3

ORTHOGONALLY-DECOUPLED SPARSE GAUSSIAN PROCESSES WITH SPHERICAL NEURAL NETWORK ACTIVATION FEATURES

3.1 INTRODUCTION

Gaussian processes (GPs) provide a versatile and robust framework for modeling unknown functions, offering data efficiency, flexible encoding of prior beliefs, and reliable uncertainty estimation. Their broad application in sequential decision-making makes them invaluable in diverse fields of ML and optimisation.

In spite of their many advantages, GPs are often compared unfavourably to deep NNs for their poor scalability to large datasets, and their inability to capture rich hierarchies of abstract representations [28, 186, 281]. While GPs are the infinite-width limit of NNs and therefore, in theory, have infinitely more basis functions [180], these basis functions are static and fully determined by the covariance function [152]. This makes it difficult for GPs to flexibly adapt to complex and structured data from which it is beneficial for the basis functions to learn and encode useful representations.

Considerable research effort has been devoted to sparse approximations for GPs [45, 198, 220, 228]. Not least of these is sparse variational GPs (SVGPs) [95, 96, 256], which we examined in Section 2.4. Such advances have not only improved the scalability of GPs, but also unlocked more flexibility in model specification. In particular, the use of *inter-domain* inducing variables in SVGP [136] effectively equips the GP approximation with data-dependent basis functions. Recent works have exploited this to construct a new family of SVGP models in which the basis functions correspond to activations of a feed-forward NN [65, 246]. By stacking multiple layers to form a DGP [48], the propagation of the predictive distribution accurately resembles a forward-pass through a deep NN.

In this chapter, we show that while this approach results in a posterior predictive with a more expressive mean, the variance estimate is typically less accurate and tends to be over-dispersed. Additionally, we examine some practical challenges associated with this method, such as limitations on the use of certain popular kernel and NN activation choices. To address these issues, we propose an extension that aims to mitigate these limitations. Specifically, when viewed from the function-space perspective, the posterior predictive of SVGP depends on a single set of basis functions that is determined by only a finite collection of inducing variables. Recent advances introduce an

orthogonal set of basis functions as a means of capturing additional variations remaining from the standard basis [36, 217, 224]. We extend this framework by introducing inter-domain variables to construct more flexible data-dependent basis functions for both the standard and orthogonal components. In particular, we show that incorporating NN activation inducing functions under this framework is an effective way to ameliorate the aforementioned shortcomings. Our experiments on numerous benchmark datasets demonstrate that this extension leads to improvements in predictive performance against comparable alternatives.

3.2 INTER-DOMAIN INDUCING FEATURES

Recall from Equation (2.24) that the test predictive density at unseen points $\mathbf{f}_* \triangleq f(\mathbf{X}_*)$ is

$$q(\mathbf{f}_*) = \mathcal{N}\left(\mathbf{f}_* \,|\, \mathbf{Q}_{*\mathbf{u}}\mathbf{m}_\mathbf{u}, \mathbf{K}_{**} - \mathbf{Q}_{*\mathbf{u}}(\mathbf{K}_{\mathbf{uu}} - \mathbf{C}_\mathbf{u})\mathbf{Q}_{\mathbf{u}*}\right), \tag{3.1}$$

where parameters $\mathbf{m}_\mathbf{u}$ and $\mathbf{C}_\mathbf{u}$ are free parameters. In the RKHS associated with k, this predictive density has a dual representation in which the mean and covariance share the same basis determined by $\mathbf{u}$ [36, 217]. More specifically, the basis function is effectively the vector-valued function $\mathbf{k}_\mathbf{u} : \mathcal{X} \to \mathbb{R}^M$ whose m-th component is defined as

$$[\mathbf{k}_\mathbf{u}(\mathbf{x})]_m \triangleq \operatorname{Cov}\left(f(\mathbf{x}), u_m\right). \tag{3.2}$$

In the standard definition of inducing points as presented in Section 2.4.2,

$$[\mathbf{k}_\mathbf{u}(\mathbf{x})]_m = k(\mathbf{z}_m, \mathbf{x}),$$

therefore, the basis function is solely determined by the covariance function k and the local influence of pseudo-input $\mathbf{z}_m$.

inter-domain inducing features

Inter-domain inducing features are a generalisation of standard inducing variables in which each variable u_m is defined through the transformation of $f(\cdot)$ by

$$u_m \triangleq L_m[f].$$

for some linear operator $L_m : \mathbb{R}^\mathcal{X} \to \mathbb{R}$. A particularly useful operator is the integral transform,

$$L_m[f] \triangleq \int_\mathcal{X} f(\mathbf{x})\phi_m(\mathbf{x})\,\mathrm{d}\mathbf{x},$$

which was originally employed by Lázaro-Gredilla and Figueiras-Vidal [136]. Refer to the manuscript of Wilk et al. [277] for a more thorough and contemporary treatment. A closely related form is the scalar projection of f onto some ϕ_m in the RKHS $\mathcal{H}$ of k,

$$L_m[f] \triangleq \langle f, \phi_m \rangle_\mathcal{H}, \tag{3.3}$$

which leads to

$$[\mathbf{k}_{\mathbf{u}}(\mathbf{x})]_m = \phi_m(\mathbf{x})$$

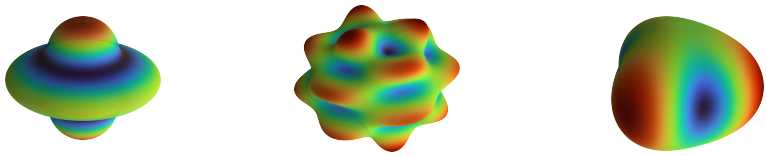
by the reproducing property of the RKHS. This, in effect, equips the GP approximation with adaptive basis functions ϕ_m that are not solely determined by a fixed kernel, and suitable choices can lead to sparser representations and considerable computational benefits [25, 64, 94, 247].

reproducing property

3.2.1 *Spherical Harmonics Inducing Features*

An instance of inter-domain features in the form of Equation (3.3) are the variational Fourier features (VFFS) [94], in which ϕ_m form an orthogonal basis of trigonometric functions. This formulation offers significant computational advantages but scales poorly beyond a small handful of dimensions. To address this, Dutordoir, Durrande, and Hensman [64] propose a generalisation of VFFS using the spherical harmonics for ϕ_m, which can be viewed as a multi-dimensional extension of the Fourier basis.

spherical harmonics



(a) Zonal: $Y_{4,0}$ (b) Tesseral: $Y_{8,3}$ (c) Sectorial: $Y_{3,3}$

Figure 3.1: A visual representation of three different *surface harmonics of the first kind*. This set of examples originates from the monograph of Efthimiou and Frye [66].

The construction relies on the Mercer's decomposition of *zonal* kernels, which can be seen as the analog of stationary kernels in Euclidean spaces, but for hyperspheres. They can be expressed as $k(\mathbf{x}, \mathbf{x}') = \kappa\left(\tilde{\mathbf{x}}^\top \tilde{\mathbf{x}}'\right)$ for some *shape function* $\kappa : [-1, 1] \to \mathbb{R}$, where $\tilde{\boldsymbol{\eta}} \triangleq \frac{\boldsymbol{\eta}}{\|\boldsymbol{\eta}\|} \in \mathbb{S}^{d-1}$ for any $\boldsymbol{\eta} \in \mathbb{R}^d$. Loosely speaking, just as stationary kernels are determined by the *distance* between inputs, zonal kernels depend only on the *angle* between inputs.

Mercer's decomposition

zonal kernels

The spherical harmonics form an orthonormal basis on $L_2(\mathbb{S}^{d-1})$ consisting of the eigenfunctions of the kernel operator $\mathcal{K}$,

$$\mathcal{K} Y_{\ell,j} = a_\ell Y_{\ell,j},$$

where $Y_{\ell,j}$ is the spherical harmonic of level ℓ and order j, and a_ℓ is the corresponding eigenvalue, or, Fourier coefficient.

A visualisation is shown in Figure 3.1. The spherical harmonics are tricky to visualise not only in higher dimensions, but even in

three dimensions, because they are, in general, complex-valued. Here we show, in three dimensions, several *surface harmonics of the first kind*, which correspond to the real or imaginary parts of the spherical harmonics, depending on the value of j. The surface harmonics can be further divided into the categories of *zonal* which are of the form $Y_{\ell,0}$, *tesseral* $Y_{\ell,j}$ for $j \neq 0$, and *sectorial* $Y_{\ell,\ell}$. Each of the surface harmonics shown here is a representative example of its respective category.

