6. Appendix

6.1. Proof of Theorem 1

Proof. We use $\boldsymbol{u}$ to denote the unit vector in the direction of $\boldsymbol{x}_i - \boldsymbol{x}_j$. By the mean value theorem, we have

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = g_{\boldsymbol{u}}(\eta\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2) = g_{\boldsymbol{u}}(0) + \eta g'_{\boldsymbol{u}}(s)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$$

for some $s \in (0, \eta\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2)$. By definition, $f(\mathbf{0}) = g_{\boldsymbol{u}}(0)$, so

$$f(\mathbf{0}) \le K(\boldsymbol{x}_i, \boldsymbol{x}_j) + \eta R\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2, \quad (6)$$

$$\text{where } R := \sup_{\theta\in\mathbb{R}, \|\boldsymbol{v}\|=1} |g'_{\boldsymbol{v}}(\theta)|. \quad (7)$$

Squaring both sides of (6) we have

$$f(\mathbf{0})^2 \le K(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 + \eta^2 R^2\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 + 2K(\boldsymbol{x}_i, \boldsymbol{x}_j)(\eta R\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2).$$

From the classical arithmetic and geometric mean inequality, we can upper bound the last term by

$$2K(\boldsymbol{x}_i, \boldsymbol{x}_j)(\eta R\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2) \le \frac{1}{2} f(\mathbf{0})^2,$$

therefore

$$\frac{f(\mathbf{0})^2}{2} \le K(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 + \eta^2 R^2\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2. \quad (8)$$

Plugging in (8) into $\mathcal{D}^{\text{kernel}}(\{\mathcal{V}_s\}_{s=1}^c)$, we have

$$\begin{aligned}
&\mathcal{D}^{\text{kernel}}(\{\mathcal{V}_s\}_{s=1}^c) \\
&\ge \sum_{s=1}^{c} \frac{1}{|\mathcal{V}_s|} \sum_{i,j\in\mathcal{V}_s} \left(\frac{f(\mathbf{0})^2}{2} - \eta^2 R^2\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 \right) \\
&\ge \frac{nf(\mathbf{0})^2}{2} - \eta^2 R^2 \sum_{s=1}^{c} \frac{1}{|\mathcal{V}_s|} \sum_{i,j\in\mathcal{V}_s} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2,
\end{aligned}$$

which proves the desired bound (3). □

6.2. Proof of Theorem 2

Proof. To prove this theorem, we use the ϵ-net theorem in (Cucker & Smale, 2001). This theorem shows that when $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$ are in a ball with radius r, there exists $T = (\frac{4r}{\bar{r}})^d$ balls of radius $\bar{r}$ that cover all the data points X. Now we set T to be k, so $\bar{r} = k^{-1/d}4r$.

We consider $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_s}\}$ are data points in the s-th cluster, $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{n_y}\}$ are data points in the t-th cluster, and $n_s = |\mathcal{V}_s|, n_t = |\mathcal{V}_t|$. Our goal is to show that $G^{(s,t)}$ is low-rank, where $G^{(s,t)}_{i,j} = K(\boldsymbol{x}_i, \boldsymbol{y}_j)$. Assume r_t is the radius of the t-th cluster, therefore we can find k balls with $\bar{r} = k^{-1/d}4r_t$ to cover $\{\boldsymbol{y}_j\}_{j=1}^{n_t}$.

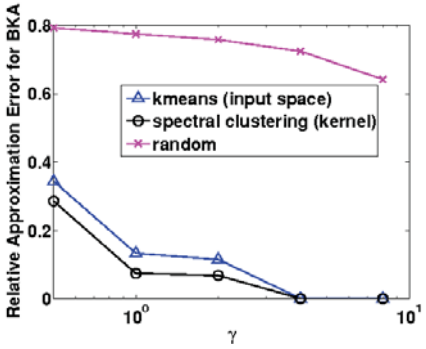


Figure 5. The Gaussian kernel approximation error of BKA using different ways to generate five partitions on 500 samples from covtype. k-means in the input space performs similar to spectral clustering on kernel matrix but is much more efficient.

