Ratio Trace Formulation of Wasserstein Discriminant Analysis

Hexuan Liu, Yunfeng Cai, You-Lin Chen, Ping Li
Cognitive Computing Lab
Baidu Research
No.10 Xibeiwang East Road, Beijing 100193, China
10900 NE 8th St. Bellevue, Washington 98004, USA
{lhxuan93, yunfengcai09, cyoulin.tw, pingli98}@gmail.com

Abstract

We reformulate the Wasserstein Discriminant Analysis (WDA) as a ratio trace problem and present an eigensolver-based algorithm to compute the discriminative subspace of WDA. This new formulation, along with the proposed algorithm, can be served as an efficient and more stable alternative to the original trace ratio formulation and its gradient-based algorithm. We provide a rigorous convergence analysis for the proposed algorithm under the self-consistent field framework, which is crucial but missing in the literature. As an application, we combine WDA with low-dimensional clustering techniques, such as K-means, to perform subspace clustering. Numerical experiments on real datasets show promising results of the ratio trace formulation of WDA in both classification and clustering tasks.

1 Introduction

Wasserstein Discriminant Analysis (WDA) [13] is a supervised linear dimensionality reduction technique that generalizes the classical Fisher Discriminant Analysis (FDA) [16] using the optimal transport distances [41]. Many existing works [44, 29, 11, 4] have addressed the issue that FDA only considers global information. In particular, [49] proposed a new formula relaying on worst-case distance; [37] developed a localized version of FDA; [22] provided an adaptive method for learning local structure from data. The recently proposed WDA [13] has the advantage of adaptively capturing both local and global information, and shows competitive performance in classification tasks compared to other supervised dimensionality reduction techniques.

WDA as developed in [13] used the trace ratio formulation to maximize the ratio of the inter-class's regularized Wasserstein distances to the intra-class's regularized Wasserstein distances. Formally, they aimed to solve $\max_{\mathbf{P}} \text{Trace}(\mathbf{P}^T\mathbf{C}_b(\mathbf{T})\mathbf{P})/\text{Trace}(\mathbf{P}^T\mathbf{C}_w(\mathbf{T})\mathbf{P})$ where $\mathbf{C}_b$ and $\mathbf{C}_w$ are the inter-class and intra-class covariance matrices, respectively, and are functions of the optimal transport matrix $\mathbf{T}$. The optimal transport matrix $\mathbf{T}$ quantifies how important the distance between two samples should be in order to obtain a good projection matrix $\mathbf{P}$. The authors in [13] derived the gradient of the objective function with respect to $\mathbf{P}$ and also utilized automatic differentiation to compute the gradients. The difficulties of their approach are 1) the optimization objective is non-convex and non-smooth; and 2) $\mathbf{C}_b$ and $\mathbf{C}_w$ are functions of $\mathbf{T}$ and $\mathbf{T}$ is an implicit function on $\mathbf{P}$. Thus WDA is a bi-level optimization problem [8] and requires solving an optimal transport problem in every step of gradient descent. Due to these complications, theoretical guarantees on the convergence are lacking. Vanilla gradient descent gets stuck easily in the non-smooth region, especially for real datasets, due to the natural structure of the data such as low rank or sparsity. In practice, the approach introduced in [13] can be sensitive to initialization and may take many iterations or even fail to reach convergence. All these issues raise concerns when WDA is applied to real data.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

In this paper, we circumvent the aforementioned challenges by reformulating WDA as a ratio trace problem, which has a closed-form solution and can be solved by the generalized eigenvalue decomposition if $\mathbf{T}$ is given. For algorithms of dimensionality reduction, it is common to use ratio trace formulation to approximate trace ratio problems [46]. For example, in Fisher Discriminant Analysis (FDA), these two formulations are both defined and are both served as criterion to maximize inter-class distance while minimizing intra-class distance [15]. Although there are many comparisons between these two formulations when the inter-class and intra-class covariance matrices are fixed [43, 28, 17, 27], they do not concern with the case of the covariance matrices being functions of the discriminative subspace as in WDA. We give numerical comparisons between these two formulations in terms of classification accuracy on simulated as well as real data, on which the proposed formulation either is comparable or outperforms the original formulation.

Specifically, we solve the ratio trace problem: $\operatorname{argmax}_{\mathbf{P}} \operatorname{Trace}(\mathbf{P}^T\mathbf{C}_b(\mathbf{T})\mathbf{P}(\mathbf{P}^T\mathbf{C}_w(\mathbf{T})\mathbf{P})^{-1})$ instead of the original WDA formulation: $\operatorname{argmax}_{\mathbf{P}} \operatorname{Trace}(\mathbf{P}^T\mathbf{C}_b(\mathbf{T})\mathbf{P})/\operatorname{Trace}(\mathbf{P}^T\mathbf{C}_w(\mathbf{T})\mathbf{P})$. We propose an algorithm: **WDA-eig**, to solve the ratio trace problem using the self-consistent field iteration (SCF), and establish a convergence analysis for the general SCF framework with specific application to the WDA context. The SCF iteration was originally used for solving Kohn-Sham equation arising in electronic structure calculations [7]. Most works on SCF concern with the standard eigenvalue problem [47, 24, 6], while convergence analysis for the generalized eigenvalue problem has not appeared in current literature. Our numerical examples demonstrate that the algorithm based on SCF iteration usually converges within a few iterations in practice and is less sensitive to initialization compared to the original approach. We also give a convergent analysis under the SCF framework, which not only provides convergence guarantee to the ratio trace WDA problem but is also applicable to other eigenvector-dependent generalized eigenvalue problem.

As an application, we extend **WDA-eig** to unsupervised clustering. Since WDA requires class labels to calculate the inter- and intra-class Wasserstein distances, a natural solution is to combine WDA with low-dimensional clustering techniques, which requires iteratively applying WDA given updated label information. The new algorithm has a fast convergence compared to the original approach and aid in iteratively applying WDA to find the most discriminative subspace. Several methods [10, 48, 42] that are closely related to our work leverage label information by combining FDA with Kmeans. Our numerical experiments show that the WDA-Kmeans has promising performance compared to existing subspace clustering techniques on real-world datasets.

Our contribution in this paper is three-fold. First, we present a ratio trace formulation of the WDA problem. Second, we propose to solve the problem using the SCF iteration, and provide a convergent analysis for the SCF framework as well as specific application to the WDA context. Last but not least, we iteratively apply WDA and low-dimensional clustering technique to perform clustering. We emphasize that we do not attempt solving the original trace ratio formulation of WDA with the proposed algorithm. A better solution to the original formulation is not the focus of this paper.

Notations. We use $\|\cdot\|$ to denote the 2-norm of a matrix or vector. $\mathbf{I}_n$ is used to denote the identity matrix of order n. For any matrix $\mathbf{X}$, let x_i denote its ith column vector and $x_{i,j}$ denote the (i,j)th entry. For any $\mathbf{A}$, $\mathbf{B} \in \mathbb{R}^{m\times n}$, $\langle \mathbf{A}, \mathbf{B}\rangle$ is the inner product of $\mathbf{A}$ and $\mathbf{B}$, i.e., $\langle \mathbf{A}, \mathbf{B}\rangle = \operatorname{trace}(\mathbf{A}^T\mathbf{B})$. Let $\mathbb{S}^n = \{A \in \mathbb{R}^{n\times n} | A = A^T\}$ be the set of symmetric matrices. For a symmetric matrix pair $(\mathbf{A}, \mathbf{B})$, $\mathbf{A}, \mathbf{B} \in \mathbb{S}^n$ with $\mathbf{B}$ being positive definite, we denote the generalized eigenvalues of $(\mathbf{A}, \mathbf{B})$ by $\lambda_{\min}(\mathbf{A}, \mathbf{B}) = \lambda_n(\mathbf{A}, \mathbf{B}) \leq \cdots \leq \lambda_1(\mathbf{A}, \mathbf{B}) = \lambda_{\max}(\mathbf{A}, \mathbf{B})$. Let $\mathbb{O}^{d\times p}$ represent the set of orthogonal $d \times p$ matrices, i.e., $\mathbb{O}^{d\times p} = \{\mathbf{A} \in \mathbb{R}^{d\times p} \mid \mathbf{A}^T\mathbf{A} = \mathbf{I}_d\}$.

2 Methodology

In this section, we first review the existing supervised WDA problem and its gradient-based solver, and reformulate the problem as a nonlinear generalized eigenvalue problem. We then present an algorithm that solves the problem using the self-consistent field iteration.

2.1 Background

Wasserstein distance (also known as the optimal transport distance, earth mover distance) is a distance between probability measures that preserves the underlying geometry of the space based on principles from the optimal transport theory [41]. The regularized Wasserstein distance is the solution of the following entropy-smoothed optimal transport problem:

$$\mathbf{T}_\lambda \triangleq \operatorname*{argmin}_{\mathbf{T}\in\mathbb{U}_{nm}} \lambda\langle \mathbf{T}, \mathbf{M}_{\mathbf{X},\mathbf{Z}}\rangle - \Omega(\mathbf{T}), \tag{1}$$

where $\lambda \geq 0$ is the Wasserstein regularization parameter, $\mathbf{M}_{\mathbf{X},\mathbf{Z}}$ denotes the pairwise squared Euclidean distance matrix between samples in $\mathbf{X} \in \mathbb{R}^{n\times d}$ and $\mathbf{Z} \in \mathbb{R}^{n\times d}$: $\mathbf{M}_{\mathbf{X},\mathbf{Z}} \triangleq [\|x_i - z_j\|_2^2]$, and $\Omega(\mathbf{T})$ is the entropy of $\mathbf{T}$: $\Omega(\mathbf{T}) \triangleq -\sum_{ij} t_{ij}\log(t_{ij})$. $\mathbb{U}_{mn}$ is the polytope of $m \times n$ nonnegative matrices with row and column sums being equal to $\mathbf{1}_m/m$ and $\mathbf{1}_n/n$ respectively: $\mathbb{U}_{mn} \triangleq \{\mathbf{T} \in \mathbb{R}_+^{m\times n} \mid \mathbf{T}\mathbf{1}_n = \mathbf{1}_m/m, \mathbf{T}^T\mathbf{1}_m = \mathbf{1}_n/n\}$.

