# Making Large-Scale Nyström Approximation Possible

**Mu Li** LIMU.CN@GMAIL.COM
Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

**James T. Kwok** JAMESK@CSE.UST.HK
Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong

**Bao-Liang Lu** BLLU@SJTU.EDU.CN
Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
MOE-MS Key Lab. for Intel. Comp. and Intel. Sys., Shanghai Jiao Tong University, Shanghai 200240, China

## Abstract

The Nyström method is an efficient technique for the eigenvalue decomposition of large kernel matrices. However, in order to ensure an accurate approximation, a sufficiently large number of columns have to be sampled. On very large data sets, the SVD step on the resultant data submatrix will soon dominate the computations and become prohibitive. In this paper, we propose an accurate and scalable Nyström scheme that first samples a large column subset from the input matrix, but then only performs an approximate SVD on the inner submatrix by using the recent randomized low-rank matrix approximation algorithms. Theoretical analysis shows that the proposed algorithm is as accurate as the standard Nyström method that directly performs a large SVD on the inner submatrix. On the other hand, its time complexity is only as low as performing a small SVD. Experiments are performed on a number of large-scale data sets for low-rank approximation and spectral embedding. In particular, spectral embedding of a MNIST data set with 3.3 million examples takes less than an hour on a standard PC with 4G memory.

## 1. Introduction

Eigenvalue decomposition is of central importance in science and engineering, and has numerous applications in diverse areas such as physics, statistics, signal processing, machine learning and data mining. In machine learning for example, eigenvalue decomposition is used in kernel principal component analysis and kernel Fisher discriminant analysis for the extraction of nonlinear structures and decision boundaries from the kernel matrix. The eigenvectors of the kernel or affinity matrix are also used in many spectral clustering (von Luxburg, 2007) and manifold learning algorithms (Belkin & Niyogi, 2002; Tenenbaum et al., 2000) for the discovery of the intrinsic clustering structure or low-dimensional manifolds.

However, standard algorithms for computing the eigenvalue decomposition of a dense $n \times n$ matrix take $O(n^3)$ time, which can be prohibitive for large data sets. Alternatively, when only a few leading (or trailing) eigenvalues/eigenvectors are needed, one may perform a partial singular value decomposition (SVD) using the Arnoldi method (Lehoucq et al., 1998). However, empirically, the time reduction is significant only when the matrix is sparse or very few eigenvectors are extracted (Williams & Seeger, 2001).

A more general approach to alleviate this problem is by using low-rank matrix approximations, of which the Nyström method (Drineas & Mahoney, 2005; Fowlkes et al., 2004; Williams & Seeger, 2001) is the most popular. It selects a subset of $m \ll n$ columns from the kernel matrix, and then uses the correlations between the sampled columns and the remaining columns to form a low-rank approximation of the full matrix. Computationally, it only has to decompose the much smaller $m \times m$ matrix (denoted $W$). Obviously, the more columns are sampled, the more accurate is the resultant approximation. However, there is a tradeoff between accuracy and efficiency. In particular, on very large data sets, even decomposing the small $W$

Appearing in *Proceedings of the $27^{th}$ International Conference on Machine Learning*, Haifa, Israel, 2010. Copyright 2010 by the author(s)/owner(s).

matrix can be expensive. For example, when the data set has several millions examples, sampling only 1% of the columns will lead to a $W$ that is larger than $10,000 \times 10,000$.

To avoid this explosion of $m$, Kumar *et al.* (2010) recently proposed the use of an ensemble of $n_e$ Nyström approximators. Each approximator, or *expert*, performs a standard Nyström approximation with a manageable column subset. Since the sampling of columns is stochastic, a number of such experts can be run and the resultant approximations are then linearly combined together. Empirically, the resultant approximation is more accurate than that of a single expert as in standard Nyström. Moreover, its computational cost is (roughly) only $n_e$ times the cost of standard Nyström. However, as will be shown in Section 3, it is essentially using a block diagonal matrix to approximate the inverse of a very large $W$. Since the inverse of a block diagonal matrix is another block diagonal matrix, this approximation can be poor unless $W$ is close to block diagonal. However, this is highly unlikely in typical applications of the Nyström method.

Recently, a new class of randomized algorithms are proposed for constructing approximate, low-rank matrix decompositions (Halko et al., 2009). It also extends the Monte Carlo algorithms in (Drineas et al., 2006) on which the analysis of the Nyström method in (Drineas & Mahoney, 2005) is based. Unlike the standard Nyström which simply samples a column subset for approximation, it first constructs a low-dimensional subspace that captures the action of the input matrix. Then, a standard factorization is performed on the matrix which is restricted to that subspace. Though being a randomized algorithm, it is shown that this can yield an accurate approximation with very high probability. On the other hand, the algorithm needs to have at least one pass over the whole input matrix. This is thus more expensive than the Nyström method (and its ensemble variant) which only accesses a column subset. On very large data sets, this performance difference can be significant.

In this paper, we combine the merits of the standard Nyström method and the randomized SVD algorithm. The standard Nyström is highly efficient but requires a large enough number of columns to be sampled, while the randomized SVD algorithm is highly accurate but less efficient. Motivated by the observation that the ensemble Nyström algorithm is essentially using a block diagonal matrix approximation for $W^+$, we will adopt a large column subset and then speed up the inner SVD step by randomized SVD. Both theoretical analysis and experimental results confirm that the error in the randomized SVD step is more than compensated for by the ability to use a large column subset, leading to an efficient and accurate eigenvalue decomposition even for very large input matrices. Moreover, unlike the ensemble Nyström method which resorts to a learner and needs to attend to the consequent model selection issues, the proposed method is very easy to implement and can be used to obtain approximate eigenvectors.

The rest of this paper is organized as follows. Section 2 gives a short introduction on the standard/ensemble Nyström method and the randomized SVD algorithm. Section 3 then describes the proposed algorithm. Experimental results are presented in Section 4, and the last section gives some concluding remarks.

**Notations** The transpose of vector/matrix is denoted by the superscript $^T$. Moreover, $\mathrm{Tr}(A)$ denotes the trace of matrix $A = [A_{ij}]$, $A^+$ is its pseudo-inverse, $\mathrm{ran}(A)$ is the range of $A$, $\|A\|_2 = \max\{\sqrt{\lambda} : \lambda \text{ is eigenvalue of } A^TA\}$ is its spectral norm, $\|A\|_F = \sqrt{\mathrm{Tr}(A^TA)}$ is its Frobenius norm, and $\sigma_i(A)$ denotes the $i$th largest singular value of $A$.

## 2. Related Works

### 2.1. Nyström Method

The Nyström method approximates a symmetric positive semidefinite (psd) matrix $G \in \mathbb{R}^{n\times n}$ by a sample $C$ of $m \ll n$ columns from $G$. Typically, this subset of columns are randomly selected by uniform sampling without replacement (Williams & Seeger, 2001; Kumar et al., 2009). Recently, more sophisticated non-uniform sampling schemes have also been pursued (Drineas & Mahoney, 2005; Zhang et al., 2008).

After selecting $C$, the rows and columns of $G$ can be rearranged such that $C$ and $G$ are written as:

$$C = \begin{bmatrix} W \\ S \end{bmatrix} \quad \text{and} \quad G = \begin{bmatrix} W & S^T \\ S & B \end{bmatrix}, \tag{1}$$

where $W \in \mathbb{R}^{m\times m}, S \in \mathbb{R}^{(n-m)\times m}$ and $B \in \mathbb{R}^{(n-m)\times(n-m)}$. Assume that the SVD of $W$ is $U\Lambda U^T$, where $U$ is an orthonormal matrix and $\Lambda = \mathrm{diag}(\sigma_1, \ldots, \sigma_m)$ is the diagonal matrix containing the singular values of $W$ in non-increasing order. For $k \le m$, the rank-$k$ Nyström approximation is

$$\tilde{G}_k = CW_k^+C^T, \tag{2}$$

where $W_k^+ = \sum_{i=1}^k \sigma_i^{-1} U^{(i)} U^{(i)T}$, and $U^{(i)}$ is the $i$th column of $U$. The time complexity is $O(nmk + m^3)$. Since $m \ll n$, this is much lower than the $O(n^3)$ complexity required by a direct SVD on $G$.

