Machine Learning in Structure Formation: Phase one (Data)

Amirhossein Samandar
Ali Saraer
Armita Kazemi Najafabadi
Fatemeh Farhangian

30 Mehr

1 Introduction

The population of dark matter halos and their formation initiated from early universe density fluctuations have received attention. In the ΛCDM paradigm of cosmological structure formation, galaxy formation proceeds within the potential wells of extended haloes of dark matter. The dark matter haloes' assembly history and internal properties directly impact the later growth of the galaxies within their cores. Our observation about the redshift of structure formation zone confirms our idea about the dark matter halo's role. However, this is a complex problem due to the highly non-linear nature of the haloes' dynamics. The more numerical method to investigate structure formation is to run extensive simulations, but it has heavy computational expenses.
Press-Schechter [1] method and more novel methods like Excursion Set Theory [2][3] are the most successful semi-analytical models for describing halo collapse. However, these models have their problems:

- Notably, dependence on the cosmological model
- Press-Schechter mass function overestimates halos with lower masses and underestimates massive halos.

This project repeats what people have done in paper [4], applying machine learning methods and solving our problems.
Compared with older methods, the most significant advantage is that it is not a model base that makes us able even to test the models, and it can be applied in all regimes. Also, there is no need to run a heavy simulation for each initial condition. We train the machine once, and we use it over and over again.
It is essential to mention that this method is tender to the simulation premises. The main goal is to train the machine to find a conformity between initial conditions and final halos.

2 Data

2.1 Data Source

Our primary source of data is G-evolution simulations with Neutrinos run by Dr. Farbod Hassani in Oslo university. This simulation run with "Tiling factor = 128,0,0", "$N-grid = 512$", "Box size = 160 Mpc/h" at the Norwegian supercomputer on more than 500 cores. Of course, because of some technical problems, this data has not been prepared before phase one's deadline, so we decided to work with a similar set of data until the original data will be generated.
Therefore, to analyze the formation of halo dark matter in the present universe, we have used the "Virgo-Millennium" database, in which Saba Etezad Razavi and her colleagues have made some significant changes in the rare form of data.
The simulation was performed using a modified version of the publicly available code "GADGET-2" (Springel 2005). A millennium run uses 10^{10} particles, each mass 8.6108 $h^{-1}M$, to follow the evolution of the dark matter distribution within a cubic region of side $500h^{-1}Mpc$ from redshift z = 127 (snapNumber = 0). The Benchmark model [1] is assumed for this simulation which the cosmological parameters in it are $\Omega_m = \Omega_{DM} + \Omega_b = 0.25$, $\Omega_b = 0.045$, $\Omega_\Lambda = 0.75$, h = 0.73, $\sigma_8 = 0.9$ and $n = 1$ with standard definitions for all quantities.

2.2 Data Contents

The data set contains the data of a $603h^{-3}Mpc^3$ box with about 20 million particles and 5kpc resolution at redshift z=127.
These 2 GB data from" Millimil" [5] includes the position of each particle, its velocity vectors, particle ID, the Peano-Hilbert key(we do not need this information about particles, so we drop it.), halo's mass(the mass of potential halo, which includes the particle in the present universe.) in one snapshot, which is in redshift 127.
Data also include two other columns, which are "in-halo" and "in-halo-log". We do not need these pieces of information in our project, so we dropped them.
You can see a further description and Data sample in our Github repository[5].

3 Data Analysis

3.1 Dark Matter Halos

As you can see in the haloes mass histogram, the population of the light haloes is more than the massive haloes. Moreover, the light haloes dominate the haloes population. (Figure 1)

3.2 Position and Velocity Distribution

In order to visualize our data, we have plotted the distribution of the particles in space for z= 127. As shown in figure 2, a homogeneous mass density field can be seen very well because redshift is large enough not to see any sign of structures yet. Also, the particle velocity histogram shows that their distribution looks like a gaussian distribution(figure 3). However, as we expected, there is no sign of the correlation between velocity and position of particles(Figure 3). Also, the velocity field of particles has shown in Figures 4 and 5.

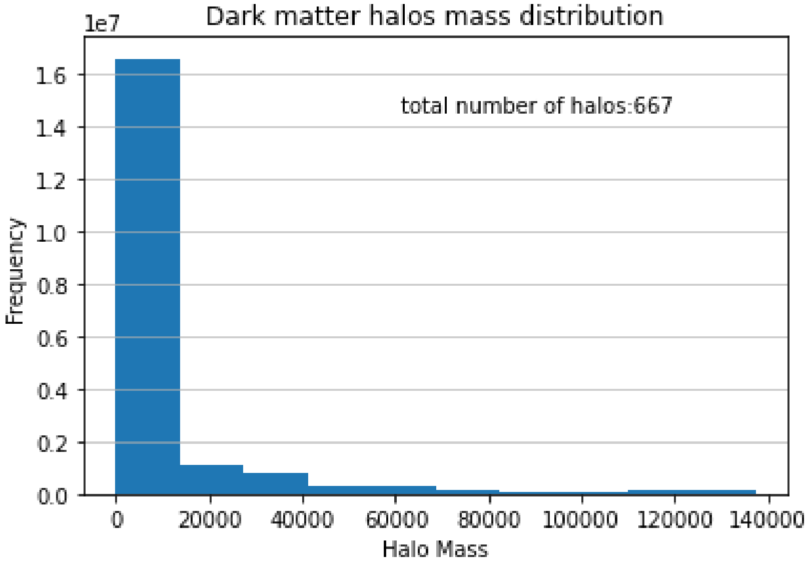

Figure 1: The haloes mass histogram.

