Supplementary material: Harmonizable mixture kernels with variational Fourier features

Zheyang Shen Markus Heinonen Samuel Kaski

Aalto University
Helsinki Institute for Information Technology HIIT

1 Proof of theorem 2

In this section, we prove the expressiveness of stationary spectral kernels.

Theorem 1. *Let h be a complex-valued positive definite, continuous and integrable function. Than the family of* generalized spectral kernels

$$k_{GS}(\boldsymbol{\tau}) = \sum_{q=1}^{Q} \alpha_q h(\boldsymbol{\tau} \circ \boldsymbol{\gamma}_q) e^{2i\pi \boldsymbol{\omega}_q^\top \boldsymbol{\tau}}. \tag{1}$$

with $\circ$ denoting the Hadamard product, $\alpha_q \in \mathbb{R}_+$, $\boldsymbol{\omega}_k \in \mathbb{R}^D$, $\boldsymbol{\gamma}_k \in \mathbb{R}_+^D$, $Q \in \mathbb{N}_+$ is dense in the family of stationary, complex-valued kernels with respect to pointwise convergence of functions.

Proof. We know from the uniform convergence of random Fourier features (Rahimi and Recht, 2008), that for an arbitrary stationary kernel $k_0(\mathbf{x}, \mathbf{x}') = k_0(\mathbf{x} - \mathbf{x}')$, for all compact subset $\mathcal{M} \in \mathbb{R}^D$, and for all $\epsilon > 0$, there exists a feature map $\zeta_\omega(\mathbf{x}) = \left(\alpha_q e^{2\pi \boldsymbol{\omega}_q^\top \mathbf{x}}\right)_{q=1}^{Q}$, such that $|\zeta_\omega(\mathbf{x})\zeta_\omega(\mathbf{x}')^* - k_0(\mathbf{x} - \mathbf{x}')| < \epsilon$. The uniform convergence of random Fourier features suggests the expressiveness of a generalized form of sparse spectrum kernel $k_{\text{SS}}(\mathbf{x} - \mathbf{x}') = \sum_{q=1}^{Q} \alpha_q e^{2\pi \boldsymbol{\omega}_q^\top (\mathbf{x} - \mathbf{x}')}$.

For an arbitrary continuous, integrable kernel h, consider the function $\tilde{k}(\boldsymbol{\tau}) = \dfrac{h(\boldsymbol{\tau} \circ \boldsymbol{\gamma})}{h(\mathbf{0})} k_{\text{SS}}(\boldsymbol{\tau}), \boldsymbol{\gamma} \succeq \mathbf{0}$. Because of the continuity of function h, $\tilde{k}$ uniformly approximates k_{SS} as $\boldsymbol{\gamma} \downarrow \mathbf{0}$, and thus can be used to approximate any stationary covariance k_0.

$\tilde{k}(\boldsymbol{\tau})$ uniformly approximates any stationary kernel k_0 on arbitrary compact subset $\mathcal{M}$ of $\mathbb{R}^D$. We can therefore construct a sequence of $\tilde{k}_n$ by setting $\epsilon_n = \dfrac{1}{n}$, $\mathcal{M}_n = \mathcal{B}(0, n) = \{v | \, \|v\| \leq n\}$, $n = 1, 2, 3, \cdots$. $\{\tilde{k}_n\}_{n=1}^{\infty}$ converges pointwise to k_0. k_{GS} takes a more general form, and thus has the same level of expressiveness as $\tilde{k}$. □

We can see from the reasoning that sparse spectrum kernel and spectral mixture kernel both weakly span stationary covariances, and thus sharing the same level of expressiveness. But the sparse spectrum kernel only encodes a finite dimensional feature mapping, which reduces a GP regression with a sparse spectrum kernel to a Bayesian linear regression with trigonometric basis expansions. The spectral mixture kernel alleviates overfitting by using Gaussian mixture on the spectral distribution, which implicitly assumes certain level of smoothness of the unknown spectral distribution being modeled – the Gaussian mixture also leads to an infinite-dimensional feature mapping which does not render a GP regression degenerate.

