Graph Fourier MMD for Signals on Graphs

Samuel Leone\textsuperscript{1} Aarthi Venkat\textsuperscript{2} Guillaume Huguet\textsuperscript{3} Alexander Tong \textsuperscript{4} Guy Wolf\textsuperscript{3} Smita Krishnaswamy\textsuperscript{6}

Abstract—While numerous methods have been proposed for computing distances between probability distributions in Euclidean space, relatively little attention has been given to computing such distances for distributions on graphs. However, there has been a marked increase in data that either lies on graph (such as protein interaction networks) or can be modeled as a graph (single cell data), particularly in the biomedical sciences. Thus, it becomes important to find ways to compare signals defined on such graphs. Here, we propose Graph Fourier MMD (GFMMD), a novel distance between distributions and signals on graphs. GFMMD is defined via an optimal witness function that is both smooth on the graph and maximizes the difference in expectation between the pair of distributions on the graph. We find an analytical solution to this optimization problem as well as an embedding of distributions that results from this method. We also prove several properties of this method including scale invariance and applicability to disconnected graphs. We showcase it on graph benchmark datasets as well on single cell RNA-sequencing data analysis. In the latter, we use the GFMMMD-based gene embeddings to find meaningful gene clusters. We also present a new distance that belongs to the family of integral probability metrics (IPMs) \cite{Hsu2017}. If $H$ is a Reproducing Kernel Hilbert Space (RKHS) of measurable, bounded functions on $X$, and two measures $P$ and $Q$ on $X$, the IPM between $P$ and $Q$ is defined as

$$\gamma_H(P, Q) = \sup_{f \in H} \mathbb{E}_P(f) - \mathbb{E}_Q(f).$$

Here, $F$ is a family of “witness function” since it emphasizes the differences between $P$ and $Q$, choosing a certain $F$ determines the IPMs. For certain classes of $F$, the resulting distance is called an kernel Maximum Mean Discrepancy (MMD) \cite{L2006}. If $H$ is a Reproducing Kernel Hilbert Space...
(RKHS) of functions on \( X \) (equipped with norm \( \| \cdot \|_H \)), then the IPM corresponding to \( \mathcal{F} = \{ f : \| f \|_H \leq 1 \} \) is an MMD. Numerous distances between distributions are IPMs, given a suitable choice of \( \mathcal{F} \). For example, the Wasserstein distance is an IPM where \( \mathcal{F} \) corresponds to the family of Lipschitz functions.

**b) Wasserstein Distance:** The Earth Mover’s Distance (EMD), also known as the 1-Wasserstein distance, is a distance between probability distributions designed to measure the least amount of “work” it takes to move mass from one distribution to another. Formally, we are given two distributions \( P \) and \( Q \) on a measure space \( (\Omega, \mathcal{F}, \mu) \) and a distance \( d : X \times X \to \mathbb{R} \). Most commonly, \( \Omega \) might be a Riemann manifold, \( \mathbb{R}^d \), or in our case, a finite graph. We define the space of couplings of \( P \) and \( Q \), denoted \( \Pi(P, Q) \), to be the set of joint probability distributions whose marginals are equal to \( P \) and \( Q \).

**Definition 2.** The 1-Wasserstein Distance between \( P \) and \( Q \) is defined to be:

\[
W(P, Q) = \min_{\pi \in \Pi(P, Q)} \int_{X \times X} d(x, y)\pi(dx, dy).
\]

The supremum joint distribution \( \pi \) would then be called the *optimal transport plan*. In the case that \( \Omega \) is finite (say of size \( n \)), \( \Pi(P, Q) \) could be thought of as the set of \( n \times n \) matrices \( \pi \) for which \( \pi 1 = P \) and \( 1 \pi = Q \). Then we could represent distances in an \( n \times n \) matrix \( D \), and the EMD is given by \( \min_{\pi} \pi \cdot D \). Typical solutions to EMD in its *primal form* are found using linear programming. The Kantorovich-Rubinstein Theorem, however, provides a dual formulation in terms of smooth functions:

**Theorem 1.** (Kantorovich-Rubinstein) The EMD is an IPM with \( \mathcal{F} \) the space of 1-Lipschitz functions

\[
W(P, Q) = \sup_{\| f \| \leq 1} \mathbb{E}_P(f) - \mathbb{E}_Q(f).
\]

We refer to [6] for a proof of the previous theorem. Intuitively, we can think of suitable functions \( f \) as being varying slowly over \( X \). The 1-Lipschitz constraint prevents witnesses from behaving too erratically over the space. As we will see, duality provides a valuable intuition for using smooth functions to compare functionals in abstract spaces.

c) The Graph Laplacian: For a weighted graph \( G = (V, E, w) \) on \( n \) vertices, we have a number of associated matrices. The first of which is an adjacency / affinity matrix \( A \) for which, given vertices \( a \) and \( b \), \( A(a, b) = w(a, b) \); for our purposes, we assume \( w(a, b) \geq 0 \). In the case when \( V \) belongs to a metric space \( (X, d) \), we have an associated distance matrix \( M \) for which \( M(a, b) = d(a, b) \) for all \( a, b \in V \). Oftentimes, the affinity matrix \( A \) is generated by a nonlinear kernel function \( k(\cdot) \) so that \( A(a, b) = k(M(a, b)) \). For our purposes, if \( A \) is generated in this way, we will call \( G \) a *affinity graph*. There is also a diagonal degree matrix for which \( D(a, a) = \sum_{b \in V} w(a, b) \). Finally, we define the combinatorial Laplacian \( L = D - A \). It can be shown that for any function on the vertices \( f, f^\top L f = \sum_{(a,b) \in E} w(a,b) (f(a) - f(b))^2 \). From this, it’s clear that \( L \) is positive semi-definite, and thus has a spectrum \( \{ \lambda_1, \psi_1 \ldots \lambda_n, \psi_n \} \), where \( \lambda_1 \geq \ldots \geq \lambda_n \).

d) Effective Resistances: In [15], graphs are regarded as electrical circuits with edge weights providing capacities. In such a graph, the *effective resistance* \( \text{Re}(a, b) \) between vertices \( a \) and \( b \) is equal to \( \| L^{-\frac{1}{2}} \delta_a - L^{-\frac{1}{2}} \delta_b \|_2^2 \), where \( \delta_a, \delta_b \) are the one-hot encodings of vertices \( a \) and \( b \), respectively. Effective resistances provide valuable information about a graph. For instance, Spielman & Srivastava use effective resistances between adjacent vertices to sparsify a graph. We will show, if we view \( \delta_a, \delta_b \) as probability densities concentrated at \( a, b \), Graph Fourier MMD provides an extension of resistances to arbitrary probability distributions on graphs. Effective resistances have been shown to be related to commute times, thus this provides a generalization of commute time when the initial and final position are not localized to a single node [1].

**B. Related Work**

The closest related work is that of [19], which constructs a family of spectral distances between graph signals based on weighted Fourier transforms. Algebraically, our distance \( G\mathcal{F}_{MMD} \) resembles a special case of these distances. Another similar distance is Diffusion EMD [17], which involves diffusion graph signals to different scales using a diffusion operator (similar to that of a diffusion map [3]) to create multiscale density estimates of the data. Then Diffusion EMD computes weighted \( L^1 \) distance between the multiscale density estimates of different signals. While this method is faster than most primal methods for EMD computation, it can be inaccurate unless the graph is significantly large.

In [10], [11], [8], the authors consider the EMD between distributions defined on a *distance* graph, that is the edge weights define the cost of moving mass from one node to another. The authors in [10], [11] provide a closed-form solution that relies on a graph shortest path distance. In this setting, there is no sparse approximation to diffusion distances in terms of graph shortest path. We consider a different problem where the edges of the graph are affinities. Among methods for MMD, the most common method has been a sampling based method that also forms a 2-sample Kernel test based on defining a kernel between empirical observations [9]. Note that semantically this takes distances between point clouds themselves by modeling them as a data graph with vertices as points. We define a method of taking signals which generalizes to an arbitrary graph, on a point cloud or otherwise, and demonstrate its effectiveness both when the graph lies in a metric space and when adjacencies are binary.

