Inter-domain Deep Gaussian Processes

Tim G. J. Rudner
University of Oxford
timr@robots.ox.ac.uk

Dino Sejdinovic
University of Oxford
dino.sejdinovic@stats.ox.ac.uk

Abstract

We propose a novel variational inference method for deep Gaussian processes (GPs), which combines doubly stochastic variational inference with variational Fourier features, an inter-domain approach that replaces inducing points-based inference with a framework that harnesses RKHS Fourier features. First experiments have shown that inter-domain deep Gaussian processes are able to achieve levels of predictive performance superior to shallow GPs and alternative deep GP models.

1 Introduction

Gaussian process (GP) methods are powerful tools for function approximation. They are non-parametric probabilistic models and as such they are flexible, robust to overfitting, and provide well-calibrated predictive uncertainty estimates [1]. Deep Gaussian processes (DGPs) are multi-layer hierarchical generalizations of GPs and promise to overcome the limitations of traditional GPs without compromising their advantages [14].

Under certain conditions, a GP can be viewed as a neural network with a single infinite-dimensional layer of hidden units [13], and deep GPs have been argued to be a type of infinitely-wide, deep neural network [6]. In fact, it was shown that a deep neural network with arbitrary depth and non-linearities, with dropout applied before every weight layer, is mathematically equivalent to an approximation to a deep GP [7]. Deep GPs hence offer a promising framework that allows us to combine deep architectures with a principled, Bayesian approach to obtaining predictive uncertainties.

We propose a novel variational inference procedure that combines the approaches presented in [14] and [9], making it the first inter-domain inference method for deep GPs. We leverage the many desirable properties of doubly stochastic variational inference (DSVI) for deep GPs and augment the inference procedure by replacing the standard inducing points methodology with an inter-domain approach that allows us to transform the space of inducing variables and capture more information about the underlying process.

2 Background

Deep GPs were first developed in [4] and have since been extended to improve their stability and scalability [1, 2, 3, 6, 10, 14]. Deep GPs are compositions of GPs in which the output of the previous layer is used as the input to the next layer. Similar to deep neural networks, the hidden layers of a deep GP learn representations of the input data, but in contrast to neural networks, they allow uncertainty to be propagated through the hierarchy and learn the hidden layer representations variationally.

Let $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]^\top$ denote the $N \times P$-dimensional training data, and let $\mathbf{y}$ denote the $N \times D$-dimensional noisy observations of the response variable. Formally, to describe a deep GP, we consider a nested structure of the form

$$\mathbf{y} = \mathbf{F}^{1:L} + \boldsymbol{\epsilon} = f^L(f^{L-1}(...f(\mathbf{X}))...) + \boldsymbol{\epsilon}, \tag{1}$$

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

where L is the number of layers, $\mathbf{F}^0 = \mathbf{X}$, and each $\mathbf{F}^\ell = f^\ell(\mathbf{F}^{\ell-1})$ in the composition $\mathbf{F}^{1:L}$ is a draw from the ℓth-layer GP, obtained by pointwise evaluation of f^ℓ at $\mathbf{F}^{\ell-1}$. We follow the notation in [14] and absorb the noise between layers, which is assumed to be i.i.d. Gaussian, into the kernel so that $k_{noisy}(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j) + \sigma_l^2 \delta_{ij}$, where δ_{ij} is the Kronecker delta and σ_l^2 is the noise variance between layers.

2.1 Doubly stochastic variational inference for deep Gaussian processes

Our approach closely follows the inference procedure in [14], which diverges from earlier inference methods for deep GPs in that it does not assume independence or Gaussianity between layers. It maintains the correlations between layers and has the same non-linear structure as the full model, making it analytically intractable. Figure 1a shows a graphical representation of the model structure.

Let $\mathbf{Z}^{\ell-1} = [\mathbf{z}_1^{\ell-1}, ..., \mathbf{z}_M^{\ell-1}]^\top$ be the matrix of $M \times P^{\ell-1}$ inducing inputs and $\mathbf{U}^\ell = [\mathbf{u}_1^\ell, ..., \mathbf{u}_M^\ell]^\top$ be the $M \times P^\ell$-dimensional matrix of inducing variables with $\mathbf{U}^\ell = f(\mathbf{Z}^{\ell-1})$. Define the covariance function as $\mathbf{K_{vw}} \equiv k(\mathbf{V}, \mathbf{W})$ with $[k(\mathbf{V}, \mathbf{W})]_{ij} = k(\mathbf{v}_i, \mathbf{w}_j)$ for any input matrices $\mathbf{V}, \mathbf{W}$. To perform stochastic variational inference in a deep GP model, [14] propose a variational posterior with three properties: first, conditioned on $\mathbf{U}^\ell$, the variational posterior maintains the exact model $q(\mathbf{F}^\ell, \mathbf{U}^\ell) = p(\mathbf{F}^\ell \,|\, \mathbf{U}^\ell) q(\mathbf{U}^\ell)$; second, the posterior distribution of $\{\mathbf{U}^\ell\}_{\ell=1}^L$ factorizes across layers (and dimensions), which implies that the variational posterior takes the form

