Journal of Machine Learning Research 18 (2018) 1-55 Submitted 7/15; Revised 1/18; Published 4/18

A Study of the Classification of Low-Dimensional Data with Supervised Manifold Learning

Elif Vural[*] VELIF@METU.EDU.TR
Department of Electrical and Electronics Engineering
Middle East Technical University
Ankara, 06800, Turkey

Christine Guillemot CHRISTINE.GUILLEMOT@INRIA.FR
Centre de Recherche INRIA Bretagne Atlantique
Campus Universitaire de Beaulieu
35042 Rennes, France

Editor: Gert Lanckriet

Abstract

Supervised manifold learning methods learn data representations by preserving the geometric structure of data while enhancing the separation between data samples from different classes. In this work, we propose a theoretical study of supervised manifold learning for classification. We consider nonlinear dimensionality reduction algorithms that yield linearly separable embeddings of training data and present generalization bounds for this type of algorithms. A necessary condition for satisfactory generalization performance is that the embedding allow the construction of a sufficiently regular interpolation function in relation with the separation margin of the embedding. We show that for supervised embeddings satisfying this condition, the classification error decays at an exponential rate with the number of training samples. Finally, we examine the separability of supervised nonlinear embeddings that aim to preserve the low-dimensional geometric structure of data based on graph representations. The proposed analysis is supported by experiments on several real data sets.

Keywords: Manifold learning, dimensionality reduction, classification, out-of-sample extensions, RBF interpolation

1. Introduction

In many data analysis problems, data samples have an intrinsically low-dimensional structure although they reside in a high-dimensional ambient space. The learning of low-dimensional structures in collections of data has been a well studied topic of the last two decades (Tenenbaum et al., 2000), (Roweis and Saul, 2000), (Belkin and Niyogi, 2003), (He and Niyogi, 2004), (Donoho and Grimes, 2003), (Zhang and Zha, 2005). Following these works, many classification methods have been proposed in the recent years to apply such manifold learning techniques to learn classifiers that are adapted to the geometric structure of low-dimensional data (Hua et al., 2012), (Yang et al., 2011), (Zhang et al., 2012), (Sugiyama, 2007), (Raducanu and Dornaika, 2012). The common approach in such works

∗. Most part of the work was performed while the first author was at INRIA.

is to learn a data representation that enhances the between-class separation while preserving the intrinsic low-dimensional structure of data. While many efforts have focused on the practical aspects of learning such supervised embeddings for training data, the generalization performance of these methods as supervised classification algorithms has not been investigated much yet. In this work, we aim to study nonlinear supervised dimensionality reduction methods and present performance bounds based on the properties of the embedding and the interpolation function used for generalizing the embedding.

Several supervised manifold learning methods extend the Laplacian eigenmaps algorithm (Belkin and Niyogi, 2003), or its linear variant LPP (He and Niyogi, 2004) to the classification problem. The algorithms proposed by Hua et al. (2012), Yang et al. (2011), Zhang et al. (2012) provide a supervised extension of the LPP algorithm and learn a linear projection that preserves the proximity of neighboring samples from the same class, while increasing the distance between nearby samples from different classes. The method by Sugiyama (2007) proposes an adaptation of the Fisher metric for linear manifold learning, which is in fact shown to be equivalent to the above methods by Yang et al. (2011), Zhang et al. (2012). In (Li et al., 2013), (Cui and Fan, 2012), (Wang and Chen, 2009), some other similar Fisher-based linear manifold learning methods are proposed. In (Raducanu and Dornaika, 2012) a method relying on a similar formulation as in (Hua et al., 2012), (Yang et al., 2011), (Zhang et al., 2012) is presented, which, however, learns a nonlinear embedding. The main advantage of linear dimensionality reduction methods over nonlinear ones is that the generalization of the learnt embedding to novel (initially unavailable) samples is straightforward. However, nonlinear manifold learning algorithms are more flexible as the possible data representations they can learn belong to a wider family of functions, e.g., one can always find a nonlinear embedding to make training samples from different classes linearly separable. On the other hand, when a nonlinear embedding is used, one must also determine a suitable interpolation function to generalize the embedding to new samples, and the choice of the interpolator is critical for the classification performance.

The common effort in all of these supervised dimensionality reduction methods is to learn an embedding that increases the separation between different classes, while preserving the geometric structure of data. It is interesting to note that supervised manifold learning methods achieve separability by reducing the dimension of data, while kernel methods in traditional classifiers achieve this by increasing the dimension of data. Meanwhile, making training data linearly separable in supervised manifold learning does not mean much only by itself. Assuming that the data are sampled from a continuous distribution (hence two samples coincide with 0 probability), it is almost always possible to separate a discrete set of samples from different classes with a nonlinear embedding, e.g., even with a simple embedding such as the one mapping each sample to a vector encoding its class label. What actually matters is how the embedding generalizes to test data, i.e., where the test samples will be mapped to in the low-dimensional domain of embedding and how well the performance will be. The generalization for test data is straightforward for kernel methods, it is determined by the underlying main algorithm. However, in nonlinear supervised manifold learning, this question has rather been overlooked so far. In this work we aim to fill this gap and look into the generalization capabilities of supervised manifold learning algorithms. We study the conditions that must be satisfied by the embedding of the training samples and the interpolation function for satisfactory generalization of the classifier. We then ex-

amine the rates of convergence of supervised manifold learning algorithms that satisfy these conditions.

In Section 2, we consider arbitrary supervised manifold learning algorithms that compute a linearly separable embedding of training samples. We study the generalization capability of such algorithms for two types of out-of-sample interpolation functions. We first consider arbitrary interpolation functions that are Lipschitz-continuous on the support of each class, and then focus on out-of-sample extensions with radial basis function (RBF) kernels, which is a popular family of interpolation functions. For both types of interpolators, we derive conditions that must be satisfied by the embedding of the training samples and the regularity of the interpolation function that generalizes the embedding to test samples, when a nearest neighbor or linear classifier is used in the low-dimensional domain of embedding. These conditions enforce the Lipschitz constant of the interpolator to be sufficiently small, in comparison with the separation margin between training samples from different classes in the low-dimensional domain of embedding. The practical value of these results resides in their implications about what must really be taken into account when designing a supervised dimensionality reduction algorithm: Achieving a good separation margin does not suffice by itself; the geometric structure must also be preserved so as to ensure that a sufficiently regular interpolator can be found to generalize the embedding to the whole ambient space. We then particularly consider Gaussian RBF kernels and show the existence of an optimal value for the kernel scale by studying the condition in our main result that links the separation with the Lipschitz constant of the kernel.

Our results in Section 2 also provide bounds on the rate of convergence of the classification error of supervised embeddings. We show that the misclassification error probability decays at an exponential rate with the number of samples, provided that the interpolation function is sufficiently regular with respect to the separation margin of the embedding. These convergence rates are higher than those reported in previous results on RBF networks (Niyogi and Girosi, 1996), (Lin et al., 2014), (Hernández-Aguirre et al., 2002), and regularized least-squares regression algorithms (Caponnetto and De Vito, 2007), (Steinwart et al., 2009). The essential difference between our results and such previous works is that those assume a general setting and do not focus on a particular data model, whereas our results are rather relevant to settings where the support of each class admits some certain structure, so as to allow the existence of an interpolator that is sufficiently regular on the support of each class. Moreover, in contrast with these previous works, our bounds are independent of the ambient space dimension and vary only with the intrinsic dimensions of the class supports as they characterize the error in terms of the covering numbers of the supports.

The results in Section 2 assume an embedding that makes training samples from different classes linearly separable. Even if most nonlinear dimensionality reduction methods are observed to yield separable embeddings in practice, we aim to verify this theoretically in Section 3. In particular, we focus on the nonlinear version of the supervised Laplacian eigenmaps embeddings (Raducanu and Dornaika, 2012), (Hua et al., 2012), (Yang et al., 2011), (Zhang et al., 2012). Supervised Laplacian eigenmaps methods embed the data with the eigenvectors of the linear combination of two graph Laplacian matrices that encode the links between neighboring samples from the same class and different classes. In such a data representation, the coordinates of neighboring data samples change slowly within the same

class and rapidly across different classes. We study the conditions for the linear separability of these embeddings and characterize their separation margin in terms of some graph and algorithm parameters.

In Section 4, we evaluate our results with experiments on several real data sets. We study the implications of the condition derived in Section 2 on the separability margin - interpolator regularity tradeoff. The experimental comparison of several supervised dimensionality reduction algorithms shows that this compromise between the separation and interpolator regularity can indeed be related to the practical classification performance of a supervised manifold learning algorithm. This suggests that, one can possibly improve the accuracy of supervised dimensionality reduction algorithms by considering more carefully the generalization capability of the embedding during the learning. We then study the variation of the classification performance with parameters such as the sample size, the RBF kernel scale, and the dimension of the embedding, in view of the generalization bounds presented in Section 2. Finally, we conclude in Section 5.

2. Performance Bounds for Supervised Manifold Learning Methods

2.1 Notation and Problem Formulation

Consider a setting with M data classes where the samples of each class $m \in \{1, \ldots, M\}$ are drawn from a probability measure ν_m in a Hilbert space H such that ν_m has a bounded support $\mathcal{M}_m \subset H$. Let $\mathcal{X} = \{x_i\}_{i=1}^N \subset H$ be a set of N training samples such that each x_i is drawn from one of the probability measures ν_m, and the samples drawn from each ν_m are independent and identically distributed. We denote the class label of x_i by $C_i \in \{1, 2, \ldots, M\}$.

Let $Y = \{y_i\}_{i=1}^N \subset \mathbb{R}^d$ be a d-dimensional embedding of $\mathcal{X}$, where each y_i corresponds to x_i. We consider supervised embeddings such that Y is linearly separable. Linear separability is defined as follows:

Definition 1 *The data representation Y is linearly separable with a margin of $\gamma > 0$, if for any two classes $k, l \in \{1, 2, \ldots, M\}$, there exists a separating hyperplane defined by $\omega_{kl} \in \mathbb{R}^d$, $\|\omega_{kl}\| = 1$ and $b_{kl} \in \mathbb{R}$ such that*

$$\begin{aligned} \omega_{kl}^T y_i + b_{kl} &\geq \gamma/2 \qquad \text{if } C_i = k \\ \omega_{kl}^T y_i + b_{kl} &\leq -\gamma/2 \qquad \text{if } C_i = l. \end{aligned} \tag{1}$$

The above definition of separability implies the following. For any given class m, there exists a set of hyperplanes $\{\omega_{mk}\}_{k \neq m} \subset \mathbb{R}^d$, $\|\omega_{mk}\| = 1$, and a set of real numbers $\{b_{mk}\}_{k \neq m} \subset \mathbb{R}$ that separate class m from other classes, such that for all y_i of class $C_i = m$

$$\omega_{mk}^T y_i + b_{mk} > \gamma/2, \quad \forall k \neq m \tag{2}$$

and for all y_i of class $C_i \neq m$, there exists a k such that

$$\omega_{mk}^T y_i + b_{mk} < -\gamma/2. \tag{3}$$

These hyperplanes are obtained by setting $\omega_{km} = -\omega_{mk}$, $b_{km} = -b_{mk}$.

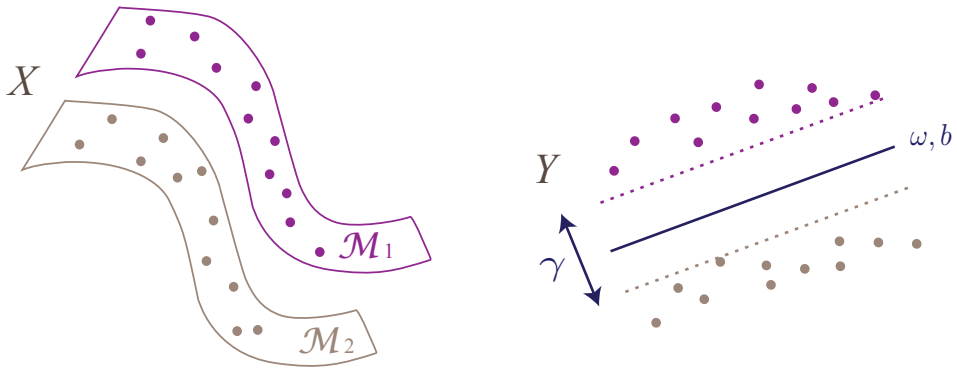


Figure 1: Illustration of a linearly separable embedding. Data in X are sampled from two different classes with supports $\mathcal{M}_1$, $\mathcal{M}_2$. The samples X are mapped to the coordinates Y with a low-dimensional embedding, where the two classes become linearly separable with margin γ with the hyperplane given by ω, b.

Figure 1 shows an illustration of a linearly separable embedding of data samples from two classes. Manifold learning methods typically compute a low-dimensional embedding Y of training data $\mathcal{X}$ in a pointwise manner, i.e., the coordinates y_i are computed only for the initially available training samples x_i. However, in a classification problem, in order to estimate the class label of a new data sample x of unknown class, x needs to be mapped to the low-dimensional domain of embedding as well. The construction of a function $f : H \to \mathbb{R}^d$ that generalizes the learnt embedding to the whole space is known as the out-of-sample generalization problem. Smooth functions are commonly used for out-of-sample interpolation, e.g. as in (Qiao et al., 2013), (Peherstorfer et al., 2011).

Now let x be a test sample drawn from the probability measure ν_m, hence, the true class label of x is m. In our study, we consider two basic classification schemes in the domain of embedding:

Linear classifier. The embeddings of the training samples are used to compute the separating hyperplanes, i.e., the classifier parameters $\{\omega_{mk}\}$ and $\{b_{mk}\}$. Then, mapping x to the low-dimensional domain as $f(x) \in \mathbb{R}^d$, the class label of x is estimated as $\hat{C}(x) = l$ if there exists $l \in \{1, \ldots, M\}$ such that

$$\omega_{lk}^T f(x) + b_{lk} > 0, \quad \forall k \in \{1, \ldots, M\} \setminus \{l\}. \tag{4}$$

Note that the existence of such an l is not guaranteed in general for any x, but for a given x there cannot be more than one l satisfying the above condition. Then x is classified correctly if the estimated class label agrees with the true class label, i.e., $\hat{C}(x) = l = m$.

Nearest neighbor classification. The test sample x is assigned the class label of the closest training point in the domain of embedding, i.e., $\hat{C}(x) = C_{i'}$, where

$$i' = \arg \min_{i=1,\ldots,N} \|y_i - f(x)\|.$$

In the rest of this section, we study the generalization performance of supervised dimensionality reduction methods. We first consider in Section 2.2 interpolation functions that

vary regularly on each class support and we search for a lower bound on the probability of correctly classifying a new data sample in terms of the regularity of f, the separation of the embedding, and the sampling density. Then in Section 2.3, we study the classification performance for a particular type of interpolation functions, namely RBF interpolators, which is one of the most popular ones (Peherstorfer et al., 2011), (Chin and Suter, 2008). We focus particularly on Gaussian RBF interpolators in Section 2.4 and derive some results regarding the existence of an optimal kernel scale parameter. Lastly, we discuss our results in comparison with previous literature in Section 2.5.

In the results in Sections 2.2-2.4, we keep a generic formulation and simply treat the supports $\{\mathcal{M}_m\}$ as arbitrary bounded subsets of H, each of which represents a different data class. Nevertheless, from the perspective of manifold learning, our results are of interest especially when the data is assumed to have an underlying low-dimensional structure. In Section 2.5, we study the implications of our results for the setting where $\mathcal{M}_m$ are low-dimensional manifolds. We then examine how the proposed bounds vary in relation to the intrinsic dimensions of $\{\mathcal{M}_m\}$.

2.2 Out-of-Sample Interpolation with Regular Functions

Let $f : H \to \mathbb{R}^d$ be an out-of-sample interpolation function such that $f(x_i) = y_i$ for each training sample x_i, $i = 1, \ldots, N$. Assume that f is Lipschitz continuous with constant $L > 0$ when restricted to any one of the supports $\mathcal{M}_m$; i.e., for any $m \in \{1, \ldots, M\}$ and any $u, v \in \mathcal{M}_m$

$$\|f(u) - f(v)\| \leq L \|u - v\|,$$

where $\|\cdot\|$ denotes above the ℓ_2-norm if the argument is in $\mathbb{R}^d$, and the norm induced from the inner product in H if the argument is in H.

We will find a relation between the classification accuracy and the number of training samples via the covering number of the supports $\mathcal{M}_m$. Let $B_\epsilon(x) \subset H$ denote an open ball of radius ϵ around x

$$B_\epsilon(x) = \{u \in H : \|x - u\| < \epsilon\}.$$

The covering number $\mathcal{N}(\epsilon, A)$ of a set $A \subset H$ is defined as the smallest number of open balls B_ϵ of radius ϵ whose union contains A (Kulkarni and Posner, 1995)

$$\mathcal{N}(\epsilon, A) = \inf\{k : \exists\, u_1, \ldots, u_k \in H \text{ s.t. } A \subset \bigcup_{i=1}^{k} B_\epsilon(u_i)\}.$$

We assume that the supports $\mathcal{M}_m$ are totally bounded, i.e., $\mathcal{M}_m$ has a finite covering number $\mathcal{N}(\epsilon, \mathcal{M}_m)$ for any $\epsilon > 0$.

We state below a lower bound for the probability of correctly classifying a sample x drawn from ν_m, in terms of the number of training samples drawn from ν_m, the separation of the embedding and the regularity of f.

Theorem 2 *For some ϵ with $0 < \epsilon \leq \gamma/(2L)$, let the training set $\mathcal{X}$ contain at least N_m samples drawn i.i.d. according to a probability measure ν_m such that*

$$N_m \geq \mathcal{N}(\epsilon/2, \mathcal{M}_m).$$

Let Y be an embedding of the training samples $\mathcal{X}$ that is linearly separable with margin larger than γ, and let f be an interpolation function that is Lipschitz continuous with constant L on the support $\mathcal{M}_m$. Then the probability of correctly classifying a test sample x drawn from ν_m independently from the training samples with the linear classifier (4) *is lower bounded as*

$$P\left(\hat{C}(x) = m\right) \geq 1 - \frac{\mathcal{N}(\epsilon/2, \mathcal{M}_m)}{2N_m}.$$

The proof of the theorem is given in Appendix A.1. Theorem 2 establishes a link between the classification performance and the separation of the embedding of the training samples. In particular, due to the condition $\epsilon \leq \gamma/(2L)$, the increase in the separation γ allows a larger value for ϵ, provided that the interpolator regularity is not affected much. This reduces the covering number $\mathcal{N}(\epsilon/2, \mathcal{M}_m)$ in return and increases the probability of correct classification. Similarly, from the condition $\epsilon \leq \gamma/(2L)$, one can also observe that at a given separation γ, a smaller Lipschitz constant L for the interpolation function allows the parameter ϵ to take a larger value. This reduces the covering number $\mathcal{N}(\epsilon/2, \mathcal{M}_m)$ and therefore increases the correct classification probability. Thus, choosing a more regular interpolator at a given separation helps improve the classification performance. If the ϵ parameter is fixed, the Lipschitz constant of the interpolator is allowed to increase only proportionally to the separation margin. The condition that the interpolator must be sufficiently regular in comparison with the separation suggests that increasing the separation too much at the cost of impairing the interpolator regularity may degrade the classifier performance. In the case that the supports $\mathcal{M}_m$ are low-dimensional manifolds, the covering number $\mathcal{N}(\epsilon/2, \mathcal{M}_m)$ increases at a geometric rate with the intrinsic dimension D of the manifold, since a D-dimensional manifold is locally homeomorphic to $\mathbb{R}^D$. Therefore, from the condition on the number of samples, N_m should increase at a geometric rate with D.

In Theorem 2 the probability of misclassification decreases with the number N_m of training samples at a rate of $O(N_m^{-1})$. In the rest of this section, we show that it is in fact possible to obtain an exponential convergence rate with linear and NN-classifiers under certain assumptions. We first present the following lemma.

Lemma 3 *Let $\mathcal{X} = \{x_i\}_{i=1}^N \subset H$ be a set of training samples such that each x_i is drawn i.i.d. from one of the probability measures $\{\nu_m\}_{m=1}^M$. Let x be a test sample randomly drawn according to the probability measure ν_m of class m. Let*

$$A = \{x_i \in \mathcal{X} : x_i \in B_\delta(x), x_i \sim \nu_m\} \tag{5}$$

be the set of samples in $\mathcal{X}$ that are in a δ-neighborhood of x and also drawn from the measure ν_m. Assume that A contains $|A| = Q$ samples. Then

$$P\left(\|f(x) - \frac{1}{Q}\sum_{x_j \in A} f(x_j)\| \leq L\delta + \sqrt{d}\epsilon\right) \geq 1 - 2d\exp\left(-\frac{Q\,\epsilon^2}{2L^2\delta^2}\right). \tag{6}$$

Lemma 3 is proved in Appendix A.2. The inequality in (6) shows that as the number Q of training samples falling in a neighborhood of a test point x increases, the probability of the deviation of $f(x)$ from its average within the neighborhood decreases. The parameter ϵ captures the relation between the amount and the probability of deviation.

