Fast Spectral Clustering via the Nyström Method

Anna Choromanska[1⋆], Tony Jebara[2], Hyungtae Kim[2], Mahesh Mohan[3], and Claire Monteleoni[3]

[1]Department of Electrical Engineering, Columbia University, NY, USA
[2]Department of Computer Science, Columbia University, NY, USA
[3]Department of Computer Science, George Washington University, DC, USA
{aec2163, tj2008, hk2561}@columbia.edu, {mahesh_mohan, cmontel}@gwu.edu

Abstract. We propose and analyze a fast spectral clustering algorithm with computational complexity linear in the number of data points that is directly applicable to large-scale datasets. The algorithm combines two powerful techniques in machine learning: spectral clustering algorithms and Nyström methods commonly used to obtain good quality low rank approximations of large matrices. The proposed algorithm applies the Nyström approximation to the graph Laplacian to perform clustering. We provide theoretical analysis of the performance of the algorithm and show the error bound it achieves and we discuss the conditions under which the algorithm performance is comparable to spectral clustering with the original graph Laplacian. We also present empirical results.

Keywords: spectral clustering, Nyström method, large-scale clustering, sampling, sparsity, performance guarantees, error bounds, unsupervised learning

1 Introduction

Clustering is one of the fundamental problems in machine learning. The recent widespread development of sensors, data-storage and data-acquisition devices has helped make large data-sets common place. This, however, poses a serious computational challenge for the existing clustering techniques. Spectral clustering techniques (Luxburg, 2007) are widely used, due to their simplicity and empirical performance advantages compared to other clustering methods, such as k-means or single-linkage algorithms. However, a significant obstacle to scaling up spectral clustering to large datasets is that it requires building an affinity matrix between pairs of data points which becomes computationally prohibitive for large data-sets.

There have been several attempts to address this problem and make spectral clustering algorithms more applicable to large-scale problems. Here we study an

⋆ Main contact author: Anna Choromanska, e-mail: aec2163@columbia.edu, mailing adress: Department of Electrical Engineering, Columbia University, CEPSR 624, 1214 Amsterdam Avenue, 10027, NY, USA.

approach that extends the spectral clustering algorithm, described in Ng et al. (2001), via Nyström approximation techniques. Our work is most related to Williams and Seeger (2001); Fowlkes et al. (2004); Li et al. (2011), which use the Nyström method to sample the columns of the affinity matrix and further approximate the full matrix by using correlations between the sampled columns and the remaining columns (Fowlkes et al., 2004). However, these works did not provide performance guarantees; that is our primary contribution.

Other approaches to scaling up spectral clustering include work by Yan et al. (2009), which used the k-means clustering algorithm (Lloyd, 1982) as a preprocessing step to spectral clustering, to reduce its computational complexity. The analysis assumes the data are generated by a mixture model (the same assumption is made in the work by Lashkari and Golland (2007)). Related work by Drineas and Mahoney (2005) performs non-uniform sampling of the Gram matrix and provides a bound on the approximation error, however in order to achieve good performance one may need to sample large number of columns (in special cases even $\mathcal{O}(n)$) and, furthermore, the practicality of this technique for massive datasets may be limited (Yan et al., 2009). Several other works on constructing approximations that are tighter than the Nyström method's, for sparse graph Laplacians, have also emerged (Fung et al., 2011; Spielman and Teng, 2011). However, the computational complexity of these methods depends highly on the number of edges in the graph. In contrast, the Nyström method has a fixed complexity that is dependent only on the number of vertices n and the number of sampled columns l. This is potentially more useful in cases where a large number of edges exist but only a few have significantly large weights, as is the case in many sparse datasets that arise in applications, such as collaborative filtering.

This paper combines the spectral clustering algorithm (Ng et al., 2001) with the Nyström approximation method by using a Nyström approximation to the graph Laplacian. Our analysis differs from the approach of Belkin and Niyogi (2007) in that we focus on the finite sample analysis whereas Belkin and Niyogi (2007) emphasized asymptotic results. In particular, they show that if points are sampled uniformly at random from an unknown sub-manifold $\mathcal{M} \in \mathbb{R}^N$, then the eigenvectors of a suitably constructed graph Laplacian converge to the eigenfunctions of the Laplace-Beltrami operator on $\mathcal{M}$. Our approach leads to a practical algorithm with complexity linear in the number of data points n. We provide performance guarantees for this algorithm by combining Nyström approximation analysis, using a uniform random sampling without replacement scheme due to Kumar et al. (2009), with perturbation theory analysis (Ng et al., 2001). We discuss conditions under which the algorithm's performance is comparable to spectral clustering with the original graph Laplacian.

2 Approach

2.1 Spectral Clustering Algorithm

Algorithm 1 Spectral clustering

Input: dataset $S = \{s_1, s_2..., s_n\} \in \mathbb{R}^d$, number of clusters k, kernel function $\kappa : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$
Output: k-clustering of S

$A \in \mathbb{R}^{n \times n}$ s.t. $A_{ij} = \delta[i \neq j]\kappa(s_i, s_j)$
$D \in \mathbb{R}^{n \times n}$ s.t. $D_{ij} = \delta[i = j] \sum_{j=1}^{n} A_{ij}$
$L = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$
$\mathcal{X} \in \mathbb{R}^{n \times k}$ s.t. $\mathsf{SmallestEigenVectors}(L, k)$
$\mathcal{Y} \in \mathbb{R}^{n \times k}$ s.t. $\mathcal{Y}_{im} = \mathcal{X}_{im} / \sqrt{\sum_m \mathcal{X}_{im}^2}$
$\mathcal{K} = \mathsf{ClusterRows}(\mathcal{Y})$ (via any k-clustering algorithm minimizing distortion, i.e. k-means)

In general, spectral clustering methods can be interpreted as graph partitioning algorithms and the above algorithm (Algorithm 1) can be seen as graph partitioning with a normalized-cut cost function. Algorithm 1 shows the widely used normalized spectral clustering algorithm presented in Ng et al. (2001). Given the set of n points $S = \{s_1, s_2..., s_n\}$ the algorithm first builds an $n \times n$ affinity matrix A, i.e.:

$$A_{ij} = \kappa(s_i, s_j) \text{ if } i \neq j \text{ and 0 otherwise.}$$

Here A_{ij} corresponds to the i'th row and j'th column of the affinity matrix and κ is any kernel function accepting two input data-points and returning a scalar output. Once the affinity matrix is computed, the normalized graph Laplacian L can be constructed. The first k eigenvectors of L are then normalized and clustered. It was shown in Ng et al. (2001) that one can perform spectral k-clustering using a perturbed version $\tilde{A}$ of the ideal affinity matrix A. Under certain assumptions, the clusterings obtained using A and $\tilde{A}$ will be similar. Our goal is to extend these assumptions and show that using 'close' to ideal graph Laplacian L and its Nyström r-rank approximation, $\tilde{L}$ will also give similar clustering results. Based on the analysis in Ng et al. (2001) we know that if the four assumptions listed below are satisfied then using either $\tilde{A}$ or A to perform spectral clustering will give similar partitionings of the dataset (and also similar to the true clustering of the dataset):

