Accelerating Deep Gaussian Processes Inference with Arc-Cosine Kernels

Kurt Cutajar
EURECOM, France
kurt.cutajar@eurecom.fr

Edwin V. Bonilla
University of New South Wales, Australia
e.bonilla@unsw.edu.au

Pietro Michiardi
EURECOM, France
pietro.michiardi@eurecom.fr

Maurizio Filippone
EURECOM, France
maurizio.filippone@eurecom.fr

Abstract

Deep Gaussian Processes (DGPs) are probabilistic deep models obtained by stacking multiple layers implemented through Gaussian Processes (GPs). Although attractive from a theoretical point of view, learning DGPs poses some significant computational challenges that arguably hinder their application to a wider variety of problems for which Deep Neural Networks (DNNs) are the preferred choice. We aim to bridge the gap between DGPs and DNNs by showing how random feature approximations to DGPs can leverage the key strengths of DNNs, while retaining a probabilistic formulation for accurate quantification of uncertainty. In particular, we show how DGPs with arc-cosine kernels can be approximated by DNNs with Rectified Linear Unit (ReLU) activation functions, leading to competitive performance and faster inference compared to state-of-the-art DGPs inference approaches.

1 Introduction

Deep Gaussian Processes (DGPs) are deep probabilistic models obtained by stacking multiple layers implemented through Gaussian Processes (GPs). This construction to model composition naturally resembles that of Deep Neural Networks (DNNs). The connection between DGPs and DNNs has been extensively investigated in the literature, and DGPs with a variety of kernel functions can be interpreted as DNNs with an infinite number of neurons at each layer and specific activation functions (see, e.g., [6,11]). In contrast to DNNs, DGPs provide an elegant way to deal with the model-selection problem of determining a suitable number of neurons, as they are inherently nonparametric learning machines. Furthermore, they allow for a principled probabilistic framework to carry out learning of latent representations and hyper-parameters.

Although attractive from a theoretical point of view, learning DGPs poses some significant computational challenges that arguably hinder their application to a wider variety of problems. In contrast, DNNs have been extremely successful in areas such as computer vision because of their amenability to GPU and distributed computations, automatic differentiation tools, and mature developments of regularization techniques, such as low-rank weight representations and dropout [14]. We aim to bridge the gap between DGPs and DNNs by showing how DGPs can be learned at scale by borrowing the key strengths of DNNs, while retaining a probabilistic formulation for accurate quantification of uncertainty. We made a significant step in this direction in [4], by showing how random Fourier feature approximations for DGPs with Radial Basis Function (RBF) kernels lead to trigonometric DNNs with low-rank weight matrices. We conveniently implemented stochastic variational inference [8] in TensorFlow [1] to infer the parameters of such approximate DGPs.

Workshop on Bayesian Deep Learning, NIPS 2016, Barcelona, Spain.
In this paper, we aim to take this idea one step further by showing how DGP with arc-cosine kernels can be approximated by DNNs with Rectified Linear Unit (ReLU) activations. Such activation functions are much cheaper to evaluate and differentiate than the trigonometric functions used in the approximation of DGP with RBF kernels, leading to faster DGP inference. We experimentally validate this acceleration effect, while showing that the proposed approximation leads to competitive performance compared with the approximation using trigonometric activations.

2 Random features for Deep Gaussian Processes with Arc-cosine Kernels

Let $X = [x_1, \ldots, x_n]^T$ be a set of input vectors associated with a set of labels $Y = [y_1, \ldots, y_n]^T$, where $x_i \in R^{D_{in}}$ and $y_i \in R^{D_{out}}$. Let $F^{(0)} := X$ and consider a DGP where we define $F^{(l)}$ as the multivariate latent representation at layer $l$, which is a collection of zero-mean GPs with covariance $k(\cdot, \cdot; \theta)$. Each layer of DGP can be approximated based on a finite set of basis functions $F^{(l+1)} = \Phi^{(l)} W^{(l)}$, leading to a DNN representation of DGPs \cite{2}. The set of basis functions $\Phi^{(l)}$ and the distribution of $W^{(l)}$ depend on the choice of the kernel (covariance) function. For example, in the RBF case, Bochner’s theorem suggests an expansion using trigonometric functions as basis functions \cite{12} and $W_{ij}^{(l)} \sim \mathcal{N}(0, 1)$.

