Reliable Representation of Data on Manifolds

Jun Li, Pengwei Hao
Queen Mary, University of London, Mile End, London E1 4NS
{junjy, phao}@dcs.qmul.ac.uk

Abstract

The manifold learning algorithms are promising data analysis tools. However, to fit an unseen point in a learned model, the point must be located in the training set, which limits its scalability. In this paper, we discuss how to select landmarks from the data to help locate the test points. Our method is for data on manifolds: the way the landmarks represent the data in the ambient space should resemble the way they represent the data on the manifold. Compared to the previous research, (i) Our test foregoes the requirement of knowing the intrinsic manifold dimension and thus is more applicable and robust. (ii) Our selection implies a provable topology preservation property. (iii) We also provide a way to improve existing landmarks. Experiments on the synthetic data and the real data have been done. The results support the proposed properties and algorithms.

1 Introduction

Manifold learning refers to a family of algorithms that analyze the variables non-linearly underlying the distribution of the points in the high dimensional data space [8]. However, unlike the classical linear methods, e.g., PCA, these methods learn the embedding, but not the explicit mapping between the manifold coordinates and the data points. Thus they are not readily generalized to unseen points [4], and rerun the algorithm for each new sample is expensive. Based on the established connection between manifold learning and kernel density estimation [2], we can estimate the low dimensional coordinates for a new point with the help of the learned embedding [4]. To apply these estimating schemes, the near neighbours of the input point on the manifold need to be found. In other words, it is to be "located" on the manifold. This entails a cost proportional to the volume of the training data.

To answer a nearest neighbour query quickly, it is natural to use part of the samples as *landmarks* to lead the search. If they are chosen appropriately, a sample may be safely located to the proper area on the manifold, by being compared with only those landmarks. The proper area will contain the position at which the sample should have been located, were the query conducted in the full set of the data.

When we compare a test point to the landmarks, only the Euclidean distance can be easily computed. Thus if we find the nearest landmark to the test point in the ambient space, we want they are near on the manifold as well. Let the geodesic distance from a point to its nearest landmark on the manifold be d and the *geodesic* distance to the nearest landmark in the *ambient* space be d'. It is straightforward to see that $d' \geq d$. However, it is worth asking:

Question 1 *Is d' significantly larger than d?*

In this paper, we analyze Question 1 and answer it with a test. Compare to the previous work, our test is more applicable and robust, because it does not require to know the

intrinsic dimension of the manifold. We also prove its validity in terms of preserving the topology of the neighbourhood graph of the data point cloud. When the landmark set cannot represent the manifold perfectly, besides the obvious solution of choosing more points as landmarks, we have also developed an optimization-based algorithm to adjust the existing landmarks. The optimization procedure is less computationally demanding than performing the test at each time of adding a new landmark.

In Section 2, we give a brief review of the related background. In Section 3, we first present our analysis of the problem, and then propose our condition for the test, as well as the properties justifying our condition. In Section 4, the optimization algorithm is presented. In Section 5, we report experimental results supporting our test and the optimization method. Finally, we conclude our paper and discuss the possible future directions in Section 6.

2 Related Work

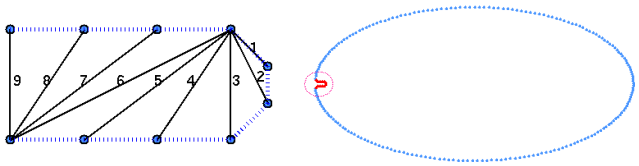
In the past few years, many manifold learning approaches emerged [8]. These methods find the embedding of the data but not the mapping. Methods of generalizing the learned embedding beyond the training samples have been developed. Some of them approximate the embedding linearly with a linear map [7; 15]. To generalize the learned embedding nonlinearly, techniques of kernel density estimation (KDE) have been exploited to extend the domain of the coordinating function to the whole data space [9; 4].

Properly set up landmarks in the data may help these kernel extension by speeding up the nearest neighbour searching. In a recent work [11], the authors proposed a criterion whether a set of landmarks preserves the topology of a manifold, given the intrinsic dimension of the manifold. However, the intrinsic dimension may not be easy to estimate for data manifolds in practical applications. Landmarks are also used for learning the embedding for the training data [12; 6; 14]. These randomly selected or error minimizing landmarks improve computational efficiency of the *training*, but they do not necessarily respect the topology of the manifold for locating *testing* points. Keeping the topology of a manifold with only a subset of the samples has been discussed in terms of surface reconstruction or mesh simplification in computer graphics [1]. These algorithms deal with the special case of 2D manifold in 3D ambient space.

In terms of nearest neighbour searching, our work is also related to the *spatial accessing methods (SAM)* [13](and references there in). Compared to these general methods, we pay special attention to the topology of the data manifold. Thus the found landmarks may be used for information propagation or exploring active learning in the data.