- Assumption A1: $\exists_{\gamma>0} \forall_{i=\{1,2,...,k\}} \lambda_2^i \leq 1 - \gamma$, where λ_2^i is the second largest eigenvalue of L^i, where L^i is the subblock of L corresponding to cluster i.
- Assumption A2: $\exists_{\epsilon_1>0} \forall_{i_1,i_2=\{1,2,...,k\}, i_1 \neq i_2} \sum_{j \in S_{i_1}} \sum_{l \in S_{i_2}} \frac{\tilde{A}_{jl}^2}{\tilde{d}_j \tilde{d}_l} \leq \epsilon_1$, where $\tilde{d}_j = \sum_{m \in S_{i_1}} \tilde{A}_{jm}$ and $\tilde{d}_l = \sum_{m \in S_{i_2}} \tilde{A}_{lm}$ and S_i is the set of points belonging to the i^{th} cluster.
- Assumption A3: $\exists_{\epsilon_2>0} \forall_{i=\{1,2,...,k\}, j \in S_i} \frac{\sum_{l: l \notin S_i} \tilde{A}_{jl}}{\tilde{d}_j} \leq \epsilon_2 (\sum_{l,m \in S_i} \frac{\tilde{A}_{lm}^2}{\tilde{d}_l \tilde{d}_m})^{-\frac{1}{2}}$.
- Assumption A4: $\exists_{\mathcal{C}>0} \forall_{i=\{1,2,...,k\}, j=\{1,2,...,n_i\}} \tilde{d}_j \geq (\sum_{l=1}^{n_i} \tilde{d}_l)/(\mathcal{C} n_i)$.

Assumption A1 guarantees each cluster to be tight. Assumption A2 and A3 require data points within a cluster to be more connected to each other than

they are with data points from any other cluster. Finally, the last assumption requires that the points in any cluster can never be much less' connected than other points in the same cluster. The similarity of the clusterings obtained using A and $\tilde{A}$ is then assured via Theorem 1. Let y_j^i be the j^{th} row of $\mathcal{Y}^i$ from Algorithm 1, where $\mathcal{Y}^i$ is the subblock of $\mathcal{Y}$ corresponding to cluster i. Then the following theorem holds.

Theorem 1 (Ng et al. (2001)). *Let assumptions A1, A2, A3 and A4 hold. Set* $\epsilon = \sqrt{k(k-1)\epsilon_1 + k\epsilon_2^2}$. *If* $\gamma > (2+\sqrt{2})\epsilon$, *then there exist* k *orthonormal vectors* $r_1, r_2, \ldots, r_k$ *such that* $\mathcal{Y}$ *in Algorithm 1 satisfies*

$$\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\|y_j^i - r_i\|^2 \leq 4C(4+2\sqrt{k})^2\frac{\epsilon^2}{(\gamma-\sqrt{2}\epsilon)^2}.$$

2.2 Nyström Method for Matrix Approximation

Algorithm 2 Nyström method for matrix approximation

1: Input: matrix L, l - number of columns sampled, r - rank approximation ($r \leq l << n$)
2: Output: $\tilde{\Sigma}$ and $\tilde{U}$ such that $\tilde{L} = \tilde{U}\tilde{\Sigma}\tilde{U}^\top$

3: $\mathcal{L} \leftarrow$ indices of l columns sampled
4: $C \leftarrow G(:, \mathcal{L})$
5: $W \leftarrow C(\mathcal{L}, :)$
6: $W_r \leftarrow$ best r-rank approximation to W
7: $\tilde{\Sigma} = \frac{n}{l}\Sigma_{W_r}$ and $\tilde{U} = \sqrt{\frac{l}{n}}CU_{W_r}\Sigma_{W_r}^{-1}$, where $W_r = U_{W_r}\Sigma_{W_r}U_{W_r}^\top$

We now explicate the Nyström r-rank approximation for any symmetric positive semidefinite (SPSD) matrix $L \in \mathbb{R}^{n\times n}$. After performing sampling (we will only be using uniform sampling without replacement schemes), create matrix $C \in \mathbb{R}^{n\times l}$ from the sampled columns. Then, form matrix $W \in \mathbb{R}^{l\times l}$ matrix consisting of the intersection of these l columns with the corresponding l rows of L. Let $W = U\Sigma U^\top$, where U is orthogonal and $\Sigma = diag(\sigma_1, \sigma_2, \ldots, \sigma_l)$ is a real diagonal matrix with the diagonal sorted in decreasing order. Let W_r^+ be the pseudo-inverse of the best rank-r approximation to W ($W_r^+ = \sum_{t=1}^{r}\sigma_t^{-1}U^{(t)}U_{(t)}$, where $U^{(t)}$ and $U_{(t)}$ are respectively the t^{th} column and row of U). Then the Nyström approximation $\tilde{L}$ of L can be obtained as follows: $\tilde{L} = CW_r^+C^\top$. Furthermore if we represent $\tilde{L}$ as $\tilde{L} = \tilde{U}\tilde{\Sigma}\tilde{U}^\top$ then $\tilde{\Sigma} = \frac{n}{l}\Sigma_{W_r}$ and $\tilde{U} = \sqrt{\frac{l}{n}}CU_{W_r}\Sigma_{W_r}^{-1}$, where $W_r = U_{W_r}\Sigma_{W_r}U_{W_r}^\top$. Theorem 2 due to Kumar et al. (2009) shows the performance bounds for the Nyström method when used with uniform sampling without replacement. In Kumar et al. (2009) the authors also

compare the quality of obtained Nyström approximations, on the experiments with large-scale datasets, when using uniform and non-uniform sampling strategies (they consider both sampling with and without replacement). They consider two most popular non-uniform sampling techniques: column-norm sampling and diagonal sampling. They show that uniform sampling without replacement is not only more effcient both in time and space but also improves the accuracy of the Nyström method.

