Foundations of Data Science
©American Institute of Mathematical Sciences
Volume **1**, Number **3**, September **2019**
doi:10.3934/fods.2019012
pp. **271–291**

LEARNING BY ACTIVE NONLINEAR DIFFUSION

Mauro Maggioni

Department of Mathematics, Department of Applied Mathematics and Statistics
Mathematical Institute of Data Sciences
Institute of Data Intensive Engineering and Science
Johns Hopkins University
Baltimore, MD 21218, USA

James M. Murphy*

Department of Mathematics
Tufts University
Medford, MA 02155, USA

Abstract. This article proposes an active learning method for high-dimensional data, based on intrinsic data geometries learned through diffusion processes on graphs. Diffusion distances are used to parametrize low-dimensional structures on the dataset, which allow for high-accuracy labelings with only a small number of carefully chosen training labels. The geometric structure of the data suggests regions that have homogeneous labels, as well as regions with high label complexity that should be queried for labels. The proposed method enjoys theoretical performance guarantees on a general geometric data model, in which clusters corresponding to semantically meaningful classes are permitted to have nonlinear geometries, high ambient dimensionality, and suffer from significant noise and outlier corruption. The proposed algorithm is implemented in a manner that is quasilinear in the number of unlabeled data points, and exhibits competitive empirical performance on synthetic datasets and real hyperspectral remote sensing images.

1. **Introduction.** Statistical and machine learning techniques are revolutionizing the sciences. Advances in medical diagnosis [27], automatic game playing [57], and computer vision [35] have been sparked by seismic advances in computational power and innovative learning algorithms and architectures. However, many state-of-the-art machine learning approaches are predicated on the availability of huge labeled data sets that may be used to train the parameters of the underlying models. Unfortunately, many important scientific problems do not have large, accurately labeled training sets readily available. This limits the practicality of many state-of-the-art supervised methods. Moreover, several fields—medicine and remote sensing, for example—are not amenable to easily generating new labeled data points at scale, due to the high cost of labeling data points. This renders the applicability of many state-of-the-art supervised learning algorithms—including modern deep

2010 *Mathematics Subject Classification.* Primary: 58F15, 58F17; Secondary: 53C35.

Key words and phrases. Active learning, statistical learning, diffusion geometry, machine learning, spectral graph theory.

This research is supported by NSF-DMS-125012, NSF-DMS-1724979, NSF-DMS-1708602, NSF-ATD-1737984, AFOSR FA9550-17-1-0280, NSF-IIS-1546392, NSF-DMS 1912737, and NSF-DMS 1924513.

* Corresponding author: James M. Murphy.

learning methods which may depend on millions of parameters—problematic, as generating sufficient training data may be resource-intensive.

When training datasets do not exist or are burdensome to generate, alternative methods may be used to exploit the glut of unlabeled data. *Data augmentation* [59, 62] may be used to generate new labeled training points by, for example, perturbing existing training points in a suitable manner. *Unsupervised methods*—those using no training (labeled) data at all—are ideal when insufficient training data is available, as they work entirely on the unlabeled data. However, unsupervised methods may be inadequate for highly complex data. Indeed, such approaches enjoy performance guarantees only when rigid geometrical or statistical properties are made on the data [5, 6, 54, 42, 39, 30]. Methods that are *semi-supervised* [16] provide a middle ground between the supervised (abundant labeled data for training) and unsupervised (no labeled data for training) regimes, taking advantage of large quantities of unlabeled data while still allowing labeled points to influence classification. When the unsupervised structure of the data (e.g., its geometric or statistical properties) is compatible with the labels of the data, semisupervised learning may improve over unsupervised learning and also over classical supervised learning with the same fixed labeled training data.

This article proposes an *active learning scheme* for high-dimensional datasets exhibiting intrinsically low-dimensional structure. Active learning is a form of semi-supervised learning in which an algorithm uses the unlabeled data to determine which data points to query for labels. In the proposed method, the geometry of the data is parametrized through diffusion processes defined on a data-dependent graph [23, 22], which are robust to high ambient dimensionality, noise, and non-spherical cluster shapes. The inferred geometry—which is computed without supervision—is then analyzed to determine which data points should be queried for labels; the query points are chosen to have maximum impact, so that relatively few are needed to achieve good empirical performance. The proposed active learning scheme is called *learning by active nonlinear diffusion (LAND)*.

1.1. **Major contributions and article outline.**

The major contributions of this article are twofold. First, LAND is proposed and is proven to perform well for data generated according to a flexible geometric data model. With only a small number of queries, LAND achieves perfect accuracy even for data that is high-dimensional, contains classes that are highly nonlinear or non-compact, and is corrupted by significant noise and outliers. The theoretical results are derived from an analysis of the underlying diffusion distances, which in turn are amenable to analysis using techniques from spectral graph theory and the analysis of Markov chains.

Second, the proposed method is implemented numerically. Taking advantage of fast nearest neighbor search algorithms and eigensolvers for sparse matrices, the proposed method is proven to enjoy *quasilinear complexity* in the number of sample points under the proposed data model, which supposes that the underlying data has intrinsically small dimensionality (in the sense of lying close to a low-dimensional manifold, for example). LAND is demonstrated on synthetic datasets as well as real hyperspectral images, demonstrating its suitability for high-dimensional geometric data.

The remainder of the article is organized as follows. Background on active learning and diffusion geometry are presented in Section 2. The geometric data model and algorithm are proposed and analyzed in Section 3. Comparisons with related

works are also presented in Section 3. Numerical experiments are in Section 4. Conclusions and future research directions are in Section 5.

2. Background.

The proposed active learning algorithm exploits the underlying diffusion geometry of data to efficiently determine points to query for labels. In this section, we review active learning as well as diffusion geometry.

2.1. Background on active learning.

Active learning is a type of semisupervised learning in which unlabeled data is analyzed to determine which points to query for labels [55]. It differs from traditional semisupervised learning in that the labeling algorithm is permitted to *ask for the labels of certain points*, instead of being provided with a random sample of labeled points. Under certain data models and methods for parsimoniously selecting query points, the active learning approach can perform as well as traditional semi-supervised or even supervised learning, with far fewer labels [21, 26]. The crucial theoretical question is how to determine which data points should be queried for labels. The active learning framework assumes there is an underlying budget that can be spent to label points. This budget should be spent carefully, in order to only query points that are most likely to prove significant for the overall labeling of the data.

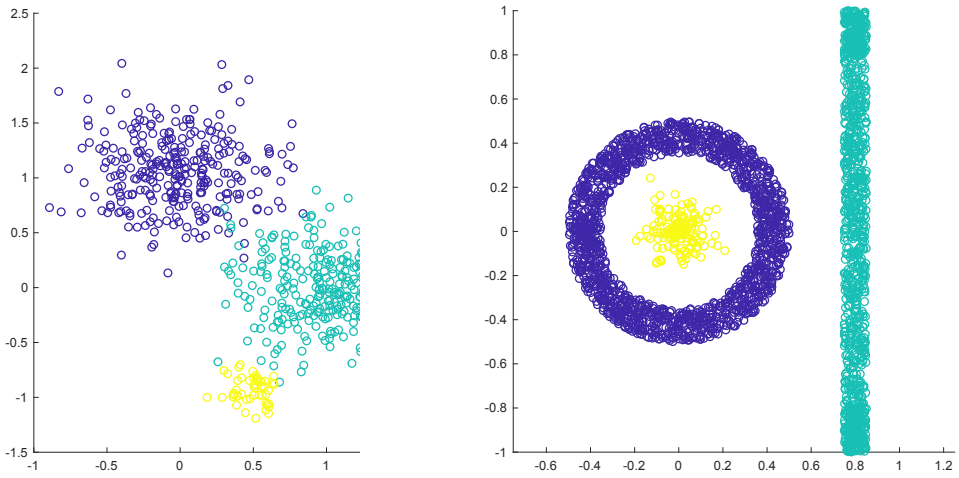
Approaches to active learning may be categorized into two general strategies: hypothesis space reduction and cluster exploitation [24]. The first category conceives of supervised learning as a process of using training points to select a "good" classifier from a large space of possible classifiers. Asymptotically, as the number of labeled sample points $n_\ell \to \infty$, a consistent supervised learning procedure converges to an optimal classifier. In practice, the rate of convergence in n_ℓ is relevant—the faster the rate of convergence, the better the learning algorithm. In this framework, active learning is a family of methods for selecting query points such that the convergence rate towards a good classifier is fast in n_ℓ, in particular faster than passive sampling methods, for example sampling labels uniformly at random. That is, query points should be influential in distinguishing between different possible classifiers, and should allow for convergence towards the "optimal" classifier with fewer points than if the labeled points were selected uniformly at random. These active learning approaches can, in certain cases, significantly improve the expected error rate of the classifier as a function of n_ℓ [9, 26, 14, 8, 34].

