Exercise VII Numerical Linear Algebra

Large-scale matrix approximation

November 2023

In this session, we will approximate large-scale matrices by $rank - r$ CUR decomposition. To be precise, for a matrix $A \in \mathbb{K}^{m\times n}$, we approximate A by

$$A \approx CUR,$$

where

- $C = A_{:J} \in \mathbb{K}^{m\times r}$ contains the columns of A at the indices in $J \subseteq [n] := \{1, \dots, n\}$ with full column rank,
- $R = A_{I:} \in \mathbb{K}^{r\times n}$ contains the rows of A at the indices in $I \subseteq [m] := \{1, \dots, m\}$ with full row rank,
- $U \in \mathbb{K}^{r\times r}$ depends on the choice of C and R. Once C and R are determined, U is unique.

There are two main approaches:

- orthogonal projection: $A \approx C(C^\dagger A R^\dagger)R = (CC^\dagger)A(R^\dagger R)$;
- cross approximation: let $S_I^\top \in \mathbb{K}^{r\times m}$ and $S_J \in \mathbb{K}^{n\times r}$ select the rows in I and the columns in J respectively,

$$A \approx C A_{IJ}^{-1} R = (C\,(S_I^\top C)^{-1} S_I^\top) A (S_J\,(R S_J)^{-1} R).$$

In the following text, we will focus on the second approach, where the main task is how to choose proper index subsets I and J such that the approximation error can be guaranteed.

1 Adaptive Cross Approximation

The outline is as follows,

1 Initialise $R = A$.

2 For $s = 1, 2, \dots, r$:

   a Determine a good pivot (i, j) of R (e.g., partial pivoting, rook pivoting and full pivoting).

b Set $I \leftarrow I \cup \{i\}$ and $J \leftarrow J \cup \{j\}$.

c Formally set $R = R - R_{:,j} R_{i,j}^{-1} R_{i,:}$ (i.e., subtract the interpolatory rank-1 tensor or cross).

3 The outputs are a set of r column indices J and r row indices I.

The details please see the slides P32–P37.

2 Derandomized Row/Column-subset Selection

For simplification, we just state the results and algorithms in terms of row-subset section. The column-subset selection is the same by dealing with $A^\top$. For $I \subseteq [m]$ with cardinality r, span(I) denote the linear span of the rows of A with indices in I and $\pi_I(A) \in \mathbb{K}^{m\times n}$ denote the matrix obtained by projecting each row onto span(I). Thus $A - \pi_I(A) \in \mathbb{K}^{m\times n}$ is the matrix obtained by orthogonal projection of each row of A to span(I).

Let's consider each row selection as a random variable X with row indices $[m]$ as sample space and $X(i) = i$ for $i \in [m]$. Then one possible rank-r row selection can be denoted as an r-tuple $(X_1, X_2, \dots, X_r)$. Denoting $I = \{i_1, i_2, \dots, i_r\}$, the r-volume probability density function V_r is given by

$$V_r := \mathbb{P}(X_1 = i_1, \dots, X_r = i_r) = \begin{cases} \frac{\det(A_{I:} A_{I:}^\top)}{r! \sum\limits_{S\subseteq[m]:|S|=r} \det(A_{S:} A_{S:}^\top)} & \text{if } i_1, i_2, \dots, i_r \text{ are distinct;} \\ 0 & \textit{otherwise.} \end{cases}$$

The second approach is inspired by the following Theorem.

Theorem 1. [1, Theorem 1.3] *Given* $A \in \mathbb{K}^{m\times n}$,

$$\mathbb{E}_{V_r}\left[\|A - \pi_{\{X_1,\dots,X_r\}}(A)\|_F^2\right] \le (r+1) \sum_{k=r+1}^{m} \sigma_k^2,$$

where $\|\cdot\|_F$ *denotes the Frobenius norm, and* σ_k *is the* k*-th singular value of* A *(in descending order).*

Due to the fact that expectation is a convex combination, Theorem 1 shows there exists at least one choice $I \subseteq [m]$ such that the error bound holds. The basic idea is that, for $t = 1, \dots, r$, we iteratively choose

$$\arg\min_{i_t\in[m]} \mathbb{E}_{V_r}\left[\|A - \pi_{\{X_1,\dots,X_r\}}(A)\|_F^2 | X_1 = i_1, \dots, X_{t-1} = i_{t-1}\right]. \tag{1}$$

