Johann Radon Institute for
Computational and Applied Mathematics
Austrian Academy of Sciences (ÖAW)

Data based construction of kernels for semi-supervised learning with less labels

H. Mhaskar, S. Pereverzyev, V.Y. Semenov, E.V. Semenova

RICAM-Report 2018-25

Data based construction of kernels for semi-supervised learning with less labels

Hrushikesh Mhaskar[*], Sergei V. Pereverzyev[†], Vasyl Yu. Semenov[‡] and Evgeniya V. Semenova[§]

[*]Institute of Mathematical Sciences, Claremont Graduate University, Claremont, CA91711, USA

[†]Johann Radon Institute for Computational and Applied Mathematics, Linz, Austria

[‡]R&D department, Scientific and Production Enterprise "Delta SPE", Kiev, Ukraine

[§]Institute of Mathematics of NASU, Kiev, Ukraine

August 24, 2018

Abstract

This paper deals with the problem of semi-supervised learning using a small number of training samples. Traditional kernel based methods utilize either a fixed kernel or a combination of judiciously chosen kernels from a fixed dictionary. In contrast, we construct a data-dependent kernel utilizing the Mercer components of different kernels constructed using ideas from diffusion geometry, and use a regularization technique with this kernel with adaptively chosen parameters. Our algorithm is illustrated using well-known data sets as well as a data set for automatic gender identification. For some of these data sets, we obtain a zero test error using only a minimal number of training samples.

Keywords: machine learning, semi-supervised learning, Reproducing Kernel Hilbert Spaces, Tikhonov regularization, Laplace-Beltrami operator, gender identification, line spectral frequencies

1 Introduction

The problem of learning from labeled and unlabeled data (semi-supervised learning) has attracted considerable attention in recent years. A variety of machine learning algorithms use Tikhonov single penalty or multiple penalty schemes for regularizing with different approaches to data analysis. Many of these are kernel based algorithms that provide regularization in Reproducing Kernel Hilbert Spaces (RKHS). The problem of finding a suitable kernel for learning a real-valued function by regularization is considered, in particular, in the papers [18], [19] (see also the references therein), where different approaches were proposed. All the methods mentioned in these papers deal with some set of kernels that appear as a result of parametrization of classical kernels or linear combination of some functions. Such approaches lead to the problem of multiple kernel learning. In this way, the kernel choice problem is somehow shifted to the problem of a description of a set (a dictionary of kernels), on which a multiple kernel learning is performed.

In the present paper we propose an approach to construct a kernel directly from observed data rather than choosing one from a given kernel dictionary in advance. The approach uses ideas from diffusion geometry (see, e.g. [1, 2, 3, 5, 12]), where the eigenvectors of the graph

Laplacian associated to the unlabeled data are used to mimic the geometry of underlying manifold that is usually unknown. The literature on this subject is too large to be cited extensively. The special issue [7] of Applied and Computational Harmonic Analysis is devoted to an early review of this subject. Most relevant to the current paper are the papers [5], [6], where the graph Laplacian associated to the data has been used to form additional penalty terms in a multi-parameter regularization functional of Tikhonov type. In contrast to [5], [6], we use eigenvectors and eigenfunctions of the corresponding family of graph Laplacians (rather than a combination of these graph Laplacians) to construct a data-dependent kernel that directly generates an RKHS.

The paper is organized as follows: in the next two sections we present the main theoretical background. Then, we give the numerical algorithms for the implementation of the proposed method. Finally, we provide experimental results with their discussion.

2 Background

The subject of diffusion geometry seeks to understand the geometry of the data $\{x_i\} \subset \mathbb{R}^D$ drawn randomly from an unknown probability distribution μ, where D is typically a large ambient dimension. It is assumed that the support of μ is a smooth sub-manifold of $\mathbb{R}^D$ having a small manifold dimension d. It is shown in [11] that a local coordinate chart of the manifold can be described in terms of the values of the heat kernel, respectively, those of some of the eigenfunctions, of the so called Laplace-Beltrami operator on the unknown manifold. However, since the manifold is unknown, one needs to approximate the Laplace-Beltrami operator. One way to do this is using a graph Laplacian as follows.