When studying the classification accuracy in the main result below, we will use the following generalized definition of the linear separation.

Definition 4 *Let Y be a linearly separable embedding with margin γ such that each pair (k,l) of classes are separated with the hyperplanes given by ω_{kl}, b_{kl} as defined in Definition 1. We say that the linear classifier given by $\{\omega_{kl}\}$, $\{b_{kl}\}$ has a Q-mean separability margin of $\gamma_Q > 0$ if any choice of Q samples $\{y_{k,i}\}_{i=1}^{Q} \subset Y$ from class k and Q samples $\{y_{l,i}\}_{i=1}^{Q} \subset Y$ from class l, $l \neq k$, satisfies*

$$\begin{aligned} \omega_{kl}^T \left(\frac{1}{Q} \sum_{i=1}^{Q} y_{k,i} \right) + b_{kl} &\geq \gamma_Q/2 \\ \omega_{kl}^T \left(\frac{1}{Q} \sum_{i=1}^{Q} y_{l,i} \right) + b_{kl} &\leq -\gamma_Q/2. \end{aligned} \tag{7}$$

The above definition of separability is more flexible than the one in Definition 1. Clearly, an embedding that is linearly separable with margin γ has a Q-mean separability margin of $\gamma_Q \geq \gamma$ for any Q. As in the previous section, we consider that the test sample x is classified with the linear classifier (4) in the low-dimensional domain, defined with respect to the set of hyperplanes given by $\{\omega_{mk}\}$ and $\{b_{mk}\}$ as in (2) and (3).

In the following result, we show that an exponential convergence rate can be obtained with linear classifiers in supervised manifold learning. We define beforehand a parameter depending on δ, which gives the smallest possible measure of the δ-neighborhood $B_\delta(x)$ of a point x in support $\mathcal{M}_m$.

$$\eta_{m,\delta} := \inf_{x \in \mathcal{M}_m} \nu_m(B_\delta(x)).$$

Theorem 5 *Let $\mathcal{X} = \{x_i\}_{i=1}^{N} \subset H$ be a set of training samples such that each x_i is drawn i.i.d. from one of the probability measures $\{\nu_m\}_{m=1}^{M}$. Let Y be an embedding of $\mathcal{X}$ in $\mathbb{R}^d$ that is linearly separable with a Q-mean separability margin larger than γ_Q. For a given $\epsilon > 0$ and $\delta > 0$, let f be a Lipschitz-continuous interpolator such that*

$$L\delta + \sqrt{d}\epsilon \leq \frac{\gamma_Q}{2}. \tag{8}$$

Consider a test sample x randomly drawn according to the probability measure ν_m of class m. If $\mathcal{X}$ contains at least N_m training samples drawn i.i.d. from ν_m such that

$$N_m > \frac{Q}{\eta_{m,\delta}},$$

then the probability of correctly classifying x with the linear classifier given in (4) is lower bounded as

$$P\left(\hat{C}(x) = m\right) \geq 1 - \exp\left(-\frac{2\left(N_m\,\eta_{m,\delta} - Q\right)^2}{N_m}\right) - 2d\exp\left(-\frac{Q\,\epsilon^2}{2L^2\delta^2}\right). \tag{9}$$

Theorem 5 is proved in Appendix A.3. The theorem shows how the classification accuracy is influenced by the separation of the classes in the embedding, the smoothness of the out-of-sample interpolant, and the number of training samples drawn from the density of each class. The condition in (8) points to the tradeoff between the separation and the regularity of the interpolation function. As the Lipschitz constant L of the interpolation function f increases, f becomes less "regular", and a higher separation γ_Q is needed to meet the condition. This is coherent with the expectation that, when f becomes irregular, the classifier becomes more sensitive to the perturbations of the data, e.g., due to noise. The requirement of a higher separation is then for ensuring a larger margin in the linear classifier, which compensates for the irregularity of f. From (8), it is also observed that the separation should increase with the dimension d as well, and also with ϵ, whose increase improves the confidence of the bound (9). Note that the condition in (8) implies also the following: When computing an embedding, it is not advisable to increase the separation of training data unconditionally. In particular, increasing the separation too much may violate the preservation of the geometry and yield an irregular interpolator. Hence, when designing a supervised dimensionality reduction algorithm, one must pay attention to the regularity of the resulting interpolator as much as the enhancement of the separation margin.

Next, we discuss the roles of the parameters Q and δ. The term $\exp(-Q\,\epsilon^2/(2L^2\delta^2))$ in the correct classification probability bound (9) shows that, for fixed δ, the confidence increases with the value of Q. Meanwhile, due to the numerator of the term $\exp(-2\,(N_m\,\eta_{m,\delta}-Q)^2/N_m)$, for a high confidence, the number of samples N_m should also be relatively big with respect to Q to have a high overall confidence. Similarly, at fixed Q, δ should be made smaller to increase the confidence due to the term $\exp(-(Q\,\epsilon^2)/(2L^2\delta^2))$, which then reduces the parameter $\eta_{m,\delta}$ and eventually requires the number of samples N_m to take a sufficiently large value in order to make the term $\exp(-2\,(N_m\,\eta_{m,\delta}-Q)^2/N_m)$ small and have a high confidence. Therefore, these two parameters Q and δ behave in a similar way, and determine the relation between the number of samples and the correct classification probability, i.e., they indicate how large N_m should be in order to have a certain confidence of correct classification.

Theorem 5 studies the setting where the class labels are estimated with a linear classifier in the domain of embedding. We also provide another result below that analyses the performance when a nearest-neighbor classifier is used in the domain of embedding.

Theorem 6 *Let $\mathcal{X} = \{x_i\}_{i=1}^N \subset H$ be a set of training samples such that each x_i is drawn i.i.d. from one of the probability measures $\{\nu_m\}_{m=1}^M$. Let Y be an embedding of $\mathcal{X}$ in $\mathbb{R}^d$ such that*

$$\begin{aligned}&\|y_i - y_j\| < D_\delta, \ \text{if } \|x_i - x_j\| \le \delta \ \text{and } C_i = C_j\\ &\|y_i - y_j\| > \gamma, \quad \text{if } C_i \neq C_j,\end{aligned}$$

hence, nearby samples from the same class are mapped to nearby points, and samples from different classes are separated by a distance of at least γ in the embedding.

For given $\epsilon > 0$ and $\delta > 0$, let f be a Lipschitz-continuous interpolation function such that

$$L\delta + \sqrt{d}\epsilon + D_{2\delta} \le \frac{\gamma}{2}. \tag{10}$$

Consider a test sample x randomly drawn according to the probability measure ν_m of class m. If $\mathcal{X}$ contains at least N_m training samples drawn i.i.d. from ν_m such that

$$N_m > \frac{Q}{\eta_{m,\delta}},$$

then the probability of correctly classifying x with nearest-neighbor classification in $\mathbb{R}^d$ is lower bounded as

$$P\left(\hat{C}(x) = m\right) \geq 1 - \exp\left(-\frac{2\left(N_m\,\eta_{m,\delta} - Q\right)^2}{N_m}\right) - 2d\exp\left(-\frac{Q\,\epsilon^2}{2L^2\delta^2}\right). \tag{11}$$

Theorem 6 is proved in Appendix A.4. Theorem 6 is quite similar to Theorem 5 and can be interpreted similarly. Unlike in the previous result, the separability condition of the embedding is based on the pairwise distances of samples from different classes here. The condition (10) suggests that the result is useful when the parameter $D_{2\delta}$ is sufficiently small, which requires the embedding to map nearby samples from the same class in the ambient space to nearby points.

In this section, we have characterized the regularity of the interpolation functions via their rates of variation when restricted to the supports $\mathcal{M}_m$. While the results of this section are generic in the sense that they are valid for any interpolation function with the described regularity properties, we have not examined the construction of such functions. In a practical classification problem where one uses a particular type of interpolation functions, one would also be interested in the adaptation of these results to obtain performance guarantees for the particular type of function used. Hence, in the following section we focus on a popular family of smooth functions; radial basis function (RBF) interpolators, and study the classification performance of this particular type of interpolators.

2.3 Out-of-Sample Interpolation with RBF Interpolators

Here we consider an RBF interpolation function $f : H \to \mathbb{R}^d$ of the form

$$f(x) = [f^1(x)\, f^2(x)\, \ldots f^d(x)],$$

such that each component f^k of f is given by

$$f^k(x) = \sum_{i=1}^{N} c_i^k\, \phi(\|x - x_i\|),$$

where $\phi : \mathbb{R} \to \mathbb{R}^+$ is a kernel function, $c_i^k \in \mathbb{R}$ are coefficients, and x_i are kernel centers. In interpolation with RBF functions, it is common to choose the set of kernel centers as the set of available data samples. Hence, we assume that the set of kernel centers $\{x_i\}_{i=1}^N$ is selected to be the same as the set of training samples $\mathcal{X}$. We consider a setting where the coefficients c_i^k are set such that $f(x_i) = y_i$, i.e., f maps each training point in $\mathcal{X}$ to its embedding previously computed with supervised manifold learning.

We consider the RBF kernel ϕ to be a Lipschitz continuous function with constant $L_\phi > 0$, hence, for any $u, v \in \mathbb{R}$

$$|\phi(u) - \phi(v)| \leq L_\phi\, |u - v|.$$

Also, let $\mathcal{C}$ be an upper bound on the coefficient magnitudes such that for all $k = 1, \ldots, d$

$$\sum_{i=1}^{N} |c_i^k| \leq \mathcal{C}.$$

In the following, we analyze the classification accuracy and extend the results in Section 2.2 to the case of RBF interpolators. We first give the following result, which probabilistically bounds how much the value of the interpolator f at a point x randomly drawn from ν_m may deviate from the average interpolator value of the training points of the same class within a neighborhood of x.

Lemma 7 *Let $\mathcal{X} = \{x_i\}_{i=1}^N \subset H$ be a set of training samples such that each x_i is drawn i.i.d. from one of the probability measures $\{\nu_m\}_{m=1}^M$. Let x be a test sample randomly drawn according to the probability measure ν_m of class m. Let*

$$A = \{x_i \in \mathcal{X} : x_i \in B_\delta(x), x_i \sim \nu_m\} \tag{12}$$

be the set of samples in $\mathcal{X}$ that are in a δ-neighborhood of x and also drawn from the measure ν_m. Assume that A contains $|A| = Q$ samples. Then

$$P\left(\|f(x) - \frac{1}{Q}\sum_{x_j \in A} f(x_j)\| \leq \sqrt{d}\mathcal{C}(L_\phi \delta + \epsilon)\right) \geq 1 - 2N \exp\left(-\frac{(Q-1)\,\epsilon^2}{2L_\phi^2 \delta^2}\right). \tag{13}$$

The proof of Lemma 7 is given in Appendix A.5. The lemma states a result similar to the one in Lemma 3; however, is specialized to the case where f is an RBF interpolator.

We are now ready to present the following main result.

Theorem 8 *Let $\mathcal{X} = \{x_i\}_{i=1}^N \subset H$ be a set of training samples such that each x_i is drawn i.i.d. from one of the probability measures $\{\nu_m\}_{m=1}^M$. Let Y be an embedding of $\mathcal{X}$ in $\mathbb{R}^d$ that is linearly separable with a Q-mean separability margin larger than γ_Q. For a given $\epsilon > 0$ and $\delta > 0$, let f be an RBF interpolator such that*

$$\sqrt{d}\,\mathcal{C}\,(L_\phi \delta + \epsilon) \leq \frac{\gamma_Q}{2}. \tag{14}$$

Consider a test sample x randomly drawn according to the probability measure ν_m of class m. If $\mathcal{X}$ contains at least N_m training samples drawn i.i.d. from ν_m such that

$$N_m > \frac{Q}{\eta_{m,\delta}},$$

then the probability of correctly classifying x with the linear classifier given in (4) is lower bounded as

$$P\left(\hat{C}(x) = m\right) \geq 1 - \exp\left(-\frac{2\,(N_m\,\eta_{m,\delta} - Q)^2}{N_m}\right) - 2N \exp\left(-\frac{(Q-1)\,\epsilon^2}{2L_\phi^2 \delta^2}\right). \tag{15}$$

The theorem is proved in Appendix A.6. The theorem bounds the classification accuracy in terms of the smoothness of the RBF interpolation function and the number of samples. The condition in (14) characterizes the compromise between the separation and the regularity of the interpolator, which depends on the Lipschitz constant of the RBF kernels and the coefficient magnitude. As the Lipschitz constant L_ϕ and the coefficient magnitude parameter $\mathcal{C}$ increase (i.e., f becomes less "regular"), a higher separation γ_Q is required to provide a performance guarantee. When the separation margin of the embedding and the interpolator satisfy the condition in (14), the misclassification probability decays exponentially as the number of training samples increases, similarly to the results in Section 2.2.

Theorem 8 studies the misclassification probability when the class labels in the low-dimensional domain are estimated with a linear classifier. We also present below a bound on the misclassification probability when the nearest-neighbor classifier is used in the low-dimensional domain.

Theorem 9 *Let $\mathcal{X} = \{x_i\}_{i=1}^N \subset H$ be a set of training samples such that each x_i is drawn i.i.d. from one of the probability measures $\{\nu_m\}_{m=1}^M$. Let Y be an embedding of $\mathcal{X}$ in $\mathbb{R}^d$ such that*

$$\begin{aligned} &\|y_i - y_j\| < D_\delta, \text{ if } \|x_i - x_j\| \leq \delta \text{ and } C_i = C_j \\ &\|y_i - y_j\| > \gamma, \quad \text{if } C_i \neq C_j. \end{aligned}$$

For given $\epsilon > 0$ and $\delta > 0$, let f be an RBF interpolator such that

$$\sqrt{d}\,\mathcal{C}\,(L_\phi \delta + \epsilon) + D_{2\delta} \leq \frac{\gamma}{2}. \tag{16}$$

Consider a test sample x randomly drawn according to the probability measure ν_m of class m. If $\mathcal{X}$ contains at least N_m training samples drawn i.i.d. from ν_m such that

$$N_m > \frac{Q}{\eta_{m,\delta}},$$

then the probability of correctly classifying x with nearest-neighbor classification in $\mathbb{R}^d$ is lower bounded as

$$P\left(\hat{C}(x) = m\right) \geq 1 - \exp\left(-\frac{2\,(N_m\,\eta_{m,\delta} - Q)^2}{N_m}\right) - 2N \exp\left(-\frac{(Q-1)\,\epsilon^2}{2L_\phi^2 \delta^2}\right). \tag{17}$$

Theorem 9 is proved in Appendix A.7. While it provides the exact convergence rate as in Theorem 8, the necessary condition in (16) includes also the parameter $D_{2\delta}$. Hence, if the embedding maps nearby samples from the same class to nearby points, and a compromise is achieved between the separation and the interpolator regularity, the misclassification probability can be upper bounded.

2.4 Optimizing the Scale of Gaussian RBF Kernels

In data interpolation with RBFs, it is known that the accuracy of interpolation is quite sensitive to the choice of the shape parameter for several kernels including the Gaussian kernel (Baxter, 1992). The relation between the shape parameter and the performance of interpolation has been an important problem of interest (Piret, 2007). In this section, we focus on the Gaussian RBF kernel, which is a popular choice for RBF interpolation due to its smoothness and good spatial localization properties. We study the choice of the scale parameter of the kernel within the context of classification.

We consider the RBF kernel given by

$$\phi(r) = e^{-\frac{r^2}{\sigma^2}},$$

where σ is the scale parameter of the Gaussian function. We focus on the condition (14) in Theorem 8

$$\sqrt{d}\,\mathcal{C}\,(L_\phi \delta + \epsilon) \leq \gamma_Q/2,$$

(or equivalently the condition (16) if the nearest neighbor classifier is used), which relates the interpolation function properties with the separation. In particular, for a given separation margin, this condition is satisfied more easily when the term on the left hand side of the inequality is smaller. Thus, in the following, we derive an expression for the left hand side of the above inequality by deriving the Lipschitz constant L_ϕ and the coefficient bound $\mathcal{C}$ in terms of the scale parameter σ of the Gaussian kernel. We then study the scale parameter that minimizes $\sqrt{d}\,\mathcal{C}\,(L_\phi \delta + \epsilon)$.

Writing the condition $f(x_i) = y_i$ in a matrix form for each dimension $k = 1, \ldots, d$, we have

$$\Phi c^k = y^k, \tag{18}$$

where $\Phi \in \mathbb{R}^{N \times N}$ is a matrix whose (i,j)-th entry is given by $\Phi_{ij} = \phi(\|x_i - x_j\|)$, $c^k \in \mathbb{R}^{N \times 1}$ is the coefficient vector whose i-th entry is c_i^k, and $y^k \in \mathbb{R}^{N \times 1}$ is the data coordinate vector giving the k-th dimensions of the embeddings of all samples, i.e., $y_i^k = Y_{ik}$. Assuming that the embedding is computed with the usual scale constraint $Y^T Y = I$, we have $\|y^k\| = 1$. The norm of the coefficient vector can then be bounded as

$$\|c^k\| \leq \|\Phi^{-1}\| \|y^k\| = \|\Phi^{-1}\|. \tag{19}$$

In the rest of this section, we assume that the data $\mathcal{X}$ are sampled from the Euclidean space, i.e., $H = \mathbb{R}^n$. We first use a result by Narcowich et al. (1994) in order to bound the norm $\|\Phi^{-1}\|$ of the inverse matrix. From (Narcowich et al., 1994, Theorem 4.1) we get[1]

$$\|\Phi^{-1}\| \leq \beta\, \sigma^{-n} e^{\alpha \sigma^2}, \tag{20}$$

where $\alpha > 0$ and $\beta > 0$ are constants depending on the dimension n and the minimum distance between the training points $\mathcal{X}$ (separation radius) (Narcowich et al., 1994). As the

1. The result stated in (Narcowich et al., 1994, Theorem 4.1) is adapted to our study by taking the measure as $\beta(\rho) = \delta(\rho - \rho_0)$ so that the RBF kernel defined in (Narcowich et al., 1994, (1.1)) corresponds to a Gaussian function as $F(r) = \exp(-\rho_0\, r^2)$. The scale of the Gaussian kernel is then given by $\sigma = {\rho_0}^{-1/2}$.

ℓ_1-norm of the coefficient vector can be bounded as $\|c^k\|_1 \leq \sqrt{N}\|c^k\|$, from (19) one can set the parameter $\mathcal{C}$ that upper bounds the coefficients magnitudes as

$$\mathcal{C} = a\sigma^{-n}e^{\alpha\sigma^2},$$

where $a = \beta\sqrt{N}$.

Next, we derive a Lipschitz constant for the Gaussian kernel $\phi(r)$ in terms of σ. Setting the second derivative of ϕ to zero

$$\frac{d^2\phi}{dr^2} = e^{-\frac{r^2}{\sigma^2}}\left(\frac{4r^2}{\sigma^4} - \frac{2}{\sigma^2}\right) = 0,$$

we get that the maximum value of $|d\phi/dr|$ is attained at $r = \sigma/\sqrt{2}$. Evaluating $|d\phi/dr|$ at this value, we obtain

$$L_\phi = \sqrt{2}e^{-\frac{1}{2}}\sigma^{-1}.$$

Now rewriting the condition (14) of the theorem, we have

$$\sqrt{d}\,\mathcal{C}\,(L_\phi\delta + \epsilon) = a_1\sigma^{-n-1}e^{\alpha\sigma^2} + a_2\sigma^{-n}e^{\alpha\sigma^2} \leq \gamma_Q/2,$$

where $a_1 = \sqrt{2d}\,a\,e^{-1/2}\delta$ and $a_2 = \sqrt{d}\,a\,\epsilon$. We thus determine the Gaussian scale parameter σ that minimizes

$$F(\sigma) = a_1\sigma^{-n-1}e^{\alpha\sigma^2} + a_2\sigma^{-n}e^{\alpha\sigma^2}.$$

First, notice that as $\sigma \to 0$ and $\sigma \to \infty$, the function $F(\sigma) \to \infty$. Therefore, it has at least one minimum. Setting

$$\frac{dF}{d\sigma} = e^{\alpha\sigma^2}\sigma^{-n-2}\big(2\alpha a_2\sigma^3 + 2\alpha a_1\sigma^2 - a_2 n\sigma - a_1(n+1)\big) = 0,$$

we need to solve

$$2\alpha a_2\sigma^3 + 2\alpha a_1\sigma^2 - a_2 n\sigma - a_1(n+1) = 0. \tag{21}$$

The leading and the second-degree coefficients are positive, while the first-degree and the constant coefficients are negative in the above cubic polynomial. Then, the sum of the roots is negative and the product of the roots is positive. Therefore, there is one and only one positive root σ_{opt}, which is the unique minimizer of $F(\sigma)$.

