Discrete-continuum interplay: formulations for supervised and semi-supervised learning

Franca Hoffmann

Hausdorff Center for Mathematics
University of Bonn

July 28, 2021

LMS/ICMS Workshop
Analytic and Geometric Approaches to Machine Learning

Goals

- Graph Laplacians as tools for data analysis;
- Understanding parameter choices in
  - spectral clustering algorithms,
  - semi-supervised learning (SSL) algorithms;
- Continuum Limits of Graph Laplacians & their properties:
  - weighted elliptic operators (PDE theory),
  - insights on discrete algorithms,
  - new continuum algorithms;

collaboration with:
Bamdad Hosseini, Assad A. Oberai, Andrew M. Stuart (preprint 2020)
Bamdad Hosseini, Zhi Ren, Andrew M. Stuart (JMLR 2020)

Graph-Based Clustering

What is spectral clustering?

$X = \{x_1, \ldots, x_N\} \subset \Omega \subset \mathbb{R}^d$, W_{ij} =measure of similarity between x_i and x_j.

- **Input:** Similarity graph (X, W).
- **Output:** Clusters $A_1, \ldots, A_K$

Two steps of spectral clustering:

1. Embedding step $\mathcal{F}_N : X \to \mathbb{R}^K$.
2. Clustering step on $\mathcal{F}_N(x_1), \ldots, \mathcal{F}_N(x_N)$ (e.g. K-means)

Question: How to choose $\mathcal{F}_N$?

What is spectral clustering?

$X = \{x_1, \ldots, x_N\} \subset \Omega \subset \mathbb{R}^d$, W_{ij} =measure of similarity between x_i and x_j.

- **Input:** Similarity graph (X, W).
- **Output:** Clusters $A_1, \ldots, A_K$

Two steps of spectral clustering:

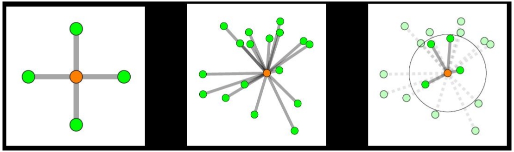
1. Embedding step $\mathcal{F}_N : X \to \mathbb{R}^K$.
2. Clustering step on $\mathcal{F}_N(x_1), \ldots, \mathcal{F}_N(x_N)$ (e.g. K-means)

Question: How to choose $\mathcal{F}_N$?

$\Rightarrow$ **Low-lying eigenfunctions of graph Laplacian:** $\mathcal{F}_N(x_i) = (u_1(x_i), ..., u_K(x_i))^T$

Graph Laplacians for Data Clustering

- N vertices $\{x_j\}_{j=1}^N \in \Omega \subset \mathbb{R}^d$.
- Suitable kernel $\eta : \mathbb{R}^d \mapsto \mathbb{R}$.
- Edge weights $\tilde{W}_{ij} = \eta\left(|x_i - x_j|\right)$.
- Degree matrix $\tilde{D} = \mathrm{diag}(\tilde{d}_i)$, $\tilde{d}_i = \sum_j \tilde{W}_{ij}$.
- Reweighted similarity matrix: $W_{ij} := \tilde{W}_{ij} / \left(\tilde{d}_i^\alpha \tilde{d}_j^\alpha\right)$.
- Reweighted degrees: $D = \mathrm{diag}(d_i)$.



Graph Laplacian

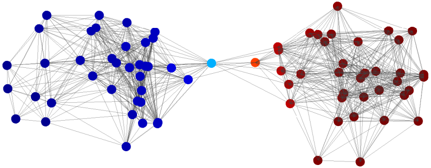
For $s, t \in \mathbb{R}$,

$$L := D^{-s}\left(D - W\right) D^{-t}$$

Dirichlet energy ($s = 0, t = 0$):

$$\langle \mathbf{u}, L\mathbf{u} \rangle = \frac{1}{2} \sum_{i,j} W_{ij} \left|u_i - u_j\right|^2 .$$

