On Relative Convergence Properties of Principal Component Analysis Algorithms

Chanchal Chatterjee, *Member, IEEE,* Vwani P. Roychowdhury, and Edwin K. P. Chong, *Senior Member, IEEE*

***Abstract*— We investigate the convergence properties of two different stochastic approximation algorithms for principal component analysis, and analytically explain some commonly observed experimental results. In our analysis, we use the theory of stochastic approximation, and in particular the results of Fabian, to explore the asymptotic mean square errors (AMSE's) of the algorithms. This study reveals the conditions under which the algorithms produce smaller AMSE's, and also the conditions under which one algorithm has a smaller AMSE than the other. Experimental study with multidimensional Gaussian data corroborate our analytical findings. We next explore the convergence rates of the two algorithms. Our experiments and an analytical explanation reveals the conditions under which the algorithms converge faster to the solution, and also the conditions under which one algorithm converges faster than the other. Finally, we observe that although one algorithm has a larger computation in each iteration, it leads to a smaller AMSE and converges faster for the minor eigenvectors when compared to the other algorithm.**

***Index Terms*—Adaptive eigen-decomposition algorithms, PCA algorithms, rates of convergence.**

I. Introduction

We investigate the convergence properties of different algorithms for principal component analysis (PCA). Existing literature on PCA (see [1]) reveals a number of algorithms that are derived from: 1) anti-Hebbian learning [1], [6]; 2) Hebbian learning [1], [10]–[12]; 3) lateral interaction algorithms [1], [6], [13]; and 4) gradient-based learning [1], [10]. Since there are multiple algorithms for PCA, and each one requires a different amount of computation at each update, yet leads to the same set of final results, we are led to ask the question: "which algorithm converges faster than others?" An answer to this question will help us select the appropriate algorithm for a given task.

Among the many algorithms for PCA, we focus on two algorithms that are commonly used in most applications. It can be shown [1], [6], [10] that several other algorithms for PCA are related to these basic procedures. In both these algorithms, we are given a sequence of random matrices $\{A_k \in \Re^{d\times d}\}$, with $\lim_{k\to\infty} E[A_k] = A$, where A_k represents the online observation of an application. Note that one common realization of A_k is from a sequence of random vectors $\{\mathbf{x}_k\}$ as $A_k = \mathbf{x}_k\mathbf{x}_k^T$. For each sample A_k, we compute a matrix $W_k \in \Re^{d\times p}(p \le d)$ by these algorithms, such that W_k tends, with probability one (w.p. 1), to a matrix Φ_p whose columns are the $p(\le d)$ eigenvectors of A ordered by decreasing eigenvalue. Conforming to the existing literature, we refer to Φ_p as the *principal eigenvector matrix* of A. The first column of Φ_p is the first principal eigenvector of A, and so on.

The first algorithm which we consider is due to Oja *et al.* [10], [11], and Sanger [14], and can be derived from Hebbian learning with a multiunit network [1]. This algorithm is

$$W_{k+1} = W_k + \eta_k\big(A_kW_k - W_k\mathrm{UT}_\gamma\big[W_k^TA_kW_k\big]\big) \qquad (1)$$

where $\{\eta_k\}$ is a sequence of scalar gains. Here, $\mathrm{UT}_\gamma[\cdot]$ sets all elements below the diagonal of its matrix argument to zero, and multiplies all elements above the diagonal with $\gamma(\ge 1)$.

The second algorithm is due to Xu *et al.* [16], [17], and can be derived from a least mean square error criterion. The algorithm is

$$W_{k+1} = W_k + \eta_k\big(2A_kW_k - W_k\mathrm{UT}_\gamma\big[W_k^TA_kW_k\big] - A_kW_k\mathrm{UT}_\gamma\big[W_k^TW_k\big]\big). \qquad (2)$$

Note that algorithm (2) uses nearly twice as much computation than algorithm (1) for each update of W_k. Hence, it is natural to ask the question: "what benefits can we derive from algorithm (2) at the cost of higher computation?"

For algorithms (1) and (2), the existing literature and experimental analyzes (also see Section IV) provide the following observations [6], [10]: 1) adding the parameter $\gamma(>1)$ leads to a faster convergence of W_k to Φ_p for the minor eigenvectors; 2) algorithm (2) converges faster than algorithm (1), especially for minor eigenvectors; and 3) increasing $\{\eta_k\}$, up to a maximum limit, leads to a faster convergence of W_k to Φ_p. In this paper, we analytically and experimentally investigate these observations.

The first observation was explained by Oja [10] by exploring the convergence of the last eigenvector (i.e., the pth column of Φ_p). Oja analyzed the stability of the ordinary differential equation (ODE) corresponding to the last eigenvector, assuming that the previous eigenvectors have converged. This analysis shows that the linear part of the ODE converges faster for $\gamma > 1$.

In our analyzes, we use the fact that both algorithms (1) and (2) are stochastic approximation procedures [2], [7], [8], [11]. We utilize the theory of stochastic approximation, in particular the results of Fabian [3], to analyze the *asymptotic mean square errors* (AMSE's) for the algorithms. This analysis

Manuscript received December 17, 1996; revised August 25, 1997 and November 27, 1997.

C. Chatterjee is with GDE Systems Inc., San Diego, CA 92150 USA.

V. P. Roychowdhury is with the Electrical Engineering Department, University of California, Los Angeles, CA 90095 USA.

E. K. P. Chong is with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA.

Publisher Item Identifier S 1045-9227(98)01810-4.

reveals the following facts: 1) by using the parameter $\gamma > 1$, both algorithms lead to a larger AMSE for the minor eigenvectors than for $\gamma = 1$; 2) by increasing $\{\eta_k\}$, up to an upper bound, we increase the AMSE's of all eigenvectors for both algorithms; and 3) algorithm (2) leads to a smaller AMSE than algorithm (1).

We next investigate the convergence rates of algorithms (1) and (2). Our experiments (see Section IV) suggest that by increasing $\gamma(>1)$ and $\{\eta_k\}$, we obtain a faster convergence for both algorithms. Furthermore, algorithm (2) converges faster than algorithm (1). In order to provide an explanation for the faster convergence for algorithm (2), we analyze the ODE's for the two algorithms in the neighborhood of the solution.

In summary, the above study yields the following results: 1) both algorithms produce larger AMSE's for increasing $\{\eta_k\}$ and $\gamma(>1)$; 2) algorithm (2) leads to a smaller AMSE of the estimates than algorithm (1); 3) both algorithms converge faster to the solution for increasing $\{\eta_k\}$ and $\gamma(>1)$; and 4) there is a tradeoff between AMSE and convergence rate for both algorithms, i.e., larger values of $\{\eta_k\}$ and γ lead to faster convergence but higher AMSE and vice versa. From 2) and 4), a clear tradeoff between computation and convergence is made explicit. By using algorithm (2) with $\gamma > 1$, we perform more computation, but produce a smaller AMSE of the estimates, and also converge faster when compared to algorithm (1).

In Section II we provide a mathematical background of the PCA algorithms. In Section III, we compare the asymptotic mean squared errors of the estimated eigenvectors. Section IV has the experimental results, and Section V has the concluding remarks.

II. Background and Definitions

In this section, we provide a mathematical background of the PCA algorithms, and give the definitions that are used here. We are given a sequence $\{A_k \in \Re^{d\times d}\}$, with $\lim_{k\to\infty} E[A_k \mid A_0, \cdots, A_{k-1}] = A$. The *p principal eigenvectors* of A are the p eigenvectors of A corresponding to the p largest eigenvalues. In the following discussion, we denote $\lambda_1 > \lambda_2 > \cdots > \lambda_p \geq \cdots \geq \lambda_d > 0$ as the eigenvalues of A, and ϕ_i as the eigenvector corresponding to λ_i such that $\phi_1, \cdots, \phi_d$ are orthonormal. Let $\Phi = [\phi_1 \cdots \phi_d]$ and $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_d)$ denote the matrix of eigenvectors and eigenvalues of A. Note that if ϕ_i is an eigenvector, then $a_i\phi_i$ for $a_i = \pm 1$ is also an eigenvector.

Let the ith column of W_k be $\mathbf{w}_k^i$ Then, algorithm (1) can be written as

$$\mathbf{w}_{k+1}^i = \mathbf{w}_k^i + \eta_k \left(A_k\mathbf{w}_k^i - \mathbf{w}_k^i\mathbf{w}_k^{i^T} A_k\mathbf{w}_k^i - \gamma \sum_{j=1}^{i-1} \mathbf{w}_k^j\mathbf{w}_k^{j^T} A_k\mathbf{w}_k^i \right) \quad \text{for } i = 1, \cdots, p. \tag{3}$$

Here parameter γ is greater than or equal to one. Algorithm (2) can be similarly written as

$$\mathbf{w}_{k+1}^i = \mathbf{w}_k^i + \eta_k \left(2A_k\mathbf{w}_k^i - \mathbf{w}_k^i\mathbf{w}_k^{i^T} A_k\mathbf{w}_k^i - \gamma \sum_{j=1}^{i-1} \mathbf{w}_k^j\mathbf{w}_k^{j^T} A_k\mathbf{w}_k^i - A_k\mathbf{w}_k^i\mathbf{w}_k^{i^T}\mathbf{w}_k^i - \gamma \sum_{j=1}^{i-1} A_k\mathbf{w}_k^j\mathbf{w}_k^{j^T}\mathbf{w}_k^i \right) \quad \text{for } i = 1, \cdots, p. \tag{4}$$

Since it is clear that we require more computation for (4) than (3) at each update, it is necessary to determine the relative convergence rates to compare the benefits at the cost of computation. For both algorithms, we have the following assumptions and propositions.

