2nd Symposium on Advances in Approximate Bayesian Inference, 2019 1–12

Spectral Mixture Kernel Approximation Using Reparameterized Random Fourier Feature

Yohan Jung and **Jinkyoo Park** {becre1776, jinkyoo.park}@kaist.ac.kr
Industrial & Systems Engineering, KAIST, Daejeon, Republic of Korea

1. Introduction

Gaussian Process (GP) has been widely used to model the complex input and output relationship of the data due to its expressive power (Rasmussen, 2004). Selecting a kernel to determine the structure of the covariance is a crucial factor governing the performance of GP model. Spectral Mixture (SM) kernel devised by Wilson (Wilson and Adams, 2013) can be one candidate kernel because SM kernel can approximate any stationary kernel. In spite of its expressive power, training SM kernel takes more time due to many SM kernel hyperparameters. This drawback would prevent this SM kernel from being widely applied to practical problems for a large volume of data. Generally, two approaches have been mainly studied for the scalability of GP model: inducing inputs method (Titsias, 2009; Hensman et al., 2013; Salimbeni et al., 2018) and sparse spectrum based on Random Fourier Feature (Lázaro-Gredilla et al., 2010; Gal and Turner, 2015; Hoang et al., 2017).

In this paper, we propose SM kernel approximation method by Reparameterized Random Fourier Feature (R-RFF). Also, we develop the regularized sparse spectrum approximation for kernel learning. Specifically, we apply the reparameterization trick (Kingma and Welling, 2013) to Random Fourier Feature (Rahimi and Recht, 2008) in the sense of both general parameter and natural parameter. In the process, we develop a robust sampling algorithm to consider the number of spectral points dependent on the ratio of weight parameters of SM kernel. Based on developed SM kernel approximation, we propose a training method employing Stochastic Gradient Variational Bayes (SGVB) to regularized lower bound. This approach allows us to scalably train GP model using SM kernel based on a stochastic optimization framework.

2. Background

2.1. Spectral Mixture (SM) Kernel

Bochner's theorem states that stationary kernel $k(\tau)$ for target function f can be obtained as the Fourier transform of spectral density $p(S)$ (Bochner, 1959).

$$k(\tau) = \int e^{2\pi i S^{\mathrm{T}}\tau} p(S) dS \tag{1}$$

where $\tau = |x_1 - x_2|$ between two inputs $x_1, x_2 \in R^P$. Wilson (Wilson and Adams, 2013) devises spectral mixture (SM) kernel by considering $p(S)$ as the symmetric weighed sum of diagonal Gaussian distribution, i.e, $p(S) = \sum_{q=1}^{Q} w_q \left(\frac{N(s|\mu_q,\Sigma_q)+N(-s|\mu_q,\Sigma_q)}{2} \right)$ with $\mu_q = (\mu_q^{(1)}, .., \mu_q^{(P)})$ and $\Sigma_q = \text{diag}(\sigma_q^{2(1)}, .., \sigma_q^{2(P)})$. Then, the SM kernel is obtained as

$$k_{SM}(\tau) = \sum_{q=1}^{Q} w_q \cos\left(2\pi\tau^T\mu_q\right) \prod_{p=1}^{P} \exp\left(-2\pi^2\tau_p^2\sigma_q^{(p)}\right) \tag{2}$$

where τ_p is the p th components in the τ.

2.2. Random Fourier Feature (RFF)

Random Fourier Feature method (Rahimi and Recht, 2008) approximates the stationary covariance function $k(x-y)$ by applying Monte Carlo integration to the Bochner's theorem.

$$\begin{aligned} k(x-y) &\approx \frac{1}{M}\sum_{i=1}^{M} \cos(2\pi {s_i}^T x)\cos(2\pi {s_i}^T y) + \sin(2\pi {s_i}^T x)\sin(2\pi {s_i}^T y) && (3) \\ &= \phi(x_s)^T\phi(y_s) && (4) \end{aligned}$$

where $\boldsymbol{s} = \{s_i\}_{i=1}^{M}$ is the M sampled spectral points from the spectral density $p(S)$ and $\phi(x^{\boldsymbol{s}}) = \frac{1}{\sqrt{M}}\left[\cos 2\pi s_1^T x, .., \cos 2\pi s_M^T x, \sin 2\pi s_1^T x, .., \sin 2\pi s_M^T x\right] \in R^{1\times 2M}$.