A second category of active learning approaches seek to exploit cluster structure in the data in order to emphasize sampling near complex regions of the data with heterogeneous labels, and to avoid oversampling near simple, homogeneous regions of the data. Indeed, if a cluster—detected through a prescribed clustering algorithm—can be estimated as relatively pure with respect to its labels, then it may be efficient to simply give all points in the cluster the same label and to focus the limited querying resources in more ambiguous regions. A crucial problem in this framework is to tap the budget in a way that balances two different tasks: confirming the label homogeneity of particular data regions and exploring new data regions. Methods based on iteratively pruning hierarchical clustering trees have been proposed [25] and analyzed in terms of label smoothness with respect to the scales of the tree [61].

The method proposed in this paper is related to the second category, and exploits the underlying geometry of the data sample in order to estimate the most impactful points to query for labels. In order to develop notions of cluster geometry that are robust to being embedded in a high-dimensional space, to being non-spherical in

shape, and to corruption by noise and outlier points, the diffusion geometry of the underlying data is estimated and used as the basis for all subsequent pairwise comparisons. This provides a set of (essentially) geometrically intrinsic coordinates for the data that are robust to dimensionality, nonlinearity, and noise.

An example of synthetic toy data for which diffusion geometry notably decreases the number of active learning queries necessary for good accuracy appears in Figure 1. The role of diffusion geometry is crucial to the proposed method, and it is reviewed in detail in Section 2.2.



(a) Gaussian data

(b) Nonlinear and elongated data

Figure 1. Data colored by class label. Both the data in (a) and (b) exhibit cluster structure, which can guide active learning in the case that the labels are constant on these clusters. Indeed, on the left, using a simple clustering algorithm such as K-means suggests that only 3 labels are necessary to correctly label the entire dataset. For the data on the right, many more than three labels are necessary if K-means is used for the underlying clustering, since the clusters are highly elongated and nonlinear. Indeed, K-means will split the annular and elongated clusters. On the other hand, if pairwise comparisons are made with distances other than Euclidean distances, it may be possible that active learning achieves near perfect results with only 3 labels. The proposed active learning scheme gains robustness to class shape via diffusion geometry, and is suitable for data in both (a) and (b).

2.2. **Background on diffusion geometry.** Let $X = \{x_i\}_{i=1}^n \subset \mathbb{R}^D$ be discrete data. The *diffusion geometry* of X is learned through Markov diffusion processes defined on a graph with nodes corresponding to the points $\{x_i\}_{i=1}^n$ and transition probabilities proportional to the similarities of these points in some metric [23, 22]. That is, points that are nearby have high probabilities of pairwise transition, and points that are far apart have low probabilities of transition. By analyzing the diffusion process across time scales, natural geometric structure in the data can be inferred.

More precisely, let $\mathcal{G} = (X, W)$ be a weighted, undirected graph with nodes X and weight $W_{ij} \in [0, 1]$ between $x_i, x_j \in X$. Typically $W_{ij} = \mathcal{K}(x_i, x_j)$ for some symmetric, radial kernel $\mathcal{K} : \mathbb{R}^D \times \mathbb{R}^D \to [0, 1]$. The weight matrix W is normalized to produce a Markov transition matrix $P = D^{-1}W$, where D is the

diagonal degree matrix with $D_{ii} = \sum_{j=1}^{n} W_{ij}$. The matrix P is row-stochastic, and *diffusion distances* measure how similar points are according to their transition probabilities in P.

Definition 2.1. Let P be a Markov transition matrix defined on $X = \{x_i\}_{i=1}^n$. Let $p_t(x_i, x_j) = (P^t)_{ij}$. The *diffusion distance between x_i and x_j at time t with respect to weight $w : X \to [0, \infty)$* is

$$D_t(x_i, x_j) = \|p_t(x_i, \cdot) - p_t(x_j, \cdot)\|_{l^2(w)} = \sqrt{\sum_{\ell=1}^{n} (p_t(x_i, x_\ell) - p_t(x_j, x_\ell))^2 w(x_\ell)}.$$

The time parameter t is a global time scale at which the diffusion process runs. For small t, the process has run for a short amount of time, which may prevent important, large-scale geometric structures in the data from impacting the diffusion distances. On the other extreme, the diffusion distances all collapse to 0 as $t \to \infty$, under the assumption that P is ergodic, since P^t converges to the rank 1 matrix with the stationary distribution π as rows, where $\pi P = \pi$. When the data has underlying geometric structure, t parametrizes multiscale hierarchy, with small t realizing fine-scale structures and t large realizing coarse-scale structures [46, 31, 40].

While P is not symmetric, it is diagonally conjugate to a symmetric matrix: $D^{\frac{1}{2}} P D^{-\frac{1}{2}} = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$. Hence, P admits a spectral decomposition, which can be exploited for the computation of diffusion distances. More precisely, let $\{(\lambda_\ell, \phi_\ell)\}_{\ell=1}^n$ be the eigenvalues and eigenvectors of $D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$, sorted so that $1 = \lambda_1 > |\lambda_2| \geq \cdots \geq |\lambda_n|$. Then

$$(P^t)_{ij} = \sum_{\ell=1}^{n} \lambda_\ell^t \psi_\ell(x_i) \varphi_\ell(x_j),$$

where $\psi_\ell(x_i) = \phi_\ell(x_i)/\sqrt{\pi(x_i)}, \varphi_\ell(x_j) = \phi_\ell(x_j)\sqrt{\pi(x_j)}$. If $\{\psi_\ell\}_{\ell=1}^n, \{\varphi_\ell\}_{\ell=1}^n$ are understood as column vectors, this is equivalent to the decomposition $P^t = \sum_{\ell=1}^n \lambda_\ell^t \psi_\ell \varphi_\ell^\top$. In particular, $\{\varphi_\ell\}_{\ell=1}^n$ is an orthonormal basis for $l^2(1/\pi)$, so that diffusion distances with respect to the weight $w(x_i) = 1/\pi(x_i)$ may be written in terms of $\{\psi_\ell\}_{\ell=1}^n$:

$$D_t(x_i, x_j) = \|p_t(x_i, \cdot) - p_t(x_j, \cdot)\|_{l^2(1/\pi)} = \sqrt{\sum_{\ell=1}^{n} \lambda_\ell^{2t} (\psi_\ell(x_i) - \psi_\ell(x_j))^2}.$$

If the underlying transition matrix P is approximately low rank, the modulus of the eigenvalues $\{\lambda_\ell\}_{\ell=1}^n$ decays rapidly, so that for t sufficiently large, this sum may be truncated after $M = O(1)$ eigenpairs yielding the approximate diffusion distances

$$D_t(x_i, x_j) \approx \sqrt{\sum_{\ell=1}^{M} \lambda_\ell^{2t} (\psi_\ell(x_i) - \psi_\ell(x_j))^2}.$$

This truncation has the added benefit of denoising the diffusion distances, since the eigenvectors associated with eigenvalues away from 1 in modulus (in some sense the high frequency eigenvectors) correspond not to intrinsic geometric structures in the data, but to random fluctuations produced by sampling [29]. The embedding

$$x_i \mapsto (\lambda_1^t \psi_1(x_i), \lambda_2^t \psi_2(x_i), \ldots, \lambda_M^t \psi_M(x_i))$$

may be understood as a form of nonlinear dimension reduction, and also as a set of (essentially) geometrically intrinsic coordinates for the data [36].

3. **Proposed algorithm and analysis.** Let $\{x_i\}_{i=1}^n \subset \mathbb{R}^D$. The LAND algorithm requires determining which points should be queried for labels. This is done by estimating modes of the nonlinear clusters in the data through a combination of density estimation and the diffusion geometry of the data.

Let $p : \mathbb{R}^D \to [0, \infty)$ be a kernel density estimator, for example

$$p(x) = \frac{1}{Z} \sum_{y \in \mathrm{NN}_k(x)} \exp(-\|x - y\|_2^2/\sigma_0^2),$$

where $\mathrm{NN}_k(x)$ are the k-nearest neighbors of x in Euclidean distance, σ_0 is a scaling parameter, and Z is a normalization constant so that $\sum_{x \in X} p(x) = 1$. Let D_t be the diffusion distance metric on X, and let

$$\rho_t(x) = \begin{cases} \min\{D_t(x, y) \ | p(y) \geq p(x), x \neq y\}, & x \neq \arg\max_z p(z) \\ \max_{y \in X} D_t(x, y), & x = \arg\max_z p(z) \end{cases} \tag{1}$$

be the (t-dependent) diffusion distance between a point and its nearest diffusion neighbor of higher density if x is not the maximizer of $p(x)$, and the maximum diffusion distance to another point if x is the maximizer of $p(x)$. The modes of the data are determined through the quantity

