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 μ a positive function such that

$$k(x, y) = \int_\omega e^{i\omega^\top(x-y)}\mu(\omega)$$

Since μ 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 $\phi(x) = [\phi_1(x) \dots \phi_d(x)]$, where $\phi_i(x) = \cos(\omega_i^\top x + b_i)$ (where $b_i \sim \texttt{Uniform}[0, 2\pi]$). Let $\widehat{k}(x, y) = \sum_i^d \widehat{k}_i(x, y) = \frac{1}{d}\sum_i^d \phi_i(x)\phi_i(y)^\top = \frac{1}{d}\phi(x)\phi(y)^\top$. 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 $\widehat{K}$, $\widehat{K}_i$, $\Phi_i (= \phi_i(X))$ and $\Phi (= \phi(X))$.

With this notation we have

$$\widehat{K} = \sum_i^d \widehat{K}_i = \sum_i^d \Phi_i\Phi_i^\top = \Phi\Phi^\top \tag{1}$$

We notice that $\widehat{K}_i$ are i.i.d. thus matrix concentration results apply to it.

To this end we want to use

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

$$\mathbb{E}\|\sum_i Z_i\| \le \sigma\sqrt{3\log(2n)} + R\log(2n) \tag{2}$$

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

$$\mathbb{E}\|\widehat{K} - K\| \le \sqrt{\frac{3N^2\log N}{d}} + \frac{2N\log N}{d} \tag{3}$$

Proof [1] Then $\widehat{K}_i$ are independent and we know that $\mathbb{E}[\widehat{K}] = K$.

$$E = \widehat{K} - K = \sum_i^d E_i, \quad E_i = \frac{1}{d}(\widehat{K}_i - K) \tag{4}$$

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

First we must show that each are bounded

$$\|E_i\| = \frac{1}{d}\|\Phi_i\Phi_i^\top - \mathbb{E}[\Phi\Phi^\top]\| \le \frac{1}{d}(\|\Phi_i\|^2 + \mathbb{E}[\|\Phi\|^2] \le \frac{1}{d}(\|\Phi_i\|^2 + \|\mathbb{E}[\Phi]\|^2) \le \frac{2B}{d} \tag{5}$$

[1] This is from [3] reproduced for a self-contained understanding of our main results.

where we used first the definitions of $\widehat{K}_i$ and K, 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 ϕ is constructed.

Then the variance of E_i is

$$\mathbb{E}[E_i^2] = \frac{1}{d^2}\mathbb{E}[(\Phi_i\Phi^\top - K)^2] \tag{6}$$

$$= \frac{1}{d^2}\mathbb{E}[(\|\Phi_i\|^2\Phi_i\Phi_i^\top - \Phi_i\Phi_i^\top K - K\Phi_i\Phi_i^\top + K^2)] \tag{7}$$

$$\preccurlyeq \frac{1}{d^2}[BK - 2K^2 + K^2] \preccurlyeq \frac{BK}{d^2} \tag{8}$$

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

$$\|\mathbb{E}[E^2]\| \leq \|\sum_i^d \mathbb{E}[E_i^2]\| \leq \frac{1}{d}B\|K\| \tag{9}$$

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

$$\mathbb{E}\|\widehat{K} - K\| \leq \sqrt{\frac{3B\|K\|\log N}{d}} + \frac{2B\log N}{d} \tag{10}$$

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

$$\widetilde{K} = K - K(I + MK)^{-1}MK \tag{11}$$

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

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

To this end we define

$$\overline{K} = \widehat{K} - \widehat{K}(I + M\widehat{K})^{-1}M\widehat{K} \tag{12}$$

and

$$\breve{K} = \Phi(I + \Phi^\top M\Phi)^{-1}\Phi^\top \tag{13}$$

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

$$(I + UV^\top)^{-1} = I - U(I + V^\top U)^{-1}V^\top \tag{14}$$

Proposition 2 *With the definitions above*

$$\overline{K} = \breve{K} \tag{15}$$

Proof

$$\overline{K} = \widehat{K} - \widehat{K}(I + M\widehat{K})^{-1}M\widehat{K} \tag{16}$$