2 Derivation of harmonizable mixture kernel

In this section we derive the parametric form of hramonizable mixture kernel. The GSD of a locally stationary Gaussian kernel follows a generalized Wiener-Khintchin relation, as noticed in (Silverman, 1957). This relation is easily noticed when subtituting $\mathbf{x}$ and $\mathbf{x}'$ with new variables $\widetilde{\mathbf{x}} = (\mathbf{x} + \mathbf{x}')/2$ and $\boldsymbol{\tau} = \mathbf{x} - \mathbf{x}'$.

$$k_{\text{LSG}}(\mathbf{x}, \mathbf{x}') = e^{-2\pi^2 \widetilde{\mathbf{x}}^\top \boldsymbol{\Sigma}_1 \widetilde{\mathbf{x}}} e^{-2\pi^2 \boldsymbol{\tau}^\top \boldsymbol{\Sigma}_2 \boldsymbol{\tau}}, \tag{2}$$

$$S_{k_{\text{LSG}}}(\boldsymbol{\omega}, \boldsymbol{\xi}) = \iint k_{\text{LSG}}(\mathbf{x}, \mathbf{x}') e^{-2i\pi(\boldsymbol{\omega}^\top \mathbf{x} - \boldsymbol{\xi}^\top \mathbf{x}')} \, \mathrm{d}\mathbf{x} \mathrm{d}\mathbf{x}' \tag{3}$$

$$= \iint e^{-2\pi^2 \widetilde{\mathbf{x}}^\top \boldsymbol{\Sigma}_1 \widetilde{\mathbf{x}} - 2i\pi(\boldsymbol{\omega} - \boldsymbol{\xi})^\top \widetilde{\mathbf{x}}} e^{-2\pi^2 \boldsymbol{\tau}^\top \boldsymbol{\Sigma}_2 \boldsymbol{\tau} - i\pi(\boldsymbol{\omega} + \boldsymbol{\xi})^\top \boldsymbol{\tau}} \, \mathrm{d}\widetilde{\mathbf{x}} \mathrm{d}\boldsymbol{\tau} \tag{4}$$

$$= \int e^{-2\pi^2 \widetilde{\mathbf{x}}^\top \boldsymbol{\Sigma}_1 \widetilde{\mathbf{x}} - 2i\pi(\boldsymbol{\omega} - \boldsymbol{\xi})^\top \widetilde{\mathbf{x}}} \, \mathrm{d}\widetilde{\mathbf{x}} \int e^{-2\pi^2 \boldsymbol{\tau}^\top \boldsymbol{\Sigma}_2 \boldsymbol{\tau} - i\pi(\boldsymbol{\omega} + \boldsymbol{\xi})^\top \boldsymbol{\tau}} \, \mathrm{d}\boldsymbol{\tau} \tag{5}$$

$$= \mathcal{N}\left(\boldsymbol{\omega} - \boldsymbol{\xi} \middle| 0, \boldsymbol{\Sigma}_1\right) \mathcal{N}\left(\frac{\boldsymbol{\omega} + \boldsymbol{\xi}}{2} \middle| 0, \boldsymbol{\Sigma}_2\right). \tag{6}$$

The Wigner transform of k_{LSG} is straightforward as the kernel factors into two parts.

$$W_{k_{\text{LSG}}}(\mathbf{x}, \boldsymbol{\omega}) = \int k\left(\mathbf{x} + \frac{\boldsymbol{\tau}}{2}, \mathbf{x} - \frac{\boldsymbol{\tau}}{2}\right) e^{-2i\pi \boldsymbol{\tau}^\top \boldsymbol{\omega}} \tag{7}$$

$$= e^{-2\pi^2 \mathbf{x}^\top \boldsymbol{\Sigma}_1 \mathbf{x}} \int e^{-2\pi^2 \boldsymbol{\tau}^\top \boldsymbol{\Sigma}_2 \boldsymbol{\tau} - 2i\pi \boldsymbol{\tau}^\top \boldsymbol{\omega}} \, \mathrm{d}\boldsymbol{\tau} \tag{8}$$

$$= e^{-2\pi^2 \mathbf{x}^\top \boldsymbol{\Sigma}_1 \mathbf{x}} \mathcal{N}(\boldsymbol{\omega} | 0, \boldsymbol{\Sigma}_2). \tag{9}$$

