Random Feature Expansions for Deep Gaussian Processes

Kurt Cutajar[1], Edwin V. Bonilla[2], Pietro Michiardi[1], Maurizio Filippone[1]

1 - EURECOM, Sophia Antipolis, France 2 - University of New South Wales, Sydney, Australia

Deep Gaussian Processes

- Deep probabilistic models;
- Composition of functions:

$$\mathrm{f}(\mathrm{x}) = \left(\mathrm{h}^{(N_\mathrm{h}-1)}\left(\theta^{(N_\mathrm{h}-1)}\right) \circ \ldots \circ \mathrm{h}^{(0)}\left(\theta^{(0)}\right)\right)(\mathrm{x});$$

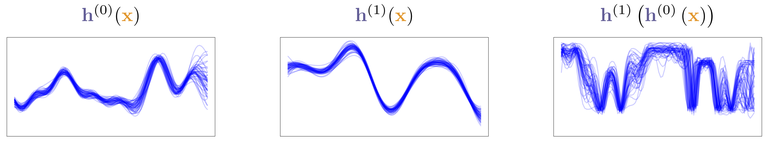

Fig. 1: Illustration of how stochastic processes may be composed.

- Inference requires calculating the marginal likelihood:

$$p(Y|X,\theta) = \int p\left(Y|F^{(N_\mathrm{h})},\theta^{(N_\mathrm{h})}\right) \times p\left(F^{(N_\mathrm{h})}|F^{(N_\mathrm{h}-1)},\theta^{(N_\mathrm{h}-1)}\right) \times \ldots \times p\left(F^{(1)}|X,\theta^{(0)}\right) \mathrm{d}F^{(N_\mathrm{h})} \ldots \mathrm{d}F^{(1)};$$

- Extremely challenging!

DGPs with Random Features

- GPs are single-layered Neural Nets with an infinite number of hidden units;
- Taking a weight-space view of a GP:

$$F = \Phi W;$$

- The priors over the weights are:

$$p(W_{\cdot i}) = \mathcal{N}(\mathbf{0}, I);$$

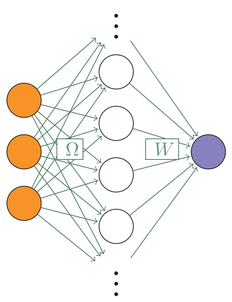

Fig. 2: Single-layered GP.

- The RBF kernel:

$$k_\mathrm{rbf}(\mathrm{x},\mathrm{x}') = \exp\left[-\frac{1}{2}(\mathrm{x}-\mathrm{x}')^\top(\mathrm{x}-\mathrm{x}')\right] = \int p(\omega)\exp\left(\iota(\mathrm{x}-\mathrm{x}')^\top\omega\right)\mathrm{d}\omega$$

can be approximated using trigonometric functions:

$$\Phi_\mathrm{rbf} = \sqrt{\frac{\sigma^2}{N_\mathrm{RF}}}\left[\cos(F\Omega), \sin(F\Omega)\right] \quad \text{with} \quad p(\Omega_{\cdot j}|\theta) = \mathcal{N}\left(\mathbf{0}, \Lambda^{-1}\right),$$

allowing for scaling factors σ^2 and $\Lambda = \mathrm{diag}(\ell_1^2, \ldots, \ell_d^2)$ for the kernel and the features;

