

Efficient Approximate Inference for Stationary Kernel on Frequency Domain

Yohan Jung¹, Kyungwoo Song², and Jinkyoo Park¹

¹ SI Lab, Korea Advanced Institute of Science & Technology (KAIST)

² MLAI Lab, University of Seoul



Motivation

- Stationary kernel $k(\tau)$ is a class of kernel defined as the function of the difference $\tau = x_1 - x_2$ between inputs x_1 and x_2 .
- The stationary kernel has a theoretical background called **Bochner's theorem** that explains the construction of stationary kernel.

Bochner's theorem

Let $k(\tau)$ be a complex-valued function defined on $\tau \in R^d$.

If $k(\tau)$ is the covariance function of weakly stationary process, $k(\tau)$ can be represented as

$$\underbrace{k(\tau)}_{\substack{\text{stationary} \\ \text{kernel}}} = \int e^{i2\pi s^T \tau} \underbrace{p(s)}_{\substack{\text{spectral} \\ \text{density}}} ds$$

where $p(s)$ denotes the spectral density defined on frequency domain $s \in R^d$.

- This theorem implies that any stationary kernel $k(\tau)$ can be quantified by specifying the spectral density $p(s)$.

For example, the RBF kernel $k_{\text{RBF}}(\tau) = \exp\left(-\frac{1}{2} \left(\frac{2\pi\tau}{l}\right)^2\right)$ is specified by setting $p(s) = N(s; 0, l^2 I)$.

- Wilson et al. devised a spectral mixture (SM) kernel $k_{\text{SM}}(\tau)$ which can approximate any stationary kernel. They quantified the spectral density $p(s)$ as a weighted Gaussian mixture density as follows:

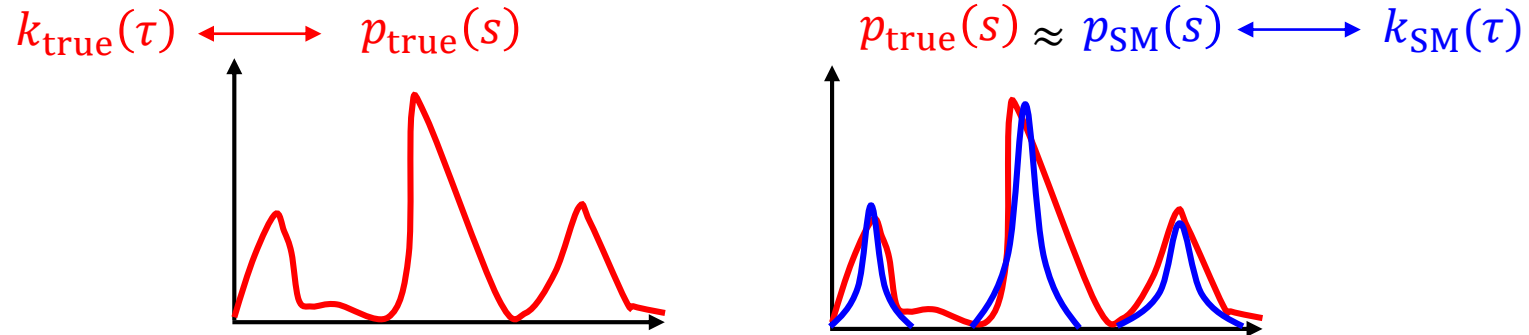
spectral density : $p(s) = \sum_{q=1}^Q w_q p_q(s; \mu_q, \sigma_q^2)$ where $p_q(s; \mu_q, \sigma_q^2) = \frac{1}{2} \left(N(s; \mu_q, \text{Diag}(\sigma_q^2)) + N(-s; \mu_q, \text{Diag}(\sigma_q^2)) \right)$