For $\epsilon > 0$, $x, y \in \mathbb{R}^D$, let

$$W^\varepsilon(x,y) := \exp\left(-\frac{\|x-y\|^2}{4\varepsilon}\right). \tag{1}$$

We consider the points $\{x_i\}_{i=1}^n$ as vertices of an undirected graph with the edge weight between x_i and x_j given by $W^\varepsilon(x_i, x_j)$, thereby defining a weighted adjacancy matrix, denoted by $\mathbf{W}^\varepsilon$. We define $\mathbf{D}^\varepsilon$ to be the diagonal matrix with the i-th entry on the diagonal given by $\sum_{j=1}^n W^\varepsilon(x_i, x_j)$. The graph Laplacian is defined by the matrix

$$\mathbf{L}^\varepsilon = \frac{1}{n}\left\{\mathbf{D}^\varepsilon - \mathbf{W}^\varepsilon\right\}. \tag{2}$$

We note that for any real numbers $a_1, \cdots, a_n$,

$$\sum_{i,j=1}^n a_i a_j L^\varepsilon_{i,j} = \frac{1}{2n}\sum_{i,j=1}^n W^\varepsilon_{i,j}(a_i - a_j)^2.$$

We conclude that the eigenvalues of $\mathbf{L}^\varepsilon$ are all real and non-negative, and therefore, can be ordered as

$$0 = \lambda^\varepsilon_0 < \lambda^\varepsilon_1 \le \ldots \le \lambda^\varepsilon_{n-1}. \tag{3}$$

It is convenient to consider the eigenvector corresponding to λ^ε_k to be a function on $\{x_j\}_{j=1}^n$ rather than a vector in $\mathbb{R}^n$, and denote it by ϕ^ε_k, thus,

$$\lambda^\varepsilon_k \phi^\varepsilon_k(x_i) = \sum_{j=1}^n L^\varepsilon_{i,j}\phi^\varepsilon_k(x_j) = \frac{1}{n}\left(\phi^\varepsilon_k(x_i)\sum_{j=1}^n W^\varepsilon(x_i,x_j) - \sum_{j=1}^n W^\varepsilon(x_i,x_j)\phi^\varepsilon_k(x_j)\right), \quad i = 1, \cdots, n. \tag{4}$$

Since the function W^ε is defined on the entire ambient space, one can extend the function ϕ_k^ε to the entire ambient space using (4) in an obvious way (the Nyström extension). Denoting this extended function by Φ_k^ε, we have

$$\lambda_k^\varepsilon \Phi_k^\varepsilon(x) = \frac{1}{n}\left(\Phi_k^\varepsilon(x)\sum_{j=1}^n W^\varepsilon(x,x_j) - \sum_{j=1}^n W^\varepsilon(x,x_j)\phi_k^\varepsilon(x_j)\right), \qquad x \in \mathbb{R}^D. \tag{5}$$

More explicitly, (cf. [23])

$$\Phi_k^\varepsilon(x) = \frac{\sum_{j=1}^n W^\varepsilon(x,x_j)\phi_k^\varepsilon(x_j)}{\sum_{j=1}^n W^\varepsilon(x,x_j) - n\lambda_k^\varepsilon}, \tag{6}$$

for all $x \in \mathbb{R}^D$ for which the denominator is not equal to 0. The condition that the denominator of (6) is not equal to 0 for any x can be verified easily for any given ε. The violation of this condition for a particular k can be seen as a sign that for given amount of data n the approximations of the eigenvalue λ_k of the corresponding Laplace-Beltrami operator by eigenvalues λ_k^ε cannot be guaranteed with a reasonable accuracy.

We end this section with a theorem ([4, Theorem 2.1]) regarding the convergence of the extended eigenfunctions Φ_k^ε, restricted to a smooth manifold X to the actual eigenfunctions of the Laplace-Beltrami operator on X. We note that each Φ_k^ε is constructed from a randomly chosen data $\{x_i\}_{i=1}^n$ from some unknown manifold X, and is therefore, itself a random variable.

Theorem 1. *Let X be a smooth, compact manifold with dimension d, and μ be the Riemannian volume measure on X, normalized to be a probability measure. Let $\{x_i\}_{i=1}^n$ be chosen randomly from μ, Φ_k^ε be as in (6), and Φ_k be the eigenfunction of the Laplace-Beltrami operator on X that has the same ordering number as k, corresponding to the eigenvalue λ_k. Then there exists a sequence $\varepsilon_n \to 0$, such that*

$$\lim_{n\to\infty} \frac{1}{\varepsilon^{1+d/2}} |\lambda_k^{\varepsilon_n} - \lambda_k| = 0, \tag{7}$$

and

$$\lim_{n\to\infty} \|\Phi_k^{\varepsilon_n} - \Phi_k\| = 0, \tag{8}$$

where the norm is the L^2 norm, and the limits are taken in probability generated by μ.

3 Numerical algorithms for semi-supervised learning

The approximation theory utilizing the eigen-decomposition of the Laplace-Beltrami operator is well developed, even in greater generality than this setting, in [15, 10, 17, 16, 9]. In practice, the correct choice of ε in the approximate construction of these eigenvalues and eigenfunctions is a delicate matter that affects greatly the performance of the kernel based methods based on these quantities. Some heuristic rules for choosing ε have been proposed in [12, 8]. These rules are not applicable universally; they need to be chosen according to the data set and the application in consideration.

