Cautious Active Clustering

A. Cloninger, H. N. Mhaskar

Abstract

We consider a set of points sampled from an unknown probability measure on a Euclidean space, each of which points belongs to one of the finitely many classes. We study the question of querying the class label at a very small number of judiciously chosen points so as to be able to attach the appropriate class label to every point in the set. Our approach is to consider the unknown probability measure as a convex combination of the conditional probabilities for each class. Our technique involves the use of a highly localized kernel constructed from Hermite polynomials, and use them to create a hierarchical estimate of the supports of the constituent probability measures. We do not need to make any assumptions on the nature of any of the probability measures nor know in advance the number of classes involved. We give theoretical guarantees measured by the $F$-score for our classification scheme. Examples include classification in hyper-spectral images, separation of distributions, and MNIST classification.

1 Introduction

In this paper, we demonstrate how certain ideas originating in our previous work on blind source signal separation can be modified and applied to a problem in machine learning, which we have termed cautious active learning in this paper. In Section 1.1, we describe the problem in the theory of machine learning. In Section 1.2, we describe briefly our work on blind source signal separation that motivates our current paper, and provides a prototype for the results in this paper. Section 1.3 explains the difficulties involved in adapting the approach in Section 1.2 and give a preview of the kind of results expected with our solution presented in this paper. In Section 1.4, we discuss connections with a few other works related to the problem and our solution to the same. The outline of the paper is given in Section 1.5. For the convenience of exposition, the notation used in this section is not the same as the one used in the rest of this paper after Section 2.

1.1 Cautious active learning

An important task of machine learning is to classify various objects into a finite number of classes. Typically, this task is formulated as follows. We are given data of the form $\{(x_i, y_i)\}_{i=1}^M$ where $x_i$’s are in some Euclidean space $\mathbb{R}^q$, and $y_i \in \{1, \cdots, K\}$ for some integer $K \geq 1$. In supervised learning, we have to build a model $P$ such that for any vector $x \in \mathbb{R}^q$ (or a compact subset thereof), $P(x)$ gives reliably the class to which $x$ belongs. In semi-supervised learning, the labels $y_i$ are known only for a small number of $x_i$’s, and the problem is to extend this labeling to the rest of the data set. It is assumed that the data set is known in advance; it is not expected to build a model for points not in the original data set. In unsupervised learning, no information is known about the labels, and the best that can be done is to find the right clusters in the dataset.

Active learning is a relatively recent area of machine learning that combines aspects of all of three paradigms above. We do not know any labels to begin with, but are allowed to seek labels on judiciously chosen points $x_i$, as few as needed to construct a model $P$ as in the case of supervised learning. Clearly, this must be done in the beginning using clustering as in unsupervised learning, based on some model. We then “purify” this clustering using a small number of queries for the label. In the end, we have started in the unsupervised regime, and then collected a small number of labeled data as in the semi-supervised regime, and finally built a model as in the supervised regime. However, in semi-supervised learning, we cannot control the set of points at which the label is known, and

\*Department of Mathematics and Halicioglu Data Science Institute, University of California San Diego, San Diego, CA 92093, U.S.A.. The research of this author is supported in part by the National Science Foundation project DMS-1818945. email: acloninger@ucsd.edu

\†Institute of Mathematical Sciences, Claremont Graduate University, Claremont, CA 91711, U.S.A.. email: hrushikesh.mhaskar@cgu.edu
we do not expect a model for the points not in the original data set. In contrast, in active learning, we get to choose which points to query the label at, and a model is expected as the end-product.

It is customary to assume that the data is drawn from an unknown probability distribution. Obviously,

\[ \text{Prob}(x, k) = \text{Prob}(k|x) \text{Prob}(x) = \text{Prob}(x|k) \text{Prob}(k). \]  

(1.1)

The first equation leads us to discriminative models. The class \( k \) of a given point \( x \) is

\[ \arg \max_{k=1,\ldots,K} \text{Prob}(k|x) \text{Prob}(x). \]

In [34], we have explored this approach in further detail, giving theoretically well founded criteria to determine how to estimate the class reliably.

In this paper, we take a closer look at the second equation in (1.1); i.e., use the fact that

\[ \text{Prob}(x) = \sum_{k=1}^{K} \text{Prob}(x|k) \text{Prob}(k). \]  

(1.2)

In measure theoretic notation, one can write

\[ \mu^* = \sum_{k=1}^{K} \mu_k, \]

(1.3)

where \( \mu^* \) is the marginal distribution from which the points \( x \) are chosen, and \( \mu_k \) represents the \( k \)-th term in (1.2), i.e., some positive measure. The task is to separate the supports of the component measures \( \mu_k \) given random samples taken from the probability law \( \mu^* \), which is not known. The intuition is that once the supports of each \( \mu_k \) is known, we just need one sample from each to complete the task of classification.

Because of overlapping class boundaries, it is not reasonable to assume that the classes; i.e., the supports of the constituent measures, are well separated. In this paper, we propose a hierarchical classification scheme, where the minimal separation among the supports of \( \mu_k \)'s is decreased step by step. The accuracy of our hierarchical clustering schemes is proved using the classical F-score as the measurement of quality of clustering.

1.2 Motivation for our approach

Let \( q \geq 1 \) be an integer, \( \mathbb{T}^{q} = \mathbb{R}^{q}/(2\pi \mathbb{Z}^{q}) \). For \( x, y \in \mathbb{T}^{q} \), we define (in this section only) \( |x - y| = \max_{1 \leq k \leq q} |(x_k - y_k) \mod 2\pi| \). One formulation of the problem of blind source signal separation is the following. Let \( \mu^* = \sum_{k=1}^{K} a_k \delta_{x_k} \) be a (signed) measure supported at points \( x_k \in \mathbb{T}^{q} \), where \( \delta_{x_k} \) is the Dirac delta measure supported at \( x \). We are given the Fourier moments \( \hat{\mu}^*(j) = \sum_{k=1}^{K} a_k \exp(-ij \cdot x_k) \) for \( |j|_{\infty} < N \) for some \( N \), and wish to recuperate the number \( K \) of components, the point sources \( x_k \) and the (signed, complex) amplitudes \( a_k \). Our solution to this problem described in [31, 29, 35] is the following. We consider a filter \( H : \mathbb{R} \to [0, 1] \) that is an infinitely differentiable, even function, with \( H(t) = 0 \) for \( |t| \geq 1 \). We then consider an operator

\[ \mathcal{T}_n(f)(x) = h_n \sum_{j \in \mathbb{Z}^{q}} H \left( \frac{|j|}{n} \right) f(j) \exp(ij \cdot x), \quad x \in \mathbb{T}^{q}, \]

where

\[ h_n = \left( \sum_{j \in \mathbb{Z}^{q}} H \left( \frac{|j|}{n} \right) \right)^{-1}. \]

With

\[ \Phi_n^T(x - y) = \sum_{j \in \mathbb{Z}^{q}} H \left( \frac{|j|}{n} \right) \exp(ij \cdot (x - y)), \]

it is not difficult to verify that

\[ \mathcal{T}_n(f)(x) = \frac{h_n}{(2\pi)^q} \int_{\mathbb{T}^{q}} \Phi_n^T(x - y) d\mu^*(y). \]  

(1.4)

The following theorem from [3] serves as a precursor of our research described in this paper, where we use the notation \( \eta \) to denote the minimal separation among the points \( x_k \), and \( m \) to denote the minimum of the \(|a_k|\)'s.
Theorem 1.1 For sufficiently large $n$ (depending upon $\eta$), the set of $x \in \mathbb{T}^q$ at which $|T_n(f(x))| \geq m/2$ is a disjoint union of exactly $K$ sets $G_k$, $1 \leq k \leq K$, each containing exactly one point $x_k$, and each with diameter $\leq c/n$ for some positive constant $c$ with $c, n$ for some positive constant $c$ with $c/n$.

The key ingredient in the proof of Theorem 1.1 is the localization estimate
\[
|\Phi_n^T(x-y)| \leq \frac{c(H,S)}{\max(1, (n|x-y|^2)^{\frac{1}{2}})}, \quad x, y \in \mathbb{T}^q,
\]
where $c(H,S) > 0$ is a constant independent of $n, x, y$.

1.3 Separation of measures

The basic idea in our paper is to use an analogous localized kernel $\Phi_{n,q}$ to be defined in (2.14) below (cf. [33] [5]) based on Hermite polynomials. It is not difficult to verify using known results about these kernels that $\int_{\mathbb{R}^q} \Phi_{n,q}(x,y) d\mu^*(y) \to d\mu^*(x)$ in a weak-star sense, and the rate of approximation is optimal in the case when $\mu^*$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^q$ with a smooth density. Therefore, it is reasonable to expect that
\[
\int_{\mathbb{R}^q} \Phi_{n,q}(x,y) d\mu^*(y) = \sum_{k=1}^{K} \int_{\mathbb{R}^q} \Phi_{n,q}(x,y) d\mu_k(y)
\]
will split into clusters of $x$ belonging to the supports of the measures $\mu_k$.

This optimality of approximation of measures however requires that the kernel $\Phi_{n,q}$ is not a positive kernel. When $\mu^*$ is discretely supported, then the localization properties of the kernel ensure that near any one point of the support, the contribution to the integral from other points is negligible. This is not the case when the measure is supported on a continuum. Therefore, the problem of finding the support of $\mu^*$ is different from the problem of finding $\mu^*$ itself. In our paper, we are interested only finding the supports, not the measures themselves. So, we will use the kernel $\Phi_{n,q}^2$ instead.

Apart from this technicality, there are many inherent barriers which makes the problem in our setting similar, yet very different, from the problem of super-resolution as described. We illustrate with an example.

Example 1.1 We consider a mixture of two distributions, 2/3-rd part a uniform distribution on a 2D ball, and 1/3-rd part a uniform distribution on a 1D line, with minimal separation $\delta \in \{0, 0.1, 0.2\}$ between the distributions. In Figure 1, we show the results of our method based on a total of 1000 points, with the value of the parameter $n \in \{3, 7\}$. Our method estimates the relative supports of the two distributions, and is able to maintain the separation between the two distributions even for small minimal separation, and at low degree $n$. This is of note because there is no assumption on the dimension of the support of the distributions, and similarly no assumption on the nature of the constituent distributions.

To summarize, we are interested in extending the theory summarized in Section 1.2 to overcome the following problems in particular.

1. Instead of having a linear combination of Dirac deltas, we have a linear combination of arbitrary probability measures, whose supports may be continuuas. In turn, this requires the coefficients of these constituent distributions to be positive.

2. Instead of having values of $f(j)$, we have random samples chosen from the distribution $\mu^*$. In some sense, this simplifies matters, since we could then discretize the integral in (1.4) directly using the samples.

3. There is no minimal separation anymore. We will replace it by a multiscale notion where we consider the supports of the constituent measures to be separated by $\eta$ for different values of $\eta$, with the remainder having a smaller and smaller probability.

4. There is no analogue of minimal magnitude $m$ here. Although one could pose the problem as the separation of a convex combination of probability measures, and assume a minimum on the coefficients involved, the probability measures themselves may be close to 0 on continuuas.
1.4 Relation to prior work

A main difficulty in the theory of unsupervised learning is to define what one should understand by a cluster. For example, the correct number of clusters is sometimes defined in terms of graph cuts [22, 25]. This definition of clustering is not necessarily intuitive and leads to arbitrary bifurcations in the geometric structure of the data. It is pointed out in [14] that the notion of a cluster (in an unsupervised setting) needs be defined hierarchically. We follow the philosophy in [14] by defining a hierarchical clustering that is tied to the order of our localized kernel $\Phi_n$, with the benefit that $\Phi_n$ provides a smooth decay with known decay rates.

