APPENDIX

A High Probability Bound for RFFMaps

Here we extend the analysis of Lopez *et al.* [19] to show that the Fourier Random Features of Rahimi and Recht [22] approximate the spectral error with their approximate Gram matrix within εn with high probability.

A.1 High Probability Bound "for all" Bound for RFFMaps

In our proof we use the Bernstein inequality on sum of zero-mean random matrices.

Matrix Bernstein Inequality: Let $X_1, \cdots, X_d \in \mathbb{R}^{n\times n}$ be independent random matrices such that for all $1 \le i \le d$, $E[X_i] = 0$ and $\|X_i\|_2 \le R$ for a fixed constant R. Define variance parameter as $\sigma^2 = \max\{\|\sum_{i=1}^d \mathbf{E}[X_i^T X_i]\|, \|\sum_{i=1}^d \mathbf{E}[X_i X_i^T]\|\}$. Then for all $t \ge 0$, $\mathbf{Pr}\left[\left\|\sum_{i=1}^d X_i\right\|_2 \ge t\right] \le 2n \cdot \exp\left(\frac{-t^2}{3\sigma^2+2Rt}\right)$. Using this inequality, [19] bounded $\mathbf{E}[\|G-\hat{G}\|_2]$. Here we employ similar ideas to improve this to a bound on $\|G-\hat{G}\|_2$ with high probability.

Lemma A.1. *For n points, let $G = \Phi\Phi^T \in \mathbb{R}^{n\times n}$ be the exact gram matrix, and let $\hat{G} = ZZ^T \in \mathbb{R}^{n\times n}$ be the approximate kernel matrix using $m = O((1/\varepsilon^2)\log(n/\delta))$ RFFMaps. Then $\|G-\hat{G}\| \le \varepsilon n$ with probability at least $1-\delta$.*

Proof. Consider m independent random variables $E_i = \frac{1}{m}G - z_i z_i^T$. Note that $\mathbf{E}[E_i] = \frac{1}{m}G - \mathbf{E}[z_i z_i^T] = 0^{n\times n}$ [22]. Next we can rewrite

$$\|E_i\|_2 = \left\|\frac{1}{m}G - z_i z_i^T\right\|_2 = \left\|\frac{1}{m}\mathbf{E}[ZZ^T] - z_i z_i^T\right\|_2$$

and thus bound

$$\|E_i\|_2 \le \frac{1}{m}\|\mathbf{E}[ZZ^T]\|_2 + \|z_i z_i^T\|_2 \le \frac{1}{m}\mathbf{E}[\|Z\|_2^2] + \|z_i\|^2 \le \frac{2n}{m} + \frac{2n}{m} = \frac{4n}{m}$$

The first inequality is correct because of triangle inequality, and second inequality is achieved using Jensen's inequality on expected values, which states $\|\mathbf{E}[X]\| \le \mathbf{E}[\|X\|]$ for any random variable X. Last inequality uses the bound on the norm of z_i as $\|z_i\|^2 \le \frac{2n}{m}$, and therefore $\|Z\|_2^2 \le \|Z\|_F^2 \le 2n$.

To bound σ^2, due to symmetry of matrices E_i, simply $\sigma^2 = \|\sum_{i=1}^m \mathbf{E}[E_i^2]\|_2$. Expanding

$$\mathbf{E}[E_i^2] = \mathbf{E}\left[\left(\frac{1}{m}G - z_i z_i^T\right)^2\right] = \mathbf{E}\left[\frac{G^2}{m^2} + \|z_i\|^2 z_i z_i^T - \frac{1}{m}(z_i z_i^T G + G z_i z_i^T)\right]$$

it follows that

$$\mathbf{E}[E_i^2] \le \frac{G^2}{m^2} + \frac{2n}{m}\mathbf{E}[z_i z_i^T] - \frac{1}{m}(\mathbf{E}[z_i z_i^T]G + G\,\mathbf{E}[z_i z_i^T]) = \frac{1}{m^2}(G^2 + 2nG - 2G^2) = \frac{1}{m^2}(2nG - G^2)$$

The first inequality holds by $\|z_i\|^2 \le 2n/m$, and second inequality is due to $\mathbf{E}[z_i z_i^T] = \frac{1}{m}G$. Therefore

$$\sigma^2 = \left\|\sum_{i=1}^m \mathbf{E}[E_i^2]\right\|_2 \le \left\|\frac{1}{m}(2n\,G - G^2)\right\|_2 \le \frac{2n}{m}\|G\|_2 + \frac{1}{m}\|G^2\|_2 \le \frac{2n^2}{m} + \frac{1}{m}\|G\|_2^2 \le \frac{3n^2}{m}$$

the second inequality is by triangle inequality, and the last inequality by $\|G\|_2 \le \mathsf{Tr}(G) = n$. Setting $M = \sum_{i=1}^m E_i = \sum_{i=1}^m (\frac{1}{m}G - z_{:,i} z_{:,i}^T) = G - \hat{G}$ and using Bernstein inequality with $t = \varepsilon n$ we obtain

$$\mathbf{Pr}\left[\|G-\hat{G}\|_2 \ge \varepsilon n\right] \le 2n\exp\left(\frac{-(\varepsilon n)^2}{3(\frac{3n^2}{m}) + 2(\frac{4n}{m})\varepsilon n}\right) = 2n\exp\left(\frac{-\varepsilon^2 m}{9+8\varepsilon}\right) \le \delta$$