3 Topological Safety Test

Let us first introduce some denotations: We have observed N data points in $\mathbb{R}^D$, $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$. The data are drawn from a d-dimensional manifold $\mathscr{M}$ embedded in $\mathbb{R}^D$, where $d < D$. In our setting, the landmarks are chosen from $\mathbf{X}$. Let their indices be $\mathbf{P} = \{p_1, p_2, \dots\}$. Then we have $\mathbf{x}_{p_k}$ be the k-th landmark. We write it as $\mathbf{p}_k$ when there is no ambiguity, and use $\mathbf{P}$ interchangeably for the set of landmarks and their indices in the data set. $\mathbf{X}$ is organized as the nodes in a neighbourhood graph $\mathbf{G}$ and there is a (weighted) edge between each node and its k (number) or ε (Euclidean distance in $\mathbb{R}^D$) nearest neighbours. We use "$\sim$" to denote the adjacency relation in the graph. Two subsets of nodes are *adjacent* when there are at least one pair of adjacent nodes belonging to each of the



(a) (b)

Figure 1: Which are topology-changing links?

(a) Dotted links: edges in the graph; Solid links: to be judged . (b) The global manifold from which (red circled area) points in (a) are sampled.

subsets respectively. $d_E(\cdot,\cdot)$ stands for the Euclidean distance in $\mathbb{R}^D$. $d_M(\cdot,\cdot)$ measures the geodesic distance $\mathscr{M}$ [1]. Given $\mathbf{x} \in \mathscr{M}$, $L_E^n(\mathbf{x};\mathbf{P})$ denotes its n^{th} nearest landmark w.r.t. $d_E(\cdot,\cdot)$, and $L_M^n(\mathbf{x};\mathbf{P})$ is the one w.r.t. $d_M(\cdot,\cdot)$ on $\mathscr{M}$. In the following, we write L_E^1 as L_E and L_M^1 as L_M, and without writing the landmark set $\mathbf{P}$ explicitly if there is no ambiguity. The inverse maps of $L_{E,M}$ are $\text{Cell}_{E,M}(\mathbf{p}) = \{\mathbf{x}|L_{E,M}(\mathbf{x}) = \mathbf{p}\}$. Cell_E can be considered as the intersection of the manifold and the *Voronoi cell* of a landmark. While Cell_M can be considered similarly, however, with the Voronoi tessellation done on the manifold directly w.r.t. the geodesic distance.

To answer Question 1, we must make clear the criterion for "significantly", i.e., whether a connection between an $\mathbf{x}$ and $L_E(\mathbf{x})$ is a short-circuit on the manifold. In [11], the criterion is heuristic and depends on the intrinsic dimension of the manifold, which is generally unavailable in practice.

We argue that whether a link on a manifold causes a short-circuit depends on the context: the manifold geometry, the point cloud sampling, the neighbourhood graph and the distribution of the other landmarks. We will show this by an example. Figure 1(a) shows some points from a U-shaped manifold, as well as the dotted links between the neighbouring points on the manifold and several numbered solid links. We are to judge if they are short-circuits. Link 1 should not be considered as a "short-circuit", otherwise the sampling is inappropriate anyway. From link 2 to 9, the pairs of linked points are farther and farther away on the manifold. Which of them should be taken as a "short-circuit"? This question should be answered with care. Even for link 9, which collapses the whole U-shaped structure and seems to be a definite "short-circuit", if one concerns a larger scale, the link might become acceptable. E.g., in (b), we show a case where the U-shaped manifold is actually an unimportant part of a larger structure.

Therefore, we propose

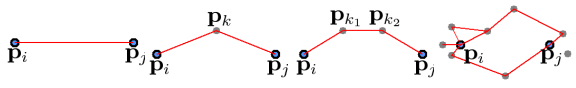
Condition 1 $\forall \mathbf{p}_1, \mathbf{p}_2 \in \mathbf{P}$, $\text{Cell}_E(\mathbf{p}_1) \cap \text{Cell}_M(\mathbf{p}_2) \neq \emptyset$ *implies* $\mathbf{p}_1 = \mathbf{p}_2$ *or* $\text{Cell}_M(\mathbf{p}_1) \sim \text{Cell}_M(\mathbf{p}_2)$.

The condition means that there are two cases for a point $\mathbf{x}$ to be considered as near enough to $L_E(\mathbf{x})$ on the manifold: (i) $L_E(\mathbf{x}) = L_M(\mathbf{x})$; and (ii) otherwise, $\mathbf{p}_1 = L_E(\mathbf{x})$, $\mathbf{p}_2 = L_M(\mathbf{x})$ and $\mathbf{p}_1$ and $\mathbf{p}_2$ are distinct. Then their geodesic Voronoi cells $\text{Cell}_M(\mathbf{p}_1)$ and $\text{Cell}_M(\mathbf{p}_2)$ must be adjacent on the manifold. We call the points at which Condition 1 does not hold *topological error points* (TEPs). In the following, we will show how this condition implements the concept that "the landmarks represent the data properly".