Our paper also ties into the general field of active learning and machine teaching, which has grown rapidly in recent years with a large number of applications. For the sake of relevance, we will focus on the subset of papers with mathematical guarantees for the proposed algorithm and that focus on assumptions on the data geometry [28, 40] rather than low-complexity classifiers [19, 20]. Many of these results either establish lower bounds on the number of labels needed, or establish very conservative criteria of where to query labels in order to avoid sampling bias. A general overview can be found in [38, 41, 26].

A number of works by Dasgupta and his collaborators [11, 2, 12, 13] have examined active learning over a class of hypotheses (i.e. classifiers) for minimax bounds on the generalization error, with probabilistic methods of choosing the points to sample. The errors are in terms of the VC-dimension of the hypothesis class. The closest connection to our work is the paper [12], which assumes that there exists a hierarchical tree on the data structure and samples randomly from various bins.

Our paper also examines active learning problems in the context of hierarchical clusters. The major tool in this research is localized kernels [30], which have been applied in a variety of contexts [35, 32, 31, 27, 9, 4, 8]. Localized kernels also play a central role in the determination of components of a blind source signal, whether stationary or time-variant [5, 27]. The super-resolution problem has been considered in a hierarchical context [24]. Our paper uses the super-resolution aspect of this theory with the harmonic analysis aspects in the context of active learning.

Our paper also connects to the analysis considered in two sample testing [3, 34], where we used the theory of localized kernels and quadrature formulas on Euclidean spaces to determine where two probability distributions deviate from one another. The theory in [34] is used in this paper to determine which distribution is dominant at each point, as well as a measure of uncertainty in the classification. The approach in [34] also helps to extend the theory of the witness function to multi-class classification. Our paper builds on the witness function approach by constructing an indicator of where each cluster dominates while knowing only a few label samples. We also use the witness function to determine the classification of uncertain points (see Section 4.2).

We wish to note in particular [28], which studies the active clustering framework for diffusion distance between
points. This establishes conditions under which the clusters are “well-enough” separated that each cluster will have a smaller in-radius for diffusion distance than the inter-cluster distance. Unlike \[28\], which constructs a single density estimate for the data, our algorithm constructs a hierarchical density estimate and, at each scale, throws away low-density points in order to guarantee well separated clusters. We will compare to \[28\], and to the hyperspectral imaging variant \[36\], in Section 5.

1.5 Outline

We introduce the notation to be used in this paper in Section 2. The main theorems are given in Section 3, and proved in Section 6. We describe our algorithm to implement the main theorem in Section 4 (together with Appendix A) and illustrate the same using several examples in Section 5.

2 Notation and definitions

In this paper, \(q \geq 1\) is a fixed integer. For \(x = (x_1, \cdots, x_q) \in \mathbb{R}^q\), we denote by \(|\cdot|_p\) the \(p\)th norm of \(x\). For \(x \in \mathbb{R}^q\), \(r > 0\), we denote
\[
\mathcal{B}(x, r) = \{ y \in \mathbb{R}^q : |x - y|_\infty \leq r \}. \tag{2.1}
\]
If \(A, B \subseteq \mathbb{R}^q\), \(x \in \mathbb{R}^q\), \(r > 0\), then
\[
\text{dist}(x, A) = \inf_{y \in A} |x - y|_\infty, \quad \mathcal{B}(A, r) = \{ x \in \mathbb{R}^q : \text{dist}(x, A) \leq r \}, \quad \text{dist}(A, B) = \inf_{x \in A} \text{dist}(x, B). \tag{2.2}
\]

The term measure will mean a positive, Borel measure on \(\mathbb{R}^q\). The support of a measure \(\mu\), denoted by \(\text{supp}(\mu)\) is the set of all \(x \in \mathbb{R}^q\) for which \(\mu(\mathcal{B}(x, r)) > 0\) for all \(r > 0\). We will fix a probability measure \(\mu^*\) on \(\mathbb{R}^q\). For \(r_0, A, \alpha > 0\), let
\[
S(r_0; A, \alpha) = \{ x \in \mathbb{R}^q : \mu^*(\mathcal{B}(x, r)) \geq Ar^\alpha, \ 0 < r \leq r_0 \}. \tag{2.3}
\]
Clearly, if \(r_0 < r_1\) then \(S(r_1; A, \alpha) \subseteq S(r_0; A, \alpha) \subseteq \text{supp}^*(\mu)\). The set \(S(r_0; A, \alpha)\) is clearly a set of points where the measure \(\mu^*\) is “large”.

Constant convention

In this paper, the symbols \(c, c_1, \cdots\) will denote generic constants depending only on the fixed quantities under consideration, such as \(q\), \(\mu^*\), and parameters \(H, S\) to be introduced later. Their values may be different at different occurrences, even within a single formula. There are occasions when we need to retain the values of some constants. Those constants whose value depend only on \(q\) and \(H\) will be denoted by \(\kappa, \kappa_1, \cdots\), those whose value depends upon the measure as well will be denoted by \(C, C^*, C_1, \cdots\).

Definition 2.1 Let \(\mu^*\) be a probability measure on \(\mathbb{R}^q\). The measure \(\mu^*\) is called detectable if each of the following conditions is satisfied.

1. (Compact support condition) The support of \(\mu^*\) is compact; in particular,
\[
\text{supp}(\mu^*) \subseteq \{ x \in \mathbb{R}^q : |x|_\infty \leq C \}.
\]

2. (Ball measure condition) There exist \(C_2\) and \(\alpha > 0\) such that
\[
\mu^*(\mathcal{B}(x, r)) \leq C_2r^\alpha. \tag{2.4}
\]

3. (Density condition) There exist \(C_1\) such that (with \(\alpha\) as in (2.4)),
\[
\lim_{r_0 \downarrow 0} \mu^*(\mathbb{R}^q \setminus S(r_0; C_1, \alpha)) = 0. \tag{2.5}
\]

The measure \(\mu^*\) is said to have fine structure if it is detectable, and each of the following conditions is satisfied.
1. There exists $\eta_0 > 0$ such that every $\eta \in (0, \eta_0]$, there is in integer $K_\eta \geq 1$ and a partition $S_{k,\eta}$, $k = 1, \ldots, K_\eta+1$ of $\text{supp}(\mu^*)$ such that

$$S_{k,\eta} \subseteq \text{supp}(\mu_k), \ k = 1, \ldots, K_\eta, \ S_{K_\eta+1,\eta} = \text{supp}(\mu^*) \setminus \bigcup_{k=1}^{K_\eta} S_{k,\eta}. \quad (2.6)$$

2. (Minimal separation condition)

$$\text{dist}(S_{k,\eta}, S_{j,\eta}) \geq 2\eta, \quad k \neq j, \ k, j = 1, \ldots, K_\eta. \quad (2.7)$$

3. (Exhaustion condition)

$$\lim_{\eta \downarrow 0} \mu^*(S_{K_\eta+1,\eta}) = 0.$$ 

Example 2.1 Let $X$ be a compact, smooth, sub-manifold (without boundary) of $\mathbb{R}^d$ with dimension $d$, $\mu^* = f d\mu$, where $\mu$ is the Riemannian volume element of $X$, and $f : X \rightarrow \mathbb{R}$ satisfies $0 < c_1 \leq f(x) \leq c_2$ for $x \in X$. Obviously, $\mu^*$ satisfies the compact support condition. We assume that $\mu^*$ satisfies the ball measure condition with $\alpha = q$. It can then be proved that $\mu^*$ satisfies the density condition with $S(1; C_2, q) = X$. Thus, $\mu^*$ is detectable, and our theorems in Section 3 explain how to find the support of $\mu^*$. If $\mu^* = \sum_{k=1}^{K} \mu_k$ with disjointly supported $\mu_k$’s, then necessarily, $\min_{j \neq k} \text{dist}(\text{supp}(\mu_j), \text{supp}(\mu_k)) \geq \eta_0$ for some $\eta_0 > 0$. Then $\mu^*$ has a fine structure, with $K_\eta = K$ and $S_{K_\eta+1,\eta} = \emptyset$ for every $\eta \in (0, \eta_0]$.

Example 2.2 Let $\mu^* = \sum_{k=1}^{K} a_k \delta_{x_k}$, where $a_k$’s are positive, $\sum_k a_k = 1$, and $x_k \in \mathbb{R}^d$. Then $\mu^*$ is compactly supported, and the ball measure condition is satisfied with $S(1; C_2) = X$ and $\alpha = 0$. Let $\eta_0 = (1/2) \min_{j \neq k} |x_j - x_k|_{\infty}$, and $m = \min_k a_k$. Then $S(\eta_0/3; m, 0) = \{x_1, \ldots, x_K\} = \text{supp}(\mu^*)$. Therefore, it easy to verify that $\mu^*$ has fine structure.

Our goal in this paper is to detect the support of $\mu^*$, and in the case when $\mu^*$ has a fine structure, to separate the supports of $\mu_k$. When the support of each $\mu_k$ is the set of all data points for which the true label $k$ is attached, then we need to discuss a measurement to assess the quality of our approximations to the supports as classification tools. We recall the $F$–measure described for a finite data set in [37]. If $\{C_1, \ldots, C_N\}$ are the obtained clusters from a certain clustering algorithm, and $\{L_1, \ldots, L_K\}$ is a partition of the data according to the (ground-truth) class labels (i.e., $L_k$ is the set of all points in the data set with the class label $k$), then one defines

$$F_D(C_j) = 2 \max_{1 \leq k \leq K} \frac{|C_j \cap L_k|}{|C_j| + |L_k|}, \quad j = 1, \ldots, N.$$ 

The (micro–averaged) $F$–measure is then defined by

$$F_D = \frac{\sum_j |C_j| F_D(C_j)}{\sum_j |C_j|}. \quad (2.8)$$

We interpret the cardinalities above as probabilities. In this paper, the total data is $\text{supp}(\mu^*)$: the partition according to labels in $\{\text{supp}(\mu_k)\}_{k=1}^{K}$. Therefore, the $F$ measure for the clusters $\{C_j\}_{j=1}^{N}$ can be defined as follows: The analogue of $F_D(C_j)$ above is:

$$F(C_j) = 2 \max_{1 \leq k \leq K} \frac{\mu^*(C_j \cap \text{supp}(\mu_k))}{\mu^*(C_j) + \mu^*(\text{supp}(\mu_k))}. \quad (2.9)$$

Then

$$F \left( \left\{ C_j \right\}_{j=1}^{N} \right) = \frac{\sum_{j=1}^{N} \mu^*(C_j) F(C_j)}{\mu^* \left( \bigcup_{j=1}^{N} C_j \right)}. \quad (2.10)$$

Clearly, $F \left( \left\{ C_j \right\}_{j=1}^{N} \right) \leq 1$. If $N = K$, and $C_k = \text{supp}(\mu_k)$ for each $k$, then $F \left( \left\{ C_j \right\}_{j=1}^{N} \right) = 1$. Thus, the closer the quantity $F$ is to 1, the better the quality of clustering with respect to the labels.
Our main tool in this paper are Hermite polynomials. In the univariate case, it is convenient to define the orthonormalized Hermite polynomial \( h_k \) of degree \( k \) recursively by
\[
 xh_{j-1}(x) = \sqrt{\frac{j}{2}} h_j(x) + \sqrt{\frac{j-1}{2}} h_{j-2}(x), \quad j = 2, 3, \ldots, \\
h_0(x) = \pi^{-1/4}, \quad h_1(x) = \sqrt{2\pi}^{-1/4}x.
\] (2.11)

Writing \( \psi_k(x) = h_k(x) \exp(-x^2/2) \), one has the orthogonality relation for \( k, j \in \mathbb{Z}_+ \),
\[
 \int_{\mathbb{R}} \psi_k(x)\psi_j(x)dx = \begin{cases} 
 1, & \text{if } k = j, \\
 0, & \text{if } k \neq j.
 \end{cases}
\] (2.12)

In multivariate case, we adopt the notation \( x = (x_1, \ldots, x_q) \). The orthonormalized Hermite function is defined by
\[
 \psi_k(x) = \prod_{j=1}^q \psi_k(x_j).
\] (2.13)

In general, when univariate notation is used in multivariate context, it is to be understood in the tensor product sense as above; e.g., \( k! = \prod_{j=1}^q k_j! \), \( x^k = \prod_{j=1}^q x_j^{k_j} \), etc.