Fiedler vector



Graph Laplacians for Data Clustering

Graph Laplacian

For $s, t \in \mathbb{R}$,

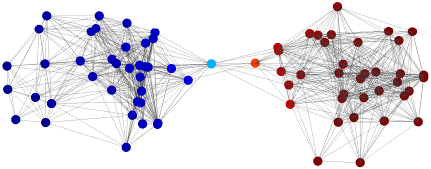
$$L := D^{-s}\,(D - W)\,D^{-t}$$

Dirichlet energy:

$$\langle \mathbf{u}, L\mathbf{u} \rangle_{(s,t)} = \langle D^{s-t}\mathbf{u}, L\mathbf{u} \rangle$$

$$= \frac{1}{2} \sum_{i,j} W_{ij} \left| \frac{u_i}{d_i^t} - \frac{u_j}{d_j^t} \right|^2$$

Fiedler vector

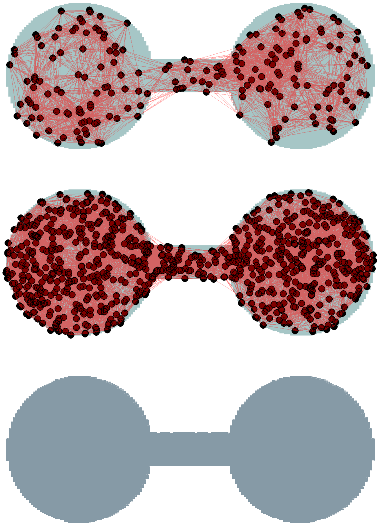


If X has K disconnected components:

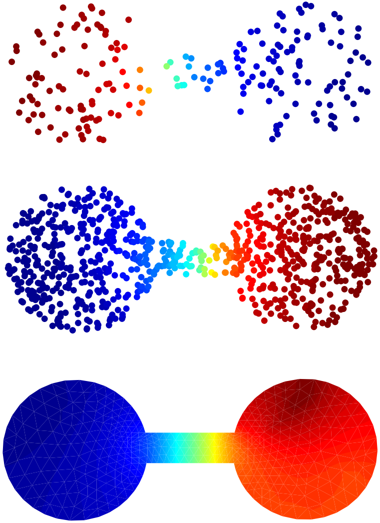
- Eigenvalues: $0 = \lambda_1^N = \ldots = \lambda_K^N < \lambda_{K+1}^N \leq \ldots \leq \lambda_N^N$
- Eigenvectors: $u_{1,N}, \ldots, u_{K,N}$ proportional to $D^{-t}\mathbb{1}$ on components.

Continuum Limits of Graph Laplacians

Ω and G



Fiedler vector



L

N
$\downarrow$
∞

$\mathcal{L}$

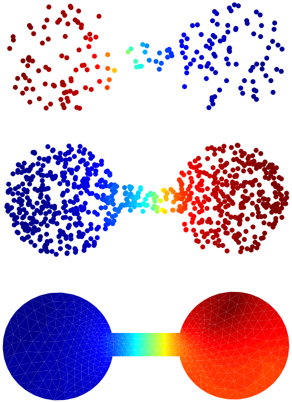
Continuum Limit of Graph Laplacians

- Vertices $x_j \overset{iid}{\sim} \rho$.
- Graph Laplacian:

$$L = D^{-s}(D - W)D^{-t}.$$

- Weighted Elliptic Operator:

$$\mathcal{L} : u \mapsto -\frac{1}{\rho^p}\operatorname{div}\left(\rho^q \nabla\left(\frac{u}{\rho^r}\right)\right) \text{ on } \Omega,$$
$$\rho^q \frac{\partial}{\partial n}\left(\frac{u}{\rho^r}\right) = 0 \text{ on } \partial\Omega.$$



$\Longrightarrow$ **Goal:** Explore properties of $\mathcal{L}$ for continuum data clustering and classification algorithms.

