

---

# Gaussian Process Models with Parallelization and GPU acceleration

---

Zhenwen Dai\*

Andreas Damianou\*

James Hensman\*

Neil Lawrence\*

Department of Computer Science

University of Sheffield

{z.dai, andreas.damianou, j.hensman, n.lawrence}@sheffield.ac.uk

## Abstract

In this work, we present an extension of Gaussian process (GP) models with sophisticated parallelization and GPU acceleration. The parallelization scheme arises naturally from the modular computational structure w.r.t. datapoints in the sparse Gaussian process formulation. Additionally, the computational bottleneck is implemented with GPU acceleration for further speed up. Combining both techniques allows applying Gaussian process models to millions of datapoints. The efficiency of our algorithm is demonstrated with a synthetic dataset. Its source code has been integrated into our popular software library GPy.

## 1 Introduction

Gaussian processes (GPs), as non-parametric, data driven approaches, are very popular for regression and dimension reduction problems. Their formulation is responsible for their power but also their memory and computational limitations. Considering a regression problem with observed input and output data, collected by rows in matrices  $\mathbf{Y} \in \mathbb{R}^{N \times D}$  and  $\mathbf{X} \in \mathbb{R}^{N \times Q}$  respectively, the input and output are related according to a set of latent functions  $f_d$  plus Gaussian noise:

$$y_{n,d} = f_d(\mathbf{x}_n) + \epsilon_{n,d}, \epsilon_{n,d} \sim \mathcal{N}(0, \beta^{-1}), \quad (1)$$

where  $y_{n,d}$  denotes the  $d$ th dimension of the  $n$ th output point and  $\mathbf{x}_n$  denotes the  $n$ th input point. A GP prior with covariance function  $k_f$  is placed on the noise-free observation  $f_d(\mathbf{x}_n)$ , which can later be integrated out. This will couple the data points in a  $N \times N$  covariance matrix  $\mathbf{K}_{ff} = k_f(\mathbf{X}, \mathbf{X})$ , the inversion of which scales with  $\mathcal{O}(N^3)$ . The main line of work in the literature attempting to speed up GPs is related to low rank approximations [1, 2, 3, 4, 5, 6] which decouple the function instantiations  $\mathbf{F} = f(\mathbf{X})$  given a set of  $M$  auxiliary (or *inducing*) function input-outputs pairs, denoted by  $\mathbf{z}_m$  and  $\mathbf{u}_m$  respectively, that result in a low rank approximation of  $\mathbf{K}_{ff}$ . However, even if the resulting computational cost is  $\mathcal{O}(NM^2)$ , in big data domain (millions of data points), the computational bottlenecks encountered in practice are actually associated with large  $N$ , since the number of inducing points  $M$  can be kept small.

In this paper we are interested in scaling up inference of GP models using parallel computations with respect to the data points. To achieve this, we exploit the independence assumptions induced by using auxiliary variables as well as the full independence assumption in  $p(y_{n,d} | \mathbf{f}_d(\mathbf{x}_n))$  of the noise model of equation (1). In practice, this means that all operations involving data points can be written as sums over  $N$ . This observation has recently been exploited by Hensman et al. [7] for **stochastic variational inference** and by Gal et al. [8] for **distributed variational inference**. Here we adopt a similar formulation, but focus on presenting a distributed implementation which combines large-scale parallelization (with MPI) and GPU acceleration. It is transparently embedded for sparse GP based models in GPy, such as Bayesian GP-LVM [9, 10], MRD [11] and deep GPs [12]. With

---

\*also at Sheffield Institute for Translational Neuroscience, SITraN.

experiments on synthetic data, we show that our inference algorithm can efficiently make use of hundreds of computer nodes, which allows for the consideration of genuine big data.

## 2 Distributed inference in GP models

The sparse inference method employed in GPy follows the variational formulation of [6], which constructs a variational lower bound to the true log. marginal likelihood of the GP using inducing point representations. Specifically, we have,  $\log p(\mathbf{Y}|\mathbf{X}) \geq \sum_{d=1}^D \mathcal{F}_d$ , with:

$$\begin{aligned} \mathcal{F}_d = & \log \frac{\beta^{N/2} |\mathbf{K}_{uu}|^{1/2}}{(2\pi)^{N/2} |\mathbf{K}_{uu} + \beta \Phi|^{1/2}} - \frac{\beta}{2} \mathbf{y}_d^\top \mathbf{y}_d \\ & + \frac{\beta^2}{2} \mathbf{y}_d^\top \mathbf{K}_{fu} (\beta \Phi + \mathbf{K}_{uu})^{-1} \mathbf{K}_{fu}^\top \mathbf{y}_d - \frac{\beta \phi}{2} + \frac{\beta}{2} \text{tr}(\mathbf{K}_{uu}^{-1} \Phi), \end{aligned} \quad (2)$$

