On Sampling-based Approximate Spectral Decomposition

Sanjiv Kumar SANJIVK@GOOGLE.COM
Google Research, New York, NY

Mehryar Mohri MOHRI@CS.NYU.EDU
Courant Institute of Mathematical Sciences and Google Research, New York, NY

Ameet Talwalkar AMEET@CS.NYU.EDU
Courant Institute of Mathematical Sciences, New York, NY

Abstract

This paper addresses the problem of approximate singular value decomposition of large dense matrices that arises naturally in many machine learning applications. We discuss two recently introduced sampling-based spectral decomposition techniques: the Nyström and the Column-sampling methods. We present a theoretical comparison between the two methods and provide novel insights regarding their suitability for various applications. We then provide experimental results motivated by this theory. Finally, we propose an efficient adaptive sampling technique to select *informative* columns from the original matrix. This novel technique outperforms standard sampling methods on a variety of datasets.

1. Introduction

Several common methods in machine learning, such as spectral clustering (Ng et al., 2001) and manifold learning (de Silva & Tenenbaum, 2003), require the computation of singular values and singular vectors of symmetric positive semi-definite (SPSD) matrices. Similarly, kernel-based methods can be sped up by using low-rank approximations of SPSD kernel matrices, which can be achieved via spectral decomposition (Williams & Seeger, 2000; Zhang et al., 2008). The computational complexity of Singular Value Decomposition (SVD) of $n \times n$ SPSD matrices is $O(n^3)$, which presents a major challenge in large-scale applications. Large-scale data sets have been used in several recent studies (Platt, 2004; Chang et al., 2008; Talwalkar et al., 2008). The size of such datasets can be in the order of millions and the $O(n^3)$ complexity is infeasible at this scale.

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009.

Many of the aforementioned applications require only some of the top or bottom singular values and singular vectors. If the input matrix is sparse, one can use efficient iterative methods, e.g., Jacobi or Arnoldi. However, for large dense matrices that arise naturally in many applications, e.g., manifold learning and kernel methods, iterative techniques are also quite expensive. In fact, for many real-world problems, even storing the full dense matrix becomes infeasible. For example, an input set containing 1 million points requires storage of a 4 TB SPSD matrix.

Sampling-based methods provide a powerful alternative for approximate spectral decomposition. They operate on a small part of the original matrix and often eliminate the need for storing the full matrix. In the last decade, two sampling-based approximation techniques have been introduced (Frieze et al., 1998; Williams & Seeger, 2000; Drineas & Mahoney, 2005). However, their similarities and relative advantages have not been well studied. Also, there exist no clear guidelines on which method to use for specific applications. This work introduces a theoretical framework to compare these methods and provides the first exhaustive empirical comparison on a variety of datasets. The analysis and subsequent experiments reveal the counter-intuitive behavior of these methods for different tasks.

Another important component of sampling-based approximations is the sampling strategy used to select *informative* columns of the original matrix. Among the techniques that sample columns according to some fixed distribution, uniform sampling has been shown to be quite effective in practice (Williams & Seeger, 2000; de Silva & Tenenbaum, 2003; Kumar et al., 2009), and a bound on approximation error has also been derived (Kumar et al., 2009). As an improvement over

the fixed-distribution techniques, an adaptive, error-driven sampling technique with better theoretical approximation accuracy was recently introduced (Deshpande et al., 2006). However, this technique requires the full matrix to be available at each step, and is thus infeasible for large matrices. In the second part of this work, we propose a simple and efficient algorithm for adaptive sampling that uses only a small submatrix at each step. In comparison to non-adaptive sampling techniques, we show that the proposed adaptive sampling can provide more accurate low-rank approximation, particularly for higher ranks.

2. Approximate spectral decomposition

Two different sampling-based methods, i.e., Nyström and Column-sampling, have recently been introduced for spectral decomposition of large matrices using subsets of their columns. Let G be an SPSD matrix of size $n \times n$ with spectral decomposition $G = U_G \Sigma_G U_G^\top$, where Σ_G contains the singular values of G and U_G the associated singular vectors. Suppose we randomly sample $l \ll n$ columns of G uniformly *without* replacement.[1] Let C be the $n \times l$ matrix of these sampled columns, and W be the $l \times l$ matrix consisting of the intersection of these l columns with the corresponding l rows of G. Since G is SPSD, W is also SPSD. Without loss of generality, we can rearrange the columns and rows of G such that:

$$G = \begin{bmatrix} W & G_{21}^\top \\ G_{21} & G_{22} \end{bmatrix} \quad \text{and} \quad C = \begin{bmatrix} W \\ G_{21} \end{bmatrix}. \qquad (1)$$

The approximation techniques discussed next use the SVD of W and C to generate approximations of U_G and Σ_G.

2.1. Nyström method

The Nyström method was presented in (Williams & Seeger, 2000) to speed up kernel machines and has been used in applications ranging from manifold learning to image segmentation (Platt, 2004; Fowlkes et al., 2004; Talwalkar et al., 2008). The Nyström method uses W and C from (1) to approximate G as:

$$G \approx \tilde{G} = C W^{-1} C^\top. \qquad (2)$$

If W is not invertible, its pseudoinverse can be used (Drineas & Mahoney, 2005). If $W = U_w \Sigma_w U_w^\top$, the approximate singular values and singular vectors of G are (Williams & Seeger, 2000):

$$\Sigma_{nys} = \left(\frac{n}{l}\right)\Sigma_w \quad \text{and} \quad U_{nys} = \sqrt{\frac{l}{n}} C U_w \Sigma_w^{-1}. \qquad (3)$$

[1] Other sampling schemes are possible (see Section 4).