Solving for m we get $m \ge \frac{9+8\varepsilon}{\varepsilon^2}\log(2n/\delta)$, so with probability at least $1-\delta$ for $m = O(\frac{1}{\varepsilon^2}\log(n/\delta))$, then $\|G-\hat{G}\|_2 \le \varepsilon n$. □

A.2 For Each Bound for RFFMaps

Here we bound $\left|\|\Phi^T x\|^2 - \|Z^T x\|^2\right|$, where Φ and Z are mappings of data to RKHS by RFFMaps, respectively and x is a *fixed* unit vector in $\mathbb{R}^n$.

Note that Lemma A.1 essentially already gave a stronger proof, where using $m = O((1/\varepsilon^2)\log(n/\delta))$ the bound $\|G-\hat{G}\|_2 \le \varepsilon n$ holds along all directions (which makes progress towards addressing an open question of constructing oblivious subspace embeddings for Gaussian kernel features spaces, in [1]). The advantage of this proof is that the bound on m will be independent of n. Unfortunately, in this proof, going from the "for each" bound to the stronger "for all" bound would seem to require a net of size $2^{O(n)}$ and a union bound resulting in a worse "for all" bound with $m = O(n/\varepsilon^2)$.

On the other hand, main objective of Test time procedure, which is mapping a single data point to the

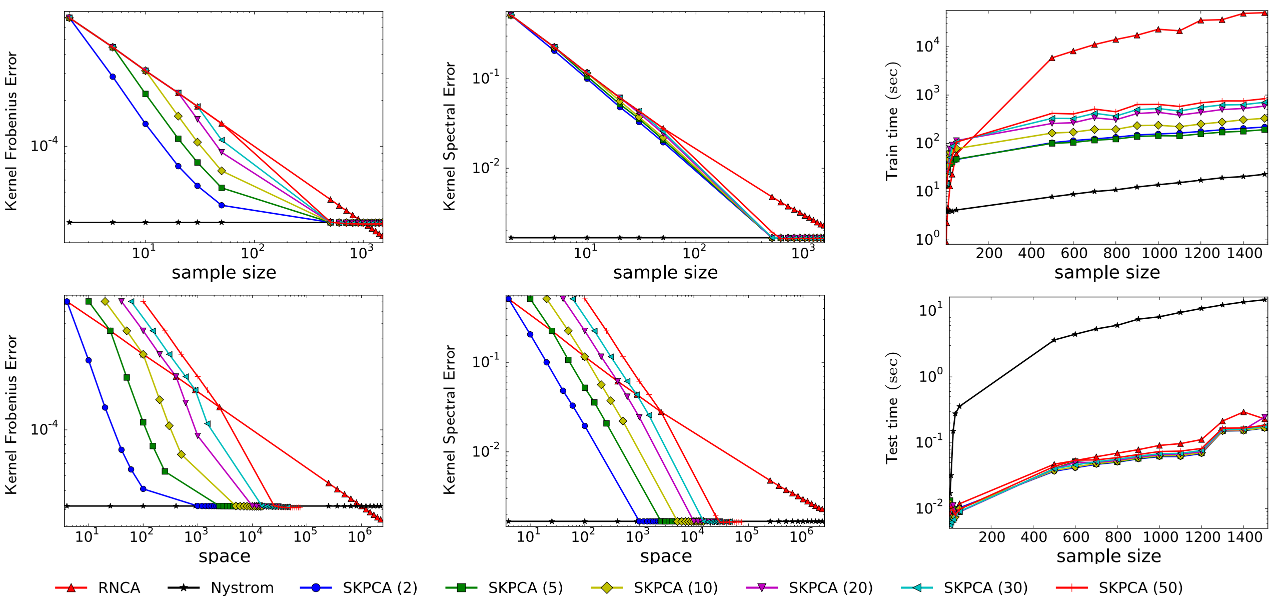

Figure 4: Results for FOREST dataset. Row 1: Kernel Frobenius Error (left), Kernel Spectral Error (middle) and TRAIN TIME (right) vs. SAMPLE SIZE. Row 2: Kernel Frobenius Error (left), Kernel Spectral Error (middle) vs. SPACE, and TEST TIME vs. SAMPLE SIZE (right)

D-dimensional or k-dimensional kernel space is already interesting for what the error is expected to be for a single vector x. This scenario corresponds to the "for each" setting that we will prove in this section.