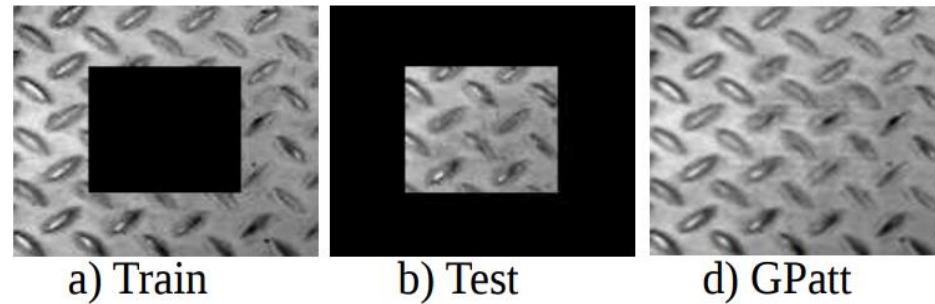
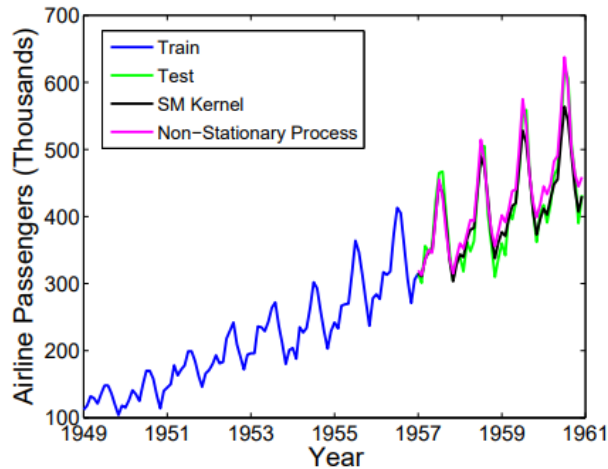
stationary kernel : $k_{\text{SM}}(\tau) = \sum_{q=1}^Q w_q \exp\left(-\frac{1}{2} (2\pi)^2 \tau^T \text{Diag}(\sigma_q^2) \tau\right) \cos 2\pi \mu_q^T \tau$.

Motivation

- For the given stationary process with its kernel $k_{\text{true}}(\tau)$ and spectral density $p_{\text{true}}(s)$, the weighted Gaussian mixture density $p(s)$ can model the true spectral density $p_{\text{true}}(s)$ as shown below figure.



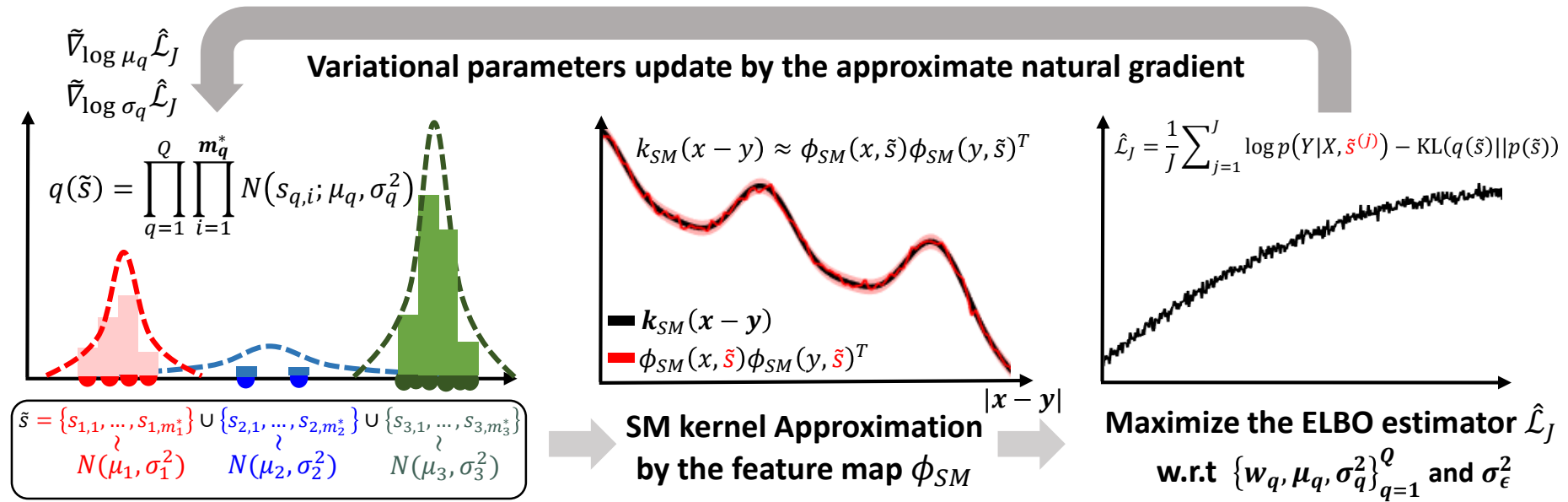
- For modeling the real dataset, SM kernel has shown great prediction performance on extrapolation task as well.



