Partial differential equations and graph-based learning

Jeff Calder
School of Mathematics
University of Minnesota

April 10, 2023
Applied Mathematics Colloquium, Columbia University

Research supported by NSF-DMS 1713691, 1944925, the Alfred P. Sloan Foundation, and the McKnight Foundation
Graph-based learning

Let \((\mathcal{X}, \mathcal{W})\) be a graph.

- Vertices \(\mathcal{X} \subset \mathbb{R}^d\).
- Nonnegative edge weights \(\mathcal{W} = (w_{xy})_{x,y \in \mathcal{X}}\).

Some common graph-based learning tasks:
1. Clustering
2. Semi-supervised learning
3. Data Depth
4. Link prediction
5. Ranking

Applications of graph-based learning:
1. Image classification
2. Social media networks
3. Biological networks
4. Drug discovery
5. Wireless networks

\[
\text{div}(\rho^2 \nabla u) = \lambda u
\]
Similarity graphs

MNIST: 70000 digits

- Each image is a datapoint
  \[ x \in \mathbb{R}^{28 \times 28} = \mathbb{R}^{784}. \]
- Geometric weights:
  \[ w_{xy} = \eta \left( \frac{|x - y|}{\varepsilon} \right) \]
  Often \( \eta(t) = e^{-t^2} \).
- \( k \)-nearest neighbor graph:
  \[ w_{xy} = \eta \left( \frac{|x - y|}{\varepsilon_k(x)} \right) \]
Similarity graphs via deep learning

Set \( w_{xy} = \eta \left( \frac{|\Psi(x) - \Psi(y)|}{\varepsilon} \right) \) where \( \Psi : \mathbb{R}^d \rightarrow \mathbb{R}^N \) is a learned feature map.

Synthetic Aperture Radar (SAR) Images


Graph distances and eikonal equations

Let $G$ be a connected graph on $\mathcal{X} = \{x_1, \ldots, x_n\}$ with edge weights $w_{ij} = w_{x_i x_j}$.

**Graph eikonal equation:**

$$
\begin{cases}
\max_{x_j \in \mathcal{X}} w_{ji} (u(x_i) - u(x_j)) = f(x_i), & \text{if } x_i \in \mathcal{X} \setminus \Gamma \\
\quad u(x_i) = 0, & \text{if } x_i \in \Gamma.
\end{cases}
$$

(1)

**Weighted graph distances:** We have

$$u(x) = d_{G,f}(x, \Gamma) := \min_{x_j \in \Gamma} d_{G,f}(x_i, x_j),$$

where

$$d_{G,f}(x_i, x_j) := \min_{m \geq 1} \sum_{\tau_1 = i, \tau_m = j}^{m-1} w_{\tau_i, \tau_{i+1}}^{-1} f(x_{\tau_{i+1}}).$$

(2)

It is common to choose $f = \hat{\rho}^{-\alpha}$, for some density estimation $\hat{\rho}$. 

Calder (UMN)
Prior work on graph distances

Applications of graph distances:

- Dimensionality reduction (e.g., ISOMAP) (Tenenbaum et al., 2000)
- Graph classification (Borgwardt and Kriegel, 2005)
- Data depth (Calder, Park and Slepcev, 2022) (Molina-Fructuoso and Murray, 2022)

Discrete to continuum:

- $k$-nn graphs (Alamgir and Von Luxburg, 2012)
- Geodesic manifold distance (Hwang, Damelin, and Hero, 2016)
- Geodesic distance on Euclidean domains (Bungert, Calder, and Roith, 2022)
Graph distances on point clouds

Figure: Plots of the solution to the graph eikonal equation for $n = 10^4$ for both the box and ball domains, and error plots for varying $\varepsilon$ averaged over 100 trials. The red points indicate the detected boundary points used in solving the PDE.
MNIST: Depth from eikonal equations

Lack of robustness to corrupted edges

Figure: From left to right we added an increasing number of corrupted edges (0, 10, 50, and 200) with edge weight $w_{ij} = 1$ (graph has 1M edges, so 200 edges is 0.02%).
The p-eikonal equation

For $p > 0$, we define the $p$-eikonal operator $A_{G,p} : F(\mathcal{X}) \to F(\mathcal{X})$ by

$$
(3) \quad A_{G,p}u(x_i) = \sum_{j=1}^{n} w_{ji}(u(x_i) - u(x_j))^p_+,
$$

where $a_+ := \max\{a, 0\}$ is the positive part. For $\Gamma \subset \mathcal{X}$ and $f : \mathcal{X} \to \mathbb{R}$, we consider the $p$-eikonal equation

$$
(4) \quad \begin{cases}
A_{G,p}u = f, & \text{in } \mathcal{X} \setminus \Gamma \\
 u = 0, & \text{on } \Gamma.
\end{cases}
$$

Note: When $p \to \infty$ we recover the graph eikonal equation and graph distance function.