In contrast to the traditional literature, where a fixed value of ε is used for all the eigenvalues and eigenfunctions, we propose in this paper the construction of kernel of the form

$$K_n(x,t) = \sum_{k=1}^{n-1} (n\lambda_k^{\varepsilon_k})^{-1} \Phi_k^{\varepsilon_k}(x)\Phi_k^{\varepsilon_k}(t); \tag{9}$$

i.e., we select the eigenvalues and the corresponding eigenfunctions from different kernels of the form W^ε to construct our kernel. We note again that in contrast to the traditional method of

combining different kernels from a fixed dictionary, we are constructing a single kernel using the Mercer components of different kernels from a dictionary.

Our rule for selecting ε_k's is based on the well-known quasi-optimality criterion [22] that is one of the simplest and the oldest but still quite efficient strategy for choosing a regularization parameter. According to that strategy, one selects a suitable value of ε (regularization parameter) from a sequence of admissible values $\{\varepsilon_j\}$, which usually form a geometric sequence, i.e. $\varepsilon_j = q^j, j = 1, 2, \ldots, M$; $q < 1$. We propose to employ the quasi-optimality criterion in the context of the approximation of the eigenvalues of the Laplace-Beltrami operator. Then by analogy to [22] for each particular k we calculate the sequence of approximate eigenvalues $\lambda_k^{\varepsilon_j}, j = 1, 2, \ldots, M$, and select $\varepsilon_k \in \{\varepsilon_j\}$ such that the differences $|\lambda_k^{\varepsilon_j} - \lambda_k^{\varepsilon_{j-1}}|$ attain their minimal value at $j = k$.

The Algorithm 1 below describes the combination of the approximation (6) with quasi-optimality criterion.

Algorithm 1 Algorithm to generate reproducing kernel from data

Given data $\{x_i\}_{i=1}^n \subset X$, $\quad \{x_i, y_i\}_{i=1}^m$ are the labeled examples.
Introduce the grid for parameter ε: $\varepsilon_j = q^j, j = 1..M$.
for $(j = 1 : M)$ **do**

Compute L^{ε_j} as in (2), and eigensystem $(\Phi_k^{\varepsilon_j}, \lambda_k^{\varepsilon_j})$ $k = 1..n-1$
end for
for $(k = 1 : n-1)$ **do**

Find $\varepsilon_k = \arg\min_{\varepsilon_j} |\lambda_k^{\varepsilon_j} - \lambda_k^{\varepsilon_{j-1}}|$.
$\lambda_k := \lambda_k^{\varepsilon_k}, \Phi_k = \Phi_k^{\varepsilon_k}$
end for
Compute

$$\Phi_k^{\varepsilon_k}(x) = \frac{\sum_{j=1}^n W^{\varepsilon_k}(x, x_j)\phi_k(x_j)}{\sum_{j=1}^n W^{\varepsilon_k}(x, x_j) - n\lambda_k}$$

Form kernel function

$$K_n(x, t) = \sum_{k=1}^{n-1} \frac{1}{n\lambda_k^{\varepsilon_k}} \Phi_k^{\varepsilon_k}(x)\Phi_k^{\varepsilon_k}(t) \tag{10}$$

Algorithm 2 uses the constructed kernel (10) in kernel ridge regression from labeled data. The regression is performed in combination with discrepancy based principle for choosing the regularization parameter α.

4 Experimental results

4.1 Two moons dataset

We start the experimental section with two moons dataset. The software and data were borrowed from https://github.com/jaejun-yoo/shallow-DANN-two-moon-dataset.

For two moons dataset we take $\{x_i\}_{i=1}^n$ with $n = 200, 150, 100, 80$ and subsets $\{x_i\}_{i=1}^m \subset \{x_i\}_{i=1}^n$ with $m = 2, 4, 6$ labeled points. The goal of semi-supervised data classification problems is to assign correct labels for remaining points $\{x_i\}_{i=1}^n \setminus \{x_i\}_{i=1}^m$. For every dataset (defined by pair (n, m)) we performed 50 trials with randomly chosen labeled examples.

As follows from experiments, the accuracy of the classification is improving with the growth of the number of unlabeled points. In particular, for $n \geq 150$, to label all points without error, it is

Algorithm 2 Algorithm for kernel ridge regression with the constructed kernel (10)

Given data $\{x_i\}_{i=1}^n \in X, \quad \{x_i, y_i\}_{i=1}^m$ are the labeled examples; $y = \{y\}_{i=1}^m$.
Form kernel using Algorithm 1
Introduce the grid for parameter α: $\alpha_k = p^k, k = 1, 2, \ldots, N$
Calculate Gram matrix $\widehat{K}_m$ consisting of the sub-matrix $\{K_n(x_i, x_j)\}_{i,j=1}^m$ (10) in labeled points
for k=1:N **do**
Calculate C_{α_k} as

$$C_{\alpha_k} = (\alpha_k I + \widehat{K}_m)^{-1} y,$$

Find the α_{min} such that $\|\widehat{K}_m C_{\alpha_k} - y\|$ is minimized.
end for
The decision-making function is

$$f_n^*(x) = \sum_{i=1}^{m} (C_{\alpha_{min}})_i K_n(x, x_i).$$

Table 1: Results of testing for two moons dataset

n:	m	average error of 50 trials
200	2	0
150	2	0
100	2	21.24%
100	4	9.08%
100	6	4.68%
80	2	34.9%
80	4	17.75%
80	6	10.35%

