AFFINE PROJECTION ALGORITHMS WITH ADAPTIVE REGULARIZATION MATRIX

Young-Seok Choi, Hyun-Chool Shin†, and Woo-Jin Song**

*Dept. of Electronics and Computer Engineering, Pohang University of Science and Technology
†Dept. of Biomedical Engineering, Johns Hopkins School of Medicine
E-mail: wjsong@postech.ac.kr

ABSTRACT

We propose a family of novel affine projection algorithms (APA) with adaptive regularization matrix. Conventional regularized APA (R-APA) uses a fixed and weighted identity matrix for regularization. The proposed algorithms incorporate a non-identity regularization matrix which is also dynamically updated. The matrix adaptation is based on the normalized stochastic-gradient of mean-square error. As a result, the efficient and robust algorithms are derived. Throughout experiments, we illustrate that the proposed algorithms show better performance than the conventional R-APA and comparable to the RLS algorithm in terms of the convergence rate and the misadjustment error.

1. INTRODUCTION

The normalized least mean square (NLMS) is the most frequently used adaptive algorithm due to its simplicity and ease of implementation. However, its convergence rate is significantly reduced for colored input signals [1]–[3]. To overcome this problem, the affine projection algorithm (APA) was proposed by Ozeki and Umeda [4]. While the NLMS updates the weights based only on the current input vectors, the APA updates the weights on the basis of the last K input vectors [4][5]. In the APA, the inversion of a rank deficient matrix may give rise to the singularity. To avoid this situation, a positive constant δ called the *regularization parameter* is used. We use the regularized APA (R-APA) as opposed to the simple APA in order to highlight the presence of the regularization parameter δ; the terminology APA is reserved for the case $\delta = 0$. Recently, it was known that the regularization parameter also plays a critical role in the convergence performance of the R-APA [7][8]. In the R-APA, the regularization parameter δ governs the rate of convergence and the misadjustment error. To meet the conflicting requirements of fast convergence and low misadjustment error, the regularization parameter needs to be optimized and updated.

In this paper, we propose a family of novel APA with adaptive regularization matrix. Conventional R-APA uses a fixed and weighted identity matrix for regularization. The key point of the paper is in the use of adaptive non-identity matrix instead of conventional fixed and weighted identity matrix. The proposed algorithms incorporate a non-identity regularization matrix which is also dynamically updated. The matrix adaptation is based on the normalized stochastic-gradient of mean-square error. As a result, the efficient and robust algorithms are derived. We show that the proposed algorithm has low additional complexity compared to the conventional R-APA. Throughout experiments, we illustrate that the proposed algorithms show better performance than the conventional R-APA and comparable to the Recursive Least Square (RLS) algorithm in terms of the convergence rate and the misadjustment error.

This work was supported by the Brain Korea (BK) 21 Program funded by the Ministry of Education and HY-SDR Research Center at Hanyang University under the ITRC Program of MIC, Korea.

2. PROPOSED R-APA

Consider data $d(i)$ that arise from the system identification model

$$d(i) = \mathbf{u}_i \mathbf{w}^\circ + v(i), \tag{1}$$

where $\mathbf{w}^\circ$ is a column vector for the impulse response of an unknown system that we wish to estimate, $v(i)$ accounts for measurement noise and $\mathbf{u}_i$ denotes the $1 \times M$ input vector,

$$\mathbf{u}_i = [u(i) \;\; u(i-1) \;\; \cdots u(i-M+1)]. \tag{2}$$

2.1. Conventional Regularization in R-APA

Let $\mathbf{w}_i$ be an estimate for $\mathbf{w}^\circ$ at iteration i. The R-APA computes $\mathbf{w}_i$ via

$$\mathbf{w}_i = \mathbf{w}_{i-1} + \mu U_i^* (U_i U_i^* + \delta I)^{-1} \mathbf{e}_i, \tag{3}$$

where

$$U_i = \begin{bmatrix} \mathbf{u}_i \\ \mathbf{u}_{i-1} \\ \vdots \\ \mathbf{u}_{i-K+1} \end{bmatrix} \quad \mathbf{d}_i = \begin{bmatrix} d(i) \\ d(i-1) \\ \vdots \\ d(i-K+1) \end{bmatrix},$$