Conveniently, by the Funk-Hecke theorem, the Fourier coefficient a_ℓ amounts to the one-dimensional integral

$$a_\ell = \frac{\Omega_d}{C_\ell^{(\alpha)}(1)} \int_{-1}^{1} \kappa(t) C_\ell^{(\alpha)}(t)(1-t^2)^{\frac{d-3}{2}} \, \mathrm{d}t,$$

where $C_\ell^{(\alpha)}$ is the Gegenbauer polynomial of degree ℓ and $\alpha \triangleq (d-1)/2$. Now, the number $J(d,\ell)$ of spherical harmonics that exist at a given level ℓ is determined by the multiplicity of eigenvalue a_ℓ.

Thus, $\kappa(t)$ can be represented by

$$\kappa(t) = \|\boldsymbol{\xi}\| \|\boldsymbol{\xi}'\| \sum_{\ell=0}^{\infty} \sum_{j=1}^{J(d,\ell)} a_\ell Y_{\ell,j}(\tilde{\boldsymbol{\xi}}) Y_{\ell,j}(\tilde{\boldsymbol{\xi}}'), \tag{3.4}$$

where $t \triangleq \tilde{\boldsymbol{\xi}}^\top \tilde{\boldsymbol{\xi}}'$ for $\boldsymbol{\xi}, \boldsymbol{\xi}' \in \mathbb{R}^d$. We refer the reader to the manuscript of Dutordoir et al. [65, Appendix B] for a concise summary of spherical harmonics in multiple dimensions.

Importantly, Equation (3.4) directly yields a Mercer decomposition for zonal kernels. In particular, let λ_ℓ denote the Fourier coefficients associated with kernel k. This gives rise to the inter-domain features $\phi_m \triangleq Y_{\ell,j}$, where m indexes the pairs (ℓ, j). Crucially, because the spherical harmonics constitute an orthogonal system, this leads to a diagonal covariance

$$[\mathbf{K}_{\mathbf{uu}}]_{mm'} \triangleq \mathrm{Cov}\left(u_m, u_{m'}\right) = \lambda_m^{-1} \delta_{mm'},$$

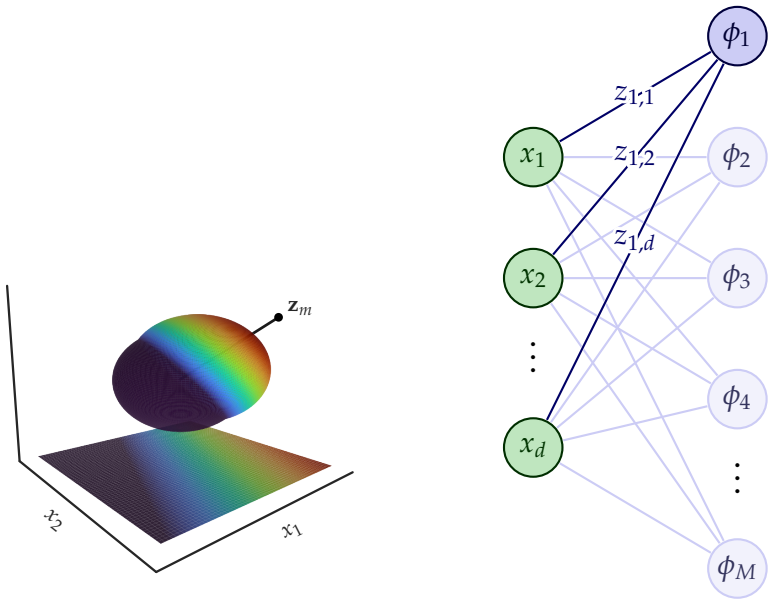
where $\lambda_m \triangleq \lambda_\ell$ and δ denotes the Kronecker delta.

3.2.2 *Spherical Neural Network Inducing Features*

The recent works of Dutordoir et al. [65] and Sun, Shi, and Grosse [246] aim to construct inter-domain features ϕ_m such that $\mathbf{k}_{\mathbf{u}}(\mathbf{x})$ in Equation (3.2) corresponds to a hidden layer in a feed-forward NN: $\sigma(\boldsymbol{\beta}\mathbf{x})$, for some $\boldsymbol{\beta} \in \mathbb{R}^{M \times d}$ and activation σ such as the SOFTPLUS or the rectified linear unit (RELU) function.

In particular, let $H_m : \mathcal{X} \to \mathbb{R}$ denote the output of the m-th hidden unit. Additionally, let us project this function onto the unit hypersphere,

$$H_m(\mathbf{x}) \triangleq \|\mathbf{z}_m\| \|\mathbf{x}\| \cdot \sigma\left(\frac{\mathbf{z}_m^\top \mathbf{x}}{\|\mathbf{z}_m\| \|\mathbf{x}\|}\right). \tag{3.5}$$



(a) An example RELU-activated hidden unit $H_m : \mathcal{X} \to \mathbb{R}$ visualised on the unit sphere in 3D and projected onto a plane.

(b) Architectural diagram of the basis functions $\mathbf{k_u}(\mathbf{x})$, which corresponds to the hidden layer of a feedforward NN when each $\phi_m \triangleq H_m$ represents a hidden unit.

Figure 3.2: Basis functions $\mathbf{k_u} : \mathcal{X} \to \mathbb{R}^M$ as hidden layers of a feedforward NN.

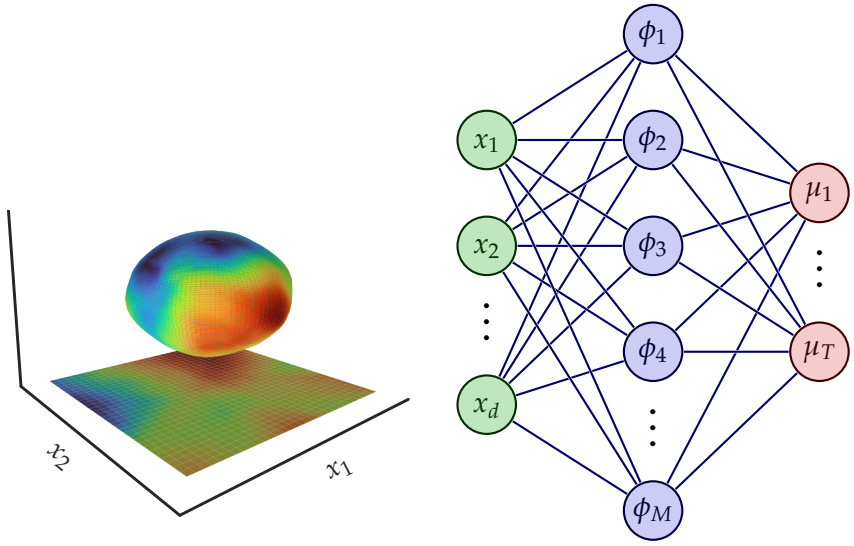
See Figure 3.2 for a visualisation of this function. Now, since this function is itself zonal, it can be represented in terms of the spherical harmonics as in Equation (3.4). Let ς_ℓ denote its associated Fourier coefficient. Thus, the inter-domain features can be defined as $\phi_m \triangleq H_m$, which leads to the covariance

$$[\mathbf{K_{uu}}]_{mm'} = \sum_{\substack{\ell=0: \\ \lambda_\ell \neq 0}}^{\infty} \frac{\varsigma_\ell^2}{\lambda_\ell} \frac{\ell + \alpha}{\alpha} C_\ell^{(\alpha)} \left(\frac{\mathbf{z}_m^\top \mathbf{z}_{m'}}{\|\mathbf{z}_m\| \|\mathbf{z}_{m'}\|} \right), \tag{3.6}$$

where λ_ℓ denotes the Fourier coefficients associated with kernel k.

We refer to this construction as the ACTIVATED SVGP. Notably, the posterior predictive mean of the ACTIVATED SVGP is equivalent to the output of a single-layer feedforward NN, as illustrated in Figure 3.3. Through this perspective, one can reason by analogy that the posterior predictive variance serves as a measure of uncertainty in the predictions of the NN. The ACTIVATED SVGP has been shown to produce competitive results, especially when multiple layers are composed to form a DGP [48]. In this configuration, the propagation of the predictive means closely emulates a forward-pass through a deep NN.