$$\mathcal{Q} = q(\{\mathbf{H}^\ell, \mathbf{U}^\ell\}_{\ell=1}^L) = \prod_{\ell=1}^L p(\mathbf{F}^\ell \,|\, \mathbf{U}^\ell, \mathbf{F}^{\ell-1}) q(\mathbf{U}^\ell); \tag{2}$$

and third, assume $q(\mathbf{U}^\ell)$ is Gaussian with mean $\boldsymbol{\mu}^\ell$ and variance $\boldsymbol{\Sigma}^\ell$. These properties allow us to marginalize $q(\mathbf{U}^\ell)$ from $\mathcal{Q}$ and get

$$q(\{\mathbf{F}^\ell\}_{\ell=1}^L) = \prod_{\ell=1}^L q(\mathbf{F}^\ell \,|\boldsymbol{\mu}^\ell, \boldsymbol{\Sigma}^\ell, \mathbf{F}^{\ell-1}) = \prod_{\ell=1}^L \mathcal{N}(\mathbf{F}^\ell \,|\widetilde{\mathbf{m}}^\ell, \widetilde{\mathbf{S}}^\ell), \tag{3}$$

where

$$\widetilde{\mathbf{m}}^\ell \equiv \widetilde{\mathbf{m}}(\mathbf{F}^\ell) = \mathbf{m}_{\mathbf{u}^\ell} - \mathbf{K}_{\mathbf{f}^\ell \mathbf{u}^\ell} \mathbf{K}_{\mathbf{u}^\ell \mathbf{u}^\ell}^{-1} (\boldsymbol{\mu}^\ell - \mathbf{m}_{\mathbf{u}^\ell}), \tag{4}$$

$$\widetilde{\mathbf{S}}^\ell \equiv \widetilde{\mathbf{S}}(\mathbf{F}^\ell, \mathbf{F}^\ell) = \mathbf{K}_{\mathbf{f}^\ell \mathbf{f}^\ell} - \mathbf{K}_{\mathbf{f}^\ell \mathbf{u}^\ell} \mathbf{K}_{\mathbf{u}^\ell \mathbf{u}^\ell}^{-1} (\mathbf{K}_{\mathbf{u}^\ell \mathbf{u}^\ell} - \boldsymbol{\Sigma}^\ell) \mathbf{K}_{\mathbf{u}^\ell \mathbf{u}^\ell}^{-1} \mathbf{K}_{\mathbf{u}^\ell \mathbf{f}^\ell}, \tag{5}$$

with mean functions $\mathbf{m}_{\mathbf{f}^\ell} \equiv m(\mathbf{F}^{\ell-1})$ and $\mathbf{m}_{\mathbf{u}^\ell} \equiv m(\mathbf{Z}^{\ell-1})$. Since, within each layer, the marginals only depend on the corresponding inputs, the nth marginal of the final layer of the variational deep GP posterior can be expressed as

$$q(\mathbf{f}_n^L) = \int \prod_{\ell=1}^{L-1} q(\mathbf{f}_n^\ell \,|\boldsymbol{\mu}^\ell, \boldsymbol{\Sigma}^\ell, \mathbf{f}_n^{\ell-1}) \, \mathrm{d}\mathbf{f}_n^\ell, \tag{6}$$

where $\mathbf{f}_n^\ell$ is the nth row of $\mathbf{F}^\ell$. This quantity is easy to compute using the reparameterization trick, which lets us sample from the nth instances of the variational posteriors across layers by defining

