between 3–4 s, at 6 s, 14 s, and 31 s. All the divergence events are due to double-talk, except at 3–4 s, where only interference is present.

robustness is achieved by using a control parameter whose optimal value is independent of the power of the input and interference signals and instead depends only on the filter misalignment. This allows the instantaneous learning rate to react very quickly even though the control parameter cannot.

## REFERENCES

[1] Benveniste, Adaptive Algorithms and Stochastic Approximation, 1990.

[2] V. Mathews and Z. Xie, "A stochastic gradient adaptive filter with gradient adaptive step size," *IEEE Trans. Signal Process.*, vol. 41, no. 6, pp. 2075–2087, Jun. 1993.

[3] W.-P. Ang and B. Farhang-Boroujeny, "A new class of gradient adaptive step-size LMS algorithms," *IEEE Trans. Signal Process.*, vol. 49, no. 4, pp. 805–810, Apr. 2001.

[4] D. Mandic, "A generalized normalized gradient descent algorithm," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 115–118, Feb. 2004.

[5] J.-M. Valin, "On adjusting the learning rate in frequency domain echo cancellation with double-talk," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1030–1034, Mar. 2007.

[6] S. Goh and D. Mandic, "A class of gradient-adaptive step size algorithms for complex-valued nonlinear neural adaptive filters," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2005, vol. V, pp. 253–256.