HARMONIC ANALYSIS ON GRAPHS AND DATASETS

NICHOLAS MARSHALL

1. Introduction.

Introduction. Harmonic Analysis is the study of what happens when one decomposes a function $f : [0, 2\pi] \to \mathbb{R}$ into sums of sines and cosines. It has been a backbone of mathematics ever since its invention with many applications in other fields. Recently, motivated by the necessity of having to analyze large complicated datasets people have started using Harmonic Analysis to study graphs. The main idea is simple: sine and cosine arise naturally as eigenfunctions of the Laplace operator (or Laplacian) $-\Delta = -d^2/dx^2$ on the interval $[0, 2\pi]$. The Laplacian can be considered as measuring the deviation of a function from its local average, for example, think of the finite difference approximation for the Laplacian in $\mathbb{R}^2$:

$$-\Delta f(x, y) \approx \frac{4}{\varepsilon^2}\left(f(x, y) - \frac{f(x+\varepsilon, y) + f(x-\varepsilon, y) + f(x, y+\varepsilon) + f(x, y-\varepsilon)}{4}\right).$$

The Laplacian is positive for convex functions (the function on a neighborhood is slightly larger on average) and negative on concave functions (the function is on average slightly smaller on a neighborhood), and 0 for linear functions. These considerations motivate the following definition for the Laplacian on graphs.

Definitions. Let $G = (V, E)$ be a graph on n vertices. A function $f : V \to \mathbb{R}$ is simply a map that assigns a real number to every vertex. We note that this space is now a simple finite-dimensional vector space isomorphic to $\mathbb{R}^n$. The graph Laplacian $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}^n$ is defined

$$(\mathcal{L}f)(v) = \frac{1}{\#\text{ of neighbors of } v} \sum_{(u,v)\in E} (f(v) - f(u)).$$

This operator is linear and thus can be identified with an $n \times n$ matrix. Such a matrix has only real eigenvalues and we call those the spectrum of the Graph Laplacian. Moreover, the associated eigenvectors, which can be interpreted as maps from the vertices to the Euclidean space, are the natural discrete analogues of sine and cosine adapted to the geometric structure given by the graph.

2. Problems

This is a very active field of research and there are many conceivable problems; we believe that the following few problems are natural starting points for an inquiry.

2.1. **Building a Graph.** In practice, we are rarely given an explicit graph: a more realistic scenario is that we are given a set $\{x_j\}_{j=1}^n$ of n points in $\mathbb{R}^N$ and have to construct the graph ourselves. How would one construct a graph that best

approximates this set of points? There are two canonical ways of doing it: one could connect all points whose distance is less than a certain threshold

$$(x_j, x_k) \in E \iff \|x_j - x_k\|_{\ell^2(\mathbb{R}^2)} < \delta,$$

where $\delta > 0$ is a fixed parameter. Such a construction is illustrated below.

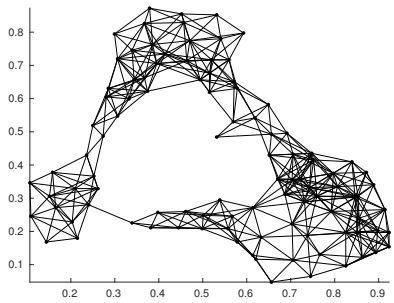

FIGURE 1. A graph built by connecting points of distance less than 1/8 apart.

Another method is to simply connect each point to its k-nearest neighbors (with $k \in \mathbb{N}$ being another parameter). From a computational point of view, it is advantageous to define the sparsest graph, that is to say, the graph with the least edges, which nonetheless captures the "geometry" of the set. The specific notion of geometry depends on the goal of the analysis. Recent work from the Yale Applied Math Program [3] describes a randomized method for building a graph with few edges, while ensuring the graph remains "sufficiently connected". This construction is much sparser than the two standard constructions outlined above but not much beyond its basic theory is known and there are many other interesting randomized graph models that could be considered. Possible problems include:

(1) A detailed investigation of various types of random constructions and the advantages and disadvantages of each.
(2) How many edges does one need? The best case would be to only use $|E| = \mathcal{O}(|V|)$ edges; can this be achieved?
(3) Which constructions are most stable under moving the points a little bit in a random direction ('stability under noise')?
(4) Can Random Graphs with good expansion properties (Ramanujan Graphs, Expanders) be explicitly used as a construction tool? Can the mathematics behind them be used to study this setting?
(5) [3] is asking whether Ulam-type random graphs could be more stable than Erdős-Renyi random graphs; can this be made rigorous?