2.3. Sparse Spectrum Approximation in Gaussian Process

Sparse spectrum GP (Lázaro-Gredilla et al., 2010) is a scalable method of GP regression with the approximated kernel by RFF. Given the dataset $X = \{x_1, .., x_n\}$ and $Y = \{y_1, .., y_n\}$ with the sampled spectral points $\boldsymbol{s}$ and the corresponding feature map $\Phi_{\boldsymbol{s}}(X) = [\phi(x_1^{\boldsymbol{s}}); ...; \phi(x_n^{\boldsymbol{s}})] \in R^{n\times 2M}$, this method trains the model to maximize the $\log p(Y|X, \boldsymbol{s})$.

$$\log p(Y|X, \boldsymbol{s}) = -\frac{1}{2}Y^T(\Phi_{\boldsymbol{s}}(X)\Phi_{\boldsymbol{s}}(X)^T + {\sigma_\epsilon}^2 I)^{-1}Y - \frac{1}{2}\log|\Phi_{\boldsymbol{s}}(X)\Phi_{\boldsymbol{s}}(X)^T + {\sigma_\epsilon}^2 I| - \frac{n}{2}\log 2\pi \tag{5}$$

This method reduces the computation time of computing inversion and determinant of kernel matrix to $O(nM^2)$ from $O(n^3)$.

3. Proposed Methodology

3.1. Generalized SM Kernel

Considering the spectral density of SM kernel (Wilson and Adams, 2013) is considered as the weighted sum of symmetric Gaussian distribution, we generalize SM Kernel by assuming each component's spectral density as exponential family distribution $p_q(S)$.

$$p_q(S;\theta_q) = h(s)\exp\left(\eta(\theta_q)\cdot T(S) - A(S)\right) \tag{6}$$

where $\eta(\theta_q)$ natural parameter, $T(S)$ sufficient statistic, and $A(S)$ normalizer. Then, the generalized SM kernel $k_{\text{GSM}}(x-y)$ can be defined as

$$k_{\text{GSM}}(x-y) = \sum_{q=1}^{Q} w_q k_q(x-y) \tag{7}$$

where $k_q(x-y)$ is corresponding kernel generated from (1) and (6).

3.2. Generalized SM Kernel Approximation

To approximate the generalized SM kernel, we use reparameterized Random Fourier Feature (R-RFF) for spectral S, i.e. $s \sim p(S)$ such that $s = g_q(\theta_q, \epsilon)$ for genereal parameter θ_q and random variable ϵ. Under the mild condition, this reparametrization can also be applied by natural parameter η_q because of one-to-one relation between θ_q and η_q, i.e., $\theta_q = \theta_q(\eta_q)$ and $g_q(\theta_q(\eta_q), \epsilon) = h_q(\eta_q, \epsilon)$ for some differentiable function h (Wainwright et al., 2008; Ruiz et al., 2016).

Using R-RFF, the feature map $\phi(x) = \left[\sqrt{w_1}\phi_{g_1(\theta_1,\epsilon)}(x), .., \sqrt{w_Q}\phi_{g_Q(\theta_Q,\epsilon)}(x)\right]$ can approximate $k_{\mathrm{GSM}}(x-y)$ as

$$k_{\mathrm{GSM}}(x-y) \approx \phi(x)\phi(y)^T \tag{8}$$

3.3. SoftMax Sampling for Spectral points

To reduce the variance of the unbiased estimator $\phi(x)\phi(y)^T$, we consider that each number of sampled spectral points is proportional to weight parameters of generalized SM kernel.

Theorem 1 (SoftMax sampling for Spectral points)
Given the defined estimator $\phi(x)\phi(y)^T$ above, let m_q be the number of sampled spectral points from $p_q(S)$ with $m = \sum_{q=1}^{Q} m_q$. The optimal ratio $p_q^* = \frac{m_q^*}{m}$ to minimize $\mathrm{Var}(\phi(x)\phi(y)^T)$ satisfies the following condition. Under the mild condition, p_q^* is proportional to the normalized weight parameters.

$$p_q^* = \frac{w_q \ \mathrm{std}(\cos 2\pi s_{q,1}^T(x-y))}{\sum_{q=1}^{Q} w_q \ \mathrm{std}(\cos 2\pi s_{q,1}^T(x-y))} \approx \frac{w_q}{\sum_{q=1}^{Q} w_q} \tag{9}$$

$$= \mathrm{SoftMax}([\log w_1, .., \log w_Q]) \tag{10}$$

Proof *See Appendix* ■

Example 1 (SM kernel Approximation in sense of General Parameter)
Given the SM kernel parameters $\{w_q, \mu_q, \sigma_q\}_{q=1}^{Q}$, let $\mathbf{s} = \cup_{q=1}^{Q}\{s_{q,i}\}_{i=1}^{m_q}$ be sampled spectral points by reparameterization $s_{q,i} = \mu_q + \sigma_q \circ \epsilon_i$ with $\epsilon_i \sim N(\epsilon; 0, I_P)$. The defined feature map $\phi_{SM}(x) = [\sqrt{w_1}\phi_1(x), .., \sqrt{w_Q}\phi_Q(x)] \in R^{2M}$ can approximate $k_{SM}(x-y)$ as follows

$$k_{SM}(x-y) \approx \phi_{SM}(x)\phi_{SM}(y)^T \tag{11}$$

Proof *See Appendix* ■

Example 2 (SM kernel Approximation in sense of Natural Parameter)
Let η_1 and η_2 be the natural parameter of Gaussian distribution. Then, η_1 and η_2 can represent the general parameter μ and Σ as

$$\Sigma = -\frac{1}{2}(\eta_2)^{-1} \quad , \quad \mu = -\frac{1}{2}(\eta_2)^{-1}\eta_1 \tag{12}$$