Despite these favorable properties, ACTIVATED SVGPS have several limitations when it comes to their use with common covariance functions. Before elaborating on them in Section 5.3, we discuss the orthogonally-decoupled GP framework on which our proposed extension relies.



(a) The predictive mean of an ACTIVATED SVGP model with RELU activation features, visualised on the unit sphere in 3D and projected onto a plane.

(b) Architectural diagram of the predictive means of a (multi-output) ACTIVATED SVGP model.

Figure 3.3: The predictive mean of an ACTIVATED SVGP model corresponds to a single-layer feedforward NN.

3.3 ORTHOGONALLY DECOUPLED INDUCING POINTS

Recent work has improved the efficiency of sparse GP methods through the structured decoupling of inducing variables [36, 217, 224]. This not only enables the use of more variables at a reduced computational expense but also allows for more flexibility in modelling the predictive mean and covariance independently. We focus on the general framework of Shi, Titsias, and Mnih [224] under which its predecessors can be subsumed.

In particular, let the random function $f(\mathbf{x})$ from Equation (2.14) be decomposed into the sum of two independent GPs,

$$f(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x}),$$

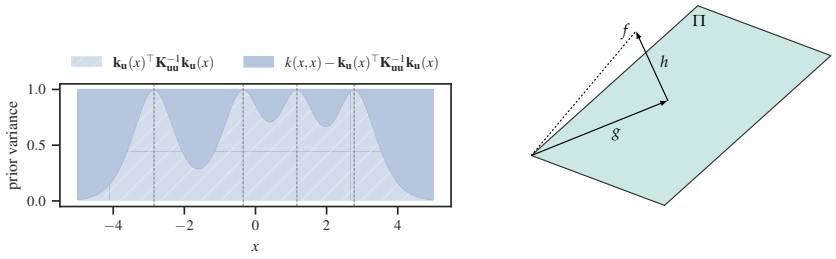
where

$$g(\mathbf{x}) \sim \mathcal{GP}(0, \mathbf{k}_\mathbf{u}^\top(\mathbf{x})\mathbf{K}_\mathbf{uu}^{-1}\mathbf{k}_\mathbf{u}(\mathbf{x}')), \quad h(\mathbf{x}) \sim \mathcal{GP}(0, s(\mathbf{x}, \mathbf{x}'))$$

and let the covariance function $s(\mathbf{x}, \mathbf{x}')$ be defined according to the Schur complement of $\mathbf{K}_\mathbf{uu}$,

$$s(\mathbf{x}, \mathbf{x}') \triangleq k(\mathbf{x}, \mathbf{x}') - \mathbf{k}_\mathbf{u}^\top(\mathbf{x})\mathbf{K}_\mathbf{uu}^{-1}\mathbf{k}_\mathbf{u}(\mathbf{x}'),$$

where $\mathbf{k}_\mathbf{u}$ is defined in Equation (3.2). Intuitively, one can view g as the projection of f onto $\mathbf{u}$, and $h \perp g$, i.e. h is *orthogonal* to g [94] in the statistical sense of linear independence [208]. See Figure 3.4



(a) Prior variance decomposed. The prior variance of $f(\mathbf{x})$ is $k(\mathbf{x}, \mathbf{x}) = \alpha$ for kernel amplitude $\alpha = 1$, which can be decomposed as the sum of the prior variances of $g(\mathbf{x})$ and $h(\mathbf{x})$. Vertical dashed lines indicate the location of inducing inputs $\mathbf{z}_m$ for $m = 1, \ldots, 4$. At these locations, the variance of $g(\mathbf{x})$ is one while that of $h(\mathbf{x})$ is zero.

(b) Orthogonal decomposition of function f wrt the hyperplane $\Pi \triangleq \{\boldsymbol{\alpha}^\top \mathbf{k_u}(\cdot); \boldsymbol{\alpha} \in \mathbb{R}^M\}$; function g is the orthogonal projection of f onto Π and function h is the residual component perpendicular to Π.

Figure 3.4: Function $f(\mathbf{x})$ decomposed as the sum of two independent GPs.

	W	**X**	**Z**
$f(\cdot)$	$\mathbf{v}$	$\mathbf{f}$	$\mathbf{u}$
$h(\cdot)$	$\mathbf{v}'$	$\mathbf{h}$	-

Table 3.1: Summary of notation: relationships between input locations and output variables.

for an illustration of the priors of $g(\mathbf{x})$ and $h(\mathbf{x})$ and a geometric interpretation in terms of vector subspaces.

Let $\mathbf{h}$ be the values of h at observed inputs $\mathbf{X}$, i. e. $\mathbf{h} \triangleq h(\mathbf{X})$. Then we have

$$p(\mathbf{h}) = \mathcal{N}(\mathbf{h} \,|\, \mathbf{0}, \mathbf{S_{ff}}),$$

where $\mathbf{S_{ff}} \triangleq \mathbf{K_{ff}} - \mathbf{Q_{ff}}$. This allows one to reparameterize $\mathbf{f} \sim p(\mathbf{f} \,|\, \mathbf{u})$ from Equation (2.22), for a given $\mathbf{u}$, as

$$\mathbf{f} = \mathbf{Q_{fu}}\mathbf{u} + \mathbf{h}, \qquad \mathbf{h} \sim p(\mathbf{h}). \tag{3.7}$$

The model's joint distribution can now be written as

$$p(\mathbf{y}, \mathbf{h}, \mathbf{u}) = p(\mathbf{y} \,|\, \mathbf{h}, \mathbf{u}) p(\mathbf{h}) p(\mathbf{u}),$$

where the likelihood is now

$$p(\mathbf{y} \,|\, \mathbf{h}, \mathbf{u}) = \mathcal{N}(\mathbf{y} \,|\, \mathbf{Q_{fu}}\mathbf{u} + \mathbf{h}, \beta^{-1}\mathbf{I}).$$

Next, *orthogonal* inducing variables $\mathbf{v}$, which represent the values of f at a collection of K orthogonal inducing locations $\mathbf{W} \triangleq [\mathbf{w}_1 \cdots \mathbf{w}_K]^\top$, are introduced. Similarly, inducing variables $\mathbf{v}'$ represent the values of h at $\mathbf{W}$. The reader may find it helpful to refer to Table 3.1 for a summary of the relationships between the input locations and the output variables defined thus far.

Now, by definition, $\mathbf{v}$ is linearly dependent on $\mathbf{v}'$

$$\mathbf{v} = \mathbf{Q}_{\mathbf{vu}}\mathbf{u} + \mathbf{v}', \tag{3.8}$$

where $\mathbf{Q}_{\mathbf{vu}} \triangleq \mathbf{K}_{\mathbf{vu}}\mathbf{K}_{\mathbf{uu}}^{-1}$, which is analogous to the relationship between $\mathbf{f}$ and $\mathbf{h}$ in Equation (3.7). Therefore, one need only be concerned with the treatment of $\mathbf{v}'$. The joint distribution of the model augmented by the variables $\mathbf{v}'$ now becomes

$$p(\mathbf{y}, \mathbf{h}, \mathbf{u}, \mathbf{v}') = p(\mathbf{y} \mid \mathbf{h}, \mathbf{u})p(\mathbf{u})p(\mathbf{h}, \mathbf{v}'),$$

where $p(\mathbf{h}, \mathbf{v}') = p(\mathbf{h} \mid \mathbf{v}')p(\mathbf{v}')$ for $p(\mathbf{v}') = \mathcal{N}(\mathbf{0}, \mathbf{S}_{\mathbf{vv}})$ and $p(\mathbf{h} \mid \mathbf{v}') = \mathcal{N}(\mathbf{h} \mid \mathbf{S}_{\mathbf{fv}}\mathbf{S}_{\mathbf{vv}}^{-1}\mathbf{v}', \mathbf{S}_{\mathbf{ff}} - \mathbf{S}_{\mathbf{fv}}\mathbf{S}_{\mathbf{vv}}^{-1}\mathbf{S}_{\mathbf{vf}})$, with

$$\mathbf{S}_{\mathbf{vf}} \triangleq \mathbf{K}_{\mathbf{vf}} - \mathbf{Q}_{\mathbf{vf}}, \quad \mathbf{Q}_{\mathbf{vf}} \triangleq \mathbf{Q}_{\mathbf{vu}}\mathbf{K}_{\mathbf{uu}}\mathbf{Q}_{\mathbf{uf}}, \tag{3.9}$$

$$\mathbf{S}_{\mathbf{vv}} \triangleq \mathbf{K}_{\mathbf{vv}} - \mathbf{Q}_{\mathbf{vv}}, \quad \mathbf{Q}_{\mathbf{vv}} \triangleq \mathbf{Q}_{\mathbf{vu}}\mathbf{K}_{\mathbf{uu}}\mathbf{Q}_{\mathbf{uv}}. \tag{3.10}$$

