

Discriminative Subspace Method for Minimum Error Pattern Recognition

Hideyuki WATANABE and Shigeru KATAGIRI

ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan

Tel.: +81-774-95-1383

Fax.: +81-774-95-1308

E-mail: watanabe@itl.atr.co.jp

Abstract

Subspace Method (SM) is one of fundamental frameworks for pattern recognition. In particular, its discriminative learning version, called Learning Subspace Method (LSM), has been shown quite useful in various applications. However, this important design method leaves much room for further analysis due to the lack of a link between LSM and the ultimate goal of pattern recognition, i.e. the minimum error situation. In this light, we investigate in this paper SM from the viewpoint of the Minimum Classification Error/Generalized Probabilistic Descent method (MCE/GPD). Applying MCE/GPD to SM, we formalize a new discriminative subspace method, called the Minimum Error Learning Subspace method (MELS), which enables one to directly pursue the minimum error recognition. This paper also provides a rigorous analysis of the MELS's learning mechanism as well as a comparison between the conventional LSM and MELS.

1 Introduction

Subspace Method (SM), especially its discriminative learning version called Learning Subspace Method (LSM), has been shown quite useful in a wide range of pattern recognitions because of its computational simplicity and robustness to statistical pattern variations [1, 2, 3]. However, due to the lack of rigorous analysis from the viewpoint of the Bayes decision theory, this valuable recognizer design method still leaves much room for further mathematical investigation.

In the meantime, the Minimum Classification Error/Generalized Probabilistic Descent method (MCE/GPD) has provided new, general math-

emational bases for designing pattern recognizers aimed at minimizing recognition errors, or in other words, achieving the *optimal* minimum error situation [4]. Actually, it has been shown that the widely-used Learning Vector Quantization (LVQ) is a simple version of the MCE/GPD-based distance classifier [5]. The analogy between LVQ and LSM naturally suggests an analysis of LSM from the MCE/GPD viewpoint.

In this paper we investigate how the SM's discriminative power can be increased in the MCE/GPD framework. We also formalize a new discriminative subspace method, named *Minimum Error Learning Subspace method (MELS)*. This paper provides detailed analysis of the learning nature of MELS and proves that this proposed method leads to at least a locally optimal recognition situation. A comparison between the conventional LSM and MELS is also discussed.

2 Conventional Subspace Methods

2.1 Pattern Recognition by Subspace Method

We consider the problem of classifying a d -dimensional input pattern $\mathbf{x} \in \mathcal{R}^d$ into one of the K classes $\{\mathbf{C}_s\}_{s=1}^K$ using the following decision rule:

$$\mathbf{C}(\mathbf{x}) = \mathbf{C}_i \quad \text{if } i = \arg \max_s g_s(\mathbf{x}), \quad (1)$$

where $g_s(\mathbf{x})$, called the *discriminant function*, represents the degree to which $\mathbf{x}$ belongs to $\mathbf{C}_s$ and $\mathbf{C}(\cdot)$ is a recognition operation. In the SM framework, an individual *class subspace* in the d -dimensional pattern space is designed for each class, and each input pattern $\mathbf{x}$ is classified as the class giving the maximum value of the orthogonal projection of $\mathbf{x}$. That is to say, the discriminant function of the s -th class ($s = 1, 2, \dots, K$) is defined as

$$g_s(\mathbf{x}; \mathbf{U}_s) = \|\mathbf{P}_{\mathbf{U}_s} \mathbf{x}\|^2, \quad (2)$$

$$\mathbf{U}_s = [\mathbf{u}_{s,1} \ \mathbf{u}_{s,2} \ \dots \ \mathbf{u}_{s,p_s}], \quad (3)$$

$$\mathbf{u}_{s,i} = [u_{s,1,i} \ u_{s,2,i} \ \dots \ u_{s,d,i}]^T \quad (i = 1, 2, \dots, p_s), \quad (4)$$

where $\|\cdot\|$ denotes the Euclidean norm, $\mathbf{U}_s$ is the $d \times p_s$ ($p_s < d$) full-rank *base matrix* whose column vectors span the s -th class subspace $\mathcal{V}_s$, and the $d \times d$ matrix $\mathbf{P}_{\mathbf{U}_s}$ is the (*orthogonal*) *projection matrix* onto $\mathcal{V}_s$ and is expressed as

$$\mathbf{P}_{\mathbf{U}_s} = \mathbf{U}_s (\mathbf{U}_s^T \mathbf{U}_s)^{-1} \mathbf{U}_s^T, \quad (5)$$

where the superscript T and $^{-1}$ denote the transpose and the inverse, respectively. Consequently, the discriminant function is rewritten as

$$g_s(\mathbf{x}; \mathbf{U}_s) = \mathbf{x}^T \mathbf{U}_s (\mathbf{U}_s^T \mathbf{U}_s)^{-1} \mathbf{U}_s^T \mathbf{x}, \quad (6)$$

and it turns out that the method of designing each class subspace $\{\mathcal{V}_s\}_{s=1}^K$, or base matrix $\{\mathbf{U}_s\}_{s=1}^K$, determines the quality of recognition decision.