The existence of an optimal scale parameter $0 < \sigma_{opt} < \infty$ for the RBF kernel can be intuitively explained as follows. When σ takes too small values, the support of the RBF function concentrated around the training points does not sufficiently cover the whole class supports $\mathcal{M}_m$. This manifests itself in (14) with the increase in the term L_ϕ, which indicates that the interpolation function is not sufficiently regular. This weakens the guarantee that a test sample will be interpolated sufficiently close to its neighboring training samples from the same class and mapped to the correct side of the hyperplane in the linear classifier. On the other hand, when σ increases too much, the stability of the linear system (18) is impaired and the coefficients c increase too much. This results in an overfitting of the interpolator and, therefore, decreases the classification performance. Hence, the analysis in this section provides a theoretical justification of the common knowledge that σ should be set to a sufficiently large value while avoiding overfitting.

Remark: It is also interesting to observe how the optimal scale parameter changes with the number of samples N. In the study (Narcowich et al., 1994), the constants α and β in (20) are shown to vary with the separation radius q at rates $\alpha = O(q^{-2})$ and $\beta = O(q^n)$, where the separation radius q is proportional to the smallest distance between two distinct training samples. Then a reasonable assumption is that the separation radius q should typically decrease at rate $O(N^{-1/n})$ as N increases. Using this relation, we get that α and β should vary at rates $\alpha = O(N^{2/n})$ and $\beta = O(N^{-1})$ with N. It follows that $a = \beta\sqrt{N} = O(N^{-1/2})$, and the parameters a_1, a_2 of the cubic polynomial in (21) also vary with N at rates $a_1 = O(N^{-1/2})$, $a_2 = O(N^{-1/2})$. The equation (21) in σ can then be rearranged as

$$b_3\sigma^3 + b_2\sigma^2 - b_1\sigma - b_0 = 0,$$

such that the constants vary with N at rates $b_3 = O(N^{2/n})$, $b_2 = O(N^{2/n})$, $b_1 = O(1)$, $b_0 = O(1)$. We can then inspect how the roots of this equation change with N as N increases. Since b_3 and b_2 dominate the other coefficients for large N, three real roots will exist if N is sufficiently large, two of which are negative and one is positive. The sum of the pairwise products of the roots is negative and it decays with N at rate $O(N^{-2/n})$, and the product of the roots also decays with N. Then at least two of the roots must decay with N. Meanwhile, the sum of the three roots is $O(1)$ and negative. This shows that one of the negative roots is $O(1)$, i.e., does not decay with N. From the product of three roots, we then observe that the product of the two decaying roots is $O(N^{-2/n})$. However, their sum also decays at the same rate (from the sum of the pairwise products), which is possible if their dominant terms have the same rate and cancel each other. We conclude that both of the decaying roots vary at rate $O(N^{-1/n})$, one of which is the positive root and the optimal value σ_{opt} of the scale parameter.

This analysis shows that the scale parameter of the Gaussian kernel should be adapted to the number of training samples, and a smaller kernel scale must be preferred for a larger number of training samples. In fact, the relation $\sigma_{opt} = O(N^{-1/n})$ is quite intuitive, as the average or typical distance between two samples will also decrease at rate $O(N^{-1/n})$ as the number of samples N increases in an n-dimensional space. Then the above result simply suggests that the kernel scale should be chosen as proportional to the average distance between the training samples.

2.5 Discussion of the Results in Relation with Previous Results

In Theorems 8 and 9, we have presented a result that characterizes the performance of classification with RBF interpolation functions. In particular, we have considered a setting where an RBF interpolator is fitted to each dimension of a low-dimensional embedding where different classes are separable. Our study has several links with RBF networks or least-squares regression algorithms. In this section, we interpret our findings in relation with previously established results.

Several previous works study the performance of learning by considering a probability measure ρ defined on $X \times Y$, where X and Y are two sets. The "label" set Y is often taken as an interval $[-L, L]$. Given a set of data pairs $\{(x_j, y_j)\}_{j=1}^N$ sampled from the distribution

ρ, the RBF network estimates a function $\hat{f}$ of the form

$$\hat{f}(x) = \sum_{i=1}^{R} c_i \, \phi \left(\frac{\|x - t_i\|}{\sigma_i} \right). \tag{22}$$

The number of RBF terms R may be different from the number of samples N in general. The function $\hat{f}$ minimizes the empirical error

$$\hat{f} = \arg\min_{f} \sum_{j=1}^{N} \left(f(x_j) - y_j \right)^2 .$$

The function $\hat{f}$ estimated from a finite collection of data samples is often compared to the regression function (Cucker and Smale, 2002)

$$f_o(x) = \int_Y y \, d\rho(y|x),$$

where $d\rho(y|x)$ is the conditional probability measure on Y. The regression function f_o minimizes the expected risk as

$$f_o = \arg\min_{f} \int_{X \times Y} \big(f(x) - y \big)^2 d\rho.$$

As the probability measure ρ is not known in practice, the estimate $\hat{f}$ of f_o is obtained from data samples. Several previous works have characterized the performance of learning by studying the approximation error (Niyogi and Girosi, 1996), (Lin et al., 2014)

$$\mathbb{E}[(f_o - \hat{f})^2] = \int_X (f_o(x) - \hat{f}(x))^2 d\rho_X(x), \tag{23}$$

where ρ_X is the marginal probability measure on X. This definition of the approximation error can be adapted to our setting as follows. In our problem the distribution of each class is assumed to have a bounded support, which is a special case of modeling the data with an overall probability distribution ρ. If the supports $\mathcal{M}_m$ are assumed to be nonintersecting, the regression function f_o is given by

$$f_o(x) = \sum_{m=1}^{M} m \, I_m(x),$$

which corresponds to the class labels $m = 1, \ldots, M$, where I_m is the indicator function of the support $\mathcal{M}_m$. It is then easy to show that the approximation error $\mathbb{E}[(f_o - \hat{f})^2]$ can be bounded as a constant times the probability of misclassification $P(\hat{C}(x) \neq m)$. Hence, we can compare our misclassification probability bounds in Section 2.3 with the approximation error in other works.

The study in (Niyogi and Girosi, 1996) assumes that the regression function is an element of the Bessel potential space of a sufficiently high order and that the sum of the coefficients

$|c_i|$ is bounded. It is then shown that for data sampled from $\mathbb{R}^n$, with probability greater than $1-\delta$ the approximation error in (23) can be bounded as

$$\mathbb{E}[(f_o - \hat{f})^2] \leq O\left(\frac{1}{R}\right) + O\left(\sqrt{\frac{Rn\log(RN) - \log(\delta)}{N}}\right), \tag{24}$$

where R is the number of RBF terms.

The analysis by Lin et al. (2014) considers families of RBF kernels that include the Gaussian function. Supposing that the regression function f_o is of Sobolev class W_2^r, and that the number of RBF terms is given by $R = N^{\frac{n}{n+2r}}$ in terms of the number of samples N, the approximation error is bounded as

$$\mathbb{E}[(f_o - \hat{f})^2] \leq O(N^{-\frac{2r}{n+2r}}\log^2(N)). \tag{25}$$

Next, we overview the study by Hernández-Aguirre et al. (2002), which studies the performance of RBFs in a Probably Approximately Correct (PAC)-learning framework. For $X \subset \mathbb{R}^n$, a family $\mathcal{F}$ of measurable functions from X to $[0,1]$ is considered and the problem of approximating a target function f_0 known only through examples with a function in $\hat{f} \in \mathcal{F}$ is studied. The authors use a previous result from (Vidyasagar, 1997) that relates the accuracy of empirical risk minimization to the covering number of $\mathcal{F}$ and the number of samples. Combining this result with the bounds on covering number estimates of Lipschitz continuous functions (Kolmogorov and Tihomirov, 1961), the following result is obtained for PAC function learning with RBF neural networks with Gaussian kernel. Let the coefficients be bounded as $|c_i| \leq A$, a common scale parameter be chosen as $\sigma_i = \sigma$, and $\mathbb{E}[|f_0 - \hat{f}|]$ be computed under a uniform probability measure ρ. Then if the number of samples satisfies

$$N \geq \frac{8}{\varepsilon^2}\log\left(\frac{\sqrt{2}RnA}{e^{-1/2}\sigma\zeta}\right), \tag{26}$$

an approximation of the target function is obtained with accuracy parameter ε and confidence parameter ζ:

$$P(\mathbb{E}[|f_0 - \hat{f}|] > \varepsilon) \leq \zeta. \tag{27}$$

In the above expression, the expectation is over the test samples, whereas the probability is over the training samples; i.e., over all possible distributions of training samples, the probability of having the average approximation error larger than ε is bounded. Note that, our results in Theorems 8 and 9, when translated into the above PAC-learning framework, correspond to a confidence parameter of $\zeta = 0$. This is because the misclassification probability bound of a test sample is valid for any choice of the training samples, provided that the condition (14) (or the condition (16)) holds. Thus, in our result the probability running over the training samples in (27) has no counterpart. When we take $\zeta = 0$, the above result does not provide a useful bound since $N \to \infty$ as $\zeta \to 0$. By contrast, our result is valid only if the conditions (14), (16) on the interpolation function holds. It is easy to show that, assuming nonintersecting class supports $\mathcal{M}_m$, the expression $\mathbb{E}[|f_0 - \hat{f}|]$ is given by a constant times the probability of misclassification. The accuracy parameter ε can then be seen as the counterpart of the misclassification probability upper bound given on the right hand sides of (15) and (17) (the expression subtracted from 1). At fixed N, the dependence

of the accuracy on the kernel scale parameter is monotonic in the bound (26); ε decreases as σ increases. Therefore, this bound does not guide the selection of the scale parameter of the RBF kernel, while the discussion in Section 2.4 (confirmed by the experimental results in Section 4.2) suggests the existence of an optimal scale.

Finally, we mention some results on the learning performance of regularized least squares regression algorithms. In (Caponnetto and De Vito, 2007) optimal rates are derived for the regularized least squares method in a Reproducing Kernel Hilbert Space (RKHS) in the minimax sense. It is shown that, under some hypotheses concerning the data probability measure and the complexity of the family of learnt functions, the maximum error (yielded by the worst distribution) obtained with the regularized least squares method converges at a rate of $O(1/N)$. Next, the work in (Steinwart et al., 2009) shows that, in regularized least squares regression over a RKHS, if the eigenvalues of the kernel integral operator decay sufficiently fast, and if the ℓ_∞-norms of regression functions can be bounded, the error of the classifier converges at a rate of up to $O(1/N)$ with high probability. Steinwart et al. also examine the learning performance in relation with the exponent of the function norm in the regularization term and show that the learning rate is not affected by the choice of the exponent of the function norm.

We now overview the three bounds given in (24), (25), and (26) in terms of the dependence of the error on the number of samples. The results in (24) and (25) provide a useful bound only in the case where the number of samples N is larger than the number of RBF terms R, contrary to our study where we treat the case $R = N$. If it is assumed that N is sufficiently larger than R, the result in (24) predicts a rate of decay of only $O(\sqrt{\log(N)/N})$ in the misclassification probability. The bound in (25) improves with the Sobolev regularity of the regression function; however, the dependence of the error on the number of samples is of a similar nature to the one in (24). Considering ε as a misclassification error parameter in the bound in (26), the error decreases at a rate of $O(N^{-1/2})$ as the number of samples increases. The analysis in (Caponnetto and De Vito, 2007) and (Steinwart et al., 2009) also provide the similar rates of convergence of $O(N^{-1})$. Meanwhile, our results in Theorems 8 and 9 predict an exponential decay in the misclassification probability as the number of samples N increases (under the reasonable assumption that $N_m = O(N)$ for each class m). The reason why we arrive at a more optimistic bound is the specialization of the analysis to the considered particular setting, where the support of each class is assumed to be restricted to a totally bounded region in the ambient space, as well as the assumed relations between the separation margin of the embedding and the regularity of the interpolation function.

Another difference between these previous results and ours is the dependence on the dimension. The results in (24), (25), and (26) predict an increase in the error at the respective rates of $O(\sqrt{n})$, $O(e^{-1/n})$, and $O(\sqrt{\log n})$ with the ambient space dimension n. While these results assume that the data $\mathcal{X} \subset \mathbb{R}^n$ is in an Euclidean space of dimension n, our study assumes the data $\mathcal{X}$ to be in a generic Hilbert space H. The results in Theorems 5-8 involve the dimension d of the low-dimensional space of embedding and does not explicitly depend on the dimension of the ambient Hilbert space H (which could be infinite-dimensional). However, especially in the context of manifold learning, it is interesting to analyze the dependence of our bound on the intrinsic dimension of the class supports $\mathcal{M}_m$.

In order to put the expressions (15), (17) in a more convenient form, let us reduce one parameter by setting $Q = N_m \eta_{m,\delta}/2$. Then the misclassification probability is of

$$O\left(\exp(-N_m \eta_{m,\delta}^2) + N \exp\left(-\frac{N_m\, \eta_{m,\delta}\, \epsilon^2}{L_\phi^2\, \delta^2}\right)\right).$$

We can relate the dependence of this expression on the intrinsic dimension as follows. Since the supports $\mathcal{M}_m$ are assumed to be totally bounded, one can define a parameter Θ that represents the "diameter" of $\mathcal{M}_m$, i.e., the largest distance between any two points on $\mathcal{M}_m$. Then the measure $\eta_{m,\delta}$ of the minimum ball of radius δ in $\mathcal{M}_m$ is of $O((\delta/\Theta)^D)$, where D is the intrinsic dimension of $\mathcal{M}_m$. Replacing this in the above expression gives the probability of misclassification as

$$O\left(\exp\left(-\frac{N_m\, \delta^{2D}}{\Theta^{2D}}\right) + N \exp\left(-\frac{N_m\, \delta^{D-2}\, \epsilon^2}{L_\phi^2\, \Theta^D}\right)\right).$$

This shows that in order to retain the correct classification guarantee, as the intrinsic dimension D grows, the number of samples N_m should increase at a geometric rate with D. In supervised manifold learning problems, data sets usually have a low intrinsic dimension, therefore, this geometric rate of increase can often be tolerated. Meanwhile the dimension of the ambient space is typically high, so that performance bounds independent of the ambient space dimension are of particular interest. Note that generalization bounds in terms of the intrinsic dimension have been proposed in some previous works as well (Bickel and Li, 2007), (Kpotufe, 2011), for the local linear regression and the K-NN regression problems.

3. Separability of Supervised Nonlinear Embeddings

In the results in Section 2, we have presented generalization bounds for classifiers based on linearly separable embeddings. One may wonder if the separability assumption is easy to satisfy when computing structure-preserving nonlinear embeddings of data. In this section, we try to answer this question by focusing on a particular family of supervised dimensionality reduction algorithms, i.e., supervised Laplacian eigenmaps embeddings, and analyze the conditions of separability. We first discuss the supervised Laplacian eigenmaps embeddings in Section 3.1 and then present results in Section 3.2 about the linearly separability of these embeddings.

3.1 Supervised Laplacian Eigenmaps Embeddings

Let $\mathcal{X} = \{x_i\}_{i=1}^N \subset H$ be a set of training samples, where each x_i belongs to one of M classes. Most manifold learning algorithms rely on a graph representation of data. This graph can be a complete graph in some works, in which case an edge exists between each pair of samples. Meanwhile, in some manifold learning algorithms, in order to better capture the intrinsic geometric structure of data, each data sample is connected only to its nearest neighbors in the graph. In this case, an edge exists only between neighboring data samples.

In our analysis, we consider a weighted data graph G each vertex of which represents a point x_i. We write $x_i \sim x_j$, or simply $i \sim j$ if the graph contains an edge between the

data samples x_i, x_j. We denote the edge weight as $w_{ij} > 0$. The weights w_{ij} are usually determined as a positive and monotonically decreasing function of the distance between x_i and x_j in H, where the Gaussian function is a common choice. Nevertheless, we maintain a generic formulation here without making any assumption on the neighborhood or weight selection strategies.

Now let G_w and G_b represent two subgraphs of G, which contain the edges of G that are respectively within the same class and between different classes. Hence, G_w contains an edge $i \sim_w j$ between samples x_i and x_j, if $i \sim j$ and $C_i = C_j$. Similarly, G_b contains an edge $i \sim_b j$ if $i \sim j$ and $C_i \neq C_j$. We assume that all vertices of G are contained in both G_w and G_b; and that G_w has exactly M connected components such that the training samples in each class form a connected component[2]. We also assume that G_w and G_b do not contain any isolated vertices; i.e., each data sample x_i has at least one neighbor in both graphs.

The $N \times N$ weight matrices W_w and W_b of G_w and G_b have entries as follows.

$$W_w(i,j) = \begin{cases} w_{ij} & \text{if } i \sim j \text{ and } C_i = C_j \\ 0 & \text{otherwise} \end{cases}$$

$$W_b(i,j) = \begin{cases} w_{ij} & \text{if } i \sim j \text{ and } C_i \neq C_j \\ 0 & \text{otherwise} \end{cases}$$

Let $d_w(i)$ and $d_b(i)$ denote the degrees of x_i in G_w and G_b

$$d_w(i) = \sum_{j \sim_w i} w_{ij}, \qquad d_b(i) = \sum_{j \sim_b i} w_{ij},$$

and D_w, D_b denote the $N \times N$ diagonal degree matrices given by $D_w(i,i) = d_w(i)$, $D_b(i,i) = d_b(i)$. The normalized graph Laplacian matrices L_w and L_b of G_w and G_b are then defined as

$$L_w := D_w^{-1/2}(D_w - W_w)D_w^{-1/2}, \qquad L_b := D_b^{-1/2}(D_b - W_b)D_b^{-1/2}.$$

Supervised extensions of the Laplacian eigenmaps and LPP algorithms seek a d-dimensional embedding of the data set $\mathcal{X}$, such that each x_i is represented by a vector $y_i \in \mathbb{R}^{d \times 1}$. Denoting the new data matrix as $Y = [y_1\, y_2\, \ldots\, y_N]^T \in \mathbb{R}^{N \times d}$, the coordinates of data samples are computed by solving the problem

$$\text{“Minimize } \mathrm{tr}(Y^T L_w Y) \text{ while maximizing } \mathrm{tr}(Y^T L_b Y).\text{”} \tag{28}$$

The reason behind this formulation can be explained as follows. For a graph Laplacian matrix $L = D^{-1/2}(D - W)D^{-1/2}$, where D and W are respectively the degree and the weight matrices, defining the coordinates $Z = D^{-1/2}Y$ normalized with the vertex degrees, we have

$$\mathrm{tr}(Y^T L\, Y) = \mathrm{tr}(Z^T(D - W)Z) = \sum_{i \sim j} \|z_i - z_j\|^2 w_{ij}, \tag{29}$$

2. The straightforward application of common graph construction strategies, like connecting each training sample to its K-nearest neighbors or to its neighbors within a given distance, may result in several disconnected components in a single class in the graph if there is much diversity in that class. However, this difficulty can be easily overcome by introducing extra edges to bridge between graph components that are originally disconnected.

where z_i is the i-th row of Z giving the normalized coordinates of the embedding of the data sample x_i. Hence, the problem in (28) seeks a representation Y that maps nearby samples in the same class to nearby points, while mapping nearby samples from different classes to distant points. In fact, when the samples x_i are assumed to come from a manifold $\mathcal{M}$, the term $y^T L y$ is the discrete equivalent of

$$\int_{\mathcal{M}} \|\nabla f(x)\|^2 dx,$$

where $f : \mathcal{M} \to \mathbb{R}$ is a continuous function on the manifold that extends the one-dimensional coordinates y to the whole manifold. Hence, the term $\mathrm{tr}(Y^T L Y)$ captures the rate of change of the learnt coordinate vectors Y over the underlying manifold. Then, in a setting where the samples of different classes come from M different manifolds $\{\mathcal{M}_m\}_{m=1}^M$, the formulation in (28) looks for a function that has a slow variation on each manifold $\mathcal{M}_m$, while having a fast variation "between" different manifolds.

