Large-Scale Data-Dependent Kernel Approximation

Appendix

This appendix presents the additional detail and proofs associated with the main paper [1].

1 Introduction

Let \( k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+ \) be a positive definite translation invariant function e.g. a Gaussian kernel \( k(x, y) = \exp(-\gamma \|x-y\|^2) \).

By Bochner’s theorem there exists \( \mu \) a positive function such that

\[
k(x, y) = \int e^{i\omega^\top (x-y)} \mu(\omega)
\]

Since \( \mu \) is positive we can use it to draw i.i.d. samples \( \omega_i \sim \mu \) which allows us to define a random feature map such that \( \phi(x) = [\phi_1(x) \ldots \phi_d(x)] \), where \( \phi_i(x) = \cos(\omega_i^\top x + b_i) \) (where \( b_i \sim \text{Uniform}[0, 2\pi] \)). Let \( \tilde{k}(x, y) = \sum_i \tilde{k}_i(x, y) = \frac{1}{d} \sum_i \phi_i(x)\phi_i(y)^\top = \frac{1}{d} \phi(x)\phi(y)^\top \). This is a standard construction; see [2, 3] for more details.

Let \( X \) be a fixed data matrix \( N \times p \) corresponding to \( N \) data points in \( \mathbb{R}^p \) and let the matrix counterparts of the above notation applied to \( X \) be \( K(i, j) = k(X(i, :), X(j, :)) \), as well as \( \tilde{K}, \tilde{K}_i, \tilde{\Phi}_i(= \phi_i(X)) \) and \( \Phi(= \phi(X)) \).

With this notation we have

\[
\tilde{K} = \sum_i \tilde{K}_i = \sum_i \tilde{\Phi}_i\tilde{\Phi}_i^\top = \Phi\Phi^\top
\]

We notice that \( \tilde{K}_i \) are i.i.d. thus matrix concentration results apply to it.

To this end we want to use

**Theorem 1 (Matrix Bernstein [4])** Let \( Z_1 \ldots Z_m \) be independent \( n \times n \) Hermitian random matrices with \( \mathbb{E}[Z_i] = 0 \) and \( \|Z_i\| \leq R \). Let \( \sigma^2 = \max\{\|\sum_i \mathbb{E}[Z_i^* Z_i]\|, \|\sum_i \mathbb{E}[Z_i Z_i^*]\|\} \), where \( \|\cdot\| \) is the operator norm. Then

\[
\mathbb{E}\left\| \sum_i Z_i \right\| \leq \sigma \sqrt{3\log(2n)} + R \log(2n)
\]

**Theorem 2 (\( \tilde{K} \) convergence [3])** Let \( \tilde{K} \) be an \( d \) term random feature approximation of the kernel matrix \( K \in \mathbb{R}^{N \times N} \)

\[
\mathbb{E}\|\tilde{K} - K\| \leq \sqrt{\frac{3N^2 \log N}{d}} + \frac{2N \log N}{d}
\]

**Proof** Then \( \tilde{K}_i \) are independent and we know that \( \mathbb{E}[\tilde{K}] = K \).

\[
E = \tilde{K} - K = \sum_i E_i, \ E_i = \frac{1}{d}(\tilde{K}_i - K)
\]

Thus \( \mathbb{E}[E_i] = 0 \) and \( E_i \) are i.i.d. as well.

First we must show that each are bounded

\[
\|E_i\| = \frac{1}{d} \left\| \tilde{\Phi}_i\tilde{\Phi}_i^\top - \mathbb{E}[\Phi\Phi^\top] \right\| \leq \frac{1}{d} (\|\tilde{\Phi}_i\|^2 + \mathbb{E}[\|\Phi\|^2]) \leq \frac{1}{d} (\|\Phi_i\|^2 + \|\mathbb{E}[\Phi]\|^2) \leq \frac{2B}{d}
\]

\[1\]This is from [3] reproduced for a self-contained understanding of our main results.
where we used first the definitions of $\hat{K}_i$ and $K_i$, followed by the triangle inequality, then Jensen for the expected value. $B$ is a finite bound for $\|\phi\| (\|\phi\|^2 \leq B)$. We know that such a bound exists, by the way $\phi$ is constructed.

Then the variance of $E_i$ is

$$E[E_i^2] = \frac{1}{d^2} E[(\Phi_i\Phi_i^T - K)^2]$$

(6)

$$= \frac{1}{d^2} E[(\|\Phi_i\|^2 \Phi_i\Phi_i^T - K\Phi_i\Phi_i^T + K^2)]$$

(7)

$$\approx \frac{1}{d^2} [BK - 2K^2 + K^2] \leq BK$$

(8)

where we unravel the square, then use $E[\hat{K}_i] = E[\Phi_i\Phi_i^T] = K$. The second $\approx$ is due to $K$ being positive definite.

