Deeply Non-Stationary Gaussian Processes

Hugh Salimbeni
Imperial College London
hrs13@ic.ac.uk

Marc Peter Deisenroth
Imperial College London
m.deisenroth@imperial.ac.uk

Abstract

We propose a new Deep Gaussian Process (DGP) model, constructed from a hierarchy of non-stationary processes. In the standard DGP formulation, the output of one GP is the input to the next in the hierarchy. This construction can lead to undesirable properties such as exponential multimodality of the posterior, non-local correlations, and degeneracy. Instead, we use a non-stationary kernel for each layer, and use the Gaussian process of the previous layer to model the lengthscale. We apply recently proposed doubly stochastic variational inference, which scales to any size of data and retains the conditional dependencies of the exact model.

1 Introduction

Deep Gaussian processes (DGP) [Damianou and Lawrence 2013] are multilayer generalizations of Gaussian process (GP) models. One way to chain GPs together is to use the outputs of each layer as the inputs to the next. This has an appealing parallel with the construction of feed-forward neural networks, but leads to undesirable properties. One problem with the direct function composition approach is that the GP mapping for typical kernels is highly non-injective. The activation functions used in neural network models are typically monotonic, but a GP non-linearity can never be monotonic as the marginals of a monotonic function cannot be Gaussian. As many inputs are mapped to the same outputs, each layer reduces the degrees of freedom of the space, resulting in highly degenerate covariances. The inclusion of a identity mean function [Salimbeni and Deisenroth 2017] or forward propagating the inputs [Duvenaud et al. 2014] mitigates the problem to an extent, but the pathological behaviour is still attainable under certain hyperparameter settings. See Fig. 1 for a demonstration of this effect.

We propose an alternative construction that completely avoids this pathological behaviour, based on non-stationary covariance functions. A non-stationary covariance function can be constructed from any stationary covariance using the approach of Paciorek and Schervish [2004]. Each point is associated with a local lengthscale, with the covariance between two points dependent on both the input locations and the lengthscales at each point. We use another GP to model the (log of) the lengthscale values. The special case of a single two dimensional hidden layer has been extensively studied in the geostatistics community using discretized representations [Paciorek 2003, Damian et al. 2001, Fuglstad et al. 2015, Roininen et al. 2016]. We extend these approaches to multi-layered hierarchies with high dimensional latent layers.

We use the approach of Salimbeni and Deisenroth [2017] for inference, employing stochastic sparse variational inference with the reparameterization trick Rezende et al. [2014], Kingma and Welling [2013] for gradients. Our sparse inference decouples the function values at each layer conditioned on a reduced set of inducing points Hensman et al. [2015]. A subtlety we need to consider in the non-stationary model is how to model the lengthscale at the inducing points (in the composition DGP we only need the inputs, which are variational parameters). An important property of the sparse approximation is that the input locations are variational parameters and free to take any values. To achieve inference in the DGP model it is not necessary to couple the inducing points to the outputs of the previous layer, greatly simplifying inference without weakening the quality of the approximation. In the DNSGP we can use the same idea and choose to model the inducing input lengthscales as

Second workshop on Bayesian Deep Learning (NIPS 2017), Long Beach, CA, USA.

variational parameters. This is valid as we are free to define the model in this way without changing the distribution over the data. The key point is that the GPs themselves are independent a priori: it is only the *evaluation locations* which couple the layers.

