AN ANALYSIS OF THE EXPONENTIATED GRADIENT DESCENT ALGORITHM

Simon I. Hill

Telecommunications Engineering, RSISE
Australian National University
Canberra, 0200, Australia
simon.hill@bigfoot.com

Robert C. Williamson

Department of Engineering
Australian National University
Canberra, 0200, Australia
bob.williamson@anu.edu.au

ABSTRACT

This paper analyses three algorithms recently studied in the Computational Learning Theory community: the Gradient Descent (GD) Algorithm, the Exponentiated Gradient Algorithm with Positive and Negative weights (EG± algorithm) and the Exponentiated Gradient Algorithm with Unnormalised Positive and Negative weights (EGU± algorithm). The analysis is of the form used in the signal processing community and is 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. Trials involving simulated acoustic echo cancellation are conducted whereby learning rates for the algorithms are selected such that they converge to the same steady state MSE. These trials demonstrate that, in the case that the target is sparse, the EG± algorithm typically converges more quickly than the GD or EGU± algorithms which perform very similarly.

1. INTRODUCTION

Gradient descent algorithms, such as LMS, are widely used in signal processing. In this paper the relationship between learning rate and steady state error is analysed for a broad class of variants of LMS which includes the exponentiated gradient algorithm of [6]. Using this relationship we then compare the exponentiated gradient algorithm to traditional LMS in an application, Acoustic Echo Cancellation (AEC), where often the target weight vector is sparse.

The input to the on-line learning algorithm at the beginning of the tth trial is $\mathbf{x}_t$, a length k vector. The algorithm maintains another length k vector of weights, $\mathbf{w}_t$ (ith entry written $w_{t,i}$). Actual system outputs are denoted y_t. Predictions of this output are $\hat{y}_t = \mathbf{w}_t \cdot \mathbf{x}_t$. Updates to the weight vector are based on prediction accuracy, given by the loss function, $L(y_t, \hat{y}_t) = (y_t - \hat{y}_t)^2$. It is assumed that there is some predictor, $\mathbf{u}$ 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 Unnormalised Positive and Negative Weights (EGU± Algorithm) and the Exponentiated Gradient Algorithm with Positive and Negative Weights (EG± Algorithm). As described in [6], the GD algorithm takes the form $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla_{\mathbf{w}_t} L(y_t, \hat{y}_t)$ where $\nabla_{\mathbf{w}_t}$ denotes the gradient with respect to $\mathbf{w}_t$. In this expression 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 [6] introduced and described the EG± algorithm which requires the maintenance of two vectors, $\mathbf{w}_t^+$ and $\mathbf{w}_t^-$ $(\mathbf{w}_t = \mathbf{w}_t^+ - \mathbf{w}_t^-)$. Updates of these vectors are made by, $w_{t+1,i}^+ = U \frac{w_{t,i}^+ r_{t,i}^+}{\sum_{j=1}^k (w_{t,j}^+ r_{t,j}^+ + w_{t,j}^- r_{t,j}^-)}$ and $w_{t+1,i}^- = U \frac{w_{t,i}^- r_{t,i}^-}{\sum_{j=1}^k (w_{t,j}^+ r_{t,j}^+ + w_{t,j}^- r_{t,j}^-)}$ where, $U \geq \|\mathbf{u}\|_1$ $\left(\|\mathbf{u}\|_p = \sqrt[p]{\sum_{i=1}^k |u_i|^p}\right)$, $r_{t,i}^+ = e^{-\eta \nabla_{\mathbf{w}_t} L(y_t, \hat{y}_t)}$ and $r_{t,i}^- = e^{\eta \nabla_{\mathbf{w}_t} L(y_t, \hat{y}_t)}$. Updates of the EGU± algorithm are similarly made through $w_{t+1,i}^+ = w_{t,i}^+ r_{t,i}^+$ and $w_{t+1,i}^- = w_{t,i}^- r_{t,i}^-$.