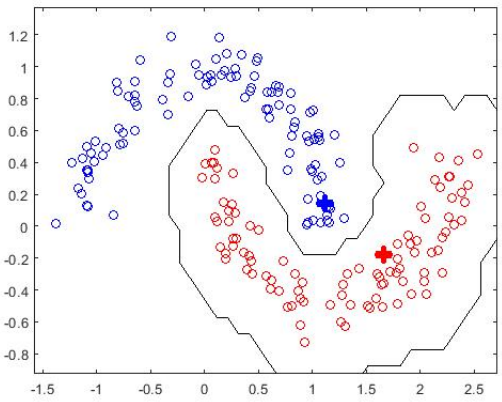

Figure 1: Classification of "two moons" dataset with extrapolation region. The values of parameters are $n = 200$, $m = 2, M = 70, p = 0.5, q$=0.9.

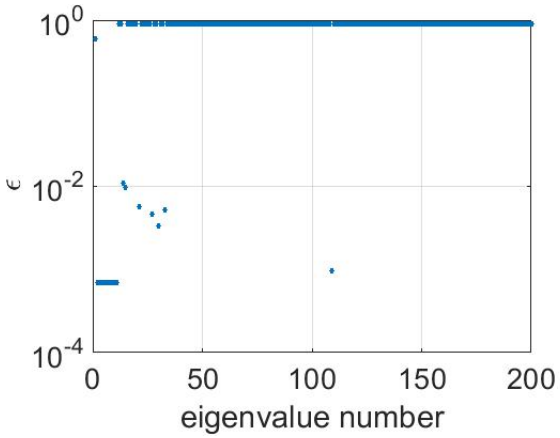

Figure 2: Logarithmic plot of adaptively chosen ε for two-moon dataset. The values of parameters are $n = 200, m = 2$, $M = 70, p = 0.5, q$=0.9.

enough to take only one labeled point for each of two classes ($m = 2$). At the same time, if the set of unlabeled points is not big enough, then for increasing the accuracy of prediction we should take more labeled points. The result of the experiment for the two moons dataset with $m = 2$ is shown at Figure 1. The big crosses correspond to the labeled data and other points are colored according to the constructed predictors. The logarithmic dependance of ε_k on the eigenvalue number k is shown in Figure 2. The parameters for both Figure 1 and Figure 2 were m=2, M=70, q=0.9.

More details about the proposed strategy are given in Figures 3, 4, where we show the plots $\lambda = \lambda_i^{\varepsilon_k}$, $\Delta\lambda = |\lambda_i^{\varepsilon_j} - \lambda_i^{\varepsilon_{j-1}}|$, for $n = 200, i = 2, 20$, $j = 2, 3, \ldots 70$.

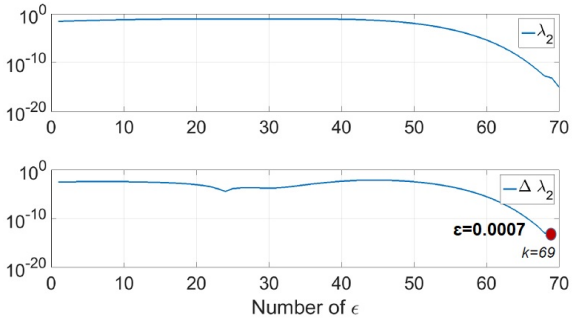

Figure 3: $\Delta\lambda_2 = |\lambda_2^{\varepsilon_j} - \lambda_2^{\varepsilon_{j-1}}|$; Algorithm 1 choses $\lambda_2 = \lambda_2^{\varepsilon_k}, k = 69, \varepsilon_k = 0.0007$.

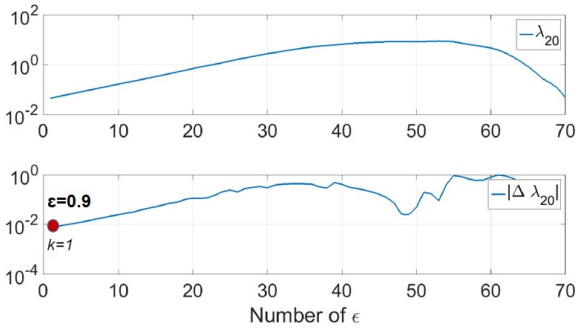

Figure 4: $\Delta\lambda_{20} = |\lambda_{20}^{\varepsilon_j} - \lambda_{20}^{\varepsilon_{j-1}}|$; Algorithm 1 chooses $\lambda_{20} = \lambda_{20}^{\varepsilon_k}, k = 1, \varepsilon_k = 0.9$.

