Appendix

Roadmap In Section A, we give an algorithm to compute a subspace embedding for the Gaussian kernel using Theorem 4.8. In Section B, we characterize a large class of kernels based on the coefficients in their Taylor expansion, and develop fast algorithms for different scenarios. In Section C, we apply our results in Section B to the Neural Tangent kernel. In Section D, we use our sketch in conjunction with another sketch to compute a good preconditioner for the Gaussian kernel. In Section E, we compose our sketch with our sketching matrices to solve Kernel Ridge Regression.

Notation We use $\widetilde{O}(f)$ to denote $f \operatorname{poly}(\log f)$ and use $\widetilde{\Omega}(f)$ to denote $f/\operatorname{poly}(\log f)$.

For an integer n, let $[n]$ denote the set $\{1, 2, \cdots, n\}$. For two scalars a and b, we say $a \approx_\epsilon b$ if $(1-\epsilon)b \leq a \leq (1+\epsilon)b$. We say a square matrix A is positive semidefinite (PSD) if $\forall x$, $x^\top Ax \geq 0$. For two PSD matrices A and B, we define $A \approx_\epsilon B$ if $(1-\epsilon)B \preceq A \preceq (1+\epsilon)B$. For a matrix A, we use $\|A\|_F = (\sum_{i,j} A_{i,j}^2)^{1/2}$ to denote its Frobenius norm and use $\|A\|_{\mathrm{op}}$ to denote its operator (spectral) norm. For a square matrix A, we use $\mathrm{tr}[A]$ to denote the trace of A. For a square matrix A, we use $\lambda_{\min}(A)$, $\lambda_{\max}(A)$ to denote its smallest and largest eigenvalues, respectively. For a rectangular matrix A, we use $\sigma_{\min}(A), \sigma_{\max}(A)$ to denote its smallest and largest singular values.

A. Gaussian Kernel

We apply Algorithm 1 to compute a subspace embedding to the Gaussian kernel matrix $G \in \mathbb{R}^{n\times n}$ defined over n data points of dimension d, denoted by $X \in \mathbb{R}^{d\times n}$. Our method has the advantage that when d is large and the matrix X is dense, its leading factor depends nearly linearly on nd, which makes it useful for certain biological and NLP tasks.

We remark that our construction of a sketch for the Gaussian kernel and its corresponding analysis is inspired by (Ahle et al., 2020), and is thus similar to the proof of Theorem 5 in their paper. For completeness, we include a proof here.

Theorem A.1 (Gaussian Kernel, formal version of Theorem 6.1)**.** *Let $r \in \mathbb{R}_+$ and $X \in \mathbb{R}^{d\times n}$ be such that $\|x_i\|_2 \leq r$ for all $i \in [n]$, where x_i is the i-th column of X. Suppose $G \in \mathbb{R}^{n\times n}$ is the Gaussian kernel matrix given in Definition 2.14. For any accuracy parameter $\epsilon \in (0,1)$ and for any failure probability $\delta \in (0,1)$, there exists an algorithm running in time:*

$$O(\epsilon^{-2} n^2 q^3 \cdot \log^3(nd/\epsilon\delta) + nd \log(nd/\epsilon\delta))$$

and outputting a matrix $W_g(X) \in \mathbb{R}^{m\times n}$ such that

$$\Pr\left[W_g(X)^\top W_g(X) \approx_\epsilon G\right] \geq 1 - \delta$$

where $m = \Omega(\epsilon^{-2} n q^3 \log^3(nd/\epsilon\delta))$ and $q = \Theta(r^2 + \log(n/\epsilon))$.

Proof. By definition of the Gaussian kernel matrix $G_{i,j} = \exp(-\|x_i - x_j\|_2^2/2)$, we can rewrite it as $G = DKD$, where D is an $n \times n$ diagonal matrix with i^{th} diagonal entry equal to $\exp(-\|x_i\|_2^2/2)$ and $K \in \mathbb{R}^{n\times n}$ is a positive definite kernel matrix defined as $K_{i,j} = \exp(x_i^\top x_j)$. Note the Taylor series expansion for kernel K gives

$$K = \sum_{l=0}^{\infty} \frac{(X^{\otimes l})^\top X^{\otimes l}}{l!}.$$

Let $q = C \cdot (r^2 + \log(n/\epsilon))$ for a sufficiently large constant C, and let $Q = \sum_{l=0}^{q} \frac{(X^{\otimes l})^\top X^{\otimes l}}{l!}$ be the first q terms of K. Then by the triangle inequality we have:

$$\begin{aligned}
\|K - Q\|_{\mathrm{op}} &\leq \sum_{l>q} \left\| \frac{(X^{\otimes l})^\top X^{\otimes l}}{l!} \right\|_{\mathrm{op}} \\
&\leq \sum_{l>q} \left\| \frac{(X^{\otimes l})^\top X^{\otimes l}}{l!} \right\|_{F} \\
&\leq \sum_{l>q} \frac{n \cdot r^{2l}}{l!} \\
&\leq \epsilon/2.
\end{aligned}$$

Then Q is a positive definite kernel matrix and $\|D\|_{\text{op}} \leq 1$. Therefore, in order to get a subspace embedding for G it is sufficient to satisfy the following with probability $1-\delta$:

$$(1-\epsilon/2)\cdot DQD \preceq W_g(X)^\top W_g(X) \preceq (1+\epsilon/2)\cdot DQD.$$

For each term $(X^{\otimes l})^\top X^{\otimes l}$ in Q, we run Algorithm 1 to approximate $X^{\otimes l}$. Let $Z_l \in \mathbb{R}^{m_l \times n}$ be the resulting matrix $\mathcal{Z}(S,T,X)$, where

$$m_l = \Omega(\epsilon^{-2} n l^2 \cdot \log^2(nd/\epsilon\delta)\cdot \log(n/\delta)).$$