Assumption (A1): Each A_k is bounded with probability one, symmetric, real, nonnegative definite, and $\lim_{k\to\infty} E[A_k \mid A_0, \cdots, A_{k-1}] = A$, where A is positive definite.

Assumption (A2): $\{\eta_k \in \Re^+\}$ is a decreasing sequence, such that $\sum_{k=1}^{\infty} \eta_k = \infty$, $\sum_{k=0}^{\infty} \eta_k^r < \infty$ for some $r > 1$, and $\lim_{k\to\infty} \sup(\eta_k^{-1} - \eta_{k-1}^{-1}) < \infty$.

Assumption (A3): The p largest eigenvalues of A are positive and each of unit multiplicity.

Proposition (P1): For algorithms (3) and (4) let assumptions A1–A3 hold. Then $\mathbf{w}_k^i$ tends either to $+\phi_i$ or $-\phi_i$ with probability one as $k \to \infty$.

Proof: See [11] for algorithm (3), and [16] for algorithm (4). □

Proposition (P2): For algorithms (3) and (4) let assumptions A1–A3 hold. Assume that $\mathbf{w}_0^i$ is bounded. Then, there exists a uniform upper bound of η_k such that $\mathbf{w}_k^i$ is uniformly bounded w.p. 1.

Proof: See [9], [11] for algorithm (3). A similar analysis can be done for algorithm (4). □

III. Comparison of the Asymptotic Mean Squared Errors

In order to compare the asymptotic mean squared errors of the two algorithms, we use the stochastic approximation results of Fabian [3]. Fabian established rates of convergence with $\eta_k = \delta k^{-\alpha}$ for $\delta > 0$ and $1/2 < \alpha \leq 1$. In the following Theorem, we specialize Fabian's result to suit the present analysis.

Theorem 1 (Fabian): Consider the equation

$$\mathbf{e}_{k+1} = \mathbf{e}_k - k^{-\alpha}\Gamma_k\mathbf{e}_k + k^{-(\alpha+\beta)/2}\mathbf{v}_k \tag{5}$$

where $\mathbf{e}_k$, $\mathbf{v}_k \in \Re^d$, α, $\beta \in \Re$ $(0 < \alpha \leq 1)$ and $\Gamma_k \in \Re^{d\times d}$. Let $\Im_k$ be a nondecreasing sequence of σ-fields. Suppose Γ_k and $\mathbf{v}_{k-1}$ are $\Im_k$-measurable. Let

$$\Gamma_k \to \Gamma, \quad E_{\Im_k}[\mathbf{v}_k] = \mathbf{0}, \quad \text{and} \quad E_{\Im_k}[\mathbf{v}_k\mathbf{v}_k^Y] \to \Sigma \text{ as } k \to \infty \tag{6}$$

where Σ, $\Gamma \in \Re^{d\times d}$ and Γ is positive definite. Let $P \in \Re^{d\times d}$ be orthonormal such that $P^T\Gamma P = \Theta$, where $\Theta =$

$\mathrm{diag}(\theta_1,\cdots,\theta_d)$. Suppose that $\theta = \min\{\theta_1,\cdots,\theta_d\}$, then there exists $\beta \geq 0$ and $\beta_+ < 2\theta$ such that

$$\beta_+ = \beta \quad \text{if } \alpha = 1, \quad \text{and} \quad \beta_+ = 0 \quad \text{if } \alpha \neq 1. \tag{7}$$

Then the asymptotic distribution of $k^{\beta/2}\mathbf{e}_k$ is Gaussian with mean $\mathbf{0}$ and covariance PMP^T where

$$M_{ij} = (P^T \Sigma P)_{ij}(\theta_i + \theta_j - \beta_+)^{-1}. \tag{8}$$

Proof: See [3]. □

For each algorithm, let us consider $\mathbf{e}_k^i = \mathbf{w}_k^i - a_i\phi_i$ $(a_i = \pm 1)$ as the error of the estimates. Then, following Theorem 1, we define AMSE as follows:

$$\mathrm{AMSE}_i = \lim_{k\to\infty} k^{\beta} E\big[\big\|\mathbf{e}_k^i\big\|^2\big].$$

Outline of Approach and Summary of Results: In the following two sections, we first present each algorithm [i.e., algorithms (1) and (2)] in the format of (5). We next evaluate the parameters given in Theorem 1, and finally, compute the asymptotic mean squared error for each algorithm. Based on the analyzes, our primary results can be summarized in the theorems below.

Theorem 2: For both algorithms (1) and (2), the asymptotic mean squared errors (AMSE's) for the minor eigenvectors increase for larger values of γ, with the smallest AMSE for $\gamma = 1$. Furthermore, by increasing δ ($\eta_k = \delta k^{-\alpha}$ where $\delta > 0$ and $1/2 < \alpha \leq 1$), up to an upper bound, we increase the AMSE's of all eigenvectors for both algorithms.

Theorem 3: If $\gamma > 1$, then algorithm (2) leads to a smaller AMSE than algorithm (1) for the minor eigenvectors.

Proof: The proofs of Theorems 2 and 3 follow from (25) and (33) below. □

A. Analysis for Algorithm (1)

Step 1: Formulation of Algorithm (1) as (5): We now analyze algorithm (1) in the framework of (5) in Theorem 1. In the representation of algorithm (1) as given in (3), we have $\eta_k = \delta k^{-\alpha}$ for $1/2 < \alpha \leq 1$. Notice that this choice of η_k satisfies assumption A2. We define

$$\mathbf{h}\big(\mathbf{w}_k^i, A_k\big) = A_k\mathbf{w}_k^i - \mathbf{w}_k^i\mathbf{w}_k^{i^T}A_k\mathbf{w}_k^i - \gamma\sum_{j=1}^{i-1}\mathbf{w}_k^j\mathbf{w}_k^{j^T}A_k\mathbf{w}_k^i. \tag{9}$$

Let $\Im_k$ be the nondecreasing σ-algebra generated by $A_0,\cdots,A_{k-1}$. Furthermore, let $E_{\Im_k}[A_k] = \bar{A}_k$ such that as $k\to\infty$, $\bar{A}_k \to A$ w.p. 1. Define

$$\bar{\mathbf{h}}\big(\mathbf{w}_k^i\big) = E_{\Im_k}\big[\mathbf{h}\big(\mathbf{w}_k^i, A_k\big)\big].$$

Then,

$$\bar{\mathbf{h}}\big(\mathbf{w}_k^i\big) = \bar{A}_k\mathbf{w}_k^i - \mathbf{w}_k^i\mathbf{w}_k^{i^T}\bar{A}_k\mathbf{w}_k^i - \gamma\sum_{j=1}^{i-1}\mathbf{w}_k^j\mathbf{w}_k^{j^T}\bar{A}_k\mathbf{w}_k^i. \tag{10}$$

We now represent (3) as follows:

$$\begin{aligned}\mathbf{w}_{k+1}^i - a_i\phi_i = \mathbf{w}_k^i - a_i\phi_i + \eta_k\bar{\mathbf{h}}\big(\mathbf{w}_k^i\big)\\ + \eta_k\big(\mathbf{h}\big(\mathbf{w}_k^i, A_k\big) - \bar{\mathbf{h}}\big(\mathbf{w}_k^i\big)\big)\end{aligned} \tag{11}$$

where $a_i = \pm 1$. By using the mean value theorem [15], we obtain from (10)

$$\bar{\mathbf{h}}\big(\mathbf{w}_k^i\big) = \frac{d\bar{\mathbf{h}}\big(\bar{\mathbf{w}}_k^i\big)}{d\mathbf{w}_k^i}\big(\mathbf{w}_k^i - a_i\phi_i\big) \tag{12}$$

where $\bar{\mathbf{w}}_k^i$ is on the line segment joining $\mathbf{w}_k^i$ and $a_i\phi_i$ for $a_i = \pm 1$.

Let $\mathbf{e}_k^i = \mathbf{w}_k^i - a_i\phi_i$ $(a_i = \pm 1)$ be the error of $\mathbf{w}_k^i$ from its asymptotically stable point $a_i\phi_i$. Then, from (11), (12) and $\eta_k = \delta k^{-\alpha}$, we obtain

$$\begin{aligned}\mathbf{e}_{k+1}^i = \mathbf{e}_k^i + \delta k^{-\alpha}\frac{d\bar{\mathbf{h}}\big(\bar{\mathbf{w}}_k^i\big)}{d\mathbf{w}_k^i}\mathbf{e}_k^i\\ + \delta k^{-(\alpha+\beta)/2}k^{-(\alpha-\beta)/2}\big(\mathbf{h}\big(\mathbf{w}_k^i, A_k\big) - \bar{\mathbf{h}}\big(\mathbf{w}_k^i\big)\big).\end{aligned} \tag{13}$$

Comparing (13) with (5), we obtain

$$\begin{aligned}\Gamma_k^i = -\delta\frac{d\bar{\mathbf{h}}\big(\bar{\mathbf{w}}_k^i\big)}{d\mathbf{w}_k^i} = \delta\Bigg(\bar{\mathbf{w}}_k^{i^T}\bar{A}_k\bar{\mathbf{w}}_k^i I + 2\bar{\mathbf{w}}_k^i\bar{\mathbf{w}}_k^{i^T}\bar{A}_k\\ + \gamma\sum_{j=1}^{i-1}\bar{\mathbf{w}}_k^j\bar{\mathbf{w}}_k^{j^T}\bar{A}_k - \bar{A}_k\Bigg)\end{aligned} \tag{14}$$

and

$$\begin{aligned}\mathbf{v}_k^i = \delta k^{-(\alpha-\beta)/2}\Bigg(I - \mathbf{w}_k^i\mathbf{w}_k^{i^T} - \gamma\sum_{j=1}^{i-1}\mathbf{w}_k^j\mathbf{w}_k^{j^T}\Bigg)\\ \times (A_k - \bar{A}_k)\mathbf{w}_k^i.\end{aligned} \tag{15}$$

Step 2: Optimal Choices of α, β, β_+^i, and θ^i: From (15), we see that for $E_{\Im_k}[\mathbf{v}_k^i\mathbf{v}_k^{i^T}]$ to converge, a sufficient condition is $\alpha \geq \beta$. In addition, we need β to be as large as possible, in order to obtain the highest rate of convergence, since the rate of convergence is proportional to $k^{\beta/2}$ by Theorem 1. Since $\alpha \leq 1$, we obtain the "best rate" of convergence for $\alpha = \beta = 1$.

