Geometric deep learning: going beyond Euclidean data

Michael M. Bronstein, Universita della Svizzera Italiana (Facebook Research)
Joan Bruna, Univ. Politec. de Catalunya (Facebook Research)
Yann LeCun, Facebook Research
Arthur Szlam, Facebook Research
Pierre Vandergheynst, Swiss Federal Institute of Technology

IEEE Signal Processing Magazine, 2017, 34(4):18-42.

Abstract

- Many scientific fields study data with an underlying structure that is a **non-Euclidean** space.
    - Graph and Manifold
    - Social networks, sensor networks, brain imaging
- Deep neural networks have recently proven to be powerful tools for a broad range of problems.
    - Data with an underlying Euclidean or grid-like structure
- **Geometric deep learning** is an umbrella term for emerging techniques attempting to generalize (structured) deep neural models to non-Euclidean domains

Introduction

- Deep learning
  - *Learning complicated concepts by building them from simpler ones in a hierarchical or multi-layer manner*
  - Neural networks are popular realizations of such deep multi-layer hierarchies

  - Qualitative breakthroughs
    - Speech recognition, machine translation…
    - The growing computational power and availability of large training datasets

Introduction

- **Why it's successful**
  - Ability to leverage statistical properties of the data
    - Stationarity and Compositionality (local statistics)
  - E.g., consider images as functions on the Euclidean space
    - **Stationarity** is owed to shift-invariance;
    - **Locality** is due to the local connectivity;
    - **Compositionality** stems from the multi-resolution structure of the grid.
- These properties are exploited by convolutional architectures
  - Extracting local features and reduces the number of parameters
  - Imposes some priors about the data

Introduction

- **Non-Euclidean geometric data**
  - In social networks, the characteristics of users can be modeled as **signals** on the vertices of the social graph
  - Sensor networks are graph models of distributed interconnected sensors, whose readings are modelled as time-dependent **signals** on the vertices
  - In genetics, gene expression data are modeled as **signals** defined on the regulatory network
- Non-Euclidean data implies no familiar properties
  - global parameterization, vector space structure, shift-invariance.
- The purpose of our paper is to show different methods of translating the key ingredients of successful deep learning methods such as convolutional neural networks to non-Euclidean data.

GEOMETRIC LEARNING PROBLEMS

- **Two classes of geometric learning problems**
  - 1. To characterize the structure of the data
  - 2. Analyzing functions defined on non-Euclidean domain
- First class: structure of the domain (embedding)
  - *Manifold learning* or *non-linear dimensionality reduction*
    - Data points with underlying lower dimensional structure embedded into a high-dimensional Euclidean space
    - Recovering the lower dimensional structure
  - Two steps of Manifold learning
    - Constructing a representation of local affinity of the data points
    - Data points are embedded into a low-dimensional space trying to preserve some criterion of the original affinity.

GEOMETRIC LEARNING PROBLEMS

- Second class: analyzing functions defined on a given non-Euclidean domain
  - Prediction: e.g., assume that we are given the geographic coordinates of the users of a social network, represented as **a time-dependent signal on the vertices** of the social graph. An important application in location-based social networks is **to predict the position of the user** given his or her past behavior, as well as that of his or her friends.

DEEP LEARNING ON EUCLIDEAN DOMAINS

- **Geometric priors**
  - 1. *Stationarity*
    - Translation operator

$$\mathcal{T}_v f(x) = f(x - v), \quad x, v \in \Omega,$$

    - The assumption is that the function y is either invariant or equivariant with respect to translations
  - $$y(\mathcal{T}_v f) = y(f)$$

DEEP LEARNING ON EUCLIDEAN DOMAINS

- **Geometric priors**
  - 2. *Local deformations and scale separation*
    - Deformation operator

$$\mathcal{L}_\tau f(x) = f(\underline{x - \tau(x)})$$

local translations, changes in point of view, rotations

$$|y(\mathcal{L}_\tau f) - y(f)| \approx \underline{\|\nabla \tau\|}$$

smoothness of a given deformation field

    - The quantity to be predicted does not change much if the input image is slightly deformed

DEEP LEARNING ON EUCLIDEAN DOMAINS

- **Geometric priors**
  - 2. *Local deformations and scale separation*
    - Long-range dependencies can be broken into Multi-scale local interaction terms

$$Y(x_1, x_2; v) = \mathrm{Prob}(f(u) = x_1 \ \text{and} \ f(u+v) = x_2)$$

      - *a*. this joint distribution will not be separable for any v
      - *b*. $Y(x_1, x_2; v) \approx Y(x_1, x_2; v(1+\epsilon))$ for small ϵ