Let \( H : [0, \infty) \rightarrow [0, 1] \) be a \( C^\infty \) function, \( H(t) = 1 \) if \( t \in [0, 1/2] \), \( H(t) = 0 \) if \( t \geq 1 \). We define the localized kernel by
\[
 \Phi_n(H; x, y) = \Phi_n(x, y) = \sum_{k \in \mathbb{Z}_+^q} H\left(\frac{\sqrt{|k|_1}}{n}\right) \psi_k(x)\psi_k(y).
\] (2.14)

The localization property is made precise in \([6.3]\) below.

### 3 Main theorems

The constants \( \kappa_3 \) and \( C^* \) below will be defined later in Section \([6]\) (Proposition \([6.1]\) and Lemma \([6.3]\)). In Section \([1.2]\) the quantity \( m \) played several roles: the minimum value of the measure on arbitrarily small balls around points of its support and the threshold in Theorem \([1.1]\). Here, the first role is contained in the definition of the sets \( S(\kappa_3/n, C_1, \alpha) \). We will take a multiscale approach by varying the minimal separation \( \eta \) as defined in Definition \([2.1]\) and the threshold \( \Theta \) to be used to determine sets of significant probabilities.

The first theorem describes the location of the support of \( \mu^* \).

**Theorem 3.1** Let \( \mu^* \) be detectable, \( S > \alpha, 0 < \Theta \leq c_1 \). \( n \geq c_2 \) be large enough, so that
\[
 \text{supp } (\mu^*) \subseteq \mathbb{R}(0, \kappa n), \quad S(\kappa_3/n; C_1, \alpha) \neq \emptyset,
\] (3.1)

With \( M \geq c_3 n^{2\alpha} \log n \), let \( \mathcal{C} = \{x_1, \ldots, x_M\} \) be independently sampled from the probability law \( \mu^* \). We define
\[
 \mathcal{G}_n(\Theta, \mathcal{C}) = \left\{ x \in \mathbb{R}^q : \sum_{j=1}^M \Phi_n(x, x_j)^2 \geq \Theta \max_{1 \leq k \leq M} \sum_{j=1}^M \Phi_n(x_k, x_j)^2 \right\}.
\] (3.2)

Then with probability at least \( 1 - c_4/M^{c_5} \),
\[
 S(\kappa_3/n, C_1, \alpha) \subseteq \mathcal{G}_n(\Theta, \mathcal{C}) \subseteq \left\{ x \in \mathbb{R}^q : \text{dist}(x, \text{supp } (\mu^*)) \leq \frac{c_6}{\Theta^{1/(S-\alpha)n}} \right\}.
\] (3.3)

**Theorem 3.2** We assume the set-up as in Theorem \([3.1]\). In addition, we assume that \( \mu^* \) has a fine structure, and that
\[
 n \geq c(\eta \Theta)^{-1}, \quad n^\alpha \mu^*(\mathbf{S}_{k+1, \eta}) \leq c_1 \Theta.
\] (3.4)

Let
\[
 \mathcal{G}_{k, n}(\Theta, \mathcal{C}) = \mathcal{G}_n(\theta, \mathcal{C}) \cap \left\{ x \in \mathbb{R}^q : \text{dist}(x, \mathbf{S}_{k, \eta}) \leq \frac{c_2}{n \Theta^{1/(S-\alpha)}} \right\}.
\] (3.5)
Then with probability exceeding \(1 - c_3M^{-c_4}\), the set \(\mathcal{G}_n(\Theta, \mathcal{C})\) is a disjoint union of sets \(\mathcal{G}_{k,\eta,n}(\Theta, \mathcal{C}), k = 1, \cdots, K_\eta\), such that
\[
\text{dist}(\mathcal{G}_{k,\eta,n}(\Theta, \mathcal{C}), \mathcal{G}_{j,\eta,n}(\Theta, \mathcal{C})) \geq \eta, \quad k \neq j, \quad k, j = 1, \cdots, K_\eta,
\]
and for \(k = 1, \cdots, K_\eta\),
\[
S(k_3/n, C_1, \alpha) \cap \left\{ \mathbf{x} \in \mathbb{R}^q : \text{dist}(\mathbf{x}, \mathbf{S}_{k,\eta}) \leq \frac{c_2}{n^{\Theta_1/(S - \alpha)}} \right\} \subseteq \mathcal{G}_{k,\eta,n}(\Theta, \mathcal{C}) \subseteq \left\{ \mathbf{x} \in \mathbb{R}^q : \text{dist}(\mathbf{x}, \mathbf{S}_{k,\eta}) \leq \frac{c_2}{n^{\Theta_1/(S - \alpha)}} \right\}.
\]

**Example 3.1** We continue the set-up as in Example 2.1. Then for \(\eta \leq \eta_0\), the second condition in (3.4) is satisfied trivially and both the conditions in (3.1) are satisfied for sufficiently large \(n\). In particular, \(K_\eta\) and the sets \(\mathcal{G}_{k,\eta,n}(\Theta)\) do not depend upon \(\eta\) if \(\eta \leq \eta_0\). The parameter \(n\) controls how close one can get to the supports of the measures \(\mu_k\).

**Example 3.2** The same remarks as in Example 3.1 apply also in the set-up of Example 2.2. In this case, the definition of \(\mathcal{G}_{k,\eta,n}(\Theta)\) shows that the diameter of each of these sets is \(\leq 2\Theta/n\). In particular, if
\[
\hat{\mathbf{x}}_k = \arg \max_{\mathbf{x} \in \mathcal{G}_{k,\eta,n}(\Theta)} \sum_{k=1}^K a_k \Phi_n(\mathbf{x}, \mathbf{x}_k)
\]
satisfies \(|\hat{\mathbf{x}}_k - \mathbf{x}_k|_\infty \leq 2\Theta/n\). We note finally that the sum expression in the above expression can be computed using the Hermite moments of \(\mu^*\); the precise location of \(\hat{\mathbf{x}}_k\)'s or the values of \(a_k\) need not be known. More impressively, the value of \(K\) is found automatically rather than being required at the outset. This is consistent with the results in [6].

**Theorem 3.3** We assume the set-up as in Theorem 3.2. With probability \(\geq 1 - cM^{-c_1}\), the clusters \(\mathcal{G}_{k,\eta,n}(\Theta, \mathcal{C})\) satisfy
\[
\lim_{n \to \infty} F\left( \{\mathcal{G}_{k,\eta,n}(\Theta, \mathcal{C})\}_{k=1}^{K_\eta} \right) = 1.
\]

**Remark 3.1** In Theorem 3.3, it is understood implicitly that the quantities \(M\) and \(\eta\) (and also \(\Theta\)) change with \(n\) so as to satisfy the various conditions of Theorem 3.2.

**Remark 3.2** The statement of Theorem 3.3 is valid in a deterministic sense if the set-up of Theorem 6.3 is assumed. In this case, we have, instead of (3.8),
\[
\lim_{n \to \infty} F\left( \{\mathcal{S}_{k,\eta,n}(\theta)\}_{k=1}^{K_\eta} \right) = 1.
\]

### 4 Algorithmic considerations

In order to apply the theory in Section 3 to classification problems in practice, one needs to develop several further details. The theorems do not give a clear algorithm to find the clusters \(\mathcal{G}_{k,\eta,n}\), and the choice of the parameters \(n, \eta, \Theta\) need to be fixed experimentally in each application. We develop these details in this section. In Section 4.1, we discuss how to decide which points lie in a single cluster \(\mathcal{G}_{k,\eta,n}\), and which of these one should query a label for. The theory suggests that we then assign the same label to every point in \(\mathcal{G}_{k,\eta,n}\). In Section 4.2, we explain how to extend the known labels to the remaining points in the data set using the witness function approach in [34]. To take advantage of the multiscale nature of the theory, we describe in Section 4.3 how to transfer sampled labels at a coarse level to inform the clustering and label propagation at finer and finer levels. This discussion is summarized in an outline form in Algorithm 1.

We note that while this algorithm allows for labels that change at various scales \(\eta\), we will describe the algorithm as though there’s a fixed set of labels that one can look up from per point. This means that the label won’t change as \(\eta\) and \(n\) vary. We also note that although the theory in Section 3 (and Section 6) allows one to decide what label if any should be given to any point in the Euclidean space, it is convenient to assume in this section that all the points at which we wish to assign labels are already collected in a data set \(\mathcal{C}\), analogous to the semi-supervised setting.
4.1 Connecting points in $G_{k,\eta,n}$

Assume that there exists some unknown label function $f : \mathbb{R}^d \to \mathbb{R}$. Further, we assume it corresponds to a consistent clustering scheme such that, for small enough $\eta$ and $x \in S_{k,\eta}$, we have that $f(x) = k$. The problem of active learning boils down to learning an estimate $\hat{f}(x)$ of $f(x)$ for all $x \in \bigcup_{k=1}^{K_n} S_{k,\eta}$ given only a small set of points $A \subset \mathcal{C}$ at which $f$ is actually known. The key difference between this and semi-supervised learning is that, in our case, $A$ can be chosen in a data-dependent fashion prior to querying the function. We note here, the choice of label function $f$ may be any layer of some hierarchical tree of labels \cite{15}, but we assume that once we begin the algorithm $f$ is fixed.

The construction in Theorem 3.2 guarantees a construction of sets $\tilde{G}_{k,\eta,n}(\Theta, \mathcal{C})$ that satisfy a minimal separation condition \cite{3.6}. However, in order to use this result to propagate learned cluster labels effectively, it is important to determine which data points $x \in \mathcal{C}$ are in a particular cluster, $x \in \mathcal{C} \cap \tilde{G}_{k,\eta,n}(\Theta, \mathcal{C})$. This is necessary both to:

1. decide the set $A$ of points $x \in \mathcal{C}$ we wish to query for a label $f$, and
2. propagate the labels from $A$ to the rest of $\mathcal{G}_n(\Theta, \mathcal{C})$ in a way that guarantees the estimate $\hat{f}(x)$ agrees with $f(x)$ itself on points $x \in \mathcal{G}_n(\Theta, \mathcal{C})$.