For a neighbourhood graph, Condition 1 ensures that the partition Cell_E is *topologically similar* to Cell_M. Let us first make clear the meaning of "topologically similar".

Given a partition of the neighbourhood graph $L : \mathbf{X} \to \mathbf{P}$, a *minor* $\mathbf{H}$ of $\mathbf{G}$ can be defined by the following contraction:

[1] In practice, this is approximated by the shortest path in $\mathbf{G}$.



(a) (b) (c) (d)

Figure 2: Graph similarity

1. $\mathbf{G}^{(0)} \leftarrow \mathbf{G}$.
2. Find an edge $\mathbf{x} \sim \mathbf{y}$ in $\mathbf{G}^{(k)}$, make $\mathbf{G}^{(k+1)}$ as following:
   - if $L(\mathbf{x}) = L(\mathbf{y})$ and $\mathbf{x}, \mathbf{y} \notin \mathbf{P}$, contract $\mathbf{x}$ and $\mathbf{y}$ to $\mathbf{x}$.
   - if $\mathbf{y} \notin \mathbf{P}$ and $L(\mathbf{y}) = \mathbf{x}$, or vice versa, contract them to the landmark.
3. Repeat Step 2 until no more edges are contractable.

The contraction results in a minor $\mathbf{H}$ of $\mathbf{G}$ consisting of all the landmarks. Let $\mathbf{H}_M = \text{Contract}(\mathbf{G}; L_M)$ and $\mathbf{H}_E = \text{Contract}(\mathbf{G}; L_E)$. Then it can be proved that

Proposition 1 *Condition 1 implies that $\mathbf{H}_E$ is similar to $\mathbf{H}_M$, where "similar" means:*

1. *If there is an edge between two landmarks $\mathbf{p}_i$ and $\mathbf{p}_j$ in $\mathbf{H}_E$, $\mathbf{p}_i$ and $\mathbf{p}_j$ are connected in $\mathbf{H}_M$ by a path of the length less than or equal to 3 (Figure 2(a)-(c)).*
2. *If there is an edge between two landmarks $\mathbf{p}_i$ and $\mathbf{p}_j$ in $\mathbf{H}_M$, these is a path in $\mathbf{H}_E$ between $\mathbf{p}_i$ and $\mathbf{p}_j$, and the path contains only landmarks in the neighbourhood of $\mathbf{p}_i$ and $\mathbf{p}_j$ in $\mathbf{H}_M$ (Figure 2(d)).*

Of the two cases, case 1 is more important. It means that for a point $\mathbf{x}$, compared to the optimal representation by $\mathrm{L}_M(\mathbf{x})$, the Euclidean representation $\mathrm{L}_E(\mathbf{x})$ does not take significantly more risk of making a "short-circuit" on the manifold. The proof is provided in Appendix A.

3.1 Implementation of the Test

Testing Condition 1 involves finding the nearest landmarks for each node in $\mathbf{G}$. It takes $\mathscr{O}(kKN \log N)$ [11], where k is the neighbourhood size of $\mathbf{G}$ and $K = |\mathbf{P}|$.

We construct a landmark set by repeatedly selecting TEPs as landmarks until Condition 1 holds everywhere. As in [11], we do not re-perform the test after each adding. For each $\mathbf{x}$, we record $L_E(\mathbf{x})$, $\rho_E(\mathbf{x}) = d_E(L_E(\mathbf{x}), \mathbf{x})$ $L_M(\mathbf{x})$ and $\rho_M(\mathbf{x}) = d_M(L_M(\mathbf{x}), \mathbf{x})$. We also record the adjacency matrix $\mathbf{H}_M$ to track whether two landmarks' cells are adjacent on the manifold. With this information, our test is:

$$L_E(\mathbf{x}) = L_M(\mathbf{x}) \quad \text{OR} \quad \mathbf{H}_M(L_E(\mathbf{x}), L_M(\mathbf{x})) = 1 \tag{1}$$

After a new landmark $\mathbf{p}_N$ is added, $L_E(\mathbf{x})$ and $\rho_E(\mathbf{x})$ can be updated easily in $\mathscr{O}(N)$. To adjust $L_M(\mathbf{x})$ and $\rho_M(\mathbf{x})$ in according to $\mathbf{p}_N$, we use Dijkstra algorithm to compute the shortest path length from $\mathbf{p}_N$ to each $\mathbf{x}$, with stopping condition:

$d_M^N(\mathbf{x}) \leftarrow \infty$ for all $\mathbf{x}$; $d_M^N(\mathbf{p}_N) \leftarrow 0$; $\mathbf{Y} \leftarrow \mathbf{X}$
while $\min(d_M^N(\mathbf{Y})) < \infty$ **do**
  $\mathbf{y} \leftarrow \text{argmin}_{\mathbf{y}' \in \mathbf{Y}} d_M^N(\mathbf{y}')$; $L_M(\mathbf{y}) \leftarrow \mathbf{p}_N$
  $\rho_M(\mathbf{y}) \leftarrow d_M^N(\mathbf{y})$
  Update d_M^N according to edges connected to $\mathbf{y}$
  Remove $\{\mathbf{y}\} \cup \{\mathbf{x} | \rho_M(\mathbf{x}) < d_M^N(\mathbf{y})\}$ from $\mathbf{Y}$
end while

The algorithm can be interpreted as: Start from the new landmark $\mathbf{p}_N$. Compute the shortest paths to the other points using Dijkstra algorithm. Once having computed the shortest path to one point $\mathbf{x}$, add $\mathbf{x}$ to the cell of $\mathbf{p}_N$ by letting $L_M(\mathbf{x}) \leftarrow \mathbf{p}$. As all points that are nearer to $\mathbf{p}_N$ than $\mathbf{x}$ have already been in the cell, we remove all $\{\mathbf{y}|\rho_M(\mathbf{y}) < d_M(\mathbf{x},\mathbf{p}_N)\}$ from the searching set. Keep expanding the "region" of $\mathbf{p}_N$ until the searching set become empty. When manipulating $\mathbf{Y}$ using a Fibonacci heap, the complexity is $\mathscr{O}(nN\log N)$. However, in practice, step 6 generally deletes most nodes in the first few steps. And thus this updating algorithm is faster than that in [11].

4 Landmark Optimization

In the last section, we have discussed how to test the criterion for the landmarks, and make it satisfied by simply adding new landmarks. One natural question may arise is that whether the extent to which the proposed criterion is satisfied (or violated) by a set of landmarks can be estimated quantitatively without performing the test procedure. If we can explicitly estimate the (dis-)satisfaction of condition in terms of the set of landmarks, we will be able to adjust or improve the existing landmark set by optimization quickly. Furthermore, the size of the landmark set can be controlled.

Ideally, we need a function $\mathscr{E} : \mathbf{P} \rightarrow \mathbb{N}$: given the landmarks, $\mathscr{E}$ returns the number of TEPs. Obviously, directly minimizing the $\mathscr{E}$ over $\mathbf{P}$ will lead to a difficult combinatorial programming problem. We can use $\sum d_M^2(\mathbf{x}, L_E(\mathbf{x}))$ as a heuristics of the number of TEPs. It is the *geodesic distances* between the landmarks and the points in the corresponding *Euclidean* Voronoi cell. We argue without proof that this heuristic objective function should favor a landmark distribution that makes the radius of each $\mathrm{Cell}_E(\mathbf{p})$ minimized and thus reduces the TEPs.

However, pre-computing $\forall \mathbf{x}, \mathbf{y}, d_M(\mathbf{x},\mathbf{y})$ requires $\mathscr{O}(N^2 \log N)$ time and $\mathscr{O}(N^2)$ storage, both of which are expensive when the training data are rich. Thus we turn to the estimate of d_M in explicit forms, which allow us to evaluate $d_M(\mathbf{x},\mathbf{p})$ for $\forall \mathbf{x} \in \mathrm{Cell}_E(\mathbf{p})$ when $\mathbf{p}$ is changing.

4.1 Objective Function and Optimization

For example, we can use the eigenvector f of the second smallest eigenvalue of the graph Laplacian of $\mathbf{G}$ as a *location indicator* [3]. For a graph with N nodes, f is of N-dimensional, with each entry corresponding to a node. We use $(f(\mathbf{x}) - f(\mathbf{y}))^2$ to estimate $d_M^2(\mathbf{x},\mathbf{y})$. Therefore, we can write the objective as

$$\mathscr{E}(\mathbf{P};\mathbf{G}) = \sum_{\mathbf{p}} \sum_{\mathbf{x} \in \mathrm{Cell}_E(\mathbf{p})} (f(\mathbf{x}) - f(\mathbf{p}))^2 \tag{2}$$

However, searching for a $\mathbf{P}$ that optimizes Eq(2) still involves complex combinatorial programming. To make gradient-based optimization possible, we apply a "soft" border cells of different landmarks, inspired by [10]: a point $\mathbf{x}$ is linked to a landmark $\mathbf{p}_k$ by $w_k(\mathbf{x})$:

$$w_k(\mathbf{x}) = \frac{\exp(-\beta \|\mathbf{x} - \mathbf{p}_k\|_2^2/2)}{\sum_j \exp(-\beta \|\mathbf{x} - \mathbf{p}_j\|_2^2/2)} \tag{3}$$

where β is the "hardness" parameter. Therefore we rewrite our objective function as

$$\mathscr{E}(\mathbf{P};\mathbf{G}) = \sum_{i=1}^{N}\sum_{k=1}^{K} w_k(\mathbf{x}_i)(f(\mathbf{x}_i) - f(\mathbf{p}_k))^2 \tag{4}$$