$$= \Phi\Phi^\top - \Phi\Phi^\top(I + M\Phi\Phi^\top)^{-1}M\Phi\Phi^\top \qquad \text{by (1)} \tag{17}$$

$$= \Phi(I - \Phi^\top(I + M\Phi\Phi^\top)^{-1}M\Phi)\Phi^\top \tag{18}$$

$$= \Phi(I + \Phi^\top M\Phi)^{-1}\Phi^\top \tag{19}$$

$$= \breve{K} \qquad \text{by (13)} \tag{20}$$

Where (19) comes by applying (14) with $U = \Phi^\top$ and $V = \Phi^\top M$ and using the symmetry of M.

So $\widetilde{\Phi} = \Phi(I + \Phi^\top M\Phi)^{-1/2}$ but given (15) we can use $\overline{K}$ instead of $\breve{K}$ for the convergence proofs. Now the goal is to obtain a bound on $\mathbb{E}\|\overline{K} - \widetilde{K}\|$.

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

$$\mathbb{E}\|\overline{K} - \widetilde{K}\| \leq \mathbb{E}\|K - \widehat{K}\|(1 + T + RT + R) \tag{21}$$

Proof

$$\|\overline{K} - \widetilde{K}\| = \|\widehat{K} - \widehat{K}(I + M\widehat{K})^{-1}M\widehat{K} - K + K(I + MK)^{-1}MK\| \tag{22}$$

$$\leq \|\widehat{K} - K\| + \|\widehat{K}(I + M\widehat{K})^{-1}M\widehat{K} - K(I + MK)^{-1}MK\| \tag{23}$$

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,

$$A = \widehat{K}(I + MK)^{-1}MK - K(I + MK)^{-1}MK \tag{24}$$

$$B = \widehat{K}(I + M\widehat{K})^{-1}MK - \widehat{K}(I + MK)^{-1}MK \tag{25}$$

$$C = \widehat{K}(I + M\widehat{K})^{-1}M\widehat{K} - \widehat{K}(I + M\widehat{K})^{-1}MK \tag{26}$$

we obtain the following,

$$\|\widehat{K}(I + M\widehat{K})^{-1}M\widehat{K} - K(I + MK)^{-1}MK\| \leq \|\widehat{K}(I + MK)^{-1}MK - K(I + MK)^{-1}MK\| \tag{27}$$

$$+ \|\widehat{K}(I + M\widehat{K})^{-1}MK - \widehat{K}(I + MK)^{-1}MK\| \tag{28}$$

$$+ \|\widehat{K}(I + M\widehat{K})^{-1}M\widehat{K} - \widehat{K}(I + M\widehat{K})^{-1}MK\| \tag{29}$$

For $\|A\|$ we obtain the following bound,

$$\|\widehat{K}(I + MK)^{-1}MK - K(I + MK)^{-1}MK\| \leq \|\widehat{K} - K\|\|(I + MK)^{-1}MK\| \tag{30}$$

For $\|B\|$ we obtain the following bound,

$$\|\widehat{K}(I + M\widehat{K})^{-1}MK - \widehat{K}(I + MK)^{-1}MK\| = \|\widehat{K}(I + M\widehat{K})^{-1}M(\widehat{K} - K)(I + MK)^{-1}MK\| \tag{31}$$

$$\leq \|\widehat{K}(I + M\widehat{K})^{-1}M\|\|\widehat{K} - K\|\|(I + MK)^{-1}MK\| \tag{32}$$

$$= \|\widehat{K}M - \widehat{K}M(I + \widehat{K}M)^{-1}\widehat{K}M\|\|\widehat{K} - K\|\|(I + MK)^{-1}MK\| \tag{33}$$

$$= \|\widehat{K}M(I + \widehat{K}M)^{-1}\|\|\widehat{K} - K\|\|(I + MK)^{-1}MK\| \tag{34}$$

In order to obtain eq. (31) we apply the identity $XZ^{-1}Y - XW^{-1}Y = XZ^{-1}(W - Z)W^{-1}Y$ with $W = I + MK$, $X = \widehat{K}$, $Y = MK$ and $Z = I + M\widehat{K}$. 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 $Q = \widehat{K}M$.