The key to connecting clusters comes from $\mu^*$ having fine structure of minimal separation of $2\eta$ at some scale, and the guarantee that our found clusters satisfy \cite{3.4} for a finite $n$, and that increasing $n$ will allow us to decrease $\eta$. For this reason, we will fix $\Theta$ and consider a minimal separation $\eta_n$ that depends on $n$ in a way that satisfies \cite{3.3}. Given this separation, we construct a nearest neighbor graph $G$ on $\mathcal{G}_n(\Theta, \mathcal{C})$ with edge set $E = \{(x, y) \in G_n(\Theta, \mathcal{C}) : |x - y|_2 < \eta_n/2\}$. In this way, we are guaranteed that each connected component $G_{\ell,\eta_n,n} \subseteq \mathcal{C}$ actually satisfies $G_{\ell,\eta_n,n} \subseteq \tilde{G}_{k,\eta_n,n}(\Theta, \mathcal{C})$ for some $k$. This implies that if we query and obtain a label $f(x)$ at some point $x \in G_{\ell,\eta_n,n}$, then we are guaranteed that all other points in $G_{\ell,\eta_n,n}$ also have label $f(x)$. Note that the number of connected components, which we'll call $\tilde{K}_n$, satisfies $\tilde{K}_n \geq K_{\eta_n}$. This is because a single class can consist of multiple connected components of $G$ at separation $\eta_n$.

The only other problem to address in this framework is to select the points $A$ at which to query a label. While theoretically any point in the connected component would be sufficient, the most reliable point to choose is the mode of the cluster, i.e.,

$$x^* = \arg \max_{x \in G_{\ell,\eta_n,n}} \sum_{j=1}^{M} \Phi_n^2(x, x_j).$$

(4.1)

The argument for this choice is a heuristic one; if there do exist points in $G_{\ell,\eta_n,n}$ with the incorrect label (i.e., some cluster couldn’t be fully resolved at level $\eta_n$) then they are more likely to lie at the boundary of the connected component, and thus have a lower empirical density.

4.2 Classification of the remaining points

For any $\eta$ there may exist low density points that lie outside $\mathcal{G}_n(\Theta, \mathcal{C})$ that may not be classified at level $n$. While this isn’t an issue in the continuum limit due to Theorem 3.3 in most applications we are constrained to finite budget of labels to be sampled. This implies that we must use a finite $n$, as we cannot realistically split $\mathcal{C}$ into $M$ different clusters and sample each point’s label separately; defeating thereby the purpose of the exercise. Because of this, we must have a trade-off between scaling $n$ until we’ve classified all points accurately, and stopping at a finite $n$ to make our best predictions of labels for points in $S_{K_{\eta_n}+1}$.

We propose to classify these additional points through the witness function approach created in \cite{34}. To summarize in this context, we define each class estimate to be $\hat{S}_{k,\eta_n} = \{x_i \in \mathcal{C} : \hat{f}(x_i) = k\}$. Then we construct a witness function for each class

$$\hat{F}_k(x) = \frac{1}{|\hat{S}_{k,\eta_n}|} \sum_{x_j \in \hat{S}_{k,\eta_n}} \Phi_n(x, x_j), \quad \text{for } x \in \mathcal{C} \setminus \mathcal{G}_n(\Theta, \mathcal{C}).$$

(4.2)

The proposed algorithm assigns a label to $x$ given by

$$\hat{f}(x) = \arg \max_{1 \leq k \leq K_n} \hat{F}_k(x),$$

(4.3)

and determine the certainty of classification through a permutation test as in \cite{34}. Note that we will refer to the points assigned in $\mathcal{G}_n(\Theta, \mathcal{C})$ as confident points, and the points assigned by the witness function, $\mathcal{C} \setminus \mathcal{G}_n(\Theta, \mathcal{C})$, as uncertain points. These uncertain points fall outside our guarantees in Theorem 3.2 other than the fact that the set becomes empty as $n \to \infty$. 
Algorithm 1 Cautious Active Clustering

Input: \( C \subseteq \mathbb{R}^d, n_{\text{max}}, \Theta, \tau \)
Output: \( f, G_{n_{\text{max}}} (\Theta, C) \)

\( A \leftarrow \emptyset \)  \hspace{1cm} \triangleright \text{Set of points at which label is queried, together with the corresponding labels.}

\( \text{while } n \leq n_{\text{max}} \) do

\( C_{n,\Theta} \leftarrow G_n (\Theta, C) \cap C \) using \( (3.2) \).
\( E \leftarrow \{(x_i, x_j) : x_i, x_j \in C_{n,\Theta} \text{ and } |x_i - x_j|_2 < \eta_n/2\} \) for \( \eta_n \) as a function of \( n \) as in \( (3.4) \).

Construct graph \( G = (C_{n,\Theta}, E) \).

\( \{C_{n,\ell}\}_{\ell=1}^{K_n} \leftarrow \text{connectedComponents}(G) \) (See Section 4.1).

Set flag(\( \ell \)) = 0 for all \( \ell \).

\( \triangleright \text{When we are done with this } n, \text{ flag}(\ell) = 1 \text{ for all } \ell. \)

\( \text{for } \ell \leq K_n \) do

\( \text{if } C_{n,\ell} \cap A = \emptyset \) then

\( x_i \leftarrow \arg \max_{x_i \in C_{n,\ell}} \sum_{j=1}^M \Phi_n(x_i, x_j)^2 \)
\( A \leftarrow A \cup \{x_i, f(x_i)\} \)
\( \hat{f}(x_j) \leftarrow f(x_j) \forall x_j \in C_{n,\ell} \)
\( \text{flag}(\ell) = 1 \)  \hspace{1cm} \triangleright \text{Point at which label is sought}

\( \triangleright \text{Update } A; \text{ this set does not lose points.} \)
\( \triangleright \text{Extend label to the whole component} \)
\( \triangleright \text{Done for this value of } \ell. \)

else

\( \text{if } \forall x_i \in C_{n,\ell} \cap A, f(x_i) = c_\ell \) then

\( \hat{f}(x_j) \leftarrow c_\ell \forall x_j \in C_{n,\ell} \)
\( \text{flag}(\ell) = 1 \)  \hspace{1cm} \triangleright \text{Extend label to the whole component}
\( \triangleright \text{Done for this value of } \ell. \)

\( \text{end if} \)

\( \text{end if} \)

\( \text{if flag}(\ell) = 1 \text{ for all } \ell, \text{ then} \)

\( n \leftarrow n + \text{step} \)  \hspace{1cm} \triangleright \text{Done for this pass, go to next level}

\( \triangleright \text{to ensure that we captured all points that could be captured.} \)

else

Increase threshold \( \Theta \leftarrow \tau \Theta \), (See Section 4.3).  \hspace{1cm} \triangleright \text{Prune the graph.}

\( \text{end if} \)

\( \text{end while} \)

\( C_{K_n+1} \leftarrow C \setminus G_{n_{\text{max}}} (\Theta_{n_{\text{max}}}, C) \) using \( (3.2) \).  \hspace{1cm} \triangleright \text{Uncertain points}

\( \tilde{S}_{k,\eta_{n_{\text{max}}}} \leftarrow \{x : f(x) = k\} \)
\( \tilde{f}(x_j) \leftarrow \arg \max_{k} \left( \frac{1}{|\tilde{S}_{k,\eta_{n_{\text{max}}}}|} \sum_{x_i \in \tilde{S}_{k,\eta_{n_{\text{max}}}}} \Phi_{n_{\text{max}}}(x_j, x_i) \right) \forall x_j \in C_{K_n+1} \)
\( \triangleright \text{Extend labels to uncertain points using witness function (See Section 4.2).} \)

4.3 Learning across layers

As described to this point, the algorithm for learning \( \tilde{f} \) is computed independently at each \( n \). However, this is not efficient from a label sampling perspective, as there may be significant information already learned at \( n_0 \) for \( n > n_0 \). We consider an increasing hierarchy of the parameter \( n, \{n_i\}_{i=1}^{\infty} \). This similarly determines a hierarchy of decreasing \( \eta, \{\eta_i\}_{i=1}^{\infty} \) such that \( \eta_j < \eta_i \) if \( j > i \).

Let \( A_i \subseteq \bigcup_{k=1}^{n_i} C_{k,\eta_k} \) be the small collection of points at which \( f \) was sampled. By definition of \( \mu^* \) being detectable, \( A_i \subseteq \bigcup_{k=1}^{n_i} C_{k,\eta_k} \) as well for \( j > i \). This means that many of the connected components \( \{C_{k,\eta_k}\}_{k=n_i}^{\infty} \) already have a member \( x \in C_{k,\eta_k} \) such that \( x \in A_i \). Thus, we must only sample the \( K_n \) clusters that do not already contain a sample.

We also wish to comment on the stability of the connected component separation across levels. While we are examining a minimal separation of \( \eta_j \), this is not a known value a priori. Even when estimated, it is possible that two clusters have separation just greater than \( \eta_j \), and that removing low-density points with threshold \( \Theta \) does not help increase the separation of clusters in \( G_{n_j}(\Theta, C) \) sufficiently. Fortunately, this can be easily detected in the situation that subsets were disconnected at \( n_i \) for \( j > i \). In this situation, \( A_i \) will contain two points \( x, y \) with different labels from level \( n_i \) such that \( x, y \in C_{k,\eta_k} \). When this occurs, it is a simple fix to slightly increase \( \Theta \) until \( x \) and \( y \) fall in different clusters. This can be done with a parameter \( \tau > 1 \) that is described in Algorithm
basically increasing the thresholding of low density points before redefining the clusters. This disagreement can thus be easily fixed at level $n_j$ and allows for a more robust clustering that must remain consistent across levels. Similarly, one could decrease the estimate of $n_j$ and rerun the algorithm.

As a final note, it’s possible to increase $\Theta$ or decrease $\eta$ only for points in $C_{\ell,n_j}$ rather than on all $C$. This will lead to a different $\Theta, \eta$ in different regions of space, but the set of points and neighborhoods will be a proper subset of $G_n(\Theta, C)$. However, this no longer guarantees Theorem 3.3.

5 Applications

In this section, we consider a number of applications to both synthetic and real data sets. For the synthetic data, we consider problems that either do not have a minimal separation, or has a very small separation relative to the inter-cluster radius (Section 5.1). This is a particularly difficult set of examples for clustering and active learning problems because many algorithms, such as $k$-means, expect clusters to be somewhat isotropic (i.e., similar variance in all directions). We also consider the latent space of a variational autoencoder that embedded the MNIST data set into a 2D latent space (Section 5.2). This problem again poses difficulty for traditional clustering algorithms, as we have purposefully chosen a latent space dimension that leads to no minimal separation between some label clusters, and even partial overlap of different labels. Even in this setting, we demonstrate strong classification accuracy off of a small number of samples.

As our main set of applications, we consider our active clustering framework on hyperspectral image pixel classification (Section 5.3). Traditionally, this is an application that requires non-Euclidean clustering methods, and a very large number of pixel labels. Similarly, there is rarely a minimal separation between clusters, made worse by the fact that pixels can even be a mix of multiple labels. We compare our algorithms to the current state-of-the-art active clustering algorithm on HSI, the LAND algorithm [28], which uses a Gaussian kernel density estimate and diffusion geometry to define the cluster centers and boundaries.

5.1 Synthetic examples without minimal separation

We examine the problem of learning with few labels on synthetic data that violates traditional clustering assumptions. In this first example in Figure 2, we use data that does not have a minimal separation between clusters. This is a setting in which filtering by density significantly benefits the clustering algorithm, as the clusters in Figure 2 have long tails of low density.

![Figure 2](image)

Figure 2: (Left) Clustered data with no minimal separation. Color corresponds to label. (Center) Example of cautious clustering approach with 4 labels queried and $n = 4$. Dark blue labels are uncertain points, i.e., points below the density threshold. (Right) Different measures of error after witness function propagation from confident points (our algorithm cautious active clustering), and comparison to LAND [28].