In \cite{2}, we explored the random feature approximation of RBF kernels using trigonometric functions. Here we propose arc-cosine kernels to accelerate learning of DGPs with the random features flavor. The arc-cosine kernel of order $p$ is defined as:

$$k_{\text{arc}}^{(p)}(x, x') = \frac{1}{\pi} \|x\|^{p} \|x'\|^{p} J_{p} \left( \frac{x \cdot x'}{\|x\| \|x'\|} \right) \quad \text{with} \quad (1)$$

$$J_{p}(\alpha) = (-1)^{p} (\sin \alpha)^{2p+1} \frac{1}{\sin \alpha} \frac{\partial}{\partial \alpha} \left( \frac{\pi - \alpha}{\sin \alpha} \right). \quad (2)$$

Let $\Theta(\cdot)$ be the Heaviside function. Following \cite{3}, an integral representation of this kernel is:

$$k_{\text{arc}}^{(p)}(x, x') = 2 \int \Theta(\omega \cdot x) \Theta(\omega \cdot x') \omega \cdot x)^{p} \omega \cdot x')^{p} \mathcal{N}(\omega|0, I) \, d\omega \quad (3)$$

This integral formulation immediately suggests a random feature approximation for equation (1), noticing that the kernel can be seen as an expectation of the product of the same function applied to the inputs to the kernel. We can extend this by scaling the kernel with a $\sigma^2$ parameter, and features through $\Lambda = \text{diag}(\lambda_{1}^{2}, \ldots, \lambda_{d}^{2})$ for automatic relevance determination \cite{9}. In the case $p = 1$, the application of the random feature approximation of DGPs with arc-cosine kernels leads to DNNs with ReLU activations:

$$\Phi^{(l)}_{\text{arc}} = \sqrt{2(\sigma^2)^{(l)}} \max_{\Omega^{(l)}_{j}} \mathcal{N} \left( 0, \Omega^{(l)}_{j} \right), \quad \text{with} \quad \Omega^{(l)}_{j} \sim \mathcal{N} \left( 0, \left(\Lambda^{(l)}\right)^{-1} \right), \quad (4)$$

which are cheaper to evaluate and differentiate compared to trigonometric functions in the RBF case. Similarly to \cite{2}, we learn an approximate posterior over all weight matrices $\Omega^{(l)}$ and $W^{(l)}$ and optimize kernel parameters $(\sigma^2)^{(l)}$ and $\Lambda^{(l)}$ using stochastic variational inference.

3 Results

We follow the same experimental setup of \cite{2}, where we evaluate the performance of DGP approaches on regression and classification tasks. We compare the approximate DGP with arc-cosine kernel (DGP-ARC) with the approximate DGP with RBF kernel (DGP-RBF) \cite{2}, and the recently proposed DGP model approximated using expectation propagation (DGP-EP) \cite{2}. All models exploit mini-batch-based stochastic gradient optimization. To ensure fairness in the comparison, we use the same configurations across all models, namely one three-dimensional hidden layer, number of inducing points/number of random features equal to 100 and same mini-batch size. An exception is made for the MNIST case, where greater dimensionality and more inducing points/random features are required (see \cite{2}). We evaluate the error rate and mean negative log-likelihood (MNLL) on withheld
we also assessed the performance of the arc-cosine kernel with degree two, but this case performed poorly compared to the arc-cosine kernel of degree one. Finally, we note that for the arc-cosine kernel with degree zero, the activation functions are Heaviside functions that are unsuitable for our inference scheme given that they yield systematically zero gradients.

Summarizing, we regard these results as a further step in the direction of bridging the gap between DGPs and DNNs, hoping that this will foster a wider adoption of DGPs in applications.

Acknowledgements

MF gratefully acknowledges support from the AXA Research Fund.

References


