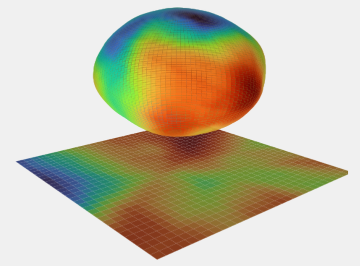


# Spherical Inducing Features for Orthogonally-Decoupled Gaussian Processes


Louis Tiao [1] Vincent Dutordoir [3,2] Victor Picheny [2]

[1]University of Sydney [2]Secondmind [3]University of Cambridge


Secondmind

## Gaussian Processes: The Story So Far

Gaussian processes (GPS) are renowned for their exceptional data efficiency, reliable uncertainty estimation, flexibility, and built-in mechanisms to mitigate against overfitting. However, they are often unfavorably compared to deep learning approaches due to limited scalability and their inability to capture hierarchies of abstract representations.

Sparse variational GPS (SVGPS) [5] address scalability by introducing auxiliary *inducing variables* $\mathbf{u} \triangleq f(\mathbf{Z}) \in \mathbb{R}^M$ at pseudo-inputs $\mathbf{Z} = [\mathbf{z}_1 \cdots \mathbf{z}_M]^\top$. Approximate the posterior $p(\mathbf{f}, \mathbf{u} \mid \mathbf{y})$ with $q^\star(\mathbf{f}, \mathbf{u}) = \arg\min_q \mathrm{KL}\left[q(\mathbf{f}, \mathbf{u}) \parallel p(\mathbf{f}, \mathbf{u} \mid \mathbf{y})\right]$ where $q(\mathbf{f}, \mathbf{u}) \triangleq p(\mathbf{f} \mid \mathbf{u}) q(\mathbf{u})$ and $q(\mathbf{u}) \triangleq \mathcal{N}(\mathbf{m_u}, \mathbf{C_u})$.

Leads to predictive density:

$$q(f(\mathbf{x})) = \mathcal{GP}\left(\mathbf{k}_\mathbf{u}^\top(\mathbf{x})\mathbf{K}_\mathbf{uu}^{-1}\mathbf{m}_\mathbf{u}, k(\mathbf{x}, \mathbf{x}') - \mathbf{k}_\mathbf{u}^\top(\mathbf{x})\mathbf{K}_\mathbf{uu}^{-1}(\mathbf{K}_\mathbf{uu} - \mathbf{C}_\mathbf{u})\mathbf{K}_\mathbf{uu}^{-1}\mathbf{k}_\mathbf{u}(\mathbf{x}')\right) \tag{1}$$

where $[\mathbf{K_{uu}}]_{mm'} \triangleq \mathrm{Cov}(u_m, u_{m'})$.

- *Reduces cost* from $\mathcal{O}(N^3)$ to $\mathcal{O}(M^3)$ (assuming $M \ll N$)
- Unlocks *greater flexibility* in model specification

Basis functions are effectively $\mathbf{k_u} : \mathcal{X} \to \mathbb{R}^M$ where each element $[\mathbf{k_u}(\mathbf{x})]_m \triangleq \mathrm{Cov}(f(\mathbf{x}), u_m)$

1. Standard Inducing Points are values of $f$ evaluated at pseudo-inputs
$$u_m \triangleq f(\mathbf{z}_m) \quad \Rightarrow \quad [\mathbf{K_{uu}}]_{mm'} = k(\mathbf{z}_m, \mathbf{z}_{m'}) \quad \text{and} \quad [\mathbf{k_u}(\mathbf{x})]_m = k(\mathbf{z}_m, \mathbf{x})$$
Mapping is *static*: solely determined by *fixed* kernel $k$ and *local* influence of $\mathbf{z}_m$.
2. Inter-domain Inducing Features [3] are a generalisation involving scalar projection of $f$ onto some $\phi_m$ in the reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ of $k$,
$$u_m \triangleq \langle f, \phi_m \rangle_\mathcal{H} \quad \Rightarrow \quad [\mathbf{K_{uu}}]_{mm'} = \langle \phi_m, \phi_{m'} \rangle_\mathcal{H} \quad \text{and} \quad [\mathbf{k_u}(\mathbf{x})]_m = \phi_m(\mathbf{x})$$
Mapping is *adaptive*: can result in *sparser representations* that lead to *greater scalability*.