### 2.2. Ensemble Nyström Algorithm

Since the Nyström method relies on random sampling of columns, it is stochastic in nature. The ensemble Nyström method (Kumar et al., 2010) employs an ensemble of $n_e \geq 1$ Nyström approximators for improved performance. It first samples $mn_e$ columns from $G$, which can be written as $C = [C_1, \ldots, C_{n_e}] \in \mathbb{R}^{n \times mn_e}$ with each $C_i \in \mathbb{R}^{n \times m}$. The standard Nyström method is then performed on $C_i$, obtaining a rank-$k$ approximation $\tilde{G}_{i,k}$ $(i = 1, \ldots, n_e)$. Finally, these are weighted to form the ensemble approximation

$$\tilde{G}^{ens} = \sum_{i=1}^{n_e} \mu_i \tilde{G}_{i,k}, \tag{3}$$

where $\mu_i$'s are the mixture weights. A number of choices have been used in setting these weights, including uniform weights, exponential weights and by ridge regression. Empirically, the best method is ridge regression. This, however, needs to sample an additional $s$ columns from $G$ as the training set, and another $s'$ columns as the hold-out set for model selection. The total time complexity is $O(n_e nmk + n_e m^3 + C_{\boldsymbol{\mu}})$, where $C_{\boldsymbol{\mu}}$ is the cost of computing the mixture weights.

Another disadvantage of the ensemble Nyström method is that, unlike the standard Nyström method, approximate eigenvectors of $G$ cannot be easily obtained. As can be seen from (3), the eigenvectors of each of the $\tilde{G}_{i,k}$'s in (3) are in general different and so cannot be easily combined together. Hence, the ensemble Nyström method cannot be used with spectral clustering and manifold learning algorithms.

### 2.3. Randomized Low-Rank Approximation

Recently, a class of simple but highly efficient randomized algorithms are proposed for constructing approximate, low-rank matrix decompositions (Halko et al., 2009). In general, they can be used on complex-valued rectangular matrices. In the following, we focus on obtaining a rank-$k$ SVD from a symmetric matrix $W \in \mathbb{R}^{m \times m}$ (Algorithm 1).

**Algorithm 1** Randomized SVD (Halko et al., 2009).

```
Input: $m \times m$ symmetric matrix $W$, scalars $k$, $p$, $q$.
Output: $U$, $\Lambda$.
1: $\Omega \leftarrow$ a $m \times (k+p)$ standard Gaussian random matrix.
2: $Z \leftarrow W\Omega$, $Y \leftarrow W^{q-1}Z$.
3: Find an orthonormal matrix $Q$ (e.g., by QR decomposition) such that $Y = QQ^TY$.
4: Solve $B(Q^T\Omega) = Q^TZ$.
5: Perform SVD on $B$ to obtain $V\Lambda V^T = B$.
6: $U \leftarrow QV$.
```

In general, there are two computational stages in this class of algorithms. In the first stage (steps 1 to 3), an orthonormal matrix $Q \in \mathbb{R}^{m \times (k+p)}$ is constructed which serves as an approximate, low-dimensional basis for the range of $W$ (i.e., $W \simeq QQ^TW$). Here, $p$ is an over-sampling parameter (typically set to 5 or 10) such that the rank of $Q$ is slightly larger than the desired rank ($k$), and $q$ is the number of steps of a power iteration (typically set to 1 or 2) which is used to speed up the decay of the singular values of $W$. In the second stage (steps 4 to 6), the input matrix matrix is restricted to the above subspace and a standard SVD is then computed on the reduced matrix

$$B = Q^TWQ \tag{4}$$

to obtain $B = V\Lambda V^T$. Finally, the SVD of $W$ can be approximated as $W \simeq U\Lambda U^T$, where $U = QV$.

Computationally, it takes $O(m^2k)$ time[1] to compute $Z$ and $Y$, $O(mk)$ time for the QR decomposition, $O(mk^2)$ time to obtain $B$, and $O(k^3)$ time for the SVD. Hence, the total time complexity is $O(m^2k + k^3)$, which is quadratic in $m$. Moreover, it needs to have at least one pass over the whole input matrix.

## 3. Algorithm

### 3.1. Combining Nyström and Randomized SVD

Obviously, the more columns are sampled, the more accurate is the Nyström approximation. Hence, the ensemble Nyström method samples $mn_e$ columns instead of $m$ columns. In the following, we abuse notations and denote the corresponding $W$ matrix by $W_{(n_em)} \in \mathbb{R}^{mn_e \times mn_e}$. However, there is a tradeoff between accuracy and efficiency. If the standard Nyström method were used, this would have taken $O(n_e^3m^3)$ time for the SVD of $W_{(n_em)}$. The ensemble Nyström method alleviates this problem by replacing this expensive SVD by $n_e$ SVDs on $n_e$ smaller $m \times m$ matrices. Our key observation is that, by using (2), the ensemble Nyström approximation in (3) can be rewritten as

$$\tilde{G}^{ens} = C \operatorname{diag}(\mu_1 W_{1,k}^+, \ldots, \mu_{n_e} W_{n_e,k}^+) C^T, \tag{5}$$

where $W_{i,k} \in \mathbb{R}^{m \times m}$ is the $W$ matrix in (1) corresponding to $\tilde{G}_{i,k}$, and $\operatorname{diag}(\mu_1 W_{1,k}^+, \ldots, \mu_{n_e} W_{n_e,k}^+)$ is

[1] Here, we compute $Y$ by multiplying $W$ to a sequence of $m \times (k+p)$ matrices, as $WZ, W(WZ), \ldots, W(W^{q-2}Z)$.

the block diagonal matrix $\begin{bmatrix} \mu_1 W^+_{1,k} & & \\ & \ddots & \\ & & \mu_{n_e} W^+_{n_e,k} \end{bmatrix}$. In other words, the ensemble Nyström algorithm can be equivalently viewed as approximating $W^+_{(n_e m)}$ by the block diagonal $\text{diag}(\mu_1 W^+_{1,k}, \ldots, \mu_{n_e} W^+_{n_e,k})$. Despite the resultant computational simplicity, the inverse of a block diagonal matrix is another block diagonal matrix. Hence, no matter how sophisticated the mixture weights $\mu_i$'s are estimated, this block diagonal approximation is rarely valid unless $W_{(n_e m)}$ is block diagonal. This, however, is highly unlikely in typical applications of the Nyström method.

Since the ensemble Nyström method attains better performance by sampling more columns, our method will also sample more columns, or, equivalently, use a $m$ larger than is typically used in the standard Nyström method. However, instead of using a block diagonal matrix approximation for solving the subsequent large-SVD problem, we will use a more accurate procedure. In particular, we will adopt the randomized low-rank matrix approximation technique introduced in Section 2.3.