Assume centers of the balls are $\{\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_k\}$, then we can form a low-rank matrix $\bar{G}^{(s,t)} = \bar{U}\bar{V}^T$, where for all $i = 1, \ldots, n_s, j = 1, \ldots, n_t$, and $s = 1, \ldots, k$,

$$\bar{U}_{i,s} = K(\boldsymbol{x}_i, \boldsymbol{m}_s) \text{ and } \bar{V}_{j,s} = \begin{cases} 1 & \text{if } \boldsymbol{y}_j \in Ball(\boldsymbol{m}_s), \\ 0 & \text{otherwise.} \end{cases}$$

Assume $\boldsymbol{y}_j$ is in ball s, then

$$\begin{aligned}
(G^{(s,t)}_{ij} - \bar{G}^{(s,t)}_{ij})^2 &= (f(\boldsymbol{x}_i - \boldsymbol{y}_j) - f(\boldsymbol{x}_i - \boldsymbol{m}_s))^2 \\
&\le C^2\|(\boldsymbol{x}_i - \boldsymbol{y}_j) - (\boldsymbol{x}_i - \boldsymbol{m}_s)\|^2 \\
&= C^2\|\boldsymbol{y}_j - \boldsymbol{m}_s\|_2^2 \\
&\le C^2\bar{r}^2.
\end{aligned}$$

Therefore, if $(G^{(s,t)})^*$ is the best rank k approximation for $G^{(s,t)}$, then

$$\|G^{(s,t)} - (G^{(s,t)})^*\|_F \le \|G^{(s,t)} - \bar{G}^{(s,t)}\|_F \le Ck^{-1/d}4r_t\sqrt{n_s n_t}. \quad (9)$$

Similarly, by dividing $\{\boldsymbol{x}_i\}_{i=1}^{m_1}$ to k balls we can get the following inequality:

$$\|G^{(s,t)} - (G^{(s,t)})^*\|_F \le Ck^{-1/d}4r_s\sqrt{n_s n_t}. \quad (10)$$

Combining (9) and (10) we can prove Theorem 2.

□

6.3. Empirical observation on low rank structure after k-means clustering

Theorem 2 suggests that each block of the kernel matrix will be low rank if we find the partition by k-means in the input space. In the following we show some empirical justification. We present the numerical rank for each block, where numerical rank for a m

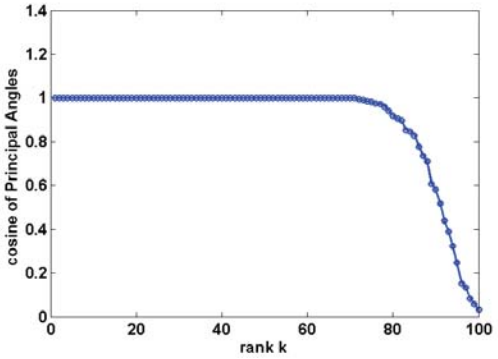


Figure 6. The cosine of the principal angles between the basis of diagonal and off-diagonal blocks of a Gaussian kernel($\gamma = 1$ and 1000 random samples from covtype) with respect to different ranks. The cosine values of the principal angles are close to 1 showing that two basis are similar.

by n matrix A is defined as the number of singular values with magnitude larger than $\max(n, m)\|A\|_2\delta$. We sample 4000 data points from the ijcnn1 dataset and generate 5 clusters by k-means and random partition. Table 5 shows the numerical rank for each block using k-means, while Table 6 shows the numerical rank for each block when the partitions are random. We observe that by using k-means, the rank for each block is fairly small.

16	14	13	7	7
14	29	13	9	9
13	13	20	10	10
7	9	10	29	11
7	9	10	11	28

Table 5. Rank of each block(from a subsampled ijcnn1 data set) using k-means clustering.

139	99	101	44	45
99	116	86	43	44
101	86	131	46	47
44	43	46	47	45
45	44	47	45	49

Table 6. Rank of each block(from a subsampled ijcnn1 data set) using random partition.