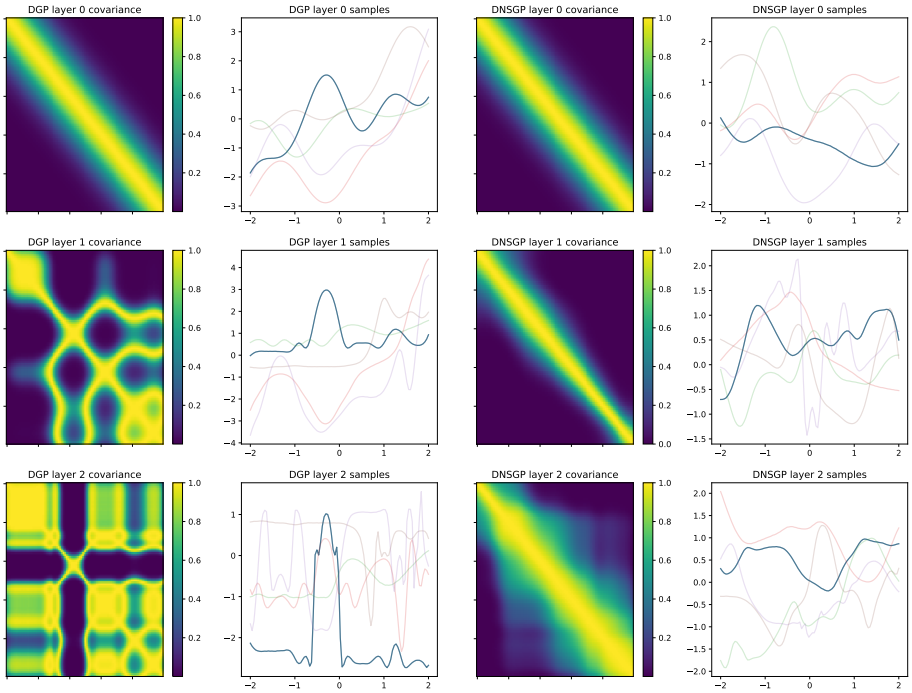

Figure 1: Three layer prior samples for the DGP (left) and DNSGP (right). A single sample is propagated through the layers, top to bottom. Further samples to give a more representative idea of the prior are shown faintly behind. Note that in all but the first layer the faint samples correspond to different covariances (they are independent samples from the whole model). The DGP covariance has characteristic 'loops' at the second layer due to the non-injectivity of the first layer mapping. Subsequent layers have increasingly degenerate covariances. The DNSGP does not suffer from this problem and creates richer covariances without pathologically collapse

2 Model

While composing GPs directly in the manner of the DGP is a recent idea, models involving more than one GP in a hierarchy have been around for a long time. In particular, the Non-Stationary Gaussian Process (NSGP) was developed by Paciorek [2003], extending ideas from Sampson and Guttorp [1992]. The key idea is that any stationary kernel $k(\mathbf{a}, \mathbf{b}) = \phi(r(\mathbf{a}, \mathbf{b}))$, where ϕ is a scalar function and $r^2(\mathbf{a}, \mathbf{b}) = \sum_{d=1}^{D}(a_d - b_d)^2 l_d^{-2}$, can be extended to a non-stationary version via

$$k^{\text{NS}}(\mathbf{a}, \mathbf{b}) = q(\mathbf{a}, \mathbf{b})\phi(s(\mathbf{a}, \mathbf{b}))\,, \tag{1}$$

where

$$s(\mathbf{a}, \mathbf{b}) = \sum_{d=1}^{D}(a_d - b_d)^2 \left(l_d(\mathbf{a})^2 + l_d(\mathbf{b})^2\right)^{-1}\,, \tag{2}$$

i.e. the Euclidean distance with constant lengthscale replaced by a input dependent term, and

$$q(\mathbf{a}, \mathbf{b}) = \prod_{d=1}^{D} \sqrt{2 l_d(\mathbf{a}) l_d(\mathbf{b}) \left(l_d(\mathbf{a})^2 + l_d(\mathbf{b})^2\right)^{-1}}\,, \tag{3}$$

where $l_d(\cdot)$ is a positive scalar function. We model l_d with a GP passed through a monotonic positive function, e.g. softplus. We need a constant mean function for each component to set the a priori average lengthscale for each layer. For inference we follow the approach of Salimbeni and Deisenroth [2017]. If we use variational parameters for the inducing point lengthscales we can use the approach

without any changes to the code other than the implementation of the non-stationary kernel. The only difference we need is to change the mean functions: in Salimbeni and Deisenroth [2017] the mean functions at all but the final layers are the identity. Here we can use the zero mean function as the GP is modelling the lengthscale. We omit all the details and refer the reader to Salimbeni and Deisenroth [2017].

3 Results

We apply the DNSGP to 8 UCI regression datasets. The results from 5 fold cross validation are shown in Fig 2. On three of the datasets (boston, wine-red and naval) all models perform similarly. As discussed in Salimbeni and Deisenroth [2017], this is because these datasets are well-modelled by a single layer GP and the deep models recover the single layer models. This is in contrast to non-Bayesian models which typically suffer from overfitting without careful regularization. On the remaining 5 dataset the deep models significantly outperform the single layer models. On the power and concrete datasets it seems that DNSGP is a superior model, albeit only slightly.