Theorem 2 (Kumar et al. (2009)). *Let $G \in \mathbb{R}^{n\times n}$ be an SPSD matrix. Assume that l columns of G are sampled uniformly at random without replacement, let $\tilde{G}_r$ be the rank-r Nyström approximation to G and let G_r be the best rank-r approximation to G. Let $\epsilon > 0$, $l \geq 64r/\epsilon^4$ and $\eta = \sqrt{\frac{\log(2/\delta)\xi(l,n-l)}{l}}$, where $\xi(m,u) = \frac{mu}{m+u-1/2}\frac{1}{1-1/(2\max\{m,u\})}$. Then with probability at least $1-\delta$,*

$$\|G - \tilde{G}_r\|_F \leq \|G - G_r\|_F + \epsilon \left[\left(\frac{n}{l} \sum_{i\in D(l)} G_{ii} \right) \sqrt{n \sum_{i=1}^{n} G_{ii}^2} + \eta \max(nG_{ii}) \right]^{\frac{1}{2}},$$

where $\|\cdot\|_F$ is the Frobenius norm, $\sum_{i\in D(l)} G_{ii}$ is the sum of the largest l diagonal entries of G.

3 Fast Spectral Clustering Algorithm

Algorithm 3 Fast spectral clustering

Input: dataset $S = \{s_1, s_2..., s_n\} \in \mathbb{R}^d$, k - number of clusters, l - number of columns sampled, r - rank approximation ($k \leq r \leq l << n$)
Output: k-clustering of S

$\mathcal{L} \leftarrow$ indices of l columns sampled (uniformly without replacement)
$\hat{A} \leftarrow A(:, \mathcal{L})$
$D \in \mathbb{R}^{n\times n}$ s.t. $D_{ij} = \delta[i=j]1/\sqrt{\sum_{j=1}^{l} \hat{A}_{ij}}$
$\Delta \in \mathbb{R}^{l\times l}$ s.t. $\Delta_{ij} = \delta[i=j]1/\sqrt{\sum_{i=1}^{n} \hat{A}_{ij}}$
$C \leftarrow \hat{I} - \sqrt{\frac{l}{n}} D \times \hat{A} \times \Delta$: $I^{'}$ - matrix of columns of I indexed by $\mathcal{L}$
$W \leftarrow C(\mathcal{L}, :)$
$W_r \leftarrow$ best r-rank approximation to W
$\tilde{\Sigma} = \frac{n}{l}\Sigma_{W_r}$ and $\tilde{U} = \sqrt{\frac{l}{n}} C U_{W_r} \Sigma_{W_r}^{-1}$, where $W_r = U_{W_r}\Sigma_{W_r}U_{W_r}^{\top}$
$\mathcal{X} \leftarrow \mathsf{SmallestEigenVectors}(\tilde{U}, k)$
$\mathcal{Y} \leftarrow \mathsf{NormalizeRows}(\mathcal{X}) : \mathcal{Y}_{im} = \mathcal{X}_{im}/(\sum_m \mathcal{X}_{im}^2)^{\frac{1}{2}}$
$\mathcal{K} \leftarrow \mathsf{ClusterRows}(Y)$ (use any k-clustering algorithm minimizing distortion, i.e. k-means)

For large-scale fast spectral clustering, we propose Algorithm 3. The algorithm chooses l columns sampled uniformly at random from the affinity matrix. It therefore never builds the entire $n \times n$ affinity matrix which would be computationally prohibitive. It then computes two sparse diagonal degree matrices D AND Δ. Subsequently, the matrix C is recovered which is n x l. Matrix C plays the role of the sampled graph Laplacian. We then follow the steps of Algorithm 2 to obtain the approximate eigensystem of the graph Laplacian and finally the first k eigenvectors are normalized and clustered. Clearly, Algorithm 3 performs sampling of the affinity matrix. This is in contrast to the more computationally expensive approach of computing the complete $n \times n$ affinity matrix and then obtaining matrix C by sampling directly from the graph Laplacian. We provide Theorem 3 to show that for appropriate values of l, both of these algorithms will give similar clustering results. First, let us introduce some additional notation. We consider two scenarios: sampling the graph Laplacian and sampling the affinity matrix. Let $\mathcal{L}$ be the set of indices of sampled columns. Let $I^{'}$ be the matrix of columns of I that are indexed by $\mathcal{L}$. Notice that for $i \in \{1, 2, \ldots, n\}$ and $j \in \mathcal{L}$, any entry in the sampled graph Laplacian has the following form:

$$C^{'}_{ij} = I^{'}_{ij} - \frac{A_{ij}}{\sqrt{(\sum_{a=1}^{n} A_{aj})(\sum_{b=1}^{n} A_{ib})}}.$$

On the other hand, matrix C in Algorithm 3 has the following form:

$$C_{ij} = I^{'}_{ij} - \sqrt{\frac{l}{n}} \frac{A_{ij}}{\sqrt{(\sum_{a=1}^{n} A_{aj})(\sum_{b \in \mathcal{L}} A_{ib})}}.$$

The difference lies in the second term in the denominator and the scaling factor $\sqrt{\frac{l}{n}}$. Consider Theorem 3.

Theorem 3. *Let A_{ij}'s be iid*[1] *scalar random variables (bounded in $[0, 1]$) whose expectation is μ. With probability at least $1 - \delta$ the following holds:*

$$C^{'}_{ij} \sqrt{\frac{\mu}{\mu + \delta^{'}}} \leq \lim_{n \to \infty} C_{ij} \leq C^{'}_{ij} \sqrt{\max(\frac{\mu}{\mu - \delta^{'}}, 1)},$$

where $\delta^{'} = \frac{1}{\sqrt{l}} \sqrt{\log(2/\delta)}$.

Proof. By the law of large numbers we have that $\lim_{n \to \infty} \frac{1}{n} \sum_{b=1}^{n} A_{ib} = \mu$. By Hoeffding's inequality we have that with probability at least $1 - \delta$, $|\frac{1}{l} \sum_{b \in \mathcal{L}} A_{ib} - \mu| \leq \frac{1}{\sqrt{l}} \sqrt{\log(2/\delta)}$. Therefore

$$\lim_{n \to \infty} C_{ij} = \lim_{n \to \infty} \left[I^{'}_{ij} - \sqrt{\frac{l}{n}} \sqrt{\frac{n}{l}} (I^{'}_{ij} - C^{'}_{ij}) \sqrt{\frac{\frac{1}{n} \sum_{b=1}^{n} A_{ib}}{\frac{1}{l} \sum_{b \in \mathcal{L}} A_{ib}}} \right]$$

[1] The iid assumption is made only for the purpose of this section.

$$= \lim_{n\to\infty} \left[C_{ij}^{'} \sqrt{\frac{\frac{1}{n}\sum_{b=1}^{n} A_{ib}}{\frac{1}{l}\sum_{b\in\mathcal{L}} A_{ib}}} \right]$$

Finally, if $\frac{1}{l}\sum_{b\in\mathcal{L}} A_{ib} \geq \mu$:

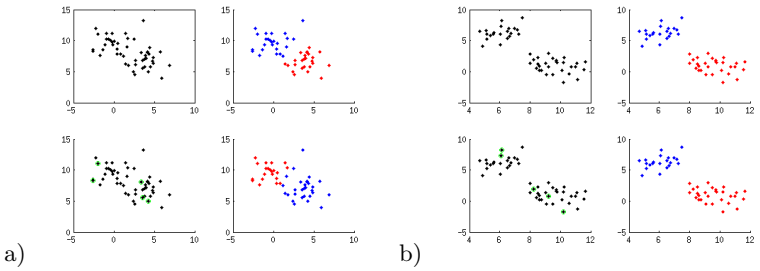
$$C_{ij}^{'}\sqrt{\frac{\mu}{\mu+\delta^{'}}} \leq \lim_{n\to\infty} \left[C_{ij}^{'} \sqrt{\frac{\frac{1}{n}\sum_{b=1}^{n} A_{ib}}{\frac{1}{l}\sum_{b\in\mathcal{L}} A_{ib}}} \right] \leq C_{ij}^{'},$$

and if $\frac{1}{l}\sum_{b\in\mathcal{L}} A_{ib} < \mu$:

$$C_{ij}^{'} \leq \lim_{n\to\infty} \left[C_{ij}^{'} \sqrt{\frac{\frac{1}{n}\sum_{b=1}^{n} A_{ib}}{\frac{1}{l}\sum_{b\in\mathcal{L}} A_{ib}}} \right] \leq C_{ij}^{'}\sqrt{\frac{\mu}{\mu-\delta^{'}}},$$

where $\delta^{'} = \frac{1}{\sqrt{l}}\sqrt{\log(2/\delta)}$. Combining both cases gives the theorem.

Theorem 3 shows that, for sufficiently large l, the two algorithms under consideration (Algorithm 3 sampling the affinity matrix and the slower alternative of sampling the graph Laplacian) should produce similar C matrices and thus yield similar clustering results. Furthermore, in batch settings with finite n, Algorithm 3 is still applicable, i.e. consider the example presented on Figure 1 showing the partitionings of two simple datasets obtained by spectral clustering algorithm of Ng et al. (2001) using the full affinity matrix and, for comparison, our Algorithm 3. The size of both datasets is very small ($n = 50$), but the performance of both algorithms is very similar. Finally, in our theoretical analysis we will focus on the scenario when the graph Laplacian is being sampled. This analysis is easier than considering sampling the affinity matrix which, for instance, does not need to be PSD. We end this section with Theorem 4 showing the computational complexity of the proposed Algorithm 3.



a) b)

Fig. 1. Result of spectral clustering on two datasets (a and b), $n = 50$, $l = 10\% * n$, $r = l$. Top row: the dataset (left) and the partitioning obtained by spectral clustering using the full affinity matrix (right). Bottom row: the dataset with sampled data points (green) (left) and the partitioning obtained by Algorithm 3 (using the sampled affinity matrix) (right).

Theorem 4. *The computational complexity of Algorithm 3 is* $\mathcal{O}(nl\max(r,c))+\Gamma$*, where* c *is the cost of evaluating a single kernel function between two data points and* Γ *is the cost of the clustering algorithm minimizing distortion used to obtain the final clustering.*

4 Performance Guarantees

As was mentioned before the theoretical analysis considers the case where we sample the graph Laplacian built from the $n \times n$ affinity matrix and thus C is an $n \times l$ matrix of sampled columns. Furthermore, matrix W is an $l \times l$ matrix consisting of the intersection of these l columns with the corresponding l rows of the graph Laplacian. Our theoretical analysis will consider the performance of the proposed algorithm in the case where the affinity matrix and the corresponding graph Laplacian are 'close' to block diagonal matrices. In particular we will require that $\|L - L_r^{'}\|_F \leq \|L^{'} - L_r^{'}\|_F \leq \epsilon n$, where L is the ideal graph Laplacian, $L^{'}$ is the true, 'close' to diagonal, graph Laplacian that is sampled and $L_r^{'}$ is its best rank r Nyström approximation. Two more conditions that we assume holds will be introduced later. This section is organized as follows: we will first show the main result (Theorem 5) and then we will show the technical lemmas and proofs that led to this result.

4.1 Main Result

Let A be the ideal affinity matrix that gave rise to the ideal graph Laplacian L. Let $\tilde{L}$ be the Nyström r-rank approximation to $L^{'}$ and let $\tilde{A}$ be the affinity matrix that would give rise to $\tilde{L}$ in case when no Nyström approximation was used. We will now present our main result, Theorem 5. Let y_j^i be the j^{th} row of $\mathcal{Y}^i$ from Algorithm 3, where $\mathcal{Y}^i$ is the subblock of $\mathcal{Y}$ corresponding to cluster i. Then the following theorem holds.

Theorem 5. *Let* $\epsilon > 0$*,* $l \geq 64r/\epsilon^4$ *and* $\eta = \sqrt{\frac{\log(2/\delta)\xi(l,n-l)}{l}}$*, where* $\xi(m,u) = \frac{mu}{m+u-1/2} \cdot \frac{1}{1-1/(2\max\{m,u\})}$*. Let* γ*,* ϵ_1*,* ϵ_2 *and* $\mathcal{C}$ *be defined as in Lemma 1, 2, 3 and 4. Set* $\epsilon^{'} = \sqrt{k(k-1)\epsilon_1 + k\epsilon_2^2}$*. If* $\gamma > (2+\sqrt{2})\epsilon^{'}$*, then with probability at least* $1-\delta$*, there exist* k *orthogonal vectors* $r_1, r_2, \ldots, r_k$ *(*$r_i^\top r_j = 1$ *if* $i = j$*,* 0 *otherwise) so that* $\mathcal{Y}$ *in Algorithm 3 satisfies:*

$$\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\|y_j^i - r_i\|^2 \leq 4\mathcal{C}(4+2\sqrt{k})^2\frac{\epsilon^{'2}}{(\gamma-\sqrt{2}\epsilon^{'})^2}$$

Theorem 5 is generalization of Theorem 1. It differs from Theorem 1 in that it extends the four assumptions used in Theorem 1 which result from the fact that $\tilde{A}$ is a very special version of the perturbed ideal A, in particular it is an affinity matrix that gave rise to the Nyström r-rank approximation to the graph

Laplacian. The assumption on each λ_2^i ensures that each cluster is tight enough such that after sampling the clusters will still remain tight (γ can be interpreted as the measure of tigtness of each cluster after sampling). This assumption also shows that when we decrease the number of sampled columns l, we expect the original clusters to be tighter in order for the clusters obtained after sampling to also be tight enough such that the dataset is still k-clusterable.