If $k \leq l$ singular values and singular vectors are needed, the run time of this algorithm is $O(l^3 + nlk)$: l^3 for SVD on W and nlk for multiplication with C.

2.2. Column-sampling method

The Column-sampling method was introduced to approximate the SVD of any rectangular matrix (Frieze et al., 1998). It approximates the spectral decomposition of G by using the SVD of C directly. If $C = U_c \Sigma_c V_c^\top$ and we assume uniform sampling, the approximate singular values and singular vectors of G are given as:

$$\Sigma_{col} = \sqrt{\frac{n}{l}}\Sigma_c \quad \text{and} \quad U_{col} = U_c = C V_c \Sigma_c^{-1}. \qquad (4)$$

The runtime of the Column-sampling method is dominated by the SVD of C. Even when only k singular values and singular vectors are required, the algorithm takes $O(nl^2)$ time and is thus more expensive than Nyström. Often, in practice, the SVD of $C^\top C$ ($O(l^3)$) is performed instead of the SVD of C. However, it is still substantially more expensive than the Nyström method due to the additional cost of computing $C^\top C$.

3. Nyström vs Column-sampling

Given that two sampling-based techniques exist to approximate the SVD of SPSD matrices, we pose a natural question: which method should one use to approximate singular values, singular vectors and low-rank approximations? We first analyze the form of these approximations and then empirically evaluate their performance in Section 3.3 on a variety of datasets.

3.1. Singular values and singular vectors

As shown in (3) and (4), the singular values of G are approximated as the scaled singular values of W and C, respectively. The scaling terms are quite rudimentary and are primarily meant to *compensate* for the 'small sample size' effect for both approximations. However, the form of singular vectors is more interesting. The Column-sampling singular vectors (U_{col}) are orthonormal since they are the singular vectors of C. In contrast, the Nyström singular vectors (U_{nys}) are approximated by *extrapolating* the singular vectors of W as shown in (3), and are *not* orthonormal. It is easy to verify that $U_{nys}^\top U_{nys} \neq I_l$, where I_l is the identity matrix of size l. As we show in Section 3.3, this adversely affects the accuracy of singular vector approximation from the Nyström method.

It is possible to orthonormalize the Nyström singular vectors by using QR decomposition. Since $U_{nys} \propto$

$CU_w\Sigma_w^{-1}$, where U_w is orthogonal and Σ_w is diagonal, this simply implies that QR decomposition creates an orthonormal span of C rotated by U_w. However, the complexity of QR decomposition of U_{nys} is the same as that of the SVD of C. Thus, the computational cost of orthogonalizing U_{nys} would nullify the computational benefit of Nyström over Column-sampling.

3.2. Low-rank approximation

Several studies have shown that the accuracy of low-rank approximations of kernel matrices is tied to the performance of kernel-based learning algorithms (Williams & Seeger, 2000; Talwalkar et al., 2008; Zhang et al., 2008). Hence, accurate low-rank approximations are of great practical interest in machine learning. Suppose we are interested in approximating G with a matrix of rank $k \ll n$, denoted as G_k. It is well-known that the G_k that minimizes the Frobenius norm of the error, i.e., $||G - G_k||_F$, is given by,

$$\hat{G}_k = U_{G,k}\Sigma_{G,k}U_{G,k}^\top = U_{G,k}U_{G,k}^\top G = GU_{G,k}U_{G,k}^\top \quad (5)$$

where $U_{G,k}$ contains the singular vectors of G corresponding to top k singular values contained in $\Sigma_{G,k}$. We refer to $U_{G,k}\Sigma_{G,k}U_{G,k}^\top$ as *Spectral Reconstruction*, since it uses both the singular values and vectors, and $U_{G,k}U_{G,k}^\top G$ as *Matrix Projection*, since it uses only singular vectors to compute the projection of G onto the space spanned by vectors $U_{G,k}$. These two low-rank approximations are equal only if $\Sigma_{G,k}$ and $U_{G,k}$ contain the true singular values and singular vectors of G. Since this is not the case for approximate methods such as Nyström and Column-sampling, these two measures generally give different errors. Thus, we analyze each measure separately in the following sections.

3.2.1. Matrix projection

For Column-sampling, using (4), the low-rank approximation via matrix projection is

$$G_k^{col} = U_{col,k}U_{col,k}^\top G = U_{c,k}U_{c,k}^\top G = C(C^\top C)_k^{-1}C^\top G, \quad (6)$$

where $U_{x,k}$ are the first k vectors of U_x and $(C^\top C)_k = V_{C,k}\Sigma_{C,k}^{-2}V_{C,k}^\top$. Clearly, if $k = l$, $(C^\top C)_k = C^\top C$. Similarly, using (3), the Nyström matrix projection is

$$G_k^{nys} = U_{nys,k}U_{nys,k}^\top G = C(\frac{l}{n}W_k^{-2})C^\top G, \quad (7)$$

where $W_k = U_{w,k}\Sigma_{w,k}U_{w,k}^\top$, and if $k = l$, $W_k = W$.

