Convergence of Exponentiated Gradient Algorithms

Simon I. Hill and Robert C. Williamson*

Abstract—**This paper studies three related algorithms: the (traditional) Gradient Descent (GD) Algorithm, the Exponentiated Gradient Algorithm with Positive and Negative weights (EG± algorithm) and the Exponentiated Gradient Algorithm with Unnormalized Positive and Negative weights (EGU± algorithm). These algorithms have been previously analyzed using the "mistake-bound framework" in the computational learning theory community. In this paper we perform a traditional signal processing analysis in terms of the mean square error.**

A relationship between the learning rate and the mean squared error (MSE) of predictions is found for the family of algorithms. This is used to compare the performance of the algorithms by choosing learning rates such that they converge to the same steady state MSE. We demonstrate that if the target weight vector is sparse, the EG± algorithm typically converges more quickly than the GD or EGU± algorithms which perform very similarly. A side effect of our analysis is a reparametrization of the algorithms that provides insights into their behavior. The general form of the results we obtain are consistent with those obtained in the mistake-bound framework [1].

The application of the algorithms to acoustic echo cancellation is then studied and it is shown in some circumstances that the EG± algorithm will converge faster than the other two algorithms.

Keywords— **Exponentiated Gradient Descent, Learning rate, Mean Squared Error, Echo, Room acoustics, Acoustic echo cancellation.**

I. Introduction

THERE are many variants of the classical LMS algorithm (see e.g. [2], [3]) but new ones continue to appear. The reason for the variants is that the problems to which the algorithms are applied vary greatly. In this paper we will study two algorithms new to the signal processing community which appear to offer improved performance under situations where the unknown target weight vector is sparse. In contrast to other approaches to improving the convergence speed of LMS type algorithms in such situations (see e.g. [4], [5] and references therein) the algorithms we study are simply a nonlinearly reparametrized standard LMS algorithm.

The present paper derives a relationship between learning rate and steady-state performance for a family of on-line learning algorithms. This allows comparison of the performances of algorithms in that family by observing their relative speeds of convergence to a desired final accuracy as they learn various targets. Sample comparisons are made with the algorithms being used for simulated acoustic echo cancellation.

At the beginning of the tth trial, an a k-length vector $\mathbf{x}_t$ is input to the algorithm. The algorithm maintains another k-length vector of weights, $\mathbf{w}_t$, the subscripts in both cases indicating that the vectors are current for that time interval, t. The ith entry in the weight vector will be denoted $w_{t,i}$ and similarly for other vectors. The algorithm employed tries to predict the output of the system whose characteristics are to be learned. The actual outputs are denoted y_t. Predictions of output are denoted $\hat{y}_t$ and are made by $\hat{y}_t = \mathbf{w}_t \cdot \mathbf{x}_t$. Updates to the weight vector are made based on the accuracy of the predictions, this is measured by the loss function, $L(y_t, \hat{y}_t) = (y_t - \hat{y}_t)^2$. It is assumed that there is some optimal predictor, $\mathbf{u}$ which remains constant and which has a minimal cumulative loss function. It is the aim of an on-line learning algorithm to learn $\mathbf{u}$.

Different algorithms use different methods to update the weight vector. Three such algorithms are considered in this paper; the Gradient Descent (GD) Algorithm, the Exponentiated Gradient Algorithm with Unnormalized Positive and Negative Weights (EGU± Algorithm) and the Exponentiated Gradient Algorithm with Positive and Negative Weights (EG± Algorithm). As described in [1], the GD algorithm takes the form

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla_{\mathbf{w}_t} L(y_t, \hat{y}_t) \tag{1}$$

or, with the given definition of $L(\cdot,\cdot)$,

$$\mathbf{w}_{t+1} = \mathbf{w}_t - 2\eta(\hat{y}_t - y_t)\mathbf{x}_t. \tag{2}$$

In these expressions η is the *learning rate*. The greater η, the more the algorithm adjusts to discrepancies between a prediction and an actual result, hence the more the algorithm learns from a particular trial.

Kivinen and Warmuth [1] also introduce and describe the derivation of the EG± algorithm which requires the maintenance of two vectors, $\mathbf{w}_t^+$ and $\mathbf{w}_t^-$. Updates of these vectors are made by,

$$w_{t+1,i}^+ = U \frac{w_{t,i}^+ r_{t,i}^+}{\sum_{j=1}^k (w_{t,j}^+ r_{t,j}^+ + w_{t,j}^- r_{t,j}^-)}, \quad i = 1, \ldots, k \tag{3}$$

$$w_{t+1,i}^- = U \frac{w_{t,i}^- r_{t,i}^-}{\sum_{j=1}^k (w_{t,j}^+ r_{t,j}^+ + w_{t,j}^- r_{t,j}^-)} \quad i = 1, \ldots, k, \tag{4}$$

where,

$$r_{t,i}^+ = e^{-\eta \nabla_{\mathbf{w}_t} L(y_t, \hat{y}_t)} \text{ and } r_{t,i}^- = e^{\eta \nabla_{\mathbf{w}_t} L(y_t, \hat{y}_t)},$$

U is a constant chosen such that $U \geq \|\mathbf{u}\|_1$ and for $p \geq 1$, $\left(\|\mathbf{u}\|_p = \sqrt[p]{\sum_{i=1}^k |u_i|^p}\right)$. The vectors $\mathbf{w}_{t+1}^+$ and $\mathbf{w}_{t+1}^-$ combine to give $\mathbf{w}_{t+1} = \mathbf{w}_{t+1}^+ - \mathbf{w}_{t+1}^-$.

The derivation of an algorithm known as the Exponentiated Gradient algorithm with unnormalized weights (EGU Algorithm) is also described in [1]. This algorithm maintains a vector of non-negative weights. By maintaining two such vectors, as for the EG± algorithm, a more generally useful algorithm is created; the EGU± algorithm. Updates of $\mathbf{w}_t = \mathbf{w}_t^+ - \mathbf{w}_t^-$ are made through

$$w_{t+1,i}^+ = w_{t,i}^+ e^{-\eta \nabla_{\mathbf{w}_t} L(y_t, \hat{y}_t)} \text{ and } w_{t+1,i}^- = w_{t,i}^- e^{\eta \nabla_{\mathbf{w}_t} L(y_t, \hat{y}_t)}. \tag{5}$$

This work was supported by the Australian Research Council.