Under the assumption of diagonal Σ *with* $\mathrm{diag}(\Sigma) = \sigma^2$*, we can also sample the* $s_{q,i}$ *as*

$$s_{q,i} = -\frac{1}{2}(\eta_2^q)^{-1} \circ \eta_1^q + \sqrt{-\frac{1}{2}(\eta_2^q)^{-1}} \circ \epsilon_i \tag{13}$$

where η_1^q *and* η_2^q *are the natural parameters of* $N(S;\mu_q,\sigma_q^2)$ *and* $\eta_2^q \in R^P$*.*

3.4. Regularized Sparse Spectrum Approximation by R-RFF and SGVB

Based on R-RFF, SGVB (Kingma and Welling, 2013) facilitates to stochastically train the kernel hyperparameters of GP model by the following regularized lower bound.

$$\begin{aligned} \log p(Y|X) &= \log \int p(Y|f)p(f|X,S)p(S)df\,dS && (14)\\ &\geq \int \log p(Y|X,S)q(S)dS - KL(q(S)||P(S)) = \mathcal{L} && (15)\\ &\approx \frac{1}{K}\sum_{k=1}^{K} \log p(Y|X,\boldsymbol{s}^{(k)}) - KL(q(S)||p(S)) = \hat{\mathcal{L}}_K && (16) \end{aligned}$$

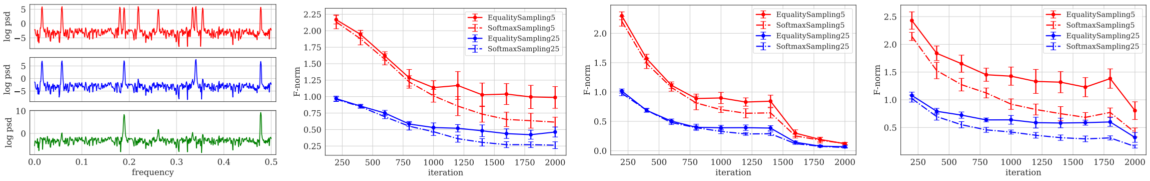
where $\boldsymbol{s}^{(k)}$ is the k th sampled spectral points from $q(S)$ by reparametrization trick. The $p(S)$ is the prior distribution of spectral density, whose parameter can be tuned by using empirical spectral density. $-KL(q(S)||p(S))$ prevents each spectral density distribution from collapsing during training.

4. Experiments

We generate the synthetic data of 10 combination of sinusoidal waveform with different amplitude by using Dirichlet distribution, i.e, $\frac{1}{\sum_{q=1}^{10} w_q}(w_1,..,w_{10}) \sim Dir_{10}(\theta)$.

$$y(t) = \sum_{q=1}^{10} w_q \sin(\alpha_q 2\pi t) + \beta_q \epsilon_t \tag{17}$$

where $\alpha_q \sim U(0,1)$, $\beta_q \sim U(0,.1)$ and $\epsilon_t \sim N(0,1)$. We consider three types of θ generated by Dirichlet distribution;'Almost Equal', 'Half Equal', and 'Rarely Equal'.



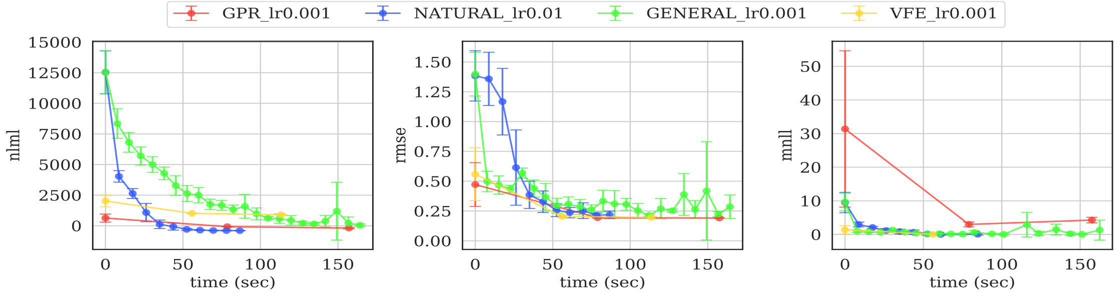
(a) log PSD (b) Almost Equal (c) Half Equal (d) Rarely Equal

Figure 1: (a) shows log power spectral density of synthetic dataset ; 'Almost Equal', 'Half Equal' and 'Rarely Equal'. (b), (c), and (d) reveal the effect on SM kernel approximation by SoftMax sampling under two cases; 'small' ($2 \times 5 \times Q$) for 'red' and 'big' ($2 \times 25 \times Q$) for 'blue' with $Q = 10$

In the first experiment, we validate that our proposed SoftMax sampling for spectral points could help approximate the SM kernel compared to the naive sampling approach that does not consider the weight of each mixture component spectral density $\{w_q\}_{q=1}^{Q}$.