Then by Theorem 5.1, we get

$$(1-\epsilon/2)(X^{\otimes l}D)^\top X^{\otimes l} D \preceq (\Pi^l X^{\otimes l} D)^\top \Pi^l X^{\otimes l} D \preceq (1+\epsilon/2)(X^{\otimes l}D)^\top X^{\otimes l} D \tag{4}$$

with probability at least $1-\frac{\delta}{q+1}$. Moreover, Z_l can be computed in time

$$O(\epsilon^{-2} n^2 l^2 \cdot \log^2(nd/\epsilon\delta)\cdot \log(n/\delta)).$$

Our algorithm will simply compute Z_l from $l=0$ to q, normalize each Z_l by $\frac{1}{\sqrt{l!}}$, and then multiply by D. More precisely, the approximation $W_g(X)$ will be

$$W_g(X) = \Big(\bigoplus_{l=0}^{q} \frac{Z_l}{\sqrt{l!}}\Big) D$$

where we use $A \oplus B$ to denote the matrix $\begin{bmatrix} A \\ B \end{bmatrix}$ if A and B have the same number of columns. Notice $W_g(X) \in \mathbb{R}^{m\times n}$. The following holds for $W_g(X)^\top W_g(X)$:

$$\begin{aligned} W_g(X)^\top W_g(X) &= D\Big(\sum_{l=0}^{q} \frac{Z_l^\top Z_l}{l!}\Big) D \\ &= \sum_{l=0}^{q} \frac{(Z_l D)^\top Z_l D}{l!}. \end{aligned}$$

By combining terms in (4) and using a union bound over all $0 \leq l \leq q$, we obtain that with probability at least $1-\delta$, we have the following:

$$(1-\epsilon/2)\cdot DQD \preceq W_g(X)^\top W_g(X) \preceq (1+\epsilon/2)\cdot DQD.$$

Thus, we conclude that

$$(1-\epsilon)\cdot G \preceq W_g(X)^\top W_g(X) \preceq (1+\epsilon)\cdot G.$$

Note the target dimension of W_g is

$$\begin{aligned} m &= m_0 + m_1 + \cdots + m_q \\ &= \Omega(\epsilon^{-2} n q^3 \cdot \log^2(nd/\epsilon\delta)\cdot \log(n/\delta)). \end{aligned}$$

Also, by Theorem 5.1, the time to compute $W_g(X)$ is

$$\begin{aligned} t &= t_0 + t_1 + \cdots + t_q \\ &= O(\epsilon^{-2} n^2 q^3 \cdot \log^2(nd/\epsilon\delta)\cdot \log(n/\delta)). \end{aligned}$$

Notice we will have to pay an additive $nd\log(nd/\epsilon\delta)$ due to line 2 of Algorithm 1, when applying the SRHT to X. However, we only need to perform this operation once for the term with the highest degree, or the terms with lower degree that can be formed by combining nodes computed with the highest degree. Thus, the final runtime is

$$O(\epsilon^{-2} n^2 q^3 \cdot \log^2(nd/\epsilon\delta)\cdot \log(n/\delta) + nd\log(nd/\epsilon\delta)).$$

□

B. General p-Convergent Sequences

We consider general p-convergent kernels defined below in Definition B.1. We apply our proposed Algorithm 1 to compute a subspace embedding with a fast running time.

B.1. General Theorem for $p > 1$

In this section, we state a general theorem for $p > 1$. The proof is similar to the proof for Theorem A.1. We start by restating the definition of a p-convergent kernel.

Definition B.1 (p-convergent kernel matrix, formal version of Definition 6.2)**.** *Given an input matrix $X \in \mathbb{R}^{d\times n}$, we say the kernel matrix $K \in \mathbb{R}^{n\times n}$ is p-convergent if its corresponding Taylor expansion series can be written as follows:*

$$K = \sum_{l=0}^{\infty} C_l \cdot (X^{\otimes l})^\top X^{\otimes l},$$

where the positive coefficients $C_l > 0$ are a function of l, and C_l satisfies

$$C_l = (l+1)^{-\Theta(p)}.$$

Theorem B.2 (Sketch for p-convergent Kernels, formal version of Theorem 6.3)**.** *Let $r \in \mathbb{R}_+$ and $p > 1$ be an integer, and let $X \in \mathbb{R}^{d\times n}$ be such that $\|x_i\|_2 \leq r$ for all $i \in [n]$, where x_i is the i-th column of X. Suppose that K is a p-convergent kernel matrix. For any $p > 1$, we choose $m = \Omega(\epsilon^{-2} n q^3 \log^3(nd/\epsilon\delta))$ and $q = \Theta(r^2 + (n/\epsilon)^{1/p})$. There exists an algorithm which computes a matrix $W_g(X) \in \mathbb{R}^{m\times n}$ in time*

$$O(\epsilon^{-2} n^2 q^3 \cdot \log^3(nd/\epsilon\delta) + nd \cdot \log(nd/\epsilon\delta))$$

such that

$$\Pr\left[W_g(X)^\top W_g(X) \approx_\epsilon G\right] \geq 1 - \delta.$$

Proof. Similar to the Gaussian kernel, here we use the first q terms to approximate the kernel matrix K.

Let $q = C \cdot (r^2 + (n/\epsilon)^{1/p})$ for a sufficiently large constant C, and let $Q = \sum_{l=0}^{q} C_l (X^{\otimes l})^\top X^{\otimes l}$ be the first q terms of K. By the triangle inequality, we have

$$\begin{aligned}
\|K - Q\|_{\mathrm{op}} &\leq \sum_{l>q} C_l \left\|(X^{\otimes l})^\top X^{\otimes l}\right\|_{\mathrm{op}} \\
&\leq \sum_{l>q} C_l \left\|(X^{\otimes l})^\top X^{\otimes l}\right\|_F \\
&\leq \sum_{l>q} C_l \cdot n \cdot r^{2l} \\
&\leq \epsilon/2.
\end{aligned}$$

The proof is identical to the proof of Theorem A.1, with the target dimension of W_g being $m = m_0 + m_1 + \cdots + m_q = \Omega(\epsilon^{-2} n q^3 \log(nd/\delta\epsilon) \log(n/\delta))$.