Simon Hill was with the Research School of Information Science and Engineering, Australian National University. He is now with the Department of Engineering, Cambridge University, Cambridge, UK. E-mail: simon.hill@bigfoot.com .

Robert Williamson is with the Department of Engineering, Australian National University, Canberra, 0200 Australia. E-mail: bob.williamson@anu.edu.au .

It is the aim to relate the learning rate of algorithms to their final accuracy. The measure of final accuracy used is the mean squared error (MSE) ($= \langle (y_t - \hat{y}_t)^2 \rangle$ where $\langle \cdot \rangle$ signifies time average) of predictions. As the loss function is defined to be the squared error then, by definition, the MSE obtained from predictions made by $\mathbf{u}$ is a minimum, denoted ξ_{min}. The MSE of an algorithm after initial convergence is affected by noise and choice of learning rate. Noise will cause elements of the weight vector, $\mathbf{w}_t$, to jiggle about elements of $\mathbf{u}$. The higher the learning rate the more influence the noise will have hence the higher the MSE.

By assuming η is very small it is possible to derive an (approximate) theoretical relationship between MSE and η. This was done in the 1970s for the GD algorithm [6], [7]. The final relationship takes the form,

$$\text{MSE}_{GD} = \xi_{GD} = \xi_{min}(1 + \eta \text{Tr}(R)) \tag{6}$$

where,

$$R = E[\mathbf{x}_t \mathbf{x}_t^T] \tag{7}$$

and $\text{Tr}(R)$ denotes the trace of R.

This paper generalizes, with some constraints, the above result across to all three algorithms. Trials are then conducted with the algorithms learning in identical environments to the same final accuracy. These trials demonstrate that the EG$\pm$ algorithm learns more quickly than the other two, given a target vector predominately made up of zeroes.

Thus the general conclusion of the analysis in [1] is recovered albeit in a different framework.

II. Algorithm Reparameterisation

The GD, EGU$\pm$ and EG$\pm$ algorithms can be shown to have some distinctive similarities in their updates. To see the similarities requires reparameterisation of the algorithms. Reparameterisation involves the introduction of a new k-length vector $\mathbf{z}_t$ and a *parameterisation function*, $\psi(\cdot)$, through which $\mathbf{w}_t$ is related to $\mathbf{z}_t$,

$$\mathbf{w}_t = \psi(\mathbf{z}_t). \tag{8}$$

For each algorithm considered in this paper $\psi(\cdot)$ can be chosen such that, in each case, updates to $\mathbf{z}_t$ take the form

$$\mathbf{z}_{t+1} = \mathbf{z}_t - \eta \nabla_{\mathbf{w}_t} L(y_t, \hat{y}_t) = \mathbf{z}_t - 2\eta(\hat{y}_t - y_t)\mathbf{x}_t. \tag{9}$$

The case of the GD algorithm is a trivial one: $\psi_{GD}(\mathbf{z}_t) = \mathbf{z}_t$.

Reparameterisation, as described, of the EGU$\pm$ algorithm cannot immediately be done due to the complete independence of starting conditions on the vectors $\mathbf{w}_t^+$ and $\mathbf{w}_t^-$. Consider the imposition of the initial constraint

$$w_{t,i}^+ = \frac{1}{w_{t,i}^-} \quad i = 1, \ldots, k.$$

Then, for some $\mathbf{z}_{t,i}$,

$$w_{t,i}^+ = e^{z_{t,i}} \text{ and } w_{t,i}^- = e^{-z_{t,i}} \ \forall i,$$

hence, from equation (5)

$$w_{t+1,i} = 2\sinh(z_{t,i} - \eta \nabla_{\mathbf{w}_t} L(y_t, \hat{y}_t)) = 2\sinh(z_{t+1,i})$$

(recall $\sinh(x) = \frac{1}{2}(e^x - e^{-x})$ and $\cosh(x) = \frac{1}{2}(e^x + e^{-x})$). By induction the parameterisation function under this constraint is given by

$$w_{t,i} = [\psi_{EGU\pm}(\mathbf{z}_t)]_i = 2\sinh(z_{t,i}) \ \forall i, t. \tag{10}$$

The EG$\pm$ algorithm can be reparameterised also. The comments made regarding the independence of $\mathbf{w}_t^+$ and $\mathbf{w}_t^-$ apply equally when considering the EG$\pm$ case. For this algorithm, suppose that there is some vector $\mathbf{z}_t$ such that

$$w_{t,i}^+ = U\frac{e^{z_{t,i}}}{2\sum_{j=1}^k \cosh(z_{t,j})} \text{ and } w_{t,i}^- = U\frac{e^{-z_{t,i}}}{2\sum_{j=1}^k \cosh(z_{t,j})}.$$

From this and equations (3) and (4) then

$$w_{t+1,i}^+ = U\frac{\sinh(z_{t+1,i})}{\sum_{j=1}^k \cosh(z_{t+1,j})}. \tag{11}$$

By induction the parameterisation function is given by

$$w_{t,i} = [\psi_{EG\pm}(\mathbf{z}_t)]_i = U\frac{\sinh(z_{t,i})}{\sum_{j=1}^k \cosh(z_{t,j})}. \tag{12}$$

For both the EG$\pm$ and the EGU$\pm$ algorithms future discussion will assume that the described parameterisations are possible, i.e. the initial condition constraints given hold. A suitable starting point for the EGU$\pm$ algorithm is $w_{0,i}^+ = w_{0,i}^- = 1$ for all i. This implies $\mathbf{z}_0 = \mathbf{0}$. For the EG$\pm$ algorithm $w_{0,i}^+ = w_{0,i}^- = \frac{U}{2k}$, $\mathbf{z}_0 = \mathbf{0}$ is the starting estimation used in [1] and also satisfies the applicable initial conditions. From these results it is clear that the three (possibly constrained) algorithms belong to a greater family of algorithms, all of which can be seen to have an additive update at their heart.

The reparametrisation used is similar in spirit, but not in detail, to the general link functions in [8] (cf. [9]).

III. MSE / Learning Rate Relationships

In realizing a generalized expression for relating learning rate to MSE first consider continuous time on-line learning algorithms. Recall the form of the updates to $\mathbf{z}_t$,

$$\mathbf{z}_{t+1} = \mathbf{z}_t - \eta \nabla_{\mathbf{w}_t} L(y_t, \hat{y}_t). \tag{13}$$

A continuous time equivalent to this exists, [10]. In this case t is a continuous time index, updates take the form

$$\dot{\mathbf{z}}_t = -\eta \nabla_{\mathbf{w}_t} L(y_t, \hat{y}_t). \tag{14}$$

An alternative way of considering discrete on-line learning algorithms is to consider them found by replacing $\dot{\mathbf{z}}_t$ by its Euler discretisation, $(\mathbf{z}_{t+h} - \mathbf{z}_t)/h$. With the usual discrete time notation then equation (14) becomes (13).

