.

BAYESIAN INFERENCE OF THE QUASAR ABSORPTION LINES FROM THE THREE-DIMENSIONAL Lyα FOREST

TFM

Author:

Gonzalo Vargas

Supervisors:

Francisco-Shu Kitaura

Francesco Sinigaglia

MASTER'S DEGREE IN ASTROPHYSICS

JULY 2023

Contents

List of Figures

SUMMARY

The Lyman-α forest is composed of a series of absorption lines in the spectrum of distant quasars, which are produced due to the interaction of ultraviolet radiation with neutral hydrogen clouds in the intergalactic medium. By analyzing the distribution in space and the statistical properties of these lines, valuable information can be inferred about the distribution of gas density in the intergalactic medium, which also allows to infer the distribution of dark matter in these regions. The Lyman-α forest also allows measurements of baryon acoustic oscillations which are particularly important to investigate the expansion history of the universe and to constrain cosmological parameters.

To reduce systematic errors, selection effects, and compare observations with theoretical models, Lyman-α forest observations are compared with synthetic spectra generated through full cosmological hydrodynamic N-body simulations. However, the fast advance in technology and instrumentation allows the detection of regions at ever greater cosmological distances and at resolutions beyond the capacity of current simulations. For this reason, it is of particular importance to improve alternative and computationally efficient methods for the generation of mock catalogs of the Lyman-α forest.

In this research it is proposed the application of an efficient Hamiltonian Monte Carlo (hereafter HMC) scheme inspired by the work by Kitaura et al. 2012b, to generate high-resolution line-of-sight absorption spectra from the data obtained from the GADGET3-OSAKA cosmological simulation (Aoyama et al. 2018; Shimizu et al. 2019). The HMC method allows to explore the parameter space more efficiently, saving the maximum computational resources for the generation of mock catalogs. Also, this study includes a model of selection effects, such as the completeness, which can affect the measurement of absorption lines in quasar spectra.

From the implementation of the HMC method, it was possible to generate absorption spectra of the Lyman-α forest with a precision of $\sim 5\%$ up to a scale of $k \sim 1.0\,h\,\mathrm{Mpc}^{-1}$, imposing an arbitrary 1D power spectrum (i.e., along the line of sight) and preserving the 3D power spectrum (i.e., over the whole simulation box). This can allow the generation of more precise Lyman-α forest catalogs, ensuring the correct spatial correlations and taking into account selection effects such as completeness.

RESUMEN

El bosque de Lyman-α está compuesto por una serie de líneas de absorción en el espectro de cuásares distantes, que se producen debido a la interacción de la radiación ultravioleta con el hidrógeno neutro en el medio intergaláctico. Mediante el análisis de la distribución en el espacio y de las propiedades estadísticas de estas lineas se puede inferir información valiosa sobre la distribución de la densidad de gas en el medio intergálactico, que a su vez permite inferir la distribución de materia oscura en estas regiones. El bosque de Lyman-α también permite la medición de las oscilaciones acústicas bariónicas que son particularmente importantes para explorar la expansión del universo y constreñir parámetros cosmológicos.

Para reducir errores sistemáticos, errores de selección, y comparar las observaciones con modelos teóricos, las observaciones del bosque de Lyman-α son comparadas con espectros sintéticos generados a partir de sofisticadas simulaciones hidrodinámicas o de N-cuerpos. Sin embargo, el rápido avance en instrumentación permite la detección de regiones a distancias cosmológicas cada vez mayores y a resoluciones mas allá de la capacidad de las simulaciones actuales, por lo cual es de particular importancia recurrir a métodos alternativos y computacionalmente eficientes para la generación de catálogos del bosque de Lyman-α.

En esta investigación se propone la aplicación del método HMC (Hamiltonian Monte Carlo) expuesto en (Kitaura et al 2012b) para generar espectros de absorción en la linea de visión de alta resolución a partir de los datos obtenidos de la simulación cosmológica GADGET3-OSAKA (Aoyama et al. 2018; Shimizu et al. 2019). El método HMC permite explorar el espacio de parámetros de manera más eficiente, ahorrando al máximo el gasto computacional requerido para la generación de catálogos. De igual manera, en este estudio se incluye la modelación de efectos de selección, como la completitud, que puede afectar la medición de las líneas de absorción en los espectros de cuásares.

A partir de la implementación del método HMC se logró generar espectros de absorción del bosque de Lyman-α con una precisión del $\sim 5\%$ hasta una escala de $k = 1.0\,h\,\mathrm{Mpc}^{-1}$ que además tienen la característica de conservar las correlaciones en una y tres dimensiones de los datos de partida dando lugar a la generación de catálogos del bosque de Lyman-α mas precisos y teniendo en cuenta efectos de selección como la completitud.

1 INTRODUCTION

The study of the large-scale structure of the universe is essential to understand its origin and evolution. Currently observed structures began as small perturbations in the primordial plasma caused by quantum fluctuations at early cosmological time. Observations from Lyman-α (hereafter Lyα) Forest indicate that galaxies and clusters of galaxies are distributed in structures such as filaments forming the well-known cosmic web (Bond and Wadsley 1997). Besides, the luminous matter only represents a small fraction of the total components of the universe. The comparison between theory and observations leads to the postulation of the existence of dark matter and dark energy. The composition of dark matter is not clear and have a significant influence in the formation of structures. Also, the accelerated expansion of the universe (Riess et al. 1998) is attributed to dark energy, whose properties remain largely unknown. The understanding of the nature and equation of state of dark energy is one of the biggest challenges in Cosmology.

An important cosmological tool in the understanding of dark energy are the Baryon Acoustic Oscillations (BAOs), which are a characteristic pattern observed in the large-scale distribution of matter in the universe. BAOs arise from sound waves that propagated in the early universe when photons and baryons were tightly coupled. These sound waves left an imprint on the matter distribution, creating a preferred scale known as the sound horizon, of about 150 Mpc. The characteristic scale of BAOs can be used as a standard ruler to estimate cosmological distances. Also, BAOs precision measure allows to constrain cosmological parameters, such as the density of dark matter, the Hubble parameter H_0 or the primordial power spectrum. Comparing the observed BAOs signature at different redshifts it is possible to determine the expansion history of the universe and investigate the equation of state of dark energy.

BAOs can be measured through the statistical analysis of the distribution of galaxies at large scales. From the estimation of the two-point correlation function the BAOs can be observed as a distinctive peak at a scale corresponding to the sound horizon. Another method in the measure of BAOs comes from the determination and study of the angular power spectrum of the anisotropies of the CMB, which captures the statistical properties of temperature fluctuations, related to the characteristic scale of the primordial density fluctuations that gave rise to BAOs.

A technique applied in last years for the detection of BAOs comes from the Lyα forest, the main topic of this investigation, which consists in a collection of absorption lines in the spectrum of distant quasars. These absorption lines are caused by the Lyα resonant scattering of photons with the neutral hydrogen atoms present in the intergalactic medium. By conducting a detailed analysis of the statistical properties and spatial distribution of these absorption lines, it is possible

to reconstruct the matter distribution on large scales. This includes studying the formation and evolution of structures such as galaxies and galaxy clusters. Additionally, the Lyα forest provides an opportunity to study the nature of dark matter and dark energy since its properties and distribution are influenced by gravity at small scales where baryon physics is relevant, and at large scales is influenced by the expansion of the universe.

There are several surveys which measured hundreds of thousands of Lyα forest spectra, including the Sloan Digital Sky Survey (SDSS), the Baryon Oscillation Spectroscopic Survey (BOSS), the Extended Baryon Oscillation Spectroscopic Survey (eBOSS), and the Dark Energy Spectroscopic Instrument (DESI). Through the use of high-precision telescopes and spectrographs, these surveys have mapped vast volumes of the universe and provided detailed line-of-sight (LOS) absorption spectra from sources at different redshift. Then, observations are compared with mock catalogs (hereafter mocks for shortness), which are synthetic data obtained from cosmological simulations that describe the large-scale evolution of the universe, incorporating detailed physical models that account for the physics of neutral hydrogen and its interaction with radiation.

Lyα forest mocks provide valuable insights into the formation of large-scale structures in the high-redshift universe. Additionally, mocks are used to tighten the constrains on cosmological models through the comparison of the results obtained from simulations with observations. The improvements in the detection instruments allows to obtain LOS spectra with increasingly higher resolutions, at the same time even more distant objects can be observed. This implies the implementation of numerical simulations of large cosmological volumes at very high resolutions, making impractical the use of N-body simulations or hydrodynamic simulations due to the high computational cost required. An alternative is to apply computational methods based on statistical models that accurately reproduce the observed large-scale matter distribution at the resolution required by current surveys.

The aim of this investigation is to develop a scheme for the generation of high-resolution absorption spectra in the LOS from synthetic data from hydrodynamic simulations through the implementation of a Hamiltonian Monte Carlo sampling method (HMC). The novel approach of this study is to ensure that the generated spectra accurately preserve correlations in both one and three dimensions. To this purpose, it is of particular interest the study of the properties of the one- and three-dimensional power spectrum of the Lyα forest absorption flux. This accuracy is crucial for creating data mocks that can be compared with surveys, enabling a more precise and comprehensive understanding of the characteristics of the Lyα forest.

2 FRAMEWORK

2.1 Large Scale Structure of the Universe

The ΛCDM model is the currently accepted cosmological model in the description of the observable universe, it is consistent with observations including the properties of the CMB, the distribution of galaxies, and the observations from supernovae. In this cosmological description the universe starts with the Big Bang about 13.7 billion years ago. As the universe expanded and cooled, matter collapses to form galaxies and stars. Observations indicate that the actual distribution of matter in the universe is not uniform, galaxies form large structures such as clusters and there are low-density regions known as voids. However, at scales greater than 100 Mpc, the Cosmological Principle (CP) can be considered valid.

The CP states that the universe is homogeneous and isotropic on large scales, and has important implications for the study of the large-scale structure of the universe and its evolution over time. It allows to model the universe as an homogeneous fluid and apply the tools of fluid dynamics to study its behavior. The theoretical framework given by the General Relativity theory allows a description of the global geometry of the universe and relate it with the energy density components through Friedmann's equations:

$$\left(\frac{\dot{a}}{a}\right) = \frac{8\pi G}{3}\rho - \frac{\kappa c^2}{a} + \frac{\Lambda c^2}{3} \tag{1}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2}\right) + \Lambda c^2 \tag{2}$$

where $\rho(t)$ is the energy density, $p(t)$ the pressure, and the Hubble expansion rate can be written as $H = \dot{a}/a$. The differential equations (1), (2) have two free parameters κ and Λ, and $a(t)$, $\rho(t)$ and $p(t)$ are unknown, then, to find a solution a third equation is needed. This comes from the equation of state of the fluids:

$$p = w\rho \tag{3}$$

With $w = 0$ for non-relativistic matter, $w = 1/3$ for radiation and $w = -1$ for the description of dark energy as a cosmological constant Λ. According to the values of κ and Λ, and the contribution of the different components of the universe to the values of $p(t)$ and $\rho(t)$, different solutions to the Friedmann's equations can be obtained, which leads to different scenarios in the large-scale evolution of the universe. The energy density of each component of the universe is typically expressed relative to the critical energy density $\Omega_i = \rho_i/\rho_c$. Where ρ_c is defined as the energy density that gives place to a non-curvature universe ($\kappa = 0$):

$$\rho_c = \frac{3H_0^2}{8\pi G} \tag{4}$$

Where H_0 is the Hubble constant. According to the measurements obtained from Planck satellite (Planck Collaboration, Aghanim et al. 2020) $H_0 = 67.7\,\mathrm{km/s/Mpc}$, the adimensional density of baryonic matter $\Omega_b = 0.0482$, of dark matter $\Omega_M = 0.307$ and dark energy $\Omega_\Lambda = 0.6928$.

In an isotropic and homogeneous universe, the formation of structures started from primordial quantum fluctuations formed at early cosmological time that were amplified during the inflation period with a characteristic spectrum of amplitudes across different length scales. This density fluctuations can be expressed as a relative deviations of density δ respect to the homogeneous mean density $\bar{\rho}$:

$$\delta(\vec{r}) = \frac{\rho(\vec{r}) - \bar{\rho}}{\bar{\rho}} = \frac{\rho(\vec{r})}{\bar{\rho}} - 1 \tag{5}$$

A good approximation for early cosmological times is consider that the density fluctuations $\rho(\vec{r})$ are small compared with the mean density $\bar{\rho}$, that means $\delta(\vec{r}) << 1$. This assumption is known as linear perturbation theory and allows to neglect the second-order terms in the evolution of the fluctuations. For the case of domain of non-relativistic matter ($w = 0$) one gets:

$$\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} - \frac{c_s^2}{a^2}\nabla^2\delta = 4\pi G\bar{\rho}\delta \tag{6}$$

Where $c_s = \sqrt{|\partial p/\partial\rho|}$ is the sound speed. A detailed derivation of this equation can be found in (Peebles 1980). The first term in the l.h.s. of equation (6) represents the density growth, the second term is the Hubble's friction since the expansion of the universe opposes to the concentration of matter, the third term corresponds to the pressure contribution, and the term on the r.g.s. is the gravitational contribution.

Proposing a solution to equation (6) in the form of a fluctuation density field within a finite volume $V = L^3$, which can be decomposed into a Fourier series expansion:

$$\delta(\vec{r}) = \sum_k \delta(\vec{k}) e^{i\vec{k}\cdot\vec{r}} \tag{7}$$

Where $\vec{k}$ is the wave vector given by $|\vec{k}| = 2\pi/\lambda$ with λ the characteristic wavelength of the fluctuation. The boundary conditions implies $|\vec{k}| = 2\pi n/L$ with $n = \{0, 1, 2, ...\}$. Then, the fluctuation field can be considered as a collection of modes with different wavelengths and its can be studied independently. Tiny values of λ (large values of k) correspond to small-scale structures, while huge values of λ (small values of k) correspond to large-scale structures.

2.2 The Power Spectrum

A statistical analysis of the distribution of matter in the universe can be very useful in the understanding of structure formation and the validation of cosmological models. The two-point correlation function $\xi(\vec{r})$ is a statistical measure that provides information about the spatial correlations between pairs of objects in the universe. $\xi(\vec{r})$ describes the excess of probability of finding two objects at a given separation distance, compared to the case of a random distribution. It is defined as:

$$\xi(\vec{r}) = \langle \delta(\vec{r})\delta(\vec{r}+\vec{x}) \rangle \tag{8}$$

The two-point correlation function provides insights into the underlying physical processes that shape the large-scale structure of the universe and is usually measured for galaxies or other cosmological tracers like the Lyα Forest. There are physical processes that affect fluctuations on particular scales, such as gravity, so is useful the estimation of the power spectrum, which contains valuable information about the amplitude of density fluctuations as a function of the scale $\vec{k}$. The power spectrum can be defined as the variance of the Fourier modes $\vec{k}$ of density fluctuations (Linder 1997):

$$P(\vec{k}) \equiv \langle \delta_k \delta_k' \rangle = \langle |\delta_k|^2 \rangle \tag{9}$$

Where the angle brackets denotes average over all modes. The power spectrum is related with the two-point correlation function through the Fourier transform:

$$\xi(\vec{r}) = \langle \delta(\vec{r})\delta(\vec{r}+\vec{x}) \rangle = (2\pi)^{-3} \int P(\vec{k}) e^{i\vec{k}\vec{r}} d^3k \tag{10}$$

Measuring the power spectrum at different redshifts reveals key information about cosmological parameters of the universe, such as dark matter and energy densities, cosmic expansion rate, and the nature of primordial fluctuations. Also, detect effects of fundamental physics in the universe, such as the existence of exotic particles and predictions by theories beyond the Λ CDM model.

In the linear approximation each mode evolves independently, this means that density fluctuations have variances that are scale-invariant, so its statistics do not change with the characteristic scale. Perturbations caused by a random phase process like quantum fluctuations will follow a Gaussian probability distribution:

$$f(\delta_k) = (2\pi)^{-3/2} P(\vec{k})^{-3/2} \exp\left(\frac{-|\delta_k|^2}{2P(\vec{k})}\right) \tag{11}$$

However, as the universe evolves, denser regions exert a stronger gravitational pull, causing a non-linear growth of perturbations. This produces deviations from gaussianity in the current

observed distribution of matter on scales where gravity is relevant, and generates correlations between fluctuations of different wavelengths λ and amplitudes.

