Appendix for Classification of Sparse and Irregularly Sampled Time Series with Mixtures of Expected Gaussian Kernels and Random Features

Steven Cheng-Xian Li **Benjamin Marlin**
University of Massachusetts Amherst
Amherst, MA 01003
{cxl,marlin}@cs.umass.edu

A Derivation of Expected Gaussian Kernels

Let $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the Gaussian density function

$$(2\pi)^{-D/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),$$

where D is the dimensionality of the random variable $\mathbf{x}$.

We can verify that the integral of the product of two Gaussians is in the form of another Gaussian:

$$\int \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\,\mathrm{d}\mathbf{x} = (2\pi)^{-D/2}|\widetilde{\boldsymbol{\Sigma}}|^{-1/2} \exp\left(-\frac{1}{2}\widetilde{\boldsymbol{\mu}}^\top \widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{\mu}}\right) \quad (9)$$

where $\widetilde{\boldsymbol{\mu}} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$ and $\widetilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j$.

Note that (9) can be expressed in several equivalent ways

$$\mathcal{N}(\boldsymbol{\mu}_i; \boldsymbol{\mu}_j, \widetilde{\boldsymbol{\Sigma}}) = \mathcal{N}(\boldsymbol{\mu}_j; \boldsymbol{\mu}_i, \widetilde{\boldsymbol{\Sigma}}) = \mathcal{N}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j; \mathbf{0}, \widetilde{\boldsymbol{\Sigma}}).$$

Applying (9) twice with the rearrangement above, we have

$$\iint \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\mathcal{N}(\mathbf{x}_j; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\mathcal{N}(\mathbf{x}_i; \mathbf{x}_j, \boldsymbol{\Sigma})\,\mathrm{d}\mathbf{x}_i\,\mathrm{d}\mathbf{x}_j = \mathcal{N}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j; \mathbf{0}, \boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j + \boldsymbol{\Sigma}).$$

This double integral is actually $\mathbb{E}_{\mathbf{x}_i\mathbf{x}_j}\mathcal{N}(\mathbf{x}_i; \mathbf{x}_j, \boldsymbol{\Sigma})$ given that $\mathbf{x}_i$ and $\mathbf{x}_j$ are independently Gaussian distributed. Therefore, the expected Gaussian kernel can be computed as following:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}_i\mathbf{x}_j}\left[\exp\Big(-\frac{1}{2}(\mathbf{x}_i-\mathbf{x}_j)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\mathbf{x}_j)\Big)\right] \\
&\quad = \mathbb{E}_{\mathbf{x}_i\mathbf{x}_j}\left[(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}\mathcal{N}(\mathbf{x}_i; \mathbf{x}_j, \boldsymbol{\Sigma})\right] \\
&\quad = (2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}\mathcal{N}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j; \mathbf{0}, \boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j + \boldsymbol{\Sigma}) \\
&\quad = \sqrt{\frac{|\boldsymbol{\Sigma}|}{|\widetilde{\boldsymbol{\Sigma}}|}}\exp\left(-\frac{1}{2}\widetilde{\boldsymbol{\mu}}^\top\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{\mu}}\right)
\end{aligned}$$

where $\widetilde{\boldsymbol{\mu}} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$ and $\widetilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j + \boldsymbol{\Sigma}$.

B Derivation of Random Fourier Features for Expected Gaussian Kernels

Due to the independence assumption, the expected Gaussian kernel can be approximated as the follows

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}_i\mathbf{x}_j}\mathcal{K}_{\mathrm{G}}(\mathbf{x}_i, \mathbf{x}_j) &\approx \mathbb{E}_{\mathbf{x}_i\mathbf{x}_j}\left[\mathbf{z}(\mathbf{x}_i)^\top\mathbf{z}(\mathbf{x}_j)\right] \\
&= \mathbb{E}_{\mathbf{x}_i}[\mathbf{z}(\mathbf{x}_i)]^\top\,\mathbb{E}_{\mathbf{x}_j}[\mathbf{z}(\mathbf{x}_j)],
\end{aligned}$$

in which the ith entry of $\mathbb{E}_{\mathbf{x}}\mathbf{z}(\mathbf{x})$ is $\sqrt{\frac{2}{m}}\,\mathbb{E}_{\mathbf{x}}\cos(\mathbf{w}_i^\top\mathbf{x}+b_i)$.