Then the variance of $E_i$ is

$$E[E_i^2] = \frac{1}{d^2} \sum_i E[E_i^2] \leq \frac{1}{d} B\|K\|$$

(9)

where we first used Jensen’s inequality, then the semi-definite bound above with $d$ terms.

Given these bounds on the variance and the norm of the random variables, we can apply (2) to get

$$E\|\hat{K} - K\| \leq \sqrt{\frac{3B\|K\| \log N}{d}} + \frac{2B \log N}{d}$$

(10)

2 Data-Dependent Kernel

Let $L$ be the normalized Laplacian i.e. $L = I - D^{-1/2}WD^{-1/2}$ with $W$ again some fixed positive definite function of the data and $D$ a diagonal matrix with the sum of each row of $W$. Let $M = L$ or some positive power of the Laplacian $M = \alpha L^c$. Then we define

$$\hat{K} = K - K(I + MK)^{-1}MK$$

(11)

as a new kernel, similarly to the one defined in [5].

So the goal is to obtain $\Phi$ with both some guarantees of consistency and a large deviation bound, in order to characterize the speed of convergence.

To this end we define

$$\mathcal{K} = \hat{K} - \hat{K}(I + M\hat{K})^{-1}M\hat{K}$$

(12)

and

$$\tilde{K} = \Phi(I + \Phi^T M\Phi)^{-1}\Phi^T$$

(13)

The Sherman-Morrison-Woodbury (SMW) identity in its simplest form states that if both $I + UV^T$ and $I + V^TU$ are invertible then

$$(I + UV^T)^{-1} = I - U(I + V^TU)^{-1}V^T$$

(14)

Proposition 2 With the definitions above

$$\mathcal{K} = \tilde{K}$$

(15)

Proof

$$\mathcal{K} = \hat{K} - \hat{K}(I + M\hat{K})^{-1}M\hat{K}$$

(16)

$$= \Phi\Phi^T - \Phi\Phi^T (I + M\Phi\Phi^T)^{-1}M\Phi\Phi^T$$

(17)

$$= \Phi(I - \Phi^T (I + M\Phi\Phi^T)^{-1}M\Phi)\Phi^T$$

(18)

$$= \Phi(I + \Phi^T M\Phi)^{-1}\Phi^T$$

(19)

$$= \tilde{K}$$

(20)

Where (19) comes by applying (14) with $U = \Phi^T$ and $V = \Phi^T M$ and using the symmetry of $M$. 
So \( \hat{\Phi} = \Phi(I + \Phi^T M \Phi)^{-1/2} \) but given [15] we can use \( \hat{K} \) instead of \( \hat{K} \) for the convergence proofs. Now the goal is to obtain a bound on \( \mathbb{E}||\hat{K} - \hat{K}|| \).

**Lemma 3** Let \( \hat{K} \) and \( \hat{K} \) defined as above and denoting \( \mathbb{E}||\hat{K} M (I + \hat{K} M)^{-1}|| \leq R \) and \( \mathbb{E}||(I + MK)^{-1} MK|| \leq T \), with \( R, T \) constants we have that

\[
\mathbb{E}||\hat{K} - \hat{K}|| \leq \mathbb{E}||K - \hat{K}||(1 + TR + R) \tag{21}
\]

**Proof**

\[
||\hat{K} - \hat{K}|| = ||\hat{K} - \hat{K}(I + M\hat{K})^{-1} MK - K + (I + MK)^{-1} MK|| \tag{22}
\]

\[
\leq ||\hat{K} - K|| + ||\hat{K}(I + M\hat{K})^{-1} MK - K + (I + MK)^{-1} MK|| \tag{23}
\]

If we apply the triangle inequality for the second term in the right side of inequality (23) in the form of \( ||A + B + C|| \leq ||A|| + ||B|| + ||C|| \) with,

\[
A = \hat{K}(I + MK)^{-1} MK - K(I + MK)^{-1} MK \tag{24}
\]

\[
B = \hat{K}(I + MK)^{-1} MK - \hat{K}(I + MK)^{-1} MK \tag{25}
\]

\[
C = \hat{K}(I + MK)^{-1} MK - \hat{K}(I + MK)^{-1} MK \tag{26}
\]

we obtain the following,

\[
||\hat{K}(I + MK)^{-1} MK - K(I + MK)^{-1} MK|| \leq ||\hat{K}(I + MK)^{-1} MK - K(I + MK)^{-1} MK|| + ||\hat{K}(I + MK)^{-1} MK - \hat{K}(I + MK)^{-1} MK|| + ||\hat{K}(I + MK)^{-1} MK - \hat{K}(I + MK)^{-1} MK|| \tag{27}
\]

For \( ||A|| \) we obtain the following bound,

\[
||\hat{K}(I + MK)^{-1} MK - K(I + MK)^{-1} MK|| \leq ||\hat{K} - K||(I + MK)^{-1} MK|| \tag{30}
\]

