Proposal Outline: Approximately measuring smoothness of functions with graphs, and consequences for learning.

Alden Green

September 27, 2019

1 Introduction.

(Provide context. How have people traditionally measured smoothness using graphs. How is what you're doing different?)

In this work, we examine alternative assumptions on the smoothness of functions we consider, and demonstrate that we can still 'approximately' measure smoothness of functions, with the approximation being sufficiently useful for subsequent learning tasks.

2 Setup.

2.1 Neighborhood graphs.

We observe $X = \{x_1, \ldots, x_n\}$, independently sampled from P, where we assume P is a distribution with density f in $\mathbb{R}^d$. Given a connectivity radius $r > 0$, we form the undirected neighborhood graph $G = G_{n,r}$ over the data X, by connecting points x_i and x_j in G whenever $\|x_i - x_j\| \leq r$. We define A to be the adjacency matrix of G, the $n \times n$ matrix with entries $A_{ij} = \mathbf{1}(\|x_i - x_j\| \leq r)$; D to be the degree matrix, and $W = \frac{1}{2}(I + D^{-1}A)$ to be the lazy random walk matrix over G.

3 Recovery of density clusters with PPR.

We first review the relevant concepts of density clusters and PPR in brief.

3.1 Density clusters and cluster recovery.

Given a distribution P with density function f, for a given threshold $\lambda > 0$, the upper level set of f at λ is $\mathbb{U}_f(\lambda) = \bigl\{x \in \mathbb{R}^d : f(x) > \lambda\bigr\}$. Suppose $\mathbb{C}_f(\lambda)$ has M connected components, $\mathcal{C}_1, \ldots, \mathcal{C}_M$. Define the *density cluster at* λ to be the collection of these connected components, $\mathbb{C}_f(\lambda) = \{\mathcal{C}_1, \ldots, \mathcal{C}_\lambda\}$.

The *density clustering* task involves estimating $\mathbb{C}_f(\lambda)$; roughly, finding a collection $\{\hat{C}_1, \ldots, \hat{C}_M\}$ such that the discrepancy between $\{\widehat{C}_1, \ldots, \widehat{C}_M\}$ and $\mathbb{C}_f(\lambda)$ is small. We consider a local analogue to the density clustering task, which we term *local density clustering*. In local density clustering, the goal is to provide an estimator $\widehat{C}$ such that the discrepancy between $\widehat{C}$ and $\mathcal{C}_m$ is small, for some $m = 1, \ldots, M$.

To measure our performance, we will use two notions of discrepancy. For a given density cluster $\mathcal{C} \in \mathbb{C}_f(\lambda)$, we call $\mathcal{C}[X] = \mathcal{C} \cap X$ the *empirical density cluster*. The size of the symmetric set difference between estimated and empirical clusters is a commonly used metric to quantify cluster estimation error (**???**).

Definition 3.1 (Symmetric set difference)**.** For an estimator $\widehat{C} \subseteq X$ and set $\mathcal{S} \subseteq \mathbb{R}^d$, we define

$$\Delta(\widehat{C}, \mathcal{S}) := \left|\widehat{C} \setminus \mathcal{S}[X] \cup \mathcal{S}[X] \setminus \widehat{C}\right|, \tag{1}$$

the cardinality of the symmetric set difference between $\widehat{C}$ and $\mathcal{S} \cap X = \mathcal{S}[X]$.

However, the symmetric set difference does not measure whether $\widehat{C}$ can distinguish any two distinct clusters $\mathcal{C}, \mathcal{C}' \in \mathbb{C}_f(\lambda)$. We therefore also study a second notion of cluster estimation, first introduced by **?**, and defined asymptotically.

Definition 3.2 (Consistent density cluster estimation)**.** For an estimator $\widehat{C} \subseteq X$ and cluster $\mathcal{C} \in \mathbb{C}_f(\lambda)$, we say $\widehat{C}$ is a consistent estimator of $\mathcal{C}$ if for all $\mathcal{C}' \in \mathbb{C}_f(\lambda)$ with $\mathcal{C} \neq \mathcal{C}'$, the following holds as $n \to \infty$:

$$\mathcal{C}[X] \subseteq \widehat{C} \quad \text{and} \quad \widehat{C} \cap \mathcal{C}'[X] = \emptyset, \tag{2}$$

with probability tending to 1.

Consistent cluster recovery roughly ensures that, for a given λ, the estimated cluster $\widehat{C}$ contains all points in a true density cluster $\mathcal{C}$, and simultaneously does not contain any points in any other density cluster $\mathcal{C}'$.

We now turn to defining the PPR vector, and a natural local clustering algorithm based on thresholding the PPR vector.