We further observe from (15) that $E_{\Im_k}[\mathbf{v}_k^i] = \mathbf{0}$ Since by Proposition P1, we have $\mathbf{w}_k^i \to \pm\phi_i$ w.p. 1 as $k\to\infty$, and since Γ_k^i is a continuous function of $\mathbf{w}_k^i$, we obtain from (14)

$$\Gamma_k^i \xrightarrow{k} \Gamma_i = \delta\Bigg(\lambda_i I + 2\lambda_i\phi_i\phi_i^T + \gamma\sum_{j=1}^{i-1}\lambda_j\phi_j\phi_j^T - A\Bigg). \tag{16}$$

We see from (16) that the eigenvectors of Γ_i are $\phi_1, \phi_2, \cdots, \phi_d$, and thus, $P_i = \Phi$. The eigenvalues of Γ_i are given by

$$\Gamma_i\phi_q = \begin{cases}\delta(\lambda_i + (\gamma-1)\lambda_q)\phi_q & \text{for } q < i\\ 2\delta\lambda_i\phi_q & \text{for } q = i\\ \delta(\lambda_i - \lambda_q)\phi_q & \text{for } q > i.\end{cases} \tag{17}$$

Hence, Θ_i is given by (we assume $i > 1$, and the results for $i = 1$ are similar)

$$\begin{aligned}\Theta_i = \delta\mathrm{diag}(\lambda_i + (\gamma-1)\lambda_1,\cdots,\lambda_i + (\gamma-1)\lambda_{i-1}, 2\lambda_i, \lambda_i\\ - \lambda_{i+1},\cdots,\lambda_i - \lambda_d).\end{aligned} \tag{18}$$

Therefore,

$$\theta^i = \delta \min\{\lambda_i + (\gamma-1)\lambda_1, \cdots, \lambda_i + (\gamma-1)\lambda_{i-1}, 2\lambda_i, \lambda_i - \lambda_{i+1}, \cdots, \lambda_i - \lambda_d\} = \delta(\lambda_i - \lambda_{i+1})$$

and

$$\beta_+^i < 2\theta^i = 2\delta(\lambda_i - \lambda_{i+1}).$$

This gives us two conditions for the choice of α, β, and β_+^i as shown below.

Condition 1) If $\delta(\lambda_i - \lambda_{i+1}) \le 1/2$ then choose $\alpha = \beta \ne 1$, and $\beta_+^i = 0$.

Condition 2: If $\delta(\lambda_i - \lambda_{i+1}) > 1/2$ then choose $\alpha = \beta = 1$, and $\beta_+^i = 1$.

Step 3: Computation of the Asymptotic Mean Squared Error (AMSE): By using Proposition P1 and Assumption A1, we obtain from (15)

$$\begin{aligned}\Sigma_i &= \lim_{k\to\infty} E_{\Im_k}\left[\mathbf{v}_k^i \mathbf{v}_k^{i^T}\right] \\ &= \delta^2 \left[I - \phi_i\phi_i^T - \gamma \sum_{j=1}^{i-1} \phi_j\phi_j^T\right] \lim_{k\to\infty} E_{\Im_k} \\ &\quad \times \left[(A_k - A)\phi_i\phi_i^T(A_k - A)\right]\left[I - \phi_i\phi_i^T - \gamma \sum_{j=1}^{i-1} \phi_j\phi_j^T\right].\end{aligned} \tag{19}$$

The diagonal elements m_{jj}^i for $j = 1, \cdots, d$, of the covariance matrix M_i of the asymptotic Gaussian distribution are obtained from (8) as

$$m_{jj}^i = \begin{cases} (\Phi^T\Sigma_i\Phi)_{jj}\left(2\delta(\lambda_i + (\gamma-1)\lambda_j) - \beta_+^i\right)^{-1} & \text{for } j < i \\ (\Phi^T\Sigma_i\Phi)_{jj}\left(4\delta\lambda_i - \beta_+^i\right)^{-1} & \text{for } j = i \\ (\Phi^T\Sigma_i\Phi)_{jj}\left(2\delta(\lambda_i - \lambda_j) - \beta_+^i\right)^{-1} & \text{for } j > i. \end{cases} \tag{20}$$

Note that AMSE $= \mathrm{tr}[M_i]$. In order to compute this, we define

$$Z_i = \lim_{k\to\infty} E_{\Im_k}\left[(A_k - A)\phi_i\phi_i^T(A_k - A)\right].$$

From (19), we obtain

$$(\Phi^T\Sigma_i\Phi)_{jj} = \begin{cases} \delta^2(1-\gamma)^2\phi_j^T Z_i \phi_j & \text{for } j < i \\ 0 & \text{for } j = i \\ \delta^2\phi_j^T Z_i \phi_j & \text{for } j > i. \end{cases} \tag{21}$$

In order to evaluate the terms $\phi_j^T Z_i \phi_j$ in (21), we observe that

$$\begin{aligned}\phi_j^T Z_i \phi_j &= \lim_{k\to\infty} E_{\Im_k}\left[\left(\phi_j^T(A_k - A)\phi_i\right)^2\right] \\ &= \lim_{k\to\infty} E_{\Im_k}\left[\left(\phi_j^T A_k \phi_i\right)^2\right]\end{aligned} \tag{22}$$

since $\phi_j^T A \phi_i = 0$ for $j \ne i$. If we further assume that $A_k = \mathbf{x}_k\mathbf{x}_k^T$, and $\mathbf{x}_k \in \Re^d$ are from an asymptotically Gaussian distribution with zero mean and covariance A, then we obtain

$$\begin{aligned}\phi_j^T Z_i \phi_j &= \lim_{k\to\infty} E_{\Im_k}\left[\left(\phi_j^T A_k \phi_i\right)^2\right] \\ &= \lim_{k\to\infty} E_{\Im_k}\left[\left(\phi_j^T \mathbf{x}_k\mathbf{x}_k^T \phi_i\right)^2\right] \\ &= \lim_{k\to\infty} E_{\Im_k}\left[\left(\phi_j^T \mathbf{x}_k\right)^2\right] \lim_{k\to\infty} E_{\Im_k}\left[\left(\phi_i^T \mathbf{x}_k\right)^2\right] = \lambda_i\lambda_j.\end{aligned} \tag{23}$$

Under these assumptions, we have from (21) and (23)

$$(\Phi^T\Sigma_i\Phi)_{jj} = \begin{cases} \delta^2(1-\gamma)^2\lambda_i\lambda_j & \text{for } j < i \\ 0 & \text{for } j = i. \\ \delta^2\lambda_i\lambda_j & \text{for } j > i \end{cases} \tag{24}$$

The AMSE for algorithm (1) is now given by

$$\begin{aligned}\mathrm{AMSE}_i = \mathrm{tr}[M_i] &= \sum_{j=1}^{i-1} \frac{\delta^2(1-\gamma)^2\lambda_i\lambda_j}{2\delta(\lambda_i + (\gamma-1)\lambda_j) - \beta_+^i} \\ &\quad + \sum_{j=i+1}^{d} \frac{\delta^2\lambda_i\lambda_j}{2\delta(\lambda_i - \lambda_j) - \beta_+^i} \quad \text{for } i = 2, \cdots, d.\end{aligned} \tag{25}$$