where  $\phi = \text{tr}(\mathbf{K}_{ff})$  and  $\Phi = \mathbf{K}_{fu}^\top \mathbf{K}_{fu}$ . Here,  $\mathbf{K}_{fu}$  contains the cross-covariances between the training and inducing inputs ( $\mathbf{X}$  and  $\mathbf{Z}$ ) and  $\mathbf{K}_{uu}$  is the matrix of (co)variances between points in  $\mathbf{Z}$ .

As can be seen, the variational bound is factorised with respect to data dimensions. However, for many practical applications we would like parallelization with respect to the number of data,  $N$ . Towards this direction, we can use the fact that the covariance matrices can be calculated in a decomposable way, as mentioned in [9]. Specifically, we can write  $\phi = \sum_{n=1}^N k_f(\mathbf{x}_n, \mathbf{x}_n)$  and  $\Phi = \sum_{n=1}^N (\mathbf{K}_{fu})_n^\top (\mathbf{K}_{fu})_n$ , where  $(\mathbf{K}_{fu})_n$  is an  $M$ -dimensional vector with its  $m$ th element given by  $k_f(\mathbf{x}_n, \mathbf{z}_m)$ . By additionally introducing  $\Psi = \sum_{n=1}^N (\mathbf{K}_{fu})_n^\top \mathbf{y}_n$ , we can re-write the full variational bound (for all dimensions) as  $\log p(\mathbf{Y}|\mathbf{X}) \geq \mathcal{F}$ , with:

$$\begin{aligned} \mathcal{F} = & D \log \frac{\beta^{N/2} |\mathbf{K}_{uu}|^{1/2}}{(2\pi)^{N/2} |\mathbf{K}_{uu} + \beta \Phi|^{1/2}} - \frac{\beta}{2} \sum_{n=1}^N \mathbf{y}_n \mathbf{y}_n^\top \\ & + \frac{\beta^2}{2} \Psi^\top (\beta \Phi + \mathbf{K}_{uu})^{-1} \Psi - \frac{\beta D \phi}{2} + \frac{\beta D}{2} \text{tr}(\mathbf{K}_{uu}^{-1} \Phi). \end{aligned} \quad (3)$$

Notice that the above bound is not *fully* factorised, due to the term  $(\beta \Phi + \mathbf{K}_{uu})^{-1}$ . However, that term is an  $M \times M$  matrix which is cheap to invert for the usual choices of  $M$ , whereas all the other expensive computations associated with data points can be parallelised. In the unsupervised version of this model, namely the Bayesian GP-LVM,  $\mathbf{X}$  is treated as a *latent variable* and is also integrated out, after introducing a prior distribution  $p(\mathbf{X})$  and a variational posterior distribution  $q(\mathbf{X})$ . Both of these distributions are Gaussian and factorised with respect to  $N$ . Then, similarly to the sparse GP regression case, we can define a variational lower bound on  $\log p(\mathbf{Y})$  where:

$$\log p(\mathbf{Y}) \geq \langle \mathcal{F} \rangle_{q(\mathbf{X})} - \sum_{n=1}^N \int q(\mathbf{x}_n) \log \frac{p(\mathbf{x}_n)}{q(\mathbf{x}_n)} d\mathbf{x}_n. \quad (4)$$

Due to the factorisation of  $q(\mathbf{X})$  w.r.t. datapoints, the summations over  $N$  are maintained in the unsupervised version. Specifically, the only difference between  $\mathcal{F}$  and  $\langle \mathcal{F} \rangle_{q(\mathbf{X})}$  is that  $\phi$ ,  $\Psi$  and  $\Phi$  are turned into expectations over  $q(\mathbf{X})$ , e.g.  $\phi = \sum_{n=1}^N \int_{\mathbf{x}_n} k_f(\mathbf{x}_n, \mathbf{x}_n) q(\mathbf{x}_n)$  and similarly for  $\Phi$  and  $\Psi$ . Compared to the approach taken in [7], the formulation presented here is not fully factorised but results in a tighter bound and, additionally, extends to the unsupervised case straightforwardly. This formulation is similar to the distributed variational inference proposed by Gal et al. [8]. Using the above variational bound as an objective function means that we have the following parameters to optimise: the kernel parameters  $\theta$ , the noise parameter  $\beta$  and the inducing inputs  $\mathbf{Z}$ . For the rest of our analysis we assume the more general unsupervised scenario, which also involves the parameters of  $\{q(\mathbf{x}_n)\}_{n=1}^N$ , namely  $\{\mu_n, \mathbf{S}_n\}_{n=1}^N$ .