As the entropy-smoothed optimal transport problem is strictly convex, the solution to (1) exists and is unique. Numerically, $\mathbf{T}_\lambda$ can be obtained very efficiently using algorithms such as the Sinkhorn's fixed-point iterations [18, 9], the Greenkhorn algorithm [2, 1], or APDAMD [23]. The regularization parameter λ can be used to balance the global (at the distribution scale) and the local (at the samples' scale) interactions between different classes.

The original Wasserstein Discriminant Analysis solves the following bi-level optimization problem:

$$\begin{aligned} &\max_{\mathbf{P}\in\mathbb{O}^{d\times p}} J(\mathbf{P}, \mathbf{T}(\mathbf{P})) = \frac{\sum_{c,c'>c}\langle \mathbf{P}\mathbf{P}^T, \mathbf{C}^{c,c'}\rangle}{\sum_c \langle \mathbf{P}\mathbf{P}^T, \mathbf{C}^{c,c}\rangle} = \frac{\langle \mathbf{P}\mathbf{P}^T, \mathbf{C}_b\rangle}{\langle \mathbf{P}\mathbf{P}^T, \mathbf{C}_w\rangle}, \\ &\text{where } \mathbf{C}^{c,c'} = \sum_{i,j} t_{i,j}^{c,c'}(x_i^c - x_j^{c'})(x_i^c - x_j^{c'})^T, \quad \forall c, c', \\ &\text{s.t. } \mathbf{T}^{c,c'} = \arg\min_{\mathbf{T}\in U_{n_c n_{c'}}} \lambda\langle \mathbf{T}, \mathbf{M}_{\mathbf{X}^c\mathbf{P},\mathbf{X}^{c'}\mathbf{P}}\rangle - \Omega(\mathbf{T}), \end{aligned} \tag{2}$$

where $\mathbf{X}^c \in \mathbb{R}^{n_c\times d}$ is the data matrix of the samples from class c, and $\mathbf{X}^c\mathbf{P}$ is the matrix of projected samples from class c. $\mathbf{C}_b = \sum_{c,c'>c}\mathbf{C}^{c,c'}$ and $\mathbf{C}_w = \sum_c \mathbf{C}^{c,c}$ are the between and within cross-covariance matrices, and they both depend on $\mathbf{T}(\mathbf{P})$.

In [13], the gradient $G^k = \nabla_{\mathbf{P}} J(\mathbf{P}, \mathbf{T}(\mathbf{P}))$ at iteration k was computed using automatic differentiation [25], and the optimization problem is solved using pymanopt solvers such as the projected gradient descent and trust region methods on the Stiefel manifold [3, 39]. In practice, due to the complication of the problem formulation and the structures of data, the gradient-based approach often has a slow convergence and is sensitive to parameters and initialization. We will illustrate these difficulties in Section 4 with numerical experiments.

2.2 The Nonlinear Eigensolver-based Approach

For (2), once $\mathbf{T}^{c,c'}$ is computed, the problem becomes a trace ratio problem:

$$\max_{\mathbf{P}\in\mathbb{O}^{d\times p}} J(\mathbf{P}) = \frac{\text{Trace}(\mathbf{P}^T\mathbf{C}_b\mathbf{P})}{\text{Trace}(\mathbf{P}^T\mathbf{C}_w\mathbf{P})}, \tag{3}$$

where $\mathbf{C}_b$ and $\mathbf{C}_w$ depend on $\mathbf{P}$. We approximate the problem by solving a ratio trace problem:

$$\max_{\mathbf{P}\in\mathbb{R}^{d\times p}} J_{rt}(\mathbf{P}) = \text{Trace}((\mathbf{P}^T\mathbf{C}_b\mathbf{P})(\mathbf{P}^T\mathbf{C}_w\mathbf{P})^{-1}), \tag{4}$$

Problem (4) can be efficiently solved by the generalized eigenvalue decomposition:

$$\mathbf{C}_b(\mathbf{P})\mathbf{P} = \mathbf{C}_w(\mathbf{P})\mathbf{P}\Lambda, \tag{5}$$

where the optimal $\mathbf{P}$ is the matrix of eigenvectors corresponding to the p largest generalized eigenvalues. The generalized eigenvector-dependent nonlinear eigenvalue problem (which we refer to as NLEP from now on) can be solved via the self-consistent field (SCF) iteration [26, 32]: given $\mathbf{P}_{t-1}$, we first construct $\mathbf{C}_b(\mathbf{P}_{t-1})$ and $\mathbf{C}_w(\mathbf{P}_{t-1})$, then solve the generalized eigenvalue problem $\mathbf{C}_b(\mathbf{P}_{t-1})v = \mu\mathbf{C}_w(\mathbf{P}_{t-1})v$. Let v_j be the eigenvector corresponding to the jth largest generalized eigenvalue, then $\mathbf{P}_t$ is updated as an orthonormal basis for $[v_1, \ldots, v_p]$. Compared to the gradient-based approach, the new formulation with SCF iteration could drastically reduce the number of iterations. We therefore propose Algorithm 1 for solving supervised WDA.

Algorithm 1 WDA-eig algorithm

```
Input: De-meaned data $X$, class labels $\hat{y}$, initial subspace $\mathbf{P}_0 \in \mathbb{O}^{d\times p}$, tolerance $\epsilon$,
       maximum number of iterations $N$
for $k = 1$ to $N$ do
    for each pair of classes $c$, $c'$ do
        Compute $\mathbf{T}^{c,c'}(\mathbf{P}_{k-1})$ using the Sinkhorn iteration
    end for
    Construct $\mathbf{C}_b(\mathbf{P}_{k-1})$ and $\mathbf{C}_w(\mathbf{P}_{k-1})$
    Compute the generalized eigenvalue problem: $\mathbf{C}_b(\mathbf{P}_{k-1})\mathbf{P} = \mathbf{C}_w(\mathbf{P}_{k-1})\mathbf{P}\Lambda$, and obtain
        $\mathbf{P}_k \in \mathbb{O}^{d\times p}$ as an orthonormal basis for the eigenvector matrix corresponding to the
        $p$ largest generalized eigenvalues
    if the change in $\mathbf{P}_k$ is sufficiently small then
        Break
    end if
end for
```

From a computational complexity point of view, suppose that for each class there are n samples and d features. For each SCF iteration, complexity is dominated by constructing $\mathbf{C}_b$ and $\mathbf{C}_w$, which are $\mathcal{O}(n^2d^2)$. Solving the generalized eigenvalue problem has complexity $\mathcal{O}(d^3)$, but it is possible to only run a few iteration to reach certain tolerance. Each Sinkhorn iteration is of $\mathcal{O}(n^2)$ and we run a fixed number of iterations. The memory complexity is $\mathcal{O}(d^2)$ by storing the matrices $\mathbf{C}_b$ and $\mathbf{C}_w$.

Note that it is also interesting to investigate if the Riemannian optimization can be applied to Problem (4), similarly to [45]. This is, however, nontrivial as the constraint manifold varies with iterations in this case. We leave it to future work. Another line of interesting future work is to develop kernelized version of **WDA-eig** and randomized algorithms to speed up the computations [20, 21]. Moreover, in the future one can also re-visit **WDA-eig** by considering sparsity constraints [5].

3 Analysis

In this section we first give a convergence analysis for the SCF framework for solving generalized NLEP, followed by an analysis for the proposed **WDA-eig** in Algorithm 1.

3.1 Convergence of SCF

Consider the generalized NLEP $A(\mathbf{P})\mathbf{V} = B(\mathbf{P})\mathbf{V}\Lambda$, where $\mathbf{V} = [v_1, \ldots, v_p]$ and $\mathbf{P}$ is an orthonormal basis of $\mathbf{V}$ that spans the same subspace as $\mathbf{V}$. $A(\mathbf{P})$, $B(\mathbf{P})$ are symmetric matrix-valued function and $B(\mathbf{P})$ is positive definite. $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$, where $\lambda_1 \geq \cdots \geq \lambda_p$ are the p largest eigenvalues of $(A(\mathbf{P}), B(\mathbf{P}))$ corresponding to eigenvectors $v_1, \ldots, v_p$. We emphasize that $A(\mathbf{P})$, $B(\mathbf{P})$ are invariant to orthogonal transformation of $\mathbf{P}$, i.e., $A(\mathbf{P}) \equiv A(\mathbf{P}Q)$, $B(\mathbf{P}) \equiv B(\mathbf{P}Q)$ for any orthogonal matrix $Q \in \mathbb{R}^{p\times p}$.