B. Analysis for Algorithm (2)

Step 1: Formulation of Algorithm (2) as (5): We now repeat the above analysis for algorithm (2). From (4), we obtain

$$\begin{aligned}\mathbf{h}(\mathbf{w}_k^i, A_k) &= 2A_k\mathbf{w}_k^i - \mathbf{w}_k^i\mathbf{w}_k^{i^T}A_k\mathbf{w}_k^i - \gamma\sum_{j=1}^{i-1}\mathbf{w}_k^j\mathbf{w}_k^{j^T}A_k\mathbf{w}_k^i \\ &\quad - A_k\mathbf{w}_k^i\mathbf{w}_k^{i^T}\mathbf{w}_k^i - \gamma\sum_{j=1}^{i-1}A_k\mathbf{w}_k^j\mathbf{w}_k^{j^T}\mathbf{w}_k^i\end{aligned} \tag{26}$$

and therefore

$$\begin{aligned}\bar{\mathbf{h}}(\mathbf{w}_k^i) &= 2\bar{A}_k\mathbf{w}_k^i - \mathbf{w}_k^i\mathbf{w}_k^{i^T}\bar{A}_k\mathbf{w}_k^i - \gamma\sum_{j=1}^{i-1}\mathbf{w}_k^j\mathbf{w}_k^{j^T}\bar{A}_k\mathbf{w}_k^i \\ &\quad - \bar{A}_k\mathbf{w}_k^i\mathbf{w}_k^{i^T}\mathbf{w}_k^i - \gamma\sum_{j=1}^{i-1}\bar{A}_k\mathbf{w}_k^j\mathbf{w}_k^{j^T}\mathbf{w}_k^i\end{aligned} \tag{27}$$

and

$$\begin{aligned}\Gamma_k^i = -\delta\frac{d\bar{\mathbf{h}}(\bar{\mathbf{w}}_k^i)}{d\mathbf{w}_k^i} &= \delta\Bigg(\bar{\mathbf{w}}_k^{i^T}\bar{A}_k\bar{\mathbf{w}}_k^i I + 2\bar{\mathbf{w}}_k^i\bar{\mathbf{w}}_k^{i^T}\bar{A}_k \\ &\quad + \gamma\sum_{j=1}^{i-1}\bar{\mathbf{w}}_k^j\bar{\mathbf{w}}_k^{j^T}\bar{A}_k + \bar{A}_k\bar{\mathbf{w}}_k^{i^T}\bar{\mathbf{w}}_k^i + 2\bar{A}_k\bar{\mathbf{w}}_k^i\bar{\mathbf{w}}_k^{i^T} \\ &\quad + \gamma\sum_{j=1}^{i-1}\bar{A}_k\bar{\mathbf{w}}_k^j\bar{\mathbf{w}}_k^{j^T} - 2\bar{A}_k\Bigg).\end{aligned} \tag{28}$$

Due to the w.p. 1 convergence of $\mathbf{w}_k^i$ to $\pm\phi_i$ by Proposition P1, and since Γ_k^i is a continuous function of $\mathbf{w}_k^i$, we obtain from (28)

$$\Gamma_k^i \xrightarrow{k} \Gamma_i = \delta\left(\lambda_i I + 4\lambda_i\phi_i\phi_i^T + 2\gamma\sum_{j=1}^{i-1}\lambda_j\phi_j\phi_j^T - A\right). \tag{29}$$

As in algorithm (1), the eigenvectors of Γ_i are $\phi_1, \phi_2, \cdots, \phi_d$, and thus, $P_i = \Phi$. The eigenvalues of Γ_i are given by

$$\Gamma_i\phi_q = \begin{cases} \delta(\lambda_i + (2\gamma-1)\lambda_q)\phi_q & \text{for } q < i \\ 4\delta\lambda_i\phi_q & \text{for } q = i \\ \delta(\lambda_i - \lambda_q)\phi_q & \text{for } q > i \end{cases} \tag{30}$$

and $\theta_i = \delta(\lambda_i - \lambda_{i+1})$, and $\beta_+^i < 2\theta_i = 2\delta(\lambda_i - \lambda_{i+1})$. The two conditions for the choice of α, β, and β_+^i as given in Section III-A are still valid.

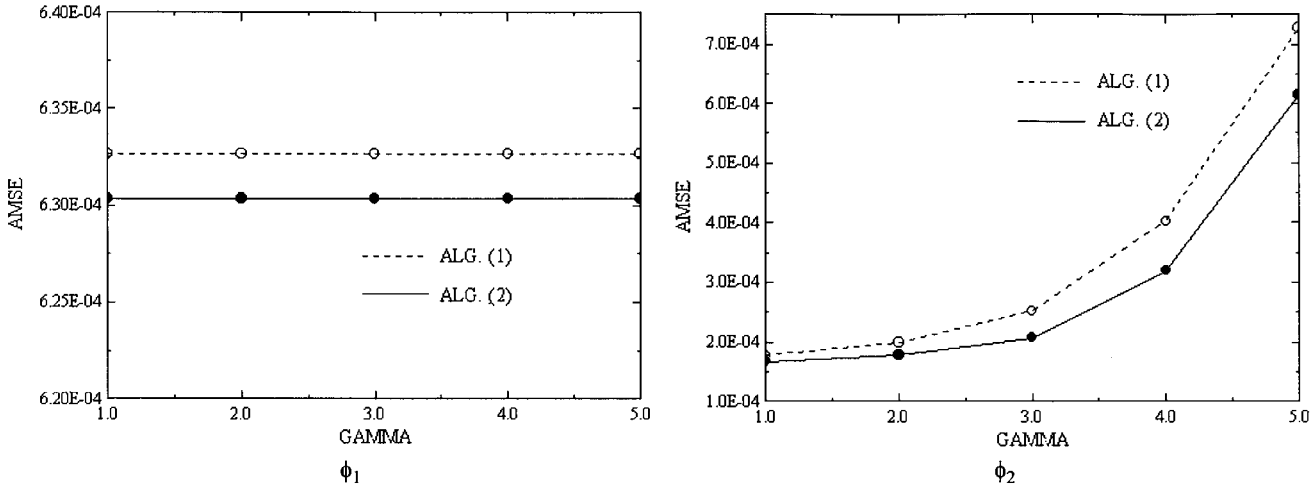

Fig. 1. AMSE's for ϕ_1 and ϕ_2 for different γ for algorithms (1) and (2) with $\eta_k = 0.5/(1000 + k)$.

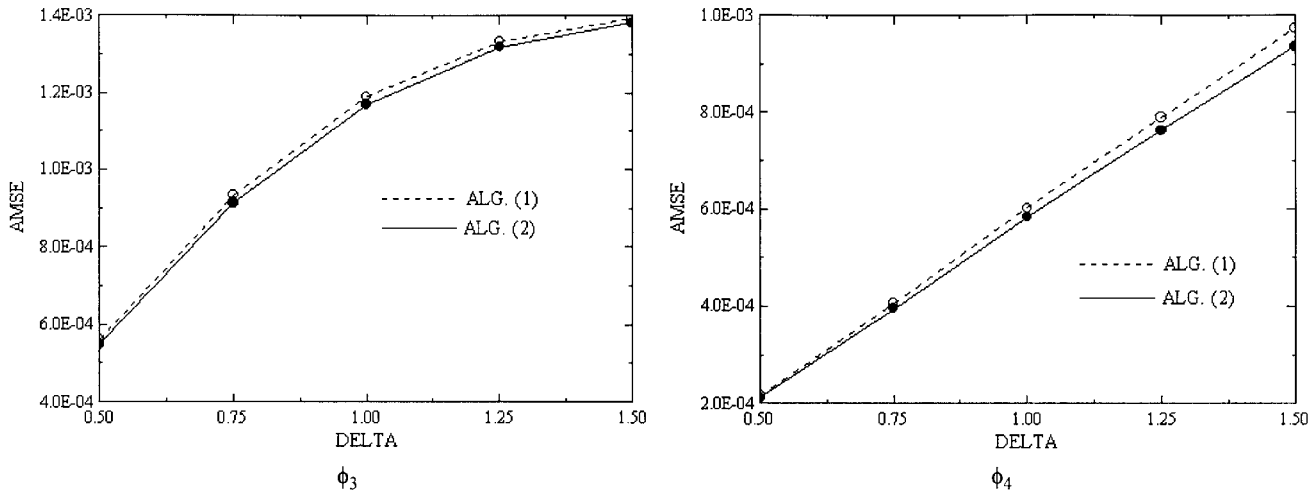

Fig. 2. AMSE's for ϕ_3 and ϕ_4 for different δ for algorithms (1) and (2) with $\gamma = 2$.

TABLE I
AMSE'S FOR ALGORITHMS (1) AND (2) FOR ϕ_1 THROUGH ϕ_4 WITH $\gamma = 2$

	ϕ_1		ϕ_2		ϕ_3		ϕ_4	
δ	Alg. (1)	Alg. (2)	Alg. (1)	Alg. (2)	Alg. (1)	Alg. (2)	Alg. (1)	Alg. (2)
0.50	63.26E-5	63.04E-5	16.66E-5	16.50E-5	55.89E-5	54.80E-5	21.44E-5	21.05E-5
0.75	104.60E-5	104.19E-5	30.97E-5	30.62E-5	93.26E-5	91.35E-5	40.57E-5	39.48E-5
1.00	133.53E-5	133.08E-5	45.97E-5	45.38E-5	119.02E-5	116.98E-5	60.10E-5	58.28E-5
1.25	149.92E-5	149.53E-5	60.72E-5	59.77E-5	133.34E-5	131.91E-5	78.95E-5	76.28E-5
1.50	158.10E-5	157.78E-5	75.68E-5	74.23E-5	139.42E-5	139.11E-5	97.34E-5	93.57E-5

Step 3: Computation of the AMSE: We now study the evaluation of $\Sigma_i = \lim_{k\to\infty} E_{\Im_k}[\mathbf{v}_k^i \mathbf{v}_k^{i^T}]$. We obtain from (4)

$$\mathbf{v}_k^i = \delta k^{-(\alpha-\beta)/2}\left(I - \mathbf{w}_k^i \mathbf{w}_k^{i^T} - \gamma \sum_{j=1}^{i-1} \mathbf{w}_k^j \mathbf{w}_k^{j^T}\right) \times (A_k - \bar{A}_k)\mathbf{w}_k^i + \delta k^{-(\alpha-\beta)/2}(A_k - \bar{A}_k) \times \left(I - \mathbf{w}_k^i \mathbf{w}_k^{i^T} - \gamma \sum_{j=1}^{i-1} \mathbf{w}_k^j \mathbf{w}_k^{j^T}\right)\mathbf{w}_k^i. \quad (31)$$

