Report: Improved Nyström Low Rank Approximation and Error Analysis [ZTK08]

Sofiya Burova, Martin Gjorgjevski

March 2022

1 Introduction

Given an n by n positive semi-definite matrix K, the classical Nyström Method for low rank approximation consists of selecting m columns $K_{i_1}, K_{i_2}, ..., K_{i_m}$ (or rows by symmetry) of the matrix K, and approximating K by $EW^{-1}E^{'}$ where E is the n by m matrix whose $j-th$ column is K_{i_j} and W is the m by m submatrix of K given by $[W]_{l,s} = [K]_{i_l,i_s}$. When m is small or moderate relative to n, the method, if succsesful, seems very promising, as it significantly reduces computational complexity- it only requires to deal with nm parametars for E and the inversion of the matrix W is not too costly as it is only m by m matrix. Ideally the columns should be chosen so that the matrix W has large eigenvalues but in practice one can not know which choice of W is the best without explicitly computing all possible choices. This is clearly not practical as there are $\binom{n}{k}$ choices for W and computing them in order to compare them defeats the purpose of low-rank computationally efficient approximation of the matrix K. Hence m columns are sampled uniformly at random and averaging over several samples is sometimes used in order to reduce the variance.

2 Historical motivation and heuristics

Originally, the Nyström method was designed for numerical solutions of integral equations [WS00a].

In a very simple form, Mercer's theorem states that if $K : [a, b] \to \mathbb{R}$ is a positive semi definite kernel then the operator defined on $L^2([a, b], pdx)$

$$T_K(f)(x) = \int K(x, y)f(y)p(y)dy$$

has orthogonal eigenfunctions e_i with nonnegative eigenvalues λ_i, and in particular there is a decomposition of K as

$$K(x, y) = \sum_{i=1}^{\infty} \lambda_i e_i(x)e_i(y)$$

The Nyström method refers to the following probabilistic approximation scheme

$$\int K(x, y)e_i(y)p(y)dy \sim \frac{1}{q}\sum_{k=0}^{q} K(x, X_k)e_i(X_k)$$

where X_k are i.i.d. drawn according to p. The strong law of large number states that this approximation is asymptotically correct. By plugging in X_j in the equation above, one obtains a q by q matrix K^q such that the spectral decomposition of K^q is expected to aid in finding the solution to the integral equation problem. When the eigenvalues of this operator decay rapidly this approximation can work really well [WS00a]. For different sample sizes q, one gets different approximations.

Inspired by this heuristic, one may try in general to approximate n by n positive semi definite matrix by sampling m columns and computing the expression $EW^{-1}E^{'}$.

3 Error Analysis of a clustered model

Prior to the article [ZTK08] the analysis of the Nyström method was limited in the literature. The main contribution of the paper is analyzing the error between the original matrix K and its low rank Nyström approximation $EW^{-1}E^{'}$ in Frobenius norm, $\mathcal{E} = ||K - EW^{-1}E^{'}||_F$. The main assumption is that the matrix K is derived from a kernel which acts on vector data, and that this kernel satisfies a bound of the form

$$(k(a,b) - k(c,d))^2 \leq C_{\mathcal{X}}^k(||a-c||^2 + ||b-d||^2)$$

where $C_{\mathcal{X}}^k$ is a constant that depends on the data and the kernel k. In many commonly used kernels this constant does not depend on the data.

Under this assumption, by assuming a clustered model in which there are m clusters, each cluster consisting of T points and a selected landmark point as a representative for each cluster the following error bound is established:

$$\mathcal{E} \leq 4T\sqrt{mC_{\mathcal{X}}^k eT} + mC_{\mathcal{X}}^k Te\|W^{-1}\|_F$$

where e is the quantization error obtained as the total distance between each data point and it's associated landmark point in its respective cluster.

The main idea of the proof is to decompose the matrix K into smaller blocks and to bound the Frobenius error norm on each such block, then the theorem follows from a triangle inequality.

4 Improvement proposal and experiments

Inspired by this result, the authors propose running the k-means algorithm on the data with $k = m$. This algorithm aims at minimizing a total quantization error: initially a random set of k points is chosen and the initial clusters are formed by Voronoi tessellation, each point is assigned to the closest landmark (initial centroid). Then the following steps are iterated: for each cluster one computes the mean of points inside it and after that another Voronoi tessellation with the updated centroids determines the next clusters. This algorithm is designed to minimize the quantization error that appears in the main result of [ZTK08]. The authors propose to use the obtained centroids as landmark points. This is the novelty of the paper. Interestingly, while the classical low rank Nyström approximation guarantees exact recoveries of entries $K(z_i, z_j)$ whenever z_i and z_j are landmark points, this property is lost in the new method.

The authors report on a wide range of numerical experiments on real world data that demonstrate that in terms of Frobenius norm error, the Improved Nystrom method outperforms the classical Nystrom method as well as other low rank estimation algorithms such as Incomplete Cholesky Decomposition.

References

[WS00a] Christopher Williams and Matthias Seeger. "The Effect of the Input Density Distribution on Kernel-based Classifiers". In: *Proceedings of the 17th International Conference on Machine Learning*. Morgan Kaufmann, 2000, pp. 1159–1166.

[WS00b] Christopher Williams and Matthias Seeger. "Using the Nyström Method to Speed Up Kernel Machines". In: *Advances in Neural Information Processing Systems*. Ed. by T. Leen, T. Dietterich, and V. Tresp. Vol. 13. MIT Press, 2000. URL: https://proceedings.neurips.cc/paper/2000/file/19de10adbaa1b2ee13f77f679fa1483a-Paper.pdf.

[ZTK08] Kai Zhang, Ivor W. Tsang, and James T. Kwok. "Improved Nyström Low-Rank Approximation and Error Analysis". In: *Proceedings of the 25th International Conference on Machine Learning*. ICML '08. Helsinki, Finland: Association for Computing Machinery, 2008, pp. 1232–1239. ISBN: 9781605582054. DOI: 10.1145/1390156.1390311. URL: https://doi.org/10.1145/1390156.1390311.