Definitions. Let $\mathcal{X}$ and $\mathcal{Y}$ be two p-dimensional subspaces of $\mathbb{R}^n$. Let the columns of X form an orthonormal basis for $\mathcal{X}$ and the columns of Y form an orthonormal basis for $\mathcal{Y}$. We use $\|\sin\Theta(\mathcal{X},\mathcal{Y})\|$ as in [35] to measure the distance between $\mathcal{X}$ and $\mathcal{Y}$, where

$$\Theta(\mathcal{X},\mathcal{Y}) = \operatorname{diag}(\theta_1(\mathcal{X},\mathcal{Y}), \ldots, \theta_p(\mathcal{X},\mathcal{Y})). \tag{6}$$

Here, $\theta_j(\mathcal{X},\mathcal{Y})$'s denote the *canonical angles* between $\mathcal{X}$ and $\mathcal{Y}$ [p. 43][35], which is defined as

$$0 \leq \theta_j(\mathcal{X},\mathcal{Y}) \triangleq \arccos\sigma_j \leq \frac{\pi}{2} \quad \text{for } 1 \leq j \leq k, \tag{7}$$

where σ_j's are the singular values of X^TY. Similar to the Crawford number for symmetric definite matrix pair (A, B) [Chapter 8.7] [40], we define the Crawford number for the generalized NLEP as

$$c \triangleq \min_{\mathbf{P}\in\mathbb{O}^{d\times p}} \min_{x\in\mathbb{C}^d, \|x\|=1} (x^T(A(\mathbf{P}) + iB(\mathbf{P}))x),$$

where i is the imaginary unit. Define $C \triangleq \max_{\mathbf{P}\in\mathbb{O}^{d\times p}} \sqrt{\|A(\mathbf{P})^2 + B(\mathbf{P})^2\|}$. At the kth SCF iteration, one computes an approximation to the eigenvector matrix $\mathbf{V}_k$ associated with the p largest eigenvalues of $(A(\mathbf{P}_{k-1}), B(\mathbf{P}_{k-1}))$, where $\mathbf{P}_{k-1}$ is an orthonormal basis for $\mathbf{V}_{k-1}$, and then $\mathbf{V}_k$ is used as the next approximation to the solution. Let $A_k = A(\mathbf{P}_k)$, $B_k = B(\mathbf{P}_k)$, $\mu_{i,k} = \arctan \lambda_i(A(\mathbf{P}_k), B(\mathbf{P}_k))$. We also define $s_k \triangleq \|\sin\Theta(\mathbf{P}_k, \mathbf{P}_{k-1})\|$ as the distance between subspaces $\mathbf{P}_k$ and $\mathbf{P}_{k-1}$.

We study the convergence of SCF iteration under the following assumptions:

A1: For any $\mathbf{P}_1, \mathbf{P}_2 \in \mathbb{O}^{d\times p}$, assume that there exist positive constants ξ_a, ξ_b such that

$$\|A(\mathbf{P}_1) - A(\mathbf{P}_2)\| \le \xi_a \|\sin\Theta(\mathbf{P}_1, \mathbf{P}_2)\|, \qquad \|B(\mathbf{P}_1) - B(\mathbf{P}_2)\| \le \xi_b \|\sin\Theta(\mathbf{P}_1, \mathbf{P}_2)\|;$$

A2: For $k = 1, 2, \cdots$, there exists an $\eta > 0$ such that

$$\mu_{p,k} - \mu_{p+1,k} \ge \eta.$$

We state the convergence theorems below and give proofs in the supplementary material. By global convergence we mean that the algorithm converges to some stationary points [34] and does not guarantee convergence to a global optimum for all initial points. The algorithm converges when the change in subspace is sufficiently small, i.e., s_k is within some user-specified tolerance.

Theorem 1. ***(Global Convergence)*** *Let $s_1 = \|\sin\Theta(\boldsymbol{P}_0, \boldsymbol{P}_1)\|$. Assume **A1** and **A2**, and $s_1\sqrt{\xi_a^2 + \xi_b^2} < c$. If*

$$\eta > \arcsin(\rho C\sqrt{\xi_a^2 + \xi_b^2}/c^2) + \arctan(s_1\sqrt{\xi_a^2 + \xi_b^2}/c)$$

for some constant $\rho > 1$, then SCF converges linearly at the rate of $\frac{1}{\rho}$.

With relaxed assumption on the arctangent gap, we can show local convergence if the initial subspace is close enough to the true subspace $\mathbf{P}^*$:

A3: Let μ_i^* denote $\arctan \lambda_i(A(\mathbf{P}^*), B(\mathbf{P}^*))$. There exists an $\eta > 0$ such that

$$\mu_p^* - \mu_{p+1}^* \ge \eta.$$

Theorem 2. ***(Local Convergence)*** *Let $\hat{s}_0 = \|\sin\Theta(\boldsymbol{P}_0, \boldsymbol{P}^*)\|$. Assume **A1** and **A3**, and $\hat{s}_0\sqrt{\xi_a^2 + \xi_b^2} < c$. If*

$$\eta > \arcsin(\rho C\sqrt{\xi_a^2 + \xi_b^2}/c^2) + \arctan(\hat{s}_0\sqrt{\xi_a^2 + \xi_b^2}/c)$$

for some constant $\rho > 1$, then SCF is locally convergent at $\boldsymbol{P}^*$ at the rate of $\frac{1}{\rho}$.

Theorems 1 and 2 characterize how the eigenspace varies when the matrix pair undergoes a small perturbation. The sensitivity of the matrix pair as functions of $\mathbf{P}$ is quantified by the Lipschitz constants in A1. A2 and A3 are assumptions to guarantee that a discriminative subspace exists. In the following section we give more concrete examples in the WDA context for these assumptions.

3.2 Analysis for Supervised WDA

In the context of WDA, $A(\mathbf{P})$ is the inter-class covariance matrix $\mathbf{C}_b(\mathbf{P})$ and $B(\mathbf{P})$ is the intra-class covariance matrix $\mathbf{C}_w(\mathbf{P})$. For each iteration in **WDA-eig**, a fixed number of Sinkhorn iterations is computed to obtain an approximation to the optimal transport distance $\mathbf{T}$. $\mathbf{T}(\mathbf{P})$ can be expressed as an implicit function using the optimality conditions of the equation defining the optimal $\mathbf{T}$, and $\frac{\partial \mathbf{T}}{\partial \mathbf{P}}$ exists and is bounded. Therefore it is safe to assume that $\mathbf{T}$ is Lipschitz continuous in $\mathbf{P}$.

Corollary 1. *Suppose that the optimal transport matrix $\boldsymbol{T}^{c,c'}$ satisfies a Lipschitz-like condition:*

$$\|\boldsymbol{T}^{c,c'}(\boldsymbol{P}_1) - \boldsymbol{T}^{c,c'}(\boldsymbol{P}_2)\| \le \xi^{c,c'}\|\sin\Theta(\boldsymbol{P}_1, \boldsymbol{P}_2)\|,$$

For a given p, let

$$\eta = \min_k\{\eta_k | \eta_k = \mu_{p,k} - \mu_{p+1,k}\}.$$

Denote $\xi_a = \sum_{c,c'>c} \xi^{c,c'} \| \sum_{i,j} (x_i^c - x_j^{c'})(x_i^c - x_j^{c'})^T \|$, $\xi_b = \sum_c \xi^{c,c} \| \sum_{i,j} (x_i^c - x_j^c)(x_i^c - x_j^c)^T \|$. *If*

$$\eta > \arcsin(\rho C \sqrt{\xi_a^2 + \xi_b^2}/c^2) + \arctan(s_1 \sqrt{\xi_a^2 + \xi_b^2}/c)$$

for some constant $\rho > 1$, *then* ***wda-eig*** *converges linearly at the rate* $\frac{1}{\rho}$.

Corollary 1 implies that given a data matrix, the convergence rate of **WDA-eig** depends on the initialization, the subspace dimension p and ξ_a, ξ_b. ξ_a and ξ_b are functions of $\xi^{c,c'}$ and depends on the Wasserstein regularization parameter λ. When $\lambda = 0$, $t^{c,c'}$ is a constant matrix and $\xi^{c,c'} = 0$. For a fixed λ, the arctangent gap η depends on the inherent structure of the data matrix and whether a discriminative subspace exists. For example, given two clusters of data generated from 2D normal distributions as shown in Figure 1, η depends on the separation of these two clusters. We can calculate $\eta^* \triangleq \mu_p^* - \mu_{p+1}^*$ since we know the true subspace $\mathbf{P}^*$, and we also run **WDA-eig** on a random initialization to get η. We observe that η is close to 0 when the clusters overlap and is a monotonically increasing function of the Euclidean distance between the mean of the two clusters.

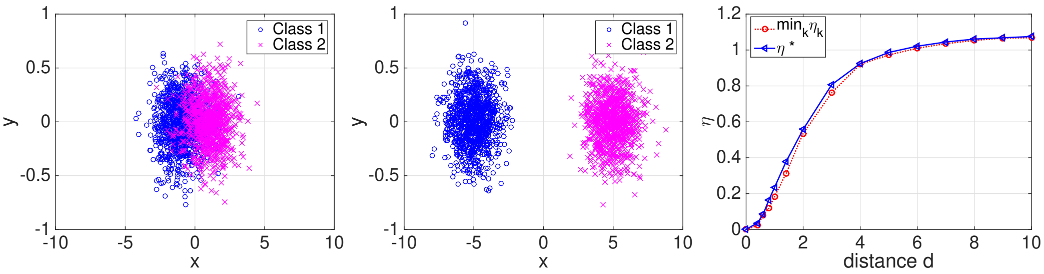

Figure 1: Left and middle: two classes of data generated from two random normal distributions: $X_i \mathcal{N} \in (\mu_i, \Sigma)$, $\mu = (\pm x, 0)$ where $x \in [0, 5]$. $x = 1$ in the Left and $x = 5$ in the Middle. Right: Arctangent gap η as a function of the distance between the means $d \triangleq \|\mu_1 - \mu_2\|$.

By applying the algorithm to a simulated dataset with 3 classes and 2 discriminative dimensions, we draw log plots of the distances between the subspaces subject to these components in Figure 2 to illustrate linear convergence rates. On the left we show s_k with different values of the subspace dimension p and with $\lambda = 0.1$ fixed. With $p = 2$ the algorithm achieves the fastest rate because the dimension of the true discriminative subspace is 2. In FDA, since $\mathbf{C}_b$ has rank $Nc - 1$ (where Nc is the number of classes), p has to be $\leq Nc - 1$. In **WDA-eig**, p is less restrictive, but choosing $p \geq Nc$ may still slow down or prevent convergence if λ is small. In the middle we show s_k with different values of the Wasserstein regularizer λ and with $p = 2$ fixed. When λ is small, the matrices $\mathbf{C}_w$ and $\mathbf{C}_b$ in WDA can be viewed as the matrices in FDA with a small perturbation, and in such cases the Lipschitz constants ξ_a and ξ_b are close to zero so the algorithm is guaranteed to converge. We also observe that a larger λ corresponds to a slower convergence rate. On the right we illustrate the effect of initialization for local convergence. We use the converged solution as an approximation to the true discriminative subspace $\mathbf{P}^*$ and plot the distance $\| \sin \Theta(\mathbf{P}^*, \mathbf{P}_{k-1}) \|$ for each iteration k, with varying $\hat{s}_0 = \| \sin \Theta(\mathbf{P}^*, \mathbf{P}_0) \|$. We observe that initialization has little effect on the convergence rate and that the algorithm converges in most cases except for the case where $\hat{s}_0 \approx 1$.