Continuum Limit of Graph Laplacians

- $\{x_j\}_{j=1}^N$ i.i.d. from density ρ on $\Omega \subset \mathbb{R}^d$.
- $\tilde{W}_{ij} = \eta_\delta(|x_i - x_j|), \quad \eta_\delta = \frac{1}{\delta^d}\eta\left(\frac{|\cdot|}{\delta}\right)$.
- Graph Laplacian:

$$L = D^{-s}(D - W)D^{-t}, \qquad W = \tilde{D}^{-\alpha}\tilde{W}\tilde{D}^{-\alpha}.$$

- Weighted Elliptic Operator:

$$\mathcal{L} : u \mapsto -\frac{1}{\rho^p}\mathrm{div}\left(\rho^q \nabla\left(\frac{u}{\rho^r}\right)\right).$$

Theorem

The new family of operators $\mathcal{L}$ arises from L in the limit $N \to \infty$, $\delta \to 0$ with

$$s = \frac{p-1}{q-1}, \quad t = \frac{r}{q-1}, \quad \alpha = 1 - q/2.$$

[García Trillos, Slepčev 2016 (ACHA)], [H., Hosseini, Oberai, Stuart (preprint)]

Sketch Proof: Limits of Quadratic Forms on Graphs

Limiting Discrete Dirichlet Energy $(p, q, r) = (1, 2, 0) \Leftrightarrow (s, t, \alpha) = (0, 0, 0)$

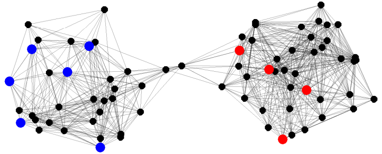
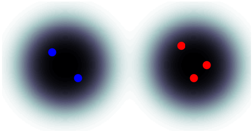
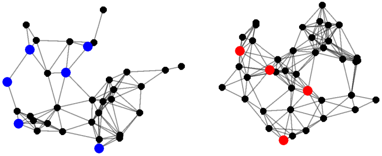
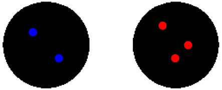
$$\langle \mathbf{u}, L\mathbf{u} \rangle \propto \frac{1}{N^2\delta^2} \sum_{j \sim k} \eta_\delta\big(x_j - x_k\big)\big|u(x_j) - u(x_k)\big|^2;$$

$$N \to \infty \approx \int_\Omega \int_\Omega \eta_\delta\big(x - y\big)\left|\frac{u(x) - u(y)}{\delta}\right|^2 \rho(x)\rho(y)dxdy;$$

$$\delta \to 0 \approx C(\eta) \int_\Omega |\nabla u(x)|^2 \rho(x)^2 dx \propto \langle u, \mathcal{L}u \rangle_{L^2_\rho}.$$

Perturbation Analysis

- Perturbed operators: $\mathcal{L}_\varepsilon = -\frac{1}{\rho_\varepsilon^p}\operatorname{div}\left(\rho_\varepsilon^q \nabla\left(\frac{u}{\rho_\varepsilon^r}\right)\right)$

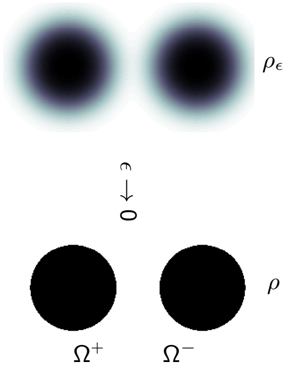
	Discrete: Perturbation of W	Continuum: Perturbation of ρ	
G_ε			ρ_ε
	$\varepsilon \to 0$	$\varepsilon \to 0$	
G			ρ

[H., Hosseini, Ren, Stuart 2020 (JMLR)]

[H., Hosseini, Oberai, Stuart (preprint)]

Spectrum of $\mathcal{L}_\varepsilon$: Two Clusters

$$\mathcal{L}_\varepsilon = -\frac{1}{\rho_\varepsilon^p}\mathrm{div}\left(\rho_\varepsilon^q \nabla\left(\frac{u}{\rho_\varepsilon^r}\right)\right)$$



