3. *Nyström method:* In class, we have shown that every manifold learning algorithm can be regarded as Kernel PCA on graphs: (1) given N data points, define a neighborhood graph with N nodes for data points; (2) construct a positive semidefinite kernel K; (3) pursue spectral decomposition of K to find the embedding (using top or bottom eigenvectors). However, this approach might suffer from the expensive computational cost in spectral decomposition of K if N is large and K is non-sparse, e.g. ISOMAP and MDS.

To overcome this hurdle, Nyström method leads us to a scalable approach to compute eigenvectors of low rank matrices. Suppose that an N-by-N positive semidefinite matrix $K \succeq 0$ admits the following block partition

$$K = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}. \tag{1}$$

where A is an n-by-n block. Assume that A has the spectral decomposition $A = U\Lambda U^T$, $\Lambda = \mathrm{diag}(\lambda_i)$ $(\lambda_1 \geq \lambda_2 \geq \ldots \lambda_k > \lambda_{k+1} = \ldots = 0)$ and $U = [u_1, \ldots, u_n]$ satisfies $U^T U = I$.

(a) Assume that $K = XX^T$ for some $X = [X_1; X_2] \in \mathbb{R}^{N \times k}$ with the block $X_1 \in \mathbb{R}^{n \times k}$. Show that X_1 and X_2 can be decided by:

$$X_1 = U_k \Lambda_k^{1/2}, \tag{2}$$

$$X_2 = B^T U_k \Lambda_k^{-1/2}, \tag{3}$$

where $U_k = [u_1, \ldots, u_k]$ consists of those k columns of U corresponding to top k eigenvalues λ_i $(i = 1, \ldots, k)$.

(b) Show that for general $K \succeq 0$, one can construct an approximation from (2) and (3),

$$\hat{K} = \begin{bmatrix} A & B \\ B^T & \hat{C} \end{bmatrix}. \tag{4}$$

where $A = X_1 X_1^T$, $B = X_1 X_2^T$, and $\hat{C} = X_2 X_2^T = B^T A^\dagger B$, $A^\dagger$ denoting the Moore-Penrose (pseudo-) inverse of A. Therefore $\|\hat{K} - K\|_F = \|C - B^T A^\dagger B\|_F$. Here the matrix $C - B^T A^\dagger B =: K/A$ is called the (generalized) *Schur Complement* of A in K.

(1) $A = U\Lambda U^T$. $\Lambda = \mathrm{diag}(\lambda_i)$

$\Rightarrow A = U_k \Lambda_k U_k^T$.

where $U_k = [u_1, \cdots, u_k]$ consists of top-k eigenvalues

$\Lambda_k = \mathrm{diag}(\lambda_1, \cdots, \lambda_k)$ (λ_i non-zero)

$X = (X_1 \; X_2)^T$

$K = XX^T = \begin{pmatrix} X_1 X_1^T & X_1 X_2^T \\ X_2 X_1^T & X_2 X_2^T \end{pmatrix}$

① $X_1 X_1^T = A = U_k \Lambda_k U_k^T$

$X_1 X_1^T = U_k \Lambda_k U_k^T$

$$= U_k \Lambda_k^{\frac{1}{2}} \Lambda_k^{\frac{1}{2}} U_k^T$$

$$= (U_k \Lambda_k^{\frac{1}{2}})(U_k \Lambda_k^{\frac{1}{2}})^T$$

Def $X_1 \triangleq U_k \Lambda_k^{\frac{1}{2}}$. Then $A = X_1 X_1^T$

② $X_1 X_2^T = B$

$$X_2^T = X_1^{-1} B = (\Lambda_k^{\frac{1}{2}})^{-1} U_k^{-1} B$$

$$= \Lambda_k^{-\frac{1}{2}} U_k^T B$$

$$X_2 = B^T U_k \Lambda_k^{-\frac{1}{2}}$$

(2) $X_1 X_1^T = A$

$X_1 X_2^T = B$

$X_2 X_1^T = (X_1 X_2^T)^T = B^T$

$X_2 X_2^T = B^T U_k \Lambda_k^{-1} U_k^T B$

Def $G \triangleq U_k \Lambda^{-1} U_k^T$. Then

① $GAG = (U_k \Lambda_k^{-1} U_k^T)(U_k \Lambda_k U_k^T)(U_k \Lambda_k^{-1} U_k^T)$

$= U_k \Lambda_k^{-1} U_k^T$

$= G$

② $AGA = (U_k \Lambda_k U_k^T)(U_k \Lambda_k^{-1} U_k^T)(U_k \Lambda_k U_k^T)$

$= U_k \Lambda_k U_k^T$

$= A$

③ $(AG)^T = (U_k \Lambda_k U_k^T U_k \Lambda_k^{-1} U_k^T)^T =] = AG$

By definition, G is the unique Moore-Penrose inverse of A. $G = A^\dagger$

Therefore,

$$\|K - \hat{K}\|_F = \sqrt{\|A - A\|_F^2 + \|B - B\|_F^2 + \|B^T - B^T\|_F^2 + \|C - B^T A^\dagger B\|_F^2}$$

$$K = \begin{bmatrix} A & B \\ B^T & \hat{C} \end{bmatrix}$$