4 Numerical Experiments

In this section we evaluate the performance of the proposed Algorithm 1 on classification tasks by applying it to a simulated dataset and the MNIST dataset. We refer to our proposed algorithm as **WDA-eig** and refer to the original implementation in [12] with projected gradient descent as **WDA**. **WDA** converges when the norm of the gradient is below 10^{-6}, and **WDA-eig** converges when the distance between two consecutive subspaces is less than 10^{-6}.

4.1 Simulated dataset

We first compare **WDA-eig** with **WDA** on a simulated dataset. We use the same setup as given in [13], where the data belongs to 3 non-linearly separable classes and is generated using 2 discrim-

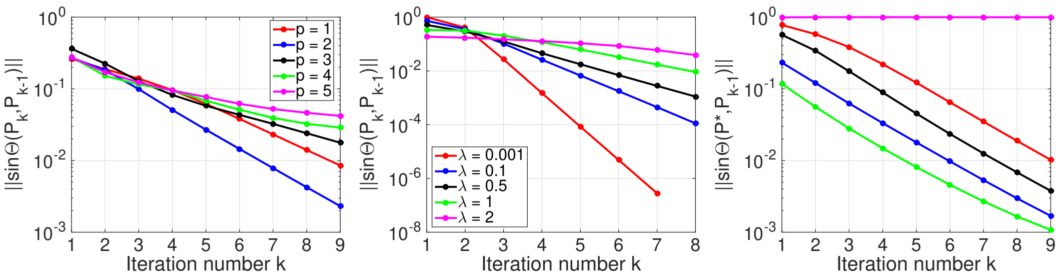

Figure 2: Left and Middle: Distances between subspaces $\|\sin\Theta(\mathbf{P}_k, \mathbf{P}_{k-1})\|$ as a function of iteration number k, with varying λ and p, respectively. Right: $\|\sin\Theta(\mathbf{P}^*, \mathbf{P}_{k-1})\|$ as a function of iteration number k, with varying initialization $\mathbf{P}_0$.

Table 1: Comparison between **WDA** and **WDA-eig**

Param λ	Algo	Prob. of Convergence	Avg. Acc.(std.)	Converged Acc.(std.)	CPU time
$\lambda = 0.1$	**wda**	25%	0.712(0.152)	0.977(0)	72
	wda-eig	100%	0.968(0)	0.968(0)	0.677
$\lambda = 1.0$	**wda**	78%	0.908(0.149)	0.987(0)	6.37
	wda-eig	100%	0.986(0)	0.986(0)	1.17
$\lambda = 5.0$	**wda**	73%	0.885(0.164)	0.985(0)	7.04
	wda-eig	100%	0.985(0)	0.985(0)	1.61

inant features and 8 dimensions of Gaussian noise. We apply these two algorithms with varying regularization parameter λ, and compare their computational efficiency and classification accuracy with a K-Nearest-Neighbors classifier (KNN) on the projected data ($k = 10$). For each λ, we run each algorithms for 100 randomly-initialized trials, and the results are shown in Table 1. The third column of the table shows the probability of convergence over 100 trials, and the fourth column shows the accuracy averaged over trials. For $\lambda = 0.1,\ 1,\ 5$, **WDA-eig** converges in all the trials with zero standard deviations and achieves higher accuracy scores on average, while **WDA** has high standard deviation due to the low probability of convergence. The fifth column shows the accuracy averaged only for the converged trials, and **WDA-eig** and **WDA** have comparable performances in accuracy, which indicates that the ratio trace formulation can serve as a good approximation to the trace ratio formulation. The last column shows the efficiency measured by averaged CPU time in seconds over 100 trials. **WDA-eig** takes shorter running time than **WDA** since the former only requires a few iterations to converge and the running time per SCF iteration is comparable to the running time per gradient descent iteration. Even in cases where most trials converge for both solvers (e.g., when $\lambda = 1$), **WDA** takes more iterations to converge on average.

4.2 MNIST dataset

Next, we test the classification performance on a real dataset and also evaluate the generalization ability of the proposed approach. We extract 1000 samples in the MNIST dataset as the training set and use 10000 samples in the test set. We measure the KNN prediction error on the projected data as a function of the subspace dimension p, the number of nearest neighbors K, and the Wasserstein regularization parameter λ respectively in Figure 3. On the left we show the prediction error of full data/PCA/FDA/**WDA/WDA-eig**+KNN applied to the original data as a function of p, with $\lambda = 0.01$ and $K = 10$ fixed. In implementation of FDA/**WDA-eig** we add a small perturbation term ϵI_p on $\mathbf{C}_w$ to make the denominator positive definite, and we choose $\epsilon = 2$ in this setting, which removes the restriction of $p \leq 9$ for FDA. In the middle we show the performance of these methods as a function of K. Another approach to avoid $\mathbf{C}_w$ being semidefinite is to project away the null space of the data matrix before applying discriminant analysis. To achieve this end, we first apply PCA to the original data matrix and retain only the first 20 principal components. We then apply PCA/FDA/LFDA [37]/**WDA/WDA-eig** on the dimension-reduced data to obtain a subspace of dimension $p = 9$ without any regularization on $\mathbf{C}_w$, and the results are shown on the right.

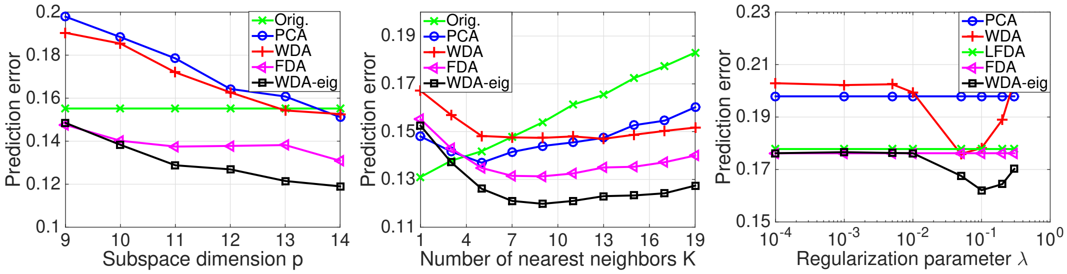

Figure 3: Prediction error as a function of the subspace dimension p, the number of nearest neighbors K, and the Wasserstein regularization parameter λ.

5 Unsupervised WDA

Since WDA is a dimensionality reduction technique, it could also be integrated with a low-dimensional clustering technique to do high-dimensional clustering. Here we propose Algorithm 2 to extend WDA to the unsupervised setting.

5.1 Clustering Algorithm

Algorithm 2 Iterative WDA clustering

```
Input: De-meaned dataset $X$, $\mathbf{P}_0 \in \mathbb{O}^{d\times p}$, tolerance $\epsilon$, max number of iterations $N$
for $k = 0, 1, \cdots, N$ do
    Compute $y = X\mathbf{P}_k \in \mathbb{R}^{n\times p}$
    Cluster $y$ into $K$ classes and obtain the class labels $\hat{y}$
    Call Algorithm 1 to update $\mathbf{P}$ : $\mathbf{P}_{k+1} = \textbf{WDA-eig}(X, \hat{y}, \mathbf{P}_k)$
    if the change in $\mathbf{P}_k$ is sufficiently small then
        Break
    end if
end for
```

We start with an initial guess and adaptively improve its labeling by performing clustering in the projected space. The goal is to converge to a discriminative subspace that will render the most accurate labels. Algorithm 2 solves the following optimization problem:

$$\max_{\mathbf{P}} \hat{J}(\mathbf{P}, \mathbf{T}, \hat{y}) \quad \text{s.t. } \mathbf{T}^{c,c'} = \operatorname*{argmin}_{\mathbf{T}\in U_{n_c n_{c'}}} E_1(\mathbf{T}, \mathbf{P}, \hat{y}),\ \hat{y} = \operatorname*{argmin}_{\hat{y}} E_2(\mathbf{P}, \hat{y}), \tag{8}$$

and E_2 is the objective of any specific low-dimensional clustering technique. The algorithm uses an alternating optimization scheme: for each iteration, given the class labels $\hat{y}$, $\mathbf{P}$ is chosen to maximize the ratio-trace problem $\hat{J}$, and then given the subspace, it finds the optimal labeling according to the clustering objective E_2. The objectives E_2 and $\hat{J}$ do not always align. A special case is FDA-Kmeans (or LDA-Kmeans) [10], where minimizing E_2 is equivalent to maximizing $\hat{J}$. It is derived that iteratively applying FDA and K-means is the same as alternating optimization in a unified objective [42], and that combining FDA and K-means is equivalent to kernel K-means in the original space with a specific kernel Gram matrix [48].

However, there is no theoretical guarantee that a larger objective value corresponds to a better clustering result in terms of external evaluation criteria. We observe that for FDA-Kmeans, the adjusted random index (ARI) [30] does not increase monotonically with the iteration number and could even converge to a worse result compared to the initial guess. For WDA, K-Means in the projected space does not maximize $\hat{J}$, but empirically we observe that several iterations with K-Means does improve clustering result in terms of external evaluation criteria such as ARI. Since the performance of FDA degrades when class distributions are multimodal, FDA could perform poorly given the wrong labels even if the true underlying distribution is Gaussian. On the other hand, we numerically observe that WDA is more robust to noisy labels due to a balance of local and global information (see supplementary material).

Table 2: Clustering results.