4.2 Theoretical Analysis

We will first present Theorem 6 which is a version of Theorem 2 when the sampled matrix is a graph Laplacian L. Theorem 6 relies on the fact that L is a SPSD matrix that is 'close' to block diagonal.

Theorem 6. *Let $L \in \mathbb{R}^{n\times n}$ be an ideal graph Laplacian and $L^{'}$ be the 'close' to block diagonal graph Laplacian defined before. Assume that l columns of $L^{'}$ are sampled uniformly at random without replacement and let $\tilde{L}$ be the best rank-r Nyström approximation to $L^{'}$. Let $\epsilon > 0$, $l \geq 64r/\epsilon^4$ and $\eta = \sqrt{\frac{\log(2/\delta)\xi(l,n-l)}{l}}$, where $\xi(m,u) = \frac{mu}{m+u-1/2} \cdot \frac{1}{1-1/(2\max\{m,u\})}$. Then with probability at least $1-\delta$,*

$$\|L - \tilde{L}\|_F \leq \epsilon n\sqrt{1+\eta}.$$

Recall a useful theorem (Theorem 7) that we will need later. It can be found i.e. in Kannan and Vempala (2009). Intuitively Theorem 7 implies that if two matrices are close (in terms of the squared Frobenius norm of their difference), then their singular values should also be close too.

Theorem 7. *For any two $n \times n$ symmetric matrices A and B,*

$$\sum_{t=1}^{n}(\sigma_t(A) - \sigma_t(B))^2 \leq \|A - B\|_F^2$$

We now proceed with the theoretical analysis that will lead to Theorem 5. We aim to make use of Theorem 1 and then Theorem 6 to provide theoretical guarantees on the performance of spectral clustering when using the Nyström approximation to the 'close' to ideal ('close' to block diagonal) graph Laplacian. We will focus on extending assumptions A1, A2, A3 and A4 used in Theorem 1. We will present Lemma 1, 3, 4 and 2. Applying them to Theorem 1 yields our main result captured in Theorem 5. We will also make three additional assumptions that we will assume hold throughout the entire analysis. First of all we will assume that $(\tilde{\lambda}_2^i - \lambda_2^i)^2 \leq \frac{1}{n}\sum_{t=1}^{r}(\tilde{\lambda}_t^i - \lambda_t^i)^2$. Secondly we will assume that $\frac{1}{h} \leq \frac{n_i}{\tilde{n}_i} \leq h$, where n_i is the number of data points assigned to cluster i when using L, $\tilde{n}_i$ is the number of data points assigned to cluster i when using $\tilde{L}$ rather than L and $h \geq 1$ is some positive constant. Finally since the original Laplacian is

assumed to be 'close' to block diagonal, we would like its Nyström approximation to be 'close' to block diagonal by assuming that the following two conditions holds: $\sum_{j\in S_{i_1}}\sum_{l\in S_{i_2}}\|\tilde{L}_{j'l}-L_{j'l}\|^2 = n_{i_1}\sum_{l\in S_{i_2}}\|\tilde{L}_{j'l}-L_{j'l}\|^2 \leq \frac{f}{n}\|\tilde{L}-L\|^2$, where $i_1 \neq i_2$, $j^{'} = \arg\max_{j\in S_{i_1}}\sum_{l\in S_{i_2}}\|\tilde{L}_{jl}-L_{jl}\|^2$ and $f>1$ is some constant, and $\sum_{j,l\in S_i}\|\tilde{L}_{jl}-L_{jl}\|^2 \leq \frac{g}{n}\|\tilde{L}-L\|^2$, where $g>1$ is some constant.

Lemma 1. *Let λ_2^i be the second largest eigenvalue of L^i, where L^i is the subblock of L corresponding to cluster i, and let $\tilde{\lambda}_2^i$ be the second largest eigenvalue of $\tilde{L}^i$, where $\tilde{L}^i$ is the subblock of $\tilde{L}$ corresponding to cluster i. If $\lambda_2^i \leq 1-c\epsilon\sqrt{r(1+\eta)}$ ($c>1$ is some constant) then with probability at least $1-\delta$, $\exists_{\gamma>0}\tilde{\lambda}_2^i \leq 1-\gamma$.*

Proof. We know that

$$(\tilde{\lambda}_2^i-\lambda_2^i)^2 \leq \frac{1}{n}\sum_{t=1}^{r}(\tilde{\lambda}_t^i-\lambda_t^i)^2 \leq \frac{1}{n}\|\tilde{L}^i-L^i\|_F^2 \leq \frac{1}{n}\|\tilde{L}-L\|_F^2, \tag{1}$$

where the first inequality comes from Theorem 7. By applying Jensen's inequality to the left hand side of Equation 1, we obtain

$$\sum_{t=1}^{r}|\tilde{\lambda}_t^i-\lambda_t^i| \leq \sqrt{r}\frac{1}{n}\|\tilde{L}-L\|_F. \tag{2}$$

Then in particular the following holds:

$$|\tilde{\lambda}_2^i-\lambda_2^i| \leq \sqrt{r}\frac{1}{n}\|\tilde{L}-L\|_F. \tag{3}$$

By assumption, we know that $\lambda_2^i \leq 1-c\epsilon\sqrt{r(1+\eta)}$. Now, if $\tilde{\lambda}_2^i \leq \lambda_2^i$ then lemma holds. If $\tilde{\lambda}_2^i > \lambda_2^i$, then we can rewrite Equation 3 as:

$$\tilde{\lambda}_2^i \leq \lambda_2^i + \sqrt{r}\frac{1}{n}\|\tilde{L}-L\|_F \tag{4}$$

Since $\lambda_2^i \leq 1-c\epsilon\sqrt{r(1+\eta)}$ and by Theorem 6 with probability at least $1-\delta$ the following holds: $\sqrt{r}\frac{1}{n}\|\tilde{L}-L\|_F \leq \epsilon\sqrt{r(1+\eta)}$, then we can write that with probability at least $1-\delta$, $\exists_{\gamma>0}\tilde{\lambda}_2^i \leq 1-\gamma$, where $\gamma=(c-1)\epsilon\sqrt{r(1+\eta)}$.

