Large Graph Limits of Learning Algorithms

Andrew M Stuart

Computing and Mathematical Sciences, Caltech

Andrea Bertozzi, Michael Luo (UCLA)
and Kostas Zygalakis (Edinburgh)
⋆
Matt Dunlop (Caltech), Dejan Slepčev (CMU)
and Matt Thorpe (CMU)

References

- X Zhu, Z Ghahramani, J Lafferty, *Semi-supervised learning using Gaussian fields and harmonic functions*, ICML, 2003. Harmonic Functions.
- C Rasmussen and C Williams, *Gaussian processes for machine learning*, MIT Press, 2006. Probit.
- AL Bertozzi and A Flenner, *Diffuse interface models on graphs for classification of high dimensional data*, SIAM MMS, 2012. Ginzburg-Landau.
- MA Iglesias, Y Lu, AM Stuart, *Bayesian level set method for geometric inverse problems*, Interfaces and Free Boundaries, 2016. Level Set.
- AL Bertozzi, M Luo, AM Stuart and K Zygalakis, *Uncertainty quantification in the classification of high dimensional data*, `https://arxiv.org/abs/1703.08816`, 2017. Probit on a graph.
- N Garcia-Trillos and D Slepčev, *A variational approach to the consistency of spectral clustering*, ACHA, 2017.
- M Dunlop, D Slepčev, AM Stuart and M Thorpe, *Large data and zero noise limits of graph based semi-supervised learning algorithms*, In preparation, 2017.
- N Garcia-Trillos, D Sanz-Alonso, *Continuum Limit of Posteriors in Graph Bayesian Inverse Problems*, `https://arxiv.org/abs/1706.07193`, 2017.

Talk Overview

Learning and Inverse Problems

Optimization

Theoretical Properties

Probability

Conclusions

Talk Overview

Learning and Inverse Problems

Optimization

Theoretical Properties

Probability

Conclusions

Regression

- Let $D \subset \mathbb{R}^d$ be a bounded open set.
- Let $D' \subset D$.

Ill-Posed Inverse Problem

Find $u : D \mapsto \mathbb{R}$ given

$$y(x) = u(x), \quad x \in D'.$$

- Strong prior information needed.

Classification

- Let $D \subset \mathbb{R}^d$ be a bounded open set.
- Let $D' \subset D$.

Ill-Posed Inverse Problem

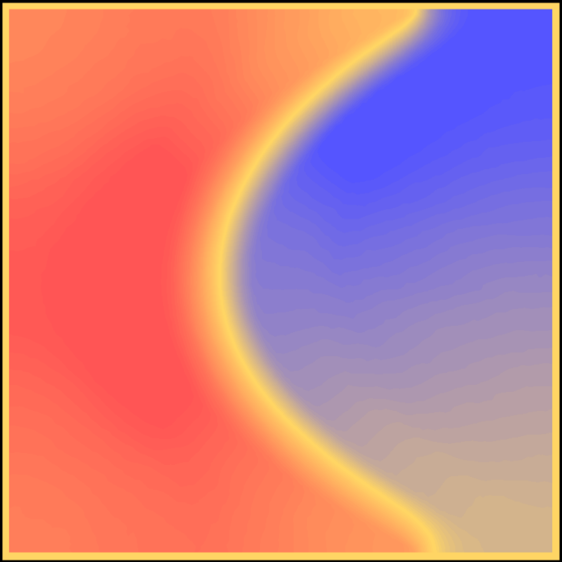
Find $u : D \mapsto \mathbb{R}$ given

$$y(x) = \mathrm{sign}\big(u(x)\big), \quad x \in D'.$$

- Strong prior information needed.

$y = \text{sign}(u)$. Red$= 1$. Blue$= -1$. Yellow: no information.

Reconstruction of the function u on D



Talk Overview

Learning and Inverse Problems

Optimization

Theoretical Properties

Probability

Conclusions

Graph Laplacian

Graph Laplacian:

- Similarity graph G with n vertices $Z = \{1, \ldots, n\}$.
- Weighted adjacency matrix $W = \{w_{j,k}\}$, $\left(w_{j,k} = \eta_\varepsilon(x_j - x_k).\right)$
- Diagonal $D = \text{diag}\{d_{jj}\}$, $d_{jj} = \sum_{k \in Z} w_{j,k}$.
- $L = s_n(D - W)$ (unnormalized); $L' = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$ (normalized).