2.2 A Brief Review of LSM

The most fundamental algorithm for designing subspaces is the *CLAFIC* method, which designs each class subspace by using Karhunen-Loève expansion [1]. Since it designs each class subspace independently without considering the influences of other competing classes, it is not necessarily the case that the *CLAFIC* method can reduce the recognition errors efficiently. To alleviate this inadequacy, the *Learning Subspace Method (LSM)* was developed by Kohonen [1]. In LSM, each subspace is trained according to the recognition result of each design pattern so that the projection onto its true class gets larger while that onto each different competing class decreases.

Let us consider the subspace design problem using a labeled training sample set $\Omega = \{\mathbf{x}_n\}$. Suppose that an input pattern $\mathbf{x}_t$ is selected randomly from Ω at the t -th iteration. In LSM, if $\mathbf{x}_t$ belongs to the k -th class, each orthonormalized base matrix $\mathbf{U}_s$ is adjusted according to the following rule:

$$\tilde{\mathbf{U}}_k^{(t)} = (\mathbf{I} + \mu_k \mathbf{x}_t \mathbf{x}_t^T) \mathbf{U}_k^{(t-1)} \quad (\text{the true class}), \quad (7)$$

$$\tilde{\mathbf{U}}_j^{(t)} = (\mathbf{I} - \mu_j \mathbf{x}_t \mathbf{x}_t^T) \mathbf{U}_j^{(t-1)} \quad (\text{the competing class}) \quad (8)$$

$$\left(j = \arg \max_{s, s \neq k} g_s(\mathbf{x}_t; \mathbf{U}_s^{(t-1)}) \right),$$

$$\mathbf{U}_s^{(t)} = \tilde{\mathbf{U}}_s^{(t)} \mathbf{T}_s^{(t)} \quad (s = k, j), \quad (9)$$

$$\mathbf{U}_i^{(t)} = \mathbf{U}_i^{(t-1)} \quad (i = 1, 2, \dots, K; i \neq k, j), \quad (10)$$

where $\mathbf{I}$ denotes the identity matrix, μ_k and μ_j are the positive real numbers called the learning coefficients, and the $p_s \times p_s$ matrix $\mathbf{T}_s^{(t)}$ represents the orthonormalization operator for the column vectors of $\tilde{\mathbf{U}}_s^{(t)}$. Detailed discussions of the properties of LSM are made in [1].

As mentioned in [1], the optimal (in the sense of minimum recognition errors) values of μ_k and μ_j are very difficult to find. Actually, LSM was vigorously investigated with regard to convergence properties. However, its capability to attain the minimum error situation has not yet been sufficiently analyzed. These values are thus usually specified heuristically, e.g. by performing certain preliminary experiments. Obviously, a mathematical method is needed to determine these learning coefficients so that one can directly pursue the minimum recognition error probability by using the above formula.

3 Discriminative Subspace Method for Minimum Error Recognition

3.1 Formalization of MELS

The learning mechanism of MCE/GPD is based on gradient search optimization. The correction vector at each iteration step can be derived by differentiating the loss function in the recognizer parameters. Detailed discussions on MCE/GPD are made in [4, 6].

The first step in the MELS formulation is to derive the first-order derivative of each discriminant function. Its result is summarized in the following theorem.

Theorem 1 *The first-order derivative of $g_s(\mathbf{x}; \mathbf{U}_s)$ with respect to $\mathbf{U}_s$ is given as*

$$\nabla_{\mathbf{U}_s} g_s(\mathbf{x}; \mathbf{U}_s) = 2\mathbf{P}_{\mathbf{U}_s}^\perp \mathbf{x} \mathbf{x}^T \mathbf{U}_s (\mathbf{U}_s^T \mathbf{U}_s)^{-1}, \quad (11)$$

where the matrix differentiation is defined as $\nabla_{\mathbf{A}} = (\nabla_{a_{i,j}})$ ($\mathbf{A} = (a_{i,j})$), and

$$\mathbf{P}_{\mathbf{U}_s}^\perp = \mathbf{I} - \mathbf{U}_s (\mathbf{U}_s^T \mathbf{U}_s)^{-1} \mathbf{U}_s^T, \quad (12)$$

which is an orthogonal projection operator onto the orthogonal complement of $\mathcal{V}_s = \text{range } \mathbf{U}_s$.