**Algorithm 2** The proposed algorithm.

```
Input: Psd matrix $G \in \mathbb{R}^{n \times n}$, number of columns $m$,
    rank $k$, over-sampling parameter $p$, power parameter $q$.
Output: $\hat{G}$, an approximation of $G$.
1: $C \leftarrow m$ columns of $G$ sampled uniformly at random without replacement.
2: $W \leftarrow m \times m$ matrix defined in (1).
3: $[\tilde{U}, \Lambda] \leftarrow \text{randsvd}(W, k, p, q)$ using Algorithm 1.
4: $U \leftarrow C\tilde{U}\Lambda^+$.
5: $\hat{G} \leftarrow \left(\sqrt{\frac{m}{n}} U\right) \left(\frac{n}{m}\Lambda\right) \left(\sqrt{\frac{m}{n}} U^T\right)$.
```

The proposed algorithm is shown in Algorithm 2. Essentially, it combines the high efficiency of the Nyström method, which however requires a large enough column subset for accurate approximation, with the ability of the randomized algorithm to produce a very accurate SVD but still relatively efficient approximation. Note from step 5 that $\hat{G} = C\tilde{U}\Lambda^+\tilde{U}^T C^T$. In turn, from Algorithm 1 and (4), $\tilde{U}\Lambda\tilde{U}^T = QBQ^T = Q(Q^T WQ)Q^T$. Hence, instead of relying on the block diagonal matrix approximation in (5), $\hat{G}$ is now more accurately approximated as

$$\hat{G} = CQ(Q^T WQ)^+ Q^T C. \tag{6}$$

Besides, instead of using the randomized SVD algorithm for the inner SVD, one might want to apply other approximations, such as using the standard Nyström method again. However, the Nyström method is not good at approximating the trailing eigenvalues, which are important in computing the inverse of $W$. Preliminary experiments show that in order for the resultant approximation on $G$ to be accurate, the inner Nyström needs to sample close to $m$ columns, which, however, will lead to little speedup over a standard SVD. Moreover, recall from Section 2.1 that there are different column sampling strategies. Here, we will focus on uniform sampling without replacement (Kumar et al., 2009). Extension to other sampling schemes will be studied in the future.

*Table 1.* Time complexities for the various methods to obtain a rank-$k$ Nyström approximation of an $n \times n$ matrix. Here, $m$ is the number of columns sampled.

| METHOD | TIME COMPLEXITY |
|---|---|
| NYSTRÖM | $O(nmk + m^3)$ |
| ENSEMBLE NYSTRÖM | $O(nmk + n_e k^3 + C_{\mu})$ |
| RANDOMIZED SVD | $O(n^2 k + k^3)$ |
| PROPOSED METHOD | $O(nmk + k^3)$ |

The time complexity required is $O(nmk + k^3)$. A summary of the time complexities[2] of the various methods is shown in Table 1. Recall that typically $n \gg m \geq k$. As can be seen, all the methods except randomized SVD scale linearly with $n$. Moreover, the proposed method has a comparable complexity as the ensemble Nyström method as both only scale cubically with $k$, but not with $m$.

### 3.2. Error Analysis

Let the column sampling matrix be $S \in \{0,1\}^{n \times k}$, where $S_{ij} = 1$ if the $i$th column of $G$ is chosen in the $j$ random trial, and $S_{ij} = 0$ otherwise. Then, $C = GS$ and $W = S^T GS$. Moreover, since $G$ is psd, we can write it as

$$G = X^T X, \tag{7}$$

for some $X \in \mathbb{R}^{d \times n}$. In the proof, we will also need the column-sampled and rescaled version of $X$:

$$H = \kappa XS, \tag{8}$$

where $\kappa = \sqrt{n/m}$ is the scaling factor. Then,

$$C = \kappa^{-1} X^T H, \quad W = \kappa^{-2} H^T H. \tag{9}$$

The error analysis will depend on a number of results in (Halko et al., 2009; Kumar et al., 2009; Stewart,

[2] In order to be consistent with the other methods, the *total* number of columns sampled in the ensemble Nyström method is now $m$ (not $n_e m$ in Section 2.2).

1990). For the readers' convenience, these are listed in the appendix.

### 3.2.1. Spectral Norm

For the matrix $W$ in step 3, we will first compute its (expected) approximation error $\mathbb{E}\,\|W - QQ^TW\|_2$. Since the input matrix $G$ is psd, $W$ is also psd. The following proposition can then be obtained by using a more general result in (Halko et al., 2009).

**Proposition 1.** *Given a psd matrix $W$, the $Q$ obtained in Algorithm 1 satisfies*

$$\mathbb{E}\,\|W - QQ^TW\|_2 \le \zeta^{1/q}\sigma_{k+1}(W), \tag{10}$$

*where $\zeta = 1 + \sqrt{\frac{k}{p-1}} + \frac{e\sqrt{k+p}}{p}\sqrt{m-k}$.*

The main theorem is stated below.

**Theorem 1.** *For the $\hat{G}$ obtained in Algorithm 2,*

$$\mathbb{E}\,\|G-\hat{G}\|_2 \le \zeta^{1/q}\|G-G_k\|_2+(1+\zeta^{1/q})\frac{n}{\sqrt{m}}G^*_{ii}, \tag{11}$$

*where $G_k$ is the best rank-$k$ approximation of $G$, $G^*_{ii} = \max_i G_{ii}$, and $\zeta = 1 + \sqrt{\frac{k}{p-1}} + \frac{e\sqrt{k+p}}{p}\sqrt{m-k}$.*

As in (Halko et al., 2009), the power iteration drives $\zeta^{1/q}$ towards 1 exponentially fast as $q$ increases, and so the error in (11) decreases with the number of sampled columns $m$. In particular, if we replace $\zeta^{1/q}$ by 1, then (11) becomes $\|G - G_k\|_2 + \frac{2n}{\sqrt{m}}G^*_{ii}$, which is the same[3] as that for the standard Nyström method using $m$ columns. In other words, Algorithm 2 is as accurate as performing a large SVD in standard Nyström.

### 3.2.2. Frobenius Norm

A similar bound can be obtained for the approximation error in terms of the Frobenius norm. Since there is no analogous theory for power iteration w.r.t. the Frobenius norm (cf. remark 10.1 of (Halko et al., 2009)), the analysis here is restricted to $q = 1$ and thus the resultant bound is quite loose. However, as will be seen in Section 4, empirically the approximation with just $q = 2$ is already very good.

**Theorem 2.** *For the $\hat{G}$ obtained in Algorithm 2,*

$$\begin{aligned}&\mathbb{E}\,\|G - \hat{G}\|_F \\ &\quad\le \frac{2(k+p)}{\sqrt{p-1}}\|G - G_k\|_F + \left(1 + \frac{4(k+p)}{\sqrt{m(p-1)}}\right)nG^*_{ii}.\end{aligned}$$

*Table 2.* Data sets used.

| | DATA | #SAMPLES | DIM |
|---|---|---|---|
| LOW-RANK APPROX | SATIMAGE | 4,435 | 36 |
| | RCV1 | 23,149 | 47,236 |
| | MNIST | 60,000 | 784 |
| | COVTYPE | 581,012 | 54 |
| EMBEDDING | MNIST-8M | 3,276,294 | 784 |

## 4. Experiments

In this section, we study the efficiency of the proposed method in solving large dense eigen-systems. Experiments are performed on low-rank approximation (Section 4.1) and spectral embedding (Section 4.2). All the implementations are in Matlab. Experiments are run on a PC with 2.4GHz Core2 Duo CPU and 4G memory.