In a second example in Figure 3, we consider data that has a very small minimal separation, but the density remains relatively constant between the middle of the clusters and their tails. In this setting, it is critical to have a highly localized kernel for density estimation and defining similarity between points. Because the origin is close to all three of the clusters, using a kernel with poor localization would lead to points near the origin having a higher estimated density than any of the points sampled from the actual distributions. This would lead to sampling points far away from the centers of the clusters.

In a final example in Figure 4, we consider data that has a very small minimal separation compared to their internal maximum radius. In this setting, it is important to have a flexible method for connecting points within cluster, like connected components, that allows for connecting far apart points as long as there exists a path between
Figure 3: (Left) Clustered data with small minimal separation and no density peaks. Color corresponds to label. (Center) Example of cautious clustering approach with 3 labels queried and $n = 4$. Dark blue labels are uncertain points, i.e., points below the density threshold. (Right) Different measures of error after witness function propagation from confident points (our algorithm cautious active clustering), and comparison to LAND [28].

Figure 4: (Left) Clustered data with small minimal separation compared to inner radius. Color corresponds to label. (Center) Example of cautious clustering approach with 2 labels queried and $n = 4$. Dark blue labels are uncertain points, i.e., points below the density threshold. (Right) Different measures of error after witness function propagation from confident points (our algorithm cautious active clustering), and comparison to LAND [28].

them. Without this, one requires a large number of points with queried labels spaced throughout the cluster in order to propagate the labels effectively.

5.2 MNIST generative models

The next set of experiments revolve around estimating regions of space corresponding to given classes, and determining which regions of the latent space correspond to which digit labels. This problem has been of great interest in recent years with the growth of generative networks, namely various variants of generative adversarial networks (GANs) [17] and variational autoencoders (VAEs) [21]. Each has a low-dimensional latent space in which new points are sampled, and mapped to $\mathbb{R}^q$ through a neural network. While GANs have been more popular in literature in recent years, we focus on VAEs in this paper because it is possible to query the locations of training points in the latent space. A good tutorial on VAEs can be found in [16].

We examine this problem with the well known MNIST data set [23]. This is a set of handwritten digits 0–9, each scanned as a 28 × 28 pixel image. There are 50000 images in the training data set, and 10000 in the test data.

In order to select the latent space for this data set, we construct a three layer VAE with encoder $E(x)$ with architecture 784 – 500 – 500 – 2 and a decoder/generator $G(z)$ with architecture 2 – 500 – 500 – 784, and for clarity consider the latent space to be the 2D middle layer. We have purposely chosen a 2D latent space because this leads to varying levels of separation between label clusters, including overlapping clusters of commonly confused digits (e.g., mixing 4’s and 9’s, and mixing 3’s, 5’s, and 8’s). For this reason, we look at both a fine grained and coarse grained label set. In fine grained, we have the traditional 10 labels, in coarse grained we create 7 labels by merging the 4’s and 9’s and merging 3’s, 5’s, and 8’s. This creates an example of the hierarchical label structure as described in the paper, as well as clusters with no minimal separation.

5.3 Hyperspectral imagery

Hyperspectral imagery (HSI) is an imaging modality that captures radiation reflected from a surface across a number of different wavelengths (also called bands). This results in each pixel in an image being represented by its energy in $q$ different bands, where $q$ is sometimes in the hundreds, as the bands can range from low infrared to high
ultraviolet. Each pixel can cover a significant area of the surface, usually several square meters. Because of this it isn’t always advantageous to use spatial similarity to aid in classification and clustering, since neighboring pixels could easily have different labels \[1\]. Instead, we will consider only the spectral similarity among the pixels \[1, 10\].

Hyperspectral pixels are difficult to collect labels for, as it requires physically inspecting the surface to determine its label. Because of this, active learning has become very popular in the remote sensing community \[39, 36\].

Another issue with labeling pixels is that clusters are inherently hierarchical in nature. For example, in agricultural settings, one not only has to distinguish stone from trees (large separation between classes), but also distinguish a particular crop after 4 weeks of growth from crops after 5 or 6 weeks of growth (small separation between classes). Because of this, there exists a hierarchical relationship to the labels, and the level of specificity desired can change the number of clusters and queried labels that are necessary.

Mathematically, each pixel can be thought of as being a data point in \(\mathbb{R}^q\), so that an image with \(M\) pixels can be organized as \(q \times M\) matrix, where the \(j\)-th column represents the \(q\)-band spectral observation on the \(j\)-th pixel.

For all of the examples below, we begin by taking the top principal components of this matrix resulting in a choice of dimension that captures 80% of the variance of the data in place of \(q\).

We demonstrate this hierarchical relationship in an example using the Salinas-A data set of hyperspectral

![Figure 5](image)

**Figure 5**: (Left) Density estimate of 2-dimensional VAE latent space, colored by each MNIST label. (Middle) Comparison of cautious clustering approach using coarse MNIST labels (merging overlapping clusters as described in Section 5.2). (Right) Fscore as a function of number of queries for cautious clustering approach for both coarse MNIST labels and for fine MNIST labels. Certain points refers to Fscore on confident points only.

![Figure 6](image)

**Figure 6**: Salinas-A HSI data. (Top Left) Ground Truth of pixels with color corresponding to one of six labels (fine labels). (Top middle) Same scene with all lettuce labels placed in single class (coarse labels). (Top right) Worst classification accuracy on a class using our algorithm as a function of \(n\). Note that we clearly separate all classes much faster under coarse labeling. (Bottom) Examples of pixels in \(\mathbb{R}^{204}\), colored by label.
imagery\footnote{http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes} Salinas-A is an 83 \times 86 image of three different agricultural crops being grown (broccoli, corn, lettuce). Beyond this, there is a sub-classification of which week of growth the lettuce is in (4 weeks, 5 weeks, 6 weeks, 7 weeks). Each pixel collects 204 bands, and we initially reduce the dimension to 20 using PCA. Thus, there are 7138 points in \( \mathbb{R}^{204} \), which are projected to 7138 points in \( \mathbb{R}^{20} \). We run our cautious clustering algorithm on this data in two settings, and display summary results in Figure 6. We plot the classification accuracy on the worst class as a function of \( n \). We can see that running our cautious hierarchical clustering algorithm on the coarse labeling (broccoli, corn, lettuce) begins to yield correct classification on all classes for much smaller degree \( n \) than for the same data with fine clustering (broccoli, corn, lettuce4, lettuce5, lettuce6, lettuce7). This establishes that the effective minimal separation between all clusters is not constant, but varies depending on the desired level of specificity for the labels.

The main purpose of this data set is to examine the HSI active learning problem, and determine the maximum classification accuracy given a budget of only sampling \( k \) labels. We wish to emphasize that our algorithm returns an additional advantage over most active learning algorithms, namely the set \( \mathcal{G}_{n_{\text{max}}} (\Theta, \mathcal{C}) \) on which we are confident in our classification. This means we can attain near perfect accuracy on these points, as well as use them to estimate the class on \( \mathcal{C} \setminus \mathcal{G}_{n_{\text{max}}} (\Theta, \mathcal{C}) \) using the witness function. We display the accuracy on \( \mathcal{G}_{n_{\text{max}}} (\Theta, \mathcal{C}) \) in Figure 7 as well as the classification accuracy on the full data set after propagating labels to \( \mathcal{C} \setminus \mathcal{G}_{n_{\text{max}}} (\Theta, \mathcal{C}) \) with the witness function.

We also examine a second data set, which is a 57 \times 41 subset of the Indian Pines hyperspectral data set\footnote{http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes}. Each pixel has 220 features, and we initially reduce the dimension to 20 using PCA. The subset we focus on contains three general materials; tilled corn, stone-steel, and soybeans. Furthermore, soybeans are subdivided into tilled, no till, and clean sublabels. This leads to five labels at the finest level of label resolution. We compare the active learning classification accuracy for our algorithm in Figure 8. While this is clearly a more difficult data set than Salinas-A as evidenced by the lower classification accuracy as a function of the number of labels queried, our algorithm still compares favorably to LAND.

![Salinas-A HSI data. (Left) Comparison of our algorithm of cautious active clustering to LAND and random sampling of labels. (Center) Classification accuracy on \( \mathcal{G}_{n_{\text{max}}} (\Theta, \mathcal{C}) \) only, and fraction of points in \( \mathcal{G}_{n_{\text{max}}} (\Theta, \mathcal{C}) \). (Right) F score for \( \mathcal{G}_{n_{\text{max}}} (\Theta, \mathcal{C}) \), and for all points \( \mathcal{C} \). Certain only refers to our Fscore on confident points.](http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes)

![Indian Pines HSI (Left) Ground truth of the small segment of Indian Pines image. (Center) Comparison of our cautious active clustering algorithm to LAND and random sampling of labels. (Right) F score for \( \mathcal{G}_{n_{\text{max}}} (\Theta, \mathcal{C}) \) and for all points \( \mathcal{C} \). Certain only refers to our Fscore on confident points.](http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes)
6 Proofs

In this section, we prove all the theorems in Section 3. The required background is given in Section 6.1. Section 6.2 develops some preparatory results. In Section 6.3, we prove first the deterministic analogues of Theorems 3.1 and Theorems 3.2 in Theorems 6.2 and Theorem 6.3 respectively, and then complete the proofs of the theorems presented in Section 3.

6.1 Background

We recall various properties of the kernel $\Phi_n = \Phi_{n,q}$ defined in (2.14) (cf. [6]).

**Proposition 6.1** There exist $\kappa, \kappa_1, \cdots, \kappa_4 > 0$ depending only on $q$ and $H$ such that

\[
\kappa_n^{2q} \leq \Phi_n(x, x)^2 \leq \kappa_1 n^{2q}, \quad |x|_\infty, |y|_\infty \leq \kappa n, \quad (6.1)
\]

\[
|\Phi_n(x, y)^2 - \Phi_n(x, x)^2| < (1/2)\Phi_n(x, x)^2, \quad |x - y|_\infty \leq \kappa_3/n, \quad |x|_\infty \leq \kappa n, \quad (6.2)
\]

and

\[
|\Phi_n(x, y)^2| \leq \frac{\kappa_4 n^{2q}}{\max(1, (n|x - y|_\infty)^3)}, \quad x, y \in \mathbb{R}^q. \quad (6.3)
\]

For $n > 0$ (not necessarily an integer), let $\Pi_{n,a} = \{x \mapsto P(x) \exp(-a|x|^2) : P \text{ polynomial of total degree } < n^2\}$. We will omit the mention of $a$ when $a = 1$. Members of $\Pi_n$ will be called $(q$-variable) weighted polynomials. The symbol $\| \cdot \|$ will denote the supremum norm on the space $C_0(\mathbb{R}^q)$. The following proposition states two important facts about weighted polynomials, obtained by applying corresponding univariate results in [30, 33] one variable at a time to the multi-variate case.

**Proposition 6.2** Let $n \geq 1$, $P \in \Pi_{n,a}$.

(a) (MRS identity) We have

\[
\|P\| = \max_{x \in [-n/a, n/a]^q} |P(x)|. \quad (6.4)
\]

(b) (Bernstein inequality) There is a positive constant $\kappa_5$ depending only on $q$ such that

\[
\| \nabla P \| \leq \kappa_5 \frac{n}{a} \|P\|. \quad (6.5)
\]

The following corollary is easy to prove:

**Corollary 6.1** Let $n > 0$, $C \subset [-n/a, n/a]^q$ be a finite set satisfying

\[
\max_{x \in [-n/a, n/a]^q} \min_{y \in C} |x - y|_\infty \leq a/(2\kappa_5 n). \quad (6.6)
\]

Then for any $P \in \Pi_{n,a}$,

\[
\max_{y \in C} |P(y)| \leq \|P\| \leq 2\max_{y \in C} |P(y)|. \quad (6.7)
\]

There exists a set $C$ as above with $|C| \sim n^{2q}$.

