Progress Report: Learning Cosmological Observables with Deep Neural Networks

Jamie Martin

November 2024

1 Overview

Analysis on the cosmological observables (e.g. power spectrum from the Cosmic Microwave Background radiation) requires the observables to be calculated from a variety of cosmological constants (i.e. in multiple dimensions). However numerical algorithms are unable to interpolate efficiently in more than a few dimensions, making the analysis time consuming. The inverse problem of finding cosmological constants from observables is crucial for analysis. Using regular interpolation methods this used to take days to solve the inverse problem.

Deep learning techniques are now established as a solution to the low efficiency of interpolation in high dimensions, computing the power spectrum in less than 50ms compared to several seconds using numerical methods (e.g. CAMB or CLASS). This can be used to solve the inverse problem and calculate cosmological constants from power spectra in a few minutes, dramatically reducing time in analysis of spectra and allowing it to be performed on a laptop. [1]

Emulators have been developed recently (e.g. CosmoPower) but these architectures have not been optimised yet. This goal of this project is to improve upon the previous techniques by using a systematic hyperparameter search and studying testing training strategies, focussing specifically on the power spectrum from the CMB radation. Once the most efficient architecture of the deep learning algorithm has been determined and implemented it will be incorporated into the Large Language Model for CAMB. [1]

2 Physics

The structure of our universe is largely due to gravitational instabilities, where matter collects slightly more or less in different regions of the universe. Despite the small initial differences in density ($\mathcal{O}(10^{-4})$) matter eventually accumulated in clumps in the universe forming structures such as galaxies. [4]

As well as the attractive force of gravity, there are two other repulsive effects: the expansion of the universe and pressure from baryons (normal matter) and photons. The expansion of the universe pulls all particles apart from each other

so in areas of faster expansion structures grow slower and in non-expanding space the growth of the structure will increase exponentially (assuming no pressure). Baryons and photons exert pressure and gas moves in the direction of lower pressure which slows down the growth of a structure.[4]

There are three main stages of evolution of the cosmological perturbations: early, when all modes are outside the horizon (radiation dominated), intermediate, when wavelengths start to enter the horizon (radiation domination transitions to matter domination) and late time, when most modes are sub-horizon (matter dominated). [4]

The power spectrum describes the distribution of amplitude of fluctuations at different length scales. It is normally plotted against $k = \frac{2\pi}{\lambda}$ for different times (z's), with $z = 0$ being now and $z > 0$ going in the past. It is calculated by applying a transfer function and growth factor again and again to the primordial curvature perturbation to "take the perturbations through time". There are two types of power spectrum that can be calculated: the linear and non-linear for when fluctuations are small or large, respectively. The equation for the linear power spectrum is given by: [4]

$$P_L(k,a) = \frac{8\pi^2}{25}\frac{A_s}{\Omega_m^2}D_+^2(a)T^2(k)\frac{k^{n_s}}{H_0^4 k_p^{n_s-1}} \tag{1}$$

A_s is the amplitude or primordial scalar perturbations (quantifies the strength of density fluctuations that grew into large-scale structures). Ω_m is the density parameter for matter, including dark matter and baryonic (normal) matter, H_0 is the current Hubble constant which gives the rate of expansion of the universe at this time. k_p is the pivot wavenumber, i.e. the scale the power spectrum is normalised against and n_s is the scalar spectral index which indicates how initial density perturbations are distributed across different scales ($n_s = 1$ means all scales contribute equally to perturbations. [4]

From this equation it is clear to see the two functions that we need to calculate are the transfer function ($T(k)$) and the growth factor ($D_+(a)$). These can be calculated analytically or numerically, the analytical solution stems from solving the Boltzmann equations for matter and an expression for gravitational potential energy within the sub-horizon regime (small-scales) using appropriate approximations. The equations we want to solve are: [4]

$$\delta_c' + iku_c = -3\Phi', \tag{2}$$

$$u_c' + \frac{a'}{a}u_c = ik\Phi, \tag{3}$$

$$k^2\Phi = 4\pi G a^2\left[\rho_c\delta_c + 4\rho_r\Theta_{r,0} + \frac{3aH}{k}\left(i\rho_c u_c + 4\rho_r\Theta_{r,1}\right)\right]. \tag{4}$$

