Beyond the Nyström Approximation: Speeding up Spectral Clustering using Uniform Sampling and Weighted Kernel k-means

Mahesh Mohan, Claire Monteleoni
George Washington University, Washington D.C.
mahesh_mohan@gwu.edu, cmontel@gwu.edu

Abstract

In this paper we present a framework for spectral clustering based on the following simple scheme: sample a subset of the input points, compute the clusters for the sampled subset using weighted kernel k-means (Dhillon et al. 2004) and use the resulting centers to compute a clustering for the remaining data points. For the case where the points are sampled uniformly at random without replacement, we show that the number of samples required depends mainly on the number of clusters and the diameter of the set of points in the kernel space. Experiments show that the proposed framework outperforms the approaches based on the Nyström approximation both in terms of accuracy and computation time.

1 Introduction

Clustering is one of the fundamental problems in machine learning. Spectral clustering techniques [Von Luxburg, 2007] have been shown to outperform other clustering methods, such as k-means or single-linkage algorithms. However, two significant obstacles to scaling up spectral clustering to large datasets include building an affinity matrix between pairs of data points and computing the eigenvectors of the resulting normalized graph Laplacian, both of which become computationally prohibitive for large data-sets.

Approaches to scaling up spectral clustering typically focus on improving the efficiency of computing the eigenvectors using various matrix approximation schemes. A popular technique for matrix approximation is the Nyström method [Williams and Seeger, 2001] which constructs a low rank approximation by sampling a subset of the columns of the affinity matrix. Various sampling schemes have been proposed in the literature to sample this subset of columns [Kumar *et al.*, 2012; Gittens and Mahoney, 2013], offering a tradeoff between the complexity of the sampling distribution and quality of the approximation. However other than uniform sampling, the practicality of these approaches for massive datasets may be limited [Yan *et al.*, 2009]. The major disadvantage of these methods is that if l columns are sampled, they involve the computation of eigenvectors of a sub-matrix of size $l \times l$, which can itself be computationally expensive when l is large. Also, the subsequent k-means step used for rounding the cluster assignment has to be performed for all the input points.

The connection between spectral clustering and weighted kernel k-means was first introduced in [Dhillon *et al.*, 2004]. In this paper, we use this connection to propose a simple framework for spectral clustering that samples a subset S of the input points, computes the clusters for S using weighted kernel k-means and uses the resulting centers to compute a clustering for rest of the data points. The proposed framework does not require computation of the entire affinity matrix and does not require the computation of eigenvectors of the selected sub-matrix. Sampling algorithms developed for the Nyström approximation can be leveraged to obtain a tradeoff between accuracy and computational efficiency. In fact, our experiments show that the use of sampling followed by weighted kernel k-means outperforms sampling followed by the Nyström approximation. Since the proposed method uses weighted kernel k-means to compute the clusters, it is easy to parallelize. The contributions of this paper are:

- Prior work on the Nyström approximation such as [Gittens and Mahoney, 2013; Talwalkar and Rostamizadeh, 2010] provide approximation guarantees with respect to various matrix norms, but not with respect to the spectral clustering objective itself. We show that matrix norm errors are not always indicative of the quality of spectral clustering.
- We propose a framework for spectral clustering that applies sampling followed by the weighted kernel k-means algorithm instead of approximating the affinity matrix or the Laplacian.
- Without making any assumptions about the data, we show that when points are sampled uniformly at random without replacement, the following theorem holds.

Theorem 1. *Let $0 < \delta < 1$, $\alpha \geq 1$, $0 < \beta < 1$ and $\epsilon > 0$ be approximation parameters. Let $\mathcal{A}$ be an α-approximation algorithm for the weighted kernel k-means problem. Given a set of n points, V, suppose we sample a subset $S \subset V$ of size s uniformly at random without replacement such that,*

$$s \geq ln\left(\frac{1}{\delta}\right)\left(1+\frac{1}{n}\right) / \left(\frac{2\beta^2\epsilon^2}{\Delta^2\alpha^2} + \frac{1}{n} ln\left(\frac{1}{\delta}\right)\right)$$

where $\Delta = \max_{i,j} w_i w_j \|\phi(a_i) - \phi(a_j)\|^2$. If we run algorithm $\mathcal{A}$ with input S, then for the solution C^* obtained, with *probability at least* $1-\delta$*,*

$$NCut(G, C^*) \leq 4(\alpha+\beta)NCut(G, C_{opt}) + \epsilon$$

In other words, by choosing sufficient number of samples s, we can obtain centers that provide a good approximation to the optimal cluster centers. As the input size n increases, the term $(1/n)$ vanishes, resulting in a bound for s that is independent of n. To the best of our knowledge, the relation between the number of samples and the spectral clustering objective studied in the proposed work has not been explored in the literature.