The measure of final accuracy of predictors used is the mean squared error (MSE)($= \langle (y_t - \hat{y}_t)^2 \rangle$ where $\langle \cdot \rangle$ signifies time average) of predictions. By definition, the MSE obtained from predictions made by $\mathbf{u}$ is a minimum, denoted ξ_{min}. Note that the greater η the more susceptible the algorithm is to noise and so the greater the movement of the elements of $\mathbf{w}_t$ about those of $\mathbf{u}$, leading to a greater steady state MSE.

2. ALGORITHM REPARAMETERISATION

In this section it will be shown how, by a suitable reparameterisation, the GD, EGU± and EG± algorithms can

This work was supported by the Australian Research Council

be viewed and analysed in a common framework. Suppose $\mathbf{z}_t \in \mathbb{R}^k$. A function $\psi(\cdot)$ is introduced such that $\mathbf{w}_t = \psi(\mathbf{z}_t)$. This function can be chosen so $\mathbf{z}_t$ is updated according to

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

The case of the GD algorithm is a trivial one, $\psi_{GD}(\mathbf{z}_t) = \mathbf{z}_t$. Reparameterisation, as described, of the EGU± algorithm cannot immediately be done due to the complete independence of starting conditions on the vectors $\mathbf{w}_t^+$ and $\mathbf{w}_t^-$. Through induction it can be shown that the imposition of the initial constraint $w_{t,i}^+ = \frac{1}{w_{t,i}^-}$ leads to the parameterisation function $\psi_{EGU\pm}$,

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

The EG± algorithm can be reparameterised also. The comments made regarding the independence of $\mathbf{w}_t^+$ and $\mathbf{w}_t^-$ also apply when considering the EG± case. For this algorithm, consider some vector $\mathbf{z}_t$ such that $w_{t,i}^+ = U\frac{e^{z_{t,i}}}{2\sum_{j=1}^k \cosh(z_{t,j})}$ and $w_{t,i}^- = U\frac{e^{-z_{t,i}}}{2\sum_{j=1}^k \cosh(z_{t,j})}$. From this, and the EG± update rule, one can show inductively that the parameterisation function is given by

$$w_{t,i} = [\psi_{EG\pm}(\mathbf{z}_t)]_i = U\frac{\sinh(z_{t,i})}{\sum_{j=1}^k \cosh(z_{t,i})}. \quad (3)$$

A suitable starting point for both algorithms which makes the parameterisation possible is $\mathbf{z}_0 = \mathbf{0}$. For the EGU± algorithm then $w_{0,i}^+ = w_{0,i}^- = 1$, for the EG± algorithm $w_{0,i}^+ = w_{0,i}^- = \frac{U}{2k}$, also the starting point in [6]. From this perspective the three algorithms belong to a greater family, all of which have an additive update at their heart.

The reparametrisation used is similar in spirit, but not in detail, to the general link functions in [7] (cf. [5]).

3. MSE / LEARNING RATE RELATIONSHIPS

Discrete time on-line learning algorithms can be considered to be derived from continuous time ones [3]. Consider the continuous time update $\dot{\mathbf{z}}_t = -\eta\nabla_{\mathbf{w}_t} L(y_t, \hat{y}_t)$. Replacing $\dot{\mathbf{z}}_t$ by its Euler discretisation, $\frac{\mathbf{z}_{t+h} - \mathbf{z}_t}{h}$ will realise the update in (1). Similarly, an additive approximation to an update of a general algorithm can be derived from $\dot{\mathbf{w}}_t = -\eta J(\psi(\mathbf{z}_t))\nabla_{\mathbf{w}_t} L(y_t, \hat{y}_t)$. An Euler discretisation gives the approximation,

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta J(\psi(\mathbf{z}_t))\nabla_{\mathbf{w}_t} L(y_t, \hat{y}_t). \quad (4)$$

This is used in the proof of theorem 3.1 (see appendix A).