Lemma 1 extends assumption A1 from Ng et al. (2001). Before we proceed to the next lemma, let us first introduce some more notation. We know that $\tilde{A}$ is defined as the affinity matrix that would give rise to graph Laplacian $\tilde{L}$ in case when no Nyström approximation was used and thus $\tilde{L}=I-\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}$ (in this case $\tilde{D}$ is the diagonal matrix whose (i,i)-element is the sum of $\tilde{A}$'s i^{th} row). Let $j \in S_i$, where $i \in \{1,2,\ldots,k\}$. Define the following: $d_{(j)} = \sum_{m=1}^{n}A_{jm}$, $\tilde{d}_{(j)} = \sum_{m=1}^{n}\tilde{A}_{jm}$, $d_j = \sum_{m\in S_i}A_{jm}$, $\tilde{d}_j = \sum_{m\in S_i}\tilde{A}_{jm}$. Notice that $d_{(j)} \geq d_j$. Let $\forall_{i\in\{1,2,\ldots,k\}}\tilde{d}^{*(i)} = \min_{j\in S_i}\tilde{d}_j$ and $\tilde{d}^* = \min_{i\in\{1,2,\ldots,k\}}\tilde{d}^{*(i)}$. Also, let $\forall_{i\in\{1,2,\ldots,k\}}\tilde{D}^{*(i)} = \max_{j\in S_i}\tilde{d}_{(j)}$ and $\tilde{D}^* = \max_{i\in\{1,2,\ldots,k\}}\tilde{D}^{*(i)}$. At

this point we will make a reasonable assumption that $\frac{\tilde{D}^*}{\tilde{d}^*}$ is a bounded positive constant. Assuming the dataset has balanced clusters (i.e., no cluster is significantly bigger/smaller than any other) and in particular the datasets have no outliers, this assumption will be naturally satisfied. Furthermore, let $\alpha_{S_{i_1}S_{i_2}} = \min_{j\in S_{i1}, l\in S_{i2}, i_1,i_2\in\{1,2,\ldots,k\}} \frac{\tilde{d}_j\tilde{d}_l}{\tilde{d}_{(j)}\tilde{d}_{(l)}}$ and let $\alpha = \min_{i_1,i_2\in\{1,2,\ldots,k\}} \alpha_{S_{i_1}S_{i_2}}$. Note that $\alpha \in (0,1]$ and in the ideal case $\alpha = 1$. We are now ready to state and prove Lemma 3.

Lemma 2. $\exists_{\mathcal{C}>0}\forall_{i=\{1,2,\ldots,k\}, j=\{1,2,\ldots,n_i\}} \tilde{d}_j \geq (\sum_{l=1}^{n_i} \tilde{d}_l)/(\mathcal{C}n_i)$.

Proof. Consider any $i \in \{1,2,\ldots,k\}$ and any $j,l \in S_i$. It is true that

$$\frac{\tilde{d}_j}{\sum_{l=1}^{n_i}\tilde{d}_l} \geq \frac{\tilde{d}^*}{\tilde{D}^* n_i} \geq \frac{1}{\mathcal{C}n_i}, \tag{5}$$

where $\mathcal{C} = \frac{\tilde{D}^*}{\tilde{d}^*}$ is a bounded positive constant as was already discussed before.

Lemma 2 extends assumption A4 from Ng et al. (2001).

Lemma 3. *With probability at least* $1-\delta$, $\forall_{i_1,i_2=\{1,2,\ldots,k\}, i_1\neq i_2} \sum_{j\in S_{i_1}}\sum_{l\in S_{i_2}} \frac{\tilde{A}_{jl}^2}{\tilde{d}_j\tilde{d}_l} \leq \epsilon_1$, *where* $\epsilon_1 = \epsilon\mathcal{C}^2 f\sqrt{1+\eta}$ *(*$f>1$ *is some constant).*

Proof. Let $j' = \arg\max_{j\in S_{i_1}} \sum_{l\in S_{i_2}} \|\tilde{L}_{jl} - L_{jl}\|^2$. We know that:

$$\sum_{j\in S_{i_1}}\sum_{l\in S_{i_2}} \|\tilde{L}_{jl} - L_{jl}\|^2 \leq \sum_{j\in S_{i_1}}\sum_{l\in S_{i_2}} \|\tilde{L}_{j'l} - L_{j'l}\|^2 \leq \frac{f}{n}\|\tilde{L} - L\|^2 \tag{6}$$

The left-hand side of Equation 6 can be further expressed as

$$\sum_{j\in S_{i_1}}\sum_{l\in S_{i_2}} |\frac{\tilde{A}_{jl}}{\sqrt{\tilde{d}_{(j)}\tilde{d}_{(l)}}} - \frac{A_{jl}}{\sqrt{d_{(j)}d_{(l)}}}|^2 = \sum_{j\in S_{i_1}}\sum_{l\in S_{i_2}} |\frac{\tilde{A}_{jl}}{\sqrt{\tilde{d}_{(j)}\tilde{d}_{(l)}}}|^2. \tag{7}$$

Combining this result with Equation 6 we have

$$\sum_{j\in S_{i_1}}\sum_{l\in S_{i_2}} \frac{\tilde{A}_{jl}^2}{\tilde{d}_{(j)}\tilde{d}_{(l)}} \leq \frac{f}{n}\|\tilde{L} - L\|_F. \tag{8}$$

Rewrite Equation 8 as:

$$\sum_{j\in S_{i_1}}\sum_{l\in S_{i_2}} \frac{\tilde{A}_{jl}^2}{\tilde{d}_j\tilde{d}_l}\frac{\tilde{d}_j\tilde{d}_l}{\tilde{d}_{(j)}\tilde{d}_{(l)}} \leq \frac{f}{n}\|\tilde{L} - L\|_F. \tag{9}$$

The left-hand side of Equation 9 is lower-bounded by $\alpha\sum_{j\in S_{i_1}}\sum_{l\in S_{i_2}} \frac{\tilde{A}_{jl}^2}{\tilde{d}_j\tilde{d}_l}$ and thus

$$\sum_{j\in S_{i_1}}\sum_{l\in S_{i_2}} \frac{\tilde{A}_{jl}^2}{\tilde{d}_j\tilde{d}_l} \leq \frac{1}{\alpha}\frac{f}{n}\|\tilde{L} - L\|_F. \tag{10}$$

Again, by Theorem 6 we can write that with probability at least $1-\delta$ the following holds:

$$\sum_{j\in S_{i_1}}\sum_{l\in S_{i_2}}\frac{\tilde{A}_{jl}^2}{\tilde{d}_j\tilde{d}_l}\leq\frac{f}{n}\frac{\epsilon n}{\alpha}\sqrt{1+\eta}\leq\epsilon f\left(\frac{\tilde{D}^*}{\tilde{d}^*}\right)^2\sqrt{1+\eta}=\epsilon\mathcal{C}^2 f\sqrt{1+\eta}, \tag{11}$$

where the last inequality comes from the fact that $\alpha\geq(\frac{\tilde{d}^*}{\tilde{D}^*})^2$.