Where δ_c is the density perturbation in the cold dark matter, u_c is the velocity perturbation of cold dark matter, Φ is the gravitational potential energy, a(t) is a measure of the universe's history (0 and the Big Band and 1 at the current epoch) ρ_c is the cold dark matter density, ρ_r is the energy density of radiation

and $\Theta_{r,0}$ and $\Theta_{r,1}$ are temperature fluctuations in large scales and small scales, respectively. The last two are assumed to be 0 in the matter dominated regime and the small scale leads to the last term in (4) going to 0. (2) and (3) can be turned into a second order equation and simplified using (4) to make the Meszaros equation: [4]

$$\frac{d^2\delta_c}{dy^2} + \frac{2+3y}{2y(y+1)}\frac{d\delta_c}{dy} - \frac{3}{2y(y+1)}\delta_c = 0. \tag{5}$$

Where $y = \frac{a}{a_{eq}} = \frac{\rho_c}{\rho_r}$ (a_{eq} is a when the universe transition from radiation dominated to matter dominated).This equation can be solved to find 2 solutions for δ_c, therefore:

$$\delta_c(k, y) = C_1 D_+(y) + C_2 D_-(y) \tag{6}$$

D_+ is the growth factor (only valid for $a \lesssim 0.1$). C_1 and C_2 can be found, but in the limit $a \gg a_{eq}$ C_2 can be approximated as 0. Using the equation:

$$\delta_m(k, a) = \frac{2}{5}\frac{k^2}{\Omega_m H_0^2}\mathcal{R}(k)T(k)D_+(a) \tag{7}$$

The transfer can be calculated can be calculated to be:

$$T(k) = 12.0\frac{k_{eq}^2}{k^2}\ln 0.12\frac{k}{k_{eq}} \tag{8}$$

This is valid in the limit of $k \gg k_{eq}$ (k_{eq} is the wavenumber of the universe at the epoch of matter-radiation equality). This is the derivation of the transfer function in specific limits. The growth factor generally must be calculated numerically unless in very specific scenarios. Generally both the transfer function and the growth factor are solved numerically using code such as CAMB, however this takes a lot of time and power to do for many sets of parameters. [4]

3 What's Been Done Previously

The crucial work done previously in this field is [] which used the emulation framework CosmoPower to construct and release neural network emulators of cosmological observables, including the matter power spectrum. The P(k) emulators are accurate to better than 0.5% out to $k = Mpc^{-1}$, which is sufficient for Stage-III data analysis (e.g. parameter estimation or testing theoretical models).

The paper uses emulators for the CMB temperature, polarization power spectra, matter power spectrum, distance-redshift relation, baryon acoustic oscillation and redshift-space distortion, however this project will focus purely on the matter power spectrum. The emulators made and trained for the matter

power spectrum are called PKL and PKNL for the linear and non-linear matter power spectra, respectively.

Six different parameters were used to build the emulators:

- ω_b, the physical baryon density;
- $\omega_c dm$, the physical dark matter density;
- H_0, the Hubble parameter today;
- τ, the reionization optical depth;
- n_s, the scalar spectral index;
- $\ln 10^{10} A_s$; where A_s is the amplitude of primordial scalar perturbations.

Parameter	Min. Value	Max. Value
$ln10^{10}As$	2.5	3.5
$\Omega_{cmb}h^2$	0.08	0.10
$\Omega_b h^2$	0.01933	0.02533
H_0 [km/s/Mpc]	39.99	100.01
n_s	0.8812(0.8)	1.0492(1.2)
τ	0.02	0.12
z	0	5

Table 1: Parameter ranges used to generate the Latin hypercube of cosmological parameters used to compute the training and test data. [1]

The ranges for all parameters are shown in Table 1. A Latin hypercube (LHC) of the parameter space was generated with $N_S = 128,000$ as the number of samples, with 80% for training and 20% for training. For the matter power spectrum another input was added: the redshift at which the power spectra are to be calculated, which is sampled between 0 and 5. The matter power spectrum of each set of parameters was then calculated from $k_{min} = 10^{-4} Mpc^{-1}$ to $k_{max} = 50 Mpc^{-1}$ with 500 points.

The relative difference between the linear and non-linear power spectra with respect to CAMB in percent was show for three redshifts: $z = 0, 2.5, 5$. It was calculated for three different methods of calculation: the emulator, high-precision CLASS and CLASS ultra-high precision. Focusing on the linear spectrum, the emulator agreed with the high precision CLASS predictions to within 0.05% across the whole k range and 0.3% with the ultra-high CLASS predictions. The emulator is most accurate at low k then deviates, with high precision CLASS, at large k. Ultra-high precision is consistently more accurate than both methods (compared to CAMB). Accuracy generally increases as z increases. The difference in spectra has a similar shape for all z values, with a dip at k_{eq} then decreasing after than, slowly at first then quickly at large k, as shown in figure .... For non-linear power spectra the difference between CAMB and COSMOPOWER is more significant, rising to 1.5% for z=5.

