

# NORMALIZED NATURAL GRADIENT ADAPTIVE FILTERING FOR SPARSE AND NON-SPARSE SYSTEMS

Steven L. Gay<sup>1</sup> and Scott C. Douglas<sup>2</sup>

<sup>1</sup>Acoustics and Speech Research  
Bell Labs, Lucent Technologies  
Murray Hill, NJ, USA 07974  
email: [slg@research.bell-labs.com](mailto:slg@research.bell-labs.com)

<sup>2</sup>Department of Electrical Engineering  
Southern Methodist University  
Dallas, TX, USA 75275  
email: [douglas@enr.smu.edu](mailto:douglas@enr.smu.edu)

## ABSTRACT

This paper introduces a class of normalized natural gradient algorithms (NNGs) for adaptive filtering tasks. Natural gradient techniques are useful for generating relatively simple adaptive filtering algorithms where the space of the adaptive coefficients is curved or warped with respect to Euclidean space. The advantage of normalizing gradient adaptive filters is that constant rates of convergence for signals with wide dynamic ranges may be achieved. We show that the so-called proportionate normalized least mean squares (PNLMS) algorithm, an adaptive filter that converges quickly for sparse solutions, is in fact an NNG on a certain parameter space warping. We also show that by choosing a warping that favors diverse or dense impulse responses, we may obtain a new adaptive algorithm, the inverse proportionate NLMS (INLMS) algorithm. This procedure converges quickly to and accurately tracks non-sparse impulse responses.

## 1. INTRODUCTION

Classic adaptive filtering algorithms such as least-mean-square (LMS) and normalized LMS (NLMS) are essential for many signal processing tasks. Despite their broad usefulness, these methods can converge slowly in certain situations. One example is in network echo cancellation, where the unknown echo path has a large time extent. It is not unusual to employ FIR adaptive filters with over 500 filter coefficients in these cases.

To improve performance, researchers have developed various algorithmic modifications to gradient adaptive filters. One intriguing modification is described in [1,2,3]. Called the proportionate NLMS (PNLMS) algorithm, this approach alters the NLMS update by individually-scaling each coefficient update according to an affine transformation of the absolute coefficient value. This algorithm has been shown to converge quickly and track unknown impulse responses that are sparse, i.e. most of the converged coefficients are nearly zero. Although some theoretical justifications for this algorithm has been given, a formal derivation of this method has not appeared in the literature. Such a derivation would likely enable the design of

algorithms for other scenarios, e.g. the unknown coefficients are non-sparse or diverse.

The purpose of this paper is three-fold. Firstly, we introduce a class of normalized natural gradient algorithms for adaptive filtering tasks. The natural gradient technique introduced by Amari [4] is useful for generating relatively simple adaptive filtering algorithms where the space of the adaptive coefficients is curved or warped with respect to Euclidean space. Normalized gradient adaptive filters provide constant rates of convergence for signals with wide dynamic ranges. By a novel extension of natural gradient adaptation to a posteriori error minimization in a system identification task, we obtain an algorithm family that can be viewed as the normalized step size version of a generic natural gradient procedure. Secondly, we show that the PNLMS algorithm can be derived as a normalized natural gradient procedure in a warped coefficient space that is defined by a particular sparse coefficient metric. Hence, PNLMS is a true gradient procedure, with corresponding well-behaved convergence properties, that is ideally suited to sparse system identification. Thirdly, by choosing a different warping of coefficient space that favors diverse or dense impulse responses, we obtain a new adaptive algorithm, the inverse proportionate NLMS (INLMS) algorithm. This procedure converges quickly to and accurately tracks non-sparse impulse responses. Simulations verify the capabilities of the proposed methods in their respective system identification tasks.

In this paper we will use the following definitions:

- $x_n$  is the system excitation,
- $\mathbf{x}_n = [x_n, \dots, x_{n-L+1}]^T$  is the excitation vector,
- $\mathbf{h}_S = [h_{0,n}, \dots, h_{L-1,n}]^T$  is the true system impulse response vector,
- $v_n$  is noise,
- $y_n = \mathbf{x}_n^T \mathbf{h}_S + v_n$  is the combination of the system response and the noise signals,
- $\mathbf{h}_n = [h_{0,n}, \dots, h_{L-1,n}]^T$  is the adaptive filter coefficient vector,
- $e_n = y_n - \mathbf{x}_n^T \mathbf{h}_{n-1}$  is the error signal,

- $\mu$  is the "step-size" parameter and is usually chosen in the range  $0 < \mu < 1$ , and
- $\delta$  is the regularization parameter.