2.3 Baryon Acoustic Oscillations

The relevance of the pressure and the gravitational terms in equation (6) depends on the value of the characteristic length of the fluctuations λ respect to a critical wavelength known as the Jeans length defined as:

$$\lambda_J = c_s \sqrt{\frac{\pi}{(1+w)(1+3w)\bar{\rho}}} \tag{12}$$

For non-relativistic matter $\lambda_J = c_s\sqrt{\pi/\bar{\rho}}$. For $\lambda > \lambda_J$, the gravitational term will dominate over the pressure term, leading to a solution where the fluctuations can grow. On the other hand, for $\lambda < \lambda_J$, the pressure term dominates, and the solution corresponds to a damped harmonic oscillator due to Hubble's friction, therefore the fluctuations will oscillate without growing. This is the case before recombination when the Universe consisted in a strongly coupled photon-baryon fluid in thermodynamic equilibrium. The interplay between gravitational attraction and radiation pressure led to the formation of an oscillatory pattern around the density fluctuations.

At the time of recombination at $z \approx 1500$, the decrease in the density of free electrons causes an increase in the mean free path of the photons, reaching a point at which the mean scattering time of the photons exceeds the Hubble time H^{-1} (which represents the characteristic time of the expansion rate at that time). This leads to the decoupling of photons at $z \approx 1150$, which are responsible for the cosmic background radiation that is detected today. At that moment, baryons are not longer affected by the radiation pressure, and the oscillations in the perturbations freeze, leaving an imprint on the distribution of matter, an excess of density at a characteristic length $r_s = 150\,\mathrm{Mpc}$, this phenomena is known as Baryon Acoustic Oscillations.

BAOs can be identify as a peak in the correlation function and is large enough to not be affected by non-linear processes of structure formation. Since the position of the acoustic peak in the correlation function in comoving coordinates is practically the same from the time of recombination to the present, it can be used as a standard ruler to measure distances on large scales (Eisenstein et al. 2005). Unlike other standard rules, BAOs is not an observable physical object, instead its scale is inferred from statistical measurements of the distribution of objects and is necessary surveying large cosmological volumes.

Since BAOs is printed in the distribution of matter in three dimensions, its characteristic scale can be measured both in the radial direction, which is related to the redshift, and in the transverse

direction as the angle it subtends in the sky, this is illustrated in Figure 1. As it can be assumed the value of r_s in comoving coordinates as invariant, knowing the angular measure of BAOs, the angular distance $D_A(z)$ can be obtained (Basset and Hlozek, 2009).

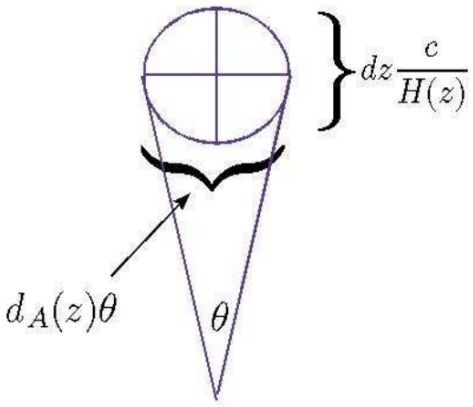

Figure 1: Radial length and transverse size in terms of the Hubble distance $D_H(z)$ and angular distance $D_A(z)$ respectively. Figure from (Basset and Hlozek, 2009).

$D_A(z)$ is defined as the associated distance to the euclidean relation of the arc length. For an object of size L at a redshift z which subtends an angle $d\theta$ on the sky is given by:

$$D_A = \frac{L}{d\theta} \tag{13}$$

This relation is valid at low redshift in which the curvature and the expansion of the universe can be neglected. However a more general expression can be derived from the Robertson-Walker metrics and Friedmann's equations:

$$D_A(z) = \frac{c}{H_0(1+z)\sqrt{-\Omega_k}} \sin\left(\sqrt{-\Omega_k}\chi(z)\right) \tag{14}$$

Where Ω_k is the adimensional curvature energy density and $\chi(z)$ is the comoving distance defined as:

$$\chi(z) = c\int_0^z \frac{1}{H(z)}dz \tag{15}$$

For the case of a flat universe, which is a good approximation supported by observations and form part of the ΛCDM model, one has $\Omega_k = 0$ and (13) can be considered valid, taking into account the dependence of L with redshift and its relation with the comoving distance, one leads to:

$$D_A(z) = \frac{\chi(z)}{1+z} \tag{16}$$

On the other hand, r_s on the radial direction represents a redshift difference dz from which a measure of the expansion rate $H(z)$ can be obtained. This can be achieved from the definition of

Hubble distance:

$$D_H(z) = \frac{c}{H(z)} \tag{17}$$

$D_H(z)$ represents the distance at which the recession speed is equal to the speed of light. And sets a limit to the distance over objects can be observed. $H(z)$ is related with the energy components of the universe as:

$$H = H_0 \sqrt{\sum_i \Omega_{i0}(1+z)^{3(1+w_i)}} \tag{18}$$

In the description of our universe from ΛCDM model:

$$H(z) = H_0 \sqrt{\Omega_{r0}(1+z)^4 + \Omega_{m0}(1+z)^3 + \Omega_\Lambda f(z)} \tag{19}$$

Has been taken into account that the curvature is negligible. The function $f(z)$ determines the evolution of dark energy, for the case of the cosmological constant $f(z) = 1$. Therefore, by measuring the characteristic scale of BAOs in the radial direction as a redshift difference dz, the value of $D_H(z)$ can be inferred:

$$D_H(z) = \frac{r_s}{1+z} dz \tag{20}$$

Likewise, from the measurement of BAOs in the transversal direction, the value of $d\theta$ is inferred, and the value of the angular distance can be obtained:

$$D_A(z) = \frac{r_s}{1+z} \frac{1}{d\theta} \tag{21}$$

One approach is to treat r_s as unknown, then a determination of BAO scale at different redshifts are necessary. By comparing ratios of distances $D_A(z)/D_A(z')$ and $D_H(z)/D_H(z')$ between two different measurements of BAO, a model-independent approach can be employed. Other approach is to determine r_s from the theoretical model and an estimation of cosmological parameters from a different method like the measurements of the CMB, enabling distance calculations. This approach is more model-dependent than the first one and can allow to constrain models of dark energy with a redshift-dependent equation of state $f(z)$ (Seo and Eisenstein 2003). Both approaches can lead to the determination of $H(z)$ at different redshifts and and can be applied to reconstruct the expansion history of the universe (wang 2006).

2.4 Lyα Forest

The Lyα forest was predicted by James Gunn and Bruce Peterson (Gunn and Peterson 1965), and was one of the first evidences of the existence of the intergalactic medium (IGM). The neutral hydrogen HI of the IGM produces an absorption line in the ultraviolet radiation emitted

by distant quasars at a wavelength of $\lambda_{L\alpha} = 121.6\,\text{nm}$, which corresponds to the Lyα atomic transition of an electron from the ground state ($n = 1$) to $n = 2$ orbital. Since the interacting gas regions are located at different redshift from the observer, this phenomena is detected as a series of absorption lines. Due to the extinction of the Earth's atmosphere, which absorbs light with wavelengths below 360 nm, the detection of the Lyα forest mostly comes from quasars at $z > 2$. Most of the detected lines are within a wavelength range of 400 nm to 900 nm, as is shown in Figure 2.

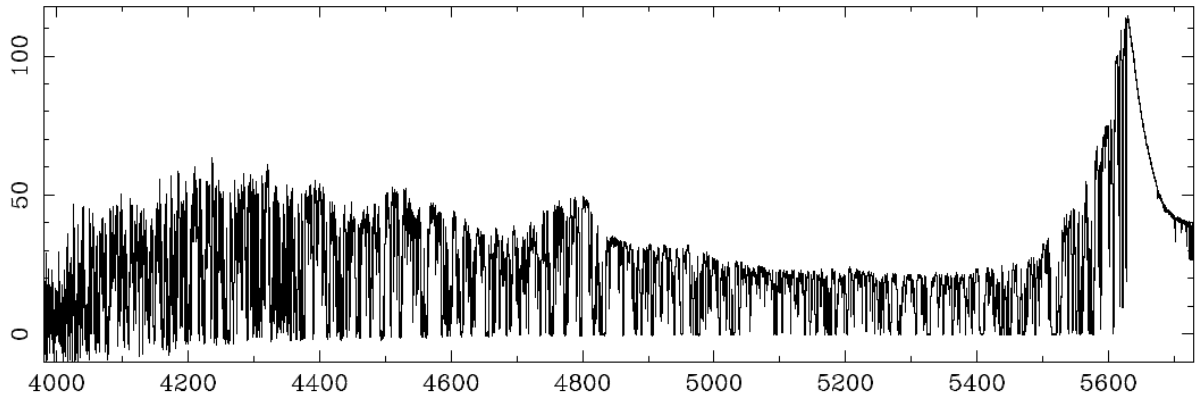

Figure 2: Lyα forest absorption lines in a quasar spectrum at $z = 3.63$ detected with the HIRES spectrograph on the Keck $10\,m$ telescope in Hawaii. Figure from Womble et al. 1996.

Gunn and Peterson predicted that at a high enough redshift this absorption would manifest as a distinctive absorption trough in the spectrum of a quasars, this is known as the Gunn-Peterson effect. Subsequent surveys focused on studying QSOs revealed that a high redshift the collective emission of ionizing radiation from all known sources produces a powerful ultraviolet radiation field intense enough to maintain the hydrogen in a highly ionized state. The Gunn-Peterson trough is a evidence of the re-ionization epoch that take place between $z \sim 20$ and $z \sim 6$. During that period the formation of stars begin inside galaxies and produce a high UV radiation emission capable of ionize the neutral hydrogen.

The Lyα forest can be quantified according to the density column that produces it (Weymann et al. 1981). Most of the detected absorption lines corresponds to column densities among $N \sim 10^{13}\,cm^{-2}$ and $N \sim 10^{16}\,cm^{-2}$. Regions with a high density ($N \sim 10^{17}\,cm^{-2}$) will behave optically thick to radiation and a discontinuity will be observed in the Lyman limit at $91.2\,nm$, regions like this are called Lyman Limit Systems (LLS). At even higher densities ($N \sim 10^{19}\,cm^{-2}$), the radiation is rapidly absorbed in the external regions keeping the gas in the central regions mostly neutral, this is known as self-shielding. In Damped Lyman Systems (DLA) the Lorentzian profile of the absorption line can be detected. On the other hand, when a quasar expels a

considerable amount of matter In the LOS, a broadening of absorption lines of the Lyα and other metals like carbon CIII or silicon SiIV is produced. This broadening is caused by the dispersion velocity of the exiting gas, resulting in the emergence of broad absorption lines (BAL) within the quasar's spectrum.

The Lyα forest can provide valuable insights of the properties of the IGM. The temperature and density of the gas regions depends on the heating produced by photo-ionization and the adiabatic cooling. Most of the gas that give place to the Lyα forest is in mildly density, diffuse and not shock-heated regions, which leads to a relation between T and ρ in the form of a power law:

$$T = T_0 \left(\frac{\rho}{\bar{\rho}}\right)^{\gamma} \tag{22}$$

where T_0 and γ are related with the UV radiation background and the reionization history of the universe and its values have been estimated around $4000\,K < T_0 < 10000\,K$ and $0.3 < \gamma < 0.6$ (Hui and Gnedin 1997). Under the assumption of photoionization equilibrium, the recombination rate of protons and electrons to form HI is in balance with the reionization rate caused by the UV radiation:

$$\mathrm{n_e n_p}\alpha(\mathrm{T}) = \mathrm{n_{HI}}\Gamma_{\mathrm{HI}} \tag{23}$$

where Γ_{HI} is the photoionization rate of the neutral hydrogen and $\alpha(T) \propto T^{-0.7}$. The fraction of neutral hydrogen is given by $\mathrm{x_{HI}} = \mathrm{n_{HI}/n_H}$. Neglecting the abundance of helium and metals in the IGM $\mathrm{n_H} = \rho/\mathrm{m_p}$, with m_p the mass of the proton. And considering that $\mathrm{n_e} \approx \mathrm{n_H}$, equation (23) leads to (Rauch 1998):

$$\mathrm{n_{HI}} = \frac{\alpha(\mathrm{T})}{\Gamma_{\mathrm{HI}}\mathrm{m_p^2}}\rho^2 \tag{24}$$

Then, the density of neutral hydrogen will be proportional to $\rho^2 T^{-0.7}/\Gamma_{\mathrm{HI}}$.

On the other hand, the optical depth of the IGM can be related with the density of neutral hydrogen n_{HI} as (Gunn and Peterson, 1965):

$$\tau = \frac{\pi e^2}{m_e c}\frac{f_{L\alpha}\lambda_{L\alpha}}{H(z)} n_{HI} \tag{25}$$

where $f_{L\alpha}$ is the Lyα oscillator strength. Equations (24) and (25) lead to an expression linking the optical depth τ and $\rho/\bar{\rho}$ known as the Fluctuating Gunn-Peterson Approximation, (FGPA hereafter, Rauch 1998):

$$\tau = A\left(\frac{\rho}{\bar{\rho}}\right)^{\beta} \tag{26}$$

where $\beta \equiv 0.7-2.0$ and A is a normalization constant. The FGPA is widely used in the generation of fast approximated Lyα forest simulations. It is an extension of the original Gunn-Peterson approximation, which assumes a fully neutral IGM. The main idea behind the FGPA is that during the cosmic reionization epoch, the IGM is not uniformly ionized but exhibits spatial variations in the HI fraction. These variations results in fluctuations in the absorption of light by neutral hydrogen. The absorption flux F and the optical depth τ can be related as:

$$F_T = \frac{F}{F_C} = e^{-\tau} \tag{27}$$

where F_T is the transmitted flux, obtained by normalizing the observed flux F to the (fitted) quasar continuum F_C.

The statistical analysis of the Lyα forest contains valuable information that can be used to constrain cosmological parameters (McDonald et al. 1999), the position of BAO peak and the expansion history of the universe (McDonald and Eisenstein 2007), and the mass of neutrinos (Seljak et al. 2005). The increase in the last years in the number of detected sources and the resolution of the spectrographs made it possible a more precise reconstruction of the distribution of gas of the IGM at cosmological distances, along with the advance in models based on simulations allows to study the distribution of matter in the large scale structure and make inferences of the properties of dark matter (Kitaura et al. 2012b; Sinigaglia et al. 2021).

2.5 Surveys and Mock Catalogs

Surveys of QSOs spectra at high redshift are valuable tools for investigating the distribution of matter on large scales and exploring the processes that influenced the formation and evolution of structures and refining the comprehension of the fundamental forces and dynamics governing the expansion of the universe. This section is a brief review of the main Lyα forest surveys, and then, a description of mock catalogs and its relevance in recent research in Cosmology.

The Sloan Digital Sky Survey (SDSS) has mapped large volumes of the universe the Data Release 16 (DR16) in September 2021 includes data from all these surveys and had observed and cataloged over $\sim 500,000$ quasars. The Baryon Oscillation Spectroscopic Survey (BOSS) (Dawson et al. 2013) is an extension of SDSS that specifically focuses on detecting BAOs. The BOSS Data Release 12 (DR12) in July 2015 include the spectra of $\sim 160,000$ quasar data set to measure BAOs. The Extended Baryon Oscillation Spectroscopic Survey (eBOSS) is a continuation of BOSS focused on the study of the expansion history of the universe and investigate the nature of dark energy (Zhao et al. 2021).

The novel Dark Energy Spectroscopic Instrument (DESI) (Levi et al. 2019) consists of thousands of robotically controlled optical fibers that are positioned to capture light from specific targets. DESI upcoming data releases are going to considerably increase the number of spectra from galaxies and QSOs and also increase the resolution of the detected absorption LOS. This improvement in the large-scale measurements also requires an improvement in the modeling of the Lyα forest.