As shown in (6) and (7), the two methods have similar expressions for matrix projection, except that $C^\top C$ is replaced by a scaled W^2. Furthermore, the scaling term appears only in the Nyström expression. We now present Theorem 1 and Observations 1 and 2, which provide further insights about these two methods.

Theorem 1. *The Column-sampling and Nyström matrix projections are of the form $U_cRU_c^\top G$, where $R \in \mathbb{R}^{l\times l}$ is SPSD. Further, Column-sampling gives the lowest reconstruction error (measured in $\|\cdot\|_F$) among all such approximations if $k = l$.*

Proof. From (6), it is easy to see that

$$G_k^{col} = U_{c,k}U_{c,k}^\top G = U_cR_{col}U_c^\top G, \quad (8)$$

where $R_{col} = \left[\begin{smallmatrix} I_k & 0 \\ 0 & 0 \end{smallmatrix}\right]$. Similarly, from (7) we can derive

$$G_k^{nys} = U_cR_{nys}U_c^\top G \text{ where } R_{nys} = Y\Sigma_{w,k}^{-2}Y^\top, \quad (9)$$

and $Y = \sqrt{l/n}\Sigma_cV_c^\top U_{w,k}$. Note that both R_{col} and R_{nys} are SPSD matrices. Furthermore, if $k = l$, $R_{col} = I$. Let E be the (squared) reconstruction error for an approximation of the form $U_cRU_c^\top G$, where R is an arbitrary SPSD matrix. Hence, when $k = l$, the difference in reconstruction error between the generic and the Column-sampling approximations is

$$\begin{aligned} E - E_{col} =& \|G - U_cRU_c^\top G\|_F^2 - \|G - U_cU_c^\top G\|_F^2 \\ =& \mathrm{Tr}\big[G^\top(I - U_cRU_c^\top)^\top(I - U_cRU_c^\top)G\big] \\ &- \mathrm{Tr}\big[G^\top(I - U_cU_c^\top)^\top(I - U_cU_c^\top)G\big] \\ =& \mathrm{Tr}\big[G^\top(U_cR^2U_c^\top - 2U_cRU_c^\top + U_cU_c^\top)G\big] \\ =& \mathrm{Tr}\big[((R - I)U_c^\top G)^\top((R - I)U_c^\top G)\big] \\ \geq& 0. \end{aligned} \quad (10)$$

We used the fact that $U_c^\top U_c = I$, and that $A^\top A$ is SPSD for any matrix A. □

Observation 1. *For $k = l$, matrix projection for Column-sampling reconstructs C exactly. This can be seen by block-decomposing G as: $G = [C \quad \bar{C}]$, where $\bar{C} = [G_{21} \quad G_{22}]^\top$, and using (6):*

$$G_l^{col} = C(C^\top C)^{-1}C^\top G = [C \quad C(C^\top C)^{-1}C^\top \bar{C}]. \quad (11)$$

Observation 2. *For $k = l$, the span of the orthogonalized Nyström singular vectors equals the span of U_{col}, as discussed in Section 3.1. Hence, matrix projection is identical for Column-sampling and Orthonormal Nyström for $k = l$.*

From an application point of view, matrix projection approximations tend to be more accurate than the spectral reconstruction approximations discussed in the next section (results omitted due to space constraints). However, these low-rank approximations are not necessarily symmetric and require storage of all entries of G. For large-scale problems, this storage requirement may be inefficient or even infeasible.

3.2.2. Spectral reconstruction

Using (3), the Nyström spectral reconstruction is:

$$G_k^{nys} = U_{nys,k}\Sigma_{nys,k}U_{nys,k}^\top = CW_k^{-1}C^\top. \quad (12)$$

When $k = l$, this approximation perfectly reconstructs three blocks of G, and G_{22} is approximated as $G_{21}W^{-1}G_{21}$, which is the Schur Complement of W in G (Williams & Seeger, 2000):

$$G_l^{nys} = CW^{-1}C^\top = \begin{bmatrix} W & G_{21}^\top \\ G_{21} & G_{21}W^{-1}G_{21}^\top \end{bmatrix}. \quad (13)$$

The Column-sampling spectral reconstruction has a similar form as (12):

$$G_k^{col} = U_{col,k}\Sigma_{col,k}U_{col,k}^\top = C\big(\frac{l}{n}(C^\top C)_k\big)^{-\frac{1}{2}}C^\top. \quad (14)$$

In contrast with matrix projection, the scaling term now appears in the Column-sampling reconstruction. To analyze the two approximations, we consider an alternative characterization based on the fact that since G is SPSD, there exists an $X \in \mathbb{R}^{m\times n}$ such that $G = X^\top X$. Similar to (Drineas & Mahoney, 2005), we define a zero-one sampling matrix, $S \in \mathbb{R}^{n\times l}$, that selects l columns from G, i.e., $C = GS$. Each column of S has exactly one non-zero entry per column. Further, $W = S^\top GS = (XS)^\top XS = X'^\top X'$, where $X' \in \mathbb{R}^{m\times l}$ contains l sampled columns of X and $X' = U_{X'}\Sigma_{X'}V_{X'}^\top$ is the SVD of X'. We use these definitions to present Theorems 2 and 3.