Reference : Covariance Kernels for Fast Automatic Pattern Discovery and Extrapolation with Gaussian Processes - AG Wilson

Methodology

- However, employing the SM kernel has some limitations due to training:
 - ✓ Training lots of hyperparameters $\{w_q, \mu_q, \sigma_q^2\}_{q=1}^Q$ could induce a **over-fitting**.
 - ✓ Training lots of hyperparameters $\{w_q, \mu_q, \sigma_q^2\}_{q=1}^Q$ requires a **longer training time and large memory (automatic differentiation)**, especially for large-scale dataset.
- In this work, we **propose the approximate inference for SM kernel (RFF + Sampling-based VI)** to alleviate the mentioned issue.



- Step1.** Define the distribution of M spectral points, with $M = \sum_{q=1}^Q m_q$, which is used as variational distribution.
- Step2.** Construct the approximate SM kernel by using the random spectral points sampled from its distribution.
- Step3.** Optimize the parameters of M spectral points distribution by using the ELBO estimator $\hat{\mathcal{L}}_J$ as training loss.

- The proposed inference is **scalable** because it uses $\hat{K}_{SM}(X, X)$ instead of $K_{SM}(X, X)$,
- The proposed inference **could relaxes overfitting** because it uses $\hat{K}_{SM}(X, X)$ having randomness and KL term as regularizer.

Special method 1. weighted sampling

Special method 2. approximate natural gradient

Special method 1: weighted sampling

Questions. Can we make the ELBO estimator $\hat{\mathcal{L}}_J(\Theta)$ stable during training ?

Suggestions. We analyze the ELBO estimator error $\hat{\mathcal{L}}_J(\Theta)$ (Proposition 2), and propose the weighted sampling (Proposition 3).

- We first bound the ELBO estimator error, i.e., $\log p(Y|X) - \hat{\mathcal{L}}_J$ via terms of spectral norm $E[\|K_{SM}(X, X) - \hat{K}_{SM}(X, X)\|_2]$.
- To reduce the ELBO estimator error, we should reduce $E[\|K_{SM}(X, X) - \hat{K}_{SM}(X, X)\|_2]$, which requires large $M = \sum_{q=1}^Q m_q$ spectral points.

Proposition 2. The error of ELBO estimator $\hat{\mathcal{L}}_J$ in Eq. (3.8) is bounded as

$$0 \leq \log p(Y|X) - \hat{\mathcal{L}}_J \leq \left(\frac{\|Y\|_2^2 + N\sigma_\epsilon^2}{2\sigma_\epsilon^4} \right) E[\|K_{SM} - \hat{K}_{SM}\|_2] + \text{KL}(q(\tilde{s})||p(\tilde{s}))$$

where $E[\|K_{SM} - \hat{K}_{SM}\|_2]$ is integrated over the spectral point \tilde{s} with $q(\tilde{s})$.

- For the fixed $M = \sum_{q=1}^Q m_q$ spectral points, we focus on reducing $E[\|K_{SM}(X, X) - \hat{K}_{SM}(X, X)\|_2]$ by allocating $\{m_q\}_{q=1}^Q$ spectral points.

Proposition 3. Given inputs $X = \{x_n\}_{n=1}^N$, let m_q be the number of spectral points sampled from $N(\mu_q, \sigma_q^2)$, and $M = \sum_{q=1}^Q m_q$ be the total number of spectral points. Let $p_q = \frac{m_q}{M}$ be the ratio of spectral points. Then, the optimal $p_1^*, \dots, p_Q^* = \arg \min_{p_1, \dots, p_Q} E[\|\hat{K}_{SM} - K_{SM}\|_F]$ is given

$$p_q^* = \frac{w_q \left[\sum_{i=1}^N \sum_{i < j} g_q(x_i - x_j) \right]^{1/2}}{\sum_{q=1}^Q w_q \left[\sum_{i=1}^N \sum_{i < j} g_q(x_i - x_j) \right]^{1/2}}$$

where $g_q(\tau) = 1 + k_q(2\tau) - 2k_q^2(\tau)$ and $k_q(\tau)$ denotes the q -th component term in SM kernel related to $\{\mu_q, \sigma_q^2\}$. The optimal spectral point m_q^* is obtained as the integer closest to $\max\{1, Mp_q^*\}$.