2 Preliminaries

2.1 Weighted Kernel k-means

Given a set of n points $V = \{v_1, v_2..., v_n\}$ with associated weights $\{w_1, w_2...w_n\}$ and a kernel matrix K, the weighted kernel k-means objective for clusters $V_1, V_2, ...V_k$ with centers $C = \{C_1, C_2, ..C_k\}$ is defined as:

$$\mathbb{W}(V, C) = \sum_{i=1}^{k} \sum_{v_j \in V_i} w_j \|\phi(v_j) - C_i\|^2$$

where $C_i = \frac{\sum_{v_j \in V_i} w_j \phi(v_j)}{\sum_{v_j \in V_i} \|w_j\|}$. Here $\phi(v)$ is the kernel function that maps the point v to a higher dimensional feature space. It was shown in [Dhillon *et al.*, 2004],

$$\|\phi(v_i) - C_j\|^2 \quad = \quad K_{ii} - \frac{2\tau_1(i,j)}{deg(V_j)} + \frac{\tau_2(V_j)}{(deg(V_j))^2}$$

where $deg(V_i) = \sum\limits_{v_j \in V_i} w_j$, $\tau_1(i,j) = \sum\limits_{v_l \in V_j} w_l K_{il}$ and $\tau_2(V_j) = \sum\limits_{v_l, v_m \in V_j} w_l w_m K_{lm}$.

Definition 1. *Let $\mathcal{A}$ be an approximation algorithm for the weighted kernel k-means objective. Let $\alpha \geq 1$. $\mathcal{A}$ is an α-approximation algorithm if the set of centers, C, returned by $\mathcal{A}$ satisfies,*

$$\mathbb{W}(V, C) \leq \alpha \mathbb{W}(V, C_{opt})$$

where C_{opt} is the set of optimal centers.

2.2 Spectral Clustering Using Norm-Cuts

In spectral clustering, we are given a graph $G = (V, A)$, which is made up of a set of n vertices V. The affinity matrix A is $n \times n$ whose entries represent the similarity between vertices. If V_1, V_2 are subsets of V, let $links(V_1, V_2) = \sum_{i \in V_1, j \in V_2} A_{ij}$.

Furthermore, let $degree(V_1) = links(V_1, V)$. The graph partitioning problem seeks to partition the graph into k disjoint clusters $V_1, ..., V_k$. A number of different graph partitioning objectives have been proposed and studied. In this paper, we will focus on the normalized cut objective. The goal is to minimize the following objective over all possible clusterings $\{V_1..V_k\}$,

$$NCut(G, \{V_1..V_k\}) = \sum_{i=1}^{k} \frac{links(V_i, V \backslash V_i)}{degree(V_i)}$$

2.3 Relation Between Weighted Kernel k-means And Spectral Clustering

To convert a spectral clustering problem to a weighted kernel k-means problem, it was shown in [Dhillon *et al.*, 2004] that we can set $W = D$ and $K = \sigma D^{-1} + D^{-1}AD^{-1}$, where $\sigma = k/(n-k)$. Here the term σD^{-1} is added to ensure K is positive definite and does not change the optimal clustering.

We note that a set of centers C for the weighted kernel k-means problem induces a clustering $V_1...V_k$ for the spectral clustering problem. Since both objectives are equivalent, we use $\mathbb{W}(V, C)$, $NCut(G, \{V_1..V_k\})$ and $NCut(G, C)$ interchangeably throughout the remainder of this paper. We also use $NCut(G, k) = \min_C NCut(G, C)$ to represent the optimal partitioning of G into k clusters.

3 Why Matrix Norms Are Insufficient For Norm Cuts-based Spectral Clustering

The Nyström approximation has been extensively studied with respect to various matrix error norms, such as the Frobenius norm, trace norm and the spectral norm [Gittens and Mahoney, 2013; Talwalkar and Rostamizadeh, 2010]. In this section, we show that approximation of these norms is neither necessary nor sufficient for producing a good clustering with respect to the norm cuts objective. For simplicity, we focus on the trace norm error. However the examples can be extended to any matrix norm. Also, for this section, we let $NCut(A, k)$ represent the optimal partitioning of the graph with affinity matrix A into k clusters.

Lemma 1. *Preserving trace norm is not sufficient for an approximation to be good with respect to the norm cuts objective.*

We construct two matrices that are extremely similar in terms of their trace norms, but have a significantly different norm cuts objective. Here, the trace norm of a matrix A is denoted as $\|A\|_* = \sum_{i=1}^{n} |\lambda_i(A)|$, where $\lambda_i(A)$ is the ith eigenvalue of A.

