Sampling Techniques for the Nyström Method

Sanjiv Kumar
Google Research
sanjivk@google.com

Mehryar Mohri
Courant Institute and Google Research
mohri@cs.nyu.edu

Ameet Talwalkar
Courant Institute, NYU
ameet@cs.nyu.edu

Abstract

The Nyström method is an efficient technique to generate low-rank matrix approximations and is used in several large-scale learning applications. A key aspect of this method is the distribution according to which columns are sampled from the original matrix. In this work, we present an analysis of different sampling techniques for the Nyström method. Our analysis includes both empirical and theoretical components. We first present novel experiments with several real world datasets, comparing the performance of the Nyström method when used with uniform versus non-uniform sampling distributions. Our results suggest that uniform sampling without replacement, in addition to being more efficient both in time and space, produces more effective approximations. This motivates the theoretical part of our analysis which gives the first performance bounds for the Nyström method precisely when used with uniform sampling without replacement.

1 Introduction

A common problem in many areas of large-scale machine learning involves deriving a useful and efficient approximation of a large matrix. This matrix may be a kernel matrix used with support vector machines (Boser *et al.*, 1992; Cortes and Vapnik, 1995), kernel principal component analysis (Schölkopf *et al.*, 1998) or manifold learning (Platt, 2003; Talwalkar *et al.*, 2008). Large matrices also naturally arise in other applications such as clustering. For these large-scale problems, the number of matrix entries can be in the order of tens of thousands to millions, leading to difficulty in operating on, or even storing the matrix.

An attractive solution to this problem involves using the Nyström method to generate a low-rank approximation of the original matrix from a subset of its columns (Williams and Seeger, 2000). A key aspect of the Nyström method is the distribution according to which the columns are sampled. This method was first introduced to the machine learning community (Williams and Seeger, 2000) using uniform sampling without replacement, and this remains the sampling method most commonly used in practice (de Silva and Tenenbaum, 2002; Fowlkes *et al.*, 2004; Platt, 2003; Talwalkar *et al.*, 2008). More recently, the Nyström method has been theoretically analyzed assuming a non-uniform sampling of the columns: Drineas and Mahoney (2005) provided bounds for the Nyström approximation while sampling with replacement from a distribution with weights proportional to the diagonal elements of the input matrix.

This paper presents an analysis of different sampling techniques for the Nyström method[1]. Our analysis includes both empirical and theoretical components. We first present novel experiments with several real-world datasets, comparing the performance of the Nyström method when used with uniform versus non-uniform sampling distributions. Although previous works have compared uniform and non-uniform distributions in a more restrictive setting (Drineas *et al.*, 2001; Zhang *et al.*, 2008), our results are the first to compare uniform sampling with the sampling technique for which the Nyström method has theoretical guarantees. Our results suggest that uniform sampling, in addition to being more efficient both in time and space, produces more effective approximations. We further show the benefits of sampling without replacement. These empirical findings motivate the theoretical part of our analysis. We give the

Appearing in Proceedings of the 12^{th} International Confe-rence on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5. Copyright 2009 by the authors.

[1] In this work, we consider only those sampling methods for which the distribution over columns remains fixed throughout the procedure. There exist other *adaptive sampling* techniques which tend to perform better but are usually quite expensive in practice (Deshpande *et al.*, 2006).

first performance bounds for the Nyström method as it is used in practice, i.e., using uniform sampling without replacement.

The remainder of the paper is organized as follows. Section 2 introduces basic definitions and gives a brief presentation of the Nyström method. In Section 3, we provide an extensive empirical comparison of various sampling methods used with the Nyström method. Section 4 presents our novel bound for the Nyström method in the scenario of uniform sampling without replacement, and provides an analysis of the bound.

2 Preliminaries

Let $G \in \mathbb{R}^{n\times n}$ be a symmetric positive semidefinite (SPSD) Gram (or kernel) matrix. For any such Gram matrix, there exists an $X \in \mathbb{R}^{m\times n}$ such that $G = X^\top X$. We define $X^{(j)}$, $j = 1 \dots n$, as the jth column vector of X and $X_{(i)}$, $i = 1 \dots m$, as the ith row vector of X, and denote by $\|\cdot\|$ the l_2 norm of a vector. Using singular value decomposition (SVD), the Gram matrix can be written as $G = U\Sigma U^\top$, where U is orthogonal and $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_n)$ is a real diagonal matrix with diagonal entries sorted in decreasing order. For $r = \mathrm{rank}(G)$, the pseudo-inverse of G is defined as $G^+ = \sum_{t=1}^r \sigma_t^{-1} U^{(t)} U_{(t)}$. Further, for $k \le r$, $G_k = \sum_{t=1}^k \sigma_t U^{(t)} U_{(t)}$ is the 'best' rank-k approximation to G, or the rank-k matrix with minimal $\|\cdot\|_F$ distance to G, where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix.

The Nyström method generates low-rank approximations of G using a subset of the columns of the matrix (Williams and Seeger, 2000). Suppose we randomly sample $l \ll n$ columns of G uniformly without replacement.[2] 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 based on this sampling such that:

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

The Nyström method uses W and C from (1) to construct a rank-k approximation $\tilde{G}_k$ to G for $k \le l$. When used with uniform sampling, the Nyström approximation is:

$$\tilde{G}_k = C W_k^+ C^\top \approx G. \tag{2}$$

The Frobenius distance between G and $\tilde{G}_k$, $\|G - \tilde{G}_k\|_F$, is one standard measurement of the accuracy of the Nyström method. The runtime of this algorithm is O(l^3+nlk): $O(l^3)$ for SVD on W and $O(nlk)$ for multiplication with C.