Theorem 2. *Column-sampling and Nyström spectral reconstructions are of the form $X^\top U_{X',k}ZU_{X',k}^\top X$, where $Z \in \mathbb{R}^{k\times k}$ is SPSD. Further, among all approximations of this form, neither the Column-sampling nor the Nyström approximation is optimal (in $\|\cdot\|_F$).*

Proof. If $\alpha = \sqrt{n/l}$, then starting from (14) and expressing C and W in terms of X and S, we have

$$\begin{aligned} G_k^{col} &= \alpha GS(S^\top G^2 S)_k^{-1/2}S^\top G^\top \\ &= \alpha X^\top X'\big(V_{C,k}\Sigma_{C,k}^2V_{C,k}^\top\big)^{-1/2}X'^\top X \\ &= X^\top U_{X',k}Z_{col}U_{X',k}^\top X, \end{aligned} \quad (15)$$

where $Z_{col} = \alpha\Sigma_{X'}V_{X'}^\top V_{C,k}\Sigma_{C,k}^{-1}V_{C,k}^\top V_{X'}\Sigma_{X'}$. Similarly, from (12) we have:

$$\begin{aligned} G_k^{nys} &= GS(S^\top GS)_k^{-1}S^\top G^\top \\ &= X^\top X'\big(X'^\top X'\big)_k^{-1}X'^\top X \\ &= X^\top U_{X',k}U_{X',k}^\top X. \end{aligned} \quad (16)$$

Clearly, $Z_{nys} = I$. Next, we analyze the error, E, for an arbitrary Z, which yields the approximation G_k^Z:

$$E = \|G - G_k^Z\|_F^2 = \|X^\top(I - U_{X',k}ZU_{X',k}^\top)X\|_F^2. \quad (17)$$

Let $X = U_X\Sigma_XV_X^\top$ and $Y = U_X^\top U_{X',k}$. Then,

$$\begin{aligned} E &= \mathrm{Tr}\left[\big((I - U_{X',k}ZU_{X',k}^\top)U_X\Sigma_X^2U_X^\top\big)^2\right] \\ &= \mathrm{Tr}\left[\big(U_X\Sigma_XU_X^\top(I - U_{X',k}ZU_{X',k}^\top)U_X\Sigma_XU_X^\top\big)^2\right] \\ &= \mathrm{Tr}\left[\big(U_X\Sigma_X(I - YZY^\top)\Sigma_XU_X^\top\big)^2\right] \\ &= \mathrm{Tr}\left[\Sigma_X(I - YZY^\top)\Sigma_X^2(I - YZY^\top)\Sigma_X\big)\right] \\ &= \mathrm{Tr}\left[\Sigma_X^4 - 2\Sigma_X^2YZY^\top\Sigma_X^2 + \Sigma_XYZY^\top\Sigma_X^2YZY^\top\Sigma_X\big)\right]. \end{aligned} \quad (18)$$

To find $\hat{Z}$, the Z that minimizes (18), we set:

$$\partial E/\partial Z = -2Y^\top\Sigma_X^4Y + 2(Y^\top\Sigma_X^2Y)\hat{Z}(Y^\top\Sigma_X^2Y) = 0$$

and solve for $\hat{Z}$:

$$\hat{Z} = (Y^\top\Sigma_X^2Y)^{-1}(Y^\top\Sigma_X^4Y)(Y^\top\Sigma_X^2Y)^{-1}.$$

Clearly $\hat{Z}$ is different from Z_{col} and Z_{nys}, and since $\Sigma_X^2 = \Sigma_G$, $\hat{Z}$ depends on the spectrum of G. □

While Theorem 2 shows that the optimal approximation is data-dependent and may differ from the Nyström and Column-sampling approximations, Theorem 3 reveals that in certain instances the Nyström method is optimal. In contrast, the Column-sampling method enjoys no such guarantee.

Theorem 3. *Suppose $r = \mathrm{rank}(G) \le k \le l$ and $\mathrm{rank}(W) = r$. Then, the Nyström approximation is exact for spectral reconstruction. In contrast, Column-sampling is exact iff $W = \big((l/n)C^\top C\big)^{1/2}$. Furthermore, when this very specific condition holds, Column-Sampling trivially reduces to the Nyström method.*

Proof. Since $G = X^\top X$, $\mathrm{rank}(G) = \mathrm{rank}(X) = r$. Similarly, $W = X'^\top X'$ implies $\mathrm{rank}(X') = r$. Thus the columns of X' span the columns of X and $U_{X',r}$ is an orthonormal basis for X, i.e., $I - U_{X',r}U_{X',r}^\top \in \mathrm{Null}(X)$. Since $k \ge r$, from (16) we have

$$\|G - G_k^{nys}\|_F = \|X^\top(I - U_{X',r}U_{X',r}^\top)X\|_F = 0. \quad (19)$$