Special method 2: approximate natural gradient update

Questions. Can we reduce the training time taken for obtaining the hyperparameters?

Suggestions. We propose the [approximate natural gradient \(Proposition 4\)](#), to expedite the convergence of parameter inference.

- For the probability density $p_\theta(z)$ parameterized by θ , the natural gradient is defined via KL divergence.

$$\tilde{\nabla}_\theta \mathcal{L}(\theta) = \arg \min_{\{\Delta\theta; \text{KL}(p_\theta \| p_{\theta+\Delta\theta})=\epsilon\}} \mathcal{L}(\theta + \Delta\theta).$$

- The natural gradient can be understood as the second order optimization using Fisher Information matrix $F_\theta = \mathbb{E}_{p_\theta(z)}[\nabla_\theta \log p_\theta(z) \nabla_\theta \log p_\theta(z)^T]$

$$\tilde{\nabla}_\theta \mathcal{L}(\theta) = \arg \min_{\Delta\theta} \mathcal{L}(\theta) + \nabla_\theta \mathcal{L}(\theta)^T \Delta\theta + \frac{\lambda}{2} \Delta\theta^T F_\theta \Delta\theta,$$

- We propose the approximate natural gradient to make the Kernel Gram matrix remain P.S.D and numerically stable during training.

Proposition 4. Let $\mu_q^{(t)}$ and $\sigma_q^{(t)}$ be the t -th iterated parameters of $N(\mu_q, \sigma_q^2)$ which is q -th component distribution for $q(\tilde{s})$. The natural gradient of $\hat{\mathcal{L}}_J$ w.r.t μ_q and σ_q in log domain, i.e. $\tilde{\nabla}_{\log \mu_q} \hat{\mathcal{L}}_J$ and $\tilde{\nabla}_{\log \sigma_q} \hat{\mathcal{L}}_J$, can be approximated as

$$\begin{aligned} \tilde{\nabla}_{\log \mu_q} \hat{\mathcal{L}}_J &\approx \left(\frac{\sigma_q^{(t+1)}}{\mu_q^{(t)}} \right)^2 \circ \nabla_{\log \mu_q} \hat{\mathcal{L}}_J \\ \tilde{\nabla}_{\log \sigma_q} \hat{\mathcal{L}}_J &\approx \frac{1}{2} \nabla_{\log \sigma_q} \hat{\mathcal{L}}_J, \end{aligned}$$

for $\left| \left(\frac{\sigma_q^{(t+1)}}{\mu_q^{(t)}} \right)^2 \circ \nabla_{\log \mu_q} \hat{\mathcal{L}}_N \right| < 1$ and $\left| \nabla_{\log \sigma_q} \hat{\mathcal{L}}_N \right| < 1$ in element-wise sense.

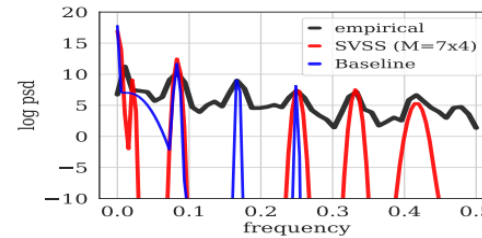
Experiments

- We conduct a regression task on airline dataset to validate that weighted sampling improves the approximate inference.

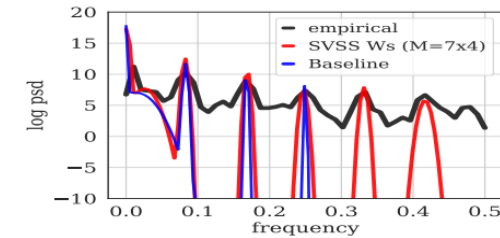
Quantitative result

	RMSE	MNLL	RMSE (Top 1-5)	MNLL (Top 1-5)
$M = 28 (Q = 7, m = 4)$				
Baseline (MLE-2)	61.85 ± 22.54	5.66 ± 0.49	26.88 ± 1.29	4.63 ± 0.02
SVSS	94.38 ± 22.30	5.53 ± 0.20	34.11 ± 1.68	4.92 ± 0.04
SVSS-Ws	54.00 ± 11.55	5.62 ± 0.34	28.62 ± 1.12	4.92 ± 0.03
$M = 28 (Q = 28, m = 1)$				
SVSS	101.04 ± 13.33	5.85 ± 0.15	78.72 ± 6.92	5.72 ± 0.21