6.4. The principal angles between the basis of diagonal and off-diagonal blocks

In MEKA, we use the diagonal blocks' basis to approximate the off-diagonal blocks' basis to reduce memory requirements. Furthermore, we observe that the principal angles between the basis of diagonal and off-diagonal blocks are small, which provides empirical justification to reuse the basis. In Figure 6, we randomly sampled 1000 data points from the covtype dataset and generated 5 clusters by k-means. The blue line shows the cosine values of the principal angles between a basis of a diagonal block $G^{(s,s)}$ and that of an off-diagonal block $G^{(s,t)}$ for different rank k, where s and t are randomly chosen. We can observe that most of the cosine values are close to 1, showing that the two basis are similar.

6.5. The comparison of MEKA and Nys with CSI

Figure 7 compares our proposed method with the standard Nyström(Nys), and incomplete Cholesky with side information (CSI) for approximating the Gaussian kernel on the wine and cpusmall datasets. All the settings are the same with Table 3. We observe that both MEKA and Nys are much faster than CSI for kernel regression.

6.6. Influence of ϵ in MEKA

We test the influence of thresholding parameter ϵ on the ijcnn1 data (Figure 8). Recall that we set $L^{(s,t)} = 0$ if $K(\boldsymbol{m}_s, \boldsymbol{m}_t) \leq \epsilon$. For large ϵ, we will set more off-diagonal blocks in L to be 0. In this case, although MEKA yields higher approximation error(because it omits more off-diagonal information), it is faster. On the other hand, for small ϵ, when more off-diagonal information is considered, we will notice an increase in time and decrease in approximation error. In the rest of our experiments, we set ϵ to be 0.1.

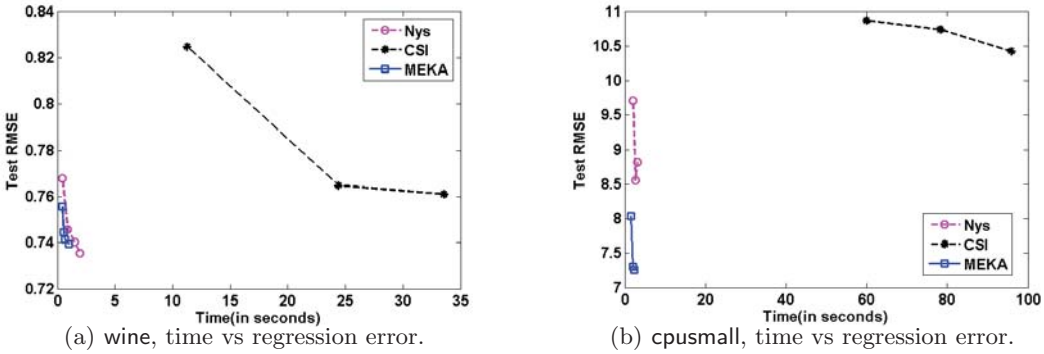
6.7. Time cost for each step in MEKA

In Figure 9, we show the time cost for each step of MEKA on ijcnn1 dataset. Here the parameter settings are $\gamma = 1$ and $k = 100$.

The execution time of our proposed algorithm mainly consists of three parts:(1) time for performing k-means clustering, T_C; (2) time for forming the "basis", W from the diagonal blocks, T_W (3) Time to compute the link matrix L from off-diagonal blocks, T_L. From Figure 9, we observe that when the number of clusters c is small, T_W will dominant the whole process. As c increases, the time for computing the link matrix L, T_L, increases. This is because the number of off-diagonal blocks increases quadratically as c increases. Since the time complexity for k-means is $O(ncd)$, T_C will increase as c increases.

6.8. Proof of Theorem 3

Proof. Let B denote the matrix formed by the diagonal block of G, that is, $B = G^{(1)} \oplus G^{(2)} \oplus \cdots \oplus G^{(c)}$. According to the definition of Δ, $G = B + \Delta$. In MEKA,



(a) wine, time vs regression error. (b) cpusmall, time vs regression error.

Figure 7. Kernel ridge regression results on wine and cpusmall datasets for Nys, CSI, and MEKA. Methods with regression error above the top of y-axis are not shown in the figures.

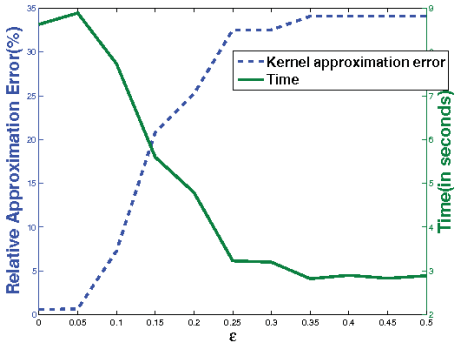


Figure 8. Time cost (in seconds) and kernel approximation quality of MEKA when varying the thresholding parameter ϵ for setting off-diagonal blocks in L to be zero.