The technique of using an Euler discretisation to derive an additive discrete time algorithm from a continuous time expression can be applied to a general case of the family of algorithms just

discussed. Consider the original starting point, equation (14). Multiplying both sides by the Jacobian $J(\psi(\mathbf{z}_t))$ gives

$$\dot{\mathbf{w}}_t = -\eta J(\psi(\mathbf{z}_t))\nabla_{\mathbf{w}_t} L(y_t, \hat{y}_t).$$

Using an Euler discretisation gives

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta J(\psi(\mathbf{z}_t))\nabla_{\mathbf{w}_t} L(y_t, \hat{y}_t), \tag{15}$$

an additive approximation to an update of a general algorithm. As previously mentioned, a relationship between learning rate and MSE exists for the GD algorithm (which is of the same form as (15) with the Jacobian replaced by the identity matrix). The derivation of this relationship can be generalized across algorithms with diagonal Jacobians provided the elements of the input, $\mathbf{x}_t$ are uncorrelated and zero mean. The general relationship between learning rate and MSE is presented in theorem 1 (the proof of which is in the appendix).

Theorem 1: Consider an on-line learning algorithm which can be reparameterised through $\mathbf{w}_t = \psi(\mathbf{z}_t)$ and $\mathbf{z}_{t+1} = \mathbf{z}_t - \eta\nabla_{\mathbf{w}_t} L(y_t, \hat{y}_t)$. Let the input to the algorithm $\mathbf{x}_t$ be made up of uncorrelated zero mean elements $x_{t,i}$. Let

$$\begin{aligned} \mathbf{z}^* &= \lim_{\mathbf{w}\to\mathbf{u}} \big(\psi^{-1}(\mathbf{w})\big) \\ R &= \mathrm{E}[\mathbf{x}_t \mathbf{x}_t^T] \end{aligned}$$

where $\mathbf{w}_t \to \mathbf{u}$ means $\|\mathbf{w}_t - \mathbf{u}\|_2 \to 0$. If the Jacobian $J(\psi(\mathbf{z}^*))$ is diagonal then the mean squared error (ξ) of that algorithm is given in the learning rate limit by

$$\lim_{\eta\to 0}(\xi) = \xi_{min} + \eta\xi_{min}\mathrm{Tr}\big(J(\psi(\mathbf{z}^*))R\big). \tag{16}$$

Of the three algorithms considered in this paper, the GD and EGU± algorithms have diagonal Jacobians. For the GD algorithm the Jacobian is the identity matrix and equation (16) reduces to the familiar expression in equation (6),

$$\mathrm{MSE}_{GD} = \xi_{GD} = \xi_{min}(1 + \eta\mathrm{Tr}(R)).$$

For the EGU± algorithm the Jacobian can be shown to have elements such that,

$$[J(\psi_{EGU\pm}(\mathbf{z}^*))]_{i,j} = \begin{cases} \frac{u_i^2 + u_i\sqrt{u_i^2+4}+4}{u_i+\sqrt{u_i^2+4}}, & i = j \\ 0, & i \neq j. \end{cases}$$

Note that $J(\psi_{EGU\pm}(\mathbf{z}^*))$ is diagonal, as desired. The Jacobian, $J(\psi_{EG\pm}(\mathbf{z}^*))$ is not, however, diagonal. For the case that U is of a similar magnitude to $\|\mathbf{u}\|_1$ the Jacobian is such that,

$$[J(\psi_{EG\pm}(\mathbf{z}^*))]_{i,j} = \begin{cases} |u_i| - \frac{1}{U}u_i^2, & i = j \\ -\frac{1}{U}u_i u_j, & i \neq j. \end{cases}$$

Recall that $U \geq \|\mathbf{u}\|_1$. Even with the constraint that U is of a similar magnitude to $\|\mathbf{u}\|_1$ then typically it would be expected that $U \gg |u_i|$. In this case the Jacobian is diagonally dominant with the $|u_i|$ components of the diagonal terms being the largest in the entire expression. In order to apply theorem 1 to the EG± algorithm $J(\psi_{EG\pm}(\mathbf{z}^*))$ is approximated by $J_d(\psi_{EG\pm}(\mathbf{z}^*))$ which is a diagonal matrix with the diagonal components of $J(\psi_{EG\pm}(\mathbf{z}^*))$.

It would be strange for an algorithm such as the GD, EGU± or EG± not to be diagonally dominant. This can be seen by considering the form of the update to $z_{t,i}$,

$$z_{t+1,i} = z_{t,i} - \eta\frac{\partial L(y_t, \hat{y}_t)}{\partial w_{t,i}}.$$

This update is based on the variation of $L(y_t, \hat{y}_t)$ with $w_{t,i}$. It is logical to expect that, subsequently, $w_{t+1,i}$ will be strongly dependent on $z_{t+1,i}$. If this is not the case then the updates of $w_{t,i}$ are dependent on aspects of $L(y_t, \hat{y}_t)$ with which it has no connection, an approach seemingly lacking in logic.

IV. Experimental Results

Experiments have been conducted to investigate the comparative predictive ability of theorem 1 across the three algorithms. Two cases have been investigated, that of a sparse target vector (a sparse vector contains mostly zero elements) and that of a non-sparse target. Each element of $\mathbf{x}_t$ was generated as an independent Gaussian random variable with mean $= \mu_x = 0$ and standard deviation $= \sigma_x = 3$. The noise added was another zero-mean Gaussian random variable with standard deviation $= \sigma_n = 0.5$. With such input concise MSE predictions can be made. For the GD algorithm, from equation (6)

$$\xi_{GD} = (1 + \eta k\sigma_x^2)\sigma_n^2. \tag{17}$$

For the EGU± algorithm, from equation (16) and the Jacobian,

$$\xi_{EGU\pm} = \Big(1 + \eta\sigma_x^2 \sum_{i=1}^{k} \frac{u_i^2 + u_i\sqrt{u_i^2+4}+4}{u_i+\sqrt{u_i^2+4}}\Big)\sigma_n^2. \tag{18}$$

The case of the EG± algorithm sees, with $U \approx \|\mathbf{u}\|_1$,

$$\xi_{EG\pm} = \Big(1 + \eta\big(\|\mathbf{u}\|_1 - \frac{\|\mathbf{u}\|_2^2}{\|\mathbf{u}\|_1}\big)\sigma_x^2\Big)\sigma_n^2 \tag{19}$$

from equation (16) and the Jacobian.