Robustness

(a) Graph distance function with corrupted edges

(b) $p$-eikonal equation with $p = 1$ with corrupted edges
Robustness

Theorem (Calder & Ettehad, 2022)

Let $\delta W \in \mathbb{R}^{n \times n}$ such that $\tilde{W} := W + \delta W \geq 0$ and $\tilde{G} := (\mathcal{X}, \tilde{W})$ is connected. Let $\Gamma \subset \mathcal{X}$, $f \in F(\mathcal{X})$ with $f > 0$ and let $u, \tilde{u} \in F(\mathcal{X})$ satisfy

\[
\begin{aligned}
\mathcal{A}_{\tilde{G},p} \tilde{u}(x_i) &= \mathcal{A}_{G,p} u(x_i) = f(x_i), \quad \text{if } x_i \in \mathcal{X} \setminus \Gamma \\
\tilde{u}(x_i) &= u(x_i) = 0, \quad \text{if } x_i \in \Gamma.
\end{aligned}
\]

Then for all $x_i \in \mathcal{X}$ we have

\[
- \left( \max_{\mathcal{X} \setminus \Gamma} \frac{\mathcal{A}_{\delta G_- ,p} \tilde{u}}{f} \right)^{\frac{1}{p}} \leq \frac{u(x_i) - \tilde{u}(x_i)}{\min\{u(x_i), \tilde{u}(x_i)\}} \leq \left( \max_{\mathcal{X} \setminus \Gamma} \frac{\mathcal{A}_{\delta G_+ ,p} u}{f} \right)^{\frac{1}{p}},
\]

where $\delta G_{\pm} = (\mathcal{X}, \pm \delta W_{\pm})$.

- The theorem can be simplified to give the weaker bound

\[
\frac{u(x_i) - \tilde{u}(x_i)}{\min\{u(x_i), \tilde{u}(x_i)\}} \leq C \left( \frac{f_{\max}}{f_{\min}} \right)^{\frac{1}{p}} \|\delta W\|_{1}^{\frac{1}{p}}.
\]

Discrete to continuum

Let $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ be an i.i.d sample on $\Omega \subset \mathbb{R}^d$ with density $\rho$ and let

$$
\begin{cases}
\mathcal{A}_{n,\varepsilon} u_{n,\varepsilon}(x) = f(x) & \text{if } x \in \mathcal{X} \setminus \Gamma \\
u_{n,\varepsilon}(x) = 0 & \text{if } x \in \Gamma.
\end{cases}
$$

where $\Gamma \subset \Omega$ is a finite set of points and

$$
\mathcal{A}_{n,\varepsilon} u(x) = \frac{1}{n \sigma_p \varepsilon^{p+d}} \sum_{y \in \mathcal{X}} \eta \left( \frac{|x - y|}{\varepsilon} \right) (u(x) - u(y))^+.
$$

**Continuum limit**: State-constrained eikonal equation

$$
\begin{cases}
\rho |\nabla u|^p = f & \text{in } \Omega \setminus \Gamma \\
u = 0 & \text{on } \Gamma.
\end{cases}
$$

**Variational interpretation**: The solution $u$ is given by

$$
u(x) = d_g(x, \Gamma) := \min_{y \in \Gamma} d_g(x, y), \quad g = \rho^{-\frac{1}{p}} f^{\frac{1}{p}},$$

where

$$
d_g(x, y) := \inf \left\{ \int_0^1 g(\gamma(t))|\gamma'(t)| \, dt : \gamma \in C^1([0, 1]; \overline{\Omega}), \gamma(0) = x, \text{ and } \gamma(1) = y \right\}.
$$
Theorem (Calder & Ettehad, 2022)

If $\epsilon$ is sufficiently small then with probability at least $1 - 6n^2 \exp\left(-cn\epsilon^{d+1}\right)$ we have

$$\max_{x \in \mathcal{X}} |d_g(x, \Gamma) - u_{n,\epsilon}(x)| \leq C \left( \sqrt{\epsilon} + \left(n\epsilon^{p+d}\right)^{\frac{1}{p}} \right)$$

Main ideas in proof:

- Pointwise consistency $A_{n,\varepsilon} \varphi(x) \approx \rho |\nabla \varphi|^p$ for smooth $\varphi$, with high probability.
- The $O(\sqrt{\varepsilon})$ rate comes from a doubling variables argument in the viscosity solutions framework.
- Rate requires Lipschitzness of $u_{n,\varepsilon}$, we show that

$$|u_{n,\varepsilon}(x) - u_{n,\varepsilon}(y)| \leq c_p \gamma_p^{-1} \max_{x} f^{\frac{1}{p}} \| \Omega(x, y) \| + \gamma_p \left( n\varepsilon^{p+d} \right)^{\frac{1}{p}}, \quad \text{for all } x, y \in \mathcal{X}$$

with probability at least $1 - n^2 \exp \left( - \frac{c_d r^d}{2^{2d+3} \rho_{\min} n \varepsilon^d} \right)$. The proof uses a geodesic cone barrier function with an additional spike:

$$v_{\beta, y}(x) := \beta (1 - \delta_y(x)) + d_\Omega(x, y)$$

- State constrained boundary condition handled with domain perturbation results.
Back to the MNIST Median

Eikonal Median digits

$p$-eikonal Median digits ($p = 1$)
Data depth

Recall the **geometric median**:

\[ x_* \in \arg \min_{x \in \mathbb{R}^d} \sum_{i=1}^{n} |x_i - x|. \]

Generalizations to other metrics are called **Karcher means** or barycenters. For the \( p \)-eikonal equation we define

\[ x_{p,\alpha} \in \arg \min_{x \in \mathcal{X}} \sum_{x_i \in \mathcal{X}} d_{x}(x_i). \]

where

\[
\begin{align*}
AG,p d_x &= \hat{\rho}^{-\alpha}, \quad \text{in } \mathcal{X} \setminus \{x\} \\
\hat{d}_x(x) &= 0.
\end{align*}
\]

(7)

Then we can define data depth as the distance to the median

\[ \text{depth}_{p,\alpha}(x) = \max_{\mathcal{X}} d_{x_{p,\alpha}} - d_{x_{p,\alpha}}(x). \]
Data depth

Figure: The $p$-eikonal data depth on 3D toy datasets sampled from manifolds embedded in $\mathbb{R}^3$. We use $p = 1$ and $\alpha = 1$. 
Data depth

(a) Deepest images (median)

(b) Shallowest images (outliers)

Figure: Comparison of deepest (median) images to shallowest (outlier) images from each MNIST digit.
Data depth

Figure: Comparison of deepest (median) images to shallowest (outlier) images from each FashionMNIST class.
Data depth

Figure: Paths from shallowest point to median for each class.

Graph-based semi-supervised learning

**Given:** Graph \((\mathcal{X}, \mathcal{W})\), labeled nodes \(\Gamma \subset \mathcal{X}\), and labels \(g : \Gamma \to \mathbb{R}^k\).

**Task:** Extend the labels to the rest of the graph \(\mathcal{X} \setminus \Gamma\).

**Semi-supervised:** Goal is to use both the labeled and unlabeled data.
Graph-based semi-supervised learning

**Given:** Graph $(\mathcal{X}, \mathcal{W})$, labeled nodes $\Gamma \subset \mathcal{X}$, and labels $g : \Gamma \to \mathbb{R}^k$.

**Task:** Extend the labels to the rest of the graph $\mathcal{X} \setminus \Gamma$.

**Semi-supervised:** Goal is to use both the labeled and unlabeled data.
Graph-based semi-supervised learning

**Given:** Graph \( (\mathcal{X}, \mathcal{W}) \), labeled nodes \( \Gamma \subset \mathcal{X} \), and labels \( g : \Gamma \rightarrow \mathbb{R}^k \).

**Task:** Extend the labels to the rest of the graph \( \mathcal{X} \setminus \Gamma \).

**Semi-supervised:** Goal is to use both the labeled and unlabeled data.
Laplacian regularization

Laplacian regularized semi-supervised learning solves the Laplace equation

\[
\begin{aligned}
    \mathcal{L}u &= 0 \quad \text{in } \mathcal{X} \setminus \Gamma, \\
    u &= g \quad \text{on } \Gamma,
\end{aligned}
\]

where \( u : \mathcal{X} \rightarrow \mathbb{R}^k \), and \( \mathcal{L} \) is the graph Laplacian

\[
\mathcal{L}u(x) = \sum_{y \in \mathcal{X}} w_{xy}(u(x) - u(y)).
\]

The label decision for vertex \( x \in \mathcal{X} \) is determined by the largest component of \( u(x) \)

\[
\ell(x) = \operatorname{argmax}_{j \in \{1, \ldots, k\}} \{u_j(x)\}.
\]

Variational Interpretation:

\[
\min_{u : \mathcal{X} \rightarrow \mathbb{R}^k} \left\{ \sum_{x, y \in \mathcal{X}} w_{xy}|u(x) - u(y)|^2 : u(x) = g(x) \text{ for all } x \in \Gamma \right\}.
\]

Active learning

**Problem**: How to choose the best training data points for a particular task?

**Active learning** chooses the training data points in a sequential (often online) setting, using information from the classifier and unlabeled data.

