Sparse Gaussian Processes with Spherical Harmonic Features

Vincent Dutordoir[1], Nicolas Durrande[1] and James Hensman[2]

[1] PROWLER.io, [2]Amazon (Work completed while JH was at PROWLER.io)

International Conference of Machine Learning – 2020

Contribution

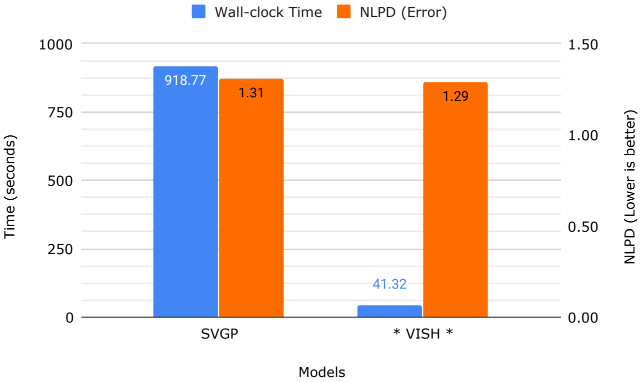
We improve the scaling of Sparse GPs with #datapoints and #inputs

Airline dataset:

- Regression problem
- 6.10^6 datapoints
- 8 input dimensions

Setup

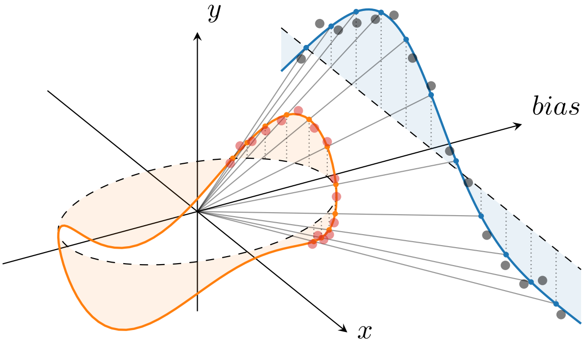
- GTX 1070 GPU



Variational Inference with Spherical Harmonics (VISH)

Gist of method:

- make inputs $d+1$ dimensional
- project data radially on $\mathbb{S}^d$
- Fast SVGP on the sphere
- map predictions on $\mathbb{S}^d$ back to the original space

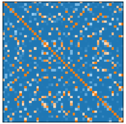


The efficiency of VISH comes from using *spherical harmonics as inducing functions* for the SVGP on the sphere.

From inducing points to inducing features

Inducing Points

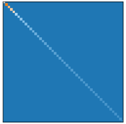
$$u_m = f(z_m)$$

$K_{\mathbf{uu}} =$ 

$K_{\mathbf{uu}}^{-1}$ is $\mathcal{O}(M^3)$

VISH

$$u_m = \langle f, \phi_m \rangle_{\mathcal{H}}$$

$K_{\mathbf{uu}} =$ 

$K_{\mathbf{uu}}^{-1}$ is $\mathcal{O}(M)$

Orthogonality of the basisfunctions ϕ leads to diagonal $K_{\mathbf{uu}}$ and $\mathcal{O}(M)$ inversion

Deep-dive

Sparse Variational Gaussian processes

Scalable and flexible

- Capture the GP by a set of inducing variables $\mathbf{u} = f(Z)$, at locations $\mathbf{z}_1, \ldots, \mathbf{z}_M$.

Sparse Variational Gaussian processes

Scalable and flexible

- Capture the GP by a set of inducing variables $\mathbf{u} = f(Z)$, at locations $\mathbf{z}_1, \dots, \mathbf{z}_M$.
- Minimise KL-divergence from $p(f(\cdot) \,|\, y)$ to $q(f(\cdot)) = \mathcal{GP}(\mu(\cdot), \nu(\cdot, \cdot'))$

$$\begin{cases} \mu(\cdot) = k_{\mathbf{u}}^{\top}(\cdot) K_{\mathbf{uu}}^{-1} \mathbf{m} \\ \nu(\cdot, \cdot') = k(\cdot, \cdot') - k_{\mathbf{u}}^{\top}(\cdot) K_{\mathbf{uu}}^{-1} (K_{\mathbf{uu}} - S) K_{\mathbf{uu}}^{-1} k_{\mathbf{u}}(\cdot') \end{cases},$$

where $[K_{\mathbf{uu}}]_{m,m'} = \mathrm{Cov}(u_m, u_{m'})$ and $[k_{\mathbf{u}}(\cdot)]_m = \mathrm{Cov}(u_m, f(\cdot))$.