The trials conducted saw the sparse target being [(36 zeroes),-2,3,0,1,(50 zeroes)], the non-sparse [0,0,(34 ones),-2,3,0,-3,(50 twos)]. Each steady state MSE result obtained with a particular algorithm learning a particular target with a particular learning rate was averaged over 300,000 trials, the results can be seen in figure 1.

Generally the predictions illustrated appear of similar quality, accurate at low learning rate but less so as the learning rate increases. This is not surprising as the derivation of the relationship assumes a low learning rate. The most reassuring aspect of figure 1 is that, from a casual observation at least, the predictions for the EGU± and EG± algorithms appear of a similar quality to those made for the GD algorithm. This is comforting as the GD prediction is commonly acknowledged as suitably accurate for many applications, despite its obvious failings at high learning rates.

A quantitative comparison of predictions is given by the percentage squared error in prediction (PSEP).

$$\mathrm{PSEP} = \frac{(\text{Actual MSE - Predicted MSE})^2}{(\text{Actual MSE - Minimum MSE})^2} \times 100.$$

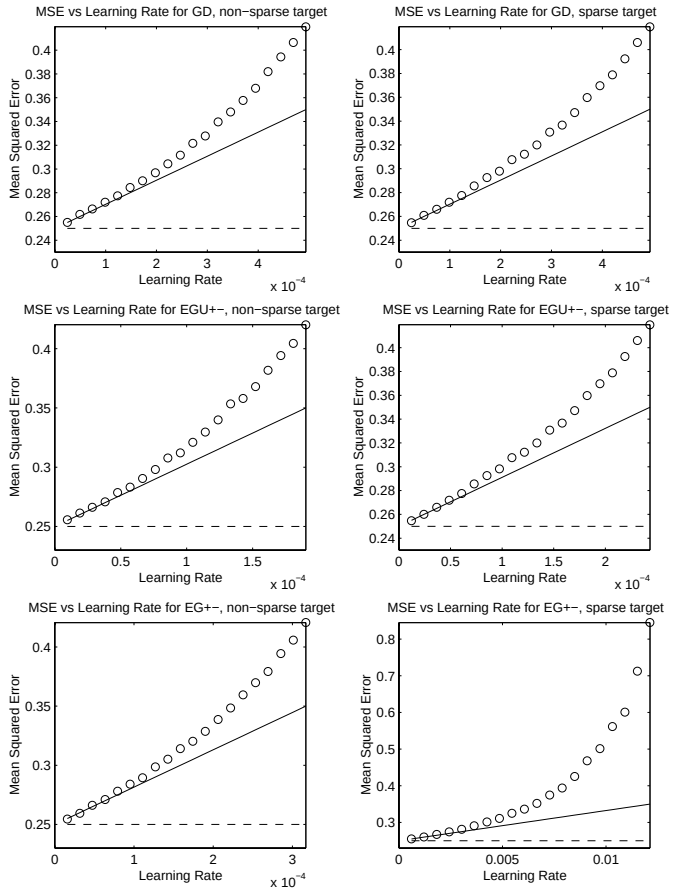


Fig. 1. Results of trials of theorem 1 *Two trials have been conducted of each algorithm, one with a sparse target and one with a non-sparse target. The experimental MSE results are given as points, the MSE prediction as a solid line and the minimum MSE as a dashed line.*

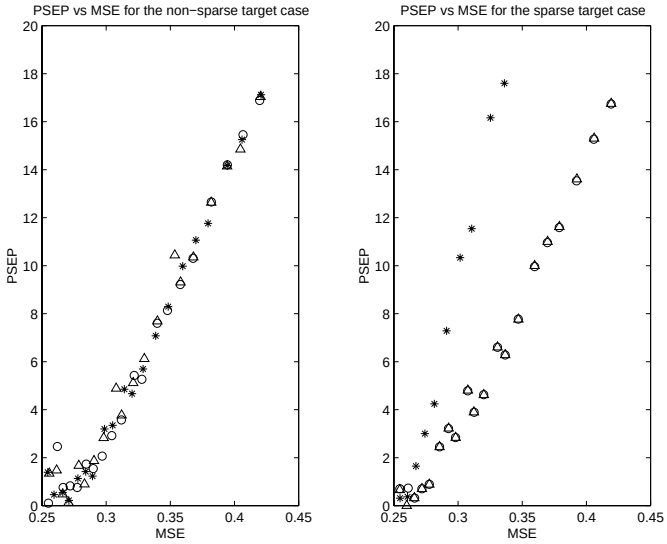


Fig. 2. Percentage squared error in MSE predictions vs MSE *The percentage squared error in an MSE prediction is the actual MSE minus the predicted MSE divided by the actual MSE minus the minimum MSE, all squared. The GD results are indicated by '$\circ$', EG$\pm$ by '*' and EGU$\pm$ by '$\triangle$'*

The PSEP corresponding to the results in figure 1 is plotted for the three algorithms in figure 2.

Figure 2 reinforces the impression obtained from figure 1, namely that at low learning rates the predictions are all of a similar quality. An indicator of approximate accuracy obtainable from figure 2 is that the prediction of MSE appears accurate to within 10% for MSE up to 30% greater than the minimum for all algorithms.

Note from figure 2 that the accuracy of the prediction of the MSE of the EG$\pm$ algorithm in learning a sparse target is clearly worse than for other cases. This is likely due, in part at least, to the decreasing accuracy of the approximation of $J(\psi_{EG\pm}(\mathbf{z}_t))$ by $J_d(\psi_{EG\pm}(\mathbf{z}_t))$.

A. Algorithm Comparison

By having an expression for steady-state MSE the algorithms can be compared through experiment. The comparison is straightforward; a desired final steady-state MSE result is chosen, from this and the MSE / learning rate relationships corresponding learning rates are derived. Experiments are then conducted and the algorithm to converge most quickly to the desired final accuracy is taken to be the better algorithm for that particular case.