The supervised learning problem in (28) has so far been studied by several authors with slight variations in their problem formulations. Raducanu and Dornaika (2012) minimize a weighted difference of the within-class and between-class similarity terms in (28) in order to learn a nonlinear embedding. Meanwhile, linear dimensionality reduction methods pose the manifold learning problem as the learning of a linear projection matrix $P \in \mathbb{R}^{d \times n}$; therefore, solve the problem in (28) under the constraint $y_i = P\, x_i$, where $x_i \in \mathbb{R}^{n \times 1}$ and $d < n$. Hua et al. (2012) formulate the problem as the minimization of the difference of the within-class and the between-class similarity terms in (28) as well. Thus, their algorithm can be seen as the linear version of the method by Raducanu and Dornaika (2012). Sugiyama (2007) proposes an adaptation of the Fisher discriminant analysis algorithm to preserve the local structures of data. Data sample pairs are weighted with respect to their affinities in the construction of the within-class and the between-class scatter matrices in Fisher discriminant analysis. Then the trace of the ratio of the between-class and the within-class scatter matrices is maximized to learn a linear embedding. Meanwhile, the within-class and the between-class local scatter matrices are closely related to the two terms in (28) as shown by Yang et al. (2011). The terms $Y^T L_w Y$ and $Y^T L_b Y$, when evaluated under the constraint $y_i = P\, x_i$, become equal to the locally weighted within-class and between-class scatter matrices of the projected data. Cui and Fan (2012) and Wang and Chen (2009) propose to maximize the ratio of the between-class and the within-class local scatters in the learning. Yang et al. (2011) optimize the same objective function, while they construct the between-class graph only on the centers of mass of the classes. Zhang et al. (2012) similarly optimize a Fisher metric to maximize the ratio of the between- and within-class scatters; however, the total scatter is also taken into account in the objective function in order to preserve the overall manifold structure.

All of the above methods use similar formulations of the supervised manifold learning problem and give comparable results. In our study, we base our analysis on the following formal problem definition

$$\min_{Y} \mathrm{tr}(Y^T L_w Y) - \mu\, \mathrm{tr}(Y^T L_b Y) \text{ subject to } Y^T Y = I, \tag{30}$$

which minimizes the difference of the within-class and the between-class similarity terms as in works such as (Raducanu and Dornaika, 2012) and (Hua et al., 2012). Here I is the

$d \times d$ identity matrix and $\mu > 0$ is a parameter adjusting the weights of the two terms. The condition $Y^T Y = I$ is a commonly used constraint to remove the scale ambiguity of the coordinates. The solution of the problem (30) is given by the first d eigenvectors of the matrix

$$L_w - \mu L_b$$

corresponding to its smallest eigenvalues.

Our purpose in this section is then to theoretically study the linear separability of the learnt coordinates of training data, with respect to the definition of linear separability given in (1). In the following, we determine some conditions on the graph properties and the weight parameter μ that ensure the linear separability. We derive lower bounds on the margin γ and study its dependence on the model parameters. Let us give beforehand the following definitions about the graphs G_w and G_b.

Definition 10 *The volume of the subgraph of G_w that corresponds to the connected component containing samples from class k is*

$$V_k := \sum_{i:\, C_i = k} d_w(i).$$

We define the maximal within-class volume as

$$V_{max} := \max_{k=1,\ldots,M} V_k.$$

The volume of the component of G_b containing the edges between the samples of classes k and l is [3]

$$V_{kl}^b := \sum_{\substack{i \sim_b j \\ C_i = k, C_j = l}} 2\, w_{ij}.$$

We then define the maximal pairwise between-class volume as

$$V_{max}^b := \max_{k \neq l} V_{kl}^b.$$

In a connected graph, the distance between two vertices x_i and x_j is the number of edges in a shortest path joining x_i and x_j. The diameter of the graph is then given by the maximum distance between any two vertices in the graph (Chung, 1996). We define the diameter of the connected component of G_w corresponding to class k as follows.

Definition 11 *For any two vertices x_i and x_j such that $C_i = C_j = k$, consider a within-class shortest path joining x_i and x_j, which contains samples only from class k. Then the diameter D_k of the connected component of G_w corresponding to class k is the maximum number of edges in the within-class shortest path joining any two vertices x_i and x_j from class k.*

Definition 12 *The minimum edge weight within class k is defined as*

$$w_{min,k} := \min_{\substack{i \sim_w j \\ C_i = C_j = k}} w_{ij}.$$

3. In order to keep the analogy with the definition of V_k, a 2 factor is introduced in this expression as each edge is counted only once in the sum.

3.2 Separability Bounds for Two Classes

We now present a lower bound for the linear separability of the embedding obtained by solving (30) in a setting with two classes $C_i \in \{1, 2\}$. We first show that an embedding of dimension $d = 1$ is sufficient to achieve linear separability for the case of two classes. We then derive a lower bound on the separation in terms of the graph parameters and the algorithm parameter μ.

Consider a one-dimensional embedding $Y = y = [y_1\, y_2\, \ldots\, y_N]^T \in \mathbb{R}^{N\times 1}$, where $y_i \in \mathbb{R}$ is the coordinate of the data sample x_i in the one-dimensional space. The coordinate vector y is given by the eigenvector of $L_w - \mu L_b$ corresponding to its smallest eigenvalue. We begin with presenting the following result, which states that the samples from the two classes are always mapped to different halves (nonnegative or nonpositive) of the real line.

Lemma 13 *The learnt embedding y of dimension $d = 1$ satisfies*

$$\begin{aligned} y_i &\le 0 \qquad \text{if } C_i = 1 \text{ (or respectively } C_i\text{=2)} \\ y_i &\ge 0 \qquad \text{if } C_i = 2 \text{ (or respectively } C_i\text{=1)} \end{aligned}$$

for any $\mu > 0$ and for any choice of the graph parameters.

Lemma 13 is proved in Appendix B.1. The lemma states that in one-dimensional embeddings of two classes, samples from different classes always have coordinates with different signs. Therefore, the hyperplane given by $\omega = 1$, $b = 0$ separates the data as $\omega^T y_i \le 0$ for $C_i = 1$ and $\omega^T y_i \ge 0$ for $C_i = 2$ (since the embedding is one dimensional, the vector ω is a scalar in this case). However, this does not guarantee that the data is separable with a positive margin $\gamma > 0$. In the following result, we show that a positive margin exists and give a lower bound on it. In the rest of this section, we assume without loss of generality that classes 1 and 2 are respectively mapped to the negative and positive halves of the real axis.

Theorem 14 *Defining the normalized data coordinates $z = D_w^{-1/2} y$, let*

$$z_{1,max} := \max_{i:\, C_i = 1} z_i \qquad z_{2,min} := \min_{i:\, C_i = 2} z_i$$

denote the maximum and minimum coordinates that classes 1 and 2 are respectively mapped to with a one-dimensional embedding learnt with supervised Laplacian eigenmaps. We also define the parameters

$$\overline{w}_{min} = \min_{k\in\{1,2\}} \frac{w_{min,k}}{D_k}\,, \qquad \beta_i = \frac{d_w(i)}{d_b(i)}\,, \qquad \beta_{max} = \max_i \beta_i\,,$$

where D_k is the diameter of the graph corresponding to class k as defined in Definition 11. Then, if the weight parameter is chosen such that $0 < \mu < \overline{w}_{min}/(\beta_{max} V^b_{max})$, any supervised Laplacian embedding of dimension $d \ge 1$ is linearly separable with a positive margin lower bounded as below:

$$z_{2,min} - z_{1,max} \ge \frac{1}{\sqrt{V_{max}}} \left(1 - \sqrt{\frac{\mu \beta_{max} V^b_{max}}{\overline{w}_{min}}}\right). \tag{31}$$

The proof of Theorem 14 is given in Appendix B.2. The proof is based on a variational characterization of the eigenvector of $L_w - \mu L_b$ corresponding to its smallest eigenvalue, whose elements are then bounded in terms of the parameters of the graph such as the diameters and volumes of its connected components.

Theorem 14 states that an embedding learnt with the supervised Laplacian eigenmaps method makes two classes linearly separable if the weight parameter μ is chosen sufficiently small. In particular, the theorem shows that, for any $0 < \delta < {V_{max}}^{-1/2}$, a choice of the weight parameter μ satisfying

$$0 < \mu \leq \frac{\overline{w}_{min}}{\beta_{max}\, V_{max}^b} \left(1 - \sqrt{V_{max}}\, \delta\right)^2$$

guarantees a separation of $z_{2,min} - z_{1,max} \geq \delta$ between classes 1 and 2 at $d = 1$. Here, we use the symbol δ to denote the separation in the normalized coordinates z. In practice, either one of the normalized eigenvectors z or the original eigenvectors y can be used for embedding the data. If the original eigenvectors y are used, due to the relation $y = D_w^{1/2} z$, we can lower bound the separation as $y_{2,min} - y_{1,max} \geq \sqrt{d_{w,min}}(z_{2,min} - z_{1,max})$ where $d_{w,min} = \min_i d_w(i)$. Thus, for any embedding of dimension $d \geq 1$, there exists a hyperplane that results in a linear separation with a margin γ of at least

$$\gamma \geq \sqrt{\frac{d_{w,min}}{V_{max}}} \left(1 - \sqrt{\frac{\mu \beta_{max} V_{max}^b}{\overline{w}_{min}}}\right).$$

Next, we comment on the dependence of the separation on μ. The inequality in (31) shows that the lower bound on the separation $z_{2,min} - z_{1,max}$ has a variation of $O(1 - \sqrt{\mu})$ with the weight parameter μ. The fact that the separation decreases with the increase in μ seems counterintuitive at first; this parameter weights the between-class dissimilarity in the objective function. This can be explained as follows. When μ is high, the algorithm tries to increase the distance between neighboring samples from different classes as much as possible by moving them away from the origin (remember that different classes are mapped to the positive and the negative sides of the real line). However, since the normalized coordinate vector z has to respect the equality $z^T D_w z = 1$, the total squared norm of the coordinates cannot be arbitrarily large. Due to this constraint, setting μ to a high value causes the algorithm to map non-neighboring samples from different classes to nearby coordinates close to the origin. This occurs since the increase in μ reduces the impact of the first term $y^T L_w y$ in the overall objective and results in an embedding with a weaker link between the samples of the same class. This causes a polarization of the data and eventually reduces the separation. Hence, the μ parameter should be carefully chosen and should not take too large values.

Theorem 14 characterizes the separation at $d = 1$ in terms of the distance between the supports of the two classes. Meanwhile, it is also interesting to determine the individual distances of the supports of the two classes to the origin. In the following corollary, we present a lower bound on the distance between the coordinates of any sample and the origin.

Corollary 15 *The distance between the supports of the first and the second classes and the origin in a one-dimensional embedding is lower bounded in terms of the separation between the two classes as*

$$\min\{|z_{1,max}|,\ |z_{2,min}|\} \geq \frac{1}{2}\frac{\beta_{min}}{\beta_{max}}(z_{2,min} - z_{1,max})$$

where

$$\beta_{min} = \min_i \beta_i, \qquad \beta_{max} = \max_i \beta_i.$$

Corollary 15 is proved in Appendix B.3. The proof is based on a Lagrangian formulation of the embedding as a constrained optimization problem, which then allows us to establish a link between the separation and the individual distances of class supports to the origin. The corollary states a lower bound on the portion of the overall separation lying in the negative or the positive sides of the real line. In particular, if the vertex degrees are equal for all samples in G_w and G_b (which is the case, for instance, if all vertices have the same number of neighbors and a constant weight of $w_{ij} = 1$ is assigned to the edges), since $\beta_{min} = \beta_{max}$, the portions of the overall separation in the positive and negative sides of the real line will be equal.

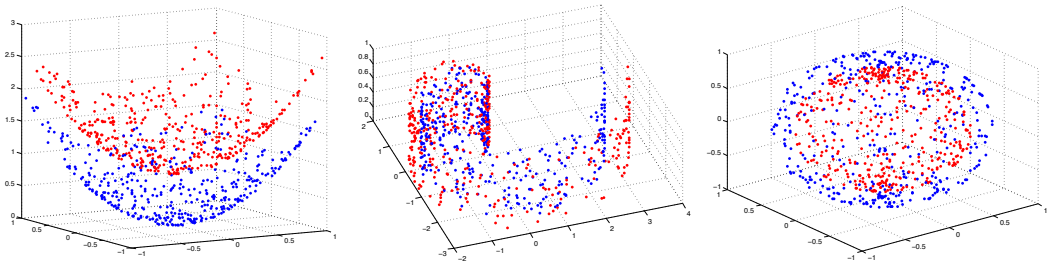
We have examined the linear separability of supervised Laplacian embeddings for the case of two classes in this section. An extension of these results to the case of multiple classes under some assumptions is available in the accompanying technical report (Vural and Guillemot, 2016b).

4. Experimental Results

In this section, we present results on synthetical and real data sets. We compare several supervised manifold learning methods and study their performances in relation with our theoretical results.

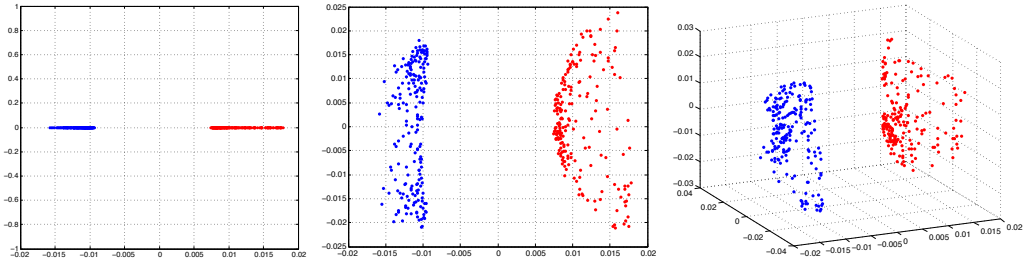
4.1 Separability of Embeddings with Supervised Manifold Learning

We first present results on synthetical data in order to study the embeddings obtained with supervised dimensionality reduction. We test the supervised Laplacian eigenmaps algorithm in a setting with two classes. We generate samples from two nonintersecting and linearly nonseparable surfaces in $\mathbb{R}^3$ that represent two different classes. We experiment on three different types of surfaces; namely, quadratic surfaces, Swiss rolls and spheres. The data sampled from these surfaces are shown in Figure 2. We choose $N = 200$ samples from each class. We construct the graph G_w by connecting each sample to its K-nearest neighbors from the same class, where K is chosen between 20 and 30. The graph G_b is constructed similarly, where each sample is connected to its $K/5$ nearest neighbors from the other class. The graph weights are determined as a Gaussian function of the distance between the samples. The embeddings are then computed by minimizing the objective function in (30). The one-dimensional, two-dimensional, and three-dimensional embeddings obtained for the quadratic surface are shown in Figure 3, where the weight parameter is taken as $\mu = 0.57$



(a) Quadratic surfaces (b) Swissrolls (c) Spheres

Figure 2: Data sampled from two-dimensional synthetical surfaces. Red and blue colors represent two different classes.

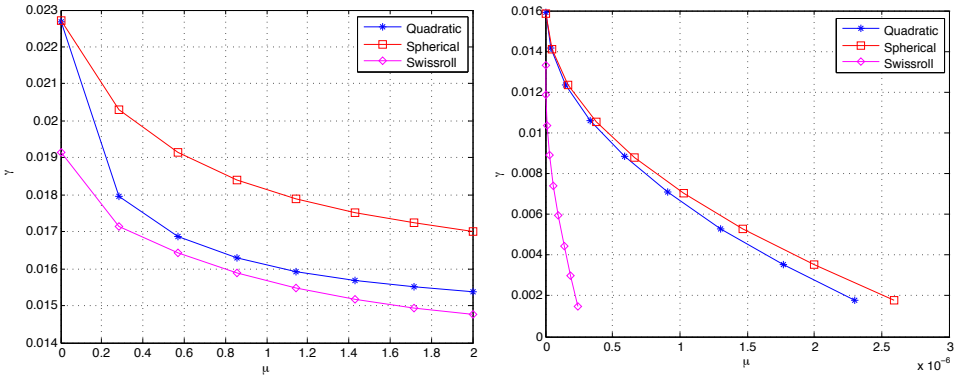


(a) 1-D embedding (b) 2-D embedding (c) 3-D embedding

Figure 3: Supervised Laplacian embeddings of data sampled from quadratic surfaces.

(to have a visually clear embedding for the purpose of illustration). Similar results are obtained on the Swiss roll and the spherical surface. One can observe that the data samples that were initially linearly nonseparable become linearly separable when embedded with the supervised Laplacian eigenmaps algorithm. The two classes are mapped to different (positive or negative) sides of the real line in Figure 3(a) as predicted by Lemma 13. The separation in the 2-D and 3-D embeddings in Figure 3 is close to the separation obtained with the 1-D embedding.

We then compute and plot the separation obtained at different values of μ. Figure 4(a) shows the experimental value of the separation $\gamma = z_{2,min} - z_{1,max}$ obtained with the 1-D embedding for the three types of surfaces. Figure 4(b) shows the theoretical upper bound for μ in Theorem 14 that guarantees a separation of at least γ. Both the experimental value and the theoretical bound for the separation γ decrease with the increase in the parameter μ. This is in agreement with (31), which predicts a decrease of $O(1-\sqrt{\mu})$ in the separation with respect to μ. The theoretical bound for the separation is seen to decrease at a relatively faster rate with μ for the Swiss roll data set. This is due to the particular structure of this data set with a nonuniform sampling density where the sampling is sparser away from the spiral center. The parameter $\overline{w}_{min}$ then takes a small value, which consequently leads to a



(a) Experimental value of the separation γ

(b) Theoretical upper bound for μ that guarantees a separation of at least γ

Figure 4: Variation of the separation γ between the two classes with the parameter μ for the synthetic data sets

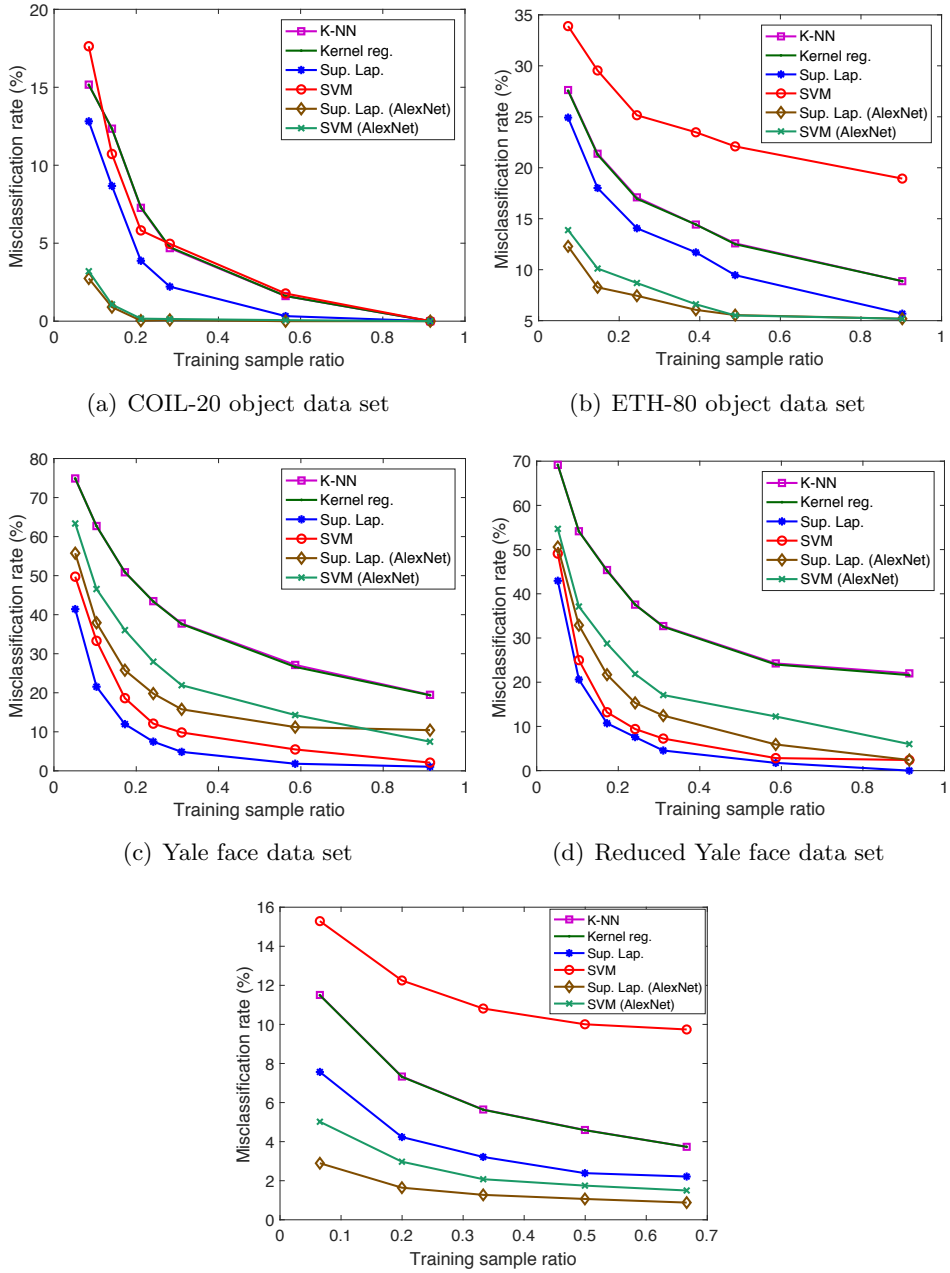
fast rate of decrease for the separation due to (31). Comparing Figures 4(a) and 4(b), one observes that the theoretical bounds for the separation are numerically more pessimistic than their experimental values, which is a result of the fact that our results are obtained with a worst-case analysis. Nevertheless, the theoretical bounds capture well the actual variation of the separation margin with μ.