Lyα forest mock catalogs consist in a collection of synthetic absorption LOS obtained from simulations based on a theoretical model. Mocks play a crucial role to validate cosmological models and also identify systematic errors in the surveys. Simulations are implemented considering the physics of the ionization of hydrogen by radiation sources such as stars and quasars, as well as the radiative cooling of gas and recombination processes. Once the Physics has been established, the synthetic spectra are obtained by tracing absorption LOS across the simulation volume and applying an interpolation method to calculate the optical depth based on the thermal and ionization history of the gas. Then, the generated synthetic spectra are compared to actual observations.

Earliest hydrodynamic simulations of the IGM and the absorption of Lyα forest incorporate a structure formation model, galaxy evolution and the properties of the gas (Miralda-Escudé et al. 1996). These simulations revealed that the fundamental characteristics of the IGM are slightly influenced by the cosmological model. Over time, other numerical approaches have been applied to model the properties of the IGM, like Smoothed Particle Hydrodynamics in Eulerian and also Lagrangian description (Zhang et al. 1995).

To reproduce realistic absorption lines consistent with observations, recent hydrodynamic simulations take into account different effects such as HI self-shielding, redshift space distortions, thermal broadening, and nonlinear corrections on small scales. This considerably increases the time and computational cost necessary to reproduce the behavior of the IGM in the cosmological volumes required for the surveys. To address this issue different approaches have been used. One method consists in applying the FGPA to independent dark matter fields obtained from DM-only N-body simulations. Other methods apply approximated gravity solvers and Gaussian random fields. Also, methods focused on the correction of shell crossing (Kitaura & Hess 2013). Alternative techniques focusing on matching the Lyα forest probability distribution function (PDF) and the power spectrum have been also developed. In general, a high level of precision in modeling two-point and three-point statistics is crucial to reproduce the statistical properties of the Lyα forest (Sinigaglia et al. 2022, and references therein).

3 METHODOLOGY

The aim of this research consists in developing a scheme to produce absorption spectra along the LOS at the resolution required for the generation of mock catalogs for current Lyα forest surveys with high precision and a consistent 1D and 3D power spectrum based on synthetic spectra of lower resolution obtained from dark matter hydrodynamic simulations. Two methods were tested. First, I applied a lognormal transformation to obtain a gaussian component of the optical depth and added to it a gaussian noise at small scales, iteratively adjusted. The second method consists in the implementation of a Hamiltonian Monte Carlo scheme and the incorporation of the Leapfrog scheme as a numerical integrator to achieve efficient exploration of the parameter space (Kitaura et al. 2010). This chapter describes the implementation of both methods, first a description of the accuracy and computational advantages of the gaussian noise model, and then, the HMC sampling scheme. Finally, in section 3.4 a description of a model for the completeness, a selection effect relevant in the Lyα forest.

3.1 Transmitted flux in the line of sight

The Gaussian noise model and the HMC scheme were tested with data from the cosmological smoothed-particle hydrodynamic (SPH) code GADGET3-OSAKA (Aoyama et al. 2018; Shimizu et al. 2019) at a comoving volume of $(500\,\mathrm{Mpc}\,h^{-1})^3$ at $z = 2$ with $N_c = 128$ cells in each dimension with the values of the transmitted flux field F_T, which results in a resolution of $4\mathrm{Mpc}h^{-1}$.

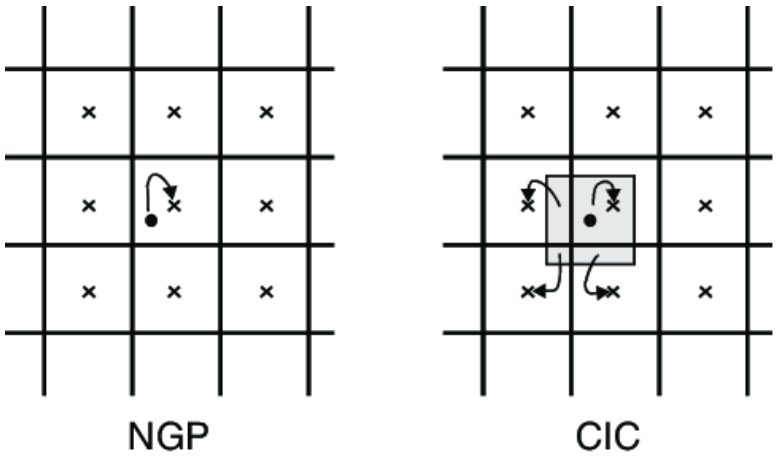

Figure 3: Graphic description of the NGP and the CIC interpolation methods.

The absorption flux from simulations is represented on a discretized 3D mesh, while the observations are one-dimensional spectra in a certain range of wavelengths. In order to obtain a realistic spectrum that can be compared with observations, the first step is to establish LOS

along the volume once the position of the observer and the position of the quasar have been defined. To determine the flux along the LOS from the field values in each cell, an interpolation method is applied, in which a mass assignment function $w(d)$ is introduced to establish the value to each bin of the spectrum according to its proximity to the cells of the computational volume. In the Nearest Grid Point (NGP) interpolation the mass assignment function is defined as:

$$w(d) = \begin{cases} 1 & d \leq H_c/2 \\ 0 & \text{otherwise} \end{cases} \tag{28}$$

Where H_c is the cell length defined as the physical comoving length over the number of cells N_c. On the other hand, in the Cloud in Cell (CIC) interpolation $w(d)$ is defined as:

$$w(d) = \begin{cases} 1 - \frac{d}{H_c} & d \leq H_c \\ 0 & \text{otherwise} \end{cases} \tag{29}$$

Both methods are illustrated in figure 3 for the case of a two dimensional grid. Figure 4 shows an example of two LOS obtained from the simulation volume with the NGP and the CIC.

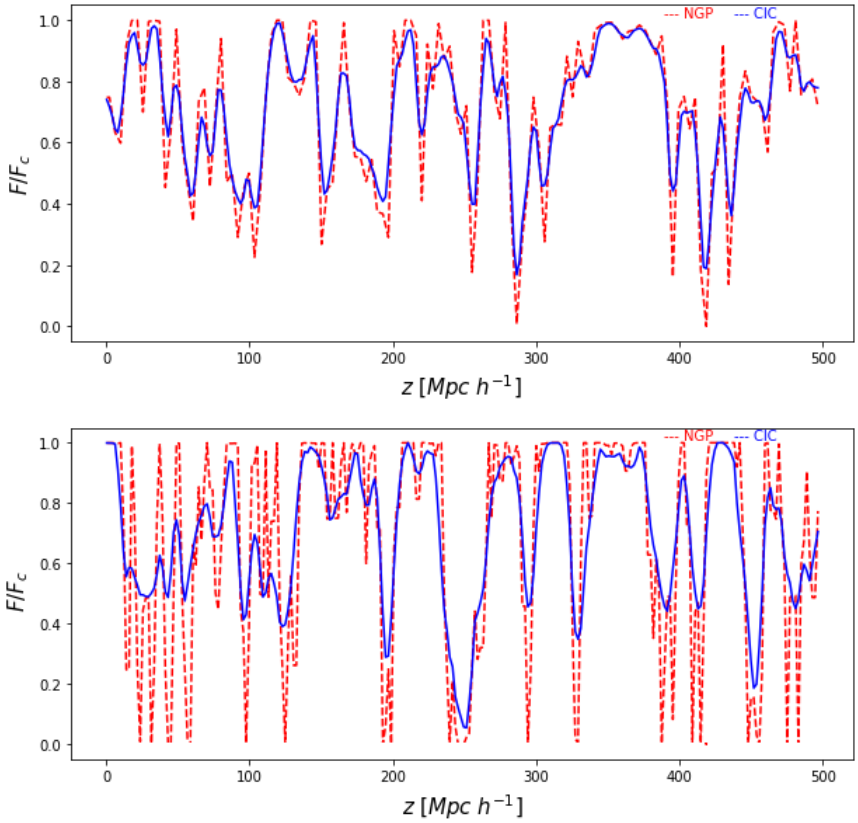

Figure 4: Absorption flux for two random LOS defined in the comoving volume from reference simulation. The NGP method is shown in red and CIC in blue.

It can be noticed that the CIC method produces an apparent loss of information compared to the NGP, due to the dependence of the mass assignment function on the parameter H_c in each method, implying that the CIC applies a larger smoothing in the assignment of the absorption flux values. Therefore, the NGP interpolation is adopted in this investigation.

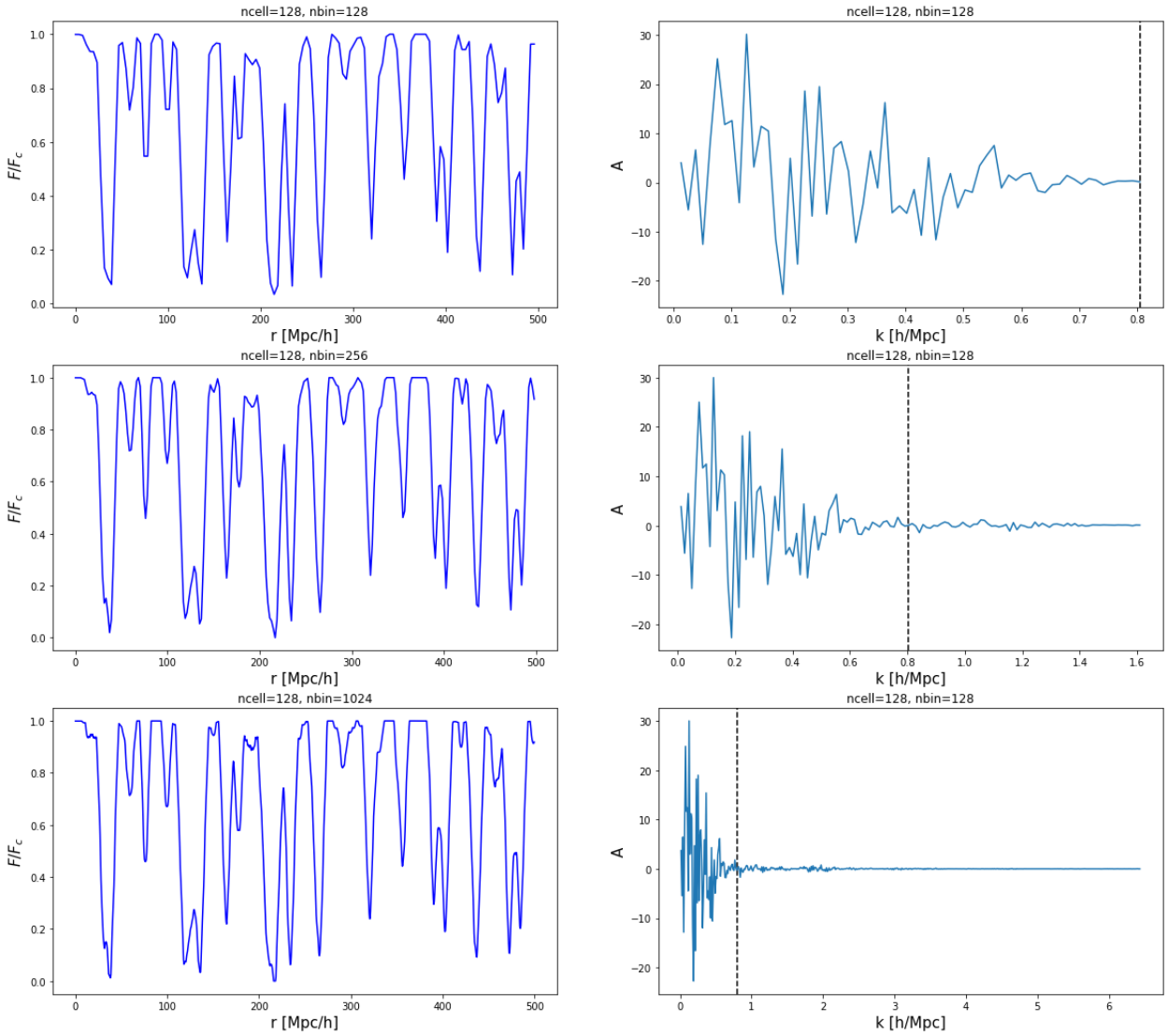

Figure 5: upsampling of a LOS from the reference simulation. The upper panels shows a skewer from the reference simulation at a resolution of $4\,\mathrm{Mpc}\,h^{-1}$. On the left in configuration space and right in Fourier space. The middle panels shows the same skewer after an expansion of twice the resolution $2\,\mathrm{Mpc}\,h^{-1}$ and the bottom ones a four times the resolution $1\,\mathrm{Mpc}\,h^{-1}$. The dotted line corresponds to the Nyquist frequency of the low resolution skewer.

3.2 Gaussian Noise Model

The Gaussian probability distribution is widely used in statistics because it describes the random variation observed in many natural phenomena. The central limit theorem states that the sum of a large number of independent random variables, each with any distribution, tends towards a Gaussian distribution. This property makes the Gaussian noise method an useful approximation for modeling complex signals with the advantage of being computationally efficient.

In the Lyα forest the observable data is the absorption flux F_T. From simulations one can obtained an absorption flux field F_T in a comoving volume, which in this context represents a coarse-grained version of the high-resolution signal. The first step is to apply the interpolation method described in section 3.1 to obtain the absorption flux in the LOS (skewers hereafter), then upsample the low-resolution LOS to match the desired high-resolution. In this particular case the tests were performed on data at a resolution of $4\,\mathrm{Mpc}\,h^{-1}$ to obtain a sample of twice the resolution $2\,\mathrm{Mpc}\,h^{-1}$. The analysis was limited to an improvement of resolution by just a factor of 2 to save computational time, in practice one seeks to reach a resolutions of $0.05\,\mathrm{Mpc}\,h^{-1}$

The physical information of the Lyα forest extracted from the reference simulation is contained in the large scales up to the resolution of the simulation allows, that is, up to the so-called Nyquist frequency:

$$K_n = \frac{\pi N_c}{L} \tag{30}$$

with $N_c = 128$ and $L = 500\,\mathrm{Mpc}\,h^{-1}$, which in the case of this investigation leads to a value of $K_n = 0.8\,h\,\mathrm{Mpc}^{-1}$. The upsampling provides high-resolution skewers, but the relevant physical information is still limited up to the Nyquist frequency at this point, as is shown in figure 5. Also, the upsampling introduces an artificial noise beyond the Nyquist frequency that can be noticed in the right panels of figure 5.

The next step is to add a Gaussian field to the upsampled data. The field is generated following a random Gaussian distribution with carefully chosen parameters obtained from an optimization method. The random field should have appropriate amplitudes to represent the high-frequency details that are missing in the low-resolution data. Since the optical depth τ is positive-defined and from equation (27) it can be seen that F_T is restricted to assume values between 0 and 1, this makes a sampling in flux or optical depth complicated to perform. An alternative is to define the signal as a lognormal transformation from flux as follows.

According to equation (27) the optical depth is given as: $\tau = -\ln(F_T)$. And the optical depth

contrast as:

$$\delta\tau = \frac{\tau - \bar{\tau}}{\bar{\tau}} \tag{31}$$

with $\bar{\tau}$ is the mean of the optical depth field. Under the assumption that the optical depth follows a lognormal distribution, its underlying gaussian distribution is given by:

$$\delta\tau_G = \ln(1 + \delta\tau) - \mu \tag{32}$$

where μ is the mean of the $\delta\tau$ field:

$$\mu = \langle \ln(1 + \delta\tau) \rangle \quad . \tag{33}$$

Then $\delta\tau_G$ is gaussian distributed, can take positive and negative values making it the suitable option for sampling. Then, the field is added to $\delta\tau_G$ and is generated from a random Gaussian distribution with zero mean and a standard deviation σ_G dependent on the optical depth as:

$$\sigma_G = \sigma_0 \left(\frac{\tau}{\bar{\tau}} \right)^c \tag{34}$$

Where σ_0 and c are free parameters that were obtained from the implementation of a Markov Chain Monte Carlo optimization method. The values of $\sigma_0 = 0.7$ and $c = 1.1$ leads to an accuracy of 5% till $k = 1\,h\,\mathrm{Mpc}^{-1}$ in the 1D power spectrum as will be shown in chapter 4. To preserve the 3D correlations that arise from the data and guarantee that the addition of the small scales do not spoil the 3D power spectrum, the Hoffman-Ribak condition is applied (Hoffman-Ribak 1991). In practice this condition establishes that the addition of Gaussian noise must keep the mean of the values of $\delta\tau_G$ from the reference data.