Spectral Properties:

- L is positive semi-definite: $\langle u, Lu \rangle_{\mathbb{R}^n} \propto \sum_{j \sim k} w_{j,k} |u_j - u_k|^2$.
- $Lq_j = \lambda_j q_j$;
- Fully connected $\Rightarrow \lambda_1 > \lambda_0 = 0$. Fiedler Vector: q_1.

Problem Statement (Optimization)

Semi-Supervised Learning

- **Input**:
  - Unlabelled data $\{x_j \in \mathbb{R}^d, \quad j \in Z := \{1, \dots, n\}\}$;
  - Labelled data $\{y_j \in \{\pm 1\}, \quad j \in Z' \subseteq Z\}$.
- **Output**:
  - Labels $\{y_j \in \{\pm 1\}, \quad j \in Z\}$.

Classification based on $\mathrm{sign}(u)$, u the optimizer of:

$$J(u; y) = \frac{1}{2}\langle u, C^{-1}u\rangle_{\mathbb{R}^n} + \Phi(u; y).$$

- u is an $\mathbb{R}-$valued function on the graph nodes.
- $C = (L + \tau^2 I)^{-\alpha}$ $\Big($from unlabelled data: $w_{j,k} = \eta_\varepsilon(x_j - x_k)$.$\Big)$
- $\Phi(u; y)$ links real-valued u to the binary-valued labels y.

Example: Voting Records

U.S. House of Representatives 1984, 16 key votes. For each congress representative we have an associated feature vector $x_j \in \mathbb{R}^{16}$ such as

$$x_j = (1, -1, 0, \cdots, 1)^T;$$

1 is "yes", -1 is "no" and 0 abstain/no-show. Hence $d = 16$ and $n = 435$.

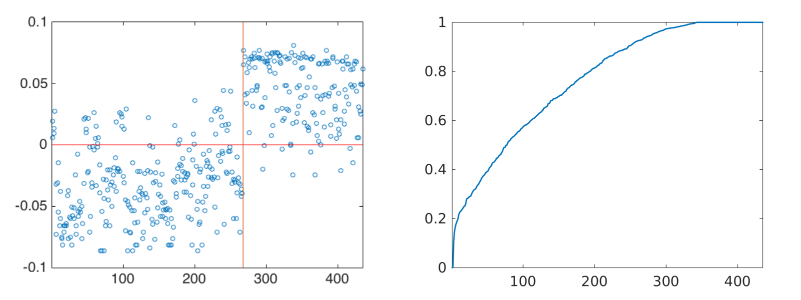


Figure: Fiedler Vector and Spectrum (Normalized Case)

Probit

Rasmussen and Williams, 2006. (MIT Press)

Bertozzi, Luo, Stuart and Zygalakis, 2017. (arXiv)

Probit Model

$$\mathsf{J}_{\mathrm{p}}^{(n)}(u;y) = \frac{1}{2}\langle u, C^{-1}u\rangle_{\mathbb{R}^n} + \Phi_{\mathrm{p}}^{(n)}(u;y).$$

Here

$$C = (L + \tau^2 I)^{-\alpha},$$

$$\Phi_{\mathrm{p}}^{(n)}(u;y) := -\sum_{j\in Z'} \log\big(\Psi(y_j u_j \ ;\gamma)\big)$$

and

$$\Psi(v;\gamma) = \frac{1}{\sqrt{2\pi\gamma^2}} \int_{-\infty}^{v} \exp\big(-t^2/2\gamma^2\big)\,dt.$$

Level Set

Iglesias, Lu and Stuart, 2016. (IFB)

Level Set Model

$$\mathsf{J}_{\mathrm{ls}}^{(n)}(u;y) = \frac{1}{2}\langle u, C^{-1}u\rangle_{\mathbb{R}^n} + \Phi_{\mathrm{ls}}^{(n)}(u;y).$$

Here

$$C = (L + \tau^2 I)^{-\alpha},$$

and

$$\Phi_{\mathrm{ls}}^{(n)}(u;y) := \frac{1}{2\gamma^2}\sum_{j\in Z'}\left|y_j - \mathrm{sign}\left(u_j\right)\right|^2.$$