## 2. NORMALIZED NATURAL GRADIENT ADAPTATION

Gradient descent is a well-known and popular procedure for system identification tasks. This procedure uses local knowledge about a chosen cost function,  $J(\mathbf{h}_n)$ , to compute updates to the coefficients. Over time, the parameters converge to a local minimum of the cost function. In most implementations, no knowledge about the unknown parameters is used within the gradient descent procedure, other than a chosen initial estimate.

Consider now the case where the unknown parameters are sparse or nearly so; *i.e.* most of the parameter values to be estimated are nearly zero. Then, the space of parameter solutions is restricted to points in  $L$ -dimensional coefficient space that lie near one or more coefficient axes. This knowledge does not lend itself to a linearly constrained adaptation procedure, because the constraints cannot be specified easily. What is needed is an adaptation method that takes advantage of the near-sparseness of the parameter solution space while still allowing for movements within general  $L$ -dimensional parameter space.

To construct the coefficient updates, a procedure that takes into account the non-isotropic nature of the parameter space is needed. Natural gradient adaptation offers the potential for the design of these updates [5]. Natural gradient adaptation is a modified gradient search that changes the standard gradient update procedure according to the non-Euclidean nature of the parameter space. The resulting updates are based on a "non-straight-line" distance metric that is defined by the Riemannian geometry of the parameter space, when such geometry exists. Natural gradient adaptation produces no spurious stationary points and is often simple to implement. For a parameter vector  $\mathbf{h}_n$ , the natural gradient update procedure is defined by

$$\mathbf{h}_n = \mathbf{h}_{n-1} + \mu \mathbf{G}^{-1}(\mathbf{h}_{n-1}) \frac{\partial J(\mathbf{h}_{n-1})}{\partial \mathbf{h}} \quad (1)$$

where  $\mathbf{G}(\mathbf{h}_{n-1})$  is the Riemannian metric tensor that describes the local curvature of the parameter space at  $\mathbf{h}_{n-1}$ ,  $\partial J(\mathbf{h}_{n-1})/\partial \mathbf{h}$  is the gradient of the cost function,  $J(\mathbf{h}_{n-1})$ , evaluated at  $\mathbf{h}_{n-1}$ , and  $\delta$  is a small constant. We specify the natural gradient adaptation procedure by specifying  $\mathbf{G}(\mathbf{h}_{n-1})$ . This matrix is determined by the distance metric that defines gradient changes within parameter space.

We now develop a normalized version of the natural gradient procedure that yields fast and stable adaptive behavior for a fixed range of step size values. To do so, consider the general form of an adaptive filter's error calculation and coefficient update,

$$e_n = y_n - \mathbf{x}_n^T \mathbf{h}_{n-1} \quad (2)$$

$$\mathbf{h}_n = \mathbf{h}_{n-1} + \mu \mathbf{r}_n \quad (3)$$

The form of the update vector,  $\mathbf{r}_n$ , determines the algorithm that we use to adjust the coefficients. To derive NLMS, we pose the following problem: Minimize the quantity

$$M(\mathbf{h}_n, \mathbf{h}_{n-1}) = \mu^{-2} \|\mathbf{h}_n - \mathbf{h}_{n-1}\|_2^2 = \mathbf{r}_n^T \mathbf{r}_n \quad (4)$$

subject to the constraint that the a posteriori error,

$$\dot{e}_n = y_n - \mathbf{x}_n^T \mathbf{h}_n = e_n - \mu \mathbf{x}_n^T \mathbf{r}_n, \quad (5)$$

is zero when  $\mu$  is one. Notice that a Euclidean or "straight-line" distance metric is minimized within this derivation, and it yields a modified LMS procedure.

To derive a normalized natural gradient procedure, we choose a distance metric that is not Euclidean, but Riemannian. Many such metrics could be chosen, but for purposes of this paper, we choose the metrics of the form,

$$\sum_{i=0}^{L-1} |f(h_{i,n-1} + r_{i,n}) - f(h_{i,n-1})|^2 \quad (6)$$

where  $f(h_{i,n})$  is a sign-preserving nonlinear scalar transformation of the coefficient value  $h_{i,n}$ . The  $\mathbf{r}_n$  that minimizes Eq. (6) depends on the *current* coefficient estimate  $\mathbf{h}_{n-1}$ . By using Taylor series approximations in Eq. (6) we find a local approximation of the metric and arrive at a metric of the form,

$$M(\mathbf{h}_n, \mathbf{h}_{n-1}) = \mathbf{r}_n^T \mathbf{G}(\mathbf{h}_{n-1}) \mathbf{r}_n \quad (7)$$

By choosing different nonlinear transformations,  $f(h_{i,n})$ , we may generate different forms of  $\mathbf{G}(\mathbf{h}_{n-1})$ , resulting in algorithms with different adaptation properties.