To evaluate our method, we measure F-norm between true SM kernel and the approximate kernel applied by our SoftMax sampling during training, i.e, $\frac{\|K_{\theta_t} - \hat{K}_{\theta_t}\|_F}{\|K_{\theta_t}\|_F}$ for t iteration.

In Figure 1, (a) describes the log power spectral density for each setting; 'Almost Equal', 'Half Equal', and 'Rarely Equal'. Figures (b), (c), and (d) shows that applying SoftMax sampling for SM kernel approximation could reduce the error of approximation in both small and large number of sampled spectral points.



(a) Half Equal

Figure 2: Training Comparison with Benchmark Models; GPR for 'red', VFE for 'yellow', proposed General parameter approach for 'lime green', and Natural parameter approach for 'blue'

In the second experiment, we compare the proposed learning method with GP regression using SM kernel (GPR) (Rasmussen, 2004; Wilson and Adams, 2013) and Variational Inducing Inputs (VFE) (Titsias, 2009) to reveal that our approximation could learn SM kernel faster.

We use 'Half Equal' dataset defined in the first experiment and conduct the extrapolation task. For training, 2000 and 400 data points are used for training and test. We measure the Negative Log marginal Likelihood (nlml) of the training set, Root Mean Square Error (rmse), and Negative Mean Log Loss (nmll) of the test set in training. For experiment setting, we use 300 spectral points $(2 \times 15 \times 10(Q))$ for our methods. For VFE, we set 300 inducing points for fair computation comparison. We find the proper learning rate for each approach and set to 0.001 except for the natural parameter approach as 0.01 because it can exceptionally learn the dataset with a relatively fast learning rate.

Figure 2 shows that our approaches perform much more iteration during the training time. Also, the natural parameter approach converges the local optimal faster than other methods in nlml and rmse.

5. Discussion

We show using R-RFF in the sense of natural parameter is likely to train the SM kernel faster. We think that geometric information of the natural parameter approach makes the model learnable at fast learning rates and then leads the fast convergence. In a further study, we will focus on explaining why natural parameter approximation could lead the faster convergence and then validate its strong points in real data experiments.

References

Salomon Bochner. *Lectures on Fourier integrals.* Princeton University Press, 1959.

Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39(1):1–49, 2002.

Yarin Gal and Richard Turner. Improving the gaussian process sparse spectrum approximation by representing uncertainty in frequency inputs. 2015.

James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013.

Quang Minh Hoang, Trong Nghia Hoang, and Kian Hsiang Low. A generalized stochastic variational bayesian hyperparameter learning framework for sparse spectrum gaussian process regression. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Miguel Lázaro-Gredilla, Joaquin Quiñonero-Candela, Carl Edward Rasmussen, and Aníbal R Figueiras-Vidal. Sparse spectrum gaussian process regression. *Journal of Machine Learning Research*, 11:1865–1881, 2010.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

Francisco R Ruiz, Michalis Titsias RC AUEB, and David Blei. The generalized reparameterization gradient. In *Advances in neural information processing systems*, pages 460–468, 2016.

Hugh Salimbeni, Ching-An Cheng, Byron Boots, and Marc Deisenroth. Orthogonally decoupled variational gaussian processes. In *Advances in neural information processing systems*, pages 8711–8720, 2018.

Jeanette P Schmidt, Alan Siegel, and Aravind Srinivasan. Chernoff–hoeffding bounds for applications with limited independence. *SIAM Journal on Discrete Mathematics*, 8(2): 223–250, 1995.

Dougal J Sutherland and Jeff Schneider. On the error of random fourier features. *arXiv preprint arXiv:1506.02785*, 2015.

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

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Andrew Wilson and Ryan Adams. Gaussian process kernels for pattern discovery and extrapolation. In *International Conference on Machine Learning*, pages 1067–1075, 2013.