4 Data Analytics Training

Ultimately, we need the initial condition of each particle and neutrinos and whether it is in a halo or not at z=0 (with our definition of halo). we should have derived this data from both halo and particle data and then classified our data in two classes," IN" and "OUT." In the next step, we will train initial conditions and "In" or "Out" categories. The machine will help us understand whether a new particle with a given initial condition will be bound in a halo or not.

However, there are some theoretical. Difficulties. For instance, according to J. M. Bardeen[3], Cosmological density fluctuations are often assumed to be Gaussian random fields. The local maxima of such fields are apparent sites for the formation of nonlinear structures. Therefore, one way to face this problem is to part the whole universe into boxes and find the local maxima in every box and specify whether these maxims are responsible for dark matter halo formation or not. The problem with this approach is the size of the boxes. However, we may handle it by using optimization algorithms.

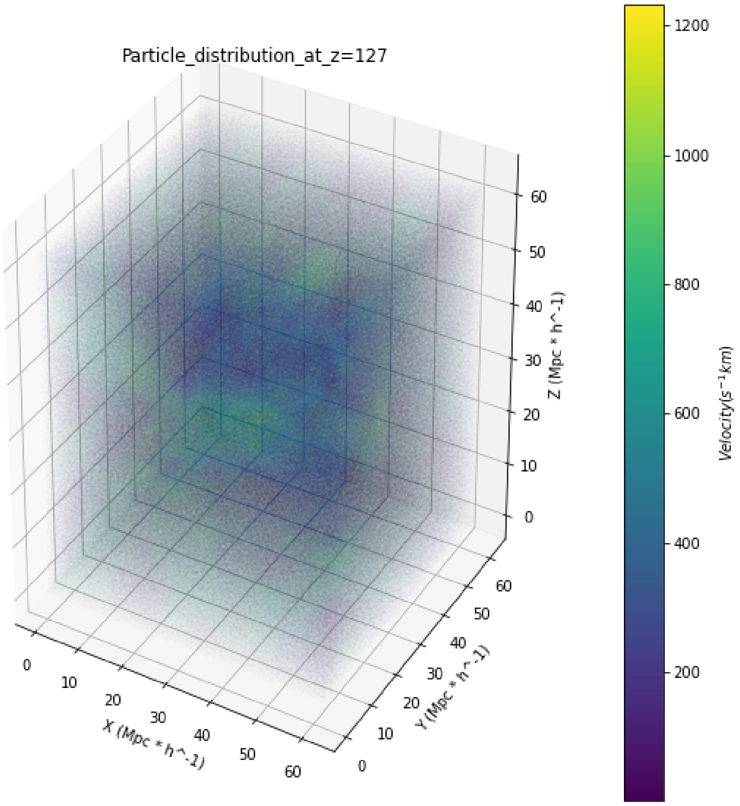

Figure 2: The haloes mass field.

5 Conclusion

We received the data from ”Millimil”(from Saba Etezad Razavi et al.) and made necessary corrections. For a better insight into the data, we visualized the data and compared them with the theoretical point of view. We learned how to work with large databases, terminals, teamwork, and Github and then pushed all the codes and plots on our repository. For finding codes, sample data, graphs, and a further description, check out our Github repository: https://github.com/as2c/MLP—Cosmology