Similar to Theorem A.1, we have to pay an extra $nd\log(nd/\epsilon\delta)$ term to apply the SRHT to X, so the final running time is

$$t_0 + t_1 + \cdots + t_q + nd\log(nd/\epsilon\delta) = O(\epsilon^{-2} n^2 q^2 \log(nd/\delta\epsilon)\log(n/\delta) + nd\log(nd/\epsilon\delta)).$$

□

Remark B.3. *Recall our setting is when $d = \mathrm{poly}(n)$, so if $p \geq 3$, Theorem B.2 gives a running time of $\widetilde{O}(n^3/\epsilon^2 + nd)$, which is better than the classical result of $O(n^2 d)$ as long as $d > n/\epsilon^2$. However, if $p \in (1, 3)$, Theorem B.2 gives a worse dependence on n, which can be further optimized.*

B.2. Sampling Scheme for $1 < p < 3$

We next describe a novel sampling scheme if $p \in (2, 3)$, with a better dependence on n compared to Theorem B.2. We first state some probability tools.

Theorem B.4 (Matrix Bernstein Inequality(Tropp, 2015)). *Let $S_1, \ldots, S_n$ be independent, zero-mean random matrices with common size $d_1 \times d_2$, and assume each one is uniformly bounded:*

$$\mathbb{E}[S_k] = 0, \|S_k\|_{\text{op}} \leq L, k \in [n]$$

Let $Z = \sum_{k=1}^{n} S_k$, and let $\mathbf{Var}[Z] = \max\{\|\mathbb{E}[Z^\top Z]\|_{\text{op}}, \|\mathbb{E}[ZZ^\top]\|_{\text{op}}\}$. Then for all $t > 0$,

$$\Pr\left[\|Z\|_{\text{op}} \geq t\right] \leq (d_1 + d_2) \cdot \exp\left(\frac{-t^2/2}{\mathbf{Var}[Z] + Lt/3}\right).$$

Theorem B.5 (Sampling Scheme for $2 < p < 3$). *Let $p > 1$ be an integer and $X \in \mathbb{R}^{d \times n}$ be such that $\|x_i\|_2 = 1$ for all $i \in [n]$, where x_i is the i-th column of X, suppose K is a p-convergent kernel matrix. For any $p \in (2, 3)$, there exists an algorithm which computes a matrix $W_g(X)$ with n columns in expected running time*

$$O((n/\epsilon)^{2+6/(1+2p)} \cdot \text{poly}(\log(nd/\epsilon\delta)) + nd\log(nd/\epsilon\delta))$$

such that

$$\Pr\left[W_g(X)^\top W_g(X) \approx_\epsilon G\right] \geq 1 - \delta.$$

Proof. Let q be the degree used in Theorem B.2 where $q = \Theta((n/\epsilon)^{1/p})$, and let s be some positive integer smaller than q. We will consider the following scheme:

- For the first s terms in the Taylor expansion, we approximate each term directly using Theorem B.2.
- For each of the next $q - s$ terms, we sample proportional to their coefficient C_l, taking only s samples in total.

Correctness proof We will show that

$$s = \Theta((n/\epsilon)^{2/(1+2p)} \cdot \text{poly}(\log(nd/\epsilon\delta)))$$

samples suffice. Let P be the sum of the first s terms in the Taylor expansion of K, and let R be the remaining $q - s$ terms. Our goal is to have $\|R\|_{\text{op}} \leq \epsilon\|K\|_{\text{op}}$. We first calculate $\|R\|_{\text{op}}$:

$$\begin{aligned}\|R\|_{\text{op}} &\leq \sum_{l=s+1}^{q} C_l \cdot n \\ &= \sum_{l=s+1}^{q} \frac{1}{l^p} \cdot n \\ &\leq \frac{n}{s^p}.\end{aligned}$$

Notice that if $\|K\|_{\text{op}}$ is large, then it suffices to use the first s terms. Specifically, if $\|K\|_{\text{op}} \geq \frac{n}{\epsilon} s^{-p}$, then we are done. Otherwise, suppose $\|K\|_{\text{op}} \leq \frac{n}{\epsilon} s^{-p}$. We will invoke Theorem B.4 to do the sampling. Let $T = \sum_{l=s+1}^{q} \frac{1}{l^p}$ and $p_i = \frac{C_i}{T}$. Define the random variable S_i as follows: with probability p_l, we sample $R - \frac{C_l}{p_l}(X^{\otimes l})^\top X^{\otimes l}$ for $l = s + 1, \ldots, q$. First notice that S_i is unbiased:

$$\mathbb{E}[S_i] = R - \sum_{i=s+1}^{q} p_i \frac{C_i}{p_i}(X^{\otimes i})^\top X^{\otimes i} = 0.$$

Using the triangle inequality, we have

$$\begin{aligned}\|S_i\|_{\mathrm{op}} &\leq T \cdot \|(X^{\otimes i})^\top X^{\otimes i}\|_{\mathrm{op}} + \|R\|_{\mathrm{op}} \\ &\leq ns^{-p} + ns^{-p} \\ &= 2ns^{-p}.\end{aligned}$$

We now consider the operator norm of the expectation of $S_i^\top S_i$:

$$\begin{aligned}\|\mathbb{E}[S_i^\top S_i]\|_{\mathrm{op}} &= \left\| \frac{C_i^2}{p_i}(X^{\otimes i})^\top X^{\otimes i}(X^{\otimes i})^\top X^{\otimes i} + R^\top R - \frac{C_i}{p_i}(X^{\otimes i})^\top X^{\otimes i} R - \frac{C_i}{p_i} R (X^{\otimes i})^\top X^{\otimes i} \right\|_{\mathrm{op}} \\ &\leq \|C_i T (X^{\otimes i})^\top X^{\otimes i}(X^{\otimes i})^\top X^{\otimes i}\|_{\mathrm{op}} + \|R^\top R\|_{\mathrm{op}} + T \cdot (\|(X^{\otimes i})^\top X^{\otimes i} R\|_{\mathrm{op}} + \|R(X^{\otimes i})^\top X^{\otimes i}\|_{\mathrm{op}}) \\ &\leq C_i T n^2 + (ns^{-p})^2 + 2Tn \cdot ns^{-p} \\ &= 4n^2 s^{-2p}.\end{aligned}$$