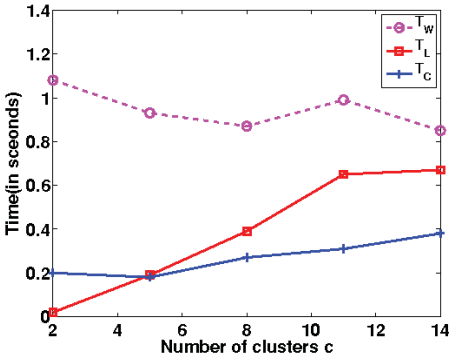


Figure 9. Time cost (in seconds) for performing each step of MEKA when varying the number of clusters c.

the error $\|\tilde{G} - G\|_2$ consists of two components,

$$\|\tilde{G} - G\|_2 = \|\tilde{B} - B + (\tilde{\Delta} - \Delta)\| \leq \|\tilde{B} - B\| + \|\tilde{\Delta} - \Delta\| \tag{11}$$

where $\tilde{B}$ and $\tilde{\Delta}$ are the approximations for B and Δ in MEKA respectively.

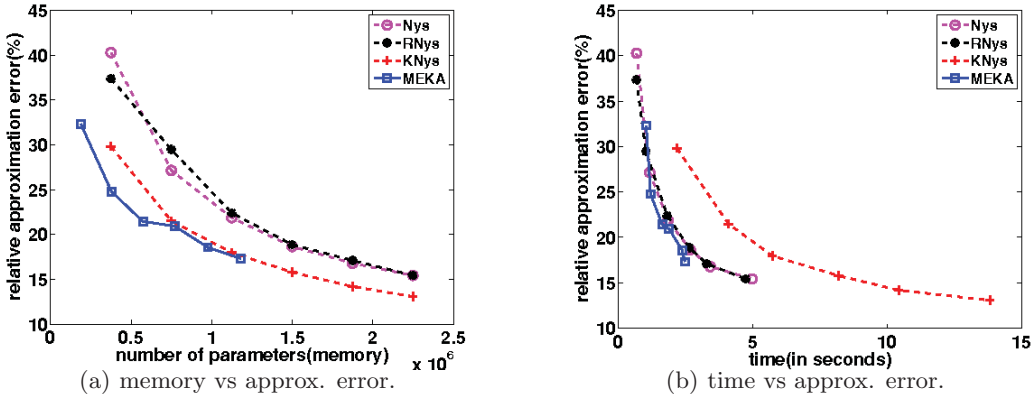
Let us first consider the error in approximating the diagonal blocks $\|\tilde{B} - B\|_2$. Since we sample cm benchmark points from n data points uniformly at random without replacement and distribute them according to the partition coming from k-means, the s-th cluster now has m_s benchmark points with $\sum_{s=1}^{s=c} m_s = cm$. For the s-th diagonal block $G^{(s)}$, we will perform the rank-k_s approximation using standard Nyström, so we have $G^{(s)} \approx E^{(s)}(M_{k_s}^{(s)})^+ E^{(s)}$, where $E^{(s)}$ denotes the matrix formed by m_s sampled columns from $G^{(s)}$ and $M_{k_s}^{(s)}$ is a $m_s \times m_s$ matrix consisting of the intersection of sampled m_s columns.

