UNIVERSITY
OF OSLO

Master's thesis

-sometitle-

Classifying N-body simulations with and without relativistic corrections
using machine learning techniques

Johan Mylius Kroken

Computational Science: Astrophysics
60 ECTS study points

Institute of Theoretical Astrophysics
Faculty of Mathematics and Natural Sciences

Autumn 2023

Johan Mylius Kroken

-sometitle-

Classifying N-body simulations with and without relativistic corrections using machine learning techniques

Supervisors:
A David Fonseca Mota
B Julian Adamek
C Francisco Antonio Villaescusa Navarro

Abstract

On large scales, comparable to the horizon, relativistic effects will affect the cosmological observables. In order to solve for these effects, one need to consistently solve for the metric, velocities and densities in a particular gauge. When simulating large-scale structures we use N-body simulations, which are usually performed in the Newtonian limit. However, it is not obvious that Newtonian gravity yield a good global description of an inhomogeneous cosmology across all scales (Jeong, Schmidt and Hirata 2012). However, literature suggest that Newtonian simulations are still solving the dynamics correctly, even on large scales close to the horizon where relativistic effects are important but may be corrected for (Chisari and Zaldarriaga 2011) (Green and Wald 2012).

Recently, Adamek et al. 2016 developed a relativistic N-body code, `gevolution`, which evolves large scales structures based on the weak field expansion in GR. I investigate the differences in the gravitational dynamics between structures evolved with and without relativistic effects, with focus on the gravitational potential Φ. This is a good choice for comparison as Φ is gauge invariant and the Newtonian and relativistic simulations are performed in different gauges.

The investigation is done by running 2000 simulations using identical ΛCDM cosmologies for the two gravity theories. The simulations are run using 64^4 particles on a 256^3 grid each with dimension 5120 Mpc/h which is a compromise in order to include both large and nonlinear scales. The data analysis consist of a preliminary analysis using conventional summary statistics, with focus on the bispectrum of Φ. There is a difference in the two cases for low redshifts in the equilateral and squeezed configurations. However, the main idea is to train a Convolutional Neural Network (CNN) to classify the two cases, given snapshots of Φ. The main analysis then involves interpretability of the CNN, which may be done by considering for instance saliency maps (Alqaraawi et al. 2020) or Grad-CAM (Selvaraju et al. 2020). In either case, revealing the features separating the two cases may help us understand the differences in the gravitational dynamics between the two theories. I expect that such a network is able to find relativistic corrections to the Newtonian snapshots that are of higher order than those obtained from power spectra and bispectra analysis. Further, it may also reveal which configurations of Fourier modes $\mathbf{k}$ yield the highest bispectral power, which for now is mainly trial and error.

Contents

List of Figures

List of Tables

Preface

Here comes your preface, including acknowledgments and thanks.

Part II

Machine Learning

Chapter 6

Fundamental Elements of Machine Learning

6.1 Introduction

In this chapter I will give a brief introduction into machine learning. This includes a mathematical description of some fundamental concepts common across numerous machine learning models. The more advanced models will be dealt with at a later stage. If not otherwise stated, the following chapter is based on Goodfellow, Bengio and Courville 2016 and Hastie, Tibshirani and Friedman 2009.

6.2 Linear Algebra

maybe

6.3 Probability and Information Theory

maybe

6.4 Basic Machine Learning

TODO: Fill more here

6.4.1 Optimisation and Generalisation

Optimisation Optimsation problems are problems in which we want to minimise some error, given some data. In other words, we want to optimise an algorithm or model given a specific dataset. We care about the error of our model for that specific dataset only.

Generalisation The concept of *generalisation* is what makes a machine learning model different from an optimisation model. We still *train* the machine learning model on some specific data, but we measure how good the model is based on how it performs on a different set of data, which it was not trained on. I.e. we need a generalised model, which is not restricted to the data it was trained on, and has the ability to perform well on unobserved data.

6.4.2 Data and Fitting

Data The key ingredient to any machine learning algorithm is the data fed into it. In an optimisation problem we deal with one single set of data on which we train the model. In a machine learning scenario we want to quantify how well the model is generalised. This is done by training the model on part of the data only, called the *training data*. The model is then assessed on the remaining data, on which it was not trained, in order to determine how general it is. I will come back to how the models are trained, but in essence they are learning features of the data they are trained on. Thus, in order to have a general model we need the data to have some inherent properties. The most important property is that the data is *independent and identically distribution* (i.i.d.). In other words, both the training and testing data are drawn from the same probability distribution.

Training and testing When training the model, we want to minimise some error with respect to the training data, which is essentially an optimisation problem. The model is then tested by measuring the same error with respect to the testing date. It is the latter error we want to be low in order to call the model general. The discrepancy between the two are summaries by the concept of fitting.