We will need the following facts from probability theory. Theorem 6.1 is proved as [34, Theorem 6.1].

**Theorem 6.1** Let $\mathcal{X}$ be a topological space, $W$ be a linear subspace of $C_0(\mathcal{X})$. We assume that there is a finite set $\mathcal{C}$ (norming set) satisfying

\[
\sup_{x \in \mathcal{X}} |f(x)| \leq n(W, \mathcal{C}) \sup_{y \in \mathcal{C}} |f(y)|, \quad f \in W. \quad (6.8)
\]

Let $(\Omega, \mathcal{B}, \mu)$ be a probability space, and $Z : \Omega \to W$. We assume further that for any $x \in \mathcal{X}$, $\omega \in \Omega$, $|Z(\omega)(x)| \leq R$ for some $R > 0$. Then for any $\delta > 0$, integer $M \geq 1$, and independent sample $\omega_1, \cdots, \omega_M$, we have

\[
\text{Prob}_\mu \left( \sup_{x \in \mathcal{X}} \left| \frac{1}{M} \sum_{j=1}^M Z(\omega_j)(x) - \mathbb{E}_\mu(Z(\cdot)(x)) \right| \geq 4n(W, \mathcal{C})R \sqrt{\frac{\log(2|\mathcal{C}|/\delta)}{M}} \right) \leq \delta. \quad (6.9)
\]
The following proposition summarizes the multiplicative Chernoff bounds in the form we need them (cf., e.g., [18 Eqn (7)] for an elementary proof).

**Proposition 6.3** Let \( M \geq 1 \), \( 0 \leq p \leq 1 \), and \( X_1, \ldots, X_M \) be random variables taking values in \( \{0,1\} \), with \( \text{Prob}(X_k = 1) = p \). Then for \( \epsilon \in (0,1] \),

\[
\text{Prob}\left( \sum_{k=1}^{M} X_k \leq (1 - \epsilon)Mp \right) \leq \exp(-\epsilon^2 Mp/2). \tag{6.10}
\]

### 6.2 Preparatory results

In this section and the next, we will assume that \( \mu^* \) is a detectable measure with parameters as described in Definition 2.1. We denote

\[
I_n = \sup_{z \in \text{supp}(\mu^*)} \int_{\mathbb{R}^q} \Phi_n(z,y)^2 d\mu^*(y). \tag{6.11}
\]

**Lemma 6.1** Let \( d > 0 \), \( n \geq 1 \), \( x \in \mathbb{R}^q \). Then there exist \( C_3, C_4, C_5 \) such that

\[
\int_{\mathbb{R}^q \setminus B(x,d)} \Phi_n(x,y)^2 d\mu^*(x) \leq C_3 \kappa_4 n^{2q - \alpha} \min\{1, (nd)^{\alpha-S}\}. \tag{6.12}
\]

If \( S(\kappa_3/n; C_1, \alpha) \cap \mathbb{B}(0,\kappa n) \neq \emptyset \) then

\[
n^{2q - \alpha}/C_4 \leq \inf_{x \in S(\kappa_3/n; C_1, \alpha) \cap \mathbb{B}(0,\kappa n)} \int_{\mathbb{R}^q} \Phi_n(x,y)^2 d\mu^*(y) \leq \max_{x \in \mathbb{R}^q} \int_{\mathbb{R}^q} \Phi_n(x,y)^2 d\mu^*(y) \leq C_3 n^{2q - \alpha}. \tag{6.13}
\]

In particular,

\[
I_n = \sup_{z \in \text{supp}(\mu^*)} \int_{\mathbb{R}^q} \Phi_n(z,y)^2 d\mu^*(y) \sim \sup_{x \in \mathbb{R}^q} \int_{\mathbb{R}^q} \Phi_n(x,y)^2 d\mu^*(y) \sim n^{2q - \alpha}. \tag{6.14}
\]

**Proof.** First, let \( d \geq 1/n \), and for \( k \in \mathbb{Z}_+ \), \( A_k = \{ y : 2^k d < |x - y|_\infty \leq 2^{k+1} d \} \). Then (2.4) shows that \( \mu^*(A_k) \leq 2^{\alpha} C_2 (2kd)^\alpha \). Hence, (6.3) leads to

\[
\int_{\mathbb{R}^q \setminus B(x,d)} \Phi_n(x,y)^2 d\mu^*(x) \leq \kappa_4 n^{2q - S} \int_{\mathbb{R}^q \setminus B(x,d)} \frac{d\mu^*(y)}{|x - y|_\infty^S} = \kappa_4 n^{2q - S} \sum_{k=0}^{\infty} \int_{A_k} \frac{d\mu^*(y)}{|x - y|_\infty^S} \leq \kappa_4 n^{2q - S} d^{-S} \sum_{k=0}^{\infty} 2^{-kS} \mu^*(A_k) \leq 2^{\alpha} C_2 \kappa_4 n^{2q - \alpha} (nd)^{\alpha - S} \sum_{k=0}^{\infty} 2^{k(\alpha - S)} = \frac{2^{\alpha}}{1 - 2^{S - \alpha}} C_2 \kappa_4 n^{2q - \alpha} (nd)^{\alpha - S}. \tag{6.15}
\]

Using this estimate with \( d = 1/n \), and using (6.3) and (2.4) again, we obtain that

\[
\int_{\mathbb{R}^q} \Phi_n(x,y)^2 d\mu^*(x) = \int_{\mathbb{B}(x,1/n)} \Phi_n(x,y)^2 d\mu^*(x) + \int_{\mathbb{R}^q \setminus B(x,d)} \Phi_n(x,y)^2 d\mu^*(x) \leq C_2 \kappa_4 n^{2q - \alpha} + \frac{2^{\alpha}}{1 - 2^{S - \alpha}} C_2 \kappa_4 n^{2q - \alpha}.
\]

This shows both the third inequality in (6.13), and together with (6.15) also (6.12) (with the same \( C_3 \)).

Let \( x_0 \in S(\kappa_3/n; C_1, \alpha) \cap \mathbb{B}(0,\kappa n) \). Then \( \mu^*(\mathbb{B}(x_0,\kappa_3/n)) \geq C_1 n^{-\alpha} \). In view of (6.1) and (6.2), we have \( \Phi_n(x_0, y)^2 \geq (\kappa_1/2)n^{2q} \) for all \( y \in \mathbb{B}(x_0,\kappa_3/n) \). Therefore,

\[
\int_{\mathbb{B}(x_0,\kappa_3/n)} \Phi_n(x_0, y)^2 d\mu^*(y) \geq (C_1 \kappa_1/2)n^{2q - \alpha}.
\]

This leads to the first inequality in (6.13). \( \square \)
Lemma 6.2 Let $n \geq 1$ be large enough so that $S(\kappa_3/n; C_1, \alpha) \cap \mathbb{B}(0, \kappa n) \neq \emptyset$, \{x_j\}$_{j=1}^{M}$ be independent samples with $\mu^*$ as the probability law. Then

\[ \text{Prob}\left( \sup_{x \in \mathbb{R}^q} \left| \frac{1}{M} \sum_{j=1}^{M} \varphi_n(x, x_j)^2 - \int_{\mathbb{R}^q} \varphi_n(x, y)^2 d\mu^*(y) \right| \geq cn^2 \sqrt{\frac{\log n}{M}} I_n \right) \leq \frac{1}{2n}. \]  

(6.16)

In particular, if $\beta > 0$, and with $c$ as in (6.16),

\[ M \geq (c^2/\beta^2)n^{2\alpha} \log n, \]  

(6.17)

then with probability $\geq 1 - 1/(2n)$, for $x \in \mathbb{R}^q$,

\[ \int_{\mathbb{R}^q} \varphi_n(x, y)^2 d\mu^*(y) - \beta I_n \leq \frac{1}{M} \sum_{j=1}^{M} \varphi_n(x, x_j)^2 \leq \int_{\mathbb{R}^q} \varphi_n(x, y)^2 d\mu^*(y) + \beta I_n, \]  

(6.18)

\[ (1 - \beta)I_n \leq \max_{x \in \mathbb{R}^q} \frac{1}{M} \sum_{j=1}^{M} \varphi_n(x, x_j)^2 \leq (1 + \beta)I_n. \]  

(6.19)

PROOF. We use Theorem 6.1 with the following choices: $\mu^*$ in place of $\mu$, $x_j$ in place of $\omega_j$, $\delta = 1/(2n)$, $Z(\omega)(x) = \Phi_n(x, \omega)^2$ (so that $W = \Pi_{2n}$, and with $C$ as in Corollary 6.1). This yields

\[ \text{Prob}\left( \sup_{x \in \mathbb{R}^q} \left| \frac{1}{M} \sum_{j=1}^{M} \varphi_n(x, x_j)^2 - \int_{\mathbb{R}^q} \varphi_n(x, y)^2 d\mu^*(y) \right| \geq cn^2 \sqrt{\frac{\log n}{M}} I_n \right) \leq \frac{1}{2n}. \]  

The proof is completed using (6.13). \hfill \Box

Lemma 6.3 Let $n \geq c_1$, $0 < \beta < 1$, $M \geq c_2 \beta^{-2}n^{2\alpha} \log n$. There exist $C^*, c_1, c_2 > 0$ with the following property. Then for any sample \{x_j\}$_{j=1}^{M}$ with $\mu^*$ as the probability law,

\[ \text{Prob}\left( \max_{1 \leq k \leq M} \frac{1}{M} \sum_{j=1}^{M} \varphi_n(x_k, x_j)^2 \leq C^* I_n \right) \leq 1/n; \]  

(6.20)

i.e., with probability $\geq 1 - 1/n$, (cf. (6.19))

\[ C^* I_n \leq \max_{1 \leq k \leq M} \frac{1}{M} \sum_{j=1}^{M} \varphi_n(x_k, x_j)^2 \leq (1 + \beta)I_n. \]  

(6.21)

PROOF. In this proof, let $P \in \Pi_{2n}$ be defined by

\[ P(x) = \int_{\mathbb{R}^q} \varphi_n(x, y)^2 d\mu^*(y). \]

Let $n$ be large enough so that $\text{supp} (\mu^*) \subset \mathbb{B}(0, \kappa n)$ and $S(\kappa_3/n; C_1, \alpha) \neq \emptyset$. Then (6.13) shows that

\[ \sup_{x \in S(\kappa_3/n; C_1, \alpha)} |P(x)| = P(x^*) \geq cI_n \]  

(6.22)

for some $x^* \in S(\kappa_3/n; C_1, \alpha)$. Therefore, using the Bernstein inequality (6.5), we obtain for $x \in \mathbb{R}^q$,

\[ |P(x^*) - P(x)| \leq cn|x^* - x|_{\infty} I_n \leq cn|x^* - x|_{\infty} \sup_{x \in S(\kappa_3/n; C_1, \alpha)} |P(x)| = cn|x^* - x|_{\infty} P(x^*); \]

i.e.,

\[ P(x) \geq (1 - cn|x^* - x|_{\infty}) P(x^*) \geq c_3 (1 - cn|x^* - x|_{\infty}) I_n, \quad x \in \mathbb{B}(x^*, (cn)^{-1}). \]  

(6.23)
Now we consider the following random variables: for $k = 1, \cdots, M$, we take $X_k = 1$ if $x_k \in \mathbb{B}(x^*, \kappa_3/n^2)$, and 0 otherwise, so that the probability $p$ that $X_k = 1$ is given by $p = \mu^*(\mathbb{B}(x^*, \kappa_3/n^2)) \geq cn^{-2\alpha}$. We then use multiplicative Chernoff bound (6.10) with $\epsilon = 1$ to obtain for $M \geq cn^{2\alpha} \log n$,

$$\text{Prob}\left(\sum_{k=1}^{M} X_k \leq 0\right) \leq \exp(-Mp/2) \leq 1/(2n).$$

Thus, with probability exceeding $1 - 1/(2n)$, there exists $x_\ell \in \mathbb{B}(x^*, \kappa_3/n^2)$. Together with (6.23) this shows that with probability exceeding $1 - 1/(2n)$,

$$\max_{1 \leq k \leq M} P(x_k) \geq P(x_\ell) \geq 2C^* I_n. \tag{6.24}$$

Lemma 6.2 shows that if $M \geq cn^{2\alpha} \log n$, then with $C^*$ as in (6.24), and probability exceeding $1 - 1/n$,

$$\max_{1 \leq k \leq M} \frac{1}{M} \sum_{j=1}^{M} \Phi_n(x_k, x_j)^2 \geq \max_{1 \leq k \leq M} P(x_k) - (C^*/2)I_n \geq C^* I_n. \tag{6.28}$$

\boxend

### 6.3 Proofs of the main theorems

We first state and prove some theorems in the non-noisy case.