According to [1], the EG$\pm$ algorithm appears likely to outperform the GD algorithm when targets are sparse and when the comparison of performance is made by that paper's, slightly different, basis of comparison. This method of comparison centers around the cumulative loss function. The result motivated the development of theory presented in this paper such that the algorithms could be compared in an acoustic echo cancellation context. Impulse responses typical of an acoustic echo, when sampled and expressed in a vector, generally create sparse vectors. Unfortunately the condition that elements of $\mathbf{x}_t$ (which is now a vector containing a history of sampled input sound) be uncorrelated does not really sit well with a problem potentially involving speech. In such a case there will be some correlation between elements and so R cannot be assumed diagonal. Regardless theorem 1 will still be used to attempt to gain some insight into the operation of the algorithms.

Example impulse responses have been generated using the image method [11]. The image method generates impulse responses of a signal traveling from a particular speaker to a particular microphone based on room dimensions and wall, floor, and ceiling reflection coefficients. The amount of energy reflected at each time depends on a material-dependent coefficient, β [12].

Impulse responses for idealized rooms, with given β everywhere, of $(10 \times 6 \times 2.5)$m with a microphone at position $(1, 1, 2)$ and speaker at $(8, 4, 1)$ are shown in figure 3. These impulse responses have been generated for signals containing frequencies between 100 and 3000Hz and using a sampling rate of 10kHz. The impulse responses vary from the reasonably sparse case to the decidedly non-sparse. In each case reflections were considered until the contribution from a single image from a further reflection dropped below 2% of the direct path strength.

Average MSE values corresponding to the learning of the impulse responses are shown in figure 4 against iteration. The input to the algorithm to produce these results saw $\mu_x = 0$, $\sigma_x = 3$,

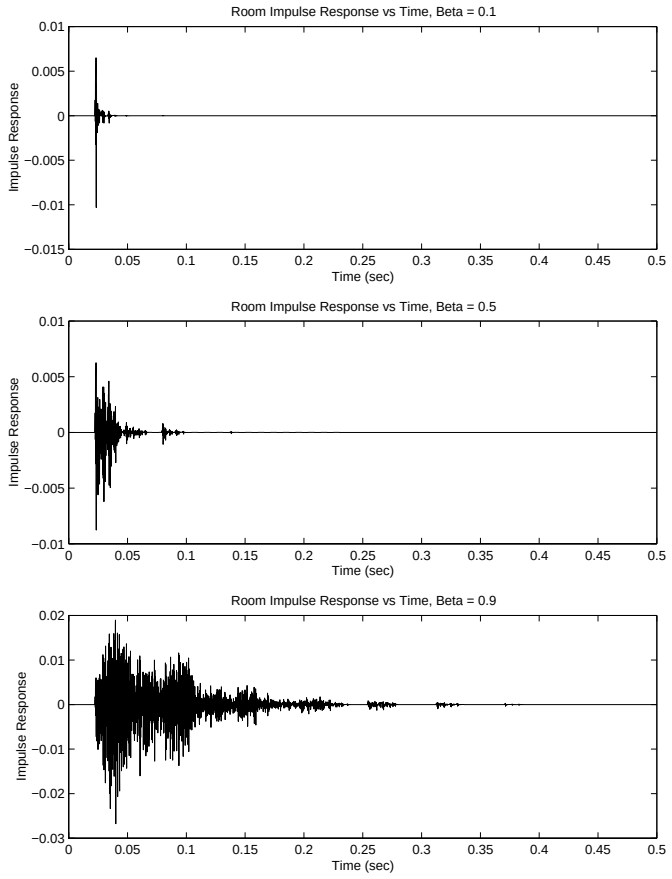


Fig. 3. Impulse responses of idealized rooms. *The rooms are (*$10 \times 6 \times 2.5$*)m with a microphone at position* $(1, 1, 2)$ *and speaker at* $(8, 4, 1)$. *The Beta values for all surfaces are as indicated.*

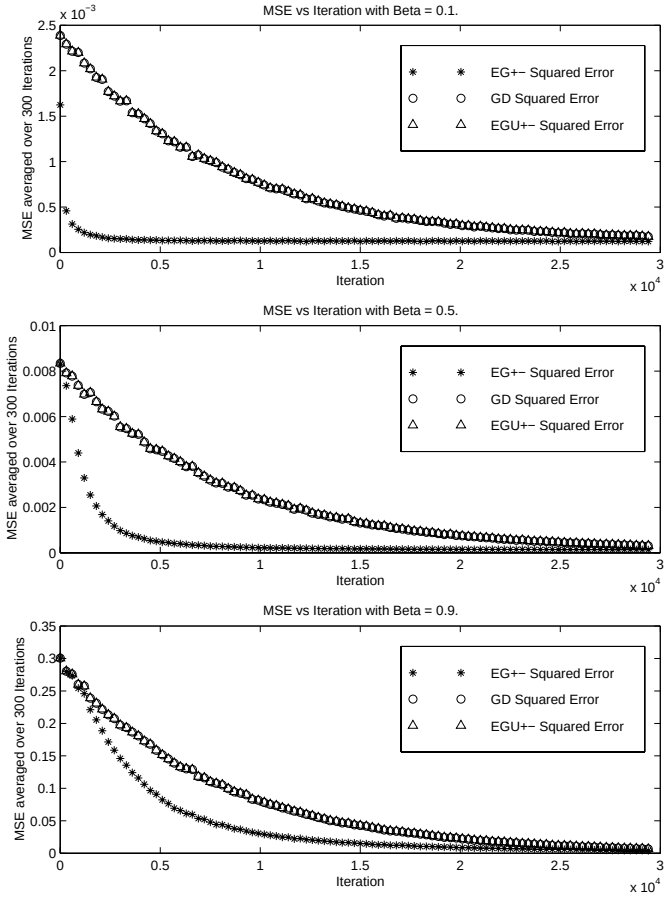


Fig. 4. A comparison of the learning of the impulse responses in figure 3. *The more sparse the target vector the faster the EG*$\pm$ *algorithm converges in comparison with the GD and EGU*$\pm$ *algorithms.*

$\mu_n = 0$, $\sigma_n = 0.01$, and the learning rate set to achieve a target MSE 20% greater than the minimum. The whole learning process was conducted 30 times to obtain an average result. Also lots of 300 iterations were averaged together. From the resulting plots it is clear that the EG$\pm$ algorithm performs far better than the other two algorithms in the sparse case, as anticipated.

An interesting outcome is that the GD and EGU$\pm$ results are virtually identical. This is likely due to the very small size of the updates and the linear nature of $\psi_{EGU\pm}(\mathbf{z}_t)$ about zero.