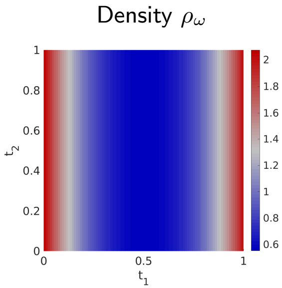
Theorem

If $p + r > 0$ $q > 0$ and $\epsilon \ll 1$, then

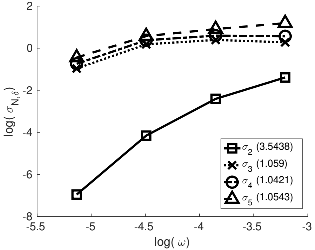
- $\lambda_{1,\varepsilon} = 0$
- $\lambda_{2,\varepsilon} \asymp \varepsilon^q$
- If $q > p + r$: $\lambda_{3,\varepsilon} \gtrsim \varepsilon^{2(q-p-r)}$
  If $q = p + r$: $\lambda_{3,\varepsilon} \geq \Lambda > 0$ (uniform spectral gap!)
  If $q < p + r$: $\lambda_{3,\varepsilon} \gtrsim \varepsilon^{p+r-q}$
- $\mathrm{span}\{\phi_{1,\varepsilon}, \phi_{2,\varepsilon}\} \approx \mathrm{span}\{\rho_\varepsilon^r \mathbf{1}_{\Omega^+}, \rho_\varepsilon^r \mathbf{1}_{\Omega^-}\}$.

[H., Hosseini, Oberai, Stuart (preprint)]

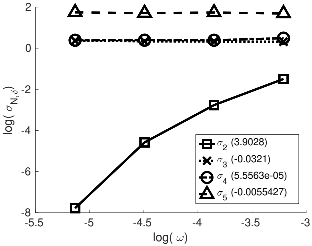
Numerical Illustration: Spectrum of graph Laplacian L



$q > p + r$ (1/2,2,1/2)



$q = p + r$ (1,2,1)



Multiple Clusters

Conjecture

If the data density ϱ_ε concentrates on $K \geq 2$ clusters as $\epsilon \to 0$, then

$$\sigma_{K,\varepsilon} \asymp \varepsilon^q, \qquad \frac{\sigma_{K,\varepsilon}}{\sigma_{K+1,\varepsilon}} \asymp \varepsilon^{\min\{q,p+r\}}.$$

Multiple Clusters

Conjecture

If the data density ϱ_ε concentrates on $K \geq 2$ clusters as $\epsilon \to 0$, then

$$\sigma_{K,\varepsilon} \asymp \varepsilon^q, \qquad \frac{\sigma_{K,\varepsilon}}{\sigma_{K+1,\varepsilon}} \asymp \varepsilon^{\min\{q,p+r\}}.$$

Theorem ($K = 2$)

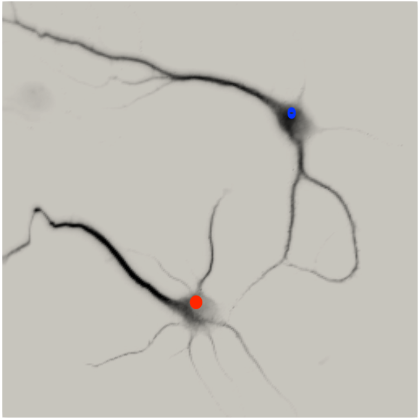
If $p + r > 0$ $q > 0$ and $\epsilon \ll 1$, then

- $\lambda_{1,\varepsilon} = 0$
- $\lambda_{2,\varepsilon} \asymp \varepsilon^q$
- If $q > p + r$: $\lambda_{3,\varepsilon} \sim \varepsilon^{q-(p+r)}$
  If $q = p + r$: $\lambda_{3,\varepsilon} \geq \Lambda > 0$ (uniform spectral gap!)
  If $q < p + r$: $\lambda_{3,\varepsilon} \geq \Lambda > 0$ (uniform spectral gap!)
- $\mathrm{span}\{\phi_{1,\varepsilon}, \phi_{2,\varepsilon}\} \approx \mathrm{span}\{\rho_\varepsilon^r \mathbf{1}_{\Omega^+}, \rho_\varepsilon^r \mathbf{1}_{\Omega^-}\}$.