Method	Dataset	ARI	NMI	Homogeneity	Completeness	FMI
Baseline	MNIST	0.334 ± 0.007	0.475 ± 0.006	0.473 ± 0.006	0.477 ± 0.007	0.403 ± 0.007
PCAKm	MNIST	0.334 ± 0.005	0.471 ± 0.005	0.469 ± 0.004	0.473 ± 0.007	0.402 ± 0.005
FDAKm	MNIST	0.360 ± 0.006	0.500 ± 0.007	0.497 ± 0.006	0.503 ± 0.008	0.426 ± 0.005
WDAKm	MNIST	$\mathbf{0.398 \pm 0.008}$	$\mathbf{0.526 \pm 0.006}$	$\mathbf{0.524 \pm 0.006}$	$\mathbf{0.528 \pm 0.006}$	$\mathbf{0.459 \pm 0.008}$
Baseline	KTH	0.424 ± 0.035	0.576 ± 0.035	0.556 ± 0.035	0.598 ± 0.033	0.535 ± 0.028
PCAKm	KTH	0.470 ± 0.017	0.616 ± 0.009	0.596 ± 0.011	0.637 ± 0.007	0.571 ± 0.011
FDAKm	KTH	0.481 ± 0.022	0.635 ± 0.018	0.614 ± 0.019	0.657 ± 0.017	0.580 ± 0.016
WDAKm	KTH	$\mathbf{0.488 \pm 0.020}$	$\mathbf{0.643 \pm 0.015}$	$\mathbf{0.623 \pm 0.016}$	$\mathbf{0.663 \pm 0.014}$	$\mathbf{0.584 \pm 0.014}$
Baseline	15scene	0.161 ± 0.009	0.352 ± 0.009	0.336 ± 0.009	0.369 ± 0.009	0.233 ± 0.008
PCAKm	15scene	0.160 ± 0.007	0.351 ± 0.006	0.334 ± 0.007	0.368 ± 0.006	0.232 ± 0.005
FDAKm	15scene	0.150 ± 0.009	0.350 ± 0.011	0.324 ± 0.010	0.376 ± 0.014	0.234 ± 0.010
WDAKm	15scene	$\mathbf{0.170 \pm 0.010}$	$\mathbf{0.366 \pm 0.012}$	$\mathbf{0.350 \pm 0.011}$	$\mathbf{0.384 \pm 0.012}$	$\mathbf{0.243 \pm 0.009}$
Baseline	20ng	0.081 ± 0.011	0.235 ± 0.021	0.217 ± 0.020	0.255 ± 0.023	0.151 ± 0.009
PCAKm	20ng	0.097 ± 0.003	0.247 ± 0.005	0.238 ± 0.005	0.255 ± 0.005	0.152 ± 0.003
FDAKm	20ng	0.113 ± 0.013	0.298 ± 0.019	0.275 ± 0.019	0.322 ± 0.020	0.185 ± 0.010
WDAKm	20ng	$\mathbf{0.128 \pm 0.011}$	$\mathbf{0.302 \pm 0.014}$	$\mathbf{0.283 \pm 0.013}$	$\mathbf{0.322 \pm 0.015}$	$\mathbf{0.194 \pm 0.010}$

5.2 Experiments on WDA Clustering

In this section we evaluate the proposed Algorithm 2 and compare with other subspace clustering techniques. In what follows, let Nc denote the number of classes, n be the number of observations and d be the number of features.

We use four real world datasets to evaluate the proposed method: the MNIST dataset for digits recognition, the 15-scene dataset [19] for multi-class image recognition, the KTH action recognition database [33] for multi-class video recognition, and the 20 newsgroup dataset for text classification. To avoid the singularity of the $\mathbf{C}_w$ matrix in FDA and WDA, we first do a dimension reduction on the original dataset using PCA and retain the first $2 \times Nc$ principal components. We refer to this data as the dimension-reduced data. We apply four different clustering methods to the four dataset: (1) K-means on the original data (Baseline); (2) K-means on dimension-reduced data (PCAKm); (3) FDA-Kmeans (FDAKm) on the dimension-reduced data; (4) WDA-kmeans (Algorithm 2 combined with K-means) (WDAKm) on the dimension-reduced data. For (3) and (4) we use the subspace obtained by PCA as initialization and $p = Nc-1$ as the subspace dimensions. No regularization term is added to $\mathbf{C}_w$. The Wasserstein regularizer λ is coarsely tuned, where we choose $\lambda = 0.01$ for MNIST and 15-scene, $\lambda = 10$ for KTH, and $\lambda = 5$ for 20ng. The results are averaged over 20 trials. We use five external evaluation criteria to evaluate the quality of the clustering solutions [30, 36, 31, 14]. The results in Table 2 show that WDAKm achieves the best performance on all four datasets, in terms of the 5 external metrics we use. We also notice that in 15-scene dataset the performance of FDAKm is worse than the baseline method, which means with some wrong tags, FDA tends to overly separate data and decrease the clustering quality. In contrast, WDA always improve the clustering even in the difficult case such as the 15-scene dataset.

6 Conclusion

In this paper, we present a ratio trace formulation of the Wasserstein Discriminant Analysis and an eigensolver-based algorithm: **WDA-eig** to solve the problem. Unlike the original trace ratio formulation, the ratio trace formulation has a closed-form solution that is readily obtainable by the generalized eigenvalue decomposition once the regularized optimal transport problem is solved. We give a convergent analysis for **WDA-eig** under the SCF framework and numerically test the efficiency and convergence properties of the proposed algorithm. Although **WDA-eig** solves a slightly different problem, the ratio trace formulation can be served as an efficient alternative for the trace ratio formulation of WDA. **WDA-eig** also takes less time to converge on average and is less sensitive to initialization and parameters compared to **WDA**. As a supervised dimensionality reduction technique, WDA can also be combined with clustering techniques and applied iteratively to perform unsupervised learning. Numerical experiments show that the WDA clustering algorithm performs well on a set of real-world problems.

Broader Impact

In the era of big data, business providers, data scientists, and governments try to explore opportunities in the large scale and high-dimensional datasets. Nevertheless, several major computational challenges arise and prevent practitioners from constructing effective algorithms or tools to analyze their datasets. Dimensionality reduction (DR) plays an essential role in supervised and unsupervised learning tasks when the datasets are high dimensional. One benefit of reducing the data dimension before classification or clustering is to save storage and reduce computational cost for the later steps, however, the DR technique itself can be costly. We study a recently proposed and promising DR technique, the Wasserstein discriminant analysis, and propose a different formulation that could achieve comparable or better results with less computational cost. We also analyze the problem from a different perspective that was originated from electronic structure calculations, which could be of interest to a broader audience in the machine learning community.

Acknowledgments and Disclosure of Funding

We thank the anonymous Referees and Area Chair for their constructive comments. The work is supported by Baidu Research.

References

[1] Mokhtar Z. Alaya, Maxime Berar, Gilles Gasso, and Alain Rakotomamonjy. Screening sinkhorn algorithm for regularized optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 12169–12179, Vancouver, Canada, 2019.

[2] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1964–1974, Long Beach, CA, 2017.

[3] Nicolas Boumal, Bamdev Mishra, Pierre-Antoine Absil, and Rodolphe Sepulchre. Manopt, a matlab toolbox for optimization on manifolds. *J. Mach. Learn. Res.*, 15(1):1455–1459, 2014.

[4] Deng Cai, Xiaofei He, Kun Zhou, Jiawei Han, and Hujun Bao. Locality sensitive discriminant analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 708–713, Hyderabad, India, 2007.

[5] Yunfeng Cai and Ping Li. An inverse-free truncated rayleigh-ritz method for sparse generalized eigenvalue problem. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, Online [Palermo, Sicily, Italy], 2020.

[6] Yunfeng Cai, Lei-Hong Zhang, Zhaojun Bai, and Ren-Cang Li. On an eigenvector-dependent nonlinear eigenvalue problem. *SIAM Journal on Matrix Analysis and Applications*, 39(3):1360–1382, 2018.

[7] Eric Cancès. Self-consistent field algorithms for kohn–sham models with fractional occupation numbers. *The Journal of Chemical Physics*, 114(24):10616–10622, 2001.

[8] Benoît Colson, Patrice Marcotte, and Gilles Savard. An overview of bilevel optimization. *Ann. Oper. Res.*, 153(1):235–256, 2007.

[9] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2292–2300, Lake Tahoe, NV, 2013.

[10] Chris H. Q. Ding and Tao Li. Adaptive dimension reduction using discriminant analysis and K-means clustering. In *Proceedings of the Twenty-Fourth International Conference (ICML)*, pages 521–528, Corvallis, OR, 2007.

[11] Zizhu Fan, Yong Xu, and David Dapeng Zhang. Local linear discriminant analysis framework using sample neighbors. *IEEE Trans. Neural Networks*, 22(7):1119–1132, 2011.

[12] R'emi Flamary and Nicolas Courty. Pot python optimal transport library, 2017.

[13] Rémi Flamary, Marco Cuturi, Nicolas Courty, and Alain Rakotomamonjy. Wasserstein discriminant analysis. *Mach. Learn.*, 107(12):1923–1945, 2018.

[14] E. B. Fowlkes and C. L. Mallows. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78(383):553–569, 1983.

[15] Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Elsevier, 2013.

[16] Trevor J. Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning:Data Mining, Inference, and Prediction*. Springer, New York, NY, 2001.

[17] Yangqing Jia, Feiping Nie, and Changshui Zhang. Trace ratio problem revisited. *IEEE Trans. Neural Networks*, 20(4):729–735, 2009.

[18] Philip A. Knight. The sinkhorn-knopp algorithm: Convergence and applications. *SIAM J. Matrix Anal. Appl.*, 30(1):261–275, 2008.

[19] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2169–2178, New York, NY, 2006.

[20] Ping Li. Linearized GMM kernels and normalized random Fourier features. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 315–324, 2017.

[21] Xiaoyun Li, Jie Gui, and Ping Li. Randomized kernel multi-view discriminant analysis. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI)*, pages 1276–1284, Santiago de Compostela, Spain, 2020.

[22] Xuelong Li, Mulin Chen, Feiping Nie, and Qi Wang. Locality adaptive discriminant analysis. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2201–2207, Melbourne, Australia, 2017.