Proof. See Appendix A. $\square$

Next, by applying the MCE/GPD adjustment rule to the above derivatives and operating the orthonormalization, an MCE/GPD-based subspace design algorithm, i.e. the MELS method, is formulated as follows:

The MELS method

At the t -th iteration, if a randomly-selected training pattern $\mathbf{x}_t$ belongs to the k -th class, then

$$\tilde{\mathbf{U}}_s^{(t)} = \mathbf{U}_s^{(t-1)} - \varepsilon_t \nabla_{\mathbf{U}_s} \ell_k(\mathbf{x}_t; \mathbf{A}^{(t-1)}), \quad (13)$$

$$\mathbf{U}_s^{(t)} = \tilde{\mathbf{U}}_s^{(t)} \mathbf{T}_s^{(t)} \quad (s = 1, 2, \dots, K), \quad (14)$$

where

$$\nabla_{\mathbf{U}_s} \ell_k(\mathbf{x}_t; \mathbf{A}^{(t-1)}) = \ell'_k(\mathbf{x}_t; \mathbf{A}^{(t-1)}) \rho_{k,s}(\mathbf{x}_t; \mathbf{A}^{(t-1)}) \nabla_{\mathbf{U}_s} g_s(\mathbf{x}_t; \mathbf{U}_s^{(t-1)}), \quad (15)$$

$$\nabla_{\mathbf{U}_s} g_s(\mathbf{x}_t; \mathbf{U}_s^{(t-1)}) = 2\mathbf{P}_{\mathbf{U}_s^{(t-1)}}^\perp \mathbf{x}_t \mathbf{x}_t^T \mathbf{U}_s^{(t-1)}, \quad (16)$$

$$\mathbf{A}^{(t)} = \begin{bmatrix} \mathbf{U}_1^{(t)} & \mathbf{U}_2^{(t)} & \dots & \mathbf{U}_K^{(t)} \end{bmatrix} \in \mathcal{R}^{d \times (\sum_{s=1}^K p_s)}, \quad (17)$$

$\varepsilon_t > 0$, $\ell'_k(\mathbf{x}_t; \mathbf{A}^{(t-1)}) (> 0)$ denotes the derivative of the *loss function* [4], $\rho_{k,s}(\mathbf{x}_t; \mathbf{A}^{(t-1)})$ denotes the derivative of the *misclassification*

measure [4] ($\rho_{k,k} < 0$, $\rho_{k,s} > 0$ ($s \neq k$)), and each $\mathbf{T}_s^{(t)}$ denotes the orthonormalization matrix applied to $\tilde{\mathbf{U}}_s^{(t)}$.

If we take the L_∞ norm in the misclassification measure (see [4]), the MELS algorithm leads to its special form named MELS*:

The MELS* method

At the t -th iteration, if a randomly-selected training pattern $\mathbf{x}_t$ belongs to the k -th class, then

$$\tilde{\mathbf{U}}_k^{(t)} = \left(\mathbf{I} + 2\varepsilon_t \ell'_k(\mathbf{x}_t; \mathbf{A}^{(t-1)}) \mathbf{P}_{\mathbf{U}_k^{(t-1)}}^\perp \mathbf{x}_t \mathbf{x}_t^T \right) \mathbf{U}_k^{(t-1)}, \quad (18)$$

$$\begin{aligned} \tilde{\mathbf{U}}_j^{(t)} &= \left(\mathbf{I} - 2\varepsilon_t \ell'_k(\mathbf{x}_t; \mathbf{A}^{(t-1)}) \mathbf{P}_{\mathbf{U}_j^{(t-1)}}^\perp \mathbf{x}_t \mathbf{x}_t^T \right) \mathbf{U}_j^{(t-1)} \\ &\quad \left(j = \arg \max_{s, s \neq k} g_s(\mathbf{x}_t; \mathbf{U}_s^{(t-1)}) \right), \end{aligned} \quad (19)$$

$$\mathbf{U}_s^{(t)} = \tilde{\mathbf{U}}_s^{(t)} \mathbf{T}_s^{(t)} \quad (s = k, j), \quad (20)$$

$$\mathbf{U}_i^{(t)} = \mathbf{U}_i^{(t-1)} \quad (i = 1, 2, \dots, K; i \neq k, j). \quad (21)$$

It can be seen that this MELS* algorithm closely resembles Kohonen's LSM. The difference in their forms is that MELS*, unlike LSM, includes the projection matrix $\mathbf{P}_{\mathbf{U}_s}^\perp$.