The first term of (31) is same as in (15) in algorithm (1), whereas the expectation of the second term is zero. Since $\mathbf{v}_k^i$ is bounded for all k, by an application of the bounded convergence theorem [15], we obtain the same expression for Σ_i as in (19). We, therefore, obtain the same expressions of $(\Phi^T \Sigma_i \Phi)_{jj}$ as in (21) and (24). The diagonal elements m_{jj}^i for $j = 1, \cdots, d$, of the covariance matrix M_i of the asymptotic Gaussian distribution are

$$m_{jj}^i = \begin{cases} (\Phi^T \Sigma_i \Phi)_{jj}\left(2\delta(\lambda_i + (2\gamma - 1)\lambda_j) - \beta_+^i\right)^{-1} & \text{for } j < i \\ (\Phi^T \Sigma_i \Phi)_{jj}\left(8\delta\lambda_i - \beta_+^i\right)^{-1} & \text{for } j = i \\ (\Phi^T \Sigma_i \Phi)_{jj}\left(2\delta(\lambda_i - \lambda_j) - \beta_+^i\right)^{-1} & \text{for } j > i. \end{cases} \quad (32)$$

The asymptotic mean squared error (AMSE) for algorithm (2)

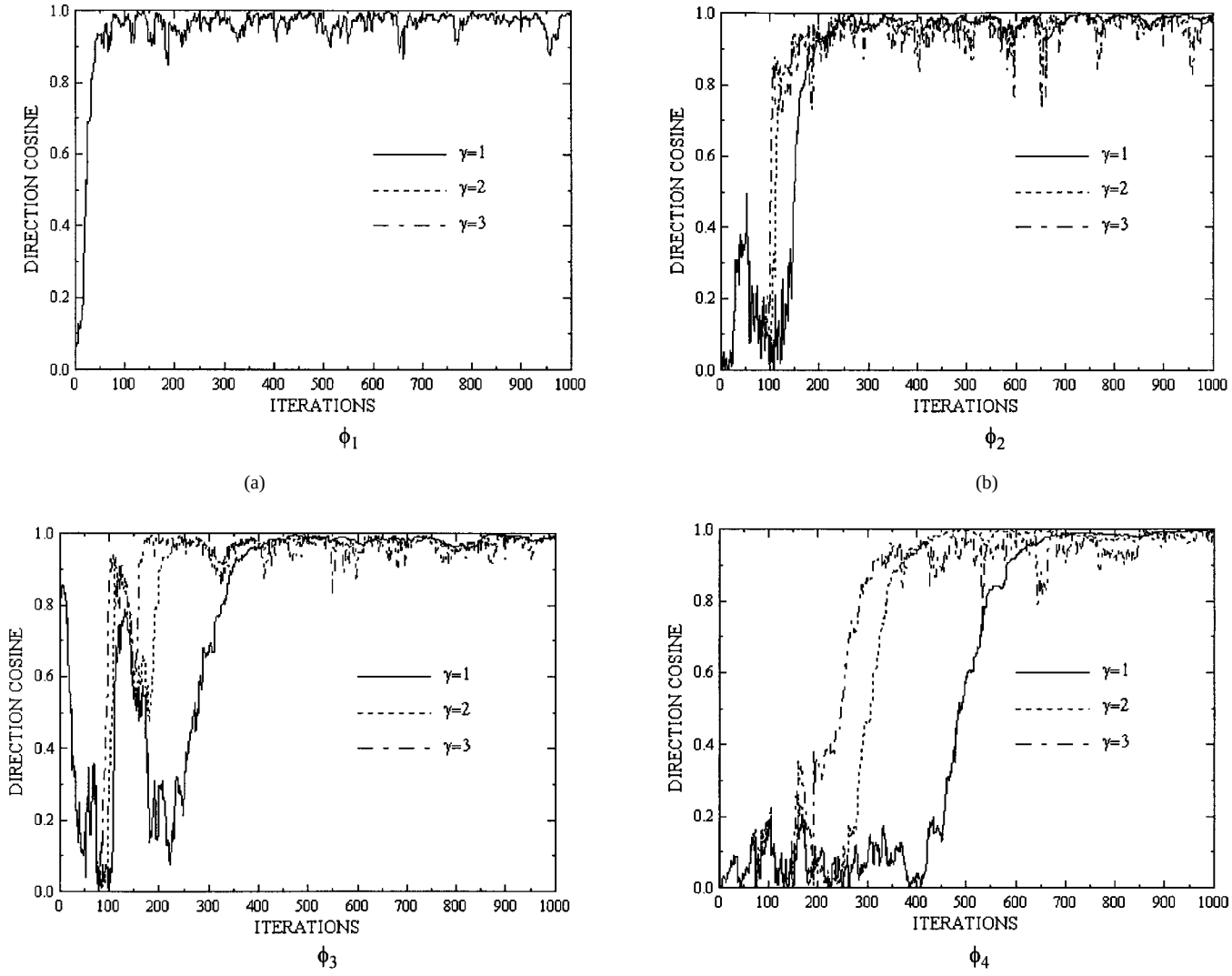

Fig. 3. Iterates of ϕ_1 through ϕ_4 for different γ for algorithm (1) with $\eta_k = 1/(1000 + k)$.

is now given by

$$\text{AMSE}_i = \text{tr}[M_i] = \sum_{j=1}^{i-1} \frac{\delta^2(1-\gamma)^2\lambda_i\lambda_j}{2\delta(\lambda_i + (2\gamma-1)\lambda_j) - \beta_+^i} + \sum_{j=i+1}^{d} \frac{\delta^2\lambda_i\lambda_j}{2\delta(\lambda_i - \lambda_j) - \beta_+^i} \quad \text{for } i = 2, \cdots, d. \tag{33}$$

IV. Experimental Results

We generate ten-dimensional Gaussian data from the first covariance matrix in [12], with the covariance matrix scaled by 20 (as given below) for data in $\Re^{10}$, as shown in the matrix at the bottom of the page. The ten eigenvalues of A are 117.996, 55.644, 34.175, 20.589, 7.873, 5.878, 1.743, 1.423, 1.213, and 1.007.

$$A = 20 \begin{bmatrix} 0.091 & & & & & & & & & \\ 0.038 & 0.373 & & & & & & & & \\ -0.053 & 0.018 & 1.430 & & & & & & & \\ -0.005 & -0.028 & 0.017 & 0.084 & & & & & & \\ 0.010 & -0.011 & 0.055 & -0.005 & 0.071 & & & & & \\ -0.136 & -0.367 & -0.450 & 0.016 & 0.088 & 5.720 & & & & \\ 0.155 & 0.154 & -0.038 & 0.042 & 0.058 & -0.544 & 2.750 & & & \\ 0.030 & -0.057 & -0.298 & -0.022 & -0.069 & -0.248 & -0.343 & 1.450 & & \\ 0.002 & -0.031 & -0.041 & 0.001 & -0.008 & 0.005 & -0.011 & 0.078 & 0.067 & \\ 0.032 & -0.065 & -0.030 & 0.005 & 0.003 & 0.095 & -0.120 & 0.028 & 0.015 & 0.341 \end{bmatrix}$$