Note that "two moons" dataset from Figure 1 has been also used for testing the performance of a manifold learning algorithm realized as a multi-parameter regularization [14]. The comparison of Table 1 with Table 6 of [14] shows that on the dataset from Figure 1 the Algorithms 1-2 based on the kernel (10) outperform the algorithm from [14], where the graph Laplacian has been used as the second penalization operator.

In our next experiment, we follow [6] and embed the two moons dataset in $\mathbb{R}^{100}$ by adding 98-dimensional zero-mean Gaussian random vectors with standard deviation σ. Then the Algorithms

1-2 have been applied to the transformed data set, which means that in (1) the symbol $\|\cdot\|$ is staying for $\mathbb{R}^{100}$-norm. The results of the experiment with only two labeled points, $m = 2$, are presented in Table 2. The performance displayed in this table is comparable to the one reported in [6], but the above performance has been achieved with minimal admissible number of labeled points, i.e. $m = 2$.

Table 2: Results of testing for two moons dataset embedded in $\mathbb{R}^{100}$, $n = 200$, $m = 2$.

σ	average error of 10 trials
0	0
10^{-4}	0.15%
10^{-3}	0.05%
10^{-2}	0.25%
5×10^{-2}	45.5%
10^{-1}	48.5%

4.2 Two circles datasets

In this section, we consider two "two-circles" datasets. Each circle has unit radius and contains 100 points so that $n = 200$. These datasets are depicted at Figures 5 and 6 respectively.

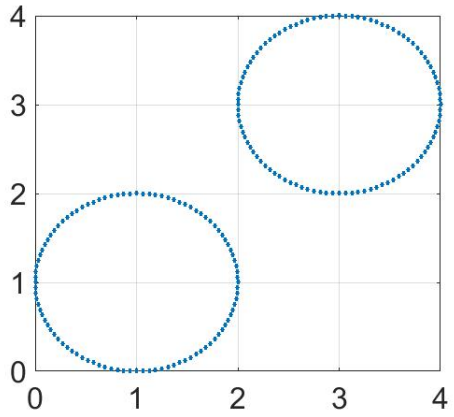

Figure 5: Testing dataset of two non-intersecting circles.

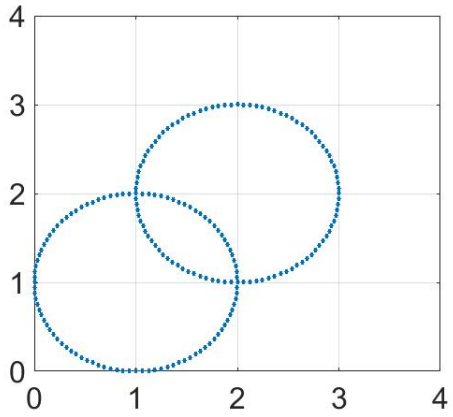

Figure 6: Testing dataset of two intersecting circles.

Below we consider classification of these datasets by the proposed method. As can be seen from Figure 7, for non-intersecting circles only $m = 2$ labeled points are enough for a correct classification of the given set. The log-plot of ε_k suggested by Algorithm 1 is shown in Figure 8.

Figure 9 shows the classification results for $m = 20$ labeled points at the intersected circles. The logarithmic plot for ε_k is shown in Figure 10. The big crosses correspond to the labeled data and other points are colored according to the constructed predictors. The error percentage for different m is shown in Table 2. It can be seen that the classification error decreases with the growth of the number m of labeled points. The fact that not all points are correctly classified can be explained by the non-smoothness of the considered manifold.

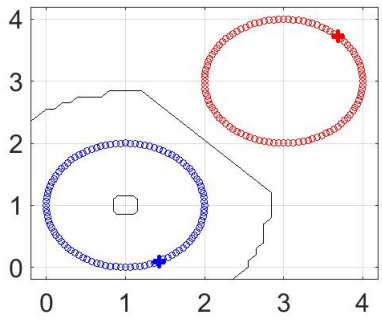

Figure 7: Classification of "two non-intersecting circles" dataset. The values of parameters are $n = 200, m = 2$, $M = 70, p = 0.5, q = 0.9$.