Sparse Variational Gaussian processes

Scalable and flexible

- Capture the GP by a set of inducing variables $\mathbf{u} = f(Z)$, at locations $\mathbf{z}_1, \ldots, \mathbf{z}_M$.
- Minimise KL-divergence from $p(f(\cdot) \,|\, y)$ to $q(f(\cdot)) = \mathcal{GP}(\mu(\cdot), \nu(\cdot, \cdot'))$

$$\begin{cases} \mu(\cdot) = k_{\mathbf{u}}^{\top}(\cdot) K_{\mathbf{uu}}^{-1} \mathbf{m} \\ \nu(\cdot, \cdot') = k(\cdot, \cdot') - k_{\mathbf{u}}^{\top}(\cdot) K_{\mathbf{uu}}^{-1} (K_{\mathbf{uu}} - S) K_{\mathbf{uu}}^{-1} k_{\mathbf{u}}(\cdot') \end{cases},$$

  where $[K_{\mathbf{uu}}]_{m,m'} = \mathrm{Cov}(u_m, u_{m'})$ and $[k_{\mathbf{u}}(\cdot)]_m = \mathrm{Cov}(u_m, f(\cdot))$.
- A more flexible (e.g. non-Gaussian likelihoods) and scalable (e.g. mini-batching) model at a cost of $\mathcal{O}(M^3 + M^2 N)$.

Sparse Variational Gaussian processes

Scalable and flexible

- Capture the GP by a set of inducing variables $\mathbf{u} = f(Z)$, at locations $\mathbf{z}_1, \dots, \mathbf{z}_M$.
- Minimise KL-divergence from $p(f(\cdot) \,|\, y)$ to $q(f(\cdot)) = \mathcal{GP}(\mu(\cdot), \nu(\cdot, \cdot'))$

$$\begin{cases} \mu(\cdot) = k_{\mathbf{u}}^\top(\cdot) K_{\mathbf{uu}}^{-1} \mathbf{m} \\ \nu(\cdot, \cdot') = k(\cdot, \cdot') - k_{\mathbf{u}}^\top(\cdot) K_{\mathbf{uu}}^{-1}(K_{\mathbf{uu}} - S) K_{\mathbf{uu}}^{-1} k_{\mathbf{u}}(\cdot') \end{cases},$$

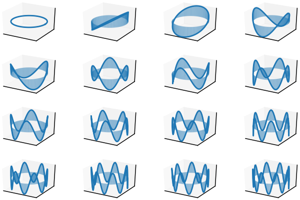
  where $[K_{\mathbf{uu}}]_{m,m'} = \mathrm{Cov}(u_m, u_{m'})$ and $[k_{\mathbf{u}}(\cdot)]_m = \mathrm{Cov}(u_m, f(\cdot))$.
- A more flexible (e.g. non-Gaussian likelihoods) and scalable (e.g. mini-batching) model at a cost of $\mathcal{O}(M^3 + M^2 N)$.
- Speedup through structure in the $K_{\mathbf{uu}}$ matrix (e.g. Hensman et al 2017, VFF).

Outline

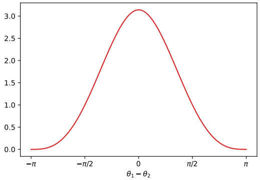
- Gaussian processes on the circle and hypersphere
- Spherical harmonics as inducing features
- Linear projection data on the hyper-sphere

Gaussian processes on the circle

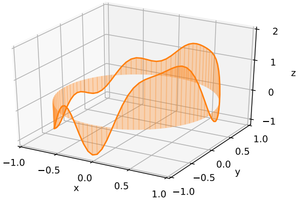
$$\Phi(\theta) = [\cos(i\theta), \sin(i\theta)]_{i=0}^{\infty}$$



$$k(\theta_1, \theta_2) = \sum_{i=0}^{\infty} \lambda_i \phi_i(\theta_1)\phi_i(\theta_2)$$

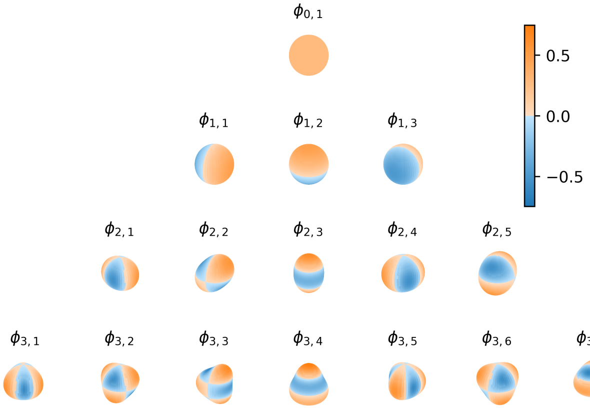


$$f = \sum_i \xi_i \phi_i(\theta), \text{ with } \xi_i \sim \mathcal{N}(0, \lambda_i)$$



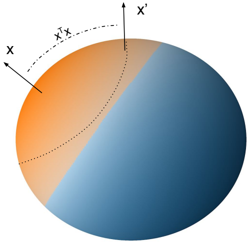
Spherical Harmonics

- Orthonormal basis on the hyper sphere
- Eigenfunctions the Laplace-Beltrami operator $\Delta^{\mathbb{S}^{d-1}} \phi_i = \lambda_i \, \phi_i$
- Eigenfunction of zonal kernels



Mercer's theorem for zonal kernels on the sphere

- Zonal kernels are the spherical counterpart of stationary kernels $k(x, x') = k'(\text{distance}(x, x'))$.

