PhD project on
Data Intensive Artificial Intelligence (DIAI)

Simulation-based cosmological inference of large galaxy surveys

2021 - 2024

Project Description

Context

The upcoming generation of wide-field optical cosmological galaxy surveys (for example the European space mission Euclid, the Rubin Observatory LSST, the Roman space telescope) aim to shed some much needed light on the physical nature of dark energy and dark matter, by mapping the Universe in great detail and on an unprecedented scale. However, as the data volumes become orders of magnitude larger compared to current experiments, the cosmological analysis of these surveys becomes increasingly subtle, and standard analysis techniques based on analytically modeling the measured cosmological signals are reaching significant limits. To probe models of modified gravity, and to account for complex systematic effects, large parameter spaces need to be explored.

The standard paradigm for cosmological inference relies on Bayesian inference performed with Markov Chain Monte-Carlo approaches with explicit analytic likelihoods. This approach becomes impractical as the cosmological and observational models become more complex and no longer tractable analytically.

The goal of this PhD thesis is to develop novel methodologies to analyse weak-lensing data observables from the UNIONS/CFIS and Euclid surveys, based on the radically different approach of simulating the survey data. The data volume of these surveys is in the petabyte regime, providing images of hundreds of millions of galaxies observed over thousands of degrees sky area.

The promise of a **simulation-based inference** approach is to circumvent the difficulties faced by analytic models, enabling an optimal analysis of weak-lensing data. We propose to develop a generic methodology enabling end-to-end Bayesian inference in situations where only a black-box simulator is available, which includes dimensionality reduction, active sampling strategies, and density estimation.

Objectives

1. Most cosmological analyses are currently based on traditional Bayesian sampling of an analytical likelihood function. Such a function however does in general only very approximately describe the non-Gaussian distributed data, and to model complex systematic effects analytically can be insurmountably difficult. To overcome these difficulties, the first axis of this PhD is to develop ***simulation-efficient*** neural density estimators and to apply them to this likelihood-free inference problem. While most of the deep-learning literature has been focused on developing density estimators (generative models) in the large-dataset regime, far less effort has gone towards building regularized density estimators that perform well in the small-dataset regime (in our case a small number of simulations). In this project, the student will explore various approaches for both active learning and regularization of the neural density estimators, with the goal of achieving an optimal convergence rate of the density estimation task as a function of number of simulations.
2. To account for the very high dimensionality of typical observables, data compression strategies have to be employed. In this second axis of the PhD project, we will explore several neural compression methods that will aim at training a deep neural network to extract a small set of sufficient statistics for the ultimate cosmological inference task. A few approaches have been proposed in the litterature, based on training summaries to maximize the Fisher information, but this approach can be unstable and is not suited to multi-modal posteriors. We will explore and compare approaches based on mutual information lower bounds and contrastive learning results. One particular aspect unique to physical inference problems that we will particularly focus on, is to build so-called ***nuisance-hardened*** summary statistics, that preserve the cosmological information while being insensitive/robust to unknown nuisance/systematic errors present in the data. Beyond our cosmological inference problem, such compression methods should find useful applications for any problems requiring a minimal and robust representation of high-dimensional data.

3. The novel methods developed here will be applied to existing weak-lensing simulations and data. The PhD candidate will have access to the UNIONS/CFIS survey, which is processed by CosmoStat. They will also use large volumes of Euclid simulations, as an important step to validate the new methods. Finally, cosmological inference of Euclid data, available in 2024, is expected to significantly contribute to the scientific analysis of this large upcoming experiment. The focus will be to discriminate standard cosmology from alternative models, potentially **modifying the gravitational force** and the lensing potential.

The likelihood-free inference methods developed by the student here will employ the latest deep-learning methods, and be implemented in python Jax. This project will provide a

user-friendly interface and documentation, and the resulting package will be made open source and hosted on the CosmoStat github organisation. The student will have access to various GPU clusters available in France to test the performance and scalability of the tools to the volume of data expected from upcoming cosmological surveys.

Work environment

This thesis will be carried out in the CosmoStat laboratory at the Département d'Astrophysique (DAp) at CEA Paris-Saclay, under the supervision of Dr. Martin Kilbinger, Dr. Valeria Pettorino, and Dr. François Lanusse. The CosmoStat group gathers experts in astrophysics, signal processing and data science, to work on challenging problems in cosmology such as weak gravitational lensing, cosmological inference, and machine learning. Dr. Martin Kilbinger has been a pioneer in exploring alternative inference techniques for weak lensing. Dr. Valeria Pettorino is a leading scientist in analysing cosmological data sets, and an expert in models of dark energy and modified gravity. Dr. Francois Lanusse is an expert in deep-learning techniques ranging from density estimation to image processing for cosmology. This PhD project will build on the expertise within CosmoStat on applying machine-learning and Bayesian methods to cosmological data.

Timeliness

CosmoStat has major responsibilities in the weak-lensing data processing and science analysis of the upcoming ESA Euclid space satellite. Planned for launch in 2022/2023, Euclid will deliver its first data one year later, during the third year of this PhD, in time for the methods developed during this project. CosmoStat is leading the processing of the state-of-the-art weak-lensing survey UNIONS/CFIS, for which data will be available at the start of this PhD, as a test bed for Euclid.