**II. METHODS**

Given two probability distributions \( P \) and \( Q \) on an arbitrary graph, we are interested in taking a meaningful distance between them in a way that incorporates graph structure. Unlike the distance setting, there is no obvious notion of Lipschitzness. However, there is still a notion of smoothness. Indeed, we define Graph Fourier MMD as the MMD induced by witness functions \( f \) which are *smooth over the graph*. That is, they have a low value in \( f^\top L f \).
Theorem 2. Let \( \mathcal{G} = (\mathcal{V}, \mathcal{E}, \omega) \) be a finite graph with Laplacian \( \mathbf{L} \) and \( P, Q \) be two bounded probability distributions on \( \mathcal{V} \). The Graph Fourier MMD between \( P \) and \( Q \) is

\[
\mathcal{GFMMD}(P, Q) = \max_{f : \mathbf{L}f \leq 1} \mathbb{E}_P(f) - \mathbb{E}_Q(f).
\]

Note that this definition holds for any construction of a positive semi-definite Laplacian matrix \( \mathbf{L} \). We can show that, under reasonable conditions, \( \mathcal{GFMMD} \) is finite and simple to compute. First, we need to establish a property which incorporates graph structure.

Definition 3. Let \( \mathcal{G} = (\mathcal{V}, \mathcal{E}, \omega) \) be a finite graph with Laplacian \( \mathbf{L} \) and \( P, Q \) be two probability distributions. Two probability distributions \( P \) and \( Q \) differ, but smoothness requires that it do so while incorporating graph structure.

Definition 4. Two probability distributions \( P \) and \( Q \) are said to have equal component mass if, for all connected components \( S \) of \( \mathcal{G} \), \( \mathbb{E}[P|v \in S] = \mathbb{E}[Q|v \in S] \).

Theorem 2. Let \( P \) and \( Q \) be bounded probability distributions defined on \( \mathcal{V} \). If \( P \) and \( Q \) have equal component mass, then \( \mathcal{GFMMD}(P, Q) = \| \mathbf{L}^{-\frac{1}{2}}(P - Q) \|_2 \). And otherwise, \( \mathcal{GFMMD}(P, Q) = +\infty \).

Proof. Suppose first that \( P \) and \( Q \) do not have the equal mass property. Then there exists a connected component \( S \) for which,

\[
\sum_{v \in S} P(v) < \sum_{v \in S} Q(v)
\]

In particular, we can write \( \sum_{v \in S} P(v) = \sum_{v \in S} Q(v) - c \) for some \( c > 0 \). Now, let \( f_\alpha \) be a signal such that \( f_\alpha(v) = \alpha \) if \( v \in S \) and \( f_\alpha(v) = 0 \) otherwise. Then certainly, \( f_\alpha^\top \mathbf{L} f_\alpha = 0 \), since it is known that indicator functions for connected components are in the null space of \( \mathbf{L} \). And so \( f_\alpha^\top \mathbf{L} f_\alpha \leq 1 \), yet,

\[
\mathbb{E}_P(f_\alpha) - \mathbb{E}_Q(f_\alpha) = \sum_{v \in \mathcal{V}} P(v) f_\alpha(v) - \sum_{v \in \mathcal{V}} Q(v) f_\alpha(v)
= \sum_{v \in S} \alpha P(v) - \sum_{v \in S} \alpha Q(v)
= \alpha c
\]

As \( \mathcal{GFMMD} \) is defined as \( \sup_{f : \mathbf{L}f \leq 1} \mathbb{E}_P(f) - \mathbb{E}_Q(f) \), we have \( \mathcal{GFMMD}(P, Q) \geq \alpha c \). Taking \( \alpha \to \infty \), we have \( \mathcal{GFMMD}(P, Q) = +\infty \).

