Convergence Analysis of Local Algorithms for Blind Decorrelation

Scott C. Douglas
Dept. of Electrical Engineering
University of Utah
Salt Lake City, UT 84112 USA

Andrzej Cichocki
Brain Information Processing Group
Frontier Research Program, RIKEN
Wako-Shi, Saitama 351-01 JAPAN

Abstract

In this paper, we analyze and extend a class of adaptive networks for second-order blind decorrelation of instantaneous signal mixtures. Firstly, we compare the performance of the decorrelation neural network employing global knowledge of the adaptive coefficients in [27] with a similar structure whose coefficients are adapted via local output connections in [8]. Through statistical analyses, the convergence behaviors and stability bounds for the algorithms' step sizes are studied and derived. Secondly, we analyze the behaviors of locally-adaptive multilayer decorrelation networks and quantify their performances for poorly-conditioned signal mixtures. Thirdly, we derive a robust locally-adaptive network structure based on *a posteriori* output signals that remains stable for any step size value. Finally, we present an extension of the locally-adaptive network for linear-phase temporal and spatial whitening of multichannel signals. Simulations verify the analyses and indicate the usefulness of the locally-adaptive networks for decorrelating signals in space and time.

1 INTRODUCTION

Blind signal separation is useful for numerous problems in acoustics, communications, biomedical signal analysis, and image processing. In blind source separation of instantaneous signal mixtures, a set of measured signals $\{x_i(k)\}$, $1 \leq i \leq n$ is assumed to be generated from a set of unknown stochastic, independent sources $\{s_i(k)\}$, $1 \leq i \leq n$, as

$$\mathbf{x}(k) \quad = \quad \mathbf{A}\mathbf{s}(k), \tag{1}$$

where $\mathbf{x}(k) = [x_1(k) \cdots x_n(k)]^T$, $\mathbf{s}(k) = [s_1(k) \cdots s_n(k)]^T$, and $\mathbf{A}$ is an unknown matrix of n^2 mixing coefficients a_{ij}. Implicit in this model is the assumption that

the number of sensors measuring the signals $x_i(k)$ equals the number of sources $s_i(k)$. The measured sensor signals are processed by a linear single-layer feedforward network as

$$\mathbf{y}(k) \quad = \quad \mathbf{W}(k)\mathbf{x}(k), \tag{2}$$

where $\mathbf{W}(k)$ is an $(n \times n)$-dimensional synaptic weight matrix. Ideally, $\mathbf{W}(k)$ is adjusted iteratively such that

$$\lim_{k \to \infty} \mathbf{W}(k)\mathbf{A} \quad = \quad \mathbf{PD}, \tag{3}$$

where $\mathbf{P}$ is an $(n \times n)$-dimensional permutation matrix with a single unity entry in any of its rows or columns and $\mathbf{D}$ is a diagonal nonsingular scaling matrix.

Recently, several simple, efficient, and iterative algorithms for adjusting $\mathbf{W}(k)$ have been proposed for the blind signal separation task [1, 5, 8, 10, 14]. Such methods use higher-order statistical information about the source signals to iteratively adjust the coefficient matrix $\mathbf{W}(k)$. A large class of these on-line adaptive algorithms can be represented by the generalized algorithm given by

$$\mathbf{W}(k+1) \quad = \quad \mathbf{W}(k) + \eta(k)\mathbf{G}(k), \tag{4}$$

where $\mathbf{G}(k)$ is a matrix that depends on $\mathbf{y}(k)$ and/or $\mathbf{W}(k)$ and $\eta(k)$ is a step size sequence. Some choices of $\mathbf{G}(k)$ are

$$\mathbf{G}(k) \quad = \quad \left(\mathbf{I} - \mathbf{f}(\mathbf{y}(k))\mathbf{g}(\mathbf{y}^T(k))\right)\mathbf{W}(k) \tag{5}$$

$$\mathbf{G}(k) \quad = \quad \left(\mathbf{I} - \mathbf{y}(k)\mathbf{y}^T(k) + \mathbf{g}(\mathbf{y}(k))\mathbf{y}^T(k) - \mathbf{y}(k)\mathbf{g}(\mathbf{y}^T(k))\right)\mathbf{W}(k) \tag{6}$$

$$\mathbf{G}(k) \quad = \quad \mathbf{I} - \mathbf{f}(\mathbf{y}(k))\mathbf{g}(\mathbf{y}^T(k)), \tag{7}$$

where $\mathbf{f}(\mathbf{y}(k))$ and $\mathbf{g}(\mathbf{y}(k))$ are vector-valued nonlinear or linear functions of the elements of $\mathbf{y}(k)$ whose forms are related to the statistics of the source signals $\{s_i(k)\}$. While the steady-state performances of these algorithms can sometimes be characterized [5], it is quite challenging to determine their transient behaviors, particularly as they depend on any convergence parameters for the chosen structure and algorithm. Without this understanding, choosing $\eta(k)$ to obtain stable behaviors of these iterative methods is a difficult and time-consuming task.

A related task to blind signal separation is that of multichannel signal decorrelation, in which the autocorrelation matrix given by

$$\mathbf{R}_{xx}(k) \quad = \quad E\{\mathbf{x}(k)\mathbf{x}^T(k)\} \tag{8}$$

is well-defined and $E\{\cdot\}$ denotes statistical expectation. In this case, the goal is to adjust $\mathbf{W}(k)$ in (2) such that the elements of $\mathbf{y}(k)$ are uncorrelated as $k \to \infty$. Typically, we desire

$$\lim_{k \to \infty} \mathbf{R}_{yy}(k) \quad = \quad \mathbf{I}, \tag{9}$$

where $\mathbf{R}_{yy}(k) = E\{\mathbf{y}(k)\mathbf{y}^T(k)\}$. The value of $\mathbf{W}(k)$ that achieves this result is

$$\lim_{k \to \infty} \mathbf{W}(k) \quad = \quad \overline{\mathbf{Q}}\mathbf{R}_{xx}^{-1/2}(k), \tag{10}$$

where $\overline{\mathbf{Q}}$ is any Hermitian matrix and $\mathbf{R}_{xx}^{-1/2}(k)$ is the symmetric square-root factor of the inverse of $\mathbf{R}_{xx}(k)$. Signal decorrelation can be used as a preprocessing step in systems such as adaptive filters and multilayer neural networks to improve their adaptation performances [17, 18].

Silva and Almeida propose and analyze in [27, 28] the following algorithm for multichannel signal decorrelation:

$$\mathbf{W}(k+1) \quad = \quad \mathbf{W}(k) + \eta(k)(\mathbf{I} - \mathbf{y}(k)\mathbf{y}^T(k))\mathbf{W}(k). \tag{11}$$

For $\mathbf{f}(\mathbf{y}(k)) = \mathbf{g}(\mathbf{y}(k)) = \mathbf{y}(k)$, (11) is equivalent to the algorithm in (4) where $\mathbf{G}(k)$ is as chosen in (5) or (6). The algorithm in (11) is a first-order in $\eta(k)$ instantaneous approximation to the iterative Potter formula for finding $\mathbf{R}_{xx}^{-1/2}$ [26]. The algorithm also possesses the so-called "equivariance property" such that its average performance does not depend on the eigenvalues of $\mathbf{R}_{xx}(k) = \mathbf{R}_{xx}$ [5]. A similar algorithm for multichannel decorrelation is a special case of (4) where $\mathbf{G}(k)$ is given by (7) with $\mathbf{f}(\mathbf{y}(k)) = \mathbf{g}(\mathbf{y}(k)) = \mathbf{y}(k)$. This update is [8]

$$\mathbf{W}(k+1) \quad = \quad \mathbf{W}(k) + \eta(k)(\mathbf{I} - \mathbf{y}(k)\mathbf{y}^T(k)). \tag{12}$$

In addition to being simpler to implement than (11), the localized memory requirements of the algorithm in (12) make it ideal for hardware and VLSI implementation. However, the performances of (11) and (12) are not the same, and it is not clear how to choose both $\mathbf{W}(0)$ and $\eta(k)$ to obtain the best performance from each system. A theoretical performance comparison of these two algorithms can also give some insight as to the general performance characteristics of the blind signal separation algorithms in (4)-(7).

In addition to the above issues, two extensions of the above structures merit further study. In [8], it is demonstrated through simulations that a multilayer network of cascaded blind signal separation systems of the form in (4) with $\mathbf{G}(k)$ chosen as in (7) has an overall performance that is similar to that of the equivariant algorithm with $\mathbf{G}(k)$ chosen as in (5). In another open problem, it is not clear how the decorrelation network in (12) can be extended to the task of multichannel spatial and temporal decorrelation of sequences such that

$$\lim_{k \to \infty} E\{\mathbf{y}(k)\mathbf{y}^T(k-i)\} \quad = \quad \mathbf{I}\delta(i), \quad -L < i < L \tag{13}$$

for a given integer value of L is obtained.

The purpose of this paper is fourfold. Firstly, we compare the average behaviors of the two decorrelation networks in (11) and (12), respectively, assuming that $E\{\mathbf{x}(k)\mathbf{x}^T(k-i)\} = \mathbf{R}_{xx}\delta(i)$. Our analysis techniques are similar to those in [28]; however, we provide additional insight as to the practical choice of a *time-varying* $\eta(k)$ to provide stable, fast, and robust adaptation of the coefficient matrix $\mathbf{W}(k)$. Secondly, we analyze the multilayer network in [8] for the case $\mathbf{f}(\mathbf{y}) = \mathbf{g}(\mathbf{y}) = \mathbf{y}$, showing that by cascading several linear distributed systems, each with the localized learning rule in (12), the overall mean convergence behavior of the system is faster than that of a single-layer system. Thirdly, we develop an *a posteriori* version of (12) that has similar properties in simulation to those of *a posteriori* algorithms in adaptive filters; namely, stable behavior for any positive-valued step size sequence $\eta(k)$ [16]. Finally, we provide an extension of the algorithm in (12) to a multichannel FIR filter structure that provides joint spatial and temporal decorrelation of multichannel convolved signals. Simulations show the effectiveness of the algorithms in multichannel and time-series decorrelation tasks.

