# SBI lab report

Mohammed Farhan Hassan, Janis Koehler & Morgane Krauth

## 1. Physics context

The early Universe is a big mystery captivating cosmologists to this day in many aspects. Naturally the initial conditions of the Universe are of great relevance to any cosmological theory. The most widely accepted theory indicates these initial conditions are drawn from a Gaussian random field (GRF) with a nearly scale-invariant power spectrum. The Cosmic Microwave Background maps temperature fluctuations in the early universe and has been measured to be a nearly perfect GRF. Although the universe's current density field departs more and more from gaussianity, as gravitational effects introduce nonlinearities, this provides a useful opportunity to study a technique called simulation-based inference (SBI). SBI uses simulation procedures to create data, from which machine learning algorithms are able to learn inferring crucial parameters used in the simulation process. These trained algorithms can later be used to infer the parameters of the actual data, which is obtained from measurements.

To this end, we simulated multiple instances of GRFs and used this data to train and test a network detecting the two parameters of the power law governing the data generation process. Although inferring parameters of the early universe would require additional complications, due to the inherent nonlinearity of the evolving density field, this provides a useful toy example to compare analytical and SBI based approaches.

## 2. Methods and results

In this lab, we were introduced to SBI, the practicalities of calculating the likelihood function of the GRF, and aimed to compare the analytical results from using a maximum likelihood estimate to those yielded from SBI.

Using the likelihood and a prior density, the posterior probability density function (PDF) can be calculated. It summarizes the scientist's state of knowledge of the model's parameter values in posterity of the data. Given the direct mathematical link between the likelihood and the posterior, any model is expected to fit either of those equally well. Here we chose to model the posterior distribution.

As opposed to analytical methods that require writing down the likelihood or posterior to start with, the SBI method allows us to infer parameters by choosing those that maximize the likelihood function or the posterior without explicitly specifying them. In this lab, we used a GRF to use as data to train and test our network. We create the random field by providing the desired power spectrum of the data as a power law depending on two parameters, where the actual field in fourier space is then randomly distributed around these frequency powers.. We then use a Fourier transform to compute the real space representation of the data and thus to visualize the field.

After creating the field, we created an autoencoder network and trained it on the data to reconstruct the field. The use of an autoencoder results in reducing the number of network parameters. This leads to faster training and enhances generalization potential. As can be seen from the training loss over time, the model yields good results after a relatively low number of epochs. Of course, reducing the number of parameters and dimensions could potentially reduce the capability of learning deeper subtleties of the structure, as the resolution decreases. We did not investigate and balance the aforementioned competing effects due to the given time constraints.

We then passed the encoded representation of the data through a normalizing flow, which, using successive fully connected layers, breaks down the probability distribution of the data to a simple distribution. This allowed us to infer useful information from the data without making any assumptions on the form of the likelihood. We did not quite have the time to get the normalizing flow to work. Our progress is attached as code in the git of this group.

## 3. Conclusion and outlook

In the end, we did not get to implement SBI fully due to time constraints so all we can say is that we applied the analytical method successfully to the

data. Of course, it would be interesting to carry out SBI and compare the results. Likewise, more techniques could be used instead of or paired with SBI to improve the outcome, like principal component analysis (PCA) for example, which would reduce the dimensionality, making it easier to work with the data. Moreover, we could have experimented with different types of autoencoders and normalizing flows. Finally, this lab was a simplified example of the actual cosmological evolution. More challenges, but also more opportunities would arise from creating more sophisticated simulation techniques and applying those to real cosmological data. The process could be improved by modifying the autoencoder and normalizing flow architecture, encoding strategy, simulation process and introducing additional machine learning techniques.