Let $Z = \sum_{i=1}^m S_i$. Since each sample is sampled independently, we have

$$\begin{aligned}\|\mathbb{E}[Z^\top Z]\|_{\mathrm{op}} &= \left\| \mathbb{E}\left[\sum_{i=1}^s S_i^\top S_i \right] \right\|_{\mathrm{op}} \\ &\leq \sum_{i=1}^s \|\mathbb{E}[S_i^\top S_i]\|_{\mathrm{op}} \\ &\leq 4mn^2 s^{-2p}.\end{aligned}$$

Let $t = m\epsilon\|K\|_{\mathrm{op}}$. Applying Theorem B.4, we get that

$$\Pr\left[\|Z\|_{\mathrm{op}} \geq m\epsilon\|K\|_{\mathrm{op}}\right] \leq 2n \cdot \exp\left(\frac{-m^2\epsilon^2\|K\|_{\mathrm{op}}^2/2}{4mn^2 s^{-2p} + 2m\epsilon\|K\|_{\mathrm{op}} n s^{-p}/3} \right).$$

Picking $m = \Theta(\epsilon^{-2} n^2 s^{-2p} \log(n/\delta))$, and then averaging over m samples, we get that

$$\Pr\left[\left\| \frac{1}{m} \sum_{i=1}^m S_i \right\|_{\mathrm{op}} \geq \epsilon \|K\|_{\mathrm{op}} \right] \leq \delta, \tag{5}$$

where we use the fact that the operator norm of K is at least 1, by our choice of s. We now compute the expected running time of this algorithm.

Runtime part 1: Computing the first s terms For the first s terms, we can apply the same reasoning as in Theorem B.2 to get a running time of $O(\epsilon^{-2} n^2 s^3 \operatorname{poly}(\log(nd/\epsilon\delta)) + nd\log(nd/\epsilon\delta))$.

Runtime part 2: Sampling the next s terms For the sampling part, we consider the *expected degree* D of the sample we will be working with:

$$\begin{aligned}D &= \sum_{l=s+1}^q p_l \cdot l \\ &= \frac{\sum_{l=s+1}^q l^{1-p}}{\sum_{l=s+1}^q l^{-p}} \\ &\leq \frac{s^{1-p}}{s^{-p} - q^{-p}} \\ &= s + \frac{s \cdot q^{-p}}{s^{-p} - q^{-p}} \\ &= s + \frac{s}{(\frac{s}{q})^p - 1} \\ &= \widetilde{O}((n/\epsilon)^{2/(1+2p)}).\end{aligned}$$

Now we are ready to compute the expected running time of the sampling phase:

$$m \cdot D^2 n^2/\epsilon^2 = (n/\epsilon)^{2+6/(1+2p)} \cdot \mathrm{poly}(\log(nd/\epsilon\delta)).$$

Additionally, we need to apply the SRHT to X at most twice, once for the initial phase, and once for the sampling phase, so the final running time is

$$(n/\epsilon)^{2+6/(1+2p)} \cdot \mathrm{poly}(\log(nd/\epsilon\delta)) + nd\log(nd/\epsilon\delta).$$

□

When $p \in (1, 2]$, we use the largest degree q as an upper bound for analyzing our running time.

Corollary B.6 (Sampling Scheme for $1 < p \leq 2$)**.** *Let $p > 1$ be an integer and $X \in \mathbb{R}^{d\times n}$ be such that $\|x_i\|_2 = 1$ for all $i \in [n]$, where x_i is the i-th column of X. Suppose K is a p-convergent kernel matrix. If $p \in (1, 2]$, then there exists an algorithm which computes a matrix $W_g(X)$ with n columns in time*

$$\epsilon^{-(2+6/(3+2p))} n^{2+6(1+1/p)/(3+2p)} \cdot \mathrm{poly}(\log(nd/\epsilon\delta)) + nd\log(nd/\epsilon\delta)$$

such that

$$\Pr\left[W_g(X)^\top W_g(X) \approx_\epsilon G\right] \geq 1 - \delta.$$

Remark B.7. *In addition, using that $p \in (1, 2]$, the first part of our running time can be upper bounded by*

$$\epsilon^{-3.2} n^{4.4} \cdot \mathrm{poly}(\log(nd/\epsilon\delta)).$$

Proof. The proof is almost identical to the proof of Theorem B.5. The only difference is when considering the expected degree D, we use q as an upper bound. The number of terms s we approximate in the initial phase will be

$$\widetilde{\Theta}\Big(\frac{n^{(2+2/p)/(3+2p)}}{\epsilon^{2/(3+2p)}}\Big).$$

The final runtime will be

$$\epsilon^{-2} n^2 \,\mathrm{poly}\log(nd/\epsilon\delta) s^3 = (n^{2+3(2+2/p)/(3+2p)}/\epsilon^{2+6/(3+2p)})\,\mathrm{poly}\log(nd/\epsilon\delta) + nd\log(nd/\epsilon\delta).$$

□

Simplifying the Exponent For the exponent of ϵ, we have

$$(2 + 6/(3 + 2p)) = 4 - \underbrace{\frac{4p}{3+2p}}_{f_1(p)}.$$

For any $p \in (1, 2]$, we have

$$4 - f_1(p) \in \Big[2 + \frac{6}{7}, 3 + \frac{1}{5}\Big).$$

For the exponent on n, we have

$$\begin{aligned} 2 + 6(1 + 1/p)/(3 + 2p) &= 2 + \frac{6p+6}{2p^2+3p} \\ &= 5 - \underbrace{\frac{6p^2+3p-6}{2p^2+3p}}_{f_2(p)}. \end{aligned}$$

For any $p \in (1, 2]$, we have

$$5 - f_2(p) \in \Big[3 + \frac{2}{7}, 4 + \frac{2}{5}\Big).$$

C. Properties of the Neural Tangent Kernel

We discuss an application of our sampling algorithm for $p \in (1, 2]$ (Corollary B.6) to the Neural Tangent Kernel (NTK). We will first formally define the NTK, then consider its Taylor expansion, and then use a p-convergent kernel to bound it.