2 DERIVATION OF THE ALGORITHMS

The two decorrelation networks in (11) and (12) can be derived as stochastic gradient versions of modified steepest descent procedures on a suitably-chosen cost function. Both updates are of the form [1]

$$\mathbf{W}(k+1) \quad = \quad \mathbf{W}(k) - \eta(k)\frac{\partial J(\mathbf{W}(k))}{\partial \mathbf{W}(k)}\mathbf{M}(k) \tag{14}$$

$$\frac{\partial J(\mathbf{W}(k))}{\partial \mathbf{W}(k)} \quad = \quad \mathbf{y}(k)\mathbf{x}^T(k) - \mathbf{W}^{-T}(k), \tag{15}$$

where $J(\mathbf{W})$ is the instantaneous approximation to the Kullback-Leibler divergence between two zero-mean normal distributions with covariances $E\{\mathbf{y}(k)\mathbf{y}^T(k)\}$ and $\mathbf{I}$, respectively [5], and $\mathbf{M}(k)$ is a preconditioning matrix chosen to simplify the implementation and improve the convergence characteristics of the system. By choosing $\mathbf{M}(k) = \mathbf{W}^T(k)\mathbf{W}(k)$ and $\mathbf{M}(k) = \mathbf{W}^T(k)$, respectively, one obtains the algorithms in (11) and (12). Both choices of $\mathbf{M}(k)$ guarantee the stability of the system for suitably small step sizes $\eta(k)$ if $\mathbf{W}(0)$ is a positive-definite matrix. Moreover, if $\mathbf{W}(0)$ is chosen to be symmetric, then $\mathbf{W}(k)$ is guaranteed to be symmetric for each algorithm, such that only the upper diagonal elements of $\mathbf{W}(k)$ need to be computed. Without any *a priori* knowledge of $\mathbf{R}_{xx}$, $\mathbf{W}(0) = c\mathbf{I}$ is typically chosen, where c is a positive constant.

The update in (12) has an interesting property in that it also converges for suitable sequences of *negative* step sizes $\eta(k)$ if $\mathbf{W}(0)$ is a negative semi-definite matrix. To see this result, multiply both sides of (12) by (-1). By defining $\underline{\mathbf{W}}(k) = -\mathbf{W}(k)$, $\underline{\mathbf{y}}(k) = -\mathbf{y}(k) = \underline{\mathbf{W}}(k)\mathbf{x}(k)$, and $\underline{\eta}(k) = -\eta(k)$,

$$\underline{\mathbf{W}}(k+1) \quad = \quad \underline{\mathbf{W}}(k) + \underline{\eta}(k)(\mathbf{I} - \underline{\mathbf{y}}(k)\underline{\mathbf{y}}^T(k)). \tag{16}$$

This algorithm is algebraically-equivalent to that in (12), and thus the coefficient matrix $\underline{\mathbf{W}}(k)$ tends towards the solution obtained by $-\mathbf{W}(k)$ in the original algorithm. The convergence conditions on $\underline{\eta}(k)$ are the same as those for $-\eta(k)$ in the original algorithm.

3 ANALYSIS OF MULTICHANNEL DECORRELATION NETWORKS

We now analyze the algorithms in (11) and (12). Our study of the transient behaviors of these algorithms yields answers to practical issues such as (i) the evolutionary behavior of the mean of the coefficient matrix $\mathbf{W}(k)$ in terms of the signal statistics, (ii) the range of stable step sizes $\eta(k)$ to provide convergence, and (iii) the optimal step sizes for fastest convergence.

In our analyses, we study the behavior of the averaged systems given by

$$E\{\mathbf{W}(k+1)\} = E\{\mathbf{W}(k)\} + \eta(k)(\mathbf{I} - E\{\mathbf{W}(k)\}\mathbf{R}_{xx}E\{\mathbf{W}^T(k)\})E\{\mathbf{W}(k)\} \tag{17}$$

and

$$E\{\mathbf{W}(k+1)\} = E\{\mathbf{W}(k)\} + \eta(k)(\mathbf{I} - E\{\mathbf{W}(k)\}\mathbf{R}_{xx}E\{\mathbf{W}^T(k)\}), \tag{18}$$

respectively. In adaptive filtering, such analyses are used to characterize the mean trajectories of the filter coefficients for small step size values [17]. These assumptions effectively imply that (i) the input signal vector $\mathbf{x}(k)$ is independent of $\mathbf{x}(m)$ for $k \neq m$, and (ii) fluctuations in the elements of $\mathbf{W}(k)$ are small such that terms of the form $w_{ij}(k)w_{mn}(k)$ and $w_{ij}(k)w_{mn}(k)w_{pq}(k)$ can be replaced by $E\{w_{ij}(k)\}E\{w_{mn}(k)\}$ and $E\{w_{ij}(k)\}E\{w_{mn}(k)\}E\{w_{pq}(k)\}$, respectively. While these assumptions are rarely true in practice, they yield analysis equations that are reliable predictors of the networks' performances for small step size values.

3.1 MEAN BEHAVIOR OF GLOBALLY-ADAPTIVE NETWORK

In [28], the behavior of (17) is studied assuming that the step size $\eta(k) = \eta$ is fixed. Under these assumptions, fixed stability bounds on η are derived for $\mathbf{W}(0) = \mathbf{I}$. In what follows, we extend this analysis to derive time-varying step size bounds for

$\eta(k)$ that depend on the approximate eigenvalues of $\mathbf{R}_{yy}(k)$. We also determine useful step size values and initial conditions for $\mathbf{W}(0)$ from this analysis.

Using (17), we can determine an update for the errors in the n eigenvalues of the matrix $\overline{\mathbf{R}}_{yy}(k)$ defined as

$$\overline{\mathbf{R}}_{yy}(k) \quad = \quad E\{\mathbf{W}(k)\}\mathbf{R}_{xx}E\{\mathbf{W}^T(k)\}. \tag{19}$$

If the fluctuations in the elements of $\mathbf{W}(k)$ are small, $\overline{\mathbf{R}}_{yy}(k)$ is a reasonable approximation to $E\{\mathbf{y}(k)\mathbf{y}^T(k)\}$. The analysis is performed in Appendix A, in which it is shown that the error in the ith eigenvalue $\lambda_i(k)$ of $\overline{\mathbf{R}}_{yy}(k)$, defined as

$$\widetilde{\lambda}_i(k) \quad = \quad \lambda_i(k) - 1, \tag{20}$$

evolves as

$$\widetilde{\lambda}_i(k+1) \quad = \quad \Big(((1-\eta(k)\lambda_i(k))^2 - \eta^2(k)\lambda_i(k)\Big)\,\widetilde{\lambda}_i(k). \tag{21}$$

Note that as $\widetilde{\lambda}_i(k)$ tends to zero, $\overline{\mathbf{R}}_{yy}(k)$ tends to the identity matrix, and thus the elements of the output vector $\mathbf{y}(k)$ become decorrelated with respect to one another.

From (21), we can determine stability conditions on $\eta(k)$ to guarantee convergence of each $\lambda_i(k)$ to unity. This analysis is also given in Appendix A, for which sufficient conditions on $\eta(k)$ for stability of (21) are

$$0 \;<\; \eta(k) \;<\; \eta_{max}(k), \tag{22}$$

where

$$\eta_{max}(k) = \begin{cases} \dfrac{1}{\sqrt{\lambda_{max}(k)}} & \text{if } 0 < \lambda_{max}(k) \leq 3 + 2\sqrt{2} \\ \dfrac{2}{\lambda_{max}(k) - 1} & \text{if } \lambda_{max}(k) > 3 + 2\sqrt{2}, \end{cases} \tag{23}$$

and $\lambda_{max}(k)$ is the maximum eigenvalue of $\overline{\mathbf{R}}_{yy}(k)$.

Equation (21) indicates for the algorithm in (11) that the average convergence of each of the eigenvalues of $\overline{\mathbf{R}}_{yy}(k)$ only depends on its value and the value of $\eta(k)$. In other words, convergence of this system does not depend on the eigenvalues λ_i, $1 \leq i \leq n$, of $\mathbf{R}_{xx}$. However, since $\lambda_i(k)$ is related to λ_i as

$$\lambda_i(k) \quad = \quad \lambda_i \lambda_{w,i}^2(k) \tag{24}$$

where $\lambda_{w,i}(k)$ is the ith eigenvalue of $E\{\mathbf{W}(k)\}$, the eigenvalues of $\overline{\mathbf{R}}_{yy}(0)$ depend on both $\mathbf{R}_{xx}$ and the initial value $\mathbf{W}(0)$ of the coefficient matrix.

From (21), the step size value for (17) that achieves one-step convergence of the ith eigenvalue of $\overline{\mathbf{R}}_{yy}(k)$ at time k such that $\widetilde{\lambda}_i(k+1) = 0$ is

$$\eta_{i,opt}(k) \quad = \quad \frac{(1/\sqrt{\lambda_i(k)}) - 1}{1 - \lambda_i(k)}. \tag{25}$$

Figure 1 plots $\eta_{i,opt}(k)$ as a function of $\lambda_i(k)$. While $\eta_{i,opt}(k) > 0$ for all $\lambda_i(k) > 0$, the value of $\eta_{i,opt}(k)$ varies by several orders of magnitude over the range of $\lambda_i(k)$ from 0.001 to 10 for this system, complicating the initial choice of $\eta(k)$ in practice. Near the optimum solution, it is desirable to choose $\eta(k) = 1/2$ for fastest convergence of (21), as all $\lambda_i(k) \approx 1$ in this case [28].

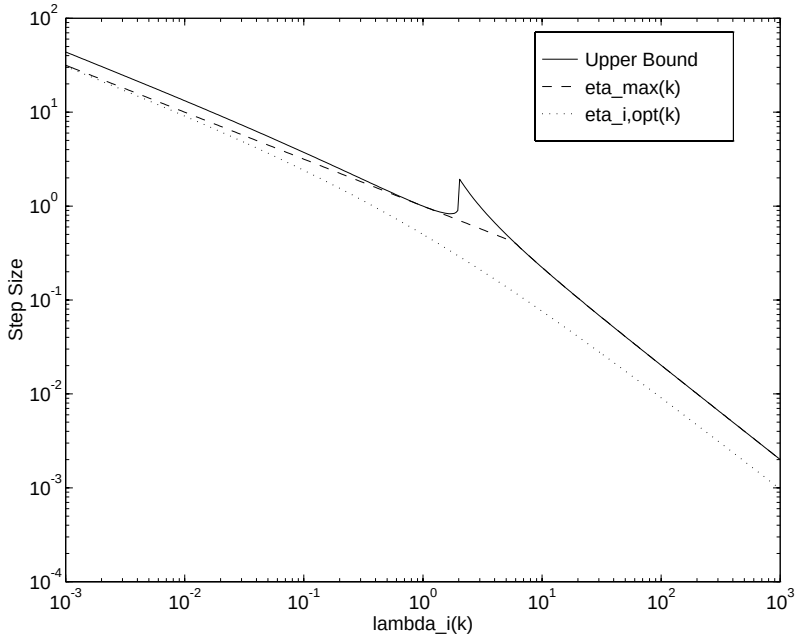

Figure 1: Upper step size bounds, sufficient bounds $\eta_{max}(k)$, and optimum step size $\eta_{i,opt}(k)$ vs. $\lambda_i(k)$ for the globally-adaptive network.