### 4.1. Low-rank Approximation

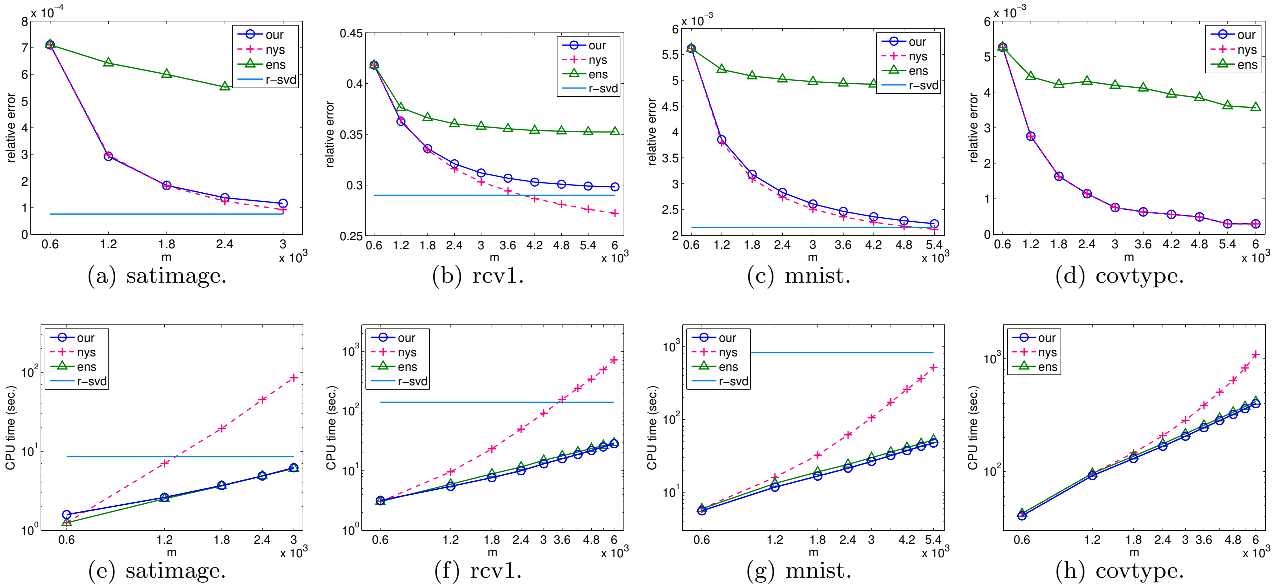
We use a number of data sets from the LIBSVM archive[4] (Table 2). The linear kernel is used on the RCV1 text data set, while the Gaussian kernel is used on all the others. The following methods are compared in the experiments:

1. Standard Nyström method (denoted nys);
2. Ensemble Nyström method (denoted ens): As in (Kumar et al., 2010), an additional $s = 20$ columns are used for training the mixture weights by ridge regression, and another $s' = 20$ columns are used for choosing the regularization parameters. For the covtype data set, $s$ and $s'$ are reduced to 2 so as to speed up computation. Moreover, we set $n_e = m/k$.
3. The proposed method (denoted our): We fix the over-sampling $p$ to 5, and the power parameter $q$ to 2.
4. Randomized SVD (denoted r-svd): Similar to the proposed method, we also use $p = 5$ and $q = 2$.

The first three methods are based on the Nyström method, and $m$ columns are uniformly sampled without replacement. Due to randomness in the sampling process, we perform 10 repetitions and report the averaged result. On the other hand, the randomized SVD algorithm does not perform sampling and the whole

[3] This bound can be obtained in (Kumar et al., 2010) by combining their (6) and (10).

[4] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/



(a) satimage. (b) rcv1. (c) mnist. (d) covtype.

(e) satimage. (f) rcv1. (g) mnist. (h) covtype.

*Figure 1.* Performance of the various methods. Top: Low-rank approximation error; Bottom: CPU time. (The randomized SVD algorithm cannot be run on the covtype data set because it is too large).

input matrix is always used. Besides, the best rank-$k$ approximation could have been obtained by a direct SVD on the whole input matrix. However, this is computationally expensive even on medium-sized data sets and so is not compared here.

#### 4.1.1. Different Numbers of Columns

In the first experiment, we fix $k = 600$ and gradually increase the number of sampled columns ($m$). Figure 1 shows the the relative approximation error[5] $\|G - \hat{G}\|_F / \|G\|_F$ and the CPU time. As can be seen, the randomized SVD algorithm is often the most accurate, albeit also the most expensive. Standard Nyström can be as accurate as randomized SVD when $m$ is large enough. However, since Nyström takes $O(m^3)$ time for the SVD step, it also quickly becomes computationally infeasible. As for the ensemble Nyström method, it degenerates to the standard Nyström when $n_e = 1$ (the left endpoint of the curve). Its approximation error decreases when the ensemble has more experts[6], which is consistent with the results in (Kumar et al., 2010). However, as discussed in Section 3.1, the ensemble Nyström method approximates the large SVD problem with a crude block diagonal matrix approximation. Hence, its accuracy is much inferior to that of the standard Nyström (which performs the large SVD directly). On the other hand, the proposed method is almost as accurate as standard Nyström, while its CPU time is comparable or even smaller than that of the ensemble Nyström method.

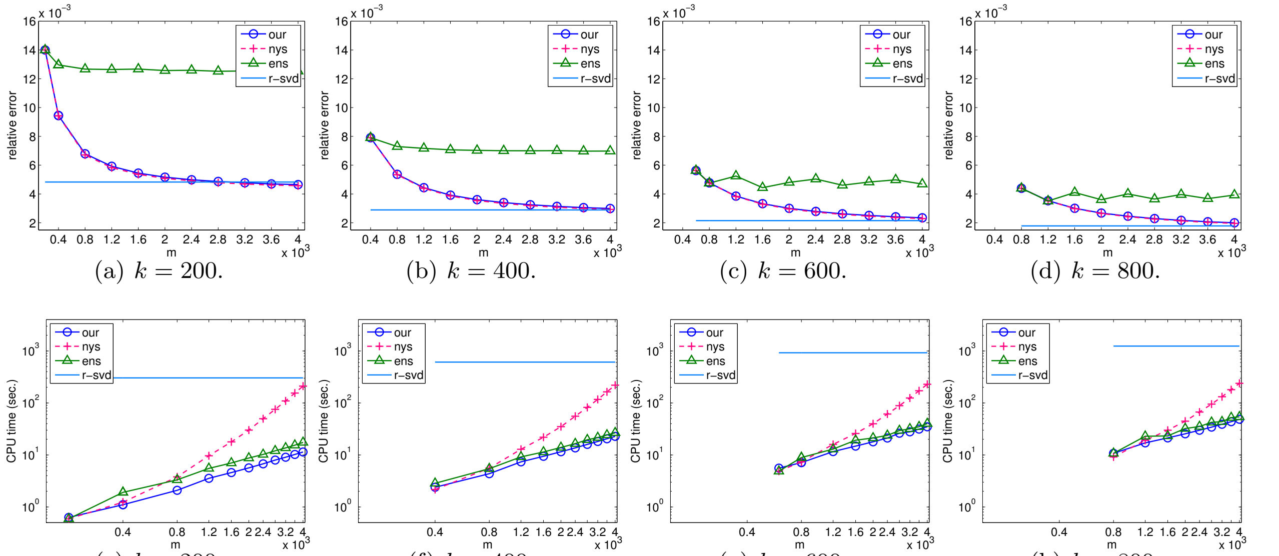
The accuracy of the ensemble Nyström method can be improved, at the expense of more computations. Recall that in our setup, all Nyström-based methods have access to $m$ columns. For the ensemble Nyström, these columns are divided by the $m/k$ experts, each receiving a size-$k$ subset. In general, let $r$ be the number of columns used by each expert. Obviously, the larger the $r$, the better the ensemble approximation. Indeed, in the extreme case where $r = m$, the ensemble Nyström method degenerates to the standard Nyström. Hence, accuracy of the ensemble Nyström method can be improved by using fewer experts, with each expert using a larger column subset. However, the time for performing $\frac{m}{r}$ size-$r$ SVD's is $O(\frac{m}{r} r^3) = O(mr^2)$. Figure 2 shows the resultant tradeoff between approximation error and CPU time on the satimage data set. As can be seen, in order for the ensemble Nyström method to have comparable speed with the proposed algorithm, this justifies our choice of $r = k$.