3.2 Convergence Property of MELS

One may think it trivial that MELS can achieve the minimum recognition error situation in the same way as MCE/GPD because it is simply derived by applying MCE/GPD to the SM framework. However, the fact that MELS includes the orthonormalization operation, which was not used in the original differentiation-based GPD adjustment, seems to require a further investigation of the learning convergence.

We first introduce the following theorem.

Theorem 2 *Assume that $|\nabla_{u_{s,j,i}} \ell_k(\mathbf{x}_t; \mathbf{A}^{(t-1)})| < \infty$ ($s = 1, 2, \dots, K; j = 1, 2, \dots, d; i = 1, 2, \dots, p_s$). Then, if the Gram-Schmidt orthonormalization (GSO) is applied to the MELS parameter adjustment, the parameter sequence $\{\mathbf{U}_s^{(t)} \ (t = 1, 2, \dots)\}$ ($s = 1, 2, \dots, K$) obeys the following rule:*

$$\mathbf{U}_s^{(t)} = \mathbf{U}_s^{(t-1)} - \varepsilon_t \nabla_{\mathbf{U}_s} \ell_k(\mathbf{x}_t; \mathbf{A}^{(t-1)}) + O(\varepsilon_t^2). \quad (22)$$

Proof. See Appendix B. □

With this theorem and the theorems described in [6], we next verify that MELS can attain the minimum recognition error situation. Define

the following vector which corresponds to $\mathbf{A} = [\mathbf{U}_1 \mathbf{U}_2 \dots \mathbf{U}_K]$ one to one:

$$\boldsymbol{\lambda} = [u_{1,1,1} \dots u_{s,j,i} \dots u_{K,d,p_K}]^T \in \mathcal{R}^{d(\sum_{s=1}^K p_s)}. \quad (23)$$

Accordingly, the rule (22) can be expressed in terms of $\boldsymbol{\lambda}$ as

$$\boldsymbol{\lambda}^{(t)} = \boldsymbol{\lambda}^{(t-1)} - \varepsilon_t \nabla_{\boldsymbol{\lambda}} \ell_k(\mathbf{x}_t; \boldsymbol{\lambda}^{(t-1)}) + O(\varepsilon_t^2). \quad (24)$$

There exist a $d(\sum_{s=1}^K p_s)$ -dimensional vector $\mathbf{d}_t$ ($\|\mathbf{d}_t\| < \infty$) such that

$$\boldsymbol{\lambda}^{(t)} = \boldsymbol{\lambda}^{(t-1)} - \varepsilon_t \left(\nabla_{\boldsymbol{\lambda}} \ell_k(\mathbf{x}_t; \boldsymbol{\lambda}^{(t-1)}) - \varepsilon_t \mathbf{d}_t \right). \quad (25)$$

For simplicity, we define

$$\Delta \boldsymbol{\lambda}^{(t)} = \nabla_{\boldsymbol{\lambda}} \ell_k(\mathbf{x}_t; \boldsymbol{\lambda}^{(t-1)}) - \varepsilon_t \mathbf{d}_t. \quad (26)$$

Furthermore, it can be seen that the Hessian matrix exists,

$$\mathbf{H}_k(\mathbf{x}; \boldsymbol{\lambda}) = \nabla_{\boldsymbol{\lambda}}^2 \ell_k(\mathbf{x}; \boldsymbol{\lambda}), \quad (27)$$

since each orthonormalized matrix $\mathbf{U}_s^{(t)}$ is full-rank. Then, we reach the following theorem guaranteeing the MELS convergence:

Theorem 3 *Make the following assumptions: for all $t \in \{1, 2, \dots\}$,*

- a1) $|\nabla_{u_{s,j,i}} \ell_k(\mathbf{x}_t; \boldsymbol{\lambda}^{(t-1)})| < \infty$
 $(s = 1, 2, \dots, K; j = 1, 2, \dots, d; i = 1, 2, \dots, p_s),$
- a2) $\langle \Delta \boldsymbol{\lambda}^{(t)} | \mathbf{H}_k(\mathbf{x}_t; \boldsymbol{\lambda}^{(t-1)} - \theta_t \varepsilon_t \Delta \boldsymbol{\lambda}^{(t)}) \Delta \boldsymbol{\lambda}^{(t)} \rangle < \infty,$
where $\langle \cdot | \cdot \rangle$ denotes the inner product and $0 \leq \theta_t \leq 1$.