Consider the following expectation

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[e^{i(\mathbf{w}^\top\mathbf{x}+b)}\right] \\
&= e^{ib}\,\mathbb{E}_{\mathbf{x}}\left[e^{i\mathbf{w}^\top\mathbf{x}}\right] \\
&= e^{ib}e^{i\mathbf{w}^\top\boldsymbol{\mu}-\frac{1}{2}\mathbf{w}^\top\boldsymbol{\Sigma}\mathbf{w}} \\
&= e^{-\frac{1}{2}\mathbf{w}^\top\boldsymbol{\Sigma}\mathbf{w}+i(\mathbf{w}^\top\boldsymbol{\mu}+b)} \\
&= e^{-\frac{1}{2}\mathbf{w}^\top\boldsymbol{\Sigma}\mathbf{w}}\Big(\cos(\mathbf{w}^\top\boldsymbol{\mu}+b) + i\sin(\mathbf{w}^\top\boldsymbol{\mu}+b)\Big). \quad (10)
\end{aligned}$$

In the second step, we use the analytic form of the characteristic function for Gaussian random vectors.

Since

$$\mathbb{E}\left[e^{i(\mathbf{w}^\top\mathbf{x}+b)}\right] = \mathbb{E}\left[\cos(\mathbf{w}^\top\mathbf{x}+b) + i\sin(\mathbf{w}^\top\mathbf{x}+b)\right],$$

we know that $\mathbb{E}\left[\cos(\mathbf{w}^\top\mathbf{x}+b)\right]$ is the real part of $\mathbb{E}\left[e^{i(\mathbf{w}^\top\mathbf{x}+b)}\right]$. Therefore, from (10) we have

$$\mathbb{E}\left[\cos(\mathbf{w}^\top\mathbf{x}+b)\right] = \exp\left(-\frac{1}{2}\mathbf{w}^\top\boldsymbol{\Sigma}\mathbf{w}\right)\cos(\mathbf{w}^\top\boldsymbol{\mu}+b).$$

C Proof of Theorem 1

To analyze the concentration of the kernel approximation, we apply the Hermitian matrix Bernstein inequality [Tropp, 2012]. Note that $\|\cdot\|$ denotes the spectral norm when taking on a matrix, and the L^2 norm when taking on a vector.

Theorem 2. *(Matrix Bernstein: Hermitian Case [Tropp, 2012]). Consider a finite sequence* $\{\mathbf{X}_k\}$ *of independent random Hermitian matrices with dimension* d*. Assume that* $\mathbb{E}\mathbf{X}_k = \mathbf{0}$ *and* $\lambda_{\max}(\mathbf{X}_k) \le R$ *for all* k*. Let* $\mathbf{Y} = \sum_k \mathbf{X}_k$*. Define the variance parameter* $\sigma^2 = \|\mathbb{E}(\mathbf{Y}^2)\|$*. Then*

$$\mathbb{E}\lambda_{\max}(\mathbf{Y}) \le \sqrt{2\sigma^2 \log d} + \frac{1}{3} R \log d.$$

Proof of Theorem 1. We follow the derivation of Lopez-Paz et al. [2014] with refinement to obtain a tighter bound.

Let the n-dimensional random vector $\mathbf{z}_k = [z_{k1}, \ldots, z_{kn}]^\top$ denote the collection of the kth random feature (sharing the same random projection parameters, $\mathbf{w}_k, b_k$) of each of the n examples. Let $\mathbf{S}_k = \mathbf{z}_k \mathbf{z}_k^\top / m$. The approximate kernel can be expressed as the sum of m independent matrices $\widehat{\mathbf{K}} = \sum_{k=1}^m \mathbf{S}_k$.