Given the following lemma (you don't need to prove it),

Lemma 2. [2, Lemma 11&12] *For* $A \in \mathbb{K}^{m\times n}$, *we have*

$$\sum_{S\subseteq[m]:|S|=r} \det(A_S A_S^\top) = \sum_{1\le i_1<\dots<i_r\le m} \sigma_{i_1}^2 \cdots \sigma_{i_r}^2 = |c_{m-r}(AA^\top)| = |c_{n-r}(A^\top A)|,$$

where $c_{m-r}(AA^\top)$ *is the coefficient of the characteristic polynomial, i.e.,*

$$\det(x\,\mathbb{I}_{m\times m} - AA^\top) = \sum_{r=0}^{m} c_{m-r}(AA^\top)x^{m-r} = \prod_{i=1}^{m}(x-\sigma_i^2).$$

Suppose $S, T \subseteq [m]$ *satisfying* $S \cap T = \emptyset$ *and* $B := A - \pi_S(A)$*,*

$$\det(A_{S\cup T}\, A_{S\cup T}^\top) = \det(A_S\, A_S^\top)\det(B_T B_T^\top).$$

Let $S = \{i_1, i_2, \ldots, i_{t-1}\}$ **and** $B = A - \pi_S(A)$**. Please use Lemma 2 to derive the following computable form of the conditional expectation in** (1)**,**

$$\mathbb{E}_{V_r}\left[\|A - \pi_{\{X_1,\ldots,X_r\}}(A)\|_F^2 | X_1 = i_1, \ldots, X_{t-1} = i_{t-1}\right] = \frac{(r-t+2)\,|c_{m-r+t-2}(BB^\top)|}{|c_{m-r+t-1}(BB^\top)|}. \quad (2)$$

Use (1) **and** (2) **to design an algorithm to obtain the row selection** $I = i_1, \ldots, i_r$**.**

3 Application: Kernel Interpolation

The kernel interpolation has been widely applied in uncertainty quantification and machine learning.

Definition 3 (reproducing kernel Hilbert space)**.** *The Hilbert space* $H(K)$ *with inner product* $< \cdot, \cdot >_H$ *is a reproducing kernel Hilbert space (RKHS) with kernel* $K : [0,1]^d \times [0,1]^d \to \mathbb{R}$ *if:*

- $K(\cdot, \boldsymbol{x}) \in H$ *for all* $\boldsymbol{x} \in [0,1]^d$*,*
- $f(\boldsymbol{x}) = \langle f, K(\cdot, \boldsymbol{x})\rangle_H$ *for all* $\boldsymbol{x} \in [0,1]^d$ *and all* $f \in H$*.*

In addition, reproducing kernel K *has the following properties:*

- (*symmetry*) $K(\boldsymbol{x}, \boldsymbol{y}) = K(\boldsymbol{y}, \boldsymbol{x})$ *for all* $\boldsymbol{x}, \boldsymbol{y} \in [0,1]^d$*,*
- (*positive semidefiniteness*) $\sum_{i=1}^N \sum_{k=1}^N a_i a_k K(\boldsymbol{x}_i, \boldsymbol{x}_k) \geq 0$ *for all* $N \geq 1$*, all* $a_i \in \mathbb{R}$ *and all* $\boldsymbol{x}_i \in [0,1]^d$*.*

For any given sample set $X_n = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{n-1}\} \subset [0,1]^d$ of n distinct points, where both n and d can be very large, the kernel interpolation of $f \in H(K)$ is given by

$$A_n(f)(\boldsymbol{y}) := \sum_{k=1}^{n} a_k K(\boldsymbol{x}_k, \boldsymbol{y}), \quad \boldsymbol{y} \in [0,1]^d,$$

which interpolates f at sample points $\boldsymbol{x}_k$, $k = 1, \ldots, n$, i.e.,

$$A_n(f)(\boldsymbol{x}_k) = f(\boldsymbol{x}_k), \quad \text{for all } k = 1, \ldots, n.$$

The coefficients a_k, $k = 0, \ldots, n-1$ are obtained by solving the resulting linear system,

$$f(\boldsymbol{x}_\ell) = \sum_{k=0}^{n-1} a_k\, K(\boldsymbol{x}_k, \boldsymbol{x}_\ell) \quad \text{for all } \ell = 0, \ldots, n-1.$$