## Spherical Inducing Features

Spherical harmonics, an extension of the Fourier basis to multiple dimensions, can be used to form $\phi_m$ [1]. Orthogonality leads to *diagonal* covariance:

$$[\mathbf{K_{uu}}]_{mm'} = \lambda_m^{-1}\delta_{mm'}$$

- Reduces cost from $\mathcal{O}(M^3)$ to $\mathcal{O}(M)$ (!)

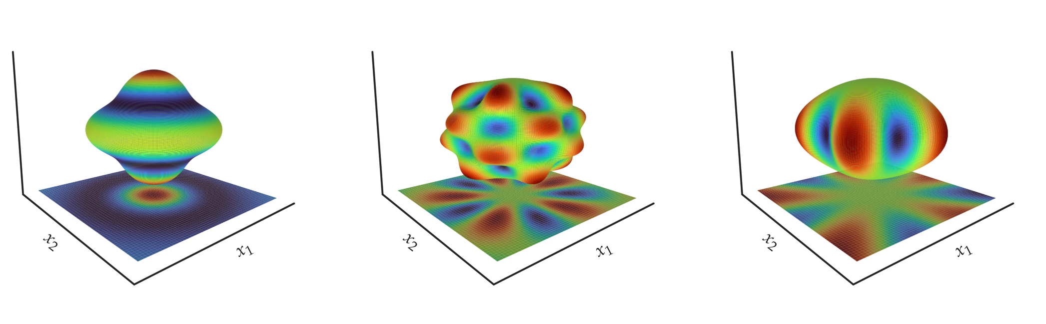


Figure 1: A few example *spherical harmonics* in 3D.

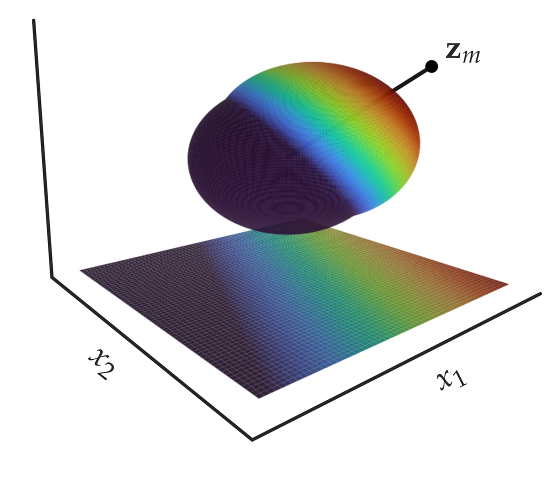


Figure 2: A RELU-activated hidden unit on the sphere in 3D.

Spherical neural network (NN) activations. To make $\mathbf{k_u}(\mathbf{x})$ resemble a hidden layer in a feedfoward NN [2], define $\phi_m$ as the $m$th hidden unit with nonlinear activation $\sigma$,

$$\phi_m(\mathbf{x}) \triangleq \|\mathbf{z}_m\|\|\mathbf{x}\| \cdot \sigma\left(\frac{\mathbf{z}_m^\top \mathbf{x}}{\|\mathbf{z}_m\|\|\mathbf{x}\|}\right)$$

- Predictive mean becomes a single-layer feedforward NN:
$$\mathbf{k_u}(\mathbf{x})^\top \mathbf{K}_\mathbf{uu}^{-1}\mathbf{m_u} = \sum_{m=1}^{M} \beta_m \phi_m(\mathbf{x}), \qquad \boldsymbol{\beta} \triangleq \mathbf{K}_\mathbf{uu}^{-1}\mathbf{m_u} \in \mathbb{R}^M$$
- When stacked to form a deep GP (DGP), the propagation of predictive means emuates forward pass through a deep NN (DNN)
- Obtain predictive uncertainty in DNNS for free as a byproduct (!)