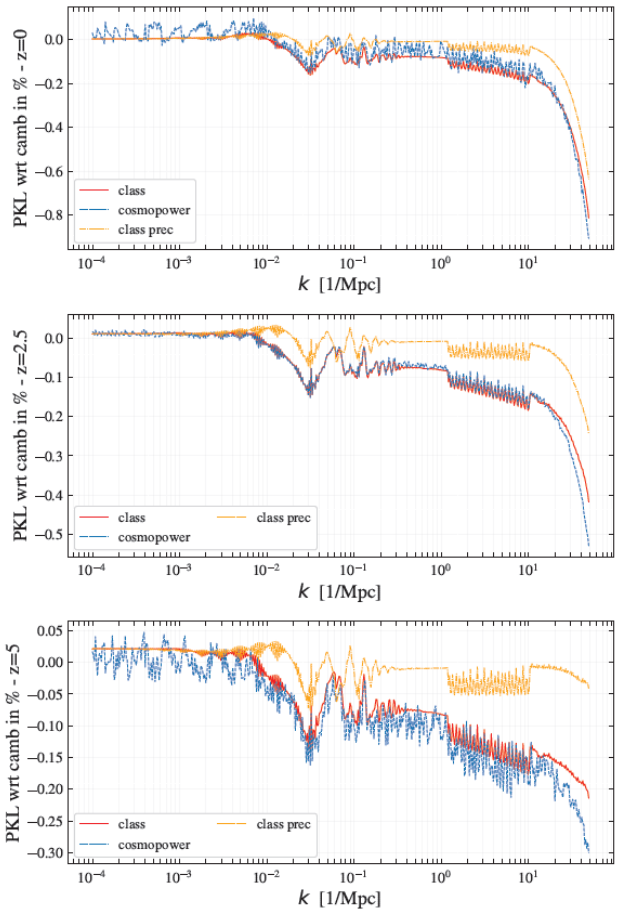

Figure 1: Relative difference of the linear matter power spectra between two different settings of CLASS and COSMOPOWER with respect to CAMB in percent at different redshifts. [1]

The emulator was then compared against the high-precision CLASS prediction fro various parameters. In each case all but one parameters were kept constant, with the one being varied and calculating different confidence levels for the error rate (indicated by the shaded regions). The non-linear matter power spectrum was the least accurately emulated observable with the linear power spectrum getting to 0.40% difference compared to the "true" CLASS value.

COSMOPOWER provides two ways of emulating power spectra: cosmopower_NN and cosmopower_PCAplusNN. Cosmopower_NN directly maps cosmological parameters to log-power spectra, with 7 inputs and 500 outputs (one for each k value of the power spectrum). The cosmopower_PCAplusNN network maps cosmological parameters to coefficients of a Principal Component Analysis (PCA)

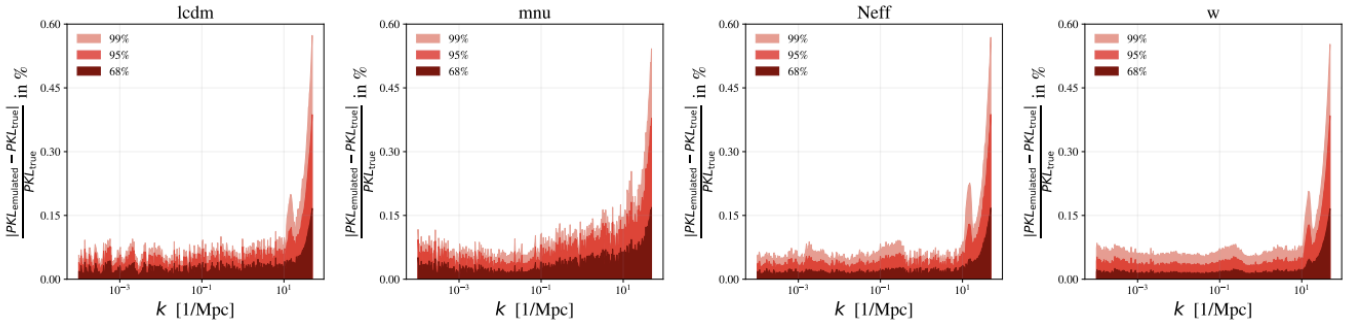

Figure 2: Relative difference between COSMOPOWER and the "true" high-precision CLASS prediction for the linear matter power spectra in percent, keeping one parameter constant in each graph. [1]

of the log-power spectra, creating a function which is the sum of PCA functions of different weights but this is not as accurate the cosmopower_NN. [3] [4]