Appendix A. Proof of Theorem 1

Proof Let m_q be the number of spectral points sampled from $p(S;\mu_q,\sigma_q)$. We define the estimator $\psi(m_1,..,m_Q)$ as

$$\psi(m_1,..,m_Q) = \sum_{q=1}^{Q} \frac{w_q}{m_q} \sum_{i=1}^{m_q} \cos\left(2\pi s_{q,i}^T (x-y)\right)$$

so that $\mathrm{E}_{q(S)}\left[\psi(m_1,..,m_Q\right] = k_{\mathrm{SM}}(x-y)$. Our objective is to find the optimal $\{m_1^*,..,m_Q^*\}$ to minimize the variance $\psi(m_1,..,m_Q)$ by solving the following optimization problem:

$$\begin{aligned} \min_{m_1,..,m_Q} \quad & \mathrm{Var}\left(\psi(m_1,..,m_Q)\right) \\ \text{s.t} \quad & \sum_{q=1}^{Q} m_q = m \quad \forall m_q \in \mathbb{Z}^+ \end{aligned}$$

Since this optimization is somewhat tricky integer programming, we take relaxation by transforming variable as $p_q = \frac{m_q}{m}$ which induces the condition $\sum_{q=1}^{Q} p_q = 1$ from $\sum_{q=1}^{Q} m_q = m$. Then, above optimization problem is modified as

$$\begin{aligned} \min_{m_1,..,m_Q} \quad & \mathrm{Var}\left(\psi(p_1,..,p_Q)\right) \\ \text{s.t} \quad & \sum_{q=1}^{Q} p_q = 1 \quad \forall p_q \in [0,1] \end{aligned}$$

where $\mathrm{Var}\left(\psi(p_1,..,p_Q)\right) = \sum_{q=1}^{Q} \frac{w_q^2}{m p_q^2} \mathrm{Var}\left(\cos 2\pi s_{q,1}^T (x-y)\right)$ because $\{s_{q,i}\}$ for $i = 1,..,m_q$ is independently sampled. The optimal solution of this problem can be obtained by applying Lagrangian method because this optimization is convex optimization problem. Let $\mathcal{L}(m_1,..,m_Q,\lambda)$ be a Lagrangian operator with the multiplier λ.

$$\mathcal{L}(m_1,..,m_Q,\lambda) = \sum_{q=1}^{Q} \frac{w_q^2}{m_q} \mathrm{Var}\left(\cos 2\pi s_{q,1}^T (x-y)\right) + \lambda \left(\sum_{q=1}^{Q} p_i - 1 \right)$$

Solving the following conditions $\frac{\partial \mathcal{L}(m_1,..,m_Q,\lambda)}{\partial \lambda} = 0$ and $\frac{\partial \mathcal{L}(m_1,..,m_Q,\lambda)}{\partial p_q} = 0$ for $q = 1,..,Q$ leads to the optimal solution $\{p_1^*,..,p_Q^*\}$.

$$p_q^* = \frac{w_q \ \mathrm{std}(\cos 2\pi s_{q,1}^T(x-y))}{\sum_{q=1}^{Q} w_q \ \mathrm{std}(\cos 2\pi s_{q,1}^T(x-y))}$$

where std $\left(\cos 2\pi s_{q,1}^T(x-y)\right)$ is obtained as $\sqrt{\frac{1}{2}\Big(1 + k_q\left(2(x-y)\right) - 2k_q^2\left(x-y\right)\Big)}$ (Sutherland and Schneider, 2015).

■

Appendix B. Proof of Example 1

$$\Pr\left(\sup_{x,y\in\mathcal{X}} \left|\phi_{\mathrm{SM}}(x)\phi_{\mathrm{SM}}(y)^T - k_{\mathrm{SM}}(x-y)\right| \geq \epsilon\right) \ \leq \ 2^8\left(\frac{\sigma_p l}{\epsilon}\right)^2 \exp\left(-\frac{M\epsilon^2}{8(d+2)}\right)$$

Proof This proof follows the basic structure of proof (Rahimi and Recht, 2008). Given the finite dataset $\mathcal{M}$, we define $\mathcal{M}_\tau = \{\tau \mid \tau = x - y \ \forall x, y \in \mathcal{M}\}$. Since $\mathcal{M}$ is compact set because of $\mathcal{M}$'s finiteness, we can define the finite ϵ-net $\{B_{\tau_i}(r)\}_{i=1}^K$ with the center τ_i and the radius r such that $\mathcal{M}_\tau \subset \bigcup_{i=1}^K B_{\tau_i}(r)$ where the number of ϵ-net K is bounded by $4(\frac{\mathrm{diam}(\mathcal{M})}{r})^d$ (Cucker and Smale, 2002).

