A Proof of Theorem 3

The error analysis for the uniform+adaptive^2 algorithm relies on Lemma 7, which guarantees the error incurred by its uniform sampling step. The proof of Lemma 7 essentially follows Gittens (2011). We prove Lemma 7 using probability inequalities and some techniques of Boutsidis et al. (2011); Gittens (2011); Gittens and Mahoney (2013); Tropp (2012); the proof is in Appendix A.1.

Lemma 7 (Uniform Column Sampling)**.** *Given an $m \times n$ matrix $\mathbf{A}$ and a target rank k, let μ_k denote the matrix coherence of $\mathbf{A}$. By sampling*

$$c \;=\; \frac{\mu_k k \log(k/\delta)}{\theta \log \theta - \theta + 1},$$

columns uniformly without replacement to construct $\mathbf{C}$, the following inequality

$$\big\|\mathbf{A} - \mathcal{P}_{\mathbf{C},k}\mathbf{A}\big\|_F^2 \quad\leq\quad \big(1 + \delta^{-1}\theta^{-1}\big)\big\|\mathbf{A} - \mathbf{A}_k\big\|_F^2.$$

holds with probability at least $1 - 2\delta$. Here $\delta \in (0, 0.5)$ and $\theta \in (0, 1)$ are arbitrary real numbers.

The error analysis for the two adaptive sampling steps of the uniform+adaptive^2 algorithm relies on Lemma 8, which follows immediately from (Wang and Zhang, 2013, Corollary 7 and Section 4.5).

Lemma 8. *Given an $m \times m$ symmetric matrix $\mathbf{A}$ and a target rank k, we let $\mathbf{C}_1$ contain the c_1 columns of $\mathbf{A}$ selected by a column sampling algorithm such that the following inequality holds:*

$$\big\|\mathbf{A} - \mathcal{P}_{\mathbf{C}_1}\mathbf{A}\big\|_F^2 \quad\leq\quad f\big\|\mathbf{A} - \mathbf{A}_k\big\|_F^2.$$

Then we select $c_2 = kf\epsilon^{-1}$ columns to construct $\mathbf{C}_2$ and $c_3 = (c_1 + c_2)\epsilon^{-1}$ columns to construct $\mathbf{C}_3$, both using the adaptive sampling according to the residual $\mathbf{B}_1 = \mathbf{A} - \mathcal{P}_{\mathbf{C}_1}\mathbf{A}$ and $\mathbf{B}_2 = \mathbf{A} - \mathcal{P}_{[\mathbf{C}_1,\mathbf{C}_2]}\mathbf{A}$, respectively. Let $\mathbf{C} = [\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3]$, we have that

$$\mathbb{P}\left\{\frac{\big\|\mathbf{A} - \mathbf{C}\big(\mathbf{C}^\dagger\mathbf{A}(\mathbf{C}^\dagger)^T\big)\mathbf{C}^T\big\|_F}{\big\|\mathbf{A} - \mathbf{A}_k\big\|_F} \;\geq\; 1 + s\epsilon\right\} \;\leq\; \frac{1+\epsilon}{1+s\epsilon},$$

where s is an arbitrary constant greater than 1.

Finally Theorem 3 is proved by combining Lemma 7 and Lemma 8. The proof is in Appendix A.2.

A.1 Proof of Lemma 7

Proof. We use uniform column sampling to select c column of $\mathbf{A}$ to construct $\mathbf{C} = \mathbf{AS}$. Here the $n \times c$ random matrix $\mathbf{S}$ has one entry equal to one and the rest equal to zero in each column, and at most one nonzero entry in each row, and $\mathbf{S}$ is uniformly distributed among $\binom{n}{c}$ such kind of matrices. Applying Lemma 7 of Boutsidis et al. (2011), we get

$$\begin{aligned}&\big\|\mathbf{A} - \mathcal{P}_{\mathbf{C},k}\mathbf{A}\big\|_F^2\\ &\quad\leq \big\|\mathbf{A} - \mathbf{A}_k\big\|_F^2 + \big\|(\mathbf{A} - \mathbf{A}_k)\mathbf{S}\big\|_F^2\,\big\|(\mathbf{V}_{\mathbf{A},k}^T\mathbf{S})^\dagger\big\|_2^2. \quad (3)\end{aligned}$$

Now we bound $\big\|(\mathbf{A} - \mathbf{A}_k)\mathbf{S}\big\|_2^2$ and $\big\|(\mathbf{V}_{\mathbf{A},k}^T\mathbf{S})^\dagger\big\|_2^2$ respectively using the techniques of Gittens (2011); Gittens and Mahoney (2013); Tropp (2012).

Let $\mathcal{I} \subset [n]$ be a random index set corresponding to $\mathbf{S}$. The support of $\mathcal{I}$ is uniformly distributing among all the index sets in $2^{[n]}$ with cardinality c. According to Gittens and Mahoney (2013), the expectation of $\big\|(\mathbf{A} - \mathbf{A}_k)\mathbf{S}\big\|_F^2$ can be written as

$$\begin{aligned}\mathbb{E}\big\|(\mathbf{A} - \mathbf{A}_k)\mathbf{S}\big\|_F^2 \;&=\; \mathbb{E}\big\|(\mathbf{A} - \mathbf{A}_k)_{\mathcal{I}}\big\|_F^2\\ &=\; c\mathbb{E}\big\|(\mathbf{A} - \mathbf{A}_k)_i\big\|_F^2 \;=\; \frac{c}{n}\big\|\mathbf{A} - \mathbf{A}_k\big\|_F^2.\end{aligned}$$

Applying Markov's inequality, we have that

$$\begin{aligned}&\mathbb{P}\left\{\big\|(\mathbf{A} - \mathbf{A}_k)\mathbf{S}\big\|_F^2 \geq \frac{c}{n\delta}\big\|\mathbf{A} - \mathbf{A}_k\big\|_F^2\right\}\\ &\qquad\leq\; \frac{\mathbb{E}\big\|(\mathbf{A} - \mathbf{A}_k)\mathbf{S}\big\|_F^2}{\frac{c}{n\delta}\big\|\mathbf{A} - \mathbf{A}_k\big\|_F^2} \;=\; \delta. \qquad (4)\end{aligned}$$

Here $\delta \in (0, 0.5)$ is a real number defined later.