Now consider the *harmonizable mixture kernel*,

$$k_{\text{HM}}(\mathbf{x}, \mathbf{x}') = \sum_{p=1}^{P} k_p(\mathbf{x} - \mathbf{x}_p, \mathbf{x}' - \mathbf{x}_p), \tag{10}$$

$$k_p(\mathbf{x}, \mathbf{x}') = k_{\text{LSG}}(\mathbf{x} \circ \boldsymbol{\gamma}_p, \mathbf{x}' \circ \boldsymbol{\gamma}_p) \phi_p(\mathbf{x})^\top \mathbf{B}_p \phi_p(\mathbf{x}') \tag{11}$$

$$= k_{\text{LSG}}(\mathbf{x} \circ \boldsymbol{\gamma}_p, \mathbf{x}' \circ \boldsymbol{\gamma}_p) \sum_{1 \le i,j \le Q_p} e^{2i\pi(\boldsymbol{\mu}_{pi}^\top \mathbf{x} - \boldsymbol{\mu}_{pj}^\top \mathbf{x}')}. \tag{12}$$

We know from the Fourier transform $\hat{f}(\boldsymbol{\xi}) = \int f(\mathbf{x}) e^{-2i\pi \mathbf{x}^\top \boldsymbol{\xi}} \, \mathrm{d}\mathbf{x}$, that the translation in the input leads to closed form Fourier transforms: for $g(\mathbf{x}) = f(\mathbf{x} \circ \boldsymbol{\gamma})$, $\hat{g}(\boldsymbol{\xi}) = \frac{1}{\prod \gamma_d} \hat{f}(\boldsymbol{\xi} \oslash \boldsymbol{\gamma})$, and for $h(\mathbf{x}) = f(\mathbf{x} - \mathbf{x}_0)$, $\hat{h}(\boldsymbol{\xi}) = \hat{f}(\boldsymbol{\xi}) e^{-2i\pi \boldsymbol{\xi}^\top \mathbf{x}_0}$. The generalized Fourier transform to obtain GSD is equivalent to a Fourier transform of the concatenated vector $\begin{pmatrix} \mathbf{x} \\ -\mathbf{x}' \end{pmatrix}$. Using the above observations, we can obtain the GSD of the harmonizable mixture kernel.

$$S_{k_{\text{HM}}}(\boldsymbol{\omega}, \boldsymbol{\xi}) = \sum_{p=1}^{P} S_{k_p}(\boldsymbol{\omega}, \boldsymbol{\xi}) e^{-2i\pi \mathbf{x}_p^\top (\boldsymbol{\omega} - \boldsymbol{\xi})}, \tag{13}$$

$$S_{k_p}(\boldsymbol{\omega}, \boldsymbol{\xi}) = \frac{1}{\prod_{d=1}^{D} \gamma_{pd}^2} \sum_{1 \le i,j \le Q_p} b_{pij} S_{pij}(\boldsymbol{\omega}, \boldsymbol{\xi}), \tag{14}$$

$$S_{pij}(\boldsymbol{\omega}, \boldsymbol{\xi}) = S_{k_{\text{LSG}}}((\boldsymbol{\omega} - \boldsymbol{\mu}_{pi}) \oslash \boldsymbol{\gamma}_p, (\boldsymbol{\xi} - \boldsymbol{\mu}_{pj}) \oslash \boldsymbol{\gamma}_p). \tag{15}$$

The Wigner transform of a k_{HM} requires an additional step of reverting the subscript.

$$k_p(\mathbf{x}, \mathbf{x}') = k_{\mathrm{LSG}}(\mathbf{x} \circ \boldsymbol{\gamma}_p, \mathbf{x}' \circ \boldsymbol{\gamma}_p) \sum_{1 \le i,j \le Q_p} \beta_{pij} e^{2i\pi(\boldsymbol{\mu}_{pi}^\top \mathbf{x} - \boldsymbol{\mu}_{pj}^\top \mathbf{x}')} \tag{16}$$

$$= \frac{1}{2} k_{\mathrm{LSG}}(\mathbf{x} \circ \boldsymbol{\gamma}_p, \mathbf{x}' \circ \boldsymbol{\gamma}_p) \sum_{1 \le i,j \le Q_p} \beta_{pij} \left(e^{2i\pi(\boldsymbol{\mu}_{pi}^\top \mathbf{x} - \boldsymbol{\mu}_{pj}^\top \mathbf{x}')} + e^{2i\pi(\boldsymbol{\mu}_{pj}^\top \mathbf{x} - \boldsymbol{\mu}_{pi}^\top \mathbf{x}')} \right) \tag{17}$$