3.3 Hamiltonian Monte Carlo Sampling

Bayesian Inference

Frequentist statistics focus on the analysis of data and making inferences based on the frequency or proportion of events. It is based on the principles of probability theory and assumes that the parameters in a statistical model are fixed but unknown. On the other hand, Bayesian statistics provides a framework for updating beliefs and making inferences about unknown quantities using prior knowledge and observed data. Bayesian framework is based on Bayes' Theorem, which relates a given data collection $\mathbf{d}$ with the signal one wants to improve $\mathbf{s}$ as:

$$P(\mathbf{s}|\mathbf{d}) = \frac{L(\mathbf{d}|\mathbf{s})}{E(\mathbf{s})} P(\mathbf{s}) \tag{35}$$

Where:

$P(\mathbf{s}|\mathbf{d})$ is the probability of the signal $\mathbf{s}$ given the data $\mathbf{d}$, and it is known as the posterior.

$L(\mathbf{d}|\mathbf{s})$ is the probability distribution of the data $\mathbf{d}$ given the signal $\mathbf{s}$. This is the model of the data and it is known as the likelihood.

$P(\mathbf{s})$ is the probability distribution of $\mathbf{s}$ before considering the evidence $E(\mathbf{s})$, and it is known as the prior.

$E(\mathbf{s})$ is the marginal probability of the data over the parameters, and it is known as the evidence, which is given by:

$$E(\mathbf{s}) = \int P(\mathbf{s})L(\mathbf{d}|\mathbf{s})d\mathbf{s} \tag{36}$$

In this context, $\mathbf{s}$ represents the unknown parameters of the model, and $\mathbf{d}$ represents the observed data. The theorem allows to estimate the posterior probability distribution $P(\mathbf{s}|\mathbf{d})$ given the observed data and the initial knowledge represented by the prior. The evidence acts as a normalization constant that can be neglected since the goal is to sample the posterior distribution from a given power spectrum (Kitaura et al. 2012b), then:

$$P(\mathbf{s}|\mathbf{d}) \propto L(\mathbf{d}|\mathbf{s})P(\mathbf{s}) \tag{37}$$

HMC Scheme

The HMC sampling introduced by Duane et al. (1987) allows to efficiently generate samples from a target distribution and its particularly well-suited for sampling in high-dimensional parameter space compared to traditional Markov Chain Monte Carlo (MCMC) methods like the Metropolis-Hastings algorithm. Apply the Hamiltonian dynamics to propose new samples allows to avoid random walk behavior and a more coherent exploration of the parameter space, leading to an acceptance rate of 100% with small deviations due to numerical errors. This results in an efficient and faster convergence to the target distribution.

The Hamiltonian represents the total energy of a system and provides the basis for analyzing and understanding the dynamics of the system within the phase-space, in which each point corresponds to a position $\mathbf{s}$ and momentum $\mathbf{p}$ for all particles in the system. The phase-space allows an analysis of the dynamics and system's behavior, such as the conservation of energy. In the HMC approach, the position $\mathbf{s}$ represents the signal one wants to improve moving in the

phase-space through the momenta $\mathbf{p}$. The Hamiltonian is defined as:

$$H(\mathbf{s}, \mathbf{p}) = K(\mathbf{p}) + E(\mathbf{s}) \tag{38}$$

$K(\mathbf{p})$ is the kinetic energy that can be defined as:

$$K(\mathbf{p}) \equiv \frac{1}{2} \sum_{ij} p_i M_{ij}^{-1} p_j \tag{39}$$

where M_{ij} is the Hamiltonian mass, a symmetric and positive-defined matrix which represents the covariance of the momenta, is the degree of freedom of the Hamiltonian sampler and its election is very relevant in the efficiency of the HMC. $E(\mathbf{s})$ is the potential energy which can be defined from the posterior distribution:

$$E(\mathbf{s}) = -\ln[P(\mathbf{s}|\mathbf{d})] = -\ln[L(\mathbf{d}|\mathbf{s})P(\mathbf{s})] \quad . \tag{40}$$

According to equation (32), the election of the signal as $\mathbf{s} = \delta\tau_G$ leads to the election of a gaussian prior for the HMC scheme:

$$P(s) = \frac{1}{\sqrt{(2\pi)_c^N \det C}} \exp\left(-\frac{1}{2} s^\dagger C^{-1} s\right) \tag{41}$$

which is a multivariate gaussian distribution with zero mean, N_c is the number of grid cells along the LOS and $C =< s^\dagger s >$ is the covariance matrix which is diagonal for a non-coupled modes. The likelihood represents the model of the data and is given by a Poisson/Gamma probability distribution, which considers that noise is independent between cells (Kitaura et al 2012b). In the particular case of this research, the data is the optical depth $\mathbf{d} = \tau$, and the likelihood:

$$L(\mathbf{d}|\lambda_i) = \prod_i \frac{\lambda_i^{N_i} e^{-\lambda_i}}{N_i!} \tag{42}$$

where λ_i is the expectation value of optical depth in the i cell and is related with the signal as:

$$\lambda_i = \bar{\tau} e^{b(\delta\tau_G + \mu)} \tag{43}$$

where b is a bias, in practice is a free parameter that can be estimated with an optimization method, and the mean μ can be obtained from the $\delta\tau_G$ field see the appendix A of Kitaura et al. (2012b):

$$\mu = -\ln(\langle \exp(\delta\tau_G) \rangle) \tag{44}$$

Once the prior and the likelihood have been defined, from equation (40) the potential energy of the Hamiltonian reads:

$$E(\mathbf{s}) = -\ln[P(\mathbf{s})] - \ln[L(\mathbf{d}|\mathbf{s})] = \frac{1}{2}\delta\tau_G^{\dagger} C^{-1}\delta\tau_G + \sum_i \lambda_i - N_i \ln(\lambda_i) + c \tag{45}$$

where c contains all the constant terms. To sample the posterior, the HMC is based on an analogy with the Hamiltonian dynamics, where the evolution of the system due to the momenta $\mathbf{p}$ allows to obtain new values of the position $\mathbf{s}$. The evolution of $\mathbf{s}$ and $\mathbf{p}$ over time are described with Hamilton's equations:

$$\frac{ds_i}{dt} = \frac{\partial}{\partial p_i} H(\mathbf{s}, \mathbf{p}) \tag{46}$$

$$\frac{dp_i}{dt} = -\frac{\partial}{\partial s_i} H(\mathbf{s}, \mathbf{p}) \tag{47}$$

From the definition of canonical distribution, the Hamiltonian can be related to the joint distribution function of the signal and the momenta as:

$$P(\mathbf{s}, \mathbf{p}) = \frac{1}{Z} e^{-H(\mathbf{s},\mathbf{p})} \tag{48}$$

with Z the partition function. The joint distribution function can be expressed as the product of two independent probability functions $P(\mathbf{s})$ and $P(\mathbf{p})$:

$$P(\mathbf{s}, \mathbf{p}) = P(\mathbf{s})P(\mathbf{p}) = \left[\frac{1}{Z_k} e^{-K(\mathbf{p})}\right]\left[\frac{1}{Z_E} e^{-E(\mathbf{s})}\right] \quad . \tag{49}$$

Taking into account equations (39) and (45), Hamilton's equations can be expressed as:

$$\frac{ds_i}{dt} = \frac{\partial}{\partial p_i} K(\mathbf{p}) = \sum_j M_{ij}^{-1} p_j \tag{50}$$

$$\frac{dp_i}{dt} = -\frac{\partial}{\partial s_i} E(\mathbf{s}) = C^{-1}\delta\tau_G + b(\bar{\tau} e^{b(\delta\tau_G + \mu)} - N_i) \tag{51}$$

Due to the conservation of energy, as $\mathbf{s}$ and $\mathbf{p}$ evolved with time the Hamiltonian is conserved $\dot{H} = 0$. However, in practice Hamilton's equations must be solved numerically and unavoidably an error is introduced. Therefore, the choice of the numerical method is relevant in the efficiency of the HMC. The Leapfrog discretization is a suitable option since it conserves the volume in the phase-space ensuring ergodicity. Furthermore, it has the characteristic of being time-reversible. In the Leapfrog scheme, the values of $\mathbf{s}$ and $\mathbf{p}$ for the next iteration are given by:

$$p_i\left(t + \frac{\epsilon}{2}\right) = p_i(t) - \frac{\epsilon}{2}\frac{\partial E(s)}{\partial s_i} \tag{52}$$

$$s_i\left(t+\epsilon\right) = s_i\left(t\right) + \epsilon M_{ij}^{-1} p_i\left(t+\frac{\epsilon}{2}\right) \tag{53}$$

$$p_i\left(t+\epsilon\right) = p_i\left(t+\frac{\epsilon}{2}\right) - \frac{\epsilon}{2}\frac{\partial E(s)}{\partial s_i} \tag{54}$$

During each time step ϵ, the Leapfrog scheme alternates between updating the positions and momenta of the particles. This involves computing the "force" which comes from the gradient of the potential energy acting on a particle based on its current position to update the momenta for a half step $\epsilon/2$ followed by updating the position based on the updated momenta and then other half step for the momenta. This process is repeated a number of time iterations N_{steps}. To avoid resonant trajectories and ensure an efficient exploration of the parameter space, ϵ and N_{steps} are randomly chosen.

Since the numerical approach introduces an error in the conservation of the Hamiltonian, and the initial signal has not been sampled from the correct distribution, it is necessary to introduce the Metropolis-Hastings acceptance criterion, in which the probability of acceptance of a new state (s', p') from an old state (s, p) is given by:

$$P_{accept} = \min\left[1, e^{-[H(s',p')-H(s,p)]}\right] \tag{55}$$

3.4 Estimation of the Completeness

Completeness is a type of selection effect that arises when some objects or structures in the universe are systematically undersampled or not detected by the survey due to various factors such as observational thresholds, instrumental limitations, or specific survey design. This introduces biases and distortions in the observed data. In the context of Lyα forest survey, the detection of many absorption features as possible allows to study statistical properties such as the distribution of density fluctuations. Then, completeness refers to the degree to which the survey has successfully identified and measured the properties of the absorption flux.

Different factors can affect the completeness of a Lyα forest survey, like the sensitivity of the observations, which determines the minimum detectable absorption strength. If the sensitivity is low, weak absorbers may go undetected, leading to lower completeness. Also, the presence of noise in the data and the specific data analysis techniques employed in the survey. An estimation of the completeness can be performed through simulations and mock data sets. By comparing synthetic absorption features with known properties with the observational data, mocks can assess the efficiency and limitations of the survey in recovering these features.

Then, modelling the completeness in the implementation of mock catalogs allows to identify

systematic and selection effects that can be used to calibrate the surveys. Let us define a model for the absorption flux following Kitaura et al. (2012b):

$$F_T^{\rm obs} = \mathbf{R}F_T + \epsilon_n \tag{56}$$

where $F_T^{\rm obs}$ is the observable absorption flux, ϵ is a random noise component and $\mathbf{R}$ is the response operator given by the completeness. $\mathbf{R}$ is in general a diagonal matrix and can be written as:

$$R_{ij} = w_i \delta_{ij}^{\rm K} \tag{57}$$

where δ_{ij}^K is the Kroenecker delta and w_i is the completeness in the i cell. The value of w_i can be calculated from the number of LOS intersecting that particular cell. Then, those cells in which no spectral line have been detected will have a completeness $w_i = 0$ and the cell which is crossed by the largest number of spectral lines will have $w_i = 1$ and is taken as a reference to establish the value of the completeness in the other cells.

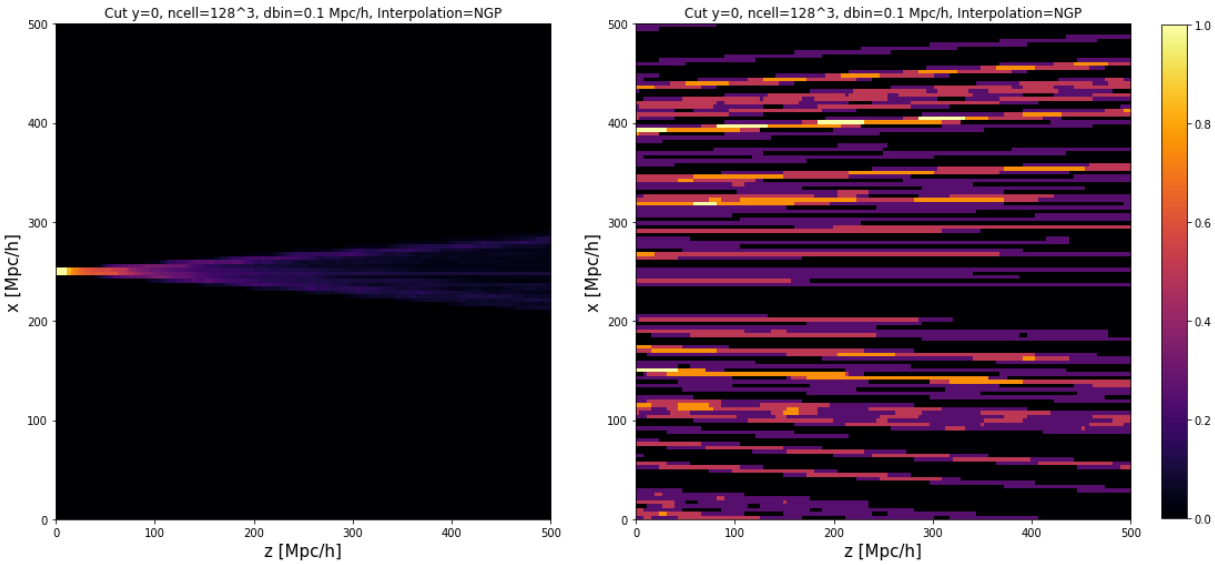

Figure 6: Model for the completeness in the Lyα Forest. Left: In the case of a simulation volume at redshift 0 with the observer at $x = 250\,{\rm Mpc}\,h^{-1}$ and $z = 0$. Right: Simulation volume at redshift 2.

A model for the completeness was estimated by randomly sampling LOS along a volume of 128^3 cells and applying the NGP interpolation method discussed in section 3.1 to estimate the number of counts in a cell caused by different spectra LOS. Figure 6 shows two slices at $y = 0$ of the 3D model of the completeness. When zooming in the $z = 0$ regions of the comoving volume (left panel of Figure 6), it can be seen how in the cells closest to the observer the number of counts saturates, making the completeness in other regions negligible.

4 RESULTS AND ANALYSIS

Different tests were performed to study the efficiency and accuracy of the Gaussian noise model and the Hamiltonian Monte Carlo scheme in sampling high resolution absorption flux. The training dataset was the absorption flux field from the cosmological SPH code GADGET3-OSAKA, in a box of 128^3 cells in a comoving volume of $500\,\mathrm{Mpc}\,h^{-1}$. From this box a NGP interpolation was applied to extract low-resolution skewers assuming a plane-parallel approximation, meaning that the observer is far enough to assume that the skewers are parallel to each other, and only intersect the cells along one of the dimensions of the box. Under this assumption completeness can be ignored. Then, the small scales are printed in the low resolution skewers. To fulfill the statistical properties of the Lyα forest, two quantities are of particular interest:

1. The mean 1D power spectrum $P(k)_{1D}$, the average power spectrum of each of the skewers.

2. The 3D power spectrum, obtained by averaging the modes in Fourier space in spherical shells.

The performed tests consisted of sampling skewers with twice the resolution of the reference data, that is, start from a resolution of $4\,\mathrm{Mpc}\,h^{-1}$ and reach a resolution of $2\,\mathrm{Mpc}\,h^{-1}$. As anticipated previously, the resolution upsampling is limited to a factor of 2 to minimize the use of computational resources and save computational time, since this work work is mainly meant to provide a proof of concept. To assess the accuracy of the applied methods in controlled conditions, the available absorption flux field from the reference simulation in the same comoving volume in a box of 256^3 cells were used as a reliable data to make a comparison. Then, the $P(k)_{1D}$ and $P(k)_{3D}$ of the sampled high resolution skewers were compared with the obtained from the box of 256^3 cells. Figure 7 shows the absorption flux along the same LOS from the low resolution reference box compared to the values from the high resolution box to illustrate the presence of small-scale fluctuations.