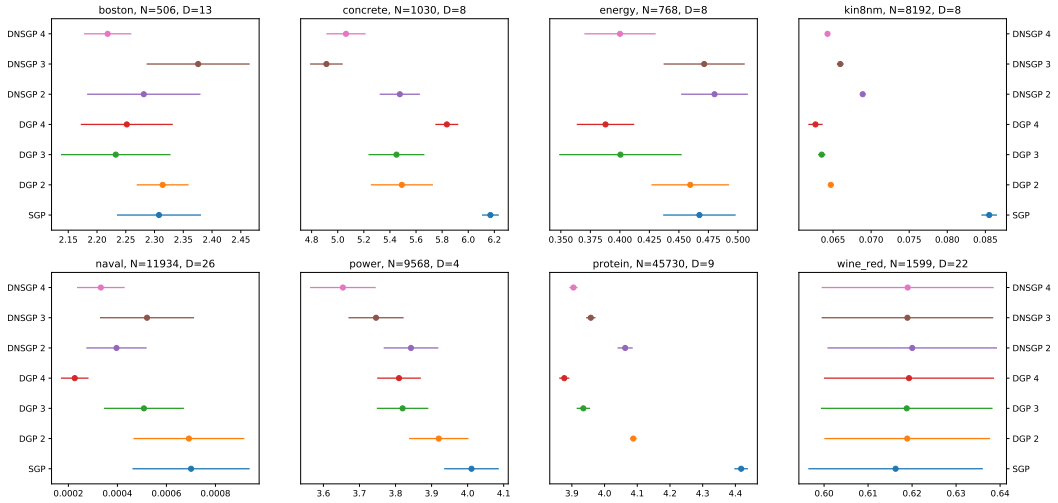

Figure 2: Regression test RMSE results with 5 fold cross validation. Lower (to the left) is better). The DNSGP outperforms the DGP on the power and concrete datasets, and performs similarly or slightly worse on protien and kin8nm, but both substantially outperform the single layer model. On wine, naval and boston the models all perform similarly

4 Discussion

Unlike the direct input warping approach of the DGP, the DNSGP does not require a 'fix' to overcome pathological behaviour. The DNSGP provides a more interpretable way to combine Gaussian processes in a multi-layer hierarchy. On preliminary experiments, we have shown that the DNSGP GP can perform at least as well as the DGP on regression tasks. There are some settings where the DGP prior may be more appropriate, for example when data has periodicities that are not captured by the kernel. In this setting the DGP can warp the space to create periodic covariances at the final layer. In other situations, however, we suggest that the DNSGP prior might be more appropriate as it more closely preserves the properties of the final layer GP.

A central challenge faced by the Bayesian deep learning community is building rich priors that both represent reasonable prior beliefs and admit effective inference. We suggest that the DNSGP fulfils both these criteria. The DNSGP is naturally suited to spatial data where smoothness can be assumed but discontinuities might exist at locations that are not known a priori. Such data are difficult to model with a single layer GP. The DGP can model the discontinuities, but introduces the potential for non-local correlations which might not be an appropriate. The DNSGP in contrast can model exactly this sort of data. In future work we will apply the DNSGP to spatial situations such as demographic data.

References

D. Damian, P. D. Sampson, and P. Guttorp. Bayesian Estimation of Semi-Parametric Non-Stationary Spatial Covariance Structures. *Environmetrics*, 2001.

A. Damianou and N. Lawrence. Deep Gaussian Processes. *Artificial Intelligence and Statistics*, 2013.

D. Duvenaud, O. Rippel, R. Adams, and Z. Ghahramani. Avoiding Pathologies in Very Deep Networks. *Artificial Intelligence and Statistics*, 2014.

G.-A. Fuglstad, D. Simpson, F. Lindgren, and H. Rue. Does Non-Stationary Spatial Data Always Require Non-Stationary Random Fields? *Spatial Statistics*, 2015.

J. Hensman, A. Matthews, M. Fillipone, and Z. Ghahramani. MCMC for Variationally Sparse Gaussian Processes. *Advances in Neural Information Processing Systems*, 2015.

D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. *arXiv preprint: 1312.6114*, 2013.

C. J. Paciorek. Nonstationary Gaussian Processes for Regression and Spatial Modelling. *PhD Thesis*, 2003.

C. J. Paciorek and M. J. Schervish. Nonstationary Covariance Functions for Gaussian Process Regression. *Advances in Neural Information Processing Systems*, 2004.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. *arXiv preprint: 1401.4082*, 2014.

L. Roininen, M. Girolami, S. Lasanen, and M. Markkanen. Hyperpriors for Matérn Fields with Applications in Bayesian Inversion. *arXiv preprint:1612.02989*, 2016.

H. Salimbeni and M. Deisenroth. Doubly stochastic variational inference for deep gaussian processes. *Advances in Neural Information Processing Systems*, 2017.

P. Sampson and P. Guttorp. Nonparametric Estimation of Nonstationary Spatial Covariance Structure. *Journal Of The American Statistical Association*, 1992.