$$= k_{\mathrm{LSG}}(\mathbf{x} \circ \boldsymbol{\gamma}_p, \mathbf{x}' \circ \boldsymbol{\gamma}_p) \sum_{1 \le i,j \le Q_p} \beta_{pij} \left(\cos\left(2\pi \left(\frac{\boldsymbol{\mu}_{pi} + \boldsymbol{\mu}_{pj}}{2} \right)^\top \boldsymbol{\tau} \right) \cos(2\pi(\boldsymbol{\mu}_{pi} - \boldsymbol{\mu}_{pj})^\top \widetilde{\mathbf{x}}) + i g(\widetilde{\mathbf{x}}, \boldsymbol{\tau}) \right). \tag{18}$$

The imaginary part $g(\widetilde{\mathbf{x}}, \boldsymbol{\tau})$ is an odd function with respect to $\boldsymbol{\tau}$: $g(\widetilde{\mathbf{x}}, \boldsymbol{\tau}) = -g(\widetilde{\mathbf{x}}, -\boldsymbol{\tau})$, and thus has an integral of 0 with Wigner transform. The above derivation gives a separable kernel formulation with respect to $\widetilde{\mathbf{x}}$ and $\boldsymbol{\tau}$

$$W_{k_{\mathrm{HM}}}(\mathbf{x}, \boldsymbol{\omega}) = \sum_{p=1}^{P} W_{k_p}(\mathbf{x} - \mathbf{x}_p, \boldsymbol{\omega}), \tag{19}$$

$$W_{k_p}(\mathbf{x}, \boldsymbol{\omega}) = \frac{1}{\prod_{d=1}^{D} \gamma_{pd}} \sum_{1 \le i,j \le Q_p} W_{pij}(\mathbf{x}, \boldsymbol{\omega}), \tag{20}$$

$$W_{pij}(\mathbf{x}, \boldsymbol{\omega}) = W_{k_{\mathrm{LSG}}}\left(\mathbf{x} \circ \boldsymbol{\gamma}_p, (\boldsymbol{\omega} - (\boldsymbol{\mu}_{pi} + \boldsymbol{\mu}_{pj})/2) \oslash \boldsymbol{\gamma}_p\right) \cos(2\pi(\boldsymbol{\mu}_{pi} - \boldsymbol{\mu}_{pj})^\top \mathbf{x}). \tag{21}$$

2.1 Derivation of variational Fourier features

For a GP with an integrable harmonizable kernel k, we can derive the cross-covariances between the primary GP f and its Fourier transform $\hat{f}$:

$$\begin{aligned}
\mathbf{cov}(\hat{f}(\boldsymbol{\omega}), f(\mathbf{x})) &= \mathbb{E}\left\{ \int f(\mathbf{t}) f(\mathbf{x}) e^{-2i\pi\boldsymbol{\omega}^\top \mathbf{t}} \, \mathrm{d}\mathbf{t} \right\} \\
&= \int_{\mathbb{R}^D} k(\mathbf{t}, \mathbf{x}) e^{-2i\pi\boldsymbol{\omega}^\top \mathbf{t}} \, \mathrm{d}\mathbf{t} \qquad (22) \\
\mathbf{cov}(f(\mathbf{x}), \hat{f}(\boldsymbol{\omega})) &= \mathrm{cov}(\hat{f}(\boldsymbol{\omega}), f(\mathbf{x}))^* \\
\mathbf{cov}(\hat{f}(\boldsymbol{\omega}), \hat{f}(\boldsymbol{\xi})) &= \mathbb{E}\left\{ \iint f(\mathbf{x}) f(\mathbf{x}') e^{-2i\pi(\boldsymbol{\omega}^\top \mathbf{x} - \boldsymbol{\xi}^\top \mathbf{x}')} \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{x}' \right\} \\
&= \iint k(\mathbf{x}, \mathbf{x}') e^{-2i\pi(\boldsymbol{\omega}^\top \mathbf{x} - \boldsymbol{\xi}^\top \mathbf{x}')} \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{x}' \\
&= S_k(\boldsymbol{\omega}, \boldsymbol{\xi}). \qquad (23)
\end{aligned}$$