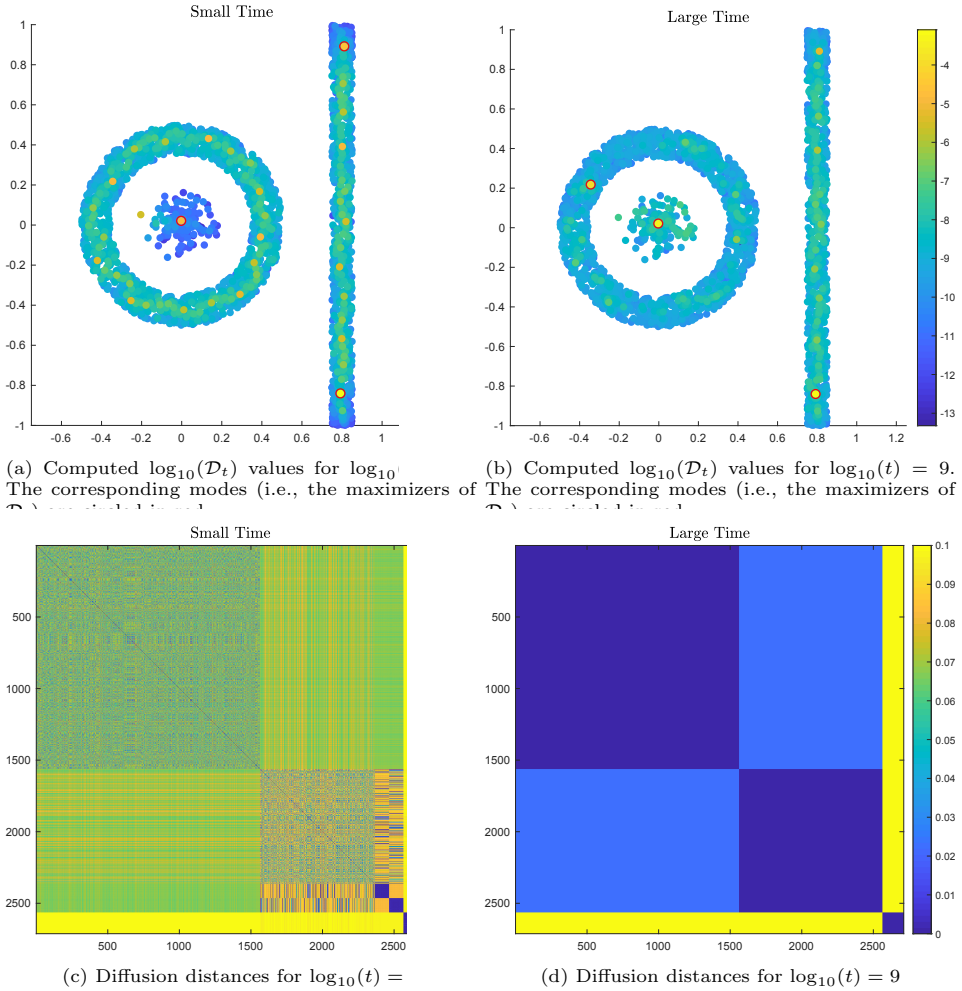
$$\mathcal{D}_t(x) = p(x)\rho_t(x).$$

Points will have a large $\mathcal{D}_t$ value if they are high density and are D_t-far from other high density points. Following [40], we characterize the modes of X as the maximizers of $\mathcal{D}_t$. This notion is robust to data geometry—as captured by diffusion distances—and provides a multiscale hierarchy to the structure of the data. See Figure 2 for an illustration of how $\mathcal{D}_t$ changes with time.

3.1. **Learning by unsupervised nonlinear diffusion.** In [40], the maximizers of $\mathcal{D}_t$ were proposed as cluster modes, and diffusion distances and density were used to label all other points relative to these modes. We summarize this unsupervised learning algorithm, called *learning by unsupervised nonlinear diffusion (LUND)* in Algorithm 1. This algorithm was proven to perfectly cluster certain data, for an appropriate choice of time parameter t, and is robust to non-spherical data geometries and cluster overlap.

It was shown that, depending on the well-connectedness of the clusters compared to their separations, the range of t for which Algorithm 1 performs well may be large [40]. However, developing methods for estimating an appropriate choice of t without using any labeled data is an important and only partially addressed problem. Indeed, if the data admits hierarchical cluster structure, then several choices of t may be appropriate, leading to different reasonable clusterings. In this context, querying a small number of points for labels can disambiguate between these different clusterings.

3.2. **Learning by active nonlinear diffusion.** In the active learning setting, we characterize potential classes as being composed of D_t-orbits around the maximizers of $\mathcal{D}_t$. These orbits partition the data, and are comparable to elements of a Voronoi tessellation [7]. In the case that the labels for the data are smooth with respect to this partition, querying the maximizers of $\mathcal{D}_t$ is a more efficient use of a sampling budget than uniform random sampling. The proposed algorithm, denoted *Learning by Active Nonlinear Diffusion (LAND)* appears in Algorithm 2.



(a) Computed $\log_{10}(\mathcal{D}_t)$ values for $\log_{10}$ The corresponding modes (i.e., the maximizers of

(b) Computed $\log_{10}(\mathcal{D}_t)$ values for $\log_{10}(t) = 9$. The corresponding modes (i.e., the maximizers of

(c) Diffusion distances for $\log_{10}(t) =$

(d) Diffusion distances for $\log_{10}(t) = 9$

Figure 2. In (a) and (b), the values of $\log_{10}(\mathcal{D}_t)$ are shown for synthetic geometric data from Figure 1 (b) for $\log_{10}(t) = 2$ and $\log_{10}(t) = 9$, respectively. We see that for small values of t, the mode estimation incorrectly places the first three modes on the highly elongated cluster. For larger time values, the underlying random walk reaches a mesoscopic equilibrium and correct mode estimation is achieved. The emergence of mesoscopic equilibria is apparent in (c), (d), which show the matrix of diffusion distances at time $\log_{10}(t) = 2$ and $\log_{10}(t) = 9$, respectively. When $\log_{10}(t) = 2$, P^t has not mixed, and there are still substantial within-cluster distances. For $\log_{10}(t) = 9$, P^t has reached mesoscopic equilibria, so that within-cluster distances are quite small, yet between-cluster distances are still large [40].

3.2.1. *Analysis of LAND.* From a theoretical standpoint, it is of interest to know when querying a small number of points (Algorithm 2) offers substantial be benefit compared to unsupervised learning (Algorithm 1). Suppose that the underlying data consists of distinct classes $X = \bigcup_{k=1}^{K} X_k$, with all points in X_k having label k. Let

$$D_t^{\text{in}} = \max_{k=1,\dots,K} \max_{x,y \in X_k} D_t(x,y),$$

Algorithm 1 Learning by Unsupervised Nonlinear Diffusion (LUND)

Input:

- $\{x_i\}_{i=1}^n$ (Unlabeled Data)
- $\{(\lambda_\ell, \psi_\ell)\}_{\ell=1}^M$ (Spectral Decomposition of P)
- $\{p(x_i)\}_{i=1}^n$ (Empirical Density Estimate)
- $\{\rho_t(x_i)\}_{i=1}^n$ (1)
- t (Time Parameter)

Output:

- $\hat{K}$ (Estimated Number of Clusters)
- Y (Labels)

Compute $\mathcal{D}_t(x_i) = p(x_i)\rho_t(x_i)$.
Sort the data in order of decreasing $\mathcal{D}_t$ value to acquire the ordering $\{x_{m_i}\}_{i=1}^n$.
Estimate $\hat{K} = \arg\max_i \left(\mathcal{D}_t(x_{m_i})/\mathcal{D}_t(x_{m_{i+1}})\right)$.
for $k = 1 : \hat{K}$ **do**
  $Y(x_{m_k}) = k$.
end for
Sort X according to $p(x)$ in decreasing order as $\{x_{\ell_i}\}_{i=1}^n$.
for $i = 1 : n$ **do**
  if $Y(x_{\ell_i}) = 0$ **then**
    $Y(x_{\ell_i}) = Y(z_i)$, $z_i = \arg\min_z \{D_t(z, x_{\ell_i}) \mid p(z) > p(x_{\ell_i}) \text{ and } Y(z_i) > 0\}$.
  end if
end for

$$D_t^{\text{btw}} = \min_{k \neq k'} \min_{x \in X_k, y \in X_{k'}} D_t(x, y)$$

be the maximum within-class and minimum between-class diffusion distances at time t, respectively. Let

$$\mathcal{M} = \{x \in X \mid \exists k \text{ such that } x = \arg\max_{y \in X_k} p(y)\},$$

$\max(\mathcal{M}) = \max_{x \in \mathcal{M}} p(x), \min(\mathcal{M}) = \min_{x \in \mathcal{M}} p(x)$ be the density maximizers of the distinct classes, the maximum density among such classwise maximizers, and the minimum density among the classwise maximizers, respectively. In [40], it is shown that if

$$D_t^{\text{in}}/D_t^{\text{btw}} < \min(\mathcal{M})/\max(\mathcal{M}), \tag{2}$$

then the data can be labeled in a fully unsupervised manner by Algorithm 1. However, the underlying density conditions may not be satisfied in practice, particularly if there are strong discrepancies between the density of the most dense point in each cluster. Moreover, (2) depends strongly on t. Introducing the active learning scheme allows to bypass this potentially stringent density condition and still achieve perfect accuracy, at the cost of querying the labels of a small number of points.