Simulation results of experiments conducted in attempts to generate more realistic rooms can be found in figures 5 and 6. The only difference between these experiments and the previous experiments is a more realistic choice of the absorption coefficients of the walls, floor and ceiling and the choice of room size [12]. Three room extremes have been chosen, a concrete cell with all six sides concrete, an office with carpeted floor, wood paneling and a plaster ceiling, and a recording studio with carpet on the floor and walls and a plaster ceiling. In each case the room is $(3 \times 3 \times 2.5)$m with a microphone at position $(1, 1.5, 2)$ and speaker at $(2, 2, 1)$.

It is apparent from figure 5 that the impulse responses of realistic rooms do not approach the sparseness suggested in figure 3. Subsequently the vastly superior performance of the EG$\pm$ algorithm seen in the top graph of figure 4 can be seen only as an example of this algorithm's ability with sparse targets but not as a result with much relevance to acoustic echo cancellation. Regardless it would appear that the EG$\pm$ algorithm may well be a better choice algorithm in an acoustic echo cancellation setting considering its consistently faster initial convergence as also evidenced in figure 6. This is perhaps more so given that there will be the additional (variable) delay due to the network which will cause extra sparseness; see the examples and the discussion in [5].

An experiment has also been conducted to find the impulse response of a (quite reverberant) real room, measuring approximately $(6 \times 4 \times 3)$m. This is shown in figure 7. This impulse response was also used in trials identical to those described previously except that, instead of 30 trials only 3 were used. The resulting plots of convergence serve to further illustrate that the EG$\pm$ algorithm could well be a preferred algorithm in learning such an impulse response.

V. Conclusion

The GD, EGU$\pm$ and EG$\pm$ algorithms have been seen to belong to a family of algorithms which have an additive update at their heart. This has led to the derivation of a general expression relating MSE in predictions made at steady-state by the algorithms in the family to the learning rate or step size used. This

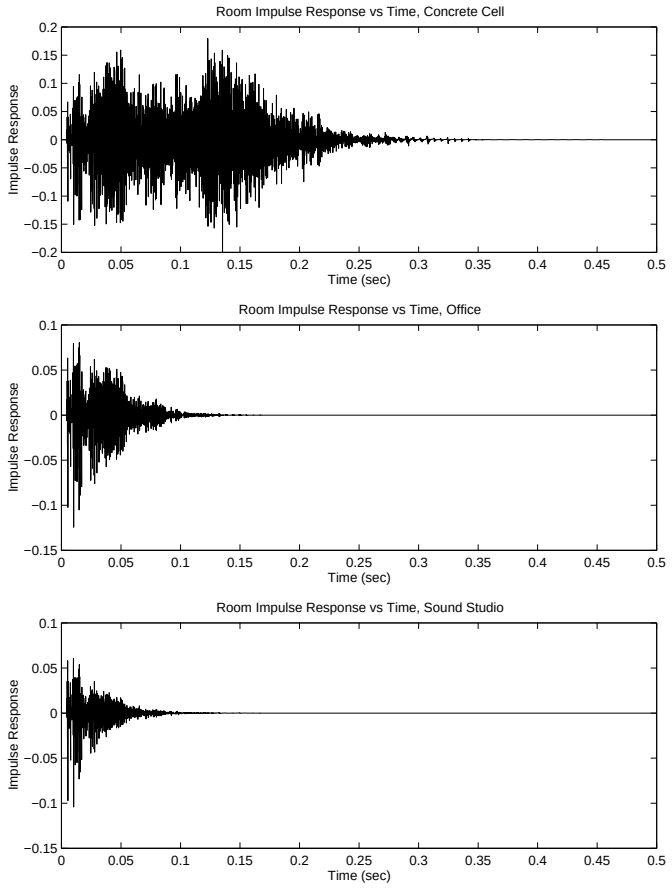


Fig. 5. Approximate impulse responses for the rooms indicated. *Based on a room size* $(3 \times 3 \times 2.5)$*m with a microphone at position* $(1, 1.5, 2)$ *and speaker at* $(2, 2, 1)$ *and with realistic absorption coefficients.*

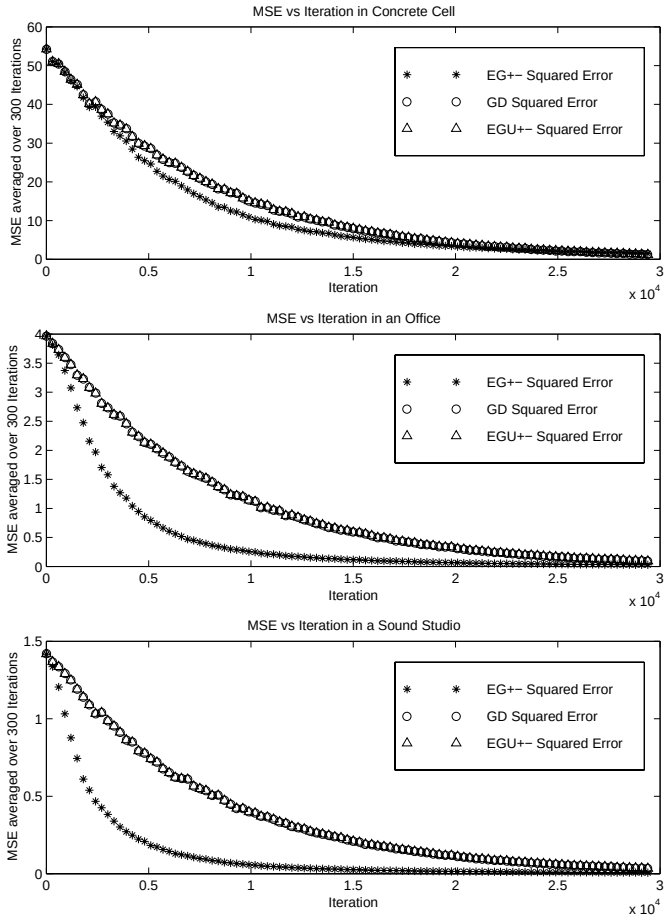


Fig. 6. A comparison of the learning of the impulse responses in figure 5. *The more sparse the target vector the better the EG± algorithm performs in comparison with the GD and EGU± algorithms.*

general expression requires that the elements of the input vector be uncorrelated and zero-mean and also that the Jacobian relating rates of change of the parameterising vector to rates of change of the weight vector be diagonal. Although the latter is not always the case it was shown that it would be strange for the Jacobian not to be at least diagonally dominant.