    - long-range dependencies can be captured and down-sampled at different scales

DEEP LEARNING ON EUCLIDEAN DOMAINS

- **Convolutional neural networks**
  - Convolutional layers $\mathbf{g} = C_{\Gamma}(\mathbf{f})$

nonlinearity

Convolutional filter

$$g_l(x) = \xi\left(\sum_{l'=1}^{p}(f_{l'} \star \gamma_{l,l'})(x)\right)$$

q-dimensional output

p-dimensional input

DEEP LEARNING ON EUCLIDEAN DOMAINS

- **Convolutional neural networks**
  - Pooling layers

a neighborhood around x

$$g_l(x) = P(\{f_l(x') : x' \in \mathcal{N}(x)\}), \ \ l = 1, \ldots, q,$$

permutation-invariant function such as a L_p-norm (in the latter case, the choice of $p = 1, 2$ or ∞ results in *average-*, *energy-*, or *max-pooling*).

DEEP LEARNING ON EUCLIDEAN DOMAINS

- **Convolutional neural networks**
  - Generic hierarchical representation

$$U_\Theta(f) = (C_{\Gamma(K)} \cdots P \cdots \circ C_{\Gamma(2)} \circ C_{\Gamma(1)})(f)$$

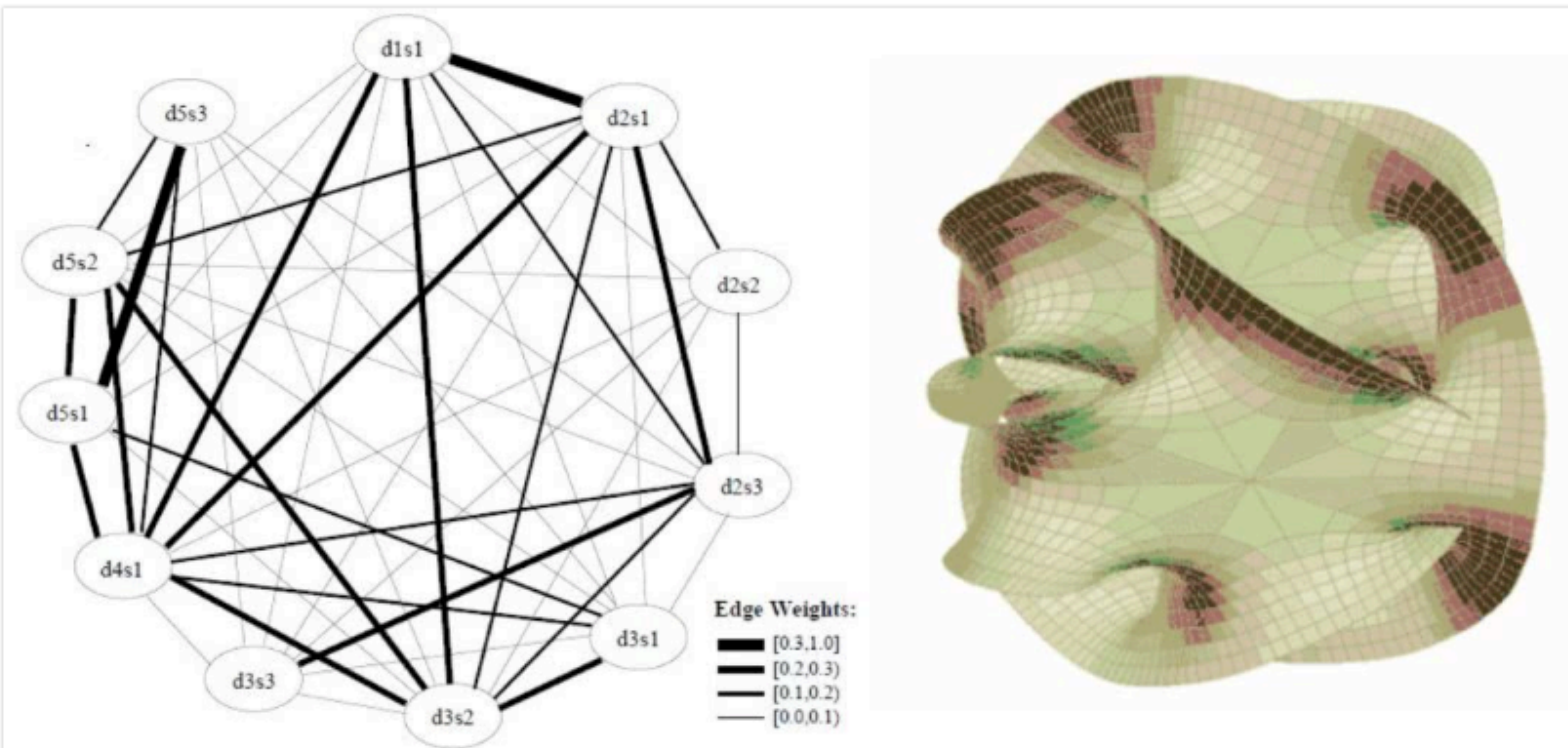
  - Supervised learning

$$\min_{\Theta} \sum_{i \in \mathcal{I}} L(U_\Theta(f_i), y_i),$$

  - A key advantage of CNNs explaining their success is that the **geometric priors** on which CNNs are based result in a learning complexity that avoids the curse of dimensionality.

THE GEOMETRY OF MANIFOLDS AND GRAPHS

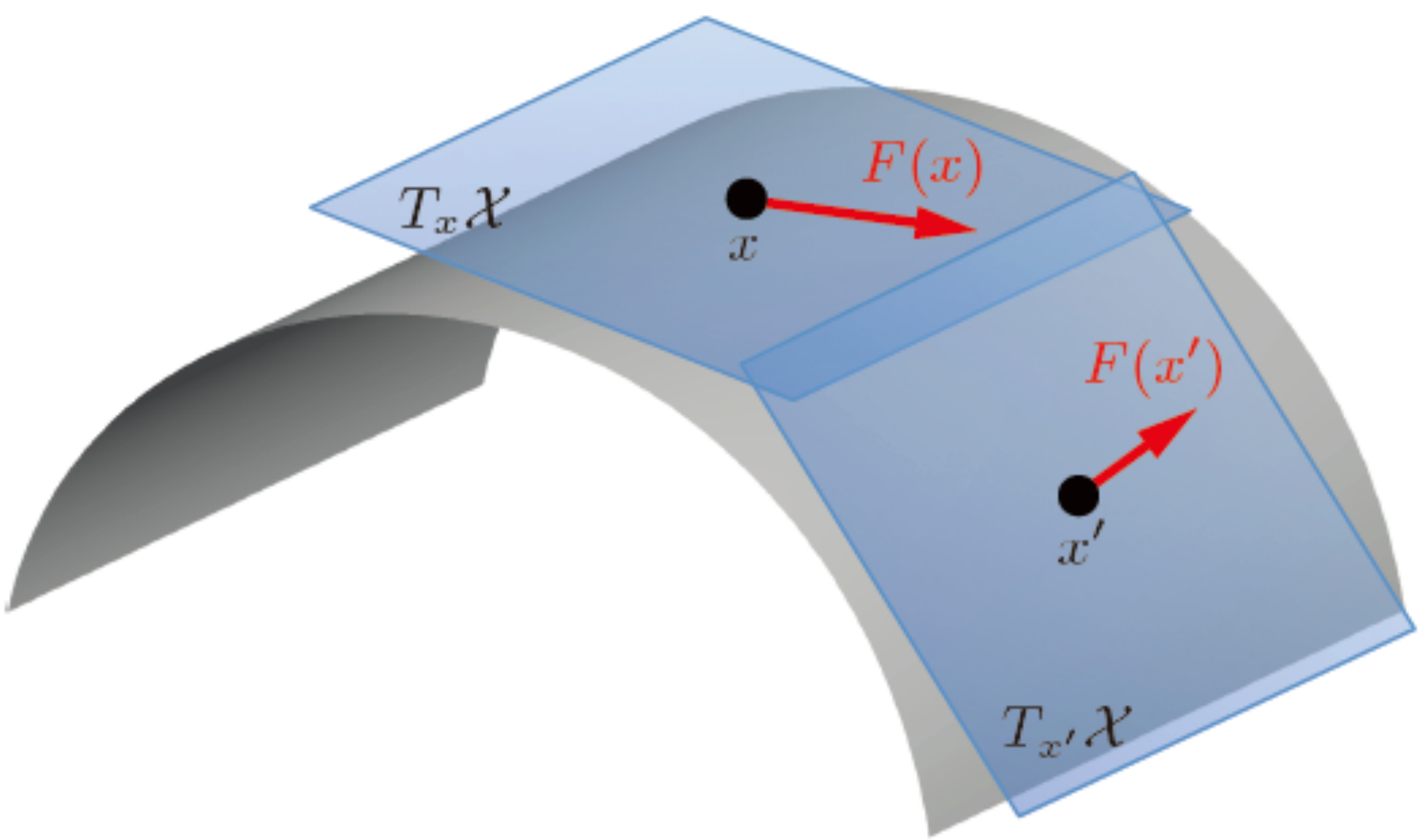
- Two prototypical structures**: manifolds and graphs**



THE GEOMETRY OF MANIFOLDS AND GRAPHS

- **Manifolds**
  - Roughly, a manifold is a space that is locally Euclidean.