Theorem 3.1. *Let $X = \bigcup_{k=1}^K$ be data to classify. Suppose that $D_t^{in} < D_t^{btw}$, and that the B maximizers of $\mathcal{D}_t$ include the elements of $\mathcal{M}$. Then LAND with a budget of size B achieves perfect classification accuracy.*

Proof. If the B maximizers of $\mathcal{D}_t$ include all the density maximizers of the distinct classes, that is, the elements of $\mathcal{M}$, then the LAND queries guarantee these points

Algorithm 2 Learning by Active Nonlinear Diffusion (LAND)

Input:

- $\{x_i\}_{i=1}^n$ (Unlabeled Data)
- $\{(\lambda_\ell, \psi_\ell)\}_{\ell=1}^M$ (Spectral Decomposition of P)
- $\{p(x_i)\}_{i=1}^n$ (Kernel Density Estimate)
- $\{\rho_t(x_i)\}_{i=1}^n$ (1)
- t (Time Parameter)
- B (Budget)
- $\mathcal{O}$ (Labeling Oracle)

Output:

- Y (Labels)

Compute $\mathcal{D}_t(x_i) = p(x_i)\rho_t(x_i)$.
Sort the data in order of decreasing $\mathcal{D}_t$ value to acquire the ordering $\{x_{m_i}\}_{i=1}^n$.
for $i = 1 : B$ **do**
  Query $\mathcal{O}$ for the label $L(x_{m_i})$ of x_{m_i}.
  Set $Y(x_{m_i}) = L(x_{m_i})$.
end for
Sort X according to $p(x)$ in decreasing order as $\{x_{\ell_i}\}_{i=1}^n$.
for $i = 1 : n$ **do**
  if $Y(x_{\ell_i}) = 0$ **then**
    $Y(x_{\ell_i}) = Y(z_i)$, $z_i = \arg\min_z \{D_t(z, x_{\ell_i}) \mid p(z) > p(x_{\ell_i}) \text{ and } Y(z) > 0\}$.
  end if
end for

are all labeled correctly. Then the result follows by induction on the data points sorted in order of decreasing $p(x)$ value. Indeed, for an unlabeled point $x \in X_k$, its nearest diffusion neighbor of higher density, x^*, must be in the same class X_k, since $D_t^{\text{in}} < D_t^{\text{btw}}$. Moreover, that point is already labeled as $Y(x^*) = k$, since $p(x^*) \geq p(x)$. Hence, $Y(x) = k$. □

Theorem 3.1 asserts that the LAND algorithm achieves perfect accuracy as long as $D_t^{\text{in}} < D_t^{\text{btw}}$ and B is large enough so that all elements of $\mathcal{M}$ are among the B maximizers of $\mathcal{D}_t$. Compared to LUND, LAND does not require that $D_t^{\text{in}}/D_t^{\text{btw}} < \min(\mathcal{M})/\max(\mathcal{M})$ to guarantee strong performance. This is an important point in practice, since the density between different regions of the data may vary considerably. Ultimately, active learning is most useful when the budget B may be taken very small compared to n; we shown in Section 4 that even a budget of just a few points may significantly improve accuracy on synthetic and real datasets.

3.3. **Comparison with related methods.** It is natural to compare LAND with related cluster-based active learning methods, as well as its unsupervised variant LUND.

3.3.1. *Comparisons with related active learning methods.* As discussed in Section 2.1, active learning methods may be categorized as falling into two broad classes: those based on refining the hypothesis space of classifiers, and those based on exploiting cluster structure in the data. LAND falls into the second category; it is thus natural to compare it with existing cluster-based active learning algorithms.

Many active learning algorithms that exploit cluster structure in the data proceed by constructing a hierarchical clustering on the data, often represented in the form of a dendrogram [28]. Given such a structure, sample queries are made in order to explore heterogeneous regions of the tree (leaves with highly mixed labels) and to avoid sampling from homogeneous regions of the data (leaves that consist mostly of a single class). The key challenge is to balance the cost of exploring ambiguous regions of the data with establishing the homogeneity of other regions.

Efficient algorithms that are statistically consistent have been proposed [25] and analyzed using the notion of "probabilistic Lipschitzness," which quantifies purity of leaves of the hierarchical clustering [61]. These approaches make analyzing the hierarchical tree the central problem; the problem of whether or not a particular method for constructing a hierarchical tree is appropriate or not is not directly considered. Indeed, it is common to construct the underlying hierarchical tree with standard methods, for example average-linkage clustering [25] or single linkage clustering [28]. Despite their pervasiveness, these methods for constructing hierarchical trees suffer from a lack of robustness to pernicious chains in the data (single-linkage) and geometric distortion (average-linkage). Active learning based on hierarchical trees performs well when the leaves of the tree become pure quickly when descending from the root node; if the underlying tree does not exhibit pure leaves until relatively deep in the tree, many samples are required for active learning, and the method may not improve substantially over random sampling.

Unlike average linkage and single linkage clustering, the proposed LAND method explicitly incorporates the underlying geometry of the data to construct clusters of multiscale granularity, which can then be exploited for active querying. The LAND algorithm may be interpreted as a method for constructing the underlying hierarchical tree, which has the desirable property that the leaves are essentially robust to geometric transformations of the underlying clusters (i.e., to making the clusters elongated or nonlinear). Indeed, given a number of clusters K, one can run a variant of LUND in which K is input as a parameter; see Algorithm 3.

It is then natural to compare the purity of the nodes of a hierarchical tree at scale K, with the purity of the clusters learned by Algorithm 3 with number of clusters equal to K. More generally, let $\mathcal{C} = \{C_k\}_{k=1}^K$ be a clustering of labeled data $\{(x_i, y_i)\}_{i=1}^n$. Let $\bar{y}_k$ be the most common label among the points in C_k. The *purity of the clustering* $\mathcal{C}$ is defined as

$$\mathcal{P}(\mathcal{C}) = \frac{1}{n} \sum_{k=1}^{K} |\{x_i \in C_k \mid y_i = \bar{y}_k\}|.$$

Given a hierarchical clustering $\{\mathcal{C}_\ell\}_{\ell=1}^n$—that is, $\mathcal{C}_1$ consists of 1 cluster with all points, $\mathcal{C}_n$ consists of n singleton clusters, and $\mathcal{C}_{\ell+1}$ is the same clustering as $\mathcal{C}_\ell$, but with two of the clusters split— the purity of the clustering at the ℓ^{th} scale is $\mathcal{P}(\mathcal{C}_\ell)$. Clearly $\mathcal{P}(\mathcal{C}_\ell)$ is non-decreasing as a function of ℓ, and $\mathcal{P}(\mathcal{C}_n) = 1$. If the growth of $\mathcal{P}(\mathcal{C}_\ell)$ towards 1 is rapid in ℓ, then an active sampler does not need to search deeply into the tree to find regions with homogeneous labels. In Figure 3, a plot of $\mathcal{P}(\mathcal{C}_\ell)$ is shown for three synthetic datasets with three different families of clusterings: single linkage clusters, average linkage clusters, and the clusters learned by Algorithm 3.

We see that for the geometric data, the clusters learned from average linkage clustering achieve high purity much later than the clusterings learned with single linkage clustering. This is due to the inability of average linkage to account for the nonlinear and elongated shapes of these clusters. Indeed, the opposite ends of

Algorithm 3 LUND, K Known

Input:

- $\{x_i\}_{i=1}^n$ (Unlabeled Data)
- $\{(\lambda_\ell, \psi_\ell)\}_{\ell=1}^M$ (Spectral Decomposition of P)
- $\{p(x_i)\}_{i=1}^n$ (Empirical Density Estimate)
- $\{\rho_t(x_i)\}_{i=1}^n$ (1)
- t (Time Parameter)
- K (Number of Clusters)

Output:

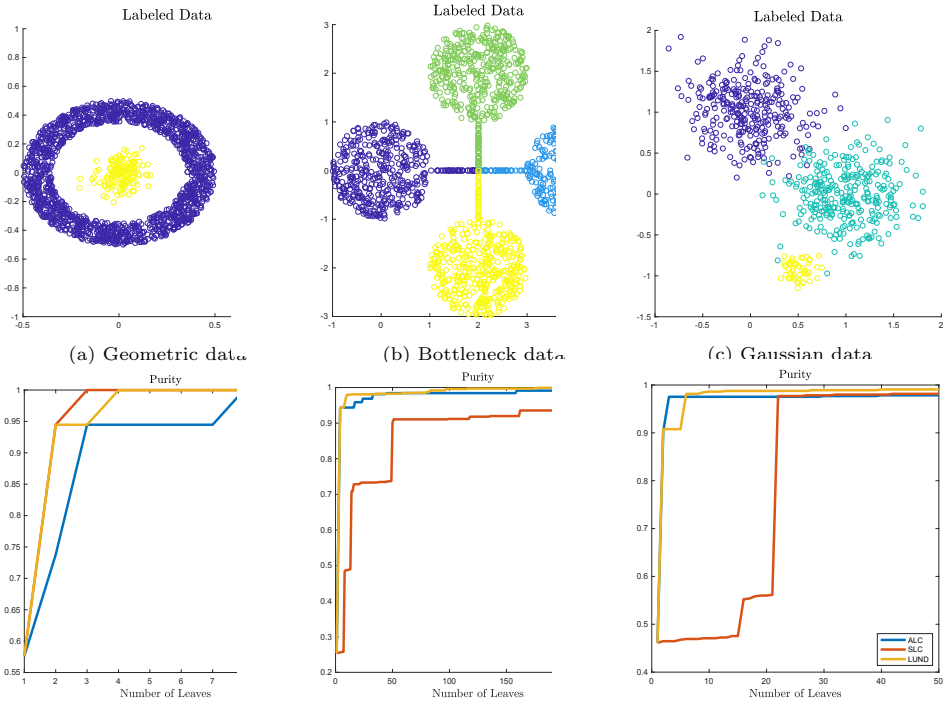
- Y (Labels)