[23] Tianyi Lin, Nhat Ho, and Michael I. Jordan. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 3982–3991, Long Beach, CA, 2019.

[24] Xin Liu, Xiao Wang, Zaiwen Wen, and Yaxiang Yuan. On the convergence of the self-consistent field iteration in kohn–sham density functional theory. *SIAM Journal on Matrix Analysis and Applications*, 35(2):546–558, 2014.

[25] Dougal Maclaurin, David Duvenaud, and Ryan P Adams. Autograd: Effortless gradients in numpy. In *ICML 2015 AutoML Workshop*, volume 238, 2015.

[26] Richard M Martin and Richard Milton Martin. *Electronic structure: basic theory and practical methods*. Cambridge university press, 2004.

[27] Thanh T. Ngo, Mohammed Bellalij, and Yousef Saad. The trace ratio optimization problem. *SIAM Rev.*, 54(3):545–569, 2012.

[28] Feiping Nie, Shiming Xiang, Yangqing Jia, Changshui Zhang, and Shuicheng Yan. Trace ratio criterion for feature selection. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (AAAI)*, pages 671–676, Chicago, IL, 2008.

[29] Feiping Nie, Shiming Xiang, and Changshui Zhang. Neighborhood minmax projections. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 993–998, Hyderabad, India, 2007.

[30] William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[31] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic, 2007.

[32] Yousef Saad, James R. Chelikowsky, and Suzanne M. Shontz. Numerical methods for electronic structure calculations of materials. *SIAM Rev.*, 52(1):3–54, 2010.

[33] Christian Schüldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local SVM approach. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, pages 32–36, Cambridge, UK, 2004.

[34] Bharath K. Sriperumbudur and Gert R. G. Lanckriet. On the convergence of the concave-convex procedure. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1759–1767, Vancouver, Canada, 2009.

[35] Gilbert W Stewart and Ji-Guang Sun. *Matrix Perturbation Theory*. Academic Press, Boston, 1990.

[36] Alexander Strehl and Joydeep Ghosh. Cluster ensembles — A knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3:583–617, 2002.

[37] Masashi Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *J. Mach. Learn. Res.*, 8:1027–1061, 2007.

[38] Ji-guang Sun. The perturbation bounds for eigenspaces of a definite matrix-pair. *Numerische Mathematik*, 41(3):321–343, 1983.

[39] James Townsend, Niklas Koep, and Sebastian Weichwald. Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation. *J. Mach. Learn. Res.*, 17:137:1–137:5, 2016.

[40] Charles F Van Loan and Gene H Golub. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, 4th edition, 2012.

[41] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[42] De Wang, Feiping Nie, and Heng Huang. Unsupervised feature selection via unified trace ratio formulation and k-means clustering (TRACK). In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, pages 306–321, Nancy, France, 2014.

[43] Huan Wang, Shuicheng Yan, Dong Xu, Xiaoou Tang, and Thomas S. Huang. Trace ratio vs. ratio trace for dimensionality reduction. In *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, MN, 2007.

[44] Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.*, 10:207–244, 2009.

[45] Zhiqiang Xu and Ping Li. A practical riemannian algorithm for computing dominant generalized eigenspace. In *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI)*, page 354, 2020.

[46] Shuicheng Yan, Dong Xu, Benyu Zhang, HongJiang Zhang, Qiang Yang, and Stephen Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(1):40–51, 2007.

[47] Chao Yang, Weiguo Gao, and Juan C Meza. On the convergence of the self-consistent field iteration for a class of nonlinear eigenvalue problems. *SIAM journal on matrix analysis and applications*, 30(4):1773–1788, 2009.

[48] Jieping Ye, Zheng Zhao, and Mingrui Wu. Discriminative k-means for clustering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1649–1656, Vancouver, Canada, 2007.

[49] Yu Zhang and Dit-Yan Yeung. Worst-case linear discriminant analysis. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2568–2576, Vancouver, Canada, 2010.

7 Proof of Theorem 1. (Global Convergence of SCF)

Consider the generalized NLEP $A(\mathbf{P})\mathbf{V} = B(\mathbf{P})\mathbf{V}\Lambda$, where $\mathbf{V} = [v_1, \dots, v_p]$ and $\mathbf{P}$ is an orthonormal basis of $\mathbf{V}$. $A(\mathbf{P})$, $B(\mathbf{P})$ are symmetric matrix-valued function and $B(\mathbf{P})$ is positive definite. $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_p)$, where $\lambda_1 \geq \cdots \geq \lambda_p$ are the p largest eigenvalues of $(A(\mathbf{P}), B(\mathbf{P}))$ corresponding to eigenvectors $v_1, \dots, v_p$. We emphasize that $A(\mathbf{P})$, $B(\mathbf{P})$ are invariant to orthogonal transformation of $\mathbf{P}$, i.e., $A(\mathbf{P}) \equiv A(\mathbf{P}Q)$, $B(\mathbf{P}) \equiv B(\mathbf{P}Q)$ for any orthogonal matrix $Q \in \mathbb{R}^{p\times p}$.

Definitions. Let $\mathcal{X}$ and $\mathcal{Y}$ be two p-dimensional subspaces of $\mathbb{R}^n$. Let the columns of X form an orthonormal basis for $\mathcal{X}$ and the columns of Y form an orthonormal basis for $\mathcal{Y}$. We use $\|\sin\Theta(\mathcal{X}, \mathcal{Y})\|$ as in [35] to measure the distance between $\mathcal{X}$ and $\mathcal{Y}$, where

$$\Theta(\mathcal{X}, \mathcal{Y}) = \mathrm{diag}(\theta_1(\mathcal{X}, \mathcal{Y}), \dots, \theta_p(\mathcal{X}, \mathcal{Y})).$$

Here, $\theta_j(\mathcal{X}, \mathcal{Y})$'s denote the *canonical angles* between $\mathcal{X}$ and $\mathcal{Y}$ [p. 43][35], which is defined as

$$0 \leq \theta_j(\mathcal{X}, \mathcal{Y}) \triangleq \arccos \sigma_j \leq \frac{\pi}{2} \quad \text{for } 1 \leq j \leq k,$$

where σ_j's are the singular values of $X^T Y$. Similar to the Crawford number for symmetric definite matrix pair (A, B) [Chapter 8.7] [40], we define the Crawford number for the generalized NLEP as

$$c \triangleq \min_{\mathbf{P}\in\mathbb{O}^{d\times p}} \min_{x\in\mathbb{C}^d, \|x\|=1} (x^T(A(\mathbf{P}) + iB(\mathbf{P}))x).$$

Define $C \triangleq \max_{\mathbf{P}\in\mathbb{O}^{d\times p}} \sqrt{\|A(\mathbf{P})^2 + B(\mathbf{P})^2\|}$. At the kth SCF iteration, one computes an approximation to the eigenvector matrix $\mathbf{V}_k$ associated with the p largest eigenvalues of $(A(\mathbf{P}_{k-1}), B(\mathbf{P}_{k-1}))$, where $\mathbf{P}_{k-1}$ is an orthonormal basis for $\mathbf{V}_{k-1}$, and then $\mathbf{V}_k$ is used as the next approximation to the solution. Let $A_k = A(\mathbf{P}_k)$, $B_k = B(\mathbf{P}_k)$, $\mu_{i,k} = \arctan \lambda_i(A(\mathbf{P}_k), B(\mathbf{P}_k))$.

We study the convergence of SCF iteration under the following assumptions:

A1: For any $\mathbf{P}_1, \mathbf{P}_2 \in \mathbb{O}^{d\times p}$, assume that there exist positive constants ξ_a, ξ_b such that

$$\|A(\mathbf{P}_1) - A(\mathbf{P}_2)\| \leq \xi_a \|\sin\Theta(\mathbf{P}_1, \mathbf{P}_2)\|, \qquad \|B(\mathbf{P}_1) - B(\mathbf{P}_2)\| \leq \xi_b \|\sin\Theta(\mathbf{P}_1, \mathbf{P}_2)\|;$$

A2: For $k = 1, 2, \cdots$, there exists an $\eta > 0$ such that

$$\mu_{p,k} - \mu_{p+1,k} \geq \eta.$$

Theorem 1. ***(Global Convergence)*** *Let $s_1 = \|\sin\Theta(\boldsymbol{P}_0, \boldsymbol{P}_1)\|$. Assume **A1** and **A2**, and $s_1\sqrt{\xi_a^2 + \xi_b^2} < c$. If*

$$\eta > \arcsin(\rho C\sqrt{\xi_a^2 + \xi_b^2}/c^2) + \arctan(s_1\sqrt{\xi_a^2 + \xi_b^2}/c)$$

for some constant $\rho > 1$, then SCF converges linearly at the rate of $\frac{1}{\rho}$.

In order to show Theorem 1, we need the following three lemmas. The first lemma gives some fundamental results for $\|\sin\Theta(X, Y)\|$, which can be verified via definition.

Lemma 1. *Let $[X,\ X_\mathrm{c}]$ and $[Y,\ Y_\mathrm{c}]$ be two orthogonal matrices with $X, Y \in \mathbb{R}^{n\times k}$. Then*

$$\|\sin\Theta(X, Y)\| = \|X_\mathrm{c}^T Y\| = \|X^T Y_\mathrm{c}\| = \|XX^T - YY^T\|.$$

The next lemma gives perturbation bound for the eigenvalues of definite matrix pair.

Lemma 2. *[Theorem 8.7.3][40] Let A, B be symmetric, B be positive definite. Let the eigenvalues of (A, B) be $\lambda_1 \geq \cdots \geq \lambda_n$. Let $c(A, B)$ be the Crawford number of $\{A, B\}$:*

$$c(A, B) \equiv \min_{x\in\mathbb{C}^n, \|x\|=1} |x^T(A + iB)x|.$$

Suppose E and F are symmetric matrices that satisfy

$$\epsilon^2 = \|E\|^2 + \|F\|^2 < c^2(A,B).$$

Then $B+F$ is positive definite, and the eigenvalues $\tilde{\lambda}_1 \geq \ldots \tilde{\lambda}_n$ of $(A+E, B+F)$ satisfy

$$|\arctan \tilde{\lambda}_i - \arctan \lambda_i| \leq \arctan \frac{\epsilon}{c(A,B)},\ \forall 1 \leq i \leq n.$$

The following lemma gives perturbation bound for the eigenspace of definite matrix pair, which is rewritten from [Theorem 2.1] [38].