#### 4.1.2. Different Ranks

In the second experiment, we study the approximation performance when the rank $k$ varies. Because of the

[5] Results are only reported for the Frobenius norm because the approximation error w.r.t. the spectral norm is computationally difficult to compute, especially on large data sets. Nevertheless, this is still a good indication of the approximation performance as the spectral norm is upper-bounded by the Frobenius norm (Lütkepohl, 1996).

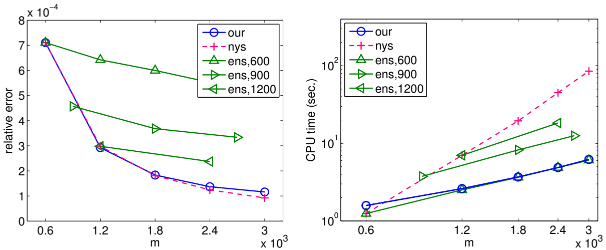
[6] Recall that we use $n_e = m/k$, and so the number of experts increases with $m$.



(a) $k = 200$. (b) $k = 400$. (c) $k = 600$. (d) $k = 800$.

(e) $k = 200$. (f) $k = 400$. (g) $k = 600$. (h) $k = 800$.

*Figure 3.* Performance at different $k$'s on the MNIST data set. Top: Low-rank approximation error; Bottom: CPU time.



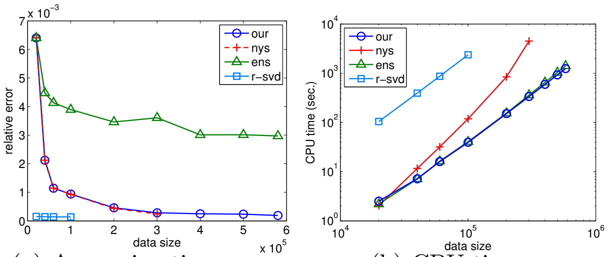
(a) Approximation error. (b) CPU time.

*Figure 2.* Low-rank approximation performance for the ensemble Nyström method, with varying number of columns used by each expert.

lack of space, results are only reported on the MNIST data set. As can be seen from Figure 3, when $k$ increases, the approximation error decreases while the CPU time increases across all methods. Hence, there is a tradeoff between accuracy and efficiency. Nevertheless, the relative performance comparison among the various methods is still the same as in Section 4.1.1.

### 4.1.3. Input Matrices of Different Sizes

In this experiment, we examine how the performance scales with the size of the input matrix. The various methods are run on subsets of the covtype data set. Here, $k$ is fixed to 600 and $m$ to $0.03n$. Results are shown in Figure 4. Note that the slopes of the curves in Figure 4(b) determines their scalings with $n$. As can be seen, the standard Nyström method scales cubically with $n$, while all the other methods scale quadratically (because $m$ also scales linearly with $n$ here). Moreover, similar to the results in the previous sections, the ensemble Nyström and the proposed method are most scalable, while the proposed method is as accurate as the standard Nyström that performs a large SVD.



(a) Approximation error. (b) CPU time.

*Figure 4.* Low-rank approximation performance at different sizes of the input matrix on the covtype data set.

## 4.2. Spectral Embedding

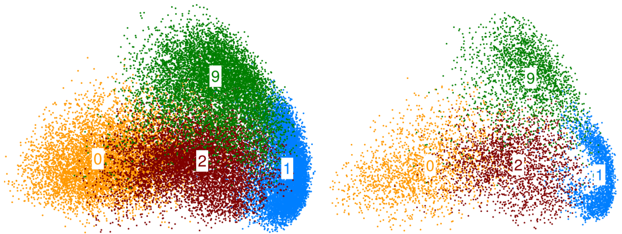
In this section, we perform spectral embedding using the Laplacian eigenmap (Belkin & Niyogi, 2002). The Gaussian kernel is used to construct the affinity matrix. For easy visualization, the data are projected onto the two singular vectors of the normalized Laplacian with the second and third smallest singular values.

Experiments are performed on the MNIST-8M data set[7], which contains 8.1M samples constructed by elastic deformation of the original MNIST training set. To avoid clutter of the embedding results, we only use digits 0, 1, 2 and 9, which result in a data set with about

[7] http://leon.bottou.org/papers/loosli-canu-bottou-2006

3.3M samples. Because of this sheer size, neither standard SVD nor Nyström can be run on the whole set. Moreover, neither can the ensemble Nyström method be used as it cannot produce approximate eigenvectors (Section 2.2). Hence, the full set can only be run with the proposed method, with $m = 4000$, $k = 400$ and the Gaussian kernel. For comparison, we also run standard SVD on a random subset of 8,000 samples.

Results are shown in Figure 5. As can be seen, the two embedding results are very similar. Besides, for the proposed method, this embedding of 3.3M samples is obtained within an hour on our PC.



(a) Proposed method. (b) SVD on a data subset.

*Figure 5.* Embedding results for the digits 0,1,2,9 in the MNIST-8M data set.

## 5. Conclusion

In this paper, we proposed an accurate and scalable Nyström approximation scheme for very large data sets. It first samples a large column subset from the input matrix, and then performs an approximate SVD on the inner submatrix by using the recent randomized low-rank matrix approximation algorithms. Both theory and experiments demonstrate that the proposed algorithm is as accurate as the standard Nyström method that directly performs a large SVD on the inner submatrix. On the other hand, its time complexity is only as low as the ensemble Nyström method. In particular, spectral embedding of a MNIST data set with 3.3 million examples takes less than an hour on a standard PC with 4G memory.

## Acknowledgments

This research was supported in part by the Research Grants Council of the Hong Kong Special Administrative Region (Grant 614508), the National Natural Science Foundation of China (Grant No. 60773090 and Grant No. 90820018), the National Basic Research Program of China (Grant No. 2009CB320901), and the National High-Tech Research Program of China (Grant No. 2008AA02Z315).

## References

Belkin, M. and Niyogi, P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS 14*, 2002.

Drineas, P. and Mahoney, M.W. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6:2175, 2005.

Drineas, P., Kannan, R., and Mahoney, M.W. Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix. *SIAM Journal on Computing*, 36(1):158–183, 2006.

Fowlkes, C., Belongie, S., Chung, F., and Malik, J. Spectral grouping using the Nyström method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):214–225, February 2004.

Halko, N., Martinsson, P.-G., and Tropp, J.A. Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions. Technical report, 2009.

Kumar, S., Mohri, M., and Talwalkar, A. Sampling techniques for the Nyström method. In *AISTATS*, 2009.

Kumar, S., Mohri, M., and Talwalkar, A. Ensemble Nyström method. In *NIPS 22*, 2010.

Lehoucq, R.B., Sorensen, D.C., and Yang, C. *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*. SIAM, 1998.

Lütkepohl, H. *Handbook of Matrices*. John Wiley and Sons, 1996.

Stewart, G.W. Matrix perturbation theory. *SIAM Review*, 1990.

Tenenbaum, J.B., de Silva, V., and Langford, J.C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

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

Williams, C.K.I. and Seeger, M. Using the Nyström method to speed up kernel machines. In *NIPS 13*, 2001.

Zhang, K., Tsang, I.W., and Kwok, J.T. Improved Nyström low rank approximation and error analysis. In *ICML*, 2008.

## A. Existing Results

The following proposition is used in the proof of Proposition 8.6 in (Halko et al., 2009).