$\mathbf{e}_i = \mathbf{d}_i - U_i \mathbf{w}_{i-1}$, μ is the step-size, δ is the regularization parameter and * denotes the Hermitian transpose. The obvious effect of the regularization parameter not only employs to avoid the inversion of a rank deficient matrix $U_i U_i^*$, but also plays a critical role in the convergence performance of the R-APA. A large regularization parameter will ensure small effective step-size and thus the R-APA results in small misadjustment error in steady state, but converges slowly. On the other hand, a small regularization parameter will provide large effective step-size and thus the R-APA converges fast but results in large misadjustment error. Along this line of thought we may expect performance improvement by optimizing and updating the regularization parameter instead of a fixed δ.

2.2. Proposed R-APA with Adaptive Regularization Matrix

To achieve this purpose, we incorporate a non-identity regularization matrix which is also dynamically updated so that $J(i) = \frac{1}{2} e^2(i)$ is minimized where $e(i) = d(i) - \mathbf{u}_i \mathbf{w}_{i-1}$. We start with a APA formulation with a non-identity regularization matrix, Δ_i:

$$\mathbf{w}_i = \mathbf{w}_{i-1} + \mu U_i^* (U_i U_i^* + \Delta_i)^{-1} \mathbf{e}_i, \tag{4}$$

where Δ_i is a $K \times K$ diagonal regularization matrix defined by

$$\Delta_i = \text{diag}\left[\delta_0(i), \delta_1(i), \ldots, \delta_{K-1}(i)\right]. \tag{5}$$

To adapt the regularization parameter, we use a stochastic gradient descent approach which can recursively minimize $E[e^2(i)]$, i.e.,

$$\begin{aligned}\delta_j(i) &= \delta_j(i-1) - \rho \nabla_\delta J(i) \\ &\text{for } j = 0, 1, \ldots, K-1,\end{aligned} \tag{6}$$

where ρ is a small positive learning rate parameter. The gradient of $J(i)$ with respect to $\delta_j(i-1)$, $\nabla_\delta J(i)$, is given by

$$\nabla_\delta J(i) = \frac{\partial J(i)}{\partial e(i)} \cdot \frac{\partial e(i)}{\partial \mathbf{w}_{i-1}} \cdot \frac{\partial \mathbf{w}_{i-1}}{\partial \delta_j(i-1)}. \tag{7}$$

Each term of right-hand side (RHS) of (7) is simply derived as

$$\frac{\partial J(i)}{\partial e(i)} = e(i), \qquad \frac{\partial e(i)}{\partial \mathbf{w}_{i-1}} = -\mathbf{u}_i, \tag{8a}$$

$$\begin{aligned}\frac{\partial \mathbf{w}_{i-1}}{\partial \delta_j(i-1)} &= -\mu U_{i-1}^*(U_{i-1}U_{i-1}^* + \Delta_{i-1})^{-1} \cdot \\ &\cdot \frac{\partial \Delta_{i-1}}{\partial \delta_j(i-1)}(U_{i-1}U_{i-1}^* + \Delta_{i-1})^{-1}\mathbf{e}_{i-1}.\end{aligned} \tag{8b}$$

More detailed derivation of (8b) is in Appendix. Then we have

$$\nabla_\delta J(i) = \mu e(i) \mathbf{u}_i U_{i-1}^* \Gamma_j \mathbf{e}_{i-1}, \tag{9}$$

where we are defining

$$\Gamma_j \triangleq (U_{i-1}U_{i-1}^* + \Delta_{i-1})^{-1} \frac{\partial \Delta_{i-1}}{\partial \delta_j(i-1)} (U_{i-1}U_{i-1}^* + \Delta_{i-1})^{-1}.$$

Note that $\Delta\delta_j(i) = \delta_j(i) - \delta_j(i-1)$ is a function of $e(i)$ and $|\Delta\delta_j(i)|$ is proportional to $|e(i)|$ since

$$|\Delta\delta_j(i)| = \rho\mu|e(i)| \cdot |\mathbf{u}_i U_{i-1}^* \Gamma_j \mathbf{e}_{i-1}|.$$

This implies that a small $e(i)$ after the initial convergence results in too small change in $\Delta\delta_j(i)$ and correspondingly $\delta_j(i)$ undergoes small variation. This is undesirable since the regularization parameters should be increasingly larger along with iteration to guarantee the lower misadjustment error.

