Deep Gaussian Processes for Large Datasets

James Hensman **Andreas Damianou** **Neil Lawrence**
Sheffield Institute for Translational Neuroscience and Dept. of Computer Science, University of Sheffield

Big data and Bayesian non-parametric modelling constitute two of the most important foci of modern machine learning research. In this preliminary work we propose a neat solution for combining the aforementioned domains into a single principled framework based on Gaussian processes. Speficically, we invistigate algorithms for training deep generative models with hidden layers connected with non-linear Gaussian process (GP) mappings. Building on recent developments on (stochastic) variational approximations, the models are fitted on massive data and the hidden variables are marginalised out in a Bayesian manner to allow for efficient propagation of the uncertainty throughout the network of variables.

Defining deep Gaussian process networks is challenging even for few data. Consider n observed datapoints $\mathbf{x}^{(0)}$ being generated from a cascade of l hidden layers of variables $\mathbf{x}^{(i)}$ connected via GP mappings $\mathbf{f}^{(i)}$, $i = 1, \dots, l$. One would like to obtain the marginal likelihood $p(\mathbf{x}^{(0)})$ by integrating out all unobserved layers. The marginalisation $\int_{\mathbf{x}^{(i)}} p(\mathbf{f}^{(i)}|\mathbf{x}^{(i)})p(\mathbf{x}^{(i)})$ associated with each layer is intractable, because $\mathbf{f}^{(i)}(\mathbf{x}^{(i)})$ is non-linear. Damianou et al. [2013] extended the variational framework of Titsias [2009] to allow the definition of a tractable lower bound on the $p(\mathbf{x}^{(0)})$ through the incorporation of m auxiliary inputs $\mathbf{z}^{(i)}$ that replace $\mathbf{x}^{(i)}$ in the intractable terms. However, even though the aforementioned variational approximation is tractable, the complexity is $\mathcal{O}(lnm^2)$, because of the coupling in the function instantiations $\mathbf{f}^{(i)}$. For standard GPs, Hensman et al. [2013] showed how this coupling can be broken by treating the function instatiations $\mathbf{u}^{(i)}$ corresponding to each input $\mathbf{z}^{(i)}$ as *global* variables. These are updated after considering small batches rather than the whole available dataset, thus extending the aforementioned variational inference in the stochastic case.

Here we start by marrying the approaches of [Hensma et al. 2013] and [Damianou et al. 2013]. The optimisation space of our models contains the global parameters $\mathbf{u}$, GP kernel parameters and variational parameters Θ_k associated with each batch k of the data. On top of this very non-convex optimisation space, it is well known that the learning rates used in the stochastic optimization greatly affect the overall training. Therefore, the final vital piece of our framework is a novel, principled way of setting the learning rate in each iteration. Specifically, we seek to find a learning rate $r^{(j)}$ at iteration j, so that the *expected* KL divergence between $p(\mathbf{u}|\Theta^{(j)})$ and $p(\mathbf{u}|\Theta^{(j+1)})$ is minimised. Then, we can solve for the optimal (in this sense) learning rate which is shown to follow an expression similar to the one found in [Ranganath et al. 2013] but also involving the gradients of the variational bound with respect to the natural parameters of the global variables.

Finally, we can define our training algorithm as follows: in each learning iteration j, we load a batch of data k and estimate the approximate posterior of all hidden variables $\mathbf{x}$ given the data, through optimisation of the variational parameters with the global parameters being held fixed. Then, the variational parameters associated with $\mathbf{x}$ are held fixed and we move r units towards the natural gradient direction of the global parameters and the gradient direction of the kernel parameters. Finally, the learnint rate r is adjusted according to our novel scheme.

In preliminary experiments we considered 20K motion capture examples. For the moment we experimented only with 1-level architectures, although extension to deep structures is only a matter of further implementation. Even in this case, our algorithm "learned" the global human motion concept better than modelling subsets of the training examples with the Bayesian GP-LVM, giving a lower error in reconstructing parts of the body from novel motions.

References

A. Damianou, N. Lawrence. Deep Gaussian processes. *AISTATS*, 2013

M. Titsias. Variational learning of inducing variables in sparse Gaussian processes. *AISTATS*, 2009

J. Hensman, N. Fusi, N. Lawrence. Gaussian processes for big data. *UAI*, 2013.

R. Ranganath, C. Wang, D. Blei, E. Xing. An Adaptive Learning Rate for Stochastic Variational Inference, *ICML*