Now we establish the bound for $\mathbb{E}\big\|\mathbf{\Omega}_2^\dagger\big\|_2^2$ as follows. Let $\lambda_i(\mathbf{X})$ be the i-th largest eigenvalue of $\mathbf{X}$. Following the proof of Lemma 1 of Gittens (2011), we have

$$\begin{aligned}\big\|(\mathbf{V}_{\mathbf{A},k}^T\mathbf{S})^\dagger\big\|_2^2 \;&=\; \lambda_k^{-1}\Big(\mathbf{V}_{\mathbf{A},k}^T\mathbf{S}\mathbf{S}^T\mathbf{V}_{\mathbf{A},k}\Big)\\ &=\; \lambda_k^{-1}\Big(\sum_{i=1}^c \mathbf{X}_i\Big) \;\leq\; \lambda_{\min}^{-1}\Big(\sum_{i=1}^c \mathbf{X}_i\Big), \quad (5)\end{aligned}$$

where the random matrices $\mathbf{X}_1, \cdots, \mathbf{X}_c$ are chosen uniformly at random from the set $\Big\{\big(\mathbf{V}_{\mathbf{A},k}^T\big)_i\big(\mathbf{V}_{\mathbf{A},k}^T\big)_i^T\Big\}_{i=1}^n$ without replacement. The random matrices are of size $k \times k$. We accordingly define

$$R = \max_i \lambda_{\max}(\mathbf{X}_i) = \max_i \big\|\big(\mathbf{V}_{\mathbf{A},k}^T\big)_i\big\|_2^2 = \frac{k}{n}\mu_k,$$

where μ_k is the matrix coherence of $\mathbf{A}$, and define

$$\begin{aligned}\beta_{\min} \;&=\; \lambda_{\min}\Big(\mathbb{E}\sum_{i=1}^c \mathbf{X}_i\Big) \;=\; \lambda_{\min}\big(c\mathbb{E}\mathbf{X}_1\big)\\ &=\; \lambda_{\min}\Big(\frac{c}{n}\mathbf{V}_{\mathbf{A},k}^T\mathbf{V}_{\mathbf{A},k}\Big) \;=\; \frac{c}{n}.\end{aligned}$$

Then we apply Lemma 9 and obtained the following inequality:

$$\mathbb{P}\Big[\lambda_{\min}\Big(\sum_{i=1}^c \mathbf{X}_i\Big) \leq \frac{\theta c}{n}\Big] \quad\leq\quad k\left[\frac{e^{\theta-1}}{\theta^\theta}\right]^{\frac{c}{k\mu_k}} \;\triangleq\; \delta, (6)$$

where $\theta \in (0,1]$ is a real number, and it follows that

$$c \;=\; \frac{\mu_k k \log(k/\delta)}{\theta \log\theta - \theta + 1}.$$

Applying (5) and (6), we have

$$\mathbb{P}\Big\{ \big\| (\mathbf{V}_{\mathbf{A},k}^T \mathbf{S})^\dagger \big\|_2^2 \geq \frac{n}{\theta c} \Big\} \quad \leq \quad \delta. \tag{7}$$

Combining (4) and (7) and applying the union bound, we have the following inequality:

$$\mathbb{P}\Big\{ \big\| (\mathbf{A} - \mathbf{A}_k)\mathbf{S} \big\|_F^2 \geq \frac{c}{n\delta} \big\| \mathbf{A} - \mathbf{A}_k \big\|_F^2 \\ \text{or} \quad \big\| (\mathbf{V}_{\mathbf{A},k}^T \mathbf{S})^\dagger \big\|_2^2 \geq \frac{n}{\theta c} \Big\} \;\leq\; 2\delta. \tag{8}$$

Finally, from (3) and (8) we have that the inequality

$$\big\| \mathbf{A} - \mathcal{P}_{\mathbf{C},k} \mathbf{A} \big\|_F^2 \quad \leq \quad \big(1 + \delta^{-1}\theta^{-1}\big) \big\| \mathbf{A} - \mathbf{A}_k \big\|_F^2$$

holds with probability at least $1 - 2\delta$, by which the lemma follows. □

Lemma 9 (Theorem 5.1.1 of Tropp (2012))**.** *We are given l independent random $d \times d$ SPSD matrices $\mathbf{X}_1, \cdots, \mathbf{X}_l$ with the property*

$$\lambda_{\max}(\mathbf{X}_i) \leq R \quad \textit{for } \; i = 1, \cdots, l.$$

We define $\mathbf{Y} = \sum_{i=1}^l \mathbf{X}_i$ and $\beta_{\min} = \lambda_{\min}\big(\mathbb{E}\mathbf{Y}\big)$. Then for any $\theta \in (0,1]$, the following inequality holds:

$$\mathbb{P}\Big\{ \lambda_{\min}(\mathbf{Y}) \leq \theta \beta_{\min} \Big\} \;\leq\; k \left[\frac{e^{\theta - 1}}{\theta^\theta} \right]^{\frac{\beta_{\min}}{R}}.$$

A.2 Proof of the Theorem

Proof. The matrix $\mathbf{C}_1$ consists of c_1 columns selected by uniform sampling, and $\mathbf{C}_2 \in \mathbb{R}^{m \times c_2}$ and $\mathbf{C}_3 \in \mathbb{R}^{m \times c_3}$ are constructed by adaptive sampling. We set $\delta = 1/\sqrt{5}$ and $\theta = \sqrt{5}/4$ for Lemma 7, then we have

$$\begin{aligned} f \;&=\; 1 + \delta^{-1}\theta^{-1} \;=\; 5, \\ c_1 \;&=\; \frac{\mu_k k \log(k/\delta)}{\theta \log\theta - \theta + 1} \;=\; 8.7 \mu_k k \log(\sqrt{5}k). \end{aligned}$$

Then we set

$$\begin{aligned} c_2 \;&=\; k f \epsilon^{-1} \;=\; 5k\epsilon^{-1}, \\ c_3 \;&=\; (c_1 + c_2)\epsilon^{-1}, \end{aligned}$$

according to Lemma 8. Letting $s > 1$ be an arbitrary constant, we have that

$$\begin{aligned} &\mathbb{P}\left\{ \frac{\big\| \mathbf{A} - \mathbf{C}\mathbf{U}\mathbf{C}^T \big\|_F}{\big\| \mathbf{A} - \mathbf{A}_k \big\|_F} \leq 1 + s\epsilon \right\} \\ &\geq \; \mathbb{P}\left\{ \frac{\big\| \mathbf{A} - \mathbf{C}\mathbf{U}\mathbf{C}^T \big\|_F}{\big\| \mathbf{A} - \mathbf{A}_k \big\|_F} \leq 1 + s\epsilon \;\middle|\; \frac{\big\| \mathbf{A} - \mathcal{P}_{\mathbf{C}_1}\mathbf{A} \big\|_F^2}{\big\| \mathbf{A} - \mathbf{A}_k \big\|_F^2} \leq f \right\} \\ &\qquad \cdot \; \mathbb{P}\left\{ \frac{\big\| \mathbf{A} - \mathcal{P}_{\mathbf{C}_1}\mathbf{A} \big\|_F^2}{\big\| \mathbf{A} - \mathbf{A}_k \big\|_F^2} \leq f \right\} \\ &\geq \; \Big(1 - \frac{1+\epsilon}{1+s\epsilon}\Big)\Big(1 - 2\delta\Big). \end{aligned}$$

where the last inequality follows from Lemma 7 and Lemma 8.