## Orthogonally-Decoupled Gaussian Processes

1. Decouple GP as sum of two independent GPS [4]:
$$f(\mathbf{x}) \sim \mathcal{GP}\left(0, k(\mathbf{x}, \mathbf{x}')\right) \quad \Leftrightarrow \quad f(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x}),$$
where
$$g(\mathbf{x}) \sim \mathcal{GP}\left(0, \mathbf{k}_\mathbf{u}^\top(\mathbf{x})\mathbf{K}_\mathbf{uu}^{-1}\mathbf{k}_\mathbf{u}(\mathbf{x}')\right),$$
$$h(\mathbf{x}) \sim \mathcal{GP}\left(0, s(\mathbf{x}, \mathbf{x}')\right)$$
for $s(\mathbf{x}, \mathbf{x}') \triangleq k(\mathbf{x}, \mathbf{x}') - \mathbf{k}_\mathbf{u}^\top(\mathbf{x})\mathbf{K}_\mathbf{uu}^{-1}\mathbf{k}_\mathbf{u}(\mathbf{x}')$.
2. Introduce *orthogonal* inducing variables $\mathbf{v} \triangleq f(\mathbf{W}), \mathbf{v}' \triangleq h(\mathbf{W}) \in \mathbb{R}^K$ at pseudo-inputs $\mathbf{W} \triangleq [\mathbf{w}_1 \cdots \mathbf{w}_K]^\top$.
3. Approximate posterior $q(\mathbf{v}') \triangleq \mathcal{N}(\mathbf{m_v}, \mathbf{C_v})$.

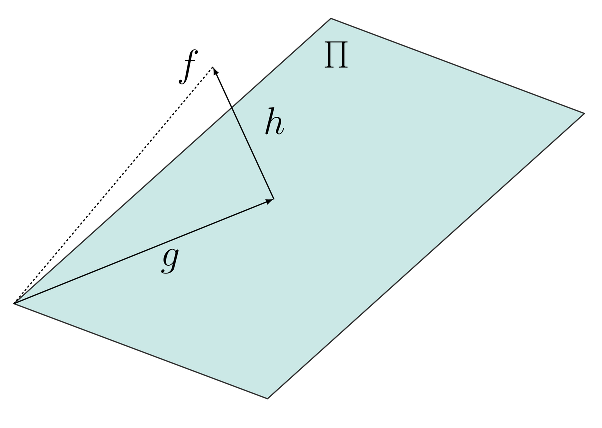




Figure 3: Hyperplane $\Pi \triangleq \{\boldsymbol{\alpha}^\top \mathbf{k_u}(\cdot); \boldsymbol{\alpha} \in \mathbb{R}^M\}$

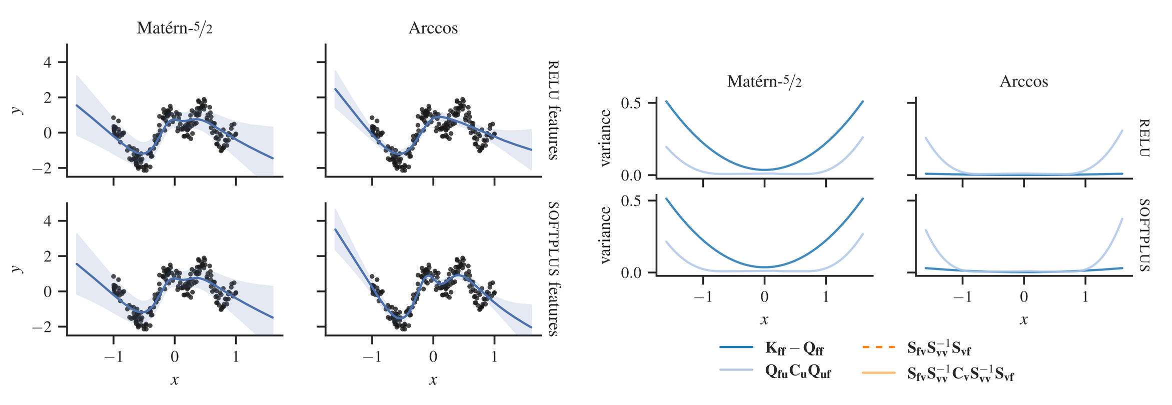
Leads to predictive density:

$$q(\mathbf{f}_*) = \mathcal{N}(\mathbf{K}_{*\mathbf{u}}\mathbf{K}_\mathbf{uu}^{-1}\mathbf{m_u} + \underbrace{\mathbf{S}_{*\mathbf{v}}\mathbf{S}_\mathbf{vv}^{-1}}_{\text{orthogonal bases}}\mathbf{m_v}, \mathbf{K}_{**} - \mathbf{K}_{*\mathbf{u}}\mathbf{K}_\mathbf{uu}^{-1}(\mathbf{K_{uu}} - \mathbf{C_u})\mathbf{K}_\mathbf{uu}^{-1}\mathbf{K}_{\mathbf{u}*} - \underbrace{\mathbf{S}_{*\mathbf{v}}\mathbf{S}_\mathbf{vv}^{-1}(\mathbf{S_{vv}} - \mathbf{C_v})\mathbf{S}_\mathbf{vv}^{-1}\mathbf{S}_{\mathbf{v}*}}_{\text{orthogonal bases}}) \tag{2}$$

## Technical Issues with Spherical NN Activation Features

In practice, several widely-used kernels and activation functions are incompatible:

- Spectra mismatch. For the Matérn kernel, discrepancies in its Fourier coefficients (nonzero) with those of the activation features (zero) lead to overestimation of the predictive variance.
- RKHS inner product. For the RELU activation features, its (squared) Fourier coefficients decay at the same rate as those of numerous kernels, resulting in an indeterminate RKHS inner product.



(a) Predictive densities. (b) Predictive variances, deconstructed.

Figure 4: Posterior of SVGPS with various kernels and activations; at $L = 8$ levels.

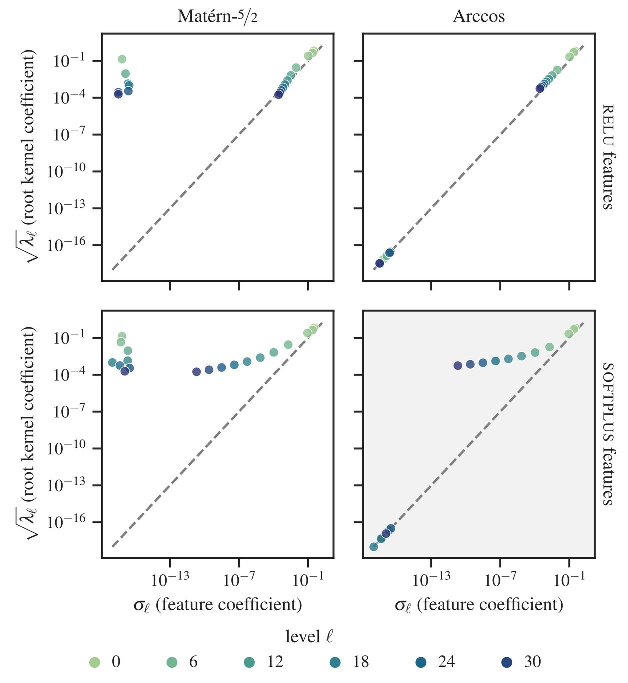


Figure 5: Comparison of the Fourier coefficients of various kernels and activation features with increasing levels.

## Our Solution

Extend the orthogonally-decoupled GP framework with inter-domain inducing features: let

$$u_m \triangleq \langle f, \phi_m \rangle_\mathcal{H}, \quad \text{and} \quad v_k \triangleq \langle f, \psi_k \rangle_\mathcal{H}$$

for some choices of $\phi_m, \psi_k \in \mathcal{H}$.
To obtain $\mathbf{S_{vv}}, \mathbf{S_{vf}}$ in eq. 2, need to compute prior covariances $\mathbf{K_{vf}}, \mathbf{K_{vu}}, \mathbf{K_{vv}}$

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{u} \\ \mathbf{v} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K_{ff}} & \mathbf{K_{uf}}^\top & \mathbf{K_{vf}}^\top \\ \mathbf{K_{uf}} & \mathbf{K_{uu}} & \mathbf{K_{vu}}^\top \\ \mathbf{K_{vf}} & \mathbf{K_{vu}} & \mathbf{K_{vv}} \end{bmatrix}\right)$$

In this work:

- $\phi_m$: $m$th unit of the spherical activation layer
- $\psi_k(\mathbf{x}) \triangleq k(\mathbf{w}_k, \mathbf{x})$

Leads to covariances:

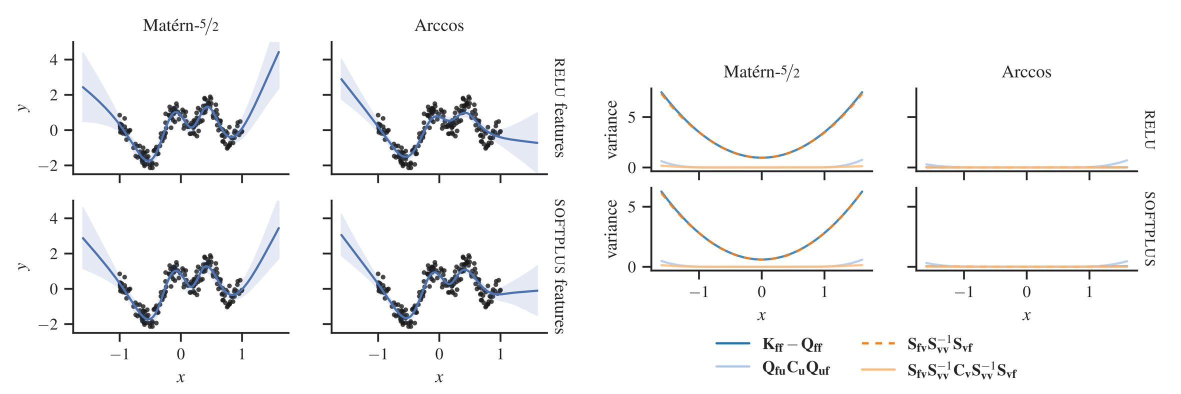
$$[\mathbf{K_{vf}}]_{kn} \triangleq \mathrm{Cov}(v_k, f(\mathbf{x}_n)) = k(\mathbf{w}_k, \mathbf{x}_n),$$
$$[\mathbf{K_{vu}}]_{km} \triangleq \mathrm{Cov}(v_k, u_m) = \phi_m(\mathbf{w}_k),$$
$$[\mathbf{K_{vv}}]_{kk'} \triangleq \mathrm{Cov}(v_k, v_{k'}) = k(\mathbf{w}_k, \mathbf{w}_{k'}).$$

Cross-covariance $\mathbf{K_{vu}}$ consists of forward-pass of pseudo-inputs $\mathbf{w}_k$ through neurons $\phi_m$

## Experimental Results

### Regression on Synthetic 1D Dataset

Incorporating a small handful of $K = 8$ orthogonal inducing variables costs roughly the same as doubling the truncation level $L$ but leads to substantial improvements.



(a) Predictive densities. (b) Predictive variances, deconstructed.

Figure 6: Posterior of SVGPS with various kernels and activations, $K = 8$ orthogonal bases; at $L = 8$ levels. New term $\mathbf{S_{fv}}\mathbf{S}_\mathbf{vv}^{-1}\mathbf{S_{vf}}$ offsets errors from the original basis.

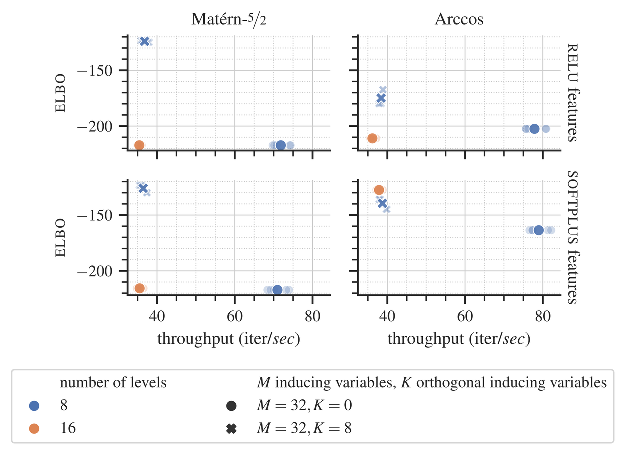


Figure 7: Evidence lower bound (ELBO) and training throughput for the various kernels and activation features visualised in Figures 4 and 6.