To prove the second part of the theorem, we note that $\mathrm{rank}(C) = r$. Thus, $C = U_{C,r}\Sigma_{C,r}V_{C,r}^\top$ and $(C^\top C)_k^{1/2} = (C^\top C)^{1/2} = V_{C,r}\Sigma_{C,r}V_{C,r}^\top$ since $k \ge r$. If $W = (1/\alpha)(C^\top C)^{1/2}$, then the Column-sampling and Nyström approximations are identical and hence exact. Conversely, to exactly reconstruct G, Column-sampling necessarily reconstructs C exactly. Using $C^\top = [W \quad G_{21}^\top]$ in (14):

$$\begin{aligned} G_k^{col} = G \quad &\Longrightarrow \quad \alpha C(C^\top C)_k^{-1/2}W = C \\ &\Longrightarrow \quad \alpha U_{C,r}V_{C,r}^\top W = U_{C,r}\Sigma_{C,r}V_{C,r}^\top \\ &\Longrightarrow \quad \alpha V_{C,r}V_{C,r}^\top W = V_{C,r}\Sigma_{C,r}V_{C,r}^\top \quad (20) \\ &\Longrightarrow \quad W = \frac{1}{\alpha}(C^\top C)^{1/2}. \quad (21) \end{aligned}$$

Dataset	Data	n	d	Kernel
PIE-2.7K	faces	2731	2304	linear
PIE-7K	faces	7412	2304	linear
MNIST	digits	4000	784	linear
ESS	proteins	4728	16	RBF

Table 1. Description of the datasets used in our experiments (Sim et al., 2002; LeCun & Cortes, 2009; Talwalkar et al., 2008). 'n' denotes the number of points and 'd' denotes the number of features in input space.

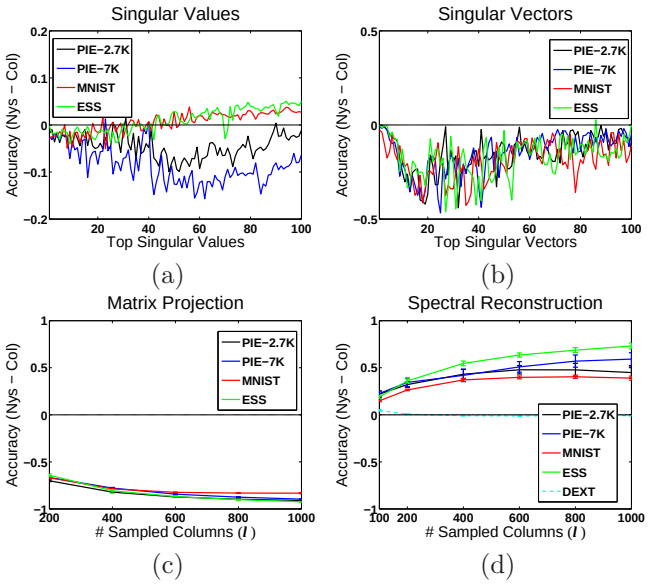


Figure 1. Differences in accuracy between Nyström and Column-Sampling. Values above zero indicate better performance of Nyström and vice-versa. (a) Top 100 singular values with $l = 600$. (b) Top 100 singular vectors with $l = 600$. (c) Matrix projection accuracy for $k = 100$. (d) Spectral reconstruction accuracy for $k = 100$.

In (20) we use $U_{C,r}^\top U_{C,r} = I$, while (21) follows since $V_{C,r}V_{C,r}^\top$ is an orthogonal projection onto the span of the rows of C and the columns of W lie within this span implying $V_{C,r}V_{C,r}^\top W = W$. □

3.3. Empirical comparison

To test the accuracy of singular values/vectors and low-rank approximations for different methods, we used several kernel matrices arising in different applications, as described in Table 1. We worked with datasets containing less than ten thousand points to be able to compare with exact SVD. We fixed k to be 100 in all the experiments, which captures more than 90% of the spectral energy for each dataset.

For singular values, we measured percentage accuracy of the approximate singular values with respect to the exact ones. For a fixed l, we performed 10 trials by selecting columns uniformly at random from G. We show in Figure 1(a) the difference in mean percentage accuracy for the two methods for $l = 600$. The Column-sampling method generates more accurate singular values than the Nyström method for the top 50 singular values, which contain more than 80% of the spectral energy for each of the datasets. A similar trend was observed for other values of l.

For singular vectors, the accuracy was measured by the dot product i.e., cosine of principal angles between the exact and the approximate singular vectors. Figure 1(b) shows the difference in mean accuracy between Nyström and Column-sampling methods. The top 100 singular vectors were all better approximated by Column-sampling for all datasets. This trend was observed for other values of l as well. This result is not surprising since the singular vectors approximated by the Nyström method are not orthonormal.

Next we compared the low-rank approximations generated by the two methods using matrix projection and spectral reconstruction as described in Section 3.2.1 and Section 3.2.2, respectively. We measured the accuracy of reconstruction relative to the optimal rank-k approximation, $\hat{G}_k$, as:

$$\text{relative accuracy} = \frac{||G - \hat{G}_k||_F}{||G - G_k^{nys/col}||_F}. \qquad (22)$$

The relative accuracy will approach one for good approximations. Results are shown in Figure 1(c) and (d). As motivated by Theorem 1 and consistent with the superior performance of Column-sampling in approximating singular values and vectors, Column-sampling generates better reconstructions via matrix projection. This was observed not only for $l = k$ but also for other values of l. In contrast, the Nyström method produces superior results for spectral reconstruction. These results are somewhat surprising given the relatively poor quality of the singular values/vectors for the Nyström method, but they are in agreement with the consequences of Theorem 3. We also note that for both reconstruction measures, the methods that exactly reconstruct subsets of the original matrix when $k = l$ (see (11) and (13)) generate better approximations. Interestingly, these are also the two methods that do not contain scaling terms (see (6) and (12)).

