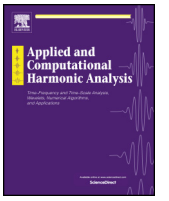
Appl. Comput. Harmon. Anal. 54 (2021) 44–74

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha



# Cautious active clustering

A. Cloninger [a,1], H.N. Mhaskar [b,*,2]

[a] *Department of Mathematics and Halicioglu Data Science Institute, University of California San Diego, San Diego, CA 92093, USA*
[b] *Institute of Mathematical Sciences, Claremont Graduate University, Claremont, CA 91711, USA*

ARTICLE INFO

*Article history:*
Received 5 May 2020
Received in revised form 7 December 2020
Accepted 22 February 2021
Available online 5 March 2021
Communicated by Charles K. Chui

Dedicated to Charles K. Chui on the occasion of his 80th birthday

*Keywords:*
Active clustering
Hierarchical clustering
Hyperspectral pixel classification

ABSTRACT

We consider the problem of classification of points sampled from an unknown probability measure on a Euclidean space. 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, in order 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 and MNIST classification.



## 1. Introduction

The main purpose of this paper is to study a classification problem in the theory of machine learning, which we have called cautious active learning, by modifying certain ideas originating in our previous work on blind source signal separation. In Section 1.1, we describe the classification problem in the theory of machine learning which we are interested in. 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 gives 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.

* Corresponding author.
*E-mail addresses:* acloninger@ucsd.edu (A. Cloninger), hrushikesh.mhaskar@cgu.edu (H.N. Mhaskar).

[1] The research of this author is supported in part by the NSF DMS grants 2012266, 1819222, and Sage Foundation Grant 2196.
[2] The research of this author is supported in part NSF DMS grant 2012355.

https://doi.org/10.1016/j.acha.2021.02.002
1063-5203/

### 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 (e.g., [5,1]). We are given data of the form $\{(\mathbf{x}_i, y_i)\}_{i=1}^M$ where $\mathbf{x}_i$'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 a build a model $P$ such that for any vector $\mathbf{x} \in \mathbb{R}^q$ (or a compact subset thereof), $P(\mathbf{x})$ gives reliably the class to which $\mathbf{x}$ belongs. In semi-supervised learning, the labels $y_i$ are known only for a small number of $\mathbf{x}_i$'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 $\mathbf{x}_i$, 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 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,

$$\mathsf{Prob}(\mathbf{x}, k) = \mathsf{Prob}(k|\mathbf{x})\mathsf{Prob}(\mathbf{x}) = \mathsf{Prob}(\mathbf{x}|k)\mathsf{Prob}(k). \tag{1.1}$$

The first equation leads us to discriminative models. The class $k$ of a given point $\mathbf{x}$ is

$$\arg\max_{k=1,\cdots,K} \mathsf{Prob}(k|\mathbf{x})\mathsf{Prob}(\mathbf{x}).$$

In [37], 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

$$\mathsf{Prob}(\mathbf{x}) = \sum_{k=1}^{K} \mathsf{Prob}(\mathbf{x}|k)\mathsf{Prob}(k). \tag{1.2}$$

In measure theoretic notation, one can write

$$\mu^* = \sum_{k=1}^{K} \mu_k, \tag{1.3}$$

where $\mu^*$ is the marginal distribution from which the points $\mathbf{x}$ are chosen, and $\mu_k$ represents the $k$-th term in (1.2); i.e., some positive measure. (It is understood that $\mu^*$ is a probability measure, and it is not claimed that each $\mu_k$ is a probability measure; indeed, it represents $k$-th summand in (1.2), which accounts for the proportion of samples in class $k$.) The task is to separate the supports of the unknown component measures $\mu_k$ given random samples taken from the unknown probability distribution $\mu^*$. The intuition is that once the supports of each $\mu_k$ are known, we just need one sample from each to complete the task of classification using only the smallest number of samples.

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. We focus on the transductive learning case, in which we are generalizing from the small labeled examples to the specific unlabeled data that is available [46].

We note that in the absence of minimal separation among the supports of the measures $\mu_k$, the decomposition (1.3) is not uniquely defined; e.g., we may group the $K$ measures in many different ways, resulting in $< K$ components. In an unsupervised setting, this is only natural. For example, a data base of images can be classified at different levels as that of an animate or non-animate object, as that of a human, or animal, or movable or unmovable object, etc.; ultimately viewing each image as its own class. Accordingly, our definitions and theorems will not assume a prior knowledge of the number of classes (equivalently, the measures $\mu_k$). The role of active learning is to reconcile this with a given classification problem with increasing confidence. Thus, having separated the clusters at different levels of minimal separation, we seek labels from each of them, and combine or further subdivide them so as to achieve the known number of classes.

### *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 $\mathbf{x}, \mathbf{y} \in \mathbb{T}^q$, we define (in this section only) $|\mathbf{x}-\mathbf{y}| = \max_{1\leq k\leq q} |(x_k - y_k) \bmod 2\pi|$. One formulation of the problem of blind source signal separation is the following. Let $\mu^* = \sum_{k=1}^K a_k \delta_{\mathbf{x}_k}$ be a (signed) measure supported at points $\mathbf{x}_k \in \mathbb{T}^q$, where $\delta_{\mathbf{x}}$ is the Dirac delta measure supported at $\mathbf{x}$. The goal is to recuperate the number $K$ of components, the point sources $\mathbf{x}_k$ and the (signed, complex) amplitudes $a_k$, given the Fourier moments $\widehat{\mu^*}(\mathbf{j}) = \sum_{k=1}^K a_k \exp(-i\mathbf{j}\cdot\mathbf{x}_k)$ for $|\mathbf{j}|_\infty < N$ for some $N$. Our solution to this problem described in [8,32,38] 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 $|t| \geq 1$. For integer $n \geq 1$, we then consider an operator

$$\mathcal{T}_n(f)(\mathbf{x}) = \hbar_n \sum_{\mathbf{j}\in\mathbb{Z}^q} H\left(\frac{|\mathbf{j}|}{n}\right) f(\mathbf{j}) \exp(i\mathbf{j}\cdot\mathbf{x}), \qquad \mathbf{x} \in \mathbb{T}^q,$$

where

$$\hbar_n = \left(\sum_{\mathbf{j}\in\mathbb{Z}^q} H\left(\frac{|\mathbf{j}|}{n}\right)\right)^{-1}.$$

With

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

it is not difficult to verify that

$$\mathcal{T}_n(f)(\mathbf{x}) = \frac{\hbar_n}{(2\pi)^q} \int_{\mathbb{T}^q} \Phi_n^T(\mathbf{x}-\mathbf{y}) d\mu^*(\mathbf{y}). \tag{1.4}$$

The following theorem from [8] 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 $\mathbf{x}_k$ (i.e., $\eta = \min_{1\leq k<j\leq K} |\mathbf{x}_k - \mathbf{x}_j|$), and $\mathfrak{m}$ to denote the minimum of the $|a_k|$'s. Applications of theorems of this sort to direction finding in phased array antennas are discussed in [38].

**Theorem 1.1.** *For sufficiently large $n$ (depending upon $\eta$), the set of $\mathbf{x} \in \mathbb{T}^q$ at which $|\mathcal{T}_n(f)(\mathbf{x})| \geq \mathfrak{m}/2$ is a disjoint union of* ***exactly*** *$K$ sets $\mathcal{G}_k$, $1 \leq k \leq K$, each containing exactly one point $\mathbf{x}_k$, and each with diameter $\leq c/n$ for some positive constant $c$ with each of the following properties. (i) The minimal separation among the sets $\mathcal{G}_k$ is at least $c/n$, (ii) If $\widehat{\mathbf{x}_k}$ is the highest peak of the power spectrum $|\mathcal{T}_n(f)|$ in $\mathcal{G}_k$, then (clearly) $|\widehat{\mathbf{x}_k} - \mathbf{x}_k| \leq c/n$, and (iii) $|\mathcal{T}_n(f)(\widehat{\mathbf{x}_k}) - a_k| \leq c_1/n$.*

The key ingredient in the proof of Theorem 1.1 is the localization estimate

$$|\Phi_n^T(\mathbf{x} - \mathbf{y})| \leq \frac{c(H,S)}{\max(1, (n|\mathbf{x} - \mathbf{y}|)^S)}, \qquad \mathbf{x}, \mathbf{y} \in \mathbb{T}^q, \tag{1.5}$$

where $c(H,S) > 0$ is a constant independent of $n, \mathbf{x}, \mathbf{y}$. We note that $n$ is the degree (order) of the trigonometric polynomial $\Phi_n^T$. In contrast to the kernel estimators in statistics, the localization here is achieved by letting $n \to \infty$.

### *1.3. Separation of measures*

The basic idea in our paper is to use an analogous localized kernel $\Phi_n$ to be defined in (2.12) below (cf. [36,9]) based on Hermite polynomials. It is not difficult to verify using known results about these kernels that $\int_{\mathbb{R}^q} \Phi_n(\mathbf{x}, \mathbf{y}) d\mu^*(\mathbf{y}) \to d\mu^*(\mathbf{x})$ 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

$$\int_{\mathbb{R}^q} \Phi_n(\mathbf{x}, \mathbf{y}) d\mu^*(\mathbf{y}) = \sum_{k=1}^{K} \int_{\mathbb{R}^q} \Phi_n(\mathbf{x}, \mathbf{y}) d\mu_k(\mathbf{y})$$

will split into clusters of $\mathbf{x}$ belonging to the supports of the measures $\mu_k$.

This optimality of approximation of measures however *requires* that the kernel $\Phi_n$ 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 $\Phi_n^2$ 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 two examples.

**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 Fig. 1, we show the results of our method based on a total of 1000 points, with the value of the parameter $n = 7$. 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 given by $n^2$. 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.

We compare our support detection algorithm to a standard Gaussian kernel density estimate with the same kernel bandwidth in Fig. 1. As a comparison, we display an indicator function of whether the density estimate is greater than $\frac{1}{4}$ the maximum peak of the density estimate,

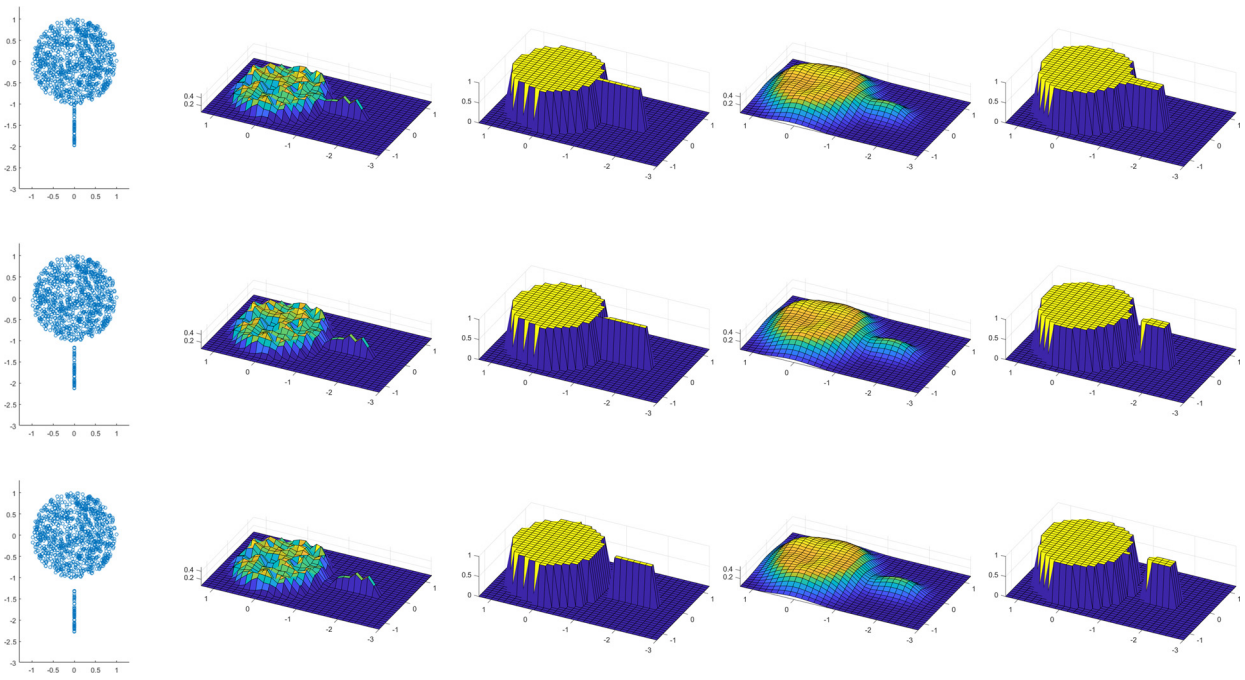


**Fig. 1.** (Column 1) Data points, (Column 2) $\Phi_n^2$ density approximation for $n = 7$, (Column 3) Indicator of whether $\mathbb{E}_{x_i}[\Phi_n^2(x, x_i)] > 0.25 \cdot \max(\mathbb{E}_{x_i}[\Phi_n^2(x, x_i)])$, (Column 4) Gaussian kernel density estimate, (Column 5) Indicator of whether $KDE(x) > 0.25 \cdot \max(KDE(x))$. (Top) No gap between different clusters, (Middle) Small gap between different clusters, (Bottom) Large gap between different clusters. For all kernels, the bandwidth $\sigma = 0.25$ was chosen to be half the median distance between points.

$$\frac{1}{M}\sum_{i=1}^{M}\Phi_n^2(x, x_i) > 0.25 \cdot \max_x \left(\frac{1}{M}\sum_{i=1}^{M}\Phi_n^2(x, x_i)\right). \tag{1.6}$$

This is a proxy for the estimate of the support of the clusters. It is clear from the figures that our $\Phi_n^2$ kernel defines a sharper boundary for the indicator function, detecting all points regions within the support of the density while still maintaining a separation between the clusters. Similarly, the support estimate from our kernel provides a tighter estimate normal to the 1D line than a Gaussian KDE, and captures a better estimate of the minimal separation. □

**Example 1.2.** We consider a mixture of two distributions in the so-called two moons data set. These are point clouds concentrated near one-dimensional manifolds and the clouds are not linearly separable. In Fig. 2 we show the results of our method with 1000 points, with the value of the parameter $n = 6$. Our method estimates the relative supports of the two distributions, and demonstrates a minimal separation between the distributions. On top of this, we show the selection of only two well chosen labeled points leads to perfect classification of the entire data set using our algorithm. □

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 continua. In turn, this requires the coefficients of these constituent distributions to be positive.
2. Instead of having values of $f(\mathbf{j})$, 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.