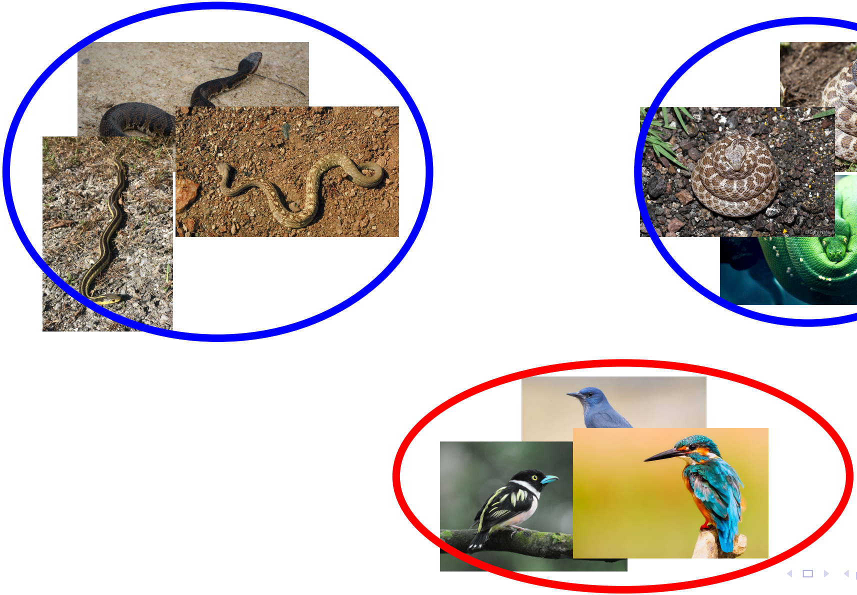
Data Classification: Semi-Supervised Learning

Adding Label Information: Image Segmentation

- Grayscale image ρ.
- Small number of labelled pixels.
- Segment the image consistently.



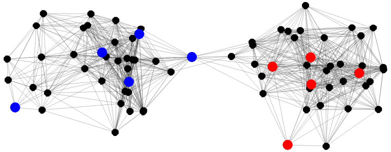
Clustering vs. Semi-Supervised Learning



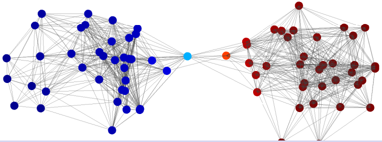
Key Idea

(spectral geometric content) $+$ (observed labels) $\rightarrow$ find all labels.

Labels



Fiedler vector



Binary Classification

- Labels $\{y_1, \ldots, y_J\} \in \{-1, +1\}$.
- $J \leq N$ number of observed labels.

Inverse Problem

Model

Given

- graph G,
- observed labels $y_1, \ldots, y_J \in \{-1, +1\}$, $J < N$,

find **ground truth** $\{u_1^\dagger, \cdots, u_N^\dagger\} \in \mathbb{R}$ s.t.

$$y_j = \mathrm{sgn}(u_j^\dagger + \eta_j), \qquad \eta_j \overset{iid}{\sim} \psi_\gamma, \qquad j \in \{1, \cdots, J\}.$$

- $\gamma > 0$ standard deviation of observation noise.
- Severely ill-posed inverse problem.

Convex Relaxation of Binary Semi-Supervised Learning

Probit optimization problem

Given graph G, labels $\mathbf{y}$, likelihood potential Φ_γ, parameters $\tau \in \mathbb{R}$, $\beta > 0$, find

$$\mathbf{u}^* = \operatorname{argmin}_{\mathbf{u}\in\mathbb{R}^N} \underbrace{\frac{1}{2}\langle \mathbf{u}, \tau^{-2\beta}(L + \tau^2 I_N)^\beta \mathbf{u}\rangle}_{\text{Convex regularization}} + \underbrace{\Phi_\gamma(\mathbf{u};\mathbf{y})}_{\text{Misfit}}.$$

- **Bayesian formulation**: connection between probability and optimization

$$\mathbb{P}(\mathbf{u}|\mathbf{y}) \propto \mathbb{P}(\mathbf{u}) \times \mathbb{P}(\mathbf{y}|\mathbf{u}) \propto \mathrm{N}(0, C) \times \exp\big(-\Phi_\gamma(\mathbf{u};\mathbf{y})\big)$$
$$\propto \exp\left(-\left[\frac{1}{2}\langle \mathbf{u}, C^{-1}\mathbf{u}\rangle + \Phi_\gamma(\mathbf{u};\mathbf{y})\right]\right)$$

