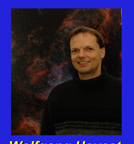
A Sampler for multidimensional parameter space, integrated in ProC

http://planck.mpa-garching.mpg.de

DLR MPAC



Wolfgang Hovest

The Sampler Team at MPA:

- Gordana Stojceska
- Torsten Ensslin
- Wolfgang Hovest
- Ilya Saverchenko
- Thomas Riller
- Uwe Doerl
- Hans-Martin Adorf

A thorough statistical analysis of astrophysical obsrvations often requires searching and sampling huge parameter spaces. A case in point is the data of the future ESA Planck satellite that will measure the cosmic microwave background (CMB) in order to estimate fundamental parameters of our universe, such as the densities of gas, dark matter and dark energy.

We organize the data analysis process using a scientific workflow engine, the Planck Process Coordinator. In order to tackle the sampling problem a generic control element, called a Sampler, was developed and applied to the study of the initial conditions for merging galaxies.

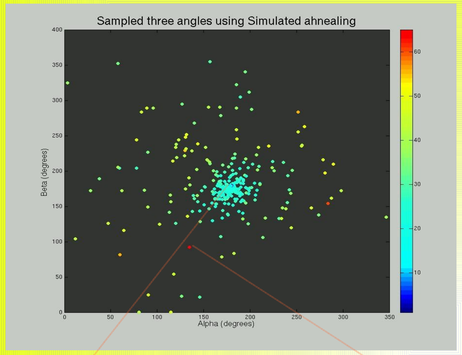
THE IDEA AND MOTIVATION

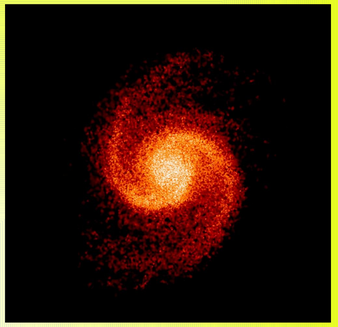
Our goal is to address inverse problems as one finds in astrophysics, physics, and engineering. In such problems one knows how to compute (i.e. through a computational expensive workflow) observables from a set of initial parameters, however one would like to infer initial parameters from observations. Therefore one has to search and sample the parameter space with efficient or intelligent algorithms, in order to identify and map parameter space regions reproducing the observations.

We describe the ongoing implementation of a generic sampler control element in the ProC workflow engine, which will allow to attack such inverse problems with a variety of algorithms ranging from grid searches, over Markov-Chain Monte Carlo sampling, to genetic algorithms. This sampler will be easily extendable by the plug in of additional algorithms written in different programming languages.

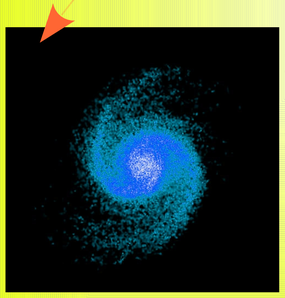
Moreover, this software tool should support exploration of the given parameter spaces on a fast and efficient way.

Sampling three projection angles

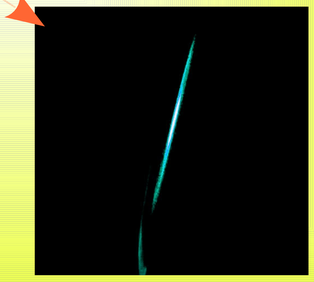




The target Galaxy (2D-picture from observations)



Here is the galaxy which was produced by the simulation for the best sampled parameters. It is very similar to the target one.



Here is the galaxy produced by the simulation for the worst sampled parameters. It is not similar at all to the target one.

ABOUT THE SAMPLER

The Sampler is a new Control Element for the ProC, which supports exploration of parameter spaces by supplying varying input parameters to
a subpipeline and evaluating its output to determine future parameter sets.

