Low Rank Approximations for Kernel Matrices: Nystrom, ALS

Vibhuti Mahajan (12792) and Ayush Sekhari (12185)
Indian Institute of Technology Kanpur, India
Course Project : Learning with Kernels
Instructor : Prof. Harish Karnick

1 Introduction

Kernel trick is widely used to model non-linear data. A hard to miss example for the same is that of Ridge Regression: For a feature map $\phi : \mathcal{X} \to \mathcal{H}$ from the input space to the feature space modifies the conventional objective function to:

$$\min_{\omega} \lambda\|\omega\|^2 + \sum_{i=1}^{L}(\omega^T\phi(x_i) - y_i)^2 \tag{1}$$

where L is the size of the learning set. Solution for the above problem :

$$\omega = \frac{1}{2\lambda}\sum_{i=1}^{L}\alpha_i x_i \tag{2}$$

$$\alpha = 2\lambda(K + \lambda I)^{-}y \tag{3}$$

$y = \begin{bmatrix} y_1, ..., y_L \end{bmatrix}^T, \alpha = \begin{bmatrix} \alpha_1, ..., \alpha_L \end{bmatrix}^T, K_{L\times L}$ kernel matrix . Also,

$$y_{predicted} = \frac{1}{2\lambda}\sum_{i=1}^{L}\alpha_i K(x_i, x_test) \tag{4}$$

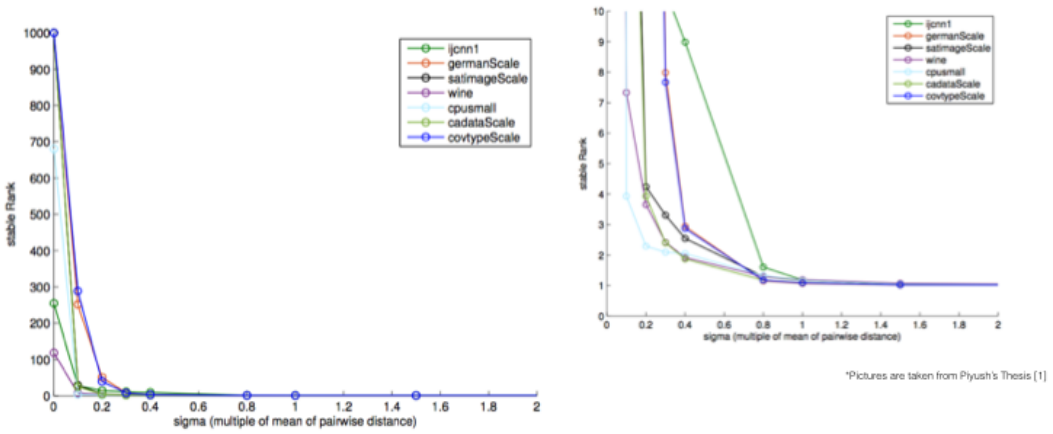
But the major challenge involved in using the Kernel Methods pertains to the high cost involved in computation and storage of the kernel matrix($\mathcal{O}(n^2d)$and $\mathcal{O}(n^2)$respectively; d is the dimension of input space). This motivates the usage of an approximate sketch of K, using an approximation of the form $K \approx UU^T$, where the rank of U is $\leq r(r < n)$.
Thus we can substitute the following in the Ridge Regression problem:

$$\lambda(K + \lambda I)^{-}y = I - U(\lambda I + U^TU)^{-}U^T \tag{5}$$

which can be computed in $\mathcal{O}(nr^2)$time.
Another motivation for using the low rank approximation comes from the effect of bandwidth parameter on the nature of the gaussian kernel Matrix: $K(x, y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2}$. The stable rank of a matrix is defined as $R_s = \frac{\|M\|_F^2}{\|M\|^2} = 1 + \frac{\sigma_2^2}{\sigma_1^2} + ... + \frac{\sigma_n^2}{\sigma_1^2}$. It gives an idea how quickly the spectrum decays relative to σ_1. The quick descent to stable rank 1 as σ is increased indicates, the largest eigen value starts dominating the other eigen values. For σ based on the mean of pairwise distance between data (σ_{mean}), the kernel matrix has a low rank nature as stable rank approaches 1. This further motivates the use of low rank approximation since the tuned σ value for ridge regression is close to σ_{mean}. Figure 1 shows stable rank vs σ for different datasets that strengthens our claim.