Motivated by this fact, we normalize the gradient, $\nabla_\delta J(i)$, by its norm. By introducing the normalized gradient, the regularization parameter $\delta_j(i)$ becomes robust to variation of $e(i)$ since the normalized version of gradient $\nabla_\delta J(i)$ with a fixed ρ always makes the same stride, independent of how steep the slope of $J(i)$ is. This property makes the regularization parameter $\delta_j(i)$ relatively stable when $\nabla_\delta J(i)$ is very small. Thus the regularization parameter $\delta_j(i)$ is recursively updated by

$$\delta_j(i) = \delta_j(i-1) - \rho \frac{\nabla_\delta J(i)}{|\nabla_\delta J(i)|}. \tag{10}$$

Then, $\frac{\nabla_\delta J(i)}{|\nabla_\delta J(i)|}$ in (10) can be rewritten by

$$\frac{\nabla_\delta J(i)}{|\nabla_\delta J(i)|} = \text{sgn}(\nabla_\delta J(i)), \tag{11}$$

where $\text{sgn}(\cdot)$ is the signum function which takes the sign of variable.

From (9), (10) and (11), the proposed R-APA with adaptive regularization matrix is summarized as

$$\delta_j'(i) = \delta_j(i-1) - \rho \text{sgn}(\mu e(i) \mathbf{u}_i U_{i-1}^* \Gamma_j \mathbf{e}_{i-1}) \tag{12a}$$

$$\delta_j(i) = \begin{cases} \delta_j'(i) & \text{if } \delta_j'(i) \geq \delta_{\min} \\ \delta_{\min} & \text{if } \delta_j'(i) < \delta_{\min} \end{cases} \tag{12b}$$

$$\Delta_i = \text{diag}\left(\delta_0(i),\, \delta_1(i), \ldots, \delta_{K-1}(i)\right) \tag{12c}$$

$$\mathbf{w}_i = \mathbf{w}_{i-1} + \mu U_i^* (U_i U_i^* + \Delta_i)^{-1} \mathbf{e}_i, \tag{12d}$$

where $\delta_{\min}$ is a minimum allowable value of $\delta_j(i)$.

By setting

$$\delta(i) = \delta_0(i) = \cdots = \delta_{K-1}(i), \tag{13}$$

we can get a simpler form of (12d):

$$\begin{aligned}&\delta'(i) = \delta(i-1) - \\ &-\rho \text{sgn}\Big(\mu e(i) \mathbf{u}_i U_{i-1}^* \left(U_{i-1}U_{i-1}^* + \delta(i-1)I\right)^{-2} \mathbf{e}_{i-1}\Big)\end{aligned} \tag{14a}$$

$$\delta(i) = \begin{cases} \delta'(i) & \text{if } \delta'(i) \geq \delta_{\min} \\ \delta_{\min} & \text{if } \delta'(i) < \delta_{\min} \end{cases} \tag{14b}$$

$$\mathbf{w}_i = \mathbf{w}_{i-1} + \mu U_i^* (U_i U_i^* + \delta(i) I)^{-1} \mathbf{e}_i, \tag{14c}$$

in which a weighted identity matrix is used for regularization like the conventional R-APA but the weighted identity matrix is adaptive here.