**Theorem 6.2** Let $\mu^*$ be detectable, $S > \alpha$, $\theta > 0$, and

$$S = S_n(\theta) = \left\{ \int_{\mathbb{R}^q} \Phi_n(x, y)^2 d\mu^*(y) \geq 4\theta \sup_{x \in \text{supp} (\mu^*)} \int_{\mathbb{R}^q} \Phi_n(x, y)^2 d\mu^*(y) \right\}. \tag{6.25}$$

We assume (3.1) and

$$0 < \theta \leq \min\left( (4C_3C_4)^{-1}, C_3C_4\kappa_4 \right). \tag{6.26}$$

Then with

$$d(\theta) = \left( \frac{\kappa_4 C_3 C_4}{\theta} \right)^{1/(S-\alpha)} \tag{6.27}$$

$$S(\kappa_3/n, C_1, \alpha) \subseteq S \subseteq \left\{ x \in \mathbb{R}^q : \text{dist}(x, \text{supp} (\mu^*)) \leq \frac{d(\theta)}{n} \right\}. \tag{6.28}$$

\textbf{Proof.} We note that $S(\kappa_3/n, C_1, \alpha) \subseteq \text{supp} (\mu^*) \subseteq \mathbb{B}(0, \kappa n)$. So, (6.13) shows that for $x \in S(\kappa_3/n, C_1, \alpha)$,

$$\int_{\mathbb{R}^q} \Phi_n(x, y)^2 d\mu^*(y) \geq \frac{I_n}{C_3 C_4}. \tag{6.29}$$

Since $\theta \leq (4C_3C_4)^{-1}$, this shows the first inclusion in (6.28).

Let $\text{dist}(x, \text{supp} (\mu^*)) \geq d(\theta)/n$. The condition (6.20) shows that $d(\theta) \geq 1$. So, (6.12) leads to

$$\int_{\mathbb{R}^q} \Phi_n(x, y)^2 d\mu^*(y) \geq \int_{\text{supp} (\mu^*)} \Phi_n(x, y)^2 d\mu^*(y)$$

$$\leq \int_{\mathbb{R}^q \setminus \mathbb{B}(x, d(\theta)/n)} \Phi_n(x, y)^2 d\mu^*(x) \leq C_3 \kappa_4 n^{\alpha - 2\alpha} (\theta)^{\alpha - S}$$

$$\leq C_3 C_4 \kappa_4 d(\theta)^{\alpha - S} I_n = \theta I_n. \tag{6.29}$$

This proves the second inclusion in (6.28). \boxend

The next theorem shows the detection of the supports $S_{k, \eta}$ of the components $\mu_k$ of $\mu^*$. 

18
Theorem 6.3  We assume the set-up as in Theorem 6.2. In addition, we assume that \( \mu^* \) has a fine structure, and that
\[
n \geq 2d(\theta)/\eta, \quad \kappa_1 C_4 n^\alpha \mu^*(S_{K_{n+1}}) \leq \theta. \tag{6.30}
\]
Let
\[
S_{k,\eta,n}(\theta) = S_n(\theta) \cap \{ x \in \mathbb{R}^q : \text{dist}(x, S_{k,\eta}) \leq d(\theta)/n \}.
\]
Then the set \( S_n(\theta) \) is a disjoint union of sets \( S_{k,\eta,n}(\theta), k = 1, \ldots, K_\eta \) such that
\[
\text{dist}(S_{k,\eta,n}(\theta), S_{j,\eta,n}(\theta)) \geq \eta, \quad k \neq j, \quad k, j = 1, \ldots, K_\eta,
\]
and for \( k = 1, \ldots, K_\eta \),
\[
S(\kappa_3/n, C_1, \alpha) \cap \{ x : \text{dist}(x, S_{k,\eta}) \leq d(\theta)/n \} \subseteq S_{k,\eta,n}(\theta) \subseteq \{ x \in \mathbb{R}^q : \text{dist}(x, S_{k,\eta}) \leq d(\theta)/n \}.
\]

**Proof.**
The minimal separation condition and the first condition in (6.30) implies that the sets \( S_{k,\eta,n} \) are disjoint, and in fact, satisfy (6.32). Also, (6.28) implies the first inclusion in (6.33).

Let \( x \in S_n(\theta) \). Then we deduce using (6.1), (6.13), and (6.30) that
\[
\int_{S_{K_{n+1}}} \Phi_n(x, y)^2 d\mu^*(y) \leq \kappa_1 n^{2\alpha} \mu^*(S_{K_{n+1}}) \leq \kappa_1 C_4 n^\alpha \mu^*(S_{K_{n+1}}) I_n \leq \theta I_n.
\]

In this proof, we will denote
\[
S = \bigcup_{k=1}^{K_\eta} S_{k,\eta,n}.
\]
If \( \text{dist}(x, S) \geq d(\theta)/n \) then we obtain as in the proof of Theorem 6.2 that
\[
\int_{S} \Phi_n(x, y)^2 d\mu^*(y) \leq \int_{\mathbb{R}^q \setminus B(x, d(\theta)/n)} \Phi_n(x, y)^2 d\mu^*(x) \leq C_3 C_4 n^{\alpha - 2\alpha} d(\theta)^{\alpha - S} \leq C_3 C_4 d(\theta)\alpha - S I_n = \theta I_n.
\]
Together with (6.34), this implies that \( S_n(\theta) \subseteq \{ x : \text{dist}(x, S) \leq d(\theta)/n \} \). Since \( d(\theta)/n \leq \eta/2 \), the minimal separation condition shows that for any \( x \) with \( \text{dist}(x, S) \leq d(\theta)/n \), there exists a unique \( k, k = 1, \ldots, K_\eta \), such that \( \text{dist}(x, S_{k,\eta}) \leq d(\theta)/n \). Thus, \( S_n(\theta) = \bigcup_{k=1}^{K_\eta} S_{k,\eta,n}(\theta) \), and the second inclusion of (6.33) is proved. \( \square \)

The following lemma helps us to connect Theorems 6.2 and 6.3 with Theorems 3.1 and 3.2.

**Lemma 6.4** Let \( 0 < \Theta \leq 1 \), \( M, n \geq 2 \) be integers, \( M \geq 2 \) and \( C = \{ x_1, \ldots, x_M \} \) be independently sampled from the probability law \( \mu^* \). Let \( G_n(\Theta, C) \) be defined by (3.2). There exist constants \( c, c_1, c_2 \) such that if \( M \geq c n^{2\alpha} \sqrt{\log n} \) then with probability \( \geq 1 - c_1 M^{-c_2} \),
\[
S_n \left( \frac{(1 + C^*) \Theta}{4} \right) \subseteq G_n(\Theta, C) \subseteq S_n(C^* \Theta/8). \tag{6.35}
\]

**Proof.** All statements below hold with probability \( \geq 1 - c_1 M^{-c_2} \), although the values of \( c_1, c_2 \) might be different at different occurrences as usual. In applying Lemma 6.2 we use \( \beta = C^* \Theta/2 \),
\[
t_1 = \frac{2\Theta + C^* \Theta(1 + \Theta)}{8} = \frac{(1 + \beta) \Theta + \beta}{4}.
\]
Let \( x \in S_n(t_1) \). Using (6.18) and (6.21), we deduce that
\[
\frac{1}{M} \sum_{j=1}^{M} \Phi_n(x, x_j)^2 \geq \int_{\mathbb{R}^q} \Phi_n(x, y)^2 d\mu^*(y) - \beta I_n \geq 4t_1 I_n - \beta I_n = (4t_1 - \beta) I_n
\]
\[
\geq \frac{4t_1 - \beta}{1 + \beta} \max_{1 \leq k \leq M} \frac{1}{M} \sum_{j=1}^{M} \Phi_n(x_k, x_j)^2 = \Theta \max_{1 \leq k \leq M} \frac{1}{M} \sum_{j=1}^{M} \Phi_n(x_k, x_j)^2.
\]
Thus,
\[
S_n \left( \frac{2\Theta + C^* \Theta(1 + \Theta)}{8} \right) \subseteq G_n(\Theta, C).
\]
Since
\[(1 + C^*)\Theta/2 \geq \frac{2\Theta + C^*\Theta(1 + \Theta)}{8},\]
this proves the first inclusion in (6.35).

Next, let \(x \in \mathcal{G}_n(\Theta, C)\). Using (6.18) and (6.21), we deduce that
\[
\int_{\mathbb{R}^q} \Phi_n(x, y)^2 d\mu^*(y) \geq \frac{1}{M} \sum_{j=1}^M \Phi_n(x, x_j)^2 - \beta I_n \geq \Theta \max_{1 \leq k \leq M} \frac{1}{M} \sum_{j=1}^M \Phi_n(x_k, x_j)^2
\]
\[
\geq (C^*\Theta - \beta) I_n = 4(C^*\Theta/8)I_n.
\]
This proves the second inclusion in (6.35). \(\square\)

**Proofs of Theorems 3.1 and 3.2**

Theorems 3.1 and 3.2 follow immediately from Theorems 6.2 and 6.3 respectively using Lemma 6.4. \(\square\)

**Proof of Theorem 3.3**

In this proof, we write \(\mathcal{G}_{k,n}\) in place of \(\mathcal{G}_{k,n}(\Theta, C)\), and with the constants \(c_1, c_2\) as in Theorem 3.2
\[
d_n = \frac{c_2}{n\Theta^{1/(S_0 - 1)}}, \quad \epsilon_n = \mu^*(\text{supp}(\mu^*) \setminus S(\kappa_3/n; C_1, \alpha))|, \quad \theta_n = c_1\Theta/n^{\alpha}.
\]

Since \(S_{k,n} \subseteq \text{supp}(\mu_k)\) are disjoint sets, (3.4) shows that
\[
|\mu^*(\text{supp}(\mu_k)) - \mu^*(S_{k,n})| = \mu^*(\text{supp}(\mu_k) \setminus S_{k,n}) \leq \mu^*(S_{K_k+1,n}) \leq c\theta_n.
\] (6.36)