Then, if an infinite sequence of random observations $\mathbf{x}_t$ is presented for training and the parameter adjustment rule of (13,14) is utilized with a sequence ε_t that satisfies $\sum_{t=1}^{\infty} \varepsilon_t \rightarrow \infty$ and $\sum_{t=1}^{\infty} \varepsilon_t^2 < \infty$, then the parameter sequence $\{\boldsymbol{\lambda}^{(t)}; t = 1, 2, \dots\}$ according to MELS converges with probability one to a $\boldsymbol{\lambda}^$ which results in a local minimum of $L(\boldsymbol{\lambda})$ under the orthonormality constraints, where $L(\boldsymbol{\lambda})$, called the expected loss, is defined as the expectation of $\ell_k(\mathbf{x}; \boldsymbol{\lambda})$ over the whole pattern space $\mathcal{X}$ and approximates (arbitrarily closely) the recognition error probability in the case of the smoothed 0-1 loss function (see [4, 6]).*

Proof. Taking Taylor expansion,

$$\begin{aligned} \ell_k(\mathbf{x}_t; \boldsymbol{\lambda}^{(t)}) &= \ell_k(\mathbf{x}_t; \boldsymbol{\lambda}^{(t-1)}) - \varepsilon_t \|\nabla_{\boldsymbol{\lambda}} \ell_k(\mathbf{x}_t; \boldsymbol{\lambda}^{(t-1)})\|^2 \\ &\quad + \varepsilon_t^2 \langle \mathbf{d}_t | \nabla_{\boldsymbol{\lambda}} \ell_k(\mathbf{x}_t; \boldsymbol{\lambda}^{(t-1)}) \rangle \\ &\quad + \frac{1}{2} \varepsilon_t^2 \langle \Delta \boldsymbol{\lambda}^{(t)} | \mathbf{H}_k(\mathbf{x}_t; \boldsymbol{\lambda}^{(t-1)} - \theta_t \varepsilon_t \Delta \boldsymbol{\lambda}^{(t)}) \Delta \boldsymbol{\lambda}^{(t)} \rangle, \end{aligned} \quad (28)$$

where the last term denotes the Lagrange's remainder term. By the Cauchy-Schwarz inequality,

$$|\langle \mathbf{d}_t | \nabla_{\boldsymbol{\lambda}} \ell_k(\mathbf{x}_t; \boldsymbol{\lambda}^{(t-1)}) \rangle| \leq \|\mathbf{d}_t\| \cdot \|\nabla_{\boldsymbol{\lambda}} \ell_k(\mathbf{x}_t; \boldsymbol{\lambda}^{(t-1)})\| < \infty. \quad (29)$$

Then, the same proof as in [6] can be applied. $\square$

4 Conclusions

We have studied the discriminative learning nature of SM from the viewpoint of minimizing recognition errors. Based on MCE/GPD, we have also formalized a new discriminative SM algorithm called the Minimum Error Learning Subspace method (MELS). MELS is a slightly different version of the conventional LSM and has been proven to achieve the minimum error situation. This result significantly increases the SM's applicability by providing the mathematical guarantee of training optimality.

Appendix

A Proof of Theorem 1

First we show the following formulae related to matrix differentiation without proofs: with proper-sized matrices $\mathbf{C}$ and $\mathbf{X}$,

$$\nabla_{\mathbf{X}} \text{tr}(\mathbf{C}\mathbf{X}) = \nabla_{\mathbf{X}} \text{tr}(\mathbf{X}\mathbf{C}) = \mathbf{C}^T, \quad (30)$$

$$\nabla_{\mathbf{X}} \text{tr}(\mathbf{C}\mathbf{X}^T) = \nabla_{\mathbf{X}} \text{tr}(\mathbf{X}^T \mathbf{C}) = \mathbf{C}, \quad (31)$$

where tr denotes the trace operator; if a matrix $\mathbf{C}(t)$ is non-singular and is parametrized by a scalar t ,

$$\frac{d}{dt} \mathbf{C}(t)^{-1} = -\mathbf{C}(t)^{-1} \frac{d\mathbf{C}(t)}{dt} \mathbf{C}(t)^{-1}. \quad (32)$$

The orthogonal projection defined in (6) can be rewritten as

$$g(\mathbf{x}; \mathbf{U}) = \text{tr} \left[\mathbf{U}(\mathbf{U}^T \mathbf{U})^{-1} \mathbf{U}^T \mathbf{x} \mathbf{x}^T \right]. \quad (33)$$

