Optimal Rates for Density and Mode Estimation with Expand-and-sparsify Representations

Kaushik Sinha*
Wichita State University
Wichita, Kansas, USA
kaushik.sinha@wichita.edu

Christopher Tosh*
Memorial Sloan Kettering Cancer Center
New York City, New York, USA
christopher.j.tosh@gmail.com

Abstract

Expand-and-sparsify representations are a class of theoretical models that capture sparse representation phenomena observed in the sensory systems of many animals. At a high level, these representations map an input $x \in \mathbb{R}^d$ to a much higher dimension $m \gg d$ via random linear projections before zeroing out all but the $k \ll m$ largest entries. The result is a k-sparse vector in $\{0,1\}^m$. We study the suitability of this representation for two fundamental statistical problems: density estimation and mode estimation. For density estimation, we show that a simple linear function of the expand-and-sparsify representation produces an estimator with minimax-optimal ℓ_∞ convergence rates. In mode estimation, we provide simple algorithms on top of our density estimator that recover single or multiple modes at optimal rates up to logarithmic factors under mild conditions.

1 INTRODUCTION

Neural circuits that transform dense, lower-dimensional signals to sparse, higher-dimensional signals are a recurring motif in animal sensory systems. In the olfactory system of Drosophila, for example, a dense signal consisting of 50 glomeruli in the antennal lobe is lifted, via random projections, to an approximately 2000-dimensional signal where only the top 5% of coordinates are non-zero (Turner et al., 2008; Caron et al., 2013). Similar phenomena are observed in the olfactory system of rodents, the visual cortex of cats, and the electrosensory system of electric fish (Stettler and Axel, 2009; Olshausen and Field, 2004; Chacron et al., 2011). The widespread nature of these transformations suggests they are useful in representing certain types of data.

Expand-and-sparsify representations are a class of mathematical formalisms to capture and analyze these ubiquitous neural phenomena (Dasgupta and Tosh, 2020). At a high level, an expand-and-sparsify representation consists of a random projection into a higher-dimensional space followed by a sparsification operation. Recent theoretical work has shown that these expand-and-sparsify representations are well-adapted to a variety of statistical applications, including locality-sensitive hashing (Dasgupta et al., 2017), count sketching (Dasgupta et al., 2022), function approximation (Dasgupta and Tosh, 2020), and non-parametric classification (Sinha, 2024).

In this work, we identify another fundamental statistical task that expand-and-sparsify representations are well-suited for: non-parametric density estimation. Specifically, we show that given n *i.i.d.* samples from a smooth density f whose support is restricted to $\mathcal{S}^{d-1}$, there is an estimator $\hat{f}_n$ that is a linear function of the expand-and-sparsify representation that achieves minimax-optimal convergence rates in ℓ_∞ distance. Our bounds hold with high probability, simultaneously for all x in the support of f. Thus, this paper bridges two important areas of study: sparse neural representations of data and density estimation. The expand-and-sparsify representation is common in nature, but its properties are poorly understood. By showing what types of tasks it can solve, we can gain insight on what kinds of circuits might be present in nature. On the other hand, density estimation is a fundamental statistical task that has been studied for decades. Previous theoretical work in this area has been dominated by

*Both authors contributed equally to this work.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

k-NN and kernel-based methods. Our work shows a completely different approach based on sparse random representations achieves optimal convergence rates.

Beyond density estimation, we present simple algorithms, utilizing the expand-and-sparsify representation density estimator $\hat{f}_n$, for another fundamental statistical task, namely, non-parametric mode estimation. For a unimodal density, this algorithm simply returns the maximizer of $\hat{f}_n$ over the finite sample that was used to construct $\hat{f}_n$. We show that with high probability, the ℓ_2 distance between the mode and the estimated mode decays at an optimal rate (up to logarithmic factors). For a density having multiple unknown modes, we present a simple iterative algorithm that outputs a set of candidate modes by taking the maximizer of $\hat{f}_n$ over appropriately chosen level sets. We show that for any mode of f that is well separated from the rest of the modes by low density regions we can always find, with high probability, a mode from the set of candidate modes returned by our algorithm whose ℓ_2 distance to this well-separated mode decays at an optimal rate (up to logarithmic factors).

We summarize our main contributions below:

- We demonstrate the suitability of expand-and-sparsify representation for two fundamental statistical problems: density estimation and mode estimation.
- For density estimation, we show that our proposed density estimator converges at a minimax-optimal rate $O\left(\left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d-1}}\right)$ in ℓ_∞ distance, where β is the Hölder smoothness constant of the distribution to be approximated.
- For mode estimation, we show that under mild conditions, our proposed simple algorithms on top of our estimated density can recover modes at an optimal rate $\tilde{O}\left(n^{-\frac{1}{d+3}}\right)$ (up to logarithmic factors).

The rest of the paper is organized as follows. We discuss related work in Section 2. We summarize various notations used in this paper in Section 3. We present our proposed density estimator using expand-and-sparsify presentation in Section 4 and derive density estimation rates in Section 5. We present our mode estimation results in Section 6, empirical evaluations in Section 7, and conclude in Section 8.

2 RELATED WORK

Density estimation. Non-parametric density estimation is a well-studied problem with known optimal rates. The estimator considered in the present paper bears a strong resemblance to the well-known k-nearest neighbor (kNN) estimator and the kernel density estimator (KDE), both of which are known to achieve minimax optimal rates under general conditions (Dasgupta and Kpotufe, 2014; Rinaldo and Wasserman, 2010; Jiang, 2017). Specifically, in d-dimensions, both the kNN estimator and KDE achieve convergence rates of $\tilde{O}\left(n^{-\frac{\beta}{2\beta+d}}\right)$ in ℓ_∞ distance, where β is the Hölder smoothness constant of the distribution to be approximated.

One interesting way in which the expand-and-sparsify estimator differs from standard density estimators such as kNN and KDE is its use of randomization. Note that once a dataset and a hyperparameter have been fixed, the KDE and kNN estimators are deterministic. However, the expand-and-sparsify density estimator has an additional source of randomness in its choice of high-dimensional projection. This extra randomness is a property shared with a recently proposed random forest density estimator that builds a set of random k-d trees that partition the space and count the number of points that fall into the leaf nodes (Wen and Hang, 2022). That work demonstrated that the estimator has expected convergence rate $\tilde{O}\left(n^{-\frac{1-4^{-\beta}}{1-4^{-\beta}+d\log 2}}\right)$ in ℓ_2 distance.

Expand-and-sparsify representations. As discussed in the introduction, there have been many recent theoretical works exploring the utility of expand-and-sparsify representations for statistical tasks, ranging from classification (Sinha and Ram, 2021; Ram and Sinha, 2022; Sinha, 2024) to locality-sensitive hashing (Dasgupta et al., 2017).

Of particular relevance to the present paper is the work of Dasgupta and Tosh (2020), which showed that for any smooth function, there is a linear function of the corresponding expand-and-sparsify representation that achieves low ℓ_∞ error. While this result demonstrates the existence of a good expand-and-sparsify density estimator in principle, it does not provide an algorithm to construct one from a finite sample. Indeed, as an unsupervised learning problem, it is unclear a priori how to estimate the corresponding coefficients from data. This is a core contribution of the present paper.

Mode estimation. There is an extensive literature on mode estimation, and we review some of the relevant works here. A classical approach to mode estimation consists of estimating the single mode x_0 of f as $\hat{x}_0 = \operatorname{argsup}_{x\in\mathbb{R}^d} f_n(x)$ where f_n is an estimate of f (the kernel density estimator is a typi-

cal choice for f_n). Note that this mode estimator is generally hard to implement in practice. Nevertheless, a series of works on this approach (Parzen, 1962; Chernoff, 1964; Eddy, 1980) establish consistency as well as as convergence rates for this procedure under various regularity assumptions on the distribution. More recent works (Grund and Hall, 1995; Klemelä, 2005; Tsybakov, 1990; Donoho and Liu, 1991) address the problem of optimal choice of kernel bandwidth to adaptively achieve minimax risk for mode estimation estimation. In particular, for a κ-times differentiable f, Tsybakov (1990); Donoho and Liu (1991) independently established minimax risk of the form $n^{-\frac{\kappa-1}{2\kappa+d}}$. Another line of work estimates the mode from practical statistics of the data (Grenander, 1965; Abraham et al., 2010; Dasgupta and Kpotufe, 2014). In particular, Abraham et al. (2010); Dasgupta and Kpotufe (2014) use a simple and practical mode estimator $\text{argmax}_{x\in X_n} f_n(x)$, where $X_n = \{x_1, \ldots, x_n\}$ are sampled *i.i.d.* according to f. When f_n is the kernel density estimator, Abraham et al. (2010) shows that this mode estimator is consistent. On the other hand, when f_n is the kNN density estimator, Dasgupta and Kpotufe (2014) shows that this mode estimator is not only consistent but also converges at a minimax-optimal rate. The mode estimation results in the present paper are similar to the results presented in Dasgupta and Kpotufe (2014) when f is restricted to $\mathcal{S}^{d-1}$.

When the distribution is not unimodal, the best known practical approach for estimating all the modes is the mean-shift algorithm and its variations (Cheng, 1995; Fukunaga and Hostetler, 1975; Comaniciu and Meer, 2002; Li et al., 2007; Arias-Castro et al., 2016), which essentially consists of gradient ascent of a sufficiently smooth f_n starting from every sample point. Under the milder condition of requiring only that f is well-approximated by a quadratic in a neighborhood of each mode, Dasgupta and Kpotufe (2014) obtained a finite sample minimax-optimal rate on $\|\hat{x}_0 - x_0\|$, where $\hat{x}_0$ is estimate of any mode x_0, of the form $O\left(n^{-\frac{1}{4+d}}\right)$ that holds with high probability. Again, our multiple mode estimation results in the present paper are similar to the results presented in Dasgupta and Kpotufe (2014) when f is restricted to $\mathcal{S}^{d-1}$.

3 NOTATIONS

Throughout this paper, we assume access to $X_n = \{x_1, \ldots, x_n\}$ drawn *i.i.d.* from a distribution $\mathcal{F}$ over $\mathcal{S}^{d-1}$, with Lebesgue density function f. For any $A \subset \mathcal{S}^{d-1}$, we use $f(A) = \int_A f(x)dx$ to denote the probability mass of A under f. We use $\|\cdot\|$ to denote ℓ_2 norm and $\|\cdot\|_\infty$ to denote ℓ_∞ norm. For any $x \in \mathcal{S}^{d-1}$ and $r > 0$, we use $B(x, r) = \{x' \in \mathcal{S}^{d-1} : \|x - x'\| \leq r\}$ to denote the closed ball of radius r and centered at x, that intersects with $\mathcal{S}^{d-1}$. For any $A \subset \mathcal{S}^{d-1}$ and $r > 0$, we write $A_r = \{x \in \mathcal{S}^{d-1} : \inf_{y\in A} \|y - x\| \leq r\}$. We use $\tilde{O}(\cdot)$ to denote the 'big-Oh' notation that hides logarithmic factors. For clear exposition, most of our proofs are deferred to the supplementary material which is organized section-wise. However, when appropriate, we provide proof sketches for important theoretical results.

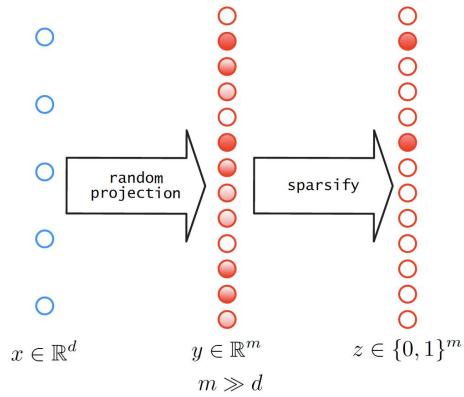


Figure 1: Visualization of the expand-and-sparsify architecture (Dasgupta and Tosh, 2020).

4 EXPAND-AND-SAPRSIFY DENSITY ESTIMATION

The expand-and-sparsify representation of an input $x \in \mathcal{X} \subset \mathbb{R}^d$ is given by the following transformation (Fig. 1).

1. A high-dimensional linear mapping, $y = \Theta x \in \mathbb{R}^m$.
2. A sparsification scheme to a binary vector $z \in \{0,1\}^m$, where

$$z_j = \begin{cases} 1 & \text{if } y_j \text{ is one of } k \text{ largest entries in } y \\ 0 & \text{otherwise} \end{cases}$$

Here, $\Theta \in \mathbb{R}^{m\times d}$ is a random matrix with rows $\theta_1, \ldots, \theta_m$ drawn i.i.d. from some measure ν. We denote the mapping $x \mapsto z$ via the function $h(x) \in \{0,1\}^m$, where each coordinate j is indexed as $h_j(x) = z_j$.

Observe that for any vector $x \in \mathbb{R}^d$, we have $h(cx) = h(x)$ for any $c > 0$. Thus, we will restrict ourselves to the setting where elements of $\mathcal{X}$ have unit length, i.e. $\mathcal{X} \subset \mathcal{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$. We will also focus on the case where ν is uniform over $\mathcal{S}^{d-1}$.

Let $C_j = \{x \in \mathcal{X} : h_j(x) = 1\}$ denote the inputs satisfying that the j-th coordinate in the expand-and-sparsify representation is set to 1. Since $\mathcal{X}$ and $\mathrm{supp}(\nu)$ are both subsets of $\mathcal{S}^{d-1}$, C_j consists exactly of those inputs x whose k closest elements of $\theta_1, \ldots, \theta_m$ contains θ_j.

Let f be some density over $\mathcal{X}$, and suppose $x_1, \ldots, x_n$ are drawn *i.i.d.* from f. Let $f_n(C_j) = \frac{1}{n}\#\{i : x_i \in C_j\}$ denote the fraction of sample points whose jth coordinate is set to 1. When the measure ν is uniform over $\mathcal{S}^{d-1}$, the expand-and-sparsify density estimator is given by

$$\begin{aligned}\hat{f}_n(x) &= \frac{m}{k^2 S_{d-1}} \sum_{j=1}^{m} f_n(C_j) h_j(x) \\ &= \frac{m}{k^2 S_{d-1}} \sum_{j:h_j(x)=1} f_n(C_j), \end{aligned} \tag{1}$$

where $S_{d-1} = \frac{2\pi^{d/2}}{\Gamma(d/2)}$ is the surface area of $\mathcal{S}^{d-1}$.

To see where this estimator comes from, consider the following (computationally intractable*) estimator:

$$\tilde{f}_n(x) = \frac{1}{k}\sum_{j=1}^{m} \frac{f_n(C_j)}{\mathrm{vol}(C_j)} h_j(x) = \frac{1}{k}\sum_{j:h_j(x)=1} \frac{f_n(C_j)}{\mathrm{vol}(C_j)},$$

where $\mathrm{vol}(C_j)$ is the $d-1$-dimensional volume of C_j, i.e. the corresponding hypersurface area. In this light, $f_n(C_j)$ is an estimate of the integral $f(C_j) = \int_{C_j} f(x)\,dx$, and normalizing this value by $\mathrm{vol}(C_j)$ gives us an estimate of $f(u)$ for u drawn uniformly from C_j. If f is smooth, C_j is small in diameter, and $x \in C_j$, then $f_n(C_j)/\mathrm{vol}(C_j)$ will be close to $f(x)$. Averaging over all j such that $x \in C_j$ gives us the estimator $\tilde{f}$. Moving from $\tilde{f}_n$ to $\hat{f}_n$ is a matter of showing that $\mathrm{vol}(C_j) \approx \frac{k}{m} S_{d-1}$ when ν is uniform over $\mathcal{S}^{d-1}$, which we demonstrate in Equation (2), below.