Lemma 3 extends assumption A2 from Ng et al. (2001). Notice that exactly the same proof technique could be used to show that $\sum_{l\in S_{i_2}}\frac{\tilde{A}_{j'l}^2}{\tilde{d}_j\tilde{d}_l}\leq\frac{\epsilon_1}{n_{i_1}}$. This result will be used in the next lemma.

Define $\beta_{S_i}=\max_{j\in S_i,l\notin S_i}\frac{\tilde{d}_l}{\tilde{d}_j}$ and $\beta=\max_{i\in\{1,2,\ldots,k\}}\beta_{S_i}$. We can now proceed to the next lemma.

Lemma 4. *With probability at least* $1-\delta$, $\forall_{i=\{1,2,\ldots,k\},j\in S_i}\frac{\sum_{l:l\notin S_i}\tilde{A}_{jl}}{\tilde{d}_j}$ $\leq\epsilon_2(\sum_{l,m\in S_i}\frac{\tilde{A}_{lm}^2}{\tilde{d}_l\tilde{d}_m})^{-\frac{1}{2}}$, *where* $\epsilon_2=\mathcal{C}^{\frac{5}{2}}\sqrt{fk(1+\eta)}\sqrt{\epsilon(\epsilon g\sqrt{1+\eta}+2h-1)}$.

Proof. Consider any $i\in\{1,2,\ldots,k\}$ and $j\in S_i$. Let $j'=\arg\min_{j\in S_i}\sum_{l\in S_{i_2}}\|\tilde{L}_{jl}-L_{jl}\|^2$. We will consider the expression:

$$\left[\frac{\sum_{l:l\notin S_i}\tilde{A}_{jl}}{\tilde{d}_j}\right]\times\left[\sum_{l,m\in S_i}\frac{\tilde{A}_{lm}^2}{\tilde{d}_l\tilde{d}_m}\right]^{\frac{1}{2}} \tag{12}$$

The first term in the above expression can be upper-bounded by Jensen's inequality as follows

$$\frac{\sum_{l:l\notin S_i}\tilde{A}_{jl}}{\tilde{d}_j}=\sum_{l:l\notin S_i}\frac{\tilde{A}_{jl}}{\tilde{d}_j}=\left[\left(\sum_{l:l\notin S_i}\frac{\tilde{A}_{jl}}{\tilde{d}_j}\right)^2\right]^{\frac{1}{2}}\leq\left[n_i\sum_{l:l\notin S_i}\left(\frac{\tilde{A}_{jl}}{\tilde{d}_j}\right)^2\right]^{\frac{1}{2}}. \tag{13}$$

The right-hand side of Equation 13 can be rewritten and bounded as

$$\left[n_i\sum_{l:l\notin S_i}\left(\frac{\tilde{A}_{jl}^2}{\tilde{d}_j\tilde{d}_l}\times\frac{\tilde{d}_l}{\tilde{d}_j}\right)\right]^{\frac{1}{2}}\leq\left[n_i\beta\sum_{l:l\notin S_i}\left(\frac{\tilde{A}_{jl}^2}{\tilde{d}_j\tilde{d}_l}\right)\right]^{\frac{1}{2}}\leq\left[n_i\beta\sum_{l:l\notin S_i}\left(\frac{\tilde{A}_{j'l}^2}{\tilde{d}_j'\tilde{d}_l}\right)\right]^{\frac{1}{2}}\leq\sqrt{k\beta\epsilon_1}. \tag{14}$$

Combining these results together and applying Equation 11 we see that, with probability at least $1-\delta$,

$$\frac{\sum_{l:l\notin S_i}\tilde{A}_{jl}}{\tilde{d}_j}\leq\sqrt{k\beta\epsilon_1}. \tag{15}$$

Now focus on bounding the second term in Expression 12. Recall that

$$\sum_{l,m\in S_i}\|\tilde{L}_{lm}-L_{lm}\|^2\leq\frac{g}{n}\|\tilde{L}-L\|_F. \tag{16}$$

Similarly, as in previous paragraph, we can write that

$$\sum_{l,m\in S_i} \|\tilde{L}_{lm} - L_{lm}\|^2 = \sum_{l,m\in S_i} |\frac{\tilde{A}_{lm}}{\sqrt{\tilde{d}_{(l)}\tilde{d}_{(m)}}} - \frac{A_{lm}}{\sqrt{d_{(l)}d_{(m)}}}|^2$$
$$= \sum_{l,m\in S_i} |\frac{\tilde{A}_{lm}}{\sqrt{\tilde{d}_{(l)}\tilde{d}_{(m)}}} - \frac{1}{n_i}|^2, \tag{17}$$

The last equality uses the fact that $\forall_{l,m\in S_i} A_{lm} = 1$ and $d_{(l)} = d_{(m)} = d_l = d_m = n_i$. We can then expand the right-hand side of Equation 17:

$$\sum_{l,m\in S_i} |\frac{\tilde{A}_{lm}}{\sqrt{\tilde{d}_{(l)}\tilde{d}_{(m)}}} - \frac{1}{n_i}|^2 = \sum_{l,m\in S_i} \left(\frac{\tilde{A}^2_{lm}}{\tilde{d}_{(l)}\tilde{d}_{(m)}} - \frac{2\tilde{A}_{lm}}{n_i\sqrt{\tilde{d}_{(l)}\tilde{d}_{(m)}}} + \frac{1}{n_i^2} \right)$$

$$= \sum_{l,m\in S_i} \left(\frac{\tilde{A}^2_{lm}}{\tilde{d}_l\tilde{d}_m} \times \frac{\tilde{d}_l\tilde{d}_m}{\tilde{d}_{(l)}\tilde{d}_{(m)}} - \frac{2\tilde{A}_{lm}}{n_i\sqrt{\tilde{d}_{(l)}\tilde{d}_{(m)}}} \right) + 1. \tag{18}$$

Equation 18 can be lower-bounded as:

$$\sum_{l,m\in S_i} \left(\frac{\tilde{A}^2_{lm}}{\tilde{d}_l\tilde{d}_m} \times \frac{\tilde{d}_l\tilde{d}_m}{\tilde{d}_{(l)}\tilde{d}_{(m)}} - \frac{2\tilde{A}_{lm}}{n_i\sqrt{\tilde{d}_{(l)}\tilde{d}_{(m)}}} \right) + 1 \geq \alpha \sum_{l,m\in S_i} \frac{\tilde{A}^2_{lm}}{\tilde{d}_l\tilde{d}_m} - 2\frac{n_i}{\tilde{n}_i} + 1$$

$$\geq \alpha \sum_{l,m\in S_i} \frac{\tilde{A}^2_{lm}}{\tilde{d}_l\tilde{d}_m} - 2h + 1 \tag{19}$$