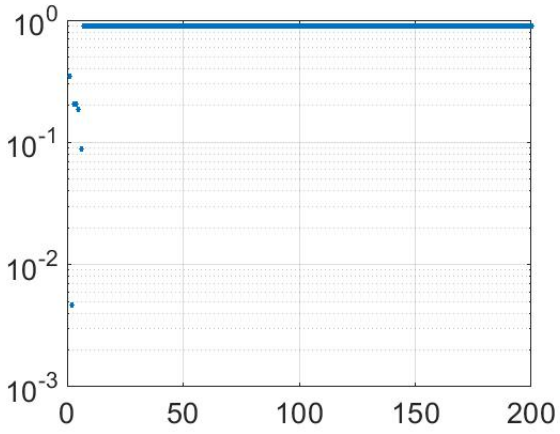

Figure 8: Logarithmic plot of adaptively chosen ε for "two non-intersecting circles" dataset. The values of parameters are $n = 200, m = 2, M = 70, p = 0.5, q = 0.9$.

Table 3: Results of testing for "two intersecting circles" dataset

m	average error of 50 trials
2	49.5%
4	26.5%
8	21.24%
10	16.58%
12	15.5%

4.3 Multiple classification. Three moons dataset

Three moons dataset has been simulated from three half-circles by adding a gaussian noise with mean zero and deviation 0.14. Each half-circle contains 100 points so that $n = 300$. Figure 11 shows the classification results for $m = 3$ labeled points with other parameters $M = 60; q = 0.9$. The logarithmic plot of ε_k suggested by Algorithm 1 is shown at Figure 12. The big crosses correspond to the labeled data and other points are colored according to the constructed predictors. It can be seen that for $m = 3$ (just one labeled point per circle) the classification is performed correctly.

4.4 Automatic gender identification

We also investigate the application of the proposed classification approach to the problem of automatic gender identification [21]. Having the audio recording of some speaker, the task is to determine the speaker's gender: male or female.

The gender classification task is usually performed on frame-by-frame basis as follows. The speech signal is divided onto the segments (frames) of 20 ms (160 samples for sampling frequency 8000 Hz). For every such frame a set of voice parameters is calculated. It is necessary to use such parameters that provide distinct description of male and female voices. We used two-dimensional ($d = 2$) parameters vector consisting of pitch period T_0 [21] and difference of the first two line

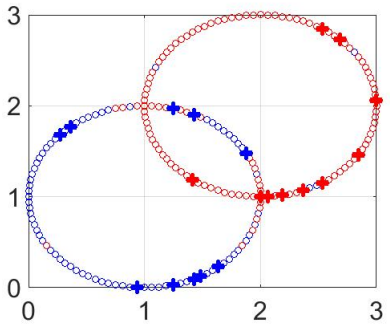

Figure 9: Classification of "two intersecting circles" dataset. The values of parameters are $n = 200, m = 20, M = 70, p = 0.5, q = 0.9$.

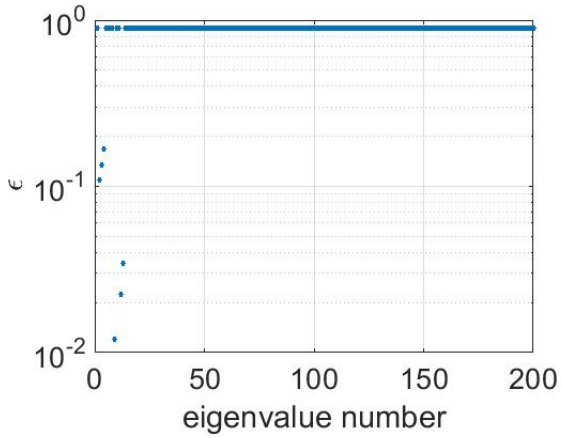

Figure 10: Logarithmic plot of adaptively chosen ε for "two intersecting circles" dataset. The values of parameters are $n = 200, m = 20, M = 70, p = 0.5, q = 0.9$.

spectral frequencies (LSF) $d = \omega_2 - \omega_1($ for the definition, properties and computation of line spectral frequencies, see e.g. [20]).

For the training we used 240 audio recordings with total duration of 14 minutes. Male and female speakers of English and Russian languages were participating. The total number of the considered parameter vectors was 8084 for male speakers and 8436 for female speakers. To make the problem computationally tractable, we selected 128 "typical" parameter vectors both for male and female parts which were determined by k-means algorithm [13]. In experiments $l = 2$ points both for "male" and "female" manifolds were labeled. Parameter ε was selected according to the proposed adaptive strategy. The value 1 of decision-making function was assigned to male speakers and the value -1 was assigned to female speakers.

The distribution of test parameters vectors and the results of their classification is shown in Figures 13 and 14 respectively.

Then the independent testing was performed on a set of 257 audio recordings including English, German, Hindi, Hungarian, Japanese, Russian and Spanish speakers (all of these speakers did not take part in the training database). The decision male/female for an audio recording was made by majority of the decisions among all its frames. As the result of this independent testing, the classification errors for male and female speakers were 12.6% and 6.5% respectively.