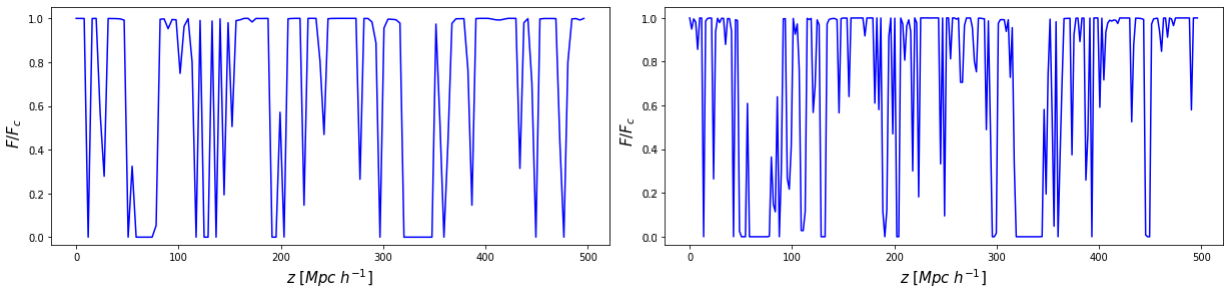

Figure 7: Skewers obtained in plane-parallel approximation from the same coordinates in the comoving volume of the reference simulation. Left: Skewer from low-resolution box (128^3). Right: Skewer from high resolution box (256^3). In can be seen the effect of small-scale fluctuations on the absorption flux

The optical depth along the line of sight can be estimated from the absorption flux. The exponential in equation (27) produces a non-linear relation between F_T and τ, which implies that an accurate reconstruction of the small scales in optical depth does not necessarily imply that the absorption flux are correctly reconstructed.

4.1 PDF of the absorption Flux and Optical Depth

The statistical properties of the Lyα forest can be analyzed by inspecting the probability density function (one-point statistics), the power spectrum (two-point statistics) and higher order statistics like the bispectrum (three-point statistics). The PDF of the Lyα forest provides valuable information about the underlying physical processes that shape the distribution of neutral hydrogen in the IGM. It characterizes the statistical behavior of the absorption lines, revealing the typical range of the absorption flux, the presence of peaks or troughs, and any other patterns or deviations from a smooth distribution.

Figure 8 shows the frequency distribution of the absorption flux values from the reference data at 128^3 cells and 256^3 cells. It can be seen that the PDF of the absorption fluxes is highly bimodal since the observations of the Lyα forest mostly comes from regions of low density where $F \simeq 1$ and regions of high density where $F \simeq 0$. This behavior makes it difficult to work out an analytical model for the PDF of the flux, and offers another reason in favor of choosing a lognormal transformation of the optical depth $\delta\tau_G$ as the candidate signal for the HMC scheme.

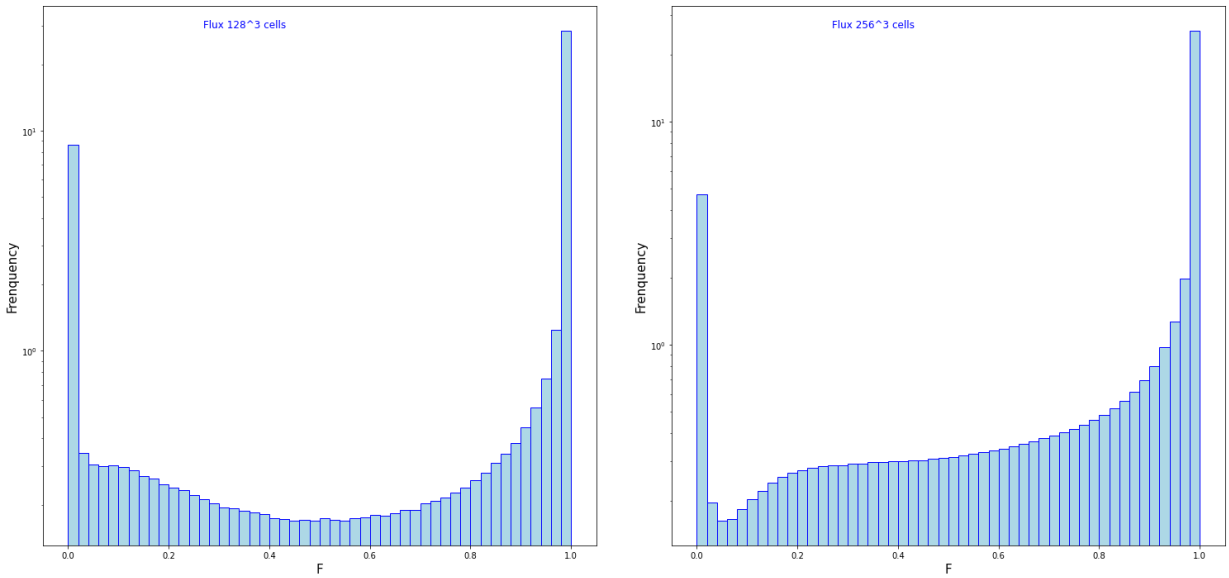

Figure 8: Frequency distribution of $\log(F_T)$. Left: Low resolution box. Right: High resolution box

Figure 9 shows the frequency distribution of the optical depth from the box of 256^3 cells. It can

be seen that more frequent values of τ are in the range of values $[0, 10^2]$, which are the values that determines the absorption flux. For $\tau \simeq 0$, $F_T \simeq 1$. For $\tau \simeq 10^2$, $F_T \simeq 10^{-44}$. Intermediate values of τ exhibit a linear frequency distribution up to higher values, the becoming less frequent. The frequency distribution of $\log_{10}(\tau)$ is shown in the right panel of figures 9. Taking the logarithm allows to appreciate in more detail the PDF of the optical depth, which exhibits a multimodal behavior with peaks at different values and the most relevant for the Lyα forest are between $-2 \leq \log(\tau) \leq 2$.

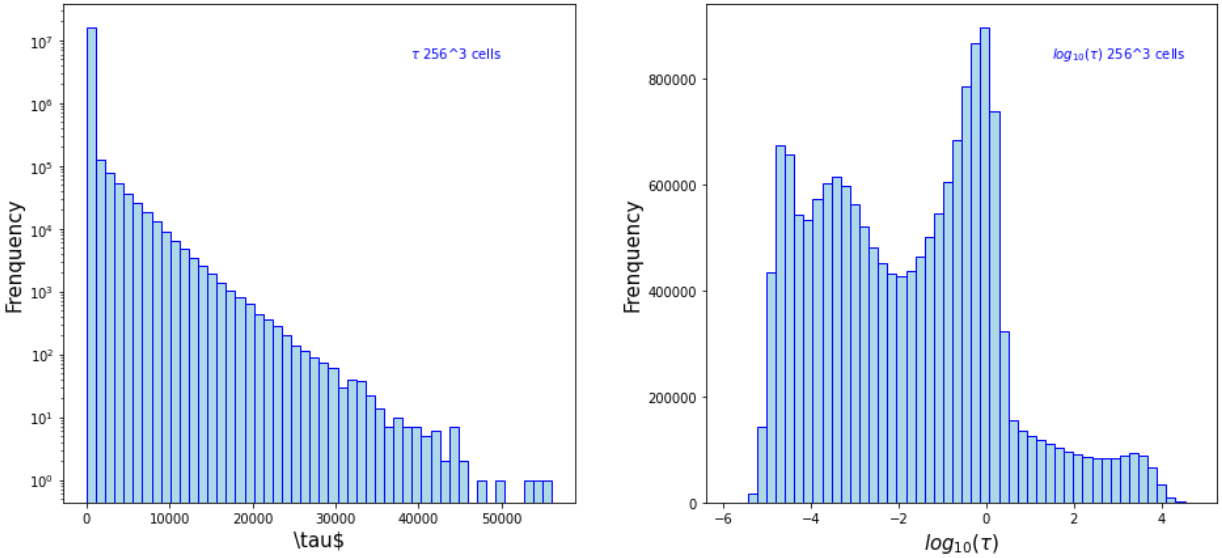

Figure 9: For high resolution box. Left: Frequency distribution of τ. Right: Frequency distribution of $\log(\tau)$

Lognormal Transformation

Due to the restriction in the values of the absorption flux F_T and the difficulty of modeling its frequency distribution, it turns out to be useful sampling the small scales taking as a training data the lognormal transformation of the optical depth $\delta\tau_G$ (equation (32)). The assumption that the optical depth follows a lognormal distribution has been applied in other investigations and generation of mocks of the Lyα forest (Font-Ribera et al. 2012; Lukić et al. 2015; Farr et al. 2020). Also a lognormal transformation with a Gaussian prior in the HMC scheme has been implemented in other studies (Kitaura & Angulo 2012; Hernández-Sánchez et al. 2021).

The efficiency of the HMC scheme is highly dependent on the definition of the PDF of the signal $\delta\tau_G$ which is related to the data τ according to equation (43):

$$\lambda_i = \bar{\tau} e^{b(\delta\tau_G + \mu)}$$

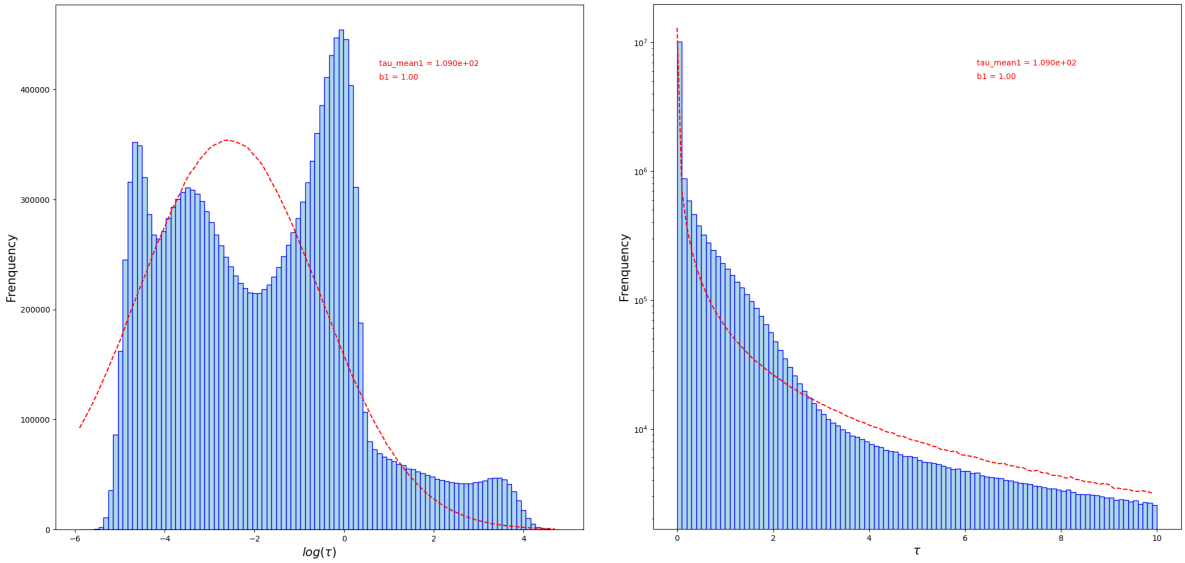

Figure 10: Lognormal model for the PDF of the optical depth. The values of the free parameters $\bar{\tau} = 109.0$ and $b = 1.0$ where obtained with the EMCEE optimization method. Left: The lognormal model in the PDF of $\log(\tau)$ corresponds to a Gaussian distribution. Right: Comparison between the lognormal model and frequency distribution from the data for the values of τ between $[0, 10]$.

The free parameters $\bar{\tau}$ and b were estimated using a Monte Carlo Markov Chain (MCMC) optimization method implemented in Python with the lognormal model described in chapter 3 and taking as baseline the PDF obtained from the optical depth of the reference simulation. The right panel of Figure 10 shows the lognormal model for $\bar{\tau} = 109.0$ and $b = 1.00$ compared with the PDF of τ from the reference simulation. It can be seen that this model is a good approximation for the values of τ between 0 and 10. The left panel shows the PDF of $\log(\tau)$ and the lognormal model with the same values of the free parameters, which corresponds to a Gaussian distribution. Figure 11 shows an accurate definition of the lognormal model with $\bar{\tau} = 350.0$ and $b = 0.78$, however this model produces and excess of probability in the values of τ greater than 1.

Several tests with the MCMC optimization method and the lognormal model revealed that a single lognormal distribution can accurately fit only a specific range of the PDF of τ. A bimodal or even a multimodal lognormal probability distribution may allow to obtain a more precise model for the PDF of the optical depth. Figure 12 shows the results of the optimization with a two lognormal model (equation (58)) and six free parameters $\bar{\tau_1} = 110.0$, $\bar{\tau_2} = 310.0$, $b_1 = 0.90$, $b_2 = 1.70$, $B_1 = 8.0$, $B_2 = 0.85$. This model accurately fits the frequency distribution of the values of τ that determine the absorption flux.

$$\log(\tau) = B_1 - \bar{\tau_1} \exp[b_1(\delta\tau_G + \mu_G)] + B_2 - \bar{\tau_2} \exp[b_2(\delta\tau_G + \mu_G)] \tag{58}$$

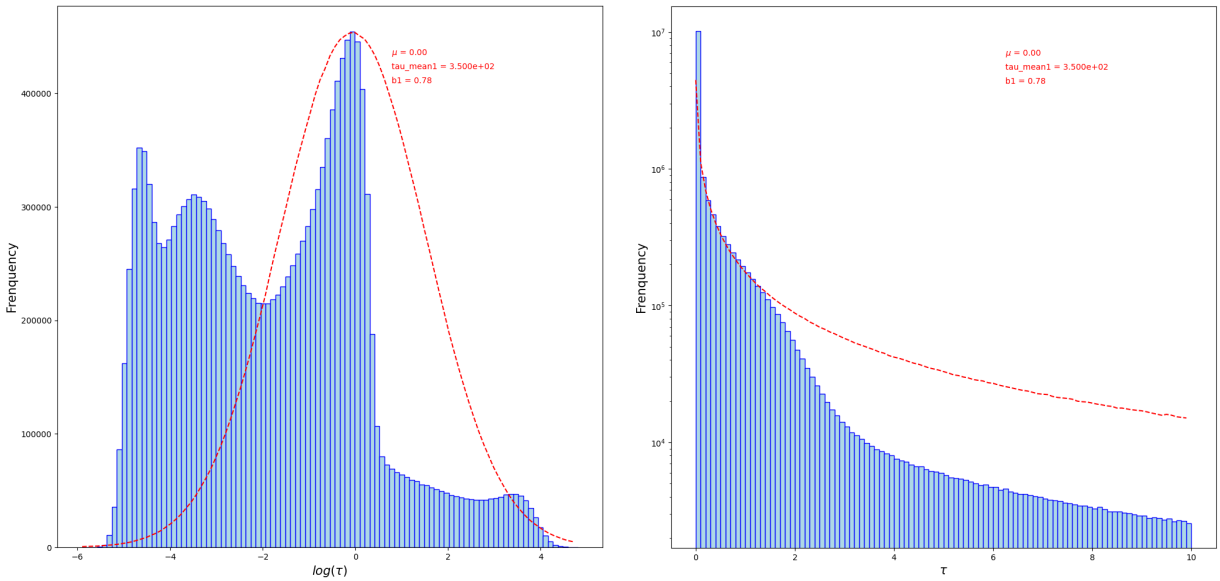

Figure 11: Lognormal model with $\bar{\tau} = 109.0$ and $b = 1.0$ obtained from the EMCEE optimization method. Left: PDF of $\log(\tau)$ which is a Gaussian distribution. Right: Comparison of data and model for the values of τ between $[0, 10]$.