2.2. **Boundary conditions.** In the continuous setting, equations involving the Laplace operator have to be considered with boundary conditions to be rigorous. Specifically, given a domain Ω with boundary $\partial\Omega$ there are two common boundary conditions which are considered:

$$\begin{array}{cc} \text{Dirichlet} & \text{Neumann} \\ \left\{ \begin{array}{ll} \Delta u = \lambda u & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega, \end{array} \right. \text{ and } & \left\{ \begin{array}{ll} \Delta u = \lambda u & \text{in } \Omega \\ \frac{\partial}{\partial n} u = 0 & \text{on } \partial\Omega, \end{array} \right. \end{array}$$

where n denotes an exterior normal to the boundary.

Given a set $\{x_j\}_{j=1}^n$ of n points sampled from a domain $\Omega \subset \mathbb{R}^N$ with a boundary $\partial\Omega$ it is interesting to consider methods of constructing an $n \times n$ matrix operator which approximates the continuous Laplace operator with specific boundary conditions. For example, a recent work in the Applied Math Program [1] considers a method of construction an approximation of the Neumann Laplacian with a bi-stochastic matrix, that is to say, a matrix with positive entries whose rows and columns all sum up to 1. Constructing a robust approximation of the Dirichlet Laplacian operator has not been treated in the literature. Specifically, it becomes necessary to generalize the notion of a point being on the boundary of data in a robust way: is it possible to define a 'boundary' of a Graph in a way that makes sense on meshes?

2.3. **Fourier Features.** Most applications of Harmonic Analysis to analyzing a given set of points $\{x_j\}_{j=1}^n \subset \mathbb{R}^N$ involve defining a graph and constructing some type of discrete analog of the Laplace operator. However, there is a rather direct method related to the classical Fourier Transform which is studied in recent work by Google Research [4] and looks very promising. Specifically, given a set $\{x_j\}_{j=1}^n \subset \mathbb{R}^N$, we choose a set of m points $\{\xi_k\}_{k=1}^m \subset \mathbb{R}^N$ from some probability density $p(\xi)$ and considers the $n \times m$ matrix F defined by

$$F(j,k) = \{e^{2\pi i x_j \cdot \xi_k}\},$$

where $\cdot$ denotes the Euclidean inner product. This matrix turns out to be very interesting and useful. First, even if the set $\{x_j\}_{j=1}^n$ is deterministic, the set $\{\xi_j\}_{k=1}^m$ is random, which adds some randomness to the matrix F. Second, multiplying F by an $m \times 1$ vector v

$$Fv(j) = \sum_{k=1}^{m} e^{2\pi i x_j \cdot \xi_k} v(k) \approx \int_{\mathbb{R}^N} v(\xi) e^{2\pi i x_j \cdot \xi} p(\xi) d\xi$$

is related to an integral which is related to the N-dimensional Fourier transform by Monte Carlo integration. It would be interesting to consider ways to understand and extend this work.

2.4. **Real Life Applications.** Many of these techniques are so new that they have never been applied to real data; a recent project of the Applied Math Group used them to understand relevant factors in Global Trade behavior [2]. It would be desirable to have more real life examples showing that mathematics can really be useful for this type of analysis.

Requirements. Math 225 or 230/231 and some coding experience. Math 305, 320, 325 can be helpful but are not strictly required. A basic understanding of probability and a good grasp of matrices is helpful.

References

[1] R. R. Coifman, N. F. Marshall. Manifold learning with bi-stochastic kernels. *arXiv:1711.06711*, 2017.

[2] Y. Li, T. Wu, N. Marshall and S. Steinerberger, Extracting geography from trade data, Physica A, p. 205–212.

[3] G. C. Linderman, G. Mishne, Y. Kluger, S. Steinerberger. Randomized Near Neighbor Graphs, Giant Components, and Applications in Data Science. *arXiv:1711.04712*, 2017.

[4] F. X. Yu, A. T. Suresh, K. Choromanski, D. Holtmann-Rice, S. Kumar Orthogonal Random Features. *arXiv:1610.09072*, 2016.