Hearafter we omit the suffix s for simplicity. Then, using the above-listed formulae, it can be easily shown that the derivative of $g(\mathbf{x}; \mathbf{U})$ with respect to $\mathbf{U}$ is expressed as

$$\nabla_{\mathbf{U}} g(\mathbf{x}; \mathbf{U}) = 2\mathbf{x} \mathbf{x}^T \mathbf{U}(\mathbf{U}^T \mathbf{U})^{-1} + \{\nabla_{\mathbf{V}} F(\mathbf{V})\}_{\mathbf{V}=\mathbf{U}}, \quad (34)$$

$$F(\mathbf{V}) = \text{tr}(\mathbf{A}\mathbf{Y}), \quad (35)$$

$$\mathbf{A} = (\mathbf{V}^T \mathbf{V})^{-1}, \quad (36)$$

$$\mathbf{Y} = \mathbf{U}^T \mathbf{x} \mathbf{x}^T \mathbf{U}. \quad (37)$$

Let $v_{i,j}$ be the (i, j) entry of the matrix $\mathbf{V}$ and $a_{m,n}$ be the (m, n) entry of $\mathbf{A}$. Considering that $\mathbf{A}$ is a function of $\mathbf{V}$, the partial derivative of F with respect to each entry $v_{i,j}$ is derived using a chain rule as

$$\frac{\partial F(\mathbf{V})}{\partial v_{i,j}} = \sum_{m=1}^p \sum_{n=1}^p \frac{\partial F}{\partial a_{m,n}} \frac{\partial a_{m,n}}{\partial v_{i,j}} = \text{tr} \left[\left(\frac{\partial F}{\partial \mathbf{A}} \right)^T \frac{\partial \mathbf{A}}{\partial v_{i,j}} \right]. \quad (38)$$

Using the formulae (30) and (32),

$$\frac{\partial F}{\partial \mathbf{A}} = \mathbf{Y}^T, \quad (39)$$

$$\frac{\partial \mathbf{A}}{\partial v_{i,j}} = -(\mathbf{V}^T \mathbf{V})^{-1} \left(\frac{\partial(\mathbf{V}^T \mathbf{V})}{\partial v_{i,j}} \right) (\mathbf{V}^T \mathbf{V})^{-1}. \quad (40)$$

Thus, using above equations, the following equation can be derived:

$$\begin{aligned} \frac{\partial F(\mathbf{V})}{\partial v_{i,j}} &= -\text{tr} \left[\mathbf{Y}(\mathbf{V}^T \mathbf{V})^{-1} \left(\frac{\partial(\mathbf{V}^T \mathbf{V})}{\partial v_{i,j}} \right) (\mathbf{V}^T \mathbf{V})^{-1} \right] \\ &= -\left\{ \frac{\partial}{\partial w_{i,j}} \text{tr} \left[(\mathbf{W}^T \mathbf{W})(\mathbf{V}^T \mathbf{V})^{-1} \mathbf{Y}(\mathbf{V}^T \mathbf{V})^{-1} \right] \right\}_{w_{i,j}=v_{i,j}} \end{aligned} \quad (41)$$

where $w_{i,j}$ is the (i, j) entry of $\mathbf{W}$. Expressing the above equation in a matrix form and using the formulae (30,31), we can get

$$\begin{aligned} \nabla_{\mathbf{V}} F(\mathbf{V}) &= -\left\{ \frac{\partial}{\partial \mathbf{W}} \text{tr} \left[(\mathbf{W}^T \mathbf{W})(\mathbf{V}^T \mathbf{V})^{-1} \mathbf{Y}(\mathbf{V}^T \mathbf{V})^{-1} \right] \right\}_{\mathbf{W}=\mathbf{V}} \\ &= -2\mathbf{V}(\mathbf{V}^T \mathbf{V})^{-1} \mathbf{Y}(\mathbf{V}^T \mathbf{V})^{-1}. \end{aligned} \quad (42)$$

Consequently, substituting the above equation into (34), we can get

$$\begin{aligned} \nabla_U g(\mathbf{x}; \mathbf{U}) &= 2\mathbf{x}\mathbf{x}^T \mathbf{U}(\mathbf{U}^T \mathbf{U})^{-1} - 2\mathbf{U}(\mathbf{U}^T \mathbf{U})^{-1} \mathbf{U}^T \mathbf{x}\mathbf{x}^T \mathbf{U}(\mathbf{U}^T \mathbf{U})^{-1} \\ &= 2\mathbf{P}_{\mathbf{U}}^\perp \mathbf{x}\mathbf{x}^T \mathbf{U}(\mathbf{U}^T \mathbf{U})^{-1}. \end{aligned} \quad (43)$$

This completes the proof.