Repeating the sampling procedure for t times and letting $\mathbf{C}_{[i]}$ and $\mathbf{U}_{[i]}$ be the i-th sample, we obtain an upper error bound on the failure probability:

$$\begin{aligned} &\mathbb{P}\left\{ \min_{i \in [t]} \left\{ \frac{\big\| \mathbf{A} - \mathbf{C}_{[i]}\mathbf{U}_{[i]}\mathbf{C}_{[i]}^T \big\|_F}{\big\| \mathbf{A} - \mathbf{A}_k \big\|_F} \right\} \;\geq\; 1 + s\epsilon \right\} \\ &\qquad \leq \; \left(1 - \Big(1 - \frac{1+\epsilon}{1+s\epsilon}\Big)\Big(1 - 2\delta\Big) \right)^t \\ &\qquad = \; \left(1 + \frac{(s-1)(1-2\delta)}{\epsilon^{-1} + 1 + 2\delta(s-1)} \right)^{-t} \;\triangleq\; p. \end{aligned}$$

Taking logarithm of both sides of the equality and applying $\log(1+x) \approx x$ when x is small, we have

$$\begin{aligned} t \;&=\; \left[\log\Big(1 + \frac{(1-2\delta)(s-1)}{\epsilon^{-1} + 1 + 2\delta(s-1)} \Big) \right]^{-1} \log\frac{1}{p} \\ &\approx \; \frac{\epsilon^{-1} + 1 + 2\delta(s-1)}{(1-2\delta)(s-1)} \log\frac{1}{p}. \end{aligned}$$

Setting $s = 2$, we have that $t \approx (10\epsilon^{-1} + 18)\log(1/p)$.

Hence by sampling totally

$$c \;=\; \big(1 + \epsilon^{-1}\big)\big(5k\epsilon^{-1} + 8.7\mu_k k \log(\sqrt{5}k)\big)$$

columns and repeating the procedure for

$$t \;\geq\; (10\epsilon^{-1} + 18)\log(1/p)$$

times, the algorithm attains the upper error bound

$$\big\| \mathbf{A} - \mathbf{C}\big(\mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T\big)\mathbf{C}^T \big\|_F \;\leq\; \big(1 + 2\epsilon\big) \big\| \mathbf{A} - \mathbf{A}_k \big\|_F$$

with probability at least $1 - p$. Substituting 2ϵ by ϵ' yields the error bound in the theorem.

Time complexity and space complexity of Algorithm 1 is calculated as follows. The uniform sampling costs $\mathcal{O}(m)$ time; the first adaptive sampling round costs $\mathcal{O}(mc_1^2) + T_{\text{Multiply}}(m^2 c_1)$ time; the second adaptive sampling round costs $\mathcal{O}(m(c_1+c_2)^2) + T_{\text{Multiply}}(m^2(c_1+c_2))$ time; computing the intersection matrix costs $\mathcal{O}(mc^2) +$

$T_{\text{Multiply}}(m^2c)$ time in general. So the total time complexity is $\mathcal{O}(mc^2) + T_{\text{Multiply}}(m^2c)$ without using Theorem 4, or $\mathcal{O}(m(c_1+c_2)^2) + T_{\text{Multiply}}(m^2c)$ using Theorem 4. As for the space complexity, the Moore-Penrose inverse of an $m \times c$ matrix demands $\mathcal{O}(mc)$ space, and multiplying a $c \times m$ matrix $\mathbf{C}^\dagger$ by an $m \times m$ matrix $\mathbf{A}$ costs $\mathcal{O}(mc)$ space by partition $\mathbf{A}$ into small blocks of size smaller than $m \times c$ and loading one block into RAM at a time to perform matrix multiplication. □

B Proof of Theorem 4

Proof. Let $\mathbf{C} \in \mathbb{R}^{m\times c}$ consists of a subset of columns of $\mathbf{A}$. By row permutation $\mathbf{C}$ can be expressed as

$$\mathbf{PC} = \begin{bmatrix} \mathbf{W} \\ \mathbf{A}_{21} \end{bmatrix}.$$

Then according to Lemma 10, the Moore-Penrose inverse of $\mathbf{C}$ can be written as

$$\mathbf{C}^\dagger = \mathbf{W}^{-1}\big(\mathbf{I}_c + \mathbf{S}^T\mathbf{S}\big)^{-1}\begin{bmatrix} \mathbf{I}_c & \mathbf{S}^T \end{bmatrix}\mathbf{P},$$

where $\mathbf{S} = \mathbf{A}_{21}\mathbf{W}^{-1}$. Then the intersection matrix of modified Nyström approximation to $\mathbf{A}$ can be expressed as

$$\begin{aligned}
\mathbf{U} &= \mathbf{C}^\dagger\mathbf{A}\big(\mathbf{C}^\dagger\big)^T \\
&= \mathbf{W}^{-1}\big(\mathbf{I}_c + \mathbf{S}^T\mathbf{S}\big)^{-1}\begin{bmatrix} \mathbf{I}_c & \mathbf{S}^T \end{bmatrix}\mathbf{P}\mathbf{A}\mathbf{P}^T \\
&\qquad \begin{bmatrix} \mathbf{I}_c \\ \mathbf{S} \end{bmatrix}\big(\mathbf{I}_c + \mathbf{S}^T\mathbf{S}\big)^{-1}\mathbf{W}^{-1} \\
&= \mathbf{W}^{-1}\big(\mathbf{I}_c + \mathbf{S}^T\mathbf{S}\big)^{-1}\begin{bmatrix} \mathbf{I}_c & \mathbf{S}^T \end{bmatrix} \\
&\qquad \begin{bmatrix} \mathbf{W} & \mathbf{A}_{21}^T \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}\begin{bmatrix} \mathbf{I}_c \\ \mathbf{S} \end{bmatrix}\big(\mathbf{I}_c + \mathbf{S}^T\mathbf{S}\big)^{-1}\mathbf{W}^{-1} \\
&= \mathbf{W}^{-1}\big(\mathbf{I}_c + \mathbf{S}^T\mathbf{S}\big)^{-1}\Big(\mathbf{W} + \mathbf{A}_{21}^T\mathbf{S} + (\mathbf{A}_{21}^T\mathbf{S})^T \\
&\qquad + \mathbf{S}^T\mathbf{A}_{22}\mathbf{S}\Big)\big(\mathbf{I}_c + \mathbf{S}^T\mathbf{S}\big)^{-1}\mathbf{W}^{-1} \\
&\triangleq \mathbf{T}_1\big(\mathbf{W} + \mathbf{T}_2 + \mathbf{T}_2^T + \mathbf{T}_3\big)\mathbf{T}_1^T.
\end{aligned}$$