Name	Type	n	d	Kernel
PIE-2.7K	faces (profile)	2731	2304	linear
PIE-7K	faces (front)	7412	2304	linear
MNIST	digit images	4000	784	linear
ESS	proteins	4728	16	RBF
ABN	abalones	4177	8	RBF

Table 1: Description of the datasets and kernels used in our experiments (Asuncion and Newman, 2007; Gustafson *et al.*, 2006; LeCun and Cortes, 2009; Sim *et al.*, 2002). 'd' denotes the number of features in input space.

3 Comparison of Sampling Methods

Since the Nyström method operates on a subset of G, i.e., C, the selection of columns can significantly influence the accuracy of approximation. Thus, in this section we discuss various sampling options used to select columns from G.

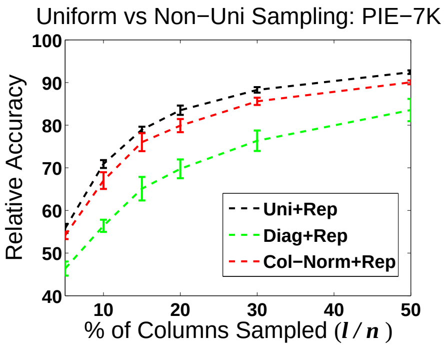
3.1 Description of Sampling Methods

The most basic sampling technique involves *uniform* sampling of the columns. Alternatively, the ith column can be sampled non-uniformly with weight proportional to either its corresponding diagonal element G_{ii} (*diagonal sampling*) or the l_2 norm of the column (*column-norm sampling*) (Drineas and Mahoney, 2005; Drineas *et al.*, 2006b). There are additional computational costs associated with these non-uniform sampling methods: $O(n)$ time and space requirements for diagonal sampling and $O(n^2)$ time and space for column-norm sampling. These non-uniform sampling techniques are often presented using sampling with replacement to simplify theoretical analysis. Column-norm sampling has been used to analyze a general SVD approximation algorithm. Further, diagonal sampling with replacement was used by Drineas and Mahoney (2005) to bound the reconstruction error of the Nyström method,[3] though the authors of that work suggest that column-norm sampling would be a better sampling assumption for the analysis of the Nyström method.

Two other techniques have also been introduced for sampling-based techniques to generate low-rank approximations. The first method adaptively samples columns of G while the second performs k-means clustering as a preprocessing step to construct informative columns (Deshpande *et al.*, 2006; Zhang *et al.*, 2008). Although these methods show good empirical accuracy on small datasets,

[2] Other sampling schemes are also possible as we discuss in Section 3. The formulation of the Nyström method under these sampling schemes is identical to the one presented here, modulo an additional step to normalize the approximation by the probabilities of the selected columns (Drineas and Mahoney, 2005).

[3] Although Drineas and Mahoney (2005) claimed to weight each column proportional to G_{ii}^2, they in fact use the diagonal sampling we present in this work, i.e., weights proportional to G_{ii} (Drineas, 2008).



(a)

l/n	Dataset	Uniform+Rep	Diag+Rep	Col-Norm+Rep
5%	PIE-2.7K	**38.8** ($\pm$**1.5**)	38.3 ($\pm$0.9)	37.0 ($\pm$0.9)
	PIE-7K	**55.8** ($\pm$**1.1**)	46.4 ($\pm$1.7)	54.2 ($\pm$0.9)
	MNIST	**47.4** ($\pm$**0.8**)	46.9 ($\pm$0.7)	45.6 ($\pm$1.0)
	ESS	**45.1** ($\pm$**2.3**)	-	41.0 ($\pm$2.2)
	ABN	**47.3** ($\pm$**3.9**)	-	44.2 ($\pm$1.2)
20%	PIE-2.7K	**72.3** ($\pm$**0.9**)	65.0 ($\pm$0.9)	63.4 ($\pm$1.4)
	PIE-7K	**83.5** ($\pm$**1.1**)	69.8 ($\pm$2.2)	79.9 ($\pm$1.6)
	MNIST	**80.8** ($\pm$**0.5**)	79.4 ($\pm$0.5)	78.1 ($\pm$0.5)
	ESS	**80.1** ($\pm$**0.7**)	-	75.5 ($\pm$1.1)
	ABN	**77.1** ($\pm$**3.0**)	-	66.3 ($\pm$4.0)

(b)

Figure 1: (a) Nyström relative accuracy for various sampling techniques on PIE-7K. (b) Nyström relative accuracy for various sampling methods for two values of l/n with $k = 100$. Values in parentheses show standard deviations for 10 different runs for a fixed l. '+Rep' denotes sampling with replacement. No error ('-') is reported for diagonal sampling with RBF kernels since diagonal sampling is equivalent to uniform sampling in this case.

they are both computationally inefficient for large-scale problems. Adaptive sampling requires a full pass through G on each iteration, while k-means clustering quickly becomes intractable for moderately large n. For this reason, in this work we focus on fixed distributions – either uniform or non-uniform – over the set of columns.