Further, as stated in Theorem 2, the optimal spectral reconstruction approximation is tied to the spectrum of G. Our results suggest that the relative accuracies of Nyström and Column-sampling spectral reconstructions are also tied to this spectrum. When we analyzed spectral reconstruction performance on a sparse kernel

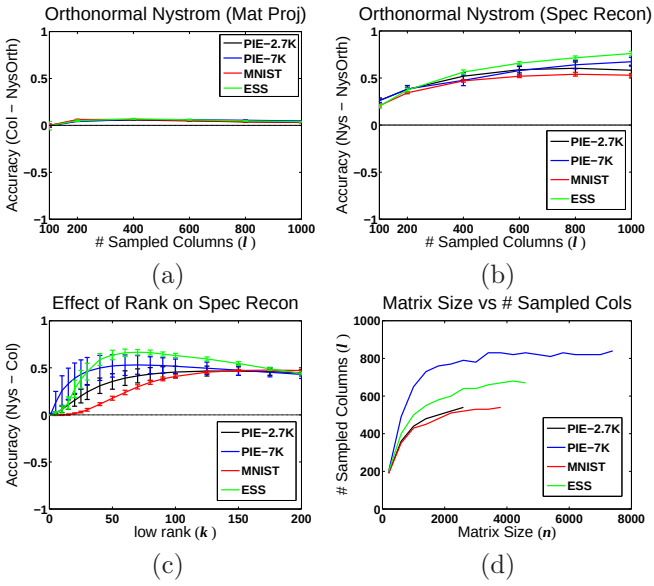


Figure 2. (a) Difference in matrix projection between Column-sampling and Orthonormal Nyström ($k = 100$). Values above zero indicate better performance of Column-sampling. (b) Difference in spectral reconstruction between Nyström and Orthonormal Nyström ($k = 100$). Values above zero indicate better performance of Nyström method. (c) Difference in spectral reconstruction accuracy between Nyström and Column-sampling for various k and fixed $l = 600$. Values above zero indicate better performance of Nyström method. (d) Number of columns needed to achieve 75% relative accuracy for Nyström spectral reconstruction as a function of n.

matrix with a slowly decaying spectrum, we found that Nyström and Column-sampling approximations were roughly equivalent ('DEXT' in Figure 1(d)). This result contrasts the results for dense kernel matrices with exponentially decaying spectra arising from the other datasets used in the experiments.

One factor that impacts the accuracy of the Nyström method for some tasks is the non-orthonormality of its singular vectors (Section 3.1). When orthonormalized, the error in resulting singular vectors is reduced (not shown) and the corresponding Nyström matrix projection error is reduced considerably as shown in Figure 2(a). Further, as discussed in Observation 2 and seen in Figure 2(a) when $l = 100$, Orthonormal Nyström is identical to Column-sampling when $k = l$. However, since orthonormalization is computationally costly, it is avoided in practice. Moreover, the accuracy of Orthogonal Nyström spectral reconstruction is actually worse relative to the standard Nyström approximation, as shown in Figure 2(b). This surprising result can be attributed to the fact that orthonormalization of the singular vectors leads to the loss of some of the unique properties described in Section 3.2.2. For instance, Theorem 3 no longer holds and the scaling terms do not cancel out, i.e., $G_k^{nys} \neq CW_k^{-1}C^\top$.

Even though matrix projection tends to produce more accurate approximations, spectral reconstruction is of great practical interest for large-scale problems since, unlike matrix projection, it does not use all entries in G to produce a low-rank approximation. Thus, we further expand upon the results from Figure 1(d). We first tested the accuracy of spectral reconstruction for the two methods for varying values of k and a fixed l. We found that the Nyström method outperforms Column-sampling across all tested values of k, as shown in Figure 2(c). Next, we addressed another basic issue: how many columns do we need to obtain reasonable reconstruction accuracy? For very large matrices ($n \approx O(10^6)$), one would wish to select only a small fraction of the samples. Hence, we performed an experiment in which we fixed k and varied the size of our dataset (n). For each n, we performed grid search over l to find the minimal l for which the relative accuracy of Nyström spectral reconstruction was at least 75%. Figure 2(d) shows that the required l does not grow linearly with n. The $\frac{l}{n}$ ratio actually decreases quickly as n increases, lending support to the use of sampling-based algorithms for large-scale data.

4. Sampling Techniques

In Section 3, we focused on uniformly sampling columns to create low-rank approximations. Since approximation techniques operate on a small subset of G, i.e., C, the selection of columns can significantly influence the accuracy of approximation. In this section we discuss various sampling options that aim to select informative columns from G. The most common sampling techniques select columns using a fixed probability distribution, with *uniform* sampling being the most basic of these *non-adaptive* approaches. Alternatively, the ith column can be sampled non-uniformly with a weight that is proportional to either the corresponding diagonal element, G_{ii} (*diagonal*) or the squared L_2 norm of the column (*column-norm*). The non-adaptive sampling methods have been combined with SVD approximation algorithms to bound the reconstruction error (Drineas & Mahoney, 2005; Kumar et al., 2009). Interestingly, the non-uniform approaches are often outperformed by uniform sampling for dense matrices (Kumar et al., 2009). Contrary to the non-adaptive sampling methods, an adaptive sampling technique with better theoretical approximation accuracy (*adaptive-full*) was proposed in (Deshpande et al., 2006). It requires a full pass through G in each

iteration. Another interesting technique that selects informative columns based on k-means clustering has been shown to give good empirical accuracy (Zhang et al., 2008). However, both methods are computationally inefficient for large G.

