Doubly Stochastic Variational Inference for Deep Gaussian Processes

Matthew Ashman James Branigan Ionnis Tsetis Rui Xia

Gaussian Processes

Given noisy observations $\mathcal{D} = \{\mathbf{x}_n, y_n\}_{n=1}^N$ with Gaussian likelihood $p(y_n|f_n) = \mathcal{N}(y_n; f_n, \sigma_y^2)$, and a GP prior over functions $f(\mathbf{x})$, the complete probabilistic model is: $p(\mathbf{y}, \mathbf{f}|\mathbf{X}) = p(\mathbf{f}; \mathbf{X}) \prod_{n=1}^N p(y_n|f_n)$ where inference over test locations $\mathbf{x}^*$ is

$$f(\mathbf{x}^*)|\mathbf{y} \sim \mathcal{GP}\left(k_{\mathbf{f}^*\mathbf{f}}(K_{\mathbf{ff}} + \sigma_y^2 I)^{-1}\mathbf{y}, \right.$$
$$\left. K_{\mathbf{f}^*\mathbf{f}^*} - k_{\mathbf{f}^*\mathbf{f}}(K_{\mathbf{ff}} + \sigma_y^2 I)^{-1}k_{\mathbf{ff}^*}\right)$$

Limitations:
1. Computation is $\mathcal{O}(N^3)$. **2.** Restricted to Gaussian functionality. **3.** Difficult and time-consuming to design kernels without underlying knowledge of $\mathcal{D}$.

Our solution: DSVI DGPs.

Sparse Gaussian Processes

The VFE approach [2] introduces pseudo-points $\mathbf{u} = f(\mathbf{Z})$ and forms a lower bound to the marginal likelihood using the variational distribution $q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f}|\mathbf{u}; \mathbf{X}, \mathbf{Z})q(\mathbf{u})$

$$\log p(\mathbf{y}) \geq \mathbb{E}_{q(\mathbf{f},\mathbf{u})}\left[\log \frac{p(\mathbf{y}, \mathbf{f}, \mathbf{u})}{q(\mathbf{f}, \mathbf{u})}\right]$$
$$= \mathbb{E}_{q(\mathbf{f},\mathbf{u})}\left[\log \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{u})}{q(\mathbf{u})}\right] = \mathcal{L}_{ELBO},$$

where $q(\mathbf{u}) = \mathcal{N}(\mathbf{u}; \mathbf{m}, \mathbf{S})$. $\mathcal{L}_{ELBO}$ can be maximised with respect to variational parameters $\{\mathbf{Z}, \mathbf{m}, \mathbf{S}\}$ and kernel hyperparameters.

Doubly Stochastic VI for Deep GPs

DSVI [1] introduces pseudo-points $\{\mathbf{U}^l\}_{l=1}^L$ for each layer. The approximate posterior factorises between layers as

$$q\left(\left\{\mathbf{F}^l, \mathbf{U}^l\right\}_{l=1}^L\right) = \prod_{l=1}^L p\left(\mathbf{F}^l|\mathbf{U}^l; \mathbf{F}^{l-1}, \mathbf{Z}^{l-1}\right) q\left(\mathbf{U}^l\right).$$

Marginalising $\mathbf{U}^l$ from each layer is analytical and we have the property that the marginal $q\left(\mathbf{f}_i^l\right)$ depends only on $\mathbf{f}_i^{l-1}$. The lower bound to the log-marginal likelihood simplifies as

$$\mathcal{L} = \sum_{n=1}^N \mathbb{E}_{q(f_n^L)}\left[\log p\left(y_n|f_n^L\right)\right] - \sum_{l=1}^L \mathrm{KL}\left[q\left(\mathbf{U}^l\right) \| p\left(\mathbf{U}^l\right)\right]$$

where $q\left(f_n^L\right) = \int \prod_{l=1}^{L-1} q\left(\mathbf{f}_n^l|\mathbf{m}^l, \mathbf{S}^l; \mathbf{f}_n^{l-1}, \mathbf{Z}^{l-1}\right) d\mathbf{f}_n^l$.

The expectation is approximated using Monte Carlo, which recursively draws samples from $\hat{\mathbf{f}}_n^l \sim q\left(\mathbf{f}_n^l|\mathbf{m}^l, \mathbf{S}^l; \hat{\mathbf{f}}_n^{l-1}, \mathbf{Z}^{l-1}\right)$ using the reparameterisation trick. Scalability is achieved through minibatching the data.