Theorem 3.1 *Consider an on-line learning algorithm which can be reparameterised through* $\mathbf{w}_t = \psi(\mathbf{z}_t)$ *and* $\mathbf{z}_{t+1} = \mathbf{z}_t - \eta\nabla_{\mathbf{w}_t} L(y_t, \hat{y}_t)$. *Suppose the input to the algorithm* $\mathbf{x}_t$ *comprises of uncorrelated zero mean elements* $x_{t,i}$. *Let*

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

where $\mathbf{w} \to \mathbf{u}$ *means* $\|\mathbf{w} - \mathbf{u}\|_p \to 0$ *for some* $p < \infty$. *Suppose that 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*

$$\lim_{\eta\to 0}(\xi) = \xi_{min} + \eta\xi_{min}\mathrm{Tr}\big(J(\psi(\mathbf{z}^*))R\big). \quad (5)$$

For the GD algorithm the Jacobian is the identity matrix and equation (5) reduces to $\text{MSE}_{GD} = \xi_{GD} = \xi_{min}(1 + \eta\mathrm{Tr}(R))$. This is the oft cited relationship, regularly found in textbooks.

For the EGU± algorithm the Jacobian $J(\psi_{EGU\pm}(\mathbf{z}^*))$ is diagonal with (i,i)th entry $= \frac{u_i^2 + u_i\sqrt{u_i^2+4}+4}{u_i + \sqrt{u_i^2+4}}$. However $J(\psi_{EG\pm}(\mathbf{z}^*))$ is not diagonal. If $U \approx \|\mathbf{u}\|_1$, the Jacobian has (i,i)th entry $= |u_i| - \frac{1}{U}u_i^2$ and (i,j)th entry $(i \neq j)$ $= -\frac{1}{U}u_iu_j$. 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 3.1 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 $z_{t+1,i}$.

4. EXPERIMENTAL RESULTS

The predictive ability of theorem 3.1 across the three algorithms has been experimentally investigated. Cases of a sparse (mostly zero elements) (=[(36 zeroes),-2,3,0,1,(50 zeroes)]) and of a non-sparse (=[0,0,(34 ones),-2,3,0,-3,(50 twos)]) target vector were examined. Each element of $\mathbf{x}_t$ was generated as an independent Gaussian random variable, as was the noise, with means $\mu_x = \mu_n = 0$ and standard deviations $\sigma_x = 3$ and $\sigma_n = 0.5$. With such input concise MSE predictions can be made. For the GD algorithm, $\xi_{GD} = (1 + \eta k\sigma_x^2)\sigma_n^2$, for the EGU± algorithm, $\xi_{EGU\pm} = \left(1 + \eta\sigma_x^2\sum_{i=1}^k \frac{u_i^2 + u_i\sqrt{u_i^2+4}+4}{u_i + \sqrt{u_i^2+4}}\right)\sigma_n^2$, and for the EG± algorithm (with $U \approx \|\mathbf{u}\|_1$), one can show that

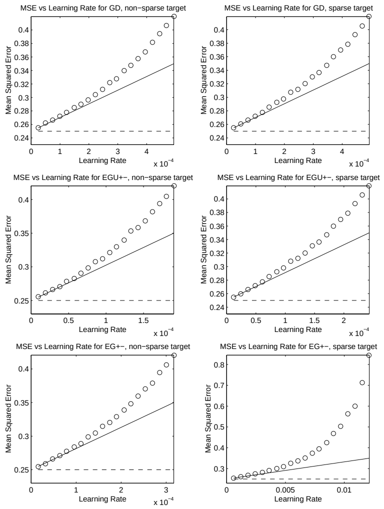

Figure 1: Experimental test of 5 *The experimental MSE results are given as points, the MSE prediction as a solid line and ξ_{min} as a dashed line.*

$\xi_{EG\pm} = \Big(1 + \eta\big(\|\mathbf{u}\|_1 - \frac{\|\mathbf{u}\|_2^2}{\|\mathbf{u}\|_1}\big)\sigma_x^2\Big)\sigma_n^2$. Each MSE result (shown with corresponding prediction in figure 1) is the result of averaging over 300,000 trials.