Lemma 3. *Let A, B, $\widetilde{A}$, $\widetilde{B}$ be symmetric, B and $\widetilde{B}$ be positive definite. Let the eigenvalues of (A,B) and $(\widetilde{A},\widetilde{B})$ be $\tan\mu_1 \geq \cdots \geq \tan\mu_n$, $\tan\tilde{\mu}_1 \geq \cdots \geq \tan\tilde{\mu}_n$, respectively, the corresponding eigenvectors be $v_1,\ldots,v_n$, and $\tilde{v}_1,\ldots,\tilde{v}_n$, respectively. Assume that there are $\alpha \geq 0$ and $\delta > 0$ satisfying $\alpha+\delta \leq 1$, and a real number γ such that*

$$|\sin(\gamma-\mu_i)| \leq \alpha, \text{ for } i = 1,\ldots,p,$$
$$|\sin(\gamma-\tilde{\mu}_j)| \geq \alpha+\delta, \text{ for } j = p+1,\ldots,n$$

(or vice-versa). Let $V_1 = [v_1,\ldots,v_p]$, $\widetilde{V}_1 = [\tilde{v}_1,\ldots,\tilde{v}_p]$. Then

$$\|\sin\Theta(V_1,\widetilde{V}_1)\| \leq \frac{p(\alpha,\delta;\gamma)\sqrt{\|A^2+B^2\|}}{c(A,B)c(\widetilde{A},\widetilde{B})} \times \frac{\sqrt{\|(\widetilde{A}-A)^2+(\widetilde{B}-B)^2\|}}{\delta},$$

where

$$p(\alpha,\delta;\gamma) \triangleq \frac{q(\gamma)(\alpha+\delta)\sqrt{1-\alpha^2}+\alpha\sqrt{1-(\alpha+\delta)^2}}{2\alpha+\delta},$$

with $q(\gamma) = \sqrt{2}$ for $\gamma \neq 0$ and $q(0) = 1$.

Now we are ready to prove **Theorem 1**.

Proof of Theorem 1. Denote $A_k = A(\mathbf{P}_k)$, $B_k = B(\mathbf{P}_k)$, $E_k = A_k - A_{k-1}$, $F_k = B_k - B_{k-1}$. Without loss of generality, we assume that A_k is also positive definite. Otherwise, we let $A_k = A_k + tB_k$ for sufficiently large t, then A_k is positive definite and the sequence $\{\mathbf{V}_k\}$ produced by SCF iteration remains unchanged. Let $\Lambda_k = \text{diag}(\lambda_{1,k},\ldots,\lambda_{p,k})$, $\mathbf{V}_k = [v_{1,k},\ldots,v_{p,k}]$, where $\lambda_{i,k}$ is the i^{th} largest eigenvalue of (A_k,B_k), $v_{i,k}$ is the corresponding eigenvector. Also denote $s_k = \|\sin\Theta(\mathbf{P}_k,\mathbf{P}_{k-1})\| = \|\mathbf{P}_k\mathbf{P}_k^T - \mathbf{P}_{k-1}\mathbf{P}_{k-1}^T\|$ as the distance between subspaces.

By assumption **A1**, we have

$$\begin{aligned}&\sqrt{\|A_k-A_{k-1}\|^2+\|B_k-B_{k-1}\|^2}\\ \leq&\sqrt{\xi_a^2+\xi_b^2}\|\sin\Theta(\mathbf{P}_k,\mathbf{P}_{k-1})\| = s_k\sqrt{\xi_a^2+\xi_b^2}.\end{aligned}$$

Now consider $k=1$. By assumption, $s_1\sqrt{\xi_a^2+\xi_b^2} < c$, then we may apply Lemma 2, which gives

$$\begin{aligned}|\mu_{i,1}-\mu_{i,0}| \leq& \arctan\frac{\sqrt{\|A_1-A_0\|^2+\|B_1-B_0\|^2}}{c}\\ \leq& \arctan(s_1\sqrt{\xi_a^2+\xi_b^2}/c),\ \forall 1 \leq i \leq n.\end{aligned}$$

It follows that

$$\begin{aligned}\mu_{p,1}-\mu_{p+1,0} =& \mu_{p,1}-\mu_{p+1,1}+\mu_{p+1,1}-\mu_{p+1,0}\\ \geq& \eta - \arctan(s_1\sqrt{\xi_a^2+\xi_b^2}/c)\\ >& \arcsin(\rho C\sqrt{\xi_a^2+\xi_b^2}/c^2) \geq 0,\end{aligned} \tag{9}$$

where the last inequality uses the assumption

$$\eta > \arcsin(\rho C\sqrt{\xi_a^2 + \xi_b^2}/c^2) + \arctan(s_1\sqrt{\xi_a^2 + \xi_b^2}/c).$$

Now let

$$\gamma = \frac{\mu_{1,1} + \mu_{p,1}}{2},\ \alpha = \sin\frac{\mu_{1,1} - \mu_{p,1}}{2},\ \alpha + \delta = \sin(\frac{\mu_{1,1} + \mu_{p,1}}{2} - \mu_{p+1,0}),$$

then for all $1 \le i \le p$ and $p + 1 \le j \le n$, we have

$$|\sin(\gamma - \mu_{i,1})| \le |\sin\frac{\mu_{1,1} - \mu_{p,1}}{2}| = \alpha, \tag{10a}$$

$$|\sin(\gamma - \mu_{j,0})| \ge |\sin(\frac{\mu_{1,1} + \mu_{p,1}}{2} - \mu_{p+1,0})| = \alpha + \delta, \tag{10b}$$

$$\alpha + \delta \le 1, \qquad \gamma > 0, \tag{10c}$$

$$\begin{aligned}\delta =& \sin(\frac{\mu_{1,1} + \mu_{p,1}}{2} - \mu_{p+1,0}) - \sin\frac{\mu_{1,1} - \mu_{p,1}}{2}\\ =& 2\cos\frac{\mu_{1,1} - \mu_{p+1,0}}{2}\sin\frac{\mu_{p,1} - \mu_{p+1,0}}{2}.\end{aligned} \tag{10d}$$

By calculations, we obtain

$$\begin{aligned}p(\alpha, \delta; \gamma) &= \frac{(\alpha + \delta)\sqrt{1 - \alpha^2} + \alpha\sqrt{1 - (\alpha + \delta)^2}}{2\alpha + \delta}\\ &= \frac{\sin\frac{\mu_{1,1} - \mu_{p,1}}{2}\cos(\frac{\mu_{1,1} + \mu_{p,1}}{2} - \mu_{p+1,0}) + \cos\frac{\mu_{1,1} - \mu_{p,1}}{2}\sin(\frac{\mu_{1,1} + \mu_{p,1}}{2} - \mu_{p+1,0})}{\sin\frac{\mu_{1,1} - \mu_{p,1}}{2} + \sin(\frac{\mu_{1,1} + \mu_{p,1}}{2} - \mu_{p+1,0})}\\ &= \frac{\sin(\mu_{1,1} - \mu_{p+1,0})}{\sin\frac{\mu_{1,1} - \mu_{p,1}}{2} + \sin(\frac{\mu_{1,1} + \mu_{p,1}}{2} - \mu_{p+1,0})}\\ &= \frac{2\sin\frac{\mu_{1,1} - \mu_{p+1,0}}{2}\cos\frac{\mu_{1,1} - \mu_{p+1,0}}{2}}{2\sin\frac{\mu_{1,1} - \mu_{p+1,0}}{2}\cos\frac{\mu_{p,1} - \mu_{p+1,0}}{2}}\\ &= \frac{\cos\frac{\mu_{1,1} - \mu_{p+1,0}}{2}}{\cos\frac{\mu_{p,1} - \mu_{p+1,0}}{2}}.\end{aligned} \tag{11}$$

Using Lemma 3, we have

$$s_2 \le \frac{p(\alpha, \delta; \gamma)C}{c^2} \cdot \frac{\sqrt{\|(A_1 - A_0)^2 + (B_1 - B_0)^2\|}}{\delta} \le \frac{p(\alpha, \delta; \gamma)C}{c^2} \cdot \frac{\sqrt{\xi_a^2 + \xi_b^2}}{\delta}s_1. \tag{12}$$

Substituting (10d), (11) into (12), and using (9), we have

$$s_2 \le \frac{1}{\rho}s_1. \tag{13}$$

where

$$\rho = \frac{c^2\sin(\mu_{p,1} - \mu_{p+1,0})}{C\sqrt{\xi_a^2 + \xi_b^2}} > 1. \tag{14}$$

For general $k = 2$, noticing the following holds

$$\begin{aligned}\eta &> \arcsin(\rho C\sqrt{\xi_a^2 + \xi_b^2}/c^2) + \arctan(s_1\sqrt{\xi_a^2 + \xi_b^2}/c)\\ &> \arcsin(\rho C\sqrt{\xi_a^2 + \xi_b^2}/c^2) + \arctan(s_2\sqrt{\xi_a^2 + \xi_b^2}/c).\end{aligned}$$

Similar to the proof for $k = 1$, we can conclude $s_3 \le \frac{1}{\rho}s_2$. By induction, $s_{k+1} \le \frac{1}{\rho}s_k$, thus completing the proof. □

8 Proof of Theorem 2. (Local Convergence of SCF)

With relaxed assumptions, we can show local convergence if the initial subspace is close enough to the true subspace **P**∗:

A3: Let μ_i^* denote $\arctan \lambda_i(A(\mathbf{P}^*), B(\mathbf{P}^*))$. There exists an $\eta > 0$ such that

$$\mu_p^* - \mu_{p+1}^* \geq \eta.$$

Theorem 2. ***(Local Convergence)*** *Let* $\hat{s}_0 = \|\sin\Theta(\boldsymbol{P}_0, \boldsymbol{P}^*)\|$. *Assume* ***A1*** *and* ***A3****, and* $\hat{s}_0\sqrt{\xi_a^2+\xi_b^2} < c$. *If*

$$\eta > \arcsin(\rho C\sqrt{\xi_a^2+\xi_b^2}/c^2) + \arctan(\hat{s}_0\sqrt{\xi_a^2+\xi_b^2}/c)$$

for some constant $\rho > 1$*, then SCF is locally convergent at* $\boldsymbol{P}^*$ *at the rate of* $\frac{1}{\rho}$.