In addition, when $K = 1$, we get a new regularized NLMS (R-NLMS) algorithm. From (12a) or (14a), the regularized NLMS with adaptive regularization parameter reduces to

$$\begin{aligned}\delta'(i) &= \delta(i-1) - \rho \text{sgn}\left(\mu \frac{e(i)e(i-1)\mathbf{u}_i \mathbf{u}_{i-1}{}^*}{\left(\|\mathbf{u}_{i-1}\|^2 + \delta(i-1)\right)^2}\right) \\ &= \delta(i-1) - \rho \text{sgn}(e(i)e(i-1)\mathbf{u}_i \mathbf{u}_{i-1}{}^*)\end{aligned} \tag{15a}$$

$$\delta(i) = \begin{cases} \delta'(i) & \text{if } \delta'(i) \geq \delta_{\min} \\ \delta_{\min} & \text{if } \delta'(i) < \delta_{\min}. \end{cases} \tag{15b}$$

$$\mathbf{w}_i = \mathbf{w}_{i-1} + \mu \frac{\mathbf{u}_i^*}{\|\mathbf{u}_i\|^2 + \delta(i)} e(i). \tag{15c}$$

Table 1 lists the number of multiplications, additions of the computation of the adaptive regularization matrix at each iteration. It is known [3] that the complexity of the conventional R-APA is $(K^2 + 2K)M + K^3 + K^2$ multiplications and $(K^2 + 2K)M + K^3$ additions.

2.3. Stability

To guarantee the stability of the proposed algorithms, we need to set $\delta_{\min}$. Let us define the *a posteriori* estimation error as

$$\mathbf{r}_i = \mathbf{d}_i - U_i \mathbf{w}_i, \tag{16}$$

i.e., the error in estimating $\mathbf{d}_i$ by using the new weight estimate. Since $U_i\mathbf{w}_i$ will be a better estimate for $\mathbf{d}_i$ than $U_i\mathbf{w}_{i-1}$, the property $\|\mathbf{r}_i\|^2 \leq \|\mathbf{e}_i\|^2$ (with equality only when $\mathbf{e}_i = 0$) should be satisfied. Assuming a scalar regularization parameter as (14c), it holds that

$$\mathbf{r}_i = \left(I - \mu U_i U_i^* (U_i U_i^* + \delta(i) I)^{-1}\right) \mathbf{e}_i, \tag{17}$$

and

$$\|\mathbf{r}_i\|^2 = \mathbf{e}_i^* A^* A \mathbf{e}_i \leq \mathbf{e}_i^* I \mathbf{e}_i = \|\mathbf{e}_i\|^2, \tag{18}$$

Table 1. Computational complexity of adaptive regularization matrix

Algorithm	multiplications	additions
Proposed (12a)	$KM + K^2 + 3K + 1$	$KM + K^2 - K$
Proposed (14a)	$KM + K^2 + 2K + 2$	$KM + K^2 - K$

where we are defining

$$A \triangleq \left(I - \mu U_i U_i^* (U_i U_i^* + \delta(i) I)^{-1}\right).$$

Therefore, $\|\mathbf{r}_i\|^2 \leq \|\mathbf{e}_i\|^2$, if and only if the matrix $(I - A^* A)$ is positive-definite by (18). In addition, let $U_i U_i^* = V_i \Lambda_i V_i^*$ denotes the eigen-decomposition of the matrix $U_i U_i^*$, where V_i is unitary and $\Lambda_i = \text{diag}\,[\lambda_o(i), \lambda_1(i), \cdots, \lambda_{K-1}(i)]$ contains the corresponding eigenvalues. Then

$$U_i U_i^* + \delta(i) I = V_i (\Lambda_i + \delta(i) I) V_i^*, \tag{19}$$

and

$$(U_i U_i^* + \delta(i) I)^{-1} = V_i (\Lambda_i + \delta(i) I)^{-1} V_i^*. \tag{20}$$