Formally stating, given the input $(x_1, ..., x_n) \in \mathcal{R}^d$ and an symmetric positive semi definite kernel function$K : \mathcal{R}^d \times \mathcal{R}^d \to \mathcal{R}$ such that $K_{ij} = K(x_i, x_j)$. The problem for r-rank approximation of K can thus be stated as:

$$\min_{U\in\mathcal{R}^{n\times r}} \sum_{ij}(K_{ij} - e_i^TUU^Te^j)^2 \tag{6}$$



where e_i is a vector whose value is 1 at i-th coordinate and 0 otherwise and U is a matrix of size $n \times r$ such that $K \approx UU^T$. Hence, the space complexity becomes $\mathcal{O}(nr)$, and time complexity for r-rank approximation becomes $\mathcal{O}(nr^2)$. In this report, we study approximation schemes using singular value decomposition(SVD), random feature base approximation, approximation using nystrom methods and alternating least squares.

2 Sngular Value Decomposition

The singular value decomposition of an $m \times n$ real or complex matrix M is a factorization of the form $M = U\Sigma V^T$, where U is an $m \times m$real or complex unitary matrix, Σ is an $m \times n$ rectangular diagonal matrix with non-negative real numbers on the diagonal, and V^T is an $n \times n$ real or complex unitary matrix. The diagonal entries Σ_i are known as the singular values of M. The m columns of U and the n columns of V are called the left-singular vectors and right-singular vectors of M, respectively.
Thus the problem of finding the best rank r can be posed as :

$$\min_{rank(\hat{M}) \leq r} \|M - \hat{M}\| : rank(\hat{M}) = r \tag{7}$$

where $\|.\|$ is the spectral/ frobenius norm and $r(1 \leq r \leq k = rank(M))$ is given
The complete SVD of M is given by:

$$M = U_{m\times m}\Sigma_{m\times n}V^T_{n\times n} \tag{8}$$

A best $r - rank$ approximation $\hat{M}_r$ is given by zeroing out the $k - r$ trailing singular values of M. That is:

$$\hat{M}_r = U\hat{\Sigma}_r V^T, \hat{\Sigma}_r = diag(\sigma_1, ..., \sigma_r, 0, ..., 0) \tag{9}$$

Now, frobenius and spectral norm are invariant by rotation of input and output spaces, i.e.:

$$\|U^T BV\|_F = \|B\|_F; \|U^T BV\|_2 = \|B\|_2 \tag{10}$$

for any matrix B, and orthogonal matrices U, V of appropriate sizes. Hence the optimal frobenius and spectral error is given by:

$$\|M - \hat{M}_r\|_F = \sqrt{\sigma^2_{r+1} + ... + \sigma^2_k} \tag{11}$$

$$\|M - \hat{M}_r\|_2 = \sigma_{r+1} \tag{12}$$

The approximation can thus be fabricated by choosing the top-r left and right singular vectors of the form $\hat{M}_r = U_{m\times r}\Sigma_{r\times r}V^T_{r\times n}$. Computing the top-r singualr vectors requires $\mathcal{O}(nmr)$ time and $\mathcal{O}(nm)$ space. This method is prohibitive in both space and time due to the requirement of computing the entire kernel matrix.

3 Random Feature Based Approximations

The seminal work of Recht and Rahimi proposes to map the input data to a randomized low-dimensional feature space and then applying existing fast linear methods. The set of random features consists of random fourier bases $\cos(\omega' x + b)$ where $\omega \in \mathcal{R}^d$ and $b \in \mathcal{R}$ are random variables. These mappings project data points on a randomly chosen line, and then pass the resulting scalar through a sinusoidal function.
This transformation is based on Bochener's theorem that states: *A continuous kernel $k(x, y) = k(x - y)$ on $\mathcal{R}^d$ is positive definite iff $k(\delta)$ is the Fourier transform of a non-negative measure.*
If shift-invariant kernel $k(\delta)$ is properly scaled, Bochner's theorem guarantees that its Fourier transform $p(\omega)$ is a proper probability distribution. Defining $\zeta_\omega(x) = e^{j\omega' x}$, we have

$$k(x - y) = \int_{\mathcal{R}^d} p(\omega) e^{j\omega'(x-y)} d\omega = E_\omega\big[\zeta_\omega(X)\zeta_\omega(y)^*\big] \tag{13}$$