For \( ||B|| \) we obtain the following bound,

\[
||\hat{K}(I + MK)^{-1} MK - \hat{K}(I + MK)^{-1} MK|| = ||\hat{K}(I + MK)^{-1} MK - K(I + MK)^{-1} MK|| \tag{31}
\]

\[
\leq ||\hat{K}(I + MK)^{-1} MK - K(I + MK)^{-1} MK|| \tag{32}
\]

\[
= ||\hat{K}M - \hat{K}M(I + \hat{K} M)^{-1} \hat{K}M|| ||\hat{K} - K||(I + MK)^{-1} MK|| \tag{33}
\]

In order to obtain eq. (31) we apply the identity \( XYZ^{-1} Y = XW^{-1} Y = XZ^{-1}(W - Z)W^{-1} Y \) with \( W = I + MK, X = \hat{K}, Y = MK \) and \( Z = I + M\hat{K} \). To reach (33) we apply the SMW identity; for eq. (34) we apply the identity \( Q - Q(I + Q)^{-1} Q = Q(I + Q)^{-1} \) with \( Q = \hat{K}M \).

For \( ||C|| \) we have the following bound,

\[
||\hat{K}(I + MK)^{-1} MK - \hat{K}(I + MK)^{-1} MK|| \leq ||\hat{K}(I + MK)^{-1} MK - \hat{K}|| \tag{35}
\]

\[
= ||\hat{K}M - \hat{K}M(I + \hat{K}M)^{-1} \hat{K}M|| ||\hat{K} - \hat{K}|| \tag{36}
\]

\[
= ||\hat{K}M(I + \hat{K}M)^{-1}|| ||\hat{K}|| \tag{37}
\]
For eqs. (36) and (37) we follow the same proof as for eqs. (33) and (34). We will focus on the first term of the right side of (37).

\[ \| \tilde{K} M (I + \tilde{K} M)^{-1} \| \leq \| \tilde{K} \| \| M \| \| (I + \tilde{K} M)^{-1} \| \] (38)

We seek to provide a bound for \( \| (I + \tilde{K} M)^{-1} \| \). We know that \( \sigma_{\text{max}}((I + \tilde{K} M)^{-1}) = \frac{1}{\sigma_{\text{min}}(I + \tilde{K} M)} \), with \( \sigma_{\text{max}}(.) \) and \( \sigma_{\text{min}}(.) \) being the maximum and minimum singular values, respectively. From [6] (with direct reference to their eq. 3.12) we can write the following inequality (which is valid for any non-singular complex matrix of order \( N \), in our case \( I + \tilde{K} M \)), with \( \| . \|_F \) being the Frobenius norm

\[ \sigma_{\text{min}}(I + \tilde{K} M) \geq | \det(I + \tilde{K} M) | \left( \frac{\sqrt{N - 1}}{\| I + \tilde{K} M \|_F} \right)^{N-1} \] (39)

For \( | \det(I + \tilde{K} M) | \) we have the following bound, where \( \lambda_i(.) \) is the \( i^{th} \) eigenvalue

\[ | \det(I + \tilde{K} M) | = | \prod_i \lambda_i(I + \tilde{K} M) | \] (40)
\[ = | \prod_i (1 + \lambda_i(\tilde{K} M)) | \] (41)
\[ \geq 1 \] (42)

The last inequality results due to the fact that \( \tilde{K} M \) is positive semi-definite. Thus, (39) becomes

\[ \sigma_{\text{min}}(I + \tilde{K} M) \geq \left( \frac{\sqrt{N - 1}}{\| I + \tilde{K} M \|_F} \right)^{N-1} \] (43)
\[ \sigma_{\text{max}}((I + \tilde{K} M)^{-1}) \leq \left( \frac{\sqrt{N - 1}}{\| I + \tilde{K} M \|_F} \right)^{N-1} \] (44)

We know that the right hand side of (44) is bounded, as \( N \) is the number of data samples, and \( \tilde{K} M \) is positive semi-definite. Given the bounds of \( \| A \|, \| B \| \) and \( \| C \| \), we substitute them in (23). Applying the expectations on both sides, leads to the claim.

**Proposition 3** Given the results before we can claim \( \mathbb{E} \| \tilde{K} - \tilde{K} \| \) \( \leq \left( \sqrt{\frac{3N^2 \log N}{d}} + \frac{2N \log N}{d} \right) (1 + T + RT + R) \)

**Proof** Given the bound for \( \mathbb{E} \| \tilde{K} - \tilde{K} \| \), the claim for deviation is

\[ \mathbb{E} \| \tilde{K} - \tilde{K} \| \leq \mathbb{E} \| \tilde{K} - K \| (1 + T + RT + R) \] (45)
\[ \leq \left( \sqrt{\frac{3N^2 \log N}{d}} + \frac{2N \log N}{d} \right) (1 + T + RT + R) \] by (3) (46)

Finally note that a convergence rate immediately follows once \( T \) and \( R \) are determined. However, these will depend on the explicit forms of \( K \) and \( M \), which is beyond the scope of this analysis.
References