B Proof of Theorem 2

The equation (22) is rewritten in terms of column vectors as

$$\begin{aligned} \mathbf{u}_{s,i}^{(t)} &= \mathbf{u}_{s,i}^{(t-1)} - \varepsilon_t \nabla_{s,i} \ell_k(\mathbf{x}_t; \mathbf{A}^{(t-1)}) + O(\varepsilon_t^2) \\ (s &= 1, 2, \dots, K; i = 1, 2, \dots, p_s), \end{aligned} \quad (44)$$

where

$$\nabla_{s,i} \ell_k(\mathbf{x}; \mathbf{A}) = \nabla_{\mathbf{u}_{s,i}} \ell_k(\mathbf{x}; \mathbf{A}) \quad (45)$$

$$= \mathbf{P}_{\mathbf{U}_s}^\perp (2\ell'_k(\mathbf{x}; \mathbf{A}) \rho_{k,s}(\mathbf{x}; \mathbf{A}) \mathbf{x}\mathbf{x}^T) \mathbf{u}_{s,i}. \quad (46)$$

Before getting into the proof, let us state the following lemma.

Lemma 1 *In each class C_s ($s = 1, 2, \dots, K$), every base vector $\{\mathbf{u}_{s,i}\}_{i=1}^{p_s}$ is orthogonal to every gradient vector $\{\nabla_{s,i} \ell_k(\mathbf{x}; \mathbf{A})\}_{i=1}^{p_s}$.*

Proof. It can be easily verified by (46). $\square$

The proof of Theorem 2 (or (44)) follows the mathematical induction. First we give the proof in the case $i = 1$. Since the first base vector $\tilde{\mathbf{u}}_{s,1}$ is merely normalized by the GSO,

$$\tilde{\mathbf{u}}_{s,1}^{(t)} = \mathbf{u}_{s,1}^{(t-1)} - \varepsilon_t \nabla_{s,1} \ell_k(\mathbf{x}_t; \mathbf{A}^{(t-1)}), \quad (47)$$

$$\mathbf{u}_{s,1}^{(t)} = \frac{\tilde{\mathbf{u}}_{s,1}^{(t)}}{\|\tilde{\mathbf{u}}_{s,1}^{(t)}\|}. \quad (48)$$

From (47) we can get

$$\|\tilde{\mathbf{u}}_{s,1}^{(t)}\|^2 = 1 + \varepsilon_t^2 \|\nabla_{s,1} \ell_k(\mathbf{x}_t; \mathbf{A}^{(t-1)})\|^2 \quad (49)$$

because of the orthonormality of $\mathbf{U}_s^{(t-1)}$ and Lemma 1. Therefore,

$$\frac{1}{\|\tilde{\mathbf{u}}_{s,1}^{(t)}\|} = 1 - \frac{1}{2} \varepsilon_t^2 M + (1 + \varepsilon_t^2 M)^{-\frac{1}{2}} - \left(1 - \frac{1}{2} \varepsilon_t^2 M\right), \quad (50)$$

where $M = \|\nabla_{s,1} \ell_k(\mathbf{x}_t; \mathbf{A}^{(t-1)})\|^2$. Then, by (48) and (50), and because of the boundedness of $\|\nabla_{s,1} \ell_k(\mathbf{x}_t; \mathbf{A}^{(t-1)})\|$, $\mathbf{u}_{s,1}^{(t)}$ becomes

$$\mathbf{u}_{s,1}^{(t)} = \tilde{\mathbf{u}}_{s,1}^{(t)} + O(\varepsilon_t^2) + \varepsilon_t^2 \gamma_{s,1}^{(t)} \tilde{\mathbf{u}}_{s,1}^{(t)}, \quad (51)$$

$$\gamma_{s,1}^{(t)} = \frac{F(\varepsilon_t^2)}{\varepsilon_t^2}, \quad (52)$$

$$F(\varepsilon_t^2) = (1 + \varepsilon_t^2 M)^{-\frac{1}{2}} - \left(1 - \frac{1}{2} \varepsilon_t^2 M\right). \quad (53)$$

Lemma 2 $\lim_{\varepsilon_t^2 \rightarrow 0} \gamma_{s,1}^{(t)} = 0$.