In other words, we show that for any value of $\epsilon > 0$, there exists n, k, A and B such that, A and B represent graphs G_A, G_B with n vertices, and $\|A - B\|_* \leq (1+\epsilon)\|A\|_*$ and $NCut(A, k) = O(k) \cdot NCut(B, k)$. Note, that the maximum value of $NCut(A, k)$ is k.

Let A and B be the affinity matrices for undirected, unweighted graphs that have n vertices and k and $k-1$ equal sized components, respectively. Further, we assume that each component is completely connected with no self loops. Since each component in A is a complete graph on n/k vertices, its eigenvalues are -1 with multiplicity $(n/k)-1$ and $(n/k)-1$ with multiplicity 1. Since each component in A is regular, the eigen values of the normalized symmetric Laplacian of A (denoted by $L(A)$) are given as, $\lambda_i(L(A)) = 1 - (1/\Delta)\lambda_i(A)$ where $\Delta = (n/k) - 1$. In other words, $L(A)$ has eigenvalues $n/(n-k)$ with multiplicity $n-k$ and 0 with multiplicity k. Similarly, $L(B)$ has eigenvalues $n/(n-k+1)$ with multiplicity $n-k+1$ and 0 with multiplicity $k-1$. A more general statement in this context can be found in [Butler, 2007].

Now we compare the trace norm of the difference of the

Algorithm 1 Proposed Framework for Spectral Clustering

```
Input: Affinity Matrix $A$, number of clusters $k$, number of samples $s$
Output: Matrix $\hat{Y}$ ($\hat{Y}_{ij} \neq 0$ only if point $i$ belongs to cluster $j$)
Procedure:
$sub \leftarrow sample(s)$
$A_{sub} \leftarrow A(sub, sub)$
$Y \leftarrow weighted\_kernel\_kmeans(A_{sub}, k)$
$\hat{Y} \leftarrow diffuse(Y, A, sub)$
```

two Laplacians, $L(A)$ and $L(B)$, as follows,

$$\begin{aligned} \|L(A)\|_* - \|L(B)\|_*^2 &= \sum_{i=1}^{n-k} |\lambda_i(L(A)) - \lambda_i(L(B))| \\ &= \sum_{i=1}^{n-k} \left(\frac{n}{n-k} - \frac{n}{n-k'} \right) + \frac{n}{n-k'} \\ &= \frac{2n}{n-k+1} \end{aligned}$$

where, $k' = k - 1$. Given a value of ϵ, we can choose appropriate values of n, k such that $\|L(A)\|_*^2 - \|L(B)\|_*^2 \leq (1+\epsilon)\|L(A)\|_*^2$. Now we examine the norm cuts ratio when we try to partition these graphs into exactly k clusters. Thus, $NCut(A, k) = 0$, since the graph has exactly k components. However for B, one of the $k-1$ components will have to be split into two equal parts to minimize the norm cuts ratio. This results in a cut that involves $\frac{n}{k-1}(\frac{n}{k-1}+1)/2$ edges. Thus the norm cuts ratio for B is given as,

$$\begin{aligned} NCut(B, k) &= k - \frac{\left(n(n+2k-2)/(k+1)^2\right)}{n(n+k-1)/2(k+1)^2} \\ &= k - \frac{n+2k-2}{4(n+k-1)} \end{aligned}$$

The second term reduces to a constant for a sufficiently large value of n, resulting in a norm cuts ratio of $O(k)$.

Lemma 2. *Preserving trace norm is not necessary for an approximation to be good with respect to the norm cuts objective.*

Proof. To show that preserving matrix norms is not necessary for an approximation to be good with respect to the norm cuts objective, we consider the case of a block diagonal matrix A with k blocks. Let $L(A)$ be the corresponding normalized Laplacian. The result of norm cuts based spectral clustering will depend on the eigenvectors of $L(A)$ and the relative order of the corresponding eigenvalues. We note that $L(A)$ has the same eigenvectors as $L(A)^2$, $L(A)^3$, etc. Thus, $L(A)$, $L(A)^2$ return the same spectral clustering. However $\|L(A)\|_* - \|L(A)^i\|_*$ can be arbitrarily high. □

4 Proposed Framework

The proposed framework for spectral clustering is described in Algorithm 1. The overall procedure consists of three stages. In the first stage, a subset of the input points are sampled according to some distribution. This subset is indicated as sub. In the second stage, the clusters are computed using weighted kernel k-means. However, instead of using the entire affinity matrix, as in [Dhillon *et al.*, 2004], we only use the principal sub-matrix induced by indices in sub. The final step involves using the cluster centers returned by the weighted kernel k-means procedure to assign clusters to the remaining data points. The procedure for this step is described in Algorithm 2.