5 DENSITY ESTIMATION RATES

In this section, we analyze the convergence rate of the expand-and-sparsify estimator $\hat{f}_n$. We first give a result for general densities with no smoothness assumptions, before specializing this result to a class of generalized Hölder smooth densities. Finally, we give a concrete matching lower bound for density estimation on the sphere, demonstrating that the expand-and-sparsify estimator is minimax optimal.

*C_j is a subset of a high-dimensional hypersphere that is only implicitly specified, and computing the volume of a high-dimensional set with a membership oracle is a known hard problem (Füredi and Bárány, 1986).

5.1 General density estimation rates

Our result for estimating general densities will be framed using the following notions of modulus of continuity that captures how much f changes in a neighborhood of any x.

Definition 5.1. For $x \in \mathcal{S}^{d-1}$ and $\epsilon > 0$, define $\hat{r}(\epsilon, x) \triangleq \sup\left\{r : \sup_{\|x-x'\|\le r} f(x') - f(x) \le \epsilon\right\}$, and $\check{r}(\epsilon, x) \triangleq \sup\left\{r : \sup_{\|x-x'\|\le r} f(x) - f(x') \le \epsilon\right\}$.

Our first observation is that with high probability each region C_j is sandwiched between two balls of nearly equal volume.

Lemma 5.2. *If $\delta > 0$ and $m > k > C_\delta^2 d \log m$, then with probability at least $1-\delta$ the following holds. For every $j = 1, \ldots, m$:*

$$B(\theta_j, r_{\text{inner}}) \subset C_j \subset B(\theta_j, r_{\text{outer}}),$$

where $C_\delta = c_0 \log(2/\delta)$ for some absolute constant c_0 and

$$\begin{aligned}\mathrm{vol}(B(\theta_j, r_{\text{inner}})) &= \frac{k}{m} S_{d-1}\left(1 - C_\delta\sqrt{\frac{d\log m}{k}}\right) \\ \mathrm{vol}(B(\theta_j, r_{\text{outer}})) &= \frac{k}{m} S_{d-1}\left(1 + C_\delta\sqrt{\frac{d\log m}{k}}\right).\end{aligned}$$

An immediate consequence of Lemma 5.2 is that with high probability, each region C_j satisfies $\mathrm{vol}(C_j) \approx S_{d-1}k/m$. Specifically, we have

$$1 - C_\delta\sqrt{\frac{d\log m}{k}} \le \frac{\mathrm{vol}(C_j)}{S_{d-1}k/m} \le 1 + C_\delta\sqrt{\frac{d\log m}{k}}. \tag{2}$$

Our second observation is that the empirical f_n-mass of each C_j will tightly approximate its true f-mass, essentially a consequence of Bernstein's inequality.

Lemma 5.3. *Suppose $\delta > 0$. With probability at least $1-\delta$,*

$$|f_n(C_j) - f(C_j)| \le \sqrt{\frac{2f(C_j)\log(m/\delta)}{n}} + \frac{2\log(m/\delta)}{3n}$$

for all $j = 1, \ldots, m$.

Given these two results, we have the following general approximation result.

Lemma 5.4. *Pick $\delta > 0, k \ge 4C_\delta^2 d\log m$ and assume that ν is uniform over $\mathcal{S}^{d-1}$. Let $\gamma_n = \alpha_n/(S_{d-1}(k/m))$, where $\alpha_n = 2\sqrt{\frac{kS_{d-1}\|f\|_\infty \log(m/\delta)}{mn}} + \frac{2\log(m/\delta)}{3n}$. If the conclusions of Lemma 5.2 and*

Lemma 5.3 hold, the following is true for all $x \in \mathcal{S}^{d-1}$ and $\epsilon > 0$,

$$(i)\ \hat{f}_n(x) \leq \left(1 + C_\delta \sqrt{\frac{d \log m}{k}}\right)(f(x) + \epsilon) + \gamma_n$$

$$(ii)\ \hat{f}_n(x) \geq \left(1 - C_\delta \sqrt{\frac{d \log m}{k}}\right)(f(x) - \epsilon) - \gamma_n$$

provided $\operatorname{diam}(C_j) \leq \min\{\check{r}(\epsilon, x), \hat{r}(\epsilon, x)\}$ *for every* $j = 1, \ldots, m$.

5.2 Smooth density estimation rates

Moving to the smooth setting, we will prove convergence rates for the expand-and-sparsify density estimator for densities satisfying the following generalization of Hölder smoothness.

Definition 5.5. We say that $f : \mathcal{X} \to \mathbb{R}$ is L, β, r-smooth if for all $x, x' \in \mathcal{X}$ satisfying $\|x - x'\| \leq r$, we have $|f(x) - f(x')| \leq L\|x - x'\|^\beta$. We say f is L, β-smooth if it is L, β, r-smooth for any $r > 0$.

Our analysis here is based on applying Lemma 5.4, which shows that so long as the diameters of the regions C_j are sufficiently small, then $\hat{f}_n$ provides an accurate estimate of f. The following result, which follows readily from Lemma 5.2, provides bounds on these diameters.

Lemma 5.6. *Let $k \geq C_\delta^2 d \log m$. If the conclusion of Lemma 5.2 holds, then*

$$\operatorname{diam}(C_j) \leq \frac{4}{\sqrt{3}} \left(\frac{6\sqrt{d}k}{m}\right)^{1/(d-1)}$$

for all $j = 1, \ldots, m$.

Combining Lemmas 5.4 and 5.6 gives us the following result.

Theorem 5.7. *Say f is L, β, r-smooth and ν is uniform over $\mathcal{S}^{d-1}$. Pick $\delta > 0$,*
$k \geq \max\{4C_\delta^2 d \log m, L^{-2} r^{-2\beta}\}$, and
$m \geq 6\sqrt{d}\left(\frac{4}{\sqrt{3}}\right)^{d-1} L^{(d-1)/\beta} k^{1+\frac{d-1}{2\beta}}$. Then with probability at least $1 - 2\delta$,

$$\|f - \hat{f}_n\|_\infty \leq \left(\|f\|_\infty C_\delta \sqrt{d \log m} + 2\right) k^{-1/2} + \gamma_n,$$

where γ_n was defined in Lemma 5.4.

To give an intuition on how the above pieces fit together, we provide here a short sketch of the proof of Theorem 5.7.

Proof sketch. By Definitions 5.1 and 5.5, we have $\hat{r}(\epsilon, x), \check{r}(\epsilon, x) \geq \left(\frac{\epsilon}{L}\right)^{1/\beta}$ for all $x \in \mathcal{S}^{d-1}$ and $\epsilon > 0$. Applying Lemma 5.6 with large enough m, we can then guarantee that:

$$\operatorname{diam}(C_j) \leq \left(\frac{\epsilon}{L}\right)^{1/\beta} \leq \min\{\hat{r}(\epsilon, x), \check{r}(\epsilon, x)\}.$$

With this bound in hand, we can then apply Lemma 5.4 to finish the proof. □

Instantiating Theorem 5.7 with $m = \Theta(n)$ and $k = \Theta\left(m^{\frac{2\beta}{2\beta+(d-1)}} (\log m)^{\frac{d-1}{2\beta+(d-1)}}\right)$ leads to

$$\sup_{x \in \mathcal{S}^{d-1}} |f(x) - \hat{f}_n(x)| = O\left(\left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+(d-1)}}\right).$$

As we will show in the next section, this rate is in fact minimax optimal.

5.3 Lower bounds on density estimation rates

Next, we derive a matching density estimation lower bound for any L, β-smooth density, whose support is restricted to $\mathcal{S}^{d-1}$. Our main result is as follows.

Theorem 5.8. *Let $\mathcal{P}$ be the class of probability distributions whose densities are supported on $\mathcal{S}^{d-1}$ and are L, β smooth. For any $P \in \mathcal{P}$, we denote its density by f_P. Let $x_1, \ldots, x_n$ be a sample from some distribution $P \in \mathcal{P}$ and let $\hat{f}_n = \hat{f}_n(x_1, \ldots, x_n)$ be an estimator of f_P. Then, for $d > 4$ and $n \geq 16$, the following holds.*

$$\inf_{\hat{f}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_P\left(\|\hat{f}_n - f_P\|_\infty\right) = \Omega\left(\left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+(d-1)}}\right).$$

We present a short sketch of the proof here. It is structurally similar to other lower bound proofs for smooth density estimation in $\mathbb{R}^d$, e.g. (Khas'minskii, 1979; Tsybakov, 2008), but adapted here to the setting of the sphere.

Proof sketch. The high level idea is to find a packing of the unit sphere and construct a finite set of smooth functions, using this packing, with non-overlapping support. We then construct a finite family of L, β smooth densities using the functions from above. We can show that these densities are close to each other in KL divergence but far from each other in ℓ_∞ distance. An application of Fano's inequality (Tsybakov, 2008) finishes the proof. □

6 MODE ESTIMATION

We now present analyses of two simple algorithms that use the expand-and-sparsify representation for estimating modes of an unknown density. We start with the following definition of modes.

Definition 6.1. We denote the set of modes of a density f by:

$$\mathcal{M} = \{x : \exists r > 0, \forall x' \in B(x,r), f(x') < f(x)\}.$$

We follow the general approach of (Dasgupta and Kpotufe, 2014) for our mode estimation results and make the following assumption.

Assumption 6.2. f is twice differentiable in a neighborhood of every $x \in \mathcal{M}$. We denote the gradient and Hessian of f by ∇f and $\nabla^2 f$. Furthermore, $\nabla^2 f(x)$ is negative definite at all $x \in \mathcal{M}$.

Assumption 6.2, like most previous work, considers only interior modes and excludes modes at the boundary of the support of f, where it is not continuously differentiable. A direct implication of Assumption 6.2 is that for any $x \in \mathcal{M}, \nabla f$ is continuous in a neighborhood of x, with $\nabla f(x) = 0$. Additionally, negative definiteness of $\nabla^2 f(x)$ at x allows f to be well approximated by a quadratic in a neighborhood of x as detailed below.

Lemma 6.3 (Dasgupta and Kpotufe (2014))**.** *Let f satisfy Assumption 6.2. Consider any $x \in \mathcal{M}$. Then there exists a neighborhood $B(x,r), r > 0$, and constants $\hat{C}_x, \check{C}_x > 0$ such that, for all $x' \in B(x,r)$, we have*

$$\check{C}_x \|x' - x\|^2 \leq f(x) - f(x') \leq \hat{C}_x \|x' - x\|^2. \quad (3)$$

The above lemma allows us to parameterize any mode $x \in \mathcal{M}$ locally as defined below, which in turn ensures that $A_x \cap \mathcal{M} = \{x\}$.

Definition 6.4. For every mode $x \in \mathcal{M}$, there exists a $r_x > 0$, such that $B(x, r_x)$ is connected in a set A_x, satisfying (i) A_x is a connected component* of a level set $\mathcal{S}^{d-1}_{\lambda_x} \triangleq \{x' \in \mathcal{X} : f(x') > \lambda_x\}$ for some $\lambda_x > 0$, and, (ii) $\exists \hat{C}_x, \check{C}_x > 0, \forall x' \in A_x, \check{C}_x \|x' - x\|^2 \leq f(x) - f(x') \leq \hat{C}_x \|x' - x\|^2$.

Finally, we assume that every hill in f corresponds to a mode in $\mathcal{M}$.

Assumption 6.5. Each connected component of any level set $\mathcal{S}^{d-1}_{\lambda} = \{x \in \mathcal{S}^{d-1} : f(x) \geq \lambda\}, \lambda > 0$, contains a mode in $\mathcal{M}$.

6.1 Mode recovery guarantee for a unimodal density

We start with a simple setting when f has a single mode x_0 (that is, $|\mathcal{M}| = 1$). Our mode estimation algorithm is simple. We estimate the mode of f to be the maximizer of $\hat{f}_n$ over X_n, where $\hat{f}_n$ is computed using Equation (1).

$$\hat{x}_0 = \operatorname*{argmax}_{x \in X_n} \hat{f}_n(x) \quad (4)$$

We provide some intuition for why Equation (4) is a reasonable strategy. Assumption 6.2, specifically Definition 6.4, allows us to easily express an upper and a lower bound of density f at any point x in a small neighborhood of x_0 relative to $f(x_0)$. In addition, Lemma 5.4 allows us to approximate f at different scales in different parts of the space by varying ϵ. Judiciously combining these two facts allows us to ensure that the infimum of $\hat{f}_n(x)$ for any x within a small ball around x_0 is strictly larger than the supremum of $\hat{f}_n(x)$ for any x outside this ball. Our main result of this section for single mode recovery is the following.

Theorem 6.6. *Let $\delta > 0$. Assume that f has a single mode x_0 and satisfies Assumption 6.2. Suppose k satisfies*

$$(i) \min\left\{\alpha \cdot C_\delta^2 d \log m, C_\delta^2 d \log n\right\} \leq k$$

$$ii) k \leq \min\{\beta_1 \cdot m^{\frac{4}{d+3}}, \beta_2 \cdot n^{\frac{4}{d+3}}\}$$

where, $\alpha = \max\left\{4, \left(\frac{f(x_0)}{3}\right)^2, \left(\frac{32 f(x_0)}{\check{C}_{x_0} r_{x_0}^2}\right)^2\right\}, \beta_1 = \frac{1}{6}\left(\left(\frac{3 f(x_0) C_\delta}{16 \hat{C}_{x_0}}\right)^2 \log m\right)^{\frac{d-1}{d+3}}$, *and* $\beta_2 = \left(\frac{S_{d-1}}{12\sqrt{d}}\right)^{\frac{4}{d+3}} (f(x_0))^{\frac{2d+2}{d+3}} \left(\left(\frac{3 C_\delta}{4 \hat{C}_{x_0}}\right)^2 d \log m\right)^{\frac{d-1}{d+3}}$. *Let $\hat{x}_0$ be our mode estimate using Equation (4). If $n \geq \frac{9m \log(m/\delta) \max\{\|f\|_\infty, 1\}}{S_{d-1}(f(x_0) C_\delta)^2 d \log m}$, then with probability at least $1 - 2\delta$, we have,*

$$\|\hat{x}_0 - x_0\| \leq 6\sqrt{\frac{C_\delta}{\check{C}_{x_0}} f(x_0)} \left(\frac{d \log m}{k}\right)^{1/4}.$$

Remark 6.7. In Theorem 6.6, the upper-bound on k is due to two different events: the first term roughly translates to $k = \tilde{O}\left(m^{\frac{4}{d+3}}\right)$ and ensures that all $\text{diam}(C_j)$ are sufficiently small so that Lemma 5.4 can be applied, while the second term, which roughly translates to $k = \tilde{O}\left(n^{\frac{4}{d+3}}\right)$, ensures that there are enough data points from the sample X_n that fall within a small ball centered at the mode x_0,

Proof sketch. We define, $r_n(x_0) = \inf\{r : B(x_0, r) \cap X_n \neq \emptyset\}$ and choose $\tau \in (0,1)$ and $\tilde{r}$ satisfying $2r_n(x_0)/\tau \leq \tilde{r} \leq r_{x_0}$. Since $\tilde{r} \geq 2r_n(x_0)$, the crux of our proof technique relies on establishing,

$$\sup_{x \in \mathcal{S}^{d-1} \setminus B(x_0, \tilde{r})} \hat{f}_n(x) < \inf_{x \in B(x_0, r_n(x_0))} \hat{f}_n(x) \quad (5)$$

*For any $S \subset \mathcal{S}^{d-1}$, we say that x and y are connected is S if $x = \mathcal{P}(0)$ and $y = \mathcal{P}(1)$ where $\mathcal{P}$ is any continuous (path) function $\mathcal{P} : [0,1] \to S$. We say S is connected if it has a single connected component.

as this will ensure that $\|\hat{x}_0 - x\| \leq \tilde{r}$. To see this, note that if $\sup_{x \in B(x_0,\tilde{r}) \setminus B(x_0, r_n(x_0))} \hat{f}_n(x) > \sup_{x \in \mathcal{S}^{d-1} \setminus B(x_0,\tilde{r})} \hat{f}_n(x)$, then it must be the case that $\hat{x}_0 \in B(x_0, \tilde{r})$. On the other hand, if $\sup_{x \in B(x_0,\tilde{r}) \setminus B(x_0, r_n(x_0))} \hat{f}_n(x) \leq \sup_{x \in \mathcal{S}^{d-1} \setminus B(x_0,\tilde{r})} \hat{f}_n(x)$, then $\hat{x}_0 \in B(x_0, r_n(x_0)) \subset B(x_0, \tilde{r})$. When $\tilde{r}$ is appropriately chosen, this yields the claimed result.