**Proposition 2.** *Suppose that $R$ is an orthogonal projector, $D$ is a nonnegative diagonal matrix, and integer $q \geq 1$. Then, $\|RDR\|_2^q \leq \|RD^qR\|_2$.*

**Theorem 3.** *[Theorem 10.6, (Halko et al., 2009)] Suppose that $A \in \mathbb{R}^{m\times n}$ with singular values $\sigma_1 \geq \sigma_2 \geq \ldots$. Choose a target rank $k$ and an oversampling parameter $p \geq 2$, where $k+p \leq \min\{m,n\}$. Draw an $n \times (k+p)$ standard Gaussian matrix $\Omega$, and construct the sample matrix $Y = A\Omega$. Then,*

$$\begin{aligned}
&\mathbb{E}\,\|(I-P_Y)A\|_2 \\
&\quad\leq \left(1+\sqrt{\frac{k}{p-1}}\right)\sigma_{k+1} + \frac{e\sqrt{k+p}}{p}\left(\sum_{j>k}\sigma_j^2\right)^{1/2},
\end{aligned}$$

*and*

$$\mathbb{E}\,\|(I-P_Y)A\|_F \leq \left(1+\frac{k}{p-1}\right)^{1/2}\left(\sum_{i>k}\sigma_i^2\right)^{1/2}.$$

**Proposition 3.** *[Corollary 2, (Kumar et al., 2009)] Suppose $A \in \mathbb{R}^{m\times n}$. Choose a set $S$ of size $m$ uniformly at random without replacement from $\{1,\ldots,n\}$, and let $C$ equal the columns of $A$ corresponding to indices in $S$ scaled by $\sqrt{n/m}$. Then,*

$$\mathbb{E}\,\|AA^T - CC^T\|_F \leq \frac{n}{\sqrt{m}}\left(\max_i \|A^i\|\right)^2,$$

*where $A^i$ is the $i$th column of $A$.*

**Proposition 4.** *(Stewart, 1990) Given matrices $A \in \mathbb{R}^{n\times n}$ and $E \in \mathbb{R}^{n\times n}$,*

$$\begin{aligned}
\max_i |\sigma_i(A+E) - \sigma_i(A)| &\leq \|E\|_2, \\
\sum_i \Big(\sigma_i(A+E) - \sigma_i(A)\Big)^2 &\leq \|E\|_F^2.
\end{aligned}$$

## B. Preliminaries

In this section, we introduce some useful properties of the spectral and Frobenius norms that will be heavily used in the analysis.

**Lemma 1.** *(Lütkepohl, 1996)*

1. *For any square $A$, $\|AA^T\|_2 = \|A^TA\|_2 = \|A\|_2^2$.*
2. *For any orthogonal $U \in \mathbb{R}^{m\times m}$, orthogonal $V \in \mathbb{R}^{n\times n}$, and matrix $A \in \mathbb{R}^{m\times n}$, $\|UAV\|_2 = \|A\|_2$.*
3. *For any $A$, $\|A\|_2 = \|A^T\|_2$ and $\|A\|_2 \leq \|A\|_F$.*
4. *For any $A$ and $B$, $\|AB\|_F \leq \|A\|_F\|B\|_2$.*

**Definition** A matrix $P$ is an *orthogonal projector* if $P = P^T = P^2$.

Given a matrix $A$,

$$P_A = A(A^TA)^+A^T = U_AU_A^T, \tag{12}$$

where $U_A$ is an orthonormal basis of ran$(A)$, is an orthogonal projector. For orthogonal projector $P$, $I-P$ is also an orthogonal projector. Moreover, since $\|P\|_2^2 = \|P^TP\|_2 = \|P^2\|_2 = \|P\|_2$, so $\|P\|_2 = 0$ or 1.

**Lemma 2.** *For $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times m}$, $\|AB\|_2 = \|BA\|_2$.*

*Proof.* Assume, without loss of generality, that $n \geq m$. Recall that if $\lambda_1,\ldots,\lambda_m$ are the eigenvalues of $AB$, then $\lambda_1,\ldots,\lambda_m,0,\ldots,0$ are the eigenvalues of $BA$ (Lütkepohl, 1996). Hence,

$$\begin{aligned}
\|AB\|_2 &= \max\{\sqrt{\lambda} \,:\, \lambda \text{ is eigenvalue of } B^TA^TAB\} \\
&= \max\{\sqrt{\lambda} \,:\, \lambda \text{ is eigenvalue of } ABB^TA^T\} \\
&= \|BA\|_2.
\end{aligned}$$

□

**Lemma 3.** *For any $A \in \mathbb{R}^{n\times n}$, and any $v \in \mathbb{R}^n$ with $\|v\| = 1$, then $v'Av \leq \|A\|_2$.*

*Proof.* $\|A\|_2 = \|A^{\frac{1}{2}}A^{\frac{1}{2}}\|_2 = \|A^{\frac{1}{2}}\|_2^2 = \max_{\|v\|=1}\|A^{\frac{1}{2}}v\|^2 = \max_{\|v\|=1} v'Av$. □

**Lemma 4.** *For any positive semidefinite (psd) $A \in \mathbb{R}^{n\times n}$, $\|A\|_F \leq \mathrm{Tr}(A)$.*

*Proof.* Since $A$ is psd, $\sigma_i(A) \geq 0$. Thus,

$$\begin{aligned}
\|A\|_F^2 &= \mathrm{Tr}(A^TA) = \sum_{i=1}^n \sigma_i(A^TA) = \sum_{i=1}^n \sigma_i^2(A) \\
&\leq \left(\sum_{i=1}^n \sigma_i(A)\right)^2 = (\mathrm{Tr}(A))^2.
\end{aligned}$$

□

**Lemma 5.** *For matrix $A$ and orthogonal projector $P$, $\|AP\|_F \leq \|A\|_F$.*

*Proof.* Since $\|P\|_2 \leq 1$, then, on using property 4 of Lemma 1, $\|AP\|_F \leq \|A\|_F\|P\|_2 \leq \|A\|_F$. □

## C. Proofs of the Results in Section 3.2.1

**Lemma 6.** *Let $P$ be an orthogonal projector, and $S$ be a psd matrix. For integer $q \geq 1$, $\|PS\|_2 \leq \|PS^q\|_2^{1/q}$.*

*Proof.* Let the SVD of $S$ be $S = U\Sigma U^T$. On using properties 1 and 2 of Lemma 1, we have

$$\|PS\|_2^{2q} = \|PSSP\|_2^q = \|(U^TPU)\Sigma^2(U^TPU)\|_2^q. \tag{13}$$

Note that $U^TPU = (U^TPU)^T = (U^TPU)(U^TPU)$. Hence, $U^TPU$ is also an orthogonal projector. Using Proposition 2 and that $\Sigma^2$ is diagonal, (13) becomes

$$\begin{aligned} \|PS\|_2^{2q} &\leq \|(U^TPU)\Sigma^{2q}(U^TPU)\|_2 = \|PS^{2q}P\|_2 \\ &= \|PS^q\|_2^2. \end{aligned}$$

□

**Proof of Proposition 1:**

*Proof.* Recall that $Q$ is an orthonormal basis of $Y$, thus, $P_Y = QQ^T$. First, consider $q = 1$. Using Theorem 3 and that $\sum_{i=k+1}^m \sigma_i^2(W) \leq (m-k)\sigma_{k+1}^2(W)$, we obtain (10). Now, for integer $q > 1$,

$$\mathbb{E}\,\|(I - P_Y)W\|_2 \leq \left(\mathbb{E}\,\|(I-P_Y)W\|_2^q\right)^{1/q},$$

on using the Hölder's inequality. Note that $I - P_Y$ is also an orthogonal projector. Hence, on using Lemma 6, we have

$$\mathbb{E}\,\|(I - P_Y)W\|_2 \leq \left(\mathbb{E}\,\|(I-P_Y)B\|_2\right)^{1/q},$$

where $B = W^q$. Using (10) with $q = 1$ (which has just been proved), we obtain

$$\mathbb{E}\,\|(I - P_Y)W\|_2 \leq \left(\zeta\sigma_{k+1}(B)\right)^{1/q} = \zeta^{1/q}\sigma_{k+1}(W).$$

