0.1 LOL @jovo

This month we finalized the real data analysis using LOL. In particular, we considered four datasets, the Prostate and Colon datasets have extensively been studied in the sparse literature. LOL yields better perforance, and for Colon, with lower dimensionality. MNIST is an even more prominent dataset, LOL achieves the best performance for all dimensions. MRN is a new dataset that we generated; it has over 500 million features, and 112 samples. We subsampled to 100 samples for cross-validation purposes. To our knowledge, no other machine learning tool is capable of even operating on 500 million features. Moreover, we demonstrate that our implementation outperforms first doing PCA on the data, for any number of dimensions that we embed into. We then also investigate the amount of time it takes to run LOL on very wide datasets. For a 128 million dimensional dataset, with 2000 samples, requiring nearly half a terabyte of space just to store, we have an approximate implementation that runs on a single machine and only takes about 3 minutes.

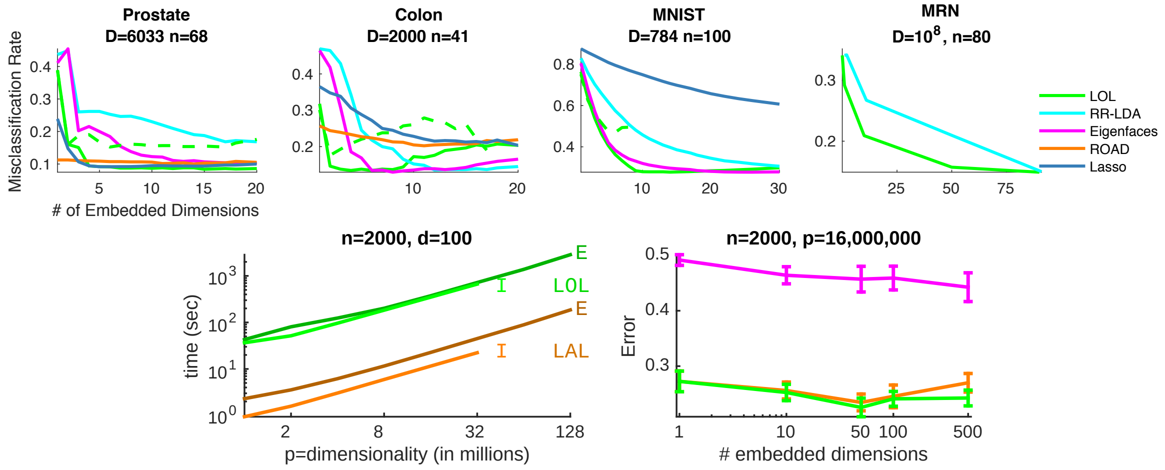

Figure 1: Top: LOL outperforms essentially all other methods for essentially all problems and number of dimensions to embed into. Bottom: LOL (green) and LAL (brown, using random projections to approximate) for both fully in memory (IE) and semi-external memory (E) implementations. Magenta compares performance to our scalable implementation of eigenfaces on the same problem.