Algorithm 2 Diffuse

```
Input: Matrix $\hat{Y}$, Affinity Matrix $A$, set of sampled indices $S$
Output: Matrix $Y$ ($Y_{ij} \neq 0$ only if point $i$ belongs to cluster $j$)
Procedure:
for each $i \in S$ do
  for j = 1:k do
    $Y_{ij} \leftarrow \hat{Y}_{ij}$
    if $Y_{ij} \neq 0$ then
      $\pi_j \leftarrow \pi_j \cup a_i$
    end if
  end for
end for
for each $i \notin S$ do
  for $c = 1 : k$ do
    $d(a_i, m_c) \leftarrow K_{ii} - \frac{2 \sum_{a_j \in \pi_c} w_j K_{ij}}{\sum_{a_j \in \pi_c} w_j} + \frac{\sum_{a_j, a_l \in \pi_c} w_j w_l K_{jl}}{(\sum_{a_j \in \pi_c} w_j)^2}$
  end for
  $j \leftarrow \arg\min_{c=1}^{k} d(a_i, m_c)$
  $Y_{ij} \leftarrow 1$
end for
```

This is scalable to larger datasets because it does not require explicit computation of the eigenvectors. Also it only requires the computation of a sub-matrix of size $n \times k$, which can result in significant savings in memory. Weighted kernel k-means is extremely easy to parallelize.

5 Analysis

In this section, we present the proof of Theorem 1. We extend the analysis in [Czumaj and Sohler, 2010] to the case of sampling without replacement. Specifically, uniform sampling without replacement allows us to use the sharper bound offered by the Serfling inequality stated below. We refer the reader to [Bardenet *et al.*, 2015] for a more thorough discussion.

Lemma 3. *[Serfling, 1974; Bardenet* et al.*, 2015] Let $X_1..X_s$ be random variables sampled uniformly without replacement.*

Let $X = X_1 + ... + X_n$ and $0 \leq X_i \leq \Delta$. Then for all $\gamma > 0$,

$$P\left[\sum_{t=1}^{s} \frac{X_t - E[X_t]}{s} \geq \gamma\right] \leq \exp\left(-\frac{2s\gamma^2}{(1-(s-1)/n)\Delta^2}\right)$$

Definition 2. *Let $\beta > 0$ and $\alpha \geq 1$. A set of k centers C is a β-bad, α-approximation if*

$$\mathbb{W}(V, C) > (\alpha + \beta)\mathbb{W}(V, C_{opt})$$

If not, C is said to be a β-good, α-approximation.

We will need the following lemmas.

Lemma 4. *Let S be a subset of V of size s such that,*

$$s \geq \ln\left(\frac{1}{\delta}\right)\left(1 + \frac{1}{n}\right) / \left(\frac{2\beta^2 m^2}{\Delta^2\alpha^2} + \frac{\ln(1/\delta)}{n}\right)$$

If an α-approximation algorithm for weighted kernel k-means $\mathcal{A}$ is run on input S, then for the set of centers, $C_{\mathcal{A}}$, obtained, with probability $\geq 1 - \delta$,

$$\mathbb{W}(S, C_{\mathcal{A}}) \leq 4(\alpha + \beta)\frac{s}{n}\mathbb{W}(V, C_{opt})$$

where C_{opt} is the set of optimal centers for V and $m = \mathbb{W}(V, C_{opt})/n$.

Proof. For each point $v_i \in S$, let $X_i = w_i\|\phi(v_i) - C_i\|^2$ where C_i the nearest center to $\phi(v_i)$ in C_{opt}. Then $\mathbb{W}(S, C_{opt}) = \sum\limits_{i=1}^{s} X_i$. We also know that $E[X_i] = m$. Let X_i be bounded in the interval $[0, \Delta]$.

$$\begin{aligned} & Pr[\mathbb{W}(S, C_{opt}) > (1 + \frac{\beta}{\alpha})\frac{s}{n}\mathbb{W}(V, C_{opt})] \\ = \quad & Pr\left[\sum X_i > (1 + \frac{\beta}{\alpha})sm)\right] \\ \leq \quad & e^{-\frac{2s\cdot\frac{\beta^2}{\alpha^2}\cdot m^2}{(1-s/n)\Delta^2}} \end{aligned}$$

The last step is obtained by applying the Serfling inequality. Note, $e^{-\frac{2s\cdot\frac{\beta^2}{\alpha^2}\cdot m^2}{(1-s/n)\Delta^2}} \leq \delta$ if

$$s \geq \ln\left(\frac{1}{\delta}\right)\left(1 + \frac{1}{n}\right) / \left(\frac{2\beta^2 m^2}{\Delta^2\alpha^2} + \frac{\ln(1/\delta)}{n}\right)$$