Qualitative results



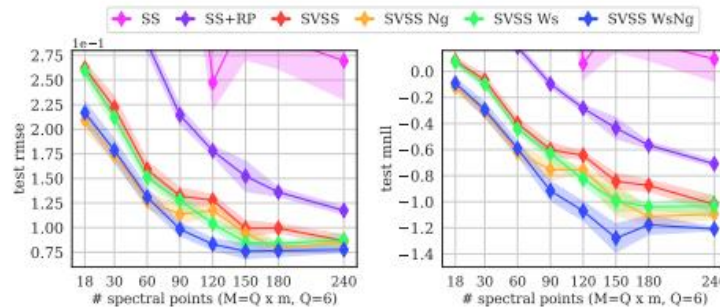
(c) $p(s)$ via SVSS



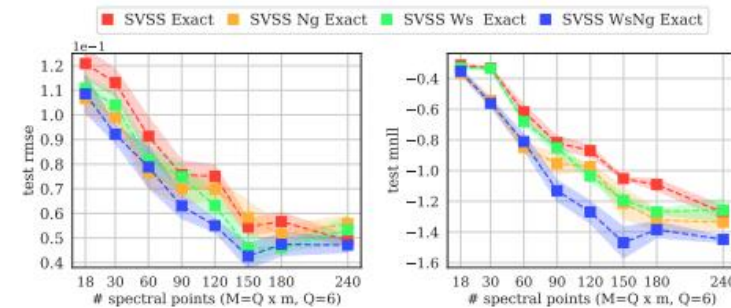
(d) $p(s)$ via SVSS-Ws

- We conduct an ablation study via regression task on UCI set to validate the effectiveness of the tricks used in the inference.

Prediction on test set



(a) prediction by the approximate kernel \hat{K}_{SM}



(b) prediction by the exact kernel K_{SM}

- Further details are explained our paper.
- If you are interested in our work, please check our paper!

Efficient Approximate Inference for Stationary Kernel on Frequency Domain

Yohan Jung¹ Kyungwoo Song² Jinkyoo Park¹

Abstract

Based on the Fourier duality between a stationary kernel and its spectral density, modeling the spectral density using a Gaussian mixture density enables one to construct a flexible kernel, known as a Spectral Mixture kernel, that can model any stationary kernel. However, despite its expressive power, training this kernel is typically difficult because scalability and overfitting issues often arise due to a large number of training parameters. To resolve these issues, we propose an approximate inference method for estimating the Spectral mixture kernel hyperparameters. Specifically, we approximate this kernel by using the finite random spectral points based on Random Fourier Feature and optimize the parameters for the distribution of spectral points by sampling-based variational inference. To improve this inference procedure, we analyze the training loss and propose two special methods: a sampling method of spectral points to reduce the error of the approximate kernel in training, and an approximate natural gradient to accel-

stationary kernel $k(x_1 - x_2)$ for inputs x_1 and x_2 can be expressed by an inverse Fourier transform of its spectral density $p(s)$ as

$$k(x_1 - x_2) = \int e^{i2\pi s^T(x_1 - x_2)} p(s) ds. \quad (1)$$

For example, the RBF kernel, which is defined as $k(x_1 - x_2) = \exp(-\frac{1}{2} \|\frac{2\pi(x_1 - x_2)}{l}\|^2)$, is obtained when specifying $p(s) = N(s; 0, l^{-2}I)$ with RBF hyperparameter l in Eq. (1). Based on the Fourier duality, selecting a specific kernel is equivalent to modeling the form of $p(s)$. If the model assumption for the spectral density is not flexible enough to express the spectral density of the true kernel, the \mathcal{GP} model with the induced kernel may not explain the dataset well. Thus, a flexible model for $p(s)$ is necessary in order to find a stationary kernel that can describe the dataset well.