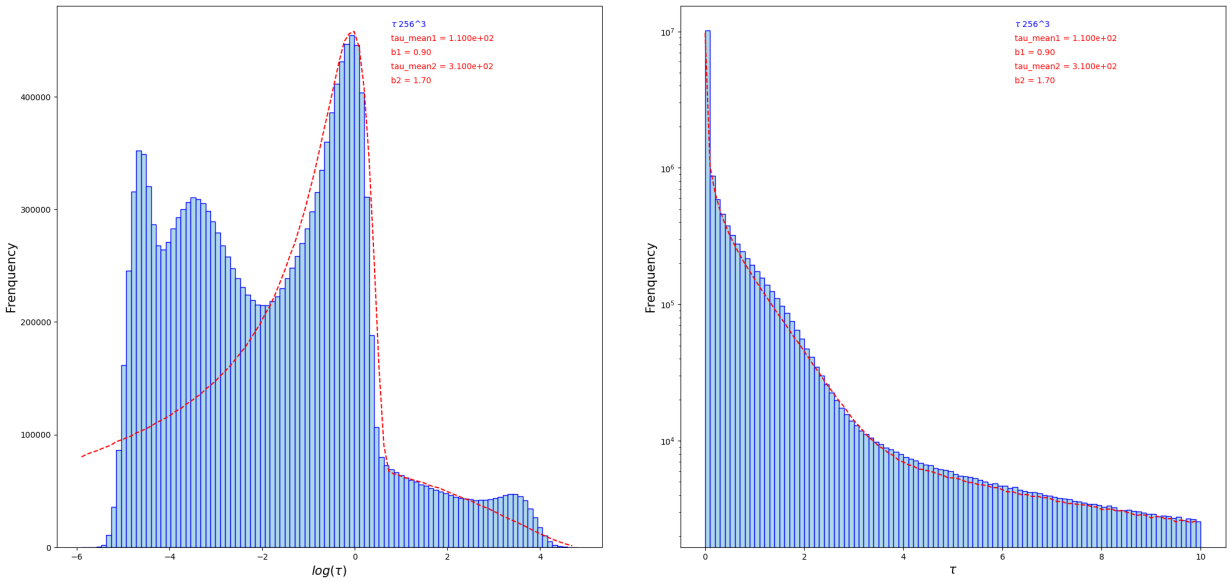

Figure 12: Two Lognormals model for the PDF of logarithm of the optical depth. The values of the free parameters $\bar{\tau_1} = 110.0$, $b_1 = 0.90$, $B_1 = 8.0$, $\bar{\tau_2} = 310.0$, $b_2 = 1.70$, $B_2 = 0.85$ where obtained with the EMCEE optimization method. Left: The two lognormals model in the PDF of $\log(\tau)$. Right: Comparison between the model and frequency distribution from the data for the values of τ between $[0, 10]$.

The definition of a multimodal lognormal approximation for the signal may be an even more

accurate model for the PDF of optical depth. However, the HMC scheme may be get stuck in one of the modes, producing an inefficient exploration of the parameter space. An improvement for future investigations consists on the implementation of a tempered HMC scheme, where an additional term is introduced into the Hamiltonian to enable a more effective exploration of parameter space.

4.2 Gaussian Noise Sampling Results

The method was implemented in Python, the skewers from the low resolution (LR) data were obtained with the method described in section 3.1 and stored in a collection of arrays of dimension $N_{\mathrm{LR}} = 128$. To increase the resolution of a skewer an upsampling is applied that consists of replicate each value of the low resolution array to create an array of dimension $N_{\mathrm{HR}} = 256$, in general $F_{\mathrm{HR}}[n * i : n * i + 1] = F_{\mathrm{LR}}[i]$ for the i element in the LR array and $n = 2$ the expansion factor. Figure 14 shows the 1D power spectrum $P(k)_{1\mathrm{D}}$ and the 3D power spectrum $P(k)_{3\mathrm{D}}$ of the optical depth after the upsampling compared with the $P(k)_{1\mathrm{D}}$ and $P(k)_{3\mathrm{D}}$ from the high resolution (HR) reference data.

Since the upsampling does not reconstruct the information encoded in the small scales, a lack of power can be seen at high k in the 3D power spectrum, which also has an effect in all scales in the 1D power spectrum. This is known as aliasing and can lead to systematic errors in the generation of mock catalogs and observational errors in the BAO measurement. Sampling small scales accurately allows to correct the aliasing and preserves the statistical properties of the absorption flux. The Gaussian Noise model consists in iteratively adding a HR resolution signal in $\delta\tau_G$ to correct the lack of power on small scales and applying the Hoffman-Ribak condition in each iteration to ensure that the addition of Gaussian noise does not change the 3D power spectrum.

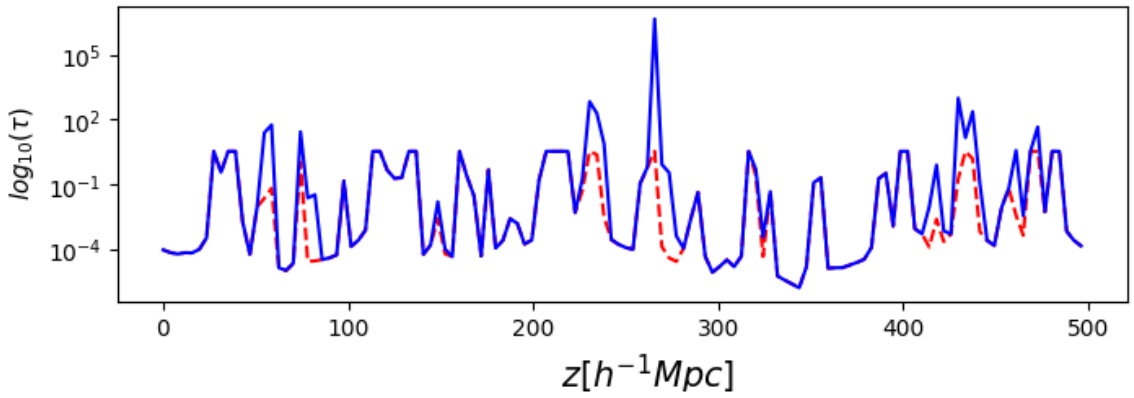

Figure 13:Values of $log(\tau)$ of a skewer from the HR reference data (red) and from the Gaussian noise sampling method after 20 iterations (blue).

It can be seen from figure 14 that aliasing produces a systematic aliasing in the 1D power spectrum that is around $\sim 0.37 \sim 1/\sqrt{8}$. In the iterative method applied to add the Gaussian noise, a normalization constant was also introduced to correct this bias. From different tests an estimation of the optimal values of the free parameters of equation (34) were found: $\sigma_0 = 0.9$ and $c = 1.3$.

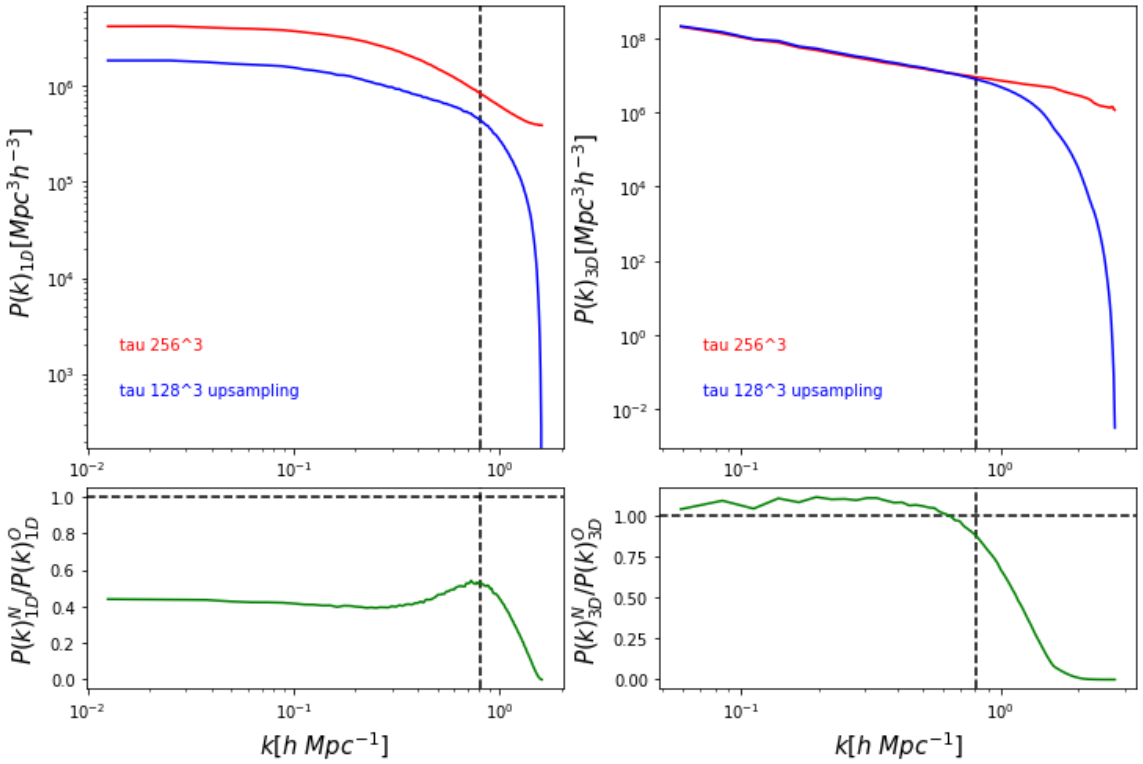

Figure 14: Comparison between the optical depth power spectrum of the HR reference data (red line) and the upsampled data (blue line). Left panels: The 1D power spectrum and the ratio between the $P(k)_{1D}$ of the upsampled data and the $P(k)_{1D}$ of the HR reference data. Right panels: The 3D power spectrum and the ratio of the upsampled and the HR reference data.

With this values of the free parameters the Gaussian noise model was applied to 128^2 skewers obtained in plane-parallel approximation from the LR data. Figure 16 shows the result of the sampling method in a random skewer compared with the same skewer obtained from HR reference data. It can be seen that the method is not entirely accurate and presents deviations to the values of optical depth in some regions. The $P(k)_{1D}$ and $P(k)_{3D}$ obtained from the 128^2 skewers are shown in figure 15. The model shows a precision of 5% in the $1D$ power spectrum at the scales between $0.01 \leq k \leq 1$.

In comparison with other studies in which only the $1D$ power spectrum of the absorption flux is studied and the effect of the sampled small scales in the $3D$ power spectrum is not explicitly

addressed (see e.g., see for instance Farr et al. 2020, where instead the two point correlation function is studied), the result achieved here demonstrates that the modelling of the small scales preserve a high accuracy also in the 3D power spectrum.

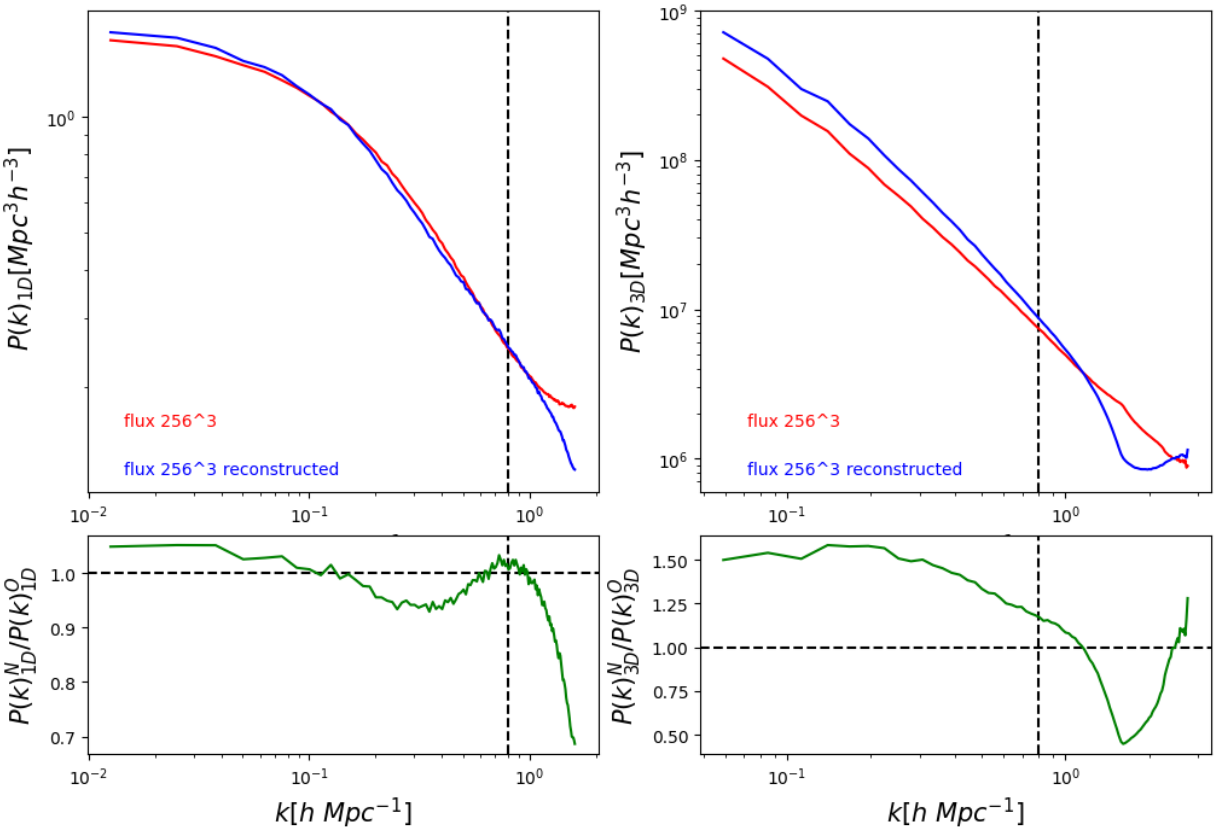

Figure 15: Comparison between the flux power spectrum of the HR reference data (red line) and the sampled skewers with the Gaussian noise model (blue line). Left panels: The 1D power spectrum and the ratio between the $P(k)_{1D}$ of the sampled skewers and the HR reference data. Right panels: The 3D power spectrum and the ratio of the $P(k)_{3D}$.

To verify that the method preserves the statistical properties of the flux absorption from the reference LR data, a downsampling of the sampled HR skewers was performed and the 1D and 3D power spectra were compared with those of the reference data (Figure 16). It can be seen an error of less than 5% in the 1D power spectrum and an error of around 10% in the 3D power spectrum. This method has the advantage of being computationally efficient and saves computational time since it does not require the calculation of expensive operations. This can be useful to sample at even higher resolutions, which is the case of the generation of mock catalogs for Lyα forest surveys.