For small step sizes where $0 < \eta(k)\lambda_{max}(k) \ll 2$, we neglect all terms in (21) that are of $O(\eta^2(k))$. In this case, the error in $\lambda_i(k)$ approximately evolves according to

$$\widetilde{\lambda}_i(k+1) \quad = \quad (1 - 2\eta(k)\lambda_i(k))\widetilde{\lambda}_i(k). \tag{26}$$

This result indicates that the absolute error of the ith eigenvalue of $\overline{\mathbf{R}}_{yy}(k)$ is slow to converge if both $\lambda_i(k)$ and $\eta(k)$ are small. For this reason, it is desirable to choose the values of $\lambda_i(0)$ to be large such that fast adaptation can occur for smaller step sizes. Noting the relationship in (24), we see that the practical choice of $\mathbf{W}(0) = c\mathbf{I}$ yields $\lambda_i(0) = \lambda_i c^2$. In this case, one should choose c for the globally-adaptive algorithm in (11) such that $(1 - 2\eta(k)c^2\lambda_{min})$ is significantly smaller than one for reasonable choices of $\eta(k)$ satisfying (22)

3.2 MEAN BEHAVIOR OF LOCALLY-ADAPTIVE NETWORK

We now study the averaged algorithm in (18). In this case, it is convenient to consider the evolution of the ith eigenvalue of $E\{\mathbf{W}(k)\}$ for this update, defined as $\lambda_{w,i}(k)$. Define the error in $\lambda_{w,i}(k)$ as

$$\widetilde{\lambda}_{w,i}(k) \quad = \quad \lambda_{w,i}(k) - \frac{1}{\sqrt{\lambda_i}}, \tag{27}$$

where λ_i is the ith eigenvalue of $\mathbf{R}_{xx}$. Then, it is shown in Appendix B that $\widetilde{\lambda}_{w,i}(k)$ for (18) evolves according to

$$\widetilde{\lambda}_{w,i}(k+1) \quad = \quad \left(1 - \eta(k)\sqrt{\lambda_i}\left(1 + \sqrt{\lambda_i(k)}\right)\right)\widetilde{\lambda}_{w,i}(k), \tag{28}$$

where $\lambda_i(k)$ is defined in (24).

We can use (28) to determine stability conditions on $\eta(k)$ to guarantee convergence of $\lambda_{w,i}(k)$ to $1/\sqrt{\lambda_i}$ for the update in (18). This analysis is also provided in Appendix B and yields the necessary and sufficient conditions

$$0 \ < \ \eta(k) \ < \ \frac{2}{\sqrt{\lambda_{max}}\left(1+\sqrt{\lambda_{max}(k)}\right)}, \tag{29}$$

where λ_{max} and $\lambda_{max}(k)$ are the maximum eigenvalues of $\mathbf{R}_{xx}$ and $\overline{\mathbf{R}}_{yy}(k)$ for the locally-adaptive network, respectively.

From (28), we see that the convergence of $\widetilde{\lambda}_{w,i}(k)$ to zero at time $k+1$ occurs for the update in (18) if

$$\eta_{i,opt}(k) \ = \ \frac{1}{\sqrt{\lambda_i}\left(1+\sqrt{\lambda_i(k)}\right)}. \tag{30}$$

Near the vicinity of $1/\sqrt{\lambda_i}$, the best step size for convergence is

$$\eta_i(k) \ = \ \frac{1}{2\sqrt{\lambda_i}}, \tag{31}$$

a value that is different for each eigenvalue of $\mathbf{R}_{xx}$. Since $\mathbf{R}_{xx}$ has a large eigenvalue spread for highly-correlated signal mixtures, it is difficult to obtain fast convergence of all the modes of the system in this situation.

Although the local update algorithm in (12) does not possess the equivariant adaptation properties of the global update algorithm in (11), the former algorithm is less sensitive to the choice of $\mathbf{W}(0)$ than is the latter algorithm. Indeed, in the extreme case where $\mathbf{W}(0) = \mathbf{0}$, the constant premultiplying $\widetilde{\lambda}_{w,i}(k)$ in (28) varies from $(1-\eta\sqrt{\lambda_i})$ initially to approximately $(1-2\eta\sqrt{\lambda_i})$ at convergence. This result implies that the initial value of $\mathbf{W}(0)$ is less critical to obtaining good adaptation performance of (12) for fixed step sizes.

3.3 ANALYTICAL COMPARISON OF PERFORMANCE

In this section, we compare the potential convergence speeds of the two decorrelation networks. Using (28), we can find an evolution equation for $\lambda_i(k)$ as defined in (24) for the locally-adaptive network as given by

$$\widetilde{\lambda}_i(k+1) \ = \ \left(\left(1-\eta(k)\sqrt{\lambda_i\lambda_i(k)}\right)^2 - \eta^2\lambda_i\right)\widetilde{\lambda}_i(k). \tag{32}$$

Since both of the updates in (21) and (32) are of the form $\widetilde{\lambda}_i(k+1) = \alpha_i(k)\widetilde{\lambda}_i(k)$, we can study the sensitivity of the convergence speed of each of the algorithms to the eigenvalues of $\mathbf{R}_{xx}$ and $\mathbf{R}_{yy}(0)$ by studying the nature of $\alpha_i(k)$ in each case. In particular, we can calculate the magnitude of $\alpha_i(k)$ as a function of $\lambda_i(k)$ and $\eta(k)$, where values close to zero for a large range of $\lambda_i(k)$ and $\eta(k)$ suggest potentially fast convergence of $\overline{\mathbf{R}}_{yy}(k)$ to the identity matrix for the particular network.

Figure 2a depicts logarithmically-spaced contours of the absolute value of the factor $|\alpha_i(k)|$ as a function of $\lambda_i(k)$ and $\eta_i(k)$ for the globally-adaptive network. Although the value of $|\alpha_i(k)|$ does not depend on the eigenvalues of $\mathbf{R}_{xx}$, the initial distribution of eigenvalues $\lambda_i(0)$ depends on the choice of $\mathbf{W}(0)$ and the eigenvalues of $\mathbf{R}_{xx}$ through (24). The most desirable $\mathbf{W}(0)$ would cluster all of the $\lambda_i(k)$ in one region of this contour plot, so that all of these eigenvalues would converge in a similar fashion with the same step size sequence $\eta(k)$. Figure 2b plots contours of

the factor $|\alpha_i(k)|$ for the locally-adaptive network for six values of λ_i in the range $0.001 \leq \lambda_i \leq 100$. From the differences in these plots, it is seen that the behavior of each $\lambda_i(k)$ depends strongly on both the value of $\lambda_i(k)$ and the eigenvalues of $\mathbf{R}_{xx}$, and thus it is impossible to get similar convergence speeds for different $\lambda_i(k)$ using the locally-adaptive network if the condition number of $\mathbf{R}_{xx}$ is large.

3.4 IMPLEMENTATION ISSUES

The analyses we have presented indicate that the step size sequence $\eta(k)$ plays an important role in the success of the decorrelation networks in (11) and (12). Moreover, because of the nonlinear forms of these coefficient updates, the convergence behaviors of these algorithms deviate from the exponential behaviors of algorithms that are based on quadratic error criteria such as the least-mean-square and recursive least-squares FIR adaptive filters [17]. Thus, it is even more critical in these networks to choose time-varying step size sequences that give fast initial convergence and accurate estimates of the decorrelation matrix $\mathbf{R}_{xx}^{-1/2}$ at convergence. In this section, we describe practical choices of the step size sequences $\eta(k)$ to obtain good adaptation behaviors of the two systems.

It is important to recognize that our previous statistical analyses do not characterize the fluctuations of the coefficients $\mathbf{W}(k)$ for either algorithm in steady-state. Such mean-square analyses of the algorithms' transient behaviors are challenging to perform due to the nonlinear form of the coefficient updates in each case. However, we can make several observations about the relationship between the choice of $\eta(k)$ and the qualitative behavior obtained by each system:

- By choosing $\eta(k)$ to be somewhat smaller than the upper bounds computed for the specific algorithm, the coefficients of each adaptive network converge quickly and track statistical changes in the input signals $\mathbf{x}(k)$. However, because of the nature of the averaging analysis, the actual stability bounds predicted by (22) and (29) are not exact predictors of the stability range of $\eta(k)$ required for mean-square convergence of the algorithms in each case. Thus, one must select values of $\eta(k)$ that are less than a fraction of the upper bounds for stability as predicted by our analyses.
- By choosing $\eta(k)$ to be a small value relative to the upper stability bound computed for each system, the fluctuations in the elements of $\mathbf{W}(k)$ are reduced for data with fixed correlation statistics $\mathbf{R}_{xx}(k) = \mathbf{R}_{xx}$. However, the speeds of convergence of the algorithms are subsequently slower as well in this case.
- Choosing a value of $\eta(k)$ near its optimum value for fastest convergence of the ith mode as in (25) or (30) for each algorithm is likely to lead to divergence of another mode of the system. Moreover, the step size value needed for accurate estimation of $\mathbf{R}_{xx}^{-1/2}$ in steady-state for fixed correlation statistics is typically a fraction of these optimum values.

Given these facts, we suggest the following methods for computing the step size sequences for the two algorithms.

Globally-Adaptive Algorithm:

$$\eta(k) = \begin{cases} \dfrac{0.5\eta_0(k)}{\delta + \sqrt{\beta_y(k)}} & \text{if } 0 < \beta_y(k) \leq 3 + 2\sqrt{2} \\ \dfrac{\eta_0(k)}{\delta - 1 + \beta_y(k)} & \text{if } \beta_y(k) > 3 + 2\sqrt{2}, \end{cases} \tag{33}$$