Formally speaking, a (differentiable) *d-dimensional manifold* $\mathcal{X}$ is a topological space where each point x has a neighborhood that is topologically equivalent (homeomorphic) to a d-dimensional Euclidean space, called the *tangent space* and denoted by $T_x\mathcal{X}$ (see Figure 1, top).



On each tangent space, we define an inner product

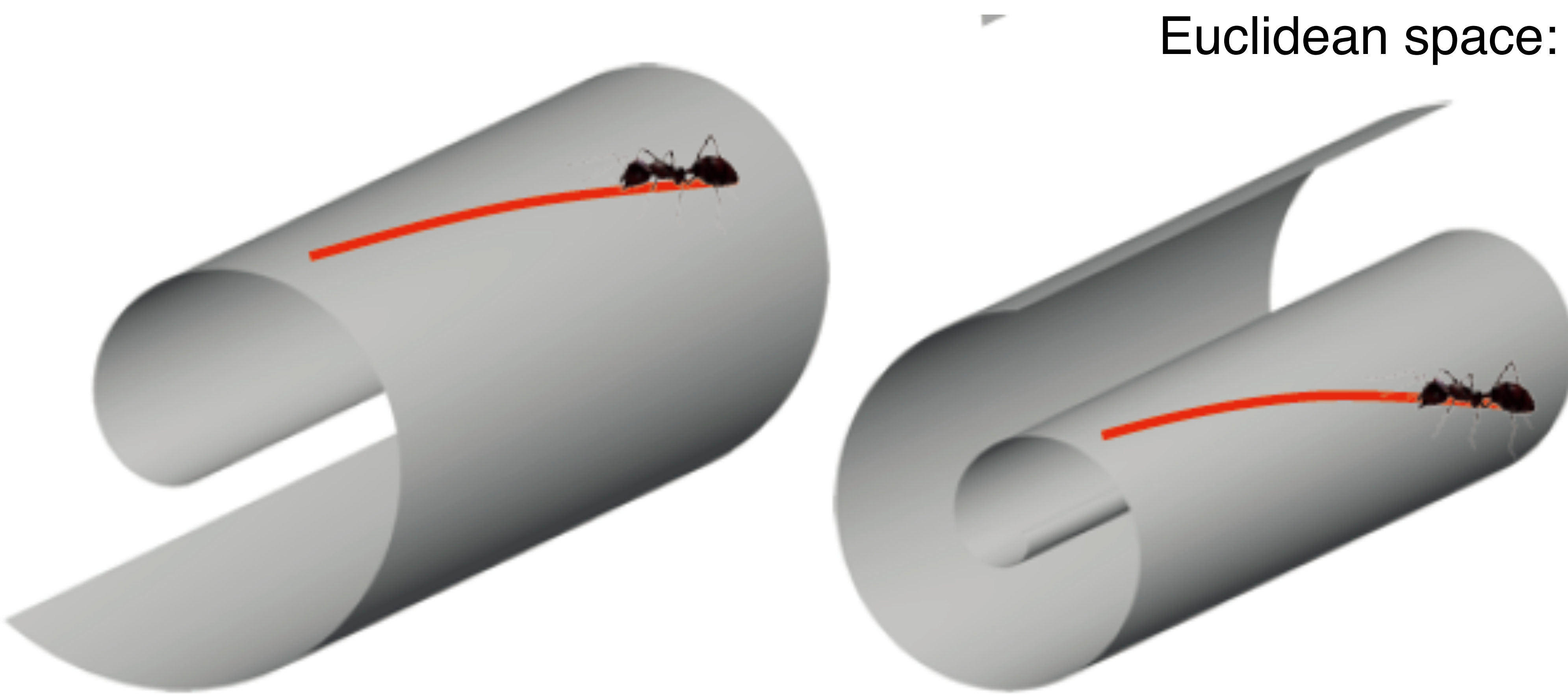
$$\langle \cdot, \cdot \rangle_{T_x\mathcal{X}} : T_x\mathcal{X} \times T_x\mathcal{X} \to \mathbb{R},$$