Here the intermediate matrices are computed by

$$\begin{aligned}
\mathbf{T}_0 &= \mathbf{A}_{21}^T\mathbf{A}_{21}, \\
\mathbf{T}_1 &= \mathbf{W}^{-1}\big(\mathbf{I}_c + \mathbf{S}^T\mathbf{S}\big)^{-1} \\
&= \mathbf{W}^{-1}\Big(\mathbf{I}_c + \mathbf{W}^{-1}\mathbf{T}_0\mathbf{W}^{-1}\Big)^{-1}, \\
\mathbf{T}_2 &= \mathbf{A}_{21}^T\mathbf{S} = \mathbf{A}_{21}^T\mathbf{A}_{21}\mathbf{W}^{-1} = \mathbf{T}_0\mathbf{W}^{-1}, \\
\mathbf{T}_3 &= \mathbf{S}^T\mathbf{A}_{22}\mathbf{S} = \mathbf{W}^{-1}\Big(\mathbf{A}_{21}^T\mathbf{A}_{22}\mathbf{A}_{21}\Big)\mathbf{W}^{-1}.
\end{aligned}$$

The matrix inverse operations are on $c \times c$ matrices which costs $\mathcal{O}(c^3)$ time. The matrix multiplication $\mathbf{A}_{21}^T\mathbf{A}_{22}\mathbf{A}_{21}$ requires time $T_{\text{Multiply}}\big((m-c)^2c\big)$. □

Lemma 10 (The Moore Penrose Inverse of Partitioned Matrices (Ben-Israel and Greville, 2003, Page 179))**.** *Given a matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$ of rank of at least c which has a nonsingular $c \times c$ submatrix $\mathbf{X}_{11}$. By rearrangement of columns and rows by permutation matrices $\mathbf{P}$ and $\mathbf{Q}$, the submatrix $\mathbf{X}_{11}$ can be bought to the top left corner of $\mathbf{X}$, that is,*

$$\mathbf{PXQ} = \begin{bmatrix} \mathbf{X}_{11} & \mathbf{X}_{12} \\ \mathbf{X}_{21} & \mathbf{X}_{22} \end{bmatrix}.$$

Then the Moore-Penrose inverse of $\mathbf{X}$ is

$$\begin{aligned}
\mathbf{X}^\dagger = \mathbf{Q}\begin{bmatrix} \mathbf{I}_c \\ \mathbf{T}^T \end{bmatrix}\big(\mathbf{I}_c + \mathbf{T}\mathbf{T}^T\big)^{-1}\mathbf{X}_{11}^{-1} \\
\big(\mathbf{I}_c + \mathbf{S}\mathbf{S}^T\big)^{-1}\begin{bmatrix} \mathbf{I}_c & \mathbf{S}^T \end{bmatrix}\mathbf{P},
\end{aligned}$$

where $\mathbf{T} = \mathbf{X}_{11}^{-1}\mathbf{X}_{12}$ and $\mathbf{S} = \mathbf{X}_{21}\mathbf{X}_{11}^{-1}$.

C The Proof of Theorem 5

Proof. Suppose that $\text{rank}(\mathbf{W}) = \text{rank}(\mathbf{A})$. We have that $\text{rank}(\mathbf{W}) = \text{rank}(\mathbf{C}) = \text{rank}(\mathbf{A})$ because

$$\text{rank}(\mathbf{A}) \geq \text{rank}(\mathbf{C}) \geq \text{rank}(\mathbf{W}). \tag{9}$$

Thus there exists a matrix $\mathbf{X}$ such that

$$\begin{bmatrix} \mathbf{A}_{21}^T \\ \mathbf{A}_{22} \end{bmatrix} = \mathbf{C}\mathbf{X}^T = \begin{bmatrix} \mathbf{W}\mathbf{X}^T \\ \mathbf{A}_{21}\mathbf{X}^T \end{bmatrix},$$

and it follows that $\mathbf{A}_{21} = \mathbf{X}\mathbf{W}$ and $\mathbf{A}_{22} = \mathbf{A}_{21}\mathbf{X}^T = \mathbf{X}\mathbf{W}\mathbf{X}^T$. Then we have that

$$\begin{aligned}
\mathbf{A} &= \begin{bmatrix} \mathbf{W} & (\mathbf{X}\mathbf{W})^T \\ \mathbf{X}\mathbf{W} & \mathbf{X}\mathbf{W}\mathbf{X}^T \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{I} \\ \mathbf{X} \end{bmatrix}\mathbf{W}\begin{bmatrix} \mathbf{I} & \mathbf{X}^T \end{bmatrix}, \qquad (10) \\
\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T &= \begin{bmatrix} \mathbf{W} \\ \mathbf{X}\mathbf{W} \end{bmatrix}\mathbf{W}^\dagger\begin{bmatrix} \mathbf{W} & (\mathbf{X}\mathbf{W})^T \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{I} \\ \mathbf{X} \end{bmatrix}\mathbf{W}\begin{bmatrix} \mathbf{I} & \mathbf{X}^T \end{bmatrix}. \qquad (11)
\end{aligned}$$

Here the second equality in (11) follows from $\mathbf{W}\mathbf{W}^\dagger\mathbf{W} = \mathbf{W}$. We obtain that $\mathbf{A} = \mathbf{C}\mathbf{W}^\dagger\mathbf{C}$. Then we show that $\mathbf{A} = \mathbf{C}\mathbf{C}^\dagger\mathbf{A}(\mathbf{C}^\dagger)^T\mathbf{C}^T$.