Compute $\mathcal{D}_t(x_i) = p(x_i)\rho_t(x_i)$.
Sort the data in order of decreasing $\mathcal{D}_t$ value to acquire the ordering $\{x_{m_i}\}_{i=1}^n$.
for $k = 1 : K$ **do**
  $Y(x_{m_k}) = k$.
end for
Sort X according to $p(x)$ in decreasing order as $\{x_{\ell_i}\}_{i=1}^n$.
for $i = 1 : n$ **do**
  if $Y(x_{\ell_i}) = 0$ **then**
    $Y(x_{\ell_i}) = Y(z_i)$, $z_i = \arg\min_z \{D_t(z, x_{\ell_i}) \mid p(z) > p(x_{\ell_i}) \text{ and } Y(z) > 0\}$.
  end if
end for

the elongated cluster are quite far apart when measured with the average linkage metric, but are much closer when diffusion distances are used. On the other hand, the bottleneck and Gaussian data illustrate how single linkage clusters may take a long time to achieve high purity, due to the fact that single linkage clustering is guided only by density, and is not robust to adversarial paths of points connecting two otherwise well-separated clusters. Compared to single linkage and average linkage clustering, the clusters learned by LUND are robust to geometric distortions, adversarial paths, and noise.

3.3.2. *Comparison with LUND.* The proposed LAND algorithm (Algorithm 2) integrates an active learning criterion into the LUND algorithm (Algorithm 1). It has been shown that when the the classes of the data X are sufficiently coherent and pairwise well-separated, LUND with a good choice of t perfectly labels all data points [40]. The unsupervised LUND algorithm depends critically on t, and the robustness of LUND to this choice of parameter suggests its usefulness. However, developing practical methods for estimating a good choice of t may be challenging in data that admits hierarchical cluster structure. Indeed, consider the data in Figure 4. For this data, it is ambiguous whether there are two or four clusters. Indeed, as shown in Figure 4 (c), if $\log_{10}(t) \in [1, 3]$, LUND estimates there are 4 clusters. If the time parameter satisfies $\log_{10}(t) \in [4, 6]$, LUND estimates there are 2 clusters. This is a fundamental ambiguity in unsupervised clustering, and one can view the ability of hierarchical clustering algorithms, and of LUND (depending on the time scale t) to detect the different possibilities for the number of clusters as a strength. Partial supervision allows for disambiguation in these situations.

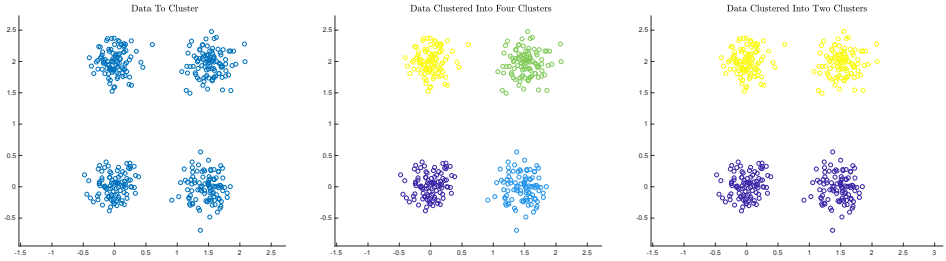
Indeed, with a very small number (4) of labeled queries, LAND is able to overcome this obstacle and determine the labels of the data. This is because even for



(a) Geometric data (b) Bottleneck data (c) Gaussian data

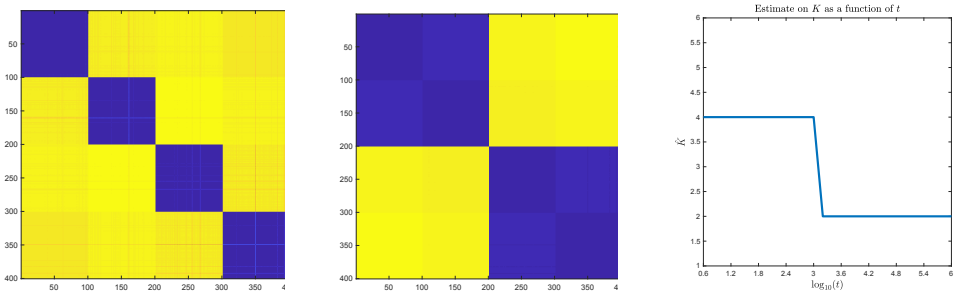
(d) Hierarchical purity of geometric data (e) Hierarchical purity of bottleneck data (f) Hierarchical purity of Gaussian data

Figure 3. *Top row*: Three different synthetic datasets in two dimensions are shown, categorized as geometric, bottleneck and Gaussian. *Bottom row:* Plots of node purity for three different multiscale, hierarchical methods of clustering: average linkage clustering (ALC), single linkage clustering (SLC) and learning by unsupervised nonlinear diffusion (LUND). As the number of leaves/clusters increases, purity is non-decreasing. The purity of the LUND clusters converges more rapidly to the optimal value 1, indicating that high accuracy can be gained by correctly labeling a smaller number of clusters, compared to ALC and SLC.



(a) Hierarchical clustered data (b) Labels with four classes (c) Labels with two classes

Figure 4. In (a), data with natural hierarchical structure is exhibited. The four Gaussians have means $(0,0),(0,2),\left(\frac{3}{2},0\right),\left(\frac{3}{2},2\right)$. While at one level of granularity there are 4 clusters (shown in (b)), at a coarser level of granularity the top 2 and bottom 2 Gaussians form clusters, leading to a clustering with only 2 clusters (shown in (c)).



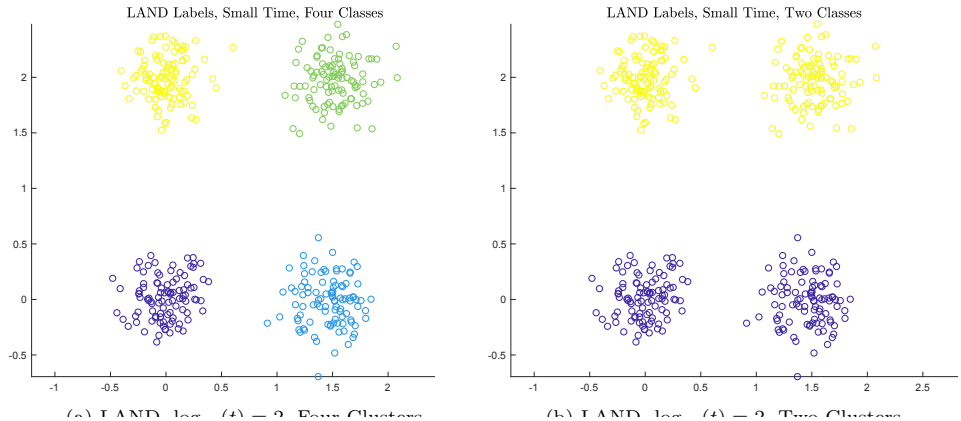
(a) Diffusion distance matrix, $\log_{10}(t) = 1.5$ (b) Diffusion distance matrix, $\log_{10}(t) = 5$ (c) Estimates on K

Figure 5. The matrix of pairwise diffusion distances for $\log_{10}(t) = 1.5$ and $\log_{10}(t) = 5$ are shown in (a) and (b), respectively, illustrating the hierarchical cluster structure in the data. This hierarchical structure introduces ambiguities into the estimation of the number of clusters K in LUND, as shown (c). For small time, $\hat{K} = 4$, while for larger time $\hat{K} = 2$.

large t, the top four values of $\mathcal{D}_t$ correspond to the modes of the four Gaussian clusters, and the diffusion distances within these clusters are quite small. In the unsupervised case, for large t, the gap between the within-cluster and between-cluster distances for the four clusters are dwarfed by the the gap between the within-cluster and between-cluster distances for the two clusters, leading to ambiguity. That is, when the underlying data is grouped into 2 clusters, $D_t^{\text{in}}/D_t^{\text{btw}}$ is large for large t and small for small t; when the underlying data is grouped into 4 clusters, $D_t^{\text{in}}/D_t^{\text{btw}}$ is large for small t and small for large t. These lead to inherent ambiguity in how to choose t in a fully unsupervised manner. However, by bringing in just 4 labels, LAND is able to correctly label the dataset for both large and small t values, as can be seen from Figure 6. In this sense, LAND introduces *robustness to the time parameter* that may be problematic in LUND, at the cost of querying a small number of points.