For $\|C\|$ we have the following bound,

$$\|\widehat{K}(I + M\widehat{K})^{-1}M\widehat{K} - \widehat{K}(I + M\widehat{K})^{-1}MK\| \leq \|\widehat{K}(I + M\widehat{K})^{-1}M\|\|K - \widehat{K}\| \tag{35}$$

$$= \|\widehat{K}M - \widehat{K}M(I + \widehat{K}M)^{-1}\widehat{K}M\|\|K - \widehat{K}\| \tag{36}$$

$$= \|\widehat{K}M(I + \widehat{K}M)^{-1}\|\|K - \widehat{K}\| \tag{37}$$

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).

$$\|\widehat{K}M(I+\widehat{K}M)^{-1}\| \leq \|\widehat{K}\|\|M\|\|(I+\widehat{K}M)^{-1}\| \tag{38}$$

We seek to provide a bound for $\|(I+\widehat{K}M)^{-1}\|$. We know that $\sigma_{max}((I+\widehat{K}M)^{-1}) = \frac{1}{\sigma_{min}(I+\widehat{K}M)}$, with $\sigma_{max}(.)$ and $\sigma_{min}(.)$ 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 $I+\widehat{K}M$), with $\|.\|_F$ being the Frobenius norm

$$\sigma_{min}(I+\widehat{K}M) \geq \left|\det(I+\widehat{K}M)\right| \left(\frac{\sqrt{N-1}}{\|I+\widehat{K}M\|_F}\right)^{N-1} \tag{39}$$

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

$$\begin{aligned} \left|\det(I+\widehat{K}M)\right| &= \left|\prod_i \lambda_i(I+\widehat{K}M)\right| && (40)\\ &= \left|\prod_i (1+\lambda_i(\widehat{K}M))\right| && (41)\\ &\geq 1 && (42) \end{aligned}$$

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

$$\begin{aligned} \sigma_{min}(I+\widehat{K}M) &\geq \left(\frac{\sqrt{N-1}}{\|I+\widehat{K}M\|_F}\right)^{N-1} && (43)\\ \sigma_{max}((I+\widehat{K}M)^{-1}) &\leq \left(\frac{\|I+\widehat{K}M\|_F}{\sqrt{N-1}}\right)^{N-1} && (44) \end{aligned}$$

We know that the right hand side of (44) is bounded, as N is the number of data samples, and $\widehat{K}M$ 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* $\mathbb{E}\|\overline{K}-\widetilde{K}\| \leq \left(\sqrt{\frac{3N^2\log N}{d}} + \frac{2N\log N}{d}\right)(1+T+RT+R)$

Proof Given the bound for $\mathbb{E}\|\widehat{K}-K\|$, the claim for deviation is

$$\begin{aligned} \mathbb{E}\|\overline{K}-\widetilde{K}\| &\leq \mathbb{E}\|\widehat{K}-K\|(1+T+RT+R) && (45)\\ &\leq \left(\sqrt{\frac{3N^2\log N}{d}} + \frac{2N\log N}{d}\right)(1+T+RT+R) \quad \text{by (3)} && (46) \end{aligned}$$

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

[1] C. Ionescu, A.-I. Popa, and C. Sminchisescu, "Large-scale data-dependent kernel approximation," in *AISTATS*, 2017.

[2] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *NIPS*, 2007.

[3] D. Lopez-Paz, S. Sra, A. Smola, Z. Ghahramani, and B. Schölkopf, "Randomized nonlinear component analysis," *arXiv preprint arXiv:1402.0119*, 2014.

[4] L. Mackey, M. I. Jordan, R. Y. Chen, B. Farrell, J. A. Tropp, *et al.*, "Matrix concentration inequalities via the method of exchangeable pairs," *The Annals of Probability*, vol. 42, no. 3, pp. 906–945, 2014.

[5] V. Sindhwani, P. Niyogi, and M. Belkin, "Beyond the point cloud: from transductive to semi-supervised learning," in *ICML*, 2005.

[6] H.-B. Li, T.-Z. Huang, and H. Li, "Some new results on determinantal inequalities and applications," *Journal of Inequalities and Applications*, vol. 2010, no. 1, p. 1, 2010.