Improved Nyström Low Rank Approximation and Error Analysis

Sofiya Burova, Martin Gjorgjevski

ENS Lyon
M2 Advanced Mathematics

March 2022

Introduction

- Nyström method for low rank approximation
- Historical motivation and heuristics
- Error analysis of a clustered model
- New k-means based method
- Empirical test of performance

Nyström Method, what is that?

Let K be a kernel on a sample set $\mathcal{X} = \{x_i, i \in [n]\}$. Then the Nyström method applied to K using a randomly chosen landmark set $\mathcal{Z} = \{z_i, i \in [m]\}$ approximates the full kernel matrix by

Williams and Seeger, 2001

$$K \sim EW^{-1}E'$$

where E is an $n \times m$ matrix given by $E_{ij} = k(x_i, z_j)$ and W is an $m \times m$ matrix given by $W_{ij} = k(z_i, z_j)$.

In practice this method promises a lot. The most popular sampling scheme is random sampling where a random subset of m points is chosen.

Numerical solutions to integral equations

- Mercer's theorem states that $K(x,y) = \sum_{k=1}^{\infty} \lambda_k \phi_k(x)\phi_k(y)$ with ϕ_k orthogonal, λ_k nonnegative and decreasing
- if x_j are sampled i.i.d. according to $p(x)$, $j = 1, 2, ..., q$, then

$$\int K(x,y)p(y)\phi_k(y)dy \approx \frac{1}{q}\sum_{j=1}^{q} K(x,x_j)\phi_k(x)$$

  by the law of large numbers.
- Solve for the eigenvalues of $K^{(q)} = [K(x_i,x_j)]_{1\le i,j\le q}$
- $\frac{\lambda_k^{(q)}}{q} \approx \lambda_k$, $\lambda_k^{(q)}$ being the k-th largest eigenvalue of $K^{(q)}$[1]
- In addition, if $K(x,y) = K(x-y)$ we have

$$\frac{1}{q}\sum_{k=1}^{q} \lambda_k^{(q)} = \sum_{k=1}^{\infty} \lambda_k$$

[1] in fact we have convergence as $q \to \infty$

Nyström Approximation

This approximation is achieved by carrying out an eigendecomposition on a smaller system of size $m < n$:

- The method can work well for matrices with rapidly decaying spectra
- In practice it is not possible to know which subset of landmark points will have the highest eigenvalue in the spectral decomposition
- Error analysis for this method has been limited in the literature

Our objective is to bound the approximation error given by

$$\mathcal{E} = \|K - EW^{-1}E'\|_F$$

where $\|\cdot\|_F$ denotes the Frobenious norm.

Main result, assumption

Assumption A

Assume that for all a, b, c, d,

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

where $C_{\mathcal{X}}^k$ is a positive constant depending on k and the sample set $\mathcal{X}$.

Remark. This assumption is valid on many commonly used kernels:

- Gaussian kernel, $C_{\mathcal{X}}^k = \frac{1}{2\sigma^2}$
- Laplacian kernel, $C_{\mathcal{X}}^k = \frac{1}{\sigma^2}$
- Inverse distance kernel, $C_{\mathcal{X}}^k = \frac{1}{\sigma^2\epsilon^4}$

Main result, statement

Partition $\mathcal{X}$ into m disjoint clusters $S_k, k \in [m]$.

Set $c(i) = \text{argmin}_{j \in [m]} \|x_i - z_j\|$.

Theorem

If the kernel satisfies Assumption A, the error of the Nyström approximation is bounded by

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

where $T = \max_k |S_k|$ and $e = \sum_{i=1}^n \|x_i - z_{c(i)}\|^2$ is the total quantization error of coding each sample $x_i \in \mathcal{X}$ with the closest landmark point $z_j \in \mathcal{Z}$.

Main result, sketch of proof

The main idea is to decompose the kernel matrix into blocks of size $m \times m$ and bound each *partial approximation error*. This is done via the following sampling process:

- at each time t pick a sample from each cluster and denote the resulting set by $\mathcal{X}_{\mathcal{I}_t}$;
- $\mathcal{X} = \cup_{t \in [T]} \mathcal{X}_{\mathcal{I}_t}$;
- $K_{\mathcal{I}_i, \mathcal{I}_j}$ and $E_{\mathcal{I}_i, \mathcal{Z}}$ are the $m \times m$ similarity matrices defined resp. on $(\mathcal{X}_{\mathcal{I}_i}, \mathcal{X}_{\mathcal{I}_j})$ and $(\mathcal{X}_{\mathcal{I}_i}, \mathcal{Z})$.
- $\mathcal{E}_{\mathcal{I}_i, \mathcal{I}_j} = \| K_{\mathcal{I}_i, \mathcal{I}_j} - E_{\mathcal{I}_i, \mathcal{Z}} W^{-1} E'_{\mathcal{I}_j, \mathcal{Z}} \|_F$ is the partial approximation error.