4.2 Classification Performance of Supervised Manifold Learning Algorithms

We now study the overall performance of classification obtained in a setting with supervised manifold learning, where the out-of-sample generalization is achieved with smooth RBF interpolators. We evaluate the theoretical results of Section 2 on several real data sets: the COIL-20 object database (Nene et al., 1996), the Yale face database (Georghiades et al., 2001), the ETH-80 object database (Leibe and Schiele, 2003), and the MNIST handwritten digit database (LeCun et al., 1998). The COIL-20, Yale face, ETH-80, and MNIST databases contain a total of 1420, 2204, 3280, and 70046 images from 20, 38, 8, and 10 image classes respectively. The images in the COIL-20, Yale and ETH-80 data sets are converted to greyscale, normalized, and downsampled to a resolution of respectively 32×32, 20×17, and 20×20 pixels.

4.2.1 Comparison of Supervised Manifold Learning to Baseline Classifiers

We first compare the performance of supervised manifold learning with some reference classification methods. The performances of SVM, K-NN, kernel regression, and the supervised Laplacian eigenmaps methods are evaluated and compared. Figure 5 reports the results obtained on the COIL-20 data set, the ETH-80 data set, the Yale data set, a subset of the Yale data set consisting of its first 10 classes (reduced Yale data set), and the MNIST data set. The SVM, K-NN, and kernel regression algorithms are applied in the original



(a) COIL-20 object data set

(b) ETH-80 object data set

(c) Yale face data set

(d) Reduced Yale face data set

(e) MNIST data set

Figure 5: Comparison of the performance of several supervised classification methods

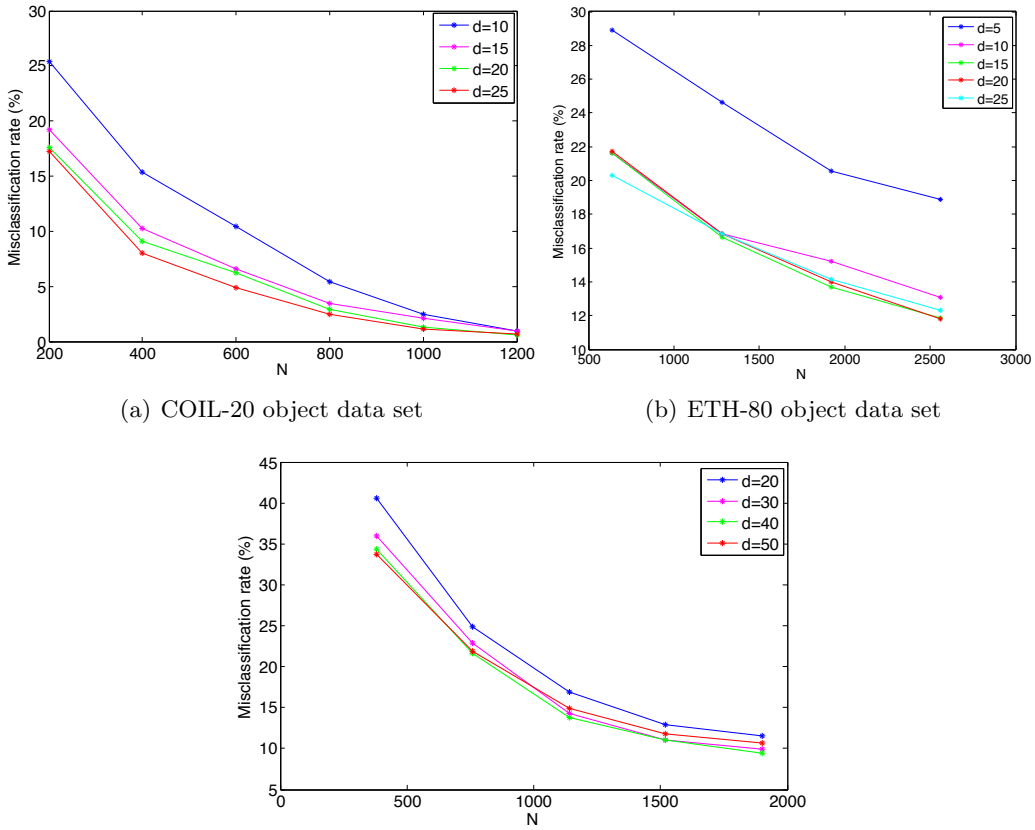
domain and their hyperparameters are optimized with cross-validation. In the supervised Laplacian eigenmaps method, the embedding of the training images into a low-dimensional space is computed. Then, an out-of-sample interpolator with Gaussian RBFs is constructed that maps the training samples to their embedded coordinates as described in Section 2.3.

Test samples are mapped to the low-dimensional domain via the RBF interpolator and the class labels of test samples are estimated via nearest-neighbor classification in the low-dimensional domain. The supervised Laplacian eigenmaps and the SVM methods are also tested over an alternative representation of the image data sets based on deep learning. The images are provided as input to the pretrained AlexNet convolutional neural network proposed in (Krizhevsky et al., 2012), and the activation values at the second fully connected layer are used as the feature representations of the images. The feature representations of training and test images are then provided to the supervised Laplacian eigenmaps and the SVM methods. The plots in Figure 5 show the variation of the misclassification rate of test samples in percentage with the ratio of the number of training samples in the whole data set. The results are the average of 5 repetitions of the experiment with different random choices for the training and test samples.

The results in Figure 5 show that the best results are obtained with the supervised Laplacian eigenmaps algorithm in general. The performances of the algorithms improve with the number of training images as expected. In the COIL-20 and ETH-80 object data sets, the supervised Laplacian eigenmaps and the SVM algorithms yield significantly smaller error when applied to the feature representations of the images obtained with deep learning. Meanwhile, in the Yale face data set these two methods perform better on raw image intensity maps. This can be explained with the fact that the AlexNet model may be more successful in extracting useful features for object images rather than face images as it is trained on many common object and animal classes. It is interesting to compare Figures 5(c) and 5(d). While the performances of the supervised Laplacian eigenmaps and the SVM methods are closer in the reduced version of the Yale database with 10 classes, the performance gap between the supervised Laplacian eigenmaps method and the other methods is larger for the full data set with 38 classes. This can be explained with the fact that the linear separability of different classes degrades as the number of classes increases, thus causing a degradation in the performance of the classifiers in comparison. Meanwhile, the performance of the supervised Laplacian eigenmaps method is not much affected by the increase in the number of classes. The K-NN and kernel regression classifiers are seen to give almost the same performance in the plots in Figure 5. The number of neighbors is set as $K = 1$ for the K-NN algorithm in these experiments, where it has been observed to attain its best performance; and the scale parameter of the kernel regression algorithm is optimized to get the best accuracy, which has turned out to take relatively small values. Hence the performances of these two classifiers practically correspond to that of the nearest-neighbor classifier in the original domain.

4.2.2 Variation of the Error with Algorithm Parameters and Sample Size

We first study the evolution of the classification error with the number of training samples. Figures 6(a)- 6(c) show the variation of the misclassification rate of test samples with respect to the total number of training samples N for the COIL-20, ETH-80 and Yale data sets. Each curve in the figure shows the errors obtained at a different value of the dimension d of the embedding. The decrease in the misclassification rate with the number of training samples is in agreement with the results in Section 2 as expected.



(a) COIL-20 object data set

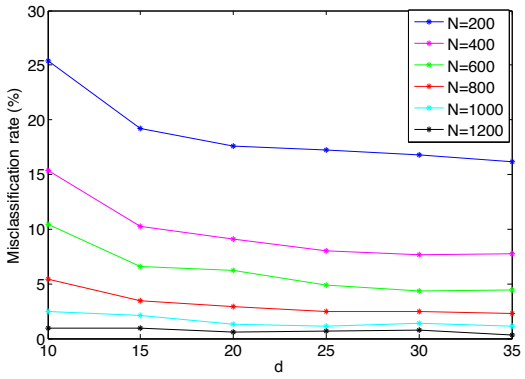
(b) ETH-80 object data set

(c) Yale face data set

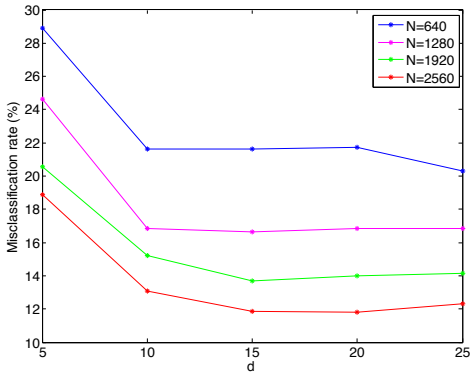
Figure 6: Variation of the misclassification rate with the number of training samples

The results of Figure 6 are replotted in Figure 7, where the variation of the misclassification rate is shown with respect to the dimension d of the embedding at different N values. It is observed that there may exist an optimal value of the dimension that minimizes the misclassification rate. This can be interpreted in light of the conditions (14) and (16) in Theorems 8 and 9, which impose a lower bound on the separability margin γ_Q in terms of the dimension d of the embedding. In the supervised Laplacian eigenmaps algorithm, the first few dimensions are critical and effective for separating different classes. The decrease in the error with the increase in the dimension for small values of d can be explained with the fact that the separation increases with d at small d, thereby satisfying the conditions (14), (16). Meanwhile, the error may stagnate or increase if the dimension d increases beyond a certain value, as the separation does not necessarily increase at the same rate.

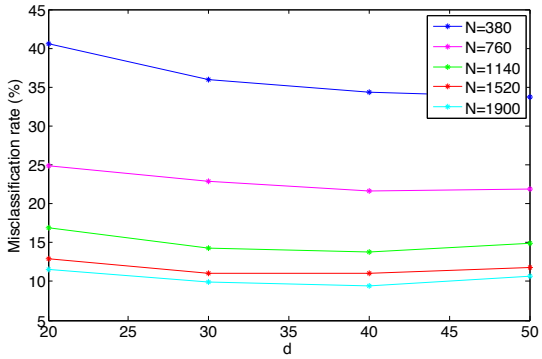
We then examine the variation of the misclassification rate with the separation. We obtain embeddings at different separation values γ by changing the parameter μ of the supervised Laplacian eigenmaps algorithm. Figure 8 shows the variation of the misclassification rate with the separation γ. Each curve is obtained at a different value of the scale parameter σ of the RBF kernels. It is seen that the misclassification rate decreases



(a) COIL-20 object data set

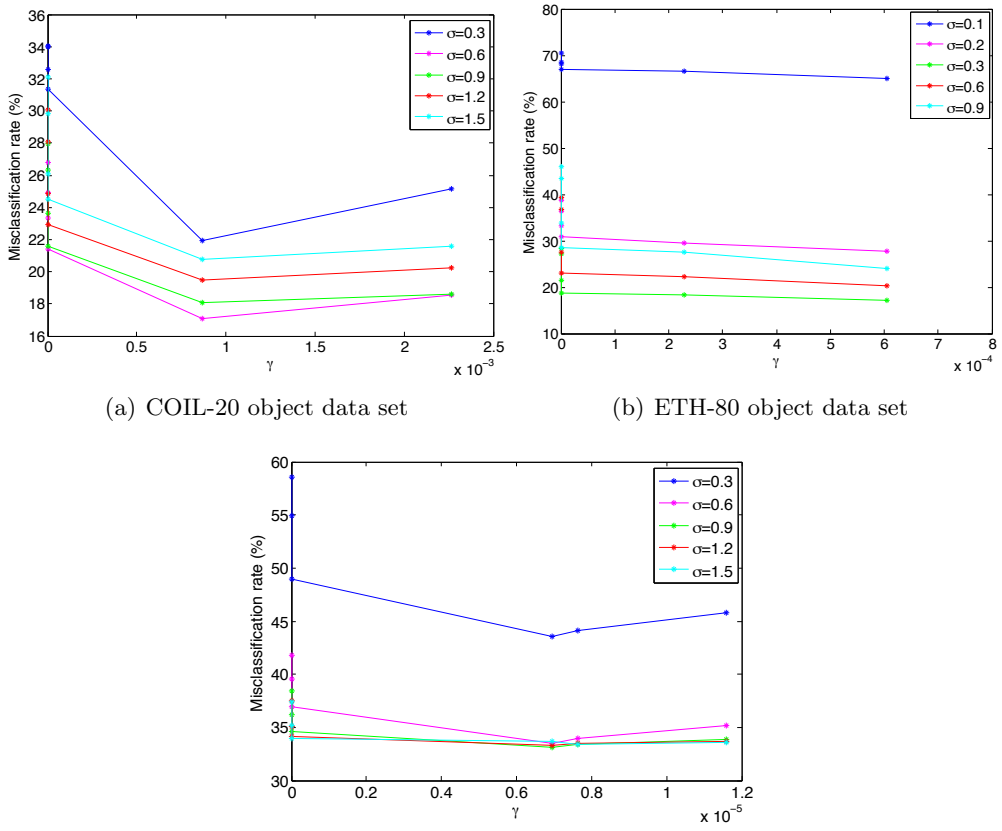


(b) ETH-80 object data set



(c) Yale face data set

Figure 7: Variation of the misclassification rate with the dimension of the embedding



(a) COIL-20 object data set

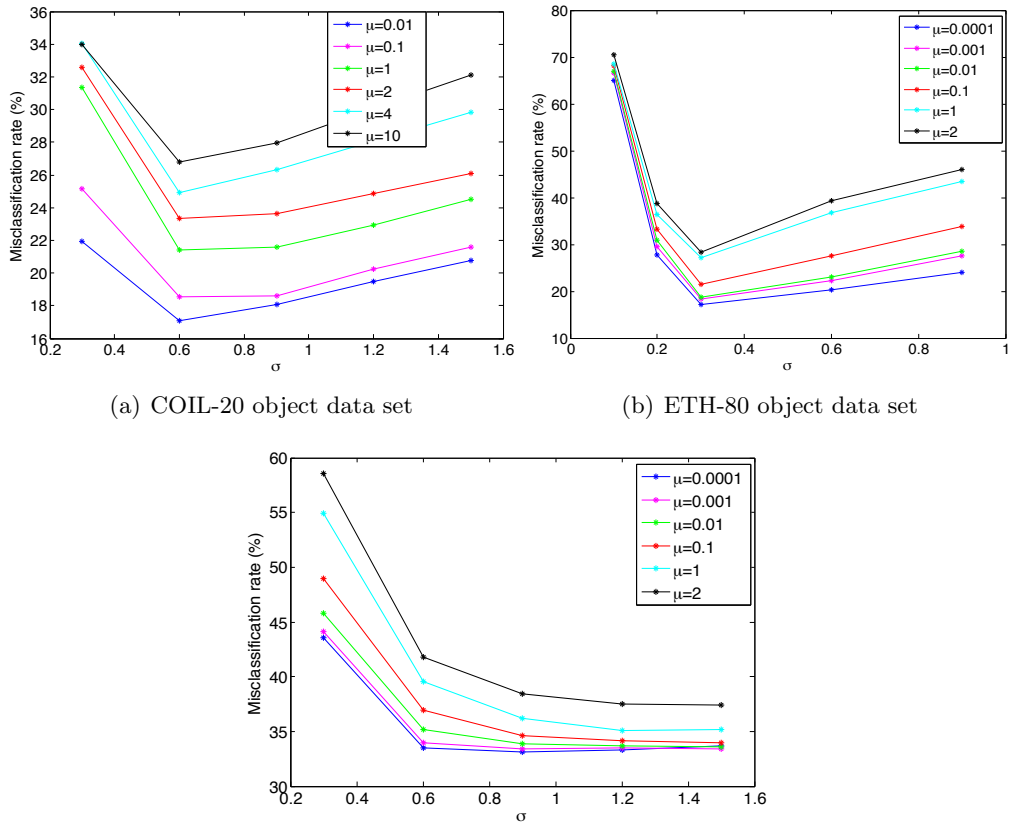
(b) ETH-80 object data set

(c) Yale face data set

Figure 8: Variation of the misclassification rate with the separation

in general with the separation for small γ values. This is in agreement with our results, as the conditions (14), (16) require the separation to be higher than a threshold. On the other hand, the possible increase in the error at relatively large values of the separation is due to the following. These parts of the plots are obtained at very small μ values, which typically result in a deformed embedding with a degenerate geometry. The deformation of structure at too small values of μ may cause the interpolation function to be irregular and hence result in an increase in the error. The tradeoff between the separation and the interpolation function regularity is further studied in Section 4.2.3.

Finally, Figure 9 shows the relation between the misclassification error and the scale parameter σ of the Gaussian RBF kernels. Each curve is obtained at a different value of the μ parameter. The optimum value of the scale parameter minimizing the misclassification error can be observed in most experiments. These results confirm the findings of Section 2.4, suggesting that there exists a unique value of σ that minimizes the left hand side of the conditions (14), (16), which probabilistically guarantee the correct classification of data.



(a) COIL-20 object data set

(b) ETH-80 object data set

(c) Yale face data set

Figure 9: Variation of the misclassification rate with the scale parameter

4.2.3 Performance Analysis of Several Supervised Manifold Learning Algorithms

Next, we compare several supervised manifold learning methods. We aim to interpret the performance differences of different types of embeddings in light of our theoretical results in Section 2.3. First, remember from Theorem 8 that the condition

$$\sqrt{d}\,\mathcal{C}\,(L_\phi \delta + \epsilon) \leq \gamma/2 \tag{32}$$

needs to be satisfied (or, equivalently the condition (16) from Theorem 9) in order for the generalization bounds to hold. This preliminary condition basically states that a compromise must be achieved between the regularity of the interpolation function, captured via the terms $\mathcal{C}$ and L_ϕ, and the separation γ of the embedding of training samples, in order to bound the misclassification error. In other words, increasing the separation too much in the embedding of training samples does not necessarily lead to good classification performance if the interpolation function has poor regularity.

Hence, when comparing different embeddings in the experiments of this section, we define a condition parameter given by

$$\frac{\sqrt{d}\mathcal{C}L_\phi}{\gamma},$$

which represents the ratio of the left and right hand sides of (32) (by fixing the probability parameters δ and ϵ). Setting the Lipschitz constant of the Gaussian RBF kernel as $L_\phi = \sqrt{2}e^{-\frac{1}{2}}\sigma^{-1}$ (see Section 2.4 for details), we can equivalently define the condition parameter as

$$\kappa = \frac{\sqrt{d}\mathcal{C}}{\sigma\gamma} \tag{33}$$

and study this condition parameter for the supervised dimensionality methods in comparison. Note that a smaller condition parameter means that the necessary conditions of Theorems 8 and 9 are more likely to be satisfied, hence hinting at the expectation of a better classification accuracy.

We compare the following supervised embeddings:

- Supervised Laplacian eigenmaps embedding obtained by solving (30):

$$\min_Y \operatorname{tr}(Y^T L_w Y) - \mu \operatorname{tr}(Y^T L_b Y) \text{ subject to } Y^T Y = I.$$

- Fisher embedding[4], obtained by solving

$$\max_Y \frac{\operatorname{tr}(Y^T L_b Y)}{\operatorname{tr}(Y^T L_w Y)}. \tag{34}$$

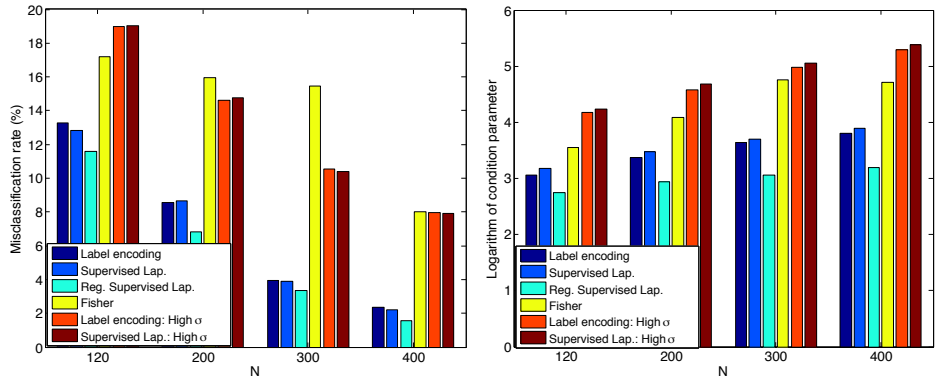
- Label encoding, which maps each data sample to its label vector of the form

$$[0\ 0 \ldots 1 \ldots 0],$$

  where the only nonzero entry corresponds to its class.