3.2 PPR

Given a seed node $v \in G$, and a teleportation parameter $\alpha \in [0, 1]$, the PPR vector is defined to be the solution to the following linear equation:

$$p_v = \alpha e_v + (1 - \alpha) pW. \tag{3}$$

3.3 Work already completed.

- Finite sample and asymptotic, respectively, bounds on the normalized cut and mixing time of a set $\mathcal{C}$. These measure internal and external connectivity of $\mathcal{C}$, and in turn internal and external smoothness of $\mathbf{1}_{\mathcal{C}}$ with respect to the density f.
- A set of sufficient conditions on a density cluster $\mathcal{C}$, and a difficulty function $\kappa(\mathcal{C})$, such that PPR has small symmetric set difference with $\mathcal{C}_\sigma$.

3.4 Proposed work.

- Lower bound. We will exhibit a density function f with density cluster $\mathcal{C}$ such that the symmetric set difference $\left|\Delta(\mathcal{C}[X] \cap \widehat{C})\right|$ is provably large, when the hyperparameters of PPR are 'adversarially tuned'. We will extend these results to show that, under a range of possible hyperparameters, the symmetric set difference $\left|\Delta(\mathcal{C}[X] \cap \widehat{C})\right|$ remains large.
- Difference between global and local spectral clustering. To better demonstrate the need for separate theory regarding local spectral clustering in a statistical context, we will exhibit a density function f with density cluster $\mathcal{C}$ such that the symmetric set difference $\left|\Delta(\mathcal{C}[X] \cap \widehat{C}_{\mathrm{PPR}})\right|$ is small, but the symmetric set difference $\left|\Delta(\mathcal{C}[X] \cap \widehat{C}_{\mathrm{spec}})\right|$ is large.
- We will make $\kappa(\mathcal{C})$ an *intrinsic* quantity – meaning a quantity which is a function only of the distribution P – by eliminating factors of r.
- We will prove a finite sample bound on the mixing time $\tau_\infty(\mathcal{C})$. We will prove bounds on the mixing time $\tau_\infty(\mathcal{C})$ in the case where $\mathcal{C}_\sigma$ is not assumed to be a biLipschitz deformation of a convex set.

3.5 Finite sample bounds on mixing time.

We desire two improvements on our current bound on mixing time:

1. Replacing our asymptotic results with finite sample results, and
2. Adding a bound which does not depend on $\mathcal{C}_\sigma$ be a biLipschitz deformation of a convex set.

3.5.1 Finite sample results on mixing time.

3.5.2 BiLipschitz deformation of a convex set.

We assume now only that $\mathcal{C}_\sigma = \mathcal{C} + B(0, \sigma)$ for $\mathcal{C}$ a connected set, and not that $\mathcal{C}_\sigma = f(\mathcal{K})$ for f a biLipschitz function and $\mathcal{K}$ a convex set. We prove the following isoperimetric inequality, which will be useful in lower bounding the conductance of $\mathcal{C}_\sigma$.

Lemma 1. *Let $\mathcal{A} = \bigcup_{i=1}^{M} B(x_i, \sigma)$. Further, assume*

$$\nu(I(x, x_{i+1})) =: V > 0, \quad \text{for all } i = 1, \ldots, M-1,$$

where $I(x, y) = B(x, \sigma) \cap B(y, \sigma)$. Then, letting $S_1 \cup S_2 \cup S_3 = \mathcal{C}_\sigma$ be a partition of $\mathcal{A}$, the following inequality holds:

$$\nu(S_3) \geq \frac{\operatorname{dist}(S_1, S_2)}{6M\sigma} \min\{\nu(S_1), \nu(S_2)\} \frac{V}{\nu_d \sigma^d} \tag{4}$$

4 Graph Testing with Minimal Assumptions.

4.1 Work already completed.

- We have studied the goodness-of-fit testing problem in a regression setting. We have exhibited a test statistic based on the spectrum of the neighborhood graph G, and shown that when r is appropriately scaled to 0, the resulting test is minimax optimal when the alternative hypothesis is constrained to lie within the unit ball of the Sobolev space $W^{2,1}$, when $d < 4$. The test does not require that the alternative hypothesis satisfy any classical notion of smoothness; i.e we do not rely on the Sobolev embedding theorem, which holds only over Sobolev spaces $W^{2,s}$ such that $s < d/2$.

4.2 Proposed work

- We will modify the proposed test statistic, and extend the previous results to hold for $s > 1, d > 4$, with a particular focus on making minimal restrictions on s relative to d.
- We will also consider a proposed test statistic (more similar to the one Ryan originally proposed) and study its properties in the goodness-of-fit and two-sample testing regimes.
- We will consider testing against bounded variation alternatives.

5 Graph Regression with Minimal Assumptions.