C.1. Taylor Expansion of NTK

In this section, we give the Taylor expansion of the NTK, by first examining its corresponding function in a single variable, and then extend it to the matrix case.

Consider a simple two-layer (an alternative name is one-hidden-layer) ReLU network with input layer initialized to standard Gaussians, activation function ReLU, and output layer initialized to uniform and independent Rademacher ($\{-1, 1\}$) random variables. Suppose we fix the output layer. Then the neural network can be characterized by a function

$$f(W, x) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \sigma(w_r^\top x)$$

where $W \in \mathbb{R}^{d \times m}$, and $w_r \in \mathbb{R}^d$ denotes the r-th column of W, for each $r \in [m]$.

The above formulation is standard for convergence analysis of neural networks (Du et al., 2019; Song & Yang, 2019; Brand et al., 2021; Huang et al., 2021).

The NTK ((Jacot et al., 2018)) is

$$\mathsf{K}(x, z) = \mathbb{E}\left[\langle \frac{\partial f(W, x)}{\partial W}, \frac{\partial f(W, z)}{\partial W} \rangle\right].$$

For the sake of simplicity, assume all $|a_r| = 1, \forall r \in [m]$, and consider an individual summand, which gives rise to

$$\mathsf{K}(x, z) = \int_{w \sim N(0,I)} \sigma'(w^\top x)\sigma'(w^\top z)x^\top z \, \mathrm{d}w.$$

If $w \in \mathbb{R}^d$ is chosen uniformly on a sphere, then we will get the following closed-form for this kernel (Cho & Saul, 2009; Xie et al., 2017):

$$\mathsf{K}(x, z) = \left(\frac{1}{2} - \frac{\arccos x^\top z}{2\pi}\right) \cdot x^\top z.$$

Fact C.1. *Let function $f : \mathbb{R} \to \mathbb{R}$ be defined as $f(x) := (\frac{1}{2} - \frac{\arccos x}{2\pi}) \cdot x$. Then the Taylor expansion of f is*

$$\begin{aligned} f(x) &= \frac{x}{4} + \left(\sum_{n=0}^{\infty} \frac{(2n)!}{2^{2n}(n!)^2} \frac{x^{2n+2}}{(2n+1)(2\pi)}\right) \\ &= \frac{x}{4} \left(\sum_{n=0}^{\infty} \binom{2n}{n} \frac{1}{2^{2n}} \frac{x^{2n+2}}{(2n+1)(2\pi)}\right). \end{aligned}$$

Fact C.2. *The Taylor expansion of the NTK is*

$$\mathsf{K} = \frac{X^\top X}{4} \sum_{l=0}^{\infty} \binom{2l}{l} \frac{1}{2^{2l}} \frac{(X^{\otimes 2l+2})^\top X^{\otimes 2l+2}}{(2l+1)2\pi}.$$

C.2. Approximating the NTK

In this section, we will use a p-convergent kernel to bound the NTK, then apply Corollary B.6 to approximate it.

Corollary C.3 (Fast Subspace Embedding for the NTK)**.** *Let $X \in \mathbb{R}^{d \times n}$ where $\|x_i\|_2 = 1$ for all $i \in [n]$, where x_i is the i^{th} column of X. Suppose $\mathsf{K} \in \mathbb{R}^{n \times n}$ is the NTK matrix. Then there exists an algorithm which computes a matrix $W_g(X)$ in time*

$$\epsilon^{-3} n^{11/3} \cdot \mathrm{poly}(\log(nd/\epsilon\delta)) + nd \log(nd/\epsilon\delta)$$

such that

$$\Pr\left[W_g(X)^\top W_g(X) \approx_\epsilon \mathsf{K}\right] \geq 1 - \delta.$$

Proof. Let C_l denote the coefficient of the l^{th} term in the Taylor expansion of the NTK:

$$C_l = \binom{2l}{l} \frac{1}{2^{2l}} \frac{1}{(2l+1)2\pi}.$$

The term $\binom{2l}{l}$ is the central binomial coefficient. We will use the following bound on it:

$$\frac{4^l}{\sqrt{4l}} \leq \binom{2l}{l} \leq \frac{4^l}{\sqrt{3l+1}}.$$

This gives upper and lower bounds on C_l:

- Upper bound:

$$\begin{aligned} C_l &\leq \frac{4^l}{\sqrt{3l+1}} \frac{1}{4^l} \frac{1}{(2l+1)2\pi} \\ &= \frac{1}{\sqrt{3l+1}(2l+1)2\pi}. \end{aligned}$$

- Lower bound:

$$C_l \geq \frac{1}{\sqrt{4l}(2l+1)2\pi}.$$

Thus, $C_l = \Theta(\frac{1}{l^{1.5}})$, and we can use a 1.5-convergent kernel for our approximation. Using Corollary B.6 with $p = 1.5$, we obtain an ϵ-approximation in time

$$\epsilon^{-3} n^{11/3} \cdot \mathrm{poly}(\log(nd/\epsilon\delta)) + nd \log(nd/\epsilon\delta).$$