As an attempt to construct a flexible kernel, [Wilson & Adams \(2013\)](#) represent $p(s)$ as a Gaussian mixture density $\sum_{q=1}^Q w_q N(s; \mu_q, \text{diag}(\sigma_q^2))$, where $\mu_q \in R^D$ and $\text{diag}(\sigma_q^2) \in R^{D \times D}$ are the mean and the diagonal covariance matrix, respectively. The Gaussian mixture density can represent any $p(s)$ flexibly based on the universal ap-

Appendix: Intuition of Kernel Approximation

- We reconsider that the SM kernel can be represented as

$$k_{\text{SM}}(\tau) = \sum_{q=1}^Q w_q \underbrace{\int e^{i2\pi s^T \tau} p_q(s; \mu_q, \text{Diag}(\sigma_q^2)) ds}_{q\text{-th integral}}$$

where q -th spectral density is defined as $p_q(s; \mu_q, \text{Diag}(\sigma_q^2)) = \frac{1}{2} (N(s; \mu_q, \text{Diag}(\sigma_q^2)) + N(-s; \mu_q, \text{Diag}(\sigma_q^2)))$.

- This implies that if we approximate q th integral term, then we can approximate SM kernel.
- For approximating each q -th integral, we can use monte-carlo integration that needs N_q spectral points $\{s_{q,i}\}_{i=1}^{N_q} \sim N(\mu_q, \text{Diag}(\sigma_q^2))$ as

$$\underbrace{\int e^{i2\pi s^T \tau} p_q(s; \mu_q, \text{Diag}(\sigma_q^2)) ds}_{q\text{-th integral}} \cong \frac{1}{N_q} \sum_{i=1}^{N_q} \cos 2\pi s_{q,1}^T \tau \triangleq \phi_q(x_1; \{s_{q,i}\}_{i=1}^{N_q})^T \phi_q(x_2; \{s_{q,i}\}_{i=1}^{N_q})$$

where $\phi_q(x; \{s_{q,i}\}_{i=1}^{N_q}) \triangleq \frac{1}{\sqrt{N_q}} [\cos 2\pi s_{q,1}^T x, \sin 2\pi s_{q,1}^T x, \dots, \cos 2\pi s_{q,N_q}^T x, \sin 2\pi s_{q,N_q}^T x] \in R^{N_q \times 1}$

- This implies that if we assign the randomness on finites spectral points $\{s_{q,i}\}_{i=1}^{N_q}$, we can generate the random estimator to approximate SM kernel.

Appendix: Step 1. Define the distribution of spectral points

- Based on this intuition, for q -th mixture component, we define the distribution of N_q spectral points $q(s_{q,1}, \dots, s_{q,N_q})$ having shared parameters $\{\mu_q, \text{Diag}(\sigma_q^2)\}$, as

$$q(s_{q,1}, \dots, s_{q,N_q}) = \prod_{i=1}^{N_q} N(s_{q,i}; \mu_q, \text{Diag}(\sigma_q^2))$$

- Then, we construct the random estimator ϕ_q such that

$$\phi_q(x; \{s_{q,i}\}_{i=1}^{N_q}) \triangleq \frac{1}{\sqrt{N_q}} \left[\cos 2\pi s_{q,1}^T, \sin 2\pi s_{q,1}^T, \dots, \cos 2\pi s_{q,N_q}^T, \sin 2\pi s_{q,N_q}^T \right] \text{ with } \{s_{q,i}\}_{i=1}^{N_q} \sim q(s_{q,1}, \dots, s_{q,N_q})$$

- For Q mixture components, we define the distribution of M spectral points as

$$q(\tilde{s}) = \prod_{q=1}^Q \underbrace{\prod_{i=1}^{N_q} N(s_{q,i}; \mu_q, \text{Diag}(\sigma_q^2))}_{q\text{-th component}}$$

where $\tilde{s} = (\underbrace{s_{1,1}, \dots, s_{1,N_1}}_{1\text{-th component}}, \dots, \underbrace{s_{Q,1}, \dots, s_{Q,N_Q}}_{Q\text{-th component}})$ and $M = \sum_{q=1}^Q m_q$.