Let the variational distribution now be $q(\mathbf{h}, \mathbf{u}, \mathbf{v}') = p(\mathbf{h} \mid \mathbf{v}')q(\mathbf{u}, \mathbf{v}')$, where $q(\mathbf{u}, \mathbf{v}') \triangleq q(\mathbf{u})q(\mathbf{v}')$ and $q(\mathbf{v}') \triangleq \mathcal{N}(\mathbf{m}_{\mathbf{v}}, \mathbf{C}_{\mathbf{v}})$ for variational parameters $\mathbf{m}_{\mathbf{v}} \in \mathbb{R}^K$ and $\mathbf{C}_{\mathbf{v}} \in \mathbb{R}^{K \times K}$ s. t. $\mathbf{C}_{\mathbf{v}} \succeq 0$. This gives the test predictive density $q(\mathbf{f}_*) = \mathcal{N}(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_{**})$, where

$$\boldsymbol{\mu}_* \triangleq \mathbf{Q}_{*\mathbf{u}}\mathbf{m}_{\mathbf{u}} + \mathbf{S}_{*\mathbf{v}}\mathbf{S}_{\mathbf{vv}}^{-1}\mathbf{m}_{\mathbf{v}}, \tag{3.11}$$

$$\begin{aligned}\boldsymbol{\Sigma}_{**} \triangleq \mathbf{K}_{**} &+ \mathbf{Q}_{*\mathbf{u}}(\mathbf{C}_{\mathbf{u}} - \mathbf{K}_{\mathbf{uu}})\mathbf{Q}_{\mathbf{u}*} \\ &+ \mathbf{S}_{*\mathbf{v}}\mathbf{S}_{\mathbf{vv}}^{-1}(\mathbf{C}_{\mathbf{v}} - \mathbf{S}_{\mathbf{vv}})\mathbf{S}_{\mathbf{vv}}^{-1}\mathbf{S}_{\mathbf{v}*}.\end{aligned} \tag{3.12}$$

Thus seen, prediction incurs a cost of $\mathcal{O}(M^3 + K^3)$ in this framework.

Like the so-called ODVGP framework of Salimbeni et al. [217], when seen from the dual RKHS perspective, the predictive mean can be decomposed into a component that shares the same standard basis as the covariance, in addition to another component that is *orthogonal* to the standard basis. However, this framework extends ODVGP further by also decomposing the predictive covariance into parts corresponding to the standard and orthogonal bases. Accordingly, setting $\mathbf{C}_{\mathbf{v}} = \mathbf{S}_{\mathbf{vv}}$ recovers the ODVGP framework, and further setting $\mathbf{m}_{\mathbf{v}} = \mathbf{0}$ recovers the standard SVGP framework.

COVARIANCE STRUCTURE. Now, unlike $q(\mathbf{u}, \mathbf{v}')$, which factorizes according to the mean-field assumption, $q(\mathbf{u}, \mathbf{v})$ has a full covariance structure by virtue of the relationship described in Equation (3.8). Specifically, we have $q(\mathbf{u}, \mathbf{v}) = \mathcal{N}(\mathbf{m}_{\mathbf{uv}}, \mathbf{C}_{\mathbf{uv}})$, where

$$\mathbf{m}_{\mathbf{uv}} \triangleq \begin{bmatrix} \mathbf{m}_{\mathbf{u}} \\ \mathbf{Q}_{\mathbf{vu}}\mathbf{m}_{\mathbf{u}} + \mathbf{m}_{\mathbf{v}} \end{bmatrix},$$

and

$$\mathbf{C}_{\mathbf{uv}} \triangleq \begin{bmatrix} \mathbf{C}_{\mathbf{u}} & \mathbf{C}_{\mathbf{u}}\mathbf{Q}_{\mathbf{uv}} \\ \mathbf{Q}_{\mathbf{vu}}\mathbf{C}_{\mathbf{u}} & \mathbf{C}_{\mathbf{v}} + \mathbf{Q}_{\mathbf{vu}}\mathbf{C}_{\mathbf{u}}\mathbf{Q}_{\mathbf{uv}} \end{bmatrix}.$$

3.4 METHODOLOGY

We begin this section by outlining some of the limitations of ACTIVATED SVGPS that preclude the use of numerous kernels and inducing features, not the least of which being popular choices of kernels such as the SE kernel and the Matérn family of kernels, combined with NN inducing features with RELU activations.

The root cause of these issues can be seen in Figure 3.5, where the Fourier coefficients of various combinations of kernels and activation features are visualized. Specifically, for each combination, we compare the (root of the) kernel coefficients $\sqrt{\lambda_\ell}$ against the feature coefficients ς_ℓ at increasing levels $\ell = 1, \ldots, 35$. The posterior predictives that result from fitting ACTIVATED SVGP models with these combinations are shown in Figure 3.6. We consider the Matérn-5/2 kernel as our running example, but the analysis extends to all stationary kernels.

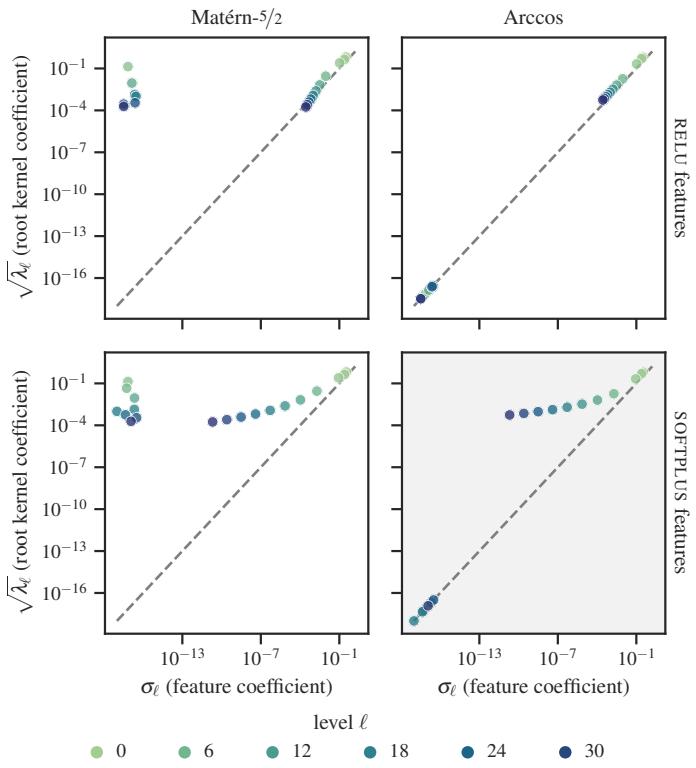


Figure 3.5: Comparison of the Fourier coefficients of various kernels and activation features for increasing levels $\ell = 1, \ldots, 35$.

SPECTRA MISMATCH. For the Matérn kernel (left column of panes in Figures 3.5 and 3.6), we see that there are multiple levels ℓ at which the feature coefficients are zero while the corresponding kernel coefficients are nonzero. Such discrepancies in the spectra yields a poor Nyström approximation $\mathbf{Q}_{\mathbf{ff}}$ that fails to fully capture the prior covariance $\mathbf{K}_{\mathbf{ff}}$ induced by the kernel, which subsequently leads to the

overestimation of the predictive variance and therefore a suboptimal ELBO. In contrast, the Arccos kernel does not suffer from this pathology.