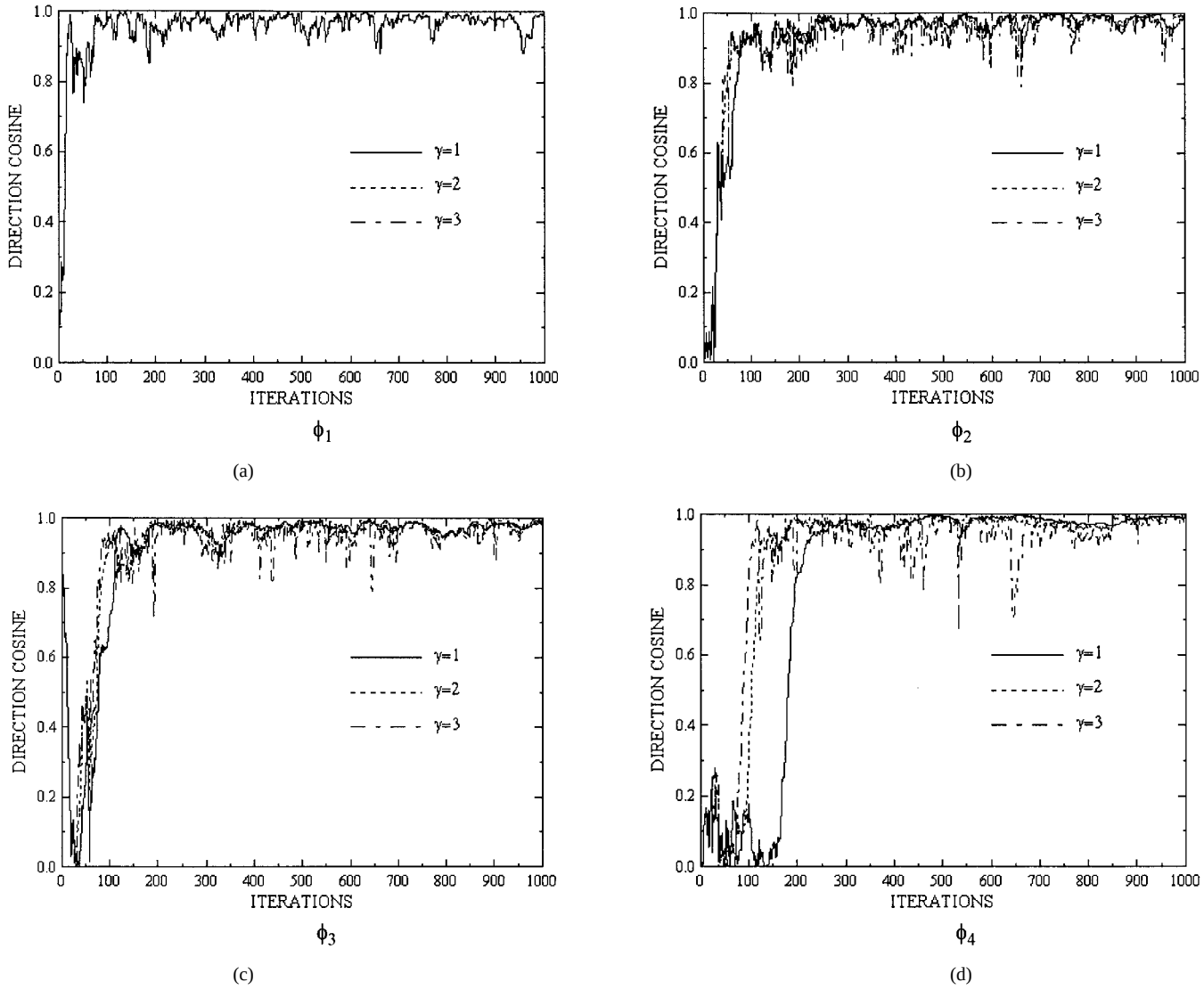

Fig. 4. Iterates of ϕ_1 through ϕ_4 for different γ for algorithm (2) with $\eta_k = 1/(1000 + k)$.

A. Experiments on Asymptotic Mean Squared Error (AMSE)

We generated 1000 samples of zero-mean multivariate Gaussian data in $\Re^{10}$. Let the kth sample be denoted by $\mathbf{x}_k$. We computed A_k as $\mathbf{x}_k\mathbf{x}_k^T$, and then obtained an estimate W of the eigenvector matrix Φ by algorithms (1) and (2) for all 1000 samples. We refer to one application of each algorithm for one sample A_k as *one iteration* of the algorithm. We refer to one application of all 1000 samples as one *epoch* of the algorithms.

In order to compute the AMSE's of the eigenvectors, we applied algorithms (1) and (2) for the first four principal eigenvectors for one through ten epochs. For each epoch, we collected the results for the last m iterations where m is varied from 10- to 50-in increments of ten. After computing the mean squared error (MSE) for the last m iterations, we plotted the MSE against m for each epoch. We found that the MSE remains constant (up to five places of decimal) for $m \leq 20$ for five epochs. Hence, we consider the MSE computed from the last 20 samples after five epochs as the AMSE of the estimates. We consider $\eta_k = \delta/(1000 + k)$.

We changed γ from 1- to 5-in increments of one with $\delta = 0.5$, and computed the AMSE's for the first two principal eigenvectors. We show the results of this experiment in Fig. 1, which reflects the following results stated in Theorem 2. We observe that:

1) the smallest AMSE is for $\gamma = 1$ for all eigenvectors;
2) the AMSE's increase as γ increases for the minor eigenvectors;
3) algorithm (2) leads to a smaller AMSE than algorithm (1) for the minor eigenvectors for $\gamma > 1$.

We next computed the AMSE's for the first four principal eigenvectors for δ between 0.5- and 1.5-in increments of 0.25 with $\gamma = 2$. We show the numerical results of this experiment in Table I.

Since the difference between the AMSE's for the two algorithms is significant only for ϕ_3 and ϕ_4, we show the graphical results for ϕ_3 and ϕ_4 only in Fig. 2.

Fig. 2 shows that the AMSE's increase as δ increases for ϕ_3 and ϕ_4 for both algorithms as stated in Theorem 2. The same result is obtained for ϕ_1 and ϕ_2.

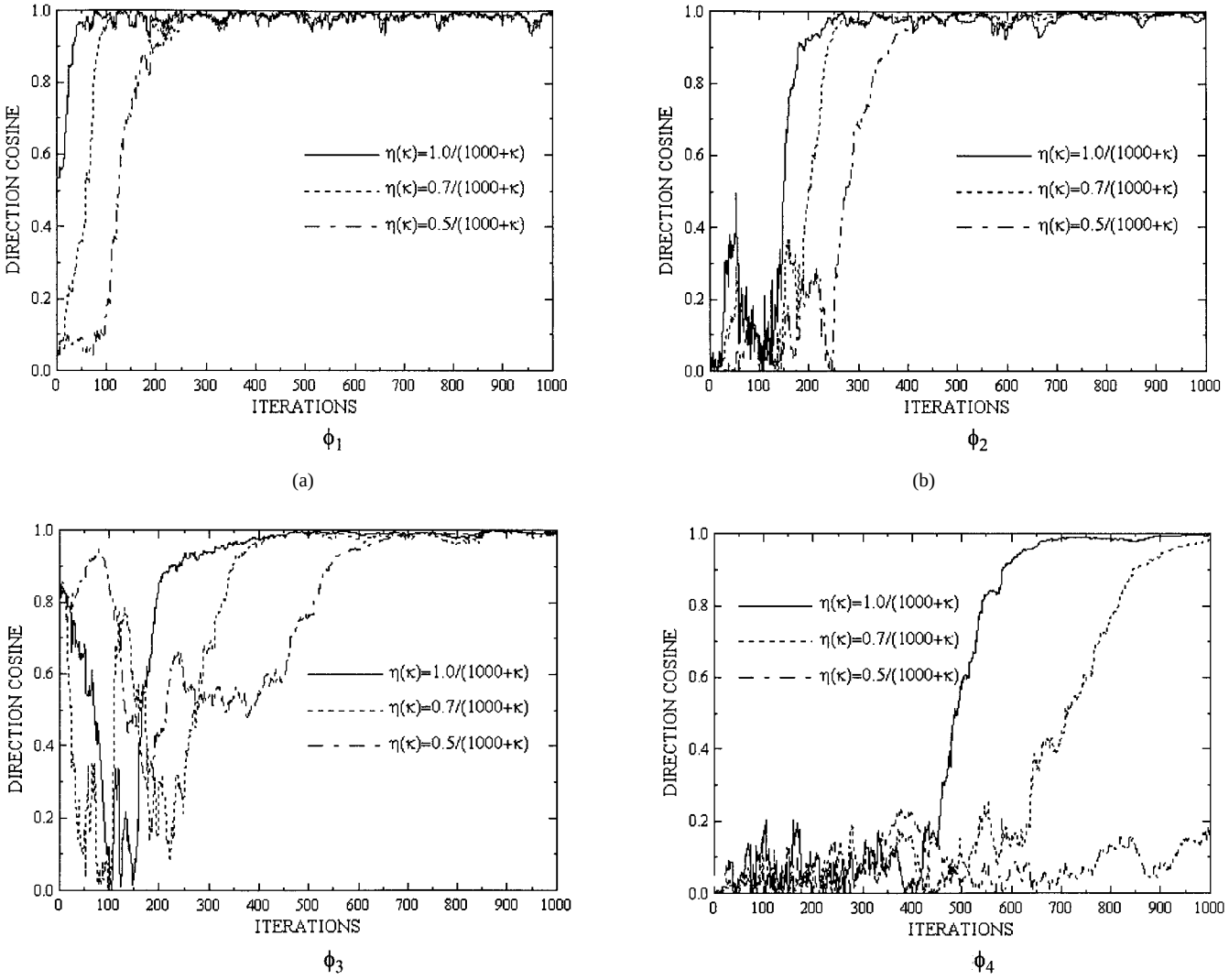

Fig. 5. Iterates of ϕ_1 through ϕ_4 for different δ for algorithm (1) with $\gamma = 1$.

B. Observed Convergence versus AMSE Tradeoffs

We next analyze the convergence of $\mathbf{w}_k^i$ to ϕ_i for different iterations of algorithms (1) and (2). In order to estimate the error for each eigenvector for each iteration, we compute the direction cosine given by

$$\text{Direction Cosine } (k) = \text{DC}(k) = \left|\mathbf{w}_k^{iT}\phi_i\right| / \left\|\mathbf{w}_k^i\right\| \|\phi_i\| \quad (34)$$

where $\mathbf{w}_k^i$ is the estimated ith principal eigenvector at the kth iteration of the adaptive algorithms, and ϕ_i is the actual ith principal eigenvector computed from all collected samples by a standard method [4]. Note that the maximum value of the direction cosine is one when $\mathbf{w}_k^i$ is exactly same as ϕ_i. Further note that the direction cosine at the kth iteration [denoted by DC(k)] is related to the error of the estimates at the kth iteration [denoted by error(k)] by the following formula:

$$\text{error}(k) = \left\| \frac{\mathbf{w}_k^i}{|\mathbf{w}_k^i|} - \phi_i \right\| = \sqrt{2(1 - \text{DC}(k))}. \quad (35)$$

In the following experiments, we present the results for one epoch.

Fig. 3 shows the iterates of the first four principal eigenvectors ϕ_1 through ϕ_4 for 1000 samples for choices of $\gamma = 1, 2$ and 3 for algorithm (1) with $\eta_k = 1/(1000 + k)$. Fig. 4 shows the same results for algorithm (2). From these figures, we observe that the estimates converge faster, i.e., the direction cosines goes to one faster as γ increases.