□

D. Preconditioning to Solve a Kernel Linear System

In this section, we illustrate how to construct a preconditioner for a kernel linear system. Specifically, we provide an algorithm to solve a Gaussian kernel linear system. Let $G = Z^\top Z$ be the Gaussian kernel. By Theorem A.1, we can compute an approximation to G, denoted $W_g(X)^\top W_g(X)$. In (Brand et al., 2021) (see Algorithm 2 and Section 4.1 there), Brand, Peng, Song and Weinstein show that if we compute the QR decomposition of $W_g(X) = QR^{-1}$, where Q has orthonormal columns and $R \in \mathbb{R}^{n \times n}$, then R is a good preconditioner for Z, i.e., ZR has constant condition number. However, in our setup where d is large, it is not feasible to compute Z directly, which takes $O(n^2 d)$ time. Instead, we notice that $W_g(X)$ is fast to compute and has only an $\widetilde{O}(n/\epsilon^2)$ number of rows. Our algorithm will sketch $W_g(X)$, and then use gradient descent to solve the optimization problem

$$\min_{x \in \mathbb{R}^n} \|W_g(X)^\top W_g(X) x - y\|_2.$$

In our result, we follow a similar approach as in (Brand et al., 2021) and the proof is similar to the proof of Lemma 4.2 in their paper. The main novelty of our framework is that we use a spectral approximation to the kernel matrix and analyze the error and runtime under our approximation. For completeness, we include a proof in this setting.

Theorem D.1 (Sketching as a Preconditioner, formal version of Theorem 6.6)**.** *Let $G \in \mathbb{R}^{n \times n}$ be the Gaussian kernel matrix for $X \in \mathbb{R}^{d \times n}$. Write $G = Z^\top Z$, and let κ denote the condition number of Z. If we assume for all $i \in [n]$ that $\|x_i\|_2 \leq 1$, then Algorithm 2, with probability at least $1 - \delta$, computes an $\widehat{x}$ satisfying the following:*

$$\|G\widehat{x} - y\|_2 \leq \epsilon \|y\|_2.$$

Moreover, $\widehat{x}$ can be computed in time

$$\epsilon^{-2} n^2 \log(\kappa/\epsilon) \cdot \mathrm{poly}(\log(nd/\epsilon\delta)) + n^\omega + nd \log(nd/\epsilon\delta),$$

where ω is the matrix multiplication exponent.

Algorithm 2 Fast Regression for the Gaussian Kernel

procedure PRECONDITIONEDGRADIENTDESCENT(X, y) ▷ Theorem D.1
  Let $m = O(n \log^2(nd/\epsilon\delta) \log(n/\delta)/\epsilon^2)$
  Let $W_g(X) \in \mathbb{R}^{m \times n}$ be the approximate Gaussian kernel in Theorem A.1
  Let $S \in \mathbb{R}^{l/\epsilon_0^2 \times m}$ be an SRHT matrix. Compute $SW_g(X)$, where $l = \Omega\left(n \log(mn/\epsilon_0\delta) \log(n/\delta)\right)$
  Compute R such that $SW_g(X)R$ has orthonormal columns via a QR decomposition ▷ $R \in \mathbb{R}^{n \times n}$
  $z_0 \leftarrow \mathbf{0}_n \in \mathbb{R}^n$
  while $\|W_g(X)^\top W_g(X) R z_t - y\|_2 \geq \epsilon$ **do**
    $z_{t+1} \leftarrow z_t - (R^\top W_g(X)^\top W_g(X) R)^\top (R^\top W_g(X)^\top W_g(X) R z_t - R^\top y)$
  end while
  return Rz_t
end procedure

Before the proof, we define some notation and corresponding facts specifically about a PSD matrix.

Fact D.2 (Inequality for condition numbers)**.** *Let A, B be conforming square matrices. Then the following inequality holds:*

$$\kappa(B) \leq \kappa(AB)\kappa(A),$$

where $\kappa(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$ is the condition number of A.

We will make use of Lemma B.2 in (Brand et al., 2021).

Lemma D.3 (Lemma B.2 in (Brand et al., 2021))**.** *Consider the regression problem:*

$$\min_{x \in \mathbb{R}^n} \|Bx - y\|_2^2.$$

Suppose B is a PSD matrix for which $\frac{3}{4} \leq \|Bx\|_2 \leq \frac{5}{4}$ holds for all $\|x\|_2 = 1$. Using gradient descent for t iterations, we obtain

$$\|B(x_t - x^*)\|_2 \leq c^t \|B(x_0 - x^*)\|_2,$$

where x_0 is our initial guess, x^* is the optimal solution, and $c \in (0, 0.9]$.

Proof of Theorem D.1. Throughout the proof, we will set $\widehat{\epsilon} = \epsilon/4$. By Theorem A.1, we can compute an ϵ-approximation to Z and $W_g(X)$ in time

$$O(\epsilon^{-2} n^2 \cdot \mathrm{poly}(\log(nd/\epsilon\delta)) + nd \log(nd/\epsilon\delta)).$$

If we solve the problem:

$$\min_{x \in \mathbb{R}^n} \|W_g(X)^\top W_g(X) x - y\|_2 \tag{6}$$

with solution $\widehat{x}$, then we have

$$\|W_g(X)^\top W_g(X)\widehat{x} - y\|_2 \leq (1 + \widehat{\epsilon}) \min_{x \in \mathbb{R}^n} \|Z^\top Z x - y\|_2.$$

This means the optimal solution for the sketched problem gives an $\widehat{\epsilon}$-approximation to the optimal solution to the original problem. We will now show that Algorithm 2 computes the desired solution. By Theorem 2.11, with probability at least $1 - \delta$, for any $x \in \mathbb{R}^n$, we have

$$\|SW_g(X)x\|_2 = (1 + \epsilon_0)\|W_g(X)x\|_2.$$

Suppose R is the $n \times n$ matrix computed via a QR decomposition, so that $SW_g(X)R$ has orthonormal columns. Then for any $\|x\|_2 = 1$, we have

$$\|W_g(X)Rx\|_2 = (1 + \epsilon_0)\|SW_g(x)Rx\|_2 = 1 + \epsilon_0.$$

Hence,

$$\|R^\top W_g(X)^\top W_g(X) R x\|_2 \leq (1+\epsilon_0)^2.$$

Now, pick $\epsilon_0 = 0.1$ and solve the following regression problem:

$$\min_{z \in \mathbb{R}^n} \|R^\top W_g(X)^\top W_g(X) R z - R^\top y\|_2. \tag{7}$$