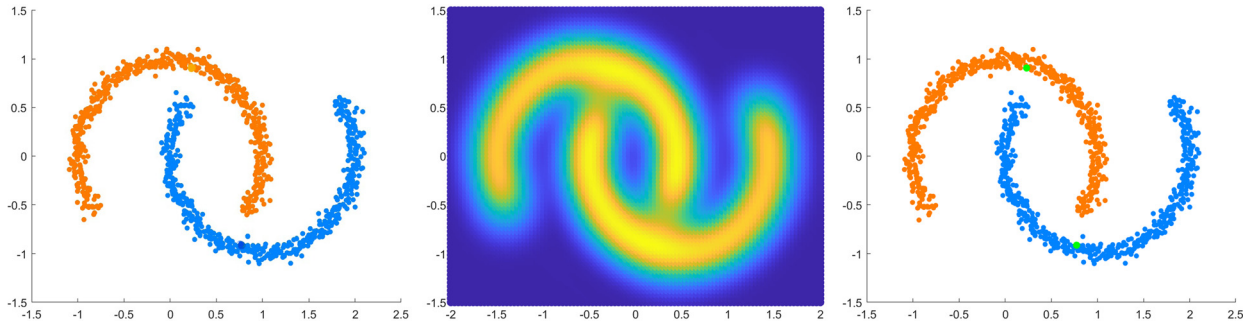


**Fig. 2.** (Left) Data points, (Center) Density Approximation for $n = 6$, (Right) Class prediction using our method with two labeled points. Labeled points are highlighted in green. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

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. The problem is inherently ill-posed; the number of constituent measures may be different at different levels of minimal separation.
5. There is no analogue of minimal magnitude $\mathfrak{m}$ 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 continua.

### *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 [25,28]. 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 [17] that the notion of a cluster (in an unsupervised setting) needs be defined hierarchically. We follow the philosophy in [17] 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 [31,45] rather than low-complexity classifiers [22,23]. 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 [43,47,29].

A number of works by Dasgupta and his collaborators [14,4,15,16] 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 [15], 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 [33], which have been applied in a variety of contexts [38,35,34,30,9,7,12]. Localized kernels also play a central role in the determination of components of a blind source signal, whether stationary or time-variant [8,11,10,9]. The super-resolution problem has been considered in a hierarchical context [27]. 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 [6,37], 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 [37] 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 [37] 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 [31], 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 [31], 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 [31], and to the hyperspectral imaging variant [40], 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 $\mathbf{x} = (x_1, \cdots, x_q) \in \mathbb{R}^q$, we denote by $|\cdot|_p$ the $\ell^p$ norm of $\mathbf{x}$. For $\mathbf{x} \in \mathbb{R}^q$, $r > 0$, we denote

$$\mathbb{B}(\mathbf{x}, r) = \{\mathbf{y} \in \mathbb{R}^q : |\mathbf{x} - \mathbf{y}|_\infty \leq r\}. \tag{2.1}$$

If $A, B \subseteq \mathbb{R}^q$, $\mathbf{x} \in \mathbb{R}^q$, $r > 0$, then

$$\mathsf{dist}(\mathbf{x}, A) = \inf_{\mathbf{y} \in A} |\mathbf{x} - \mathbf{y}|_\infty, \ \mathbb{B}(A, r) = \{\mathbf{x} \in \mathbb{R}^q : \mathsf{dist}(\mathbf{x}, A) \leq r\}, \ \mathsf{dist}(A, B) = \inf_{\mathbf{x} \in A} \mathsf{dist}(\mathbf{x}, B). \tag{2.2}$$

In Section 2.1, we describe the measure theoretic notions used in this paper. In Section 2.2, we develop a measurement to test the accuracy our classification algorithms. The bare minimum definition of our localized kernel is developed in Section 2.3; the details of the properties of this kernel will be developed in Section 6.

### *2.1. Measures*

The term measure will mean a positive, Borel measure on $\mathbb{R}^q$. The support of a measure $\mu$, denoted by $\mathsf{supp}(\mu)$ is the set of all $\mathbf{x} \in \mathbb{R}^q$ for which $\mu(\mathbb{B}(\mathbf{x}, r)) > 0$ for all $r > 0$. We will fix a probability measure $\mu^*$ on $\mathbb{R}^q$.

We will use the following convention regarding generic positive constants.

**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 depends 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 $\mathbb{R}^q$.

(a) 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,
$$\mathsf{supp}(\mu^*) \subseteq \{\mathbf{x} \in \mathbb{R}^q : |\mathbf{x}|_\infty \le C\}.$$
2. (**Ball measure condition**) There exist $C_2$ and $\alpha \ge 0$ such that
$$\mu^*(\mathbb{B}(\mathbf{x}, r)) \le C_2 r^\alpha. \tag{2.3}$$
3. (**Density condition**) There exist $C_1 > 0$, $r_0 > 0$ such that (with $\alpha$ as in (2.3),
$$\mu^*\left(\mathbb{B}(\mathbf{x}, r)\right) \ge C_1 r^\alpha, \qquad \mathbf{x} \in \mathsf{supp}\ (\mu^*),\ 0 < r \le r_0. \tag{2.4}$$

(b) The measure $\mu^*$ is said to have **fine structure** if it is detectable, and there exists $\eta_0 > 0$ with the property that for every $\eta \in (0, \eta_0]$, there is in integer $K_\eta \ge 1$ and a partition $\mathbf{S}_{k,\eta}$, $k = 1, \cdots, K_\eta + 1$ of $\mathsf{supp}(\mu^*)$ such that each of the following conditions is satisfied.
1. (**Cluster minimal separation condition**)
$$\mathsf{dist}(\mathbf{S}_{k,\eta}, \mathbf{S}_{j,\eta}) \ge 2\eta, \qquad k \ne j,\ k, j = 1, \cdots, K_\eta. \tag{2.5}$$
2. (**Exhaustion condition**)
$$\lim_{\eta \downarrow 0} \mu^*(\mathbf{S}_{K_\eta+1,\eta}) = 0.$$

(c) The $\mu^*$ is said to have **fine structure in the classical sense** if it has a fine structure with $K_\eta = K$ for all $\eta \le \eta_0$, and there are measures $\mu_k$ such that $\mu^* = \sum_{k=1}^K \mu_k$ such that each $\mathbf{S}_{k,\eta} \subseteq \mathsf{supp}\ (\mu_k)$, $k = 1, \cdots, K$.

**Remark 2.1.** Necessarily, a non-atomic, detectable probability measure $\mu^*$ has fine structure. The qualification of measures having fine structure in the classical sense is not an intrinsic property of such measures, but depends upon how the points in $\mathsf{supp}\ (\mu^*)$ are labeled. In the **unsupervised setting** in the absence of minimal separation among various clusters, the notion of fine structure still enables us to construct reliable labels for various clusters by assigning at each level $\eta$, the label $k$ with $\mathbf{S}_{k,\eta}$, $k = 1, \cdots, K_\eta$. The exhaustion condition requires that the $\mu^*$-measure of the support of the "left-over" part of $\mathsf{supp}\ (\mu^*)$ tends to 0 as $\eta \to 0$. □

**Example 2.1.** Let $\mu^* = \sum_{k=1}^K a_k \delta_{\mathbf{x}_k}$, where $a_k$'s are positive, $\sum_k a_k = 1$, and $\mathbf{x}_k \in \mathbb{R}^q$. Then $\mu^*$ is compactly supported, and the ball measure condition is satisfied with $C_2 = 1$ and $\alpha = 0$. The density condition is satisfied with $C_1 = \min_k a_k$. It is easy to verify that $\mu^*$ has fine structure in the classical setting, with $\eta_0 = (1/2) \min_{j \ne k} |\mathbf{x}_j - \mathbf{x}_k|_\infty$. □

**Example 2.2.** With the setting as in Section 1.2, we let $q = 1$, write $\mathbb{T} = \mathbb{T}^1$, for $x \in \mathbb{T}$, $r > 0$. For $x \in \mathbb{T}$, $k \ge 1$, we let $g_k(x) = (2/\pi)\chi_{\mathbb{B}(\pi/2^k, \pi/2^{k+1}]}$, and define $\mu^*$ by $d\mu^*(x) = \left(\sum_{k=1}^\infty g_k(x)\right) dx$. If $x \in \mathsf{supp}\ (\mu^*)$,

it is not hard to verify that for any $r \in (0, \pi/8]$, $cr \le \mu^*(\mathbb{B}(x,r)) \le 2r$ for a suitable constant $c > 0$. So, $\mu^*$ is detectable. If $L \ge 1$ is an integer, and $\eta = \pi/2^{L+2}$, we may write $K_\eta = L$, and consider the partition of the support of $\mu^*$ given by $\mathbb{B}(\pi/2^k, \pi/2^{k+1}]$, $k = 1, \cdots, L$, with $\mathbf{S}_{K_\eta+1}$ equal to the remainder of the support. Clearly $\mu^*$ has a fine structure. Since the number of elements in the partition depends upon $\eta$, this is the unsupervised setting. On the other hand, we may view this also as a classical setting with $K = 2$ as follows. Let $\mu_1$, $\mu_2$ be defined by $d\mu_1(x) = \left(\sum_{k=1}^{\infty} g_{2k}(x)\right) dx$, $d\mu_2(x) = \left(\sum_{k=1}^{\infty} g_{2k-1}(x)\right) dx$. For integer $L \ge 2$, let $\mu_{1,L}$, $\mu_{2,L}$ be defined by $d\mu_{1,L}(x) = \left(\sum_{k=1}^{L} g_{2k}(x)\right) dx$, $d\mu_{2,L}(x) = \left(\sum_{k=1}^{L} g_{2k-1}(x)\right) dx$. Then $\mathsf{dist}(\mathsf{supp}\ (\mu_{1,L}), \mathsf{supp}\ (\mu_{2,L})) = \pi/2^{2L+1}$, and

$$\lim_{L\to\infty} \mu^* \left(\mathbb{T} \setminus (\mathsf{supp}\ (\mu_{1,L}) \cup \mathsf{supp}\ (\mu_{2,L}))\right) = 0.$$

Thus, only queries about labels on points can decide whether to interpret a measure with fine structure in the unsupervised or classical setting. □

**Example 2.3.** With the setting as in Example 2.2, let $\mu^* = \sum_{k=1}^{\infty} 2^{-k}\delta_{\pi/2^k}$. Then $\mu^*$ satisfies the compact support condition as well as the ball measure condition with $\alpha = 0$, but not with any $\alpha > 0$. The density condition is also not satisfied with $\alpha = 0$. Thus, $\mu^*$ is not detectable. □

**Example 2.4.** Let $K \ge 2$, $\mathbb{X}_k$, $k = 1, \cdots, K$ be mutually disjoint, compact, smooth, sub-manifolds (without boundary) of $\mathbb{R}^q$ each having the dimension $d$. Let $\mu_k$ be the volume measure of $\mathbb{X}_k$, and $\mu^* = \sum_{k=1}^{K} \mu_k$ normalized to be a probability measure. For a large class of manifolds $\mathbb{X}_k$, it is known that there exist constants $c_1, c_2 > 0$ such that for any point $\mathbf{x} \in \mathbb{X}_k$ and $r > 0$, $c_1 r^d \le \mu_k(\mathbb{B}(\mathbf{x}, r)) = \mu^*(\mathbb{B}(\mathbf{x}, r)) \le c_2 r^d$. Therefore, $\mu^*$ is detectable. It is trivial to see that $\mu^*$ has a fine structure in the classical setting. □

### *2.2. F-score*

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 components $\mathbf{S}_{k,\eta}$. If we attach the label $k$ with every data point in $\mathbf{S}_{k,\eta}$, then we need to discuss a measurement to assess the quality of our algorithms as classification tools. We recall the $F$-score described for a finite data set in [42]. If $\{C_1, \cdots, C_N\}$ are the obtained clusters from a certain clustering algorithm, and $\{L_1, \cdots, L_K\}$ 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

$$F_D(C_j) = 2 \max_{1\le k\le K} \frac{|C_j \cap L_k|}{|C_j| + |L_k|}, \qquad j = 1, \cdots, N.$$

The (micro-averaged) $F$-score is then defined by

$$\mathcal{F}_D = \frac{\sum_j |C_j| F_D(C_j)}{\sum_j |C_j|}. \tag{2.6}$$

We interpret the cardinalities above as probabilities. In this paper, the total data is $\mathsf{supp}(\mu^*)$; the labeled sets (corresponding to $L_k$'s above) at separation level $\eta$ is $\{\mathbf{S}_{k,\eta}\}_{k=1}^{K_\eta}$. Therefore, the $F$-score for the clusters $\{\mathcal{C}_j\}_{j=1}^{N}$ can be defined as follows: The analogue of $F_D(C_j)$ above is:

$$F_\eta(\mathcal{C}_j) = 2 \max_{1\le k\le K} \frac{\mu^*(\mathcal{C}_j \cap \mathbf{S}_{k,\eta})}{\mu^*(\mathcal{C}_j) + \mu^*(\mathbf{S}_{k,\eta})}. \tag{2.7}$$

Then

$$\mathcal{F}_\eta\left(\{\mathcal{C}_j\}_{j=1}^N\right) = \frac{\sum_{j=1}^N \mu^*(\mathcal{C}_j)F(\mathcal{C}_j)}{\mu^*\left(\bigcup_{j=1}^N \mathcal{C}_j\right)}. \tag{2.8}$$

Clearly, $\mathcal{F}_\eta\left(\{\mathcal{C}_j\}_{k=1}^N\right) \le 1$. If $N = K$, and $\mathcal{C}_k = \mathbf{S}_{k,\eta}$ for each $k$, then $\mathcal{F}_\eta\left(\{\mathcal{C}_j\}_{j=1}^N\right) = 1$. Thus, the closer the quantity $\mathcal{F}_\eta$ is to 1, the better the quality of clustering with respect to the labels.