3.3.3. *Comparisons with graph subsampling methods.* The LAND algorithm maybe understood as a method of acquiring a subsample $\tilde{X}$ from the full data set X, so that accurate labels on X may be inferred from $\tilde{X}$ alone. Indeed, the maximizers of the function $\mathcal{D}_t : X \to [0, \infty)$ determine $\tilde{X}$, whose labels are then propagated to all of X using diffusion distances. In this sense, LAND bears similarity to graph subsampling algorithms [56, 17]. In particular, a range of approaches for sampling smooth (bandlimited) functions on graphs have been proposed [48, 49, 17, 2, 4, 3], including those based on adaptive sampling driven by the localization properties of low-frequency Laplacian eigenfunctions [50, 51].

Compared to these methods, LAND explicitly incorporates density into the sampling procedure and does not rely solely on the spectral properties of the underlying graph Laplacian (or random walk matrix P). In addition, the time parameter t for diffusion distances parametrizes multiscale structure in LAND. Note that t is implicitly related to the smoothness of the labeling functions LAND can learn: as t increases, LAND will only be able to learn labeling distributions that are increasingly smooth with respect to the underlying graph. This is because high-frequency eigenfunctions contribute negligibly to diffusion distances for large t. Moreover, the sampling procedure in LAND is deterministic, while many state-of-the-art graph



(c) LAND, $\log_{10}(t) = 5$, Four Clusters

(d) LAND, $\log_{10}(t) = 5$, Two Clusters

Figure 6. LAND labelings of the data with four queries under two different scenarios: small diffusion time (top row) and large diffusion time (bottom row), and four latent clusters (first column) and two latent clusters (second column). The clusters are closer in the horizontal direction than in the vertical direction, from whence the hierarchical structure is derived. In all cases, LAND is able to to correctly label the data with just four queries, one for each Gaussian.

subsampling methods are random with a sampling distribution that is non-uniform and biased in favor of points on which low-frequency Laplacian eigenfunctions localize. An interesting topic of future work is to consider a *randomized* version of LAND in which the sampling procedure is not deterministic, but random with nonuniform sampling distribution proportional to $1/\mathcal{D}_t(x)$.

3.4. **Computational complexity and implementation.** The proposed Algorithm 2 has computational complexity depending crucially on the number of data points to label (n), the ambient dimensionality of the data (D), and the intrinsic dimensionality of the data (d).

Theorem 3.2. *Let $\{x_i\}_{i=1}^n \subset \mathbb{R}^D$ be data to label. Suppose all except for $O(\log(n))$ points have a higher density point within its $O(\log(n))$ D_t-nearest neighbors. In the case that a k_{NN}-sparse matrix P is used, the LAND algorithm has complexity $O(C_{NN} + nk_{NN} + n\log(n)))$, where C_{NN} is the cost of computing all k_{NN} nearest neighbors.*

Proof. The construction of the Markov transition matrix P has complexity $O(C_{\mathrm{NN}})$. The subsequent kernel density estimation for all points is then $O(nk_{\mathrm{NN}})$. The computation of ρ_t for all points is $O(n\log(n))$, where we assume that all except for $O(\log(n))$ points have a higher density point within their $O(\log(n))$ D_t-nearest neighbors. To estimate the modes from $\mathcal{D}_t$ requires sorting n values, so has complexity $O(n\log(n))$. Once the modes are estimated, labeling all points has complexity $O(n\log(n))$ by the assumption that all except for $O(\log(n))$ points has a higher density point within its $O(\log(n))$ D_t-nearest neighbors. The result follows. □

In the worst case, $C_{\mathrm{NN}} = n^2$, so that LAND has quadratic complexity in n. When the data has intrinsically low-dimensional structure, fast nearest neighbor searches reduce this complexity to be quasilinear in n.

Corollary 3.1. *Let $\{x_i\}_{i=1}^n \subset \mathbb{R}^D$ be data to label. When the underlying data is intrinsically d-dimensional structure (in the sense of doubling dimension) and when $k_{NN} \ll \log(n)$, LAND has computational complexity $O(DC^d n\log(n)^2)$.*

Proof. In the case that the data has intrinsically low-dimensional structure in the sense of doubling dimension, the cover tree indexing structure [10] may be used so that to compute each points k_{NN} has complexity $O(DC^d k_{\mathrm{NN}} n\log(n))$. The result follows. □

Corollary 3.1 suggests that the proposed algorithm is appropriate for large numbers of data points n in high dimension, provided that the intrinsic dimensionality of the data is small.

4. Experimental analysis.

We perform experiments on three representative synthetic datasets, as well as two real hyperspectral images[1]. Comparisons are made between LAND and two related methods:

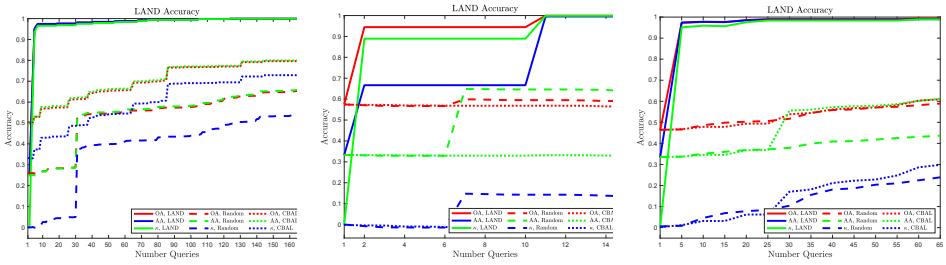
1. *LAND with random query points.* This algorithm consists of Algorithm 2, but with random points selected for querying, rather than the maximizers of $\mathcal{D}_t$. Comparison with LAND will suggest if the query points determined by diffusion geometry and density—as captured by $\mathcal{D}_t$—are actually of significant value.
2. *Cluster-based active learning (CBAL).* This algorithm [25] is implemented using a hierarchical tree constructed from average linkage clustering.

Three performance metrics are used to compare the active learning results. *Overall accuracy (OA)* is the ratio of correctly labeled pixels to the total number of pixels. *Average accuracy (AA)* averages the OA of each class, equalizing the significance of small and large classes. *Cohen's κ-statistic* (κ) is a measure of agreement between two labelings that is robust to random chance [20].

4.1. Experiments on synthetic data.

Experimental results on the three synthetic datasets introduced in Figure 3 are shown in Figure 7, illustrating the efficacy of LAND. In all cases, LAND achieves near perfect accuracy with fewer than 10 labels, while the comparison methods converge to high accuracy much more gradually.

[1] http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes



(a) Bottleneck data results (b) Geometric data results (c) Gaussian data results

Figure 7. Experimental results on the synthetic datasets introduced in Figure 3. We see that LAND achieves rapid convergence to perfect labeling accuracy, compared to much slower convergence for the two comparison methods.

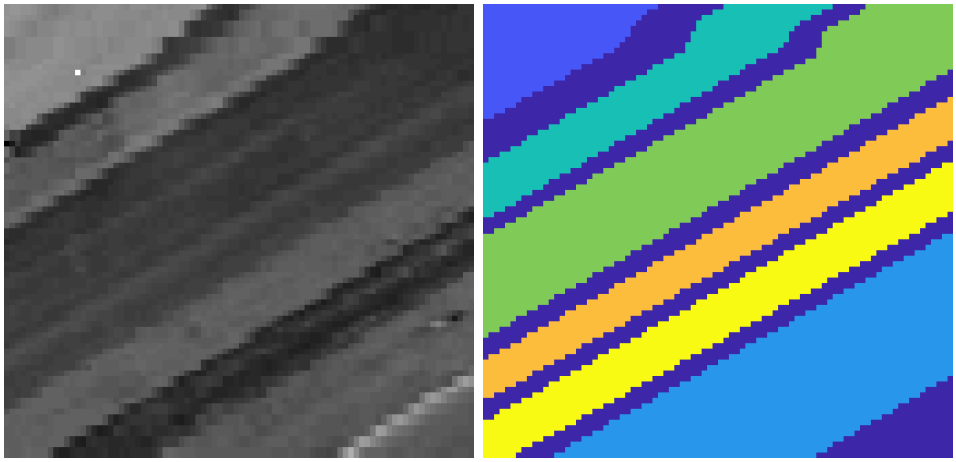
4.2. **Experiments on hyperspectral data.** In order to illustrate the efficacy of LAND on real data, we demonstrate its performance on *hyperspectral imagery (HSI)*, which constitutes an important data type in the remote sensing of the environment [15]. An HSI is an image consisting of D spectral bands, each localized to a narrow electromagnetic range. The concatenation of these D spectral bands provides highly detailed information about the materials being imaged, and can allow for precise discrimination of specific objects in the scene. While nominally a 3-dimensional tensor, an HSI is often analyzed by collapsing the spatial coordinates to produce a dataset $\{x_i\}_{i=1}^n \subset \mathbb{R}^D$, where n is the total number of pixels in the image and D is the total number of spectral bands. When large training sets of labeled pixels are available, classification of an HSI scene may be effectively performed using a range of techniques, including support vector machines [41], deep learning [18], and random forests [33].