RKHS INNER PRODUCT. The RKHS inner product associated with zonal kernels in general is a series consisting of ratios of Fourier coefficients. Since the RELU feature coefficients (top row of panes in Figures 3.5 and 3.6) decay at the same rate as the square root of the kernel coefficients, this results in a divergent series which in turn renders the RKHS inner product indeterminate. In contrast, the feature coefficients of the comparatively smoother SOFTPLUS activation (bottom row of panes in Figures 3.5 and 3.6) decay at a much faster rate, and thus yields a well-defined RKHS inner product. For the reasons outlined above, the work of Dutordoir et al. [65] restricted its scope to the use of the Arccos kernel in conjunction with the SOFTPLUS activation (pane highlighted in gray in Figure 3.5).

TRUNCATION ERROR. Lastly, as expected, the truncation of the series in Equation (3.6) at some finite number L of spherical harmonic levels often leads to overly smooth predictive response surfaces and overestimation of the variance.

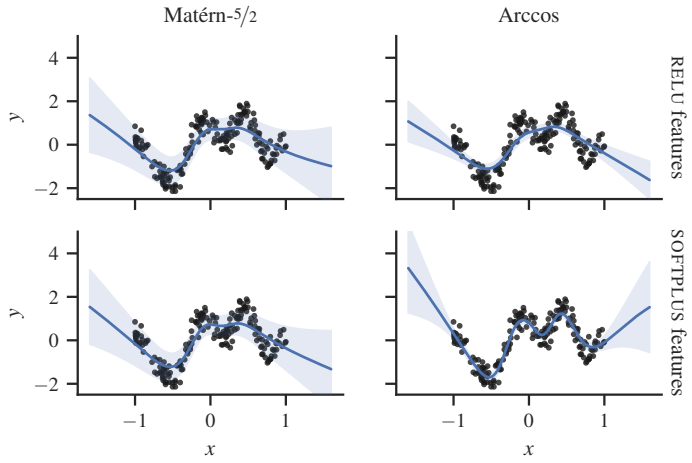


Figure 3.6: Posterior predictives of ACTIVATED SVGP models various kernels and activation features and $L = 16$ levels.

Spherical Features for Orthogonally-Decoupled GPs

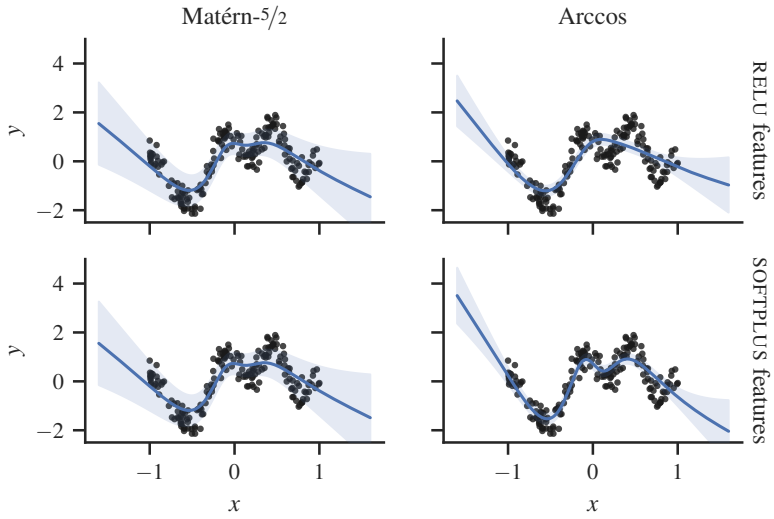
We propose extending the orthogonally-decoupled GP framework (Section 3.3) to use inter-domain inducing features. Accordingly, let $u_m \triangleq \langle f, \phi_m \rangle_{\mathcal{H}}$ and $v_k \triangleq \langle f, \psi_k \rangle_{\mathcal{H}}$ for some arbitrary choices of $\phi_m, \psi_k \in \mathcal{H}$. This generalizes the framework of Shi, Titsias, and Mnih [224] since, by the reproducing property, setting $\phi_m : \mathbf{x} \mapsto k(\mathbf{z}_m, \mathbf{x})$ and $\psi_k : \mathbf{x} \mapsto k(\mathbf{w}_k, \mathbf{x})$ leads to standard inducing points, $u_m = f(\mathbf{z}_m), v_k = f(\mathbf{w}_k)$.

In particular, we define $\phi_m \triangleq H_m$, the m-th unit of the spherical activation layer (Equation (3.5)) described in Section 3.2.2, and $\psi_k(\mathbf{x}) \triangleq k(\mathbf{w}_k, \mathbf{x})$. The posterior predictive of the model described in Section 3.3, summarized by Equations (3.11) and (3.12), is fully determined by the covariances $\mathbf{K_{ff}}, \mathbf{K_{uf}}, \mathbf{K_{vf}}, \mathbf{K_{uu}}, \mathbf{K_{vu}}$ and $\mathbf{K_{vv}}$. Recall that $[\mathbf{K_{uf}}]_{mn} = [\mathbf{k_u}(\mathbf{x}_n)]_m$ and $\mathbf{K_{uu}}$ is precisely as expressed in Equation (3.6). We have

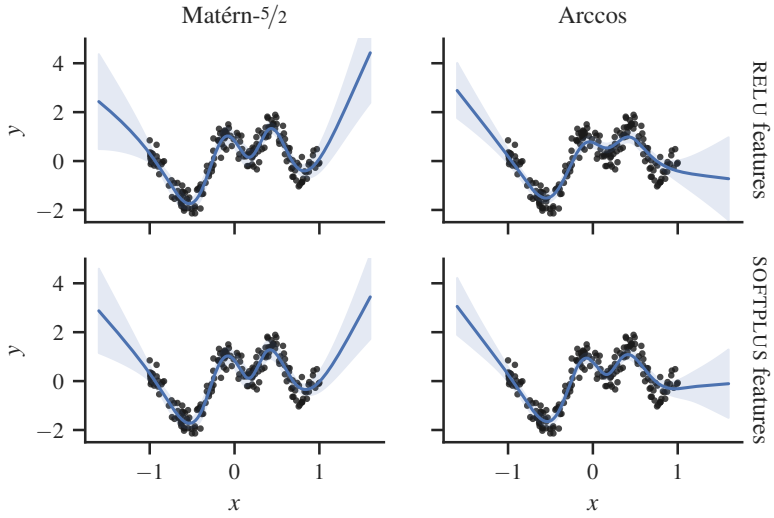
$$
\begin{aligned}
[\mathbf{K_{vf}}]_{kn} &\triangleq \mathrm{Cov}\left(v_k, f(\mathbf{x}_n)\right) = k(\mathbf{w}_k, \mathbf{x}_n), \\
[\mathbf{K_{vu}}]_{km} &\triangleq \mathrm{Cov}\left(v_k, u_m\right) = \phi_m(\mathbf{w}_k), \\
[\mathbf{K_{vv}}]_{kk'} &\triangleq \mathrm{Cov}\left(v_k, v_{k'}\right) = k(\mathbf{w}_k, \mathbf{w}_{k'}).
\end{aligned}
$$

Note that the cross-covariance $\mathbf{K_{vu}}$ between $\mathbf{u}$ and $\mathbf{v}$ can be interpreted as the forward-pass of the orthogonal pseudo-input $\mathbf{w}_k$ through the NN activation H_m. Crucially, these terms constitute the orthogonal basis and provide additional degrees of flexibility, through free parameters $\mathbf{W}$, that can compensate for errors remaining from the original basis – in both the predictive mean and variance. Suffice it to say, this is not the only possible choice but is one that possesses a number of appealing properties.

As discussed in Section 3.3, the addition of K inducing variables incurs a cost of $\mathcal{O}(M^3 + K^3)$. More precisely: suppose the exact cost is $C \cdot (M^3 + K^3)$ operations for some constant C wrt M, K. Further, suppose $K = B \cdot M$ for some $B > 0$. Then there are a total of $(B+1) \cdot M$ inducing variables (orthogonal or otherwise) and the cost becomes $(B^3 + 1)C \cdot M^3$. By comparison, incorporating the same number of inducing variables in SVGP costs $(B+1)^3 C \cdot M^3$. That is, this approach leads to a $(B^3 + 1)$-fold increase in the constant rather than a $(B+1)^3$-fold increase. Concretely, this means that doubling the number of inducing variables doubles the constant in this approach, but leads to an *eight-fold* increase in SVGP. While such a difference vanishes asymptotically for large M and K, it still has a considerable impact for modest sizes ($M, K < 1,000$) that are feasible in practice. Thus seen, incorporating an orthogonal basis spanned by K inducing variables is a more cost-effective strategy for improving ACTIVATED SVGP than increasing M or the truncation level L.