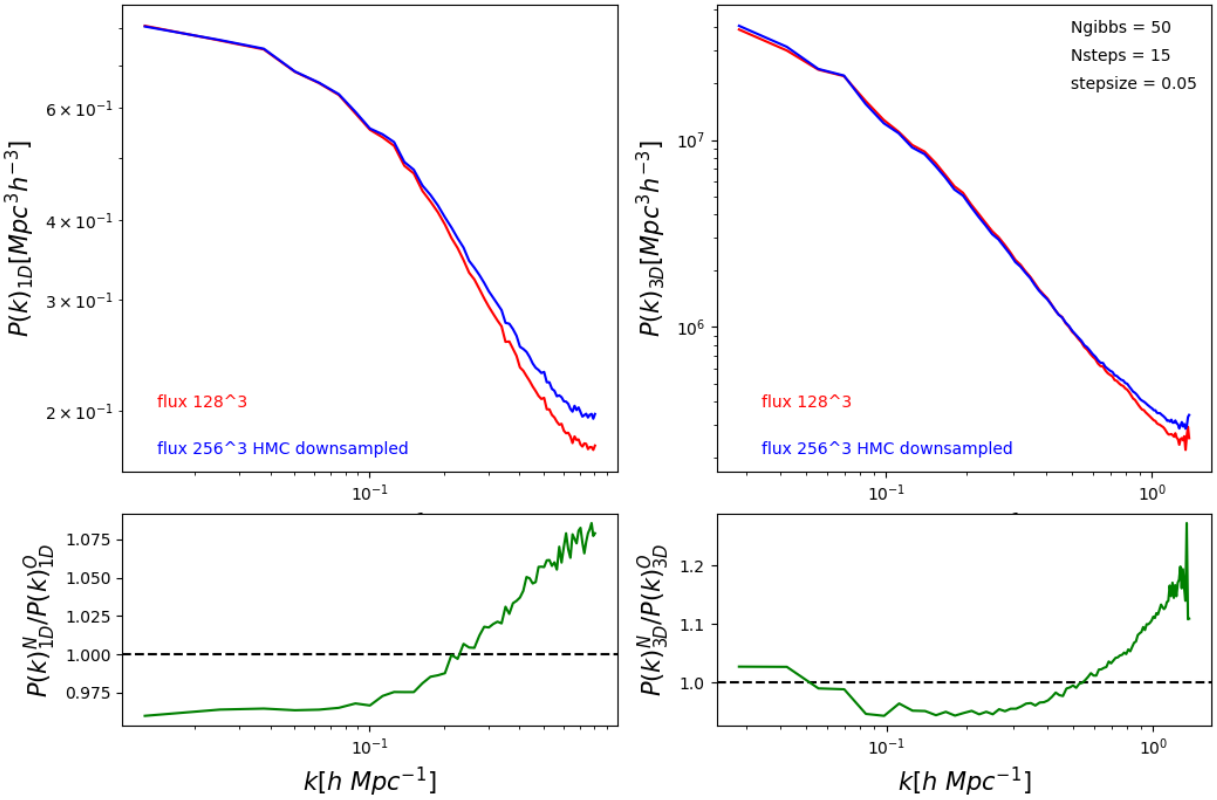

Figure 16: $1D$ and $3D$ power spectrum of the downsampled skewers for comparison

4.3 Hamiltonian Sampling Results

The Hamiltonian Markov Chain Montecarlo was implemented in Python based on the investigation from Hernández-Sánchez et al. (2021). The HMC scheme can perform a more precise sampling of the HR scales in comparison with the random Gaussian field sampling since it allows to include the information from the PDF of the optical depth in the data given the signal model (equation (43)). This relation is contained in the Poisson/Gamma likelihood which allows to have control over the errors introduced in the scheme.

In the same way, the definition of the prior according to the lognormal nature of the optical depth is a reasonable approximation for the small scales which are desired to sample. The Hamiltonian mass matrix from equation (41) which represents the covariance of the signal can be obtained from the 1D power spectrum of the HR reference data according to its relation from equation (10). In practice, the information extracted from the $P(k)_{1\mathrm{D}}$ is printed in the initial signal through a convolution in Fourier space. Then, with the definition of the posterior from the prior and the likelihood the scheme was performed to sample HR skewers and also, the Hoffman-Ribak

condition was applied in order to keep the 3D power spectrum from the LR data.

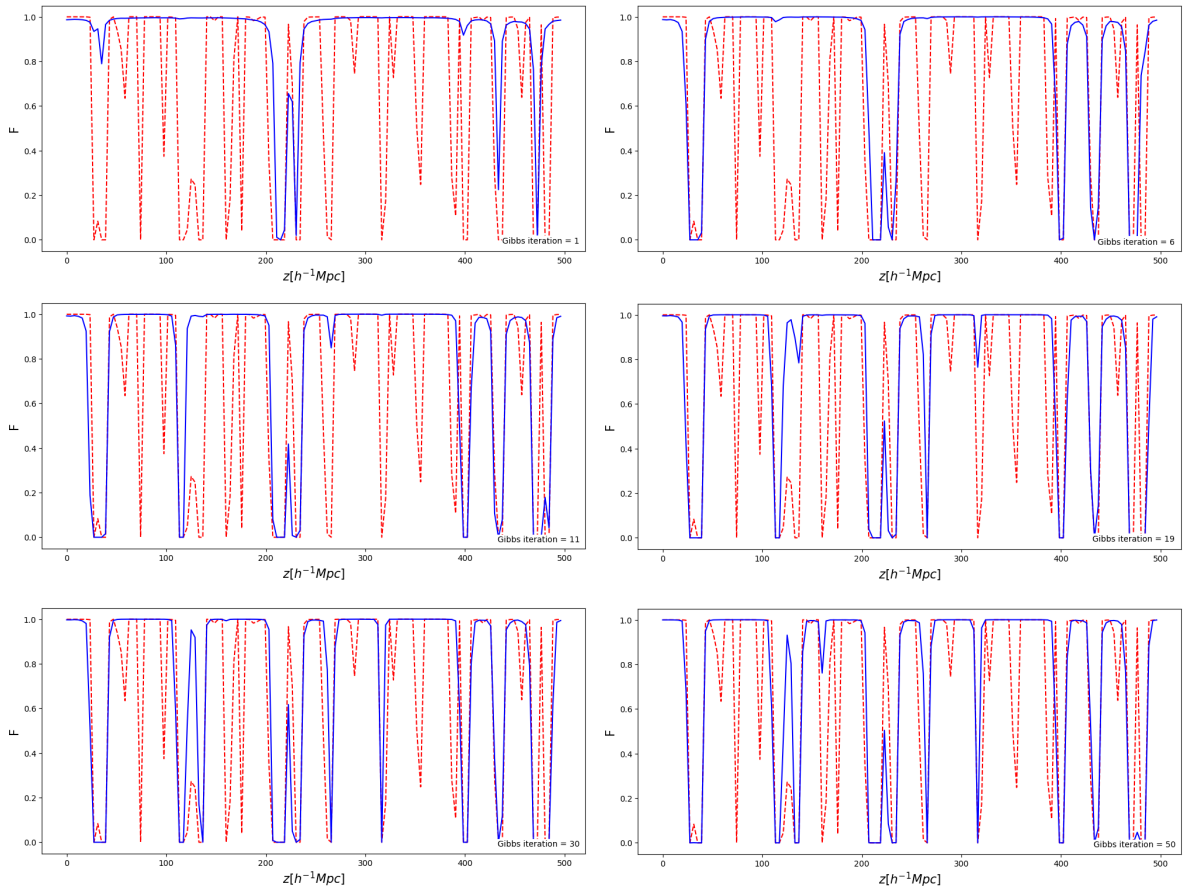

Figure 17: Sampling of an absorption flux skewer (blue line) with the HMC scheme an free parameters $\bar{\tau} = 109.0$ y $b = 1.0$ compared with the reference data (red line). From top to bottom: $it = 1$, $it = 6$, $it = 11$, $it = 19$, $it = 30$, $it = 50$.

First, different test were carried out to assess the efficiency of the HMC scheme in the reconstruction of a single HR skewer with the values of the free parameters of equation (43) obtained from the MCMC optimization method: $\bar{\tau} = 109.0$ and $b = 1.0$. Figure 17 shows the sampling of an absorption flux skewer at different iterations of the HMC. As can be seen the scheme reach convergence at around 50 iterations, also the sampled skewer accurately reproduces the flux values close to 0 and close to 1, while present deviations in intermediate values. This is due to the choice of the free parameters from the model of the PDF shown in figure 12, which approximately fits the low and high values of $\log(\tau)$ that corresponds to the flux values of F =1 and $F = 0$. However, it does not fit the intermediate values of $\log(\tau)$ which determine the flux between 0 and 1.

The accuracy of the results from the Hamiltonian sampling are very sensitive to the chosen value

of the time step ϵ and the number of evaluations N_{steps} of the leapfrog numerical integrator. A large value of ϵ can leads to an inefficient exploration of the parameter space while a small value of ϵ allows a detailed exploration of the parameter space. A low number of evaluations N_{steps} leads to unexplored regions, and a high N_{steps} results in increase the computational time, which is an important issue to take into account given the high number of skewers needed to be sampled.

N_{steps}	ϵ	iteration of convergence	MSE	Number of HMC evaluations
5	0.01	48	0.56	194
5	0.05	50	0.53	199
5	0.1	47	0.49	199
15	0.01	48	0.67	294
15	0.05	50	0.76	302
15	0.1	47	0.69	318
20	0.01	48	0.75	457
20	0.05	47	0.74	463
20	0.1	50	0.77	471

Table 1: Results from the performed tests to find the optimal values of N_{steps} and ϵ.

Different tests were doing to estimate the optimal values for ϵ and N_{steps}, two criteria were analyzed, the mean squared root (MSE) between the reference HR data and the sampled signal, and the number of evaluations of the HCM. The results are shown in table 1, it can be noticed that the number of iterations of convergence for all tests were around 50. To find an equilibrium between an optimal number of evaluations of the HMC and accuracy respect to the reference data, the scheme was run with the values of $N_{\text{steps}} = 15$ and $\epsilon = 0.05$.

To compensate the lack of power due to aliasing in the 1D power spectrum, a kernel was defined as the ratio between the power spectrum from the HMC sampled data and the obtained from the reference HR data:

$$K(k) = \frac{[P(k)_{1\text{D}}]_{\text{HMC}}}{[P(k)_{1\text{D}}]_{\text{ref}}} \tag{59}$$

The kernel was iteratively applied to the $P(k)_{1\text{D}}$ used to estimate the Hamiltonian mass matrix, which represents the variance of the signal and has an important effect in the coherency of the scheme keeping the statistical properties of the LR data. Figure 18 shows the effect of the kernel in the sampling of a skewer with the HMC scheme, note the improvement in the accuracy respect to the Gaussian noise model (Figure 13).

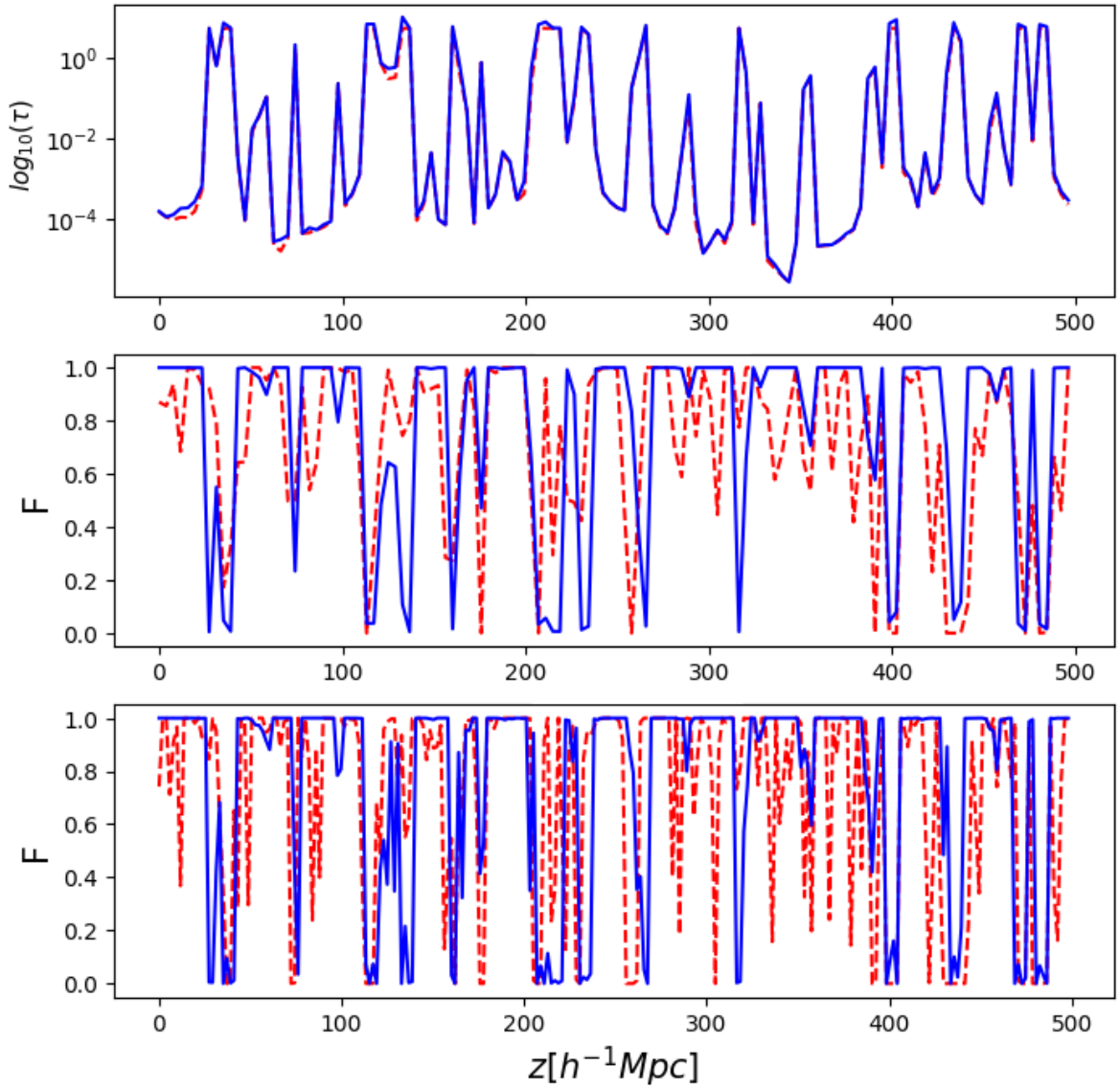

Figure 18: Sampling of a skewer with the HMC scheme and the Kernel application. Top: the values of $\log \tau$ along the skewer. Middle: The LR flux absorption for comparison. Bottom: HR flux absorption

Figure 19 shows the $P(k)_{1D}$ and $P(k)_{3D}$ of the absorption flux from 256^2 sampled skewers with the HMC scheme. The convergence is reached after 20 iterations, which represents an improvement with respect to the Gaussian sampler. The calculation of the gradients of the Hamiltonian and the implementation of the second order leapfrog numerical integrator leads to an increase in the accuracy in the reconstruction of the absorption flux but it requires more computational time, a factor of around 16 compared with the Gaussian noise method.

To study the effect of aliasing along one dimension, the scheme was applied to sampling 128^2 skewers in plane-parallel approximation from a resolution of 128 to 256 bins. To compare with the absorption flux of the LR reference data, a one-dimensional downsampling was performed. The systematic aliasing introduced by three-dimensional downsampling was studied in previous sections and requires a normalization factor of $\sqrt{8}$. While the performed tests point out to a normalization factor of $\sqrt{2}$ in the one-dimensional case. Then, to compare the LR reference data with the 1D downsampled data, a normalization factor of $\sqrt{8}/\sqrt{2} = 2$ is introduced in the 1D and 3D power spectrum (Figure 20).