**Goal** is to achieve good results with as few labeled examples as possible.
Acquisition functions

Graph-based active learning methods usually choose the next data point $x_{k+1}$ to label by minimizing (or maximizing) an acquisition function $A_k : \mathcal{X} \rightarrow \mathbb{R}$:

$$x_{k+1} = \arg \min_{x \in \mathcal{X} \setminus \Gamma_k} A_k(x) \quad \text{and} \quad \Gamma_{k+1} = \Gamma_k \cup \{x_{k+1}\}.$$

**Previous work:**

- **Uncertainty sampling:** $A_k(x)$ is the uncertainty of the classifier at node $x$.

- **(Ji & Han 2012):** Variance minimization (V-OPT): Acquisition function $A_k$ involves full inversion of $\mathcal{L}_{\Gamma_k^c \Gamma_k^c}$ (minimizes $\text{Trace}(\mathcal{L}_{\Gamma_k^c \Gamma_k^c}^{-1})$).

- **(Ma et al. 2013):** $\Sigma$-optimality: Similar to V-OPT but minimizes $1^T \mathcal{L}_{\Gamma_k^c \Gamma_k^c}^{-1} 1$.

- **(Dasarathy, Nowak, & Zhu, 2015):** $S^2$ (Shortest-shortest path)

- **(Murphy & Maggioni, 2019):** Learning by Active Non-linear Diffusion (LAND)


- **(Cloninger & Mhaskar, 2021):** Cautious Active Learning (CAL)
The exploration vs exploitation tradeoff
Continuum perspective

Let $x_1, x_2, \ldots, x_n$ be i.i.d random variables on $\Omega \subset \mathbb{R}^d$ with density $\rho$ and set

$$
\mathcal{L}_{n,\varepsilon} u(x) = \frac{1}{n\varepsilon^{d+2} \sigma_{\eta}} \sum_{j=1}^{n} \eta \left( \frac{|x-x_j|}{\varepsilon} \right) (u(x_j) - u(x)).
$$

Then we can compute (via concentration inequalities and Taylor expansion)

$$
\mathcal{L}_{n,\varepsilon} u(x) = \frac{1}{\varepsilon^{d+2} \sigma_{\eta}} \int_{B(x,\varepsilon)} \eta \left( \frac{1}{\varepsilon} |x - y| \right) (u(y) - u(x)) \rho(y) \, dy + O \left( \frac{\sqrt{\sigma^2}}{n} \right)
$$

$$
= \rho^{-1} \text{div}(\rho^2 \nabla u) + O \left( \varepsilon^2 + \sqrt{\frac{1}{n\varepsilon^{d+2}}} \right).
$$

Thus, the continuum limit for Laplace learning is

$$
\begin{align*}
\text{div}(\rho^2 \nabla u) &= 0, \quad \text{in } \Omega \setminus \Gamma \\
u &= g, \quad \text{on } \Gamma.
\end{align*}
$$

(8)

$$
\min_{u|\Gamma=g} \int_{\Omega} \rho^2 |\nabla u|^2 \, dx.
$$

This equation is ill-posed when $\Gamma$ contains isolated points.
Previous work

- Higher-order regularization: (Zhou and Belkin, 2011), (Dunlop et al., 2019)
- $p$-Laplace regularization: (Alaoui et al., 2016), (Calder 2018, 2019), (Slepcev & Thorpe 2019)
- Re-weighted Laplacians: (Shi et al., 2017), (Calder & Slepcev, 2020)
- Poisson learning (Calder et al., 2020, 2022)
Poisson Reweighted Laplace Learning (PWLL)

We recently developed Poisson ReWeighted Laplace Learning (PWLL) which solves

\[
\begin{align*}
\mathcal{L}_\gamma u &= 0 \quad \text{on } \mathcal{X} \setminus \Gamma, \\
\quad u &= g \quad \text{on } \Gamma,
\end{align*}
\]

where

\[
\mathcal{L}_\gamma = \sum_{y \in \Gamma} \left( \delta_y - \frac{1}{n} \right) \quad \text{on } \mathcal{X},
\]

and

\[
\mathcal{L}_\gamma u(x) = \sum_{y \in \mathcal{X}} \gamma(x) \gamma(y) w_{xy}(u(x) - u(y)).
\]