### *2.3. Localized kernel*

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

$$\begin{aligned} xh_{j-1}(x) &= \sqrt{\frac{j}{2}}h_j(x) + \sqrt{\frac{j-1}{2}}h_{j-2}(x), \quad j = 2, 3, \cdots, \\ h_0(x) &= \pi^{-1/4},\ h_1(x) = \sqrt{2}\pi^{-1/4}x. \end{aligned} \tag{2.9}$$

Writing $\psi_k(x) = h_k(x)\exp(-x^2/2)$, one has the orthogonality relation for $k, j \in \mathbb{Z}_+$,

$$\int_{\mathbb{R}} \psi_k(x)\psi_j(x)dx = \begin{cases} 1, & \text{if } k = j, \\ 0, & \text{if } k \ne j. \end{cases} \tag{2.10}$$

In multivariate case, we adopt the notation $\mathbf{x} = (x_1, \cdots, x_q)$. The orthonormalized Hermite function is defined by

$$\psi_{\mathbf{k}}(\mathbf{x}) = \prod_{j=1}^q \psi_{k_j}(x_j). \tag{2.11}$$

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

Let $H : [0, \infty) \to [0, 1]$ be a $C^\infty$ function, $H(t) = 1$ if $t \in [0, 1/2]$, $H(t) = 0$ if $t \ge 1$. We define the *localized kernel* by

$$\Phi_n(H; \mathbf{x}, \mathbf{y}) = \Phi_n(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{k}\in\mathbb{Z}_+^q} H\left(\frac{\sqrt{|\mathbf{k}|_1}}{n}\right)\psi_{\mathbf{k}}(\mathbf{x})\psi_{\mathbf{k}}(\mathbf{y}). \tag{2.12}$$

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

## 3. Main theorems

In Section 1.2, the quantity $\mathfrak{m}$ 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 played by the density condition. 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 determination of the support of $\mu^*$. For the sake of interpretability, we briefly discuss the important parameters associated with this theorem. We fix $n$, which corresponds to the localization parameter for $\Phi_n$, and leads to an eventual clustering of all points as $n \to \infty$. We set a threshold

$\Theta$ to determine the level that a density estimate at $\mathbf{x}$ must attain for the fixed $n$ to be considered in the subsequent clustering. The parameter $\alpha$ corresponds to the effective dimension of the support of $\mu^*$, and determines how $n$ relates to the number of points $M$, the detectable minimal separation between clusters, and algorithmically determines how slowly we increase the localization of $\Phi_n$ and the number of points retained for that level of clustering. These are the main parameters that appear in Algorithm 1, though we use $\eta_n$ rather than $\alpha$ as a parameter by utilizing Eq. (3.4).

**Theorem 3.1.** *Let $\mu^*$ be detectable, $S > \alpha$, $0 < \Theta \leq c_1$, $n \geq c_2$ be large enough, so that*

$$\text{supp } (\mu^*) \subseteq \mathbb{B}(\mathbf{0}, \kappa n). \tag{3.1}$$

*With $M \geq c_3 n^{2\alpha} \log n$, let $\mathcal{C} = \{\mathbf{x}_1, \cdots, \mathbf{x}_M\}$ be independently sampled from the probability distribution $\mu^*$. We define*

$$\mathcal{G}_n(\Theta, \mathcal{C}) = \left\{ \mathbf{x} \in \mathbb{R}^q : \sum_{j=1}^{M} \Phi_n(\mathbf{x}, \mathbf{x}_j)^2 \geq \Theta \max_{1 \leq k \leq M} \sum_{j=1}^{M} \Phi_n(\mathbf{x}_k, \mathbf{x}_j)^2 \right\}. \tag{3.2}$$

*Then with probability at least $1 - c_4/M^{c_5}$,*

$$\text{supp } (\mu^*) \subseteq \mathcal{G}_n(\Theta, \mathcal{C}) \subseteq \left\{ \mathbf{x} \in \mathbb{R}^q : \text{dist}(\mathbf{x}, \text{supp } (\mu^*)) \leq \frac{c_6}{\Theta^{1/(S-\alpha)} n} \right\}. \tag{3.3}$$

**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*

$$n \geq c(\eta\Theta)^{-1}, \;\; n^\alpha \mu^*(\mathbf{S}_{K_\eta+1,\eta}) \leq c_1 \Theta. \tag{3.4}$$

*Let*

$$\mathcal{G}_{k,\eta,n}(\Theta, \mathcal{C}) = \mathcal{G}_n(\theta, \mathcal{C}) \cap \left\{ \mathbf{x} \in \mathbb{R}^q : \text{dist}(\mathbf{x}, \mathbf{S}_{k,\eta}) \leq \frac{c_2}{n\Theta^{1/(S-\alpha)}} \right\}. \tag{3.5}$$

*Then with probability exceeding $1 - c_3 M^{-c_4}$, 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*

$$\text{dist}(\mathcal{G}_{k,\eta,n}(\Theta, \mathcal{C}), \mathcal{G}_{j,\eta,n}(\Theta, \mathcal{C})) \geq \eta, \qquad k \neq j, \;\; k, j = 1, \cdots, K_\eta, \tag{3.6}$$

*and for $k = 1, \cdots, K_\eta$,*

$$\text{supp } (\mu^*) \cap \left\{ \mathbf{x} \in \mathbb{R}^q : \text{dist}(\mathbf{x}, \mathbf{S}_{k,\eta}) \leq \frac{c_2}{n\Theta^{1/(S-\alpha)}} \right\} \subseteq \mathcal{G}_{k,\eta,n}(\Theta, \mathcal{C}) \subseteq \left\{ \mathbf{x} \in \mathbb{R}^q : \text{dist}(\mathbf{x}, \mathbf{S}_{k,\eta}) \leq \frac{c_2}{n\Theta^{1/(S-\alpha)}} \right\}. \tag{3.7}$$

Theorems 3.1 and 3.2 combine to guarantee that, for a fixed $n$, the training data kept at a given threshold will be contained in a small tube around the true partitions $S_{k,\eta}$. Similarly, points belonging to different clusters will have a minimal separation of at least $\eta$ (half the minimal separation of the true partitions). This is a critical aspect to our algorithm, as we use a simple distance parameter to build a kNN network between training points, and use connected components to define the clusters. Theorem 3.2 guarantees that if we correctly choose this distance parameter, no connected component will span multiple true partitions.

**Example 3.1.** We continue the set-up as in Example 2.4. Then for $\eta \le \eta_0 = \min_{1\le k\ne j\le K} \mathsf{dist}(\mathbb{X}_k, \mathbb{X}_j)$, 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 $\eta \le \eta_0$. 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 2.1. In this case, the definition of $\mathcal{G}_{k,\eta,n}(\Theta)$ shows that the diameter of each of these sets is $\le 2\Theta/n$. 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 $|\hat{\mathbf{x}}_k - \mathbf{x}_k|_\infty \le 2\Theta/n$. We note finally that the sum expression in the above expression can be computed using the Hermite moments of $\mu^*$; the precise location of $\mathbf{x}_k$'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 [9]. □

**Remark 3.1.** In a broad sense, we may view $(1/M)\sum_{j=1}^{M} \Phi_n(\mathbf{x}, \mathbf{x}_j)^2$ as a probability density estimator with kernel $\Phi_n(\mathbf{x}, \mathbf{y})^2$ that localizes as $n \to \infty$. However, there are a number of important differences. First, the measure $\mu^*$ is allowed to be singular with respect to the Lebesgue measure on $\mathbb{R}^q$, so that there is no density to estimate. Second, we have demonstrated in [37] that in the case when $\mu^*$ is absolutely continuous, the use of the special kernel $\Phi_n$ leads to a better estimation of the density than the customary Gaussian kernel, with sharper boundaries and faster empirical convergence in terms of the number of points. Indeed, in applying this theorem, we may use both the bandwidth parameter as in the Gaussian kernel as well as the degree parameter $n$ to have a superior localization property. Third, in this paper we are not interested in approximating the measures themselves, just in finding their support. For the purpose of approximation, the oscillatory behavior of $\Phi_n$ leads to provably optimal approximation results. However, this very ability may lead to certain points in the support having zero values for the estimator. Therefore, to find the support accurately, we need to use a positive kernel. We use $\Phi_n(\mathbf{x}, \mathbf{y})^2$ because of the ease of evaluating certain integrals. □

**Theorem 3.3.** *We assume the set-up as in Theorem 3.2. In addition, we assume that*

$$\lim_{\eta\downarrow 0} \frac{\mu^*(\mathbf{S}_{K_\eta+1,\eta})}{\min_{1\le k\le K_\eta} \mu^*(\mathbf{S}_{k,\eta})} = 0. \tag{3.8}$$

*With probability $\ge 1 - cM^{-c_1}$, the clusters $\mathcal{G}_{k,\eta,n}(\Theta, \mathcal{C})$ satisfy*

$$\lim_{\eta\to 0} \mathcal{F}_\eta \left( \{\mathcal{G}_{k,\eta,n}(\Theta, \mathcal{C})\}_{k=1}^{K_\eta} \right) = 1. \tag{3.9}$$

**Remark 3.2.** In the classical setting (Remark 2.1), $K_\eta = K$ for all $\eta$, and $\mathbf{S}_{k,\eta} \subseteq \mathsf{supp}\ (\mu_k)$. The exhaustion condition implies that $\mu^*(\mathsf{supp}\ (\mu_k) \setminus \mathbf{S}_{k,\eta}) \to 0$ as $\eta \to 0$. In particular, the condition (3.8) holds trivially. Moreover, in the definition (2.7), we may replace $\mathbf{S}_{k,\eta}$ by $\mathsf{supp}\ (\mu_k)$. Thus, Theorem 3.3 guarantees that with high probability, the clusters of the high-density regions from each $n$ will grow to contain the entire $\mathrm{supp}(\mu_k)$ for all $k$ and attain a perfect $F$-score (Eq. (2.8)). □

**Remark 3.3.** 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 3.4.** 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 (3.9),

$$\lim_{n\to\infty} \mathcal{F}\left(\{\mathcal{S}_{k,\eta,n}(\theta)\}_{k=1}^{K_\eta}\right) = 1. \tag{3.10}$$

## 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.

We will describe the algorithm assuming that the data distribution $\mu^*$ has a fine structure in the classical sense defined in Definition 2.1(c); i.e., we assume that there is a fixed set of labels involved, that does not change as the various parameters $n$, $\eta$, $\Theta$ change. 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.

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 [37]. 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. Finally, in Section 4.4 we discuss the computational complexity of the algorithm.

### *4.1. Connecting points in $\mathcal{G}_{k,\eta,n}$*

In the classical setting, we may assume that there exists some unknown label function $f : \mathbb{R}^q \to \mathbb{R}$. Further, we assume it corresponds to a consistent clustering scheme such that, for small enough $\eta$ and $\mathbf{x} \in \mathbf{S}_{k,\eta}$, we have that $f(\mathbf{x}) = k$ for $1 \le k \le K$. The problem of active learning boils down to learning an estimate $\widehat{f}(\mathbf{x})$ of $f(\mathbf{x})$ for all $\mathbf{x} \in \bigcup_{k=1}^{K} \mathbf{S}_{k,\eta}$ given only a small set of points $\mathcal{A} \subset \mathcal{C}$ at which $f$ is actually known. The key difference between this and semi-supervised learning is that, in our case, $\mathcal{A}$ can be chosen in a data-dependent fashion prior to querying the function. We note that the choice of label function $f$ may be any layer of some hierarchical tree of labels [18], but we assume that $f$ is fixed at the start of the algorithm.