Proof. By assumption **A1**, we have

$$\sqrt{\|A_0 - A^*\|^2 + \|B_0 - B^*\|^2} \leq \sqrt{\xi_a^2+\xi_b^2}\|\sin\Theta(\mathbf{P}_0, \mathbf{P}^*)\| = \hat{s}_0\sqrt{\xi_a^2+\xi_b^2}.$$

Applying Lemma 2, we have

$$\begin{aligned} |\mu_{p,0} - \mu_p^*| \leq & \arctan\frac{\sqrt{\|A_0 - A^*\|^2 + \|B_0 - B^*\|^2}}{c} \\ \leq & \arctan(\hat{s}_0\sqrt{\xi_a^2+\xi_b^2}/c),\ \forall 1 \leq p \leq n. \end{aligned}$$

By assumption **A3** it follows that

$$\begin{aligned} \mu_{p,0} - \mu_{p+1}^* = \mu_{p,0} - \mu_p^* + \mu_p^* - \mu_{p+1}^* \geq \eta - \arctan(\hat{s}_0\sqrt{\xi_a^2+\xi_b^2}/c) \\ > \arcsin(\rho C\sqrt{\xi_a^2+\xi_b^2}/c^2) \geq 0. \end{aligned} \tag{15}$$

Following the same procedures as in the proof for **Theorem 1**, we arrive that

$$\|\sin\Theta(\mathbf{P}_{k-1}, \mathbf{P}^*)\| \leq \frac{1}{\rho}\|\sin\Theta(\mathbf{P}_k, \mathbf{P}^*)\|,$$

where $\rho = \frac{c^2\sin(\mu_{p,0} - \mu_{p+1}^*)}{C\sqrt{\xi_a^2+\xi_b^2}}$. □

9 Proof of Corollary 1. (Convergence of WDA-eig)

Corollary 1. *Suppose that the optimal transport matrix $\boldsymbol{T}^{c,c'}$ satisfies a Lipschitz-like condition:*

$$\|\boldsymbol{T}^{c,c'}(\boldsymbol{P}_1) - \boldsymbol{T}^{c,c'}(\boldsymbol{P}_2)\| \le \xi^{c,c'}\|\sin\Theta(\boldsymbol{P}_1,\boldsymbol{P}_2)\|,$$

For a given p, let

$$\eta = \min_k\{\mu_{p,k} - \mu_{p+1,k}\}.$$

Denote $\xi_a = \sum_{c,c'>c}\xi^{c,c'}\|\sum_{i,j}(x_i^c - x_j^{c'})(x_i^c - x_j^{c'})^T\|$, $\xi_b = \sum_c \xi^{c,c}\|\sum_{i,j}(x_i^c - x_j^c)(x_i^c - x_j^c)^T\|$. If

$$\eta > \arcsin(\rho C\sqrt{\xi_a^2 + \xi_b^2}/c^2) + \arctan(s_1\sqrt{\xi_a^2 + \xi_b^2}/c)$$

*for some constant $\rho > 1$, then **wda-eig** converges linearly at the rate $\frac{1}{\rho}$.*

Proof. We first note that in **WDA-eig**, $\mathbf{C}_b$ and $\mathbf{C}_w$ are invariant to orthogonal transformation of $\mathbf{P}$, i.e., $\mathbf{C}_b(\mathbf{P}) \equiv \mathbf{C}_b(\mathbf{P}Q)$, $\mathbf{C}_w(\mathbf{P}) \equiv \mathbf{C}_w(\mathbf{P}Q)$ for any orthogonal matrix $Q \in \mathbb{R}^{p\times p}$, since

$$\begin{aligned}\mathbf{C}_b(\mathbf{P}) &= \sum_{c,c'>c}\sum_{i,j} t_{ij}^{c,c'}(\mathbf{P})(x_i^c - x_j^{c'})(x_i^c - x_j^{c'})^T,\\ \mathbf{C}_w(\mathbf{P}) &= \sum_{c}\sum_{i,j} t_{ij}^{c,c}(\mathbf{P})(x_i^c - x_j^{c})(x_i^c - x_j^{c})^T,\end{aligned}$$

and $\mathbf{T}^{c,c'}(\mathbf{P})$ and $\mathbf{T}^{c}(\mathbf{P})$ are invariant to orthogonal transformation of $\mathbf{P}$:

$$\begin{aligned}\mathbf{T}^{c,c'}(\mathbf{P}Q) &\triangleq \operatorname*{argmin}_{\mathbf{T}\in U_{n_c n_{c'}}} \lambda\langle \mathbf{T}, \mathbf{M}_{X^c\mathbf{P}Q, X^{c'}\mathbf{P}Q}\rangle + \sum_{i,j} t_{ij}\log(t_{ij})\\ &= \operatorname*{argmin}_{\mathbf{T}\in U_{n_c n_{c'}}} \lambda\sum_{i,j} t_{ij}\|(x_i^c - x_j^{c'})^T\mathbf{P}Q\| + \sum_{i,j} t_{ij}\log(t_{ij})\\ &= \operatorname*{argmin}_{\mathbf{T}\in U_{n_c n_{c'}}} \lambda\sum_{i,j} t_{ij}\|(x_i^c - x_j^{c'})^T\mathbf{P}\| + \sum_{i,j} t_{ij}\log(t_{ij}) = \mathbf{T}^{c,c'}(\mathbf{P}).\end{aligned}$$

By the assumption on $\mathbf{T}^{c,c'}$, $\mathbf{C}_b$ satisfies

$$\begin{aligned}\|\mathbf{C}_b(\mathbf{P}_1) - \mathbf{C}_b(\mathbf{P}_2)\| &= \|\sum_{c,c'>c}\sum_{i,j}(t_{ij}^{c,c'}(\mathbf{P}_1) - t_{ij}^{c,c'}(\mathbf{P}_2))(x_i^c - x_j^{c'})(x_i^c - x_j^{c'})^T)\|\\ &\le \sum_{c,c'>c}\max_{i,j}|t_{ij}^{c,c'}(\mathbf{P}_1) - t_{ij}^{c,c'}(\mathbf{P}_2)|\|\sum_{i,j}(x_i^c - x_j^{c'})(x_i^c - x_j^{c'})^T\|\\ &\le \sum_{c,c'>c}\xi^{c,c'}\|\sin\Theta(\mathbf{P}_1,\mathbf{P}_2)\|\|\sum_{i,j}(x_i^c - x_j^{c'})(x_i^c - x_j^{c'})^T\|\\ &\triangleq \xi_a\|\sin\Theta(\mathbf{P}_1,\mathbf{P}_2)\|.\end{aligned}$$

The last inequality holds since

$$\max_{i,j}|t_{ij}^{c,c'}(\mathbf{P}_1) - t_{ij}^{c,c'}(\mathbf{P}_2)| \le \|\mathbf{T}^{c,c'}(\mathbf{P}_1) - \mathbf{T}^{c,c'}(\mathbf{P}_2)\| \le \xi^{c,c'}\|\sin\Theta(\mathbf{P}_1,\mathbf{P}_2)\|.$$

Similarly,

$$\|\mathbf{C}_w(\mathbf{P}_1) - \mathbf{C}_w(\mathbf{P}_2)\| \le \sum_c \xi^{c,c}\|\sum_{i,j}(x_i^c - x_j^c)(x_i^c - x_j^c)^T\|\|\sin\Theta(\mathbf{P}_1,\mathbf{P}_2)\| \triangleq \xi_b\|\sin\Theta(\mathbf{P}_1,\mathbf{P}_2)\|.$$

For all iteration number k, $\mu_{p,k} - \mu_{p+1,k} \ge \eta$. Since **A1** and **A2** are satisfied, the result follows directly from **Theorem 1**. □

10 Sensitivity to Noisy Labels

For iterative subspace clustering, performing K-Means on the projected data may not render accurate labels in the first few iterations, especially if we initialize with random subspace. We therefore investigate how the subspace changes when we perturb the labels. The results in Table 3 illustrate the sensitivity to noisy labels of FDA (same as **WDA-eig** with $\lambda = 0$), **WDA-eig** ($\lambda = 1.0$) and local FDA (LFDA) [37] with the number of neighbors$= 1$. We use the simulated dataset introduced in the Main Paper, Section 4.1 and add noisy labels to the data. The first column of Table 3 shows the percentage of wrong labels added. The rest of the columns show the distance of the subspace $\mathbf{P}$ obtained by FDA/**WDA-eig**/LFDA under the noisy labels to the original subspaces $\mathbf{P}^*$ measured by $\|\sin\Theta(\mathbf{P},\mathbf{P}^*)\|$, where the original subspaces are approximated by the converged solution of FDA/**WDA-eig**/LFDA under true labels. The results are averaged over 20 trials. We observe that WDA is more robust to noisy labels than both FDA and local FDA.

Table 3: Sensitivity to Noisy Labels.

% wrong labels	FDA dist. to $\mathbf{P}^*$	WDA dist. to $\mathbf{P}^*$	LFDA dist. to $\mathbf{P}^*$
1%	0.21	**0.01**	0.04
5%	0.32	**0.02**	0.08
10%	0.59	**0.05**	0.11
20%	0.84	**0.07**	0.15