In the case of harmonizable mixture kernels, we need to compute closed form $\int k_p(\mathbf{t}, \mathbf{x}) e^{-2i\pi\boldsymbol{\xi}^\top \mathbf{t}} \, \mathrm{d}\mathbf{t}$ for the cross-covariances in variational Fourier features which is derived below:

$$\begin{aligned}
\int k_p(\mathbf{t}, \mathbf{x}) e^{-2i\pi\boldsymbol{\xi}^\top \mathbf{t}} \, \mathrm{d}\mathbf{t} &= \sum_{1 \le i,j \le Q_p} \beta_{pij} \exp\left(-2\pi^2 \mathbf{x}^\top \left(\frac{\boldsymbol{\Sigma}_1}{4} + \boldsymbol{\Sigma}_2 \right) - 2i\pi \boldsymbol{\mu}_{pj}^\top \mathbf{x} \right) \\
&\quad \times \int \exp\left(-2\pi^2 (\mathbf{t} - \mathbf{x}_0)^\top \left(\frac{\boldsymbol{\Sigma}_1}{4} + \boldsymbol{\Sigma}_2 \right) (\mathbf{t} - \mathbf{x}_0) + 2i\pi \mu_{pi}^\top \mathbf{x} - 2i\pi \boldsymbol{\xi}^\top \mathbf{x} \right) \, \mathrm{d}\mathbf{x} \qquad (24) \\
&= \sum_{1 \le i,j \le Q_p} \beta_{pij} \exp\left(-2\pi^2 \mathbf{x}^\top \left(\frac{\boldsymbol{\Sigma}_1}{4} + \boldsymbol{\Sigma}_2 \right) - 2i\pi \boldsymbol{\mu}_{pj}^\top \mathbf{x} - 2i\pi \mathbf{x}_0^\top \boldsymbol{\xi} \right) \\
&\quad \times \mathcal{N}\left((\boldsymbol{\xi} - \boldsymbol{\mu}_{pi}) \oslash \gamma_p \,\middle|\, 0, \frac{\boldsymbol{\Sigma}_1}{4} + \boldsymbol{\Sigma}_2 \right), \qquad (25) \\
\mathbf{x}_0 &= (\boldsymbol{\Sigma}_1 + 4\boldsymbol{\Sigma}_2)^{-1} (4\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1) \mathbf{x}. \qquad (26)
\end{aligned}$$

3 Proof of theorem 3

Theorem 2. *Given a continuous, integrable kernel k_{LS} with a valid* generalized spectral density*, the harmonizable mixture kernel*

$$k_h(\mathbf{x}, \mathbf{x}') = \sum_{p=1}^{P} k_p(\mathbf{x} - \mathbf{x}_p, \mathbf{x}' - \mathbf{x}_p), \tag{27}$$

$$k_p(\mathbf{x}, \mathbf{x}') = k_{LS}(\mathbf{x} \circ \boldsymbol{\gamma}_p, \mathbf{x}' \circ \boldsymbol{\gamma}_p)\phi_p(\mathbf{x})^\dagger \mathbf{B}_p \phi_p(\mathbf{x}'), \tag{28}$$

where $P \in \mathbb{N}_+$, $(\phi_p(\mathbf{x}))_q = e^{2i\pi \boldsymbol{\mu}_{pq}^\top \mathbf{x}}$, $q = 1, \ldots, Q_p$, $\boldsymbol{\gamma}_p \in \mathbb{R}_+^D$, $\mathbf{x}_p \in \mathbb{R}^D$, $\boldsymbol{\mu}_{pq} \in \mathbb{R}^D$, B_p as positive definite Hermitian matrices, is dense in the family of harmonizable covariances with respect to pointwise convergence of functions.