Let C' be the set of k centers in S obtained by replacing each $C \in C_{opt}$ by its nearest neighbor in S. From the weakened triangle inequality, it can be shown that $\mathbb{W}(S, C') \leq 4\mathbb{W}(S, C_{opt})$. This implies with probability at least $1 - \delta$, S contains k centers with cost at most $4(1 + \frac{\beta}{\alpha})\frac{s}{n}\mathbb{W}(V, C_{opt})$. In other words,

$$\mathbb{W}(S, C') \leq 4(1 + \frac{\beta}{\alpha})\frac{s}{n}\mathbb{W}(V, C_{opt})$$

If C'' is the set of optimum centers for $\mathbb{W}(S, C)$, then

$$\mathbb{W}(S, C'') \leq \mathbb{W}(S, C')$$

Since $\mathcal{A}$ is an α-approximation algorithm,

$$\mathbb{W}(S, C_{\mathcal{A}}) \leq \alpha\mathbb{W}(S, C'')$$

This implies,

$$\mathbb{W}(S, C_{\mathcal{A}}) \leq \alpha\mathbb{W}(S, C')$$

The lemma now follows. □

Lemma 5. *Let S be a subset of V of size s, sampled uniformly at random such that,*

$$s \geq ln\left(\frac{1}{\delta}\right)\left(1 + \frac{1}{n}\right) / \left(\frac{\beta^2 m^2}{\Delta^2} + \frac{1}{n}ln\left(\frac{1}{\delta}\right)\right)$$

Let χ be the set of 12β-bad 4α-approximations of $\mathbb{W}(V, C_{opt})$ then,

$$Pr[C_b \subset S \textit{ and } C_b \in \chi \textit{ and } \mathbb{W}(S, C_b) \leq 4(\alpha + \beta)m] \leq \delta$$

Proof. Let $s \geq \frac{4\alpha+5\beta}{\beta}k$. Let $C_b \in \chi$ be a 12β-bad 4α-approximation of $\mathbb{W}(V, C_{opt})$. If $C_b \subset S$, let S^* be the subset of S with C_b removed, such that $|S^*| = s - k$. Thus,

$$\begin{aligned} & Pr[C_b \subset S \text{ and } \frac{\mathbb{W}(S, C_b)}{s} \leq 4(\alpha + \beta)m] \\ & = Pr\left[\frac{\mathbb{W}(S, C_b)}{s} \leq 4(\alpha + \beta)m \;\middle|\; C_b \subset S\right] \cdot Pr[C_b \subset S] \\ & \leq Pr\left[\frac{\mathbb{W}(S^*, C_b)}{s-k} \leq 4\frac{s}{s-k}(\alpha + \beta)m\right] \cdot Pr[C_b \subset S] \\ & \leq Pr\left[\frac{\mathbb{W}(S^*, C_b)}{s-k} \leq (4\alpha + 5\beta)m\right] \cdot Pr[C_b \subset S] \qquad (1) \end{aligned}$$

The second step holds because the mean remains unchanged at each step when the elements are chosen uniformly at random without replacement and $\mathbb{W}(S, C_b) = \mathbb{W}(S^*, C_b)$ when $C_b \subset S$. The third step holds since $s \geq \frac{4\alpha+5\beta}{\beta}k$.

Let Y_i denote the distance of point $v_i \in S$ from its closest center in C_b. Since $C_b \in \chi$, $E[Y_i] \geq (4\alpha + 12\beta)m$. Thus,

$$Pr\left[\frac{\mathbb{W}(S^*, C_b)}{s-k} \leq (4\alpha + 5\beta)m\right]$$

$$\begin{aligned} = \quad & Pr\left[\sum_{i=1}^{s-k} Y_i \leq (4\alpha + 5\beta)m\right] \\ \leq \quad & Pr\left[\sum_{i=1}^{s-k} Y_i \leq \frac{4\alpha + 5\beta}{4\alpha + 12\beta}E[\sum_{i=1}^{s-k} Y_i]\right] \\ \leq \quad & Pr\left[\sum_{i=1}^{s-k} Y_i \leq \left(1 - \frac{7\beta}{4\alpha + 12\beta}\right)E[\sum_{i=1}^{s-k} Y_i]\right] \\ \leq \quad & Pr\left[\sum_{i=1}^{s-k} Y_i - E[Y_i] \leq \left(-\frac{7\beta}{4\alpha + 12\beta}\right)E[\sum_{i=1}^{s-k} Y_i]\right] \\ \leq \quad & \exp\left(-2(s-k)\frac{\left(\frac{7\beta E[\sum_{i=1}^{s-k} Y_i]}{4\alpha+12\beta}\right)^2}{(1-(s-1)/n)\Delta^2}\right) \qquad (2) \\ \leq \quad & \exp\left(-2(s-k)\frac{49\beta^2 m^2}{(1-(s-1)/n)\Delta^2}\right) \end{aligned}$$

