Optimising Earth observation parameter retrieval system

Submission to the RSD Call for Projects

Dr J L Gómez-Dans, (RA, Dept. of Geography)[‡]
Prof P Lewis (Professor, Dept. of Geography)

February 2015

1 Abstract

Earth Observation (EO) data provides global coverage of the land surface on a timely and spatially detailed manner. The data need to be interpreted using physical models that are computationally expensive. This results in an inverse problem that is solved either using quasi-Newton minimisation methods or other data assimilation (DA) approaches (Kalman filters, particle filters, ...). In either approach, the physical model needs to be evaluated for many times, typically within a poorly-conditioned iterative algorithm. This approach impractical for large data volumes coming from satellites. To improve on this, we have started using Gaussian process (GP) emulators: these are surrogate functions that produce the same output as the complex model, but in a fraction of the time. Current efforts have resulted in a $\sim 10^4\times$ speed increase in model runtime (including calculating the partial derivatives of the model, an important requirement when using gradient descent optimisation methods). However, although the surrogates are now faster, iterative gradient descent minimisation algorithms and the very large dimensionality and poor conditioning of the problems still hinder the use of these advanced systems with the very large EO archives. We believe that parallel architectures like GPUs show promise to improve the current limitations of DA systems, not only in evaluating the emulators, but also in bringing in advanced minimisation or pre-conditioning techniques that exploit parallel architectures. Improving on this would potentially lead to the uptake of these techniques and software by e.g. space agencies.

[*]Principal Investigator
[†]Corresponding author, j.gomez-dans@ucl.ac.uk,Tel. (0)20 767 90590

2 Introduction

2.1 Introduction to research area

EO data are critical to monitor the land surface within a context of climate change. However, EO data are indirect measurements of land surface state, and this requires physical models that describe the physics of radiation absorption and scattering by the land surface that give rise to the observations. These models need to be "inverted" to match the observations, resulting in an inverse problem that is hard to tackle numerically. The inversion results in the parameters of the physical model that would make said model fit the observations, with quantified uncertainty.

DA techniques solve the previous problem, and are widely used in e.g. numerical weather prediction (NWP). Our group has been working on this area for some time, funded by the European Space Agency (ESA) and the NERC National Centre for Earth Observation (NCEO). We have used so-called variational methods, which rely on the minimisation of a compound cost function by (typically) quasi-Newton methods. These approaches have necessitated the use of automatic differentiation tools, that coupled with computationally expensive numerical models, have lead only to proof-of-concept studies [1, 2], using the EO-LDAS Python toolbox.

We now use "emulators" as surrogate models of the physical models that are the bottleneck of the current system. These are in effect regressions between model input and output pairs that allow us to approximate both the value of the complex (very non-linear) physical model and its gradient (providing an estimate of the approximation uncertainty) from a limited set of full model runs. Gaussian processes (GPs) have been widely used for this task[3]. They are convenient because of their simplicity and good performance with models we use. They are also fast to train and extremely fast in prediction mode. However, for practical applications, even the speeds obtained with GPs are insufficient to cope with large data amounts. This is coupled to the problem that the evaluation of the GPs is embedded in an iterative minimisation algorithm. We think that both the GP implementation and the cost minimisation strategies should be reviewed to account for GPU or other massively parallel architectures. We have already developed a Python GP implementation code and a streamlined version of EO-LDAS package (see here).

2.2 Introduction to research group

The Environmental Modelling and Observation Group in the Department of Geography has a strong background in computational problems applied to remote sensing, from ray tracing models to complex DA schemes and canopy reconstruction. The authors of this proposal have solid understanding of Python, C, Fortran and Bash, and have developed large projects using these languages, as well as common free software libraries such as GDAL, GSL, Lapack, FFTW, etc. At the Department, we both teach a Master's level course on scientific computing, focused on Python and geospatial data. We mostly use github (see here and here for our repositories). We have published about this research earlier (see [1, 2]), and are currently working on a paper on the use of GP emulators for these problems.