- Meanwhile, the first order Arc-Cosine kernel:

$$k_\mathrm{arc}(\mathrm{x},\mathrm{x}') = \frac{1}{\pi}\|\mathrm{x}\|\|\mathrm{x}'\|J(\alpha) = 2\int \max(\mathbf{0},\omega^\top\mathrm{x})\max(\mathbf{0},\omega^\top\mathrm{x}')\mathcal{N}(\omega|\mathbf{0},I)\,\mathrm{d}\omega \quad \text{where}$$

$$J(\alpha) = \sin\alpha + (\pi-\alpha)\cos\alpha \quad \text{and} \quad \alpha = \cos^{-1}\left(\frac{\mathrm{x}^\top\mathrm{x}'}{\|\mathrm{x}\|\|\mathrm{x}'\|}\right)$$

can be approximated using Rectified Linear Units (ReLU):

$$\Phi_\mathrm{arc} = \sqrt{\frac{2\sigma^2}{N_\mathrm{RF}}}\max(\mathbf{0}, F\Omega) \quad \text{with} \quad p(\Omega_{\cdot j}|\theta) = \mathcal{N}\left(\mathbf{0}, \Lambda^{-1}\right).$$

Model Architecture

- DGPs with random features become DNNs with low-rank weight matrices!

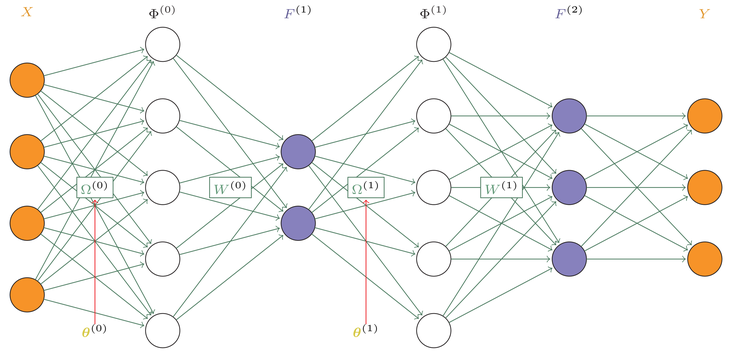

Fig. 3: Diagram of the proposed DGP model with random features.

Stochastic Variational Inference

- Define $\Psi = \left(\Omega^{(0)}, \ldots, \Omega^{(L)}, W^{(0)}, \ldots, W^{(L)}\right)$;
- Lower bound on marginal likelihood:

$$\log\left[p(Y|\theta)\right] \geq \mathrm{E}_{q(\Psi)}\left(\log\left[p\left(Y|\Psi\right)\right]\right) - \mathrm{D}_\mathrm{KL}\left[q(\Psi)\|p\left(\Psi|\theta\right)\right],$$

where $q(\Psi)$ approximates $p(\Psi|Y,\theta)$;

- Factorized approximate posterior:

$$q(\Psi) = \prod_{ijl} q\left(\Omega_{ij}^{(l)}\right)\prod_{ijl} q\left(W_{ij}^{(l)}\right),$$

with

$$q\left(W_{ij}^{(l)}\right) = \mathcal{N}\left(\mu_{ij}^{(l)}, (\sigma^2)_{ij}^{(l)}\right) \quad \text{and} \quad q\left(\Omega_{ij}^{(l)}\right) = \mathcal{N}\left(m_{ij}^{(l)}, (s^2)_{ij}^{(l)}\right);$$

- Assuming factorized likelihood, we can use mini-batch stochastic gradient optimization:

$$\frac{n}{m}\sum_{k\in\mathcal{I}_m}\mathrm{E}_{q(\Psi)}(\log[p(y_k|\Psi)]) - \mathrm{D}_\mathrm{KL}[q(\Psi)\|p(\Psi|\theta)];$$

- The expectation can be estimated using Monte Carlo:

$$\mathrm{E}_{q(\Psi)}\left(\log\left[p(y_k|\Psi)\right]\right) \approx \frac{1}{N_\mathrm{MC}}\sum_{r=1}^{N_\mathrm{MC}}\log[p(y_k|\tilde{\Psi}_r)],$$

with $\tilde{\Psi}_r \sim q(\Psi)$;

- Computational cost dominated by low-rank matrix multiplication - no inverses required;
- Various optimization strategies for Ω available.