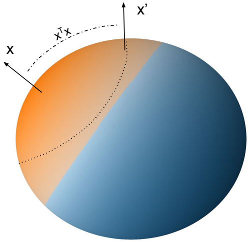


Mercer's theorem for zonal kernels on the sphere

- Zonal kernels are the spherical counterpart of stationary kernels $k(x, x') = k'(\text{distance}(x, x'))$.

- Mercer's decomposition: Any zonal kernel k on the hyper-sphere can be decomposed as

$$k(\mathbf{x}, \mathbf{x}') = \sum_{i=0}^{\infty} \lambda_i \, \phi_i(\mathbf{x}) \, \phi_i(\mathbf{x}').$$

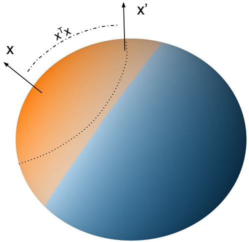


Mercer's theorem for zonal kernels on the sphere

- Zonal kernels are the spherical counterpart of stationary kernels $k(x, x') = k'(\text{distance}(x, x'))$.

- Mercer's decomposition: Any zonal kernel k on the hypersphere can be decomposed as

$$k(\mathbf{x}, \mathbf{x}') = \sum_{i=0}^{\infty} \lambda_i \, \phi_i(\mathbf{x}) \, \phi_i(\mathbf{x}').$$



- Karhunen–Loève expansion: A GP f on the hypersphere with zonal covariance k can be written $f = \sum_i \xi_i \phi_i$ with $\xi_i \sim \mathcal{N}(0, \lambda_i)$:

$f = \xi_0 \cdot$  $+ \xi_1 \cdot$  $+ \xi_2 \cdot$  $+ \xi_3 \cdot$  $+ \xi_4 \cdot$  $\ldots$

Spherical harmonics as inducing features in SVGPs

- Define the kernel's RKHS $\mathcal{H}$ with reproducing inner-product:

$$\langle k(\mathbf{x}, \cdot), h(\cdot) \rangle_{\mathcal{H}} = h(\mathbf{x})$$

Spherical harmonics as inducing features in SVGPs

- Define the kernel's RKHS $\mathcal{H}$ with reproducing inner-product:

$$\langle k(\mathbf{x}, \cdot), h(\cdot) \rangle_{\mathcal{H}} = h(\mathbf{x})$$

- Approximate posterior constructed out of inducing features

$$\boxed{u_m = \langle f, \phi_m \rangle_{\mathcal{H}}}$$

Spherical harmonics as inducing features in SVGPs

- Define the kernel's RKHS $\mathcal{H}$ with reproducing inner-product:

$$\langle k(\mathbf{x}, \cdot), h(\cdot) \rangle_{\mathcal{H}} = h(\mathbf{x})$$

- Approximate posterior constructed out of inducing features

$$\boxed{u_m = \langle f, \phi_m \rangle_{\mathcal{H}}}$$

$\implies$ Diagonal covariance matrix: $[K_{\mathbf{uu}}]_{m,m'} = \mathrm{Cov}(u_m, u_{m'}) = \langle \phi_m, \phi_{m'} \rangle_{\mathcal{H}} = \lambda_m^{-1} \, \delta_{mm'}$

Spherical harmonics as inducing features in SVGPs

- Define the kernel's RKHS $\mathcal{H}$ with reproducing inner-product:

$$\langle k(\mathbf{x}, \cdot), h(\cdot) \rangle_{\mathcal{H}} = h(\mathbf{x})$$

- Approximate posterior constructed out of inducing features

$$\boxed{u_m = \langle f, \phi_m \rangle_{\mathcal{H}}}$$

$\implies$ Diagonal covariance matrix: $[K_{\mathbf{uu}}]_{m,m'} = \mathrm{Cov}(u_m, u_{m'}) = \langle \phi_m, \phi_{m'} \rangle_{\mathcal{H}} = \lambda_m^{-1}\, \delta_{mm'}$