Also, we define the feature map $\phi_q(x)$ induced from $N(S;\mu_q,\Sigma_q)$ which

$$\phi_q(x) = \sqrt{\frac{1}{m_q}}\left[\cos\left(2\pi s_{q,1}^T x\right), .., \cos\left(2\pi s_{q,m_q}^T x\right), \sin\left(2\pi s_{q,1}^T x\right), .., \sin\left(2\pi s_{q,m_q}^T x\right)\right]$$

define the ϕ_{SM} such that $\mathrm{E}\left[\phi_{\mathrm{SM}}(x)^T\phi_{\mathrm{SM}}(y)\right] = k_{\mathrm{SM}}(x-y)$.

$$\phi_{\mathrm{SM}}(x) = [\sqrt{w_1}\phi_1(x), \sqrt{w_2}\phi_2(x), .., \sqrt{w_Q}\phi_Q(x)] \in R^{1\times 2(\sum_{q=1}^{Q} m_q)}$$

$$\begin{aligned}
\mathrm{E}\left[\phi_{\mathrm{SM}}(x)^T\phi_{\mathrm{SM}}(y)\right] &= \mathrm{E}\left[\sum_{q=1}^{Q}\frac{w_q}{m_q}\sum_{i=1}^{m_q}\cos\left(2\pi s_{q,i}^T(x-y)\right)\right] \\
&= \sum_{q=1}^{Q} w_q \mathrm{E}\left[\frac{1}{m_q}\sum_{i=1}^{m_q}\cos\left(2\pi s_{q,i}^T(x-y)\right)\right] \\
&= \sum_{q=1}^{Q} w_q k_q(x-y) = k_{\mathrm{SM}}(x-y)
\end{aligned}$$

Here, what we are going to prove is $\left|\phi_{\mathrm{SM}}(x)\phi_{\mathrm{SM}}(y)^T - k_{\mathrm{SM}}(x-y)\right| \leq \epsilon$ for $\forall x, y \in \mathcal{M}$ in probability convergence sense, as $m = \sum_{q=1}^{Q} m_q \to \infty$.

To show this statement, we define the error function $f(x-y) = \phi_{\mathrm{SM}}(x)\phi_{\mathrm{SM}}(y)^T - k(x-y)$ on $\mathcal{M}_\tau$ and denote L_f to be Lipschitz constant of f, i.e, $L_f = \sup_{\tau\in\mathcal{M}_\tau}\|\nabla f(\tau)\|$ because f is continuously differentiable function with respect to τ. Then, we will verify the following two conditions in probability:

1. $\big|f(\tau_i)\big| \leq \frac{\epsilon}{2}$ for all $i = 1, .., K$
2. $L_f \leq \frac{\epsilon}{2r}$

After proving these two conditions, intuitively, we can show that for $\forall \tau \in \mathcal{M}_\tau$, there exists for some τ_i s.t $\tau \in B_{\tau_i}(r)$ and $|f(\tau_i)| \leq \frac{\epsilon}{2}$. Then, $|f(\tau)| \leq |f(\tau) - f(\tau_i)| + |f(\tau_i)| \leq L_f \|\tau - \tau_i\| + |f(\tau_i)| \leq \frac{\epsilon}{2r} r + \frac{\epsilon}{2} \leq \epsilon$.

Proof for the condition 1

Applying the Chernoff-Hoeffding's inequality (Schmidt et al., 1995) to the event set $\big\{|f(\tau)| \geq \frac{\epsilon}{2}\big\}$ for $\tau \in \mathcal{M}_\tau$ can bound the probability of the event $\big\{|f(\tau)| \geq \frac{\epsilon}{2}\big\}$ where each term of $\phi_{\text{SM}}(x)\phi_{\text{SM}}(y)^T = \sum_{q=1}^{Q} \frac{w_q}{m_q} \sum_{i=1}^{m_q} \cos\Big(2\pi s_{q,i}^T (x-y)\Big)$ is bounded as $\big|\frac{w_q}{m_q} \cos\Big(2\pi s_{q,i}^T (x-y)\Big)\big| \leq \frac{w_q}{m_q}$ for $q = 1, .., Q$ and $i = 1, .., M$.

$$\Pr\left(\big|f(\tau)\big| \geq \frac{\epsilon}{2}\right) \leq 2\exp\left(\frac{-\epsilon^2}{8\sum_{q=1}^{Q} \frac{w_q^2}{m_q}}\right)$$

Complement event $\left(\bigcap_{i=1}^{T} \left\{|f(\Delta_{x_i,y_i})| \leq \frac{\epsilon}{2}\right\}\right)^c = \bigcup_{i=1}^{T} \left\{|f(\Delta_{x_i,y_i})| \geq \frac{\epsilon}{2}\right\}$ with the previous result and the union bound induces the following bound:

$$\Pr\left(\bigcup_{i=1}^{K} \left\{|f(\tau_i)| \geq \frac{\epsilon}{2}\right\}\right) \leq \sum_{i=1}^{K} \Pr\left(\big|f(\tau_i)\big| \geq \frac{\epsilon}{2}\right) \leq 2K\exp\left(\frac{-\epsilon^2}{8\sum_{q=1}^{Q} \frac{w_q^2}{m_q}}\right)$$

This result implies that $|f(\tau_i)| \leq \frac{\epsilon}{2}$ for all $i = 1, .., K$ with the probability $1 - 2K\exp\left(\frac{-\epsilon^2}{8\sum_{q=1}^{Q} \frac{w_q^2}{m_q}}\right)$.