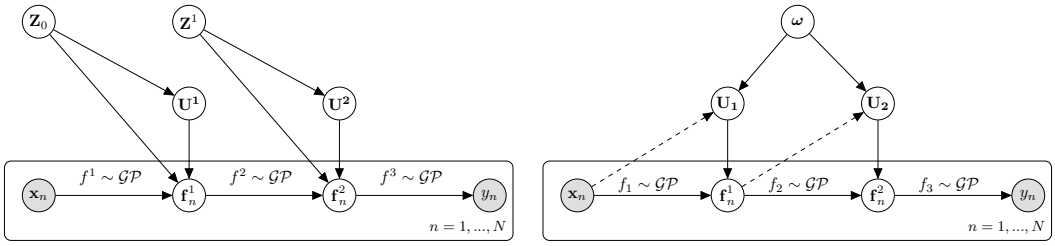
$$\hat{\mathbf{f}}_n^\ell = \widetilde{\mathbf{m}}(\hat{\mathbf{f}}_n^{\ell-1}) + \boldsymbol{\epsilon}_n^\ell \odot \sqrt{\widetilde{\mathbf{S}}(\hat{\mathbf{f}}_n^{\ell-1}, \hat{\mathbf{f}}_n^{\ell-1})} \tag{7}$$

and sampling from $\boldsymbol{\epsilon}_n^\ell \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{P^\ell})$ [11, 14]. Moreover, the reparameterization trick allows us to compute unbiased gradients of the marginal likelihood bound with respect to the variational and model parameters.

2.2 Variational Fourier features

The central idea behind using Fourier features in an inducing points-based variational inference framework is that they let us generalize the conventional inducing-inputs approach by allowing a different decomposition of the underlying process. This can be achieved by replacing the inducing variables $\mathbf{u}_m = f(\mathbf{z}_m)$ with the projection

$$\mathbf{u}_m = \int f(\mathbf{x}) g(\mathbf{x}, \mathbf{s}_m) \, \mathrm{d}\mathbf{x}, \tag{8}$$



(a) Graphical model representation of a deep GP model with inducing inputs, inducing variables, and two hidden layers, $\mathbf{F}^1$ and $\mathbf{F}^2$, for $n = 1, ..., N$.

(b) Graphical model representation of a deep GP model with sparse Fourier feature inducing variables and two hidden layers, $\mathbf{F}^1$ and $\mathbf{F}^2$, for $n = 1, ..., N$.

Figure 1: Graphical representation of conventional inducing-points deep GPs and inter-domain deep GPs.

where $\mathbf{x} \in \mathbb{R}^P$ and $\mathbf{u}_m \in \mathbb{R}^D$. The feature extraction function $g(\mathbf{x}, \mathbf{s}_m)$ used in the integral defines the transformed domain in which the inducing dataset lies. The inducing variables obtained this way can be seen as projections of the target function $f(\mathbf{x})$ on the feature extraction function over the entire input space [12]. As such, each of the inducing variables contains information about the behavior of $f(\mathbf{x})$ everywhere on the input space and thus becomes more informative about the posterior [9, 12].

The usefulness of inducing variables mostly relies on their covariance with the remainder of the process, which, for inducing points-based variational inference, is encoded in the vector-valued function $\mathbf{k_u}(\mathbf{x}) = [k(\mathbf{z}_1, \mathbf{x}), k(\mathbf{z}_2, \mathbf{x}), ..., k(\mathbf{z}_M, \mathbf{x})]$. $\mathbf{K_{uu}}$ and $\mathbf{k_u}(\mathbf{x})$ are central to inducing points-based variational inference for GPs, which is exemplified by the use of $\mathbf{K}_{\mathbf{u}^\ell\mathbf{u}^\ell}$ and $\mathbf{K}_{\mathbf{u}^\ell\mathbf{f}^\ell}$ in $\widetilde{\mathbf{m}}(\mathbf{F}^\ell)$ and $\widetilde{\mathbf{S}}(\mathbf{F}^\ell, \mathbf{F}^\ell)$ in DSVI for deep GPs described above. Variational Fourier features, in contrast, are based on a variational inference procedure that uses Reproducing Kernel Hilbert Space (RKHS) theory to construct inter-domain alternatives to $\mathbf{K_{uu}}$ and $\mathbf{k_u}(\mathbf{x})$ by projecting f onto the truncated Fourier basis,

$$\boldsymbol{\phi}(x) = [1, \cos(\omega_1(x-a)), ..., \cos(\omega_M(x-a)), \sin(\omega_1(x-a)), ..., \sin(\omega_M(x-a))]^\top, \tag{9}$$

where x is a single, one-dimensional input, and the mth frequency ω_m is defined as $\omega_m = \frac{2\pi m}{b-a}$ for some interval $[a, b]$. More specifically, for an RKHS $\mathcal{H}$, the coordinate of the projection of a function $h \in \mathcal{H}$ onto $\boldsymbol{\phi}_m(x)$ is given by