$\implies$ Spherical Harmonics as features $[k_{\mathbf{u}}(\cdot)]_m = \mathrm{Cov}(u_m, f(\cdot)) = \phi_m(\cdot)$

Spherical harmonics as inducing features in SVGPs

- Define the kernel's RKHS $\mathcal{H}$ with reproducing inner-product:

$$\langle k(\mathbf{x}, \cdot), h(\cdot) \rangle_{\mathcal{H}} = h(\mathbf{x})$$

- Approximate posterior constructed out of inducing features

$$\boxed{u_m = \langle f, \phi_m \rangle_{\mathcal{H}}}$$

$\Longrightarrow$ Diagonal covariance matrix: $[K_{\mathbf{uu}}]_{m,m'} = \mathrm{Cov}(u_m, u_{m'}) = \langle \phi_m, \phi_{m'} \rangle_{\mathcal{H}} = \lambda_m^{-1}\, \delta_{mm'}$

$\Longrightarrow$ Spherical Harmonics as features $[k_{\mathbf{u}}(\cdot)]_m = \mathrm{Cov}(u_m, f(\cdot)) = \phi_m(\cdot)$

$\Longrightarrow$ A $\mathcal{O}(M^2 N)$ approximate GP $q(f(\cdot))$

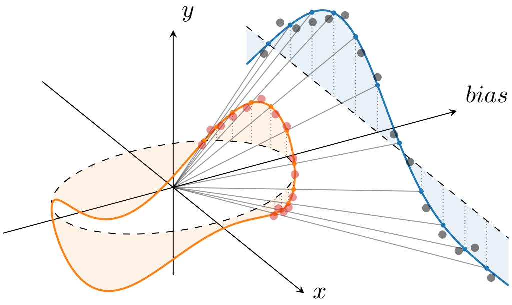
$$\mathcal{GP}\Big(\mathbf{\Phi}^\top(\cdot)\, \mathbf{m}; \quad k(\cdot, \cdot') - \mathbf{\Phi}^\top(\cdot)(\mathbf{\Lambda} - S)\mathbf{\Phi}(\cdot')\Big),$$

where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \dots, \lambda_M)$ and $\mathbf{\Phi}(\cdot) = [\phi_1(\cdot), \dots, \phi_M(\cdot)]$.

Linear mapping to the hypersphere

Most datasets do not correspond to data on a hypersphere...

The proposed solution is to augment the inputs with a constant variable (bias) before projecting it radially onto the hypersphere.

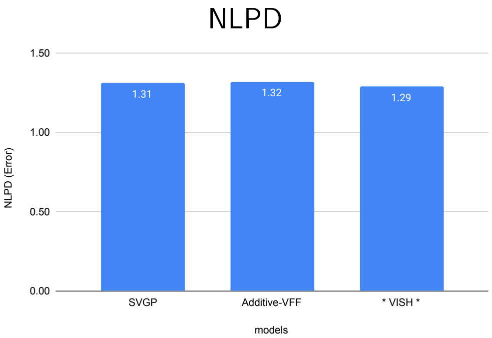


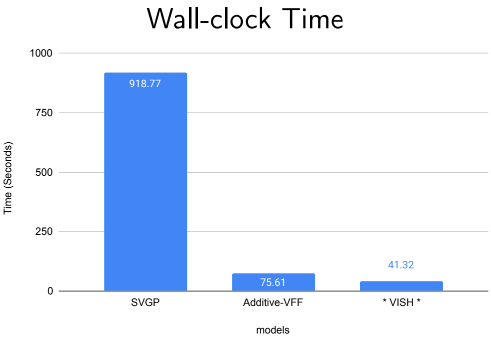
Although such construction may seem arbitrary, it is used implicitly in the Arc-Cosine kernel [Cho & Saul, 2009]:

$$k(\mathbf{x}, \mathbf{x}') = \underbrace{\|\mathbf{x}\| \|\mathbf{x}'\|}_{\text{radial}} \underbrace{(\sin\theta + (\pi - \theta)\cos\theta)}_{\text{angular}} \quad \text{with } \theta = \arccos \frac{\mathbf{x}^\top \mathbf{x}'}{\|\mathbf{x}\| \|\mathbf{x}'\|}.$$

Experiment

Airline dataset: 6,000,000 datapoints regression task fitted in 40 seconds on a single cheap GTX 1070 GPU





Conclusion

Summary of the advantages

- It is the fastest SVGP model to date
  $\Rightarrow$ No need for expensive hardware
- The natural ordering of spherical harmonics makes our model scale nicely with the input dimension
  $\Rightarrow$ Does not suffer from the curse of dimensionality as VFF
- Similarities with Arc-cosine kernel makes extrapolation properties similar to Neural Networks

Reach out to have a chat if you want to know more!