Proof for the condition 2
To bound the probability of the event $\{L_f \geq \frac{\epsilon}{2r}\}$, the Markov inequality leads that the bound of $\mathrm{E}[L_f^2]$ can bound the probability of the event $\{L_f \geq \frac{\epsilon}{2r}\}$.

$$\Pr\left(L_f \geq \frac{\epsilon}{2r}\right) \leq \Pr\left(L_f^2 \geq (\frac{\epsilon}{2r})^2\right) \leq (\frac{\epsilon}{2r})^{-2}\mathbb{E}[L_f^2]$$

Let $\tau^* = x^* - y^*$ be the optimal elements in $\mathcal{M}_\tau$ to satisfy $L_f = \|\nabla f(\tau^*)\|$ and the existence of τ^* can be verified by the compactness for $\mathcal{M}_\tau$.

$$\begin{aligned}
\mathbb{E}[L_f^2] &= \mathbb{E}[\|\nabla f(\tau^*)\|^2] = \mathbb{E}[\left\|\nabla \phi_{\mathrm{SM}}(x^*)\phi_{\mathrm{SM}}(y^*)^T - \nabla k_{\mathrm{SM}}(x^* - y^*)\right\|^2] \\
&= \mathbb{E}[\left\|\nabla \phi_{\mathrm{SM}}(x^*)\phi_{\mathrm{SM}}(y^*)^T\right\|^2] - \mathbb{E}[\|\nabla k(x^*, y^*)\|]^2 \\
&\leq \mathbb{E}[\left\|\nabla \phi_{\mathrm{SM}}(x^*)\phi_{\mathrm{SM}}(y^*)^T\right\|^2] \\
&= \mathbb{E}\left[\left\|\sum_{q=1}^{Q} \frac{w_q}{m_q} \sum_{i=1}^{m_q} -\sin\left(2\pi s_{q,i}^T(x^* - y^*)\right) \circ 2\pi s_{q,i}\right\|^2\right] \\
&\leq \mathbb{E}\left[\left(\sum_{q=1}^{Q}\sum_{i=1}^{m_q} \frac{w_q}{m_q} \left\|-\sin\left(2\pi s_{q,i}^T(x^* - y^*)\right) \circ 2\pi s_{q,i}\right\|\right)^2\right] \\
&= \mathbb{E}\left[\sum_{q=1}^{Q}(\frac{w_q}{m_q})^2 \left(\sum_{i=1}^{m_q} \left\|-\sin\left(2\pi s_{q,i}^T(x^* - y^*)\right) \circ 2\pi s_{q,i}\right\|\right)^2\right] \\
&= \sum_{q=1}^{Q}(\frac{{w_q}^2}{m_q})\mathbb{E}\left[\left\|-\sin\left(2\pi s_{q,1}^T(x^* - y^*)\right) \circ 2\pi s_{q,1}\right\|^2\right] \\
&\leq \sum_{q=1}^{Q}(\frac{{w_q}^2}{m_q})\mathbb{E}\left[\left\|1 \circ 2\pi s_{q,1}\right\|^2\right] = \sum_{q=1}^{Q}(\frac{4\pi^2{w_q}^2}{m_q})\left(\|\mu_q\|^2 + \|\sigma_q\|^2\right)
\end{aligned}$$

Thus, the probability of the event $\{L_f \geq \frac{\epsilon}{2r}\}$ is bounded as

$$\Pr(L_f \geq \frac{\epsilon}{2r}) \leq \left(\frac{2r}{\epsilon}\right)^2 \sum_{q=1}^{Q}(\frac{4\pi^2{w_q}^2}{m_q})\left(\|\mu_q\|^2 + \|\sigma_q\|^2\right)$$

The combination of the result for the condition 1 and condition 2 proves the following statement with $K \leq 4(\frac{\mathrm{diam}(\mathcal{M})}{r})^d$

$$\begin{aligned}
&\Pr\left(\sup_{x,y\in\mathcal{M}} \left|\phi_{\mathrm{SM}}(x)\phi_{\mathrm{SM}}(y)^T - k_{\mathrm{SM}}(x-y)\right| \leq \epsilon\right) \\
&\leq 1 - 8(\frac{\mathrm{diam}(\mathcal{M})}{r})^d \exp\left(\frac{-\epsilon^2}{8\sum_{q=1}^{Q}\frac{w_q^2}{m_q}}\right) - \left(\frac{2r}{\epsilon}\right)^2 \sum_{q=1}^{Q}(\frac{4\pi^2{w_q}^2}{m_q})\left(\|\mu_q\|^2 + \|\sigma_q\|^2\right)
\end{aligned}$$