Using the eigen-decomposition of $U_i U_i^*$ and (20), the following holds:

$$\begin{aligned} A^* A &= (I - \mu V_i \Lambda_i^{'} V_i^*)^* (I - \mu V_i \Lambda_i^{'} V_i^*) \\ &= I - 2\mu V_i \Lambda_i^{'} V_i^* + \mu^2 V_i \Lambda_i^{'^2} V_i^*, \end{aligned} \tag{21}$$

where $\Lambda_i^{'} = \Lambda_i (\Lambda_i + \delta(i) I)^{-1}$. So, it holds that

$$I - A^* A = \mu V_i \Lambda_i^{'} (2I - \mu \Lambda_i^{'}) V_i^*. \tag{22}$$

To satisfy $(I - A^* A)$ is positive-definite, we find

$$\begin{aligned} &2I - \mu \Lambda_i^{'} = 2I - \mu \Lambda_i (\Lambda_i + \delta(i) I)^{-1} \\ &= 2I - \mu\, \text{diag}\left(\frac{\lambda_o(i)}{\lambda_o(i) + \delta(i)}, \cdots, \frac{\lambda_{K-1}(i)}{\lambda_{K-1}(i) + \delta(i)}\right) > 0. \end{aligned} \tag{23}$$

Then, we get the lower bound of the regularization parameter for the stability of the proposed algorithms as

$$\delta_{\min} > \lambda_{\max}(i) \left(\frac{\mu}{2} - 1\right), \tag{24}$$

where $\lambda_{\max}(i)$ is a maximum value of $\lambda_k(i)$ with $1 \leq k \leq K - 1$. Also it is known that the convergence in the mean of R-APA is guaranteed for any μ satisfying [2][3]

$$0 < \mu < 2. \tag{25}$$

3. EXPERIMENTAL RESULTS

We illustrate the performance of the proposed algorithms by carrying out computer simulations in a channel identification scenario. The unknown channel $H(z)$ has 16 taps and is randomly generated. The adaptive filter and the unknown channel are assumed to have the same number of taps. A Gaussian distributed signal is used for the input signal. The input signal is obtained by filtering a white, zero-mean. Gaussian random sequence through a first-order system $G(z) = 1/(1 - 0.9z^{-1})$. The signal-to-noise ratio (SNR) is calculated by SNR $= 10 \log_{10}(E[y^2(i)]/E[v^2(i)])$,

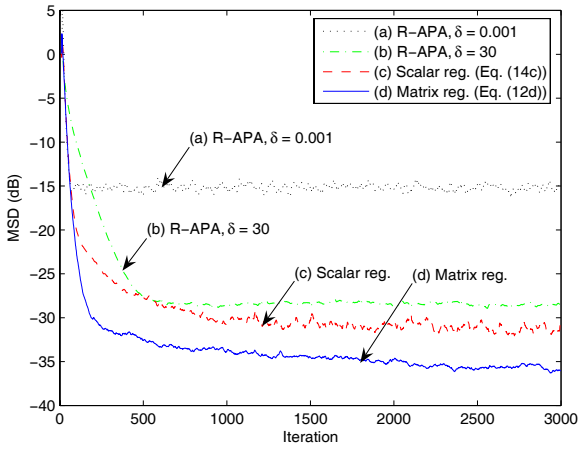

Fig. 1. Performance of the proposed APAs in (12d) and (14c), and conventional R-APA (K=8)

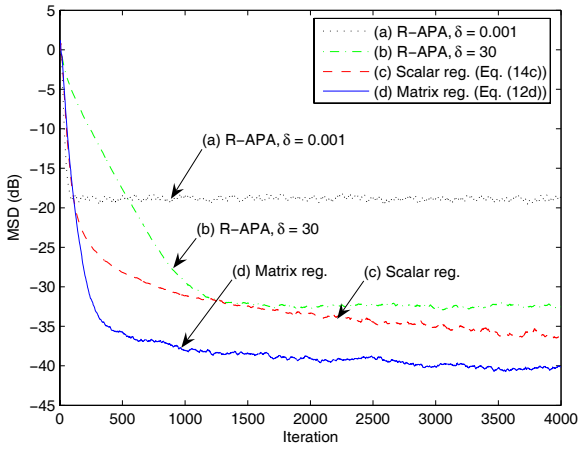

Fig. 2. Performance of the proposed APAs in (12d) and (14c), and conventional R-APA (K=4)

where $y(i) = \mathbf{u}_i \mathbf{w}^\circ$. The measurement noise $v(i)$ is added to $y(i)$ such that SNR $=$ 30dB. The mean square deviation (MSD), $E\|\mathbf{w}^\circ - \mathbf{w}_i\|^2$, is taken and averaged over 100 independent trials. The initial value $\delta_j(0)$ is set to 0.001 and $\delta_{\min}$ is chosen to 0.0001 for all experiments. The step-size is always $\mu = 0.5$.