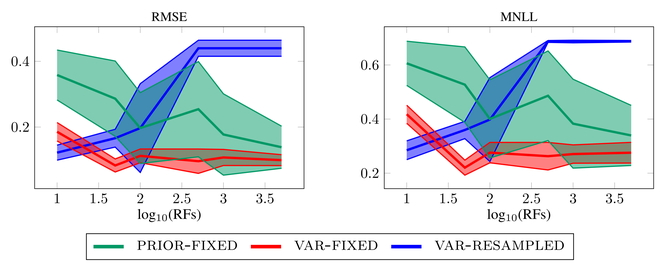

Fig. 4: Performance of different strategies for dealing with $\boldsymbol{\Omega}$ as a function of the number of random features. These can be fixed (PRIOR-FIXED), or treated variationally (with fixed randomness VAR-FIXED or resampled at each iteration VAR-RESAMPLED).

Experimental Setup and Results

- Competing methods:
  - DGP-RF with RBF Kernel (DGP-RBF);
  - DGP-RF with first order Arc-Cosine Kernel (DGP-ARC);
  - DGP with Expectation Propagation (DGP-EP);
  - DNN with Dropout (DNN);
  - Sparse GP with Variational Inference (VAR-GP).

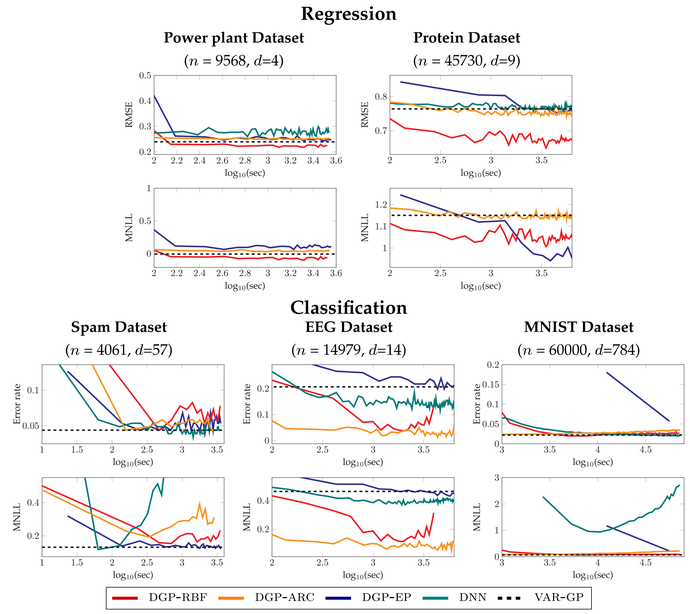

Fig. 5: Progression of RMSE and MNLL on test data over time for competing models.

- Easily extendable architecture with option to feed-forward inputs.

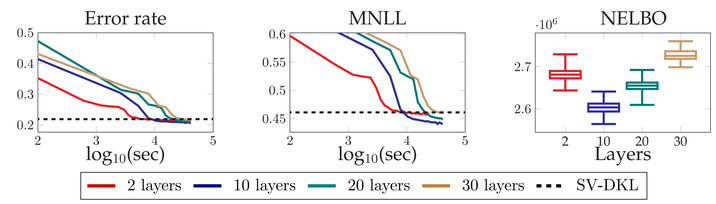

Fig. 6: Left and center - Performance of our model on the AIRLINE dataset as function of time for different depths. Right - The box plot of the negative evidence lower bound confirms this is a suitable objective for model selection.

Conclusions

- Our contributions:
  - ✓ Complete specification and evaluation of DGPs based on random features;
  - ✓ Scalable and practical DGP inference - no matrix inverses;
  - ✓ Synchronous/asynchronous distributed implementation available;
- Ongoing work:
  - Fastfood and other kernels;
  - Convolutional GP layers for image processing.

References

[1] A. Damianou and N. Lawrence. *Deep Gaussian Processes*, AISTATS 2013.

[2] Y. Gal and Z. Gharamani. *Dropout as a Bayesian Approximation*, ICML 2016.