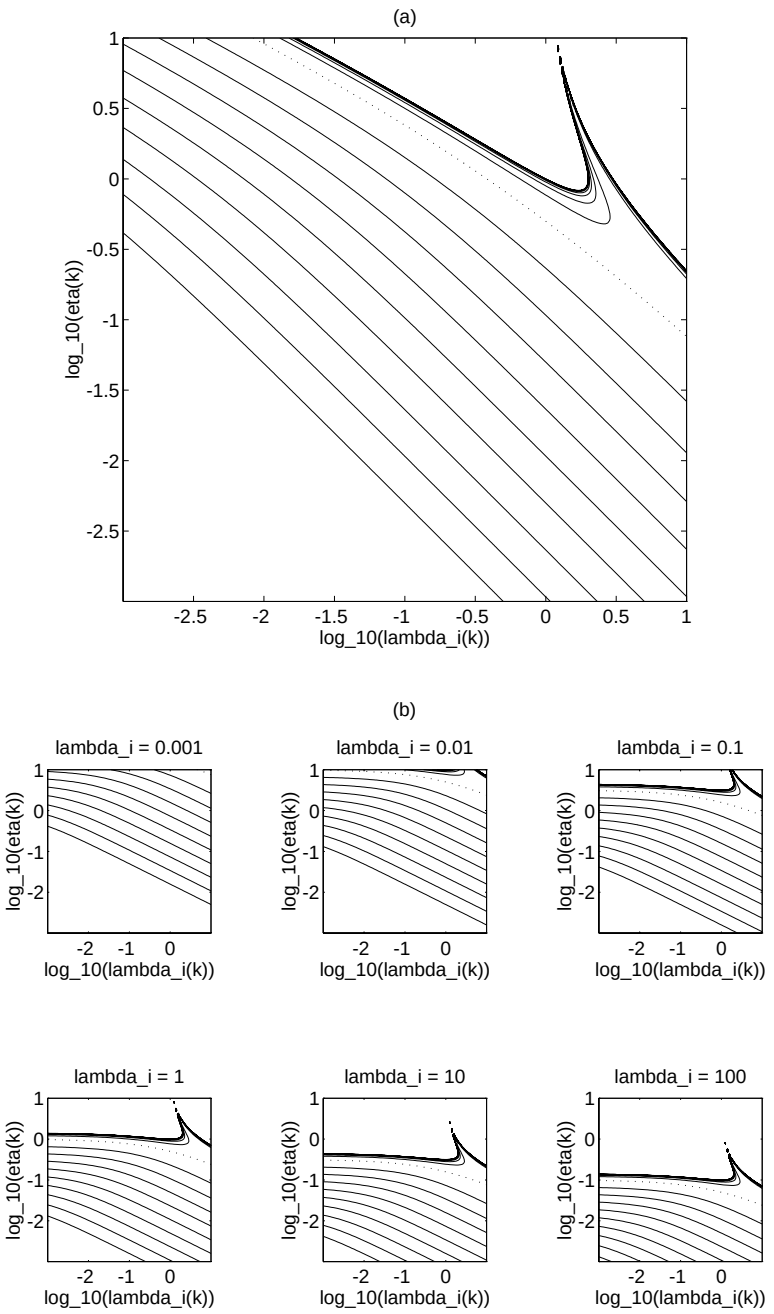

Figure 2: Contour plots of $|\alpha_i(k)|$ as a function of $\lambda_i(k)$ and $\eta(k)$, a) globally-adaptive network and b) locally-adaptive network for different λ_i. The dashed lines indicate the locus of points for which $\alpha_i(k) = 0$.

Locally-Adaptive Algorithm:

$$\eta(k) = \frac{\eta_0(k)}{\delta + \sqrt{\beta_x(k)}\left(1 + \sqrt{\beta_y(k)}\right)}, \tag{34}$$

where $0 < \eta_0(k) < 0.1$, δ is a small positive constant, and $\beta_x(k)$ and $\beta_y(k)$ are computed recursively as

$$\beta_x(k) = \lambda_x \beta_x(k-1) + (1-\lambda_x)||\mathbf{x}(k)||^2 \tag{35}$$

$$\beta_y(k) = \lambda_y \beta_y(k-1) + (1-\lambda_y)||\mathbf{y}(k)||^2, \tag{36}$$

where λ_x and λ_y satisfy $0 \ll \{\lambda_x, \lambda_y\} < 1$.

Both (33) and (34) comfortably satisfy the sufficient bounds for convergence of the respective averaged systems as given in (23) and (29), respectively, for $0 < \eta_0(k) < 0.1$. These bounds use the fact that the maximum eigenvalues of $\mathbf{R}_{xx}(k)$ and $\mathbf{R}_{yy}(k)$ are upper bounded by $E\{||\mathbf{x}(k)||^2\}$ and $E\{||\mathbf{y}(k)||^2\}$, respectively. Time averages are then used to compute estimates of these quantities iteratively over time. Both $\beta_x(0)$ and $\beta_y(0)$ should be chosen large enough to prevent initial divergence of either algorithm but not so large as to prevent fast initial estimation of $E\{||\mathbf{x}(k)||^2\}$ and $E\{||\mathbf{y}(k)||^2\}$. If the statistics of the input signal are changing slowly or are constant, the value of λ_x can be chosen close to one to give good estimates of $E\{||\mathbf{x}(k)||^2\}$. However, since $\mathbf{y}(k)$ is highly nonstationary while $\mathbf{W}(k)$ is converging, the value of λ_y should be chosen somewhat less than one to allow forgetting of the past coefficient values contained in the past output vectors $\mathbf{y}(k)$.

4 ANALYSIS OF MULTILAYER DECORRELATION NETWORKS

In this section, we consider a multilayer network of N cascaded, locally-adaptive decorrelation systems of the form in (12) as originally proposed in [8] in the more-general blind signal separation case. In this structure, the mth output of the system is given by

$$\mathbf{y}^{(m)}(k) = \mathbf{W}^{(m)}(k)\mathbf{y}^{(m-1)}(k), \tag{37}$$

where $\mathbf{y}^{(0)}(k) = \mathbf{x}(k)$, and $\mathbf{W}^{(m)}(k)$, $1 \le m \le N$ is updated as

$$\mathbf{W}^{(m)}(k+1) = \mathbf{W}^{(m)}(k) + \eta(k)\left(\mathbf{I} - \mathbf{y}^{(m-1)}(k)\mathbf{y}^{(m-1)T}(k)\right). \tag{38}$$

Simulations in [8] indicate that a similar multilayered network for blind signal separation provides better separation capabilities as compared to that provided by the single-layer network in (4) with $\mathbf{G}(k)$ given by (7).

Using (32), it is straightforward to show that the approximate convergence behavior of the absolute error in the ith eigenvalue of $\overline{\mathbf{R}}_{yy}^{(m)}(k) = E\{\mathbf{W}^{(m)}(k)\}\overline{\mathbf{R}}_{yy}^{(m-1)}(k)E\{\mathbf{W}^{(m)T}(k)\}$ is determined by

$$\widetilde{\lambda}_i^{(m)}(k+1) = \left(\left(1 - \eta^{(m)}(k)\sqrt{\lambda_i^{(m-1)}(k)\lambda_i^{(m)}(k)}\right)^2 - \eta^{(m)2}(k)\lambda_i^{(m-1)}(k)\right)\widetilde{\lambda}_i^{(m)}(k), \tag{39}$$

where $\lambda_i^{(m-1)}(k)$ is the ith eigenvalue of $\overline{\mathbf{R}}_{yy}^{(m-1)}(k)$. Comparing this equation with (21), we see that if $\lambda_i^{(m-1)}(k) \approx \lambda_i^{(m)}(k)$, then the evolution of $\lambda_i^{(m)}(k)$ is approximately the same as that of $\lambda_i(k)$ for the algorithm in (11).

Of course, $\lambda_i^{(m-1)}(k) \neq \lambda_i^{(m)}(k)$ in practice, and thus the cascaded network of locally-adaptive decorrelators does not behave like the system in (11). However, from (39), one can easily see the benefits of the cascaded structure. The decorrelation provided by the network at the $(m-1)$th stage reduces the eigenvalue spread of $\overline{\mathbf{R}}_{yy}^{(m-1)}(k)$, thus making it easier to select a step size $\eta^{(m)}(k)$ that greatly reduces all values of $|\widetilde{\lambda}_i^{(m)}(k)|$ for $1 \leq i \leq n$ at the mth stage of the system. In effect, the improvements in the convergence behaviors of previous decorrelation stages are compounded in subsequent decorrelation stages, and the increased adaptation speed is noticeable even for systems with $N = 2$ stages.

Although the multilayer structure brings potential benefits in convergence speed, there are two issues that can limit the performance of this system in practice. Firstly, since the cascaded structure has $n^2(N-1)$ redundant parameters, fluctuations in these parameters for non-zero adaptation speeds increase the observed level of error at the system outputs, an effect that is not characterized by our analysis of the mean behavior of the system. This effect can be mitigated by choosing step sizes $\eta^{(m)}(k)$ that are somewhat smaller than that used for the single-layer structure, where the convergence benefits obtained by the nonlinear form of the coefficient updates can be realized even for these smaller adaptation parameters. Secondly, the cascaded structure is an interconnection of several systems, each of which has its own memory and coefficient updates; therefore, it is a nonlinear dynamical system. By simulating (39) for this structure with non-zero adaptation speeds for all layers of the system, it is observed that the outputs of this system do not monotonically converge to decorrelated signals. This behavior is observed even for step sizes that satisfy bounds of the form in (29) for each layer, because the objective functions of the second and subsequent layers change according to the adaptation performances of the first and preceding layers, respectively. In other words, the system can experience overshoot, ringing, and other effects common in nonlinear dynamical systems. One can overcome this difficulty by using a multitiered adaptation strategy whereby only one coefficient layer is allowed to adapt at any one time. As an example, for an interconnected system of two layers, the first coefficient layer $\mathbf{W}^{(1)}(k)$ is allowed to partially converge for a fixed number of iterations, at which time $\eta^{(1)}(k)$ is set to zero and the second coefficient layer $\mathbf{W}^{(2)}(k)$ is allowed to adapt. In this way, the partial decorrelation provided by previous stages improves the convergence speed of the mth stage without affecting the monotonic convergence of the coefficients of the mth stage. Such a methodology does require setting adaptation switching times, and if the structure of $\mathbf{R}_{xx}$ changes instantaneously, additional logic would be required to monitor the quality of the decorrelated output and alternate the adaptation of each layer in turn.

5 AN *A POSTERIORI* LOCALLY-ADAPTIVE NETWORK

The analyses of the locally-adaptive algorithm in (12) in previous sections provide some guidance as to the choice of the step size sequence $\eta(k)$ that provides stable adaptation behavior. However, because these analyses assume that the input vector sequence $\mathbf{x}(k)$ is independent from time instant to time instant, they do not provide true stability bounds for $\eta(k)$ in terms of mean-square performance and in the more-realistic case of time-correlated input signal vectors. In this section, we derive an

algorithm that is based on an *a posteriori* error criterion. In adaptive filtering, the *a posteriori* version of the LMS algorithm yields a form of the normalized LMS algorithm that is guaranteed not to diverge for any positive step size value [16]. Simulations described in the next section indicate that the *a posteriori* version of the locally-adaptive network for blind decorrelation derived here also appears not to diverge for any step size value, and it provides good decorrelation performance as the step size is reduced.

Our proposed algorithm is a modification of the locally-adaptive algorithm in (12) and can be compactly stated as follows:

$$\mathbf{W}(k+1) \quad = \quad \mathbf{W}(k) + \eta_0(k) \left(\mathbf{I} - \check{\mathbf{y}}(k) \check{\mathbf{y}}^T(k) \right), \tag{40}$$

where $\check{\mathbf{y}}(k)$ is the *a posteriori* output vector given by

$$\check{\mathbf{y}}(k) \quad = \quad \mathbf{W}(k+1)\mathbf{x}(k). \tag{41}$$

Since $\check{\mathbf{y}}(k)$ depends on $\mathbf{W}(k+1)$, the equation in (40) does not represent a coefficient update. However, via suitable manipulations, we can develop a relation that computes $\mathbf{W}(k+1)$ from its past value and signals that are available at time k. The derivation is given in Appendix C, and the resulting update for $\mathbf{W}(k)$ is

$$\mathbf{W}(k+1) = \mathbf{W}(k) + \eta_0(k) \left(\mathbf{I} - \frac{\mathbf{z}(k)\mathbf{z}^T(k)}{\eta_0(k)\mathbf{x}^T(k)\mathbf{z}(k) + \left(1 + \sqrt{1 + 4\eta_0(k)\mathbf{x}^T(k)\mathbf{z}(k)}\right)/2} \right), \tag{42}$$

where

$$\mathbf{z}(k) \quad = \quad \mathbf{y}(k) + \eta_0(k)\mathbf{x}(k). \tag{43}$$

Note that as $\eta_0(k) \rightarrow 0$, the update in (42) approaches that for the coefficients of the locally-adaptive network in (12).