In the remainder of this section we present novel experimental results comparing the performance of these sampling methods on several data sets. Previous works have compared uniform and non-uniform in a more restrictive setting, using fewer types of kernels and focusing only on column-norm sampling (Drineas *et al.*, 2001; Zhang *et al.*, 2008). However in this work we provide the first comparison that includes diagonal sampling, the sampling technique for which the Nyström method has theoretical guarantees.

3.2 Datasets

We used 5 datasets from a variety of applications, e.g., computer vision and biology, as described in Table 1. SPSD kernel matrices were generated by mean centering the datasets and applying either a linear kernel or RBF kernel. The diagonals (respectively column norms) of these kernel matrices were used to calculate diagonal (respectively column-norm) distributions. Note that the diagonal distribution equals the uniform distribution for RBF kernels since diagonal entries of RBF kernel matrices always equal one.

3.3 Experiments

We used the datasets described in the previous section to test the approximation accuracy for each sampling method. Low-rank approximations of G were generated using the Nyström method along with these sampling methods, and accuracies were measured relative to the best rank-k approximation (G_k) as follows:

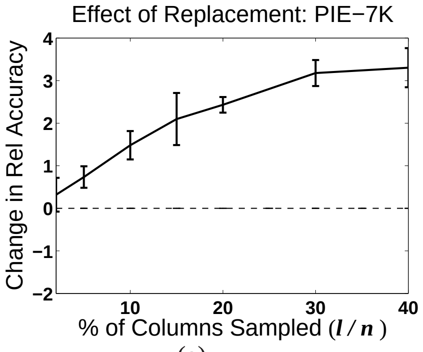
$$\text{relative accuracy} = \frac{\|G - G_k\|_F}{\|G - \tilde{G}_k\|_F}.$$

Note that relative accuracy is upper bounded by 1 and approaches 1 for good approximations. We fixed $k = 100$ for all experiments, a value that captures more than 90% of the spectral energy for each dataset. We first compared the effectiveness of the three sampling techniques using sampling with replacement. The results for PIE-7K are presented in Figure 1(a) and summarized for all datasets in Figure 1(b). The results across all datasets show that uniform sampling outperforms all other methods, while being much cheaper computationally and space-wise. Thus, while non-uniform sampling techniques might be effective in extreme cases where a few columns of G dominate in terms of $\|\cdot\|$, this situation does not tend to arise with real-world data, where uniform sampling is most effective.

Next, we compared the performance of uniform sampling with and without replacement. Figure 2(a) illustrates the effect of replacement for the PIE-7K dataset for different $\frac{l}{n}$ ratios. Similar results for the remaining datasets are summarized in Figure 2(b). The results show that uniform sampling without replacement improves the accuracy of the Nyström method over sampling with replacement, even when sampling less than 5% of the total columns.

4 Improved Nyström Bound

The experimental results from Section 3 show that uniform sampling is the cheapest and most efficient sampling technique across several datasets. Further, it is the most commonly used method in practice. However, there does not currently exist a formal analysis of the accuracy of the Nyström approximation when using uniform sampling without replacement. We next present a theoretical analy-



(a)

Dataset	5%	10%	15%	30%
PIE-2.7K	0.8 ($\pm$.6)	1.7 ($\pm$.3)	2.3 ($\pm$.9)	4.4 ($\pm$.4)
PIE-7K	0.7 ($\pm$.3)	1.5 ($\pm$.3)	2.1 ($\pm$.6)	3.2 ($\pm$.3)
MNIST	1.0 ($\pm$.5)	1.9 ($\pm$.6)	2.3 ($\pm$.4)	3.4 ($\pm$.4)
ESS	0.9 ($\pm$.9)	1.8 ($\pm$.9)	2.2 ($\pm$.6)	3.7 ($\pm$.7)
ABN	0.7 ($\pm$1.2)	1.3 ($\pm$1.8)	2.6 ($\pm$1.4)	4.5 ($\pm$1.1)

(b)

Figure 2: Comparison of uniform sampling with and without replacement measured by the difference in relative accuracy. (a) Improvement in relative accuracy for PIE-7K when sampling without replacement. (b) Improvement in relative accuracy when sampling without replacement across all datasets for various l/n percentages.

sis of the Nyström method using the more reasonable assumption of *uniform sampling without replacement*. We first introduce a general concentration bound for sampling without replacement (Section 4.1), and use it to derive a general bound on approximate matrix multiplication in the setting of sampling without replacement (Section 4.2). In Section 4.3, following Drineas and Mahoney (2005), we show the connection between the Nyström method and approximate matrix multiplication and present our main result: a general bound for the Nyström method in the scenario of uniform sampling without replacement.

4.1 Concentration Bound for Sampling Without Replacement

We will be using the following concentration bound for sampling without replacement shown by Cortes *et al.* (2008) which holds for *symmetric functions*. A function $\phi\colon \mathcal{X}^m \to \mathbb{R}$ defined over a set $\mathcal{X}$ is said to be symmetric if $\phi(x_1, \ldots, x_m) = \phi(x_{\tau(1)}, \ldots, x_{\tau(m)})$ for any $x_1, \ldots, x_m \in X$ and any permutation τ of $(1, \ldots, m)$.

