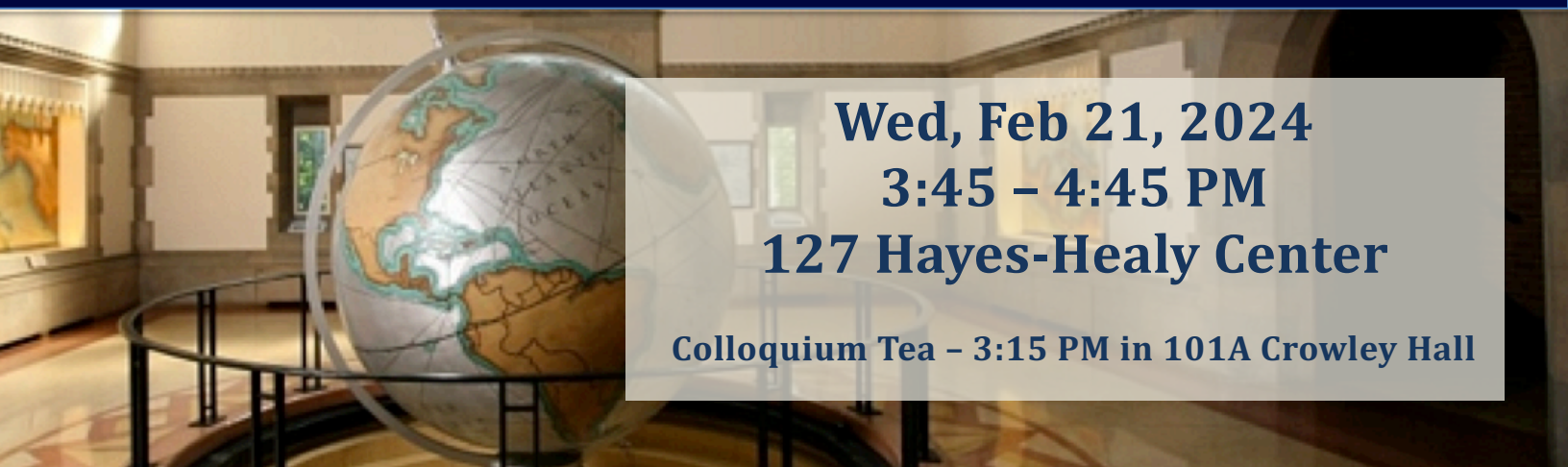
Department of Applied and Computational Mathematics and Statistics Colloquium

Soham Jana

Princeton University

Resolving the robust clustering problem for general data distributions

Clustering is a fundamental tool in statistical machine learning in the presence of heterogeneous data. Many recent results, such as the performances of the Lloyd algorithm and spectral clustering techniques, focus primarily on optimal mislabeling guarantees when data are distributed around centroids with sub-Gaussian errors. Yet, the restrictive sub-Gaussian model is often invalid in practice since various real-world applications exhibit heavy tail distributions around the centroids or suffer from possible adversarial attacks that call for robust clustering with a robust data-driven initialization. In this work, we introduce novel hybrid clustering techniques to produce optimal mislabeling guarantees under a weak initialization condition for general error distributions around the centroids. In addition, our approach also produces optimal mislabeling even in the presence of adversarial outliers. Our results reduce to the sub-Gaussian case when errors follow sub-Gaussian distributions. To solve the problem thoroughly, we also present a novel data-driven robust initialization technique and show that, with probabilities approaching one, these initial centroid estimates are sufficiently good for the subsequent clustering algorithms to achieve optimal guarantees. Both simulated and real data examples support our robust initialization procedure and clustering algorithms.



Wed, Feb 21, 2024
3:45 – 4:45 PM
127 Hayes-Healy Center

Colloquium Tea – 3:15 PM in 101A Crowley Hall