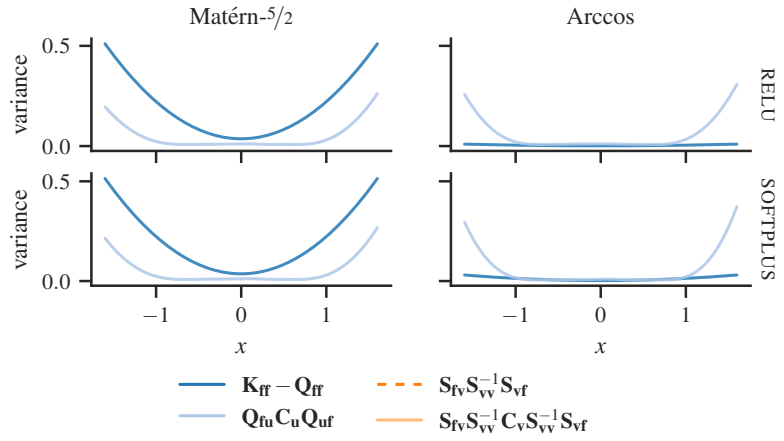


(a) Inducing activation features with $L = 8$ levels.

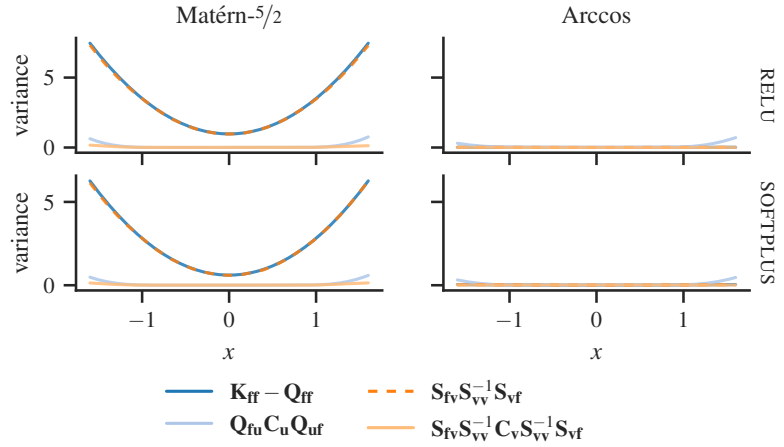


(b) Inducing activation features with $L = 8$ levels and $K = 8$ orthogonal bases (*our method*).

Figure 3.7: Posterior predictives of ACTIVATED SVGP with various kernels and activation features on the 1D Snelson dataset; *black circular markers* represent the observations; *blue solid lines* and *shaded regions* denote the mean and the ± 2 standard deviations, resp.



(a) Inducing activation features with $L = 8$ levels.



(b) Inducing activation features with $L = 8$ levels and $K = 8$ orthogonal bases (*our method*).

Figure 3.8: Decomposition of the posterior predictive variances of SVGP with various kernels and activation features on the 1D SNELSON dataset (see Figure 3.7) into its constituent terms; the additive terms that constitute the predictive variance are indicated by *solid* lines, while the subtractive terms are indicated by *dashed* lines; terms that constitute the predictive variance of the original SVGP model [256] have a *blue* hue, while additional terms introduced by the orthogonally-decoupled model [224] have an *orange* hue.

3.5 EXPERIMENTS

We describe the experiments conducted to empirically validate our approach. The open-source implementation of our method can be found on GitHub at: ltiao/spherical-orthogonal-gaussian-processes. Further information concerning the experimental set-up and various implementation details can be found in Section 3.A.

3.5.1 *Synthetic 1D Dataset*

We highlight some notable properties of our method on the one-dimensional dataset of Snelson and Ghahramani [229].

First we fit ACTIVATED SVGP models with different combinations of kernels and activation features using $L = 8$ truncation levels. The resulting posterior predictives are shown in Figure 3.7. More specifically, in Figure 3.7a, we see that none of the model fits are particularly tight due in part to truncation errors, since we are using relatively few levels. This is especially true of the Matérn kernel (left column of panes), which results in a posterior that is not only too smooth but also clearly suffering from an overestimation of the variance. A conceptually straightforward way to improve performance is to increase the truncation level. Accordingly, Figure 3.6 (introduced earlier in Section 5.3) showed results from effectively the exact same set-up, but with twice the number of levels ($L = 16$). With this increase, we see a clear improvement in the Arccos-SOFTPLUS case, but no discernible difference in the other combinations. Notably, the overestimation of the variances in the Matérn kernel persists. By comparison, Figure 3.7b shows results from using $L = 8$ truncation levels, but with the addition of $K = 8$ orthogonal inducing variables. Remarkably, incorporating just a handful of these variables produces substantial improvements, not least for the Matérn kernel.

Figure 3.8 offers a deeper insight into the underlying mechanisms that contribute to these improvements. Here we plot the predictive variance (Equation (3.12)) in terms of its constituent parts. In Figure 3.8a, we see that the variance estimate with Matérn kernels is heavily distorted by large spurious contributions in the $\mathbf{K_{ff}} - \mathbf{Q_{ff}}$ term (*dark blue solid line*), which is caused by the pathology described in Section 5.3. On the other hand, in Figure 3.8b, such spurious contributions also appear, but are offset by the subtractive term $\mathbf{S_{fv}}\mathbf{S_{vv}}^{-1}\mathbf{S_{vf}}$ (*dark orange dashed line*). This term constitutes the orthogonal basis, and provides added flexibility that is effective at nullifying errors introduced by the original basis.

Each of the three variations discussed above are repeated 5 times, and some quantitative results are summarized in Figure 3.9. Specifically, we report the ELBO and the *throughput*, i.e. the average number of optimisation iterations completed per second. The ACTIVATED SVGP

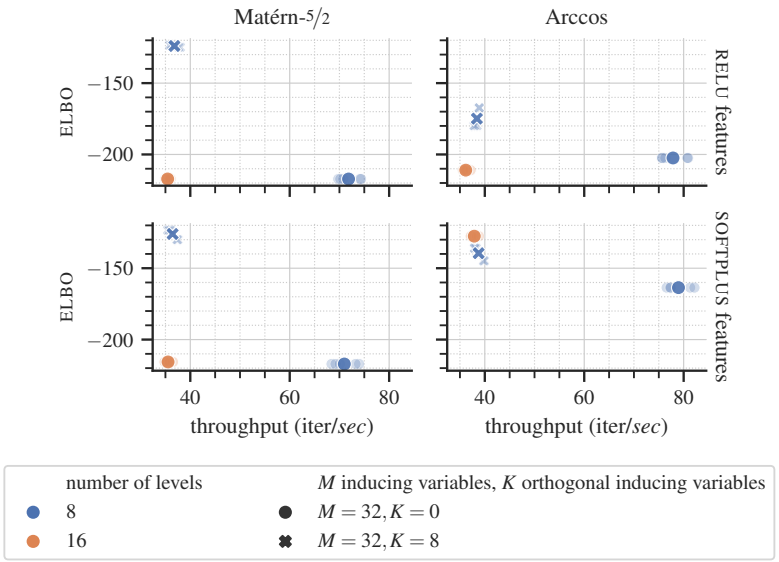


Figure 3.9: The ELBO and throughput (of model fitting) for various kernels and activation features and the configurations visualized in Figure 3.6, Figures 3.7a and 3.8a, and Figures 3.7b and 3.8b; markers with low opacity represent the individual runs, while markers with high opacity represent the mean of each group.

with $L = 8$ truncation levels, as seen in Figures 3.7a and 3.8a, is represented by the *blue circular markers*. The model resulting from doubling the number of levels $L = 16$, as seen in Figure 3.6, is represented by the *orange circular markers*. As discussed, this leads to an improvement in the Arccos-SOFTPLUS case, but to modest or no improvements otherwise. However, we can now see that this has come at a significant computational expense, as the throughput has reduced by roughly half. On the other hand, the model resulting from retaining the same truncation level but incorporate an orthogonal basis consisting of $K = 8$ variables, as seen in Figures 3.7b and 3.8b, is represnted by the *blue cross markers*. This can be seen to have roughly the same footprint as doubling the truncation level, but leads to a considerably improved model fit, especially in cases involving the Matérn kernel (the only exception is in the Arccos-SOFTPLUS case, where doubling the truncation level retains a slight advantage). All told, incorporating an orthogonal basis has roughly the same cost as doubling the truncation level but leads to significantly better performance improvements.