Trials have demonstrated the ability of the generalized expression to predict MSE across the three algorithms with comparative accuracy. Further trials have also shown that in learning sparse targets, the EG± algorithm appears able to learn to the desired accuracy more quickly than the other two algorithms which performed reasonably similarly to each other. This ability deteriorates with decreasing sparseness.

Simulations of rooms performed using the image method demonstrated that the echo characteristics are typically sufficiently sparse such that the EG± algorithm will converge to some final MSE more quickly than the other two algorithms. Even though the generalized expression calls for an input vector containing uncorrelated elements, which is not accurate in an acoustic echo scenario, it is expected that these trials provide some insight into relative performance generally..

We believe that the range of different nonlinear parametrisations of the LMS algorithm being studied in [8], [10], [9], [13] offer many opportunities in signal processing for algorithms more closely tailored to specific problems. The present paper is but one step along this route. We expect further developments and more refined analyses to provide an improved toolkit for designing online adaptive algorithms in situations where one has some prior knowledge of the target weight vector.

APPENDIX

I. THE PROOF OF THEOREM 1

The starting point is the definition of the mean squared error. Recall that the error signal is given by $e_t = y_t - \hat{y}_t$. The estimate $\hat{y}_t$ is given by $\hat{y}_t = \mathbf{w}_t \cdot \mathbf{x}_t$. An alternative expression is $\hat{y}_t = \mathbf{w}_t^T \mathbf{x}_t$. The mean squared error for a given $\mathbf{w}_t$ is given by

$$\xi \triangleq \mathrm{E}[y_t^2] - 2\mathrm{E}[y_t \mathbf{x}_t^T]\mathbf{w}_t + \mathbf{w}_t^T \mathrm{E}[\mathbf{x}_t \mathbf{x}_t^T]\mathbf{w}_t. \tag{20}$$

To simplify this expression [6] introduces $P = \mathrm{E}[y_t \mathbf{x}_t]$ and $R = \mathrm{E}[\mathbf{x}_t \mathbf{x}_t^T]$. The aim of an on-line learning algorithm is to minimize ξ. The gradient is given by:

$$\nabla_{\mathbf{w}_t} \xi = -2P + 2R\mathbf{w}_t.$$

Setting this to zero gives the target vector,

$$\mathbf{u} = R^{-1}P. \tag{21}$$

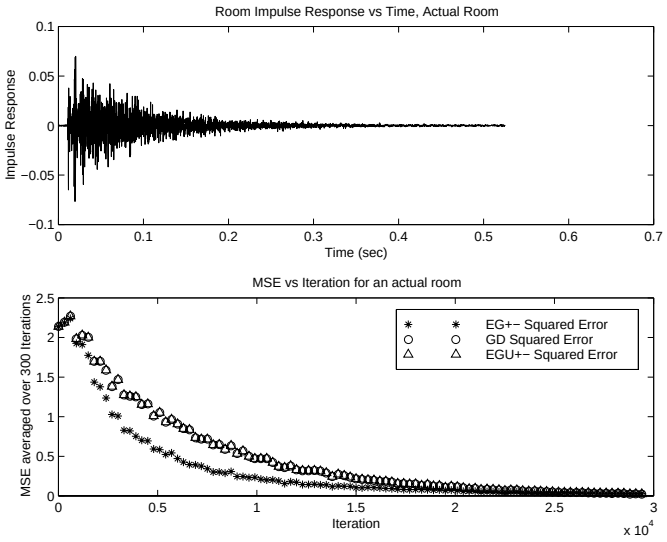


Fig. 7. The impulse response and convergence results for an actual, room. *The impulse response was experimentally determined and trials, as previously conducted to learn it.*

Substituting equation (21) into equation (20) gives, from [6],

$$\xi_{min} = \mathrm{E}[y_t^2] - P^T \mathbf{u}. \tag{22}$$

Substituting (22) into (20) gives

$$\xi = \xi_{min} + (\mathbf{w}_t - \mathbf{u})^T R(\mathbf{w}_t - \mathbf{u}).$$

Let $\mathbf{v}_t = \mathbf{w}_t - \mathbf{u}$ then

$$\xi = \xi_{min} + \mathbf{v}_t^T R \mathbf{v}_t \tag{23}$$

and

$$\nabla_{\mathbf{w}_t} \xi = 2R\mathbf{v}_t. \tag{24}$$

In all algorithms considered an estimate is made of $\nabla_{\mathbf{w}_t} \xi$,

$$\nabla_{\mathbf{w}_t} \xi \approx -\nabla_{\mathbf{w}_t} L(y_t, \hat{y}_t) = \nabla_{\mathbf{w}_t} \xi + \mathbf{n}_t \tag{25}$$

where $\mathbf{n}_t$ is gradient noise. Gradient noise is a measure of the inaccuracy in the knowledge of $\nabla_{\mathbf{w}_t} \xi$. This inaccuracy is due to the approximation of this average result by a single sample. If $\nabla_{\mathbf{w}_t} \xi = 0$ then $\nabla_{\mathbf{w}_t} L(y_t, \hat{y}_t) = \mathbf{n}_t$. Once initial convergence has been completed and in the case of a suitably small learning rate, $\mathbf{w}_t \approx \mathbf{u}$, then $\nabla_{\mathbf{w}_t} \xi \approx 0$. In this case

$$\mathbf{n}_t = 2(\hat{y}_t - y_t)\mathbf{x}_t$$

and from this the covariance matrix $\mathrm{cov}[\mathbf{n}_t] = \mathrm{E}[\mathbf{n}_t \mathbf{n}_t^T]$) can be found:

$$\mathrm{cov}[\mathbf{n}_t] = 4\xi_{min} R. \tag{26}$$

For the case that the elements of $\mathbf{x}_t$ are uncorrelated and zero mean $R = \mathrm{cov}(\mathbf{x}_t)$ is diagonal.

From equations (25) and (15) approximate additive updates can be expressed as

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \eta J(\psi(\mathbf{z}_t))(\nabla_{\mathbf{w}_t} \xi - \mathbf{n}_t).$$

For the steady state case and with sufficiently small learning rate $\mathbf{w}_t \approx \mathbf{u}$. The Jacobian is approximated by $J(\psi(\mathbf{z}_t^*))$. For convenience this is written as J in this proof.