The above bound has the form $1-k_1r^{-d}-k_2r^2$. Setting $r = \frac{k_1}{k_2}^{\frac{1}{d=2}}$ turn this to $1-2k_2^{\frac{d}{d+2}}k_1^{\frac{2}{d+2}}$ where $k_1 = 8\text{diam}(\mathcal{M})^d \exp \frac{-\epsilon^2}{8\sum_{q=1}^{Q}\frac{w_q^2}{m_q}}$ and $k_2 = \frac{4}{\epsilon^2}\sum_{q=1}^{Q}\frac{4\pi^2 w_q^2}{m_q}\left(\|\mu_q\|^2 + \|\sigma_q\|^2\right)$

$$\leq 1 - 2^6\pi^2\left(\sqrt{\sum_{q=1}^{Q}(\frac{4\pi^2 {w_q}^2}{m_q})\left(\|\mu_q\|^2 + \|\sigma_q\|^2\right)}\text{diam}(\mathcal{M})\right)^2 \exp\left(\frac{-\epsilon^2}{4(d+2)\sum_{q=1}^{Q}\frac{w_q^2}{m_q}}\right)$$

■

Appendix C. Derivation of Regularized Sparse Spectrum Approximation by R-RFF and SGVB

Proof We consider the lower bound of log marginal likelihood with the candidate distribution $q(S)$ where $S = (S_{1,1},..,S_{1,m_1},..,S_{Q,1},..,S_{Q,m_Q})$ is the random sample used in R-RFF for kernel approximation. Then, we can derive the lower bound $\mathcal{L}$ as follows:

$$\begin{aligned}
\log p(Y|X) &= \log \iint p(Y|f)p(f|X,S)\frac{p(S)}{q(S)}q(S)df\,dS \\
&= \log \int p(Y|X,S)\frac{p(S)}{q(S)}q(S)dS \\
&\geq \int \log \left(p(Y|X,S)\frac{p(S)}{q(S)}\right)q(S)dS \\
&= \int \log p(Y|X,S)q(S) + \log \frac{p(S)}{q(S)}q(S)dS \\
&= \int \log p(Y|X,S)q(S)dS - KL(q(S)||P(S)) = \mathcal{L}
\end{aligned}$$

Applying the Stochastic Gradient Variational Bayes (SGVB) (Kingma and Welling, 2013) to $\mathcal{L}$ with the reparametrizable distribution $q(S)$, leads to the following unbiased estimator $\hat{\mathcal{L}_K}$.

$$\hat{\mathcal{L}}_K = \frac{1}{K}\sum_{i=1}^{K}\log p(Y|X,s^{(i)}) - KL(q(S)||P(S))$$

where $s^{(i)}$ is i-th sampled spectral points from $q(S)$. ■

Appendix D. Algorithm

Algorithm 1: Example 1 learning by SGVB

```
Input: $X, Y, \theta = \{w_q, \mu_q, \sigma_q\}_{q=1}^{Q}$,$m$, and $K$
Output: $\theta^* = \{w_q^*, \mu_q^*, \sigma_q^*\}_{q=1}^{Q}$
for $t$=1,..,$T$ do
    Set the temperature $\tau = T/t$
    Sample the spectral points $\{\boldsymbol{S}^{(k)}\}_{k=1}^{K}$
    for $k$ = 1,...,$K$ do
        for $q$ = 1,...,$Q$ do
            Get $\#m_q$ by $\mathrm{SoftMax}_\tau(\log w_1, .., \log w_Q)$
            Sample $\#m_q$ spectral points from $q_q(S_q)$
                $$\epsilon_i \sim N(\epsilon; 0, I)$$
                $$s_{q,i} = \mu_q + \sigma_q \circ \epsilon_i$$
        end
        $\boldsymbol{s}^{(k)} = \cup_{q=1}^{Q} \{s_{q,i}\}_{i=1}^{m_q}$
    end
    Get Monte-Carlo estimated gradients of $\hat{\mathcal{L}}_K$ by
    the sampled spectral points $\{\boldsymbol{s}^{(k)}\}_{k=1}^{K}$
        $$\left\{ \frac{\partial \hat{\mathcal{L}}_K}{\partial w_q}, \frac{\partial \hat{\mathcal{L}}_K}{\partial \mu_q}, \frac{\partial \hat{\mathcal{L}}_K}{\partial \sigma_q} \right\}_{q=1}^{Q}$$
    Update $\{w_q, \mu_q, \sigma_q\}_{q=1}^{Q}$ by ADAM method with the estimated
    gradients
end
```