Talk Overview

Learning and Inverse Problems

Optimization

Theoretical Properties

Probability

Conclusions

Infimization

Recall that both optimization problems have the form

$$\mathsf{J}^{(n)}(u;y) = \frac{1}{2}\langle u, C^{-1}u\rangle_{\mathbb{R}^n} + \Phi^{(n)}(u;y).$$

Indeed:

$$\Phi_{\mathrm{p}}^{(n)}(u;y) := -\sum_{j\in Z'} \log\big(\Psi(y_j\, u_j\ ;\gamma)\big)$$

and

$$\Phi_{\mathrm{ls}}^{(n)}(u;y) := \frac{1}{2\gamma^2}\sum_{j\in Z'} |y_j - \mathrm{sign}\big(u_j\big)|^2.$$

Theorem 1

- Probit: J_{p} is convex.
- Level Set: J_{ls} does not attain its infimum.

Limit Theorem for the Dirichlet Energy

Garcia-Trillos and Slepčev, 2016. (ACHA)

Unlabelled data $\{x_j\}$ sampled i.i.d. from density ρ supported on bounded $D \subset \mathbb{R}^d$. Let

$$\mathcal{L}u = -\frac{1}{\rho}\nabla \cdot \left(\rho^2 \nabla u\right) \quad x \in D, \quad \frac{\partial u}{\partial n} = 0, \quad x \in \partial D.$$

Theorem 2

Let $s_n = \frac{2}{C(\eta)n\varepsilon^2}$. Then under connectivity conditions on $\varepsilon = \varepsilon(n)$ in η_ε, the scaled Dirichlet energy $\Gamma-$ converges in the TL^2 metric:

$$\frac{1}{n}\langle u, Lu\rangle_{\mathbb{R}^n} \to \langle u, \mathcal{L}u\rangle_{L^2_\rho} \quad \text{as} \quad n \to \infty.$$

Sketch Proof: Quadratic Forms on Graphs

Discrete Dirichlet Energy

$$\langle u, Lu \rangle_{\mathbb{R}^n} \propto \sum_{j \sim k} w_{j,k} |u_j - u_k|^2.$$

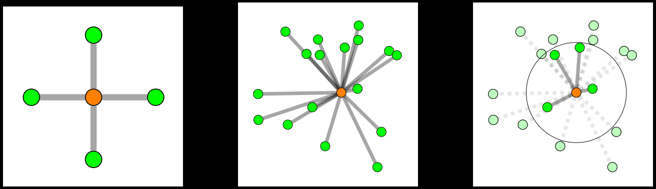


Figure: Connectivity Stencils For Orange Node: PDE, Data, Localized Data.

Sketch Proof: Limits of Quadratic Forms on Graphs

Garcia-Trillos and Slepčev, 2016. (ACHA)

- $\{x_j\}_{j=1}^n$ i.i.d. from density ρ on $D \subset \mathbb{R}^d$.
- $w_{jk} = \eta_\varepsilon(x_j - x_k), \quad \eta_\varepsilon = \frac{1}{\varepsilon^d}\eta\left(\frac{|\cdot|}{\varepsilon}\right)$.

Limiting Discrete Dirichlet Energy

$$\langle u, Lu\rangle_{\mathbb{R}^n} \propto \frac{1}{n^2\varepsilon^2}\sum_{j\sim k}\eta_\varepsilon\left(x_j - x_k\right)\left|u(x_j) - u(x_k)\right|^2;$$

$$n \to \infty \approx \int_D\int_D \eta_\varepsilon\left(x - y\right)\left|\frac{u(x) - u(y)}{\varepsilon}\right|^2\rho(x)\rho(y)dxdy;$$

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

Limit Theorem for Probit

M. Dunlop, D Slepčev, AM Stuart and M Thorpe, In preparation 2017.

Let $D^{\pm}$ be two disjoint bounded subsets of D, define $D' = D^{+} \cup D^{-}$ and

$$y(x) = +1,\ x \in D^{+}; \quad y(x) = -1,\ x \in D^{-}.$$

For $\alpha > 0$, define $\mathcal{C} = (\mathcal{L} + \tau^2 I)^{-\alpha}$. Recall that $C = (L + \tau^2 I)^{-\alpha}$.