Riemannian metric in differential geometry

THE GEOMETRY OF MANIFOLDS AND GRAPHS

- **Manifolds**
  - A manifold equipped with a metric is called a **Riemannian manifold**
  - **Nash Embedding Theorem**: any sufficiently smooth Riemannian manifold can be realized in a Euclidean space of sufficiently high dimension

Riemannian metric: intrinsic
Euclidean space: extrinsic



THE GEOMETRY OF MANIFOLDS AND GRAPHS

- **Graphs and discrete differential operators**

weighted undirected graphs, formally defined as a pair $(\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \ldots, n\}$ is the set of n vertices, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges, where the graph being undirected implies that $(i, j) \in \mathcal{E}$ iff $(j, i) \in \mathcal{E}$. Furthermore, we associate a weight $a_i > 0$ with each vertex $i \in \mathcal{V}$, and a weight $w_{ij} \geq 0$ with each edge $(i, j) \in \mathcal{E}$.

THE GEOMETRY OF MANIFOLDS AND GRAPHS

- **Graphs and discrete differential operators**

Real functions $f : \mathcal{V} \rightarrow \mathbb{R}$ and $F : \mathcal{E} \rightarrow \mathbb{R}$ on the vertices and edges of the graph, respectively, are roughly the discrete analogy of continuous scalar and tangent vector fields in differential geometry.[4] We can define Hilbert spaces $L^2(\mathcal{V})$ and $L^2(\mathcal{E})$ of such functions by specifying the respective inner products,

$$\langle f, g \rangle_{L^2(\mathcal{V})} = \sum_{i \in \mathcal{V}} a_i f_i g_i; \tag{20}$$

$$\langle F, G \rangle_{L^2(\mathcal{E})} = \sum_{i \in \mathcal{E}} w_{ij} F_{ij} G_{ij}. \tag{21}$$

THE GEOMETRY OF MANIFOLDS AND GRAPHS

- **Graphs and discrete differential operators**
  - Graph gradient: mapping functions defined on vertices to functions defined on edges

$$\nabla : L^2(\mathcal{V}) \to L^2(\mathcal{E})$$

$$(\nabla f)_{ij} \quad = \quad f_i - f_j$$

  - Graph divergence

$$L^2(\mathcal{E}) \to L^2(\mathcal{V})$$

$$(\mathrm{div} F)_i \quad = \quad \frac{1}{a_i} \sum_{j:(i,j)\in\mathcal{E}} w_{ij} F_{ij}$$

THE GEOMETRY OF MANIFOLDS AND GRAPHS

- **Graphs and discrete differential operators**
  - Graph Laplacian $\Delta = -\mathrm{div}\nabla.$

$$\Delta : L^2(\mathcal{V}) \to L^2(\mathcal{V})$$

$$(\Delta f)_i \quad = \quad \frac{1}{a_i} \sum_{(i,j)\in\mathcal{E}} w_{ij}(f_i - f_j).$$

  - **Intuitive geometric interpretation**: the difference between the local average of a function around a point and the value of the function at the point itself.

$$\mathbf{\Delta f} \quad = \quad \mathbf{A}^{-1}(\mathbf{D} - \mathbf{W})\mathbf{f}.$$

THE GEOMETRY OF MANIFOLDS AND GRAPHS

- **Fourier analysis on non-Euclidean domains**
  - The Laplacian operator is a self-adjoint positive-semidefinite operator

$$\Delta\phi_i = \lambda_i\phi_i, \quad i = 0, 1, \ldots$$

orthonormal eigenfunctions

eigenvalues

  - The eigenfunctions are the smoothest functions in the sense of the Dirichlet energy and can be interpreted as a generalization of the standard **Fourier basis** to a non-Euclidean domain.