To prove Equation (5), we repeatedly invoke Lemma 5.4 with different ϵ values. In particular, we find the upper bound of the l.h.s. of Equation (5) by invoking Lemma 5.4 with properly tuned ϵ, and show that this is less than the lower bound of the r.h.s. of Equation (5) obtained by invoking again Lemma 5.4 with properly tuned (but different than before) ϵ. □

Now, with the appropriate choice of hyperparameters, we immediately get the following corollary.

Corollary 6.8. *Setting $m = \Theta(n)$, and $k = O\left((d \log m)^{\frac{d-1}{d+3}} n^{\frac{4}{d+3}}\right)$ in Theorem 6.6, we have that with probability at least $1 - 2\delta$,*

$$\|\hat{x}_0 - x_0\| = O\left(\left(\frac{d \log m}{n}\right)^{\frac{1}{d+3}}\right) = \tilde{O}\left(n^{-\frac{1}{d+3}}\right).$$

When f is restricted to $\mathcal{S}^{d-1}$, there are effectively $d-1$ degrees of freedom, and we note that the rate of Corollary 6.8, ignoring logarithmic factors, is optimal for any twice-differentiable density f (Tsybakov, 1990; Donoho and Liu, 1991) and matches the results of Dasgupta and Kpotufe (2014).

6.2 Mode recovery guarantee in presence of multiple modes

Next, we address the mode estimation problem when f has an unknown number of modes. Our solution technique relies on estimating the connected components (CCs) of various level sets of the unknown density, so that, when the modes are sufficiently well separated from each other by apropriate low density regions, the maximizer of $\hat{f}_n$ within any CC of an appropriate level set will be a good approximation of the mode contained within that CC. It is well known that CCs of the mutual k-NN graph and its variants constructed using samples drawn *i.i.d.* from f are good estimates of the CCs of the corresponding levels sets of the unknown density f (Kpotufe and von Luxburg, 2011; Chaudhuri and Dasgupta, 2010; Balakrishnan et al., 2013; Dasgupta and Kpotufe, 2014). In order to estimate the CCs of level sets of f, we define the set of nested graphs $G(\lambda)$ using $\hat{f}_n$ as given in Equation (1), which are sub-graphs of a mutual k-NN graph on the sample X_n.

Algorithm 1 Recovery of any mode

Input: $X_n, \tilde{\epsilon}$
Initialize $\mathcal{M}_n \leftarrow \emptyset$.
for $i = 1$ to n, in decreasing order of $\hat{f}_n(x_i)$ **do**
  $\lambda \leftarrow \hat{f}_n(x_i)$
  Let $\{\tilde{A}_j\}_{j=1}^{m_i}$ be the CC of $G(\lambda - \tilde{\epsilon})$ disjoint from $\mathcal{M}_n$
  $\mathcal{M}_n \leftarrow \mathcal{M}_n \cup \left\{x'_j \triangleq \text{argmax}_{x \in \tilde{A}_j \cap X_{n,\lambda}} \hat{f}_n(x)\right\}_{j=1}^{m_i}$
end for
Return $\mathcal{M}_n$

Definition 6.9. Given $\lambda \in \mathbb{R}$, let $G(\lambda)$ denotes the graph whose vertices are in $X_{n,\lambda} \triangleq \{x \in X_n : \hat{f}_n(x) \geq \lambda\}$, and whose vertices x, x' are connected by an edge if $\|x - x'\| \leq \alpha \cdot \min\{r_k(x), r_k(x')\}$, for some $\alpha \geq \sqrt{2}$, where $r_k(y)$ is the distance from y to its kth nearest neighbor in X_n.

When the density f has multiple modes, we guarantee that the modes which are well separated by other modes by appropriate low density regions – termed as *salient* modes – can be recovered under mild conditions. In order to define salient modes we first define the notion of r-separation as in Chaudhuri and Dasgupta (2010).

Definition 6.10 (r-separation)**.** $A, A' \subset \mathcal{S}^{d-1}$ are r-separated if there exists a separating set $S \subset \mathcal{S}^{d-1}$ such that every path from A to A' crosses S and $\sup_{x \in S_r} f(x) < \inf_{x \in A \cup A'} f(x)$.

Following Dasgupta and Kpotufe (2014), we define a mode x to be r-salient if the critical set A_x of Definition 6.4 is well separated from all components at the level where it appears.

Definition 6.11 (r-salient modes)**.** A mode x of f is said to be r-salient for $r > 0$ if the following holds. There exists A_x as in Definition 6.4, which is a CC of say $\mathcal{S}^{d-1}_{\lambda_x} \triangleq \{x \in \mathcal{S}^{d-1} : f(x) \geq \lambda_x\}$ and A_x is r-separated from $\mathcal{S}^{d-1}_{\lambda_x} \setminus A_x$.

Our proposed approach is summarized in Alg. 1, where, $\tilde{\epsilon} = \tilde{\epsilon}(n)$ is an user defined hyper-parameter which is a decreasing function of n. In particular, choosing $\tilde{\epsilon} = \gamma_n$, where γ_n was defined in Lemma 5.4 suffices. Alg. 1 outputs a set of estimated modes $\mathcal{M}_n$. Our next theorem guarantees that any salient mode in $\mathcal{M}$ can be recovered by Alg. 1 at the optimal rate $\tilde{O}\left(n^{-\frac{1}{d+3}}\right)$ for the choice of $k = \tilde{O}\left(n^{\frac{4}{d+3}}\right)$ and $m = \Theta(n)$.

Theorem 6.12. *Assume f satisfies Assumptions 6.2 and 6.5. Suppose k satisfies (i) $\min\left\{\alpha_1 \cdot C_\delta^2 d \log m, C_\delta^2 d \log n\right\} \leq k$, and (ii) $k \leq \min\{\beta_1 \cdot m^{\frac{4}{d+3}}, \beta_2 \cdot n^{\frac{4}{d+3}}\}$, where,*

$$\alpha_1 = \max\left\{4, \left(\frac{f(x_0)}{3}\right)^2, \left(\frac{40 f(x_0)}{\bar{C}_{x_0} \min\left\{\left(\frac{r_{x_0}}{2}\right)^2, \left(\frac{r}{\alpha}\right)^2\right\}}\right)^2\right\},$$

$\beta_1 = \frac{1}{6}\left(\left(\frac{15\lambda_{x_0}C_\delta}{64\check{C}_{x_0}}\right)^2 \log m\right)^{\frac{d-1}{d+3}}$, *and* $\beta_2 = \left(\frac{S_{d-1}}{6\sqrt{d}}\right)^{\frac{4}{d+3}} (\lambda_{x_0})^{\frac{2d+2}{d+3}} \left(\left(\frac{15C_\delta}{16\check{C}_{x_0}}\right)^2 d\log m\right)^{\frac{d-1}{d+3}}$. *Let x_0 be an r-salient mode for some $r > 0$ and $\mathcal{M}_n$ be the set of modes returned by Alg. 1. If $n \geq \frac{9m\log(m/\delta)\max\{\|f\|_\infty,1\}}{S_{d-1}(f(x_0)C_\delta)^2 d\log m}$, then with probability at least $1-2\delta$, there exists $\hat{x}_0 \in \mathcal{M}_n$ such that,*

$$\|\hat{x}_0 - x_0\| \leq 7\sqrt{\frac{C_\delta}{\check{C}_{x_0}} f(x_0)}\left(\frac{d\log m}{k}\right)^{1/4}.$$

7 EMPIRICAL EVALUATIONS

Since our results hold for data lying on the unit sphere, for our empirical evaluations, we choose distributions that are supported on the unit sphere, namely, von Mises-Fisher distributions. For any $x \in \mathcal{S}^{d-1}$, the probability density of a von Mises-Fisher distribution is:

$$f(x) = \frac{\kappa^{\frac{d}{2}-1}}{(2\pi)^{\frac{d}{2}} I_{\frac{d}{2}-1}(\kappa)} \exp\left(\kappa\mu^\top x\right) \tag{6}$$

where, $\mu \in \mathcal{S}^{d-1}$ is the mean direction, $\kappa > 0$ is the concentration parameter, d is the dimension, and I is the modified Bessel function of the first kind. The distribution becomes narrower around μ with increasing κ. For our empirical evaluations, we sample data points from a mixture of two von Mises-Fisher distributions given by:

$$f(x) = wf_1(x) + (1-w)f_2(x) \tag{7}$$

where, for $i = 1, 2$, $f_i(x)$ is specified, using (6), by mean direction μ_i, concentration parameter κ_i and $w \in (0,1)$ is the mixture coefficient. We consider data dimensionality $d = 2$ and 3 for our experiments. We compare our proposed density estimation method (EaSDE) with kNN based density estimation method (KNNDE) of Dasgupta and Kpotufe (2014) and kernel density estimation method (KDE). For each of these methods we train the respective density estimators using the training set, select hyper-parameters using the validation set* and evaluate the performance on the test set. We measure the accuracy of the density estimators using empirical total variation distance (ETV) defined by $ETV(f,\hat{f}) = \frac{1}{2M}\sum_{i=1}^{M}|f(x_i) - \hat{f}(x_i)|$, where, f is the true density, $\hat{f}$ is the estimated density and $x_1, \ldots, x_M$ is the test set. We present our results in Figure **??**, where we plot expansion factor, which is m/d, on the x-axis and plot the ETV on the y-axis.

*See Appendix C for details of hyper-parameter selection, data generation process and other experimental details.

As can be seen from Figure **??**, ETV for EaSDE generally decreases with increasing expansion factor. In sub-figure (a) and (b), we use smaller κ_1 and κ_2 values resulting in a flatter true distribution and using a reasonably smaller expansion factor, EaSDE achieves comparable performance to that of KNNDE and KDE. In sub-figure (c) and (d), we used higher values for κ_1 and κ_2 resulting in a concentrated component distributions around the component means. In this case, ETV is much higher for all methods, and KNNDE in particular performs poorly. Here EaSDE requires a larger expansion factor to reduce ETV, more so for higher d.

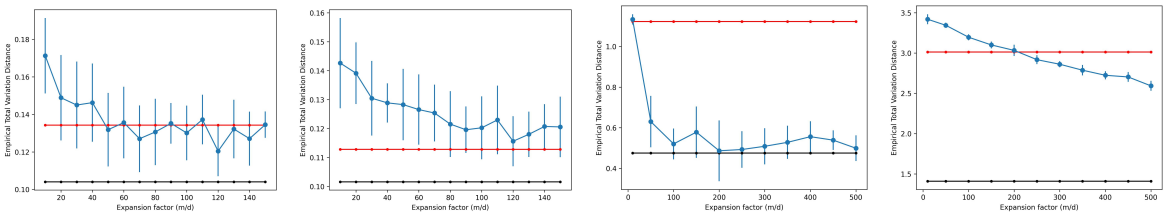
8 CONCLUSIONS AND LIMITATIONS

In this paper we studied the suitability of expand-and-sparsify representations for two fundamental statistical problems – density estimation and mode estimation. For density estimation, we demonstrated that given n *i.i.d.* examples from a density f, there is an estimator $\hat{f}_n$ that is a linear function of the expand-and sparsify representation that achieves a minimax-optimal convergence rate in ℓ_∞ distance. Our bounds hold with high probability, simultaneously for all x in the support of f. For mode estimation, we presented simple algorithms on top of this density estimator and demonstrated that under mild conditions, modes of f can be recovered at a rate $\tilde{O}\left(n^{-\frac{1}{d+3}}\right)$ with high probability.

For mode estimation, we assume that f is twice differentiable in a neighborhood of each mode and our mode recovery rate is optimal up to a logarithmic factor. It is not immediately clear whether this extra logarithmic factor is an artifact of our proof or it is a limitation of our algorithm.

If the data is on a manifold that is not supported on the sphere, then our rates do not hold, since the sparsification scheme implies that $h(cx) = h(x)$ for any $c > 0$. But if the data comes from a submanifold of the sphere, then our rates still hold, but they are not adaptive to the intrinsic dimension of the submanifold. It is an interesting future direction to consider expand-and-sparsify schemes that adapt to the intrinsic dimension of a dataset.

The expand-and-sparsify representation also bears a strong resemblance to certain algorithms arising in the hyperdimensional computing literature. As a future research direction, it would be very interesting to explore the computational benefits of the expand-and-sparsify representation on non-traditional computer architectures.



(a) $d=2, \kappa_1=10, \kappa_2=5$ (b) $d=3, \kappa_1=10, \kappa_2=5$ (c) $d=2, \kappa_1=80, \kappa_2=100$ (d) $d=3, \kappa_1=80, \kappa_2=100$

Figure 2: Color code: Red – KNNDE, Black – KDE, BLUE (error bar) – EaSDE.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback. KS gratefully acknowledges funding from the "NSF AI Institute for Foundations of Machine Learning (IFML)" (FAIN:2019844).

References

C. Abraham, G. Biau, and B. Cadre. On the asymptotic properties of a simple estimate of the mode. *ESAIM: Probability and Statistics*, 8:1–11, 3 2010.

E. Arias-Castro, D. Mason, and B. Pelletier. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *Journal of Machine Learning Research*, 17(43):1–28, 2016.

S. Balakrishnan, S. Narayanan, A. Rinaldo, A. Singh, and L. Wasserman. Cluster trees on manifolds. In *Advances in Neural Information Processing Systems*, pages 2679–2687, 2013.

S. J. C. Caron, V. Ruta, L. F. Abbott, and R. Axel. Random convergence of olfactory inputs in the drosophila mushroom body. *Nature*, 497(7447):113–117, 2013.

M. Chacron, A. Longtin, and L. Maler. Efficient computation via sparse coding in electrosensory neural networks. *Current Opinion in Neurobiology*, 21(5): 752–760, 2011.

K. Chaudhuri and S. Dasgupta. Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems*, pages 343—351, 2010.

Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, 1995.

H. Chernoff. Estimation of the mode. *Annals of the Institute of Statistical Mathematics*, 16:31–41, 1964.

D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

S. Dasgupta and S. Kpotufe. Optimal rates for k-nn density and mode estimation. In *Advances in Neural Information Processing Systems*, pages 2555–2563, 2014.

S. Dasgupta and C. Tosh. Expressivity of expand-and-sparsify representations. *arXiv preprint arXiv:2006.03741*, 2020.

S. Dasgupta, C. F. Stevens, and S. Navlakha. A neural algorithm for a fundamental computing problem. *Science*, 358(6364):793–796, 2017.

S. Dasgupta, D. Hattori, and S. Navlakha. A neural theory for counting memories. *Nature communications*, 13(1):5961, 2022.

D. L. Donoho and R. C. Liu. Geometrizing Rates of Convergence, III. *The Annals of Statistics*, 19(2): 668 – 701, 1991.

W. F. Eddy. Optimum Kernel Estimators of the Mode. *The Annals of Statistics*, 8(4):870 – 882, 1980.

K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975.