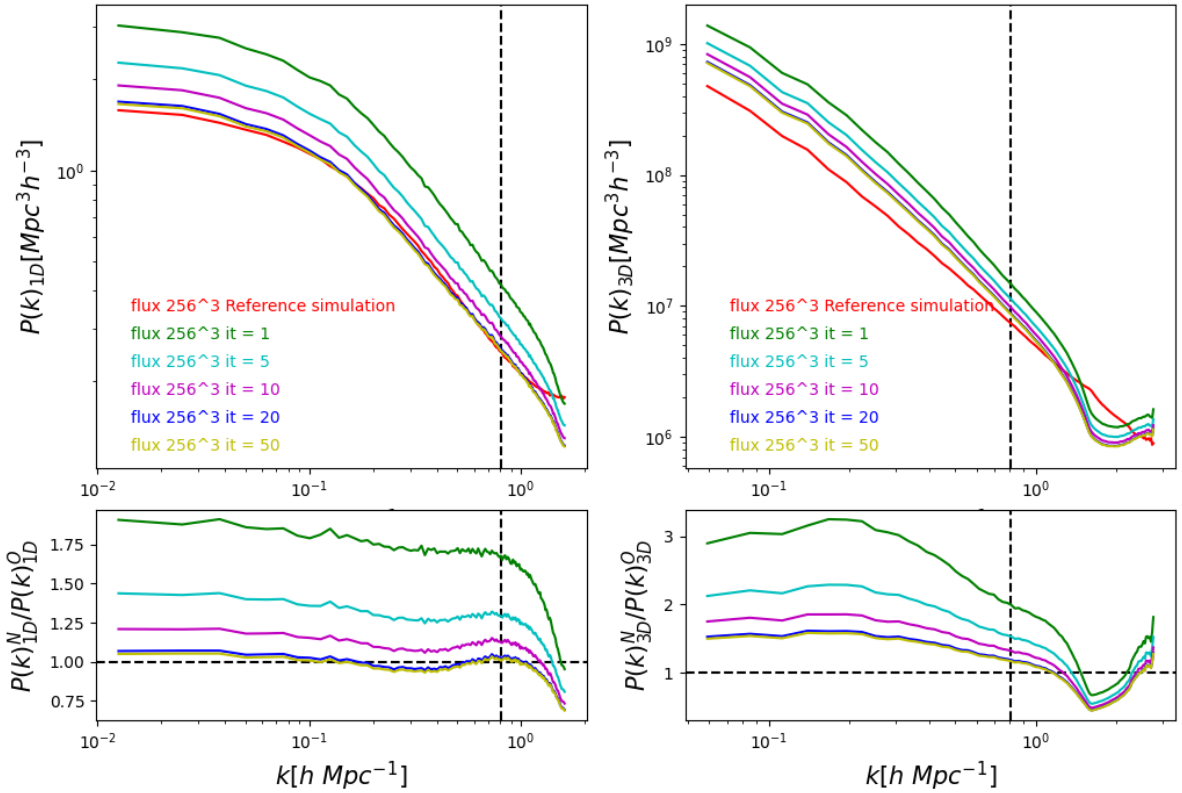

Figure 19: $P(k)_{1D}$ and $P(k)_{3D}$ of flux from the sampled skewers with the HMC scheme for 1, 5, 10, 20 and 50 iterations.

It can be seen that the sampling model reproduce the 1D and 3D power spectrum of the absorption flux from the reference simulation with high accuracy, with an error of around 5% up to scales of $k = 1\,\mathrm{Mpc}\,h^{-1}$. The achieved precision at these scales are comparable with the obtained in other investigations (Lukić et al. 2014; Farr et al. 2020; Walther et al. 2021; Chabanier et al. 2022), with the advantage that the sampling of the high-resolution scales takes into account the three-dimensional power spectrum $P(k)_{3D}$.

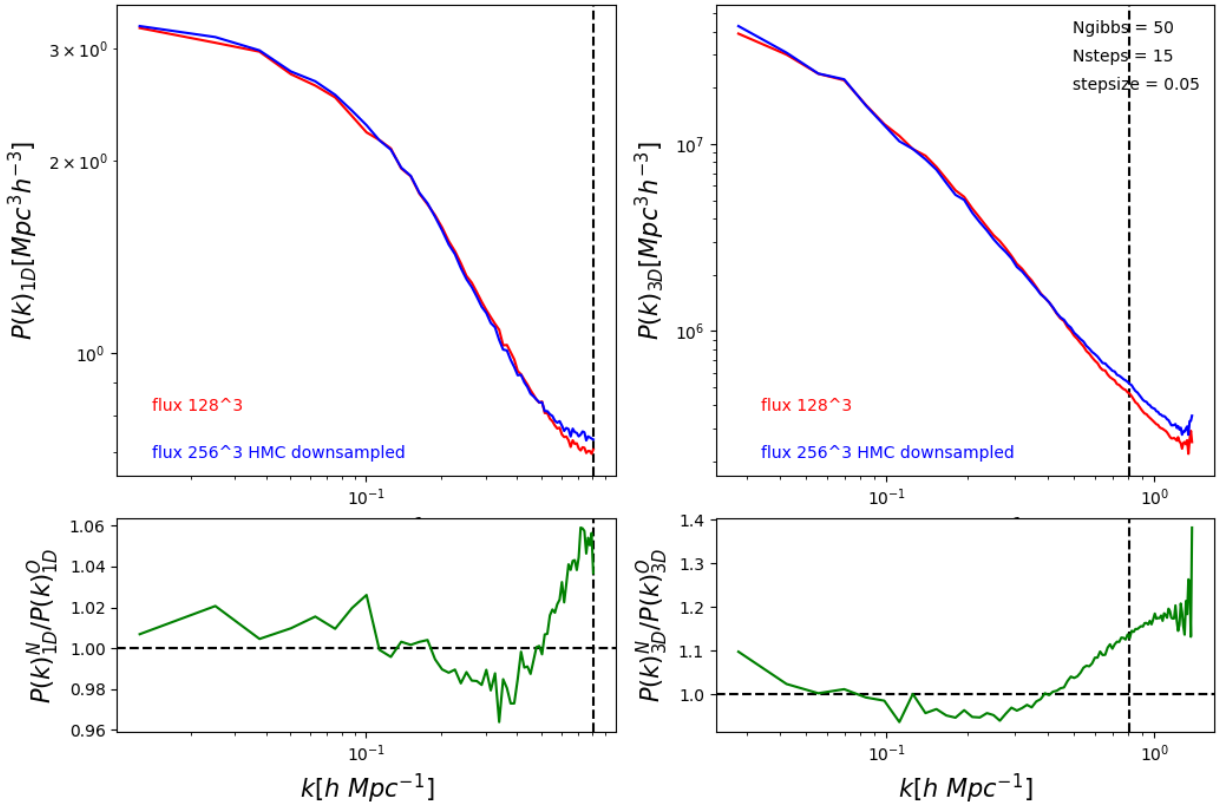

Figure 20: 1D and 3D power spectrum of the downsampled skewers from the HMC scheme

A relevant result of this investigation from several tests with the Hamiltonian sampler is that the Hoffman-Ribak condition can be neglected since the HMC is based on the model of the data given the signal from the study of the PDF of the optical depth, and the 1D power spectrum of the HR reference data. Then, the definition of a new sample preserves the statistical information of the low-resolution reference data.

Future investigations consists on apply the sampling methods at the required resolution of Lyα forest surveys which is around $40\,\mathrm{kpc}\,h^{-1}$, that means sampling high resolution skewers at ~ 12500 cells, which is unreachable for hydrodynamic or N-body simulations. The Gaussian noise and HMC scheme become more important due to their efficiency and low computational consumption. Improvements to the applied method include a better lognormal model for the optical depth PDF taking into account its multimodal nature, which in turn requires the implementation of a tempered Hamiltonian Monte Carlo scheme (see Graham & Storkey 2016).

5 CONCLUSIONS

In this investigation I described the fundamental aspects of the Lyα forest and its applications in the understanding of the properties of the intergalactic medium, the distribution of dark matter, the equation of state of dark energy and the expansion history of the universe through BAO measure. Given the cosmological importance of an accurate detection of the Lyα forest absorption flux the creation of mock catalogs at high resolution scales and considering all kind error sources are needed.

In this context, two different schemes to sample HR absorption flux from data of a reference hydrodynamic simulation were explored. To obtain skewers along the LOS from the flux field of the reference simulation, an interpolation method was described. All the performed tests point to the choice of a Nearest Grid Point (NGP) interpolation over a Cloud In Cell (CIC) in order to minimize the aliasing, an effect affecting the estimation of the $1D$ and $3D$ power spectra of the absorption flux. A selection effect was studied, the completeness, which constitutes one of the main sources of error in the detection of the Lyα forest and must be taken into account in the implementation of mock catalogs.

The LR resolution skewers obtained from the absorption flux field of the reference simulation were used as a training data for the implemented schemes, and a lognormal transformation was applied given the restrictions to the values of the absorption flux. The latter turned out to be a good approximation for the definition of a model for the probability density function of the optical depth. Improvements for future works includes the consideration of the multimodal nature of τ.

The first scheme consists on the addition of a Gaussian noise to the LR data. The results presented in Figure 17 allows to conclude that the scheme reaches a precision of $\sim 5\%$ in the reconstruction of the power spectrum 1D while it presents a greater error in the reconstruction of the power spectrum 3D. The advantage of this method is the low computational time required for its execution, which is a plus given the high number of skewers needed to be sampled ($\sim 256^2$) and more for the implementation of realistic mock catalogs of the Lyα forest. The performed tests shows a computational time of a factor 16 times lower than the required for the HMC scheme.

The HMC scheme requires a longer computational time for its execution, but it is a more accurate method is based on the Hamiltonian Mechanics analogy to explore the parameter space through the calculation of the gradients of the Hamiltonian using the Leapfrog numerical integrator. The

choice of a Gaussian prior with a lognormal transformation was justified by the statistical nature of the optical depth. On other hand, the definition of the data model given the signal from a Poisson/Gamma likelihood allows to reach a precision of $\sim 5\%$ in sampling skewers from a resolution of 128 cells to 256 cells. Future studies can be focused on exploring other models for the likelihood and the definition of the model given the signal, also applied to the sampling scheme at higher resolutions.

In this investigation, all the tests were focused on sampling skewers with a resolution of 256 cells, so the choice of a second order leapfrog over fourth order results more efficient given its advantage in sampling with small systems in the space of parameters. This allows to save computational time. The results of HMC scheme presented in Figure 20 show a better accuracy in the reconstruction of the $P(k)_{1\mathrm{D}}$ of the absorption flux, and also an improvement in the $P(k)_{3\mathrm{D}}$ with respect to the Gaussian noise scheme, which is under the $\sim 5\%$ and is comparable with the results of other studies focus on sampling high resolution skewers of the Lyα Forest. Also, Figure 18 shows an accurate reconstruction of the small scales, since the Hamiltonian sampler can make use of the information known from the PDF of optical depth.

Finally, the novelty of this investigation was the joint study of the 1D power spectrum and the 3D power spectrum of the Lyα forest absorption flux in order to accurately reproduce the statistical properties of the Lyα forest from the application of a Bayesian inference scheme based on the data from state-of-the-art hydrodynamic dark matter simulations. The analysis of the one-point and the two-point statistics leads to the generation of accurate mock catalogs for current Lyα forest forest surveys and contribute to the understanding of the large scale structure of the Universe.

References

[1] Aoyama, S. et al. (2018). Cosmological simulation with dust formation and destruction. Monthly Notices of the Royal Astronomical Society, 478(4), 4905–4921.

[2] Basset B. & Hlozek R. (2009). Baryon Acoustic Oscillations. Dark Energy: Observational and Theoretical Approaches. https://arxiv.org/abs/0910.5224

[3] Bond, J. R., & Wadsley, J. (1997). Lyα Absorption in the Cosmic Web. https://arxiv.org/pdf/astro-ph/9710102.pdf

[4] Chabanier, S. et al. (2022). Modelling the Lyman-α forest with Eulerian and SPH hydrodynamical methods. Monthly Notices of the Royal Astronomical Society, 518(3), 3754–3776.

[5] Duane, S., Kennedy, A., Pendleton, B. J., & Roweth, D. (1987). Hybrid Monte Carlo. Physics Letters B, 195(2), 216–222.

[6] Eisenstein D.J., et al. (2005). Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies. The Astrophysical Journal, 633(2), 560–574.

[7] Farr, J., et al. (2020). LyaCoLoRe: synthetic datasets for current and future Lyα forest BAO surveys. Journal of Cosmology and Astroparticle Physics, 2020(03), 068–068.

[8] Font-Ribera, A., et al. (2012). The large-scale cross-correlation of Damped Lyman alpha systems with the Lyman alpha forest: first measurements from BOSS. Journal of Cosmology and Astroparticle Physics, 2012(11), 059–059.

[9] Graham, M. & Storkey, J. (2016). Continuously Tempered Hamiltonian Monte Carlo. Conference on Uncertainty in Artificial Intelligence. https://arxiv.org/abs/1704.03338.

[10] Gunn, J. E., & Peterson, B. A. (1965). On the Density of Neutral Hydrogen in Intergalactic Space. The Astrophysical Journal, 142, 1633.

[11] Hernández-Sánchez, M., Kitaura, F. S., Ata, M., & Dalla Vecchia, C. (2021). Higher order Hamiltonian Monte Carlo sampling for cosmological large-scale structure analysis. Monthly Notices of the Royal Astronomical Society, 502(3), 3976–3992.

[12] Hoffman, Y., & Ribak, E. (1991). Constrained realizations of Gaussian fields - A simple algorithm. The Astrophysical Journal, 380, L5.

[13] Hui, L., & Gnedin, N. Y. (1997). Equation of state of the photoionized intergalactic medium. Monthly Notices of the Royal Astronomical Society, 292(1), 27–42.

[14] Kitaura, F. S., & Angulo, R. E. (2012). Linearization with cosmological perturbation theory. Monthly Notices of the Royal Astronomical Society, 425(4), 2443–2454.

[15] Kitaura, F. S., Gallerani, S., & Ferrara, A. (2012b). Multiscale inference of matter fields and baryon acoustic oscillations from the Lyα forest. Monthly Notices of the Royal Astronomical Society, 420(1), 61–74.

[16] Kitaura, F. S., & Heß, S. (2013). Cosmological structure formation with augmented Lagrangian perturbation theory. Monthly Notices of the Royal Astronomical Society: Letters, 435(1), L78–L82.

[17] Kitaura, F. S., Jasche, J., & Metcalf, R. B. (2010). Recovering the non-linear density field from the galaxy distribution with a Poisson-lognormal filter. Monthly Notices of the Royal Astronomical Society, 403(2), 589–604.

[18] Linder, E. V. (1997). First Principles of Cosmology. Pearson Education

[19] Lukić, Z. et al. (2014). The Lyα forest in optically thin hydrodynamical simulations. Monthly Notices of the Royal Astronomical Society, 446(4), 3697–3724.

[20] McDonald, P., et al. (2000). The Observed Probability Distribution Function, Power Spectrum, and Correlation Function of the Transmitted Flux in the Lyα Forest. The Astrophysical Journal, 543(1), 1–23.

[21] McDonald, P., & Eisenstein, D. J. (2007). Dark energy and curvature from a future baryonic acoustic oscillation survey using the Lyman-α forest. Physical Review D, 76(6).

[22] Peebles, P. J. E. (1980). The Large-Scale Structure of the Universe. Princeton University Press.

[23] Planck Collaboration, Aghanim, N., et al. (2020) Planck 2018 Results. VI. Cosmological Parameters. Astronomy & Astrophysics, 641, Article No. A6, 67 p.

[24] Rauch, M. (1998). The Lyman Alpha Forest in the Spectra of QSOs. Annual Review of Astronomy and Astrophysics, 36(1), 267–316.

[25] Riess, A. G. et al. (1998). Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant. The Astronomical Journal, 116(3), 1009–1038.

[26] Seo, H., & Eisenstein, D. J. (2003). Probing Dark Energy with Baryonic Acoustic Oscillations from Future Large Galaxy Redshift Surveys. The Astrophysical Journal, 598(2), 720–740.

[27] Seljak, U. et al (2005). Cosmological parameter analysis including SDSS Lyα forest and galaxy bias: Constraints on the primordial spectrum of fluctuations, neutrino mass, and dark energy. Physical Review D, 71(10).

[28] Shimizu, I. et al. (2019). Osaka feedback model: isolated disc galaxy simulations. Monthly Notices of the Royal Astronomical Society, 484(2), 2632–2655.

[29] Sinigaglia, F. et al. (2021). The Bias from Hydrodynamic Simulations: Mapping Baryon Physics onto Dark Matter Fields. The Astrophysical Journal, 921(1), 66.

[30] Sinigaglia, F. et al. (2022). Mapping the Three-dimensional Lyα Forest Large-scale Structure in Real and Redshift Space. The Astrophysical Journal, 927(2), 230.

[31] Walther, M. et al. (2021). Simulating intergalactic gas for DESI-like small scale Lyman α forest observations. Journal of Cosmology and Astroparticle Physics, 2021(04), 059.

[32] Wang, Y. (2006). Dark Energy Constraints from Baryon Acoustic Oscillations. The Astrophysical Journal, 647(1), 1–7.