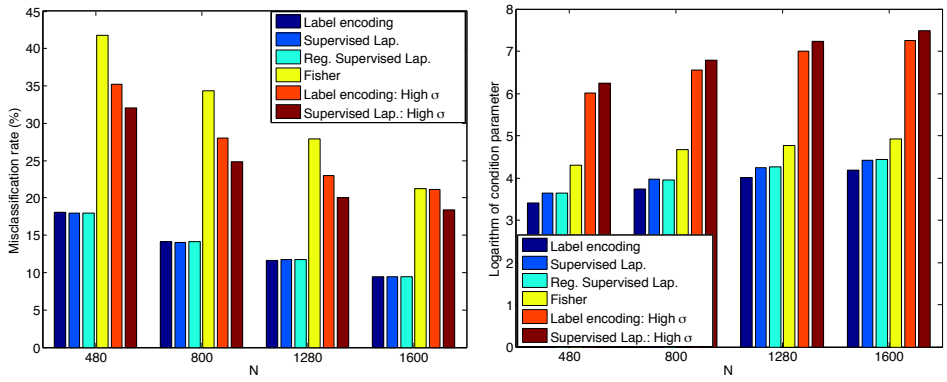
The label encoding method is included in the experiments to provide a reference, which can also be regarded as a degenerate supervised manifold learning algorithm that provides maximal separation between data samples from different classes. In all of the above methods the training samples are embedded into the low-dimensional domain, and test samples are mapped via Gaussian RBF interpolators and assigned labels via nearest neighbor classification in the low-dimensional domain. The scale parameter σ of the RBF kernel is set to a reference value in each data set within the typical range $[0.5, 1]$ where the best accuracy is attained. We have fixed the weight parameter as $\mu = 0.01$ in all setups, and set the dimension of the embedding as equal to the number of classes. In order to study the properties of the interpolation function in relation with the condition parameter in (33), we also test the supervised Laplacian eigenmaps and the label encoding methods under RBF

4. We use a nonlinear version of the formulation in (Wang and Chen, 2009) by removing the constraint that the embedding be given by a linear projection of the data.



(a) (b)

Figure 10: Misclassification rates and the condition parameters of the embeddings for the COIL-20 object data set

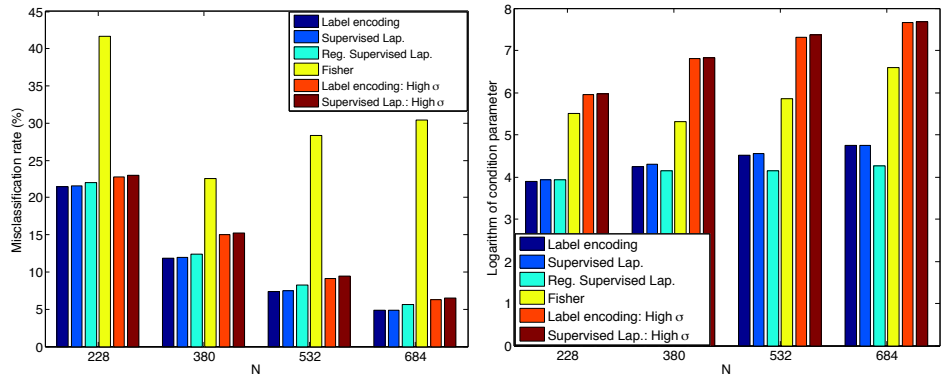


(a) (b)

Figure 11: Misclassification rates and the condition parameters of the embeddings for the ETH-80 object data set

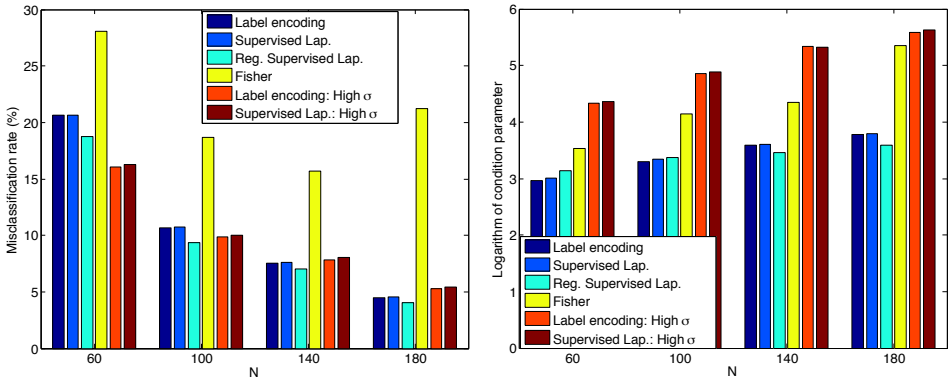
interpolators with high scale parameters, which are chosen as a few times the reference σ value giving the best results. Finally, we also include in the comparisons a regularized version of the supervised Laplacian eigenmaps embedding by controlling the magnitude of the interpolation function.

The results obtained on the COIL-20, ETH-80, Yale and reduced Yale data sets are reported respectively in Figures 10-13. In each figure, panel (a) shows the misclassification rates of the embeddings and panel (b) shows the condition parameters of the embeddings at different total number of training samples (N). The logarithm of the condition parameter is plotted for ease of visualization.



(a) (b)

Figure 12: Misclassification rates and the condition parameters of the embeddings for the Yale face data set



(a) (b)

Figure 13: Misclassification rates and the condition parameters of the embeddings for the reduced Yale face data set

The plots in Figures 10-13 show that the label encoding, supervised Laplacian eigenmaps, and the regularized supervised Laplacian eigenmaps embeddings yield better classification accuracy than the other three methods (supervised Fisher, and the embeddings with high scale parameters) in all experiments, with the only exception of the cases $N = 60$ and $N = 100$ for the reduced Yale data set. Meanwhile, examining the condition parameters of the embeddings, we observe that label encoding, supervised Laplacian eigenmaps, and the regularized supervised Laplacian eigenmaps embeddings always have a smaller condition parameter than the other three methods. This observation confirms the intuition provided by the necessary conditions of Theorems 8 and 9: A compromise between the separation and the interpolator regularity is required for good classification accuracy. The increase in the condition parameter as N increases is since the coefficient bound $\mathcal{C}$ involves a summation over all training samples. The reason why the embeddings with high σ parameters yield better classification accuracy than the other ones in the cases $N = 60$ and $N = 100$ for the reduced Yale data set is that a larger RBF scale helps better cover up the ambient space when the number of training samples is particularly low.

In the COIL-20 and the reduced Yale data sets, the best classification accuracy is obtained with the regularized supervised Laplacian eigenmaps method, while this is also the method having the smallest condition number, except for the smallest two values of N in the reduced Yale data set. In the ETH-80 and Full Yale data sets, the classification accuracy of label encoding attains that of the supervised Laplacian eigenmaps method. The condition parameter of the label encoding embedding is relatively small in these two data sets; in fact, in ETH-80 the label encoding embedding has the smallest condition number among all methods. This may be useful for explaining why this simple classification method has quite favorable performance in this data set. Likewise, if we leave aside the versions of the methods with high-scale interpolators, the Fisher embedding has the highest misclassification rate compared to label encoding, the supervised Laplacian, and the regularized supervised Laplacian embeddings, while it also has the highest condition parameter among these methods. [5]

To conclude, the results in this section suggest that the experimental findings are in agreement with the main results in Section 2.3, justifying the pertinence of the conditions (14) and (16) to classification accuracy, hence suggesting that a balance must be sought between the separability margin of the embedding and the regularity of the interpolation function in supervised manifold learning.

5. Conclusions

Most of the current supervised manifold learning algorithms focus on learning representations of training data, while the generalization properties of these representations have not been understood well yet. In this work, we have proposed a theoretical analysis of the performance of supervised manifold learning methods. We have presented generalization bounds for nonlinear supervised manifold learning algorithms and explored how the classification accuracy relates to several setup parameters such as the linear separation

5. The formulation in (34) has been observed to give highly polarized embeddings in (Vural and Guillemot, 2016a), where the samples of only few classes stretch out along each dimension and all the other classes are mapped close to zero.

margin of the embedding, the regularity of the interpolation function, the number of training samples, and the intrinsic dimensions of the class supports (manifolds). Our results suggest that embeddings of training data with good generalization capacities must allow the construction of sufficiently regular interpolation functions that extend the mapping to new data. We have then examined whether the assumption of linear separability is easy to satisfy for structure-preserving supervised embedding algorithms. We have taken the supervised Laplacian eigenmaps algorithms as reference, and showed that these methods can yield linearly separable embeddings. Providing insight about the generalization capabilities of supervised dimensionality reduction algorithms, our findings can be helpful in the classification of low-dimensional data sets.

Acknowledgments

We would like to thank Pascal Frossard and Alhussein Fawzi for the helpful discussions that contributed to this study.

Appendix A. Proof of the Results in Section 2

A.1 Proof of Theorem 2

Proof Given x, let $x_i \in \mathcal{X}$ be the nearest neighbor of x in $\mathcal{X}$ that is sampled from ν_m

$$i = \arg\min_j \|x - x_j\| \text{ s.t. } x_j \sim \nu_m.$$

Due to the separation hypothesis,

$$\omega_{mk}^T y_i + b_{mk} > \gamma/2, \quad \forall k = 1, \ldots, M-1.$$

We have

$$\begin{aligned}\omega_{mk}^T f(x) + b_{mk} &= \omega_{mk}^T f(x_i) + b_{mk} + \omega_{mk}^T (f(x) - f(x_i)) \\ &\geq \omega_{mk}^T y_i + b_{mk} - \left|\omega_{mk}^T (f(x) - f(x_i))\right| \\ &> \gamma/2 - \|f(x) - f(x_i)\| \ \geq \ \gamma/2 - L\|x - x_i\|.\end{aligned}$$

Then if the condition $L\|x - x_i\| \leq \gamma/2$ is satisfied, from the above inequality we have $\omega_{mk}^T f(x) + b_{mk} > 0$ for all $k = 1, \ldots, M-1$. This gives $\hat{C}(x) = m$ and thus ensures that x is classified correctly.

In the sequel, we lower bound the probability that the distance $\|x - x_i\|$ between x and its nearest neighbor from the same class is smaller than $\gamma/2$. We employ the following result by Kulkarni and Posner (1995). It is demonstrated in the proof of Theorem 1 in (Kulkarni and Posner, 1995) that, if $\mathcal{X}$ contains at least N_m samples drawn i.i.d. from ν_m such that $N_m \geq \mathcal{N}(\epsilon/2, \mathcal{M}_m)$ for some $\epsilon > 0$, then the probability of $\|x - x_i\|$ being larger than ϵ can be upper bounded in terms of the covering number of $\mathcal{M}_m$ as

$$P(\|x - x_i\| > \epsilon) \leq \frac{\mathcal{N}(\epsilon/2, \mathcal{M}_m)}{2N_m}.$$

Therefore, for any ϵ such that $\epsilon \leq \gamma/(2L)$ and $N_m \geq \mathcal{N}(\epsilon/2, \mathcal{M}_m)$, with probability at least $1 - \mathcal{N}(\epsilon/2, \mathcal{M}_m)/(2N_m)$, we have

$$\|x - x_i\| \leq \epsilon \leq \gamma/(2L),$$

thus, the class label of x is correctly estimated as $\hat{C}(x) = m$ due to the above discussion. ■

A.2 Proof of Lemma 3

Proof We first bound the deviation of $f(x)$ from the sample average of f in the neighborhood of x as

$$\left\| f(x) - \frac{1}{Q} \sum_{x_j \in A} f(x_j) \right\| \leq \|f(x) - m_f\| + \left\| \frac{1}{Q} \sum_{x_j \in A} f(x_j) - m_f \right\|, \tag{35}$$

where m_f is the conditional expectation of $f(u)$, given $u \in B_\delta(x)$

$$m_f = \mathbb{E}_u\big[f(u) \,|\, u \in B_\delta(x)\big] = \frac{1}{\nu_m\big(B_\delta(x)\big)} \int_{B_\delta(x)} f(u)\, d\nu_m(u).$$

The first term in (35) can be bounded as

$$\begin{aligned} \|f(x) - m_f\| &= \left\| \frac{1}{\nu_m\big(B_\delta(x)\big)} \int_{B_\delta(x)} \big(f(x) - f(u)\big)\, d\nu_m(u) \right\| \\ &\leq \frac{1}{\nu_m\big(B_\delta(x)\big)} \int_{B_\delta(x)} \|f(x) - f(u)\|\, d\nu_m(u) \leq \frac{1}{\nu_m\big(B_\delta(x)\big)} \int_{B_\delta(x)} L\|x - u\|\, d\nu_m(u) \\ &\leq \frac{1}{\nu_m\big(B_\delta(x)\big)} \int_{B_\delta(x)} L\delta\, d\nu_m(u) = L\delta, \end{aligned} \tag{36}$$

where the second inequality follows from the fact that f is Lipschitz continuous on the support $\mathcal{M}_m$, where the measure ν_m is nonzero.

The second term in (35) is given by

$$\left\| \frac{1}{Q} \sum_{x_j \in A} f(x_j) - m_f \right\| = \left(\sum_{k=1}^{d} \left| \frac{1}{Q} \sum_{x_j \in A} f^k(x_j) - m_f^k \right|^2 \right)^{1/2}, \tag{37}$$

where m_f^k denotes the k-th component of m_f, for $k = 1, \ldots, d$. Consider the random variables $f^k(x_j)$. Defining

$$f^k_{\min} = \inf_{u \in B_\delta(x)} f^k(u), \qquad f^k_{\max} = \sup_{u \in B_\delta(x)} f^k(u),$$

it follows that $f^k_{\max} - f^k_{\min} \leq 2L\delta$ due to the Lipschitz continuity of f. Then from Hoeffding's inequality, we have

$$P\left(\left| \frac{1}{Q} \sum_{x_j \in A} f^k(x_j) - m_f^k \right| \geq \epsilon \right) \leq 2\exp\left(-\frac{2Q\epsilon^2}{(f^k_{\max} - f^k_{\min})^2} \right) \leq 2\exp\left(-\frac{Q\epsilon^2}{2L^2\delta^2} \right).$$

From the union bound, we get that with probability at least $1 - 2d \exp\left(-\frac{Q\epsilon^2}{2L^2\delta^2}\right)$, for all k

$$\left| \frac{1}{Q} \sum_{x_j \in A} f^k(x_j) - m_f^k \right| \leq \epsilon,$$

which yields from (37)

$$\left\| \frac{1}{Q} \sum_{x_j \in A} f(x_j) - m_f \right\| \leq \sqrt{d}\epsilon.$$

Combining this result with the bound in (36), we conclude that with probability at least $1 - 2d \exp\left(-\frac{Q\epsilon^2}{2L^2\delta^2}\right)$

$$\left\| f(x) - \frac{1}{Q} \sum_{x_j \in A} f(x_j) \right\| \leq L\delta + \sqrt{d}\epsilon.$$

■

A.3 Proof of Theorem 5

Proof Given the test sample x and a training sample x_i drawn i.i.d. with respect to ν_m, the probability that x_i lies within a δ-neighborhood of x is given by

$$P(x_i \in B_\delta(x)) = \nu_m(B_\delta(x)) \geq \eta_{m,\delta}.$$

Then, among the N_m samples drawn with respect to ν_m, the probability that $B_\delta(x)$ contains at least Q samples is given by

$$\begin{aligned} P(|A| \geq Q) &= \sum_{q=Q}^{N_m} \binom{N_m}{q} \Big(\nu_m(B_\delta(x))\Big)^q \Big(1 - \nu_m(B_\delta(x))\Big)^{N_m - q} \\ &\geq \sum_{q=Q}^{N_m} \binom{N_m}{q} (\eta_{m,\delta})^q \, (1 - \eta_{m,\delta})^{N_m - q}, \end{aligned}$$

where the set A is defined as in (5). The last expression above is the probability of having at least Q successes out of N_m realizations of a Bernoulli random variable with probability parameter $\eta_{m,\delta}$. This probability can be lower bounded using a tail bound for binomial distributions. We thus have

$$P(|A| \geq Q) \geq 1 - \exp\left(\frac{-2\,(N_m\,\eta_{m,\delta} - Q)^2}{N_m}\right),$$

which simply follows from interpreting $|A|$ as the sum of of N_m i.i.d. observations of a Bernoulli distributed random variable and then applying Hoeffding's inequality as shown by Herbrich (1999), under the hypothesis that $N_m > Q/\eta_{m,\delta}$.

Assuming that $B_\delta(x)$ contains at least Q samples, Lemma 3 states that with probability at least

$$1 - 2d \exp\left(-\frac{|A|\epsilon^2}{2L^2\delta^2}\right) \geq 1 - 2d \exp\left(-\frac{Q\epsilon^2}{2L^2\delta^2}\right)$$

the deviation between $f(x)$ and the sample average of its neighbors is bounded as

$$\left\| f(x) - \frac{1}{|A|} \sum_{x_j \in A} f(x_j) \right\| \leq L\delta + \sqrt{d}\epsilon.$$

Hence, with probability at least

$$\begin{aligned} &\left(1 - \exp\left(\frac{-2\left(N_m\, \eta_{m,\delta} - Q\right)^2}{N_m}\right)\right)\left(1 - 2d \exp\left(-\frac{Q\epsilon^2}{2L^2\delta^2}\right)\right) \\ &\qquad \geq 1 - \exp\left(\frac{-2\left(N_m\, \eta_{m,\delta} - Q\right)^2}{N_m}\right) - 2d \exp\left(-\frac{Q\epsilon^2}{2L^2\delta^2}\right) \end{aligned}$$

we have

$$\left\| f(x) - \frac{1}{|A|} \sum_{x_j \in A} f(x_j) \right\| \leq L\delta + \sqrt{d}\epsilon. \tag{38}$$

The class label of a test sample x drawn from ν_m is correctly estimated with respect to the classifier (4) if

$$\omega_{mk}^T f(x) + b_{mk} > 0, \quad \forall k = 1, \ldots, M-1,\, k \neq m.$$

If the condition in (38) is satisfied, for all $k \neq m$, we have

$$\begin{aligned} \omega_{mk}^T f(x) + b_{mk} &= \omega_{mk}^T \frac{1}{|A|} \sum_{x_j \in A} f(x_j) + b_{mk} + \omega_{mk}^T \left(f(x) - \frac{1}{|A|} \sum_{x_j \in A} f(x_j) \right) \\ &\geq \omega_{mk}^T \frac{1}{|A|} \sum_{x_j \in A} f(x_j) + b_{mk} - \| f(x) - \frac{1}{|A|} \sum_{x_j \in A} f(x_j) \| \\ &> \gamma_Q/2 - \| f(x) - \frac{1}{|A|} \sum_{x_j \in A} f(x_j) \| \geq \gamma_Q/2 - L\delta - \sqrt{d}\epsilon \geq 0. \end{aligned}$$

Here, we obtain the second inequality from the hypothesis that the embedding is Q-mean separable with margin larger than γ_Q, which implies that the embedding is also R-mean separable with margin larger than γ_Q, for $R > Q$. Then the last inequality is due to the condition (8) on the interpolation function in the theorem. We thus get that with probability at least

$$1 - \exp\left(\frac{-2\left(N_m\, \eta_{m,\delta} - Q\right)^2}{N_m}\right) - 2d \exp\left(-\frac{Q\epsilon^2}{2L^2\delta^2}\right),$$

$\omega_{mk}^T f(x) + b_{mk} > 0$ for all $k \neq m$, hence, the sample x is correctly classified. This concludes the proof of the theorem. ■

A.4 Proof of Theorem 6

Proof Remember from the proof of Theorem 5 that with probability at least

$$1 - \exp\left(\frac{-2\left(N_m\,\eta_{m,\delta} - Q\right)^2}{N_m}\right) - 2d\exp\left(-\frac{Q\epsilon^2}{2L^2\delta^2}\right)$$

the δ-neighborhood $B_\delta(x)$ of a test sample x from class m contains at least Q samples from class m, and

$$\left\| f(x) - \frac{1}{|A|}\sum_{x_j\in A} f(x_j)\right\| \leq L\delta + \sqrt{d}\epsilon, \tag{39}$$

where A is the set of training samples in $B_\delta(x)$ from class m.