Theorem 3

Let $s_n = \frac{2}{C(\eta) n \varepsilon^2}$. Then under connectivity conditions on $\varepsilon = \varepsilon(n)$ the scaled probit objective function $\Gamma-$converges in the TL^2 metric:

$$\frac{1}{n} \mathsf{J}_{\mathrm{p}}^{(n)}(u; y) \to \mathsf{J}_{\mathrm{p}}(u; y) \quad \text{as} \quad n \to \infty,$$

where

$$\mathsf{J}_{\mathrm{p}}(u; y) = \frac{1}{2} \langle u, \mathcal{C}^{-1} u \rangle_{L^2_\rho} + \Phi_{\mathrm{p}}(u; y),$$

$$\Phi_{\mathrm{p}}(u; y) := -\int_{D'} \log\Big(\Psi\big(y(x)\, u(x)\ ; \gamma\big)\Big) \rho(x) dx.$$

Talk Overview

Learning and Inverse Problems

Optimization

Theoretical Properties

Probability

Conclusions

Problem Statement (Bayesian Formulation)

Semi-Supervised Learning

- **Input**:
  - Unlabelled data $\{x_j \in \mathbb{R}^d, \quad j \in Z := \{1, \ldots, n\}\}$; **prior**
  - Labelled data $\{y_j) \in \{\pm 1\}, \quad j \in Z' \subseteq Z\}$. **likelihood**
- **Output**:
  - Labels $\{y_j \in \{\pm 1\}, \quad j \in Z\}$. **posterior**

Connection between probability and optimization:

$$J^{(n)}(u; y) = \frac{1}{2}\langle u, C^{-1}u\rangle_{\mathbb{R}^n} + \Phi^{(n)}(u; y).$$

$$\begin{aligned}
\mathbb{P}(u|y) &\propto \exp\big(-J^{(n)}(u; y)\big) \\
&\propto \exp\big(-\Phi^{(n)}(u; y)\big) \times \mathsf{N}(0, C) \\
&\propto \mathbb{P}(y|u) \times \mathbb{P}(u).
\end{aligned}$$

Example of Underlying Gaussian (Voting Records)

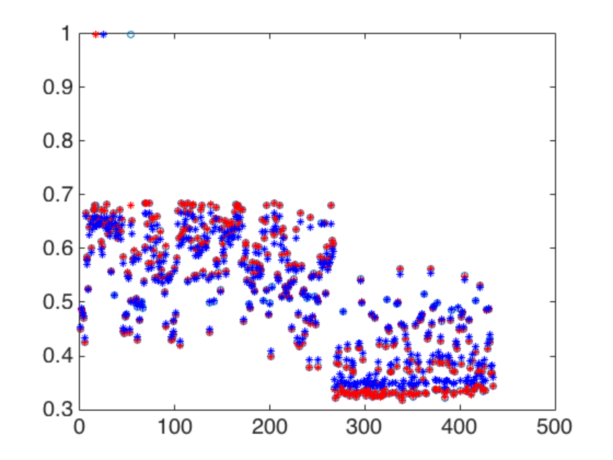


Figure: Two point correlation of sign(u) for 3 democrats

Probit (Continuum Limit)

Let $\alpha > \frac{d}{2}$.

Probit Probabilistic Model

- Prior: Gaussian $\mathbb{P}(du) = \mathsf{N}(0, \mathcal{C})$.
- Posterior: $\mathbb{P}_\gamma(du|y) \propto \exp\big(-\Phi_\mathrm{p}(u;y)\big)\mathbb{P}(du)$.

$$\Phi_\mathrm{p}(u;y) := -\int_{D'} \log\Big(\Psi\big(y(x)\,u(x)\ ;\gamma\big)\Big)\rho(x)dx.$$

Level Set (Continuum Limit)

Let $\alpha > \frac{d}{2}$.