[IN3] Physical interpretation of Laplacian eigenfunctions: Given a function f on the domain $\mathcal{X}$, the *Dirichlet energy*

$$\mathcal{E}_{\text{Dir}}(f) = \int_{\mathcal{X}} \|\nabla f(x)\|^2_{T_x\mathcal{X}} dx = \int_{\mathcal{X}} f(x)\Delta f(x)dx, \quad (27)$$

measures how smooth it is (the last identity in (27) stems from (19)). We are looking for an orthonormal basis on $\mathcal{X}$, containing k smoothest possible functions (FIGS3), by solving the optimization problem

$$\begin{aligned}
&\min_{\phi_0} \mathcal{E}_{\text{Dir}}(\phi_0) \quad \text{s.t.} \quad \|\phi_0\| = 1 && (28)\\
&\min_{\phi_i} \mathcal{E}_{\text{Dir}}(\phi_i) \quad \text{s.t.} \quad \|\phi_i\| = 1, \quad i = 1, 2, \dots k-1\\
&\qquad\qquad\qquad\qquad\quad\; \phi_i \perp \text{span}\{\phi_0, \dots, \phi_{i-1}\}.
\end{aligned}$$

In the discrete setting, when the domain is sampled at n points, problem (28) can be rewritten as

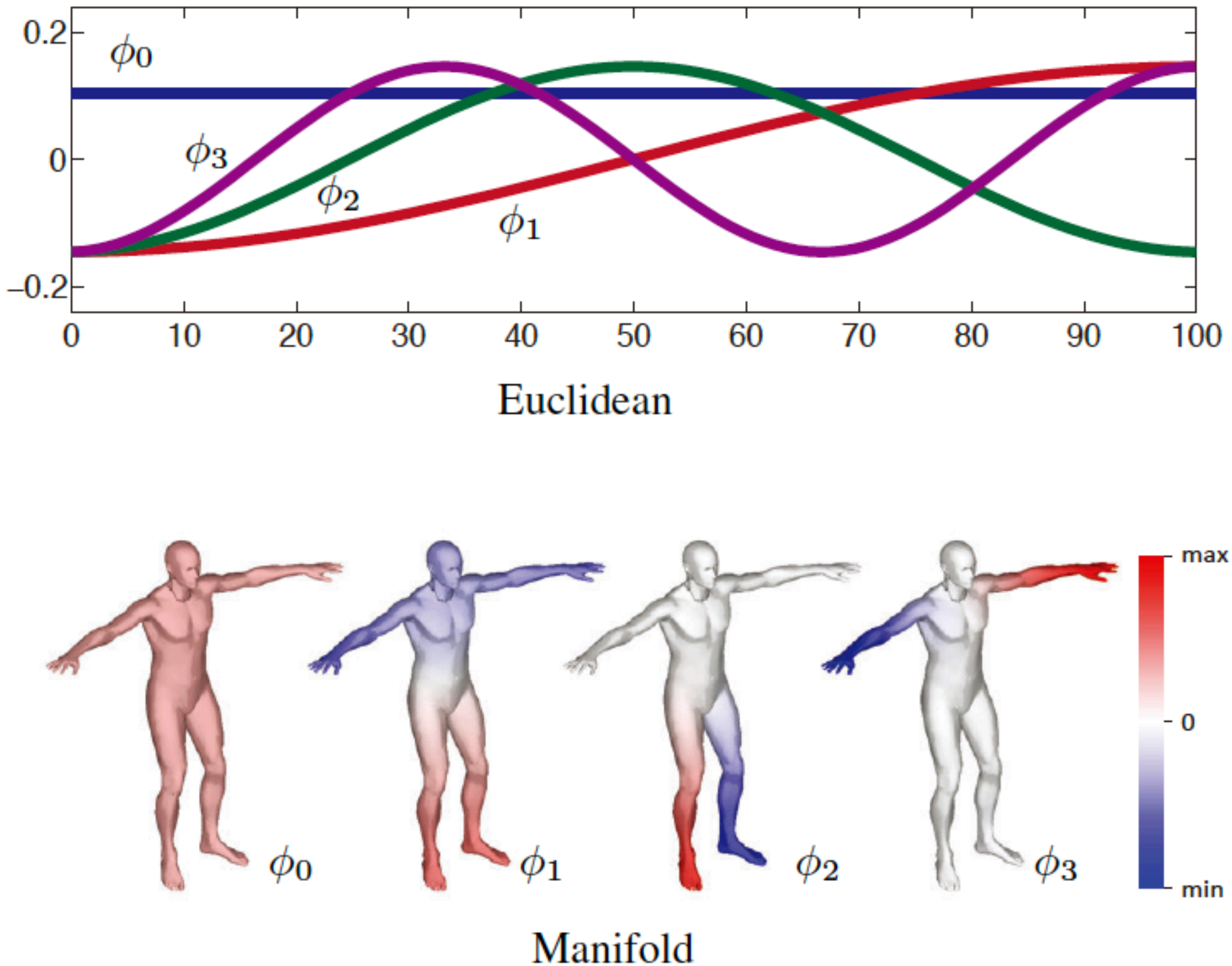
$$\min_{\mathbf{\Phi}_k \in \mathbb{R}^{n\times k}} \text{trace}(\mathbf{\Phi}_k^\top \mathbf{\Delta} \mathbf{\Phi}_k) \quad \text{s.t.} \quad \mathbf{\Phi}_k^\top \mathbf{\Phi}_k = \mathbf{I}, \quad (29)$$