We can optimize Eq(4) w.r.t. $\mathbf{p}_k$ by taking the gradient

$$\frac{\partial \mathscr{E}}{\partial \mathbf{p}_k} = \sum_{i=1}^{N}\sum_{j=1}^{K} [(f(\mathbf{x}_i) - f(\mathbf{p}_j))^2 \frac{\partial w_j(\mathbf{x}_i)}{\partial \mathbf{p}_k} - 2w_j(\mathbf{x}_i)(f(\mathbf{x}_i) - f(\mathbf{p}_j))\frac{\partial f(\mathbf{p}_j)}{\partial \mathbf{p}_k}] \tag{5}$$

where [10] $\frac{\partial w_j(\mathbf{x})}{\partial \mathbf{p}_k} = \beta w_k(\mathbf{x})(\delta_{jk} - w_j(\mathbf{x}))(\mathbf{x} - \mathbf{p}_k)$ and $\frac{\partial f(\mathbf{p}_j)}{\partial \mathbf{p}_k} = \delta_{jk}\nabla f|_{\mathbf{p}_k}$, δ_{jk} is 1 for $j = k$ and 0 otherwise, and ∇f can be pre-computed numerically from the training data. Note that for calculating Eq(5), the computation only needs to be done for $w_j(x_k) \neq 0$. If the border parameter β is "hard", for most j, $w_j(\cdot)$ only has one non-zero element.

In the t-th step, the potential update for the k-th landmark $\mathbf{p}_k^{(t+1)*}$ is found by searching the neighbourhood of $\mathbf{p}_k^{(t)}$ in $\mathbf{G}$ and finding the neighbour to which the vector from $\mathbf{p}_k^{(t)}$ best matches the direction of the gradient in Eq(5).

$$\mathbf{p}_k^{(t+1)*} = \underset{\mathbf{y}\sim\mathbf{p}_k^{(t)}}{\text{argmax}}\ \cos(< \mathbf{y} - \mathbf{p}_k^{(t)}, -\frac{\partial \mathscr{E}^{(t)}}{\partial \mathbf{p}_k^{(t)}} >) \tag{6}$$

A learning rate can be set to decide whether $\mathbf{p}_k^{(t+1)}$ remains the same as $\mathbf{p}_k^{(t)}$ or is updated to $\mathbf{p}_k^{(t+1)*}$. In our implementation, we always update it.

Link to manifold learning One may notice that the f happens to be the results of the non-linear dimensionality reduction of the manifold into 1-D coordinate space using Laplacian eigenmaps [3]. This observation reveals that we are essentially looking for a landmark set that results similar *Euclidean* Voronoi tessellations in both the ambient space and the low-dimensional[2] global manifold coordinate space. Therefore, other manifold coordinate functions (may be vector-valued) may be used for estimating d_M as well. Note that our landmarks are for generalizing learned models, thus requiring the low-dimensional embedding of the data points to be known does not result overheads in practice.

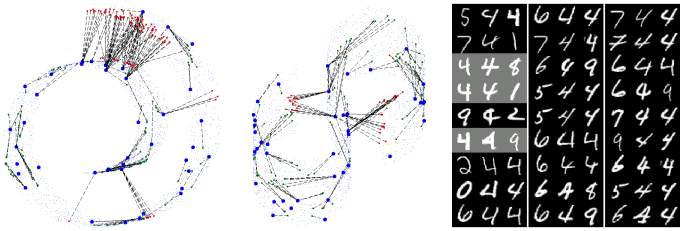
5 Experiments

TEP detection In Figure 3, we show the results of testing Cond. 1 on two synthetic data sets. We randomly select 40 landmarks out of 2000 points as initialization, and the neighbourhood size is 8 for constructing $\mathbf{G}$. In the figure, we link $(\mathbf{x}, L_E(\mathbf{x}))$ when $L_E(\mathbf{x}) \neq L_M(\mathbf{x})$. Out of those $\mathbf{x}$-s, we mark TEPs with red dots. We can see that Condition 1 is in consistence with the intuition whether the link between $\mathbf{x}$ and $L_E(\mathbf{x})$ is a "short-circuit" on the manifold.

We then apply the test on $12,000$ handwritten digit images [3]. Those 28×28 images are preprocessed with PCA and the first 100 principal components are used. The neighborhood size is set to 3 for constructing $\mathbf{G}$. In this experiment, 100 randomly selected

[2] In this case, it is 1-D.