4.1. Proposed Adaptive Sampling Method

Instead of sampling all l columns from a fixed distribution, adaptive sampling alternates between selecting a set of columns and updating the distribution over all the columns. Starting with an initial distribution over the columns, $s < l$ columns are chosen to form a set C'. The probabilities are then updated as a function of previously chosen columns and s new columns are sampled and incorporated in C'. This process is repeated until l columns have been selected.

Input: $n \times n$ SPSD matrix G, number of columns to be chosen (l), initial probability distribution over $[1 \dots n]$ (P_0), number of columns selected at each iteration (s)
Output: l indices corresponding to columns of G

SAMPLE-ADAPTIVE(G, n, l, P_0, s)
1 $R \leftarrow$ set of s indices sampled according to P_0
2 $t \leftarrow \frac{l}{s} - 1$ ▷ number of iterations
3 **for** $i \in [1 \dots t]$ **do**
4 $\quad P_i \leftarrow$ UPDATE-PROBABILITY-PARTIAL(R)
5 $\quad R_i \leftarrow$ set of s indices sampled according to P_i
6 $\quad R \leftarrow R \cup R_i$
7 **return** R

UPDATE-PROBABILITY-PARTIAL(R)
1 $C' \leftarrow$ columns of G corresponding to indices in R
2 $k' \leftarrow$ CHOOSE-RANK() ▷ low rank (k) or $\frac{|R|}{2}$
3 $\Sigma_{nys}, U_{nys} \leftarrow$ DO-NYSTRÖM (C', k') ▷ see eq (3)
4 $C'_{nys} \leftarrow$ Spectral reconstruction using Σ_{nys}, U_{nys}
5 $E \leftarrow C' - C'_{nys}$
6 **for** $j \in [1 \dots n]$ **do**
7 $\quad$ **if** $j \in R$ **then**
8 $\quad\quad P_j \leftarrow 0$ ▷ sample without replacement
9 $\quad$ **else** $P_j \leftarrow ||E_j||_2^2$
10 $P \leftarrow \frac{P}{||P||_2}$
11 **return** P

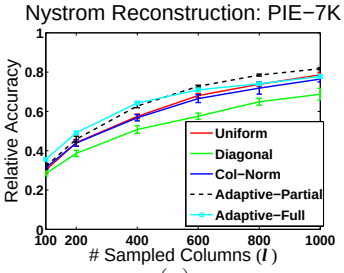
Figure 3. The proposed adaptive sampling technique that uses only a small subset of the original matrix G to compute probability distribution over columns. Note that it does not need to store or run a full pass over G.

We propose a simple sampling technique (*adaptive-partial*) that incorporates the advantages of adaptive sampling while avoiding the computational and storage burdens of the technique in (Deshpande et al., 2006). At each iterative step, we measure the reconstruction error for each *row* of C' and the distribution over corresponding *columns* of G is updated proportional to this error. Unlike (Deshpande et al., 2006), we compute the error for C', which is much smaller than G, thus avoiding the $O(n^2)$ computation. As described in (13), if k' is fixed to be the number of columns in C', it will lead to $C'_{nys} = C'$ resulting in perfect reconstruction of C'. So, one must choose a smaller k' to generate non-zero reconstruction errors from which probabilities can be updated (we used $k' = (\#\text{ columns in } C')/2$ in our experiments). One artifact of using a k' smaller than the rank of C' is that all the columns of G will have a non-zero probability of being selected, which could lead to the selection of previously selected columns in the next iteration. However, sampling *without* replacement strategy alleviates this problem. Working with C' instead of G to iteratively compute errors makes this algorithm significantly more efficient than that of (Deshpande et al., 2006), as each iteration is $O(nlk' + l^3)$ and requires at most the storage of l columns of G. The details of the proposed sampling technique are outlined in Figure 3.

4.2. Sampling Experiments

We used the datasets already shown in Table 1, and compared the effect of different sampling techniques on the relative accuracy of Nyström spectral reconstruction for $k = 100$. The results for PIE-7K are presented in Figure 4(a) for varying values of l. The results across datasets (Figure 4(b)) show that our adaptive-partial sampling technique outperforms all non-adaptive methods. They show that adaptive-partial performs roughly the same as adaptive-full for smaller l and outperforms it for larger l, while being much cheaper computationally (Figure 5(a)).