Z. Füredi and I. Bárány. Computing the volume is difficult. In *Proceedings of the Eighteenth Annual ACM Symposium on Theory of Computing*, pages 442—447, 1986.

U. Grenander. Some Direct Estimates of the Mode. *The Annals of Mathematical Statistics*, 36(1):131 – 138, 1965.

B. Grund and P. Hall. On the minimisation of L^p error in mode estimation. *The Annals of Statistics*, 23(6): 2264 – 2284, 1995.

H. Jiang. Uniform convergence rates for kernel density estimation. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1694–1703, 2017.

R. Z. Khas'minskii. A lower bound on the risks of non-parametric estimates of densities in the uniform metric. *Theory of Probability & Its Applications*, 23 (4):794–798, 1979.

J. Klemelä. Adaptive estimation of the mode of a multivariate density. *Journal of Nonparametric Statistics*, 17(1):83–105, 2005.

S. Kpotufe and U. von Luxburg. Pruning nearest neighbor cluster trees. In *Proceedings of the 28th International Conference on Machine Learning*, pages 225—232, 2011.

J. Li, S. Ray, and B. G. Lindsay. A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*, 8 (59):1687–1723, 2007.

B. A. Olshausen and D. J. Field. Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14(4):481–487, 2004.

E. Parzen. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3):1065 – 1076, 1962.

P. Ram and K. Sinha. Federated nearest neighbor classification with a colony of fruit-flies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8097–8105, 2022.

A. Rinaldo and L. Wasserman. Generalized density clustering. *The Annals of Statistics*, 38(5):2678 – 2722, 2010.

K. Sinha. Non-parametric classification via expand-and-sparsify representation. In *Advances on Neural Information Processing Systems*, pages 29657–29696, 2024.

K. Sinha and P. Ram. Fruit-fly inspired neighborhood encoding for classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 1470–1480, 2021.

D. D. Stettler and R. Axel. Representations of odor in the piriform cortex. *Neuron*, 63(6):854–864, 2009.

A. B. Tsybakov. Recursive estimation of the mode of a multivariate distribution. *Probl. Peredachi Inf.*, 26 (1):38 – 45, 1990.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 1st edition, 2008.

G. C. Turner, M. Bazhenov, and G. Laurent. Olfactory representations by drosophila mushroom body neurons. *Journal of Neurophysiology*, 99(2):734–746, 2008.

H. Wen and H. Hang. Random forest density estimation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 23701–23722, 2022.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] Have not provided the code but provided sufficient details to reproduce the results.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Optimal Rates for Density and Mode Estimation with Expand-and-sparsify Representations: Supplementary Materials

A Proofs from section 5

A.1 Proof of Lemma 5.2

For a set $B \subset \mathcal{S}^{d-1}$, let

$$\nu_m(B) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{I}[\theta_j \in B],$$

where $\theta_1, \ldots, \theta_m$ are the random vectors drawn from ν.

Lemma A.1 (Chaudhuri and Dasgupta (2010))**.** *Assume $k \geq d \log m$ and fix $\delta > 0$. Then there is an absolute constant c_o such that with probability at least $1-\delta$, the following holds for every ball $B \subset \mathbb{R}^d$:*

$$\nu(B) \geq \frac{k}{m}\left(1 + C_\delta\sqrt{\frac{d\log m}{k}}\right) \implies \nu_m(B) \geq \frac{k}{m}$$

$$\nu(B) \leq \frac{k}{m}\left(1 - C_\delta\sqrt{\frac{d\log m}{k}}\right) \implies \nu_m(B) < \frac{k}{m},$$

where $C_\delta = 2c_o \log(2/\delta)$.

We can use Lemma A.1 to prove Lemma 5.2.

Proof of Lemma 5.2. First we observe that because ν is uniform over $\mathcal{S}^{d-1}$, for any value $p \in (0,1)$, there is a unique radius $r \in (0,2)$ such that $\nu(B(x,r)) = p$ for any $x \in \mathcal{S}^{d-1}$. Observe that the bounds on k imply that

$$0 < \frac{k}{m}\left(1 - C_\delta\sqrt{\frac{d\log m}{k}}\right) \leq \frac{k}{m}\left(1 + C_\delta\sqrt{\frac{d\log m}{k}}\right) < 2.$$

The proof that $C_j \subset B(\theta_j, r_{\text{outer}})$ follows directly from Lemma 11 in Dasgupta and Tosh (2020), but we include the proof here for completeness. Let us condition on the events of Lemma A.1 occurring.

By definition, every ball of radius r_{outer} centered at some $x \in \mathcal{S}^{d-1}$ has ν-mass

$$\nu(B(x, r_{\text{outer}})) = \frac{k}{m}\left(1 + C_\delta\sqrt{\frac{d\log m}{k}}\right).$$

By Lemma A.1, this implies that there are at least k θ_j's in each $B(x, r_{\text{outer}})$. Another way to say this is that every $x \in \mathcal{S}^{d-1}$ will have its nearest k θ_j's within distance r_{outer}. Thus, any $h_j(x)$ will only be activated for points x within distance r_{outer} of θ_j, i.e. $C_j \subset B(\theta_j, r_{\text{outer}})$.

To see that $B(\theta_j, r_{\text{inner}}) \subset C_j$ observe again that by definition, every ball of radius r_{inner} centered at some $x \in \mathcal{S}^{d-1}$ has ν-mass

$$\nu(B(x, r_{\text{inner}})) = \frac{k}{m}\left(1 - C_\delta\sqrt{\frac{d\log m}{k}}\right).$$

Thus, by Lemma A.1, every ball $B(x, r_{\text{inner}})$ will contain fewer than k points. Thus, if $x \in B(\theta_j, r_{\text{inner}})$ then $\theta_j \in B(x, r_{\text{inner}})$, and so θ_j must be among the k closest θ's to x, and so satisfy that $x \in C_j$.

To finish off the proof, we observe that $\nu(C) = \frac{\text{vol}(C)}{S_{d-1}}$ for any $C \subset \mathcal{S}^{d-1}$. □

A.2 Proof of Lemma 5.3

We will use the following form of Bernstein's bound.

Lemma A.2 (Bernstein bound)**.** *Let $X_1, \ldots, X_n$ be i.i.d. and suppose that $|X_i| \leq c$ and $\mathbb{E}(X_i) = \mu$. Pick $0 < \delta < 1$. Then with probability at least $1 - \delta$,*

$$\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| \leq \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}} + \frac{2c \log(1/\delta)}{3n}$$

where, $\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \mathrm{Var}(X_i)$.

Proof of Lemma 5.3. Fix any $j \in [m]$ and let $z_i = \mathbb{1}[x_i \in C_j]$. Then $|z_i| \leq 1$ and

$$\begin{aligned} \mathrm{Var}(z_i) &= \mathbb{E}(z_i^2) - (\mathbb{E} z_i)^2 = f(C_j) - (f(C_j))^2 \\ &= f(C_j)(1 - f(C_j) \leq f(C_j). \end{aligned}$$

Combining Lemma A.2 with a union bound, we can say that with probability at least $1 - \delta$,

$$|f_n(C_j) - f(C_j)| \leq \sqrt{\frac{2 f(C_j) \log(m/\delta)}{n}} + \frac{2 \log(m/\delta)}{3n}.$$

□

A.3 Proof of Lemma 5.4

Proof. Pick any $x \in \mathcal{S}^{d-1}$ and assume that $x \in C_j$ for some $j \in \{1, \ldots, m\}$. Since $\mathrm{diam}(C_j) \leq \hat{r}(\epsilon, x)$, we have $C_j \subset B(x, \hat{r}(\epsilon, x))$. This implies $f(C_j) \leq \mathrm{vol}(C_j)(f(x) + \epsilon)$. Next, note that, with probability at least $1 - \delta$, (2) holds. Conditioning on this occurring, and for the choice of k as given in the Lemma statement, we have that for every $j = 1, \ldots, m$,

$$f(C_j) \leq \|f\|_\infty \, \mathrm{vol}(C_j) \leq \frac{2k S_{d-1} \|f\|_\infty}{m}$$

Using this bound, Lemma 5.3 ensures that with probability at least $1 - \delta$,

$$|f_n(C_j) - f(C_j)| \leq 2\sqrt{\frac{k S_{d-1} \|f\|_\infty \log(m/\delta)}{mn}} + \frac{2 \log(m/\delta)}{3n} = \alpha_n$$

We condition on this occurring. We first show the upper bound.

$$\begin{aligned} \hat{f}_n(x) &= \frac{m}{k^2 S_{d-1}} \sum_{j : h_j(x) = 1} f_n(C_j) \leq \frac{1}{k} \sum_{j : h_j(x) = 1} \left(\frac{f(C_j)}{S_{d-1} k/m} + \frac{\alpha_n}{S_{d-1} k/m} \right) \\ &\leq \frac{1}{k} \sum_{j : h_j(x) = 1} \left(\frac{f(C_j)}{\mathrm{vol}(C_j)} \left(1 + C_\delta \sqrt{\frac{d \log m}{k}} \right) + \frac{\alpha_n}{S_{d-1} k/m} \right) \\ &\leq \frac{1}{k} \sum_{j : h_j(x) = 1} \left(\frac{\mathrm{vol}(C_j)(f(x) + \epsilon)}{\mathrm{vol}(C_j)} \left(1 + C_\delta \sqrt{\frac{d \log m}{k}} \right) + \frac{\alpha_n}{S_{d-1} k/m} \right) \\ &= \left(1 + C_\delta \sqrt{\frac{d \log m}{k}} \right) (f(x) + \epsilon) + \gamma_n \end{aligned}$$

The lower bound is shown in a similar manner. Pick any $x \in \mathcal{S}^{d-1}$ and assume that $x \in C_j$ for some $j \in \{1, \ldots, m\}$. Since $\mathrm{diam}(C_j) \leq \check{r}(\epsilon, x)$, we have $C_j \subset B(x, \check{r}(\epsilon, x))$.

This implies $f(C_j) \geq \mathrm{vol}(C_j)(f(x) - \epsilon)$. Then we have,

$$\begin{aligned}
\hat{f}_n(x) &= \frac{m}{k^2 S_{d-1}} \sum_{j:h_j(x)=1} f_n(C_j) \geq \frac{1}{k} \sum_{j:h_j(x)=1} \left(\frac{f(C_j)}{S_{d-1}k/m} - \frac{\alpha_n}{S_{d-1}k/m} \right) \\
&\geq \frac{1}{k} \sum_{j:h_j(x)=1} \left(\frac{f(C_j)}{\mathrm{vol}(C_j)} \left(1 - C_\delta \sqrt{\frac{d \log m}{k}} \right) - \frac{\alpha_n}{S_{d-1}k/m} \right) \\
&\geq \frac{1}{k} \sum_{j:h_j(x)=1} \left(\frac{\mathrm{vol}(C_j)(f(x)-\epsilon)}{\mathrm{vol}(C_j)} \left(1 - C_\delta \sqrt{\frac{d \log m}{k}} \right) - \frac{\alpha_n}{S_{d-1}k/m} \right) \\
&= \left(1 - C_\delta \sqrt{\frac{d \log m}{k}} \right) (f(x) - \epsilon) - \gamma_n.
\end{aligned}$$

□

A.4 Proof of Lemma 5.6

We will make use of the following result.

Lemma A.3 (Lemma 12 Dasgupta and Tosh (2020))**.** *Suppose ν is uniform over $\mathcal{S}^{d-1}$, $\theta \in \mathcal{S}^{d-1}$, and $r \in (0,1)$. Then*

$$r \leq \frac{2}{\sqrt{3}} \left(3\sqrt{d}\nu(B(x,r)) \right)^{1/(d-1)}.$$

Proof of Lemma 5.6. For any C_j, observe that by the conclusion of Lemma 5.2, we have $C_j \subset B(\theta_j, r_{\text{outer}})$, where

$$\nu(B(\theta_j, r_{\text{outer}})) \leq \frac{k}{m} \left(1 + C_\delta \sqrt{\frac{d \log m}{k}} \right) \leq \frac{2k}{m},$$

where we have used our lower bound on k. Thus, by Lemma A.3,

$$\begin{aligned}
\mathrm{diam}(C_j) &\leq 2r_{\text{outer}} \\
&\leq \frac{4}{\sqrt{3}} \left(3\sqrt{d}\nu(B(x, r_{\text{outer}}) \right)^{1/(d-1)} \\
&\leq \frac{4}{\sqrt{3}} \left(\frac{6\sqrt{d}k}{m} \right)^{1/(d-1)}.
\end{aligned}$$

□

A.5 Proof of Theorem 5.7

Proof. First observe that with probability at least $1 - 2\delta$, the conclusions of Lemma 5.2 and Lemma 5.3 hold. Assume that this is so, in which case Lemma 5.4 and Lemma 5.6 both hold as well.

We will select $\epsilon = \sqrt{\frac{1}{k}}$. Now observe that by the condition of smoothness, if we have $\epsilon \leq Lr^\beta$, then

$$\hat{r}(\epsilon, x), \check{r}(\epsilon, x) \geq \left(\frac{\epsilon}{L} \right)^{1/\beta}$$

for all $x \in \mathcal{S}^{d-1}$. Our choice of ϵ, k ensures the above does hold. Applying Lemma 5.6, we have

$$\mathrm{diam}(C_j) \leq \frac{4}{\sqrt{3}} \left(\frac{6\sqrt{d}k}{m} \right)^{1/(d-1)} \leq L^{-1/\beta} k^{-1/2\beta} = \left(\frac{\epsilon}{L} \right)^\beta,$$

where the second inequality follows from our lower bound on m and canceling terms and the final equality follows from our choice ϵ. Thus, the conditions of Lemma 5.4 are met, and we have that for any $x \in \mathcal{S}^{d-1}$,

$$\begin{aligned}
|f(x) - \hat{f}_n(x)| &\leq f(x) C_\delta \sqrt{\frac{d \log m}{k}} + \epsilon + \epsilon C_\delta \sqrt{\frac{d \log m}{k}} + \gamma_n \\
&\leq \left(\|f\|_\infty C_\delta \sqrt{d \log m} + 2 \right) k^{-1/2} + \gamma_n,
\end{aligned}$$

where we have used our choice of $\epsilon = \sqrt{1/k}$ and our lower bound of $k \geq C_\delta^2 d \log m$. □

Instantiating Theorem 5.7 with $m = \Theta(n)$ and $k = \Theta\left(m^{\frac{2\beta}{2\beta+(d-1)}} (\log m)^{\frac{d-1}{2\beta+(d-1)}}\right)$ leads to $\sup_{x\in\mathcal{S}^{d-1}} |f(x) - \hat{f}_n(x)| = \|f - \hat{f}_n\|_\infty = O\left(\left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+(d-1)}}\right)$.

A.6 Proof of Theorem 5.8

Proof. The high level idea of our proof is to first find a packing of the unit sphere, then construct a finite set of smooth functions, using such packing, with non-overlapping support. Finally, construct a finite family of L, β smooth densities using the smooth functions constructed before for which we can use Fano's minimax bound (Theorem A.5). Now we provide the details. For any positive integer k, we use the notation $[k]$ to denote the set $\{1, \ldots, k\}$.

Fix any $a \in \mathcal{S}^{d-1}$ and let $k : \mathcal{S}^{d-1} \to \mathbb{R}$ be a bounded function defined on $B(a, 1/2) \cap \mathcal{S}^{d-1}$ such that,

$$\int_{x\in B(a,1/2)\cap\mathcal{S}^{d-1}} k(x)dx = 0,\ \forall x, x' \in \mathcal{S}^{d-1}, |k(x) - k(x')| \leq \|x - x'\|.$$

Note that, based on our function definition, $k(x) = 0$ for any $x \notin B(a, 1/2) \cap \mathcal{S}^{d-1}$. Also, let $k_{\max} = \max_{x\in\mathcal{S}^{d-1}} |k(x)|$.