2.3 Introduction to code to be worked on

The main bottlenecks are the scaling of the an iterative minimisation problem driven by GPs to massive data spaces (of dimensions $> 10^8$), where conditioning is generally poor, and we believe that (i) parallel evaluation of the GPs and (ii) parallel optimisation algorithms will be useful in this regard. Parallelisation of the GP code[1] is likely to be simple for an expert, although a solution that bridges both parts of the problem would be better. The code is Python with the Numpy and Scipy libraries as dependencies. The code introduces a Gaussian Process class, which does the GP "training" (fitting hyperparameters to a limited set of full model runs input/output pairs), as well as a `predict` method that actually evaluates the model approximation and the partial derivatives. This method should be able to approximate the emulated model (and associated gradient) for a very large number of input vectors, and since these evaluations are all independent, we think it might be good to focus on this part of the code. The code has been tested on its own (by emulating the full model for a large number of randomly chosen input vectors) and also as part of a complete rewrite of the original EO-LDAS codebase, `eoldas_ng`, with very satisfactory results. In most cases, the emulation error can be safely ignored in the inversions, as it is one or two orders of magnitude less than the observational error.

3 Suggested Objectives for The Project

We believe that the approach presented in [1] is sound. However, to fully deliver on the promise of consistent land surface parameter estimates form all available sensors, the bottleneck of inefficient minimisation of surrogate models needs to be addressed. As mentioned in Section 2.3, we envisage two areas where vast improvements are likely to happen:

1. Parallel evaluation of the GP predictions for very large numbers of input vectors using e.g. GPUs.
2. Minimisation algorithms on parallel architectures efficient for a large number of dimensions.
3. Pre-conditioning strategies to redeuce the number of iterations on iterative minimisation schemes.

4 Impact of the project

4.1 Potential for use of software beyond originating group ans sustainability

Within NCEO, the `eoldas_ng` tool is seen as strategic in bridging the gap between land surface and climate modellers and the EO community, and will be providing opportunities to train upcoming PhD students and others in using this tool. The CEMS/JASMIN infrastructure is accessible and widely used by NCEO-funded scientists, and we envisage that given the interest in this approach from the European Space Agency, a practical way to process the vast data archives held in CEMS would be a major opportunity for impact in many communities. Being

[1] An working version of the code is available from github. We have a newer, faster and better documented version, but not on github yet.

able to process large amounts of data is a core commitment we have towards NCEO, and part of our national capability providing DA tools for EO. Finally, this would provide members of our group an understanding of parallel architectures, which would have a bearing in other research activities we do, such as raytracing or canopy reconstruction. We would seek collaboration from NCEO and ESA to continue funding developments on this project.

5 Justification for application

5.1 Justification for use of RSDT staff

Currently, the staff working on this project are working in other areas of the problem, and lack the expertise to implement these changes themselves. Although one part of the proposal is probably fairly simple to implement by an expert in parallel systems, we think that useful advice on the overall problem of minimising a cost function in a highly dimensional space would be very valuable, and this is expertise that we currently do not have, and that is not available within the land EO community.

5.2 Justification for use of free project

Having this new implementation would open up new opportunities to use the software on big data archives, responding to calls from ESA and the EC, in particular in the light of the Sentinels (a set of operational EO satellites) and COPERNICUS initiatives (the new European EO programme), where we think the DA approach we have developed will be perfectly suited, if only we can deal with large data problems efficiently, which is what this project will try to address. We would be asking for support from NCEO to use the EMERALD GPU cluster in conjunction with the data archives present in the CEMS system, both located at Rutherford Appleton Labs.

References

[1] Lewis, P., J. Gómez-Dans, T. Kaminski, Jeffrey Settle, Tristan Quaife, N. Gobron, J. Styles, and M. Berger. "An earth observation land data assimilation system (EO-LDAS)." Remote Sensing of Environment 120 (2012): 219-235.

[2] Lewis, P., J. Gómez-Dans, T. Kaminski, J. Settle, T. Quaife, N. Gobron, J. Styles, and M. Berger. "Data assimilation of Sentinel-2 observations: preliminary results from EO-LDAS and outlook." In Proc."First Sentinel-2 Preparatory Symposium", Frascati, Italy, pp. 23-27. 2012.

[3] O'Hagan, Anthony. "Bayesian analysis of computer code outputs: a tutorial." Reliability Engineering & System Safety 91, no. 10 (2006): 1290-1300.