Proof. Discrete measures are dense in the Banach space of complex-valued measures on $\mathbb{R}^D \times \mathbb{R}^D$. And the same can be extended to the denseness of discrete positive definite bimeasures (a subset of measures on $\mathbb{R}^D \times \mathbb{R}^D$) in positive definite bimeasures. Intuitively, a harmonizable kernel $k_0 : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{C}$ with a generalized spectral density $S(\boldsymbol{\omega}, \boldsymbol{\xi}) = \dfrac{\partial^2 \Psi(\boldsymbol{\omega}, \boldsymbol{\xi})}{\partial \boldsymbol{\omega} \partial \boldsymbol{\xi}}$ can be expressed in the following form:

$$k_0(\mathbf{x}, \mathbf{x}') = \iint S(\boldsymbol{\omega}, \boldsymbol{\xi}) e^{2i\pi(\boldsymbol{\omega}^\top \mathbf{x} - \boldsymbol{\xi}^\top \mathbf{x}')} \, \mathrm{d}\boldsymbol{\omega} \, \mathrm{d}\boldsymbol{\xi}. \tag{29}$$

Consider the Darboux sum with respect to a grid of frequencies $\boldsymbol{\omega}_0 < \boldsymbol{\omega}_2 < \ldots < \boldsymbol{\omega}_Q$

$$\sum_{1 \le u,v \le Q} e^{2i\pi(\boldsymbol{\omega}_v^\top \mathbf{x} - \boldsymbol{\omega}_u^\top \mathbf{x}')} \Psi([\boldsymbol{\omega}_{u-1}, \boldsymbol{\omega}_u], [\boldsymbol{\omega}_{v-1}, \boldsymbol{\omega}_v]) = \sum_{1 \le u,v \le Q} \alpha_{uv} e^{2i\pi(\boldsymbol{\omega}_u^\top \mathbf{x} - \boldsymbol{\omega}_v^\top \mathbf{x})}. \tag{30}$$

Given the positive definiteness of $\Psi(\cdot, \cdot)$, the matrix $(\alpha_{uv})_{u,v=1}^Q$ is positive semidefinite. the Darboux sum takes a "generalized sparse spectrum" form: $k_{\mathrm{GSS}}(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\dagger \mathbf{B} \phi(\mathbf{y})$. It is an uniform approximator of the double integral on a compact set $[\boldsymbol{\omega}_0, \boldsymbol{\omega}_Q] \times [\boldsymbol{\omega}_0, \boldsymbol{\omega}_Q]$, which converges to k_0 as $[\boldsymbol{\omega}_0, \boldsymbol{\omega}_Q] \times [\boldsymbol{\omega}_0, \boldsymbol{\omega}_Q]$ covers the entire frequency domain.

Given the expressiveness of the generalized sparse spectrum kernel, we can similarly smooth the spectral representation by multiplying with $k_{\mathrm{LS}}(\mathbf{x} \circ \gamma, \mathbf{x}' \circ \gamma)$, and add more flexibility by translating the input, which gives the final harmonizable mixture kernel form. □

It is worth noting that the theorem can be strengthened from positive semidefinite Hermitian matrices $\mathbf{B}_p$, to non-negative valued positive semidefinite matrices. This is an immediate result from the "phase shift" of the Fourier transform.

4 Expressiveness of product spectral kernels

The spectral mixture product (SMP) kernel (Wilson et al., 2014) is a variant of the spectral mixture kernel, where the inner product inside the cosine function is decomposed into a product of cosines, which makes each spectral component a product kernel.

$$k_{\mathrm{SMP}}(\boldsymbol{\tau}) = \sum_{q=1^Q} w_q^2 \prod_{d=1}^{D} e^{-2\pi^2 \sigma_d^2 \tau_d^2} \cos(2\pi \mu_{qd} \tau_d). \tag{31}$$

Spectral mixture product kernel is used in multidimensional pattern discovery for its added scalability (Wilson et al., 2014). However, it is not as expressive as the original spectral mixture kernel. We see the product of cosines can be decomposed as follows

$$\prod_{d=1}^{D} \cos(2\pi \mu_{qd} \tau_d) = \frac{1}{2^D} \sum_{\mathbf{b} \in \{-1,1\}^D} e^{2i\pi(\mathbf{b} \circ \boldsymbol{\mu})^\top \boldsymbol{\tau}}. \tag{32}$$

Therefore, product spectral kernels are spectral mixture kernel with additional symmetry constraint: $\psi_k(\omega) = \psi_k(b \circ \omega), \forall b \in \{-1, 1\}^D$. Note that this constraint is stricter than the constraint for an arbitrary stationary kernel $\psi_k(\omega) = \psi_k(-\omega)$. We conclude that spectral mixture product kernel shall behave as well as spectral mixture kernel when we underlying covariance has a spectral distribution that is symmetrical with respect to every "axis".