Suppose on the other hand that \( P \) and \( Q \) do have the equal mass property. If we let \( \mathbb{1}_{S_1}, \ldots, \mathbb{1}_{S_m} \) be indicator functions for connected components \( S_1, \ldots, S_m \), the equal mass property insists that \( P^\top \mathbb{1}_{S_i} = Q^\top \mathbb{1}_{S_i} \) for all \( i \). And thus, \( (P - Q)^\top \mathbb{1}_{S_i} = 0 \). Since it is known that these indicator functions form a basis for the kernel of \( \mathbf{L} \), it follows that \( P - Q \in \ker(\mathbf{L})^\perp \). Now, any function \( f \) such that \( f^\top \mathbf{L} f \) can be broken up into \( f = f_1 + f_2 \), where \( f_1 \in \ker(\mathbf{L}) \) and \( f_2 \in \ker(\mathbf{L}) \). Finally, observe that we can view \( P \) and \( Q \) as probability vectors indexed over \( \mathcal{V} \). Thus,

\[
\mathbb{E}_P(f) - \mathbb{E}_Q(f) = P^\top f - Q^\top f = (P - Q)^\top (f_1 + f_2)
\]

\[
= (P - Q)^\top f_1 + (P - Q)^\top f_2
\]

\[
= (P - Q)^\top f_2
\]

Furthermore, \( f^\top \mathbf{L} f = f_2^\top \mathbf{L} f_2 \). Combined, these observations tell us that we may assume, without loss of generality, that \( f \in \ker(\mathbf{L})^\perp \). And thus,

\[
\mathcal{GFMMD}(P, Q) = \sup_{f : \mathbf{L}f \leq 1, f \in \ker(\mathbf{L})^\perp} (P - Q)^\top f
\]

For any such \( f \), we can define \( y = \mathbf{L}^{-\frac{1}{2}} f \). And thus, \( f^\top \mathbf{L} f = f^\top \mathbf{L}^{\frac{1}{2}} \mathbf{L}^{-\frac{1}{2}} f = \| y \|_2^2 \). Furthermore, since \( f \in \ker(\mathbf{L})^\perp \), \( f = \mathbf{L}^{-\frac{1}{2}} y \). Here, \( \mathbf{L}^{-\frac{1}{2}} \mathbf{L}^{\frac{1}{2}} \) is the square root of the Moore-Penrose pseudoinverse \( \mathbf{L}^\dagger \) of \( \mathbf{L} \). Thus,

\[
\mathcal{GFMMD}(P, Q) = \sup_{y : \| y \|_2 \leq 1} (P - Q)^\top \mathbf{L}^{-\frac{1}{2}} y
\]

Which clearly, by Cauchy Schwarz, is simply equal to \( \| (P - Q)^\top \mathbf{L}^{-\frac{1}{2}} \mathbf{L}^{\frac{1}{2}} (P - Q) \|_2 \), as desired. □

The effect of equal component mass is highly intuitive: we would expect “infinite effort” to move a probability distribution between disconnected sets of vertices. \( \mathcal{GFMMD} \) possesses a set of convenient properties. Namely, we have a representation in terms of an explicit feature map \( \mathbf{L}^{-\frac{1}{2}} \). So to compute pairwise distances, it is sufficient to apply the feature map and then take Euclidean distances. Also, the distance value is a true distance (particularly an MMD).

Lemma 3. (i) \( \mathcal{GFMMD}(\cdot, \cdot) \) defines a valid distance on the probability distributions acting on \( \mathcal{V} \). Furthermore, (ii) \( \mathcal{GFMMD}(P, Q) \) is a Maximum Mean Discrepancy with explicit feature map \( \mathbf{L}^{-\frac{1}{2}} \), when it is finite.