For any $h > 0$, using Lemma A.6, let $m = \frac{\sqrt{d}/2}{h^{d-1}}$ be the packing number of $\mathcal{S}^{d-1}$. Then there exists $S = \{x_1, \ldots, x_m\} \subset \mathcal{S}^{d-1}$ such that, $\|x_i - x_j\| > h$ and $C_i \cap C_j = \emptyset$ for $i \neq j$, where $C_i = B(x_i, h/2) \cap \mathcal{S}^{d-1}$ for $i = 1, \ldots, m$. For any such $x_j \in S$, define

$$g_j(x) = h^\beta k\left(\frac{x - x_j}{h} + a\right).$$

Let $y = \frac{x-x_j}{h} + a$. Then $dy = \left(\frac{1}{h}\right)^d dx$. It is easy to check that g_j has support in C_i and therefore,

$$\begin{aligned}\int_{\mathcal{S}^{d-1}} g_j(x)dx &= h^\beta \int_{\mathcal{S}^{d-1}} k\left(\frac{x - x_j}{h} + a\right) dx \\ &= h^{\beta+d}\left(\int_{y\in B(a,1/2)\cap\mathcal{S}^{d-1}} k(y)dy + \int_{y\notin B(a,1/2)\cap\mathcal{S}^{d-1}} k(y)dy\right) = 0\end{aligned}$$

Thus, g_js have non-overlapping support and each g_j is 1-Lipschitz. To see this, observe that for $\forall x, x' \in \mathcal{S}^{d-1}$,

$$|g_j(x) - g_j(x')| = h^\beta \left|k\left(\frac{x - x_j}{h} + a\right) - k\left(\frac{x' - x_j}{h} + a\right)\right| \leq h^\beta \left(\frac{\|x - x'\|}{h}\right)^\beta = \|x - x'\|^\beta$$

Next, we define the function class, where $S_{d-1} = \mathrm{vol}(\mathcal{S}^{d-1})$.

$$\mathcal{M} = \left\{f_j(x) = \frac{1}{S_{d-1}} + \frac{g_j(x)}{2S_{d-1}k_{\max}} : j = 1, \ldots, m\right\} \tag{8}$$

For $h < 1$, it is easy to see that $f_j(x) > 0$ for all $x \in \mathcal{S}^{d-1}$ and $j = 1, \ldots, m$. Also,

$$\int_{\mathcal{S}^{d-1}} f_j(x)dx = \frac{1}{S_{d-1}} \int_{\mathcal{S}^{d-1}} dx + \frac{1}{2S_{d-1}k_{\max}} \int_{\mathcal{S}^{d-1}} g_j(x)dx = 1$$

Therefore, each $f_i \in \mathcal{M}$ is a probability density function whose support is $\mathcal{S}^{d-1}$. It is easy to verify that that each $f_i \in \mathcal{M}$ is $\frac{1}{S_{d-1}k_{\max}}, \beta$ smooth. For each $f_i \in \mathcal{M}$ we denote by P_i its associated probability distribution.

In order to apply Fano's inequality, we need to compute, (i) an upper bound of $\max_{i\neq j} \mathtt{KL}(P_i, P_j)$, and (ii) a lower bound of $\min_{i\neq j} \|f_i - f_j\|_\infty$. Let's start with the KLdivergence first. Note that for any $x \in C_i$ and $i \in [m]$,

$$f_i(x) = \frac{1}{S_{d-1}} + \frac{g_i(x)}{2S_{d-1}k_{\max}}, f_j(x) = \frac{1}{S_{d-1}} \tag{9}$$

Note that for $h \leq 1$, for any $i \in [m]$, $f_i(x) \geq \frac{1}{S_{d-1}} - \frac{k_{\max}}{2S_{d-1}k_{\max}} = \frac{1}{2S_{d-1}}$. Using this lower bound, we have,

$$\begin{aligned}
\mathtt{KL}(P_i, P_j) \leq \chi^2(f_i, f_j) &= \int_{\mathcal{S}^{d-1}} \left(\frac{f_i(x)}{f_j(x)} - 1\right)^2 f_i(x)dx = \int_{S^{d-1}} \frac{(f_i(x) - f_j(x))^2}{f_j(x)} dx \\
&\leq 2S_{d-1} \int_{S^{d-1}} (f_i(x) - f_j(x))^2 dx \\
&= 2S_{d-1} \left(\int_{C_i \cup C_j} (f_i(x) - f_j(x))^2 dx + \int_{\mathcal{S}^{d-1} \setminus \{C_i \cup C_j\}} (f_i(x) - f_j(x))^2 dx \right) \\
&\overset{a}{=} 2S_{d-1} \int_{C_i \cup C_i} (f_i(x) - f_j(x))^2 dx \\
&= 2S_{d-1} \left(\int_{C_i} (f_i(x) - f_j(x))^2 dx + \int_{C_j} (f_i(x) - f_j(x))^2 dx \right) \\
&\overset{b}{\leq} 2S_{d-1} \left(\int_{C_i} \left(\frac{h^\beta k_{\max}}{2S_{d-1}k_{\max}}\right)^2 dx + \int_{C_j} \left(\frac{h^\beta k_{\max}}{2S_{d-1}k_{\max}}\right)^2 dx \right) \\
&= \left(\frac{h^{2\beta}}{2S_{d-1}}\right) \left(\int_{C_i} dx + \int_{C_j} dx \right) = \left(\frac{h^{2\beta}}{2S_{d-1}}\right) (\mathrm{vol}(C_i) + \mathrm{vol}(C_j)) \\
&\overset{c}{\leq} \left(\frac{h^{2\beta}}{S_{d-1}}\right) \frac{2S_{d-1}}{\sqrt{d}} h^{d-1} \\
&= \left(\frac{2}{\sqrt{d}}\right) h^{2\beta + (d-1)}.
\end{aligned} \tag{10}$$

where, equality a follows from the fact that outside any $C_i \cup C_j$, f_i and f_j are identical. Inequality b follows from (9) and using the maximum possible value of g_i. Inequality c follows from Lemma A.7.

Now, for $i \neq j$, we have,

$$\begin{aligned}
\|f_i - f_j\|_\infty &= \sup_{x \in \mathcal{S}^{d-1}} |f_i(x) - f_j(x)| = \sup_{x \in C_i \cup C_j} |f_i(x) - f_j(x)| \\
&\overset{a}{=} \sup_{x \in C_i \cup C_j} \left| \frac{g(x)}{2S_{d-1}k_{\max}} \right| = \frac{h^\beta}{2S_{d-1}}
\end{aligned} \tag{11}$$

where, equality a follows from (9).

Now set $h = \kappa \left(\frac{\log n}{n}\right)^{\frac{1}{2\beta + (d-1)}}$ for some $\kappa \leq \min\left\{\frac{1}{16}, \left(\frac{\sqrt{d}}{16\left(1 + \frac{2\beta}{d-1}\right)}\right)^{\frac{1}{2\beta+(d-1)}}\right\}$. Then it follows from Lemma A.4 that $\frac{n \max_{i \neq j} \mathtt{KL}(P_i, P_j) + \log 2}{\log m} \leq \frac{1}{2}$. The desired lower bound immediately follows from an application of Theorem A.5. □

Lemma A.4. *Suppose $d > 4$, $n \geq 16$, and $\kappa \leq \min\left\{\frac{1}{16}, \left(\frac{\sqrt{d}}{16\left(1 + \frac{2\beta}{d-1}\right)}\right)^{\frac{1}{2\beta+(d-1)}}\right\}$. Set $h = \kappa \left(\frac{\log n}{n}\right)^{\frac{1}{2\beta+(d-1)}}$ and $m = \frac{\sqrt{d}/2}{h^{d-1}}$. Then the following holds.*

$$\frac{n \max_{i \neq j} \mathtt{KL}(P_i, P_j) + \log 2}{\log m} \leq \frac{1}{2}.$$

Proof. Since $h = \kappa \left(\frac{\log n}{n}\right)^{\frac{1}{2\beta+(d-1)}}$, we have,

$$\max_{i \neq j} \mathtt{KL}(P_i, P_j) \leq \frac{2}{\sqrt{d}} h^{2\beta + (d-1)} = \frac{2\kappa^{2\beta+(d-1)}}{\sqrt{d}} \left(\frac{\log n}{n}\right).$$

Also we have,

$$\log m = \log\left(\frac{\sqrt{d}/2}{h^{d-1}}\right) = \log\left(\left(\frac{\sqrt{d}}{2}\right)\left(\frac{1}{\kappa}\right)^{d-1}\left(\frac{n}{\log n}\right)^{\frac{d-1}{2\beta+(d-1)}}\right)$$
$$= \frac{1}{2}\log\left(\frac{d}{4}\right) + (d-1)\log\left(\frac{1}{\kappa}\right) + \frac{(d-1)}{2\beta+(d-1)}\left(\log n - \log\log n\right)$$

Therefore,

$$\begin{aligned}
\frac{n\max_{i\neq j}\mathtt{KL}(P_i,P_j)+\log 2}{\log m} &\leq \frac{\frac{2}{\sqrt{d}}\kappa^{2\beta+(d-1)}\log n + \log 2}{\frac{1}{2}\log\left(\frac{d}{4}\right) + (d-1)\log\left(\frac{1}{\kappa}\right) + \frac{(d-1)}{2\beta+(d-1)}\left(\log n - \log\log n\right)} \\
&\leq \frac{\frac{2}{\sqrt{d}}\kappa^{2\beta+(d-1)}\log n}{\frac{(d-1)}{2\beta+(d-1)}\left(\log n - \log\log n\right)} + \frac{\log 2}{(d-1)\log\left(\frac{1}{\kappa}\right)} \\
&= \frac{2}{\sqrt{d}}\left(1+\frac{2\beta}{d-1}\right)\kappa^{2\beta+(d-1)}\frac{1}{\left(1-\frac{\log\log n}{\log n}\right)} + \frac{\log 2}{(d-1)\log\left(\frac{1}{\kappa}\right)} \\
&\leq \frac{4}{\sqrt{d}}\left(1+\frac{2\beta}{d-1}\right)\kappa^{2\beta+(d-1)} + \frac{\log 2}{(d-1)\log\left(\frac{1}{\kappa}\right)} \\
&\leq \frac{1}{4} + \frac{\log 2}{(d-1)\log\left(\frac{1}{\kappa}\right)} \\
&\leq \frac{1}{4} + \frac{\log 2}{\log\left(\frac{1}{\kappa}\right)} \leq \frac{1}{4} = \frac{1}{4} = \frac{1}{2}.
\end{aligned}$$

□

Theorem A.5 (Fano's minimax bound (Tsybakov, 2008))**.** *Let $\mathcal{P}$ be a set of distributions and let $x_1,\ldots,x_n$ be a sample from some distribution $P\in\mathcal{P}$. Let $\hat{\theta}=\hat{\theta}(x_1,\ldots,x_n)$ is an estimator of the parameter $\theta(P)$ taking values in a metric space with metric d. Let $F=\{P_1,\ldots,P_N\}\subset\mathcal{P}$. Then,*

$$\inf_{\hat{\theta}}\sup_{P\in\mathcal{P}}\mathbb{E}_P\left(d\left(\hat{\theta}-\theta(P)\right)\right) \geq \frac{s}{2}\left(1-\frac{n\alpha+\log 2}{\log N}\right)$$

where $s=\min_{i\neq j} d\left(\theta(P_i),\theta(P_j)\right)$, and $\alpha=\max_{i\neq j}\mathit{KL}(P_i,P_j)$.

Lemma A.6. *Pick any $h\in(0,1)$. Let N be the largest set of h-separated points on $\mathcal{S}^{d-1}$ using ℓ_2 norm, that is, there exists $x_1,\ldots,x_N\in\mathcal{S}^{d-1}$ such that for any $i\neq j$, $\|x_i-x_j\|>h$. Then, $N\leq\frac{(\sqrt{d}/2)}{h^{d-1}}$.*

Proof. It is clear that, for $i=1,\ldots,N$, $B(x_i,h/2)$ are disjoint. Then, it must be the case that, $\sum_{i=1}^N \mathrm{vol}(B(x_i,h/2))\leq\mathrm{vol}(B(\mathcal{S}^{d-1}))$. Using Lemma A.7, we get the desired bound on N. □

Lemma A.7. *Suppose $d\geq 2$ and $h\in(0,1)$. Let $\mathrm{vol}(\mathcal{S}^{d-1})=S_{d-1}$. Then, for any $x\in\mathcal{S}^{d-1}$, we have,*

$$\mathrm{vol}(B(x,h)) \leq \frac{2S_{d-1}}{\sqrt{d}}h^{d-1}.$$

Proof. Let ν is the uniform distribution over $\mathcal{S}^{d-1}$.We first show that $\nu(B(x,h))\leq\frac{2}{\sqrt{d}}h^{d-1}$, and then use the fact that $\mathrm{vol}(B(x,h))=\nu(B(x,h))S_{d-1}$. Towards bounding $\nu(B(x,h))$ from above, we can generate a random sample $\theta=(\theta_1,\ldots,\theta_d)\sim\nu$ by first drawing $Y_1,\ldots,Y_d$ independently from a standard normal distribution, and then taking

$$\theta_i = \frac{Y_i}{(Y_1^2+\cdots+Y_d^2)^{1/2}}.$$

This works because the distribution $Y=(Y_1,\ldots,Y_d)$ is spherically symmetric. Now for any k,the sum $Y_1^2+\cdots,Y_d^2$ is a chi-squared distribution with k degrees of freedom , denoted by $\chi^2(k)$. It is well known that if

$A \sim \chi^2(k)$ and $B \sim \chi^2(l)$ are independent, then $A/(A+B)$ has a Beta$(k/2, l/2)$ distribution. Thus $\theta_1^2 = Y_1^2/(Y_1^2 + \cdots + Y_d^2)$ follows a Beta$(1/2, (d-1)/2)$ distribution. Without loss of generality, assume $x = e_1$. Then,

$$\theta \in B(x,h) \iff \|\theta - e_1\|^2 \le h^2 \iff \theta_1 \ge 1 - h^2/2.$$

Letting $\epsilon = 1 - (1 - h^2/2)^2 = h^2(1 - h^2/4)$ we get,

$$\begin{aligned}
\nu(B(x,h)) &= \frac{1}{2} \cdot \Pr\left(\theta_1^2 \ge (1 - (h^2/2))^2\right) = \frac{1}{2} \cdot \Pr\left(\theta_1^2 \ge 1 - \epsilon\right) \\
&\le \frac{1}{2} \cdot \frac{1}{B(1/2, (d-1)/2)} \cdot \frac{\epsilon^{(d-1)/2}}{(d-1)/2} \cdot (1-\epsilon)^{-1/2} \\
&= \frac{1}{2} \cdot \frac{\Gamma(d/2)}{\Gamma(1/2)\Gamma((d-1)/2)(d-1)/2} \cdot h^{d-1} \cdot (1 - h^2/4)^{(d-1)/2} \cdot \frac{1}{(1 - h^2/2)} \\
&\le \frac{\Gamma(d/2)}{\Gamma(1/2)\Gamma((d-1)/2)(d-1)/2} \cdot h^{d-1} \\
&= \frac{1}{\sqrt{\pi}(d-1)/2} \cdot \frac{\Gamma(d/2)}{\Gamma(d/2 - 1/2)} \cdot h^{d-1} \\
&\le \frac{\sqrt{d/2}}{\sqrt{\pi}(d-1)/2} \cdot h^{d-1} \\
&\le \frac{\sqrt{d/2}}{\sqrt{\pi}(d/4)} \cdot h^{d-1} \le \frac{2}{\sqrt{d}} \cdot h^{d-1}
\end{aligned}$$

where, the first inequality is due to Lemma A.8, the second inequality is due to the fact that $h \in (0,1)$, the third inequality is due the fact that $\Gamma(x)/\Gamma(x - 0.5) \le \sqrt{x}$ and the fourth inequality is due to the condition on d. □