Minimizing Eq. (7) under the constraint of Eq. (5) we arrive at the overall cost function,

$$C_n = \delta \mathbf{x}_n^T \mathbf{G}_n \mathbf{r}_n + (e_n - \mathbf{x}_n^T \mathbf{r}_n)^2 \quad (8)$$

Setting the derivative of this to zero and then using the matrix inversion lemma, it can be seen that (8) is minimized by,

$$\mathbf{r}_n = \mu \mathbf{G}_n^{-1} \mathbf{x}_n (\mathbf{x}_n^T \mathbf{G}_n^{-1} \mathbf{x}_n + \delta)^{-1} e_n \quad (9)$$

So, (3) becomes,

$$\mathbf{h}_n = \mathbf{h}_{n-1} + \mu \mathbf{G}_n^{-1} \mathbf{x}_n (\mathbf{x}_n^T \mathbf{G}_n^{-1} \mathbf{x}_n + \delta)^{-1} e_n \quad (10)$$

which is a normalized version of Eq. (1). Note that when  $\mathbf{G}_n = \mathbf{I}$ , Eq. (10) reduces to the standard regularized NLMS coefficient update. Moreover, when the regularization parameter,  $\delta$ , is large, then Eq. (10) behaves as the standard natural gradient update.

## 3. DERIVATION OF PNLMs AND INLMs

We now show that the normalized natural gradient algorithm derived in the last section is identical to the PNLMs update when a particular metric  $M(\mathbf{h}_n, \mathbf{h}_{n-1})$  is chosen. This metric is determined by a warping transformation on the coefficient space given by

$$f(h_{i,n}) = \sqrt{\bar{h}_n (|h_{i,n}| + \beta_n)} \operatorname{sgn}(h_{i,n}) \quad (11)$$

where,

$$\beta_n = \rho (\|\mathbf{h}_n\|_2 + \delta_p) \quad (12)$$

and

$$\bar{h}_n = \frac{1}{L} \sum_{i=0}^{L-1} |h_{i,n}| + \beta_n. \quad (13)$$

Typically,  $\rho \approx .01$ , and  $\delta_p \approx 5/L$ .

The warping of the parameter space defined by Eq. (11) through (13) is somewhat complicated by some necessary regularization ( $\beta_n$ ) and normalization ( $\bar{h}_n$ ) terms, but the essential idea is that the square root in Eq. (11) warps the parameter space in such a way as to make distances in the direction *orthogonal* to coordinate axes *near* those coordinate axes larger than Euclidean distances. Thus, once adaptive coefficient vectors come close to coordinate axes, they tend to stay close, resulting in a tendency toward sparse solutions.

The metric we will minimize under the a posteriori constraint is

$$M(\mathbf{h}_n, \mathbf{h}_{n-1}) = \|F(\mathbf{h}_n) - F(\mathbf{h}_{n-1})\|_2^2 \quad (14)$$

where we define  $F(\mathbf{h}_{n-1})$  as the vector operator,

$$F(\mathbf{h}_n) = \bar{h}_n \begin{bmatrix} \sqrt{\bar{h}_n (|h_{0,n}| + \beta_n)} \operatorname{sgn}(h_{0,n}) \\ \sqrt{\bar{h}_n (|h_{1,n}| + \beta_n)} \operatorname{sgn}(h_{1,n}) \\ \vdots \\ \sqrt{\bar{h}_n (|h_{L-1,n}| + \beta_n)} \operatorname{sgn}(h_{L-1,n}) \end{bmatrix}. \quad (15)$$