Since $\mathbf{C}^\dagger = (\mathbf{C}^T\mathbf{C})^\dagger\mathbf{C}^T$, we have that

$$\mathbf{C}^\dagger = \big(\mathbf{W}(\mathbf{I} + \mathbf{X}^T\mathbf{X})\mathbf{W}\big)^\dagger\mathbf{W}\,[\mathbf{I}, \mathbf{X}^T],$$

and thus

$$\begin{aligned}
&\mathbf{C}^\dagger\mathbf{A}(\mathbf{C}^\dagger)^T\mathbf{W} \\
&= \big(\mathbf{W}(\mathbf{I} + \mathbf{X}^T\mathbf{X})\mathbf{W}\big)^\dagger\mathbf{W}(\mathbf{I} + \mathbf{X}^T\mathbf{X})\Big[\mathbf{W}(\mathbf{I} + \mathbf{X}^T\mathbf{X}) \\
&\qquad \mathbf{W}\big(\mathbf{W}(\mathbf{I} + \mathbf{X}^T\mathbf{X})\mathbf{W}\big)^\dagger\mathbf{W}\Big] \\
&= \big(\mathbf{W}(\mathbf{I} + \mathbf{X}^T\mathbf{X})\mathbf{W}\big)^\dagger\mathbf{W}(\mathbf{I} + \mathbf{X}^T\mathbf{X})\mathbf{W},
\end{aligned}$$

where the second equality follows from Lemma 11 because $(\mathbf{I}+\mathbf{X}^T\mathbf{X})$ is positive definite. Similarly we have

$$\begin{aligned}&\mathbf{W}\mathbf{C}^\dagger\mathbf{A}(\mathbf{C}^\dagger)^T\mathbf{W}\\ &= \mathbf{W}\big(\mathbf{W}(\mathbf{I}+\mathbf{X}^T\mathbf{X})\mathbf{W}\big)^\dagger\mathbf{W}(\mathbf{I}+\mathbf{X}^T\mathbf{X})\mathbf{W} = \mathbf{W}.\end{aligned}$$

Thus we have

$$\begin{aligned}\mathbf{C}\mathbf{C}^\dagger\mathbf{A}(\mathbf{C}^\dagger)^T\mathbf{C} &= \begin{bmatrix}\mathbf{I}\\ \mathbf{X}\end{bmatrix}\mathbf{W}\mathbf{C}^\dagger\mathbf{A}(\mathbf{C}^\dagger)^T\mathbf{W}\begin{bmatrix}\mathbf{I} & \mathbf{X}^T\end{bmatrix}\\ &= \begin{bmatrix}\mathbf{I}\\ \mathbf{X}\end{bmatrix}\mathbf{W}\begin{bmatrix}\mathbf{I} & \mathbf{X}^T\end{bmatrix}. \qquad (12)\end{aligned}$$

It follows from Equations (10) (11) (12) that $\mathbf{A} = \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T = \mathbf{C}\mathbf{C}^\dagger\mathbf{A}(\mathbf{C}^\dagger)^T\mathbf{C}^T$.

Conversely, when $\mathbf{A} = \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T$, we have that $\mathrm{rank}(\mathbf{A}) \le \mathrm{rank}(\mathbf{W}^\dagger) = \mathrm{rank}(\mathbf{W})$. By applying (9) we have that $\mathrm{rank}(\mathbf{A}) = \mathrm{rank}(\mathbf{W})$.

When $\mathbf{A} = \mathbf{C}\mathbf{C}^\dagger\mathbf{A}(\mathbf{C}^\dagger)^T\mathbf{C}^T$, we have $\mathrm{rank}(\mathbf{A}) \le \mathrm{rank}(\mathbf{C})$. Thus there exists a matrix $\mathbf{X}$ such that

$$\begin{bmatrix}\mathbf{A}_{21}^T\\ \mathbf{A}_{22}\end{bmatrix} = \mathbf{C}\mathbf{X}^T = \begin{bmatrix}\mathbf{W}\mathbf{X}^T\\ \mathbf{A}_{21}\mathbf{X}^T\end{bmatrix},$$

and therefore $\mathbf{A}_{21} = \mathbf{X}\mathbf{W}$. Then we have that

$$\mathbf{C} = \begin{bmatrix}\mathbf{W}\\ \mathbf{A}_{21}\end{bmatrix} = \begin{bmatrix}\mathbf{I}\\ \mathbf{X}\end{bmatrix}\mathbf{W},$$

so $\mathrm{rank}(\mathbf{C}) \le \mathrm{rank}(\mathbf{W})$. Apply (9) again we have $\mathrm{rank}(\mathbf{A}) = \mathrm{rank}(\mathbf{W})$. □

Lemma 11. $\mathbf{X}^T\mathbf{V}\mathbf{X}\big(\mathbf{X}^T\mathbf{V}\mathbf{X}\big)^\dagger\mathbf{X}^T = \mathbf{X}^T$ *for any positive definite matrix* $\mathbf{V}$.

Proof. Since the positive definite matrix $\mathbf{V}$ have a decomposition $\mathbf{V} = \mathbf{B}^T\mathbf{B}$ for some nonsingular matrix $\mathbf{B}$, so we have

$$\begin{aligned}&\mathbf{X}^T\mathbf{V}\mathbf{X}\big(\mathbf{X}^T\mathbf{V}\mathbf{X}\big)^\dagger\mathbf{X}^T\\ &= (\mathbf{B}\mathbf{X})^T\Big(\mathbf{B}\mathbf{X}\big((\mathbf{B}\mathbf{X})^T(\mathbf{B}\mathbf{X})\big)^\dagger\Big)(\mathbf{B}\mathbf{X})^T\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\\ &= (\mathbf{B}\mathbf{X})^T\big((\mathbf{B}\mathbf{X})^T\big)^\dagger(\mathbf{B}\mathbf{X})^T(\mathbf{B}^T)^{-1}\\ &= (\mathbf{B}\mathbf{X})^T(\mathbf{B}^T)^{-1}\\ &= \mathbf{X}^T.\end{aligned}$$

□

D Proof of Theorem 6

In Section D.1 we provide two key lemmas, and then in Section D.2 we prove Theorem 6 using the two lemmas.