We can write the above linear system as

$$\boldsymbol{f}_{X_n} = \mathcal{K}\,\boldsymbol{a}, \quad \text{with} \quad \mathcal{K} := \left[K(\boldsymbol{x}_k, \boldsymbol{x}_\ell)\right]_{\ell,\, k=0,\dots,n-1}, \quad \boldsymbol{a} := \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n-1} \end{pmatrix}, \quad \boldsymbol{f}_{X_n} := \begin{pmatrix} f(\boldsymbol{x}_0) \\ f(\boldsymbol{x}_1) \\ \vdots \\ f(\boldsymbol{x}_{n-1}) \end{pmatrix}.$$

If $\mathcal{K}$ has full rank, then the inverse $\mathcal{K}^{-1}$ exists and the solution $\boldsymbol{a}$ is unique, i.e., $\boldsymbol{a} = \mathcal{K}^{-1}\boldsymbol{f}_{X_n}$.

One example of RKHS is the weighted Korobov space with reproducing kernel characterised by smoothness parameter α ($\alpha > 1$ is required for continuity of functions in the space) and positive weights $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_d)$,

$$K(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\boldsymbol{h} \in \mathbb{Z}^d} \frac{\mathrm{e}^{2\pi\mathrm{i}\boldsymbol{h}\cdot(\boldsymbol{x}-\boldsymbol{y})}}{r_{d,\alpha,\boldsymbol{\gamma}}(\boldsymbol{h})}, \quad \text{with} \quad r_{d,\alpha,\boldsymbol{\gamma}}(\boldsymbol{h}) := \prod_{\substack{j=1 \\ h_j \neq 0}}^{d} \frac{|h_j|^\alpha}{\gamma_j}.$$

For integer $q > 1$, we have (see [4, (24.8.3)])

$$\text{periodic Bernoulli polynomials} \;\; B_q(y) = \frac{q!}{-(2\pi\mathrm{i})^q} \sum_{h \in \mathbb{Z}\setminus\{0\}} \frac{\mathrm{e}^{2\pi\mathrm{i}hy}}{h^q} \quad \text{for} \quad y \in [0,1],$$

Show that for even $\alpha > 1$, the kernel can be expressed as follows,

$$K(\boldsymbol{x}, \boldsymbol{y}) = \prod_{j=1}^{d} \left[1 + (-1)^{\alpha/2+1} \frac{(2\pi)^\alpha}{(\alpha)!} \gamma_j \, B_\alpha\left(\{x_j - y_j\}\right)\right], \quad \text{for even } \alpha.$$

where the braces denote taking the fractional part of the input.

Now we can play with the kernel matrix $\mathcal{K} := \left[K(\boldsymbol{x}_k, \boldsymbol{x}_\ell)\right]_{\ell,\, k=0,\dots,n-1}$:

- **α can be set as 2, 4 or 8; (What happens when α becomes large?)**
- **weight parameters $(\gamma_j)_{j\in[d]}$ can be set as $\gamma_j = \frac{0.9^{j-1}}{\pi^\alpha}$;**
- **randomly generate n distinct sample points $X_n = \{\boldsymbol{x}_0, \dots, \boldsymbol{x}_{n-1}\} \subset [0,1]^d$, where n can be 2^{10}, 2^{15} or 2^{20}, and d can be 10, 50 or 100;**
- **compute the rank-r CUR decomposition $\widetilde{\mathcal{K}}_r$ using the random sample set X_n by both *adaptive cross approximation* and *derandomized row/column-subset selection* with various r, and then compare the time cost and the error in Frobenius norm.**

References

[1] Deshpande, A., Rademacher, L., Vempala, S., Wang, G.: Matrix approximation and projective clustering via volume sampling. Theory Comput. **2**, 255–247 (2006).

[2] Deshpande, A., Rademacher, L.,: Efficient volume sampling for row/column subset selection. in 2010 IEEE 51st Annual Symposium on Foundations of Computer Science (2010).

[3] Kaarnioja, V., Kazashi, Y., Kuo, F.Y., Nobile, F., Sloan, I.H.: Fast approximation by periodic kernel-based lattice-point interpolation with application in uncertainty quantification. Numer. Math. **150**, 33–77 (2022).

[4] Olver, F.W.J., et al.: NIST Digital Library of Mathematical Functions. http://dlmf.nist.gov/ (2023). Accessed 08 May 2023.