### Regression on UCI Repository Datasets

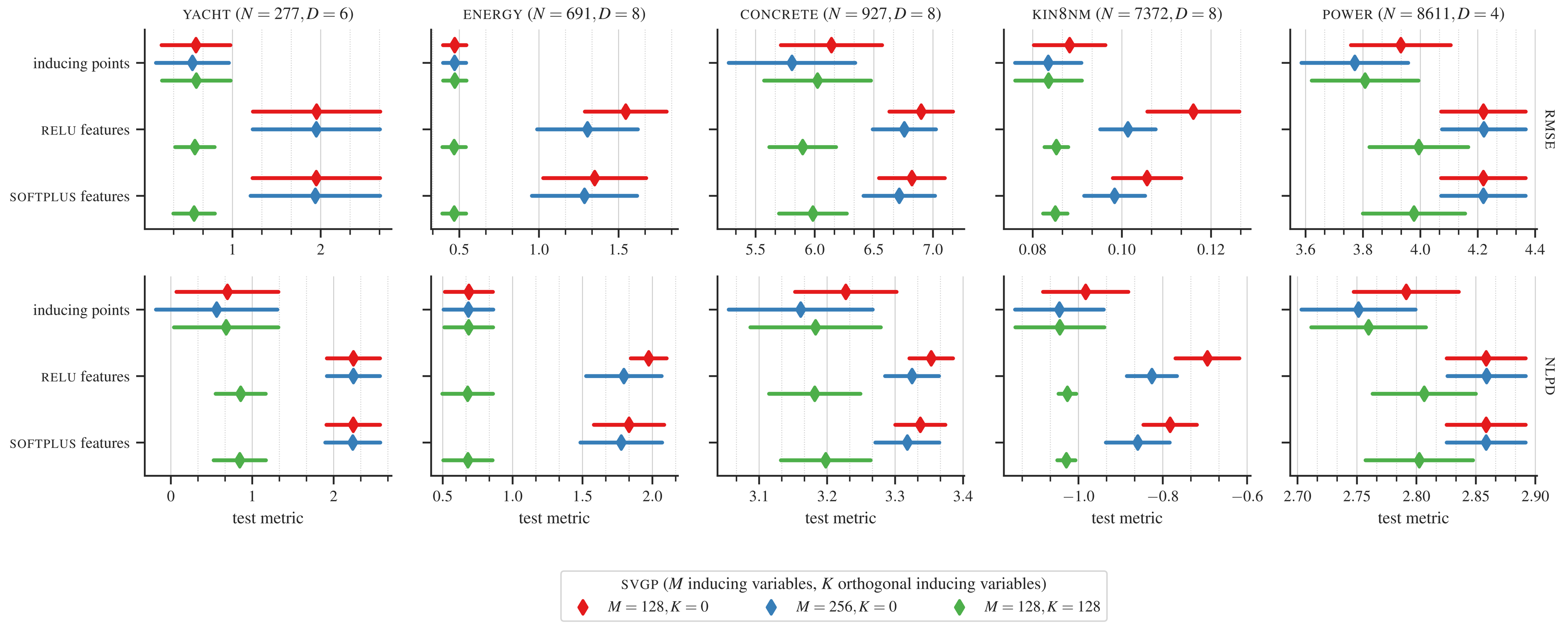


Figure 8: Test metrics, RMSE and NLPD, on the UCI regression datasets using the Arccos kernel with various activation features. Along the rows labeled *"inducing points"*, the red and blue markers (♦, ♦) represent the original SVGP model [5], while the green markers (♦) represent SOLVEGP [4]. Along the remaining rows, the red and blue markers (♦, ♦) represent the activated SVGP [2], while the green markers (♦) represent our proposed approach.

## References


[1] Vincent Dutordoir, Nicolas Durrande, and James Hensman. Sparse Gaussian processes with spherical harmonic features. In *International Conference on Machine Learning*, pages 2793–2802. PMLR, 2020.

[2] Vincent Dutordoir, James Hensman, Mark van der Wilk, Carl Henrik Ek, Zoubin Ghahramani, and Nicolas Durrande. Deep neural networks as point estimates for deep Gaussian processes. *Advances in Neural Information Processing Systems*, 34, 2021.

[3] Miguel Lázaro-Gredilla and Anibal Figueiras-Vidal. Inter-domain Gaussian processes for sparse inference using inducing features. *Advances in Neural Information Processing Systems*, 22, 2009.

[4] Jiaxin Shi, Michalis Titsias, and Andriy Mnih. Sparse orthogonal variational inference for Gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, pages 1932–1942. PMLR, 2020.

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