so, $\zeta_\omega(X)\zeta_\omega(y)^*$ is an unbiased estimate of $k(x, y)$ when ω is drawn from p.
Setting $z_\omega(x) = \sqrt{2}cos(\omega' + b)$, we obtain the mapping

$$E\big[z_\omega(x) z_\omega(y)\big] = k(x, y) \tag{14}$$

where ω is drawn from $p(\omega)$ and b from $\big[0, 2\pi\big]$. Hence, for a fixed pair of x and y, $-\sqrt{2} \leq z\omega \leq \sqrt{2}$.
(Uniform Convergence of Fourier Features)*Let $\mathcal{M}$ be a compact subset of $\mathcal{R}^d$ with diameter $diam(\mathcal{M})$. Then, for the mapping z we have*

$$Pr\big[\sup_{x,y\in\mathcal{M}} |z(x)'z(y) - k(y, x)| \geq \in \big] \leq 2^8 \Big(\frac{\sigma_p(diam\mathcal{M})}{\in}\Big)^2 \exp\Big(-\frac{D \in^2}{4(d+2)}\Big) \tag{15}$$

where $\sigma_p^2 \equiv E_p\big[\omega'\omega\big]$ is the second moment of fourier transform of k.
Further, $\sup_{x,y\in\mathcal{M}} |z(x)'z(y) - k(y, x)| \leq \in$ *with any constant probability when* $D = \Omega\Big(\frac{d}{\in^2} \log \frac{\sigma_p diam(\mathcal{M})}{\in}\Big)$ This guarantees that $z(x)'z(y)$ is close to $k(x - y)$ for the centers of an $\in -$net over $\mathcal{M} \times \mathcal{M}$. This result is then extended to the entire space using the fact that the feature map is smooth with high probability. But the number of features required to get a good approximation of $K(x, y)$ with $\in$ accuracy is rather very large. By extending the bernstein inequalities, it is established that

$$Pr\Big(\|K - \hat{K}_D\| \leq \in \Big) \geq 1 - \delta \tag{16}$$

where $\hat{K}_D$ is an approximation of K formed by using D random Fourier features and $D \geq \frac{4n^2 + 2n\in}{\in^2} \log \frac{2n}{\delta}$. This defeats the purpose of approximation since the number of Fourier features to be computed are $\mathcal{O}(n^2)$. For rank r random Fourier features for r rank approximation requires $\mathcal{O}(nrd)$ (d is the dimension of input space) time and $\mathcal{O}(nr)$ space for storing parameters.

4 Nystrom Methods

The approximation $\hat{K}$ is obtained by choosing m rows/columns of K, and then setting $\hat{K} = K_{n\times m} K^-_{m\times m} K_{m\times n}$ where $K_{n\times m}$ is the block of original matrix K. In theory, kernel machines can be related to an expansion in a feature space of dimension N, such that

$$k(x, y) = \sum_{i=1}^{N} \lambda_i \phi_i(x) \phi_i(y) \tag{17}$$

where $N \leq \infty, \lambda_1 \geq \lambda_2 ... \geq 0$are eigen values and ϕ_2, ϕ_2 denotes the eigen functions of the operator, whose kernel is k, so that

$$\int k(x, y) \phi_i(x) p(x) dx = \lambda_i \phi_i(y) \tag{18}$$

where $p(x)$ s the probability density of input vector x. Given an iid sample $\{x_1, x_2, ..., x_q\}$ from $p(x)$, (4.2) can be approximated using nystrom expansion of the integral (i.e. mapping integral to a weighted summation):

$$\frac{1}{q}\sum_{k=1}^{q} k(y, x_k)\phi_i(x) \approx \lambda_i \phi_i(y) \tag{19}$$

The matrix eigen problem now becomes:

$$K^{(q)}U^{(q)} = U^{(q)}\Lambda^{(q)} \tag{20}$$

where $K^{(q)}$ is the $q \times q$ Gram matrix with elements $K_{ij}^{(q)}; i, j = 1, .., q, U^{(q)} \in \mathcal{R}^{q\times q}, \Lambda^{(q)} = diag(\lambda_1^{(q)} \geq \lambda_2^{(q)} \geq ... \geq \lambda_q^{(q)}) \geq 0$. Putting x_j for y in (19) and equating with (20), we get:

$$\lambda_i \approx \frac{\lambda_i^{(q)}}{q} \tag{21}$$

$$\phi_i(x_j) \approx \sqrt{q}U_{ji}^{(q)} \rightarrow \phi_i(y) \approx \frac{\sqrt{q}}{\lambda_i^{(q)}}\sum_{k=1}^{q} k(y, x_k)U_{ki}^{(q)} = \frac{\sqrt{q}}{\lambda_i^{(q)}} k_y.u_i^{(q)} \tag{22}$$

where $k_y = (k(x_i, y), ..., k(x_q, y))^T, u_i^{(q)}$is the i-th column of $U^{(q)}$. It follows that:

$$\hat{\lambda_i}^{(n)} = \frac{n}{m}\lambda_i^{(m)} \tag{23}$$

$$\hat{u_i}^{(n)} = \sqrt{\frac{m}{n}}\frac{1}{\lambda_i^{(m)}} K_{n\times m} u_i^{(m)} \tag{24}$$

where $u_i^{(m)}$is the i-th eigenvector of th $m \times m$eigen problem and $K_{n\times m}$ is the appropriate $n \times m$ submatrix of K. We consider the quality of approximation $\hat{K}$ at the m points used for eigen decomposition. Hence,

$$\hat{K}_m = K_{n\times m} K_{m\times m}^{-} K_{m\times n} \tag{25}$$

where $K_{m\times m}$is teh best r-rank approximation of K Work of Drineas and Mahoney states the convergence of $\hat{K}_r$ with high probability if m is atleast $\mathcal{O}(4c \in^{-2})$ where c is a constant depending upon the high probability parameter. i.e.

$$\|K - \hat{K}_r\|_{2/F} \leq \|K - K_r\|_{2/F} + \in \sum_{i=1}^{n} K_{ii}^2 \tag{26}$$

where $\hat{K}_r$ is the r-rank approximation using nystrom methods and K_r is the optimal r-rank approximation of K. The time complexity is $\mathcal{O}(m^3 + nmr), \mathcal{O}(m^3)$ for solving the $m \times m$ eigen problem and $\mathcal{O}(nmr)$ for a rank r approximation.
NOTE: Nystrom method can be further extended by using non-uniform sampling probabilities, k-means clustering to select the columns.

5 Matrix Approximations using Alternate Least Squares

5.1 Leveraged Element Low Rank Approximation

The problem of Low rank approximation with limited number of passes over the matrix was first introduced by Frieze et. al. There has since been a long track of work in this field. The current state of the art algorithm, knows as *leveraged element low rank approximation* was given by Bhojanapalli et. al . Their algorithm works in input sparsity time $\mathcal{O}(nnz(M)$ and gives guarantees in spectral Norm.
LELA works by non-uniformly sampling a few entries of the matrix and then doing a weighted alternate least squares to perform matrix completion. Random samples are chosen based on an weighted distribution using the column norms of the matrix and the absolute value of the entries of

Algorithm 1 Basic Algorithm with ALS

Input: data X, target rank r

Output: $\hat{U}_{n\times r}$

1: Sample K to create $R_\Omega(K)$

2: Initialize $\hat{U}^{(0)}$

3: **for** $t = 0, \ldots, T-1$ **do**

4: $\quad \hat{V}^{(t+1)} \leftarrow \text{argmin}_{V_{n\times r}} \sum_{(i,j)\in\Omega} w_{ij} \left(K_{ij} - e_i^T \hat{U}^{(t)} V^T e_j\right)^2 + \lambda ||V||_F^2$

5: $\quad \hat{U}^{(t+1)} \leftarrow \frac{1}{2}\left(\hat{V}^{(t+1)} + \hat{U}^{(t)}\right)$

the matrix. Let $M_{n\times d}$ be the matrix to be approximated. Every element of M is sampled using a non-uniform distribution given by :

$$q_{ij} = m(0.5\frac{\|M_i\|^2 + \|M^j\|^2}{(n+d)\|M\|_F^2} + 0.5\frac{|M|_{ij}}{\|M\|_{1,1}}) \tag{27}$$

where m is the desired sample size, M_i denotes a vector of elements of the i^{th} row of M, M^j denotes a vector of elements of the j^{th} column of M and $\|M\|_{1,1} = \sum_{i,j} |M|_{i,j}$. This type of distribution is carefully chosen so that the L1 term gives importance to the heavier element of the matrix, thus accounting for the possible high-rank nature of the matrix.