Combining Equation 16 and 19 we obtain:

$$\sum_{l,m\in S_i} \frac{\tilde{A}^2_{lm}}{\tilde{d}_l\tilde{d}_m} \leq \frac{1}{\alpha}(\|\tilde{L} - L\|_F + 2h - 1). \tag{20}$$

After applying Theorem 6 we obtain that with probability at least $1 - \delta$

$$\sum_{l,m\in S_i} \frac{\tilde{A}^2_{lm}}{\tilde{d}_l\tilde{d}_m} \leq \frac{1}{\alpha}(\frac{g}{n}\epsilon n\sqrt{1+\eta} + 2h - 1). \tag{21}$$

Combining Equation 15 and 21 we get the following:

$$\left[\frac{\sum_{l:l\notin S_i} \tilde{A}_{jl}}{\tilde{d}_j}\right] \times \left[\sum_{l,m\in S_i} \frac{\tilde{A}^2_{lm}}{\tilde{d}_l\tilde{d}_m}\right]^{\frac{1}{2}} \leq \sqrt{k\beta\epsilon_1 \times \frac{1}{\alpha}(\epsilon g\sqrt{1+\eta} + 2h - 1)}$$

$$\leq \mathcal{C}^{\frac{5}{2}}\sqrt{fk(1+\eta)}\sqrt{\epsilon(\epsilon g\sqrt{1+\eta} + 2h - 1)} \tag{22}$$

where the last inequality uses the fact that $\alpha \geq (\frac{\tilde{d}^*}{\tilde{D}^*})^2$ and $\beta \leq \frac{\tilde{D}^*}{\tilde{d}^*}$.

5 Experiments

To evaluate the proposed algorithms empirically, we consider the four datasets described in Ng et al. (2001). We used a Gaussian kernel to build the affinity matrix ($\kappa(s_i, s_j) = \exp(-\|s_i - s_j\|^2/2\sigma^2)$). The parameters σ and r were manually tuned to obtain the best performance. Figure 2 shows the datasets with plots of the error versus the percent of the columns sampled (l/n). We used uniform sampling without replacement throughout. Note that both the choice of columns as well as the initialization of the k-means clustering algorithm[2] slightly affect the performance. Thus, we show two types of results: the curves in the second row on Figure 2 obtained by averaging over 10,000 runs and the curves underneath showing the most frequently obtained performance (i.e. the median case). Also we performed two sets of experiments where r was held constant as well as where r was tuned for each value of l. In the first case, we set $r = \tau$ (the value of τ for each dataset is provided under Figure 2) and when $l \leq \tau$ we set $\tau = l$. In the second case, we observed that tuning r for each value of l (when l increases, r should decrease) can improve the performance but the improvement is relatively small and not worth presenting here.

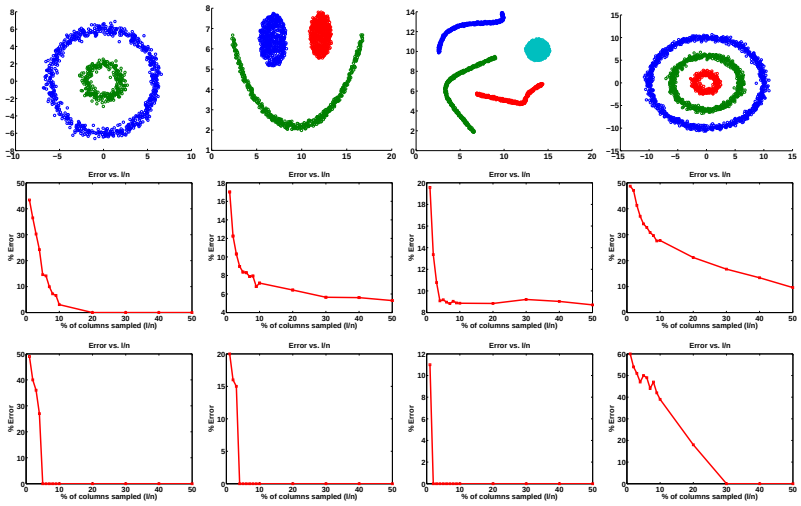


Fig. 2. Top row: the datasets with color-coded clusters. Second Row: error curves vs % of columns sampled with the error averaged over 10,000 runs. Third row: error curves vs % of columns sampled with the most frequent result being displayed. The parameters of interest for each experiment (from left to right) were: a) $n = 1000$; $\sigma = 1$; $\tau = 50$, b) $n = 1500$; $\sigma = 1$; $\tau = 20$, c) $n = 2000$; $\sigma = 1$; $\tau = 50$, d) $n = 2000$; $\sigma = 1$; $\tau = 50$.

[2] There was no significant difference in the choice of the distortion-minimizing algorithm we use in the last step of our spectral clustering algorithm, be it Lloyd's algorithm, k-means++ and k-means#.

Acknowledgments. The authors thank Sanjiv Kumar for helpful suggestions.

References

Belkin, M. and Niyogi, P. (2007). Convergence of Laplacian eigenmaps. In *NIPS 2006*, pages 129–136. MIT Press.

Drineas, P. and Mahoney, M. W. (2005). On the Nyström Method for Approximating a Gram Matrix for Improved Kernel-Based Learning. *Journal of Machine Learning Research*, 6:2005.

Fowlkes, C., Belongie, S., Chung, F., and Malik, J. (2004). Spectral grouping using the nyström method. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(2):214–225.

Fung, W. S., Hariharan, R., Harvey, N. J., and Panigrahi, D. (2011). A general framework for graph sparsification. In *STOC*.

Kannan, R. and Vempala, S. (2009). Spectral algorithms. *Foundations and Trends in Theoretical Computer Science*, 4(3-4):157–288.

Kumar, S., Mohri, M., and Talwalkar, A. (2009). Sampling techniques for the nyström method. *Journal of Machine Learning Research*, 5:304–311.

Lashkari, D. and Golland, P. (2007). Convex clustering with exemplar-based models. In *NIPS 2007*.

Li, M., Lian, X.-C., Kwok, J. T., and Lu, B.-L. (2011). Time and space efficient spectral clustering via column sampling. In *24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011*, pages 2297–2304. IEEE.

Lloyd, S. P. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137.

Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In *NIPS 2001*, pages 849–856. MIT Press.

Spielman, D. A. and Teng, S.-H. (2011). Spectral sparsification of graphs. *SIAM Journal on Computing*, 40(4):981–1025.

Williams, C. and Seeger, M. (2001). Using the Nyström method to speed up kernel machines. In *NIPS 2000*, pages 682–688. MIT Press.

Yan, D., Huang, L., and Jordan, M. I. (2009). Fast approximate spectral clustering. In *ACM SIGKDD*, pages 907–916. ACM.