Let $x_i, x_j \in A$ be two training samples from class m in $B_\delta(x)$. As $\|x_i - x_j\| \leq 2\delta$, by the hypothesis on the embedding, we have $\|y_i - y_j\| = \|f(x_i) - f(x_j)\| \leq D_{2\delta}$, which gives

$$\|f(x_i) - \frac{1}{|A|}\sum_{x_j\in A} f(x_j)\| = \left\|\frac{1}{|A|}\sum_{x_j\in A}\big(f(x_i) - f(x_j)\big)\right\| \leq \frac{1}{|A|}\sum_{x_j\in A}\|f(x_i) - f(x_j)\| \leq D_{2\delta}.$$

Then, for any $x_i \in B_\delta(x)$,

$$\begin{aligned}\|f(x) - f(x_i)\| &= \|f(x) - \frac{1}{|A|}\sum_{x_j\in A} f(x_j) + \frac{1}{|A|}\sum_{x_j\in A} f(x_j) - f(x_i)\| \\ &\leq \|f(x) - \frac{1}{|A|}\sum_{x_j\in A} f(x_j)\| + D_{2\delta}.\end{aligned}$$

Combining this with (39), we get that with probability at least

$$1 - \exp\left(\frac{-2\left(N_m\,\eta_{m,\delta} - Q\right)^2}{N_m}\right) - 2d\exp\left(-\frac{Q\epsilon^2}{2L^2\delta^2}\right)$$

$B_\delta(x)$ will contain at least Q samples x_i from class m such that

$$\|f(x) - f(x_i)\| \leq L\delta + \sqrt{d}\epsilon + D_{2\delta}. \tag{40}$$

Now, assuming (40), let x_i' be a training sample from another class (other than m). We have

$$\|f(x) - f(x_i')\| \geq \|f(x_i) - f(x_i')\| - \|f(x) - f(x_i)\| > \gamma - (L\delta + \sqrt{d}\epsilon + D_{2\delta}),$$

which follows from (40) and the hypothesis on the embedding that $\|f(x_i) - f(x_i')\| > \gamma$.

It follows from the condition (10) that $\gamma \geq 2L\delta + 2\sqrt{d}\epsilon + 2D_{2\delta}$. Using this in the above equation, we get

$$\|f(x) - f(x_i')\| > L\delta + \sqrt{d}\epsilon + D_{2\delta}.$$

This means that the distance of $f(x)$ to the embedding of any other sample from another class is more than $L\delta + \sqrt{d}\epsilon + D_{2\delta}$, while there are samples from its own class within a distance of $L\delta + \sqrt{d}\epsilon + D_{2\delta}$ to $f(x)$. Therefore, x is classified correctly with nearest-neighbor classification in the low-dimensional domain of embedding.

■

A.5 Proof of Lemma 7

Proof The deviation of each component $f^k(x)$ of the interpolator from the sample average in the neighborhood of x is given by

$$\left| f^k(x) - \frac{1}{Q} \sum_{x_j \in A} f^k(x_j) \right| = \left| \sum_{i=1}^{N} c_i^k \left(\phi(\|x - x_i\|) - \frac{1}{Q} \sum_{x_j \in A} \phi(\|x_j - x_i\|) \right) \right|. \tag{41}$$

We thus proceed by studying the term

$$\phi(\|x - x_i\|) - \frac{1}{Q} \sum_{x_j \in A} \phi(\|x_j - x_i\|), \tag{42}$$

which will then be used in the above expression to arrive at the stated result.

Now let $x_i \in \mathcal{X}$ be any training sample. In order to study the term in (42), we first look at

$$\left| \phi(\|x - x_i\|) - \mathbb{E}_u\big[\phi(\|u - x_i\|) \,|\, u \in B_\delta(x)\big] \right|,$$

where $\mathbb{E}_u\big[\phi(\|u - x_i\|) \,|\, u \in B_\delta(x)\big]$ denotes the conditional expectation of $\phi(\|u - x_i\|)$ over u, given $u \in B_\delta(x)$. The conditional expectation is given by

$$\mathbb{E}_u\big[\phi(\|u - x_i\|) \,|\, u \in B_\delta(x)\big] = \frac{1}{\nu_m\big(B_\delta(x)\big)} \int_{B_\delta(x)} \phi(\|u - x_i\|)\, d\nu_m(u).$$

We have

$$\begin{aligned}
&\left| \phi(\|x - x_i\|) - \mathbb{E}_u\big[\phi(\|u - x_i\|) \,|\, u \in B_\delta(x)\big] \right| \\
&\qquad = \frac{1}{\nu_m\big(B_\delta(x)\big)} \left| \int_{B_\delta(x)} \big(\phi(\|x - x_i\|) - \phi(\|u - x_i\|)\big)\, d\nu_m(u) \right| \\
&\qquad \le \frac{1}{\nu_m\big(B_\delta(x)\big)} \int_{B_\delta(x)} \big|\phi(\|x - x_i\|) - \phi(\|u - x_i\|)\big|\, d\nu_m(u).
\end{aligned}$$

The term in the integral is bounded as

$$\big|\phi(\|x - x_i\|) - \phi(\|u - x_i\|)\big| \le L_\phi \big|\|x - x_i\| - \|u - x_i\|\big| \le L_\phi \|x - u\|.$$

Using this in the above term, we get

$$\begin{aligned}
&\left| \phi(\|x - x_i\|) - \mathbb{E}_u\big[\phi(\|u - x_i\|) \,|\, u \in B_\delta(x)\big] \right| \\
&\qquad \le \frac{L_\phi}{\nu_m\big(B_\delta(x)\big)} \int_{B_\delta(x)} \|x - u\|\, d\nu_m(u) = L_\phi\, \mathbb{E}_u\big[\|u - x\| \,|\, u \in B_\delta(x)\big] \\
&\qquad \le L_\phi\, \delta.
\end{aligned} \tag{43}$$

We now analyze the term in (42) for a given x_i for two different cases, i.e., for $x_i \notin B_\delta(x)$ and $x_i \in B_\delta(x)$. We first look at the case $x_i \notin B_\delta(x)$. For $x_j \in B_\delta(x)$, let

$$\zeta_j := \phi(\|x_j - x_i\|).$$

The observations ζ_j are i.i.d. (since x_j are i.i.d.) with mean $m_\zeta = \mathbb{E}_u\big[\phi(\|u-x_i\|) \,|\, u \in B_\delta(x)\big]$ and take values in the interval $\zeta_{\min} \leq \zeta_j \leq \zeta_{\max}$, where

$$\zeta_{\min} := \inf_{u \in B_\delta(x)} \phi(\|u - x_i\|), \qquad \zeta_{\max} := \sup_{u \in B_\delta(x)} \phi(\|u - x_i\|).$$

Since for any $u_1, u_2 \in B_\delta(x)$, $\|u_1 - u_2\| \leq 2\delta$, it follows from the Lipschitz continuity of ϕ that $\zeta_{\max} - \zeta_{\min} \leq 2L_\phi\delta$. Using this together with the Hoeffding's inequality, we get

$$P\bigg(\bigg|\frac{1}{Q}\sum_{x_j \in A} \zeta_j - m_\zeta\bigg| \geq \epsilon\bigg) \leq 2\exp\bigg(-\frac{2\,Q\,\epsilon^2}{(\zeta_{\max} - \zeta_{\min})^2}\bigg) \leq 2\exp\bigg(-\frac{Q\,\epsilon^2}{2L_\phi^2\delta^2}\bigg). \tag{44}$$

We have

$$\bigg|\phi(\|x - x_i\|) - \frac{1}{Q}\sum_{x_j \in A} \phi(\|x_j - x_i\|)\bigg| \leq \big|\phi(\|x - x_i\|) - m_\zeta\big| + \bigg|m_\zeta - \frac{1}{Q}\sum_{x_j \in A} \phi(\|x_j - x_i\|)\bigg|.$$

Using (43) and (44) in the above equation, it holds with probability at least

$$1 - 2\exp\bigg(-\frac{Q\,\epsilon^2}{2L_\phi^2\delta^2}\bigg)$$

that

$$\bigg|\phi(\|x - x_i\|) - \frac{1}{Q}\sum_{x_j \in A} \phi(\|x_j - x_i\|)\bigg| \leq L_\phi\delta + \epsilon.$$

Next, we study the case $x_i \in B_\delta(x)$. For any fixed $x_i \in B_\delta(x)$, hence $x_i \in A$, we have

$$\begin{aligned}
&\bigg|\phi(\|x - x_i\|) - \frac{1}{Q}\sum_{x_j \in A} \phi(\|x_j - x_i\|)\bigg| \\
&= \bigg|\frac{1}{Q}\phi(\|x - x_i\|) + \frac{Q-1}{Q}\phi(\|x - x_i\|) - \frac{1}{Q}\phi(\|x_i - x_i\|) - \frac{1}{Q}\sum_{x_j \in A\setminus\{x_i\}} \phi(\|x_j - x_i\|)\bigg| \\
&\leq \frac{1}{Q}\bigg|\phi(\|x - x_i\|) - \phi(\|x_i - x_i\|)\bigg| + \frac{Q-1}{Q}\bigg|\phi(\|x - x_i\|) - \frac{1}{Q-1}\sum_{x_j \in A\setminus\{x_i\}} \phi(\|x_j - x_i\|)\bigg|.
\end{aligned}$$

The first term above is bounded as

$$\frac{1}{Q}\bigg|\phi(\|x - x_i\|) - \phi(\|x_i - x_i\|)\bigg| \leq \frac{L_\phi\delta}{Q}.$$

Next, similarly to the analysis of the case $x_i \neq B_\delta(x)$, we get that for $x_i \in B_\delta(x)$ with probability at least

$$1 - 2\exp\bigg(-\frac{(Q-1)\,\epsilon^2}{2L_\phi^2\delta^2}\bigg)$$

it holds that

$$\bigg|\phi(\|x - x_i\|) - \frac{1}{Q-1}\sum_{x_j \in A\setminus\{x_i\}} \phi(\|x_j - x_i\|)\bigg| \leq L_\phi\delta + \epsilon,$$

hence

$$\left|\phi(\|x-x_i\|)-\frac{1}{Q}\sum_{x_j\in A}\phi(\|x_j-x_i\|)\right|\leq\frac{L_\phi\delta}{Q}+\frac{Q-1}{Q}(L_\phi\delta+\epsilon)\leq L_\phi\delta+\epsilon.$$

Combining the analyses of the cases $x_i\neq B_\delta(x)$ and $x_i\in B_\delta(x)$, we conclude that for any given $x_i\in\mathcal{X}$,

$$P\left(\left|\phi(\|x-x_i\|)-\frac{1}{Q}\sum_{x_j\in A}\phi(\|x_j-x_i\|)\right|\leq L_\phi\delta+\epsilon\right)\geq 1-2\exp\left(-\frac{(Q-1)\,\epsilon^2}{2L_\phi^2\delta^2}\right).$$

Therefore, applying the union bound on all N samples x_i in $\mathcal{X}$, we get that with probability at least

$$1-2N\exp\left(-\frac{(Q-1)\,\epsilon^2}{2L_\phi^2\delta^2}\right)$$

it holds that

$$\left|\phi(\|x-x_i\|)-\frac{1}{Q}\sum_{x_j\in A}\phi(\|x_j-x_i\|)\right|\leq L_\phi\delta+\epsilon \tag{45}$$

for all $x_i\in\mathcal{X}$.

We can now use this in (41) to bound the deviation of $f^k(x)$ from the empirical mean of f^k in the neighbourhood of x. Assuming that the condition (45) holds for all $x_i\in\mathcal{X}$, we obtain

$$\begin{aligned}\left|f^k(x)-\frac{1}{Q}\sum_{x_j\in A}f^k(x_j)\right|&=\left|\sum_{i=1}^N c_i^k\left(\phi(\|x-x_i\|)-\frac{1}{Q}\sum_{x_j\in A}\phi(\|x_j-x_i\|)\right)\right|\\&\leq(L_\phi\delta+\epsilon)\sum_{i=1}^N|c_i^k|\leq\mathcal{C}(L_\phi\delta+\epsilon),\end{aligned}$$

which gives

$$\|f(x)-\frac{1}{Q}\sum_{x_j\in A}f(x_j)\|=\left(\sum_{k=1}^d\left(f^k(x)-\frac{1}{Q}\sum_{x_j\in A}f^k(x_j)\right)^2\right)^{1/2}\leq\sqrt{d}\mathcal{C}(L_\phi\delta+\epsilon).$$

We thus get

$$P\left(\|f(x)-\frac{1}{Q}\sum_{x_j\in A}f(x_j)\|\leq\sqrt{d}\mathcal{C}(L_\phi\delta+\epsilon)\right)\geq 1-2N\exp\left(-\frac{(Q-1)\,\epsilon^2}{2L_\phi^2\delta^2}\right),$$

which completes the proof. ■

A.6 Proof of Theorem 8

Proof

Remember from the proof of Theorem 5 that

$$P(|A| \geq Q) \geq 1 - \exp\left(\frac{-2\left(N_m\,\eta_{m,\delta} - Q\right)^2}{N_m}\right).$$

Lemma 7 states that, if $B_\delta(x)$ contains at least Q samples from class m, i.e., $|A| \geq Q$, then

$$\begin{aligned} P\left(\|f(x) - \frac{1}{|A|}\sum_{x_j \in A} f(x_j)\| \leq \sqrt{d}\mathcal{C}(L_\phi\delta + \epsilon)\right) &\geq 1 - 2N\exp\left(-\frac{(|A|-1)\,\epsilon^2}{2L_\phi^2\delta^2}\right) \\ &\geq 1 - 2N\exp\left(-\frac{(Q-1)\,\epsilon^2}{2L_\phi^2\delta^2}\right). \end{aligned}$$

Hence, combining these two results (multiplying both probabilities), we get that with probability at least

$$\begin{aligned} &\left(1 - \exp\left(\frac{-2\left(N_m\,\eta_{m,\delta} - Q\right)^2}{N_m}\right)\right)\left(1 - 2N\exp\left(-\frac{(Q-1)\,\epsilon^2}{2L_\phi^2\delta^2}\right)\right) \\ &\quad\geq 1 - \exp\left(\frac{-2\left(N_m\,\eta_{m,\delta} - Q\right)^2}{N_m}\right) - 2N\exp\left(-\frac{(Q-1)\,\epsilon^2}{2L_\phi^2\delta^2}\right) \end{aligned}$$

it holds that

$$\|f(x) - \frac{1}{|A|}\sum_{x_j \in A} f(x_j)\| \leq \sqrt{d}\mathcal{C}(L_\phi\delta + \epsilon). \tag{46}$$

A test sample x drawn from ν_m is classified correctly with the linear classifier if

$$\omega_{mk}^T\, f(x) + b_{mk} > 0, \quad \forall k = 1, \ldots, M-1,\ k \neq m.$$

If the condition in (46) is satisfied, for all $k \neq m$, we have

$$\begin{aligned} \omega_{mk}^T\, f(x) + b_{mk} &= \omega_{mk}^T \frac{1}{|A|}\sum_{x_j \in A} f(x_j) + b_{mk} + \omega_{mk}^T\left(f(x) - \frac{1}{|A|}\sum_{x_j \in A} f(x_j)\right) \\ &\geq \omega_{mk}^T \frac{1}{|A|}\sum_{x_j \in A} f(x_j) + b_{mk} - \|f(x) - \frac{1}{|A|}\sum_{x_j \in A} f(x_j)\| \\ &> \gamma_Q/2 - \|f(x) - \frac{1}{|A|}\sum_{x_j \in A} f(x_j)\| \geq \gamma_Q/2 - \sqrt{d}\mathcal{C}(L_\phi\delta + \epsilon) \geq 0. \end{aligned}$$

We thus conclude that with probability at least

$$1 - \exp\left(\frac{-2\left(N_m\,\eta_{m,\delta} - Q\right)^2}{N_m}\right) - 2N\exp\left(-\frac{(Q-1)\,\epsilon^2}{2L_\phi^2\delta^2}\right),$$

$\omega_{mk}^T\, f(x) + b_{mk} > 0$ for all $k \neq m$, hence, the class label of x is estimated correctly. ■

A.7 Proof of Theorem 9

Proof First, recall from the proof of Theorem 8 that, with probability at least

$$1 - \exp\left(\frac{-2\left(N_m\,\eta_{m,\delta} - Q\right)^2}{N_m}\right) - 2N\exp\left(-\frac{(Q-1)\,\epsilon^2}{2L_\phi^2\delta^2}\right)$$

the δ-neighborhood $B_\delta(x)$ of a test sample x from class m contains at least Q samples from class m, and

$$\|f(x) - \frac{1}{|A|}\sum_{x_j\in A} f(x_j)\| \leq \sqrt{d}\mathcal{C}(L_\phi\delta + \epsilon), \tag{47}$$

where A is the set of training samples in $B_\delta(x)$ from class m.

Then it is easy to show that (as in the proof of Theorem 6), with probability at least

$$1 - \exp\left(\frac{-2\left(N_m\,\eta_{m,\delta} - Q\right)^2}{N_m}\right) - 2N\exp\left(-\frac{(Q-1)\,\epsilon^2}{2L_\phi^2\delta^2}\right),$$

$B_\delta(x)$ will contain at least Q samples x_i from class m such that

$$\|f(x) - f(x_i)\| \leq \sqrt{d}\mathcal{C}(L_\phi\delta + \epsilon) + D_{2\delta}. \tag{48}$$

Hence, for a training sample x_i' from another class (other than m), we have

$$\|f(x) - f(x_i')\| \geq \|f(x_i) - f(x_i')\| - \|f(x) - f(x_i)\| > \gamma - (\sqrt{d}\mathcal{C}(L_\phi\delta + \epsilon) + D_{2\delta}),$$

which follows from (48) and the hypothesis on the embedding that $\|f(x_i) - f(x_i')\| > \gamma$.

Due to the condition (16), we have $\gamma \geq 2\sqrt{d}\,\mathcal{C}\,(L_\phi\delta + \epsilon) + 2D_{2\delta}$. Using this above equation, we obtain

$$\|f(x) - f(x_i')\| > \sqrt{d}\mathcal{C}(L_\phi\delta + \epsilon) + D_{2\delta}.$$

Therefore, the distance of $f(x)$ to the embedding of the samples from other classes is more than $\sqrt{d}\mathcal{C}(L_\phi\delta + \epsilon) + D_{2\delta}$, while there are samples from its own class within a distance of $\sqrt{d}\mathcal{C}(L_\phi\delta + \epsilon) + D_{2\delta}$ to $f(x)$. We thus conclude that the class label of x is estimated correctly with nearest-neighbor classification in the low-dimensional domain of embedding. ■

Appendix B. Proof of the Results in Section 3

B.1 Proof of Lemma 13

Proof The coordinate vector y is the eigenvector of the matrix $L_w - \mu L_b$ corresponding to its minimum eigenvalue. Hence,

$$y = \arg\min_{\substack{\xi \\ \|\xi\|=1}} \xi^T(L_w - \mu L_b)\xi.$$

Equivalently, defining the degree-normalized coordinates $z = D_w^{-1/2} y$, and thus replacing the above ξ by $D_w^{1/2}\xi$, we have

$$\begin{aligned} z &= \arg \min_{\substack{\xi \\ \xi^T D_w \xi = 1}} N(\xi) \\ N(\xi) &= \xi^T D_w^{1/2} (L_w - \mu L_b) D_w^{1/2} \xi \\ &= \xi^T (D_w - W_w)\xi - \mu\, \xi^T (D_w D_b^{-1})^{1/2} \, (D_b - W_b) \, (D_b^{-1} D_w)^{1/2} \xi. \end{aligned} \tag{49}$$

Then, denoting $\beta_i = d_w(i)/d_b(i)$, the term $N(\xi)$ can be rearranged as

$$\begin{aligned} N(\xi) &= \sum_i \xi_i \Big(d_w(i)\, \xi_i - \sum_{j \sim_w i} \xi_j \, w_{ij} \Big) - \mu \sum_i \xi_i \Big(d_w(i)\, \xi_i - \sum_{j \sim_b i} \xi_j \, w_{ij} \, \sqrt{\beta_i \beta_j} \Big) \\ &= \sum_i \xi_i \sum_{j \sim_w i} (\xi_i - \xi_j) w_{ij} - \mu \sum_i \xi_i \sum_{j \sim_b i} (\beta_i \xi_i - \sqrt{\beta_i \beta_j}\, \xi_j) w_{ij} \\ &= \sum_i \sum_{j \sim_w i} (\xi_i^2 - \xi_i \xi_j) w_{ij} - \mu \sum_i \sum_{j \sim_b i} (\beta_i \xi_i^2 - \sqrt{\beta_i \beta_j} \xi_i \xi_j) w_{ij}, \end{aligned}$$

which gives [6]

$$N(\xi) = \sum_{i \sim_w j} (\xi_i - \xi_j)^2 \, w_{ij} - \mu \sum_{i \sim_b j} \Big(\sqrt{\beta_i} \xi_i - \sqrt{\beta_j} \xi_j \Big)^2 \, w_{ij} \tag{50}$$

by grouping the neighboring (i, j) pairs in the inner sums. Now, for any $\xi \in \mathbb{R}^{N \times 1}$ such that $\xi^T D_w \xi = 1$, we define ξ^* as follows

$$\xi_i^* = \begin{cases} -|\xi_i| & \text{if } C_i = 1 \\ |\xi_i| & \text{if } C_i = 2. \end{cases} \tag{51}$$

Clearly, ξ^* also satisfies $(\xi^*)^T D_w \xi^* = 1$. From (50), it can be easily checked that $N(\xi^*) \leq N(\xi)$ for any ξ, Then, a minimizer z of the problem (49) has to be of the separable form defined in (51); otherwise z^* would yield a smaller value for the function N, which would contradict the fact that z is a minimizer. Note that the equality $N(z^*) = N(z)$ holds only if $z = z^*$ or $z = -z^*$, thus when z is separable. Therefore, the embedding z satisfies the condition

$$z_i \leq 0 \;\; \text{if } C_i = 1, \qquad\qquad z_i \geq 0 \;\; \text{if } C_i = 2,$$

or the equivalent condition

$$z_i \leq 0 \;\; \text{if } C_i = 2, \qquad\qquad z_i \geq 0 \;\; \text{if } C_i = 1.$$

Finally, since $y_i = \sqrt{d_w(i)}\, z_i$, the same property also holds for the embedding y.