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

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

The insight for constructing the clusters $\mathcal{G}_{k,\eta,n}$ comes from three observations: (1) $\mu^*$ has a fine structure of minimal separation of $2\eta$ for some $\eta$, (2) the theoretical guarantee that the clusters $\mathcal{G}_{k,\eta,n}$ satisfy (3.6) for a finite $n$, and (3) we can decrease $\eta$ by increasing $n$ as in (3.4). For this reason, we will fix $\Theta$ and consider a minimal separation $\eta_n$ that depends on $n$ in a way that satisfies (3.4). Given this separation, we construct a

nearest neighbor graph $G$ on $\mathcal{G}_n(\Theta, \mathcal{C})$ with edge set $E = \{(\mathbf{x}, \mathbf{y})\} \in \mathcal{G}_n(\Theta, \mathcal{C}) : |\mathbf{x} - \mathbf{y}|_2 < \eta_n/2\}$. In this way, we are guaranteed that each connected component $C_{\ell,\eta_n,n} \subset \mathcal{C}$ actually satisfies $C_{\ell,\eta_n,n} \subset \mathcal{G}_{k,\eta_n,n}(\Theta, \mathcal{C})$ for some $k$. This implies that if we query and obtain a label $f(\mathbf{x})$ at some point $\mathbf{x} \in C_{\ell,\eta_n,n}$, then we are guaranteed that all other points in $C_{\ell,\eta_n,n}$ also have label $f(\mathbf{x})$. Note that the number of connected components, which we'll call $\widetilde{K}_n$, satisfies $\widetilde{K}_n \geq K$. 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 $\mathcal{A}$ 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.,

$$\mathbf{x}^* = \arg \max_{\mathbf{x} \in C_{\ell,\eta_n,n}} \sum_{j=1}^{M} \Phi_n^2(x, x_j). \tag{4.1}$$

The argument for this choice is a heuristic one; if there do exist points in $C_{\ell,\eta_n,n}$ 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 is not an issue in the continuum limit due to Theorem 3.3, we are constrained in most applications 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 $S_{K+1}$.

We propose to classify these additional points through the witness function approach developed in [37]. To summarize in this context, we define each class estimate to be $\widehat{S}_{k,\eta_n} = \{x_i \in \mathcal{C} : \widehat{f}(x_i) = k\}$. Then we construct a witness function for each class

$$\widehat{F}_k(\mathbf{x}) = \frac{1}{|\widehat{S}_{k,\eta_n}|} \sum_{\mathbf{x}_j \in \widehat{S}_{k,\eta_n}} \Phi_n(\mathbf{x}, \mathbf{x}_j), \text{ for } \mathbf{x} \in \mathcal{C} \setminus \mathcal{G}_n(\Theta, \mathcal{C}). \tag{4.2}$$

The proposed algorithm assigns a label to $\mathbf{x}$ given by

$$\widehat{f}(\mathbf{x}) = \arg \max_{1 \leq k \leq K} \widehat{F_k}(\mathbf{x}), \tag{4.3}$$

and determines the certainty of classification through a permutation test as in [37]. Note that we will refer to the points assigned in $\mathcal{G}_n(\Theta, \mathcal{C})$ and labeled according to Section 4.1 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 $n \to \infty$.

### *4.3. Learning across layers*

As described to this point, the algorithm for learning $\widehat{f}$ 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 $j > i$.

**Algorithm 1** Cautious active clustering.

**Input:** $\mathcal{C} \subset \mathbb{R}^q$, $nmax$, $\Theta$, $\tau$
**Output:** $\widehat{f}$, $\mathcal{G}_{nmax}(\Theta, \mathcal{C})$

$\mathcal{A} \leftarrow \emptyset$ ▷ Set of points at which label is queried, together with the corresponding labels.
**while** $n \leq nmax$ **do**
  $\mathcal{C}_{n,\Theta} \leftarrow \mathcal{G}_n(\Theta, \mathcal{C}) \cap \mathcal{C}$ using (3.2).
  $E \leftarrow \{(x_i, x_j) : x_i, x_j \in \mathcal{C}_{n,\Theta}$ and $|x_i - x_j|_2 < \eta_n/2\}$ for $\eta_n$ as a function of $n$ as in (3.4)
  Construct graph $G = (\mathcal{C}_{n,\Theta}, E)$
  $\{C_{n,\ell}\}_{\ell=1}^{\widetilde{K}_n} \leftarrow$connectedComponents$(G)$ (See Section 4.1)
  Set flag$(\ell) = 0$ for all $\ell$ ▷ When we are done with this $n$, flag$(\ell) = 1$ for all $\ell$.
  **for** $\ell \leq \widetilde{K}_n$ **do**
    **if** $C_{n,\ell} \cap \mathcal{A} = \emptyset$ **then**
      $x_i \leftarrow \arg\max_{x_i \in C_{n,\ell}} \sum_{j=1}^{M} \Phi_n(x_i, x_j)^2$ ▷ Point at which label is sought
      $\mathcal{A} \leftarrow \mathcal{A} \cup (x_i, f(x_i))$ ▷ Update $\mathcal{A}$; this set does not lose points.
      $\widehat{f}(x_j) \leftarrow f(x_i)\ \forall x_j \in C_{n,\ell}$ ▷ Extend label to the whole component
      flag$(\ell) = 1$ ▷ Done for this value of $\ell$.
    **else**
      **if** $\forall x_i \in C_{n,\ell} \cap \mathcal{A},\ f(x_i) = c_\ell$ **then**
        $\widehat{f}(x_j) \leftarrow c_\ell\ \forall x_j \in C_{n,\ell}$ ▷ Extend label to the whole component
        flag$(\ell) = 1$ ▷ Done for this value of $\ell$.
      **end if**
    **end if**
  **end for**
  **if** flag$(\ell) = 1$ for all $\ell$, **then**
    $n \leftarrow n + \text{step}$ ▷ Done for this pass, go to next level
    ▷ to ensure that we captured all points that could be captured.
  **else**
    Increase threshold $\Theta \leftarrow \tau\Theta$, (See Section 4.3). ▷ Prune the graph.
  **end if**
**end while**
$\mathcal{C}_{\widetilde{K}_{nmax}+1} \leftarrow \mathcal{C} \setminus \mathcal{G}_{nmax}(\Theta_{nmax}, \mathcal{C})$ using (3.2) ▷ Uncertain points
$\widehat{S}_{k,\eta_{nmax}} \leftarrow \{x_i : \widehat{f}(x_i) = k\}$
$\widehat{f}(x_j) \leftarrow \arg\max_k \frac{1}{|\widehat{S}_{k,\eta_{nmax}}|} \sum_{x_i \in \widehat{S}_{k,\eta_{nmax}}} \Phi_{nmax}(x_j, x_i)$ for $x_j \in \mathcal{C}_{\widetilde{K}_{nmax}+1}$
▷ Extend labels to uncertain points using witness function (See Section 4.2).

Let $\mathcal{A}_i \subset \bigcup_{\ell=1}^{\widetilde{K}_{n_i}} C_{\ell,\eta_i,n_i}$ be the small collection of points at which $f$ was sampled. By definition of $\mu^*$ being detectable, $\mathcal{A}_i \subset \bigcup_{\ell=1}^{\widetilde{K}_{n_j}} C_{\ell,\eta_j,n_j}$ as well for $j > i$. This means that many of the connected components $\{C_{\ell,\eta_j,n_j}\}_{\ell=1}^{\widetilde{K}_{n_j}}$ already have a member $\mathbf{x} \in C_{\ell,\eta_j,n_j}$ such that $\mathbf{x} \in \mathcal{A}_i$. Thus, we must only sample the $\widetilde{K}_{n_j} - \widetilde{K}_{n_i}$ 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 $\mathcal{G}_{n_j}(\Theta, \mathcal{C})$ sufficiently. Fortunately, this can be easily detected in the situation that subsets were disconnected at $n_i$ for $j > i$. In this situation, $\mathcal{A}_i$ will contain two points $\mathbf{x}, \mathbf{y}$ with different labels from level $n_i$ such that $\mathbf{x}, \mathbf{y} \in C_{\ell,\eta_j,n_j}$. When this occurs, it is a simple fix to slightly increase $\Theta$ until $\mathbf{x}$ and $\mathbf{y}$ fall in different clusters. This can be done with a parameter $\tau > 1$ that is described in Algorithm 1, 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 $\eta_j$ and rerun the algorithm.

As a final note, it's possible to increase $\Theta$ or decrease $\eta$ only for points in $C_{\ell,\eta_j,n_j}$ rather than on all $\mathcal{C}$. 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 $\mathcal{G}_n(\Theta, \mathcal{C})$.

### 4.4. Computational complexity

The computational complexity of this algorithm depends mostly on construction of the kernel $\Phi_n$ at multiple $n$. The recurrence formula (2.9) enables us to compute $\psi_n$ for any $n$, keeping in storage only the values of $\psi_{n-2}$ and $\psi_{n-1}$. This means the computation depends only on the largest $n$, and depends quadratically in the number of points $M$, as do most kernel methods. This leads to a complexity of $O(n^2M^2q)$. This can be precomputed, and is followed by Equation (A.3), which is complexity $O(n^4M^2)$ per projection, resulting in a final computation for the kernel with $O(qn^6M^2)$ flops. We note that $\Phi_n$ is a $q$-variate weighted polynomial of total degree $n^2$ in each variable. Therefore, one expects a complexity of $\mathcal{O}(n^{2q}M^2)$ in the computation, as it would be if a monomial or tensor product Chebyshev polynomial basis was used to construct the kernel. However, the special property of Hermite polynomials given in Equation (A.3) enables us to evaluate the kernel with complexity linear in $q$. In practice and in Section 5, the $n$ can remain quite small while still resulting in significantly improved performance. As with most kernel methods, this can be improved below a quadratic dependence on $M$ using nearest neighbor or $\varepsilon$-ball algorithms to estimate where contribution will be negligible. Beyond this computation, the most expensive aspect is the connected components computation after thresholding $\mathcal{C}_{n,\Theta}$, which is $O(M + M^2)$ for most connected component algorithms because the number of edges is proportional to $M^2$. All other aspects depend linearly only on the number of connected components $\widetilde{K}_n$, which is at most $M$ and in practical terms significantly smaller.

## 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 based on 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 [31], 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 the first example in Fig. 3, we use the 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 Fig. 3 have long tails of low density.

In a second example in Fig. 4, 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

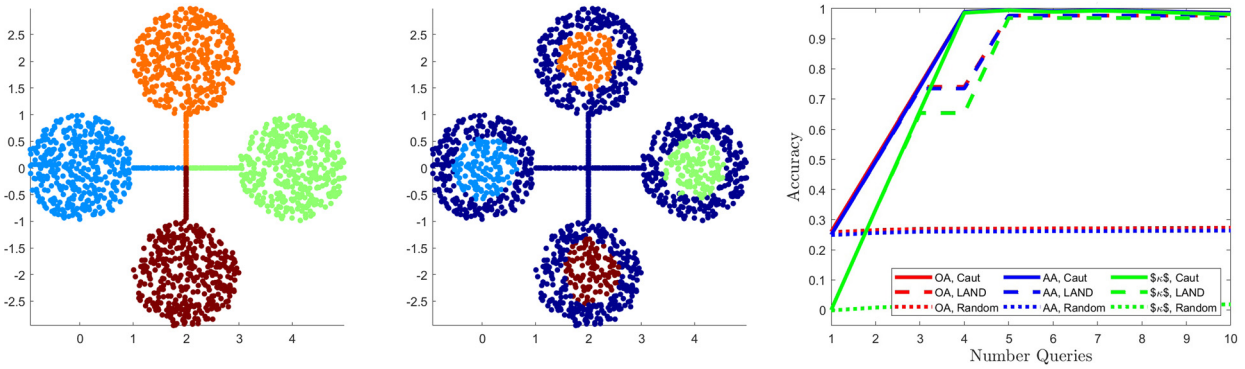



**Fig. 3.** (Left) Clustered data with no minimal separation. Color corresponds to label. (Center) Example of cautious clustering approach with 4 labels queried and $n = 4$. Dark blue labels are uncertain points, i.e., points below the density threshold. (Right) Different measures of error after witness function propagation from confident points (our algorithm cautious active clustering), and comparison to LAND [31]. Note that LAND performs similarly by the time 5 labels have been queried, and plateaus at a similar overall accuracy to our algorithm.