The idea for development and implementation of the *Sampler* was given as a Master Thesis topic to Gordana Stojceska, student of "Computational Sciences and Engineering" Master Programm at the Technische Universitaet Muenchen. Together with other Planck scientists and developers, she designed, developed and integrated the Sampler into ProC. The Sampler should succesfully sample parameter values that later on will be used to compare the simulated results with results got from the observations.

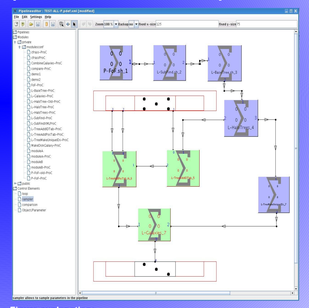


Figure explanation:
A new ProC-Layout where the sampler contorl element is included. The boxes with green and purple color represent the modules. The Sampler Control element influences the execution of the green modules.

PROC INFRASTRUCTURE

The ProC allows assembly of data processing pipelines from individual modules (written in Fortran, C, C++ or Java, IDL is projected), combined with high level control elements (e.g. loops) and subpipelines. Since these subpipelines are themselves pipelines, complex pipelines can be built in a hierarchical manner.

The ProC executes the pipelines by using the Grid Application Toolkit
(GAT), either in a local cluster through scheduling systems like PBS or Sun Grid Engine, or in the Grid through the Globus Toolkit or Unicore.In addition to the main results, all parameters, all provisional results of the single modules of the pipeline, and any information on the processing itself will be traced and saved in a Data Management Component (DMC), which is being developed side by side with the ProC.

The Pipeline is the main building component in ProC. It can be built from subpipeline(s), module(s) and control elements.
The integrated Sampler plays the role of control element. It should be applied to a subpipeline, so that it will sample chosen parameters and then give back the sampled values to the subpipeline.

CASE STUDY: GALAXY COLLISION

Is it possible to see how an observed system of two or more interacting galaxies develops over time? To answer the question we use GADGET - a cosmological simulation code and a set of additional tools. These programs are combined together to form a workflow. We work with ProC because it provides a friendly environment for workflow definition and efficient execution.
A sampler control element, which is a part of ProC, includes a variety of sampling algorithms such as Markov chain Monte Carlo or Genetic Algorithms. The element is able to efficiently sample high dimensional parameter space. By iterating over the samples we are able to accurately estimate initial parameters for numerical simulation of a system of several interacting galaxies similar to the one we can observe.

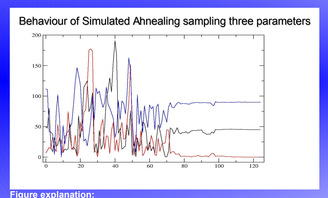


Figure explanation:
Y-axis: angles between 0-180 degrees, X-axis: number of steps
Each projection angle has different colour: red, blue and black.
The sampled angle values are known after 120 steps of the algorithm (in this case simulated ahnealing)

THE PLANCK MISSION

Planck is the third Medium-Sized Mission (M3) of ESA's Horizon-2000 Scientific Programme. Planned for 2007, Planck will obtain full-sky maps in nine frequency bands in the microwave regime between 30 and 857 GHz. The primary goal of Planck is to map the Cosmic Microwave Background (CMB) with unprecedented resolution and sensitivity. The accurately measured angular power spectrum of the CMB will allow the precise determination of all relevant cosmological parameters. Planck will also test the inflationary model of the early universe.

Planck is being planned, developed and built by a European/North American Consortium under the direction of ESA. The Max-Planck-Institut für Astrophysik (MPA) represents Germany in that consortium. One of the tasks of the MPA is the development of the *Process Coordinator (ProC)* for the automatic operation of data-simulation and -analysis pipelines. The ProC allows to build up pipelines from individual modules (written in FORTRAN, C, C++ or Java) and execute them in a distributed heterogenous network in an automatic fashion. Besides the main results all parameters, all provisional results of the single modules of the pipeline and any information on the processing itself will be traced and saved in an object-oriented database, or alternatively in FITS files. More about the ProC one can find under: http://planck.mpa-garching.mpg.de.