Next, we wished to identify the situations where adaptive sampling is most effective. It is well-known that most matrices arising in real-world applications exhibit a fast decaying singular value spectrum. For these matrices, sampling based spectral decomposition methods generally provide accurate estimates of the top few singular values/vectors. However, the accuracy generally deteriorates for subsequent singular values/vectors. To test this behavior, we conducted an experiment with the PIE-7K dataset, where the relative accuracy for Nyström spectral reconstruction was measured by varying k for a fixed l. As shown in Figure 5(b), the relative accuracy for all sampling methods decreases as k is increased, thus verifying the deterioration in quality of singular value/vector approximation as k increases. However, by performing error-driven sampling, the proposed adaptive sampling



(a)

l	Dataset	Uniform	Diagonal	Col-Norm	Adapt-Part	Adapt-Full
400	PIE-2.7K	67.2 (±1.1)	62.1 (±0.9)	59.7 (±1.0)	70.4 (±0.9)	**72.6** (±**1.0**)
	PIE-7K	57.5 (±1.1)	50.8 (±1.9)	56.8 (±1.6)	62.8 (±0.9)	**64.3** (±**0.7**)
	MNIST	67.4 (±0.7)	67.4 (±0.4)	65.3 (±0.5)	**69.3** (±**0.7**)	69.2 (±0.7)
	ESS	61.0 (±1.7)	61.5 (±1.5)	57.5 (±1.9)	**65.0** (±**1.0**)	63.9 (±0.9)
800	PIE-2.7K	84.1 (±0.5)	77.8 (±0.6)	73.9 (±1.0)	86.5 (±0.4)	**87.7** (±**0.4**)
	PIE-7K	73.8 (±1.2)	64.9 (±1.8)	71.8 (±3.0)	**78.5** (±**0.5**)	74.1 (±0.6)
	MNIST	83.3 (±0.3)	83.0 (±0.3)	80.4 (±0.4)	**84.2** (±**0.4**)	80.7 (±0.5)
	ESS	78.1 (±1.0)	79.2 (±0.9)	75.4 (±1.2)	**80.6** (±**1.1**)	74.8 (±0.8)

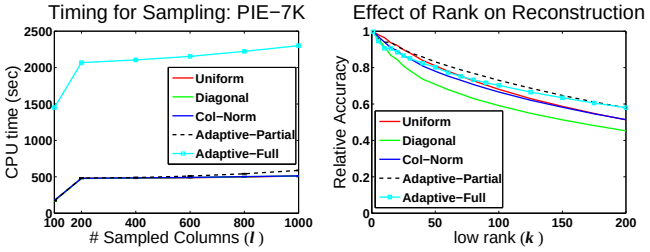
(b)

Figure 4. Nyström spectral reconstruction accuracy for various sampling methods for $k = 100$. (a) Results for PIE-7K using several values of l. (b) Results for all datasets for two values of l with $k = 100$. Numbers in parenthesis indicate the standard deviations for 10 different runs for each l.

provides better relative accuracy as k increases.

5. Conclusions and Future Work

We presented an analysis of two sampling-based techniques for approximating SVD on large dense SPSD matrices, and provided a theoretical and empirical comparison. Although the Column-sampling method generates more accurate singular values/vectors and low-rank matrix projections, the Nyström method constructs better low-rank spectral approximations, which are of great practical interest as they do not use the full matrix. We also presented a novel adaptive sampling technique that results in improved performance over standard techniques and is significantly more efficient than the existing adaptive method. An important question left to study is how different properties of SPSD matrices, e.g., sparsity and singular value spectrum, affect the quality of SVD approximations and the effectiveness of various sampling techniques.



(a) (b)

Figure 5. Results on PIE-7K for different sampling techniques. Left: Empirical run times (Matlab) for Nyström method for $k = 100$. Right: Mean Nyström spectral reconstruction accuracy for varying k and fixed $l = 600$.

References

Chang, E., Zhu, K., Wang, H., Bai, H., Li, J., Qiu, Z., & Cui, H. (2008). Parallelizing support vector machines on distributed computers. *NIPS* (pp. 257–264).

de Silva, V., & Tenenbaum, J. (2003). Global versus local methods in nonlinear dimensionality reduction. *NIPS* (pp. 705–712).

Deshpande, A., Rademacher, L., Vempala, S., & Wang, G. (2006). Matrix approximation and projective clustering via volume sampling. *SODA* (pp. 1117–1126).

Drineas, P., & Mahoney, M. W. (2005). On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *JMLR*, *6*, 2153–2175.

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

Frieze, A., Kannan, R., & Vempala, S. (1998). Fast Monte-Carlo algorithms for finding low-rank approximations. *FOCS* (pp. 370–378).

Kumar, S., Mohri, M., & Talwalkar, A. (2009). Sampling techniques for the Nyström method. *AISTATS* (pp. 304–311). Clearwater Beach, Florida: JMLR: W&CP 5.

LeCun, Y., & Cortes, C. (2009). The MNIST database of handwritten digits, http://yann.lecun.com/exdb/mnist/.

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2001). On spectral clustering: analysis and an algorithm. *NIPS* (pp. 849–856).

Platt, J. C. (2004). Fast embedding of sparse similarity graphs. *NIPS*.

Sim, T., Baker, S., & Bsat, M. (2002). The CMU PIE database. *Conference on Automatic Face and Gesture Recognition*.

Talwalkar, A., Kumar, S., & Rowley, H. (2008). Large-scale manifold learning. *CVPR*.

Williams, C. K. I., & Seeger, M. (2000). Using the Nyström method to speed up kernel machines. *NIPS* (pp. 682–688).

Zhang, K., Tsang, I., & Kwok, J. (2008). Improved Nyström low-rank approximation and error analysis. *ICML* (pp. 273–297).