We can write Eq. (14) as,

$$M(\mathbf{h}_n, \mathbf{h}_{n-1}) = \sum_{i=0}^{L-1} \left\| \sqrt{\bar{h}_{n-1} (|h_{i,n-1}| + r_{i,n}) + \beta_{n-1}} - \left\| \sqrt{\bar{h}_{n-1} (|h_{i,n-1}| + \beta_{n-1})} \right\|_2 \right\|^2 \quad (16)$$

where we assume that  $\|\mathbf{h}_{n-1}\| \gg r_{i,n}$ . Consider the two cases,  $h_{i,n-1}r_{i,n} > 0$  and  $h_{i,n-1}r_{i,n} < 0$ , respectively. In the first case, Eq. (16) becomes

$$M(\mathbf{h}_n, \mathbf{h}_{n-1}) = \sum_{i=0}^{L-1} \left\| \sqrt{\bar{h}_{n-1} (|h_{i,n-1}| + |r_{i,n}| + \beta_{n-1})} - \left\| \sqrt{\bar{h}_{n-1} (|h_{i,n-1}| + \beta_{n-1})} \right\|_2 \right\|^2. \quad (17)$$

Expressing the square root with a Taylor series expansion and keeping only the first two terms, we can approximate Eq. (17) with

$$\sum_{i=0}^{L-1} \frac{r_{i,n}^2 \bar{h}_{n-1}}{\left( |h_{i,n-1}| + \beta_{n-1} \right)}. \quad (18)$$

In the second case, we have

$$\sum_{i=0}^{L-1} \left\| \sqrt{\bar{h}_{n-1} (|h_{i,n-1}| - |r_{i,n}| + \beta_{n-1})} - \left\| \sqrt{\bar{h}_{n-1} (|h_{i,n-1}| + \beta_{n-1})} \right\|_2 \right\|^2 \quad (19)$$

which again, using the truncated Taylor series, gives us Eq. (18). We can write Eq. (18) in vector form as,

$$\bar{h}_{n-1} \mathbf{r}_n^T (\mathbf{F}_{n-1} + \beta_{n-1} \mathbf{I})^{-1} \mathbf{r}_n \quad (20)$$

where,

$$\mathbf{F}_{n-1} = \begin{bmatrix} |h_{0,n-1}| & & & \\ & |h_{1,n-1}| & & \\ & & \ddots & \\ & & & |h_{L-1,n-1}| \end{bmatrix}. \quad (21)$$

So, we recognize for our chosen metric that the Riemannian metric tensor is simply,

$$\mathbf{G}(\mathbf{h}_{n-1}) = (\mathbf{F}_{n-1} + \beta_{n-1} \mathbf{I})^{-1}. \quad (22)$$

To obtain the normalized natural gradient for this non-Euclidean coefficient metric, we substitute the  $\mathbf{G}(\mathbf{h}_{n-1})$  expression above into Eq. (10) to get

$$\mathbf{h}_n = \mathbf{h}_{n-1} + (\mathbf{F}_{n-1} + \beta_{n-1} \mathbf{I}) \mathbf{x}_n \bullet \quad (23)$$

$$\left[ \mathbf{x}_n^T (\mathbf{F}_{n-1} + \beta_{n-1} \mathbf{I}) \mathbf{x}_n + \delta \bar{h}_{n-1} \right]^{-1} e_n$$

which is very similar to Gay's version of the PNLMS algorithm [2].

As stated previously, choosing different parameter space transformations yields different algorithms. If we replace  $f(h_{i,n})$  in Eq. (11) with another sparse favoring transform,

$$f_l(h_{i,n}) = \bar{h}_n \log(|h_{i,n}| + \beta_{n-1}) \operatorname{sgn}(h_{i,n}), \quad (24)$$

we obtain an alternative algorithm for sparse system identification. Using the NNG derivation above, we obtain an algorithm that has a similar to that in Eq. (23):

$$\mathbf{h}_n = \mathbf{h}_{n-1} + (\mathbf{F}_{n-1} + \beta_{n-1} \mathbf{I})^2 \mathbf{x}_n \bullet \quad (25)$$

$$\left[ \mathbf{x}_n^T (\mathbf{F}_{n-1} + \beta_{n-1} \mathbf{I})^2 \mathbf{x}_n + \delta \bar{h}_{n-1}^2 \right]^{-1} e_n.$$

Alternatively, one may specify a *diversity* favoring transformation on the parameter space. One such transformation is,

$$f_d(h_{i,n}) = \bar{h}_n (|h_{i,n}| + \beta_{n-1})^{3/2} \operatorname{sgn}(h_{i,n}). \quad (26)$$