□

**Lemma 7.** *For $G$ in (7) and $H$ in (8), $\mathbb{E}\,\|XX^T - HH^T\|_2 \leq \frac{n}{\sqrt{m}}G_{ii}^*$, where $G_{ii}^* = \max_i G_{ii}$.*

*Proof.* Using property 3 of Lemma 1, $\|XX^T - HH^T\|_2 \leq \|XX^T - HH^T\|_F$. Since $G = X^TX$, $G_{ii} = \|X^{(i)}\|^2$, where $X^{(i)}$ is the $i$th column of $X$. Thus, $\left(\max_i \|X^{(i)}\|\right)^2 = \max_i \|X^{(i)}\|^2 = \max_i G_{ii} = G_{ii}^*$. Result follows on applying Proposition 3. □

**Lemma 8.** *Let $U$ be an orthonormal basis of the range of matrix $R \in \mathbb{R}^{n\times k}$. Then for any $X \in \mathbb{R}^{n\times n}$,*

$$\|XX^T - XUU^TX^T\|_2 \leq \|XX^T - RR^T\|_2.$$

*Proof.* Let $P_R = UU^T$. On using property 1 of Lemma 1,

$$\begin{aligned} \|X^TX - X^TUU^TX\|_2 &= \|X - P_RX\|_2^2 \\ &= \max_{\|v\|=1} \|v^T(X - P_RX)\|^2. \end{aligned} \tag{14}$$

Decompose $v$ as $v = \alpha y^T + \beta z^T$, where $y \in \mathrm{ran}(R)$, $z \in \mathrm{ran}(R)^\perp$, $\|y\| = \|z\| = 1$ and $\alpha^2 + \beta^2 = 1$. Obviously, $y^TP_R = y^T$, and $z^TP_R = 0$. Thus,

$$\begin{aligned} &\|X - P_RX\|_2 \\ &\leq \max_{y\in\mathrm{ran}(R),\|y\|=1} \|y^T(X - P_RX)\| \\ &\quad + \max_{z\in\mathrm{ran}(R)^\perp,\|z\|=1} \|z^T(X-P_RX)\| \\ &\leq \max_{z\in\mathrm{ran}(R)^\perp,\|z\|=1} \|z^TX\|. \end{aligned} \tag{15}$$

Since $z \in \mathrm{ran}(R)^\perp$,

$$\begin{aligned} \|z^TX\|^2 &= z^TXX^Tz \\ &= z^TRR^Tz + z^T(XX^T - RR^T)z \\ &= z^T(XX^T - RR^T)z \\ &\leq \|XX^T - RR^T\|_2, \end{aligned}$$

on using Lemma 3. Result follows on combining this with (14) and (15). □

**Proof of Theorem 1:**

*Proof.* Let $R = HQ$ and $U_R$ an orthonormal basis of $\mathrm{ran}(R)$. From (6) and (12),

$$\begin{aligned} \hat{G} &= CQ(Q^TWQ)^+Q^TC^T \\ &= X^THQ(Q^TH^THQ)^+Q^TH^TX \\ &= X^TP_{HQ}X = X^TU_RU_R^TX. \end{aligned}$$

Using Lemmas 2 and 8, we have

$$\begin{aligned} \|G - \hat{G}\|_2 &= \|X^TX - X^TU_RU_R^TX\|_2 \\ &= \|X - U_RU_R^TX\|_2^2 = \|X - XU_RU_R^T\|_2^2 \\ &= \|XX^T - XU_RU_R^TX^T\|_2 \\ &\leq \|XX^T - RR^T\|_2 \\ &\leq \|XX^T - HH^T\|_2 + \|HH^T - RR^T\|_2. \end{aligned}$$

Again on using Lemma 2 and (9),

$$\begin{aligned} \|HH^T - RR^T\|_2 &= \|H(I - QQ^T)H^T\|_2 \\ &= \|(I - QQ^T)H^TH\|_2 \\ &= \kappa^2\|W - QQ^TW\|_2. \end{aligned}$$

Then by Proposition 1,

$$\begin{aligned}&\mathbb{E}\,\|HH^T - RR^T\|_2\\ &\quad= \kappa^2\,\mathbb{E}\,\|W - QQ^TW\|_2 \leq \kappa^2\zeta^{1/q}\sigma_{k+1}(W)\\ &\quad= \zeta^{1/q}\sigma_{k+1}(HH^T)\\ &\quad\leq \zeta^{1/q}\sigma_{k+1}(XX^T) + \zeta^{1/q}\|XX^T - HH^T\|_2,\end{aligned}$$

where the last step is due to Proposition 4. Moreover, note from Proposition 1 that $H$ is assumed to be fixed and the expectation above is only taken over the random variable $Q$ (i.e., randomness due to the Gaussian random matrix). Putting all these together, and taking expectation over both $Q$ and $H$ (i.e., also including the randomness in selecting the columns), we have

$$\begin{aligned}&\mathbb{E}\,\|G - \hat{G}\|_2\\ &\quad\leq \mathbb{E}\,\|XX^T - HH^T\|_2 + \mathbb{E}\left(\mathbb{E}_{|H}\,\|HH^T - RR^T\|_2\right)\\ &\quad\leq \zeta^{1/q}\sigma_{k+1}(XX^T) + (1+\zeta^{1/q})\,\mathbb{E}\,\|XX^T - HH^T\|_2\\ &\quad\leq \zeta^{1/q}\|G - G_k\|_2 + (1+\zeta^{1/q})\frac{n}{\sqrt{m}}G^*_{ii}.\end{aligned}$$

The last step uses the fact that $\|G - G_k\|_2 = \sigma_{k+1}(G) = \sigma_{k+1}(XX^T)$ and Lemma 7. □

## D. Error Analysis for the Frobenius Norm

In this section, we obtain a similar bound for the approximation error in terms of the Frobenius norm. Since there is no analogous theory for power iteration w.r.t. the Frobenius norm (remark 10.1 of (Halko et al., 2009)), the analysis is restricted to $q = 1$ and the resultant bound is quite loose. Nevertheless, as will be seen in the experiments, the approximation error (with $q > 1$) is very small.

As in Section 3.2.1, we first consider the error in approximating $W$ from Algorithm 1.

**Corollary 1.** *For the $W$ and $Q$ obtained in Algorithm 1,*

$$\mathbb{E}\,\|W - QQ^TW\|_F \leq \left(1 + \frac{k}{p-1}\right)^{1/2}\left(\sum_{i>k}\sigma_i^2(W)\right)^{1/2}.$$

*Proof.* This is a direct application of Theorem 3. □

**Lemma 9.** *Given matrices $A \in \mathbb{R}^{n\times t}$ and $B \in \mathbb{R}^{n\times s}$, with $n \geq \max\{s,t\}$. Then, for any $k \leq \min\{s,t\}$,*

$$\sum_{i=1}^k\left(\sigma_i^2(A) - \sigma_i^2(B)\right) \leq \sqrt{k}\|AA^T - BB^T\|_F.$$