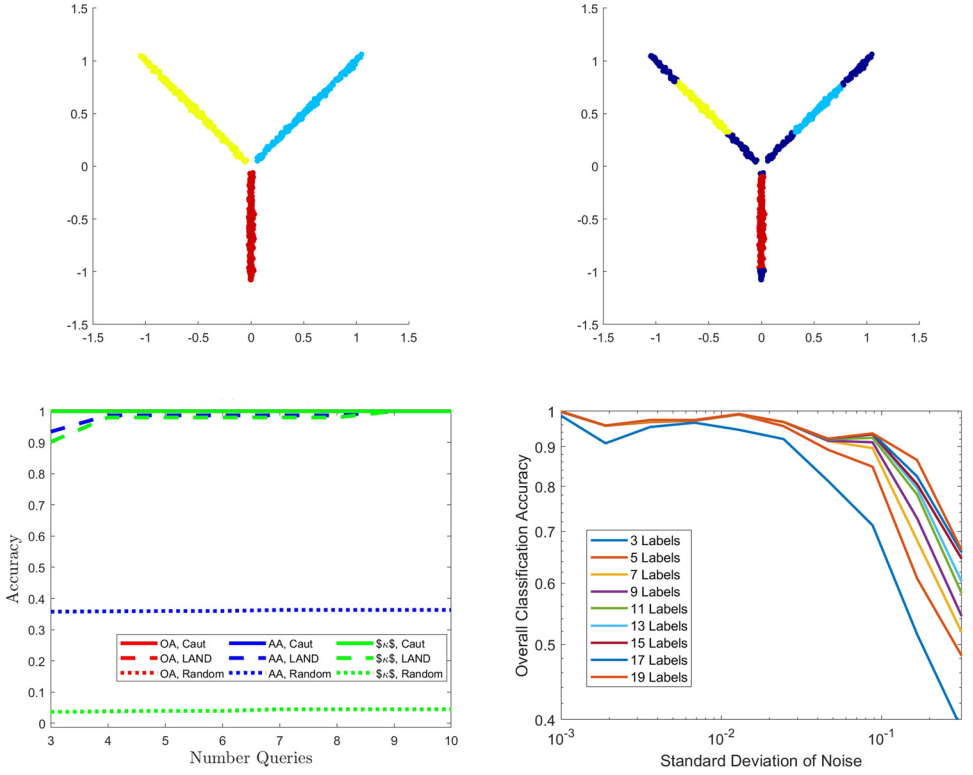



**Fig. 4.** (Top Left) Clustered data with small minimal separation and no density peaks. Color corresponds to label. (Top Right) Example of cautious clustering approach with 3 labels queried and $n = 4$. Dark blue labels are uncertain points, i.e., points below the density threshold. (Bottom Left) Different measures of error for our cautious active clustering algorithm, and comparison to LAND [31]. Note that LAND requires 9 labels to attain perfect classification, as opposed to 3 for our algorithm. (Bottom Right) Accuracy of our algorithm on same data with varying levels of Gaussian noise added to the points. Different curves correspond to the allowed budget of sampled points, and curves are averaged over 25 realizations.

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 Fig. 5, 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

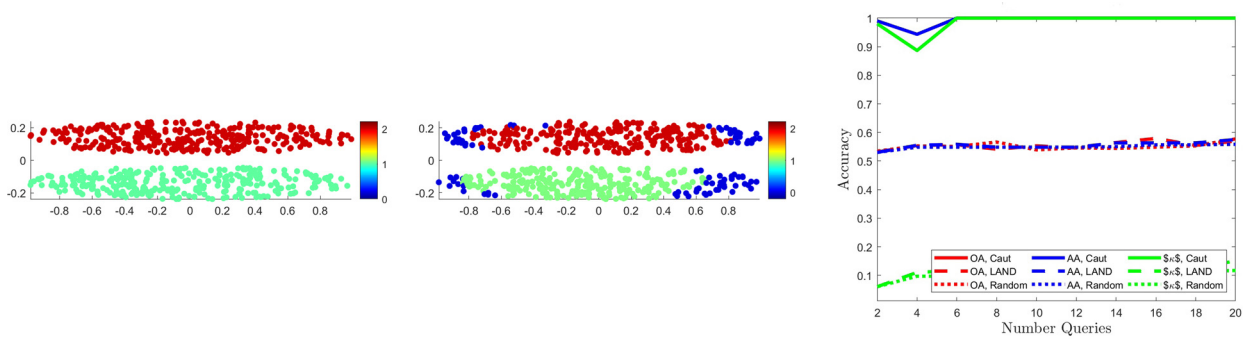



**Fig. 5.** (Left) Clustered data with small minimal separation compared to inner radius. Color corresponds to label. (Center) Example of cautious clustering approach with 2 labels queried and $n = 4$. Dark blue labels are uncertain points, i.e., points below the density threshold. (Right) Different measures of error after witness function propagation from confident points (our algorithm cautious active clustering), and comparison to LAND [31]. Note that LAND performs significantly worse than our algorithm in this case, mostly due to the small minimal separation between clusters that leads to false edges between the two clusters without first performing thresholding.

there exists a path between 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 revolves around estimating regions of space corresponding to given classes, and determining which regions of the latent space correspond to which class 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) [20] and variational autoencoders (VAEs) [24]. 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 [19].

We examine this problem with the well known MNIST data set [26]. This is a set of handwritten digits $0 \cdots 9$, each scanned as a $28 \times 28$ 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 $E(\mathbf{x})$ with architecture $784 - 500 - 500 - 2$ and a decoder/generator $G(\mathbf{z})$ 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 (Fig. 6).

### *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 long wavelength infrared ($\approx 2500$ nm wavelength) to ultraviolet ($\approx 400$ nm wavelength). Each pixel can cover a significant area of the surface, usually several square meters. Because of this, it is not always

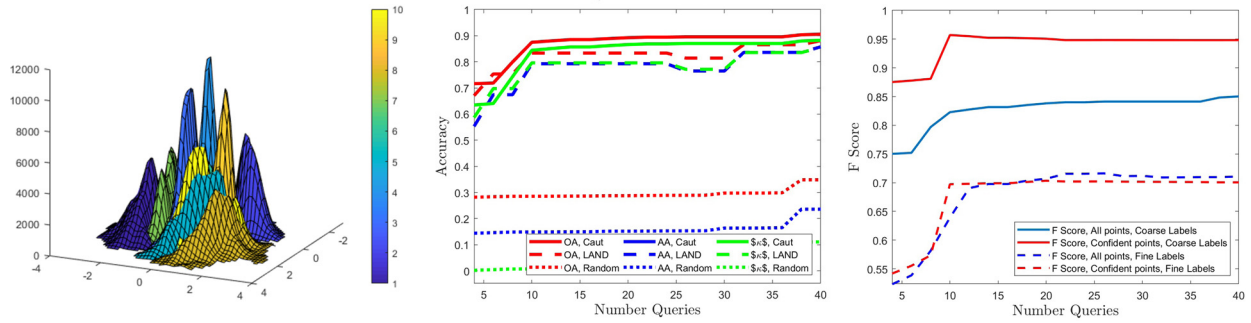



**Fig. 6.** (Left) Density estimate of 2-dimensional VAE latent space, colored by each MNIST label. (Middle) Comparison of cautious clustering approach using coarse MNIST labels (merging overlapping clusters as described in Section 5.2). (Right) $F$-score as a function of number of queries for cautious clustering approach for both coarse MNIST labels and for fine MNIST labels.

advantageous to use spatial similarity to aid in classification and clustering, since neighboring pixels could easily have different labels [2]. Instead, we will consider only the spectral similarity among the pixels [2,13].

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 [44,40].

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 also note that for these experiments, we only plan to use spectral similarity between HSI pixels. In certain HSI algorithms, the spatial relationship between the pixels is also incorporated. This could be done using $\Phi_n$ by considering data $\mathbf{x} = (\mathbf{x}_{spec}, \mathbf{x}_{spat})$ for spectral information $\mathbf{x}_{spec}$ and spatial information $\mathbf{x}_{spat}$, and taking a product kernel $\Phi_n(\mathbf{x}_{spec}, \mathbf{y}_{spec}) \cdot \Phi_m(\mathbf{x}_{spat}, \mathbf{y}_{spat})$. Such spectral-spatial product kernels have been considered on HSI [13,3], and spatial-spectral kernels have been considered in the context of HSI active learning [39,41].

We demonstrate this hierarchical relationship in an example using the Salinas-A data set of hyperspectral imagery.[3]

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 10 using PCA. Thus, there are 7138 points in $\mathbb{R}^{204}$, which are projected to 7138 points in $\mathbb{R}^{10}$. We run our cautious clustering algorithm on this data in two settings, and display summary results in Fig. 7. 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.

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

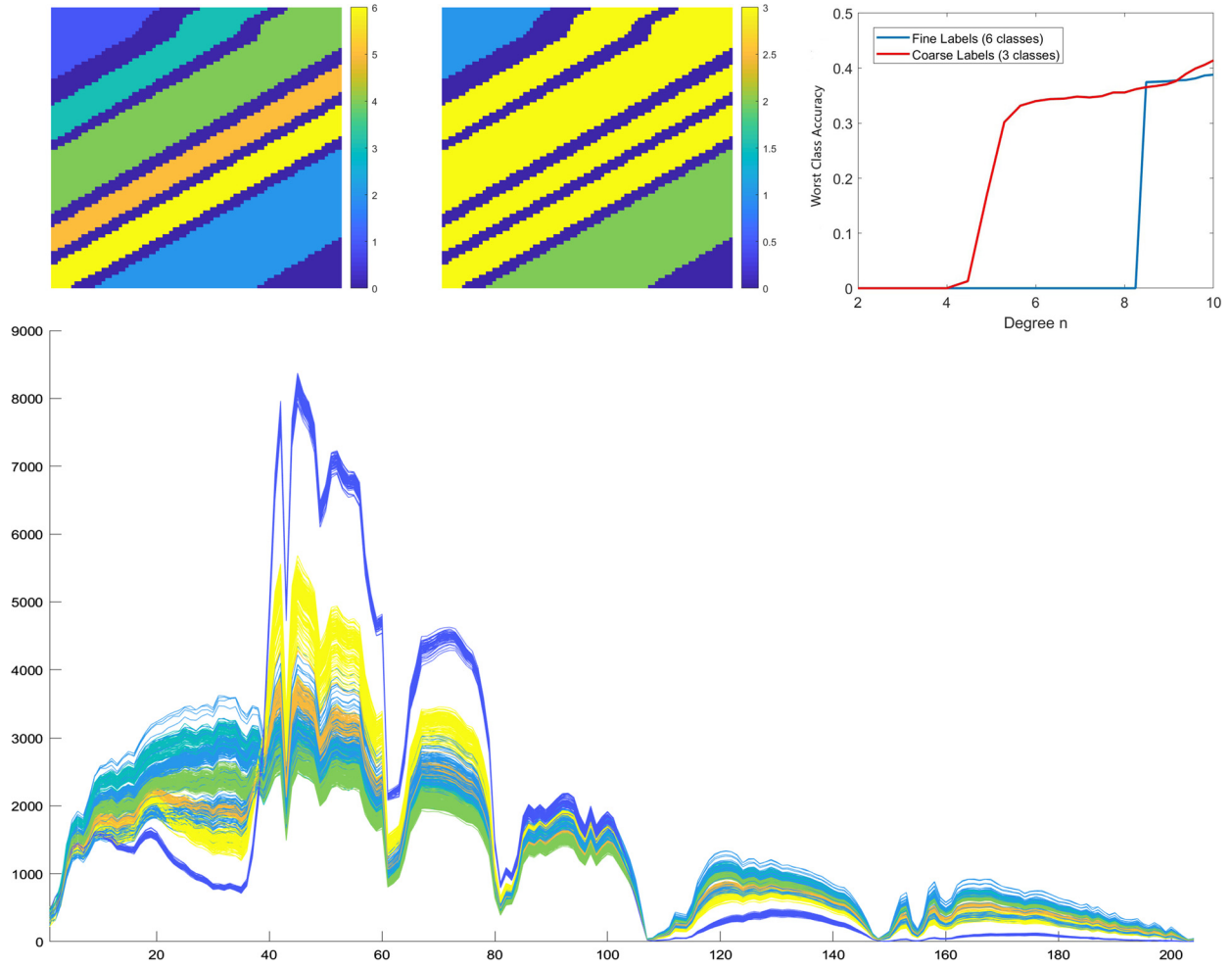



**Fig. 7.** Salinas-A HSI data. (Top Left) Ground Truth of pixels with color corresponding to one of six labels (fine labels). (Top middle) Same scene with all lettuce labels placed in single class (coarse labels). (Top right) Worst classification accuracy on a class using our algorithm as a function 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, the $x$-axis denotes the band number and the $y$-axis denotes the corresponding spectral radiance in that band.

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 $\mathcal{G}_{nmax}(\Theta, \mathcal{C})$ 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 $\mathcal{C} \setminus \mathcal{G}_{nmax}(\Theta, \mathcal{C})$ using the witness function. We display the accuracy on $\mathcal{G}_{nmax}(\Theta, \mathcal{C})$ in Fig. 8, as well as the classification accuracy on the full data set after propagating labels to $\mathcal{C} \setminus \mathcal{G}_{nmax}(\Theta, \mathcal{C})$ with the witness function. We note that there is a slight dip in the accuracy and $F$-score for approximately 6 queries. This is due to the fact that the newest labels added begin to repeat classes (multiple samples from one label) prior to adding a first sample of the final class of points missing. Similarly, when the minimal separation is still large (small $n$), it is possible that one cluster still contains multiple classes and thus has an artificially low classification accuracy that rebounds as $n$ increases. Finally, in Fig. 8 we display the classification accuracy with a varied PCA dimension for pre-processing the data. Reducing the dimension corresponds to removing in additional directions of variance from the HSI pixels, some of which may contain signal in determining a separation between clusters. Despite this, our algorithm still performs well in this reduced dimension.