A. Relationship with f-Spectral Distances

The \( f \) spectral distance between two signals \( P \) and \( Q \) per [19] is defined as \( \sum_i f(\lambda_i) (\hat{P}(i) - \hat{Q}(i))^2 \), where \( f \) is some monotone increasing or decreasing function, and \( \hat{P}, \hat{Q} \) denotes the Fourier transform of \( P \) and \( Q \). Likewise, \( \mathcal{GFMMD}(P, Q) = \sum_{\lambda \neq 0} \frac{1}{\lambda} (\hat{P}(i) - \hat{Q}(i))^2 \). Algebraically, \( \mathcal{GFMMD} \) resembles the \( f \)-spectral distance for \( f(x) = 1/x \), although there is different treatment of components in \( \ker(\mathbf{L}) \), which arises from the fact that we only consider probability distributions.
B. Relationship with Resistive Embeddings

Graph Fourier MMD provides a natural extension of resistive embeddings through application of the same feature map. In fact, resistive embeddings can be viewed as a special case of Graph Fourier MMD for dirac distributions. On the other hand, theorem 2 offers a variational characteristic of resistive embeddings, which could be used for lower bounds. Finally, it can be shown that Graph Fourier MMD can be related to the the expected effective resistance of $X \sim P$ and $Y \sim Q$. Not only does this provide an interpretation of GFMD, but it illustrates that spectral sparsification algorithms such as [15] preserve GFMD.

Theorem 4. If $X \sim P$ and $Y \sim Q$, not necessarily independent, then $\mathcal{GFMMD}(P, Q)^2 \leq \mathbb{E}_{X,Y}[\text{Re}(X,Y)]$.

C. Graph Fourier MMD for Signal Localization

By fixing a probability distribution $P$ and considering its distance to the uniform distribution $U = \frac{1}{n} \mathbb{1}$, we obtain a measure of how much $P$ concentrates on the graph. For a high GFMD to the uniform distribution, we say a signal is localized, and otherwise we say it is dispersed. Signal localization simply corresponds to length in feature space.

Definition 5. The Localization Score $s(P)$ of a signal $P$ is defined as $\mathcal{GFMMD}(P, U) = \| L^{-\frac{1}{2}} (P - U) \|_2 = \| L^{-\frac{1}{2}} P \|_2$.

Interestingly, this is related to pre-existing notions of spread on a graph. For example, [18] defines the best spreader as the vertex $v$ minimizing $\sum_i |L^i v|^2$. This vertex also is also that whose point mass distribution has the lowest localization score & greatest dispersion.

D. Computational Complexity and Speedup

Consider the task of getting the pairwise GFMD between a collection of probability distributions. Computation of the pseudoinverse or $L^{-\frac{1}{2}}$ is roughly $O(n^3)$. An alternate approach would be calculate $L^{-\frac{1}{2}} (P - Q)$ by computing the $y$ that solves $Ly = P - Q$ via conjugate gradient descent, and then computing $\mathcal{GFMMD}(P, Q) = (P - Q)^\top y$. We elect to use Chebyshev polynomials to approximate $L^{-\frac{1}{2}} P$ and $L^{-\frac{1}{2}} Q$, then take Euclidean distances. This is a large improvement in the particular case where the graph $G$ is sparse i.e. $|E| = O(n \log n)$. In such cases, we present an $O(n \log n)$ algorithm for the computation of GFMD which is substantially faster than naive implementations based on a Chebyshev polynomial approximation of the filter in Algorithm 1 as well as a KNN kernel. The steps for an arbitrary graph are the same, but with $W$ provided.

This gives a total fine grained time complexity of $O((k_1 + t)n \log n + k_2 nm \log m)$ and a space complexity of $O(n \log n + nm)$ space. Here, $t$ is the order of the Chebyshev polynomial, $k_1$ is the threshold for number of nearest-neighbors in constructing $G$, and $k_2$ is the number of nearest-distributions we’d like to calculate. More simply, for fixed Chebyshev order, and number of neighbors, the time to estimate distances between all distributions is $O(n \log n + nm^2)$.