Fig. 5 shows the iterates of the first four principal eigenvectors ϕ_1 through ϕ_4 for 1000 samples for choices of $\delta = 1, 0.7$, and 0.5 (where $\eta_k = \delta k^{-\alpha}$) for algorithm (1) with $\gamma = 1$. Once again, the direction cosine goes to one faster for larger values of δ up to an upper bound of δ. Fig. 6 shows the same results for algorithm (2). Once again, we observe that the estimates converge faster as δ increases.

These experiments and several others show us a convergence-AMSE tradeoff in choosing γ and δ, i.e., increasing the values of γ and δ give us a faster convergence but a larger AMSE, and vice versa. An explanation for this

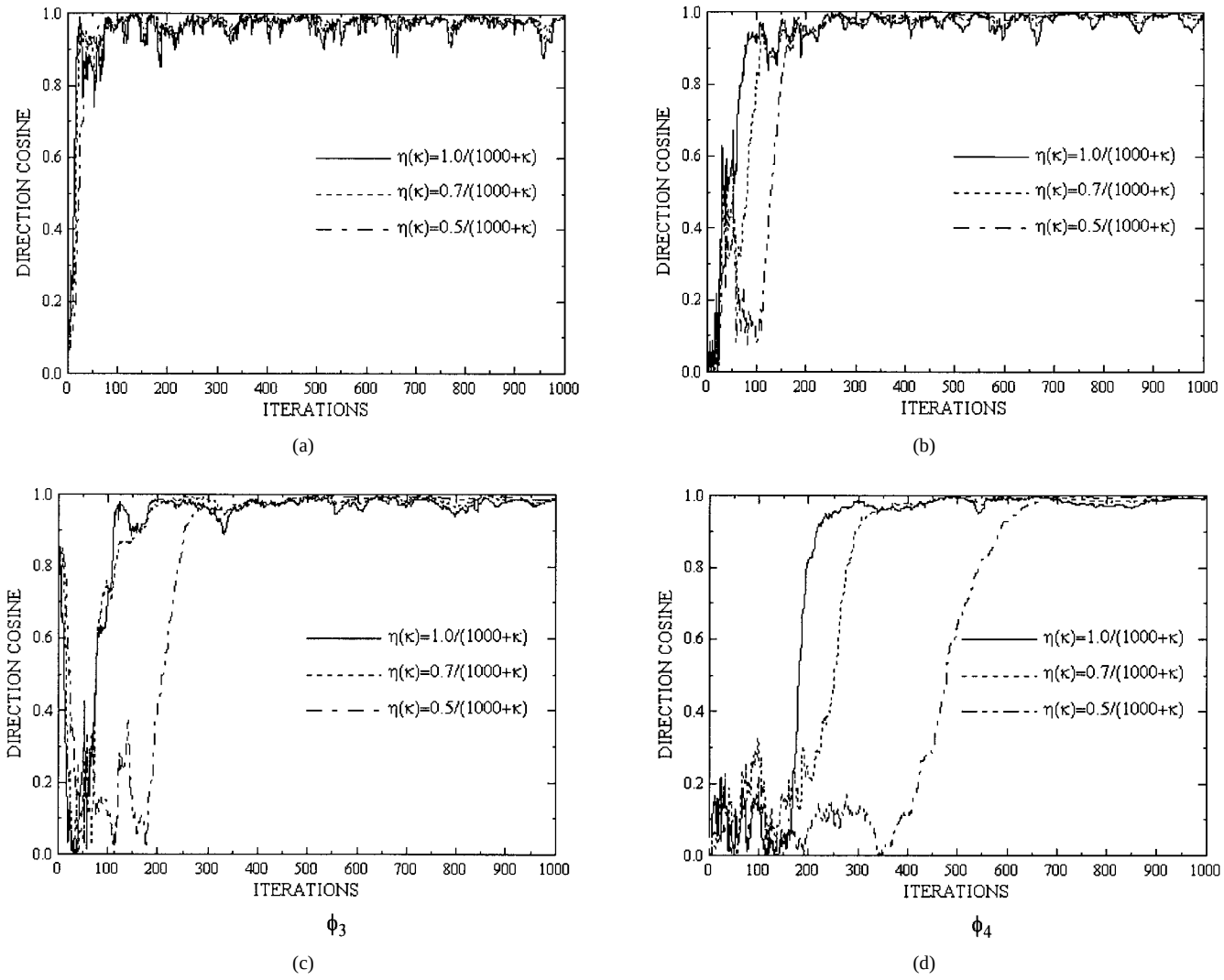

Fig. 6. Iterates of ϕ_1 through ϕ_4 for different δ for algorithm (2) with $\gamma = 1$.

observation (i.e., the dependence of convergence rate on γ) is given in Section IV-C.

C. Relative Rates of Convergence for Algorithms (1) and (2) (ODE Analysis)

A comparison of algorithms (1) and (2) for 1000 iterations with various γ and δ can be obtained by comparing Figs. 3 and 4 and also Figs. 5 and 6. This comparison with $\gamma = 1$ and $\delta = 1$ is illustrated in Fig. 7 for ϕ_1 through ϕ_4.

We see that algorithm (2) converges faster than algorithm (1) for the minor eigenvectors. A possible explanation for this observation (with a given γ) can be obtained by analyzing the ODE's associated with the stochastic approximation procedures (3) and (4). A similar analysis was done by Oja [10] for algorithm (1). We extend Oja's analysis to compare algorithms (1) and (2).

It can be shown [2], [7], [8] that the asymptotic limits of algorithms (3) and (4) can be solved by analyzing the corresponding continuous-time ODE's. Denoting the continuous time counterpart of $\mathbf{w}_k^i$ as $\mathbf{w}_i(t)$, with t denoting continuous time, the ODE's for the two algorithms are as follows.

Algorithm (3):

$$\frac{d\mathbf{w}_i(t)}{dt} = A\mathbf{w}_i - \mathbf{w}_i\mathbf{w}_i^T A\mathbf{w}_i - \gamma \sum_{j=1}^{i-1} \mathbf{w}_j\mathbf{w}_j^T A\mathbf{w}_i \quad (i = 1, \cdots, p). \quad (36)$$

Algorithm (4):

$$\frac{d\mathbf{w}_i(t)}{dt} = 2A\mathbf{w}_i - \mathbf{w}_i\mathbf{w}_i^T A\mathbf{w}_i - \gamma \sum_{j=1}^{i-1} \mathbf{w}_j\mathbf{w}_j^T A\mathbf{w}_i - A\mathbf{w}_i\mathbf{w}_i^T\mathbf{w}_i - \gamma \sum_{j=1}^{i-1} A\mathbf{w}_j\mathbf{w}_j^T\mathbf{w}_i \quad (i = 1, \cdots, p). \quad (37)$$

To illustrate the behavior of the two ODE's near the solution, consider only the last vector $\mathbf{w}_p(t)$. Since the previous vectors $\mathbf{w}_1, \cdots, \mathbf{w}_{p-1}$ are independent of $\mathbf{w}_p$, it can be assumed that they have already converged to $\phi_1, \cdots, \phi_{p-1}$ respectively. This assumption is supported in Figs. 5 and 6. For example, for ϕ_4 (see Fig. 5), algorithm (1) starts converging at $k > 400$, whereas ϕ_1 through ϕ_3 are within 95% of their final value at