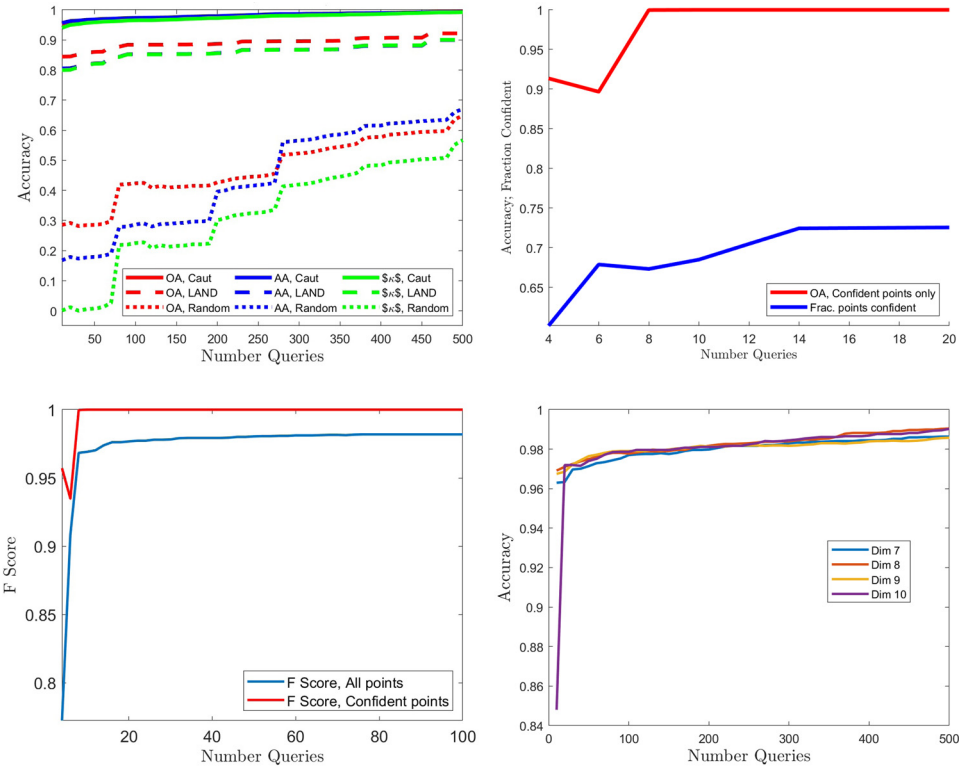



**Fig. 8.** Salinas-A HSI data. (Top Left) Comparison of our algorithm of cautious active clustering to LAND and random sampling of labels. (Top Right) Classification accuracy on $\mathcal{G}_{nmax}(\Theta, \mathcal{C})$ only, and fraction of points in $\mathcal{G}_{nmax}(\Theta, \mathcal{C})$. (Bottom Left) F score for $\mathcal{G}_{nmax}(\Theta, \mathcal{C})$, and for all points $\mathcal{C}$. (Bottom Right) The classification accuracy of our algorithm when varying the dimension of the PCA pre-processing.

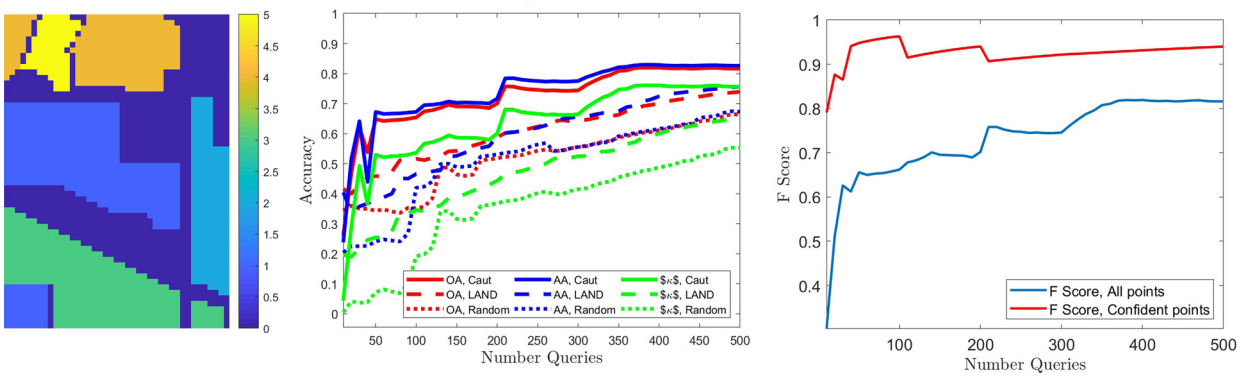



**Fig. 9.** Indian Pines HSI (Left) Ground truth of the small segment of Indian Pines image. (Center) Comparison of our cautious active clustering algorithm to LAND and random sampling of labels. (Right) F score for $\mathcal{G}_{nmax}(\Theta, \mathcal{C})$ and for all points $\mathcal{C}$.

We also examine a second data set, which is a $57 \times 41$ subset of the Indian Pines hyperspectral data set.[4]

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 Fig. 9. 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.

[4] 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 $\Phi_n$ defined in (2.12) (cf. [9]).

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

$$\kappa_1 n^{2q} \le \Phi_n(\mathbf{x},\mathbf{x})^2 \le \kappa_2 n^{2q}, \qquad |\mathbf{x}|_\infty, |\mathbf{y}|_\infty \le \kappa n, \tag{6.1}$$

$$|\Phi_n(\mathbf{x},\mathbf{y})^2 - \Phi_n(\mathbf{x},\mathbf{x})^2| < (1/2)\Phi_n(\mathbf{x},\mathbf{x})^2, \qquad |\mathbf{x}-\mathbf{y}|_\infty \le \kappa_3/n,\ |\mathbf{x}|_\infty \le \kappa n, \tag{6.2}$$

*and*

$$|\Phi_n(\mathbf{x},\mathbf{y})|^2 \le \frac{\kappa_4 n^{2q}}{\max(1,(n|\mathbf{x}-\mathbf{y}|_\infty)^S)}, \qquad \mathbf{x},\mathbf{y}\in\mathbb{R}^q. \tag{6.3}$$

For $n > 0$ (not necessarily an integer), let $\Pi^q_{n,a} = \{\mathbf{x} \mapsto P(\mathbf{x})\exp(-a|\mathbf{x}|_2^2) : P \text{ polynomial of total degree } < n^2\}$. We will omit the mention of $a$ when $a = 1$. Members of $\Pi^q_n$ will be called ($q$-variate) 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 [33,36] one variable at a time to the multi-variate case.

**Proposition 6.2.** *Let $n \ge 1$, $P \in \Pi^q_{n,a}$.*

(a) ***(MRS identity)*** *We have*

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

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

$$\||\nabla P|\| \le \kappa_5 \frac{n}{a}\|P\|. \tag{6.5}$$

The following corollary is easy to prove:

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

$$\max_{\mathbf{x}\in[-n/a,n/a]^q} \min_{\mathbf{y}\in\mathcal{C}} |\mathbf{x}-\mathbf{y}|_\infty \le a/(2\kappa_5 n). \tag{6.6}$$

*Then for any $P \in \Pi^q_{n,a}$,*

$$\max_{\mathbf{y}\in\mathcal{C}} |P(\mathbf{y})| \le \|P\| \le 2\max_{\mathbf{y}\in\mathcal{C}} |P(\mathbf{y})|. \tag{6.7}$$

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

We will need the following facts from probability theory. Theorem 6.1 is proved as [37, Theorem 6.1].

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

$$\sup_{x\in\mathbb{X}} |f(x)| \le \mathfrak{n}(W,\mathcal{C}) \sup_{y\in\mathcal{C}} |f(y)|, \qquad f \in W. \tag{6.8}$$

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

$$\mathsf{Prob}_\mu \left( \sup_{x\in\mathbb{X}} \left| \frac{1}{M} \sum_{j=1}^{M} Z(\omega_j)(x) - \mathbb{E}_\mu(Z(\circ)(x)) \right| \ge 4\mathfrak{n}(W,\mathcal{C}) R \sqrt{\frac{\log(2|\mathcal{C}|/\delta)}{M}} \right) \le \delta. \tag{6.9}$$

The following proposition summarizes the multiplicative Chernoff bounds in the form we need them (cf., e.g., [21, Eqn. (7)] for an elementary proof).

**Proposition 6.3.** *Let $M \ge 1$, $0 \le p \le 1$, and $X_1, \cdots, X_M$ be random variables taking values in $\{0,1\}$, with $\mathsf{Prob}(X_k = 1) = p$. Then for $\epsilon \in (0,1]$,*

$$\mathsf{Prob}\left( \sum_{k=1}^{M} X_k \le (1-\epsilon)Mp \right) \le \exp(-\epsilon^2 Mp/2). \tag{6.10}$$

*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

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

**Lemma 6.1.** *Let $d > 0$, $n \ge 1$, $\mathbf{x} \in \mathbb{R}^q$, and $\mathsf{supp}\ (\mu^*) \subseteq \mathbb{B}(\mathbf{0}, \kappa n)$. Then there exist $C_3, C_4$ such that*

$$\int_{\mathbb{R}^q \setminus \mathbb{B}(\mathbf{x},d)} \Phi_n(\mathbf{x},\mathbf{y})^2 d\mu^*(\mathbf{x}) \le C_3 n^{2q-\alpha} \min\left(1, (nd)^{\alpha-S}\right), \tag{6.12}$$

*and*

$$n^{2q-\alpha}/C_4 \le \inf_{\mathbf{x}\in \mathsf{supp}\ (\mu^*)} \int_{\mathbb{R}^q} \Phi_n(\mathbf{x},\mathbf{y})^2 d\mu^*(\mathbf{y}) \le \max_{\mathbf{x}\in\mathbb{R}^q} \int_{\mathbb{R}^q} \Phi_n(\mathbf{x},\mathbf{y})^2 d\mu^*(\mathbf{y}) \le C_3 n^{2q-\alpha}. \tag{6.13}$$

*In particular,*

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

**Proof.** First, let $d \ge 1/n$, and for $k \in \mathbb{Z}_+$, $A_k = \{\mathbf{y} : 2^k d < |\mathbf{x}-\mathbf{y}|_\infty \le 2^{k+1} d\}$. Then (2.3) shows that $\mu^*(A_k) \le 2^\alpha C_2 (2^k d)^\alpha$. Hence, (6.3) leads to

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

Using this estimate with $d = 1/n$, and using (6.3) and (2.3) again, we obtain that

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

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

Let $\mathbf{x}_0 \in \mathsf{supp}\ (\mu^*)$. Then $\mu^*(\mathbb{B}(\mathbf{x}_0,\kappa_3/n)) \ge C_1\kappa_3^\alpha n^{-\alpha}$. In view of (6.1) and (6.2), we have $\Phi_n(\mathbf{x}_0,\mathbf{y})^2 \ge (\kappa_1/2)n^{2q}$ for all $\mathbf{y} \in \mathbb{B}(\mathbf{x}_0,\kappa_3/n)$. Therefore,

$$\int_{\mathbb{B}(\mathbf{x}_0,\kappa_3/n)} \Phi_n(\mathbf{x}_0,\mathbf{y})^2 d\mu^*(\mathbf{y}) \ge (C_1\kappa_1/2)\kappa_3^\alpha n^{2q-\alpha}.$$

This leads to the first inequality in (6.13). □

**Lemma 6.2.** *Let $n \ge 1$ be large enough so that $\mathsf{supp}\ (\mu^*) \subseteq \mathbb{B}(\mathbf{0},\kappa n)$, $\{\mathbf{x}_j\}_{j=1}^M$ be independent samples with $\mu^*$ as the probability distribution. Then*

$$\mathsf{Prob}\left(\sup_{\mathbf{x}\in\mathbb{R}^q} \left|\frac{1}{M}\sum_{j=1}^M \Phi_n(\mathbf{x},\mathbf{x}_j)^2 - \int_{\mathbb{R}^q} \Phi_n(\mathbf{x},\mathbf{y})^2 d\mu^*(\mathbf{y})\right| \ge cn^\alpha\sqrt{\frac{\log n}{M}} I_n\right) \le \frac{1}{2n}. \tag{6.16}$$

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

$$M \ge (c^2/\beta^2) n^{2\alpha}\log n, \tag{6.17}$$

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

$$\int_{\mathbb{R}^q} \Phi_n(\mathbf{x},\mathbf{y})^2 d\mu^*(\mathbf{y}) - \beta I_n \le \frac{1}{M}\sum_{j=1}^M \Phi_n(\mathbf{x},\mathbf{x}_j)^2 \le \int_{\mathbb{R}^q} \Phi_n(\mathbf{x},\mathbf{y})^2 d\mu^*(\mathbf{y}) + \beta I_n, \tag{6.18}$$

$$(1-\beta)I_n \le \max_{\mathbf{x}\in\mathbb{R}^q} \frac{1}{M}\sum_{j=1}^M \Phi_n(\mathbf{x},\mathbf{x}_j)^2 \le (1+\beta)I_n. \tag{6.19}$$

**Proof.** We use Theorem 6.1 with the following choices: $\mu^*$ in place of $\mu$, $\mathbf{x}_j$ in place of $\omega_j$, $\delta = 1/(2n)$, $Z(\circ)(\mathbf{x}) = \Phi_n(\mathbf{x},\circ)^2$ (so that $W = \Pi_{2n}^q$, and with $\mathcal{C}$ as in Corollary 6.1, $\mathfrak{n}(W,\mathcal{C}) = 2$, $|\mathcal{C}| \sim n^{2q}$). This yields

$$\mathsf{Prob}\left(\sup_{\mathbf{x}\in\mathbb{R}^q} \left|\frac{1}{M}\sum_{j=1}^M \Phi_n(\mathbf{x},\mathbf{x}_j)^2 - \int_{\mathbb{R}^q} \Phi_n(\mathbf{x},\mathbf{y})^2 d\mu^*(\mathbf{y})\right| \ge cn^{2q}\sqrt{\frac{\log n}{M}}\right) \le \frac{1}{2n}.$$