Using equation 24 and the fact that $\mathbf{v}_t = \mathbf{w}_t - \mathbf{u}$ then;

$$\mathbf{v}_{t+1} = \mathbf{v}_t - 2\eta JR\mathbf{v}_t + J\mathbf{n}_t$$

This can be expressed,

$$\mathbf{v}_{t+1} = (I - 2\eta JR)\mathbf{v}_t + J\mathbf{n}_t$$

where I is the identity matrix. From equation (23), with a diagonal R, the excess MSE = $\sum_{j=1}^{k} \langle v_{t,j}^2 \rangle R_{jj}$ where R_{ij} = the element in R in the ith row and jth column. To find $\langle v_{t,j}^2 \rangle$ the covariance matrix of $\mathbf{v}_t$ is found,

$$\mathrm{cov}[\mathbf{v}_{t+1}] = (I - 2\eta JR)\mathrm{cov}[\mathbf{v}_t](I - 2\eta JR) + \eta^2 J \mathrm{cov}[\mathbf{n}_t] J^T. \tag{27}$$

At steady state $\mathrm{cov}[\mathbf{v}_{t+1}] = \mathrm{cov}[\mathbf{v}_t]$. Equation (27) can be rewritten,

$$\begin{aligned} \mathrm{cov}[\mathbf{v}_t] \quad = \quad & \mathrm{cov}[\mathbf{v}_t] - 2\eta JR\mathrm{cov}[\mathbf{v}_t] - \mathrm{cov}[\mathbf{v}_t]2\eta JR \\ & + 4\eta^2 JR\mathrm{cov}[\mathbf{v}_t]JR + 4\eta^2 \xi_{min} JRJ^T. \end{aligned}$$

Assuming η is small, the term $4\eta^2 JR\mathrm{cov}[\mathbf{v}_t]JR$ is neglected. As R and J are diagonal then $\mathrm{cov}[\mathbf{v}_t]$ must be diagonal and this becomes

$$4\eta JR\mathrm{cov}[\mathbf{v}_t] = 4\eta^2 \xi_{min} JRJ^T.$$

Dividing through by 4 and η and multiplying both sides on the left hand side by $R^{-1}J^{-1}$ gives

$$\mathrm{cov}[\mathbf{v}_t] = \eta \xi_{min} J^T = \eta \xi_{min} J. \tag{28}$$

From equation (23) then

$$\xi = \xi_{min} + \eta \xi_{min} \mathrm{tr}(JR).$$

As stated the condition in theorem 1 that elements of $\mathbf{x}_t$ be uncorrelated is a necessary one for a proof of this type. The proof in [6] finds a result for more general R. This is through diagonalisation of R. A similar approach cannot be taken here due to the Jacobian. The Jacobian appears in the equations in such a way that diagonalisation of R or JR or any combination of R and J does not provide the desired solution.

REFERENCES

[1] Kivinen, J. Warmuth, M.K., "Additive versus exponentiated gradient updates for linear prediction," *Information and Computation*, vol. 132, no. 1, pp. 1–64, January 1997.

[2] Johnson, C.R. Jr, *Lectures on Adaptive Parameter Estimation*, Prentice Hall, Englewood Cliffs, New Jersey, 1988.

[3] Clarkson, P.M., *Optimal and Adaptive Signal Processing*, CRC Press, Inc., Boca Raton, Florida, 1993.

[4] Homer, J. Mareels, I. Bitmead, R. Wahlberg, B. Gustafsson, F., "Least mean squares estimation via structural detection," *IEEE Transactions on Signal Processing*, vol. 46, no. 10, pp. 2651–2663, October 1998.

[5] A. Sugiyama, H. Sato, A. Hirano, and S. Ikeda, "A fast convergence algorithm for adaptive fir filters under computational constraint for adaptive tap-position control," *IEEE Transactions on Circuits and Systems II*, vol. 43, no. 9, pp. 629–636, 1996.

[6] Widrow, B. McCool, J.M. Larimore, M.G. Johnson, C.R. Jr., "Stationary and nonstationary learning characteristics of the least mean squares adaptive filter," *Proceedings of the IEEE*, vol. 64, no. 8, pp. 1151–1162, August 1976.

[7] Widrow, B. Glover, J.R. Jr. McCool, J.M. Kaunitz, J. Williams, C.S. Hearn, R.H. Zeidler, J.R. Dong, E. Jr. Goodlin, R.C., "Adaptive noise cancelling: Principles and applications," *Proceedings of the IEEE*, vol. 63, no. 12, pp. 1692–1716, December 1975.

[8] Kivinen, J. Warmuth, M.K., "Relative loss bounds for multidimensional regression problems," *Advances in Neural Information Processing Systems 10 (NIPS '97), MIT Press*, pp. 287–293, 1998.

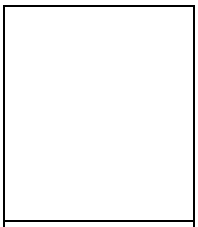
[9] Grove, A.J. Littlestone, N. Schuurmans, D., "General convergence results for linear discriminant updates," *Proceedings Tenth Annual Conference on Computational Learning Theory, ACM Press*, pp. 171–183, 1997.

[10] Warmuth, M.K. Jagota, A.K., "Continuous versus discrete-time non-linear gradient descent: Relative loss bounds and convergence," *International Symposium on Artificial Intelligence and Mathematics*, January 1997.

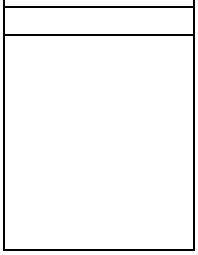
[11] Allen, J.B. Berkley, D.A., "Image method for efficiently simulating small room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, April 1979.

[12] Beranek, L.L., *Acoustics*, Acoustical Society of America, 2nd edition, 1996.

[13] Geoffrey J. Gordon, "Regret bounds for prediction problems," in *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, New York, 1999, pp. 29–40, ACM Press.



Simon Hill Obtained the BE degree from the ANU in 1999. He is now studying towards a PhD at Cambridge University. His research interests include signal processing, learning systems and the theory of mokes.



Bob Williamson received the BE degree from QIT in 1984, and the M.Eng.Sc (1986) and PhD (1990) from the University of Queensland (all in Electrical Engineering). Since 1990 he has been at the Australian National University where he is a reader in the Department of Engineering and teaches undergraduate courses in electrotechnology, signals and systems and telecommunications. His current research activities are focussed on microphone arrays and learning systems (especially support vector machines and related methods). His home page is spigot.anu.edu.au/~williams/home.shtml