The continuum limit of PWLL should be the equations

\[
\begin{align*}
\text{div}(\rho^2 \gamma^2 \nabla u) &= 0, \quad \text{in } \Omega \setminus \Gamma \\
\quad u &= g \quad \text{on } \Gamma
\end{align*}
\]

Provided \( \gamma(x)^2 \sim \text{dist}(x, \Gamma)^{-\alpha} \) with \( \alpha > d - 2 \), then (9) is well-posed.
Uncertainty norm active learning

We propose uncertainty norm active learning which solves

\[
\begin{cases}
\tau u + \mathcal{L}_\gamma u = 0 & \text{on } \mathcal{X} \setminus \Gamma_k, \\
 u = g & \text{on } \Gamma_k,
\end{cases}
\]

and selects the next point \( x_k \) by minimizing the acquisition function

\[
A_k(x) = \|u(x)\|^2.
\]

Intuitively, the additional \( \tau u \) term localizes the solution around the labeled data points. In the 1D case \( \tau u - u'' = 0 \), the solution decays like \( e^{-\sqrt{\tau}x} \) away from labels.
Uncertainty norm active learning

(a) Clusters and Init. Labeled

(b) Ground Truth Classification

(c) $\tau = 0$
(d) $\tau = 10^{-9}$
(e) $\tau = 10^{-7}$
Uncertainty norm active learning

Uncertainty norm active learning uses the acquisition function $A(x) = \| u(x) \|_2^2$:

$$\tau u + \mathcal{L}_\gamma u = 0 \quad \iff \quad \tau u - \rho^{-1} \text{div}(\rho^2 \gamma^2 \nabla u) = 0.$$ 

Discrete $\iff$ Continuum

**Theorem (Miller & Calder, 2022)**

Let $\alpha > d - 2$. Given a clusterability assumption on $\rho$, for $\tau$ sufficiently large we have

1. **On any unexplored cluster $\mathcal{D}$ we have**

   $$\sup_{\mathcal{D}} A \leq \left( \sqrt{\# \text{Classes}} \right) \exp \left( -\frac{s}{4} \sqrt{\frac{\tau}{\delta}} \right),$$

   where $\delta = \max_{\partial \mathcal{D} + B_{2s}} \rho$. 

2. **For $r > 0$ sufficiently small:**

   $$\inf_{\Gamma + B_r} A \geq 1 - Cr^\beta,$$

   where $\beta = \frac{1}{2}(\alpha + 2 - d)$.

This is an **exploration guarantee**:

- When $\tau \gg 0$, uncertainty norm sampling will explore new clusters before selecting a point within $r$ of an existing labeled data point.

- The parameter $\tau$ controls the **exploration** ($\tau \gg 0$) vs **exploitation** ($\tau \ll 1$) tradeoff.

---

Active Learning Results

Ground Truth

8 Labels

100 Labels

Ground Truth

15 Labels

50 Labels

Calder (UMN)
PDEs and graphs

Columbia

34 / 39
Active learning results

**MNIST-mod3:**
- We group the classes modulo 3.
- Our method is Unc. (Norm) in blue and yellow.

---

Active learning results

**FashionMNIST-mod3:**
- We group the classes modulo 3.
- Our method is Unc. (Norm) in blue and yellow.

---

Active learning results

**EMNIST-mod5**: Extended MNIST with letters and numbers (47 classes)
- We group the classes modulo 5.
- Our method is Unc. (Norm) in blue and yellow.

---

Active learning results

ISOLET: Spoken letter dataset (audio)
- 26 classes, 7800 letters with 150 different speakers.
- Our method is Unc. (Norm) in blue and yellow.

---

Ji, Ming, and Jiawei Han. *A variance minimization criterion to active learning on graphs*. Artificial Intelligence and Statistics. PMLR, 2012.
Future work, papers, and code

**Future Work:**

1. *p*-eikonal equation: Manifold setting and applications (e.g., ISOMAP)
2. Poisson reweighted Laplace learning: Discrete to continuum, consistency, and clustering.
3. Active learning: Batch active learning.

**Papers:**


**Code:** All code uses the GraphLearning python package

   [https://github.com/jwcalder/GraphLearning](https://github.com/jwcalder/GraphLearning) (pip install graphlearning)