Traditional supervised learning has led to strong empirical performance for HSI classification. However, supervised learning for HSI—particularly state-of-the-art deep learning—is predicated on the availability of large labeled training sets, which must be collected and annotated, typically by human experts. The need for large training sets is exacerbated by the high-dimensionality of the data. The collection of large training sets may not be practical in the context of HSI, where there is a huge number of possible classes and large variabilities are introduced by sensing conditions. Indeed, the task of generating huge training sets for general HSI is quite onerous, and may even require the deployment of humans to observe physically the scene that has been remotely sensed, which is very resource intensive. It is thus crucial to develop methods that can label HSI with no labeled training data [1, 47, 11, 13, 32, 64, 19, 66, 63, 45, 44] or a combination of labeled and unlabeled data [12, 53, 38].

Active learning for HSI is an important method for achieving high-accuracy classification results, without requiring large labeled training sets [52, 60, 37, 58, 65, 43]. These methods typically query for labels points near the boundaries of classes, thus improving the convergence of the learning algorithm towards a good classifier. LAND, on the other hand, exploits cluster structure in the data.

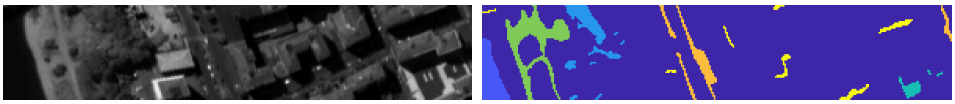
4.2.1. *Experimental results for HSI.* We perform active learning experiments on two real HSI datasets, shown in Figure 8 and 9, respectively.

Experimental results for the three methods on the Salinas A and Pavia datasets are shown in Figure 10. For the Salinas A dataset, accuracy with LAND is strong,



(a) Salinas A (b) Ground Truth

Figure 8. The Salinas A dataset consists of $83 \times 86 = 7138$ pixels in $D = 224$ dimensions. The image has spatial resolution 3.7m/pixel, and was recorded over Salinas, USA by the Aviris sensor. The six labelled classes are arranged in diagonal rows, and are quite spatially regular. The sum across all spectral bands is shown in (a), and the labeled ground truth is shown in (b), with pixels having the same class being given the same color.



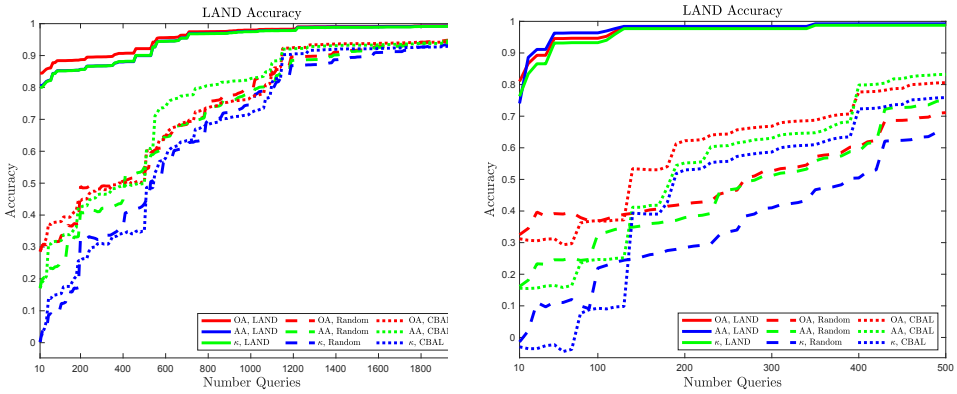
(a) Pavia (b) Ground Truth

Figure 9. Pavia data consists of a $270 \times 50 = 13500$ subset of the full Pavia data set. The image has spatial resolution 1.3m/pixel, and was recorded over Pavia, Italy by the ROSIS sensor. It consists of 6 spatial classes, some of which are quite well-spread out in the image. The sum across all spectral bands is shown in (a), and the labeled ground truth is shown in (b), with pixels having the same class being given the same color.

with only 10 labels leading to highly accurate empirical results, and subsequent labels leading to rapid improvement towards perfect accuracy. In particular, compared to using random query labels or CBAL, the improvement of LAND as a function of the number of queries is fast. For the Pavia dataset, there is a similar early jump in accuracy for LAND, while the improvement is slower for the comparison methods.

5. **Conclusions and future work.** The LAND algorithm integrates diffusion geometry and density estimation to efficiently estimate query points that are highly impactful on overall labeling accuracy in the active learning setting. Our theoretical and empirical analyses show LAND's robustness to geometric distortions of the underlying data classes, and our experiments on real-world HSI demonstrate its effectiveness in accurately labeling high-dimensional datasets with a very small number of query points.

In the context of HSI, developing active learning methods that incorporate spatial proximity into the underlying diffusion process is of interest. This information may



(a) Salinas A results

(b) Pavia results

Figure 10. The active learning results for the Salinas A and Pavia datasets are shown in (a) and (b), respectively. In both cases, the LAND algorithm strongly outperforms the modified LAND variant using randomly selected training data, and the CBAL algorithm. In particular, LAND is able to achieve a significant improvement in accuracy with a very small number of labels.

suggest that it is useful to query information in a spatially homogeneous region, where it can be most impactful. The integration of spatial information into a variant of the LUND algorithm adapted for HSI has proven effective [43, 44], and it is likely that such information would similarly boost the effectiveness of LAND.

It is of interest to develop a cross-validation scheme that exploits the active learning queries in order to iteratively update the optimal choice of time parameter t. Indeed, as argued in Section 3.3.2, the use of a very small (essentially $O(K)$) active learning queries can be used to achieve robustness to the parameter t, which is critically important in the LUND algorithm. However, it may be possible to update the time parameter in an iterative fashion, by selecting at each time step a time scale that separates all the modes learned so far, before querying a new point. This has the potential to require fewer queries to learn all the classes, since the parameter is being adaptively optimized at each time step, rather than after all queries have been made.

Acknowledgements. We are grateful to the anonymous reviewer for many helpful comments and suggestions which significantly improved the manuscript.

REFERENCES

[1] N. Acito, G. Corsini and M. Diani, An unsupervised algorithm for hyperspectral image segmentation based on the gaussian mixture model, in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, **6** (2003), 3745–3747.

[2] A. Anis, A. Gadde and A. Ortega, Towards a sampling theorem for signals on arbitrary graphs, in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2014, 3864–3868.

[3] A. Anis, A. Gadde and A. Ortega, Efficient sampling set selection for bandlimited graph signals using graph spectral proxies, *IEEE Transactions on Signal Processing*, **64** (2016), 3775–3789.

[4] A. Anis, A. E. Gamal, S. Avestimehr and A. Ortega, Asymptotic justification of bandlimited interpolation of graph signals for semi-supervised learning, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2015, 5461–5465.

[5] E. Arias-Castro, Clustering based on pairwise distances when the data is of mixed dimensions, *IEEE Transactions on Information Theory*, **57** (2011), 1692–1706.
[6] E. Arias-Castro, G. Lerman and T. Zhang, Spectral clustering based on local PCA, *Journal of Machine Learning Research*, **18** (2017), 1–57.
[7] F. Aurenhammer, Voronoi diagrams—a survey of a fundamental geometric data structure, *ACM Computing Surveys (CSUR)*, **23** (1991), 345–405.
[8] M.-F. Balcan, A. Beygelzimer and J. Langford, Agnostic active learning, *Journal of Computer and System Sciences*, **75** (2009), 78–89.
[9] M.-F. Balcan, A. Broder and T. Zhang, Margin based active learning, in *International Conference on Computational Learning Theory*, Springer, **4359** (2007), 35–50.
[10] A. Beygelzimer, S. Kakade and J. Langford, Cover trees for nearest neighbor, in *Proceedings of the 23rd International Conference on Machine Learning*, ACM, 2006, 97–104.
[11] N. Cahill, W. Czaja and D. Messinger, Schroedinger eigenmaps with nondiagonal potentials for spatial-spectral clustering of hyperspectral imagery, in *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XX*, vol. 9088, International Society for Optics and Photonics, 2014, 908804.
[12] G. Camps-Valls, T. Marsheva and D. Zhou, Semi-supervised graph-based hyperspectral image classification, *IEEE Transactions on Geoscience and Remote Sensing*, **45** (2007), 3044–3054.
[13] C. Cariou and K. Chehdi, Unsupervised nearest neighbors clustering with application to hyperspectral images, *IEEE Journal of Selected Topics in Signal Processing*, **9** (2015), 1105–1116.
[14] R. Castro and R. Nowak, Minimax bounds for active learning, *IEEE Transactions on Information Theory*, **54** (2008), 2339–2353.
[15] C.-I. Chang, *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*, vol. 1, Springer Science & Business Media, 2003.
[16] O. Chapelle, B. Scholkopf and A. Zien, *Semi-supervised Learning*, MIT Press, 2006.
[17] S. Chen, R. Varma, A. Sandryhaila and J. Kovačević, Discrete signal processing on graphs: Sampling theory, *IEEE Transactions on Signal Processing*, **63** (2015), 6510–6523.
[18] Y. Chen, Z. Lin, X. Zhao, G. Wang and Y. Gu, Deep learning-based classification of hyperspectral data, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **7** (2014), 2094–2107.
[19] Y. Chen, S. Ma, X. Chen and P. Ghamisi, Hyperspectral data clustering based on density analysis ensemble, *Remote Sensing Letters*, **8** (2017), 194–203.
[20] J. Cohen, A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, **20** (1960), 37–46.
[21] D. Cohn, L. Atlas and R. Ladner, Improving generalization with active learning, *Machine Learning*, **15** (1994), 201–221.
[22] R. Coifman and S. Lafon, Diffusion maps, *Applied and Computational Harmonic Analysis*, **21** (2006), 5–30.
[23] R. Coifman, S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner and S. Zucker, Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps, *Proceedings of the National Academy of Sciences of the United States of America*, **102** (2005), 7426–7431.
[24] S. Dasgupta, Two faces of active learning, *Theoretical Computer Science*, **412** (2011), 1767–1781.
[25] S. Dasgupta and D. Hsu, Hierarchical sampling for active learning, in *Proceedings of the 25th International Conference on Machine Learning*, ACM, 2008, 208–215.
[26] S. Dasgupta, D. Hsu and C. Monteleoni, A general agnostic active learning algorithm, in *Advances in neural information processing systems*, 2008, 353–360.
[27] A. Esteva, B. Kuprel, R. Novoa, J. Ko, S. Swetter, H. Blau and S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, *Nature*, **542** (2017), 115–118.
[28] J. Friedman, T. Hastie and R. Tibshirani, *The Elements of Statistical Learning*, vol. 1, Springer series in Statistics Springer, Berlin, 2001.
[29] N. Garcia Trillos, M. Gerlach, M. Hein and D. Slepcev, Error estimates for spectral convergence of the graph Laplacian on random geometric graphs towards the Laplace–Beltrami operator, arXiv:1801.10108.
[30] N. Garcia Trillos, F. Hoffmann and B. Hosseini, Geometric structure of graph Laplacian embeddings, arXiv:1901.10651.