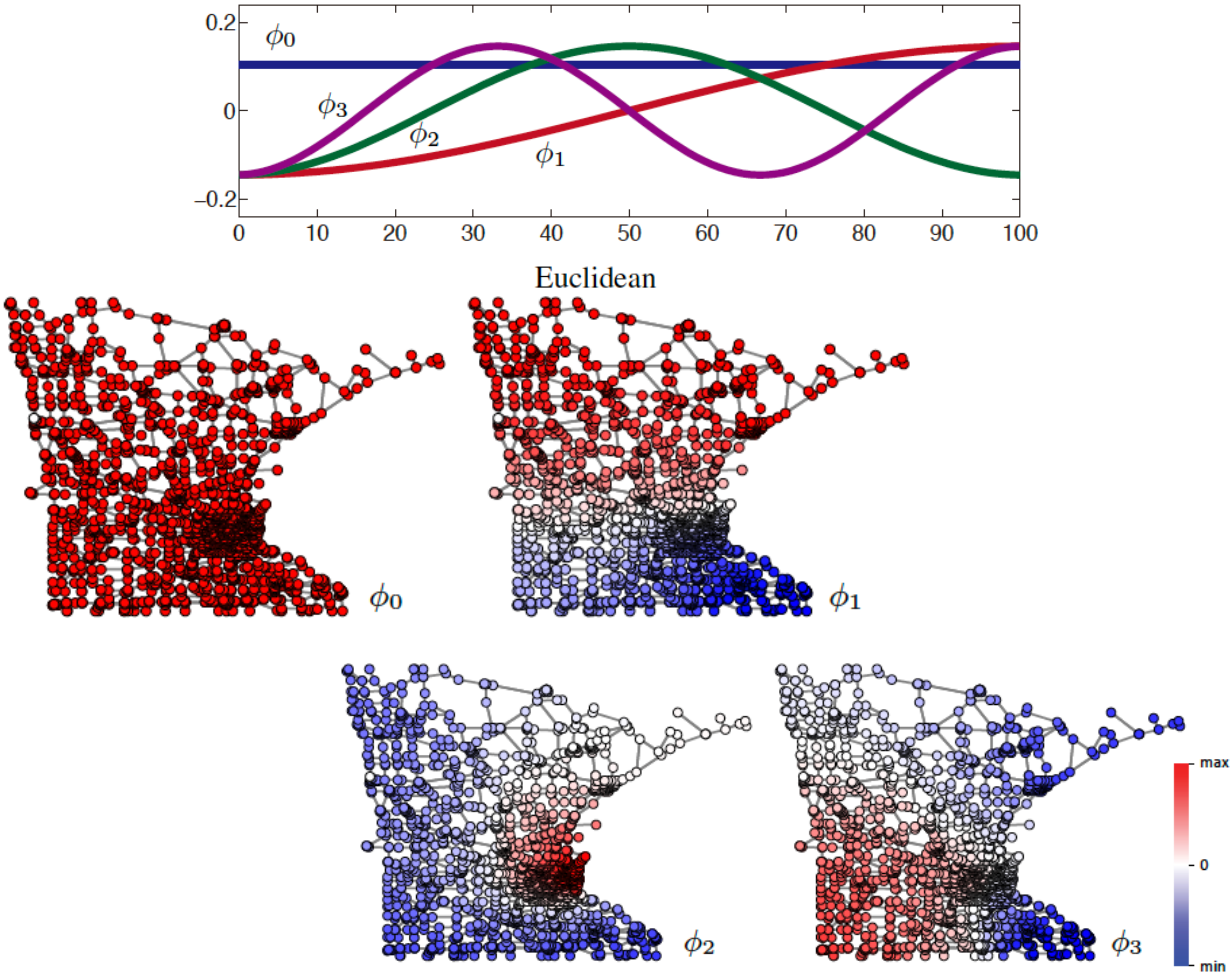
where $\mathbf{\Phi}_k = (\phi_0, \dots \phi_{k-1})$. The solution of (29) is given by the first k eigenvectors of $\mathbf{\Delta}$ satisfying

$$\mathbf{\Delta}\mathbf{\Phi}_k = \mathbf{\Phi}_k \mathbf{\Lambda}_k, \tag{30}$$

where $\mathbf{\Lambda}_k = \mathrm{diag}(\lambda_0, \dots, \lambda_{k-1})$ is the diagonal matrix of corresponding eigenvalues. The eigenvalues $0 = \lambda_0 \leq \lambda_1 \leq \dots \lambda_{k-1}$ are non-negative due to the positive-semidefiniteness of the Laplacian and can be interpreted as 'frequencies', where $\phi_0 = \mathrm{const}$ with the corresponding eigenvalue $\lambda_0 = 0$ play the role of the DC.