Notice that Algorithm 2 implements gradient descent. Using Lemma D.3, after $t = \log(1/\widehat{\epsilon})$ iterations, we have

$$\|R^\top W_g(X)^\top W_g(X) R(z_t - z^*)\|_2 \leq \widehat{\epsilon} \|R^\top W_g(X)^\top W_g(X) R(z_0 - z^*)\|_2, \tag{8}$$

where $z^* = (R^\top W_g(X)^\top W_g(X) R)^{-1} R^\top y$ is the optimal solution to Equation (7). We will show the following for $x_t = R z_t$:

$$\|W_g(X)^\top W_g(X) x_t - y\|_2 \leq \kappa \widehat{\epsilon} \|y\|_2.$$

Recalling that $z_0 = 0$, plugging into Eq. (8) we get

$$\|R^\top W_g(X)^\top W_g(X) x_t - R^\top y\|_2 \leq \widehat{\epsilon} \|R^\top y\|_2 \leq \widehat{\epsilon} \cdot \sigma_{\max}(R^\top) \|y\|_2.$$

On the other hand,

$$\|R^\top W_g(X)^\top W_g(X) x_t - R^\top y\|_2 = \|R^\top (W_g(X)^\top W_g(X) x_t - y)\|_2 \geq \sigma_{\min}(R^\top) \|W_g(X)^\top W_g(X) x_t - y\|_2.$$

Putting everything together, we get

$$\begin{aligned}
\|W_g^\top W_g x_t - y\|_2 &\leq \widehat{\epsilon} \kappa(R^\top) \|y\|_2 \\
&\leq \widehat{\epsilon} \kappa(R) \|y\|_2 \\
&\leq \widehat{\epsilon} \kappa(W_g(X) R) \kappa(W_g(X)) \|y\|_2 \\
&\leq 2\widehat{\epsilon} \kappa(W_g(X)) \|y\|_2 \\
&\leq 2\widehat{\epsilon} \kappa \frac{1+\widehat{\epsilon}}{1-\widehat{\epsilon}} \|y\|_2 \\
&\leq 2\kappa \widehat{\epsilon} \|y\|_2.
\end{aligned}$$

The second inequality uses that R is a square matrix, the third inequality uses Fact D.2, and the second-to-last inequality uses that we have a $(1 \pm \widehat{\epsilon})$-subspace embedding. This means by setting the number of iterations to $t = \log(\kappa/\epsilon)$, we obtain

$$\|W_g(X)^\top W_g(X) x_t - y\|_2 \leq 2\widehat{\epsilon} \|y\|_2.$$

Now, recall that for any $x, y \in \mathbb{R}^n$, we have

$$\|W_g(X)^\top W_g(X) x - y\|_2 \leq (1+\widehat{\epsilon}) \|Z^\top Z x - y\|_2.$$

As a consequence, we get

$$\begin{aligned}
\|Z^\top Z x_t - y\|_2 &\leq (1+\widehat{\epsilon}) \|W_g(X)^\top W_g(X) x_t - y\|_2 \\
&\leq (1+\widehat{\epsilon}) 2\widehat{\epsilon} \|y\|_2 \\
&\leq \epsilon \|y\|_2.
\end{aligned}$$

Now we analyze the runtime.

- Computing $W_g(X)$, by Theorem A.1, takes time

$$\epsilon^{-2} n^2 \cdot \mathrm{poly}(\log(nd/\epsilon\delta)) + nd \log(nd/\epsilon\delta).$$

- Applying S to $W_g(X)$, using the FFT algorithm, takes time

$$\epsilon^{-2} n^2 \cdot \mathrm{poly}(\log(nd/\epsilon\delta)).$$

- A QR decomposition algorithm, due to (Demmel et al., 2007), can be computed in time n^ω.

The cost of each iteration is bounded by the cost of taking a matrix-vector product, which is at most $\widetilde{O}(n^2/\epsilon^2)$, and there are $O(\log(\kappa/\epsilon))$ iterations in total. Thus, we obtain a final runtime of

$$\epsilon^{-2} n^2 \cdot \mathrm{poly}(\log(nd/\epsilon\delta)) \cdot \log(\kappa/\epsilon) + n^\omega + nd \log(nd/\epsilon\delta).$$

□

E. Kernel Ridge Regression

In this section, we show how to compose our sketch with other sketches whose dimensions depend on the statistical dimension of K instead of n. Before proceeding, we introduce the notion of the *statistical dimension*.

Definition E.1 (Statistical Dimension)**.** *Given $\lambda \geq 0$, for every positive semi-definite matrix $K \in \mathbb{R}^{n \times n}$, we define the λ-statistical dimension of K to be*

$$s_\lambda(K) := \mathrm{tr}[K(K + \lambda I_n)^{-1}].$$

Solving ridge regression with runtime depending on thhe statistical dimension is done in a number of works, for example (Rahimi & Recht, 2007; Alaoui & Mahoney, 2015; Avron et al., 2017c;a; Musco & Musco, 2017).

We state and prove our main result in this section below.

Theorem E.2 (Kernel Ridge Regression, formal version of Theorem 6.9)**.** *Let $\epsilon \in (0, 1)$, $p > 1$ be an integer, and $X \in \mathbb{R}^{d \times n}$. If K is a degree-p polynomial kernel with statistical dimension $s_\lambda(K)$ with $\lambda < \epsilon^{-2}\lambda_{\max}(K)$, then we can compute $Z \in \mathbb{R}^{t \times n}$ such that $Z^\top Z$ is a $1 \pm \epsilon$ spectral approximation to K in $\widetilde{O}(\epsilon^{-2}p^2n^2 + nd)$ time and $t = \widetilde{O}(\epsilon^{-2}p^2 n)$.*

Moreover, there exists a matrix S with $m = \widetilde{O}(\epsilon^{-1}s_\lambda(K))$ rows such that if x^* is the optimal solution to $\|S(Z^\top Zx - y)\|_2^2 + \lambda\|Zx\|_2^2$, then

$$\|Kx^* - y\|_2^2 + \lambda\|X^{\otimes p}x^*\|_2^2 \leq (1+\epsilon) \min_{x \in \mathbb{R}^n} \|Kx - y\|_2^2 + \lambda\|X^{\otimes p}x\|_2^2.$$

Finally, The time to solve above KRR *is $\widetilde{O}(\epsilon^{-2}p^2n(n + m^2) + n^\omega)$.*

Before starting the proof, we introduce a key lemma regarding using the SRHT to approximate the solution of KRR.