Algorithm 1 GFMD in Metric Space

Input: A set of $n$ points $X \subseteq \mathbb{R}^d$, $m$ probability distributions $f_i : X \rightarrow \mathbb{R}$ in an $n \times m$ matrix $F$, and a kernel function $k : X \times X \rightarrow \mathbb{R}$

Output: An $m \times n$ embedding matrix $E$ in which $\| E_i - E_j \| = \mathcal{GFMMD}(f_i, f_j)$. $M$ and a distance matrix in which $M_{ij} = \mathcal{GFMMD}(f_i, f_j)$

Create a thresholded K-Nearest Neighbor graph $G$ over $X$ with $O(n \log n)$ edges

$W_{ij} \leftarrow k(X_i, X_j)$ for all $(i, j) \in E$

$L \leftarrow D - W$

$E_i \leftarrow L^{-\frac{1}{2}} f_i$ by exact or Chebyshev approximation of the filter $h(\lambda) = \lambda^{-\frac{1}{2}}$

$M_{ij} \leftarrow \| E_i - E_j \|$ for all $i, j \in [n]$

return $M, E$

III. Experimental Results & Applications

In our experiments, signals are always nonnegative and normalized to be interpreted as probability distributions. However, the metric induced by $L^{-\frac{1}{2}}$ always induces a valid seminorm on graph signals (but in particular, a norm on probability distributions).

A. Identifying Distributions on the Swiss Roll

In this experiment, we generate random point clouds centered at points on the swiss roll. More specifically, we sample $n = 100$ points on the swiss roll $x_1, x_2, \ldots, x_n$ and around each of these points, generate a point cloud $d_i$ of size $m = 100$ points from a multivariate normal distribution centered at $x_i$. The result is $nm$ points in $\mathbb{R}^{10}$. For each $i, j \in [n]$, we have a known geodesic distance between $x_i$ and $x_j$. Across the different measures, we can see how well the distance between the point clouds $d_i$ and $d_j$ compares to the geodesic distance between their corresponding centers $x_i$ and $x_j$. We compare the induced probability distributions on point clouds using: 1) computation of earth Mover’s Distance between point clouds in ambient space, 2) Sinkhorn algorithm [4], 3) Diffusion EMD [17], 4) Kernel MMD [9] between all pairs via random sampling (20 points from each distribution with replacement), 5) Graph Fourier MMD between $p_i, p_j$, using both the exact calculation and approximation via Chebyshev polynomials. We then take correlation between the estimated nearest distributions and geodesic distance between centers. Results are shown in Table I. As we see, Graph MMD outperforms all other methods in accuracy and speed. In the appendix, the corresponding feature maps are visualized alongside geodesic distance, where GFMD visually outperforms other methods at extracting manifold nonlinearities.

B. Single cell Analysis with GFMD

To demonstrate the utility of Graph Fourier MMD for biological analysis, we leverage publicly available single-cell RNA sequencing dataset of CD8-positive T cells [21]. CD8-positive T cells are adaptive immune cells known to be critical for mediating immune response in infection, cancer, and other
The Chebyshev approximated Graph MMD (Chebyshev, t) is extremely fast and almost as performant at even low orders t. GFMMD is most performant but requires a eigen-decomposition. 100 points each on a swiss roll manifold. The exact Graph MMD is most performant but requires a eigen-decomposition. of 100 points each on a swiss roll manifold. The exact Graph MMD is most performant but requires a eigen-decomposition. We propose the use of the localization score of the genes, viewed as probability distributions over the cells. Based on this score, in Figure 3, we visualize first most local gene, 10th most local gene, and 20th most local. Here, we compare localization scores between housekeeping genes and the gene signatures for naive CD8+ T cells. Housekeeping genes are expressed highly in many systems, but are not known to have a function that contributes strongly to cell-cell variation for T cells [7], [20], [5]. By contrast, cells enriched for the naive CD8+ T cell signature are a subset of T cells along the T cell differentiation axis. We show that the localization score is an order of magnitude higher for the naive gene signature versus the housekeeping signature Figure 3 (see the appendix), validating our intuition about localized genes.

IV. Conclusion

In this paper we have introduced Graph Fourier MMD, a framework for taking distances between signals on graphs and generating embeddings in which these distances hold. We have shown its intuitive performance in both the Riemannian and abstract graphical setting for known distributions, as well as its advantage in speed, and ability to capture global properties of the underlying data manifold compared to alternative methods like Earth Mover’s Distance and Diffusion EMD. Its rapidity makes it particularly useful for high dimensional datasets, such as single cell data, where we have showed its ability to capture the natural trajectories of gene expression.
REFERENCES