The proof is completed using (6.13). □

**Lemma 6.3.** *There exist $C^*, c_1, c_2 > 0$ with the following property. Let $n \geq c_1$, $0 < \beta < 1$, $\mathsf{supp}\ (\mu^*) \subseteq \mathbb{B}(\mathbf{0}, \kappa n)$, $M \geq c_2 \beta^{-2} n^{2\alpha} \log n$. If $\{\mathbf{x}_j\}_{j=1}^M$ is a random sample with $\mu^*$ as the probability distribution, then*

$$\mathsf{Prob}\left( \max_{1\leq k\leq M} \frac{1}{M} \sum_{j=1}^{M} \Phi_n(\mathbf{x}_k, \mathbf{x}_j)^2 \leq C^* I_n \right) \leq 1/(2n); \tag{6.20}$$

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

$$C^* I_n \leq \max_{1\leq k\leq M} \frac{1}{M} \sum_{j=1}^{M} \Phi_n(\mathbf{x}_k, \mathbf{x}_j)^2 \leq (1+\beta) I_n. \tag{6.21}$$

**Proof.** In this proof, let $P \in \Pi_{2n}^q$ be defined by

$$P(\mathbf{x}) = \int_{\mathbb{R}^q} \Phi_n(\mathbf{x}, \mathbf{y})^2 d\mu^*(\mathbf{y}).$$

Let $n$ be large enough so that $\mathsf{supp}\ (\mu^*) \subset \mathbb{B}(\mathbf{0}, \kappa n)$. Then (6.13) shows that

$$\sup_{\mathbf{x}\in \mathsf{supp}\ (\mu^*)} |P(\mathbf{x})| = P(\mathbf{x}^*) \geq c I_n \tag{6.22}$$

for some $\mathbf{x}^* \in \mathsf{supp}\ (\mu^*)$. Therefore, using the Bernstein inequality (6.5), we obtain for $\mathbf{x} \in \mathbb{R}^q$,

$$|P(\mathbf{x}^*) - P(\mathbf{x})| \leq cn|\mathbf{x}^* - \mathbf{x}|_\infty I_n \leq cn|\mathbf{x}^* - \mathbf{x}|_\infty \sup_{\mathbf{x}\in \mathsf{supp}\ (\mu^*)} |P(\mathbf{x})| = cn|\mathbf{x}^* - \mathbf{x}|_\infty P(\mathbf{x}^*);$$

i.e.,

$$P(\mathbf{x}) \geq (1 - cn|\mathbf{x}^* - \mathbf{x}|_\infty)\, P(\mathbf{x}^*) \geq c_3\, (1 - cn|\mathbf{x}^* - \mathbf{x}|_\infty)\, I_n, \qquad \mathbf{x} \in \mathbb{B}(\mathbf{x}^*, (cn)^{-1}). \tag{6.23}$$

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

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

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

$$\max_{1\leq k\leq M} P(\mathbf{x}_k) \geq P(\mathbf{x}_\ell) \geq 2C^* I_n. \tag{6.24}$$

Using the first estimate in (6.18) with $C^*/2$ in place of $\beta$, we see that if $M \geq cn^{2\alpha} \log n$, then with $C^*$ as in (6.24), and probability exceeding $1 - 1/(2n)$,

$$\max_{1\leq k\leq M} \frac{1}{M} \sum_{j=1}^{M} \Phi_n(\mathbf{x}_k, \mathbf{x}_j)^2 \geq \max_{1\leq k\leq M} P(\mathbf{x}_k) - (C^*/2) I_n \geq C^* I_n.$$

This proves the first inequality in (6.21). The second inequality is immediate from the second inequality in (6.19). □

*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 for $n \geq 1$,*

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

*We assume (3.1) and*

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

*Then with*

$$d(\theta) = \left(\frac{C_3C_4}{\theta}\right)^{1/(S-\alpha)} \tag{6.27}$$

$$\mathsf{supp}\ (\mu^*) \subseteq \mathcal{S} \subseteq \left\{ \mathbf{x} \in \mathbb{R}^q : \mathsf{dist}(\mathbf{x}, \mathsf{supp}\ (\mu^*)) \leq \frac{d(\theta)}{n} \right\}. \tag{6.28}$$

**Proof.** The estimate (6.13) shows that for $\mathbf{x} \in \mathsf{supp}\ (\mu^*)$,

$$\int_{\mathbb{R}^q} \Phi_n(\mathbf{x},\mathbf{y})^2 d\mu^*(y) \geq \frac{I_n}{C_3C_4}.$$

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

Let $\mathsf{dist}(\mathbf{x}, \mathsf{supp}\ (\mu^*)) \geq d(\theta)/n$. The condition (6.26) shows that $d(\theta) \geq 1$. So, (6.12) leads to

$$\begin{aligned}\int_{\mathbb{R}^q} \Phi_n(\mathbf{x},\mathbf{y})^2 d\mu^*(y) &= \int_{\mathsf{supp}\ (\mu^*)} \Phi_n(\mathbf{x},\mathbf{y})^2 d\mu^*(y) \\ &\leq \int_{\mathbb{R}^q\setminus \mathbb{B}(\mathbf{x},d(\theta)/n)} \Phi_n(\mathbf{x},\mathbf{y})^2 d\mu^*(\mathbf{x}) \leq C_3 n^{q-2\alpha} d(\theta)^{\alpha-S} \\ &\leq C_3C_4 d(\theta)^{\alpha-S} I_n = \theta I_n. \end{aligned} \tag{6.29}$$

This proves the second inclusion in (6.28). □

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

**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*

$$n \geq 2d(\theta)/\eta,\ \ \kappa_1 C_4 n^\alpha \mu^*(\mathbf{S}_{K_\eta+1,\eta}) \leq \theta. \tag{6.30}$$

*Let*

$$\mathcal{S}_{k,\eta,n}(\theta) = \mathcal{S}_n(\theta) \cap \{\mathbf{x} \in \mathbb{R}^q : \mathsf{dist}(\mathbf{x}, \mathbf{S}_{k,\eta}) \le d(\theta)/n\}. \tag{6.31}$$

*Then the set $\mathcal{S}_n(\theta)$ is a disjoint union of sets $\mathcal{S}_{k,\eta,n}(\theta)$, $k = 1, \cdots, K_\eta$ such that*

$$\mathsf{dist}(\mathcal{S}_{k,\eta,n}(\theta), \mathcal{S}_{j,\eta,n}(\theta)) \ge \eta, \qquad k \ne j,\ k, j = 1, \cdots, K_\eta, \tag{6.32}$$

*and for $k = 1, \cdots, K_\eta$,*

$$\mathsf{supp}\ (\mu^*) \cap \{\mathbf{x} : \mathsf{dist}(\mathbf{x}, \mathbf{S}_{k,\eta}) \le d(\theta)/n\} \subseteq \mathcal{S}_{k,\eta,n}(\theta) \subseteq \{\mathbf{x} \in \mathbb{R}^q : \mathsf{dist}(\mathbf{x}, \mathbf{S}_{k,\eta}) \le d(\theta)/n\}. \tag{6.33}$$

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

Let $\mathbf{x} \in \mathcal{S}_n(\theta)$. Then we deduce using (6.1) (6.13), and (6.30) that

$$\int_{\mathbf{S}_{K_\eta+1}} \Phi_n(\mathbf{x}, \mathbf{y})^2 d\mu^*(y) \le \kappa_1 n^{2q} \mu^*(\mathbf{S}_{K_\eta+1}) \le \kappa_1 C_4 n^\alpha \mu^*(\mathbf{S}_{K_\eta+1}) I_n \le \theta I_n. \tag{6.34}$$

In this proof, we will denote

$$\mathbf{S} = \bigcup_{k=1}^{K_\eta} \mathbf{S}_{k,\eta}.$$

If $\mathsf{dist}(\mathbf{x}, \mathbf{S}) \ge d(\theta)/n$ then we obtain as in the proof of Theorem 6.2, that

$$\int_{\mathbf{S}} \Phi_n(\mathbf{x}, \mathbf{y})^2 d\mu^*(y) \le \int_{\mathbb{R}^q \setminus \mathbb{B}(\mathbf{x}, d(\theta)/n)} \Phi_n(\mathbf{x}, \mathbf{y})^2 d\mu^*(\mathbf{x}) \le C_3 n^{q-2\alpha} d(\theta)^{\alpha - S} \le C_3 C_4 d(\theta)^{\alpha-S} I_n = \theta I_n.$$

Together with (6.34), this implies that $\mathcal{S}_n(\theta) \subseteq \{\mathbf{x} : \mathsf{dist}(\mathbf{x}, \mathbf{S}) \le d(\theta)/n\}$. Since $d(\theta)/n \le \eta/2$, the minimal separation condition shows that for any $\mathbf{x}$ with $\mathsf{dist}(\mathbf{x}, \mathbf{S}) \le d(\theta)/n$, there exists a unique $k$, $k = 1, \cdots, K$, such that $\mathsf{dist}(\mathbf{x}, \mathbf{S}_{k,\eta}) \le d(\theta)/n$. Thus, $\mathcal{S}_n(\theta) = \bigcup_{k=1}^K \mathcal{S}_{k,\eta,n}(\theta)$, and the second inclusion of (6.33) is proved. □

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

**Lemma 6.4.** *Let $0 < \Theta \le 1$, $M, n \ge 2$ be integers, $M \ge 2$ and $\mathcal{C} = \{\mathbf{x}_1, \cdots, \mathbf{x}_M\}$ be independently sampled from the probability distribution $\mu^*$. Let $\mathcal{G}_n(\Theta, \mathcal{C})$ be defined by (3.2). There exist constants $c, c_1, c_2$ such that if $M \ge cn^{2\alpha}\sqrt{\log n}$ then with probability $\ge 1 - c_1 M^{-c_2}$,*

$$\mathcal{S}_n\left(\frac{(1+C^*)\Theta}{4}\right) \subseteq \mathcal{G}_n(\Theta, \mathcal{C}) \subseteq \mathcal{S}_n(C^*\Theta/8). \tag{6.35}$$

**Proof.** All statements below hold with probability $\ge 1 - c_1 M^{-c_2}$, 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 $\mathbf{x} \in \mathcal{S}_n(t_1)$. Using (6.18) and (6.21), we deduce that

$$\frac{1}{M}\sum_{j=1}^{M}\Phi_n(\mathbf{x},\mathbf{x}_j)^2 \geq \int_{\mathbb{R}^q}\Phi_n(\mathbf{x},\mathbf{y})^2 d\mu^*(\mathbf{y}) - \beta I_n \geq 4t_1 I_n - \beta I_n = (4t_1-\beta)I_n$$
$$\geq \frac{4t_1-\beta}{1+\beta}\max_{1\leq k\leq M}\frac{1}{M}\sum_{j=1}^{M}\Phi_n(\mathbf{x}_k,\mathbf{x}_j)^2 = \Theta\max_{1\leq k\leq M}\frac{1}{M}\sum_{j=1}^{M}\Phi_n(\mathbf{x}_k,\mathbf{x}_j)^2.$$

Thus,

$$\mathcal{S}_n\left(\frac{2\Theta+C^*\Theta(1+\Theta)}{8}\right)\subseteq \mathcal{G}_n(\Theta,\mathcal{C}).$$

Since

$$(1+C^*)\Theta/2 \geq \frac{2\Theta+C^*\Theta(1+\Theta)}{8},$$

this proves the first inclusion in (6.35).

