Noise Thresholds for Spectral Clustering

Sivaraman Balakrishnan

Poster: W056

Min Xu

Akshay Krishnamurthy

Aarti Singh

School of Computer Science
Carnegie Mellon University

Traditional analyses of Spectral Clustering

- k-way spectral clustering
  - Compute **L** = **D** – **W**, $\mathbf{v}_1, \ldots, \mathbf{v}_k \leftarrow$ smallest k eigenvectors of **L**
  - **E**mbed each data point i into k-dim space $\mathbf{x}(i) = [\mathbf{v}_1(i), \ldots, \mathbf{v}_k(i)]$
  - Run k-means on embedded data points

High-level justification: Connection to graph cut, random walks on graph, electric network theory, Laplace-Beltrami operator on manifold – don't translate to cluster recovery guarantees

Perturbation Analysis: Rohe et. al. (2010) and McSherry (2001) – spectral algorithms for planted partition (structured random graph) model (constant block similarities, low rank)

Jordan, Weiss (2001), Huang, Yan, Jordan, Taft (2009) – eigenvectors are **stable in l_2-norm** (Davis-Kahan Theorem) under small similarity perturbations

Our contributions – 1/2

- Study hierarchical spectral clustering and traditional k-way spectral clustering
- Characterization of general similarity conditions under which true eigenvectors reflect cluster structure, including **eigenvectors of hierarchically-structured high-rank matrices**
- Stability of eigenvectors in l_∞-norm under sub-Gaussian perturbation
- Precise characterization of **total clustering error** of k-way and hierarchical spectral clustering
  - As a function of noise variance, number of objects, size of clusters and within v/s between cluster similarity gap

Our contributions – 2/2

- **Information theoretic (minimax) optimality** of signal-to-noise thresholds
  - **Minimax lower bound**: No clustering method can succeed if

$$\sigma = \omega\left(\gamma\sqrt{\frac{\log n}{n}}\right)$$

σ - Noise std. dev. of perturbation, n - number of objects

γ - Gap between inter and intra cluster similarity, γ/σ - SNR

  - **Ratio min-cut** (combinatorial) achieves this rate up to constants
  - **Spectral clustering** succeeds if

$$\sigma = o\left(\gamma\sqrt[4]{\frac{\log n}{n}}\right)$$

- Remarks:
  - Price of computational efficiency: ratio min-cut (combinatorial) outperforms spectral clustering (efficient)
  - Conjecture rate can be improved under different conditions on noise