We can analyze the behavior of the *a posteriori* update in (40) using similar assumptions as in our previous analyses. We study the behavior of the averaged system given by

$$E\{\mathbf{W}(k+1)\} + \eta_0(k) E\{\mathbf{W}(k+1)\} \mathbf{R}_{xx} E\{\mathbf{W}^T(k+1)\} = E\{\mathbf{W}(k)\} + \eta_0(k)\mathbf{I}. \tag{44}$$

Using similar decomposition techniques as in Appendix B, we obtain the following update relationship for the ith eigenvalue of $E\{\mathbf{W}(k)\}$:

$$\lambda_{w,i}(k+1) + \eta_0(k)\lambda_i\lambda_{w,i}^2(k+1) \quad = \quad \lambda_{w,i}(k) + \eta_0(k). \tag{45}$$

Solving for the positive root of $\lambda_{w,i}(k+1)$ gives

$$\lambda_{w,i}(k+1) \quad = \quad \frac{1}{2\eta_0(k)\lambda_i} \left(-1 + \sqrt{1 + 4\eta_0(k)\lambda_i(\lambda_{w,i}(k) + \eta_0(k))} \right). \tag{46}$$

For non-negative choices of $\lambda_{w,i}(0)$, it can be shown that (46) has a stationary point at $\lambda_{w,i}(k) = \lambda_i^{-1/2}$ for constant step size sequences $\eta_0(k) = \eta_0 > 0$. Since the actual coefficient updates in (42) use the input signal measurements and not the input signal autocorrelation matrix, a small value of $\eta_0(k)$ is desirable to obtain an accurate estimate of $\mathbf{W}(k)$ as effectively averaged across successive time instants within the network.

6 EXTENSION TO TIME-SERIES DECORRELATION

In this section, we consider the task of decorrelating a multichannel time series signal $\mathbf{x}(k)$ with an unknown autocorrelation function $\mathbf{R}_{xx}(m,n) = E\{\mathbf{x}(m)\mathbf{x}^T(n)\}$ using a multichannel finite-impulse-response (FIR) filter network such that

$$\mathbf{y}(k) \quad = \quad \sum_{p=0}^{L} \mathbf{W}_p(k)\mathbf{x}(k-p), \tag{47}$$

where $\mathbf{W}_p(k)$, $0 \leq p \leq L$ are $(n \times n)$-dimensional matrices containing the $n^2(L+1)$ coefficients of the network.

For this task, we propose the following generalized locally-adaptive algorithm for adjusting the linear neural network's synaptic weights:

$$\begin{aligned} \mathbf{W}_p(k+1) \quad = \quad & \mathbf{W}_p(k) + \frac{\eta(k)}{2}\left(2\mathbf{I}\delta\left(p - \frac{L}{2}\right)\right. \\ & \left. -\mathbf{y}\left(k - \frac{L}{2}\right)\mathbf{y}^T(k-p) - \mathbf{y}(k-L+p)\mathbf{y}\left(k - \frac{L}{2}\right)\right). \end{aligned} \tag{48}$$

This algorithm employs delayed output signals in its update to avoid recalculation of the output signals at time $k-p$, $1 \leq p \leq L$ using the most-recent coefficient values. Using delayed coefficient values also allows the algorithm to be computed causally from signals available at time k. The algorithm maintains spatial and temporal symmetry of the filter coefficients by insuring that

$$\mathbf{W}_p(k) \quad = \quad \mathbf{W}_{L-p}^T(k). \tag{49}$$

such that the multidimensional discrete-time transfer function $W_k(\omega)$, $-\pi \leq \omega < \pi$ of the network has the form

$$W_k(\omega) \quad = \quad \sum_{p=0}^{L} \mathbf{W}_p(k)e^{-j\omega p} \quad = \quad A_k(\omega)e^{-j\omega L/2}, \tag{50}$$

where $A_k(\omega)$ is an n-dimensional matrix with real eigenvalues. Simulations indicate that, so long as $A_k(\omega)$ has positive eigenvalues at each iteration, then a positive fixed step size value $\eta(k)$ can be chosen to maintain the stability of the algorithm. Such a condition is equivalent to insuring that $\mathbf{M}(k)$ in (14) is positive-definite for the locally-adaptive network for separation of instantaneous signal mixtures.

While an analysis of the convergence behavior and stability properties of the algorithm in (48) are beyond the scope of this paper, simulations of the algorithm indicate that it performs as desired, providing joint temporal and spatial decorrelation of multichannel signals. In addition, we can make two comments concerning its performance:

- Because of the constraint in (49), this multichannel decorrelation network has a constant group delay. Thus, the temporal shapes of the input signal waveforms are approximately maintained at the network's outputs, a useful feature for some applications. Note that other methods for multichannel decorrelation such as those based on linear prediction do not produce a linear-phase decorrelation system.
- Because the coefficient updates at time k are computed using delayed outputs $\mathbf{y}(k-i)$, $0 \leq i \leq L$, the algorithm step size required for stability generally decreases as the length L of this multichannel FIR filter is increased. Similar issues govern the selection of adaptation parameters for

the delayed LMS algorithm used in hardware implementations of adaptive FIR filters [23] and the filtered-X LMS algorithm for feedforward active noise control [21].

7 SIMULATIONS

7.1 PERFORMANCE COMPARISON OF MULTICHANNEL DECORRELATION NETWORKS

In this section, we compare the performances of the globally-adaptive algorithm in (11), the locally-adaptive algorithm in (12), a two-layer network structure of the form in (37)–(38), and the *a posteriori* locally-adaptive algorithm in (42) using both theory and simulation. In each case, we use the theoretical results derived for each network to predict its simulated performance on jointly Gaussian signals generated as in (1), where $\mathbf{s}(k) = [s_1(k)\ s_2(k)\ s_3(k)]^T$ is a zero-mean jointly Gaussian vector sequence with $E\{\mathbf{s}(k)\mathbf{s}(k-i)\} = \mathbf{I}\delta(i)$. For each system, we have computed step sizes according to our recommended methods in Section 3.4 for each algorithm, where $\lambda_x = \lambda_y = 0.99$, $\delta = 0.01$, and $\eta_0(k)$ is as specified in each case. To gauge the performance of each network, we define the performance factor $\rho(k)$ as

$$\rho(k) \quad = \quad ||\mathbf{I} - \mathbf{W}(k)\mathbf{R}_{xx}\mathbf{W}^T(k)||_F^2 \tag{51}$$

where $||\cdot||_F$ denotes the matrix Frobenius norm and $\mathbf{R}_{xx} = \mathbf{A}\mathbf{A}^T$ for our signal model. For each network, we have selected $\mathbf{W}(0) = 0.1\mathbf{I}$. One hundred simulations have been run and ensemble averages have been calculated to estimate the value of $E\{\rho(k)\}$ and $E\{\mathbf{W}(k)\}$ in each case.

For our first set of simulations, we have chosen

$$\mathbf{A} \quad = \quad \begin{bmatrix} 0.1 & 0.5 & 0.3 \\ 0.4 & 0.5 & 0.1 \\ 0.1 & 0.9 & 0.0 \end{bmatrix} \tag{52}$$

yielding a condition number of $\lambda_{max}/\lambda_{min} = 24.2$ for $\mathbf{R}_{xx}$. Figure 3 plots the average value of $\rho(k)$ obtained from simulation for each decorrelation system, where we have chosen constant step size values $\eta_0(k)$ of 0.017, 0.01, and 0.0046 for the globally-adaptive, locally-adaptive, and *a posteriori* locally-adaptive decorrelation networks, respectively, such that each network gives approximately the same average value of $\rho(k)$ in steady-state. For the two-layer decorrelation network, we have chosen $\mathbf{W}^{(1)}(0) = 0.1\mathbf{I}$, $\mathbf{W}^{(2)}(0) = \mathbf{I}$, and $\eta_0^{(1)}(k) = 0.017$ for the first two hundred iterations of the system, at which time the adaptation of the first layer is halted and the second layer is adapted using $\eta_0^{(2)}(k) = 0.014$. From the plots, we see that the multilayer network provides the fastest adaptation, followed by the globally-adaptive, locally-adaptive, and *a posteriori* locally-adaptive single-layer networks, respectively. Note that the convergence speed of the globally-adaptive network is initially slower than that of the locally-adaptive network, as predicted by our analysis of these systems, although the overall adaptation performance of the former system is better than that of the latter system. In addition, the speed of convergence afforded by the multilayer structure indicates that the partial decorrelation provided by the first layer can enable fast adaptation of the second-layer coefficients of the system, verifying the usefulness of this structure. While the *a posteriori* locally-adaptive system has the worst adaptation performance of all of the systems in this case, it also proved to be stable for any sequence $\mathbf{x}(k)$ and any positive step size value $\eta_0(k)$ in these simulations, and the steady-state estimation performance of this system increases as $\eta_0(k)$ is decreased, as expected.

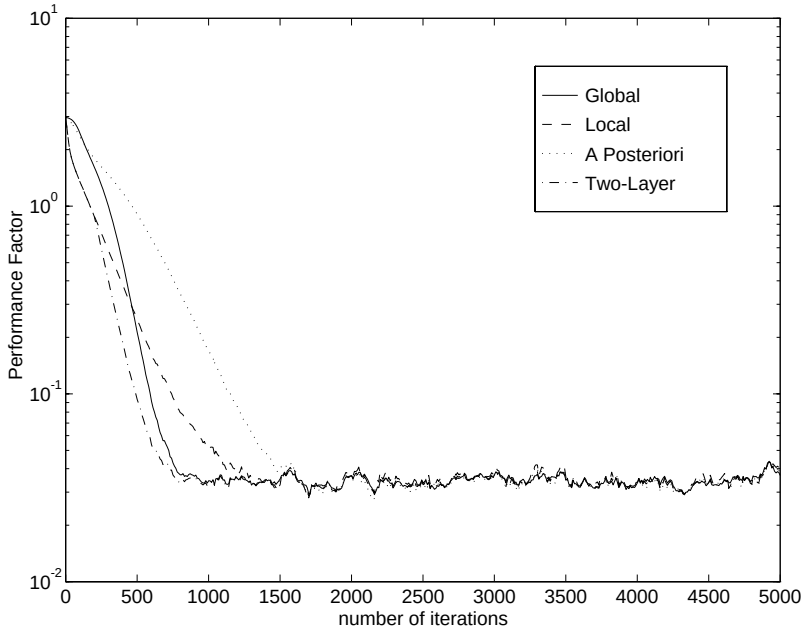

Figure 3: Convergence of average value of performance factor $\rho(k)$: globally-adaptive, locally-adaptive, *a posteriori* locally-adaptive, and two-layer networks, correlated Gaussian input signals, $\lambda_{max}/\lambda_{min} = 24.2$.