Next, let $\mathbf{x}\in\mathcal{G}_n(\Theta,\mathcal{C})$. Using (6.18) and (6.21), we deduce that

$$\int_{\mathbb{R}^q}\Phi_n(\mathbf{x},\mathbf{y})^2 d\mu^*(\mathbf{y}) \geq \frac{1}{M}\sum_{j=1}^{M}\Phi_n(\mathbf{x},\mathbf{x}_j)^2 - \beta I_n \geq \Theta\max_{1\leq k\leq M}\frac{1}{M}\sum_{j=1}^{M}\Phi_n(\mathbf{x}_k,\mathbf{x}_j)^2$$
$$\geq (C^*\Theta-\beta)I_n = 4(C^*\Theta/8)I_n.$$

This proves the second inclusion in (6.35). □

**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. □

**Proof of Theorem 3.3.** In this proof, we write $\mathcal{G}_{k,\eta,n}$ in place of $\mathcal{G}_{k,\eta,n}(\Theta,\mathcal{C})$. In view of (3.7), we have for $k=1,\cdots,K_\eta$, $\mathsf{supp}\ (\mu^*)\cap\mathbf{S}_{k,\eta}\subset\mathcal{G}_{k,\eta,n}$. Hence,

$$\mu^*(\mathcal{G}_{k,\eta,n})\geq\mu^*(\mathbf{S}_{k,\eta}),\qquad k=1,\cdots,K_\eta. \tag{6.36}$$

With $c_1$, $c_2$ as in Theorem 3.2, let

$$d_n = c_2/(n\Theta^{1/(S-\alpha)}).$$

Then The estimate (3.7) implies again that

$$\mu^*\left(\{\mathbf{x}\in\mathbb{R}^q : \mathsf{dist}(\mathbf{x},\mathbf{S}_{k,\eta})\leq d_n\}\right)$$
$$= \mu^*\left(\mathsf{supp}(\mu^*)\cap\{\mathbf{x}\in\mathbb{R}^q : \mathsf{dist}(\mathbf{x},\mathbf{S}_{k,\eta})\leq d_n\}\right) = \mu^*(\mathbf{S}_{k,\eta}\cup\mathbf{S}_{K_\eta+1,\eta})\leq\mu^*(\mathbf{S}_{k,\eta})+\mu^*(\mathbf{S}_{K_\eta+1,\eta}).$$

Hence, the second inclusion in (3.7) implies that for $k=1,\cdots,K_\eta$,

$$\mu^*(\mathcal{G}_{k,\eta,n})\leq\mu^*(\mathbf{S}_{k,\eta})+\mu^*(\mathbf{S}_{K_\eta+1,\eta}). \tag{6.37}$$

In view of (6.36) and (6.37), for each $k=1,\cdots,K$,

$$F_\eta(\mathcal{G}_{k,\eta,n})\geq 2\frac{\mu^*(\mathbf{S}_{k,\eta})}{2\mu^*(\mathbf{S}_{k,\eta})+\mu^*(\mathbf{S}_{K_\eta+1,\eta})}\geq 1-\frac{\mu^*(\mathbf{S}_{K_\eta+1,\eta})}{2\mu^*(\mathbf{S}_{k,\eta})+\mu^*(\mathbf{S}_{K_\eta+1,\eta})}. \tag{6.38}$$

Since $\{\mathcal{G}_{k,\eta,n}\}$ is a partition of $\mathcal{G}_n(\Theta,\mathcal{C}) \supseteq \mathsf{supp}\ (\mu^*)$,

$$\sum_{k=1}^{K_\eta} \mu^*(\mathcal{G}_{k,\eta,n}) = 1.$$

Therefore, (6.38) and (6.37) imply that

$$\sum_{k=1}^{K_\eta} \mu^*(\mathcal{G}_{k,\eta,n}) F\eta(\mathcal{G}_{k,\eta,n}) \geq 1 - \mu^*(\mathbf{S}_{K_\eta+1,\eta}) \sum_{k=1}^{K_\eta} \frac{\mu^*(\mathcal{G}_{k,\eta,n})}{2\mu^*(\mathbf{S}_{k,\eta}) + \mu^*(\mathbf{S}_{K_\eta+1,\eta})}. \tag{6.39}$$

Therefore,

$$1 \geq \mathcal{F}_\eta\left(\{\mathcal{G}_{k,\eta,n}\}_{k=1}^K\right) \geq 1 - \frac{\mu^*(\mathbf{S}_{K_\eta+1,\eta})}{2\min_{1\leq k\leq K_\eta} \mu^*(\mathbf{S}_{k,\eta})}.$$

In view of (3.8), this completes the proof. □

## 7. Conclusions

We have introduced the concept that the machine learning problem of classification can be considered in a manner analogous to the problem in signal processing of separating point sources. We have pointed out the various similarities and differences which makes the problem much harder than that of separation of point sources, in particular, because of overlapping class boundaries. We have introduced a localized kernel based on Hermite polynomials, and demonstrated its use in separating supports of the components of the data distribution corresponding to different classes. We have constructed a multiscale in which the number of classes can be defined hierarchically, and shown that the $F$-score for our classification scheme converges to the optimal value of 1. Having separated the supports of the components, one sample per component leads in theory to a complete classification based on a minimal number of label queries. We have given an algorithm to determine which points in the data set should be queried for labels in an optimal and reliable manner. The algorithm is demonstrated in several synthetic examples as well as the MNIST data set and some hyperspectral image data sets.

## Appendix A. Constructing localized kernel $\Phi_n$

With

$$\mathsf{Proj}_m(\mathbf{x},\mathbf{y}) = \sum_{|\mathbf{k}|_1=m} \psi_{\mathbf{k}}(\mathbf{x})\psi_{\mathbf{k}}(\mathbf{y}), \tag{A.1}$$

we observe that

$$\Phi_n(H;\mathbf{x},\mathbf{y}) = \sum_{\mathbf{k}\in\mathbb{Z}_+^q} H\left(\frac{\sqrt{|\mathbf{k}|_1}}{n}\right)\psi_{\mathbf{k}}(\mathbf{x})\psi_{\mathbf{k}}(\mathbf{y}) = \sum_{m=0}^{\infty} H\left(\frac{\sqrt{m}}{n}\right)\mathsf{Proj}_m(\mathbf{x},\mathbf{y}). \tag{A.2}$$

In [37], we have observed using the so-called Mehler identity that

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

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

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

and

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

Therefore, the procedure to compute the kernel $\Phi_n(H;\mathbf{x},\mathbf{y})$ is simple. We use the recurrence relations (2.9) to compute the univariate Hermite functions $\psi_j$, use these together with (A.3) to compute $\mathsf{Proj}_m(\mathbf{x},\mathbf{y})$ for $|\mathfrak{m}|_1 \leq n^2$, and finally compute $\Phi_n(H;\mathbf{x},\mathbf{y})$ using (A.2).

## References


[1] M. Anthony, P.L. Bartlett, Neural Network Learning: Theoretical Foundations, 1st edition, Cambridge University Press, New York, NY, USA, 2009.
[2] C.M. Bachmann, T.L. Ainsworth, R.A. Fusina, Exploiting manifold geometry in hyperspectral imagery, IEEE Trans. Geosci. Remote Sens. 43 (3) (2005) 441–454.
[3] J.J. Benedetto, W. Czaja, J. Dobrosotskaya, T. Doster, K. Duke, Spatial-spectral operator theoretic methods for hyperspectral image classification, GEM Int. J. Geomath. 7 (2) (2016) 275–297.
[4] A. Beygelzimer, S. Dasgupta, J. Langford, Importance weighted active learning, in: Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 49–56.
[5] O. Chapelle, B. Scholkopf, A. Zien, Semi-supervised learning (Chapelle, O. et al., eds. 2006) [book reviews], IEEE Trans. Neural Netw. 20 (3) (2009) 542.
[6] X. Cheng, A. Cloninger, R.R. Coifman, Two-sample statistics based on anisotropic kernels, arXiv preprint, arXiv:1709.05006, 2017.
[7] C.K. Chui, F. Filbir, H.N. Mhaskar, Representation of functions on big data: graphs and trees, Appl. Comput. Harmon. Anal. 38 (3) (2015) 489–509.
[8] C.K. Chui, H.N. Mhaskar, Signal decomposition and analysis via extraction of frequencies, Appl. Comput. Harmon. Anal. 40 (1) (2016) 97–136.
[9] C.K. Chui, H.N. Mhaskar, A Fourier-invariant method for locating point-masses and computing their attributes, Appl. Comput. Harmon. Anal. 45 (2018) 436–452.
[10] C.K. Chui, H.N. Mhaskar, A unified method for super-resolution recovery and real exponential-sum separation, Appl. Comput. Harmon. Anal. (2018), published online February 21, 2018.
[11] C.K. Chui, H.N. Mhaskar, M.D. van der Walt, Data-driven atomic decomposition via frequency extraction of intrinsic mode functions, GEM Int. J. Geomath. 7 (1) (2016) 117–146.
[12] C.K. Chui, H.N. Mhaskar, X. Zhuang, Representation of functions on big data associated with directed graphs, Appl. Comput. Harmon. Anal. 44 (2018) 165–188.
[13] A. Cloninger, W. Czaja, T. Doster, Operator analysis and diffusion based embeddings for heterogeneous data fusion, in: 2014 IEEE Geoscience and Remote Sensing Symposium, IEEE, 2014, pp. 1249–1252.
[14] S. Dasgupta, Coarse sample complexity bounds for active learning, in: Advances in Neural Information Processing Systems, 2006, pp. 235–242.
[15] S. Dasgupta, D. Hsu, Hierarchical sampling for active learning, in: Proceedings of the 25th International Conference on Machine Learning, 2008, pp. 208–215.
[16] S. Dasgupta, D. Hsu, S. Poulis, X. Zhu, Teaching a black-box learner, in: International Conference on Machine Learning, 2019, pp. 1547–1555.
[17] S. Dasgupta, P.M. Long, Performance guarantees for hierarchical clustering, J. Comput. Syst. Sci. 70 (4) (2005) 555–569.
[18] O. Dekel, J. Keshet, Y. Singer, Large margin hierarchical classification, in: Proceedings of the Twenty-First International Conference on Machine Learning, 2004, p. 27.
[19] C. Doersch, Tutorial on variational autoencoders, arXiv preprint, arXiv:1606.05908, 2016.
[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
[21] T. Hagerup, C. Rüb, A guided tour of Chernoff bounds, Inf. Process. Lett. 33 (6) (1990) 305–308.
[22] S. Hanneke, A bound on the label complexity of agnostic active learning, in: Proceedings of the 24th International Conference on Machine Learning, 2007, pp. 353–360.
[23] S. Hanneke, L. Yang, Minimax analysis of active learning, J. Mach. Learn. Res. 16 (1) (2015) 3487–3602.

[24] D.P. Kingma, M. Welling, Auto-encoding variational Bayes, arXiv preprint, arXiv:1312.6114, 2013.
[25] S. Lafon, A.B. Lee, Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization, IEEE Trans. Pattern Anal. Mach. Intell. 28 (9) (2006) 1393–1403.
[26] Y. LeCun, C. Cortes, C. Burges, Mnist handwritten digit database, in: AT&T Labs [Online], 2, 2010, Available: http://yann.lecun.com/exdb/mnist.
[27] W. Li, W. Liao, A. Fannjiang, Super-resolution limit of the esprit algorithm, IEEE Trans. Inf. Theory (2020).
[28] S. Ling, T. Strohmer, Certifying global optimality of graph cuts via semidefinite relaxation: a performance guarantee for spectral clustering, Found. Comput. Math. (2019) 1–55.
[29] W. Liu, B. Dai, X. Li, Z. Liu, J.M. Rehg, L. Song, Towards black-box iterative machine teaching, arXiv preprint, arXiv:1710.07742, 2017.
[30] M. Maggioni, H.N. Mhaskar, Diffusion polynomial frames on metric measure spaces, Appl. Comput. Harmon. Anal. 24 (3) (2008) 329–353.
[31] M. Maggioni, J. Murphy, Learning by active nonlinear diffusion, Found. Data Sci. 1 (01 2019) 1–21.
[32] H. Mhaskar, J. Prestin, On the detection of singularities of a periodic function, Adv. Comput. Math. 12 (2–3) (2000) 95–131.
[33] H.N. Mhaskar, Introduction to the Theory of Weighted Polynomial Approximation, vol. 56, World Scientific, Singapore, 1996.
[34] H.N. Mhaskar, On the representation of smooth functions on the sphere using finitely many bits, Appl. Comput. Harmon. Anal. 18 (3) (2005) 215–233.
[35] H.N. Mhaskar, Polynomial operators and local smoothness classes on the unit interval, II, Jaen J. Approx. 1 (1) (2009) 1–25.
[36] H.N. Mhaskar, Local approximation using Hermite functions, in: Progress in Approximation Theory and Applicable Complex Analysis, Springer, 2017, pp. 341–362.
[37] H.N. Mhaskar, A. Cloninger, X. Cheng, A witness function based construction of discriminative models using Hermite polynomials, Arxiv preprint, arXiv:1901.02975, 2019.
[38] H.N. Mhaskar, J. Prestin, On local smoothness classes of periodic functions, J. Fourier Anal. Appl. 11 (3) (2005) 353–373.
[39] J.M. Murphy, Spatially regularized active diffusion learning for high-dimensional images, Pattern Recognit. Lett. (2020).
[40] J.M. Murphy, M. Maggioni, Unsupervised clustering and active learning of hyperspectral images with nonlinear diffusion, IEEE Trans. Geosci. Remote Sens. 57 (3) (2018) 1829–1845.
[41] J.M. Murphy, M. Maggioni, Spectral-spatial diffusion geometry for hyperspectral image clustering, IEEE Geosci. Remote Sens. Lett. (2019).
[42] V. Satuluri, S. Parthasarathy, Symmetrizations for clustering directed graphs, in: Proceedings of the 14th International Conference on Extending Database Technology, ACM, 2011, pp. 343–354.
[43] B. Settles, Active learning literature survey, Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.
[44] D. Tuia, F. Ratle, F. Pacifici, M.F. Kanevski, W.J. Emery, Active learning methods for remote sensing image classification, IEEE Trans. Geosci. Remote Sens. 47 (7) (2009) 2218–2232.
[45] C. Xiong, D.M. Johnson, J.J. Corso, Active clustering with model-based uncertainty reduction, IEEE Trans. Pattern Anal. Mach. Intell. 39 (1) (2016) 5–17.
[46] K. Yu, J. Bi, V. Tresp, Active learning via transductive experimental design, in: Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 1081–1088.
[47] X. Zhu, A. Singla, S. Zilles, A.N. Rafferty, An overview of machine teaching, arXiv preprint, arXiv:1801.05927, 2018.