The examples of the decisions for frames of a audio recording are shown in Figures 15 and 16 for male and female speakers respectively. Each record was divided onto frames of 20 msec. On every frame the vector of features was calculated and then classified by the proposed approach. It can be seen that in the case of male speaker the most of the decision-making function values are grouped near value "1". It provides correct classification of this record as "male". Similarly, most of the decision-making function values for a female recording are grouped around are grouped around value "-1".

The obtained results are promising and encourage to test the proposed approach on a larger variety of signals.

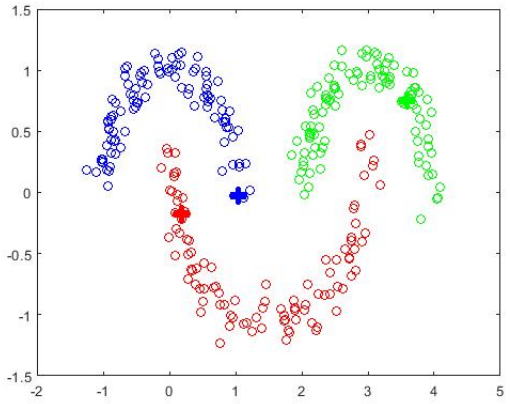

Figure 11: Classification of "three moons" dataset. The values of parameters are $n = 300, m = 3, M = 60, p = 0.5, q$=0.9.

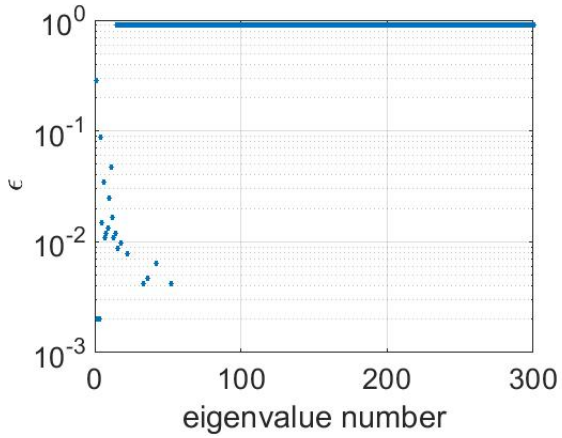

Figure 12: Logarithmic plot of adaptively chosen ε for three-moon dataset. The values of parameters are $n = 300$, $m = 3, M = 60, p = 0.5, q$=0.9.

Acknowledgments

Sergei V. Pereverzyev and Evgeniya Semenova gratefully acknowledge the support of the consortium AMMODIT funded within EU H2020-MSCA-RICE. The research of Hrushikesh Mhaskar is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2018-18032000002.

References

[1] M. Belkin, I. Matveeva, and P. Niyogi. Regularization and semi-supervised learning on large graphs. In *Learning theory*, pages 624–638. Springer, 2004.

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

[3] M. Belkin and P. Niyogi. Semi-supervised learning on Riemannian manifolds. *Machine learning*, 56(1-3):209–239, 2004.

[4] Mikhail Belkin and Partha Niyogi. Convergence of laplacian eigenmaps. *Adv. neur. inform. process. 19*, 129, 2007.

[5] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, 7:2399–2434, 2006.

[6] Andrea L. Bertozzi, Xiyang Luo, Andrew M. Stuart, and Konstantinos C. Zygalakis. Uncertainty quantification in the classification of high dimensional data. *CoRR*, abs/1703.08816, 2017.

[7] C. K. Chui and D. L. Donoho. Special issue: Diffusion maps and wavelets. *Appl. and Comput. Harm. Anal.*, 21(1), 2006.

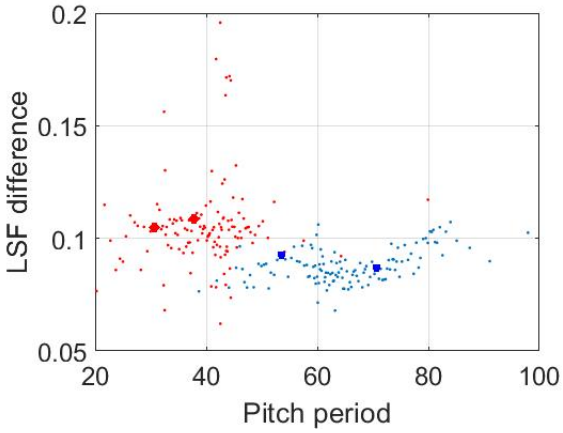

Figure 13: The distribution of 128 considered vectors for male (blue color) and female (red color) speakers. The four labeled points are marked.