To verify the robust adaptation behavior of the *a posteriori* locally-adaptive network, we applied this system to numerous input signals generated using the above input signal model for fixed step sizes $\eta_0(k)$ ranging from 0.01 to 100. Although the average value of $\rho(k)$ in steady-state varied from approximately 0.081 to 1.9×10^9 in these simulations, the system never diverged in any of our tests, verifying the stable adaptation behavior of this network.

Figures 4(a), (b), (c), and (d) depict the mean values over time of the nine coefficients for each of the decorrelation networks. In the case of the multilayer structure, we have plotted the values of $\mathbf{W}(k) = \mathbf{W}^{(2)}(k)\mathbf{W}^{(1)}(k)$ for our given adaptation strategy for this system. Shown on the plots are the optimal steady-state values, given by the elements of $\mathbf{R}_{xx}^{1/2}$, for this decorrelation task. Also shown on these plots are the predicted trajectories of the network coefficients as computed from our mean analyses, showing that our analytical descriptions of the mean behaviors of these systems are accurate.

We now investigate the performances of these networks for the signal model in (1) when $\mathbf{A}$ is chosen as

$$\mathbf{A} = \begin{bmatrix} 0.1 & 0.2 & 0.3 \\ 0.2 & 0.1 & 0.2 \\ 0.5 & 0.1 & 0.1, \end{bmatrix} \tag{53}$$

which yields a condition number of $\lambda_{max}/\lambda_{min} = 422.4$ for $\mathbf{R}_{xx}$. Figure 5 shows the averaged performance factors for each system, where we have selected step sizes of 0.034, 0.02, and 0.016 for the globally-adaptive, locally-adaptive, and *a posteriori* locally-adaptive networks, respectively. For the two-layer network, we have selected step sizes of $\eta_0^{(1)}(k) = 0.034$ and $\eta_0^{(2)}(k) = 0.03$ over their respective periods of adap-

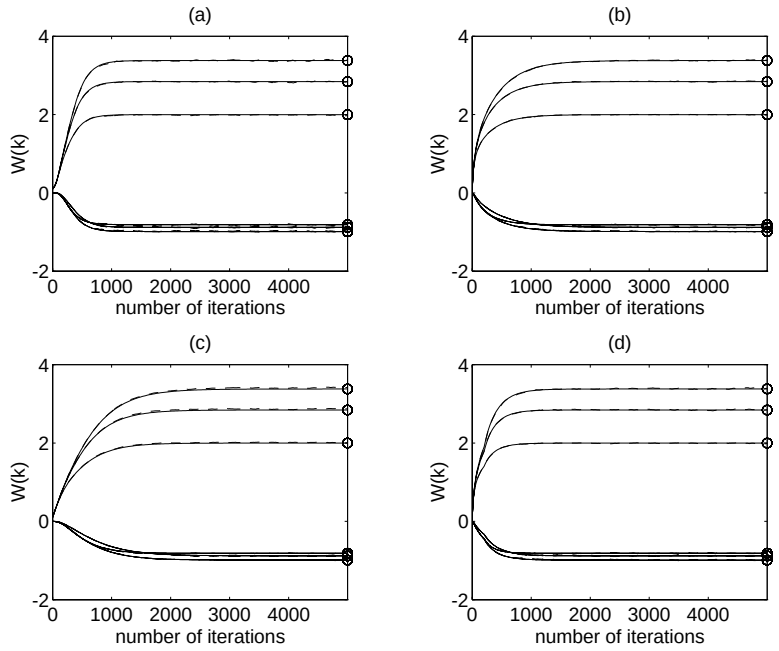

Figure 4: Evolution of averaged coefficient matrices $\mathbf{W}(k)$, theory (solid lines), simulation (dashed lines), and optimum values (circles): (a) globally-adaptive, (b) locally-adaptive, (c) *a posteriori* locally-adaptive, and (d) two-layer networks, correlated Gaussian input signals, $\lambda_{max}/\lambda_{min} = 24.2$.

tation, and we have switched adaptation from the first to the second layer at time $k = 300$. From these results, we see that the globally-adaptive network performs the best, followed by the multilayer network, the locally-adaptive, and the *a posteriori* locally-adaptive networks, respectively. The similar convergence behaviors of the globally-adaptive network in both simulation cases is due to the equivariant property of this system, implying that this network performs well in a variety of situations. Although the other networks based on local adaptation strategies do not perform as well, they do converge to the proper coefficient solutions, and in the case of the *a posteriori*-based algorithm, its stability is robust to step size choice as well.

7.2 SPATIAL AND TEMPORAL DECORRELATION OF SIGNALS

We now explore the performance of the proposed system in (48) for combined spatial and temporal decorrelation of multichannel signals. For our first example, we generate a spatially- and temporally-correlated signal using the first-order linear system given by

$$\mathbf{x}(k) \quad = \quad \mathbf{A}\mathbf{x}(k-1) + \mathbf{B}\mathbf{s}(k), \tag{54}$$

where $\mathbf{s}(k) = [s_1(k)\ s_2(k)\ s_3(k)]^T$ is sequence of zero-mean jointly Gaussian random vectors with $E\{\mathbf{s}(k)\mathbf{s}(k-i)\} = \mathbf{I}\delta(i)$ and $\mathbf{A}$ and $\mathbf{B}$ are given by

$$\mathbf{A} \ = \ \begin{bmatrix} 0.4 & 0.1 & 0.3 \\ 0.4 & 0.2 & 0.5 \\ 0.2 & 0.1 & 0.5 \end{bmatrix} \quad \text{and} \quad \mathbf{B} \ = \ \begin{bmatrix} 0.9 & 0.1 & 0.7 \\ 0.8 & 0.7 & 0.9 \\ 0.5 & 0.4 & 0.8 \end{bmatrix}, \tag{55}$$

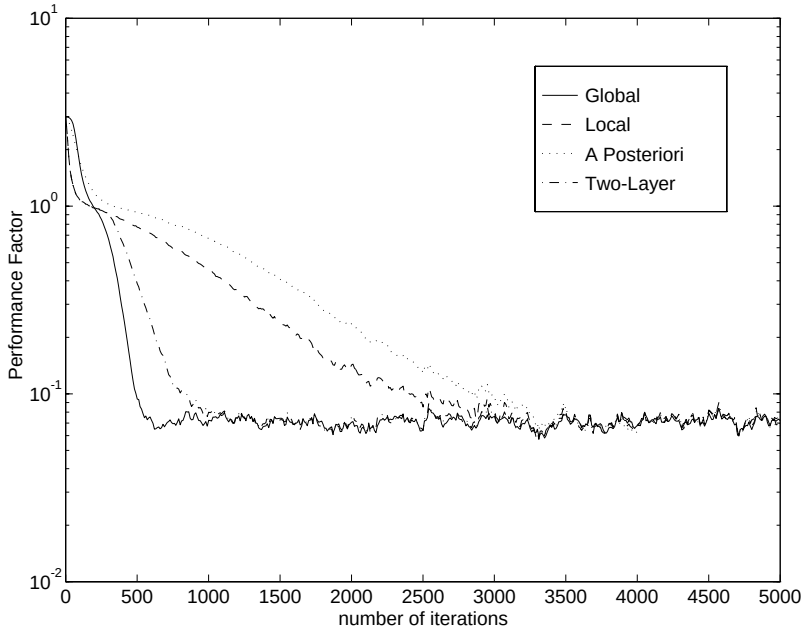

Figure 5: Convergence of average value of performance factor $\rho(k)$: globally-adaptive, locally-adaptive, *a posteriori* locally-adaptive, and two-layer networks, correlated Gaussian input signals, $\lambda_{max}/\lambda_{min} = 422.4$.

respectively. It can be shown that $\lim_{k\to\infty} E\{\mathbf{x}(k)\mathbf{x}(k)\} = \mathbf{R}_{xx}$ is the solution of the discrete-time algebraic Riccati equation given by [19]

$$\mathbf{R}_{xx} \quad = \quad \mathbf{A}\mathbf{R}_{xx}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T, \tag{56}$$

and from this solution we can compute the limiting value of $\mathbf{R}_{xx}(k, k-i) = E\{\mathbf{x}(k)\mathbf{x}(k-i)\}$ for $i > 0$ as

$$\lim_{k\to\infty} \mathbf{R}_{xx}(k, k-i) \quad = \quad \mathbf{A}^i\mathbf{R}_{xx}. \tag{57}$$

To quantify the performance of the algorithm, we compute the performance factor $\rho(k)$ as

$$\rho_T(k) \quad = \quad \sum_{p=-L}^{L} \left|\left|\mathbf{I}\delta(p) - \mathcal{W}(k)\mathcal{R}(p)\mathcal{W}^T(k)\right|\right|_F^2 \tag{58}$$

where $\mathcal{W}(k) = [\mathbf{W}_0(k)\ \mathbf{W}_1(k)\ \cdots\ \mathbf{W}_L(k)]$ and $\mathcal{R}(p)$ is an $(n(L+1) \times n(L+1))$ block Toeplitz matrix whose (i,j)th block entry is the limiting value of $\mathbf{R}_{xx}(k, k+i-j-p)$ in (57). In this case, we have generated twenty data sets from this model and averaged the ensembles of the estimates of $\rho_T(k)$ from the multichannel decorrelation network to estimate the value of $E\{\rho_T(k)\}$ at each time instant.

Figure 6 shows the averaged value of $\rho_T(k)$ as a function of time for a three-channel decorrelator with a length of $L = 7$, where we have chosen $\eta(k) = 0.0005$ and $\mathbf{W}_p(0) = \mathbf{0}$ for $0 \le p \le L$. As can be seen, the average value of $\rho_T(k)$ decreases from a value of approximately three to an average value of 0.0537 in steady-state, indicating that the network in (48) is performing combined spatial and temporal decorrelation of the multichannel signals in this case.

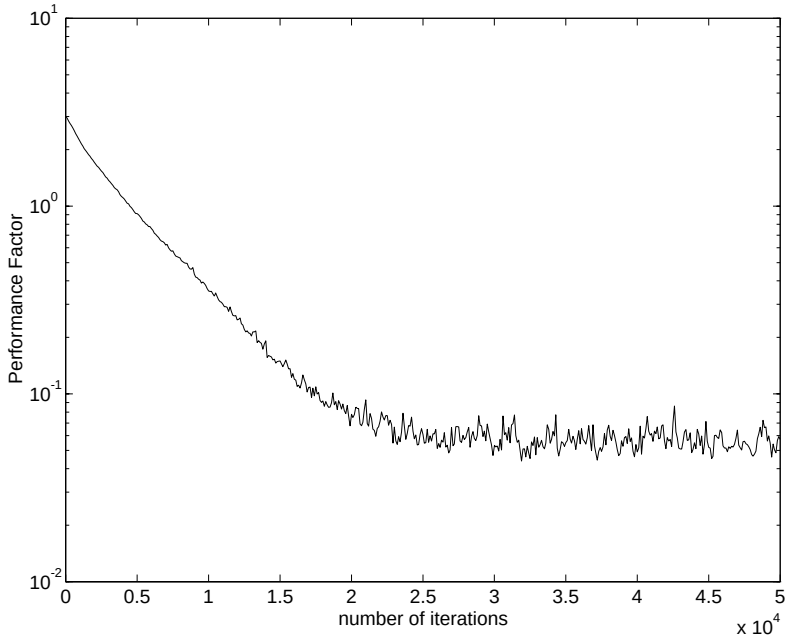

Figure 6: Convergence of average value of performance factor $\rho_T(k)$: multichannel spatial and temporal decorrelation network, correlated Gaussian input signals, $L = 7$, $n = 3$.