Probit Likelihood Potential Φ_γ

$$y_j = \mathrm{sgn}(u_j + \eta_j), \quad \eta_j \overset{iid}{\sim} \psi_\gamma, \quad \forall j \in \{1, \cdots, J\}.$$

Likelihood

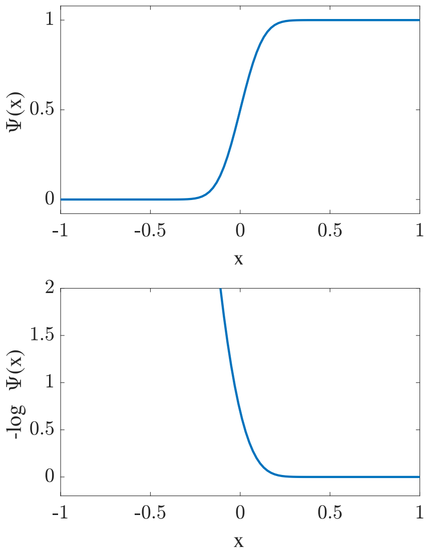
$$\mathbb{P}(\mathbf{y}|\mathbf{u}) \propto \Pi_{j=1}^{J} \Psi_\gamma(u_j y_j).$$

- Symmetric log-concave density ψ_γ on $\mathbb{R}$.
- Ψ_γ = CDF of ψ_γ.

Misfit: Negative Log-Likelihood

$$\Phi_\gamma(\mathbf{u}; \mathbf{y}) = -\sum_{j=1}^{J} \log \Psi_\gamma(u_j y_j).$$

[Rasmussen, Williams 2006],
[Bertozzi, Luo, Stuart, Zygalakis 2018]



Probit optimization problem

Given graph G, labels $\mathbf{y}$, likelihood potential Φ_γ, parameters $\tau \in \mathbb{R}$, $\beta > 0$, find

$$\mathbf{u}^* = \operatorname{argmin}_{\mathbf{u}\in\mathbb{R}^N} \underbrace{\frac{1}{2}\langle \mathbf{u}, \tau^{-2\beta}(L+\tau^2 I_N)^\beta \mathbf{u}\rangle}_{\text{Convex regularization}} + \underbrace{\Phi_\gamma(\mathbf{u};\mathbf{y})}_{\text{Misfit}}.$$

Theorem

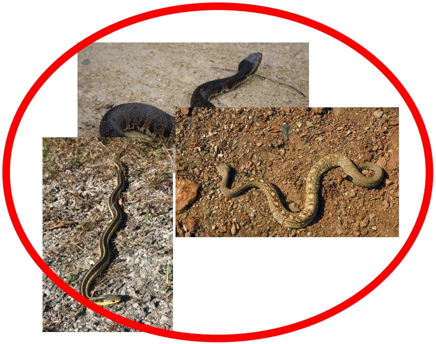
Existence and uniqueness of the probit minimizer $\mathbf{u}^*$.

- Asymptotic consistency: as $\gamma \downarrow 0$,

$$\operatorname{sgn}(u_j^*) \to \operatorname{sgn}(u_j^\dagger), \qquad \forall j \in \{1, \cdots, N\},$$

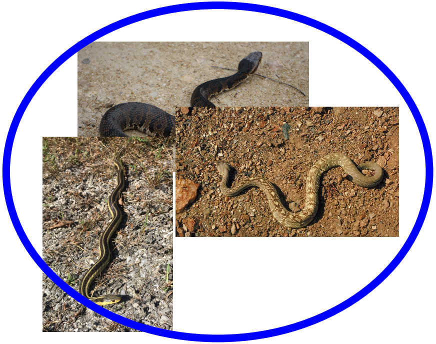
  in a suitable sense [H., Hosseini, Ren, Stuart 2020 (JMLR)]
- Also common to use probit with $\tau = 0$, $\beta = 1$ and constrain $\mathbf{u} \perp \mathbb{1}$. [Bertozzi, Luo, Stuart, Zygalakis 2018]
- **We do not constrain $\mathbf{u} \perp \mathbb{1}$.**