Overfitting Overfitting is when we train the model in such a way that the training error becomes too low. This optimises the model too much with respect to the training data, effectively reconstructing the training data point by point and thereby loosing trends in the data features. The result of this is poor performance on the testing data, and as a result a non-general model.

Underfitting Underfitting is the opposite, when the model is unable to achieve a low error on the training data, not capturing the main features of the dataset. This is also bad, as the training error would also be large. This also results in a non-general model.

TODO: Figure explaining over- and underfitting

6.4.3 Estimators, Bias, Variance and Error

Estimators Based on the assumption that there exists some true parameter(s) $\boldsymbol{\theta}$ which remain unknown,[1] we are able to make predictions and estimations of such parameter(s). Let's say we have m independent and identically distributed (i.i.d.) random variables $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ drawn from the same probability distribution $p(\mathbf{x})$. An *estimator* of the true values $\boldsymbol{\theta}$ is any function of the data such that $\hat{\boldsymbol{\theta}}_m = g(\mathbf{x}_1, \ldots, \mathbf{x}_m)$, where $\hat{\boldsymbol{\theta}}$ is the estimate of $\boldsymbol{\theta}$. This is known as point estimation, as we are estimating a single value. This definition does not pose any restrictions on the function g. However, a good estimator would yield an estimate $\hat{\boldsymbol{\theta}}_m$ that is close to the true value $\boldsymbol{\theta}$.

Function estimators Say we want to predict a variable $\mathbf{y}$ given some vector $\mathbf{x}$. We assume the true variable $\mathbf{y}$ is given by some function approxiamation $f(\mathbf{x})$ plus some error $\boldsymbol{\epsilon}$: $\mathbf{y} = f(\mathbf{x}) + \boldsymbol{\epsilon}$. The aim is then to estimate the function f with the estimator $\hat{f}$. If we then realise that $\hat{f}$ is really just a point estimator in function space, the two above concepts are equivalent.

[1] This is the frequentist perspective of statistics

Bias The bias of the estimator $\hat{\boldsymbol{\theta}}_m$ is defined as the difference between the expected value of the estimator and the true value of the parameter: $\text{Bias}\left[\hat{\boldsymbol{\theta}}_m\right] = \mathbb{E}\left[\hat{\boldsymbol{\theta}}_m\right] - \boldsymbol{\theta}$. An unbiased estimator has zero bias, i.e. $\mathbb{E}\left[\hat{\boldsymbol{\theta}}_m\right] = \boldsymbol{\theta}$. An estimator is asymptotically unbiased if its bias approaches zero as the number of data points m approaches infinity, i.e. $\lim_{m\to\infty} \mathbb{E}\left[\hat{\boldsymbol{\theta}}_m\right] = \boldsymbol{\theta}$.

Variance Variance serves as a crucial metric in assessing the variability of an estimator concerning changes in the input data. Denoted as $\text{Var}\left[\hat{\theta}\right]$, where $\hat{\theta}$ represents the training set, low variance is desirable in an estimator. In essence, variance quantifies the extent to which we anticipate the estimator to fluctuate based on different datasets.

Standard Error The standard error, denoted as $\text{SE}\left[\hat{\theta}\right]$, is a fundamental concept closely related to variance, being the square root of the latter. Specifically, for the estimator of the mean, $\hat{\mu}$, the standard error is defined by the formula:

$$\text{SE}\left[\hat{\mu}\right] = \sqrt{\text{Var}\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right]} = \frac{\sigma}{\sqrt{m}}, \tag{6.1}$$

Here, σ^2 represents the true variance. The standard error of $\hat{\mu}$ is pivotal in gauging the precision of the estimator for the mean, especially when applied to testing data, aiding in the evaluation of a model's generalization performance.

Mean Squared Error The Mean Squared Error (MSE) provides a comprehensive measure for evaluating the performance of an estimator. Expressed as:

$$\text{MSE} = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] = \text{Bias}\left[\hat{\theta}\right]^2 + \text{Var}\left[\hat{\theta}\right], \tag{6.2}$$

MSE encompasses both the bias and variance components. The MSE offers a unified perspective, capturing the overall accuracy and variability of the estimator in estimating the true parameter θ.

6.4.4 Maximum Likelihood Estimation

6.4.5 Bayesian Statistics

6.4.6 Supervised Learning

6.4.7 Unsupervised Learning

Chapter 7

Neural Networks

7.1 Forward pass - Prediction

7.1.1 Architecture

The simplest architecture of a neural network is the fully connected feed forward (FCNN), which consist of L layers in total, the first being an input layer and the remaining $L-1$ layers are called *hidden layers*, $\mathbf{h}$. Each hidden layer has an *activation* $\mathbf{a}$, used as an input to an *activation function*, g. Mathematically we may write this architecture as:

$$\begin{aligned} \mathbf{h}^0 &= \mathbf{x}^{(i)} \\ \mathbf{h}^1 &= g_1(\mathbf{a}^1) \\ \mathbf{h}^2 &= g_2(\mathbf{a}^2) \\ &\vdots \\ \mathbf{h}^L &= g_L(\mathbf{a}^L), \end{aligned} \tag{7.1}$$

where $\mathbf{h}^L$ and g_L are the output layer and output function respectively. The parameter L governs the depth of the neural network.