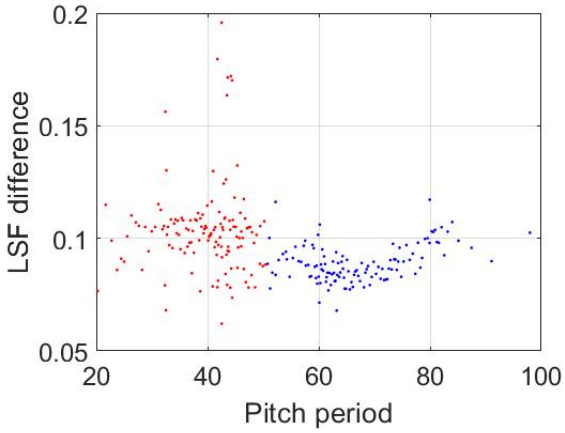

Figure 14: The results of the vectors' classification. Blue and red colors correspond to male and female speakers respectively. The values of parameters are $n = 128, m = 4, M = 30, p = 0.5, q$=0.9.

[8] R. R. Coifman and M. J. Hirn. Diffusion maps for changing data. *Applied and Computational Harmonic Analysis*, 36(1):79–107, 2014.

[9] M. Ehler, F. Filbir, and H. N. Mhaskar. Locally learning biomedical data using diffusion frames. *Journal of Computational Biology*, 19(11):1251–1264, 2012.

[10] F. Filbir and H. N. Mhaskar. Marcinkiewicz–Zygmund measures on manifolds. *Journal of Complexity*, 27(6):568–596, 2011.

[11] P. W. Jones, M. Maggioni, and R. Schul. Universal local parametrizations via heat kernels and eigenfunctions of the Laplacian. *Ann. Acad. Sci. Fenn. Math.*, 35:131–174, 2010.

[12] S. S. Lafon. *Diffusion maps and geometric harmonics*. PhD thesis, Yale University, Yale, 2004.

[13] Yoseph Linde, Andres Buzo, and Robert Gray. An algorithm for vector quantizer design. *IEEE Trans. Comm.*, 28(1):84–95, 1980.

[14] Shuai Lu and Sergei V. Pereverzyev. Multi-parameter regularization and its numerical realization. *Numerische Mathematik*, 118(1):1–31, May 2011.

[15] M. Maggioni and H. N. Mhaskar. Diffusion polynomial frames on metric measure spaces. *Applied and Computational Harmonic Analysis*, 24(3):329–353, 2008.

[16] H. N. Mhaskar. A generalized diffusion frame for parsimonious representation of functions on data defined manifolds. *Neural Networks*, 24(4):345–359, 2011.

[17] Hrushikesh N. Mhaskar. Eignets for function approximation on manifolds. *Appl. Comput. Harm. Anal.*, 29(1):63–87, 2010.

[18] Charles A. Micchelli and Massimiliano Pontil. Learning the kernel function via regularization. *J.Mach.Learn.Res.*, 6:10127–10134, 2005.

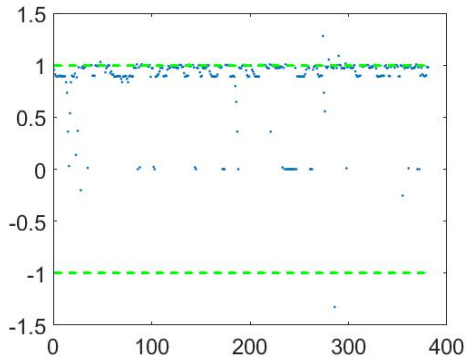

Figure 15: The example of decisions for frames of an audio recording of a male speaker.

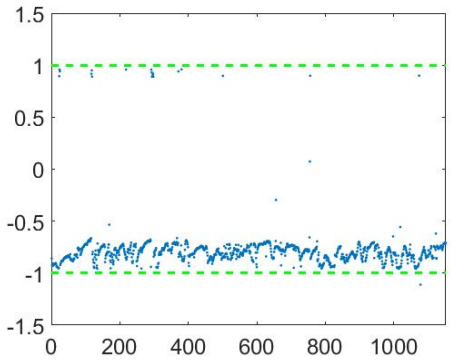

Figure 16: The example of decisions for frames of an audio recording of a female speaker.

[19] Sergei V. Pereverzyev and Pavlo Tkachenko. Regularization by the linear functional strategy with multiple kernels. *Frontiers in Applied Mathematics and Statistics*, 3:1, 2017.

[20] Vasyl Semenov. A novel approach to calculation of line spectral frequencies based on inter-frame ordering property. In *Proc. IEEE ICASSP*, pages 1072–1075, 2006.

[21] Vasyl Semenov. Automatic determination of speaker's gender based on gaussian mixtures. In *Proceedings of Acoustical Conference "Consonans"*, pages 189–194, 2009.

[22] A.N. Tikhonov and V.B. Glasko. Use of regularization method in non-linear problems. *Zh. Vychisl. Mat. Mat. Fiz.*, 5:463–473, 1965.

[23] Ulrike von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *Ann. Statist.*, 36(2):555–586, 2008.