Lemma E.3 (Corollary 15 of (Avron et al., 2017a))**.** *Let $A \in \mathbb{R}^{n \times t}$ and $\epsilon \in (0, 1)$. Suppose $\lambda < \epsilon^{-2}\sigma_{\max}^2(A)$. Suppose*

$$m = \Omega(\epsilon^{-1}(s_\lambda(A) + \log(1/\epsilon)) \log(s_\lambda(A)/\epsilon))$$

and $S \in \mathbb{R}^{m \times n}$ is a SRHT *matrix (Definition 2.7) and let $\widehat{x} = \arg\min_{x \in \mathbb{R}^t} \|S(Ax - b)\|_2^2 + \lambda\|x\|_2^2$. Then with probability at least 0.99, we have*

$$\|A\widehat{x} - b\|_2^2 + \lambda\|\widehat{x}\|_2^2 \leq (1+\epsilon) \min_{x \in \mathbb{R}^t} \|Ax - b\|_2^2 + \lambda\|x\|_2^2.$$

Proof of Theorem E.2. Throughout the proof, we assume K has full rank and set S to be a SRHT matrix with $m = \widetilde{O}(\epsilon^{-1}s_\lambda(K))$ rows. We also use A to denote $Z^\top Z$.

The proof consists of 3 parts:

- Part 1: Provide a construction of matrix Z;
- Part 2: Provide a sketching matrix S with the solution guarantee;

- Part 3: Provide a runtime analysis for solving KRR.

Note that part 1 can be solved using Theorem 5.1. As a side note, since $Z^\top Z$ is a $1 \pm \epsilon$ approximation to K, with high probability it also has full rank. Consequently, Z has full rank as well.

To show part 2, we will show the following:

- The optimal solution to $\|S(Z^\top Zx - y)\|_2^2 + \lambda\|Zx\|_2^2$ is a $(1 \pm \epsilon)$ approximation to the optimum of $\|Z^\top Zx - y\|_2^2 + \lambda\|Zx\|_2^2$;
- The optimum of $\|Z^\top Zx - b\|_2^2 + \lambda\|Zx\|_2^2$ is a $(1 \pm \epsilon)$ approximation to the optimum of $\|Kx - y\|_2^2 + \lambda\|X^{\otimes p}x\|_2^2$.

From $\|S(Z^\top Zx - y)\|_2^2 + \lambda\|Zx\|_2^2$ to $\|Z^\top Zx - y\|_2^2 + \lambda\|Zx\|_2^2$ Recall that Z has full rank. Therefore, we can set $z = Zx$ and the sketched problem becomes

$$\|S(Z^\top z - y)\|_2^2 + \lambda\|z\|_2^2,$$

which can be solved using Lemma E.3. The only thing we need to justify is that the statistical dimension of A gives a good approximation to the statistical dimension of K. Note that

$$\begin{aligned} s_\lambda(A) &= \sum_{i=1}^n \frac{\lambda_i(A)}{\lambda_i(A) + \lambda} \\ &\le \sum_{i=1}^n \frac{(1+\epsilon)\lambda_i(K)}{(1-\epsilon)\lambda_i(K) + \lambda} \\ &\le \sum_{i=1}^n \frac{(1+\epsilon)\lambda_i(K)}{(1-\epsilon)(\lambda_i(K) + \lambda)} \\ &= \frac{1+\epsilon}{1-\epsilon} \cdot s_\lambda(K) \\ &\le (1+3\epsilon) \cdot s_\lambda(K). \end{aligned}$$

Thus, the dimension $O(\epsilon^{-1}s_\lambda(K)) = O(\epsilon^{-1}s_\lambda(A))$, which means we can invoke Lemma E.3.

From $\|Z^\top Zx - y\|_2^2 + \lambda\|Zx\|_2^2$ to $\|Kx - y\|_2^2 + \lambda\|X^{\otimes p}x\|_2^2$ To prove this part, we define matrix $\widehat{A}$ and $\widehat{K}$:

$$\widehat{A} := \begin{bmatrix} A \\ \sqrt{\lambda}Z \end{bmatrix}, \widehat{K} := \begin{bmatrix} K \\ \sqrt{\lambda}X^{\otimes p} \end{bmatrix}.$$

Similar to the first part, it suffices to show that for any $x \in \mathbb{R}^n$, we have

$$\|\widehat{A}x\|_2 \le (1+\epsilon)\|\widehat{K}x\|_2.$$

We start by computing the LHS:

$$\begin{aligned} \|\widehat{A}x\|_2^2 &= \|Ax\|_2^2 + \lambda x^\top Z^\top Zx \\ &\le (1+\epsilon)\|Kx\|_2^2 + (1+\epsilon)\lambda\|X^{\otimes p}x\|_2^2. \end{aligned}$$

This completes our proof for part 2.

For the final part, note that applying the sketch takes $\widetilde{O}(\epsilon^{-2}p^2n^2)$ time. To solve the regression problem, we instead solve:

$$\min_{z \in \mathbb{R}^t} \|SZ^\top z - Sy\|_2^2 + \lambda\|z\|_2^2.$$

Since Z has full rank, we know the argument z realizing the minimum is the x^* we are looking for. To output an x, we can simply solve the linear system $Zx = z$, which takes $\widetilde{O}(nt + n^\omega)$ time. Finally, solving the above regression problem takes $\widetilde{O}(m^2t)$ time (see (Saunders et al., 1998)). This concludes our runtime analysis. □