The Laplacian eigendecomposition can be carried out in two ways. First, equation (30) can be rewritten as a generalized eigenproblem $(\mathbf{D} - \mathbf{W})\mathbf{\Phi}_k = \mathbf{A}\mathbf{\Phi}_k\mathbf{\Lambda}_k$, resulting in $\mathbf{A}$-orthogonal eigenvectors, $\mathbf{\Phi}_k^\top \mathbf{A}\mathbf{\Phi}_k = \mathbf{I}$. Alternatively, introducing a change of variables $\mathbf{\Psi}_k = \mathbf{A}^{1/2}\mathbf{\Phi}_k$, we can obtain a standard eigendecomposition problem $\mathbf{A}^{-1/2}(\mathbf{D} - \mathbf{W})\mathbf{A}^{-1/2}\mathbf{\Psi}_k = \mathbf{\Psi}_k\mathbf{\Lambda}_k$ with orthogonal eigenvectors $\mathbf{\Psi}_k^\top\mathbf{\Psi}_k = \mathbf{I}$. When $\mathbf{A} = \mathbf{D}$ is used, the matrix $\mathbf{\Delta} = \mathbf{A}^{-1/2}(\mathbf{D} - \mathbf{W})\mathbf{A}^{-1/2}$ is referred to as the *normalized symmetric Laplacian.*


0.2
0
–0.2
0
10
20
30
40
50
60
70
80
90
100
ϕ_0
ϕ_1
ϕ_2
ϕ_3
Euclidean
ϕ_0
ϕ_1
ϕ_2
ϕ_3
max
0
min
Manifold



0.2
0
−0.2
0
10
20
30
40
50
60
70
80
90
100
ϕ_0
ϕ_1
ϕ_2
ϕ_3
Euclidean
ϕ_0
ϕ_1
ϕ_2
ϕ_3
max
0
min

THE GEOMETRY OF MANIFOLDS AND GRAPHS

A square-integrable function f on $\mathcal{X}$ can be decomposed into *Fourier series* as

$$f(x) \;=\; \sum_{i \geq 0} \underbrace{\langle f, \phi_i \rangle_{L^2(\mathcal{X})}}_{\hat{f}_i} \phi_i(x), \tag{32}$$

analysis stage

synthesis stage

SPECTRAL METHODS

- **Spectral CNN (SCNN)**

$$\mathbf{g}_l = \xi \left(\sum_{l'=1}^{q} \mathbf{\Phi}_k \mathbf{\Gamma}_{l,l'} \mathbf{\Phi}_k^\top \mathbf{f}_{l'} \right)$$

$\mathbf{\Gamma}_{l,l'}$ is a $k \times k$ diagonal matrix of spectral multipliers

- Using only the first *k* eigenvectors in sets a cutoff frequency which depends on the intrinsic regularity of the graph and also the sample size.

SPECTRAL METHODS

- **Spectral CNN (SCNN)**
  - A fundamental limitation of the spectral construction is its limitation to a single domain
    - Basis dependent

Domain	$\mathcal{X}$	$\mathcal{X}$	$\mathcal{Y}$
Basis	$\mathbf{\Phi}$	$\mathbf{\Phi}$	$\mathbf{\Psi}$
Signal	$\mathbf{f}$	$\mathbf{\Phi W \Phi}^\top \mathbf{f}$	$\mathbf{\Psi W \Psi}^\top \mathbf{f}$

SPECTRAL METHODS

- **Spectral CNN (SCNN)**
  - A fundamental limitation of the spectral construction is its limitation to a single domain
    - Basis dependent
    - It is possible to construct compatible orthogonal bases **across different domains** resorting to a joint diagonalization procedure
      - Social graph
      - Computer graphics
  - Another limitation: too many parameters

$$pqk = O(n)$$

    - On Euclidean domains, this is achieved by learning convolutional kernels with small spatial support, which enables the model to learn a number of parameters independent of the input size.