Theorem 1. *Let m and u be positive integers, $x_1, \ldots, x_m$ a sequence of random variables sampled from an underlying set $\mathcal{X}$ of $m+u$ elements without replacement, and let $\phi : \mathcal{X}^m \mapsto \mathbb{R}$ be a symmetric function such that for all $i \in [1, m]$ and for all $x_1, \ldots, x_m \in \mathcal{X}$ and $x'_1, \ldots, x'_m \in \mathcal{X}$,*

$$|\phi(x_1, \ldots, x_m) - \phi(x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots, x_m)| \leq \Delta,$$

where Δ is a positive real number. Then $\forall\, \epsilon > 0$,

$$\Pr\left[\left|\phi - \mathrm{E}[\phi]\right| \geq \epsilon\right] \leq 2 \exp\left(\frac{-2\epsilon^2}{\alpha(m,u)\Delta^2}\right), \tag{3}$$

where $\alpha(m,u) = \frac{mu}{m+u-1/2} \cdot \frac{1}{1-1/(2\max\{m,u\})}$.

4.2 Concentration Bound for Matrix Multiplication

To derive a bound for the Nyström method using uniform sampling without replacement, we first present a generalization of a bound on approximate matrix multiplication given by Drineas *et al.* (2006a) to the more complex setting of uniform sampling without replacement. This generalization is not trivial since previous inequalities hinge upon a key i.i.d. assumption which clearly does not hold when sampling without replacement.

Theorem 2. *Suppose $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$, $1 \leq l \leq n$. Choose a set (S) of size l uniformly at random without replacement from $\{1 \ldots n\}$, and let C (R) equal the columns of A (rows of B) corresponding to indices in S scaled by $\sqrt{n/l}$. Then CR is an approximation to AB, i.e.,*

$$AB = \sum_{t=1}^{n} A^{(t)} B_{(t)} \approx \sum_{t=1}^{l} C^{(t)} R_{(t)} = \frac{n}{l} \sum_{t \in S} A^{(t)} B_{(t)} = CR,$$

and,

$$\mathrm{E}\left[\|AB - CR\|_F\right] \leq \sqrt{\frac{n}{l} \sum_{t=1}^{n} \|A^{(t)}\|^2 \|B_{(t)}\|^2}. \tag{4}$$

Further, let $\delta \in (0,1)$, $t^* = \mathrm{argmax}_t \|A^{(t)}\| \|B_{(t)}\|$, and $\eta = \sqrt{\frac{\log(2/\delta)\alpha(l, n-l)}{l}}$, with $\alpha(l, n-l)$ defined in Theorem 1. Then, with probability at least $1-\delta$,

$$\|AB - CR\|_F \leq \sqrt{\frac{n}{l} \sum_{t=1}^{n} \|A^{(t)}\|^2 \|B_{(t)}\|^2} + \frac{\sqrt{2}\eta n}{\sqrt{l}} \|A^{(t^*)}\| \|B_{(t^*)}\|. \tag{5}$$

We note that for even moderately sized l and n, $\alpha(l, n-l) \approx l(1 - l/n)$ and thus $\eta \approx \sqrt{\log(2/\delta)(1 - l/n)}$.

Corollary 1. *If $A = B^\top$ and $t^* = \mathrm{argmax}_t \|A^{(t)}\|$, then*

$$\mathrm{E}\left[\|AA^\top - CC^\top\|_F\right] \leq \sqrt{\frac{n}{l} \sum_{t=1}^{n} \|A^{(t)}\|^4}. \tag{6}$$

Further, let $\delta \in (0,1)$ *and* $\eta = \sqrt{\frac{\log(2/\delta)\alpha(l,n-l)}{l}}$. *Then, with probability at least* $1-\delta$,

$$\|AA^\top - CC^\top\|_F \leq \sqrt{\frac{n}{l}\sum_{t=1}^{n}\|A^{(t)}\|^4} + \frac{\eta n}{\sqrt{l}}\|A^{(t^*)}\|^2. \quad (7)$$

In this special case, we use the tighter Lipschitz condition defined in (26). Further, since $\sum_{t=1}^{n}\|A^{(t)}\|^4 \leq n\|A^{(t^*)}\|^4$ we can simplify Corollary 1 as follows:

Corollary 2. *If* $A = B^\top$ *then*

$$\mathrm{E}\left[\|AA^\top - CC^\top\|_F\right] \leq \frac{n}{\sqrt{l}}\|A^{(t^*)}\|^2. \quad (8)$$

Further, let $\delta \in (0,1)$, $t^* = \operatorname{argmax}_t \|A^{(t)}\|$, *and* $\eta = \sqrt{\frac{\log(2/\delta)\alpha(l,n-l)}{l}}$. *Then, with probability at least* $1-\delta$,

$$\|AA^\top - CC^\top\|_F \leq (1+\eta)\frac{n}{\sqrt{l}}\|A^{(t^*)}\|^2. \quad (9)$$

The proof of this theorem and its corollaries involves bounding an expectation, determining a Lipschitz condition and using the concentration bound of Theorem 1. These three steps are presented in detail below.