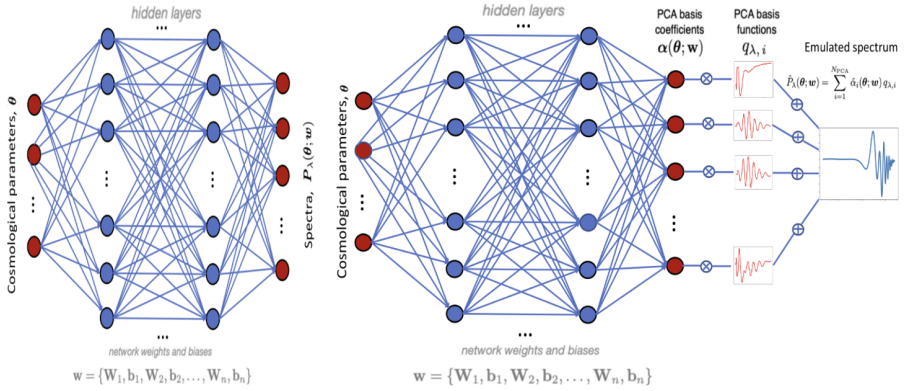

Figure 3: Visual representations of cosmopower_NN (left) and cosmopower_PCAplusNN (right) neural networks. [4]

4 Methods to Improve Performance

There are two main ways the performance can be optimised: performing a systematic hyperparameter search and investigating different training strategies.

The hyperparameter search includes various parts of the network, such as the number of middle layers; number of nodes in each layer; and choice of activation function (swish for this project). These can be searched using Optuna [6]; hyperparameter optimisation software that works by using a define-by-run approach to dynamically construct and explore search spaces. It optimises the

objective function, which evaluates model performance for a given set of hyperparameters by employing algorithms such as Tree-structure Parzen Estimator or a random search. It can stop early on unpromising trials, handle conditional parameters and provide tools for visualisation of results.

The architecture will be compared to the cosmopower_NN network as that is the more accurate, which we will call the z-method. This neural network turns 7 inputs into 500 outputs, one for each point on the spectrum, however if multiple spectra are needed with different redshift but same cosmological parameters otherwise this network must be run for however many different redshifts are needed (e.g. 100), which takes a long time. A method of improving this architecture is by generating a 2D grid of spectra with wavenumber on one axis and redshift on the other and flattening this 2D array into a 1D array so the spectra of each redshift are sequentially stored, this is called the grid-method. This is a neural network with 7 inputs going to 500*100 outputs (for 100 redshifts), which is a comparison of a smaller neural network run 100 times or a larger neural network run once. Another method to improve accuracy is to use numerical methods to determine most of the power spectrum and then use the neural network to fill in the parts that are more difficult to approximate (e.g. around k_{eq}).

We can then use these emulators for application in Simulation Based Inference, or Symbolic Regression. Simulation Based Inference, in this context, is using neural networks to infer cosmological parameters from observed power spectra (i.e. solving the inverse problem to the original). COSMOPOWER uses the neural networks to accelerate Bayesian inference by developing Bayesian inference pipelines using machine learning. Symbolic Regression discovrs interpretable mathematical expressions that approximate the relationship between cosmological parameters are CMB matter power spectra, so neural network emulators aren't relied on as black-boxes. [5]

The plan for my project is shown in the figure 4 below, with exact dates shown in table 2:

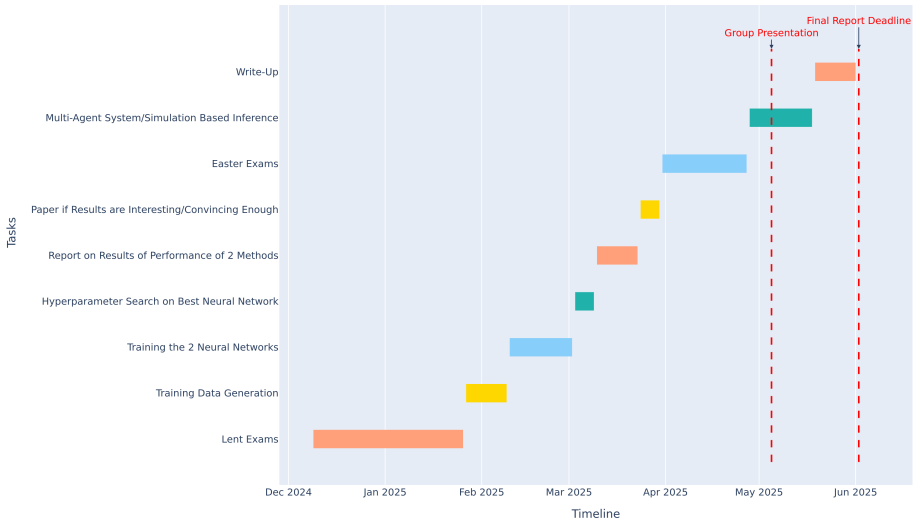

Figure 4: Gantt chart showing the plan and important dates for my project.

Task	**Start Date**	**Finish Date**
Lent Exams	2024-12-09	2025-01-26
Training Data Generation	2025-01-27	2025-02-09
Training the 2 Neural Networks	2025-02-10	2025-03-02
Hyperparameter Search on Best Neural Network	2025-03-03	2025-03-09
Report on Results of Performance of 2 Methods	2025-03-10	2025-03-23
Paper if Results are Interesting/Convincing Enough	2025-03-24	2025-03-30
Easter Exams	2025-03-31	2025-04-27
Multi-Agent System/Simulation Based Inference	2025-04-28	2025-05-18
Write-Up	2025-05-19	2025-06-01

Table 2: Exact dates for each task in the project

So far I have written code that produces training and test data for the neural network for both the z-method and the grid-method. I used the CAMB repository to calculate the power spectra and used the ranges for parameters from Table 1. I will run these codes on the DIRAC high performance computer [7], to reduce CPU time, to calculate the data for the neural networks. I will run the z-method 1,000,000 times, choosing a random set of parameters every time using the LHC, and the grid method 100,000 times with 10 z values for each set of random parameters. The z values in the grid-method will be chosen to increase from 0 to 5 on a logarithmically uniform distribution.

Boris Bolliet (Supervisor)

James Martin (Student)

References

[1] Bolliet, Boris, Spurio Mancini, Alessio, Hill, J. Colin, Madhavacheril, Mathew, Jense, Hidde T., Calabrese, Erminia, and Dunkley, Jo. *High-accuracy emulators for observables in ΛCDM, N_{eff}, Σm_ν, and w cosmologies*. Monthly Notices of the Royal Astronomical Society, Vol. 531, No. 1, pp. 1351–1370, 2024.
DOI: 10.1093/mnras/stae1201.

[2] S. Dodelson and F. Schmidt, *Modern Cosmology*, 2nd edition, Chapter 8, Academic Press, Elsevier, 2021. ISBN: 978-0-12-815948-4.

[3] J. Alsing, H. Peiris, J. Leja, C. Hahn, R. Tojeiro, D. Mortlock, B. Leistedt, B. D. Johnson, and C. Conroy, "Speculator: Emulating stellar population synthesis for fast and accurate galaxy spectra and photometry," *The Astrophysical Journal Supplement Series*, vol. 249, no. 1, 2020. DOI: https://doi.org/10.3847/1538-4365/ab917f.

[4] A. Spurio Mancini, D. Piras, J. Alsing, B. Joachimi, and M. P. Hobson, "CosmoPower: Emulating cosmological power spectra for accelerated Bayesian inference from next-generation surveys," *Monthly Notices of the Royal Astronomical Society*, vol. 511, no. 2, pp. 1771–1788, 2022. DOI: https://doi.org/10.1093/mnras/stac064.

[5] C. Sui, D. J. Bartlett, S. Pandey, H. Desmond, P. G. Ferreira, and B. D. Wandelt, "syren-new: Precise formulae for the linear and nonlinear matter power spectra with massive neutrinos and dynamical dark energy," *arXiv preprint arXiv:2410.14623*, 2024. DOI: https://doi.org/10.48550/arXiv.2410.14623.

[6] T. Ohta, M. Koyama, and the Optuna team, "Optuna: A Next-generation Hyperparameter Optimization Framework," *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2019. DOI:https://doi.org/10.1145/3292500.3330701.

[7] "DiRAC High Performance Computing Facility," University of Cambridge, 2024. Available at: https://www.dirac.ac.uk.