LELA uses the above sampled elements to perform a matrix completion step. Let $R_\Omega(M)$ be the sampled matrix with m sampled entries and Ω be the set of sampled entries, then LELA solves the following optimization problem over factors $U_{n\times r}$ and $V_{d\times r}$:

$$\min_{U,V} \sum_{\Omega} \frac{(M_{ij} - e_i^T U V^T e_j)}{q_{ij}} \tag{28}$$

The problem is inherently non-convex in U and V but becomes convex if one of U or V is fixed and thus the above equation is optimised using alternate least squares. For $M_{n\times d}$, LELA returns a rank r approximation in $\mathcal{O}(nnz(M) + mrr)$ time, where m is the number of randomly samples entries of M. The number of parameters required to store the approximation is $\mathcal{O}(n+d)r$.

5.2 Extending LELA

Given n input points $x_1, x_2, ..., x_n$ in input space $\mathcal{X} \subseteq \mathcal{R}^d$ and an SPSD kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathcal{R}$, the associated kernel matrix is an $n \times n$ matrix with $K_{ij} = K(x_i, x_j)$. The problem of rank r approximation of matrix K can be states as finding $U_{n\times r}$ such that

$$\min_{U \epsilon \mathcal{R}^{n\times r}} \sum_{i,j} (K_{i,j} - e_i^T U U^T e_j)^2) \tag{29}$$

where e_i is the $n \times 1$ column vector with 1 at the i^{th} coordinate and 0 at the other coordinates.

As LELA is a general matrix approximation algorithm it provides an approximation of the form UV^T. However for SPSD kernel matrix we need factorisation of the form UU^T. This issue is tackled by solving the optimization problem for U and V and then taking the average of U and V. It can be shown that this strategy leads to low errors.

In the basic algorithm, in step 1, $R_\Omega(K)$ is the sparse matrix created based on some sampling scheme which are discussed later. The solution is initialised as $\hat{U}^{(0)}$ following some initialisation strategies in step 2. Step 4 and 5 are iterative ALS optimization steps. Weights are used while forming the convex optimization problem ($w_{ij} = \frac{1}{p_{ij}}$), where p_{ij} is the probability with which the $(i,j)^{th}$ entry

is sampled. One advantage of taking a weighted objective function is that in expectation the error value is same as that of non weights objective function created by sampling all the entries. For ease of computation the optimization problem in **??** can be sub-divided into n independent least squares problems as:

$$\min \sum_{j=1}^{n} \sum_{i:(i,j)\epsilon\Omega} w_{ij}(K_{ij} - e_i^T \hat{U}^{(t)} V^j)^2 + \lambda \|V^j\|^2 \tag{30}$$

Each of these n problems can be solved independently for $V^j, j = 1, ..., n$. This lets us use a parallel solver. In the equation 30, we also used a regularisor on V. This is to ensure that while optimizing using least squares, the non- sampled terms do not blow up arbitrarily.

5.2.1 Time Complexity

Solving each of the n optimization problems in step 4 of the basic algorithm takes $\mathcal{O}(|\Omega_j|r^2 + r^3)$ time, where $|\Omega_j|$ is the number of entries sampled from the jth column of K. Thus, the overall sequential time complexity of running each iteration of ALS is $\mathcal{O}(|\Omega|r^2 + nr^3)$. The space complexity is $\mathcal{O}(rn)$. The algorithm can be made much faster after parallelisation.

5.2.2 Theoretical Analysis

The below given theoretical analysis is for the case with uniform sampling and ALS initialisation using Nystrom Approximations.

Definition .1. A matrix $M\epsilon\mathcal{R}^{n\times m}$ is incoherent with the parameter μ_r if:

$$\|u^{(i)}\| \leq \frac{\mu_r^2\sqrt{r}}{\sqrt{n}}, \forall i\epsilon[n], \|v^{(i)}\| \leq \frac{\mu_r^2\sqrt{r}}{\sqrt{m}}, \forall i\epsilon[m] \tag{31}$$

where $M = U\sum V^T$ is the SVD of M and $u^{(i)}$, $v^{(j)}$ denote the ith row of U and the jth row of V respectively.

It is trivial to see that the following bounds holds for μ_r

$$1 \leq \mu_r^2 \leq \frac{n}{r} \tag{32}$$

The above notion defines coherence of a matrix(with respect to canonical basis). we are interested in it because of the following result by candes et al

A0: The coherenences obey $max(\mu(U), \mu(V)) \leq \mu_0$ for some positive μ_0.