$$\mathcal{P}_{\boldsymbol{\phi}_m}(h) = \langle h, \boldsymbol{\phi}_m \rangle_{\mathcal{H}} \tag{10}$$

and defines a projection between domains. It has been shown that if $\mathcal{H}$ is a Matérn RKHS of functions over $[a, b]$, then the span of $\boldsymbol{\phi}$ belongs to $\mathcal{H}$, which ensures that the inner product between h and $\boldsymbol{\phi}_m$ is defined [5, 9]. With these results, we can construct the inducing variables by defining $u_m = \mathcal{P}_{\boldsymbol{\phi}_m}(f)$, which yields

$$\text{cov}(u_m, f(x)) = \boldsymbol{\phi}_m(x), \tag{11}$$

$$\text{cov}(u_m, u_{m'}) = \langle \boldsymbol{\phi}_m, \boldsymbol{\phi}_{m'} \rangle_{\mathcal{H}}, \tag{12}$$

for both of which there are closed-form expressions for the half-integer members of the Matérn family of kernels. The resulting operators $\mathbf{k}_{\mathbf{u}}^{\phi}(x) = \boldsymbol{\phi}(x)$ and $\mathbf{K}_{\mathbf{uu}}^{\phi} = \mathbf{K}_{\boldsymbol{\phi}\boldsymbol{\phi}}$, where $\mathbf{K}_{\boldsymbol{\phi}\boldsymbol{\phi}}$ is the Gram Matrix of $\boldsymbol{\phi}$ in $\mathcal{H}$, represent generalized, inter-domain alternatives to the $\mathbf{k_u}(x)$ and $\mathbf{K_{uu}}$ operators used in conventional inducing-points approaches.

There are several ways to apply this RKHS Fourier feature approach to multidimensional inputs [9]. In this work, we used additive kernels, which, for shallow GP models, are defined by

$$f(\mathbf{x}) = \sum_{p=1}^{P} f_p(x_p), \qquad f_p \sim \mathcal{GP}(0, k_p(x_p, x_p')), \tag{13}$$

where x_p is the pth component of $\mathbf{x}$ and $k_p(\cdot, \cdot)$ is a kernel defined on a scalar input space.

3 Inter-domain deep Gaussian processes

In inter-domain deep GPs, we marry the two previously introduced approaches. In DSVI for deep GPs, integrating out the hidden variables is straightforward due to the functional form of $q(\mathbf{f}_n^L)$ and the use of the reparameterization trick. This property allows us to use the inter-domain operators $\mathbf{K}_{\mathbf{u}^\ell \mathbf{f}^\ell}^\phi$ in the likelihood expectation without having to analytically convolve $\mathbf{K}_{\mathbf{u}^\ell \mathbf{f}^\ell}^\phi$ with the distributions of the hidden variables, which would have been necessary in previous deep GP inference methods.

Our approach constructs the variational posterior through sinusoidals by using the inter-domain operators introduced in the previous section to compute $\widetilde{\mathbf{m}}^\ell$ and $\widetilde{\mathbf{S}}^\ell$ (see Figure 1b). For RKHS Fourier feature-based inducing variables, the deep GP's joint density is then

$$p(\mathbf{y}, \{\mathbf{F}^\ell, \mathbf{U}^\ell\}_{\ell=1}^L) = \prod_{n=1}^N p(y_n|\, \mathbf{f}_n^L) \prod_{\ell=1}^L p(\mathbf{F}^\ell \,|\, \mathbf{U}^\ell, \mathbf{F}^{\ell-1}, \{\boldsymbol{\omega}_m\}_{m=1}^M) p(\mathbf{U}^\ell \,|\{\boldsymbol{\omega}_m\}_{m=1}^M). \tag{14}$$

All model properties of the inducing points-based DSVI presented in [14] are being preserved in this inter-domain framework.

To optimize inter-domain deep GPs, we maximize the expected lower bound (ELBO) on the marginal likelihood $\log(\mathbf{y})$ given by

$$\mathcal{L} = \sum_{n=1}^N \mathbb{E}_{q(\mathbf{f}_n^L)} \left[\log p(y_n|\, \mathbf{f}_n^L)\right] - \sum_{\ell=1}^L \mathrm{KL}(q(\mathbf{U}^\ell)||p(\mathbf{U}^\ell)), \tag{15}$$

which resembles the standard DSVI marginal likelihood, but, unlike in [14], the expectation is taken over the variational distribution $q(\mathbf{f}_n^L)$ constructed from the inter-domain operators.