[3] http://www.cs.toronto.edu/ roweis/data.html



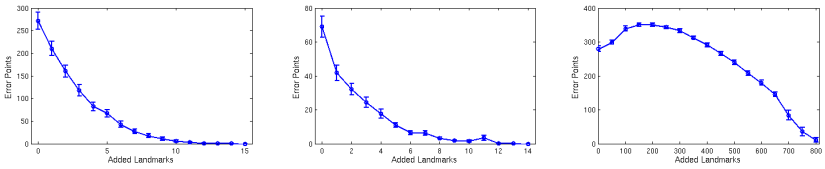
(a) (b) (c)

	0	1	2	3	4	5	6	7	8	9	Total
Bad Points	25	1	82	33	27	35	30	15	53	13	314
π_E^{Err}	25	1	81	29	24	29	27	13	45	10	284
$\pi_E^{Err}\pi_M^{Corr}$	18	1	57	14	18	21	21	10	22	1	183
$\pi_E^{Corr}\pi_M^{Err}$	0	0	1	4	3	6	1	0	5	2	22

(d)

Figure 3: Topological Error Points

(a) and (b): Big blue dots: landmarks; Red dots: TEPs; Green dots: $\pi_E \neq \pi_M$, but NOT TEPs. Links are made between a point and its π_E. (c): TEPs found for "4". Table (d): See text



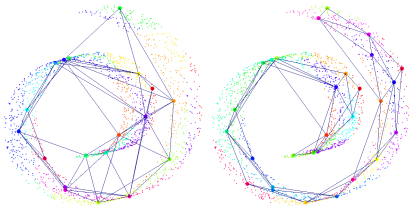
Swiss roll S-shaped Digits

Figure 4: Added Landmarks and TEPs

images are taken as landmarks. In Figure 3(c), we show the TEPs found for the digit "4". In each triplet of the images, the middle one is the sample $\mathbf{x}$, the left is $Ł_E(\mathbf{x})$, and the right is $Ł_M(\mathbf{x})$. We can see that in most cases, $Ł_M(\mathbf{x})$ is the proper landmark for $\mathbf{x}$ to be represented. The *exceptions* are highlighted. In table (b), for each digit , we list the statistics of: (i) number of TEPs detected, (ii) $Ł_E(\mathbf{x})$ with incorrect label, (iii) erroneous $Ł_E(\mathbf{x})$ but correct $Ł_M(\mathbf{x})$ (potential gains) and (iv) correct $Ł_E(\mathbf{x})$ but erroneous $Ł_M(\mathbf{x})$ (potential loss). Most of the TEPs indicate mis-classified samples, and can be corrected by representing the TEP with the corresponding $Ł_M(\mathbf{x})$.

Eliminate TEPs by adding In Figure 4, we show the number of TEPs versus the number of the added landmarks for two synthetic data sets and the hand-written data, respectively. The error bar is computed in 10 runs. Note that for the hand-written data, the number of TEPs increases during the first few iterations. A possible explanation is that as the landmarks increase, the partition of the manifold becomes finer. Therefore, some structural details on the manifold may come out (considering the case of the small U-shaped structure we discussed in Section 3).

In Figure 5 we draw the initial and the resulting landmarks on Swiss roll data. In the figure, two landmarks are linked by an edge if their Euclidean Voronoi cells are adjacent. It shows that the minor graph generated by the initial landmarks contains many edges which connect remote parts of the manifold, while the minor graph generated by the resulting landmarks preserves the topology of the manifold.



Initial Result

Figure 5: Adjacency of Voronoi Cells of Landmarks

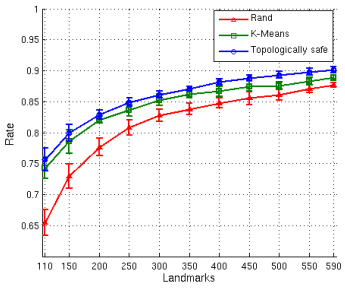


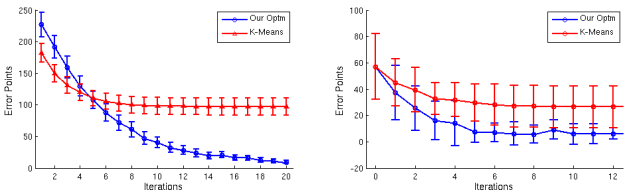
Figure 6: Handwritten Digits Recognition with Landmarks

Landmarks for classification In Figure 6, we use different numbers of landmarks for nearest neighbour classifier on 10000 test digits. The landmarks are selected randomly, by K-Means and by eliminating the TEPs, respectively. The error bar is computed from 10 runs. Our proposed topology safe landmarks perform constantly better.

Optimizing the landmarks We randomly choose 30 landmarks in the data points, and then apply both our optimization and the K-Means. After each iteration, we find TEPs in the data points. The results are shown in Figure 7. The landmarks found by our optimization represent the topology of the data manifold more reliably than those selected by K-Means.