According to Rahimi and Recht [2007], the random Fourier feature is unbiased. Specifically, for $z(\mathbf{x}) = \sqrt{2}\cos(\mathbf{w}^\top \mathbf{x} + b)$ where $\mathbf{w}$ draws from the distribution induced by the kernel and $b \sim \text{uniform}(0, 2\pi)$, we have

$$\mathcal{K}_{\text{G}}(\mathbf{x}_i, \mathbf{x}_j) = \mathbb{E}_{\mathbf{w},b}[z(\mathbf{x}_i)^\top z(\mathbf{x}_j)]. \tag{11}$$

As a result, the random feature for the expected Gaussian kernel is also unbiased as shown below. Therefore, when m random features are used, we have $\mathbb{E}\mathbf{S}_k = \mathbf{K}/m$ and $\mathbb{E}\widehat{\mathbf{K}} = \mathbf{K}$.

$$\begin{aligned}\mathcal{K}_{\text{EG}}(\mathcal{N}_i, \mathcal{N}_j) &= \mathbb{E}_{\mathbf{x}_i \sim \mathcal{N}_i, \mathbf{x}_j \sim \mathcal{N}_j} \mathbb{E}_{\mathbf{w},b}[z(\mathbf{x}_i)^\top z(\mathbf{x}_j)] \\ &= \mathbb{E}_{\mathbf{w},b}\left[\mathbb{E}_{\mathbf{x}_i \sim \mathcal{N}_i}[z(\mathbf{x}_i)]^\top \mathbb{E}_{\mathbf{x}_j \sim \mathcal{N}_j}[z(\mathbf{x}_j)]\right],\end{aligned}$$

where $\mathbb{E}_{\mathbf{x}}[z(\mathbf{x})]$ is in the form of $\sqrt{2/m}\,\mathbb{E}_{\mathbf{x}}[\cos(\mathbf{w}^\top \mathbf{x} + b)]$ with its absolute value bounded by $\sqrt{2/m}$. As a result, there exists a constant B such that $\|\mathbf{z}_k\|^2 \le B \le 2n/m$.

The error matrix $\widehat{\mathbf{K}} - \mathbf{K}$ can then be expressed as the sum of m independent zero-mean matrices:

$$\widehat{\mathbf{K}} - \mathbf{K} = \sum_{k=1}^m (\mathbf{S}_k - \mathbb{E}\mathbf{S}_k).$$

Since $\widehat{\mathbf{K}} - \mathbf{K}$ is symmetric, the singular values are the absolute values of its eigenvalues. Therefore,

$$\begin{aligned}\|\widehat{\mathbf{K}} - \mathbf{K}\| &= \max\left\{\lambda_{\max}(\widehat{\mathbf{K}} - \mathbf{K}), -\lambda_{\min}(\widehat{\mathbf{K}} - \mathbf{K})\right\} \\ &= \max\left\{\lambda_{\max}(\widehat{\mathbf{K}} - \mathbf{K}), \lambda_{\max}(\mathbf{K} - \widehat{\mathbf{K}})\right\}.\end{aligned}$$

In order to apply matrix Bernstein inequality, we need to bound both $\lambda_{\max}(\mathbf{S}_k - \mathbb{E}\mathbf{S}_k)$ and $\lambda_{\max}(\mathbb{E}\mathbf{S}_k - \mathbf{S}_k)$.

$$\lambda_{\max}(\mathbf{S}_k - \mathbb{E}\mathbf{S}_k) \le \lambda_{\max}(\mathbf{S}_k) = \|\mathbf{S}_k\| = \frac{1}{m}\|\mathbf{z}_k\|^2 \le \frac{B}{m}.$$

The first relation holds because both $\mathbf{S}_k$ and $\mathbb{E}\mathbf{S}_k$ are symmetric and positive semidefinite[5]. Similarly,

$$\lambda_{\max}(\mathbb{E}\mathbf{S}_k - \mathbf{S}_k) \le \lambda_{\max}(\mathbb{E}\mathbf{S}_k) = \frac{1}{m}\|\mathbf{K}\| \le \frac{B}{m}$$

where we bound $\|\mathbf{K}\|$ using Jensen's inequality:

$$\|\mathbf{K}\| = \|\mathbb{E}[\mathbf{z}\mathbf{z}^\top]\| \le \mathbb{E}\|\mathbf{z}\mathbf{z}^\top\| = \mathbb{E}[\|\mathbf{z}\|^2] \le B.$$