In Fig. 1, we show the MSD curves for $K = 8$ and $\rho = 1.0$. Dashed lines indicate the results of R-APA with fixed regularization parameters where we choose $\delta = 0.001$ and 30. As can be seen, the proposed R-APA has the faster convergence and the lower misadjustment error. In addition, the proposed R-APA with adaptive non-identity regularization matrix has a improved performance than with adaptive scalar regularization parameter as expected. In Fig. 2, we choose $K = 4$ and $\rho = 0.9$. Fig. 3 shows the performance of the proposed R-NLMS where $\rho = 0.005$. For the comparison purpose, the GNGD [8] is presented using same ρ. A similar result of Fig. 1 is observed in Fig. 2 and Fig. 3.

Finally, Fig. 4 demonstrates the performance comparison of the

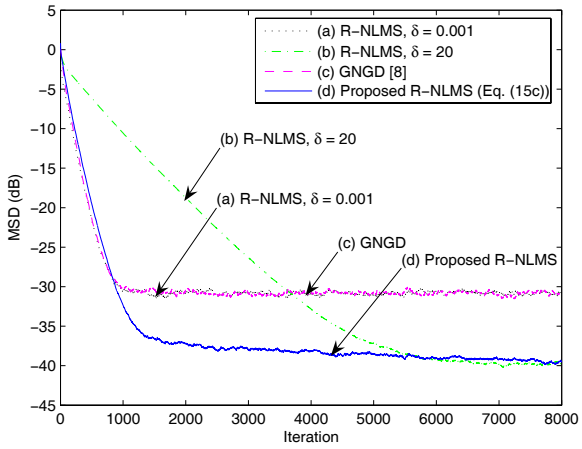

Fig. 3. Performance comparison of the proposed NLMS in (15c), GNGD [8], and conventional R-NLMS (K=1)

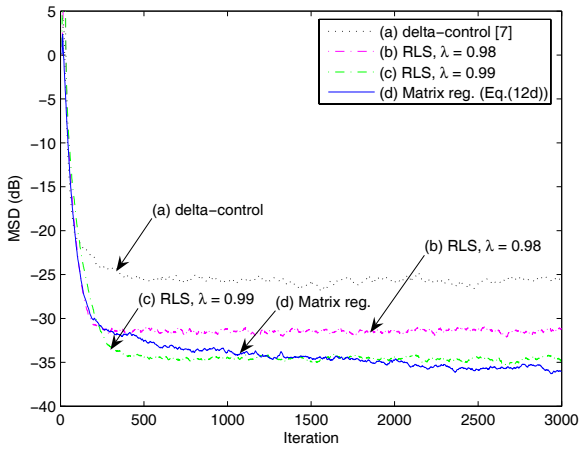

Fig. 4. Performance comparison of the proposed APA in (12d) (K=8), R-APA with delta-control [7] (K=8), and RLS

proposed R-APA, R-APA with the delta-control method [7], and the RLS where $K = 8$. We choose the forgetting factor as $\lambda = 0.98$ and 0.99 for the RLS. In the figure, we know that the proposed R-APA outperforms the delta-control method and is comparable to the RLS.

4. CONCLUSION

We have presented a family of novel R-APA with adaptive regularization matrix. The matrix is more general and robust than the conventional R-APA in that the matrix is non-identity and dynamically updated. As a result, we highly improved the convergence performance, which is even comparable to the RLS. Although here we limited to diagonal regularization matrix, we may expect further improvement by extending it to general square matrix. Also computational reduction in adaptation of the regularization matrix is a challenging subject.