As further evidence of the capabilities of the proposed decorrelation network, Figure 7 shows the power spectra of a human speech signal sampled at $f_s = 8192$Hz both before and after processing by a single-channel, $L = 30$-length FIR decorrelation system adapted using (48). These spectra have been estimated over the latter two seconds of a four-second continuous speech segment. In this case, the mean input signal power is 4.6×10^{-3}, and we have chosen $\eta(k) = 0.005$ and $w_p(0) = 0$, $0 \leq p \leq L$ for the decorrelation system. The power spectrum of the output signal is approximately flat over the usable frequency range of $100 \leq f \leq 3000$ for this signal, indicating that the decorrelation system works as intended.

8 CONCLUSIONS

In this paper, we have analyzed and compared the performance of several networks for the multichannel decorrelation of signals. Using analyses of the averaged versions of these systems, we provide useful techniques for choosing the step sizes for these algorithms in an on-line system to provide fast, accurate, and robust adaptation behavior in each case. An analysis of a multilayer decorrelation network employing local adaptation rules indicates that the decorrelation provided by successive stages of this system improves the convergence performance and decorrelation properties of the overall network. Finally, we have derived two new adaptive networks based on the local adaptation method of [8]: a robust system for multichannel decorrelation using an *a posteriori* error criterion, and a system for spatial and temporal decorrelation of multichannel time series. Simulations show that all of these systems work as designed, yielding decorrelated outputs in space and, in the case of the latter algorithm, in space and time. Because of their mathematical and architectural simplicity, these decorrelation methods are expected to have wide use in a number

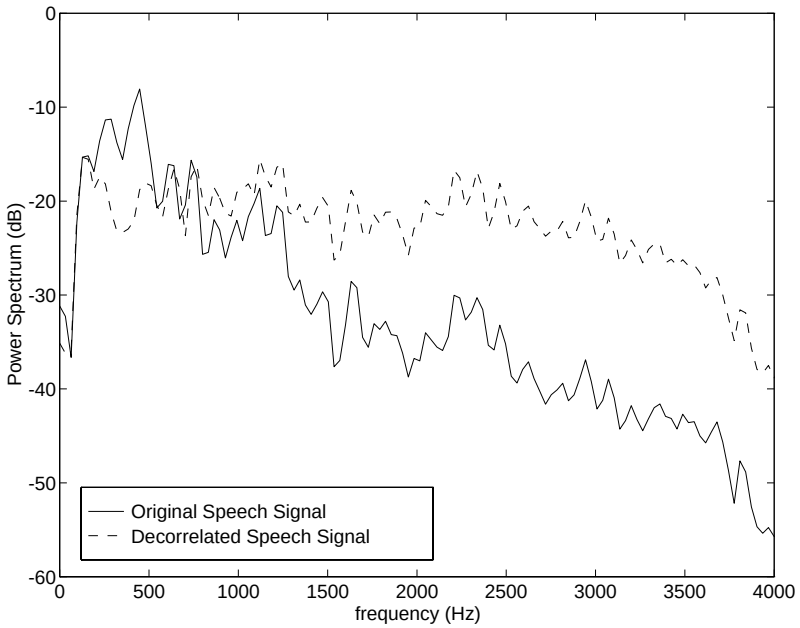

Figure 7: Normalized power spectra of original and decorrelated speech signals, single-channel decorrelation system, $L = 30$.

of communications, control, and signal processing applications.

Appendix A

In this section, we analyze the convergence behavior of the averaged version of the globally-adaptive decorrelation network as given in (17). For our analysis, we post-multiply both sides of (17) by the respective transposes of both sides of (17), pre-multiplied by $\mathbf{R}_{xx}$. From these operations, we determine an update for $\overline{\mathbf{R}}_{yy}(k)$, defined in (19), that is given by

$$\overline{\mathbf{R}}_{yy}(k+1) \;=\; (1+\eta(k))^2\overline{\mathbf{R}}_{yy}(k) - 2\eta(k)(1+\eta(k))\overline{\mathbf{R}}^2_{yy}(k) + \eta^2(k)\overline{\mathbf{R}}^3_{yy}(k). \quad (59)$$

As is shown in [28], this update can be diagonalized, yielding the expression

$$\mathbf{\Lambda}_{yy}(k+1) \;=\; (1+\eta(k))^2\mathbf{\Lambda}_{yy}(k) - 2\eta(k)(1+\eta(k))\mathbf{\Lambda}^2_{yy}(k) + \eta^2(k)\mathbf{\Lambda}^3_{yy}(k), \quad (60)$$

where the diagonal matrix $\mathbf{\Lambda}_{yy}(k) = \mathbf{Q}_y^T\overline{\mathbf{R}}_{yy}(k)\mathbf{Q}_y = \mathrm{diag}\{\lambda_1(k), \lambda_2(k), \ldots, \lambda_n(k)\}$ contains the eigenvalues of $\overline{\mathbf{R}}_{yy}(k)$. Thus, we study the set of n scalar equations given by

$$\lambda_i(k+1) \;=\; (1+\eta(k))^2\lambda_i(k) - 2\eta(k)(1+\eta(k))\lambda^2_i(k) + \eta^2(k)\lambda^3_i(k). \quad (61)$$

We can determine stability conditions on $\eta(k)$ to guarantee convergence of (61). For this calculation, we subtract one from both sides of (61). After some simplification, we obtain the relationship in (21), where $\widetilde{\lambda}_i(k)$ is as defined in (20). To obtain uniform convergence, we require that $|\widetilde{\lambda}_i(k+1)| < |\widetilde{\lambda}_i(k)|$. For $0 < \lambda_i(k) \leq 2$, the resulting constraints on $\eta(k)$ are

$$0 \;<\; \eta(k) \;<\; \frac{1}{\lambda_i(k)-1}\left(1 - \sqrt{\frac{2}{\lambda_i(k)} - 1}\right). \quad (62)$$

The upper bound in (62) is always positive for $0 < \lambda_i(k) \leq 2$. In the case where $\lambda_i(k) > 2$, we have the conditions

$$0 \;<\; \eta(k) \;<\; \frac{2}{\lambda_i(k) - 1}. \tag{63}$$

Figure 1 plots the upper bound on $\eta(k)$ as a function of $\lambda_i(k)$ as depicted in (62)–(63). As can be seen, the function is not monotonically-decreasing with $\lambda_i(k)$. Since knowledge of each $\lambda_i(k)$ is difficult to obtain in practice, we desire a sufficient bound on $\eta(k)$ that guarantees uniform convergence of all the eigenvalues of $\overline{\mathbf{R}}_{yy}(k)$. To this end, we introduce the function $\eta_{max}(k)$ in (23). This function is also shown in Figure 1 assuming $\lambda_i(k) = \lambda_{max}(k)$. It can be seen that $\eta_{max}(k)$ satisfies the bounds on $\eta(k)$ in (62) and (63) over their respective valid ranges for all values of $\lambda_i(k)$ satisfying $0 < \lambda_i(k) \leq \lambda_{max}(k)$. Thus, we can guarantee uniform convergence of $\lambda_i(k)$ for the averaged version of the globally-adaptive network if we choose $\eta(k)$ to satisfy (22).

In cases where the average behavior of the coefficient matrix $E\{\mathbf{W}(k)\}$ of this network is desired, we can diagonalize (17) directly by pre- and post-multiplying this equation by $\mathbf{Q}_x^T$ and $\mathbf{Q}_x$, respectively, where $\mathbf{R}_{xx} = \mathbf{Q}_x \mathbf{\Lambda}_{xx} \mathbf{Q}_x^T$, and $\mathbf{\Lambda}_{xx}$ is a diagonal matrix of eigenvalues of $\mathbf{R}_{xx}$. Noting that $\mathbf{W}(0) = c\mathbf{I}$ in practice, we find that the ith eigenvalue of $E\{\mathbf{W}(k)\}$ for this system evolves as

$$\lambda_{w,i}(k+1) \;=\; (1 + \eta(k)(1 - \lambda_i(k)))\, \lambda_{w,i}(k), \tag{64}$$

where $\lambda_i(k)$ is as given in (24).

Appendix B

In this section, we analyze the convergence behavior of the averaged version of the locally-adaptive decorrelation network as given in (18). Consider the eigenvalue decomposition of $\mathbf{R}_{xx}$ as $\mathbf{Q}_x \mathbf{\Lambda}_{xx} \mathbf{Q}_x^T$, where λ_i is the ith diagonal entry of $\mathbf{\Lambda}_{xx}$. By pre- and post-multiplying (18) by $\mathbf{Q}_x$ and $\mathbf{Q}_x^T$, we produce the update

$$\mathbf{\Lambda}_w(k+1) \;=\; \mathbf{\Lambda}_w(k) + \eta(k)(\mathbf{I} - \mathbf{\Lambda}_w(k)\mathbf{\Lambda}_{xx}\mathbf{\Lambda}_w(k)), \tag{65}$$

where $\mathbf{\Lambda}_w(k) = \mathbf{Q}_x^T E\{\mathbf{W}(k)\}\mathbf{Q}_x$. Thus, if we set $\mathbf{W}(0) = c\mathbf{I}$, then $\mathbf{\Lambda}_w(k)$ is a diagonal matrix, and the eigenvalues of $E\{\mathbf{W}(k)\}$ approximately evolve according to

$$\lambda_{w,i}(k+1) \;=\; \lambda_{w,i}(k) + \eta(k)(1 - \lambda_i \lambda_{w,i}^2(k)), \tag{66}$$

where λ_i is the ith eigenvalue of $\mathbf{R}_{xx}$. Subtracting $1/\sqrt{\lambda_i}$ from both sides of (66) and simplifying the resulting expression, we obtain the update

$$\widetilde{\lambda}_{w,i}(k+1) \;=\; \left(1 - \eta(k)\left(\sqrt{\lambda_i} + \lambda_i \lambda_{w,i}(k)\right)\right) \widetilde{\lambda}_{w,i}(k), \tag{67}$$

where $\widetilde{\lambda}_{w,i}(k)$ as defined in (27). Noting the relationship of $\lambda_i(k)$ in (24), we can rewrite (67) in the form given by (28).