3.1 Experiment with large-scale, non-stationary data

We tested the performance of our model on the U.S. flight delay prediction example, a large-scale regression problem that has reached a status of a standard test in GP regression due to its massive size of 5,929,413 observations ($P = 8$) and its non-stationary nature, both of which are challenging for GPs [9]. We found that a 2-layer inter-domain deep GP model was able to achieve levels of predictive performance superior to state-of-the-art shallow and deep GP models, including DSVI for deep GPs and variational Fourier features (see Table 1).

Table 1: Predictive MSEs with one standard error for U.S. flight delay prediction. Standard errors were computed by taking ten pseudo-random subsamples from the data, splitting them into training and test data, and using the same subsample splits for each model. Models: DSVI for deep GPs [14], Approximate Expectation Propagation for deep GPs (AEP DGP) [1], Stochastic Variational Inference for GPs (SVGP) [8], Variational Fourier Features for GPs (VFF) [9]. NA: computationally too expensive or, in the case of AEP DGP, computationally infeasible.
† denotes that MSEs were obtained from [9].

N	10,000	1,000,000	5,929,413
DSVI DGP (2L)	0.893 ± 0.16	0.847 ± 0.01	$0.83\pm$ NA
AEP DGP	0.933 ± 0.17	0.963 ± 0.05	NA
SVGP†	0.900 ± 0.14	0.830 ± 0.01	$0.837\pm$ NA
VFF	0.892 ± 0.15	0.823 ± 0.01	$0.826\pm$ NA
this work (2L)	$\mathbf{0.869 \pm 0.15}$	$\mathbf{0.809 \pm 0.01}$	$\mathbf{0.810\pm}$ **NA**

4 Conclusion

Deep GPs show a lot of promise, and recent advances have demonstrated their strong performance on challenging problems [2, 14]. Inter-domain deep GPs show great experimental potential and offer exciting avenues for future research in Bayesian deep learning.

Acknowledgments

I gratefully acknowledge support from the Rhodes Trust and the EPSRC-funded Center for Doctoral Training in Autonomous Intelligent Machines & Systems.

References

[1] Thang Bui, Daniel Hernandez-Lobato, Jose Hernandez-Lobato, Yingzhen Li, and Richard Turner. Deep Gaussian processes for regression using approximate expectation propagation. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1472–1481, 20–22 Jun 2016.

[2] Kurt Cutajar, Edwin V. Bonilla, Pietro Michiardi, and Maurizio Filippone. Random feature expansions for deep Gaussian processes. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 884–893, 06–11 Aug 2017.

[3] Zhenwen Dai, Andreas C. Damianou, Javier González, and Neil D. Lawrence. Variational auto-encoded deep Gaussian processes. *CoRR*, abs/1511.06455, 2015.

[4] Andreas Damianou and Neil Lawrence. Deep Gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215, 29 April 2013.

[5] Nicolas Durrande, James Hensman, Magnus Rattray, and Neil D. Lawrence. Detecting periodicities with Gaussian processes. *PeerJ Computer Science*, 4:0, 00 2016.

[6] David Duvenaud, Oren Rippel, Ryan Adams, and Zoubin Ghahramani. Avoiding pathologies in very deep networks. In *Artificial Intelligence and Statistics*, pages 202–210, 2 April 2014.

[7] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML 2016, pages 1050–1059, 2016.

[8] James Hensman, Alexander G. de G. Matthews, and Zoubin Ghahramani. Scalable Variational Gaussian Process Classification. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12*, 2015.

[9] James Hensman, Nicolas Durrande, and Arno Solin. Variational Fourier features for Gaussian processes. *arXiv preprint arXiv:1611.06740*, 2016.

[10] James Hensman and Neil D Lawrence. Nested variational compression in deep Gaussian processes. *arXiv preprint arXiv:1412.1370*, 2014.

[11] Diederik P Kingma and Max Welling. Stochastic gradient VB and the variational auto-encoder. In *Second International Conference on Learning Representations, International Conference on Learning Representations*, 2014.

[12] Miguel Lázaro-Gredilla and Aníbal R. Figueiras-Vidal. Inter-domain Gaussian processes for sparse inference using inducing features. In *Proceedings of the 22Nd International Conference on Neural Information Processing Systems*, pages 1087–1095, 2009.

[13] Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996.

[14] Hugh Salimbeni and Marc Deisenroth. Doubly stochastic variational inference for deep Gaussian processes. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017.