Bound on Expectation

To obtain a bound for $\mathrm{E}\left[\|AB - CR\|_F\right]$, we first calculate expressions for the mean and variance of the (i,j)th component of CR, i.e., $(CR)_{ij}$. For any set S of distinct elements in $\{1\ldots n\}$, $|S| = l$, we define $\pi(S)$ as the probability that a randomly chosen subset of l elements equals S. There are a total of $\binom{n}{l}$ distinct sets and in the uniform case, $\pi(S) = 1/\binom{n}{l}$. Furthermore, each element in $\{1\ldots n\}$ appears in l/n of these distinct sets. Thus, the following equalities hold:

$$\mathrm{E}[(CR)_{ij}] = \sum_{k=1}^{\binom{n}{l}}\pi(S_k)\cdot\left[\sum_{t\in S_k}\frac{n}{l}A_{it}B_{tj}\right] \quad (10)$$

$$= \frac{n}{l\binom{n}{l}}\sum_{t=1}^{n}\frac{l\binom{n}{l}}{n}A_{it}B_{tj} \quad (11)$$

$$= (AB)_{ij}. \quad (12)$$

Further, we have

$$\mathrm{E}[(CR)_{ij}]^2 = (AB)_{ij}^2 = \left(\sum_{t=1}^{n}A_{it}B_{tj}\right)^2 \quad (13)$$

and

$$\mathrm{E}[(CR)_{ij}^2] = \sum_{k=1}^{\binom{n}{l}}\pi(S_k)\cdot\left[\sum_{t\in S_k}\frac{n}{l}A_{it}B_{tj}\right]^2 \quad (14)$$

$$= \frac{n^2}{l^2}\sum_{k=1}^{\binom{n}{l}}\pi(S_k)\left[\sum_{t\in S_k}A_{it}B_{tj}\right]^2. \quad (15)$$

Since all sets (S_k) have equal probability and each element appears in $\frac{l}{n}$ of these sets, when we expand $\left[\sum_{t\in S_k}A_{it}B_{tj}\right]^2$ we find that the coefficient for each $(A_{it}B_{tj})^2$ term is $\frac{l}{n}$. Further, to find the coefficients for the cross terms, we calculate the probability that two distinct elements appear in the same set. If we fix elements t and t' with $t \neq t'$ and define set S_k such that $t \in S_k$, then $\mathbf{Pr}[t' \in S_k] = \frac{l-1}{n-1}$. Thus,

$$\mathrm{E}[(CR)_{ij}^2] = \frac{n}{l}\sum_{t=1}^{n}(A_{it}B_{tj})^2 + \quad (16)$$

$$\frac{l-1}{l}\frac{n}{n-1}\sum_{t=1}^{n}\sum_{t'\neq t}^{n}A_{it}B_{tj}A_{it'}B_{t'j}$$

$$= \frac{n}{l}\sum_{t=1}^{n}(A_{it}B_{tj})^2 + \quad (17)$$

$$\frac{l-1}{l}\frac{n}{n-1}\Big((AB)_{ij}^2 - \sum_{t=1}^{n}(A_{it}B_{tj})^2\Big)$$

$$\leq \frac{n}{l}\sum_{t=1}^{n}(A_{it}B_{tj})^2 + \frac{l-1}{l}(AB)_{ij}^2, \quad (18)$$

where the inequality follows since $\|x\|_1 \leq \sqrt{n}\|x\|$ for $x \in \mathbb{R}^n$. We can now bound the variance as:

$$\mathbf{Var}[(CR)_{ij}] = \mathrm{E}[(CR)_{ij}^2] - \mathrm{E}[(CR)_{ij}]^2 \quad (19)$$

$$\leq \frac{n}{l}\sum_{t=1}^{n}(A_{it}B_{tj})^2 - \frac{1}{l}(AB)_{ij}^2. \quad (20)$$

Now, we can bound the expectation as:

$$\mathrm{E}\left[\|AB - CR\|_F^2\right] = \sum_{i=1}^{m}\sum_{j=1}^{p}\mathrm{E}[(AB - CR)_{ij}^2]$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{p}\mathbf{Var}[(CR)_{ij}]$$

$$\leq \frac{n}{l}\sum_{t=1}^{n}(\sum_{i}A_{it}^2)(\sum_{j}B_{tj}^2) - \frac{1}{l}\|AB\|_F^2$$

$$\leq \frac{n}{l}\sum_{t=1}^{n}\|A^{(t)}\|^2\|B_{(t)}\|^2.$$

By the concavity of $\sqrt{\cdot}$ and Jensen's inequality, $\mathrm{E}\left[\|AB - CR\|_F\right] \leq \sqrt{\mathrm{E}\left[\|AB - CR\|_F^2\right]}$. Thus,

$$\mathrm{E}\left[\|AB - CR\|_F\right] \leq \sqrt{\frac{n}{l}\sum_{t=1}^{n}\|A^{(t)}\|^2\|B_{(t)}\|^2}. \quad (21)$$

Lipschitz Bound

Consider the function Φ defined by $\Phi(S) = \|AB - CR\|_F$, where S is the set of l indices chosen uniformly at random without replacement from $\{1 \dots n\}$ to construct C and R. If we create a new set S' of indices by exchanging $i \in S$ for some $i' \notin S$, then we can construct the corresponding C' and R' from this new set of indices. We are interested in finding a Δ such that

$$|\Phi(S) - \Phi(S')| \le \Delta. \tag{22}$$

Using the triangle inequality, we see that

$$\Big| \|AB - CR\|_F - \|AB - C'R'\|_F \Big| \le \|CR - C'R'\|_F.$$

We next observe that the difference between CR and $C'R'$ depends only on indices i and i',[4] thus

$$\begin{aligned} \|CR - C'R'\|_F &= \frac{n}{l} \|A^{(i)} B_{(i)} - A^{(i')} B_{(i')}\|_F \qquad (23)\\ &\le \frac{n}{l} \left(\|A^{(i)}\| \|B_{(i)}\| + \|A^{(i')}\| \|B_{(i')}\| \right) \\ &\le \frac{2n}{l} \|A^{(t^*)}\| \|B_{(t^*)}\|, \qquad (24) \end{aligned}$$

where we use the triangle inequality and the identity $\|A^{(i)} B_{(i)}\|_F = \|A^{(i)}\| \|B_{(i)}\|$ to obtain (24).