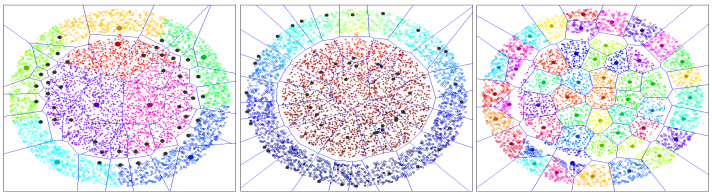
Semi-supervised clustering We use the data shown in Figure 8. In each data set, there are $5,000$ training samples and another $5,000$ for testing. The neighborhood size is 8. The experiment scheme is as follows:

1. Select a few points as the initial landmarks $\mathbf{P}_{Init}$ and label them.

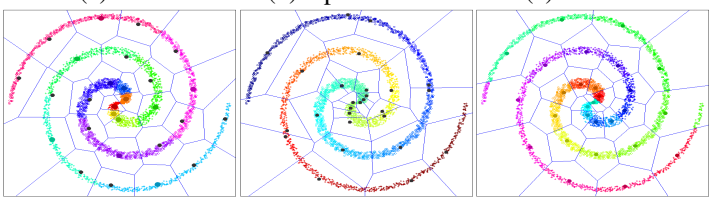


Swiss roll S-shaped

Figure 7: Optimizing Landmarks: TEPs v.s. Iteration Number



Error: 5 (a) Add | Error: 3 (b) Optimization | Error: 116 (c) K-Means



Error: 10 (d) Add | Error: 36 (e) Optimization | Error: 150 (f) K-Means

Figure 8: Semi-Supervised Classification

(a) and (d): initial landmarks (the big dots, color for initial and black for added), samples are colourized with L_M in the initial landmarks. (b) and (e): colourized with f. (c) and (f): colourized with L_E in the resulting landmarks. In each pane, "+"/"o" indicate the true classes.

2. Detect and eliminate the TEPs, when selecting a TEP as a new landmark, label it with L_M. Recorded the number of added landmarks N_{Add}.
3. Run the optimization algorithm with N_{Add} random selected landmarks and $\mathbf{P}_{Init}$ labeled.
4. Initialized with the same landmarks, run K-Means.

Then we use each of these resulting landmark sets to classify 5000 test points. The errors are shown in the figure.

6 Conclusion

In this paper, a criterion of landmarks for reliably representing the data manifold is proposed, as well as an optimization scheme to improve the existing landmark sets. Our condition ensures that the individual Voronoi cells generated by the landmarks in the ambient space do NOT intersect the manifold at faraway locations and thus preserves the manifold structure. They can help reliably locate novel samples to the correct regions on the manifold. Our method is more robust and applicable than the previous work, because it does not require to know the manifold dimensionality.

In the future work, we will study the expected risk by taking a statistical point of view of the data. We will also consider the possibility that generalizing this to multiple metric spaces as well.

Acknowledgement

The authors would like to express their gratitude to their colleague Dr. Fabrizio and Dr. Howarth, who gave very helpful advices and helped much with the writing.

A Proof of Proposition 1

Proposition 1 (1) *If* $\mathbf{p}_i \sim \mathbf{p}_j$ *in* $\mathbf{H}_E$*, according to the contraction procedure, we have* $\mathrm{Cell}_E(\mathbf{p}_i) \sim \mathrm{Cell}_E(\mathbf{p}_j)$ *in* $\mathbf{G}$ *(see below). Therefore, there exists equidistant points on* $\mathscr{M}$ *to* $\mathbf{p}_i$ *and* $\mathbf{p}_j$*. Take one of these points,* $\mathbf{x}$*, and let* $\mathrm{L}_M(\mathbf{x}) = \mathbf{p}_k$*. By Condition 1,* $\{\mathbf{x}\} \subset \mathrm{Cell}_E(\mathbf{p}_i) \cap \mathrm{Cell}_M(\mathbf{p}_k) \neq \emptyset$ *implies* $\mathrm{Cell}_M(\mathbf{p}_i)$ *and* $\mathrm{Cell}_M(\mathbf{p}_k)$ *are adjacent, and thus* $\mathbf{p}_i \sim \mathbf{p}_k$ *in* $\mathbf{H}_M$*. Using the same deduction, we have* $\mathbf{p}_j \sim \mathbf{p}_k$ *in* $\mathbf{H}_M$ *as well. If* $k = i$ *or* $k = j$*, then* $\mathbf{p}_i \sim \mathbf{p}_j$ *in* $\mathbf{H}_M$ *(Figure 2(a)). Otherwise,* $\mathbf{p}_i$ *and* $\mathbf{p}_j$ *are connected in* $\mathbf{H}_M$ *by a path* $(\mathbf{p}_i, \mathbf{p}_k, \mathbf{p}_j)$ *(Figure 2(b)).*