As mentioned above, all the computations w.r.t. datapoints are combined according to a couple of summations. Therefore, data parallelism is naturally applicable, so that all datapoints are distributed across computer nodes and every node only performs the computation w.r.t. its local portion of data, i.e. the computation of  $\phi$ ,  $\Phi$ ,  $\Psi$  (and also the second term of equation (4) in the Bayesian GP-LVM case). Although all the computations w.r.t. datapoints are parallelizable, to recover the exact lower bound, a few steps of indistributable computations have to be done after collecting the intermediate

---

Table 1: The GPU function for computing  $\Psi$  or  $\Phi$ 

```

1: for each inducing input  $m$  (each pair  $(m_1, m_2)$  for  $\Phi$ ) do                                ▷ distributed across GPU blocks
2:   for each datapoint  $n$  do                                                                ▷ distributed across GPU threads
3:     Compute  $\Psi_{nm}$  or  $\Phi_{m_1 m_2}^{(n)}$ .
4:     Write  $\Psi_{nm}$  into global GPU memory, when computing  $\Psi$ .
5:     Write  $\Phi_{m_1 m_2}^{(n)}$  into shared local memory, when computing  $\Phi$ .
6:   end for
7:   Sum  $\Phi_{m_1 m_2}^{(n)}$  from all threads and write into global memory, when computing  $\Phi$ .
8: end for

```

Table 2: Computing the gradients that depend on  $\Psi$  or  $\Phi$ 

**Require:**  $\frac{\partial L}{\partial \Psi}$  or  $\frac{\partial L}{\partial \Phi}$

```

1: for each input dimension  $q$  do
2:   Set  $(\frac{\partial L}{\partial Z})_{mq} = 0$ .
3:   for each datapoint  $n$  do                                                                ▷ distributed across GPU threads
4:     for each inducing input  $m$  (each pair  $(m_1, m_2)$  for  $\Phi$ ) do                            ▷ distributed across GPU blocks
5:       Compute  $\frac{\partial \Psi_{nm}}{\partial \mu_{nq}}, \frac{\partial \Psi_{nm}}{\partial S_{nq}}, \frac{\partial \Psi_{nm}}{\partial Z_{mq}}, \frac{\partial \Psi_{nm}}{\partial \theta_q}$ , or  $\frac{\partial \Phi_{m_1 m_2}^{(n)}}{\partial \mu_{nq}}, \frac{\partial \Phi_{m_1 m_2}^{(n)}}{\partial S_{nq}}, \frac{\partial \Phi_{m_1 m_2}^{(n)}}{\partial Z_{m_1 q}}, \frac{\partial \Phi_{m_1 m_2}^{(n)}}{\partial \theta_q}$ .
6:       Compute the derivatives of  $L$  w.r.t. parameters by combining  $\frac{\partial L}{\partial \Psi}$  or  $\frac{\partial L}{\partial \Phi}$ .
7:       Add  $(\frac{\partial L}{\partial Z})_{nmq}$  into  $(\frac{\partial L}{\partial Z})_{mq}$ .
8:     end for
9:     Sum the intermediate results for  $(\frac{\partial L}{\partial \mu})_{nq}$  and  $(\frac{\partial L}{\partial S})_{nq}$ .
10:   end for
11: end for
12: Sum the intermediate results for  $\frac{\partial L}{\partial \theta}$ .

```

---

results from individual nodes, e.g.,  $(\beta\Phi + \mathbf{K}_{uu})^{-1}$  in particular. After getting the lower bound, the derivatives w.r.t. model parameters can be estimated locally. In principle, with a distributed optimizer, all the local parameters can be determined locally, while the global parameters such as kernel parameters can be determined by synchronizing their gradients among all the computer nodes. However, to make use of the existing optimizers like L-BFGS-B in Scipy, we currently collect the gradients of both local and global parameters into one node, estimate the new parameters according to the chosen optimizer, and spread the new parameters across the rest computer nodes.