For multidimensional harmonizable spectral kernel, we can utilize enhanced scalability when we similarly replace the cosine term with a product of cosines with respect to every dimension, which leads to similar stronger symmetry of the generalized spectral distribution $\Psi(\omega, \xi) = \Psi(b_1 \circ \omega, b_2 \circ \xi), \forall b_1, b_2 \in \{-1, 1\}^D$.

When we use product spectral kernel in replacement of original spectral kernels, there is a tradeoff between scalability and expressiveness: product spectral kernels offer additional scalability for the cost of reduced expressiveness based on symmetry of the (generalized) spectral distribution.

5 Interpreting generalized spectral mixture kernel

The *generalized spectral mixture kernel* (GSM) (Remes et al., 2017) is a nonstationary generalization of the stationary spectral mixture kernel. The functional formulation makes the kernel able to handle complex structure in the input. It is formulated as

$$k_{\text{GSM}}(x, x') = \sum_{q=1}^{Q} w_q(x) w_q(x') k_{\text{Gibbs, q}}(x, x') \cos(2\pi(\mu_q(x)x - \mu_q(x')x')), \tag{33}$$

$$k_{\text{Gibbs, q}}(x, x') = \sqrt{\frac{2 l_q(x) l_q(x')}{l_q(x)^2 + l_q(x')^2}} \exp(-\frac{(x - x')^2}{l_q(x)^2 + l_q(x')^2}), \tag{34}$$

where functions $w_q(x)$, $\mu_q(x)$, $l_q(x)$ have GP priors, encoding a spectrogram with $w_q(x)$ denoting the magnitude of the frequency, $\mu_q(x)$, and $l_q(x)$ denoting the mean and variance of the frequency components. We propose that this kernel first projects input using some unknown feature map, and then assume stationary in the projected space and fit a stationary spectral mixture kernel. Consider the kernel $k_{FSS}(\mathbf{x}, \mathbf{x}') = \cos(g(\mathbf{x}) - g(\mathbf{x}'))$ with an arbitrary function $g : \mathbb{R}^D \mapsto \mathbb{R}$. Assuming $g(\cdot)$ lies within some RKHS $\mathcal{H}$, then $g(x) = \langle g, K(x, \cdot) \rangle_{\mathcal{H}}$ is an inner product between a "constant vector" g and the projected input $K(x, \cdot)$, therefore the kernel k_{FSS} generalizes sparse spectrum kernel by projecting the data with a feature map first. The GSM kernel then multiplies k_{FSS} with a Gibbs kernel, implying an unknown mixture model on the spectrum induced by the projected space.

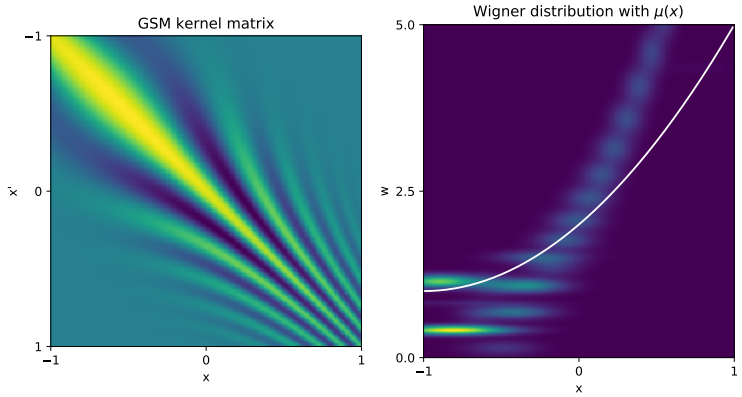

Figure 1: Wigner distribution of the approximation of a GSM kernel

The white line denotes the $\mu(x)$ corresponding to frequency of the spectrogram.

However, the intuitive interpretation of the underlying spectrogram might be an inaccurate way to interpret GSM kernel. When we approximate a GSM kernel with HM kernel, the Wigner distribution of the HM kernel does not quite correspond to the spectrogram interpretation: the mean of the frequency components are "stretched", when x approaches 1, the actual local frequency is higher than what the function $\mu(x)$ suggests. GSM kernel seems to keep a biased account of the frequency information.