Further, if $A = B^\top$, we can obtain a tighter bound. If $a = A^{(i)}$ and $a' = A^{(i')}$, we have:

$$\begin{aligned} \|aa^\top - a'a'^\top\|_F &= \sqrt{\mathrm{Tr}\left[(aa^\top - a'a'^\top)^\top (aa^\top - a'a'^T) \right]} \\ &= \sqrt{\|a\|^4 + \|a'\|^4 - 2(a^\top a')^2} \\ &\le \sqrt{\|a\|^4 + \|a'\|^4}. \qquad (25) \end{aligned}$$

Combining (23) with (25), the condition in (22) is satisfied for any Δ such that,

$$\Delta \ge \frac{\sqrt{2}n}{l} \|A^{(t^*)}\|^2. \tag{26}$$

Concentration Bound

Using the bound on the expectation and the Lipschitz bound just shown, by Theorem 1, for any $\epsilon > 0$ and $\delta > 0$, the following inequality holds:

$$\begin{aligned} \mathbf{Pr}\left[\|AB - CR\|_F \ge \sqrt{\frac{n}{l} \sum_{t=1}^{n} \|A^{(t)}\|^2 \|B_{(t)}\|^2} + \epsilon \right] \\ \le 2 \cdot \exp\left(\frac{-2\epsilon^2}{\alpha(l, n-l)\Delta^2} \right). \qquad (27) \end{aligned}$$

Setting δ to match the right-hand side and choosing $\epsilon = \Delta \sqrt{\frac{\log(2/\delta)\alpha(l,n-l)}{2}}$ yields the statement of Theorem 2.

4.3 Bound for Nyström Method

We now present a bound on the accuracy of the Nyström method when columns are chosen uniformly at random without replacement.[5]

Theorem 3. *Let $G \in \mathbb{R}^{n \times n}$ be an SPSD matrix. Assume that l columns of G are sampled uniformly at random without replacement, let $\tilde{G}_k$ be the rank-k Nyström approximation to G as described in (2), and let G_k be the best rank-k approximation to G. For $\epsilon > 0$, if $l \ge 64k/\epsilon^4$, then*

$$\begin{aligned} \mathrm{E}\left[\|G - \tilde{G}_k\|_F \right] \le \|G - G_k\|_F + \\ \epsilon \left[\left(\frac{n}{l} \sum_{i \in D(l)} G_{ii} \right) \sqrt{n \sum_{i=1}^{n} G_{ii}^2} \right]^{\frac{1}{2}}, \end{aligned}$$

where $\sum_{i \in D(l)} G_{ii}$ is the sum of the largest l diagonal entries of G. Further, if $\eta = \sqrt{\frac{\log(2/\delta)\alpha(l,n-l)}{l}}$, with $\alpha(l, n-l)$ defined in Theorem 1 and if $l \ge 64k/\epsilon^4$ then with probability at least $1 - \delta$,

$$\begin{aligned} \|G - \tilde{G}_k\|_F \le \|G - G_k\|_F + \\ \epsilon \left[\left(\frac{n}{l} \sum_{i \in D(l)} G_{ii} \right) \left(\sqrt{n \sum_{i=1}^{n} G_{ii}^2} + \eta \max\left(nG_{ii} \right) \right) \right]^{\frac{1}{2}}. \end{aligned}$$

Recall that for even moderately sized l and n, $\alpha(l, n-l) \approx l(1 - l/n)$ and thus $\eta \approx \sqrt{\log(2/\delta)(1 - l/n)}$. To prove this theorem, we use Corollary 1 (see proof for further details). If we instead use Corollary 2, we obtain the following weaker, yet more intuitive bound.[6]

Corollary 3. *Let $G \in \mathbb{R}^{n \times n}$ be an SPSD matrix. Assume that l columns of G are sampled uniformly at random without replacement, let $\tilde{G}_k$ be the rank-k Nyström approximation to G as described in (2), and let G_k be the best rank-k approximation to G. For $\epsilon > 0$, if $l \ge 64k/\epsilon^4$, then*

$$\mathrm{E}\left[\|G - \tilde{G}_k\|_F \right] \le \|G - G_k\|_F + \epsilon \cdot \max\Big(nG_{ii} \Big). \tag{28}$$

Further, if $\eta = \sqrt{\frac{\log(2/\delta)\alpha(l,n-l)}{l}}$, with $\alpha(l, n-l)$ defined in Theorem 1 and if $l \ge 64k(1+\eta)^2/\epsilon^4$ then with probability at least $1 - \delta$,

$$\|G - \tilde{G}_k\|_F \le \|G - G_k\|_F + \epsilon \cdot \max\Big(nG_{ii} \Big) \tag{29}$$

Proof. The theorem and its corollary follow from applying Lemma 2 to Lemma 1 and using Jensen's inequality. Note that when using these lemmas to prove Theorem 3, we use the fact that if $G = X^\top X$ then $\sum_{i \in D(l)} G_{ii} = \|X^{(1:l*)}\|_F^2$, where $X^{(1:l*)}$ are the largest l columns of X with respect to $\|\cdot\|$. We next state and prove these lemmas. □

[4] A similar argument is made in Drineas *et al.* (2006a) using the assumption of sampling independently and with replacement.