### 3 GPU acceleration

The scheme mentioned above dramatically scales up computations by exploiting data parallelism. However, we notice that for some models like the Bayesian GP-LVM, the quantities  $\Phi$  and  $\Psi$  constitute the computational bottleneck. **Computing these quantities requires to go through all datapoints multiple times, something which can take more than 99% of inference time for large datasets.** To further speed up the inference, we make use of GPU acceleration. GPU is a type of specialized computation hardware, which consists of a large number of small processing units (e.g., GTX TITAN black has about 2,880 cores). It is extremely efficient at solving a large number of similar small tasks, which is ideal for our inference algorithm. Therefore, we take advantage of the specialized architecture by shifting the computation of  $\Phi$  and  $\Psi$  onto GPU. The key for making full use of its computational power is to properly divide computational workloads for parallelization within GPU. Due to the specific architecture of GPU, there are a couple of constraints for programming, e.g., **different computing blocks are not synchronized within a GPU function**<sup>1</sup>, and **synchronization of writing GPU global memory is very expensive**. The optimal division of workloads depends on the specific type of data, e.g. the size difference among  $N$ ,  $M$ ,  $D$  and  $Q$ . Our implementation tries to balance between different choices and gives a generic design, which is suitable for most cases.

In particular, we assign each computing block of GPU a subset of inducing inputs for computing  $\Psi$  and a subset of inducing input pairs for computing  $\Phi$ . We assign each thread within a computing

---

<sup>1</sup>More sophisticated synchronization schemes are supported with new cards, but, to support a wide range of GPU cards, we currently stick to compute capability 2.0.

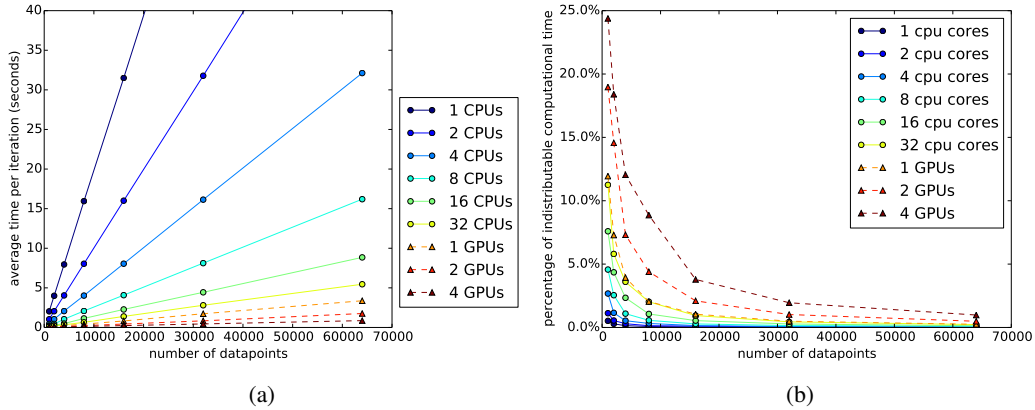


Figure 1: (a) The scaling of our parallel inference with both CPUs and GPUs. It shows the average time per iteration for different sizes of datasets. (b) The percentage of time used for indistributable computation per iteration.

block a subset of datapoints. With this division, for computing  $\Psi$ , each thread can write its result directly into the GPU memory, and for  $\Phi$  the intermediate results w.r.t. datapoints are stored in shared local memory; these results are later summed together across different threads and written into the GPU memory (see Table 1). For computing the gradients w.r.t. parameters that depend on  $\Psi$  and  $\Phi$ , the same division of workloads is applied. Due to the different sizes of individual parameters, their gradients are produced at different levels of loops (see Table 2). On top of GPU acceleration, the parallelism scheme mentioned in the previous section can be easily integrated. The computation of the lower bound can be distributed across multiple GPU cards by assigning each GPU card a subset of datapoints. Then, the computation of  $\Phi$  and  $\Psi$  for the local portion is shifted onto individual GPU cards, and the local results are combined as mentioned before.

## 4 Experiments

The performance of our parallelization and GPU acceleration algorithms are evaluated through a synthetic dataset. The dataset was generated by randomly sampling 64k 1D datapoints, and mapping into 3D space by sampling according to a Radius Basis Function (RBF) kernel function. The task is to recover the 1D latent representations from these 3D datapoints by applying the Bayesian GP-LVM as a dimension reduction algorithm. Therefore, the dimensionality of latent space is set to be one, and the number of inducing inputs is set to be 100. A list of datasets with different sizes, ranging from 1k to 64k, are used with a list of different parallelization configurations. All the experiments ran on a single machine with 4 AMD Opteron processors (32 cores in total) and 4 NVIDIA GTX 480 GPU cards, but can directly run on clusters without any changes on the code. The efficiency of our parallel inference algorithm is measured by the average computational time per iteration (shown in Fig. 1a). The computational time scales linearly with the number of datapoints, which is expected from the computational complexity, and the speed of inference scales roughly linearly with the number of CPUs/GPUs, which shows the efficiency of our parallel implementation, in which the communication overhead is negligible. Note that the speed with a single GPU card is significantly higher than a 32-core computer node. Additionally, the percentage of time used for indistributable computation is shown in Fig. 1b. It shows that most of time is spent on distributable computation, which means that with more computer resources the speed can further increase.