To compute the variance parameter σ^2, we start with the expectation $\mathbb{E}[(\widehat{\mathbf{K}} - \mathbf{K})^2]$:

$$\begin{aligned}\mathbb{E}[(\widehat{\mathbf{K}} - \mathbf{K})^2] &= \mathbb{E}[\widehat{\mathbf{K}}^2] - \mathbf{K}^2 \\ &= \mathbb{E}\left[\left(\sum_{k=1}^m \mathbf{S}_k\right)^2\right] - \mathbf{K}^2 \\ &= \sum_{i=1}^m \sum_{j=1}^m \mathbb{E}\left[\mathbf{S}_i \mathbf{S}_j\right] - \mathbf{K}^2 \\ &= \left(\sum_{k=1}^m \mathbb{E}\left[\mathbf{S}_k^2\right]\right) + \frac{m^2 - m}{m^2}\mathbf{K}^2 - \mathbf{K}^2 \\ &= \left(\sum_{k=1}^m \mathbb{E}\left[\mathbf{S}_k^2\right]\right) - \frac{1}{m}\mathbf{K}^2\end{aligned}$$

where

$$\mathbb{E}[\mathbf{S}_k^2] = \mathbb{E}\left[\left(\frac{1}{m}\mathbf{z}_k\mathbf{z}_k^\top\right)^2\right] = \frac{1}{m^2}\mathbb{E}\left[\|\mathbf{z}_k\|^2 \mathbf{z}_k \mathbf{z}_k^\top\right] \preccurlyeq \frac{B\mathbf{K}}{m^2}$$

in which the expression $\mathbf{A} \preccurlyeq \mathbf{B}$ means that $\mathbf{B} - \mathbf{A}$ is positive semidefinite. Therefore,

$$\mathbb{E}[(\widehat{\mathbf{K}} - \mathbf{K})^2] \preccurlyeq \left(\sum_{k=1}^m \frac{B\mathbf{K}}{m^2}\right) - \frac{\mathbf{K}^2}{m} = \frac{B\mathbf{K}}{m} - \frac{\mathbf{K}^2}{m} \preccurlyeq \frac{B\mathbf{K}}{m}.$$

The last step holds due to $\mathbf{K}^2 \succcurlyeq \mathbf{0}$. Since both $\mathbf{K}$ and $\mathbb{E}[(\widehat{\mathbf{K}} - \mathbf{K})^2]$ are symmetric and positive semidefinite, we have

$$\sigma^2 = \|\mathbb{E}[(\widehat{\mathbf{K}} - \mathbf{K})^2]\| \le \frac{B\|\mathbf{K}\|}{m}.$$

Given that $\lambda_{\max}(\mathbf{S}_k - \mathbb{E}\mathbf{S}_k)$ and $\lambda_{\max}(\mathbb{E}\mathbf{S}_k - \mathbf{S}_k)$ are both bounded by B/m, we obtain the same bound for $\mathbb{E}\lambda_{\max}(\widehat{\mathbf{K}} - \mathbf{K})$ and $\mathbb{E}\lambda_{\max}(\mathbf{K} - \widehat{\mathbf{K}})$ when plugging $R \le B/m$ and $\sigma^2 \le B\|\mathbf{K}\|/m$ into the matrix Bernstein inequality in Theorem 2. This leads to the bound on the expected norm:

$$\mathbb{E}\|\widehat{\mathbf{K}} - \mathbf{K}\| \le \sqrt{\frac{2B\|\mathbf{K}\| \log n}{m}} + \frac{B \log n}{3m}.$$

With $\|\mathbf{K}\| \le B \le 2n/m$, we attain

$$\mathbb{E}\|\widehat{\mathbf{K}} - \mathbf{K}\| \le \frac{2n}{m}\sqrt{\frac{2\log n}{m}} + \frac{2n \log n}{3m^2}.$$

□

[5] Given Hermitian positive semidefinite matrices $\mathbf{A}$ and $\mathbf{B}$, let $\mathbf{u} = \operatorname{argmax}_{\|\mathbf{v}\|=1} \mathbf{v}^\top(\mathbf{B} - \mathbf{A})\mathbf{v}$, then $\lambda_{\max}(\mathbf{B} - \mathbf{A}) = \mathbf{u}^\top(\mathbf{B} - \mathbf{A})\mathbf{u} \le \mathbf{u}^\top \mathbf{B}\mathbf{u} \le \lambda_{\max}(\mathbf{B}) = \|\mathbf{B}\|$.

References

Lopez-Paz, D., Sra, S., Smola, A. J., Ghahramani, Z., and Schölkopf, B. (2014). Randomized nonlinear component analysis. In *ICML*.

Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184.

Tropp, J. A. (2012). User-friendly tools for random matrices: An introduction. Technical report, DTIC Document.