*Proof.* Using the Cauchy-Schwarz inequality,

$$\begin{aligned}&\sum_{i=1}^k\left(\sigma_i^2(A) - \sigma_i^2(B)\right)\\ &\quad\leq \sqrt{k}\left[\sum_{i=1}^k\left(\sigma_i^2(A) - \sigma_i^2(B)\right)^2\right]^{1/2}\\ &\quad= \sqrt{k}\left[\sum_{i=1}^k\left(\sigma_i(AA^T) - \sigma_i(BB^T)\right)^2\right]^{1/2}\\ &\quad\leq \sqrt{k}\left[\sum_{i=1}^n\left(\sigma_i(AA^T) - \sigma_i(BB^T)\right)^2\right]^{1/2}\\ &\quad\leq \sqrt{k}\|AA^T - BB^T\|_F,\end{aligned}$$

on using Proposition 4. □

**Lemma 10.** *For matrices $A \in \mathbb{R}^{n\times k}$ and $B \in \mathbb{R}^{n\times n}$, with $n \geq k$. Let $U$ be an orthonormal basis of* ran$(A)$*. Then,*

$$\sum_{i=1}^k\sigma_i^2(A) - \|U^TB\|_F^2 \leq \sqrt{k}\|AA^T - BB^T\|_F.$$

*Proof.* First, note that $\sum_{i=1}^k\sigma_i^2(A) = \sum_{i=1}^k\sigma_i(AA^T) = \text{Tr}(U^TAA^TU)$. Let $U^{(i)}$ be the $i$th column of $U$. Then

$$\begin{aligned}&\sum_{i=1}^k\sigma_i^2(A) - \|U^TB\|_F^2\\ &\quad= \text{Tr}(U^TAA^TU) - \text{Tr}(U^TBB^TU)\\ &\quad= \sum_{i=1}^k U^{(i)T}(AA^T - BB^T)U^{(i)}\\ &\quad\leq \sqrt{k}\left[\sum_{i=1}^k\left(U^{(i)T}(AA^T - BB^T)U^{(i)}\right)^2\right]^{1/2}\\ &\quad\leq \sqrt{k}\left[\sum_{i=1}^k\sigma_i^2(AA^T - BB^T)\right]^{1/2}\\ &\quad\leq \sqrt{k}\|AA^T - BB^T\|_F,\end{aligned}$$

where the first inequality follows from the Cauchy-Schwarz inequality. □

**Lemma 11.** *Given $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times n}$, and $k \leq n$,*

$$\left|\sqrt{\sum_{i>k}\sigma_i^2(A)} - \sqrt{\sum_{i>k}\sigma_i^2(B)}\right| \leq \|A - B\|_F. \qquad (16)$$

*Proof.* Using the triangle inequality,

$$\begin{aligned}\left|\sqrt{\sum_{i>k}\sigma_i^2(A)}-\sqrt{\sum_{i>k}\sigma_i^2(B)}\right| &\leq \sqrt{\sum_{i>k}(\sigma_i(A)-\sigma_i(B))^2}\\ &\leq \sqrt{\sum_{i=1}^{n}(\sigma_i(A)-\sigma_i(B))^2}.\end{aligned}$$

On using Proposition 4, we obtain (16). □

**Proof of Theorem 2:**

*Proof.* Let $R = HQ$ and $U_R$ be an orthonormal basis of $\text{ran}(R)$. Then similar to Theorem 1,

$$\|G-\hat{G}\|_F = \|X^TX - X^TU_RU_R^TX\|_F.$$

Since $I - U_RU_R^T$ is an orthogonal projector, it is psd. Thus, for any vector $u$, $u^TX^T(I-U_RU_R^T)Xu = (Xu)^T(I - U_RU_R^T)(Xu) \geq 0$, and so $X^TX - X^TU_RU_R^TX$ is also psd. Using Lemma 4,

$$\begin{aligned}\|G-\hat{G}\|_F &\leq \text{Tr}(X^TX - X^TU_RU_R^TX)\\ &= \|X\|_F^2 - \|U_R^TX\|_F^2. \qquad (17)\end{aligned}$$

Using Lemmas 9 and 10,

$$\begin{aligned}&\|X\|_F^2 - \|U_R^TX\|_F^2\\ &= \|X\|_F^2 - \sum_{i=1}^{k+p}\sigma_i^2(X) + \sum_{i=1}^{k+p}\sigma_i^2(X) - \sum_{i=1}^{k+p}\sigma_i^2(H)\\ &\quad + \sum_{i=1}^{k+p}\sigma_i^2(H) - \sum_{i=1}^{k+p}\sigma_i^2(R) + \sum_{i=1}^{k+p}\sigma_i^2(R) - \|U_R^TX\|_F^2\\ &\leq \sum_{i>k+p}\sigma_i^2(X) + \sqrt{k+p}\|XX^T - HH^T\|_F\\ &\quad + \sqrt{k+p}\|HH^T - RR^T\|_F\\ &\quad + \sqrt{k+p}\|XX^T - RR^T\|_F\\ &\leq \sum_{i>k+p}\sigma_i^2(X) + 2\sqrt{k+p}\|XX^T - HH^T\|_F\\ &\quad + 2\sqrt{k+p}\|HH^T - RR^T\|_F. \qquad (18)\end{aligned}$$

Note that $\sum_{i>k+p}\sigma_i^2(X)$ can be bounded as

$$\sum_{i>k+p}\sigma_i^2(X) \leq \sum_{i=1}^{n}\sigma_i^2(X) = \text{Tr}(X^TX) \leq nG_{ii}^*. \quad (19)$$

Moreover, let $P = I - QQ^T$, which is an orthogonal projector. Then

$$\begin{aligned}\|HH^T - RR^T\|_F^2 &= \|HH^T - HQQ^TH^T\|_F^2\\ &= \text{Tr}\left(HPH^THPH^T\right)\\ &= \text{Tr}\left((HP)(PH^THPPH^T)\right)\\ &= \text{Tr}(PH^THPPH^THP)\\ &= \|PH^THP\|_F\\ &\leq \|PH^TH\|_F\\ &= \|H^TH - QQ^TH^TH\|_F^2\\ &= \kappa^4\|W - QQ^TW\|_F^2,\end{aligned}$$

where the inequality is due to Lemma 5. Using Corollary 1, and let $\zeta_F = \left(1+\frac{k}{p-1}\right)^{1/2}$, then

$$\begin{aligned}&\mathbb{E}\,\|HH^T - RR^T\|_F\\ &= \kappa^2\,\mathbb{E}\,\|W - QQ^TW\|_F\\ &\leq \kappa^2\zeta_F\left[\sum_{i>k}\sigma_i^2(W)\right]^{1/2}\\ &= \zeta_F\left[\sum_{i>k}\sigma_i^2(HH^T)\right]^{1/2}\\ &\leq \zeta_F\left[\sum_{i>k}\sigma_i^2(XX^T)\right]^{1/2} + \zeta_F\|XX^T - HH^T\|_F\\ &= \zeta_F\|G - G_k\|_F + \zeta_F\|XX^T - HH^T\|_F, \qquad (20)\end{aligned}$$

on using Lemma 11. Combining (17) with (18), (19), and (20), we have

$$\begin{aligned}\mathbb{E}\,\|G-\hat{G}\|_F \leq\, &nG_{ii}^* + 2\sqrt{k+p}\,\mathbb{E}\,\|XX^T - HH^T\|_F\\ &+2\sqrt{k+p}\,\mathbb{E}\left(\mathbb{E}_{|H}\,\|HH^T - RR^T\|\right)\\ \leq\, &nG_{ii}^* + 2\zeta_F\sqrt{k+p}\|G - G_k\|_F\\ &+2(1+\zeta_F)\sqrt{k+p}\,\mathbb{E}\,\|XX^T - HH^T\|_F.\end{aligned}$$

Finally, using Proposition 7 and property 3 of Lemma 1, we obtain

$$\begin{aligned}\mathbb{E}\,\|G-\hat{G}\|_F &\leq nG_{ii}^* + 2\zeta_F\sqrt{k+p}\|G-G_k\|_F\\ &\quad +2(1+\zeta_F)\sqrt{k+p}\frac{n}{\sqrt{m}}G_{ii}^*\\ &\leq \frac{2(k+p)}{\sqrt{p-1}}\|G-G_k\|_F\\ &\quad +\left(1+\frac{4(k+p)}{\sqrt{m(p-1)}}\right)nG_{ii}^*.\end{aligned}$$

□