In practice, however, there are generally no such equidistant points in the discrete point set sampled from $\mathscr{M}$*. Nevertheless, because* $\mathrm{Cell}_E(\mathbf{p}_i) \sim \mathrm{Cell}_E(\mathbf{p}_j)$ *in* $\mathbf{G}$*,* $\exists \mathbf{x}_1 \sim \mathbf{x}_2$ *in* $\mathbf{G}$*, where* $\mathrm{L}_E(\mathbf{x}_1) = \mathbf{p}_i$ *and* $\mathrm{L}_E(\mathbf{x}_2) = \mathbf{p}_j$*. Consider* $\mathrm{L}_M(\mathbf{x}_1) = \mathbf{p}_{k_1}$ *and* $\mathrm{L}_M(\mathbf{x}_2) = \mathbf{p}_{k_2}$*, because* $\mathbf{x}_1 \sim \mathbf{x}_2$*, we have* $\mathrm{Cell}_M(\mathbf{p}_{k_1}) \sim \mathrm{Cell}_M(\mathbf{p}_{k_2})$ *in* $\mathbf{G}$ *and equivalently* $\mathbf{p}_{k_1} \sim \mathbf{p}_{k_2}$ *in* $\mathbf{H}_M$*. And as discussed in the continuous case above, we have* $\mathbf{p}_i \sim \mathbf{p}_{k_1}$ *and* $\mathbf{p}_j \sim \mathbf{p}_{k_2}$ *in* $\mathbf{H}_M$*. Depending on whether (a)* $k_1 = i, k_2 = j$*, (b-i)* $k_1 = i, k_2 \neq j$*, (b-ii)* $k_1 \neq i, k_2 = j$ *or (d)* $k_1 \neq i, k_2 \neq j$*, the path between* $\mathbf{p}_i$ *and* $\mathbf{p}_j$ *in* $\mathbf{H}_M$ *corresponds to one of the cases (a)-(c) in Figure 2.*

(2) *If* $\mathbf{p}_i \sim \mathbf{p}_j$ *in* $\mathbf{H}_M$*, consider the shortest path* $(\mathbf{p}_i = \mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_n = \mathbf{p}_j)$ *in* $\mathbf{G}$*. It is not difficult to see, all points on the path is within either* $\mathrm{Cell}_M(\mathbf{p}_i)$ *or* $\mathrm{Cell}_M(\mathbf{p}_j)$*. Thus by Condition 1, if* $\mathrm{Cell}_E(\mathbf{p})$ *contains these points, either (i)* $\mathbf{p}$ *is* $\mathbf{p}_i, \mathbf{p}_j$ *or (ii)* $\mathrm{Cell}_M(\mathbf{p})$ *is adjacent to* $\mathrm{Cell}_M(\mathbf{p}_i)$ *or* $\mathrm{Cell}_M(\mathbf{p}_j)$*. Let* $\mathbf{U}^M_{i,j} = \{\mathbf{p} | \mathbf{p} \sim \mathbf{p}_i \text{ or } \mathbf{p} \sim \mathbf{p}_j\}$ *in* $\mathbf{H}_M$*. There exists a path between* $\mathbf{p}_i$ *and* $\mathbf{p}_j$ *in* $\mathbf{H}_E$ *consisting only nodes in* $\mathbf{U}^M_{i,j}$ *(Figure 2(d)).*

References

[1] N. Amenta, M. Bern, and M. Kamvysselis. A new voronoi-based surface reconstruction algorithm. In *SIGGRAPH*, 1998.

[2] M. Belkin. *Problems of learning on manifolds*. PhD thesis, University of Chicago, 2003.

[3] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 2003.

[4] Y. Bengio, J. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In *NIPS*, 2004.

[5] F. R. K. Chung. *Spectral Graph Theory*. 1997.

[6] V. de Silva and J. B. Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. In *NIPS*, 2002.

[7] X. He, S. Yan, Y. Hu, and P. Niyogi. Face recognition using laplacianfaces. *TPAMI*, 2005.

[8] X. Huo and A. K. Smith. Performance analysis of a manifold learning algorithm in dimension reduction. Technical report, Georgia Institute of Technology, 2006.

[9] R. Jenssen, D. Erdogmus, J. Principe, and T. Eltoft. The laplacian pdf distance: A cost function for clustering in a kernel feature space. In *NIPS*. 2005.

[10] S. Lazebnik and M. Raginsky. Learning nearest-neighbor quantizers from labeled data by information loss minimization. In *AISTATS*, 2007.

[11] J. Li and P. Hao. Hierarchical structuring of data on manifolds. In *CVPR*, 2007.

[12] J.G. Silva and J.S. Marques. Selecting landmark points for sparse manifold learning. In *NIPS*, 2005.

[13] P. van Oosterom. *Spatial Access Methods*, pages 385–400. Wiley, 1999.

[14] Weinberg05. Mvu. In *AISTAT*, 2005.

[15] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *TPAMI*, 2007.