Lemma A.8 (Lemma 13 Dasgupta and Tosh (2020))**.** *Suppose Z has Beta(α, β) distribution with $\alpha \le 1$ and $\beta \ge 1$. For any $0 < \epsilon < 1$,*

$$\frac{1}{B(\alpha,\beta)} \cdot \frac{\epsilon^\beta}{\beta} \le \Pr(Z \ge 1 - \epsilon) \le \frac{1}{B(\alpha,\beta)} \cdot \frac{\epsilon^\beta}{\beta} \cdot (1-\epsilon)^{\alpha - 1}.$$

B Proofs from Section 6

B.1 Proof of Theorem 6.6

Proof. Let $x_0 \in \mathcal{M}$ and define

$$r_n(x_0) = \inf\{r : B(x_0, r) \cap X_n \ne \emptyset\}.$$

Let $\tau \in (0,1)$ be a constant to be fixed later. Assume for now that $r_n(x_0) \le \frac{\tau}{2} r_{x_0}$. In due course, we will show that this holds with high probability. Also, consider $\tilde{r}$, to be appropriately chosen later, satisfying the following:

$$\frac{2r_n(x_0)}{\tau} \le \tilde{r} \le r_{x_0}.$$

Since $\tilde{r} \ge 2r_n(x_0)$, our goal is to show

$$\sup_{x \in \mathcal{S}^{d-1} \setminus B(x_0, \tilde{r})} \hat{f}_n(x) < \inf_{x \in B(x_0, r_n(x_0))} \hat{f}_n(x)$$

as this will ensure that $\|\hat{x}_0 - x\| \le \tilde{r}$. To see this, note that if $\sup_{x \in B(x_0,\tilde{r}) \setminus B(x_0, r_n(x_0))} \hat{f}_n(x) > \sup_{x \in \mathcal{S}^{d-1} \setminus B(x_0,\tilde{r})} \hat{f}_n(x)$, then $\hat{x}_0 \in B(x_0, \tilde{r})$. On the other hand, if $\sup_{x \in B(x_0,\tilde{r}) \setminus B(x_0, r_n(x_0))} \hat{f}_n(x) \le \sup_{x \in \mathcal{S}^{d-1} \setminus B(x_0,\tilde{r})} \hat{f}_n(x)$, then $\hat{x}_0 \in B(x_0, r_n(x_0)) \subset B(x_0, \tilde{r})$.

We first bound $\sup_{x \in \mathcal{S}^{d-1} \setminus B(x_0,\tilde{r})} \hat{f}_n(x)$ from above. Using definition 6.4, observe that,

$$\sup_{x \in A_{x_0} \setminus B(x_0, \tilde{r}/2)} f(x) \le f(x_0) - \check{C}_{x_0}(\tilde{r}/2)^2 \triangleq \hat{F} \tag{12}$$

To apply Lemma 5.4, note the following. For any $x \in \mathcal{S}^{d-1} \setminus B(x_0, \tilde{r}/2), f(x) \leq \hat{F}$. This is because A_{x_0} is a level set of unimodal f, that is, $\sup_{x \notin A_{x_0}} f(x) \leq \inf_{x \in A_{x_0}} f(x)$. Now, for any $x \in \mathcal{S}^{d-1} \setminus B(x_0, \tilde{r})$, let $\epsilon = \hat{F} - f(x)$. Then it is easy to see that $\hat{r}(\epsilon, x) \geq \tilde{r}/2$. Therefore, for every $j = 1., \ldots, m$, if we can somehow restrict $\text{diam}(C_j) \leq \frac{\tilde{r}}{2}$, then it automatically ensures that $\text{diam}(C_j) \leq \hat{r}(\epsilon, x)$. Applying Lemma 5.4, the following holds with probability at least $1 - \delta$,

$$\begin{aligned}
\sup_{x \in \mathcal{S}^{d-1} \setminus B(x_0, \tilde{r})} \hat{f}_n(x) &\leq \left(1 + C_\delta \sqrt{\frac{d \log m}{k}}\right)(f(x) + \epsilon) + \gamma_n \\
&= \left(1 + C_\delta \sqrt{\frac{d \log m}{k}}\right) \hat{F} + \gamma_n \\
&= \left(1 + C_\delta \sqrt{\frac{d \log m}{k}}\right)(f(x_0) - \check{C}_{x_0}(\tilde{r}/2)^2) + \gamma_n
\end{aligned} \tag{13}$$

Now we bound $\inf_{x \in B(x_0, r_n(x_0)} \hat{f}_n(x)$ from below. Since $\tau \tilde{r} \leq \tau r_{x_0} \leq r_{x_0}$, using definition 6.4, observe that,

$$\inf_{x \in B(x_0, \tau \tilde{r})} f(x) \geq f(x_0) - \hat{C}_{x_0}(\tau \tilde{r})^2 \triangleq \check{F} \tag{14}$$

Now for any $x \in B(x_0, r_n(x_0))$, let $\epsilon = f(x) - \check{F}$. Then we have $\check{r}(\epsilon, x) \geq \tau \check{r} - r_n(x_0) \geq \tau \tilde{r} - \tau \tilde{r}/2 = \tau \tilde{r}/2$. Therefore, for every $j = 1., \ldots, m$, if we can somehow restrict $\text{diam}(C_j) \leq \frac{\tau \tilde{r}}{2}$, then it automatically ensures that $\text{diam}(C_j) \leq \check{r}(\epsilon, x)$. Again, applying Lemma 5.4, the following holds with probability at least $1 - \delta$,

$$\begin{aligned}
\inf_{x \in B(x_0, r_n(x_0))} \hat{f}_n(x) &\geq \left(1 - C_\delta \sqrt{\frac{d \log m}{k}}\right)(f(x) - \epsilon) + \gamma_n \\
&= \left(1 - C_\delta \sqrt{\frac{d \log m}{k}}\right) \check{F} - \gamma_n \\
&= \left(1 - C_\delta \sqrt{\frac{d \log m}{k}}\right)(f(x_0) - \hat{C}_{x_0}(\tau \tilde{r})^2) - \gamma_n
\end{aligned} \tag{15}$$

Now, we can pick τ and $\tilde{r}$ such that the r.h.s. of (13) is less than the r.h.s. of (15). It suffices to pick $\tau^2 = \frac{\check{C}_{x_0}}{8\hat{C}_{x_0}}$ and $\tilde{r}^2 \geq \frac{16}{\check{C}_{x_0}}\left(f(x_0) C_\delta \sqrt{\frac{d \log m}{k}} + \gamma_n\right)$. Using Lemma B.2 and the Theorem's assumption on k, it is easy to check that if $n \geq \frac{9m \log(m/\delta) \max\{\|f\|_\infty, 1\}}{S_{d-1}(f(x_0)C_\delta)^2 d \log m}$, then γ_n is at most $f(x_0) C_\delta \sqrt{\frac{d \log m}{k}}$ and therefore $\tilde{r}^2 \geq 32 f(x_0) \frac{C_\delta}{\check{C}_{x_0}} \sqrt{\frac{d \log m}{k}}$. Therefore, we are free to choose $\tilde{r} \geq \max\left\{\sqrt{32 f(x_0) \frac{C_\delta}{\check{C}_{x_0}} \sqrt{\frac{d \log m}{k}}}, \frac{2 r_n(x_0)}{\tau}\right\}$.

All that remains to show is that, (i) $r_n(x_0) \leq \frac{\tau}{2} r_{x_0}$ with high probability, and (ii) $\text{diam}(C_j) \leq \min\{\tilde{r}/2, \tau \tilde{r}/2\}$ for every $j = 1., \ldots, m$. Let us focus on the first of the two. Let $r = \sqrt{32 f(x_0) \frac{C_\delta}{\check{C}_{x_0}} \sqrt{\frac{d \log m}{k}}}$. The Theorem's assumption on k ensures that $r \leq r_{x_0}$. Next, we show that with high probability, $|B(x_0, \tau r/2) \cap X_n| \geq \frac{k}{n} > 0$

which implies that $r_n(x_0) \le \tau r/2$. We first bound the probability mass of $f(B(x_0, \tau r/2))$ from below as follows:

$$\begin{aligned}
f(B(x_0,\tau r/2)) &\ge \mathrm{vol}(B(x_0,\tau r/2)) \inf_{x\in B(x_0,\tau r/2)} f(x) \\
&\overset{a}{\ge} \frac{1}{3\sqrt{d}}(3/4)^{\frac{d-1}{2}} S_{d-1}(\tau r/2)^{d-1} \inf_{x\in B(x_0,\tau r/2)} f(x) \\
&= \frac{1}{3\sqrt{d}}(3/4)^{\frac{d-1}{2}} S_{d-1}\left(f(x_0)\frac{C_\delta}{\hat{C}_{x_0}}\right)^{\frac{d-1}{2}}\left(\frac{d\log m}{k}\right)^{\frac{d-1}{4}} \inf_{x\in B(x_0,\tau r/2)} f(x) \\
&\overset{b}{\ge} \frac{1}{3\sqrt{d}} S_{d-1}\left(\frac{3f(x_0)}{4}\frac{C_\delta}{\hat{C}_{x_0}}\right)^{\frac{d-1}{2}}\left(\frac{d\log m}{k}\right)^{\frac{d-1}{4}}\left(f(x_0)-\hat{C}_{x_0}(\tau r/2)^2\right) \\
&= \frac{1}{3\sqrt{d}} S_{d-1}\left(\frac{3f(x_0)}{4}\frac{C_\delta}{\hat{C}_{x_0}}\right)^{\frac{d-1}{2}}\left(\frac{d\log m}{k}\right)^{\frac{d-1}{4}} f(x_0)\left(1-C_\delta\sqrt{\frac{d\log m}{k}}\right) \\
&\ge \frac{1}{6\sqrt{d}} S_{d-1}\left(\frac{3f(x_0)}{4}\frac{C_\delta}{\hat{C}_{x_0}}\right)^{\frac{d-1}{2}}\left(\frac{d\log m}{k}\right)^{\frac{d-1}{4}} f(x_0) \\
&= \frac{S_{d-1}}{6\sqrt{d}}\left(f(x_0)\right)^{\frac{d+1}{2}}\left(\frac{3C_\delta}{4\hat{C}_{x_0}}\right)^{\frac{d-1}{2}}\left(\frac{d\log m}{k}\right)^{\frac{d-1}{4}} \qquad (16)
\end{aligned}$$

where, inequality a follows from Lemma A.3 and the fact that $\nu(C) = \frac{\mathrm{vol}(C)}{S_{d-1}}$ for any $C \subset \mathcal{S}^{d-1}$. Inequality b follows from Definition 6.4.

If the r.h.s. of (16) is at least $\frac{k}{n} + \frac{C_\delta\sqrt{kd\log n}}{n}$, then using Lemma 7 of Chaudhuri and Dasgupta (2010), with probability at least $1-\delta$, $|B(x_0,\tau r/2)\cap X_n| \ge k$. Therefore, $r_n(x_0) \le \tau r/2$. This happens whenever $k \ge C_\delta^2 d\log n$ and $k \le \left(\frac{nS_{d-1}}{12\sqrt{d}}\right)^{\frac{4}{d+3}}(f(x_0))^{\frac{2d+2}{d+3}}\left(\left(\frac{3C_\delta}{4\hat{C}_{x_0}}\right)^2 d\log m\right)^{\frac{d-1}{d+3}}$. Further, using Theorem's assumption on k, it follows that, $r_n(x_0) \le \tau r/2 \le \tau r_{x_0}/2$. It is now clear that we can choose $\tilde{r} = r$. Once the choice of $\tilde{r}$ is finalized, using the fact that $\tau < 1$, Lemma B.1 and the Theorem's assumption on k, it is immediate that $\mathrm{diam}(C_j) \le \tau\tilde{r}/2$, for every $j = 1,\ldots,m$. □

B.2 Proof of Theorem 6.12

Since the proof is a bit involved, before presenting the detailed proof, we first provide a sketch of proof which is easier to follow.

Proof sketch: We define

$$r_n(x_0) = \inf\{r : B(x_0,r)\cap X_n \ne \emptyset\}$$

and choose $\tau \in (0,1)$ and $\tilde{r}$ satisfying the following:

$$\frac{2r_n(x_0)}{\tau} \le \tilde{r} \le \bar{r} \le r_{x_0}$$

Then, $B(x_0,\tilde{r}) \subset B(x_0,\bar{r}) \subset A_{x_0}$. It is clear that if Alg. 1 picks a maximizer of $\hat{f}_n$ from a set contained in A_{x_0} and intersecting $B(x_0, r_n(x)))$, then this maximizer must be within $\tilde{r}$ of x_0. If $\tilde{r}$ is appropriately chosen, then that would imply the result. Thus, our proof strategy is to show that Alg. 1 indeed picks a maximizer from such a set. To this end, let $\lambda \ge \inf_{x\in B(x_0,r_n(x_0))}\hat{f}_n(x)$ be the first level in the iteration of Alg. 1 containing a point from $B(x_0, r_n(x_0))$. Also, let S be the set that r-separates A_{x_0} from $\mathcal{S}^{d-1}_{\lambda_{x_0}} \setminus A_{x_0}$. Define

$$\mathcal{S}^{d-1}_{S\rightsquigarrow x_0} \triangleq \{x : \exists \text{ a path } \mathcal{P} \text{ from } x \text{ to } x_0, \mathcal{P}\cap S = \emptyset\}$$

to be the set of points reachable from x_0 without crossing S. Clearly, $\mathcal{S}^{d-1}_{S\rightsquigarrow x_0} \supseteq A_{x_0}$ and by Definition 6.4, $\mathcal{S}^{d-1}_{S\rightsquigarrow x_0} \cap \mathcal{S}^{d-1}_{\lambda_{x_0}} = A_{x_0}$. The crux of our proof strategy is to show that, (i) at the level $\lambda - \tilde{\epsilon}$, $G(\lambda-\tilde{\epsilon})$ contains

no point from $\mathcal{S}^{d-1}_{S\rightsquigarrow x_0} \setminus B(x_0, \bar{r})$, and (ii) there does not exist any edge between any $x \in B(x_0, \bar{r})$ and points in $\mathcal{S}^{d-1} \setminus \mathcal{S}^{d-1}_{S\rightsquigarrow x_0}$ in $G(\lambda - \tilde{\epsilon})$. Therefore, it follows that there exists a CC $\tilde{A}_j$ of $G(\lambda - \tilde{\epsilon})$ that is disconnected from $\mathcal{S}^{d-1} \setminus \mathcal{S}^{d-1}_{S\rightsquigarrow x_0}$, and thus any previous mode found by Alg. 1 such that $\tilde{A}_j \cap X_{n,\lambda}$ contains only points from A_{x_0} and intersects $B(x_0, r_n(x_0))$. Therefore, the procedure picks a point $\hat{x}_0 = \text{argmax}_{x \in \tilde{A}_j \cap X_{n,\lambda}} \hat{f}_n(x)$ satisfying $\|\hat{x}_0 - x_0\| \leq \tilde{r}$.