Modelling Assumptions



$$\tau = 0, \qquad \mathbf{u} \perp \mathbb{1}.$$

Modeling Assumptions



$$\tau > 0, \qquad \mathbf{u} \in \mathbb{R}^N.$$

($N < \infty$) Discrete probit on N vertices $X \in \mathbb{R}^{d \times N}$:

- Ground truth function $\mathbf{u}^\dagger \in \mathbb{R}^N$.
- Data $y_j = \mathrm{sgn}(u_j^\dagger + \gamma \eta_j), \qquad j \in \{1, \cdots, J\}$.
- Recover sign of $\mathbf{u}^\dagger$ by solving

$$\mathbf{u}^* = \mathrm{argmin}_{\mathbf{u} \in \mathbb{R}^N} \frac{1}{2} \langle \mathbf{u}, \tau^{-2\beta} (L + \tau^2 I)^\beta \mathbf{u} \rangle - \sum_{j=1}^J \log \Psi_\gamma (u_j y_j)$$

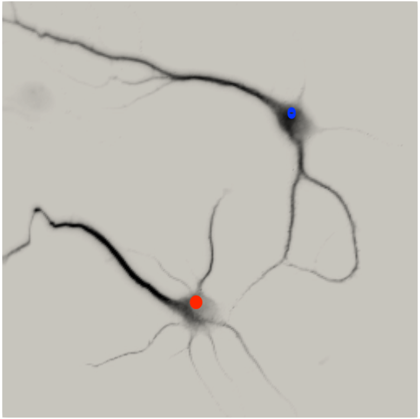
($N = \infty$) Continuum probit on $\Omega \subset \mathbb{R}^d$:

- Probability density ρ on Ω.
- Ground truth function $u^\dagger : \Omega \mapsto \mathbb{R}$.
- Fixed observed points $\{x_j\}_{j=1}^J \in \Omega$.
- Observed data $y_j = \mathrm{sgn}(u^\dagger(x_j) + \gamma \eta_j), \quad j \in \{1, \cdots, J\}$.
- Recover sign of $u^\dagger$ by solving

$$u^* = \mathrm{argmin}_{u \in \mathcal{H}^\beta(\Omega)} \frac{1}{2} \langle u, \tau^{-2\beta} (\mathcal{L} + \tau^2 I)^\beta u \rangle_\rho - \sum_{j=1}^J \log \Psi_\gamma (u(x_j) y_j)$$

An Application In Image Segmentation

- Grayscale image ρ.
- Small number of labelled pixels.
- Segment the image consistently.

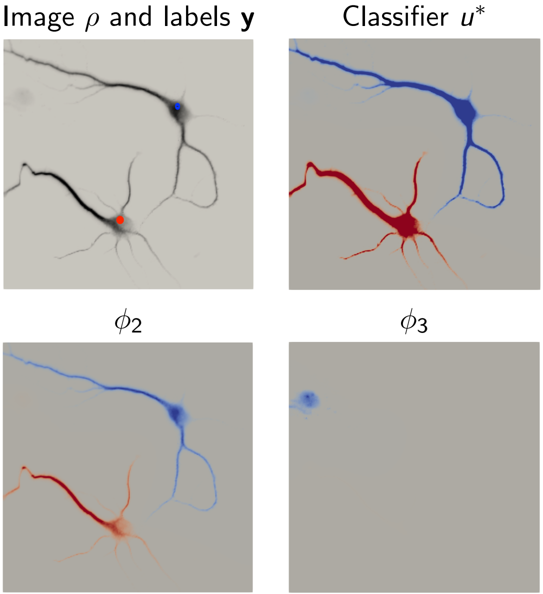


An Application In Image Segmentation

- Laplacian parameters $(p, q, r) = (1, 2, 1)$

$$\mathcal{L}u = -\frac{1}{\rho}\text{div}\left(\rho^2\nabla\left(\frac{u}{\rho}\right)\right)$$

- Solve continuum probit with eigenvalue problem solver in FEniCS.
- Probit minimizer u^* segments image.



QUESTIONS!

Discussion

Questions for you:

- How to leverage continuum formulations for algorithm design and implementations?
- How to evaluate and compare these implementations?