Suppose we use the singular value based approach to choose k_s for s-th cluster as described in Section 4.2, and $M_{ck}^+ = (M_{k_1}^{(1)})^+ \oplus (M_{k_2}^{(2)})^+ \oplus \cdots \oplus (M_{k_c}^{(c)})^+$, where M is the $cm \times cm$ block diagonal matrix that consists of the intersection of the sampled cm columns. Then we can see that approximating the diagonal blocks B is equivalent to directly performing standard Nyström on B by sampling cm benchmark points uniformly at random without replacement to achieve the rank-ck approximation. The standard Nyström's norm-2 and Frobenius error bound are given in (Kumar et al., 2009), so $\|B - \tilde{B}\|_2$ can be bounded with probability at least $1 - \delta$ as

$$\|B - \tilde{B}\|_2 \leq \|B - B_{ck}\|_2 + \tag{12}$$
$$\frac{2n}{\sqrt{cm}} B_{max}\left[1 + \sqrt{\frac{n - cm}{n - 0.5}\frac{1}{\beta(cm, n)} log\frac{1}{\delta}} d_{max}^B / B_{max}^{\frac{1}{2}}\right],$$

where B_{ck} denotes the best rank-ck approximation to B; $B_{max} = max_i B_{ii}$; d_{max}^B represents the distance $max_{ij}\sqrt{B_{ii} + B_{jj} - 2B_{ij}}$.

To bound $\|\tilde{\Delta} - \Delta\|_2$, recall that some off-diagonal blocks in MEKA are set to 0 by thresholding and $\mathbf{0}$ is



(a) memory vs approx. error. (b) time vs approx. error.

Figure 10. Low-rank Laplacian kernel approximation results for pendigit.

one special solution of least squares problem to compute $L^{(s,t)}$, we have $\|\tilde{\Delta} - \Delta\|_2 \leq \|\Delta\|_2$.

Furthermore, according to perturbation theory (Stewart & Ji-Guang, 1990), we have

$$\|B - B_{ck}\|_2 \leq \|G - G_{ck}\|_2 + \|\Delta\|_2. \tag{13}$$

The inequality in (12) combined with (13) gives a bound on $\|\tilde{G} - G\|_2$ as,

$$\begin{aligned}
&\|\tilde{G} - G\|_2 \\
&\leq \|B - B_{ck}\|_2 + \|\Delta\|_2 + \\
&\quad \frac{2n}{\sqrt{cm}} B_{max}[1 + \sqrt{\frac{n-cm}{n-0.5}\frac{1}{\beta(cm,n)}log\frac{1}{\delta}}d^B_{max}/B^{\frac{1}{2}}_{max}] \\
&\leq \|G - G_{ck}\|_2 + 2\|\Delta\|_2 + \\
&\quad \frac{2n}{\sqrt{cm}} B_{max}[1 + \sqrt{\frac{n-cm}{n-0.5}\frac{1}{\beta(cm,n)}log\frac{1}{\delta}}d^B_{max}/B^{\frac{1}{2}}_{max}] \\
&\leq \|G - G_{ck}\|_2 + 2\|\Delta\|_2 + \\
&\quad \frac{2n}{\sqrt{cm}} G_{max}[1 + \sqrt{\frac{n-cm}{n-0.5}\frac{1}{\beta(cm,n)}log\frac{1}{\delta}}d^G_{max}/G^{\frac{1}{2}}_{max}] \\
&\leq \|G - G_{ck}\|_2 + 2\|\Delta\|_2 + \\
&\quad \frac{1}{\sqrt{c}}\frac{2n}{\sqrt{m}} G_{max}[1 + \sqrt{\frac{n-m}{n-0.5}\frac{1}{\beta(m,n)}log\frac{1}{\delta}}d^G_{max}/G^{\frac{1}{2}}_{max}],
\end{aligned}$$

where G_{ck} denotes the best rank-ck approximation to G; $G_{max} = max_i G_{ii}$; d^G_{max} represents the distance $max_{ij}\sqrt{G_{ii} + G_{jj} - 2G_{ij}}$. The third inequality is because $G = B + \Delta$, $B_{max} \leq G_{max}$ and $d^B_{max} \leq d^G_{max}$. The last inequality is because $n \gg m$ and $n \gg cm$.

Similarly by using perturbation theory and upper bounds for the Frobenius error of standard Nyström, the result follows. □

6.9. The performance of MEKA on Laplacian kernel on pendigit dataset

Figure 10 compares our proposed method with the standard Nyström(Nys), Randomized Nyström(RNys), and Kmeans Nyström(KNys) for approximating the Laplacian kernel on the pendigit data, where $c = 3$ and $\gamma = 2^{-7}$. Similar to Gaussian kernel, we observe that MEKA is more memory efficient and faster than other methods for approximating the Laplacian kernel.