Note that the increasing accuracy of prediction with decreasing η is not surprising, given (5) holds only in the limit $\eta \to 0$. Reassuringly the predictions for the EGU± and EG± algorithms appear of similar quality to those commonly used for the GD algorithm.

4.1. Algorithm Comparison

By having an expression for steady-state MSE the algorithms can be fairly 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 chosen. 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.

Consider the problem of a speaker phone, the loudspeaker broadcasts into the room, the sound echoes around and acts as noise on the signal heard by the microphone. Impulse responses describing the echoing in such a situation typically form sparse vectors. By learning the impulse response in this and other similar echo problems the echoing can be cancelled. According to [6], the EG± algorithm appears likely to outperform the GD algorithm when targets are sparse. Unfortunately the condition imposed by theorem 3.1 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. Regardless the theorem can still be used to attempt to gain some insight into the operation of the algorithms.

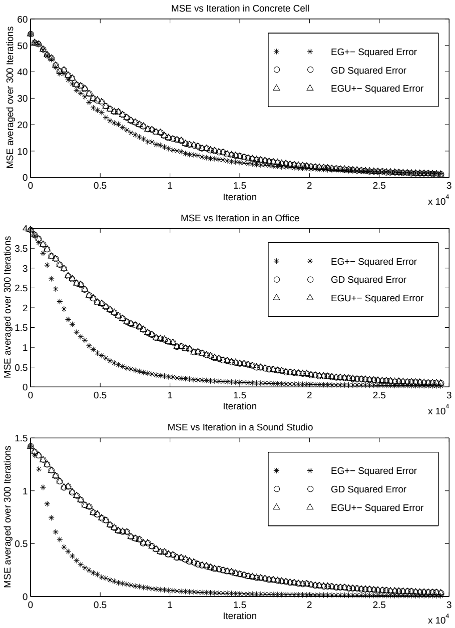

Figure 2: A comparison of the learning of the room impulse responses. *The more sparse the target vector the better the EG± algorithm performs in comparison with the GD and EGU± algorithms.*

Example impulse responses of simulated rooms were generated using the image method [2]. Three room extremes were modelled, in order of increasing sparseness of their impulse responses; a concrete cell with all six sides concrete, an office with carpeted floor, wood panelling 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 $(2, 2, 1)$. Approximate absorption coefficients were found with reference to [1].

Trials involving the learning of these impulse responses were conducted to compare the algorithms. The inputs used were similar to those described above, the only change being $\sigma_n = 0.01$. Theorem 3.1 was used to set final MSE of all algorithms at 20% greater than the minimum. It would appear from the results in figure 2 that the EG± algorithm may well be a better choice algorithm in AEC considering its consistently faster initial convergence.

5. CONCLUSION

The GD, EGU± and EG± 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 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 generalised expression to predict MSE across the three algorithms with comparative accuracy. Simulations of rooms performed using the image method demonstrated that the echo characteristics are typically suitably sparse such that the EG± algorithm will converge to some final MSE more quickly than the other two algorithms. This ability deteriorates with decreasing sparseness. Unfortunately the generalised expression calls for an input vector containing uncorrelated elements which is not accurate in an acoustic echo scenario. However it is expected that these trials provide some insight into relative performance generally.

6. REFERENCES

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

[2] Allen, J.B. Berkley, D.A. Image method for efficiently simulating small room acoustics. *Journal of the Acoustical Society of America*, 65(4):943–950, April 1979.

[3] 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.

[4] 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*, 64(8):1151–1162, August 1976.

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

[6] Kivinen, J. Warmuth, M.K. Additive versus exponentiated gradient updates for linear prediction. *Information and Computation*, 132(1):1–64, January 1997.