Similar to the previous lemma, we have used the Serfling bound to obtain (2). Thus it follows that,

$$Pr\left[\frac{\mathbb{W}(S^*, C_b)}{s-k} \leq (4\alpha + 5\beta)m\right] \leq e^{-\frac{s\beta^2 m^2}{(1-(s-1)/n)\Delta^2}} \qquad (3)$$

We note, $Pr[C_b \subset S] \leq (s/n)^k$ and $|\chi| \leq n^k$. Using this and (3) in (1), we have,

$$
\begin{aligned}
&Pr[C_b \subset S \text{ and } C_b \in \chi \text{ and } \frac{\mathbb{W}(S, C_b)}{s} \leq 4(\alpha+\beta)m] \\
&\leq \sum_{C_b \in \chi} Pr[C_b \subset S \text{ and } \frac{\mathbb{W}(S, C_b)}{s} \leq 4(\alpha+\beta)m] \\
&\leq \sum_{C_b \in \chi} Pr\left[\frac{\mathbb{W}(S^*, C_b)}{s-k} \leq (4\alpha+5\beta)m\right] \cdot (s/n)^k \\
&\leq n^k \cdot e^{-\frac{s\beta^2 m^2}{(1-(s-1)/n)\Delta^2}} \cdot (s/n)^k \\
&\leq s^k \cdot e^{-\frac{s\beta^2 m^2}{(1-(s-1)/n)\Delta^2}}
\end{aligned}
$$

which is upper bounded by δ when,

$$
s \geq ln\left(\frac{1}{\delta}\right)\left(1+\frac{1}{n}\right) / \left(\frac{\beta^2 m^2}{\Delta^2} + \frac{1}{n} ln\left(\frac{1}{\delta}\right)\right)
$$

□

Next, we use these lemmas to prove Theorem 1.

Proof. Let β^* be some constant. Let $S \subseteq V$ be a subset of s points chosen uniformly at random such that,

$$
s \geq ln\left(\frac{1}{\delta}\right)\left(1+\frac{1}{n}\right) / \left(\frac{2(\beta^*)^2 m^2}{\Delta^2 \alpha^2} + \frac{1}{n} ln\left(\frac{1}{\delta}\right)\right)
$$

Then from Lemma 3, we have, if $C \subseteq S$ and $C \in \chi$, with probability at least $1-\delta$,

$$
\mathbb{W}(S, C) > 4(\alpha+\beta^*)m
$$

On the other hand, if we run algorithm $\mathcal{A}$ for set S, then by Lemma 2, the resulting set of centers satisfies, with probability at least $1-\delta$,

$$
\mathbb{W}(S, C) \leq 4(\alpha+\beta^*)m
$$

This implies with probability at least $1-2\delta$ the set C is a (12β)-good (4α)-approximation. In order to remove the dependence on the number of samples on m,

- If $m \leq \epsilon$, we simply set $\beta = (1/\sqrt{2})\epsilon/m$. In this case we will obtain the following bound on s.

$$
s \geq ln\left(\frac{1}{\delta}\right)\left(1+\frac{1}{n}\right) / \left(\frac{\epsilon^2}{\Delta^2 \alpha^2} + \frac{1}{n} ln\left(\frac{1}{\delta}\right)\right)
$$

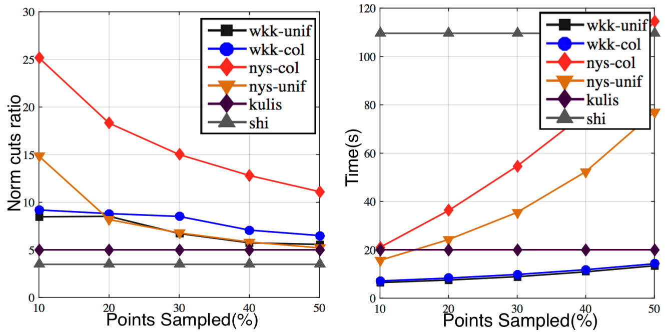
- If $m > \epsilon$, we set $\beta = 3\beta^*$ to obtain,

$$
s \geq ln\left(\frac{1}{\delta}\right)\left(1+\frac{1}{n}\right) / \left(\frac{2\beta^2 \epsilon^2}{\Delta^2 \alpha^2} + \frac{1}{n} ln\left(\frac{1}{\delta}\right)\right)
$$

The theorem now follows. □

Name	Instances	Attributes	Classes
Aggregation (Agg)	788	2	7
Flame (Fla)	240	2	2
A.K‘s toy problem (AK)	373	2	2

Table 1: The synthetic datasets used in our experiments.