Main result, sketch of proof

Observe that the total error $\mathcal{E} \leq \sum_{i,j=1}^{T} \mathcal{E}_{\mathcal{I}_i,\mathcal{I}_j}$. Hence, the following Lemma concludes the proof:

Lemma

If the kernel satisfies Assumption A, the partial approximation error is bounded by

$$\mathcal{E}_{\mathcal{I}_i,\mathcal{I}_j} \leq \sqrt{2mC_{\mathcal{X}}^{k}(e_{\mathcal{I}_i} + e_{\mathcal{I}_j})} + \sqrt{mC_{\mathcal{X}}^{k} e_{\mathcal{I}_i}} + \sqrt{mC_{\mathcal{X}}^{k} e_{\mathcal{I}_j}} + m\sqrt{mC_{\mathcal{X}}^{k} e_{\mathcal{I}_i}}\sqrt{e_{\mathcal{I}_i} e_{\mathcal{I}_j}}\|W^{-1}\|_F$$

where $e_{\mathcal{I}_i}$ is the quantization error induced by coding each sample in $\mathcal{X}_{\mathcal{I}_i}$ i.e.

$$e_{\mathcal{I}_i} = \sum_{x_i \in \mathcal{X}_{\mathcal{I}_i}} \|x_i - z_{c(i)}\|^2.$$

Conclusion

For a number of commonly used kernels,in order to minimize the total error of approximation, it suffices to minimize the quantization error

$$e = \sum_{i=1}^{n} \|x_i - z_{c(i)}\|^2.$$

Naturally, we propose the centers obtained from the k-means as the landmark points.

Example

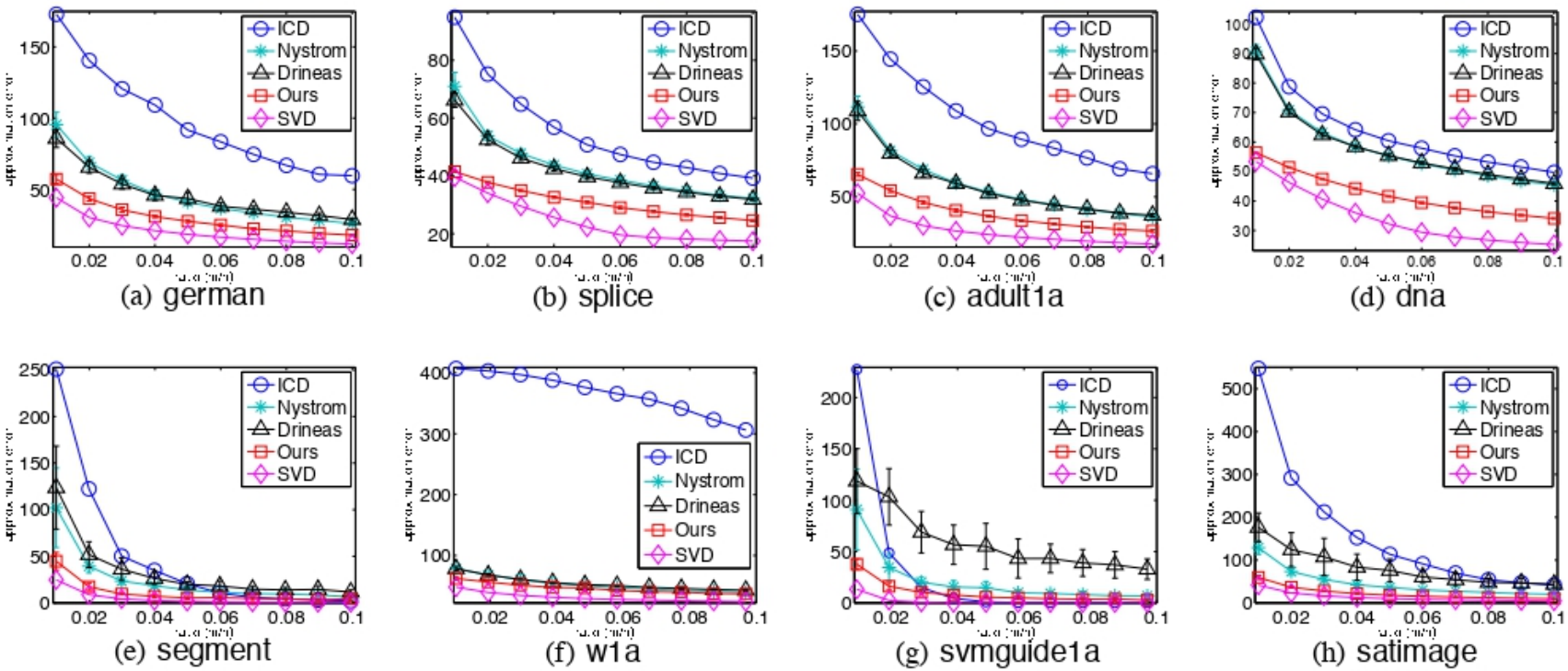

Figure 1. Approximation errors (in terms of the Frobenius norm) on the kernel matrix by different low-rank approximation schemes.