D.1 Key Lemmas

Lemma 12. *For an $m\times m$ matrix $\mathbf{B}$ with diagonal entries equal to one and off-diagonal entries equal to α, the error incurred by the modified Nyström method is lower bounded by*

$$\begin{aligned}&\|\mathbf{B}-\tilde{\mathbf{B}}_c^{mod}\|_F^2\\ &\quad\ge (1-\alpha)^2(m-c)\bigg(1+\frac{2}{c}-(1-\alpha)\frac{1+o(1)}{\alpha cm/2}\bigg).\end{aligned}$$

Proof. Without loss of generality, we assume the first c column of $\mathbf{B}$ are selected to construct $\mathbf{C}$. We partition $\mathbf{B}$ and $\mathbf{C}$ as:

$$\mathbf{B} = \begin{bmatrix}\mathbf{W} & \mathbf{B}_{21}^T\\ \mathbf{B}_{21} & \mathbf{B}_{22}\end{bmatrix} \quad\text{and}\quad \mathbf{C} = \begin{bmatrix}\mathbf{W}\\ \mathbf{B}_{21}\end{bmatrix}.$$

Here the matrix $\mathbf{W}$ can be expressed by $\mathbf{W} = (1-\alpha)\mathbf{I}_c+\alpha\mathbf{1}_c\mathbf{1}_c^T$. We apply the Sherman-Morrison-Woodbury formula

$$(\mathbf{A}+\mathbf{B}\mathbf{C}\mathbf{D})^{-1} = \mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1}+\mathbf{D}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{D}\mathbf{A}^{-1}$$

to compute $\mathbf{W}^{-1}$, yielding

$$\mathbf{W}^{-1} = \frac{1}{1-\alpha}\mathbf{I}_c - \frac{\alpha}{(1-\alpha)(1-\alpha+c\alpha)}\mathbf{1}_c\mathbf{1}_c^T. \qquad (13)$$

We expand the Moore-Penrose inverse of $\mathbf{C}$ by Lemma 10 and obtain

$$\mathbf{C}^\dagger = \mathbf{W}^{-1}\big(\mathbf{I}_c+\mathbf{S}^T\mathbf{S}\big)^{-1}\begin{bmatrix}\mathbf{I}_c & \mathbf{S}^T\end{bmatrix}$$

where

$$\mathbf{S} = \mathbf{B}_{21}\mathbf{W}^{-1} = \frac{\alpha}{1-\alpha+c\alpha}\mathbf{1}_{m-c}\mathbf{1}_c^T.$$

It is easily verified that $\mathbf{S}^T\mathbf{S} = \big(\frac{\alpha}{1-\alpha+c\alpha}\big)^2(m-c)\mathbf{1}_c\mathbf{1}_c^T$.

Now we express the matrix constructed by the modified Nyström method in a partitioned form:

$$\begin{aligned}\tilde{\mathbf{B}}_c^{\text{mod}} &= \mathbf{C}\mathbf{C}^\dagger\mathbf{B}\big(\mathbf{C}^\dagger\big)^T\mathbf{C}^T\\ &= \begin{bmatrix}\mathbf{W}\\ \mathbf{B}_{21}\end{bmatrix}\mathbf{W}^{-1}\big(\mathbf{I}_c+\mathbf{S}^T\mathbf{S}\big)^{-1}\begin{bmatrix}\mathbf{I}_c & \mathbf{S}^T\end{bmatrix}\mathbf{B}\\ &\qquad\begin{bmatrix}\mathbf{I}_c\\ \mathbf{S}\end{bmatrix}\big(\mathbf{I}_c+\mathbf{S}^T\mathbf{S}\big)^{-1}\mathbf{W}^{-1}\begin{bmatrix}\mathbf{W}\\ \mathbf{B}_{21}\end{bmatrix}^T\\ &= \begin{bmatrix}\big(\mathbf{I}_c+\mathbf{S}^T\mathbf{S}\big)^{-1}\\ \mathbf{B}_{21}\mathbf{W}^{-1}\big(\mathbf{I}_c+\mathbf{S}^T\mathbf{S}\big)^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{I}_c & \mathbf{S}^T\end{bmatrix}\mathbf{B}\\ &\qquad\begin{bmatrix}\mathbf{I}_c\\ \mathbf{S}\end{bmatrix}\begin{bmatrix}\big(\mathbf{I}_c+\mathbf{S}^T\mathbf{S}\big)^{-1}\\ \mathbf{B}_{21}\mathbf{W}^{-1}\big(\mathbf{I}_c+\mathbf{S}^T\mathbf{S}\big)^{-1}\end{bmatrix}^T. \qquad (14)\end{aligned}$$

We then compute the submatrices $\big(\mathbf{I}_c+\mathbf{S}^T\mathbf{S}\big)^{-1}$ and $\mathbf{B}_{21}\mathbf{W}^{-1}\big(\mathbf{I}_c+\mathbf{S}^T\mathbf{S}\big)^{-1}$ respectively as follows. We apply

the Sherman-Morrison-Woodbury formula to compute $\big(\mathbf{I}_c + \mathbf{S}^T\mathbf{S}\big)^{-1}$, yielding

$$\begin{aligned}\big(\mathbf{I}_c + \mathbf{S}^T\mathbf{S}\big)^{-1} &= \Big(\mathbf{I}_c + \Big(\frac{\alpha}{1-\alpha+c\alpha}\Big)^2(m-c)\mathbf{1}_c\mathbf{1}_c^T\Big)^{-1} \\ &= \mathbf{I}_c - \gamma_1\mathbf{1}_c\mathbf{1}_c^T, \end{aligned} \tag{15}$$

where

$$\gamma_1 = \frac{m-c}{mc + \big(\frac{1-\alpha}{\alpha}\big)^2 + \frac{2(1-\alpha)c}{\alpha}}.$$

It follows from (13) and (15) that

$$\begin{aligned}\mathbf{W}^{-1}\big(\mathbf{I}_c + \mathbf{S}^T\mathbf{S}\big)^{-1} &= (\gamma_2\mathbf{I}_c - \gamma_3\mathbf{1}_c\mathbf{1}_c^T)(\mathbf{I}_c - \gamma_1\mathbf{1}_c\mathbf{1}_c^T) \\ &= \gamma_2\mathbf{I}_c + (\gamma_1\gamma_3 c - \gamma_1\gamma_2 - \gamma_3)\mathbf{1}_c\mathbf{1}_c^T \end{aligned} \tag{16}$$

where

$$\gamma_2 = \frac{1}{1-\alpha} \quad \text{and} \quad \gamma_3 = \frac{\alpha}{(1-\alpha)(1-\alpha+\alpha c)}.$$

Then we have that

$$\begin{aligned}&\mathbf{B}_{21}\mathbf{W}^{-1}\big(\mathbf{I}_c + \mathbf{S}^T\mathbf{S}\big)^{-1} \\ &\quad= \alpha\big(\gamma_1\gamma_3c^2 - \gamma_3c - \gamma_1\gamma_2c + \gamma_2\big)\mathbf{1}_{m-c}\mathbf{1}_c^T \\ &\quad\triangleq \gamma\mathbf{1}_{m-c}\mathbf{1}_c^T, \end{aligned} \tag{17}$$

where

$$\begin{aligned}\gamma &= \alpha\big(\gamma_1\gamma_3c^2 - \gamma_3c - \gamma_1\gamma_2c + \gamma_2\big) \\ &= \frac{\alpha(\alpha c - \alpha + 1)}{2\alpha c - 2\alpha - 2\alpha^2c + \alpha^2 + \alpha^2cm + 1}. \end{aligned} \tag{18}$$