[31] M. Gavish and B. Nadler, Normalized cuts are approximately inverse exit times, *SIAM Journal on Matrix Analysis and Applications*, **34** (2013), 757–772.
[32] N. Gillis, D. Kuang and H. Park, Hierarchical clustering of hyperspectral images using rank-two nonnegative matrix factorization, *IEEE Transactions on Geoscience and Remote Sensing*, **53** (2015), 2066–2078.
[33] J. Ham, Y. Chen, M. Crawford and J. Ghosh, Investigation of the random forest framework for classification of hyperspectral data, *IEEE Transactions on Geoscience and Remote Sensing*, **43** (2005), 492–501.
[34] S. Hanneke, Rates of convergence in active learning, *The Annals of Statistics*, **39** (2011), 333–361.
[35] A. Krizhevsky, I. Sutskever and G. Hinton, Imagenet classification with deep convolutional neural networks, *Communications of the ACM*, **60** (2017), 84–90.
[36] S. Lafon and A. Lee, Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28** (2006), 1393–1403.
[37] J. Li, J. Bioucas-Dias and A. Plaza, Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning, *IEEE Transactions on Geoscience and Remote Sensing*, **48** (2010), 4085–4098.
[38] J. Li, J. Bioucas-Dias and A. Plaza, Semisupervised hyperspectral image classification using soft sparse multinomial logistic regression, *IEEE Geoscience and Remote Sensing Letters*, **10** (2013), 318–322.
[39] A. Little, M. Maggioni and J. Murphy, Path-based spectral clustering: Guarantees, robustness to outliers, and fast algorithms, arXiv:1712.06206.
[40] M. Maggioni and J. Murphy, Learning by unsupervised nonlinear diffusion, arXiv:1810.06702.
[41] F. Melgani and L. Bruzzone, Classification of hyperspectral remote sensing images with support vector machines, *IEEE Transactions on geoscience and remote sensing*, **42** (2004), 1778–1790.
[42] D. Mixon, S. Villar and R. Ward, Clustering subgaussian mixtures by semidefinite programming, *Information and Inference: A Journal of the IMA*, **6** (2017), 389–415.
[43] J. Murphy and M. Maggioni, Iterative active learning with diffusion geometry for hyperspectral images, in *9th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, IEEE, 2018, 1–5.
[44] J. Murphy and M. Maggioni, Spectral-spatial diffusion geometry for hyperspectral image clustering, arXiv:1902.05402.
[45] J. Murphy and M. Maggioni, Unsupervised clustering and active learning of hyperspectral images with nonlinear diffusion, *IEEE Transactions on Geoscience and Remote Sensing*, **57** (2019), 1829–1845.
[46] B. Nadler and M. Galun, Fundamental limitations of spectral clustering, in *Advances in Neural Information Processing Systems*, 2007, 1017–1024.
[47] A. Paoli, F. Melgani and E. Pasolli, Clustering of hyperspectral images based on multiobjective particle swarm optimization, *IEEE Transactions on Geoscience and Remote Sensing*, **47** (2009), 4175–4188.
[48] I. Pesenson, Sampling in paley-wiener spaces on combinatorial graphs, *Transactions of the American Mathematical Society*, **360** (2008), 5603–5627.
[49] I. Pesenson and M. Pesenson, Sampling, filtering and sparse approximations on combinatorial graphs, *Journal of Fourier Analysis and Applications*, **16** (2010), 921–942.
[50] G. Puy and P. Pérez, Structured sampling and fast reconstruction of smooth graph signals, *Information and Inference: A Journal of the IMA*, **7** (2018), 657–688.
[51] G. Puy, N. Tremblay, R. Gribonval and P. Vandergheynst, Random sampling of bandlimited signals on graphs, *Applied and Computational Harmonic Analysis*, **44** (2018), 446–475.
[52] S. Rajan, J. Ghosh and M. Crawford, An active learning approach to hyperspectral data classification, *IEEE Transactions on Geoscience and Remote Sensing*, **46** (2008), 1231–1242.
[53] F. Ratle, G. Camps-Valls and J. Weston, Semisupervised neural networks for efficient hyperspectral image classification, *IEEE Transactions on Geoscience and Remote Sensing*, **48** (2010), 2271–2282.
[54] G. Schiebinger, M. Wainwright and B. Yu, The geometry of kernelized spectral clustering, *The Annals of Statistics*, **43** (2015), 819–846.
[55] B. Settles, *Active Learning Literature Survey*, Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[56] D. Shuman, S. Narang, P. Frossard, A. Ortega and P. Vandergheynst, The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains, *IEEE Signal Processing Magazine*, **3** (2013), 83–98.

[57] D. Silver, A. Huang, C. Maddison, A. Guez, L. Sifre, G. V. D. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam and M. Lanctot, Mastering the game of go with deep neural networks and tree search, *nature*, **529** (2016), 484–489.

[58] S. Sun, P. Zhong, H. Xiao and R. Wang, Active learning with gaussian process classifier for hyperspectral image classification, *IEEE Transactions on Geoscience and Remote Sensing*, **53** (2015), 1746–1760.

[59] M. Tanner and W. Wong, The calculation of posterior distributions by data augmentation, *Journal of the American statistical Association*, **82** (1987), 528–540.

[60] D. Tuia, F. Ratle, F. Pacifici, M. Kanevski and W. Emery, Active learning methods for remote sensing image classification, *IEEE Transactions on Geoscience and Remote Sensing*, **47** (2009), 2218–2232.

[61] R. Urner, S. Wulff and S. Ben-David, PLAL cluster-based active learning, in *Conference on Learning Theory*, 2013, 376–397.

[62] D. Van Dyk and X.-L. Meng, The art of data augmentation, *Journal of Computational and Graphical Statistics*, **10** (2001), 1–111.

[63] H. Zhai, H. Zhang, L. Zhang, P. Li and A. Plaza, A new sparse subspace clustering algorithm for hyperspectral remote sensing imagery, *IEEE Geoscience and Remote Sensing Letters*, **14** (2017), 43–47.

[64] H. Zhang, H. Zhai and L. Z. P. Li, Spectral–spatial sparse subspace clustering for hyperspectral remote sensing images, *IEEE Transactions on Geoscience and Remote Sensing*, **54** (2016), 3672–3684.

[65] Z. Zhang, E. Pasolli, M. Crawford and J. Tilton, An active learning framework for hyperspectral image classification using hierarchical segmentation, *IEEE J-STARS*, **9** (2016), 640–654.

[66] W. Zhu, V. Chayes, A. Tiard, S. Sanchez, D. Dahlberg, A. Bertozzi, S. Osher, D. Zosso and D. Kuang, Unsupervised classification in hyperspectral imagery with nonlocal total variation and primal-dual hybrid gradient algorithm, *IEEE Transactions on Geoscience and Remote Sensing*, **55** (2017), 2786–2798.

E-mail address: mauro.maggioni@jhu.edu
E-mail address: jm.murphy@tufts.edu