This leads to a new algorithm, the inversely proportionate NLMS (INLMS) given by,

$$\mathbf{h}_n = \mathbf{h}_{n-1} + (\mathbf{F}_{n-1} + \beta_{n-1} \mathbf{I})^{-1} \mathbf{x}_n \bullet \quad (27)$$

$$\left[ \mathbf{x}_n^T (\mathbf{F}_{n-1} + \beta_{n-1} \mathbf{I})^{-1} \mathbf{x}_n + \delta \bar{h}_{n-1}^{-2} \right]^{-1} e_n.$$

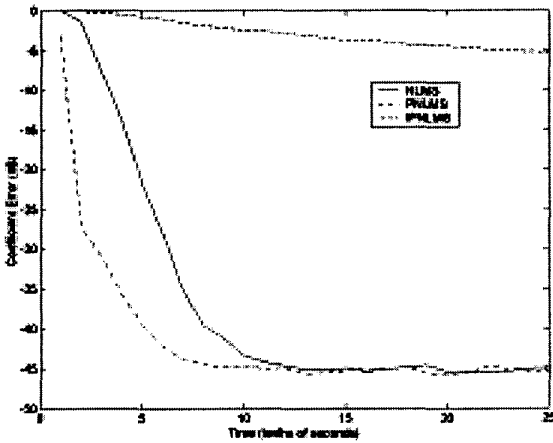


Figure 1: Convergence curves for NLMS, PNLMs and INLMS for a sparse system.

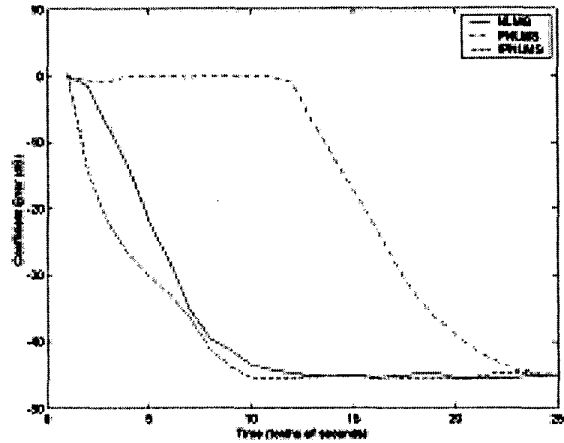


Figure 2: Convergence curves for NLMS, PNLMs and INLMS for a diverse system.

As shown below, INLMS converges faster than NLMS for non-sparse impulse responses.

The PNLMs++ [2] technique may be employed to combine coefficient update methods that favor different solutions. For instance, we may combine the PNLMs, INLMS, and NLMS. This way fast convergence may be achieved for solutions that may take on any number of properties.

#### 4. SIMULATIONS

We now explore the behaviors of the PNLMs and INLMS algorithms through simulations. In both of these simulations, the stepsize,  $\mu$  was set to 0.4, the length of the adaptive filter was 512 coefficients, and the system output signal to noise ration was 39 dB.

Fig. 1 shows the coefficient convergence of NLMS, PNLMs, and INLMS for a sparse system identification problem. The system consisted of 512 coefficients – all zero except for one set to one. PNLMs converges much faster than NLMS while INLMS doesn't converge much at all.

Fig. 2 shows the comparison of the convergences when the system to be identified has all its coefficients set to one – very non-sparse. INLMS now converges faster than NLMS (which converges at the same rate for every type of system) while PNLMs takes a while to begin its convergence.

#### 5. CONCLUSIONS

This paper has introduced a class of normalized natural gradient algorithms (NNGs) for adaptive filtering tasks. We demonstrated that PNLMs is in fact an NNG on a certain parameter space warping. Using a warping that favors impulse responses, we derived a new algorithm, INLMS that converges quickly to and accurately tracks non-sparse impulse responses.

#### REFERENCES

- [1] D. L. Duttweiler, "Proportionate normalized least-mean-squares adaptation in echo cancellers," *IEEE Trans. on Speech and Audio Proc.*, Vol.: 8 Issue: 5, Sept. 2000, pp. 508 -518
- [2] S. L. Gay, "An efficient, fast converging adaptive filter for network echo cancellation," Asilomar, Pacific Grove, CA, November 1998.
- [3] J. Benesty, S. L. Gay, "An improved PNLMs algorithm," *Proc. of ICASSP*, Orlando FL., 2002.
- [4] S. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, pp. 251-276, 1998.
- [5] S.C. Douglas and S. Amari, "Natural gradient adaptation," in *Unsupervised Adaptive Filtering, Vol. I: Blind Source Separation*, S. Haykin, ed (New York: Wiley, 2000), pp. 13-61.