Hence,
\[
\mu^*(S_{k,n} \setminus S(\kappa_3/n; C_1, \alpha)) \leq \epsilon_n,
\]
and
\[
\mu^*(S(\kappa_3/n; C_1, \alpha) \cap S_{k,n}) = \mu^*(S_{k,n}) - \mu^*(S_{k,n} \setminus S(\kappa_3/n; C_1, \alpha)) \geq \mu^*(S_{k,n}) - \epsilon_n.
\]

Together with the first inclusion in (3.5) and (6.36), this implies that for \(k = 1, \ldots, K\),
\[
\mu^*(\mathcal{G}_{k,n}) \geq \mu^*(\mathcal{G}_{k,n} \cap \text{supp}(\mu_k)) \geq \mu^*(\text{supp}(\mu_k)) - \epsilon_n - c\theta_n.
\] (6.37)

The estimate (6.36) implies again that for \(k = 1, \ldots, K\),
\[
\mu^*(\{x \in \mathbb{R}^q : \text{dist}(x, S_{k,n}) \leq d_n\})
\]
\[
= \mu^*(\text{supp}(\mu_k) \cap \{x \in \mathbb{R}^q : \text{dist}(x, S_{k,n}) \leq d_n\}) = \mu^*(S_{k,n} \cup S_{K_{k+1,n}}) \leq \mu^*(S_{k,n}) + c\theta_n
\]
\[
\leq \mu^*(\text{supp}(\mu_k)) + c\theta_n.
\]

Hence, the second inclusion in (3.5) implies that for \(k = 1, \ldots, K\),
\[
\mu^*(\mathcal{G}_{k,n}) \leq \mu^*(\text{supp}(\mu_k)) + c\theta_n.
\] (6.38)

In view of (6.37) and (6.38), for each \(k = 1, \ldots, K\),
\[
F(\mathcal{G}_{k,n}) \geq \frac{\mu^*(\text{supp}(\mu_k)) - \epsilon_n - c\theta_n}{2\mu^*(\text{supp}(\mu_k)) + c\theta_n} \geq 1 - \frac{2\epsilon_n - c\theta_n}{\mu^*(\text{supp}(\mu_k))}.
\]

Using (6.37) again,
\[
\sum_{k=1}^K \mu^*(\mathcal{G}_{k,n}) F(\mathcal{G}_{k,n}) \geq \sum_{k=1}^K \mu^*(\mathcal{G}_{k,n}) - (2\epsilon_n - c\theta_n)(1 - \epsilon_n - c\theta_n) \sum_{k=1}^K \frac{1}{\mu^*(\text{supp}(\mu_k))}
\] (6.39)

In view of (3.3)
\[
\sum_{k=1}^K \mu^*(\mathcal{G}_{k,n}) = \mu^*(\mathcal{G}_n(\Theta, C)) \geq 1 - \mu^*(\text{supp}(\mu^*) \setminus S(\kappa_3/n; C_1, \alpha)) \leq 1 - \epsilon_n.
\]

(6.39) leads to
\[
1 \geq F\left(\left\{\mathcal{G}_{k,n}\right\}_{k=1}^K\right) \geq 1 - \frac{(2\epsilon_n - c\theta_n)(1 - \epsilon_n - c\theta_n)}{1 - \epsilon_n} \sum_{k=1}^K \frac{1}{\mu^*(\text{supp}(\mu_k))}.
\]

This completes the proof. \(\square\)
A Constructing Localized Kernel $\Phi_n$

With

$$\text{Proj}_m(x, y) = \sum_{|k|_1 = m} \psi_k(x) \psi_k(y), \quad (A.1)$$

we observe that

$$\Phi_n(H; x, y) = \sum_{k \in \mathbb{Z}^n_+} H \left( \frac{\sqrt{|k|_1}}{n} \right) \psi_k(x) \psi_k(y) = \sum_{m=0}^{\infty} H \left( \frac{\sqrt{m}}{n} \right) \text{Proj}_m(x, y). \quad (A.2)$$

In [34], we have observed using the so-called Mehler identity that

$$\text{Proj}_m(x, y) = \sum_{j=0}^{m} \psi_j(|x|) \psi_j(|y| \cos \theta) \sum_{\ell=0}^{m-j} \psi_\ell(0) \psi_\ell(|y| \sin \theta) D_{q-2;m-j-\ell}, \quad (A.3)$$

where $\theta$ is the acute angle between $x$ and $y$, and

$$\psi_\ell(0) = \begin{cases} \pi^{-1/4} (-1)^\ell/2 \frac{\sqrt{\ell!}}{2^{\ell/2} (\ell/2)!}, & \text{if } \ell \text{ is even}, \\ 0, & \text{if } \ell \text{ is odd}, \end{cases} \quad (A.4)$$

and

$$D_{q-2;r} = \begin{cases} \pi^{1-r/2} \Gamma(q/2 + r/2 - 1) \Gamma(q/2 - 1)(r/2)! \Gamma(q/2 - 1)(r/2)! \Gamma(q/2), & \text{if } r \text{ is even, } q \geq 3, \\ 0, & \text{if } r \text{ is odd, } q \geq 3, \\ 1, & \text{if } q \leq 2. \end{cases} \quad (A.5)$$

Therefore, the procedure to compute the kernel $\Phi_n(H; x, y)$ is simple. We use the recurrence relations (2.11) to compute the univariate Hermite functions $\psi_j$, use these together with (A.3) to compute $\text{Proj}_m(x, y)$ for $|m|_1 \leq n^2$, and finally compute $\Phi_n(H; x, y)$ using (A.2).

Acknowledgments

This work was supported by NSF grants DMS-2012355 and DMS-2012266. AC was also partially supported by NSF grant DMS-1819222, and Russel Sage Foundation Grant 2196. Thanks to James Murphy for sharing code for LAND.

References

[1] C. M. Bachmann, T. L. Ainsworth, and R. A. Fusina. Exploiting manifold geometry in hyperspectral imagery. IEEE transactions on Geoscience and Remote Sensing, 43(3):441–454, 2005.

[2] A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In Proceedings of the 26th annual international conference on machine learning, pages 49–56, 2009.

[3] X. Cheng, A. Cloninger, and R. R. Coifman. Two-sample statistics based on anisotropic kernels. arXiv preprint arXiv:1709.05006, 2017.

[4] C. K. Chui, F. Filbir, and H. N. Mhaskar. Representation of functions on big data: graphs and trees. Applied and Computational Harmonic Analysis, 38(3):489–509, 2015.

[5] C. K. Chui and H. N. Mhaskar. Signal decomposition and analysis via extraction of frequencies. Applied and Computational Harmonic Analysis, 40(1):97–136, 2016.

[6] C. K. Chui and H. N. Mhaskar. A Fourier-invariant method for locating point-masses and computing their attributes. Appl. Comput. Harmon. Anal., 45:436–452, 2018.
[7] C. K. Chui and H. N. Mhaskar. A unified method for super-resolution recovery and real exponential-sum separation. *Appl. Comput. Harmon. Anal.*, published online February 21, 2018, 2018.

[8] C. K. Chui, H. N. Mhaskar, and M. D. van der Walt. Data-driven atomic decomposition via frequency extraction of intrinsic mode functions. *GEM-International Journal on Geomathematics*, 7(1):117–146, 2016.

[9] C. K. Chui, H. N. Mhaskar, and X. Zhuang. Representation of functions on big data associated with directed graphs. *Applied and Computational Harmonic Analysis*, 44:165–188, 2018.

[10] A. Cloninger, W. Czaja, and T. Doster. Operator analysis and diffusion based embeddings for heterogeneous data fusion. In *2014 IEEE Geoscience and Remote Sensing Symposium*, pages 1249–1252. IEEE, 2014.

[11] S. Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in neural information processing systems*, pages 235–242, 2006.

[12] S. Dasgupta and D. Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th international conference on Machine learning*, pages 208–215, 2008.

[13] S. Dasgupta, D. Hsu, S. Poulis, and X. Zhu. Teaching a black-box learner. In *International Conference on Machine Learning*, pages 1547–1555, 2019.

[14] S. Dasgupta and P. M. Long. Performance guarantees for hierarchical clustering. *Journal of Computer and System Sciences*, 70(4):555–569, 2005.

[15] O. Dekel, J. Keshet, and Y. Singer. Large margin hierarchical classification. In *Proceedings of the twenty-first international conference on Machine learning*, page 27, 2004.

[16] C. Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[18] T. Hagerup and C. Rüb. A guided tour of chernoff bounds. *Information processing letters*, 33(6):305–308, 1990.

[19] S. Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the 24th international conference on Machine learning*, pages 353–360, 2007.

[20] S. Hanneke and L. Yang. Minimax analysis of active learning. *The Journal of Machine Learning Research*, 16(1):3487–3602, 2015.

[21] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[22] S. Lafon and A. B. Lee. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(9):1393–1403, 2006.

[23] Y. LeCun, C. Cortes, and C. Burges. Mnist handwritten digit database. *AT&T Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[24] W. Li, W. Liao, and A. Fannjiang. Super-resolution limit of the esprit algorithm. *IEEE Transactions on Information Theory*, 2020.

[25] S. Ling and T. Strohmer. Certifying global optimality of graph cuts via semidefinite relaxation: A performance guarantee for spectral clustering. *Foundations of Computational Mathematics*, pages 1–55, 2019.

[26] W. Liu, B. Dai, X. Li, Z. Liu, J. M. Rehg, and L. Song. Towards black-box iterative machine teaching. *arXiv preprint arXiv:1710.07742*, 2017.

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

[28] M. Maggioni and J. M. Murphy. Learning by unsupervised nonlinear diffusion. *Journal of Machine Learning Research*, 20(160):1–56, 2019.
[29] H. Mhaskar and J. Prestin. On the detection of singularities of a periodic function. *Advances in Computational Mathematics*, 12(2-3):95–131, 2000.

[30] H. N. Mhaskar. *Introduction to the theory of weighted polynomial approximation*, volume 56. World Scientific Singapore, 1996.

[31] H. N. Mhaskar. On the representation of smooth functions on the sphere using finitely many bits. *Applied and Computational Harmonic Analysis*, 18(3):215–233, 2005.

[32] H. N. Mhaskar. Polynomial operators and local smoothness classes on the unit interval, ii. *Jaén J. of Approx.*, 1(1):1–25, 2009.

[33] H. N. Mhaskar. Local approximation using Hermite functions. In *Progress in Approximation Theory and Applicable Complex Analysis*, pages 341–362. Springer, 2017.

[34] H. N. Mhaskar, A. Cloninger, and X. Cheng. A witness function based construction of discriminative models using hermite polynomials. *Arxiv preprint, arXiv:1901.02975*, 2019.

[35] H. N. Mhaskar and J. Prestin. On local smoothness classes of periodic functions. *Journal of Fourier Analysis and Applications*, 11(3):353–373, 2005.

[36] J. M. Murphy and M. Maggioni. Unsupervised clustering and active learning of hyperspectral images with nonlinear diffusion. *IEEE Transactions on Geoscience and Remote Sensing*, 57(3):1829–1845, 2018.

[37] V. Satuluri and S. Parthasarathy. Symmetrizations for clustering directed graphs. In *Proceedings of the 14th International Conference on Extending Database Technology*, pages 343–354. ACM, 2011.

[38] B. Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[39] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery. Active learning methods for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 47(7):2218–2232, 2009.

[40] C. Xiong, D. M. Johnson, and J. J. Corso. Active clustering with model-based uncertainty reduction. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):5–17, 2016.

[41] X. Zhu, A. Singla, S. Zilles, and A. N. Rafferty. An overview of machine teaching. *Arxiv preprint arXiv:1801.05927*, 2018.