Level Set Probabilistic Model

- Prior: Gaussian $\mathbb{P}(du) = \mathsf{N}(0, \mathcal{C})$.
- Posterior: $\mathbb{P}_\gamma(du|y) \propto \exp\big(-\Phi_{\mathrm{ls}}(u;y)\big)\mathbb{P}(du)$.

$$\Phi_{\mathrm{ls}}(u;y) := \int_{D'} \frac{1}{2\gamma^2}\Big|y(x) - \mathrm{sign}\big(u(x)\big)\Big|^2 \rho(x)dx.$$

Connecting Probit, Level Set and Regression

M. Dunlop, D Slepčev, AM Stuart and M Thorpe, In preparation 2017.

Theorem 4

Let $\alpha > \frac{d}{2}$. We have $\mathbb{P}_\gamma(u|y) \Rightarrow \mathbb{P}(u|y)$ as $\gamma \to 0$ where

$$\mathbb{P}(du|y) \propto \mathbf{1}_A(u)\mathbb{P}(du), \quad \mathbb{P}(du) = \mathsf{N}(0, \mathcal{C})$$

$$A = \{u : \mathrm{sign}\big(u(x)\big) = y(x), \quad x \in D'\}.$$

Compare with regression (Zhu, Ghahramani, Lafferty 2003, (ICML):)

$$A_0 = \{u : u(x) = y(x), \quad x \in D'\}.$$

Example (PDE Two Moons – Unlabelled Data)

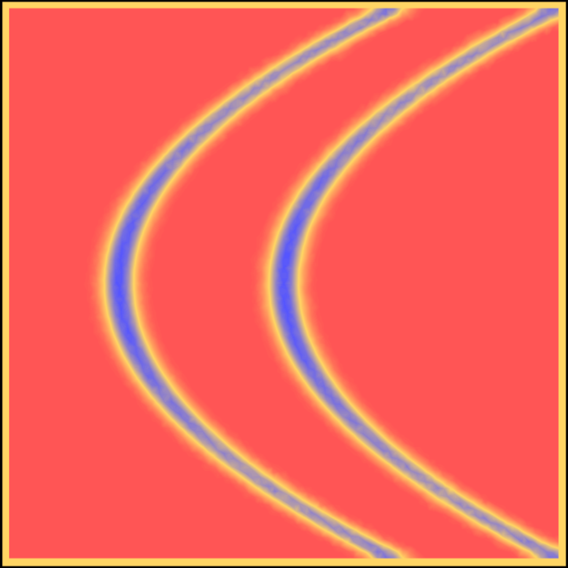


Figure: Sampling density ρ of unlabelled data.

Example (PDE Two Moons – Label Data)

Figure: Labelled Data.

Example (PDE Two Moons – Fiedler Vector of $\mathcal{L}$)

Figure: Fiedler Vector.

Example (PDE Two Moons – Posterior Labelling)

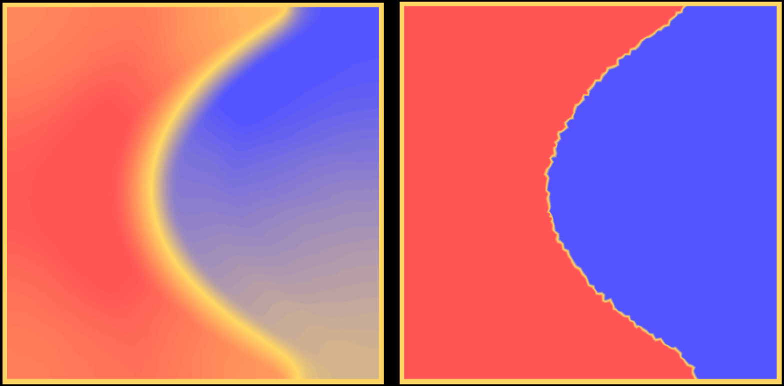


Figure: Posterior mean of u and $\mathrm{sign}(u)$.

Example (One Data Point Makes All The Difference)

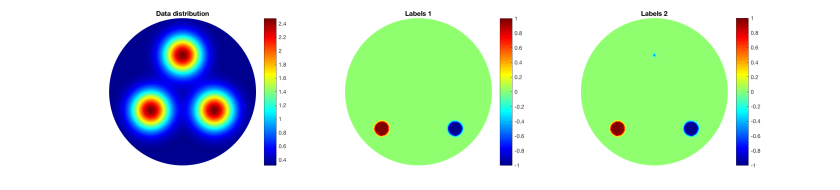


Figure: Sampling density, Label Data 1, Label Data 2.