7.1.2 Activation

The activation $\mathbf{a}^l$ of a layer l is an affine transformation of the output of the previous layer, $\mathbf{h}^{l-1}$. The intercept of this affine transformation is known as the bias $\mathbf{b}^l$, [1] typically used to ensure that no activation becomes zero. The activation takes the form:

$$\mathbf{a}^l = (\mathbf{W}^{l-1\to l})^{\mathrm{T}}\mathbf{h}^{l-1} + \mathbf{b}^l, \tag{7.2}$$

where $\mathbf{W}^{l-1\to l} \in \mathbb{R}^{\dim_{l-1}\times\dim_l}$ is the matrix of weights describing the mapping from layer $l-1$ to layer l. Each layer l has dimension (or neurons) $\dim_l$ which governs the width of each layer.

[1] Must not be confused with the statistical bias of an estimator.

7.1.3 Activation functions

7.1.4 Loss functions

7.2 Backpropagation - Training

7.2.1 Chain rule

Basics Lets say we have a function $f(x)$ that is composed of two functions $g(x)$ and $h(x)$, such that $f(x) = g(h(x))$. The chain rule states that the derivative of $f(x)$ with respect to x is the product of the derivative of $g(x)$ with respect to $h(x)$ and the derivative of $h(x)$ with respect to x:

$$\frac{df}{dx} = \frac{dg}{dh}\frac{dh}{dx}. \tag{7.3}$$

This can be generalised into the vector case where $f(\mathbf{x}) = g(h(\mathbf{x}))$ and $\mathbf{x} \in \mathbb{R}^n$ and TODO: write more ahaha

7.2.2 Gradient descent

7.2.3 Optimizers

7.2.4 Regularization

Chapter 8

Convolutional Neural Networks

8.1 Convolution

8.2 New Layers

8.2.1 Convolutional layers

8.2.2 Pooling layers

8.2.3 Dropout layers

Bibliography

Adamek, Julian et al. (29th July 2016). 'gevolution: a cosmological N-body code based on General Relativity'. In: *Journal of Cosmology and Astroparticle Physics* 2016.7, pp. 053–053. ISSN: 1475-7516. DOI: 10.1088/1475-7516/2016/07/053. arXiv: 1604.06065[astro-ph,physics:gr-qc,physics:physics]. URL: http://arxiv.org/abs/1604.06065 (visited on 23/08/2023).

Alqaraawi, Ahmed et al. (3rd Feb. 2020). *Evaluating Saliency Map Explanations for Convolutional Neural Networks: A User Study*. arXiv: 2002.00772[cs]. URL: http://arxiv.org/abs/2002.00772 (visited on 23/10/2023).

Chisari, Nora Elisa and Matias Zaldarriaga (2nd June 2011). 'Connection between Newtonian simulations and general relativity'. In: *Physical Review D* 83.12, p. 123505. ISSN: 1550-7998, 1550-2368. DOI: 10.1103/PhysRevD.83.123505. arXiv: 1101.3555[astro-ph,physics:gr-qc]. URL: http://arxiv.org/abs/1101.3555 (visited on 23/08/2023).

Goodfellow, Ian, Yoshua Bengio and Aaron Courville (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press.

Green, Stephen R. and Robert M. Wald (15th Mar. 2012). 'Newtonian and Relativistic Cosmologies'. In: *Physical Review D* 85.6, p. 063512. ISSN: 1550-7998, 1550-2368. DOI: 10.1103/PhysRevD.85.063512. arXiv: 1111.2997[astro-ph,physics:gr-qc]. URL: http://arxiv.org/abs/1111.2997 (visited on 23/08/2023).

Hastie, T., R. Tibshirani and J.H. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer. ISBN: 9780387848846. URL: https://books.google.no/books?id=eBSgoAEACAAJ.

Jeong, Donghui, Fabian Schmidt and Christopher M. Hirata (4th Jan. 2012). 'Large-scale clustering of galaxies in general relativity'. In: *Physical Review D* 85.2, p. 023504. ISSN: 1550-7998, 1550-2368. DOI: 10.1103/PhysRevD.85.023504. arXiv: 1107.5427[astro-ph]. URL: http://arxiv.org/abs/1107.5427 (visited on 23/08/2023).

Selvaraju, Ramprasaath R. et al. (Feb. 2020). 'Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization'. In: *International Journal of Computer Vision* 128.2, pp. 336–359. ISSN: 0920-5691, 1573-1405. DOI: 10.1007/s11263-019-01228-7. arXiv: 1610.02391[cs]. URL: http://arxiv.org/abs/1610.02391 (visited on 23/10/2023).