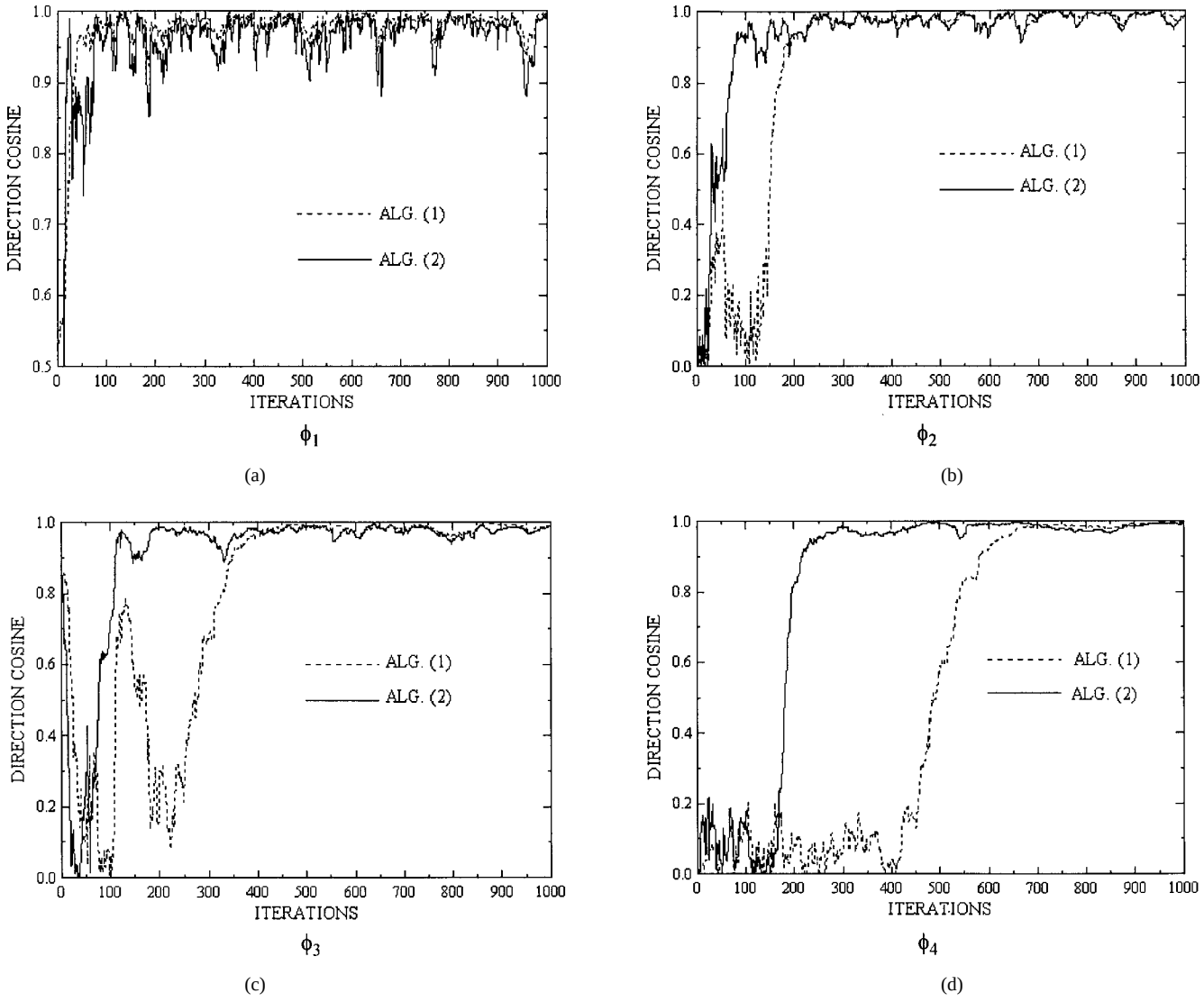

Fig. 7. Iterates of ϕ_1 through ϕ_4 for algorithms (1) and (2) with $\gamma = 1$ and $\eta_k = 1/(1000 + k)$.

$k = 400$. Defining error $\mathbf{e}_p(t) = \mathbf{w}_p(t) - a_p\phi_p$ ($a_p = \pm 1$), we have from (36)

$$\frac{d\mathbf{e}_p(t)}{dt} = \left(A - \lambda_p I - 2\lambda_p \phi_p \phi_p^T - \gamma \sum_{j=1}^{p-1} \lambda_j \phi_j \phi_j^T \right) \mathbf{e}_p + \mathbf{g}(\mathbf{e}_p) \quad (38)$$

where $\mathbf{g}(\mathbf{e}_p)$ has the properties $\mathbf{g}(\mathbf{0}) = \mathbf{0}$ and $\partial \mathbf{g}(\mathbf{0})/\partial \mathbf{e}_p = 0$. Following Theorem 2.4 of Hale [5, p. 86], the asymptotic stability of zero as the solution of (38) is determined by the linear part. If at a certain moment t

$$\mathbf{e}_p(t) = \sum_{j=1}^{d} b_j(t)\phi_j \quad (39)$$

then from (38) we have

$$\frac{db_j(t)}{dt} = \begin{cases} -((\gamma - 1)\lambda_j + \lambda_p)b_j & \text{for } j < p \\ -2\lambda_j b_j & \text{for } j = p \\ -(\lambda_p - \lambda_j)b_j & \text{for } j > p. \end{cases} \quad (40)$$

By repeating the above analysis for algorithm (2), we obtain

$$\frac{db_j(t)}{dt} = \begin{cases} -((2\gamma - 1)\lambda_j + \lambda_p)b_j & \text{for } j < p \\ -4\lambda_j b_j & \text{for } j = p \\ -(\lambda_p - \lambda_j)b_j & \text{for } j > p. \end{cases} \quad (41)$$

Equations (40) and (41) show that for the choice of $\gamma > 1$, the components in the direction of the more dominant eigenvectors die out faster for the ODE corresponding to algorithm (2) than for algorithm (1). These equations [i.e., (40) and (41)] also provide a possible explanation for faster convergence of both algorithms for $\gamma > 1$.

V. Concluding Remarks

We discuss the convergence properties for two important principal component analysis algorithms. We show that in-

creasing the values of parameters γ and δ results in larger AMSE's for both algorithms. On the other hand, experimental results and an analytical study suggests that increasing the values of these parameters make the convergence faster for both algorithms. Hence, we obtain a tradeoff: increasing the values of γ and δ results in larger AMSE's but faster convergence and vice versa.

Moreover, although algorithm (2) has a larger computation in each iteration, it leads to a smaller AMSE and faster convergence for the minor eigenvectors. Thus, we demonstrate a computation versus convergence tradeoff for the two algorithms.

REFERENCES

[1] P. Baldi and K. Hornik, "Learning in linear neural networks: A survey," *IEEE Trans. Neural Networks*, vol. 6, pp. 837–857, 1995.

[2] A. Benveniste, A. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*. New York: Springer-Verlag, 1990.

[3] V. Fabian, "On asymptotic normality in stochastic approximation," *Ann. Math. Statist.*, vol. 39, no. 4, pp. 1327–1332, 1968.

[4] G. H. Golub and C. F. VanLoan, *Matrix Computations*. Baltimore, MD: Johns Hopkins Univ. Press, 1983.

[5] J. K. Hale, *Ordinary Differential Equations*. New York: Wiley, 1969.

[6] S. Haykin, *Neural Networks—A Comprehensive Foundation*. New York: Macmillan, 1994.

[7] L. Ljung, "Analysis of recursive stochastic algorithms," *IEEE Trans. Automat. Contr.*, vol. AC-22, pp. 551–575, 1977.

[8] L. Ljung, G. Pflug, and H. Walk, *Stochastic Approximation and Optimization of Random Systems*. Boston, MA: Birkhauser, 1992.

[9] J. Karhunen, "Stability of Oja's PCA subspace rule," *Neural Computa.*, vol. 6, pp. 739–747, 1994.

[10] E. Oja, "Principal components, minor components, and linear neural networks," *Neural Networks*, vol. 5, pp. 927–935, 1992.

[11] E. Oja and J. Karhunen, "On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix," *J. Math. Anal. Applicat.*, vol. 106, pp. 69–84, 1985.

[12] T. Okada and S. Tomita, "An optimal orthonormal system for discriminant analysis," *Pattern Recognition*, vol. 18, no. 2, pp. 139–144, 1985.

[13] J. Rubner and P. Tavan, "A self-organizing network for principal component analysis," *Europhys. Lett.*, vol. 10, no. 7, pp. 693–698, 1989.

[14] T. D. Sanger, "Optimal unsupervised learning in a single-layer linear feedforward neural network," *Neural Networks*, vol. 2, pp. 459–473, 1989.

[15] R. L. Wheeden and A. Zygmund, *Measure and Integral—An Introduction to Real Analysis*. New York: Marcel Dekker, 1977.

[16] L. Xu, "Least mean square error reconstruction principle for self-organizing neural-nets," *Neural Networks*, vol. 6, pp. 627–648, 1993.

[17] L. Xu and A. L. Yuille, "Robust principal component analysis by self-organizing rules based on statistical physics approach," *IEEE Trans. Neural Networks*, vol. 6, pp. 131–143, 1995.

Chanchal Chatterjee (M'88) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Kanpur, India, and the M.S.E.E. degree from Purdue University, West Lafayette, IN, in 1983 and 1984, respectively. In 1996, he received the Ph.D. degree in electrical and computer engineering from Purdue.

Between 1985 and 1995 he worked at Machine Vision International and Medar Inc., both in Detroit, MI. He is currently Senior Algorithms Specialist at GDE Systems Inc., San Diego, CA. He is also affiliated with the Electrical Engineering Department at the University of California, Los Angeles. His areas of interest include image processing, computer vision, neural networks, and adaptive algorithms and systems for pattern recognition and signal processing.

Vwani P. Roychowdhury received the B.Tech. degree from the Indian Institute of Technology, Kanpur, India and the Ph.D. degree from Stanford University, Stanford, CA, in 1982 and 1989, respectively, all in electrical engineering.

He is currently a Professor in the Department of Electrical Engineering at the University of California, Los Angeles. From August 1991, until June 1996, he was a faculty member at the School of Electrical and Computer Engineering at Purdue University. He has coauthored several books including *Discrete Neural Computation: A Theoretical Foundation* (Englweood Cliffs, NJ: Prentice-Hall, 1995) and *Theoretical Advances in Neural Computation and Learning* (Boston, MA: Kluwer, 1994). His research interests include parallel algorithms and architectures, design and analysis of neural networks, application of computational principles to nanoelectronics, special purpose computing arrays, VLSI design, and fault-tolerant computation.

Edwin K. P. Chong (S'87–M'91–SM'96) received the B.E.(Hons.) degree with First Class Honors from the University of Adelaide, South Australia, in 1987, graduating top of his class, and the M.A. and Ph.D. degrees in 1989 and 1991, respectively, from Princeton University, NJ, where he held an IBM Graduate Fellowship.

Since August 1991, he has been with the School of Electrical and Computer Engineering at Purdue University, West Lafayette, IN, where he is currently an Associate Professor. He is coauthor of a book, *An Introduction to Optimization* (New York: Wiley, 1996. His research interests include optimization, control, modeling, and learning.

Dr. Chong serves as an Associate Editor for the IEEE TRANSACTIONS ON AUTOMATIC CONTROL. He received a Faculty Early Career Development (CAREER) Award from the National Science Foundation in 1995.