Talk Overview

Learning and Inverse Problems

Optimization

Theoretical Properties

Probability

Conclusions

Summary

- Single optimization framework for classification algorithms.
- Single Bayesian framework for classification algorithms.
- Comparison of related optimization problems.
- Probit and Level Set have same small noise limit.
- This limit generalizes previous regression-based methods.
- Fast mixing MCMC algorithms.
- Fast per MCMC step approximations.
- Infinite data limit identifies appropriate parameter choices.
- Infinite data limit to (S)PDEs, conditioned Gaussian measure.

References

- X Zhu, Z Ghahramani, J Lafferty, *Semi-supervised learning using Gaussian fields and harmonic functions*, ICML, 2003. Harmonic Functions.
- C Rasmussen and C Williams, *Gaussian processes for machine learning*, MIT Press, 2006. Probit.
- AL Bertozzi and A Flenner, *Diffuse interface models on graphs for classification of high dimensional data*, SIAM MMS, 2012. Ginzburg-Landau.
- MA Iglesias, Y Lu, AM Stuart, *Bayesian level set method for geometric inverse problems*, Interfaces and Free Boundaries, 2016. Level Set.
- AL Bertozzi, M Luo, AM Stuart and K Zygalakis, *Uncertainty quantification in the classification of high dimensional data*, `https://arxiv.org/abs/1703.08816`, 2017. Probit on a graph.
- N Garcia-Trillos and D Slepčev, *A variational approach to the consistency of spectral clustering*, ACHA, 2017.
- M Dunlop, D Slepčev, AM Stuart and M Thorpe, *Large data and zero noise limits of graph based semi-supervised learning algorithms*, In preparation, 2017.
- N Garcia-Trillos, D Sanz-Alonso, *Continuum Limit of Posteriors in Graph Bayesian Inverse Problems*, `https://arxiv.org/abs/1706.07193`, 2017.

pCN

$$\alpha(u, v) = \min\{1, \exp(\Phi(u) - \Phi(v)\}.$$

The preconditioned Crank-Nicolson (pCN) Method

1: **while** $k < M$ **do**
2: $\quad v^{(k)} = \sqrt{1-\beta^2}u^{(k)} + \beta\xi^{(k)}$, where $\xi^{(k)} \sim \mathsf{N}(0, C)$.
3: $\quad$ Accept: $u^{(k+1)} = v^{(k)}$ with probability $\alpha(u^{(k)}, v^{(k)})$, otherwise
4: $\quad$ Reject: $u^{(k+1)} = u^{(k)}$.
5: **end while**

Why pCN?

- For given acceptance probability, β is independent of $N = |Z|$.
- Can exploit approximation of graph Laplacian (Nyström) and $\cdots$

Example of UQ (Two Moons)

Recall that $d = 10^2, N = 2 \times 10^3$.

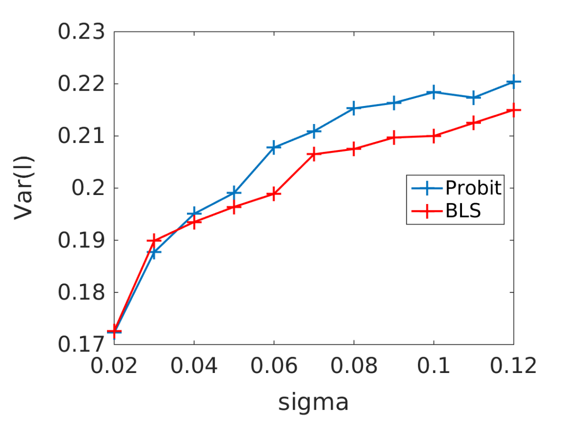


Figure: Average Label Posterior Variance vs σ, feature vector noise.

Example of UQ (MNIST)

Here $d = 784$ and $N = 4000$.

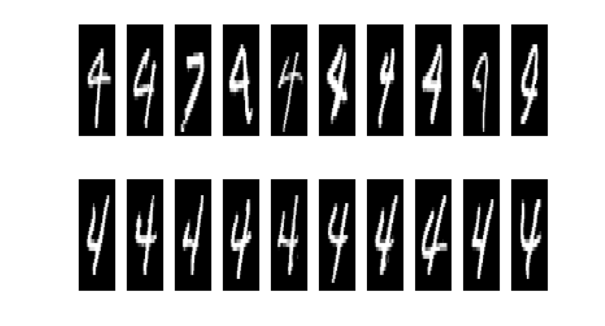


Figure: “Low confidence” vs “High confidence” nodes in MNIST49 graph.