3.5.2 *Regression on UCI Repository Datasets*

We evaluate our method on a number of well-studied regression problems from the UCI repository of datasets [60]. In particular, we consider the YACHT, CONCRETE, ENERGY, KIN8NM and POWER datasets.

Additional results on the larger datasets from this collection can be found in Section 3.B.2.

We fit variations of SVGP with the Arccos, Matérn, and SE kernels, and (a) standard inducing points, and inter-domain inducing features based on (b) RELU- and (c) SOFTPLUS-activated inducing features. For each of these variants, we consider three combinations of base and orthogonal inducing variables: (i-ii) 128 and 256 base inducing variables (and no orthogonal inducing variables), and (iii) 128 base inducing variables with 128 orthogonal inducing variables. The activation features are truncated at $L = 6$ levels. Our proposed method is represented by the combinations consisting of RELU- and SOFTPLUS-activated features with orthogonal inducing variables (b-c,iii). The remaining combinations, against which we benchmark, correspond to the original SVGP (a,i-ii) [256], SOLVEGP (a,iii) [224], and ACTIVATED SVGP (b-c,i-ii) [65].

To quantitatively assess performance, we report the test root-mean-square error (RMSE) and negative log predictive density (NLPD), shown in Figures 3.10 and 3.11, respectively. Unless otherwise stated, for each method and problem, we perform random sub-sampling validation by aggregating results from 5 repetitions across 10% held-out test sets. Within the training set, the inputs and outputs are standardized, i. e. scaled to have zero mean and unit variance and subsequently restored to the original scale at test time.

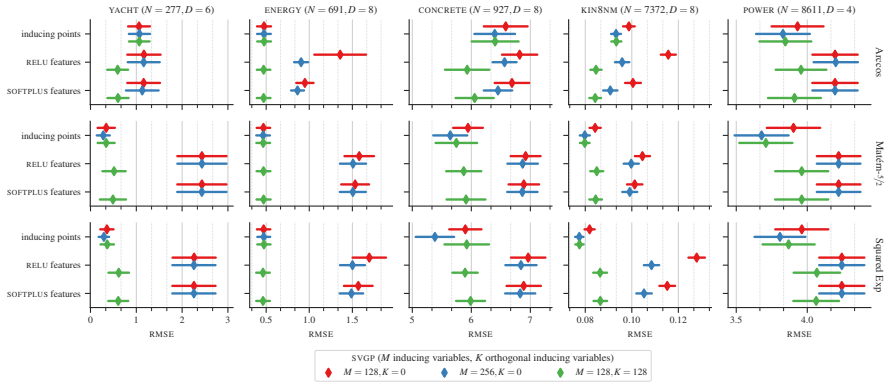


Figure 3.10: Test RMSE on regression problems from the UCI repository of datasets for various kernels and activation features. Along the rows labeled *"inducing points"*, the red and blue markers (♦, ♦) represent the original SVGP model [256], while the green markers (♦) represent SOLVEGP [224]. Along the remaining rows, the red and blue markers (♦, ♦) represent the ACTIVATED SVGP [65], while the green markers (♦) represent our proposed approach.

We observe that, irrespective of the choice of kernel, when using activation features, whether RELU- or SOFTPLUS-activated, augmenting the model with orthogonal bases significantly improves performance, notably even more so than doubling the number of base inducing variables. This can readily be seen across all datasets on both the NLPD

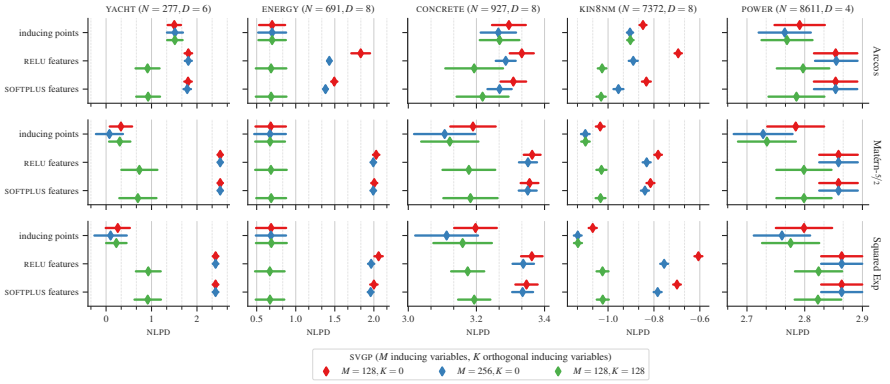


Figure 3.11: Test NLPD on regression problems from the UCI repository of datasets for various kernels and activation features. Along the rows labeled *"inducing points"*, the red and blue markers (♦, ♦) represent the original SVGP model [256], while the green markers (♦) represent SOLVEGP [224]. Along the remaining rows, the red and blue markers (♦, ♦) represent the ACTIVATED SVGP [65], while the green markers (♦) represent our proposed approach.

and RMSE metrics. Further, with the Arccos kernel, it outperforms its counterparts based on standard inducing points across most datasets (the exception being the POWER dataset). With the Matérn and SE kernels, it achieves results comparable to its standard inducing points counterparts in most datasets.

3.5.3 *Large-scale Regression on Airline Delays Dataset*

Finally, we consider a large-scale regression dataset concerning U.S. commercial airline delays in 2008. The task is to forecast the duration of delays in reaching the destination of a given flight, utilizing information such as the route distance, airtime, scheduled month, day of the week, and other relevant factors, as well as characteristics of the aircraft such as its age (number of years since deployment). The complete dataset encompasses 5,929,413 flights, of which we randomly select 1M observations without replacement to form a subset that is more manageable but still considerable in scale. Results on a reduced 100K subset can be found in Section 3.B.1.

To quantitatively assess performance, we report the test RMSE and NLPD evaluated on a 1/3 held-out test set. The results are shown in the top and bottom rows of Figure 3.12, respectively. Within the training set, the inputs and outputs are standardized, i. e. scaled to have zero mean and unit variance and subsequently restored to the original scale at test time.

Given the immense volume of data at hand, we are compelled to utilize mini-batch training for stochastic optimisation [95]. To this end, we use the Adam optimizer [123] with its typical default settings

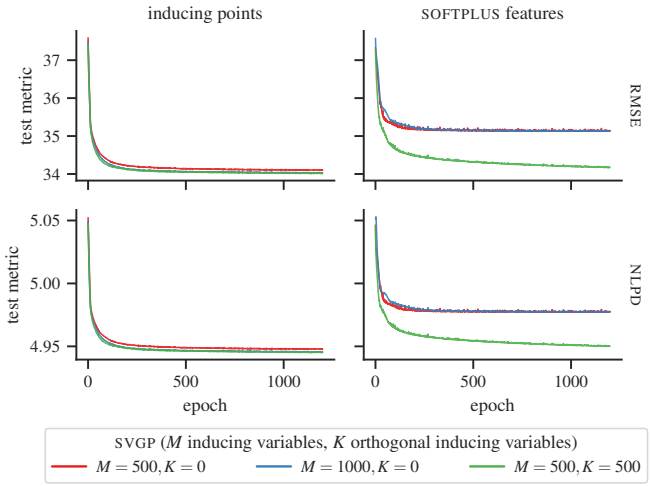


Figure 3.12: Test metrics, RMSE and NLPD, on the large-scale 2008 U.S. airline delays dataset using the *Arccos* kernel with standard inducing points and SOFTPLUS-activated features. Along the column labeled "inducing points", the red and blue lines (— and —) represent the mini-batch SVGP [95], while the green line (—) represents SOLVEGP [224]. Along the column labeled "SOFTPLUS features", the red and blue lines (— and —) represent the ACTIVATED SVGP [65], while the green line (—) represents our proposed approach.

(learning rate 1×10^{-3}, $\beta_1 = 0.9, \beta_2 = 0.999$). Our batch size is set to 5,000, and we train the models for a total of 1,200 epochs.