To guarantee convergence of $\widetilde{\lambda}_{w,i}$ to zero, we require that $|\widetilde{\lambda}_{w,i}(k+1)| < |\widetilde{\lambda}_{w,i}(k)|$ at each iteration. Using (28), this constraint leads to the stability condition for the locally-adaptive system as

$$\left|1 - \eta(k)\sqrt{\lambda_i}\left(1 + \sqrt{\lambda_i(k)}\right)\right| \;<\; 1. \tag{68}$$

Convergence of $\lambda_{w,i}(k)$ is thus guaranteed if

$$0 \;<\; \eta(k) \;<\; \frac{2}{\sqrt{\lambda_i}\left(1 + \sqrt{\lambda_i(k)}\right)}. \tag{69}$$

Since these bounds must hold for all λ_i, we have the sufficient bounds in (29) for stability of the averaged system.

Appendix C

In this section, we derive a causal coefficient update for the *a posteriori* locally-adaptive network given in (40). To begin, we move all terms that depend on $\mathbf{W}(k+1)$ to the left-hand-side of (40), giving

$$\mathbf{W}(k+1)\left(\mathbf{I}+\eta_0(k)\mathbf{x}(k)\check{\mathbf{y}}^T(k)\right) \quad = \quad \mathbf{W}(k)+\eta_0(k)\mathbf{I}. \tag{70}$$

Using the matrix inversion lemma [19], it can be shown that

$$\left(\mathbf{I}+\eta_0(k)\mathbf{x}(k)\check{\mathbf{y}}^T(k)\right)^{-1} \quad = \quad \mathbf{I}-\frac{\eta_0(k)\mathbf{x}(k)\check{\mathbf{y}}^T(k)}{1+\eta_0(k)\mathbf{x}^T(k)\check{\mathbf{y}}(k)}. \tag{71}$$

Post-multiplying each side of (70) by the respective sides of (71) and simplifying, we find after some algebra that

$$\mathbf{W}(k+1) \quad = \quad \mathbf{W}(k)+\eta_0(k)\left(\mathbf{I}-\frac{\mathbf{z}(k)\check{\mathbf{y}}^T(k)}{1+\eta_0(k)\mathbf{x}^T(k)\check{\mathbf{y}}(k)}\right), \tag{72}$$

where $\mathbf{z}(k)$ is defined in (43).

To continue, we move the term on the right-hand-side of (72) that depends on $\mathbf{W}(k+1)$ to the left-hand side of the equation. Assuming that $\mathbf{W}(k+1)=\mathbf{W}^T(k+1)$, we see that

$$\left(\mathbf{I}+\frac{\eta_0(k)\mathbf{z}(k)\mathbf{x}^T(k)}{1+\eta_0(k)\mathbf{x}^T(k)\check{\mathbf{y}}(k)}\right)\mathbf{W}(k+1) \quad = \quad \mathbf{W}(k)+\eta(k)\mathbf{I}. \tag{73}$$

Again, by using the matrix inversion lemma on the matrix premultiplying $\mathbf{W}(k+1)$ in (73), we pre-multiply both sides of this equation by the inverse of this matrix and simplify the result, yielding

$$\mathbf{W}(k+1)=\mathbf{W}(k)+\eta_0(k)\left(\mathbf{I}-\frac{\mathbf{z}(k)\mathbf{z}^T(k)}{1+\eta_0(k)\mathbf{x}^T(k)\mathbf{z}(k)+\eta_0(k)\mathbf{x}^T(k)\check{\mathbf{y}}(k)}\right). \tag{74}$$

Now, by pre- and post-multiplying the relation in (40) by $\mathbf{x}^T(k)$ and $\mathbf{x}(k)$, respectively, we find the equation

$$\eta_0(k)(\mathbf{x}^T(k)\check{\mathbf{y}}(k))^2+\mathbf{x}^T(k)\check{\mathbf{y}}(k)-\mathbf{x}^T(k)\mathbf{z}(k) \quad = \quad 0. \tag{75}$$

This equation is quadratic in $\mathbf{x}^T(k)\check{\mathbf{y}}(k)$, where the solution corresponding to $\mathbf{x}^T(k)\check{\mathbf{y}}(k)>0$ is chosen if $\eta_0(k)>0$ to maintain positive definiteness of $\mathbf{W}(k+1)$. Therefore, we solve for the positive root of $\mathbf{x}^T(k)\check{\mathbf{y}}(k)$, which is

$$\mathbf{x}^T(k)\check{\mathbf{y}}(k) \quad = \quad \frac{1}{2\eta_0(k)}\left(-1+\sqrt{1+4\eta_0(k)\mathbf{x}^T(k)\mathbf{z}(k)}\right). \tag{76}$$

Substituting this relationship into (74) and simplifying the expression produces the update in (42).

Acknowledgements

Portions of this work were supported by the Japanese Frontier Research Program, RIKEN, and by the Federal Bureau of Investigation under Contract No. JFBI95017.

References

[1] S.-I. Amari, A. Cichocki, and H.H. Yang, "A new learning algorithm for blind signal separation," *Adv. Neural Inform. Proc. Sys. 8* (Cambridge, MA: MIT Press, 1996), pp. 757-763.

[2] S.-I. Amari, A. Cichocki, and H.H. Yang, "Recurrent neural networks for blind separation of sources," *Proc. Int. Symp. Nonlinear Theory App.,* Las Vegas, NV, pp. 37-42, December 1995.

[3] A.J. Bell and T.J. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Computation,* vol. 7, no. 6, pp. 1129-1159, November œ1995.

[4] J.F. Cardoso, A. Belouchrani, and B. Laheld, "A new composite criterion for adaptive and iterative blind source separation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Adelaide, Australia, vol. 4, pp. 273-276, April 1994.

[5] J.F. Cardoso and B. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Processing,* vol. 44, no. 12, to appear December 1996.

[6] D.C.B. Chan, P.J.W. Rayner, and S.J. Godsill, "Multi-channel signal separation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Atlanta, GA, vol. 2, pp. 649-652, May 1996.

[7] A. Cichocki and R. Unbehauen, *Neural Networks for Optimization and Signal Processing,* 2nd ed. (New York: Wiley, 1994), pp. 461-471.

[8] A. Cichocki, W. Kasprzak, and S.-I. Amari, "Multi-layer neural networks with a local adaptive learning rule for blind separation of source signals," *Proc. Int. Symp. Nonlinear Theory App.,* Las Vegas, NV, pp. 61-65, December 1995.

[9] A. Cichocki, W. Kasprzak, and S.-I. Amari, "Neural network approach to blind separation and enhancement of images," *Proc. Signal Processing VIII (EUSIPCO)* (Amsterdam: EURASIP/Elsevier Science, 1996), vol. 1, pp. 579-582.

[10] A. Cichocki, R Unbehauen, and E. Rummert, "Robust learning algorithm for blind separation of signals," *Electron. Lett.,* vol. 30, no. 17, pp. 1386-1387, August 1994.

[11] A. Cichocki, R. Unbehauen, L. Moszczyński, and E. Rummert, "A new online adaptive learning algorithm for blind separation of source signals," *Proc. Int. Symp. Artificial Neural Networks,* Tainan, Taiwan, pp. 406-411, December 1994.

[12] A. Cichocki, R. Swiniarski, and R.E. Bogner, "Hierarchical neural network for robust PCA of complex-valued signals," *Proc. World Congress Neural Networks,* San Diego, CA, pp. 818-821, September 1996.

[13] P. Comon, C. Jutten, and J. Herault, "Blind separation of sources, part II: Problem statement," *Signal Processing,* vol. 24, no. 1, pp. 11-20, July 1991.

[14] P. Comon, "Independent component analysis: A new concept?" *Signal Processing,* vol. 36, no. 3, pp. 287-314, April 1994.

[15] K.I. Diamantaras and S.Y. Kung, *Principal Component Neural Networks: Theory and Applications* (New York: Wiley, 1996).

[16] S.C. Douglas and T.H.-Y. Meng, "Normalized data nonlinearities for LMS adaptation," *IEEE Trans. Signal Processing,* vol. 42, no. 6, pp. 1352-1365, June 1994.

[17] S. Haykin, *Adaptive Filter Theory,* 3rd ed. (Upper Saddle River, NJ: Prentice-Hall, 1996).

[18] S. Haykin, *Neural Networks: A Comprehensive Foundation* (New York: Macmillan, 1994).

[19] T. Kailath, *Linear Systems* (Englewood Cliffs, NJ: Prentice-Hall, 1981).

[20] J. Karhunen and J. Joutsensalo, "Representation and separation of signals using nonlinear PCA type learning," *Neural Networks,* vol. 7, no. 1, pp. 113-127, 1994.

[21] S.M. Kuo and D.R. Morgan, *Active Noise Control Systems: Algorithms and DSP Implementations* (New York: Wiley-Interscience, 1996).

[22] J.L. Lacoume and P. Ruiz, "Separation of independent sources from correlated inputs," *IEEE Trans. Signal Processing,* vol. 40, no. 12, pp. 3074-3078, December 1992.

[23] G. Long, F. Ling, and J.G. Proakis, "The LMS algorithm with delayed coefficient adaptation," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 37, no. 9, pp. 1397-1405, September 1989; vol. 40, no. 1, pp. 230-232, January 1992 (corrections).

[24] L. Molgedey, and H. G. Schuster, "Separation of a mixture of independent signals using time delayed correlations," *Phys. Rev. Lett.,* vol. 72, pp. 3634-3637, June 1994.

[25] E. Oja and J. Karhunen, "Signal separation by nonlinear Hebbian learning," in *Computational Intelligence - A Dynamic System Perspective*, ed. M. Palaniswami *et al*, (New York: IEEE Press, 1995), pp. 83-97.

[26] J.E. Potter, "New statistical formulas," Tech. Rept., Instr. Lab., Mass. Inst. Tech., 1963.

[27] F.M. Silva and L.B. Almeida, "A distributed solution for data orthonormalization," *Proc. Int. Conf. Artificial Neural Networks,* Espoo, Finland, vol. 1, pp. 943-948, June 1991.

[28] F.M. Silva and L.B. Almeida, "A distributed decorrelation algorithm," in *Neural Networks: Advances and Applications,* ed. E. Gelenbe, (Amsterdam: Elsevier Science, 1991), pp. 145-163.

[29] H. N. Thi and C. Jutten, "Blind source separation of convolutive mixtures," *Signal Processing,* vol. 45, no. 2, pp. 209-229, August 1995.

[30] D. Van Compernolle and S. Van Gerven, "Signal separation in a symmetric adaptive noise canceler by output decorrelation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* San Francisco, CA, vol. 4, pp. 221-224, March 1992.

[31] E. Weinstein, M. Feder, and A.V. Oppenheim, "Multi-channel signal separation by decorrelation," *IEEE Trans. Speech Audio Processing,* vol. 1, no. 4, pp. 405-413, October 1993.

[32] D. Yellin and E. Weinstein, "Multichannel signal separation: Methods and analysis," *IEEE Trans. Signal Processing,* vol. 44, no. 1, pp. 106-118, January 1996.