(a) (b)

Figure 2: Figure (a) shows the norm cuts ratio, whereas (b) shows the time taken by different methods for a graph with 230400 nodes. This shows that methods using weighted kernel k-means outperform their Nyström counterparts with respect to computation time, while obtaining a comparable norm cuts ratio.

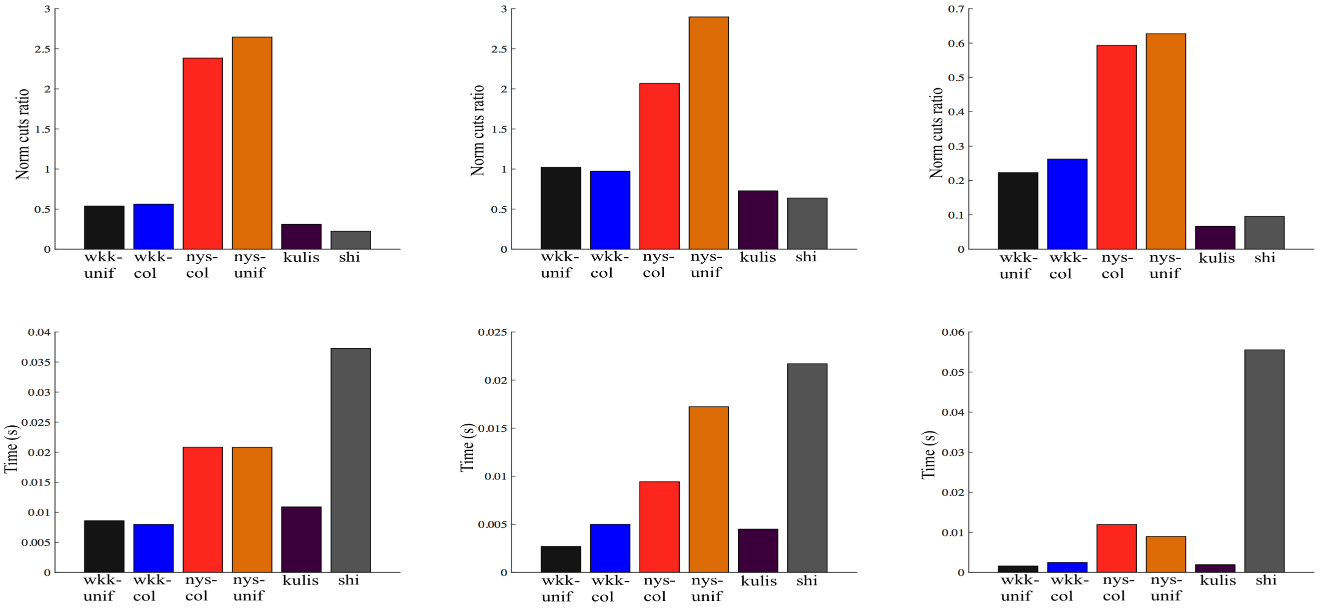
6 Experiments

We now present experimental results to show that the use of sampling followed by weighted kernel k-means outperforms sampling followed by the Nyström approximation. We evaluated the performance of two variants of the proposed framework. The first variant uses uniform sampling without replacement to sample a subset of the input points (denoted as wkk-unif). The second variant samples points based on the norm of the corresponding column in the affinity matrix (denoted as wkk-col). These results were compared against approaches described in [Dhillon *et al.*, 2004] (denoted as kulis) and [Yu and Shi, 2003] (denoted as shi). Both of these methods use the entire affinity matrix and do not perform any sampling. We also compare against the Nyström approximation when the columns are sampled uniformly at random (denoted as nys-unif) and sampled based on their column norm (denoted as nys-col).

6.1 Synthetic Data

We evaluated the norm cuts ratio and the computation time on six commonly used synthetic datasets [Bouneffouf and Birol, 2015], described in Table 1, and repeated our evaluations 10 times. We measured the clustering quality of each algorithm using the average accuracy across different datasets.

The results are shown in Figure 1. The first row shows the norm cuts ratio obtained by various approaches. It can be seen that norm cuts ratio for wkk-unif is comparable to kulis. This supports the result presented in Theorem 1. Furthermore, the second row in Figure 1 shows the time taken by each approach. This highlights the speedup obtained by using wkk-unif and wkk-col over shi and kulis. For all the datasets, the



(d) Agg

(e) R15

(f) Fla

Figure 1: The first and second rows show the norm cuts ratio and the time taken for each dataset. It can be seen that wkk-unif and wkk-col, which are variants of the proposed framework yield results similar to kulis and shi, which use the entire affinity matrix. wkk-unif and wkk-col also outperform nys-col and nys-unif in terms of accuracy and computation time.

approaches based on combining sampling and the weighted kernel k-means algorithm outperform approaches based on the Nyström approximation both in terms of accuracy and computation time.