[5] Bounds for the l_2 norm can obtained using similar techniques. They are omitted due to space constraints.

[6] Corollary 3 can also be derived from Theorem 3 by noting that $\sum_{i \in D(l)} G_{ii} \le l \max\left(G_{ii}\right)$ and $\sum_{i=1}^{n} G_{ii}^2 \le n \max\left(G_{ii}^2\right)$.

Lemma 1. *Let $G \in \mathbb{R}^{n\times n}$ be an SPSD matrix and define $X \in \mathbb{R}^{m\times n}$ such that $G = X^\top X$. Further, let S, $|S| = l$ be any set of indices chosen without replacement from $\{1 \dots n\}$. Let $\tilde{G}_k$ be the rank-k Nyström approximation of G constructed from the columns of G corresponding to indices in S. Define $C_X \in \mathbb{R}^{m\times l}$ as the columns in X corresponding to the indices in S scaled by $\sqrt{n/l}$. Then*

$$\|G - \tilde{G}_k\|_F^2 \le \|G - G_k\|_F^2 + 4\sqrt{k}\|XX^\top XX^\top - C_X C_X^\top C_X C_X^\top\|_F.$$

Proof. The proof of this lemma in Drineas and Mahoney (2005) does not require any assumption on the distribution from which the columns are sampled and thus holds in the case of uniform sampling without replacement. Indeed, the proof relies on the ability to decompose $G = X^\top X$. To make this presentation self-contained, we next review the main steps of the proof of this lemma.

Let $X = U\Sigma V^\top$ and $C_X = \hat{U}\hat{\Sigma}\hat{V}^\top$ denote the the singular value decompositions of X and C_X. Further, let $\hat{U}_k$ denote the top k left singular vectors of C_X and define $E = \|XX^\top XX^\top - C_X C_X^\top C_X C_X^\top\|_F$. Then the following inequalities hold:

$$\begin{aligned}
\|G - \tilde{G}_k\|_F^2 &= \|X^\top X - X^\top \hat{U}_k \hat{U}_k^\top X\|_F^2 \\
&= \|X^\top X\|_F^2 - 2\|XX^\top \hat{U}_k\|_2^F + \|\hat{U}_k^\top XX^\top \hat{U}_k\|_F^2 \\
&\le \|X^\top X\|_F^2 - \sum_{t=1}^{k} \sigma_t^4(C_X) + 3\sqrt{k}\|E\|_F \\
&\le \|X^\top X\|_F^2 - \sum_{t=1}^{k} \sigma_t^2(X^\top X) + 4\sqrt{k}\|E\|_F \\
&= \|G - G_k\|_F^2 + 4\sqrt{k}\|E\|_F.
\end{aligned}$$

Refer to Drineas and Mahoney (2005) for further details. □

Lemma 2. *Suppose $X \in \mathbb{R}^{m\times n}$, $1 \le l \le n$ and construct C_X from X as described in Theorem 2. Let $E = XX^\top XX^\top - C_X C_X^\top C_X C_X^\top$ and define $X^{(1:l*)} \in \mathbb{R}^{m\times l}$ as the largest l columns of X with respect to $\|\cdot\|$. Then,*

$$\mathrm{E}\left[\|E\|_F\right] \le \frac{2n}{l}\|X^{(1:l*)}\|_F^2 \cdot \sqrt{\frac{n}{l}\sum_{t=1}^{n}\|X^{(t)}\|^4}. \tag{30}$$

Further, let $\delta \in (0,1)$ and $\eta = \sqrt{\frac{\log(2/\delta)\alpha(l,n-l)}{l}}$. Then with probability at least $1-\delta$,

$$\|E\|_F \le \frac{2n}{l}\|X^{(1:l*)}\|_F^2 \cdot \left(\sqrt{\frac{n}{l}\sum_{t=1}^{n}\|X^{(t)}\|^4} + \frac{\eta n}{\sqrt{l}}\|X^{(t^*)}\|^2\right). \tag{31}$$

Proof. We first expand E as follows:

$$\begin{aligned}
E &= XX^\top XX^\top - XX^\top C_X C_X^\top + XX^\top C_X C_X^\top - \\
&\quad C_X C_X^\top C_X C_X^\top \\
&= XX^\top(XX^\top - C_X C_X^\top) + (XX^\top - C_X C_X^\top) C_X C_X^\top.
\end{aligned}$$

Using the triangle inequality we have:

$$\begin{aligned}
\|E\|_F &\le \left(\|X\|_F^2 + \|C_X\|_F^2\right) \cdot \|XX^\top - C_X C_X^\top\|_F \\
&\le \frac{2n}{l}\|X^{(1:l*)}\|_F^2 \cdot \|XX^\top - C_X C_X^\top\|_F.
\end{aligned}$$