To achieve the first goal, it is enough to show that,

$$\sup_{x \in \mathcal{S}^{d-1}_{S\rightsquigarrow x_0} \setminus B(x_0, \bar{r})} \hat{f}_n(x) \leq \inf_{x \in B(x_0, r_n(x_0))} \hat{f}_n(x) - \gamma_n \tag{17}$$

since for $\tilde{\epsilon} \leq \gamma_n$, that would imply:

$$\sup_{x \in \mathcal{S}^{d-1}_{S\rightsquigarrow x_0} \setminus B(x_0, \bar{r})} \hat{f}_n(x) \leq \inf_{x \in B(x_0, r_n(x_0))} \hat{f}_n(x) - \gamma_n \leq \inf_{x \in B(x_0, r_n(x_0))} \hat{f}_n(x) - \tilde{\epsilon} \leq \lambda - \tilde{\epsilon}$$

To prove (17), we repeatedly invoke Lemma 5.4 with different ϵ values. In particular, we find the upper bound of the l.h.s. of (17) by invoking Lemma 5.4 with properly tuned ϵ, and show that this is less than the lower bound of the r.h.s. of (17) obtained by invoking again Lemma 5.4 with properly tuned (but different than before) ϵ.

To achieve the second goal, we ensure that for any point $x \in B(x_0, \bar{r})$, the ball $B(x, \tilde{r}/2)$ contains at least k points such that $r_k(x) \leq \tilde{r}/2$. Further, appropriately setting $\tilde{r}$, in comparison to separation parameter r and α given in Definition 6.9, we ensure that there can not be an edge between $x \in B(x_0, \bar{r})$ and points in $\mathcal{S}^{d-1} \setminus \mathcal{S}^{d-1}_{S\rightsquigarrow x_0}$ in $G(\lambda - \tilde{\epsilon})$. □

Now we present the actual proof.

Proof. Let $r_n(x_0) = \inf\{r : B(x_0, r) \cap X_n \neq \emptyset\}$ be the smallest radius around x_0 containing a sample from X_n. Similar to the proof of Theorem 6.6, let $\tau \in (0, 1)$ be a constant to be fixed later. Assume for now that $r_n(x_0) \leq \frac{\tau}{2} r_{x_0}$. In due course, we will show that this holds with high probability. Also, consider $\tilde{r}$, to be appropriately chosen later, satisfying the following:

$$\frac{2r_n(x_0)}{\tau} \leq \tilde{r} \leq r_{x_0}.$$

Here, in addition we use a new scale $\bar{r} = r_{x_0}/2$ satisfying the following:

$$\frac{2r_n(x_0)}{\tau} \leq \tilde{r} \leq \bar{r} \leq r_{x_0}.$$

Clearly, $B(x_0, \bar{r}) \subset B(x_0, r_{x_0} \subset A_{x_0}$. Note that, for appropriate choice of $\tilde{r}$, if the maximizer of $\hat{f}_n$ comes from a set contained in A_{x_0} and intersecting $B(x_0, r_n(x_0))$, then that maximizer must be within $\tilde{r}$ of x_0. Next, we show that Alg. 1 indeed picks a maximizer out of such a set.

To this end, let $\lambda \geq \inf_{x \in B(x_0, r_n(x_0))} \hat{f}_n(x)$ be the first level in the iteration of Alg. 1 containing a point from $B(x_0, r_n(x_0))$. Also, let S be the set that r-separates A_{x_0} from $\mathcal{S}^{d-1}_{\lambda_{x_0}} \setminus A_{x_0}$. Define

$$\mathcal{S}^{d-1}_{S\rightsquigarrow x_0} \triangleq \{x : \exists \text{ a path } \mathcal{P} \text{ from } x \text{ to } x_0, \mathcal{P} \cap S = \emptyset\}$$

to be the set of points reachable from x_0 without crossing S. Clearly, $\mathcal{S}^{d-1}_{S\rightsquigarrow x_0} \supseteq A_{x_0}$ and by Definition 6.4, $\mathcal{S}^{d-1}_{S\rightsquigarrow x_0} \cap \mathcal{S}^{d-1}_{\lambda_{x_0}} = A_{x_0}$.

Our goal is to show that, (i) at the level $\lambda - \tilde{\epsilon}$, $G(\lambda - \tilde{\epsilon})$ does not contain any point from $\mathcal{S}^{d-1}_{S\rightsquigarrow x_0} \setminus B(x_0, \bar{r})$, and (ii) there does not exist any edge between any $x \in B(x_0, \bar{r})$ and points in $\mathcal{S}^{d-1} \setminus \mathcal{S}^{d-1}_{S\rightsquigarrow x_0}$ in $G(\lambda - \tilde{\epsilon})$. Whenever the above goals are satisfied, it follows that there exists a CC $\tilde{A}_j$ of $G(\lambda - \tilde{\epsilon})$ that is disconnected from $\mathcal{S}^{d-1} \setminus \mathcal{S}^{d-1}_{S\rightsquigarrow x_0}$, and thus is disconnected from any previous mode found by Alg. 1 such that $\tilde{A}_j \cap X_{n,\lambda}$ contains only points from A_{x_0} and intersects $B(x_0, r_n(x_0))$. Therefore, Alg. 1 picks a point $\hat{x}_0 = \text{argmax}_{x \in \tilde{A}_j \cap X_{n,\lambda}} \hat{f}_n(x)$ satisfying $\|\hat{x}_0 - x_0\| \leq \tilde{r}$.

Rest of the proof ensures that the above two goals are satisfied. We start with the first goal. Let $\bar{F} = f(x_0) - \check{C}_{x_0}(\bar{r}/2)^2$. Note that $\forall x \in \mathcal{S}^{d-1}_{S\rightsquigarrow x_0} \setminus B(x_0, \bar{r})$ we have $\sup_{x' \in B(x,\tilde{r}/2)} f(x') \le \bar{F}$. To see this, first observe that,

$$B(x, \tilde{r}/2) \subset \left(\mathcal{S}^{d-1}_{S\rightsquigarrow x_0} \setminus B(x_0, \bar{r}/2)\right) \cup S_r$$

as long as $\tilde{r}/2 \le r$. We separately show that $\sup_{x \in \mathcal{S}^{d-1}_{S\rightsquigarrow x_0} \setminus B(x_0,\bar{r}/2)} f(x) \le \bar{F}$ and $\sup_{x \in S_r} f(x) \le \bar{F}$. The set $\mathcal{S}^{d-1}_{S\rightsquigarrow x_0} \setminus B(x_0, \bar{r}/2)$ can be written as $\mathcal{S}^{d-1}_{S\rightsquigarrow x_0} \setminus B(x_0, \bar{r}/2) = A_{x_0} \setminus B(x_0, \bar{r}/2) \cup \mathcal{S}^{d-1}_{S\rightsquigarrow x_0} \setminus A_{x_0}$. It follows immediately from Definition 6.4 that $\sup_{x \in A_{x_0} \setminus B(x_0,\bar{r}/2)} \le \bar{F}$. Also, from the definition of $\mathcal{S}^{d-1}_{S\rightsquigarrow x_0}$, it follows that $\sup_{x \in \mathcal{S}^{d-1}_{S\rightsquigarrow x_0} \setminus A_{x_0}} f(x) < \lambda_{x_0} \le \bar{F}$. Next, by Definition 6.10, $\sup_{x \in S_r} f(x) < \lambda_{x_0} \le \bar{F}$.

For any $x \in \mathcal{S}^{d-1}_{S\rightsquigarrow x_0} \setminus B(x_0, \bar{r})$, let $\epsilon = \bar{F} - f(x)$. Then it is easy to see that $\hat{r}(\epsilon, x) \ge \bar{r}/2 \ge \tilde{r}/2$. Therefore, for every $j = 1., \ldots, m$, if we can somehow restrict $\mathrm{diam}(C_j) \le \frac{\tilde{r}}{2}$, then it automatically ensures that $\mathrm{diam}(C_j) \le \hat{r}(\epsilon, x)$. Applying Lemma 5.4, the following holds with probability at least $1 - \delta$,

$$\begin{aligned}
\sup_{x \in \mathcal{S}^{d-1}_{S\rightsquigarrow x_0} \setminus B(x_0,\bar{r})} \hat{f}_n(x) &\le \left(1 + C_\delta \sqrt{\frac{d \log m}{k}}\right)(f(x) + \epsilon) + \gamma_n \\
&= \left(1 + C_\delta \sqrt{\frac{d \log m}{k}}\right)\bar{F} + \gamma_n \\
&= \left(1 + C_\delta \sqrt{\frac{d \log m}{k}}\right)(f(x_0) - \check{C}_{x_0}(\bar{r}/2)^2) + \gamma_n \\
&\le \left(1 + C_\delta \sqrt{\frac{d \log m}{k}}\right)(f(x_0) - \check{C}_{x_0}(\tilde{r}/2)^2) + \gamma_n
\end{aligned} \tag{18}$$

Now, for any $x \in S_r$, by r-separation, $f(x) \le \inf_{x' \in A_{x_0}} f(x') \le \bar{F}$. Next for any $x \in S_{r/2}$, let $\epsilon = \bar{F} - f(x)$. Then it is easy to see that $\hat{r}(\epsilon, x) \ge r/2 \ge \tilde{r}/2$, provided $\tilde{r} \le r$. Then using the same argument as before the following holds with probability at least $1 - \delta$,

$$\sup_{x \in S_{r/2}} \hat{f}_n(x) \le \left(1 + C_\delta \sqrt{\frac{d \log m}{k}}\right)(f(x_0) - \check{C}_{x_0}(\tilde{r}/2)^2) + \gamma_n \tag{19}$$

Now we bound $\inf_{x \in B(x_0, r_n(x_0))} \hat{f}_n(x)$ from below. Since $\tau\tilde{r} \le \tau r_{x_0} \le r_{x_0}$, using definition 6.4, observe that,

$$\inf_{x \in B(x_0, \tau\tilde{r})} f(x) \ge f(x_0) - \hat{C}_{x_0}(\tau\tilde{r})^2 \triangleq \check{F} \tag{20}$$

Now for any $x \in B(x_0, r_n(x_0))$, let $\epsilon = f(x) - \check{F}$. Then we have $\check{r}(\epsilon, x) \ge \tau\check{r} - r_n(x_0) \ge \tau\check{r} - \tau\tilde{r}/2 = \tau\tilde{r}/2$. Therefore, for every $j = 1., \ldots, m$, if we can somehow restrict $\mathrm{diam}(C_j) \le \frac{\tau\tilde{r}}{2}$, then it automatically ensures that $\mathrm{diam}(C_j) \le \check{r}(\epsilon, x)$. Again, applying Lemma 5.4, the following holds with probability at least $1 - \delta$,

$$\begin{aligned}
\inf_{x \in B(x_0, r_n(x_0))} \hat{f}_n(x) &\ge \left(1 - C_\delta \sqrt{\frac{d \log m}{k}}\right)(f(x) - \epsilon) + \gamma_n \\
&= \left(1 - C_\delta \sqrt{\frac{d \log m}{k}}\right)\check{F} - \gamma_n \\
&= \left(1 - C_\delta \sqrt{\frac{d \log m}{k}}\right)(f(x_0) - \hat{C}_{x_0}(\tau\tilde{r})^2) - \gamma_n
\end{aligned} \tag{21}$$

Noting that $\tilde{\epsilon} \le \gamma_n$, we can pick τ and $\tilde{r}$ such that

$$\sup_{x \in \mathcal{S}^{d-1}_{S\rightsquigarrow x_0} \setminus B(x_0,\bar{r})} \hat{f}_n(x) \le \inf_{x \in B(x_0, r_n(x_0))} \hat{f}_n(x) - \gamma_n \le \inf_{x \in B(x_0, r_n(x_0))} \hat{f}_n(x) - \tilde{\epsilon} \le \lambda - \tilde{\epsilon}$$

and

$$\sup_{x\in S_{r/2}} \hat{f}_n(x) \leq \inf_{x\in B(x_0,r_n(x_0))} \hat{f}_n(x) - \gamma_n \leq \inf_{x\in B(x_0,r_n(x_0))} \hat{f}_n(x) - \tilde{\epsilon} \leq \lambda - \tilde{\epsilon}$$

ensuring that the graph $G(\lambda-\tilde{\epsilon})$ contains no points from $\mathcal{S}^{d-1}_{S\rightsquigarrow x_0} \setminus B(x_0,\bar{r})$ and no pints from $S_{r/2}$. It suffices to pick $\tau^2 = \frac{\check{C}_{x_0}}{8\hat{C}_{x_0}}$ and $\tilde{r}^2 \geq \frac{16}{\check{C}_{x_0}}\left(f(x_0)C_\delta\sqrt{\frac{d\log m}{k}} + 3\gamma_n/2\right)$. Using Lemma B.2 and the Theorem's assumption on k, it is easy to check that if $n \geq \frac{9m\log(m/\delta)\max\{\|f\|_\infty,1\}}{S_{d-1}(f(x_0)C_\delta)^2 d\log m}$, then γ_n is at most $f(x_0)C_\delta\sqrt{\frac{d\log m}{k}}$ and therefore $\tilde{r}^2 \geq 40f(x_0)\frac{C_\delta}{\check{C}_{x_0}}\sqrt{\frac{d\log m}{k}}$. Therefore, we are free to choose $\tilde{r} \geq \max\left\{\sqrt{40f(x_0)\frac{C_\delta}{\check{C}_{x_0}}\sqrt{\frac{d\log m}{k}}}, \frac{2r_n(x_0)}{\tau}\right\}$.

Let's set $\tilde{r} = \sqrt{40f(x_0)\frac{C_\delta}{\check{C}_{x_0}}\sqrt{\frac{d\log m}{k}}}$.

Next, for any $x \in B(x_0,\bar{r})$, the ball $B(x,\tilde{r}/2)$ is fully contained in A_{x_0} since $\bar{r}+\tilde{r}/2 \leq 3\bar{r}/2 = 3r_{x_0}/4 < r_{x_0}$. We seek to bound $f(B(x,\tilde{r}/2))$ from below as follows

$$\begin{aligned}
f(B(x,\tilde{r}/2)) &\geq f(B(x,\tau\tilde{r}/2)) \\
&\geq \mathrm{vol}(B(x,\tau\tilde{r}/2)) \inf_{x\in B(x,\tau\tilde{r}/2)} f(x) \\
&\overset{a}{\geq} \frac{1}{3\sqrt{d}}(3/4)^{\frac{d-1}{2}} S_{d-1}(\tau\tilde{r}/2)^{d-1} \inf_{x\in B(x,\tau\tilde{r}/2)} f(x) \\
&\geq \frac{1}{3\sqrt{d}}(3/4)^{\frac{d-1}{2}} S_{d-1}(\tau\tilde{r}/2)^{d-1}\lambda_{x_0} \\
&\overset{b}{\geq} \frac{S_{d-1}}{3\sqrt{d}}\left(\frac{15\lambda_{x_0}C_\delta}{16\hat{C}_{x_0}}\right)^{\frac{d-1}{2}}\left(\frac{d\log m}{k}\right)^{\frac{d-1}{4}}
\end{aligned} \tag{22}$$

where, inequality a follows from Lemma A.3 and the fact that $\nu(C) = \frac{\mathrm{vol}(C)}{S_{d-1}}$ for any $C \subset \mathcal{S}^{d-1}$. Inequality b follows from the fact that $f(x_0) \geq \lambda_{x_0}$. If the r.h.s. of (22) is at least $\frac{k}{n} + \frac{C_\delta\sqrt{kd\log n}}{n}$, then using Lemma 7 of Chaudhuri and Dasgupta (2010), with probability at least $1-\delta$, $|B(x_0,\tilde{r}/2)\cap X_n| \geq k$. Therefore, $r_k(x) \leq \tilde{r}/2$. This happens whenever $k \geq C_\delta^2 d\log n$ and $k \leq \left(\frac{S_{d-1}}{6\sqrt{d}}\left(\frac{15f(x_0)C_\delta}{64\check{C}_{x_0}}\right)^2 d\log m\right)^{\frac{d-1}{d+3}} n^{\frac{4}{d+3}}$. If in addition $\tilde{r} \leq r/\alpha$, then it is easy to see that $\alpha r_k(x) \leq \alpha\tilde{r}/2 \leq r/2$. Therefore, $B(x,\alpha r_k(x)) \subset \mathcal{S}^{d-1}_{S\rightsquigarrow x_0}$ and there can not be any edge between $x \in B(x_0,\bar{r})$ and points in $\mathcal{S}^{d-1}\setminus\mathcal{S}^{d-1}_{S\rightsquigarrow x_0}$ in $G(\lambda-\tilde{\epsilon})$.