Appendix

We start with

$$\frac{\partial \mathbf{w}_{i-1}}{\partial \delta_j(i-1)} = \mu \frac{\partial U_{i-1}^* (U_{i-1}U_{i-1}^* + \Delta_{i-1})^{-1}\mathbf{e}_{i-1}}{\partial \delta_j(i-1)}$$
$$= \mu U_{i-1}^* \frac{\partial (U_{i-1}U_{i-1}^* + \Delta_{i-1})^{-1}}{\partial \delta_j(i-1)} \mathbf{e}_{i-1}. \quad (26)$$

Let us define

$$Y \triangleq (U_{i-1}U_{i-1}^* + \Delta_{i-1}).$$

From [9], we know that by differentiating $YY^{-1} = I$ with respect to $\delta_j(i-1)$

$$\frac{\partial Y}{\partial \delta_j(i-1)} Y^{-1} + Y \frac{\partial Y^{-1}}{\partial \delta_j(i-1)} = 0. \quad (27)$$

Then, we get

$$\frac{\partial Y^{-1}}{\partial \delta_j(i-1)} = -Y^{-1} \frac{\partial Y}{\partial \delta_j(i-1)} Y^{-1}. \quad (28)$$

Now we substitute $Y = (U_{i-1}U_{i-1}^* + \Delta_{i-1})$ into (28), then

$$\frac{\partial (U_{i-1}U_{i-1}^* + \Delta_{i-1})^{-1}}{\partial \delta_j(i-1)}$$
$$= -(U_{i-1}U_{i-1}^* + \Delta_{i-1})^{-1} \frac{\partial (U_{i-1}U_{i-1}^* + \Delta_{i-1})}{\partial \delta_j(i-1)} \cdot$$
$$\cdot (U_{i-1}U_{i-1}^* + \Delta_{i-1})^{-1}$$
$$= -(U_{i-1}U_{i-1}^* + \Delta_{i-1})^{-1} \frac{\partial \Delta_{i-1}}{\partial \delta_j(i-1)} (U_{i-1}U_{i-1}^* + \Delta_{i-1})^{-1}. \quad (29)$$

Using (26) and (29), we obtain (8b).

5. REFERENCES

[1] B. Widrow and S. D. Sterns, *Adaptive Signal Processing*, Englewood Cliffs, NJ: Prentice Hall, 1985.

[2] S. Haykin, *Adaptive Filter Theory*, 4th edition, Upper Saddle River, NJ: Prentice Hall, 2002.

[3] A. H. Sayed, *Fundamentals of Adaptive Filtering*, New York: Wiley, 2003.

[4] K. Ozeki and T. Umeda, "An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties," *Electro. Commun. Jpn.*, vol. 67-A, no. 5, pp. 19–27, 1984.

[5] H.-C. Shin and A. H. Sayed, "Mean-square peformance of a family of affine projection algorithms," *IEEE Trans. Signal Processing*, vol. 52, pp. 90–102, Jan. 2004.

[6] S. C. Douglas, "Generalized gradient adaptive step sizes for stochastic gradient adaptive filters," in *Proc. IEEE Int. Conf. on Accoustics, Speech, and Signal Processing, ICASSP'95*, vol. 2, pp. 1396–1399, 1995.

[7] V. Myllylä and G. Schmidt, "Pseudo-optimal regulariztion for affine projection algorithms," in *Proc. IEEE Int. Conf. on Accoustics, Speech, and Signal Processing, ICASSP'02*, Orlando, Florida, May 2002, pp. 1917–1920.

[8] D. P. Mandic, "A generalized normalized gradient descent algorithm," *IEEE Signal Processing Lett.*, vol. 11, No. 2, pp. 115–118, Feb. 2004.

[9] T. K. Moon and W. C. Stirling, *Mathmatical Methods and Algorithms for Signal Processing*, Upper Saddle River, NJ: Prentice Hall, 1999.