While the harmonizable mixture kernel handles nonstationarity in the input directly, the GSM kernel is equally valid – it projects the input space to a feature space, and then assumes stationarity on the feature space.

6 Experiment details

The models are implemented in Python using the GPFlow framework (Matthews et al., 2017). We implemented the *harmonizable mixture kernel*, two sparse GP models with variational Fourier features (namely the variational lower bound for sparse GP regression (Titsias, 2009) and the stochastic variational Gaussian process (Hensman et al., 2017)), and a natural gradient optimizer accepting complex-valued variational parameters.

6.1 Kernel recovery

For kernel recovery, we perform stochastic gradient descent using Adam (Kingma and Ba, 2014), using mean square error of random batches of data as objective function.

6.2 GP classification

For GP classification using banana dataset, we selected a subset of data containing 500 data points, and trained a variational GP model. The full variational model is then approximated using sparse GP with inducing points and inducing frequencies.

The inducing points are initialized using K-means clustering, and the inducing frequencies are initialized using the frequency means suggested in the trained HMK, with an added Gaussian noise. We ran each model with 5 random initializations and pick the model with highest classification accuracy on the training set.

For the training of sparse GP model, we first trained the variational parameters with natural gradients for 200 iterations. We then jointly train the inducing variables and variational parameters with 700 alternating rounds of optimization using respective natural gradient optimizers and Adam (such approach is suggested in (Salimbeni et al., 2018)).

6.3 GP regression

For GP regression with solar irrandiance, we used the same partition of training and test set in experiments in (Gal and Turner, 2015) and (Hensman et al., 2017). We further standardize the X-axis for numerical stability of the variational Fourier features. We used sparse GP regression (Titsias, 2009), where the model is modified to allow for VFF with the harmonizable mixture kernel.

For GP regression with Gaussian kernel, we used 50 inducing points initialized with K-Means, and initialized the kernel hyperparameters using 5 increasing lengthscales. The model is chosen using log-likelihoods on the training set.

With an assumption of smoothness of the underlying data, we used the residual value of the training data minus the predicted value of the previous model, and used a discrete Fourier transform on 6 subdivisions of data. The SM kernel has 3 frequency components initialized with respectively the highest two frequency in the discrete Fourier transform and the 0 frequency. This is initialization is then added with Gaussian noise and optimized.

The HMK for GP regression has a total of $P = 6$ components, with $Q = 3$ frequency values for each components. The input shifts $\mathbf{x}_p$ are initialized using K-means clustering, and the frequency values are 0, the

highest density frequency obtained in discrete Fourier transform, and random values. We ran the sparse GP model with inducing points for some iterations and then ran variational Fourier features centered around the frequency values.

References

Yarin Gal and Richard Turner. Improving the Gaussian process sparse spectrum approximation by representing uncertainty in frequency inputs. In *International Conference on Machine Learning*, pages 655–664, 2015.

James Hensman, Nicolas Durrande, and Arno Solin. Variational Fourier features for Gaussian processes. *The Journal of Machine Learning Research*, 18(1):5537–5588, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alexander G. de G. Matthews, Mark van der Wilk, Tom Nickson, Keisuke. Fujii, Alexis Boukouvalas, Pablo León-Villagrá, Zoubin Ghahramani, and James Hensman. GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research*, 18(40):1–6, Apr 2017. URL `http://jmlr.org/papers/v18/16-537.html`.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2008.

Sami Remes, Markus Heinonen, and Samuel Kaski. Non-stationary spectral kernels. In *Advances in Neural Information Processing Systems*, pages 4642–4651, 2017.

Hugh Salimbeni, Stefanos Eleftheriadis, and James Hensman. Natural gradients in practice: Non-conjugate variational inference in Gaussian process models. *arXiv preprint arXiv:1803.09151*, 2018.

R. Silverman. Locally stationary random processes. *IRE Transactions on Information Theory*, 3(3):182–187, 1957.

Michalis Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009.

Andrew G. Wilson, Elad Gilboa, Arye Nehorai, and John P. Cunningham. Fast kernel learning for multidimensional pattern extrapolation. In *Advances in Neural Information Processing Systems*, pages 3626–3634, 2014.