Since $\mathbf{B}_{21} = \alpha\mathbf{1}_{m-c}\mathbf{1}_c^T$ and $\mathbf{B}_{22} = (1-\alpha)\mathbf{I}_{m-c} + \alpha\mathbf{1}_{m-c}\mathbf{1}_{m-c}^T$, it is easily verified that

$$\begin{aligned}&\begin{bmatrix}\mathbf{I}_c & \mathbf{S}^T\end{bmatrix}\mathbf{B}\begin{bmatrix}\mathbf{I}_c \\ \mathbf{S}\end{bmatrix} \\ &\quad= \begin{bmatrix}\mathbf{I}_c & \mathbf{S}^T\end{bmatrix}\begin{bmatrix}\mathbf{W} & \mathbf{B}_{21}^T \\ \mathbf{B}_{21} & \mathbf{B}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{I}_c \\ \mathbf{S}\end{bmatrix} \\ &\quad= (1-\alpha)\mathbf{I}_c + \lambda\mathbf{1}_c\mathbf{1}_c^T, \end{aligned} \tag{19}$$

where

$$\lambda = \frac{\alpha(3\alpha m - \alpha c - 2\alpha + \alpha^2c - 3\alpha^2m + \alpha^2 + \alpha^2m^2 + 1)}{(\alpha c - \alpha + 1)^2}$$

It follows from (14), (15), (17), and (19) that

$$\begin{aligned}&\tilde{\mathbf{B}}_c^{\text{mod}} \\ &= \begin{bmatrix}\mathbf{I}_c - \gamma_1\mathbf{1}_c\mathbf{1}_c^T \\ \gamma\mathbf{1}_{m-c}\mathbf{1}_c^T\end{bmatrix}\Big((1-\alpha)\mathbf{I}_c + \lambda\mathbf{1}_c\mathbf{1}_c^T\Big)\begin{bmatrix}\mathbf{I}_c - \gamma_1\mathbf{1}_c\mathbf{1}_c^T \\ \gamma\mathbf{1}_{m-c}\mathbf{1}_c^T\end{bmatrix}^T \\ &\triangleq \begin{bmatrix}\tilde{\mathbf{B}}_{11} & \tilde{\mathbf{B}}_{21}^T \\ \tilde{\mathbf{B}}_{21} & \tilde{\mathbf{B}}_{22}\end{bmatrix}, \end{aligned}$$

where

$$\begin{aligned}\tilde{\mathbf{B}}_{11} &= (1-\alpha)\mathbf{I}_c + \big[(1-\gamma_1c) \\ &\quad (\lambda - \lambda\gamma_1c - (1-\alpha)\gamma_1) - (1-\alpha)\gamma_1\big]\mathbf{1}_c\mathbf{1}_c^T \\ &= (1-\alpha)\mathbf{I}_c + \eta_1\mathbf{1}_c\mathbf{1}_c^T, \\ \tilde{\mathbf{B}}_{21} &= \tilde{\mathbf{A}}_{12}^T = \gamma(1-\gamma_1c)(1-\alpha+\lambda c)\mathbf{1}_{m-c}\mathbf{1}_c^T \\ &= \eta_2\mathbf{1}_{m-c}\mathbf{1}_c^T, \\ \tilde{\mathbf{B}}_{22} &= \gamma^2c(1-\alpha+\lambda c)\mathbf{1}_{m-c}\mathbf{1}_{m-c}^T \\ &= \eta_3\mathbf{1}_{m-c}\mathbf{1}_{m-c}^T, \end{aligned}$$

where

$$\begin{aligned}\eta_1 &= (1-\gamma_1c)(\lambda - \lambda\gamma_1c - (1-\alpha)\gamma_1) - (1-\alpha)\gamma_1, \\ \eta_2 &= \gamma(1-\gamma_1c)(1-\alpha+\lambda c), \\ \eta_3 &= \gamma^2c(1-\alpha+\lambda c), \end{aligned}$$

By dealing with the four blocks of $\tilde{\mathbf{B}}_c^{\text{mod}}$ respectively, we finally obtain that

$$\begin{aligned}&\|\mathbf{B} - \tilde{\mathbf{B}}_c^{\text{mod}}\|_F^2 \\ &= \|\mathbf{W} - \tilde{\mathbf{B}}_{11}\|_F^2 + 2\|\mathbf{B}_{21} - \tilde{\mathbf{B}}_{21}\|_F^2 + \|\mathbf{B}_{22} - \tilde{\mathbf{B}}_{22}\|_F^2 \\ &= c^2(\alpha-\eta_1)^2 + 2c(m-c)(\alpha-\eta_2)^2 \\ &\quad + (m-c)(m-c-1)(\alpha-\eta_3)^2 + (m-c)(1-\eta_3)^2 \\ &= (m-c)(\alpha-1)^2\big(\alpha^4c^2m^2 - 4\alpha^4c^2m + 4\alpha^4c^2 \\ &\quad + 2\alpha^4cm^2 - 4\alpha^4cm + \alpha^4c + \alpha^4m - \alpha^4 + 4\alpha^3c^2m \\ &\quad - 8\alpha^3c^2 + 2\alpha^3cm + 2\alpha^3c - 2\alpha^3m + 2\alpha^3 + 4\alpha^2c^2 \\ &\quad + 2\alpha^2cm - 7\alpha^2c + \alpha^2m + 4\alpha c - 2\alpha + 1\big)/\big(2\alpha c \\ &\quad - 2\alpha - 2\alpha^2c + \alpha^2 + \alpha^2cm + 1\big)^2 \\ &= (m-c)(\alpha-1)^2\bigg(1 + \frac{2}{c} - \frac{(1-\alpha)}{c}\big(6\alpha c - 6\alpha \\ &\quad - 12\alpha^2c + 6\alpha^3c + 6\alpha^2 - 2\alpha^3 + 3\alpha^2c^2 - 3\alpha^3c^2 \\ &\quad + 2\alpha^3c^2m + 3\alpha^2cm - 3\alpha^3cm + 2\big)/\big(2\alpha c - 2\alpha \\ &\quad - 2\alpha^2c + \alpha^2 + \alpha^2cm + 1\big)^2\bigg) \\ &= (m-c)(\alpha-1)^2\bigg(1 + \frac{2}{c} - \big(1+o(1)\big)\frac{1-\alpha}{\alpha cm/2}\bigg). \end{aligned}$$