We fit variations of SVGP with the Arccos kernel and (a) standard inducing points and (b) inter-domain inducing features based on SOFTPLUS-activated inducing features. For each of these variants, we consider three combinations of base and orthogonal inducing variables: (i-ii) 500 and 1,000 base inducing variables (and no orthogonal inducing variables), and (iii) 500 base inducing variables with 500 orthogonal inducing variables. The activation features are truncated at $L = 6$ levels. Our proposed method is represented by the combination consisting of SOFTPLUS-activated features with orthogonal inducing variables (b,iii). The remaining combinations, against which we benchmark, correspond to the mini-batch SVGP (a,i-ii) [95], SOLVEGP (a,iii) [224], and ACTIVATED SVGP (b,i-ii) [65].

The outcomes are as expected when employing standard inducing points (left). In particular, doubling the number of base inducing points from 500 to 1,000 demonstrates significant improvements. Furthermore, by using 500 base inducing points alongside 500 orthogonal inducing points, we achieve comparable performance to having 1,000 base inducing points, while enjoying improved computationally efficiency. In contrast, when examinining the ACTIVATED SVGP model with SOFTPLUS features (right), it's apparent that it underperforms compared to the original SVGP counterparts. Furthermore, doubling the number of inducing features from 500 to 1,000 has virtually no effect.

However, by incorporating orthogonal bases into the ACTIVATED SVGP model with 500 features following our proposed approach, we witness substantial improvements and achieve comparable performance to its standard inducing points counterparts.

3.6 SUMMARY

We considered the use of inter-domain inducing features in the orthogonally-decoupled SVGP framework, specifically, the spherical activation features, and showed that this alleviates some of the practical issues and shortcomings associated with the ACTIVATED SVGP model. We demonstrated the effectiveness of this approach by conducting empirical evaluations on several problems, and showed that this leads to enhanced predictive performance over more computationally demanding alternatives such as increasing the truncation levels or the number of inducing variables.

Future work will explore alternative designs of inter-domain inducing features to construct new standard and orthogonal bases that provide additional complementary benefits.

ADDENDUM

3.A EXPERIMENTAL SET-UP AND IMPLEMENTATION DETAILS

3.A.1 *Hardware*

All experiments were carried out on a consumer-grade laptop computer with an Intel Core™ i7-11800H (8 Cores) @ 4.6GHz Processor, 16GB Memory, and a NVIDIA GeForce RTX™ 3070 Laptop (Mobile/Max-Q) Graphics Card.

3.A.2 *Software*

Our method is implemented by extending functionality from the GPFlow software library [163]. The code will be released as open-source software upon publication. Additional software dependencies upon which our implementation relies, either directly or indirectly, are enumerated in Table 3.A.1.

Table 3.A.1: Key software dependencies.

Method	Software Library	URL (`github.com/*`)
SVGP [256]	GPFlow	`GPflow/GPflow`
ODVGP [217]	-	`hughsalimbeni/orth_decoupled_var_gps`
SOLVEGP [224]	-	`thjashin/solvegp`
VISH [64]	Spherical Harmonics	`vdutor/SphericalHarmonics`
ACTIVATED SVGP [65]	-	`vdutor/ActivatedDeepGPs`
-	Bayesian Benchmarks	`hughsalimbeni/bayesian_benchmarks`

3.A.3 *Hyperparameters*

We adopt sensible defaults across all problems and datasets; no hand-tuning is applied to any specific one. The choices of the hyperparameters and other relevant dependencies are summarized as follows:

OPTIMISATION. We use the L-BFGS optimizer [26, 293] with the default settings from `scipy.optimize` [269].

LIKELIHOOD. The Gaussian likelihood variance is initialized to 1.0 across all experiments.

KERNEL PARAMETER INITIALISATION. All stationary kernels are initialized with unit lengthscale and amplitude.

VARIATIONAL PARAMETER INITIALISATION. The variational distributions $q(\mathbf{u})$, $q(\mathbf{v}')$ are initialized with zero mean and identity covariance $\mathbf{m} = \mathbf{0}, \mathbf{C} = \mathbf{I}$.

WHITENED PARAMETERISATION. We do *not* use the whitened parameterisation (as used, for example, by Hensman et al. [97] and Murray and Adams [177]) in either $q(\mathbf{u})$ or $q(\mathbf{v}')$.

INDUCING POINT INITIALISATION. We make our best effort to ensure a fair comparison against baselines involving standard inducing points. To this end, we adopt the best practice of first optimizing the variational parameters, not least the inducing input locations $\mathbf{Z}$ (and $\mathbf{W}$ where applicable), before jointly optimizing all of the free parameters. This initialisation phase is done for up to 100 iterations of the L-BFGS algorithm.

3.B ADDITIONAL RESULTS

3.B.1 *Regression on Airline Delays Dataset*

We repeat the experiment outlined in Section 3.5.3, focusing on a reduced subset of the 2008 U.S. airline delays dataset that consists of 100K randomly selected observations. Unlike the previous experimental set-up, the parameters are optimised for a total of 1,000 epochs. Additionally, we report aggregated results from 5 repetitions across $1/3$ held-out test sets. The results are shown in Figure 3.B.1.

3.B.2 *Extra UCI Repository Datasets*

Results on a few larger regression datasets from the UCI repository can be found in Figure 3.B.2. In this analysis, we adopted the same combination of activation features and sparse GP models as described in Section 3.5.2. However, in contrast to Section 3.5.2, we restrict our focus to the Arccos kernel.

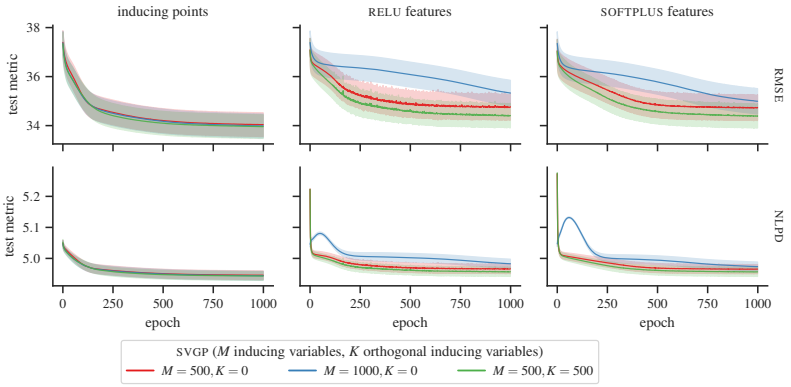


Figure 3.B.1: Test metrics, RMSE and NLPD, aggregated across 5 random subsampling test splits on a 100K subset of the 2008 U.S. airline delays dataset. Results are shown for models using the *Arccos* kernel with standard inducing points and various activation features. Along the column labeled "inducing points", the red and blue lines (— and —) represent the mini-batch SVGP [95], while the green line (—) represents SOLVEGP [224]. Along the column labeled "SOFTPLUS features", the red and blue lines (— and —) represent the ACTIVATED SVGP [65], while the green line (—) represents our proposed approach.

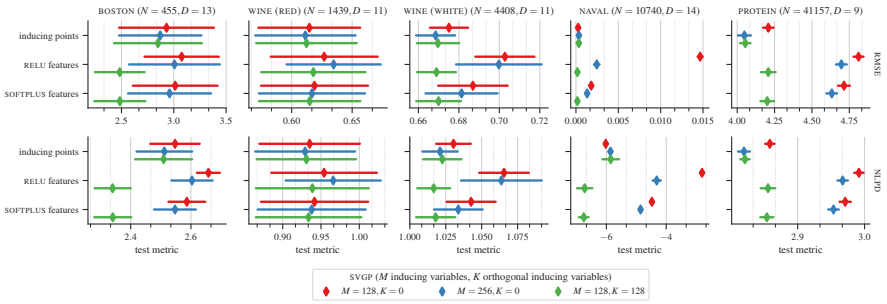


Figure 3.B.2: Test metrics, RMSE and NLPD, on an extra set of larger regression problems from the UCI dataset repository for the *Arccos* kernel and various activation features. Along the rows labeled "*inducing points*", the red and blue markers (♦, ♦) represent the original SVGP model [256], while the green markers (♦) represent SOLVEGP [224]. Along the remaining rows, the red and blue markers (♦, ♦) represent the ACTIVATED SVGP [65], while the green markers (♦) represent our proposed approach.