A1: The $n_1 \times n_2$ matrix $\sum_{1\leq k\leq r} u_k v_k^*$ has a maximum entry bounded by $mu_1\sqrt{\frac{r}{n_1 n_2}}$ in absolute value for some positive mu_1.

Theorem .2. *Let M be an $n_1 \times n_2$ matrix of rank r obeying* **A0** *and* **A1** *and put n = $max(n_1, n_2)$. Suppose we observe m entries of M with locations sampled uniformly at random. Then there exists constants C, c such that if*

$$m \geq Cmax(\mu_1^2, \mu_0^{1/2}\mu_1, \mu_0 n^{1/4})nr(\beta log(n))$$

for some $\beta > 2$, then the minimiser to 30 is unique and equal to M with probability at least $1 - cn^{-\beta}$.

We can conclude from the above theorem that small coherence means that we can learn a low rank approximation with lesser number of sample points. This incoherent matrices are easy to be recovered.

We also try to quantify the distance between subspaces using the idea of Principal Angles:

Definition .3. Given two matrices $\hat{U}, \hat{W} \epsilon \mathcal{R}^{n \times r}$, the principal angle based distance between the subspaces spanned by the columns of $\hat{U}$ and $\hat{W}$ is given by :

$$dist\hat{U}, \hat{(W)} = \|U_{\perp}^{T} W\| = \|W_{\perp}^{T} U\| \tag{33}$$

where U and W are orthonormal bases spanning the column space of $\hat{U}$ and $\hat{W}$ respectively and $U_{\perp}$ and $W_{\perp}$ denote the orthonormal basis spanning the spaces perpendicular to $\hat{U}$ and $\hat{W}$ respectively

We prove the correctness of the algorithm using the following induction steps :

1. **Base Case**: Show that U^0 is close to U^* and U^0 is incoherent.
2. **Inductive Hypothesis** : Assume that U^t is close to U^* and U^t is incoherent
3. **Inductive Step**: Show that U^{t+1} is close to U^* and U^{t+1} is incoherent.

INITIALIZATION STEPS
Our initialisation method is standard Nystom Approximation. We will construct our proof on the following result :

Theorem .4. *For rank r Nystrom Approximation, the distance between the optimal subspace, U^* and the subspace spanned by the columns of $\hat{K}, U_{\hat{K}}$ is bounded by:*

$$\sqrt{\frac{\sigma_n(K)}{\Sigma_1(\hat{K})}} \|U_{B'_{(n-r)}}\| \leq dist(U^*, U_{\hat{K}}) \leq \sqrt{\frac{\sigma_{r+1}(K)}{\sigma_r(\hat{K})}} \tag{34}$$

We use the following results from Yang et al with the above theorem:

Theorem .5. *let $\Delta = \frac{\sigma_r(K)}{\sigma_{r+1}(K)}$ be the ratio of consecutive eigenvalues of K. If*

$$\sigma_r(K) - \sigma_{r+1}(K) \geq 12 \frac{ln(2/\delta)}{\sqrt{r}} \tag{35}$$

then with probability $1 - \delta$

$$dist(U^*, U_{\hat{K}}) \leq \sqrt{\frac{3}{1 + 1\Delta}} \tag{36}$$

We use the following theorem by Yang et al. to state the incoherence of our initial assumption

Theorem .6. *Suppose U^* is incoherent with parameter μ and $U_{\hat{K}}$ is the orthonormal basis of the rank r Nystrom Approximation $\hat{K}$, then $\hat{K}$ is incoherent with parameter $\sqrt{\frac{\sigma_1(K)}{\sigma_r(\hat{K})}}\mu$. Further, if*

$$\sigma_r(K) - \sigma_{r+1}(K) \geq 12 \frac{ln(2/\delta)}{\sqrt{r}} \tag{37}$$

then with probability $1 - \delta$, $\hat{K}$ is incoherent with parameter $\sqrt{1.5\mathcal{K}}\mu$ where $\mathcal{K} = \frac{\sigma_1(K)}{\sigma_r(K)}$ is the condition number of best rank r approximation of K.

This directly leads to the following

$$\mu_r(\hat{K}) \leq \sqrt{\frac{\sigma_1(K)}{\sigma_r(K)}} \mu(K) \tag{38}$$

Thus proving the consistency of our starting point.