All that remains to show is, (i) $\tilde{r} \leq \bar{r} = r_{x_0}/2$, (ii) $r_n(x_0) \leq \tau\tilde{r}/2 \leq \tau r_{x_0}/2$, (iii) $\tilde{r} \leq r/\alpha$, and (iv) $\mathrm{diam}(C_j) \leq \min\{\tilde{r}/2, \tau\tilde{r}/2\}$. If $k \geq \left(\frac{40f(x_0)C_\delta}{\check{C}_{x_0}(\min\{r_{x_0}/2, r/\alpha\})^2}\right)^2 d\log m$, using the value of $\tilde{r}$, the first and the third conditions are satisfied. Note that (22) hold for any $x \in B(x_0,\bar{r})$, in particular, for x_0. Therefore, $r_n(x_0) \leq \tau\tilde{r}/2$ and the second condition is satisfied. Finally, using the fact that $\tau < 1$, Lemma B.1 it is immediate that $\mathrm{diam}(C_j) \leq \tau\tilde{r}/2$ for every $j = 1,\ldots,m$ whenever $k \leq \frac{1}{6}\left(\left(\frac{3}{16}f(x_0)\frac{C_\delta}{\hat{C}_{x_0}}\right)^2\log m\right)^{\frac{d-1}{d+3}} m^{\frac{4}{d+3}}$.

□

B.3 Auxiliary proofs

Lemma B.1. *Let $x_0 \in \mathcal{M}$. Set $\tau^2 = \frac{\check{C}_{x_0}}{8\hat{C}_{x_0}}$, $\tilde{r}_1^2 = 32f(x_0)\frac{C_\delta}{\check{C}_{x_0}}\sqrt{\frac{d\log m}{k}}$, and $\tilde{r}_2^2 = 40f(x_0)\frac{C_\delta}{\check{C}_{x_0}}\sqrt{\frac{d\log m}{k}}$. Assume that the conclusion of the Lemma 5.2 holds. Then, $\mathrm{diam}(C_j) \leq \tau\tilde{r}_1/2$ for every $j = 1,\ldots,m$, provided the following holds.*

$$k \leq \frac{1}{6}\left(\left(\frac{3}{16}f(x_0)\frac{C_\delta}{\hat{C}_{x_0}}\right)^2\log m\right)^{\frac{d-1}{d+3}} m^{\frac{4}{d+3}}.$$

Similarly, $\mathrm{diam}(C_j) \le \tau\tilde{r}_2/2$ *for every* $j = 1, \ldots, m$*, provided the following holds.*

$$k \le \frac{1}{6}\left(\left(\frac{5}{64}\lambda_{x_0}\frac{C_\delta}{\hat{C}_{x_0}}\right)^2 \log m\right)^{\frac{d-1}{d+3}} m^{\frac{4}{d+3}}.$$

Proof. Since conclusion of Lemma 5.2 holds, we can apply Lemma 5.6 to get

$$\mathrm{diam}(C_j) \le \frac{4}{\sqrt{3}}\left(\frac{6\sqrt{d}k}{m}\right)^{\frac{1}{(d-1)}}$$

Note that, $\frac{\tau\tilde{r}_1}{2} = \sqrt{f(x_0)\frac{C_\delta}{\hat{C}_{x_0}}}\left(\frac{d\log m}{k}\right)^{1/4}$. Then $\mathrm{diam}(C_j) \le \tau\tilde{r}_1/2$ for every $j = 1, \ldots, m$ provided

$$\begin{aligned}
&\frac{4}{\sqrt{3}}\left(\frac{6\sqrt{d}k}{m}\right)^{\frac{1}{(d-1)}} \le \sqrt{f(x_0)\frac{C_\delta}{\hat{C}_{x_0}}}\left(\frac{d\log m}{k}\right)^{1/4} \\
\implies &\frac{16}{3}\left(\frac{6\sqrt{d}k}{m}\right)^{\frac{2}{d-1}} \le f(x_0)\frac{C_\delta}{\hat{C}_{x_0}}\sqrt{\frac{d\log m}{k}} \\
\implies &\frac{256}{9}\left(\frac{6\sqrt{d}k}{m}\right)^{\frac{4}{d-1}} \le \left(f(x_0)\frac{C_\delta}{\hat{C}_{x_0}}\right)^2 \frac{d\log m}{k} \\
\implies &k^{\frac{d+3}{d-1}} \le \left(\left(\frac{3}{16}f(x_0)\frac{C_\delta}{\hat{C}_{x_0}}\right)^2 \log m\right) d^{\frac{d-3}{d-1}} m^{\frac{4}{d-1}}/6^{\frac{4}{d-1}} \\
\implies &k \le \left(\left(\frac{3}{16}f(x_0)\frac{C_\delta}{\hat{C}_{x_0}}\right)^2 \log m\right)^{\frac{d-1}{d+3}} d^{\frac{d-3}{d+3}} m^{\frac{4}{d+3}}/6^{\frac{4}{d+3}}
\end{aligned}$$

Noting that, $d^{\frac{d-3}{d+3}} = d^{1-\frac{6}{d+3}}$ is an increasing function of d, we have $d^{\frac{d-3}{d+3}} \ge \frac{1}{2^{1/5}}$ for $d \ge 2$. Similarly, noting that $6^{\frac{4}{d+3}}$ is a decreasing function of d, we have $6^{\frac{4}{d+3}} \le 6^{4/5}$ for $d \ge 2$. Thus, $d^{\frac{d-3}{d+3}}/6^{\frac{4}{d+3}} \ge \frac{1}{2^{1/5}} \cdot \frac{1}{6^{\frac{4}{5}}} = \frac{1}{2^{1/5}} \cdot \frac{1}{2^{4/5}3^{4/5}} = \frac{1}{2} \cdot \frac{1}{3^{4/5}} \ge \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$. Plugging in this value, the choice of k as given in the Lemma statement suffices.

Similarly, note that, $\frac{\tau\tilde{r}_2}{2} = \sqrt{\frac{5}{4}f(x_0)\frac{C_\delta}{\hat{C}_{x_0}}}\left(\frac{d\log m}{k}\right)^{1/4} \ge \sqrt{\frac{5\lambda_{x_0}C_\delta}{4\hat{C}_{x_0}}}\left(\frac{d\log m}{k}\right)^{1/4}$. Using similar steps as above, the result follows. □

Lemma B.2. *Let* $x_0 \in \mathcal{M}$*,* $k \ge \frac{(f(x_0)C_\delta)^2}{9} d\log m$*, and* $\gamma_n = \alpha_n/(S_{d-1}(k/m))$*, where* $\alpha_n = 2\sqrt{\frac{kS_{d-1}\|f\|_\infty \log(m/\delta)}{mn}} + \frac{2\log(m/\delta)}{3n}$*. Then,* $\gamma_n \le f(x_0)C_\delta\sqrt{\frac{d\log m}{k}}$ *provided the following holds,*

$$n \ge \frac{9m\log(m/\delta)\max\{\|f\|_\infty, 1\}}{S_{d-1}(f(x_0)C_\delta)^2 d\log m}.$$

Proof. Let $\beta = \frac{m\log(m/\delta)}{S_{d-1}kn}$. Then,

$$\begin{aligned}
\gamma_n = \frac{\alpha_n}{S_{d-1}k/m} &= \frac{\left(2\sqrt{\frac{kS_{d-1}\|f\|_\infty \log(m/\delta)}{mn}} + \frac{2\log(m/\delta)}{3n}\right)}{S_{d-1}k/m} \\
&= 2\sqrt{\beta\|f\|_\infty} + \frac{2}{3}\beta \\
&\leq 2\sqrt{\beta\max\{\|f\|_\infty, 1\}} + \frac{2}{3}\beta\max\{\|f\|_\infty, 1\} \\
&\overset{a}{\leq} 3\sqrt{\beta\max\{\|f\|_\infty, 1\}} \\
&\overset{b}{\leq} f(x_0)C_\delta\sqrt{\frac{d\log m}{k}}
\end{aligned}$$

Inequality a holds, if $\beta\max\{\|f\|_\infty, 1\} \leq 1$, or in other words,

$$n \geq \frac{m\log(m/\delta)\max\{\|f\|_\infty, 1\}}{kS_{d-1}} \tag{23}$$

Inequality b holds, if,

$$n \geq \frac{9m\log(m/\delta)\max\{\|f\|_\infty, 1\}}{S_{d-1}(f(x_0)C_\delta)^2 d\log m} \tag{24}$$

Therefore, both the inequalities hold, if,

$$n \geq \frac{m\log(m/\delta)\max\{\|f\|_\infty, 1\}}{S_{d-1}} \times \max\left\{\frac{1}{k}, \frac{9}{(f(x_0)C_\delta)^2 d\log m}\right\}.$$

Since $k \geq \frac{(f(x_0)C_\delta)^2}{9} d\log m$, choosing n as given in the Lemma statement suffices. □

C Empirical evaluations details

Since our density and mode estimation results hold for data lying on the unit sphere, for our empirical evaluations, we choose distributions that are supported on the unit sphere. A well known distribution defined on the unit sphere is von Mises-Fisher distribution which is a directional distribution on the surface of the unit sphere $\mathcal{S}^{d-1}$. For any $x \in \mathcal{S}^{d-1}$, the probability density of a von Mises-Fisher distribution is given as in (6) where, $\mu \in \mathcal{S}^{d-1}$ is the mean direction, $\kappa > 0$ is the concentration parameter, d is the dimension, and I is the modified Bessel function of the first kind. The distribution becomes narrower around μ with increasing κ. In particular, the reciprocal value $1/\kappa$ resembles the variance parameter of the normal distribution. We chose a mixture of two von Mises-Fisher distributions given in (7) where, for $i = 1, 2$, $f_i(x)$ is specified, using (6) by mean direction μ_i, concentration parameter κ_i and $w \in (0, 1)$ is the mixture coefficient. In Fig.**??**, we visually depict a mixture of von Mises-Fisher distribution in $\mathcal{S}^2$, where $\mu_1 = \left(-\sqrt{0.9}, -\sqrt{0.1}, 0\right) \in \mathcal{S}^2$, $\kappa_1 = 10$, $\mu_2 = \left(-\sqrt{0.01}, -\sqrt{0.99}, 0\right) \in \mathcal{S}^2$, $\kappa_2 = 5$, and $w = 0.3$.

C.1 Data generation process and computing environment

For our empirical evaluations, we generate samples from a mixture of von Moses-Fisher distribution as given in (7). Specifically, we use `scipy`* to generate samples from an individual von Mises-Fisher distribution. To generate samples from a mixture of two von Mises-Fisher distributions we first need to specify the component means μ_1 and μ_2. We randomly pick μ_1 from $\mathcal{S}^{d-1}$ and then randomly pick $\mu_2 \in \mathcal{S}^{d-1}$ that makes an angle $\pi/4$ with μ_1 (basically, once we have chosen μ_1, we keep on sampling points uniformly at random from $\mathcal{S}^{d-1}$ until we find a point that makes an angle $\pi/4$ with μ_1.) We use two sets of values for κ_1 and κ_2. In one set we use $\kappa_1 = 10$ and $\kappa 2 = 5$. In another set we use $\kappa_1 = 80$ and $\kappa_2 = 100$. For larger κ values individual von Mises-Fisher distributions becomes narrower around their respective means. We set mixture parameter $w = 0.3$.

*`https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.vonmises_fisher.html`

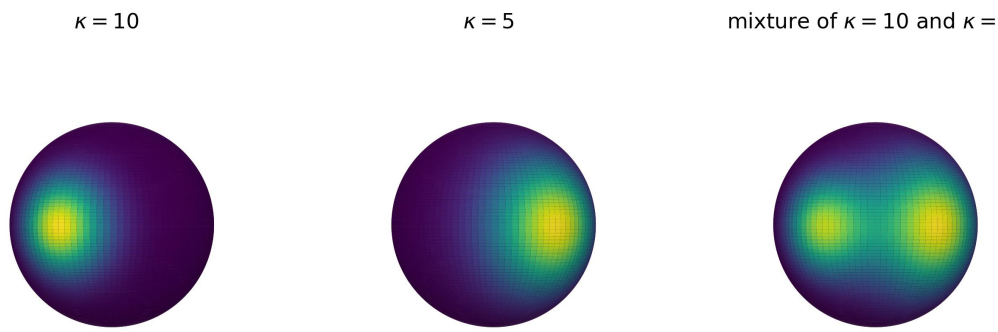


Figure 3: Individual components and a mixture of von Mises-Fisher distributions on $\mathcal{S}^2$. The left panel shows probability density of von Mises-Fisher distribution with parameters $\mu_1 = \left(-\sqrt{0.9}, -\sqrt{0.1}, 0\right) \in \mathcal{S}^2$ and $\kappa_1 = 10$. The middle panel shows the probability density function of von Mises-Fisher distribution with parameters $\mu_2 = \left(-\sqrt{0.01}, -\sqrt{0.99}, 0\right) \in \mathcal{S}^2$ and $\kappa_2 = 5$. The right panel shows the probability density function of a mixture of these two von Mises-Fisher distributions where the mixing coefficient $w = 0.3$. In all these picture bright yellow color represents high density regions and the shifting shades from yellow to green represents gradual low density regions.

Once we specify $d, \mu_1, \mu_2 \in \mathcal{S}^{d-1}, \kappa_1, \kappa_2 > 0$ and $w \in (0, 1)$ we can easily sample data points from the resulting mixture of von Mises-fisher distribution using `scipy`. We sample three sets of iid samples from mixture of von Mises-Fisher distribution. The first set contains 10000 samples and is used to train the density estimators (training set). The second set contains 2000 samples and is used to choose hyper-parameters of various estimators (validation set) and the last set contains 10000 samples and is used to test the performance of the density estimators (test set).

We run our experiments on a laptop with Intel Xeon W-10855M Processor, 64GB memory and NVDIAQuadro RTX 5000 Mobile GPU (with 16GB memory).

C.2 Hyper-parameter selection

For each of the three estimators KNNDE, KDE and EaSDE, we select hyper-parameters using the validation set. If there are n samples in the training set, then for KNNDE, we choose the best value of k, the number of nearest neighbors, from the set $\{1, 10, 50, \log(n), \sqrt{n}, n/8, n/4, n/2, 3n/4\}$ for which empirical total variation distance of KNNDE is smallest on the validation set. For any data point, we find its k nearest neighbors from the training set using `sklear.neighbors` from `scikit-learn`.*

We implemented KDE using `scikit-learn`'s `KernelDensity` package. We used Gaussian kernel and the bandwidth h of the Gaissian kernel was chosen using the validation set. We selected the best h value from the set `np.logspace(-2, 0, 20)` containing 20 values in the log space from the interval $[10^{-2}, 1]$.

For EasDE, we varied m via varying the expansion factor which is defined as m/d. For each fixed m, we selected the value of k using the validation set. The various values of k that were considered were from the set $[d\log(m) - 8, d\log(m) + 8]$.

Eroor bars of EaSDE we computed from 10 independent runs of EaSDE, each time using a different random projection matrix Θ.

*`https://scikit-learn.org/stable/`