- Then, we can define SM kernel random estimator $\phi_{SM}(x)$ using M spectral points as follows:

$$\phi_{SM}(x) = \left[\underbrace{\sqrt{w_1} \phi_1(x; \{s_{1,i}\}_{i=1}^{N_1})}_{1\text{-th component}}, \dots, \underbrace{\sqrt{w_Q} \phi_1(x; \{s_{Q,i}\}_{i=1}^{N_Q})}_{Q\text{-th component}} \right] \in R^{1 \times M}.$$

Appendix: Step 2. Approximate SM Kernel

- Using the defined random estimator $\phi_{SM}(x)$, we can approximate the SM kernel in unbiased manner:

$$k_{SM}(x_1 - x_2) = E_{q(\mathcal{S})}[\phi_{SM}(x_1)^T \phi_{SM}(x_2)] \cong \phi_{SM}(x_1)^T \phi_{SM}(x_2)$$

- For the dataset $X = \{x_n\}_{n=1}^N$, we can define random feature matrix $\Phi^{SM}(X)$ satisfying $E[\Phi^{SM}(X)^T \Phi^{SM}(X)] = K_{SM}(X, X) \in R^{N \times N}$.

$$\Phi^{SM}(X) = [\phi_{SM}(x_1); \dots, \phi_{SM}(x_N)] \in R^{N \times 2M}$$

- To answer the provable guarantee for the SM kernel approximation, we present the error bound of $\Phi^{SM}(X)^T \Phi^{SM}(X)$ as follows:

Proposition 1. Let $W_0 = \sqrt{\sum_{q=1}^Q w_q^2}$ and $M = Qm_1$ under the assumption $m_1 = \dots = m_Q$. Then, for a small $\epsilon > 0$, the error bound for the estimator $\hat{K}_{SM}(X, X) := \Phi^{SM}(X)\Phi^{SM}(X)^T$ is obtained as

$$\Pr\left(\|\hat{K}_{SM}(X, X) - K_{SM}(X, X)\|_2 \geq \epsilon\right) \leq N \exp\left(\frac{-3\epsilon^2 M}{NW_0 Q(6\|K_{SM}(X, X)\|_2 + 3NW_0\sqrt{Q} + 8\epsilon)}\right).$$

where $\|\cdot\|_2$ denotes the matrix spectral norm.

- Roughly speaking, as more spectral points (large $M = \sum_{q=1}^Q N_q$) are sampled, its kernel matrix approximation error is likely to be small.

Appendix: Step 3. Compute the ELBO estimator and update the parameters

- For training, we use the ELBO estimator $\hat{\mathcal{L}}(\Theta)$, with $\Theta = \{w_q, \mu_q, \sigma_q^2\}_{q=1}^Q$, which uses the distribution of spectral points $q(\tilde{s})$ as variational distribution

$$\begin{aligned} \log p(Y|X) &\geq \int \log \left(p(Y|X, \tilde{s}) q(\tilde{s}) \left(\frac{p(\tilde{s})}{q(\tilde{s})} \right) \right) d\tilde{s} \\ &\cong \underbrace{\frac{1}{J} \sum_{j=1}^J \log p(Y|X, \tilde{s}_{(j)})}_{\text{log marg. lik estimator}} - \underbrace{\text{KL}(q(\tilde{s})||p(\tilde{s}))}_{\text{regulaizer}} \triangleq \hat{\mathcal{L}}_J(\Theta) \quad \tilde{s}_{(j)} = (s_{1,1}^{(j)}, \dots, s_{Q,N_Q}^{(j)}) \sim q(\tilde{s}) \end{aligned}$$

where $\log p(Y|X, \tilde{s}_{(j)}) = \log N(Y; \hat{K}_{\text{SM}}^{(j)}(X, X) + \sigma_\epsilon^2 I)$ and $\hat{K}_{\text{SM}}^{(j)}(X, X) = \Phi_{\text{SM}}(X; \tilde{s}_{(j)})^T \Phi_{\text{SM}}(X; \tilde{s}_{(j)})$

$$\text{KL}(q(\tilde{s})||p(\tilde{s})) = \sum_{q=1}^Q \text{KL} \left(N(s_{q,1}; \mu_q, \sigma_q^2) || N(s_{q,1}; \mu_{q_0}, \sigma_{q_0}^2) \right) \text{ with } q(\tilde{s}): N(s_{q,1}; \mu_q, \sigma_q^2) \text{ and } p(\tilde{s}): N(s_{q,1}; \mu_{q_0}, \sigma_{q_0}^2)$$

- The $q(\tilde{s})$ denotes the variational distribution of spectral points, which is defined in Step 1.
- The $p(\tilde{s})$ denotes the prior distribution of $q(\tilde{s})$ having the same form with $q(\tilde{s})$.
- The parameters $\{\mu_{q_0}, \sigma_{q_0}^2\}_{q=1}^Q$ of $p(\tilde{s})$ are initialized by using the empirical spectral density of dataset or inductive bias of the task.