ITERATIVE STEPS
We use the following theorems from Jain et al which holds for uniform sampling case.

Algorithm	Space Complexity	Rank	Time Complexity
Nystrom	O(nr)	r	O(m^3+nmr)=O(nr^2)
SVD	O(nr)	r	O(rn^2)
ALS	O(nr)	r	O(nr^3 + |Ω|r^2) =O(nlognr^3)

Figure 1: Complexity comparison for the three algorithms

Theorem .7. *In the sampling step, let every entry of K be samples uniformly and independently with probability,*

$$p \geq C \frac{\frac{\sigma_1(K)}{\sigma_2(K)}^2 \mu^2 r^{2.5} log(n) log(\frac{r\|K\|_F}{\epsilon})}{n\delta_{2r}^2} \tag{39}$$

where $\delta_{2r} \leq \frac{\Sigma_r(K)}{12r\sigma_1(K)}$ *and* $C \geq 0$ *is a global constant. Then* $(t+1)^{th}$ *iterate* $\hat{V}^{t+1}$ *satisfies the following with high probability :*

$$dist(\hat{V}^{t+1}, U^*) \leq \frac{dist(\hat{U}^{(t)}, U^*)}{4} \tag{40}$$

Using the above theorem, we have

$$dist(\hat{U}^{t+1}, U^*) = \|\tfrac{1}{2}(U_\perp^* \hat{V}^{(t+1)} + U_\perp^* \hat{U}^{(t)})\|$$

$$\leq \tfrac{1}{2}\|U_\perp^* \hat{V}^{(t+1)}\| + \|U_\perp^* \hat{U}^{(t)})\|$$

$$= \tfrac{1}{2}(dist((\hat{V})^{t+1}, U^*) + dist(\hat{U}^{(t)}, U^*))$$
$$\leq \tfrac{5}{8} dist(\hat{U}^{(t)}, U^*)$$

The above result shows that with each iteration the distance between the real solution and our estimate decreases with a factor of at-least $\frac{5}{8}$. Finally, we show the incoherence of $\hat{U}^{(t+1)}$. The incoherence of $\hat{V}^{(t+1)}$ follows directly from the following theorem by Jain et. al

Theorem .8. *Let* $\hat{U}^t$ *be* μ_1 *incoherent. Then with probability at least* $1 - \frac{1}{n^3}$*, iterate* $\hat{V}^{t+1}$ *is also* μ_1 *coherent*

The above theorem could easilty be extended to prove that $\hat{U}^{(t+1)}$ is also μ_1 coherent thus completing the proof of correctness of our algorithm.

6 Conclusion

Kernel approximation scales up kernel machines in an effective way. The literature provides several techniques for kernel approximation, out of which we study SVD, Nystrom approximations, Feature

based approximations and an extension of Alternating Least Squares. Complexity analysis is given in the table below.

The space complexity for SVD, Nystrom approximation and ALS is $\mathcal{O}(nr)$. Empirical results show that the performance of SVD is optimal but the nature of variation of errors with the number of parameters is similar for ALS and SVD. Nystrom methods is the fastest to solve and ALS is the slowest owing to the overheads in sampling and running the iterative least square.

References

1. Piyush Bhardwaj, M.Tech Thesis (2015) Efficient Low Rank Approximations via Alternate Least Squares for Scalable Kernel Learning

2. Ali Rahimi and benjamin Recht, NIPS (2007) Random Features for large Scale Kernel machines

3. CKI Williams, Matthias Seeger, Using the Nystrom Method to speed up Kernel machines

4. Petros Drineas, Michael W. Mahoney, On the Nystrom Methods for Approximating a Gram matrix for Improved Kernel-Based learning

5. Emmanuel J Candes, Benjamin Recht, Exact Matrix Completion via Convex optimizatoion

6. Mehrdad Madhavi, Tianbao Yang and Rong Jin, An improved bound for the nystrom method using large eigen-gap

7. Prateek Jain, Praneeth Netrapalli, Sujay Sanghavi, Low Rank Matrix Completion using alternating Minimization.

8. Alan Frieze, Ravi Kannan, Santosh Vempala, Fast Monte-Carlo algorithms for finding low-rank approximations

9. Srinadh Bhojanapalli, Prateek Jain, Sujay Sanghavi, Tighter Low-Rank approximation via sampling the leveraged element.