□

Lemma 13 (Lemma 19 of Wang and Zhang (2013))**.** *Given m and k, we let $\mathbf{B}$ be an $\frac{m}{k} \times \frac{m}{k}$ matrix whose diagonal entries equal to one and off-diagonal entries equal to $\alpha \in [0,1)$. We let $\mathbf{A}$ be an $m \times m$ block-diagonal matrix*

$$\mathbf{A} = \mathsf{diag}(\underbrace{\mathbf{B}, \cdots, \mathbf{B}}_{k \text{ blocks}}). \tag{20}$$

Let $\mathbf{A}_k$ be the best rank-k approximation to the matrix $\mathbf{A}$, then we have that

$$\|\mathbf{A} - \mathbf{A}_k\|_F = (1-\alpha)\sqrt{m-k}.$$

D.2 Proof of the Theorem

Now we prove Theorem 6 using Lemma 12 and Lemma 13.

Proof. Let $\mathbf{C}$ consist of c column sampled from $\mathbf{A}$ and $\hat{\mathbf{C}}_i$ consist of c_i columns sampled from the i-th block diagonal matrix in $\mathbf{A}$. Without loss of generality, we assume $\hat{\mathbf{C}}_i$ consists of the first c_i columns of $\mathbf{B}$. Then the intersection matrix $\mathbf{U}$ is computed by

$$\begin{aligned}\mathbf{U} &= \mathbf{C}^\dagger \mathbf{A} \big(\mathbf{C}^T\big)^\dagger \\ &= \big[\mathsf{diag}\big(\hat{\mathbf{C}}_1, \cdots, \hat{\mathbf{C}}_k\big)\big]^\dagger \mathbf{A} \big[\mathsf{diag}\big(\hat{\mathbf{C}}_1^T, \cdots, \hat{\mathbf{C}}_k^T\big)\big]^\dagger \\ &= \mathsf{diag}\Big(\hat{\mathbf{C}}_1^\dagger \mathbf{B} \big(\hat{\mathbf{C}}_1^\dagger\big)^T, \cdots, \hat{\mathbf{C}}_k^\dagger \mathbf{B} \big(\hat{\mathbf{C}}_k^\dagger\big)^T\Big).\end{aligned}$$

The modified Nyström approximation to $\mathbf{A}$ is

$$\begin{aligned}\tilde{\mathbf{A}}_c^{\mathrm{mod}} &= \mathbf{C}\mathbf{U}\mathbf{C}^T \\ &= \mathsf{diag}\Big(\hat{\mathbf{C}}_1 \hat{\mathbf{C}}_1^\dagger \mathbf{B} \big(\hat{\mathbf{C}}_1^\dagger\big)^T \hat{\mathbf{C}}_1^T, \cdots, \hat{\mathbf{C}}_k \hat{\mathbf{C}}_k^\dagger \mathbf{B} \big(\hat{\mathbf{C}}_k^\dagger\big)^T \hat{\mathbf{C}}_k^T\Big),\end{aligned}$$

and thus the approximation error is

$$\begin{aligned}\big\|\mathbf{A} - \tilde{\mathbf{A}}_c^{\mathrm{mod}}\big\|_F^2 &= \sum_{i=1}^k \Big\|\mathbf{B} - \hat{\mathbf{C}}_i \hat{\mathbf{C}}_i^\dagger \mathbf{B} \big(\hat{\mathbf{C}}_i^\dagger\big)^T \hat{\mathbf{C}}_i^T\Big\|_F^2 \\ &\geq (1-\alpha)^2 \sum_{i=1}^k (p - c_i)\Big(1 + \frac{2}{c_i} - (1-\alpha)\Big(\frac{1+o(1)}{\alpha c_i p/2}\Big)\Big) \\ &= (1-\alpha)^2 \Big(\sum_{i=1}^k (p - c_i) \\ &\qquad + \sum_{i=1}^k \frac{2(p-c_i)}{c_i}\Big(1 - \frac{(1-\alpha)(1+o(1))}{\alpha p}\Big)\Big) \\ &\geq (1-\alpha)^2 (m-c)\Big(1 + \frac{2k}{c}\Big(1 - \frac{k(1-\alpha)(1+o(1))}{\alpha m}\Big)\Big),\end{aligned}$$

where the former inequality follows from Lemma 12, and the latter inequality follows by minimizing over $c_1, \cdots, c_k$. Finally we apply Lemma 13, and the theorem follows by setting $\alpha \to 1$. □

E Supplementary Experiments

We have mentioned in Remark 1 that the resulting approximation accuracy is insensitive to the parameter μ in Algorithm 1, and setting μ to be exactly the matrix coherence does not in general give rise to the highest accuracy. To demonstrate this point of view, we conduct experiments on an RBF kernel matrix of the Letters Dataset with $\sigma = 0.2$, and we set $k = 10$.

We compare the uniform+adaptive^2 algorithm with different settings of μ; we also employ the adaptive-full algorithm of Kumar et al. (2012), the near-optimal+adaptive algorithm of Wang and Zhang (2013), and the uniform sampling algorithm for comparison. The experiment settings are the same to Section 6. Here the adaptive-full algorithm also has three steps: one uniform sampling and two adaptive sampling steps, and we set $c_1 = c_2 = c_3 = c/3$ according to Kumar et al. (2012). We plot the approximation errors in Figure 5.

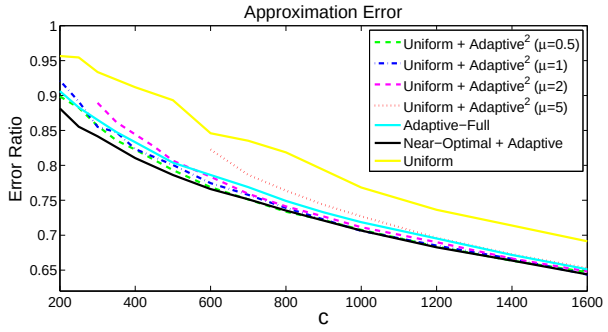

Figure 5: Effect of the parameter μ in Algorithm 1.

We can see from Figure 5 that different settings of μ does not have big influence on the approximation accuracy. We can also see that it is unnecessary to set μ to be exactly the matrix coherence; in this set of experiments, the uniform+adaptive^2 algorithm achieves the higher accuracy when $\mu = 0.5$ (the actual matrix coherence is $\mu_{10} = 62.05$).