In our proof, we use a variant of Chernoff-Hoeffding inequality, stated next. Consider a set of r independent random variables $\{X_1, \cdots, X_r\}$ where $0 \leq X_i \leq \Delta$. Let $M = \sum_{i=1}^r X_i$, then for any $\alpha \in (0, 1/2)$, $\textbf{Pr}\left[|M - \textbf{E}[M]| > \alpha\right] \leq 2 \exp\left(\frac{-2\alpha^2}{r\Delta^2}\right)$.

For this proof we are more careful with notation about rows and column vectors. Now matrix $Z \in \mathbb{R}^{n \times m}$ can be written as a set rows $[z_{1,:}; z_{2,:}; \ldots, z_{n,:}]$ where each $z_{i,:}$ is a vector of length m or a set of columns $[z_{:,1}, z_{:,2}, \ldots, z_{:,d}]$, where each $z_{:,j}$ is a vector of length n. We denote the (i,j)-th entry of this matrix as $z_{i,j}$.

Theorem A.1. *For n points in any arbitrary dimension and a shift-invariant kernel, let $G = \Phi\Phi^T \in \mathbb{R}^{n \times n}$ be the exact gram matrix, and $\hat{G} = ZZ^T \in \mathbb{R}^{n \times n}$ be the approximate kernel matrix using $m = O((1/\varepsilon^2)\log(1/\delta))$ RFFMAPS. Then for any fixed unit vector $x \in \mathbb{R}^n$, it holds that $\left|\|\Phi^T x\|^2 - \|Z^T x\|^2\right| \leq \varepsilon n$ with probability at least $1-\delta$.*

Proof. Note $\mathbb{R}^n$ is not the dimension of data. Consider any unit vector $x \in \mathbb{R}^n$. Define m independent random variables $\{X_i = \langle z_{:,i}, x\rangle^2\}_{i=1}^m$. We can bound each X_i as $0 \leq X_i \leq \|z_{:,i}\|^2 \leq 2n/m$ therefore $\Delta = 2n/m$ for all X_is. Setting $M = \sum_{i=1}^m X_i = \|Z^T x\|^2$, we observe

$$\begin{aligned}
\textbf{E}[M] &= \sum_{i=1}^m \textbf{E}\left[\langle z_{:,i}, x\rangle^2\right] = \sum_{i=1}^m \textbf{E}\left[\left(\sum_{j=1}^n z_{ji}\, x_j\right)^2\right] \\
&= \sum_{i=1}^m \textbf{E}\left[\sum_{j=1}^n (z_{ji}\, x_j)^2 + 2\sum_{j=1}^n \sum_{k>j}^n z_{ji}\, z_{ki}\, x_j\, x_k\right] \\
&= \sum_{j=1}^n x_j^2\, \textbf{E}\left[\sum_{i=1}^m z_{ji}^2\right] + 2\sum_{j=1}^n \sum_{k>j}^n x_j\, x_k\, \textbf{E}\left[\sum_{i=1}^m z_{ji}\, z_{ki}\right] \\
&= \sum_{j=1}^n x_j^2\, \textbf{E}\left[\langle z_{j,:}, z_{j,:}\rangle\right] + 2\sum_{j=1}^n \sum_{k>j}^n x_j\, x_k\, \textbf{E}\left[\langle z_{j,:}, z_{k,:}\rangle\right] \\
&= \sum_{j=1}^n x_j^2\, \langle \phi_{j,:}\,, \phi_{j,:}\rangle + 2\sum_{j=1}^n \sum_{k>j}^n x_j\, x_k\, \langle \phi_{j,:}, \phi_{k,:}\rangle \\
&= \sum_{j=1}^n x_j^2 \sum_{i=1}^D \phi_{ji}^2 + 2\sum_{j=1}^n \sum_{k>j}^n x_j\, x_k \sum_{i=1}^D \phi_{ji}\, \phi_{ki} \\
&= \sum_{i=1}^D \left(\sum_{j=1}^n x_j^2\, \phi_{ji}^2 + 2\sum_{j=1}^n \sum_{k>j}^n x_j\, x_k\, \phi_{ji}\, \phi_{ki}\right) \\
&= \sum_{i=1}^D \langle \phi_{:,i}, x\rangle^2 = \|\Phi^T x\|^2
\end{aligned}$$

Since x is a fixed unit vector, it is pulled out of all expectations. Using the Chernoff-Hoeffding bound and setting $\alpha = \varepsilon n$ yields $\textbf{Pr}\left[\left|\|\Phi^T x\|^2 - \|Z^T x\|^2\right| > \varepsilon n\right] \leq 2\exp\left(\frac{-2(\varepsilon n)^2}{m(2n/m)^2}\right) = 2\exp\left(-2\varepsilon^2 m\right) \leq \delta$. Then we solve for $m = (1/(2\varepsilon^2))\ln(2/\delta)$ in the last inequality. ∎