6.2 Image Segmentation

One of the most popular applications of spectral clustering is image segmentation. In this section, we describe results obtained on an image segmentation benchmark [Yu and Shi, 2003]. The affinity matrix and the final discretization were computed using the approach of [Yu and Shi, 2003].

Experimental results in Figure 2 show that wkk-unif and wkk-col obtain the same accuracy as the Nyström based approaches, while taking significantly less time. In cases where only 10% of the columns were sampled, wkk-unif and wkk-col achieved a lower norm cuts ratio compared to nys-unif and nys-col. The baseline methods shi and kulis use the entire affinity matrix. Thus they can achieve lower norm cuts ratio, but have a significantly higher computation time.

In order to study how the proposed approach scales to larger graphs, we vary the size of the input images. The results are shown in Figure 3. For shi, the time taken scales cubically with the size of the graph. Hence it is unsuitable for large graphs. In contrast, kulis, wkk-unif and wkk-col scale almost linearly with the size of the graph. Both wkk-unif and wkk-col outperform kulis in terms of execution time. They also outperform the Nyström-based approaches both in terms of accuracy and computation time. These results show that uniform sampling followed by the weighted kernel k-means algorithm scales well in practice.

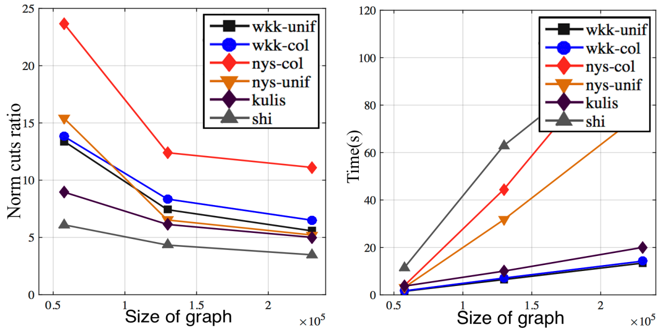


Figure 3: Behavior of the various approaches when the graph size increases. This shows that by using only 50% of the points, wkk-unif and wkk-col achieve accuracy comparable to kulis and shi while taking significantly less time. wkk-unif and wkk-col also outperform nys-unif and nys-col both in terms of accuracy and computation time.

7 Conclusion

This paper presented a framework for spectral clustering based on combining sampling and the weighted kernel k-means algorithm. If the points are sampled uniformly at random without replacement, we show that for large datasets, the number of samples required, is independent of the input size and depends only on the number of clusters and the diameter of the set of points in the kernel space. Experiments show that approaches based on the proposed framework outperform approaches based on the Nyström approximation.

References

[Bardenet *et al.*, 2015] Rémi Bardenet, Odalric-Ambrym Maillard, et al. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361–1385, 2015.

[Bouneffouf and Birol, 2015] Djallel Bouneffouf and Inanc Birol. Sampling with minimum sum of squared similarities for nystrom-based large scale spectral clustering. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, 2015.

[Butler, 2007] Steve Butler. Interlacing for weighted graphs using the normalized laplacian. *Electronic Journal of Linear Algebra*, 16(1):8, 2007.

[Czumaj and Sohler, 2010] Artur Czumaj and Christian Sohler. Sublinear-time algorithms. In *Property testing*, pages 41–64. Springer, 2010.

[Dhillon *et al.*, 2004] Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. *A unified view of kernel k-means, spectral clustering and graph cuts*. Citeseer, 2004.

[Gittens and Mahoney, 2013] Alex Gittens and Michael W Mahoney. Revisiting the nystrom method for improved large-scale machine learning. *arXiv preprint arXiv:1303.1849*, 2013.

[Kumar *et al.*, 2012] Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Sampling methods for the nyström method. *The Journal of Machine Learning Research*, 13(1):981–1006, 2012.

[Serfling, 1974] Robert J Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, pages 39–48, 1974.

[Talwalkar and Rostamizadeh, 2010] Ameet Talwalkar and Afshin Rostamizadeh. Matrix coherence and the nystrom method. *arXiv preprint arXiv:1004.2008*, 2010.

[Von Luxburg, 2007] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[Williams and Seeger, 2001] Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Proceedings of the 14th Annual Conference on Neural Information Processing Systems*, number EPFL-CONF-161322, pages 682–688, 2001.

[Yan *et al.*, 2009] Donghui Yan, Ling Huang, and Michael I Jordan. Fast approximate spectral clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 907–916. ACM, 2009.

[Yu and Shi, 2003] Stella X Yu and Jianbo Shi. Multiclass spectral clustering. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 313–319. IEEE, 2003.