■

6. In our notation, the terms $i \sim_w j$ and $i \sim_b j$ in the summation indices as in (50) refer to edges rather than neighboring (i, j)-pairs; i.e., each pair is counted only once in the summation.

B.2 Proof of Theorem 14

Proof From (49) and (50), we have

$$z = \arg \min_{\substack{\xi \\ \xi^T D_w \xi = 1}} \sum_{i \sim_w j} (\xi_i - \xi_j)^2 \, w_{ij} - \mu \sum_{i \sim_b j} \left(\sqrt{\beta_i} \xi_i - \sqrt{\beta_j} \xi_j \right)^2 w_{ij}. \tag{52}$$

Thus, at the optimal solution z the objective function takes the value

$$N(z) = \sum_{i \sim_w j} (z_i - z_j)^2 \, w_{ij} - \mu \sum_{i \sim_b j} \left(\sqrt{\beta_i} z_i - \sqrt{\beta_j} z_j \right)^2 w_{ij}. \tag{53}$$

In the following, we derive a lower bound for the first sum and an upper bound for the second sum in (53). We begin with the first sum. Let $i_{1,min}$, $i_{1,max}$, $i_{2,min}$ and $i_{2,max}$ denote the indices of the data samples in class 1 and class 2 that are respectively mapped to the extremal coordinates $z_{1,min}$, $z_{1,max}$, $z_{2,min}$, $z_{2,max}$, where

$$z_{k,min} = \min_{i:\, C_i = k} z_i \,, \qquad\qquad z_{k,max} = \max_{i:\, C_i = k} z_i \,.$$

Let $P_1 = \{(x_{k_{i-1}}, x_{k_i})\}_{i=1}^{L_1}$ be a shortest path of length L_1 joining $x_{i_{1,min}}$ and $x_{i_{1,max}}$ and $P_2 = \{(x_{n_{i-1}}, x_{n_i})\}_{i=1}^{L_2}$ be a shortest path of length L_2 joining $x_{i_{2,min}}$ and $x_{i_{2,max}}$. We have

$$\begin{aligned} \sum_{i \sim_w j} (z_i - z_j)^2 \, w_{ij} &\geq \sum_{i=1}^{L_1} (z_{k_i} - z_{k_{i-1}})^2 \, w_{k_{i-1} k_i} + \sum_{i=1}^{L_2} (z_{n_i} - z_{n_{i-1}})^2 \, w_{n_{i-1} n_i} \\ &\geq w_{min,1} \sum_{i=1}^{L_1} (z_{k_i} - z_{k_{i-1}})^2 + w_{min,2} \sum_{i=1}^{L_2} (z_{n_i} - z_{n_{i-1}})^2 \,, \end{aligned} \tag{54}$$

where the first inequality simply follows from the fact that the set of edges making up $P_1 \cup P_2$ are contained in the set of all edges in G_w. For a sequence $\{a_i\}_{i=0}^{L}$, the following inequality holds

$$\begin{aligned} (a_L - a_0)^2 &= \sum_{i=1}^{L} (a_i - a_{i-1})^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^{L} (a_i - a_{i-1})(a_j - a_{j-1}) \\ &\leq \sum_{i=1}^{L} (a_i - a_{i-1})^2 + \frac{1}{2} \sum_{\substack{i,j=1 \\ i \neq j}}^{L} \left((a_i - a_{i-1})^2 + (a_j - a_{j-1})^2 \right) = L \sum_{i=1}^{L} (a_i - a_{i-1})^2. \end{aligned}$$

Hence,

$$\sum_{i=1}^{L} (a_i - a_{i-1})^2 \geq \frac{1}{L} (a_L - a_0)^2.$$

Using this inequality in (54), we get

$$\sum_{i \sim_w j} (z_i - z_j)^2 \, w_{ij} \geq \frac{w_{min,1}}{L_1} (z_{1,max} - z_{1,min})^2 + \frac{w_{min,2}}{L_2} (z_{2,max} - z_{2,min})^2.$$

Since the path lengths L_1 and L_2 are upper bounded by the diameters D_1 and D_2, we finally obtain the lower bound

$$\sum_{i\sim_w j}(z_i - z_j)^2\, w_{ij} \geq \frac{w_{min,1}}{D_1}(z_{1,max} - z_{1,min})^2 + \frac{w_{min,2}}{D_2}(z_{2,max} - z_{2,min})^2. \tag{55}$$

Next, we find an upper bound for the second sum in (53). Using Lemma 13, we obtain the following inequality

$$\begin{aligned}\sum_{i\sim_b j}\left(\sqrt{\beta_i}z_i - \sqrt{\beta_j}z_j\right)^2 w_{ij} &\leq \sum_{i\sim_b j}(z_{2,max} - z_{1,min})^2\, \beta_{max}\, w_{ij} \\ &= \frac{1}{2}(z_{2,max} - z_{1,min})^2\, \beta_{max} V^b_{max}.\end{aligned} \tag{56}$$

Now, since the solution z in (52) minimizes the objective function $N(\xi)$, we have

$$N(z) = \lambda_{\min}(L_w - \mu L_b),$$

where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ respectively denote the minimum and the maximum eigenvalues of a matrix. For two Hermitian matrices A and B, the inequality $\lambda_{\min}(A+B) \leq \lambda_{\min}(A) + \lambda_{\max}(B)$ holds. As L_w and L_b are graph Laplacian matrices, we have $\lambda_{\min}(L_w) = \lambda_{\min}(L_b) = 0$ and thus

$$N(z) = \lambda_{\min}(L_w - \mu L_b) \leq \lambda_{\min}(L_w) + \lambda_{\max}(-\mu L_b) = \lambda_{\min}(L_w) - \mu\lambda_{\min}(L_b) = 0.$$

Using in (53) the above inequality and the lower and upper bounds in (55) and (56), we obtain

$$\begin{aligned}0 \geq N(z) &= \sum_{i\sim_w j}(z_i - z_j)^2\, w_{ij} - \mu\sum_{i\sim_b j}\left(\sqrt{\beta_i}z_i - \sqrt{\beta_j}z_j\right)^2 w_{ij} \\ &\geq \frac{w_{min,1}}{D_1}(z_{1,max} - z_{1,min})^2 + \frac{w_{min,2}}{D_2}(z_{2,max} - z_{2,min})^2 \\ &\quad - \frac{1}{2}\mu(z_{2,max} - z_{1,min})^2\, \beta_{max} V^b_{max}.\end{aligned}$$

Hence

$$\frac{w_{min,1}}{D_1}(z_{1,max} - z_{1,min})^2 + \frac{w_{min,2}}{D_2}(z_{2,max} - z_{2,min})^2 \leq \frac{1}{2}\mu(z_{2,max} - z_{1,min})^2\, \beta_{max} V^b_{max}. \tag{57}$$

The RHS of the above inequality is related to the overall support $z_{2,max} - z_{1,min}$ of the data, whereas the terms on the LHS are related to the individual supports $z_{1,max} - z_{1,min}$ and $z_{2,max} - z_{2,min}$ of the two classes in the learnt embedding. Meanwhile, the separation $z_{2,min} - z_{1,max}$ between the two classes is given by the gap between the overall support and the sums of the individual supports. In order to use the above inequality in view of this observation, we first derive a lower bound on the RHS term. Since $z^T D_w z = 1$, we have

$$\begin{aligned}1 = \sum_i z_i^2\, d_w(i) &= \sum_{i:\, C_i=1} z_i^2\, d_w(i) + \sum_{i:\, C_i=2} z_i^2\, d_w(i) \\ &\leq z^2_{1,min}\sum_{i:\, C_i=1} d_w(i) + z^2_{2,max}\sum_{i:\, C_i=2} d_w(i) = z^2_{1,min} V_1 + z^2_{2,max} V_2.\end{aligned}$$

This gives

$$z_{1,min}^2 + z_{2,max}^2 \geq \frac{1}{V_{max}}.$$

Hence, we obtain the following lower bound on the overall support

$$(z_{2,max} - z_{1,min})^2 \geq z_{2,max}^2 + z_{1,min}^2 \geq \frac{1}{V_{max}}. \tag{58}$$

Denoting the supports of class 1 and class 2 and the overall support as

$$S_1 = z_{1,max} - z_{1,min}, \qquad S_2 = z_{2,max} - z_{2,min}, \qquad S = z_{2,max} - z_{1,min},$$

we have from (57)

$$\overline{w}_{min}(S_1^2 + S_2^2) \leq \frac{1}{2}\,\mu\, S^2\, \beta_{max} V_{max}^b,$$

which yields the following upper bound on the total support of the two classes

$$S_1 + S_2 \leq \sqrt{2(S_1^2 + S_2^2)} \leq S\sqrt{\frac{\mu \beta_{max} V_{max}^b}{\overline{w}_{min}}}.$$

We can thus lower bound the separation $z_{2,min} - z_{1,max}$ as

$$z_{2,min} - z_{1,max} = S - (S_1 + S_2) \geq S\left(1 - \sqrt{\frac{\mu \beta_{max} V_{max}^b}{\overline{w}_{min}}}\right),$$

provided that $\mu < \overline{w}_{min}/(\beta_{max} V_{max}^b)$. From the lower bound on the overall support in (58), we lower bound the separation as follows

$$z_{2,min} - z_{1,max} \geq \frac{1}{\sqrt{V_{max}}}\left(1 - \sqrt{\frac{\mu \beta_{max} V_{max}^b}{\overline{w}_{min}}}\right).$$

Finally, since the separation of any embedding with dimension $d \geq 1$ is at least as much as the separation $z_{2,min} - z_{1,max}$ of the embedding of dimension $d = 1$, the above lower bound holds for any $d \geq 1$ as well. ■

B.3 Proof of Corollary 15

Proof The one-dimensional embedding z is given as the solution of the constrained optimization problem

$$z = \arg\min N(\xi) \text{ s.t. } D(\xi) = 1,$$

where

$$N(\xi) = \xi^T D_w^{1/2}(L_w - \mu L_b) D_w^{1/2}\xi, \qquad D(\xi) = \xi^T D_w \xi.$$

Defining the Lagrangian function

$$\Lambda(\xi, \lambda) = N(\xi) + \lambda(D(\xi) - 1)$$

at the optimal solution z, we have

$$\nabla_\xi \Lambda = \nabla_\lambda \Lambda = 0,$$

where ∇_ξ and ∇_λ respectively denote the derivatives of Λ with respect to ξ and λ. Thus, at $\xi = z$,

$$\frac{\partial \Lambda}{\partial \xi_i} = \frac{\partial N(\xi)}{\partial \xi_i} + \lambda \frac{\partial D(\xi)}{\partial \xi_i} = 0$$

for all $i = 1, \ldots, N$. From (50), the derivatives of $N(\xi)$ and $D(\xi)$ at z are given by

$$\begin{aligned}
\left.\frac{\partial N(\xi)}{\partial \xi_i}\right|_{\xi=z} &= \sum_{j \sim_w i} 2(z_i - z_j) w_{ij} - \mu \sum_{j \sim_b i} 2\left(\sqrt{\beta_i} z_i - \sqrt{\beta_j} z_j\right) \sqrt{\beta_i}\, w_{ij} \\
\left.\frac{\partial D(\xi)}{\partial \xi_i}\right|_{\xi=z} &= 2\, d_w(i)\, z_i,
\end{aligned}$$

which yields

$$\sum_{j \sim_w i} (z_i - z_j) w_{ij} - \mu \sum_{j \sim_b i} \left(\sqrt{\beta_i} z_i - \sqrt{\beta_j} z_j\right) \sqrt{\beta_i}\, w_{ij} + \lambda\, d_w(i)\, z_i = 0 \tag{59}$$

for all i. At $i = i_{1,max}$, as z attains its maximal value $z_{1,max}$ for class 1, we have

$$\begin{aligned}
\lambda\, d_w(i_{1,max})\, z_{1,max} = &\sum_{j \sim_w i_{1,max}} (z_j - z_{1,max}) w_{i_{1,max} j} \\
&+ \mu \sum_{j \sim_b i_{1,max}} \left(\sqrt{\beta_{i_{1,max}}} z_{1,max} - \sqrt{\beta_j} z_j\right) \sqrt{\beta_{i_{1,max}}}\, w_{i_{1,max} j} \\
&\leq -\mu\, \beta_{min}\, (z_{2,min} - z_{1,max})\, d_b(i_{1,max}).
\end{aligned}$$

Hence

$$|z_{1,max}| = -z_{1,max} \geq \frac{\mu\, \beta_{min}\, (z_{2,min} - z_{1,max}) d_b(i_{1,max})}{\lambda\, d_w(i_{1,max})} \geq \frac{\mu\, \beta_{min}\, (z_{2,min} - z_{1,max})}{\lambda\, \beta_{max}}. \tag{60}$$

We proceed by deriving an upper bound for λ. The gradients of $N(\xi)$ and $D(\xi)$ are given by

$$\nabla_\xi N = 2 D_w^{1/2} (L_w - \mu L_b) D_w^{1/2} \xi, \qquad \nabla_\xi D = 2 D_w \xi.$$

From the condition $\nabla_\xi \Lambda = 0$ at $\xi = z$, we have

$$\begin{aligned}
D_w^{1/2} (L_w - \mu L_b) D_w^{1/2} z + \lambda D_w z &= 0 \\
(L_w - \mu L_b) y + \lambda y &= 0.
\end{aligned}$$

Since $y = D_w^{1/2} z$ is the unit-norm eigenvector of $L_w - \mu L_b$ corresponding to its smallest eigenvalue, the Lagrangian multiplier λ is given by

$$\lambda = -\lambda_{\min}(L_w - \mu L_b).$$

We can lower bound the minimum eigenvalue as

$$\lambda_{\min}(L_w - \mu L_b) \geq \lambda_{\min}(L_w) + \lambda_{\min}(-\mu L_b) = 0 - \mu\lambda_{\max}(L_b) \geq -2\mu$$

since the eigenvalues of a graph Laplacian are upper bounded by 2. This gives $\lambda \leq 2\mu$. Using this upper bound on λ in (60), we obtain

$$|z_{1,max}| \geq \frac{1}{2}\frac{\beta_{min}}{\beta_{max}}(z_{2,min} - z_{1,max}).$$

Repeating the same steps for $i = i_{2,min}$ following (59), one can similarly show that

$$z_{2,min} \geq \frac{1}{2}\frac{\beta_{min}}{\beta_{max}}(z_{2,min} - z_{1,max}).$$

■

References

B. J. C. Baxter. *The interpolation theory of radial basis functions*. PhD thesis, Cambridge University, Trinity College, 1992.

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, June 2003.

P. J. Bickel and B. Li. Local polynomial regression on unknown manifolds. *Lecture Notes-Monograph Series*, 54:177–186, 2007.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

T. J. Chin and D. Suter. Out-of-sample extrapolation of learned manifolds. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(9):1547–1556, 2008.

F. R. K. Chung. *Spectral Graph Theory (CBMS Regional Conference Series in Mathematics, No. 92)*. American Mathematical Society, December 1996.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2002.

Y. Cui and L. Fan. A novel supervised dimensionality reduction algorithm: Graph-based fisher analysis. *Pattern Recognition*, 45(4):1471–1481, 2012.

D. L. Donoho and C. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences of the United States of America*, 100(10):5591–5596, May 2003.

A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 23(6):643–660, 2001.

X. He and P. Niyogi. Locality preserving projections. In *Advances in Neural Information Processing Systems 16*, Cambridge, MA, 2004. MIT Press.

R. Herbrich. Exact tail bounds for binomial distributed variables. *Online: Available at http://research.microsoft.com/apps/pubs/default.aspx?id=66854*, 1999.

A. Hernández-Aguirre, C. Koutsougeras, and B. P. Buckles. Sample complexity for function learning tasks through linear neural networks. *International Journal on Artificial Intelligence Tools*, 11(4):499–511, 2002.

Q. Hua, L. Bai, X. Wang, and Y. Liu. Local similarity and diversity preserving discriminant projection for face and handwriting digits recognition. *Neurocomputing*, 86:150–157, 2012.

A. N. Kolmogorov and V. M. Tihomirov. ε-entropy and ε-capacity of sets in functional spaces. *Amer. Math. Soc. Transl.*, 2(17):277–364, 1961.

S. Kpotufe. k-NN regression adapts to local intrinsic dimension. In *Proc. Advances in Neural Information Processing Systems 24*, pages 729–737, 2011.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. Advances in Neural Information Processing Systems 25*, pages 1106–1114, 2012.

S. R. Kulkarni and S. E. Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, 41(4):1028–1039, 1995.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324, 1998.

B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object categorization. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003)*, pages 409–415, 2003.

B. Li, J. Liu, Z. Zhao, and W. Zhang. Locally linear representation fisher criterion. In *The 2013 International Joint Conference on Neural Networks*, pages 1–7, 2013.

S. Lin, X. Liu, Y. Rong, and Z. Xu. Almost optimal estimates for approximation and learning by radial basis function networks. *Machine Learning*, 95(2):147–164, 2014.

F. J. Narcowich, N. Sivakumar, and J. D. Ward. On condition numbers associated with radial-function interpolation. *Journal of Mathematical Analysis and Applications*, 186(2): 457 – 485, 1994.

S. A. Nene, S. K. Nayar, and H. Murase. Columbia Object Image Library (COIL-20). Technical report, Feb 1996.

P. Niyogi and F. Girosi. On the relationship between generalization error, hypothesis complexity, and sample complexity for radial basis functions. *Neural Computation*, 8:819–842, 1996.

B. Peherstorfer, D. Pflüger, and H. J. Bungartz. A sparse-grid-based out-of-sample extension for dimensionality reduction and clustering with laplacian eigenmaps. In *AI 2011: Proc. Advances in Artificial Intelligence - 24th Australasian Joint Conference*, pages 112–121, 2011.

C. Piret. *Analytical and Numerical Advances in Radial Basis Functions*. PhD thesis, University of Colorado, 2007.

H. Qiao, P. Zhang, D. Wang, and B. Zhang. An explicit nonlinear mapping for manifold learning. *IEEE T. Cybernetics*, 43(1):51–63, 2013.

B. Raducanu and F. Dornaika. A supervised non-linear dimensionality reduction approach for manifold learning. *Pattern Recognition*, 45(6):2432–2444, 2012.

S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *The 22nd Conference on Learning Theory*, 2009.

M. Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *Journal of Machine Learning Research*, 8:1027–1061, 2007.

J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, December 2000.

M. Vidyasagar. *A Theory of Learning and Generalization*. Springer-Verlag, Secaucus, NJ, USA, 2nd edition, 1997.

E. Vural and C. Guillemot. Out-of-sample generalizations for supervised manifold learning for classification. *IEEE Transactions on Image Processing*, 25(3):1410–1424, March 2016a.

E. Vural and C. Guillemot. A study of the classification of low-dimensional data with supervised manifold learning. *Technical report*. Available at https://arxiv.org/abs/1507.05880, 2016b.

R. Wang and X. Chen. Manifold discriminant analysis. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 429–436, 2009.

W. Yang, C. Sun, and L. Zhang. A multi-manifold discriminant analysis method for image feature extraction. *Pattern Recognition*, 44(8):1649–1657, 2011.

Z. Zhang and H. Zha. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *SIAM Journal of Scientific Computing*, 26:313–338, 2005.

Z. Zhang, M. Zhao, and T. Chow. Marginal semi-supervised sub-manifold projections with informative constraints for dimensionality reduction and recognition. *Neural Networks*, 36:97–111, 2012.