The lemma now follows by applying Corollary 1. □

4.4 Analysis of Bound

In the previous section we presented a new bound for the Nyström method, assuming columns are sampled uniformly without replacement. We now compare this bound with one presented in Drineas and Mahoney (2005), in which columns are sampled non-uniformly with replacement using a diagonal distribution. We compare the relative tightness of the bounds assuming that the diagonal entries of G are uniformly distributed, in which case Theorem 3 reduces to Corollary 3. This is the case for any normalized kernel matrix (K') constructed from an initial kernel matrix (K) as follows:

$$K'(x,y) = \frac{K(x,y)}{\sqrt{K(x,x)K(y,y)}}. \tag{32}$$

The diagonals of kernel matrices are also identical in the case of the RBF kernels, which Williams and Seeger (2000) suggests are particularly amenable to the Nyström method since their eigenvalues decay rapidly. When the diagonals are equal, the form of the bound in Drineas and Mahoney (2005) is identical to that of Corollary 3, and hence we can compare the bounds by measuring the value of the minimal allowable ϵ as a function of the fraction of columns used for approximation, i.e., the l/n ratio. Both bounds are tightest when the inequalities involving l, e.g., $l \ge 64k(1+\eta)^2/\epsilon^4$ for Corollary 3, are set to equalities, so we use these equalities to solve for the minimal allowable epsilon. In our analysis, we fix the confidence parameter $\delta = 0.1$ and set $k = .01 \times n$. The plots displayed in Figure 3 clearly show that the bound from Theorem 3 is tighter than that of Drineas and Mahoney (2005).

5 Conclusion

The Nyström method is used in a variety of large-scale learning applications, in particular in dimensionality reduction and image segmentation. This method is commonly used with uniform sampling without replacement, though non-uniform distributions have been used to theoretically analyze the Nyström method.

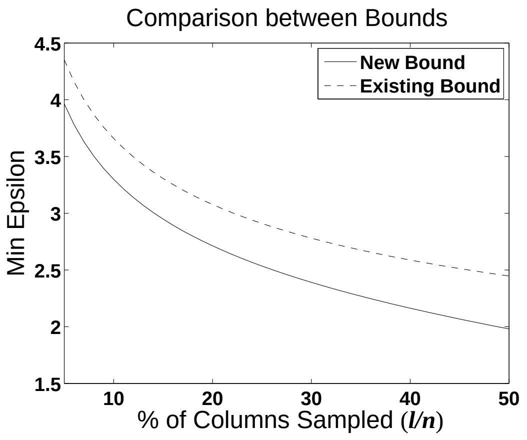


Figure 3: Comparison of the bound given by Drineas and Mahoney (2005) and our bound based on sampling without replacement.

In this work, we gave a series of clear empirical results supporting the use of uniform over non-uniform sampling, as uniform sampling tends to be superior in both speed and accuracy in several data sets. We then bridged the gap between theory and practical use of the Nyström method by providing performance bounds for the Nyström method when used with uniform sampling without replacement. Our analysis gives the first theoretical justification for the use of uniform sampling without replacement in this context. Our experiments and comparisons further demonstrate that the qualitative behavior of our bound matches empirical observations. Our bounds and theoretical analysis are also of independent interest for the analysis of other approximations in this setting.

Acknowledgments

This work was supported in part by the New York State Office of Science Technology and Academic Research (NYSTAR) and by a Google Research Grant.

References

A. Asuncion and D.J. Newman. UC Irvine machine learning repository, http://www.ics.uci.edu/∼mlearn/MLRepository.html, 2007.

Bernhard E. Boser, Isabelle Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *COLT*, 1992.

Corinna Cortes and Vladimir N. Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.

Corinna Cortes, Mehryar Mohri, Dmitry Pechyony, and Ashish Rastogi. Stability of transductive regression algorithms. In *ICML*, 2008.

Vin de Silva and Joshua B. Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. In *NIPS*, 2002.

Amit Deshpande, Luis Rademacher, Santosh Vempala, and Grant Wang. Matrix approximation and projective clustering via volume sampling. In *Symposium on Discrete Algorithms*, 2006.

Petros Drineas and Michael W. Mahoney. On the Nyström Method for Approximating a Gram Matrix for Improved Kernel-Based Learning. *Journal of Machine Learning Research*, 6:2153–2175, 2005.

Petros Drineas, Eleni Drinea, and Patrick S. Huggins. An experimental evaluation of a Monte-Carlo algorithm for SVD. In *Panhellenic Conference on Informatics*, 2001.

Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast Monte Carlo Algorithms for Matrices I: Approximating Matrix Multiplication. *SIAM J. Comput.*, 36(1):132–157, 2006.

Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix. *SIAM J. Comput.*, 36(1), 2006.

Petros Drineas. Personal communication, 2008.

Charless Fowlkes, Serge Belongie, Fan Chung, and Jitendra Malik. Spectral grouping using the Nyström method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2), 2004.

A. Gustafson, E. Snitkin, S. Parker, C. DeLisi, and S. Kasif. Towards the identification of essential genes using targeted genome sequencing and comparative analysis. *BMC:Genomics*, 7:265, 2006.

Yann LeCun and Corinna Cortes. The MNIST database of handwritten digits, http://yann.lecun.com/exdb/mnist/, 2009.

John C. Platt. Fast embedding of sparse similarity graphs. In *NIPS*, 2003.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

Terence Sim, Simon Baker, and Maan Bsat. The CMU pose, illumination, and expression database. In *Conference on Automatic Face and Gesture Recognition*, 2002.

Ameet Talwalkar, Sanjiv Kumar, and Henry Rowley. Large-scale manifold learning. In *CVPR*, 2008.

Christopher K. I. Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *NIPS*, 2000.

Kai Zhang, Ivor W. Tsang, and James T. Kwok. Improved Nyström low-rank approximation and error analysis. In *ICML*, 2008.