## 5 Conclusion

We have presented an efficient parallelised implementation of Gaussian process models and, additionally, the first algorithm with GPU acceleration in this domain. Our results constitute a counter argument against the prejudice that GPs are not applicable to big data, as long as efficient implementations are considered. As future work, we plan to scale up other GP-based models already implemented in GPy and compare our results to methods that currently dominate the domain of big data, like deep learning.

## Acknowledgments

We acknowledge funding by the RADIANT and WYSIWYD (EU FP7-ICT) projects. JH was funded by an MRC fellowship.

## References

- [1] L. Csató and M. Opper, “Sparse on-line Gaussian processes,” *Neural Computation*, vol. 14, no. 3, pp. 641–668, 2002.
- [2] M. Seeger, C. K. I. Williams, and N. D. Lawrence, “Fast forward selection to speed up sparse Gaussian process regression,” in *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics* (C. M. Bishop and B. J. Frey, eds.), (Key West, FL), 3–6 Jan 2003.
- [3] E. Snelson and Z. Ghahramani, “Sparse Gaussian processes using pseudo-inputs,” in *Advances in Neural Information Processing Systems* (Y. Weiss, B. Schölkopf, and J. C. Platt, eds.), vol. 18, (Cambridge, MA), MIT Press, 2006.
- [4] J. Quiñero Candela and C. E. Rasmussen, “A unifying view of sparse approximate Gaussian process regression,” *Journal of Machine Learning Research*, vol. 6, pp. 1939–1959, 2005.
- [5] N. D. Lawrence, M. Seeger, and R. Herbrich, “Fast sparse Gaussian process methods: The informative vector machine,” in *Advances in Neural Information Processing Systems* (S. Becker, S. Thrun, and K. Obermayer, eds.), vol. 15, (Cambridge, MA), pp. 625–632, MIT Press, 2003.
- [6] M. K. Titsias, “Variational learning of inducing variables in sparse Gaussian processes,” in *Proceedings of the Twelfth International Workshop on Artificial Intelligence and Statistics* (D. van Dyk and M. Welling, eds.), vol. 5, (Clearwater Beach, FL), pp. 567–574, JMLR W&CP 5, 16-18 April 2009.
- [7] J. Hensman, N. Fusi, and N. D. Lawrence, “Gaussian processes for big data,” in *Uncertainty in Artificial Intelligence* (A. Nicholson and P. Smyth, eds.), vol. 29, AUAI Press, 2013.
- [8] Y. Gal, M. van der Wilk, and C. E. Rasmussen, “Distributed variational inference in sparse Gaussian process regression and latent variable models,” in *Advances in Neural Information Processing Systems* (Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds.), vol. 27, (Cambridge, MA), 2014.
- [9] M. K. Titsias and N. D. Lawrence, “Bayesian Gaussian process latent variable model,” in *Proceedings of the Thirteenth International Workshop on Artificial Intelligence and Statistics* (Y. W. Teh and D. M. Titterton, eds.), vol. 9, (Chia Laguna Resort, Sardinia, Italy), pp. 844–851, JMLR W&CP 9, 13-16 May 2010.
- [10] A. C. Damianou, M. K. Titsias, and N. D. Lawrence, “Variational inference for uncertainty on the inputs of gaussian process models,” *arXiv preprint arXiv:1409.2287*, 2014.
- [11] A. Damianou, C. H. Ek, M. K. Titsias, and N. D. Lawrence, “Manifold relevance determination,” in *Proceedings of the International Conference in Machine Learning* (J. Langford and J. Pineau, eds.), vol. 29, (San Francisco, CA), Morgan Kaufman, 2012.
- [12] A. Damianou and N. D. Lawrence, “Deep Gaussian processes,” in *Proceedings of the Sixteenth International Workshop on Artificial Intelligence and Statistics* (C. Carvalho and P. Ravikumar, eds.), vol. 31, (AZ, USA), JMLR W&CP 31, 2013.