Proof. Since $\lim_{\varepsilon_t^2 \rightarrow 0} F(\varepsilon_t^2) = 0$, and by de l'Hospital's theorem,

$$\begin{aligned} \lim_{\varepsilon_t^2 \rightarrow 0} \gamma_{s,1}^{(t)} &= \lim_{\varepsilon_t^2 \rightarrow 0} \frac{dF(\varepsilon_t^2)/d(\varepsilon_t^2)}{d(\varepsilon_t^2)/d(\varepsilon_t^2)} \\ &= \lim_{\varepsilon_t^2 \rightarrow 0} \left(-\frac{1}{2} (1 + \varepsilon_t^2 M)^{-\frac{3}{2}} + \frac{1}{2} \right) M = 0. \quad \square \end{aligned}$$

Accordingly, considering Lemma 2, we can say that

$$\mathbf{u}_{s,1}^{(t)} = \mathbf{u}_{s,1}^{(t-1)} - \varepsilon_t \nabla_{s,1} \ell_k(\mathbf{x}_t; \mathbf{A}^{(t-1)}) + O(\varepsilon_t^2). \quad (54)$$

Next, *suppose* that in each case $i = 1, 2, \dots, j$ the following equation holds:

$$\mathbf{u}_{s,i}^{(t)} = \mathbf{u}_{s,i}^{(t-1)} - \varepsilon_t \nabla_{s,i} \ell_k(\mathbf{x}_t; \mathbf{A}^{(t-1)}) + O(\varepsilon_t^2), \quad (55)$$

and consider the case $i = j + 1$. According to the GSO, the $(j + 1)$ -th normalized base vector $\mathbf{u}_{s,j+1}^{(t)}$ is derived from $\{\mathbf{u}_{s,i}^{(t)}\}_{i=1}^j$ as

$$\mathbf{v}_{s,j+1}^{(t)} = \left(\mathbf{I} - \sum_{i=1}^j \mathbf{u}_{s,i}^{(t)} \mathbf{u}_{s,i}^{(t)T} \right) \tilde{\mathbf{u}}_{s,j+1}^{(t)}, \quad (56)$$

$$\mathbf{u}_{s,j+1}^{(t)} = \frac{\mathbf{v}_{s,j+1}^{(t)}}{\|\mathbf{v}_{s,j+1}^{(t)}\|}. \quad (57)$$

Note that

$$\tilde{\mathbf{u}}_{s,j+1}^{(t)} = \mathbf{u}_{s,j+1}^{(t-1)} - \varepsilon_t \nabla_{s,j+1} \ell_k(\mathbf{x}_t; \mathbf{A}^{(t-1)}). \quad (58)$$

From (55) and (58) we can get ($i = 1, 2, \dots, j$)

$$\mathbf{u}_{s,i}^{(t)T} \tilde{\mathbf{u}}_{s,j+1}^{(t)} = O(\varepsilon_t^2) \quad (59)$$

because of the orthonormality of $\mathbf{U}_s^{(t-1)}$ and Lemma 1. Therefore, from (56), (58) and (59),

$$\mathbf{v}_{s,j+1}^{(t)} = \mathbf{u}_{s,j+1}^{(t-1)} - \varepsilon_t \nabla_{s,j+1} \ell_k(\mathbf{x}_t; \mathbf{A}^{(t-1)}) + O(\varepsilon_t^2), \quad (60)$$

and accordingly we can get

$$\|\mathbf{v}_{s,j+1}^{(t)}\|^2 = 1 + \varepsilon_t^2 \|\nabla_{s,j+1} \ell_k(\mathbf{x}_t; \mathbf{A}^{(t-1)})\|^2 + O(\varepsilon_t^2) \quad (61)$$

because of the orthonormality of $\mathbf{U}_s^{(t-1)}$ and Lemma 1. Then, according to the logic analogous to the case $i = 1$, we reach

$$\mathbf{u}_{s,j+1}^{(t)} = \mathbf{u}_{s,j+1}^{(t-1)} - \varepsilon_t \nabla_{s,j+1} \ell_k(\mathbf{x}_t; \mathbf{A}^{(t-1)}) + O(\varepsilon_t^2). \quad (62)$$

This completes the proof according to the mathematical induction.

References

- [1] E. Oja, *Subspace Methods of Pattern Recognition*, Research Studies Press, 1983.
- [2] Y. Ariki and K. Doi, "Speaker recognition based on subspace methods", *Proc. ICSLP 94*, Yokohama, pp.1859-1862, Sep. 1994.
- [3] Y. Takebayashi, "Speech recognition based on the subspace method: AI class-description learning viewpoint", *J. Acoust. Soc. Jpn. (E)*, vol. 13, No. 6, pp.429-439, Nov. 1992.
- [4] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification", *IEEE Trans. Signal Processing*, vol. 40, No. 12, pp. 3043-3054, Dec. 1992.
- [5] S. Katagiri, C.-H. Lee, and B.-H. Juang, "New discriminative training algorithms based on the generalized probabilistic descent method", in *Proc. 1991 IEEE Workshop on Neural Networks for Signal Processing*, pp. 299-308, Princeton, NJ, Sept. 1991.
- [6] W. Chou and B.-H. Juang, "Adaptive discriminative learning in pattern recognition", Technical Report of AT&T Bell Laboratory.