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

A. THE PROOF OF THEOREM 3.1

The mean squared error for a given $\mathbf{w}_t$ is given by

$$\xi \triangleq \mathrm{E}[y_t^2] - 2\mathrm{E}[y_t\mathbf{x}_t^T]\mathbf{w}_t + \mathbf{w}_t^T\mathrm{E}[\mathbf{x}_t\mathbf{x}_t^T]\mathbf{w}_t. \qquad (6)$$

Following [4], let $P = \mathrm{E}[y_t\mathbf{x}_t]$ and $R = \mathrm{E}[\mathbf{x}_t\mathbf{x}_t^T]$. The gradient is given by: $\nabla_{\mathbf{w}_t}\xi = -2P + 2R\mathbf{w}_t$. Setting this to zero gives the target vector, $\mathbf{u} = R^{-1}P$. Substituting this into (6) gives, (see [4]),

$$\xi_{min} = \mathrm{E}[y_t^2] - P^T\mathbf{u}. \qquad (7)$$

Substituting (7) into (6) and with $\mathbf{v}_t = \mathbf{w}_t - \mathbf{u}$ then $\xi = \xi_{min} + \mathbf{v}_t^T R\mathbf{v}_t$ and $\nabla_{\mathbf{w}_t}\xi = 2R\mathbf{v}_t$. In all algorithms considered an estimate is made of $\nabla_{\mathbf{w}_t}\xi$,

$$\nabla_{\mathbf{w}_t}\xi \approx -\nabla_{\mathbf{w}_t}L(y_t,\hat{y}_t) = \nabla_{\mathbf{w}_t}\xi + \mathbf{n}_t \qquad (8)$$

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. In steady state and with a small η, $\mathbf{w}_t \approx \mathbf{u}$, so $\nabla_{\mathbf{w}_t}\xi \approx 0$, furthermore the Jacobian is approximated by $J(\psi(\mathbf{z}_t^*))$. In this case $\mathbf{n}_t = 2(\hat{y}_t - y_t)\mathbf{x}_t$ and from this the covariance matrix cov$[\mathbf{n}_t] = \mathrm{E}[\mathbf{n}_t\mathbf{n}_t^T]$) can be found, cov$[\mathbf{n}_t] = 4\xi_{min}R$. When the elements of $\mathbf{x}_t$ are uncorrelated and have zero mean, $R =$ cov$(\mathbf{x}_t)$ is diagonal. From (8) and (4),

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

Using $\nabla_{\mathbf{w}_t}\xi = 2R\mathbf{v}_t$ 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 written as $\mathbf{v}_{t+1} = (I - 2\eta JR)\mathbf{v}_t + J\mathbf{n}_t$ where I is the identity matrix. From $\xi = \xi_{min} + \mathbf{v}_t^T R\mathbf{v}_t$, with a diagonal R, the excess MSE = $\sum_{j=1}^{k}\langle v_{t,j}^2\rangle R_{jj}$ where R_{ij} = the (i,j)th element in R. To find $\langle v_{t,j}^2\rangle$, the covariance matrix of $\mathbf{v}_t$ is found, taking into account that at steady state cov$[\mathbf{v}_{t+1}] =$ cov$[\mathbf{v}_t]$. Then,

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

Assuming η is small, the term $4\eta^2 JR$cov$[\mathbf{v}_t]JR$ is neglected. Since R and J are diagonal, cov$[\mathbf{v}_t]$ is too, giving $4\eta JR$cov$[\mathbf{v}_t] = 4\eta^2\xi_{min}JRJ^T$, which simplifies to

$$\mathrm{cov}[\mathbf{v}_t] = \eta\xi_{min}J^T = \eta\xi_{min}J. \qquad (9)$$

Substituting (9) into $\xi = \xi_{min} + \mathbf{v}_t^T R\mathbf{v}_t$ yields

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

As stated the condition in theorem 3.1 that elements of $\mathbf{x}_t$ be uncorrelated is a necessary one for a proof of this type. The proof in [4] finds a result for more general R. This is through diagonalisation of R. A similar approach cannot be taken here due to the Jacobian.