Saturation of Spectra in Applications

Karhunen-Loeve – if $Lq_j = \lambda_j q_j$ then $u \sim \mathsf{N}(0, C)$ is:

$$u = c^{\frac{1}{2}} \sum_{j=1}^{N-1} (\lambda_j + \tau^2)^{-\frac{\alpha}{2}} q_j z_j, z_j \sim \mathsf{N}(0,1) \quad \text{i.i.d.} \tag{1}$$

- Spectrum of graph Laplacian often saturates as $j \to N-1$.
- Spectral Projection $\iff \lambda_k := \infty, k \geq \ell$.
- Spectral Approximation: set λ_k to some $\bar{\lambda} < \infty$.

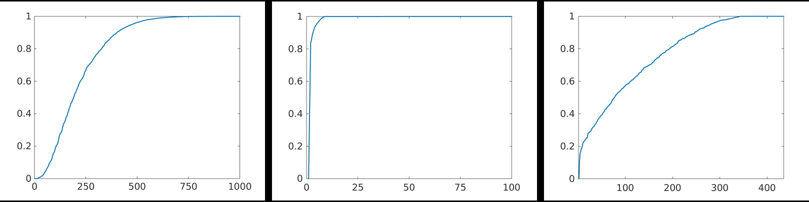


Figure: Two Moons, Hyperspectral, Voting Records.

Example of UQ (Voting)

Recall that $d = 16$ and $N = 435$.

Mean Absolute Error: *Projection*: 0.1577, *Approximation*: 0.0261.

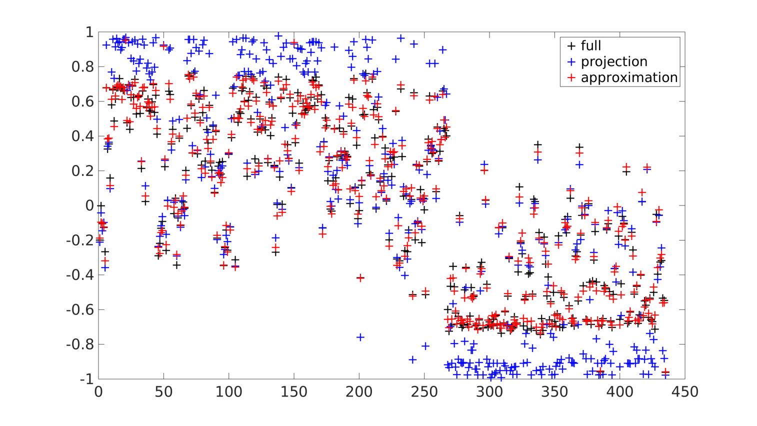


Figure: Mean Label Posterior. Compare Full (black), Spectral Approximation (red) and Spectral Projection (blue).

Example of UQ (Hyperspectral)

Here $d = 129$ and $N \approx 3 \times 10^5$. Use Nyström .

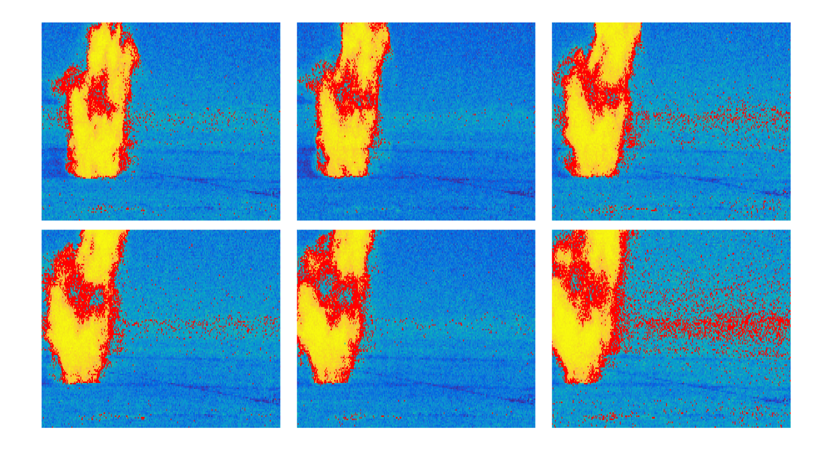


Figure: Spectral Approximation. Uncertain classification in red.