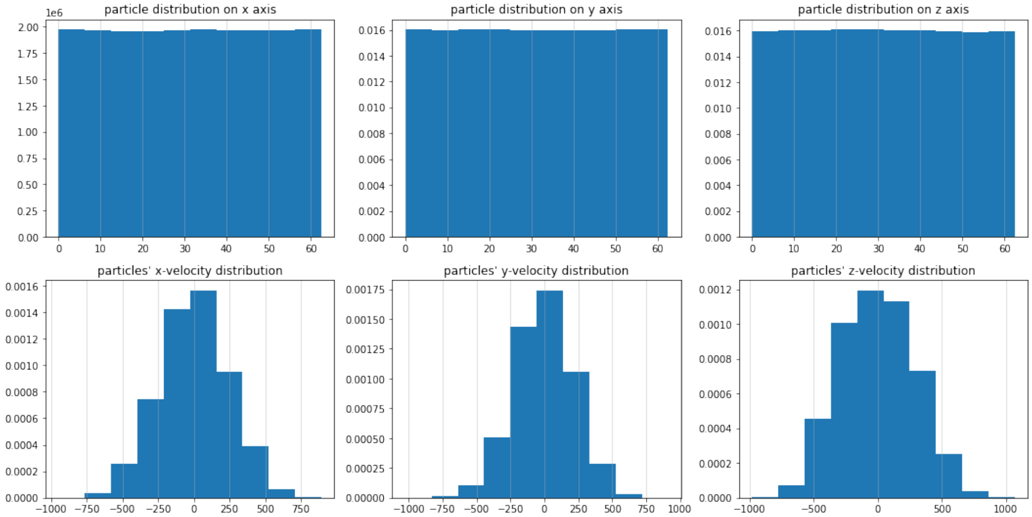

Figure 3: Position and Velocity Distribution

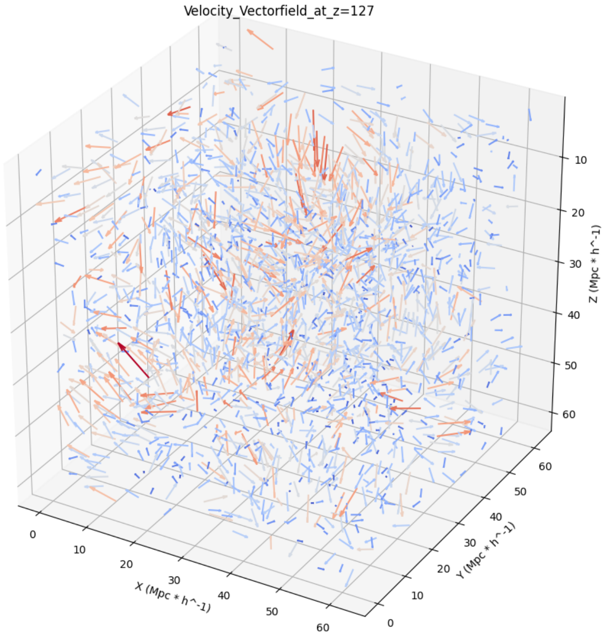

Figure 4: Particles Velocity Field 3D

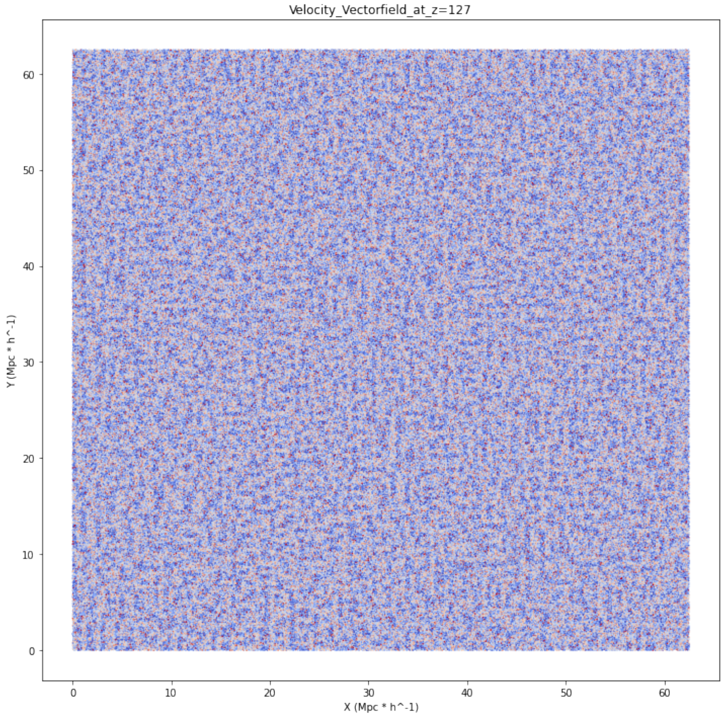

Figure 5: Particles Velocity Field 2D

6 References

1. P. Press W. H. Schechter. “Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation”. In: (1974). url:https://ui.adsabs.harvard.edu/abs/1974ApJ...187..425P/abstract

2. Andrew R. Zentner. “The Excursion Set Theory of Halo Mass Functions,Halo Clustering, and Halo Growth”. In: (2006).
url: https://arxiv.org/abs/astro-ph/0611454

3. J. M. Bardeen. ”THE STATISTICS OF PEAKS OF GAUSSIAN RANDOM FIELDS”. In:(1986).
url : https://ui.adsabs.harvard.edu/abs/1986ApJ...304...15B/abstract

4. Andrew Pontzen Luisa Lucie-Smith Hiranya V. Peiris. ”An interpretable machine learning framework for dark matter halo formation”. In: (2019). url: arXiv:1906.06339v2 [astro-ph.CO] 19 Sep 2019

5. Amirhossein Samandar - Ali Saraer - Fatemeh Farhangian - Armita Kazemi Najafabadi. ”DataPreparation.ipynb” . In :(2021)
   url: https://github.com/as2c/MLP—Cosmology