Deep Gaussian Processes

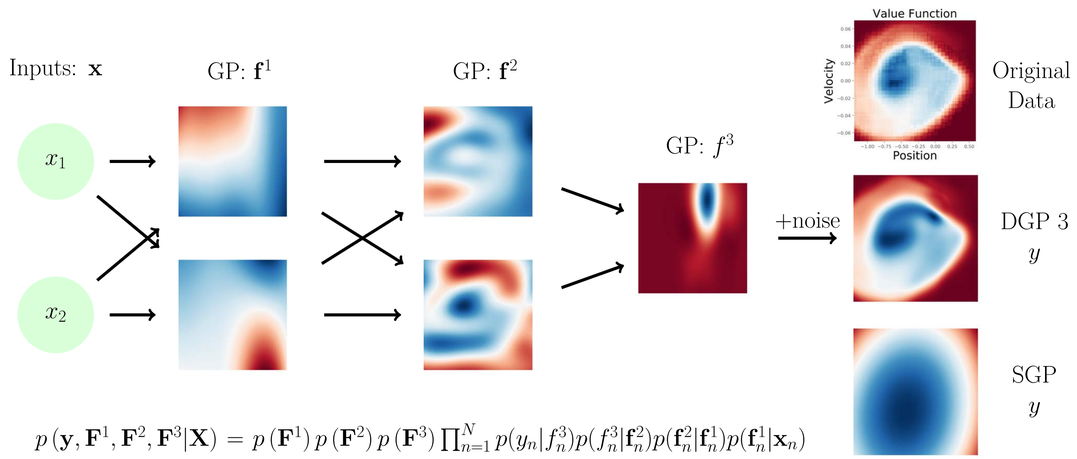

Figure: DGP and SGP modelling of the mountain car problem value function, a difficult problem for GPs due to sharp fluctuations. We see that the first layer functions, $\mathbf{f}^1$, are simple and learn to explain different parts of the input space.

Regression

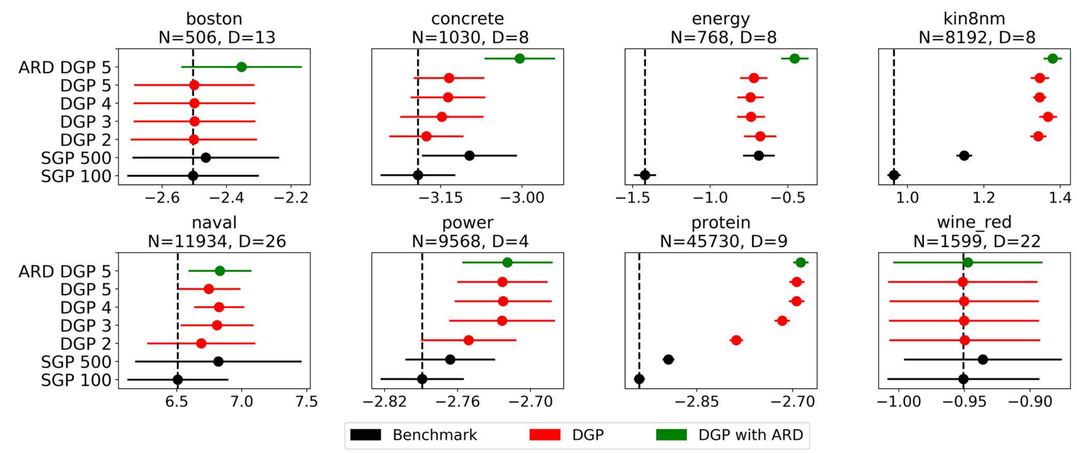

Figure: Regression test log-likelihood results on benchmark UCI datasets. The plots show the mean ± standard deviation over 20 splits.

Classification

MNIST	SGP 100	SGP 500	DGP2	DGP3	DGP3 ARD	DGP2 AEP
Log-likelihood	−0.2807	−0.2623	−0.0778	**−0.0721**	−0.0729	−0.1294
Accuracy (%)	92.26	92.88	97.89	97.98	**97.99**	96.46

Table: MNIST multi-class classification. 30 hidden layer dimensions are used in both methods.

Rectangle	SGP 100	SGP 500	DGP2	DGP3	DGP2 ARD	DGP2 AEP
Log-likelihood	−0.6575	−0.6541	−0.4817	**−0.4643**	−0.4646	−0.4815
Accuracy (%)	72.12	72.87	76.74	77.47	**77.51**	75.19

Table: Rectangles-Images binary classification.

Image Completion

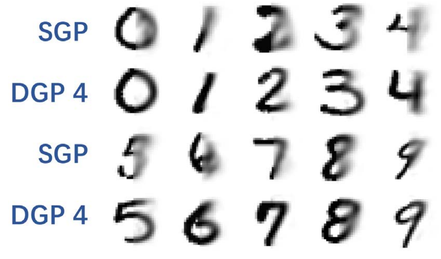

Figure: Finished image with the predicted right half.

Model	SGP	DGP 2	DGP 3	DGP 4
Metrics	0.428	0.597	0.611	0.627

Table: Structural image similarity index averaged over 10k images.

Discussion

DGPs outperform standard GPs on almost all regression, classification and image completion experiments. DGPs also produce **well calibrated uncertainty estimates** on classification tasks, whilst we found that GPs underestimate the accuracy when p is low and overestimate when p is high.

Unlike DNNs, increasing the **depth does not significantly improve DGPs' performance**. Furthermore, using **ARD kernels outperforms the original results**, and is comparable with state-of-the-art inference techniques in DGPs (with less computational cost).

However, this model cannot handle multi-modal data due to absence of posterior correlations between pseudo-points and univariate Gaussian assumption.

Future Work

1. Use of convolutional kernels. **2.** Application of DGPs in continual learning, active learning and Bayesian optimisation. **3.** Modelling correlated outputs using autoregressive DGP model. **4.** Handling multi-modal data.

References

[1] Hugh Salimbeni and Marc Deisenroth. Doubly stochastic variational inference for deep Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 4588–4599, 2017.

[2] Michalis Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009.